Cartesian Grid Active Flux Methods for Hyperbolic Conservation Laws

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

Erik Chudzik

from Chemnitz

Düsseldorf, January 2024

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. Donna Calhoun Boise State University, United States, Idaho

Date of the oral examination: 06/14/2024

Contents

Introduction

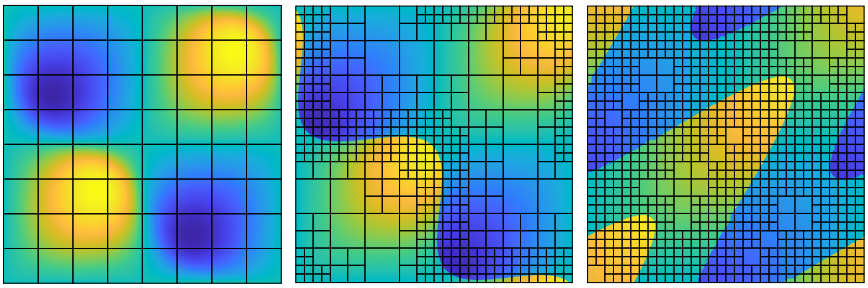
The Active Flux method is a novel finite volume method for hyperbolic conservation laws, which was proposed by Eymann and Roe in 2011 [14]. In the special case of one-dimensional advection an equivalent method is van Leer's Scheme V from the paper series "Towards the ultimate conservative difference scheme" [19]. Originally designed to be third-order accurate, the Active Flux method has recently been extended to a class of schemes of arbitrary orders of accuracy [7]. This thesis builds on the original third-order accurate version.

For the two-dimensional Active Flux method Eymann and Roe used unstructured triangular grids, we instead consider Cartesian grids as in [17] and [9]. To achieve high order accuracy, the Active Flux method enriches the stencil by adding point values along grid cell interfaces. This approach differs from traditional finite volume methods, where high accuracy is typically achieved by expanding the stencil to consider more neighboring cell averages. Using the point values and the cell averages, a continuous piecewise quadratic function is reconstructed, which serves as the initial condition for the evolution. The resulting method is truly multi-dimensional, fully discrete and gives accurate approximations even on coarse grids. Furthermore, it has a compact stencil in space and time, which from a physical point of view aligns with the desire to use as little information outside the true domain of dependence as possible [23]. This thesis consists of two main contributions:

I For complex solution structures it is often desirable to capture dynamically regions with, e.g., turbulent flow or steep gradients, more resolved, while leaving smooth regions at a lower level of resolution to reduce computational costs. For hyperbolic partial differential equations on Cartesian grids, adaptive mesh refinement was introduced by Berger, Oliger and Colella [10], [11]. When considering adaptive mesh refinement, the refinement of a cell typically involves some sort of interpolation through neighbouring cells, where the size of the stencil depends on the desired order of accuracy of the scheme. The Active Flux method however has a cellwise defined reconstruction, which can be evaluated in any point to compute point values on the refined grid.

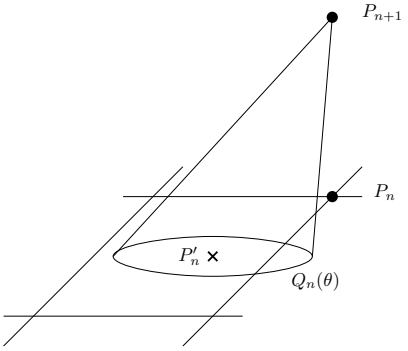
These can then be used to compute corresponding cell averages. We implemented an Active Flux method with adaptive mesh refinement in ForestClaw, which is a software for patch-based adaptive mesh refinement on a forest of quadtrees and capable of parallelization [12]. A detailed description of the adaptively refined Active Flux method can be found in [1], which is the first attached paper with title "The Cartesian Grid Active Flux Method with Adaptive Mesh Refinement".

Compared to, e.g., semi-discrete schemes, where every stage of the underlying Runge-Kutta method increases the stencil, or WENO schemes, which achieve high order of accuracy through expansive spatial stencils, the compact stencil of our method leads to significantly reduced overhead costs for data exchange between patches.



A solution to Burgers' equation at different times using adaptive mesh refinement [1]. The different squares indicate patches of Cartesian grids with 16×16 grid cells each.

II A distinctive property of the Active Flux method, compared to classical finite volume methods, is the use of point values. These are evolved in time using truly multi-dimensional evolution operators, which are derived from the quasilinear formulation of the governing equations. Cell averages are updated using the conservative formulation. For that matter we explored the method of bicharacteristics, which provides a general framework for the construction of evolution operators for linear hyperbolic systems [2]. This approach was previously used in the context of Evolution Galerkin methods[20]. A detailed description of the Active Flux method using bicharacteristics theory can be found in [2], which is the second attached paper with title "Active Flux Methods for Hyperbolic Systems using the Method of Bicharacteristics".



Update of a point value using the method of bicharacteristics for the linearised Euler equations [2].

This thesis is organized as follows: In the first chapter the original Active Flux method for one-dimensional linear advection is introduced in detail. From there we construct two- and three-dimensional Cartesian grid versions with the help of tensor bases. The resulting $3D$ Active Flux method has not been described previously.

In the second chapter we perform a linear stability analysis for one- and two-dimensional advection similar to [3], [17] but now we in addition derive exact formulas for the eigenvalues of the update matrix in the one-dimensional case.

How to construct an Active Flux method for nonlinear conservation laws via local linearisation around each point value is explained in the third chapter by means of two-dimensional Burgers' equation. While this approach can be extended to any nonlinear conservation law, it is important to note that the resulting method may not inherently achieve third-order accuracy. However first results using this approach for the Euler equations of gas dynamics shown in [2] are very promising.

High order accurate methods for nonlinear problems require some form of limiting in order to approximate shock waves or solution structures with steep gradients. There is not yet an established limiting approach for the Active Flux method but different concepts have been discussed in particular in the one-dimensional case: For example, in [24] the parabolic reconstruction is replaced by a combination of parabolas and straight lines and in [8] the author utilizes power law limiting. Both approaches are not extendable to the two- or three-dimensional case in a straight forward way. In the fourth chapter we therefore review two different limiters by Zhang and Shu, which can be extended to any dimension. We start with a bound preserving reconstruction limiter [26], which we already applied to Burgers' equation in [3]. It turns out that limiting the reconstruction does not always suppress spurious oscillations. We therefore in addition introduce a positivity preserving flux limiter motivated by [27] and illustrate the performance for advective

transport. Another interesting application, which is currently studied with Yanick Kiechle and not considered in this thesis, is the Vlasov-Poisson problem. In this situation it is important to preserve positivity of the Vlasov solution.

Attached are the two articles "The Cartesian Grid Active Flux Method with Adaptive Mesh Refinement" [1] and "Active Flux Methods for Hyperbolic Systems using the Method of Bicharacteristics" [2]. Furthermore, I contributed to article [3] as part of my Master of Science thesis and the conference proceedings [4] and [5].

Chapter 1

The Cartesian grid Active Flux method

We are interested in the solution of hyperbolic conservation laws in divergence form

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

with initial values

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

where $\Omega \subset \mathbb{R}^d$ is a domain in the d-dimensional space with interior Ω^0, $q : \Omega \times \mathbb{R}^+ \to \mathbb{R}^s$ is a vector of s conserved quantities, $f : \mathbb{R}^s \to \mathbb{R}^s \times \mathbb{R}^d$ is a flux function and $q_0 : \Omega \to \mathbb{R}^s$ is an initial condition.

In this chapter we describe the Active Flux method in one,two and three spatial dimensions.

1.1 $1D$ case

Lets consider linear advection in one dimension

$$\frac{\partial}{\partial t}q(x,t) + a\frac{\partial}{\partial x}q(x,t) = 0 \tag{1.3}$$

with advection speed $a \in \mathbb{R}$. The domain is discretised on a grid with cells $\mathcal{C}_i := [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ of constant size $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$.

The Active Flux method uses cell average values, as well as point values at cell interfaces $x_{i\pm\frac{1}{2}}$ of the conserved quantities. We denote them by

$$Q_i^n \approx \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x,t_n)\,\mathrm{d}x, \quad Q_{i\pm\frac{1}{2}}^n \approx q(x_{i\pm\frac{1}{2}},t_n). \tag{1.4}$$

To start with third order accurate approximations, we initialize point values of q_0 at cell interfaces $x_{i\pm\frac{1}{2}}$ as well as at cell centres x_i and use Simpson's rule

to calculate the cell average values $Q_i^0 \approx \frac{1}{6}\left(q_0(x_{i-\frac{1}{2}}) + 4q_0(x_i) + q_0(x_{i+\frac{1}{2}})\right)$. These are then generally updated using a finite volume method

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}}\right), \tag{1.5}$$

where Δt is the time step size. Numerical fluxes are again computed with Simpson's rule but now applied in time, i.e.,

$$F_{i\pm\frac{1}{2}} := \frac{1}{6}\left(f(Q^n_{i\pm\frac{1}{2}}) + 4f(Q^{n+\frac{1}{2}}_{i\pm\frac{1}{2}}) + f(Q^{n+1}_{i\pm\frac{1}{2}})\right) \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(q(x_{i\pm\frac{1}{2}}, t))\,\mathrm{d}t. \tag{1.6}$$

The four missing point values $Q^{n+\frac{1}{2}}_{i\pm\frac{1}{2}}$ and $Q^{n+1}_{i\pm\frac{1}{2}}$ need to be approximated with at least third order to obtain a third order scheme. In case of linear advection one can use the exact solution given by

$$q(x,t) = q_0(x - at) \tag{1.7}$$

to compute the point values at the intermediate and new time level by tracing back characteristics, i.e.,

$$Q^{n+\frac{1}{2}}_{i\pm\frac{1}{2}} \approx q\left(x_{i\pm\frac{1}{2}} - \frac{1}{2}a\Delta t, t_n\right), \quad Q^{n+1}_{i\pm\frac{1}{2}} \approx q\left(x_{i\pm\frac{1}{2}} - a\Delta t, t_n\right). \tag{1.8}$$

Next we compute a continuous, piecewise quadratic reconstruction q^{rec} using the point values at grid cell interfaces and the cell average values, which should interpolate the point values and preserve the cell average values, i.e.,

$$q^{rec}(x_{i\pm\frac{1}{2}}) = Q^n_{i\pm\frac{1}{2}}, \quad \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q^{rec}(x)\,\mathrm{d}x = Q_i^n. \tag{1.9}$$

For that matter we transform each cell to the reference cell $[-1,1]$, i.e., we define $\xi = \xi(x) := \frac{2}{\Delta x}(x - x_{i-\frac{1}{2}}) - 1$, and make the ansatz $q_i^{rec}(\xi) = Q^n_{i-\frac{1}{2}}N_{-\frac{1}{2}}(\xi) + cN_0(\xi) + Q^n_{i+\frac{1}{2}}N_{\frac{1}{2}}(\xi)$ with the Lagrange interpolating polynomials

$$N_{-\frac{1}{2}}(\xi) = \frac{1}{2}(\xi^2 - \xi), \quad N_0(\xi) = 1 - \xi^2, \quad N_{\frac{1}{2}}(\xi) = \frac{1}{2}(\xi^2 + \xi). \tag{1.10}$$

Solving for c using the preservation of the cell average and Simpson's rule we get

$$q_i^{rec}(\xi) = Q^n_{i-\frac{1}{2}}N_{-\frac{1}{2}}(\xi) + \frac{1}{4}(6Q_i^n - (Q^n_{i-\frac{1}{2}} + Q^n_{i+\frac{1}{2}}))N_0(\xi) + Q^n_{i+\frac{1}{2}}N_{\frac{1}{2}}(\xi). \tag{1.11}$$

In case of linear advection, the update of the point values now takes the form

$$Q^{n+k}_{i+\frac{1}{2}} = \begin{cases} q_i^{rec}(\ \ 1 - ak\frac{2\Delta t}{\Delta x}) & : a > 0 \\ q_{i+1}^{rec}(-1 - ak\frac{2\Delta t}{\Delta x}) & : a \le 0 \end{cases}, \tag{1.12}$$

with $k = \frac{1}{2}, 1$ and for the fluxes we observe

$$\begin{aligned}
\frac{a}{\Delta t}\int_{t_n}^{t_{n+1}} q(x_{i+\frac{1}{2}}, t)\,\mathrm{d}t &= \frac{a}{\Delta t}\int_{t_n}^{t_{n+1}} q(x_{i+\frac{1}{2}} - a(t-t_n), t_n)\,\mathrm{d}t \\
&= \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-a\Delta t}^{x_{i+\frac{1}{2}}} q(x, t_n)\,\mathrm{d}x \\
&\approx \frac{\Delta x}{2\Delta t}\int_{1-a\frac{2\Delta t}{\Delta x}}^{1} q_i^{rec}(\xi)\,\mathrm{d}\xi \\
&= \frac{a}{6}\left(q_i^{rec}\left(1 - a\frac{2\Delta t}{\Delta x}\right) + 4q_i^{rec}\left(1 - a\frac{\Delta t}{\Delta x}\right) + q_i^{rec}(1)\right) \\
&= F_{i+\frac{1}{2}}.
\end{aligned} \tag{1.13}$$

This means that an approximation is only introduced by replacing the exact solution $q(x, t_n)$ with the third order accurate piecewise quadratic function $q^{rec}(\xi)$, since under the CFL condition $\frac{a\Delta t}{\Delta x} \leq 1$ Simpson's rule leads to the exact integral. As we will see, this is no longer the case in two spatial dimensions leading to a reduced stability. Reconstruction, evolution and averaging of the Active Flux scheme for one-dimensional advection are summarized in Fig.1.1.

