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

Simon Krotsch*, Wasilij Barsukow†, Christian Klingenberg*

Abstract

We present a theoretical and numerical analysis of novel IMEX Active Flux schemes for a hyperbolic approximation of the Korteweg-de Vries (KdV) equation. For this approximation, it is important that the numerical schemes recover the limit equations (asymptotic preserving) and retain their order of accuracy in the limit (asymptotically accurate). Using formal asymptotic analysis and numerical experiments, we show that an Active Flux spatial discretisation (closely related to the standard semi-discrete Active Flux scheme), combined with an implicit-explicit Runge-Kutta scheme, is asymptotic preserving but not asymptotically accurate. We therefore propose alternative spatial discretisations that differ in the update equations of the point values. First, we derive a scheme by rescaling and eliminating terms in the upwind matrices. Second, we develop a second-order scheme by using central discretisation for the terms treated implicitly in time and present two modifications of it to third-order. Finally, we formulate a scheme with predefined upwind directions that are not chosen based on characteristics. We show that the alternative schemes are not only asymptotic preserving but also asymptotically accurate, and compare their accuracy, stability, and energy conservation numerically.

Keywords: Active Flux · Semi-discrete schemes · IMEX time discretisation · AP schemes · Formal asymptotic analysis

Mathematics Subject Classification (2020): 65M08, 65M12, 65M20, 35L65

1 Introduction

In 1834, the Scottish engineer John Scott Russell observed a special type of wave while he was following a boat along the Union Channel: "[...] 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." [36]. His observation and experiments to better comprehend the phenomena he encountered were followed by Boussinesq (1871) and Rayleigh (1876) and finally concluded in the derivation of a simple model partial differential equation in 1895 by Korteweg and de Vries [33] which admits solutions similar to Scott's observations. Today, this equation,

$$\frac{\partial}{\partial t}u(t,x) + u(t,x)\frac{\partial}{\partial x}u(t,x) + \frac{\partial^3}{\partial x^3}u(t,x) = 0, \tag{1}$$

is known as the Korteweg-de Vries (KdV) equation. 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 [35], [18]. The special type of travelling wave solution is as follows:

Theorem 1.1 (Single-soliton solution, [24])**.** *The two-parameter family of functions $u : \mathbb{R}^2 \to \mathbb{R}$ defined by*

$$u(t,x) = \tilde{u}(x-ct) = 3c\ \mathrm{sech}^2\left(\frac{\sqrt{c}}{2}(x - x_0 - ct)\right) \tag{2}$$

with $c > 0$, $x_0 \in \mathbb{R}$ satisfies the KdV equation for all $(t,x) \in \mathbb{R}^2$.

The parameter c simultaneously controls the speed and amplitude of the solution. Due to the nonlinearity of (1), one cannot expect that the sum of solutions constitutes another solution. In 1965 Zabusky and Kruskal [41] discovered that something close to a superposition principle holds true for travelling

*Institute of Mathematics, University of Würzburg, Emil-Fischer-Straße 40, 97074 Würzburg, Germany, simon.krotsch@uni-wuerzburg.de (Simon Krotsch), christian.klingenberg@uni-wuerzburg.de (Christian Klingenberg)

†Institut de Mathématiques de Bordeaux (IMB), CNRS UMR 5251, University of Bordeaux, 33405 Talence, France and Imperial College London, Abraham de Moivre Laboratory, CNRS IRL 2004, London, UK, wasilij.barsukow@math.u-bordeaux.fr

wave solutions of the kind of (2). Suppose that at $t = 0$ two such solutions are given, separated from each other and the slower, smaller solution is located to the right of the larger, faster one. After sufficient time, the faster wave catches up to the slower and, up to a shift, their interaction looks as if the faster wave would simply overtake the slower. Zabusky and Kruskal referred to this type of special travelling wave solutions as solitons [5]. The following theorem states a solution to the KdV equation consisting of two solitons.

Theorem 1.2 (Double-soliton solution, [40])**.** *The function* $u : \mathbb{R}^2 \to \mathbb{R}$*,*

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

where $\gamma_i : \mathbb{R}^2 \to \mathbb{R}$*,* $\gamma_i(t,x) = \delta_i x - 4\delta_i^3 t$ *for* $i = 1, 2$ *and* $\delta_1, \delta_2 \in \mathbb{R}$ *with* $\delta_1 < \delta_2$*, satisfies the KdV equation for all* $(t,x) \in \mathbb{R}^2$*.*

The KdV equation has an infinite number of conserved quantities (see [12]). The first two are mass M and energy E with

$$M(t) = \int_{\mathbb{R}} u(t,x)\mathrm{d}x \qquad \text{and} \qquad E(t) = \int_{\mathbb{R}} \frac{1}{2}(u(t,x))^2 \mathrm{d}x.$$

One can approximate high-order PDEs with a first-order hyperbolic system. This process is called hyperbolisation or hyperbolic relaxation. Hyperbolisation can serve multiple purposes, among them the easy implementation into existing codes and imposition of boundary conditions. In certain cases it can be used to remove numerical stiffness. Moreover, high-order PDEs allow for infinite propagation speed, which might be considered a violation of causality and is not the case for first order hyperbolic systems [32]. The hyperbolisations offer an approach to Riemann problems for PDEs that do not a priori admit discontinuous solutions. A substantial amount of research has been dedicated to hyperbolisation. For example, in recent years, hyperbolic approximations of the Navier-Stokes-Korteweg system [17], the conduit equation [29], the Serre-Green-Naghdi model [10], and a class of dispersive and diffusive-dispersive equations [9] have been presented and analysed. The latter class includes the KdV equation.

Hyperbolisations of the same PDE can have different properties, for instance concerning energy conservation. In [11] the following hyperbolic approximation of the KdV equation was presented that was later also studied in [12]:

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

$$\frac{\partial}{\partial t}\rho(t,x) - \frac{1}{\varepsilon^2}\frac{\partial}{\partial x}\rho(t,x) = -\frac{\psi(t,x)}{\varepsilon^2}, \tag{3b}$$

$$\frac{\partial}{\partial t}\ \psi(t,x) + \frac{1}{\varepsilon^2}\frac{\partial}{\partial x}u(t,x) = \frac{\rho(t,x)}{\varepsilon^2}, \tag{3c}$$

where $0 < \varepsilon < 1$ denotes the relaxation parameter. We use ϵ^2 instead of ϵ to avoid its fractional powers in the matrices and the asymptotic analysis below. In the following, we refer to these equations as the hyperbolic KdV (HKdV) equations. They were originally introduced to ease the implementation of boundary conditions. We can rewrite (3) as

$$\frac{\partial}{\partial t}q(t,x) + \frac{\partial}{\partial x}f(q(t,x)) = Sq(t,x), \tag{4}$$

where

$$q(t,x) = \begin{pmatrix} u(t,x) \\ \rho(t,x) \\ \psi(t,x) \end{pmatrix}, \qquad f(q) := \begin{pmatrix} \frac{1}{2}(u)^2 + \psi \\ -\frac{1}{\varepsilon^2}\rho \\ \frac{1}{\varepsilon^2}u \end{pmatrix}, \qquad \text{and} \qquad S = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\varepsilon^2} \\ 0 & \frac{1}{\varepsilon^2} & 0 \end{pmatrix}.$$

The eigenvalues of the Jacobian of f are

$$\lambda_1(q) = -\frac{1}{\varepsilon^2}, \qquad \lambda_2(q) = \frac{1}{2}\left(u - \sqrt{u^2 + \frac{4}{\varepsilon^2}}\right), \qquad \lambda_3(q) = \frac{1}{2}\left(u + \sqrt{u^2 + \frac{4}{\varepsilon^2}}\right) \tag{5}$$

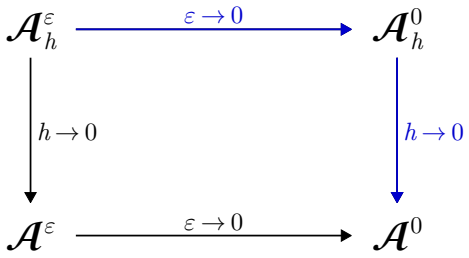


Figure 1: Illustration of the AP property.

and the eigenvectors are

$$v_1(q) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad v_2(q) = \begin{pmatrix} \frac{\varepsilon^2}{2}\left(u - \sqrt{u^2 + \frac{4}{\varepsilon^2}}\right) \\ 0 \\ 1 \end{pmatrix}, \qquad v_3(q) = \begin{pmatrix} \frac{\varepsilon^2}{2}\left(u + \sqrt{u^2 + \frac{4}{\varepsilon^2}}\right) \\ 0 \\ 1 \end{pmatrix}. \tag{6}$$

Since the eigenvectors are linearly independent, the HKdV equations are a hyperbolic system of balance laws with a stiff linear source term.

The formal limit of (3) is

$$\frac{\partial}{\partial t}u(t,x) + u(t,x)\frac{\partial}{\partial x}u(t,x) + \frac{\partial^3}{\partial x^3}u(t,x) = 0, \tag{7a}$$

$$\rho(t,x) = \frac{\partial}{\partial x}u(t,x), \tag{7b}$$

$$\psi(t,x) = \frac{\partial^2}{\partial x^2}u(t,x). \tag{7c}$$

Equation (7a) is the KdV equation (1), while ρ and ψ are the first and second derivatives of u. In [11] it was found that (3) also admits travelling wave solutions and in [12] a way to compute them numerically is presented. In particular, there exist travelling wave solutions that tend to the single-soliton solution of the KdV equation (and its derivatives) as $\varepsilon \to 0$.

Just as the KdV equation, the HKdV equations possess conserved quantities. For a finite ε, they conserve the mass M and a modified energy

$$E_{\mathrm{H}}(t) = \int_{\mathbb{R}} \frac{1}{2}\Big((u(t,x))^2 + \varepsilon^2(\rho(t,x))^2 + \varepsilon^2(\psi(t,x))^2\Big)\mathrm{d}x.$$

Upon $\varepsilon \to 0$, this energy converges formally to E.

Not every numerical method is able to obtain good approximations to the limit equations effectively by solving their hyperbolic approximations with a small ϵ. Schemes that achieve this are called asymptotic preserving (AP). This property is illustrated in Figure 1. The hyperbolic approximation is $\mathcal{A}^\varepsilon$. We assume that this approximation features a relaxation parameter ε. If $\varepsilon \to 0$, $\mathcal{A}^\varepsilon$ formally becomes the limit equations denoted by $\mathcal{A}^0$ (under suitable assumptions on the initial data). $\mathcal{A}_h^\varepsilon$ is a (consistent and stable) numerical scheme discretising $\mathcal{A}^\varepsilon$, where h denotes the discretisation parameter, e.g. $h = (\Delta x, \Delta t)$. For a fixed h, the asymptotic limit of $\mathcal{A}_h^\varepsilon$, provided it exists (under suitable assumptions on the discrete initial data), is denoted by $\mathcal{A}_h^0$.

Definition 1.3 (Asymptotic preserving, [31], [12])**.** *If $\mathcal{A}_h^0$ is a consistent and stable discretisation of $\mathcal{A}^0$, then the numerical scheme $\mathcal{A}_h^\varepsilon$ is called asymptotic preserving (AP).*

If a numerical scheme is AP and, for example, able to achieve third order accuracy for sufficiently large ε, this does not immediately imply that the same holds as $\varepsilon \to 0$. Thus, it is natural to distinguish schemes that retain their order of accuracy as $\varepsilon \to 0$ from those that do not.

Definition 1.4 (Asymptotically accurate, [12])**.** *The numerical scheme $\mathcal{A}_h^\varepsilon$ is called asymptotically accurate (AA) if it retains its order of accuracy as $\varepsilon \to 0$.*

The limit equations of hyperbolic approximations often have the structure of (7) and consist of the (relevant) high-order PDE for u and (irrelevant) constraints. It is possible that a limit scheme $\mathcal{A}_h^0$ only discretises some part of the limit equations. Therefore, we call a scheme for the HKdV equations AP for the component u if it includes a discretisation of (7a). The temporal and spatial discretisation scheme must be carefully chosen to ensure that the limit scheme $\mathcal{A}_h^0$ exists and is a discretisation of $\mathcal{A}^0$. Additionally, the temporal discretisation scheme, for a fixed Δt, has to remain stable as $\varepsilon \to 0$ (asymptotic stability).

In the early 2010s, Philip L. Roe introduced the Active Flux method [25], [26], [27] as a new numerical method to solve hyperbolic conservation laws. It is based on a numerical method for the linear advection equation in one spatial dimension presented in 1977 by Bram van Leer [38]. Roe was able to extend the idea of van Leer to multiple spatial dimensions. Today, Active Flux is a class of different methods. In addition to Roe's version of Active Flux, which is commonly referred to as the traditional or fully-discrete Active Flux method, there exists the generalised or semi-discrete Active Flux method [1], [2], [3], [4]. The different versions of Active Flux share some of the principles of the original scheme, notably (and contrary to classical finite volume or discontinuous Galerkin schemes), the Active Flux method computes the numerical solution without Riemann solvers. This is achieved by considering several types of degrees of freedom that are evolved in time independently: cell averages and point values at the cell interfaces in the case of the third order method. The point values are located at the interfaces of the cells and are shared between adjacent cells. This amounts to a globally continuous reconstruction.

The essential part of the fully-discrete Active Flux method is an evolution operator. It maps 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 good approximate evolution operators for sufficiently complicated problems in multiple spatial dimensions remains a significant challenge (with progress made in [15], [7], [23], [30]). In contrast, the semi-discrete Active Flux method does not depend on an evolution operator. Thus, a discretization determined once is immediately applicable to a variety of partial differential equations. For example, in [19] the semi-discrete Active Flux method was used to numerically solve ideal magnetohydrodynamics.

The semi-discrete version itself offers a high degree of flexibility, with the particular choice of the degrees of freedom and the way to update them shaping the specific method. For instance, different ways to discretise the spatial derivative can be selected and, especially for systems, different ways to include upwinding can be considered.

For the one-dimensional third-order semi-discrete Active Flux schemes considered here, we divide 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 with $i \in \mathbb{Z}$. We assume that all cells have the same size $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} > 0$. The degrees of freedom are cell averages and point values denoted by

$$Q_i(t) \approx \frac{1}{\Delta x} \int_{\mathcal{C}_i} q(t,x)\mathrm{d}x \qquad \text{and} \qquad q_{i+\frac{1}{2}}(t) \approx q(t, x_{i+\frac{1}{2}}).$$

Since the point values are shared between adjacent cells, this amounts to two degrees of freedom per cell: one point value and one average value. 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

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

with $\xi \in [-\frac{\Delta x}{2}, \frac{\Delta x}{2}]$, interpolates the cell averages and the point values. Given this polynomial in each cell, we can define the globally continuous function $q_{\mathrm{recon}} : \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}^3$ by $q_{\mathrm{recon}}(t,x) = q_{\mathrm{recon},i}(t, x - x_i)$ if $x \in \mathcal{C}_i$. We call this function the (global) reconstruction. The time-continuous update equations of the point values and cell averages that constitute the specific schemes considered in this work are introduced in Sections 2, 3, and 5.

It was found that the Active Flux method possesses structure preserving properties, e.g. it is stationarity preserving for the linear acoustic equations [8] and low Mach number compliant [4]. In [21], the authors considered a hyperbolic approximation of the heat equation:

$$\frac{\partial}{\partial t} p(t,\mathbf{x}) + \frac{1}{\varepsilon} \nabla \cdot u(t,\mathbf{x}) = 0, \tag{8a}$$

$$\frac{\partial}{\partial t}\, u(t,\mathbf{x}) + \frac{1}{\varepsilon}\nabla p(t,\mathbf{x}) + \frac{\sigma}{\varepsilon^2} u(t,\mathbf{x}) = 0, \tag{8b}$$

where $\sigma > 0$ and $0 < \varepsilon < 1$. In the limit $\varepsilon \to 0$, this system formally becomes

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

They found that for (8), the third-order semi-discrete Active Flux method with a point value update based on Jacobian splitting combined with implicit time discretisation is AP without any modification. In addition, in one spatial dimension they presented an alternative upwinding which resulted in an asymptotically accurate scheme. In this work we intend to further explore the ability of Active Flux to solve hyperbolic approximations of high-order partial differential equations. We are additionally interested in achieving not only the AP property, but to obtain a method that is AA.

To obtain asymptotically stable schemes, fully implicit time discretisation schemes can be used. However, in the case of nonlinear PDEs such as the HKdV equations, this amounts to solving nonlinear systems at each time step. A more economical approach is to discretise in time with implicit-explicit (IMEX) Runge-Kutta schemes. Consider an initial value problem for an ordinary differential equation,

$$\frac{\mathrm{d}}{\mathrm{d}t} v(t) = f_1(v(t)) + f_2(v(t)) \tag{9}$$

where $v : \mathbb{R}_0^+ \to \mathcal{O}$, $\mathcal{O} \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$ and $f_1, f_2 : \mathcal{O} \to \mathbb{R}^m$. Assume at time step $t^n \geq 0$, $n \in \mathbb{N}_0$, the approximation to $v(t^n)$, denoted by v^n, to be given. Denote by v^{n+1} the approximations at the next time step $t^{n+1} = t^n + \Delta t$. An IMEX Runge-Kutta scheme with $\sigma \in \mathbb{N}$ stages for (9) takes the form

$$v^{[k]} = v^n + \Delta t \left(\sum_{j=1}^{k-1} \tilde{a}_{k,j} f_1(v^{[j]}) + \sum_{j=1}^{\sigma} \hat{a}_{k,j} f_2(v^{[j]}) \right), \tag{10a}$$

$$v^{n+1} = v^n + \Delta t \left(\sum_{j=1}^{\sigma} \tilde{b}_j f_1(v^{[j]}) + \sum_{j=1}^{\sigma} \hat{b}_j f_2(v^{[j]}) \right), \tag{10b}$$

with $k = 1, \dots \sigma$ and $v^{[k]}$ the solution at stage k. The lower triangular matrix $\tilde{A} = (\tilde{a}_{k,j}) \in \mathbb{R}^{\sigma \times \sigma}$ with zero diagonal and $\tilde{b} = (\tilde{b}_j) \in \mathbb{R}^{\sigma}$ are the coefficients of the explicit scheme, while the matrix $\hat{A} = (\hat{a}_{k,j}) \in \mathbb{R}^{\sigma \times \sigma}$ and $\hat{b} = (\hat{b}_j) \in \mathbb{R}^{\sigma}$ are those of the implicit scheme. The matrices are such that the scheme is explicit in f_1 and implicit in f_2. We assume that $\hat{A}$ is a lower triangular matrix, i.e. the implicit part is a diagonally implicit Runge-Kutta scheme. This implies that the second sum in (10a) is from 1 to k.

