Active Flux Methods for Conservation Laws on Complex Geometries

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

David Christian Kerkmann
from Düsseldorf

Düsseldorf, June 2021

from the Mathematical Institute
at the Heinrich Heine University Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Supervisor:	Prof. Dr. Christiane Helzel
	Heinrich Heine University Düsseldorf
Co-supervisor:	Prof. Dr. Rupert Klein
	Freie Universität Berlin

Date of the oral examination: 09/30/2021

Abstract

We are interested in solving hyperbolic conservation laws that are models for transport phenomena for example in computational fluid dynamics. In practical applications, the domain of interest is typically not shaped simply but is rather complex. To discretize these complex geometries, we use a Cartesian grid which cells are cut along the boundary. The resulting *cut cells* can have arbitrary shapes and sizes.

A numerical method for conservation laws should be accurate, stable and conservative. Many methods have been proposed that deal with these aspects for cut cells. Each method has its own advantages and disadvantages. Explicit finite volume methods typically satisfy a time step restriction that depends on the smallest cell size. We search for a third order accurate method that is both conservative and stable for reasonable time steps that do not depend on the size of the smallest cut cells. A necessary requirement for stability is the so-called cancellation property which ensures that the update of the small cells is bounded by the order of their cell sizes.

An attractive candidate for a method that might satisfy these properties is the *Active Flux method* developed by Eymann and Roe. The Active Flux method is a finite volume method that not only uses the cell average values but also point values of the conserved quantities at the interfaces between neighboring cells. It updates the point values separately from the cell averages, thus the *flux* is computed *actively*. In each cell, the conserved quantities are reconstructed locally using only the values that belong to that cell. When reconstructing on irregular grid cells, reconstructions can become ill conditioned and lead to poor results when some values are disturbed slightly. We discuss ways to overcome this problem. Furthermore, we examine the stability properties of the Active Flux method for linear systems on Cartesian grids and cut cell grids. The method is stable on cut cell grids for time steps that are restricted only by the size of a regular grid cell. We show that the cancellation property can be achieved when using a continuous reconstruction. No further stabilization technique is required. Furthermore, we find that for our linear model problems the method shows an excellent third order accuracy and is conservative by nature.

In the second part we develop a different interpretation of the Active Flux method for Cartesian grids. This reinterpretation can be applied to every hyperbolic conservation law and is no longer restricted to linear or simple nonlinear systems. It differs from the original Active Flux method in the way the point values are updated since no exact evolution formula is known. We present results for a multitude of test cases and obtain third order accuracy, good stability and conservation. Many of the results presented in this thesis can be found in our previous publications [1, 2, 3].

Table of Contents

List of Figures

List of Tables

1 Introduction

Mesh generation on complex geometries

For many years, researchers have tried to solve partial differential equations (PDEs) on complex geometries. These geometries arise directly from the applications of interest. A typical field of application is computational fluid dynamics (CFD), where one wishes to determine the airflow around a vehicle or the flow of fluids around obstacles, for example. The solution to a given set of PDEs that lives on a complex geometry is generally unknown and, depending on the nature of the problem, there often doesn't even exist a unique solution in the classical sense. Numerical solvers are needed to determine an approximate solution to the problem. An essential part of the numerical solver is the construction of a computational mesh of the relevant region. While simple domains can easily be discretized by a Cartesian or triangular mesh or a mapped version of the just mentioned, complex domains raise the question of where the discrete values are to be defined, since it may very well not be possible in a uniform way.

The standard approach is the use of an unstructured triangularization. Here, the domain is partitioned in many triangles (in two dimensions) or tetrahedrons (in three dimensions). One tries to create similar sized cells and to avoid very small or degenerate cells. This task may sound simple but, for complex geometries, can be difficult and expensive to perform. Additionally, effort has to be made to store cell information such as size, shape and connections to other cells. The numerical method has to be able to adapt to these different cell sizes.

A much simpler way to generate a numerical mesh is to use an underlying Cartesian grid consisting of regular squares or cubes, where the cells are cut along the boundary of the domain and result in so-called cut cells. By construction, the resulting cut cells can have various shapes and arbitrarily small sizes. Figure 1.1 shows a potential cut cell grid on a domain in two spatial dimensions. The use of Cartesian grids makes it easier to construct numerical methods for the interior of the domain, where no cut cells are present. Cells can be referenced by index arrays and the same scheme can be used for any internal cell without modification. Also, it allows the easier use of local and adaptive mesh refinement, as the cells can be divided into smaller Cartesian cells in a trivial way. The difficulty in the construction of a numerical solver that uses cut cells lies in the treatment of the cut cells themselves. As the method on the regular cells in the interior of the domain often has very good accuracy and stability properties, it is mandatory to adapt the method on the cut cells along the boundary to maintain these properties as much as possible. A well established program that makes use of the

cut cell framework is Cart3D [1]. This package performs grid generation, CFD analysis and simulations of solutions in an automated way. Recently, an implicit solver for time dependent problems has been added to the previously existing solver for steady state problems. Hyperbolic conservation laws, as introduced in the next section, allow for explicit solvers which are usually more efficient and are therefore more commonly used. Thus, we focus on explicit schemes in this work.

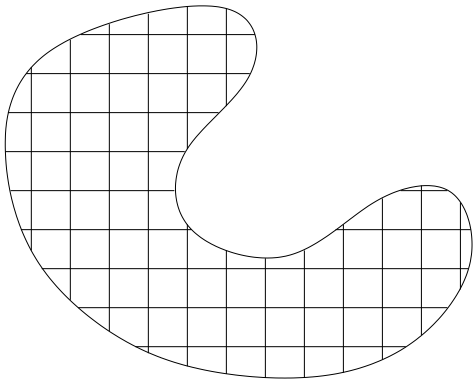

Figure 1.1.: Cut cell grid in two dimensions.

Equations and notation

In this thesis, we are concerned with hyperbolic conservation laws. Conservation laws are often found in CFD because many related problems deal with the conservation of physical quantities such as mass, momentum or energy.

We consider the conservation law

$$\frac{\partial}{\partial t} q(x,t) + \nabla_x \cdot f(q(x,t)) = 0, \quad \text{on } \Omega^o \times \mathbb{R}^+ \tag{1.1}$$

with initial values

$$q(x,0) = q_0(x), \quad x \in \Omega. \tag{1.2}$$

Here, d is a spatial dimension, $\Omega \in \mathbb{R}^d$ is a domain, Ω^o its interior and $\partial\Omega$ its boundary, $q : \Omega \times \mathbb{R}^+ \to \mathbb{R}^m$ is a vector of $m \in \mathbb{N}$ conserved quantities, $f : \mathbb{R}^m \to \mathbb{R}^m \times \mathbb{R}^d$ is a flux function and $q_0 : \Omega \to \mathbb{R}^m$ is an initial condition. Boundary values are imposed on $\partial\Omega \times \mathbb{R}^+$. Furthermore, we assume f to be sufficiently smooth. Moreover, derivatives and integrals of vector valued functions are to be understood as vectors of the derivatives and integrals of the individual components.

Let S^{d-1} be the unit sphere in $\mathbb{R}^d$. We assume that the system is hyperbolic, i.e., we have for all $\widetilde{q} \in \mathbb{R}^m$ and all $n \in S^{d-1}$ that $f'(\widetilde{q}) \cdot n$ is diagonalizable with real eigenvalues.

Many theoretical and analytical results for hyperbolic conservation laws have been found. A detailed introduction to the topic of conservation laws can, for example, be found in [4, 5, 6]. Most notably, the name conservation law stems from the conservation of the quantities q, which can be seen by integrating (1.1) over a space-time-volume

[1] `https://www.nas.nasa.gov/publications/software/docs/cart3d/`

$\Omega \times [t_n, t_{n+1}]$. After the application of the divergence theorem, one obtains

$$\int_\Omega q(x, t^{n+1}) \, \mathrm{d}x - \int_\Omega q(x, t^n) \, \mathrm{d}x + \int_{t_n}^{t_{n+1}} \int_{\partial\Omega} n \cdot f(q(x,t)) \, \mathrm{d}S \, \mathrm{d}t = 0. \tag{1.3}$$

Here, $n \in \mathbb{R}^d$ is the outer normal at each point on $\partial\Omega$. Subsequently, the integral of q during a period of time $t_{n+1} - t_n$ is only changed through the fluctuations across the boundary $\partial\Omega$. This is also the basis for the class of finite volume methods (FVM). By applying the divergence theorem on each cell of the spatial discretization, cell averages of the conserved quantities can be defined. The time-averaged true flux integral over a cell face with index k is approximated by a numerical flux F_k, so that each connecting face of two cells has a flux approximation. Through substituting the exact flux by the numerical flux for all faces in (1.3), a finite volume method is obtained. The conservation property is recovered by summing up all cell averages weighted by the cell size, leading to a cancellation of all numerical fluxes except the boundary fluxes.

Another important property of solutions is that they don't need to keep the regularity of the initial conditions. After a period of time, solutions can become discontinuous even if $q_0 \in C^\infty(\mathbb{R}^d)$ and therefore fail to fulfill (1.1) in a classical way. Weak solutions are introduced, that obey the weak formulation of the conservation law. As weak solutions are not unique, entropy conditions or other selection criteria are used as a filter for the physically relevant solution. The predominant class of methods that work directly on the (semidiscrete) weak formulation is the class of discontinuous Galerkin (DG) schemes.

Finite volume methods and discontinuous Galerkin methods are the two most commonly used types of methods for hyperbolic conservation laws.

Three important aspects of cut cell methods

Generally, there are three important aspects that need to be considered when constructing a numerical solver for conservation laws. These aspects are not only important on the regular part of the domain but need to be considered especially in and when transitioning to the cut cells.

Firstly, the method should give an accurate approximation to the true solution. This is, for smooth solutions, often measured by the order of convergence. When measuring with a norm that is taking the grid size into account, such as the L_1-norm, we can typically allow the order of the error in the cut cell region at the boundary to be one order less than the order in the interior part of the domain, since this area converges to the boundary of the domain, which is of one dimension smaller than the domain itself. Nevertheless, the domain close to the boundary often contains important aspects of the solution structure, so it is encouraged to also obtain the same order of accuracy here.

Secondly, the occurring cut cells can be of arbitrary shape and size. Typically, explicit schemes obey a stability condition which ensures that the solution does not blow up. This stability condition is given in the form of a time step restriction that limits the time step according to the cell sizes of the numerical mesh. Since the cut cells

can be orders of magnitude smaller than the regular cells in the interior of the domain, this restriction would lead to a very small time step and thereby an unfeasible method in practice. This is called the small cell stability problem. A necessary condition for stability is the cancellation property. The sum of the numerical fluxes over each cell face has to be bounded by the order of the cell size:

$$\sum_k n_k \cdot F_k = \mathcal{O}(|\Omega_i|) \tag{1.4}$$

By employing a small time step that suits the standard time step restriction for the smallest cell, this is automatically achieved. However, for cut cell grids, one seeks to construct a scheme that is stable under a stability condition that depends on the size of the regular cells rather than the cut cells.

Thirdly, and as important as the other two aspects, is the aspect of conservation. As already mentioned, a conservation law conserves the total amount of one or several quantities. The importance of covering the conservation in the numerical method is twofold: On the one hand, conservation is one of the conditions in the famous Lax-Wendroff theorem, which (under some technical assumptions) states that a converging, conservative method converges to a weak solution [7]. On the other hand, failing to conserve the quantities can lead to nonphysical solutions [6, Chapter 2.9].

Hitherto existing methods

Over the past couple of decades, many attempts have been made to cope with these three difficulties in the presence of cut cells, leading to a whole set of methods. The following enumeration is not a complete list of all developed methods but gives a good overview over different approaches.

One idea that comes to mind to overcome the small cell issue is cell merging. By fusing the small cut cells to decently sized neighboring cells, the size of the smallest cells is no longer an issue for stability. This has been done for both FVM [8] and DG [9]. The difficulties lie in the selection of the cells to be merged. This is especially true in three dimensions and a robust and automatic cell merging is yet to be discovered.

To pass the cancellation property, Colella et al. have tried to redistribute a large portion of the flux difference in cut cells to the neighboring cells [10]. That way the cancellation property is achieved while keeping the method conservative. The redistribution is usually done according to the cell sizes of the neighboring cells. This method is easily implemented and often used in practical computations, but lacks accuracy. Second order accuracy could not be obtained at the boundary.

The ideas of the two previous approaches are combined in a recent work by Berger and Giuliani [11] in a FVM and by Giuliani in a DG setting [12]. They first perform an unstable update on the whole grid and use a post-processing routine to account for stability, while maintaining conservation. Cut cells are again merged with neighboring cells, while it is possible for the new cells to overlap. The post-processing takes these new neighborhoods into account and corrects the numerical states, redistributing them to the neighboring cells. Therefore this method has the name state redistribution.

Another way to satisfy the cancellation property was done by Berger, Helzel and LeVeque [13, 14, 15]. They introduce so-called h-boxes, which are additional artificial cells tangential and normal to the cut cell boundary. Due to the overlapping of the boxes and the way they are computed, the cancellation property is satisfied. While second order accuracy is achieved, the method has not been implemented in three dimensions yet.

To ensure stability in the cut cells, Berger and May have developed a method that uses explicit time stepping in the interior of the domain and implicit time stepping at the boundary [16, 17]. They discuss how to couple both schemes and achieve up to second order accurate results.

Gokhale, Klein, Nikiforakis and Nordin-Bates propose a dimensional split approach in [18, 19], where cut cell fluxes are carefully calculated through a "local proportional flux stabilization" that makes use of local geometry and wave speed information. Second order in the L_1-norm is achieved and quite challenging problems are solved, even in three dimensions.

In the setting of finite difference methods, Tan and Shu developed a method that approximates the boundary conditions through an inverse Lax-Wendroff procedure [20]. The method is further developed to account for high order, moving boundaries and efficient implementation [21, 22, 23, 24]. While the results look very sound, the conservation property is not recovered at the boundary.

Recently, Engwer et al. proposed a new stabilisation for solving the linear advection equation in two dimensions using a DG method [25]. By adding penalty terms to the spatial discretization, the method aims at reconstructing the proper domain of dependence for the cut cells. It achieves second order accuracy in the multidimensional case, measured in the L_1-norm.

Active Flux methods

High order, stable and conservative methods for cut cells are of high interest. In the last decade, the family of Active Flux (AF) methods was introduced and developed. The Active Flux method is a FVM developed by Eymann, Roe and coauthors [26, 27, 28, 29, 30]. It originates from scheme V from van Leer's series of articles "towards the ultimate conservative difference scheme" [31]. It makes use of a very local stencil and shows excellent third order accurate results by introducing additional degrees of freedom on the cell faces. With these point values, the reconstruction is carried out locally in each cell resulting in a globally continuous function. By evolving the point values separately from the cell average values, the true multidimensional character of the equations is captured in a dimensional unsplit scheme. The method is so far mostly developed for linear systems where the exact evolution is known, although preliminary work for nonlinear equations exists [32, 33, 34]. The local stencil and continuous reconstruction make the Active Flux method an attractive candidate for use in the presence of cut cells. In particular, we will see that the Active Flux method can handle the cut cell geometry but does not need further stabilization for the cut cell update.

The stabilization is automatically achieved in contrast to other methods.

Aims of this thesis

This thesis aims at laying the foundations of the use of the Active Flux method for cut cell grids. For the construction of a cut cell method it is required to have a better understanding of the Cartesian grid method without cut cells. Since the original Active Flux method uses a triangular grid in two dimensions, a Cartesian grid version is developed. We analyze the method in terms of accuracy and stability and discuss necessary and possible changes for the incorporation of cut cells. In the second part, we are concerned with an extension of the Active Flux method on Cartesian grids to the nonlinear case.

The outline of this thesis is the following: In Chapter 2 the Active Flux method is introduced and explained in one spatial dimension. We go into detail on how the method can be used in the context of cut cells and provide results for accuracy, stability and limiting. Chapter 3 explains the method in two spatial dimensions and especially provides the Cartesian grid version of the method. Although a lot of aspects of cut cells can already be studied in the one-dimensional case, additional features appear in the multidimensional setting. Thus, a detailed description on the adaptation of the two-dimensional Active Flux method for cut cells is given. Again, accuracy, stability and limiting are investigated. For both chapters, we focus on linear problems. Chapter 4 is concerned with an extension of the original Active Flux scheme to nonlinear systems of equations. While the original version of the Active Flux method is only defined for linear or simple nonlinear problems, we extend the method to any nonlinear hyperbolic conservation law in one and two spatial dimensions on Cartesian grids. However, the development of the nonlinear method has yet to include cut cell meshes. Finally, conclusions and outlook are presented in Chapter 5. Throughout this thesis, we use our results from [1, 2, 3]. Appendix B contains a statement about the authors contribution to this previously published work. The used programs can be found on his homepage[2].

[2] `http://www.am.uni-duesseldorf.de/~kerkmann/Forschung/`

2 The Active Flux Method in One Spatial Dimension

In this chapter we investigate equation (1.1) in one spatial dimension, i.e., $d = 1$.

2.1 Equidistant Grids

This section elaborately explains the Active Flux scheme as developed by Eymann, Roe and coauthors [26, 27, 28, 29, 30]. Some parts and figures are taken and adapted from [1, Section 2] and [3, Section 2.1].

In contrast to conventional finite volume methods, the Active Flux method operates not only with cell average values, but also with point values of the unknowns. This allows us to keep high order accuracy despite having a minimal numerical domain of dependence. These point values are located exactly on the interface between two cells in the one-dimensional case. Let M be an index set and let $(x_{i+\frac{1}{2}})_{i\in M}$ be a numerical grid (interfaces). For $i \in M$ let $C_i := [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ be the cells. With $x_i = \frac{x_{i+\frac{1}{2}}+x_{i-\frac{1}{2}}}{2}$ we denote the cell centers. In addition, let $0 = t_0 < t_1 < t_2 < \dots$ be a temporal discretization with $\Delta t := t_{n+1} - t_n \ \forall n \in \mathbb{N}_0$. We use

$$Q^n_{i+\frac{1}{2}} \approx q^n_{i+\frac{1}{2}} := q(x_{i+\frac{1}{2}}, t_n) \tag{2.1}$$

for the approximate point value on the interface $x_{i+\frac{1}{2}}$ at time t_n and

$$Q^n_i \approx \bar{q}^n_i := \frac{1}{x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x, t_n)\ \mathrm{d}x \tag{2.2}$$

for the approximate cell average in cell C_i, also at time t_n. If we speak about the exact solution, we use the letter q. Likewise the analytic flux function is denoted by the letter f, and the numerical flux evaluated at an interface $x_{i+\frac{1}{2}}$ across a time interval $[t_n, t_{n+1}]$ by $F^n_{i+\frac{1}{2}}$. A graphical depiction of the degrees of freedom is given in Figure 2.1.

Since we are concerned with a certain area in practical calculations, we consider (1.1) on a fixed interval $[x_l, x_r]$. For now, we consider periodic boundary conditions. Let $x_l = x_{\frac{1}{2}}, \dots, x_{N+\frac{1}{2}} = x_r$ be an equidistant numerical grid and $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} \ \forall i \in \{1, \dots, N\}$. From the initial values we construct cell average values Q^0_i and point values

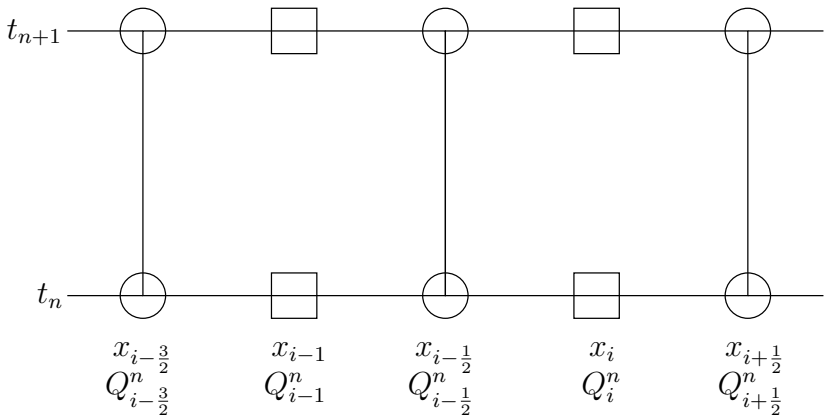

Figure 2.1.: Degrees of freedom in the one-dimensional Active Flux method.

$Q^0_{i+\frac{1}{2}}$. While the latter can be extracted through evaluation of q_0, i.e.,

$$Q^0_{i+\frac{1}{2}} = q_0(x_{i+\frac{1}{2}}), \tag{2.3}$$

we require the cell average values to be approximated to at least third order accuracy:

$$Q^0_i = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q_0(x) \, \mathrm{d}x + \mathcal{O}(\Delta x^3) \tag{2.4}$$

If the antiderivative to q_0 is not known, this can be done by an appropriate quadrature rule. We use Simpson's rule for the determination of these values, which is even a quadrature rule of fourth order accuracy.

For the complete description of the method we need to know how point values and cell average values are updated. First, we integrate (1.1) over a cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ as well as a time interval $[t_n, t_{n+1}]$ and we obtain after the application of the fundamental theorem of calculus:

$$\bar{q}_i^{n+1} = \bar{q}_i^n - \frac{1}{\Delta x} \int_{t_n}^{t_{n+1}} f(q(x_{i+\frac{1}{2}}, t)) - f(q(x_{i-\frac{1}{2}}, t)) \, \mathrm{d}t \tag{2.5}$$

To obtain a FVM, we approximate the integrals at both interfaces:

$$F^n_{i+\frac{1}{2}} \approx \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} f(q(x_{i+\frac{1}{2}}, t)) \, \mathrm{d}t \tag{2.6}$$

An analog formula holds for $F^n_{i-\frac{1}{2}}$. This integral is ought to be approximated by a quadrature rule. We remind the reader that the point value $Q^n_{i+\frac{1}{2}}$ is already known due to the choice of the degrees of freedom. Additionally, we need an update for this value for the new instant of time, i.e., $Q^{n+1}_{i+\frac{1}{2}}$. This advocates the use of a Gauss-Lobatto

quadrature rule. To obtain the desired third order, we again use Simpson's rule:

$$F^n_{i+\frac{1}{2}} := \frac{1}{6}\left(f(Q^n_{i+\frac{1}{2}}) + 4f(Q^{n+\frac{1}{2}}_{i+\frac{1}{2}}) + f(Q^{n+1}_{i+\frac{1}{2}})\right) \tag{2.7}$$

Simpson's rule gives us a fourth order accurate approximation, but the method is limited to third order because of the third order accurate reconstruction, which is described in the following. Through the use of the approximation of the cell average values we have the FVM

$$Q^{n+1}_i = Q^n_i - \frac{\Delta t}{\Delta x}\left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}\right) \tag{2.8}$$

which in turn automatically satisfies the conservation law in the discrete setting. We now discuss the computation of the point values at later times. For our method, we need the values $Q^{n+\frac{1}{2}}_{i+\frac{1}{2}}$ as well as $Q^{n+1}_{i+\frac{1}{2}}$. This is arguably the most interesting and most important aspect of the method, since the use of point values of the conserved quantities separates this method from most other finite volume methods. Let $\mathcal{L}_f(q(x,t),\Delta t)$ be the evolution operator of (1.1), i.e., $\mathcal{L}_f(q(x,t),\tau) = q(x,t+\tau)$. If we know this operator, then we can find the sought-after values with its help. We can specify the exact solution operator for any linear hyperbolic system in one dimension with the help of characteristic theory. For nonlinear equations this is generally not possible. In Chapter 4 we state an extension of our method to nonlinear hyperbolic systems and discuss which approximate evolution operator can be used in that case. The extension centeres around this step and in fact changes the original method only in this step.

As a representative for linear equations we consider the advection equation $f(q) = aq$ with positive speed of propagation $a > 0$ in the following. The evolution operator is given by $\mathcal{L}_f(q(x,t),\tau) = q(x,t+\tau) = q(x-a\tau,t)$. To determine the values at later times, we trace the characteristics back to evaluate the solution at the current time level.

To be able to evaluate the solution at any point in space, we need a global representation of the solution. We reconstruct a function Q^n_{rec} from point values and cell averages. It is piecewise defined on each cell: We use a reference cell $[0,1]$ and interpolate both interface values and require the conservation of the cell average in every cell $C_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$:

$$Q^n_{rec,i}(0) = Q^n_{i-\frac{1}{2}} \tag{2.9}$$

$$Q^n_{rec,i}(1) = Q^n_{i+\frac{1}{2}} \tag{2.10}$$

$$\int_0^1 Q^n_{rec,i}(\xi)\ \mathrm{d}\xi = Q^n_i \tag{2.11}$$

The arranging of a polynomial results in an uniquely defined parabola which can be written in the following way:

$$Q^n_{rec}(x) = Q^n_{rec,i}(\xi) \quad \forall x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \tag{2.12}$$

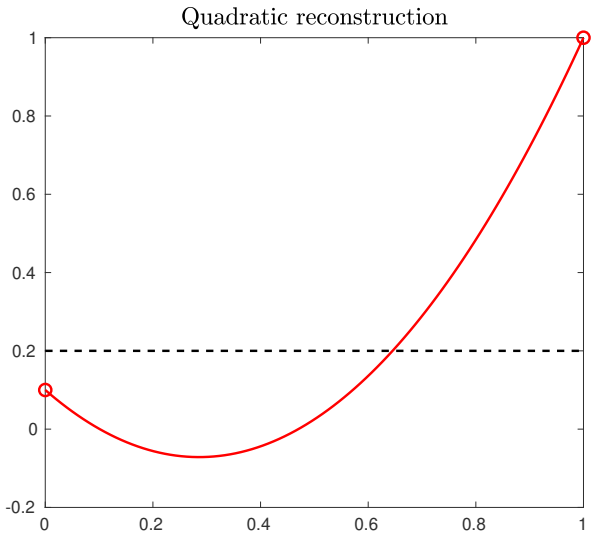

Figure 2.2.: Example reconstruction inside one cell.

with

$$\xi = \frac{x - x_{i-\frac{1}{2}}}{x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}} \tag{2.13}$$

and

$$Q^n_{rec,i}(\xi) = Q^n_{i-\frac{1}{2}}(3\xi^2 - 4\xi + 1) + Q^n_i(-6\xi^2 + 6\xi) + Q^n_{i+\frac{1}{2}}(3\xi^2 - 2\xi), \quad \xi \in [0,1]. \tag{2.14}$$

Formula (2.14) can also be displayed in a different basis. For the derivation one can compare with [1]. Figure 2.2 illustrates the reconstruction. With that, Q^n_{rec} is by definition continuous in each cell C_i and across all interfaces and therefore we have $Q^n_{rec} \in \mathcal{C}^\infty(C_i)$. Under the assumption

$$Q^n_i = \bar{q}^n_i + \mathcal{O}(\Delta x^3) \tag{2.15}$$

$$Q^n_{i+\frac{1}{2}} = q(x_{i+\frac{1}{2}}, t_n) + \mathcal{O}(\Delta x^3) \tag{2.16}$$

we furthermore have

$$Q^n_{rec}(x) = q(x, t_n) + \mathcal{O}(\Delta x^3) \tag{2.17}$$

because of the interpolation error for all $x \in [x_l, x_r]$.

We then obtain

$$Q^{n+1}_{i+\frac{1}{2}} = Q^n_{rec}(x_{i+\frac{1}{2}} - a\Delta t) \tag{2.18}$$

$$Q^{n+\frac{1}{2}}_{i+\frac{1}{2}} = Q^n_{rec}\left(x_{i+\frac{1}{2}} - a\frac{\Delta t}{2}\right) \tag{2.19}$$

by using the exact evolution. Equation (2.18) is of double importance: On the one hand, it states the update of the point values at the interfaces, on the other hand it provides the evaluation in the last node of Simpson's rule for time integration. The

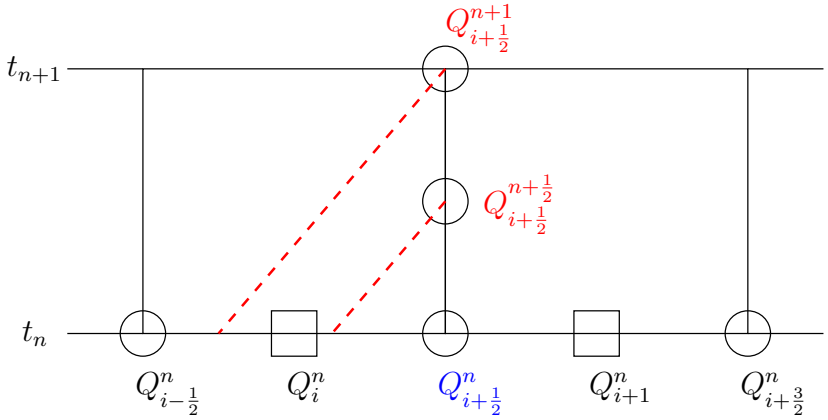

Figure 2.3.: Schematic illustration of the update of the point values at the interface $i + \frac{1}{2}$ for the advection equation.

update and the characteristics are presented in Figure 2.3.

The so presented method possesses the following properties:

- The method uses $2N + 1$ degrees of freedom (for periodic meshes $2N$).
- The order of consistency is 3.
- The used reconstruction is globally continuous.
- The method uses information solely from adjacent cells to update point values and cell average values. In particular only those cells are taken into consideration from which information are actually required. (For example advection with $a > 0$: cell C_{i+1} is not used for the update of degrees of freedom belonging to cell C_i.)
- The method converges under the time step restriction (CFL restriction)

$$\nu := \frac{|a|\Delta t}{\Delta x} \leq 1. \tag{2.20}$$

  This classical time step restriction, which occurs for explicit FVM, can be understood in that way, that characteristics which are used for the computation of the new point values and thus for the numerical flux may not leave the adjacent cells during one time step. We will later see numerically that this condition is sufficient.

Additionally to the use of point values, the locality of the numerical domain of dependence and the global continuity of the reconstruction are what distinguishes the Active Flux method from other conventional high order FVM. We will see that the continuity plays an important role in the stability of the method when applying it to a mesh with small cells. Therefore it plays an essential role in the construction of methods for complex geometries.

We now consider first numerical results of the Active Flux method. Modified equation analysis explains why methods of even order accuracy and methods of odd order

accuracy each show certain behaviors [35, pp. 235 ff.]. While methods with even order accuracy create dispersive waves, methods with odd order accuracy lead to a damping of emerging oscillations because of dissipation. The more dispersive character of the Active Flux method can be seen in comparison to the second order accurate Lax-Wendroff method in Figure 2.4. For a fair comparison we have to use the double amount of cells in the Lax-Wendroff method, respectively, since the Active Flux method contains two degrees of freedom per cell. Here, we use the linear acoustic equations:

$$q = \begin{pmatrix} p \\ v \end{pmatrix}, \quad f(q) = Aq = \begin{pmatrix} 0 & K_0 \\ \frac{1}{\rho_0} & 0 \end{pmatrix} \begin{pmatrix} p \\ v \end{pmatrix} \tag{2.21}$$

In doing so, let p be the pressure, v the velocity, ρ_0 a constant density and K_0 be the bulk modulus of compressibility. For the computations we use $\rho_0 = K_0 = 1.4$ and the initial conditions

$$q_0(x) = \begin{pmatrix} \exp(-100x^2)\sin(80(x-0.5)) \\ 0 \end{pmatrix}. \tag{2.22}$$

We use the interval $[-1, 1]$ with periodic boundary conditions, the final time $T = 7.5$ and $\nu = \frac{c_0 \Delta t}{\Delta x} \leq 0.9$, where $c_0 = \sqrt{\frac{K_0}{\rho_0}} = 1$. As the linear acoustic equations form a linear, hyperbolic system, we can decouple them through eigenvalue transformations into a diagonal system, so that every equation can be viewed as an advection equation in the characteristic variables [6, Section 2.9]. Since the newly created diagonal system again consists of conservation laws, it is possible to transform the variables before the use of the method to their characteristic form and transform them back after. In this process it is important to pay attention to the correct transformation of the cell averages. Alternatively, and equivalently, one can continue to work with the conservative, primal variables, as long as the correct characteristic speeds are used in the computation of the updates of the point values.

The solution which is obtained by the use of the Active Flux method is on one hand more accurate than the solution that is obtained by the Lax-Wendroff method and on the other hand does not exhibit a dispersive character.

The mentioned properties for accuracy and stability are not immediately evident. Before dealing with an accuracy or stability analysis of the Active Flux method, we now consider cut cell situations in one spatial dimension.

2.2 Cut Cell Grids

Because connected grids in one spatial dimension are always intervals, it is always possible to find an equidistant grid that uniformly discretizes a given interval. Nonetheless we are interested in understanding the case where a fixed grid size is given which does not divide the interval length since we thereby acquire valuable information to extend the method for actual cut cells in higher dimensions. We simulate this situation by the introduction of a small cell in an equidistant grid. This section is based on [3, Section 2.2] and amplified.

Let $x_l = x_{\frac{1}{2}}, \ldots, x_{N+\frac{1}{2}} = x_r$ and $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ $\forall i \in \{1, \ldots, N\} \setminus \{k\}$. Further-

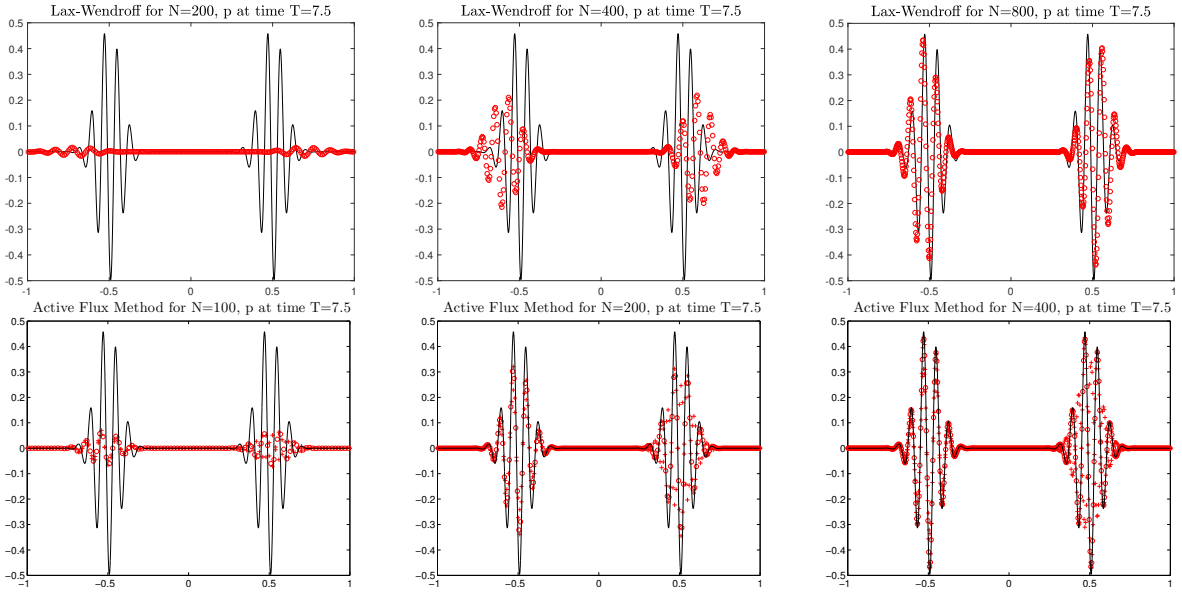

Figure 2.4.: Comparison of the Lax-Wendroff method (first row) and the Active Flux method (second row) for linear acoustics. Results of p at time $T = 7.5$ with 200 (left column), 400 (middle column) and 800 (right column) degrees of freedom are shown. The solid line is the exact solution, the red circles indicate the cell average values of p as computed by the two different methods. Point values are marked with a '+', while cell averages are marked with an 'o'.

more, let $\Delta x_k = x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}} = \alpha \Delta x$ with $0 < \alpha \leq 1$. For $\alpha = 1$ we recover the situation without a small cell. Cell C_k will from now on be called *the small cell*.

We apply the Active Flux method from the previous section. Since α can be arbitrarily small, we don't want the time step to depend on the size of the small cell. The interpretation of (2.20) lets us suspect that the stability condition changes:

$$\frac{|a|\Delta t}{\min_i x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}} = \frac{|a|\Delta t}{\alpha \Delta x} \leq 1 \tag{2.23}$$

However, this time step restriction is not necessary. The application of the method provided that simply (2.20) holds is displayed in Figure 2.5. In a natural way the characteristics that are required for interface $k + \frac{1}{2}$ now originate in cells C_{k-1} or C_k. The evolution of the point values therefore takes the respective cell of origin into account and evaluates the reconstruction at the origin. The flux update stays unchanged. Three cases arise:

1. $a\Delta t - \alpha \Delta x \leq 0$ (d. h. $\nu \leq \alpha$): Both characteristics originate in cell C_k.
2. $a\Delta t - 2\alpha \Delta x \leq 0 < a\Delta t - \alpha \Delta x$ (i.e., $\alpha < \nu \leq 2\alpha$): The characteristic of $Q_{k+\frac{1}{2}}^{n+\frac{1}{2}}$ originates in cell C_k, the characteristic of $Q_{k+\frac{1}{2}}^{n+1}$ in cell C_{k-1}.
3. $a\Delta t - 2\alpha \Delta x > 0$ (d. h. $\nu > 2\alpha$): Both characteristics originate in cell C_{k-1}.

