**No. 609** **June 2019**

# Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws

**D. Kuzmin**

**ISSN: 2190-1767**

# Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws

Dmitri Kuzmin[a]

[a] *Institute of Applied Mathematics (LS III), TU Dortmund University, Vogelpothsweg 87, D-44227 Dortmund, Germany*

**Abstract**

Using the theoretical framework of algebraic flux correction and invariant domain preserving schemes, we introduce a monolithic approach to convex limiting in continuous finite element schemes for linear advection equations, nonlinear scalar conservation laws, and hyperbolic systems. In contrast to flux-corrected transport (FCT) algorithms that apply limited antidiffusive corrections to bound-preserving low-order solutions, our new limiting strategy exploits the fact that these solutions can be expressed as convex combinations of *bar states* belonging to a convex invariant set of physically admissible solutions. Each antidiffusive flux is limited in a way which guarantees that the associated bar state remains in the invariant set and preserves appropriate local bounds. There is no free parameter and no need for limit fluxes associated with the consistent mass matrix of time derivative term separately. Moreover, the steady-state limit of the nonlinear discrete problem is well defined and independent of the pseudo-time step. In the case study for the Euler equations, the components of the bar states are constrained sequentially to satisfy local maximum principles for the density, velocity, and specific total energy in addition to positivity preservation for the density and pressure. The results of numerical experiments for standard test problems illustrate the ability of built-in convex limiters to resolve steep fronts in a sharp and nonoscillatory manner.

*Keywords:* hyperbolic conservation laws, positivity preservation, invariant domains, finite elements, algebraic flux correction, convex limiting

## 1. Introduction

The discretization of hyperbolic conservation laws using continuous finite elements requires the use of advanced shock-capturing techniques based on (generalized) discrete maximum principles. If the exact solution of an initial value problem is known to lie in a convex invariant set, numerical approximations should be constrained to stay in this set as well. Discretizations that provide

*Email address:* kuzmin@math.uni-dortmund.de (Dmitri Kuzmin)

this property are called *invariant domain preserving* [29, 25, 28]. Additionally, global bounds depending on the boundary conditions and local bounds depending on the solution values at neighboring nodes / previous time steps may need to be enforced to avoid numerical instabilities and spurious oscillations. Since both global and local maximum principles can be formulated using equivalent nonnegativity constraints, numerical schemes satisfying the above requirements are often called *positivity preserving* or simply *positive*.

Recent years have witnessed an increased interest of the finite element community in the analysis of existing and design of new bound-preserving schemes for convection-dominated transport problems. New algebraic approaches based on the *flux-corrected transport* (FCT) methodology [15, 46, 48, 70] and various generalizations of *total variation diminishing* (TVD) limiters [33, 34] were developed using the unified framework of *algebraic flux correction* (AFC) schemes [6, 10, 26, 40, 42, 54]. Moreover, a breakthrough was achieved in the field of rigorous theoretical analysis for nonlinear high-resolution AFC schemes based on continuous finite element approximations [9, 11, 26, 52].

Modern limiting techniques for continuous finite elements trace their origins to edge-based FEM developed in the 1980s and early 1990s [55, 59, 60, 62, 63, 64]. In view of the equivalence between linear finite elements and vertex-centered finite volumes [27, 65, 66], edge-based extensions of FCT and TVD-like schemes can be constructed in a fairly straightforward manner. The AFC formalism presented in [42, 43] provides algebraic interpretations of *local extremum diminishing* (LED) schemes [35, 36, 37] that use artificial diffusion operators and flux limiters to enforce discrete maximum principles. Localized element-based limiting procedures for scalar conservation laws and hyperbolic systems were proposed in [16, 18, 40, 44, 54]. The way in which the limiter is localized in element-based and edge-based FCT schemes of this kind forms the basis of what Guermond et al. [25, 27] named *convex limiting* in the context of second-order invariant domain preserving schemes for hyperbolic systems.

The theoretical studies of Barrenechea et al. [8, 9, 11] provided new insights into the properties of AFC schemes for steady convection-diffusion equations and stimulated the development of improved limiter functions [10, 40]. The corresponding theoretical results for steady and unsteady linear advection problems were recently obtained by Lohmann [52]. The design of edge-based invariant domain preserving (IDP) finite element schemes for time-dependent nonlinear conservation laws and hyperbolic systems was greatly advanced by the recent work of Guermond et al. [26, 27, 25, 28] who derived such schemes using a very general and abstract theoretical framework based on convexity arguments. The second-order versions of their IDP schemes are based on a predictor-corrector algorithm of FCT type. At the first stage, a low-order approximation is calculated using graph viscosity based on a provable upper bound for the *guaranteed maximum speed* (GMS). At the second stage, an antidiffusive correction is performed using edge-based convex limiting to maintain the IDP property.

As shown by Lohmann [52], multistep limiting procedures provide LED properties with respect to extended stencils, which may result in a lack of monotonicity. To avoid bounded phase errors within the range of IDP values, some

high-order stabilization may need to be included [26, 54]. If fractional-step algorithms are used for limiting purposes, pseudo-time stepping schemes do not converge to steady state solutions. Monolithic AFC schemes [42, 52] lead to well-posed nonlinear discrete problems but the accuracy and convergence behavior of constrained finite element approximations depends on the choice of the involved free parameters. Moreover, existing extensions of such schemes to hyperbolic systems [61] cannot be analyzed using the concept of invariant domains as long as this concept is restricted to initial value problems.

The convex limiting strategy proposed in this paper makes it possible to constrain the antidiffusive part of the continuous Galerkin discretization in a manner which guarantees IDP properties and does not inhibit convergence to steady state solutions. No free parameters are involved and local maximum principles hold with respect to compact stencils. Using the representation of the low-order method in terms of invariant domain preserving *bar states*, we use these states to define the IDP bounds for the limited antidiffusive fluxes. The resulting limiting procedure has the structure of a localized FCT algorithm in which the original bar state (rather than the nodal value of a low-order solution assembled from multiple bar states) is adjusted in the process of antidiffusive corrections. The resulting nonlinear discrete problem has a well-defined steady-state residual and exhibits the structure of a monolithic AFC discretization but the correction factors for the antidiffusive fluxes are determined using the bar states (rather than the nodal values) of the solution from the previous iteration or time step. In Sections 2 and 3, we explain the underlying design philosophy for the continuous $\mathbb{P}_1/\mathbb{Q}_1$ finite element discretization of a generic hyperbolic problem. In Sections 3 and 4, we discuss convex limiting for linear advection equations, nonlinear conservation laws, and hyperbolic systems. In the last two sections, we perform numerical studies and draw preliminary conclusions.

## 2. Low-order invariant domain preserving schemes

Let $u(\mathbf{x},t) \in \mathbb{R}^m$ denote the local density of $m \in \mathbb{N}$ conserved quantities at the space location $\mathbf{x} \in \mathbb{R}^d, d \in \{1,2,3\}$ and time $t \geq 0$. In many models of computational fluid dynamics, the evolution of such quantities is governed by linear or nonlinear hyperbolic conservation laws which can be written as

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = 0 \quad \text{in } \Omega \times \mathbb{R}_+, \tag{1}$$

where $\Omega \subseteq \mathbb{R}^d$ is the domain of interest, $\mathbf{f}(u) \in \mathbb{R}^{m\times d}$ is an array of inviscid fluxes and $(\nabla\cdot\mathbf{f})_k = \sum_{l=1}^d \frac{\partial f_{kl}}{\partial x_l}$, $k = 1,\ldots,m$. A set $\mathcal{G}$ containing the initial data

$$u(\cdot,0) = u_0 \qquad \text{in } \Omega \tag{2}$$

is called an invariant set if $u(\mathbf{x},t) \in \mathcal{G}$ for all $\mathbf{x} \in \Omega$ and $t > 0$. A formal definition of invariant sets and invariant domains for hyperbolic initial value problems in unbounded domains can be found in [25, 28, 29].

If the domain $\Omega$ is bounded with a Lipschitz boundary $\Gamma = \partial\Omega$ and unit outward normal $\mathbf{n}$, a natural boundary condition of the form

$$\mathbf{f}(u) \cdot \mathbf{n} = \sum_{l=1}^{d} f_{kl}(u) n_l = f_n(u, \hat{u}) \in \mathbb{R}^m \tag{3}$$

can be formulated using the solution of the one-dimensional Riemann problem

$$\frac{\partial u}{\partial t} + (\mathbf{n} \cdot \nabla)(\mathbf{f}(u) \cdot \mathbf{n}) = 0, \qquad u(x,0) = \begin{cases} u_L & \text{for } x < 0, \\ u_R & \text{for } x > 0 \end{cases} \tag{4}$$

with the internal state $u_L = u$ and an external state $u_R = \hat{u} \in \mathcal{G}$. Invariant sets of initial-boundary value problems depend on the boundary conditions (see Section 4.1 for a definition of $\mathcal{G}$ based on a global maximum principle). Admissible sets for stationary hyperbolic problems can be defined in terms of the inflow boundary data (incoming Riemann invariants for systems).

Integrating the weighted residuals of (1) and (3) over $\Omega$ and $\Gamma$, respectively, we consider the following weak form of the boundary value problem:

$$\int_\Omega w \left( \frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) \right) \mathrm{d}\mathbf{x} = \int_\Gamma w \left( \mathbf{f}(u) \cdot \mathbf{n} - f_n(u, \hat{u}) \right) \mathrm{ds} \qquad \forall w \in W. \tag{5}$$

Let $W_h \subset W$ be the subspace of $W = L^2(\Omega)$ corresponding to a globally continuous approximation on a conforming mesh $\mathcal{T}_h = \{K^1, \dots, K^{E_h}\}$ using linear ($\mathbb{P}_1$) or multilinear ($\mathbb{Q}_1$) finite elements. The approximate solution

$$u_h(\mathbf{x}, t) = \sum_{j=1}^{N_h} u_j(t) \varphi_j(\mathbf{x}), \qquad \mathbf{x} \in K, \quad K \in \mathcal{T}_h, \quad t \in [0, T] \tag{6}$$

is expressed in terms of basis functions $\varphi_1, \dots, \varphi_{N_h} \in W_h$ such that $\varphi_i(\mathbf{x}_j) = \delta_{ij}$ and $\varphi_i|_K \in \mathbb{P}(K)$ or $\varphi_i|_K \in \mathbb{Q}(K)$ on each element $K \in \mathcal{T}_h$.