Definition 1.5 (Types of IMEX Runge-Kutta schemes, [14])**.** *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 $a \in \mathbb{R}^s$, $s = \sigma - 1$, and an invertible sub-matrix $A \in \mathbb{R}^{s \times s}$. If, in addition, $a = 0$ and $\hat{b}_1 = 0$, the IMEX Runge-Kutta scheme is called of type ARS.

For IMEX Runge-Kutta schemes of type ARS, we can rewrite (10) as

$$\begin{aligned} v^{[0]} &= v^n, \\ v^{[k]} &= v^n + \Delta t \left(\sum_{j=1}^{k} \tilde{a}_{k+1,j} f_1(v^{[j-1]}) + \sum_{j=1}^{k} a_{k,j} f_2(v^{[j]}) \right), \\ v^{n+1} &= v^n + \Delta t \left(\sum_{j=1}^{\sigma} \tilde{b}_j f_1(v^{[j-1]}) + \sum_{j=1}^{s} b_j f_2(v^{[j]}) \right), \end{aligned} \tag{11}$$

where $k = 1, \dots s$ and $b = (\hat{b}_2, \dots, \hat{b}_\sigma)^{\mathrm{T}} \in \mathbb{R}^s$.

Definition 1.6 ((Globally) stiffly accurate, [32]). *An IMEX Runge-Kutta scheme is called globally stiffly accurate (GSA) if the corresponding diagonally implicit Runge-Kutta scheme is stiffly accurate, i.e.* $\hat{a}_{\sigma j} = \hat{b}_j$ *for* $j = 1, \dots, \sigma$ *and the explicit scheme is First Same As Last, i.e.* $\tilde{a}_{\sigma,j} = \tilde{b}_j$ *for all* $j = 1, \dots \sigma$.

An immediate consequence of this definition is the following: If an IMEX Runge-Kutta scheme is GSA, then $v^{n+1} = v^{[s]}$. This property is useful for stiff problems [6].

In [12], the authors analysed IMEX Runge-Kutta time discretisations for the HKdV equations (3). They found that an IMEX Runge-Kutta scheme of type I applied to a splitting of (3) without spatial discretisation is always AP for the u-component. If the scheme is additionally GSA, then the scheme is AP for all components. An IMEX Runge-Kutta scheme of type II applied to the same splitting is AP if the scheme is GSA and the initial data is well-prepared. In addition, they developed a structure preserving spatial discretisation based on summation-by-parts operators. If they applied IMEX Runge-Kutta schemes of type I and II to the discretisation instead of the space-continuous equations, they obtained the same results.

The main focus of this work is to analyse whether the Active Flux spatial discretisations are AP or AA. For this reason, we only consider IMEX Runge-Kutta schemes of type ARS. An additional reason for using these schemes is that their properties turn out to be important for the later analysis, e.g. the necessity for the initial data to be well-prepared can be avoided. So far, [21] is the only study of AP properties of Active Flux for a hyperbolised high-order PDE. The CFL condition associated with the heat equation still allowed time-explicit methods to be used; the time-implicit method employed was fully implicit since the entire problem was linear. The present work, focusing on a nonlinear, third-order PDE must thus go significantly beyond [21]. Explicit methods can no longer be reasonably used, and the nonlinearity requires the development of an IMEX Active Flux method. This in itself is a novelty, since so far, the semi-discrete Active Flux scheme has only ever been integrated in time fully explicitly or fully implicitly. The main contributions of this work are:

1. The development of an Active Flux scheme that can be discretised in time with IMEX Runge-Kutta schemes.

2. The formal asymptotic analysis and numerical experiments showing that the novel IMEX Active Flux scheme is without any modification AP for the u component and AP (for all variables) if the IMEX Runge-Kutta scheme is additionally GSA.

3. The development and (theoretical and numerical) analysis of improved AP IMEX Active Flux schemes with alternative update equations of the point values, which, in particular, are also AA.

The remainder of the paper is organised as follows: In Section 2, the standard time-explicit Active Flux is recalled. In Section 3, a closely related scheme is derived by individually upwinding the nonlinear and stiff part of the HKdV equations. After that, a formal asymptotic analysis for this scheme discretised in time with IMEX Runge Kutta schemes of type ARS is conducted in Section 4. In Section 5, alternative schemes are presented and analysed. Lastly, numerical results are presented in Section 6 and the findings of this work are concluded in Section 7.

2 Active Flux scheme with coupled upwinding

We integrate the HKdV equations (4) over the cell $\mathcal{C}_i$ to obtain

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

The linear source term allows us to interchange the integral and matrix S. Based on this equation, the time-continuous update equation for the cell averages is

$$\frac{\mathrm{d}}{\mathrm{d}t}Q_i(t) = -\frac{f(q_{i+\frac{1}{2}}(t))) - f(q_{i-\frac{1}{2}}(t)))}{\Delta x} + SQ_i(t). \tag{12}$$

To find an update equation for the point values, we consider the quasi-linear form (3) of the HKdV equations and approximate the spatial derivatives by finite difference formulas. We choose them to be high-order approximations of the derivatives at the location of the point values. Such formulas can be

obtained by differentiating the reconstruction. 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

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

and

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

Having two different derivatives is not a disadvantage since it allows us to include upwinding. We diagonalise the Jacobian F of f, $F(q) = R(q)\Lambda(q)R^{-1}(q)$ with $q \in \mathbb{R}^3$. $\Lambda(q)$ is a diagonal matrix with the eigenvalues (5) on the diagonal. Next, we define $F^{\pm}(q) = R(q)\Lambda^{\pm}(q)R^{-1}(q)$ with $\Lambda^{\pm}(q) = \mathrm{diag}(\lambda_1^{\pm}(q), \lambda_2^{\pm}(q), \lambda_3^{\pm}(q))$, where $\lambda_i^+(q) = \max(\lambda_i(q), 0)$ and $\lambda_i^-(q) = \min(\lambda_i(q), 0)$. The positive and negative parts of the Jacobian F are:

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

and

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

The update equation of the point values is

$$\frac{\mathrm{d}}{\mathrm{d}t} q_{i+\frac{1}{2}}(t) = -\left(F^+(q_{i+\frac{1}{2}}(t))D^+ q_{i+\frac{1}{2}}(t) + F^-(q_{i+\frac{1}{2}}(t))D^- q_{i+\frac{1}{2}}(t)\right) + S q_{i+\frac{1}{2}}(t), \tag{13}$$

Observe that the Jacobian F contains nonlinear and stiff linear terms. To distinguish the Active Flux scheme (12), (13) from other schemes developed in this work, it is referred to as the Active Flux scheme with combined upwinding or short the `combined` scheme.

The `combined` scheme can be discretised in time explicitly, for example, by a three-stage, third-order strong stability preserving Runge-Kutta (SSPRK3) scheme [37]. In this case, the time step size $\Delta t = t^{n+1} - t^n$ is subject to a CFL condition

$$\Delta t = C_{\mathrm{CFL}} \frac{\Delta x}{\lambda_{\max}}, \qquad \lambda_{\max} = \max_{i\in\{1,2,3\}} \max_{q\in\Gamma} |\lambda_i(q)|,$$

where Γ is the range of the solution q at time t^n. In practice, Γ is defined as the set of all point values at time step t^n. The eigenvalues depend on ε. As $\varepsilon \to 0$, $|\lambda_i(q)| \to \infty$ for any $q \in \mathbb{R}^3$ and as a result $\Delta t \to 0$ in order for the scheme to remain stable, i.e. the scheme is not asymptotically stable and inefficient for small values of ε. Asymptotic stability can be obtained by discretising in time with IMEX Runge-Kutta schemes. Therefore, the update equations (12) and (13) have to be split into a nonsingular and a linear singular part. The update equation of the cell averages (12) is in a form that makes splitting easily possible. It is more complicated to do the same for the update equation of the point values (13) due to the presence of the square roots.

3 Active Flux scheme with classical individual upwinding

Inspired by [12], we follow a different approach to obtain an Active Flux scheme with update equations that are split into a nonsingular and a linear singular part. First, we split the spatial derivative in (3) into the two parts and then upwind them independently from each other. We write (4) as

$$\frac{\partial}{\partial t} q(t,x) + \frac{\partial}{\partial x} f_1(q(t,x)) + \frac{\partial}{\partial x} f_2(q(t,x)) = S q(t,x), \tag{14}$$

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

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

The function f_1 is independent of ε and f_2 is linear. In quasi-linear form (14) reads

$$\frac{\partial}{\partial t}q(t,x) + F_1(q(t,x))\frac{\partial}{\partial x}q(t,x) + F_2\frac{\partial}{\partial x}q(t,x) = Sq(t,x), \tag{15}$$

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

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

are the Jacobians of f_1 and f_2. The Jacobian F_2 is independent of q. Hence, we dropped the dependence on it. Using the classical approach based on Jacobian splitting, we can determine the positive and negative parts of the two Jacobians. Starting with F_1, we have

$$F_1^+(q) = \begin{pmatrix} \max(u,0) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad F_1^-(q) = \begin{pmatrix} \min(u,0) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{16}$$

The positive and negative parts of F_2 are:

$$F_2^+ = \begin{pmatrix} \frac{1}{2\varepsilon} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} & 0 & \frac{1}{2\varepsilon} \end{pmatrix} \qquad \text{and} \qquad F_2^- = \begin{pmatrix} -\frac{1}{2\varepsilon} & 0 & \frac{1}{2} \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & -\frac{1}{2\varepsilon} \end{pmatrix}. \tag{17}$$

We define the update equations of the point values to be

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}q_{i+\frac{1}{2}}(t) = &-\left(F_1^+(q_{i+\frac{1}{2}}(t))D^+q_{i+\frac{1}{2}}(t) + F_1^-(q_{i+\frac{1}{2}}(t))D^-q_{i+\frac{1}{2}}(t)\right) \\ &-\left(F_2^+D^+q_{i+\frac{1}{2}}(t) + F_2^-D^-q_{i+\frac{1}{2}}(t)\right) + Sq_{i+\frac{1}{2}}(t) \end{aligned} \tag{18}$$

These are not the same update equations as before since F_1 and F_2 are upwinded independently of each other. However, the resulting update equations are related: Expanding the square roots appearing in $F^\pm$ as power series in ε one formally finds

$$F^+(q) = \begin{pmatrix} \frac{1}{2\varepsilon} + \frac{u}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} + \frac{1}{4\varepsilon}u & 0 & \frac{1}{2\varepsilon} \end{pmatrix} + \mathcal{O}(\varepsilon) \quad \text{and} \quad F^-(q) = \begin{pmatrix} -\frac{1}{2\varepsilon} + \frac{u}{2} & 0 & \frac{1}{2} \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} - \frac{1}{4\varepsilon}u & 0 & -\frac{1}{2\varepsilon} \end{pmatrix} + \mathcal{O}(\varepsilon). \tag{19}$$

Comparing (19) to $F_1^\pm(q) + F_2^\pm$, we observe that they differ by the additional terms $\pm\frac{1}{4\varepsilon}u$ in the third entry of the first row and by the $\mathcal{O}(\varepsilon)$ terms.

From the integral form (14), we deduce in the same way as before the update equation of the cell averages:

$$\frac{\mathrm{d}}{\mathrm{d}t}Q_i(t) = -\frac{1}{\Delta x}\left(f_1(q_{i+\frac{1}{2}}(t)) - f_1(q_{i-\frac{1}{2}}(t))\right) - \frac{1}{\Delta x}\left(f_2(q_{i+\frac{1}{2}}(t)) - f_2(q_{i-\frac{1}{2}}(t))\right) + SQ_i(t) \tag{20}$$

This is (12) split into a linear part and a nonlinear part, the latter being independent of the relaxation parameter.

Equations (18) and (20) form what will be called the Active Flux scheme with classical individual upwinding or short the `classical` scheme. It allows us to discretise in time with IMEX Runge-Kutta schemes. However, it is still possible to discretise (18) and (20) explicitly in time.

4 Formal asymptotic analysis

Next, we analyse whether the `classical` scheme discretised in time with an IMEX Runge-Kutta scheme (11) of type ARS is AP. To this end, we perform a formal asymptotic analysis, following the approaches of [21] and [12], and using Mathematica [39] to assist in the analysis. We denote the point values and cell averages at time step t^n, $n \in \mathbb{N}_0$, by

$$q_{i+\frac{1}{2}}^n = \begin{pmatrix} u_{i+\frac{1}{2}}^n \\ \rho_{i+\frac{1}{2}}^n \\ \psi_{i+\frac{1}{2}}^n \end{pmatrix} \qquad \text{and} \qquad Q_i^n = \begin{pmatrix} U_i^n \\ P_i^n \\ \Psi_i^n \end{pmatrix}$$

with $i \in \mathbb{Z}$. The vector of stage solutions is defined as

$$\overline{\rho}_{i+\frac{1}{2}} = \begin{pmatrix} \rho_{i+\frac{1}{2}}^{[1]} \\ \vdots \\ \rho_{i+\frac{1}{2}}^{[s]}, \end{pmatrix}. \tag{21}$$

and $\overline{\psi}_{i+\frac{1}{2}}, \overline{P}_i$, and $\overline{\Psi}_i$ in a similar way. In addition, the temporal step size is denoted by $\Delta t = t^{n+1} - t^n$. The time discrete update equations for the cell averages are

$$\begin{aligned} U_i^{[k]} &= U_i^n + \Delta t \sum_{j=1}^{k} \tilde{a}_{k+1,j} \left(-\frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{[j-1]} \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{[j-1]} \right)^2 \right) \right) \\ &\quad + \Delta t \sum_{j=1}^{k} a_{k,j} \left(-\frac{1}{\Delta x} \left(\psi_{i+\frac{1}{2}}^{[j]} - \psi_{i-\frac{1}{2}}^{[j]} \right) \right), \end{aligned} \tag{22a}$$

$$\begin{aligned} U_i^{n+1} &= U_i^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\frac{1}{\Delta x} \left(\frac{1}{2} \left(u_{i+\frac{1}{2}}^{[j-1]} \right)^2 - \frac{1}{2} \left(u_{i-\frac{1}{2}}^{[j-1]} \right)^2 \right) \right) \\ &\quad + \Delta t \sum_{j=1}^{s} b_j \left(-\frac{1}{\Delta x} \left(\psi_{i+\frac{1}{2}}^{[j]} - \psi_{i-\frac{1}{2}}^{[j]} \right) \right) \end{aligned} \tag{22b}$$

$$\overline{P}_i = P_i^n \mathbb{I}_s + \Delta t A \left(-\frac{1}{\varepsilon^2} \frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{[j]} + \rho_{i-\frac{1}{2}}^{[j]} \right) - \frac{1}{\varepsilon^2} \Psi_i^{[j]} \right)_{j=1,\dots,s}, \tag{22c}$$

$$P_i^{n+1} = P_i^n \mathbb{I}_s + \Delta t b \cdot \left(-\frac{1}{\varepsilon^2} \frac{1}{\Delta x} \left(-\rho_{i+\frac{1}{2}}^{[j]} + \rho_{i-\frac{1}{2}}^{[j]} \right) - \frac{1}{\varepsilon^2} \Psi_i^{[j]} \right)_{j=1,\dots,s}, \tag{22d}$$

$$\overline{\Psi}_i = \Psi_i^n \mathbb{I}_s + \Delta t A \left(-\frac{1}{\varepsilon^2} \frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{[j]} - u_{i-\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} P_i^{[j]} \right)_{j=1,\dots,s}, \tag{22e}$$

$$\Psi_i^{n+1} = \Psi_i^n \mathbb{I}_s + \Delta t b \cdot \left(-\frac{1}{\varepsilon^2} \frac{1}{\Delta x} \left(u_{i+\frac{1}{2}}^{[j]} - u_{i-\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2} P_i^{[j]} \right)_{j=1,\dots,s} \tag{22f}$$

with $k = 1, \dots s$ and $i \in \mathbb{Z}$. The ones for the point values are

$$\begin{aligned} u_{i+\frac{1}{2}}^{[k]} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^{k} \tilde{a}_{k+1,j} \left(-\max\left(u_{i+\frac{1}{2}}^{[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{[j-1]} - \min\left(u_{i+\frac{1}{2}}^{[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{[j-1]} \right) \\ &\quad + \Delta t \sum_{j=1}^{k} a_{k,j} \left(-\frac{1}{2\varepsilon} \left(D^+ u_{i+\frac{1}{2}}^{[j]} - D^- u_{i+\frac{1}{2}}^{[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{[j]} + D^- \psi_{i+\frac{1}{2}}^{[j]} \right) \right), \end{aligned} \tag{23a}$$

$$\begin{aligned} u_{i+\frac{1}{2}}^{n+1} &= u_{i+\frac{1}{2}}^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\max\left(u_{i+\frac{1}{2}}^{[j-1]}, 0 \right) D^+ u_{i+\frac{1}{2}}^{[j-1]} - \min\left(u_{i+\frac{1}{2}}^{[j-1]}, 0 \right) D^- u_{i+\frac{1}{2}}^{[j-1]} \right) \\ &\quad + \Delta t \sum_{j=1}^{s} b_j \left(-\frac{1}{2\varepsilon} \left(D^+ u_{i+\frac{1}{2}}^{[j]} - D^- u_{i+\frac{1}{2}}^{[j]} \right) - \frac{1}{2} \left(D^+ \psi_{i+\frac{1}{2}}^{[j]} + D^- \psi_{i+\frac{1}{2}}^{[j]} \right) \right) \end{aligned} \tag{23b}$$

$$\overline{\rho}_{i+\frac{1}{2}} = \rho_{i+\frac{1}{2}}^n \mathbb{I}_s + \Delta t A \left(\frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2} \psi_{i+\frac{1}{2}}^{[j]} \right)_{j=1,\dots,s}, \tag{23c}$$

$$\rho_{i+\frac{1}{2}}^{n+1} = \rho_{i+\frac{1}{2}}^{n}\mathbb{I}_s + \Delta t b \cdot \left(\frac{1}{\varepsilon^2} D^- \rho_{i+\frac{1}{2}}^{[j]} - \frac{1}{\varepsilon^2}\psi_{i+\frac{1}{2}}^{[j]} \right)_{j=1,\dots,s}, \tag{23d}$$

$$\begin{aligned}\overline{\psi}_{i+\frac{1}{2}} = \psi_{i+\frac{1}{2}}^{n}\mathbb{I}_s + \Delta t A \Bigg(&- \frac{1}{2\varepsilon^2}\left(D^+ u_{i+\frac{1}{2}}^{[j]} + D^- u_{i+\frac{1}{2}}^{[j]} \right) \\ &- \frac{1}{2\varepsilon}\left(D^+ \psi_{i+\frac{1}{2}}^{[j]} - D^- \psi_{i+\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2}\rho_{i+\frac{1}{2}}^{[j]} \Bigg)_{j=1,\dots,s},\end{aligned} \tag{23e}$$

$$\begin{aligned}\psi_{i+\frac{1}{2}}^{n+1} = \psi_{i+\frac{1}{2}}^{n}\mathbb{I}_s + \Delta t b \cdot \Bigg(&- \frac{1}{2\varepsilon^2}\left(D^+ u_{i+\frac{1}{2}}^{[j]} + D^- u_{i+\frac{1}{2}}^{[j]} \right) \\ &- \frac{1}{2\varepsilon}\left(D^+ \psi_{i+\frac{1}{2}}^{[j]} - D^- \psi_{i+\frac{1}{2}}^{[j]} \right) + \frac{1}{\varepsilon^2}\rho_{i+\frac{1}{2}}^{[j]} \Bigg)_{j=1,\dots,s},\end{aligned} \tag{23f}$$