Independently from the case we assert that the Active Flux method used on this grid is stable under condition (2.20). The next two sections deal with the accuracy and stability analysis in depth. At this point it shall be remarked that the method used on a grid with two adjacent small cells leads to an unstable approximation in case 3.

2.3 Accuracy

In numerical accuracy studies the local truncation error is frequently specified. Before doing this, we examine the method for possible error sources.

We first of all require that exact point values and cell average values are given at time t_n, i.e., (2.15) and (2.16) hold without error term. For the reconstruction it still holds (2.17). Since the exact evolution operator is used for the update of the point values, the error of approximation in the reconstruction is carried over to the new point values.

The second error source is generated in the flux computation. Suppose that ν stays constant for $\Delta t, \Delta x \to 0$. Then we obtain for the advection equation:

$$\begin{aligned} \bar{f}_{i+\frac{1}{2}}^{n+\frac{1}{2}} &:= \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} f(q(x_{i+\frac{1}{2}}, t)) \, \mathrm{d}t \\ &= \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} aq(x_{i+\frac{1}{2}} - a(t - t_n), t_n) \, \mathrm{d}t \\ &= \frac{1}{\Delta t} \int_{x_{i+\frac{1}{2}} - a\Delta t}^{x_{i+\frac{1}{2}}} q(x, t_n) \, \mathrm{d}x \end{aligned} \tag{2.24}$$

The integral in time can thus be interpreted by exact evolution using characteristics for the advection equation as an integral in space. Moreover, it now follows from inserting the reconstruction that

$$\begin{aligned} \bar{f}_{i+\frac{1}{2}}^{n+\frac{1}{2}} &= \frac{1}{\Delta t} \int_{x_{i+\frac{1}{2}} - a\Delta t}^{x_{i+\frac{1}{2}}} Q_{rec}^n(x) \, \mathrm{d}x + \mathcal{O}(\Delta x^3) \\ &= \frac{a}{6} \left(Q_{rec}^n(x_{i+\frac{1}{2}} - a\Delta t) + 4 Q_{rec}^n \left(x_{i+\frac{1}{2}} - a\frac{\Delta t}{2} \right) + Q_{rec}^n(x_{i+\frac{1}{2}}) \right) + \mathcal{O}(\Delta x^3) \\ &= F_{i+\frac{1}{2}}^n + \mathcal{O}(\Delta x^3). \end{aligned} \tag{2.25}$$

As the reconstruction is locally quadratic, no error is introduced by the use of Simpson's rule in the second step as long as $x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} \geq a\Delta t$. This condition is ensured by (2.20) for regular cells. For this reason the approximation error of the flux is completely determined by the reconstruction. But for small cells this condition doesn't hold in general (compare cases 2 and 3 from the previous section). In these cases we integrate a piecewise smooth function. The use of Simpson's rule can here also be understood as exact integration of a different reconstruction which is given by the parabola determined by the three used quadrature nodes that originate from the present reconstruction.

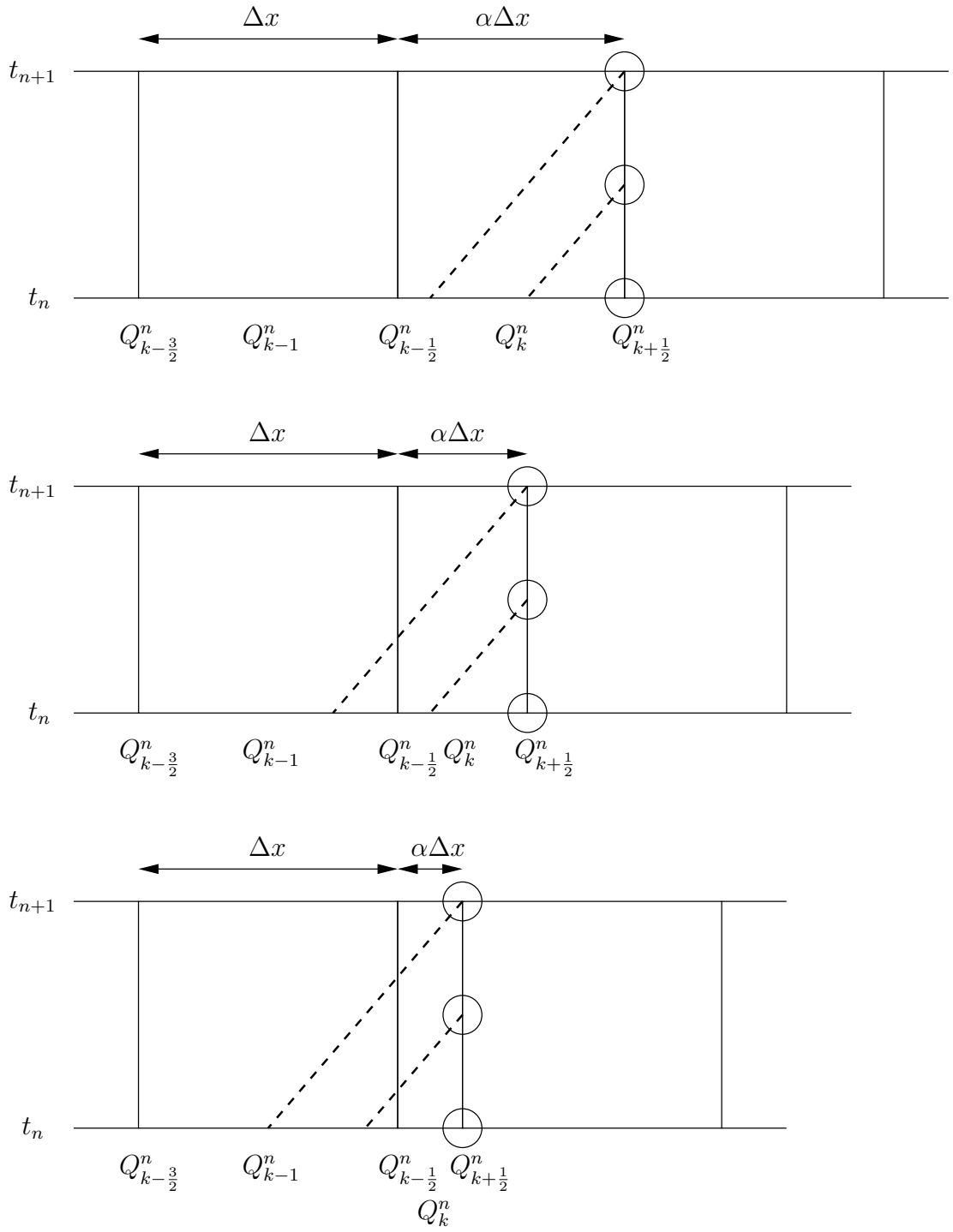

Figure 2.5.: Active Flux point value update under the presence of a small cell. Top: Case 1. Center: Case 2. Bottom: Case 3.

Since all points of the reconstruction approximate the exact solution to third order, this new parabola defines an implicit used reconstruction that is also third order accurate to the exact solution. If this implicitly given reconstruction is used in the first step of (2.25), then we immidiately see that the third order accuracy is recovered and the approximation error of the flux is given entirely by this implicitly given reconstruction as well. That reconstruction can also be considered as the use of an h-box of length $a\Delta t$ without explicitly determining its form (compare [13, 14, 15]).

We now give the usual definition of the local truncation error.

Definition 2.3.1. The *local truncation error* τ_i^n in a cell C_i at time t_n is defined by

$$\tau_i^n = \frac{\bar{q}_i^{n+1} - \bar{q}_i^n + \frac{\Delta t}{\Delta x}\left(\widetilde{F}_{i+\frac{1}{2}}^n - \widetilde{F}_{i-\frac{1}{2}}^n\right)}{\Delta t}, \tag{2.26}$$

where $\widetilde{F}_{i+\frac{1}{2}}^n$ and $\widetilde{F}_{i-\frac{1}{2}}^n$ denote the numerical fluxes given by the assumption that exact point values $Q_{i+\frac{1}{2}}^n$ and cell average values Q_i^n are used in the reconstruction.

From the previous observation we now immediately see that we only have $\tau_i^n = \mathcal{O}(\Delta t^2)$. The error does therefore not show the wanted and in practice seen third order accuracy after formally stating the method. Indeed, the local truncation error can be stated exactly for the advection equation (see [3, Lemma 1]):

Theorem 2.3.2. *Let $q \in \mathcal{C}^4(\mathbb{R})$ and let (2.20) be true. The local truncation error of the one-dimensional Active Flux method for the advection equation, as introduced in Section 2.1, on an equidistant grid for any cell C_i at time t_n reads*

$$\tau_i^n = \frac{1}{24} a \Delta x^3 \nu (1-\nu)^2 \frac{\partial^4}{\partial x^4} q(x_i, t_n) + \mathcal{O}(\Delta x^4). \tag{2.27}$$

Proof. We use the following formula to convert from point values to cell average values and vice versa [36]:

$$\bar{q}_i^n = q(x_i, t_n) + \frac{\Delta x^2}{24} q_{xx}(x_i, t_n) + \frac{\Delta x^4}{1920} q_{xxxx}(x_i, t_n) + \mathcal{O}(\Delta x^6) \tag{2.28}$$

Also let (2.20) be true. At this point, we will drop the arguments of $q(x_i, t_n)$ and all its derivatives for shorter notation. Remaining terms of Taylor series expansions will be collected in one term $\mathcal{O}(\Delta t^4) = \mathcal{O}(\Delta x^4)$. Let us consider the first part of the local

truncation error:

$$\begin{aligned}\frac{\bar{q}_i^{n+1}-\bar{q}_i^n}{\Delta t} =& \frac{q(x_i,t_{n+1})+\frac{\Delta x^2}{24}q_{xx}(x_i,t_{n+1})+\frac{\Delta x^4}{1920}q_{xxxx}(x_i,t_{n+1})}{\Delta t}\\ &-\frac{q+\frac{\Delta x^2}{24}q_{xx}+\frac{\Delta x^4}{1920}q_{xxxx}}{\Delta t}+\mathcal{O}(\Delta t^4)\\ =& \frac{q+\Delta t q_t+\frac{\Delta t^2}{2}q_{tt}+\frac{\Delta t^3}{6}q_{ttt}+\frac{\Delta t^4}{24}q_{tttt}+\frac{\Delta x^2}{24}\left(q_{xx}+\Delta t q_{xxt}+\frac{\Delta t^2}{2}q_{xxtt}\right)}{\Delta t}\\ &+\frac{\frac{\Delta x^4}{1920}q_{xxxx}-q-\frac{\Delta x^2}{24}q_{xx}-\frac{\Delta x^4}{1920}q_{xxxx}}{\Delta t}\\ =&-aq_x+\frac{\nu}{2}a\Delta x q_{xx}+\left(-\frac{\nu^2}{6}-\frac{1}{24}\right)a\Delta x^2 q_{xxx}+\left(-\frac{\nu^3}{24}+\frac{\nu}{48}\right)a\Delta x^3 q_{xxxx}\end{aligned} \tag{2.29}$$

Now let

$$\widetilde{Q}^n_{rec,i}(\xi)=q(x_{i-\frac{1}{2}},t_n)(3\xi^2-4\xi+1)+\bar{q}_i^n(-6\xi^2+6\xi)+q(x_{i+\frac{1}{2}},t_n)(3\xi^2-2\xi),\quad \xi\in[0,1], \tag{2.30}$$

be the reconstruction in any cell C_i under the use of the exact solution at time t_n. Then we get for the second part of the local truncation error:

$$\begin{aligned}&\frac{\widetilde{F}^n_{i+\frac{1}{2}}-\widetilde{F}^n_{i-\frac{1}{2}}}{\Delta x}\\ =&\frac{a}{6\Delta x}\Big[\ \widetilde{Q}^n_{rec,i}(1-\nu)+4\widetilde{Q}^n_{rec,i}\left(1-\frac{\nu}{2}\right)+q(x_{i+\frac{1}{2}},t_n)\\ &\quad-\widetilde{Q}^n_{rec,i-1}(1-\nu)-4\widetilde{Q}^n_{rec,i-1}\left(1-\frac{\nu}{2}\right)-q(x_{i-\frac{1}{2}},t_n)\Big]\\ =&\frac{a}{6\Delta x}\Big[\ (q(x_{i-\frac{1}{2}},t_n)-q(x_{i-\frac{3}{2}},t_n))\\ &\quad\cdot\left(3(1-\nu)^2-4(1-\nu)+1)+4\left(3\left(1-\frac{\nu}{2}\right)^2-4\left(1-\frac{\nu}{2}\right)+1\right)\right)\\ &\quad+(\bar{q}_i^n-\bar{q}_{i-1}^n)\left(-6(1-\nu)^2+6(1-\nu)+4(-6\left(1-\frac{\nu}{2}\right)^2+6\left(1-\frac{\nu}{2}\right)\right)\\ &\quad+(q(x_{i+\frac{1}{2}},t_n)-q(x_{i-\frac{1}{2}},t_n))\\ &\quad\cdot\left(3(1-\nu)^2-2(1-\nu)+4\left(3\left(1-\frac{\nu}{2}\right)^2-2\left(1-\frac{\nu}{2}\right)\right)+1\right)\Big]\\ =&\frac{a}{\Delta x}\Big[\ \ q(x_{i-\frac{3}{2}},t_n)(-\nu^2+\nu)+(\bar{q}_i^n-\bar{q}_{i-1}^n)(-2\nu^2+3\nu)\\ &\quad+q(x_{i-\frac{1}{2}},t_n)(\nu-1)+q(x_{i+\frac{1}{2}},t_n)(\nu^2-2\nu+1)\Big]\end{aligned} \tag{2.31}$$

$$
\begin{aligned}
&= \frac{a}{\Delta x}\Bigg[\quad \left(q - \frac{3}{2}\Delta x q_x + \frac{9}{4}\frac{\Delta x^2}{2}q_{xx} - \frac{27}{8}\frac{\Delta x^3}{6}q_{xxx} + \frac{81}{16}\frac{\Delta x^4}{24}q_{xxxx}\right)(-\nu^2+\nu) \\
&\qquad + \Bigg(q + \frac{\Delta x^2}{24}q_{xx} + \frac{\Delta x^4}{1920}q_{xxxx} - q(x_{i-1},t_n) \\
&\qquad\quad - \frac{\Delta x^2}{24}q_{xx}(x_{i-1},t_n) - \frac{\Delta x^4}{1920}q_{xxxx}(x_{i-1},t_n)\Bigg)(-2\nu^2+3\nu) \\
&\qquad + \left(q - \frac{1}{2}\Delta x q_x + \frac{1}{4}\frac{\Delta x^2}{2}q_{xx} - \frac{1}{8}\frac{\Delta x^3}{6}q_{xxx} + \frac{1}{16}\frac{\Delta x^4}{24}q_{xxxx}\right)(\nu-1) \\
&\qquad + \left(q + \frac{1}{2}\Delta x q_x + \frac{1}{4}\frac{\Delta x^2}{2}q_{xx} + \frac{1}{8}\frac{\Delta x^3}{6}q_{xxx} + \frac{1}{16}\frac{\Delta x^4}{24}q_{xxxx}\right)(\nu^2-2\nu+1)\Bigg] + \mathcal{O}(\Delta x^4) \\
&= \frac{a}{\Delta x}\Bigg[\quad \left(q - \frac{3}{2}\Delta x q_x + \frac{9}{4}\frac{\Delta x^2}{2}q_{xx} - \frac{27}{8}\frac{\Delta x^3}{6}q_{xxx} + \frac{81}{16}\frac{\Delta x^4}{24}q_{xxxx}\right)(-\nu^2+\nu) \\
&\qquad + \Bigg(q + \frac{\Delta x^2}{24}q_{xx} + \frac{\Delta x^4}{1920}q_{xxxx} - q + \Delta x q_x - \frac{\Delta x^2}{2}q_{xx} + \frac{\Delta x^3}{6}q_{xxx} - \frac{\Delta x^4}{24}q_{xxxx} \\
&\qquad\quad - \frac{\Delta x^2}{24}\left(q_{xx} - \Delta x q_{xxx} + \frac{\Delta x^2}{2}q_{xxxx}\right) - \frac{\Delta x^4}{1920}q_{xxxx}\Bigg)(-2\nu^2+3\nu) \\
&\qquad + \left(q - \frac{1}{2}\Delta x q_x + \frac{1}{4}\frac{\Delta x^2}{2}q_{xx} - \frac{1}{8}\frac{\Delta x^3}{6}q_{xxx} + \frac{1}{16}\frac{\Delta x^4}{24}q_{xxxx}\right)(\nu-1) \\
&\qquad + \left(q + \frac{1}{2}\Delta x q_x + \frac{1}{4}\frac{\Delta x^2}{2}q_{xx} + \frac{1}{8}\frac{\Delta x^3}{6}q_{xxx} + \frac{1}{16}\frac{\Delta x^4}{24}q_{xxxx}\right)(\nu^2-2\nu+1)\Bigg] + \mathcal{O}(\Delta x^4) \\
&= \frac{a}{\Delta x}\Bigg[\quad q\left(-\nu^2+\nu+\nu-1+\nu^2-2\nu+1\right) \\
&\qquad + \Delta x q_x\left(-\frac{3}{2}(-\nu^2+\nu) - 2\nu^2 + 3\nu - \frac{1}{2}(\nu-1) + \frac{1}{2}(\nu^2-2\nu+1)\right) \\
&\qquad + \Delta x^2 q_{xx}\left(\frac{9}{8}(-\nu^2+\nu) - \frac{1}{2}(-2\nu^2+3\nu) + \frac{1}{8}(\nu-1) + \frac{1}{8}(\nu^2-2\nu+1)\right) \\
&\qquad + \Delta x^3 q_{xxx}\left(-\frac{27}{48}(-\nu^2+\nu) + \frac{5}{24}(-2\nu^2+3\nu) - \frac{1}{48}(\nu-1) + \frac{1}{48}(\nu^2-2\nu+1)\right) \\
&\qquad + \Delta x^4 q_{xxxx}\left(\frac{81}{384}(-\nu^2+\nu) - \frac{1}{48}(-2\nu^2+3\nu) + \frac{1}{384}(\nu-1) + \frac{1}{384}(\nu^2-2\nu+1)\right)\Bigg] \\
&\quad + \mathcal{O}(\Delta x^4) \\
&= a q_x - \frac{\nu}{2}\Delta x q_{xx} + \left(\frac{\nu^2}{6} + \frac{1}{24}\right)\Delta x^2 q_{xxx} + \left(-\frac{\nu^3}{12} + \frac{\nu}{48}\right)\Delta x^3 q_{xxxx} + \mathcal{O}(\Delta x^4)
\end{aligned}
$$

The addition of both parts yields:

$$\tau_i^n = \frac{1}{24}a\Delta x^3\nu(1-\nu)^2\frac{\partial^4}{\partial x^4}q(x_i,t_n) + \mathcal{O}(\Delta x^4) \tag{2.32}$$

□

The method thus actually exhibits order of consistency three. That means that the

leading error term in both fluxes cancels so that the flux difference gains one order of accuracy. Practical experiments on the initial data $q_0(x) = x^3$ conform these findings.

Remark 2.3.3. Would the used point values in Simpson's rule be of fourth order accuracy, a third order accurate method would emerge due to the fourth order accuracy of Simpson's rule even without the cancellation of the leading error terms of the flux difference.

Let us now consider the situation with one small cell. Particularly interesting is case 3 since in practical applications the most relevant difficult cases are the small cells with $\alpha \ll 1$. We have (compare (c) to [3, Lemma 2]):

Theorem 2.3.4. *Let $q \in \mathcal{C}^3(\mathbb{R})$. Let there be a grid with one small cell as defined in Section 2.2. The local truncation error of the one-dimensional Active Flux method for the advection equation, as developed in Section 1.2, in the small cell C_k at time t_n reads,*

(a) if $a\Delta t - \alpha\Delta x \leq 0$ (case 1),

$$\tau_k^n = \frac{1}{24}\nu(2\nu - 1)\left(\frac{1}{\alpha} - 1\right) a\Delta x^2 \frac{\partial^3}{\partial x^3} q(x_k, t_n) + \mathcal{O}(\Delta x^3). \tag{2.33}$$

(b) if $a\Delta t - 2\alpha\Delta x \leq 0 < a\Delta t - \alpha\Delta x$ (case 2),

$$\begin{aligned}\tau_k^n = &\left(-\frac{1}{72} - \frac{\alpha}{24} - \frac{\alpha^2}{36} + \left(\frac{1}{12} + \frac{\alpha}{9} - \frac{1}{36\alpha}\right)\nu - \left(\frac{1}{8} - \frac{1}{24\alpha}\right)\nu^2\right) \\ &\cdot a\Delta x^2 \frac{\partial^3}{\partial x^3} q(x_k, t_n) + \mathcal{O}(\Delta x^3).\end{aligned} \tag{2.34}$$

(c) if $a\Delta t - 2\alpha\Delta x > 0$ (case 3),

$$\tau_k^n = \left(-\frac{5}{72} - \frac{5}{24}\alpha - \frac{5}{36}\alpha^2 + \frac{1}{4}(1+\alpha)\nu - \frac{1}{6}\nu^2\right) a\Delta x^2 \frac{\partial^3}{\partial x^3} q(x_k, t_n) + \mathcal{O}(\Delta x^3). \tag{2.35}$$

Furthermore the local truncation error in cell C_{k+1} has the same form, multiplied by $-\alpha$, and the third derivative in $q(x_{k+1}, t_n)$.

The computations can be followed in the program **adv_truncation_errors.ipynb**.

We see that the order of consistency in fact reduces to second order in the small cell and its right neighbor here.

To circumvent this reduction in order, we can replace Simpson's rule for this one flux computation by exact integration. The other integrals are already integrated exactly using Simpson's rule. One possible representation of the exact integration is an iterated Simpson's rule over the respective parts of cells C_{k-1} and C_k. Because the area of integration includes cell C_k entirely, one can also immediately use the cell average Q_k^n for this part of the iterated Simpson's rule. One other possible interpretation of this exact integration is the conception as an h-box of length $a\Delta t$, again. The wanted

N	L_1-error	EOC	L_∞-error	EOC	e_k	EOC	e_{k+1}	EOC
50	$4.1623 \cdot 10^{-5}$		$6.5860 \cdot 10^{-5}$		$1.9129 \cdot 10^{-5}$		$2.8565 \cdot 10^{-5}$	
100	$5.1223 \cdot 10^{-6}$	2.99	$8.0811 \cdot 10^{-6}$	3.00	$2.7028 \cdot 10^{-6}$	2.79	$4.1193 \cdot 10^{-6}$	2.77
200	$6.3511 \cdot 10^{-7}$	3.00	$9.9977 \cdot 10^{-7}$	3.00	$3.5630 \cdot 10^{-7}$	2.91	$5.4654 \cdot 10^{-7}$	2.90
400	$7.9065 \cdot 10^{-8}$	3.00	$1.2432 \cdot 10^{-7}$	3.00	$4.5657 \cdot 10^{-8}$	2.96	$7.0208 \cdot 10^{-8}$	2.95

Table 2.1.: Convergence study for the advection equation in the presence of a small cell (cut cell) for $\nu \approx 0.25$ and $\alpha = 0.3$ (case 1).

N	L_1-error	EOC	L_∞-error	EOC	e_k	EOC	e_{k+1}	EOC
50	$2.5625 \cdot 10^{-5}$		$4.1114 \cdot 10^{-5}$		$9.5437 \cdot 10^{-5}$		$1.5489 \cdot 10^{-5}$	
100	$3.1556 \cdot 10^{-6}$	2.99	$5.0059 \cdot 10^{-6}$	3.01	$1.3092 \cdot 10^{-6}$	2.84	$2.2303 \cdot 10^{-6}$	2.77
200	$3.9115 \cdot 10^{-7}$	3.00	$6.1730 \cdot 10^{-7}$	3.00	$1.7032 \cdot 10^{-7}$	2.93	$2.9567 \cdot 10^{-7}$	2.90
400	$4.8680 \cdot 10^{-8}$	3.00	$7.6633 \cdot 10^{-8}$	3.00	$2.1689 \cdot 10^{-8}$	2.97	$3.7966 \cdot 10^{-8}$	2.95

Table 2.2.: Convergence study for the advection equation in the presence of a small cell (cut cell) for $\nu \approx 0.5$ and $\alpha = 0.3$ (case 2).

integral value matches the cell average value of this h-box exactly. It is computed as stated above.

We check the found order reduction with the help of a numerical experiment: Let $a = 1$, $\alpha = 0.3$, $T = 0.6$, $x_l = 0$, $x_r = 1$ and $q_0(x) = \sin(2\pi x)$. We apply the Active Flux method for $\nu \approx 0.25$ (case 1), $\nu \approx 0.5$ (case 2) and $\nu \approx 0.8$ (case 3) and measure the global error in the L_1-norm und L_∞-norm as well as the error in cells C_k and C_{k+1}, here denoted by e_k and e_{k+1}, respectively. The results can be seen in Tables 2.1, 2.2 and 2.3. The time step is slightly adjusted to meet the final time and to be of uniform size. Despite second order accuracy of the local truncation error we obtain third order convergence in all cases apart from the small case and therefore the L_∞-norm in case 3. The following remark is adapted from [3, Remark 1].

Remark 2.3.5. This can be explained as follows: The new cell average Q_{k+1}^{n+1} will be used in the subsequent time step in the computation of the flux $F_{k+\frac{3}{2}}^{n+1}$. A part of the error that is proportional to ν will be transported to cell C_{k+2}. The same effect appears in cell C_{k+2} one more time step later. That means that the resulting error in cell C_{k+1} spreads to all other $n-k = \mathcal{O}(\frac{1}{\Delta x})$ cells over time. Albeit the one step error $\Delta t \tau_{k+1}^n = \mathcal{O}(\Delta x^3)$ is created in every step, the error in cell C_{k+1} (and all subsequent cells) is bounded

N	L_1-error	EOC	L_∞-error	EOC	e_k	EOC	e_{k+1}	EOC
50	$1.1207 \cdot 10^{-5}$		$1.9262 \cdot 10^{-5}$		$1.9262 \cdot 10^{-5}$		$4.9990 \cdot 10^{-6}$	
100	$1.4272 \cdot 10^{-6}$	2.94	$3.6581 \cdot 10^{-6}$	2.37	$3.6581 \cdot 10^{-6}$	2.37	$7.9021 \cdot 10^{-7}$	2.63
200	$1.7611 \cdot 10^{-7}$	3.00	$9.4140 \cdot 10^{-7}$	1.95	$9.4140 \cdot 10^{-7}$	1.95	$1.0385 \cdot 10^{-7}$	2.91
400	$2.1870 \cdot 10^{-8}$	3.00	$2.3858 \cdot 10^{-7}$	1.98	$2.3858 \cdot 10^{-7}$	1.98	$1.3279 \cdot 10^{-8}$	2.96

Table 2.3.: Convergence study for the advection equation in the presence of a small cell (cut cell) for $\nu \approx 0.8$ and $\alpha = 0.3$ (case 3).

through this harmonic property and we recover third order of consistency even in this cell. This effect does not happen in cell 3 since the cell average Q_k^n is never used in the update. For this, compare to the third plot of Figure 2.5. In the other two cases the cell average is used (left and center plot) and this effect is even present in the small cell C_k.

2.4 Linear Stability and Cancellation Property

The first part of this section is adapted from [3, Section 2.4]. To examine the linear stability of the one-dimensional Active Flux method in the presence of small cells we rewrite the method in the matrix-vector form

$$\mathbf{Q}^{n+1} = A\mathbf{Q}^n, \tag{2.36}$$

where the vector $\mathbf{Q}^n$ contains all degrees of freedom at time t_n. The matrix A describes the method. This is possible since the Active Flux method is linear. The method is Lax-Richtmyer stable if and only if $\|A^n\|$ is bounded independently of n. By the use of the Jordan decomposition of A it is possible to show that this is equivalent to the statement that $|\lambda| \leq 1$ for all eigenvalues λ of A and if $|\lambda| = 1$ then the geometric and algebraic multiplicity have to match [37]. Since we can't compute analytic expressions for the eigenvalues of A we determine approximations with the help of a program for differently sized grids. The results for the advection equation are visible in Figure 2.6. In the left plot we show the eigenvalues of matrix A which belongs to the Active Flux method on an equidistant grid. In the center plot the situation with one small cell of size $\alpha\Delta x \leq \frac{1}{2}a\Delta t$ (case 3) is shown. The right plot displays the eigenvalues of a matrix that occurs if every other cell is small regarding the same condition (case 3). In all cases we numerically see the stability of the method under the time step restriction (2.20).

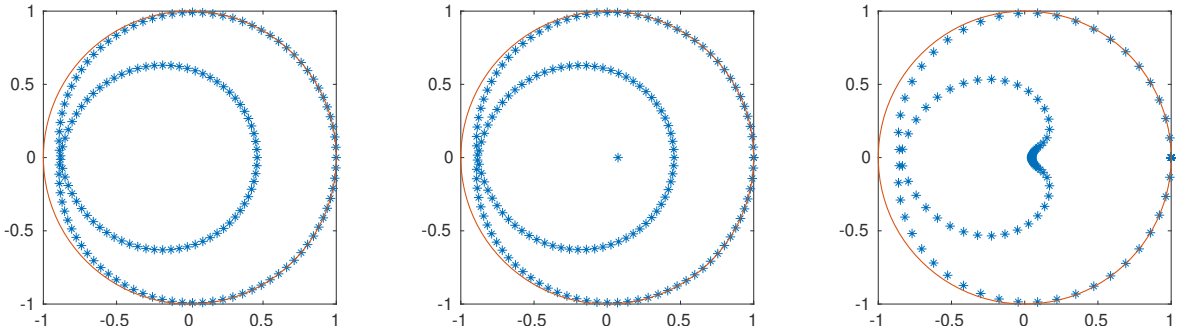

Figure 2.6.: Results of the linear stability analysis of the one-dimensional Active Flux method for the advection equation with an equidistant grid (left), one small cell (center) and a grid with alternating cell sizes (right). The used CFL number is $\nu = 0.9$. The cut cell size is $\alpha = 0.05$ for every small cell.

As introductorily mentioned, the cancellation property is a necessary condition for the stability of the method. For this purpose we show the following theorem:

Theorem 2.4.1. *For the update of the small cell C_k in the Active Flux method for the advection equation, as explained in Section 2.2, it holds:*

$$F^n_{k+\frac{1}{2}} - F^n_{k-\frac{1}{2}} = \mathcal{O}(\alpha\Delta x) \tag{2.37}$$

Proof. Let

$$F^n_{k+\frac{1}{2}} = \sum_{i=0}^{m} a_i f(Q^n_{rec}(s_i)) \tag{2.38}$$

be the used quadrature formula for the flux with weights $a_i > 0$ and nodes $s_i \in [x_{k+\frac{1}{2}} - a\Delta t, x_{k+\frac{1}{2}}]$. In our case this is Simpson's rule but the statement holds for any quadrature formula. Then we have

$$F^n_{k-\frac{1}{2}} = \sum_{i=0}^{m} a_i f(Q^n_{rec}(s_i - \alpha\Delta x)) \tag{2.39}$$

because of the exact solution operator of the advection equation. Furthermore Q^n_{rec} is locally Lipschitz continuous since it is continuous and consists of piecewise defined parabolas. Let $L_Q \geq 0$ be the corresponding Lipschitz constant. We get:

$$\begin{aligned}
\left|F^n_{k+\frac{1}{2}} - F^n_{k-\frac{1}{2}}\right| &= \left|\sum_{i=0}^{m} a_i (f(Q^n_{rec}(s_i)) - f(Q^n_{rec}(s_i - \alpha\Delta x)))\right| \\
&\leq \sum_{i=0}^{m} a_i a |Q^n_{rec}(s_i) - Q^n_{rec}(s_i - \alpha\Delta x)| \\
&\leq \sum_{i=0}^{m} a_i a L_Q |s_i - (s_i - \alpha\Delta x)| \\
&= \sum_{i=0}^{m} a_i a L_Q \alpha\Delta x = \mathcal{O}(\alpha\Delta x)
\end{aligned} \tag{2.40}$$

□

So the method fulfills the cancellation property without the need of further stabilization as long as the reconstruction is continuous.

Remark 2.4.2. Using exact integration, we see that the area of integration of the two fluxes for the update of the small cell differs only in an area of size $2\alpha\Delta x$. Hence we automatically recover the cancellation property. In doing so it is not necessary that the reconstruction is continuous across the interface. Exact integration thus provides a good alternative if it is not possible to reconstruct continuously across the interface. With it, we see an improvement in accuracy as well as a possible stabilization. This will be relevant in the limiting in Section 2.5 as well as in the multidimensional Active Flux method in Chapter 3.

Remark 2.4.3. With the use of exact integration it is possible to apply the method for arbitrarily large time steps. We can thus state a so-called *large time step* method which doesn't satisfy the condition (2.20) but is nevertheless stable. The larger time steps can even slightly speed up the method since the already known cell averages can be used while integrating over whole cells and no quadrature formula has to be used. However, this is only possible for linear equations and doesn't work for nonlinear equations.

2.5 Limiting

In this section we investigate the behavior of the Active Flux method for the advection equation for piecewise continuous initial data which involve one or more jumps. It is adopted from [1, Section 5] and presented in more detail and with additional content. The exact solution shifts in the course of time with speed a so that the discontinuities persist. We consider the interval $[0, 1]$ with periodic boundary conditions and the following initial data:

$$q_0(x) = \begin{cases} 1 + \exp(-100(x - 0.3)^2) & : x \in [0.6, 0.8] \\ \exp(-100(x - 0.3)^2) & : x \in [0, 1] \setminus [0.6, 0.8] \end{cases} \tag{2.41}$$

Obviously these inital conditons are technically speaking not periodic. The jump at the boundary is really small however so it doesn't have an effect in the computations. We furthermore use $a = 1$ and the final time $T = 1$. Also we have $\nu = 0.9$ for all time steps except the last one to meet the final time. Figure 2.7 illustrates the results for $N = 200$. While the smooth part of the solution is approximated very well as expected, some over- and undershoots appear near the discontinuities. With that we receive new maxima and minima in the solution. Indeed, the oscillations are not very distinctive. Regardless it is desirable to suppress those. We have to adapt the method. Many of the common approaches for limiting use the neighboring cells, for instance to enable a slope limiting or to pursue a WENO ansatz. Eymann and Roe propose to trace back the characteristics even further and use for the limiting in cell C_i not only the current values $Q^n_{i-\frac{1}{2}}$, Q^n_i and $Q^n_{i+\frac{1}{2}}$ but also the values of the previous time steps $Q^{n-1}_{i-\frac{1}{2}}$, Q^{n-1}_i and $Q^{n-1}_{i+\frac{1}{2}}$ [26]. The so forming numerical domain of dependence now ranges not over multiple grid cells in one time steps but over multiple time steps. First of all this method is not applicable in the first time step. Also we were not able to reproduce the shown limiting for every CFL number ν. In our work [1, 2], we consider alternative limiting possibilities that only use very local information of the current time step. Other authors describe similar approaches [32].

To conserve the locality of our method, we pursue an easier method which only works with the cell average value and the two interface values of the respective cell at the current time step and not with values from the previous time steps. Remember that the reconstruction is defined by a parabola which is uniquely determined by (2.14).

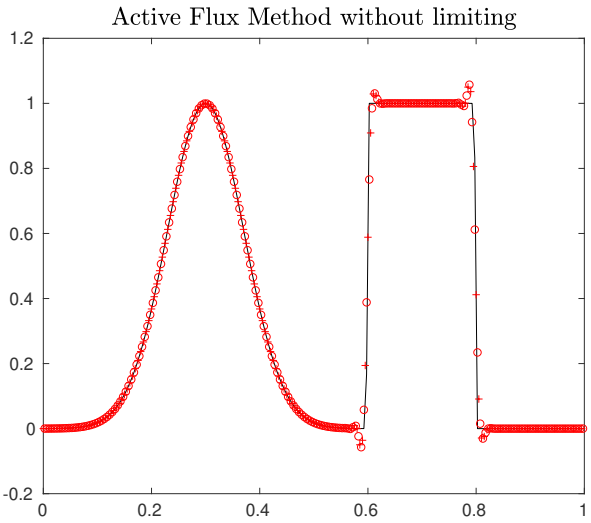

Figure 2.7.: Cell average values (o) and point values (+) of the numerical solution of the Active Flux method for the advection equation with $N = 200$ grid cells at time $T = 1$ without limiting.