Substituting (6) into the Galerkin weak form (5) and using $w \in \{\varphi_1, \dots, \varphi_{N_h}\}$, a system of differential-algebraic equations is obtained for $N_h$ discrete nodal states $u_j = (u_{j1}, \dots, u_{jm})$. For brevity, we will often call $\mathbf{x}_i,\ i = 1, \dots, N_h$ "node $i$" and $K^e,\ e = 1, \dots, E_h$ "element $e$" instead of saying that $i$ and $e$ are the numbers of $\mathbf{x}_i$ and $K^e$, respectively. Let $\mathcal{E}_i$ be the set of elements containing node $i$ and $\mathcal{N}_i$ be the set of nodes belonging to these elements. Using this stencil notation, the $m$ equations associated with node $i$ can be written as

$$\sum_{j \in \mathcal{N}_i} m_{ij} \frac{\mathrm{d}u_j}{\mathrm{d}t} + \sum_{j \in \mathcal{N}_i} u_j \sum_{e \in \mathcal{E}_i \cap \mathcal{E}_j} \int_{K^e} \varphi_i \mathbf{f}'(u_h) \cdot \nabla \varphi_j \, \mathrm{d}\mathbf{x} = b_i(u_h, \hat{u}), \tag{7}$$

where

$$m_{ij} = \sum_{e \in \mathcal{E}_i \cap \mathcal{E}_j} \int_{K^e} \varphi_i \varphi_j \, \mathrm{d}\mathbf{x} \tag{8}$$

is an entry of the consistent mass matrix. The sums of element contributions associated with the boundary integral over $\Gamma$ are stored in

$$b_i(u_h, \hat{u}) = \sum_{e \in \mathcal{E}_i} \int_{K^e \cap \Gamma} \varphi_i (\mathbf{f}(u) \cdot \mathbf{n} - f_n(u_h, \hat{u})) \, \mathrm{ds}. \tag{9}$$

In the case of a scalar conservation law ($m = 1$), the vector $\mathbf{f}'(u) \in \mathbb{R}^d$ is the characteristic velocity of wave propagation. In the case of a hyperbolic system ($m > 1$), the Jacobian $\mathbf{f}'(u)$ is an array of $m \times m$ matrices $\mathbf{f}_1'(u), \ldots, \mathbf{f}_d'(u)$. For any unit vector $\mathbf{n} = (n_1, \ldots, n_d)^\top$, the corresponding Jacobian matrix

$$\mathbf{f}'(u) \cdot \mathbf{n} := \sum_{l=1}^{d} \mathbf{f}_l'(u) n_l \in \mathbb{R}^{m \times m}$$

is diagonalizable with real eigenvalues $\lambda_1(\mathbf{n}, u), \ldots, \lambda_m(\mathbf{n}, u)$ representing the $m$ wave speeds, i.e., the projections of $m$ characteristic velocities onto $\mathbf{n}$.

The first-order invariant domain preserving schemes of Guermond and Popov [28] generalize the concept of *discrete upwinding* [42, 46] to hyperbolic systems. Their derivation is based on three conservative modifications of (7). First, the entries $m_{ij}$ of the consistent mass matrix are approximated by

$$\tilde{m}_{ij} = m_i \delta_{ij}, \qquad m_i = \sum_{j=1}^{N_h} m_{ij} = \sum_{e \in \mathcal{E}_i} \int_{K^e} \varphi_i \, \mathrm{d}\mathbf{x}. \tag{10}$$

This approximation is known as *row-sum mass lumping* and can be interpreted as calculation of $m_{ij}$ using low-order nodal quadrature. In a similar vein, let the boundary term $b_i(u_h, \hat{u})$ be replaced with the lumped approximation

$$\tilde{b}_i(u_h, \hat{u}) = (\mathbf{f}(u_i) \cdot \mathbf{n} - f_n(u_i, \hat{u}(\mathbf{x}_i))) \sum_{e \in \mathcal{E}_i} \int_{K^e \cap \Gamma} \varphi_i \, \mathrm{ds}. \tag{11}$$

Next, the Galerkin flux $\mathbf{f}(u_h)$ is approximated by the finite element interpolant

$$\mathbf{f}_h(u_h) = \sum_{j=1}^{N_h} \mathbf{f}_j \varphi_j, \qquad \mathbf{f}_j = \mathbf{f}(u_j). \tag{12}$$

Many edge-based FEM for compressible flow problems are based on this *group finite element formulation* [20, 21] which can also be interpreted as low-order nodal quadrature [12]. It greatly reduces the cost of matrix assembly without degrading the accuracy of the lumped Galerkin approximation.

Using the partition of unity property ($\sum_{i=1}^{N_h} \varphi_i \equiv 1$) of the basis functions $\varphi_i$, the resulting semi-discrete scheme can be written as [25, 43]

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} + \sum_{j \in \mathcal{N}_i^*} \mathbf{c}_{ij} \cdot (\mathbf{f}_j - \mathbf{f}_i) = \tilde{b}_i(u_h, \hat{u}), \tag{13}$$

where $\mathcal{N}_i^* = \mathcal{N}_i \backslash \{i\}$ is the set containing the nearest neighbors of node $i$. The vector-valued coefficents $\mathbf{c}_{ij}$ of the discrete gradient operator are defined by

$$\mathbf{c}_{ij} = \sum_{e \in \mathcal{E}_i \cap \mathcal{E}_j} \int_{K^e} \varphi_i \nabla \varphi_j \, \mathrm{d}\mathbf{x}. \tag{14}$$

The final modification of (7) is the addition of diffusive fluxes $d_{ij}(u_i - u_j)$ to the edge contributions $\mathbf{c}_{ij} \cdot (\mathbf{f}_j - \mathbf{f}_i)$ of the group finite element approximation. This modification preserves the discrete conservation property if $d_{ij} = d_{ji}$ for $i, j = 1, \ldots, N_h$. It leads to the system of ordinary differential equations

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \tilde{b}_i(u_h, \hat{u}) + \sum_{j \in \mathcal{N}_i^*} [d_{ij}(u_j - u_i) - \mathbf{c}_{ij} \cdot (\mathbf{f}_j - \mathbf{f}_i)]. \tag{15}$$

Let $i$ be an interior node of the finite element mesh. Then $\varphi_i = 0$ on $\Gamma$ and, therefore, $\tilde{b}_i(u_h, \hat{u}) = 0$. If time integration is performed using an explicit *strong stability preserving* (SSP) method [23], each stage is an update of the form

$$\bar{u}_i = u_i + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} \left[ d_{ij}(u_j - u_i) - \mathbf{c}_{ij} \cdot (\mathbf{f}_j - \mathbf{f}_i) \right], \tag{16}$$

where $\Delta t$ is the time increment. For sufficiently small $\Delta t$, the so-defined approximation is invariant domain preserving (IDP) w.r.t. $\mathcal{G}$ if the parameters $d_{ij}$ are chosen so that $\bar{u}_i \in \mathcal{G}$ whenever $u_j \in \mathcal{G}$ for all $j \in \mathcal{N}_i$.

To determine artificial viscosity coefficients $d_{ij}$ which provably provide the IDP property under CFL-like time step restrictions, Guermond et al. [25] considered an equivalent repesentation of (16) in terms of the *bar states*

$$\bar{u}_{ij} = \frac{u_j + u_i}{2} - \frac{\mathbf{c}_{ij} \cdot (\mathbf{f}_j - \mathbf{f}_i)}{2 d_{ij}}. \tag{17}$$

Note that $\bar{u}_{ij} = \bar{u}_{ji}$ for nodes $\mathbf{x}_i, \mathbf{x}_j \notin \Gamma$ since $\mathbf{c}_{ji} = -\mathbf{c}_{ij}$ if $\varphi_i = \varphi_j = 0$ on $\Gamma$.

The result $\bar{u}_i$ of the generic forward Euler update (16) can be written as

$$\bar{u}_i = \left( 1 - \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij} \right) u_i + \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij} \bar{u}_{ij}. \tag{18}$$

Let $\bar{u}(\mathbf{n}, u_L, u_R)$ denote the exact solution of the projected one-dimensional Riemann problem (4). As explained in [25], the bar state $\bar{u}_{ij}$ is a space average of $\bar{u}(\mathbf{n}_{ij}, u_i, u_j)$ for $\mathbf{n}_{ij} = \frac{\mathbf{c}_{ij}}{|\mathbf{c}_{ij}|}$ at the artificial time $\tau_{ij} = \frac{|\mathbf{c}_{ij}|}{2 d_{ij}}$ if

$$\lambda_{\max}(\mathbf{n}_{ij}, u_i, u_j) \leq \frac{1}{2\tau_{ij}} = \frac{d_{ij}}{|\mathbf{c}_{ij}|}, \tag{19}$$

where $\lambda_{\max}(\mathbf{n}, u_L, u_R) = \max_{\omega \in [0,1]} \operatorname{spr}(\mathbf{f}'(\omega u_L + (1-\omega) u_R) \cdot \mathbf{n})$ is the fastest wave speed. For time steps $\Delta t$ satisfying the CFL-like condition

$$\frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij} \leq 1, \tag{20}$$

the nodal state $\bar{u}_i$ defined by formula (18) is a convex combination of $u_i$ and $\bar{u}_{ij},\ j \in \mathcal{N}_i^*$. Hence, this approximation proves IDP for $d_{ij} \geq \lambda_{ij}^{\text{GMS}} |\mathbf{c}_{ij}|$, where $\lambda_{ij}^{\text{GMS}}$ is an upper bound for $\lambda_{\max}(\mathbf{n}_{ij}, u_i, u_j)$. The abbreviation GMS stands for *guaranteed maximum speed* [29, 25]. To define a consistent and conservative graph Laplacian operator, the GMS artificial viscosity coefficients

$$d_{ij} = \begin{cases} \max\{\lambda_{ij}^{\text{GMS}} |\mathbf{c}_{ij}|, \lambda_{ji}^{\text{GMS}} |\mathbf{c}_{ji}|\} & \text{if } j \in \mathcal{N}_i^*, \\ -\sum_{k \in \mathcal{N}_i^*} d_{ik} & \text{if } j = i, \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

are chosen to satisfy the symmetry condition $d_{ij} = d_{ji}$ for $i, j = 1, \ldots, N_h$ and the zero sum condition $\sum_{j=1}^{N_h} d_{ij} = 0$ for $i = 1, \ldots, N_h$ (cf. [42, 46]).