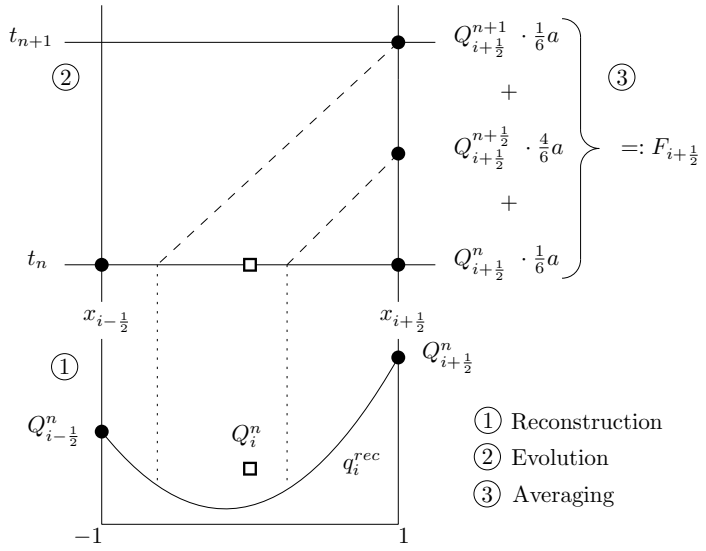


Figure 1.1: Illustration of reconstruction, evolution and averaging for the Active Flux scheme based on one-dimensional advection.

1.2 $2D$ case

In the two-dimensional case

$$\frac{\partial}{\partial t}q(x,y,t) + \frac{\partial}{\partial x}f(q(x,y,t)) + \frac{\partial}{\partial y}g(q(x,y,t)) = 0, \tag{1.14}$$

with $q : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^s$, $f, g : \mathbb{R}^s \to \mathbb{R}^s$, we use a Cartesian grid with $\Delta x = \Delta y$. The degrees of freedom of the method are again cell average values and point values, which are now located along grid cell boundaries as illustrated in Fig.1.2. We compute a piecewise quadratic and globally con-

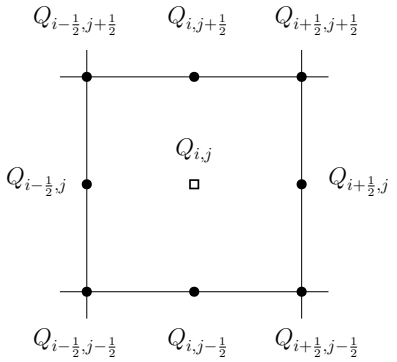


Figure 1.2: Degrees of freedom used for two-dimensional Cartesian grid Active Flux method.

tinuous reconstruction, which preserves the cell averages, by transforming each cell to the reference cell $[-1,1]^2$. Choosing the tensor basis $M_{\hat{\imath},\hat{\jmath}}(\xi,\eta) := N_{\hat{\imath}}(\xi)N_{\hat{\jmath}}(\eta), \hat{\imath},\hat{\jmath} \in \left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$ we get

$$q_{i,j}^{rec}(\xi,\eta) = \sum_{(\hat{\imath},\hat{\jmath})\neq(0,0)} Q^n_{i+\hat{\imath},j+\hat{\jmath}} M_{\hat{\imath},\hat{\jmath}}(\xi,\eta) + cM_{0,0}(\xi,\eta). \tag{1.15}$$

Preservation of the cell average $\frac{1}{\Delta x \Delta y}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q_{i,j}^{rec}(x,y)\,\mathrm{d}y\,\mathrm{d}x = Q^n_{i,j}$ defines c as

$$\begin{aligned} c = \frac{1}{16}\Big(36Q^n_{i,j} - (&Q^n_{i-\frac{1}{2},j-\frac{1}{2}} + Q^n_{i+\frac{1}{2},j-\frac{1}{2}} + Q^n_{i+\frac{1}{2},j+\frac{1}{2}} + Q^n_{i-\frac{1}{2},j+\frac{1}{2}}) \\ &-4(Q^n_{i,j-\frac{1}{2}} + Q^n_{i+\frac{1}{2},j} + Q^n_{i,j+\frac{1}{2}} + Q^n_{i-\frac{1}{2},j})\Big). \end{aligned} \tag{1.16}$$

Cell averages are again evolved in time using a finite volume method

$$Q^{n+1}_{i,j} = Q^n_{i,j} - \frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2},j} - F_{i-\frac{1}{2},j}\right) - \frac{\Delta t}{\Delta y}\left(G_{i,j+\frac{1}{2}} - G_{i,j-\frac{1}{2}}\right), \tag{1.17}$$

where the numerical fluxes are computed using the two-dimensional version of Simpson's rule, e.g., the fluxes at the right and left interface have the

form

$$F_{i\pm\frac{1}{2},j} := \frac{1}{36}\Big(f(Q^n_{i\pm\frac{1}{2},j-\frac{1}{2}}) + \ 4f(Q^n_{i\pm\frac{1}{2},j}) + \ f(Q^n_{i\pm\frac{1}{2},j+\frac{1}{2}}) \\ +4f(Q^{n+\frac{1}{2}}_{i\pm\frac{1}{2},j-\frac{1}{2}}) + 16f(Q^{n+\frac{1}{2}}_{i\pm\frac{1}{2},j}) + 4f(Q^{n+\frac{1}{2}}_{i\pm\frac{1}{2},j+\frac{1}{2}}) \\ +f(Q^{n+1}_{i\pm\frac{1}{2},j-\frac{1}{2}}) + \ 4f(Q^{n+1}_{i\pm\frac{1}{2},j}) + \ f(Q^{n+1}_{i\pm\frac{1}{2},j+\frac{1}{2}})\Big). \tag{1.18}$$

The fluxes at bottom and top interfaces are computed analogously. In the case of two-dimensional linear advection with advection speeds $a, b \in \mathbb{R}$, i.e.,

$$\frac{\partial}{\partial t}q(x,y,t) + a\frac{\partial}{\partial x}q(x,y,t) + b\frac{\partial}{\partial y}q(x,y,t) = 0, \tag{1.19}$$

we again use the exact solution $q(x,y,t) = q_0(x-at, y-bt)$ to trace back characteristics, update our point values and compute fluxes as indicated in Fig.1.3 (left). The flux computation is now equivalent to the integration

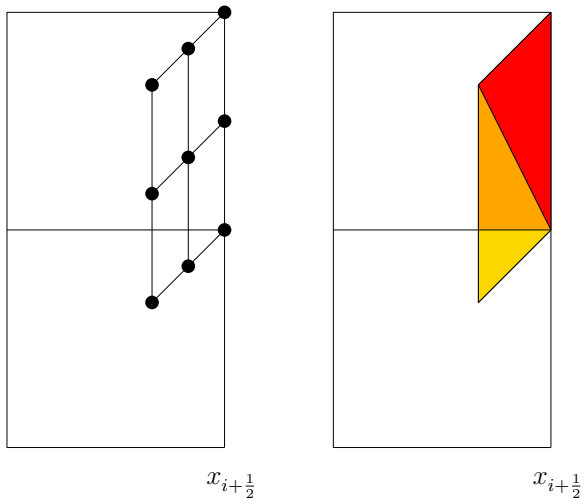


Figure 1.3: Flux computation for the two-dimensional advection equation using Simpson's rule (left) and exact integration (right).

of our reconstruction on a parallelogram in the $x-y$ plane, introducing, compared to the one-dimensional case, a second approximation. As in the one-dimensional case, the first approximation came from replacing the exact solution with our reconstruction. Since the parallelogram will typically lay in two neighbouring cells, Simpson's rule is now used to approximate the integral over a piecewise quadratic function and is no longer exact. As we will see, this leads to a reduced stability in the two-dimensional case and will also cause additional difficulties when it comes to limiting.

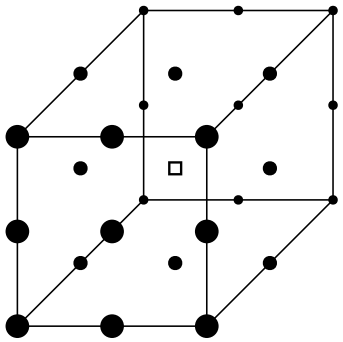
1.3 $3D$ case

We already constructed the two-dimensional Active Flux method from the one-dimensional method by choosing the tensor basis of the one-dimensional basis as the basis polynoms for the two-dimensional reconstruction. Fluxes have than been computed via a two-dimensional version of Simpson's rule and the cell averages have been updated with a finite volume method. This idea is basically extendable to any dimension.

In the three-dimensional case

$$\frac{\partial}{\partial t}q(x,y,z,t)+\frac{\partial}{\partial x}f(q(x,y,z,t))+\frac{\partial}{\partial y}g(q(x,y,z,t))+\frac{\partial}{\partial z}h(q(x,y,z,t))=0, \tag{1.20}$$

with $q:\mathbb{R}^3\times\mathbb{R}^+\to\mathbb{R}^s$, $f,g,h:\mathbb{R}^s\to\mathbb{R}^s$, we use again a Cartesian grid with $\Delta x=\Delta y=\Delta z$. The degrees of freedom of the method are cell average values and point values, which are now located along the boundary surfaces of a cell as illustrated in the figure to the right, i.e., the point values are given by the vertices, centres of edges and centres of faces of a cell. The piecewise quadratic and globally continuous reconstruction is now constructed again by transforming each cell to the reference cell $[-1,1]^3$ and with the tensor basis $P_{\hat{\imath},\hat{\jmath},\hat{k}}(\xi,\eta,\mu):=N_{\hat{\imath}}(\xi)N_{\hat{\jmath}}(\eta)N_{\hat{k}}(\mu)$, $\hat{\imath},\hat{\jmath},\hat{k}\in\left\{-\frac{1}{2},0,\frac{1}{2}\right\}$. For each cell we get



$$q^{rec}_{i,j,k}(\xi,\eta,\mu)=\sum_{(\hat{\imath},\hat{\jmath},\hat{k})\neq(0,0,0)}Q^n_{i+\hat{\imath},j+\hat{\jmath},k+\hat{k}}P_{\hat{\imath},\hat{\jmath},\hat{k}}(\xi,\eta,\mu)+cP_{0,0,0}(\xi,\eta,\mu) \tag{1.21}$$

choosing c so that the average of the reconstruction in a cell equals its cell average, i.e., we solve

$$\frac{1}{2^3}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}q^{rec}_{i,j,k}(\xi,\eta,\mu)\,\mathrm{d}\xi\,\mathrm{d}\eta\,\mathrm{d}\mu=Q^n_{i,j,k} \tag{1.22}$$

for c. The cell average update is again performed in a finite volume manner

$$\begin{aligned}Q^{n+1}_{i,j,k}=Q^n_{i,j,k}&-\frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2},j,k}-F_{i-\frac{1}{2},j,k}\right)\\&-\frac{\Delta t}{\Delta y}\left(G_{i,j+\frac{1}{2},k}-G_{i,j-\frac{1}{2},k}\right)-\frac{\Delta t}{\Delta z}\left(H_{i,j,k+\frac{1}{2}}-H_{i,j,k-\frac{1}{2}}\right),\end{aligned}$$

where all fluxes, e.g.,

$$H_{i,j,k+\frac{1}{2}}\approx\frac{1}{\Delta x\Delta y\Delta t}\int_{t_n}^{t_{n+1}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}q(x,y,z_{k+\frac{1}{2}},t)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t, \tag{1.23}$$

are computed by application of Simpson's rule to each integral.

For a three-dimensional quadratic reconstruction one only needs 10 basis polynomials. In future investigations we should find alternative reconstructions and suitable quadrature formulas. Instead of the tensor basis consisting of 27 basis polynomials and Simpson's rule one could use, e.g., only edge centre point values and the preservation of the cell average value to construct the reconstruction together with a two-point Gauss quadrature rule for the calculation of the fluxes. This would already reduce the number of basis-polynomials to 13. Since the quadrature rule is no longer closed, the point values are then updated separately and one needs to explore, whether such a newly derived scheme would indeed be more efficient and would have good stability properties. In Section 4.3 we show first numerical results using the three-dimensional version described above.

Chapter 2

Linear stability analysis

The Active Flux method for linear problems, e.g., linear advection as described in the previous section, is a linear method. This means we can write the method in the form

$$\boldsymbol{Q}^{n+1} = A\boldsymbol{Q}^n, \tag{2.1}$$

where $\boldsymbol{Q}^n$ is the vector of all degrees of freedom at the current time level t^n and A is a matrix containing row-wise the coefficients of each degree of freedom for the update of the point values and cell average values.

A criteria for linear stability of a method is Lax-Richtmyer stability, where a method is Lax-Richtmyer stable iff $||A^n||$ is bounded independently of n. Let λ be an eigenvalue of A, then this is equivalent to the condition $|\lambda| \leq 1$ and if $|\lambda| = 1$, then the geometric and algebraic multiplicity need to match [13].

In the next section we describe the update matrix A in the case of one-dimensional linear advection and derive explicit formulas for the eigenvalues of A.