Let

$$\bar{m}_i^n := \min\{Q^n_{i-\frac{1}{2}}, Q^n_{i+\frac{1}{2}}\}, \tag{2.42}$$

$$\bar{M}_i^n := \max\{Q^n_{i-\frac{1}{2}}, Q^n_{i+\frac{1}{2}}\}, \tag{2.43}$$

$$N_i := [\bar{m}_i^n, \bar{M}_i^n] \tag{2.44}$$

for all cells. The angular point of the parabola then constitutes a new extremum provided it is located inside the cell. The value of the new extremum lies outside the interval N_i. Because of the exactness of the numerical flux this can directly lead to a new extremum in the discrete values of the solution within one time step. In other words, the evolution and averaging don't produce new extrema for the advection equation as long as an exact integration is used. The limiting of the reconstruction is therefore sufficient to prevent oscillations. We thus demand:

$$Q^n_{rec,i,lim}(\xi) \in N_i \ \forall \xi \in [0,1] \tag{2.45}$$

Here, $Q^n_{rec,i,lim}$ denotes the limited reconstruction on a reference cell $[0,1]$. The transformation (2.13) is preserved and (2.12) is transferred in a canonical manner. New extrema in the reconstruction appear in two cases:

1. The cell average value lies closely to either of the two interface values. More precisely:

$$\begin{aligned} Q_i^n &< \frac{2Q_{i-\frac{1}{2}} + Q_{i+\frac{1}{2}}}{3} \wedge Q^n_{i-\frac{1}{2}} < Q^n_{i+\frac{1}{2}} \\ \text{or } Q_i^n &< \frac{Q_{i-\frac{1}{2}} + 2Q_{i+\frac{1}{2}}}{3} \wedge Q^n_{i-\frac{1}{2}} > Q^n_{i+\frac{1}{2}} \end{aligned} \tag{2.46}$$

or

$$\begin{aligned} Q_i^n &> \frac{Q_{i-\frac{1}{2}} + 2Q_{i+\frac{1}{2}}}{3} \wedge Q_{i-\frac{1}{2}}^n < Q_{i+\frac{1}{2}}^n \\ \text{or } Q_i^n &> \frac{2Q_{i-\frac{1}{2}} + Q_{i+\frac{1}{2}}}{3} \wedge Q_{i-\frac{1}{2}}^n > Q_{i+\frac{1}{2}}^n. \end{aligned} \tag{2.47}$$

2. The cell average values lies outside the interval N_i:

$$Q_i^n < m_i^n \tag{2.48}$$

or

$$Q_i^n > M_i^n. \tag{2.49}$$

Indeed, the first case is equivalent to the existence of new extrema. The simple proof of this claim is left to the reader. (2.48) clearly implies (2.46) and (2.49) clearly implies (2.47). Nonetheless, we specify case 2 separately. Since the cell average value in the case of (2.48) or (2.49) lies outside the interval N_i, we in fact expect the exact solution to feature a local extremum inside this cell. That's why we don't want to limit this case.

Would we also limit case 2, for example by a constant reconstruction $Q_{rec,i,lim}^n(x) \equiv Q_i^n$, we would get a monotone method which may lead to a loss of accuracy. By avoiding the limiting in this case we allow for local extrema at reasonable locations. Numerical experiments show that third order accuracy is recovered while a limiter is active in case 1, but not in case 2.

We have multiple options to design a limited reconstruction. In the following, we discuss some options and their respective advantages and disadvantages.

2.5.1 Limiting by Change of Basis

To keep the interpolation conditions (2.9) and (2.10) at the left and right boundary of the cell as well as the cell average (2.11) we can consult other basis functions. In (2.14) we use a basis of the polynomial space $\mathcal{P}_2([0,1])$. Two other possibilities are:

1. We use a piecewise defined function as the reconstruction. Since the cell average is close to one of the two interface values because of (2.46) or (2.47) we can assume that a discontinuity or a steep gradient in the solution is close to one of the interfaces. Let (2.46) be true with $Q_{i-\frac{1}{2}}^n < Q_{i+\frac{1}{2}}^n$. We initially approximate one part of the solution as a constant function and add a parabola for the remaining part. More precisely we have for this piecewise defined reconstruction:

$$Q_{rec,i,lim}^n(\xi) = \begin{cases} Q_{i-\frac{1}{2}}^n & : \xi \leq \xi^\star \\ Q_{i-\frac{1}{2}}^n + (Q_{i+\frac{1}{2}}^n - Q_{i-\frac{1}{2}}^n)\frac{(\xi-\xi^\star)^2}{(1-\xi^\star)^2} & : \xi > \xi^\star \end{cases} \tag{2.50}$$

where

$$\xi^\star = \frac{2Q^n_{i-\frac{1}{2}} + Q^n_{i+\frac{1}{2}} - 3Q^n_i}{Q^n_{i-\frac{1}{2}} - Q^n_{i+\frac{1}{2}}}. \tag{2.51}$$

It is easy to see that the conditions (2.9) - (2.11) are fulfilled. The case $Q^n_{i-\frac{1}{2}} > Q^n_{i+\frac{1}{2}}$ and (2.47) is treated in analogy. For this reconstruction we merely have $Q^n_{rec,i,lim} \in \mathcal{C}^1([0,1])$. Also possible is an ansatz that strings together a linear and a quadratic function or even two quadratic functions. Similar approaches are examined by Roe and coauthors [38].

2. We use a nonlinear function which is monotone. Such reconstructions are used for example by Marquina [39]. We can make the hyperbolic ansatz

$$Q^n_{rec,i,lim}(\xi) = a + \frac{b}{\xi - \frac{1}{2} + c}. \tag{2.52}$$

By requiring (2.9) - (2.11), a nonlinear system of equations for the coefficients a, b and c emerges which can be solved approximately by an iterative method. In consequence of the necessity of such an iterative method this ansatz is way more expensive than the previous. It is also ill conditioned for $Q^n_i \approx Q^n_{i-\frac{1}{2}}$ or $Q^n_i \approx Q^n_{i+\frac{1}{2}}$. However, we gain $Q^n_{rec,i,lim} \in \mathcal{C}^\infty([0,1])$, which was not true for ansatz 1.

A sample reconstruction of both approaches, compared to the standard quadratic reconstruction, is displayed in Figure 2.8. Further Figure 2.9 shows a comparison of the numerical solution of the advection equation for all three reconstructions. We see that both limitings provide a good approximation that removes the oscillations completely and resolves the discontinuities in the solution in only a few grid cells.

We notice that for both limitings an exact integration is necessary. For this, consider using Simpson's rule for either of the two limitings. It can be understood as an interpolatory quadrature formula, i.e., we form a parabola through the integrated function which is then integrated exactly. This parabola can again produce new extrema for the newly described reconstructions $Q^n_{rec,i,lim} \notin \mathcal{P}_2([0,1])$ which doesn't achieve the purpose of limiting. Exact integration is not difficult in both cases but is undesirable in general.

Next to the already mentioned disadvantages one further disadvantage stands out: None of the approaches can be extended to the multidimensional reconstruction. Furthermore, additional difficulties arise when using these reconstructions in conjunction with nonlinear equations, which are considered in Chapter 4. We therefore neglect these approaches and turn to a different idea.

2.5.2 Limiting by Discontinuous Reconstruction

Instead of changing the basis we can also decide to drop one of the conditions (2.9) - (2.11). Since the reconstruction is a global representation of the solution, we'd like

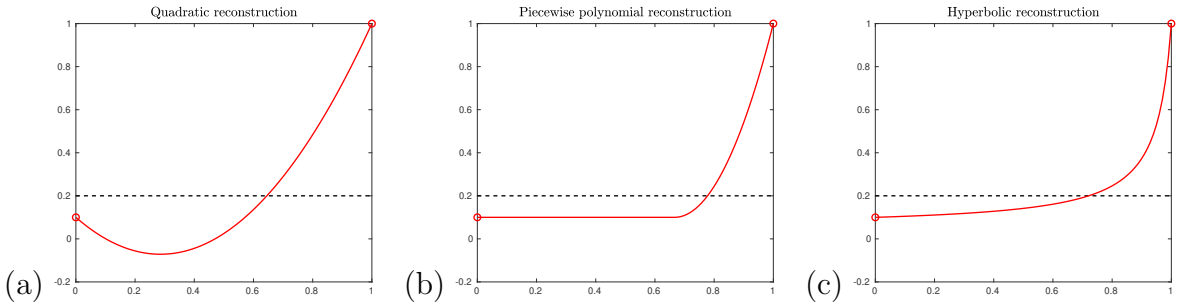

Figure 2.8.: Different reconstructions of the Active Flux method. The dashed line indicates the cell average value, the 'o' symbols at the boundaries indicate the edge values and the red solid line indicates the reconstruction. For all of these plots we used $Q^n_{i-\frac{1}{2}} = 0.1$, $Q^n_{i+\frac{1}{2}} = 1$ and $Q^n_i = 0.2$. Unlimited, quadratic reconstruction (a), piecewise polynomial reconstruction 2.50 (b), hyperbolic reconstruction 2.52 (c).

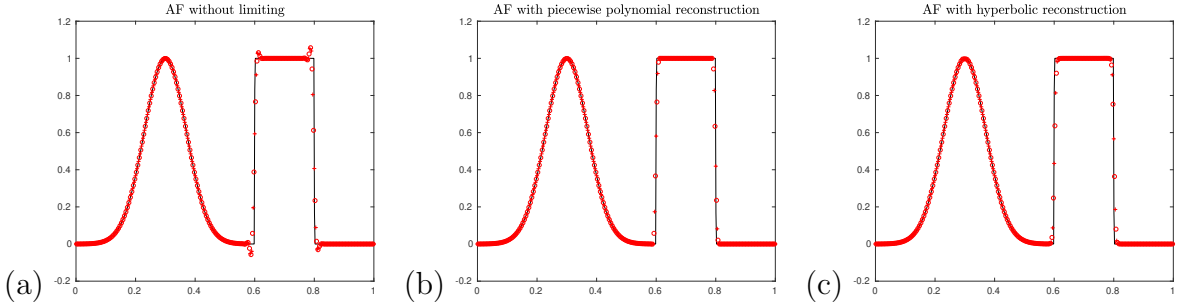

Figure 2.9.: Advection test computed with different versions of the Active Flux method using 200 grid cells. The solution is shown at time $T = 1$, i.e., after one rotation. The solution using the unlimited Active Flux method is shown in (a). In (b) and (c) we show results for the piecewise polynomial and hyperbolic reconstruction, respectively. Point values are marked with a '+', while cell averages are marked with an 'o'.

to reflect the conservation property. Thus, (2.11) shall continue to hold. We therefore have to discuss which of the conditions (2.9) - (2.10) can be dropped. The naive ansatz reads as follows: We drop the one interpolation condition which point value is further away from the cell average value. To be precise: If (2.46) with $Q^n_{i-\frac{1}{2}} < Q^n_{i+\frac{1}{2}}$ or (2.47) with $Q^n_{i-\frac{1}{2}} > Q^n_{i+\frac{1}{2}}$ holds and (2.48) doesn't hold, require only (2.9) and (2.11). Else, if (2.46) with $Q^n_{i-\frac{1}{2}} > Q^n_{i+\frac{1}{2}}$ or (2.47) with $Q^n_{i-\frac{1}{2}} < Q^n_{i+\frac{1}{2}}$ holds and (2.49) doesn't hold, require only (2.10) and (2.11).

The restriction (2.45) doesn't provide a uniquely defined quadratic reconstruction yet, but only a constraint on the third condition we can demand. Let without loss of generality (2.46) with $Q^n_{i-\frac{1}{2}} < Q^n_{i+\frac{1}{2}}$ be true and (2.48) be false. The other cases can again be treated analogously. We therefore firstly demand (2.9) and (2.11). Then it is preferable to approximate the not interpolated point value $Q^n_{i+\frac{1}{2}}$ as best as possible.

So we look for $Q^n_{rec,i,lim} \in \mathcal{P}_2([0,1])$ which solves the following minimization problem:

$$\min\big\{|Q_{i+\frac{1}{2}} - Q^n_{rec,i,lim}(1)| \;\big|\; Q^n_{rec,i,lim}(\xi) \in N_i \;\forall \xi \in [0,1]\big\}. \tag{2.53}$$

One can easily reason that this is exactly the case if and only if

$$(Q^n_{rec,i,lim})'(0) = 0. \tag{2.54}$$

This is the third additional condition which yields the following explicit representation of the reconstruction:

$$Q^n_{rec,i,lim}(\xi) = Q^n_{i-\frac{1}{2}} + 3(Q^n_i - Q^n_{i-\frac{1}{2}})\xi^2 \tag{2.55}$$

In the other case we analogously get:

$$Q^n_{rec,i,lim}(\xi) = Q^n_{i+\frac{1}{2}} + 3(Q^n_i - Q^n_{i+\frac{1}{2}})(\xi - 1)^2 \tag{2.56}$$

In contrast to the limitings in the previous section, this reconstruction is no longer globally continuous. This means that the use of the node value $Q^n_{i+\frac{1}{2}}$ (respectively $Q^n_{i-\frac{1}{2}}$) in Simpson's rule (2.7) doesn't lead to an exact integration (even without cut cells) if this interpolation condition was dropped. Equations (2.24) and (2.25) make it clear that instead the value in upwind direction has to be used: If $a > 0$, use

$$F^n_{i+\frac{1}{2}} = \frac{1}{6}\left(f(Q^n_{rec,i,lim}(1)) + 4f(Q^{n+\frac{1}{2}}_{i+\frac{1}{2}}) + f(Q^{n+1}_{i+\frac{1}{2}})\right) \tag{2.57}$$

and if $a < 0$, use

$$F^n_{i+\frac{1}{2}} = \frac{1}{6}\left(f(Q^n_{rec,i+1,lim}(0)) + 4f(Q^{n+\frac{1}{2}}_{i+\frac{1}{2}}) + f(Q^{n+1}_{i+\frac{1}{2}})\right). \tag{2.58}$$

Without limiting, this definition of the numerical flux is equivalent to (2.7) due to the continuity of the reconstruction.

Condition (2.45) in conjunction with conditions (2.9) and (2.11) can also be understood as a monotonicity condition, i.e., $(Q^n_{rec,i,lim})'(\xi) \geq 0 \;\forall \xi \in [0,1]$ or $(Q^n_{rec,i,lim})'(\xi) \leq 0 \;\forall \xi \in [0,1]$. The increased amount of degrees of freedom per cell in the multidimensional case hinders the application of this method in that case. We adapt the method by dropping the continuity at both interfaces and assemble our limiting instead on the idea of Zhang and Shu [40, 41]. It is also considered by Roe and coauthors [38].

For this, let $Q^n_{rec,i}$ resume to be the quadratic reconstruction in cell C_i. Let

$$M^n_i := \max_{\xi\in[0,1]} Q^n_{rec,i}(\xi), \tag{2.59}$$

$$m^n_i := \min_{\xi\in[0,1]} Q^n_{rec,i}(\xi). \tag{2.60}$$

Then, the limited reconstruction is defined by

$$Q^n_{rec,i,lim}(\xi) := \theta Q^n_{rec,i}(\xi) + (1-\theta)Q^n_i \tag{2.61}$$

with

$$\theta := \min\left\{\left|\frac{\bar{M}^n_i - Q^n_i}{M^n_i - Q^n_i}\right|, \left|\frac{\bar{m}^n_i - Q^n_i}{m^n_i - Q^n_i}\right|, 1\right\}. \tag{2.62}$$

The parameter θ is determined in such a way that in case of a limiting the function is compressed around the cell average value exactly so far that the new minimum or maximum is kept in the interval N_i. Otherwise, we have $\theta = 1$ and recover the original, unlimited reconstruction.

This limiting stands out in contrast to the aforementioned limitings that the use of Simpson's rule is still exact (if the proper upwind values are used) and that it can be easily extended to the multidimensional case as discussed later.

A graphical representation of both in this section explained limitings can be found in Figure 2.10. Figure 2.11 shows the comparison of the numerical solution of the advection equation under the use of these limitings. While the first limiting produces very similar results to the two limitings discussed in the previous section, the more generalizable limiting (2.61) can't resolve the discontinuities just as well.

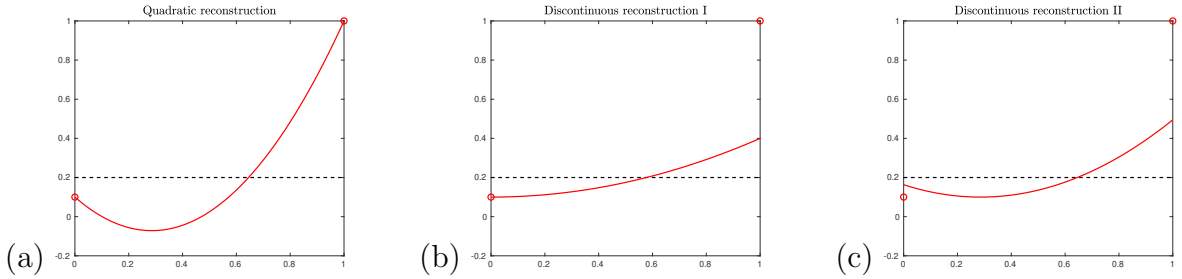

Figure 2.10.: Different reconstructions of the Active Flux method. The dashed line indicates the cell average value, the 'o' symbols at the boundaries indicate the edge values and the red solid line indicates the reconstruction. For all of these plots we used $Q^n_{i-\frac{1}{2}} = 0.1$, $Q^n_{i+\frac{1}{2}} = 1$ and $Q^n_i = 0.2$. Unlimited, quadratic reconstruction (a), discontinuous reconstruction 2.55 (b), discontinuous reconstruction 2.61 (c).

All mentioned limitings can also be used for a grid with cut cells. It is important to notice that the cancellation property is no longer forced since the continuity is lost in the reconstruction. Therefore, one has to use exact integration for the respective fluxes in cells C_{k-1} or C_k in case of a limiting.

2.6 Summary

In this chapter, we have applied the Active Flux method to one-dimensional linear conservation laws. We use point and cell average values of the conserved quantities

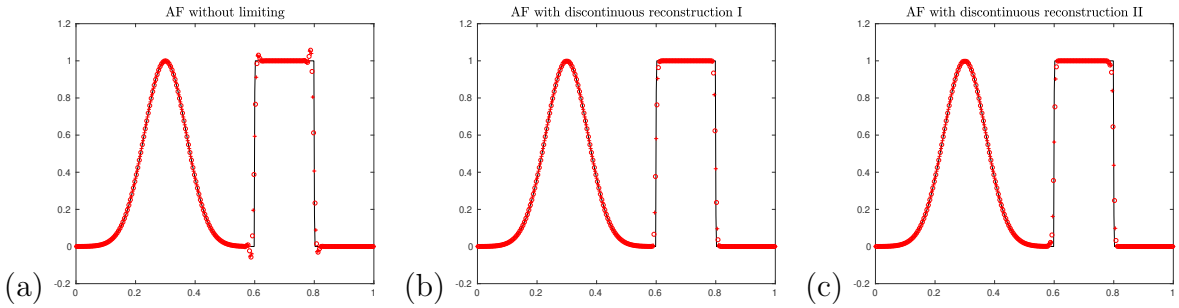

Figure 2.11.: Advection test computed with different versions of the Active Flux method using 200 grid cells. The solution is shown at time $T = 1$, i.e., after one rotation. The solution using the unlimited Active Flux method is shown in (a). In (b) and (c) we show results for the piecewise polynomial and hyperbolic reconstruction, respectively. Point values are marked with a '+', while cell averages are marked with an 'o'.

to obtain a third order accurate method. The method can be used for cut cells with some small changes, remains third order accurate and is stable for time steps chosen according to the size of the regular grid cells. Appearing oscillations near discontinuities or steep gradients in the solution are small and can be eliminated by the use of a limited reconstruction.

The special choice of degrees of freedom at the interfaces between each cell offers a good choice for a cut cell method in higher dimensions. The two-dimensional case is covered in Chapter 3. Chapter 4 deals with approaches for nonlinear systems of equations since the solution operator $\mathcal{L}_f(q(x,t), \Delta t)$ is not known in general.

3 The Active Flux Method in Two Spatial Dimensions

In the previous chapter, the Active Flux method for linear equations in one spatial dimension has been introduced. We now consider the two-dimensional case, i.e., equations (1.1) and (1.2) for $d = 2$. For a clearer distinction we write

$$\frac{\partial}{\partial t}q(x,t) + \frac{\partial}{\partial x}f(q(x,y,t)) + \frac{\partial}{\partial y}g(q(x,y,t)) = 0 \tag{3.1}$$

with $f, g : \mathbb{R}^m \to \mathbb{R}^m$.

The Active Flux method developed by Roe and coauthors has been constructed on a triangular mesh [28]. To proceed the idea of cut cells we now explain a possible adaptation of the method on Cartesian grids. This has been done by us in [1] and was independently worked out by Barsukow et al. in [42] at the same time. We restrict ourselves again to linear equations, specifically to the linear advection equation and the linear acoustic equations. At the latter we will see that the complexity of the exact solution operator for multidimensional systems strongly exceeds the complexity of solution operators in the one-dimensional case. Subsequently, we will examine the Active Flux method for the advection equation in two dimensions in the presence of cut cells.

3.1 Cartesian Grids

The notation can be extended from the one-dimensional case. We mainly follow our description in [1, Section 6]. Let Mx, My be index sets and $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})_{i\in Mx, j\in My}$ be a numerical grid, here also called corners. For $i \in Mx, j \in My$ we denote the cells by $C_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{i-\frac{1}{2}}, y_{i+\frac{1}{2}}]$. Further let $(x_i, y_j) = \left(\frac{x_{i+\frac{1}{2}}+x_{i-\frac{1}{2}}}{2}, \frac{y_{j+\frac{1}{2}}+y_{j-\frac{1}{2}}}{2}\right)$ denote the cell centers. The temporal discretization stays unchanged. In contrast to the one-dimensional case, where the boundary of each cell consists of two points that were used to define the two point values, the boundary now consists of four line segments which link the corners, here also called edges. The choice of degrees of freedom is no longer immediately apparent. We follow the choice of degrees of freedom by Eymann and Roe on triangular cells and place them on the corners and the midpoints of the edges. While this results in six degrees of freedom on the boundary for triangular grid cells

we now have eight degrees of freedom on the boundary of the Cartesian grid cell. We use

$$Q^n_{i+\frac{1}{2},j+\frac{1}{2}} \approx q^n_{i+\frac{1}{2},j+\frac{1}{2}} := q(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}, t_n) \tag{3.2}$$

for the approximate point value at a corner $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$ and

$$Q^n_{i+\frac{1}{2},j} \approx q^n_{i+\frac{1}{2},j} := q(x_{i+\frac{1}{2}}, y_j, t_n) \tag{3.3}$$

for the approximate point value at a vertical edge $(x_{i+\frac{1}{2}}, y_j)$ and

$$Q^n_{i,j+\frac{1}{2}} \approx q^n_{i,j+\frac{1}{2}} := q(x_i, y_{j+\frac{1}{2}}, t_n) \tag{3.4}$$

for the approximate point value at a horizontal edge $(x_i, y_{j+\frac{1}{2}})$, respectively. Further,

$$Q^n_{i,j} \approx \bar{q}^n_{i,j} := \frac{1}{(x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}})(y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}})} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q(x, y, t_n) \, \mathrm{d}y \, \mathrm{d}x \tag{3.5}$$

is the approximate cell average in cell $C_{i,j}$. For practical computations we now consider a rectangle $[x_l, x_r] \times [y_l, y_r]$ with double periodic boundary conditions. For this let $x_l = x_{\frac{1}{2}}, \ldots, x_{Nx+\frac{1}{2}} = x_r$ and $y_l = y_{\frac{1}{2}}, \ldots, y_{Ny+\frac{1}{2}}$ be the numerical grid and $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ $\forall i \in \{1, \ldots, Nx\}$ as well as $\Delta y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$ $\forall j \in \{1, \ldots, Ny\}$. Furthermore, for $c \in \mathbb{R}^+$ let $\Delta x = c\Delta y$, so that we can express all terms of third and higher order with $\mathcal{O}(\Delta x^3) = \mathcal{O}(\Delta x^2 \Delta y) = \mathcal{O}(\Delta x \Delta y^2) = \mathcal{O}(\Delta y^3)$. We again demand the initial values to be at least third order accurate:

$$Q^0_{i+\frac{1}{2},j+\frac{1}{2}} = q_0(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}) \tag{3.6}$$

and

$$Q^0_{i,j} = \frac{1}{(x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}})(y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}})} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q_0(x, y) \, \mathrm{d}y \, \mathrm{d}x + \mathcal{O}(\Delta x^3) \tag{3.7}$$

Figure 3.1 (left) shows the configuration of the degrees of freedom in one cell. In analogy to the one-dimensional case we have, after the use of the Divergence theorem and the fundamental theorem of calculus,

$$\begin{aligned} \bar{q}^{n+1}_{i,j} = \bar{q}^n_{i,j} &- \frac{1}{\Delta x \Delta y} \int_{t_n}^{t_{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f(q(x_{i+\frac{1}{2}}, y, t)) - f(q(x_{i-\frac{1}{2}}, y, t)) \, \mathrm{d}y \, \mathrm{d}t \\ &- \frac{1}{\Delta x \Delta y} \int_{t_n}^{t_{n+1}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} g(q(x, y_{j+\frac{1}{2}}, t)) - g(q(x, y_{j-\frac{1}{2}}, t)) \, \mathrm{d}x \, \mathrm{d}t, \end{aligned} \tag{3.8}$$

since the boundary segments of the cell (the edges) are all parallel or orthogonal to the coordinate axes. The integrals over the edges and time intervals will be approximated

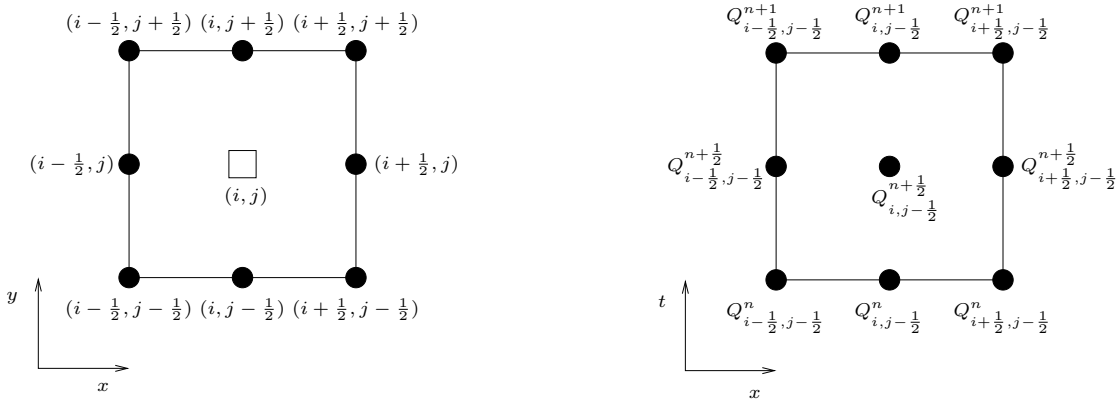

Figure 3.1.: Left: Configuration of degrees of freedom in cell $C_{i,j}$. Right: Area of integration in space-time, nodes for Simpson's rule are shown.

by a two-dimensional version of Simpson's rule:

$$\begin{aligned} F^n_{i+\frac{1}{2},j} := \frac{\Delta y}{36}\Big(& f(Q^n_{i+\frac{1}{2},j-\frac{1}{2}}) + 4f(Q^n_{i+\frac{1}{2},j}) + f(Q^n_{i+\frac{1}{2},j+\frac{1}{2}}) \\ & + 4f(Q^{n+\frac{1}{2}}_{i+\frac{1}{2},j-\frac{1}{2}}) + 16f(Q^{n+\frac{1}{2}}_{i+\frac{1}{2},j}) + 4f(Q^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2}}) \\ & + f(Q^{n+1}_{i+\frac{1}{2},j-\frac{1}{2}}) + 4f(Q^{n+1}_{i+\frac{1}{2},j}) + f(Q^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}})\Big) \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} G^n_{i,j+\frac{1}{2}} := \frac{\Delta x}{36}\Big(& g(Q^n_{i-\frac{1}{2},j+\frac{1}{2}}) + 4g(Q^n_{i,j+\frac{1}{2}}) + g(Q^n_{i+\frac{1}{2},j+\frac{1}{2}}) \\ & + 4g(Q^{n+\frac{1}{2}}_{i-\frac{1}{2},j+\frac{1}{2}}) + 16g(Q^{n+\frac{1}{2}}_{i,j+\frac{1}{2}}) + 4g(Q^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2}}) \\ & + g(Q^{n+1}_{i-\frac{1}{2},j+\frac{1}{2}}) + 4g(Q^{n+1}_{i,j+\frac{1}{2}}) + g(Q^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}})\Big). \end{aligned} \tag{3.10}$$

Therefore, we have

$$F^n_{i+\frac{1}{2},j} \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f(q(x_{i+\frac{1}{2}},y,t))\,\mathrm{d}y\,\mathrm{d}t \tag{3.11}$$

and

$$G^n_{i,j+\frac{1}{2}} \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} g(q(x,y_{i+\frac{1}{2}},t))\,\mathrm{d}x\,\mathrm{d}t. \tag{3.12}$$

This results in the finite volume method

$$Q^{n+1}_{i,j} = Q^n_{i,j} - \frac{\Delta t}{\Delta x \Delta y}\left(F^n_{i+\frac{1}{2},j} - F^n_{i-\frac{1}{2},j} - G^n_{i,j+\frac{1}{2}} + G^n_{i,j-\frac{1}{2}}\right). \tag{3.13}$$

Figure 3.1 (right) shows the quadrature points of Simpson's rule for the integration in space and time along a grid cell edge. For the update of the point values we use

the exact evolution operator. We thus restrict ourselves to equations where we can evaluate it. In Chapter 4 we consider equations where the exact evolution operator is not known.

To construct a third order accurate method we'd like to use a reconstruction which approximates q to the same order. It is sufficient to utilize the basis $\{1, x, y, x^2, xy, y^2\}$ for the reconstruction. The seven degrees of freedom of the Active Flux method on triangular grids use a basis that generates the same polynomial space and interpolates all six point values on the boundary, and additionally uses one further higher order basis function that vanishes on the boundary of the cell and simply forces the conservation of the cell average value ("bubble function"). This reconstruction yields a globally continuous reconstruction since the three degrees of freedom on each edge produce a uniquely defined parabola. Since the Active Flux method on Cartesian grids now has nine degrees of freedom it makes sense to extend the basis to $\{1, x, y, x^2, xy, y^2, x^2y, xy^2, x^2y^2\}$. The so created reconstruction interpolates all eight degrees of freedom on the boundary and conserves the cell average. It can be represented as follows. On a reference cell $[-1, 1] \times [-1, 1]$ we have

$$Q^n_{rec,i,j}(\xi, \mu) = \sum_{i=1}^{9} c_i N_i(\xi, \mu). \tag{3.14}$$

The basis functions N_i and coefficients c_i for $i \in \{1, \dots, 9\}$ are given in table 3.1. By

i	c_i	N_i
1	$Q^n_{i-\frac{1}{2},j-\frac{1}{2}}$	$\frac{1}{4}(\xi^2-\xi)(\eta^2-\eta)$
2	$Q^n_{i,j-\frac{1}{2}}$	$\frac{1}{2}(1-\xi^2)(\eta^2-\eta)$
3	$Q^n_{i+\frac{1}{2},j-\frac{1}{2}}$	$\frac{1}{4}(\xi^2+\xi)(\eta^2-\eta)$
4	$Q^n_{i+\frac{1}{2},j}$	$\frac{1}{2}(\xi^2+\xi)(1-\eta^2)$
5	$Q^n_{i+\frac{1}{2},j+\frac{1}{2}}$	$\frac{1}{4}(\xi^2+\xi)(\eta^2+\eta)$
6	$Q^n_{i,j+\frac{1}{2}}$	$\frac{1}{2}(\eta^2+\eta)(1-\xi^2)$
7	$Q^n_{i-\frac{1}{2},j+\frac{1}{2}}$	$\frac{1}{4}(\xi^2-\xi)(\eta^2+\eta)$
8	$Q^n_{i-\frac{1}{2},j}$	$\frac{1}{2}(\xi^2-\xi)(1-\eta^2)$
9	$\frac{1}{16}\Big(36Q^n_{i,j} - (Q^n_{i-\frac{1}{2},j-\frac{1}{2}} + Q^n_{i+\frac{1}{2},j-\frac{1}{2}} + Q^n_{i+\frac{1}{2},j+\frac{1}{2}} + Q^n_{i-\frac{1}{2},j+\frac{1}{2}})$ $-4(Q^n_{i,j-\frac{1}{2}} + Q^n_{i+\frac{1}{2},j} + Q^n_{i,j+\frac{1}{2}} + Q^n_{i-\frac{1}{2},j})\Big)$	$(1-\xi^2)(1-\eta^2)$

Table 3.1.: Basis functions and coefficients for the two-dimensional reconstruction (3.14).

$$Q^n_{rec}(x, y) = Q^n_{rec,i,j}(\xi, \mu) \quad \forall (x, y) \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}] \tag{3.15}$$

with

$$\xi = 2\frac{x - x_{i-\frac{1}{2}}}{\Delta x} - 1, \quad \mu = 2\frac{y - y_{j-\frac{1}{2}}}{\Delta y} - 1 \tag{3.16}$$

the reconstruction is defined on the whole grid. It is not only continuous on the point values but globally as was the case on triangular grids. We also have continuity of all partial derivatives with respect to x on the horizontal edges and continuity with respect to y on the vertical edges.

For a better illustration of the method and the used evolution operator we now consider the advection equation and the linear acoustic equations.

3.1.1 Linear Advection Equation

With $f(q) = aq, a \in \mathbb{R}$ and $g(q) = bq, b \in \mathbb{R}$, the linear advection equation in two dimensions is described. Its exact solution operator reads

$$\mathcal{L}_{f,g}(q(x,y,t),\tau) = q(x,y,t+\tau) = q(x-a\tau, y-b\tau, t). \tag{3.17}$$

We evaluate the reconstruction at the corresponding locations in upwind direction for the update of the point values:

$$\begin{aligned} Q^{n+\frac{1}{2}}_{i+\frac{1}{2},j-\frac{1}{2}} &= Q^n_{rec}\left(x_{i+\frac{1}{2}} - a\frac{\Delta t}{2}, y_{j-\frac{1}{2}} - b\frac{\Delta t}{2}\right) \\ Q^{n+1}_{i+\frac{1}{2},j-\frac{1}{2}} &= Q^n_{rec}\left(x_{i+\frac{1}{2}} - a\Delta t, y_{j-\frac{1}{2}} - b\Delta t\right) \end{aligned} \tag{3.18}$$

The time step is restricted so that the solution propagates a maximum of one grid cell per time step and define

$$\nu := \max\left(\frac{|a|\Delta t}{\Delta x}, \frac{|b|\Delta t}{\Delta y}\right) \leq 1. \tag{3.19}$$

In Section 3.1.4 we will see that the so defined method actually converges under a slightly tightened condition.

3.1.2 Linear Acoustic Equations

The linear acoustic equations in two dimensions are given by

$$\begin{aligned} \partial_t p + c\nabla \cdot \mathbf{v} &= 0 \\ \partial_t v + c\nabla p &= 0, \end{aligned} \tag{3.20}$$

where $\mathbf{v} = [u, v]^T$ is the velocity vector, p is the pressure and $c > 0$ is the speed of sound.

The here used exact evolution operator for the linear acoustic equations is based on the spherical mean [42]. The description is adapted from [2, Section 2.2]. For a scalar function $f : \mathbb{R}^2 \to \mathbb{R}$ the spherical mean over a disc with radius r, centered around

(x, y), is defined by

$$M[f](x, y, r) := \frac{1}{2\pi r} \int_0^{2\pi} \int_0^r f(x + s\cos\varphi, y + s\sin\varphi) \frac{s}{\sqrt{r^2 - s^2}} \, \mathrm{d}s \, \mathrm{d}\varphi. \tag{3.21}$$

Assuming the solution is known at time $t_0 = 0$, i.e., $\mathbf{v}_0 = (u_0, v_0)^T$ and p_0, we can express the solution at a later time t in the following way:

$$\begin{aligned} p(x, y, t) &= \partial_r \left(rM[p_0](x, y, r) \right) \Big|_{r=ct} - \frac{1}{ct} \partial_r \left(r^2 M[\mathbf{v}_0 \cdot \vec{n}](x, y, r) \right) \Big|_{r=ct} \\ \mathbf{v}(x, y, t) &= \mathbf{v}_0(x, y) - \frac{1}{ct} \partial_r \left(r^2 M[p_0 \vec{n}](x, y, r) \right) \Big|_{r=ct} \\ &\quad + \int_0^{ct} \frac{1}{r} \partial_r \left(\frac{1}{r} \partial_r \left(r^3 M[(\mathbf{v}_0 \cdot \vec{n})\vec{n}](x, y, r) \right) - rM[\mathbf{v}_0](x, y, r) \right) \, \mathrm{d}r \end{aligned} \tag{3.22}$$

The vector valued function $M[\mathbf{v}_0]$ is computed component-by-component, i.e., $M[\mathbf{v}_0] = (M[u_0], M[v_0])^T$. The scalar function $M[\mathbf{v}_0 \cdot \vec{n}]$ for $\vec{n} = (\cos\varphi, \sin\varphi)^T$ is determined by replacing $f(x + s\cos\varphi, y + \sin\varphi)$ by $\cos\varphi \, u_0(x + s\cos\varphi, y + s\sin\varphi) + \sin\varphi \, v_0(x + s\cos\varphi, y + s\sin\varphi)$ in (3.21). The two entries in $M[p_0\vec{n}]$ are determined by replacing f by $\cos\varphi \, p_0(x + s\cos\varphi, y + s\sin\varphi)$ and $\sin\varphi \, p_0(x + s\cos\varphi, y + s\sin\varphi)$. Analogously, the two entries of $M[(\mathbf{v}_0 \cdot \vec{n})\vec{n}]$ are determined.