The vector $\mathbb{I}_s \in \mathbb{R}^s$ in the above equations is defined as $\mathbb{I}_s = (1,\dots,1)^{\mathrm{T}}$. To continue with the formal asymptotic analysis, we assume that the stage solutions for $j = 0,\dots,s$ and the numerical approximations at time t^{n+1} are regular enough such that they can be expanded into power series in ε. For example,

$$u_{i+\frac{1}{2}}^{[j]} = u_{i+\frac{1}{2}}^{(0)[j]} + \varepsilon u_{i+\frac{1}{2}}^{(1)[j]} + \varepsilon^2 u_{i+\frac{1}{2}}^{(2)[j]} + \dots$$

for $j = 0,\dots,s$. We now aim to determine the equations that the terms of order 0 satisfy. Therefore we start by substituting the power series in ε into the time-discrete update equation of the cell averages (22) and ordering the resulting terms by the power of ε. This yields

$$0 = \Delta t A \left(-\frac{1}{\Delta x}\left(-\rho_{i+\frac{1}{2}}^{(0)[j]} + \rho_{i-\frac{1}{2}}^{(0)[j]} \right) - \Psi_i^{(0)[j]} \right)_{j=1,\dots,s}, \tag{24a}$$

$$0 = \Delta t b \cdot \left(-\frac{1}{\Delta x}\left(-\rho_{i+\frac{1}{2}}^{(0)[j]} + \rho_{i-\frac{1}{2}}^{(0)[j]} \right) - \Psi_i^{(0)[j]} \right)_{j=1,\dots,s}, \tag{24b}$$

and

$$0 = \Delta t A \left(-\frac{1}{\Delta x}\left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) + P_i^{(0)[j]} \right)_{j=1,\dots,s}, \tag{25a}$$

$$0 = \Delta t b \cdot \left(-\frac{1}{\Delta x}\left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) + P_i^{(0)[j]} \right)_{j=1,\dots,s}, \tag{25b}$$

and

$$\begin{aligned}U_i^{(0)[k]} = U_i^n + \Delta t \sum_{j=1}^{k} \tilde{a}_{k+1,j} \left(-\frac{1}{\Delta x}\left(\frac{1}{2}\left(u_{i+\frac{1}{2}}^{(0)[j-1]} \right)^2 - \frac{1}{2}\left(u_{i-\frac{1}{2}}^{(0)[j-1]} \right)^2 \right) \right) \\ + \Delta t \sum_{j=1}^{k} a_{k,j} \left(-\frac{1}{\Delta x}\left(\psi_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i-\frac{1}{2}}^{(0)[j]} \right) \right),\end{aligned} \tag{26a}$$

$$\begin{aligned}U_i^{(0)n+1} = U_i^n + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(-\frac{1}{\Delta x}\left(\frac{1}{2}\left(u_{i+\frac{1}{2}}^{(0)[j-1]} \right)^2 - \frac{1}{2}\left(u_{i-\frac{1}{2}}^{(0)[j-1]} \right)^2 \right) \right) \\ + \Delta t \sum_{j=1}^{s} b_j \left(-\frac{1}{\Delta x}\left(\psi_{i+\frac{1}{2}}^{(0)[j]} - \psi_{i-\frac{1}{2}}^{(0)[j]} \right) \right)\end{aligned} \tag{26b}$$

We use the invertibility of A to simplify (24a) and (25a) and obtain that the approximations of order 0 satisfy

$$\Psi_i^{(0)[j]} = \frac{1}{\Delta x}\left(\rho_{i+\frac{1}{2}}^{(0)[j]} - \rho_{i-\frac{1}{2}}^{(0)[j]} \right) \tag{27}$$

and

$$P_i^{(0)[j]} = \frac{1}{\Delta x}\left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]} \right) \tag{28}$$

for all $j = 1, \dots s$. Equations (27) and (28) reassuringly look like discretisations of (7b) and (7c). However, they cannot be used in (26) to conclude that the latter is a discretisation of (7a), since (26) contains the point values $\psi_{i+\frac{1}{2}}$, and not the averages. We thus need to repeat the process for the update equations (23) of the point values before being able to proceed.

Inserting the power series into the time-discrete update equations of the point values (23) and ordering the resulting terms by the power of ε gives

$$0 = \Delta t A \left(D^- \rho^{(0)[j]}_{i+\frac{1}{2}} - \psi^{(0)[j]}_{i+\frac{1}{2}} \right)_{j=1,\dots,s}, \tag{29a}$$

$$0 = \Delta t b \cdot \left(D^- \rho^{(0)[j]}_{i+\frac{1}{2}} - \psi^{(0)[j]}_{i+\frac{1}{2}} \right)_{j=1,\dots,s}, \tag{29b}$$

and

$$0 = \Delta t A \left(-\frac{1}{2} \left(D^+ u^{(0)[j]}_{i+\frac{1}{2}} + D^- u^{(0)[j]}_{i+\frac{1}{2}} \right) + \rho^{(0)[j]}_{i+\frac{1}{2}} \right)_{j=1,\dots,s}, \tag{30a}$$

$$0 = \Delta t b \cdot \left(-\frac{1}{2} \left(D^+ u^{(0)[j]}_{i+\frac{1}{2}} + D^- u^{(0)[j]}_{i+\frac{1}{2}} \right) + \rho^{(0)[j]}_{i+\frac{1}{2}} \right)_{j=1,\dots,s} \tag{30b}$$

and

$$\begin{aligned} u^{[k](0)}_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}} &+ \Delta t \sum_{j=1}^{k} \tilde{a}_{k+1,j} \left(- \max \left((u^{(0)[j-1]}_{i+\frac{1}{2}}, 0 \right) D^+ u^{(0)[j-1]}_{i+\frac{1}{2}} \right. \\ &\left. - \min \left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0 \right) D^- u^{(0)[j-1]}_{i+\frac{1}{2}} \right) \\ &+ \Delta t \sum_{j=1}^{k} a_{k,j} \left(-\frac{1}{2} \left(D^+ u^{(1)[j]}_{i+\frac{1}{2}} - D^- u^{(1)[j]}_{i+\frac{1}{2}} \right) - \frac{1}{2} \left(D^+ \psi^{(0)[j]}_{i+\frac{1}{2}} + D^- \psi^{(1)[j]}_{i+\frac{1}{2}} \right) \right), \end{aligned} \tag{31a}$$

$$\begin{aligned} u^{n+1(0)}_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}} &+ \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(- \max \left(u^{[j-1](0)}_{i+\frac{1}{2}}, 0 \right) D^+ u^{[j-1](0)}_{i+\frac{1}{2}} \right. \\ &\left. - \min \left(u^{[j-1](0)}_{i+\frac{1}{2}}, 0 \right) D^- u^{[j-1](0)}_{i+\frac{1}{2}} \right) \\ &+ \Delta t \sum_{j=1}^{s} b_j \left(-\frac{1}{2} \left(D^+ u^{(1)[j]}_{i+\frac{1}{2}} - D^- u^{(1)[j]}_{i+\frac{1}{2}} \right) - \frac{1}{2} \left(D^+ \psi^{(0)[j]}_{i+\frac{1}{2}} + D^- \psi^{(0)[j]}_{i+\frac{1}{2}} \right) \right) \end{aligned} \tag{31b}$$

We can again use the invertibility of A to obtain

$$\psi^{(0)[j]}_{i+\frac{1}{2}} = D^- \rho^{(0)[j]}_{i+\frac{1}{2}} \tag{32}$$

and

$$\rho^{(0)[j]}_{i+\frac{1}{2}} = \frac{1}{2} \left(D^+ u^{(0)[j]}_{i+\frac{1}{2}} + D^- u^{(0)[j]}_{i+\frac{1}{2}} \right) \tag{33}$$

for $i = 1 \dots, s$ from (29a) and (30a).

Lemma 4.1. *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*

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

This implies

$$\frac{1}{2} \left(D^+ v_{i+\frac{1}{2}} + D^- v_{i+\frac{1}{2}} \right) = \frac{1}{\Delta x} \left(v_{i-\frac{1}{2}} - 3 V_i + 3 V_{i+1} - v_{i+\frac{3}{2}} \right), \tag{34a}$$

$$\frac{1}{2} \left(D^+ v_{i+\frac{1}{2}} - D^- v_{i+\frac{1}{2}} \right) = \frac{1}{\Delta x} \left(v_{i-\frac{1}{2}} - 3 V_i + 4 v_{i+\frac{1}{2}} - 3 V_{i+1} + v_{i+\frac{3}{2}} \right). \tag{34b}$$

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

Lemma 4.2. *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

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

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

Proof. Using the Equations (27), (28), (32), and (33) as well as Lemma 4.1, we obtain

$$\begin{aligned}\psi_{i+\frac{1}{2}}^{(0)[j]} &\overset{(32)}{=} D^-\rho_{i+\frac{1}{2}}^{(0)[j]} \\ &= \frac{1}{\Delta x}\left(-4\rho_{i+\frac{1}{2}}^{(0)[j]} + 6P_{i+1}^{(0)[j]} - 2\rho_{i+\frac{3}{2}}^{(0)[j]}\right) \\ &\overset{(33),(28)}{=} \frac{1}{\Delta x}\left(-2\left(D^+u_{i+\frac{1}{2}}^{(0)[j]} + D^-u_{i+\frac{1}{2}}^{(0)[j]}\right) + \frac{6}{\Delta x}\left(u_{i+\frac{3}{2}}^{(0)[j]} - u_{i+\frac{1}{2}}^{(0)[j]}\right) - \left(D^+u_{i+\frac{3}{2}}^{(0)[j]} + D^-u_{i+\frac{3}{2}}^{(0)[j]}\right)\right) \\ &= \frac{2}{\Delta x^2}\left(-2u_{i-\frac{1}{2}}^{(0)[j]} + 6U_i^{(0)[j]} - 4u_{i+\frac{1}{2}}^{(0)[j]} - 3U_{i+1}^{(0)[j]} + 5u_{i+\frac{3}{2}}^{(0)[j]} - 3U_{i+2}^{(0)[j]} + u_{i+\frac{5}{2}}^{(0)[j]}\right).\end{aligned}$$

In a similar way, using Lemma 4.1, we can deduce

$$\begin{aligned}\frac{1}{2}\left(D^+\psi_{i+\frac{1}{2}}^{(0)[j]} + D^-\psi_{i+\frac{1}{2}}^{(0)[j]}\right) &= \frac{1}{\Delta x}\left(\psi_{i-\frac{1}{2}}^{(0)[j]} - 3\Psi_i^{(0)[j]} + 3\Psi_{i+1}^{(0)[j]} - \psi_{i+\frac{3}{2}}^{(0)[j]}\right) \\ &\overset{(32),(27)}{=} \frac{1}{\Delta x}\left(D^-\rho_{i-\frac{1}{2}}^{(0)[j]} - \frac{3}{\Delta x}\left(\rho_{i+\frac{1}{2}}^{(0)[j]} - \rho_{i-\frac{1}{2}}^{(0)[j]}\right) + \frac{3}{\Delta x}\left(\rho_{i+\frac{3}{2}}^{(0)[j]} - \rho_{i+\frac{1}{2}}^{(0)[j]}\right) - D^-\rho_{i+\frac{3}{2}}^{(0)[j]}\right) \\ &= \frac{1}{\Delta x^2}\left(-\rho_{i-\frac{1}{2}}^{(0)[j]} + 6P_i^{(0)[j]} - 8\rho_{i+\frac{1}{2}}^{(0)[j]} + 7\rho_{i+\frac{3}{2}}^{(0)[j]} - 6P_{i+2}^{(0)[j]} + 2\rho_{i+\frac{5}{2}}^{(0)[j]}\right) \\ &\overset{(33),(28)}{=} \frac{1}{\Delta x^2}\Bigg(-\frac{1}{2}\left(D^+u_{i-\frac{1}{2}}^{(0)[j]} + D^-u_{i-\frac{1}{2}}^{(0)[j]}\right) + \frac{6}{\Delta x}\left(u_{i+\frac{1}{2}}^{(0)[j]} - u_{i-\frac{1}{2}}^{(0)[j]}\right) - 4\left(D^+u_{i+\frac{1}{2}}^{(0)[j]} + D^-u_{i+\frac{1}{2}}^{(0)[j]}\right) \\ &\qquad + \frac{7}{2}\left(D^+u_{i+\frac{3}{2}}^{(0)[j]} + D^-u_{i+\frac{3}{2}}^{(0)[j]}\right) - \frac{6}{\Delta x}\left(u_{i+\frac{5}{2}}^{(0)[j]} - u_{i+\frac{3}{2}}^{(0)[j]}\right) + \left(D^+u_{i+\frac{5}{2}}^{(0)[j]} + D^-u_{i+\frac{5}{2}}^{(0)[j]}\right)\Bigg) \\ &= \frac{1}{\Delta x^3}\Big(-u_{i-\frac{3}{2}}^{(0)[j]} + 3U_{i-1}^{(0)[j]} - 14u_{i-\frac{1}{2}}^{(0)[j]} + 21U_i^{(0)[j]} + 14u_{i+\frac{1}{2}}^{(0)[j]} \\ &\quad - 45U_{i+1}^{(0)[j]} + 16u_{i+\frac{3}{2}}^{(0)[j]} + 15U_{i+2}^{(0)[j]} - 13u_{i+\frac{5}{2}}^{(0)[j]} + 6U_{i+3}^{(0)[j]} - 2u_{i+\frac{7}{2}}^{(0)[j]}\Big).\end{aligned}$$

□

Lemma 4.2 implies that (31) can be understood as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}u_{i+\frac{1}{2}}^{(0)}(t) = &- \Bigg(\max\left(u_{i+\frac{1}{2}}^{(0)}(t), 0\right)\frac{1}{\Delta x}\left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t)\right) \\ &+ \min\left(u_{i+\frac{1}{2}}^{(0)}(t), 0\right)\frac{1}{\Delta x}\left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t)\right)\Bigg) \\ &- \frac{1}{\Delta x^3}\Bigg(- u_{i-\frac{3}{2}}^{(0)}(t) + 3U_{i-1}^{(0)}(t) - 14u_{i-\frac{1}{2}}^{(0)}(t) + 21U_i^{(0)}(t) + 14u_{i+\frac{1}{2}}^{(0)}(t) \\ &- 45U_{i+1}^{(0)}(t) + 16u_{i+\frac{3}{2}}^{(0)}(t) + 15U_{i+2}^{(0)}(t) - 13u_{i+\frac{5}{2}}^{(0)}(t) + 6U_{i+3}^{(0)}(t) - 2u_{i+\frac{7}{2}}^{(0)}(t)\Bigg) \\ &- \frac{1}{\Delta x}\left(u_{i-\frac{1}{2}}^{(1)}(t) - 3U_i^{(1)}(t) + 4u_{i+\frac{1}{2}}^{(1)}(t) - 3U_{i+1}^{(1)}(t) + u_{i+\frac{3}{2}}^{(1)}(t)\right)\end{aligned} \tag{35}$$

discretised using an IMEX Runge-Kutta scheme and that, similarly, (26) is the temporal discretisation of

