

Master's thesis

Asymptotic preserving IMEX Active Flux schemes
for a hyperbolic approximation
of the Korteweg-de Vries equation



submitted by
Simon Paul Krotzsch
(2610636)

Supervisor: Prof. Dr. Christian Klingenberg

Mathematical Fluid Mechanics
Faculty of Mathematics and Computer Science
Julius-Maximilians-Universität Würzburg

Würzburg, September 4, 2026

In memory of Prof. Philip L. Roe

Kurzfassung

Diese Thesis beschäftigt sich mit dem Active-Flux-Verfahren, einem numerischen Verfahren, welches primär dazu entwickelt wurde, Erhaltungsgleichungen numerisch zu lösen. Dieses Verfahren besitzt zahlreiche positive Eigenschaften. Unter Anderem ist das klassische semi-diskrete Active-Flux-Verfahren mit einer Update-Gleichung der Punktwerte basierend auf Jacobian-Splitting angewendet auf eine hyperbolische Approximation der Wärmeleitungsgleichung asymptotic preserving. Hyperbolische Approximationen beinhalten einen Relaxationsparameter. Ein numerisches Verfahren wird asymptotic preserving genannt, wenn es für große Parameter numerische Lösungen der hyperbolischen Approximation und für signifikant kleine Lösungen der approximierten Gleichung liefert.

Der maßgebliche Beitrag dieser Arbeit besteht sowohl in der Entwicklung als auch in der theoretischen und numerischen Analyse neuartiger asymptotic preserving implizit-explizit (IMEX) Active-Flux-Verfahren zur numerischen Lösung der HKdV-Gleichungen, einer hyperbolischen Approximation der Korteweg-de Vries (KdV) Gleichung. Konkret wird, nach der Anwendung des klassischen semi-diskreten Active-Flux-Verfahrens auf die HKdV-Gleichungen, ein eng mit diesem verwandtes Classical-IMEX-Active-Flux-Verfahren entwickelt, welches als Zeitdiskretisierung IMEX-Runge-Kutta-Verfahren des Typs ARS verwendet. Um zu zeigen, dass dieses Verfahren asymptotic preserving ist, werden formale asymptotische Analyse und numerische Experimente durchgeführt.

Des Weiteren werden verschiedene Möglichkeiten entwickelt und untersucht, das zuvor entwickelte Classical-IMEX-Active-Flux-Verfahren durch die Verwendung alternativer Update-Gleichungen der Punktwerte zu verbessern. Zu diesen zählt sowohl die Weiterentwicklung des Classical-IMEX-Active-Flux-Verfahrens durch Eliminierung und Anpassung der Faktoren ausgewählter Terme in den Update-Gleichungen der Punktwerte, der Entwurf eines Verfahrens zweiter Ordnung basierend auf dem Central-Active-Flux-Verfahren und dessen Weiterentwicklung zu dritter Ordnung, als auch eine Kombination aus dem Central-IMEX-Active-Flux-Verfahrens dritter Ordnung und vorab festgelegter einseitiger Ausbreitungsrichtung der Informationen der partiellen Differentialgleichung, dem Biased-IMEX-Active-Flux-Verfahren. Die formale asymptotische Analyse und numerische Experimente dieser Verfahren zeigen, dass sie ebenfalls asymptotic preserving sind. Zudem deuten die numerischen Experimente darauf hin, dass das weiterentwickelte Classical-IMEX-Active-Flux-Verfahrens, das Central-IMEX-Active-Flux-Verfahrens dritter Ordnung und das Biased-IMEX-Active-Flux-Verfahren mit dem ARS(4, 4, 3)-IMEX-Runge-Kutta-Verfahren als Zeitdiskretisierung unabhängig vom Relaxationsparameter dritte Ordnung erreichen und damit asymptotic accurate sind. Vergleicht man diese Verfahren bezüglich Genauigkeit, Stabilität und Erhaltung einer diskreten Energie und Masse, erzielt das Biased-IMEX-Active-Flux Verfahren die besten Ergebnisse.

Durch diese Resultate bietet diese Arbeit eine Grundlage zur weiteren Entwicklung, Analyse und Optimierung von IMEX-Active-Flux-Verfahren für partielle Differentialgleichungen als auch deren hyperbolische Approximationen.

Abstract

The topic of this thesis is the Active Flux scheme, a numerical scheme primarily developed to numerically solve conservation laws. This scheme possesses several advantageous properties. For instance, the classical semi-discrete Active Flux scheme with an update equation for the point values based on Jacobian splitting applied to a hyperbolic approximation of the heat equation is asymptotic preserving. Hyperbolic approximations include a relaxation parameter. A numerical scheme is called asymptotic preserving if the scheme computes numerical solution to the hyperbolic approximation if this parameter is large and solution to the approximated equation if the parameter is small.

The main contribution of this thesis is the development as well as the theoretical and numerical analysis of novel IMEX Active Flux schemes for a hyperbolic approximation of the Korteweg-de Vries (KdV) equation, the HKdV equations. In detail, following the application of the classical semi-discrete Active Flux scheme to the HKdV equations, a closely related classical IMEX Active Flux is developed. This scheme employs an IMEX Runge-Kutta scheme of type ARS as a time discretisation. To demonstrate that this scheme is asymptotic preserving, formal asymptotic analysis as well as numerical experiments are performed.

Furthermore, different approaches are developed and analysed to improve the classical IMEX Active Flux scheme by using different update equations for the point values. This includes the further development of the classical IMEX Active Flux scheme by eliminating and rescaling selected terms in update equations of the point values, the design of a second order IMEX Active Flux schemes based on the central Active Flux scheme and the modification of this scheme to third order as well as a combination of imposing a directional bias in the update equations of the point values and the modified central IMEX Active Flux scheme, called the biased IMEX Active Flux scheme. Formal asymptotic analysis and numerical experiments indicate that the alternative IMEX Active Flux schemes are asymptotic preserving. In addition, the further developed classical IMEX Active Flux scheme, the modified central IMEX Active Flux scheme, and the biased IMEX Active Flux scheme achieve third order accuracy independent of the relaxation parameter and are, thus, asymptotic accurate. Comparing the schemes by measuring accuracy, stability, and conservation of a discrete energy and mass, the biased IMEX Active Flux scheme yields the best results.

The results of this thesis serve as a foundation for the further development, analysis, and optimisation of IMEX Active Flux schemes for partial differential equations as well as their hyperbolic approximation.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Dr. Christian Klingenberg not only for his constant and dedicated support and guidance throughout the writing of this thesis but also for the countless opportunities he has provided me, including the participation in the HYP2026 and SunHype2026 conferences as well as the privilege to meet and collaborate with outstanding scientists. His trust allowed and still allows me to grow both personally and professionally, and I am very much looking forward to continuing to work together.

I would like to express my sincere gratitude to Dr. Wasilij Barsukow for always taking his time to answer my questions and supporting me in the writing of this thesis with new inspirations and suggestions. I am excited to continue working together on many more projects.

In addition, I would like to thank the work group of Prof. Dr. Klingenberg for warmly welcoming me into the team and the many joyful experiences.

On a personal level, I wish to thank my parents and grandparents for their unconditional support and encouragement throughout my studies and my brother Julius for carefully proofreading parts of this thesis.

Contents

1	Introduction	1
2	Preliminaries	5
2.1	Hyperbolic systems of conservation laws	5
2.2	Korteweg-de Vries equation	7
2.3	One-dimensional semi-discrete Active Flux scheme	21
2.4	IMEX Runge-Kutta schemes	26
2.5	Asymptotic preserving numerical schemes	29
3	An asymptotic-preserving IMEX Active Flux scheme	32
3.1	Time-explicit Active Flux scheme	32
3.2	Classical IMEX Active Flux scheme	34
3.3	Formal asymptotic analysis	36
3.4	Implementation	51
3.5	Numerical experiments	54
3.6	Numerical results	57
4	Alternative asymptotic-preserving IMEX Active Flux schemes	67
4.1	Simplified IMEX Active Flux scheme	67
4.2	Optimised IMEX Active Flux scheme	74
4.3	Central IMEX Active Flux scheme	91
4.4	Modified central IMEX Active Flux scheme	101
4.5	Biased IMEX Active Flux scheme	112
5	Comparison	123
6	Summary, conclusion, and outlook	127
	Bibliography	128

1 Introduction

The Korteweg-de Vries (KdV) equation, which is usually stated today as

$$\frac{\partial}{\partial t}u(t, x) + u(t, x)\frac{\partial}{\partial x}u(t, x) + \frac{\partial^3}{\partial x^3}u(t, x) = 0,$$

is the result of the extensive study of long waves in shallow waters. The Scottish engineer John Scott Russel observed a special type of wave while he was following a boat along the Union Channel in 1834. In his 1845 article, "Report on Waves", he describes his observation in the following words [1]:

"I believe I shall best introduce this phenomenon by describing the circumstances of my own first acquaintance with it. I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped - not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation [...]"

According to [2] and [3], he later conducted various experiments to better comprehend the phenomena he had observed in the Union Channel. Through his experiments, he was able to empirically deduce that the volume of the water in the wave is equal to the volume of water displaced and that the speed of the wave, denoted by c , is given by

$$c^2 = g(h + a),$$

where a is the wave amplitude, h the undisturbed depth of the water, and g the acceleration of gravity. Joseph Boussinesq (1871) and Lord Rayleigh (1876) deduced independently from each other Russel's formula for the speed c from the equations of motion of an inviscid incompressible fluid. They showed that the wave profile, denoted by $u(t, x)$, where t is time and x is position, is given by

$$u(t, x) = a \operatorname{sech}^2(b(x - ct))$$

with

$$b^{-2} = \frac{4h^2(h + a)}{3a}$$

for any $a > 0$. Following their work, in 1895, Diederik Johannes Korteweg and his doctoral student Gustav de Vries stated a simple model partial differential equation, today known as the Korteweg-de Vries equation, which admits solutions of the form of u . In fact, the KdV equation in a different form had already appeared in an article by Boussinesq in 1877, before it was rediscovered by Korteweg and de Vries.

In the early 2010s, Philip L. Roe introduced the Active Flux method [4], [5], [6], a new numerical method to solve hyperbolic conservation laws. The Active Flux method is based on a numerical method for the linear advection equation in one spatial dimension presented in 1977 by Bram

van Leer [7]. Philip L. Roe was able to extend the idea of Bram van Leer to multiple spatial dimensions. In [6], he describes his motivation to develop a new method to be the shortcomings of contemporary, production-level numerical solvers used in computational fluid dynamics. The shortcomings of these methods are, in his opinion, the lack of true multi-dimensional physics, the second order accuracy, and the non-compact stencil.

Today, the Active Flux method is not a single method but a whole class of different methods, who share the same principals and ideas. To distinguish Roe's version of Active Flux from later version, it is commonly referred to as the traditional or fully-discrete Active Flux method. Philip L. Roe contributed during his life extensively to the development of the fully-discrete Active Flux method. His contributions include the extension of the fully-discrete Active Flux method to advection-diffusion problems [8], to higher order of accuracy [9], and a practical comparison between the fully-discrete Active Flux method and the Discontinuous Galerkin method [10]. His work also inspired others to further develop the traditional Active Flux method, for example in [11], [12], [13], and [14].

The essential part of the fully-discrete Active Flux method is an evolution operator. This is an operator that maps some given initial data to the exact or approximate solution of the problem at some given point in space and at some given time. For linear problems in one spatial dimension, it is straightforward to determine such an evolution operator, but finding even good approximate evolution operators for sufficiently complicated problems in multiple spatial dimensions remains a significant challenge.

A different approach is the generalised or semi-discrete Active Flux method, which does not depend on an evolution operator [15], [16], [17], [18]. Because of that, the semi-discrete version of Active Flux is more easily applicable to various partial differential equations. For example, in [19], the semi-discrete Active Flux method was used to numerically solve ideal magnetohydrodynamics.

The semi-discrete method itself allows for some freedom, especially in the semi-discrete update equations for the degrees of freedom. For example, in [20], an alternative approach to [18] is presented to derive the update equations for the point values. Additionally, it can be proven that a semi-discrete Active Flux method with certain update equations is in some sense equivalent to the Discontinuous Galerkin method [21]. Furthermore, the semi-discrete Active Flux method was extended to arbitrary order of accuracy [22] and proven to be stationarity preserving for the linear acoustic equations [23].

In [24], the authors investigated the asymptotic preserving properties of the semi-discrete Active Flux method. Therefore, they considered the hyperbolic heat equation in diffusive scaling,

$$\begin{aligned}\frac{\partial}{\partial t}p(t, \mathbf{x}) + \frac{1}{\varepsilon}\nabla \cdot u(t, \mathbf{x}) &= 0, \\ \frac{\partial}{\partial t}u(t, \mathbf{x}) + \frac{1}{\varepsilon}\nabla p(t, \mathbf{x}) + \frac{\sigma}{\varepsilon^2}u(t, \mathbf{x}) &= 0,\end{aligned}$$

where $\sigma > 0$ is some given parameter and $0 < \varepsilon < 1$ is called the relaxation parameter. In the limit $\varepsilon \rightarrow 0$, this system becomes

$$\begin{aligned}\frac{\partial}{\partial t}p(t, \mathbf{x}) &= \nabla \cdot \left(\frac{1}{\sigma} \nabla p(t, \mathbf{x}) \right), \\ u(t, \mathbf{x}) &= 0.\end{aligned}$$

It was shown that the third-order semi-discrete Active Flux method with a point value update based on Jacobian splitting applied to the one- and two-dimensional hyperbolic heat equation is without any modification asymptotic preserving. This means in simple terms that if $\varepsilon \rightarrow 0$, the Active Flux scheme for the hyperbolic heat equations automatically becomes a discretisation of the limit equation.

The general aim of this thesis is to develop asymptotic preserving implicit-explicit (IMEX) Active Flux schemes for a hyperbolic approximation of a non-linear partial differential equation, the Korteweg-De Vries equation. At the time of writing this thesis, there exists no literature about IMEX Active Flux schemes. Thus, this thesis aims to achieve several goals:

1. **Development of a classical IMEX Active Flux scheme for the hyperbolic approximation of the KdV equation:** Using time-explicit schemes to discretise the semi-discrete Active Flux method in time requires the temporal step size to scales with the relaxation parameter for the method to be stable. However, the Active Flux scheme should be able to compute numerical solutions of the hyperbolic relaxation for sufficiently small values of the relaxation parameter, which would result in an unnecessary high run-time and is therefore, in practise, not viable. The heat equation as well as its hyperbolic approximation are linear partial differential equations, and, hence, in [24] implicit time discretisation was used to overcome this problem. In case of a non-linear partial differential equation, implicit time integration would require solving non-linear systems of equations in each time step, and thus, in this thesis a different approach is followed, and implicit-explicit (IMEX) Runge-Kutta schemes are used to discretise in time. These schemes are not directly applicable to the classical semi-discrete Active Flux scheme based on Jacobian splitting applied to the HKdV equations. Therefore the first goal of this thesis is to develop an IMEX version of the classical semi-discrete Active Flux scheme for a hyperbolic approximation of the KdV equation.
2. **Theoretical and numerical analysis of the classical IMEX Active Flux scheme:** Since in [24], the Active Flux scheme for the hyperbolic approximation of the heat equations is without any modification asymptotic preserving, the second goal of this thesis is to investigate if the same holds for the classical IMEX Active Flux scheme. This includes the theoretical analysis of the scheme by performing formal asymptotic analysis as well as the implementation of the scheme to conduct various numerical experiments.
3. **Development and analysis of alternative IMEX Active Flux schemes:** In [24], the authors also considered a modified version of the Active Flux scheme with a different point value update equation. They observed that this modified scheme has preferable properties compared to the classical scheme. Thus, the third goal of this thesis is to develop and analyse IMEX Active Flux schemes with alternative update equations of the point values. This includes the theoretical analysis of these schemes by formal asymptotic analysis as well as the implementation of the schemes and the investigation of their properties through numerical experiments.
4. **Comparison of the IMEX Active Flux schemes with different point value update equations:** After introducing and analysing the alternative IMEX Active Flux schemes, the natural next step and last goal of this thesis is to compare the different schemes and identify the one with the most advantageous properties.

The thesis is structured in the following way: In the second chapter, important and necessary results are presented that are needed to develop and analyse the IMEX Active Flux schemes. The first and second goal of this thesis are covered in the third chapter, where, following a classical approach, an IMEX Active Flux scheme is derived, analysed, and implemented. In addition, numerical experiments are performed to further investigate the properties of the scheme. The third goal is addressed in the fourth chapter, in which several modified IMEX Active Flux schemes are developed and investigated numerically and theoretically. The fifth chapter summarises the findings in some sense by comparing the different schemes and determining the one with the most advantageous properties. In the last chapter, the findings of this thesis are concluded.

The numerical code used to produce the result presented in this thesis was written by the author of the thesis in the Julia programming language [25]. The following packages are utilised in the code: SparseArrays.jl [26], Interpolations.jl [27], QuadGK.jl [28], DataStructures.jl [29], LinearAlgebra.jl [30], and JLD2.jl [31]. The Figures in this thesis were created by the author with Julia and Makie.jl [32].

2 Preliminaries

This chapter presents the foundations that are needed to develop, analyse, and test the Active Flux schemes introduced later in this work. First, systems of conservation laws are defined in a general context, and after that the KdV equation as well as its hyperbolic approximation are established and some of their important properties are investigated. This includes travelling wave solution and conserved quantities. The last parts of this chapter are dedicated to the third-order semi-discrete Active Flux scheme in one-spatial dimension, IMEX Runge-Kutta schemes, and the definition of asymptotic preserving numerical schemes.

2.1 Hyperbolic systems of conservation laws

Following [33] and [34], we begin by considering $m \in \mathbb{N}$ conserved quantities $q_1, \dots, q_m$ that are functions of time $t \in [0, T)$ and space $\mathbf{x} \in \Omega$. The space $\Omega \subseteq \mathbb{R}^d$ with dimension $d \in \mathbb{N}$ is assumed to be a Lipschitz domain that is not necessarily bounded and the maximal time T is in $\mathbb{R}^+ \cup \{\infty\}$ with $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$. The conserved quantities are collected in the state vector $q : [0, T) \times \Omega \rightarrow \mathcal{O}$,

$$q(t, \mathbf{x}) = \begin{pmatrix} q_1(t, \mathbf{x}) \\ \vdots \\ q_m(t, \mathbf{x}) \end{pmatrix},$$

where we additionally required that $q(t, \mathbf{x}) \in \mathcal{O}$ and $\mathcal{O}$ is an open subset of $\mathbb{R}^m$. The integral

$$\int_D q(t, \mathbf{x}) d\mathbf{x}$$

for $t \in [0, T)$ and a bounded subset $D \subseteq \Omega$ describes the total mass of the conserved quantities in D at time t .

To continue, assume $f : \mathcal{O} \rightarrow \mathbb{R}^{m \times d}$ is a continuously differential function, which we call the flux-function. This allows us to define under which conditions the quantities q are called conserved: The quantities $q : [0, T) \times \Omega \rightarrow \mathcal{O}$ are conserved or obey a conservation law if

$$\int_D q(t_1, \mathbf{x}) d\mathbf{x} - \int_D q(t_0, \mathbf{x}) d\mathbf{x} + \int_{t_0}^{t_1} \int_{\partial D} f(q(t, \mathbf{x})) \mathbf{n}(\mathbf{x}) dS dt = 0 \quad (2.1)$$

holds for all times $0 \leq t_0 \leq t_1 < T$ and all bounded Lipschitz domains $D \subseteq \Omega$. The function $\mathbf{n} : \partial D \rightarrow \mathbb{R}^d$ is the outward unit normal to the boundary ∂D of D . For the moment, we assume that q is regular enough such that these integrals exist.

Remark 2.1. If $m = 1$ a conservation law is called scalar and if $m \geq 2$ it is referred to as a system of conservation laws.

To understand the physical meaning of (2.1), we rewrite it as

$$\int_D q(t_1, \mathbf{x}) d\mathbf{x} = \int_D q(t_0, \mathbf{x}) d\mathbf{x} - \int_{t_0}^{t_1} \int_{\partial D} f(q(t, \mathbf{x})) \mathbf{n}(\mathbf{x}) dS dt.$$

We can interpret this formulation in the following way: The total mass of q in D at time t_1 is equal to the total mass at time t_0 plus the amount of mass that flew into the domain during the time interval. This form of the conservation law is called the integral formulation, and it is perhaps the most natural form to write a conservation law since it is the mathematical formulation of the physical concepts of conservation.

However, in the literature, a conservation law is often written in its weak formulation:

$$\int_0^T \int_{\Omega} \left(q(t, \mathbf{x}) \cdot \frac{\partial}{\partial t} \phi(t, \mathbf{x}) + \sum_{j=1}^d \sum_{i=1}^m f_{ij}(q(t, \mathbf{x})) \frac{\partial}{\partial x_j} \phi_i(t, \mathbf{x}) \right) d\mathbf{x} dt = 0, \quad (2.2)$$

holds for all test functions $\phi \in C_c^\infty((0, T) \times \Omega; \mathbb{R}^m)$. The dot $\cdot$ denotes the standard scalar product in $\mathbb{R}^d$. For $q \in L^\infty((0, T) \times \Omega; \mathcal{O})$ it can be shown that the integral formulation and the weak formulation are equivalent. Thus, from here on, we assume that q is in $L^\infty((0, T) \times \Omega; \mathcal{O})$. If we impose more restrictions onto q , we can derive the strong formulation of a conservation law. Assume q is in $C^1((0, T) \times \Omega; \mathcal{O})$ and satisfies the conservation law in the weak formulation. Using the divergence theorem, we obtain

$$\int_0^T \int_\Omega \left(\frac{\partial}{\partial t} q(t, \mathbf{x}) + \nabla \cdot f(q(t, \mathbf{x})) \right) \cdot \phi(t, \mathbf{x}) d\mathbf{x} dt = 0,$$

for all $\phi \in C_c^\infty((0, T) \times \Omega; \mathbb{R}^m)$, where

$$\nabla \cdot f(q(t, \mathbf{x})) = \sum_{j=1}^d \frac{\partial}{\partial x_j} f_j(q(t, \mathbf{x}))$$

and

$$f_j(q(t, \mathbf{x})) = \begin{pmatrix} f_{1j}(q(t, \mathbf{x})) \\ \vdots \\ f_{mj}(q(t, \mathbf{x})) \end{pmatrix}.$$

Since the last equation holds for all functions $\phi \in C_c^\infty((0, T) \times \Omega; \mathbb{R}^m)$, this implies

$$\frac{\partial}{\partial t} q(t, \mathbf{x}) + \nabla \cdot f(q(t, \mathbf{x})) = 0 \quad (2.3)$$

holds for all $(t, \mathbf{x}) \in (0, T) \times \Omega$. This form is called the strong formulation of the conservation law. The converse of what we have just shown is also true:

Proposition 2.2. *A function $q \in C^1((0, T) \times \Omega; \mathcal{O})$ satisfies (2.2) for all $\phi \in C_c^\infty((0, T) \times \Omega; \mathbb{R}^m)$ if and only if it satisfies (2.3) for all $(t, \mathbf{x}) \in (0, T) \times \Omega$.*

Proof. The fact that a function $q \in C^1((0, T) \times \Omega; \mathcal{O})$ that satisfies (2.2) for all $\phi \in C_c^\infty((0, T) \times \Omega; \mathbb{R}^m)$ also satisfies (2.3) for all $(t, \mathbf{x}) \in (0, T) \times \Omega$ was shown above and the other implication follows applying the same arguments as used for the other implication but in reverse order. $\square$

Remark 2.3. One says that a function $q \in C^1((0, T) \times \Omega; \mathcal{O})$ satisfies (2.3) in the sense of distributions if (2.2) holds for all $\phi \in C_c^\infty((0, T) \times \Omega; \mathbb{R}^m)$.

The next definition states in which case a given system of conservation laws is called hyperbolic.

Definition 2.4. *Let $F_j : \mathcal{O} \rightarrow \mathbb{R}^{m \times m}$ be the Jacobian matrix of $f_j : \mathcal{O} \rightarrow \mathbb{R}^m$ for all $j = 1, \dots, d$. A system of conservation laws is called hyperbolic if, for all $q \in \mathcal{O}$ and any $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{R}^d$, $\alpha \neq 0$, the matrix*

$$A(q, \alpha) = \sum_{j=1}^d \alpha_j F_j(q)$$

has m real eigenvalues and m linearly independent corresponding eigenvectors. The system is called strictly hyperbolic if, in addition, all eigenvalues are distinct.

The Euler equations, formulated by L. Euler in the eighteenth century, which describe the motion of fluids, are perhaps the most prominent example of a system of hyperbolic conservation laws. Under the assumption that a fluid is described by (mass) density ρ , velocity u and pressure p , Euler derived what is today known as the barotropic compressible Euler system:

$$\frac{\partial}{\partial t} \rho(t, \mathbf{x}) + \nabla \cdot (\rho u)(t, \mathbf{x}) = 0,$$

$$\frac{\partial}{\partial t}(\rho u)(t, \mathbf{x}) + \nabla \cdot ((\rho u)(t, \mathbf{x}) \otimes u(t, \mathbf{x})) + \nabla p(\rho(t, \mathbf{x})) = 0,$$

where t denotes the time and $\mathbf{x}$ the space. The first equation describes the conservation of mass and the second the conservation of momentum.

The focus of this thesis are initial value problems to systems of hyperbolic conservation laws in one spatial dimension ($d = 1$). Assume $\Omega = \mathbb{R}$ and $q_0 \in L^\infty(\Omega; \mathcal{O})$ is a given function, called initial data. We want to find a function $q \in L^\infty((0, T) \times \Omega; \mathcal{O})$ that satisfies the conservation law and the initial condition

$$q(0, x) = q_0(x). \quad (2.4)$$

There exist different formulations of conservation laws and thus there also exist different notions of solution: We call a continuously differentiable function $q \in L^\infty([0, T] \times \overline{\Omega}; \mathcal{O})$ a strong solution to the initial value problem if q satisfies the strong formulation of the conservation law,

$$\frac{\partial}{\partial t}q(t, x) + \frac{\partial}{\partial x}f(q(t, x)) = 0$$

for all $(t, x) \in (0, T) \times \Omega$, and (2.4) pointwise almost everywhere on Ω . In contrast, we call a function $q \in L^\infty([0, T] \times \overline{\Omega}; \mathcal{O})$ a weak solution to the initial value problem if

$$0 = \int_0^T \int_\Omega \left(q(t, x) \cdot \frac{\partial}{\partial t}\phi(t, x) + f(q(t, x)) \cdot \frac{\partial}{\partial x}\phi(t, x) \right) dx dt + \int_\Omega q_0(x) \cdot \phi(0, x) dx$$

holds for all $\phi \in C_c^\infty([0, T] \times \overline{\Omega}; \mathbb{R}^m)$. For the relation between the two types of solutions, we have the following theorem:

Theorem 2.5. *A continuously differential function $q \in L^\infty([0, T] \times \overline{\Omega}; \mathcal{O})$ is a strong solution to the initial value problem if and only if it is a weak solution to the initial value problem.*

Proof. See [33]. □

2.2 Korteweg-de Vries equation

2.2.1 Hyperbolic approximation of the Korteweg-de Vries equation

In the literature, there exist different hyperbolic approximations of the KdV equation,

$$\frac{\partial}{\partial t}u(t, x) + u(t, x)\frac{\partial}{\partial x}u(t, x) + \frac{\partial^3}{\partial x^3}u(t, x) = 0.$$

For example, in [35], the authors present approximations to partial differential equations of the form

$$\frac{\partial}{\partial t}u(t, x) + u(t, x)\frac{\partial}{\partial x}u(t, x) - c_1\frac{\partial^2}{\partial x^2}u(t, x) - c_2\frac{\partial^3}{\partial x^3}u(t, x) = 0,$$

where $c_1 \in \mathbb{R}_0^+$ and $c_2 \in \mathbb{R} \setminus \{0\}$. Choosing $c_1 = 0$ and $c_2 = -1$, one obtains the KdV equation. In [36] a different approximation is presented:

$$\begin{aligned} \frac{\partial}{\partial t}u(t, x) + u(t, x)\frac{\partial}{\partial x}u(t, x) + \frac{\partial}{\partial x}\psi(t, x) &= 0, \\ \frac{\partial}{\partial t}\rho(t, x) - \frac{1}{\varepsilon^2}\frac{\partial}{\partial x}\rho(t, x) &= -\frac{\psi(t, x)}{\varepsilon^2}, \\ \frac{\partial}{\partial t}\psi(t, x) + \frac{1}{\varepsilon^2}\frac{\partial}{\partial x}u(t, x) &= \frac{\rho(t, x)}{\varepsilon^2}, \end{aligned}$$

where $0 < \varepsilon < 1$ denotes the relaxation parameter. In the following, this system is referred to as the hyperbolic KdV (HKdV) equations.

Remark 2.6. In [36], the authors write ε instead of ε^2 , however choosing ε^2 simplifies the later analysis of the numerical scheme.

As the name suggests, the HKdV equations are a hyperbolic system of conservation laws with an additional source term. Define $u, \rho, \psi : [0, T) \times \mathbb{R} \rightarrow \mathbb{R}$ to be the conserved quantities and $q : [0, T) \times \mathbb{R} \rightarrow \mathbb{R}^3$,

$$q(t, x) = \begin{pmatrix} u(t, x) \\ \rho(t, x) \\ \psi(t, x) \end{pmatrix}$$

to be the state vector. In addition, if the flux function $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is given by

$$f(q) := \begin{pmatrix} \frac{1}{2}(u)^2 + \psi \\ -\frac{1}{\varepsilon^2}\rho \\ \frac{1}{\varepsilon^2}u \end{pmatrix}$$

with $q = (u, \rho, \psi)^T$, we can rewrite the HkdV equation as

$$\frac{\partial}{\partial t} q(t, x) + \frac{\partial}{\partial x} f(q(t, x)) = S q(t, x),$$

where

$$S = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\varepsilon^2} \\ 0 & \frac{1}{\varepsilon^2} & 0 \end{pmatrix}.$$

To show that the homogeneous part of the system is a hyperbolic system of conservation laws in the strong formulation, we determine the Jacobian F of f , which reads

$$F(q) = \begin{pmatrix} u & 0 & 1 \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{\varepsilon^2} & 0 & 0 \end{pmatrix}$$

for $q \in \mathbb{R}^3$. With the help of Mathematica [37], the eigenvalues and eigenvectors of $F(q)$ can be computed. The eigenvalues are

$$\lambda_1(q) = -\frac{1}{\varepsilon^2}, \quad \lambda_2(q) = \frac{1}{2} \left(u - \sqrt{u^2 + \frac{4}{\varepsilon^2}} \right), \quad \lambda_3(q) = \frac{1}{2} \left(u + \sqrt{u^2 + \frac{4}{\varepsilon^2}} \right).$$

and the eigenvectors are

$$v_1(q) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad v_2(q) = \begin{pmatrix} \frac{\varepsilon^2}{2} \left(u - \sqrt{u^2 + \frac{4}{\varepsilon^2}} \right) \\ 0 \\ 1 \end{pmatrix}, \quad v_3(q) = \begin{pmatrix} \frac{\varepsilon^2}{2} \left(u + \sqrt{u^2 + \frac{4}{\varepsilon^2}} \right) \\ 0 \\ 1 \end{pmatrix}.$$

The eigenvectors and eigenvalues depend only on u and not on ρ or ψ . Moreover, they are real numbers for all choices of $\varepsilon > 0$. Since

$$\underbrace{u - \sqrt{u^2 + \frac{4}{\varepsilon^2}}}_{>0} < \underbrace{u + \sqrt{u^2 + \frac{4}{\varepsilon^2}}}_{>0},$$

it follows that

$$a_1 v_1(q) + a_2 v_2(q) + a_3 v_3(q) = (0, 0, 0)^T$$

with $a_1, a_2, a_3 \in \mathbb{R}$ implies $a_1 = 0$, $a_2 = a_3$, and

$$a_2 \left(u - \sqrt{u^2 + \frac{4}{\varepsilon^2}} \right) = a_3 \left(u - \sqrt{u^2 + \frac{4}{\varepsilon^2}} \right).$$

Hence, $a_1 = a_2 = a_3 = 0$ and the eigenvectors of F are linearly independent for all $q \in \mathbb{R}^3$ and all $\varepsilon > 0$. Thus, the homogeneous part of the HKdV equations is a hyperbolic system of conservation laws.

With formal asymptotic analysis, we can study the limit of the HKdV equations as $\varepsilon \rightarrow 0$ to investigate whether the HKdV equations approximate the KdV equation. Proceeding as in [24], we assume $u, \rho, \psi : [0, T] \times \Omega \rightarrow \mathbb{R}$ satisfy the HKdV equations on $(0, T) \times \Omega$ and, in addition, are regular enough, such that they can be expanded as power series in ε , i.e.

$$\begin{aligned} u(t, x) &= u^{(0)}(t, x) + \varepsilon u^{(1)}(t, x) + \varepsilon^2 u^{(2)}(t, x) + \dots, \\ \rho(t, x) &= \rho^{(0)}(t, x) + \varepsilon \rho^{(1)}(t, x) + \varepsilon^2 \rho^{(2)}(t, x) + \dots, \\ \psi(t, x) &= \psi^{(0)}(t, x) + \varepsilon \psi^{(1)}(t, x) + \varepsilon^2 \psi^{(2)}(t, x) + \dots, \end{aligned}$$

where $u^{(k)}, \rho^{(k)}, \psi^{(k)} : [0, T] \times \Omega \rightarrow \mathbb{R}$ are functions that are also assumed to be regular. Such an expansion is sometimes referred to as Chapman-Enskog expansion [38] or Hilbert expansion [39]. By inserting these assumptions into the HKdV equations, we obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial}{\partial x} \psi^{(0)}(t, x) \\ &\quad + \varepsilon \left(\frac{\partial}{\partial t} u^{(1)}(t, x) + u^{(1)}(t, x) \frac{\partial}{\partial x} u^{(1)}(t, x) + \frac{\partial}{\partial x} \psi^{(1)}(t, x) \right) \\ &\quad + \dots, \end{aligned}$$

and

$$\begin{aligned} 0 &= \frac{1}{\varepsilon^2} \left(-\frac{\partial}{\partial x} \rho^{(0)} + \psi^{(0)}(t, x) \right) + \frac{1}{\varepsilon} \left(-\frac{\partial}{\partial x} \rho^{(1)} + \psi^{(1)}(t, x) \right) \\ &\quad + \frac{\partial}{\partial t} \rho^{(0)}(t, x) - \frac{\partial}{\partial x} \rho^{(2)} + \psi^{(2)}(t, x) + \dots, \end{aligned}$$

as well as

$$\begin{aligned} 0 &= \frac{1}{\varepsilon^2} \left(\frac{\partial}{\partial x} u^{(0)} - \rho^{(0)}(t, x) \right) + \frac{1}{\varepsilon} \left(\frac{\partial}{\partial x} u^{(1)} - \rho^{(1)}(t, x) \right) \\ &\quad + \frac{\partial}{\partial t} \psi^{(0)}(t, x) + \frac{\partial}{\partial x} u^{(2)} - \rho^{(2)}(t, x) + \dots \end{aligned}$$

We aim to find equations that the functions $u^{(0)}, \rho^{(0)}$, and $\psi^{(0)}$ satisfy in the limit $\varepsilon \rightarrow 0$. Therefore, we order the terms after the power of ε and obtain

$$\begin{aligned} \varepsilon^{-2} : \quad &\frac{\partial}{\partial x} \rho^{(0)}(t, x) = \psi^{(0)}(t, x) \\ &\frac{\partial}{\partial x} u^{(0)}(t, x) = \rho^{(0)}(t, x) \\ \varepsilon^{-1} : \quad &\frac{\partial}{\partial x} \rho^{(1)}(t, x) = \psi^{(1)}(t, x) \\ &\frac{\partial}{\partial x} u^{(1)}(t, x) = \rho^{(1)}(t, x) \\ \varepsilon^0 : \quad &\frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial}{\partial x} \psi^{(0)}(t, x) = 0 \end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial t}\rho^{(0)}(t, x) - \frac{\partial}{\partial x}\rho^{(2)}(t, x) &= -\psi^{(2)}(t, x) \\ \frac{\partial}{\partial t}\psi^{(0)}(t, x) + \frac{\partial}{\partial x}u^{(2)}(t, x) &= \rho^{(2)}(t, x) \\ \varepsilon^1 : \dots\end{aligned}$$

Thus, the leading order functions, i.e. the functions $q^{(0)}$, $\rho^{(0)}$ and $\psi^{(0)}$, satisfy

$$\begin{aligned}\frac{\partial}{\partial t}u^{(0)}(t, x) + u^{(0)}(t, x)\frac{\partial}{\partial x}u^{(0)}(t, x) + \frac{\partial^3}{\partial t^3}u^{(0)}(t, x) &= 0, \\ p^{(0)}(t, x) &= \frac{\partial}{\partial x}u^{(0)}(t, x), \\ \psi^{(0)}(t, x) &= \frac{\partial^2}{\partial x^2}u^{(0)}(t, x).\end{aligned}$$

The first equation of this system is the KdV equation and, thus, the function $u^{(0)}$ satisfies the KdV equation, while $\rho^{(0)}$ and $\psi^{(0)}$ are the first and second derivative of $u^{(0)}$. This system is the formal limit of the HKdV equations as $\varepsilon \rightarrow 0$ and, therefore, we call it the limit equations of the HKdV equations.

2.2.2 Travelling wave solutions of the KdV equation

As stated in the introduction, the KdV equation allows for solutions of the form

$$u(t, x) = a \operatorname{sech}^2(b(x - ct)).$$

Due to their behaviour, such solutions are called travelling wave solutions. More generally, travelling wave solutions can be defined as in [40]:

Definition 2.7. Consider a partial differential equation involving the two variables $x \in \mathbb{R}$ and $t \in \mathbb{R}^+$. A function u that satisfies the partial differential equation and is of the form

$$u(t, x) = \tilde{u}(x - ct),$$

is called a travelling wave solution with speed $c \in \mathbb{R}$ and profile $\tilde{u}$.

Next, following [40], a special type of travelling wave solutions for the KdV equation is derived. A continuously differentiable function $u : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$, which can be written as $u(t, x) = \tilde{u}(x - ct)$ with $\tilde{u} : \mathbb{R} \rightarrow \mathbb{R}$ satisfies the KdV equation if $\tilde{u}$ satisfies the ordinary differential equation

$$-c\tilde{u}'(\xi) + \tilde{u}(\xi)\tilde{u}'(\xi) + \tilde{u}'''(\xi) = 0. \quad (2.5)$$

To continue, we assume $\tilde{u}(\xi), \tilde{u}'(\xi), \tilde{u}''(\xi) \rightarrow 0$ as $|\xi| \rightarrow \infty$ and define

$$\mathcal{D} = \{\xi \in \mathbb{R} : 0 < \tilde{u}(\xi) < c\}$$

Integrating (2.5) gives

$$-c\tilde{u}(\xi) + \frac{1}{2}(\tilde{u}(\xi))^2 + \tilde{u}''(\xi) = 0 \quad (2.6)$$

and multiplying the resulting equation with $\tilde{u}'(\xi)$ yields

$$-c\tilde{u}(\xi)\tilde{u}'(\xi) + \frac{1}{2}(\tilde{u}(\xi))^2\tilde{u}'(\xi) + \tilde{u}''(\xi)\tilde{u}'(\xi) = 0,$$

which can be integrated again to obtain

$$-\frac{c}{2}(\tilde{u}(\xi))^2 + \frac{1}{6}(\tilde{u}(\xi))^3 + \frac{1}{2}(\tilde{u}'(\xi))^2 = 0.$$

Due to our assumptions, the constants of integration are always equal to 0. If we assume $\xi \in \mathcal{D}$, we can rewrite the last equation as

$$\tilde{u}'(\xi) = \pm \sqrt{c(\tilde{u}(\xi))^2 - \frac{1}{3}(\tilde{u}(\xi))^3} = \pm \tilde{u}(\xi) \sqrt{c - \frac{1}{3}\tilde{u}(\xi)}.$$

For computational convenience, we only consider the above equation with a minus sign. Separation of variables gives

$$\xi = - \int_0^{\tilde{u}(\xi)} \frac{1}{v \sqrt{c - \frac{1}{3}v}} dv + k, \quad (2.7)$$

where $k \in \mathbb{R}$ is some constant assumed to be 0 from here on. If we substitute

$$v = 3c \operatorname{sech}^2(\theta),$$

we obtain

$$\xi = - \int_0^{\tilde{u}(\xi)} \frac{1}{v \sqrt{c - \frac{1}{3}v}} dv = - \int_0^\theta \frac{1}{3c \operatorname{sech}^2(z) \sqrt{(c - \frac{1}{3}) 3c \operatorname{sech}^2(z)}} \left(\frac{d}{dz} 3c \operatorname{sech}^2(z) \right) dz,$$

where θ is implicitly given by

$$\tilde{u}(\xi) = 3c \operatorname{sech}^2(\theta).$$

With the help of Mathematica [37], one can compute

$$\frac{d}{dz} 3c \operatorname{sech}^2(z) = -6c \operatorname{sech}^2(z) \tanh(z).$$

This, together with

$$\sqrt{\left(c - \frac{1}{3}\right) 3c \operatorname{sech}^2(z)} = \sqrt{c(1 - \operatorname{sech}^2(z))} = \sqrt{c} \tanh(z),$$

implies that (2.7) simplifies to

$$\xi = \frac{2}{\sqrt{c}} \theta.$$

Using the last equations, we obtain

$$\tilde{u}(\xi) = 3c \operatorname{sech}^2(\theta) = 3c \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}\xi\right).$$

Note that

$$0 < \operatorname{sech}(z) \leq 1$$

for all $z \in \mathbb{R}$, and $\operatorname{sech}(z) = 1$ if and only if $z = 0$. This implies that $\mathcal{D} = \mathbb{R} \setminus \{0\}$. However, since

$$\operatorname{sech}(z) = \frac{1}{e^z + e^{-z}},$$

$\tilde{u}$ is defined on $\mathbb{R}$ and smooth. With Mathematica [37], the derivatives of u can be computed:

$$\tilde{u}'(\xi) = -3c^{\frac{3}{2}} \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}\xi\right) \tanh\left(\frac{\sqrt{c}}{2}\xi\right),$$

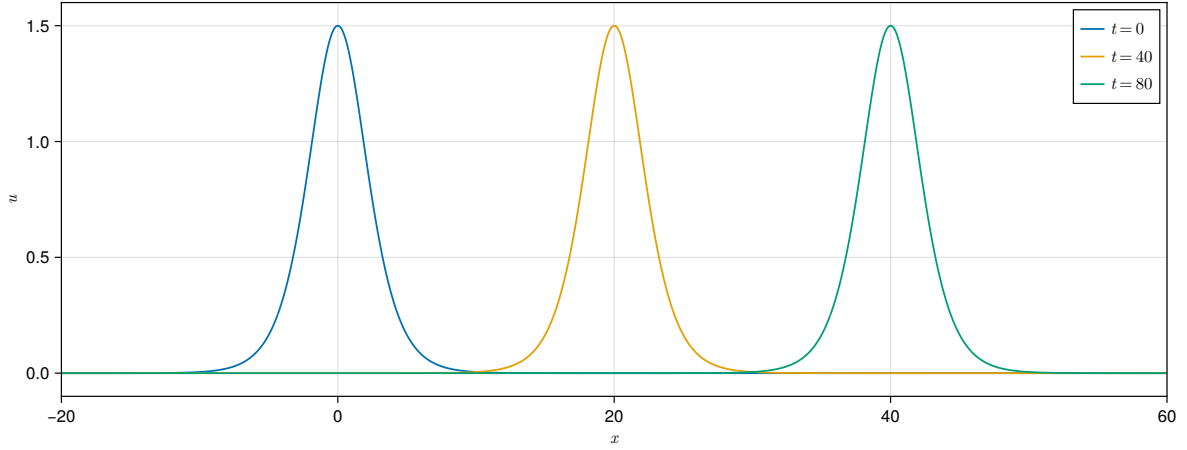


Figure 1: The soliton solution of the KdV equation at different times t .

$$\begin{aligned}\tilde{u}''(\xi) &= -\frac{3}{2} \operatorname{sech}^4\left(\frac{\sqrt{c}}{2}\xi\right) + 3c^2 \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}\xi\right) \tanh^2\left(\frac{\sqrt{c}}{2}\xi\right), \\ \tilde{u}'''(\xi) &= 6c^{\frac{5}{2}} \operatorname{sech}^4\left(\frac{\sqrt{c}}{2}\xi\right) \tanh\left(\frac{\sqrt{c}}{2}\xi\right) - 3c^{\frac{5}{2}} \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}\xi\right) \tanh^3\left(\frac{\sqrt{c}}{2}\xi\right).\end{aligned}$$

Inserting these terms into (2.5) yields

$$\begin{aligned}& 3c^{\frac{5}{2}} \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}\xi\right) \tanh\left(\frac{\sqrt{c}}{2}\xi\right) - 9c^{\frac{5}{2}} \operatorname{sech}^4\left(\frac{\sqrt{c}}{2}\xi\right) \tanh\left(\frac{\sqrt{c}}{2}\xi\right) \\ & \quad + 6c^{\frac{5}{2}} \operatorname{sech}^4\left(\frac{\sqrt{c}}{2}\xi\right) \tanh\left(\frac{\sqrt{c}}{2}\xi\right) - 3c^{\frac{5}{2}} \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}\xi\right) \tanh^3\left(\frac{\sqrt{c}}{2}\xi\right) \\ &= -3c^{\frac{5}{2}} \left(\operatorname{sech}^4\left(\frac{\sqrt{c}}{2}\xi\right) \tanh\left(\frac{\sqrt{c}}{2}\xi\right) - \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}\xi\right) \tanh\left(\frac{\sqrt{c}}{2}\xi\right) \left(1 - \tanh^2\left(\frac{\sqrt{c}}{2}\xi\right)\right) \right)\end{aligned}$$

and since

$$1 - \tanh^2\left(\frac{\sqrt{c}}{2}\xi\right) = \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}\xi\right),$$

it follows that $\tilde{u}$ satisfies (2.5) for all $\xi \in \mathbb{R}$. Consequently, we have proven the following theorem.

Theorem 2.8. *The function $u : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$ defined by*

$$u(t, x) = \tilde{u}(x - ct) = 3c \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}(x - ct)\right)$$

with $c > 0$ satisfies the KdV equation for all $(t, x) \in \mathbb{R}^+ \times \mathbb{R}$.

Remark 2.9. The parameter c in the function controls the width, amplitude, and speed of the wave.

Such a travelling wave solution is also called a soliton solution or soliton of the KdV equation. In Figure 1, one can see the function u with $c = \frac{1}{2}$ at different times t . Notice that the function travels to the right without changing its shape.

In addition, there exist travelling wave solutions that consist of multiple smooth spikes of different amplitude. In this case we refer to each of these spikes as a soliton. The solitons

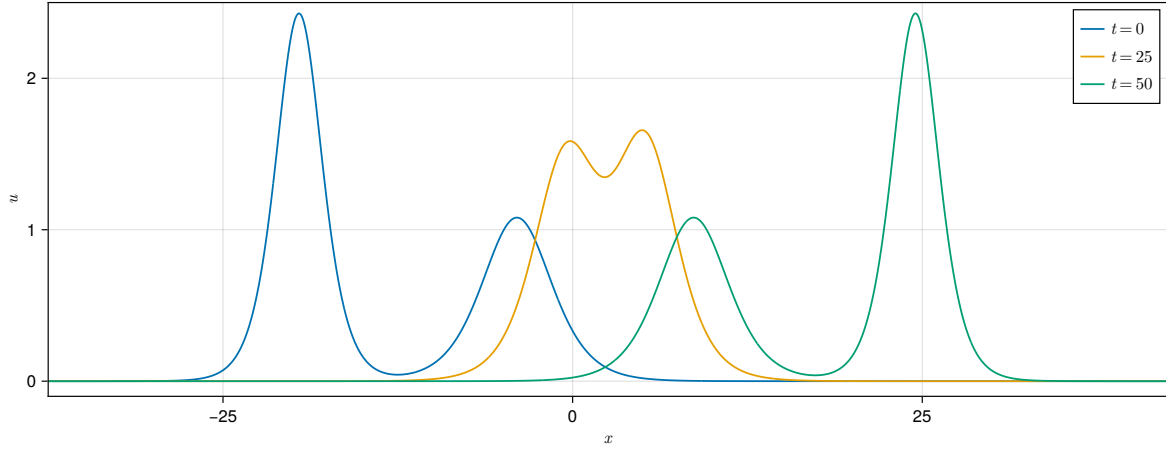


Figure 2: A solution of the KdV equation consisting of two solitons at different times t .

possibly have different amplitude and therefore also different speed. Hence, they can interact with each other since the faster one can pass the slower one. In [41], a solution to the KdV equation consisting of two different solitons is given.

Theorem 2.10. *The function $q : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$,*

$$u(t, x) = 12(\delta_2^2 - \delta_1^2) \frac{\delta_2^2 \operatorname{cosech}^2(\gamma_2(t, x)) + \delta_1^2 \operatorname{sech}^2(\gamma_1(t, x))}{(\delta_2 \coth(\gamma_2(t, x)) + \delta_1 \tanh(\gamma_1(t, x)))^2},$$

where $\gamma_i : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$,

$$\gamma_i(t, x) = \delta_i x - 4\delta_i^3 t$$

for $i = 1, 2$ and $\delta_1, \delta_2 \in \mathbb{R}$ with $\delta_1 < \delta_2$, satisfies the KdV equation on $\mathbb{R}^+ \times \mathbb{R}$.

Proof. See [41]. □

In Figure 2, one can see the function u from the last theorem with parameters $\delta_1 = 0.3$ and $\delta_2 = 0.45$ at different times t . At $t = 0$ the faster, larger soliton is at the left of the slower, smaller soliton. After some time, the faster one overtakes the slower one and they interact with each other ($t = 25$). After the interaction, at $t = 50$, their shape is the same as before the interaction.

2.2.3 Travelling wave solutions of the HKdV equations

In [36] and [39], the authors studied the existence of the travelling wave solutions of the HKdV equations. They found that the HKdV equations also admit travelling wave solutions. To follow their procedure to deduce the existence of travelling wave solution, some existence results for ordinary differential equations are needed, which can be found, for example, in [42]. In this text, travelling wave solution that are related to the soliton solution of the KdV equation are of special interest. The derivation of those travelling wave solution of the HKdV equations is similar to the one of the KdV equation conducted before.

We start by assuming that $u, \rho, \psi : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$ are continuously differentiable functions that can be written as

$$\begin{aligned} u(x, t) &= \tilde{u}(x - ct), \\ \rho(x, t) &= \tilde{\rho}(x - ct), \end{aligned}$$

$$\psi(x, t) = \tilde{\psi}(x - ct),$$

where $c > 0$ and $\tilde{q}, \tilde{\rho}, \tilde{\psi} : \mathbb{R} \rightarrow \mathbb{R}$ are, again, assumed to be continuously differentiable. Moreover, we assume $\tilde{u}(\xi), \tilde{\rho}(\xi), \tilde{\psi}(\xi) \rightarrow 0$ as $|\xi| \rightarrow \infty$. If u, ρ and ψ satisfy the HKdV equations, the functions $\tilde{u}, \tilde{\rho}$ and $\tilde{\psi}$ satisfy

$$\begin{aligned} -c\tilde{u}'(\xi) + \tilde{u}(\xi)\tilde{u}'(\xi) + \tilde{\psi}'(\xi) &= 0, \\ -c\tilde{\rho}'(\xi) - \frac{1}{\varepsilon^2}\tilde{\rho}'(\xi) &= -\frac{1}{\varepsilon^2}\tilde{\psi}(\xi), \\ -c\tilde{\psi}'(\xi) + \frac{1}{\varepsilon^2}\tilde{u}'(\xi) &= \frac{1}{\varepsilon^2}\tilde{\rho}(\xi) \end{aligned} \tag{2.8}$$

for all $\xi \in \mathbb{R}$. If we can show that there exist functions that satisfy (2.8), the existence of travelling wave solutions to the HKdV equations immediately follows. We begin by reducing (2.8) to a system of two equations. Rewriting the first equation as

$$\tilde{\psi}'(\xi) = c\tilde{u}'(\xi) - \tilde{u}(\xi)\tilde{u}'(\xi)$$

and inserting it into the third equation yields

$$(1 + \varepsilon^2 c (\tilde{u}(\xi) - c)) \tilde{u}'(\xi) = \tilde{\rho}(\xi).$$

To continue, we need to divide the equation by

$$1 + \varepsilon^2 c (\tilde{u}(\xi) - c)$$

and therefore, assume that this term is unequal to zero for all $\xi \in \mathbb{R}$. With this assumption, we obtain

$$\tilde{u}'(\xi) = \frac{1}{1 + \varepsilon^2 c (\tilde{u}(\xi) - c)} \tilde{\rho}(\xi).$$

Integrating the first equation gives

$$-c\tilde{u}(\xi) + \frac{1}{2}(\tilde{u}(\xi))^2 + \tilde{\psi}(\xi) = 0.$$

Due to the assumption, the constant of integration is zero. Rewriting the equation as

$$\tilde{\psi}(\xi) = \left(c - \frac{1}{2}\tilde{u}(\xi) \right) \tilde{u}(\xi),$$

and inserting it into the second equation, yields

$$\tilde{\rho}'(\xi) = \frac{1}{(1 + \varepsilon^2 c)} \left(c - \frac{1}{2}\tilde{u}(\xi) \right) \tilde{u}(\xi).$$

Now we are left with a system of two ordinary differential equations,

$$\begin{aligned} \tilde{u}'(\xi) &= \frac{1}{1 + \varepsilon^2 c (\tilde{u}(\xi) - c)} \tilde{\rho}(\xi), \\ \tilde{\rho}'(\xi) &= \frac{1}{(1 + \varepsilon^2 c)} \left(c - \frac{1}{2}\tilde{u}(\xi) \right) \tilde{u}(\xi). \end{aligned} \tag{2.9}$$

We call a point $\bar{y} = (\bar{u}, \bar{\rho})^T \in \mathbb{R}^2$ an equilibrium point of the system (2.9), if $\tilde{u}(\xi) = \bar{u}$ and $\tilde{\rho}(\xi) = \bar{\rho}$ are constant solutions of (2.9). The equilibrium point of system (2.9) are $(0, 0)^T$ and $(2c, 0)^T$.

Lemma 2.11. *The function $H : \mathbb{R}^2 \rightarrow \mathbb{R}$,*

$$H(v_1, v_2) = \frac{1}{2}(v_2)^2 - \frac{1}{1 + \varepsilon^2 c}(v_1)^2 \left(-\frac{\varepsilon^2 c}{8}(v_1)^2 + \frac{3\varepsilon^2 c^2 - 1}{6}v_1 + \frac{c(1 - \varepsilon^2 c^2)}{2} \right).$$

is a first integral of the system of ordinary differential equations (2.9), i.e.

$$\frac{d}{d\xi} H(\tilde{u}(\xi), \tilde{\rho}(\xi)) = 0$$

for all solutions $\tilde{u}$ and $\tilde{\rho}$ of the system (2.9).

Proof. First, we can compute the derivatives of H with respect to v_1 and v_2 :

$$\begin{aligned} \frac{\partial}{\partial v_1} H(v_1, v_2) &= \frac{1}{1 + \varepsilon^2 c} \left(\varepsilon^2 c \frac{(v_1)^3}{2} - (3\varepsilon^2 c^2 - 1) \frac{(v_1)^2}{2} - c(1 - \varepsilon^2 c^2) v_1 \right) \\ &= \frac{1}{1 + \varepsilon^2 c} \left(\varepsilon^2 c \frac{(v_1)^3}{2} - 3\varepsilon^2 c^2 \frac{(v_1)^2}{2} + \frac{(v_1)^2}{2} - cv_1 + \varepsilon^2 c^3 v_1 \right) \\ &= -\frac{1}{1 + \varepsilon^2 c} \left(\left(c - \frac{v_1}{2} \right) v_1 + \varepsilon^2 c v_1 \left(c - \frac{v_1}{2} \right) v_1 - \varepsilon^2 c^2 \left(c - \frac{v_1}{2} \right) v_1 \right) \\ &= -\frac{1}{1 + \varepsilon^2 c} \left(c - \frac{v_1}{2} \right) v_1 (1 + \varepsilon^2 c(v_1 - c)) \end{aligned}$$

and

$$\frac{\partial}{\partial v_2} H(v_1, v_2) = v_2.$$

Assuming $\tilde{u}$ and $\tilde{\rho}$ are solutions to the system (2.9), we have

$$\begin{aligned} \frac{d}{d\xi} H(\tilde{u}(\xi), \tilde{\rho}(\xi)) &= \tilde{u}'(\xi) \frac{\partial}{\partial v_1} H(\tilde{u}(\xi), \tilde{\rho}(\xi)) + \tilde{\rho}'(\xi) \frac{\partial}{\partial v_2} H(\tilde{u}(\xi), \tilde{\rho}(\xi)) \\ &= \frac{1}{(1 + \varepsilon^2 c(\tilde{u}(\xi) - c))} \tilde{\rho}(\xi) \frac{\partial}{\partial v_1} H(\tilde{u}(\xi), \tilde{\rho}(\xi)) \\ &\quad + \frac{1}{(1 + \varepsilon^2 c)} \left(c - \frac{1}{2} \tilde{u}(\xi) \right) \tilde{u}(\xi) \frac{\partial}{\partial v_2} H(\tilde{u}(\xi), \tilde{\rho}(\xi)). \\ &= -\frac{1}{1 + \varepsilon^2 c} \left(c - \frac{1}{2} \tilde{u}(\xi) \right) \tilde{u}(\xi) \tilde{\rho}(\xi) + \frac{1}{1 + \varepsilon^2 c} \left(c - \frac{1}{2} \tilde{u}(\xi) \right) \tilde{u}(\xi) \tilde{\rho}(\xi) \\ &= 0. \end{aligned}$$

Thus, H is a first integral of the system of ordinary differential equations. $\square$

Using the first integral of the system allows us to draw the phase portrait of the system of ordinary differential equations, i.e. the level curves of H . The phase portrait for $c = \frac{1}{2}$ and $\varepsilon = 0.8$ is depicted in Figure 3, where the blue line represents a curve starting and ending at the equilibrium point $(0, 0)^T$.

Theorem 2.12. *If we set $c = \frac{1}{2}$ and $\varepsilon = 0.8$, then there exists continuously differentiable functions $\tilde{u}, \tilde{\rho} : \mathbb{R} \rightarrow \mathbb{R}^2$ satisfying (2.9), $\tilde{u}(\xi), \tilde{\rho}(\xi) \rightarrow 0$ as $|\xi| \rightarrow \infty$, and*

$$(1 + \varepsilon^2 c(\tilde{u}(\xi) - c)) \neq 0$$

for all $\xi \in \mathbb{R}$.

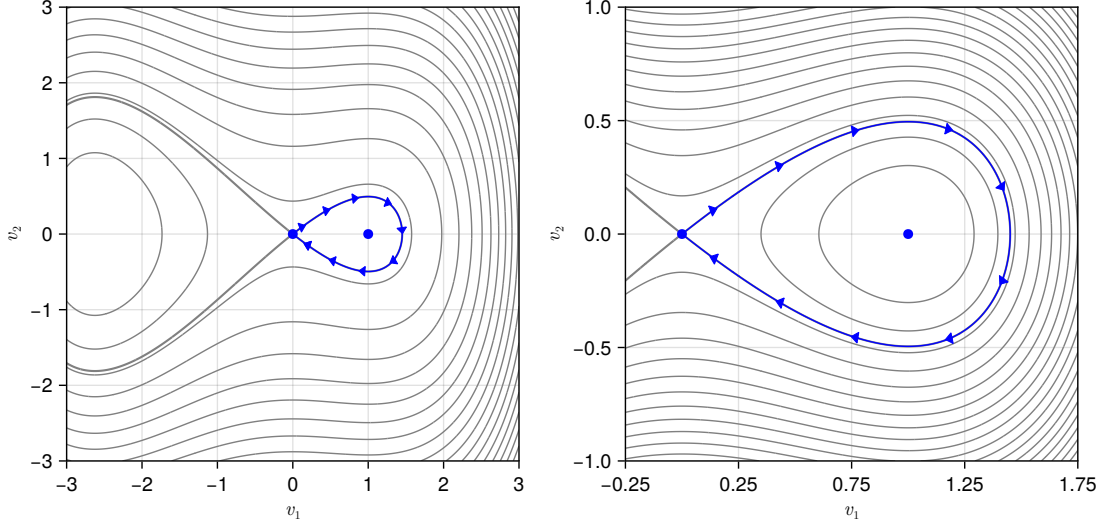


Figure 3: The phase portrait of the system of ordinary differential equations (2.9). The relaxation parameter is set to $\varepsilon = 0.8$ and $c = 1/2$.