This evolution formula can be evaluated exactly if p_0 and $\mathbf{v}_0$ are replaced by the corresponding components of the continuous, piecewise quadratic reconstruction Q^n_{rec} in each time step. The integration has to be performed separately on each cell that is part of the circle with radius $c\Delta t$.

The time step is furthermore bounded such that circles with radius $c\Delta t$ around a degree of freedom on the edge don't cross the two adjacent cells. Figure 3.2 illustrates this. This constraint is expressed by

$$\nu := \max\left(\frac{c\Delta t}{\Delta x}, \frac{c\Delta t}{\Delta y} \right) \leq \frac{1}{2}. \tag{3.23}$$

Thereby, for each edge we have two areas of integration and for each corner we have four areas of integration. Inequality (3.23) is a necessary condition for stability and

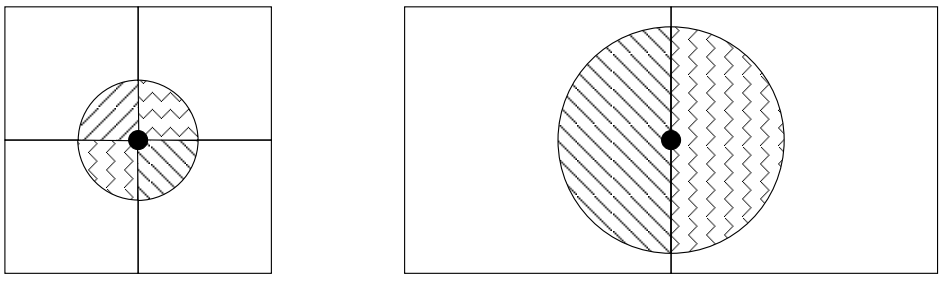

Figure 3.2.: Illustration of the computation of point values of the conserved quantities using the exact evolution formula for the acoustic equations. Left: Corner value. Right: Edge value.

converge. We will examine linear stability in Section 3.1.4.

3.1.3 Accuracy

In Section 2.3 we have presented the local truncation error for the one-dimensional advection equation. For the two-dimensional variant an analogue procedure is possible but proves to be extremely cumbersome due to the increasingly long and more extensive formulas. We therefore pass on it and instead perform an investigation similar to (2.24) and (2.25). This part is adapted from [2, Section 3.1].

The flux for the advection equation on a vertical edge has the following form:

$$
\begin{aligned}
\bar{f}^{n+\frac{1}{2}}_{i+\frac{1}{2},j} &= \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} aq(x_{i+\frac{1}{2}},y,t)\ \mathrm{d}y\ \mathrm{d}t \\
&= \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-a\Delta t}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}-\frac{b}{a}(x_{i+\frac{1}{2}}-x)}^{y_{j+\frac{1}{2}}-\frac{b}{a}(x_{i+\frac{1}{2}}-x)} q(x,y,t_n)\ \mathrm{d}y\ \mathrm{d}x \\
&\approx \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-a\Delta t}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}-\frac{b}{a}(x_{i+\frac{1}{2}}-x)}^{y_{j+\frac{1}{2}}-\frac{b}{a}(x_{i+\frac{1}{2}}-x)} Q^n_{rec}(x,y)\ \mathrm{d}y\ \mathrm{d}x \\
&\approx \frac{a\Delta y}{36}\Big(Q^n_{rec}(x_{i+\frac{1}{2}},y_{j-\frac{1}{2}}) + 4Q^n_{rec}(x_{i+\frac{1}{2}},y_j) + Q^n_{rec}(x_{i+\frac{1}{2}},y_{j+\frac{1}{2}}) \\
&+ 4Q^n_{rec}\left(x_{i+\frac{1}{2}}-\frac{a\Delta t}{2},y_{j-\frac{1}{2}}-\frac{b\Delta t}{2}\right) + 16Q^n_{rec}\left(x_{i+\frac{1}{2}}-\frac{a\Delta t}{2},y_j-\frac{b\Delta t}{2}\right) \\
&+ 4Q^n_{rec}\left(x_{i+\frac{1}{2}}-\frac{a\Delta t}{2},y_{j+\frac{1}{2}}-\frac{b\Delta t}{2}\right) + Q^n_{rec}(x_{i+\frac{1}{2}}-a\Delta t,y_{j-\frac{1}{2}}-b\Delta t) \\
&+ 4Q^n_{rec}(x_{i+\frac{1}{2}}-a\Delta t,y_j-b\Delta t) + Q^n_{rec}(x_{i+\frac{1}{2}}-a\Delta t,y_{j+\frac{1}{2}}-b\Delta t)\Big) \\
&=: F^n_{i+\frac{1}{2},j}.
\end{aligned}
\tag{3.24}
$$

Here, we have introduced two approximations. First, the exact solution $q(x,y,t_n)$ is replaced by the reconstruction $Q^n_{rec}(x,y)$. In contrast to the one-dimensional case a second approximation is introduced by applying Simpson's rule: While we integrated a quadratic function under compliance with the stability condition in the one-dimensional case and therefore recovered the true integral by Simpson's rule, the reconstruction now is only locally quadratic on each part of the two-dimensional area of integration. That makes Simpson's rule no longer exact. In this case we can analytically compute the integral by an exact integration. This can be achieved by separating the area into the different cells. For unification purposes, a triangularization of the area is possible for example. Figure 3.3 shows an exemplary area of integration as well as a depiction of Simpson's rule (left) and exact integration (right) for $a, b > 0$. Whereas no exact form of the truncation error was state we still find third order accuracy in the L_1-norm as well as in the L_∞-norm for both methods of integration. This let's us conjecture that the second order terms cancel like in the one-dimensional case. Exemplary computations are shown in Section 3.1.5 and Section 4.2.2.2.

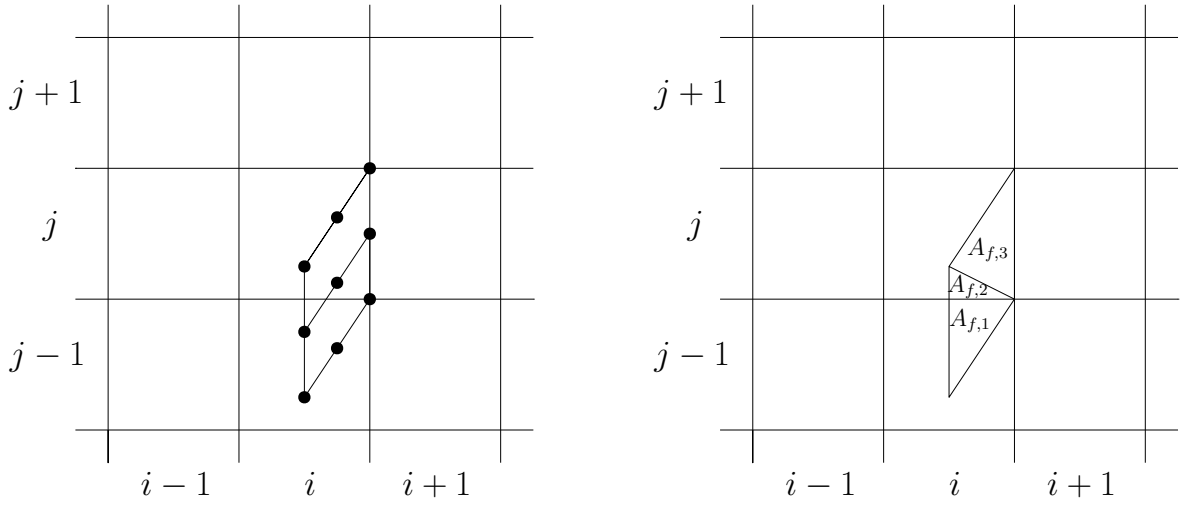

Figure 3.3.: The left plot illustrates the flux computation for the advection equation at a vertical grid cell interface using Simpson's rule. The right plot illustrates the flux computation using exact integration.

For the linear acoustic equations we can't perform an exact flux integration that easily. We thus restrict ourselves to the computation by Simpson's rule. A computation can be found in Section 4.2.2.2.

3.1.4 Linear Stability

As in the one-dimensional case we are concerned with the linear stability of the Active Flux method. Since the method is also linear in two dimensions, we pursue the same ansatz. This whole section is taken from [2, Section 3] and has been modified slightly. To simplify notation we consider a quadratic area that is discretized by a grid of size $N \times N$. In particular we have $\Delta x = \Delta y$. We again apply periodic boundary conditions. For a unique assignment we allocate to each cell $C_{i,j}$ its cell average $Q^n_{i,j}$, the left and bottom edge values $Q^n_{i-\frac{1}{2},j}$, $Q^n_{i,j-\frac{1}{2}}$ and the left bottom corner value $Q^n_{i-\frac{1}{2},j-\frac{1}{2}}$. There are of course other possibilities. Figure 3.4 shows the allocated degrees of freedom in each cell in red. With this, each cell is responsible for $4m$ degrees of freedom. We again write the method in matrix-vector form

$$\mathbf{Q}^{n+1} = A\mathbf{Q}^n, \tag{3.25}$$

where $\mathbf{Q}^n \in \mathbb{R}^{4mN^2}$ consists of all degrees of freedom at time t_n and the matrix $A \in \mathbb{R}^{(4mN^2)\times(4mN^2)}$ describes the update of the method in one time step. The number of unknowns is $m = 1$ for the advection equation and $m = 3$ for the linear acoustic equations in two dimensions. That's why we have four degrees of freedom per cell for the advection equation and twelve for the linear acoustic equations.

To explore the stability of the different versions of the Active Flux method experimentally, we plot the eigenvalues of the matrices A for a fixed grid. Since the construction of the matrix by hand, like in the one-dimensional case, is very intricate,

we describe a procedural approach that is implemented with the help of the sympy package of Python. The vector $\mathbf{Q}$ will be built up as follows:

- For each cell, arrange the degrees of freedom in the order

$$Q^n_{i-\frac{1}{2},j}, Q^n_{i-\frac{1}{2},j-\frac{1}{2}}, Q^n_{i,j-\frac{1}{2}}, Q^n_{i,j}. \tag{3.26}$$

  Each of these four vectors will contain all unknown variables at the corresponding positions. For example, the unknowns for the linear acoustic equations will be ordered as

$$\begin{aligned}\mathbf{Q}_{i,j} := [&p^n_{i-\frac{1}{2},j}, v^{1,n}_{i-\frac{1}{2},j}, v^{2,n}_{i-\frac{1}{2},j}, p^n_{i-\frac{1}{2},j-\frac{1}{2}}, v^{1,n}_{i-\frac{1}{2},j-\frac{1}{2}}, v^{2,n}_{i-\frac{1}{2},j-\frac{1}{2}},\\ &p^n_{i,j-\frac{1}{2}}, v^{1,n}_{i,j-\frac{1}{2}}, v^{2,n}_{i,j-\frac{1}{2}}, p^n_{i,j}, v^{1,n}_{i,j}, v^{2,n}_{i,j}].\end{aligned} \tag{3.27}$$

- Then, all cells will be concatenated row by row, i.e.,

$$\mathbf{Q}^n = [Q_{1,1}, Q_{2,1}, \ldots, Q_{N,1}, Q_{1,2}, Q_{2,2}, \ldots, Q_{N-1,N}, Q_{N,N}]^T. \tag{3.28}$$

For the construction of the matrix A we consider for a given cell $C_{i,j}$ all cells that have at least one degree of freedom that contributes to the update of any of the degrees of freedoms of cell $C_{i,j}$. Which of the cells actually do have a non-vanishing contribution depends strongly on the equations and its parameters, for example advection speeds (> 0 or < 0), and on the evolution operator that is used for the interface values. Generally, because the method has a local stencil, it is sufficient to investigate a 4×4 grid around cell $C_{i,j}$, i.e., cells $C_{i-1,j-1}$ to $C_{i+2,j+2}$. The red marked degrees of freedom in Figure 3.5 indicate a potential non-vanishing contribution to the update of cell $C_{i,j}$. After performing the update for both point values and cell average values the coefficients of all $16 \cdot 4 \cdot m$ degrees of freedom can be easily extracted. These coefficients will now solely depend on Δt and Δx. They are then saved in 16 matrices $Z_{k,l} \in \mathbb{R}^{4m \times 4m}$, where $k, l \in \{-1, 0, 1, 2\}$ represent the relative position to cell $C_{i,j}$. The entry $Z_{k,l}[x, y]$, $x, y \in \{1, \ldots, 4m\}$ will therefore represent the contribution of the degree of freedom y in cell $C_{i+k,j+l}$ to the update of the degree of freedom x in cell $C_{i,j}$. x and y are of course to be understood according to the order described in (3.26).

After obtaining the matrices $Z_{k,l}$, one can substitute the symbolic values by numerical values in order to test for stability.

Finally, A can be constructed by placing these 16 matrices in the correct spots through Kronecker products. Let $P_m(\delta)$ be the identity matrix $I_m \in \mathbb{R}^{m \times m}$ shifted by δ columns to the right (or equivalently by δ rows to the top), e.g.

$$P_4(1) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad P_3(-1) = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \tag{3.29}$$

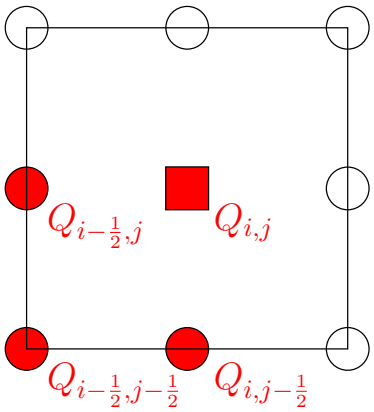

Figure 3.4.: DoF saved in cell $C_{i,j}$ marked in red.

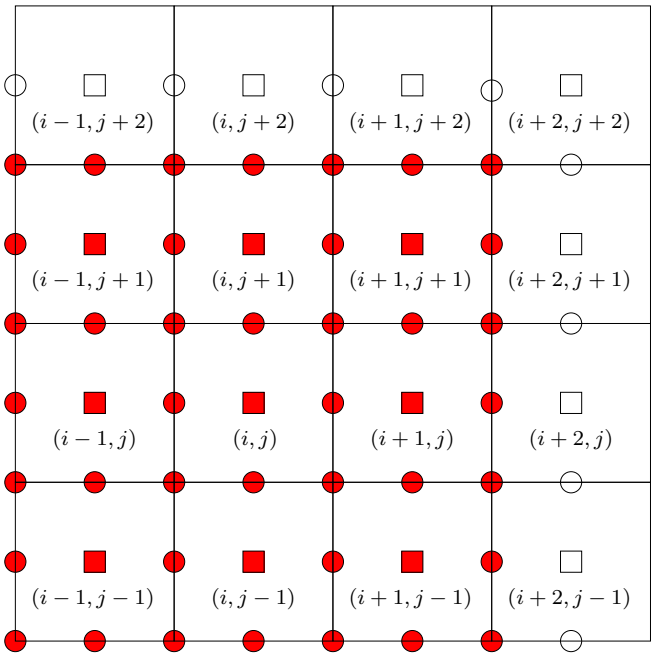

Figure 3.5.: Domain of influence for the update of all DoF of cell $C_{i,j}$.

Then, A is given by

$$A = \sum_{k,l=-1}^{2} P_m(k) \otimes P_m(l) \otimes Z_{k,l}. \tag{3.30}$$

Figure 3.6 displays the structure of the sparse matrix A for both linear problems and $N = 10$.

More details can be found in Appendix A. Figures 3.7 - 3.9 shown in the next section and the similar figures in Appendix A were done with the help of an earlier version of the stability programs by Erik Chudzik.

3.1.4.1 Linear Advection Equation

After we have built up the matrix we can inspect the eigenvalues as intended. Figure 3.7 displays the eigenvalues for $a = b$, $\nu = 0.75$ as well as $a = b$, $\nu = 0.9$ for a grid with $N = 20$. While both Simpson's rule and exact integration result in a potentially stable method for $\nu = 0.75$ (left plots), some eigenvalue lie outside the unit circle when using Simpson's rule in the case $\nu = 0.9$ (right top plot) and therefore don't result in a stable method. This is not the case when using exact integration (right bottom plot). This different behavior raises the question how the stability region of the Active Flux method with Simpson's rule in two dimensions looks like. To have a better answer to this question we again choose a fixed grid with $N = 20$ as in the previous test and vary $|a|\Delta t/\Delta x$ and $|b|\Delta t/\Delta x$. For every combination we test if an eigenvalue with modulus greater than one exists. A dot in Figure 3.8 identifies a situation with $|\lambda| \leq 1$ for all eigenvalues λ. As expected, we see that the use of Simpson's rule leads to a reduced

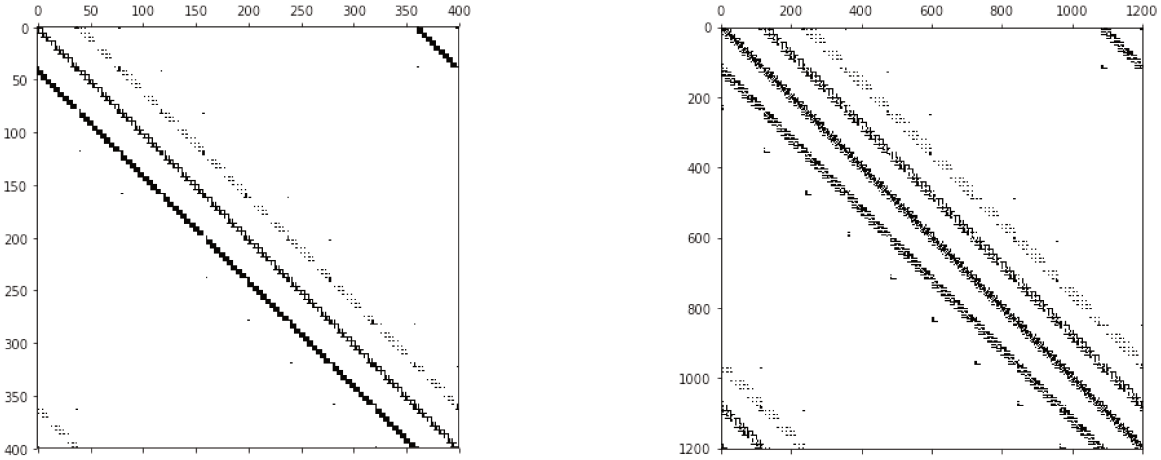

Figure 3.6.: Structure of the matrix representing the Active Flux method for advection using Simpson's rule (left) and acoustics (right) in two dimensions for $N = 10$ cells in both directions.

stability region (left plot). For a visual comparison we draw the bound

$$\frac{\sqrt{a^2 + b^2}\Delta t}{\Delta x} \leq 1. \tag{3.31}$$

as a red line. This line roughly describes the stability region. However, there exist both dots inside the region that characterize an unstable method as well as dots outside the region that indicate a potentially stable method. On the contrary, the Active Flux method using exact integration only shows potentially stable situations for all time steps that suffice (3.19). This is presented in the right plot of Figure 3.8.

In Appendix A we show results for the same test but with $N = 50$. The patterns formed by the location of the eigenvalues compare well. This suggests that the prediction of the asymptotic stability behavior from Figure 3.8, which was computed using $N = 20$, is correct for more general grid resolutions. The test for the maximal eigenvalue leads to a consistent picture to Figure 3.8, too. In Figure 3.9 we plot $\|A^n\|$ against n for the Active Flux method using Simpson's rule for different CFL numbers ν and $a = b$. The upper left plot in Figure 3.9 belongs to the asymptotically stable case $a\frac{\Delta t}{\Delta x} = b\frac{\Delta t}{\Delta y} = 0.75$. The top right and bottom left plot of Figure 3.9 illustrate the unstable situations $a\frac{\Delta t}{\Delta x} = b\frac{\Delta t}{\Delta y} = 0.8$ and $a\frac{\Delta t}{\Delta x} = b\frac{\Delta t}{\Delta y} = 0.9$.

The case $a\frac{\Delta t}{\Delta x} = b\frac{\Delta t}{\Delta y} = 1$ deserves a special consideration. The single dot in the upper right corner of Figure 3.8 (left) indicates that for Simpson's rule the magnitude of all eigenvalues is bounded by one. In the bottom right plot of Figure 3.9 we plot $\|A^n\|$ against n for this case and observe that $\|A^n\|$ grows asymptotically linearly with n. Thus the method is unstable in this case. The linear growth of $\|A^n\|$ indicates the existence of a Jordan block of size at least two for an eigenvalue of magnitude one. In this case the algebraic and geometric multiplicity of the eigenvalues of A do not match. For $N = 2$, i.e., $A \in \mathbb{R}^{16\times 16}$, this can also be verified analytically with the help of our Python code, see Appendix A for more details.

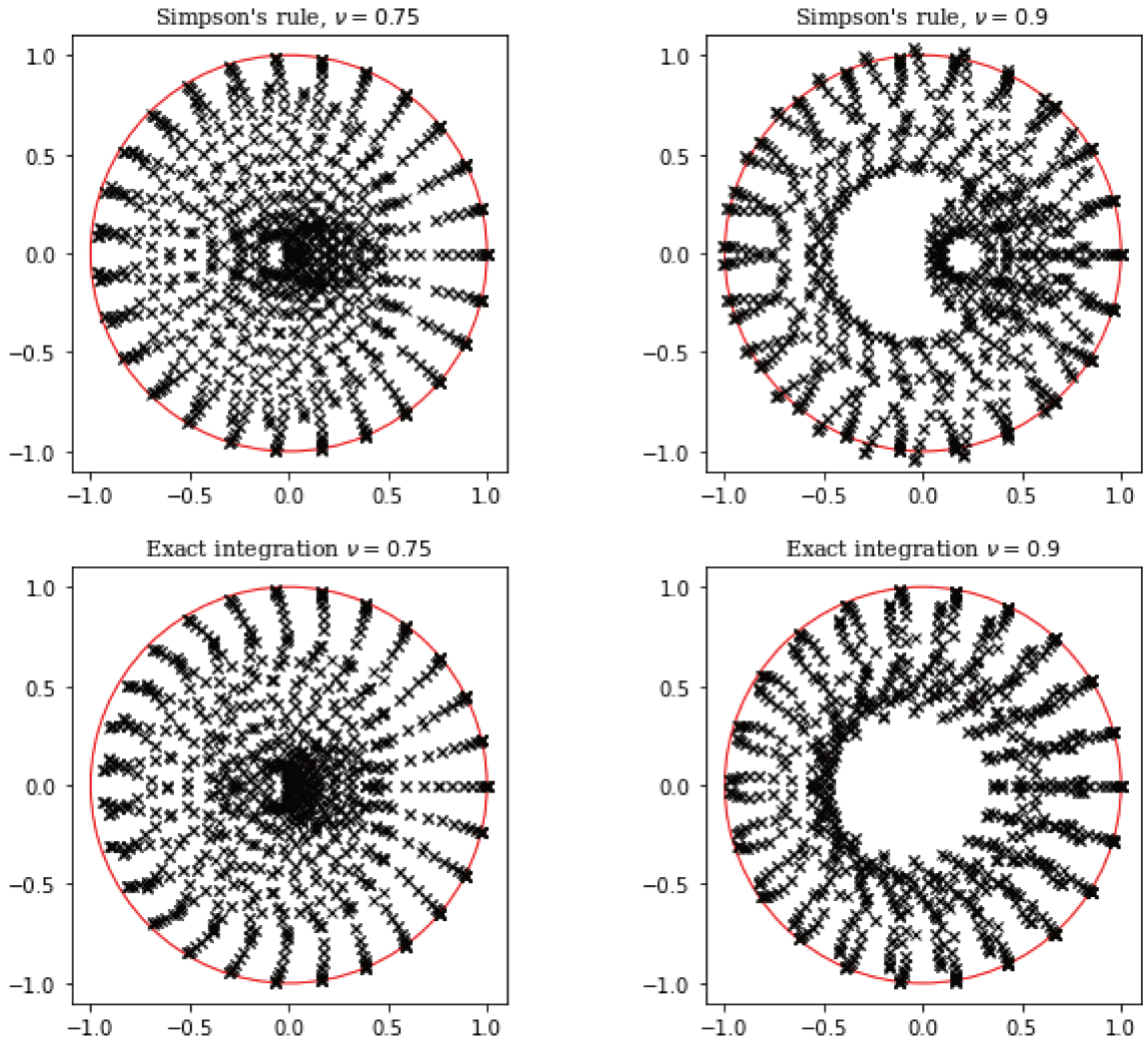

Figure 3.7.: Eigenvalues for $a = b$, $\Delta x = \Delta y = 1/20$, $N = 20$, $\nu = 0.75$ and $\nu = 0.9$ using Simpson's rule (first row) and exact integration (second row) for the computation of the cell averaged values.

If we instead use exact integration, then the cell average values and point values are simply shifted in diagonal direction. For $a = b = 1$, $\Delta x = \Delta y = \Delta t$, the matrix $Z_{-1,-1}$ is a 4×4 identity matrix and all other $Z_{k,l}$ matrices vanish. Therefore, A is a permutation matrix and it holds $\|A^n\| = 1$ for all values n. Thus, exact integration remains stable.

3.1.4.2 Linear Acoustic Equations

A similar analysis can be done for the linear acoustic equations. Our tests were done for various CFL numbers and grids, always using $\Delta x = \Delta y$. Here, we show the plots for $\nu = 0.5$ and $\nu = 0.4$ on a grid with $N = 30$ grid cells (cf. Figure 3.10). With these tests, we can numerically confirm the presumed stability 3.23, i.e., $\nu \leq 0.5$. For $\nu > 0.5$

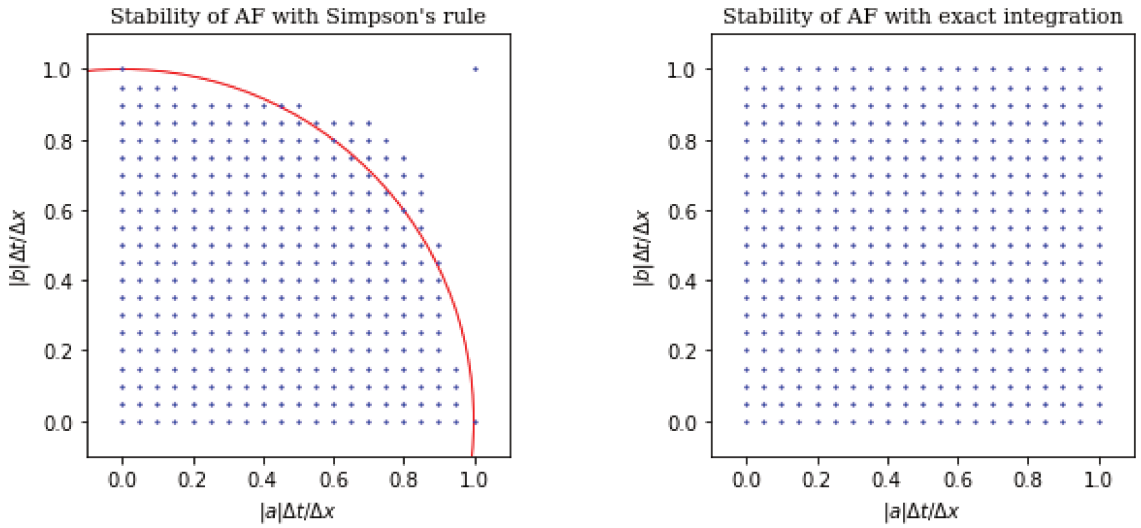

Figure 3.8.: Active Flux method eigenvalue test for different advection speeds and CFL numbers with $0 \leq \nu \leq 1$ and $N = 20$. The dots indicate potentially stable methods. Exact integration (right plot) is stable for $\nu \leq 1$, while the use of Simpson's rule (left plot) leads to a reduced stability. For a visual comparison, a quarter of the unit circle is plotted in red.

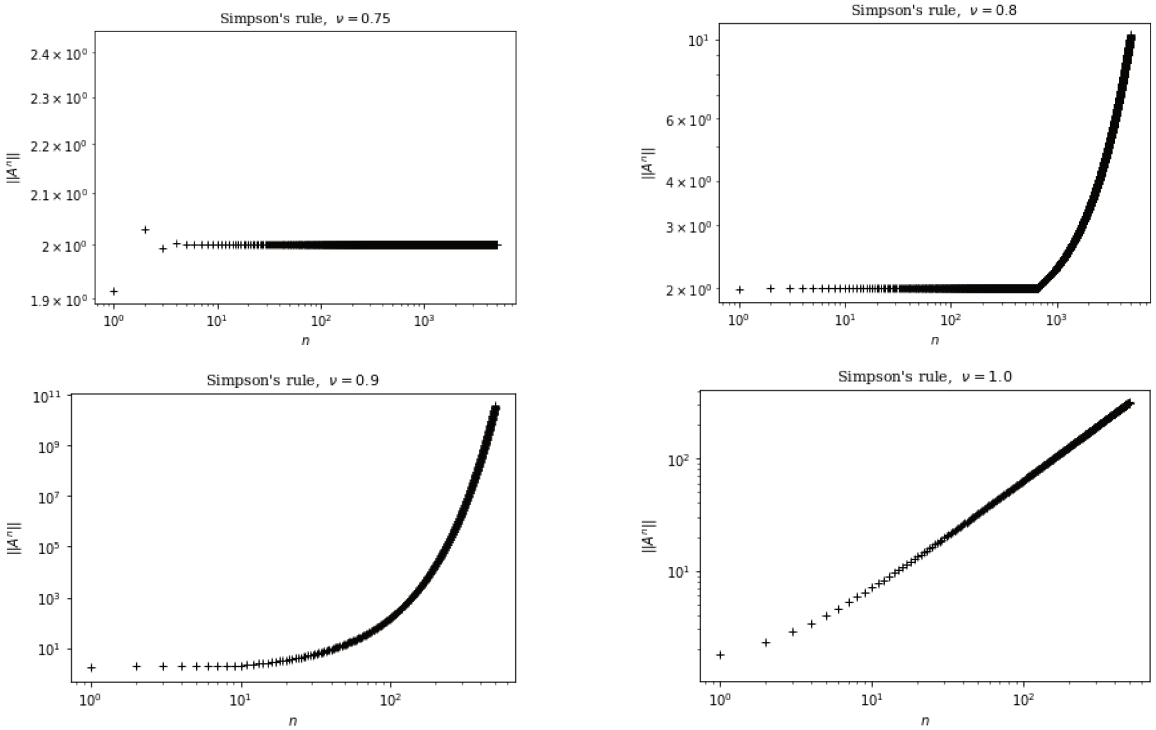

Figure 3.9.: $\|A^n\|$ against n for the Active Flux method with Simpson's rule. Top left: $\nu = 0.75$. Top right: $\nu = 0.8$. Bottom left: $\nu = 0.9$. Bottom right: $\nu = 1$.

the method is not stable as explained in Section 3.1.2. A further increase of the time step would require to integrate over larger circles. This would make the method more difficult to implement and has not been done. For the current version of the method, there is no need to consider exact integration of the flux, since the use of Simpson's rule already provides the maximal stability for all appropriate time steps.

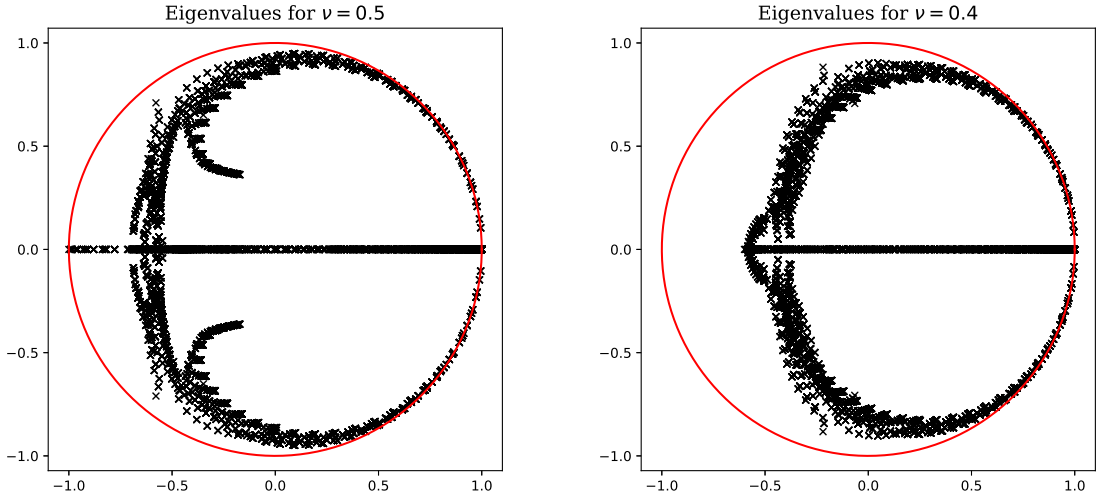

Figure 3.10.: Eigenvalues of the update matrix A in comparison to the unit circle for $N = 30$ grid cells. Time steps correspond to $\nu = 0.5$ (left) and $\nu = 0.4$ (right).

3.1.5 Limiting

We convert the limiting introduced in Section 2.5.2 to the two-dimensional case. In this whole section, we follow [2, Section 4]. Let

$$M_{i,j} := \max_{(\xi,\eta)\in[-1,1]^2} Q^n_{rec,i,j}(\xi,\eta), \tag{3.32}$$

$$m_{i,j} := \min_{(\xi,\eta)\in[-1,1]^2} Q^n_{rec,i,j}(\xi,\eta) \tag{3.33}$$

be the maximum and minimum of the reconstruction in cell $C_{i,j}$ and

$$\bar{M}_{i,j} := \max\left\{Q^n_{i-\frac{1}{2},j-\frac{1}{2}}, Q^n_{i,j-\frac{1}{2}}, Q^n_{i+\frac{1}{2},j-\frac{1}{2}}, Q^n_{i+\frac{1}{2},j}, Q^n_{i+\frac{1}{2},j+\frac{1}{2}}, Q^n_{i,j+\frac{1}{2}}, Q^n_{i-\frac{1}{2},j+\frac{1}{2}}, Q^n_{i-\frac{1}{2}j}\right\} \tag{3.34}$$

$$\bar{m}_{i,j} := \min\left\{Q^n_{i-\frac{1}{2},j-\frac{1}{2}}, Q^n_{i,j-\frac{1}{2}}, Q^n_{i+\frac{1}{2},j-\frac{1}{2}}, Q^n_{i+\frac{1}{2},j}, Q^n_{i+\frac{1}{2},j+\frac{1}{2}}, Q^n_{i,j+\frac{1}{2}}, Q^n_{i-\frac{1}{2},j+\frac{1}{2}}, Q^n_{i-\frac{1}{2}j}\right\} \tag{3.35}$$

$$N_{i,j} := [\bar{m}_{i,j}, \bar{M}_{i,j}] \tag{3.36}$$

be the maximum and minimum of the degrees of freedom on the boundary of cell $C_{i,j}$ as well as the interval between the two values. As in the one-dimensional case we pursue the following limiting strategy: If $Q^n_{i,j} \in N_{i,j}$, then we limit the reconstruction through

$$Q^n_{rec,i,j,lim}(\xi,\eta) := \theta \left(Q^n_{rec,i,j}(\xi,\eta) - Q^n_{i,j} \right) + Q^n_{i,j} \tag{3.37}$$

with

$$\theta := \min \left\{ \left| \frac{\bar{M}_{ij} - Q^n_{i,j}}{M_{i,j} - Q^n_{i,j}} \right|, \left| \frac{\bar{m}_{ij} - Q^n_{i,j}}{m_{i,j} - Q^n_{i,j}} \right|, 1 \right\}. \tag{3.38}$$

If $Q^n_{i,j} \notin N_{i,j}$, we do not apply any limiting in order to keep the accuracy of the solution near local extrema. If $Q^n_{i,j} \in N_{i,j}$, the limited reconstruction obeys

$$\bar{m}_{i,j} \leq Q^n_{rec,i,j,lim}(\xi,\eta) \leq \bar{M}_{i,j} \quad \forall (\xi,\eta) \in [-1,1]^2.$$

Again, if Simpson's rule is used to compute the fluxes, the respective values in upwind direction have to be used in order to reach a correct approximation. We now consider the accuracy of the method while using this limiting for the advection equation. Nonlinear equations can be limited in the exact same manner. In Section 4.1.2 limiting for Burgers' equation is looked at.

