Journal of Scientific Computing (2025) 102:36
https://doi.org/10.1007/s10915-024-02749-1

A Semi-discrete Active Flux Method for the Euler Equations on Cartesian Grids

Rémi Abgrall[1] · **Wasilij Barsukow**[2] · **Christian Klingenberg**[3]

Received: 29 September 2023 / Revised: 18 November 2024 / Accepted: 20 November 2024

Abstract

Active Flux is an extension of the Finite Volume method and additionally incorporates point values located at cell boundaries. This gives rise to a globally continuous approximation of the solution. Originally, the Active Flux method emerged as a fully discrete method, and required an exact or approximate evolution operator for the point value update. For nonlinear problems such an operator is often difficult to obtain, in particular for multiple spatial dimensions. We demonstrate that a new semi-discrete Active Flux method (first described in Abgrall et al., ESAIM: Mathematical Modelling and Numerical Analysis 57:991-1027, 2023 for one space dimension) can be used to solve nonlinear hyperbolic systems in multiple dimensions without requiring evolution operators. We focus here on the compressible Euler equations of inviscid hydrodynamics and third-order accuracy. We introduce a multi-dimensional limiting strategy and demonstrate the performance of the new method on both Riemann problems and subsonic flows.

Keywords Compressible Euler equations · Active flux · High-order methods

Mathematics Subject Classification 65M08 · 65M20 · 65M70 · 76M12

CK and WB acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within *SPP 2410 Hyperbolic Balance Laws in Fluid Mechanics: Complexity, Scales, Randomness (CoScaRa)*, project number 525941602.

✉ Wasilij Barsukow
wasilij.barsukow@math.u-bordeaux.fr

Rémi Abgrall
remi.abgrall@math.uzh.ch

Christian Klingenberg
klingen@mathematik.uni-wuerzburg.de

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

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

[3] Institute for Mathematics, University of Wurzburg, Emil-Fischer-Strasse 40, 97074 Wurzburg, Germany

Published online: 03 January 2025

1 Introduction

The Active Flux method uses as its degrees of freedom both cell averages and point values at cell interfaces. The point values are shared between adjacent cells. In one spatial dimension, any choice of the reconstruction in one cell that interpolates those point values will give rise to a globally continuous reconstruction. In multiple spatial dimensions, it also seems natural to ensure global continuity, i.e. continuity all along the cell interface and not just in the point values. Traditionally, discontinuous reconstructions were favored for hyperbolic conservation laws; in [2], Active Flux is found to perform better than the Discontinuous Galerkin method.

While the averages require a conservative update, the update of the point values is essentially not restricted by more than the condition that the resulting method should be stable. To this end it needs to incorporate upwinding, and the earliest version of the Active Flux method ([3], for linear advection in 1-d) traced a characteristic back to the time level t^n where a reconstruction of the data was evaluated. In [4], approximate evolution operators were derived using a Cauchy-Kowalevskaya/ADER procedure. As the reconstruction is globally only C^0, Riemann problems in the derivatives were solved. This led to Active Flux methods for nonlinear systems of conservation laws in one spatial dimension. For scalar, nonlinear conservation laws, a fixpoint iteration can be used to systematically generate approximations to the speed of the characteristic of arbitrary order of accuracy [5, 6]. This approach was extended in [5] to hyperbolic systems of conservation laws in one spatial dimension, even if they do not allow for characteristic variables. Predictor-corrector estimates of the eigenvalues and the eigenvectors of the Jacobian led to third-order accurate approximate evolution operators. They were used, for example, in [7] to solve the one-dimensional shallow water equations in presence of dry areas.

For hyperbolic systems in multiple spatial dimensions, even if they are linear, characteristic curves in general no longer exist. Also, values in general are not transported, but the solution is a convolution of the initial data with a more or less complicated kernel. For the acoustic equations with the speed of sound c, for example, the solution in $\mathbf{x}$ at time t depends on the initial data in a disc with radius ct around $\mathbf{x}$. This disc is the interior of the intersection of the hypersurface of initial data with the cone of bicharacteristics which has its vertex at $(t, \mathbf{x})$. In [8], [9], a solution operator was given for the acoustic equations, which relied on smoothness of the initial data, and in [10] a solution in the sense of distributions was obtained which could be used to solve e.g. Riemann problems. These operators can be implemented efficiently and used to update the point values in an Active Flux method for linear acoustics [11], [12]. An approximate evolution operator for linear acoustics based on bicharacteristics is used in [12]. For the Euler equations, splitting and linearization is suggested in [2]. Generally, when studying the approximation error using Taylor series for nonlinear problems, evolution operators designed using linearization are found to require an additional fix in order to achieve third-order accuracy. For the one-dimensional case, in [5] a general algorithm to achieve third order of accuracy on non-linear problems while solving only linear ones was derived. Analogous high-order approximate evolution operators for general multi-dimensional nonlinear systems of conservation laws are currently unavailable, but for the Euler equations some suggestions can be found in [13].

All these Active Flux methods are fully-discrete. In [14, 15], a semi-discrete version of Active Flux was introduced. In order to obtain an equation for the point values, the spatial derivative in the PDE is discretized using finite difference formulae. This approach is immediately applicable to all kinds of nonlinear problems without the need to derive an

evolution operator, but at the price of a reduced CFL condition. In [16], a link is made between the semi-discrete approach and the reduced CFL condition for high-order methods with a compact stencil in space.

In [1, 17] the semi-discrete Active Flux has been applied to one-dimensional nonlinear problems, and extended to arbitrary order. The aim of the present work is, maintaining 3rd order of accuracy, to extend it to the multi-dimensional Euler equations. The paper is organized as follows: Sect. 2 describes the method and Sect. 3 presents a novel multi-dimensional limiting strategy. Numerical results are shown in Sect. 4.

2 The Semi-Discrete Active Flux Method

Here, we let ourselves guide by the semi-discrete approach of [1] and extend it to multi-dimensional Cartesian grids, while aiming at third-order accuracy. Consider a hyperbolic $m \times m$ system of conservation laws in d spatial dimensions[1]

$$\partial_t q + \nabla \cdot \mathbf{f}(q) = 0 \qquad q : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}^m \tag{1}$$

For simplicity, we restrict ourselves to two spatial dimensions ($d = 2$) and write $\mathbf{f} = (f^x, f^y)$, $\nabla_q f^x = J^x$, $\nabla_q f^y = J^y$ whenever convenient.

2.1 Update of the Averages

Integrating (1) over the Cartesian cell

$$C_{ij} := \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \times \left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right] \tag{2}$$

and denoting the cell average by

$$\bar{q}_{ij}(t) := \frac{1}{\Delta x \Delta y} \int_{C_{ij}} q(t, \mathbf{x}) \mathrm{d}\mathbf{x} \tag{3}$$

one finds

$$\frac{\mathrm{d}}{\mathrm{d}t} \bar{q}_{ij} + \frac{1}{\Delta x \Delta y} \int_{\partial C_{ij}} \mathbf{n} \cdot \mathbf{f}(q) = 0 \tag{4}$$

As there are degrees of freedom located at the boundary ∂C_{ij} of cell C_{ij}, we intend to use them as quadrature points for a sufficiently accurate quadrature of the integral appearing in (4). Inspired by previous approaches (e.g. [4, 11]) we use three Gauss-Lobatto points per edge, where the extreme points (corners) are shared (see Fig. 1). This is in contrast to [18] where a distribution of point values based on a Gauss-Legendre quadrature along the edge is suggested, or [14] where only second-order accuracy is obtained. Note also that we enforce global continuity: the point values on an edge are the same as seen from either of the adjacent cells and a value at a corner is involved in the update of four cells. This is in contrast to e.g. discontinuous Galerkin methods.

On Cartesian grids it is convenient to adopt the following notation for the 8 point values on the boundary of cell C_{ij}: $q_{i-\frac{1}{2},j+\frac{1}{2}}\, q_{i,j+\frac{1}{2}}\, q_{i+\frac{1}{2},j+\frac{1}{2}}\, q_{i-\frac{1}{2},j}\, q_{i+\frac{1}{2},j}\, q_{i-\frac{1}{2},j-\frac{1}{2}}\, q_{i,j-\frac{1}{2}}\, q_{i+\frac{1}{2},j-\frac{1}{2}}$

[1] Boldface letters denote "spatial" vectors, i.e. those whose natural dimension is that of space (d). Other collections of scalars (such as the conserved quantities q) are not typeset in boldface.

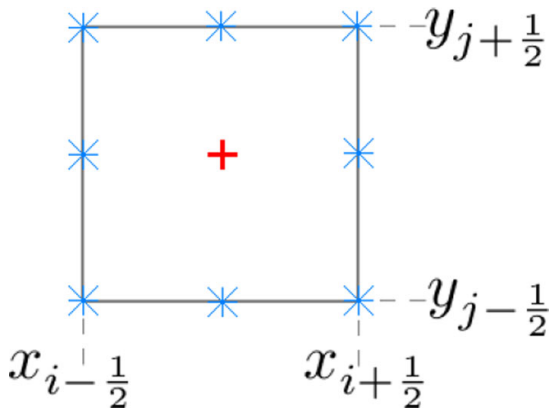

Fig. 1 Degrees of freedom employed in the third-order accurate Active Flux method on Cartesian grids. Stars denote point values, the cross denotes the cell average. Every cell has access to 8 point values and one average. Point values are shared with adjacent cells; per cell there is one average and $4 \cdot \frac{1}{4} + 4 \cdot \frac{1}{2} = 3$ point values

Then,

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

using Simpson's rule ($\omega_{-\frac{1}{2}} = \omega_{\frac{1}{2}} = \frac{1}{6}, \omega_0 = \frac{2}{3}$) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{q}_{ij}(t) + \frac{1}{\Delta x}\sum_{K=-\frac{1}{2},0,\frac{1}{2}} \omega_K\left(f^x\left(q_{i+\frac{1}{2},j+K}(t)\right) - f^x\left(q_{i-\frac{1}{2},j+K}(t)\right)\right) + \frac{1}{\Delta y}\sum_{K=-\frac{1}{2},0,\frac{1}{2}} \omega_K\left(f^y\left(q_{i+K,j+\frac{1}{2}}(t)\right) - f^y\left(q_{i+K,j-\frac{1}{2}}(t)\right)\right) = 0 \tag{6}$$

This method is conservative, with e.g. the x-flux through the cell interface $(i+\frac{1}{2}, j)$ being given by

$$\hat{f}^x_{i+\frac{1}{2},j} = \sum_{K=-\frac{1}{2},0,\frac{1}{2}} \omega_K f^x\left(q_{i+\frac{1}{2},j+K}\right) \tag{7}$$

$$= \frac{f^x(q_{i+\frac{1}{2},j-\frac{1}{2}}) + 4f^x\left(q_{i+\frac{1}{2},j}\right) + f^x\left(q_{i+\frac{1}{2},j+\frac{1}{2}}\right)}{6} \tag{8}$$

It is also at least 3rd order accurate, as it is exact for biparabolic functions.

2.2 Update of the Point Values

The update of the cell averages, as described above, now needs to be complemented by an update of the point values. In the one-dimensional case, it was proposed in [1] to replace the spatial derivatives appearing in (1) by finite differences. Here, the multi-dimensional case will be addressed. Note first that hyperbolicity of (1) implies that it is always possible to define the positive and negative parts of the Jacobians via their eigenvalues. With $J^x = R\mathrm{diag}(\lambda_1, \ldots, \lambda_m)R^{-1}$ one has

$$\left(J^x\right)^+ := R\mathrm{diag}\left(\lambda_1^+, \ldots, \lambda_m^+\right)R^{-1} \tag{9}$$

$$\left(J^x\right)^- := R\mathrm{diag}\left(\lambda_1^-, \ldots, \lambda_m^-\right) R^{-1} \tag{10}$$

where, for scalars $a \in \mathbb{R}$ the positive/negative parts are simply $a^+ = \max(0, a)$, $a^- = \min(0, a)$.

The finite difference formulae are obtained by differentiating a reconstruction. Define first the unique biparabolic polynomial

$$q_{ij,\mathrm{recon}} \in P^{2,2}, \quad q_{ij,\mathrm{recon}} \colon \left[-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right] \times \left[-\frac{\Delta y}{2}, \frac{\Delta y}{2}\right] \to \mathbb{R}^m \tag{11}$$

$$P^{2,2} = \mathrm{span}\left(1, x, x^2, y, xy, x^2y, y^2, xy^2, x^2y^2\right) \tag{12}$$

that interpolates the degrees of freedom accessible to cell ij:

$$q_{ij,\mathrm{recon}}\left(-\frac{\Delta x}{2}, \frac{\Delta y}{2}\right) = q_{i-\frac{1}{2},j+\frac{1}{2}} \qquad q_{ij,\mathrm{recon}}\left(\frac{\Delta x}{2}, \frac{\Delta y}{2}\right) = q_{i+\frac{1}{2},j+\frac{1}{2}}$$

$$q_{ij,\mathrm{recon}}\left(-\frac{\Delta x}{2}, -\frac{\Delta y}{2}\right) = q_{i-\frac{1}{2},j-\frac{1}{2}} \qquad q_{ij,\mathrm{recon}}\left(\frac{\Delta x}{2}, -\frac{\Delta y}{2}\right) = q_{i+\frac{1}{2},j-\frac{1}{2}}$$

$$q_{ij,\mathrm{recon}}\left(0, \frac{\Delta y}{2}\right) = q_{i,j+\frac{1}{2}} \qquad q_{ij,\mathrm{recon}}\left(-\frac{\Delta x}{2}, 0\right) = q_{i-\frac{1}{2},j}$$

$$q_{ij,\mathrm{recon}}\left(\frac{\Delta x}{2}, 0\right) = q_{i+\frac{1}{2},j} \qquad q_{ij,\mathrm{recon}}\left(0, -\frac{\Delta y}{2}\right) = q_{i,j-\frac{1}{2}}$$

and

$$\frac{1}{\Delta x \Delta y} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \int_{-\frac{\Delta y}{2}}^{\frac{\Delta y}{2}} q_{ij,\mathrm{recon}}(x, y)\, \mathrm{d}y \mathrm{d}x = \bar{q}_{ij} \tag{13}$$

This reconstruction has already been used in [4, 11] and is given there explicitly. Then we define the finite differences in the corner as

$$\left(D^x\right)^+_{i+\frac{1}{2},j+\frac{1}{2}} q := \partial_x q_{ij,\mathrm{recon}}\left(x, \frac{\Delta y}{2}\right)\Big|_{x=\frac{\Delta x}{2}} \tag{14}$$

$$\left(D^x\right)^-_{i+\frac{1}{2},j+\frac{1}{2}} q := \partial_x q_{i+1,j,\mathrm{recon}}\left(x, \frac{\Delta y}{2}\right)\Big|_{x=-\frac{\Delta x}{2}} \tag{15}$$

$$\left(D^y\right)^+_{i+\frac{1}{2},j+\frac{1}{2}} q := \partial_y q_{ij,\mathrm{recon}}\left(\frac{\Delta x}{2}, y\right)\Big|_{y=\frac{\Delta y}{2}} \tag{16}$$

$$\left(D^y\right)^-_{i+\frac{1}{2},j+\frac{1}{2}} q := \partial_y q_{i,j+1,\mathrm{recon}}\left(\frac{\Delta x}{2}, y\right)\Big|_{y=-\frac{\Delta y}{2}} \tag{17}$$

Observe that due to continuity,

$$\left(D^x\right)^+_{i+\frac{1}{2},j+\frac{1}{2}} q = \partial_x q_{i,j+1,\mathrm{recon}}\left(x, -\frac{\Delta y}{2}\right)\Big|_{x=\frac{\Delta x}{2}} \tag{18}$$

such that this would be an equivalent definition that gives the same result (and similarly for the other finite differences). Analogously, we define the finite differences on the edges

$$\left(D^x\right)^+_{i+\frac{1}{2},j} q := \partial_x q_{ij,\mathrm{recon}}\,(x, 0)|_{x=\frac{\Delta x}{2}} \tag{19}$$

$$\left(D^x\right)^-_{i+\frac{1}{2},j} q := \partial_x q_{i+1,j,\mathrm{recon}}\,(x, 0)|_{x=-\frac{\Delta x}{2}} \tag{20}$$

$$\left(D^y\right)_{i+\frac{1}{2},j} q := \partial_x q_{ij,\text{recon}} \left(\frac{\Delta x}{2}, y\right)\Big|_{y=0} \tag{21}$$

Observe that due to continuity, there is no distinction between $(D^y)^+_{i+\frac{1}{2},j}$ and $(D^y)^-_{i+\frac{1}{2},j}$. Here, again, the symmetric definition

$$\left(D^y\right)_{i+\frac{1}{2},j} q := \partial_y q_{i+1,j,\text{recon}} \left(-\frac{\Delta x}{2}, y\right)\Big|_{y=0} \tag{22}$$

yields the same result. The derivatives at $(i, j+\frac{1}{2})$ are obtained analogously. For reference we now state their explicit forms:

$$\begin{aligned}
\left(D^x\right)^+_{i+\frac{1}{2},j} q &= \frac{1}{4\Delta x}\Big(4\left(-9\bar{q}_{ij} + 2\left(q_{i-\frac{1}{2},j} + 2q_{i+\frac{1}{2},j}\right)\right) + 4\left(q_{i,j-\frac{1}{2}} + q_{i,j+\frac{1}{2}}\right) \\
&\quad + q_{i-\frac{1}{2},j-\frac{1}{2}} + q_{i+\frac{1}{2},j-\frac{1}{2}} + q_{i-\frac{1}{2},j+\frac{1}{2}} + q_{i+\frac{1}{2},j+\frac{1}{2}}\Big) \\
\left(D^x\right)^-_{i+\frac{1}{2},j} q &= -\frac{1}{4\Delta x}\Big(-36\bar{q}_{i+1,j} + 8\left(2q_{i+\frac{1}{2},j} + q_{i+\frac{3}{2},j}\right) + q_{i+\frac{1}{2},j-\frac{1}{2}} \\
&\quad + 4\left(q_{i+1,j-\frac{1}{2}} + q_{i+1,j+\frac{1}{2}}\right) + q_{i+\frac{3}{2},j-\frac{1}{2}} + q_{i+\frac{1}{2},j+\frac{1}{2}} + q_{i+\frac{3}{2},j+\frac{1}{2}}\Big) \\
\left(D^y\right)_{i+\frac{1}{2},j} q &= \frac{q_{i+\frac{1}{2},j+\frac{1}{2}} - q_{i+\frac{1}{2},j-\frac{1}{2}}}{\Delta y} \\
\left(D^y\right)^+_{i,j+\frac{1}{2}} q &= \frac{1}{4\Delta y}\Big(4\left(q_{i-\frac{1}{2},j} - 9\bar{q}_{ij} + q_{i+\frac{1}{2},j}\right) + q_{i-\frac{1}{2},j-\frac{1}{2}} + q_{i-\frac{1}{2},j+\frac{1}{2}} \\
&\quad + q_{i+\frac{1}{2},j-\frac{1}{2}} + q_{i+\frac{1}{2},j+\frac{1}{2}} + 8\left(q_{i,j-\frac{1}{2}} + 2q_{i,j+\frac{1}{2}}\right)\Big) \\
\left(D^y\right)^-_{i,j+\frac{1}{2}} q &= -\frac{1}{4\Delta y}\Big(4\left(q_{i-\frac{1}{2},j+1} - 9\bar{q}_{i,j+1} + q_{i+\frac{1}{2},j+1}\right) + q_{i-\frac{1}{2},j+\frac{1}{2}} \\
&\quad + q_{i+\frac{1}{2},j+\frac{1}{2}} + q_{i-\frac{1}{2},j+\frac{3}{2}} + q_{i+\frac{1}{2},j+\frac{3}{2}} + 8\left(2q_{i,j+\frac{1}{2}} + q_{i,j+\frac{3}{2}}\right)\Big) \\
\left(D^x\right)_{i,j+\frac{1}{2}} q &= \frac{q_{i+\frac{1}{2},j+\frac{1}{2}} - q_{i-\frac{1}{2},j+\frac{1}{2}}}{\Delta x} \\
\left(D^x\right)^+_{i+\frac{1}{2},j+\frac{1}{2}} q &= \frac{q_{i-\frac{1}{2},j+\frac{1}{2}} - 4q_{i,j+\frac{1}{2}} + 3q_{i+\frac{1}{2},j+\frac{1}{2}}}{\Delta x} \\
\left(D^x\right)^-_{i+\frac{1}{2},j+\frac{1}{2}} q &= \frac{4q_{i+1,j+\frac{1}{2}} - 3q_{i+\frac{1}{2},j+\frac{1}{2}} - q_{i+\frac{3}{2},j+\frac{1}{2}}}{\Delta x} \\
\left(D^y\right)^+_{i+\frac{1}{2},j+\frac{1}{2}} q &= \frac{q_{i+\frac{1}{2},j-\frac{1}{2}} - 4q_{i+\frac{1}{2},j} + 3q_{i+\frac{1}{2},j+\frac{1}{2}}}{\Delta y} \\
\left(D^y\right)^-_{i+\frac{1}{2},j+\frac{1}{2}} q &= \frac{4q_{i+\frac{1}{2},j+1} - 3q_{i+\frac{1}{2},j+\frac{1}{2}} - q_{i+\frac{1}{2},j+\frac{3}{2}}}{\Delta y}
\end{aligned}$$

However, in some situations one might be willing to employ a different reconstruction, as is, for instance, the case in Sect. 3 concerned with limiting. At this point one has to resort to the more general formulae (14)–(21).