2.1 $1D$ case

Lets consider a positive advection speed $a > 0$ and a domain which we discretise with m cells of constant size Δx. Furthermore we impose periodic boundary conditions on this domain. Associating the left point value $Q_{i-\frac{1}{2}}$ and the cell average value Q_i with each cell $\mathcal{C}_i, i = 1, \ldots, m$, the vector of all degrees of freedom $\boldsymbol{Q}^n := \boldsymbol{Q}$ is given by

$$\boldsymbol{Q} = [Q_{\frac{1}{2}}, Q_1, Q_{\frac{3}{2}}, Q_2, \ldots, Q_{m-\frac{1}{2}}, Q_m]^T \in \mathbb{R}^{2m}. \tag{2.2}$$

Updating a cell $\mathcal{C}_i$ involves the two neighbouring cells $\mathcal{C}_{i-1}$ and $\mathcal{C}_{i+1}$. The contributions of these three cells to the update of cell $\mathcal{C}_i$ are stored in matrices a_{-1}, a_0 and a_1, i.e., with $C := \text{CFL} = |\frac{a\Delta t}{\Delta x}|$, the Courant-Friedrichs-Lewy

number,

$$
\begin{aligned}
a_{-1}(C) &= \begin{pmatrix} 3C^2-2C & -6C^2+6C \\ C^3-C^2 & -2C^3+3C^2 \end{pmatrix}, \\
a_0(C) &= \begin{pmatrix} 3C^2-4C+1 & 0 \\ -C^2+C & 2C^3-3C^2+1 \end{pmatrix}, \\
a_1(C) &= \begin{pmatrix} 0 & 0 \\ -C^3+2C^2-C & 0 \end{pmatrix}.
\end{aligned} \tag{2.3}
$$

These matrices are derived in the following way. The update of a point value $Q^n_{i-\frac{1}{2}}$ is with (1.12) given by

$$
\begin{aligned}
Q^{n+1}_{i-\frac{1}{2}} &= q^{rec}_{i-1}(1-2C) \\
&= (3C^2-2C)Q^n_{i-\frac{3}{2}} + (-6C^2+6C)Q^n_{i-1} + (3C^2-4C+1)Q^n_{i-\frac{1}{2}}.
\end{aligned} \tag{2.4}
$$

Extracting the coefficients of the degrees of freedom from cells $\mathcal{C}_{i-1}$, $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$ gives the first row of the matrices. Similary the update of a cell average value Q^n_i is with (1.13) given by

$$
\begin{aligned}
Q^{n+1}_i = Q^n_i &- C\left(q^{rec}_i(1-2C) + 4q^{rec}_i(1-C) + q^{rec}_i(1)\right) \\
&- C\left(q^{rec}_{i-1}(1-2C) + 4q^{rec}_{i-1}(1-C) + q^{rec}_{i-1}(1)\right).
\end{aligned} \tag{2.5}
$$

Writing this again as a linear combination of the degrees of freedom from cells $\mathcal{C}_{i-1}$, $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$ gives the second row of the matrices.
With

$$
P_1 := \begin{pmatrix}
0 & 1 & 0 & \cdots & \cdots & 0 \\
\vdots & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \vdots & & \ddots & \ddots & 0 \\
0 & \vdots & & & 0 & 1 \\
1 & 0 & \cdots & \cdots & \cdots & 0
\end{pmatrix} \in \mathbb{R}^m, \quad P_0 = Id_{m\times m}, \quad P_{-1} := P_1^T \tag{2.6}
$$

we can now construct A through Kronecker products obtaining

$$
A = P_{-1} \otimes a_{-1}(C) + P_0 \otimes a_0(C) + P_1 \otimes a_1(C). \tag{2.7}
$$

Theorem 2.1.1. *Let $\gamma := \gamma(k,m) = \exp\left(2\pi i \frac{k}{m}\right)$, $k = 1,\dots,m$, be the m-th unit roots and*

$$
\begin{aligned}
p_1(C,k,m) &:= (\gamma-1)C^3 + 3C^2 + (-2\gamma-1)C + \gamma, \\
p_2(C,k,m) &:= \sqrt{p_3(C,k,m)}(C-1)C
\end{aligned} \tag{2.8}
$$

with

$$p_3(C,k,m) := (\gamma^2 - 2\gamma + 1)C^2 + 2(\gamma^2 + \gamma - 2)C + (-2\gamma^2 + 10\gamma + 1). \quad (2.9)$$

The eigenvalues of A are then given by

$$\begin{aligned} \lambda_1(C,k,m) &= (p_1(C,k,m) - p_2(C,k,m))\gamma(k,m)^{-1}, \\ \lambda_2(C,k,m) &= (p_1(C,k,m) + p_2(C,k,m))\gamma(k,m)^{-1}. \end{aligned} \quad (2.10)$$

Proof: The matrices P_{-1} and P_1 are diagonalizable with eigenvalues $\gamma(k,m)^{-1}$ and $\gamma(k,m)$. Since they commute, they are simultaneously diagonalizable and A is similar to a block diagonal matrix

$$\tilde{A} = \begin{pmatrix} \tilde{a}(C,1,m) & & & \\ & \tilde{a}(C,2,m) & & \\ & & \ddots & \\ & & & \tilde{a}(C,m,m) \end{pmatrix} \quad (2.11)$$

with

$$\tilde{a}(C,k,m) = \begin{pmatrix} \tilde{a}_{1,1} & \tilde{a}_{2,1} \\ \tilde{a}_{2,2} & \tilde{a}_{2,2} \end{pmatrix} \gamma(k,m)^{-1}$$

and

$$\begin{aligned} \tilde{a}_{1,1} &= (3\gamma + 3)C^2 + (-4\gamma - 2)C + \gamma, \\ \tilde{a}_{1,2} &= -6C^2 + 6C, \\ \tilde{a}_{2,1} &= (-\gamma^2 + 1)C^3 + (2\gamma^2 - \gamma - 1)C^2 + (-\gamma^2 + \gamma)C, \\ \tilde{a}_{2,2} &= (2\gamma - 2)C^3 + (-3\gamma + 3)C^2 + \gamma. \end{aligned} \quad (2.12)$$

The eigenvalues of $\tilde{a}(C,k,m)$ are $\lambda_1(C,k,m)$ and $\lambda_2(C,k,m)$. □

Corollary 2.1.1. *The Active Flux method for one-dimensional linear advection is Lax-Richtmyer stable for CFL = 1.*

Proof: It holds

$$\tilde{a}(1,k,m) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \gamma(k,m)^{-1}.$$

□

Corollary 2.1.2. *The update matrix A of the Active Flux method for one-dimensional linear advection is similar to a diagonal matrix.*

Proof: It holds $\lambda_1 = \lambda_2$ is equivalent to p_2=0. This is true for $C = 0$ or $C = 1$, for which we already know that A is similar to a diagonal matrix, or $p_3 = 0$.

We now consider three cases:
- For $\gamma = 1$ we have $p_3(C, m, m) = 9 \neq 0$.
- For $\gamma = -1$ we have $p_3(C, \frac{m}{2}, m) = 4C^2 - 4C - 11 \neq 0, C \in (0, 1)$.
- For $\gamma = \alpha + i\beta, \beta \neq 0$ holds

$$\mathrm{Im}(p_3) = 2(C^2 + 2C - 2)\alpha\beta + (-2C^2 + 2C + 10)\beta \tag{2.13}$$

and $\mathrm{Im}(p_3) = 0$ is equivalent to

$$\alpha = \frac{C^2 - C - 5}{C^2 + 2C - 2}, \tag{2.14}$$

where $\left|\frac{C^2-C-5}{C^2+2C-2}\right| > 1, C \in (0, 1)$. This is a contradiction to γ being a unit root. It follows $\lambda_1 \neq \lambda_2$ and there is an invertible matrix $S \in \mathbb{R}^{2m\times 2m}$ such that SAS^{-1} is diagonal. □

Fig.2.1 shows the eigenvalues of A for $\mathrm{CFL} = \frac{n}{10}$, $n = 1, \ldots, 10$ and $m = 32$ together with the unit circle, λ_1 is marked in red and λ_2 is marked in blue dots. Though equations (2.10) are not of practical use for a rigorous proof of the stability of the Active Flux method for one-dimensional linear advection, these plots indicate that the method is indeed stable for $\mathrm{CFL} \leq 1$, since all eigenvalues remain in the unit circle and the update matrix is similar to a diagonal matrix, meaning that the geometric and algebraic multiplicities do match.

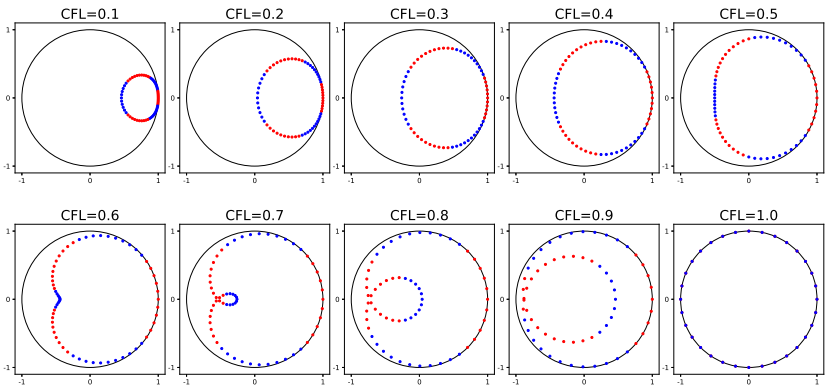


Figure 2.1: Eigenvalues of A for $\mathrm{CFL} = \frac{n}{10}$, $n = 1, \ldots, 10$ in case of one-dimensional linear advection.

2.2 $2D$ case

In this section we will review some results from [3]. Lets consider positive advection speeds $a > 0$, $b > 0$ and a quadratic domain which we

discretise with $m \times m$ cells. Furthermore we impose double periodic boundary conditions. Associating the bottom left corner, bottom face and left face point value together with the cell average value with a cell $\mathcal{C}_{i,j}$, i.e., $\mathcal{C}_{i,j} \hat{=} \left[Q_{i-\frac{1}{2},j-\frac{1}{2}}, Q_{i,j-\frac{1}{2}}, Q_{i-\frac{1}{2},j}, Q_{i,j}\right]$, the stencil for the update of a cell $\mathcal{C}_{i,j}$ consists of the cells $\mathcal{C}_{i+\hat{\imath},j+\hat{\jmath}}$, $\hat{\imath},\hat{\jmath} = -1,0,1$. We then write the vector of all degrees of freedom $\boldsymbol{Q}^n \in \mathbb{R}^{4m^2}$ by concatenating all cells row by row, i.e.,

$$\boldsymbol{Q}^n = [\mathcal{C}_{1,1}, \mathcal{C}_{2,1}, \ldots, \mathcal{C}_{m,1}, \mathcal{C}_{1,2}, \mathcal{C}_{2,2}, \ldots, \mathcal{C}_{m,2}, \ldots, \mathcal{C}_{1,m}, \mathcal{C}_{2,m}, \ldots, \mathcal{C}_{m,m}]^T, \tag{2.15}$$

so the update matrix can again be constructed through Kronecker products

$$A = \sum_{\hat{\imath},\hat{\jmath}=-1,0,1} P_{\hat{\imath}} \otimes P_{\hat{\jmath}} \otimes a_{\hat{\imath},\hat{\jmath}}(C_x, C_y) \tag{2.16}$$

with P_{-1}, P_0 and P_1 as in the one-dimensional case and $C_x = \mathrm{CFL}_x := |\frac{a\Delta t}{\Delta x}|$, $C_y = \mathrm{CFL}_y := |\frac{b\Delta t}{\Delta y}|$. The matrices $a_{\hat{\imath},\hat{\jmath}} \in \mathbb{R}^{4\times 4}$, $\hat{\imath},\hat{\jmath} = -1,0,1$ contain the contributions of cells with index $(\hat{\imath},\hat{\jmath})$ relative to the cell being updated.

Since the matrices $P_{\hat{\imath}} \otimes P_{\hat{\jmath}}$ are simultaneously diagonalizable with eigenvalues $\lambda_{\hat{\imath},\hat{\jmath}}(k_1,k_2,m) = \exp(2\pi i \frac{\hat{\imath}k_1}{m}) \exp(2\pi i \frac{\hat{\jmath}k_2}{m})$, $k_1,k_2 = 1,\ldots,m$, the update matrix A is again similar to a block diagonal matrix with blocks $\tilde{a}(C_x,C_y,k_1,k_2,m) \in \mathbb{R}^{4\times 4}$ and the eigenvalues can be computed exactly. The resulting formulas for the eigenvalues are not of practical use in general and are omitted here due to their length, however for the special case $\mathrm{CFL}_x = \mathrm{CFL}_y = 1$ we can prove the following result.