3.1.5.1 Accuracy Study on the Advection Equation for Smooth Initial Conditions

For a smooth advection problem we now compare the accuracy of the Active Flux method using either exact integration or Simpson's rule in order to compute the fluxes. We consider the advection equation with $a = b = 0.7$ and

$$q(x,y,0) = \sin(\pi x)\cos(\pi y) \tag{3.39}$$

for $(x,y) \in [0,2] \times [0,2]$ and periodic boundary conditions. We use $\Delta x = \Delta y = \Delta t$, i.e., $\nu = 0.7$ and compute solutions at time $T = 0.8$ using different grid resolutions.

In the left plot of Figure 3.11 we display the error in the L_1-norm against the mesh size $\Delta x = \Delta y$ while using the Active Flux method without limiting. While both versions of the Active Flux method are third order accurate, the Active Flux method with exact integration is slightly more accurate than the Active Flux method which uses Simpson's rule. Furthermore, we perform an accuracy study for the same test problem by replacing the Active Flux reconstruction with the bound preserving reconstruction. The use of the limiter reduces the accuracy of the method, but we still observe third order convergence in the L_1-norm. Again, we observe a sightly higher accuracy by using exact integration instead of Simpson's rule. In the right plot, we show the error in the L_∞-norm. In this norm, the error obtained by using exact integration is almost identical with the error that is obtained by using Simpson's rule. In the limited case, we now observe a loss in convergence rate.

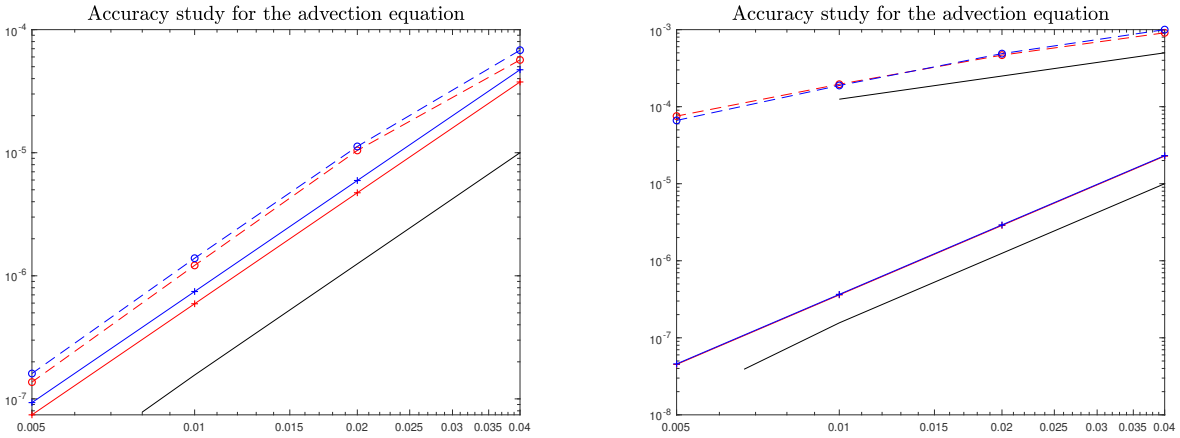

Figure 3.11.: Accuracy study for the Active Flux method using exact integration as well as Simpson's rule. The curves of the left plot show the error in the L_1-norm, the curves of the right plot show the error in the L_∞-norm. The red curves show results for exact integration in the unlimited case ('+' symbols) as well as limited case ('o' symbols). The blue curves show results for Simpson's method in the unlimited case ('+' symbols) and limited case ('o' symbols). The black line in the left plot is a reference curve for third order accuracy. In the right, the two black lines are reference lines for third as well as first order convergence.

3.1.5.2 Accuracy Study on the Advection Equation for Discontinuous Initial Conditions

We now consider the same initial value problem with initial values

$$q(x,y,0) = \begin{cases} 1 & : \quad (x,y) \in \left[\frac{1}{3},\frac{2}{3}\right] \times \left[\frac{1}{3},\frac{2}{3}\right] \\ \exp\left(-20((x-1.25)^2+(y-1.25)^2)\right) & : \quad \text{otherwise} \end{cases} \tag{3.40}$$

The other parameters are chosen as in the previous test. Figure 3.12 shows numerical results after one rotation obtained on a 100×100 grid. The numerical flux is again computed using either Simpson's rule or exact integration. In the left plots, we show results for the unlimited case and in the right plot we show results for the limited reconstruction. As expected, both flux computations lead to overshoots and undershoots near the discontinuity if no limiting is used. On the contrary, these are removed by the limiting as long as exact integration is used. Although the reconstruction doesn't produce new extrema, the flux computation using Simpson's rule is not exact for this piecewise quadratic function. Therefore, the numerical solution might produce spurious oscillations. On the 100×100 grid, the maximal cell average value observed after one full rotation is one, i.e., there are no overshoots. The minimal cell average value is about -10^{-3}, i.e., there are small undershoots. The exact flux computation in combination with the limited bound preserving reconstruction leads to accurate approximations and eliminates undershoots and overshoots up to machine precision. We summarise our results in the following theorem:

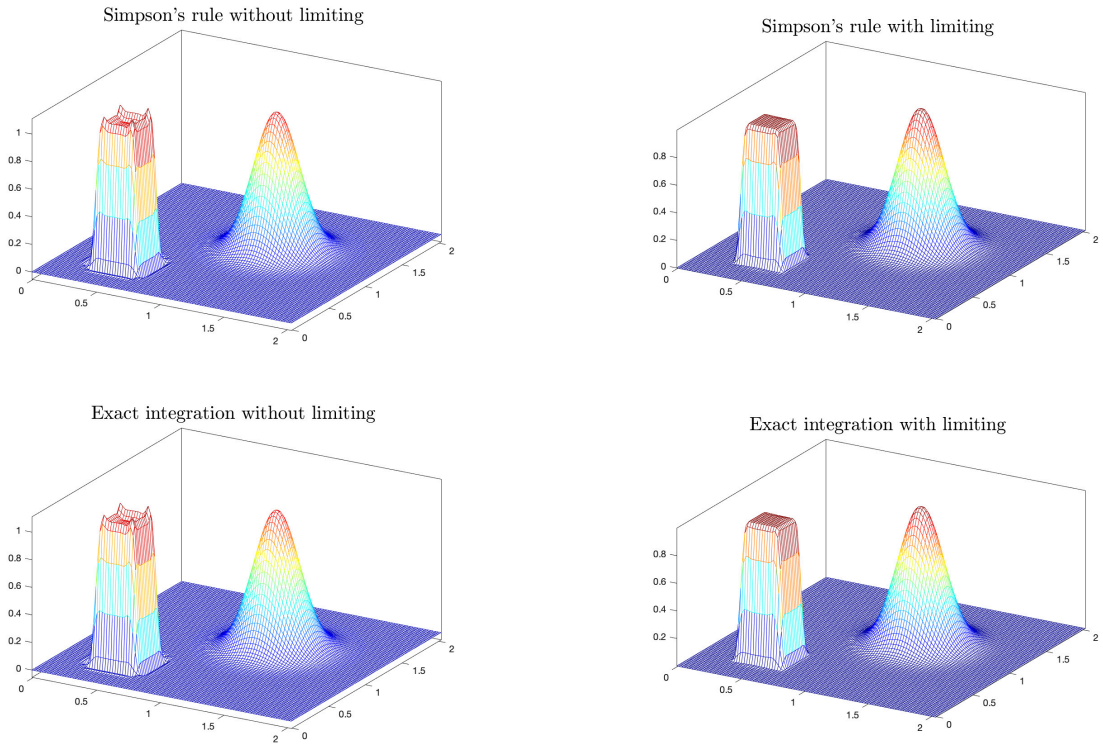

Figure 3.12.: Numerical results for the advection equation using Simpson's rule with unlimited reconstruction (top, left) and limited reconstruction (top, right), as well as exact integration with unlimited reconstruction (bottom, left) and limited reconstruction (bottom, right).

Theorem 3.1.1. *Let* $\bar{m} := \min_{i,j} \bar{m}_{i,j}$ *and* $\bar{M} := \max_{i,j} \bar{M}_{i,j}$ *describe the global minima and maxima of all point values of* q *at time* t_n*. If*

$$Q^n_{i,j} \in N_{i,j} \quad \forall (i,j), \tag{3.41}$$

then then Active Flux method for the two-dimensional advection equation with limited reconstruction (3.37) *and exact integration for the flux computation produces new cell average values and point values which satisfy*

$$Q^{n+1}_{k,l} \in [\bar{m}, \bar{M}] \quad \forall (k,l) \in \left\{ \left(i - \frac{1}{2}, j\right), \left(i - \frac{1}{2}, j - \frac{1}{2}\right), \left(i, j - \frac{1}{2}\right), (i,j) \right\}.$$

Proof. The property for the point values follows directly from the bound preserving limited reconstruction and the update of the point values described in (3.18).

We can understand the method in the framework reconstruction-evolution-averaging. The bound preserving reconstruction doesn't introduce new extrema. The evolution using exact quadrature is exact for the reconstructed data and therefore also bound preserving. As always, averaging doesn't introduce new extrema. □

Note that condition (3.41) is necessary, since our limiting strategy avoids limiting near local extrema. During the evolution, new local minima or maxima might arise in grid cells where this situation has not appeared previously. If we would combine the bound preserving reconstruction with a piecewise constant reconstruction near local minima or maxima, it would be straight forward to prove a TVD result for the Active Flux method with exact integration. However, the accuracy of such a method would be reduced.

3.2 Cut Cell Grids

While cut cells in one spatial dimension are only smaller versions of the regular cells, cut cells can in general not be obtained by a linear transformation in the multidimensional case. Through arbitrary cuts arbitrarily formed cut cells arise in the two-dimensional cells $C_{i,j}$. While some ideas and first results have already been presented in [3, Section 3] and are used in this whole section, we provide a more detailed description and discussion. To lay the foundations for the use of the Active Flux method on multidimensional cut cell grids, we make three assumptions to our cut cell grid:

- Each cell is only crossed by a maximum of one connected boundary path. This condition is normally ensured by using a fine grid in the proximity of narrow passages of the domain and can be enforced by a rougher boundary estimation in these fine grid cells. Since grids should not be fine everywhere, the use of mesh refinement is of high interest. This is a current research topic of our work group.
- The boundary of a cut cell is given by the boundary of the corresponding part of the domain. For general boundaries it is important to discuss how the boundary should be approximated. In this work we approximate the boundary by a straight line segment connecting the two points of intersection of the boundary with the cell. The considered problem has a straight boundary, so no additional approximation is introduced.
- We strictly use $\Delta x = \Delta y$. This is not obligatory but reduces the complexity of the arising cut cells and integration areas.

Up to rotations and reflections four different cell types emerge: Uncut rectangles (squares), quadrilaterals with one inclined edge, right triangles and pentagons which arise by cutting off one corner of a rectangle. The degrees of freedom of the unknowns are placed on the corners and midpoints of the edges so that Simpson's rule can be used for the flux computation along every edge. Figure 3.13 shows all possible cells.

3.2.1 Reconstruction

The rectangles and triangular cells were already described here and by Roe et al., respectively [28]. The used reconstructions are considered in a reference cell which can be obtained by an affine map from the original cell and vice versa. Because of that

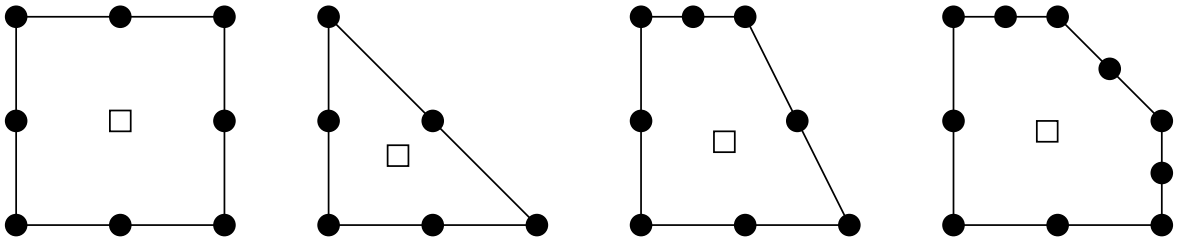

Figure 3.13.: Degrees of freedom in the two-dimensional Active Flux method for all possible cut cells. The solid dots indicate point values of the conserved quantity while the squares indicate the cell average.

the reconstruction is easily possible even in very plain triangles or rectangles. No stiff system of equations has to be solved.

We now consider the situation in a pentagon: Since there are eleven degrees of freedom the nine basis functions for the bi-parabolic reconstruction are no longer sufficient to ensure an interpolation of all ten point values. For a long time we added the basis functions x^3 and y^3 to satisfy all conditions. Through a clever choice of the basis functions it was previously possible to evaluate the reconstruction in the interior of the cell with pleasant coefficients. This property gets lost with the new approach for the pentagons. In every step a system of equations has to be solved or the previously computed analytic solution can be used which is very long and expensive. Next to the high costs two problems arise:

- First of all the reconstruction is no longer continuous across the whole edge but only in the interpolation points. This can situationally lead to a loss of the cancellation property which in turn leads to an unstable method. In practical computations this didn't seem to be a problem, although we could not prove any statement for the cancellation property when the reconstruction is not continuous.

- The condition of these interpolation problems depends strongly on the distance of the degrees of freedom. Since there is no affine transformation from the here considered quadrilateral and pentagonal cells to a global reference cell, the evaluation of the reconstruction is badly conditioned if two or more degrees of freedom lie very close to each other in a relative way. This is the case for the pentagons if and only if it doesn't differ much from the other possible cell types.

To enforce the continuity over the boundary of the cut cell that lies in the interior of the domain, one could try to use the nine previous basis functions to interpolate the nine point values that lie on this part of the boundary (all but the value on the diagonal and the cell average value). By adding two basis functions of higher order, which vanish on the boundary of the original uncut cell, one could integrate the two missing values and obtain a reconstruction that is continuous across all four internal cell boundaries. This is realized by the so-called "bubble function" in the triangular grids [28]. However, this is not possible here since the this interpolation problem is singular.

Another possibility that solves both problems is the use of a triangularization of the cut cells. Here, the pentagonal cell can be split into three triangles which results in two new point values on the new edges between the triangles and three cell average values of the triangles. We would therefore deviate from the Cartesian grid at the boundary. The reconstruction on the triangles could be done without any problems and the Active Flux method could be adjusted to fit for the newly created cells. Unfortunately, this idea suffers from severe time step restrictions when using Simpson's rule. We have observed instabilities as soon as the time step was large enough that the correct area of integration covered more than the width of the smallest triangular cell. This can make the time step arbitrarily small, again. Exact integration could solve this problem but becomes increasingly more complex the more cells are covered by the area of integration.

We therefore use the following strategy: Let $C_{i,j}$ be the pentagon in 3.13. Then, the degrees of freedom $Q^n_{i-\frac{1}{2},j+\frac{1}{2}}, Q^n_{i-\frac{1}{2},j}, Q^n_{i-\frac{1}{2},j-\frac{1}{2}}, Q^n_{i,j-\frac{1}{2}}, Q^n_{i+\frac{1}{2},j-\frac{1}{2}}$ are connected to the neighboring, most likely regular, cells $C_{i-1,j}$ and $C_{i,j-1}$. We use the basis functions from table 3.1 and choose the coefficients c_1, c_2, c_3, c_7, c_8 to the basis functions N_1, N_2, N_3, N_7, N_8, to interpolate these point values. After that, we solve a least square problem to approximate the missing five point values as best as we can with the help of the remaining basis functions N_4, N_5, N_6, i.e., we search the coefficients c_4, c_5, c_6. The coefficient c_9 of the basis function N_9 influences the point value on the midpoint of the diagonal boundary of the pentagon. To conserve the cell average value exactly we can incorporate it into the setting of the least square problem beforehand. This strategy allows for a well conditioned evaluation of the reconstruction, even if the cell is almost degenerate. However, the continuity is lost. We only recover continuity along the long faces of the pentagon. Computations show that the discontinuities along the short faces of the pentagon are really small since the used basis functions on the boundaries of the cell are the same. The results in the next section show that the stability is not affected by this. It is interesting to notice what happens in the degenerate limit, i.e., when the degrees of freedom that lie close to each other collapse. If the three point on the diagonal of the pentagon collapse, the degenerating limiting shape is the rectangle and the solution to the least square problem becomes the standard reconstruction to the Cartesian grid cells which interpolates all degrees of freedom. If on the other hand the three values on both of the two cut faces each collapse and the limiting shape is a triangle, then the least square problem becomes overdetermined but any solution yields a valid third order reconstruction in the triangle. Lastly, if the limiting shape is a quadrilateral, i.e., if only three values on one of the cut faces of the pentagon collapse, then the canonical reconstruction on quadrilaterals is recovered which is unproblematic if the remaining cut face is not too small.

Figure 3.14 shows the comparison between a naive reconstruction with two added basis functions of higher order and the mixed interpolation and least square reconstruction for a pentagonal cell whose shape is close to a rectangle (relative cell size of $1 - 2 \cdot 10^{-8}$ to a Cartesian cell) and its surrounding cells. In this test, the three point values along the diagonal cut of the pentagon differ by about 10^{-4}. An artifical error of 10^{-6} is added to one of these point values, leading to a wild reconstruction when using the naive reconstruction. Although the interpolation is achieved, the quality of the

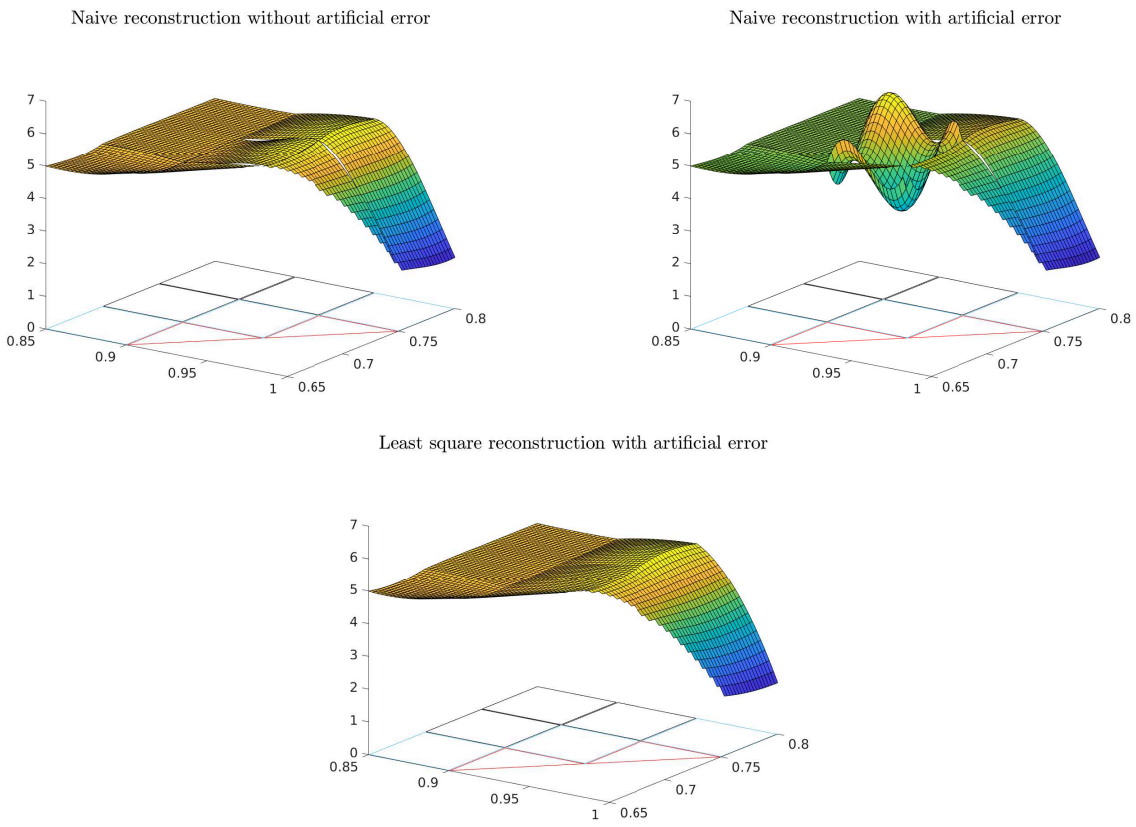

Figure 3.14.: Reconstruction in a pentagonal cell and its neighbors. Top left: Reconstruction with two added basis functions of higher order. Top right: Reconstruction with two added basis functions of higher order with a small artificial error in one of the point values on the boundary. Bottom: Reconstruction with the least square ansatz with the same artificial error.

reconstruction with the added basis functions is not satisfying. The least square approximation shows an improved reconstruction. The very small jumps in the transition between the pentagon and the neighboring triangles are not visible to the naked eye. Also, there is no visible difference between introducing the error and not introducing the error when using the mixed interpolation and least square reconstruction.

The procedure can be followed in a very similar fashion for the quadrilaterals. While the continuity would not produce a problem because of the use of the same nine basis functions, the almost degeneration of the cell to a triangle was observed even more commonly and caused the same problems for the condition. Thus, we use five of the degrees of freedom from the triangular cell to be interpolated exactly. We use the longest of the two parallel edges and the other regular edge for the choice of these five degrees of freedom. Subsequently, the sixth basis function $4\xi\eta$ that completes a third order approximation is used for a least square problem. The cell average is again conserved with the help of the "bubble function". In the degenerate limit, when three of the degrees of freedom collapse, the standard reconstruction of the limit triangular cell is recovered.

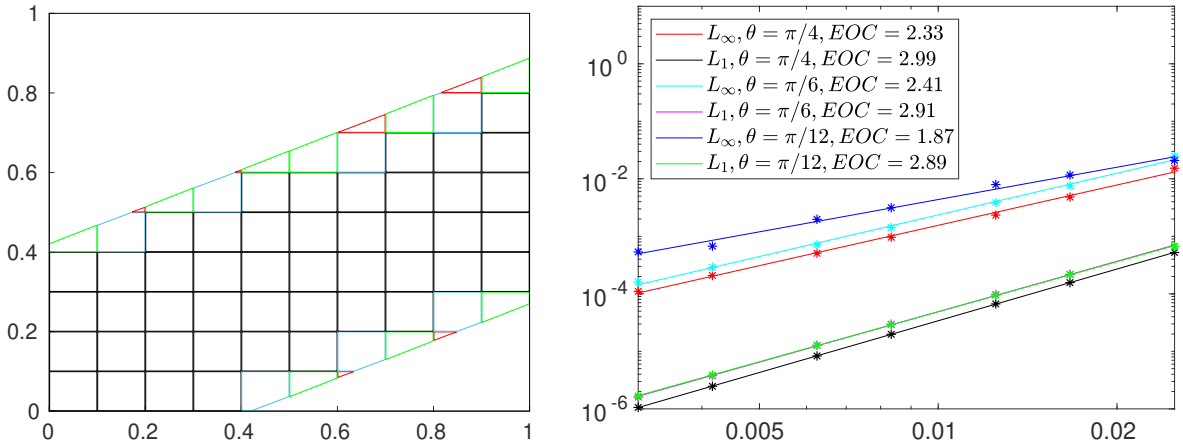

Figure 3.15.: Left: Coarse cut cell grid. Right: Errors and estimated order of convergence for $\sigma = \frac{\pi}{12}, \frac{\pi}{6}, \frac{\pi}{4}$.

In particular one can discuss when the canonical reconstruction can be used for the quadrilaterals. For nondegenerate quadrilaterals the reconstruction is well behaved if the two parallel edges are of similar length. We didn't find large differences in the accuracy of the method when switching between different reconstructions but decided to use the mixed interpolation and least square ansatz if the quotient between the length was bigger than 5 or smaller than $\frac{1}{5}$.

We now discuss the accuracy of the method.

3.2.2 Accuracy Study: Flow Along a Channel

Let $\sigma \in (0, \frac{\pi}{4}]$ be an angle, $\delta \in (0, 1)$ an offset,
$\Omega = [0,1]^2 \cap \{(x,y) \mid y - \tan(\sigma)(x-\delta) > 0 \wedge y - \delta - \tan(\sigma)x < 0\}$ a channel in two dimensions and $a = \cos(\sigma)$ and $b = \sin(\sigma)$ velocities parallel to the channel walls. The setup and an example of a cut cell grid are shown in Figure 3.15 (left).

We impose inflow boundary conditions for $x = 0$ and $y = 0$ and outflow boundary conditions for $x = 1$ and $y = 1$. Since the flow is parallel to the channel walls, there is no flow across the boundary. The initial condition reads

$$q_0(x,y) = 5\exp(-100(x+y-0.7)^2). \tag{3.42}$$

We use $\Delta t = 0.7 \max\{\frac{\Delta x}{a}, \frac{\Delta y}{b}\}$ and the final time $T = 0.4$. A similar test using a discontinuous Galerkin cut cell method is performed in [25].

We estimate the order of convergence by the solution to the least square problem that is given by fitting a straight line to the logarithmic errors. We test for various offsets δ and angles σ. The results for some values of σ and $\delta = 0.2001$ using exact integration are shown in Figure 3.15 (right). The results for all other tested values look very similar.The method remains stable and accurate for any cut cell size. Third order is achieved in the L_1 norm and second order or better is achieved in the L_∞ norm. In these tests, the size of the smallest cut cell varied between a factor of 10^{-3} to 10^{-8} compared to the regular cells.

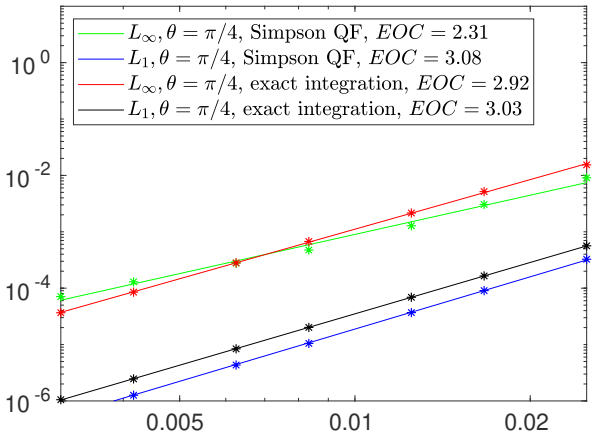

Figure 3.16.: Estimated error of convergence of the two-dimensional cut cell Active Flux method for the special grid configuration $\delta = 0.2 + \frac{\Delta x}{2}$ and $\sigma = \frac{\pi}{4}$.

It is especially worth to notice that we do not see any reduction in stability when using Simpson's rule to calculate the fluxes. The accuracy is very similar. However, it is important to notice that we don't obtain third order in the L_∞ norm. While we were able to recover the third order by using exact integration in the one dimensional case, we don't manage to do that here. We are convinced that this lies in the fact that we have different types of cells where different reconstructions with different bases are used so that the truncation error does not vanish. We support this argument by the following test problem:

Let $\delta = 0.2 + \frac{\Delta x}{2}$ and $\sigma = \frac{\pi}{4}$. Then, we only have triangular and pentagonal cells near the boundary. The triangular cells will always have relative size $\frac{1}{8}$ and the pentagons will always have relative size $\frac{7}{8}$ compared to the regular cells. We perform the same test as above with this configuration. The slight change of domain size will barely have any impact on the estimated error of convergence. Figure 3.16 shows third order convergence in this special case. This means that it is possible to recover the full third order using exact integration, provided that the used reconstructions are somewhat regular. The use of Simpson's rule results in a comparable result to the previous test with an estimated error of convergence of about 2.3 in the L_∞ norm. It is also in line with the order reduction that we saw in the one dimensional case in table 2.3.

3.2.3 Linear Stability and Cancellation Property

Since the cut cell shapes vary in size and shape for every problem, we cannot study the eigenvalues of the update matrix as in Section 3.1.4. In all of our numerical tests we found the method to be stable for the same time steps that are allowed for regular Cartesian grids (compare with the stability regions in Figure 3.8).

We can, however, study the cancellation property with the help of the test problem from the previous section. As pentagonal cells are always at least half as big as regular cells, the cancellation property is trivially achieved. Consider a triangular cell $C_{i,j}$ which edges are the diagonal cell boundary, the bottom cut interface with length $\alpha\Delta x$

and the right cut interface with length $\beta\Delta y$ for some $0 < \alpha < 1$, $0 < \beta < 1$. Since there is no flux across the boundary, the cell average update to this cell reads

$$Q_{i,j}^{n+1} = Q_{i,j}^{n} - \frac{\Delta t}{|C_{i,j}|}\left(F_{i+\frac{1}{2},j}^{n} - G_{i,j-\frac{1}{2}}^{n}\right). \tag{3.43}$$

We proof the following theorems:

Theorem 3.2.1. *The two-dimensional Active Flux method for the advection equation on cut cell grids using exact integration in the test problem from Section 3.2.2 fulfills the cancellation property in the triangular cells of size $|C_{i,j}|$, i.e.,*

$$|F_{i+\frac{1}{2},j}^{n} - G_{i,j-\frac{1}{2}}^{n}| = \mathcal{O}(|C_{i,j}|), \tag{3.44}$$

if a globally continuous reconstruction is used.

Proof. We use the flux form presented in (3.24) before applying Simpson's rule for $F_{i+\frac{1}{2},j}^{n}$ and an analogue formula for $G_{i,j-\frac{1}{2}}^{n}$. For this cut cell, we have

$$\begin{aligned} |F_{i+\frac{1}{2},j}^{n} - G_{i,j-\frac{1}{2}}^{n}| = \Bigg| & \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-a\Delta t}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}-\frac{b}{a}(x_{i+\frac{1}{2}}-x)}^{y_{j-\frac{1}{2}}+\beta\Delta y-\frac{b}{a}(x_{i+\frac{1}{2}}-x)} Q_{rec}^{n}(x,y)\ \mathrm{d}y\ \mathrm{d}x \\ & - \frac{1}{\Delta t}\int_{y_{j-\frac{1}{2}}-b\Delta t}^{y_{j-\frac{1}{2}}}\int_{x_{i+\frac{1}{2}}-\alpha\Delta x-\frac{a}{b}(y_{i-\frac{1}{2}}-y)}^{x_{i+\frac{1}{2}}-\frac{a}{b}(y_{i-\frac{1}{2}}-y)} Q_{rec}^{n}(x,y)\ \mathrm{d}x\ \mathrm{d}y\Bigg| \end{aligned} \tag{3.45}$$

The two areas of integration partly overlap and thus cancel each other up to two triangular areas of integration, one of them being the cell $C_{i,j}$ itself. The other triangle, $\triangle_C$, is obtained by the shift $(-a\Delta t, -b\Delta t)$ from $C_{i,j}$. Figure 3.17 shows both areas for an exemplary triangle. Furthermore, Q_{rec}^{n} is locally Lipschitz continuous since it is globally continuous. Let $L_Q \geq 0$ be the corresponding Lipschitz constant. Then we have:

$$\begin{aligned} |F_{i+\frac{1}{2},j}^{n} - G_{i,j-\frac{1}{2}}^{n}| &= \left|\frac{1}{\Delta t}\int_{C_{i,j}} Q_{rec}^{n}(x,y)\ \mathrm{d}x\ \mathrm{d}y - \frac{1}{\Delta t}\int_{\triangle_C} Q_{rec}^{n}(x,y)\ \mathrm{d}x\ \mathrm{d}y\right| \\ &= \frac{1}{\Delta t}\left|\int_{C_{i,j}} Q_{rec}^{n}(x,y) - Q_{rec}^{n}(x-a\Delta t, y-b\Delta t)\ \mathrm{d}x\ \mathrm{d}y\right| \\ &\leq \frac{1}{\Delta t}\int_{C_{i,j}} L_Q(a+b)\Delta t\ \mathrm{d}x\ \mathrm{d}y \\ &= L_Q(a+b)|C_{i,j}| = \mathcal{O}(|C_{i,j}|) \end{aligned} \tag{3.46}$$

□

Theorem 3.2.2. *The two-dimensional Active Flux method for the advection equation on cut cell grids using Simpson's rule in the test problem from Section 3.2.2 fulfills the*

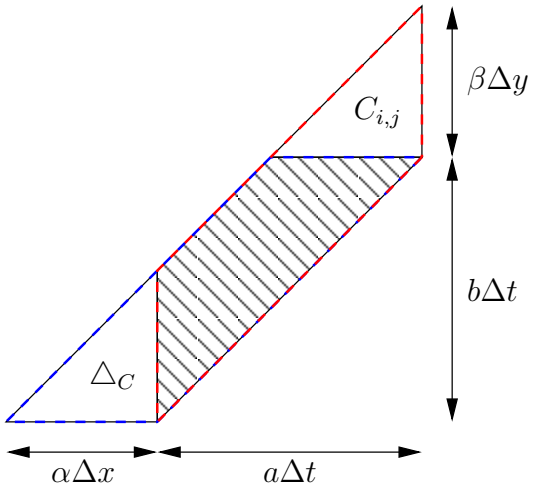

Figure 3.17.: Comparison of the areas of integration of the two fluxes of the triangular cell $C_{i,j}$, outlined in red and blue dashed lines. The two areas cancel to the triangles $C_{i,j}$ and $\triangle_C$ of total size $2|C_{i,j}|$.

cancellation property in the triangular cells of size $|C_{i,j}|$, i.e.,

$$|F^n_{i+\frac{1}{2},j} - G^n_{i,j-\frac{1}{2}}| = \mathcal{O}(|C_{i,j}|), \tag{3.47}$$

if a globally continuous reconstruction is used.

Proof. The cell size is given by $|C_{i,j}| = \frac{1}{2}\alpha\Delta x\beta\Delta y$. Since the propagation is aligned to the slope of the channel, we also have

$$\frac{\beta\Delta y}{\alpha\Delta x} = \tan(\sigma) = \frac{b}{a}. \tag{3.48}$$

Q^n_{rec} is locally Lipschitz continuous since it is globally continuous. Let $L_Q \geq 0$ be the corresponding Lipschitz constant. Let further

$$F^n_{i+\frac{1}{2},j} = \beta\Delta y \sum_{k=0}^{m} a_k f(Q^n_{rec}(s_k, t_k)) \tag{3.49}$$

be the used quadrature formula for the flux $F^n_{i+\frac{1}{2},j}$ with weights $a_k > 0$ and nodes s_k, t_k in the correct area of integration. In our case this is Simpson's rule but the statement holds for any quadrature formula. Then, we can also use the same quadrature rule for $G^n_{i,j-\frac{1}{2}}$ with the same weights a_k and the nodes $\widetilde{s}_k, \widetilde{t}_k$ where

$$\|(s_k - \widetilde{s}_k, t_k - \widetilde{t}_k)\|_2 \leq \beta\Delta y + \alpha\Delta x \quad \forall k \in \{0, \dots, m\}. \tag{3.50}$$

We obtain:

$$\begin{aligned}
|F^n_{i+\frac{1}{2},j} - G^n_{i,j-\frac{1}{2}}| &= \left| \sum_{k=0}^{m} a_k \left(\beta \Delta y a Q^n_{rec}(s_k, t_k) - \alpha \Delta x b Q^n_{rec}(\widetilde{s}_k, \widetilde{t}_k) \right) \right| \\
&\leq \sum_{k=0}^{m} a_k \beta \Delta y a \left| Q^n_{rec}(s_k, t_k) - Q^n_{rec}(\widetilde{s}_k, \widetilde{t}_k) \right| \\
&\leq \sum_{k=0}^{m} a_k \beta \Delta y a L_Q (\beta dy + \alpha \Delta x) \\
&= \sum_{k=0}^{m} 2 a_k L_Q (a+b) |C_{i,j}| = \mathcal{O}(|C_{i,j}|)
\end{aligned} \tag{3.51}$$

□

Note that our reconstruction is not necessarily continuous across the interfaces. However, as already mentioned, we did not find any instabilities when using a large time step that is suited for the regular grid.

3.3 Summary

We have seen how to expand the Active Flux method to the two-dimensional case. In contrast to the one-dimensional case there now is a difference between Simpson's rule and exact evolution, even without the presence of cut cells. The stability is slightly reduced when using Simpson's rule, too. In our examination of the Active Flux method on cut cell grids, new problems arise when trying to unify the reconstruction. We reconstruct by using a mixed interpolation and least square ansatz. We obtain a positive answer to the question of the usability of the Active Flux methods for cut cell discretizations and find very good results for accuracy and stability. In particular, no further stabilization technique has to be applied. While the foundations of the method have been laid, some questions and possibilities for further development are still at hand. A list can be found in Chapter 5.