Finally, the upwinding is defined as

$$\left(J^x D^x_{i+K,j+L}\right)^{\text{upw}} q := \left(J^x\right)^+ \left(D^x\right)^+_{i+K,j+L} q + \left(J^x\right)^- \left(D^x\right)^-_{i+K,j+L} q \tag{23}$$

with $K, L \in \{-\frac{1}{2}, 0, \frac{1}{2}\}$ and an analogous definition for J^y. We content ourselves here with simple upwinding based on the x- and y-Jacobians separately. The multi-dimensionality advertised in works of Roe and collaborators is not entirely lost, though, because the degree of freedom at the node couples the waves from different directions in a truly multi-dimensional manner (different from DG methods, say).

We propose to update the point values as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t} q_{i+\frac{1}{2},j} + \left(J^x D^x_{i+\frac{1}{2},j} \right)^{\mathrm{upw}} q + J^y D^y_{i+\frac{1}{2},j} q = 0 \tag{24}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} q_{i,j+\frac{1}{2}} + J^x D^x_{i,j+\frac{1}{2}} q + \left(J^y D^y_{i,j+\frac{1}{2}} \right)^{\mathrm{upw}} q = 0 \tag{25}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} q_{i+\frac{1}{2},j+\frac{1}{2}} + \left(J^x D^x_{i+\frac{1}{2},j+\frac{1}{2}} \right)^{\mathrm{upw}} q + \left(J^y D^y_{i+\frac{1}{2},j+\frac{1}{2}} \right)^{\mathrm{upw}} q = 0 \tag{26}$$

As the finite differences are exact for biparabolic function, one expects 3rd order of accuracy.

The complete method consists of the ODEs (6) (average update), (24)–(25) (point values at edge midpoints) and (26) (point values at nodes). We propose to integrate these with an SSP-RK3 method. In [17], it was shown for linear advection in one spatial dimension that this approach leads to a stable scheme with a maximum CFL number of 0.41, a value lower than what can be achieved if an evolution operator is available [6], but comparable e.g. to CFL numbers of DG methods. In two space dimensions we expect stability at least for half the CFL number, i.e. 0.2, or possibly higher.

3 Limiting

Existing approaches to limiting in the context of standard Finite Volume methods modify the values of the reconstruction at a cell interface. They cannot be used for Active Flux due to its global continuity and the fact that point values at cell interfaces are prescribed and cannot be modified arbitrarily. Limiting employed in [4] therefore gives up on continuity. Approaches to limiting that maintain continuity so far have only been treating the situation in which a parabolic reconstruction of monotonic discrete data (point values and average) is not monotonic, i.e. has an artificial extremum. In [19], a piecewise linear/parabolic reconstruction is used in this case, and in [5] the same situation is handled by replacing the parabola by a power law. One can show that then the reconstruction is always monotonic whenever the discrete data are. Such modified reconstructions are effective in drastically reducing spurious oscillations, but they do not guarantee to remove them entirely. This is because the update of the averages is not limited and can itself create artificial extrema in the discrete data. However, in absence of better approaches, e.g. the power-law reconstruction is a viable limiting strategy. In particular, it is not computationally intensive.

In multiple spatial dimensions, a similar strategy is presented here for the first time. The multi-dimensional case is, however, much more complex because every cell has access to 8 point values. Even monotonicity is a vague concept, because data can be monotonic along one direction and non-monotonic along the other. We focus only on the degrees of freedom accessible to a cell. As long as the cell average is between the smallest and the largest point values we propose to reconstruct in such a way that the value of the reconstruction at any point inside the cell is also between the smallest and the largest point value. This is always possible, as will be shown. We additionally impose a monotonicity constraint along any edge,

i.e. we reconstruct in a monotonic fashion if the data are monotonic. Finally, we also impose global continuity.

Consider point values at edge centers $q_{\mathrm{N}}, q_{\mathrm{S}}, q_{\mathrm{W}}, q_{\mathrm{E}}$ and at vertices $q_{\mathrm{NE}}, q_{\mathrm{SE}}, q_{\mathrm{NW}}, q_{\mathrm{SW}}$ of a (reference) Cartesian cell $c = [-\frac{\Delta x}{2}, \frac{\Delta x}{2}] \times [-\frac{\Delta y}{2}, \frac{\Delta y}{2}]$ and a cell average $\bar{q}$ to be given. We will refer to the four edges as N-edge, S-edge, W-edge and E-edge, respectively. The reconstruction will be denoted by $q_{\mathrm{recon}} : c \to \mathbb{R}$ for simplicity. There exist two ways how the reconstruction can create new extrema, which can occur independently of each other:

1. It can happen that the parabolic reconstruction along an edge (as part of a biparabolic reconstruction in the cell) overshoots/undershoots the three point values along the edge in question. For the example of an N-edge, this happens if either
   - the point values $q_{\mathrm{NW}}, q_{\mathrm{N}}, q_{\mathrm{NE}}$ are not monotonic and $q_{\mathrm{NW}} \neq q_{\mathrm{NE}}$, or if
   - they are monotonic (i.e. either $q_{\mathrm{NW}} < q_{\mathrm{N}} < q_{\mathrm{NE}}$ or $q_{\mathrm{NW}} > q_{\mathrm{N}} > q_{\mathrm{NE}}$), but

$$\left| q_{\mathrm{N}} - \frac{q_{\mathrm{NE}} + q_{\mathrm{NW}}}{2} \right| > \frac{|q_{\mathrm{NE}} - q_{\mathrm{NW}}|}{4} \tag{27}$$

     such that the parabolic reconstruction has an extremum not present in the discrete data.

   In this case the reconstruction along the edge will be chosen continuous piecewise linear ("hat"). We will say that the **reconstruction along the edge is limited**, or just that the "edge is limited". To ensure continuity, the reconstruction in any cell with a limited edge can then no longer be biparabolic, but needs to be modified as detailed below and in Section A.1.
2. Define

$$m := \min\left(q_{\mathrm{N}}, q_{\mathrm{S}}, q_{\mathrm{W}}, q_{\mathrm{E}}, q_{\mathrm{NE}}, q_{\mathrm{SE}}, q_{\mathrm{NW}}, q_{\mathrm{SW}}\right) \tag{28}$$

$$M := \max\left(q_{\mathrm{N}}, q_{\mathrm{S}}, q_{\mathrm{W}}, q_{\mathrm{E}}, q_{\mathrm{NE}}, q_{\mathrm{SE}}, q_{\mathrm{NW}}, q_{\mathrm{SW}}\right) \tag{29}$$

   It can happen that despite

$$m < \bar{q} < M \tag{30}$$

   the reconstruction q_{recon} inside the cell c has an extremum with no counterpart in the discrete data, i.e.

$$\exists \mathbf{x} \in c \text{ such that either } q_{\mathrm{recon}}(\mathbf{x}) < m \text{ or } q_{\mathrm{recon}}(\mathbf{x}) > M \tag{31}$$

   This situation will be improved by introducing a piecewise defined reconstruction with a central region where the function is constant ("plateau"), and connecting the plateau to the (parabolic or hat) reconstructions along the edges in a continuous fashion. More details are given below and in Sect. A.2; Figs. 2 and 3 show examples. This new reconstruction fulfills

$$m < q_{\mathrm{recon}}(\mathbf{x}) < M \qquad \forall \mathbf{x} \in c \tag{32}$$

   We will say that the **reconstruction inside the cell is limited**, or just that the "cell is limited".
   This situation appears already in 1-d, in which case it has been suggested in [5] to replace the parabolic reconstruction in the cell by a power law. A multi-dimensional analogue of the power law seems unfeasible, though, and we resort here to a piecewise defined, but easier function.

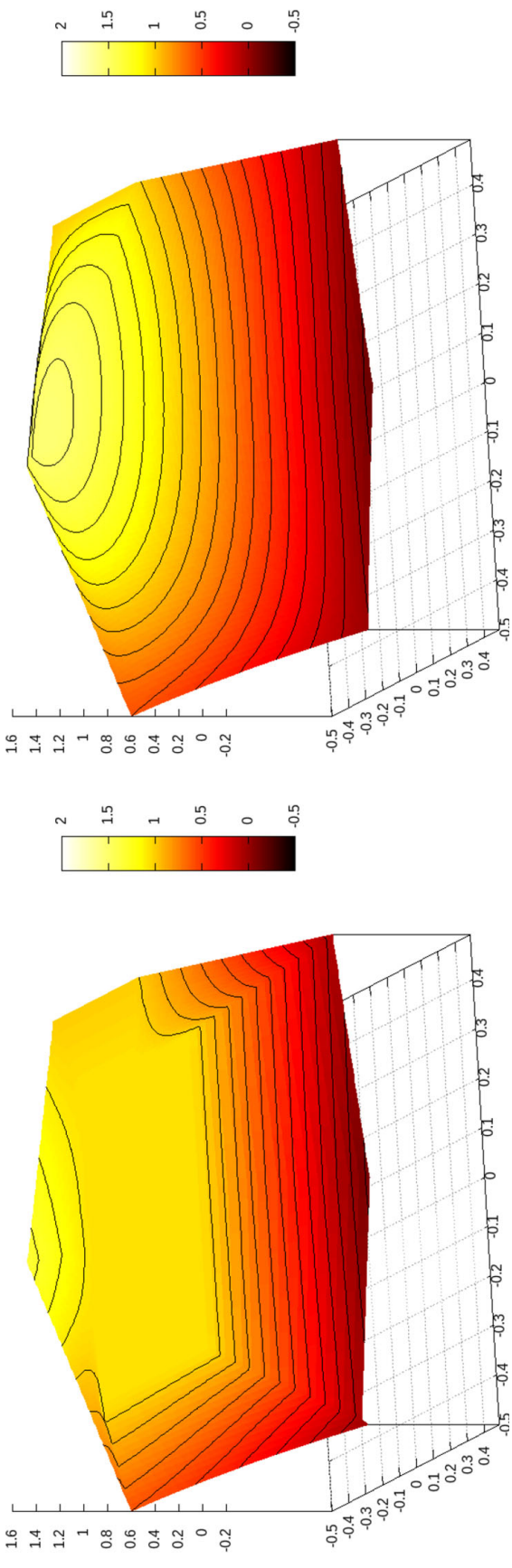

Fig. 2 *Left*: An example of a plateau reconstruction. Here, $q_{\mathrm{NW}} = 1$, $q_{\mathrm{W}} = 1.35$, $q_{\mathrm{SW}} = 0.6$, $q_{\mathrm{S}} = 0.4$, $q_{\mathrm{SE}} = 0$, $q_{\mathrm{E}} = -0.2$, $q_{\mathrm{NE}} = 0.0$, $q_{\mathrm{N}} = 1$, $\bar{q} = 0.9$ (the S-edge is on the left and $\Delta x = \Delta y = 1$). All edges but the S-edge are reconstructed as hats, the S-edge is reconstructed parabolically. *Right*: A piecewise-biparabolic reconstruction of the same data; one clearly observes an overshoot. The isolines have a spacing of 0.1

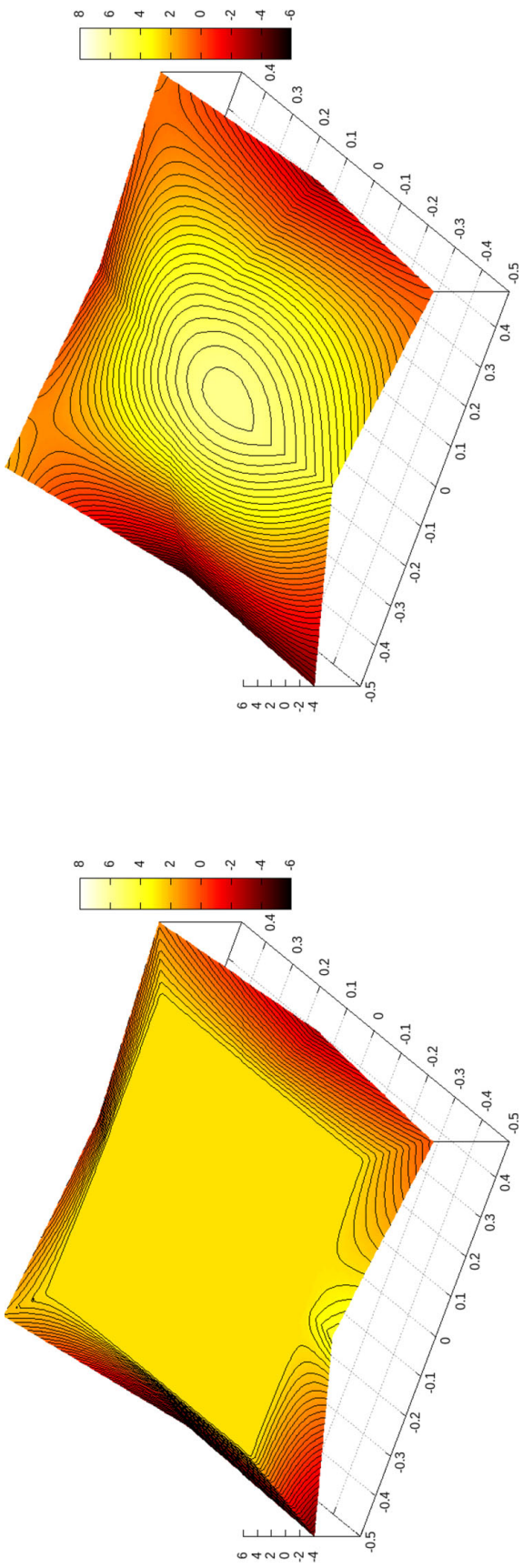

Fig. 3 *Left*: An example of a plateau reconstruction. $q_{\mathrm{NE}} = 1$, $q_{\mathrm{NW}} = 2$, $q_{\mathrm{SW}} = -4$, $q_{\mathrm{SE}} = 0$, $q_{\mathrm{N}} = -1$, $q_{\mathrm{S}} = 4$, $q_{\mathrm{W}} = -5$, $q_{\mathrm{E}} = -3$, $\bar{q} = 2$ (the W-edge is on the left and $\Delta x = \Delta y = 1$). All edges are reconstructed as hats. *Right*: A piecewise-biparabolic reconstruction of the same data; one clearly observes an overshoot. The isolines have a spacing of 0.25

The two situations are independent: any number of edges along the boundary of a cell might require limiting, and this will not generally imply anything about whether the cell itself is to be limited. The possible presence of hat functions along the boundary requires the reconstruction inside the cell to flexibly adapt to the different combinations of edge-reconstructions in order to be continuous. For instance, the plateau reconstruction needs to connect the plateau continuously to either a parabola, or a hat function (see Sect. A.2). Also, if there exists at least one edge that is reconstructed as a hat function, then one cannot use a biparabolic reconstruction inside the cell any longer, even if the cell is not limited. If at least one edge is limited, but the cell is not, a piecewise-biparabolic reconstruction will be used, detailed in Section A.1.

As we are aiming at a globally continuous reconstruction, that is computed locally from merely the cell average and the point values of the cell, the reconstruction along an edge can only depend on the three values associated to this edge, and cannot depend on other values in the cell. Indeed, if edge-reconstruction of one of edges of c were to depend on, say, the average in the cell c, then the reconstruction in the neighbouring cell c' would also need to know about the average in c.

Due to the particular choice of degrees of freedom for Active Flux the reconstruction has to fulfill two types of conditions: It is supposed to interpolate the point values at cell interfaces and its average is supposed to be equal to the given one. The latter condition – merely to simplify the calculations – will be replaced by a (yet unknown) point value q_C at cell center which is kept as a variable in the formulae. Once the type of reconstruction in all regions of the cell has been determined, their integrals over the respective domains of definition can easily be found as functions of q_C, and q_C is then determined by imposing the average of the reconstruction over the entire cell. This is a linear equation in q_C due to linearity of the interpolation problem which makes q_C enter linearly everywhere. The explicit formulae below therefore also depend on q_C, but the reconstruction in a cell in the end only depends on the point values along its boundary and on its average. This detour does not change the result but simplifies the algorithm.

The overall structure of the reconstruction algorithm is:

1. Decide for every edge of the cell whether it is reconstructed parabolically, or as a hat function.
2. Assume as hypothesis that the cell does not require limiting (i.e. that it is reconstructed in a piecewise biparabolic fashion) and compute the value of q_C that ensures that the average of the reconstruction agrees with the given cell average.
3. Check (30) and if true, decide whether the piecewise-biparabolic reconstruction obtained in 3 has an artificial extremum in the sense of what has been described above[2]
4. If this is the case, the cell needs to be limited with a plateau reconstruction. Compute the parameters η, q_p (see below) of the plateau reconstruction that ensure maximum principle preservation and the correct value of the average of the reconstruction.

A pedagogical derivation of the reconstruction algorithm is given in Sect. A. Here, we only state all the relevant results in a concise way.

Theorem 1 *The following reconstruction* $q_{recon}\colon \left[-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right] \times \left[-\frac{\Delta y}{2}, \frac{\Delta y}{2}\right] \to \mathbb{R}$ *is continuous, interpolates all the point values along the boundary of the cell, its average agrees with the given cell average and the reconstruction has the following properties:*

(i) If Condition (30), i.e. $m < \bar{q} < M$ *is fulfilled, then* $m \le q_{recon}(\boldsymbol{x}) \le M$ *for all* $\boldsymbol{x}$ *inside the cell.*

[2] This happens numerically by testing a given number of locations.

(ii) If $q_{NW} < q_N < q_{NE}$, then $q_{NW} \leq q_{recon}(\boldsymbol{x}) \leq q_{NE}$ for all $\boldsymbol{x}$ along the N-edge, and similarly for all the other edges.