Lemma 2.2.1. *The Active Flux method for two-dimensional linear advection is not Lax-Richtmyer stable for $CFL_x = CFL_y = 1$.*

Proof: For $\mathrm{CFL}_x = \mathrm{CFL}_y = 1$ the matrices $a_{\hat{\imath},\hat{\jmath}}$ are

$$a_{-1,1} = 0_{4\times 4}\,, \qquad a_{0,1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{72} & -\frac{1}{18} & 0 & 0 \end{pmatrix}, \qquad a_{1,1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\frac{1}{24} & 0 & 0 & 0 \end{pmatrix},$$

$$a_{-1,0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\frac{1}{72} & \frac{5}{18} & 0 & 0 \end{pmatrix}, \quad a_{0,0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -\frac{5}{18} & -\frac{5}{18} & \frac{1}{2} \end{pmatrix}, \quad a_{1,0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{72} & 0 & -\frac{1}{18} & 0 \end{pmatrix},$$

$$a_{-1,-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{24} & \frac{1}{18} & \frac{1}{18} & \frac{1}{2} \end{pmatrix}, \quad a_{0,-1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\frac{1}{72} & 0 & \frac{5}{18} & 0 \end{pmatrix}, \quad a_{1,-1} = 0_{4\times 4}\,.$$

Let m be even, then

$$\tilde{a}\left(1,1,\frac{m}{2},\frac{m}{2},m\right) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -\frac{4}{9} & -\frac{4}{9} & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.17}$$

meaning that the geometric and algebraic multiplicity of eigenvalue 1 do not match. □

Since the formulas for the eigenvalues are to complex in order to derive an analytic criteria for stability, we check for a fixed grid with m = 20, whether there exist eigenvalues with absolute value larger than one varying CFL_x and CFL_y from 0 to 1 as shown in Fig.2.2 (left). Dots indicate a situation, where the absolute value of all eigenvalues is less than or equal 1. Clearly not all pairs $(\mathrm{CFL}_x, \mathrm{CFL}_y)$ lead to a stable method with Simpson's rule. For $\mathrm{CFL}_x = \mathrm{CFL}_y$ the highest CFL number that can be stable is around 0.75.

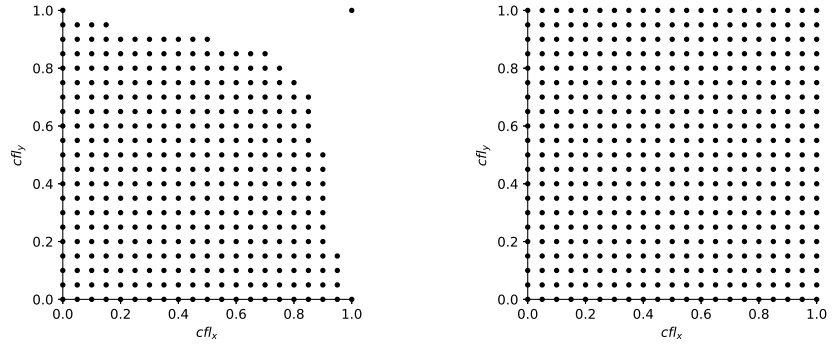


Figure 2.2: Pairs $(\mathrm{CFL}_x, \mathrm{CFL}_y)$ with eigenvalues of bounded by 1 absolute values indicated as dots for Simpson's rule (left) and exact integration (right) [3].

In this simple case of linear advection we can obtain a method which is stable for $\mathrm{CFL} \leq 1$ by calculating the numerical fluxes exactly. For that matter we split the parallelograms, over which the integration for the computation of the fluxes takes place, into three triangles as indicated in Fig.1.3 (right). We can integrate the reconstruction now exactly using Gauss-Lobatto quadrature after transforming each triangle to the reference triangle with corners $(0,0),(1,0)$ and $(0,1)$. Checking again the absolute value of the eigenvalues, Fig.2.2 (right), the method with exact integration is indeed stable for all $\mathrm{CFL} = \max(\mathrm{CFL}_x, \mathrm{CFL}_y) \leq 1$. This is further confirmed in Fig.2.3, where we show $||A^n||,\ n = 0, \ldots, 1000$ for $\mathrm{CFL}_x = \mathrm{CFL}_y = 0.75, 0.9, 1$ using Simpson's rule (top) and exact integration (bottom). With Simpson's rule $||A^n||$ is bounded for $\mathrm{CFL}_x = \mathrm{CFL}_y = 0.75$, the case $\mathrm{CFL}_x = \mathrm{CFL}_y = 0.9$

shows exponential growth indicating an unstable situation and the case $\mathrm{CFL}_x = \mathrm{CFL}_y = 1$ shows the linear growth we expected from (2.17). Using exact integration the term $||A^n||$ is bounded in all three cases and the method is stable.

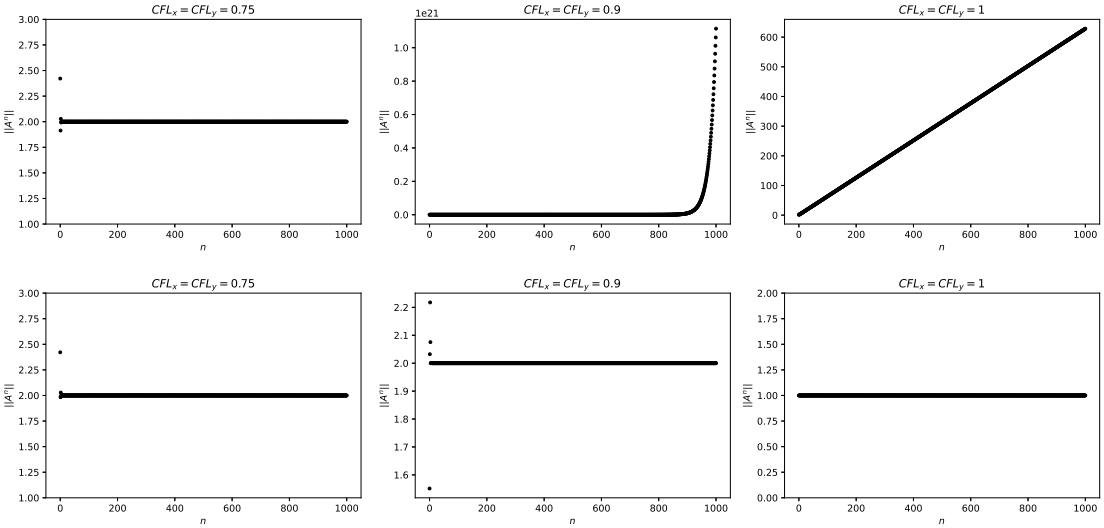


Figure 2.3: $||A^n||,\ n = 0, \ldots, 1000$ for $\mathrm{CFL}_x = \mathrm{CFL}_y = 0.75, 0.9, 1$ using Simpson's rule (top) and exact integration (bottom) [3].

Finally we show the eigenvalues of A for $\mathrm{CFL}_x = \mathrm{CFL}_y = 0.75, 0.9$, Fig.2.4. In case of Simpson's rule, some eigenvalues lay outside the unit circle for $\mathrm{CFL}_x = \mathrm{CFL}_y = 0.9$, using exact integration they are pulled back inside the unit circle leading to a stable scheme.

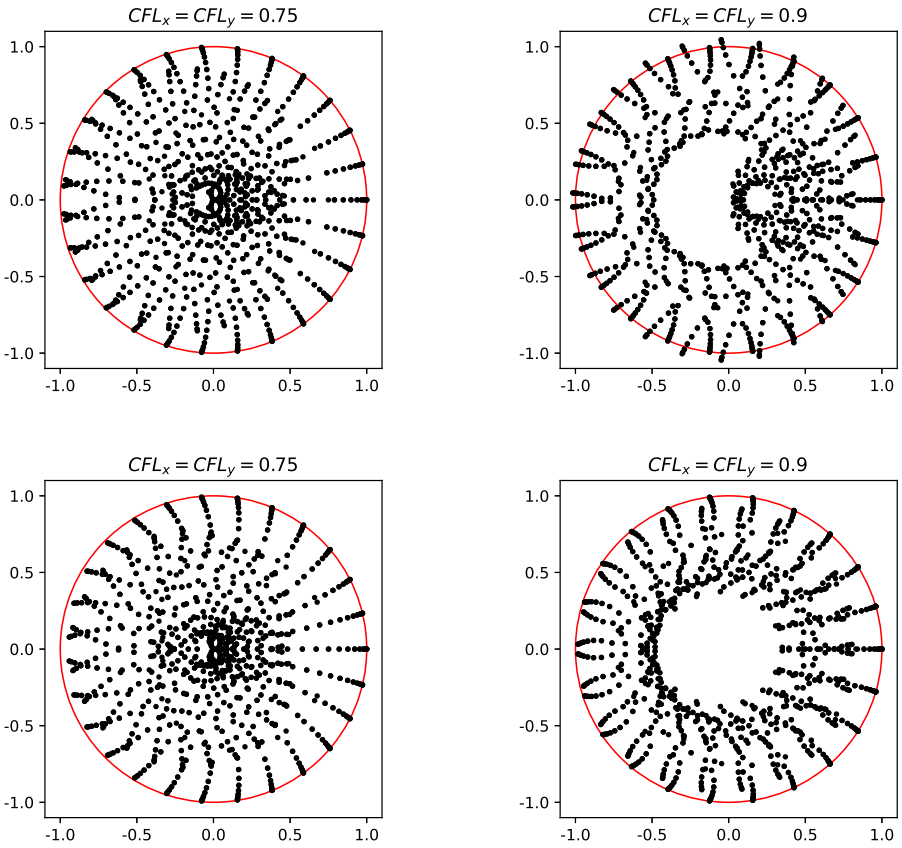


Figure 2.4: Eigenvalues for $\mathrm{CFL}_x = \mathrm{CFL}_y = 0.75, 0.9$ using Simpson's rule (top) and exact integration (bottom) [3].

Chapter 3

From linear to nonlinear conservation laws

We consider two-dimensional Burgers' equation

$$\frac{\partial}{\partial t}q(x,y,t)+\frac{\partial}{\partial x}\left(\frac{1}{2}q^2(x,y,t)\right)+\frac{\partial}{\partial y}\left(\frac{1}{2}q^2(x,y,t)\right)=0, \tag{3.1}$$

for which we in general do not have an exact evolution formula. We therefore review the iterative approach presented in [3]. Let
$q(x_{\hat{\imath}},y_{\hat{\jmath}},t),(\hat{\imath},\hat{\jmath})\in\left\{(i-\frac{1}{2},j-\frac{1}{2}),(i-\frac{1}{2},j),(i,j-\frac{1}{2})\right\}$ be some point value which needs to be evolved for a timestep τ. For q smooth, linearising (3.1) around $q':=q(x_{\hat{\imath}},y_{\hat{\jmath}},t+\tau)$ gives

$$\frac{\partial}{\partial t}q(x,y,t)+q'\frac{\partial}{\partial x}q(x,y,t)+q'\frac{\partial}{\partial y}q(x,y,t)=0, \tag{3.2}$$

for which we can use the exact evolution formula for two-dimensional linear advection. We obtain

$$\begin{aligned}q'&=q(x_{\hat{\imath}}-q'\tau,y_{\hat{\jmath}}-q'\tau,t)\\&=q(x_{\hat{\imath}},y_{\hat{\jmath}},t)-q'\tau q_x(x_{\hat{\imath}},y_{\hat{\jmath}},t)-q'\tau q_y(x_{\hat{\imath}},y_{\hat{\jmath}},t)+\mathcal{O}(\tau^2),\end{aligned} \tag{3.3}$$

which is equivalent to

$$q'=\frac{q(x_{\hat{\imath}},y_{\hat{\jmath}},t)}{1+\tau q_x(x_{\hat{\imath}},y_{\hat{\jmath}},t)+q'\tau q_y(x_{\hat{\imath}},y_{\hat{\jmath}},t)}+\mathcal{O}(\tau^2). \tag{3.4}$$

Linearising again around $q'':=\frac{q(x_{\hat{\imath}},y_{\hat{\jmath}},t)}{1+\tau q_x(x_{\hat{\imath}},y_{\hat{\jmath}},t)+q'\tau q_y(x_{\hat{\imath}},y_{\hat{\jmath}},t)}$ gives

$$\begin{aligned}&q(x_{\hat{\imath}}-q''\tau,y_{\hat{\jmath}}-q''\tau,t)\\&=q(x_{\hat{\imath}}-(q'+\mathcal{O}(\tau^2))\tau,y_{\hat{\jmath}}-(q'+\mathcal{O}(\tau^2))\tau,t)\\&=q(x_{\hat{\imath}}-q'\tau+\mathcal{O}(\tau^3),y_{\hat{\jmath}}-q'\tau+\mathcal{O}(\tau^3),t)\\&=q(x_{\hat{\imath}}-q'\tau,y_{\hat{\jmath}}-q'\tau,t)+\mathcal{O}(\tau^3)\\&=q(x_{\hat{\imath}},y_{\hat{\jmath}},t+\tau)+\mathcal{O}(\tau^3).\end{aligned} \tag{3.5}$$

We observe from (3.3) and (3.5) that each iteration increases the order of the approximation by one. This means, starting with a first order approximation q'_0 of q', the iterative approach

$$q'_l = q^{rec}(x_{\hat{\imath}} - q'_{l-1}\tau, y_{\hat{\jmath}} - q'_{l-1}\tau) \ , \ \tau \in \left\{ \frac{\Delta t}{2}, \Delta t \right\}, \ l = 1, 2, \ldots \tag{3.6}$$

gives a method of third order accuracy after two iterations, if all point values are updated in that manner. Since the piecewise quadratic reconstruction limits the achievable order of accuracy to three, we will stop here. Note that q'_2 is also the update of the point value for $\tau = \Delta t$.

In theory q'' can be computed exactly by differentiation of our cell wise defined reconstruction. However, even the one-dimensional method is unstable if the characteristic speed changes sign across a shock [17], the same holds for the two-dimensional case. An obvious initial guess would be to linearise around the point we want to update at the current time t, i.e., $q'_0 = q(x_{\hat{\imath}}, y_{\hat{\jmath}}, t)$. This initial guess is third order accurate in space and first order accurate in time. Unfortunately this approach suffers from the same instability when the characteristic speed changes sign across a shock. The reason for this instability is illustrated in Fig.3.1.