4 The ADER Interpretation of the Active Flux Method

This chapter deals with the ADER interpretation of the Active Flux method. The goal is to extend the scheme to nonlinear flux functions f. In general, an exact evolution operator for the update of the point values is not known. Therefore, an approximate evolution has to be used. In a recent work, Roe presents a way to rewrite the evolution equation for Burgers' equation in order to obtain an approximate evolution [43]. This idea will be discussed in more detail in the next section, where we consider Burgers' equation in one and two spatial dimensions and adapt the Active Flux methods for linear equations from Chs. 2 and 3 to it. We use the implicit representation for the exact evolution operator $\mathcal{L}_f$ to state an iterative process that finds an approximate update after only a few iterations and cures some failures of the originally suggested update. For more complicated nonlinear flux functions it is difficult to generalize this approach. Recently, Barsukow studied possible approximate evolutions [32]. For scalar conservation laws, the there presented idea uses a fixed point iteration to compute approximate characteristic origins just like our iterative approach. For systems of conservation laws, he then develops two approaches that extend this iterative idea and try to follow the idea of characteristics closely.

To find an approach for general conservation laws, we instead explore another approach that is based on a different interpretation of the Active Flux method. We show that this new interpretation is equivalent to the Active Flux scheme for linear systems in one dimension and carry out how an arbitrary nonlinear system of hyperbolic conservation laws can be solved with this technique. In [1], we have developed this method.

For simplicity, when the expressions become too long, we sometimes use the abbreviations

$$\frac{\partial}{\partial x}q(x,y,t) =: q_x(x,y,t), \quad \frac{\partial^2}{\partial x^2}q(x,y,t) =: q_{xx}(x,y,t), \quad \frac{\partial^2}{\partial x \partial y}q(x,y,t) =: q_{xy}(x,y,t) \tag{4.1}$$

and so on for the multidimensional derivatives.

4.1 Active Flux Methods for Burgers' Equation

In this section we deal with the first nonlinear flux function which is Burgers' equation. We adapt the already defined Active Flux method to this simplest nonlinear equation. The original idea originates from an article of Roe [43] and is worked out in more detail here. We will see that new problems arise due to the nonlinearity. We first discuss the one-dimensional case and then state the necessary changes for the two-dimensional case.

4.1.1 The One-dimensional Case

We consider Burgers' equation

$$\frac{\partial}{\partial t}q(x,t) + \frac{\partial}{\partial x}\left(\frac{q^2(x,t)}{2}\right) = 0, \tag{4.2}$$

i.e., $f(q) = \frac{q^2}{2}$. We use our description in [2, Section 2.3], adapted to the one-dimensional case. For smooth solutions this equation can also be written in the form

$$q_t(x,t) + q(x,t)q_x(x,t) = 0. \tag{4.3}$$

This form suggests the following implicitely defined evolution formula:

$$\begin{aligned} q(x,t+\tau) &= q(x - q(x,t+\tau)\tau, t) \\ &= q(x,t) - q(x,t+\tau)\tau q_x(x,y,0) + \mathcal{O}(\tau^2) \\ \Leftrightarrow\ q(x,t+\tau) &= \frac{q(x,t)}{1+\tau q_x(x,t)} + \mathcal{O}(\tau^2) \end{aligned} \tag{4.4}$$

If we drop the term $\mathcal{O}(\tau^2)$, we obtain a second order accurate approximation to $q(x,y,t+\tau)$. By plugging in this value in the first line of(4.4), we even obtain a third order accurate approximation:

$$\begin{aligned} q\left(x - \frac{q(x,t)}{1+\tau q_x(x,t)}\tau, t\right) &= q(x - (q(x,t+\tau) + \mathcal{O}(\tau^2))\tau, t) \\ &= q(x - q(x,t+\tau)\tau + \mathcal{O}(\tau^3), t) \\ &= q(x - q(x,t+\tau)\tau, t) + \mathcal{O}(\tau^3) \\ &= q(x,t+\tau) + \mathcal{O}(\tau^3) \end{aligned} \tag{4.5}$$

Roe understands the approximation (4.4) to $q(x,t+\tau)$ as a *corrected* wave speed to then find the required point values via (4.5). Using simply the interface value for $q(x,t)$ instead of (4.4) as an approximation to the wave speed, one would only obtain a first order accurate approximation to the wave speed and therefore a second order accurate approximation to the update. This approach leads to multiple problems. First of all

the reconstruction Q^n_{rec} is not differentiable at the interface. One has to state how the numerical value for $q_x(x,t)$ at the interface can be determined. For this one can use the interface value as an approximation of first order. If it is greater than 0 one uses the derivative from below, otherwise the derivative from above is used. To guarantee the CFL condition the wave speed (4.4), i.e., the slope of the characteristic, has to be small enough. If the denominator is small in modulus, the approximation considerably worsens. The time step has hence to be fitted that the modulus of the denominator is bounded from below.

An additional problem shows in the following situation (taken and adapted from [1, Section 5.5]): We'd like to compute the fluxes $F^n_{i-\frac{1}{2}}$ and $F^n_{i+\frac{1}{2}}$ and let $Q^n_{i-\frac{1}{2}} > 0$ and $Q^n_{i+\frac{1}{2}} < 0$. This situation is shown in Figure 4.1. For the update of the point value $Q^n_{i-\frac{1}{2}}$ only characteristics with positive slope are used and for the update of $Q^n_{i+\frac{1}{2}}$ only characteristics with negative slope. One can often observe this setting in the proximity of a shock wave. The cell average value Q^n_i is not using in any update and possibly grows/drops indefinitely within the next time steps. In [1] we suggested to reconstruct in a discontinuous way in this situation. For this a Riemann problem has to be solved at the interface to obtain the first approximation to the wave speed. The approximation can change signs this way and thus the characteristic will point in the right direction. To obtain a solution without using limiting to stabilize the method we propose the following method by now [2]:

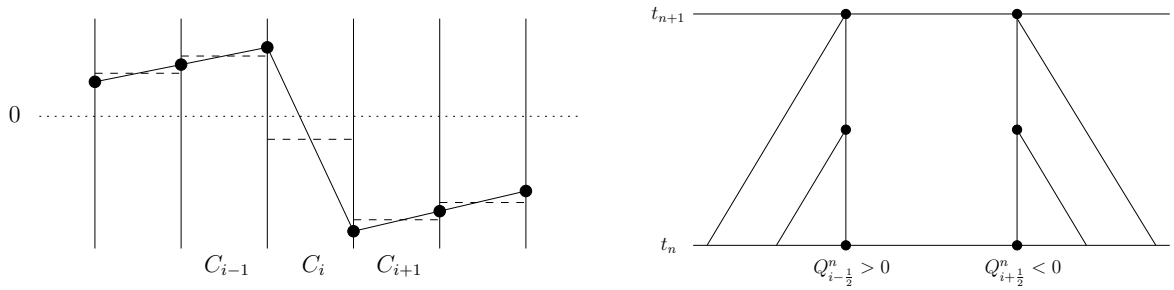

Figure 4.1.: Left: Problematic data for the Active Flux approximation applied to the Burgers' equation. Right: Characteristics for the cell C_i never origin in cell C_i.

To compute the required point values $Q^{n+\frac{1}{2}}_{i+\frac{1}{2}}$ and $Q^{n+1}_{i-\frac{1}{2}}$ let $Q^{(0)}_{i+\frac{1}{2}}$ be an approximation of first order accuracy. Then, we use the iteration

$$Q^{(l+1)}_{i+\frac{1}{2}} = Q^n_{rec}\left(x_{i+\frac{1}{2}} - \tau Q^{(l)}_{i+\frac{1}{2}}\right), \quad l = 0, 1, \ldots \tag{4.6}$$

Each iteration improves the accuracy by one order as one can see by the calculations in (4.5). We thus acquire a third order accurate approximation after two iterations. To remove the previously discussed instability we don't choose the obvious initial values $Q^{(0)}_{i+\frac{1}{2}} = Q^n_{i+\frac{1}{2}}$, but

$$Q^{(0)}_{i+\frac{1}{2}} = \frac{1}{2}\left(Q^n_i + Q^n_{i+1}\right). \tag{4.7}$$

With this, the possible domain of dependence is no longer immediately isolated to one of the two neighboring cells but we first get an adequate estimate from the two neighboring cell average values. This leads to a stronger coupling of the interface and cell average values and removes the instabilities. Since all iterations now purely exist of evaluations of the reconstruction (or in the case of the initial value of a convex combination of the cell averages), the CFL condition can be determined by the values of the current time step a priori and is not violated by possibly too big wave speeds. In order to do so it is necessary to determine the global extrema of the reconstruction. We thus use

$$\nu := \max_x |Q^n_{rec}(x)| \frac{\Delta t}{\Delta x} \leq 1. \tag{4.8}$$

To remove any oscillations, we use the limiting from Section 2.5.2. We only show numerical results for the two-dimensional case here, but compare the here discussed method to a more general approach in Section 4.2.1.3.

4.1.2 The Two-dimensional Case

The afore presented method can be extended to the multidimensional case (compare again to [2, Section 2.3]). We now consider

$$\frac{\partial}{\partial t} q(x,y,t) + \frac{\partial}{\partial x}\left(\frac{q^2(x,y,t)}{2}\right) + \frac{\partial}{\partial y}\left(\frac{q^2(x,y,t)}{2}\right) = 0 \tag{4.9}$$

as well as the formulation

$$q_t(x,y,t) + q(x,y,t)q_x(x,y,t) + q(x,y,t)q_y(x,y,t) = 0. \tag{4.10}$$

We adapt the Active Flux method presented in Chapter 3. In analogy to (4.4) and (4.5), implicit evolution formulas are derived to achieve an improvement of one order per iteration:

$$q(x,y,t+\tau) = q(x - q(x,y,t+\tau)\tau, y - q(x,y,t+\tau), t) \tag{4.11}$$

and

$$\begin{aligned} & q(x - (q(x,y,t+\tau) + \mathcal{O}(\tau^2))\tau, y - (q(x,y,t+\tau) + \mathcal{O}(\tau^2))\tau, t) \\ &= q(x - q(x,y,t+\tau)\tau + \mathcal{O}(\tau^3), y - q(x,y,t+\tau)\tau + \mathcal{O}(\tau^3), t) \\ &= q(x - q(x,y,t+\tau)\tau, y - q(x,y,t+\tau)\tau, t) + \mathcal{O}(\tau^3) \\ &= q(x,y,t+\tau) + \mathcal{O}(\tau^3). \end{aligned} \tag{4.12}$$

Therefore, the iteration reads

$$Q^{(l+1)}_{i+\frac{1}{2},j+\frac{1}{2}} = Q^n_{rec}\left(x_{i+\frac{1}{2}} - \tau Q^{(l)}_{i+\frac{1}{2},j+\frac{1}{2}}, y_{j+\frac{1}{2}} - \tau Q^{(l)}_{i+\frac{1}{2},j+\frac{1}{2}}\right), \quad l = 0,1,\ldots \tag{4.13}$$

	unlimited iterative method				limited iterative method			
	2 iterations		3 iterations		2 iterations		3 iterations	
N	L_1-error	EOC	L_1-error	EOC	L_1-error	EOC	L_1-error	EOC
50^2	$1.0680 \cdot 10^{-5}$		$2.3993 \cdot 10^{-6}$		$1.1775 \cdot 10^{-5}$		$3.4799 \cdot 10^{-6}$	
100^2	$1.4830 \cdot 10^{-6}$	2.84	$3.3487 \cdot 10^{-7}$	2.84	$1.6126 \cdot 10^{-6}$	2.86	$4.6216 \cdot 10^{-7}$	2.91
200^2	$2.1224 \cdot 10^{-7}$	2.80	$4.5392 \cdot 10^{-8}$	2.88	$2.4202 \cdot 10^{-7}$	2.73	$7.4984 \cdot 10^{-8}$	2.64

Table 4.1.: Accuracy study for smooth solutions of the two-dimensional Burgers' equation, using the iterative approach with unlimited and limited reconstruction.

The initial values of the iterations are

$$Q^{(0)}_{i+\frac{1}{2},j} = \frac{1}{2}\left(Q^n_{i,j} + Q^n_{i+1,j}\right) \tag{4.14}$$

for the edges and

$$Q^{(0)}_{i+\frac{1}{2},j+\frac{1}{2}} = \frac{1}{4}\left(Q^n_{i,j} + Q^n_{i+1,j} + Q^n_{i,j+1} + Q^n_{i+1,j+1}\right) \tag{4.15}$$

for the corners. Analog formulas are given for all other interface values. The limiting is done as in Section 3.1.5.

We perform an accuracy study with the use of the initial values

$$q_0(x, y) = \sin(2\pi x)\sin(2\pi y) + 0.1 \tag{4.16}$$

on the domain $[0, 1] \times [0, 1]$ with double periodic boundary conditions (taken from [2, Section 4.3]. We choose the final time $T = 0.05$, where the solutions are still smooth and use both the unlimited and the limited reconstruction. With the formula [44, page 257] we can estimate an experimental convergence order without knowledge of the exact solution. The results can be found in table 4.1. We obtain third order accuracy in both cases. Additionally, we state the results if using three instead of two iterations. We don't gain another order since the reconstruction is limited to third order but the accuracy is improved by quite a bit.

To study the behavior of the solution in the existence of shock waves we take a look at the solution at time $T = 0.5$. Figure 4.2 displays the solution structure for an 100×100 grid with and without limiting. We observed that limiting successfully eliminates the arising oscillations around the shock. Figure 4.3 further shows cross sections of the solutions for the sake of a better illustration of the limiting effect.

It is difficult to transfer this adaption of the Active Flux method to general nonlinear systems. Approaches that base on finding an approximate evolution operator with the help of characteristic theory are a subject of current research [32]. In the following, we develop an alternative understanding of the Active Flux method for linear equations in arbitrary dimensions that can also be used for nonlinear equations.

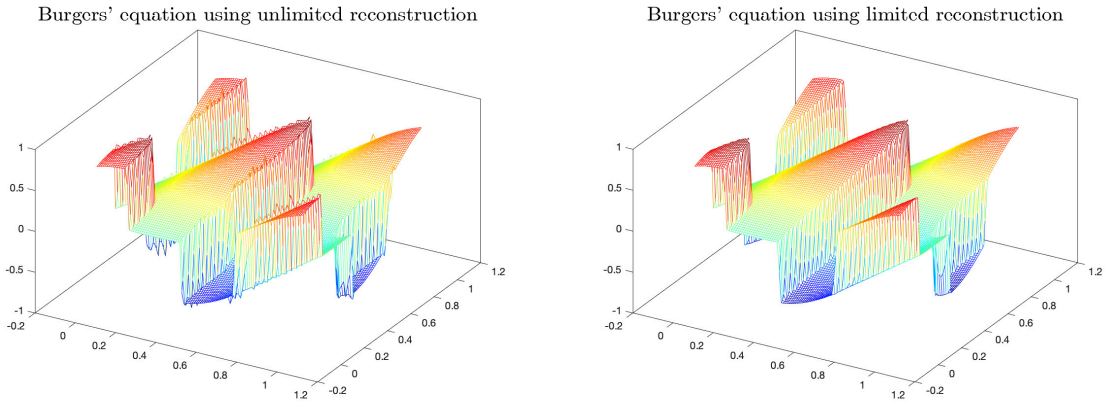

Figure 4.2.: Numerical results for Burgers' equation using the unlimited (left) and the limited (right) Active Flux method. The solution was computed on a mesh consisting of 100×100 grid cells.

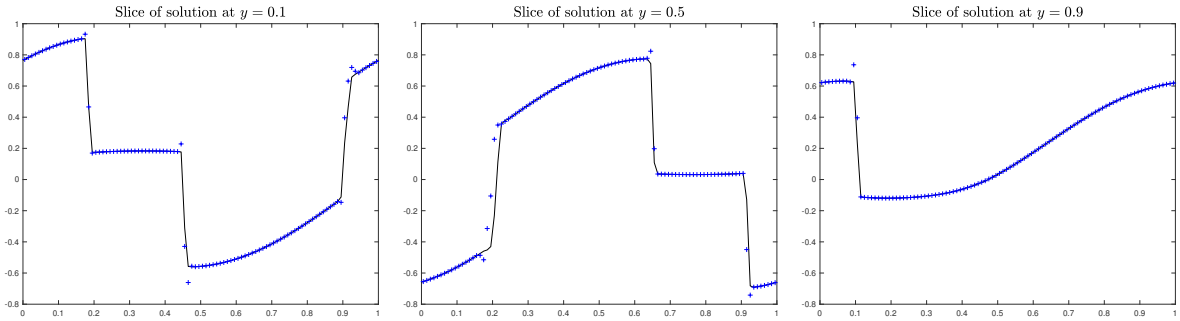

Figure 4.3.: Slices of the two-dimensional numerical solution from Figure 4.2 at $y = 0.1$, $y = 0.5$ and $y = 0.9$. The black line shows the limited solution, the blue '+' symbols the unlimited solution.

4.2 The ADER Interpretation

In the previous section we got to know an adaptation of the Active Flux method for Burgers' equation. It is based solely on the specification of an approximate evolution operator which is used to find the required new point values. The update of the cell average values is implicitly affected by the computation of the quadrature points, but no change is being made on the computation of the fluxes or the use of the finite volume method. The method presented in this section distinguishes itself only in the computation of the point values as well. The computation of the fluxes via Simpson's rule, the used degrees of freedom and the finite volume update remain unchanged. We explain the concept again on the basis of one-dimensional equations and switch to the multidimensional case afterwards.

4.2.1 One-dimensional Problems

Instead of using the exact evolution which is based on characteristic theory we instead use a Taylor series expansion in time:

$$q(x_{i+\frac{1}{2}}, t+\tau) = q(x_{i+\frac{1}{2}}, t) + \tau q_t(x_{i+\frac{1}{2}}, t) + \frac{1}{2}\tau^2 q_{tt}(x_{i+\frac{1}{2}}, t) + \mathcal{O}(\tau^3) \tag{4.17}$$

We here substitute the temporal derivatives with the help of the differential equation by spatial derivatives. This idea is used in the so-called ADER (*Arbitrary DERivative*) methods which go back to Titarev and Toro [45].

4.2.1.1 Linear Systems

We first consider the linear system

$$q_t(x,t) + Aq_x(x,t) = 0, \tag{4.18}$$

where $A \in \mathbb{R}^{m\times m}$ is a constant matrix that is diagonalizable with real eigenvalues so that the system (4.18) is hyperbolic. This section is adapted from [1, Section 3.2]. We have

$$q_t(x,t) = -Aq_x(x,t) \tag{4.19}$$

and

$$q_{tt}(x,t) = (-Aq_x(x,t))_t = (-Aq_t(x,t))_x = A^2 q_{xx}(x,t). \tag{4.20}$$

The vector-valued conserved quantities at the interface can now be approximated as follows:

$$q(x_{i+\frac{1}{2}}, t+\tau) \approx q(x_{i+\frac{1}{2}}, t) - \tau A q_x(x_{i+\frac{1}{2}}, t) + \frac{1}{2}\tau^2 A^2 q_{xx}(x_{i+\frac{1}{2}}, t) \tag{4.21}$$

Indeed, for the quadratic reconstruction this equation is exact, since all additional terms of the Taylor series expansion vanish. For $\tau = \Delta t/2$ and $\tau = \Delta t$ we get:

$$Q^{n+\frac{1}{2}}_{i+\frac{1}{2}} = Q^n_{i+\frac{1}{2}} - \frac{\Delta t}{2} A Q^n_{x,i+\frac{1}{2}} + \frac{\Delta t^2}{8} A^2 Q^n_{xx,i+\frac{1}{2}} \tag{4.22}$$

$$Q^{n+1}_{i+\frac{1}{2}} = Q^n_{i+\frac{1}{2}} - \Delta t A Q^n_{x,i+\frac{1}{2}} + \frac{\Delta t^2}{2} A^2 Q^n_{xx,i+\frac{1}{2}} \tag{4.23}$$

The values $Q^n_{x,i+\frac{1}{2}}$ and $Q^n_{xx,i+\frac{1}{2}}$ are determined by solving the Riemann problem of the following form:

$$\begin{aligned} &(Q_x)_t + A(Q_x)_x = 0 \\ &Q_x(x,t_n) = \begin{cases} \frac{1}{\Delta x}(Q^n_{rec,i})'(1) & : x < x_{i+\frac{1}{2}} \\ \frac{1}{\Delta x}(Q^n_{rec,i+1})'(0) & : x > x_{i+\frac{1}{2}} \end{cases} \end{aligned} \tag{4.24}$$

$$
\begin{aligned}
&(Q_{xx})_t + A(Q_{xx})_x = 0 \\
&Q_{xx}(x,t_n) = \begin{cases} \frac{1}{\Delta x^2}(Q^n_{rec,i})''(1) & : x < x_{i+\frac{1}{2}} \\ \frac{1}{\Delta x^2}(Q^n_{rec,i+1})''(0) & : x > x_{i+\frac{1}{2}} \end{cases}
\end{aligned} \tag{4.25}
$$

The scalings Δx respectively Δx^2 are generated from the mapping to the reference cell. At a discontinuous reconstruction one has to solve an additional Riemann problem for $Q^n_{i+\frac{1}{2}}$ because this value is no longer given uniquely in the reconstruction. This Riemann problem has the following form:

$$
\begin{aligned}
&Q_t + AQ_x = 0 \\
&Q(x,t_n) = \begin{cases} Q^n_{rec,i}(1) & : x < x_{i+\frac{1}{2}} \\ Q^n_{rec,i+1}(0) & : x > x_{i+\frac{1}{2}} \end{cases}
\end{aligned} \tag{4.26}
$$

Solving this Riemann problem corresponds to the choice of the upwind direction in 2.57 and 2.58 and here doesn't have to be only included in the flux but also in the Taylor series expansion. We call the just described method the ADER interpretation of the Active Flux method. The word interpretation originates from the following statement:

Theorem 4.2.1. *For linear hyperbolic systems in one spatial dimension the in Chapter 2 described Active Flux method and the here presented ADER interpretation are equivalent if the CFL condition* (2.20) *is satisfied.*

Proof. Let $A = RDR^{-1}$ be the eigenvector decomposition of A and $w := R^{-1}q$ the vector of characteristic variables. Let W denote the transformed values of Q. Furthermore, let λ be the vector of the diagonal of D.

We show that both methods lead to the same update for $W^{n+1}_{i+\frac{1}{2}}$. For $W^{n+\frac{1}{2}}_{i+\frac{1}{2}}$ the same argument can be used. With the help of the additional indices $k, s \in \{1, \dots, m\}$ we denote the kth or sth component of a vector or the kth or sth column of a matrix. For the in Chapter 2 described Active Flux method we then have:

$$
W^{n+1}_{i+\frac{1}{2},k} = \begin{cases} W^n_{rec,i,k}(x_{i+\frac{1}{2}} - \Delta t\lambda_k) & : \lambda_k > 0 \\ W^n_{rec,i+1,k}(x_{i+\frac{1}{2}} - \Delta t\lambda_k) & : \lambda_k < 0 \end{cases} \tag{4.27}
$$

For the here described ADER interpretation from (4.21), multiplication from the left by R^{-1} and the use of the eigenvalue decomposition we have

$$
\begin{aligned}
W^{n+1}_{i+\frac{1}{2},k} &= W^n_{i+\frac{1}{2},k} - \Delta t D_k R^{-1} Q^n_{x,i+\frac{1}{2}} + \frac{\Delta t^2}{2}(D^2)_k R^{-1} Q^n_{xx,i+\frac{1}{2}} \\
&= W^n_{i+\frac{1}{2},k} - \Delta t\lambda_k \left[R^{-1}Q^n_{x,i+\frac{1}{2}}\right]_k + \frac{\Delta t^2}{2}\lambda_k^2 \left[R^{-1}Q^n_{xx,i+\frac{1}{2}}\right]_k .
\end{aligned} \tag{4.28}
$$

The solutions to the Riemann problems(4.24) and (4.25) read

$$\begin{aligned} Q^n_{x,i+\frac{1}{2}} &= \frac{1}{\Delta x}\left((Q^n_{rec,i})'(1) + \sum_{\substack{s=1 \\ \lambda_s<0}}^{m} \left[R^{-1}((Q^n_{rec,i+1})'(0) - (Q^n_{rec,i})'(1))\right]_s R_s \right) \\ &= \frac{1}{\Delta x}\left((Q^n_{rec,i})'(1) + \sum_{\substack{s=1 \\ \lambda_s<0}}^{m} \left[(W^n_{rec,i+1})'(0) - (W^n_{rec,i})'(1)\right]_s R_s \right) \end{aligned} \tag{4.29}$$

and

$$\begin{aligned} Q^n_{xx,i+\frac{1}{2}} &= \frac{1}{\Delta x^2}\left((Q^n_{rec,i})''(1) + \sum_{\substack{s=1 \\ \lambda_s<0}}^{m} \left[R^{-1}((Q^n_{rec,i+1})''(0) - (Q^n_{rec,i})''(1))\right]_s R_s \right) \\ &= \frac{1}{\Delta x^2}\left((Q^n_{rec,i})''(1) + \sum_{\substack{s=1 \\ \lambda_s<0}}^{m} \left[(W^n_{rec,i+1})''(0) - (W^n_{rec,i})''(1)\right]_s R_s \right). \end{aligned} \tag{4.30}$$

Let e_s be the sth unit vector. Because of $R^{-1}R_s = e_s$ it follows that

$$\begin{aligned} \left[R^{-1}Q^n_{x,i+\frac{1}{2}}\right]_k &= \frac{1}{\Delta x}\left((W^n_{rec,i})'(1) + \sum_{\substack{s=1 \\ \lambda_s<0}}^{m} \left[(W^n_{rec,i+1})'(0) - (W^n_{rec,i})'(1)\right]_s e_s \right)_k \\ &= \begin{cases} \frac{1}{\Delta x}(W^n_{rec,i,k})'(1) & : \lambda_k > 0 \\ \frac{1}{\Delta x}(W^n_{rec,i+1,k})'(0) & : \lambda_k < 0 \end{cases} \end{aligned} \tag{4.31}$$

and in the same way

$$\left[R^{-1}Q^n_{xx,i+\frac{1}{2}}\right]_k = \begin{cases} \frac{1}{\Delta x^2}(W^n_{rec,i,k})''(1) & : \lambda_k > 0 \\ \frac{1}{\Delta x^2}(W^n_{rec,i+1,k})''(0) & : \lambda_k < 0 \end{cases}. \tag{4.32}$$

The substitution in (4.28) yields

$$\begin{aligned} W^{n+1}_{i+\frac{1}{2},k} &= W^n_{i+\frac{1}{2},k} - \Delta t\lambda_k \left[R^{-1}Q^n_{x,i+\frac{1}{2}}\right]_k + \frac{\Delta t^2}{2}\lambda_k^2 \left[R^{-1}Q^n_{xx,i+\frac{1}{2}}\right]_k \\ &= \begin{cases} W^n_{i+\frac{1}{2},k} - \frac{\Delta t\lambda_k}{\Delta x}(W^n_{rec,i,k})'(1) + \frac{\Delta t^2\lambda_k^2}{2\Delta x^2}(W^n_{rec,i,k})''(1) & : \lambda_k > 0 \\ W^n_{i+\frac{1}{2},k} - \frac{\Delta t\lambda_k}{\Delta x}(W^n_{rec,i+1,k})'(0) + \frac{\Delta t^2\lambda_k^2}{2\Delta x^2}(W^n_{rec,i+1,k})''(0) & : \lambda_k < 0 \end{cases}. \end{aligned} \tag{4.33}$$

This exactly matches the Taylor series expansion of (4.27) in $x_{i+\frac{1}{2}}$ due to the piecewise defined parabolic reconstruction as long as the CFL condition (2.20) holds true. □

We'd shortly like to discuss possible limitings at this point, again. In Section 2.5,

several possibilities are presented to limit the reconstruction. To keep equivalence in the limited case we need both approached to yield the same updates for the point values. The piecewise defined reconstruction can be discarded since it is not twice continuously differentiable. The hyperbolic reconstruction can also not be used: On the one hand it again needs exact integration for the flux computation to deny over- or undershoots. On the other hand it has a too small convergence radius for the Taylor expansion to yield a valid approximation. The discontinuous reconstructions can both be used. We continue to use the compressed reconstruction (2.61) that can be used in any dimension.

4.2.1.2 Nonlinear Systems

Similar to the approach for linear equations we perform the Taylor series expansion for nonlinear equations and substitute the temporal derivatives with the help of the equations by spatial derivatives. In the following, we refrain from using the (x,t)-arguments. As usual, see for example [46, p. 147], we define the application of the nth derivative at the location $q \in \mathbb{R}^m$ to the vectors $r_1, \dots, r_n \in \mathbb{R}^m$ by

$$f^{(n)}(q) \cdot (r_1, \dots, r_n) = \sum_{i_1,\dots,i_n}^{m} \frac{\partial^n f(q)}{\partial q_{i_1}, \dots \partial q_{i_n}} r_{1,i_1} \cdots r_{n,i_n} \tag{4.34}$$

and use this notation for second and higher order derivatives. This section is adapted from [1, Section 4.2]. For a general hyperbolic system in the form (1.1) in one dimension we have the following equations:

$$q_t = -f(q)_x = -f'(q)q_x \tag{4.35}$$
$$q_{tx} = q_{xt} = -(f'(q)q_x)_x = -(f''(q)\cdot(q_x,q_x) + f'(q)q_{xx}) \tag{4.36}$$
$$q_{tt} = -(f'(q)q_x)_t = -(f''(q)\cdot(q_t,q_x) + f'(q)q_{xt}) \tag{4.37}$$

Thus, the update is given by

$$\begin{aligned} q(x_{i+\frac{1}{2}}, t_n + \tau) \approx\ & Q^n_{i+\frac{1}{2}} - \tau f'(Q^n_{i+\frac{1}{2}}) Q^n_{x,i+\frac{1}{2}} \\ & + \frac{1}{2}\tau^2 \Big(f''(Q^n_{i+\frac{1}{2}}) \cdot (f'(Q^n_{i+\frac{1}{2}}) Q^n_{x,i+\frac{1}{2}}, Q^n_{x,i+\frac{1}{2}}) \\ & + f'(Q^n_{i+\frac{1}{2}}) f''(Q^n_{i+\frac{1}{2}}) \cdot (Q^n_{x,i+\frac{1}{2}}, Q^n_{x,i+\frac{1}{2}}) \\ & + (f'(Q^n_{i+\frac{1}{2}}))^2 Q^n_{xx,i+\frac{1}{2}} \Big). \end{aligned} \tag{4.38}$$

The values $Q^n_{x,i+\frac{1}{2}}$ and $Q^n_{xx,i+\frac{1}{2}}$ are again found by solving the linearized Riemann problem of the following form:

$$\begin{aligned} & (Q_x)_t + f'(Q^n_{i+\frac{1}{2}})(Q_x)_x = 0 \\ & Q_x(x,t_n) = \begin{cases} \frac{1}{\Delta x}(Q^n_{rec,i})'(1) & : x < x_{i+\frac{1}{2}} \\ \frac{1}{\Delta x}(Q^n_{rec,i+1})'(0) & : x > x_{i+\frac{1}{2}} \end{cases} \end{aligned} \tag{4.39}$$

$$(Q_{xx})_t + f'(Q^n_{i+\frac{1}{2}})(Q_{xx})_x = 0$$
$$Q_{xx}(x,t_n) = \begin{cases} \frac{1}{\Delta x^2}(Q^n_{rec,i})''(1) & : x < x_{i+\frac{1}{2}} \\ \frac{1}{\Delta x^2}(Q^n_{rec,i+1})''(0) & : x > x_{i+\frac{1}{2}} \end{cases} \tag{4.40}$$

If we reconstruct at an interface $i+\frac{1}{2}$ in a discontinuous way we again have to solve a Riemann problem for $Q^n_{i+\frac{1}{2}}$.

It is important to notice that the time step can no longer be chosen a priori so that (4.8) is no longer sufficient. Because of the nonlinearity of (4.38) for nonlinear flux functions we cannot easily predict if the so found values which are not only used for the update of the interface value but also for the flux computation still satisfy the CFL condition. Therefore, the time step has to be adjusted and possibly reduced before every update. In all of our computations this rarely posed a problem and led to almost no slowdown of the method.

We now consider two typical representatives for nonlinear equations.

4.2.1.3 Burgers' Equation

We have already seen how to solve Burgers' equation with the use of an iteration for the update of the interface values. We will now compare this ansatz with the new method. This section is based on [1, Section 4.1]. We have

$$f(q) = \frac{q^2}{2}, \quad f'(q) = q, \quad f''(q) = 1. \tag{4.41}$$

Thus, the update of the interface values reads:

$$q(x_{i+\frac{1}{2}}, t_n+\tau) \approx Q^n_{i+\frac{1}{2}} - \tau Q^n_{i+\frac{1}{2}} Q^n_{x,i+\frac{1}{2}} + \frac{1}{2}\tau^2\left(2Q^n_{i+\frac{1}{2}}(Q^n_{x,i+\frac{1}{2}})^2 + (Q^n_{i+\frac{1}{2}})^2 Q^n_{xx,i+\frac{1}{2}}\right) \tag{4.42}$$

The solutions to the linearized Riemann problems are

$$Q^n_{x,i+\frac{1}{2}} = \begin{cases} \frac{1}{\Delta x}(Q^n_{rec,i})'(1) & : Q^n_{i+\frac{1}{2}} > 0 \\ \frac{1}{\Delta x}(Q^n_{rec,i+1})'(0) & : Q^n_{i+\frac{1}{2}} < 0 \end{cases} \tag{4.43}$$

and

$$Q^n_{xx,i+\frac{1}{2}} = \begin{cases} \frac{1}{\Delta x^2}(Q^n_{rec,i})''(1) & : Q^n_{i+\frac{1}{2}} > 0 \\ \frac{1}{\Delta x^2}(Q^n_{rec,i+1})''(0) & : Q^n_{i+\frac{1}{2}} < 0 \end{cases}. \tag{4.44}$$

We compare both approaches for the initial value problem (4.2) with the initial values

$$q_0(x) = \sin(2\pi x) \tag{4.45}$$

on the interval $[-1,1]$ with periodic boundary conditions at the final time $T = 0.15$. Here, the solution is still smooth but shortly after a shock wave will form. We use $\nu \leq 0.9$ with regard to (4.8). We can see in Table 4.2 that both methods produce similar values.

	Our iterative method		ADER interpretation of the Active Flux method	
N	L_1-error	EOC	L_1-error	EOC
64	$2.9688 \cdot 10^{-4}$		$1.0725 \cdot 10^{-4}$	
128	$8.9769 \cdot 10^{-5}$	1.73	$4.1553 \cdot 10^{-5}$	1.37
256	$1.8047 \cdot 10^{-5}$	2.31	$1.1451 \cdot 10^{-5}$	1.86
512	$2.5751 \cdot 10^{-6}$	2.81	$1.4751 \cdot 10^{-6}$	2.96
1024	$3.4858 \cdot 10^{-7}$	2.89	$2.0518 \cdot 10^{-7}$	2.85
2048	$4.4386 \cdot 10^{-8}$	2.97	$2.6110 \cdot 10^{-8}$	2.97
4096	$5.5383 \cdot 10^{-9}$	3.00	$3.1904 \cdot 10^{-9}$	3.03

Table 4.2.: Convergence study for the Burgers' equation (4.2) with initial data (4.45) at time $T = 0.15$, using our iterative method as well as our ADER version of the Active Flux method.

4.2.1.4 Euler Equations

This section is adapted from [1, Section 4.2 and App. A]. We now consider the one-dimensional Euler equations which are given by the following quantities:

$$q = \begin{pmatrix} \rho \\ \rho v \\ E \end{pmatrix}, \quad f(q) = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ v(E+p) \end{pmatrix}. \tag{4.46}$$

Here, ρ, v, p and E represent the density, velocity, pressure and energy, respectively. To close the system we use the ideal gas equation

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho v^2 \tag{4.47}$$

with adiabatic exponent $\gamma = 1.4$ as the equation of state.

As before, we compute f' and f'' and uses these in the approximation (4.38). We have

$$f'(q) = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2}(\gamma-3)v^2 & (3-\gamma)u & (\gamma-1) \\ \frac{1}{2}(\gamma-1)v^3 - vH & H - (\gamma-1)v^2 & \gamma v \end{pmatrix} \tag{4.48}$$

with $H = (E+p)/\rho$. Let f_1, f_2 and f_3 be the three components of the flux f. It

further holds:

$$\begin{aligned}
\left(\frac{\partial^2 f_1}{\partial q_i \partial q_j}(q)\right)_{i,j=1,\dots,3} &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\left(\frac{\partial^2 f_2}{\partial q_i \partial q_j}(q)\right)_{i,j=1,\dots,3} &= \begin{pmatrix} (3-\gamma)\frac{v^2}{\rho} & (\gamma-3)\frac{v}{\rho} & 0 \\ (\gamma-3)\frac{v}{\rho} & (3-\gamma)\frac{1}{\rho} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\left(\frac{\partial^2 f_3}{\partial q_i \partial q_j}(q)\right)_{i,j=1,\dots,3} &= \begin{pmatrix} 2\gamma E\frac{v}{\rho^2} - 3(\gamma-1)\frac{v^3}{\rho} & -\gamma\frac{E}{\rho^2} + 3(\gamma-1)\frac{v^2}{\rho} & -\gamma\frac{v}{\rho} \\ -\gamma\frac{E}{\rho^2} + 3(\gamma-1)\frac{v^2}{\rho} & -3(\gamma-1)\frac{v}{\rho} & \gamma\frac{1}{\rho} \\ -\gamma\frac{v}{\rho} & \gamma\frac{1}{\rho} & 0 \end{pmatrix}.
\end{aligned} \tag{4.49}$$

By a linearization around the interface value and solving the Riemann problems (4.38) is fully described. In case of a limiting we solve the additional Riemann problem for $Q^n_{i+\frac{1}{2}}$ by using the Roe averaged state based on $Q^n_{rec,i}(1)$ and $Q^n_{rec,i+1}(0)$. This will also then be used to evaluate f' and f''. With this we also have the linearization around the interface value in order to solve the Riemann problems for $Q^n_{x,i+\frac{1}{2}}$ and $Q^n_{xx,i+\frac{1}{2}}$.