The definition of the reconstruction is as follows: If $m < \bar{q} < M$ is not fulfilled, or if additionally $m < q_{recon}^{pw.\ biparab.}(x, y) < M$ for all $(x, y) \in c$, then

$$q_{recon}(x, y) := q_{recon}^{pw.\ biparab.}(x, y) \tag{33}$$

otherwise

$$q_{recon}(x, y) := q_{recon}^{plateau}(x, y), \tag{34}$$

the two types of reconstruction being defined as follows:

1. *The piecewise-barabolic reconstruction:*

$$\begin{aligned} q_{recon}^{pw.\ biparab.}(x, y) := & \qquad (35)\\ & q_{recon}^{W}\left(\frac{q_{SW}-\bar{Q}}{2}, q_W-\bar{Q}, \frac{q_{NW}-\bar{Q}}{2}, x, y, S, N, W, \frac{\bar{q}-\bar{Q}}{4}\right)\\ & + q_{recon}^{S}\left(\frac{q_{SE}-\bar{Q}}{2}, q_S-\bar{Q}, \frac{q_{SW}-\bar{Q}}{2}, x, y, E, W, S, \frac{\bar{q}-\bar{Q}}{4}\right)\\ & + q_{recon}^{N}\left(\frac{q_{NW}-\bar{Q}}{2}, q_N-\bar{Q}, \frac{q_{NE}-\bar{Q}}{2}, x, y, W, E, N, \frac{\bar{q}-\bar{Q}}{4}\right)\\ & + q_{recon}^{E}\left(\frac{q_{NE}-\bar{Q}}{2}, q_E-\bar{Q}, \frac{q_{SE}-\bar{Q}}{2}, x, y, N, S, E, \frac{\bar{q}-\bar{Q}}{4}\right)\\ & + \bar{Q} \qquad (36) \end{aligned}$$

with $\bar{Q} := \frac{q_{SW}+q_W+q_{NW}+q_N+q_{NE}+q_E+q_{SE}+q_S}{8}$ (other choices are possible) and

$$q_{recon}^{S}(q_{SE}, q_S, q_{SW}, x, y, E, W, S, \bar{q}) = q_{recon}^{W}(q_{SE}, q_S, q_{SW}, y, -x, E, W, S, \bar{q}) \tag{37}$$

$$q_{recon}^{N}(q_{NW}, q_N, q_{NE}, x, y, W, E, N, \bar{q}) = q_{recon}^{W}(q_{NW}, q_N, q_{NE}, -y, x, W, E, N, \bar{q}) \tag{38}$$

$$q_{recon}^{E}(q_{NE}, q_E, q_{SE}, x, y, N, S, E, \bar{q}) = q_{recon}^{W}(q_{NE}, q_E, q_{SE}, -x, -y, N, S, E, \bar{q}) \tag{39}$$

and

$$q^{W}_{recon}\left(q_{SW}, q_W, q_{NW}, x, y, S, N, W, \bar{q}\right) = \begin{cases} (88) & \text{N,S,W parabolic} \\ (89)\text{–}(90) & \text{W parabolic, N,S hat} \\ (92)\text{–}(93) & \text{W,S parabolic, N hat} \\ (95)\text{–}(96) & \text{W,N parabolic, S hat} \\ (102) \text{ and } (100) & \text{W hat, N,S parabolic} \\ (102) \text{ and } (103)\text{–}(104) & \text{W,N hat, S parabolic} \\ (100) \text{ and } (106)\text{–}(107) & \text{W,S hat, N parabolic} \\ (103)\text{–}(104) \text{ and } (106)\text{–}(107) & \text{W,S,N hat} \end{cases} \tag{40}$$

Here, N/S/E/W denote the edges of the cell. q_C fulfills

$$\begin{cases} q_C = \frac{1}{16}\left(36\bar{q} - q_{NW} - q_{SW} - 4q_W\right) & \text{N,S,W parabolic} \\ q_C = \frac{1}{32}\left(72\bar{q} - 3q_{NW} - 3q_{SW} - 8q_W\right) & \text{W parabolic, N,S hat} \\ q_C = \frac{1}{32}\left(72\bar{q} - 3q_{NW} - 2q_{SW} - 8q_W\right) & \text{W,S parabolic, N hat} \\ q_C = \frac{1}{32}\left(72\bar{q} - 2q_{NW} - 3q_{SW} - 8q_W\right) & \text{W,N parabolic, S hat} \\ \bar{q} = \frac{2q_C}{9} + \frac{q_{SW}+q_W}{24} + \frac{2q_C}{9} + \frac{q_{NW}+q_W}{24} & \text{W hat, N,S parabolic} \\ \bar{q} = \frac{2q_C}{9} + \frac{q_{SW}+q_W}{24} + \frac{2q_C}{9} + \frac{1}{576}\left(35q_{NW} + q_{SW} + 22q_W\right) & \text{W,N hat, S parabolic} \\ \bar{q} = \frac{2q_C}{9} + \frac{q_{NW}+q_W}{24} + \frac{2q_C}{9} + \frac{1}{576}\left(q_{NW} + 35q_{SW} + 22q_W\right) & \text{W,S hat, N parabolic} \\ \bar{q} = \frac{2q_C}{9} + \frac{1}{576}\left(35q_{NW} + q_{SW} + 22q_W\right) + \frac{2q_C}{9} + \frac{1}{576}\left(q_{NW} + 35q_{SW} + 22q_W\right) & \text{W,S,N hat} \end{cases}$$

2. *The plateau reconstruction:*

$$q^{plateau}_{recon}(x, y) := \begin{cases} q_p \text{ if } (x, y) \in \left[\Delta x\left(\eta - \frac{1}{2}\right), \Delta x\left(\frac{1}{2} - \eta\right)\right] \times \left[\Delta y\left(\eta - \frac{1}{2}\right), \Delta y\left(\frac{1}{2} - \eta\right)\right] \\ q^{trapeze,W}_{recon}\left(q_{SW}, q_W, q_{NW}, x, y, W, \eta, q_p\right) \text{ if } (x, y) \in W\text{-trapeze} \\ q^{trapeze,S}_{recon}\left(q_{SE}, q_S, q_{SW}, x, y, S, \eta, q_p\right) \text{ if } (x, y) \in S\text{-trapeze} \\ q^{trapeze,N}_{recon}\left(q_{NW}, q_N, q_{NE}, x, y, N, \eta, q_p\right) \text{ if } (x, y) \in N\text{-trapeze} \\ q^{trapeze,E}_{recon}\left(q_{NE}, q_E, q_{SE}, x, y, E, \eta, q_p\right) \text{ if } (x, y) \in E\text{-trapeze} \end{cases}$$

with

$$q^{trapeze,W}_{recon}\left(q_{SW}, q_W, q_{NW}, x, y, W, \eta, q_p\right) = \begin{cases} (117) & (x, y) \in W \text{ parabolic} \\ (126)\text{–}(127) & (x, y) \in W \text{ hat} \end{cases} \tag{41}$$

defined only in

$$W\text{-trapeze} = \left\{(x, y) \text{ s.t. } x \in \left[-\frac{\Delta x}{2}, -\Delta x\left(\frac{1}{2} - \eta\right)\right] \text{ and } y \in \left[\frac{x}{\Delta y}\Delta x, -\frac{x}{\Delta y}\Delta x\right]\right\} \tag{42}$$

The reconstructions of the other trapezes are

$$q^{trapeze,S}_{recon}\left(q_{SE}, q_S, q_{SW}, x, y, S, \eta, q_p\right) = q^{trapeze,W}_{recon}\left(q_{SE}, q_S, q_{SW}, y, -x, S, \eta, q_p\right) \tag{43}$$

$$q^{trapeze,N}_{recon}\left(q_{NW}, q_N, q_{NE}, x, y, N, \eta, q_p\right) = q^{trapeze,W}_{recon}\left(q_{NW}, q_N, q_{NE}, -y, x, N, \eta, q_p\right) \tag{44}$$

$$q^{trapeze,E}_{recon}\left(q_{NE}, q_E, q_{SE}, x, y, E, \eta, q_p\right) = q^{trapeze,W}_{recon}\left(q_{NE}, q_E, q_{SE}, -x, -y, E, \eta, q_p\right) \tag{45}$$

The parameters q_p and η are found according to the procedure descrobed in Sect. A.2.4.

Proof Continuity is a consequence of Theorems 5 and 6 and of the fact that the reconstruction along any edge only involves the points on this edge. The pointwise and average interpolation property follows from Theorems 4 and 7. Preservation of the maximum principle along the edges is clear from (27) and the idea of reconstructing a hat function along the edge; preservation of the maximum principle for the cell follows from Theorem 7. □

Theorem 2 *The usage of the reconstruction from Theorem 1 in every cell leads to a globally continuous reconstruction.*

Proof It follows by construction (see A.1.1, A.2.1) that the reconstruction in a cell c continuously turns into the reconstruction along the edge as $c \ni (x, y) \to s \in \partial c$. The reconstructions along the edges only depend on the three point values located on the edge, and thus the limit as (x, y) approaches the same edge from the other cell is the same. □

4 Numerical Results

Here, the Euler equations with $q = (\rho, \rho u, \rho v, e)$,

$$f^x = \left(\rho u, \rho u^2 + p, \rho uv, u(e + p)\right) \tag{46}$$

$$f^y = \left(\rho v, \rho uv, \rho v^2 + p, v(e + p)\right) \tag{47}$$

$$e = \frac{p}{\gamma - 1} + \frac{1}{2}\rho\left(u^2 + v^2\right) \tag{48}$$

and $\gamma = 1.4$ are solved using the Active Flux method described above. Initial data are denoted by ρ_0, u_0, v_0, p_0.

4.1 Convergence Studies

For a convergence analysis, the following initial data (similar to those used in [4, 5]) are solved until $t = 0.05$ on grids of different resolution:

$$u_0(x, y) = v_0(x, y) = 0 \qquad r := \sqrt{\left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2} \tag{49}$$

$$\rho_0(x, y) = p_0(x, y) = 1 + \frac{1}{2}\exp\left(-80r^2\right) \tag{50}$$

Figure 4 shows the setup and the error, computed with respect to a reference solution obtained on a grid of 1024×1024 covering $[0, 1]^2$.

To facilitate comparison with other methods, we have also performed a convergence analysis using the setup of an isentropic ($p = \rho^\gamma$), traveling vortex as in [20], Example 3.3:

$$\begin{pmatrix} u_0(x, y) \\ v_0(x, y) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{\Gamma}{2\pi}\exp\left(\frac{1 - r^2}{2}\right)\begin{pmatrix} -y \\ x \end{pmatrix} \tag{51}$$

$$T_0(x, y) := \frac{p_0(x, y)}{\rho_0(x, y)} = 1 - \frac{(\gamma - 1)\Gamma^2}{8\gamma\pi^2}\exp\left(1 - r^2\right) \tag{52}$$

$$r := \sqrt{(x - 10)^2 + (y - 10)^2} \tag{53}$$

The exact solution is pure advection with speed $(1, 1)$. We use $\Gamma = 5$ and compute the errors at time $t = 2$. This vortex is not compactly supported, but its rotational velocity decays

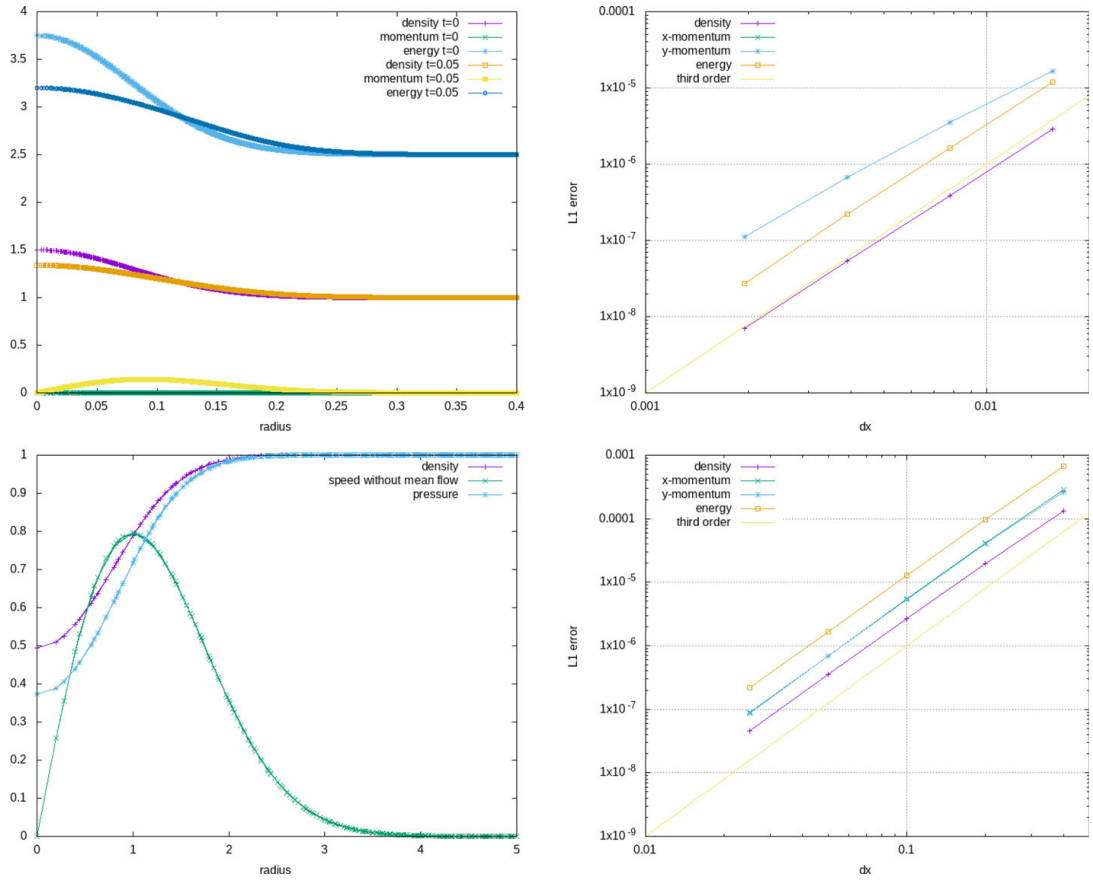

Fig. 4 Convergence study. *Top left*: Setup (49)–(50) at initial time and at $t = 0.05$, shown as a scatter plot as a function of radius, computed on a 256×256 grid. *Bottom left*: Setup (51)–(52) at initial time on a grid of 100×100, shown as a scatter plot. *Right top*: ℓ^1 error of the numerical solution of the point values, i.e. the average of $\frac{1}{|\Omega|}\sum_{ij} |q_{i+\frac{1}{2},j+\frac{1}{2}}(t^n) - q_{\text{ref}}(t^n, x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})|\Delta x \Delta y$ for the setup (49)–(50). The analogous error of the averages has virtually the same values and is not shown. *Right bottom*: Same for the setup (51)–(52)

rapidly. We choose a larger domain than [20], namely $[0, 20]^2$, to ensure that the rotational velocity of the vortex has decreased to the level of machine error at the boundary of the domain.

In both cases limiting is not used and we use a CFL number of 0.2. In Fig. 4 one observes third order accuracy in agreement with the expectation.

4.2 Spherical Shock Tube

As a first test with discontinuities, Fig. 5 shows a 2-dimensional version of Sod's shock tube:

$$\rho_0(x, y) = \begin{cases} 1 & r < 0.3 \\ 0.125 & \text{else} \end{cases} \qquad p_0(x, y) = \begin{cases} 1 & r < 0.3 \\ 0.1 & \text{else} \end{cases} \tag{54}$$

$$u_0(x, y) = v_0(x, y) = 0 \tag{55}$$

with $r = \sqrt{(x - \frac{1}{2})^2 + (y - \frac{1}{2})^2}$. We use a CFL number of 0.05, a reduction we find helpful to avoid negative pressure. One observes that the limiting is successful in suppressing oscillations at the shock and around the rarefaction. However, both the limited and the unlimited

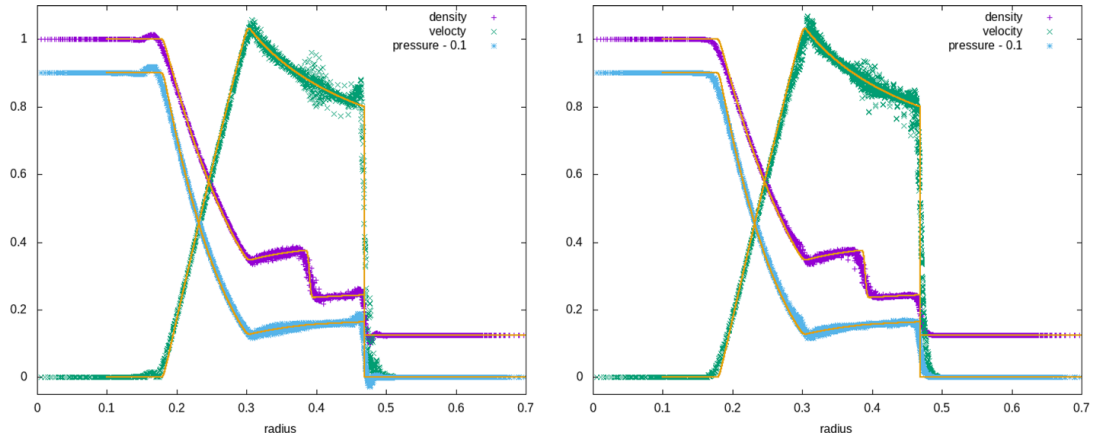

Fig. 5 Radial scatter plot of the two-dimensional version of Sod's shock tube solved on a 100×100 grid. The solid line shows a finely resolved solution of the one-dimensional, radial Euler equations obtained with a standard Finite Volume method. We show the pressure offset by 0.1 for better readability of the plot. *Left*: No limiting. *Right*: Limiting used. The limiting is successful at suppressing oscillations at the shock and around the rarefaction. As in a radial scatter plot points from different locations end up shown in the same location, around the contact wave one rather observes scatter than oscillations, i.e. a deviation from radial symmetry

versions show some scatter around the contact wave, the origin of which will be studied as future work. Global continuity does not impede Active Flux from converging to weak solutions, because the update of the averages is conservative and fulfills a version of the Lax-Wendroff-theorem [15].

4.3 Multi-dimensional Riemann Problems

In [21], particular multi-dimensional Riemann problems were studied, designed such that the one-dimensional Riemann problems outside the central interaction region result in elementary waves. Inside the interaction region these Riemann problems display a lot of sophisticated structure. They will illustrate the ability of the proposed method to solve complex interactions of shocks, rarefactions and slip lines. All the Riemann problems shown in Fig. 6 are solved on grids with $\Delta x = \Delta y = \frac{1}{200}$ (the original publication used $\frac{1}{400}$) with a domain slightly larger than the one shown (to exclude the influence of boundary conditions). A CFL number of 0.05 was used, as well as limiting, as described in Sect. 3. Figures 7 and 8 show results on even coarser meshes. Figures 9 and 10 show a comparison between results obtained with and without limiting. It seems that the intricate structures in the interaction region are not significantly smeared out by the limiting while oscillations at shocks are very efficiently suppressed.

4.4 Low Mach Number Vortex

In e.g. [22] a subsonic vortex from [23] has been used to assess the low Mach number properties of the numerical method. The velocity field associated with this stationary solution of the Euler equations is divergence-free. The Mach number of the flow is modified by changing the background pressure, see [22] for setup details. While numerical methods that are not low Mach compliant diffuse the vortex on acoustic time scales, low Mach number compliant methods apply only advective diffusion, i.e. the rate of decay of the numerical solution is asymptotically independent of the Mach number. Figure 11 shows the vortex on

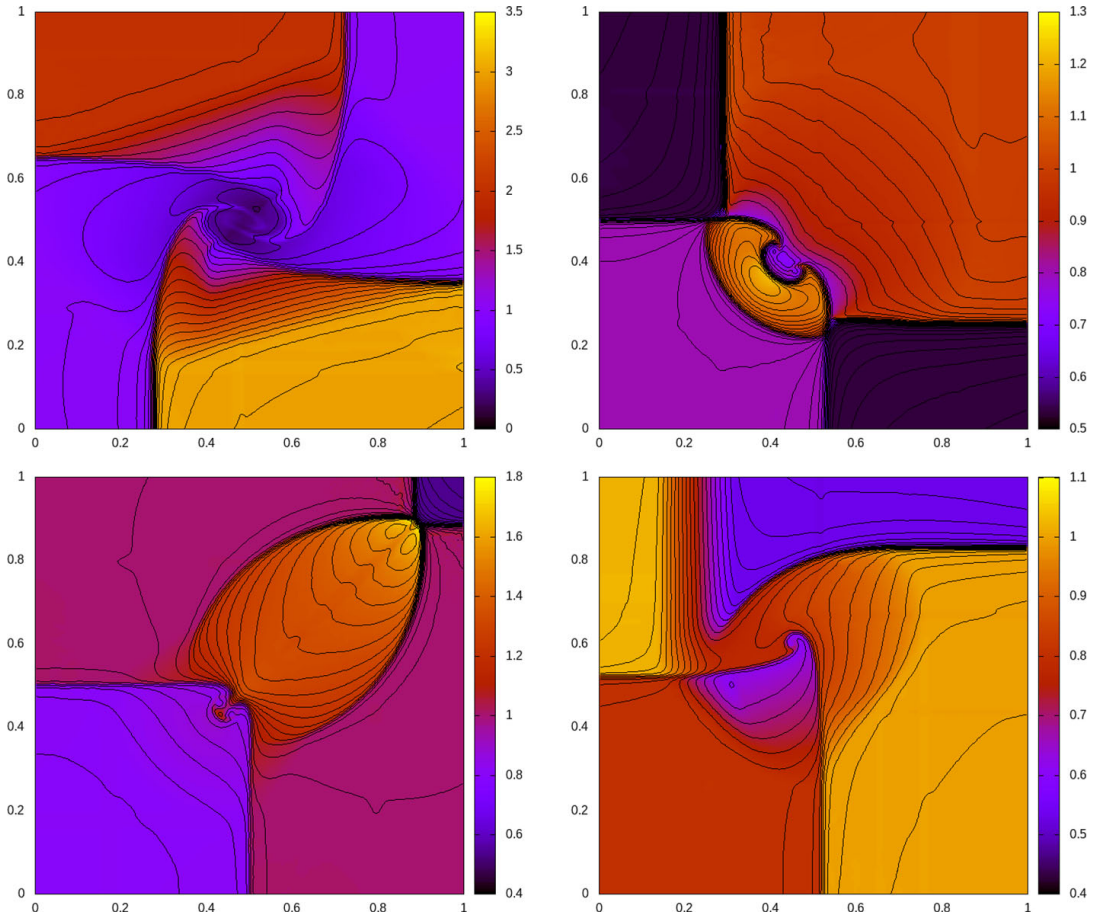

Fig. 6 Multi-dimensional Riemann problems solved on a grid with $\Delta x = \Delta y = \frac{1}{200}$ using limiting as described in Sect. 3. Configurations 6 (*top left*), 11 (*top right*), 12 (*bottom left*) and 16 (*bottom right*) from [21] are shown. Density is shown in color and contour

a 50×50 grid for different values of the Mach number until $t = 1$ with a CFL number of 0.2. One observes the independence of the results on the Mach number leading us to the conclusion that Active Flux is well-suited for subsonic/low Mach number flows. This parallels the finding in [11] where is was shown that Active Flux (upon usage of the exact evolution operator) is stationarity preserving and thus low Mach compliant for linear acoustics. We also observe that the advective diffusion leads to a (Mach number independent) smearing of the vortex, to levels that would require solving this setup on a 200×200 grid with a first-order low Mach number compliant method [24].