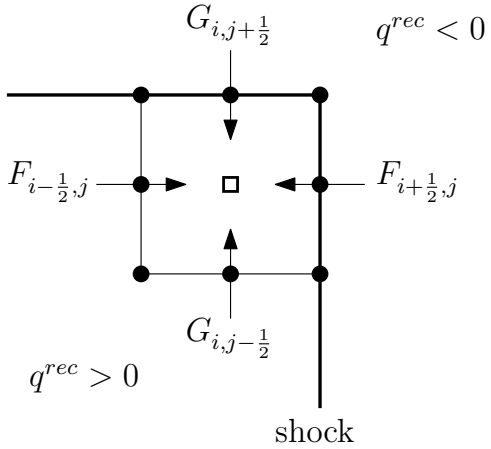


Figure 3.1: Development of an instability along a shock, where the characteristic speed changes sign.

In this scenario the absolute value of all fluxes is added to the cell average in the update process letting the cell average grow unbounded and leading to an unstable method. Similary the cell average has no lower bound, when the absolute value of all fluxes is subtracted from the cell average. We can overcome this instability by taking the cell averages of the neighbouring cells into account. We choose the mean value of the cell average values of the

cells which share the point value being updated, i.e.,

$$q_0' = \begin{cases} \frac{1}{2}(Q_{i-1,j} + Q_{i,j}) & : (\hat{\imath}, \hat{\jmath}) = (i - \frac{1}{2}, j) \\ \frac{1}{4}(Q_{i-1,j-1} + Q_{i,j-1} + Q_{i,j} + Q_{i-1,j}) & : (\hat{\imath}, \hat{\jmath}) = (i - \frac{1}{2}, j - \frac{1}{2}) \\ \frac{1}{2}(Q_{i,j-1} + Q_{i,j}) & : (\hat{\imath}, \hat{\jmath}) = (i, j - \frac{1}{2}) \end{cases} . \quad (3.7)$$

This slightly increased stencil leads to a stable scheme. The idea to linearise locally around the point values and use evolution formulas for the linearised problem can be used generally. First results for the Euler equations of gas dynamics employing evolution formulas for the linearised Euler equations are showcased in [2].

Chapter 4

Limiting

In the $1D$ advection case, limiting the reconstruction is sufficient, since the computation of the fluxes via Simpson's rule is exact. This is no longer true in more than one dimension, since the flux computation now corresponds to integration over piecewise quadratic functions, which leads to additional problems. The finite volume update itself might cause over- and undershoots and we also need to limit the fluxes to avoid these. This chapter presents two limiting approaches: A bound preserving reconstruction limiter and a positivity preserving flux limiter.

4.1 Bound preserving reconstruction

The continuous, piecewise quadratic reconstruction of the Active Flux method might introduce new extrema. This is especially the case when approximating unsteady solution structures containing shocks or contact discontinuities and leads to unphysical oscillations. To overcome this problem, we replace the original reconstruction with a discontinuous, piecewise quadratic reconstruction as illustrated in Fig.4.1. In the following we describe the limiting strategy for the two-dimensional case, which is based on Zhang and Shu [26].

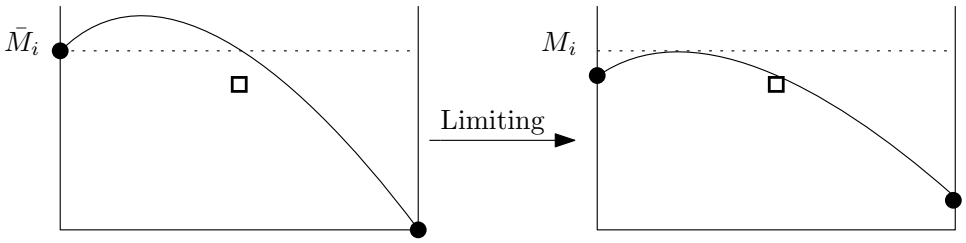


Figure 4.1: Illustration of bound-preserving limiting: The unlimited reconstruction (left) violates the maximum principle.

Lets define

$$M_{i,j} := \max_{[-1,1]^2} q_{i,j}^{rec}(\xi,\eta),\ m_{i,j} := \min_{[-1,1]^2} q_{i,j}^{rec}(\xi,\eta), \tag{4.1}$$

i.e., the maxima and minima of the reconstruction in cell $\mathcal{C}_{i,j}$ and

$$\begin{aligned} \bar{M}_{i,j} &:= \max\left\{ Q_{i+\hat{\imath},j+\hat{\jmath}} \;\middle|\; (\hat{\imath},\hat{\jmath}) \in \left\{-\frac{1}{2},0,\frac{1}{2}\right\}^2 \setminus (0,0) \right\}, \\ \bar{m}_{i,j} &:= \min\left\{ Q_{i+\hat{\imath},j+\hat{\jmath}} \;\middle|\; (\hat{\imath},\hat{\jmath}) \in \left\{-\frac{1}{2},0,\frac{1}{2}\right\}^2 \setminus (0,0) \right\}, \end{aligned} \tag{4.2}$$

i.e., the maximal and minimal point value along the boundary of cell $\mathcal{C}_{i,j}$. Replacing now cell wise the original reconstruction with a convex linear combination of reconstruction and cell average value, we define a limited reconstruction in each cell

$$q_{i,j}^{lim}(\xi,\eta) := \theta q_{i,j}^{rec}(\xi,\eta) + (1-\theta)Q_{i,j} \tag{4.3}$$

with

$$\theta := \begin{cases} \min\left\{ \left|\frac{\bar{M}_{i,j}-Q_{i,j}}{M_{i,j}-Q_{i,j}}\right|, \left|\frac{\bar{m}_{i,j}-Q_{i,j}}{m_{i,j}-Q_{i,j}}\right|, 1 \right\} & : Q_{i,j} \in [\bar{m}_{i,j},\bar{M}_{i,j}] \\ 1 & : Q_{i,j} \notin [\bar{m}_{i,j},\bar{M}_{i,j}] \end{cases}. \tag{4.4}$$

The limited reconstruction $q_{i,j}^{lim}$ in cell $\mathcal{C}_{i,j}$ has the following properties

(i) $\frac{1}{4}\int_{\mathcal{C}_{i,j}} q_{i,j}^{lim}(\xi,\eta)\,\mathrm{d}\xi\,\mathrm{d}\eta = Q_{i,j}$, i.e., we preserve the cell average.

(ii) We do not limit if $Q_{i,j} \notin [\bar{m}_{i,j},\bar{M}_{i,j}]$ in order to maintain the accuracy near local extrema of the solution structure.

(iii) $\theta = \left|\frac{\bar{M}_{i,j}-Q_{i,j}}{M_{i,j}-Q_{i,j}}\right| \Rightarrow \max\limits_{[-1,1]^2} q_{i,j}^{lim}(\xi,\eta) = \bar{M}_{i,j}$ and

$\theta = \left|\frac{\bar{m}_{i,j}-Q_{i,j}}{m_{i,j}-Q_{i,j}}\right| \Rightarrow \min\limits_{[-1,1]^2} q_{i,j}^{lim}(\xi,\eta) = \bar{m}_{i,j}$.

This property ensures not to introduce non-physical extrema in the reconstruction, i.e., the limited reconstruction is bound preserving.

Another important aspect is that the limitation will not decrease the order of our scheme as we will see in the next theorem, which is an extension to the two-dimensional case of Lemma 2.3. from [28].

Theorem 4.1.1. *The limited reconstruction $q_{i,j}^{lim}$ is a third order approximation to the unlimited reconstruction $q_{i,j}^{rec}$ for sufficiently smooth q.*

Proof: We consider $\theta = \left|\frac{\bar{M}-Q}{M-Q}\right|$ omitting the index i,j for better readability. The case $\theta = \left|\frac{\bar{m}-Q}{m-Q}\right|$ follows analogously and the case $\theta = 1$ is trivial. Since $Q \leq \bar{M} \leq M$, we have $\theta = \frac{\bar{M}-Q}{M-Q}$ and it holds

$$\begin{aligned} q^{lim}(\xi,\eta) - q^{rec}(\xi,\eta) &= \theta q^{rec}(\xi,\eta) + (1-\theta)Q - q^{rec}(\xi,\eta) \\ &= (\theta-1)(q^{rec}(\xi,\eta) - Q) \\ &= \left(\frac{\bar{M}-Q}{M-Q} - \frac{M-Q}{M-Q}\right)(q^{rec}(\xi,\eta) - Q) \\ &= \left(\bar{M}-M\right)\frac{q^{rec}(\xi,\eta)-Q}{M-Q}. \end{aligned} \tag{4.5}$$

By definition we have $\theta < 1$, i.e., an overshoot $M > \bar{M}$ with $M - \bar{M} = \mathcal{O}\left(\Delta x^3\right)$, since q^{rec} is a third order approximation. It remains to show

$$\left|\frac{q^{rec}(\xi,\eta)-Q}{M-Q}\right| \leq C \tag{4.6}$$

for some constant C. It holds

$$\max_{[-1,1]^2}\left|\frac{q^{rec}(\xi,\eta)-Q}{M-Q}\right| = \max_{[-1,1]^2}\left|\frac{q^{rec}(\xi,\eta)-Q}{\max\limits_{[-1,1]^2} q^{rec}(\xi,\eta)-Q}\right|. \tag{4.7}$$

It is sufficient to show the existence of a constant C such that

$$\left|\frac{\min\limits_{[-1,1]^2} q^{rec}(\xi,\eta)-Q}{\max\limits_{[-1,1]^2} q^{rec}(\xi,\eta)-Q}\right| \leq C. \tag{4.8}$$

This is truly the case, since $\left|\min\limits_{[-1,1]^2}(\cdot)(\xi,\eta)\right|$ and $\left|\max\limits_{[-1,1]^2}(\cdot)(\xi,\eta)\right|$ are both norms on the finite dimensional linear space spanned through our basis

polynomials modulo $Q = 0$. Hence they are equivalent and their ratio is bounded. □

In practice it is not possible to compute $M_{i,j}$ and $m_{i,j}$ exactly in two or higher-dimensional space. We therefore replace $M_{i,j}$ and $m_{i,j}$ with discrete function evaluations of our reconstruction on a $(k-1)\times(k-1)$ grid defining

$$\mathcal{D}_k := \left\{ -1 + 2\frac{s}{k} \;\middle|\; s = 1, \ldots, k-1 \right\}^2 \tag{4.9}$$

and

$$M_{i,j} \approx \max_{\mathcal{D}_k} q_{i,j}^{rec}(\xi,\eta),\ m_{i,j} \approx \min_{\mathcal{D}_k} q_{i,j}^{rec}(\xi,\eta). \tag{4.10}$$

Some important remarks:

- Theorem 4.1.1 will also hold using these approximations, since they are less restrictive, but the limiter will now not be truly bound preserving. Nevertheless we get accurate results, e.g., for $k = 6$ in case of linear advection.
- θ is computed individually for each cell. The global reconstruction is therefore not continuous at grid cell interfaces and point values are no longer unique. When computing the fluxes we need, e.g., $f(Q^n_{i-\frac{1}{2},j-\frac{1}{2}})$ and one could simply use the point value computed in the previous step. We found that this approach will cause unphysical oscillations nearby discontinuities, but we can overcome this issue in the scalar case when using the point values in upwind direction, e.g., for positive speeds we set $Q^n_{i-\frac{1}{2},j-\frac{1}{2}} = q^{lim}_{i-1,j-1}(1,1)$.
- The proof of theorem 4.1.1 does neither explicitly depend on the degree of our reconstruction nor on the dimension of the spatial space. This means that the same limiting strategy can be used if one constructs a higher order Active Flux method using, e.g., more point values for a higher polynomial reconstruction and appropriate quadrature formulas and also in any spatial dimension. However, the next theorem, which can also be found in [4], only holds in the one-dimensional case.

Theorem 4.1.2. *The one-dimensional Active Flux method for linear advection with limited reconstruction does not produce new extrema in the cell average values.*

Proof: In the one-dimensional case, the fluxes are computed exactly using Simpson's rule, since the integration takes place in a single cell with, though limited, quadratic reconstruction. Thus, the new cell average values agree with exact averages over parts of the bound preserving reconstruction and therefore are bound preserving. □

For two-dimensional linear advection we can compute the numerical fluxes exactly as explained in Section 2.2. This would lead to a bound preserving approximation of the cell average values as in the one-dimensional case. Unfortunately, this approach can not be extended to more general hyperbolic problems and it is often not sufficient to only limit the reconstruction, we need to limit the fluxes, too. Especially positivity preserving schemes are needed, when, e.g., approximating the Euler equations of gas dynamics, since pressure and density need to remain positive, or when approximating the Vlasov–Poisson equations, where the conserved quantity is a probability density function. An approach for positivity preserving flux limiting will now be explained in the next section.

4.2 Positivity preserving flux limiter

The following limiter was first proposed by Zhang and Shu [27] and further explored by Hu, Adams and Shu [18] for the approximation of the Euler equations of fluid mechanics. This scenario is more complex compared to, e.g., two-dimensional advection with a spatially and temporally varying velocity field, which we will consider here.

