Extensions of Active Flux to arbitrary order of accuracy

Rémi Abgrall, Wasilij Barsukow

Extensions of Active Flux to arbitrary order of accuracy

Rémi Abgrall[1], Wasilij Barsukow[2]

Abstract

Active Flux is a recently developed numerical method for hyperbolic conservation laws. Its classical degrees of freedom are cell averages and point values at cell interfaces. These latter are shared between adjacent cells, leading to a globally continuous reconstruction. The update of the point values includes upwinding, but without solving a Riemann Problem. The update of the cell average requires a flux quadrature at the cell interface, which can be immediately performed using the point values. This paper explores different extensions of Active Flux to arbitrarily high order of accuracy, while maintaining the idea of global continuity. We propose to either increase the stencil while keeping the same degrees of freedom, or to increase the number of point values, or to include higher moments as new degrees of freedom. These extensions have different properties, and reflect different views upon the relation of Active Flux to the families of Finite Volume, Finite Difference and Finite Element methods.

Keywords: Active Flux, high order methods, conservation laws

Mathematics Subject Classification (2010): 65M06,65M08,65M60,76N99

1 Introduction

Solutions to the initial value problem for hyperbolic $m \times m$ systems of conservation laws

$$\partial_t q + \partial_x f(q) = 0 \qquad q\colon \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}^m \tag{1}$$

$$f\colon \mathbb{R}^m \to \mathbb{R}^m \tag{2}$$

generically develop discontinuities in finite time, even if the initial data are smooth. The natural setting therefore are weak solutions. For convergence to the weak solution of (1), a numerical method needs to be conservative (Lax-Wendroff theorem).

A popular way to derive numerical methods (due to Godunov) is to introduce discontinuities at every cell interface (reconstruction step), and to evolve piecewise constant data over a short period of time (Riemann solver). Even if a more complicated reconstruction is used inside the cell, or if an approximate evolution is used instead of the exact one, the main idea of Godunov's approach remains to introduce discontinuities at every cell interface. This approach has also been adapted to Finite Element methods, which has led to the development of Discontinuous Galerkin (DG) methods. These methods are conservative.

However, convergence to the weak solution is not enough in practice. In view of the big computational effort associated with grid refinement (particularly in multi-d), there is ongoing interest in guaranteeing properties of numerical solutions for coarse grids already. Some of them are essential for the simulation to continue running, e.g. non-negativity of the density. Other shortcomings might not cause a simulation to crash, but still are a

[1]Institute for Mathematics and Computational Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[2]Institute for Mathematics, CNRS UMR 5251, University of Bordeaux, 351 Cours de la Libération, 33405 Talence, France

big problem in practice, in particular if they require the computational grid to be much finer than reasonable. For example it might turn out to be necessary to resolve a vortex with a number of grid cells that depends on how long one intends to run the simulation for. This can occur in simulations of fluid flow phenomena that are smooth (low Mach number/incompressible limit, vortices, ...). They are not well approximated by standard Godunov methods on coarse grids because the effects of numerical diffusion become more and more pronounced as the simulation runs. Grid refinement increases the time scales on which diffusion becomes overwhelming, but – besides slowing down the simulation – does not solve the problem in a fundamental way.

The difficulties of the Godunov method in regions of smooth flow are not surprising, as it introduces shocks everywhere in the computational domain. For example, the failure of the Godunov method to resolve the low Mach number limit on coarse grids has been traced back directly to the acoustic waves introduced at every time step at every cell interface (in [GM04]). For low Mach number flow, modifications of Godunov methods have been suggested which restore the correct behaviour, but they are ad hoc and cannot guarantee adequate treatment of other aspects of smooth multi-dimensional flow (for example vortices). Note at this point that in the quest for a new numerical method we expect it to be able to work well in presence of shocks as well as when they are absent.

The argumentation of [GM04] assumes that the multi-dimensional numerical method is constructed using a one-dimensional Riemann solver. The fluxes through each interface are computed while ignoring the influence of the corners of the cell. Such an approach shall be called directionally-split. Truly multi-dimensional phenomena, where contributions from different direction balance each other, tend to be a challenge for dimensionally split methods. The low Mach number/incompressible limit is one example of such a truly multi-dimensional phenomenon, because the divergence-free condition is trivial in one spatial dimension. One might therefore conjecture that the difficulties do not originate in the Godunov method per se, but in the fact that quasi-one-dimensional solvers are applied in a directionally split way, and thus multi-dimensionality is not taken into account appropriately. One might think that perhaps it would suffice to take into account all the multi-dimensional interactions in multi-dimensional Riemann problems to obtain a Godunov-type method that resolves smooth multi-dimensional problems accurately on coarse grids.

Unfortunately, the situation is not that simple. In [BK22] this line of thought has been developed completely for the linear acoustic equations, which have interesting multi-dimensional features that are not captured by standard dimensionally-split Godunov-type methods on coarse grids (e.g. an analogue of the low Mach number limit, and the involution of vorticity). The complete solution of the four-quadrant Riemann problem for linear acoustics has been obtained and a two-dimensional Godunov method on Cartesian grids constructed in [BK22], whose evolution step was thus exact and fully multi-dimensional. None of the desirable properties (low Mach number compliance, vorticity preservation) was found to hold true for this method. Taking into account all multi-dimensional interactions from multi-dimensional Riemann problems is not enough (and, of course, quite an effort).

Similar observations described in [Roe17] have sparked the development of the Active Flux method by Roe and collaborators ([ER11], based on [vL77]), a numerical method with a continuous reconstruction, whose evolution step thus requires a short-time solution of the IVP for (1) with *continuous* data.

This does not mean that this method is only applicable in the subsonic regime. Con-

tinuous data can self-steepen over the course of a time step, and in such cases, the update needs to be performed carefully, before the discontinuity is projected onto a continuous reconstruction again. In a typical simulation, however, shocks are located on sets of codimension 1 (e.g. in 1-d on countably many points), and thus in this sense almost everywhere in the computational domain the assumption of continuity is correct. Compare this to the approach of a Godunov method solving a rarefaction: instead of maintaining the continuous solution, the rarefaction is cut apart in a staircase shape at every time step, and over the course of the evolution step every jump of the staircase is evolving again into a small rarefaction.

Leaving aside conceptual questions, from a practical point of view, the answers to the following three questions matter most: Is this method stable, possibly even under an explicit time integration? Does the numerical solution obtained with this method converge to the weak solution of (1), i.e. is there an analogue of the Lax-Wendroff-Theorem? Does this method have favorable properties on smooth, multi-dimensional solutions? The answer to all three questions is yes, which makes it an interesting alternative to Godunov-type methods.

The degrees of freedom of the traditional Active Flux [vL77, ER11]) are cell averages and, additionally, point values located at cell interfaces and shared by adjacent cells. This allows for a continuous (but not necessarily differentiable) reconstruction, which passes through the given point values, and whose average agrees with the given one. The evolution of the point value requires the solution of the IVP for (1) over a short time step for continuous initial data. This building block replaces the Riemann solver in Godunov-type methods. The evolution of the average is obtained as usual by integrating the conservation law over the cell and over a time step, and by applying Gauss' theorem. The fluxes at cell interfaces can be immediately evaluated using the available point values. The evolution of the averages therefore is conservative, and an analogue of a Lax-Wendroff theorem for Active Flux has been shown in [Abg20]. Active Flux decouples the problem of average evolution from the computation of interface values, which are declared independent degrees of freedom, hence the name Active Flux, although it would be more accurate to say "Active Point Value at Cell Interface". This allows for more flexibility. The update of point values is not subject to the constraint of conservation, because conservation applies only to averages. Therefore it is possible to even switch variables and update the point values in, say, primitive variables. This has been pointed out in [Abg20]. A more thorough description to the traditional Active Flux method is given in Section 2, as well as in [ER11, ER13, BHKR19, HKS19, Bar21].

The development of a high order method requires chiseling out the distinctive features of the low order method. The aim of the present paper therefore is not only to propose practical high order extensions of Active Flux. These extensions are tied to possible interpretations of Active Flux that allow to place it amidst other families of numerical methods. The three extensions that we propose allow to see Active Flux either as a coupled Finite Volume/Finite Difference method (Section 3), or as an enriched Finite Volume method (Section 4), or as a coupled Finite Element/Finite Difference method (Section 5). Which of the three interpretations, or maybe another, will turn out to be the most fruitful one, only future can tell. Therefore at this stage we content ourselves with developing them to a level that allows to solve the Euler equations, and with stating their respective advantages and disadvantages, whereby we cannot single out any of the three extensions as overperforming the others in all respects.

This paper focuses on the one-dimensional case and defers the case of multiple dimensions to a forthcoming publication in order to keep the size of the paper reasonable. Additionally, in multiple spatial dimensions there are further aspects that need to be taken into account, and the numerical discretization also is different for Cartesian grids and unstructured ones.

Here, the grid is assumed equidistant with cells $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ where $x_{i\pm\frac{1}{2}} = x_i \pm \frac{\Delta x}{2}$, $i \in \mathbb{Z}$. As customary, indices referring to time are denoted as superscripts, i.e. t^n is the n-th time level. C^k denotes the class of k times continuously differentiable functions, and P^k the set of polynomials of degree less or equal to k.

2 Review: Third order Active Flux

In this section, an overview of the existing Active Flux method(s) in one spatial dimension shall be given. At the same time we would like to define Active Flux in the largest possible generality.

Active Flux has two distinctive features. First, the mixed type of degrees of freedom: a cell average, reminiscent of Finite Volume methods, and point values, of which at least some are located at cell boundaries. The numerical solution is thus given by the two sets

$$\{\bar{q}_i^n \in \mathbb{R}^m\}_{i\in I\subset\mathbb{Z}} \qquad\qquad \{q_{i+\frac{1}{2}}^n \in \mathbb{R}^m\}_{i\in I\subset\mathbb{Z}} \tag{3}$$

with the interpretations

$$\bar{q}_i^n \simeq \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathrm{d}x\, q(t^n, x) \qquad\qquad q_{i+\frac{1}{2}}^n \simeq q(t^n, x_{i+\frac{1}{2}}) \tag{4}$$

The same formulae are used to initialize the degrees of freedom $\{\bar{q}_i^0\}_i, \{q_{i+\frac{1}{2}}^0\}_i$ given initial data $q_0 \colon \mathbb{R} \to \mathbb{R}^m$. Traditional degrees of freedom in Active Flux are one cell average per cell, and one (shared) point value located at every cell interface.

Second, the point values are shared between adjacent cells, contrary to the approach of e.g. DG methods. The point values at cell interfaces allow to immediately write down a conservative update of the cell average. Indeed, integrating the conservation law (1) over a computational cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and applying Gauss' law yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(t,x)\,\mathrm{d}x\right) + \frac{f(q(t,x_{i+\frac{1}{2}})) - f(q(t,x_{i-\frac{1}{2}}))}{\Delta x} = 0 \tag{5}$$

The evolution of the point values can happen in many ways, such that for the moment we content ourselves with the following general definition:

Definition 2.1 (Traditional Active Flux)**.** *Inside cell i, consider a reconstruction*

$$q_{recon,i} \in C^0 \cap L^1_{loc} \qquad\qquad q_{recon,i} \colon (\mathbb{R}^m)^3 \times \left[-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right] \to \mathbb{R}^m \tag{6}$$

with the properties

$$q_{recon,i}\left(q_{i-\frac{1}{2}}, \bar{q}_i, q_{i+\frac{1}{2}}, \pm\frac{\Delta x}{2}\right) = q_{i\pm\frac{1}{2}} \tag{7}$$

$$\frac{1}{\Delta x}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \mathrm{d}x\, q_{recon,i}\left(q_{i-\frac{1}{2}}, \bar{q}_i, q_{i+\frac{1}{2}}, x\right) = \bar{q}_i \tag{8}$$

Define the global reconstruction

$$q_{recon} \in C^0 \cap L^1_{loc} \qquad\qquad q_{recon}\colon \mathbb{R} \to \mathbb{R}^m \tag{9}$$

$$q_{recon}(x) := q_{recon,i}\left(q_{i-\frac{1}{2}}, \bar{q}_i, q_{i+\frac{1}{2}}, q_{i,1}, \ldots, q_{i,k}, x - x_i\right) \qquad \text{if } x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \tag{10}$$

The traditional Active Flux method *is the following semi-discretization*

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\bar{q}_i(t) = -\dfrac{f(q_{i+\frac{1}{2}}(t)) - f(q_{i-\frac{1}{2}}(t))}{\Delta x} \\ \\ q_{i+\frac{1}{2}}(t) = \Big(\textit{solution at } x = x_{i+\frac{1}{2}} \textit{ of the IVP (1) with initial data } q_{recon}\Big) + \mathcal{O}(t^3) \end{cases} \tag{11}$$

of (1) *with the interpretations*

$$\bar{q}_i(t) \simeq \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(t,x)\mathrm{d}x \qquad\qquad q_{i+\frac{1}{2}}(t) \simeq q(t, x_{i+\frac{1}{2}}) \tag{12}$$

In the following, we often drop the dependence of $q_{\mathrm{recon},i}$ on its parameters other than x. When the time step is important, we denote by $q^n_{\mathrm{recon},i}$ the reconstruction obtained by using values of time step n.

Usually, $q_{\mathrm{recon},i} \in P^2$ but limiting might require other choices, discussed in Section 2.2. An Active Flux method with this choice of reconstruction is at most third order accurate.

In [Abg20], a closely related definition was given, which, however, does not rely on a reconstruction:

Definition 2.2 (Semidiscrete Active Flux from [Abg20])**.** *The* semidiscrete Active Flux method from [Abg20] *is the following semi-discretization*

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\bar{q}_i(t) = -\dfrac{f(q_{i+\frac{1}{2}}(t)) - f(q_{i-\frac{1}{2}}(t))}{\Delta x} \\ \\ \dfrac{\mathrm{d}}{\mathrm{d}t}q_{i+\frac{1}{2}}(t) = -R\Big(q_{i+\frac{1}{2}-k}(t), \bar{q}_{i-k+1}(t), q_{i+\frac{1}{2}-k+1}(t), \ldots, \bar{q}_{i+m}(t), q_{i+\frac{1}{2}+m}(t)\Big) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i \in \mathbb{Z}, k \geq 0, m \geq 0 \end{cases} \tag{13}$$

of (1) *with the interpretations* (12) *and* R *a consistent approximation of* $\partial_x f(q)$ *at* $x_{i+\frac{1}{2}}$.

Its order of accuracy depends on the approximation order of R (and eventually on the order of accuracy of the time integration).

Later, (Sections 4 and 5) we discuss extensions of Active Flux to higher order that include other types of degrees of freedom. In these cases, the definition is modified accordingly in order to include their respective updates. This is discussed in the corresponding Sections.

2.1 Time integration

Only third-order methods have been suggested so far in the literature, with two different time discretizations for the two methods (11) and (13). In [Abg20], system (13) is integrated with a Runge-Kutta method. The update of the point value is obtained by introducing a point value $q_{i,\text{midpoint}}(t)$ at the center of the cell and expressing the averages as $\bar{q}_i = \frac{1}{6}(q_{i-\frac{1}{2}} + 4q_{i,\text{midpoint}} + q_{i+\frac{1}{2}})$. This allows to choose for R a standard finite difference formula from [Ise82] on a grid of half the spacing:

$$D_3(q_{i-\frac{1}{2}}, q_{i,\text{midpoint}}, q_{i+\frac{1}{2}}, q_{i+1,\text{midpoint}}) := \frac{\frac{1}{6}q_{i-\frac{1}{2}} - q_{i,\text{midpoint}} + \frac{1}{2}q_{i+\frac{1}{2}} + \frac{1}{3}q_{i+1,\text{midpoint}}}{\Delta x/2} \tag{14}$$

$$\simeq \partial_x q + \mathcal{O}(\Delta x^3) \tag{15}$$

For linear advection $\partial_t q + c\partial_x q = 0$, the equation for the point value at cell interface then reads

$$\frac{\mathrm{d}}{\mathrm{d}t} q_{i+\frac{1}{2}}(t) = -c\frac{\frac{1}{6}q_{i-\frac{1}{2}}(t) - q_{i,\text{midpoint}}(t) + \frac{1}{2}q_{i+\frac{1}{2}}(t) + \frac{1}{3}q_{i+1,\text{midpoint}(t)}}{\Delta x/2} \tag{16}$$

To the same order of accuracy, it can be expressed in the form (13) with $k = 1$, $m = 1$ as

$$R(q_{i+\frac{1}{2}-k}t, \bar{q}_{i-k+1}, q_{i+\frac{1}{2}-k+1}(t), \ldots, \bar{q}_{i+m}, q_{i+\frac{1}{2}+m}) = c\frac{\frac{5}{6}q_{i-\frac{1}{2}} - 3\bar{q}_i + \frac{4}{3}q_{i+\frac{1}{2}} + \bar{q}_{i+1} - \frac{1}{6}q_{i+\frac{3}{2}}}{\Delta x} \tag{17}$$

The stability bound for this method is $\lambda_{\max} = \frac{c\Delta t_{\max}}{\Delta x} = 0.77$ (see Section A).

This method has been successfully applied to the Euler equations in [Abg20] by replacing the prefactor c in (16) by the Jacobian $J = f'$ and using two finite-difference formulae (biased in different directions) in order to include upwinding.