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}U_i^{(0)}(t) = &-\frac{1}{\Delta x}\left(\frac{1}{2}\left(u_{i+\frac{1}{2}}^{(0)}(t)\right)^2-\frac{1}{2}\left(u_{i-\frac{1}{2}}^{(0)}(t)\right)^2\right)\\ &-\frac{2}{\Delta x^3}\left(\left(-2u_{i-\frac{1}{2}}^{(0)}(t)+6U_i^{(0)}(t)-4u_{i+\frac{1}{2}}^{(0)}(t)-3U_{i+1}^{(0)}(t)+5u_{i+\frac{3}{2}}^{(0)}(t)\right.\right.\\ &\left.-3U_{i+2}^{(0)}(t)+u_{i+\frac{5}{2}}^{(0)}(t)\right)-\left(-2u_{i-\frac{3}{2}}^{(0)}(t)+6U_{i-1}^{(0)}(t)-4u_{i-\frac{1}{2}}^{(0)}(t)-3U_i^{(0)}(t)\right.\\ &\left.\left.+5u_{i+\frac{1}{2}}^{(0)}(t)-3U_{i+1}^{(0)}(t)+u_{i+\frac{3}{2}}^{(0)}(t)\right)\right)\end{aligned} \tag{36}$$

with an error in $\mathcal{O}(\Delta t^p)$, where p denotes the order of the IMEX Runge-Kutta scheme. Observe how we were able to recover time-continuous update equations for u (point values and cell averages) that do not contain point values or cell averages of ρ and ψ. If the spatial discretisations appearing therein are approximations of the spatial derivatives in (7a), then the `classical` scheme (discretised in time with an IMEX-Runge Kutta scheme of type ARS) is AP for the component u. For the scheme to be AP for all components, we need to recover discretisations of (7b) and (7c). Equations (27), (28), (32), and (33) look like discretisations of (7b) and (7c) but they only hold for stage solutions and not for the approximation at the time steps. If at this point, we assume that the IMEX Runge-Kutta scheme is GSA, then the last stage solution is equal to the numerical approximation at the time step t^{n+1}. Using Lemma 4.2 once more, we then obtain

$$\rho_{i+\frac{1}{2}}^{(0)n+1} = \frac{1}{\Delta x}\left(u_{i-\frac{1}{2}}^{(0)n+1}-3U_i^{(0)n+1}+3U_{i+1}^{(0)n+1}-u_{i+\frac{3}{2}}^{(0)n+1}\right) \tag{37}$$

and

$$\begin{aligned}\psi_{i+\frac{1}{2}}^{(0)n+1} = \frac{2}{\Delta x^2}&\left(-2u_{i-\frac{1}{2}}^{(0)n+1}+6U_i^{(0)n+1}-4u_{i+\frac{1}{2}}^{(0)n+1}\right.\\ &\left.-3U_{i+1}^{(0)n+1}+5u_{i+\frac{3}{2}}^{(0)n+1}-3U_{i+2}^{(0)n+1}+u_{i+\frac{5}{2}}^{(0)n+1}\right).\end{aligned} \tag{38}$$

and

$$P_i^{(0)n+1} = \frac{1}{\Delta x}\left(u_{i+\frac{1}{2}}^{(0)n+1}-u_{i-\frac{1}{2}}^{(0)n+1}\right) \tag{39}$$

and

$$\begin{aligned}\Psi_i^{(0)n+1} = \frac{1}{\Delta x^2}&\left(\left(u_{i-\frac{1}{2}}^{(0)n+1}-3U_i^{(0)n+1}+3U_{i+1}^{(0)n+1}-u_{i+\frac{3}{2}}^{(0)n+1}\right)\right.\\ &\left.-\left(u_{i-\frac{3}{2}}^{(0)n+1}-3U_{i-1}^{(0)n+1}+3U_i^{(0)n+1}-u_{i+\frac{1}{2}}^{(0)n+1}\right)\right).\end{aligned} \tag{40}$$

These four equations are temporal discretisations of the following time-continuous equations:

$$\rho_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{\Delta x}\left(u_{i-\frac{1}{2}}^{(0)}(t)-3U_i^{(0)}(t)+3U_{i+1}^{(0)}(t)-u_{i+\frac{3}{2}}^{(0)}(t)\right) \tag{41}$$

and

$$\psi_{i+\frac{1}{2}}^{(0)}(t) = \frac{2}{\Delta x^2}\left(-2u_{i-\frac{1}{2}}^{(0)}(t)+6U_i^{(0)}(t)-4u_{i+\frac{1}{2}}^{(0)}(t)-3U_{i+1}^{(0)}(t)\ +5u_{i+\frac{3}{2}}^{(0)}(t)-3U_{i+2}^{(0)}(t)+u_{i+\frac{5}{2}}^{(0)}(t)\right) \tag{42}$$

and

$$P_i^{(0)}(t) = \frac{1}{\Delta x}\left(u_{i+\frac{1}{2}}^{(0)}(t)-u_{i-\frac{1}{2}}^{(0)}(t)\right) \tag{43}$$

and

$$\begin{aligned}\Psi_i^{(0)}(t) = \frac{1}{\Delta x^2}&\left(\left(u_{i-\frac{1}{2}}^{(0)}(t)-3U_i^{(0)}(t)+3U_{i+1}^{(0)}(t)-u_{i+\frac{3}{2}}^{(0)}(t)\right)\right.\\ &\left.-\left(u_{i-\frac{3}{2}}^{(0)}(t)-3U_{i-1}^{(0)}(t)+3U_i^{(0)}(t)-u_{i+\frac{1}{2}}^{(0)}(t)\right)\right).\end{aligned} \tag{44}$$

If (41) and (43) are (spatial) discretisations of (7b), and if (42) and (44) are discretisations of (7c), then the scheme is AP. This is indeed the case and the next theorem establishes the AP property.

Theorem 4.3. *Assume $u^{(0)}, u^{(1)} : \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}$ are regular functions with bounded derivatives that interpolate the point values and cell averages. Under this assumption,* (35) *and* (36) *are, respectively, discretisations of*

$$\frac{\partial}{\partial t}u^{(0)}(t,x) + u^{(0)}(t,x)\frac{\partial}{\partial x}u^{(0)}(t,x) + \frac{\partial^3}{\partial x^3}u^{(0)}(t,x) = 0 \tag{45}$$

and

$$\frac{1}{\Delta x}\int_{\mathcal{C}_i}\left(\frac{\partial}{\partial t}u^{(0)}(t,x) + u^{(0)}(t,x)\frac{\partial}{\partial x}u^{(0)}(t,x) + \frac{\partial^3}{\partial x^3}u^{(0)}(t,x)\right)\mathrm{d}x = 0 \tag{46}$$

with an error $\mathcal{O}(\Delta x)$. In addition, (41) *and* (43) *are, respectively, discretisations of*

$$\rho^{(0)}(t,x) = \frac{\partial}{\partial x}u^{(0)}(t,x) \tag{47}$$

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

$$\frac{1}{\Delta x}\int_{\mathcal{C}_i}\rho^{(0)}(t,x)\mathrm{d}x = \frac{1}{\Delta x}\int_{\mathcal{C}_i}\frac{\partial}{\partial x}u^{(0)}(t,x)\mathrm{d}x. \tag{48}$$

with no discretization error. Equation (42) *is a discretisation of*

$$\psi^{(0)}(t,x) = \frac{\partial^2}{\partial x^2}u^{(0)}(t,x) \tag{49}$$

with an error $\mathcal{O}(\Delta x)$, while (44) *discretises*

$$\frac{1}{\Delta x}\int_{\mathcal{C}_i}\psi^{(0)}(t,x)\mathrm{d}x = \frac{1}{\Delta x}\int_{\mathcal{C}_i}\frac{\partial^2}{\partial x^2}u^{(0)}(t,x)\mathrm{d}x \tag{50}$$

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

Proof. Starting with (35), it follows that

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

and

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

as well as

$$\begin{aligned}
&\frac{1}{\Delta x}\left(u^{(1)}_{i-\frac{1}{2}}(t) - 3U^{(1)}_i(t) + 4u^{(1)}_{i+\frac{1}{2}}(t) - 3U^{(1)}_{i+1}(t) + u^{(1)}_{i+\frac{3}{2}}(t)\right)\\
&\quad = \frac{\Delta x^3}{30}\frac{\partial^4}{\partial x^4}u^{(1)}(t,x_{i+\frac{1}{2}}) + \frac{\Delta x^5}{630}\frac{\partial^6}{\partial x^6}u^{(1)}(t,x_{i+\frac{1}{2}}) + \mathcal{O}\left(\Delta x^7\right).
\end{aligned} \tag{51}$$

Hence, (35) is a spatial discretisation of (45) with an error in $\mathcal{O}(\Delta x)$. For the terms appearing in (36), we obtain

$$\frac{1}{\Delta x}\left(\frac{1}{2}\left(u^{(0)}_{i+\frac{1}{2}}(t)\right)^2 - \frac{1}{2}\left(u^{(0)}_{i-\frac{1}{2}}(t)\right)^2\right) = \frac{1}{\Delta x}\int_{\mathcal{C}_i}u^{(0)}(t,x)\frac{\partial}{\partial x}u^{(0)}(t,x)\mathrm{d}x$$

and

$$\begin{aligned}
&\frac{2}{\Delta x^2}\left(-2u^{(0)}_{i-\frac{1}{2}}(t) + 6U^{(0)}_i(t) - 4u^{(0)}_{i+\frac{1}{2}}(t) - 3U^{(0)}_{i+1}(t) + 5u^{(0)}_{i+\frac{3}{2}}(t) - 3U^{(0)}_{i+2}(t) + u^{(0)(t)}_{i+\frac{5}{2}}\right)\\
&\qquad = \frac{\partial^2}{\partial x^2}u^{(0)}(t,x_{i+\frac{1}{2}}) + \frac{\Delta x}{2}\frac{\partial^3}{\partial x^3}u^{(0)}(t,x_{i+\frac{1}{2}}) + \mathcal{O}\left(\Delta x^2\right),
\end{aligned} \tag{52}$$

which implies

$$\frac{2}{\Delta x^3}\Bigg(\Bigg(-2u^{(0)}_{i-\frac{1}{2}}(t)+6U^{(0)}_i(t)-4u^{(0)}_{i+\frac{1}{2}}(t)-3U^{(0)}_{i+1}(t)+5u^{(0)}_{i+\frac{3}{2}}(t)$$
$$-3U^{(0)}_{i+2}(t)+u^{(0)}_{i+\frac{5}{2}}(t)\Bigg)-\Big(-2u^{(0)}_{i-\frac{3}{2}}(t)+6U^{(0)}_{i-1}(t)-4u^{(0)}_{i-\frac{1}{2}}(t)-3U^{(0)}_i(t)$$
$$+5u^{(0)}_{i+\frac{1}{2}}(t)-3U^{(0)}_{i+1}(t)+u^{(0)}_{i+\frac{3}{2}}(t)\Big)\Bigg)=\frac{1}{\Delta x}\left(\frac{\partial^2}{\partial x^2}u^{(0)}(t,x_{i+\frac{1}{2}})-\frac{\partial^2}{\partial x^2}u^{(0)}(t,x_{i-\frac{1}{2}})\right)$$
$$+\frac{1}{2}\left(\frac{\partial^3}{\partial x^3}u^{(0)}(t,x_{i+\frac{1}{2}})-\frac{\partial^3}{\partial x^3}u^{(0)}(t,x_{i-\frac{1}{2}})\right)+\mathcal{O}(\Delta x).$$

In addition, we have

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

where we used the Lipschitz continuity of the derivatives of $u^{(0)}$. Thus, (36) is a spatial discretisation of (46) with an error in $\mathcal{O}(\Delta x)$. Equation (41) is a discretisation of (47) with an error in $\mathcal{O}(\Delta x^2)$, since

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

Furthermore,

$$\frac{1}{\Delta x}\left(u^{(0)}_{i+\frac{1}{2}}(t)-u^{(0)}_{i-\frac{1}{2}}(t)\right)=\frac{1}{\Delta x}\left(u^{(0)}(t,x_{i+\frac{1}{2}})-u^{(0)}(t,x_{i-\frac{1}{2}})\right)=\frac{1}{\Delta x}\int_{\mathcal{C}_i}\frac{\partial}{\partial x}u^{(0)}(t,x)\mathrm{d}x$$

holds and it follows that (43) is a discretisation of (48) with zero discretisation error. From (52) we immediately obtain that (42) is a discretisation of (49) with an error in $\mathcal{O}(\Delta x)$, while

$$\frac{1}{\Delta x^2}\Bigg(\left(u^{(0)}_{i-\frac{1}{2}}(t)-3U^{(0)}_i(t)+3U^{(0)}_{i+1}(t)-u^{(0)}_{i+\frac{3}{2}}(t)\right)-\left(u^{(0)}_{i-\frac{3}{2}}(t)-3U^{(0)}_{i-1}(t)+3U^{(0)}_i(t)-u^{(0)}_{i+\frac{1}{2}}(t)\right)\Bigg)$$
$$=\frac{1}{\Delta x}\int_{\mathcal{C}_i}\frac{\partial^2}{\partial x^2}u^{(0)}(t,x)\mathrm{d}x+\mathcal{O}(\Delta x^2),$$

gives that (44) is a discretisation of (50) with an error in $\mathcal{O}(\Delta x^2)$. □

To sum up the findings: The formal asymptotic analysis indicates that the `classical` scheme discretised in time with an IMEX Runge-Kutta scheme of type ARS is AP for the component u. If, in addition, the IMEX Runge-Kutta scheme is GSA, then it indicates that the scheme is AP. Moreover, the analysis suggests that the truncation error of the limit scheme is in $\mathcal{O}(\Delta x)$. Numerically we observe that the scheme is indeed only first order if the relaxation parameter is sufficiently small.

Remark 4.4. The asymptotic analysis yields the additional equation

$$0=\frac{1}{2}\left(D^+u^{(0)[j]}_{i+\frac{1}{2}}-D^-u^{(0)[j]}_{i+\frac{1}{2}}\right)_{j=1,\dots,s} \tag{53}$$

originating from (23a). If the IMEX Runge-Kutta scheme is GSA, this implies

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

which is 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 into whether and how the unnatural condition affects the AP property of `classical` scheme.

5 Alternative Active Flux schemes

Numerical experiments conducted for the `classical` scheme (discretised in time with IMEX Runge-Kutta schemes of type ARS) confirm the AP property, but also indicate that some of the advantageous properties of the scheme vanish as $\varepsilon \to 0$. In particular, the `classical` scheme is not AA, and the discrete energy is not conserved well for small ε. This leads us to develop and analyse several alternative Active Flux schemes with different point value update equations. These schemes differ from the `classical` scheme only in the way the term

$$F_2 \frac{\partial}{\partial x} q(t,x)$$

in (15) is discretised, while all other parts of the scheme remain the same.

5.1 Optimised version of the classical scheme

The first approach to develop Active Flux schemes with improved properties is to optimise the `classical` scheme by modifying the matrices $F_2^{\pm}$. Observe that the terms of factor $\frac{1}{\varepsilon}$ in (17) give rise to expressions of the form (34b) which is a discretisation of

$$\frac{\Delta x^3}{30} \frac{\partial^4}{\partial x^4} v(t, x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^5).$$

(see (51)).These appear in (23a)–(23b) and (23e)–(23f). If we replace the two $\frac{1}{\varepsilon}$ elements on the diagonal of $F_2^{\pm}$ with zeros, (18) is still a discretisation of (15). Numerical experiments show that the resulting scheme is at most second order accurate and performs worse than the `classical` scheme for ε close to 1. However, for sufficiently small ε, this scheme is slightly more accurate than the `classical` scheme. To refine this approach, we only replace the first element in the first row (the "1-1-element") of the matrices $F_2^{\pm}$ by zero, since it is the one responsible for the unnatural condition (53). The 3-3-element of $F_2^{\pm}$ scales with $\frac{1}{\varepsilon}$, while the 3-1-element scales with $\frac{1}{\varepsilon^2}$. As a result, if $\varepsilon \to 0$, the relation between the terms in update equation of the point values changes. By changing the scaling of the 3-3-element to $\frac{1}{\varepsilon^2}$, the relation of the terms in the update equation becomes independent of ε. These arguments will become clearer below when the asymptotic analysis is performed. Following this new approach, the splitting of the Jacobian now reads

$$F^+_{2,\text{optimised}} = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2\varepsilon^2} & 0 & \frac{1}{2\varepsilon^2} \end{pmatrix} \qquad \text{and} \qquad F^-_{2,\text{optimised}} = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & -\frac{1}{\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & -\frac{1}{2\varepsilon^2} \end{pmatrix}. \tag{54}$$

Observe that the eigenvalues of $F^{\pm}_{2,\text{optimised}}$ no longer have a definite sign. However, ad hoc modifications of the upwinding matrices are very successful in restoring the correct asymptotics while retaining stability in other situations already, e.g. as "low Mach fixes" in the context of the incompressible limit of the Euler equations, see e.g. [34]. We call the resulting scheme the Active Flux scheme with optimised individual upwinding or short the `optimised` scheme.