**Remark 1.** The above algebraic manipulations can also be performed using the entries of element matrices instead of the global matrix entries. In such localized algorithms [18, 44, 54], the element contributions $d_{ij}^e$ to $d_{ij}$ are defined in terms of the element contributions $\mathbf{c}_{ij}^e = \int_{K^e} \varphi_i \nabla \varphi_j \, \mathrm{d}\mathbf{x}$ to (14).

## 3. High-order invariant domain preserving schemes

By the Godunov theorem [22], the space discretization defined by (15) and (21) can be at most first-order accurate. In monolithic AFC schemes [42, 52], limited antidiffusive fluxes $f_{ij}^*$ are added to $d_{ij}(u_j - u_i)$ in order to compensate the mass lumping error and remove the artificial viscosity or replace it with high-order stabilization in smooth regions. The flux-corrected scheme

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \tilde{b}_i(u_h, \hat{u}) + \sum_{j \in \mathcal{N}_i^*} [d_{ij}(u_j - u_i) + f_{ij}^* - \mathbf{c}_{ij} \cdot (\mathbf{f}_j - \mathbf{f}_i)] \tag{22}$$

can be written in terms of the bar states defined by (17) as follows:

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \tilde{b}_i(u_h, \hat{u}) - \left( 2 \sum_{j \in \mathcal{N}_i^*} d_{ij} \right) u_i + \sum_{j \in \mathcal{N}_i^*} 2 d_{ij} \left( \bar{u}_{ij} + \frac{f_{ij}^*}{2 d_{ij}} \right). \tag{23}$$

The corresponding forward Euler step is IDP if the corrected bar states

$$\bar{u}_{ij}^* = \bar{u}_{ij} + \frac{f_{ij}^*}{2 d_{ij}} \tag{24}$$

stay in the convex invariant set $\mathcal{G}$. This is generally not the case for $f_{ij}^* = f_{ij}$, where $f_{ij}$ are the fluxes that transform (15) into a high-order *target scheme*. Hence, the fluxes $f_{ij}$ must be limited to produce IDP fluxes $f_{ij}^*$. In the remainder of this section, we discuss the definition of $f_{ij}$ and convex limiting techniques that can be used to enforce the IDP property in nonlinear AFC schemes.

*3.1. Definition of the raw antidiffusive fluxes*

It is easy to verify that the consistent-mass group Galerkin approximation

$$\sum_{j\in\mathcal{N}_i} m_{ij}\frac{\mathrm{d}u_j}{\mathrm{d}t} = \tilde{b}_i(u_h,\hat{u}) - \sum_{j\in\mathcal{N}_i^*} \mathbf{c}_{ij}\cdot(\mathbf{f}_j-\mathbf{f}_i) \tag{25}$$

can be recovered from (15) by adding the high-order antidiffusive fluxes

$$f_{ij}^H = m_{ij}\left(\dot{u}_i^H - \dot{u}_j^H\right) + d_{ij}(u_i-u_j). \tag{26}$$

In view of (25), the calculation of the time derivatives $\dot{u}_i^H = \frac{\mathrm{d}u_i}{\mathrm{d}t},\ i=1,\ldots,N_h$ requires (approximate) solution of the well-conditioned linear system

$$\sum_{j\in\mathcal{N}_i} m_{ij}\dot{u}_j^H = \tilde{b}_i(u_h,\hat{u}) - \sum_{j\in\mathcal{N}_i^*} \mathbf{c}_{ij}\cdot(\mathbf{f}_j-\mathbf{f}_i), \qquad i=1,\ldots,N_h. \tag{27}$$

The time derivatives $\dot{u}_i^L$ corresponding to the solution of (15) are given by

$$\dot{u}_i^L = \frac{1}{m_i}\left(\tilde{b}_i(u_h,\hat{u}) + \sum_{j\in\mathcal{N}_i^*}\left[d_{ij}(u_j-u_i) - \mathbf{c}_{ij}\cdot(\mathbf{f}_j-\mathbf{f}_i)\right]\right). \tag{28}$$

The replacement of $\dot{u}^H$ by $\dot{u}^L$ in (26) corresponds to using the target fluxes

$$f_{ij} = m_{ij}\left(\dot{u}_i^L - \dot{u}_j^L\right) + d_{ij}(u_i-u_j) = f_{ij}^H + f_{ij}^{\mathrm{S}}, \tag{29}$$

where the oscillatory Galerkin component $f_{ij}^H$ is stabilized by

$$f_{ij}^{\mathrm{S}} = m_{ij}(\dot{u}_i^L - \dot{u}_j^L) - m_{ij}(\dot{u}_i^H - \dot{u}_j^H). \tag{30}$$

Definition (29) of the target flux $f_{ij}$ was introduced in the context of AFC schemes [41, 42, 43]. The stabilization properties of $f_{ij}^{\mathrm{S}}$ were explored in numerical studies by Lohmann [52]. The stabilized target (29) produces well-defined approximations at steady state and is an efficient alternative to the use of dissipative fluxes $f_{ij}^S$ based on entropy viscosity [26, 25] and other kinds of high-order stabilization which may be required to achieve optimal convergence behavior [40, 54] and/or prevent convergence to entropy-violating weak solutions [26].

*3.2. Predictor-corrector limiting strategy*

The convex limiting procedure proposed by Guermond et al. [25, 27] belongs to the family of flux-corrected transport (FCT) algorithms. That is, it is based on a predictor-corrector strategy. The forward Euler update

$$\bar{u}_i = u_i + \frac{\Delta t}{m_i}\sum_{j\in\mathcal{N}_i^*}\left[d_{ij}(u_j-u_i) + f_{ij}^* - \mathbf{c}_{ij}\cdot(\mathbf{f}_j-\mathbf{f}_i)\right] \tag{31}$$

corresponding to an SSP Runge-Kutta time discretization of equation (22) for an internal node $i$ is split into the low-order prediction step

$$u_i^L = u_i + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} [d_{ij}(u_j - u_i) - \mathbf{c}_{ij} \cdot (\mathbf{f}_j - \mathbf{f}_i)] \tag{32}$$

and the high-order antidiffusive correction step

$$\bar{u}_i = u_i^L + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} f_{ij}^*. \tag{33}$$

The limited fluxes $f_{ij}^* = \alpha_{ij} f_{ij}$ are defined using the same scalar-valued correction factor $\alpha_{ij} = \alpha_{ji}$ for all components of $f_{ij} \in \mathbb{R}^m$. The proof of the IDP property for update (32) was reviewed in Section 2. The calculation of correction factors $\alpha_{ij} \in [0, 1]$ for update (32) is based on the representation

$$\bar{u}_i = \frac{1}{m_i} \sum_{j \in \mathcal{N}_i^*} \mu_{ij} u_{ij}^* \tag{34}$$

of the flux-corrected state as a convex combination of edge contributions

$$u_{ij}^* = u_i + \frac{\Delta t}{\mu_{ij}} f_{ij}^* = u_i + \frac{\Delta t}{\mu_{ij}} \alpha_{ij} f_{ij} \tag{35}$$

with nonnegative weights $\mu_{ij}$ such that $\sum_{j \in \mathcal{N}_i^*} \mu_{ij} = m_i$. The derivation of IDP limiters (i.e., of algorithms for calculating the correction factors $\alpha_{ij}$) is based on the fact that $\bar{u}_i$ is a convex combination of the edge states $u_{ij}^*$ which stay in the invariant set $\mathcal{G}$ if $u_i \in \mathcal{G}$ and $\alpha_{ij}$ are chosen sufficiently small.

An element-based *localized* FCT algorithm based on the above design philosophy was proposed in [16]. It can be interpreted as algebraic version of the Barth-Jespersen slope limiter [13] which is well suited for extensions to high-order finite elements and hyperbolic systems [18, 54, 44]. The first edge-based localized limiting procedures of this kind were introduced independently in [50] and [25]. The weights for (34) can be defined e.g., by the formula $\mu_{ij} = \frac{m_{ij} m_i}{m_i - m_{ii}}$, as proposed in Section 4.4.1.1 of [52]. Guermond et al. [25] used $\mu_{ij} = \frac{m_i}{|\mathcal{N}_i^*|}$, where $|\mathcal{N}_i^*| = \text{card}(\mathcal{N}_i^*)$ is the number of nearest neighbors of node $i$.

### *3.3. Monolithic limiting strategy*

As mentioned in the Introduction, FCT-like predictor-corrector approaches are generally not monotonicity preserving and cause convergence problems in steady state computations using pseudo-time stepping. To constrain the bar states without splitting, we propose a new limiting strategy which belongs to the family of monolithic AFC schemes. In contrast to other representatives of this family, it does not involve the use of free parameters and leads to very simple monolithic versions of convex limiting techniques designed for FCT.