The earliest version of Active Flux ([vL77, ER11]) employed a leap-frog-type time integrator for the traditional Active Flux (11). First, approximations

$$q_{i+\frac{1}{2}}^{n+\frac{\ell}{2}} = q_{i+\frac{1}{2}}\left(\ell\frac{\Delta t}{2}\right) + \mathcal{O}(\Delta t^3) \qquad \ell = 1, 2 \tag{18}$$

are computed by solving the IVP (1) as described in Definition 2.1 over half the time step and over the full time step (from the same initial data q_{recon}), and then these values are used to compute fluxes in the update of the average:

$$\bar{q}_i^{n+1} = \bar{q}_i^n - \Delta t \sum_{\ell=0}^{2} \omega_\ell \frac{f(q_{i+\frac{1}{2}}^{n+\frac{\ell}{2}}) - f(q_{i-\frac{1}{2}}^{n+\frac{\ell}{2}})}{\Delta x} \tag{19}$$

Here, $\omega_\ell = \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right)$ are the quadrature weights of Simpson's rule, which is the adequate quadrature for a third order method. This is a leap-frog-type of integration, because the new point values are already used for computing the new value of the average.

The original publication [vL77] considered Active Flux only for linear advection $\partial_t q + c\partial_x q = 0$, in which case one has for positive c

$$q_{i+\frac{1}{2}}^{n+\frac{\ell}{2}} = q_{\mathrm{recon},i}^n \left(x_{i+\frac{1}{2}} - c\frac{\ell}{2}\Delta t \right) \qquad \ell = 1, 2 \tag{20}$$

$$\bar{q}_i^{n+1} = \bar{q}_i^n - \Delta t \sum_{\ell=0}^{2} \omega_\ell c \frac{q_{i+\frac{1}{2}}^{n+\frac{\ell}{2}} - q_{i-\frac{1}{2}}^{n+\frac{\ell}{2}}}{\Delta x} \tag{21}$$

Given the unique parabolic reconstruction $q_{\mathrm{recon},i}^n$ which fulfills

$$q_{\mathrm{recon},i}^n \left(\pm\frac{\Delta x}{2} \right) = q_{i\pm\frac{1}{2}}^n \qquad\qquad \frac{1}{\Delta x} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} q_{\mathrm{recon},i}^n(x)\,\mathrm{d}x = \bar{q}_i^n \tag{22}$$

Equation (20) reads

$$q_{i+\frac{1}{2}}^{n+\frac{\ell}{2}} = q_{i+\frac{1}{2}}^n \left(1 - 2\lambda\ell + \frac{3}{4}\lambda^2\ell^2 \right) + q_{i-\frac{1}{2}}^n \left(-\lambda\ell + \frac{3}{4}\lambda^2\ell^2 \right) - \frac{3}{2}\lambda\ell\bar{q}_i^n(\lambda\ell - 2) \tag{23}$$

$$\bar{q}_i^{n+1} = \bar{q}_i^n(1+2\lambda)(1-\lambda)^2 + \bar{q}_{i-1}^n\lambda^2(3-2\lambda) - q_{i-\frac{3}{2}}^n(1-\lambda)\lambda^2 + q_{i-\frac{1}{2}}^n\lambda(1-\lambda) - q_{i+\frac{1}{2}}^n(1-\lambda)^2\lambda \tag{24}$$

with $\Delta t = \lambda\Delta x/c$.

This method is stable up to $\lambda = 1$. It has been extended to nonlinear scalar conservation laws and to hyperbolic systems of conservation laws by means of an estimate of the characteristic directions in [Bar21]. The main difficulty is to obtain a sufficiently accurate approximate solution of the IVP. In multiple spatial dimensions, the concept of characteristics (along which some quantity is constant) is generally to be replaced by that of characteristic cones, which do not allow for simple transformations to characteristic variables. For linear acoustics, an exact evolution operator has been obtained in [ER13, BK22]. It has been then applied to Active Flux on Cartesian two-dimensional grids in [BHKR19], and structure preservation properties of the resulting method have been demonstrated: The Active Flux method has been found to be vorticity preserving / low Mach compliant. Despite promising efforts towards approximate evolution operators for the multi-dimensional Euler equations in e.g. [Fan17], further research is required for hyperbolic systems in multiple spatial dimensions that is beyond the scope of the present paper.

2.2 Limiting

Limiting is necessary for Active Flux because it is a high order method, and because it is linear when applied to linear problems. The usual philosophy of limiting in the context of finite volume methods is to modify the reconstruction such that its value at the cell interface satisfies a maximum principle. Examples of such approaches are slope limiters, or the approach in [ZS10] where the reconstruction polynomial is scaled around its average. For Active Flux, it is not possible to combine such ways of limiting with continuity of the reconstruction because the values at cell interfaces are prescribed. Limiting strategies for Active Flux that give up continuity can be found in e.g. [HKS19, CHK21].

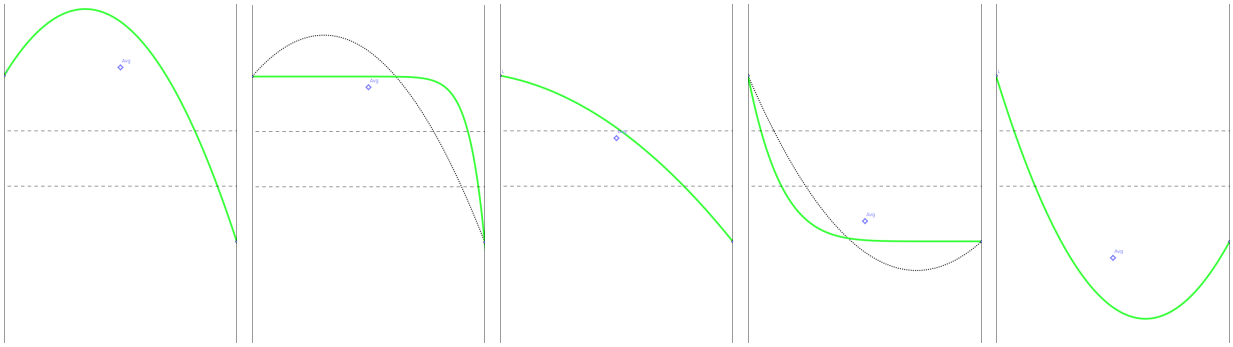

Figure 1: Different cases of limiting. The two point values remain the same, the diamond in the cell center indicates the value of the cell average. The green solid line shows the choice of reconstruction, the dotted line shows the parabolic reconstruction whenever it is not the one chosen. The dashed horizontal lines indicate the region of the cell average inside which the parabolic reconstruction is monotone. *From left to right*: The cell average is larger than the two point values, there exists is no continuous monotone reconstruction, and the parabola is accepted. The cell average is between the two point values, but the parabola is not monotone, we choose one of the power laws. The average is inside the region given by the dashed lines, we choose the (monotone) parabola. The average is below the region, we choose the other power law. The average is below both point values, we choose the parabola again.

Here, we prefer to maintain continuity. The traditional Active Flux method uses a reconstruction to evolve the point values, and it is a natural approach to limiting to replace an oscillative reconstruction by a monotone, or less oscillative function. Such limiting strategies based on piecewise defined functions (e.g. a constant and a parabola) have been proposed in [RLM15, BB20]. A simple limiting strategy has been introduced in [Bar21] where the parabolic reconstruction is replaced by a power law. This shall be taken as inspiration here.

Recall the following result from [Bar21]:

Theorem 2.1. *Consider monotone data*

$$q_{i-\frac{1}{2}} \leq \bar{q}_i \leq q_{i+\frac{1}{2}} \quad \text{or} \quad q_{i-\frac{1}{2}} \geq \bar{q}_i \geq q_{i+\frac{1}{2}} \tag{25}$$

Their unique parabolic interpolant is monotone, iff

$$r := \frac{q_{i+\frac{1}{2}} - \bar{q}_i}{\bar{q}_i - q_{i-\frac{1}{2}}} \in \left[\frac{1}{2}, 2\right] \tag{26}$$

Thus, parabolic interpolants can violate the maximum principle if the average $\bar{q}_i$ is close

to one of the point values $q_{i\pm\frac{1}{2}}$. The power law reconstructions, proposed in [Bar21]

$$q^n_{\text{recon},i,\text{power-law},1}(x) = q^n_{i-\frac{1}{2}} + (q^n_{i+\frac{1}{2}} - q^n_{i-\frac{1}{2}}) \left(\frac{x + \frac{\Delta x}{2}}{\Delta x} \right)^{\frac{q^n_{i+\frac{1}{2}} - \bar{q}^n_i}{\bar{q}^n_i - q^n_{i-\frac{1}{2}}}} \tag{27}$$

$$q^n_{\text{recon},i,\text{power-law},2}(x) = q^n_{i+\frac{1}{2}} - (q^n_{i+\frac{1}{2}} - q^n_{i-\frac{1}{2}}) \left(\frac{\frac{\Delta x}{2} - x}{\Delta x} \right)^{\frac{\bar{q}^n_i - q^n_{i-\frac{1}{2}}}{q^n_{i+\frac{1}{2}} - \bar{q}^n_i}} \tag{28}$$

$$x \in \left[-\frac{\Delta x}{2}, \frac{\Delta x}{2} \right]$$

are interpolating the same data while being monotone whenever the data are, and are thus a good replacement for the parabolic reconstruction whenever the latter fails to be monotone. Observe that the expression appearing in (26) is just the power in (27). Thus, when

$$r = \frac{q_{i+\frac{1}{2}} - \bar{q}_i}{\bar{q}_i - q_{i-\frac{1}{2}}} = 2 \qquad \Rightarrow \qquad \bar{q}_i - q_{i-\frac{1}{2}} = \frac{q_{i+\frac{1}{2}} - q_{i-\frac{1}{2}}}{3} \tag{29}$$

the power law (27) reduces to the parabolic interpolant. The same is true for (28) for

$$r = \frac{q_{i+\frac{1}{2}} - \bar{q}_i}{\bar{q}_i - q_{i-\frac{1}{2}}} = \frac{1}{2} \qquad \Rightarrow \qquad \bar{q}_i - q_{i+\frac{1}{2}} = \frac{q_{i-\frac{1}{2}} - q_{i+\frac{1}{2}}}{3} \tag{30}$$

If $r = \frac{1}{2}$, one can switch from the parabola to the power-law (28), and at $r = 2$ to the the power law (27):

Theorem 2.2. *Consider monotone data fulfilling* (25)*. The reconstruction*

$$q_{recon,i}(x) = \begin{cases} q^n_{recon,i,power\text{-}law,1}(x) & |\bar{q}_i - q_{i-\frac{1}{2}}| < \frac{|q_{i+\frac{1}{2}} - q_{i-\frac{1}{2}}|}{3} \\ q^n_{recon,i,power\text{-}law,2}(x) & |\bar{q}_i - q_{i+\frac{1}{2}}| < \frac{|q_{i+\frac{1}{2}} - q_{i-\frac{1}{2}}|}{3} \\ parabolic & else \end{cases} \qquad x \in \left[-\frac{\Delta x}{2}, \frac{\Delta x}{2} \right]$$

is continuous as a function of $\bar{q}_i$*.*

For data that are not monotone (i.e. violating (25)), a parabolic reconstruction is used. See Figure 1 for an overview.

A rational interpolation (suggested in [HKS19]) is similar in spirit to the power law reconstruction, but its coefficients cannot be computed analytically because of the condition on its average, which requires integration of a rational function.

Note that in all suggestions of limiting available in the literature, only the update of the point values is limited. This provides satisfactory results, but does not exclude the appearance of oscillations, as the finite volume step

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{q}_i(t) = -\frac{f(q_{i+\frac{1}{2}}(t)) - f(q_{i-\frac{1}{2}}(t))}{\Delta x} \tag{31}$$

remains unlimited. A limiting of the finite volume step is subject of future work.

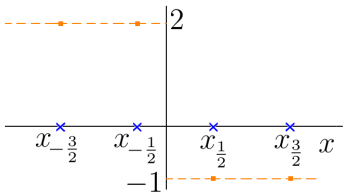

Figure 2: Setup for the illustration of problems if transonic upwinding is not performed carefully.

2.3 Transonic upwinding

In [HKS19, Bar21], the following problematic situation has been observed with early versions of Active Flux. Consider initial data

$$q_0(x) = \begin{cases} 2 & x \le 0 \\ -1 & x > 0 \end{cases} \tag{32}$$

for Burgers' equation, such that the discrete data initially are (see also Figure 2)

$$\ldots = q_{-\frac{3}{2}} = q_{-\frac{1}{2}} = 2 \qquad q_{\frac{1}{2}} = q_{\frac{3}{2}} = \ldots = -1 \tag{33}$$

$$\ldots = \bar{q}_{-2} = \bar{q}_{-1} = 2 \qquad \bar{q}_1 = \bar{q}_2 = \ldots = -1 \tag{34}$$

$$q_{\text{recon},-1} \equiv 2 \qquad q_{\text{recon},1} \equiv -1 \tag{35}$$

The value of $\bar{q}_0$ is irrelevant for the discussion.

The exact solution of (32) would be a shock moving at speed $\frac{1}{2}$. In the update of $q_{-\frac{1}{2}}$, if its value 2 is used to estimate the upwind direction, information would be taken from the left, where the reconstruction is identically equal to 2. $q_{-\frac{1}{2}}$ would thus remain stationary. The same happens to $q_{\frac{1}{2}}$, whose initial value is negative, and thus information is taken from the right. In fact, the reconstruction $q_{\text{recon},0}$ thus never contributes. The numerical method keeps all point values stationary, while the average in cell 0 will grow, because the flux difference is $(q_{\frac{1}{2}})^2/2 - (q_{-\frac{1}{2}})^2/2 \neq 0$.

The error is the wrong estimate of the upwind direction at $x_{\frac{1}{2}}$. An exact time evolution of the continuous reconstruction q_{recon} would see it form a shock, and this shock would first move left, and then right, reaching $x_{\frac{1}{2}}$ before the end of the time step. At this moment, the direction of information flow at $x_{\frac{1}{2}}$ would be reversed, and information would no longer be taken from the right (see Figure 3). Thus, the deficiency of the early versions of Active Flux is due to an erroneous estimate of the upwind direction in the point value update. This estimate needs to take into account the fact that a continuous reconstruction might self-steepen before the end of the time step.

This error is most visible in the transonic case, because the shock has time to form during the time step. Depending on the CFL condition, setups close to transonic might be affected as well. In the worst case, the consequence is a shock that is not moving (and thus violating the Rankine-Hugoniot conditions), with the cell average in its vicinity growing without bound. In setups that are not strictly transonic, an artificially large average at the location of the shock is visible in the form of a spike (see examples in [Bar21]).

In [Bar21] it has been suggested to estimate the upwind direction more carefully. In particular, when solving for the foot point of the characteristic (which, generally, requires

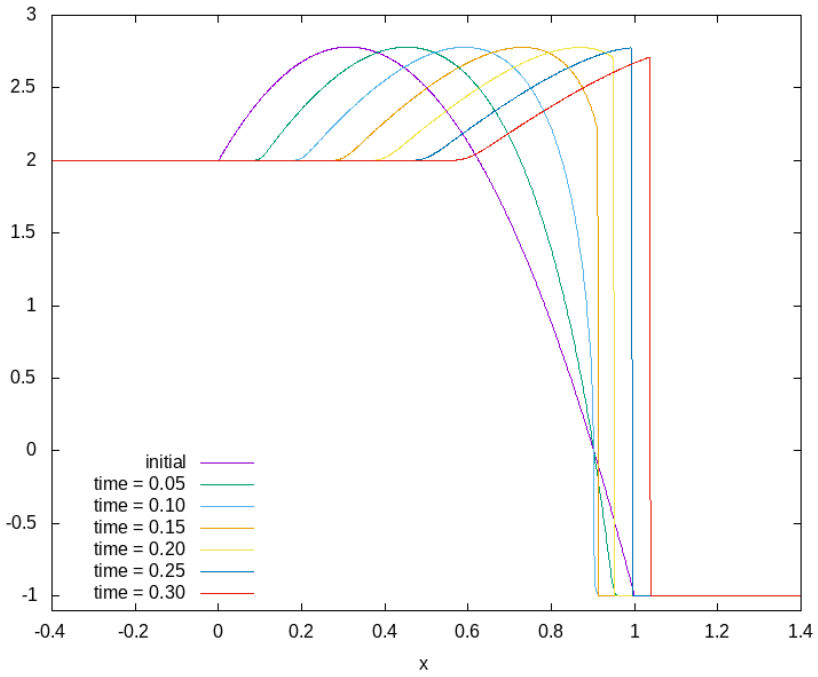

Figure 3: The evolution of continuous initial data corresponding to a parabolic reconstruction of (32) with $\Delta x = 1$ for Burgers' equation. The time step based on the maximum of the reconstruction is $\Delta t \simeq 0.357$, and even larger if based on the point values and averages alone. At $x = 1$ one observes initially an information flow from the right, but as the shock forms, it stalls and starts moving to the right. After $t \simeq 0.25$, the information is flowing from the left at $x = 1$.

to solve a nonlinear algebraic equation), it has been suggested to initialize the iteration with not only the value at the respective cell interface, but also with values from its neighbours. If a shock forms over the time step, one would thus obtain several candidate characteristics, while in a smooth setup the iteration would converge to the unique characteristic. A simple strategy that works well in practice is then to select from the candidate characteristics the one whose speed is largest in modulus. In [Bar21, BB20] it has been demonstrated that this strategy works well even for systems. Some authors (e.g. [CHK21]) suggest to average to values obtained from the candidate characteristics.