Performing a formal asymptotic analysis for the `optimised` scheme, discretised in time with an IMEX Runge-Kutta scheme of type ARS, allows us to obtain (26), (27), and (28) for the terms of order 0, since the update equation of the cell averages remains unchanged. In addition, we get (32) (since (23c)–(23d) and thus (29a)–(29b) remain valid) as well as

$$\begin{aligned} u^{[k](0)}_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}} + \Delta t \sum_{j=1}^{k} \tilde{a}_{k+1,j} \Big(&- \max\left((u^{(0)[j-1]}_{i+\frac{1}{2}}, 0\right) D^+ u^{(0)[j-1]}_{i+\frac{1}{2}} \\ &- \min\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0\right) D^- u^{(0)[j-1]}_{i+\frac{1}{2}} \Big) + \Delta t \sum_{j=1}^{k} a_{k,j} \left(-\frac{1}{2} \left(D^+ \psi^{(0)[j]}_{i+\frac{1}{2}} + D^- \psi^{(0)[j]}_{i+\frac{1}{2}} \right) \right), \end{aligned} \tag{55a}$$

$$\begin{aligned} u^{n+1(0)}_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}} + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \Big(&- \max\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0\right) D^+ u^{(0)[j-1]}_{i+\frac{1}{2}} \\ &- \min\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0\right) D^- u^{(0)[j-1]}_{i+\frac{1}{2}} \Big) + \Delta t \sum_{j=1}^{s} b_j \left(-\frac{1}{2} \left(D^+ \psi^{(0)[j]}_{i+\frac{1}{2}} + D^- \psi^{(0)[j]}_{i+\frac{1}{2}} \right) \right) \end{aligned} \tag{55b}$$

and

$$\rho_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2}\left(D^+ u_{i+\frac{1}{2}}^{(0)[j]} + D^- u_{i+\frac{1}{2}}^{(0)[j]}\right) + \frac{1}{2}\left(D^+ \psi_{i+\frac{1}{2}}^{(0)[j]} - D^- \psi_{i+\frac{1}{2}}^{(0)[j]}\right) \tag{56}$$

instead of (31) and (33), respectively. The absence of the $u^{(1)}$-terms in (55) and the presence of the $\psi^{(0)}$-terms in (56) are due to our modification. The unnatural condition (53) is no longer present. It follows in a similar way as before that (55) is a temporal discretisation (using an type ARS IMEX Runge-Kutta scheme) of

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} u_{i+\frac{1}{2}}^{(0)}(t) = & -\Bigg(\max\left(u_{i+\frac{1}{2}}^{(0)}(t), 0\right) \frac{1}{\Delta x}\left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t)\right) \\
& + \min\left(u_{i+\frac{1}{2}}^{(0)}(t), 0\right) \frac{1}{\Delta x}\left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t)\right)\Bigg) \\
& - \frac{1}{\Delta x^3}\Bigg(- u_{i-\frac{3}{2}}^{(0)}(t) + 3U_{i-1}^{(0)}(t) - 14u_{i-\frac{1}{2}}^{(0)}(t) + 21U_i^{(0)}(t) + 14u_{i+\frac{1}{2}}^{(0)}(t) \\
& - 45U_{i+1}^{(0)}(t) + 16u_{i+\frac{3}{2}}^{(0)}(t) + 15U_{i+2}^{(0)}(t) - 13u_{i+\frac{5}{2}}^{(0)}(t) + 6U_{i+3}^{(0)}(t) - 2u_{i+\frac{7}{2}}^{(0)}(t)\Bigg) \\
& - \frac{1}{\Delta x^2}\left(-\Lambda_{i-\frac{1}{2}}^{(0)}(t) - 8\Lambda_{i+\frac{1}{2}}^{(0)}(t) + 7\Lambda_{i+\frac{3}{2}}^{(0)}(t) + 2\Lambda_{i+\frac{5}{2}}^{(0)}(t)\right)
\end{aligned} \tag{57}$$

and (26) is a discretisation of

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} U_i^{(0)}(t) = & -\frac{1}{\Delta x}\left(\frac{1}{2}\left(u_{i+\frac{1}{2}}^{(0)}(t)\right)^2 - \frac{1}{2}\left(u_{i-\frac{1}{2}}^{(0)}(t)\right)^2\right) \\
& - \frac{2}{\Delta x^3}\Bigg(\Bigg(- 2u_{i-\frac{1}{2}}^{(0)}(t) + 6U_i^{(0)}(t) - 4u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + 5u_{i+\frac{3}{2}}^{(0)}(t) \\
& -3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t)\Bigg) - \Big(-2u_{i-\frac{3}{2}}^{(0)}(t) + 6U_{i-1}^{(0)}(t) - 4u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) \\
& +5u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + u_{i+\frac{3}{2}}^{(0)}(t)\Big)\Bigg) \\
& - \frac{1}{\Delta x^2}\left(\left(-4\Lambda_{i+\frac{1}{2}}^{(0)}(t) - 2\Lambda_{i+\frac{3}{2}}^{(0)}(t)\right) - \left(-4\Lambda_{i-\frac{1}{2}}^{(0)}(t) - 2\Lambda_{i+\frac{1}{2}}^{(0)}(t)\right)\right),
\end{aligned} \tag{58}$$

where

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

If the time discretisation scheme is GSA, then we additionally obtain the semi-discretisations (43) and, instead of (44)

$$\begin{aligned}
\Psi_i^{(0)}(t) = & \frac{1}{\Delta x^2}\Bigg(\left(u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) + 3U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t)\right) \\
& - \left(u_{i-\frac{3}{2}}^{(0)}(t) - 3U_{i-1}^{(0)}(t) + 3U_i^{(0)}(t) - u_{i+\frac{1}{2}}^{(0)}(t)\right)\Bigg) \\
& + \frac{1}{\Delta x^2}\Bigg(\left(\psi_{i-\frac{1}{2}}^{(0)}(t) - 3\Psi_i^{(0)}(t) + 4\psi_{i+\frac{1}{2}}^{(0)}(t) - 3\Psi_{i+1}^{(0)}(t) + \psi_{i+\frac{3}{2}}^{(0)}(t)\right) \\
& - \left(\psi_{i-\frac{3}{2}}^{(0)}(t) - 3\Psi_{i-1}^{(0)}(t) + 4\psi_{i-\frac{1}{2}}^{(0)}(t) - 3\Psi_i^{(0)}(t) + \psi_{i+\frac{1}{2}}^{(0)}(t)\right)\Bigg)
\end{aligned} \tag{59}$$

as well as

$$\begin{aligned}
\rho_{i+\frac{1}{2}}^{(0)}(t) = & \frac{1}{\Delta x}\left(u_{i-\frac{1}{2}}^{(0)}(t) - 3U_i^{(0)}(t) + 3U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t)\right) \\
& + \frac{1}{\Delta x}\left(\psi_{i-\frac{1}{2}}^{(0)}(t) - 3\Psi_i^{(0)}(t) + 4\psi_{i+\frac{1}{2}}^{(0)}(t) - 3\Psi_{i+1}^{(0)}(t) + \psi_{i+\frac{3}{2}}^{(0)}(t)\right)
\end{aligned} \tag{60}$$

and

$$\begin{aligned}
\psi_{i+\frac{1}{2}}^{(0)}(t) = & \frac{2}{\Delta x^2}\Big(-2u_{i-\frac{1}{2}}^{(0)}(t) + 6U_i^{(0)}(t) - 4u_{i+\frac{1}{2}}^{(0)}(t) - 3U_{i+1}^{(0)}(t) + 5u_{i+\frac{3}{2}}^{(0)}(t) \\
& -3U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t)\Big) + \frac{1}{\Delta x}\left(-4\Lambda_{i+\frac{1}{2}}^{(0)}(t) - 2\Lambda_{i+\frac{3}{2}}^{(0)}(t)\right)
\end{aligned} \tag{61}$$

instead of (41) and (42), respectively. Assuming $u^{(0)}, \psi^{(0)} : \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}$ are regular enough and interpolate the point values and cell averages, the additional Λ-terms are in $\mathcal{O}(\Delta x^3)$. The difference between (55) and (31) is the term (51). This term is in $\mathcal{O}(\Delta x^3)$. Therefore, a similar result as Theorem 4.3 holds and the formal asymptotic analysis indicates that the `optimised` scheme is AP for the u component. If the IMEX Runge-Kutta scheme is GSA, then the scheme is AP.

For ε close to 1 the `optimised` scheme performs similarly as the `classical` scheme, as will be seen in the experiments of Section 6. However, if $\varepsilon \to 0$, the properties of the `optimised` scheme do not deteriorate. In particular, the scheme achieves third-order accuracy independently of ε, i.e. it is AA, although the asymptotic analysis continues to show presence of terms that are only $\mathcal{O}(\Delta x)$ approximations (e.g. in (57)).

5.2 Central Active Flux schemes

For time-explicit schemes, upwinding is typically used to ensure the stability of the scheme. It is common to use a central discretisation for the term discretised in time implicitly, i.e. in our case we would be thus led to choosing $F_2^{\pm} = \frac{1}{2}F_2$. The resulting matrices read

$$F^+_{2,\text{central}} = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & -\frac{1}{2\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad F^-_{2,\text{central}} = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & -\frac{1}{2\varepsilon^2} & 0 \\ \frac{1}{2\varepsilon^2} & 0 & 0 \end{pmatrix}. \tag{62}$$

We refer to an Active Flux scheme with (62) instead of (17) in (18) as the central Active Flux scheme or short the `central` scheme. A formal asymptotic analysis for the time-discrete `central` scheme yields, besides (26), (27), (28), and (33), also

$$\begin{aligned} u^{[k](0)}_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}} &+ \Delta t \sum_{j=1}^{k} \tilde{a}_{k+1,j} \left(- \max\left((u^{(0)[j-1]}_{i+\frac{1}{2}}, 0 \right) D^+ u^{(0)[j-1]}_{i+\frac{1}{2}} \right. \\ &\left. - \min\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0 \right) D^- u^{(0)[j-1]}_{i+\frac{1}{2}} \right) \\ &+ \Delta t \sum_{j=1}^{k} a_{k,j} \left(-\frac{1}{2} \left(D^+ \psi^{(0)[j]}_{i+\frac{1}{2}} + D^- \psi^{(0)[j]}_{i+\frac{1}{2}} \right) \right), \end{aligned} \tag{63a}$$

$$\begin{aligned} u^{n+1(0)}_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}} &+ \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \left(- \max\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0 \right) D^+ u^{(0)[j-1]}_{i+\frac{1}{2}} \right. \\ &\left. - \min\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0 \right) D^- u^{(0)[j-1]}_{i+\frac{1}{2}} \right) \\ &+ \Delta t \sum_{j=1}^{s} b_j \left(-\frac{1}{2} \left(D^+ \psi^{(0)[j]}_{i+\frac{1}{2}} + D^- \psi^{(0)[j]}_{i+\frac{1}{2}} \right) \right) \end{aligned} \tag{63b}$$

instead of (31) and

$$\psi^{(0)[j]}_{i+\frac{1}{2}} = \frac{1}{2} \left(D^+ \rho^{(0)[j]}_{i+\frac{1}{2}} + D^- \rho^{(0)[j]}_{i+\frac{1}{2}} \right). \tag{64}$$

instead of (32). In particular, the scheme does not give rise to any unnatural conditions (like (53)) for the terms of order 0. We can recover the following time-continuous discretisations:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} u^{(0)}_{i+\frac{1}{2}}(t) = &- \left(\max\left(u^{(0)}_{i+\frac{1}{2}}(t), 0 \right) \frac{1}{\Delta x} \left(2u^{(0)}_{i-\frac{1}{2}}(t) - 6U^{(0)}_i(t) + 4u^{(0)}_{i+\frac{1}{2}}(t) \right) \right. \\ &\left. + \min\left(u^{(0)}_{i+\frac{1}{2}}(t), 0 \right) \frac{1}{\Delta x} \left(-4u^{(0)}_{i+\frac{1}{2}}(t) + 6U^{(0)}_{i+1}(t) - 2u^{(0)}_{i+\frac{3}{2}}(t) \right) \right) \\ &- \frac{1}{\Delta x^3} \left(u^{(0)}_{i-\frac{5}{2}}(t) - 3U^{(0)}_{i-2}(t) + 6u^{(0)}_{i-\frac{3}{2}}(t) - 6U^{(0)}_{i-1}(t) \right. \\ &- 15u^{(0)}_{i-\frac{1}{2}}(t) + 33U^{(0)}_i(t) - 33U^{(0)}_{i+1}(t) + 15u^{(0)}_{i+\frac{3}{2}}(t) \\ &\left. + 6U^{(0)}_{i-1}(t) - 6u^{(0)}_{i+\frac{5}{2}}(t) + 3U^{(0)}_{i+3}(t) - u^{(0)}_{i-\frac{7}{2}}(t) \right) \end{aligned} \tag{65}$$

instead of (35),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}U_i^{(0)}(t) = &-\frac{1}{2\Delta x}\left(\left(u_{i+\frac{1}{2}}^{(0)}(t)\right)^2-\left(u_{i-\frac{1}{2}}^{(0)}(t)\right)^2\right)-\frac{1}{\Delta x^3}\left(\left(u_{i-\frac{3}{2}}^{(0)}(t)-3U_{i-1}^{(0)}(t)\right.\right.\\ &+3u_{i-\frac{1}{2}}^{(0)}(t)+3U_i^{(0)}(t)-8u_{i+\frac{1}{2}}^{(0)}(t)+3U_{i+1}^{(0)}(t)+3u_{i+\frac{3}{2}}^{(0)}(t)\\ &\left.-3U_{i+2}^{(0)}(t)+u_{i+\frac{5}{2}}^{(0)}(t)\right)-\left(u_{i-\frac{5}{2}}^{(0)}(t)-3U_{i-2}^{(0)}(t)+3u_{i-\frac{3}{2}}^{(0)}(t)\right.\\ &\left.\left.+3U_{i-1}^{(0)}(t)-8u_{i-\frac{1}{2}}^{(0)}(t)+3U_i^{(0)}(t)+3u_{i+\frac{1}{2}}^{(0)}(t)-3U_{i+1}^{(0)}(t)+u_{i+\frac{3}{2}}^{(0)}(t)\right)\right).\end{aligned} \tag{66}$$

instead of (36), and

$$\begin{aligned}\psi_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{\Delta x^2}&\left(u_{i-\frac{3}{2}}^{(0)}(t)-3U_{i-1}^{(0)}(t)+3u_{i-\frac{1}{2}}^{(0)}(t)+3U_i^{(0)}(t)-8u_{i+\frac{1}{2}}^{(0)}(t)\right.\\ &\left.+3U_{i+1}^{(0)}(t)+3u_{i+\frac{3}{2}}^{(0)[j]}-3U_{i+2}^{(0)}(t)+u_{i+\frac{5}{2}}^{(0)}(t)\right).\end{aligned} \tag{67}$$

instead of (42) as well as (41), (43), and (44). As before, recovering the latter four requires the assumption that the IMEX Runge-Kutta scheme is GSA. Observe the new, more symmetric discretisations of $\frac{\partial^3}{\partial x^3}u$ and $\frac{\partial^2}{\partial x^2}u$ in (66) and (67), respectively. Similarly to the proof of Theorem 4.3, one can show that (65), (41), and (67) are a discretisation of (45), (47), and (49) with an error in $\mathcal{O}(\Delta x^2)$ and (66), (43), and (44) are a discretisation of (46), (48), and (50) with the same discretisation error. Hence, the `central` scheme is AP for the u component and AP for all components if the IMEX Runge-Kutta scheme is GSA. Numerically this property can be verified. We find that the `central` scheme is second order accurate for both large and small ε, i.e. the scheme is AA since it retains its order of accuracy, but it does not reach the usual third order of Active Flux methods. The present analysis does not offer a conclusive theoretical explanation for this loss of accuracy.

It is possible to modify the central scheme to obtain a third-order AA scheme by considering the following alternatives to $D^{\pm}$ inspired by [22]:

$$D_{\mathrm{mod}}^{+}v_{i+\frac{1}{2}}(t) = \frac{1}{6\Delta x}\left(5v_{i-\frac{1}{2}}(t)-18V_i(t)+8v_{i+\frac{1}{2}}(t)+6V_{i+1}(t)-v_{i+\frac{3}{2}}(t)\right) \tag{68a}$$

$$D_{\mathrm{mod}}^{-}v_{i+\frac{1}{2}}(t) = \frac{1}{6\Delta x}\left(v_{i-\frac{1}{2}}(t)-6V_i(t)-8v_{i+\frac{1}{2}}(t)+18V_{i+1}(t)-5v_{i+\frac{3}{2}}(t)\right). \tag{68b}$$

We replace $D^{\pm}$ by $D_{\mathrm{mod}}^{\pm}$ only in the part that is discretised in time implicitly. We name the scheme in which we replace all $D^{\pm}$ in the implicit part with the new formulas (68) the first modified central Active Flux scheme or `mod1-central` scheme. In fact, we find it sufficient to only make the replacement in the update equations of the point values of ρ and ψ. This scheme is referred to as the second modified central Active Flux scheme or `mod2-central` scheme.

Since the two modified central schemes are closely related, the formal asymptotic analysis of the schemes is almost identical. Following the usual approach, we obtain (26), (27), (28), and

$$\begin{aligned}u_{i+\frac{1}{2}}^{[k](0)} = u_{i+\frac{1}{2}}^{n}&+\Delta t\sum_{j=1}^{k}\tilde{a}_{k+1,j}\left(-\max\left(u_{i+\frac{1}{2}}^{(0)[j-1]},0\right)D^{+}u_{i+\frac{1}{2}}^{(0)[j-1]}\right.\\ &\left.-\min\left(u_{i+\frac{1}{2}}^{(0)[j-1]},0\right)D^{-}u_{i+\frac{1}{2}}^{(0)[j-1]}\right)\\ &+\Delta t\sum_{j=1}^{k}a_{k,j}\left(-\frac{1}{2}\left(D_{\mathrm{mod}}^{+}\psi_{i+\frac{1}{2}}^{[j](0)}+D_{\mathrm{mod}}^{-}\psi_{i+\frac{1}{2}}^{[j](0)}\right)\right),\end{aligned} \tag{69a}$$

$$\begin{aligned}u_{i+\frac{1}{2}}^{n+1(0)} = u_{i+\frac{1}{2}}^{n}&+\Delta t\sum_{j=1}^{\sigma}\tilde{b}_{j}\left(-\max\left(u_{i+\frac{1}{2}}^{[j-1](0)},0\right)D^{+}u_{i+\frac{1}{2}}^{[j-1](0)}\right.\\ &\left.-\min\left(u_{i+\frac{1}{2}}^{[j-1](0)},0\right)D^{-}u_{i+\frac{1}{2}}^{[j-1](0)}\right)\\ &+\Delta t\sum_{j=1}^{s}b_{j}\left(-\frac{1}{2}\left(D_{\mathrm{mod}}^{+}\psi_{i+\frac{1}{2}}^{[j](0)}+D_{\mathrm{mod}}^{-}\psi_{i+\frac{1}{2}}^{[j](0)}\right)\right).\end{aligned} \tag{69b}$$

as well as

$$\rho_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2}\left(D_{\mathrm{mod}}^{+}u_{i+\frac{1}{2}}^{(0)[j]}+D_{\mathrm{mod}}^{-}u_{i+\frac{1}{2}}^{(0)[j]}\right) \tag{70}$$

and

$$\psi_{i+\frac{1}{2}}^{(0)[j]} = \frac{1}{2}\left(D_{\mathrm{mod}}^{+}\rho_{i+\frac{1}{2}}^{(0)[j]} + D_{\mathrm{mod}}^{-}\rho_{i+\frac{1}{2}}^{(0)[j]}\right). \tag{71}$$

for the terms of order 0 of the `mod1-central` scheme. For the `mod2-central` scheme, the analysis yields the same equations with the exception that we get (63) instead of (69). For both schemes unnatural conditions of the kind of (53) do not appear.