Proof. Consider the point furthest right on the blue curve in Figure 3, which we denote by $(\bar{u}, 0)^T$ with $\bar{u} > 0$. We rewrite (2.9) as

$$\frac{d}{d\xi} y(\xi) = f(y(\xi)),$$

where

$$y(\xi) = \begin{pmatrix} \tilde{u}(\xi) \\ \tilde{\rho}(\xi) \end{pmatrix}$$

and $f : U \rightarrow \mathbb{R}^2$,

$$f(y) = \begin{pmatrix} \frac{1}{1+\varepsilon^2 c(\tilde{u}-c)} \tilde{\rho} \\ \frac{1}{(1+\varepsilon^2 c)} (c - \frac{1}{2}\tilde{u}) \tilde{u} \end{pmatrix}.$$

with $U = (-1, 2) \times (-1, 2)$. Since $\varepsilon^2 c = 0.32$ and $|\tilde{u} - c| \leq 1.5$ for $(\tilde{u}, \tilde{\rho})^T \in U$, it follows

$$\varepsilon^2 c |\tilde{u} - c| \leq 0.48$$

and

$$1 + \varepsilon^2 c (\tilde{u} - c) \neq 0$$

for all $(\tilde{u}, \tilde{\rho})^T \in U$. This implies that f is defined on U and continuously differentiable. Thus, there exists a $a > 0$ and a function $y : (-a, a) \rightarrow \mathbb{R}^2$ satisfying (2.9) and $y(0) = (\bar{u}, 0)^T$ (see [42], Chapter 17.1). Due to the previous lemma, H is constant along the solution curve $\xi \mapsto y(\xi)$ and, hence, the solution curve lies on the blue curve in Figure 3. Therefore, we are able to define a compact set $K \subseteq U$ such that for any solution $y : (-a, a) \rightarrow \mathbb{R}^2$, it follows that $y(\xi) \in K$ for all $\xi \in (-a, a)$. This implies that there exists a function $y : \mathbb{R} \rightarrow U$ that satisfies $y(0) = (\bar{u}, 0)^T$ and (2.9) (see [42], Chapter 17.4). Since y matches the blue curve, $y(\xi) \rightarrow 0$ as $|\xi| \rightarrow \infty$. $\square$

Given functions $\tilde{u}$ and $\tilde{\rho}$ satisfying (2.9), we can obtain a continuously differentiable function ψ with $\psi(\xi) \rightarrow 0$ for $|\xi| \rightarrow \infty$. The functions $\tilde{u}$, $\tilde{\rho}$, and ψ satisfy (2.8) and, as an immediate consequence,

$$u(x, t) = \tilde{u}(x - ct),$$

$$\begin{aligned}\rho(x, t) &= \tilde{\rho}(x - ct), \\ \psi(x, t) &= \tilde{\psi}(x - ct),\end{aligned}$$

are a continuously differentiable travelling wave solution of the HKdV equations.

Remark 2.13. We have only shown the existence of a travelling wave solution for a special choice of c and ε . However, in the same way, the existence of travelling wave solutions of the HKdV equations for different values of ε and c can be proven, provided a similar phase portrait exists.

2.2.4 Numerical computation of the HKdV travelling wave solutions

In this section, (2.9) is solved numerically to compute numerical approximations to the travelling wave solutions of the HKdV equation based on [39]. We start by rewriting the first equation of (2.9) as

$$\tilde{\rho}(\xi) = \tilde{u}'(\xi) + \varepsilon^2 c \tilde{u}'(\xi) \tilde{u}(\xi) - \varepsilon^2 c^2 \tilde{u}'(\xi).$$

Assuming that $\tilde{u}$ and $\tilde{\rho}$ are sufficiently smooth and taking the derivatives in the last equations yields

$$\tilde{\rho}'(\xi) = \tilde{u}''(\xi) + \varepsilon^2 c \frac{d}{d\xi}(\tilde{u}'(\xi) \tilde{u}(\xi)) - \varepsilon^2 c^2 \tilde{u}''(\xi).$$

We can insert this equation into the second equation of system (2.9) to obtain the scalar ordinary differential equation

$$\tilde{u}''(\xi) + \varepsilon^2 c \frac{d}{d\xi}(\tilde{u}'(\xi) \tilde{u}(\xi)) - \varepsilon^2 c^2 \tilde{u}''(\xi) = \frac{1}{(1 + \varepsilon^2 c)} \left(c - \frac{1}{2} \tilde{u}(\xi) \right) \tilde{u}(\xi)$$

or, equivalently,

$$(1 - \varepsilon^2 c^2) \tilde{u}''(\xi) - \frac{c}{1 + \varepsilon^2 c} \tilde{u}(\xi) = -\frac{1}{2(1 + \varepsilon^2 c)} (\tilde{u}(\xi))^2 - \varepsilon^2 c \frac{d}{d\xi}(\tilde{u}'(\xi) \tilde{u}(\xi)). \quad (2.10)$$

If $c \in \mathbb{R}$ fixed and $\varepsilon \rightarrow 0$, this equation becomes

$$-c \tilde{u}(\xi) + \frac{1}{2} (\tilde{u}(\xi))^2 + \tilde{u}''(\xi) = 0,$$

which is (2.6). Since this equation appeared in the derivation of the KdV soliton solution, this indicates that the HKdV travelling wave solution tends to soliton solution of the KdV equations as $\varepsilon \rightarrow 0$.

Remark 2.14. The travelling wave solutions of the HKdV equations are not only computed for demonstrative purposes, but they are also used as initial data and reference solution in the numerical experiments later in this thesis. Thus, it is of importance that the obtained numerical approximations of (2.10) are sufficiently accurate.

In theory, the solution to (2.10) exists on $\mathbb{R}$, however numerically we can only compute the solution on a closed and bounded interval $[x_{\min}, x_{\max}]$ and, thus, need additional boundary conditions. We choose

$$[x_{\min}, x_{\max}] = [-47.0, 47.0]$$

discretised with $n = 42000$ grid points and impose periodic boundary conditions, i.e. the numerical solution at the grid point $n + 1$ is equal to the numerical solution at the first grid point. Define the vector $U \in \mathbb{R}^n$ as

$$U = (u_1, \dots, u_n)^T,$$

where $u_1, \dots, u_n$ represent the approximation to the solution of (2.10) at the grid points $\xi_1, \dots, \xi_n$. To approximate the derivatives appearing in the ordinary differential equation, we use an eighth order central finite difference formula [43]. The matrices $D_1 \in \mathbb{R}^{n \times n}$ and $D_2 \in \mathbb{R}^{n \times n}$ represents the finite difference operators that, applied to U , yield approximations to the first and second derivative.

Remark 2.15. The authors in [39] used Fourier spectral differentiation. This could perhaps increase the accuracy and efficiency of this algorithm. However, using finite difference approximations and a sufficiently fine grid is sufficient for the experiments in this thesis.

Next, we define

$$L = -D_2 + \frac{c}{(1 + \varepsilon^2 c)(1 - \varepsilon^2 c^2)} \mathbb{I}_{n \times n},$$

where $\mathbb{I}_{n \times n}$ is the identity matrix and $N : \mathbb{R}^n \rightarrow \mathbb{R}^n$,

$$N(U) = \frac{1}{2(1 + \varepsilon^2 c)(1 - \varepsilon^2 c^2)} U^2 + \frac{\varepsilon^2 c}{1 - \varepsilon^2 c^2} D_1(U \odot D_1 U).$$

The multiplication $U \odot D_1 U$ denotes the point wise multiplication of the two vectors, i.e.

$$U \odot D_1 U = \begin{pmatrix} u_1(D_1 U)_1 \\ \vdots \\ u_n(D_1 U)_n \end{pmatrix}.$$

The resulting nonlinear system of equations,

$$LU = N(U), \tag{2.11}$$

is a discretisation of (2.10). To solve (2.11), a Petviashvili-type algorithm [44] is used, which is a type of fix point algorithm. As an initial guess, we employ the soliton solution of the KdV equation,

$$u_i^0 = 3c \operatorname{sech}^2 \left(\frac{\sqrt{c}}{2} \xi_i \right)$$

for all $i = 1, \dots, n$ and set

$$U^0 = (u_1^0, \dots, u_n^0)^T.$$

Starting with $i = 0$, the linear system of equations

$$LU^{i+1} = m(U^i)^2 N(U^i)$$

with stabilising factor

$$m(U^i) = \frac{LU^i \cdot U^i}{N(U^i) \cdot U^i}$$

is solved exactly until the l_∞ -error

$$\|LU^i - N(U^i)\|_\infty$$

is sufficiently small or does not decrease anymore.

Remark 2.16. The l_∞ -error of the numerical solutions of (2.11) presented in this text are smaller than $5 \cdot 10^{-10}$.

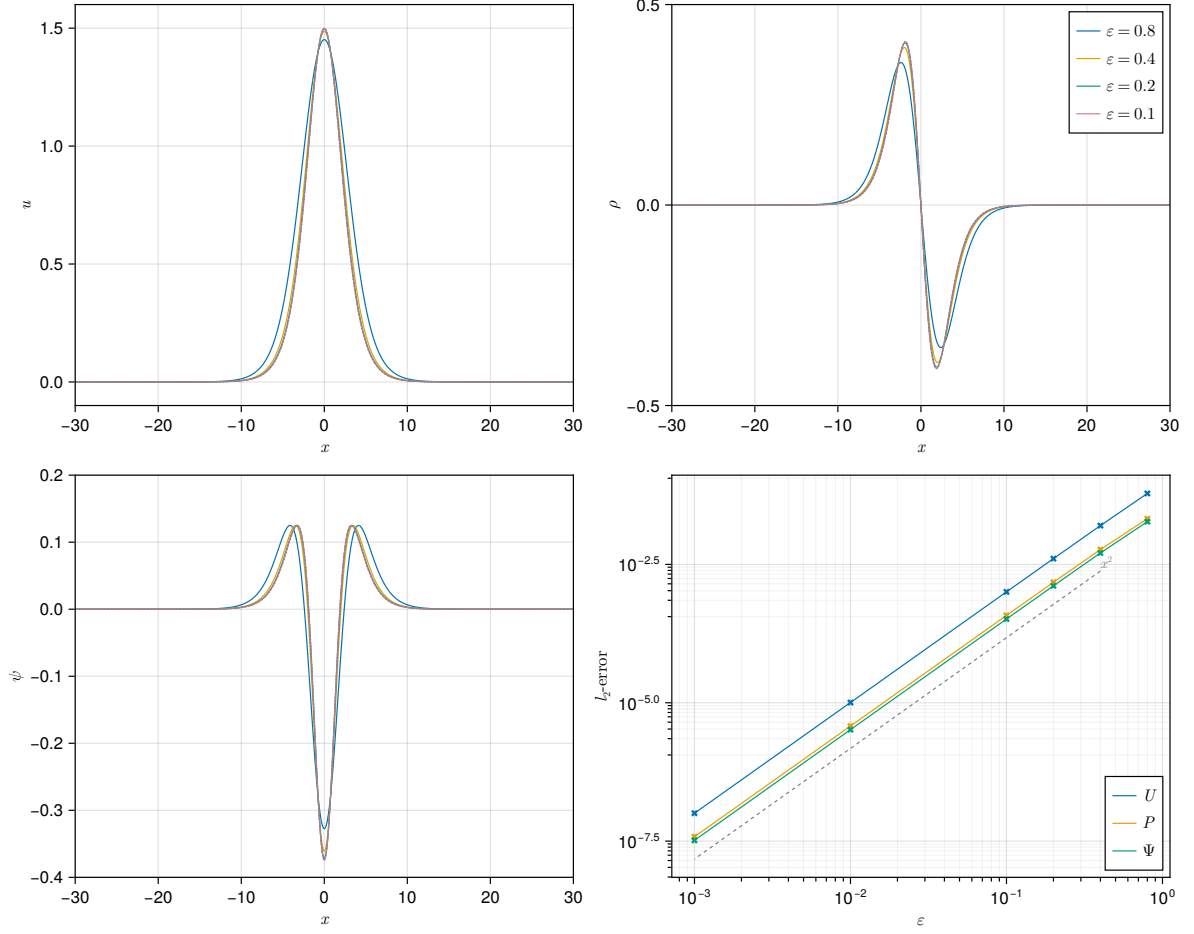


Figure 4: From top left to bottom right: $u_{\text{num}}(0, x)$, $\rho_{\text{num}}(0, x)$ and $\psi_{\text{num}}(0, x)$ in comparison the corresponding KdV soliton and its derivatives for different values of ε and the l_2 -error of U , P and Ψ with respect to ε .

After we obtained the approximations U with a sufficiently small error, approximation P and Ψ to ρ and ψ at the grid point can be computed:

$$P = (\mathbb{I}_n + c\varepsilon^2 (U - c\mathbb{I}_n)) \odot D_1 U$$

and

$$\Psi = D_1 P + c\varepsilon^2 D_1 P.$$

To obtain a function, which can be evaluated at every given point in space, we interpolate the solution at the grid points with cubic splines. The resulting functions are denoted by $\tilde{u}_{\text{num}}, \tilde{\rho}_{\text{num}}, \tilde{\psi}_{\text{num}}$. We can obtain a numerical approximation to the travelling wave solution of the HKdV equations by shifting these functions to the right with respect to time, i.e.

$$\begin{aligned} u_{\text{num}}(t, x) &= \tilde{u}_{\text{num}}(x - ct), \\ \rho_{\text{num}}(t, x) &= \tilde{\rho}_{\text{num}}(x - ct), \\ \psi_{\text{num}}(t, x) &= \tilde{\psi}_{\text{num}}(x - ct). \end{aligned}$$

In Figure 4, the functions $u_{\text{num}}, \rho_{\text{num}}$ and ψ_{num} with $c = \frac{1}{2}$ at $t = 0$ are depicted in comparison to the corresponding KdV soliton solution and its derivatives for different values of ε together with the l_2 -error of U , P , and Ψ with respect to ε . The results indicate that the numerical

HKdV travelling wave solution converges to the KdV soliton solution as $\varepsilon \rightarrow 0$. The achieved accuracy with respect to ε is approximately two.

2.2.5 Conserved quantities of the KdV equation and HKdV equations

The KdV equation as well as the HKdV equations admit conserved quantities. In fact, according to [39], the KdV equation has an infinite number of conserved quantities. Assume the solution $u : \mathbb{R}^+ \times \mathbb{R} \rightarrow \mathbb{R}$ of the KdV equation is sufficiently smooth. We denote the mass by

$$M(t) = \int_{\mathbb{R}} u(t, x) dx$$

and have

$$\begin{aligned} \frac{d}{dt} M(t) &= \frac{d}{dt} \int_{\mathbb{R}} u(t, x) dx \\ &= \int_{\mathbb{R}} \frac{\partial}{\partial t} u(t, x) dx \\ &= \int_{\mathbb{R}} -\frac{\partial}{\partial x} \left(\frac{1}{2} (u(t, x))^2 + \frac{\partial^2}{\partial x^2} u(t, x) \right) dx. \end{aligned}$$

If we also assume that $u(t, x)$ as well as its derivatives go to 0 for $|x| \rightarrow \infty$ and a fixed $t \in \mathbb{R}^+$, as it is the case for the soliton solution, the last integral is equal to zero and the mass is conserved. The energy is defined as

$$E(t) = \int_{\mathbb{R}} \frac{1}{2} (u(t, x))^2 dx$$

and we also have

$$\begin{aligned} \frac{d}{dt} E(t) &= \frac{d}{dt} \int_{\mathbb{R}} \frac{1}{2} (u(t, x))^2 dx \\ &= \int_{\mathbb{R}} \frac{1}{2} \frac{\partial}{\partial t} (u(t, x))^2 dx \\ &= \int_{\mathbb{R}} u(t, x) \frac{\partial}{\partial t} u(t, x) dx \\ &= \int_{\mathbb{R}} u(t, x) \left(-u(t, x) \frac{\partial}{\partial x} (u(t, x))^2 - \frac{\partial^3}{\partial x^3} u(t, x) \right) dx \\ &= \int_{\mathbb{R}} -\frac{\partial}{\partial x} \left(\frac{1}{3} (u(t, x))^3 + u(t, x) \frac{\partial^2}{\partial x^2} u(t, x) - \frac{1}{2} \left(\frac{\partial}{\partial x} u(t, x) \right)^2 \right) dx, \\ &= 0, \end{aligned}$$

where we used the same assumptions on u as before. Concerning the HKdV equations, one can show in a similar way that the mass M is a conserved quantity of the system. According to [39], the energy E is not exactly conserved for the HKdV equations. However, the authors define a modified energy:

$$E_H(t) = \int_{\mathbb{R}} \frac{1}{2} ((u(t, x))^2 + \varepsilon^2 (\rho(t, x))^2 + \varepsilon^2 (\psi(t, x))^2) dx.$$

This modified energy is conserved by the HKdV equations since

$$\frac{d}{dt} E_H(t) = \int_{\mathbb{R}} \frac{1}{2} \frac{\partial}{\partial t} ((u(t, x))^2 + \varepsilon^2 (\rho(t, x))^2 + \varepsilon^2 (\psi(t, x))^2) dx$$

$$\begin{aligned}
 &= \int_{\mathbb{R}} u(t, x) \frac{\partial}{\partial t} u(t, x) + \varepsilon^2 \rho(t, x) \frac{\partial}{\partial t} \rho(t, x) + \varepsilon^2 \psi(t, x) \frac{\partial}{\partial t} \psi(t, x) dx \\
 &= \int_{\mathbb{R}} u(t, x) \left(-u(t, x) \frac{\partial}{\partial x} u(t, x) - \frac{\partial}{\partial x} \psi(t, x) \right) \\
 &\quad + \rho(t, x) \left(\frac{\partial}{\partial x} \rho(t, x) - \psi(t, x) \right) \\
 &\quad + \psi(t, x) \left(-\frac{\partial}{\partial x} u(t, x) + \rho(t, x) \right) dx \\
 &= \int_{\mathbb{R}} -\frac{\partial}{\partial x} \left(\frac{1}{3} (u(t, x))^3 + u(t, x) \psi(t, x) + \frac{1}{2} (\rho(t, x))^2 \right) \\
 &= 0
 \end{aligned}$$

under the assumption that u, ρ and ψ are continuously differentiable and go to zero for a fixed t as $|x| \rightarrow \infty$.

2.3 One-dimensional semi-discrete Active Flux scheme

In this section, the third-order semi-discrete Active Flux scheme with a point value update equation based on Jacobian splitting to solve initial values problems of the form

$$\begin{aligned}
 \frac{\partial}{\partial t} q(t, x) + \frac{\partial}{\partial x} f(q(t, x)) &= 0, \quad t \in \mathbb{R}^+, x \in \mathbb{R}, \\
 q(0, x) &= q_0(x), \quad x \in \mathbb{R}
 \end{aligned}$$

is presented based on [17] and [24]. To derive the semi-discrete Active Flux scheme, we start by dividing the real line into cells

$$\mathcal{C}_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}],$$

where $x_{i\pm\frac{1}{2}} \in \mathbb{R}$ denotes the boundaries of the cell and x_i the centre of the cell $\mathcal{C}_i$ with $i \in \mathbb{Z}$. We assume all cells have the same size $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} > 0$. The degrees of freedom of the third-order semi-discrete Active Flux scheme are cell averages

$$Q_i(t) \approx \frac{1}{\Delta x} \int_{\mathcal{C}_i} q(t, x) dx$$

and point values

$$q_{i\pm\frac{1}{2}}(t) \approx q(t, x_{i\pm\frac{1}{2}})$$

located at the interfaces of the cells. The point values are shared between two adjacent cells. This amounts to two degrees of freedom per cell: one point value and one average value. In Figure 5 one can see an illustration of the degrees of freedom.

The next step is to find, locally in each cell $\mathcal{C}_i$, a quadratic polynomial that interpolates the degrees of freedom accessible to this cell. We call this polynomial the (local) reconstruction.

Proposition 2.17. *Suppose for a cell $\mathcal{C}_i$ the point values $q_{i\pm\frac{1}{2}}(t)$ and the cell average $Q_i(t)$ at time $t \geq 0$ are given. The quadratic polynomial*

$$q_{\text{recon},i}(t, \xi) = \frac{6Q_i(t) - q_{i+\frac{1}{2}}(t) - q_{i-\frac{1}{2}}(t)}{4} + \frac{q_{i+\frac{1}{2}}(t) - q_{i-\frac{1}{2}}(t)}{\Delta x} \xi + 3 \frac{q_{i-\frac{1}{2}}(t) + q_{i+\frac{1}{2}}(t) - 2Q_i(t)}{\Delta x^2} \xi^2$$

with $\xi \in [-\frac{\Delta x}{2}, \frac{\Delta x}{2}]$, interpolates the cell averages and the point values, i.e.

$$q_{\text{recon},i} \left(t, \pm \frac{\Delta x}{2} \right) = q_{i\pm\frac{1}{2}}(t), \quad \frac{1}{\Delta x} \int_{\mathcal{C}_i} q_{\text{recon},i}^n(t, \xi) d\xi = Q_i(t).$$

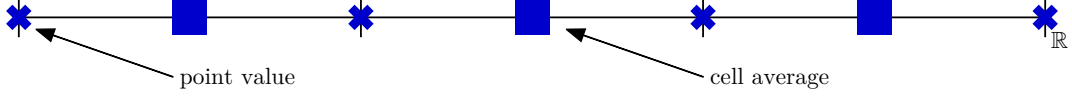


Figure 5: Illustration of the degrees of freedom of the third-order semi-discrete Active Flux scheme.

Proof. To verify that this polynomial interpolates the point values, we compute

$$\begin{aligned} q_{\text{recon},i} \left(t, \frac{\Delta x}{2} \right) &= \frac{1}{4} \left(6Q_i(t) - q_{i+\frac{1}{2}}(t) - q_{i-\frac{1}{2}}(t) \right) + \frac{1}{2} \left(q_{i+\frac{1}{2}}(t) - q_{i-\frac{1}{2}}(t) \right) \\ &\quad + \frac{3}{4} \left(q_{i-\frac{1}{2}}(t) + q_{i+\frac{1}{2}}(t) - 2Q_i(t) \right) \\ &= q_{i+\frac{1}{2}}(t) \end{aligned}$$

and

$$\begin{aligned} q_{\text{recon},i} \left(t, -\frac{\Delta x}{2} \right) &= \frac{1}{4} \left(6Q_i(t) - q_{i+\frac{1}{2}}(t) - q_{i-\frac{1}{2}}(t) \right) - \frac{1}{2} \left(q_{i+\frac{1}{2}}(t) - q_{i-\frac{1}{2}}(t) \right) \\ &\quad + \frac{3}{4} \left(q_{i-\frac{1}{2}}(t) + q_{i+\frac{1}{2}}(t) - 2Q_i(t) \right) \\ &= q_{i-\frac{1}{2}}(t). \end{aligned}$$

In addition, calculating the cell averages of the reconstruction yields

$$\begin{aligned} \frac{1}{\Delta x} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} q_{\text{recon},i}(t, \xi) d\xi &= \left[\frac{1}{4} \left(6Q_i(t) - q_{i+\frac{1}{2}}(t) - q_{i-\frac{1}{2}}(t) \right) \xi + \frac{1}{4\Delta x} \left(q_{i+\frac{1}{2}}(t) - q_{i-\frac{1}{2}}(t) \right) \xi^2 \right. \\ &\quad \left. + \frac{1}{\Delta x^2} \left(q_{i-\frac{1}{2}}(t) + q_{i+\frac{1}{2}}(t) - 2Q_i(t) \right) \xi^3 \right]_{\xi=-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \\ &= Q_i(t). \end{aligned}$$

Thus, the quadratic polynomial defined as above interpolated the point values and cell averages. $\square$

The assumption of the last theorem is not an unnatural assumption since at $t = 0$ the cell averages and point value are defined by the initial condition. Given the quadratic polynomial in each cell, we can define a function $q_{\text{recon}} : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$ by

$$q_{\text{recon}}(t, x) = q_{\text{recon},i}(t, x - x_i), \quad \text{if } x \in \mathcal{C}_i.$$

We call this function the (global) reconstruction. Since the point values at the interfaces of the cells are shared between two adjacent cells, the reconstruction is continuous, but it does not have to be differentiable across the cell interfaces. An illustration of the reconstruction is shown in Figure 6.

Remark 2.18. The reconstruction can be considered the numerical solution, i.e. the approximation to the solution of the initial value problem.

Next, we want to find update equations, i.e. semi-discretisation of the conservation law, for the cell averages and the point values. Integrating the conservation law over the cell $\mathcal{C}_i$ yields

$$\frac{d}{dt} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(t, x) dx + f(q(t, x_{i+\frac{1}{2}})) - f(q(t, x_{i-\frac{1}{2}})) = 0.$$

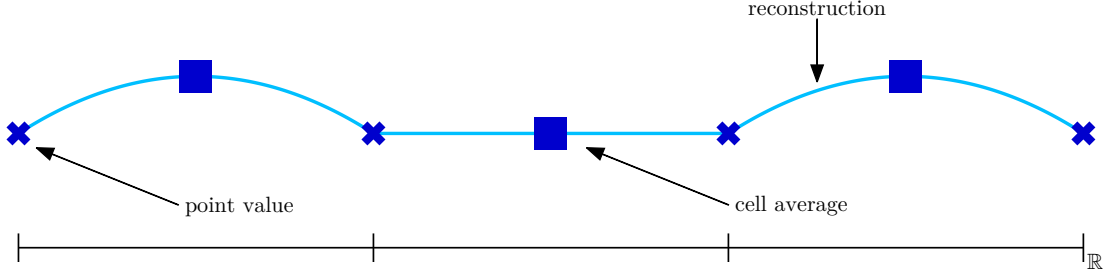


Figure 6: Illustration of the reconstruction of the third-order semi-discrete Active Flux scheme.

We can rewrite this to obtain

$$\frac{d}{dt} \left(\frac{1}{\Delta x} \int_{c_i} q(t, x) dx \right) = - \frac{f(q(t, x_{i+\frac{1}{2}})) - f(q(t, x_{i-\frac{1}{2}}))}{\Delta x}.$$

The left side of the equation is the time derivative of the exact cell average, and on the right side we have values of the solution at the location of the point values. Therefore, we can deduce an update equation for the cell average:

$$\frac{d}{dt} Q_i(t) = - \frac{f(q_{i+\frac{1}{2}}(t)) - f(q_{i-\frac{1}{2}}(t))}{\Delta x}.$$

To find an update equation for the point values, we consider the quasi-linear form of the system of hyperbolic conservation laws,

$$\frac{\partial}{\partial t} q(t, x) + F(q) \frac{\partial}{\partial x} q(t, x) = 0.$$

Our aim is to approximate the spatial derivatives in the conservation law with finite difference formulas at the location of the point values. These formulas can be obtained by differentiating the reconstruction. As stated before, the reconstruction is not differentiable at the cell interfaces. Hence, at the location of a point value, there are two different derivatives: one from the cell on the right and one from the cell on the left. We define

$$D^- q_{i+\frac{1}{2}}(t) = \frac{\partial}{\partial \xi} q_{\text{recon}, i+1} \left(t, -\frac{\Delta x}{2} \right)$$

and

$$D^+ q_{i+\frac{1}{2}}(t) = \frac{\partial}{\partial \xi} q_{\text{recon}, i} \left(t, \frac{\Delta x}{2} \right)$$

to be those two derivatives. Having two different derivatives is no disadvantage since it allows us to include upwinding.

Remark 2.19. According to [45], for hyperbolic conservation laws, we expect information to propagate as waves moving along characteristics. In case of systems, we have multiple waves that travel from different directions and with different speeds. Upwinding means that we want to take this into account in the design of numerical methods.

We diagonalise the Jacobian F of f ,

$$F(q) = R(q) \Lambda(q) R^{-1}(q)$$

with $q \in \mathcal{O} \subseteq \mathbb{R}^m$. The matrix $\Lambda(q)$ is

$$\Lambda(q) = \text{diag}(\lambda_1(q), \lambda_2(q), \lambda_3(q)),$$

where $\lambda_i(q)$ with $i = 1, \dots, m$ are the eigenvalues of $F(q)$. Since we assumed that the conservation law is hyperbolic, this is always possible. Next, we define

$$F^\pm(q) = R\Lambda^\pm(q)R^{-1}$$

with

$$\Lambda^\pm(q) = \text{diag}(\lambda_1^\pm(q), \lambda_2^\pm(q), \lambda_3^\pm(q)),$$

where

$$\lambda_i^+(q) = \max(\lambda_i(q), 0)$$

and

$$\lambda_i^-(q) = \min(\lambda_i(q), 0).$$

Due to their construction,

$$F^+(q) + F^-(q) = F(q)$$

and hence, we split F into a part with positive eigenvalues (right-going propagation) and a part with negative eigenvalues (left-going propagation). Based on this, we choose

$$\frac{d}{dt}q_{i+\frac{1}{2}}(t) = -\left(F^+(q_{i+\frac{1}{2}}(t))D^+q_{i+\frac{1}{2}}(t) + F^-(q_{i+\frac{1}{2}}(t))D^-q_{i+\frac{1}{2}}(t)\right)$$

as the update equation for the point values. The reconstruction does not need to be computed explicitly for a given time to obtain the approximations to the derivatives.

Proposition 2.20. *The derivatives of the reconstruction at the location of the point values are*

$$\begin{aligned} D^+q_{i+\frac{1}{2}}(t) &= \frac{1}{\Delta x} \left(2q_{i-\frac{1}{2}}(t) - 6Q_i(t) + 4q_{i+\frac{1}{2}}(t)\right), \\ D^-q_{i+\frac{1}{2}}(t) &= \frac{1}{\Delta x} \left(-4q_{i+\frac{1}{2}}(t) + 6Q_{i+1}(t) - 2q_{i+\frac{3}{2}}(t)\right). \end{aligned}$$

Proof. First, we compute the derivative of the local reconstruction,

$$\frac{\partial}{\partial \xi}q_{\text{recon},i}(t, \xi) = \frac{1}{\Delta x} \left(q_{i+\frac{1}{2}}(t) - q_{i-\frac{1}{2}}(t)\right) + \frac{6}{\Delta x^2} \left(q_{i-\frac{1}{2}}(t) + q_{i+\frac{1}{2}}(t) - 2Q_i(t)\right) \xi.$$

Now substituting $\xi = \pm \frac{\Delta x}{2}$, we obtain

$$\begin{aligned} D^+q_{i+\frac{1}{2}}(t) &= \frac{\partial}{\partial \xi}q_{\text{recon},i} \left(t, \frac{\Delta x}{2}\right) \\ &= \frac{1}{\Delta x} \left(2q_{i-\frac{1}{2}}(t) - 6Q_i(t) + 4q_{i+\frac{1}{2}}(t)\right) \end{aligned}$$

and

$$\begin{aligned} D^-q_{i+\frac{1}{2}}(t) &= \frac{\partial}{\partial \xi}q_{\text{recon},i+1} \left(t, -\frac{\Delta x}{2}\right) \\ &= \frac{1}{\Delta x} \left(-4q_{i+\frac{1}{2}}(t) + 6Q_{i+1}(t) - 2q_{i+\frac{3}{2}}(t)\right). \end{aligned}$$

□

We are left with a system of ordinary differential equations

$$\frac{d}{dt}U(t) = R(U(t)),$$

where

$$U(t) = \begin{pmatrix} \vdots \\ Q_i(t) \\ q_{i+\frac{1}{2}}(t) \\ \vdots \end{pmatrix}$$

and

$$R(U(t)) = \begin{pmatrix} \vdots \\ -\frac{1}{\Delta x}(f(q_{i+\frac{1}{2}}(t)) - f(q_{i-\frac{1}{2}}(t))) \\ -\left(F^+(q_{i+\frac{1}{2}}(t))D^+q_{i+\frac{1}{2}}(t) + F^-(q_{i+\frac{1}{2}}(t))D^-q_{i+\frac{1}{2}}(t)\right) \\ \vdots \end{pmatrix}.$$

To obtain a fully-discrete scheme (not to be confused with the fully-discrete Active Flux method), we need to discretise this scheme in time. Usually this is done with, for example, a third stage, third order strong stability preserving Runge-Kutta (SSPRK3) scheme [46]. Assume that an approximation U^n to $U(t^n)$ at a given time step $t^n \geq 0$, $n \in \mathbb{N}_0$ is known. The solution at the next time step $t^{n+1} = t^n + \Delta t$ with time step size Δt computed with the SSPRK3 scheme is

$$\begin{aligned} U^{(1)} &= U^n + \Delta t R(U^n, t^n), \\ U^{(2)} &= \frac{3}{4}U^n + \frac{1}{4}\left(U^{(1)} + \Delta t R\left(U^{(1)}, t^n + \Delta t\right)\right), \\ U^{n+1} &= \frac{1}{3}U^n + \frac{2}{3}\left(U^{(2)} + \Delta t R\left(U^{(2)}, t^n + \frac{1}{2}\Delta t\right)\right). \end{aligned}$$

The time step size Δt is subject to a CFL condition [47]:

$$\Delta t < \frac{\Delta x}{\lambda_{\max}},$$

where $\lambda_{\max}$ is the largest characteristic speed in absolute value evaluated at the point values. In practise we compute the time step size at each time step by

$$\Delta t = C_{\text{CFL}} \frac{\Delta x}{\lambda_{\max}},$$

where C_{CFL} is the CFL number. In case of the linear advection equation,

$$\frac{\partial}{\partial t}q(t, x) + a \frac{\partial}{\partial x}q(t, x) = 0$$

with $a \in \mathbb{R} \setminus \{0\}$, we have $\lambda_{\max} = |a|$. In [17], the von Neumann stability of the semi-discrete Active Flux scheme with SSPRK3 as time discretisation was analysed and it was found that the scheme is stable up to $C_{\text{CFL}} = 0.41$.

2.4 IMEX Runge-Kutta schemes

Consider an initial value problem for an ordinary differential equation,

$$\begin{aligned}\frac{d}{dt}v(t) &= f_1(v(t)) + f_2(v(t)), \quad t > 0, \\ v(0) &= v_0,\end{aligned}$$

where $v : \mathbb{R}_0^+ \rightarrow \mathcal{O}$, $\mathcal{O} \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$ and $f_1, f_2 : \mathcal{O} \rightarrow \mathbb{R}^m$. The function f_1 is nonlinear and independent of a relaxation parameter ε and f_2 is linear and dependent on ε . To avoid strict stability constraints for the time step size, one can discretise the nonlinear part explicitly in time and the linear part depending on ε implicitly using implicit-explicit (IMEX) Runge-Kutta schemes [48], [49]. As the name suggests, these schemes are a combination of explicit and implicit Runge-Kutta schemes. Assume at time step $t^n \geq 0$, $n \in \mathbb{N}_0$, approximation to $v(t^n)$, denoted by v^n , are given. Following [48] and [49], an $\sigma \in \mathbb{N}$ stage IMEX Runge-Kutta scheme to compute the approximations v^{n+1} at the next time step t^{n+1} takes the form

$$\begin{aligned}v^{[k]} &= v^n + \Delta t \left(\sum_{j=1}^{k-1} \tilde{a}_{k,j} f_1(v^{[j]}) + \sum_{j=1}^k \hat{a}_{k,j} f_2(v^{[j]}) \right), \\ v^{n+1} &= v^n + \Delta t \left(\sum_{j=1}^{\sigma} \tilde{b}_j f_1(v^{[j]}) + \sum_{j=1}^{\sigma} \hat{b}_j f_2(v^{[j]}) \right),\end{aligned}\tag{2.12}$$

with $k = 1, \dots, \sigma$. The values $\tilde{a}_{k,j} \in \mathbb{R}$ and $\tilde{b}_j \in \mathbb{R}$, $k, j = 1, \dots, \sigma$ are the coefficients of the explicit Runge-Kutta scheme, $\hat{a}_{k,j} \in \mathbb{R}$ and $\hat{b}_j \in \mathbb{R}$ are the coefficients of the implicit Runge-Kutta scheme, and $v^{[k]}$ are called the stage solution. It is convenient to define

$$\hat{A} = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \tilde{a}_{2,1} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \tilde{a}_{\sigma,1} & \cdots & \tilde{a}_{\sigma,\sigma-1} & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} \tilde{b}_1 \\ \vdots \\ \tilde{b}_{\sigma} \end{pmatrix}$$

as well as

$$\hat{A} = \begin{pmatrix} \hat{a}_{1,1} & 0 & \cdots & \hat{0} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ \hat{a}_{\sigma,1} & \cdots & \cdots & \hat{a}_{\sigma,\sigma} \end{pmatrix} \quad \text{and} \quad \hat{b} = \begin{pmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_{\sigma} \end{pmatrix}$$

One can characterise IMEX Runge-Kutta schemes based on the structure of $\hat{A}$ and $\hat{b}$. The following definition is taken from [48].

Definition 2.21. *An IMEX Runge-Kutta scheme is of type I if the matrix $\hat{A}$ is invertible. It is said to be of type II if the matrix $\hat{A}$ can be written as*

$$\hat{A} = \begin{pmatrix} 0 & 0 \\ a & A \end{pmatrix}$$

with $a = (\hat{a}_{2,1}, \dots, \hat{a}_{\sigma,1})^T \in \mathbb{R}^{\sigma-1}$ and the sub-matrix $A \in \mathbb{R}^{(\sigma-1) \times (\sigma-1)}$ with elements $a_{k,j} = \hat{a}_{k+1,j+1}$ for $k, j = 1, \dots, \sigma-1$. The matrix A is required to be invertible. If, in addition, $a = 0$ and $\hat{b}_1 = 0$, the IMEX Runge-Kutta scheme is called of type ARS.

The focus of this text are IMEX Runge-Kutta schemes of type ARS since the properties of these schemes turn out to be useful for the later analysis of the Active Flux schemes. For such a scheme, we define $b = (b_2, \dots, b_\sigma) \in \mathbb{R}^s$ with $s = \sigma - 1$.

If the IMEX Runge-Kutta scheme is of type ARS, can rewrite (2.12) in a form that is more convenient to work with. We start by writing (2.12) as

$$\begin{aligned} v^{[k+1]} &= v^n + \Delta t \left(\sum_{j=1}^k \tilde{a}_{k+1,j} f_1(v^{[j]}) + \sum_{j=1}^{k+1} \hat{a}_{k+1,j} f_2(v^{[j]}) \right), \\ v^{n+1} &= v^n + \Delta t \left(\sum_{j=1}^{\sigma} \tilde{b}_j f_1(v^{[j]}) + \sum_{j=1}^{\sigma} \hat{b}_j f_2(v^{[j]}) \right) \end{aligned}$$

with $k = 0, \dots, s$. Next, we use the special structure of $\hat{A}$ to obtain

$$\begin{aligned} v^{[k+1]} &= v^n + \Delta t \left(\sum_{j=1}^k \tilde{a}_{k+1,j} f_1(v^{[j]}) + \sum_{j=2}^{k+1} \hat{a}_{k+1,j} f_2(v^{[j]}) \right), \\ v^{n+1} &= v^n + \Delta t \left(\sum_{j=1}^{\sigma} \tilde{b}_j f_1(v^{[j]}) + \sum_{j=2}^{\sigma} \hat{b}_j f_2(v^{[j]}) \right). \end{aligned}$$

Now, since $a_{i,j} = \hat{a}_{i+1,j+1}$ for $i, j = 1, \dots, \sigma - 1$, we have

$$\sum_{j=2}^{k+1} \hat{a}_{k+1,j} f_2(v^{[j]}) = \sum_{j=1}^k \hat{a}_{k+1,j+1} f_2(v^{[j+1]}) = \sum_{j=1}^k a_{k,j} f_2(v^{[j+1]})$$

and

$$\sum_{j=2}^{\sigma} \hat{b}_j f_2(v^{[j]}) = \sum_{j=1}^s \hat{b}_{j+1} f_2(v^{[j+1]}) = \sum_{j=1}^s b_j f_2(v^{[j+1]}),$$

where we used $b_j = \hat{b}_{j+1}$ for $j = 1, \dots, s$. This leaves us with

$$\begin{aligned} v^{[k+1]} &= v^n + \Delta t \left(\sum_{j=1}^k \tilde{a}_{k+1,j} f_1(v^{[j]}) + \sum_{j=1}^k a_{k,j} f_2(v^{[j+1]}) \right), \\ v^{n+1} &= v^n + \Delta t \left(\sum_{j=1}^{\sigma} \tilde{b}_j f_1(v^{[j]}) + \sum_{j=1}^s b_j f_2(v^{[j+1]}) \right). \end{aligned}$$

with $k = 0, \dots, s$ or equivalently

$$\begin{aligned} v^{[1]} &= v^n, \\ v^{[k+1]} &= v^n + \Delta t \left(\sum_{j=1}^k \tilde{a}_{k+1,j} f_1(v^{[j]}) + \sum_{j=1}^k a_{k,j} f_2(v^{[j+1]}) \right), \\ v^{n+1} &= v^n + \Delta t \left(\sum_{j=1}^{\sigma} \tilde{b}_j f_1(v^{[j]}) + \sum_{j=1}^s b_j f_2(v^{[j+1]}) \right) \end{aligned}$$

with $k = 1, \dots, s$. Lastly, defining the first stage to be $v^{[0]}$ instead of $v^{[1]}$, yields

$$\begin{aligned} v^{[0]} &= v^n, \\ v^{[k]} &= v^n + \Delta t \left(\sum_{j=1}^k \tilde{a}_{k+1,j} f_1(v^{[j-1]}) + \sum_{j=1}^k a_{k,j} f_2(v^{[j]}) \right), \\ v^{n+1} &= v^n + \Delta t \left(\sum_{j=1}^{\sigma} \tilde{b}_j f_1(v^{[j-1]}) + \sum_{j=1}^s b_j f_2(v^{[j]}) \right), \end{aligned}$$

which is the form of an IMEX-Runge Kutta scheme used in this text. For simplicity, the first equation $v^{[0]} = v^n$ is not explicitly stated. Instead, it is assumed to hold if we employ a IMEX Runge-Kutta scheme of type ARS to discretise in time.

In the special case, where $\tilde{b}_{\sigma} = 0$, the last equation reduces to

$$v^{n+1} = v^n + \Delta t \left(\sum_{j=1}^s \tilde{b}_j f_1(v^{[j-1]}) + \sum_{j=1}^s b_j f_2(v^{[j]}) \right).$$

Definition 2.22. An IMEX Runge-Kutta scheme of type ARS with $\tilde{b}_{\sigma} = 0$ is called *stiffly accurate* if

$$a_{s,j} = b_j$$

for all $j = 1, \dots, s$. In addition, the scheme is called *globally stiffly accurate* if

$$\tilde{a}_{\sigma,j} = \tilde{b}_j$$

for all $j = 1, \dots, s$.

Lemma 2.23. If an IMEX Runge-Kutta scheme of type ARS is globally stiffly accurate, then

$$v^{n+1} = v^{[s]}.$$

Proof. Given a globally stiffly accurate scheme, we have

$$\begin{aligned} v^{[s]} &= v^n + \Delta t \left(\sum_{j=1}^s \tilde{a}_{\sigma,j} f_1(v^{[j-1]}) + \sum_{j=1}^s a_{s,j} f_2(v^{[j]}) \right) \\ &= v^n + \Delta t \left(\sum_{j=1}^s \tilde{b}_j f_1(v^{[j-1]}) + \sum_{j=1}^s b_j f_2(v^{[j]}) \right) \end{aligned}$$

as well as

$$v^{n+1} = v^n + \Delta t \left(\sum_{j=1}^s \tilde{b}_j f_1(v^{[j-1]}) + \sum_{j=1}^s b_j f_2(v^{[j]}) \right),$$

and, thus, $v^{n+1} = v^{[s]}$. □

The triple (s, σ, p) is used to denote different IMEX Runge-Kutta schemes, where p stands for the combined order of the scheme and s and σ are defined as before. In the case of $b_{\sigma} = 0$ we write $(s, \sigma - 1, p)$ instead. The simplest example of an IMEX Runge-Kutta scheme is perhaps the forward-backward Euler or ARS(1, 1, 1) scheme with coefficients

$$A = 1 \qquad \text{and} \qquad b = 1$$

and

$$\tilde{A} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \tilde{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

A second order, globally stiffly accurate IMEX Runge-Kutta scheme is, for example, the ARS(2, 2, 2) scheme that is obtained by the following coefficient choice:

$$A = \begin{pmatrix} \gamma & 0 \\ 1 - \gamma & \gamma \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 - \gamma \\ \gamma \end{pmatrix}$$

as well as

$$\tilde{A} = \begin{pmatrix} 0 & 0 & 0 \\ \gamma & 0 & 0 \\ \delta & 1 - \delta & 0 \end{pmatrix} \quad \text{and} \quad \tilde{b} = \begin{pmatrix} \delta \\ 1 - \delta \\ 0 \end{pmatrix}$$

with $\gamma = \frac{2-\sqrt{2}}{2}$ and $\delta = 1 - \frac{1}{2\gamma}$. A third order globally stiffly accurate scheme is the ARS(4, 4, 3) scheme with coefficients

$$A = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{3}{2} & -\frac{3}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} \frac{3}{2} \\ -\frac{3}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

and

$$\tilde{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{11}{18} & \frac{1}{18} & 0 & 0 & 0 \\ \frac{5}{6} & -\frac{5}{6} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{7}{4} & \frac{3}{4} & -\frac{7}{4} & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} \frac{1}{4} \\ \frac{7}{4} \\ \frac{3}{4} \\ -\frac{7}{4} \\ 0 \end{pmatrix}.$$

An different third order scheme is the ARS(2, 3, 3) scheme, which is not globally stiffly accurate. The coefficients of this scheme are

$$A = \begin{pmatrix} \gamma & 0 \\ 1 - 2\gamma & \gamma \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

and

$$\tilde{A} = \begin{pmatrix} 0 & 0 & 0 \\ \gamma & 0 & 0 \\ \gamma - 1 & 2(1 - \gamma) & 0 \end{pmatrix} \quad \text{and} \quad \tilde{b} = \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

with $\gamma = \frac{3+\sqrt{3}}{6}$. The presented schemes are taken from [49], where additional information and schemes can be found.

2.5 Asymptotic preserving numerical schemes

After introducing the semi-discrete Active Flux scheme to discretise in space and IMEX Runge-Kutta schemes to discretise in time, this chapter stated under which conditions a (fully-discrete) numerical scheme is called asymptotic preserving. The next definition which provides these conditions is based on [50].

Definition 2.24. Assume $\mathcal{A}^\varepsilon$ is a given partial differential equation that depends on a relaxation parameter ε . The limit of this partial differential equation as $\varepsilon \rightarrow 0$ is denoted by $\mathcal{A}^0$. Suppose $\mathcal{A}_{\Delta x}^\varepsilon$ is a (consistent and stable) numerical scheme that discretises $\mathcal{A}^\varepsilon$, where Δx denotes the spatial step size. In addition, assume the time step size Δt is dependent on Δx . If the limit of the numerical discretisation $\mathcal{A}_{\Delta x}^\varepsilon$ as $\varepsilon \rightarrow 0$, denoted by $\mathcal{A}_{\Delta x}^0$, exists and is a (consistent and stable) approximation of $\mathcal{A}^0$, then the numerical scheme $\mathcal{A}_{\Delta x}^\varepsilon$ is called asymptotic preserving.

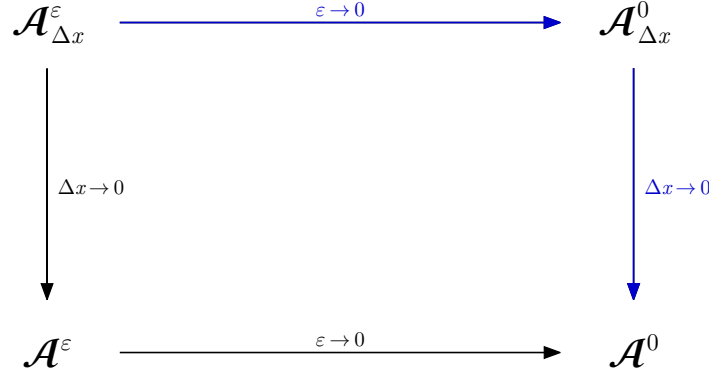


Figure 7: Illustration of the definition of asymptotic preserving numerical schemes.

This definition is illustrated in Figure 7, which was inspired by [50]. The highlighted arrows correspond to the asymptotic preserving property of the scheme. For example, in this thesis, $\mathcal{A}^\varepsilon$ stands for the HKdV equations, $\mathcal{A}_{\Delta x}^\varepsilon$ for a discretisation of the HKdV equations, for example, with an Active Flux and IMEX Runge-Kutta scheme, and $\mathcal{A}^0$ represents the limit equations of the HKdV equations.

To further illustrate asymptotic preserving property, consider the error of the numerical scheme with respect to the relaxation parameter ε . In Figure 8, four different cases are depicted. We assume the illustrated behaviour is similar for all spatial step sizes Δx . The two depictions in the top row of the figure represent schemes that are in our definition not asymptotic preserving since in the limit $\varepsilon \rightarrow 0$ the error increases or attains a constant value independent of the chosen spatial step size. The third and fourth case illustrate numerical schemes, that are asymptotic preserving. The two differ in the order of accuracy that the numerical scheme achieves for sufficiently small ε . If a numerical scheme is, for example, able to achieve third order accuracy for sufficiently large values of ε , this does not immediately imply that the same holds as $\varepsilon \rightarrow 0$. Thus, it is natural to distinguish scheme that retain their order of accuracy if $\varepsilon \rightarrow 0$ from those who do not. The next definition is taken from [39].

Definition 2.25. Assume $\mathcal{A}_{\Delta x}^\varepsilon$ is an asymptotic preserving numerical scheme. If the scheme retains its order of accuracy as $\varepsilon \rightarrow 0$, it is called asymptotic accurate.

For the theoretical analysis it is convenient to consider a weaker form of Definition 2.24. Assume $\mathcal{A}_\varepsilon$ is a system of hyperbolic conservation laws with components $q_1, \dots, q_n$, $n \in \mathbb{N}$. Inspired by [39], a numerical scheme $\mathcal{A}_{\Delta x}^\varepsilon$ is called asymptotic preserving for the component q_k , $k \in \{1, \dots, n\}$ if the limit $\mathcal{A}_{\Delta x}^0$ as $\varepsilon \rightarrow 0$ of $\mathcal{A}_{\Delta x}^\varepsilon$ exists and, in addition, the part of $\mathcal{A}_{\Delta x}^0$ associated to q_k is a (consistent and stable) approximation of the equation associated to q_k in the system of hyperbolic conservation laws.

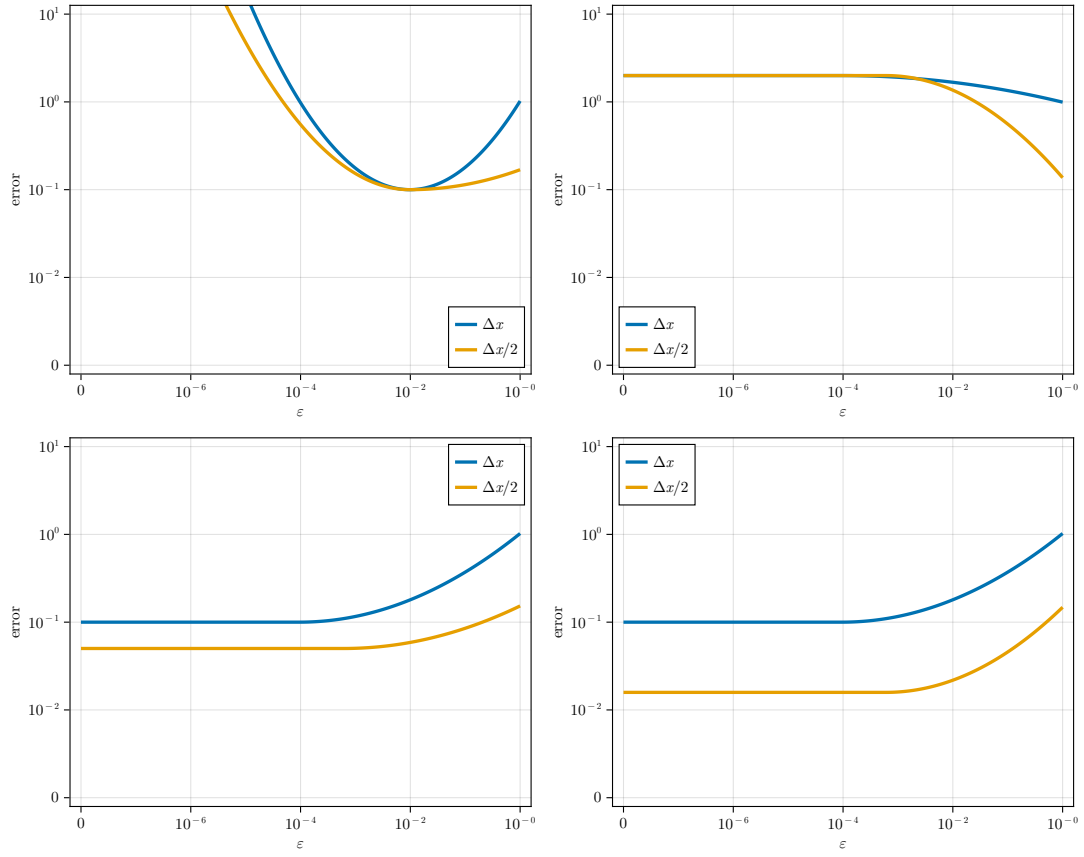


Figure 8: Illustration of error behaviour with respect to ε for different numerical schemes. Top row: Numerical schemes that are not asymptotic preserving. Bottom row: Numerical schemes that are asymptotic preserving.

3 An asymptotic-preserving IMEX Active Flux scheme

This chapter is devoted to the development of an IMEX Active Flux scheme following a classical approach based on Jacobian splitting and the investigation of its properties. However, for the sake of completeness and as an intermediate step towards the implicit-explicit scheme, this section begins by introducing a time-explicit Active Flux scheme for the HKdV equations. This allows the demonstration of the disadvantages of the time-explicit scheme and the relation between the time explicit and the IMEX Active Flux scheme.

3.1 Time-explicit Active Flux scheme

The semi-discrete Active Flux scheme presented in the previous chapter is designed for homogeneous hyperbolic conservation laws and the HKdV equations are not a homogeneous conservation law since they have an additional linear source term. However, similar as in the previous chapter, a semi-discrete Active Flux scheme that incorporates the linear source term can be derived by following the approach of [24].

Integrating the HKdV equations over one cell $\mathcal{C}_i$, where again $i \in \mathbb{Z}$, we obtain

$$\frac{d}{dt} \int_{\mathcal{C}_i} q(t, x) \, dx + f(q(t, x_{i+\frac{1}{2}})) - f(q(t, x_{i-\frac{1}{2}})) = S \int_{\mathcal{C}_i} q(t, x) \, dx.$$

From this form of the conservation law we deduce the update equations for the cell averages, which are

$$\frac{d}{dt} Q_i(t) = - \frac{f(q_{i+\frac{1}{2}}(t)) - f(q_{i-\frac{1}{2}}(t))}{\Delta x} + S Q_i(t)$$

or

$$\begin{aligned} \frac{d}{dt} U_i(t) &= - \frac{1}{\Delta x} \left(\frac{1}{2} (u_{i+\frac{1}{2}}(t))^2 + \psi_{i+\frac{1}{2}}(t) - \frac{1}{2} (u_{i-\frac{1}{2}}(t))^2 - \psi_{i-\frac{1}{2}}(t) \right), \\ \frac{d}{dt} P_i(t) &= - \frac{1}{\Delta x} \left(- \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}(t) + \frac{1}{\varepsilon^2} \rho_{i-\frac{1}{2}}(t) \right) - \frac{1}{\varepsilon^2} \Psi_i(t), \\ \frac{d}{dt} \Psi_i(t) &= - \frac{1}{\Delta x} \left(\frac{1}{\varepsilon^2} u_{i+\frac{1}{2}}(t) - \frac{1}{\varepsilon^2} u_{i-\frac{1}{2}}(t) \right) + \frac{1}{\varepsilon^2} P_i(t), \end{aligned}$$

where the cell averages are denoted by

$$Q_i(t) = (U_i(t), P_i(t), \Psi_i(t))^T$$

and the point values by

$$q_{i+\frac{1}{2}}(t) = \left(u_{i+\frac{1}{2}}(t), \rho_{i+\frac{1}{2}}(t), \psi_{i+\frac{1}{2}}(t) \right)^T.$$

To state the update equations of the latter one, we need to compute the positive and negative part of the Jacobian. Therefore, we review the eigenvalues of the Jacobian:

$$\lambda_1(q) = -\frac{1}{\varepsilon^2}, \quad \lambda_2(q) = \frac{1}{2} \left(u - \sqrt{u^2 + \frac{4}{\varepsilon^2}} \right), \quad \lambda_3(q) = \frac{1}{2} \left(u + \sqrt{u^2 + \frac{4}{\varepsilon^2}} \right).$$

The first and second eigenvalue are always negative and the third one is always positive since

$$u < \sqrt{u^2 + \frac{4}{\varepsilon^2}}$$

for all $u \in \mathbb{R}$ and $\varepsilon > 0$. Using the same notation as in the chapter before, we have

$$\Lambda^+(q) = \text{diag}(0, 0, \lambda_3(q))$$

and

$$\Lambda^-(q) = \text{diag}(\lambda_1(q), \lambda_2(q), 0).$$

The positive and negative part of the Jacobian F are

$$F^+(q) = \begin{pmatrix} \frac{(u + \sqrt{u^2 + \frac{4}{\varepsilon^2}})^2}{4\sqrt{u^2 + \frac{4}{\varepsilon^2}}} & 0 & \frac{1}{2} \left(\frac{u}{\sqrt{u^2 + \frac{4}{\varepsilon^2}}} + 1 \right) \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} \left(\frac{u}{\sqrt{u^2 + \frac{4}{\varepsilon^2}}} + 1 \right) & 0 & \frac{1}{\varepsilon^2 \sqrt{u^2 + \frac{4}{\varepsilon^2}}} \end{pmatrix}$$

and

$$F^-(q) = \begin{pmatrix} -\frac{(-u + \sqrt{u^2 + \frac{4}{\varepsilon^2}})^2}{4\sqrt{u^2 + \frac{4}{\varepsilon^2}}} & 0 & \frac{1}{2} \left(-\frac{u}{\sqrt{u^2 + \frac{4}{\varepsilon^2}}} + 1 \right) \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} \left(-\frac{u}{\sqrt{u^2 + \frac{4}{\varepsilon^2}}} + 1 \right) & 0 & -\frac{1}{\varepsilon^2 \sqrt{u^2 + \frac{4}{\varepsilon^2}}} \end{pmatrix}.$$

Mathematica [37] was employed to assist in the computation of F^+ and F^- . Given the positive and negative part of the Jacobian, the update equations of the point values are given by

$$\frac{d}{dt} q_{i+\frac{1}{2}}(t) = - \left(F^+(q_{i+\frac{1}{2}}(t)) D^+ q_{i+\frac{1}{2}}(t) + F^-(q_{i+\frac{1}{2}}(t)) D^- q_{i+\frac{1}{2}}(t) \right) + S q_{i+\frac{1}{2}}(t),$$

which is a semi-discretisation of the HKdV equations in quasi-linear form. Equivalently, we can write

$$\begin{aligned} \frac{d}{dt} u_{i+\frac{1}{2}}(t) &= - \frac{(u + \sqrt{u^2 + \frac{4}{\varepsilon^2}})^2}{4\sqrt{u^2 + \frac{4}{\varepsilon^2}}} D^+ u_{i+\frac{1}{2}}(t) + \frac{(-u + \sqrt{u^2 + \frac{4}{\varepsilon^2}})^2}{4\sqrt{u^2 + \frac{4}{\varepsilon^2}}} D^- u_{i+\frac{1}{2}}(t) \\ &\quad - \frac{1}{2} \left(\frac{u}{\sqrt{u^2 + \frac{4}{\varepsilon^2}}} + 1 \right) D^+ \psi_{i+\frac{1}{2}}(t) - \frac{1}{2} \left(-\frac{u}{\sqrt{u^2 + \frac{4}{\varepsilon^2}}} + 1 \right) D^- \psi_{i+\frac{1}{2}}(t), \\ \frac{d}{dt} \rho_{i+\frac{1}{2}}(t) &= \frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}(t) - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}(t), \\ \frac{d}{dt} \psi_{i+\frac{1}{2}}(t) &= - \frac{1}{2\varepsilon^2} \left(\frac{u}{\sqrt{u^2 + \frac{4}{\varepsilon^2}}} + 1 \right) D^+ u_{i+\frac{1}{2}}(t) - \frac{1}{2\varepsilon^2} \left(-\frac{u}{\sqrt{u^2 + \frac{4}{\varepsilon^2}}} + 1 \right) D^- u_{i+\frac{1}{2}}(t) \\ &\quad - \frac{1}{\varepsilon^2 \sqrt{u^2 + \frac{4}{\varepsilon^2}}} D^+ \psi_{i+\frac{1}{2}}(t) + \frac{1}{\varepsilon^2 \sqrt{u^2 + \frac{4}{\varepsilon^2}}} D^- \psi_{i+\frac{1}{2}}(t) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}(t). \end{aligned}$$

To discretise the semi-discrete update equations in time, we employ the SSPRK3 scheme. A requirement for the resulting fully-discrete scheme to be stable is that the time step size $\Delta t = t^{n+1} - t^n$ is chosen as

$$\Delta t = C_{\text{CFL}} \frac{\Delta x}{\lambda_{\max}},$$

where

$$\lambda_{\max} = \max_{i \in \{1,2,3\}} \max_{q \in \Gamma} |\lambda_i(q)|,$$

with Γ defined as the set of all point values at time step t^n . The eigenvalues depend on ε and as $\varepsilon \rightarrow 0$, $|\lambda_i(q)| \rightarrow \infty$ for any $q \in \mathbb{R}^3$. Thus, if $\varepsilon \rightarrow 0$, it follows that $\Delta t \rightarrow 0$, which results in an unnecessary high runtime for small values of ε .