3 High order via larger stencils

3.1 Discretization in space

The first strategy for high order accuracy that we propose involves the traditional degrees of freedom (cell averages and shared point values at cell interfaces), and uses larger stencils. The starting point therefore is the semidiscrete Active Flux of Definition 2.2, with a Runge-Kutta time integration. Observe again that the update of the averages

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{q}_i(t) = -\frac{f(q_{i+\frac{1}{2}}(t)) - f(q_{i-\frac{1}{2}}(t))}{\Delta x} \tag{36}$$

is exact, and that the order of accuracy of the method is given by the order of accuracy of the right-hand side of the point value update in (13) and of the order of accuracy of the time integration.

Consider linear finite difference p-th order approximations to the derivative of the form

$$\frac{1}{\Delta x}\sum_{j=-\ell}^{\ell}(b_j^{(0)}\bar{q}_{i+j} + b_{j+\frac{1}{2}}q_{i+j+\frac{1}{2}}) = q'(x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^p) \tag{37}$$

We shall adopt the following tableau notation for such a finite difference formula that involves point values at cell interfaces and cell averages

Notation 3.1.

$$\frac{1}{\Delta x}\sum_{j=-\ell}^{\ell}(b_j^{(0)}\bar{q}_{i+j} + b_{j+\frac{1}{2}}q_{i+j+\frac{1}{2}}) = \tag{38}$$

$b_{-\ell}^{(0)}$		$\cdots$		$b_{-1}^{(0)}$		$b_0^{(0)}$	$\|\|$	$b_1^{(0)}$		$\cdots$		$b_\ell^{(0)}$	
	$b_{-\ell+\frac{1}{2}}$		$b_{-\frac{3}{2}}$		$b_{-\frac{1}{2}}$		$b_{\frac{1}{2}}$		$b_{\frac{3}{2}}$		$b_{\ell-\frac{1}{2}}$		$b_{\ell+\frac{1}{2}}$

The vertical lines indicate cell interfaces (such that coefficients of cell averages are written "inside" the cell and coefficients of point values are associated to cell interfaces). The double vertical line indicates the cell interface at which the finite difference is supposed to provide an approximation to the derivative (of the corresponding order of accuracy). Coefficients not marked explicitly in the tableau are assumed zero. The notation will be slightly expanded in Section 5, where the superscript $^{(0)}$ *on the coefficients of the cell averages will become clear.*

Here are some examples ($a_2, \ldots, a_7$ are free parameters):

$$FD1 = \left[\begin{array}{c|c||c} & 2-2a_2 & \\ -2+a_2 & & a_2 \end{array}\right] \tag{39}$$

$$FD3a = \left[\begin{array}{c|c||c|c} & -2-\frac{3a_3}{4} & & 2-\frac{3a_3}{4} & \\ \frac{2+a_3}{4} & & a_3 & & \frac{-2+a_3}{4} \end{array}\right] \tag{40}$$

$$FD5b = \left[\begin{array}{c|c|c||c|c} & \frac{19-22a_5}{54} & & -\frac{89+76a_5}{54} & & \frac{50-11a_5}{27} & \\ \frac{a_5-1}{9} & & a_5 & & a_5 & & \frac{a_5-4}{9} \end{array}\right] \tag{41}$$

$$FD7a = \left[\begin{array}{c|c|c|c|c||c|c} & \frac{49}{72}-\frac{25a_7}{96} & & \frac{293}{72}-\frac{185a_7}{96} & & -\frac{31}{72}-\frac{185a_7}{96} & & \frac{436-75a_7}{288} & \\ \frac{-8+3a_7}{48} & & -\frac{7}{3}+a_7 & & -3+\frac{9a_7}{4} & & a_7 & & -\frac{1}{3}+\frac{a_7}{16} \end{array}\right] \tag{42}$$

Note once more that the double vertical line indicates the cell interface at which the finite difference is providing the desired approximation of the derivative. E.g. for FD1,

$$q'(x_{i+\frac{1}{2}}) \simeq (-2+a_2)q_{i-\frac{1}{2}} + (2-2a_2)\bar{q}_i + a_2 q_{i+\frac{1}{2}} \tag{43}$$

while for FD3a,

$$q'(x_{i+\frac{1}{2}}) \simeq \frac{2+a_3}{4}q_{i-\frac{1}{2}} + \left(-2-\frac{3a_3}{4}\right)\bar{q}_i + a_3 q_{i+\frac{1}{2}} + \left(2-\frac{3a_3}{4}\right)\bar{q}_{i+1} + \frac{-2+a_3}{4}q_{i+\frac{3}{2}} \tag{44}$$

The finite difference formulas are involve one degree of freedome more than it would be necessary to obtain the desired order of accuracy; the remaining free parameter is used later to optimize stability. These finite difference formulas are different from standard finite differences because they involve point values and averages. Table 1 summarizes some of the possible choices. Given a finite difference formula

$$D = \left[\begin{array}{c|c|c|c||c|c|c|c} b^{(0)}_{-\ell} & \cdots & b^{(0)}_{-1} & b^{(0)}_{0} & b^{(0)}_{1} & \cdots & b^{(0)}_{\ell} & \\ & b_{-\ell+\frac{1}{2}} & b_{-\frac{3}{2}} & b_{-\frac{1}{2}} & b_{\frac{1}{2}} & b_{\frac{3}{2}} & b_{\ell-\frac{1}{2}} & b_{\ell+\frac{1}{2}} \end{array}\right] \tag{45}$$

define the flipped formula

$$D^* = \left[\begin{array}{c|c|c|c||c|c|c|c} -b^{(0)}_{\ell} & \cdots & -b^{(0)}_{1} & -b^{(0)}_{0} & \cdots & -b^{(0)}_{-\ell} & \\ -b_{\ell+\frac{1}{2}} & -b_{\ell-\frac{1}{2}} & -b_{\frac{3}{2}} & -b_{\frac{1}{2}} & -b_{-\frac{1}{2}} & -b_{-\ell+\frac{3}{2}} & -b_{-\ell+\frac{1}{2}} \end{array}\right] \tag{46}$$

It is an approximation of the derivative at the same location and of the same order of accuracy as D, and arises because the spatial derivative changes sign upon reflection $x \mapsto -x$.

We propose to use these finite differences in the update of the point values by writing

$$\frac{\mathrm{d}}{\mathrm{d}t}q_{i+\frac{1}{2}}(t) = -\Big(f'(\tilde{q}_{i+\frac{1}{2}})^+ D + f'(\tilde{q}_{i+\frac{1}{2}})^- D^*\Big) \tag{47}$$

where D is any of the finite difference formulae (39)–(42). In the scalar case the positive and negative parts are given by

$$f'(\tilde{q}_{i+\frac{1}{2}})^+ = \max(0, f'(\tilde{q}_{i+\frac{1}{2}})) \qquad f'(\tilde{q}_{i+\frac{1}{2}})^- = \min(0, f'(\tilde{q}_{i+\frac{1}{2}})) \tag{48}$$

In the case of systems, f' is the Jacobian and they are defined via its eigenvalues:

$$f'(\tilde{q}_{i+\frac{1}{2}}) = R\mathrm{diag}\,(\lambda_1, \ldots, \lambda_m)R^{-1} \qquad f'(\tilde{q}_{i+\frac{1}{2}})^\pm := R\mathrm{diag}\,(\lambda_1^\pm, \ldots, \lambda_m^\pm)R^{-1} \tag{49}$$

For an Active Flux method of order N, it is sufficient to use a derivative approximation of order $N-1$. This offset is due to the offset in the definition of "order". For example, a 3rd order Active Flux method is exact on parabolic data, while already a 2nd order derivative approximation is exact for parabolae.

It remains to define $\tilde{q}_{i+\frac{1}{2}}$. For scalar conservation laws ($m = 1$), if f' does not switch sign in the computational domain, then the natural choice $\tilde{q}_{i+\frac{1}{2}} = q^n_{i+\frac{1}{2}}$ works. Further details are given in the next section.

The use of finite differences in the point value update is inspired by the approach in [Abg20], but different in the following crucial aspect. In [Abg20], a point value $q_{i,\mathrm{midpoint}}$ at the cell center is first estimated through Simpson's rule:

$$\bar{q}_i =: \frac{q_{i-\frac{1}{2}} + 4q_{i,\mathrm{midpoint}} + q_{i+\frac{1}{2}}}{6} \tag{50}$$

One thus obtains a grid of half spacing, on which standard finite differences can be used. (This is described in Section 2.1.) While for third order, Simpson's rule is accurate enough, for higher orders the natural expression of the cell average in terms of point values would require solving a linear system, i.e. it would become nonlocal. Indeed, a formula such as

$$\bar{q}_i =: \frac{-4q_{i-1,\mathrm{midpoint}} + 34q_{i-\frac{1}{2}} + 144q_{i,\mathrm{midpoint}} + 34q_{i+\frac{1}{2}} - q_{i+1,\mathrm{midpoint}}}{180} = \frac{1}{\Delta x}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \mathrm{d}xq + \mathcal{O}(\Delta x^6) \tag{51}$$

cannot be as easily solved for the midpoint values as (50). One could use a quadrature for the average that involves precisely one midpoint value and correspondingly more of the point values at cell interfaces:

$$\bar{q}_i =: \frac{-4q_{i-\frac{3}{2}} + 189q_{i-\frac{1}{2}} + 704q_{i,\mathrm{midpoint}} + 189q_{i+\frac{1}{2}} - q_{i+\frac{3}{2}}}{1080} = \frac{1}{\Delta x}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \mathrm{d}xq + \mathcal{O}(\Delta x^6) \tag{52}$$

It is clear, however, that the form (37) includes this case, and is more general.

3.2 Time integration and stability conditions

The algorithm consists of the ODE system (36) and (47). We propose to solve it with a Runge-Kutta time integrator, e.g. a strong stability preserving method of order 3 (SSP-RK3). For linear problems, of course, higher order Runge-Kutta methods are easily available. (Compare Figures 11 and 12, where stability of FD5b is shown for RK3 and RK5, respectively, with the maximum CFL number being only slightly higher for RK5.). For nonlinear problems, often a lower order time integrator is used with a time step small

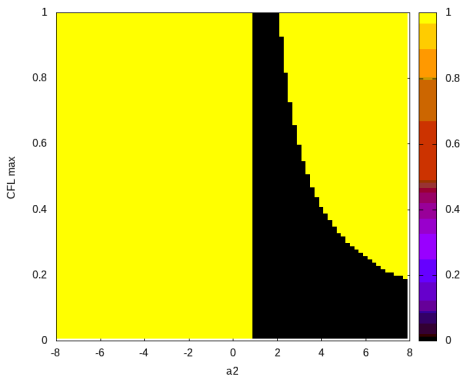

Figure 4: Stability region (RK3) of (39) (FD1) (black = stable).

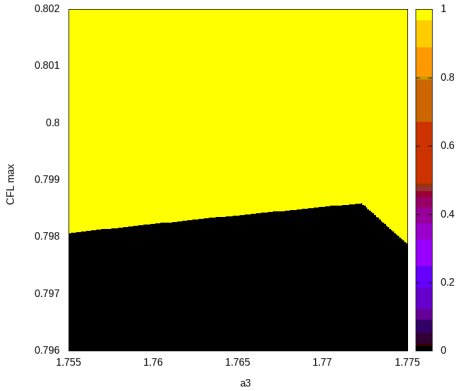

Figure 5: Stability region (RK3) of (40) (FD3a) (black = stable).

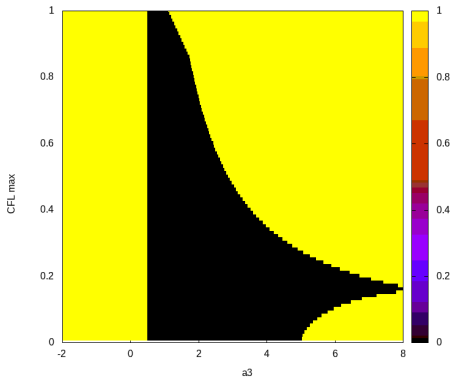

Figure 6: Stability region (RK3) of finite difference formula FD3b (black = stable).

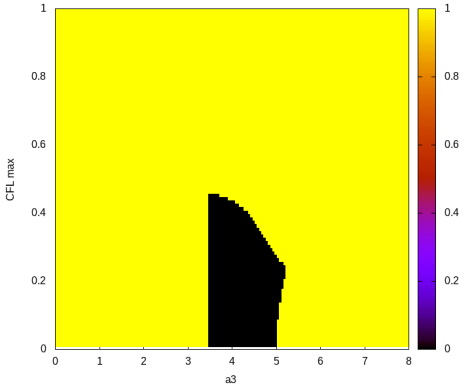

Figure 7: Stability regions (RK3) of finite difference formula FD3c (black = stable).

enough, such that the global error is dominated by the spatial accuracy. We thus content ourselves with results on RK3. The finite difference formulas considered above are typically one cell larger than it would be necessary to obtain the desired order of accuracy, such that the free parameter can be used to optimize stability. The results of the linear stability analysis (see Section A) can thus be depicted in a diagram whose axes show the CFL number and the value of the free parameter. Some of these diagrams are shown in Figures 6–15. Table 1, shows the maximum CFL numbers for parameter values optimizing stability.

3.3 Transonic upwinding

By analogy with the way how transonic upwinding has been proposed to be handled in [Bar21] (see also Section 2.3), we propose here for the case of a scalar conservation law ($m = 1$) the following choice:

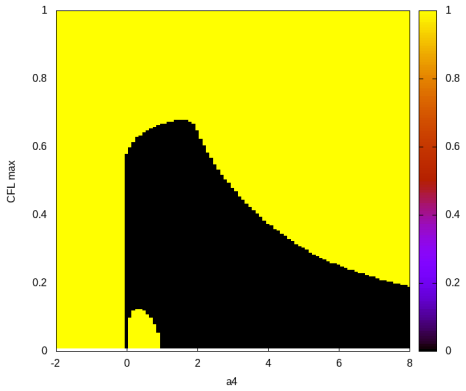

Figure 8: Stability region (RK3) of finite difference formula FD4a (black = stable).

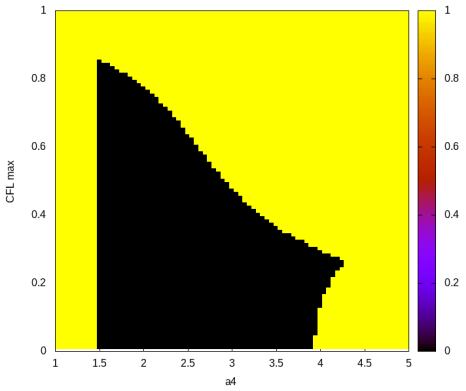

Figure 9: Stability regions (RK3) of finite difference formula FD4b (black = stable).

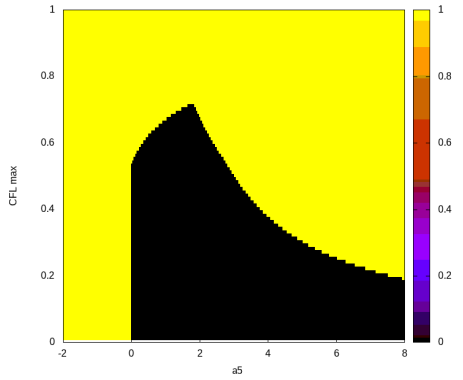

Figure 10: Stability region (RK3) of finite difference formula FD5a (black = stable).

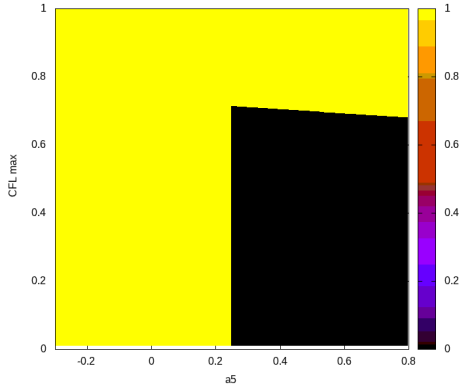

Figure 11: Stability region (RK3) of (41) (FD5b) (black = stable).

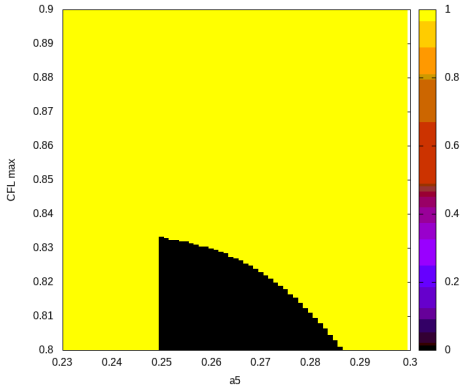

Figure 12: Stability region of (41) (FD5b) for RK5 (black = stable).

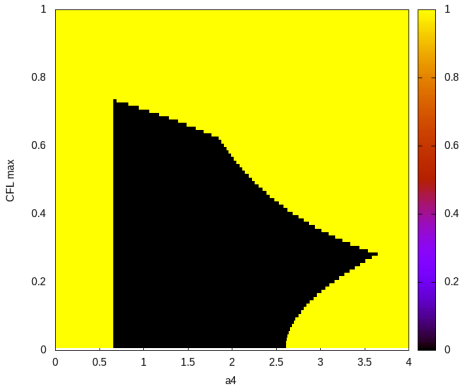

Figure 13: Stability region (RK3) of finite difference formula FD6 (black = stable).