The equations for the terms of order 0 of the `mod1-central` scheme allow us to obtain the following semi-discretisations:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}u_{i+\frac{1}{2}}^{(0)}(t) = &-\left(\max\left(u_{i+\frac{1}{2}}^{(0)}(t),0\right)\frac{1}{\Delta x}\left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t)\right)\right.\\ &\left.+\min\left(u_{i+\frac{1}{2}}^{(0)}(t),0\right)\frac{1}{\Delta x}\left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t)\right)\right)\\ &-\frac{1}{8\Delta x^3}\Big(u_{i-\frac{5}{2}}^{(0)}(t) - 4U_{i-2}^{(0)}(t) + 16u_{i-\frac{3}{2}}^{(0)}(t) - 28U_{i-1}^{(0)}(t)\\ &-35u_{i-\frac{1}{2}}^{(0)}(t) + 104U_i^{(0)}(t) - 104U_{i+1}^{(0)}(t) + 35u_{i+\frac{3}{2}}^{(0)}(t)\\ &+28U_{i+2}^{(0)}(t) - 16u_{i+\frac{5}{2}}^{(0)}(t) + 4U_{i+3}^{(0)}(t) - u_{i-\frac{7}{2}}^{(0)}(t)\Big).\end{aligned} \tag{72}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}U_i^{(0)}(t) = &-\frac{1}{\Delta x}\left(\frac{1}{2}\left(u_{i+\frac{1}{2}}^{(0)}(t)\right)^2 - \frac{1}{2}\left(u_{i-\frac{1}{2}}^{(0)}(t)\right)^2\right)\\ &-\frac{1}{4\Delta x^3}\Big(\Big(u_{i-\frac{3}{2}}^{(0)}(t) - 4U_{i-1}^{(0)}(t) + 8u_{i-\frac{1}{2}}^{(0)} + 4U_i^{(0)}(t) - 18u_{i+\frac{1}{2}}^{(0)}(t) + 4U_{i+1}^{(0)}(t)\\ &+8u_{i+\frac{3}{2}}^{(0)}(t) - 4U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t)\Big) - \Big(u_{i-\frac{5}{2}}^{(0)}(t) - 4U_{i-2}^{(0)}(t) + 8u_{i-\frac{3}{2}}^{(0)}\\ &+4U_{-1}^{(0)}(t) - 18u_{i-\frac{1}{2}}^{(0)}(t) + 4U_i^{(0)}(t) + 8u_{i+\frac{1}{2}}^{(0)}(t) - 4U_{i+1}^{(0)}(t) + u_{i+\frac{3}{2}}^{(0)}(t)\Big)\Big).\end{aligned} \tag{73}$$

and under the assumption that the time discretisation is GSA (43) and

$$\rho_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{2\Delta x}\left(u_{i-\frac{1}{2}}^{(0)}(t) - 4U_i^{(0)}(t)) + 4U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t)\right), \tag{74}$$

$$\begin{aligned}\psi_{i+\frac{1}{2}}^{(0)}(t) = \frac{1}{4\Delta x^2}&\left(u_{i-\frac{3}{2}}^{(0)}(t) - 4U_{i-1}^{(0)}(t) + 8u_{i-\frac{1}{2}}^{(0)}(t) + 4U_i^{(0)}(t) - 18u_{i+\frac{1}{2}}^{(0)}(t)\right.\\ &\left.+4U_{i+1}^{(0)}(t) + 8u_{i+\frac{3}{2}}^{(0)}(t) - 4U_{i+2}^{(0)}(t) + u_{i+\frac{5}{2}}^{(0)}(t)\right)\end{aligned} \tag{75}$$

and

$$\begin{aligned}\Psi_i^{(0)}(t) = \frac{1}{\Delta 2x^2}&\left(\left(u_{i-\frac{1}{2}}^{(0)}(t) - 4U_i^{(0)}(t) + 4U_{i+1}^{(0)}(t) - u_{i+\frac{3}{2}}^{(0)}(t)\right)\right.\\ &\left.-\left(u_{i-\frac{3}{2}}^{(0)}(t) - 4U_{i-1}^{(0)}(t) + 4U_i^{(0)}(t) - u_{i+\frac{1}{2}}^{(0)}(t)\right)\right).\end{aligned} \tag{76}$$

In case of the `mod2-central` scheme, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}u_{i+\frac{1}{2}}^{(0)}(t) = &-\left(\max\left(u_{i+\frac{1}{2}}^{(0)}(t),0\right)\frac{1}{\Delta x}\left(2u_{i-\frac{1}{2}}^{(0)}(t) - 6U_i^{(0)}(t) + 4u_{i+\frac{1}{2}}^{(0)}(t)\right)\right.\\ &\left.+\min\left(u_{i+\frac{1}{2}}^{(0)}(t),0\right)\frac{1}{\Delta x}\left(-4u_{i+\frac{1}{2}}^{(0)}(t) + 6U_{i+1}^{(0)}(t) - 2u_{i+\frac{3}{2}}^{(0)}(t)\right)\right)\\ &-\frac{1}{4\Delta x^3}\Big(u_{i-\frac{5}{2}}^{(0)}(t) - 4U_{i-2}^{(0)}(t) + 14u_{i-\frac{3}{2}}^{(0)}(t) - 20U_{i-1}^{(0)}(t)\\ &-31u_{i-\frac{1}{2}}^{(0)}(t) + 80U_i^{(0)}(t) - 80U_{i+1}^{(0)}(t) + 31u_{i+\frac{3}{2}}^{(0)}(t)\\ &+20U_{i+2}^{(0)}(t) - 14u_{i+\frac{5}{2}}^{(0)}(t) + 4U_{i+3}^{(0)}(t) - u_{i-\frac{7}{2}}^{(0)}(t)\Big),\end{aligned} \tag{77}$$

as well as (73), (43), (74), (75), and (76). Equation (72) is a discretisation of (45) with an error in $\mathcal{O}(\Delta x^2)$. The error is due to the discretisation of the nonlinear part, the discretisation error of the third

derivative is in $\mathcal{O}(\Delta x^4)$. In addition, (74), (75), (73), and (76) are discretisation of (47), (49), (46), and (50) respectively with an error in $\mathcal{O}(\Delta x^4)$. As (72), (77) is also a discretisation of (45) with an error in $\mathcal{O}(\Delta x^2)$. In contrast to (72), the error is not only due to the nonlinear part, since the discretisation error of the third derivative is in this case also in $\mathcal{O}(\Delta x^2)$. The results imply that the `mod1-central` and `mod2-central` schemes are AP for the u component. If the time discretisation is GSA, then the schemes are AP (for all components). Numerically we observe that both schemes are third order accurate and AA which justifies the modifications.

5.3 Active Flux scheme with biased individual upwinding

Although the `mod2-central` scheme delivers compelling results in the numerical experiments, we introduce an additional alternative scheme. It performs similarly to, if not better than, the `mod2-central` scheme, but shows that central discretisation is not the only viable approach. Inspired by the spatial discretisation of the HKdV equations presented in [12], we consider another choice of upwinding F_2 and define

$$F^+_{2,\text{biased}} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -\frac{1}{2\varepsilon^2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad F^-_{2,\text{biased}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\frac{1}{2\varepsilon^2} & 0 \\ \frac{1}{\varepsilon^2} & 0 & 0 \end{pmatrix}.$$

u is upwinded from the left, ψ from the right, and ρ centrally. Numerical experiments indicate that using the standard $D^\pm$ operators in the point value update equation of ρ results in a scheme that is only second order. However, using the modified operators $D^\pm_{\text{mod}}$ increases the order of accuracy to three. Due to the imposed directional bias, which is not based on characteristics, we call the resulting scheme the Active Flux scheme with biased individual upwinding or short the `biased` scheme.

By conducting a formal asymptotic analysis for this scheme (discretised in time with a type ARS IMEX Runge-Kutta scheme), we obtain (26), (27), and (28) from the update equation of the cell averages and

$$\begin{aligned} u^{[k](0)}_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}} + \Delta t \sum_{j=1}^{k} \tilde{a}_{k+1,j} \Big(&- \max\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0\right) D^+ u^{(0)[j-1]}_{i+\frac{1}{2}} \\ &- \min\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0\right) D^- u^{(0)[j-1]}_{i+\frac{1}{2}} \Big) + \Delta t \sum_{j=1}^{k} a_{k,j} \left(-D^+ \psi^{(0)[j]}_{i+\frac{1}{2}}\right), \end{aligned} \tag{78a}$$

$$\begin{aligned} u^{n+1(0)}_{i+\frac{1}{2}} = u^n_{i+\frac{1}{2}} + \Delta t \sum_{j=1}^{\sigma} \tilde{b}_j \Big(&- \max\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0\right) D^+ u^{(0)[j-1]}_{i+\frac{1}{2}} \\ &- \min\left(u^{(0)[j-1]}_{i+\frac{1}{2}}, 0\right) D^- u^{(0)[j-1]}_{i+\frac{1}{2}} \Big) + \Delta t \sum_{j=1}^{s} b_j \left(-D^+ \psi^{(0)[j]}_{i+\frac{1}{2}}\right) \end{aligned} \tag{78b}$$

and

$$\psi^{(0)[j]}_{i+\frac{1}{2}} = \frac{1}{2}\left(D^+_{\text{mod}} \rho^{(0)[j]}_{i+\frac{1}{2}} + D^-_{\text{mod}} \rho^{(0)[j]}_{i+\frac{1}{2}}\right) \tag{79}$$

and

$$\rho^{(0)[j]}_{i+\frac{1}{2}} = D^- u^{(0)[j]}_{i+\frac{1}{2}} \tag{80}$$

for all $j = 1, \ldots, s$ from the update equations of the point values. Discretising in space with the `biased` scheme does not yield any unnatural conditions of the kind of (53). As before, it follows that (26) and (78) are temporal discretisations of

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} U^{(0)}_i(t) = &- \frac{1}{\Delta x}\left(\frac{1}{2}\left(u^{(0)}_{i+\frac{1}{2}}(t)\right)^2 - \frac{1}{2}\left(u^{(0)}_{i-\frac{1}{2}}(t)\right)^2\right) \\ &- \frac{1}{\Delta x^3}\Big(\Big(3U^{(0)[j]}_i - 5u^{(0)[j]}_{i+\frac{1}{2}} + 4u^{(0)[j]}_{i+\frac{3}{2}} - 3U^{(0)[j]}_{i+2} + u^{(0)[j]}_{i+\frac{5}{2}}\Big) \\ &- \Big(3U^{(0)[j]}_{i-1} - 5u^{(0)[j]}_{i-\frac{1}{2}} + 4u^{(0)[j]}_{i+\frac{1}{2}} - 3U^{(0)[j]}_{i+1} + u^{(0)[j]}_{i+\frac{3}{2}}\Big)\Big). \end{aligned} \tag{81}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}u^{(0)}_{i+\frac{1}{2}}(t) = &-\left(\max\left(u^{(0)}_{i+\frac{1}{2}}(t),0\right)\frac{1}{\Delta x}\left(2u^{(0)}_{i-\frac{1}{2}}(t)-6U^{(0)}_{i}(t)+4u^{(0)}_{i+\frac{1}{2}}(t)\right)\right.\\ &\left.+\min\left(u^{(0)}_{i+\frac{1}{2}}(t),0\right)\frac{1}{\Delta x}\left(-4u^{(0)}_{i+\frac{1}{2}}(t)+6U^{(0)}_{i+1}(t)-2u^{(0)}_{i+\frac{3}{2}}(t)\right)\right)\\ &-\frac{1}{\Delta x^3}\left(6U^{(0)[j]}_{i-1}-34u^{(0)[j]}_{i-\frac{1}{2}}+48U^{(0)[j]}_{i}-42U^{(0)[j]}_{i+1}\right.\\ &\left.+30u^{(0)[j]}_{i+\frac{3}{2}}-12U^{(0)[j]}_{i+2}+4u^{(0)[j]}_{i+\frac{5}{2}}\right).\end{aligned} \tag{82}$$

In addition, if the IMEX Runge-Kutta scheme is GSA, we obtain from (27), (28), (79), and (80),

$$\begin{aligned}\Psi^{(0)}_{i}(t) = \frac{1}{\Delta x^2}&\left(\left(-4u^{(0)}_{i+\frac{1}{2}}(t)+6U^{(0)}_{i+1}(t)-2u^{(0)}_{i+\frac{3}{2}}(t)\right)\right.\\ &\left.-\left(-4u^{(0)}_{i-\frac{1}{2}}(t)+6U^{(0)}_{i}(t)-2u^{(0)}_{i+\frac{1}{2}}(t)\right)\right)\end{aligned} \tag{83}$$

and (43) as well as

$$\rho^{(0)}_{i+\frac{1}{2}}(t) = \frac{1}{\Delta x}\left(-4u^{(0)}_{i+\frac{1}{2}}(t)+6U^{(0)}_{i+1}(t)-2u^{(0)}_{i+\frac{3}{2}}(t)\right) \tag{84}$$

and

$$\psi^{(0)}_{i+\frac{1}{2}}(t) = \frac{1}{\Delta x^2}\left(3U^{(0)}_{i}(t)-5u^{(0)}_{i+\frac{1}{2}}(t)+4u^{(0)}_{i+\frac{3}{2}}(t)-3U^{(0)}_{i+2}(t)+u^{(0)}_{i+\frac{5}{2}}(t)\right). \tag{85}$$

Equation (82) is a discretisation of (45) with an error in $\mathcal{O}(\Delta x)$ and (81) is a discretisation of (46) with an error in $\mathcal{O}(\Delta x^2)$, implying that the `biased` scheme is AP for the component u. If the time discretisation scheme is GSA, then the scheme is AP since (43), (83), (84), and (85) are discretisations of (48), (50), (47), and (49) with an error in $\mathcal{O}(\Delta x^2)$. Our numerical experiments demonstrate that the `biased` scheme achieves third order accuracy and is AA. This shows that a much finer theoretical analysis of the order of accuracy is required.

6 Numerical experiments

The formal asymptotic analysis indicates that a GSA scheme is the preferable choice for temporal discretisation. Thus, unless otherwise specified, we use the ARS$(4,4,3)$ scheme [6], which is GSA. The coefficients of this scheme are

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

and

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

The schemes are implemented in Julia [13] and the results are visualised with Makie.jl [16].

6.1 Experiment 1: Travelling wave solution of the HKdV equations

This experiment investigates the ability of our schemes to solve an initial value problem for the HKdV equations, i.e. a value of ε close to 1. For this experiment, $\varepsilon = 0.8$ is chosen. As initial data, we use the travelling wave solution of the HKdV equation that tends to the single-soliton solution of the KdV equation and its derivatives with $c = \frac{1}{2}$ and $x_0 = 0$ at $t = 0$. Since it is not analytically available, the travelling wave solution is computed numerically as described in [12] on the interval $[x_{\min}, x_{\max}] = [-47, 47]$ with periodic boundary conditions. In contrast to [12], eighth-order central finite difference

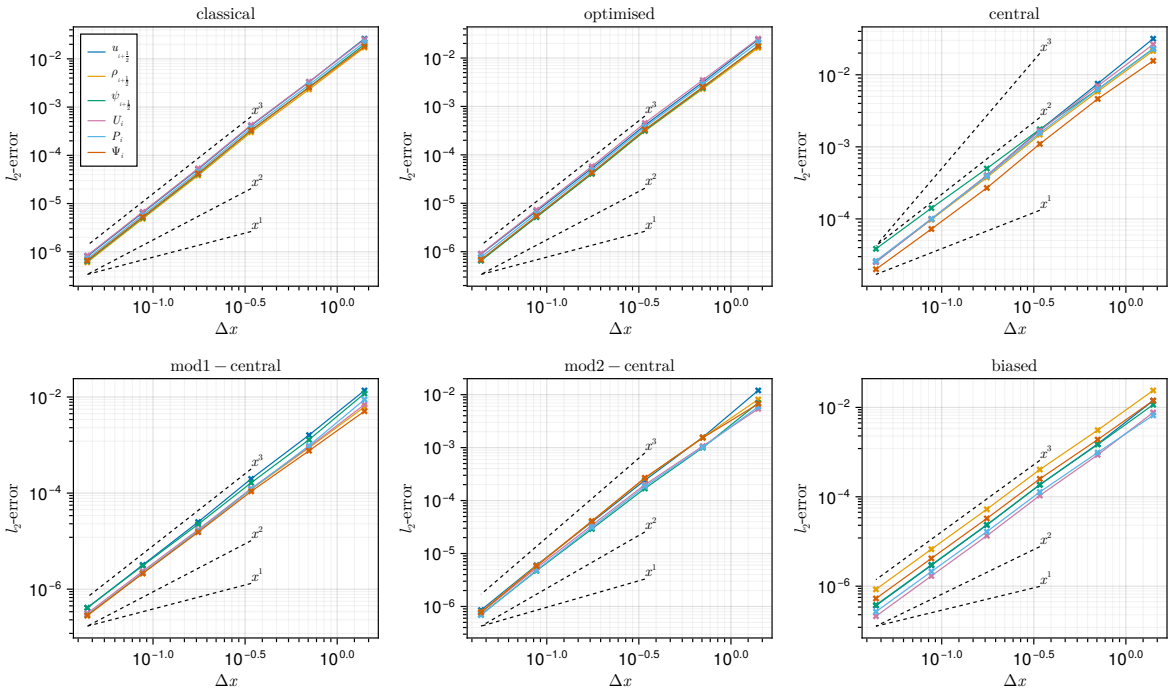


Figure 2: Experiment 1: The l_2-error of the point values and cell averages measured with respect to the spatial step size Δx with $\varepsilon = 0.8$ for the different IMEX Active Flux schemes. All schemes are third-order accurate except the `central` scheme which is only second-order accurate.

formulas [28] are used instead of Fourier spectral differentiation. We find that using high-order finite difference formulas and a sufficiently fine grid (42000 points) is sufficient for the experiments in this text. The numerical solution at the grid points is interpolated with cubic splines to obtain a continuous function $\tilde{q}_{\text{ref}} : [x_{\min}, x_{\max}] \to \mathbb{R}^3$. Shifting this function to the right with respect to time, i.e. $q_{\text{ref}}(t, x) = \tilde{q}_{\text{ref}}(x - ct)$, gives an accurate approximation to the travelling wave solution.