The good low Mach number behaviour of the proposed method is due to the way how the point values are evolved, and thus not trivial. For comparison, in Fig. 12 the results of a simulation are shown which uses

$$(J_x)^+ := \frac{1}{2}\left(J_x + s\mathbb{1}\right) \tag{56}$$

$$(J_x)^- := \frac{1}{2}\left(J_x - s\mathbb{1}\right) \qquad s := \max_k |\lambda_k| \tag{57}$$

instead of (9)–(10). One observes artefacts typical of numerical diffusion associated to the acoustic terms, and thus a method unsuitable for the low Mach number regime.

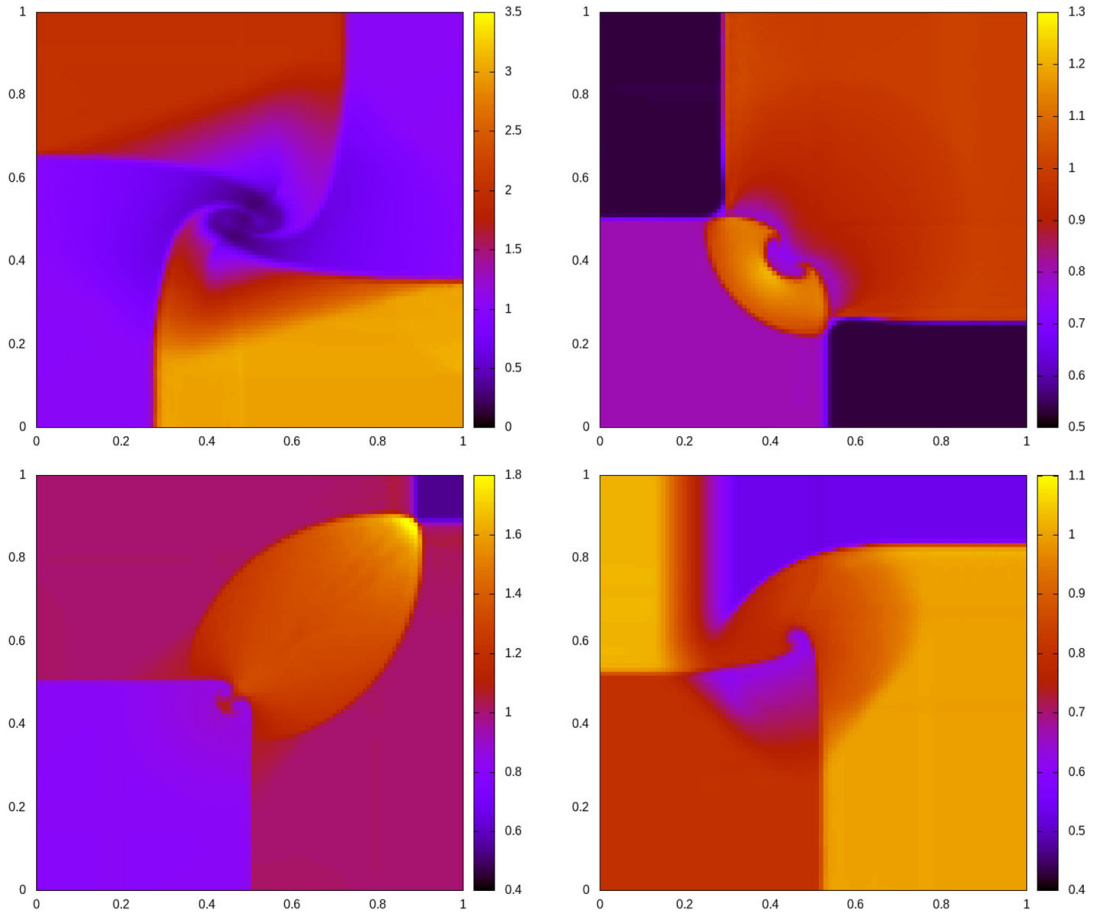

Fig. 7 Setup as in Fig. 6, but solved on a grid with $\Delta x = \Delta y = \frac{1}{100}$

4.5 Kelvin-Helmholtz Instability

A special kind of a Kelvin-Helmholtz instability triggered by the passage of an acoustic wave has been used in [25] to assess the properties of a numerical method for subsonic flow. This setup is of interest because acoustic phenomena are present on top of a low Mach number flow. On a domain $\left[-\frac{1}{\mathcal{M}}, \frac{1}{\mathcal{M}}\right] \times \left[0, \frac{2}{5\mathcal{M}}\right]$, the initial data consist of those of a right-running sound wave q^{s} on top of a vertically stratified background q^{bg}

$$q_0(x, y) = q^{\mathrm{bg}}(y) + q^{\mathrm{s}}(x) \tag{58}$$

with

$$\rho_0^{\mathrm{bg}}(y) = 1 + \varphi(y) \qquad \rho_0^{\mathrm{s}}(x) = \frac{\mathcal{M}}{5}\psi(x) \tag{59}$$

$$u_0^{\mathrm{bg}}(y) = 0 \qquad u_0^{\mathrm{s}}(x) = \sqrt{\gamma}\psi(x) \tag{60}$$

$$v_0^{\mathrm{bg}}(y) = 0 \qquad p_0^{\mathrm{s}}(x) = \frac{1}{\mathcal{M}}\gamma\psi(x) \tag{61}$$

$$p_0^{\mathrm{bg}}(y) = \frac{1}{\mathcal{M}^2} \qquad v_0^{\mathrm{s}}(x) = 0 \tag{62}$$

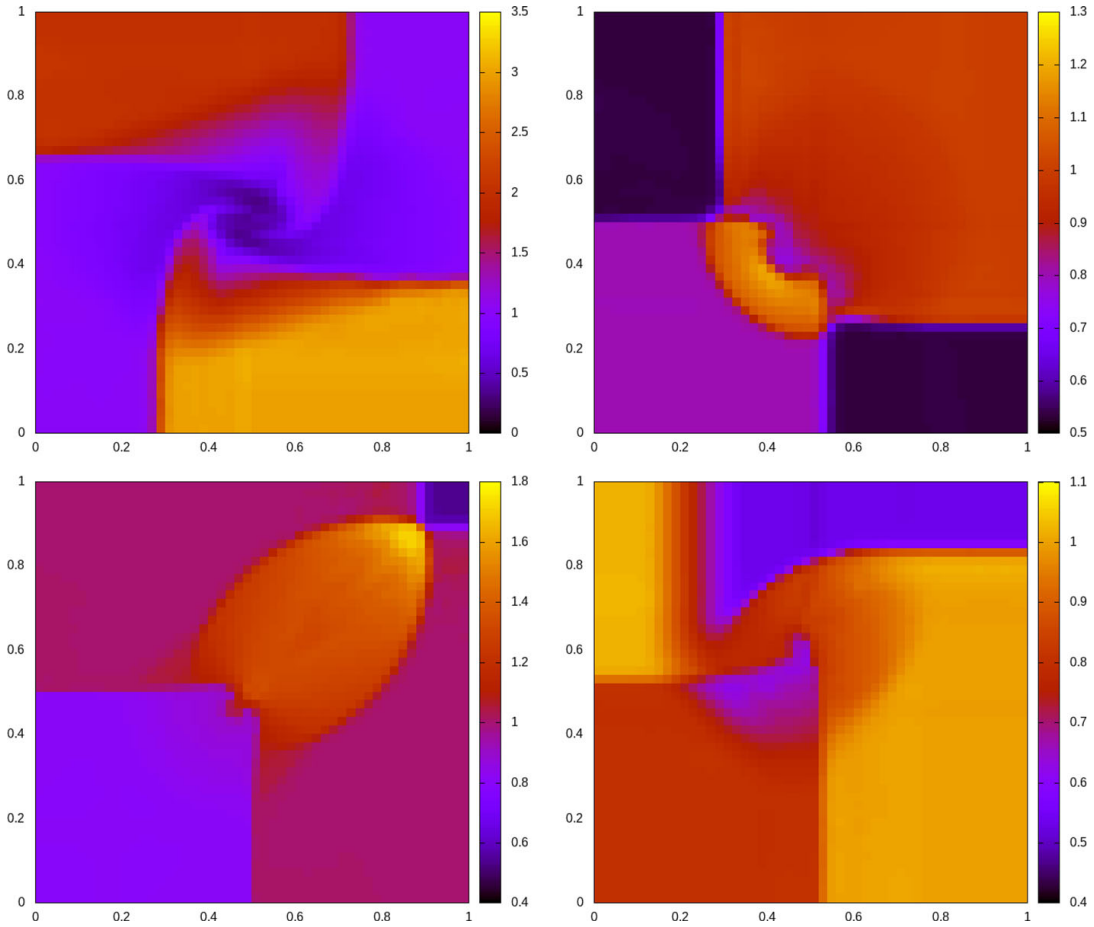

Fig. 8 Setup as in Fig. 6, but solved on a grid with $\Delta x = \Delta y = \frac{1}{50}$

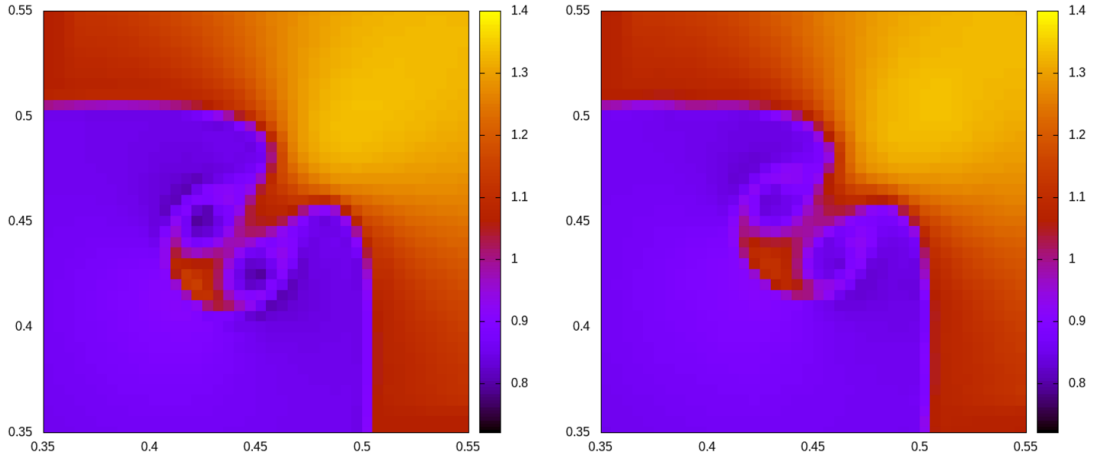

Fig. 9 Influence of limiting on the central region in Configuration 12. *Left*: Limiting off. *Right*: Limiting on. Without limiting one observes some undershoots in the vicinity of the vortices. The structure of the solution feature is, however, not degraded by applying the limiter

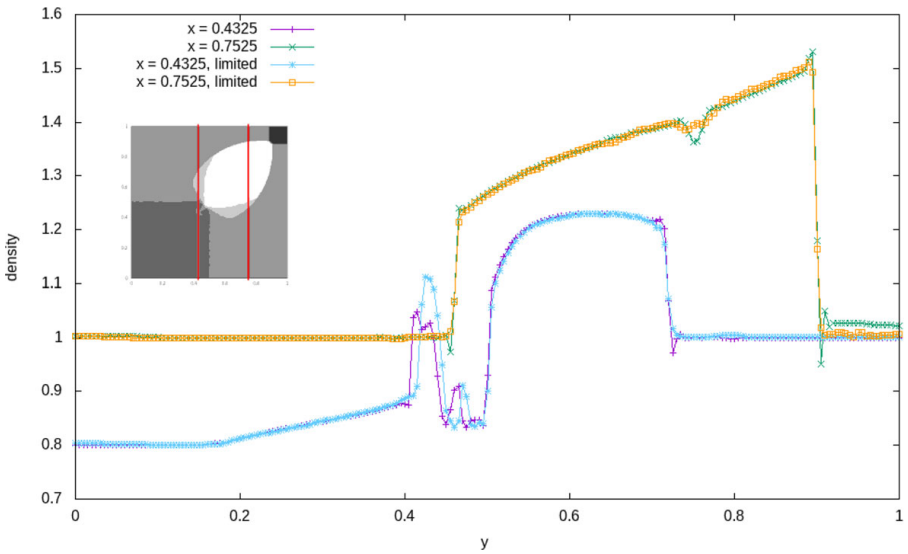

Fig. 10 Influence of limiting on Configuration 12. Density is shown along the lines $x = 0.4325$ and $x = 0.7525$ (as indicated in the inset). One observes that limiting successfully removes spurious oscillations in the vicinity of discontinuities. However, it also gently shifts the location of the central double-vortex and smears out the feature along the $x = y$ diagonal in the first quadrant

and

$$\varphi(y) := \begin{cases} 2\mathcal{M}y & y < 4 \\ 2\mathcal{M}(y-4) - 0.4 & \text{else} \end{cases} \qquad \psi(x) := 1 + \cos(\pi \mathcal{M} x) \tag{63}$$

For the linearized Euler equations

$$\partial_t \rho^{\mathrm{s}}(t,x) + \bar{\rho}\partial_x u^{\mathrm{s}}(t,x) = 0 \tag{64}$$

$$\partial_t u^{\mathrm{s}}(t,x) + \frac{1}{\bar{\rho}}\partial_x p^{\mathrm{s}}(t,x) = 0 \tag{65}$$

$$\partial_t p^{\mathrm{s}}(t,x) + \bar{\rho}c^2\partial_x u^{\mathrm{s}}(t,x) = 0 \tag{66}$$

with $c^2 = \frac{5\gamma}{\mathcal{M}^2}$ and $\bar{\rho} = \frac{1}{\sqrt{5}}$ the initial data $q_0^{\mathrm{s}}(x)$ evolve as follows

$$\rho^{\mathrm{s}}(t,x) = \rho_0^{\mathrm{s}}(x-ct) \qquad u^{\mathrm{s}}(t,x) = u_0^{\mathrm{s}}(x-ct) \qquad p^{\mathrm{s}}(t,x) = p_0^{\mathrm{s}}(x-ct). \tag{67}$$

This justifies referring to q_0^{s} as the initial data of a right-running sound wave. The non-linearity of the full Euler equations leads to its self-steepening over time.

Additionally, due to the saw-tooth-shaped change in y-direction of the background density, the sound wave is not moving the same way above and below the interface located initially at $y = 4$. A shear flow is induced, which causes a Kelvin-Helmholtz-type instability. Here we show this setup on grids of 400×80 (Fig. 13) and 800×160 (Fig. 14) with a CFL of 0.15, periodic boundaries and $\mathcal{M} = \frac{1}{20}$. No limiting was used. Figure 15 shows the results at additional times to facilitate comparison to [25]. Numerical methods that are not low Mach number compliant would add so much diffusion to the subsonic background flow that the instability would be artificially stabilized (unless the grid is refined excessively). One observes that here, the method is able to adequately[3] evolve both the instability (comparable

[3] without claiming that the solution is physically correct, of course, as we are solving only the Euler equations and are neglecting viscosity and other effects crucial for such setups.

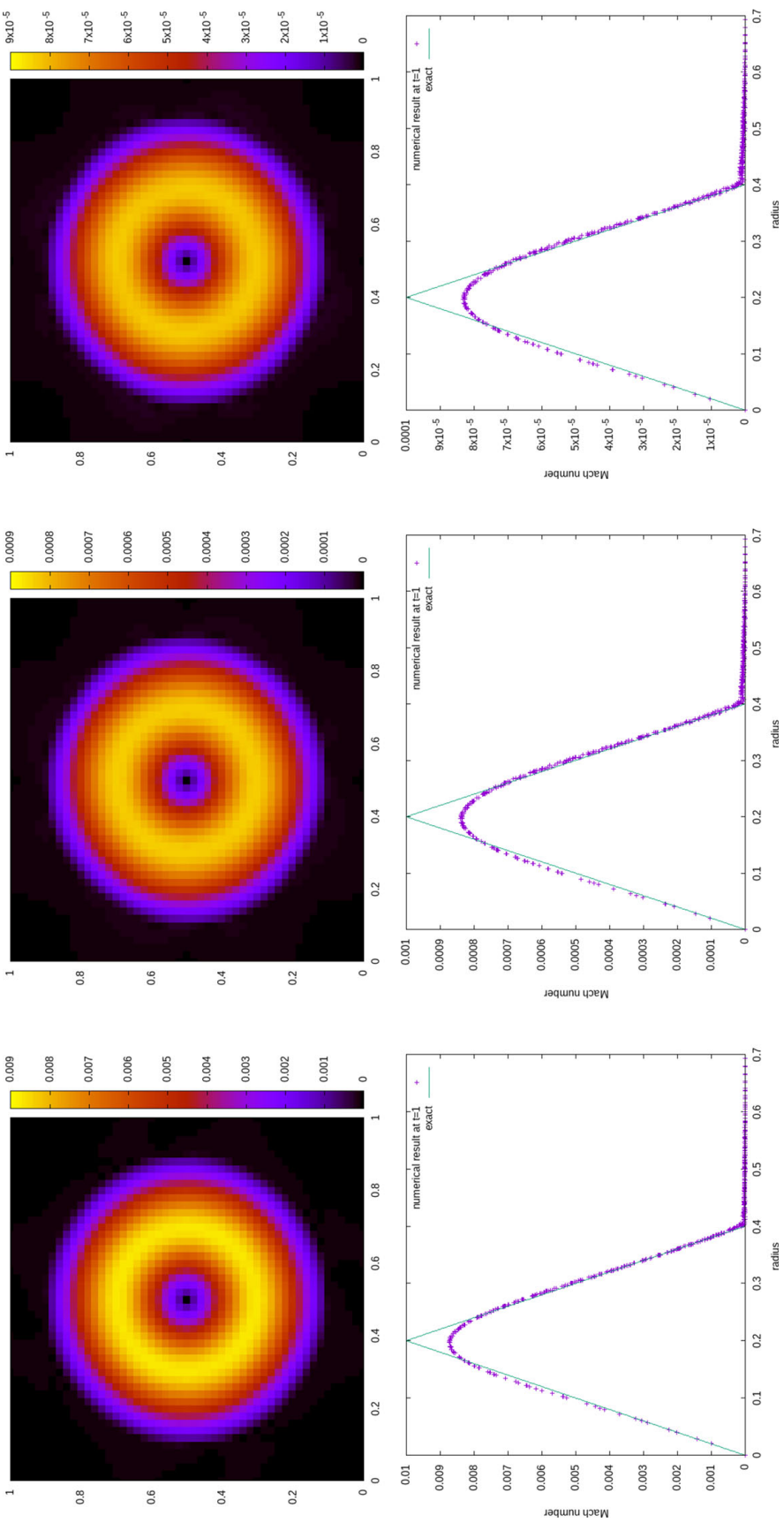

Fig. 11 The Gresho vortex at various maximal Mach numbers (10^{-2}, 10^{-3}, 10^{-4} *from left to right*). While the Mach number decreases by a factor of 10 from one plot to the other, the effect of numerical diffusion remains the same asymptotically, i.e. the observed diffusion is purely advective, not acoustic