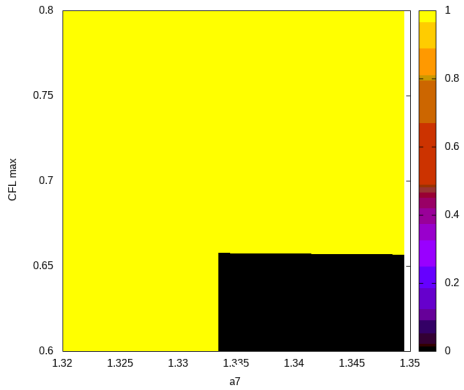

Figure 14: Stability region (RK3) of (42)(FD7a) (black = stable).

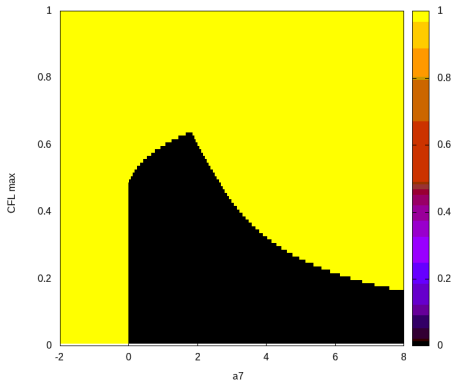

Figure 15: Stability region (RK3) of finite difference formula FD7c (black = stable).

name	FD1	FD3a	FD3b	FD3c	FD4a	FD4b
p	1 (2)	3	3	3	4	4
$b_{-3}^{(0)}$						
$b_{-\frac{5}{2}}$						
$b_{-2}^{(0)}$						
$b_{-\frac{3}{2}}$				$a_3 - 5$		$\frac{a_4-3}{3}$
$b_{-1}^{(0)}$			$\frac{1}{6}(2-a_3)$	$\frac{29}{2} - 3a_3$	$-\frac{a_4}{18}$	$\frac{19}{6} - \frac{10a_4}{9}$
$b_{-\frac{1}{2}}$	$-2 + a_2$	$\frac{2+a_3}{4}$	$a_3 - 1$	$4(a_3 - 4)$	$\frac{1+a_4}{2}$	$2(a_4 - 2)$
$b_0^{(0)}$	$2 - 2a_2$	$-2 - \frac{3a_3}{4}$	$-\frac{1}{6} - \frac{5a_3}{3}$	$\frac{13}{2} - 3a_3$	$-2 - \frac{19a_4}{18}$	$\frac{7}{6} - \frac{19a_4}{9}$
$b_{\frac{1}{2}}$	a_2	a_3	a_3	a_3	a_4	a_4
$b_1^{(0)}$		$2 - \frac{3a_3}{4}$	$\frac{1}{6}(5-a_3)$		$2 - \frac{5a_4}{9}$	$\frac{6-a_4}{9}$
$b_{\frac{3}{2}}$		$\frac{-2+a_3}{4}$			$\frac{a_4-3}{6}$	
$\mathrm{CFL}_{\max}$ (RK3)	1	0.7985	1	0.45	0.675	0.855
parameter region	$a_2 \in [1, 2]$ Fig. 4	$a_3 \simeq 1.7723$ Fig. 5	$a_3 \in [\frac{1}{2}, 1.1]$ Fig. 6	$a_3 \simeq 3.5$ Fig. 7	$a_4 \simeq 1.6$ Fig. 8	$a_4 \simeq 1.5$ Fig. 9

Table 1: Overview of finite difference formulae. The lower part shows the maximum attainable CFL number together with the corresponding value of the parameter. The formula marked as first-order accurate becomes a second-order accurate approximation upon the choice $a_2 = 4$.

name	FD5a	FD5b	FD5c	FD6	FD7a	FD7c
p	5	5	5	6	7	7
$b_{-3}^{(0)}$						
$b_{-\frac{5}{2}}$					$\frac{-8+3a_7}{48}$	
$b_{-2}^{(0)}$			$\frac{4-a_5}{12}$	$\frac{2-a_6}{48}$	$\frac{49}{72} - \frac{25a_7}{96}$	
$b_{-\frac{3}{2}}$		$\frac{a_5-1}{9}$	$a_5 - \frac{11}{3}$	$\frac{3a_6-5}{9}$	$-\frac{7}{3} + a_7$	$\frac{1+a_7}{36}$
$b_{-1}^{(0)}$	$-\frac{1+a_5}{36}$	$\frac{19-22a_5}{54}$	$\frac{302-87a_5}{36}$	$\frac{586-393a_6}{432}$	$\frac{293}{72} - \frac{185a_7}{96}$	$-\frac{28+25a_7}{216}$
$b_{-\frac{1}{2}}$	$\frac{2+a_5}{3}$	a_5	$3a_5 - 8$	$\frac{3a_6}{2} - 1$	$-3 + \frac{9a_7}{4}$	$\frac{4}{9}(2 + a_7)$
$b_0^{(0)}$	$-\frac{9}{4} - \frac{29a_5}{36}$	$-\frac{89+76a_5}{54}$	$\frac{86-87a_5}{36}$	$\frac{-494-717a_6}{432}$	$-\frac{31}{72} - \frac{185a_7}{96}$	$-\frac{5}{216}(108 + 37a_7)$
$b_{\frac{1}{2}}$	a_5	a_5	a_5	a_6	a_7	a_7
$b_1^{(0)}$	$\frac{9}{4} - \frac{29a_5}{36}$	$\frac{50-11a_5}{27}$	$\frac{5}{9} - \frac{a_5}{12}$	$\frac{730-141a_6}{432}$	$\frac{436-75a_7}{288}$	$\frac{5}{2} - \frac{185a_7}{216}$
$b_{\frac{3}{2}}$	$\frac{a_5-2}{3}$	$\frac{a_5-4}{9}$		$\frac{3a_6-14}{36}$	$-\frac{1}{3} + \frac{a_7}{16}$	$\frac{4}{9}(a_7 - 2)$
$b_2^{(0)}$	$\frac{1-a_5}{36}$					$\frac{28-25a_7}{216}$
$b_{\frac{5}{2}}$						$\frac{a_7-1}{36}$
$\mathrm{CFL}_{\max}$ (RK3)	0.7	0.713	0.56	0.73	0.657	0.62
parameter	$a_5 \simeq 1.88$	$a_5 \simeq \frac{1}{4}$	$a_5 \simeq 2.3$	$a_6 \simeq 0.68$	$a_7 \simeq \frac{4}{3}$	$a_7 \simeq 1.9$
region	Fig. 10	Fig. 11		Fig. 13	Fig. 14	Fig. 15

Table 2: Table 1 continued.

$$\tilde{q}_{i+\frac{1}{2}} = \begin{cases} q^n_{i+\frac{1}{2}} & \text{if sign } f'(q^n_{i-\frac{1}{2}}) = \text{sign } f'(q^n_{i+\frac{1}{2}}) = \text{sign } f'(q^n_{i+\frac{3}{2}}) \\ q^n_{i+\frac{1}{2}} & \text{else and if } q^n_{i-\frac{1}{2}} < q^n_{i+\frac{3}{2}} \text{transonic rarefaction} \\ q^n_{i-\frac{1}{2}} & \text{else and if } |f'(q^n_{i-\frac{1}{2}})| \geq \max(|f'(q^n_{i+\frac{1}{2}})|, |f'(q^n_{i+\frac{3}{2}})|) \\ q^n_{i+\frac{1}{2}} & \text{else and if } |f'(q^n_{i+\frac{1}{2}})| \geq \max(|f'(q^n_{i-\frac{1}{2}})|, |f'(q^n_{i+\frac{3}{2}})|) \\ q^n_{i+\frac{3}{2}} & \text{else and if } |f'(q^n_{i+\frac{3}{2}})| \geq \max(|f'(q^n_{i-\frac{1}{2}})|, |f'(q^n_{i+\frac{1}{2}})|) \end{cases} \tag{53}$$

While the original modification from [Bar21] leads to a third-order accurate method, the above suggestion degrades the accuracy at sonic points. We do not find averages such as $\tilde{q}_{i+\frac{1}{2}} = \frac{q^n_{i-\frac{1}{2}} + 2q^n_{i+\frac{1}{2}} + q^n_{i+\frac{3}{2}}}{4}$ to work satisfactorily for our high order versions.

For systems, the presence of different waves and thus inherent information transport from different parts of the initial data seems to make transonic upwinding less important. We experimentally find that the high-order version of Active Flux via finite differences in many cases, such as in the numerical examples below, works well with the simple choice $\tilde{q}_{i+\frac{1}{2}} = q^n_{i+\frac{1}{2}}$.

3.4 Limiting

We propose to gradually reduce the order of the finite difference approximation employed if a violation of monotonicity is detected. If the values

$$\bar{q}_{i-\ell}, q_{i+\frac{1}{2}-\ell}, \ldots, \bar{q}_{i+\ell}, q_{i+\frac{1}{2}+\ell} \tag{54}$$

involved in the finite difference formula are monotone, then one would expect the finite difference approximation to the derivative to have the same sign as, say, $q_{i+\frac{1}{2}} - \bar{q}_i$. If this is not the case we propose to reduce the order (e.g. as FD7a–FD6–FD5b–FD4b–FD3b–FD1 with $a_2 = 4$, which is second order), until either monotonicity is obtained, or until even the finite difference of lowest order is found to be non-monotone. In this case, a fall-back finite difference is used. We propose to take inspiration from the power-law limiting suggested in [Bar21]. Its original version amounts to a modification of the reconstruction, but a finite difference can be obtained by differentiation at the cell interfaces $x = \pm\frac{\Delta x}{2}$. The nonvanishing derivatives of the two power laws (27)–(28) yield two nonlinear finite difference approximations:

$$\frac{\mathrm{d}}{\mathrm{d}x} q^n_{\text{recon},i,\text{power-law},1}\left(-\frac{\Delta x}{2}\right) = \frac{1}{\Delta x}(q^n_{i+\frac{1}{2}} - q^n_{i-\frac{1}{2}})\frac{\bar{q}^n_i - q^n_{i-\frac{1}{2}}}{q^n_{i+\frac{1}{2}} - \bar{q}^n_i} \tag{55}$$

$$\frac{\mathrm{d}}{\mathrm{d}x} q^n_{\text{recon},i,\text{power-law},2}\left(\frac{\Delta x}{2}\right) = \frac{1}{\Delta x}(q^n_{i+\frac{1}{2}} - q^n_{i-\frac{1}{2}})\frac{q^n_{i+\frac{1}{2}} - \bar{q}^n_i}{\bar{q}^n_i - q^n_{i-\frac{1}{2}}} \tag{56}$$

These finite differences are first-order accurate, as can be checked explicitly. Up to a shift by one cell, the one is a finite difference at $x_{i+\frac{1}{2}}$ for positive speed, and the other for negative, i.e. we define

$$D_1(q_{i-\frac{1}{2}}, \bar{q}_i, q_{i+\frac{1}{2}}) := \frac{1}{\Delta x}(q^n_{i+\frac{1}{2}} - q^n_{i-\frac{1}{2}})\frac{q^n_{i+\frac{1}{2}} - \bar{q}^n_i}{\bar{q}^n_i - q^n_{i-\frac{1}{2}}} \tag{57}$$

$$D^*_1(q_{i+\frac{1}{2}}, \bar{q}_{i+1}, q_{i+\frac{3}{2}}) := \frac{1}{\Delta x}(q^n_{i+\frac{3}{2}} - q^n_{i+\frac{1}{2}})\frac{\bar{q}^n_{i+1} - q^n_{i+\frac{1}{2}}}{q^n_{i+\frac{3}{2}} - \bar{q}^n_{i+1}} \tag{58}$$

By analogy with the original version from [Bar21], this finite difference is used only if the values (54) are monotone, all higher order finite differences were found to violate monotonicity, and if condition (26) is violated. For practical reasons we also refrain from using it if the exponents of the power laws are outside the interval $[\frac{1}{50}, 50]$.

3.5 Numerical results

3.5.1 Linear advection

Figure 16 shows the experimental order of accuracy for high order versions of Active Flux using finite differences for linear advection $\partial_t q + c\partial_x q = 0$ ($c = 1$). Numerical methods of orders 3 to 7 are shown. The tests are using a Runge-Kutta integrator of 3rd order accuracy and a small CFL number of 10^{-2} in order for the total error to be dominated by the spatial error. The initial data are

$$q_0(x) = 0.8 + \exp\left(-\frac{(x-0.5)^2}{0.05^2}\right) \tag{59}$$

on grids covering $[0, 1]$. The error is shown at time $t = 0.1$. Limiting is not used. The finite differences employed are FD6 ($a_6 = 2.5$), FD5b ($a_5 = 2$), FD4b ($a_4 = 1.55$), FD3b ($a_3 = 1$), FD1 ($a_4 = 4$). We find that it is not always easy in practice to demonstrate experimentally the desired order of accuracy if the parameters of the finite difference formulas are chosen

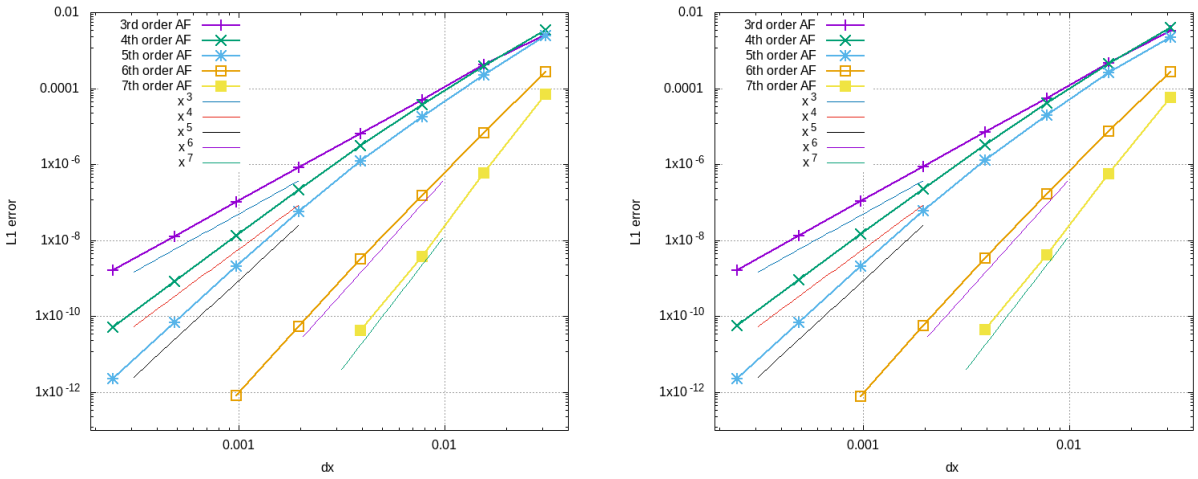

Figure 16: Convergence of the the high order extension of Active Flux via finite differences for linear advection. *Left*: L^1 error of the point values. *Right*: L^1 error of the averages.

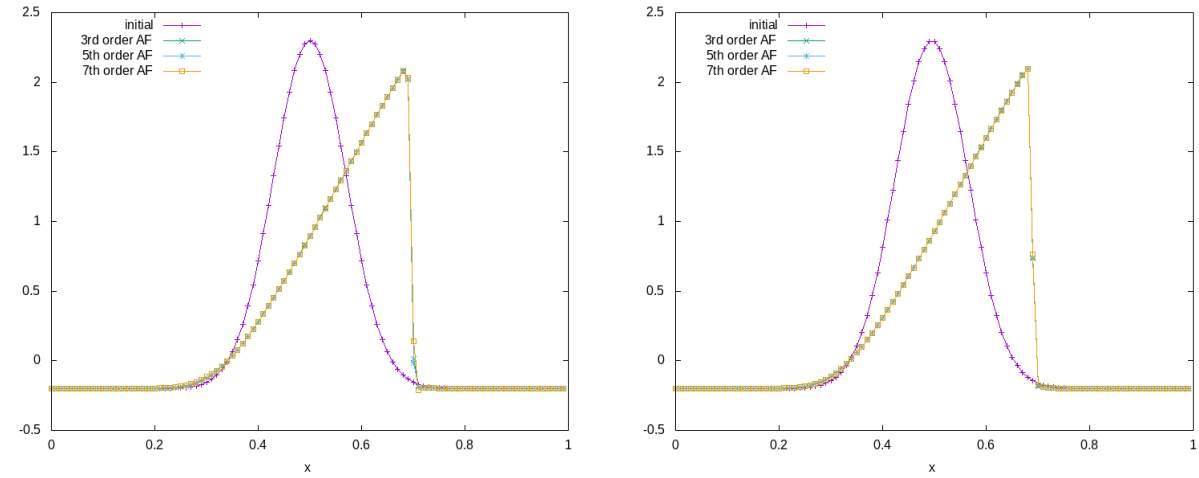

Figure 17: Numerical evolution of Gaussian initial data for Burgers' equation at time $t = 0.1$. *Left*: Point values. *Right*: Averages.

too close to the stability limit. We also find that the choice of the stencil among those of the same order of accuracy can affect the value of the error; a detailed study of this influence shall, however, remain subject of future work. Generally, we find the theoretical orders of accuracy reflected in the simulation results, as Figure 16 shows.