3.2 Classical IMEX Active Flux scheme

To avoid the unnecessary high runtime for small values of ε , the aim of this thesis is to discretise the semi-discrete Active Flux scheme in time using IMEX Runge-Kutta schemes of type ARS. A requirement to employ such schemes is that the semi-discrete update equations can be split into a nonlinear part independent of ε and a linear part that is allowed to depend on the relaxation parameter. While the update equations of the cell averages are given in a form that could be easily split into these two parts, the update equations of the point values cannot be divided directly. Thus, a different approach to derive the update equation for the degrees of freedom is used. Inspired by [39], we start by splitting the HKdV equations into two parts:

$$\frac{\partial}{\partial t} q(t, x) + \frac{\partial}{\partial x} f_1(q(t, x)) + \frac{\partial}{\partial x} f_2(q(t, x)) = Sq(t, x), \quad (3.1)$$

where q and S are defined as usual and $f_1, f_2 : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ with

$$f_1(q) = \begin{pmatrix} \frac{1}{2}u^2 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad f_2(q) = \begin{pmatrix} \psi \\ -\frac{1}{\varepsilon^2}\rho \\ \frac{1}{\varepsilon^2}u \end{pmatrix}$$

for $q = (u, \rho, \psi)^T \in \mathbb{R}^3$. Notice that f_1 is independent of ε and f_2 is linear. The general idea to obtain the update equations is to discretise the two spatial derivatives independently from each other. Computing the integral form of this equation allows us to deduce update equations for the cell averages that read

$$\frac{d}{dt} Q_i(t) = -\frac{1}{\Delta x} \left(f_1(q_{i+\frac{1}{2}}(t)) - f_1(q_{i-\frac{1}{2}}(t)) \right) - \frac{1}{\Delta x} \left(f_2(q_{i+\frac{1}{2}}(t)) - f_2(q_{i-\frac{1}{2}}(t)) \right) + SQ_i(t)$$

or

$$\begin{aligned} \frac{d}{dt} U_i(t) &= -\frac{1}{\Delta x} \left(\frac{1}{2}(u_{i+\frac{1}{2}}(t))^2 - \frac{1}{2}(u_{i-\frac{1}{2}}(t))^2 \right) - \frac{1}{\Delta x} \left(\psi_{i+\frac{1}{2}}(t) - \psi_{i-\frac{1}{2}}(t) \right), \\ \frac{d}{dt} P_i(t) &= -\frac{1}{\Delta x} \left(-\frac{1}{\varepsilon^2}\rho_{i+\frac{1}{2}}(t) + \frac{1}{\varepsilon^2}\rho_{i-\frac{1}{2}}(t) \right) - \frac{1}{\varepsilon^2} \Psi_i(t), \\ \frac{d}{dt} \Psi_i(t) &= -\frac{1}{\Delta x} \left(\frac{1}{\varepsilon^2}u_{i+\frac{1}{2}}(t) - \frac{1}{\varepsilon^2}u_{i-\frac{1}{2}}(t) \right) + \frac{1}{\varepsilon^2} P_i(t), \end{aligned}$$

These are the same update equations obtained in context of the explicit Active Flux scheme, split into a nonlinear and linear part. However, the difference lies in the update equations of the point values. In quasi-linear form (3.1) reads

$$\frac{\partial}{\partial t} q(t, x) + F_1(q(t, x)) \frac{\partial}{\partial x} q(t, x) + F_2 \frac{\partial}{\partial x} q(t, x) = Sq(t, x),$$

where $F_1 : \mathbb{R}^3 \rightarrow \mathbb{R}^{3 \times 3}$ with

$$F_1(q) = \begin{pmatrix} u & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad F_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{\varepsilon^2} & 0 & 0 \end{pmatrix}$$

are the Jacobians of f_1 and f_2 . The Jacobian F_2 is independent of q and hence we dropped the dependence on it. Using the classical approach based on Jacobian splitting, we can determine the positive and negative part of the two Jacobians, where Mathematica [37] is used to assist in the computation. Starting with F_1 , we have

$$F_1^+(q) = \begin{pmatrix} \max(u, 0) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad F_1^-(q) = \begin{pmatrix} \min(u, 0) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

For $\varepsilon > 0$, the matrix F_2 has three real and distinct eigenvalues,

$$\lambda_1 = -\frac{1}{\varepsilon^2}, \quad \lambda_2 = -\frac{1}{\varepsilon}, \quad \lambda_3 = \frac{1}{\varepsilon}$$

and three linearly independent eigenvectors

$$v_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -\varepsilon \\ 0 \\ 1 \end{pmatrix}, \quad v_3 = \begin{pmatrix} \varepsilon \\ 0 \\ 1 \end{pmatrix}.$$

This allows us to split F_2 into a positive and negative part:

$$F_2^+ = \begin{pmatrix} \frac{1}{2\varepsilon} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} & 0 & \frac{1}{2\varepsilon} \end{pmatrix} \quad \text{and} \quad F_2^- = \begin{pmatrix} -\frac{1}{2\varepsilon} & 0 & \frac{1}{2} \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & -\frac{1}{2\varepsilon} \end{pmatrix}.$$

We define the updates equations of the point values to be

$$\begin{aligned} \frac{d}{dt} q_{i+\frac{1}{2}}(t) = & -(F_1^+ D^+ q_{i+\frac{1}{2}}(t) + F_1^- D^- q_{i+\frac{1}{2}}(t)) \\ & - (F_2^+ D^+ q_{i+\frac{1}{2}}(t) + F_2^- D^- q_{i+\frac{1}{2}}(t)) + S q_{i+\frac{1}{2}}(t) \end{aligned}$$

or equivalently

$$\begin{aligned} \frac{d}{dt} u_{i+\frac{1}{2}}(t) = & -\max(u_{i+\frac{1}{2}}, 0) D^+ u_{i+\frac{1}{2}}(t) - \min(u_{i+\frac{1}{2}}, 0) D^- u_{i+\frac{1}{2}}(t) \\ & - \frac{1}{2\varepsilon} (D^+ u_{i+\frac{1}{2}}(t) - D^- u_{i+\frac{1}{2}}(t)) - \frac{1}{2} (D^+ \psi_{i+\frac{1}{2}}(t) + D^- \psi_{i+\frac{1}{2}}(t)). \\ \frac{d}{dt} \rho_{i+\frac{1}{2}}(t) = & \frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}(t) - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}(t), \\ \frac{d}{dt} \psi_{i+\frac{1}{2}}(t) = & -\frac{1}{2\varepsilon^2} (D^+ u_{i+\frac{1}{2}}(t) + D^- u_{i+\frac{1}{2}}(t)) - \frac{1}{2\varepsilon} (D^+ \psi_{i+\frac{1}{2}}(t) - D^- \psi_{i+\frac{1}{2}}(t)) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}(t). \end{aligned}$$

These are not the same update equations as if one split the Jacobian F into a positive and negative part since

$$F^\pm(q) \neq F_1^\pm(q) + F_2^\pm.$$

However, the resulting update equations are closely related. Starting from $F^\pm$, assume that the solution u to the HKdV equations is bounded and with the help of Mathematica [37] expand the square roots appearing in $F^\pm$ as power series in ε , i.e.

$$\left(u^2 + \frac{4}{\varepsilon^2}\right)^{\frac{1}{2}} = \frac{2}{\varepsilon} \left(1 + \left(\frac{\varepsilon u}{2}\right)^2\right)^{\frac{1}{2}} = \frac{2}{\varepsilon} + \frac{u^2 \varepsilon}{4} + \frac{u^4 \varepsilon^3}{64} + \mathcal{O}(\varepsilon^5)$$

and

$$\left(u^2 + \frac{4}{\varepsilon^2}\right)^{-\frac{1}{2}} = \frac{\varepsilon}{2} \left(1 + \left(\frac{\varepsilon u}{2}\right)^2\right)^{-\frac{1}{2}} = \frac{\varepsilon}{2} - \frac{u^2 \varepsilon^3}{16} + \frac{3u^4 \varepsilon^5}{256} + \mathcal{O}(\varepsilon^7).$$

This gives

$$\begin{aligned}
F^+(q) &= \begin{pmatrix} \frac{1}{4} \left(u + \frac{2}{\varepsilon} + \mathcal{O}(\varepsilon)\right)^2 \left(\frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^3)\right) & 0 & \frac{1}{2} \left(u \left(\frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^3)\right) + 1\right) \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} \left(u \left(\frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^3)\right) + 1\right) & 0 & \frac{1}{\varepsilon^2} \left(\frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^3)\right) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{2\varepsilon} + \frac{u}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} + \frac{1}{4\varepsilon}u & 0 & \frac{1}{2\varepsilon} \end{pmatrix} + \mathcal{O}(\varepsilon)
\end{aligned}$$

for the positive part and

$$\begin{aligned}
F^-(q) &= \begin{pmatrix} -\frac{1}{4} \left(-u + \frac{2}{\varepsilon} + \mathcal{O}(\varepsilon)\right)^2 \left(\frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^3)\right) & 0 & \frac{1}{2} \left(-u \left(\frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^3)\right) + 1\right) \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} \left(-u \left(\frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^3)\right) + 1\right) & 0 & -\frac{1}{\varepsilon^2} \left(\frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^3)\right) \end{pmatrix} \\
&= \begin{pmatrix} -\frac{1}{2\varepsilon} + \frac{u}{2} & 0 & \frac{1}{2} \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} - \frac{1}{4\varepsilon}u & 0 & -\frac{1}{2\varepsilon} \end{pmatrix} + \mathcal{O}(\varepsilon)
\end{aligned}$$

for the negative part. Dropping the terms of factor ε and the terms $\pm \frac{1}{4\varepsilon}u$ in the first element of the last row of both matrices as well as replacing $\frac{u}{2}$ in the first element of the first row by $\max(u, 0)$ or $\min(u, 0)$ in F^+ and F^- respectively, yields $F_1^\pm + F_2^\pm$.

To distinguish a semi-discrete Active Flux scheme with these update equations of the point values and cell averages from other semi-discrete Active Flux schemes, it is called the classical IMEX Active Flux scheme. In this case IMEX does not refer to the time-discretisation but to the splitting which directly allows the usage of an IMEX Runge-Kutta scheme to discretise in time. Nevertheless, it is possible to discretise the classical IMEX Active Flux scheme in time with an explicit Runge-Kutta scheme.

3.3 Formal asymptotic analysis

The next step and the second goal of this thesis is to analyse if the classical IMEX Active Flux scheme discretised in time with an IMEX Runge-Kutta scheme of type ARS is asymptotic preserving. Before the scheme is implemented, one can obtain a theoretical indication by performing formal asymptotic analysis. This section follows the approaches of [24] and [39] and Mathematica [37] is used to assist in the analysis.

Remark 3.1. In [39], the authors investigated with the help of formal asymptotic analysis the asymptotic preserving properties of IMEX Runge-Kutta schemes of type I and type II applied to the space continuous HKdV equations and the HKdV equations discretised in space with periodic summation-by-parts (SPB) operators. They found that the analysis indicates that an IMEX Runge Kutta scheme of type I applied to the HKdV equations or their discretisation of the HKdV equations is asymptotic preserving for the component u . If the scheme is globally stiffly accurate, then the scheme is asymptotic preserving. In addition, a IMEX Runge Kutta scheme of type II applied to the HKdV equations or their discretisation of the HKdV equations is asymptotic preserving if the IMEX Runge-Kutta scheme is globally stiffly accurate and the initial data is well prepared. In case of an Active Flux scheme, we have different degrees of freedom and different update equations. Thus, the results of this article do not immediately generalise to the classical IMEX Active Flux scheme considered here. In addition, using IMEX Runge Kutta schemes of type ARS, which are also type II schemes, leads to an indication similar to the one the authors of [39] obtained for type I schemes. In particular, in the analysis performed in this text, the assumption on the initial data to be well prepared is not needed.

Remark 3.2. Notice that the classical IMEX Active Flux is not specifically designed to be an asymptotic preserving scheme. Provided it exists, no information about the limit scheme is

known at this point. Hence, in contrast to [39], the goal of the formal asymptotic analysis is not to verify that the numerical approximations obtained by the classical IMEX Active Flux scheme tend numerical approximations of the limit equations discretised by the limit scheme. Instead, it is investigated whether a limit scheme exists and is a discretisation of the limit equations.

Remark 3.3. In [24], the authors performed formal asymptotic analysis only for the semi-discrete Active Flux scheme. However, in this thesis, it is performed for the fully-discrete scheme discretised in time with IMEX Runge-Kutta schemes of type ARS since this approach provides more informative results.

To begin with the formal asymptotic analysis, we denote the point values and cell averages at time step t^n , $n \in \mathbb{N}_0$, by

$$q_{i+\frac{1}{2}}^n = \begin{pmatrix} u_{i+\frac{1}{2}}^n \\ \rho_{i+\frac{1}{2}}^n \\ \psi_{i+\frac{1}{2}}^n \end{pmatrix}$$

and

$$Q_i^n = \begin{pmatrix} U_i^n \\ P_i^n \\ \Psi_i^n \end{pmatrix}$$

with $i \in \mathbb{Z}$. In addition, the temporal step size is denoted by $\Delta t = t^{n+1} - t^n$. Using a IMEX Runge-Kutta scheme of type ARS, approximations to cell averages at the next time step t^{n+1} are given by

$$\begin{aligned} U_i^{[k]} &= U_i^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{[j-1]} \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{[j-1]} \right)^2 \right) \right) \\ &\quad + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{\Delta x} \left(\psi_{i+\frac{1}{2}}^{[j]} - \psi_{i-\frac{1}{2}}^{[j]} \right) \right), \\ U_i^{n+1} &= U_i^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{[j-1]} \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{[j-1]} \right)^2 \right) \right) \\ &\quad + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{\Delta x} \left(\psi_{i+\frac{1}{2}}^{[j]} - \psi_{i-\frac{1}{2}}^{[j]} \right) \right) \end{aligned}$$

and

$$\begin{aligned} P_i^{[k]} &= P_i^n + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{\Delta x} \left(-\frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}^{[j]} + \frac{1}{\varepsilon^2} \rho_{i-\frac{1}{2}}^{[j]} \right) - \frac{1}{\varepsilon^2} \Psi_i^{[j]} \right), \\ P_i^{n+1} &= P_i^n + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{\Delta x} \left(-\frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}^{[j]} + \frac{1}{\varepsilon^2} \rho_{i-\frac{1}{2}}^{[j]} \right) - \frac{1}{\varepsilon^2} \Psi_i^{[j]} \right) \end{aligned}$$

and

$$\begin{aligned} \Psi_i^{[k]} &= \Psi_i^n + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{\Delta x} \left(\frac{1}{\varepsilon^2} u_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2} u_{i-\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} P_i^{[j]} \right), \\ \Psi_i^{n+1} &= \Psi_i^n + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{\Delta x} \left(\frac{1}{\varepsilon^2} u_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2} u_{i-\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} P_i^{[j]} \right) \end{aligned}$$

with $k = 1, \dots, s$ and $i \in \mathbb{Z}$. Approximations to the point values at the next time step are computed by

$$\begin{aligned} u_{i+\frac{1}{2}}^{[k]} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\max \left(u_{i+\frac{1}{2}}^{[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{[j-1]} - \min \left(u_{i+\frac{1}{2}}^{[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{[j-1]} \right) \\ &\quad + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{2\varepsilon} \left(D^+ u_{i+\frac{1}{2}}^{[j]} - D^- u_{i+\frac{1}{2}}^{[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{[j]} + D^- \psi_{i+\frac{1}{2}}^{[j]} \right) \right), \\ u_{i+\frac{1}{2}}^{n+1} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\max \left(u_{i+\frac{1}{2}}^{[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{[j-1]} - \min \left(u_{i+\frac{1}{2}}^{[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{[j-1]} \right) \\ &\quad + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{2\varepsilon} \left(D^+ u_{i+\frac{1}{2}}^{[j]} - D^- u_{i+\frac{1}{2}}^{[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{[j]} + D^- \psi_{i+\frac{1}{2}}^{[j]} \right) \right) \end{aligned}$$

and

$$\begin{aligned} \rho_{i+\frac{1}{2}}^{[k]} &= \rho_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^k a_{k,j} \left(\frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}^{[j]} \right), \\ \rho_{i+\frac{1}{2}}^{n+1} &= \rho_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^s b_j \left(\frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}^{[j]} \right) \end{aligned}$$

and

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{[k]} &= \psi_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{2\varepsilon^2} \left(D^+ u_{i+\frac{1}{2}}^{[j]} + D^- u_{i+\frac{1}{2}}^{[j]} \right) \right. \\ &\quad \left. - \frac{1}{2\varepsilon} \left(D^+ \psi_{i+\frac{1}{2}}^{[j]} - D^- \psi_{i+\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}^{[j]} \right), \\ \psi_{i+\frac{1}{2}}^{n+1} &= \psi_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{2\varepsilon^2} \left(D^+ u_{i+\frac{1}{2}}^{[j]} + D^- u_{i+\frac{1}{2}}^{[j]} \right) \right. \\ &\quad \left. - \frac{1}{2\varepsilon} \left(D^+ \psi_{i+\frac{1}{2}}^{[j]} - D^- \psi_{i+\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}^{[j]} \right) \end{aligned}$$

with $k = 1, \dots, s$ and $i \in \mathbb{Z}$. In the following, $k = 1, \dots, s$ and $i \in \mathbb{Z}$ is not stated every time, but it is henceforth assume to be given. In addition, since we discretise in time with an IMEX-Runge Kutta scheme of type ARS, the zeroth stage solution is equal to the approximation at time step t^n , e.g. $U_i^{[0]} = U_i^n$.

For the next steps it is convenient to rewrite the time-discrete update equations of the point values and cell averages corresponding to the components ρ and ψ . Therefore, we define the vector of stage solutions

$$\bar{\rho}_{i+\frac{1}{2}} = \begin{pmatrix} \rho_{i+\frac{1}{2}}^{[1]} \\ \vdots \\ \rho_{i+\frac{1}{2}}^{[s]} \end{pmatrix}$$

and $\bar{\psi}_{i+\frac{1}{2}}$, $\bar{P}_i$, and $\bar{\Psi}_i$ in a similar way. With the help of this notion, we obtain

$$\bar{P}_i = P_i^n \mathbb{I}_s + \Delta t A \left(-\frac{1}{\Delta x} \left(-\frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}^{[j]} + \frac{1}{\varepsilon^2} \rho_{i-\frac{1}{2}}^{[j]} \right) - \frac{1}{\varepsilon^2} \Psi_i^{[j]} \right)_{j=1, \dots, s},$$

$$P_i^{n+1} = P_i^n + \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(-\frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}^{[j]} + \frac{1}{\varepsilon^2} \rho_{i-\frac{1}{2}}^{[j]} \right) - \frac{1}{\varepsilon^2} \Psi_i^{[j]} \right)_{j=1, \dots, s}$$

and

$$\begin{aligned} \bar{\Psi}_i &= \Psi_i^n \mathbb{I}_s + \Delta t A \left(-\frac{1}{\Delta x} \left(\frac{1}{\varepsilon^2} u_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2} u_{i-\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} P_i^{[j]} \right)_{j=1, \dots, s}, \\ \Psi_i^{n+1} &= \Psi_i^n + \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(\frac{1}{\varepsilon^2} u_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2} u_{i-\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} P_i^{[j]} \right)_{j=1, \dots, s} \end{aligned}$$

as well as

$$\begin{aligned} \bar{\rho}_{i+\frac{1}{2}} &= \rho_{i+\frac{1}{2}}^n \mathbb{I}_s + \Delta t A \left(\frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}^{[j]} \right)_{j=1, \dots, s}, \\ \rho_{i+\frac{1}{2}}^{n+1} &= \rho_{i+\frac{1}{2}}^n + \Delta t b \cdot \left(\frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}^{[j]} \right)_{j=1, \dots, s} \end{aligned}$$

and

$$\begin{aligned} \bar{\psi}_{i+\frac{1}{2}} &= \psi_{i+\frac{1}{2}}^n \mathbb{I}_s + \Delta t A \left(-\frac{1}{2\varepsilon^2} \left(D^+ u_{i+\frac{1}{2}}^{[j]} + D^- u_{i+\frac{1}{2}}^{[j]} \right) \right. \\ &\quad \left. - \frac{1}{2\varepsilon} \left(D^+ \psi_{i+\frac{1}{2}}^{[j]} - D^- \psi_{i+\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}^{[j]} \right)_{j=1, \dots, s}, \\ \psi_{i+\frac{1}{2}}^{n+1} &= \psi_{i+\frac{1}{2}}^n + \Delta t b \cdot \left(-\frac{1}{2\varepsilon^2} \left(D^+ u_{i+\frac{1}{2}}^{[j]} + D^- u_{i+\frac{1}{2}}^{[j]} \right) \right. \\ &\quad \left. - \frac{1}{2\varepsilon} \left(D^+ \psi_{i+\frac{1}{2}}^{[j]} - D^- \psi_{i+\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}^{[j]} \right)_{j=1, \dots, s}. \end{aligned}$$

The vector $\mathbb{I}_s \in \mathbb{R}^s$ in the above equations is defined as

$$\mathbb{I}_s = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

and $\cdot$ denotes the scalar product in $\mathbb{R}^s$.

To continue with the formal asymptotic analysis, a certain degree of regularity of the numerical approximation and stage solutions needs to be assumed. We assume that they are bounded and can be expanded into power series in ε , i.e.

$$\begin{aligned} u_{i+\frac{1}{2}}^{[j]} &= u_{i+\frac{1}{2}}^{(0)[j]} + \varepsilon u_{i+\frac{1}{2}}^{(1)[j]} + \varepsilon^2 u_{i+\frac{1}{2}}^{(2)[j]} + \dots, \\ U_i^{[j]} &= U_i^{(0)[j]} + \varepsilon U_i^{(1)[j]} + \varepsilon^2 U_i^{(2)[j]} + \dots, \\ \rho_{i+\frac{1}{2}}^{[j]} &= \rho_{i+\frac{1}{2}}^{(0)[j]} + \varepsilon \rho_{i+\frac{1}{2}}^{(1)[j]} + \varepsilon^2 \rho_{i+\frac{1}{2}}^{(2)[j]} + \dots, \\ P_i^{[j]} &= P_i^{(0)[j]} + \varepsilon P_i^{(1)[j]} + \varepsilon^2 P_i^{(2)[j]} + \dots, \\ \psi_{i+\frac{1}{2}}^{[j]} &= \psi_{i+\frac{1}{2}}^{(0)[j]} + \varepsilon \psi_{i+\frac{1}{2}}^{(1)[j]} + \varepsilon^2 \psi_{i+\frac{1}{2}}^{(2)[j]} + \dots, \\ \Psi_i^{[j]} &= \Psi_i^{(0)[j]} + \varepsilon \Psi_i^{(1)[j]} + \varepsilon^2 \Psi_i^{(2)[j]} + \dots \end{aligned}$$

for $j = 0, \dots, s$, as well as

$$u_{i+\frac{1}{2}}^{n+1} = u_{i+\frac{1}{2}}^{(0)n+1} + \varepsilon u_{i+\frac{1}{2}}^{(1)n+1} + \varepsilon^2 u_{i+\frac{1}{2}}^{(2)n+1} + \dots,$$

$$\begin{aligned}
 U_i^{n+1} &= U_i^{(0)n+1} + \varepsilon U_i^{(1)n+1} + \varepsilon^2 U_i^{(2)n+1} + \dots, \\
 \rho_{i+\frac{1}{2}}^{n+1} &= \rho_{i+\frac{1}{2}}^{(0)n+1} + \varepsilon \rho_{i+\frac{1}{2}}^{(1)n+1} + \varepsilon^2 \rho_{i+\frac{1}{2}}^{(2)n+1} + \dots, \\
 P_i^{n+1} &= P_i^{(0)n+1} + \varepsilon P_i^{(1)n+1} + \varepsilon^2 P_i^{(2)n+1} + \dots, \\
 \psi_{i+\frac{1}{2}}^{n+1} &= \psi_{i+\frac{1}{2}}^{(0)n+1} + \varepsilon \psi_{i+\frac{1}{2}}^{(1)n+1} + \varepsilon^2 \psi_{i+\frac{1}{2}}^{(2)n+1} + \dots, \\
 \Psi_i^{n+1} &= \Psi_i^{(0)n+1} + \varepsilon \Psi_i^{(1)n+1} + \varepsilon^2 \Psi_i^{(2)n+1} + \dots
 \end{aligned}$$

Moreover, we define vector of stage solution of order $k \in \mathbb{N}_0$,

$$\bar{\rho}_{i+\frac{1}{2}}^{(k)} = \begin{pmatrix} \rho^{(k)[1]} \\ \vdots \\ \rho^{(k)[s]} \end{pmatrix}$$

and again, $\bar{\psi}_{i+\frac{1}{2}}^{(k)}$, $\bar{P}_i^{(k)}$, and $\bar{\Psi}_i^{(k)}$ in a similar way. Substituting the assumptions into the time-discrete update equation of the cell average gives

$$\begin{aligned}
 U_i^{(0)[k]} &= U_i^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{(0)[j-1]} \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{(0)[j-1]} \right)^2 \right) \right) \\
 &\quad + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{\Delta x} \left(\psi_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i-\frac{1}{2}}^{(0)[j]} \right) \right) + \dots, \\
 U_i^{(0)n+1} &= U_i^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{(0)[j-1]} \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{(0)[j-1]} \right)^2 \right) \right) \\
 &\quad + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{\Delta x} \left(\psi_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i-\frac{1}{2}}^{(0)[j]} \right) \right) + \dots
 \end{aligned}$$

and

$$\begin{aligned}
 \bar{P}_i^{(0)} &= P_i^{(0)n} \mathbb{I}_s + \frac{1}{\varepsilon^2} \Delta t A \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(0)[j]} + \rho_{i-\frac{1}{2}}^{(0)[j]} \right) - \Psi_i^{(0)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \frac{1}{\varepsilon} \Delta t A \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(1)[j]} + \rho_{i-\frac{1}{2}}^{(1)[j]} \right) - \Psi_i^{(1)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \Delta t A \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(2)[j]} + \rho_{i-\frac{1}{2}}^{(2)[j]} \right) - \Psi_i^{(2)[j]} \right)_{j=1,\dots,s} + \dots, \\
 P_i^{(0)n+1} &= P_i^{(0)n} + \frac{1}{\varepsilon^2} \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(0)[j]} + \rho_{i-\frac{1}{2}}^{(0)[j]} \right) - \Psi_i^{(0)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \frac{1}{\varepsilon} \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(1)[j]} + \rho_{i-\frac{1}{2}}^{(1)[j]} \right) - \Psi_i^{(1)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(2)[j]} + \rho_{i-\frac{1}{2}}^{(2)[j]} \right) - \Psi_i^{(2)[j]} \right)_{j=1,\dots,s} + \dots
 \end{aligned}$$

and

$$\begin{aligned}
 \bar{\Psi}_i^{(0)} &= \Psi_i^n \mathbb{I}_s + \frac{1}{\varepsilon^2} \Delta t A \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) + P_i^{(0)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \frac{1}{\varepsilon} \Delta t A \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(1)[j]} - u_{i-\frac{1}{2}}^{(1)[j]} \right) + P_i^{(1)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \Delta t A \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(2)[j]} - u_{i-\frac{1}{2}}^{(2)[j]} \right) + P_i^{(2)[j]} \right)_{j=1,\dots,s} + \dots,
 \end{aligned}$$

$$\begin{aligned}
 \Psi_i^{(0)n+1} &= \Psi_i^{(0)n} + \frac{1}{\varepsilon^2} \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) + P_i^{(0)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \frac{1}{\varepsilon^2} \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(1)[j]} - u_{i-\frac{1}{2}}^{(1)[j]} \right) + P_i^{(1)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(2)[j]} - u_{i-\frac{1}{2}}^{(2)[j]} \right) + P_i^{(2)[j]} \right)_{j=1,\dots,s} + \dots
 \end{aligned}$$

Inserting the assumption into the update equations of the point values yields

$$\begin{aligned}
 u_{i+\frac{1}{2}}^{(0)[k]} &= u_{i+\frac{1}{2}}^n + \frac{1}{\varepsilon} \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} - D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) \right) \\
 &\quad + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\max \left((u_{i+\frac{1}{2}}^{(0)[j-1]}, 0) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \right. \\
 &\quad \left. \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \right) \\
 &\quad + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} - D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \right) + \dots, \\
 u_{i+\frac{1}{2}}^{(0)n+1} &= u_{i+\frac{1}{2}}^n + \frac{1}{\varepsilon} \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} - D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) \right) \\
 &\quad + \Delta t \sum_{j=1}^s \tilde{b}_j \left(-\max \left((u_{i+\frac{1}{2}}^{(0)[j-1]}, 0) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \right. \\
 &\quad \left. \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \right) \\
 &\quad + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} - D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \right) + \dots
 \end{aligned}$$

and

$$\begin{aligned}
 \bar{\rho}_{i+\frac{1}{2}}^{(0)} &= \rho_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \frac{1}{\varepsilon^2} \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \frac{1}{\varepsilon} \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(1)[j]} - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s} + \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(2)[j]} - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s} + \dots, \\
 \rho_{i+\frac{1}{2}}^{(0)n+1} &= \rho_{i+\frac{1}{2}}^{(0)n} + \frac{1}{\varepsilon^2} \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s} \\
 &\quad + \frac{1}{\varepsilon} \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(1)[j]} - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s} + \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(2)[j]} - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s} + \dots
 \end{aligned}$$

and

$$\begin{aligned}
\overline{\psi}_{i+\frac{1}{2}}^{(0)} &= \psi_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \frac{1}{\varepsilon^2} \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s} \\
&\quad + \frac{1}{\varepsilon} \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) \right. \\
&\quad \left. - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} - D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s} \\
&\quad + \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) \right. \\
&\quad \left. - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(1)[j]} - D^- \psi_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s} + \dots, \\
\psi_{i+\frac{1}{2}}^{(0)n+1} &= \psi_{i+\frac{1}{2}}^{(0)n} + \frac{1}{\varepsilon^2} \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s} \\
&\quad + \frac{1}{\varepsilon} \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) \right. \\
&\quad \left. - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} - D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s} \\
&\quad + \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) \right. \\
&\quad \left. - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(1)[j]} - D^- \psi_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s} + \dots
\end{aligned}$$

Due to our assumption, we obtain, for instance, $U_i^{[j]} \rightarrow U_i^{(0)[j]}$ as $\varepsilon \rightarrow 0$. Hence, in the next step, it is investigated which equations the terms of order 0 satisfy if $\varepsilon \rightarrow 0$. To obtain these equations the terms that appear in the time-discrete update equations corresponding to the cell averages are sorted by the power of ε . For the terms of factor ε^{-2} , we have

$$\begin{aligned}
0 &= \Delta t A \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(0)[j]} + \rho_{i-\frac{1}{2}}^{(0)[j]} \right) - \Psi_i^{(0)[j]} \right)_{j=1,\dots,s}, \\
0 &= \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(0)[j]} + \rho_{i-\frac{1}{2}}^{(0)[j]} \right) - \Psi_i^{(0)[j]} \right)_{j=1,\dots,s}, \\
0 &= \Delta t A \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) + P_i^{(0)[j]} \right)_{j=1,\dots,s}, \\
0 &= \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) + P_i^{(0)[j]} \right)_{j=1,\dots,s},
\end{aligned}$$

for the ones of factor ε^{-1} , we have

$$\begin{aligned}
0 &= \Delta t A \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(1)[j]} + \rho_{i-\frac{1}{2}}^{(1)[j]} \right) - \Psi_i^{(1)[j]} \right)_{j=1,\dots,s}, \\
0 &= \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(1)[j]} + \rho_{i-\frac{1}{2}}^{(1)[j]} \right) - \Psi_i^{(1)[j]} \right)_{j=1,\dots,s}, \\
0 &= \Delta t A \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(1)[j]} - u_{i-\frac{1}{2}}^{(1)[j]} \right) + P_i^{(1)[j]} \right)_{j=1,\dots,s}, \\
0 &= \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(1)[j]} - u_{i-\frac{1}{2}}^{(1)[j]} \right) + P_i^{(1)[j]} \right)_{j=1,\dots,s},
\end{aligned}$$

and for the ones of factor ε^0 , we have

$$\begin{aligned}
 U_i^{(0)[k]} &= U_i^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{(0)[j-1]} \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{(0)[j-1]} \right)^2 \right) \right) \\
 &\quad + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{\Delta x} \left(\psi_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i-\frac{1}{2}}^{(0)[j]} \right) \right), \\
 U_i^{(0)n+1} &= U_i^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{(0)[j-1]} \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{(0)[j-1]} \right)^2 \right) \right) \\
 &\quad + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{\Delta x} \left(\psi_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i-\frac{1}{2}}^{(0)[j]} \right) \right)
 \end{aligned} \tag{3.2}$$

as well as

$$\begin{aligned}
 \overline{P}_i^{(0)} &= P_i^{(0)n} \mathbb{I}_s + \Delta t A \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(2)[j]} + \rho_{i-\frac{1}{2}}^{(2)[j]} \right) - \Psi_i^{(2)[j]} \right)_{j=1,\dots,s}, \\
 P_i^{(0)n+1} &= P_i^{(0)n} + \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{(2)[j]} + \rho_{i-\frac{1}{2}}^{(2)[j]} \right) - \Psi_i^{(2)[j]} \right)_{j=1,\dots,s}
 \end{aligned}$$

and

$$\begin{aligned}
 \overline{\Psi}_i^{(0)} &= \Psi_i^{(0)n} \mathbb{I}_s + \Delta t A \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(2)[j]} - u_{i-\frac{1}{2}}^{(2)[j]} \right) + P_i^{(2)[j]} \right)_{j=1,\dots,s}, \\
 \Psi_i^{(0)n+1} &= \Psi_i^{(0)n} + \Delta t b \cdot \left(-\frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(2)[j]} - u_{i-\frac{1}{2}}^{(2)[j]} \right) + P_i^{(2)[j]} \right)_{j=1,\dots,s}.
 \end{aligned}$$

Since the IMEX Runge-Kutta scheme is assumed to be of type ARS, we can use the invertibility of A to simplify the equations corresponding to ε^{-2} . We obtain that the approximations of order 0 satisfy

$$\Psi_i^{(0)[j]} = \frac{1}{\Delta x} \left(\rho_{i+\frac{1}{2}}^{(0)[j]} - \rho_{i-\frac{1}{2}}^{(0)[j]} \right) \tag{3.3}$$

and

$$P_i^{(0)[j]} = \frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) \tag{3.4}$$

for all $j = 1, \dots, s$. Next, sorting the terms that appear in the update equations of the point values by the power of ε yields

$$\begin{aligned}
 0 &= \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\
 0 &= \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\
 0 &= \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\
 0 &= \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}
 \end{aligned}$$

for the terms of order ε^{-2}

$$0 = \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} - D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) \right)_{j=1,\dots,s},$$

$$\begin{aligned}
0 &= \Delta tb \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} - D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) \right)_{j=1, \dots, s}, \\
0 &= \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(1)[j]} - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1, \dots, s}, \\
0 &= \Delta tb \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(1)[j]} - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1, \dots, s}, \\
0 &= \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} - D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1, \dots, s}, \\
0 &= \Delta tb \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} - D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1, \dots, s}
\end{aligned}$$

for the terms of order ε^{-1} , and

$$\begin{aligned}
u_{i+\frac{1}{2}}^{[k](0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\max \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \\
&\quad \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \\
&\quad + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} - D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(1)[j]} \right) \right), \\
u_{i+\frac{1}{2}}^{n+1(0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^s \tilde{b}_j \left(-\max \left(u_{i+\frac{1}{2}}^{[j-1](0)}, 0 \right) D^+ u_{i+\frac{1}{2}}^{[j-1](0)} \right. \\
&\quad \left. - \min \left(u_{i+\frac{1}{2}}^{[j-1](0)}, 0 \right) D^- u_{i+\frac{1}{2}}^{[j-1](0)} \right) \\
&\quad + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} - D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \right)
\end{aligned} \tag{3.5}$$

as well as

$$\begin{aligned}
\bar{\rho}_{i+\frac{1}{2}}^{(0)} &= \rho_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(2)[j]} - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}, \\
\rho_{i+\frac{1}{2}}^{(0)n+1} &= \rho_{i+\frac{1}{2}}^{(0)n} + \Delta tb \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(2)[j]} - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}
\end{aligned}$$

and

$$\begin{aligned}
\bar{\psi}_{i+\frac{1}{2}}^{(0)} &= \psi_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) \right. \\
&\quad \left. - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(1)[j]} - D^- \psi_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}, \\
\psi_{i+\frac{1}{2}}^{(0)n+1} &= \psi_{i+\frac{1}{2}}^{(0)n} + \Delta tb \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) \right. \\
&\quad \left. - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(1)[j]} - D^- \psi_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}
\end{aligned}$$

for the terms of order ε^0 . We can use the invertibility of A to obtain

$$\psi_{i+\frac{1}{2}}^{(0)[j]} = D^- \rho_{i+\frac{1}{2}}^{(0)[j]} \tag{3.6}$$

and

$$\rho_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) \tag{3.7}$$

for $i = 1, \dots, s$.

Lemma 3.4. Assume $v_{i+\frac{1}{2}}$ and V_i for $i \in \mathbb{Z}$ are point values and cell averages of an Active Flux scheme and $D^+v_{i+\frac{1}{2}}$ and $D^-v_{i+\frac{1}{2}}$ are defined as

$$\begin{aligned} D^+v_{i+\frac{1}{2}} &= \frac{1}{\Delta x} \left(2v_{i-\frac{1}{2}} - 6V_i + 4v_{i+\frac{1}{2}} \right), \\ D^-v_{i+\frac{1}{2}} &= \frac{1}{\Delta x} \left(-4v_{i+\frac{1}{2}} + 6V_{i+1} - 2v_{i+\frac{3}{2}} \right). \end{aligned}$$

This implies

$$\frac{1}{2} \left(D^+v_{i+\frac{1}{2}} + D^-v_{i+\frac{1}{2}} \right) = \frac{1}{\Delta x} \left(v_{i-\frac{1}{2}} - 3V_i + 3V_{i+1} - v_{i+\frac{3}{2}} \right)$$

and

$$\frac{1}{2} \left(D^+v_{i+\frac{1}{2}} - D^-v_{i+\frac{1}{2}} \right) = \frac{1}{\Delta x} \left(v_{i-\frac{1}{2}} - 3V_i + 4v_{i+\frac{1}{2}} - 3V_{i+1} + v_{i+\frac{3}{2}} \right).$$

Proof. The result immediately follows from the definition of $D^+v_{i+\frac{1}{2}}$ and $D^-v_{i+\frac{1}{2}}$. $\square$

Lemma 3.5. The leading order terms, i.e. the terms of order 0, satisfy

$$\psi_{i+\frac{1}{2}}^{(0)[j]} = \frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)[j]} + 6U_i^{(0)[j]} - 4u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + 5u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right)$$

and

$$\begin{aligned} \frac{1}{2} \left(D^+\psi_{i+\frac{1}{2}}^{(0)[j]} + D^-\psi_{i+\frac{1}{2}}^{(0)[j]} \right) &= \frac{1}{\Delta x^3} \left(-u_{i-\frac{3}{2}}^{(0)[j]} + 3U_{i-1}^{(0)[j]} - 14u_{i-\frac{1}{2}}^{(0)[j]} + 21U_i^{(0)[j]} + 14u_{i+\frac{1}{2}}^{(0)[j]} \right. \\ &\quad \left. - 45U_{i+1}^{(0)[j]} + 16u_{i+\frac{3}{2}}^{(0)[j]} + 15U_{i+2}^{(0)[j]} - 13u_{i+\frac{5}{2}}^{(0)[j]} + 6U_{i+3}^{(0)[j]} - 2u_{i+\frac{7}{2}}^{(0)[j]} \right) \end{aligned}$$

for $j = 1, \dots, s$.

Proof. Using the Equations (3.3), (3.4), (3.6), and (3.7) as well as Lemma 3.4, we obtain

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{(0)[j]} &= D^-\rho_{i+\frac{1}{2}}^{(0)[j]} \\ &= \frac{1}{\Delta x} \left(-4\rho_{i+\frac{1}{2}}^{(0)[j]} + 6P_{i+1}^{(0)[j]} - 2\rho_{i+\frac{3}{2}}^{(0)[j]} \right) \\ &= \frac{1}{\Delta x} \left(-2 \left(D^+u_{i+\frac{1}{2}}^{(0)[j]} + D^-u_{i+\frac{1}{2}}^{(0)[j]} \right) + \frac{6}{\Delta x} \left(u_{i+\frac{3}{2}}^{(0)[j]} - u_{i+\frac{1}{2}}^{(0)[j]} \right) \right. \\ &\quad \left. - \left(D^+u_{i+\frac{3}{2}}^{(0)[j]} + D^-u_{i+\frac{3}{2}}^{(0)[j]} \right) \right) \\ &= \frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)[j]} + 6U_i^{(0)[j]} - 4u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + 5u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right). \end{aligned}$$

In a similar way, we can deduce

$$\begin{aligned} \frac{1}{2} \left(D^+\psi_{i+\frac{1}{2}}^{(0)[j]} + D^-\psi_{i+\frac{1}{2}}^{(0)[j]} \right) &= \frac{1}{\Delta x} \left(\psi_{i-\frac{1}{2}}^{(0)[j]} - 3\Psi_i^{(0)[j]} + 3\Psi_{i+1}^{(0)[j]} - \psi_{i+\frac{3}{2}}^{(0)[j]} \right) \\ &= \frac{1}{\Delta x} \left(D^-\rho_{i-\frac{1}{2}}^{(0)[j]} - \frac{3}{\Delta x} \left(\rho_{i+\frac{1}{2}}^{(0)[j]} - \rho_{i-\frac{1}{2}}^{(0)[j]} \right) \right. \\ &\quad \left. + \frac{3}{\Delta x} \left(\rho_{i+\frac{3}{2}}^{(0)[j]} - \rho_{i+\frac{1}{2}}^{(0)[j]} \right) - D^-\rho_{i+\frac{3}{2}}^{(0)[j]} \right) \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{\Delta x^2} \left(-\rho_{i-\frac{1}{2}}^{(0)[j]} + 6P_i^{(0)[j]} - 8\rho_{i+\frac{1}{2}}^{(0)[j]} + 7\rho_{i+\frac{3}{2}}^{(0)[j]} - 6P_{i+2}^{(0)[j]} + 2\rho_{i+\frac{5}{2}}^{(0)[j]} \right) \\
 &= \frac{1}{\Delta x^2} \left(-\frac{1}{2} \left(D^+ u_{i-\frac{1}{2}}^{(0)[j]} + D^- u_{i-\frac{1}{2}}^{(0)[j]} \right) + \frac{6}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) \right. \\
 &\quad - 4 \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \frac{7}{2} \left(D^+ u_{i+\frac{3}{2}}^{(0)[j]} + D^- u_{i+\frac{3}{2}}^{(0)[j]} \right) \\
 &\quad \left. - \frac{6}{\Delta x} \left(u_{i+\frac{5}{2}}^{(0)[j]} - u_{i+\frac{3}{2}}^{(0)[j]} \right) + \left(D^+ u_{i+\frac{5}{2}}^{(0)[j]} + D^- u_{i+\frac{5}{2}}^{(0)[j]} \right) \right) \\
 &= \frac{1}{\Delta x^3} \left(-u_{i-\frac{3}{2}}^{(0)[j]} + 3U_{i-1}^{(0)[j]} - 14u_{i-\frac{1}{2}}^{(0)[j]} + 21U_i^{(0)[j]} + 14u_{i+\frac{1}{2}}^{(0)[j]} \right. \\
 &\quad \left. - 45U_{i+1}^{(0)[j]} + 16u_{i+\frac{3}{2}}^{(0)[j]} + 15U_{i+2}^{(0)[j]} - 13u_{i+\frac{5}{2}}^{(0)[j]} + 6U_{i+3}^{(0)[j]} - 2u_{i+\frac{7}{2}}^{(0)[j]} \right).
 \end{aligned}$$

□

The two lemmata imply that Equation (3.5) is a temporal discretisation using an IMEX Runge-Kutta scheme of

$$\begin{aligned}
 \frac{d}{dt} u_{i+\frac{1}{2}}^{(0)}(t) = & - \left(\max \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t) \right) \right. \\
 & + \min \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t) \right) \Big) \\
 & - \frac{1}{\Delta x^3} \left(-u_{i-\frac{3}{2}}^{(0)}(t) + 3U_{i-1}^{(0)}(t) - 14u_{i-\frac{1}{2}}^{(0)}(t) + 21U_i^{(0)}(t) + 14u_{i+\frac{1}{2}}^{(0)}(t) \right. \\
 & - 45U_{i+1}^{(0)}(t) + 16u_{i+\frac{3}{2}}^{(0)}(t) + 15U_{i+2}^{(0)}(t) - 13u_{i+\frac{5}{2}}^{(0)}(t) + 6U_{i+3}^{(0)}(t) - 2u_{i+\frac{7}{2}}^{(0)}(t) \Big) \\
 & - \frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(1)}(t) - 3U_i^{(1)}(t) + 4u_{i+\frac{1}{2}}^{(1)}(t) - 3U_{i+1}^{(1)}(t) + u_{i+\frac{3}{2}}^{(1)}(t) \right)
 \end{aligned} \tag{3.8}$$

and (3.2) is the temporal discretisation of

$$\begin{aligned}
 \frac{d}{dt} U_i^{(0)}(t) = & - \frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{(0)}(t) \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{(0)}(t) \right)^2 \right) \\
 & - \frac{2}{\Delta x^3} \left(\left(-2u_{i-\frac{1}{2}}^{(0)}(t) + 6U_i^{(0)}(t) - 4u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + 5u_{i+\frac{3}{2}}^{(0)}(t) \right. \right. \\
 & - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \Big) - \left(-2u_{i-\frac{3}{2}}^{(0)}(t) + 6U_{i-1}^{(0)}(t) - 4u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) \right. \\
 & \left. \left. + 5u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + u_{i+\frac{3}{2}}^{(0)}(t) \right) \right)
 \end{aligned} \tag{3.9}$$

with an error in $\mathcal{O}(\Delta t^p)$, where p denotes the order of the IMEX Runge-Kutta scheme.

To continue with the formal asymptotic analysis, we assume that the IMEX Runge-Kutta scheme is globally stiffly accurate. Hence, it follows

$$\begin{aligned}
 u_{i+\frac{1}{2}}^{(0)n+1} &= u_{i+\frac{1}{2}}^{(0)[s]}, & U_i^{(0)n+1} &= U_i^{(0)[s]}, \\
 \rho_{i+\frac{1}{2}}^{(0)n+1} &= \rho_{i+\frac{1}{2}}^{(0)[s]}, & P_i^{(0)n+1} &= P_i^{(0)[s]}, \\
 \psi_{i+\frac{1}{2}}^{(0)n+1} &= \psi_{i+\frac{1}{2}}^{(0)[s]}, & \Psi_i^{(0)n+1} &= \Psi_i^{(0)[s]}.
 \end{aligned}$$

Using Lemma 3.5 and Equation (3.6) and (3.7), gives

$$\begin{aligned}\rho_{i+\frac{1}{2}}^{(0)n+1} &= \rho_{i+\frac{1}{2}}^{(0)[s]} = \frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(0)[s]} - 3U_i^{(0)[s]} + 3U_{i+1}^{(0)[s]} - u_{i+\frac{3}{2}}^{(0)[s]} \right) \\ &= \frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(0)n+1} - 3U_i^{(0)n+1} + 3U_{i+1}^{(0)n+1} - u_{i+\frac{3}{2}}^{(0)n+1} \right)\end{aligned}\quad (3.10)$$

and

$$\begin{aligned}\psi_{i+\frac{1}{2}}^{(0)n+1} &= \psi_{i+\frac{1}{2}}^{(0)[s]} = \frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)[s]} + 6U_i^{(0)[s]} - 4u_{i+\frac{1}{2}}^{(0)[s]} - 3U_{i+1}^{(0)[s]} \right. \\ &\quad \left. + 5u_{i+\frac{3}{2}}^{(0)[s]} - 3U_{i+2}^{(0)[s]} + u_{i+\frac{5}{2}}^{(0)[s]} \right) = \frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)n+1} + 6U_i^{(0)n+1} \right. \\ &\quad \left. - 4u_{i+\frac{1}{2}}^{(0)n+1} - 3U_{i+1}^{(0)n+1} + 5u_{i+\frac{3}{2}}^{(0)n+1} - 3U_{i+2}^{(0)n+1} + u_{i+\frac{5}{2}}^{(0)n+1} \right).\end{aligned}\quad (3.11)$$

Equations (3.3) and (3.4) imply

$$P_i^{(0)n+1} = P_i^{(0)[s]} = \frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[s]} - u_{i-\frac{1}{2}}^{(0)[s]} \right) = \frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)n+1} - u_{i-\frac{1}{2}}^{(0)n+1} \right)\quad (3.12)$$

and

$$\begin{aligned}\Psi_i^{(0)n+1} &= \Psi_i^{(0)[s]} = \frac{1}{\Delta x} \left(\rho_{i+\frac{1}{2}}^{(0)[s]} - \rho_{i-\frac{1}{2}}^{(0)[s]} \right) = \frac{1}{\Delta x} \left(\rho_{i+\frac{1}{2}}^{(0)n+1} - \rho_{i-\frac{1}{2}}^{(0)n+1} \right) \\ &= \frac{1}{\Delta x^2} \left(\left(u_{i-\frac{1}{2}}^{(0)n+1} - 3U_i^{(0)n+1} + 3U_{i+1}^{(0)n+1} - u_{i+\frac{3}{2}}^{(0)n+1} \right) \right. \\ &\quad \left. - \left(u_{i-\frac{3}{2}}^{(0)n+1} - 3U_{i-1}^{(0)n+1} + 3U_i^{(0)n+1} - u_{i+\frac{1}{2}}^{(0)n+1} \right) \right).\end{aligned}\quad (3.13)$$

Thus, it follows that (3.5), (3.2) together with (3.10), (3.11) (3.12), and (3.13), are spatial discretisations using an IMEX Runge-Kutta scheme of type ARS of the system of ordinary equations (3.8), (3.9) and

$$\rho_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) + 3U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t) \right)\quad (3.14)$$

and

$$\begin{aligned}\psi_{i+\frac{1}{2}}^{(0)}(t) &= \frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)}(t) + 6U_i^{(0)}(t) - 4u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) \right. \\ &\quad \left. + 5u_{i+\frac{3}{2}}^{(0)}(t) - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right)\end{aligned}\quad (3.15)$$

and

$$P_i^{(0)}(t) = \frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)}(t) - u_{i-\frac{1}{2}}^{(0)}(t) \right),\quad (3.16)$$

and

$$\begin{aligned}\Psi_i^{(0)}(t) &= \frac{1}{\Delta x^2} \left(\left(u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) + 3U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t) \right) \right. \\ &\quad \left. - \left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) + 3U_i^{(0)}(t) - u_{i+\frac{1}{2}}^{(0)}(t) \right) \right).\end{aligned}\quad (3.17)$$

At this point it is useful to review the steps in the formal asymptotic analysis we have taken up to this point. We started by discretising the HKdV equations with the classical IMEX Active Flux scheme in space and an IMEX Runge-Kutta scheme of type ARS in time. After that, under the assumption that the numerical approximation and stage solution are regular,

we deduced equations that the solution of order 0 satisfy in the limit $\varepsilon \rightarrow 0$. These equations are a temporal discretisation of (3.8) and (3.9). If (3.8) and (3.9) are discretisations of the KdV equation, then we have an indication that the classical IMEX Active Flux scheme discretised in time with an IMEX Runge-Kutta scheme of type ARS is asymptotic preserving for the u component. If, in addition, the scheme is stiffly accurate, the equations for the terms of order 0 are a temporal discretisation of the semi-discretisations (3.8), (3.9), (3.14), (3.15), (3.16), and (3.17). If this semi-discretisation is a discretisation of the limit equation, then the formal asymptotic analysis indicates that the classical IMEX Active Flux scheme discretised in time with an IMEX Runge-Kutta scheme of type ARS is asymptotic preserving. The next theorem, which can be considered the core result of the formal asymptotic analysis, provides the last missing piece.

Theorem 3.6. *Equation (3.8) is a discretisation of*

$$\frac{\partial}{\partial t}u^{(0)}(t, x) + u^{(0)}(t, x)\frac{\partial}{\partial x}u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3}u^{(0)}(t, x) = 0$$

with an error in $\mathcal{O}(\Delta x)$ and (3.9) is a spatial discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t}u^{(0)}(t, x) + u^{(0)}(t, x)\frac{\partial}{\partial x}u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3}u^{(0)}(t, x)dx = 0$$

with an error in $\mathcal{O}(\Delta x)$. In addition, (3.16) is a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \rho^{(0)}(t, x)dx = \frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial x}u^{(0)}(t, x)dx$$

and (3.17) is discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \psi^{(0)}(t, x)dx = \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2}u^{(0)}(t, x)dx$$

with an error in $\mathcal{O}(\Delta x^2)$. The semi-discretisation (3.14) is a discretisation of

$$\rho^{(0)}(t, x) = \frac{\partial}{\partial x}u^{(0)}(t, x)$$

with a error in $\mathcal{O}(\Delta x^2)$ and (3.15) a discretisation of

$$\psi^{(0)}(t, x) = \frac{\partial^2}{\partial x^2}u^{(0)}(t, x)$$

with an error in $\mathcal{O}(\Delta x)$.

Proof. Assume $u^{(0)}, u^{(1)} : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$ are a smooth and bounded functions with bounded derivatives that interpolate the point values and cell averages, i.e.

$$u^{(k)}(t^n, x_{i+\frac{1}{2}}) = u_{i+\frac{1}{2}}^{(k)}(t)$$

and

$$\frac{1}{\Delta x} \int_{C_i} u^{(k)}(t, x)dx = U_i^{(k)}(t).$$

for $k = 0, 1$. It follows that

$$\frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)}(t) + 6U_i^{(0)}(t) - 4u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + 5u_{i+\frac{3}{2}}^{(0)}(t) - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right)$$

$$= \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) + \frac{\Delta x}{2} \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^2)$$

and

$$\begin{aligned} & \frac{1}{\Delta x^3} \left(-u_{i-\frac{3}{2}}^{(0)}(t) + 3U_{i-1}^{(0)}(t) - 14u_{i-\frac{1}{2}}^{(0)}(t) + 21U_i^{(0)}(t) + 14u_{i+\frac{1}{2}}^{(0)}(t) \right. \\ & \quad \left. - 45U_{i+1}^{(0)}(t) + 16u_{i+\frac{3}{2}}^{(0)}(t) + 15U_{i+2}^{(0)}(t) - 13u_{i+\frac{5}{2}}^{(0)}(t) + 6U_{i+3}^{(0)}(t) - 2u_{i+\frac{7}{2}}^{(0)}(t) \right) \\ & = \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) - \Delta x \frac{\partial^4}{\partial x^4} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^2). \end{aligned}$$

In addition, we have

$$\begin{aligned} & \frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(1)}(t) - 3U_i^{(1)}(t) + 4u_{i+\frac{1}{2}}^{(1)}(t) - 3U_{i+1}^{(1)}(t) + u_{i+\frac{3}{2}}^{(1)}(t) \right) \\ & = \frac{\Delta x^3}{30} \frac{\partial^4}{\partial x^4} u^{(1)}(t, x_{i+\frac{1}{2}}) + \frac{\Delta x^5}{630} \frac{\partial^6}{\partial x^6} u^{(1)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^7) \end{aligned}$$

and

$$\begin{aligned} & \left(\max \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) D^+ u_{i+\frac{1}{2}}^{(0)}(t) + \min \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) D^- u_{i+\frac{1}{2}}^{(0)}(t) \right) \\ & = u^{(0)}(t, x_{i+\frac{1}{2}}) \frac{\partial}{\partial x} u^{(0)}(t, x_{i+\frac{1}{2}}) - u^{(0)}(t, x_{i+\frac{1}{2}}) \frac{\Delta x^2}{12} \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^3) \end{aligned}$$

since

$$\frac{1}{\Delta x} \left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t) \right) = \frac{\partial}{\partial x} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\Delta x^2}{12} \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^3)$$

and

$$\frac{1}{\Delta x} \left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t) \right) = \frac{\partial}{\partial x} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\Delta x^2}{12} \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^3).$$

This implies that the time continuous semi-discretisation (3.8) is a discretisation of

$$\frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) = 0$$

with an error in $\mathcal{O}(\Delta x)$. Moreover, we have

$$\frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{(0)}(t) \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{(0)}(t) \right)^2 \right) = \frac{1}{\Delta x} \int_{C_i} u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) dx$$

and

$$\begin{aligned} & \frac{2}{\Delta x^3} \left(\left(-2u_{i-\frac{1}{2}}^{(0)}(t) + 6U_i^{(0)}(t) - 4u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + 5u_{i+\frac{3}{2}}^{(0)}(t) \right. \right. \\ & \quad \left. \left. - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right) - \left(-2u_{i-\frac{3}{2}}^{(0)}(t) + 6U_{i-1}^{(0)}(t) - 4u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) \right. \right. \\ & \quad \left. \left. + 5u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + u_{i+\frac{3}{2}}^{(0)}(t) \right) \right) = \frac{1}{\Delta x} \left(\frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) \\ & \quad + \frac{1}{2} \left(\frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) + \mathcal{O}(\Delta x). \end{aligned}$$

Since the function $u^{(0)}$ is assumed to be smooth with bounded derivatives, its derivatives are Lipschitz continuous. Thus, there exists a constant $L > 0$ such that

$$\left| \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i-\frac{1}{2}}) \right| \leq L\Delta x$$

and

$$\begin{aligned} \frac{1}{\Delta x} \left(\frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) + \frac{1}{2} \left(\frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) \\ = \frac{1}{\Delta x} \left(\frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) + \mathcal{O}(\Delta x) \end{aligned}$$

This implies that (3.9) is a spatial discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) dx = 0$$

with an error in $\mathcal{O}(\Delta x)$. In addition, we have

$$\begin{aligned} \frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) + 3U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t) \right) \\ = \frac{\partial}{\partial x} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\Delta x^2}{12} \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^4) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)}(t) - u_{i-\frac{1}{2}}^{(0)}(t) \right) &= \frac{1}{\Delta x} \left(u^{(0)}(t, x_{i+\frac{1}{2}}) - u^{(0)}(t, x_{i-\frac{1}{2}}) \right) \\ &= \frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial x} u^{(0)}(t, x) dx \end{aligned}$$

as well as

$$\begin{aligned} \frac{1}{\Delta x^2} \left(\left(u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) + 3U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t) \right) - \left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) \right. \right. \\ \left. \left. + 3U_i^{(0)}(t) - u_{i+\frac{1}{2}}^{(0)}(t) \right) \right) = \frac{1}{\Delta x} \left(\frac{\partial}{\partial x} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial}{\partial x} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) \\ + \frac{\Delta x}{12} \left(\frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) + \mathcal{O}(\Delta x^4) \\ = \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) dx + \mathcal{O}(\Delta x^2), \end{aligned}$$

where we again used the Lipschitz continuity of the derivatives of $u^{(0)}$. It immediately follows that Equation (3.16) and (3.17) are discretisations of

$$\begin{aligned} \frac{1}{\Delta x} \int_{C_i} \rho^{(0)}(t, x) dx &= \frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial x} u^{(0)}(t, x) dx, \\ \frac{1}{\Delta x} \int_{C_i} \psi^{(0)}(t, x) dx &= \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) dx \end{aligned}$$

with an maximum error in $\mathcal{O}(\Delta x^2)$ and (3.14) and (3.15) are discretisations of

$$\begin{aligned} \rho^{(0)}(t, x) &= \frac{\partial}{\partial x} u^{(0)}(t, x), \\ \psi^{(0)}(t, x) &= \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) \end{aligned}$$

with an error in $\mathcal{O}(\Delta x^2)$ and $\mathcal{O}(\Delta x)$, respectively. □

To sum up the findings: The formal asymptotic analysis indicates that the classical IMEX Active Flux scheme for the HKdV equations discretised in time with an IMEX Runge-Kutta scheme of type ARS is without modification asymptotic preserving for the component u . If, in addition, the IMEX Runge-Kutta scheme is globally stiffly accurate, then it indicates that the scheme is asymptotic preserving.