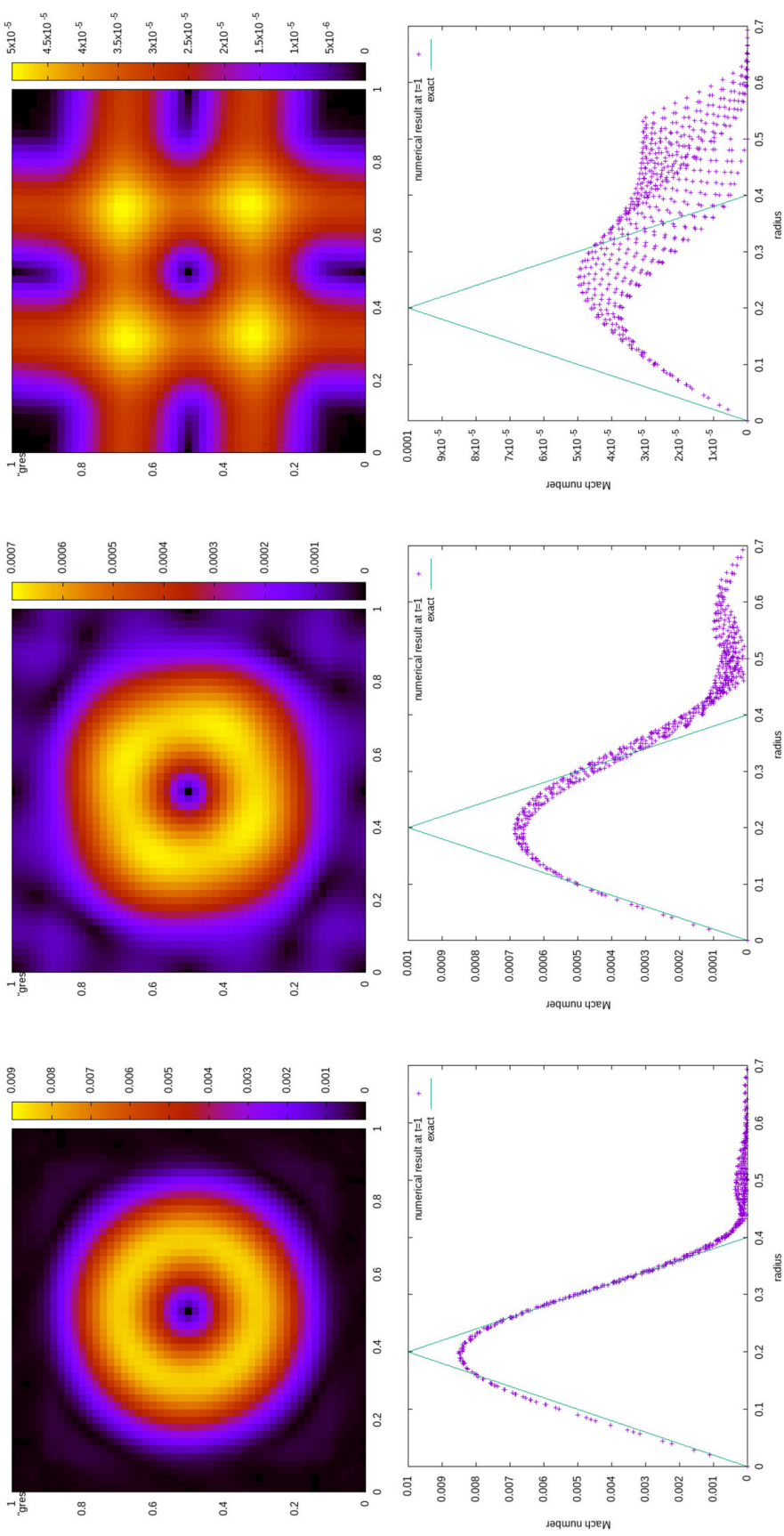

Fig. 12 The Gresho vortex as in Fig. 11 solved using (56)–(57) instead of (9)–(10). The presence of numerical diffusion in the terms associated to acoustics makes the choice (9)–(10) unsuitable for low Mach number computations

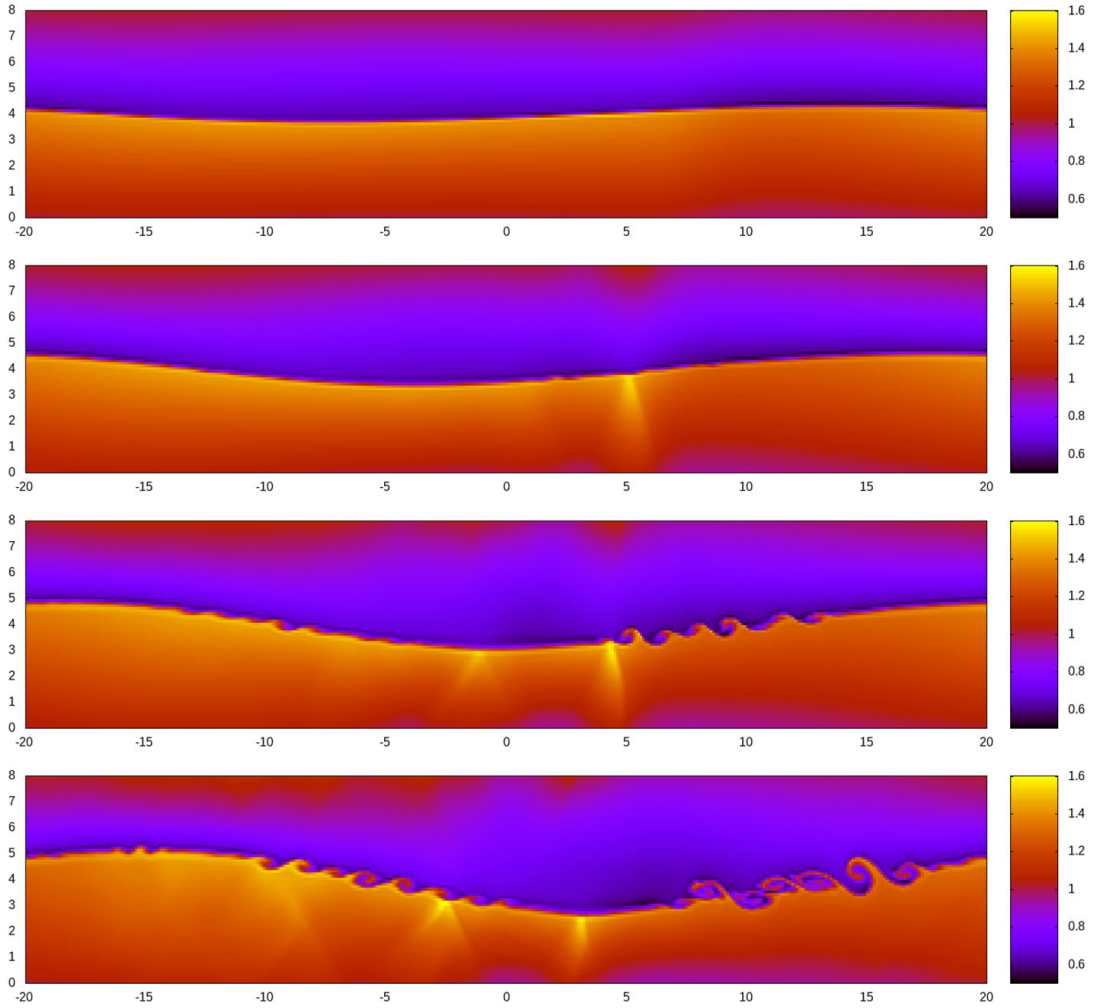

Fig. 13 A Kelvin-Helmholtz instability is triggered by the passage of acoustic waves (visible as weak shocks). The setup is computed on a 400×80 grid without using limiting. Density is shown at times $t = 3, 6, 9, 12$

to results in [26] obtained using a high-order DG method) and the sound waves (which have become weak shocks) passing through the domain. This leads to the hypothesis that the proposed Active Flux method is well-suited for all-speed regimes.

5 Conclusions

Active Flux combines aspects of Finite Volume and Finite Element methods. The evolution of cell averages ensures shock-capturing properties, while the presence of point values at cell interfaces leads to a globally continuous reconstruction, which is a great difference to Godunov methods. The incorporation of additional degrees of freedom and thus the compact nature of the stencil makes the method high-order, but efficient for parallelization or the implementation of boundary conditions. The shared degrees of freedom imply less memory cost than DG methods. Finally, point values do not need to be expressed in conservative variables, i.e. Active Flux offers more freedom than conventional approaches.

For Active Flux, an emphasis on multi-dimensional thinking was put from the beginning, e.g. in [8]. This, again, is different from Godunov methods which are rooted in

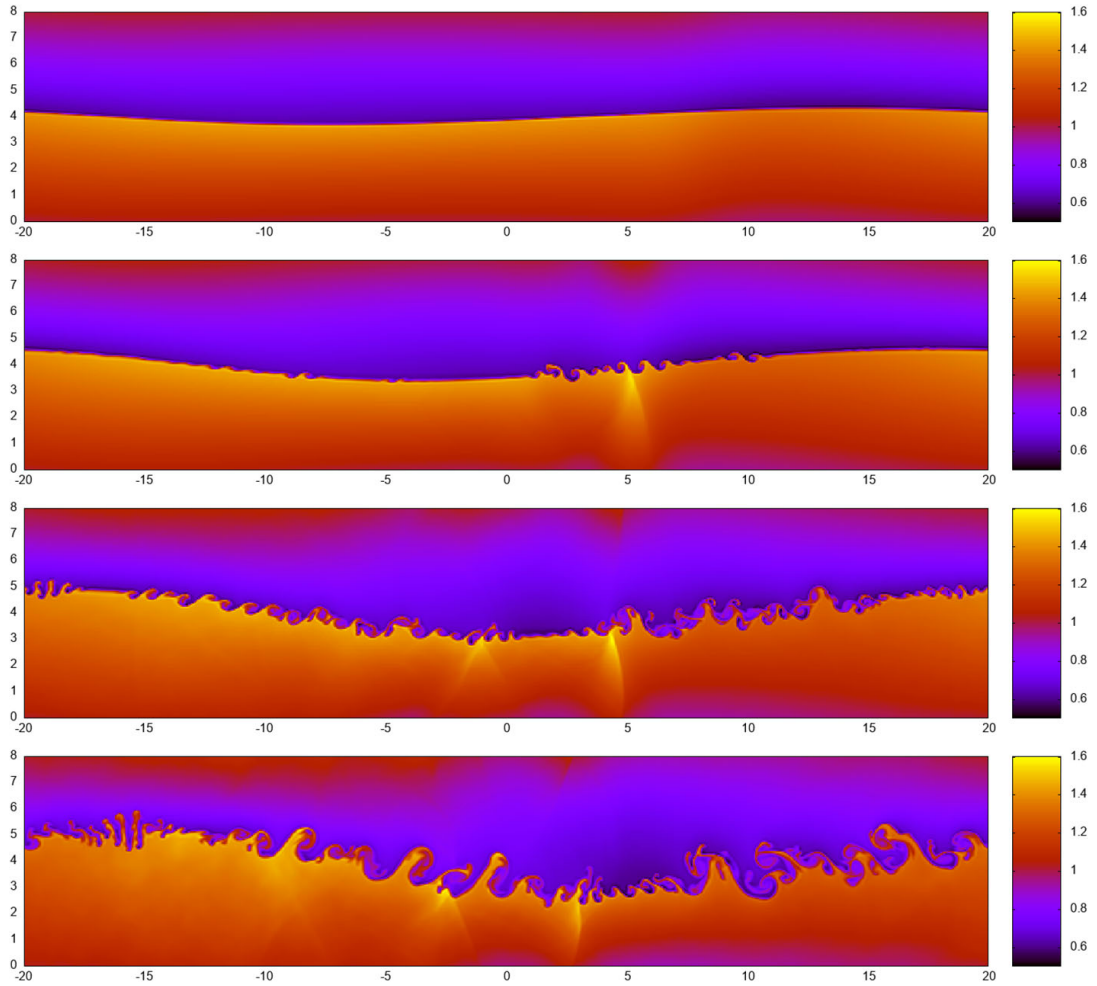

Fig. 14 Same setup as Fig. 13, but on a grid of 800×160

one-dimensional thinking even when applied in multiple dimensions. Interestingly, in the framework of Godunov methods, structure preservation is not obtained even if all the multi-dimensional Riemann problems are solved exactly (see [10]). As has been shown in [11] for linear acoustics, using an exact evolution operator for Active Flux results in a method that is stationarity preserving.

Due to the inherent high-order nature of Active Flux, sufficiently high-order evolution operators for nonlinear, multi-dimensional systems, that can be used in fully discrete Active Flux methods are non-trivial to derive. For a semi-discrete Active Flux method [1, 15], it is easier to find spatial discretizations that use the degrees of freedom of Active Flux and to immediately write down evolution equations for the point values. The semi-discrete problem can then be integrated in time using standard methods, at the price of a reduced CFL number. The incorporation of truly multi-dimensional information is also more difficult.

To show that, however, such an Active Flux method can be successfully used to solve the multi-dimensional Euler equations is the aim of the present work. One finds that, endowed with a limiting strategy, Active Flux is indeed able to solve complex flow problems. This has been demonstrated here for examples of multi-dimensional Riemann problems and for subsonic flows, highlighting in particular the ability of Active Flux to cope with low Mach number flows. This leads us to conjecture that also the semi-discrete Active Flux method possesses structure-preserving properties. A comparison between the semi-discrete and the

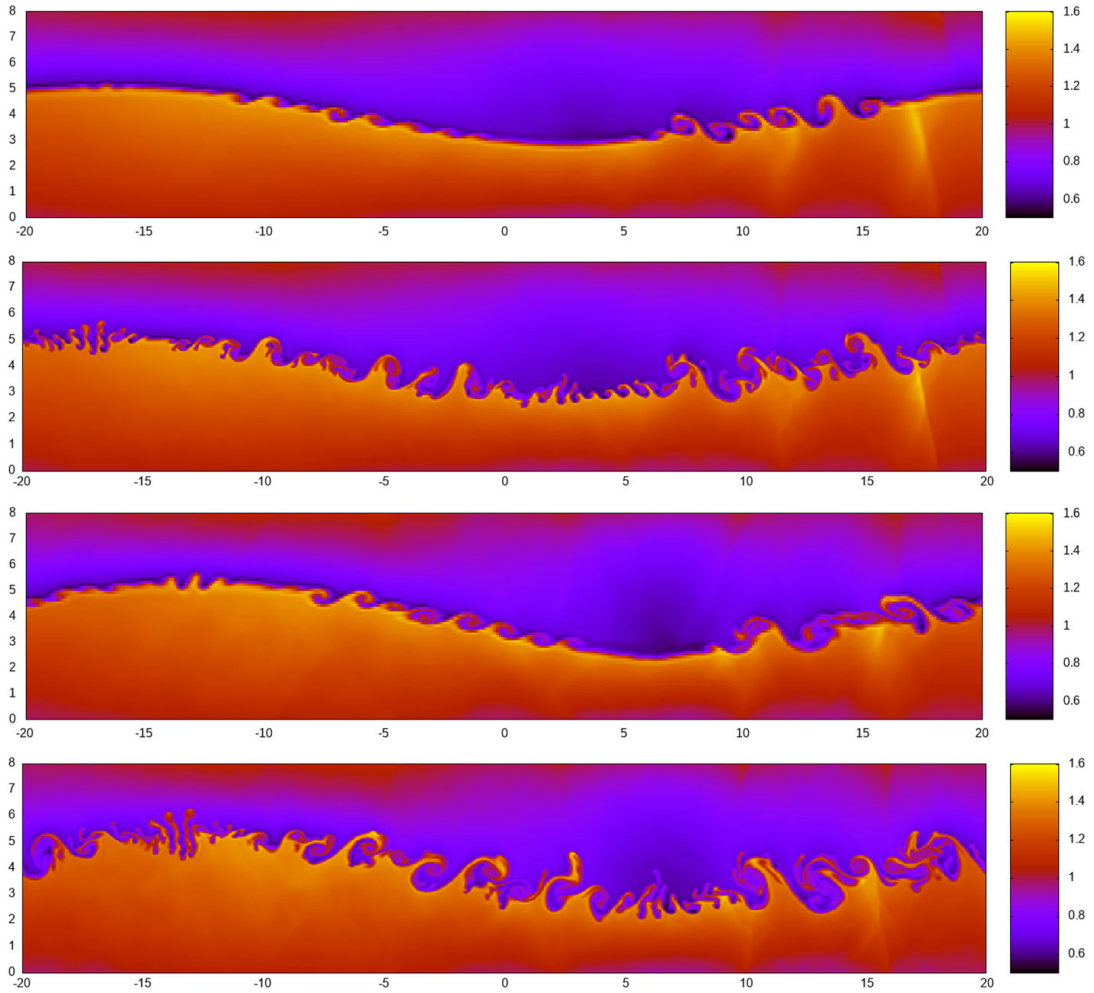

Fig. 15 From top to bottom: 400×80, $t = 11$; 800×160, $t = 11$; 400×80, $t = 14$; 800×160, $t = 14$. Density is shown. These times are the ones shown in [25]

fully discrete, classical approach to Active Flux in the setting of linear acoustics is subject of a manuscript currently in preparation.

The semi-discrete approach can in principle be applied to other hyperbolic systems of conservation laws. Further research is necessary to understand the theoretical aspects of this method, such as entropy inequalities, or to construct appropriate boundary conditions. The presented approach to limiting is rather a proof of concept than a final choice. Future work will be directed towards better limiting procedures, a study of other choices of the point value update and towards preservation of physical conditions such as positivity of the pressure. A comparison between the classical and the semi-discrete approach for nonlinear problems, and to this end further development of high-order evolution operators for multi-dimensional nonlinear systems also need to be addressed in future.

A Detailed Derivation of the Multi-Dimensional Limiting

A.1 Piecewise-Biparabolic Reconstruction

If none of the edges needs to be limited, then the natural choice of the reconstruction is biparabolic, as it has been used already since [4, 11]. However, if one of the edges is reconstructed as a hat, then something else needs to be done inside the cell in order to ensure continuity. We generally choose to subdivide the cell into regions (quadrants or halves, depending on the situation) and to reconstruct biparabolically in every such region while maintaining global continuity.

Then, if the discrete data fulfill condition (30), the reconstruction is tested (in an approximate way) to check whether $m \leq q_{\text{recon}}(x, y) \leq M$ holds inside the cell. In case not, the reconstruction is discarded and replaced by the plateau reconstruction of Sect.A.2.

Linearity of the problem (in the point values and the average) will be exploited by considering all point values apart from $q_{\text{SW}}, q_{\text{W}}, q_{\text{NW}}$ to vanish:

Definition 1 Consider all point values apart from $q_{\text{SW}}, q_{\text{W}}, q_{\text{NW}}$ to vanish. Then a reconstruction of the cell that interpolates these values pointwise and whose average agrees with $\bar{q}$ is called the **edge-basis-function** $q^{\text{W}}_{\text{recon}}$ of the W-edge:

$$q^{\text{W}}_{\text{recon}}\left(-\frac{\Delta x}{2}, \frac{\Delta y}{2}\right) = q_{\text{NW}} \qquad q^{\text{W}}_{\text{recon}}\left(-\frac{\Delta x}{2}, 0\right) = q_{\text{W}} \tag{68}$$

$$q^{\text{W}}_{\text{recon}}\left(-\frac{\Delta x}{2}, -\frac{\Delta y}{2}\right) = q_{\text{SW}} \qquad \frac{1}{\Delta x \Delta y}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}}\int_{-\frac{\Delta y}{2}}^{\frac{\Delta y}{2}} q^{\text{W}}_{\text{recon}}(x, y)\mathrm{d}x\mathrm{d}y = \bar{q} \tag{69}$$

Similar notions will be used for the other edges.

Observe that an edge-basis-function is a reconstruction of the entire cell. In the following, only the edge-basis-functions for the W-edge will be given explicitly, as those for the other edges can be obtained by rotation, as long as $\Delta y = \Delta x$ (otherwise some rescaling is necessary).

Theorem 3 *If the edge-basis-function for the W-edge is*

$$q^{W}_{recon}\left(q_{SW}, q_W, q_{NW}, x, y, S, N, W, \bar{q}\right) \tag{70}$$

then the other basis functions are

$$q^{S}_{recon}\left(q_{SE}, q_S, q_{SW}, x, y, E, W, S, \bar{q}\right) = q^{W}_{recon}\left(q_{SE}, q_S, q_{SW}, y, -x, E, W, S, \bar{q}\right) \tag{71}$$

$$q^{N}_{recon}\left(q_{NW}, q_N, q_{NE}, x, y, W, E, N, \bar{q}\right) = q^{W}_{recon}\left(q_{NW}, q_N, q_{NE}, -y, x, W, E, N, \bar{q}\right) \tag{72}$$

$$q^{E}_{recon}\left(q_{NE}, q_E, q_{SE}, x, y, N, S, E, \bar{q}\right) = q^{W}_{recon}\left(q_{NE}, q_E, q_{SE}, -x, -y, N, S, E, \bar{q}\right) \tag{73}$$

The edge-basis-function depends on $q_{\text{SW}}, q_{\text{W}}, q_{\text{NW}}$, on whether the reconstruction of the W-edge is parabolic or hat, and – this complicates things a little – on whether the neighbouring edges (S and N) are reconstructed as hats or as parabolae. This is necessary due to global continuity and because the corner values $q_{\text{SW}}, q_{\text{NW}}$ are shared with the S- and N-edges.