3.5.2 Burgers' equation

Figure 17 shows the numerical time evolution of Gaussian initial data

$$q_0(x) = 2.5 \exp\left(-\frac{(x-0.5)^2}{0.1^2}\right) - 0.2 \tag{60}$$

on a grid of 100 cells covering $[0, 1]$ subject to Burgers' equation. This setup is transonic, but is captured without artifacts due to use of limiting and transonic upwinding as described above. The CFL number used is 0.4.

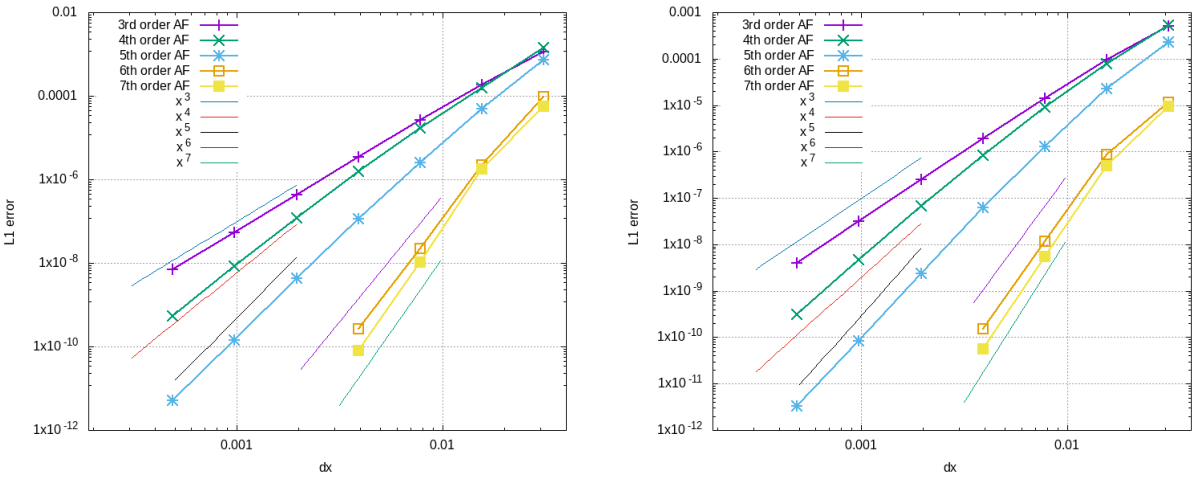

Figure 18: Convergence of the the high order extension of Active Flux via finite differences for Burgers' equation. *Left*: L^1 error of the point values. *Right*: L^1 error of the averages.

Figure 18 shows the experimental convergence on different grids covering $[0, 1]$ using initial datum (59) at time $t = 0.01$, before the formation of a shock. The CFL number is chosen to be 10^{-3}, to ensure that the error is dominated by that of spatial discretization. Here, limiting is not used.

3.5.3 Euler equations

Finally, we assess the performance of the method on the Euler equations

$$\partial_t \rho + \partial_x(\rho v) = 0 \qquad \rho\colon \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}^+ \tag{61}$$

$$\partial_t(\rho v) + \partial_x(\rho v^2 + p) = 0 \qquad v\colon \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R} \tag{62}$$

$$\partial_t e + \partial_x(\rho v) = 0 \qquad e\colon \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}^+ \tag{63}$$

with $e = \frac{p}{\gamma-1} + \frac{1}{2}\rho v^2$. Figure 19 shows the results for the Sod shock tube [Sod78] on a grid of 100 points; the CFL number is 0.25 and the 6th order method is used.

4 High order via further point values

A different viewpoint is that of Active Flux as an enriched Finite Volume method, which includes point values additionally to the cell average. Higher order can thus be obtained by including more point values. In one dimension, the cell interfaces are already "taken" and any new point values must be located inside the cell. With k such points, the generalization of the reconstruction (6) in cell i then is a polynomial $q_{\text{recon},i} \in P^{k+2}$, while the global regularity remains C^0. Therefore, the order of accuracy is expected to be $N = k+3$, while the average number of degrees of freedom per cell is $k + 2$. Denote the locations of the additional points by $\Delta x \xi_j$, $\xi_j \in [-\frac{1}{2}, \frac{1}{2}]$, and their values by $q_{i,j}^n$, $j = 1, \ldots, k$.

Such a method is thus a natural extension of the traditional Active Flux method (Definition 2.1) and is endowed with a leap-frog-type time integration: first, all point values are updated, and then the average is updated with fluxes that already use the new point values. Again, in order to achieve a high order in time quadrature of the flux, point values

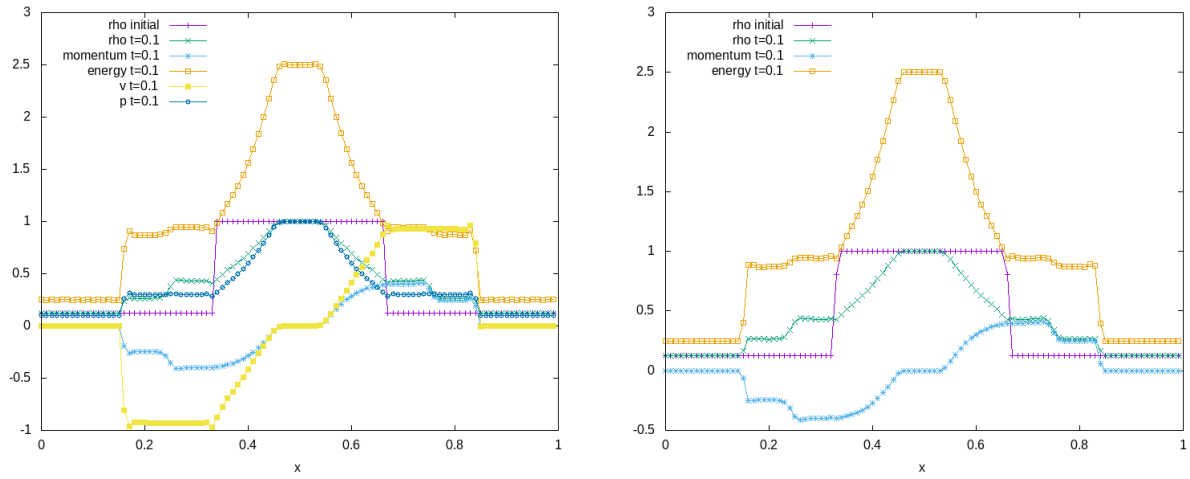

Figure 19: Numerical evolution of the Sod shock tube at time $t = 0.1$. *Left*: Point values. *Right*: Averages.

at cell interfaces are computed at intermediate times as well, while the point values inside the cell are evolved over the full time step directly. The evolution operators, as well as the limiting strategy and the transonic upwinding as described in [Bar21] require only the smallest modifications.

4.1 High-order reconstruction

The place of the traditional reconstruction (6) is taken by a reconstruction in cell i

$$q_{\text{recon},i} \in C^0 \cap L^1_{\text{loc}} \qquad q_{\text{recon},i} \colon (\mathbb{R}^m)^{3+k} \times \left[-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right] \to \mathbb{R}^m \tag{64}$$

satisfying

$$q_{\text{recon},i}\left(q_{i-\frac{1}{2}}, \bar{q}_i, q_{i+\frac{1}{2}}, \pm\frac{\Delta x}{2}\right) = q_{i\pm\frac{1}{2}} \tag{65}$$

$$q_{\text{recon},i}\left(q_{i-\frac{1}{2}}, \bar{q}_i, q_{i+\frac{1}{2}}, \Delta x \xi_j\right) = q_{i,j} \qquad j = 1, \dots, k \tag{66}$$

$$\frac{1}{\Delta x}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \mathrm{d}x \, q_{\text{recon},i}\left(q_{i-\frac{1}{2}}, \bar{q}_i, q_{i+\frac{1}{2}}, x\right) = \bar{q}_i \tag{67}$$

These are $k + 3$ conditions. It is natural to seek the reconstruction in the space P^{2+k} of polynomials which will yield an Active Flux method of N-th order with $k + 2 = N - 1$.

In this case, it is useful to rewrite this interpolation problem (which involves interpolating point values and averages) as a standard pointwise interpolation problem, in order to be able to use standard algorithms.

Theorem 4.1. *Let* $\{(x_j, y_j) \in \mathbb{R}^2\}_{j=1,\dots,N-1}$ *with* $(x_i = x_j) \Rightarrow (i = j)\, \forall i, j$ *and let* $\bar{y} \in \mathbb{R}$ *be given. Fix an arbitrary* $x_0 \neq x_j\, \forall j$*. Then there exist unique polynomials*

$$p_1 \in P^{N-2} \text{ with } p(x_j) = y_j \quad \forall j = 1, \dots, N-1 \tag{68}$$

$$p_2 \in P^{N-1} \text{ with } \begin{cases} p(x_j) = 0 & \forall j = 1, \dots, N-1 \\ p(x_0) = 1 \end{cases} \tag{69}$$

Moreover, if it exists, then the unique polynomial $p \in P^{k+2}$ fulfilling

$$p(x_j) = y_j \qquad j = 1, \dots N-1 \qquad \text{and} \qquad \frac{1}{\Delta x}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p(x)\,\mathrm{d}x = \bar{y} \tag{70}$$

is

$$p(x) = p_1(x) + p_2(x)\cdot \frac{\Delta x \bar{y} - \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_1\,\mathrm{d}x}{\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_2\,\mathrm{d}x} \tag{71}$$

Proof. p_1 and p_2 are defined as standard interpolating polynomials, they thus exist and are unique. By construction, for any $\alpha \in \mathbb{R}$

$$p(x_j) = p_1(x_j) + \alpha p_2(x_j) = y_j \quad \forall j = 1, \dots, N-1 \tag{72}$$

and upon integration

$$\Delta x \bar{y} = \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_{\text{recon}}\,\mathrm{d}x = \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_1\,\mathrm{d}x + \alpha \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_2\,\mathrm{d}x \tag{73}$$

i.e.

$$\alpha = \frac{\Delta x \bar{y} - \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_1\,\mathrm{d}x}{\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_2\,\mathrm{d}x} \tag{74}$$

if $\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_2\,\mathrm{d}x \neq 0$. □

This theorem can be used to define the high-order reconstruction through following identifications:

$$q^n_{\text{recon},i}(x) := p(x) \tag{75}$$

with

$$x_1 := -\frac{\Delta x}{2} \qquad x_j := \Delta x \xi_{j-1} \quad j = 2, \dots, N-2 \qquad x_{N-1} := \frac{\Delta x}{2} \tag{76}$$

$$y_1 := q^n_{i-\frac{1}{2}} \qquad y_j := q^n_{i,j-1} \quad j = 2, \dots, N-2 \qquad y_{N-1} := q^n_{i+\frac{1}{2}} \tag{77}$$

Theorem 4.2. *Consider the setup of Theorem 4.1 and assume the points $\{x_j\}_{j=1,\dots,N-1}$ to be located symmetrically around 0. If their number $N-1$ is odd, then $\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_2\,\mathrm{d}x = 0$, and there is no polynomial of degree $k+2$ fulfilling* (68)–(69)*.*

Proof. Observe that, with the Lagrange formula, p_2 can be given the explicit form

$$p_2(x) = \prod_{j=1}^{N-1} \frac{x - x_j}{x_0 - x_j} \tag{78}$$

Take n to be even, and call $\nu := N/2$. Order the points symmetrically as

$$x_1 < x_2 < \ldots < x_{\nu-1} < x_\nu = 0 < -x_{\nu-1} < \ldots < -x_2 < -x_1 \tag{79}$$

Then, collecting the pairs,

$$p_2(x) = \frac{x}{x_0 - x_\nu} \prod_{j=1}^{\nu-1} \frac{x^2 - x_j^2}{x_0^2 - x_j^2} \tag{80}$$

Thus, p_2 is antisymmetric, and $\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} p_2 \, \mathrm{d}x = 0$. □

This means that we are unable to construct methods of even order with the point values located symmetrically around the cell center. There exist other numerical methods where odd orders of accuracy are given preference, e.g. WENO. If one would be interested to construct an even-order version of Active Flux, one can show that this is possible upon usage of an asymmetric distribution of points.

Definition 4.1 (Active Flux with additional point values)**.** *With the reconstruction $q_{recon,i}$ in cell i defined in* (64)–(67)*, define the global reconstruction*

$$q_{recon} \in C^0 \cap L^1_{loc} \qquad\qquad q_{recon} \colon \mathbb{R} \to \mathbb{R}^m \tag{81}$$

$$q_{recon}(x) := q_{recon,i}\left(q_{i-\frac{1}{2}}, \bar{q}_i, q_{i+\frac{1}{2}}, q_{i,1}, \ldots, q_{i,k}, x - x_i\right) \qquad \text{if } x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \tag{82}$$

The Active Flux method with additional point values *is the following semi-discretization*

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\bar{q}_i(t) = -\dfrac{f(q_{i+\frac{1}{2}}(t)) - f(q_{i-\frac{1}{2}}(t))}{\Delta x} \\ q_{i+\frac{1}{2}}(t) = \left(\textit{solution at } x = x_{i+\frac{1}{2}} \textit{ of the IVP } (1) \textit{ with initial data } q_{recon}\right) + \mathcal{O}(t^{3+k}) \\ q_{i,j}(t) = \left(\textit{solution at } x = x_i + \Delta x \xi_j \textit{ of the IVP } (1) \textit{ with initial data } q_{recon}\right) + \mathcal{O}(t^{3+k}) \end{cases} \tag{83}$$

of (1) *with the interpretations*

$$\bar{q}_i(t) \simeq \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(t,x) \mathrm{d}x \qquad q_{i+\frac{1}{2}}(t) \simeq q(t, x_{i+\frac{1}{2}}) \tag{84}$$

$$q_{i,j}(t) \simeq q(t, x_i + \Delta x \xi_j) \qquad j = 1, \ldots, k \tag{85}$$

A solution operator for linear advection $\partial_t q + c\partial_x q = 0$ is easily given:

$$q_{i+\frac{1}{2}}(t) = q_{\mathrm{recon}}(x_{i+\frac{1}{2}} - ct) = \begin{cases} q_{\mathrm{recon},i}(\Delta x/2 - ct) & c > 0, ct < \Delta x \\ q_{\mathrm{recon},i+1}(-\Delta x/2 - ct) & c < 0, |c|t < \Delta x \end{cases} \tag{86}$$

A solution operator for scalar conservation laws based on a fixpoint iteration to find the characteristic is given in [Bar21] for any order of accuracy. To this end, the equation for the foot point x_0 of the characteristic

$$x_0 + f'(q_{\text{recon}}(x_0))t = x_{i+\frac{1}{2}} \tag{87}$$

is solved iteratively:

$$x_0^{(0)} := x_{i+\frac{1}{2}} \tag{88}$$

$$x_0^{(\ell)} := x_{i+\frac{1}{2}} - f'(q_{\text{recon}}(x_0^{(\ell-1)}))t \qquad \ell = 1, \dots, \ell_{\max} \tag{89}$$

An approximate solution operator is

$$q_{i+\frac{1}{2}}(t) \simeq q_{\text{recon}}(x_0^{(\ell_{\max})}) = q(t, x_{i+\frac{1}{2}}) + \mathcal{O}(t^{\ell_{\max}+1}) \tag{90}$$

A third order accurate evolution operator for hyperbolic systems of conservation laws can also be found in [Bar21], and an approximate evolution operator for inhomogeneous problems has been given in [BBK21],[BB20].

4.2 Time integration and stability conditions

The high-order extension via additional point values allows to use the very successful leap-frog-type integrator of the traditional Active Flux method.

For linear advection $\partial_t q + c\partial_x q = 0$ with $c > 0$, the algorithm extending (20)–(21) thus reads

$$q_{i+\frac{1}{2}}^{n+\frac{\ell}{M-1}} = q_{\text{recon}}^n(x_{i+\frac{1}{2}} - c\zeta_\ell \Delta t) \qquad \ell = 1, \dots, M-1 \tag{91}$$

$$q_{i,j}^{n+1} = q_{\text{recon}}^n(x_i + \Delta x \xi_j - c\Delta t) \qquad j = 1, \dots, N-3 \tag{92}$$

$$\bar{q}_i^{n+1} = \bar{q}_i^n - \Delta t \sum_{\ell=0}^{M-1} \omega_\ell \frac{f(q_{i+\frac{1}{2}}^{n+\frac{\ell}{M-1}}) - f(q_{i-\frac{1}{2}}^{n+\frac{\ell}{M-1}})}{\Delta x} \tag{93}$$

Here, M is the number of points in the time-quadrature approximating the time-step-averaged flux

$$\frac{1}{\Delta t} \int_0^{\Delta t} f(q_{i+\frac{1}{2}}(t)) \mathrm{d}t \simeq \sum_{\ell=0}^{M-1} \omega_\ell f(q_{i+\frac{1}{2}}^{n+\frac{\ell}{M-1}}) \tag{94}$$

(with $\zeta_0 = 0$, $\zeta_{M-1} = 1$), and $\{\omega_\ell\}_{\ell=0,\dots,M-1}$ the corresponding weights. It is natural to use a Gauss-Lobatto quadrature here. Clearly, $q_{i+\frac{1}{2}}^{n+\frac{\ell}{M-1}} \simeq q_{i+\frac{1}{2}}(\Delta t \zeta_\ell)$. Note that we use the notation $q_{i+\frac{1}{2}}^{n+\frac{\ell}{M-1}}$ to denote intermediate values, but we do not actually insist on them being equidistant in time. See Figure 20 for an example.