Remark 3.7. The stage solution of order 0 also satisfy

$$0 = \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} - D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) \right)_{j=1,\dots,s}$$

or equivalently

$$0 = \frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} - D^- u_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}.$$

If the scheme is additionally stiffly accurate, we obtain

$$0 = \frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)n+1} - D^- u_{i+\frac{1}{2}}^{(0)n+1} \right),$$

which is, again under the assumption that a function $u^{(0)}$ with the same properties as before exists, a discretisation of

$$\frac{\Delta x^3}{30} \frac{\partial^4}{\partial x^4} u^{(0)}(t, x) = 0$$

with an error in $\mathcal{O}(\Delta x^5)$. Since this condition does not correspond to any part of the limit equations, we refer to it as an unnatural condition. The formal asymptotic analysis gives no insight, how and if the unnatural condition affects the asymptotic preserving property of classical IMEX Active Flux scheme in the limit $\varepsilon \rightarrow 0$.

3.4 Implementation

The formal asymptotic analysis only gives an indication that the scheme is asymptotic preserving. To verify this indication, the scheme is implemented and several numerical experiments are performed. Before the results of the experiments are stated, an instruction how to implement the classical IMEX Active Flux scheme discretised in time with an IMEX Runge-Kutta scheme of type ARS is given.

Until now, we considered the initial value problem for all $x \in \mathbb{R}$ and discretised $\mathbb{R}$ with cells $\mathcal{C}_i$, $i \in \mathbb{Z}$. However, numerically it is only possible to consider a finite number of cells. Therefore, the initial value problem is restricted to the interval $[x_{\min}, x_{\max}]$, which is assumed to be divided into $N \in \mathbb{N}$ cells $\mathcal{C}_i$, $i = 1, \dots, N$ of equal length Δx . Since only a finite number of cells is considered, numerical boundary conditions are imposed. Since the initial value problems that are numerically solved in this thesis admit solutions that tend to 0 as $|x| \rightarrow \infty$ for a fixed time t , periodic boundary conditions are a natural and simple choice.

To begin with the implementation, we define the degrees of freedom that belong to cell $\mathcal{C}_i$ to be $q_{i-\frac{1}{2}}(t)$ and $Q_i(t)$ and introduce $v_i : \mathbb{R}_0^+ \rightarrow \mathbb{R}^6$,

$$v_i(t) = \begin{pmatrix} u_{i-\frac{1}{2}}(t) \\ \rho_{i-\frac{1}{2}}(t) \\ \psi_{i-\frac{1}{2}}(t) \\ U_i(t) \\ P_i(t) \\ \Psi_i(t) \end{pmatrix}$$

for $i = 1, \dots, N$. In addition, we define $v : \mathbb{R}_0^+ \rightarrow \mathbb{R}^{6N}$,

$$v(t) = \begin{pmatrix} v_1(t) \\ \vdots \\ v_i(t) \\ \vdots \\ v_N(t) \end{pmatrix}.$$

To implement the scheme, we rewrite the finitely many update equations of the classical IMEX Active Flux scheme in the form

$$\frac{d}{dt}v(t) = f_1(v(t)) + Mv(t),$$

where $f_1 : \mathbb{R}^{6N} \rightarrow \mathbb{R}^{6N}$ and $M \in \mathbb{R}^{6N \times 6N}$. First, we define the matrix M . Therefore, it is convenient to introduce the sub-matrices

$$M_1 = \begin{pmatrix} -\frac{1}{\Delta x \varepsilon} & 0 & -\frac{1}{\Delta x} & \frac{3}{\Delta x \varepsilon} & 0 & \frac{3}{\Delta x} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & -\frac{1}{\Delta x \varepsilon} & \frac{3}{\Delta x \varepsilon^2} & 0 & \frac{3}{\Delta x \varepsilon} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} -\frac{4}{\Delta x \varepsilon} & 0 & 0 & \frac{3}{\Delta x \varepsilon} & 0 & -\frac{3}{\Delta x} \\ 0 & -\frac{4}{\Delta x \varepsilon^2} & -\frac{1}{\varepsilon^2} & 0 & \frac{6}{\Delta x \varepsilon^2} & 0 \\ 0 & \frac{1}{\varepsilon^2} & -\frac{4}{\Delta x \varepsilon} & -\frac{3}{\Delta x \varepsilon^2} & 0 & \frac{3}{\Delta x \varepsilon} \\ 0 & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & -\frac{1}{\varepsilon^2} \\ \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & \frac{1}{\varepsilon^2} & 0 \end{pmatrix},$$

and

$$M_3 = \begin{pmatrix} -\frac{1}{\Delta x \varepsilon} & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{2}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ \frac{1}{\Delta x \varepsilon^2} & 0 & -\frac{1}{\Delta x \varepsilon} & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The matrix M also depends on the chosen boundary conditions and by imposing periodic boundary conditions M reads

$$M = \begin{pmatrix} A_2 & A_3 & 0_{6 \times 6} & 0_{6 \times 6} & 0_{6 \times 6} & \cdots & 0_{6 \times 6} & 0_{6 \times 6} & A_1 \\ A_1 & A_2 & A_3 & 0_{6 \times 6} & 0_{6 \times 6} & \cdots & 0_{6 \times 6} & 0_{6 \times 6} & 0_{6 \times 6} \\ 0_{6 \times 6} & A_1 & A_2 & A_3 & 0_{6 \times 6} & \cdots & 0_{6 \times 6} & 0_{6 \times 6} & 0_{6 \times 6} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ A_3 & 0_{6 \times 6} & 0_{6 \times 6} & 0_{6 \times 6} & 0_{6 \times 6} & \cdots & 0_{6 \times 6} & A_1 & A_2 \end{pmatrix},$$

where $0_{6 \times 6} \in \mathbb{R}^{6 \times 6}$ with only zero entries. Furthermore, define

$$f_{1,1,1}(v(t)) = -\max(v_{1,1}(t), 0) \frac{1}{\Delta x} (2v_{N,1}(t) - 6v_{N,4}(t) + 4v_{1,1}(t))$$

$$- \min(v_{1,1}(t), 0) \frac{1}{\Delta x} (-4v_{1,1}(t) + 6v_{1,4}(t) - 2v_{2,1}(t)),$$

and

$$\begin{aligned} f_{1,i,1}(v(t)) &= -\max(v_{i,1}(t), 0) \frac{1}{\Delta x} (2v_{i-1,1}(t) - 6v_{i-1,4}(t) + 4v_{i,1}(t)) \\ &\quad - \min(v_{i,1}(t), 0) \frac{1}{\Delta x} (-4v_{i,1}(t) + 6v_{i,4}(t) - 2v_{i+1,1}(t)) \end{aligned}$$

and

$$\begin{aligned} f_{1,N,1}(v(t)) &= -\max(v_{N,1}(t), 0) \frac{1}{\Delta x} (2v_{N-1,1}(t) - 6v_{N-1,4}(t) + 4v_{N,1}(t)) \\ &\quad - \min(v_{N,1}(t), 0) \frac{1}{\Delta x} (-4v_{N,1}(t) + 6v_{N,4}(t) - 2v_{1,1}(t)), \end{aligned}$$

where, for example, $v_{N,1}(t)$ denotes for the first element of $v_N(t)$. Furthermore,

$$f_{1,i,4}(v(t)) = -\frac{1}{\Delta x} \left(\frac{1}{2}(v_{i+1,1}(t))^2 - \frac{1}{2}(v_{i,1}(t))^2 \right)$$

and

$$f_{1,N,4}(v(t)) = -\frac{1}{\Delta x} \left(\frac{1}{2}(v_{1,1}(t))^2 - \frac{1}{2}(v_{N,1}(t))^2 \right).$$

The function f_1 is given by

$$f_1(v(t)) = \begin{pmatrix} f_{1,1}(v(t)) \\ \vdots \\ f_{1,i}(v(t)) \\ \vdots \\ f_{1,N}(v(t)) \end{pmatrix}$$

with

$$f_{1,i}(v(t)) = \begin{pmatrix} f_{1,i,1}(v(t)) \\ 0 \\ 0 \\ f_{1,i,4}(v(t)) \\ 0 \\ 0 \end{pmatrix}.$$

The time-discretisation is implemented as describes in [49]. Given the coefficients of the implicit and explicit scheme, $A, \tilde{A}, b$ and $\tilde{b}$ and assume v^n , an approximation to $v(t^n)$ at time step $t^n \geq 0$, $n \in \mathbb{N}_0$, is known. The time step size is computed by

$$\Delta t = C_{\text{CFL}} \frac{\Delta x}{\lambda_{\max}}$$

with

$$\lambda_{\max} = \max_{q \in \Gamma} |u|,$$

where Γ is the set of all point values and $C_{\text{CFL}} > 0$. Due to the implicit treatment of the terms depending on ε , $\lambda_{\max}$ and Δt are independent of ε . The algorithm to compute the approximation v^{n+1} at the next time step $t^{n+1} = t^n + \Delta t$ is shown in Algorithm 1. To compute each stage solution, a linear system must be solved. In the experiments presented in this thesis, this system is solved exactly by computing the LU factorisation of the matrix.

Remark 3.8. The matrix of the linear system of equations in Algorithm 1 only depends on $a_{i,i}$, because Δt and M are independent of $i = 1, \dots, s$. If this value is the same for all $i = 1, \dots, s$, one can precompute the factorisation and thereby increase the performance of the algorithm.

Algorithm 1 IMEX Runge-Kutta method**Require:** $\Delta t, v^n, A, \tilde{A}, b, \tilde{b}, f, M$

-
- 1: $\tilde{K}_1 := f_1(v^n)$
 - 2: **for** $k = 1, \dots, s$ **do**
 - 3: $K_i := (\mathbb{I} - \Delta t a_{k,k} M) \setminus \left(M v^n + \Delta t \sum_{j=1}^{k-1} a_{k,j} M K_j + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} M \tilde{K}_j \right)$
 - 4: $v^{[k]} := v^n + \Delta t \sum_{j=1}^k a_{k,j} K_j + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \tilde{K}_j$
 - 5: $K_{k+1} = f_1(v^{[k]})$
 - 6: **end for**
 - 7: $v^{n+1} := v^n + \Delta t \sum_{j=1}^s b_j K_j + \Delta t \sum_{j=1}^s \tilde{b}_j \tilde{K}_j$
-

3.5 Numerical experiments

The classical IMEX Active Flux scheme can be combined with different IMEX Runge-Kutta scheme and, later in this thesis, modified IMEX Active Flux schemes are considered. Thus, it is convenient to define the numerical experiments in detail once. The experiments can be divided in two types: the ones in which an initial value problem for the HKdV equations is solved numerically and the ones in which an initial value problem for the KdV equations is solved numerically.

Remark 3.9. In general, only Experiment 1 and 2 are performed for a given IMEX Active Flux scheme discretised in time with the ARS(4,4,3) scheme since the results of these two experiments already give valuable insight into the properties of the scheme. In addition, the ARS(4,4,3) scheme is globally stiffly accurate, and this property is needed to obtain the theoretical indication that a given IMEX Active Flux scheme is asymptotic preserving. However, there is no indication for the case of a not globally stiffly accurate scheme, like the ARS(2,3,3) scheme. Thus, for some IMEX Active Flux schemes, including the classical one, Experiment 1 and 2 are carried out using this time discretisation. If the results of the first two experiments suggest that the scheme possesses promising properties, Experiment 3 is performed to obtain a stronger verification of those properties.

3.5.1 Experiment 1

This experiment investigates the ability of the given scheme to solve an initial value problem for the HKdV equations, i.e. a value of ε close to 1. For this experiment, $\varepsilon = 0.8$ and $c = \frac{1}{2}$ are chosen as well as the numerically computed travelling wave solution u_{num} , ρ_{num} and ψ_{num} at $t = 0$ as initial data. The exact solution to this initial value problem is the numerically computed travelling wave solution of the HKdV equations. The numerical solution is computed on the interval

$$[x_{\min}, x_{\max}] = [-45.0, 45.0]$$

and periodic boundary conditions are imposed. Since the initial data is not a smooth periodic function, errors arise at the boundary as a consequence. The length of the interval chosen here is sufficiently large to ensure that these errors do not affect the error of the numerical solution for the spatial step sizes considered in the experiments. Using this setup, different properties of the scheme are investigated. It is convenient to define the different parts of the experiment:

- **Experiment 1.1:** The aim of this experiment is to investigate the accuracy of the scheme on a coarse grid and the behaviour of the l_2 -error of the point values and cell averages

with respect to the spatial step size Δx . Assume $v_{i+\frac{1}{2}}(t)$ are point values and $V_i(t)$ cell averages of the given scheme at time $t \geq 0$. The discrete l_2 -error of the point values is

$$\mathcal{E}_{l_2}(v, \Delta x) = \sqrt{\Delta x \sum_{i=1}^N |v_{i+\frac{1}{2}}(t) - \tilde{v}(t, x_{i+\frac{1}{2}})|^2}$$

and the discrete l_2 -error of the cell averages is

$$\mathcal{E}_{l_2}(V, \Delta x) = \sqrt{\Delta x \sum_{i=1}^N \left| V_i(t) - \frac{1}{\Delta x} \int_{\mathcal{C}_i} \tilde{v}(t, x) dx \right|^2},$$

where N denotes the number of cells and $\tilde{v}$ the exact solution of the initial value problem. In this experiment the numerical solution is computed at $t = 4$ on grids with 64, 128, 256, 512, 1024, and 2048 cells and the l_2 -error for all the point values and cell averages measured. In addition, to analyse the accuracy of the numerical solution computed on the coarse grid with 64 cells, the maximum l_2 -error is computed, i.e.

$$\mathcal{E}_{l_2, \max}(\Delta x) = \max \left(\max_{v \in \Gamma_1} \mathcal{E}_{l_2}(v, \Delta x), \max_{V \in \Gamma_2} \mathcal{E}_{l_2}(V, \Delta x) \right),$$

where in case of the HKdV equations $\Gamma_1 = \{u, \rho, \psi\}$ and $\Gamma_2 = \{U, P, \Psi\}$.

- **Experiment 1.2:** This experiment investigates the stability and long-time behaviour of the numerical solution. The numerical solution on a grid with 512 cells at the time at which the exact travelling wave solution travelled one time over the whole interval, i.e. $t = 180$ for $c = \frac{1}{2}$, is computed and the maximum l_2 -error of the numerical solution is measured. In addition, the maximum l_2 -error of the numerical solution at $t = 4$ on a grid with 512 for different values of $C_{\text{CFL}} \in (0, 1]$ is determined. This gives us a numerical estimate of the maximum value C_{CFL} for which the scheme is stable.
- **Experiment 1.3:** With this experiment, the conservation of the mass and energy is investigated. Since the scheme is conservative and periodic boundary conditions are imposed, the mass is conserved in the cell averages. Thus, to compare different schemes, the conservation of mass is measured in the point values. The discrete mass is defined as

$$M_{\text{dis}}(t) = \Delta x \sum_{i=1}^N u_{i+\frac{1}{2}}(t).$$

It is not clear if a given IMEX Active Flux schemes conserve a discrete energy and how this energy is defined. Since the main reason to measure a discrete energy in this thesis is to compare different scheme, the discrete energy is defined to be

$$E_{\text{H,dis}}(t) = \Delta x \frac{1}{2} \sum_{i=1}^N \left((u_{i+\frac{1}{2}}(t))^2 + \varepsilon^2 (\rho_{i+\frac{1}{2}}(t))^2 + \varepsilon^2 (\psi_{i+\frac{1}{2}}(t))^2 \right),$$

which can be considered a discrete counterpart of the energy of HKdV equation. To measure the conservation of discrete mass and energy, the numerical solution is computed on grids with 256, 512, and 1024 cells at different $t \in [0, 360]$. In addition, the percentage of the discrete mass and energy at $t = 360$ relative to $t = 0$ is determined.

3.5.2 Experiment 2

The overall aim of this experiment is to investigate the ability of a chosen scheme to compute numerical approximations to the limit equations of the HKdV equations. The interval

$$[x_{\min}, x_{\max}] = [-45.0, 45.0]$$

is chosen and periodic boundary conditions imposed. The functions

$$\begin{aligned} u_0(x) &= 3c \operatorname{sech}^2 \left(\frac{\sqrt{c}}{2} x \right), \\ \rho_0(x) &= -3c^{\frac{3}{2}} \operatorname{sech}^2 \left(\frac{\sqrt{c}}{2} x \right) \tanh \left(\frac{\sqrt{c}}{2} x \right), \\ \psi_0(x) &= 3c^2 \tanh^2 \left(\frac{\sqrt{c}}{2} x \right) \operatorname{sech}^2 \left(\frac{\sqrt{c}}{2} x \right) - \frac{3}{2} c^2 \operatorname{sech}^4 \left(\frac{\sqrt{c}}{2} x \right) \end{aligned}$$

with $c = \frac{1}{2}$ serve as initial data. As reference soliton, the exact solution of the initial value problem for the limit equations of the HKdV equations is used, which is the soliton solution of the KdV equations and its first two derivatives. This means that the errors are computed with respect to this reference solution and not with respect to the solution of the HKdV equations for the given initial value problem. This decision is justified since the goal of this experiment is to determine if the scheme can solve an initial value problems for the limit equations. The different parts of the experiment are:

- **Experiment 2.1:** With this experiment, the asymptotic preserving property of the given scheme is analysed numerically. Therefore, the l_2 -error of the numerical solution at $t = 4$ with respect to ε and for different spatial step sizes is computed. In addition, the l_2 -error of the numerical solution at $t = 4$ with respect to the spatial step size for a sufficiently small value of ε , is investigated. If, for a fixed spatial step size, the error becomes constant as $\varepsilon \rightarrow 0$ and for a sufficiently small value of ε the numerical solution seems to converge to the exact solution of the limit equations, then this is an experimental indication that the scheme is asymptotic preserving. Sufficiently small means in this context that the l_2 -error as a function of ε became constant before reaching this value of ε . In the experiment of this thesis, $\varepsilon = 10^{-7}$ is found to be sufficiently small. (Note that ε^2 is used in the partial differential equations and, thus, this translates to a factor of 10^{-14}). This experiment also indicates if the numerical scheme is asymptotic accurate since for an asymptotic accurate scheme, the achieved order of accuracy should not decrease for this value of ε . Furthermore, the accuracy of the numerical solution at $t = 4$ and with $\varepsilon = 10^{-7}$ computed on a coarse grid with 64 cells is analysed. Therefore, the maximum l_2 -error of the numerical solution is determined.
- **Experiment 2.2:** Similar to Experiment 1.2, this experiment investigates the stability and long-time behaviour of the numerical solution for $\varepsilon = 10^{-7}$. The l_2 -error of numerical solutions at $t = 4$ computed on a grid with 512 cells for different $C_{\text{CFL}} \in (0, 1]$ is determined and the accuracy of the numerical solution at $t = 180$ computed on a grid with 512 cells is investigated by evaluating the maximum l_2 -error.
- **Experiment 2.3:** The aim of this experiment is to analyse the conservation mass and energy. Again, $\varepsilon = 10^{-7}$ is chosen and the numerical solution is computed on a grid with 256, 512, and 1024 cells at different times $t \in [0, 360]$. A discrete mass, defined as in Experiment 1.3 and a discrete energy

$$E_{\text{dis}}(t) = \Delta x \sum_{i=1}^N \frac{1}{2} \left(u_{i+\frac{1}{2}}(t) \right)^2,$$

which corresponds to the energy that is conserved by the KdV equations, are measured at the different times. In addition, the percentage of the remaining discrete energy and mass at $t = 360$ relative to that at $t = 0$ is calculated.

3.5.3 Experiment 3

This experiment serves as an additional verification of the properties of the scheme to compute accurate solution to the limit equations of the HKdV equations and, thus, $\varepsilon = 10^{-7}$ is set for all parts of the experiment. As before,

$$[x_{\min}, x_{\max}] = [-45.0, 45.0]$$

is chosen and periodic boundary conditions are imposed. As initial data,

$$\begin{aligned} u_0(x) &= 2(\delta_2^2 - \delta_1^2) \frac{\delta_2^2 \operatorname{cosech}(\gamma_2(0, x))^2 + \delta_1^2 \operatorname{sech}(\gamma_1(0, x))^2}{(\delta_2 \coth(\gamma_2(0, x)) + \delta_1 \tanh(\gamma_1(0, x)))^2}, \\ \rho_0(x) &= Du_0(x), \\ \psi_o(x) &= D\rho_0(x) \end{aligned}$$

is used, where γ_1 and γ_2 are defined as in Theorem 2.10 and D is a fourth order central finite difference operator [43] with spatial step size $h = 0.0005$. In addition, $\delta_1 = 0.3$ and $\delta_2 = 0.45$ is selected. The analytical solution of this initial values problem for the limit equations is the function given in Theorem 2.10 and its first and second derivative. This solution acts as the reference solution in this experiment, where the first and second derivative are computed with the same fourth order central finite difference formulas. As before, the experiment can be split into different parts:

- **Experiment 3.1:** This experiment is similar to Experiment 1.1 and Experiment 2.1, as the aim of it is to investigate the accuracy of a numerical solution on a coarse grid and the behaviour of the l_2 -error of the point values and cell averages as $\Delta x \rightarrow 0$. Therefore, the numerical solution at $t = 4$ is computed on grid with 64, 128, 256, 512, 1024, and 2048 cells and the l_2 -error of the point values and cell averages measured. Additionally, the maximum l_2 -error of the numerical solution computed on the coarse grid with 64 cells is determined.
- **Experiment 3.2:** This experiment is performed to analyse the accuracy of the numerical solution at different times t before, during, and after the interaction of the two solitons. Thus, the maximum l_2 -error of numerical solution at $t = 0, 20, 40$, and 60, which is computed on a grid with 512 cells, is determined.

3.6 Numerical results

Frist, Experiment 1 is conducted using the explicit semi-discrete Active Flux scheme paired with the SSPRK3 scheme to discretise in time. If not stated otherwise, for all parts of the experiment, $C_{\text{CFL}} = 0.4$ is chosen. The result of Experiment 1.1 can be seen in Figure 9, where the l_2 -error of numerical solution computed at $t = 4$ with respect to the spatial step size and the numerical solution on a coarse grid with 64 cells in comparison to the reference solution at $t = 4$ is shown. The numerical solution obtained by using a grid with 64 cells approximates the exact solution, which is in this case the numerically computed travelling wave solution of HKdV equation, well with a maximum l_2 -error of approximately 0.0247. The coloured dots in the figure represent the point values and the coloured lines the reconstruction. As seen in the right part of Figure 9, the scheme achieves approximately third-order accuracy. The results of Experiment 1.2 can be seen in Figure 10. The numerical solution computed on a grid with 512 cells at $t = 180$ matches

the exact solution at this time. The maximum l_2 -error of the numerical solution is roughly 0.0118. Moreover, on a grid with the same number of cells, the numerical scheme appears to be stable up to approximately $C_{\text{CFL}} = 0.5$, which is depicted on the right of the figure. The conservation of the discrete energy and mass (Experiment 1.3) is depicted in Figure 11. The discrete mass at the last measured time ($t = 360$) is approximately 100% of the discrete mass at $t = 0$ for all spatial step sizes considered in the experiment. For $t = 0$ the discrete energy of the numerical solution computed on a grid with 1024 cells is approximately $E_{\text{H,dis}}(0) = 5.5917$ and for $t = 360$ we have $E_{\text{H,dis}}(360) = 5.5908$ which is approximately 99.98% of the original energy. In case of 512 cells, the discrete energy at $t = 360$ is approximately 99.87% of the discrete energy at $t = 0$ and in case of 256 cells it is approximately 98.99% of the discrete energy at $t = 0$.

Next, the classical IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation is considered. The results of Experiment 1 can be seen Figure 12, Figure 13, and Figure 14, where $C_{\text{CFL}} = 0.4$ is chosen. The numerical solution on a coarse grid has a maximum l_2 -error of roughly 0.0267 and the numerical scheme achieves approximately third order accuracy. The numerical solution at $t = 180$ computed on a grid with 512 cells visually coincides with the exact solution at this time with a maximum l_2 -error of approximately 0.0117. In contrast to the explicit scheme, the classical IMEX Active Flux scheme combined with the ARS(4, 4, 3) schemes seems to be stable for all $C_{\text{CFL}} \in (0, 1]$ considered in the experiment. The discrete mass of the numerical solution at $t = 360$ is approximately 100% of the discrete mass at $t = 0$ for all spatial step sizes considered in the experiment and the discrete energy at $t = 360$ is approximately 98.99% (256 cells), 99.87% (512 cells), and 99.98% (1024 cells) of the discrete energy at $t = 0$. To investigate the asymptotic preserving properties of the scheme, Experiment 2 with $C_{\text{CFL}} = 0.4$ is performed. The results of Experiment 2.1 can be seen in Figure 15 and Figure 16. One can observe that the l_2 -error of all point values and cell averages becomes constant as $\varepsilon \rightarrow 0$. However, it still decreases for a fixed values of ε as $\Delta x \rightarrow 0$. For $\varepsilon = 10^{-7}$, the scheme achieves roughly first order accuracy, as depicted on the right of Figure 16. The numerical solution on a coarse grid is not as accurate as the one obtained in Experiment 1.1 and admits a maximum l_2 -error of roughly 0.4581. The results of Experiment 2.2 (see Figure 17) show that the solution at $t = 180$ on a grid with 512 does not match the exact solution at that time. The maximum l_2 -error of the numerical solution is approximately 3.5441. This indicates that the classical IMEX Active Flux scheme is highly dissipative and dispersive for $\varepsilon = 10^{-7}$. Moreover, the scheme seems to be stable for all $C_{\text{CFL}} \in (0, 1]$ studied in the experiment. The results of Experiment 2.3 are shown in Figure 18. For all different number of cells, the discrete mass at $t = 360$ is approximately 100% of the discrete mass at $t = 0$. In contrast, the discrete energy is not conserved satisfactory since at $t = 360$ it is approximately 28.06% (256 cells), 38.84% (512 cells), and 52.25% (1024 cells) of the discrete energy at $t = 0$.

The result of the Experiment 1 if the ARS(2, 3, 3) scheme, which is not globally stiffly accurate, is used instead to discretise in time, can be seen in Figure 19 - Figure 21. For Experiment 1 as well as Experiment 2, $C_{\text{CFL}} = 0.4$ is selected. As one can observe, similar results as for the classical IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme are obtained: the numerical scheme achieves approximately third order accuracy and the numerical solution computed on a coarse grid at $t = 4$ is already reasonably accurate with a maximum l_2 -error of roughly 0.0267. One difference is that the classical IMEX Active Flux scheme with the ARS(2, 3, 3) scheme as time discretisation appears only to be stable up to approximately $C_{\text{CFL}} = 0.625$. The numerical solution at $t = 180$ computed on a grid with 512 cells visually matches the exact solution and admits a maximum l_2 -error of approximately 0.0118. As before, the results of Experiment 1.3 (Figure 21) show that the discrete mass is almost perfectly conserved since at the last measured time it is roughly 100% of the original discrete mass at $t = 0$. The discrete energy of the solution at $t = 360$ is approximately 98.98% (256 cells), 99.87% (512 cells), and

99.98% (1024 cells) of the discrete energy at $t = 0$.

Similar to the case of using the ARS(4, 4, 3) scheme as time discretisation, the results of Experiment 2.1 (see Figure 22 and Figure 23) indicate that the error of the numerical solution becomes constant on fixed grid if $\varepsilon \rightarrow 0$, and for $\varepsilon = 10^{-7}$ the schemes achieves roughly first order accuracy. The accuracy of the numerical solution at $t = 4$ on a coarse grid is not satisfactory with a maximum l_2 -error of approximately 0.4575. The results of Experiment 2.2 (see Figure 24) indicate that using a different time discretisation does not affect the dissipative or dispersive behaviour of the scheme since the numerical solution does not match the analytical solution at $t = 180$. The maximum l_2 -error is roughly 3.5441. However, using the ARS(2, 3, 3) scheme to discretise in time, the numerical scheme appears to be only stable up to $C_{\text{CFL}} = 0.475$. The discrete mass at $t = 360$ is approximately 100% of the original discrete mass and the discrete energy at this time is approximately 28.06% (256 cells), 38.84% (512 cells), and 52.25% (1024 cells) of the discrete energy at $t = 0$ (see Figure 25).

To conclude, the numerical experiments indicate that the classical IMEX Active Flux scheme with the ARS(4, 4, 3) scheme or the ARS(2, 3, 3) scheme as time discretisation is asymptotic preserving. Concerning accuracy of the numerical solutions, there is no difference between the globally stiffly accurate time discretisation and the other one. However, there are a few shortcomings of noticeable in the numerical experiments. First, the classical IMEX Active Flux scheme, independent of the chosen time discretisation, is not asymptotic accurate and only achieves first order accuracy for sufficiently small ε . In addition, the discrete energy is not conserved satisfactory in Experiment 2.3.

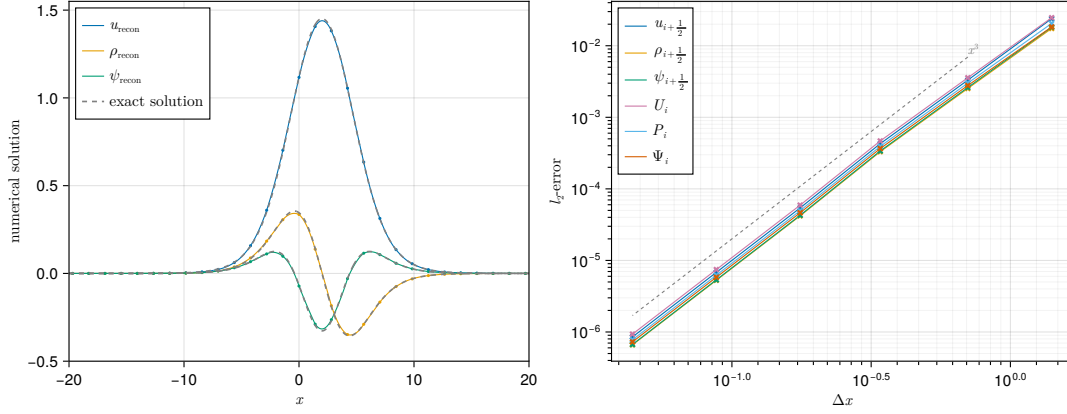


Figure 9: The results of Experiment 1.1 for the explicit Active Flux scheme with the SSPRK3 scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

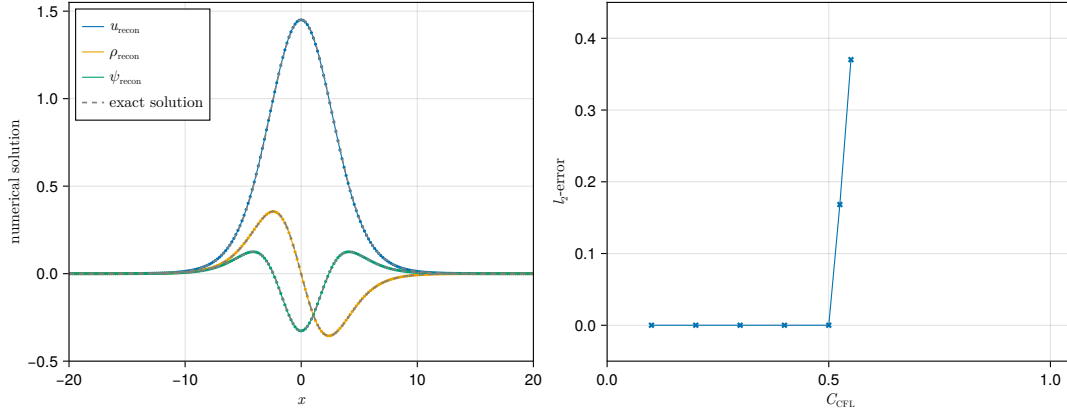


Figure 10: The results of Experiment 1.2 for the explicit Active Flux scheme with the SSPRK3 scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

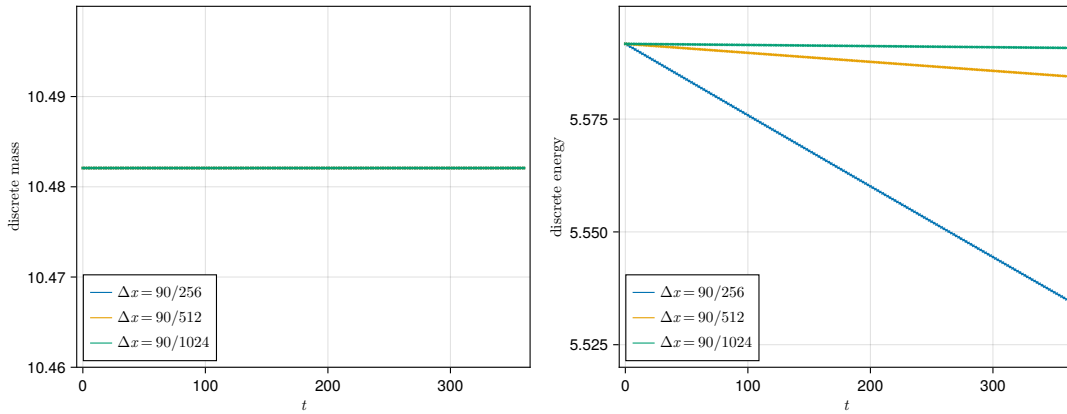


Figure 11: The results of Experiment 1.3 for the explicit Active Flux scheme with the SSPRK3 scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy $E_{\text{H,dis}}$ with respect to time t for different spatial step sizes.

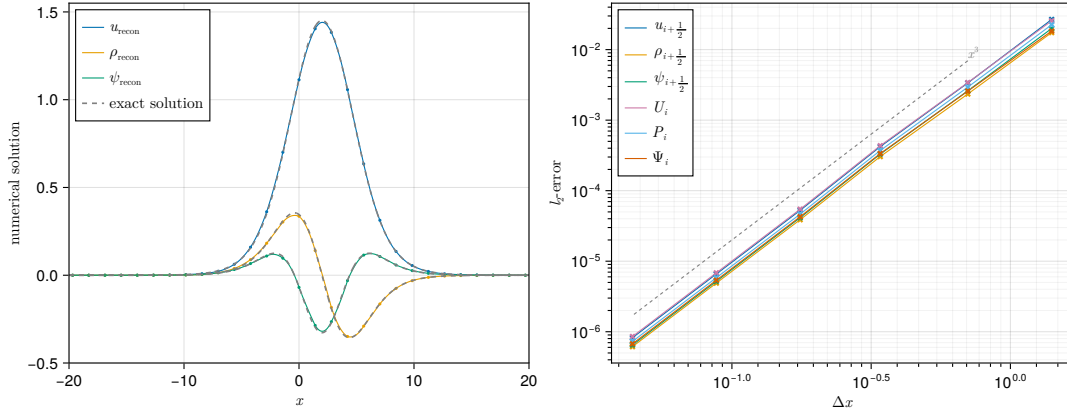


Figure 12: The results of Experiment 1.1 for the classical IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

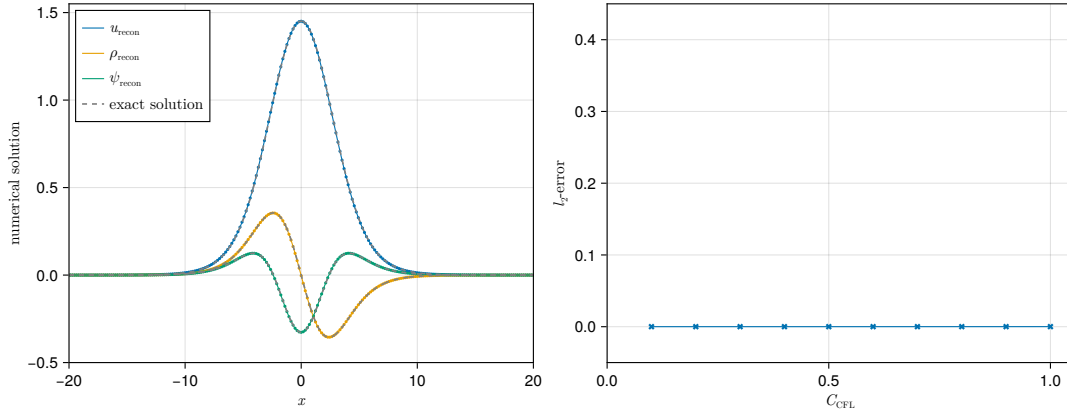


Figure 13: The results of Experiment 1.2 for the classical IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

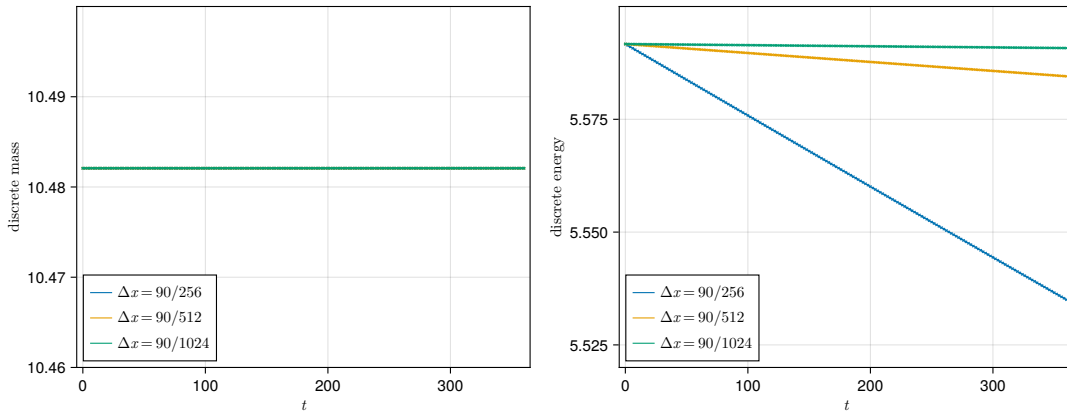


Figure 14: The results of Experiment 1.3 for the classical IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy $E_{\text{H,dis}}$ with respect to time t .

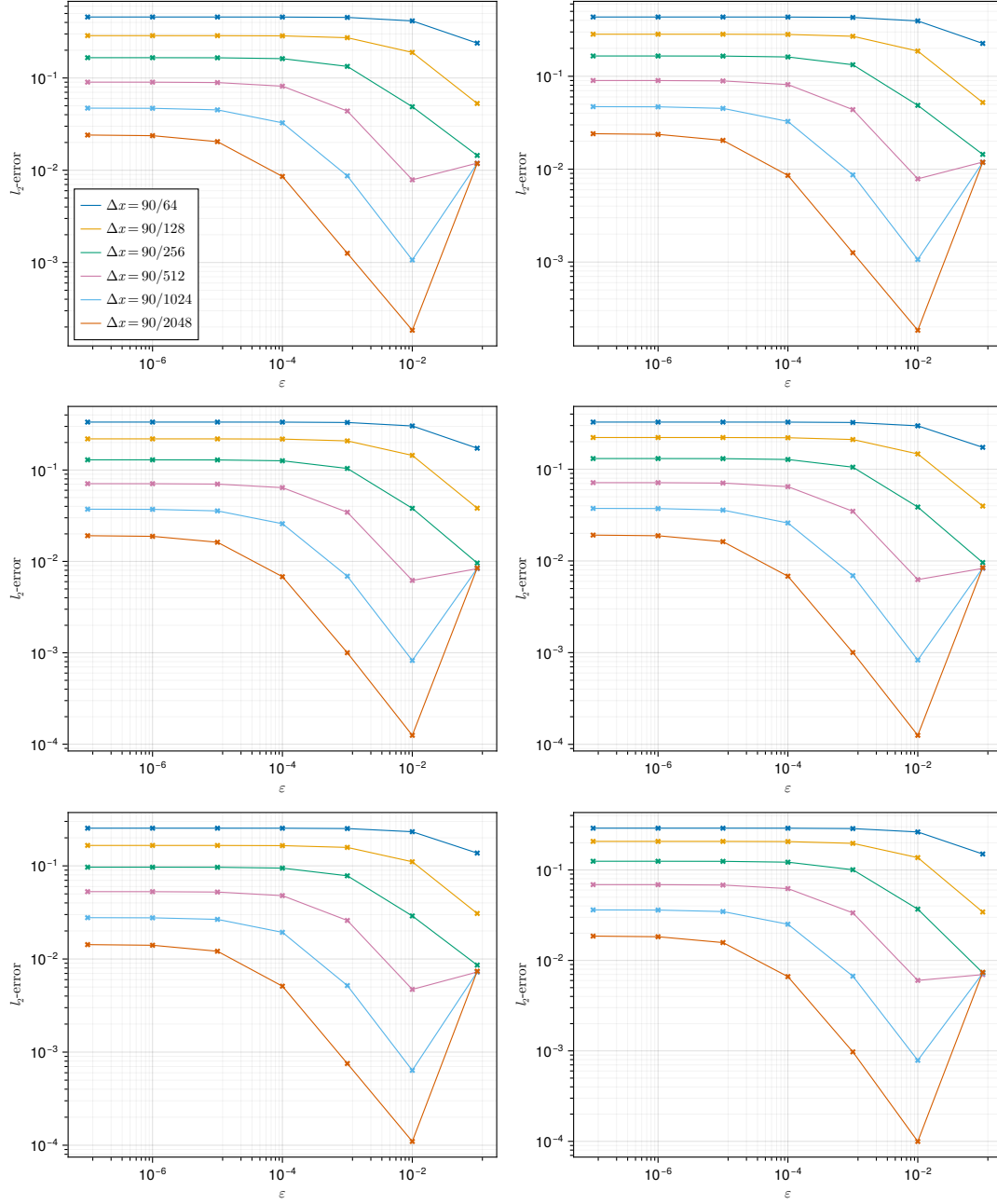


Figure 15: The results of Experiment 2.1 (first part) for the classical IMEX Active Flux scheme with the ARS(4,4,3) scheme as time discretisation. From left to bottom right: The l_2 -error of $u_{i+\frac{1}{2}}$, U_i , $\rho_{i+\frac{1}{2}}$, P_i , $\psi_{i+\frac{1}{2}}$, and Ψ_i at $t = 4$ with respect to ε for different spatial step sizes.

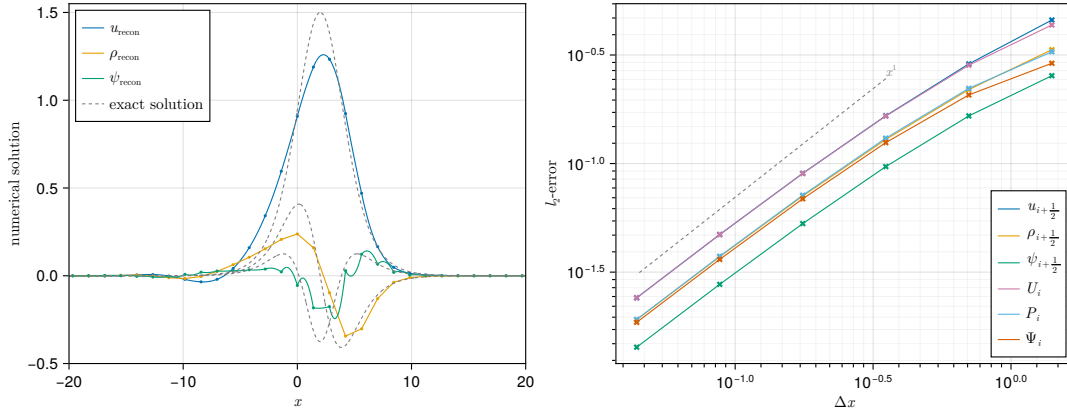


Figure 16: The results of Experiment 2.1 (second part) for the classical IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

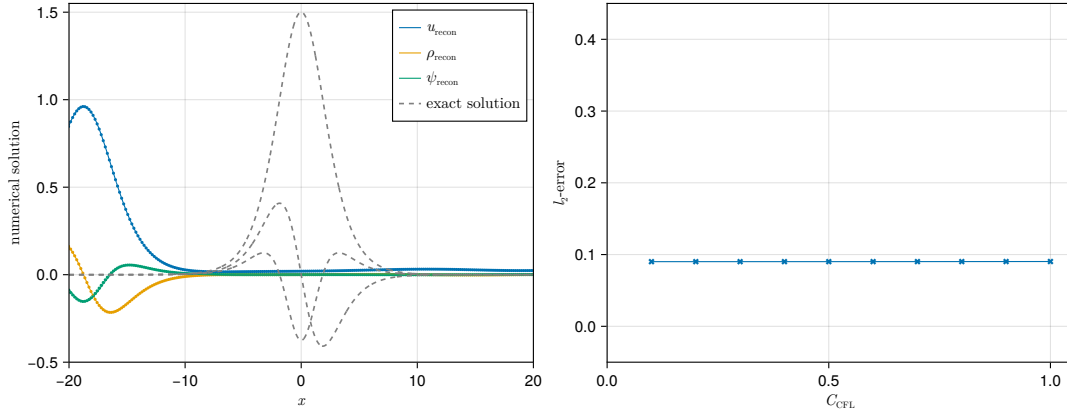


Figure 17: The results of Experiment 2.2 for the classical IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

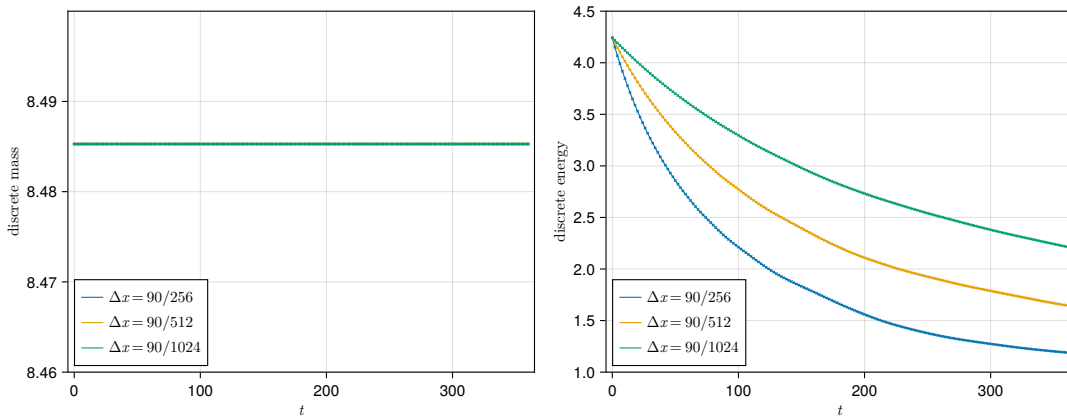


Figure 18: The results of Experiment 2.3 for the classical IMEX Active Flux scheme with the ARS(4, 4, 3) time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy E_{dis} with respect to time t .

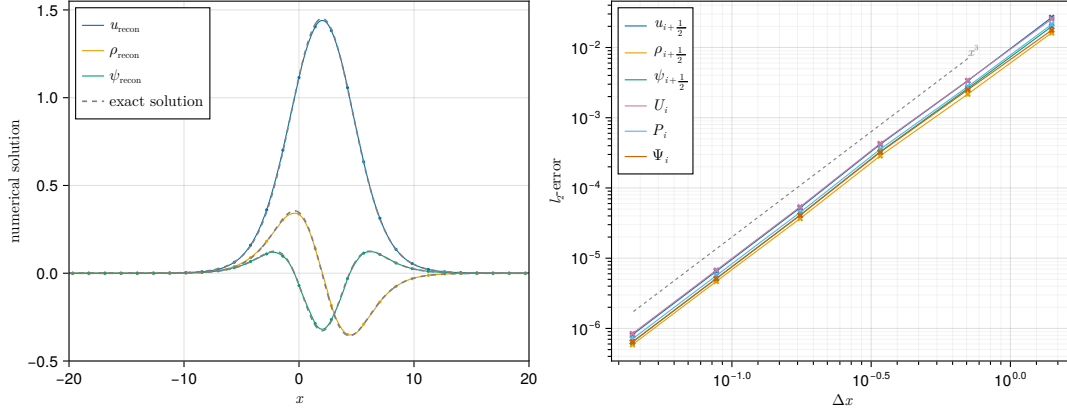


Figure 19: The results of Experiment 1.1 for the classical IMEX Active Flux scheme with the ARS(2,3,3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

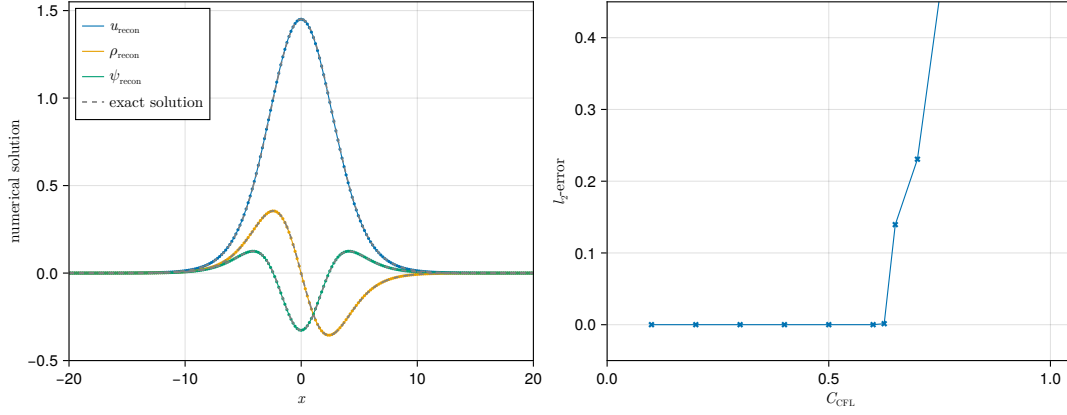


Figure 20: The results of Experiment 1.2 for the classical IMEX Active Flux scheme with the ARS(2,3,3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

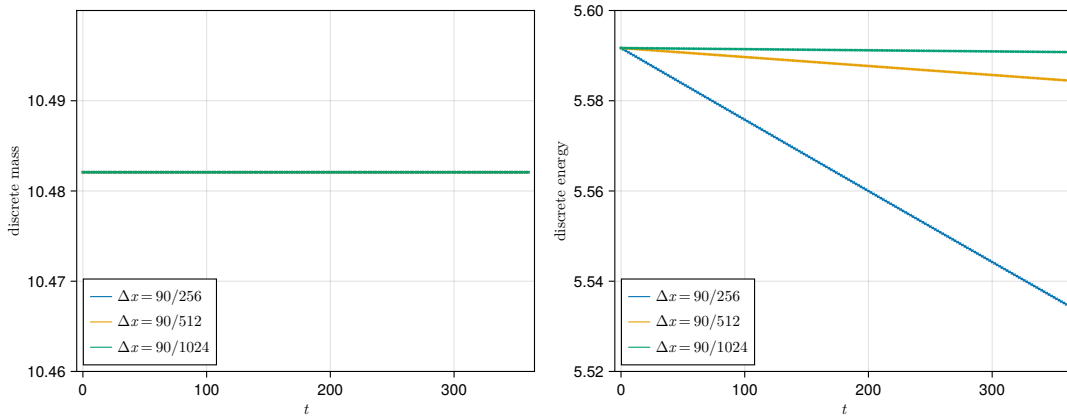


Figure 21: The results of Experiment 1.3 for the classical IMEX Active Flux scheme with the ARS(2,3,3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy $E_{\text{H,dis}}$ with respect to time t .

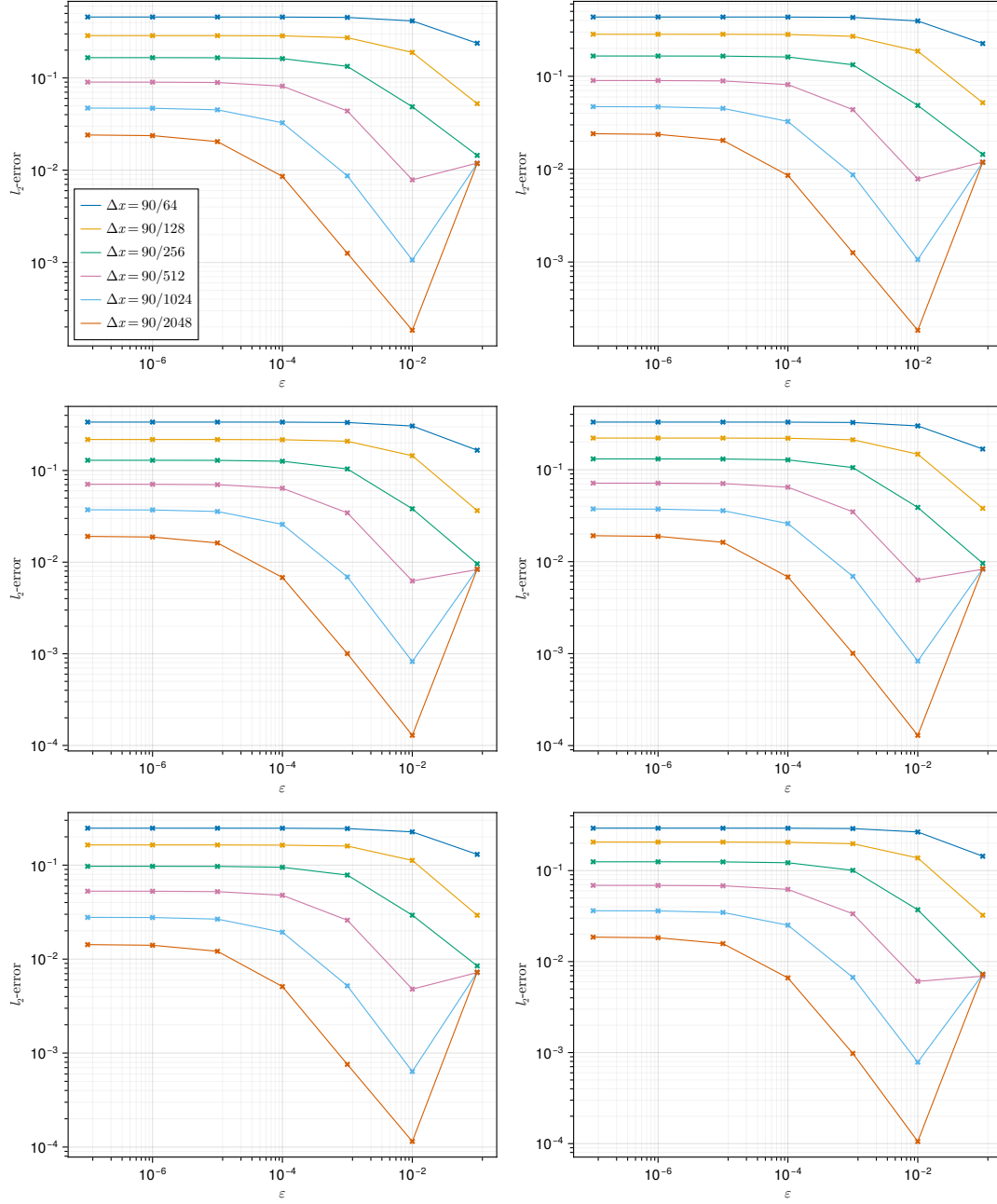


Figure 22: The results of Experiment 2.1 (first part) for the classical IMEX Active Flux scheme with the ARS(2,3,3) scheme as time discretisation. From to left to bottom right: the l_2 -error of $u_{i+\frac{1}{2}}$, U_i , $\rho_{i+\frac{1}{2}}$, P_i , $\psi_{i+\frac{1}{2}}$, and Ψ_i at $t = 4$ with respect to ε for different spatial step sizes.

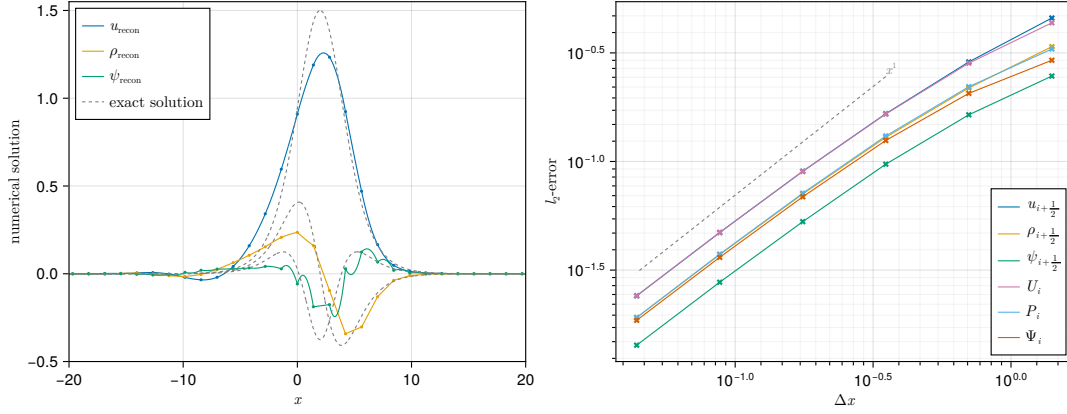


Figure 23: The results of Experiment 2.1 (second part) for the classical IMEX Active Flux scheme with the ARS(2, 3, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

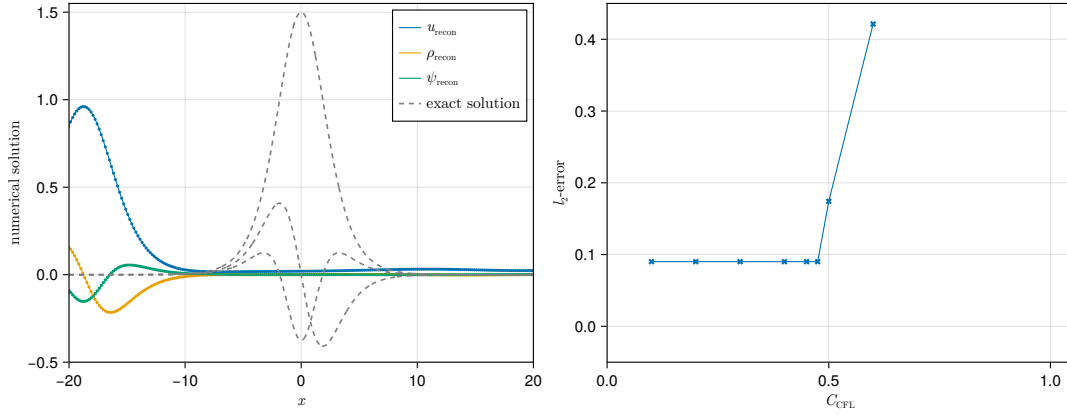


Figure 24: The results of Experiment 2.2 for the classical IMEX Active Flux scheme with the ARS(2, 3, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

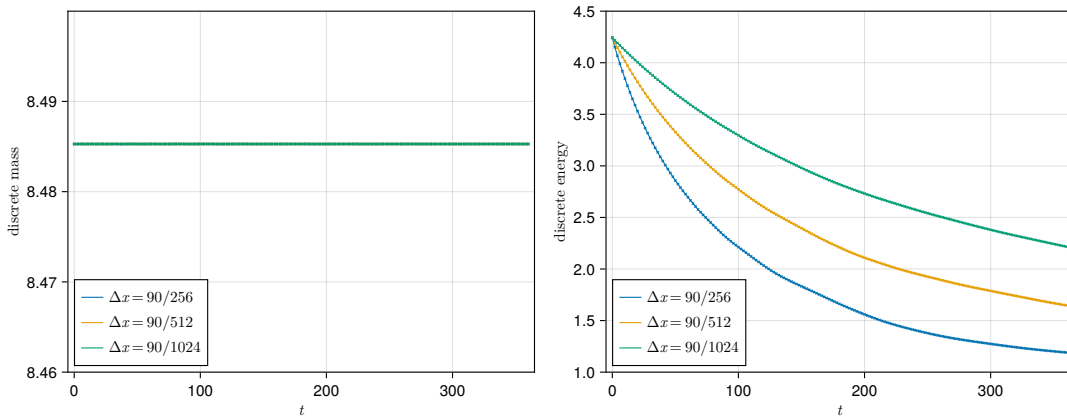


Figure 25: The results of Experiment 2.3 for the classical IMEX Active Flux scheme with the ARS(2, 3, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy E_{dis} with respect to time t .

4 Alternative asymptotic-preserving IMEX Active Flux schemes

The accuracy decrease of an Active Flux scheme for a hyperbolic approximation was also noticed by the authors in [24] in case of the hyperbolic heat equation. There, the semi-discrete Active Flux scheme (combined with implicit time discretisation) achieves only second order accuracy, if the relaxation parameter is chosen sufficiently small. For the one-dimensional heat equation, the authors propose an alternative point value update equation, and it is shown that an Active Flux scheme which uses this point value update equation achieves at least third order accuracy in all degrees of freedom, even if ε is small. For more details about the alternative point value update equation, refer to the original article [24]. The shortcomings of the classical IMEX Active Flux scheme, observed in the section before, lead to the consideration of different update equations for the point values to find an IMEX Active Flux scheme that has superior properties compared to the classical IMEX Active Flux scheme. Several different schemes that are the result of modifying the update equations of the point values, were analysed both numerically and theoretically. In the following part, some of these schemes are presented. The formal asymptotic analysis of all these schemes is similar to the one of the classical scheme. Thus, only the differences are highlighted and identical part are omitted. In addition, for all schemes presented in this chapter, Mathematica [37] is employed to assist in the formal asymptotic analysis.