Lets consider a one-dimensional scalar conservation law with positive initial data

$$\begin{aligned} \frac{\partial}{\partial t} q(x,t) + \frac{\partial}{\partial x} f(q(x,t)) = 0 \ , \\ q(x,0) = q_0(x) \geq 0 \end{aligned} \tag{4.11}$$

with $f \in C^1(\mathbb{R})$ such that $q(x,t) \geq 0$ for all t. The first order Lax-Friedrichs flux is given by [26]

$$F^{LF}_{i-\frac{1}{2}} = \frac{1}{2}\left(f(Q^n_{i-1}) + f(Q^n_i) - a_0(Q^n_i - Q^n_{i-1}))\right) \tag{4.12}$$

with $a_0 > 0$. The corresponding finite volume scheme can be written as a convex combination

$$\begin{aligned} Q^{n+1}_i &= Q^n_i - \frac{\Delta t}{\Delta x}\left(F^{LF}_{i+\frac{1}{2}} - F^{LF}_{i-\frac{1}{2}}\right) \\ &= \left(1 - \frac{\Delta t}{\Delta x} a_0\right) Q^n_i + \frac{\Delta t}{2\Delta x} a_0 \left(Q^n_{i+1} - \frac{1}{a_0} f(Q^n_{i+1})\right) \\ &\qquad + \frac{\Delta t}{2\Delta x} a_0 \left(Q^n_{i-1} + \frac{1}{a_0} f(Q^n_{i-1})\right). \end{aligned} \tag{4.13}$$

A scheme $Q^{n+1}_i = H(Q^n_{i-1}, Q^n_i, Q^n_{i+1})$ is monotone iff $\frac{\partial}{\partial Q^n_{i+\hat{\imath}}} H(Q^n_{i-1}, Q^n_i, Q^n_{i+1}) \geq 0$, $\hat{\imath} = -1, 0, 1$ [16]. Choosing a_0 as the maximum

signal speed, i.e., $a_0 = \max_i |f'(Q_i^n)|$, gives a monotone scheme for CFL ≤ 1, since for $H(Q_{i-1}^n, Q_i^n, Q_{i+1}^n)$ defined as the right hand side of (4.13) we have

$$\begin{aligned}
\frac{\partial}{\partial Q_i^n} H(Q_{i-1}^n, Q_i^n, Q_{i+1}^n) &= 1 - \frac{\Delta t}{\Delta x} a_0 \geq 0 \ , \\
\frac{\partial}{\partial Q_{i+1}^n} H(Q_{i-1}^n, Q_i^n, Q_{i+1}^n) &= \frac{\Delta t}{2\Delta x} a_0 \left(1 - \frac{1}{a_0} f'(Q_{i+1}^n)\right) \geq 0 \ , \\
\frac{\partial}{\partial Q_{i-1}^n} H(Q_{i-1}^n, Q_i^n, Q_{i+1}^n) &= \frac{\Delta t}{2\Delta x} a_0 \left(1 + \frac{1}{a_0} f'(Q_{i-1}^n)\right) \geq 0 \ .
\end{aligned} \tag{4.14}$$

Let m and M be the minimum and maximum of the exact solution to (4.11) and $Q_i^n \in [m, M]$, then

$$m = H(m, m, m) \leq Q_i^{n+1} \leq H(M, M, M) = M. \tag{4.15}$$

The idea is now to replace the Active Flux fluxes with convex linear combinations of Active Flux and first order Lax-Friedrichs fluxes

$$F_{i-\frac{1}{2}}^{lim} := \gamma F_{i-\frac{1}{2}} + (1-\gamma) F_{i-\frac{1}{2}}^{LF} \tag{4.16}$$

choosing γ in a way that adds just enough numerical dissipation to obtain a positivity preserving scheme. For that matter we rewrite the update of the cell averages (1.5) as a convex combination

$$Q_i^{n+1} = \frac{1}{2}\left(Q_i^n + 2\frac{\Delta t}{\Delta x} F_{i-\frac{1}{2}}\right) + \frac{1}{2}\left(Q_i^n - 2\frac{\Delta t}{\Delta x} F_{i+\frac{1}{2}}\right). \tag{4.17}$$

With

$$Q_i^- := Q_i^n + 2\frac{\Delta t}{\Delta x} F_{i-\frac{1}{2}} \ , \quad Q_i^+ := Q_i^n - 2\frac{\Delta t}{\Delta x} F_{i+\frac{1}{2}} \ , \tag{4.18}$$

a sufficient condition for a positivity preserving scheme is $Q_i^{\pm} \geq 0$.

We will now derive a γ that ensures these inequalities.

Let $0 < \epsilon \ll 1$, then

$$\begin{aligned} Q_i^{lim,-} &:= Q_i^n + 2\frac{\Delta t}{\Delta x} F_{i-\frac{1}{2}}^{lim} \\ &= Q_i^n + 2\frac{\Delta t}{\Delta x}\left(F_{i-\frac{1}{2}}^{LF} + \gamma^-\left(F_{i-\frac{1}{2}} - F_{i-\frac{1}{2}}^{LF}\right)\right) \\ &\geq \epsilon \end{aligned} \tag{4.19}$$

is for $F_{i-\frac{1}{2}} < F_{i-\frac{1}{2}}^{LF}$ equivalent to

$$\gamma^- \leq \frac{1}{2\frac{\Delta t}{\Delta x}} \frac{\epsilon - (Q_i^n + 2\frac{\Delta t}{\Delta x} F_{i-\frac{1}{2}}^{LF})}{F_{i-\frac{1}{2}} - F_{i-\frac{1}{2}}^{LF}}. \tag{4.20}$$

Furthermore we have

$$2\frac{\Delta t}{\Delta x}\left(F_{i-\frac{1}{2}} - F_{i-\frac{1}{2}}^{LF}\right) = Q_i^- - Q_i^{LF,-} \ , \tag{4.21}$$

where similary to (4.18) $Q_i^{LF,-} := Q_i^n + 2\frac{\Delta t}{\Delta x} F_{i-\frac{1}{2}}^{LF}$.
Thus we obtain

$$\gamma^- \leq \frac{\epsilon - Q_i^{LF,-}}{Q_i^- - Q_i^{LF,-}}. \tag{4.22}$$

This motivates the definitions

$$\gamma^- := \begin{cases} \frac{\epsilon - Q_i^{LF,-}}{Q_i^- - Q_i^{LF,-}} & : Q_i^- < \epsilon \\ 1 & : Q_i^- \geq \epsilon \end{cases}, \quad \gamma^+ := \begin{cases} \frac{\epsilon - Q_{i-1}^{LF,+}}{Q_{i-1}^+ - Q_{i-1}^{LF,+}} & : Q_{i-1}^+ < \epsilon \\ 1 & : Q_{i-1}^+ \geq \epsilon \end{cases} \tag{4.23}$$

with $\epsilon := \min\{10^{-13}, \min q_0(x)\}$. Finally we define

$$\gamma := \min(\gamma^-, \gamma^+). \tag{4.24}$$

Via construction, finite volume methods of the form (1.5) with limited fluxes (4.16) enforce positivity.

What is left to check is that our flux limiter does not affect the order of accuracy of our scheme. As we will see in the next theorem, which is based on [18], this is indeed the case, if ϵ is defined sufficiently small.

Theorem 4.2.1. *The limited fluxes $F_{i-\frac{1}{2}}^{lim}$ are third order accurate approximations to the original fluxes $F_{i-\frac{1}{2}}$ for q sufficiently smooth, i.e.,*

$$\left|F_{i-\frac{1}{2}}^{lim} - F_{i-\frac{1}{2}}\right| = \mathcal{O}(\Delta x^3). \tag{4.25}$$

Proof: It holds

$$\left|F^{lim}_{i-\frac{1}{2}} - F_{i-\frac{1}{2}}\right| = (1-\gamma)\left|F_{i-\frac{1}{2}} - F^{LF}_{i-\frac{1}{2}}\right|, \tag{4.26}$$

where $\left|F_{i-\frac{1}{2}} - F^{LF}_{i-\frac{1}{2}}\right| = \mathcal{O}(\Delta x)$. Both fluxes $F_{i-\frac{1}{2}}$ and $F^{LF}_{i-\frac{1}{2}}$ are bounded in smooth regions, this means it is enough to show $1-\gamma = \mathcal{O}(\Delta x^2)$. We consider $\gamma = \frac{\epsilon - Q^{LF,+}}{Q^+ - Q^{LF,+}}$, i.e., $Q^+ < \epsilon$, omitting the indices $i, i-1$ and $i+\frac{1}{2}$ here and in the following. The other case follows analogously. We then have

$$1-\gamma = \frac{Q^+ - \epsilon}{Q^+ - Q^{LF,+}} \approx \frac{Q^+}{Q^+ - Q^{LF,+}} \leq \frac{|Q^+|}{Q^{LF,+}}\ , \tag{4.27}$$

since ϵ is negligibly small. We already know that

$$Q^n - \frac{\Delta t}{\Delta x}\left(F^{LF}_{i+\frac{1}{2}} - F^{LF}_{i-\frac{1}{2}}\right) \geq m > 0 \tag{4.28}$$

is bounded away from zero by the minimum of the initial data m, since the first order Lax-Friedrichs method is bound preserving for CFL ≤ 1. This is also true for $F^{LF}_{i-\frac{1}{2}} = 0$, hence $Q^{LF,+} = Q^n - 2\frac{\Delta t}{\Delta x}F^{LF} \geq m > 0$ for CFL ≤ 0.5. Furthermore we have

$$\begin{aligned} Q^+ &= Q^n - 2\frac{\Delta t}{\Delta x}F \\ &= Q^n - 2\frac{\Delta t}{\Delta x}F^{exa} + \mathcal{O}(\Delta x^2) \end{aligned} \tag{4.29}$$

and since the exact flux F^{exa} is also bound preserving, we get with $Q^+ < \epsilon$ close to zero or negative

$$|Q^+| \leq \left|Q^+ - \left(Q^n - 2\frac{\Delta t}{\Delta x}F^{exa}\right)\right| = \mathcal{O}(\Delta x^2)\ , \tag{4.30}$$

which completes the proof. □

Some important remarks:

- The computation of γ is ill-conditioned and accumulating round-off errors may lead to non-positive solutions. It is therefore advisable to cut-off γ, i.e., we set

$$\gamma := \max(\min(\gamma^-, \gamma^+, 1), 0).$$

- Though the limited fluxes enforce positivity of the cell averages, it may still be necessary to also limit the reconstruction, e.g., linearising around point values with negative density for the solution of the Euler equations causes segmentation errors [2].

- As we see from the proof of theorem 4.2.1, a more restrictive condition CFL ≤ 0.5 is introduced when using the positivity preserving flux limiter.

- This limiting strategy is adaptable for any spatial dimension via a simple dimension-by-dimension approach, e.g., in the two-dimensional case we again rewrite the update of the cell averages (1.17)

$$\begin{aligned} Q_{i,j}^{n+1} = &\frac{\alpha_x}{2}\left(Q_{i,j}^n + 2\frac{\Delta t}{\alpha_x \Delta x} F_{i-\frac{1}{2},j}\right) + \frac{\alpha_x}{2}\left(Q_{i,j}^n - 2\frac{\Delta t}{\alpha_x \Delta x} F_{i+\frac{1}{2},j}\right) \\ &+ \frac{\alpha_y}{2}\left(Q_{i,j}^n + 2\frac{\Delta t}{\alpha_y \Delta y} G_{i,j-\frac{1}{2}}\right) + \frac{\alpha_y}{2}\left(Q_{i,j}^n - 2\frac{\Delta t}{\alpha_y \Delta y} G_{i,j+\frac{1}{2}}\right) \end{aligned} \tag{4.31}$$

  with $\alpha_x + \alpha_y = 1$ and $\alpha_x, \alpha_y > 0$. The four summands are then limited in the same way as in the one-dimensional case.

- In the non-scalar case we limit the fluxes consecutively, e.g., if $f : \mathbb{R}^2 \to \mathbb{R}^2$ we compute γ_1 for the first component and set

$$F_{i-\frac{1}{2}}^* = \gamma_1 F_{i-\frac{1}{2}} + (1-\gamma_1) F_{i-\frac{1}{2}}^{LF} \ . \tag{4.32}$$

  We then compute γ_2 for the second component of

$$Q_i^{*,-} := Q_i^n + 2\frac{\Delta t}{\Delta x} F_{i-\frac{1}{2}}^* \ , \quad Q_i^{*,+} := Q_i^n - 2\frac{\Delta t}{\Delta x} F_{i+\frac{1}{2}}^* \tag{4.33}$$

  and let

$$\begin{aligned} F_{i-\frac{1}{2}}^{lim} &= \gamma_2 F_{i-\frac{1}{2}}^* + (1-\gamma_2) F_{i-\frac{1}{2}}^{LF} \\ &= \gamma_1\gamma_2 F_{i-\frac{1}{2}} + (1-\gamma_1\gamma_2) F_{i-\frac{1}{2}}^{LF} \ , \end{aligned} \tag{4.34}$$

  which is again a convex combination of the two fluxes.