Note that according to the definition of q_{recon} in (82) the foot of the characteristic in (92) is evaluated in the "correct" cell: if

$$\Delta x \xi_j - c\Delta t > -\frac{\Delta x}{2} \tag{95}$$

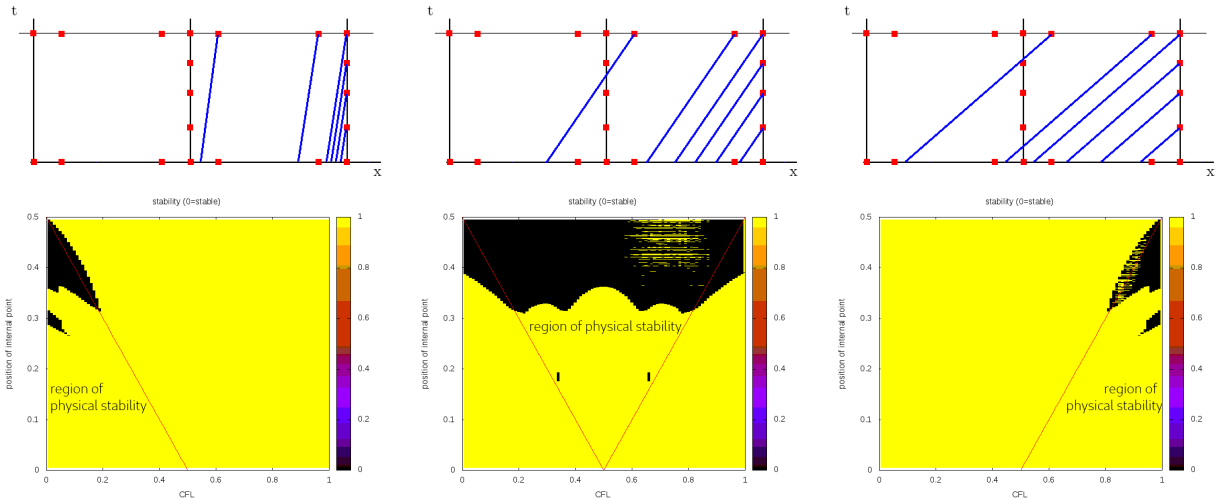

Figure 20: Two point values located symmetrically $\xi_2 = -\xi_1$ (fifth order method) with $M = 5$ equidistant quadrature points in time. Lower panel shows stability as a function of the location of the right internal point value (ξ_2) and the CFL number (black = stable). The noise in the central and right figure has to do with the sampling and with one of the eigenvalues being exactly one; physical stability regions (i.e. those combinations of point location and CFL for which the foot of the characteristic is actually in the cell whose reconstruction is used in the update) are to the left, between and to the right of the red lines, respectively. One observes that actual stability also occurs outside the region of physical stability, while certain parameter choices inside the physical stability region turn out to be actually unstable.

then information is taken from cell i, otherwise, for

$$\Delta x \xi_j - c\Delta t \leq -\frac{\Delta x}{2} \tag{96}$$

from cell $i-1$, (see Figure 20).

This guarantees optimal stability, as can be seen in Figure 20 for 5th order, which shows stability analysis associated to different cases: for small CFL numbers, both characteristics used in the update of the internal point values have their foot points in the same cell; for intermediate CFL numbers, one of them reaches out into the neighbouring cell, and for large CFL numbers, the foot points of both are in the neighbouring cell. One observes that taken together, these yield a stability result valid for all CFL numbers between 0 and 1. At the same time, one remarks that, surprisingly, the location of the internal point values is not arbitrary. For stability, the internal point values have to be distributed fairly close to the cell interfaces, with $\xi_1 = -\xi_2 \gtrsim 0.4$. We are not aware of a simple reason that would explain this finding. Once inside the stable domain, we do not find the precise location of the point values to have a visible influence on the simulation results.

At the same time, from Fig. 20 (middle) one deduces that the case of only one of the characteristics reaching out into the left cell is stable even outside the physical region of stability.

4.3 Transonic upwinding

We use the same algorithm as the one described in [Bar21]; the reader is referred to this work for more details. In short, for scalar conservation laws, the iterative approximation to the foot point of the characteristic passing through $x_{i+\frac{1}{2}}$ at time t is initialized with two values ($q^n_{i-\frac{1}{2}}$ and $q^n_{i+\frac{3}{2}}$ instead of $q^n_{i+\frac{1}{2}}$), see Equation (88). Every iteration increases the order of accuracy of the evolution operator by one. After a sufficient number of iterations, the speeds of the two estimated characteristics are compared, and the one with the largest speed in absolute value is selected. This treatment of the transonic case does not affect the accuracy of the evolution operator. For systems, the approximate evolution operator is evaluated at different initialization locations. Then, the one with the largest sum of absolute values of the eigenvalues is chosen; for further details the reader is again referred to [Bar21].

4.4 Limiting

The limiting strategy is to check whether the reconstruction exceeds the minimum/maximum of $\{\bar{q}_i, q_{i\pm\frac{1}{2}}\} \cup \{q_{i,j}\}_{j=1,\dots,k}$, and to use a parabolic/power law reconstruction in that case. As a consequence, the internal point values are successively ignored until for the parabolic/power law reconstruction only the point values at cell interfaces contribute. This latter part is again the limiting strategy from [Bar21], as described in Section 2.2.

As it is too costly to evaluate the presence of local extrema analytically for high order polynomials, the value of the reconstruction is checked at 10 locations inside the cell.

4.5 Numerical results

4.5.1 Linear advection

For linear advection with speed 1, the initial data (59) are evolved using a CFL number of 0.5 on grids covering $[0, 1]$. For 5th order accuracy, two additional points values located at $\xi_{1,2} = -0.415, 0.415$ were used, as well as a Gauss-Lobatto quadrature for (94) with $M = 4$ points in time. For 7th order, 4 additional points were located at $\{-0.48, -0.41, 0.41, 0.48\}$, and the Gauss-Lobatto quadrature used $M = 5$ points in time. The exact evolution operator (20) is used. The setup is run without limiting. Figure 21 shows the experimental convergence at time $t = 0.01$.

4.5.2 Burgers' equation

For Burgers' equation we consider again transonic initial data (60) on a grid of 100 cells covering $[0, 1]$. Figure 22 shows the numerical solution at time $t = 0.1$ after shock formation. A CFL number of 0.5 has been used. For 5th order accuracy, two additional points values located at $\xi_{1,2} = -0.415, 0.415$ were used, as well as a Gauss-Lobatto quadrature for (94) with $M = 4$ points in time. For 7th order, 4 additional points were located at $\{-0.48, -0.41, 0.41, 0.48\}$, and the Gauss-Lobatto quadrature used $M = 5$ points in time. In both cases, the approximate evolution operator (88)–(89) was used, with an adequate number of fixpoint iterations according to (90).

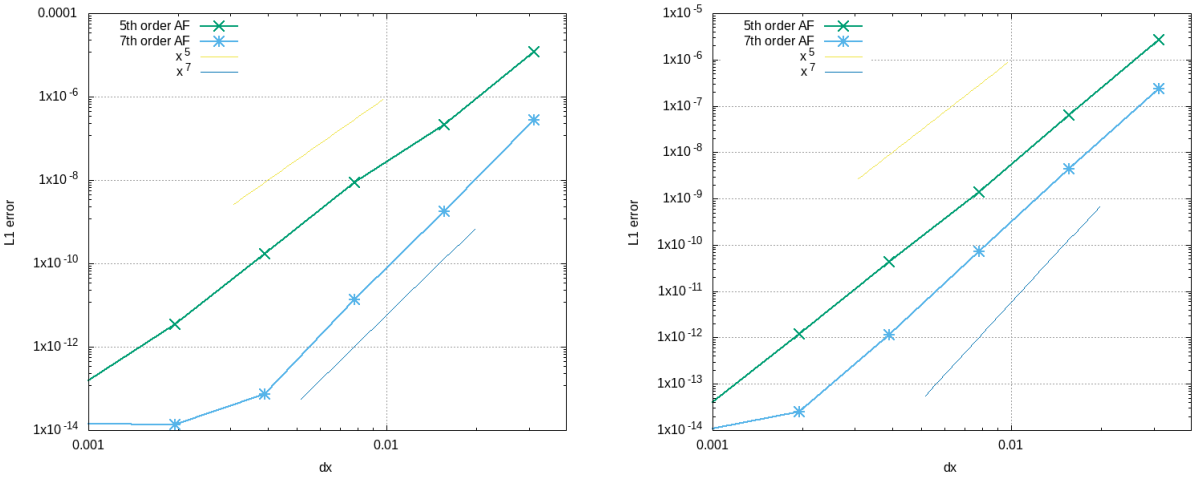

Figure 21: Convergence of the the high order extension of Active Flux via additional point values for linear advection. *Left*: L^1 error of the point values. *Right*: L^1 error of the averages.

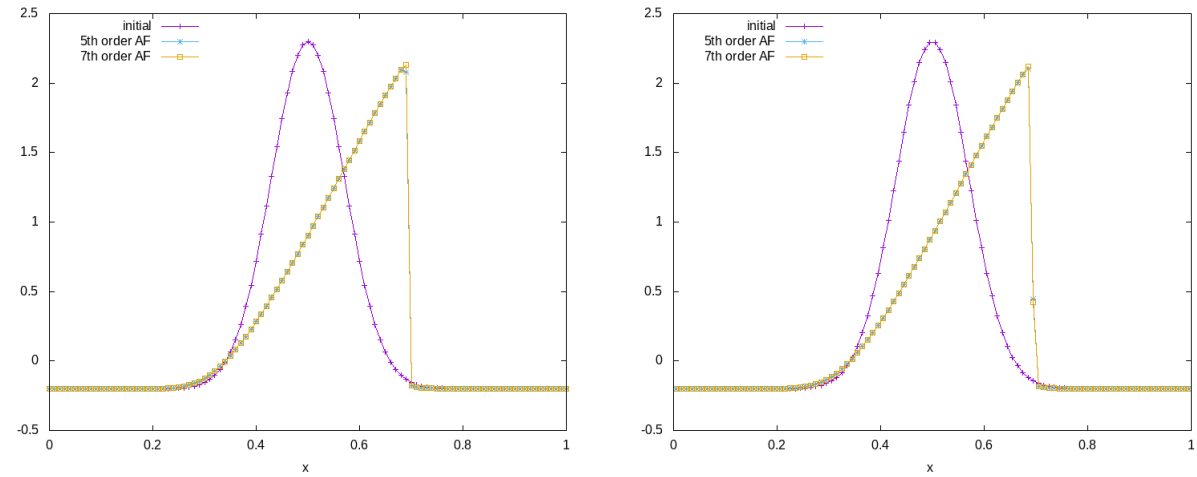

Figure 22: Numerical evolution of Gaussian initial data for Burgers' equation at time $t = 0.1$. *Left*: Point values. *Right*: Averages.

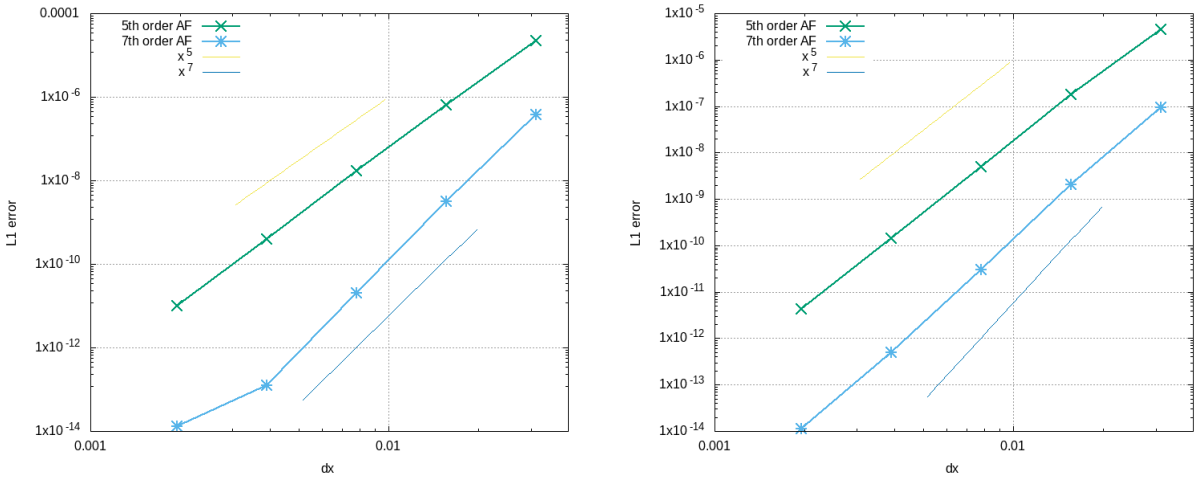

Figure 23: Convergence of the the high order extension of Active Flux via additional point values for Burgers' equation. *Left*: L^1 error of the point values. *Right*: L^1 error of the averages.

Figure 23 shows convergence results at time $t = 0.01$ prior to shock formation. Here, limiting has not been used, but the transonic upwinding was following the procedure described in 4.3.

4.5.3 Euler equations

The Sod shock tube is evolved on a grid of 100 points covering $[0, 1]$, using a CFL number of 0.5. The additional point values are located at $\xi_{1,2} = \pm 0.415$, the flux quadrature (94) is performed using a Gauss-Lobatto rule involving $M = 4$ points in time, a setup corresponding to 5th order accuracy. The approximate evolution operator is described in [Bar21]. It is formally third order accurate only; however, this can be seen in parallel to the common usage of low order time integrators (e.g. RK3) for spatially high order methods. An extension of the approximate evolution operator from [Bar21] to higher order does not seem generally impossible, but for now remains subject of future work. Numerical results at $t = 0.1$ are shown in Figure 24.

5 High order via further moments

Finally, it is possible to interpret Active Flux in a way closer to Finite Element methods, by declaring higher moments to be new degrees of freedom in a high-order version. It is thus a complementary view to the previous section.

Define the p-th moment as

$$q_i^{(p)} := A_p \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} x^p q(t, x_i + x)\,\mathrm{d}x \tag{97}$$

The normalization $A_p = \frac{(p+1)2^p}{\Delta x^{p+1}}$ is chosen such that $q(t, x) \equiv 1$ implies $q^{(p)} = 1\ \forall p \in 2\mathbb{N}_0$. The first values are $(A_p)_p = \left(\frac{1}{\Delta x}, \frac{4}{\Delta x^2}, \frac{12}{\Delta x^3}, \frac{32}{\Delta x^4}, \frac{80}{\Delta x^5}, \ldots\right)$ (sequence A001787 in the Online Encyclopedia of Integer Sequences (OEIS)).

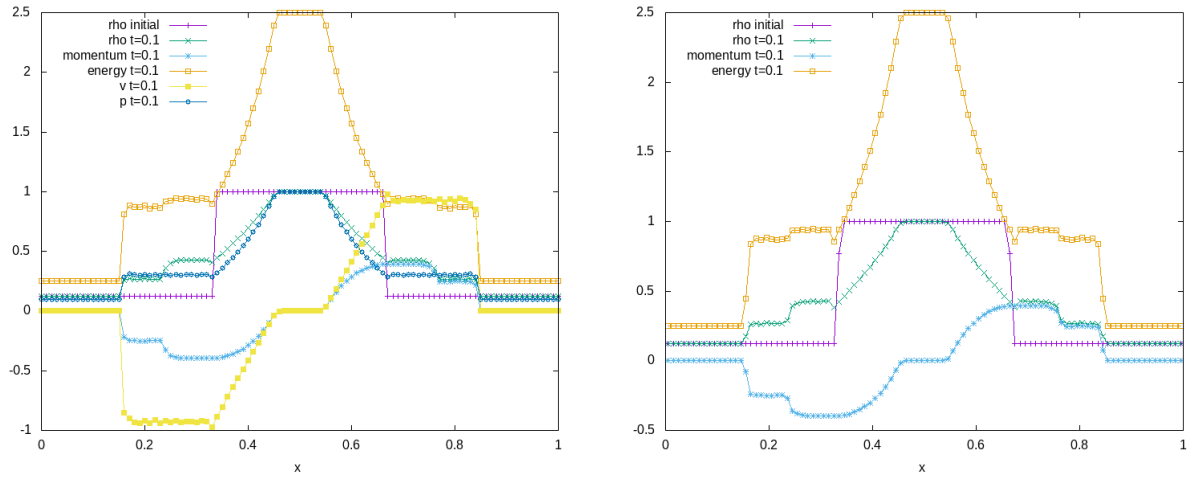

Figure 24: Numerical evolution of the Sod shock tube at time $t = 0.1$ using the high order extension of Active Flux via additional point values. *Left*: Point values. *Right*: Averages.

With k moments, every cell has access to $k+2$ pieces of information and thus, a polynomial reconstruction of degree $k+1$ can be constructed in every cell. The order of accuracy of the method is thus expected to be $k+2$. As the point values are shared, on average there are $k+1$ degrees of freedom per cell. The global reconstruction remains C^0.