4.1 Simplified IMEX Active Flux scheme

Consider the matrices $F_2^\pm$ of the classical IMEX Active Flux scheme,

$$F_2^+ = \begin{pmatrix} \frac{1}{2\varepsilon} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} & 0 & \frac{1}{2\varepsilon} \end{pmatrix} \quad \text{and} \quad F_2^- = \begin{pmatrix} -\frac{1}{2\varepsilon} & 0 & \frac{1}{2} \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & -\frac{1}{2\varepsilon} \end{pmatrix}.$$

A naive way of modifying the scheme is to remove all elements of the matrices above that might cause unnatural conditions, i.e. the first element in the first row and the third in the third row since these elements lead to terms of the form

$$\frac{1}{2} \left(D^+ v_{i+\frac{1}{2}}(t) - D^- v_{i+\frac{1}{2}}(t) \right)$$

appearing in the update equations of the point values. Removing the terms leaves us with

$$F_{2,\text{simplified}}^+ = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} & 0 & 0 \end{pmatrix} \quad \text{and} \quad F_{2,\text{simplified}}^- = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & 0 \end{pmatrix}.$$

Notice that

$$F_{2,\text{simplified}}^+ + F_{2,\text{simplified}}^- = F_2$$

as it is the case for $F_2^\pm$. If we use these matrices instead of $F_2^\pm$ and the same matrices $F_1^\pm$ as defined for the classical IMEX Active Flux scheme, we obtain point values update equations that read

$$\begin{aligned} \frac{d}{dt} u_{i+\frac{1}{2}}(t) &= -\max(u_{i+\frac{1}{2}}(t), 0) D^+ u_{i+\frac{1}{2}}(t) - \min(u_{i+\frac{1}{2}}(t), 0) D^- u_{i+\frac{1}{2}}(t) \\ &\quad - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}(t) + D^- \psi_{i+\frac{1}{2}}(t) \right), \\ \frac{d}{dt} \rho_{i+\frac{1}{2}}(t) &= \frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}(t) - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}(t), \end{aligned}$$

$$\frac{d}{dt}\psi_{i+\frac{1}{2}}(t) = -\frac{1}{2\varepsilon^2} \left(D^+ u_{i+\frac{1}{2}}(t) + D^- u_{i+\frac{1}{2}}(t) \right) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}(t).$$

A IMEX Active Flux scheme with these point value update equations and the cell average update equations of the classical scheme is referred to as the simplified IMEX Active Flux scheme or short the simplified scheme.

4.1.1 Formal asymptotic analysis

Similar to the classical IMEX Active Flux scheme, formal asymptotic analysis is performed to analyse the asymptotic-preserving properties of the simplified scheme. We proceed as in the analysis of the classical scheme and insert the power series, which we assume exists, into the equations to compute the approximation at the next time step t^{n+1} . Sorting the resulting terms by the power of ε , the update equations of the cell averages yield the same conditions as for the classical scheme since these update equations are not modified. Especially, (3.2), (3.3) and (3.4) hold. For the point values, the terms of order ε^{-2} and ε^{-1} imply

$$\begin{aligned} 0 &= \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(1)[j]} - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(1)[j]} - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s} \end{aligned}$$

and the terms of order ε^0 imply

$$\begin{aligned} u_{i+\frac{1}{2}}^{[k](0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\max \left((u_{i+\frac{1}{2}}^{(0)[j-1]}, 0) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \right. \\ &\quad \left. \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \right. \\ &\quad \left. + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \right) \right), \\ u_{i+\frac{1}{2}}^{n+1(0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\max \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \\ &\quad \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \\ &\quad + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \right) \end{aligned} \tag{4.1}$$

as well as

$$\begin{aligned}\bar{\rho}_{i+\frac{1}{2}}^{(0)} &= \rho_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(2)[j]} - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s}, \\ \rho_{i+\frac{1}{2}}^{(0)n+1} &= \rho_{i+\frac{1}{2}}^{(0)n} + \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(2)[j]} - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s}\end{aligned}$$

and

$$\begin{aligned}\bar{\psi}_{i+\frac{1}{2}}^{(0)} &= \psi_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s}, \\ \psi_{i+\frac{1}{2}}^{(0)n+1} &= \psi_{i+\frac{1}{2}}^{(0)n} + \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s}.\end{aligned}$$

The integrability of A yields (3.6) and (3.7) for all $j = 1, \dots, s$. Since these equations are satisfies as well as (3.3) and (3.4), the result of Lemma 3.5 also hold for the simplified scheme. Thus, it follows that (4.1) is a temporal discretisation using a type ARS IMEX Runge-Kutta scheme of

$$\begin{aligned}\frac{d}{dt} u_{i+\frac{1}{2}}^{(0)}(t) &= - \left(\max \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t) \right) \right. \\ &\quad + \min \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t) \right) \Big) \\ &\quad - \frac{1}{\Delta x^3} \left(-u_{i-\frac{3}{2}}^{(0)}(t) + 3U_{i-1}^{(0)}(t) - 14u_{i-\frac{1}{2}}^{(0)}(t) + 21U_i^{(0)}(t) + 14u_{i+\frac{1}{2}}^{(0)}(t) \right. \\ &\quad \left. - 45U_{i+1}^{(0)}(t) + 16u_{i+\frac{3}{2}}^{(0)}(t) + 15U_{i+2}^{(0)}(t) - 13u_{i+\frac{5}{2}}^{(0)}(t) + 6U_{i+3}^{(0)}(t) - 2u_{i+\frac{7}{2}}^{(0)}(t) \right).\end{aligned}\tag{4.2}$$

and, additionally, it follows in the same way as for the classical scheme that (3.2) is a time discretisation of (3.9). If the IMEX Runge-Kutta scheme is globally stiffly accurate, (3.10), (3.11), (3.12) and (3.13) follow. Furthermore, (4.1) and (3.2), combined with (3.10), (3.11), (3.12), and (3.13) are a time discretisations of a system of ordinary differential equations consisting of (4.2), (3.9), (3.14), (3.15), and (3.16), (3.17).

Notice that the only difference between (4.2) and (3.8) is the absence of the term

$$-\frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(1)}(t) - 3U_i^{(1)}(t) + 4u_{i+\frac{1}{2}}^{(1)}(t) - 3U_{i+1}^{(1)}(t) + u_{i+\frac{3}{2}}^{(1)}(t) \right)$$

in (4.2). In the proof of Theorem 3.6, this term only contributes to the error. Thus, the results of Theorem 3.6 hold and the formal asymptotic analysis indicates that the simplified IMEX Active Flux scheme for the HKdV equations discretised in time with a type ARS IMEX Runge-Kutta scheme is asymptotic preserving for the component u . If, in addition, the IMEX Runge-Kutta scheme is globally stiffly accurate, then it indicates that the scheme is asymptotic preserving.

4.1.2 Implementation

Since the simplified scheme is closely related to the classical scheme, the implementation of it is similar to the one of the latter one. The only difference are the matrices M_1 , M_2 , and M_3 :

$$M_1 = \begin{pmatrix} 0 & 0 & -\frac{1}{\Delta x} & 0 & 0 & \frac{3}{\Delta x} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & \frac{3}{\Delta x \varepsilon^2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -\frac{3}{\Delta x} \\ 0 & -\frac{4}{\Delta x \varepsilon^2} & -\frac{1}{\varepsilon^2} & 0 & \frac{6}{\Delta x \varepsilon^2} & 0 \\ 0 & \frac{1}{\varepsilon^2} & 0 & -\frac{3}{\Delta x \varepsilon^2} & 0 & 0 \\ 0 & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & -\frac{1}{\varepsilon^2} \\ \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & \frac{1}{\varepsilon^2} & 0 \end{pmatrix},$$

and,

$$M_3 = \begin{pmatrix} 0 & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{2}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

4.1.3 Numerical results

To verify the indications of the formal asymptotic analysis and to further analyse the properties of the simplified scheme, Experiment 1 and Experiment 2 are conducted. The globally stiffly accurate ARS(4, 4, 3) scheme is chosen as the time discretisation and $C_{\text{CFL}} = 0.4$. The results of Experiment 1 can be seen in Figure 26 - 28. The simplified IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme achieves an order of accuracy below two. At $t = 4$ the numerical solution computed on a coarse grid with 64 cells (see Figure 26) has a maximum l_2 -error of approximately 0.0534. In Figure 27 it is shown, that at $t = 180$ some dispersion and dissipation effects are visible and therefore, the numerical solution does not match the exact solution at this time. The l_2 -error of the numerical solution is approximately 0.5154. Moreover, the scheme appears to be stable up to roughly $C_{\text{CFL}} = 0.45$. The result of Experiment 1.3 is shown in Figure 28. The discrete mass is approximately 99.98% (256 cells), 100.01% (512 cells), and 100.00% (1024 cells) of the discrete energy at $t = 0$, while the discrete energy at $t = 360$ is approximately 77.13% (256 cells), 93.72% (512 cells), and 98.96% (1024 cells) of the discrete energy at $t = 0$.

In Figure 29 and 30, one can see the results of Experiment 2.1. If $\varepsilon \rightarrow 0$, one can observe that the l_2 -error of the point values and cell averages becomes constant for a fixed spatial step size. For $\varepsilon = 10^{-7}$, the simplified IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme achieves first order accuracy. These numerical experiments indicate that the simplified IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme is asymptotic preserving but not asymptotic accurate. The numerical solution computed on a coarse grid at $t = 4$ (see Figure 29) has a maximum l_2 -error of roughly 0.1427. The results of Experiment 2.2 are depicted in Figure 31, where one can observe that the numerical solution at $t = 180$ does not match the exact solution. The maximum l_2 -error of the numerical solution at this time is approximately 3.2787. Furthermore, the scheme seems to be stable up to approximately $C_{\text{CFL}} = 0.725$, as one can see on the right side of the figure. The conservation of discrete mass and energy associated to the KdV equation (Experiment 2.3), can be seen in Figure 32. The discrete mass at $t = 360$ is approximately 99.95% (256 cells), 99.98% (512 cells), and 99.99% (1024 cells) of the discrete mass at $t = 0$. The discrete energy at $t = 360$ is approximately 57.50% (256 cells), 71.95% (512 cells), and 83.31% (1024 cells) of the discrete energy at $t = 0$.

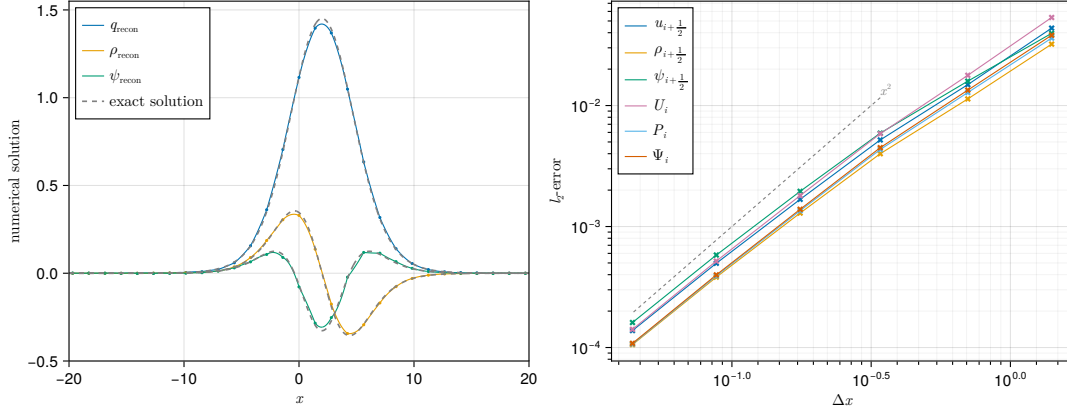


Figure 26: The results of Experiment 1.1 for the simplified IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

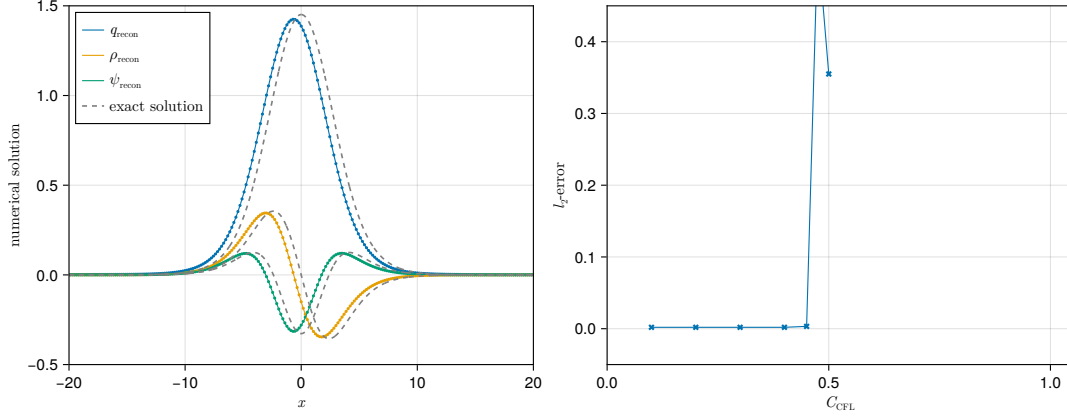


Figure 27: The results of Experiment 1.2 for the simplified IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

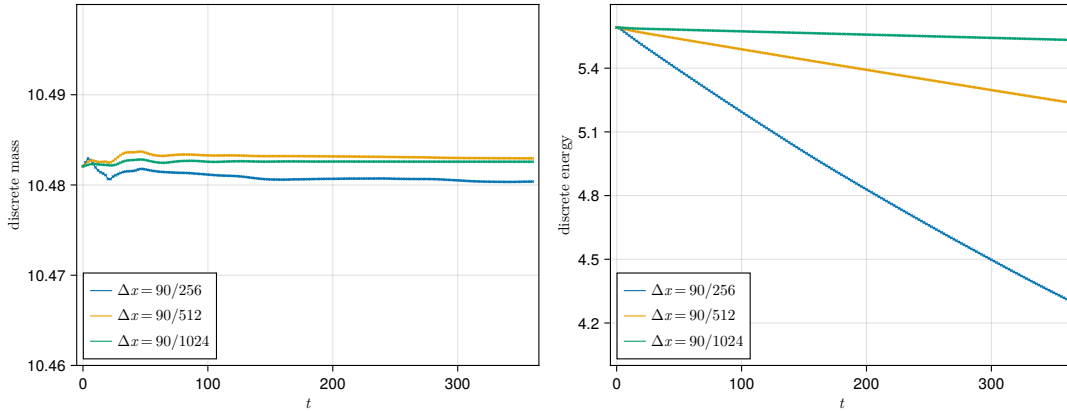


Figure 28: The results of Experiment 1.3 for the simplified IMEX Active Flux scheme with the ARS(4, 4, 3) as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy $E_{H,dis}$ with respect to time t .

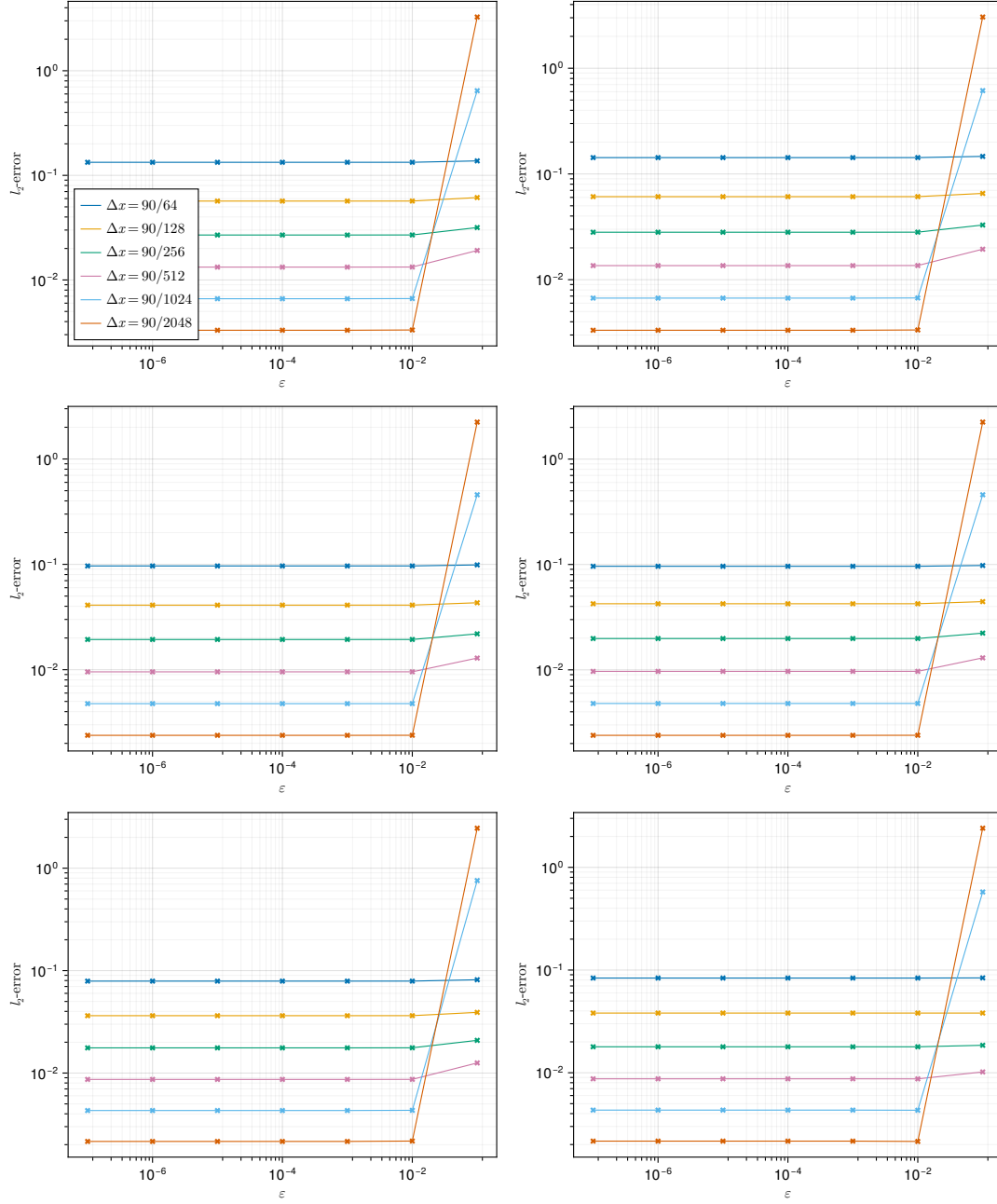


Figure 29: The results of Experiment 2.1 (first part) for the simplified IMEX Active Flux scheme with the ARS(4,4,3) scheme as time discretisation. From to left to bottom right: the l_2 -error of $u_{i+\frac{1}{2}}$, U_i , $\rho_{i+\frac{1}{2}}$, P_i , $\psi_{i+\frac{1}{2}}$, and Ψ_i with respect to ε for different spatial step sizes.

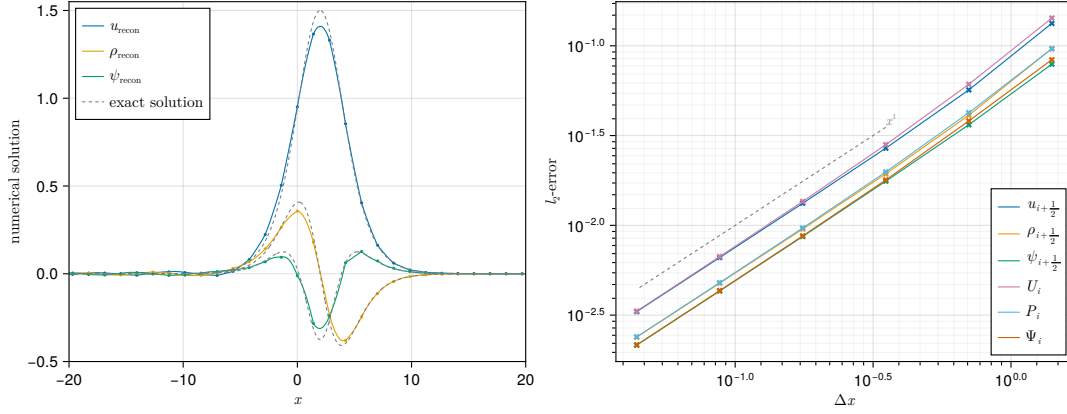


Figure 30: The results of Experiment 2.1 (second part) for the simplified IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

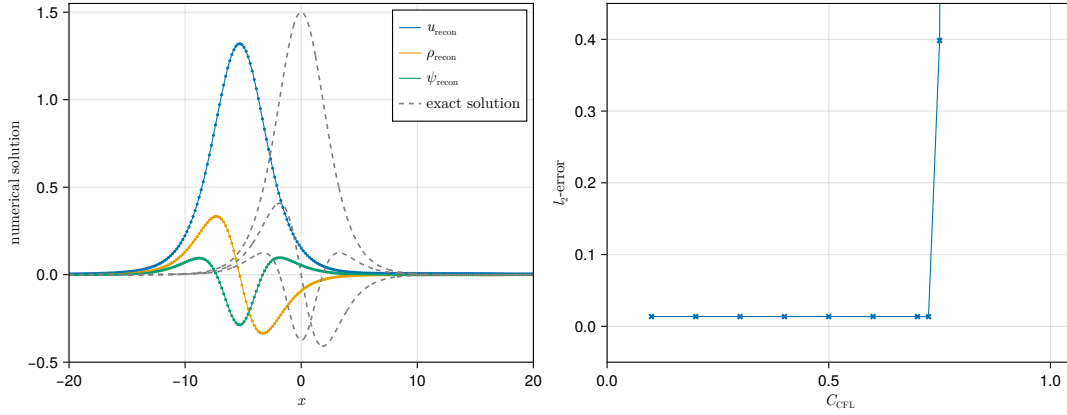


Figure 31: The results of Experiment 2.2 for the simplified IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

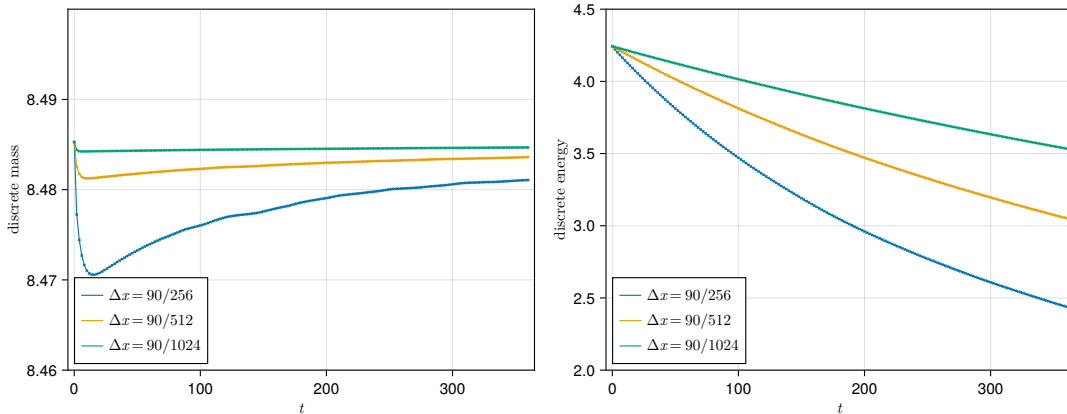


Figure 32: The results of Experiment 2.3 for the simplified IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy E_{dis} with respect to time t .

4.2 Optimised IMEX Active Flux scheme

The simplified scheme achieves in Experiment 1.1 only below second order accuracy, while, in contrast, the classical IMEX Active Flux scheme combined with the ARS(4, 4, 3) scheme or the ARS(2, 3, 3) scheme is able to achieve third order accuracy in the same experiment. To optimise the approach of modifying the elements corresponding to terms of the form

$$\frac{1}{2} \left(D^+ v_{i+\frac{1}{2}}(t) - D^- v_{i+\frac{1}{2}}(t) \right),$$

only the element that is responsible for the unnatural conditions describes in context of the classical scheme, i.e. the first element in the first row of both matrices, is removed. Notice that the third element in the third row scales with $\frac{1}{\varepsilon}$, while the other element in this row scales with $\frac{1}{\varepsilon^2}$. As a result, if $\varepsilon \rightarrow 0$, the relation between the terms in the associated update equation of the point values changes. By changing the scaling of this element to $\frac{1}{\varepsilon^2}$, the relation of the terms in the update equation is independent of ε . Following this new approach, the positive and negative part of the Jacobian read

$$F_{2,\text{optimised}}^+ = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} & 0 & \frac{1}{2\varepsilon^2} \end{pmatrix} \quad \text{and} \quad F_{2,\text{optimised}}^- = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & -\frac{1}{2\varepsilon^2} \end{pmatrix}.$$

As before we use the $F_1^\pm$ in the update equation of the point value as defined for the classical scheme. The update equations for the point values with $F_{2,\text{optimised}}^\pm$ instead of $F_2^\pm$ are

$$\begin{aligned} \frac{d}{dt} u_{i+\frac{1}{2}}(t) &= -\max(u_{i+\frac{1}{2}}, 0) D^+ u_{i+\frac{1}{2}}(t) - \min(u_{i+\frac{1}{2}}, 0) D^- u_{i+\frac{1}{2}}(t) \\ &\quad - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}(t) + D^- \psi_{i+\frac{1}{2}}(t) \right), \\ \frac{d}{dt} \rho_{i+\frac{1}{2}}(t) &= \frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}(t) - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}(t), \\ \frac{d}{dt} \psi_{i+\frac{1}{2}}(t) &= -\frac{1}{2\varepsilon^2} \left(D^+ u_{i+\frac{1}{2}}(t) + D^- u_{i+\frac{1}{2}}(t) \right) - \frac{1}{2\varepsilon^2} \left(D^+ \psi_{i+\frac{1}{2}}(t) - D^- \psi_{i+\frac{1}{2}}(t) \right) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}(t). \end{aligned}$$

The update equations of the cell averages are, again, left unchanged and an IMEX Active Flux scheme with these update equations is named the optimised IMEX Active Flux scheme.

4.2.1 Formal asymptotic analysis

As for the other IMEX Active Flux schemes considered in this text, the asymptotic preserving properties of the optimised scheme are investigated with help of formal asymptotic analysis. The first steps of the analysis, including the assumption that the numerical approximations and stage solution are regular enough and inserting these assumptions into the time-discrete update equations, are the same as for the other schemes. Ordering the terms in time-discrete update equations, the ones associated with the cell averages yields the same equations as in the case of the classical scheme. Thus, we obtain (3.2), (3.3) and (3.4). In case of the optimised scheme, collecting the terms of factor ε^{-2} gives

$$\begin{aligned} 0 &= \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} - D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \end{aligned}$$

$$0 = \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} - D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s},$$

collecting the terms of factor ε^{-1} yields

$$\begin{aligned} 0 &= \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(1)[j]} - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(1)[j]} - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(1)[j]} - D^- \psi_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(1)[j]} - D^- \psi_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \end{aligned}$$

and collecting the terms of factor ε^0 gives

$$\begin{aligned} u_{i+\frac{1}{2}}^{[k](0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\max \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \\ &\quad \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \\ &\quad + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \right), \\ u_{i+\frac{1}{2}}^{n+1(0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\max \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \\ &\quad \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \\ &\quad + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \right) \end{aligned} \tag{4.3}$$

as well as

$$\begin{aligned} \bar{\rho}_{i+\frac{1}{2}}^{(0)} &= \rho_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \Delta t A \left(D^- \rho_{i+\frac{1}{2}}^{(2)[j]} - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s}, \\ \rho_{i+\frac{1}{2}}^{(0)n+1} &= \rho_{i+\frac{1}{2}}^{(0)n} + \Delta t b \cdot \left(D^- \rho_{i+\frac{1}{2}}^{(2)[j]} - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s} \end{aligned}$$

and

$$\begin{aligned} \bar{\psi}_{i+\frac{1}{2}}^{(0)} &= \psi_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) \right. \\ &\quad \left. - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(2)[j]} - D^- \psi_{i+\frac{1}{2}}^{(2)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s}, \\ \psi_{i+\frac{1}{2}}^{(0)n+1} &= \psi_{i+\frac{1}{2}}^{(0)n} + \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) \right. \\ &\quad \left. - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(2)[j]} - D^- \psi_{i+\frac{1}{2}}^{(2)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1,\dots,s} \end{aligned}$$

Since A is assumed to be invertible, we obtain (3.6) and instead of (3.7), we have

$$\rho_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} - D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \tag{4.4}$$

for all $j = 1, \dots, s$. In the following, it is convenient to simplify the notations and therefore define

$$\Lambda_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} - D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right).$$

Since (3.7) is not satisfied, we cannot use our previous results to continue, but we are able to prove a similar result to Lemma 3.5.

Lemma 4.1. *The leading order terms, i.e. the terms of order 0, satisfy*

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{(0)[j]} &= \frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)[j]} + 6U_i^{(0)[j]} - 4u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + 5u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right), \\ &\quad + \frac{1}{\Delta x} \left(-4\Lambda_{i+\frac{1}{2}}^{(0)[j]} - 2\Lambda_{i+\frac{3}{2}}^{(0)[j]} \right) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) &= \frac{1}{\Delta x^3} \left(-u_{i-\frac{3}{2}}^{(0)[j]} + 3U_{i-1}^{(0)[j]} - 14u_{i-\frac{1}{2}}^{(0)[j]} + 21U_i^{(0)[j]} + 14u_{i+\frac{1}{2}}^{(0)[j]} \right. \\ &\quad \left. - 45U_{i+1}^{(0)[j]} + 16u_{i+\frac{3}{2}}^{(0)[j]} + 15U_{i+2}^{(0)[j]} - 13u_{i+\frac{5}{2}}^{(0)[j]} + 6U_{i+3}^{(0)[j]} - 2u_{i+\frac{7}{2}}^{(0)[j]} \right) \\ &\quad + \frac{1}{\Delta x^2} \left(-\Lambda_{i-\frac{1}{2}}^{(0)[j]} - 8\Lambda_{i+\frac{1}{2}}^{(0)[j]} + 7\Lambda_{i+\frac{3}{2}}^{(0)[j]} + 2\Lambda_{i+\frac{5}{2}}^{(0)[j]} \right) \end{aligned}$$

for $j = 1, \dots, s$.

Proof. The proof is similar to the one of Lemma 3.5 since the only difference between (3.7) and (4.4) is the additional term $\Lambda_{i+\frac{1}{2}}^{(0)[j]}$. Thus, using (3.3), (3.4), (3.6), and (4.4), we obtain

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{(0)[j]} &= D^- \rho_{i+\frac{1}{2}}^{(0)[j]} \\ &= \frac{1}{\Delta x} \left(-4\rho_{i+\frac{1}{2}}^{(0)[j]} + 6P_{i+1}^{(0)[j]} - 2\rho_{i+\frac{3}{2}}^{(0)[j]} \right) \\ &= \frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)[j]} + 6U_i^{(0)[j]} - 4u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + 5u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right) \\ &\quad + \frac{1}{\Delta x} \left(-4\Lambda_{i+\frac{1}{2}}^{(0)[j]} - 2\Lambda_{i+\frac{3}{2}}^{(0)[j]} \right) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) &= \frac{1}{\Delta x} \left(\psi_{i-\frac{1}{2}}^{(0)[j]} - 3\Psi_i^{(0)[j]} + 3\Psi_{i+1}^{(0)[j]} - \psi_{i+\frac{3}{2}}^{(0)[j]} \right) \\ &= \frac{1}{\Delta x} \left(D^- \rho_{i-\frac{1}{2}}^{(0)[j]} - \frac{3}{\Delta x} \left(\rho_{i+\frac{1}{2}}^{(0)[j]} - \rho_{i-\frac{1}{2}}^{(0)[j]} \right) \right. \\ &\quad \left. + \frac{3}{\Delta x} \left(\rho_{i+\frac{3}{2}}^{(0)[j]} - \rho_{i+\frac{1}{2}}^{(0)[j]} \right) - D^- \rho_{i+\frac{3}{2}}^{(0)[j]} \right) \\ &= \frac{1}{\Delta x^2} \left(-\rho_{i-\frac{1}{2}}^{(0)[j]} + 6P_i^{(0)[j]} - 8\rho_{i+\frac{1}{2}}^{(0)[j]} + 7\rho_{i+\frac{3}{2}}^{(0)[j]} - 6P_{i+2}^{(0)[j]} + 2\rho_{i+\frac{5}{2}}^{(0)[j]} \right) \\ &= \frac{1}{\Delta x^3} \left(-u_{i-\frac{3}{2}}^{(0)[j]} + 3U_{i-1}^{(0)[j]} - 14u_{i-\frac{1}{2}}^{(0)[j]} + 21U_i^{(0)[j]} + 14u_{i+\frac{1}{2}}^{(0)[j]} \right. \\ &\quad \left. - 45U_{i+1}^{(0)[j]} + 16u_{i+\frac{3}{2}}^{(0)[j]} + 15U_{i+2}^{(0)[j]} - 13u_{i+\frac{5}{2}}^{(0)[j]} + 6U_{i+3}^{(0)[j]} - 2u_{i+\frac{7}{2}}^{(0)[j]} \right) \\ &\quad + \frac{1}{\Delta x^2} \left(-\Lambda_{i-\frac{1}{2}}^{(0)[j]} - 8\Lambda_{i+\frac{1}{2}}^{(0)[j]} + 7\Lambda_{i+\frac{3}{2}}^{(0)[j]} + 2\Lambda_{i+\frac{5}{2}}^{(0)[j]} \right) \end{aligned}$$

□

The results of Lemma 4.1 imply that (4.3) is a temporal discretisation using an type ARS IMEX Runge-Kutta scheme of

$$\begin{aligned} \frac{d}{dt} u_{i+\frac{1}{2}}^{(0)}(t) = & - \left(\max \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t) \right) \right. \\ & + \min \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t) \right) \Big) \\ & - \frac{1}{\Delta x^3} \left(-u_{i-\frac{3}{2}}^{(0)}(t) + 3U_{i-1}^{(0)}(t) - 14u_{i-\frac{1}{2}}^{(0)}(t) + 21U_i^{(0)}(t) + 14u_{i+\frac{1}{2}}^{(0)}(t) \right. \\ & - 45U_{i+1}^{(0)}(t) + 16u_{i+\frac{3}{2}}^{(0)}(t) + 15U_{i+2}^{(0)}(t) - 13u_{i+\frac{5}{2}}^{(0)}(t) + 6U_{i+3}^{(0)}(t) - 2u_{i+\frac{7}{2}}^{(0)}(t) \Big) \\ & \left. - \frac{1}{\Delta x^2} \left(-\Lambda_{i-\frac{1}{2}}^{(0)}(t) - 8\Lambda_{i+\frac{1}{2}}^{(0)}(t) + 7\Lambda_{i+\frac{3}{2}}^{(0)}(t) + 2\Lambda_{i+\frac{5}{2}}^{(0)}(t) \right) \right) \end{aligned} \quad (4.5)$$

and (3.2) is a discretisation of

$$\begin{aligned} \frac{d}{dt} U_i^{(0)}(t) = & - \frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{(0)}(t) \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{(0)}(t) \right)^2 \right) \\ & - \frac{2}{\Delta x^3} \left(\left(-2u_{i-\frac{1}{2}}^{(0)}(t) + 6U_i^{(0)}(t) - 4u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + 5u_{i+\frac{3}{2}}^{(0)}(t) \right. \right. \\ & - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \Big) - \left(-2u_{i-\frac{3}{2}}^{(0)}(t) + 6U_{i-1}^{(0)}(t) - 4u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) \right. \\ & \left. \left. + 5u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + u_{i+\frac{3}{2}}^{(0)}(t) \right) \right) \\ & - \frac{1}{\Delta x^2} \left(\left(-4\Lambda_{i+\frac{1}{2}}^{(0)}(t) - 2\Lambda_{i+\frac{3}{2}}^{(0)}(t) \right) - \left(-4\Lambda_{i-\frac{1}{2}}^{(0)}(t) - 2\Lambda_{i+\frac{1}{2}}^{(0)}(t) \right) \right), \end{aligned} \quad (4.6)$$

where

$$\Lambda_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)}(t) - D^- \psi_{i+\frac{1}{2}}^{(0)}(t) \right).$$

Assume the IMEX-Runge Kutta scheme of type ARS is globally stiffly accurate. Since (3.3) is satisfied (3.12) immediately follows. In addition, we obtain

$$\begin{aligned} \rho_{i+\frac{1}{2}}^{(0)n+1} = & \frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(0)n+1} - 3U_i^{(0)n+1} + 3U_{i+1}^{(0)n+1} - u_{i+\frac{3}{2}}^{(0)n+1} \right) \\ & + \frac{1}{\Delta x} \left(\psi_{i-\frac{1}{2}}^{(0)n+1} - 3\Psi_i^{(0)n+1} + 4\psi_{i+\frac{1}{2}}^{(0)n+1} - 3\Psi_{i+1}^{(0)n+1} + \psi_{i+\frac{3}{2}}^{(0)n+1} \right) \end{aligned} \quad (4.7)$$

from (4.4) and

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{(0)n+1} = & \frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)n+1} + 6U_i^{(0)n+1} - 4u_{i+\frac{1}{2}}^{(0)n+1} - 3U_{i+1}^{(0)n+1} + 5u_{i+\frac{3}{2}}^{(0)n+1} \right. \\ & \left. - 3U_{i+2}^{(0)n+1} + u_{i+\frac{5}{2}}^{(0)n+1} \right) + \frac{1}{\Delta x} \left(-4\Lambda_{i+\frac{1}{2}}^{(0)n+1} - 2\Lambda_{i+\frac{3}{2}}^{(0)n+1} \right) \end{aligned} \quad (4.8)$$

form (3.6), where

$$\Lambda_{i+\frac{1}{2}}^{(0)n+1} = \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)n+1} - D^- \psi_{i+\frac{1}{2}}^{(0)n+1} \right).$$

Lastly (3.4) implies

$$\begin{aligned} \Psi_i^{(0)n+1} & \frac{1}{\Delta x^2} \left(\left(u_{i-\frac{1}{2}}^{(0)n+1} - 3U_i^{(0)n+1} + 3U_{i+1}^{(0)n+1} - u_{i+\frac{3}{2}}^{(0)n+1} \right) \right. \\ & \quad \left. - \left(u_{i-\frac{3}{2}}^{(0)n+1} - 3U_{i-1}^{(0)n+1} + 3U_i^{(0)n+1} - u_{i+\frac{1}{2}}^{(0)n+1} \right) \right) \\ & + \frac{1}{\Delta x^2} \left(\left(\psi_{i-\frac{1}{2}}^{(0)n+1} - 3\Psi_i^{(0)n+1} + 4\psi_{i+\frac{1}{2}}^{(0)n+1} - 3\Psi_{i+1}^{(0)n+1} + \psi_{i+\frac{3}{2}}^{(0)n+1} \right) \right. \\ & \quad \left. - \left(\psi_{i-\frac{3}{2}}^{(0)n+1} - 3\Psi_{i-1}^{(0)n+1} + 4\psi_{i-\frac{1}{2}}^{(0)n+1} - 3\Psi_i^{(0)n+1} + \psi_{i+\frac{1}{2}}^{(0)n+1} \right) \right) \end{aligned} \quad (4.9)$$

Thus, (4.3), (3.2), (3.12), (4.7), (4.8), and (4.9) are a temporal discretisation of the semi-discretisation (4.5), (4.6), (3.16), and

$$\begin{aligned} \Psi_i^{(0)}(t) & = \frac{1}{\Delta x^2} \left(\left(u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) + 3U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t) \right) \right. \\ & \quad \left. - \left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) + 3U_i^{(0)}(t) - u_{i+\frac{1}{2}}^{(0)}(t) \right) \right) \\ & + \frac{1}{\Delta x^2} \left(\left(\psi_{i-\frac{1}{2}}^{(0)}(t) - 3\Psi_i^{(0)}(t) + 4\psi_{i+\frac{1}{2}}^{(0)}(t) - 3\Psi_{i+1}^{(0)}(t) + \psi_{i+\frac{3}{2}}^{(0)}(t) \right) \right. \\ & \quad \left. - \left(\psi_{i-\frac{3}{2}}^{(0)}(t) - 3\Psi_{i-1}^{(0)}(t) + 4\psi_{i-\frac{1}{2}}^{(0)}(t) - 3\Psi_i^{(0)}(t) + \psi_{i+\frac{1}{2}}^{(0)}(t) \right) \right) \end{aligned} \quad (4.10)$$

as well as

$$\begin{aligned} \rho^{(0)}(t)_{i+\frac{1}{2}} & = \frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) + 3U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t) \right) \\ & + \frac{1}{\Delta x} \left(\psi_{i-\frac{1}{2}}^{(0)}(t) - 3\Psi_i^{(0)}(t) + 4\psi_{i+\frac{1}{2}}^{(0)}(t) - 3\Psi_{i+1}^{(0)}(t) + \psi_{i+\frac{3}{2}}^{(0)}(t) \right) \end{aligned} \quad (4.11)$$

and

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{(0)}(t) & = \frac{2}{\Delta x^2} \left(-2u_{i-\frac{1}{2}}^{(0)}(t) + 6U_i^{(0)}(t) - 4u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + 5u_{i+\frac{3}{2}}^{(0)}(t) \right. \\ & \quad \left. - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right) + \frac{1}{\Delta x} \left(-4\Lambda_{i+\frac{1}{2}}^{(0)}(t) - 2\Lambda_{i+\frac{3}{2}}^{(0)}(t) \right). \end{aligned} \quad (4.12)$$

With these finding we can prove the counterpart of Theorem 3.6.

Theorem 4.2. *Equations (4.5) and (4.6) are spatial discretisations of*

$$\frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) = 0$$

and

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) dx = 0$$

with an error in $\mathcal{O}(\Delta x)$. In addition, (4.10) is a semi-discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \psi^{(0)}(t, x) dx = \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) dx$$

with an error in $\mathcal{O}(\Delta x^2)$ and (4.11) and (4.12) are spatial discretisations of

$$\rho^{(0)}(t, x) = \frac{\partial}{\partial x} u^{(0)}(t, x),$$

with an error in $\mathcal{O}(\Delta x^2)$ and

$$\psi^{(0)}(t, x) = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x)$$

with an error in $\mathcal{O}(\Delta x)$, respectively.

Proof. Assume that $u^0 : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$ and $\psi^{(0)} : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$ are smooth and bounded functions with bounded derivatives that interpolate the point values and cell averages. The assumptions imply

$$\begin{aligned} \frac{1}{\Delta x} \left(-4\Lambda_{i+\frac{1}{2}}^{(0)}(t) - 2\Lambda_{i+\frac{3}{2}}^{(0)}(t) \right) &= \frac{2}{\Delta x^2} \left(-2\psi_{i-\frac{1}{2}}^{(0)}(t) + 6\Psi_i^{(0)}(t) - 9\psi_{i+\frac{1}{2}}^{(0)}(t) + 9\Psi_{i+1}^{(0)}(t) \right. \\ &\quad \left. - 6\psi_{i+\frac{3}{2}}^{(0)}(t) + 3\Psi_{i+2}^{(0)}(t) - \psi_{i+\frac{5}{2}}^{(0)}(t) \right) \\ &= -\frac{\Delta x^3}{5} \frac{\partial^4}{\partial x^4} \psi^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\Delta x^4}{15} \frac{\partial^5}{\partial x^5} \psi^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^5) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{\Delta x^2} \left(-\Lambda_{i-\frac{1}{2}}^{(0)}(t) - 8\Lambda_{i+\frac{1}{2}}^{(0)}(t) + 7\Lambda_{i+\frac{3}{2}}^{(0)}(t) + 2\Lambda_{i+\frac{5}{2}}^{(0)}(t) \right) \\ = \frac{1}{\Delta x^3} \left(-\psi_{i-\frac{3}{2}}^{(0)}(t) + 3\Psi_{i-1}^{(0)}(t) - 12\psi_{i-\frac{1}{2}}^{(0)}(t) + 27\Psi_i^{(0)}(t) - 26\psi_{i+\frac{1}{2}}^{(0)}(t) + 3\Psi_{i+1}^{(0)}(t) \right. \\ \left. + 22\psi_{i+\frac{3}{2}}^{(0)}(t) - 27\Psi_{i+1}^{(0)}(t) + 15\psi_{i+\frac{5}{2}}^{(0)}(t) - 6\Psi_{i+2}^{(0)}(t) + 2\psi_{i+\frac{7}{2}}^{(0)}(t) \right) \\ = \frac{2\Delta x^3}{5} \frac{\partial^5}{\partial x^5} \psi^{(0)}(t, x_{i+\frac{1}{2}}) + \frac{7\Delta x^4}{30} \frac{\partial^6}{\partial x^6} \psi^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^5). \end{aligned}$$

Therefore, the additional Λ -terms in (4.5) and (4.6) only contribute to the error and as in the proof of Theorem 3.6 it follows that (4.5) and (4.6) are spatial discretisations of

$$\frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) = 0$$

and

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) dx = 0$$

with an error in $\mathcal{O}(\Delta x)$. Furthermore, we obtain

$$\begin{aligned} \frac{1}{\Delta x} \left(\psi_{i-\frac{1}{2}}^{(0)}(t) - 3\Psi_i^{(0)}(t) + 4\psi_{i+\frac{1}{2}}^{(0)}(t) - 3\Psi_{i+1}^{(0)}(t) + \psi_{i+\frac{3}{2}}^{(0)}(t) \right) \\ = \frac{\Delta x^3}{30} \frac{\partial^4}{\partial x^4} \psi^{(0)}(t, x_{i+\frac{1}{2}}) + \frac{\Delta x^5}{630} \frac{\partial^6}{\partial x^6} \psi^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^7). \end{aligned}$$

and

$$\begin{aligned} \frac{1}{\Delta x} \left(\frac{1}{\Delta x} \left(u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) + 3U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t) \right) - \frac{1}{\Delta x} \left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) \right. \right. \\ \left. \left. + 3U_i^{(0)}(t) - u_{i+\frac{1}{2}}^{(0)}(t) \right) \right) + \frac{1}{\Delta x^2} \left(\left(\psi_{i-\frac{1}{2}}^{(0)}(t) - 3\Psi_i^{(0)}(t) + 4\psi_{i+\frac{1}{2}}^{(0)}(t) \right. \right. \\ \left. \left. - 3\Psi_{i+1}^{(0)}(t) + \psi_{i+\frac{3}{2}}^{(0)}(t) \right) - \left(\psi_{i-\frac{3}{2}}^{(0)}(t) - 3\Psi_{i-1}^{(0)}(t) + 4\psi_{i-\frac{1}{2}}^{(0)}(t) - 3\Psi_i^{(0)}(t) + \psi_{i+\frac{1}{2}}^{(0)}(t) \right) \right) \end{aligned}$$

$$= \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) dx + \mathcal{O}(\Delta x^2),$$

where in the last equality, we used the Lipschitz continuity of the derivatives of $u^{(0)}$ and $\psi^{(0)}$. This implies that (4.10) is a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \psi^{(0)}(t, x) dx = \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) dx$$

with an error in $\mathcal{O}(\Delta x^2)$ and (4.11) is a discretisation of

$$\rho^{(0)}(t, x) = \frac{\partial}{\partial x} u^{(0)}(t, x)$$

with an error in $\mathcal{O}(\Delta x^2)$. Lastly, using our previous results, we obtain that (4.12) is a spatial discretisation of

$$\psi^{(0)}(t, x) = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x)$$

with an error in $\mathcal{O}(\Delta x)$. □

This leaves us with the following result: The formal asymptotic analysis indicates that the optimised IMEX Active Flux scheme for the HKdV equations combined with IMEX Runge-Kutta scheme of type ARS to discretise in time is asymptotic preserving for the component u . If, in addition, the IMEX Runge-Kutta scheme is globally stiffly accurate, then the formal asymptotic analysis indicates that the scheme is asymptotic preserving.

4.2.2 Implementation

The only difference in the implementation are again the sub-matrices. The matrices M_1 , M_2 and M_3 are

$$M_1 = \begin{pmatrix} 0 & 0 & -\frac{1}{\Delta x} & 0 & 0 & \frac{3}{\Delta x} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & -\frac{1}{\Delta x \varepsilon^2} & \frac{3}{\Delta x \varepsilon^2} & 0 & \frac{3}{\Delta x \varepsilon^2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -\frac{3}{\Delta x} \\ 0 & -\frac{4}{\Delta x \varepsilon^2} & -\frac{1}{\varepsilon^2} & 0 & \frac{6}{\Delta x \varepsilon^2} & 0 \\ 0 & \frac{1}{\varepsilon^2} & -\frac{4}{\Delta x \varepsilon^2} & -\frac{3}{\Delta x \varepsilon^2} & 0 & \frac{3}{\Delta x \varepsilon^2} \\ 0 & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & -\frac{1}{\varepsilon^2} \\ \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & \frac{1}{\varepsilon^2} & 0 \end{pmatrix},$$

and,

$$M_3 = \begin{pmatrix} 0 & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{2}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ \frac{1}{\Delta x \varepsilon^2} & 0 & -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

4.2.3 Numerical results

The formal asymptotic analysis, especially the errors that appear in it, do not indicate that the optimised scheme behaves differently in the asymptotic limit. To investigate its properties further, Experiment 1 and 2 using the ARS(4, 4, 3) scheme and the ARS(2, 3, 3) scheme as the time discretisations are performed. For all the experiments $C_{\text{CFL}} = 0.35$ is used.

The result of Experiment 1 using the optimised IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme can be seen in Figure 33 - 35. In Figure 33 one can observe the numerical solution at $t = 4$ computed on a coarse grid, where the coloured lines are the reconstruction and the dots are the point values. The numerical solution has a maximum l_2 -error of approximately 0.0251 and the optimised scheme achieves approximately third order accuracy. The numerical solution at $t = 180$ computed on a grid with 512 cells visually matches the exact solution at that time (see Figure 34). The maximum l_2 -error of the numerical solution is approximately 0.0121. In the same figure, one can observe that the optimised scheme appears to be stable up to roughly $C_{\text{CFL}} = 0.45$ on a grid with 512 cells. The results of Experiment 1.3 are depicted in Figure 35. The discrete mass at $t = 360$ in this experiment is approximately 100.03% (256 cells), 100.01% (512 cells), and 100.00% (1024 cells) of the original discrete mass at $t = 0$, and the discrete energy at $t = 360$ is approximately 98.96% (256 cells), 99.87% (512 cells), and 99.98% (1024 cells) of the original discrete energy.

The results of Experiment 2.1, depicted in Figure 36 and 37, show that the l_2 -error of the point values and cell averages becomes constant as $\varepsilon \rightarrow 0$. In contrast to the other schemes considered so far in this text, the optimised scheme with the ARS(4, 4, 3) scheme as a times discretisation achieves third order accuracy for $\varepsilon = 10^{-7}$. Moreover, for this values of ε , the numerical solution at $t = 4$ computed on a coarse grid has a maximum l_2 -error of roughly 0.0837, and the numerical solution at $t = 180$ computed on a grid with 512 cells has a maximum l_2 -error of approximately 0.0578. The numerical solution at this time is shown in Figure 38, where one can also observe that the scheme is stable up to approximately $C_{\text{CFL}} = 0.4$. In addition, the results of Experiment 2.3, i.e. discrete mass and energy with respect to time t are shown Figure 39. The discrete mass at $t = 360$ is approximately 100.03% (256 cells), 100.01% (512 cells), and 100.00% (1024 cells) of the discrete mass at $t = 0$. Furthermore, the discrete energy at $t = 360$ is approximately 96.71% (256 cells), 99.57% (512 cells), and 99.95% (1024 cells) of the discrete energy at $t = 0$.

Discretising in time with the ARS(2, 3, 3) scheme does not significantly affect the results of Experiment 1, which can be seen Figure 40 - 42. The scheme is still able to achieve third order accuracy, the numerical solution on a coarse grid is almost visually indistinguishable from the exact solution with a maximum l_2 -error of approximately 0.0251, and the numerical solution at $t = 180$ visually matches the exact solution with a maximum l_2 -error of approximately 0.0121. Furthermore, the optimised scheme with this time discretisation appears to be stable up to approximately $C_{\text{CFL}} = 0.5$ (see Figure 41). The discrete mass at $t = 360$ is approximately 100.03% (256 cells), 100.01% (512 cells), and 100.00% (1024 cells) of the discrete mass at $t = 0$, while the discrete energy at the last measured time is approximately 98.96% (256 cells), 99.87% (512 cells), and 99.98% (1024 cells) of the discrete energy at $t = 0$ (see Figure 42).

Figure 43 to Figure 46 depicted the result of Experiment 2 for the optimised IMEX Active Flux scheme discretised in time with the ARS(2, 2, 3) scheme. In Figure 43 (Experiment 2.1), one can observe that the l_2 -error becomes constant as $\varepsilon \rightarrow 0$, but differently to the optimised scheme with the ARS(4, 4, 3) scheme as time discretisation, the straight lines representing the constant errors of the cell averages and point values associated to the ρ and ϕ components do not always have the same distance to each other. This results in the behaviour of the l_2 -error of the point values and cell averages with respect to Δx for $\varepsilon = 10^{-7}$ that can be observed in Figure 44. It is not clear, if the scheme achieves third order accuracy in those point values

and cell averages as $\Delta x \rightarrow 0$. The maximum l_2 -error of the numerical solution at $t = 4$ on a coarse grid, which can be seen in Figure 44, is roughly 0.0819, and the maximum l_2 -error of the numerical solution at $t = 180$ computed on a grid with 512 cells is approximately 0.0562 (see Figure 45). Furthermore, the scheme appears to be stable up to approximately $C_{\text{CFL}} = 0.5$. The discrete mass at $t = 360$ is approximately 100.03% (256 cells), 100.01% (512 cells), and 100.00% (1024 cells) of the original discrete mass, and the discrete energy at time $t = 360$ is approximately 96.78% (256 cells), 99.58% (512 cells), and 99.95% (1024 cells) of the discrete energy at $t = 0$.

The numerical experiment indicate that the optimised IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) is asymptotic preserving. However, for the optimised scheme with the ARS(2, 3, 3) scheme as time discretisation, the numerical experiments only indicate that the scheme is asymptotic preserving for the u component. There is no clear indication for the ρ and ψ components since it is not clear how the l_2 -error behaves as $\Delta x \rightarrow 0$. This corresponds to the formal asymptotic analysis, where the assumption that the time discretisation is globally stiffly accurate was needed to obtain the indication that the scheme is asymptotic preserving.

To investigate if the numerical experiments indicate that the optimised scheme discretised in time with the ARS(4, 4, 3) scheme is asymptotic accurate the achieved order of accuracy of the scheme needs to be investigated. In Figure 33 and Figure 37, one can observe that the scheme achieves approximately third order accuracy. To determine more detailed results, the experimental order of convergence is computed. Given the l_2 -errors of the point values and cell averages of a numerical solutions computed on grids with the different number of cells, the experimental order of convergence (EOC) associated to point value $v_{i+\frac{1}{2}}$ is

$$\text{EOC}_{N_1 \rightarrow N_2}(v) = \frac{1}{\log\left(\frac{N_2}{N_1}\right)} \log\left(\frac{\mathcal{E}_{l_2}\left(v, \frac{90}{N_1}\right)}{\mathcal{E}_{l_2}\left(v, \frac{90}{N_2}\right)}\right),$$

where $N_1 < N_2$ denote the number of cells, i.e. 512, 1024, and 2048. Similarly, we define the EOC associated to cell averages V_i as $\text{EOC}_{N_1 \rightarrow N_2}(V)$. These values correspond to the slope, for example, in Figure 37 between the measured l_2 -errors at $\Delta x = \frac{90}{N_1}$ and $\Delta x = \frac{90}{N_2}$. In Table 1, one can see the EOC computed from the results of Experiment 1.1 and Experiment 2.1 for 512, 1024, and 2048 cells. The difference between the EOCs associated to the point values and cell averages obtained from the results of the two Experiments is smaller than 0.05. Thus, the numerical experiments suggest that the optimised IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme is asymptotic accurate.

To further investigate the properties of the optimised IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme, Experiment 3 is performed. As for the first and second experiment, $C_{\text{CFL}} = 0.35$ is chosen. The results of Experiment 3.1 can be seen in Figure 47 and the ones of Experiment 3.2 in Figure 48. One can observe that the numerical solution on a coarse grid does not perfectly match the exact solution. Especially, the amplitude of the first numerically computed soliton is smaller than the one of the exact soliton. The maximum l_2 -error of the numerical solution computed on the grid with 64 cells is approximately 0.2963. Nevertheless, the scheme achieves approximately third order accuracy. The values of $\text{EOC}_{512 \rightarrow 1024}$ for all cell averages and point values range between approximately 2.9863 and 3.0144 and the values of $\text{EOC}_{1024 \rightarrow 2024}$ between 2.9925 and 3.0146. If the numerical solution is computed on a grid with 512 cells, it is visually indistinguishable from exact solution of the initial value problem for the limit equation for all times t , i.e. $t = 0, 20, 40$, and 60 (see Figure 48). The maximum l_2 -error of the numerical solution is approximately 0.0047 ($t = 20$), 0.0173 ($t = 40$), and 0.0443 ($t = 60$). Hence, this experiment further validates the previous indications

that the optimised IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme is asymptotic accurate.

	Experiment 1.1		Experiment 2.1	
	512 $\rightarrow$ 1024	1024 $\rightarrow$ 2048	512 $\rightarrow$ 1024	1024 $\rightarrow$ 2048
$u_{i+\frac{1}{2}}$	2.9433	2.9429	2.9836	2.9899
$\rho_{i+\frac{1}{2}}$	2.9593	2.9754	2.9909	2.9978
$\psi_{i+\frac{1}{2}}$	2.9747	2.9879	3.0160	3.0045
U_i	2.9905	2.9978	2.9890	2.9986
P_i	2.9745	2.9844	2.9947	2.9995
Ψ_i	2.9776	2.9918	3.0061	3.0006

Table 1: The experimental order of convergence computed from the results of Experiment 1.1 and Experiment 2.1 of performed for the optimised IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme.

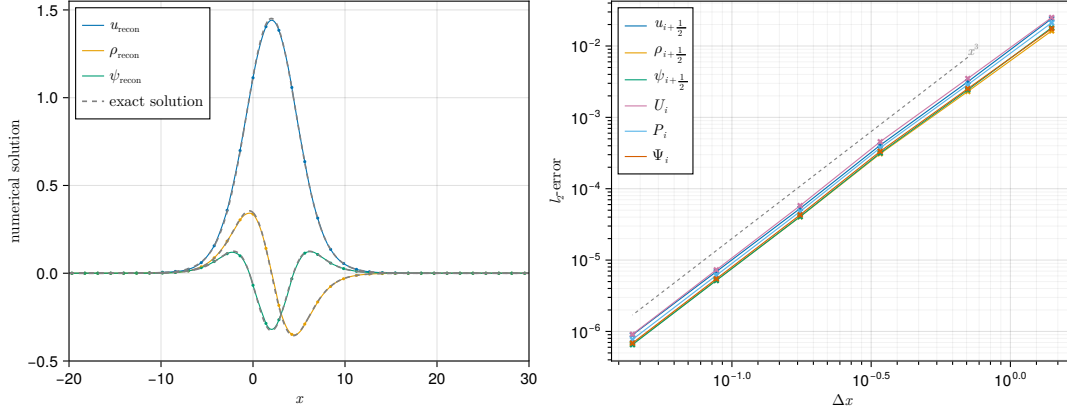


Figure 33: The results of Experiment 1.1 for the optimised IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

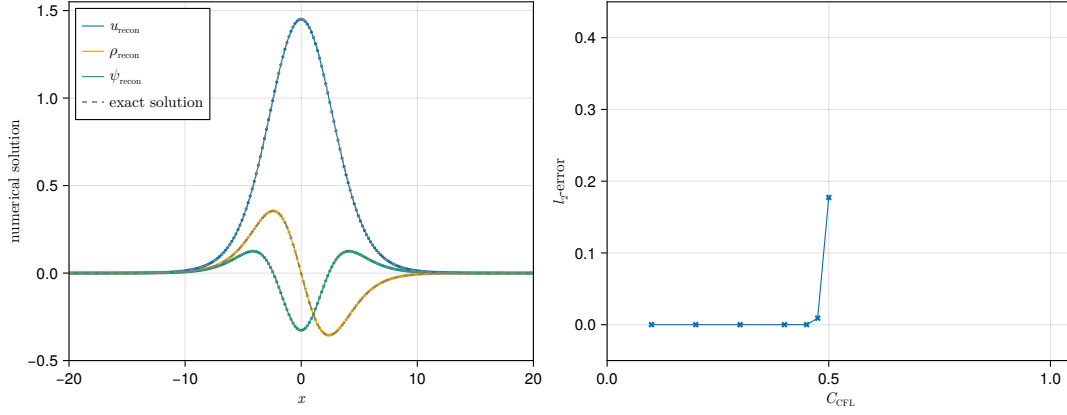


Figure 34: The results of Experiment 1.2 for the optimised IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

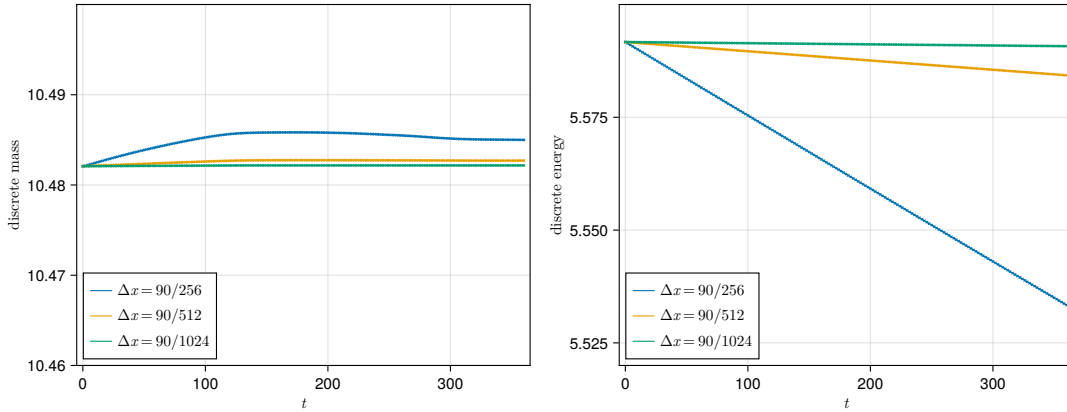


Figure 35: The results of Experiment 1.3 for the optimised IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy $E_{\text{H,dis}}$ with respect to time t .