The final reconstruction is obtained through summation:

Theorem 4 *The following reconstruction q_{recon} interpolates all the point values along the boundary of the cell and its average agrees with the given cell average:*

$$q_{recon}(x, y) := q^{W}_{recon}\left(\frac{q_{SW} - \bar{Q}}{2}, q_W - \bar{Q}, \frac{q_{NW} - \bar{Q}}{2}, x, y, S, N, W, \frac{\bar{q} - \bar{Q}}{4}\right)$$

$$+ q^{S}_{recon}\left(\frac{q_{SE}-\bar{Q}}{2}, q_S-\bar{Q}, \frac{q_{SW}-\bar{Q}}{2}, x, y, E, W, S, \frac{\bar{q}-\bar{Q}}{4}\right)$$
$$+ q^{N}_{recon}\left(\frac{q_{NW}-\bar{Q}}{2}, q_N-\bar{Q}, \frac{q_{NE}-\bar{Q}}{2}, x, y, W, E, N, \frac{\bar{q}-\bar{Q}}{4}\right)$$
$$+ q^{E}_{recon}\left(\frac{q_{NE}-\bar{Q}}{2}, q_E-\bar{Q}, \frac{q_{SE}-\bar{Q}}{2}, x, y, N, S, E, \frac{\bar{q}-\bar{Q}}{4}\right)+\bar{Q} \quad (74)$$

where $\bar{Q} := \frac{q_{SW}+q_W+q_{NW}+q_N+q_{NE}+q_E+q_{SE}+q_S}{8}$. *Moreover, as all the point values tend to* $\bar{q}$,

$$q_{recon}(x, y) \to \bar{q} \quad (75)$$

for all x, y.

Proof The pointwise interpolation property is clear because, for example,

$$q_{\text{recon}}\left(\frac{\Delta x}{2}, \frac{\Delta y}{2}\right) = q^{\text{N}}_{\text{recon}}\left(\frac{q_{\text{NW}}-\bar{Q}}{2}, q_{\text{N}}-\bar{Q}, \frac{q_{\text{NE}}-\bar{Q}}{2}, \frac{\Delta x}{2}, \frac{\Delta y}{2}, \text{W}, \text{E}, \text{N}, \frac{\bar{q}-\bar{Q}}{4}\right) \quad (76)$$
$$+ q^{\text{E}}_{\text{recon}}\left(\frac{q_{\text{NE}}-\bar{Q}}{2}, q_{\text{E}}-\bar{Q}, \frac{q_{\text{SE}}-\bar{Q}}{2}, \frac{\Delta x}{2}, \frac{\Delta y}{2}, \text{N}, \text{S}, \text{E}, \frac{\bar{q}-\bar{Q}}{4}\right)+\bar{Q}$$
$$= \frac{q_{\text{NE}}-\bar{Q}}{2}+\frac{q_{\text{NE}}-\bar{Q}}{2}+\bar{Q} = q_{\text{NE}} \quad (77)$$

The correctness of the average follows from

$$\frac{1}{\Delta x \Delta y}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}}\int_{-\frac{\Delta y}{2}}^{\frac{\Delta y}{2}} q^{\text{W}}_{\text{recon}}(x, y)\text{d}x\text{d}y = 4\cdot\frac{\bar{q}-\bar{Q}}{4}+\bar{Q} = \bar{q} \quad (78)$$

Finally, property (75) is trivial if $\bar{q} = 0$, because the reconstruction is linear in all the point values and in the average, and thus $q_{\text{recon}}(x, y) \to 0$ uniformly in this case. If the point values tend to $\bar{q} \neq 0$, then so does $\bar{Q} \to \bar{q}$ and thus

$$q_{\text{recon}}(x, y) \to 0+\bar{Q} \to \bar{q} \quad (79)$$

□

Remark 1 One might think that it would be sufficient to define the reconstruction as

$$q^{\text{W}}_{\text{recon}}\left(\frac{q_{\text{SW}}}{2}, q_{\text{W}}, \frac{q_{\text{NW}}}{2}, x, y, \text{S}, \text{N}, \text{W}, \frac{\bar{q}}{4}\right)+q^{\text{S}}_{\text{recon}}\left(\frac{q_{\text{SE}}}{2}, q_{\text{S}}, \frac{q_{\text{SW}}}{2}, x, y, \text{E}, \text{W}, \text{S}, \frac{\bar{q}}{4}\right) \quad (80)$$
$$+q^{\text{N}}_{\text{recon}}\left(\frac{q_{\text{NW}}}{2}, q_{\text{N}}, \frac{q_{\text{NE}}}{2}, x, y, \text{W}, \text{E}, \text{N}, \frac{\bar{q}}{4}\right)+q^{\text{E}}_{\text{recon}}\left(\frac{q_{\text{NE}}}{2}, q_{\text{E}}, \frac{q_{\text{SE}}}{2}, x, y, \text{N}, \text{S}, \text{E}, \frac{\bar{q}}{4}\right) \quad (81)$$

This function also has the interpolation properties in Theorem 4. However, in the limit of all the point values converging to $\bar{q}$, property (75) is not guaranteed. Linearity merely implies that in the limit, q_{recon} will be proportional to $\bar{q}$, but it can still have a non-trivial dependence on x, y. The only case where one can be sure of obtaining a uniform constant is when $\bar{q} = 0$.

Thus, one can first subtract a uniform constant k from all the discrete data, for instance the average of all the point values, reconstruct, and add it back:

$$q^{\mathrm{W}}_{\mathrm{recon}}\left(\frac{q_{\mathrm{SW}}}{2}-k, q_{\mathrm{W}}-k, \frac{q_{\mathrm{NW}}}{2}-k, x, y, \mathrm{S}, \mathrm{N}, \mathrm{W}, \frac{\bar{q}-k}{4}\right) \tag{82}$$

$$+ q^{\mathrm{S}}_{\mathrm{recon}}\left(\frac{q_{\mathrm{SE}}}{2}-k, q_{\mathrm{S}}-k, \frac{q_{\mathrm{SW}}}{2}-k, x, y, \mathrm{E}, \mathrm{W}, \mathrm{S}, \frac{\bar{q}-k}{4}\right) \tag{83}$$

$$+q^{\mathrm{N}}_{\mathrm{recon}}\left(\frac{q_{\mathrm{NW}}}{2}-k, q_{\mathrm{N}}-k, \frac{q_{\mathrm{NE}}}{2}-k, x, y, \mathrm{W}, \mathrm{E}, \mathrm{N}, \frac{\bar{q}-k}{4}\right) \tag{84}$$

$$+ q^{\mathrm{E}}_{\mathrm{recon}}\left(\frac{q_{\mathrm{NE}}}{2}, q_{\mathrm{E}}, \frac{q_{\mathrm{SE}}}{2}, x, y, \mathrm{N}, \mathrm{S}, \mathrm{E}, \frac{\bar{q}-k}{4}\right)+k \tag{85}$$

Now, if all point values tend to $\bar{q}$, k also tends to $\bar{q}$, and all the reconstructions tend to uniform 0. The choice of $k=\bar{Q}$ is a simple one, but other ones are possible.

As mentioned in Sect. 3, instead of directly imposing the correct average over the cell, the value q_{C} of the reconstruction at the cell center is imposed. Later, given the cell average, q_{C} is found as the solution of a linear equation. This is a purely algorithmic detour not affecting the results, but it makes it easier to combine all the different cases. Below, these linear equatiosn linking $\bar{q}$ and q_{C} are given together with the reconstructions.

If the reconstruction happens on the unit square, then $\Delta x=\Delta y=1$ should be used in the formulas below. The sketches of the interpolation problem are encoded as follows: ● denotes the central value q_{C}, ○ / ○ denotes a value that is not on the W edge and thus zero (gray if it is not used in the interpolation), ✕ / ✕ denotes one of the values q_{NW}, q_{W}, q_{SW} (gray if it is not used in the interpolation). Values marked with an arrow do not, in principle, need to be included in the interpolation stencil, but are included here in order to achieve continuity. The colored area denotes the support of the different functions that make up the piecewise defined reconstruction. $\underline{\text{lin}}$ denotes an edge that is reconstructed linearly, in other words, as part of the interpolation procedure, we impose that the restriction of the reconstruction onto that edge is linear (the quadratic term vanishing).

In many cases, the reconstruction is (piecewise) biparabolic, i.e. of the form

$$\left(a_0+a_1x+a_2x^2\right)+\left(a_3+a_4x+a_5x^2\right)y+\left(a_6+a_7x+a_8x^2\right)y^2 \tag{86}$$

In the following, biparabolic reconstructions are given by specifying the values of these 9 coefficients.

A.1.1 Parabolic Reconstruction on W Edge

If edges W, S and N are all reconstructed parabolically, then the W-edge-basis-function is a biparabolic function. If either S or N (or both) are reconstructed as hat functions, the reconstruction in the cell is defined piecewise: the left and the rights halves of the cell have individual biparabolic reconstructions, which are joined in a continuous fashion.

Parabolic reconstruction on both neighbouring edges

If both neighbouring edges (N and S) are reconstructed parabolically, then the reconstruction inside the cell is the trivial biparabolic reconstruction (see Fig. 16):

$$q^{\mathrm{W}}_{\mathrm{recon}}=\left\{a_0=q_{\mathrm{C}}, a_1=-\frac{q_{\mathrm{W}}}{\Delta x}, a_2=-\frac{2(2q_{\mathrm{C}}-q_{\mathrm{W}})}{\Delta x^2}, a_3=0, a_4=-\frac{q_{\mathrm{NW}}-q_{\mathrm{SW}}}{\Delta x\,\Delta y},\right. \tag{87}$$

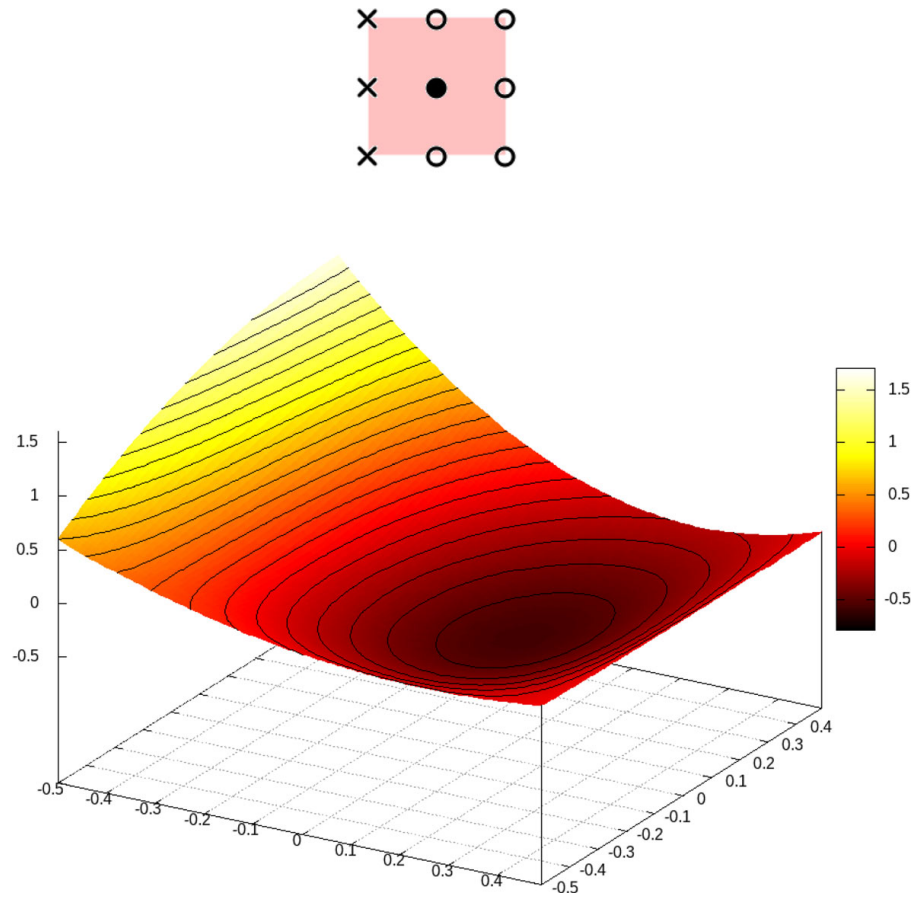

Fig. 16 All edges are reconstructed parabolically, and the corresponding edge-basis-function is a simple biparabolic interpolation. $q_{\mathrm{NW}} = 1.6$, $q_{\mathrm{W}} = 1.35$, $q_{\mathrm{SW}} = 0.6$

$$\left.\begin{aligned} a_5 &= \frac{2(q_{\mathrm{NW}} - q_{\mathrm{SW}})}{\Delta x^2 \Delta y}, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = -\frac{2(q_{\mathrm{NW}} + q_{\mathrm{SW}} - 2q_{\mathrm{W}})}{\Delta x \Delta y^2}, \\ a_8 &= \frac{4(4q_{\mathrm{C}} + q_{\mathrm{NW}} + q_{\mathrm{SW}} - 2q_{\mathrm{W}})}{\Delta x^2 \Delta y^2} \end{aligned}\right\}$$

$$q_{\mathrm{C}} = \frac{1}{16}(36\bar{q} - q_{\mathrm{NW}} - q_{\mathrm{SW}} - 4q_{\mathrm{W}}) \tag{88}$$

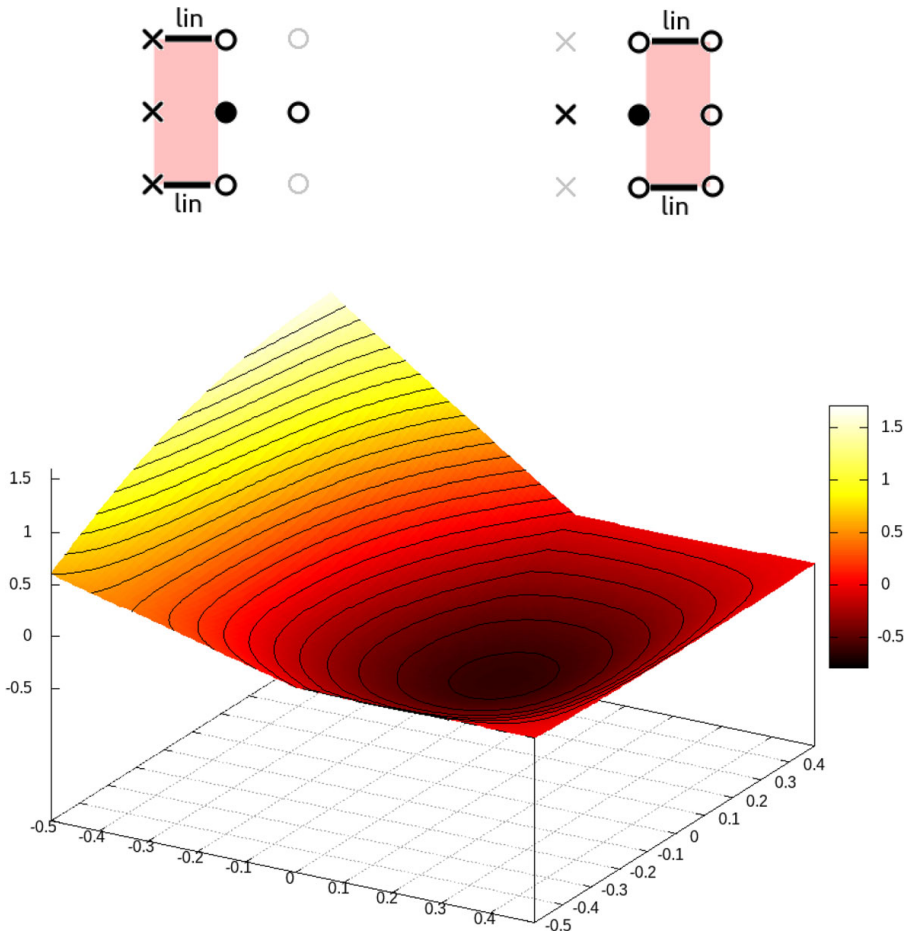

Fig. 17 *Top*: The case of both neighbouring edges reconstructed using hat functions, while the primary edge is reconstructed parabolically. *Bottom*: The W edge is reconstructed parabolically, while the two neighbouring reconstructions are hat functions. $q_{\mathrm{NW}} = 1.6$, $q_{\mathrm{W}} = 1.35$, $q_{\mathrm{SW}} = 0.6$

Hat reconstruction on both neighbouring edges

The interpolation problem is shown in Fig. 17.

$$q^{\mathrm{W}}_{\mathrm{recon}}\Big|_{x<0} = \Big\{a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = -\frac{2(q_{\mathrm{NW}} - q_{\mathrm{SW}})}{\Delta x \Delta y}, \tag{89}$$
$$a_5 = 0, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = -\frac{4(q_{\mathrm{NW}} + q_{\mathrm{SW}} - q_{\mathrm{W}})}{\Delta x \Delta y^2}, a_8 = \frac{8(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2 \Delta y^2}\Big\}$$

$$q^{\mathrm{W}}_{\mathrm{recon}}\Big|_{x\geq 0} = \Big\{a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = 0, \tag{90}$$
$$a_5 = 0, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = \frac{4q_{\mathrm{W}}}{\Delta x \Delta y^2}, a_8 = \frac{8(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2 \Delta y^2}\Big\}$$

$$q_{\mathrm{C}} = \frac{1}{32}(72\bar{q} - 3(q_{\mathrm{NW}} + q_{\mathrm{SW}}) - 8q_{\mathrm{W}}) \tag{91}$$

Hat reconstruction on just one neighbouring edge

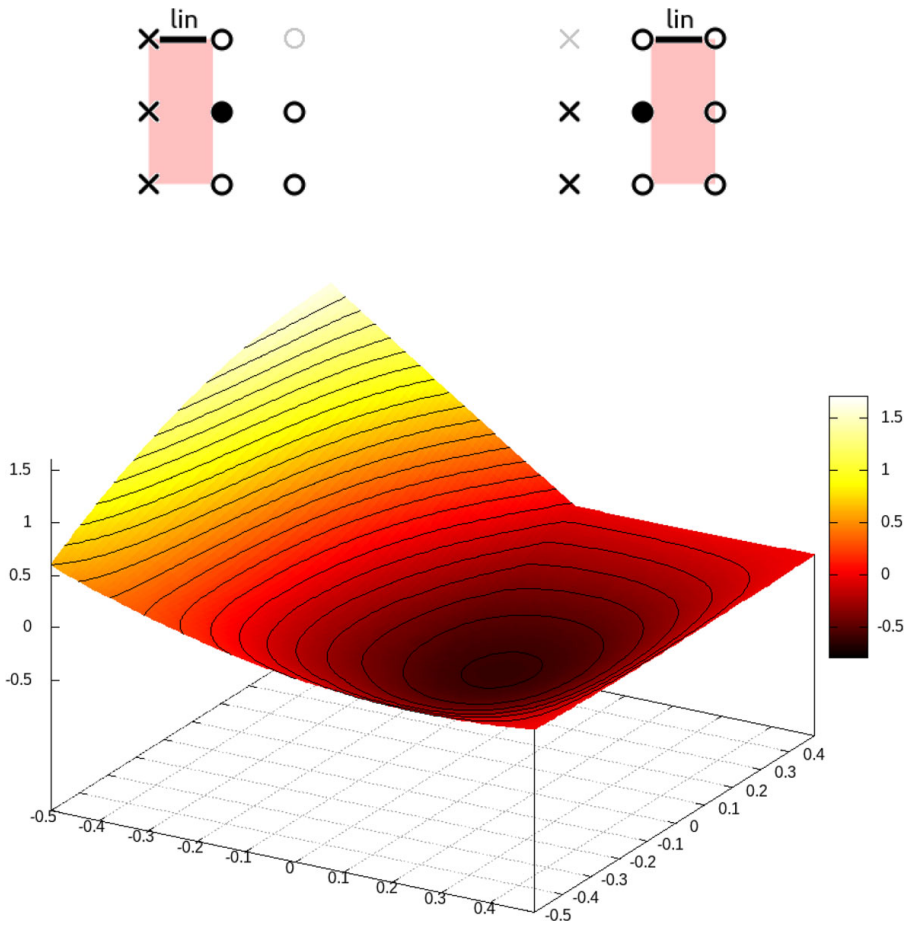

Fig. 18 *Top*: The case of the N edge reconstructed using hat functions, while the primary edge and the S-edge is reconstructed parabolically. *Bottom*: The W and S edge is reconstructed parabolically, while the N edge is reconstructed using a hat function. $q_{\mathrm{NW}} = 1.6$, $q_{\mathrm{W}} = 1.35$, $q_{\mathrm{SW}} = 0.6$

If the N edge is reconstructed using a hat function, and both the W-edge and the S-edge parabolically, then one reconstructs the cell as follows (Fig. 18):

$$q^{\mathrm{W}}_{\mathrm{recon}}\Big|_{x<0} = \Big\{a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = -\frac{2q_{\mathrm{NW}} - q_{\mathrm{SW}}}{\Delta x \Delta y}, \tag{92}$$

$$a_5 = -\frac{2q_{\mathrm{SW}}}{\Delta x^2 \Delta y}, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = -\frac{2(2q_{\mathrm{NW}} + q_{\mathrm{SW}} - 2q_{\mathrm{W}})}{\Delta x \Delta y^2},$$

$$a_8 = \frac{4(4q_{\mathrm{C}} + q_{\mathrm{SW}} - 2q_{\mathrm{W}})}{\Delta x^2 \Delta y^2}\Big\}$$

$$q^{\mathrm{W}}_{\mathrm{recon}}\Big|_{x\geq 0} = \Big\{a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = \frac{q_{\mathrm{SW}}}{\Delta x \Delta y}, \tag{93}$$

$$a_5 = -\frac{2q_{\mathrm{SW}}}{\Delta x^2 \Delta y}, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = -\frac{2(q_{\mathrm{SW}} - 2q_{\mathrm{W}})}{\Delta x \Delta y^2},$$

$$a_8 = \frac{4(4q_{\mathrm{C}} + q_{\mathrm{SW}} - 2q_{\mathrm{W}})}{\Delta x^2 \Delta y^2}\Big\}$$

$$q_{\mathrm{C}} = \frac{1}{32}(72\bar{q} - 3q_{\mathrm{NW}} - 2(q_{\mathrm{SW}} + 4q_{\mathrm{W}})) \tag{94}$$