Our representation of the unsplit semi-discrete problem (22) in the bar state form (23) reveals that the generic forward Euler step can be written as

$$\bar{u}_i = \left(1 - \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij}\right) u_i + \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij} \bar{u}_{ij}^* \tag{36}$$

with bar states $\bar{u}_{ij}^*$ defined in terms of the constrained fluxes $f_{ij}^*$ by (24). Note that the only difference between (36) and (18) is the use of $\bar{u}_{ij}^*$ instead of $\bar{u}_{ij} \in \mathcal{G}$. If the CFL number $\nu_i = \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i} d_{ij}$ satisfies condition (20) and the definition of $f_{ij}^*$ guarantees that $\bar{u}_{ij}^* \in \mathcal{G}$ for $\bar{u}_{ij} \in \mathcal{G}$ and, then the IDP property of the updated nodal state $\bar{u}_i$ follows by convexity. The trivial choice $f_{ij}^* = 0$ produces the bar state $\bar{u}_{ij}^* = \bar{u}_{ij}$ of the low-order IDP scheme. The choice $f_{ij}^* = f_{ij}$ produces the bar state of the unconstrained target scheme. Since neither of these approximations is generally satisfactory, the limited flux $f_{ij}^*$ should be chosen as close to $f_{ij}$ as possible without producing an unacceptable bar state $\bar{u}_{ij}^* \notin \mathcal{G}$ and/or violating additional limiting criteria to be defined below. The simplest way to constrain $f_{ij}^* \in \mathbb{R}^m$ in this manner is to multiply each component of $f_{ij}$ by the same correction factor $\alpha_{ij} \in [0, 1]$, as in *synchronized* FCT algorithms for hyperbolic systems [25, 48, 53]. However, the use of individually chosen correction factors for different quantities of interest [18, 44] or even inequality-constrained optimization approaches [14] to direct calculation of the fluxes $f_{ij}^*$ may be appropriate for some applications. Some practical algorithms for calculating $\alpha_{ij}$ and $f_{ij}^*$ are presented in Sections 4 and 5.

We emphasize that the analysis of (36) was used to show the IDP property in the simplest possible setting. In contrast to the FCT approach, which requires the use of explicit SSP time integrators and small time steps, the unsplit form (31) of our monolithic semi-discrete AFC scheme can be integrated in time implicitly, and the nonlinear discrete problem corresponding to the steady state limit ($\frac{\mathrm{d}u_i}{\mathrm{d}t} = 0,\ i = 1, \ldots, N_h$) can be solved using more efficient iterative methods than time marching with spatially constant $\Delta t$. Moreover, the monolithic formulation in terms of the bar states $\bar{u}_{ij}$ is amenable to theoretical analysis.

**Remark 2.** Using the weights $\mu_{ij}$ defined in Section 3.2, update (36) can also be written as $\bar{u}_i = \frac{1}{m_i} \sum_{j \in \mathcal{N}_i^*} \mu_{ij} \tilde{u}_{ij}$ with the edge states (cf. [5], p. 12)

$$\tilde{u}_{ij} = u_i + \frac{\Delta t}{\mu_{ij}} \left[\mathbf{c}_{ij} \cdot (\mathbf{f}_i + \mathbf{f}_j) + d_{ij}(u_j - u_i) + f_{ij}^*\right]. \tag{37}$$

This representation of $\bar{u}_i$ is also well suited for convex limiting but the reguirement that $\tilde{u}_{ij} \in \mathcal{G}$ for $u_{ij}^L = u_i + \frac{\Delta t}{\mu_{ij}} \left[\mathbf{c}_{ij} \cdot (\mathbf{f}_i + \mathbf{f}_j) + d_{ij}(u_j - u_i)\right] \in \mathcal{G}$ leads to FCT-like inequality constraints in which the bounds for $f_{ij}^*$ depend on the time step $\Delta t$. Convex limiters based on such constraints are not monolithic.

**Remark 3.** Similarly to element-based versions [18, 44, 54, 61] of low-order schemes for AFC, the convex limiting procedure can be further localized to edges of individual elements using the bar states $\bar{u}_{ij}^e$ and antidiffusive fluxes $f_{ij}^e$ defined in terms of the edge contributions $m_{ij}^e$ and $d_{ij}^e$ to $m_{ij}$ and $d_{ij}$.

## 4. Convex limiting for scalar conservation laws

The preservation of global bounds encoded into the definition of the convex invariant set $\mathcal{G}$ is generally insufficient to guarantee monotonicity preservation and entropy consistency [25]. As shown in [52] for the linear advection equation and symmetric tensor fields, the low-order scheme defined by (15) is not only globally positivity preserving but also local extremum diminishing (LED) under the additional assumption of an incompressible velocity field. In FCT algorithms for scalar conservation laws, the LED property of the antidiffusive correction step (33) is enforced by using limiters based on the inequality constraints

$$\min_{j\in\mathcal{N}_i} u_j^L =: u_i^{\min,L} \le \bar{u}_i \le u_i^{\max,L} := \max_{j\in\mathcal{N}_i} u_j^L. \tag{38}$$

This limiting criterion implies that $\bar{u}_i = \mathcal{G}$ if $u_j \in \mathcal{G}$ for all $j \in \mathcal{N}_i$. However, the investigations and counterexamples presented in [52] show that these FCT constraints do not guarantee the validity of the local maximum principle

$$\min_{j\in\mathcal{N}_i} u_j =: u_i^{\min} \le \bar{u}_i \le u_i^{\max} := \max_{j\in\mathcal{N}_i} u_j. \tag{39}$$

A sufficient condition for the state $\bar{u}_i$ defined by (36) to satisfy (39) is given by

$$u_i^{\min} \le \bar{u}_{ij}^* = \bar{u}_{ij} + \alpha_{ij}\frac{f_{ij}}{2d_{ij}} \le u_i^{\max} \qquad \forall j \in \mathcal{N}_i. \tag{40}$$

By definition (21) of the GMS diffusion coefficient $d_{ij}$, all scalar bar states $\bar{u}_{ij}$ defined by (17) are in the admissible range, i.e., $\bar{u}_{ij} \in [u_i^{\min}, u_i^{\max}]$ for $j \in \mathcal{N}_i^*$. To verify the validity of this local maximum principle, we notice that

$$\bar{u}_{ij} = \frac{u_j + u_i}{2} - \frac{\mathbf{c}_{ij}\cdot(\mathbf{f}_j - \mathbf{f}_i)}{2d_{ij}} = \frac{u_j + u_i}{2} - \left(\frac{\mathbf{c}_{ij}\cdot\mathbf{v}_{ij}}{d_{ij}}\right)\frac{u_j - u_i}{2}, \tag{41}$$

where

$$\mathbf{v}_{ij} = \begin{cases} \frac{\mathbf{f}_j - \mathbf{f}_i}{u_j - u_i} & \text{for } u_j \neq u_i, \\ \mathbf{f}'(\bar{u}_{ij}) & \text{for } u_j = u_j = \bar{u}_{ij} \end{cases} \tag{42}$$

is the Rankine-Hugoniot shock velocity of the Riemann problem with the initial states $u_i$ and $u_j$. By the mean value theorem, we have

$$\mathbf{c}_{ij}\cdot\mathbf{v}_{ij} = \mathbf{c}_{ij}\cdot\mathbf{f}'(\omega u_i + (1-\omega)u_j)$$

for some $\omega \in [0,1]$. The scalar version of the GMS formula (21) yields

$$d_{ij} = \max_{\omega\in[0,1]} \max\{|\mathbf{c}_{ij}\cdot\mathbf{f}'(\omega u_i + (1-\omega)u_j)|, |\mathbf{c}_{ji}\cdot\mathbf{f}'(\omega u_i + (1-\omega)u_j)|\}. \tag{43}$$

It follows that the low-order bar state $\bar{u}_{ij}$ is a convex average of the nodal states $u_i \in [u_i^{\min}, u_i^{\max}]$ and $u_j \in [u_i^{\min}, u_i^{\max}]$. In general, the existence of a correction factor $\alpha_{ij} \in [0,1]$ satisfying (40) is guaranteed if $\bar{u}_{ij} \in [u_i^{\min}, u_i^{\max}]$.

The conservation property is preserved for $\alpha_{ji} = \alpha_{ij}$. Hence, an additional requirement for the choice of $\alpha_{ij}$ is given by the inequality constraints

$$u_j^{\min} \le \bar{u}_{ji}^* = \bar{u}_{ji} - \alpha_{ij}\frac{f_{ij}}{2d_{ij}} \le u_j^{\max} \tag{44}$$

for the bar state $\bar{u}_{ji}^*$, whose low-order counterpart $\bar{u}_{ji}$ coincides with $\bar{u}_{ij}$ if $i$ and $j$ are internal nodes. In view of these considerations, we define $\alpha_{ij}$ as follows:

$$\alpha_{ij} = \begin{cases} \min\left\{1, \min\left\{\frac{2d_{ij}(u_i^{\max}-\bar{u}_{ij})}{f_{ij}}, \frac{2d_{ij}(\bar{u}_{ji}-u_j^{\min})}{f_{ij}}\right\}\right\} & \text{if } f_{ij} > 0, \\ \min\left\{1, \min\left\{\frac{2d_{ij}(u_i^{\min}-\bar{u}_{ij})}{f_{ij}}, \frac{2d_{ij}(\bar{u}_{ji}-u_j^{\max})}{f_{ij}}\right\}\right\} & \text{if } f_{ij} < 0. \end{cases} \tag{45}$$

In the case $f_{ij} = 0$, the correction factor $\alpha_{ij}$ may be chosen arbitrarily.

Note that the CFL condition (20) under which the IDP property is guaranteed for bound-preserving bar states $\bar{u}_{ij}^* \in \mathcal{G}$ is independent of $\alpha_{ij}$. This a quite remarkable fact since time step restrictions of other monolithic AFC schemes exhibit strong dependence on the design of limiter functions [52]. In steady state computations, this dependence has an adverse effect on the convergence behavior of iterative solvers for the nonlinear discrete problem.