5.1 Update of the moments

Multiplying (1) by x^p and integrating yields the evolution equation for the moments:

$$A_p \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} x^p \partial_t q \,\mathrm{d}x + A_p \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} x^p \partial_x f(q) \,\mathrm{d}x = 0 \tag{98}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} q_i^{(p)} + A_p \left(\left(\frac{\Delta x}{2} \right)^p f(q_{i+\frac{1}{2}}) - \left(-\frac{\Delta x}{2} \right)^p f(q_{i-\frac{1}{2}}) \right) - pA_p \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} x^{p-1} f(q) \,\mathrm{d}x = 0 \tag{99}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} q_i^{(p)} + (p+1) \frac{f(q_{i+\frac{1}{2}}) - (-1)^p f(q_{i-\frac{1}{2}})}{\Delta x} - pA_p \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} x^{p-1} f(q) \,\mathrm{d}x = 0 \tag{100}$$

For $p = 0$ one just has the usual equation for the average $\bar{q}_i \equiv q_i^{(0)}$

$$\frac{\mathrm{d}}{\mathrm{d}t} q_i^{(0)} + \frac{f(q_{i+\frac{1}{2}}) - f(q_{i-\frac{1}{2}})}{\Delta x} = 0 \tag{101}$$

For linear flux $f(q) = cq$, the integral in (100) can be expressed as a moment of one order less:

$$\frac{\mathrm{d}}{\mathrm{d}t} q_i^{(p)} + (p+1) c \frac{q_{i+\frac{1}{2}} - (-1)^p q_{i-\frac{1}{2}}}{\Delta x} - \frac{2(p+1)}{\Delta x} \frac{p 2^{p-1}}{\Delta x^p} c \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} x^{p-1} q \,\mathrm{d}x = 0 \tag{102}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} q_i^{(p)} + (p+1) c \frac{q_{i+\frac{1}{2}} - (-1)^p q_{i-\frac{1}{2}}}{\Delta x} - \frac{2(p+1)}{\Delta x} c q_i^{(p-1)} = 0 \quad (p \geq 1) \tag{103}$$

such that, for example, the first three moments thus evolve according to

$$\frac{\mathrm{d}}{\mathrm{d}t}q_i^{(0)} + c\frac{q_{i+\frac{1}{2}} - q_{i-\frac{1}{2}}}{\Delta x} = 0 \tag{104}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}q_i^{(1)} + 2c\frac{q_{i+\frac{1}{2}} + q_{i-\frac{1}{2}}}{\Delta x} - \frac{4}{\Delta x}cq_i^{(0)} = 0 \tag{105}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}q_i^{(2)} + 3c\frac{q_{i+\frac{1}{2}} - q_{i-\frac{1}{2}}}{\Delta x} - \frac{6}{\Delta x}cq_i^{(1)} = 0 \tag{106}$$

Observe that the evolution of the moments is exact (up to a possible quadrature error upon evaluating the integral in (100) in case of a nonlinear f). Apart from quadrature errors, the accuracy of the method is determined entirely by the choice of the point value update.

Definition 5.1 (Active Flux with higher moments)**.** *The* Active Flux method with higher moments *is the following semi-discretization*

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}q_i^{(p)} = -(p+1)\dfrac{f(q_{i+\frac{1}{2}}) - (-1)^p f(q_{i-\frac{1}{2}})}{\Delta x} + pA_p\displaystyle\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} x^{p-1}f(q)\,\mathrm{d}x \quad p = 0,\dots,k \\ \\ \dfrac{\mathrm{d}}{\mathrm{d}t}q_{i+\frac{1}{2}}(t) = -R\Big(q_{i-\frac{1}{2}}(t), q_i^{(0)}(t),\dots,q_i^{(k)}(t), q_{i+\frac{1}{2}}(t), q_{i+1}^{(0)}(t),\dots,q_{i+1}^{(k)}(t), q_{i+\frac{3}{2}}(t)\Big) \qquad i\in\mathbb{Z}, k\geq 0 \end{cases} \tag{107}$$

of (1) *with* $A_p = \frac{(p+1)2^p}{\Delta x^{p+1}}$ *and the interpretations*

$$q_i^{(p)}(t) \simeq A_p\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} x^p q(t,x)\mathrm{d}x \qquad\qquad q_{i+\frac{1}{2}}(t) \simeq q(t, x_{i+\frac{1}{2}}) \tag{108}$$

and R *a consistent approximation of* $\partial_x f(q)$ *at* $x_{i+\frac{1}{2}}$.

5.2 Update of the point values

We propose to update the point value with a finite difference approximation to the derivative, obtained as the derivative of the reconstruction. This means that this finite difference has a compact stencil, and only involves the degrees of freedom of one of the adjacent cells.

For an N-th order accurate method, define the unique polynomial $q_{\mathrm{recon},i} : \left[-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right] \to \mathbb{R}^m$ of order $N-1$ fulfilling

$$q_{\mathrm{recon},i}\left(\pm\frac{\Delta x}{2}\right) = q_{i\pm\frac{1}{2}} \qquad A_p\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} x^p q_{recon,i}(x)\,\mathrm{d}x = q_i^{(p)} \qquad p = 0,\dots,N-3 \tag{109}$$

For example, the lowest order polynomial with given values $q_{i\pm\frac{1}{2}}$ and moments 0 through 2 is

$$\frac{1}{16}q_{i+\frac{1}{2}}(2\xi+1)(3-30\xi-60\xi^2+280\xi^3) + \frac{1}{16}q_{i-\frac{1}{2}}(2\xi-1)(-3-30\xi+60\xi^2+280\xi^3) \tag{110}$$

$$+(2\xi-1)(2\xi+1)\left(\frac{15}{16}q^{(0)}(28\xi^2-3) - \frac{15}{2}q^{(1)}\xi - \frac{35}{16}q^{(2)}(20\xi^2-1)\right)$$

where $\xi = x/\Delta x$.

A finite difference approximation to the derivative at $x = \frac{\Delta x}{2}$ is obtained by evaluating $(\partial_x q_{\mathrm{recon},i})\left(\frac{\Delta x}{2}\right)$. It necessarily is a linear function of $q_{i+\frac{1}{2}}, q_{i-\frac{1}{2}}, q_i^{(0)}, \ldots, q_i^{(N-3)}$ and therefore can be written as

$$\frac{1}{\Delta x}\left(b_{\frac{1}{2}} q_{i+\frac{1}{2}} + b_{-\frac{1}{2}} q_{i-\frac{1}{2}} + \sum_{j=0}^{N-3} b_0^{(j)} q_i^{(j)}\right) = q'(x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^N) \tag{111}$$

Notation 5.1. *A finite difference approximation of the form* (111) *shall be depicted graphically as follows:*

$$\frac{1}{\Delta x}\left(b_{\frac{1}{2}} q_{i+\frac{1}{2}} + b_{-\frac{1}{2}} q_{i-\frac{1}{2}} + \sum_{j=0}^{N-3} b_0^{(j)} q_i^{(j)}\right) = \left[\begin{array}{c|c||} & b_0^{(N-3)} \\ & \vdots \\ & b_0^{(1)} \\ & b_0^{(0)} \\ b_{-\frac{1}{2}} & b_{\frac{1}{2}} \end{array}\right] \tag{112}$$

Observe the consistency of the notation with that introduced in (38)*. Note that, again, the double vertical line indicates the cell interface at which the finite difference formula is an approximation of the derivative to the given order.*

The finite differences constructed in the way described above are unique for every order of accuracy; to lowest orders they read[3]

$$\mathrm{MD3} = \left[\begin{array}{c|c||} & -6 \\ 2 & 4 \end{array}\right] \qquad \mathrm{MD5} = \left[\begin{array}{c|c||} & -35 \\ & -15 \\ & 15 \\ 4 & 16 \end{array}\right] \qquad \mathrm{MD7} = \left[\begin{array}{c|c||} & -\frac{693}{4} \\ & -\frac{315}{4} \\ & \frac{315}{2} \\ & \frac{105}{2} \\ & -\frac{105}{4} \\ 6 & 36 \end{array}\right] \tag{113}$$

Note that the coefficients of even moments and point values have to add up to 0. For linear advection $\partial_t q + c\partial_x q = 0$ for positive c, the update with MD5 amounts to

$$\partial_t q_{i+\frac{1}{2}} = -c\frac{4q_{i-\frac{1}{2}} + 15q_i^{(0)} - 15q_i^{(1)} - 35q_i^{(2)} + 16q_{i+\frac{1}{2}}}{\Delta x} \tag{114}$$

For seventh order, the derivative approximation is MD7

$$3\frac{8q_{i-\frac{1}{2}} - 35q_i^{(0)} + 70q_i^{(1)} + 210q_i^{(2)} - 105q_i^{(3)} - 231q_i^{(4)} + 48q_{i+\frac{1}{2}}}{4\Delta x} \tag{115}$$

Define again by D^* the flipped version of the finite difference D as follows:

[3]MD is supposed to mean *moment difference*.

$$D = \begin{bmatrix} \vert & b_0^{(N-3)} & \Vert \\ & \vdots & \\ & b_0^{(2)} & \\ & b_0^{(1)} & \\ & b_0^{(0)} & \\ b_{-\frac{1}{2}} & & b_{\frac{1}{2}} \end{bmatrix} \qquad D^* = \begin{bmatrix} \Vert & \pm b_0^{(N-3)} & \vert \\ & \vdots & \\ & -b_0^{(2)} & \\ & b_0^{(1)} & \\ & -b_0^{(0)} & \\ -b_{\frac{1}{2}} & & -b_{-\frac{1}{2}} \end{bmatrix} \tag{116}$$

Observe that upon a reflection of space $x \mapsto -x$, the sign of the derivative changes, but also that of odd moments.

These flipped finite differences can of course be obtained by evaluating the derivative of the polynomial reconstruction at $-\frac{\Delta x}{2}$ and by performing an index shift. For example, the derivative of (110) at $\frac{\Delta x}{2}$ is

$$\frac{4q_{i-\frac{1}{2}} + 15q_i^{(0)} - 15q_i^{(1)} - 35q_i^{(2)} + 16q_{i+\frac{1}{2}}}{\Delta x} \tag{117}$$

while the one at $-\frac{\Delta x}{2}$ is

$$\frac{-4q_{i+\frac{1}{2}} - 15q_i^{(0)} - 15q_i^{(1)} + 35q_i^{(2)} - 16q_{i-\frac{1}{2}}}{\Delta x} \tag{118}$$

Given these finite differences, the semi-discrete method is Equations (100) and

$$\frac{\mathrm{d}}{\mathrm{d}t} q_{i+\frac{1}{2}}(t) = -\Big(f'(\tilde{q}_{i+\frac{1}{2}})^+ D + f'(\tilde{q}_{i+\frac{1}{2}})^- D^* \Big) \tag{119}$$

The point value update is thus in close analogy to (47), with D now being one of the moment differences (113).

5.3 Time integration and stability results

Equations (100) and (119) constitute a system of ODEs that we propose to integrate in time using e.g. SSP-RK3. For linear advection, MD5 is found to be stable for CFL ≤ 0.13 when using RK3 (see Section A). Experiments also suggest stability CFL < 0.17 for RK5. For MD7, RK3 yields stability until CFL ≤ 0.066 (experimentally, CFL ≤ 0.1 for RK7).

5.4 Limiting

As in the previous versions, we propose to gradually decrease the order of the method upon detection of oscillations. The finite differences employed in the point value update are constructed by differentiating[4] a reconstruction. If the reconstruction is found to be non-monotone, the highest moment is neglected and a reconstruction of lower polynomial degree considered, until one either arrives at a monotone reconstruction, or at the parabolic reconstruction that employs the two point values and the average. In case it also is non-monotone, we propose once more to follow the limiting strategy of [Bar21] by replacing the

[4]In fact, for the ease of the algorithm we perform the differentiation numerically.

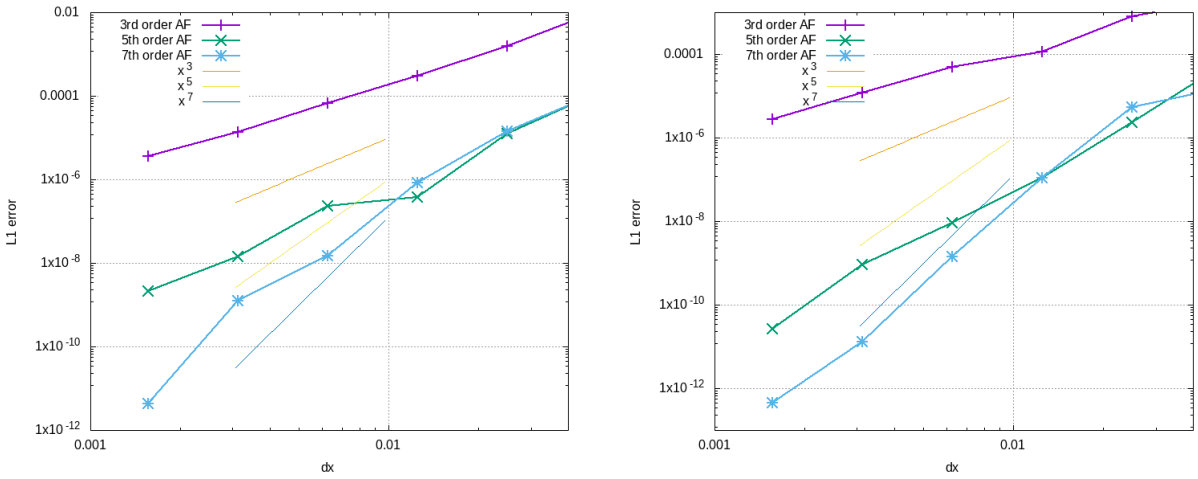

Figure 25: Convergence of the the high order extension of Active Flux via higher moments for linear advection. *Left*: L^1 error of the point values. *Right*: L^1 error of the averages.

parabola by a power law (27)–(28), whenever a monotone reconstruction is at all possible. Their derivatives at cell interfaces are given in (55)–(56). For polynomials of degree higher than 2, monotonicity is checked approximately upon evaluation of the polynomial at 5 locations inside the cell.

5.5 Transonic upwinding

We employ the same choice of $\tilde{q}_{i+\frac{1}{2}}$ as in Section 3.3.

5.6 Numerical results

5.6.1 Linear advection

Convergence results for linear advection with speed 1 are shown in Figure 25. Here, limiting is not used, and as before a very small CFL number of 10^{-4} is used with RK3 in order for the error to be dominated by that of spatial discretization. The initial data are again (59).

5.6.2 Burgers' equation

Figure 26 shows the numerical evolution of transonic initial data (60) for Burgers' equation for high order Active Flux via higher moments at $t = 0.1$. The grid consists of 100 cells covering $[0, 1]$ and the CFL number is 0.01. Limiting is used, as is the transonic upwinding modification. The shock is resolved well and no artefacts at the sonic point are visible.

Figure 27 shows the experimental order of convergence for initial data (59) at time $t = 0.01$, before the formation of a shock. No limiting is used, and the CFL number is again 10^{-4} (RK3) in order to ensure that the error is dominated by that of spatial discretization.

5.6.3 Euler equations

Finally, a Sod shock tube is solved on a grid of 100 cells covering $[0, 1]$. The initial discontinuities are located at $x = \frac{1}{3}$ and $x = \frac{2}{3}$, such that it is possible to use periodic boundaries.

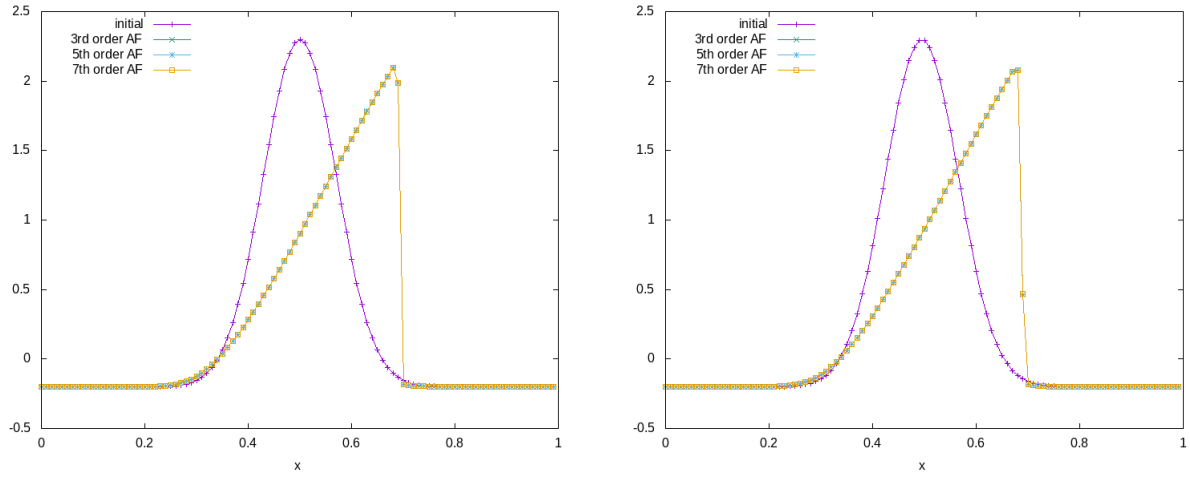

Figure 26: Numerical evolution of Gaussian initial data for Burgers' equation at time $t = 0.1$. Setup as in Section 3.5.2. *Left*: Point values. *Right*: Averages.