To study the accuracy we consider the initial values

$$\begin{aligned}
\rho(x,0) = p(x,0) &= 1 + \frac{1}{2}\exp\left(-80\left(x-\frac{1}{2}\right)^2\right) \\
v(x,0) &= 0
\end{aligned} \tag{4.50}$$

on the interval $[0,1]$ with periodic boundary conditions. The time steps satisfy $\nu \leq 0.9$. The solution structure at time $T = 0.25$ can be seen in Figure 4.4. The convergence study in table 4.3 shows the desired third order accuracy in the density. The other components are also approximated to third order.

N	L_1-error in density	EOC
32	$2.22371 \cdot 10^{-4}$	
64	$2.76821 \cdot 10^{-5}$	3.01
128	$3.55443 \cdot 10^{-6}$	2.96
256	$4.58017 \cdot 10^{-7}$	2.96
512	$5.83485 \cdot 10^{-8}$	2.97

Table 4.3.: Convergence study for Euler equations (4.46) with initial data (4.50) at time $T = 0.25$ using a reference solution with 4096 grid cells.

To study the behavior for problems that need limiting we study Sod's shock tube problem, i.e., we use the following initial values on the domain $[0,1]$:

$$(\rho, v, p)(x,0) = \begin{cases} (1,0,1) & : \ x < \frac{1}{2} \\ (0.125,0,0.1) & : \ x \geq \frac{1}{2} \end{cases} \tag{4.51}$$

Numerical approximations at final time $T = 0.17$ on a grid with 400 grid cells are

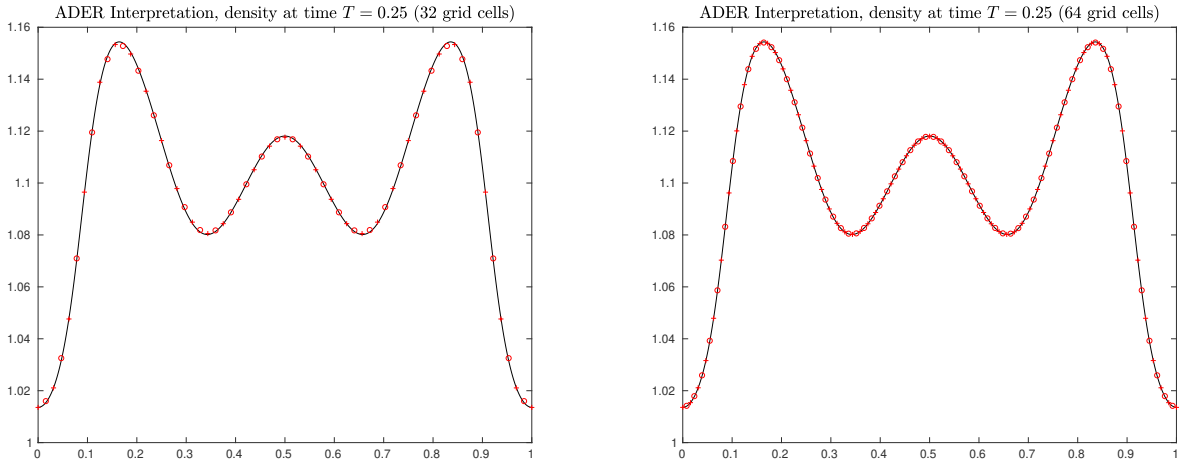

Figure 4.4.: Approximation of the Euler equations (4.46) with initial data (4.50) at time $T = 0.25$ using the ADER version of the Active Flux method. The solid line is a highly resolved reference solution computed on a grid with 4096 grid cells. Point values are marked with a '+', while cell averages are marked with an 'o'. We show results for density using 32 (left) and 64 (right) grid cells.

plotted next to a highly resolved reference solution in Figure 4.5. We compare the numerical solution with results of the third order accurate version of a one-step ADER finite volume method with space-time DG predictor that was kindly provided to us by Michael Dumbser.

As a second test problem we use the well known Shu-Osher test [47], i.e., the initial conditions read

$$(\rho, v, p)(x, 0) = \begin{cases} (3.857143, 2.629369, 10.3333) & : \quad x < -4 \\ (1 + 0.2\sin(5x), 0, 1) & : \quad x \geq -4. \end{cases} \tag{4.52}$$

In Figure 4.6 we plot the numerical results for the density at time $T = 1.8$, which were computed on grids of sizes 200, 300 and 400 cells, again compared to the ADER-DG method.

Our goal is to use the unlimited ADER version of the Active Flux method in as many grid cells as possible. For all components of the conserved quantities we use the same type of reconstruction, i.e., either the standard continuous, piecewise quadratic reconstruction, or the on both ends discontinuous, piecewise quadratic reconstruction (2.61). For the Euler equations, most of the structure is seen in density. Therefore, this quantity decides which reconstruction is used, i.e., the value of θ.

The Shu-Osher test confirms that our limited version of the Active Flux method can capture both the small-scale smooth flow features as well as the shock wave. Furthermore, it compares well with the ADER-DG method of Dumbser and Toro. The unlimited version of the Active Flux method is unstable for the two test problems considered in this section, while we found that the limited version is stable for time steps which satisfy the inequality $\nu \leq 0.5$. This is a small restriction compared to the limiting used in [1], but no additional, different reconstruction has to be used near

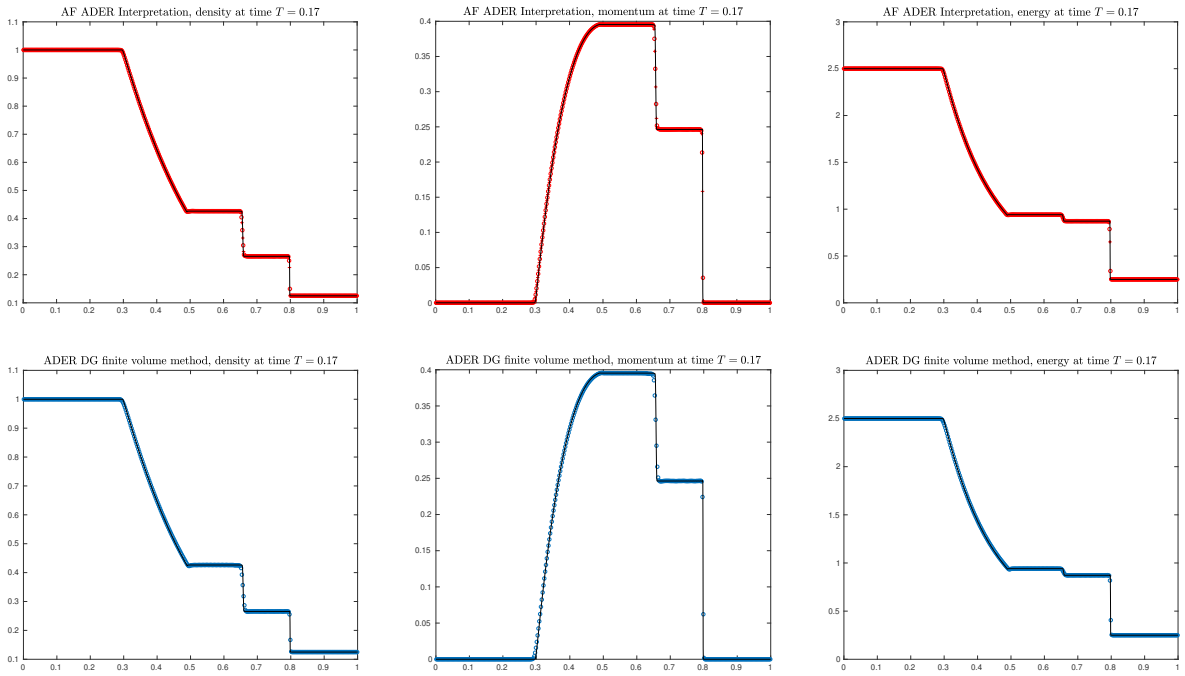

Figure 4.5.: Results for Sod's shock tube problem.
Top row: ADER interpretation of the Active Flux method using the discontinuous, limited, piecewise quadratic reconstruction if needed. Point values are marked with a '+', while cell averages are marked with an 'o'. The solution is computed using 400 grid cells. Time steps correspond to $\nu \leq 0.5$.
Bottom row: ADER-DG finite volume method of Dumbser and Toro with 400 grid cells.
The solid line is a reference solution, computed using 2000 grid cells.

shock waves.

4.2.2 Multidimensional Problems

After having explained the ADER interpretation in the one-dimensional case we now consider multidimensional problems. We now look at equation (1.1) for the two-dimensional case. The choice of degrees of freedom and the reconstruction are inherited from the previous chapter. We examine the same approach for the two-dimensional variants of the advection equation, the linear acoustic equations and the Euler equations.

4.2.2.1 Linear Advection Equation

This section is adapted from [1, Section 6.1]. We consider the equation

$$q_t + aq_x + bq_y = 0. \tag{4.53}$$

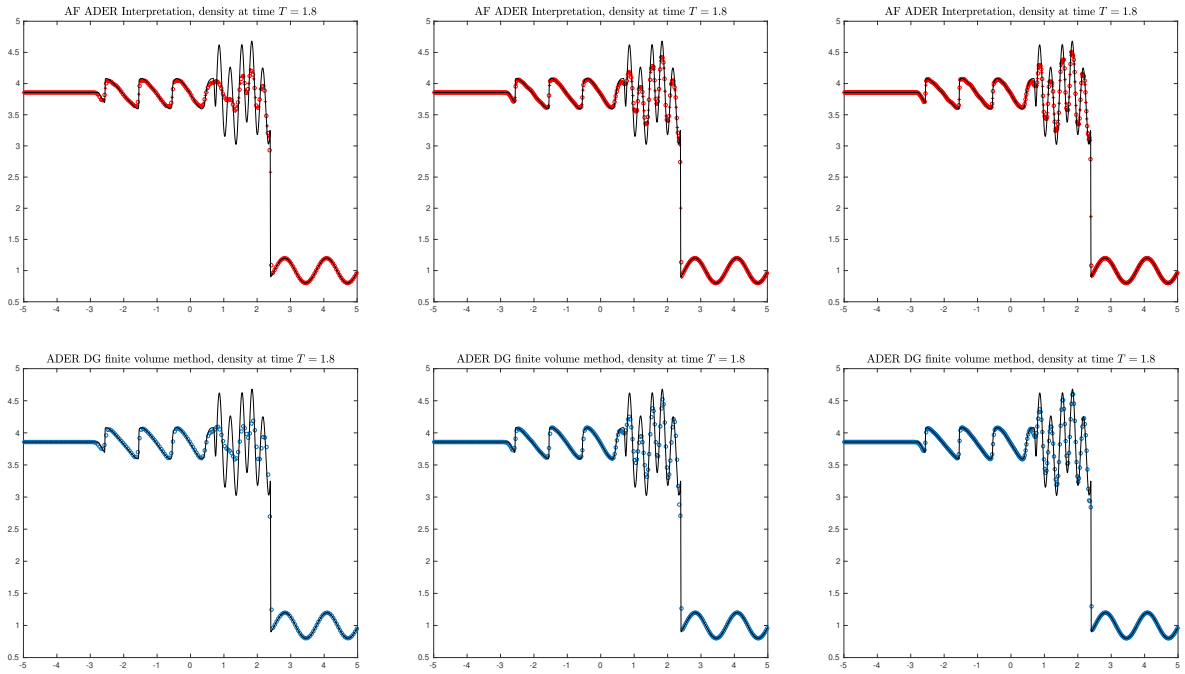

Figure 4.6.: Results for the Shu-Osher test problem.
Top row: ADER interpretation of the Active Flux method using the discontinuous, limited, piecewise quadratic reconstruction if needed. Point values are marked with a '+', while cell averages are marked with an 'o'. Time steps correspond to $\nu \leq 0.5$.
Bottom row: ADER-DG finite volume method of Dumbser and Toro.
For both methods, we use 200 (left), 300 (center) and 400 (right) grid cells. The solid line is a reference solution, which was obtained using 2000 grid cells.

We obtain

$$q_t = -aq_x - bq_y, \tag{4.54}$$
$$q_{tt} = a^2 q_{xx} + abq_{xy} + baq_{yx} + b^2 q_{yy}. \tag{4.55}$$

The Taylor series expansion in space (4.17) at a corner $(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}})$ on the interface now has the following form:

$$\begin{aligned} q(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t+\tau) \approx\ & q(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) + \tau \left(-aq_x(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) - bq_y(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) \right) \\ & + \frac{1}{2}\tau^2 \left(a^2 q_{xx}(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) + baq_{xy}(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) \right. \\ & \left. + abq_{yx}(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) + b^2 q_{yy}(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) \right) \end{aligned} \tag{4.56}$$

In the same way the analogue equations holds for the edge values. Riemann problems for all first and second order derivatives have to be solved at all corners and edges.

Let us first consider the edges: Since the reconstruction is globally continuous the derivatives with respect to x along a horizontal edge and the derivatives with respect to y along a vertical edge are uniquely defined. The values for $Q^n_{x,i,j\pm\frac{1}{2}}$ and $Q^n_{xx,i,j\pm\frac{1}{2}}$ and $Q^n_{y,i\pm\frac{1}{2},j}$ and $Q^n_{yy,i\pm\frac{1}{2},j}$ can be taken directly from the reconstruction. For the other derivatives we can solve a one-dimensional Riemann problem that is stated orthogonally to the edge, i.e., for a vertical edge $x_{i+\frac{1}{2}}$ we have a Riemann problem for $Q^n_{x,i+\frac{1}{2},j}$ of the form

$$\begin{aligned}&(Q_x)_t + a(Q_x)_x = 0\\ &Q_x(x,y,t_n) = \begin{cases} \frac{1}{\Delta x} Q^n_{rec,i,j,x}(1,0) & : x < x_{i+\frac{1}{2}} \\ \frac{1}{\Delta x} Q^n_{rec,i+1,j,x}(-1,0) & : x > x_{i+\frac{1}{2}} \end{cases}.\end{aligned} \tag{4.57}$$

We obtain the solution

$$Q^n_{x,i+\frac{1}{2},j} = \begin{cases} \frac{1}{\Delta x} Q^n_{rec,i,j,x}(1,0) & : a > 0, \\ \frac{1}{\Delta x} Q^n_{rec,i+1,j,x}(-1,0) & : a < 0. \end{cases} \tag{4.58}$$

Analog Riemann problems can be established for all other derivatives and edges.

For the corners so-called four-quadrant Riemann problems occur. Here, four values collide at a common corner. Because of the continuity of the reconstruction two times two of the values coincide for non mixed derivatives, respectively. We are then left with two values that can be used for a regular one-dimensional Riemann problem. For the mixed derivatives this does not hold in general. At the corner $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$ we then obtain a full four-quadrant Riemann problem of the following form:

$$\begin{aligned}&(Q_{xy})_t + a(Q_{xy})_x + b(Q_{xy})_y = 0\\ &Q_{xy}(x,y,t_n) = \begin{cases} \frac{1}{\Delta x \Delta y} Q^n_{rec,i,j,xy}(1,1) & : x < x_{i+\frac{1}{2}}, y < y_{j+\frac{1}{2}} \\ \frac{1}{\Delta x \Delta y} Q^n_{rec,i+1,j,xy}(-1,1) & : x > x_{i+\frac{1}{2}}, y < y_{j+\frac{1}{2}} \\ \frac{1}{\Delta x \Delta y} Q^n_{rec,i,j+1,xy}(1,-1) & : x < x_{i+\frac{1}{2}}, y > y_{j+\frac{1}{2}} \\ \frac{1}{\Delta x \Delta y} Q^n_{rec,i+1,j+1,xy}(-1,-1) & : x > x_{i+\frac{1}{2}}, y > y_{j+\frac{1}{2}} \end{cases}\end{aligned} \tag{4.59}$$

We can solve this Riemann problem for the advection equation. The correct solution is again the value in upwind direction:

$$Q^n_{xy,i+\frac{1}{2},j+\frac{1}{2}} = \begin{cases} \frac{1}{\Delta x \Delta y} Q^n_{rec,i,j,xy}(1,1) & : a > 0, b > 0 \\ \frac{1}{\Delta x \Delta y} Q^n_{rec,i+1,j,xy}(-1,1) & : a < 0, b > 0 \\ \frac{1}{\Delta x \Delta y} Q^n_{rec,i,j+1,xy}(1,-1) & : a > 0, b < 0 \\ \frac{1}{\Delta x \Delta y} Q^n_{rec,i+1,j+1,xy}(-1,-1) & : a < 0, b < 0 \end{cases} \tag{4.60}$$

The same value is obtained for $Q^n_{yx,i+\frac{1}{2},j+\frac{1}{2}}$.

In contrast to its one-dimensional form, the resulting method is no longer equivalent to the two-dimensional Active Flux method on Cartesian grids since the mixed

derivatives up to order 4 don't vanish in the two-dimensional case. If we use the terms q_{ttt} and q_{tttt} in the expansion we once again obtain an equivalent method. The high corresponding high order derivatives at the corners can be computed in the same way as just described for Q_{xy}. For third order an expansion up to q_{tt} is sufficient, however. We call the method that uses all higher order terms **ADER full** and the method that uses only the terms up to third order accuracy **ADER reduced**.

During numerical test computations it stands out that ADER reduced needs a reduced time step. We conjecture a bound by $\nu \leq 0.5$, while ν is given by (3.19). We compare both approaches to study their accuracy. Let $a = 1$, $b = 0.7$ and let

$$q_0(x, y) = \sin(4\pi(x + y)). \tag{4.61}$$

We compare the error to the exact solution at time $T = 1$ on different grids. The results can be seen in Figure 4.7. Both variants yield third order accuracy, the inclusion of the higher derivatives improves the accuracy slightly. The use of a higher CFL number further improves the accuracy.

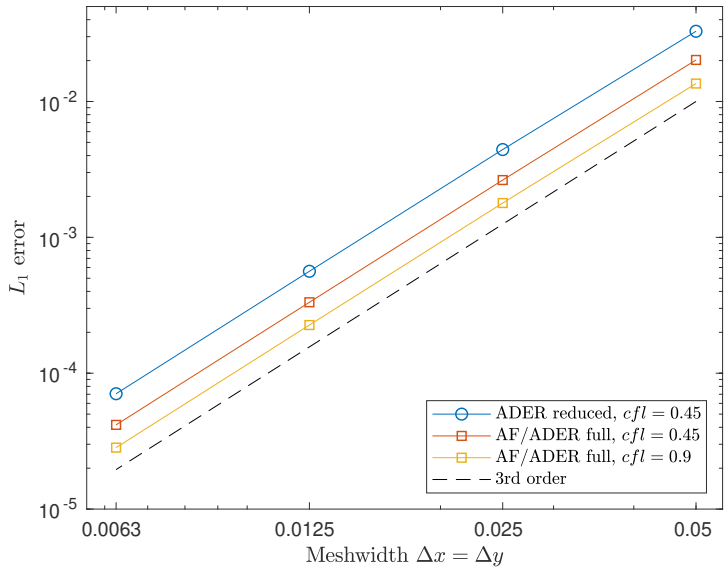

Figure 4.7.: Accuracy study for the two dimensional advection problem. The yellow curve (Active Flux / ADER full) shows the error vs. mesh if the exact evolution formula with $\nu \leq 0.9$ is used to update the interface values. For advection, the ADER method which uses all nonzero derivative terms for the update of the interface values is equivalent to using the exact evolution formula. The blue curve shows the error for the ADER update with time steps chosen such that $\nu \leq 0.45$ and using only those derivative values that are necessary in order to obtain third order. The red curve shows the error of the method that uses exact evolution of the interface values and time steps according to $\nu \leq 0.45$.

4.2.2.2 Linear Acoustic Equations

This section is based on [1, Section 6.2] and explained in more detail. The approach to approximate the solution to the linear acoustic equations (3.20) differs only slightly from the way the advection equation is treated. We rewrite the system in the form

$$q_t(x,y,t) + Aq_x(x,y,t) + Bq_y(x,y,t) = 0 \tag{4.62}$$

with

$$A = \begin{pmatrix} 0 & c & 0 \\ c & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} 0 & 0 & c \\ 0 & 0 & 0 \\ c & 0 & 0 \end{pmatrix}. \tag{4.63}$$

For the expansion, we obtain:

$$\begin{aligned} q(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t+\tau) \approx\ & q(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) + \tau \left(-Aq_x(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) - Bq_y(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) \right) \\ & + \frac{1}{2}\tau^2 \left(A^2 q_{xx}(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) + BAq_{xy}(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) \right. \\ & \left. + ABq_{yx}(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) + B^2 q_{yy}(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}, t) \right) \end{aligned} \tag{4.64}$$

Additional terms are present if higher order terms are incorporated for ADER full. The Riemann problems are solved in a similar fashion as in the advection equation. It is however important to distinguish between the AB and BA terms since the matrices A and B don't commute. Let us consider a particular Riemann problem for Q_{xy} at a vertical edge: If we solve the Riemann problem (4.57) (with matrix A instead of the velocity a) and multiply the result with AB as demanded, we lose information from the kernel of B. We therefore multiply with B before solving the Riemann problem and then solve:

$$\begin{aligned} & (Q_{xy})_t + A(Q_{xy})_x = 0 \\ & Q_{xy}(x,y,t_n) = \begin{cases} \frac{1}{\Delta x \Delta y} BQ^n_{rec,i,j,xy}(1,0) & : x < x_{i+\frac{1}{2}} \\ \frac{1}{\Delta x \Delta y} BQ^n_{rec,i+1,j,xy}(-1,0) & : x > x_{i+\frac{1}{2}} \end{cases} \end{aligned} \tag{4.65}$$

The solution is subsequently multiplied with A, not by AB. Alternatively, one can also multiply the data with AB before solving the Riemann problem as the solution to the Riemann problem is invariant against multiplications with A. In the same way we proceed with Q_{yx} at a horizontal edge. This procedure results in the correct values.

Since the four-quadrant Riemann problem for the linear acoustic equations does not have a simple solution, we again use a splitting approach. We first solve the two following Riemann problems for Q_{xy}:

$$\begin{aligned} & (Q^{\star_1}_{xy})_t + A(Q^{\star_1}_{xy})_x = 0 \\ & Q^{\star_1}_{xy}(x,y,t_n) = \begin{cases} \frac{1}{\Delta x \Delta y} BQ^n_{rec,i,j,xy}(1,1) & : x < x_{i+\frac{1}{2}} \\ \frac{1}{\Delta x \Delta y} BQ^n_{rec,i+1,j,xy}(-1,1) & : x > x_{i+\frac{1}{2}} \end{cases} \end{aligned} \tag{4.66}$$

and

$$\begin{aligned}(Q^{\star 2}_{xy})_t + A(Q^{\star 2}_{xy})_x &= 0\\ Q^{\star 2}_{xy}(x,y,t_n) &= \begin{cases} \frac{1}{\Delta x \Delta y} B Q^n_{rec,i,j+,xy}(1,-1) & : x < x_{i+\frac{1}{2}} \\ \frac{1}{\Delta x \Delta y} B Q^n_{rec,i+1,j+,xy}(-1,-1) & : x > x_{i+\frac{1}{2}} \end{cases}.\end{aligned} \tag{4.67}$$

The two solutions are then used for an additional Riemann problem alongside the other direction:

$$\begin{aligned}(Q_{xy})_t + B(Q_{xy})_x &= 0\\ Q_{xy}(x,y,t_n) &= \begin{cases} \frac{1}{\Delta x \Delta y} B Q^{\star 1}_{xy} & : y < y_{i+\frac{1}{2}} \\ \frac{1}{\Delta x \Delta y} B Q^{\star 2}_{xy} & : y > y_{i+\frac{1}{2}} \end{cases}\end{aligned} \tag{4.68}$$

This concept can be generalized for higher derivatives.

Remark 4.2.2. It turns out that Cartesian grids have one more advantage. In triangular grids, for example, arbitrarily many triangles can meet in any of the corners, leading to possibly many initial values that can influence the solution to the corresponding Riemann problem. Its solution is not more difficult to determine for the advection equation but the explained splitting approach is not applicable for general hyperbolic systems.

While we have seen the equivalence of the ADER interpretation to the Active Flux method for the advection equation even in the two-dimensional case, this is not the case for the linear acoustic equations. The given matrices are both diagonalizable as required but are not simultaneously diagonalizable. That's why no common characteristic quantities can be stated and the proof of theorem 4.2.1 doesn't work anymore. It can be shown that our procedure differs to the exact evolution 3.22 to third order accuracy only in a few terms of second order which are merely produced by the consecutive Riemann problems. These terms can not be represented by the solution of one-dimensional Riemann problems. The use of ADER full makes the errors in the higher derivatives to be solely determined by the errors in the multidimensional Riemann problems, too.

Remark 4.2.3. The permutation of Riemann problems and matrix multiplication is not necessary to obtain a method of third order accuracy. However, more terms of the truncation error are eliminated. The observant reader may have noticed that a similar case was already present in the one-dimensional Euler equations: There, the operators $f'(q)$ and $f''(q)$ don't commute, so that for the term $f'(Q^n_{i+\frac{1}{2}})f''(Q^n_{i+\frac{1}{2}}) \cdot (Q^n_{x,i+\frac{1}{2}}, Q^n_{x,i+\frac{1}{2}})$ a multiplication with $f''(Q^n_{i+\frac{1}{2}})$ before solving the Riemann problem would be possible, which would then be defined as follows:

$$\begin{aligned}(Q_{xx})_t + f'(Q^n_{i+\frac{1}{2}})(Q_{xx})_x &= 0\\ Q_{xx}(x,t_n) &= \begin{cases} \frac{1}{\Delta x \Delta x} f''(Q^n_{i+\frac{1}{2}}) \cdot (Q^n_{rec,i,j,xx}(1), Q^n_{rec,i,j,xx}(1)) & : x < x_{i+\frac{1}{2}} \\ \frac{1}{\Delta x \Delta x} f''(Q^n_{i+\frac{1}{2}}) \cdot (Q^n_{rec,i+1,j,xx}(0), Q^n_{rec,i+1,j,xx}(0)) & : x > x_{i+\frac{1}{2}} \end{cases}\end{aligned} \tag{4.69}$$

In the shown computations this has not been used.

To study the accuracy we consider the periodic initial values

$$\begin{aligned} p_0(x,y) &= -\frac{1}{c_0}\left(\sin(2\pi x) + \sin(2\pi y)\right) \\ u_0(x,y) &= 0 \\ v_0(x,y) &= 0 \end{aligned} \tag{4.70}$$

on the domain $[-1,1]\times[-1,1]$ and final time $T = 0.2$. This test problem was suggested by Lukáčová et al. [48], where the exact solution can also be found. We compare the Active Flux method with ADER full as well as ADER reduced. Figure 4.8 shows the error against the mesh width. Since the problem and the initial values are both symmetrical in the velocities u and v, their errors are of equal size. Here, we also notice a reduced stability of $\nu \leq 0.25$ for

$$\nu = \frac{c_o \Delta t}{\min\{\Delta x, \Delta y\}}. \tag{4.71}$$

While the error in the velocity components is almost the same for all methods, the ADER variants give a small improvement to the accuracy in the pressure for this test problem compared to the Active Flux method. The third order accuracy is clearly visible for all methods.

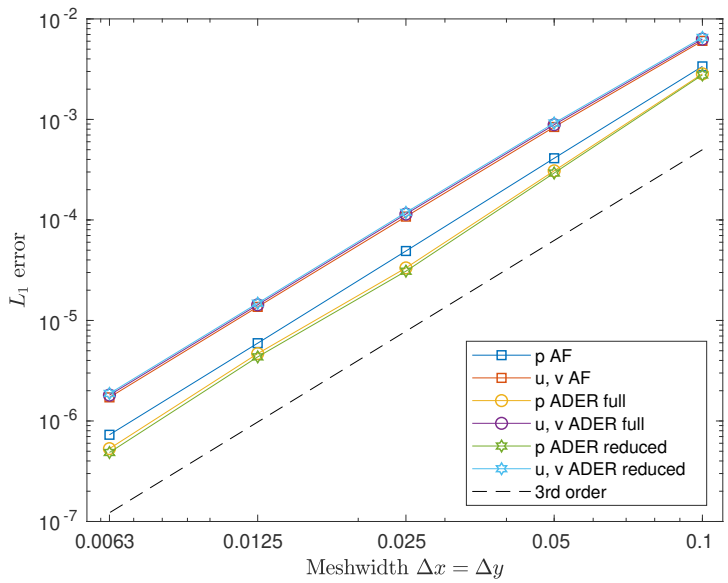

Figure 4.8.: Accuracy study for the two-dimensional acoustic equations. We compare results of our ADER approach with results where the exact evolution formula was used to update the interface values.

An important test for Roe and coauthors is the check for radial symmetry for fitting

intial values [28]. For this, we consider

$$\begin{aligned} p_0(x,y) &= 1 + \exp(-\mu((x-x_0)^2 + (y-y_0)^2)) \\ u_0(x,y) &= v_0(x,y) = 0 \end{aligned} \tag{4.72}$$

on the domain $[-2,2]\times[-2,2]$ with $x_0 = y_0 = 0$ and $\mu = 50$. We study the solutions at final time $T = 1.25$ on a 50×50 and a 150×150 grid by plotting each interface value against the radius $r = \sqrt{(x-x_0)^2 + (y-y_0)^2}$. As a reference we also draw the exact solution as a black line. Figures 4.9 and 4.10 show the resulting plots. Both the Active Flux method and the ADER interpretations show a good symmetry. In comparison to the original Active Flux method of Roe et al. on triangular grids, both methods perform equally well (compare with [28]).

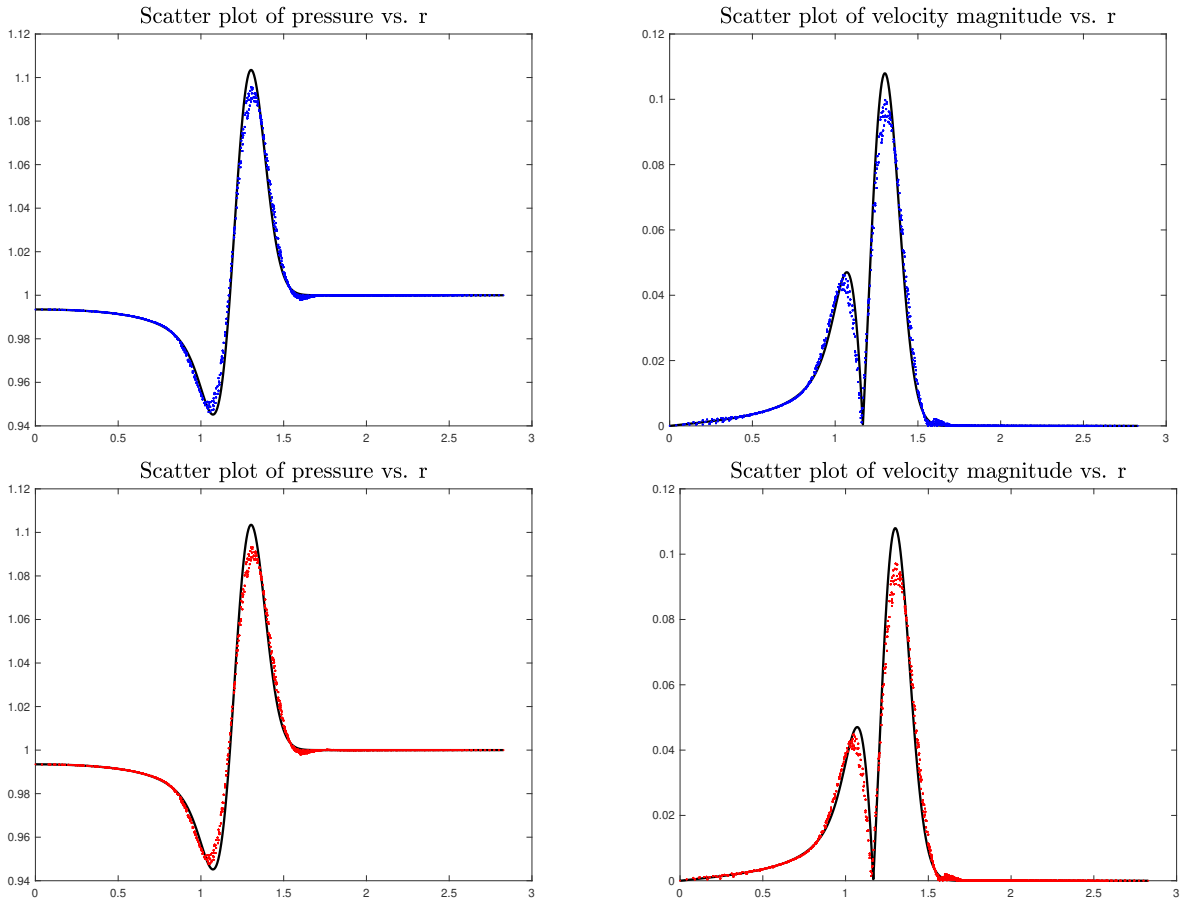

Figure 4.9.: Scatter plots of the magnitude of pressure and velocity obtained using two different methods on a 50×50 grid. The blue dots (first row) show the results obtained by using the exact evolution formula for the update of the edge values, the red dots (second row) indicate the results obtained using the ADER update formula. The black line is the exact solution.

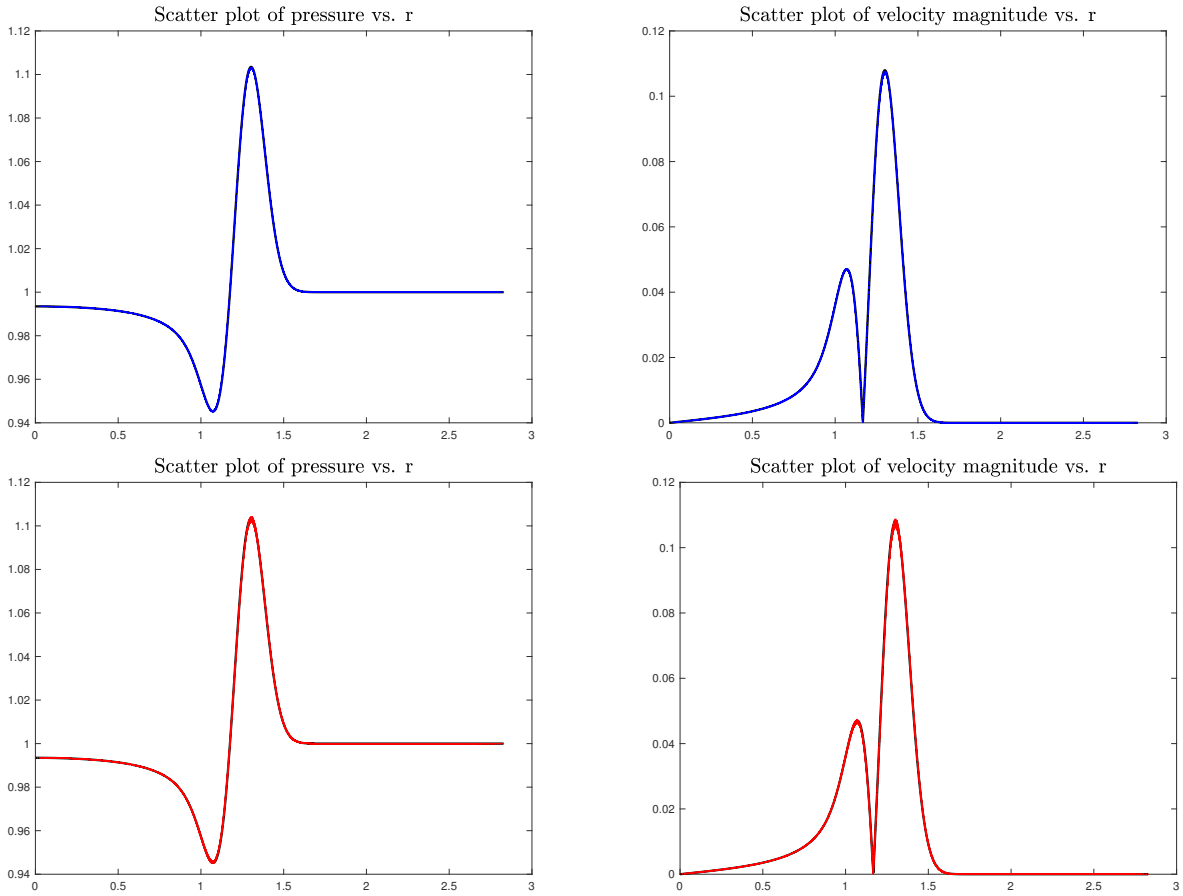

Figure 4.10.: Scatter plots of the magnitude of pressure and velocity obtained using two different methods on a 150×150 grid. The blue dots (first row) show the results obtained by using the exact evolution formula for the update of the edge values, the red dots (second row) indicate the results obtained using the ADER update formula. The black line is the exact solution.