4.3 Numerical results

Example 1: To test the behaviour of the two different limiting strategies, we consider two-dimensional advection with a spatially and temporally varying velocity field

$$\frac{\partial}{\partial t} q(x,y,t) + \frac{\partial}{\partial x}(a(x,y,t)q(x,y,t)) + \frac{\partial}{\partial y}(b(x,y,t)q(x,y,t)) = 0, \tag{4.35}$$

where the velocities are defined via a stream function $\Psi(x,y,t)$ through

$$\begin{aligned} a(x,y,t) &= -\frac{\partial}{\partial y}\Psi(x,y,t) \\ b(x,y,t) &= \ \ \frac{\partial}{\partial x}\Psi(x,y,t). \end{aligned} \tag{4.36}$$

By doing so the velocity field becomes divergence free and (4.35) is equivalent to

$$\frac{\partial}{\partial t}q(x,y,t) + a(x,y,t)\frac{\partial}{\partial x}q(x,y,t) + b(x,y,t)\frac{\partial}{\partial y}q(x,y,t) = 0. \tag{4.37}$$

The characteristics can then be approximated using an ODE solver of sufficiently high order as described in [1].

We now approximate the solution to (4.37) for

$$\Psi(x,y,t) = \frac{4}{3}\left((x-1)^2 + (y-1)^2\right)^{\frac{3}{2}}\cos(\pi t) \tag{4.38}$$

on the domain $[0,1.5]^2$ with CFL $= 0.25$ and 128×128 grid points. The factor $\cos(\pi t)$ leads to a solution which is 2-periodic in time and equal to the initial data for all $t \in \mathbb{N}$. The initial condition is given by

$$q_0(x,t) = \begin{cases} 1 & : (x,y) \in [0.5,1]^2 \\ 10^{-10} & : (x,y) \notin [0.5,1]^2 \end{cases}.$$

Figure 4.2 shows the solutions at times $t = 0.5$ (top) and $t = 1$ (bottom) using (from left to right) no limiter, bound preserving reconstruction, the positivity preserving flux limiter and a combination of both limiting strategies. The bound preserving reconstruction is computed on $\mathcal{D}_8$ as defined in (4.9). Table 4.1 shows the corresponding minimal Q^{min} and maximal Q^{max} cell average values of the four methods at time $t = 1$.

	NoLim	BP	PP	BP+PP
Q^{min}	-7.2489×10^{-2}	-5.4609×10^{-6}	8.4992×10^{-14}	5.2106×10^{-6}
Q^{max}	1.0885	9.9999×10^{-1}	1.0878	9.9999×10^{-1}

Table 4.1: Minimal Q^{min} and maximal Q^{max} cell average value of the solutions to (4.37) at time $t = 1$ using no limiter (NoLim), bound preserving reconstruction (BP), the positivity preserving flux limiter (PP) and a combination of both limiters (BP+PP).

Without any limiter the method produces over- and undershoots of size $\pm 10^{-2}$ while successfully capturing the solution structure.

Using only the bound preserving reconstruction reduces the undershoots to size $\pm 10^{-6}$ and cuts off the overshoots entirely, but clearly at the cost of a smeared solution, meaning that this limiter introduces a lot of numerical

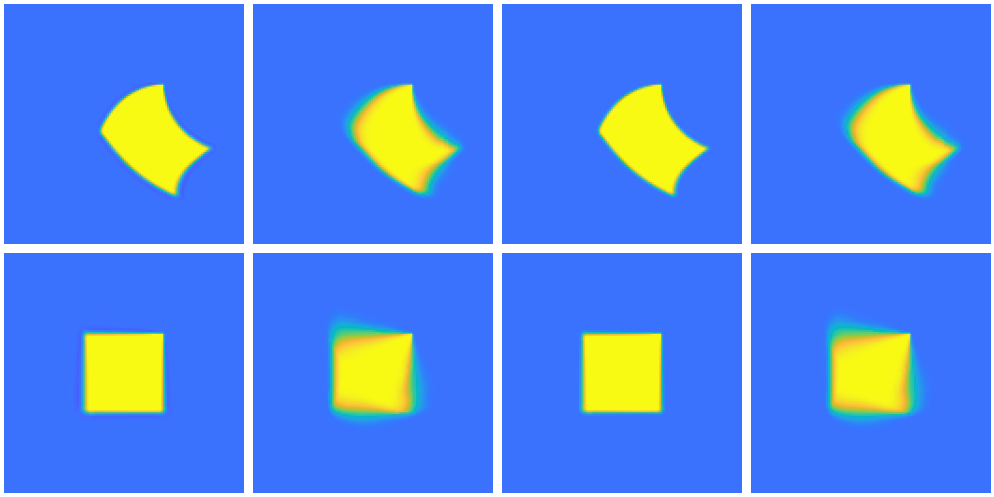


Figure 4.2: Solutions for *Example 1* at times $t = 0.5$ (top) and $t = 1$ (bottom) using (from left to right) no limiter, bound preserving reconstruction, the positivity preserving flux limiter and a combination of both limiters.

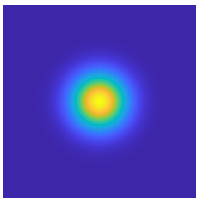
dissipation even though the extremas of each cell are approximated through values of the reconstruction on $\mathcal{D}_8$ and the limiter is less restrictive than a truly bound preserving one.

As one might expect the undershoots are cut off entirely when using only the flux limiter while leaving the overshoots untouched. Via construction this method is almost as dissipative as the unlimited one and hence captures the solution structure successfully, too, which is confirmed in this test case.

The combination of both limiting strategies gives a bound preserving scheme cutting of over- and undershoots, but again the reconstruction limiter leads to a dissipative scheme.

Example 2: To test the accuracy of the bound preserving reconstruction limiter we consider two-dimensional Burgers' equation with initial data, where the characteristic speed changes sign along a shock wave,

$$q_0(x,t) = \exp(-50\left(\left(x-\frac{1}{2}\right)^2 + \left(y-\frac{1}{2}\right)^2\right) - \frac{1}{5}$$



on the domain $[0,1]^2$ with CFL $= 0.7$ and compute the EOC for grids of size $m \times m$, $m = 64, 128, 256$ at an early time $t = 0.04$, where the solution is still smooth. The error_m for grid $m \times m$ is approximated in the L_1-norm by comparison with the cell averages of grid $2m \times 2m$.

We then compute the EOC_m via

$$\mathrm{EOC}_m = \frac{1}{\log\left(\frac{1}{2}\right)} \log\left(\frac{\mathrm{error}_{\frac{m}{2}}}{\mathrm{error}_m}\right). \tag{4.39}$$

The results without limitation and with bound preserving reconstruction on $\mathcal{D}_{10}$ are shown in table 4.2 and confirm third order accuracy of both methods, while their errors are comparable. Figure 4.4 shows the solutions

	NoLim		BP	
m	Error	EOC	Error	EOC
64	9.4147×10^{-7}	—	1.5357×10^{-6}	—
128	1.3129×10^{-7}	2.8422	2.3600×10^{-7}	2.7021
256	1.7534×10^{-8}	2.9044	3.0493×10^{-8}	2.9522

Table 4.2: Accuracy study for smooth solutions of the two-dimensional Burgers equation using the iterative approach with unlimited (NoLim) and limited reconstruction (BP).

for $m = 128$ at time $t = 1$ without limiter (left) and with bound preserving reconstruction (right). The different shades come from spurious oscillations in the unlimted case which are visible, e.g., as a dark blue layer nearby the centre of the domain. The limitation successfully avoids these oscillations while smoothing the solution.

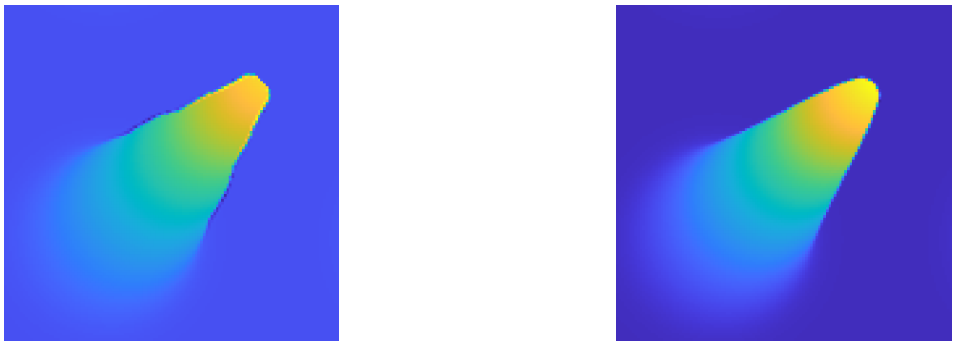


Figure 4.3: Solutions for *Example 2* at time $t = 1$ without limiter (left) and with bound preserving reconstruction (right).

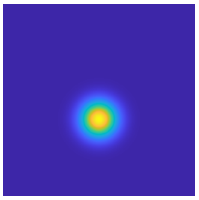
Example 3: To test the accuracy of the positivity preserving flux limiter we again consider two-dimensional advection with a spatially and temporally

varying velocity field (4.37) this time with the stream function

$$\Psi(x, y, t) = \frac{1}{\pi} \sin(\pi x)^2 \sin(\pi y)^2 \cos(\pi t)\,. \tag{4.40}$$

The solution is then 2-periodic in time and matches the initial data for all times $t \in \mathbb{N}$. The initial data are set to

$$q_0(x, t) = \exp(-118 \left(\left(x - \frac{1}{2} \right)^2 + \left(y - \frac{2}{5} \right)^2 \right) + 10^{-12}$$



on the domain $[0, 1]^2$. We then compute the error and EOC at time $t = 1$ for grids of size $m \times m$, $m = 64, 128, 256$ with CFL $= 0.25$ without limiter and with positivity preserving flux limiter. For this particular example it is sufficient to use the flux limiter only. The results are shown in tables 4.3 together with the minimal cell average Q^{min}. The errors are again com-

NoLim			
m	Error	EOC	Q^{min}
64	6.2119×10^{-5}	—	-4.4528×10^{-8}
128	8.1943×10^{-6}	2.9223	-1.1744×10^{-11}
256	1.0499×10^{-6}	2.9644	-5.3774×10^{-12}

PP			
m	Error	EOC	Q^{min}
64	1.1217×10^{-4}	—	1.0054×10^{-13}
128	1.4448×10^{-5}	2.9568	2.0239×10^{-13}
256	1.6353×10^{-6}	3.1432	3.0896×10^{-13}

Table 4.3: Accuracy study and Q^{min} for advection with a spatially and temporally varying velocity field using the iterative approach without limiter (NoLim) and with positivity preserving flux limiter (PP).

parable and both methods are third order accurate, while the flux limiter successfully avoids negative cell averages.

Fig.4.4 shows the solutions at time $t = 0.5$ for $m = 256$, which give the impression to be almost identical. However only the right one is non-negative.

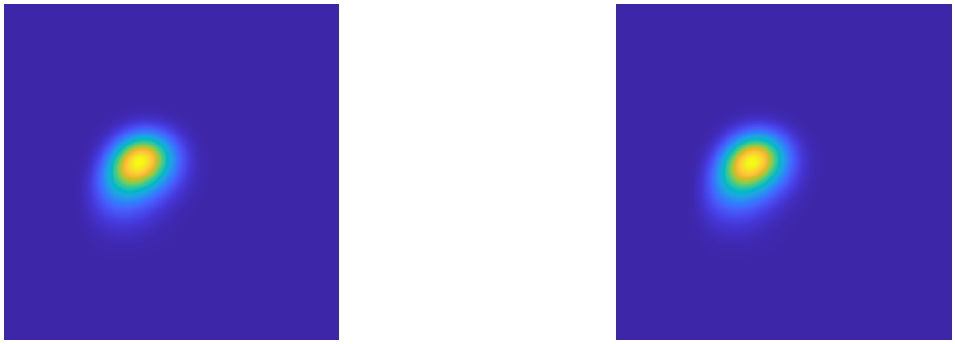


Figure 4.4: Solutions for *Example 3* at time $t = 0.5$ without limiter (left) and with positivity preserving flux limiter (right).

Example 4: Finally we show first results for our three-dimensional scheme, which have not been published yet. We consider three-dimensional advection with a spatially and temporally varying velocity field

$$\begin{aligned} &\frac{\partial}{\partial t}q(x,y,z,t) + \frac{\partial}{\partial x}(a(x,y,z,t)q(x,y,z,t)) \\ &+\frac{\partial}{\partial y}(b(x,y,z,t)q(x,y,z,t)) + \frac{\partial}{\partial z}(c(x,y,z,t)q(x,y,z,t)) = 0. \end{aligned} \tag{4.41}$$

To gain a divergence free velocity field, we modify (4.40) and define the 2-time periodic three-dimensional stream function

$$\Psi(x,y,z,t) = \frac{1}{\pi^2}\sin(\pi x)^2\sin(\pi y)^2\sin(\pi z)^2\cos(\pi t). \tag{4.42}$$

With

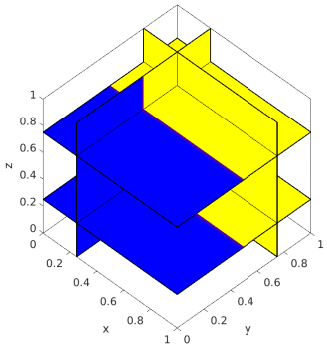
$$\begin{aligned} a(x,y,z,t) &:= 2\frac{\partial^2}{\partial y \partial z}\Psi(x,y,z,t) \\ b(x,y,z,t) &:= -\frac{\partial^2}{\partial x \partial z}\Psi(x,y,z,t) \\ a(x,y,z,t) &:= -\frac{\partial^2}{\partial x \partial y}\Psi(x,y,z,t) \end{aligned} \tag{4.43}$$

(4.41) is equivalent to

$$\begin{aligned} &\frac{\partial}{\partial t}q(x,y,z,t) + a(x,y,z,t)\frac{\partial}{\partial x}q(x,y,z,t) \\ &+b(x,y,z,t)\frac{\partial}{\partial y}q(x,y,z,t) + c(x,y,z,t)\frac{\partial}{\partial z}q(x,y,z,t) = 0 \end{aligned} \tag{4.44}$$

and characteristics can be traced back using a sufficiently high order ODE solver to update our point values. Figure 4.5 shows the behaviour of the solution for initial data

$$q_0(x,y,z) = \begin{cases} 1 & : x < \frac{1}{2} \\ 0 & : x \geq \frac{1}{2} \end{cases}$$



on the domain $[0,1]^3$ using a $64 \times 64 \times 64$ grid with triple periodic boundary conditions and CFL $= 0.7$. The solution is shown at times $t = 0.5, 1, 1.5, 2$ (from left to right) by means of slices $x = \frac{1}{4}$, $y = \frac{3}{4}$, $z = \frac{1}{4}, \frac{3}{4}$ using no limiter (top) and a three-dimensional version of the bound preserving reconstruction (bottom), where we approximate extrema with evaluations of our reconstruction on $\mathcal{E}_k := \left\{ -1 + 2\frac{s}{k} \;\middle|\; s = 1, \ldots, k-1 \right\}^3$ with $k = 20$.