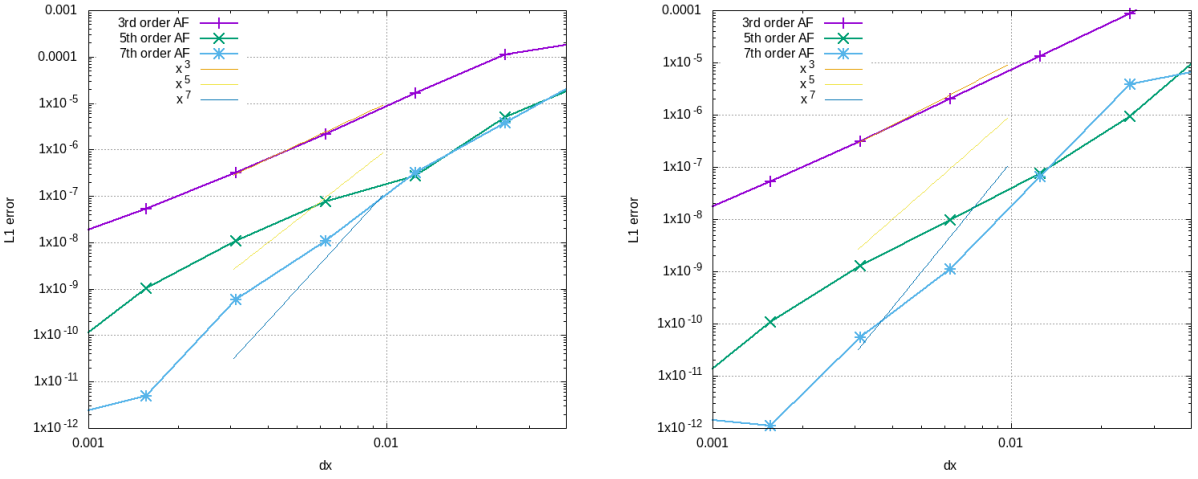

Figure 27: Convergence of the the high order extension of Active Flux via higher moments for Burgers' equation. *Left*: L^1 error of the point values. *Right*: L^1 error of the averages.

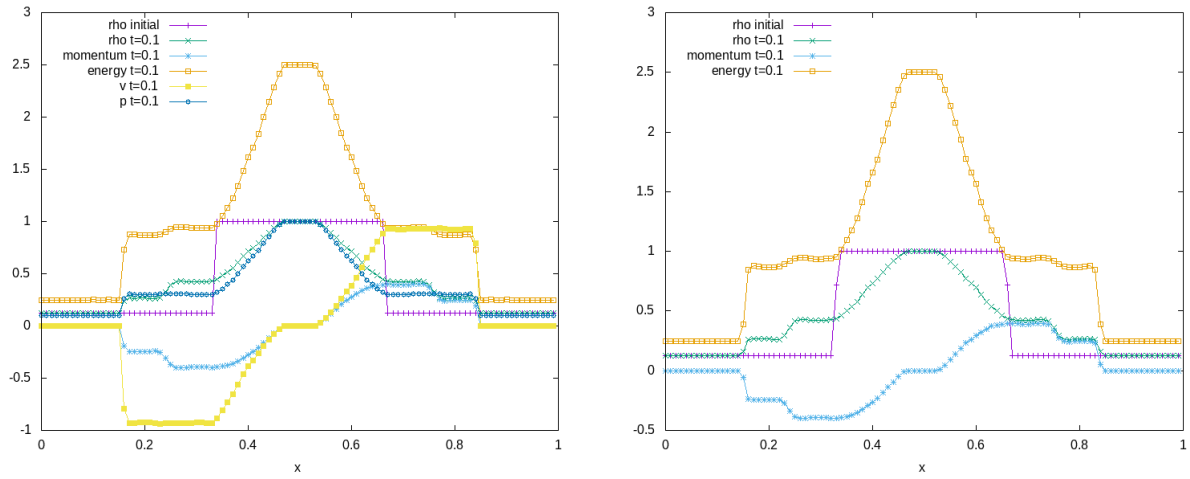

Figure 28: Numerical evolution of the Sod shock tube at time $t = 0.1$. *Left*: Point values. *Right*: Averages.

The numerical method employed is of 7th order accuracy in space; time integration is performed using RK3 with a CFL number of 0.01.

6 Higher order via higher derivatives

Yet another high-order extension of Active Flux is presented in [Roe21], where *spatial derivatives* are stored at cell interfaces. When in every cell an average is stored, and at every cell interface a point value and derivatives of order $\leq s$, the modified Hermite interpolation problem gives rise to a polynomial P^{2s+2} in every cell, and simultaneously augments the global regularity of the reconstruction to C^s. The average number of degrees of freedom per cell is only $s + 2$ and the order of the scheme is no higher than $2s + 3$. More information is stored at the cell interfaces, which reduces the memory requirements of the method. Note, though, that the spatial derivatives cannot be used to increase the order of accuracy for the flux quadrature in time. Also, the evolution of spatial derivatives may be cumbersome, and the extension of the approach to multiple spatial dimensions at the moment remains unexplored. The approximate evaluation of the Schur-Miller criterion (Section A) suggests that, for linear advection, the 5th order version of this method is stable up to CFL = 1.

7 Conclusions and outlook

Active Flux, as it was presented so far since its first appearance in [vL77], was restricted to third order accuracy. During this time, the degrees of freedom of Active Flux in one sptial dimension always have been a cell average and a point value at every cell interface, which means that a parabola could naturally be used as a reconstruction in each cell. Third order accuracy has been so natural, that any attempt to go beyond third order must question the current foundations of the method, the understanding of "what" Active Flux is. In this paper, we associate three choices of an extension to higher order with three different interpretations of Active Flux. In all these interpretations we, however, have

maintained two special properties of Active Flux, that we consider to be its distinctive features: continuity of the approximation and the presence of point values and cell averages among the degrees of freedom. The solution is considered globally continuous and, contrary to Godunov methods, Riemann Problems are never introduced in order to compute a numerical flux. The presence of point values (whose update includes upwinding) ensures stability, and allows to use non-conservative formulations of the equations for their update. At the same time, the presence of an average that is updated conservatively ensures the correct approximation of weak solutions.

The first view of Active Flux presented in this paper is that of a coupled Finite Volume/Finite Difference method (A) with a cell average updated according to the Finite Volume approach and point values, whose update includes finite difference approximations on a stencil whose size depends on the order of accuracy. The second view is that of an enriched Finite Volume method (B), with a cell average, and point values not only on cell interfaces but also inside the cell. The third view is that of a coupled Finite Element/Finite Difference method (C), with degrees of freedom being moments of the solution (starting with the average) and point values at cell interfaces that are used to update the moments.

The three high order extensions have very different properties. The characteristics-based update of the point values inside the cell for the variant B is the one that is most unorthodox, but also the one with the best stability properties. It is also closest to the time integration employed in the original Active Flux method. The usage of moments (variant C) as additional degrees of freedom maintains the elegance of a compact stencil and is the best candidate for further extensions, but also has the smallest CFL numbers among the considered variants. Both the variant A and C come at the advantage that their application to any hyperbolic system is immediate, and not even a solution of a Riemann Problem is required.

Future work will be devoted to an extension of these methods to multiple spatial dimensions. As the Active Flux method is fairly recent (at least for nonlinear problems), certain questions that might seem settled for Finite Volume methods are still far from being conclusively answered for Active Flux. This concerns limiting as well as the convergence to the correct entropy solution. Although this paper presents viable approaches, more theoretical studies will be necessary, which surely will bring further improvements.

References

[Abg20] Rémi Abgrall. A combination of residual distribution and the active flux formulations or a new class of schemes that can combine several writings of the same hyperbolic problem: application to the 1d euler equations. *arXiv preprint arXiv:2011.12572*, 2020.

[Bar21] Wasilij Barsukow. The active flux scheme for nonlinear problems. *Journal of Scientific Computing*, 86(1):1–34, 2021.

[BB20] Wasilij Barsukow and Jonas Berberich. A well-balanced Active Flux scheme for the shallow water equations with wetting and drying. *submitted*, 2020.

[BBK21] Wasilij Barsukow, Jonas P Berberich, and Christian Klingenberg. On the Active Flux scheme for hyperbolic PDEs with source terms. *SIAM Journal on Scientific Computing*, 43(6):A4015–A4042, 2021.

[BHKR19] Wasilij Barsukow, Jonathan Hohm, Christian Klingenberg, and Philip L Roe. The active flux scheme on Cartesian grids and its low Mach number limit. *Journal of Scientific Computing*, 81(1):594–622, 2019.

[BK22] Wasilij Barsukow and Christian Klingenberg. Exact solution and a truly multidimensional Godunov scheme for the acoustic equations. *ESAIM: M2AN*, 56(1), 2022.

[CHK21] Erik Chudzik, Christiane Helzel, and David Kerkmann. The cartesian grid active flux method: Linear stability and bound preserving limiting. *Applied Mathematics and Computation*, 393:125501, 2021.

[Coh22] Arthur Cohn. Über die Anzahl der Wurzeln einer algebraischen Gleichung in einem Kreise. *Mathematische Zeitschrift*, 14(1):110–148, 1922.

[ER11] Timothy A Eymann and Philip L Roe. Active flux schemes for systems. In *20th AIAA computational fluid dynamics conference*, 2011.

[ER13] Timothy A Eymann and Philip L Roe. Multidimensional active flux schemes. In *21st AIAA computational fluid dynamics conference*, 2013.

[Fan17] Duoming Fan. *On the acoustic component of active flux schemes for nonlinear hyperbolic conservation laws*. PhD thesis, University of Michigan, Dissertation, 2017.

[GM04] Hervé Guillard and Angelo Murrone. On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes. *Computers & fluids*, 33(4):655–675, 2004.

[HKS19] Christiane Helzel, David Kerkmann, and Leonardo Scandurra. A new ADER method inspired by the active flux method. *Journal of Scientific Computing*, 80(3):1463–1497, 2019.

[Ise82] Arieh Iserles. Order stars and a saturation theorem for first-order hyperbolics. *IMA Journal of Numerical Analysis*, 2(1):49–61, 1982.

[Mil71] John JH Miller. On the location of zeros of certain classes of polynomials with applications to numerical analysis. *IMA Journal of Applied Mathematics*, 8(3):397–406, 1971.

[RLM15] Philip L Roe, Tyler Lung, and Jungyeoul Maeng. New approaches to limiting. In *22nd AIAA Computational Fluid Dynamics Conference*, page 2913, 2015.

[Roe17] Philip Roe. Multidimensional upwinding. *Handbook of Numerical Analysis*, 18:53–80, 2017.

[Roe21] Philip Roe. Designing cfd methods for bandwidth—a physical approach. *Computers & Fluids*, 214:104774, 2021.

[Sch17] Issai Schur. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. *Journal für die reine und angewandte Mathematik*, 147:205–232, 1917.

[Sch18] I Schur. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. part ii. *J. Reine Angew. Math*, 148:122–145, 1918.

[Sod78] Gary A Sod. A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *Journal of computational physics*, 27(1):1–31, 1978.

[vL77] Bram van Leer. Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. *Journal of computational physics*, 23(3):276–299, 1977.

[ZS10] Xiangxiong Zhang and Chi-Wang Shu. On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. *Journal of Computational Physics*, 229(23):8918–8934, 2010.

A Stability analysis

On equidistant Cartesian grids with $x_i = i\Delta x$, consider inserting a Fourier mode[5] $\hat{q}\exp(\mathbb{i}Kx)$ with $\hat{q} \in \mathbb{C}^m$, $K \in \mathbb{R}$ as data at time t^n. This means for point values

$$q^n_{i+\frac{1}{2}} = \hat{q}\exp\left(\mathbb{i}K\frac{\Delta x}{2}\right)\exp(\mathbb{i}Ki\Delta x) =: \hat{q}_p\exp(\mathbb{i}Ki\Delta x) \tag{120}$$

$$q^n_{i+\frac{3}{2}} = \hat{q}_p\exp(\mathbb{i}K\Delta x)\exp(\mathbb{i}Ki\Delta x) \tag{121}$$

$$q^n_{i-\frac{1}{2}} = \hat{q}_p\exp(-\mathbb{i}K\Delta x)\exp(\mathbb{i}Ki\Delta x) \tag{122}$$

Thus, every shift by one grid cell appears as an algebraic factor of $\exp(\mathbb{i}K\Delta x) =: t_x$:

$$q_{i+\frac{3}{2}} = q_{i+\frac{1}{2}}t_x \qquad q_{i-\frac{1}{2}} = q_{i+\frac{1}{2}}t_x^{-1} \qquad \text{etc.} \tag{123}$$

Similarly, for cell averages one finds

$$\bar{q}_i = \hat{q}\frac{1}{\Delta x}\int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}}\exp(\mathbb{i}Kx)\,\mathrm{d}x = \hat{q}\frac{1}{\Delta x}\exp(\mathbb{i}Ki\Delta x)\frac{\exp\left(\mathbb{i}K\frac{\Delta x}{2}\right)-\exp\left(-\mathbb{i}K\frac{\Delta x}{2}\right)}{\mathbb{i}K} \tag{124}$$

$$= \hat{q}\,\mathrm{sinc}\left(K\frac{\Delta x}{2}\right)\exp(\mathbb{i}Ki\Delta x) =: \hat{q}_a\exp(\mathbb{i}Ki\Delta x) \tag{125}$$

with $\mathrm{sinc}\,x := \frac{\sin x}{x}$. Therefore, again,

$$\bar{q}_{i+1} = \bar{q}_i t_x \qquad \bar{q}_{i-1} = \bar{q}_i t_x^{-1} \qquad \text{etc.} \tag{126}$$

The very same happens for moments.

Thus, any linear finite difference formula involving point values and moments can be written as a linear combination of just the point value and moments associated to one cell. For example, the update

$$\bar{q}_i^{n+1} = \bar{q}_i^n - \frac{c\Delta t}{\Delta x}(q_{i+\frac{1}{2}} - q_{i-\frac{1}{2}}) \tag{127}$$

[5] $\mathbb{i}$ is the maginary unit

of $\bar{q}_i$ for linear advection upon inserting a Fourier mode reads

$$\bar{q}_i^{n+1} = \bar{q}_i^n - \frac{c\Delta t}{\Delta x} q_{i+\frac{1}{2}}(1 - t_x^{-1}) \tag{128}$$

In general therefore, any of the three high-order versions of Active Flux presented in this paper can be written as

$$Q^{n+1} = \mathcal{A}\, Q^n \tag{129}$$

with $A \in \mathbb{C}^{(2+N_1+N_2)\times(2+N_1+N_2)}$ a complex-valued matrix that depends on t_x, and thus on K, and

$$Q^n := \left(q_{i+\frac{1}{2}}^n, \bar{q}_i^n, q_i^{(1)}, q_i^{(2)}, \ldots, q_i^{(N_1)}, q_{i,1}, q_{i,2}, \ldots, q_{i,N_2}\right) \qquad N_1, N_2 \in \mathbb{N}_0 \tag{130}$$

For the high-order Active Flux via larger stencils, $N_1 = N_2 = 0$; for high-order Active Flux via further point values, $N_1 = 0$ and for high-order via further moments, $N_2 = 0$.

L_2, or von Neumann stability, requires the eigenvalues of $\mathcal{A}$ to be contained in the closed unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$ for all values of $K \in [-\pi, \pi]$. With the help of the following theorem, this property can be checked more easily than by actually computing the eigenvalues:

Definition A.1. *Given a polynomial $f \in \mathbb{C}[z]$, $f(z) = \sum_{j=0}^n a_j z^j$, $a_j \in \mathbb{C}$ with $a_n \neq 0$ and $f(0) \neq 0$, define*

$$f^*(z) := \sum_{j=0}^{n} \bar{a}_{n-j} z^j \qquad\qquad f_1(z) := \frac{f^*(0)f(z) - f(0)f^*(z)}{z} \tag{131}$$

f', as usual, denotes the derivative of f with respect to x, while $\bar{a}$ denotes the complex-conjugate of $a \in \mathbb{C}$.

Theorem A.1 (Schur)**.** *A polynomial $f \in \mathbb{C}[z]$ of degree ≥ 1 has all its zeros contained in the unit disc, iff either*

- *$|f^*(0)| > |f(0)|$ and either f_1 has all its zeros contained in the unit disc or f_1 is a non-vanishing constant, or*
- *$f_1 \equiv 0$ and either f' has all its zeros contained in the unit disc or f' is a non-vanishing constant.*

Note that the recursion is sure to stop, as both f_1 and f' are of one degree lower than f. This theorem was proved in [Sch17, Sch18]; its algorithmic form was given in [Coh22], and then in [Mil71], where its usefulness for stability analysis was emphasized.

Note that still this property needs to be guaranteed for all K. However, the Theorem A.1 is easily amenable to an implementation which will, for any polynomial, state whether its zeros are contained in the closed unit disc up to machine error. We then propose to sample the domain of K, as well as any other parameters on which $\mathcal{A}$ might depend (in particular Δt). This allows to obtain stability results which we find very reliable in practice.

For the high-order extension of Active Flux via larger stencils, Figures 6–15 show the thus obtained stability regions in the plane spanned by the CFL number and the free parameter.