If it is the S edge, then (Fig. 19):

$$q^{\mathrm{W}}_{\mathrm{recon}}\Big|_{x<0} = \Big\{a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = -\frac{q_{\mathrm{NW}} - 2q_{\mathrm{SW}}}{\Delta x \Delta y}, \tag{95}$$
$$a_5 = \frac{2q_{\mathrm{NW}}}{\Delta x^2 \Delta y}, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = -\frac{2(q_{\mathrm{NW}} + 2q_{\mathrm{SW}} - 2q_{\mathrm{W}})}{\Delta x \Delta y^2},$$
$$a_8 = \frac{4(4q_{\mathrm{C}} + q_{\mathrm{NW}} - 2q_{\mathrm{W}})}{\Delta x^2 \Delta y^2}\Big\}$$

$$q^{\mathrm{W}}_{\mathrm{recon}}\Big|_{x\geq 0} = \Big\{a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = -\frac{q_{\mathrm{NW}}}{\Delta x \Delta y}, \tag{96}$$
$$a_5 = \frac{2q_{\mathrm{NW}}}{\Delta x^2 \Delta y}, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = -\frac{2(q_{\mathrm{NW}} - 2q_{\mathrm{W}})}{\Delta x \Delta y^2},$$
$$a_8 = \frac{4(4q_{\mathrm{C}} + q_{\mathrm{NW}} - 2q_{\mathrm{W}})}{\Delta x^2 \Delta y^2}\Big\}$$

$$q_{\mathrm{C}} = \frac{1}{32}(72\bar{q} - 2q_{\mathrm{NW}} - 3q_{\mathrm{SW}} - 8q_{\mathrm{W}}) \tag{97}$$

Proof of continuity

It is obvious from the sketches of the interpolation problem in Figs. 16, 17, 18 and 19 that the reconstructions interpolate the values on the cell interfaces. What remains to be shown is that the piecewise defined reconstruction is continuous:

Theorem 5 *The reconstructions from Sect. A.1.1. are continuous along the line $x = 0$ where the two pieces are joined.*

Proof As is obvious from the sketches of the interpolation problems in Figs. 17, 18 and 19, the three points along $x = 0$, i.e.

$$q_{\mathrm{recon}}\left(0, \frac{\Delta y}{2}\right) = 0 \qquad q_{\mathrm{recon}}(0, 0) = q_{\mathrm{C}} \qquad q_{\mathrm{recon}}\left(0, -\frac{\Delta y}{2}\right) = 0 \tag{98}$$

are part of the interpolation. Recall that the restriction of a biparabolic function onto the straight line $x = 0$ is a parabola in y, and that the latter is uniquely defined by three points. Therefore, all the values of the reconstruction along $x = 0$ agree for all the reconstructions presented in Sect. A.1.1. □

A.1.2 Hat Reconstruction on W Edge

If the W-edge is reconstructed as a hat function, then necessarily one needs to consider a piecewise defined reconstruction with the pieces joined along $y = 0$. The reconstruction in each piece only depends on whether the other adjacent edge is reconstructed parabolically or as a hat function. One thus has less cases to consider.

Consider the top piece, i.e. the one defined on $[-\frac{\Delta x}{2}, \frac{\Delta x}{2}] \times [0, \frac{\Delta y}{2}]$. It is bordered by the N-edge. If the N-edge is reconstructed as a hat function then one needs additionally to define the reconstruction piecewise in the left and right halves (joined along $x = 0$), i.e. the reconstruction is piecewise by quadrant. This is not necessary if the N-edge is reconstructed parabolically.

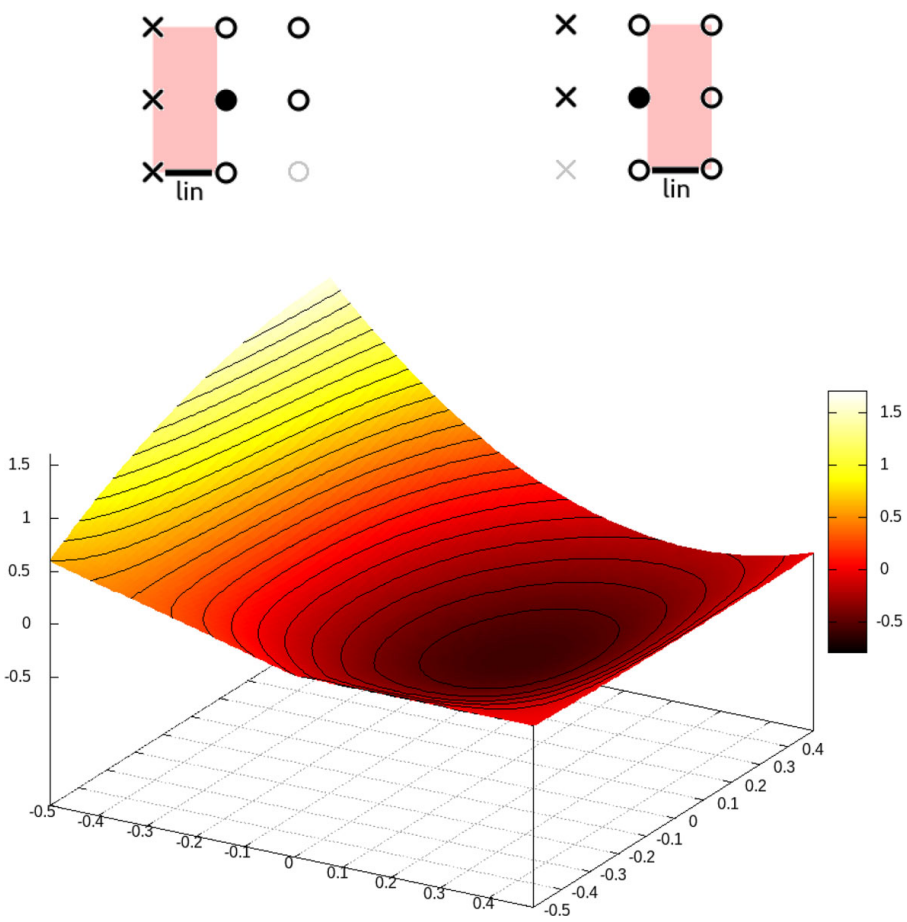

Fig. 19 *Top*: The case of the S edge reconstructed using hat functions, while the primary edge is reconstructed parabolically. *Bottom*: The W and N edge is reconstructed parabolically, while the S edge is reconstructed using a hat function. $q_{\mathrm{NW}} = 1.6$, $q_{\mathrm{W}} = 1.35$, $q_{\mathrm{SW}} = 0.6$

Parabolic Reconstruction on at Least One Neighbouring Edge

Here, the situation is considered in which either the N-edge or the S-edge are reconstructed as parabolae. Then it is possible to provide a biparabolic reconstruction of, respectively, the top or bottom half of the cell.

These cases can occur individually or simultaneously. If both the N-edge and the S-edge are reconstructed parabolically, then the entire reconstruction of the cell is given by the two pieces given in (99)–(101). If, for example, the N-edge is reconstructed parabolically, and the S-edge as a hat function, then the top piece of the reconstruction in the cell is to be taken from (99), while the bottom piece used should be the one from (106)–(107).

See Fig. 20 for the setup of the interpolation problem.

$$q^{W}_{\mathrm{recon}}\Big|_{y\geq 0} = \Big\{a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = -\frac{2(q_{\mathrm{NW}} - q_{\mathrm{W}})}{\Delta x \Delta y}, \tag{99}$$

$$a_5 = \frac{4(q_{\mathrm{NW}} - q_{\mathrm{W}})}{\Delta x^2 \Delta y}, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = 0, a_8 = \frac{16q_{\mathrm{C}}}{\Delta x^2 \Delta y^2}\Big\}$$

$$\frac{1}{\Delta x \Delta y}\int_{y\geq 0} q_{\mathrm{recon}}\,\mathrm{d}x\mathrm{d}y = \frac{2q_{\mathrm{C}}}{9} + \frac{q_{\mathrm{NW}} + q_{\mathrm{W}}}{24} \tag{100}$$

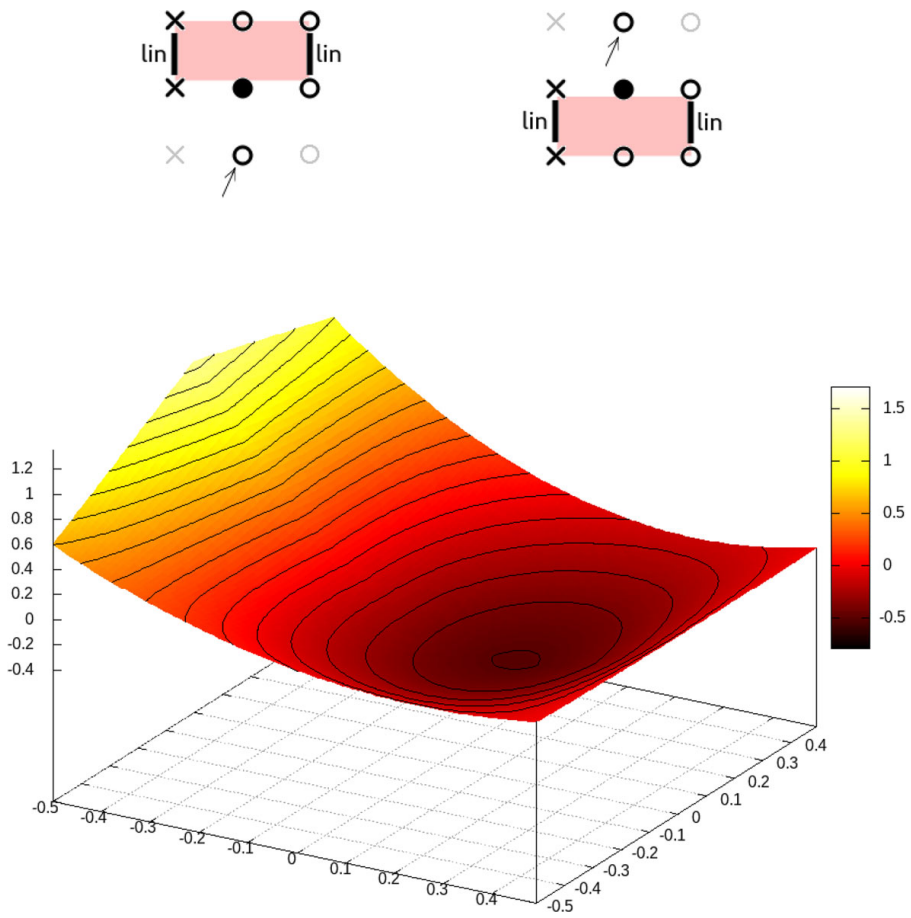

Fig. 20 *Top*: The case of the neighbouring edges reconstructed using parabolas, while the primary edge is reconstructed using the hat function. *Bottom*: The W edge is reconstructed as a hat function, the other edges are reconstructed parabolically. $q_{\mathrm{NW}} = 1$, $q_{\mathrm{W}} = 1.5$, $q_{\mathrm{SW}} = 0$

$$q^{W}_{\mathrm{recon}}\Big|_{y<0} = \Big\{ a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = \frac{2(q_{\mathrm{SW}} - q_{\mathrm{W}})}{\Delta x \Delta y}, \tag{101}$$

$$a_5 = -\frac{4(q_{\mathrm{SW}} - q_{\mathrm{W}})}{\Delta x^2 \Delta y}, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = 0, a_8 = \frac{16q_{\mathrm{C}}}{\Delta x^2 \Delta y^2} \Big\}$$

$$\frac{1}{\Delta x \Delta y} \int_{y<0} q_{\mathrm{recon}} \, \mathrm{d}x\mathrm{d}y = \frac{2q_{\mathrm{C}}}{9} + \frac{q_{\mathrm{SW}} + q_{\mathrm{W}}}{24} \tag{102}$$

Hat Reconstruction on at Least One Neighbouring Edge

In this case the reconstruction is defined piecewise on each quadrant. The biparabolic reconstructions are obtained from interpolation problems shown in Fig. 21.

If the N edge is reconstructed as a hat function, then the top half $[-\frac{\Delta x}{2}, \frac{\Delta x}{2}] \times [0, \frac{\Delta y}{2}]$ of the cell is to be reconstructed as

$$q^{\mathrm{W}}_{\mathrm{recon}}\Big|_{y\geq 0, x<0} = \Big\{ a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = -\frac{3q_{\mathrm{NW}} - 2q_{\mathrm{W}}}{\Delta x \Delta y}, \tag{103}$$

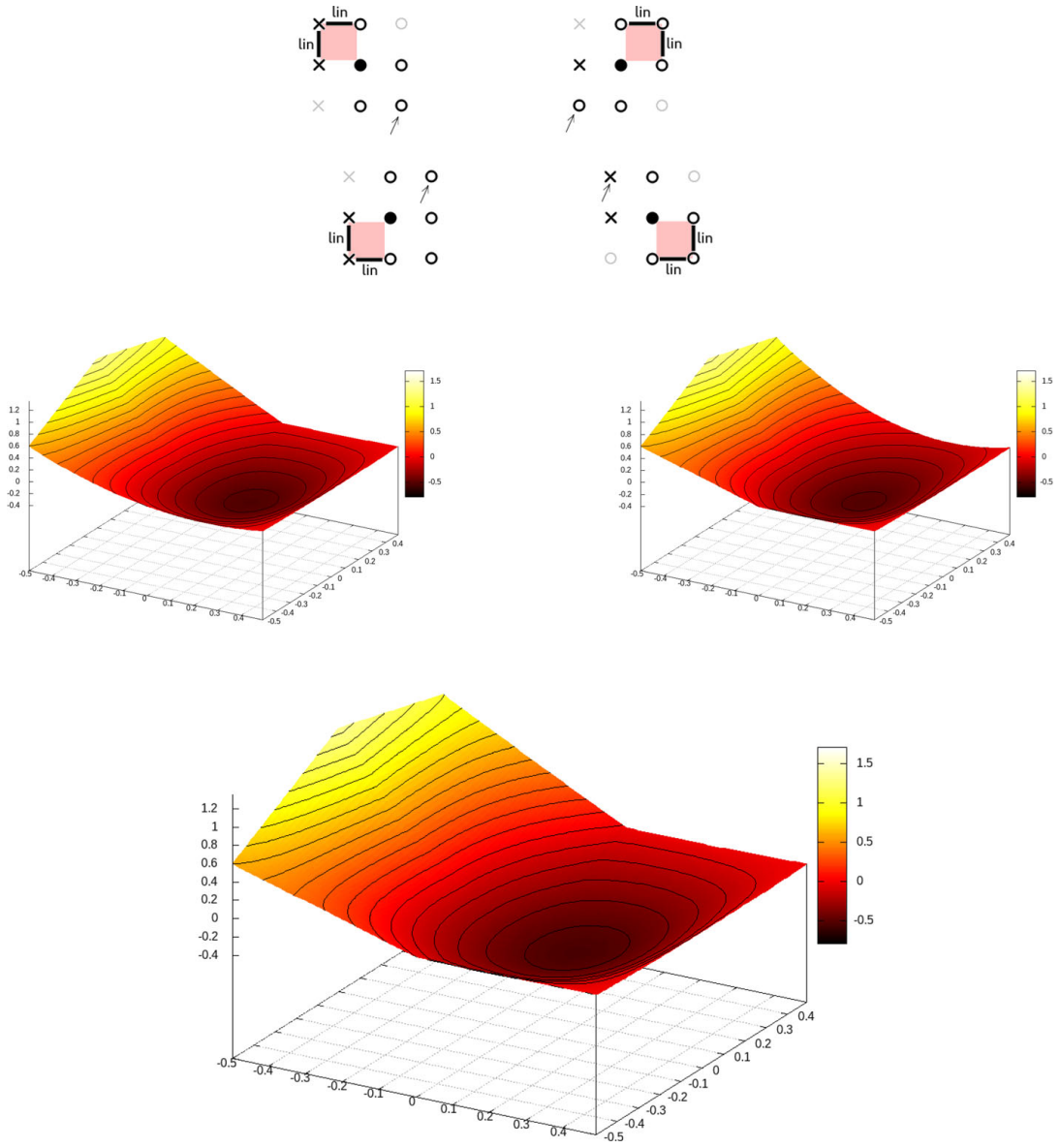

Fig. 21 *Top*: The case of (possibly) all three edges reconstructed using hat functions. The reconstruction inside the cell is defined on four quadrants. *Middle*: The W edge is reconstructed as a hat function, and also N (*left*) / S (*right*). *Bottom*: All edges are reconstructed as hat functions

$$
\begin{aligned}
a_5 = \frac{2(q_{\mathrm{NW}} - 2q_{\mathrm{W}})}{\Delta x^2 \Delta y}, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = -\frac{2q_{\mathrm{NW}}}{\Delta x \Delta y^2}, \\
a_8 = \frac{4(4q_{\mathrm{C}} - q_{\mathrm{NW}})}{\Delta x^2 \Delta y^2} \Bigg\} \\
q^{\mathrm{W}}_{\mathrm{recon}}\Big|_{y\geq 0, x\geq 0} = \Bigg\{ a_0 = q_{\mathrm{C}}, a_1 = -\frac{q_{\mathrm{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\mathrm{C}} - q_{\mathrm{W}})}{\Delta x^2}, a_3 = 0, a_4 = \frac{q_{\mathrm{SW}}}{\Delta x \Delta y}, \\
a_5 = -\frac{2q_{\mathrm{SW}}}{\Delta x^2 \Delta y}, a_6 = -\frac{4q_{\mathrm{C}}}{\Delta y^2}, a_7 = -\frac{2(q_{\mathrm{SW}} - 2q_{\mathrm{W}})}{\Delta x \Delta y^2}, \\
a_8 = \frac{4(4q_{\mathrm{C}} + q_{\mathrm{SW}} - 2q_{\mathrm{W}})}{\Delta x^2 \Delta y^2} \Bigg\}
\end{aligned}
\tag{104}
$$

$$\frac{1}{\Delta x\Delta y}\int_{y\geq 0} q_{\text{recon}}\,\mathrm{d}x\mathrm{d}y = \frac{2q_{\text{C}}}{9} + \frac{1}{576}(35q_{\text{NW}} + q_{\text{SW}} + 22q_{\text{W}}) \tag{105}$$

If the S edge is reconstructed as a hat function, then the reconstruction reads

$$\begin{aligned} q^{\text{W}}_{\text{recon}}\Big|_{y<0,x<0} = \Big\{ & a_0 = q_{\text{C}}, a_1 = -\frac{q_{\text{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\text{C}} - q_{\text{W}})}{\Delta x^2}, a_3 = 0, \\ & a_4 = -\frac{-3q_{\text{SW}} + 2q_{\text{W}}}{\Delta x\Delta y}, a_5 = -\frac{2(q_{\text{SW}} - 2q_{\text{W}})}{\Delta x^2\Delta y}, a_6 = -\frac{4q_{\text{C}}}{\Delta y^2}, \\ & a_7 = -\frac{2q_{\text{SW}}}{\Delta x\Delta y^2}, a_8 = \frac{4(4q_{\text{C}} - q_{\text{SW}})}{\Delta x^2\Delta y^2} \Big\} \end{aligned} \tag{106}$$

$$\begin{aligned} q^{\text{W}}_{\text{recon}}\Big|_{y<0,x\geq 0} = \Big\{ & a_0 = q_{\text{C}}, a_1 = -\frac{q_{\text{W}}}{\Delta x}, a_2 = -\frac{2(2q_{\text{C}} - q_{\text{W}})}{\Delta x^2}, a_3 = 0, a_4 = -\frac{q_{\text{NW}}}{\Delta x\Delta y}, \\ & a_5 = \frac{2q_{\text{NW}}}{\Delta x^2\Delta y}, a_6 = -\frac{4q_{\text{C}}}{\Delta y^2}, a_7 = -\frac{2(q_{\text{NW}} - 2q_{\text{W}})}{\Delta x\Delta y^2}, \\ & a_8 = \frac{4(4q_{\text{C}} + q_{\text{NW}} - 2q_{\text{W}})}{\Delta x^2\Delta y^2} \Big\} \end{aligned} \tag{107}$$

$$\frac{1}{\Delta x\Delta y}\int_{y<0} q_{\text{recon}}\,\mathrm{d}x\mathrm{d}y = \frac{2q_{\text{C}}}{9} + \frac{1}{576}(q_{\text{NW}} + 35q_{\text{SW}} + 22q_{\text{W}}) \tag{108}$$

Proof of Continuity

Theorem 6 *The reconstructions in Sect. A.1.2 are continuous along $x = 0$ and along $y = 0$.*

Proof In complete analogy to the proof of Theorem 5 one observes from the sketches of the interpolation problem in Figs. 20 and 21 that the points along $x = 0$ and $y = 0$ are always included. The three points along $x = 0$ and the three points along $y = 0$ each define a unique parabola. □

A.2 Plateau-Limiting

Consider a situation in which (30) is true, while the reconstruction described above exceeds m or M. In that case, the idea of a plateau reconstruction (see Fig. 22) is to introduce a rectangle a distance $\eta\Delta x$ (or $\eta\Delta y$) away from the cell boundary, i.e.

$$\left[\Delta x\left(-\frac{1}{2}+\eta\right), \Delta x\left(\frac{1}{2}-\eta\right)\right] \times \left[\Delta y\left(-\frac{1}{2}+\eta\right), \Delta y\left(\frac{1}{2}-\eta\right)\right]$$

with $\eta \in (0, \frac{1}{2})$ where the value of the reconstruction will be constant and equal to q_{p}, a value to be determined to ensure that the average of the reconstruction equals the given average (see Fig. 2 for an example). This rectangle will be referred to as **plateau**. The remaining four trapezes will be the supports of functions that continuously join the reconstruction along the edge to the plateau in the simplest possible way. Because reconstructions along edges are either parabolas or hats, every trapezoidal region is either joining the plateau to a parabola or to a hat function. η will be chosen in such a way that the maximum principle is guaranteed. It is clear that, as (30) is true, this can always be done by choosing η small enough.