Some remarks regarding the practical calculation of $f_{ij}^*$ are in order. First, the division by $d_{ij}$ in the formulas for $\bar{u}_{ij}$ and $\bar{u}_{ij}^*$ may give rise to large rounding errors. Second, the limited flux $f_{ij}^* = \alpha_{ij} f_{ij}$ is a continuous function of the nodal values $u_1, \ldots, u_{N_h}$. We calculate it directly using the formula

$$f_{ij}^* = \begin{cases} \min\left\{f_{ij}, \min\left\{2d_{ij}u_i^{\max} - \bar{w}_{ij}, \bar{w}_{ji} - 2d_{ij}u_j^{\min}\right\}\right\} & \text{if } f_{ij} > 0, \\ \max\left\{f_{ij}, \max\{2d_{ij}u_i^{\min} - \bar{w}_{ij}, \bar{w}_{ji} - 2d_{ij}u_j^{\max}\right\}\right\} & \text{otherwise,} \end{cases} \tag{46}$$

where $\bar{w}_{ij} = 2d_{ij}\frac{u_j+u_i}{2} - \mathbf{c}_{ij}\cdot(\mathbf{f}_j - \mathbf{f}_i)$ is a numerically stable representation of $2d_{ij}\bar{u}_{ij}$ for the low-order bar state $\bar{u}_{ij}$ defined by (17). The limited bar states $\bar{u}_{ij}^*$ are never calculated in practice. Following the derivation of the low-order scheme in Section 2, we introduced the bar state form (36) just to prove that definitions (45) and (46) ensure the validity of (39). In a practical implementation, the fluxes $f_{ij}^*$ should be inserted into the right-hand side of (31).

In general, the residual of the nonlinear system associated with a fully discrete version of (31) can be assembled in the above manner without calculating the bar states $\bar{u}_{ij}^*$. The antidiffusive flux corresponding to the steady-state limit of (26) is given by $f_{ij} = d_{ij}(u_i - u_j)$. Substituting it into (46), we obtain

$$f_{ij}^* = \begin{cases} d_{ij}\min\left\{u_i - u_j, 2\min\left\{u_i^{\max} - \bar{u}_{ij}, \bar{u}_{ji} - u_j^{\min}\right\}\right\} & \text{if } u_i > u_j, \\ d_{ij}\max\left\{u_i - u_j, 2\max\{u_i^{\min} - \bar{u}_{ij}, \bar{u}_{ji} - u_j^{\max}\right\}\right\} & \text{otherwise.} \end{cases} \tag{47}$$

Importantly, the continuous dependence of the limited fluxes $f_{ij}^*$ on the degrees of freedom implies existence of a solution to the nonlinear discrete problem by Brouwer's fixed-point theorem, see [9, 11, 52] for detailed analysis.

*4.1. Case study: linear advection*

To illustrate the presented ideas in a simple setting, we consider a hyperbolic conservation law of the form (1) with the linear flux function $\mathbf{f}(u) = \mathbf{v}u$, where $\mathbf{v} \in \mathbb{R}^d$ is a constant vector and $u$ is a scalar. In this linear advection model, the flux boundary condition (3) should be defined using the upwind state

$$\hat{u} = \begin{cases} u_{\text{in}} & \text{on } \Gamma_-, \\ u & \text{on } \Gamma \backslash \Gamma_-, \end{cases} \tag{48}$$

where $\Gamma_- = \{\mathbf{x} \in \Gamma : \mathbf{v} \cdot \mathbf{n}(\mathbf{x}) < 0\}$ is the inflow boundary of $\Omega$ and $u_{\text{in}}$ is a given data. Substituting the so-defined $\mathbf{f}(u)$ and $\hat{u}$ into (5), we obtain

$$\int_\Omega w \left( \frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) \right) \mathrm{d}\mathbf{x} = \int_{\Gamma_-} w(u - u_{\text{in}}) \mathbf{v} \cdot \mathbf{n} \, \mathrm{ds} \qquad \forall w \in W. \tag{49}$$

Since the exact solution of the linear advection equation is constant along the characteristics, the initial-boundary value problem has the invariant set

$$\mathcal{G} = \left\{ u \in \mathbb{R} : u^{\min} \le u \le u^{\max} \right\}, \tag{50}$$

where $u^{\min}$ and $u^{\max}$ are the global bounds defined by

$$u^{\min} = \min \left\{ \min_{\Omega} u_0, \min_{\Gamma_-} u_{\text{in}} \right\}, \tag{51}$$

$$u^{\max} = \max \left\{ \max_{\Omega} u_0, \max_{\Gamma_-} u_{\text{in}} \right\}. \tag{52}$$

The semi-discrete AFC scheme (22) with $d_{ij}$ defined by (21) becomes

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \tilde{b}_i(u_h, \hat{u}) + \sum_{j \in \mathcal{N}_i^*} [(d_{ij} - \mathbf{c}_{ij} \cdot \mathbf{v})(u_j - u_i) + f_{ij}^*] \tag{53}$$

$$= \tilde{b}_i(u_h, \hat{u}) - \left( 2 \sum_{j \in \mathcal{N}_i^*} d_{ij} \right) u_i + \sum_{j \in \mathcal{N}_i^*} 2d_{ij} \left( \bar{u}_{ij} + \frac{f_{ij}^*}{2d_{ij}} \right) \tag{54}$$

$$= \tilde{b}_i(u_h, \hat{u}) - \left( 2 \sum_{j \in \mathcal{N}_i^*} d_{ij} \right) u_i + \sum_{j \in \mathcal{N}_i^*} 2d_{ij} \bar{u}_{ij}^*, \tag{55}$$

where

$$\tilde{b}_i(u_h, \hat{u}) = (u_i - u_{\text{in}}(\mathbf{x}_i)) \int_{\Gamma_-} \varphi_i \mathbf{v} \cdot \mathbf{n} \, \mathrm{ds} \tag{56}$$

and

$$d_{ij} = \max\{|\mathbf{c}_{ij} \cdot \mathbf{v}|, |\mathbf{c}_{ji} \cdot \mathbf{v}|\}. \tag{57}$$

If $i$ and $j$ are internal nodes, then $\mathbf{c}_{ji} = -\mathbf{c}_{ij}$ and, therefore, the bar state

$$\bar{u}_{ij} = \frac{u_j + u_i}{2} - \left( \frac{\mathbf{c}_{ij} \cdot \mathbf{v}}{d_{ij}} \right) \frac{u_j - u_i}{2} = \begin{cases} u_i & \text{if } \mathbf{c}_{ij} \cdot \mathbf{v} > 0, \\ u_j & \text{if } \mathbf{c}_{ij} \cdot \mathbf{v} < 0 \end{cases} \tag{58}$$

represents the solution value at the *upwind node* of the graph edge $\{i,j\}$ in the terminology of edge-based AFC schemes [42]. In other words, the addition of the GMS artificial viscosity transforms a centered approximation of the convective flux into an upwind-type approximation. Therefore, this type of graph Laplacian stabilization is called *discrete upwinding* in the AFC literature. Geometrically, $\mathbf{x}_i$ is the upwind node of the edge connecting $\mathbf{x}_i$ to $\mathbf{x}_j$ if and only if $(\mathbf{x}_j - \mathbf{x}_i) \cdot \mathbf{v} > 0$. Hence, the algebraic and geometric definitions of upwind nodes may differ in the case $d > 1$ if $\mathbf{c}_{ij} \neq \alpha(\mathbf{x}_j - \mathbf{x}_i)$ for some $\alpha > 0$ [39].

For constant velocity fields, the Galerkin flux $\mathbf{f}(u_h) = \mathbf{v}u_h$ coincides with its group finite element approximation defined by (12). This is no longer the case if $\mathbf{v} = \mathbf{v}(\mathbf{x})$ is a spatially variable vector field. Moreover, the corresponding advection problem is not of the form (1). Since the flux $\mathbf{f}(\mathbf{x}, u) = \mathbf{v}(\mathbf{x})u$ depends not only on $u$ but also on $\mathbf{x}$, the Rankine-Hugoniot velocity $\mathbf{v}_{ij}$ defined by (42) may become infinite on edges where $u_i = u_j$ and $\mathbf{v}_i \neq \mathbf{v}_j$. Hence, the IDP properties of exact and numerical solutions require further analysis.

If the velocity field $\mathbf{v}$ is divergence-free, then an invariant set of the continuous initial-boundary value problem is still defined by (50)–(52). In the case $\nabla \cdot \mathbf{v} \neq 0$, the method of characteristics can be used to show that

$$\mathcal{G} = \{u \in \mathbb{R} \,:\, u \geq 0\} \tag{59}$$

is an invariant set, i.e., at least positivity preservation is guaranteed for such $\mathbf{v}$.

The finite element discretization of $\nabla \cdot (\mathbf{v}u)$ produces edge contributions of the form $a_{ij}(u_j - u_i)$. The standard Galerkin method yields [42]

$$a_{ij} = \int_\Omega \varphi_i \mathbf{v} \cdot \nabla \varphi_j \, \mathrm{d}\mathbf{x} + \int_\Omega \varphi_i (\nabla \cdot \mathbf{v}) \varphi_j \, \mathrm{d}\mathbf{x}. \tag{60}$$

In the case $\nabla \cdot \mathbf{v} = 0$, we have $\sum_{j=1}^{N_h} a_{ij} = 0$ and the IDP property of the resulting discrete upwind schemes can be shown for [42, 46, 52]

$$d_{ij} \geq \max\{a_{ij}, 0, a_{ji}\}, \qquad j \in \mathcal{N}_i^*. \tag{61}$$

The validity of local maximum principles for the bar states $\bar{u}_{ij}$ and $\bar{u}_{ij}^*$ can be verified as before. Lohmann's [52] analysis for general AFC discretizations of advection problems proves $\mathcal{O}(h^{1/2})$ accuracy for any choice of IDP correction factors $\alpha_{ij}$. In particular, this worst-case convergence behavior is guaranteed for the low-order scheme and limiters that define $\alpha_{ij}$ in terms of bar states.