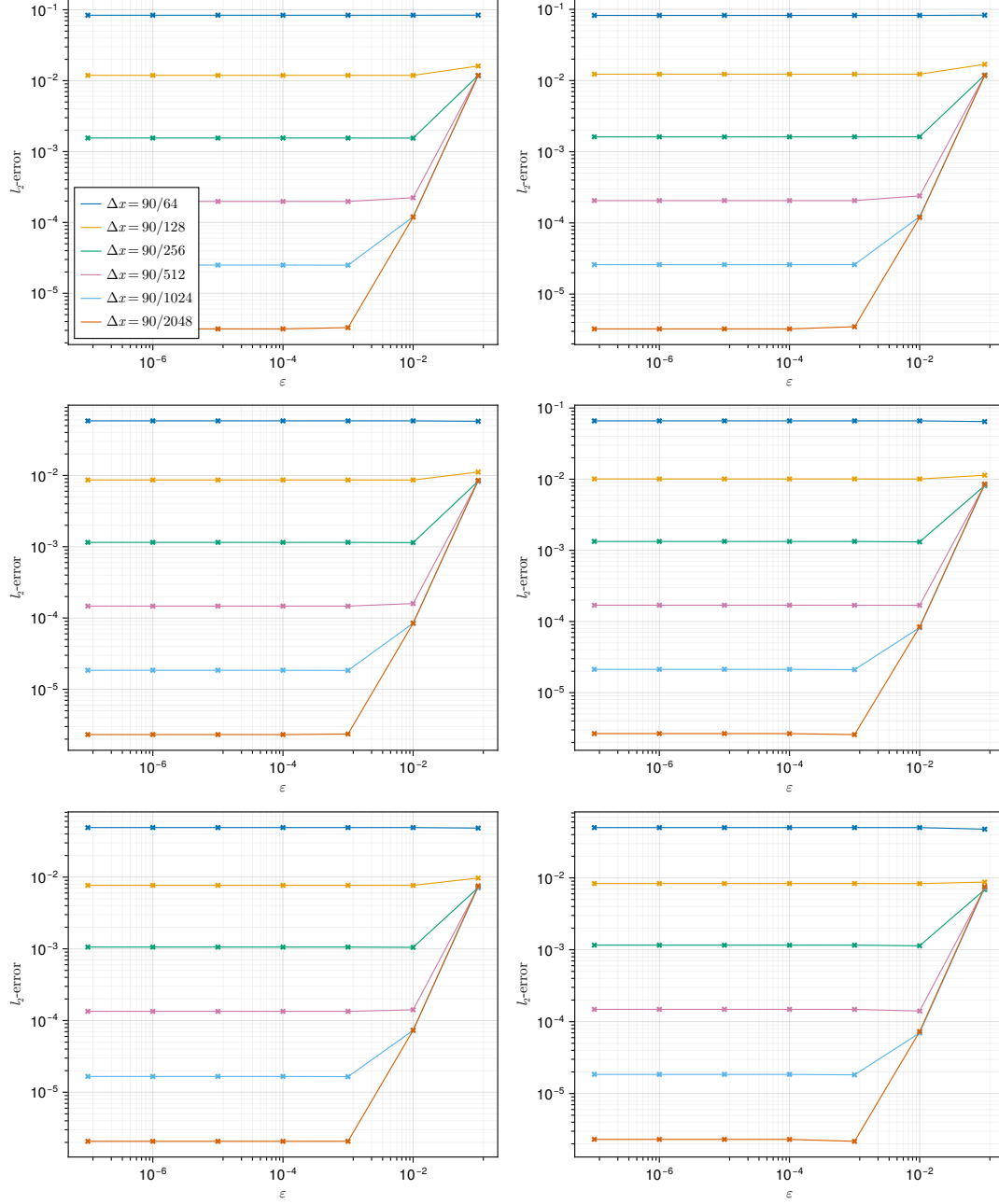


Figure 36: The results of Experiment 2.1 (first part) for the optimised IMEX Active Flux scheme with the ARS(4,4,3) scheme as time discretisation. From to left to bottom right: the l_2 -error of $u_{i+\frac{1}{2}}$, U_i , $\rho_{i+\frac{1}{2}}$, P_i , $\psi_{i+\frac{1}{2}}$, and Ψ_i with respect to ε for different spatial step sizes.

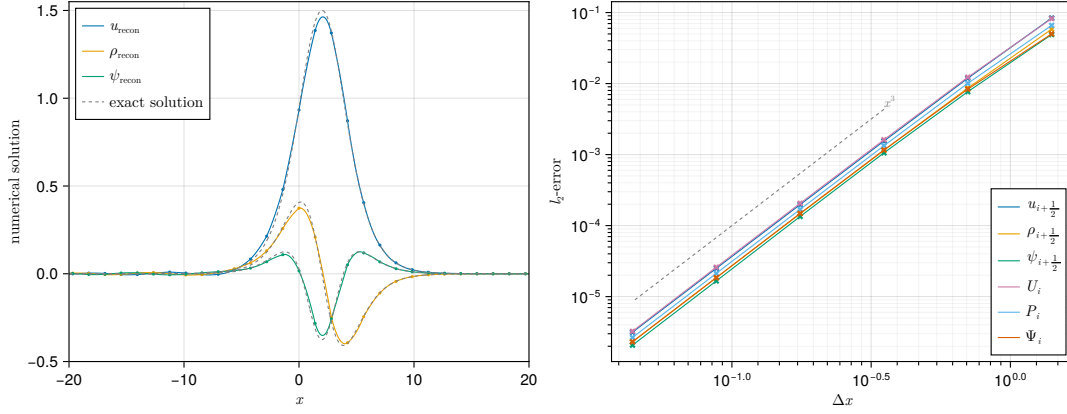


Figure 37: The results of Experiment 2.1 (second part) for the optimised IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

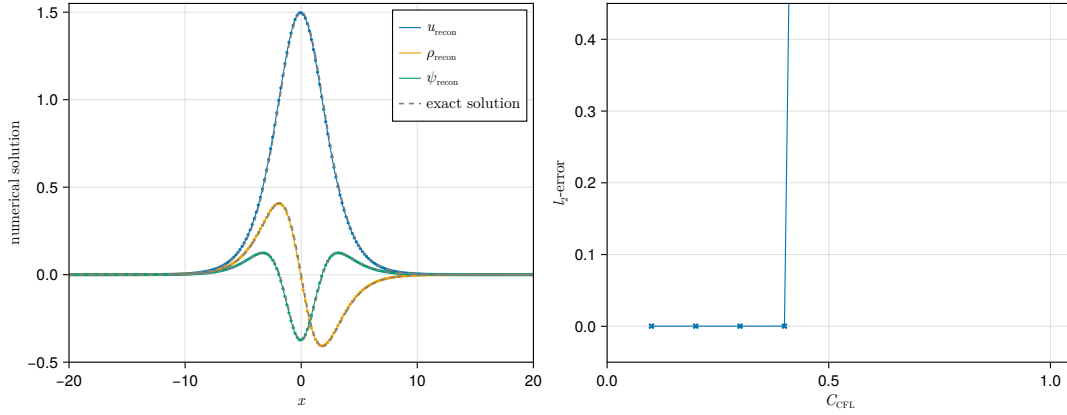


Figure 38: The results of Experiment 2.2 for the optimised IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

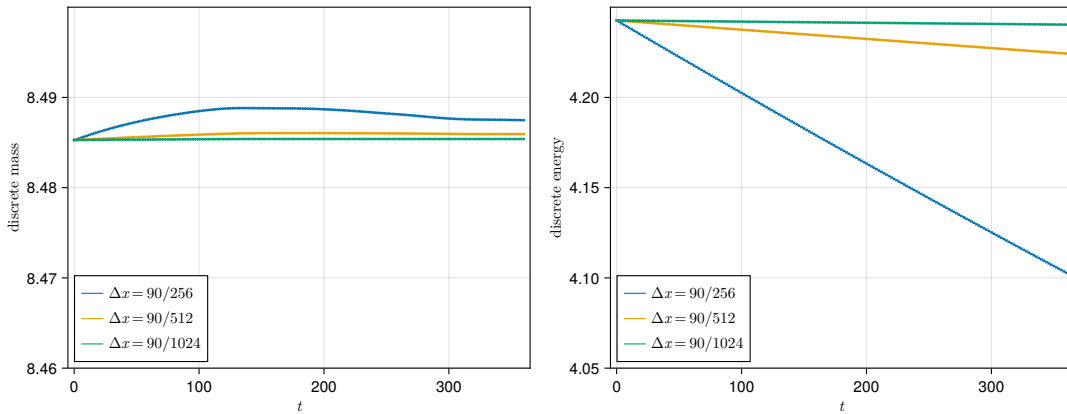


Figure 39: The results of Experiment 2.3 for the optimised IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy E_{dis} with respect to time t .

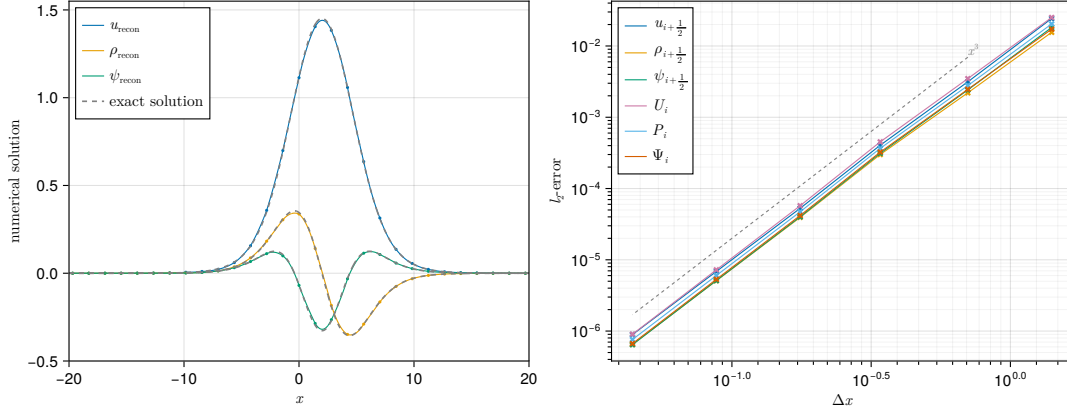


Figure 40: The results of Experiment 1.1 for the optimised IMEX Active Flux scheme with the ARS(2,3,3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

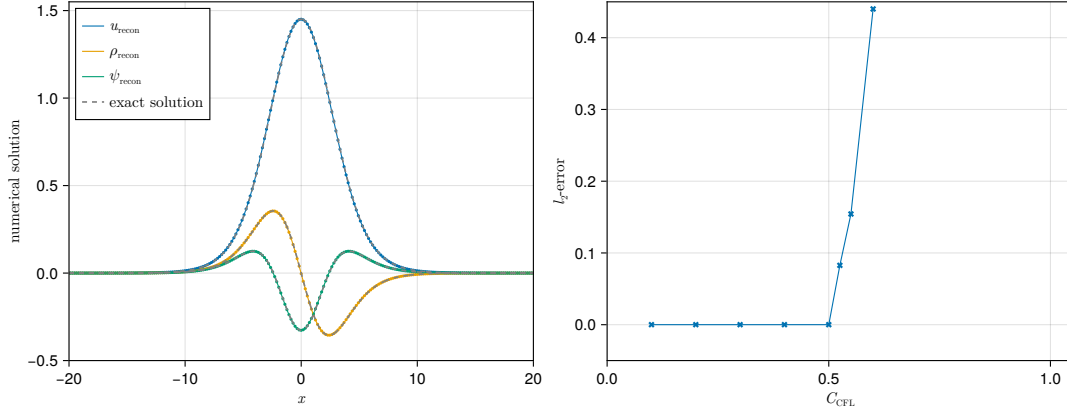


Figure 41: The results of Experiment 1.2 for the optimised IMEX Active Flux scheme with the ARS(2,3,3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

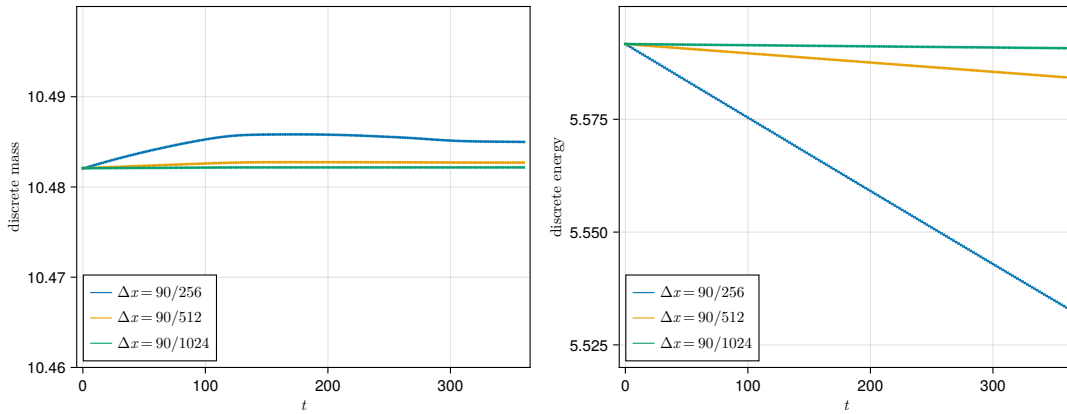


Figure 42: The results of Experiment 1.3 for the optimised IMEX Active Flux scheme with the ARS(2,3,3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy $E_{\text{H,dis}}$ with respect to time t .

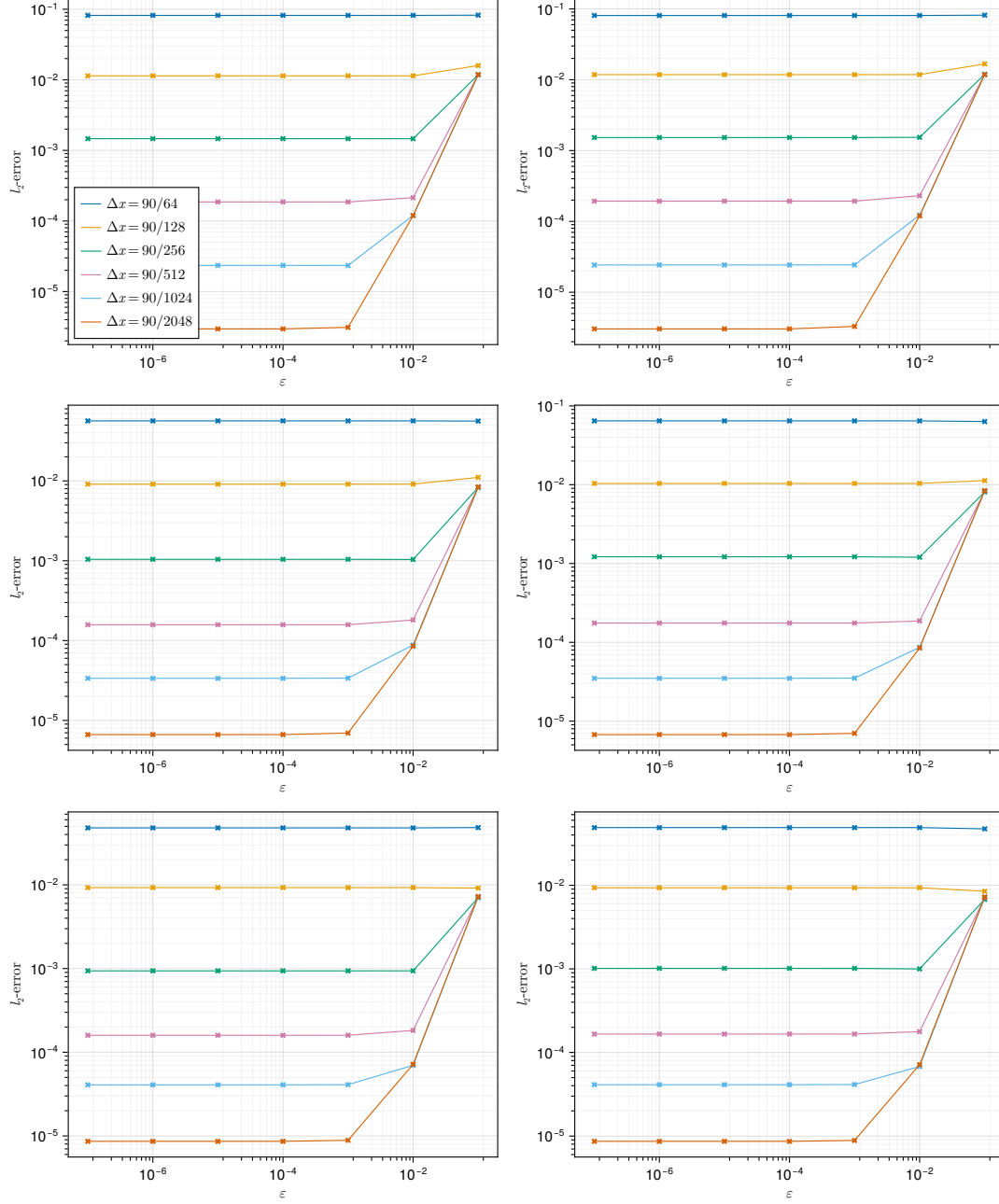


Figure 43: The results of Experiment 2.1 (first part) for the optimised IMEX Active Flux scheme with the ARS(2,3,3) scheme as time discretisation. From top to bottom left to right: the l_2 -error of $u_{i+\frac{1}{2}}$, U_i , $\rho_{i+\frac{1}{2}}$, P_i , $\psi_{i+\frac{1}{2}}$, and Ψ_i with respect to ε for different spatial step sizes.

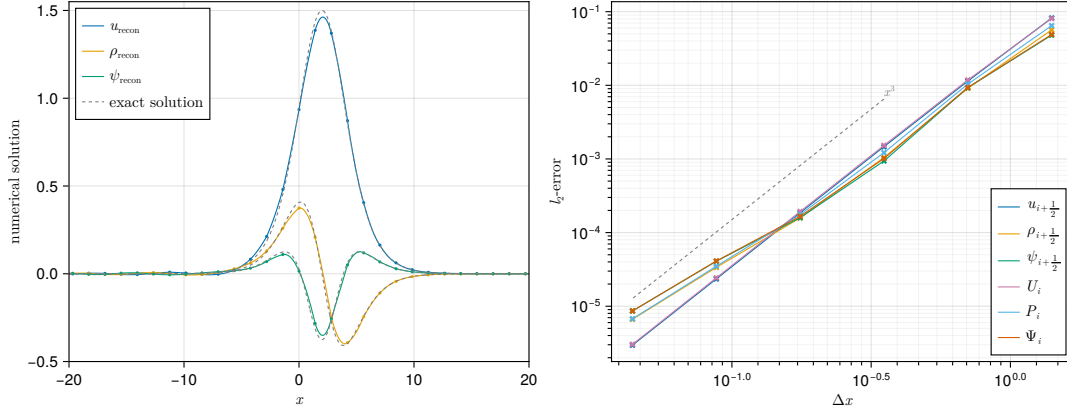


Figure 44: The results of Experiment 2.1 (second part) for the optimised IMEX Active Flux scheme with the ARS(2, 3, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

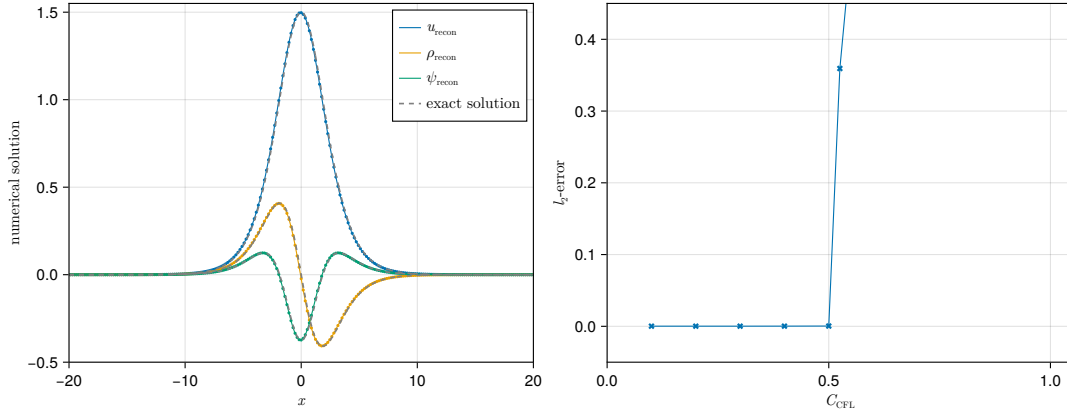


Figure 45: The results of Experiment 2.2 for the optimised IMEX Active Flux scheme with the ARS(2, 3, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

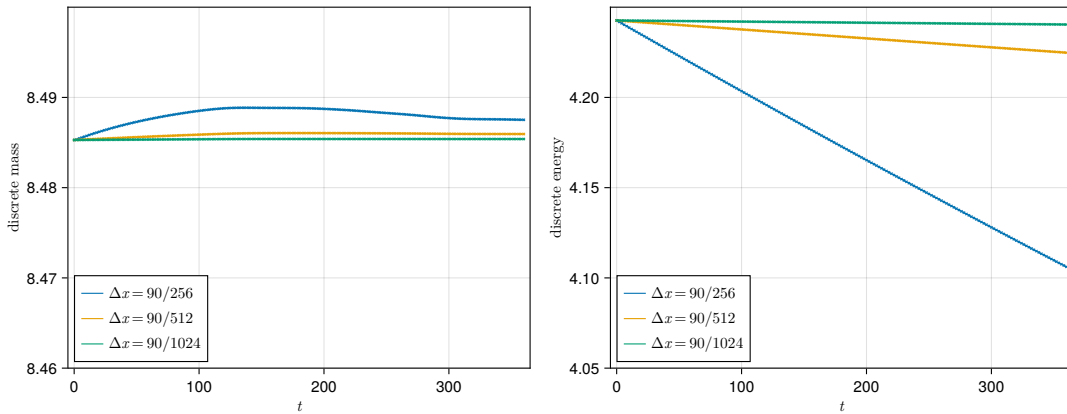


Figure 46: The results of Experiment 2.3 for the optimised IMEX Active Flux scheme with the ARS(2, 3, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy E_{dis} with respect to time t .

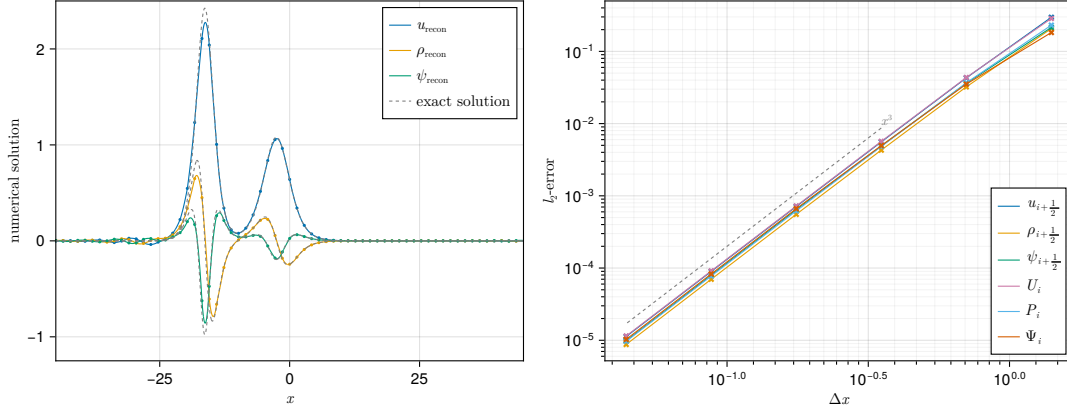


Figure 47: The results of Experiment 3.1 for the optimised IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

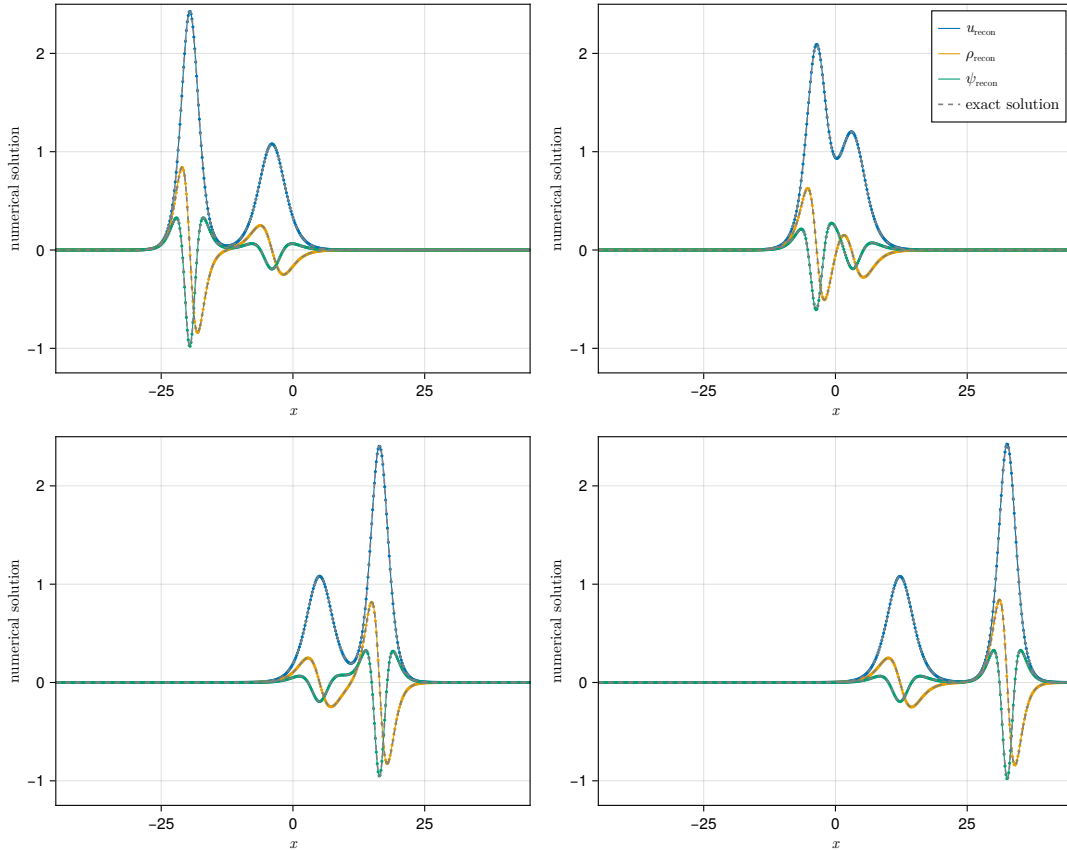


Figure 48: The results of Experiment 3.2 for the optimised IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. From top left to bottom right: The numerical solution computed on grid with 512 cells at $t = 0, 20, 40,$ and 60 in comparison to the exact solution.

4.3 Central IMEX Active Flux scheme

The general idea of the simplified and the optimised scheme is to take the positive and negative part of the Jacobian $F_2^\pm$ associated the classical IMEX Active Flux scheme and modify it by eliminating and rescaling certain terms. In [51], an alternative way to split a Jacobian into a positive and negative part is discussed:

$$F_{2,\text{central}}^\pm = \frac{1}{2}F_2.$$

If we split F_2 using this approach, we obtain

$$F_{2,\text{central}}^+ = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & -\frac{1}{2\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & 0 \end{pmatrix}, \quad F_{2,\text{central}}^- = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & -\frac{1}{2\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & 0 \end{pmatrix}.$$

Remark 4.3. The matrices $F_{2,\text{central}}^\pm$ have no elements that result in terms of the form

$$\frac{1}{2} \left(D^+ v_{i+\frac{1}{2}}(t) - D^- v_{i+\frac{1}{2}}(t) \right)$$

in the update equations of the point values.

The point value update equations using $F_1^\pm$ and $F_{2,\text{central}}^\pm$ are

$$\begin{aligned} \frac{d}{dt} u_{i+\frac{1}{2}}(t) &= -\max(u_{i+\frac{1}{2}}(t), 0) D^+ u_{i+\frac{1}{2}}(t) - \min(u_{i+\frac{1}{2}}(t), 0) D^- u_{i+\frac{1}{2}}(t) \\ &\quad - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}(t) + D^- \psi_{i+\frac{1}{2}}(t) \right), \\ \frac{d}{dt} \rho_{i+\frac{1}{2}}(t) &= \frac{1}{2\varepsilon^2} \left(D^+ \rho_{i+\frac{1}{2}}(t) + D^- \rho_{i+\frac{1}{2}}(t) \right) - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}(t), \\ \frac{d}{dt} \psi_{i+\frac{1}{2}}(t) &= -\frac{1}{2\varepsilon^2} \left(D^+ u_{i+\frac{1}{2}}(t) + D^- u_{i+\frac{1}{2}}(t) \right) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}(t). \end{aligned}$$

An IMEX Active Flux scheme with these update equations for the point values and the usual update equations for the cell averages is in this text called the central IMEX Active Flux scheme.

4.3.1 Formal asymptotic analysis

To analyse the asymptotic preserving property of the central scheme a formal asymptotic analysis is conducted by following, again, the approach of the analysis for classical scheme. Ordering the terms by the power of ε , after inserting the power series into the equations, the terms of factor ε^{-2} satisfy yield

$$\begin{aligned} 0 &= \Delta t A \left(\frac{1}{2} \left(D^+ \rho_{i+\frac{1}{2}}^{(0)[j]} + D^- \rho_{i+\frac{1}{2}}^{(0)[j]} \right) - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(\frac{1}{2} \left(D^+ \rho_{i+\frac{1}{2}}^{(0)[j]} + D^- \rho_{i+\frac{1}{2}}^{(0)[j]} \right) - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \end{aligned}$$

the terms of factor ε^{-1} the yield

$$0 = \Delta t A \left(\frac{1}{2} \left(D^+ \rho_{i+\frac{1}{2}}^{(1)[j]} + D^- \rho_{i+\frac{1}{2}}^{(1)[j]} \right) - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s},$$

$$\begin{aligned}
 0 &= \Delta t b \cdot \left(\frac{1}{2} \left(D^+ \rho_{i+\frac{1}{2}}^{(1)[j]} + D^- \rho_{i+\frac{1}{2}}^{(1)[j]} \right) - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1, \dots, s}, \\
 0 &= \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1, \dots, s}, \\
 0 &= \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(1)[j]} + D^- u_{i+\frac{1}{2}}^{(1)[j]} \right) + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1, \dots, s},
 \end{aligned}$$

and the terms of factor ε^0 yield

$$\begin{aligned}
 u_{i+\frac{1}{2}}^{[k](0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\max \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \\
 &\quad \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \\
 &\quad + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \right), \\
 u_{i+\frac{1}{2}}^{n+1(0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\max \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \\
 &\quad \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \\
 &\quad + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \right)
 \end{aligned} \tag{4.13}$$

as well as

$$\begin{aligned}
 \bar{\rho}_{i+\frac{1}{2}}^{(0)} &= \rho_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \Delta t A \left(\frac{1}{2} \left(D^+ \rho_{i+\frac{1}{2}}^{(2)[j]} + D^- \rho_{i+\frac{1}{2}}^{(2)[j]} \right) - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}, \\
 \rho_{i+\frac{1}{2}}^{(0)n+1} &= \rho_{i+\frac{1}{2}}^{(0)n} + \Delta t b \cdot \left(\frac{1}{2} \left(D^+ \rho_{i+\frac{1}{2}}^{(2)[j]} + D^- \rho_{i+\frac{1}{2}}^{(2)[j]} \right) - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}, \\
 \bar{\psi}_{i+\frac{1}{2}}^{(0)} &= \psi_{i+\frac{1}{2}}^{(0)n} \mathbb{I}_s + \Delta t A \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}, \\
 \psi_{i+\frac{1}{2}}^{(0)n+1} &= \psi_{i+\frac{1}{2}}^{(0)n} + \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u_{i+\frac{1}{2}}^{(2)[j]} + D^- u_{i+\frac{1}{2}}^{(2)[j]} \right) + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}.
 \end{aligned}$$

This implies

$$\psi_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2} \left(D^+ \rho_{i+\frac{1}{2}}^{(0)[j]} + D^- \rho_{i+\frac{1}{2}}^{(0)[j]} \right) \tag{4.14}$$

and (3.7) for all $j = 1, \dots, s$. In addition, since the usual update equations for the cell averages are used, (3.2), (3.3) and (3.4) hold. The results of Lemma 3.5 do not hold, because (3.6) is not satisfied. However, with (4.14), a similar condition is satisfied and, hence, we can prove the next lemma.

Lemma 4.4. *The leading order terms, i.e. the terms of order 0, satisfy*

$$\begin{aligned}
 \psi_{i+\frac{1}{2}}^{(0)[j]} &= \frac{1}{\Delta x^2} \left(u_{i-\frac{3}{2}}^{(0)[j]} - 3U_{i-1}^{(0)[j]} + 3u_{i-\frac{1}{2}}^{(0)[j]} + 3U_i^{(0)[j]} - 8u_{i+\frac{1}{2}}^{(0)[j]} \right. \\
 &\quad \left. + 3U_{i+1}^{(0)[j]} + 3u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right)
 \end{aligned}$$

and

$$\begin{aligned} \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) &= \frac{1}{\Delta x^3} \left(u_{i-\frac{5}{2}}^{(0)[j]} - 3U_{i-2}^{(0)[j]} + 6u_{i-\frac{3}{2}}^{(0)[j]} - 6U_{i-1}^{(0)[j]} - 15u_{i-\frac{1}{2}}^{(0)[j]} + 33U_i^{(0)[j]} \right. \\ &\quad \left. - 33U_{i+1}^{(0)[j]} + 15u_{i+\frac{3}{2}}^{(0)[j]} + 6U_{i+2}^{(0)[j]} - 6u_{i+\frac{5}{2}}^{(0)[j]} + 3U_{i+3}^{(0)[j]} - u_{i-\frac{7}{2}}^{(0)[j]} \right) \end{aligned}$$

for $j = 1, \dots, s$.

Proof. To derive the formulas we use (4.14), as well as (3.3), (3.4), and (3.7). With the help of these equations, we obtain

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{(0)[j]} &= \frac{1}{\Delta x} \left(\rho_{i-\frac{1}{2}}^{(0)[j]} - 3P_i^{(0)[j]} + 3P_{i+1}^{(0)[j]} - \rho_{i+\frac{3}{2}}^{(0)[j]} \right) \\ &= \frac{1}{\Delta x} \left(\frac{1}{\Delta x} \left(u_{i-\frac{3}{2}}^{(0)[j]} - 3U_{i-1}^{(0)[j]} + 3U_i^{(0)[j]} - u_{i+\frac{1}{2}}^{(0)[j]} \right) - \frac{3}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) \right. \\ &\quad \left. + \frac{3}{\Delta x} \left(u_{i+\frac{3}{2}}^{(0)[j]} - u_{i+\frac{1}{2}}^{(0)[j]} \right) - \frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + 3U_{i+2}^{(0)[j]} - u_{i+\frac{5}{2}}^{(0)[j]} \right) \right) \\ &= \frac{1}{\Delta x^2} \left(u_{i-\frac{3}{2}}^{(0)[j]} - 3U_{i-1}^{(0)[j]} + 3u_{i-\frac{1}{2}}^{(0)[j]} + 3U_i^{(0)[j]} - 8u_{i+\frac{1}{2}}^{(0)[j]} \right. \\ &\quad \left. + 3U_{i+1}^{(0)[j]} + 3u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) &= \frac{1}{\Delta x} \left(\psi_{i-\frac{1}{2}}^{(0)[j]} - 3\Psi_i^{(0)[j]} + 3\Psi_{i+1}^{(0)[j]} - \psi_{i+\frac{3}{2}}^{(0)[j]} \right) \\ &= \frac{1}{\Delta x} \left(\frac{1}{\Delta x^2} \left(u_{i-\frac{5}{2}}^{(0)[j]} - 3U_{i-2}^{(0)[j]} + 3u_{i-\frac{3}{2}}^{(0)[j]} + 3U_{i-1}^{(0)[j]} - 8u_{i-\frac{1}{2}}^{(0)[j]} \right. \right. \\ &\quad \left. \left. + 3U_i^{(0)[j]} + 3u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + u_{i+\frac{3}{2}}^{(0)[j]} \right) \right. \\ &\quad \left. - \frac{3}{\Delta x^2} \left(\left(u_{i-\frac{1}{2}}^{(0)[j]} - 3U_i^{(0)[j]} + 3U_{i+1}^{(0)[j]} - u_{i+\frac{3}{2}}^{(0)[j]} \right) \right. \right. \\ &\quad \left. \left. - \left(u_{i-\frac{3}{2}}^{(0)[j]} - 3U_{i-1}^{(0)[j]} + 3U_i^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) \right) \right. \\ &\quad \left. + \frac{3}{\Delta x} \left(\left(u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + 3U_{i+2}^{(0)[j]} - u_{i+\frac{5}{2}}^{(0)[j]} \right) \right. \right. \\ &\quad \left. \left. - \left(u_{i-\frac{1}{2}}^{(0)[j]} - 3U_i^{(0)[j]} + 3U_{i+1}^{(0)[j]} - u_{i-\frac{3}{2}}^{(0)[j]} \right) \right) \right. \\ &\quad \left. - \frac{1}{\Delta x^2} \left(u_{i-\frac{1}{2}}^{(0)[j]} - 3U_i^{(0)[j]} + 3u_{i+\frac{1}{2}}^{(0)[j]} + 3U_{i+1}^{(0)[j]} - 8u_{i+\frac{3}{2}}^{(0)[j]} \right. \right. \\ &\quad \left. \left. + 3U_{i+2}^{(0)[j]} + 3u_{i+\frac{5}{2}}^{(0)[j]} - 3U_{i+3}^{(0)[j]} + u_{i+\frac{7}{2}}^{(0)[j]} \right) \right) \\ &= \frac{1}{\Delta x^3} \left(u_{i-\frac{5}{2}}^{(0)[j]} - 3U_{i-2}^{(0)[j]} + 6u_{i-\frac{3}{2}}^{(0)[j]} - 6U_{i-1}^{(0)[j]} - 15u_{i-\frac{1}{2}}^{(0)[j]} + 33U_i^{(0)[j]} \right. \\ &\quad \left. - 33U_{i+1}^{(0)[j]} + 15u_{i+\frac{3}{2}}^{(0)[j]} + 6U_{i+2}^{(0)[j]} - 6u_{i+\frac{5}{2}}^{(0)[j]} + 3U_{i+3}^{(0)[j]} - u_{i-\frac{7}{2}}^{(0)[j]} \right). \end{aligned}$$

□

Hence,

$$\begin{aligned}
 \frac{d}{dt}u_{i+\frac{1}{2}}^{(0)}(t) = & - \left(\max \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t) \right) \right. \\
 & + \min \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t) \right) \Bigg) \\
 & - \frac{1}{\Delta x^3} \left(u_{i-\frac{5}{2}}^{(0)}(t) - 3U_{i-2}^{(0)}(t) + 6u_{i-\frac{3}{2}}^{(0)}(t) - 6U_{i-1}^{(0)}(t) \right. \\
 & - 15u_{i-\frac{1}{2}}^{(0)}(t) + 33U_i^{(0)}(t) - 33U_{i+1}^{(0)}(t) + 15u_{i+\frac{3}{2}}^{(0)}(t) \\
 & \left. + 6U_{i-1}^{(0)}(t) - 6u_{i+\frac{5}{2}}^{(0)}(t) + 3U_{i+3}^{(0)}(t) - u_{i-\frac{7}{2}}^{(0)}(t) \right)
 \end{aligned} \tag{4.15}$$

is a time discretisation using a type ARS IMEX Runge-Kutta scheme of (4.13) and (3.2) is a time discretisation of

$$\begin{aligned}
 \frac{d}{dt}U_i^{(0)}(t) = & - \frac{1}{2\Delta x} \left(\left(u_{i+\frac{1}{2}}^{(0)}(t) \right)^2 - \left(u_{i-\frac{1}{2}}^{(0)}(t) \right)^2 \right) - \frac{1}{\Delta x^3} \left(\left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) \right. \right. \\
 & + 3u_{i-\frac{1}{2}}^{(0)}(t) + 3U_i^{(0)}(t) - 8u_{i+\frac{1}{2}}^{(0)}(t) + 3U_{i+1}^{(0)}(t) + 3u_{i+\frac{3}{2}}^{(0)}(t) \\
 & - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \Bigg) - \left(u_{i-\frac{5}{2}}^{(0)}(t) - 3U_{i-2}^{(0)}(t) + 3u_{i-\frac{3}{2}}^{(0)}(t) \right. \\
 & \left. \left. + 3U_{i-1}^{(0)}(t) - 8u_{i-\frac{1}{2}}^{(0)}(t) + 3U_i^{(0)}(t) + 3u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + u_{i+\frac{3}{2}}^{(0)}(t) \right) \right).
 \end{aligned} \tag{4.16}$$

Suppose the IMEX Runge-Kutta scheme is globally stiffly accurate. Since (3.3), (3.4), and (3.7) are satisfied, (3.10), (3.12), and (3.13) follow in the same way as for the classical scheme. Using Lemma 4.4, (4.14) implies,

$$\begin{aligned}
 \psi_{i+\frac{1}{2}}^{(0)n+1} = & \frac{1}{\Delta x^2} \left(u_{i-\frac{3}{2}}^{(0)n+1} - 3U_{i-1}^{(0)n+1} + 3u_{i-\frac{1}{2}}^{(0)n+1} + 3U_i^{(0)n+1} - 8u_{i+\frac{1}{2}}^{(0)n+1} \right. \\
 & \left. + 3U_{i+1}^{(0)n+1} + 3u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)n+1} + u_{i+\frac{5}{2}}^{(0)n+1} \right).
 \end{aligned} \tag{4.17}$$

Therefore (4.13) and (3.2) together with (3.10), (3.12), (3.13), and (4.17) are a discretisation of (4.15), (4.16), (3.14), (3.16), (3.17), and

$$\begin{aligned}
 \psi_{i+\frac{1}{2}}^{(0)}(t) = & \frac{1}{\Delta x^2} \left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) + 3u_{i-\frac{1}{2}}^{(0)}(t) + 3U_i^{(0)}(t) - 8u_{i+\frac{1}{2}}^{(0)}(t) \right. \\
 & \left. + 3U_{i+1}^{(0)}(t) + 3u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right).
 \end{aligned} \tag{4.18}$$

Theorem 4.5. *Equation (4.15) is a time-continuous semi-discretisation of*

$$\frac{\partial}{\partial t}u^{(0)}(t, x) + u^{(0)}(t, x)\frac{\partial}{\partial x}u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3}u^{(0)}(t, x) = 0$$

with an error in $\mathcal{O}(\Delta x^2)$, (4.16) is a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t}u^{(0)}(t, x) + u^{(0)}(t, x)\frac{\partial}{\partial x}u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3}u^{(0)}(t, x)dx = 0$$

with an error on $\mathcal{O}(\Delta x^2)$, and (4.18) is a discretisation of

$$\psi^{(0)}(t, x) = \frac{\partial^2}{\partial x^2}u^{(0)}(t, x)$$

with an error in $\mathcal{O}(\Delta x^2)$.

Proof. Assume that $u^{(0)} : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$ is a smooth and bounded function with bounded derivatives that interpolates the point values and cell averages. This implies

$$\begin{aligned} \frac{1}{\Delta x^2} & \left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) + 3u_{i-\frac{1}{2}}^{(0)}(t) + 3U_i^{(0)}(t) - 8u_{i+\frac{1}{2}}^{(0)}(t) + 3U_{i+1}^{(0)}(t) \right. \\ & \left. + 3u_{i+\frac{3}{2}}^{(0)}(t) - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right) = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) + \frac{\Delta x^2}{12} \frac{\partial^4}{\partial x^4} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^4) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{\Delta x^3} & \left(u_{i-\frac{5}{2}}^{(0)}(t) - 3U_{i-2}^{(0)}(t) + 6u_{i-\frac{3}{2}}^{(0)}(t) - 6U_{i-1}^{(0)}(t) - 15u_{i-\frac{1}{2}}^{(0)}(t) + 33U_i^{(0)}(t) \right. \\ & \left. - 33U_{i+1}^{(0)}(t) + 15u_{i+\frac{3}{2}}^{(0)}(t) + 6U_{i+2}^{(0)}(t) - 6u_{i+\frac{5}{2}}^{(0)}(t) + 3U_{i+3}^{(0)}(t) - u_{i+\frac{7}{2}}^{(0)}(t) \right) \\ & = \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\Delta x^2}{2} \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^4). \end{aligned}$$

Proceeding as in the proof of Theorem 3.6, we obtain that (4.15) is a discretisation of

$$\frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) = 0$$

with an error in $\mathcal{O}(\Delta x^2)$. Moreover, we have

$$\begin{aligned} \frac{1}{\Delta x^3} & \left(\left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) + 3u_{i-\frac{1}{2}}^{(0)}(t) + 3U_i^{(0)}(t) - 8u_{i+\frac{1}{2}}^{(0)}(t) + 3U_{i+1}^{(0)}(t) \right. \right. \\ & \left. + 3u_{i+\frac{3}{2}}^{(0)}(t) - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right) - \left(u_{i-\frac{5}{2}}^{(0)}(t) - 3U_{i-2}^{(0)}(t) + 3u_{i-\frac{3}{2}}^{(0)}(t) \right. \\ & \left. + 3U_{i-1}^{(0)}(t) - 8u_{i-\frac{1}{2}}^{(0)}(t) + 3U_i^{(0)}(t) + 3u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + u_{i+\frac{3}{2}}^{(0)}(t) \right) \Big) \\ & = \frac{1}{\Delta x} \left(\frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) \\ & + \frac{\Delta x}{12} \left(\frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) + \mathcal{O}(\Delta x^3) \\ & = \frac{1}{\Delta x} \int_{C_i} u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) dx - \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) dx + \mathcal{O}(\Delta x^2). \end{aligned}$$

In the last equality we used the Lipschitz continuity of the derivatives of $u^{(0)}$. An immediate result of the last equation is that (4.16) is a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) dx = 0$$

with an error on $\mathcal{O}(\Delta x^2)$. The equation at the beginning of the proof, additionally, implies that (4.18) is a spatial discretisation of

$$\psi^{(0)}(t, x) = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x)$$

with an error in $\mathcal{O}(\Delta x^2)$. □

Remark 4.6. Combining the result of the previous theorem with the ones of Theorem 3.6, one can observe that all discretisations have an error in at least $\mathcal{O}(\Delta x^2)$. In contrast, some the discretisations that appeared in the formal analysis of the classical, simplified, and optimised scheme have an error in $\mathcal{O}(\Delta x)$.

To sum up the analysis: The formal asymptotic analysis indicates that the central IMEX Active Flux scheme for the HKdV equations discretised in time with a type ARS IMEX Runge-Kutta scheme is asymptotic preserving for u component. In addition, it is asymptotic preserving, if the IMEX Runge-Kutta scheme is globally stiffly accurate.

4.3.2 Implementation

As for the other scheme, the only difference of the implementation of the central scheme compared to the implementation of the classical scheme are the matrices M_1 , M_2 , and M_3 . In case of the central scheme, they are

$$M_1 = \begin{pmatrix} 0 & 0 & -\frac{1}{\Delta x} & 0 & 0 & \frac{3}{\Delta x} \\ 0 & \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & -\frac{3}{\Delta x \varepsilon^2} & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & \frac{3}{\Delta x \varepsilon^2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -\frac{3}{\Delta x} \\ 0 & 0 & -\frac{1}{\varepsilon^2} & 0 & \frac{3}{\Delta x \varepsilon^2} & 0 \\ 0 & \frac{1}{\varepsilon^2} & 0 & -\frac{3}{\Delta x \varepsilon^2} & 0 & 0 \\ 0 & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & -\frac{1}{\varepsilon^2} \\ \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & \frac{1}{\varepsilon^2} & 0 \end{pmatrix},$$

and,

$$M_3 = \begin{pmatrix} 0 & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

4.3.3 Numerical results

The result of Experiment 1 performed for the central IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme can be seen in Figure 49 - 51. For all parts of the experiment $C_{\text{CFL}} = 0.4$ is chosen. The results of Experiment 1.1 are depicted in Figure 49. The numerical solution computed with the central scheme matches the exact solution on a coarse grid with 64 cell at $t = 4$, has maximum l_2 -error of approximately 0.0316, and the scheme achieves roughly second order accuracy. The numerical solution computed on a grid with 512 cells at $t = 180$, which is shown in Figure 50, is visually indistinguishable from the exact solution. The l_2 -error of the numerical solution is approximately 0.0096. Furthermore, the central IMEX Active Flux scheme combined with the ARS(4, 4, 3) scheme seems to be stable up to $C_{\text{CFL}} = 0.45$ if the numerical solution is computed on a grid with 512 cells and the error is evaluated at $t = 4$. In Figure 51, the discrete mass and energy with respect to time t are shown. At $t = 360$, the discrete mass of the numerical solution is about 99.86% (256 cells), 99.96% (512 cells), and 99.99% (1024 cells) of the original discrete mass at $t = 0$, and the discrete energy at $t = 360$ is approximately 99.22% (256 cells), 99.91% (512 cells), and 99.99% (1024 cells) of the original discrete energy.

Figure 52 - Figure 55 depict the results of Experiment 2, where $C_{\text{CFL}} = 0.4$ is chosen. First, in Figure 52, where the results of Experiment 2.1 are shown, it can be observed that the l_2 -error of numerical solutions computed on a fixed grid becomes constants as the relaxation parameter tends to zero. This is the case for all point values and cell averages. For $\varepsilon = 10^{-7}$, the achieved order of accuracy of the central scheme is roughly two and the numerical solution on a coarse grid has a maximum l_2 -error of approximately 0.0759 (see Figure 53). The results of Experiment 2.2 (see Figure 54) indicate the central scheme becomes more dispersive in the limit $\varepsilon \rightarrow 0$ since the numerical solution at $t = 180$ does not match the exact solution at this time. The maximum l_2 -error of the numerical solution is approximately 0.3024. In addition, the central IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme appears to be stable up to $C_{\text{CFL}} = 0.45$. The results of Experiment 2.3 can be seen in Figure 55. It is noticeable that especially the discrete mass with respect to time of the numerical solution computed on a grid with 256 cells is not a straight line. At $t = 360$ the discrete mass is approximately 100.02% (256 cells), 99.97% (512 cells), and 99.99% (1024 cells) of the discrete mass at $t = 0$. The discrete energy is reasonably conserved since at $t = 360$ it is 89.64% (256 cells), 97.58% (512 cells), and 99.58% (1024 cells) of the discrete energy at $t = 0$.

To analyse if the central IMEX Active Flux scheme is asymptotic accurate, the experimental order of convergence is computed from the results of Experiment 1.1 and Experiment 2.1. It is shown, rounded to four decimal places, in Table 2. The experimental order of convergence for $512 \rightarrow 1024$ associated to Experiment 1.1 is not for all components the same as the one associated to Experiment 2.2. For example, the one associated to $u_{i+\frac{1}{2}}$ is more than 0.25 higher in Experiment 1.1 than in Experiment 2.1. However, for $1024 \rightarrow 2048$ cells, the difference is smaller than 0.1 for all point values and cells averages. Thus, the results cautiously suggest that the scheme is asymptotic accurate.

	Experiment 1.1		Experiment 2.1	
	512 $\rightarrow$ 1024	1024 $\rightarrow$ 2048	512 $\rightarrow$ 1024	1024 $\rightarrow$ 2048
$u_{i+\frac{1}{2}}$	2.0154	1.9648	1.7563	1.8920
$\rho_{i+\frac{1}{2}}$	1.9351	1.9257	1.6596	1.8354
$\psi_{i+\frac{1}{2}}$	1.8255	1.8730	1.8058	1.9362
U_i	2.0113	1.9729	1.8011	1.9089
P_i	1.9568	1.9380	1.7578	1.8888
Ψ_i	1.8933	1.8684	1.6374	1.8099

Table 2: The experimental order of convergence computed from the results of Experiment 1.1 and Experiment 2.1 performed for the central IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme.

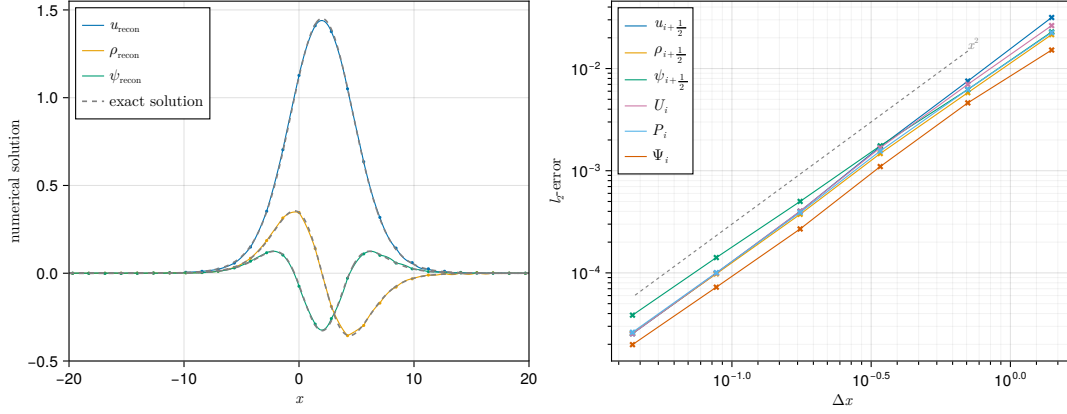


Figure 49: The results of Experiment 1.1 for the central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

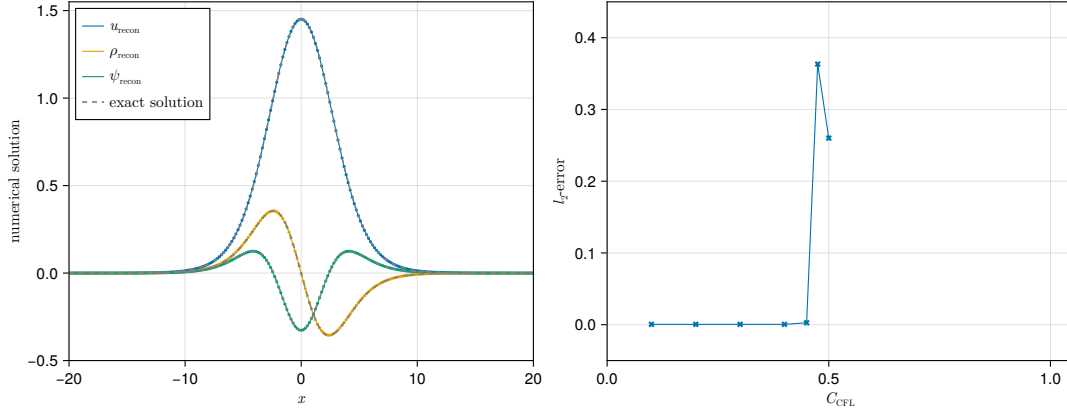


Figure 50: The results of Experiment 1.2 for the central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

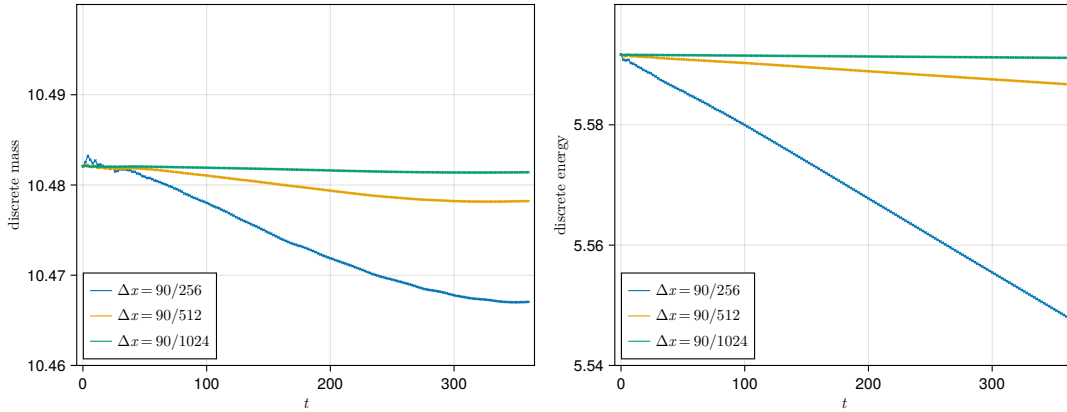


Figure 51: The results of Experiment 1.3 for the central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy $E_{\text{H,dis}}$ with respect to time t .

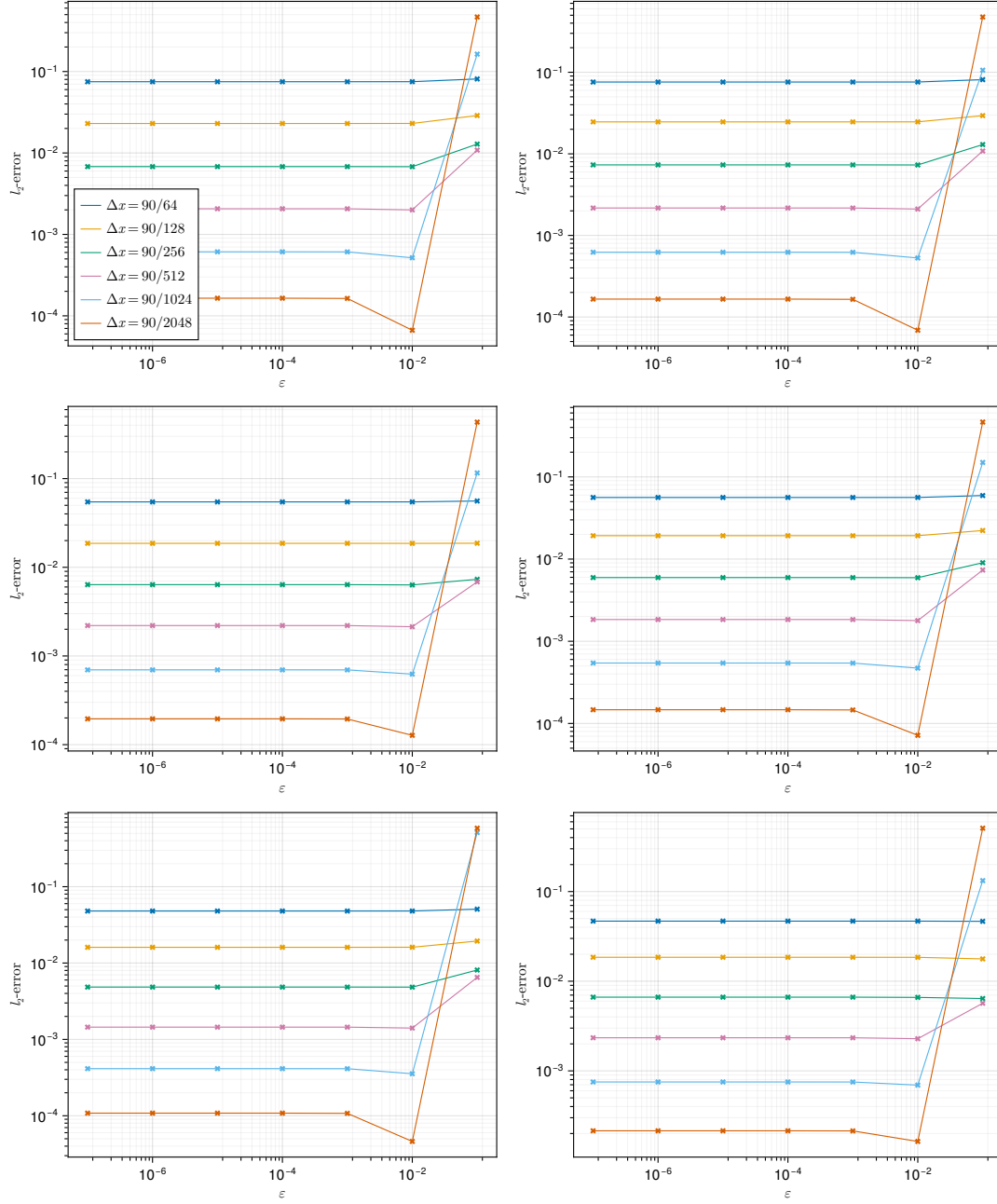


Figure 52: The results of Experiment 2.1 (first part) for the central IMEX Active Flux scheme with the ARS(4,4,3) scheme as time discretisation. From to left to bottom right: the l_2 -error of $u_{i+\frac{1}{2}}$, U_i , $\rho_{i+\frac{1}{2}}$, P_i , $\psi_{i+\frac{1}{2}}$, and Ψ_i with respect to ε for different spatial step sizes.