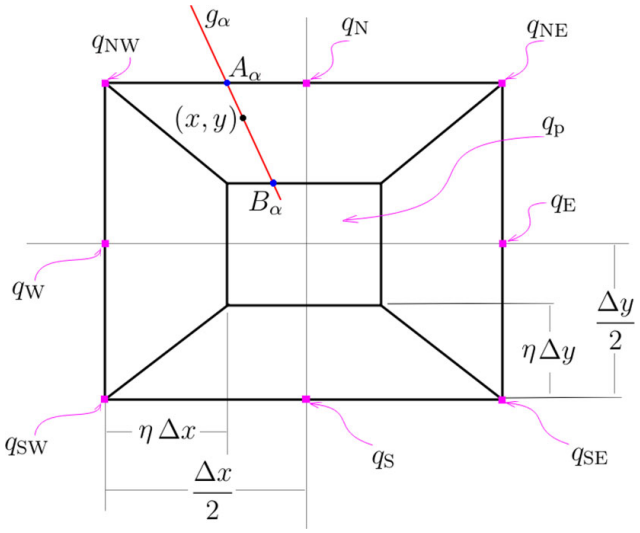

Fig. 22 Sketch of the interpolation between the plateau and the boundary of the cell

A.2.1 Interpolation in the trapezes

Consider for definiteness the northern trapeze. Define a point $A_\alpha := \left(-\frac{\Delta x}{2} + \alpha \Delta x, \frac{\Delta y}{2}\right) \in \mathbb{R}^2$ parametrized by $\alpha \in [0, 1]$. Define a point

$$B_\alpha := \left(\Delta x(-\frac{1}{2} + \eta) + \alpha \Delta x(1 - 2\eta), \Delta y(\frac{1}{2} - \eta)\right)$$

on the northern edge of the plateau. Observe that as α goes from 0 to 1, both points move all the way from the left to the right on their respective edges. The straight line

$$g_\alpha := \left\{(x, y) : \frac{x}{\Delta x} = -\frac{1}{2} + \alpha - \left(\frac{y}{\Delta y} - \frac{1}{2}\right)(1 - 2\alpha)\right\} \tag{109}$$

connects them. Obviously, given x and y there is a unique

$$\alpha = \frac{\frac{x}{\Delta x} + \frac{y}{\Delta y}}{2\frac{y}{\Delta y}} = \frac{x\Delta y + y\Delta x}{2y\Delta x} \tag{110}$$

The idea of the reconstruction is to associate to a point (x, y) the value given by a linear interpolation between the value of the reconstruction at A_α and the (constant) value q_{p} at B_α. In particular this means that the diagonal edges of the reconstruction (connections between the corners of the cell and the corners of the plateau) are reconstructed as straight lines.

The four trapezes can be reconstructed individually, because continuity along the diagonal segments where they join is already guaranteed by the above procedure. For a given trapeze, the choice of reconstruction thus merely depends on whether the adjacent edge is reconstructed parabolically (see Sect. A.2.2) or as a hat function (see Sect. A.2.3).

A.2.2 Parabolic Reconstruction Along the Edge

The parabolic reconstruction along the N-edge is given by

$$q^{\mathrm{N}}_{\mathrm{parabolic}}(x) = q_{\mathrm{N}} + \frac{x}{\Delta x}(q_{\mathrm{NE}} - q_{\mathrm{NW}}) + 2\frac{x^2}{\Delta x^2}(q_{\mathrm{NE}} + q_{\mathrm{NW}} - 2q_{\mathrm{N}}) \qquad x \in \left[-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right] \tag{111}$$

The value of this parabolic reconstruction is sought at the location ξ of point A_α with α given by (110):

$$\xi = \Delta x \left(-\frac{1}{2} + \frac{\frac{x}{\Delta x}\Delta y + y}{2y} \right) = \Delta x \frac{\frac{x}{\Delta x}}{2\frac{y}{\Delta y}} \tag{112}$$

Finally, the reconstruction at (x, y) is assigned the value

$$q^{\mathrm{N}}_{\mathrm{recon}}(x, y) := q^{\mathrm{N}}_{\mathrm{parabolic}}(\xi) + \left(y - \frac{\Delta y}{2} \right) \frac{q_{\mathrm{P}} - q^{\mathrm{N}}_{\mathrm{parabolic}}(\xi)}{-\Delta y \eta} \tag{113}$$

$$= q^{\mathrm{N}}_{\mathrm{parabolic}}(\xi) \left(1 + \frac{\frac{y}{\Delta y} - \frac{1}{2}}{\eta} \right) - \frac{\frac{y}{\Delta y} - \frac{1}{2}}{\eta} q_{\mathrm{P}} \tag{114}$$

with

$$q^{\mathrm{N}}_{\mathrm{parabolic}}(\xi) = q_{\mathrm{N}} + \frac{\hat{x}}{2\hat{y}}(q_{\mathrm{NE}} - q_{\mathrm{NW}}) + 2\left(\frac{\hat{x}}{2\hat{y}} \right)^2 (q_{\mathrm{NE}} + q_{\mathrm{NW}} - 2q_{\mathrm{N}}) \tag{115}$$

and $\hat{x} := \frac{x}{\Delta x}$ and $\hat{y} := \frac{y}{\Delta y}$. Observe that the reconstruction is not polynomial, but lies in

$$\mathrm{span}\left(1, \hat{x}, \hat{y}, \frac{\hat{x}}{\hat{y}}, \frac{\hat{x}^2}{\hat{y}}, \frac{\hat{x}^2}{\hat{y}^2} \right) \tag{116}$$

For reference we give the four reconstructions:

$$q^{\mathrm{trapeze,W}}_{\mathrm{recon}}(x, y) = q_{\mathrm{P}} \frac{1 + 2\hat{x}}{2\eta} + (-1 + 2\eta - 2\hat{x}) \left(\frac{q_{\mathrm{W}}}{2\eta} - \frac{(q_{\mathrm{NW}} - q_{\mathrm{SW}})y}{4\eta\hat{x}} + \frac{(q_{\mathrm{NW}} + q_{\mathrm{SW}} - 2q_{\mathrm{W}})\hat{y}^2}{4\eta\hat{x}^2} \right) \tag{117}$$

$$q^{\mathrm{trapeze,E}}_{\mathrm{recon}}(x, y) = q_{\mathrm{P}} \frac{1 - 2\hat{x}}{2\eta} + (-1 + 2\eta + 2\hat{x}) \left(\frac{q_{\mathrm{E}}}{2\eta} + \frac{(q_{\mathrm{NE}} - q_{\mathrm{SE}})\hat{y}}{4\eta\hat{x}} - \frac{(2q_{\mathrm{E}} - q_{\mathrm{NE}} - q_{\mathrm{SE}})\hat{y}^2}{4\eta\hat{x}^2} \right) \tag{118}$$

$$q^{\mathrm{trapeze,N}}_{\mathrm{recon}}(x, y) = q_{\mathrm{P}} \frac{1 - 2\hat{y}}{2\eta} + (-1 + 2\eta + 2\hat{y}) \left(-\frac{(2q_{\mathrm{N}} - q_{\mathrm{NE}} - q_{\mathrm{NW}})\hat{x}^2}{4\eta\hat{y}^2} + \frac{(q_{\mathrm{NE}} - q_{\mathrm{NW}})\hat{x}}{4\eta\hat{y}} + \frac{q_{\mathrm{N}}}{2\eta} \right) \tag{119}$$

$$q^{\mathrm{trapeze,S}}_{\mathrm{recon}}(x, y) = q_{\mathrm{P}} \frac{1 + 2\hat{y}}{2\eta} + (1 - 2\eta + 2\hat{y}) \left(\frac{(2q_{\mathrm{S}} - q_{\mathrm{SE}} - q_{\mathrm{SW}})\hat{x}^2}{4\eta\hat{y}^2} + \frac{(q_{\mathrm{SE}} - q_{\mathrm{SW}})\hat{x}}{4\eta\hat{y}} - \frac{q_{\mathrm{S}}}{2\eta} \right) \tag{120}$$

The integrals over the four regions are

$$\frac{1}{\Delta x \Delta y} \int_{\mathrm{trapeze,W}} q_{\mathrm{recon}} \, \mathrm{d}x\mathrm{d}y = \frac{1}{36}\eta \Big(6(3 - 4\eta)q_{\mathrm{P}} - (2\eta - 3)(4q_{\mathrm{W}} + q_{\mathrm{NW}} + q_{\mathrm{SW}}) \Big) \tag{121}$$

$$\frac{1}{\Delta x \Delta y} \int_{\mathrm{trapeze,E}} q_{\mathrm{recon}} \, \mathrm{d}x\mathrm{d}y = \frac{1}{36}\eta \Big(6(3 - 4\eta)q_{\mathrm{P}} - (2\eta - 3)(4q_{\mathrm{E}} + q_{\mathrm{NE}} + q_{\mathrm{SE}}) \Big) \tag{122}$$

$$\frac{1}{\Delta x \Delta y} \int_{\mathrm{trapeze,N}} q_{\mathrm{recon}} \, \mathrm{d}x\mathrm{d}y = \frac{1}{36}\eta \Big(6(3 - 4\eta)q_{\mathrm{P}} - (2\eta - 3)(4q_{\mathrm{N}} + q_{\mathrm{NE}} + q_{\mathrm{NW}}) \Big) \tag{123}$$

$$\frac{1}{\Delta x \Delta y} \int_{\mathrm{trapeze,S}} q_{\mathrm{recon}} \, \mathrm{d}x\mathrm{d}y = \frac{1}{36}\eta \Big(6(3 - 4\eta)q_{\mathrm{P}} - (2\eta - 3)(4q_{\mathrm{S}} + q_{\mathrm{SE}} + q_{\mathrm{SW}}) \Big) \tag{124}$$

and the integral over the plateau obviously

$$\frac{1}{\Delta x \Delta y} \int_{\mathrm{plateau}} q_{\mathrm{recon}} \, \mathrm{d}x\mathrm{d}y = (1 - 2\eta)^2 q_{\mathrm{P}} \tag{125}$$

A.2.3 Hat-Function Reconstruction Along the Edge

If an edge is reconstructed using a hat-function, then the reconstruction of the trapeze follows the algorithm outlined at the beginning of Sect. A.2, but is naturally defined in a piecewise fashion. The reconstruction of the W-trapeze is

$$q_{\text{recon}}^{\text{trapeze,W}}(x,y)\Big|_{y\geq 0} = q_{\text{W}} - \frac{\Delta x(q_{\text{NW}}-q_{\text{W}})y}{x\Delta y} + \frac{\left(\frac{\Delta x}{2}+x\right)\left(q_{\text{P}}-q_{\text{W}}+\frac{\Delta x(q_{\text{NW}}-q_{\text{W}})y}{x\Delta y}\right)}{\Delta x\eta} \tag{126}$$

$$q_{\text{recon}}^{\text{trapeze,W}}(x,y)\Big|_{y<0} = q_{\text{W}} + \frac{\Delta x(q_{\text{SW}}-q_{\text{W}})y}{x\Delta y} + \frac{\left(\frac{\Delta x}{2}+x\right)\left(q_{\text{P}}-q_{\text{W}}-\frac{\Delta x(q_{\text{SW}}-q_{\text{W}})y}{x\Delta y}\right)}{\Delta x\eta} \tag{127}$$

$$\frac{1}{\Delta x\Delta y}\int_{\text{trapeze,W}} q_{\text{recon}}\,\mathrm{d}x\mathrm{d}y = \frac{1}{6}(3-4\eta)\eta q_{\text{p}} + \frac{1}{24}\eta(2\eta-3)(q_{\text{NW}}+q_{\text{SW}}+2q_{\text{W}}) \tag{128}$$

The reconstructions of the other trapezes can be obtained by rotation as in Equations (37)–(39).

A.2.4 Choice of the Plateau Value and the Maximum Principle

Theorem 7 *There exists a choice of η such that the reconstruction is conservative and $m \leq q_{recon}(x,y) \leq M$ for all x, y inside the cell.*

Proof For any choice of $q_{\text{p}} \in (m, M)$, the reconstruction inside the cell fulfills $m \leq q_{\text{recon}} \leq M$, because the reconstructions inside the trapezes are interpolations along straight lines between q_{p} and a maximum-preserving reconstruction along the edge. For the same reason, as $\eta \to 0$, the average of the reconstruction over the cell approaches q_{p}, because the reconstructions inside the trapezes remain bounded and their contribution to the cell average thus vanishes in the limit. Thus, for all $\epsilon > 0$ sufficiently small one can find an $\eta > 0$ such that $\frac{1}{\Delta x\Delta y}\int_c q_{\text{recon}}(x,y)\,\mathrm{d}x\mathrm{d}y = q_{\text{p}} + a$ with $|a| < \epsilon$. Then, choosing $q_{\text{p}} := \bar{q} - a$ ensures conservativity of the reconstruction. At the same time, as $m < \bar{q} < M$, one simply needs to choose $\epsilon < \min(M-\bar{q}, \bar{q}-m)$ to ensure that $m < q_{\text{p}} < M$. □

For example, if all edges are reconstructed parabolically, then the average of the reconstruction over the entire cell is

$$q_{\text{p}} - \frac{1}{9}\eta(2\eta-3)\Big(4E - 6q_{\text{p}} + 2V\Big) \stackrel{!}{=} \bar{q} \tag{129}$$

(where $4V := q_{\text{NE}} + q_{\text{NW}} + q_{\text{SE}} + q_{\text{SW}}$, $4E := q_{\text{E}} + q_{\text{N}} + q_{\text{S}} + q_{\text{W}}$) which gives the value of q_{p}:

$$q_{\text{p}} = \frac{9\bar{q} + \eta(2\eta-3)(4E+2V)}{3(3-6\eta+4\eta^2)} \tag{130}$$

The polynomial in the denominator does not have real zeros.

What thus remains is the choice of η. The only bounds on η originate from the condition

$$m < q_{\text{p}} < M \tag{131}$$

The equation $q_{\text{p}} = \mu \in \{m, M\}$ is quadratic in η – and this is true in general and not just in this example. It is therefore easy to identify real, positive solutions and to take their minimum.

In practice, having established a minimum, η is chosen to be half of it. In case no real, positive solutions are identified, η is not subject to any conditions and we choose $\eta = \frac{1}{4}$.

Author Contributions All authors contributed equally to the study conception and design. All authors read and approved the final manuscript.

Funding The authors declare that no grants were received during the preparation of this manuscript.

Data Availability This work has no associated data.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Abgrall, R., Barsukow, W.: Extensions of active flux to arbitrary order of accuracy. ESAIM: Math. Modell. Numer. Anal. **57**(2), 991–1027 (2023)
2. Roe, P.L., Maeng, J., and Fan,D.: Comparing active flux and discontinuous Galerkin methods for compressble flow. In 2018 AIAA aerospace sciences meeting, pp. 0836, (2018)
3. van Leer, B.: Towards the ultimate conservative difference scheme IV. A new approach to numerical convection. J. comput. phys. **23**(3), 276–299 (1977)
4. Helzel, C., Kerkmann, D., Scandurra, L.: A new ADER method inspired by the active flux method. J. Sci. Comput. **80**(3), 1463–1497 (2019)
5. Barsukow, W.: The active flux scheme for nonlinear problems. J. Sci. Comput. **86**(1), 1–34 (2021)
6. Chudzik, E., Helzel, C., Kerkmann, D.: The cartesian grid active flux method: linear stability and bound preserving limiting. Appl. Math. Comput. **393**, 125501 (2021)
7. Barsukow, W., Berberich, J.P.: A well-balanced Active Flux method for the shallow water equations with wetting and drying. Commun. Appl. Math. Comput, pp. 1–46, (2023)
8. Eymann, T.A., Roe, P.L.: Multidimensional active flux schemes. In 21st AIAA computational fluid dynamics conference, (2013)
9. Fan, D., Roe, P.L.: Investigations of a new scheme for wave propagation. In 22nd AIAA computational fluid dynamics conference, pp. 2449, (2015)
10. Barsukow, W., Klingenberg, C.: Exact solution and a truly multidimensional Godunov scheme for the acoustic equations. ESAIM: M2AN, 56(1), (2022)
11. Barsukow, W., Hohm, J., Klingenberg, C., Roe, P.L.: The active flux scheme on Cartesian grids and its low Mach number limit. J. Sci. Comput. **81**(1), 594–622 (2019)
12. Chudzik, E., Helzel, C.: A review of cartesian grid active flux methods for hyperbolic conservation laws. In international conference on finite volumes for complex applications, pp. 93–109. Springer, (2023)
13. Maeng, J.: On the advective component of active flux schemes for nonlinear hyperbolic conservation laws. PhD thesis, University of Michigan, Dissertation, (2017)
14. Zeng, X.: A high-order hybrid finite difference-finite volume approach with application to inviscid compressible flow problems: a preliminary study. Comput. & Fluids **98**, 91–110 (2014)
15. Abgrall, R.: 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. Commun. Appl. Math. Comput., pp. 1–33, (2022)
16. Roe, P.: Designing CFD methods for bandwidth-a physical approach. Comput. & Fluids **214**, 104774 (2021)
17. Abgrall, R.: A hybrid finite element-finite volume method for conservation laws. Appl. Math. Comput. **447**, 127846 (2023)
18. Zeng, X.: Linear hybrid-variable methods for advection equations. Adv. Comput. Math. **45**(2), 929–980 (2019)
19. Roe, P.L., Lung, T., and Maeng, J.: New approaches to limiting. In 22nd AIAA computational fluid dynamics conference, pp. 2913, (2015)
20. Changqing, H., Shu, C.-W.: Weighted essentially non-oscillatory schemes on triangular meshes. J. Comput. Phys. **150**(1), 97–127 (1999)

21. Lax, P.D., Liu, X.-D.: Solution of two-dimensional riemann problems of gas dynamics by positive schemes. SIAM J. Sci. Comput. **19**(2), 319–340 (1998)
22. Barsukow, W., Edelmann, P.V.F., Klingenberg, C., Miczek, F., Röpke, F.K.: A numerical scheme for the compressible low-Mach number regime of ideal fluid dynamics. J. Sci. Comput. **72**(2), 623–646 (2017)
23. Gresho, P.M., Chan, S.T.: On the theory of semi-implicit projection methods for viscous incompressible flow and its implementation via a finite element method that also introduces a nearly consistent mass matrix. Part 2: implementation. Int. J. Numer. Methods Fluids **11**(5), 621–659 (1990)
24. Barsukow, W.: Truly multi-dimensional all-speed schemes for the Euler equations on cartesian grids. J. Comput. Phys. **435**, 110216 (2021)
25. Munz, C.-D., Roller, S., Klein, R., Geratz, K.J.: The extension of incompressible flow solvers to the weakly compressible regime. Comput. & Fluids **32**(2), 173–196 (2003)
26. Peraire, J., Persson, P.-O.: High-order discontinuous Galerkin methods for CFD. In Adaptive high-order methods in computational fluid dynamics, pp. 119–152. World Scientific, (2011)