Given the initial data and reference solution, we compute the numerical solutions on the interval $[x_{\min}, x_{\max}] = [-45, 45]$ with periodic boundary conditions using the Active Flux schemes proposed in this work. For all Active Flux schemes we choose $C_{\text{CFL}} = 0.35$ if not stated otherwise.

We assess the experimental convergence rate, the error on a given grid and for long times, stability, and energy conservation. First, we compute the numerical solution at $t = 4$ on grids with 64, 128, 256, 512, 1024, and 2048 cells and measure the l_2-error for all point values and cell averages, see Figure 2. The `central` scheme achieves an order of accuracy of about two, while the other schemes show third order accuracy.

In order to make a comparison between the schemes on the same coarse grid with 64 cells, the maximum l_2-error is computed as

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

where in case of the HKdV equations $\Gamma_1 = \{u, \rho, \psi\}$ and $\Gamma_2 = \{U, P, \Psi\}$. The maximum l_2-error of the different schemes is displayed in Table 1. The `mod2-central` scheme admits the lowest error.

To investigate the long-time accuracy of the methods, we compute the maximum l_2-error on a grid with 512 cells at the time at which the wave has travelled once over the computational domain $[-45, 45]$, i.e. at $t = 180$ for $c = \frac{1}{2}$ (see Table 1). The `biased` scheme has the lowest error for this long simulation time.

In order to experimentally investigate stability bounds of the different schemes, we determine the maximum l_2-error of the numerical solution at $t = 4$ on a grid with 512 for different values of $C_{\text{CFL}} \leq 1$. We take extreme error growth as a sign of instability. Table 1 shows the maximum value of C_{CFL} for which the error is not higher than ten times the error for $C_{\text{CFL}} = 0.1$. The `biased` scheme and the `classical` scheme are the only schemes for which this holds for all tested values of $C_{\text{CFL}} \leq 1$.

`classical`	`optimised`	`central`	`mod1-central`	`mod2-central`	`biased`
maximum l_2-error on a grid with 64 cells at $t = 4$					
0.0262	0.0251	0.0317	0.0137	**0.0120**	0.0240
maximum l_2-error on a grid with 512 cells at $t = 180$					
0.0116	0.0121	0.0094	0.0019	0.0015	**0.0006**
maximum $C_{\mathrm{CFL}} \leq 1$ on a grid with 512 cells at $t = 4$					
1.000	0.450	0.450	0.450	0.450	**1.000**
conservation defect with $\Delta x = \frac{90}{256}, \frac{90}{512}, \frac{90}{1048}$ (top to bottom) at $t = 360$					
$1.0000 \cdot 10^{-2}$	$1.0423 \cdot 10^{-2}$	$7.6617 \cdot 10^{-3}$	$1.6431 \cdot 10^{-3}$	$1.3863 \cdot 10^{-3}$	$\mathbf{4.1771 \cdot 10^{-4}}$
$1.2609 \cdot 10^{-3}$	$1.3176 \cdot 10^{-3}$	$8.6421 \cdot 10^{-4}$	$2.0559 \cdot 10^{-4}$	$1.6120 \cdot 10^{-4}$	$\mathbf{5.5125 \cdot 10^{-5}}$
$1.5779 \cdot 10^{-4}$	$1.6476 \cdot 10^{-4}$	$9.5892 \cdot 10^{-5}$	$2.5729 \cdot 10^{-5}$	$1.9574 \cdot 10^{-5}$	$\mathbf{6.9489 \cdot 10^{-6}}$

Table 1: Results of Experiment 1 performed for the different Active Flux schemes discretised in time with the ARS$(4, 4, 3)$ scheme. The best performance is highlighted in boldface.

Lastly, we investigate the conservation of energy by the different numerical schemes. It is not clear if a given Active Flux scheme conserves a discrete energy and how this energy is defined. We compare the proposed schemes by measuring

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

which can be considered a discrete counterpart of the energy of the HKdV equations. N denotes the number of cells. We define the conservation defect at time t as

$$\Delta E_{\mathrm{H,dis}}(\Delta x, t) = 1 - \frac{E_{\mathrm{H,dis}}(\Delta x, t)}{E_{\mathrm{H,dis}}(\Delta x, 0)}.$$

This value is displayed in Table 1, where we chose $t = 360$ and $\Delta x = \frac{90}{256}, \frac{90}{512}, \frac{90}{1048}$. All schemes conserve the energy satisfactorily but there is a factor of 20 in the defect between the best and the worst scheme on the same grid. The `biased` scheme delivers the best results followed by the `mod2-central` scheme.

Observe that the results obtained for the `classical` and `optimised` schemes are similar, which is not unexpected since the schemes are closely related. The results of the experiment conducted for the `combined` scheme discretised in time with the explicit SSPRK3 scheme are also comparable to the results obtained with the `classical` and `optimised` scheme: it achieves third order accuracy, the maximum l_2-error on a coarse grid is 0.0246 and for long times 0.0118, the maximum value of C_{CFL} is 0.5, and the conservation defects are $1.0123 \cdot 10^{-2}$ (256 cells), $1.2834 \cdot 10^{-3}$ (512 cells), and $1.6085 \cdot 10^{-4}$ (1024 cells).

6.2 Experiment 2: Single-soliton solution of the KdV equation

Next, we investigate the ability of the schemes to compute numerical approximations to the limit equations of the HKdV equations, i.e. we study the behaviour of the methods in the limit $\epsilon \to 0$. The interval $[x_{\min}, x_{\max}] = [-45, 45]$ is chosen, periodic boundary conditions are imposed, and the CFL number $C_{\mathrm{CFL}} = 0.35$ is used. The single-soliton solution (2) of the KdV equations (with $x_0 = 0$ and $c = \frac{1}{2}$) and its first two derivatives serve as reference solution. In particular, the initial data are

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

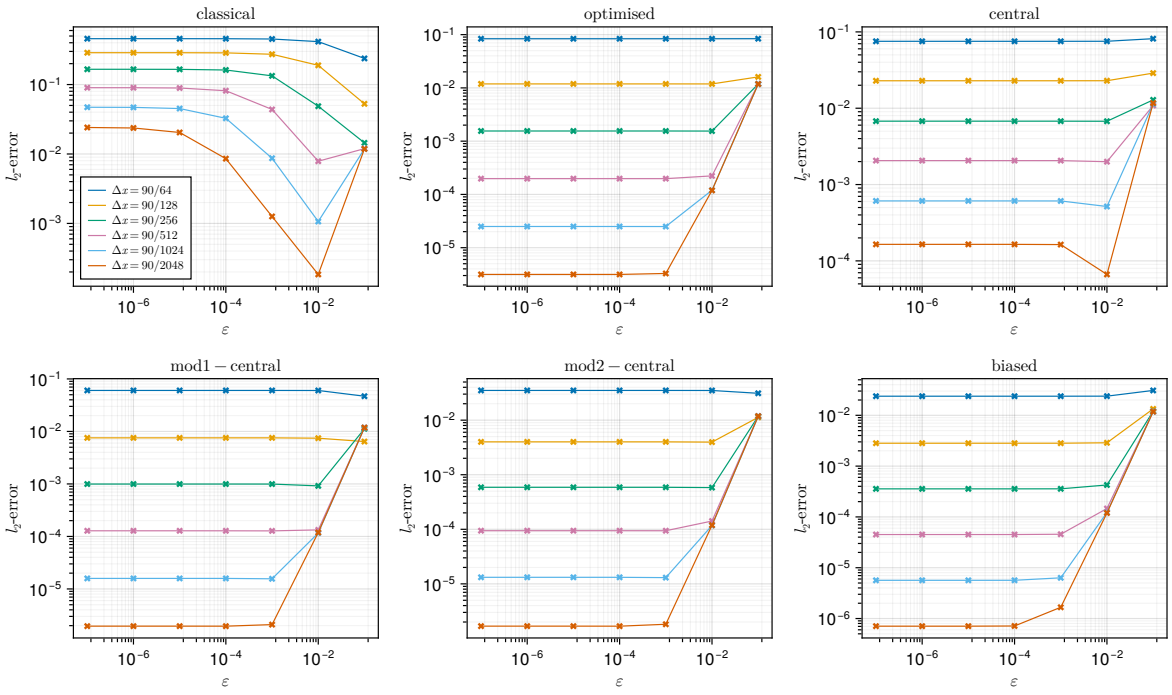


Figure 3: Experiment 2: The l_2-error of $u_{i+\frac{1}{2}}$ with respect to the relaxation parameter ε for the different IMEX Active Flux schemes and on different grids.

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 we want to determine if the scheme can solve an initial value problem for the limit equations.

In Figure 3 the l_2-error of the numerical solution at $t = 4$ is shown for different values of ε and different spatial step sizes. For simplicity only the l_2-error of the point values $u_{i+\frac{1}{2}}$ is displayed. The errors of the other point values and cell averages behave similarly. For all schemes, the error becomes constant for a fixed Δx as $\varepsilon \to 0$. This implies that for the smallest value of ε considered, i.e. $\varepsilon = 10^{-7}$, the measured error is the discretisation error of the limit scheme (cf. Figure 1). Observe that if $\Delta x \to 0$, the error decreases. This is shown in more detail in Figure 4, where the error of all point values and cell averages is displayed for fixed $\varepsilon = 10^{-7}$ and different values of Δx. Figures 3 and 4 are numerical evidence that the schemes are AP, since, as $\varepsilon \to 0$, the schemes become a (consistent and stable) discretisation of the limit scheme. For a scheme that is not AP, the measured error for small ϵ would remain constant upon grid refinement, for example.

The accuracy of the `classical` scheme drops to first-order, while all the other schemes retain the same order of accuracy as for large ε. Consequently, the numerical experiment suggests that all schemes except the `classical` scheme are AA.

Remark 6.1 (Importance of the GSA property). Figure 5 shows the l_2-error of the point values and cell averages for different spatial step sizes if the schemes are discretised in time with the ARS$(2, 3, 3)$ scheme, which is not GSA and has the following coefficients:

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

and

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

with $\gamma = \frac{3+\sqrt{3}}{6}$. The results obtained for the `classical` and `central` scheme are similar to those shown in Figure 4: the `classical` scheme achieves first order accuracy and the `central` scheme second order accuracy. The other schemes display third-order accuracy for u, but only a reduced order for ρ

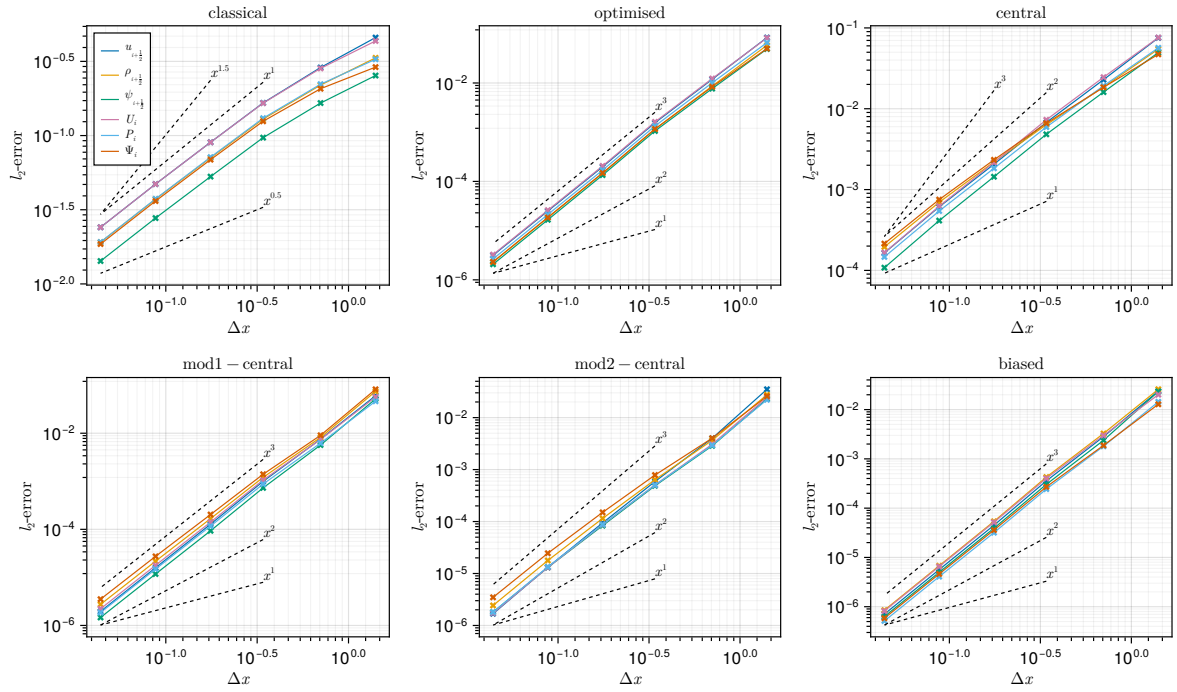


Figure 4: Experiment 2: The l_2-error of the point values and cell averages with respect to the spatial step size Δx for $\varepsilon = 10^{-7}$. The `classical` scheme only achieves first-order accuracy, the `central` scheme second-order accuracy, and the other schemes third-order accuracy.

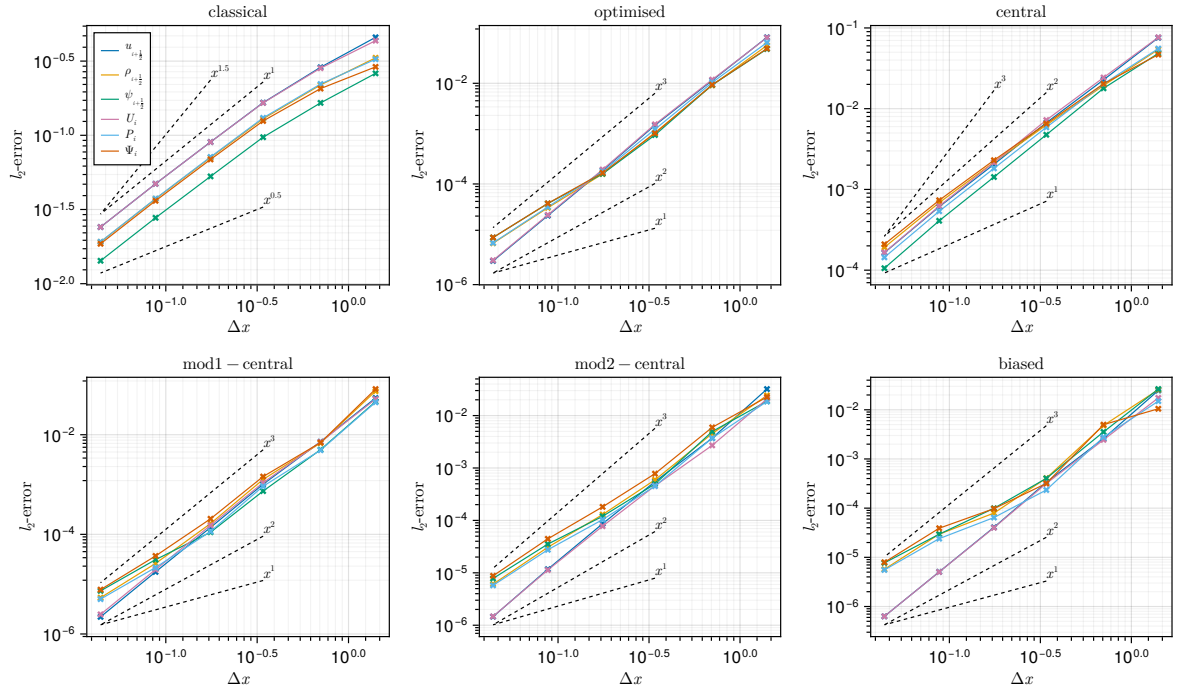


Figure 5: Experiment 2: The l_2-error of the point values and cell averages with respect to the spatial step size Δx for the different Active Flux schemes discretised in time with the ARS$(2, 3, 3)$ scheme.

`classical`	`optimised`	`central`	`mod1-central`	`mod2-central`	`biased`
maximum l_2-error on a grid with 64 cells at $t = 4$					
0.4576	0.0837	0.0760	0.0820	0.0350	**0.0258**
maximum l_2-error on a grid with 512 cells at $t = 180$					
3.5441	0.0578	0.3014	0.0190	**0.0086**	0.0123
maximum $C_{\mathrm{CFL}} \leq 1$ on a grid with 512 cells at $t = 4$					
1.000	0.400	0.450	0.425	0.425	**1.000**
conservation defect with $\Delta x = \frac{90}{256}, \frac{90}{512}, \frac{90}{1048}$ (top to bottom) at $t = 360$					
$7.1943 \cdot 10^{-1}$	$3.2949 \cdot 10^{-2}$	$1.0313 \cdot 10^{-1}$	$-9.5614 \cdot 10^{-3}$	$\mathbf{5.0983 \cdot 10^{-3}}$	$7.0442 \cdot 10^{-3}$
$6.1156 \cdot 10^{-1}$	$4.3219 \cdot 10^{-3}$	$2.4094 \cdot 10^{-2}$	$-1.3954 \cdot 10^{-3}$	$\mathbf{6.5200 \cdot 10^{-4}}$	$9.1454 \cdot 10^{-4}$
$4.7749 \cdot 10^{-1}$	$5.4399 \cdot 10^{-4}$	$4.1704 \cdot 10^{-3}$	$-1.8348 \cdot 10^{-4}$	$\mathbf{8.7434 \cdot 10^{-5}}$	$1.1553 \cdot 10^{-4}$

Table 2: Results of Experiment 2 performed for the different IMEX Active Flux scheme discretised in time with the ARS$(4, 4, 3)$ scheme. The best performance is highlighted in boldface.

and ψ, implying that the schemes are AP but with exception of the central scheme are not AA for all components.