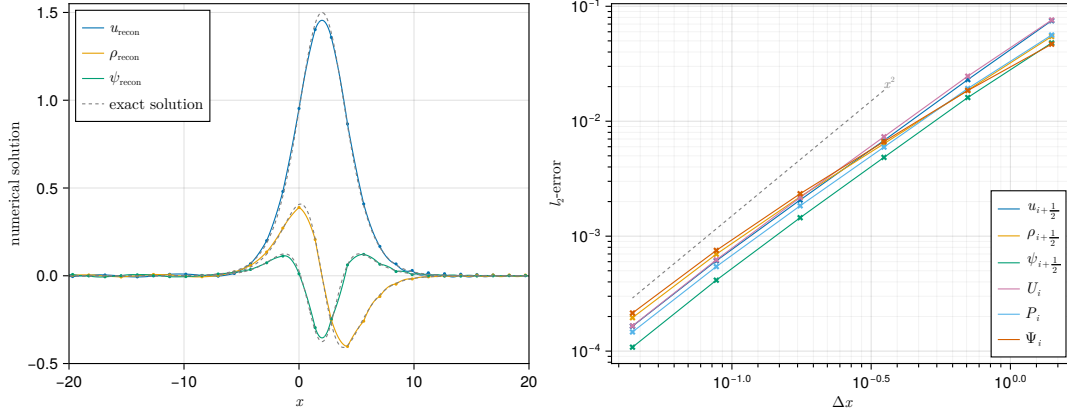


Figure 53: The results of Experiment 2.1 (second part) for the central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

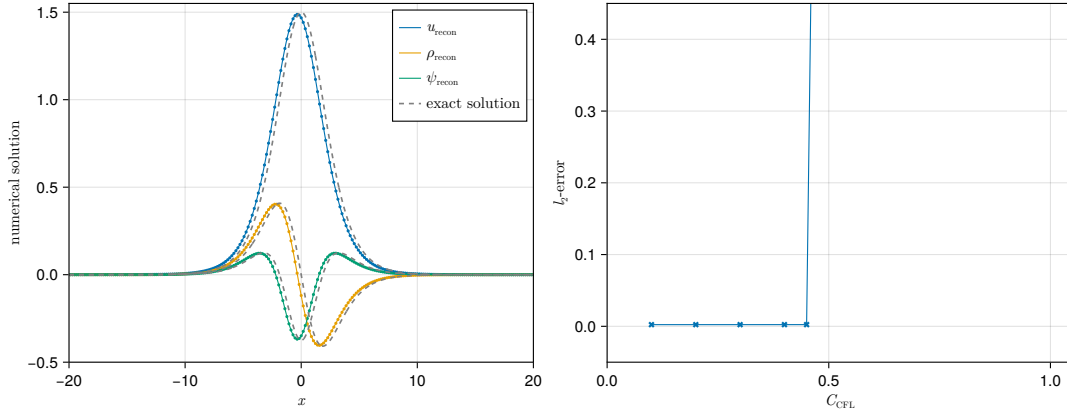


Figure 54: The results of Experiment 2.2 for the central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

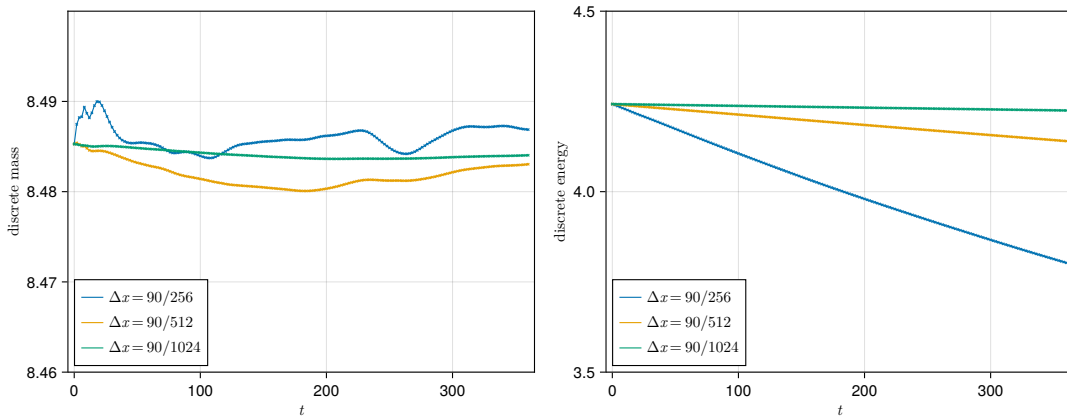


Figure 55: The results of Experiment 2.3 for the central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy E_{dis} with respect to time t .

4.4 Modified central IMEX Active Flux scheme

The central IMEX Active Flux scheme achieves second order accuracy and ideally one wants to obtain a scheme that is third order accurate since the classical IMEX Active Flux scheme achieves this order of accuracy for large ε . It is possible to increase the order of accuracy of the central IMEX Active Flux scheme by considering alternatives to $D^\pm v_{i+\frac{1}{2}}(t)$. Based on [52] and with the help of Mathematica [37], the following alternatives are considered in this text:

$$D_{\text{mod}}^+ v_{i+\frac{1}{2}}(t) = \frac{1}{6\Delta x} \left(5v_{i-\frac{1}{2}}(t) - 18V_i(t) + 8v_{i+\frac{1}{2}}(t) + 6V_{i+1}(t) - v_{i+\frac{3}{2}}(t) \right)$$

and

$$D_{\text{mod}}^- v_{i+\frac{1}{2}}(t) = \frac{1}{6\Delta x} \left(v_{i-\frac{1}{2}}(t) - 6V_i(t) - 8v_{i+\frac{1}{2}}(t) + 18V_{i+1}(t) - 5v_{i+\frac{3}{2}}(t) \right).$$

If one assumes $v : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$ is a smooth and bounded function with bounded derivatives that interpolates the point values and cell average $v_{i+\frac{1}{2}}(t)$ and $V_i(t)$, it follows

$$D_{\text{mod}}^+ v_{i+\frac{1}{2}}(t) = \frac{\partial}{\partial x} v(t, x) + \frac{\Delta x^3}{90} \frac{\partial^4}{\partial x^4} v(t, x) + \mathcal{O}(\Delta x^4)$$

and

$$D_{\text{mod}}^- v_{i+\frac{1}{2}}(t) = \frac{\partial}{\partial x} v(t, x) - \frac{\Delta x^3}{90} \frac{\partial^4}{\partial x^4} v(t, x) + \mathcal{O}(\Delta x^4).$$

This implies

$$\begin{aligned} \frac{1}{2} \left(D_{\text{mod}}^+ v_{i+\frac{1}{2}}(t) + D_{\text{mod}}^- v_{i+\frac{1}{2}}(t) \right) &= \frac{1}{2\Delta x} \left(v_{i-\frac{1}{2}}(t) - 4V_i(t) + 4V_{i+1}(t) - v_{i+\frac{3}{2}}(t) \right) \\ &= \frac{\partial}{\partial x} v(t, x) + \mathcal{O}(\Delta x^4). \end{aligned}$$

The general idea to increase the order of accuracy of the central IMEX Active Flux scheme is to replace $D^\pm$ by $D_{\text{mod}}^\pm$ in the part that is discretised in time implicitly. Following this approach, the update equations for the point values are

$$\begin{aligned} \frac{d}{dt} u_{i+\frac{1}{2}}(t) &= -\max(u_{i+\frac{1}{2}}(t), 0) D_{\text{mod}}^+ u_{i+\frac{1}{2}}(t) - \min(u_{i+\frac{1}{2}}(t), 0) D_{\text{mod}}^- u_{i+\frac{1}{2}}(t) \\ &\quad - \frac{1}{2} \left(D_{\text{mod}}^+ \psi_{i+\frac{1}{2}}(t) + D_{\text{mod}}^- \psi_{i+\frac{1}{2}}(t) \right), \\ \frac{d}{dt} \rho_{i+\frac{1}{2}}(t) &= \frac{1}{2\varepsilon^2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}(t) + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}(t) \right) - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}(t), \\ \frac{d}{dt} \psi_{i+\frac{1}{2}}(t) &= -\frac{1}{2\varepsilon^2} \left(D_{\text{mod}}^+ u_{i+\frac{1}{2}}(t) + D_{\text{mod}}^- u_{i+\frac{1}{2}}(t) \right) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}(t). \end{aligned}$$

As before, the update equations of the cell averages are not altered and a scheme with those update equations is henceforth called the modified central IMEX Active Flux scheme.

4.4.1 Formal asymptotic analysis

Since the general structure of the scheme is not modified, the formal asymptotic analysis is in many parts identical to the one for the central IMEX Active Flux scheme discretised in time with an IMEX Runge-Kutta scheme of type ARS. Especially, after ordering the terms by the power of ε , we obtain the same equations at the different order with the only difference that in the implicit part $D^\pm$ is replaced by $D_{\text{mod}}^\pm$. Thus, we do not state all of them here but refer to

the formal asymptotic analysis of the central IMEX Active Flux scheme. The counterparts of (4.13), (4.14), and (3.7) are

$$\begin{aligned}
 u_{i+\frac{1}{2}}^{[k](0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\max \left((u_{i+\frac{1}{2}}^{(0)[j-1]}, 0) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \right. \\
 &\quad \left. \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) \right) \right. \\
 &\quad \left. + \Delta t \sum_{j=1}^k a_{k,j} \left(-\frac{1}{2} \left(D_{\text{mod}}^+ \psi_{i+\frac{1}{2}}^{[j](0)} + D_{\text{mod}}^- \psi_{i+\frac{1}{2}}^{[j](0)} \right) \right), \right. \\
 u_{i+\frac{1}{2}}^{n+1(0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\max \left(u_{i+\frac{1}{2}}^{[j-1](0)}, 0) D^+ u_{i+\frac{1}{2}}^{[j-1](0)} \right. \right. \\
 &\quad \left. \left. - \min \left(u_{i+\frac{1}{2}}^{[j-1](0)}, 0) D^- u_{i+\frac{1}{2}}^{[j-1](0)} \right) \right) \right. \\
 &\quad \left. + \Delta t \sum_{j=1}^s b_j \left(-\frac{1}{2} \left(D_{\text{mod}}^+ \psi_{i+\frac{1}{2}}^{[j](0)} + D_{\text{mod}}^- \psi_{i+\frac{1}{2}}^{[j](0)} \right) \right). \right.
 \end{aligned} \tag{4.19}$$

as well as

$$\rho_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2} \left(D_{\text{mod}}^+ u_{i+\frac{1}{2}}^{(0)[j]} + D_{\text{mod}}^- u_{i+\frac{1}{2}}^{(0)[j]} \right) \tag{4.20}$$

and

$$\psi_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(0)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(0)[j]} \right). \tag{4.21}$$

Lemma 4.7. *The leading order terms, i.e. the terms of order 0, satisfy*

$$\psi_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{4\Delta x^2} \left(u_{i-\frac{3}{2}}^{(0)} - 4U_{i-1}^{(0)} + 8u_{i-\frac{1}{2}}^{(0)} + 4U_i^{(0)} - 18u_{i+\frac{1}{2}}^{(0)} + 4U_{i+1}^{(0)} + 8u_{i+\frac{3}{2}}^{(0)} - 4U_{i+2}^{(0)} + u_{i+\frac{5}{2}}^{(0)} \right),$$

and

$$\begin{aligned}
 \frac{1}{2} \left(D_{\text{mod}}^+ \psi_{i+\frac{1}{2}}^{(0)[j]} + D_{\text{mod}}^- \psi_{i+\frac{1}{2}}^{(0)[j]} \right) &= \frac{1}{8\Delta x^3} \left(u_{i-\frac{5}{2}}^{(0)} - 4U_{i-2}^{(0)} + 16u_{i-\frac{3}{2}}^{(0)} - 28U_{i-1}^{(0)} - 35u_{i-\frac{1}{2}}^{(0)} + 104U_i^{(0)} \right. \\
 &\quad \left. - 104U_{i+1}^{(0)} + 35u_{i+\frac{3}{2}}^{(0)} + 28U_{i+2}^{(0)} - 16u_{i+\frac{5}{2}}^{(0)} + 4U_{i+3}^{(0)} - u_{i+\frac{7}{2}}^{(0)} \right)
 \end{aligned}$$

for $j = 1, \dots, s$.

Proof. Equations (4.20) and (4.20) imply

$$\begin{aligned}
 \psi_{i+\frac{1}{2}}^{(0)} &= \frac{1}{2\Delta x} \left(\rho_{i-\frac{1}{2}}^{(0)} - 4P_i^{(0)} + 4P_{i+1}^{(0)} - \rho_{i+\frac{3}{2}}^{(0)} \right) \\
 &= \frac{1}{2\Delta x} \left(\frac{1}{2\Delta x} \left(u_{i-\frac{3}{2}}^{(0)} - 4U_{i-1}^{(0)} + 4U_i^{(0)} - u_{i+\frac{1}{2}}^{(0)} \right) - \frac{4}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)} - u_{i-\frac{1}{2}}^{(0)} \right) \right. \\
 &\quad \left. + \frac{4}{\Delta x} \left(u_{i+\frac{3}{2}}^{(0)} - u_{i+\frac{1}{2}}^{(0)} \right) - \frac{1}{2\Delta x} \left(u_{i+\frac{1}{2}}^{(0)} - 4U_{i+1}^{(0)} + 4U_{i+2}^{(0)} - u_{i+\frac{5}{2}}^{(0)} \right) \right) \\
 &= \frac{1}{4\Delta x^2} \left(u_{i-\frac{3}{2}}^{(0)} - 4U_{i-1}^{(0)} + 8u_{i-\frac{1}{2}}^{(0)} + 4U_i^{(0)} - 18u_{i+\frac{1}{2}}^{(0)} + 4U_{i+1}^{(0)} + 8u_{i+\frac{3}{2}}^{(0)} - 4U_{i+2}^{(0)} + u_{i+\frac{5}{2}}^{(0)} \right)
 \end{aligned}$$

and

$$\begin{aligned}
 \frac{1}{2}(D_{\text{mod}}^+ \psi_{i+\frac{1}{2}}^{(0)} + D_{\text{mod}}^- \psi_{i+\frac{1}{2}}^{(0)}) &= \frac{1}{2\Delta x} \left(\psi_{i-\frac{1}{2}}^{(0)} - 4\Psi_i^{(0)} + 4\Psi_{i+1}^{(0)} - \psi_{i+\frac{3}{2}}^{(0)} \right) \\
 &= \frac{1}{2\Delta x} \left(\frac{1}{4\Delta x^2} \left(u_{i-\frac{5}{2}}^{(0)} - 4U_{i-2}^{(0)} + 8u_{i-\frac{3}{2}}^{(0)} + 4U_{i-1}^{(0)} \right. \right. \\
 &\quad \left. \left. - 18u_{i-\frac{1}{2}}^{(0)} + 4U_i^{(0)} + 8u_{i+\frac{1}{2}}^{(0)} - 4U_{i+1}^{(0)} + u_{i+\frac{3}{2}}^{(0)} \right) \right. \\
 &\quad \left. - \frac{4}{2\Delta x^2} \left(\left(u_{i-\frac{1}{2}}^{(0)} - 4U_i^{(0)} + 4U_{i+1}^{(0)} - u_{i+\frac{3}{2}}^{(0)} \right) \right. \right. \\
 &\quad \left. \left. - \left(u_{i-\frac{3}{2}}^{(0)} - 4U_{i-1}^{(0)} + 4U_i^{(0)} - u_{i-\frac{1}{2}}^{(0)} \right) \right) \right. \\
 &\quad \left. + \frac{4}{2\Delta x} \left(\left(u_{i+\frac{1}{2}}^{(0)} - 4U_{i+1}^{(0)} + 4U_{i+2}^{(0)} - u_{i+\frac{5}{2}}^{(0)} \right) \right. \right. \\
 &\quad \left. \left. - \left(u_{i-\frac{1}{2}}^{(0)} - 4U_i^{(0)} + 4U_{i+1}^{(0)} - u_{i-\frac{3}{2}}^{(0)} \right) \right) \right. \\
 &\quad \left. - \frac{1}{4\Delta x^2} \left(u_{i-\frac{1}{2}}^{(0)} - 4U_i^{(0)} + 8u_{i+\frac{1}{2}}^{(0)} + 4U_{i+1}^{(0)} \right. \right. \\
 &\quad \left. \left. - 18u_{i+\frac{3}{2}}^{(0)} + 4U_{i+2}^{(0)} + 8u_{i+\frac{5}{2}}^{(0)} - 4U_{i+3}^{(0)} + u_{i+\frac{7}{2}}^{(0)} \right) \right) \\
 &= \frac{1}{8\Delta x^3} \left(u_{i-\frac{5}{2}}^{(0)} - 4U_{i-2}^{(0)} + 16u_{i-\frac{3}{2}}^{(0)} - 28U_{i-1}^{(0)} - 35u_{i-\frac{1}{2}}^{(0)} + 104U_i^{(0)} \right. \\
 &\quad \left. - 104U_{i+1}^{(0)} + 35u_{i+\frac{3}{2}}^{(0)} + 28U_{i+2}^{(0)} - 16u_{i+\frac{5}{2}}^{(0)} + 4U_{i+3}^{(0)} - u_{i-\frac{7}{2}}^{(0)} \right).
 \end{aligned}$$

□

It follows from the previous lemma that (4.19) is a time discretisation of

$$\begin{aligned}
 \frac{d}{dt} u_{i+\frac{1}{2}}^{(0)}(t) &= - \left(\max \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t) \right) \right. \\
 &\quad \left. + \min \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t) \right) \right) \\
 &\quad - \frac{1}{8\Delta x^3} \left(u_{i-\frac{5}{2}}^{(0)}(t) - 4U_{i-2}^{(0)}(t) + 16u_{i-\frac{3}{2}}^{(0)}(t) - 28U_{i-1}^{(0)}(t) \right. \\
 &\quad \left. - 35u_{i-\frac{1}{2}}^{(0)}(t) + 104U_i^{(0)}(t) - 104U_{i+1}^{(0)}(t) + 35u_{i+\frac{3}{2}}^{(0)}(t) \right. \\
 &\quad \left. + 28U_{i+2}^{(0)}(t) - 16u_{i+\frac{5}{2}}^{(0)}(t) + 4U_{i+3}^{(0)}(t) - u_{i-\frac{7}{2}}^{(0)}(t) \right).
 \end{aligned} \tag{4.22}$$

Furthermore, since we do not use different update equations for the cell averages, (3.2) holds and, due to Lemma 4.7, it is a temporal discretisation of

$$\begin{aligned}
 \frac{d}{dt} U_i^{(0)}(t) &= - \frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{(0)}(t) \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{(0)}(t) \right)^2 \right) \\
 &\quad - \frac{1}{4\Delta x^3} \left(\left(u_{i-\frac{3}{2}}^{(0)}(t) - 4U_{i-1}^{(0)}(t) + 8u_{i-\frac{1}{2}}^{(0)}(t) + 4U_i^{(0)}(t) - 18u_{i+\frac{1}{2}}^{(0)}(t) + 4U_{i+1}^{(0)}(t) \right. \right. \\
 &\quad \left. \left. + 8u_{i+\frac{3}{2}}^{(0)}(t) - 4U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right) - \left(u_{i-\frac{5}{2}}^{(0)}(t) - 4U_{i-2}^{(0)}(t) + 8u_{i-\frac{3}{2}}^{(0)}(t) \right. \right. \\
 &\quad \left. \left. + 4U_{i-1}^{(0)}(t) - 18u_{i-\frac{1}{2}}^{(0)}(t) + 4U_i^{(0)}(t) + 8u_{i+\frac{1}{2}}^{(0)}(t) - 4U_{i+1}^{(0)}(t) + u_{i+\frac{3}{2}}^{(0)}(t) \right) \right).
 \end{aligned} \tag{4.23}$$

If we assume the time discretisation is globally stiffly accurate, we obtain from (4.20) and (4.21)

$$\rho_{i+\frac{1}{2}}^{(0)n+1} = \frac{1}{2\Delta x} \left(u_{i-\frac{1}{2}}^{(0)n+1} - 4U_i^{(0)n+1} + 4U_{i+1}^{(0)n+1} - u_{i+\frac{3}{2}}^{(0)n+1} \right) \quad (4.24)$$

and

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{(0)n+1} = \frac{1}{\Delta x^2} & \left(u_{i-\frac{3}{2}}^{(0)n+1} - 3U_{i-1}^{(0)n+1} + 3u_{i-\frac{1}{2}}^{(0)n+1} + 3U_i^{(0)n+1} - 8u_{i+\frac{1}{2}}^{(0)n+1} \right. \\ & \left. + 3U_{i+1}^{(0)n+1} + 3u_{i+\frac{3}{2}}^{(0)n+1} - 3U_{i+2}^{(0)n+1} + u_{i+\frac{5}{2}}^{(0)n+1} \right). \end{aligned} \quad (4.25)$$

Moreover, (3.4) implies (3.12) and due to (3.3), we obtain

$$\begin{aligned} \Psi_i^{(0)n+1} = \frac{1}{2\Delta x^2} & \left(\left(u_{i-\frac{1}{2}}^{(0)n+1} - 4U_i^{(0)n+1} + 4U_{i+1}^{(0)n+1} - u_{i+\frac{3}{2}}^{(0)n+1} \right) \right. \\ & \left. - \left(u_{i-\frac{3}{2}}^{(0)n+1} - 4U_{i-1}^{(0)n+1} + 4U_i^{(0)n+1} - u_{i+\frac{1}{2}}^{(0)n+1} \right) \right). \end{aligned} \quad (4.26)$$

Hence the combined equations (4.19), (3.2), (4.24), (4.25), (3.12), and (4.26) are a time discretisation using a IMEX Runge-Kutta scheme of type ARS of (4.22), (4.23), (3.16), and

$$\rho_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{2\Delta x} \left(u_{i-\frac{1}{2}}^{(0)}(t) - 4U_i^{(0)}(t) + 4U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t) \right) \quad (4.27)$$

as well as

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{\Delta x^2} & \left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) + 3u_{i-\frac{1}{2}}^{(0)}(t) + 3U_i^{(0)}(t) - 8u_{i+\frac{1}{2}}^{(0)}(t) \right. \\ & \left. + 3U_{i+1}^{(0)}(t) + 3u_{i+\frac{3}{2}}^{(0)}(t) - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right) \end{aligned} \quad (4.28)$$

and

$$\begin{aligned} \Psi_i^{(0)}(t) = \frac{1}{\Delta 2x^2} & \left(\left(u_{i-\frac{1}{2}}^{(0)}(t) - 4U_i^{(0)}(t) + 4U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t) \right) \right. \\ & \left. - \left(u_{i-\frac{3}{2}}^{(0)}(t) - 4U_{i-1}^{(0)}(t) + 4U_i^{(0)}(t) - u_{i+\frac{1}{2}}^{(0)}(t) \right) \right). \end{aligned} \quad (4.29)$$

Theorem 4.8. Equation (4.22) is a spatial discretisation of

$$\frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) = 0$$

with an error in $\mathcal{O}(\Delta x^2)$ and (4.23) is a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) dx = 0$$

with an error on $\mathcal{O}(\Delta x^6)$. Additionally, it follows that (4.27) and (4.28) are discretisations of

$$\rho^{(0)}(t, x) = \frac{\partial}{\partial x} u^{(0)}(t, x)$$

with an error in $\mathcal{O}(\Delta x^4)$ and

$$\psi^{(0)}(t, x) = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x)$$

with an error in $\mathcal{O}(\Delta x^6)$ as well as (4.29) is a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \psi^{(0)}(t, x) dx = \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) dx.$$

with an maximum error in $\mathcal{O}(\Delta x^4)$.

Proof. Assume that $u^{(0)} : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$ is a smooth and bounded function with bounded derivatives that interpolates the point values and cell averages. The assumption implies that

$$\begin{aligned} \frac{1}{4\Delta x^2} & \left(u_{i-\frac{3}{2}}^{(0)}(t) - 4U_{i-1}^{(0)}(t) + 8u_{i-\frac{1}{2}}^{(0)} + 4U_i^{(0)}(t) - 18u_{i+\frac{1}{2}}^{(0)}(t) + 4U_{i+1}^{(0)}(t) \right. \\ & \left. + 8u_{i+\frac{3}{2}}^{(0)}(t) - 4U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right) = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) + \frac{\Delta x^6}{2160} \frac{\partial^8}{\partial x^8} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^8) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{8\Delta x^3} & \left(u_{i-\frac{5}{2}}^{(0)}(t) - 4U_{i-2}^{(0)}(t) + 16u_{i-\frac{3}{2}}^{(0)}(t) - 28U_{i-1}^{(0)}(t) - 35u_{i-\frac{1}{2}}^{(0)}(t) + 104U_i^{(0)}(t) \right. \\ & \left. - 104U_{i+1}^{(0)}(t) + 35u_{i+\frac{3}{2}}^{(0)}(t) + 28U_{i+2}^{(0)}(t) - 16u_{i+\frac{5}{2}}^{(0)}(t) + 4U_{i+3}^{(0)}(t) - u_{i+\frac{7}{2}}^{(0)}(t) \right) \\ & = \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\Delta x^4}{120} \frac{\partial^7}{\partial x^7} u^{(0)}(t, x_{i+\frac{1}{2}}) + (\mathcal{O})(\Delta x^6). \end{aligned}$$

Thus, proceeding as in the proof of Theorem 3.6, (4.22) is a spatial discretisation of

$$\frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) = 0$$

with an error in $\mathcal{O}(\Delta x^2)$. In addition, we obtain

$$\begin{aligned} \frac{1}{4\Delta x^3} & \left(\left(u_{i-\frac{3}{2}}^{(0)}(t) - 4U_{i-1}^{(0)}(t) + 8u_{i-\frac{1}{2}}^{(0)} + 4U_i^{(0)}(t) - 18u_{i+\frac{1}{2}}^{(0)}(t) + 4U_{i+1}^{(0)}(t) \right. \right. \\ & \left. \left. + 8u_{i+\frac{3}{2}}^{(0)}(t) - 4U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right) - \left(u_{i-\frac{5}{2}}^{(0)}(t) - 4U_{i-2}^{(0)}(t) + 8u_{i-\frac{3}{2}}^{(0)} \right. \right. \\ & \left. \left. + 4U_{i-1}^{(0)}(t) - 18u_{i-\frac{1}{2}}^{(0)}(t) + 4U_i^{(0)}(t) + 8u_{i+\frac{1}{2}}^{(0)}(t) - 4U_{i+1}^{(0)}(t) + u_{i+\frac{3}{2}}^{(0)}(t) \right) \right) \\ & = \frac{1}{\Delta x} \left(\frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) \\ & + \frac{\Delta x^5}{2160} \left(\frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) - \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i-\frac{1}{2}}) \right) + \mathcal{O}(\Delta x^7) \\ & = \frac{1}{\Delta x} \int_{C_i} u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) dx - \frac{\partial^3}{\partial x^3} u^{(0)}(t, x_{i+\frac{1}{2}}) dx + \mathcal{O}(\Delta x^6). \end{aligned}$$

This implies that (4.23) is a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) dx = 0$$

with an error in $\mathcal{O}(\Delta x^6)$. Proceeding as for the central scheme or the classical scheme, it follows that (4.27) and (4.28) are discretisations of

$$\rho^{(0)}(t, x) = \frac{\partial}{\partial x} u^{(0)}(t, x)$$

and with an error in $\mathcal{O}(\Delta x^4)$

$$\psi^{(0)}(t, x) = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x)$$

with with an error in $\mathcal{O}(\Delta x^6)$. Lastly, it follows that (4.29) is a spatial discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \psi^{(0)}(t, x) dx = \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) dx$$

with an maximum error in $\mathcal{O}(\Delta x^4)$. □

Remark 4.9. The error of the first discretisation in the theorem is due to the nonlinear part, where still $D^\pm$, which is second order accurate, is used.

To conclude the findings: The formal asymptotic analysis indicates that the modified central IMEX Active Flux scheme for the HKdV equations discretised in time with an IMEX Runge-Kutta scheme of type ARS is asymptotic preserving for u component. If, in addition, the IMEX Runge-Kutta scheme is globally stiffly accurate, then the modified central IMEX Active Flux scheme is indicated to be asymptotic preserving.

4.4.2 Implementation

Even though different finite difference formulas are used, the implementation is still similar to the one of the other schemes. The only difference are again the matrices M_1 , M_2 , and M_3 :

$$M_1 = \begin{pmatrix} 0 & 0 & -\frac{1}{2\Delta x} & 0 & 0 & \frac{4}{2\Delta x} \\ 0 & \frac{1}{2\Delta x \varepsilon^2} & 0 & 0 & -\frac{4}{2\Delta x \varepsilon^2} & 0 \\ -\frac{1}{2\Delta x \varepsilon^2} & 0 & 0 & \frac{4}{2\Delta x \varepsilon^2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -\frac{4}{2\Delta x} \\ 0 & 0 & -\frac{1}{\varepsilon^2} & 0 & \frac{4}{2\Delta x \varepsilon^2} & 0 \\ 0 & \frac{1}{\varepsilon^2} & 0 & -\frac{4}{2\Delta x \varepsilon^2} & 0 & 0 \\ 0 & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & -\frac{1}{\varepsilon^2} \\ \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & \frac{1}{\varepsilon^2} & 0 \end{pmatrix},$$

and,

$$M_3 = \begin{pmatrix} 0 & 0 & \frac{1}{2\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{1}{2\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ \frac{1}{2\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

4.4.3 Numerical results

Figure 56 - 58 depicted the results of Experiment 1 and Figure 59 - 62 the results of Experiment 2 conducted for the modified central IMEX Active Flux scheme combined with the ARS(4, 4, 3) scheme as a time discretisation. If necessary, $C_{\text{CFL}} = 0.35$ is chosen in all experiments. As shown in Figure 56, where the results of Experiment 1.1 are displayed, the scheme achieves third order accuracy and the numerical solution on a coarse grid is visually indistinguishable from the exact solution with a maximum l_2 -error of roughly 0.0137. Thus, using $D_{\text{mod}}^\pm$ instead of $D^\pm$ increases the numerically observed accuracy of the scheme from two to three. Furthermore, the results of Experiment 1.2 are depicted in Figure 57. The numerical solution visually matches the exact solution at $t = 180$ with a maximum l_2 -error of approximately 0.0019. In addition, the scheme is stable up to $C_{\text{CFL}} = 0.45$. In Figure 58 one can see the discrete mass and energy with respect to time, i.e. the results of Experiment 1.3. The discrete mass at $t = 360$ is approximately 100.00% of the discrete mass at $t = 0$ for all the different spatial step sizes considered in the experiment. The discrete energy is approximately 99.84% (256 cells), 99.98% (512 cells), and 100.00% (1024 cells) of the original discrete energy at $t = 0$.

The results of Experiment 2.1 (see Figure 59) indicate that the scheme is asymptotic preserving since the l_2 -error becomes constant as $\varepsilon \rightarrow 0$ but still decreases if $\Delta x \rightarrow 0$. The scheme achieves third order accuracy. However, the numerical solution computed on a coarse grid and with $\varepsilon = 10^{-7}$ does not perfectly matches the exact solution (see Figure 60). It is noticeable that the amplitude of the numerical solution is higher than the one of the exact solution. The maximum l_2 -error of the numerical solution with respect to the KdV soliton solution is approximately 0.0820. As the results of Experiment 2.2 show, computing the numerical solution at $t = 180$ on a grid with 512 cells results in a numerical solution that visually matches the exact solution with a maximum l_2 -error of roughly 0.0190 (see Figure 61). Moreover, the scheme appears to be stable up to $C_{\text{CFL}} = 0.425$. The results of Experiment 2.3 (see Figure 62) show that especially on a coarser grid, the discrete mass and energy increase as $t \rightarrow \infty$. The discrete mass at $t = 360$ is approximately 100.05% (256 cells), 100.00% (512 cells), and 100.00% (1024 cells) of the discrete mass at $t = 0$. The discrete energy is approximately 100.96% (256 cells), 100.14% (512 cells), and 100.02% (1024 cells) of the original discrete energy.

In Table 3, the experimental order of convergence computed with the results of Experiment 1.1 and Experiment 2.1 is given. The EOC computed with the results of both experiments is roughly the same, however in general the EOC associated to Experiment 2.1 is higher than the one associated to Experiment 1.1. Nevertheless, the numerical experiments indicate that the scheme is asymptotic accurate.

	Experiment 1.1		Experiment 2.1	
	512 $\rightarrow$ 1024	1024 $\rightarrow$ 2048	512 $\rightarrow$ 1024	1024 $\rightarrow$ 2048
$u_{i+\frac{1}{2}}$	2.9678	2.9685	3.0130	3.0205
$\rho_{i+\frac{1}{2}}$	2.9142	2.9378	2.9431	2.9864
$\psi_{i+\frac{1}{2}}$	2.8642	2.9049	2.9881	3.0051
U_i	2.8723	2.9134	2.9862	3.0070
P_i	2.8859	2.9136	2.9765	3.0036
Ψ_i	2.8523	2.8966	2.8915	2.9605

Table 3: The experimental order of convergence computed from the results Experiment 1.1 and Experiment 2.1 performed for the modified central IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme.

Since the result of the experiments indicate that the modified IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation has overall satisfactory properties, Experiment 3 is performed for to further validate them. Again, $C_{\text{CFL}} = 0.35$ is chosen for all parts of Experiment 3. The results of the Experiment are depicted in Figure 63 (Experiment 3.1) and Figure 64 (Experiment 3.2). On a coarse grid with 64 cells, the numerical solution at $t = 4$ differs from the exact solution. For example, the larger soliton of the numerical solution has a higher amplitude than the one of the exact solution and small oscillations are present on the left of it. The maximum l_2 -error of numerical solution is approximately 0.4963. However, the scheme achieves approximately third order accuracy with values of $\text{EOC}_{512 \rightarrow 1024}$ ranging between 2.8691 and 3.0186 as well as values of $\text{EOC}_{1024 \rightarrow 2048}$ ranging between 2.9348 and 3.0145. Furthermore, as shown in Figure 64, the numerical solution computed on a grid with 512 cells visually matches the exact solution at the different times considered in the experiment. In particular, the oscillations that can be observed in Experiment 3.1 are no longer present. The maximum l_2 -error of the numerical solution at the different times is approximately 0.0020 ($t = 20$), 0.0067 ($t = 40$), and 0.0166 ($t = 60$).

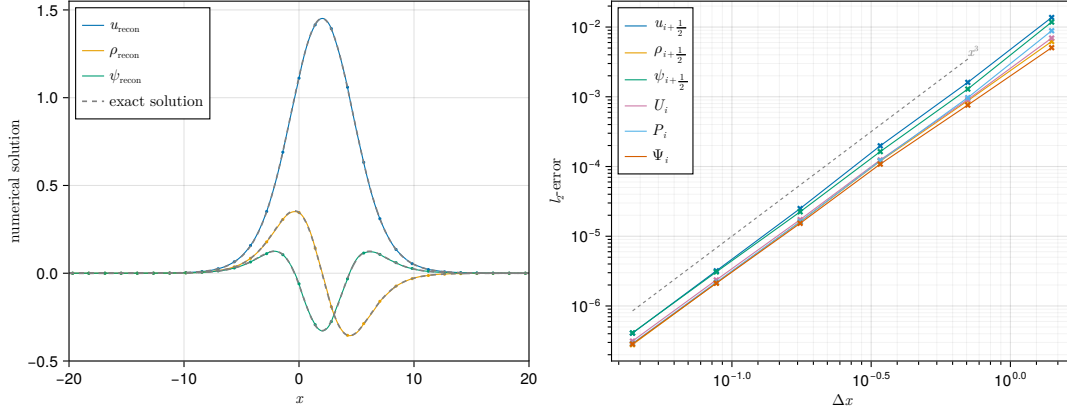


Figure 56: The results of Experiment 1.1 for the modified central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

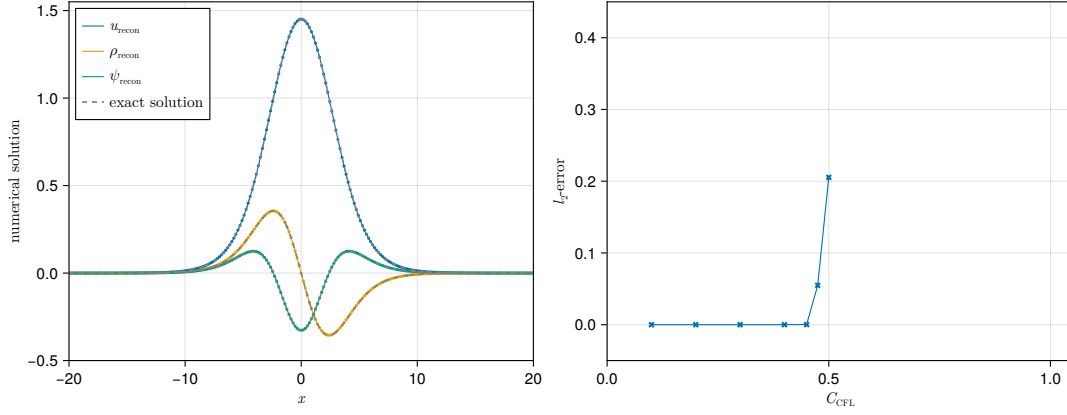


Figure 57: The results of Experiment 1.2 for the modified central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

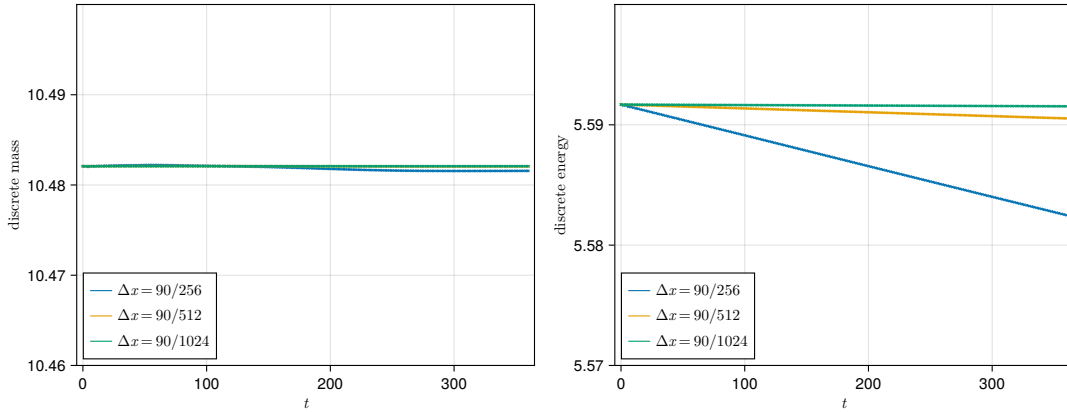


Figure 58: The results of Experiment 1.3 for the modified central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy $E_{\text{H,dis}}$ with respect to time t .

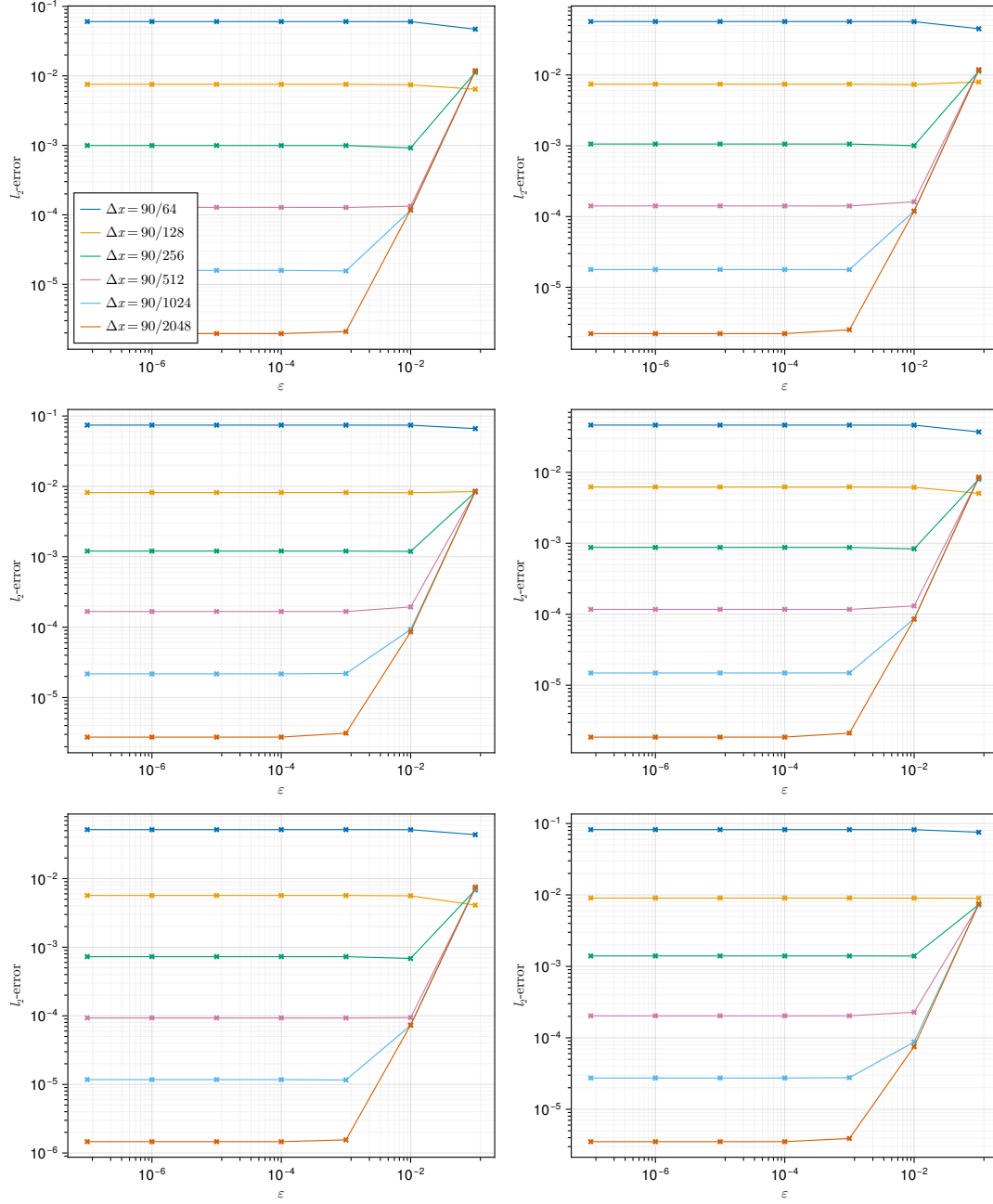


Figure 59: The results of Experiment 2.1 (first part) for the modified central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. From to left to bottom right: the l_2 -error of $u_{i+\frac{1}{2}}$, U_i , $\rho_{i+\frac{1}{2}}$, P_i , $\psi_{i+\frac{1}{2}}$ and Ψ_i with respect to ε for different spatial step sizes.

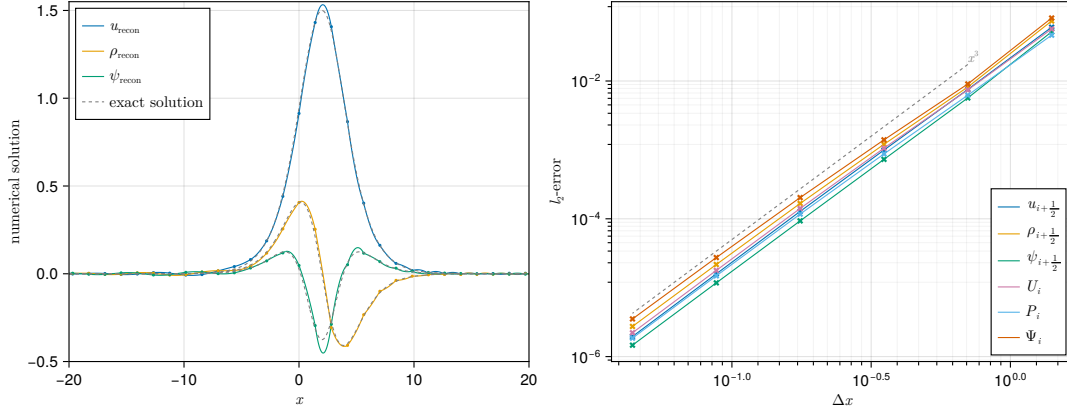


Figure 60: The results of Experiment 2.1 (second part) for the modified central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

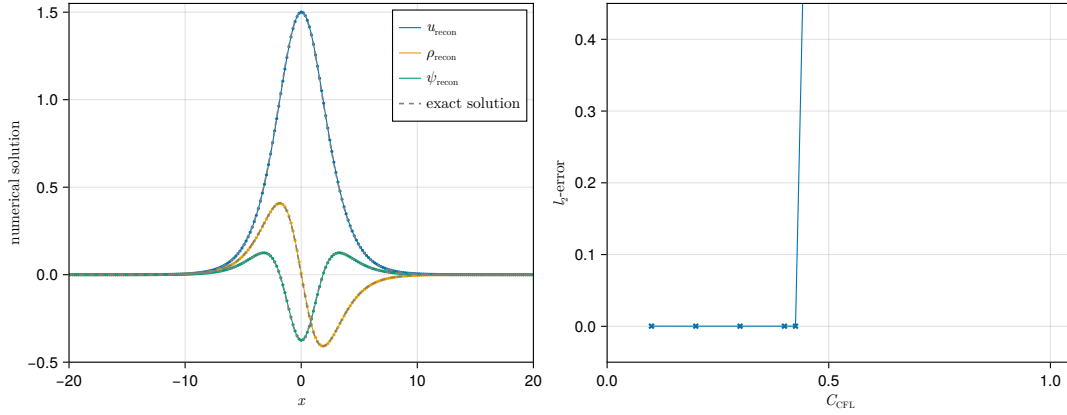


Figure 61: The results of Experiment 2.2 for the modified central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

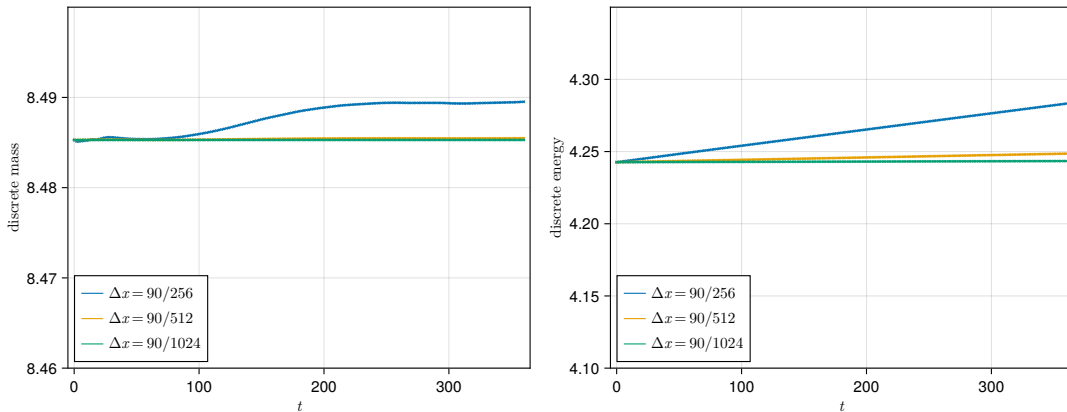


Figure 62: The results of Experiment 2.3 for the modified central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy E_{dis} with respect to time t .

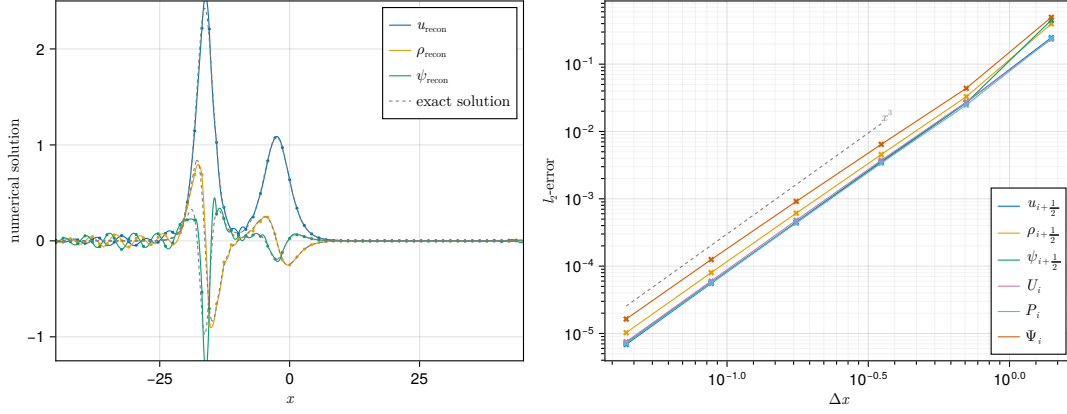


Figure 63: The results of Experiment 3.1 for the modified central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

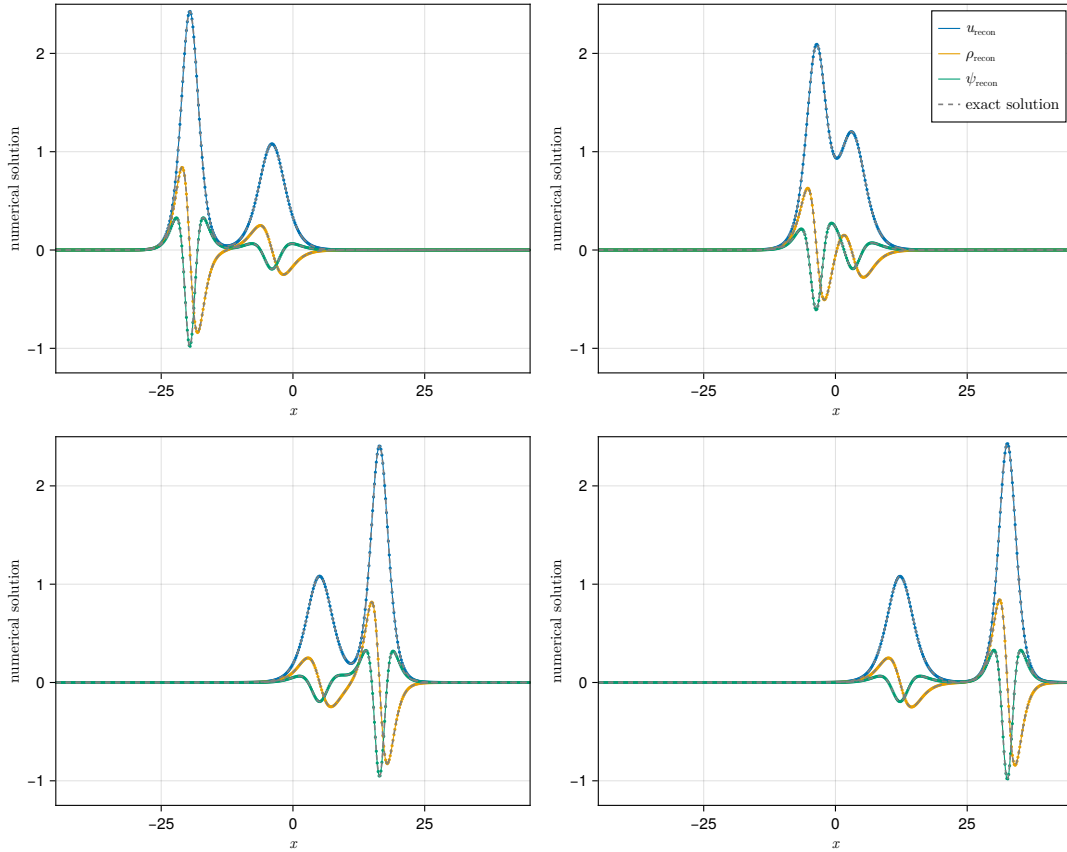


Figure 64: The results of Experiment 3.2 for the modified central IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. From top left to bottom right: The numerical solution computed on grid with 512 cells at $t = 0, 20, 40,$ and 60 in comparison to the exact solution.

4.5 Biased IMEX Active Flux scheme

The last alternative IMEX Active Flux scheme developed and introduced in this thesis combines the approach of the modified central scheme to use higher order finite difference formulas with another idea. Inspired by the spatial discretisation of HKdV equations presented in [39], the first and third equation are discretised by imposing a directional bias and the second equation by using the central approach. The resulting matrixes read

$$F_{2,\text{biased}}^+ = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -\frac{1}{2\varepsilon^2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad F_{2,\text{biased}}^- = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\frac{1}{2\varepsilon^2} & 0 \\ \frac{1}{\varepsilon^2} & 0 & 0 \end{pmatrix}.$$

In addition, the matrices $F_1^\pm$ are used in the point value update equations. Numerical experiments indicate that using the standard $D^\pm$ operators in the second update equation results in a scheme that is only second order. However, using the finding of the previous section, the modified operators $D_{\text{mod}}^\pm$ can be used instead to achieve a higher order of accuracy. The resulting update equations for the point values are

$$\begin{aligned} \frac{d}{dt} u_{i+\frac{1}{2}}(t) &= -\max(u_{i+\frac{1}{2}}(t), 0) D^+ u_{i+\frac{1}{2}}(t) - \min(u_{i+\frac{1}{2}}(t), 0) D^- u_{i+\frac{1}{2}}(t) - D^+ \psi_{i+\frac{1}{2}}(t), \\ \frac{d}{dt} \rho_{i+\frac{1}{2}}(t) &= \frac{1}{2\varepsilon^2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}(t) - D_{\text{mod}}^- \rho_{i+\frac{1}{2}}(t) \right) - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}(t), \\ \frac{d}{dt} \psi_{i+\frac{1}{2}}(t) &= -\frac{1}{\varepsilon^2} D^- u_{i+\frac{1}{2}}(t) + \frac{1}{\varepsilon^2} \rho_{i+\frac{1}{2}}(t). \end{aligned}$$

Due to the impose directional bias, a scheme with these point value update equations and the usual cell average update equations is in this text called the biased IMEX Active Flux scheme.

4.5.1 Formal asymptotic analysis

To perform the formal asymptotic analysis of this scheme, the biased scheme is discretised in time with an ARS type IMEX-Runge Kutta scheme. Ordering the resulting terms by the power of ε , after inserting the assumptions into the time-discrete update equations, yields the same equations associated to the update equation of the cell averages, including (3.2), (3.3), and (3.4). For the point values, the terms of factor ε^{-2} imply

$$\begin{aligned} 0 &= \Delta t A \left(\frac{1}{2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(0)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(0)[j]} \right) - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(\frac{1}{2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(0)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(0)[j]} \right) - \psi_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t A \left(-D^- u_{i+\frac{1}{2}}^{(0)[j]} + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(-D^- u_{i+\frac{1}{2}}^{(0)[j]} + \rho_{i+\frac{1}{2}}^{(0)[j]} \right)_{j=1,\dots,s}, \end{aligned}$$

the terms of factor ε^{-1} imply

$$\begin{aligned} 0 &= \Delta t A \left(\frac{1}{2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(1)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(1)[j]} \right) - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t b \cdot \left(\frac{1}{2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(1)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(1)[j]} \right) - \psi_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \\ 0 &= \Delta t A \left(-D^- u_{i+\frac{1}{2}}^{(1)[j]} + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1,\dots,s}, \end{aligned}$$

$$0 = \Delta t b \cdot \left(-D^- u_{i+\frac{1}{2}}^{(1)[j]} + \rho_{i+\frac{1}{2}}^{(1)[j]} \right)_{j=1, \dots, s},$$

and, lastly, the terms of factor ε^0 imply

$$\begin{aligned} u_{i+\frac{1}{2}}^{[k](0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^k \tilde{a}_{k+1,j} \left(-\max \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \\ &\quad \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) + \Delta t \sum_{j=1}^k a_{k,j} \left(-D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} \right), \\ u_{i+\frac{1}{2}}^{n+1(0)} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\max \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{(0)[j-1]} \right. \\ &\quad \left. - \min \left(u_{i+\frac{1}{2}}^{(0)[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{(0)[j-1]} \right) + \Delta t \sum_{j=1}^s b_j \left(-D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} \right) \end{aligned} \quad (4.30)$$

as well as

$$\begin{aligned} \bar{\rho}_{i+\frac{1}{2}}^{(0)} &= \Delta t A \left(\frac{1}{2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(2)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(2)[j]} \right) - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}, \\ \rho_{i+\frac{1}{2}}^{(0)n+1} &= \Delta t b \cdot \left(\frac{1}{2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(2)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(2)[j]} \right) - \psi_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}, \\ \bar{\psi}_{i+\frac{1}{2}}^{(0)} &= \Delta t A \left(-D^- u_{i+\frac{1}{2}}^{(2)[j]} + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}, \\ \psi_{i+\frac{1}{2}}^{(0)n+1} &= \Delta t b \cdot \left(-D^- u_{i+\frac{1}{2}}^{(2)[j]} + \rho_{i+\frac{1}{2}}^{(2)[j]} \right)_{j=1, \dots, s}. \end{aligned}$$

This implies

$$\psi_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(0)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(0)[j]} \right) \quad (4.31)$$

and

$$\rho_{i+\frac{1}{2}}^{(0)[j]} = D^- u_{i+\frac{1}{2}}^{(0)[j]} \quad (4.32)$$

for all $j = 1, \dots, s$. Using these conditions, we can prove the next lemma.

Lemma 4.10. *The leading order terms, i.e. the terms of order 0, satisfy*

$$\psi^{(0)[j]} = \frac{1}{\Delta x^2} \left(3U_i^{(0)[j]} - 5u_{i+\frac{1}{2}}^{(0)[j]} + 4u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right)$$

and

$$D^+ \psi^{(0)[j]} = \frac{1}{\Delta x^3} \left(6U_{i-1}^{(0)[j]} - 34u_{i-\frac{1}{2}}^{(0)[j]} + 48U_i^{(0)[j]} - 42U_{i+1}^{(0)[j]} + 30u_{i+\frac{3}{2}}^{(0)[j]} - 12U_{i+2}^{(0)[j]} + 4u_{i+\frac{5}{2}}^{(0)[j]} \right)$$

for $j = 1, \dots, s$.