4.2.2.3 Euler Equations

As a common representative for nonlinear equation we here consider the two-dimensional Euler equations which have the following form:

$$q_t + f(q)_x + g(q)_y = 0 \tag{4.73}$$

with

$$q = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix}, \quad f(q) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(E+p) \end{pmatrix}, \quad g(q) = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(E+p) \end{pmatrix}. \tag{4.74}$$

Again, ρ, u, v, p and E denote density, velocity components in x and y direction, pressure and energy. We again use the ideal gas equation

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho(u^2 + v^2) \tag{4.75}$$

with adiabatic exponent $\gamma = 1.4$.

The derivatives f', f'', g' and g'' of the flux functions are similar to the ones in the one-dimensional case. We refrain from stating them at this point. For the temporal expansion the following equations arise:

$$q_t = -f'(q)q_x - g'(q)q_y \tag{4.76}$$
$$q_{tx} = -(f''(q) \cdot (q_x, q_x) + f'(q)q_{xx}) - (g''(q) \cdot (q_y, q_x) + g'(q)q_{yx}) \tag{4.77}$$
$$q_{ty} = -(f''(q) \cdot (q_x, q_y) + f'(q)q_{xy}) - (g''(q) \cdot (q_y, q_y) + g'(q)q_{yy}) \tag{4.78}$$
$$q_{tt} = -(f''(q) \cdot (q_t, q_x) + f'(q)q_{tx}) - (g''(q) \cdot (q_t, q_y) + g'(q)q_{ty}) \tag{4.79}$$

The now appearing second order derivatives are increasingly more expensive in the computations and multiplications due to the increasing dimension of the operator. In any case one should stay away from using higher order terms like they are used in ADER full to keep the effort as low as possible.

Like in the acoustic equations we also multiply with the respective derivatives before solving the Riemann problems.

Remark 4.2.4. While in the acoustic equations only one term was affected by this change (in ADER reduced), any terms that contain multiplications of different derivatives are affected here. These are all terms from the combined q_{tt} except $f'(q)q_{xx}$ and $g'(q)q_{yy}$. It is also not required to perform these changes to obtain a third order accurate method but we still reduces the error by a small amount.

Let now $r = \sqrt{x^2 + y^2}$ and $\epsilon = 5$. We perform an accuracy study on the initial values

$$\begin{aligned}
\rho(x, y, 0) &= (1 - (\gamma - 1)\frac{\epsilon^2}{8\gamma\pi^2}\exp(1 - r^2))^{\frac{1}{\gamma-1}} \\
u(x, y, 0) &= 1 - \frac{\epsilon y}{2\pi}\exp\left(\frac{1}{2}(1 - r^2)\right) \\
v(x, y, 0) &= 1 + \frac{\epsilon x}{2\pi}\exp\left(\frac{1}{2}(1 - r^2)\right) \\
E(x, y, 0) &= \frac{1}{\gamma - 1}\rho(x, y, 0)^\gamma + \frac{1}{2}(\rho(x, y, 0)u(x, y, 0)^2 + \rho(x, y, 0)v(x, y, 0)^2)
\end{aligned} \tag{4.80}$$

on the domain $[-5, 5] \times [-5, 5]$ with double periodic boundary conditions. These initial values are also known as a vortex evolution [47]. The exact solution at time $T = 10$ matches the initial conditions. In our computations we measure that the approximate choice $\nu \leq 0.3$ leads to a stable method. Higher values lead to an unstable method. Thus, we also observe a reduced stability here. Table 4.4 shows the measured error as well as the estimated order of convergence. When resolving the solution well enough,

	ρ		ρu		ρv		E	
N^2	L_∞-error	EOC	L_∞-error	EOC	L_∞-error	EOC	L_∞-error	EOC
20^2	$3.3005 \cdot 10^{-2}$		$6.0654 \cdot 10^{-2}$		$5.5441 \cdot 10^{-2}$		$2.0167 \cdot 10^{-1}$	
40^2	$6.5423 \cdot 10^{-3}$	2.33	$1.5654 \cdot 10^{-2}$	1.95	$1.4179 \cdot 10^{-2}$	1.97	$4.0841 \cdot 10^{-2}$	2.30
80^2	$9.7585 \cdot 10^{-4}$	2.75	$2.4310 \cdot 10^{-3}$	2.69	$2.2036 \cdot 10^{-3}$	2.69	$5.9688 \cdot 10^{-3}$	2.77
160^2	$1.3284 \cdot 10^{-4}$	2.88	$3.2980 \cdot 10^{-4}$	2.88	$2.9666 \cdot 10^{-4}$	2.89	$7.5699 \cdot 10^{-4}$	2.98

Table 4.4.: Convergence study for the Euler equations (4.74) with initial data (4.80) at time $T = 10$, using the ADER interpretation of the Active Flux method.

third order accuracy is achieved. Figure 4.11 shows the density in a profile view.

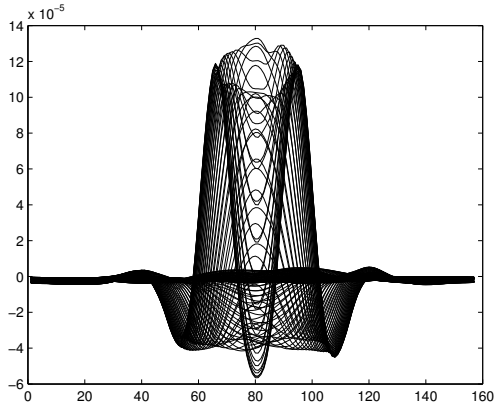

Figure 4.11.: Error profile of the density for 160×160 cells at time $T = 10$ using the ADER interpretation of the Active Flux method.

4.3 Summary

We have introduced a possibility to extend the Active Flux method to nonlinear hyperbolic equations in multiple space dimensions, the ADER interpretation of the Active Flux method. We continue to use point values of the unknowns on the boundary of each cell and compute the flux actively from the point values at later times. Only the use of an approximate evolution operator changes the method, which is based on the original idea of ADER methods. We obtain a third order method that shows convincing results for all of the considered problems and can be used on any nonlinear hyperbolic system of equations.

While the method is equivalent to the Active Flux method for one-dimensional, linear hyperbolic systems, this is no longer necessarily the case for two-dimensional systems. An exact solution to the four-quadrant Riemann problem and the use of ADER full would again result in equivalence, but the general solution to these Riemann problems is cumbersome [49]. Our splitting approach to solve the multidimensional Riemann problem eliminates this difficulty.

Barsukow et al. have shown that the Active Flux method for the linear acoustic equations on Cartesian grids is stationary preserving [42]. This property is lost using the ADER ansatz. Instead, we have constructed a method that does not rely on complicated or expensive evolution operators and can be used for any hyperbolic conservation law in multiple dimensions where exact evolution operators are not known. The important stability properties of the Active Flux method using exact integration are lost when using the ADER approach. The evolution uses a Taylor series expansion that yields extrapolated values if used in a cut cell. This makes the ADER approach unstable for cut cells so that it cannot be used for cut cell grids in this form.

5 Conclusions and Outlook

In the first part, we developed and investigated a new finite volume method for cut cell meshes that is based on the Active Flux method in one and two spatial dimensions. It used not only cell average values but also point values on the interfaces between the cells. The method was so far valid only for linear systems where an exact evolution operator can be found.

Firstly, we saw that the third order accuracy of the one-dimensional method can be maintained in the presence of a small cell if exact integration is used instead of Simpson's rule for the small cell. Nevertheless, the use of Simpson's rule reduced the order by one only in the cut cell, leading to a stable and overall third order scheme, measured in the L_1-norm. The cancellation property was achieved in both cases due to the continuous reconstruction. Limiting could be done in various ways in one dimension, but only one of the proposed strategies could be extended to the multidimensional case.

Secondly, we transformed the original Active Flux method on triangular grids to Cartesian grids for two dimensions. We found a slight reduction in stability when using Simpson's rule for the flux computation for the linear advection equation, while exact integration led to maximal stability. Since two-dimensional cut cells differentiate strongly from the Cartesian cells, additional effort had to be put into reconstruction and flux computation in these cut cells. Since we didn't find a stable way to compute a globally continuous reconstruction, a combined interpolation/least square ansatz was pursued. The resulting method was third order accurate in the L_1-norm, while we saw a slight reduction of accuracy near the boundary in the general case. The scheme was stable for time steps that depend on the regular grid size for all considered test problems, although the reconstruction could have small discontinuities across cut cell faces. Limiting could be applied if needed.

In the second part, we developed an ADER finite volume method that is equivalent to the Active Flux method for linear one-dimensional systems. It expanded the AF method for any hyperbolic conservation law, without the need of an exact evolution operator. Different examples and comparisons to previous methods were given and showed excellent accuracy. The method showed very good numerical stability on Cartesian grids and didn't suffer from a severe time step restriction. However, the method could not be used for cut cell grids in its current form due to a lack of stability for small cells.

Looking forward, there are many interesting topics that can be studied in the future:

All has not been said and done when it comes to the reconstruction in the cut cells. While the current approach has shown to be stable in all of our experiments, there is

no proof. Through a globally continuous reconstruction this might be possible, but one has yet to find such reconstruction without making the method too expensive.

The loss of accuracy near the boundary in many cut cell situations can possibly be improved. Through higher order reconstruction or filtering a full third order method in the L_∞-norm could be constructed.

One would like to be able to compute more than the advection equation on cut cell meshes. This might be doable by an extension of our ADER interpretation of the Active Flux method by the introduction of h-boxes or other locally defined helping grids. Alternatively, other ways of finding an approximate evolution operator can be tried. The first interesting problems to study could be advection with spatially varying velocity or linear acoustics on more complex geometries.

With that, one interesting question is how the boundary should be approximated by the cut cells. Besides straight line approximations one could also consider curved approximations or level-set functions. The inclusion of boundary conditions immediately falls into line as well.

As there are many interesting research problems to be solved, the Active Flux method for cut cells is a good topic for future research. The foundations on how to construct a cut cell Active Flux method have been laid in this thesis.

Appendices

A Derivation of the update matrix A for the linear stability analysis

In Section 3.1.4 we discussed the derivation of the update matrix A for the linear stability analysis. Here, we want to present the construction of the sub-matrices $Z_{k,l}$ for the advection equation in more detail and provide extended results for the eigenvalues. This Appendix is adapted from [2, Appendix A]. Recall that $Z_{k,l}[x,y]$ with $k,l \in \{-1,0,1,2\}$ and $x,y \in \{1,2,3,4\}$ represents the contribution of the DoF y in cell $C_{i+k,j+l}$ to the update of the DoF x in cell $C_{i,j}$.

Remember that the reconstruction in cell $C_{i,j}$ at time t_n can be expressed in the form

$$Q^n_{rec,i,j}(\xi,\eta) = \sum_{i=1}^{9} c_i N_i(\xi,\eta) \tag{A.1}$$

(compare to (3.14)) with c_i and N_i as described in Table 3.1 and $(\xi,\eta) \in [-1,1]\times[-1,1]$. This reconstruction interpolates the point values along the boundary and preserves the cell average $Q^n_{i,j}$. For our considerations it is more convenient to express the reconstruction in the form

$$Q^n_{rec,i,j}(\xi,\eta) = \sum_{i=1}^{9} \hat{i}_k \hat{N}_i(\xi,\eta), \tag{A.2}$$

using the DoF of the method as coefficients. Thus, we set

$$\begin{aligned}
\hat{N}_i(\xi,\eta) &:= N_i(\xi,\eta) - \frac{1}{16} N_i(\xi,\eta), && i = 1,3,5,7,\\
\hat{N}_i(\xi,\eta) &:= N_i(\xi,\eta) - \frac{1}{4} N_i(\xi,\eta), && i = 2,4,6,8,\\
\hat{N}_9(\xi,\eta) &:= \frac{9}{4} N_9(\xi,\eta) && \\
\hat{c}_i &:= c_i, && i = 1,\dots,8,\\
\hat{c}_9 &:= Q^n_{i,j}. &&
\end{aligned}$$

In our Python code we distinguish between four different cases which are specified in Table A.1.

The update of point values of the conserved quantities can now easily be expressed by appropriate entries of different Z matrices. In order to describe the update of the

case number	estimates
1	$\lvert a\rvert\Delta t/\Delta x \leq 0.5$, $\lvert b\rvert\Delta t/\Delta y \leq 0.5$
2	$\lvert a\rvert\Delta t/\Delta x \geq 0.5$, $\lvert b\rvert\Delta t/\Delta y \leq 0.5$
3	$\lvert a\rvert\Delta t/\Delta x \leq 0.5$, $\lvert b\rvert\Delta t/\Delta y \geq 0.5$
4	$\lvert a\rvert\Delta t/\Delta x \geq 0.5$, $\lvert b\rvert\Delta t/\Delta y \geq 0.5$

Table A.1.: Different cases which are considered in order to define the matrix A for the two-dimensional advection equation.

cell average for the AF method, we rewrite the finite volume update formula (3.13) in a sum of the form

$$Q_{i,j}^{n+1} = \sum_{s,t} m_{s,t} Q_{i+s,j+t}^{n},$$

where we consider all relevant DoF that contribute to the update. The precise form of $m_{s,t}$ depends of the specific choice of the method, i.e., Simpson's rule or exact integration. As an example, we consider $a, b > 0$ in a case 2 situation. Then matrix $Z_{-1,0}$, for example, has the form

$$\begin{pmatrix} \hat{N}_8(\xi_1,\eta_1) & \hat{N}_1(\xi_1,\eta_1) & \hat{N}_2(\xi_1,\eta_1) & \hat{N}_9(\xi_1,\eta_1) \\ 0 & N_7(\xi_2,\eta_2) & N_6(\xi_2,\eta_2) & 0 \\ 0 & N_7(\xi_3,\eta_3) & N_6(\xi_3,\eta_3) & 0 \\ m_{i-\frac{3}{2},j} & m_{i-\frac{3}{2},j-\frac{1}{2}} & m_{i-1,j-\frac{1}{2}} & m_{i-1,j} \end{pmatrix} \tag{A.3}$$

with

$$\begin{aligned} (\xi_1,\eta_1) &= \left(1 - 2a\frac{\Delta t}{\Delta x}, -2b\frac{\Delta t}{\Delta y}\right), \\ (\xi_2,\eta_2) &= \left(1 - 2a\frac{\Delta t}{\Delta x}, 1 - 2b\frac{\Delta t}{\Delta y}\right), \\ (\xi_3,\eta_3) &= \left(2 - 2a\frac{\Delta t}{\Delta x}, 1 - 2b\frac{\Delta t}{\Delta y}\right). \end{aligned} \tag{A.4}$$

See also Figure A.1 for an illustration of the relevant points and DoF. Once the Z matrices are defined, we can compute A using (3.30). In Fig. A.2 we show the eigenvalues for Simpson's method as well as exact integration using $a = b$, a grid with 50×50 time steps, periodic boundary conditions and two different time steps. While the AF method with Simpson's rule is stable for $\nu = 0.75$ and unstable for $\nu = 0.9$, AF with exact integration is stable for both time steps (compare to the results in Section 3.1.4).

In the special case $a = b$ and $\nu = 1$, the AF method with Simpson's rule can be

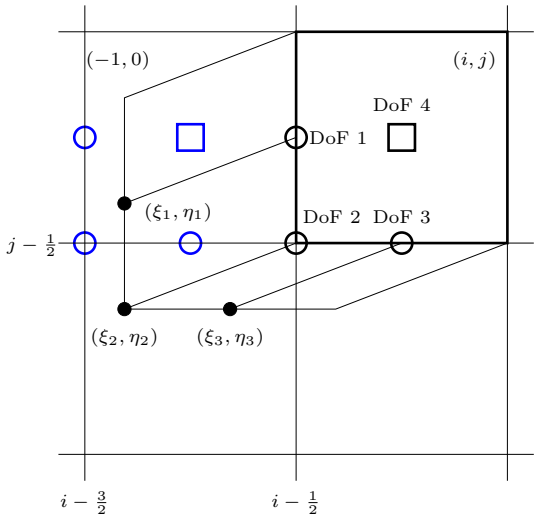

Figure A.1.: Depiction of the relevant DoF in a case 2 situation.

expressed using

$$Z_{-1,-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{18} & \frac{1}{24} & \frac{1}{18} & \frac{1}{2} \end{pmatrix}, Z_{-1,0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{72} & \frac{5}{18} & 0 \end{pmatrix}, Z_{0,-1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{5}{18} & -\frac{1}{72} & 0 & 0 \end{pmatrix}, \tag{A.5}$$

$$Z_{0,0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\frac{5}{18} & 0 & -\frac{5}{18} & \frac{1}{2} \end{pmatrix}, Z_{0,1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{22} & -\frac{1}{18} & 0 \end{pmatrix}, Z_{1,0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\frac{1}{18} & \frac{1}{72} & 0 & 0 \end{pmatrix}, \tag{A.6}$$

$$Z_{1,1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{24} & 0 & 0 \end{pmatrix} \tag{A.7}$$

$$\tag{A.8}$$

All other Z matrices are zero matrices. In Figure A.3 we show the eigenvalues for the AF method for this case using either Simpson's rule or exact integration for $N = 50$. The magnitude of all eigenvalues is bounded by one. However, Simpson's method is unstable in this situation.

In order to further analyse the stability, we need to investigate the algebraic and geometric multiplicity of the eigenvalues with magnitude one. For $N = 2$, i.e., $A \in \mathbb{R}^{16\times16}$, we can explicitly calculate the eigenvalues and eigenvectors. The matrix A has the eigenvalues $\lambda_{1,\dots,6} = -1, \lambda_{7,8} = 0, \lambda_{9,\dots,16} = 1$. There are only seven linearly independent eigenvectors which correspond to the eigenvalue 1, i.e., the geometric and

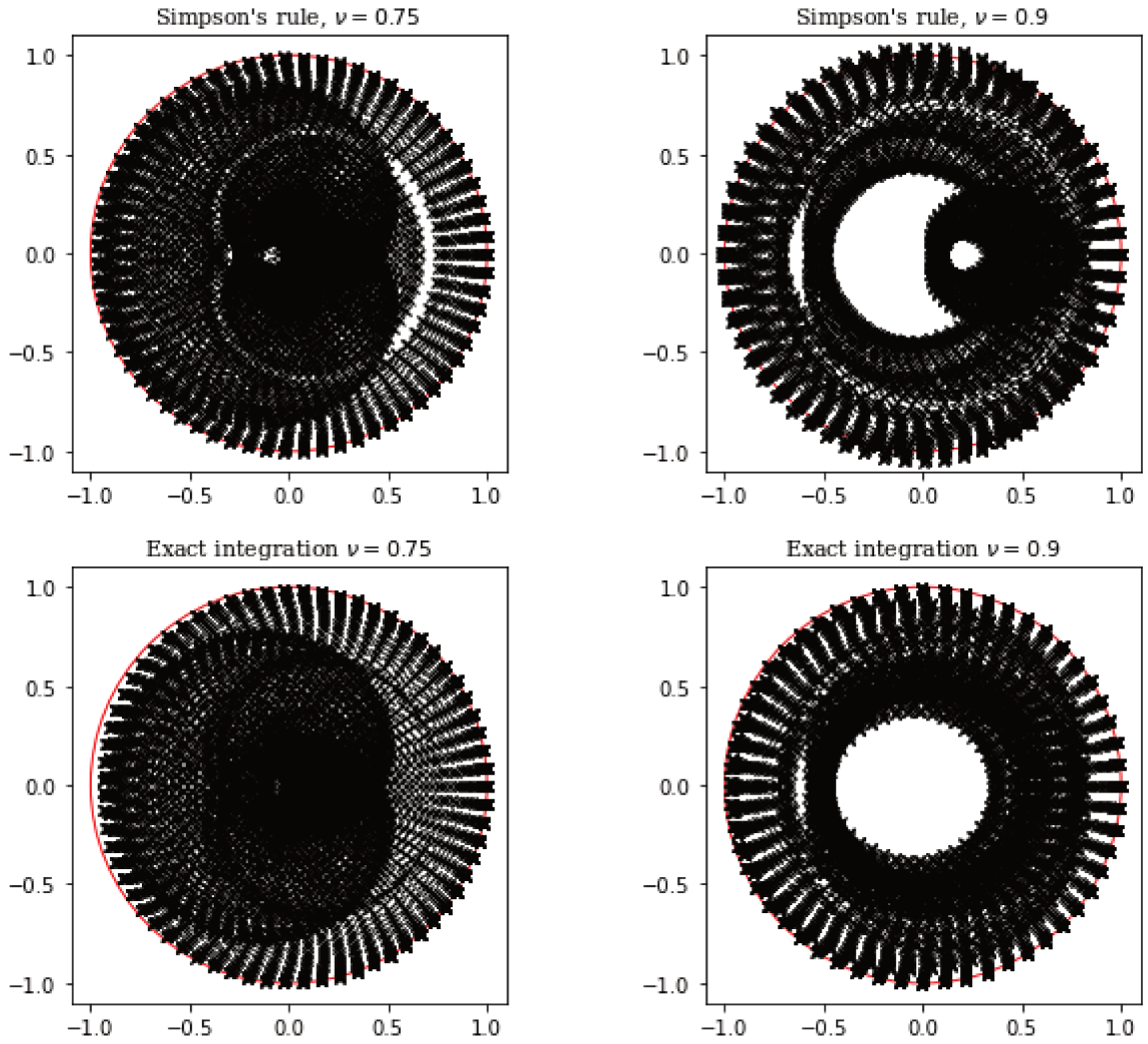

Figure A.2.: Eigenvalues for $a = b$, $\Delta x = \Delta y = 1/20$, $N = 50$, $\nu = 0.75$ and $\nu = 0.9$ using Simpson's rule (first row) and exact integration (second row) for the computation of the cell averaged values.

algebraic multiplicity doesn't match. Details can be found in the Python code. This explains the instability even for $a = b$, $\nu = 1$.

Note that while it is quite simple to present the matrices for advection, where each update coefficient for the point values requires only a single evaluation of one basis function, we omit the detailed description of the linear acoustics 12×12 matrices, where each entry has to be carefully calculated by the correct spherical mean integration. Again, this process can be followed in our Python code.

Python programs for the computation of the matrices A, both for advection as well as acoustics, can be downloaded from

`http://www.am.uni-duesseldorf.de/~kerkmann/Forschung/`

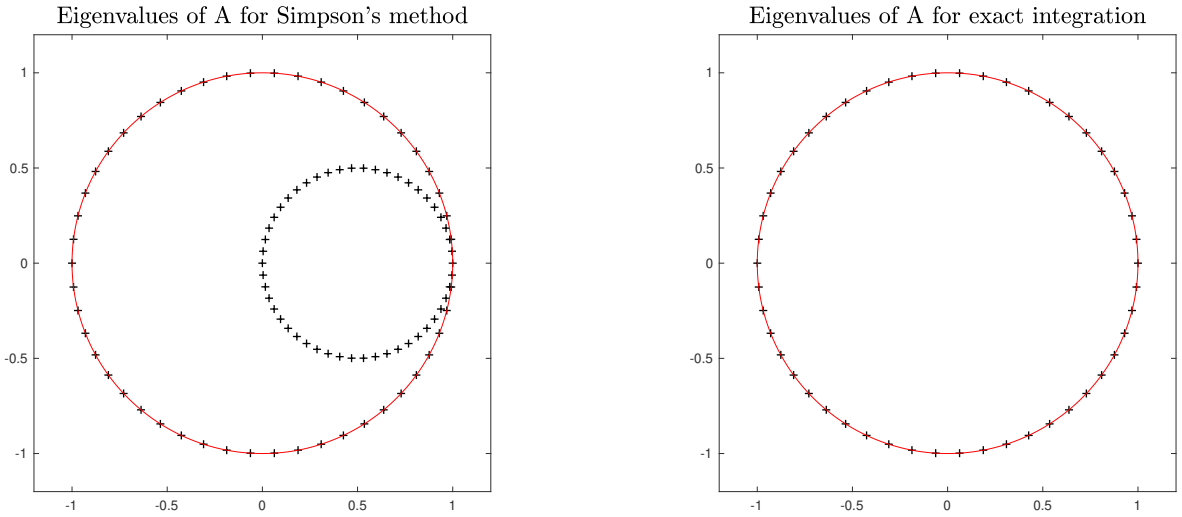

Figure A.3.: Eigenvalues for Simpson's method (left) and exact integration (right) for the AF method in the case $a = b$, $\nu = 1$ and $N = 50$.

B Statement about the Authors Contribution to Previously Published Work

The work of this thesis has been done within the context of the project "High-order accurate numerical methods for hyperbolic partial differential equations on Cartesian grids with embedded geometry", supported by the German Research Foundation through HE 4858/4-1,4-2. Two journal articles and one article in a conference proceeding have been published.

The theoretical results in [1] were derived by the authors supervisor, Christiane Helzel, and the author of this thesis in equal parts. All numerical computations were performed by the author of this thesis. An exception is [1, Section 4.2] which is due to Leonardo Scandurra who joined our group as a PostDoc at a later stage of the project.

Article [2] deals with stability and limiting of the Active Flux method in two spatial dimensions. At that time Erik Chudzik was a master student in our group. As part of his master thesis he developed a program which allowed to investigate the stability of the Active Flux method for the two-dimensional advection equation. The author of this thesis later developed a more general program for advection and linear acoustics which is described in [2, Section 3]. The computations in [2, Section 3.1] were performed by Erik Chudzik, whereas the results in [2, Section 3.2] are due to the author of this thesis. [2, Sections 1, 4] were developed by all three authors in equal parts. [2, Section 2] is due to Christiane Helzel and the author of this thesis in equal parts.

Paper [3, Section 2] deals with cut cell results for the Active Flux method. The theoretical results in [3, Section 2] were derived by both authors in equal parts. The results in [3, Section 3] and all computations are due to the author of this thesis. The entire article was written down by the author of this thesis.

Acknowledgment

This work was supported by the German Research Foundation through HE 4858/4-1,4-2.

Bibliography

[1] C. Helzel, D. Kerkmann, and L. Scandurra, "A New ADER Method Inspired by the Active Flux Method," *Journal of Scientific Computing*, vol. 80, pp. 1463–1497, 2019.

[2] E. Chudzik, C. Helzel, and D. Kerkmann, "The Cartesian Grid Active Flux Method: Linear stability and bound preserving limiting," *Applied Mathematics and Computation*, vol. 393, 125501, 2021.

[3] C. Helzel and D. Kerkmann, "An Active Flux Method for Cut Cell Grids," in *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples* (R. Klöfkorn, E. Keilegavlen, F. A. Radu, and J. Fuhrmann, eds.), (Cham), pp. 507–515, Springer International Publishing, 2020.

[4] J. S. Hesthaven, *Numerical Methods for Conservation Laws: From Analysis to Algorithms*. Computational Science and Engineering, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2018.

[5] D. Kröner, *Numerical Schemes for Conservation Laws*. Wiley-Teubner Series, Advances in Numerical Mathematics, Chichester, NY: Wiley-Teubner, 1997.

[6] R. J. LeVeque, *Finite-Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics, Cambridge: Cambridge University Press, 2002.

[7] B. Wendroff and P. Lax, "Systems of Conservation Laws," *Communications on Pure and Applied Mathematics*, vol. 13, pp. 217–237, 1960.

[8] J. J. Quirk, "An alternative to unstructured grids for computing gas dynamic flows around arbitrarily complex two-dimensional bodies," *Computers and Fluids*, vol. 23, no. 1, pp. 125–142, 1994.

[9] R. Qin and L. Krivodonova, "A discontinuous Galerkin method for solutions of the Euler equations on Cartesian grids with embedded geometries," *Journal of Computational Science*, vol. 4, pp. 24–35, 2013.

[10] P. Colella, D. T. Graves, B. J. Keen, and D. Modiano, "A Cartesian grid embedded boundary method for hyperbolic conservation laws," *Journal of Computational Physics*, vol. 211, no. 1, pp. 347–366, 2006.

[11] M. Berger and A. Giuliani, "A state redistribution algorithm for finite volume schemes on cut cell meshes," *Journal of Computational Physics*, vol. 428, 109820, 2021.

[12] A. Giuliani, "A two-dimensional stabilized discontinuous Galerkin method on curvilinear embedded boundary grids," *arXiv:2102.01857*, 2021.

[13] M. Berger and C. Helzel, "A Simplified h-box Method for Embedded Boundary Grids," *SIAM Journal on Scientific Computing*, vol. 34, no. 2, pp. A861–A888, 2012.

[14] M. J. Berger, C. Helzel, and R. J. LeVeque, "h-Box Methods for the Approximation of Hyperbolic Conservation Laws on Irregular Grids," *SIAM Journal on Numerical Analysis*, vol. 41, no. 3, pp. 893–918, 2003.

[15] C. Helzel, M. J. Berger, and R. J. Leveque, "A High-Resolution Rotated Grid Method for Conservation Laws with Embedded Geometries," *SIAM Journal on Scientific Computing*, vol. 26, no. 3, pp. 785–809, 2005.

[16] S. May and M. Berger, "An Explicit Implicit Scheme for Cut Cells in Embedded Boundary Meshes," *Journal of Scientific Computing*, vol. 71, pp. 919–943, 2017.

[17] S. May and M. Berger, "A Mixed Explicit Implicit Time Stepping Scheme for Cartesian Embedded Boundary Meshes," in *Finite Volumes for Complex Applications VII-Methods and Theoretical Aspects* (J. Fuhrmann, M. Ohlberger, and C. Rohde, eds.), vol. 77, (Cham), pp. 393–400, 2014.

[18] N. Gokhale, N. Nikiforakis, and R. Klein, "A dimensionally split Cartesian cut cell method for hyperbolic conservation laws," *Journal of Computational Physics*, vol. 364, pp. 186–208, 2018.

[19] R. Klein, K. R. Nordin-Bates, and N. Nikiforakis, "Well-balanced compressible cut-cell simulation of atmospheric flow," *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences*, vol. 367, pp. 4559–4575, 2009.

[20] S. Tan and C.-W. Shu, "Inverse Lax-Wendroff procedure for numerical boundary conditions of conservation laws," *Journal of Computational Physics*, vol. 229, no. 21, pp. 8144–8166, 2010.

[21] T. Li, C.-W. Shu, and M. Zhang, "Stability analysis of the inverse Lax-Wendroff boundary treatment for high order upwind-biased finite difference schemes," *Journal of Computational and Applied Mathematics*, vol. 299, pp. 140–158, 2016.

[22] J. Lu, C.-W. Shu, S. Tan, and M. Zhang, "An inverse Lax-Wendroff procedure for hyperbolic conservation laws with changing wind direction on the boundary," *Journal of Computational Physics*, vol. 426, 109940, 2021.

[23] S. Tan and C.-W. Shu, "Inverse Lax-Wendroff Procedure for Numerical Boundary Conditions of Hyperbolic Equations: Survey and New Developments," in *Advances in Applied Mathematics, Modeling, and Computational Science* (R. Melnik and I. S. Kotsireas, eds.), vol. 66 of *Fields Institute Communications*, (Boston, MA), pp. 41–63, Springer US, 2013.

[24] S. Tan, C. Wang, C.-W. Shu, and J. Ning, "Efficient implementation of high order inverse LaxWendroff boundary treatment for conservation laws," *Journal of Computational Physics*, vol. 231, no. 6, pp. 2510–2527, 2012.

[25] C. Engwer, S. May, A. Nüßing, and F. Streitbürger, "A Stabilized DG Cut Cell Method for Discretizing the Linear Transport Equation," *SIAM Journal on Scientific Computing*, vol. 42, no. 6, pp. A3677–A3703, 2020.

[26] T. A. Eymann and P. L. Roe, "Active Flux Schemes," in *49th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, AIAA 2011–382, 2011.

[27] T. A. Eymann and P. L. Roe, "Active Flux Schemes for Systems," in *20th AIAA Computational Fluid Dynamics Conference*, AIAA 2011–3840, 2011.

[28] T. A. Eymann and P. L. Roe, "Multidimensional Active Flux Schemes," in *21st AIAA Computational Fluid Dynamics Conference*, AIAA 2013–2940, 2013.

[29] H. Nishikawa, P. L. Roe, and T. A. Eymann, "Active Flux Schemes for Diffusion," in *44th AIAA AVIATION 2014 - 7th AIAA Theoretical Fluid Mechanics Conference*, AIAA 2014–2092, 2014.

[30] P. L. Roe, B. Maeng, and D. Fan, "Comparing Active Flux and Discontinuous Galerkin Methods for Compressible Flow," in *2018 AIAA Aerospace Sciences Meeting*, AIAA 2018–0836, 2018.

[31] B. van Leer, "Towards the Ultimate Conservative Difference Scheme. IV. A New Approach to Numerical Convection," *Journal of Computational Physics*, vol. 23, pp. 276–299, 1977.

[32] W. Barsukow, "The active flux scheme for nonlinear problems," *Journal of Scientific Computing*, vol. 86, article no. 3, 2021.

[33] D. Fan, *On the Acoustic Component of Active Flux Schemes for Nonlinear Hyperbolic Conservation Laws*. PhD thesis, University of Michigan, 2017.

[34] D. Fan and P. Roe, "Investigations of a new scheme for wave propagation," in *22nd AIAA Computational Fluid Dynamics Conference*, AIAA 2015–2449, 2015.

[35] J. D. Anderson, *Computational Fluid Dynamics: The Basics with Applications*. McGraw-Hill International Editions: Mechanical Engineering, Singapore: McGraw-Hill, 1995.

[36] P. Buchmüller and C. Helzel, "Improved Accuracy of High-Order WENO Finite Volume Methods on Cartesian Grids," *Journal of Scientific Computing*, vol. 61, pp. 343–368, 2014.

[37] J. L. M. van Dorsselaer, J. F. B. M. Kraaijevanger, and M. N. Spijker, "Linear stability analysis in the numerical solution of initial value problems," *Acta Numerica*, vol. 2, pp. 199–237, 1993.

[38] P. L. Roe, T. B. Lung, and B. Maeng, "New approaches to Limiting," in *22nd AIAA Computational Fluid Dynamics Conference*, AIAA 2015–2913, 2015.

[39] A. Marquina, "Local Piecewise Hyperbolic Reconstruction of Numerical Fluxes for Nonlinear Scalar Conservation Laws," *SIAM Journal on Scientific Computing*, vol. 15, no. 4, pp. 892–915, 1994.

[40] X. Zhang and C.-W. Shu, "Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments," *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 467, pp. 2752–2776, 2011.

[41] X. Zhang and C.-W. Shu, "On maximum-principle-satisfying high order schemes for scalar conservation laws," *Journal of Computational Physics*, vol. 229, no. 9, pp. 3091–3120, 2010.

[42] W. Barsukow, J. Hohm, C. Klingenberg, and P. L. Roe, "The Active Flux Scheme on Cartesian Grids and Its Low Mach Number Limit," *Journal of Scientific Computing*, vol. 81, pp. 594–622, 2019.

[43] P. Roe, "Is Discontinuous Reconstruction Really a Good Idea?," *Journal of Scientific Computing*, vol. 73, pp. 1094–1114, 12 2017.

[44] R. J. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2007.

[45] V. A. Titarev and E. F. Toro, "ADER: Arbitrary High Order Godunov Approach," *Journal of Scientific Computing*, vol. 17, pp. 609–618, 2002.

[46] P. Deuflhard and F. Bornemann, *Numerische Mathematik 2: Gewöhnliche Differentialgleichungen (German)*. Berlin, Boston: De Gruyter, 4th ed., 2013.

[47] C. Hu and C.-W. Shu, "Weighted Essentially Non-oscillatory Schemes on Triangular Meshes," *Journal of Computational Physics*, vol. 150, no. 1, pp. 97–127, 1999.

[48] M. Lukáčová-Medvid'ová, K. W. Morton, and G. Warnecke, "Evolution Galerkin methods for hyperbolic systems in two space dimensions," *Mathematics of Computation*, vol. 69, no. 232, pp. 1355–1384, 2000.

[49] H. Gilquin, J. Laurens, and C. Rosier, "Multi-dimensional Riemann problems for linear hyperbolic systems: Part II," in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects: Proceedings of the Fourth International Conference on Hyperbolic Problems, Taormina, Italy, April 3 to 8, 1992* (A. Donato and F. Oliveri, eds.), (Braunschweig, Wiesbaden), pp. 284–290, Vieweg, 1993.

Affidavit

I declare under oath that I have produced my thesis independently and without any undue assistance by third parties under consideration of the 'Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf'.

Düsseldorf, June 24th 2021 David Christian Kerkmann