The group finite element approximation (12) produces $a_{ij} = \mathbf{c}_{ij} \cdot \mathbf{v}_j$, and the corresponding low-order scheme can be defined using the GMS formula

$$d_{ij} = \max\{|\mathbf{c}_{ij} \cdot \mathbf{v}_i|, |\mathbf{c}_{ij} \cdot \mathbf{v}_j|, |\mathbf{c}_{ji} \cdot \mathbf{v}_i|, |\mathbf{c}_{ji} \cdot \mathbf{v}_j|\}. \tag{62}$$

For internal nodes, this definition simplifies to $d_{ij} = \max\{|\mathbf{c}_{ij} \cdot \mathbf{v}_i|, |\mathbf{c}_{ij} \cdot \mathbf{v}_j|\}$ because $\mathbf{c}_{ji} = -\mathbf{c}_{ij}$. Positivity preservation is guaranteed for any velocity field. The discrete maximum principle for solenoidal velocities holds approximately because the group finite element formulation may produce nonvanishing terms $u_i \sum_{j=1}^{N_h} a_{ij} = u_i \sum_{j=1}^{N_h} \mathbf{c}_{ij} \cdot \mathbf{v}_j \approx m_i u_i (\nabla \cdot \mathbf{v})_i$ even in the case $\nabla \cdot \mathbf{v} = 0$.

*4.2. Case study: Burgers equation*

As a nonlinear counterpart of the first model problem, we consider the inviscid Burgers equation corresponding to (1) with the flux function $\mathbf{f}(u) = \mathbf{v}\frac{u^2}{2}$, where $\mathbf{v} \in \mathbb{R}^d$ is a constant vector. Let the initial data $u_0$ be piecewise-constant with values in the range $[u^{\min}, u^{\max}]$. Define the consistent normal flux

$$f_n(\hat{u}, u) := \mathbf{f}(u) \cdot \mathbf{n} - (u - \hat{u})\hat{u}\mathbf{v} \cdot \mathbf{n} \tag{63}$$

of the weakly imposed boundary condition (3) using the exact solution $u_{\text{ex}}$ of the Riemann problem in $\mathbb{R}^d$ to determine the upwind state $\hat{u} = u_{\text{in}}$ at the inlet $\Gamma_-(t) = \{\mathbf{x} \in \Gamma \,:\, u_{\text{ex}}(\mathbf{x}, t)\mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}$. Use $\hat{u} = u$ elsewhere. Then an invariant set $\mathcal{G}$ of the continuous problem is given by (50).

Definition (63) of the upwind flux leads to the weighted residual formulation

$$\int_\Omega w \left( \frac{\partial u}{\partial t} + \frac{1}{2}\nabla \cdot (\mathbf{v}u^2) \right) \mathrm{d}\mathbf{x} = \int_{\Gamma_-(t)} w(u - u_{\text{in}})u_{\text{in}}\mathbf{v} \cdot \mathbf{n} \, \mathrm{ds} \qquad \forall w \in W. \tag{64}$$

We discretize this problem using the group finite element approximation for the advective flux. The bar state corrections produce the semi-discrete scheme

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \tilde{b}_i(u_h, \hat{u}) + \sum_{j \in \mathcal{N}_i^*} [(d_{ij} - \mathbf{c}_{ij} \cdot \mathbf{v}_{ij})(u_j - u_i) + f_{ij}^*] \tag{65}$$

where $\mathbf{v}_{ij} = \frac{\mathbf{f}_j - \mathbf{f}_i}{u_j - u_i} = \mathbf{v}\frac{u_i + u_j}{2}$ is the shock velocity defined by (42) and

$$\tilde{b}_i(u_h, \hat{u}) = (u_i - u_{\text{in}}(\mathbf{x}_i)) \int_{\Gamma_-(t)} \varphi_i u_{\text{in}}\mathbf{v} \cdot \mathbf{n} \, \mathrm{ds}. \tag{66}$$

Following the analysis of the bar state behavior for the general case and linear advection with constant velocity, the IDP property can be shown for

$$d_{ij} \geq \max\{|\mathbf{c}_{ij} \cdot \mathbf{v}_{ij}|, |\mathbf{c}_{ji} \cdot \mathbf{v}_{ij}|\} \tag{67}$$

and, in particular, for the GMS diffusion coefficient (21) which is given by

$$d_{ij} = \max\{|\mathbf{c}_{ij} \cdot \mathbf{v}|, |\mathbf{c}_{ji} \cdot \mathbf{v}|\} \max\{|u_i|, |u_j|\}. \tag{68}$$

Since $\mathbf{c}_{ji} = -\mathbf{c}_{ij}$ for each pair of internal nodes $i$ and $j$, the use of $d_{ij} = |\mathbf{c}_{ij} \cdot \mathbf{v}_{ij}|$ for such node pairs yields bar states $\bar{u}_{ij}$ corresponding to upwind values.

## 5. Convex limiting for hyperbolic systems

In extensions to systems of conservation laws, local bounds may need to be imposed on nonlinear functions of $\bar{u}_{i1}, \ldots, \bar{u}_{im}$. The constrained nodal state $\bar{u}_i$ should belong to a subset $\mathcal{G} \cap \mathcal{G}_i$ of the convex invariant set $\mathcal{G}$. In the terminology of Multi-dimensional Optimal Order Detection (MOOD) [19, 68], the sets $\mathcal{G}$ and

$\mathcal{G}_i$ may be associated with physical and numerical admissibility conditions, respectively. While an appropriate invariant set $\mathcal{G}$ is often unambiguously defined by the physics of the problem at hand, mathematical structure of the continuous problem, and initial/boundary conditions, the set $\mathcal{G}_i$ should be defined to prevent large numerical errors and nonphysical solution behavior. In addition to local maximum principles of the form (39) for conserved or derived quantities [18, 25, 42, 48, 49, 51, 53], the definition of $\mathcal{G}_i$ may be based, e.g., on objectivity requirements [31, 56, 57], smoothness criteria [17, 18], and/or the principle of linearity preservation [8, 10, 42]. In general, $\mathcal{G}$ is not a subset of $\mathcal{G}_i$ and vice versa. For example, the use of smoothness indicators may result in violations of global bounds, while the preservation of these bounds does not guarantee the absence of spurious oscillations, entropy shocks, and other numerical artifacts.

### *5.1. Case study: Euler equations*

As an important example of a nonlinear hyperbolic system that requires a careful choice of the set $\mathcal{G}_i$ and of the limiting strategy, we consider the Euler equations of gas dynamics which can be written in the form (1). The vector $u$ of conserved variables and the matrix $\mathbf{f}(u)$ of inviscid fluxes are defined by

$$u = \begin{pmatrix} \rho \\ \rho\mathbf{v} \\ \rho E \end{pmatrix} \in \mathbb{R}^{d+2}, \quad \mathbf{f}(u) = \begin{pmatrix} \rho\mathbf{v} \\ \rho\mathbf{v}\otimes\mathbf{v} + pI_d \\ \rho E\mathbf{v} + p\mathbf{v} \end{pmatrix} \in \mathbb{R}^{d+2,d}, \tag{69}$$

where $I_d$ is the $d\times d$ identity tensor, $\rho$ is the density, $\mathbf{v}$ is the velocity, and $E$ is the specific total energy. The pressure $p$ of an ideal polytropic gas is related to the internal energy $\rho e$ by the equation of state

$$p = (\gamma-1)\left(\rho E - \frac{|\rho\mathbf{v}|^2}{2\rho}\right) = (\gamma-1)\rho e \tag{70}$$

with the heat capacity ratio $\gamma > 1$. Let $s^{\min} > 0$ be an arbitrary lower bound for the specific entropy $s$ of initial and boundary data. Then

$$\mathcal{G} = \{(\rho, \rho\mathbf{v}, \rho E)^\top \,:\, \rho > 0,\ e > 0,\ s \geq s^{\min}\} \tag{71}$$

is a convex invariant set of the Euler system [25]. Let $\mathcal{G}_i$ be the set of states $\bar{u}$ satisfying the numerical admissibility conditions (cf. [18, 31, 44])

$$\rho_i^{\min} \leq \bar{\rho} \leq \rho_i^{\max}, \tag{72}$$

$$\bar{\rho} v_{i,l}^{\min} \leq \overline{(\rho\mathbf{v})}\cdot\mathbf{e}_l \leq \bar{\rho} v_{i,l}^{\max}, \qquad l = 1,\ldots,d, \tag{73}$$

$$\bar{\rho} E_i^{\min} \leq \overline{(\rho E)} \leq \bar{\rho} E_i^{\max}, \tag{74}$$

where $\mathbf{e}_1,\ldots,\mathbf{e}_d$ are unit vectors of the Cartesian coordinate systems and the following definition of the local bounds for convex limiting is adopted:

$$\rho_i^{\min} = \min_{j\in\mathcal{N}_i}\rho_j, \qquad \rho_i^{\max} = \max_{j\in\mathcal{N}_i}\rho_j, \tag{75}$$

$$v_{i,l}^{\min} = \min_{j \in \mathcal{N}_i^*} (\bar{\mathbf{v}}_{ij} \cdot \mathbf{e}_l), \qquad v_{i,l}^{\max} = \max_{j \in \mathcal{N}_i^*} (\bar{\mathbf{v}}_{ij} \cdot \mathbf{e}_l), \tag{76}$$

$$E_i^{\min} = \min_{j \in \mathcal{N}_i^*} \bar{E}_{ij}, \qquad E_i^{\max} = \max_{j \in \mathcal{N}_i^*} \bar{E}_{ij}. \tag{77}$$