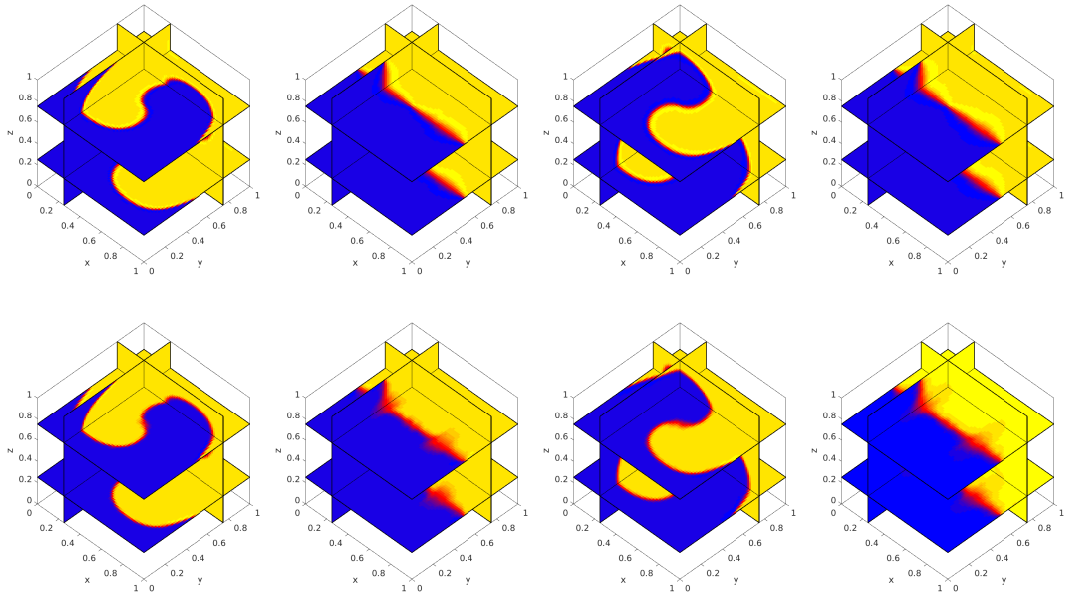


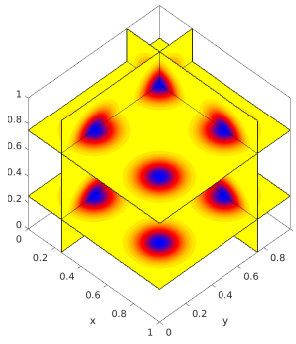
Figure 4.5: Solution to (4.44) with stream function (4.42) and Riemann initial data at times $t = 0.5, 1, 1.5, 2$ (from left to right) using no limiter (top) and the bound preserving reconstruction (bottom).

Like in the two-dimensional case the limiter leads to a more dissipative solution, but as we see from the second column undershoots indicated by a

lighter yellow and overshoots indicated by a lighter blue (top) can be sufficiently eliminated (bottom).

To test the accuracy of the method we define a bump function $\beta(x,y,z) := \exp\left(-80\left(x^2+y^2+z^2\right)\right)$ and consider smooth initial data of the form

$$q_0(x,y,z) = \sum_{(\hat{x},\hat{y},\hat{z})\in\left\{\frac{1}{4},\frac{3}{4}\right\}^3} \beta(x-\hat{x},y-\hat{y},z-\hat{z}).$$



The time periodic factor $\cos(\pi t)$ in (4.42) is replaced by $\cos(2\pi t)$ to retain a smooth solution. We then compute errors at time $t = 0.5$ using CFL $= 0.7$, where the solution equals the initial condition, for grids of size $m \times m \times m$, $m = 32, 64, 128, 256$. The results are summarized in table 4.4 and confirm third order accuracy of our method. Figure 4.6 shows the solution for $m = 256$ at times $t = 0.25, 0.5$.

m	Error	EOC
32	6.5259×10^{-4}	—
64	1.0156×10^{-4}	2.6839
128	1.4091×10^{-5}	2.8495
256	1.8538×10^{-6}	2.9263

Table 4.4: Accuracy study for (4.44) with a modified version of the stream function (4.42) at time $t = 0.5$.

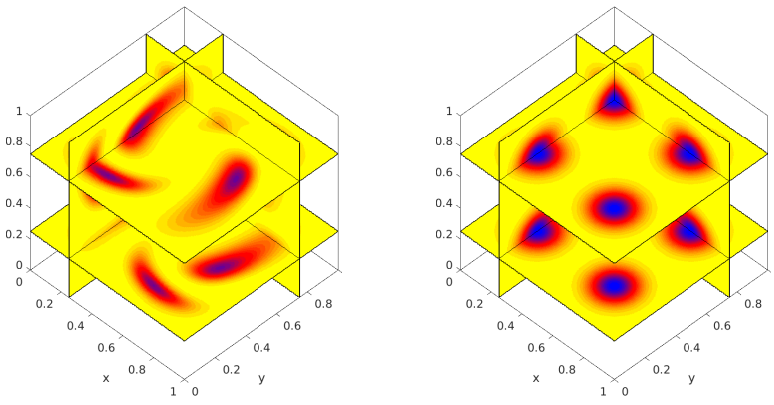


Figure 4.6: Solution to (4.44) with modified version of the stream function (4.42) and smooth initial data at times $t = 0.25$ (left) and $t = 0.5$ (right).

Appendix A

Statement about the Authors Contribution to Previous Work

In the first attached paper "The Cartesian Grid Active Flux Method with Adaptive Mesh Refinement" [1] the first implementation of the Active Flux method on adaptively refined Cartesian grids was presented. For that matter we implemented our method as a new solver in ForestClaw, a parallel algorithm for patch-based adaptive mesh refinement on a forest of quadtrees. The theoretical results 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. The authors Co-supervisor, Donna Calhoun, who is the main developer of ForestClaw, helped a lot with the installation and implementation and was a great support in the process of debugging the code. This enables us to apply the Active Flux method to complex applications.

In the second attached paper [2] we derived Active Flux methods for hyperbolic systems using the method of bicharacteristics. The theoretical results were derived by Christiane Helzel and the author of this thesis in equal parts. All numerical computations were performed by the author of this thesis. Mária Lukáčová-Medvid'ová's expertise in bicharacteristics theory was a crucial part in the development of the presented methods.

We ordered the authors alphabetically in both papers to recognise these different contributions equally.

Appendix B

Acknowledgment

This work was supported and by a subcontract with Boise State University, Boise, ID, under DARPA Cooperative Agreement HR00112120003 with Embry-Riddle Aeronautical University, Daytona Beach, FL, USA, and by the German Research Foundation through HE 4858/4–2.

Bibliography

[1] Calhoun, D., Chudzik, E., Helzel, C.: The Cartesian grid active flux method with adaptive mesh refinement. J. Sci. Comput. 94 (2023).

[2] Chudzik, E., Helzel, C., Lukáčová-Medvid'ová, M.: Active Flux Methods for Linear Hyperbolic Systems using the Method of Bicharacteristics. Accepted for publication in J. Sci. Comput. (2024).

[3] Chudzik, E., Helzel, C., Kerkmann, D.: The Cartesian Grid Active Flux Method: Linear stability and bound preserving limiting. Appl. Math. Comput. 393, pp. 125501 (2021).

[4] Chudzik, E., Helzel, C.: A Review of Cartesian Grid Active Flux Methods for Hyperbolic Conservation Laws. In: Franck, E., Fuhrmann, J., Michel-Dansac, V., Navoret, L. (eds) Finite Volumes for Complex Applications X-Volume 1, Elliptic and Parabolic Problems, FVCA 2023, Springer Proceedings in Mathematics & Statistics, vol 432, Springer, Cham, pp. 93–109 (2023).

[5] Kiechle, Y.-F., Chudzik, E., Helzel, C.: An active flux method for the Vlasov-Poisson system. In: Franck, E., Fuhrmann, J., Michel-Dansac, V., Navoret, L. (eds) Finite Volumes for Complex Applications X-Volume 2, Hyperbolic and Related Problems, FVCA 2023, Springer Proceedings in Mathematics & Statistics, vol 433, Springer, Cham, pp. 93–101 (2023).

[6] Abgrall, R.: A combination of residual distribution and the active flux formulation or a new class of schemes that can combine several writings of the same hyperbolic problem: Application to the 1d Euler equations. Commun. Appl. Math. Comput. 5, pp. 370–402 (2023).

[7] Abgrall, R., Barsukow, W.: Extensions of active flux to arbitrary order of accuracy. ESAIM: Mathematical Modelling and Numerical Analysis, № 2, pp. 991–1027 (2023).

[8] Barsukow, W.: The active flux method for nonlinear problems. J. Sci. Comput. 86:3 (2021).

[9] Barsukow, W., Hohm, J., Klingenberg, C., Roe, P.L.: The active flux scheme on Cartesian grids and its low Mach number limit. J. Sci. Comput. 81, pp. 594–622 (2019).

[10] Berger, M.J., Colella, P.: Local adaptive mesh refinement for shock hydrodynamics. J. Comput. Phys. 82, pp. 64–84 (1989)

[11] Berger, M.J., Oliger, J.: Adaptive mesh refinement for hyperbolic partial differential equations. J. Comput. Phys. 53, pp. 484–512 (1984).

[12] Calhoun, D., Burstedde, C.: ForestClaw: a parallel algorithm for patch-based adaptive mesh refinement on a forest of quadtrees. arXiv: 1703.03116 (2017).

[13] van Drosselaer, J., Kraaijevanger, J., Spijker, M.: Linear stability analysis in the numerical solution of initial value problems. Acta Numer. 2, pp. 199–237 (1993).

[14] Eymann, T.A., Roe, P.L.: Active flux schemes for systems. AIAA 2011-3840 (2011).

[15] Eymann, T.A., Roe, P.L.: Multidimensional active flux schemes. AIAA Conference Paper, June 2013.

[16] Harten, A., Hyman, J.M., Lax, P.D.: On finite-difference approximations and entropy conditions for shocks, Communications on pure and applied mathematics, Vol. XXIX (1976), pp. 297–322

[17] Helzel, C., Kerkmann, D., Scandurra, L.: A new ADER method inspired by the active flux method. J. Sci. Comput. 80, pp. 1463–1497 (2019).

[18] Hu, X.Y., Adams, N.A., Shu, C.-W.: Positivity-preserving method for high-order conservative schemes solving compressible Euler equations. J. Comput. Phys. 242, pp. 169–180 (2013).

[19] van Leer, Bram: Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. Journal of computational physics 23(3) , pp. 276–299 (1977).

[20] Lukáčová-Medvid'ová, M., Morton, K.W., Warnecke, G.: Evolution Galerkin methods for hyperbolic systems in two space dimensions. Math. Comput. 69, pp. 1355–1384 (2000).

[21] Lukáčová-Medvid'ová, M., Saibertová, J., Warnecke, G.: Finite volume evolution Galerkin methods for nonlinear hyperbolic systems. J. Comput. Phys. 183, pp. 533–562 (2002).

[22] Roe, P.: Is discontinuous reconstruction really a good idea? J. Sci. Comput. 73, pp. 1094–1114 (2017).

[23] Roe, P.: Designing CFD methods for bandwidth - A physical approach. Comput. Fluids 214, pp. 104774 (2021)

[24] Roe, P.L., Lung, T.B., Maeng, J.: New approaches to Limiting. 2015 22nd AIAA Computational Fluid Dynamics Conference.

[25] Roe, P.L., Maeng, J., Fan, D.: Comparing active flux and discontinuous Galerkin methods for compressible flow. 2018 AIAA Aerospace Science Meeting.

[26] Zhang, X., Shu, C.-W.: On maximum-principle-satisfying high order schemes for scalar conservation laws, J. Comput. Phys. 229, pp. 3091–3120 (2010).

[27] Zhang, X., Shu, C.-W.: On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes, J. Comput. Phys. 229, pp. 8918–8934 (2010).

[28] Zhang, X., Shu, C.-W.: Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. Proc. R. Soc. A 467, pp. 2752–2776 (2011).