In addition, the matrices A of the ARS$(2, 3, 3)$ and ARS$(4, 4, 3)$ schemes have the same entries on the diagonal. This allows us to precompute the LU-decomposition of the resulting matrices in the time discretisation. Hence, even though the ARS$(4, 4, 3)$ scheme has more stages than the ARS$(2, 3, 3)$ scheme, their efficiency is almost the same since computing the LU-decomposition is the most expensive part. This justifies our choice to employ the ARS$(4, 4, 3)$ scheme to discretise in time.

For the remainder of the experiment $\varepsilon = 10^{-7}$ is chosen. As in the first experiment, to compare the different schemes, we measure their maximum l_2-error of the numerical solutions at $t = 4$ computed on a coarse gird (64 cells) and at $t = 180$ on a grid with 512 cells. The results are shown in Table 2. The `biased` scheme has the smallest error on the coarse grid and the `mod2-central` scheme the lowest for long times. Additionally, we determine the highest values of C_{CFL}, for which the maximum error of a numerical solution computed on a grid with 512 cells at $t = 4$ is smaller than ten times the error with $C_{\mathrm{CFL}} = 0.1$. As shown in Table 2, only the `classical` and `biased` scheme achieve this for all $C_{\mathrm{CFL}} \leq 1$.

Lastly, we investigate the conservation of a discrete counterpart of the energy conserved by the KdV equation. We define

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

and the conservation defect

$$\Delta E_{\mathrm{H}}(\Delta x, t) = 1 - \frac{E_{\mathrm{H}}(\Delta x, t)}{E_{\mathrm{H,dis}}(\Delta x, 0)}.$$

The defect at $t = 360$ for $\Delta x = \frac{90}{256}, \frac{90}{512}, \frac{90}{1048}$ is shown in Table 2. There are 2 orders of magnitude in the measured defect between the best and the worst scheme. The `mod2-central` scheme gives the best results, followed by the `biased` scheme. Observe that the discrete energy of the numerical solution computed with the `mod1-central` scheme increases and that the defect of the classical nearly stagnates upon mesh refinement.

6.3 Experiment 3: Double-soliton solution of the KdV equation

In this experiment, we are employing the different Active Flux schemes to numerically compute the double-soliton solution of the KdV equation. We set $\varepsilon = 10^{-7}$ and $C_{\mathrm{CFL}} = 0.35$ for all parts of the experiment, choose $[x_{\min}, x_{\max}] = [-30.0, 80.0]$ divided into 512 cells, and impose periodic boundary conditions. As initial data,

$$u_0(x) = 2(\delta_2^2 - \delta_1^2) \frac{\delta_2^2 \mathrm{cosech}(\gamma_2(0, x))^2 + \delta_1^2 \mathrm{sech}(\gamma_1(0, x))^2}{(\delta_2 \coth(\gamma_2(0, x)) + \delta_1 \tanh(\gamma_1(0, x)))^2},$$

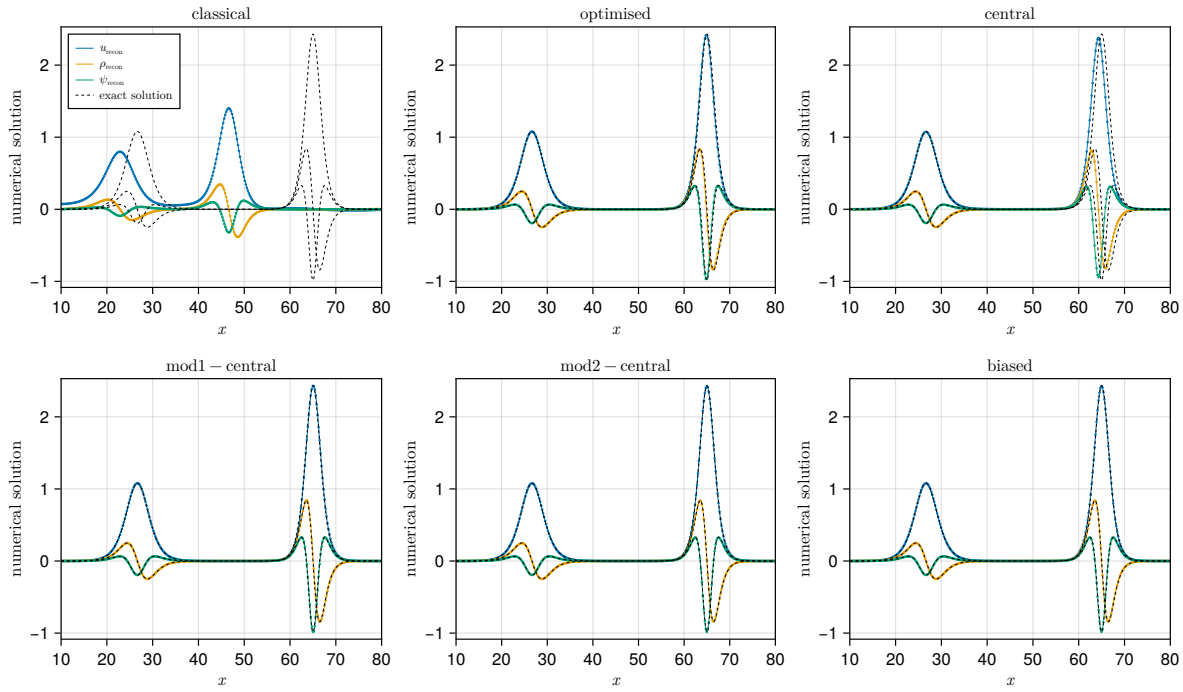


Figure 6: Experiment 3: The numerical double-solution solution computed with the different Active Flux schemes in comparison to the exact solution at $t = 100$. The faster soliton has already overtaken the slower one.

$$\rho_0(x) = Du_0(x),$$
$$\psi_0(x) = D\rho_0(x)$$

is used, where γ_1 and γ_2 are defined as in Theorem 1.2 and D is a fourth order central finite difference operator [28] with spatial step size $h = 0.0005$; $\delta_1 = 0.3$ and $\delta_2 = 0.45$ are selected. The initial data consist of two solitons, a smaller, slower one that is located on the right of a faster, larger one. The analytical solution of this initial value problem for the limit equations is the function given in Theorem 1.2 and its first and second derivative. This solution acts as the reference solution in this experiment. The first and second derivatives of it are again computed using fourth order central finite differences.

Figure 6 shows the numerical solution computed with the different Active Flux schemes at $t = 100$. At this time, the faster soliton has already overtaken the slower soliton. The l_2-error of the numerical solution for various times t is displayed in Figure 7. The `mod2-central` scheme is the most accurate out of all the different schemes, while the `classical` and the `central` schemes yield solutions that deviate strongly from the exact one.

7 Conclusion and outlook

After presenting the `combined` scheme, i.e. the standard semi-discrete Active Flux scheme for the HKdV equations, we derived a closely related scheme, the `classical` scheme, by individually upwinding the nonlinear and stiff part of the equations. This scheme allows the discretisation in time with IMEX Runge-Kutta schemes. Formal asymptotic analysis was conducted for this scheme using an IMEX Runge-Kutta scheme of type ARS. The theoretical findings show that the fully-discrete scheme is AP for the u component of the HKdV equations. If the time discretisation is globally stiffly accurate (GSA), then the scheme is AP for all components. This is confirmed experimentally, yet the scheme is only first-order accurate for small ε.

To find schemes with superior properties, we considered different ways to upwind the stiff part of the Jacobian. First, the `optimised` scheme was obtained by modifying and rescaling particular elements of the upwinding matrices of the stiff operator. In addition, using central discretisation for the terms discretised in time implicitly, the `central` scheme was developed. Since the scheme turned out to achieve only second-order accuracy, two modifications were introduced, the `mod1-central` and `mod2-central`

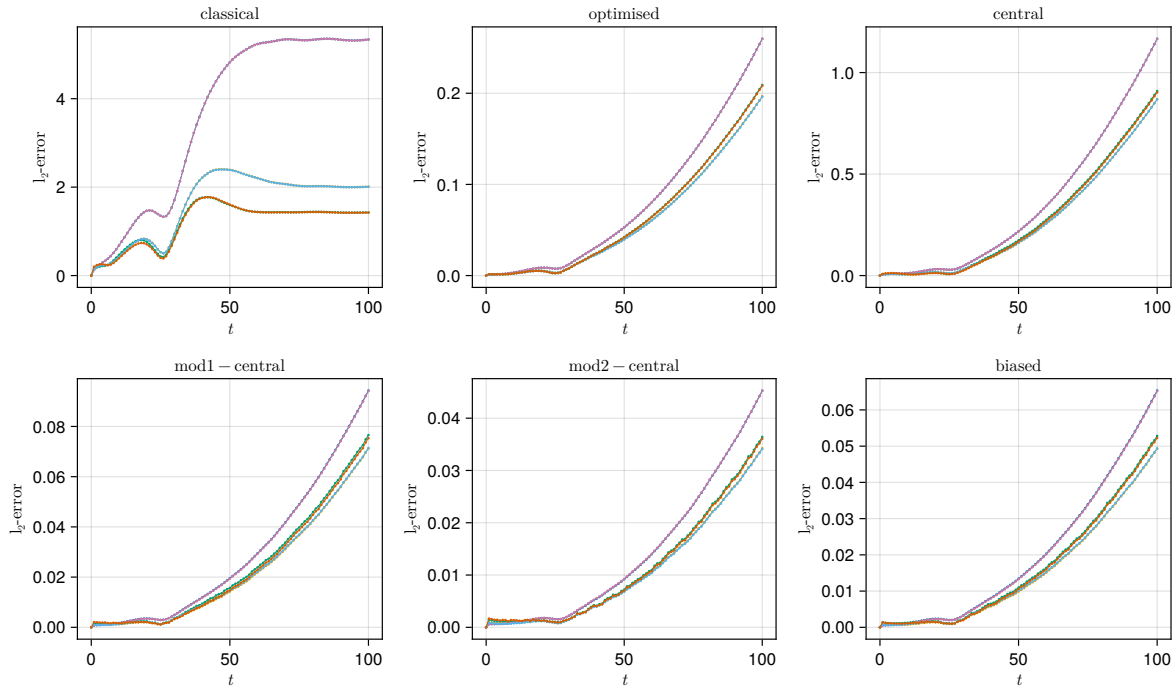


Figure 7: Experiment 3: The l_2-error of the numerical double-solution solution computed with the different Active Flux schemes as a function of time t.

schemes. Both of them achieve an order of accuracy of three. Lastly, to demonstrate an alternative to central discretisation, we presented the `biased` scheme, where we imposed predefined upwind directions in the upwinding matrices of the stiff operator. Formal asymptotic analysis for these schemes gave the same results as for the `classical` scheme, i.e. that they all are AP.

The numerical experiments conducted in this work verify that all of the schemes presented here are AP, which is generally not trivial to achieve. In fact, the experiments show that they already appear to be AP if they are discretised in time with a scheme that is not GSA. All schemes apart from the `classical` scheme are AA, a property clear from the experiments but incompletely understood theoretically. We observed this property (with the exception of the `central` scheme) only if we used a GSA time discretisation. By comparing the accuracy, maximum CFL number, and conservation of a discrete energy, the `mod2-central` scheme and the `biased` scheme emerge as the schemes with the best properties.

Further analysis of the presented alternative point value update equations and resulting schemes is a compelling topic for future research into the reasons for their superior properties compared to the `classical` scheme. One option is to implement the limit schemes and measure the rate at which the numerical approximations of the Active Flux schemes tend to the numerical approximations of the corresponding limit schemes. Moreover, it would be interesting to further study the role of the time-discretisation by using different IMEX Runge-Kutta schemes of type I and type II and comparing the results. It is also worth identifying simple criteria for a given Active Flux scheme to be AP, without having to redo the asymptotic analysis. In the future, we will apply the IMEX Active Flux schemes developed here to more complex models of dispersive free-surface waves. The applicability of IMEX schemes is not restricted to hyperbolised high-order PDEs, and we will explore their use in other contexts.

Acknowledgements

Christian Klingenberg acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within SPP 2410 Hyperbolic Balance Laws in Fluid Mechanics: Complexity, Scales, Randomness (CoScaRa), project No. 525941602.

References

[1] 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.

[2] Rémi Abgrall and Wasilij Barsukow. Extensions of Active Flux to arbitrary order of accuracy. ESAIM: M2AN, 57(2):991–1027, 2023. DOI: 10.1051/m2an/2023004.

[3] Rémi Abgrall and Wasilij Barsukow. A hybrid finite element–finite volume method for conservation laws. Applied Mathematics and Computation, 447:127846, 2023. DOI: 10.1016/j.amc.2023.127846.

[4] 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.

[5] Mark J. Ablowitz and Harvey Segur. Solitons and the Inverse Scattering Transform. Society for Industrial and Applied Mathematics, 1981. DOI: 10.1137/1.9781611970883.

[6] 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(2-3):151–167, 1997. DOI: 10.1016/S0168-9274(97)00056-1.

[7] 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.

[8] 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, 8:1478–1526, 2026. DOI: 10.1007/s42967-025-00512-3.

[9] 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.

[10] Caterina Bassi, Luca Bonaventura, Saray Busto, and Michael Dumbser. A hyperbolic reformulation of the Serre-Green-Naghdi model for general bottom topographies. Computers and Fluids, 212:104716, 2020. DOI: 10.1016/j.compfluid.2020.104716.

[11] Christophe Besse1, Sergey Gavrilyuk, Maria Kazakova, and Pascal Noble. Perfectly matched layers methods for mixed hyperbolic–dispersive equations. Water Waves, 4:313–343, 2022. DOI: 10.1007/s42286-022-00069-1.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Firas Dhaouadi and Michael Dumbser. A first order hyperbolic reformulation of the Navier-Stokes-Korteweg system based on the GPR model and an augmented Lagrangian approach. Journal of Computational Physics, 470:111544, 2022. DOI: 10.1016/j.jcp.2022.111544.

[18] Philip G. Drazin. Solitons. Cambridge University Press, Cambridge, 1983.

[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 laws – flux vector splitting and bound-preservation. SIAM Journal on Scientific Computing, 47(2):A811–A837, 2025. DOI: 10.1137/24M1658887.

[21] 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.

[22] 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.

[23] 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.

[24] Lawrence C. Evans. Partial differential equations. American Mathematical Society, Providence, 2010.

[25] 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.

[26] 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.

[27] Timothy Eymann and Philip L. Roe. Multidimensional Active Flux schemes. In: 21st AIAA Computational Fluid Dynamics Conference, 2013. DOI: 10.2514/6.2013-2940.

[28] Bengt Fornberg. Generation of finite difference formulas on arbitrarily spaced grids. Mathematics of Computation, 51(184):699-706, 1988. DOI: 10.2307/2008770.

[29] Sergey Gavrilyuk, Boniface Nkonga, and Keh-Ming Shyue. The conduit equation: Hyperbolic approximation and generalized Riemann problem. Journal of Computational Physics, 514:113232, 2024. DOI: 10.1016/j.jcp.2024.113232.

[30] Lukas Hensel, Gudrun Grünwald, Katharina Kormann, and Rainer Grauer. A split-step Active Flux method for the Vlasov-Poisson system. Journal of Computational Physics, 540:114294, 2025. DOI: 10.1016/j.jcp.2025.114294.

[31] 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. https://www.rivmat.unipr.it/vols/2012-3-2/01.html.

[32] David Ketcheson and Abhijit Biswas. Approximation of arbitrarily high-order PDEs by first-order hyperbolic relaxation. Nonlinearity, 38(5):055002, 2025. DOI: 10.1088/1361-6544/adc6e8.

[33] Diederik Korteweg and Gustav de Vries. XLI. On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 39(240):422–443, 1895. DOI: 10.1080/14786449508620739.

[34] Xue-song Li and Chun-wei Gu. Mechanism of Roe-type schemes for all-speed flows and its application. Computers & Fluids, 86:56-70, 2013. DOI:10.1016/j.compfluid.2013.07.004.

[35] François Marin. Solitons: Historical and physical introduction. In: Robert A. Meyers, editor, Mathematics of Complexity and Dynamical Systems. Springer, New York, 2012. DOI: 10.1007/978-1-4614-1806-1_100.

[36] John Scott Russell. Report on waves. In: Report of the fourteenth meeting of the British Association for the Advancement of Science. John Murray, 311-390, 1845.

[37] 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.

[38] Bram van Leer. Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. Journal of Computational Physics, 23(3):276–299, 1977. DOI: 10.1016/0021-9991(77)90095-X.

[39] Wolfram Research, Inc. Mathematica, Version 15.0, Champaign, IL, 2026.

[40] Tohru Yoneyama. The Korteweg-de Vries two-soliton solution as interacting two single solitons. Progress of Theoretical Physics, 71(4):843–846, 1984. DOI: 10.1143/PTP.71.843.

[41] Norman Zabusky and Martin Kruskal. Interaction of "Solitons" in a Collisionless Plasma and the Recurrence of Initial States. Physical Review Letters, 15(6):240–243, 1965. DOI: 10.1103/PhysRevLett.15.240.