In the monolithic version, these bounds should be calculated using the nodal states $u_i = (\rho_i, (\rho\mathbf{v})_i, (\rho E)_i)^\top$ of the numerical solution at the beginning of the Runge-Kutta stage (or previous iteration in implicit schemes and steady state solvers) to calculate the invariant domain preserving low-order bar states $\bar{u}_{ij} = (\bar{\rho}_{ij}, \overline{(\rho\mathbf{v})}_{ij}, \overline{(\rho E)}_{ij})^\top \in \mathcal{G}$ and derived quantities of the form

$$\phi_i = \frac{(\rho\phi)_i}{\rho_i}, \qquad \bar{\phi}_{ij} = \frac{\overline{(\rho\phi)}_{ij} + \overline{(\rho\phi)}_{ji}}{\bar{\rho}_{ij} + \bar{\rho}_{ji}} = \bar{\phi}_{ji}. \tag{78}$$

The low-order bar states $\bar{u}_{ij}$ are calculated using the GMS diffusion coefficient $d_{ij}$ defined by (21). A guaranteed upper bound $\lambda_{ij}$ for the maximum wave speed $\lambda_{\max}(\mathbf{n}_{ij}, u_i, u_j)$ can be found in [25, 28]. The frequently used approximation $\tilde{\lambda}_{ij} = \max\{|\mathbf{v}_i| + c_i, |\mathbf{v}_j| + c_j\}$, where $c_i$ is the local speed of sound at node $i$, corresponds to the classical Rusanov (local Lax-Friedrichs) scheme. It is essentially nonoscillatory but not provably IDP because $\tilde{\lambda}_{ij}$ may underestimate $\lambda_{\max}(\mathbf{n}_{ij}, u_i, u_j)$ in some pathological cases, see [25] for further explanations.

The above definition of local bounds guarantees that $\bar{u}_{ij} \in \mathcal{G}_i$. To make sure that the constrained bar states $\bar{u}_{ij}^* = \bar{u}_{ij} + \frac{f_{ij}^*}{2d_{ij}}$ belong to $\mathcal{G}_i$ as well, we will construct the fluxes $f_{ij}^* = (f_{ij}^{*,\rho}, f_{ij}^{*,\rho v_1}, \ldots, f_{ij}^{*,\rho v_d}, f_{ij}^{*,\rho E})^\top$ using an edge-based bar state version of the sequential limiting strategy developed in [18, 31, 44] for element-based FCT algorithms. Given the vector $f_{ij} = (f_{ij}^{\rho}, f_{ij}^{\rho v_1}, \ldots, f_{ij}^{\rho v_d}, f_{ij}^{\rho E})^\top$ of target fluxes, we first enforce the numerical admissibility conditions (72)–(74) by limiting the components of $f_{ij}$ in the following sequential manner:

1. Limit $f_{ij}^{\rho}$ and update the density

$$f_{ij}^{*,\rho} = \alpha_{ij}^{\rho} f_{ij}^{\rho}, \qquad \rho_{ij}^* = \bar{\rho}_{ij} + \frac{f_{ij}^{*,\rho}}{2d_{ij}}. \tag{79}$$

2. For $\phi \in \{v_1, \ldots, v_d, E\}$ calculate

$$f_{ij}^{*,\rho\phi} = 2d_{ij} \left[ \rho_{ij}^* \bar{\phi}_{ij} - \overline{(\rho\phi)}_{ij} \right] + \alpha_{ij}^{\rho\phi} g_{ij}^{\rho\phi} \tag{80}$$

using

$$g_{ij}^{\rho\phi} = f_{ij}^{\rho\phi} + 2d_{ij} \left[ \overline{(\rho\phi)}_{ij} - \rho_{ij}^* \bar{\phi}_{ij} \right]. \tag{81}$$

By definition, the target fluxes satisfy $f_{ji}^{\rho\phi} = -f_{ij}^{\rho\phi}$. The conservation property $f_{ji}^{*,\rho\phi} = -f_{ij}^{*,\rho\phi}$ of their limited counterparts follows from the fact that

$$\begin{aligned} g_{ij}^{\rho\phi} + g_{ji}^{\rho\phi} &= 2d_{ij} \left[ \overline{(\rho\phi)}_{ij} + \overline{(\rho\phi)}_{ji} \right] - 2d_{ij} \left[ \rho_{ij}^* + \rho_{ji}^* \right] \bar{\phi}_{ij} + \left[ f_{ij}^{\rho\phi} + f_{ji}^{\rho\phi} \right] \\ &= 2d_{ij} \left[ \overline{(\rho\phi)}_{ij} + \overline{(\rho\phi)}_{ji} \right] - 2d_{ij} \left[ \bar{\rho}_{ij} + \bar{\rho}_{ji} \right] \bar{\phi}_{ij} - \left[ f_{ij}^{\rho} + f_{ji}^{\rho} \right] \bar{\phi}_{ij} = 0 \end{aligned}$$

by definition of $\bar{\phi}_{ij}$. A correction factor $\alpha_{ij}^{\rho}$ such that $\rho_{ij}^{*} \in [\rho_i^{\min}, \rho_i^{\max}]$ can be readily computed using the scalar flux limiter (45). The validity of

$$\rho_{ij}^{*}\phi_i^{\min} \le (\rho\phi)_{ij}^{*} = \overline{(\rho\phi)}_{ij} + \frac{f_{ij}^{*,\rho\phi}}{2d_{ij}} = \rho_{ij}^{*}\bar{\phi}_{ij} + \frac{\alpha_{ij}^{\rho\phi} g_{ij}^{\rho\phi}}{2d_{ij}} \le \rho_{ij}^{*}\phi_i^{\max} \tag{82}$$

is guaranteed at least for $\alpha_{ij}^{\rho\phi} = 0$ since this choice produces $(\rho\phi)_{ij}^{*} = \rho_{ij}^{*}\bar{\phi}_{ij}$. On the other hand, the unlimited target flux $f_{ij}^{*,\rho\phi} = f_{ij}^{\rho\phi}$ can be recovered using the correction factor $\alpha_{ij}^{\rho\phi} = 1$. Clearly, neither of these choices is generally optimal. Since conditions (82) are equivalent to the inequality constraints

$$2d_{ij}\rho_{ij}^{*}(\phi_i^{\min} - \bar{\phi}_{ij}) =: g_{ij}^{\min} \le g_{ij}^{*,\rho\phi} = \alpha_{ij}^{\rho\phi} g_{ij}^{\rho\phi} \le g_{ij}^{\max} := 2d_{ij}\rho_{ij}^{*}(\phi_i^{\max} - \bar{\phi}_{ij}),$$

a better approximation $g_{ij}^{*,\rho\phi}$ to $g_{ij}^{\rho\phi}$ can be calculated similarly to (46) thus:

$$g_{ij}^{*,\rho\phi} = \begin{cases} \min\{g_{ij}^{\max,\rho\phi}, \max\{g_{ij}^{\rho\phi}, g_{ij}^{\min,\rho\phi}\}\} & \text{if } g_{ij}^{\rho\phi} > 0, \\ \max\{g_{ij}^{\min,\rho\phi}, \min\{g_{ij}^{\rho\phi}, g_{ij}^{\max,\rho\phi}\}\} & \text{otherwise.} \end{cases} \tag{83}$$

In view of the fact that $\rho_{ij}^{*} \in [\rho_i^{\min}, \rho_i^{\max}]$ and $\rho_i^{\min} \ge 0$, the preconstrained density $\rho_{ij}^{*}$ is nonnegative. If the tentative bar state $u_{ij}^{*} \in \mathcal{G}_i$ fails to satisfy the additional conditions $e_{ij}^{*} > 0$ and $s_{ij}^{*} \ge s^{\min}$, they can be readily enforced using additional synchronized limiting of the offending fluxes $f_{ij}^{*}$ [31, 44]. Let the final, physically admissible bar state $\bar{u}_{ij}^{*} \in \mathcal{G}$ be defined by the formula

$$\bar{u}_{ij}^{*} = \bar{u}_{ij} + \frac{\alpha_{ij} f_{ij}^{*}}{2d_{ij}}. \tag{84}$$

The energy constraint $\bar{e}_{ij}^{*} \ge 0$ holds if $\alpha_{ij} \in [0, 1]$ is chosen to satisfy

$$\bar{\rho}_{ij}^{*}\overline{(\rho E)}_{ij}^{*} \ge \frac{|\overline{(\rho\mathbf{v})}_{ij}^{*}|^2}{2}, \tag{85}$$

i.e., if the kinetic energy does not exceed the total energy. Invoking definition (84) of $\bar{u}_{ij}^{*} = \left[\bar{\rho}_{ij}^{*}, \overline{(\rho\mathbf{v})}_{ij}^{*}, \overline{(\rho E)}_{ij}^{*}\right]$ and introducing the scaled bar states

$$\bar{w}_{ij} = [\bar{w}_{ij}^{\rho}, \bar{\mathbf{w}}^{\rho\mathbf{v}}, \bar{w}_{ij}^{\rho E}]^{\top} = 2d_{ij}\bar{u}_{ij} \tag{86}$$

for reasons explained in Section 4, we find that (85) is equivalent to

$$P_{ij}(\alpha_{ij}) \le Q_{ij} := \bar{w}_{ij}^{\rho}\bar{w}_{ij}^{\rho E} - \frac{|\bar{\mathbf{w}}^{\rho\mathbf{v}}|^2}{2}, \tag{87}$$

where $P_{ij}(\alpha)$ is a quadratic polynomial (cf. [53, 44]) defined by

$$P_{ij}(\alpha) = \left[\frac{|\mathbf{f}_{ij}^{*,\rho\mathbf{v}}|^2}{2} - f_{ij}^{*,\rho E} f_{ij}^{*,\rho}\right]\alpha^2 + \left[\bar{\mathbf{w}}^{\rho\mathbf{v}} \cdot \mathbf{f}_{ij}^{*,\rho\mathbf{v}} - \bar{w}_{ij}^{\rho} f_{ij}^{*,\rho E} - \bar{w}_{ij}^{\rho E} f_{ij}^{*,\rho}\right]\alpha.$$