Proof. Using (4.31) and (4.32), we obtain

$$\begin{aligned} \psi_{i+\frac{1}{2}}^{(0)[j]} &= \frac{1}{2} \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(0)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(0)[j]} \right) \\ &= \frac{1}{2\Delta x} \left(\rho_{i-\frac{1}{2}}^{(0)[j]} - 4P_i^{(0)[j]} + 4P_{i+1}^{(0)[j]} - \rho_{i+\frac{3}{2}}^{(0)[j]} \right) \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{2\Delta x} \left(D^- u_{i-\frac{1}{2}}^{(0)[j]} - \frac{4}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) + \frac{4}{\Delta x} \left(u_{i+\frac{3}{2}}^{(0)[j]} - u_{i+\frac{1}{2}}^{(0)[j]} \right) - D^- u_{i+\frac{3}{2}}^{(0)[j]} \right) \\
 &= \frac{1}{\Delta x^2} \left(3U_i^{(0)[j]} - 5u_{i+\frac{1}{2}}^{(0)[j]} + 4u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right)
 \end{aligned}$$

and

$$\begin{aligned}
 D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} &= \frac{1}{\Delta x} \left(2\psi_{i-\frac{1}{2}}^{(0)[j]} - 6\Psi_i^{(0)[j]} + 4\psi_{i+\frac{1}{2}}^{(0)[j]} \right) \\
 &= \frac{1}{\Delta x} \left(\left(D_{\text{mod}}^+ \rho_{i-\frac{1}{2}}^{(0)[j]} + D_{\text{mod}}^- \rho_{i-\frac{1}{2}}^{(0)[j]} \right) - \frac{6}{\Delta x} \left(\rho_{i+\frac{1}{2}}^{(0)[j]} - \rho_{i-\frac{1}{2}}^{(0)[j]} \right) \right. \\
 &\quad \left. + 2 \left(D_{\text{mod}}^+ \rho_{i+\frac{1}{2}}^{(0)[j]} + D_{\text{mod}}^- \rho_{i+\frac{1}{2}}^{(0)[j]} \right) \right) \\
 &= \frac{1}{\Delta x} \left(\frac{1}{2\Delta x} \left(\rho_{i-\frac{3}{2}}^{(0)[j]} - 4P_{i-1}^{(0)[j]} + 4P_i^{(0)[j]} - \rho_{i+\frac{1}{2}}^{(0)[j]} \right) - \frac{6}{\Delta x} \left(\rho_{i+\frac{1}{2}}^{(0)[j]} - \rho_{i-\frac{1}{2}}^{(0)[j]} \right) \right. \\
 &\quad \left. + \frac{1}{\Delta x} \left(\rho_{i-\frac{1}{2}}^{(0)[j]} - 4P_i^{(0)[j]} + 4P_{i+1}^{(0)[j]} - \rho_{i+\frac{3}{2}}^{(0)[j]} \right) \right) \\
 &= \frac{1}{\Delta x} \left(\frac{1}{2\Delta x} \left(D^- u_{i-\frac{3}{2}}^{(0)[j]} - \frac{4}{\Delta x} \left(u_{i-\frac{1}{2}}^{(0)[j]} - u_{i-\frac{3}{2}}^{(0)[j]} \right) + \frac{4}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) - D^- u_{i+\frac{1}{2}}^{(0)[j]} \right) \right. \\
 &\quad \left. - \frac{6}{\Delta x} \left(D^- u_{i+\frac{1}{2}}^{(0)[j]} - D^- u_{i-\frac{1}{2}}^{(0)[j]} \right) \right. \\
 &\quad \left. + \frac{1}{2\Delta x} \left(D^- u_{i-\frac{1}{2}}^{(0)[j]} - \frac{4}{\Delta x} \left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) + \frac{4}{\Delta x} \left(u_{i+\frac{3}{2}}^{(0)[j]} - u_{i+\frac{1}{2}}^{(0)[j]} \right) - D^- u_{i+\frac{3}{2}}^{(0)[j]} \right) \right) \\
 &= \frac{1}{\Delta x^3} \left(6U_{i-1}^{(0)[j]} - 34u_{i-\frac{1}{2}}^{(0)[j]} + 48U_i^{(0)[j]} - 42U_{i+1}^{(0)[j]} + 30u_{i+\frac{3}{2}}^{(0)[j]} + 12U_{i+2}^{(0)[j]} + 4u_{i+\frac{5}{2}}^{(0)[j]} \right).
 \end{aligned}$$

□

The results of the lemma imply that (4.30) is a temporal discretisation using a type ARS IMEX Runge-Kutta scheme of

$$\begin{aligned}
 \frac{d}{dt} u_{i+\frac{1}{2}}^{(0)}(t) &= - \left(\max \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t) \right) \right. \\
 &\quad \left. + \min \left(u_{i+\frac{1}{2}}^{(0)}(t), 0 \right) \frac{1}{\Delta x} \left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t) \right) \right) \\
 &\quad - \frac{1}{\Delta x^3} \left(6U_{i-1}^{(0)[j]} - 34u_{i-\frac{1}{2}}^{(0)[j]} + 48U_i^{(0)[j]} - 42U_{i+1}^{(0)[j]} \right. \\
 &\quad \left. + 30u_{i+\frac{3}{2}}^{(0)[j]} - 12U_{i+2}^{(0)[j]} + 4u_{i+\frac{5}{2}}^{(0)[j]} \right)
 \end{aligned} \tag{4.33}$$

and (3.2) is a discretisation of

$$\begin{aligned}
 \frac{d}{dt} U_i^{(0)}(t) &= - \frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{(0)}(t) \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{(0)}(t) \right)^2 \right) \\
 &\quad - \frac{1}{\Delta x^3} \left(\left(3U_i^{(0)[j]} - 5u_{i+\frac{1}{2}}^{(0)[j]} + 4u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right) \right. \\
 &\quad \left. - \left(3U_{i-1}^{(0)[j]} - 5u_{i-\frac{1}{2}}^{(0)[j]} + 4u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + u_{i+\frac{3}{2}}^{(0)[j]} \right) \right).
 \end{aligned} \tag{4.34}$$

If we assume the IMEX Runge-Kutta scheme of type ARS is globally stiffly accurate then, due to (4.31) and (4.31), we obtain

$$\psi_{i+\frac{1}{2}}^{(0)n+1} = \frac{1}{\Delta x^2} \left(3U_i^{(0)n+1} - 5u_{i+\frac{1}{2}}^{(0)n+1} + 4u_{i+\frac{3}{2}}^{(0)n+1} - 3U_{i+2}^{(0)n+1} + u_{i+\frac{5}{2}}^{(0)n+1} \right) \tag{4.35}$$

and

$$\rho_{i+\frac{1}{2}}^{(0)n+1} = \frac{1}{\Delta x} \left(-4u_{i+\frac{1}{2}}^{(0)n+1} + 6U_{i+1}^{(0)n+1} - 2u_{i+\frac{3}{2}}^{(0)n+1} \right) \quad (4.36)$$

as well as (3.12) and

$$\begin{aligned} \Psi_i^{(0)n+1} = \frac{1}{\Delta x^2} & \left(\left(-4u_{i+\frac{1}{2}}^{(0)n+1} + 6U_{i+1}^{(0)n+1} - 2u_{i+\frac{3}{2}}^{(0)n+1} \right) \right. \\ & \left. - \left(-4u_{i-\frac{1}{2}}^{(0)n+1} + 6U_i^{(0)n+1} - 2u_{i+\frac{1}{2}}^{(0)n+1} \right) \right). \end{aligned} \quad (4.37)$$

Therefore (4.30), (3.2), (4.35), (4.36), (3.12), and (4.37) are a temporal discretisation of the system of ordinary differential equations composed of (4.33), (4.34), (3.16)

$$\rho_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{\Delta x} \left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t) \right), \quad (4.38)$$

$$\psi_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{\Delta x^2} \left(3U_i^{(0)}(t) - 5u_{i+\frac{1}{2}}^{(0)}(t) + 4u_{i+\frac{3}{2}}^{(0)}(t) - 3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t) \right), \quad (4.39)$$

and

$$\begin{aligned} \Psi_i^{(0)}(t) = \frac{1}{\Delta x^2} & \left(\left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t) \right) \right. \\ & \left. - \left(-4u_{i-\frac{1}{2}}^{(0)}(t) + 6U_i^{(0)}(t) - 2u_{i+\frac{1}{2}}^{(0)}(t) \right) \right). \end{aligned} \quad (4.40)$$

Next, the counterpart of Theorem 3.6 for the biased scheme gives information about the finite difference formulas in the time continuous semi-discretisations above.

Theorem 4.11. *Equation (4.22) is a discretisation of*

$$\frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) = 0$$

with an error in $\mathcal{O}(\Delta x)$ and (4.23) is a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) dx = 0$$

with an error in $\mathcal{O}(\Delta x^2)$. Concerning (4.38), (4.39), and (4.40), we have that (4.38) is a spatial discretisation of

$$\rho_{i+\frac{1}{2}}^{(0)}(t, x) = \frac{\partial}{\partial x} u^{(0)}(t, x),$$

(4.39) a discretisation of

$$\psi^{(0)}(t, x) = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x),$$

and (4.40) a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \psi^{(0)}(t, x) dx = \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) dx.$$

The error of all these discretisations is in $\mathcal{O}(\Delta x^2)$.

Proof. Assuming the existence of a smooth and bounded function with bounded derivatives that interpolates the point values and cell averages $u^{(0)} : \mathbb{R}_0^+ \times \mathbb{R} \rightarrow \mathbb{R}$, we obtain

$$\begin{aligned} \frac{1}{\Delta x^2} \left(3U_i^{(0)[j]} - 5u_{i+\frac{1}{2}}^{(0)[j]} + 4u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right) \\ = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) + \frac{\Delta x^2}{12} \frac{\partial^4}{\partial x^4} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^3) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{\Delta x^3} \left(6U_{i-1}^{(0)[j]} - 34u_{i-\frac{1}{2}}^{(0)[j]} + 48U_i^{(0)[j]} - 42U_{i+1}^{(0)[j]} + 30u_{i+\frac{3}{2}}^{(0)[j]} - 12U_{i+2}^{(0)[j]} + 4u_{i+\frac{5}{2}}^{(0)[j]} \right) \\ = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x_{i+\frac{1}{2}}) + \Delta x \frac{\partial^4}{\partial x^4} u^{(0)}(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^2). \end{aligned}$$

In addition, we have

$$\begin{aligned} \frac{1}{\Delta x^3} \left(\left(3U_{i-1}^{(0)[j]} - 5u_{i-\frac{1}{2}}^{(0)[j]} + 4u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + u_{i+\frac{3}{2}}^{(0)[j]} \right) \right. \\ \left. - \left(3U_i^{(0)[j]} - 5u_{i+\frac{1}{2}}^{(0)[j]} + 4u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]} \right) \right) = \frac{1}{\Delta x} \int_{C_i} \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) dx + \mathcal{O}(\Delta x^2). \end{aligned}$$

Thus, similar to the proof of Theorem 3.6, it follows that (4.22) is a discretisation of

$$\frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) = 0$$

with an error in $\mathcal{O}(\Delta x)$, (4.23) is a discretisation of

$$\frac{1}{\Delta x} \int_{C_i} \frac{\partial}{\partial t} u^{(0)}(t, x) + u^{(0)}(t, x) \frac{\partial}{\partial x} u^{(0)}(t, x) + \frac{\partial^3}{\partial x^3} u^{(0)}(t, x) dx = 0$$

with an error on $\mathcal{O}(\Delta x^2)$. Furthermore, we have that (4.38) is a discretisation of

$$\rho_{i+\frac{1}{2}}^{(0)}(t, x) = \frac{\partial}{\partial x} u^{(0)}(t, x),$$

with an error in $\mathcal{O}(\Delta x^2)$, (4.39) is a discretisation of

$$\psi^{(0)}(t, x) = \frac{\partial^2}{\partial x^2} u^{(0)}(t, x)$$

with an with an error in $\mathcal{O}(\Delta x^2)$, and, lastly, (4.40) is a spatial discretisation

$$\frac{1}{\Delta x} \int_{C_i} \psi^{(0)}(t, x) dx = \frac{1}{\Delta x} \int_{C_i} \frac{\partial^2}{\partial x^2} u^{(0)}(t, x) dx.$$

with an error in $\mathcal{O}(\Delta x^2)$. □

This leave us with the following results: The formal asymptotic analysis gives the indication that the biased IMEX Active Flux scheme for the HKdV equations discretised in time with a type ARS IMEX Runge-Kutta scheme is asymptotic preserving for component u . If the IMEX Runge-Kutta scheme is globally stiffly accurate, then the formal asymptotic analysis indicates that the biased scheme is asymptotic preserving.

4.5.2 Implementation

The matrices M_1 , M_2 , and M_3 are

$$M_1 = \begin{pmatrix} 0 & 0 & -\frac{2}{\Delta x} & 0 & 0 & \frac{6}{\Delta x} \\ 0 & \frac{1}{2\Delta x \varepsilon^2} & 0 & 0 & -\frac{4}{2\Delta x \varepsilon^2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 0 & 0 & -\frac{4}{\Delta x} & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\varepsilon^2} & 0 & \frac{4}{2\Delta x \varepsilon^2} & 0 \\ \frac{4}{\Delta x \varepsilon^2} & \frac{1}{\varepsilon^2} & 0 & -\frac{6}{\Delta x \varepsilon^2} & 0 & 0 \\ 0 & 0 & \frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & -\frac{1}{\varepsilon^2} \\ \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & \frac{1}{\varepsilon^2} & 0 \end{pmatrix},$$

and,

$$M_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ \frac{2}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\Delta x} & 0 & 0 & 0 \\ 0 & \frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 \\ -\frac{1}{\Delta x \varepsilon^2} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Otherwise, the implementation of the biased IMEX Active Flux scheme is the same as the implementation of the other schemes.

4.5.3 Numerical results

As for the other scheme, the biased IMEX Active Flux scheme is investigated numerically by performing the different experiments. Therefore, it is discretised in time with the ARS(4, 4, 3) scheme and $C_{\text{CFL}} = 0.4$ is chosen. The results of Experiment 1.1 can be observed in Figure 65, where one can see that the scheme achieves approximately third-order accuracy and that the numerical solution computed on a coarse grid visually coincides with the exact solution. The coloured lines are the reconstruction, and the dots are the point values. The maximum l_2 -error is approximately 0.0241. A similar result is obtained in Experiment 1.2 (see Figure 66) since the numerical solution at $t = 180$ is also visually indistinguishable from the exact solution at this time with a maximum l_2 -error of approximately 0.0007. Moreover, the scheme is stable for all values of C_{CFL} tested in the experiment. The discrete mass and discrete energy with respect to time, i.e. the results of Experiment 1.3, are given in Figure 67. For all spatial step sizes considered in the experiment, the discrete mass at $t = 360$ is approximately 100.00% of the discrete mass at $t = 0$, and the discrete energy at $t = 360$ is approximately 99.94% (256 cells), 99.99% (512 cells), and 100.00% (1024 cells) of the discrete energy at $t = 0$.

Furthermore, the results of Experiment 2.1, which can be seen in Figure 68 and Figure 69, indicate that the biased IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme is asymptotic preserving. For $\varepsilon = 10^{-7}$, the scheme achieves third order accuracy. As for Experiment 1.1, the numerical solution computed on coarse grid visually matches the exact solution with a maximum l_2 -error of 0.0260. The results of the next experiment, Experiment 2.2, can be observed in Figure 70. The numerical solution at $t = 180$ has a maximum l_2 -error

of approximately 0.0138. Additionally, the scheme is stable for all $C_{\text{CFL}} \in (0, 1]$ considered in the experiment. The results of Experiment 2.3 (see Figure 71) indicate that the discrete mass and energy are almost perfectly conserved. The discrete mass at the last measured time, i.e. $t = 360$, is for all spatial step sizes roughly 100.00% of the discrete mass at $t = 0$, and the discrete energy at $t = 360$ is approximately 99.22% (256 cells), 99.90% (512 cells), and 99.99% (1024 cells) of the discrete energy at $t = 0$.

To investigate in detail whether the biased IMEX Active Flux scheme combined with the ARS(4, 4, 3) scheme as time discretisation is asymptotic accurate, the EOC is computed from the results of Experiment 1.1 and Experiment 2.2 (see Table 4). The values associated to the two experiments are close together and, thus, the numerical experiments indicate that the scheme is asymptotic accurate.

	Experiment 1.1		Experiment 2.1	
	512 $\rightarrow$ 1024	1024 $\rightarrow$ 2048	512 $\rightarrow$ 1024	1024 $\rightarrow$ 2048
$u_{i+\frac{1}{2}}$	2.9927	2.9937	2.9812	2.9918
$\rho_{i+\frac{1}{2}}$	2.9699	2.9838	2.9864	2.9951
$\psi_{i+\frac{1}{2}}$	2.9765	2.9863	2.9477	2.9708
U_i	2.9912	2.9928	2.9796	2.9928
P_i	2.9626	2.9791	2.9521	2.9774
Ψ_i	2.9663	2.9818	2.9431	2.9760

Table 4: The experimental order of convergence computed from the results of Experiment 1.1 and Experiment 2.1 performed for the biased IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme.

Lastly, to further analyse the biased IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme, Experiment 3 is conducted, where, again, $C_{\text{CFL}} = 0.4$ is chosen. The results of the experiment can be seen in Figure 72 (Experiment 3.1) and Figure 73 (Experiment 3.2). The numerical solution computed on a coarse grid has a maximum l_2 -error of approximately 0.1827. Moreover, the scheme achieves third order accuracy as depicted in Figure 72. The values of $\text{EOC}_{512 \rightarrow 1024}$ range between 2.9205 and 2.9787 and the values of $\text{EOC}_{1024 \rightarrow 2048}$ range between 2.9618 and 2.9922. The numerical solution computed on a grid with 512 cells and at the different times considered in Experiment 3.2 has a maximum l_2 -error of approximately 0.0014 ($t = 20$), 0.0050 ($t = 40$), and 0.0126 ($t = 60$). Hence, these experiments do not contradict the previous indications.

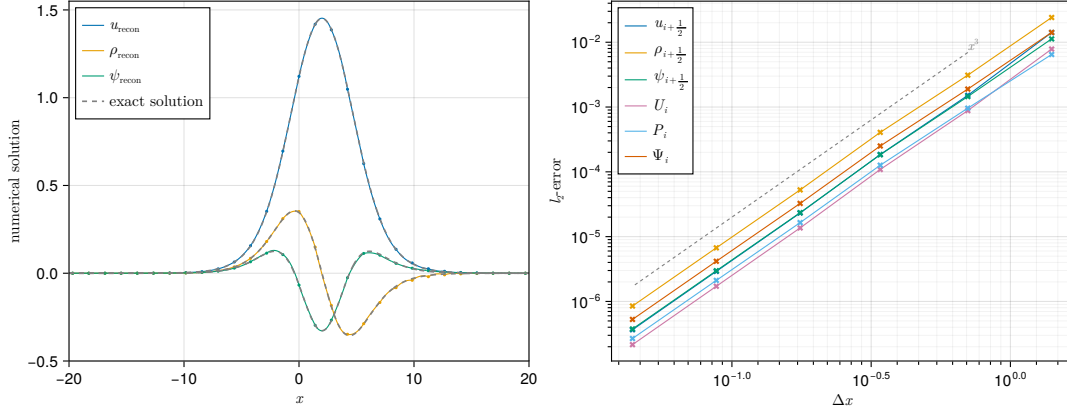


Figure 65: The results of Experiment 1.1 for the biased IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

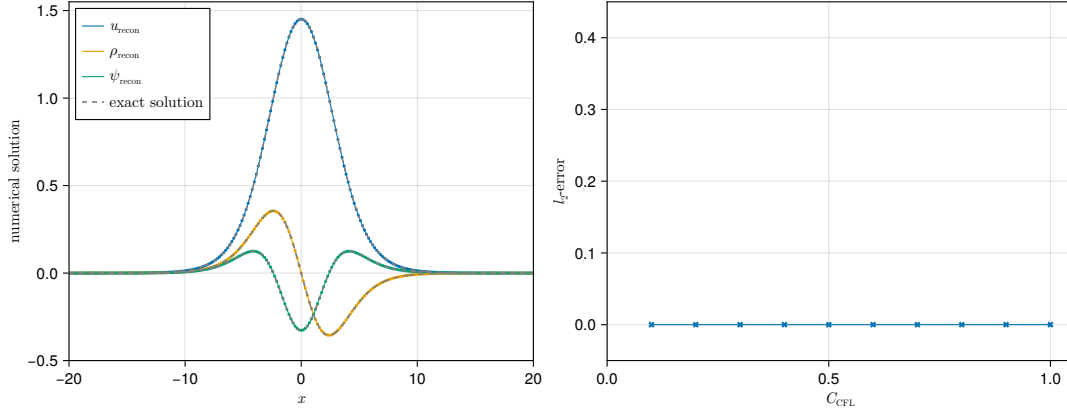


Figure 66: The results of Experiment 1.2 for the biased IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

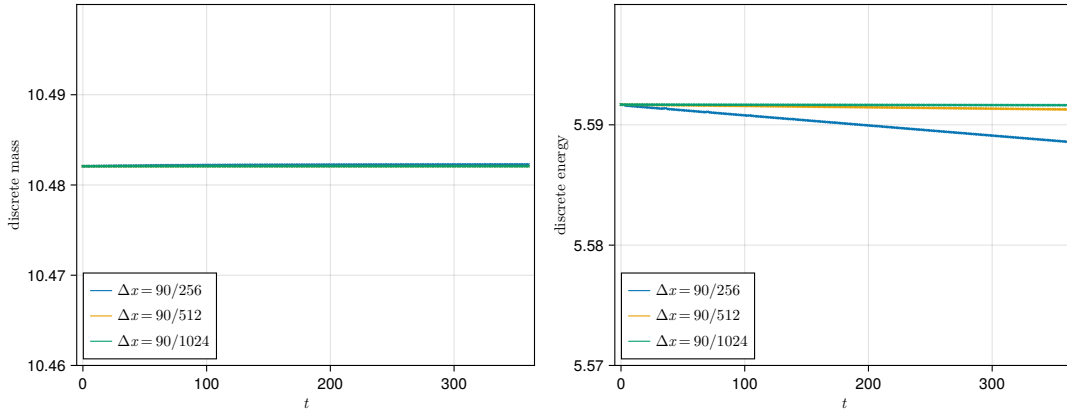


Figure 67: The results of Experiment 1.3 for the biased IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy $E_{\text{H,dis}}$ with respect to time t .

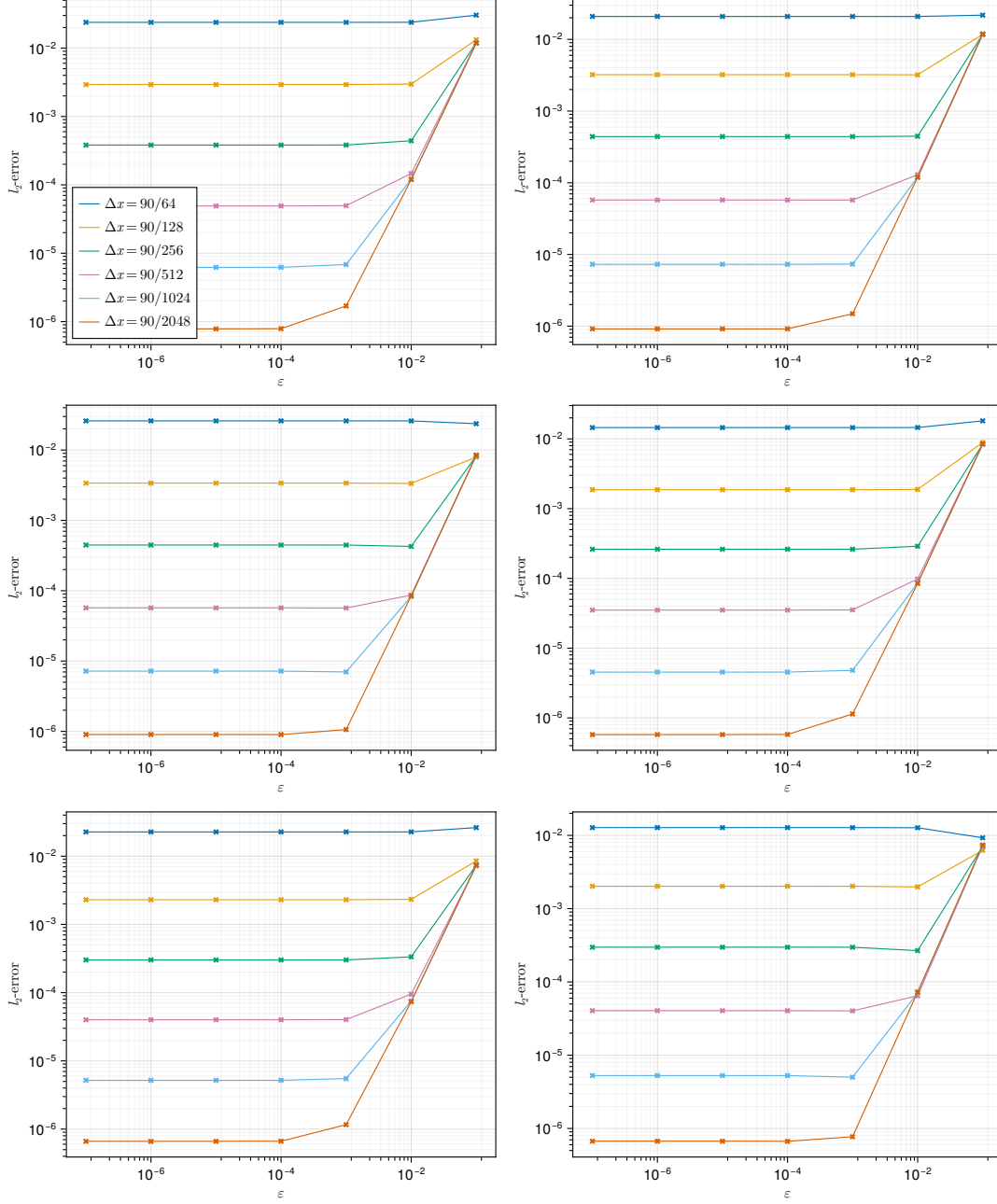


Figure 68: The results of Experiment 2.1 (first part) for the biased IMEX Active Flux scheme with the ARS(4,4,3) scheme as time discretisation. From to left to bottom right: the l_2 -error of $u_{i+\frac{1}{2}}$, U_i , $\rho_{i+\frac{1}{2}}$, P_i , $\psi_{i+\frac{1}{2}}$ and Ψ_i with respect to ε for different spatial step sizes.

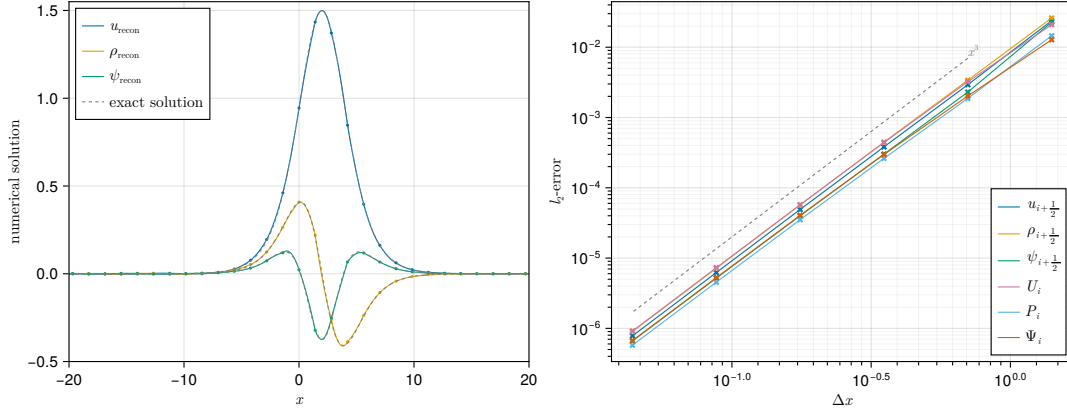


Figure 69: The results of Experiment 2.1 (second part) for the biased IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

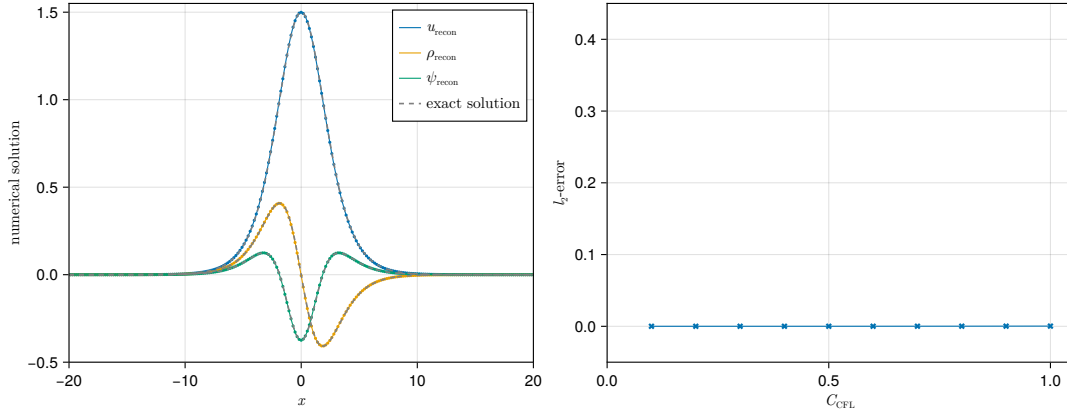


Figure 70: The results of Experiment 2.2 for the biased IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The numerical solution computed on grid with 512 cells at $t = 180$ in comparison to the exact solution. Right: The maximum l_2 -error with respect to the CFL number C_{CFL} .

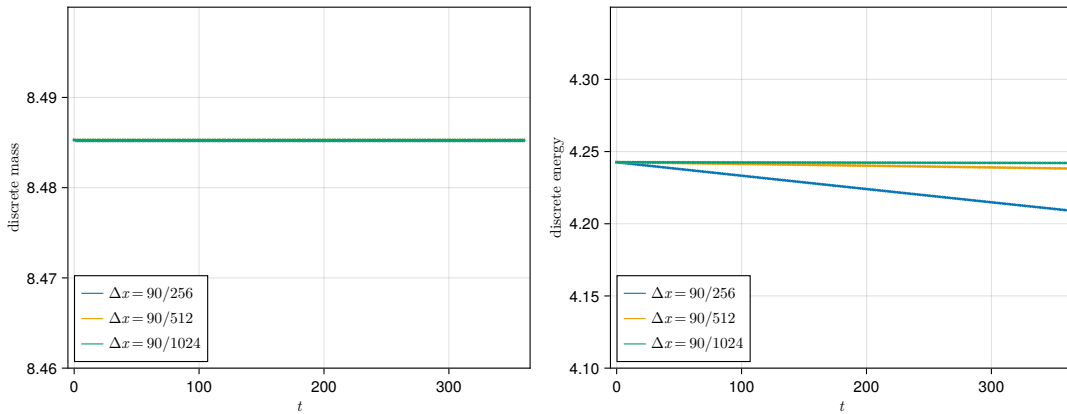


Figure 71: The results of Experiment 2.3 for the biased IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation. Left: The discrete mass M_{dis} with respect to time t . Right: The discrete energy E_{dis} with respect to time t .

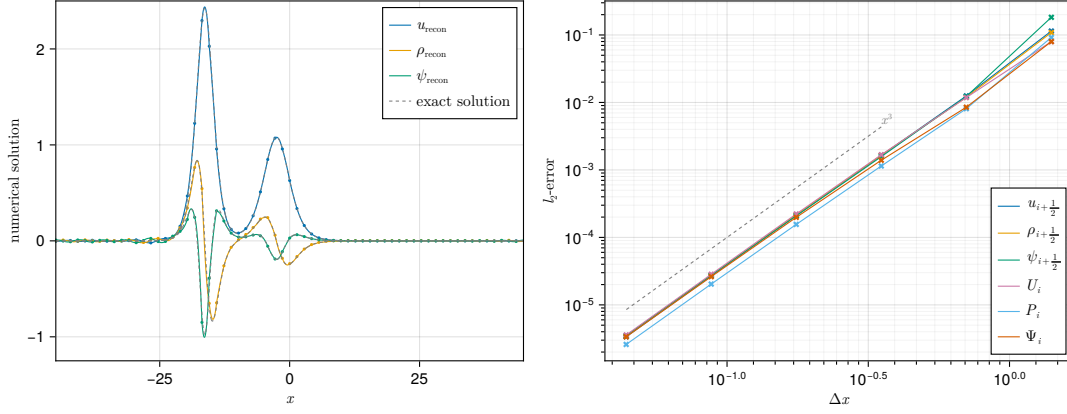


Figure 72: The results of Experiment 3.1 for the biased IMEX Active Flux scheme with the ARS(4,4,3) scheme as time discretisation. Left: The numerical solution computed on a grid with 64 cells at $t = 4$ in comparison to the exact solution. Right: The l_2 -error of the point values and cell averages at $t = 4$ with respect to the spatial step size.

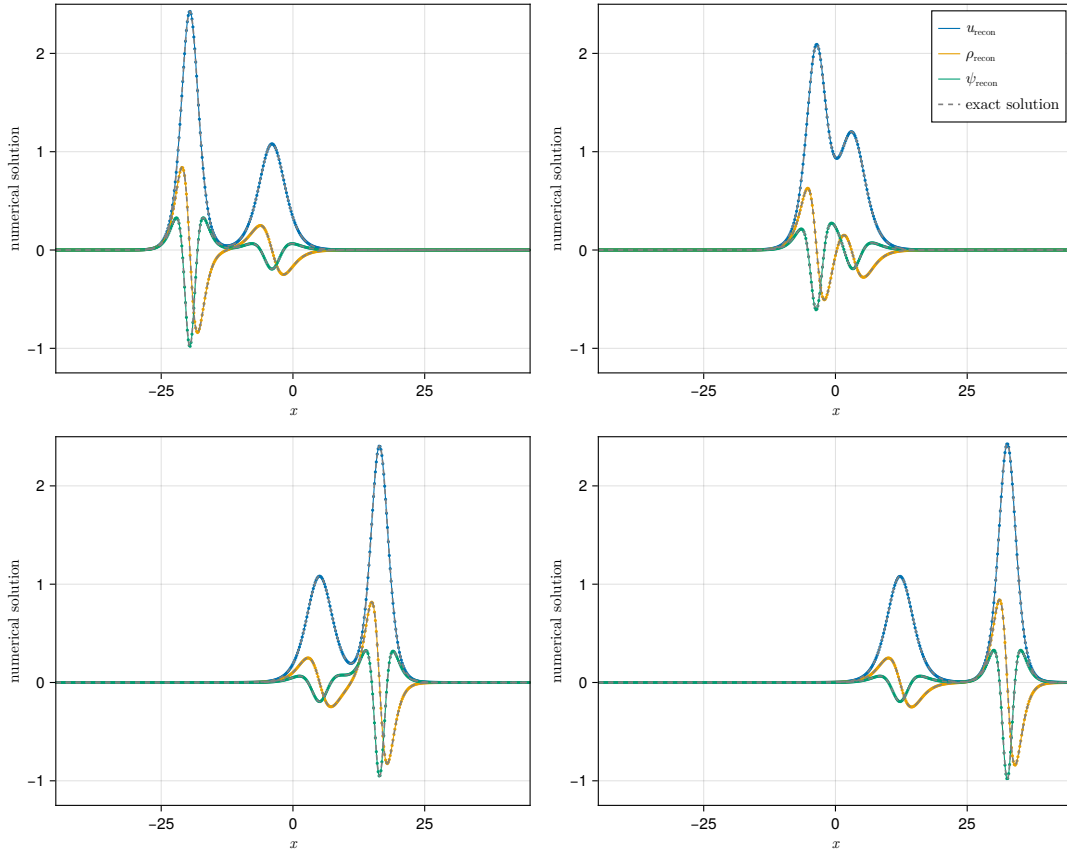


Figure 73: The results of Experiment 3.2 for the biased IMEX Active Flux scheme with the ARS(4,4,3) scheme as time discretisation. From top left to bottom right: The numerical solution computed on grid with 512 cells at $t = 0, 20, 40$, and 60 in comparison to the exact solution.

5 Comparison

After developing and analysing the different IMEX Active Flux schemes discretised in time with IMEX Runge-Kutta schemes of type ARS, the natural next step is to compare the schemes and determine the one with the most promising properties.

Concerning the different time discretisation, we performed Experiment 1 and Experiment 2 for the classical and the optimised IMEX Active Flux scheme discretised in time with the globally stiffly accurate ARS(4, 4, 3) scheme and the ARS(2, 3, 3) scheme, which is not globally stiffly accurate. In case of the classical scheme, the results of the experiments are similar. However, using the optimised scheme with the ARS(2, 3, 3) scheme as time discretisation, the l_2 -error of the point values and cell averages associated to ρ and ψ does not decrease at a fixed rate, and the scheme does not achieve third order accuracy in associated point values and cell averages. However, the scheme achieves third order accuracy in the point values and cell averages associated to the component u . This behaviour matches the indication given by the formal asymptotic analysis in some sense since the indication that the scheme is asymptotic preserving for components ρ and ψ was only given under the assumption that the IMEX Runge-Kutta scheme is globally stiffly accurate. At this point, it is important to note that even though the theoretical and numerical analysis gives no strong indication that optimised scheme with the ARS(2, 3, 3) scheme as time discretisation is asymptotic preserving or asymptotic accurate, this does not imply that scheme does not have these properties as $\varepsilon \rightarrow 0$ and $\Delta \rightarrow 0$. Nevertheless, it appears that using the ARS(4, 4, 3) scheme is the preferable choice. Due to that and to ensure a fair comparison between the different IMEX Active Flux schemes, only result of the experiments performed for schemes with the ARS(4, 4, 3) scheme as time discretisation are compared in the next part.

In Table 5, Table 6, and Table 7 one can see an overview of the result of the first Experiment, where the ability of the schemes to compute numerical solution to an initial value problem to the HKdV equations is analysed. The best results are highlighted in the tables with the exception that the EOC values are highlighted if they are approximately three.

Starting with Experiment 1.1, only the simplified and central scheme deliver inferior results compared to the classical scheme. These two scheme only achieved an order of accuracy of roughly below two. If the accuracy of the numerical solution is measured by the maximum l_2 -error, then the simplified and central scheme are less accurate than the classical scheme on a coarse grid. The other schemes, i.e. the optimised scheme, the modified central scheme and the biased scheme, achieve approximately third order accuracy like the classical scheme. The numerical solution on a coarse grid with 64 cells computed with the optimised and biased scheme is about as accurate as the numerical solution computed with the classical scheme on the same grid. However, the maximum l_2 -error of the numerical solution obtained by employing the modified central Active Flux scheme is noticeably lower than the errors admitted by the other schemes.

Continuing with Experiment 1.2, the maximum l_2 -error of the numerical solution at $t = 180$ computed with the classical, optimised, and central scheme are roughly the same, while the one corresponding to the simplified scheme is significantly higher and the ones of the numerical solutions of the modified central and biased scheme are significantly smaller. The biased scheme computes the most accurate numerical solution. Moreover, the biased scheme and the classical scheme appear to be stable for all $C_{\text{CFL}} \in (0, 1]$, which is not the case for the other schemes.

Lastly, the numerical solution of all schemes despite the simplified one, conserve the energy and mass satisfactory for the different number of cells considered in this experiment. However, again, the biased scheme delivers overall the best results followed by the modified central scheme.

The findings of Experiment 2 are shown in Table 8, Table 9, and Table 10. Formal asymptotic analysis and the numerical experiments indicate the the schemes discretised in time with the ARS(4, 4, 3) scheme are asymptotic preserving and the optimised, central, modified central, and biased scheme are asymptotic accurate.

The maximum l_2 -error of the numerical solution computed with the biased IMEX Active Flux scheme on a coarse grid is by far the lowest of all the maximum l_2 -error stated in Table 8. In contrast, the classical scheme has the highest error and is therefore the scheme with which one obtains the most inaccurate solution. Only the optimised, the modified central, and the biases scheme achieve third order accuracy, while the classical and simplified scheme attain approximately first order accuracy and the central scheme roughly second order accuracy.

Using the biased scheme, the numerical solution computed on a grid with 512 cells admits the lowest maximum l_2 -error at $t = 180$. In addition, it is noticeable that the maximum l_2 -error of the simplified and classical scheme is significantly higher than the other ones given in the table. The biases scheme as well as the classical scheme appear to be the only schemes that are stable for all $C_{CFL} \in (0, 1]$.

Concerning the results of Experiment 2.3, one can observe in Table 10 that as in Table 7 the biased scheme seems to conserve the discrete energy and mass the most satisfactory compared to the other schemes. Note that the discrete energy of the numerical solution computed with the modified scheme increases as $t \rightarrow 0$. This is not the case in Experiment 1.3.

Experiment 3 was performed for the optimised, modified central, and biased scheme discretised in time with the ARS(4, 4, 3) scheme. All schemes achieve third order accuracy in the experiment. However, the maximum l_2 -error of the numerical solution computed on a coarse grid with the biased scheme, which is approximately 0.1827, is smaller than the ones of the numerical solutions computed with the modified central scheme (0.4963) and the optimised scheme (0.2963). Furthermore, the numerical solution computed with the biased scheme on a grid with 512 cells has a maximum l_2 -error of approximately 0.0014 ($t = 20$), 0.0050 ($t = 40$), and $t = 0126$ ($t = 60$). Thus, it is the most accurate of the three schemes since the numerical solution computed with the modified central scheme has a maximum l_2 -error of approximately 0.0020 ($t = 20$), 0.0067 ($t = 40$), and $t = 0166$ ($t = 60$) and the one computed with the optimised scheme has a maximum l_2 -error of approximately 0.0047 ($t = 20$), 0.0173 ($t = 40$), and $t = 0443$ ($t = 60$).

The comparison of the different IMEX Active Flux scheme indicates that the biased IMEX Active Flux scheme discretised in time with the ARS(2, 3, 3) scheme is the scheme with the most advantageous properties out of all the schemes developed and analysed in this thesis.

	classical	simplified	optimised	central	m. central	biased
max l_2 -error	0.0267	0.0534	0.0251	0.0316	0.0137	0.0240
EOC $u_{i+\frac{1}{2}}$	2.9998	1.8478	2.9429	1.9648	2.9685	2.9937
EOC $\rho_{i+\frac{1}{2}}$	2.9933	1.8537	2.9754	1.9257	2.9378	2.9838
EOC $\psi_{i+\frac{1}{2}}$	2.9993	1.8527	2.9879	1.8730	2.9049	2.9863
EOC U_i	2.9982	1.8733	2.9978	1.9729	2.9134	2.9928
EOC P_i	2.9991	1.8705	2.9844	1.9380	2.9136	2.9791
EOC Ψ_i	2.9953	1.8791	2.9918	1.8684	2.8966	2.9818

Table 5: Results of Experiment 1.1 performed for the different IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme, i.e. the l_2 -error of the numerical solution computed on a coarse grid with 64 cells and the EOC for $1024 \rightarrow 2048$ cells.

	classical	simplified	optimised	central	m. central	biased
max l_2 -error	0.0117	0.5154	0.0121	0.0096	0.0019	0.0007
max C_{CFL}	1.000	0.450	0.450	0.450	0.450	1.000

Table 6: Results of Experiment 1.2 performed for the different IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme, i.e. the accuracy of the numerical solution at $t = 180$ and highest values of C_{CFL} for which the scheme is stable on a grid with 512 cells.

	classical	simplified	optimised	central	m. central	biased
M_{dis} (256)	100.00%	99.98%	100.03%	99.86%	100.00%	100.00%
M_{dis} (512)	100.00%	100.01%	100.01%	99.96%	100.00%	100.00%
M_{dis} (1024)	100.00%	100.00%	100.00%	99.99%	100.00%	100.00%
$E_{\text{H,dis}}$ (256)	98.99%	77.13%	98.96%	99.22%	99.84 %	99.94%
$E_{\text{H,dis}}$ (512)	99.87%	93.72%	99.87%	99.91%	99.98%	99.99%
$E_{\text{H,dis}}$ (1024)	99.98%	98.96%	99.98%	99.99%	100.00%	100.00%

Table 7: Results of Experiment 1.3 performed for the different IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme, i.e. the percentage of the discrete mass and energy of the numerical solution at $t = 360$ computed on grid of different sizes with respect to the discrete energy and mass at $t = 0$.

	classical	simplified	optimised	central	m. central	biased
max l_2 -error	0.4581	0.1427	0.0837	0.0759	0.0820	0.0260
EOC $u_{i+\frac{1}{2}}$	0.9681	1.0009	2.9899	1.8920	3.0205	2.9918
EOC $\rho_{i+\frac{1}{2}}$	0.9635	0.9989	2.9978	1.8354	2.9864	2.9951
EOC $\psi_{i+\frac{1}{2}}$	0.9615	1.0025	3.0045	1.9362	3.0051	2.9708
EOC U_i	0.9679	1.0096	2.9986	1.9089	3.0070	2.9928
EOC P_i	0.9667	1.0039	2.9995	1.8888	3.0036	2.9774
EOC Ψ_i	0.9628	1.0037	3.0006	1.8099	2.9605	2.9760

Table 8: Results of Experiment 2.1 performed for the different IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme, i.e. the l_2 -error of the numerical solution computed on a coarse grid with 64 cells and the EOC for $1024 \rightarrow 2048$ cells.

	classical	simplified	optimised	central	m. central	biased
max l_2 -error	3.5441	3.2787	0.0578	0.3024	0.0190	0.0138
max C_{CFL}	1.000	0.725	0.400	0.450	0.425	1.000

Table 9: Results of Experiment 2.2 performed for the different IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme, i.e. the accuracy of the numerical solution at $t = 180$ and highest values of C_{CFL} for which the scheme is stable on a grid with 512 cells.

	classical	simplified	optimised	central	m. central	biased
M_{dis} (256)	100.00%	99.95%	100.03%	100.02%	100.05%	100.00%
M_{dis} (512)	100.00%	99.98%	100.01%	99.97%	100.00%	100.00%
M_{dis} (1024)	100.00%	99.99%	100.00%	99.99%	100.00%	100.00%
E_{dis} (256)	28.06%	57.50%	96.71%	89.64%	100.96%	99.22%
E_{dis} (512)	38.84%	71.95%	99.57%	97.58%	100.14%	99.90%
E_{dis} (1024)	52.25%	83.31%	99.95%	99.58%	100.02%	99.99%

Table 10: Results of Experiment 2.3 performed for the different IMEX Active Flux scheme discretised in time with the ARS(4, 4, 3) scheme, i.e. the percentage of the discrete mass and energy of the numerical solution at $t = 360$ computed on grid of different sizes with respect to the discrete energy and mass at $t = 0$.

6 Summary, conclusion, and outlook

This thesis aimed to achieve multiple goals as stated in the introduction. First, after applying the classical semi-discrete Active Flux scheme based on Jacobian splitting to the HKdV equations, a closely related classical IMEX Active Flux scheme was developed. Formal asymptotic analysis and numerical experiment were conducted for this scheme discretised in time with a globally stiffly accurate IMEX Runge-Kutta scheme of type ARS, e.g. the ARS(4, 4, 3) scheme. IMEX Runge-Kutta schemes of this type were used to avoid the requirement in the formal asymptotic analysis that the initial data is well prepared. The theoretical and numerical analysis of the scheme indicates that the scheme is asymptotic preserving, even though it was not specifically designed to possess this property.

Moreover, to enhance the properties of the classical IMEX Active Flux scheme, IMEX Active schemes with alternative point value update equations were developed. The first two, the simplified and optimised IMEX Active Flux scheme, were obtained by modifying the update equations of the classical scheme by eliminating and rescaling terms in those equations. In addition, based on the central Active Flux scheme, a central IMEX Active Flux scheme for the HKdV equations was developed and, since the scheme turned out to achieve only second order, a modification was introduced, the modified central IMEX Active Flux scheme, that increased the accuracy of the scheme to third order. Lastly, a combination of imposing a directional bias in the update equations of the point values and the modified central IMEX Active Flux scheme was formulated, the biased IMEX Active Flux scheme. The formal asymptotic analysis and numerical experiments, which were conducted for all alternative schemes, indicates that these schemes are asymptotic preserving.

The optimised, modified central, and biased IMEX Active Flux scheme with the ARS(4, 4, 3) scheme as time discretisation are, additionally, asymptotic accurate. Furthermore, comparing the numerical results of the different numerical experiments suggests that the bias IMEX Active Flux scheme is the one with the most advantageous properties.

This thesis serves as a foundation for the further development, analysis, and optimisation of IMEX Active Flux scheme by not only deriving a novel IMEX Active Flux scheme closely related to the classical semi-discrete Active Flux scheme but also offering different possibilities to enhance such a scheme.

A compelling topic of future research is the further analysis of the presented alternative point value update equations and resulting schemes to obtain an improved understanding of the reasons due to which they possess superior properties compared to the classical scheme. One option to further analyse the schemes is to implement the limit schemes and measure the rate at which the numerical approximations of the IMEX Active Flux schemes tends to the numerical approximations of the corresponding limit schemes.

Moreover, extending the results of this thesis to IMEX Runge-Kutta schemes of type I and type II is a natural next step. It is also worth to investigate how the formal asymptotic analysis in this thesis generalises such that one can identify immediately if a given Active Flux scheme is indicated to be asymptotic preserving. Lastly, extending the findings of thesis to different problems, for instance, the hyperbolic approximation of more complicated partial differential equations emerges as a promising topic for future research.

Bibliography

- [1] John Scott Russel. Report on waves. In John Murray, editor, Report of the British Association for the Advancement of Science, 14:311-390, 1845.
- [2] François Marin. Solitons: Historical and physical introduction. In Robert A. Meyers, editor, Mathematics of Complexity and Dynamical Systems. Springer, New York, 2012. DOI: 10.1007/978-3-642-27737-5_506-2.
- [3] P.G. Drazin. Solitons. Cambridge University Press, Cambridge, 1983.
- [4] Timothy Eymann and Philip L. Roe. Active Flux schemes. In 49th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition, 2011. DOI: 10.2514/6.2011-382.
- [5] Timothy Eymann and Philip L. Roe. Active Flux schemes for systems. In 20th AIAA Computational Fluid Dynamics Conference, 2011. DOI: 10.2514/6.2011-3840.
- [6] Timothy Eymann and Philip L. Roe. Multidimensional Active Flux schemes. In 21st AIAA Computational Fluid Dynamics Conference, 2013. DOI: 10.2514/6.2013-2940.
- [7] Bram van Leer. Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. Journal of Computational Physics, 23(3):276299, 1977. DOI: 10.1016/0021-9991(77)90095-X.
- [8] Hiroaki Nishikawa and Philip L. Roe. Third-order Active-Flux scheme for advection diffusion: Hyperbolic diffusion, boundary condition, and Newton solver. Computers & Fluids, 125:71-81, 2016. DOI: 10.1016/j.compfluid.2015.10.020.
- [9] Iman Samani and Philip L. Roe. Acoustics on a coarse grid. In AIAA SCITECH 2023 Forum, 2023. DOI: 10.2514/6.2023-1156.
- [10] Philip L. Roe, Jungyeoul (Brad) Maeng, and Doreen Fan. Comparing Active Flux and Discontinuous Galerkin methods for compressible flow. In 2018 AIAA Aerospace Sciences Meeting, 2018. DOI: 10.2514/6.2018-0836.
- [11] Wasilij Barsukow, Jonathan Hohm, Christian Klingenberg, and Philip L. Roe. The Active Flux scheme on Cartesian grids and its low Mach number limit. Journal of Scientific Computing, 81:594-622, 2019. DOI: 10.1007/s10915-019-01031-z.
- [12] Wasilij Barsukow. An Active Flux method for the Euler equations based on the exact acoustic evolution operator. Journal of Scientific Computing, 108:87, 2026. DOI: 10.1007/s10915-026-03405-6.
- [13] Erik Chudzik and Christiane Helzel. A Review of Cartesian Grid Active Flux Methods for Hyperbolic Conservation Laws. In Emmanuel Franck, Jürgen Fuhrmann, Victor Michel-Dansac, and Laurent Navoret, editors, Finite volumes for complex applications X - Volume 1, elliptic and parabolic Problems. Springer, 2023. DOI: 10.1007/978-3-031-40864-9_6.
- [14] Karthik Duraisamy. Fully discrete Active Flux method based on transported acoustic increments for the compressible euler equations. arXiv preprint arXiv:2605.12915, 2026. DOI: 10.48550/arXiv.2605.12915.
- [15] Rémi Abgrall. A combination of residual distribution and the Active Flux formulations or a new class of schemes that can combine several writings of the same hyperbolic problem: Application to the 1D Euler equations. Communications on Applied Mathematics and Computation, 5:370-402, 2023. DOI: 10.1007/s42967-021-00175-w.

-
- [16] Rémi Abgrall and Wasilij Baruskow. Extensions of Active Flux to arbitrary order of accuracy. *ESAIM: M2AN* 57(2):9911027, 2023. DOI: 10.1051/m2an/2023004.
 - [17] Rémi Abgrall and Wasilij Barsukow. A hybrid finite elementfinite volume method for conservation laws. *Applied Mathematics and Computation*, 447:127846 2023. DOI: 10.1016/j.amc.2023.127846.
 - [18] Rémi Abgrall, Wasilij Barsukow and Christian Klingenberg. A Semi-discrete Active Flux method for the Euler equations on Cartesian grids. *Journal of Scientific Computing*, 102:36, 2025. DOI: 10.1007/s10915-024-02749-1.
 - [19] Junming Duan, Praveen Chandrashekar, and Christian Klingenberg. Active flux for ideal magnetohydrodynamics: A positivity-preserving scheme with the Godunov-Powell source term. *arXiv preprint arXiv:2506.04857*, 2025. DOI: 10.48550/arXiv.2506.04857.
 - [20] Junming Duan, Wasilij Barsukow, and Christian Klingenberg. Active Flux methods for hyperbolic conservation lawsflux vector splitting and bound-preservation. *SIAM Journal on Scientific Computing*, 47(2):A811-A837, 2025. DOI: 10.1137/24M1658887.
 - [21] Wasilij Barsukow, Christian Klingenberg, and Simon Krotzsch. On the equivalence of semi-discrete Active Flux and Discontinuous Galerkin methods and a comparison of their performance. *arXiv preprint arXiv:2603.18160*, 2026. DOI: 10.48550/arXiv.2603.18160.
 - [22] Wasilij Baruskow, Praveen Chandrashekar, Christian Klingenberg, and Lisa Lechner. A generalized Active Flux method of arbitrarily high order in two dimensions. *Computer & Fluids*, 307:106886, 2026. DOI: 10.1016/j.compfluid.2025.106886.
 - [23] Wasilij Barsukow, Janina Kern, Christian Klingenberg, and Lisa Lechner. Analysis of the multi-dimensional semi-discrete Active Flux method using the Fourier transform. *Communications on Applied Mathematics and Computation*, 8:14781526, 2026. DOI: 10.1007/s42967-025-00512-3.
 - [24] Junming Duan, Wasilij Barsukow, and Christian Klingenberg. An asymptotic-preserving active flux scheme for the hyperbolic heat equation in the diffusive scaling. *arXiv preprint arXiv:2508.05166*, 2025. DOI: 10.48550/arXiv.2508.05166
 - [25] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. *Siam Review* 59(1):65-98, 2017. DOI: 10.1137/141000671.
 - [26] JuliaSparse. SparseArrays. <https://github.com/JuliaSparse/SparseArrays.jl>, 2026.
 - [27] Mark Kitisopikul, Timothy E. Holy, and Tomas Aschan. Interpolation.jl. <https://github.com/JuliaMath/Interpolations.jl>, 2025. DOI: 10.5281/zenodo.593143.
 - [28] Steven G. Johnson. QuadGK.jl: Gauss-Kronrod integration in Julia. <https://github.com/JuliaMath/QuadGK.jl>, 2013.
 - [29] JuliaCollections. DataStructures.jl. <https://github.com/JuliaCollections/DataStructures.jl>, 2026.
 - [30] JuliaLang. LinearAlgebra. <https://github.com/JuliaLang/LinearAlgebra.jl>, 2026.
 - [31] JuliaO. JLD2. <https://github.com/juliaio/jld2.jl>, 2026.
 - [32] Simon Danisch and Julius Krumbiegel. Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software* 6(65):3349, 2021. DOI: 10.21105/joss.03349.

-
- [33] Simon Markfelder. Convex integration applied to the multi-dimensional Ccmprssible Euler equations. Springer, Cham, 2021. DOI: doi.org/10.1007/978-3-030-83785-3.
 - [34] Edwige Godlewski and Pierre-Arnaud Raviart. Numerical approximation of hyperbolic systems of conservation laws. Springer, New York, 2021. DOI: doi.org/10.1007/978-1-0716-1344-3.
 - [35] Rahul Barthwal, Firas Dhaouadi, and Christian Rohde. On hyperbolic approximations for a class of dispersive and diffusive-dispersive equations. arXiv preprint arXiv:2512.04882, 2025. DOI: 10.48550/arXiv.2512.04882.
 - [36] Christophe Besse1, Sergey Gavriluk, Maria Kazakova, and Pascal Noble. Perfectly matched layers methods for mixed hyperbolicdispersive equations. *Water Waves* 4:313-343, 2022. DOI: doi.org/10.1007/s42286-022-00069-1.
 - [37] Wolfram Research, Inc. Mathematica, Version 15.0, Champaign, IL, 2026.
 - [38] Christophe Chalons and Rodolphe Turpault. High-order asymptotic-preserving schemes for linear systems: Application to the GoldsteinTaylor equations. *Numerical Methods for Partial Differential Equations*, 35(4):1538-1561, 2019. DOI: 10.1002/num.22363.
 - [39] Abhijit Biswas, David I. Ketcheson, Hendrik Ranocha, and Jochen Schütz. Traveling-wave solutions and structure-preserving numerical methods for a hyperbolic approximation of the Korteweg-de Vries equation. *Journal of Scientific Computing*, 103:90, 2025. DOI: 10.1007/s10915-025-02898-x.
 - [40] Lawrence C. Evan. *Partial Differential Equations*. American Mathematical Society, Providence, 2010.
 - [41] Tohru Yoneyama. The Korteweg-de Vries two-soliton solution as interacting two single solitons. *Progress of Theoretical Physics*, 71(4):843846, 1984. DOI: 10.1143/PTP.71.843.
 - [42] Morris W. Hirsch, Stephen Smale, and Robert L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, San Diego, CA, 2004.
 - [43] Bengt Fornberg. Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of Computation*, 51(184):699-706, 1988. DOI: 10.2307/2008770.
 - [44] J. Álvarez and A. Durán. Petviashvili type methods for traveling wave computations: I. Analysis of convergence. *Journal of Computational and Applied Mathematics*, 266:39-51, 2014. DOI: 10.1016/j.cam.2014.01.015.
 - [45] Randall J. Leveque. *Finite-Volume methods for hyperbolic problems*. Cambridge University Press, Cambridge, 2002.
 - [46] Chi-Wang Shu and Stanley Osher. Efficient implementation of essentially non-oscillatory shock-Capturing schemes. *Journal of Computational Physics*, 77(2):439-471, 1988. DOI: 10.1016/0021-9991(88)90177-5.
 - [47] Wasilij Barsukow. The Active Flux Scheme for Nonlinear Problems. *Journal of Scientific Computing*, 86:3, 2021. DOI: 10.1007/s10915-020-01381-z.
 - [48] Sebastiano Boscarino and Giovanni Russo. Asymptotic preserving methods for quasilinear hyperbolic systems with stiff relaxation: a review. *SeMA Journal*, 81:3-49, 2024. DOI: 10.1007/s40324-024-00351-x.

- [49] Uri M. Ascher, Steven J. Ruuth, and Raymond J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Applied Numerical Mathematics*, 25(2):151-167, 1997. DOI: 10.1016/S0168-9274(97)00056-1.
- [50] Shi Jin. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. *Rivista di Matematica della Università di Parma*, 3(2):177-216, 2012.
- [51] Wasilij Barsukow, Janina Kern, Christian Klingenberg, and Lisa Lechner. Analysis of the multi-dimensional semi-discrete Active Flux method using the Fourier transform. *Communications on Applied Mathematics and Computation*, 8:1578-1526, 2026. DOI: 10.1007/s42967-025-00512-3.
- [52] Junming Duan. A compact fourth-order doubly conservative active flux method with maximum-principle-preserving limiting for degenerate convection–diffusion equations. arXiv preprint arXiv:2510.11527v2, 2026. DOI: 10.48550/arXiv.2510.11527

Titel der Masterarbeit:

Asymptotic preserving IMEX Active Flux schemes for a hyperbolic approximation of the Korteweg-de Vries equation

Thema bereitgestellt von (Titel, Vorname, Nachname, Lehrstuhl):

Prof. Dr. Christian Klingenberg, Mathematische Strömungsmechanik

Eingereicht durch (Vorname, Nachname, Matrikel):

Simon Paul Krotsch, 2610636

Ich versichere, dass ich die vorstehende schriftliche Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die benutzte Literatur sowie sonstige Hilfsquellen sind vollständig angegeben. Wörtlich oder dem Sinne nach dem Schrifttum oder dem Internet entnommene Stellen sind unter Angabe der Quelle kenntlich gemacht.

Weitere Personen waren an der geistigen Leistung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines Ghostwriters oder einer Ghostwriting-Agentur in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar Geld oder geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Arbeit stehen.

☐ Mit dem Prüfungsleiter bzw. der Prüfungsleiterin wurde abgestimmt, dass für die Erstellung der vorgelegten schriftlichen Arbeit Chatbots (insbesondere ChatGPT) bzw. allgemein solche Programme, die anstelle meiner Person die Aufgabenstellung der Prüfung bzw. Teile derselben bearbeiten könnten, entsprechend den Vorgaben der Prüfungsleiterin bzw. des Prüfungsleiters eingesetzt wurden. Die mittels Chatbots erstellten Passagen sind als solche gekennzeichnet.

Der Durchführung einer elektronischen Plagiatsprüfung stimme ich hiermit zu. Die eingereichte elektronische Fassung der Arbeit ist vollständig. Mir ist bewusst, dass nachträgliche Ergänzungen ausgeschlossen sind.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht. Ich bin mir bewusst, dass eine unwahre Erklärung zur Versicherung der selbstständigen Leistungserbringung rechtliche Folgen haben kann.

Würzburg, 04.09.2026 S. Krotsch

Ort, Datum, Unterschrift