Since $\alpha^n \leq \alpha$ for $\alpha \in [0,1]$ and $n \in \mathbb{N}$, the estimate $P_{ij}(\alpha) \leq \alpha R_{ij}$ holds for

$$R_{ij} = \max\left\{0, \bar{\mathbf{w}}^{\rho\mathbf{v}} \cdot \mathbf{f}_{ij}^{*,\rho\mathbf{v}} - \bar{w}_{ij}^{\rho} f_{ij}^{*,\rho E} - \bar{w}_{ij}^{\rho E} f_{ij}^{*,\rho} + R_{ij}^{+}\right\},$$

where

$$R_{ij}^{+} = \max\left\{0, \frac{|\mathbf{f}_{ij}^{*,\rho\mathbf{v}}|^2}{2} - f_{ij}^{*,\rho E} f_{ij}^{*,\rho}\right\}.$$

The symmetry condition $\alpha_{ij} = \alpha_{ji}$ and the internal energy constraint for node $j$ can be taken into account by using the synchronized correction factor

$$\alpha_{ij} = \begin{cases} \min\left\{\frac{Q_{ij}}{R_{ij}}, \frac{Q_{ji}}{R_{ji}}\right\} & \text{if } R_{ij} > Q_{ij},\ R_{ji} > Q_{ji}, \\ \frac{Q_{ij}}{R_{ij}} & \text{if } R_{ij} > Q_{ij},\ R_{ji} \leq Q_{ji}, \\ \frac{Q_{ji}}{R_{ji}} & \text{if } R_{ij} \leq Q_{ij},\ R_{ji} > Q_{ji}, \\ 1 & \text{otherwise.} \end{cases} \tag{88}$$

If $i$ and $j$ are internal nodes, we have $Q_{ji} = Q_{ij}$, and (88) simplifies to

$$\alpha_{ij} = \begin{cases} \frac{Q_{ij}}{\max\{R_{ij}, R_{ji}\}} & \text{if } \max\{R_{ij}, R_{ji}\} > Q_{ij}, \\ 1 & \text{otherwise.} \end{cases} \tag{89}$$

We use formula (88) in the numerical experiments below. Since it does not guarantee continuous dependence of $\alpha_{ij} f_{ij}^*$ on the data, convergence problems may occur in steady state computations according to the AFC theory developed in [9, 11, 52]. To avoid these problems, the triangle inequality estimate

$$\begin{aligned} \max\{R_{ij}, R_{ji}\} &= |\bar{\mathbf{w}}^{\rho\mathbf{v}} \cdot \mathbf{f}_{ij}^{*,\rho\mathbf{v}} - \bar{w}_{ij}^{\rho} f_{ij}^{*,\rho E} - \bar{w}_{ij}^{\rho E} f_{ij}^{*,\rho}| + R_{ij}^{+} \\ &\leq |\bar{\mathbf{w}}^{\rho\mathbf{v}}|\,|\mathbf{f}_{ij}^{*,\rho\mathbf{v}}| + |\bar{w}_{ij}^{\rho} f_{ij}^{*,\rho E}| + |\bar{w}_{ij}^{\rho E} f_{ij}^{*,\rho}| + R_{ij}^{+} =: R_{ij}^{\max} \end{aligned}$$

may be used to redefine the synchronized correction factor $\alpha_{ij}$ as follows:

$$\alpha_{ij} = \begin{cases} \frac{\min\{Q_{ij}, Q_{ji}\}}{R_{ij}^{\max}} & \text{if } R_{ij}^{\max} > Q_{ij}, \\ 1 & \text{otherwise.} \end{cases} \tag{90}$$

This modification of (88) satisfies the theoretical requirements for the design of limiter functions without producing significantly higher levels of artificial viscosity, as we show in two one-dimensional examples of Section 6.

Positivity preservation for the assembled nodal values of the internal energy and pressure follows by Jensen's inequality [18, 31, 44]. To enforce the entropy constraint $\min\{\bar{s}_{ij}, \bar{s}_{ji}\} \geq s^{\min}$, an entropy-consistent synchronized correction factor $\alpha_{ij}^s \leq \alpha_{ij}$ can be calculated using the convex limiting methodology originally developed in [25] for FCT algorithms of predictor-corrector type.

The reader may wonder why we calculate the prelimited fluxes $f_{ij}^*$ instead of multiplying $f_{ij}$ by $\alpha_{ij} \in [0,1]$ such that $\bar{u}_{ij}^* = \bar{u}_{ij} + \frac{\alpha_{ij} f_{ij}}{2d_{ij}} \in \mathcal{G}_i \cap \mathcal{G}$. The reason

for this is very simple. Using the same correction factor for all components of $f_{ij}$ generates excessively strong numerical dissipation. The prelimited bar states $u_{ij}^* = \bar{u}_{ij} + \frac{\alpha_{ij} f_{ij}}{2d_{ij}} \in \mathcal{G}_i$ are more accurate and no further correction of these states is required if $u_{ij}^* \in \mathcal{G}$. Moreover, the magnitude of the raw antidiffusive fluxes $f_{ij}^*$ is smaller and, therefore, worst-case estimates for nonlinear functions of $\alpha_{ij}$ are not as pessimistic as in fully synchronized limiters for $f_{ij}$ [44].

*5.2. Case study: tensorial advection*

Hyperbolic problems of the form (1) are also used to model advective transport of tensor fields which play an important role, e.g., in numerical simulations of fiber suspension flows and injection molding processes [49, 51, 52]. Let the states $u \in \mathbb{R}^{d(d+1)/2}$ be defined as arrays containing the independent components of symmetric tensors $U \in \mathbb{R}^{d\times d}$. The real eigenvalues of these tensors will be denoted by $\lambda_l(u),\ l = 1, \ldots, d$ and sorted in ascending order.

The flux function of the tensorial linear advection problem is $\mathbf{f} = \mathbf{v}u$. If $\nabla \cdot \mathbf{v} = 0$ holds in $\Omega$, the eigenvalues of $u$ remain bounded by [52]

$$\lambda^{\min} = \min\left\{\min_{\Omega} \lambda_1(u_0), \min_{\Gamma_-} \lambda_1(u_{\text{in}})\right\}, \tag{91}$$

$$\lambda^{\max} = \max\left\{\max_{\Omega} \lambda_d(u_0), \max_{\Gamma_-} \lambda_d(u_{\text{in}})\right\}. \tag{92}$$

Hence, $\mathcal{G} = \{u \in \mathbb{R}^{d(d+1)/2} : \lambda^{\min} \le \lambda_1(u),\ \lambda_d(u) \le \lambda^{\max}\}$ is a convex invariant set of the initial-boundary value problem. In particular, this fact implies preservation of positive semidefiniteness in the process of advection.

The set $\mathcal{G}_i$ of numerically admissible states may be defined to enforce inequality constraints depending on the nature of the problem at hand. As shown by Lohmann [49, 51, 52], the imposition of local bounds

$$\lambda_i^{\min} = \min_{j\in\mathcal{N}_i^*} \lambda_1(u_j) \ge \lambda^{\min}, \qquad \lambda_i^{\max} = \max_{j\in\mathcal{N}_i^*} \lambda_d(u_j) \le \lambda^{\max} \tag{93}$$

on the maximal and minimal eigenvalues, i.e., using $\mathcal{G}_i = \{u \in \mathbb{R}^{d(d+1)/2} : \lambda_i^{\min} \le \lambda_1(u),\ \lambda_d(u) \le \lambda_i^{\max}\} \subseteq \mathcal{G}$ is particularly well suited for AFC discretizations of fiber suspension flow models. This definition of $\mathcal{G}_i$ turned out to be a good criterion for the design of eigenvalue range preserving (ERP) limiters.

Since all components of evolving tensors are advected by the same velocity field $\mathbf{v}$, the diffusion coefficient $d_{ij}$ is defined as in Section 4.1. The corresponding low-order scheme is ERP, as shown by Lohmann [49, 51, 52]. The representation of this scheme and its limited high-order extensions in terms of the bar states makes it possible to obtain alternative proofs of the ERP property (at least for explicit SSP Runge-Kutta time discretizations) and convert the FCT algorithms developed in [49] into tensorial bar state limiters. We also envisage that a new generation of ERP tensor limiters could be designed exploiting the fact that $\bar{u}_{ij} = \bar{u}_{ji}$ for internal edges and, therefore, limiting can be performed in the principal axis system of $\bar{U}_{ij}$ using individually chosen correction factors for different eigenvalues (see [51, 52] for examples of such spectral limiters).

## 6. Numerical examples

In this section, we perform numerical studies of the monolithic convex limiting (MCL) procedures and compare some results to those obtained with other approaches (NVL: monolithic nodal variation limiter using $\alpha_{ij} = \min\{\alpha_i, \alpha_j\}$ with linearity-preserving nodal correction factors $\alpha_i$ defined as in [40], FCT: predictor-corrector scheme based on a sequential FCT algorithm [18, 31, 44]). The methods under investigation are applied to well-documented test problems for steady and unsteady linear advection, the two-dimensional inviscid Burgers equation, and the Euler equations. We demonstrate the superb shock-capturing properties of bar state limiters and perform grid convergence studies for an advection problem with a smooth steady-state solution. In all examples, we use linear or bilinear Lagrange finite elements. Time integration is performed using an explicit second-order accurate SSP Runge-Kutta scheme.

Given a reference solution $u$, we measure the errors of numerical approximations $u_h$ on successively refined meshes using the discrete $L^1$ norm [41, 42]

$$E_1(h) := \sum_{i=1}^{N_h} m_i |u(\mathbf{x}_i) - u_i| \approx \int_\Omega |u - u_h| \, \mathrm{d}\mathbf{x} = \|u - u_h\|_{L^1(\Omega)}, \tag{94}$$

where $m_i$ is a diagonal coefficient of the lumped mass defined by (10). The experimental order of convergence is determined using the formula [47]

$$p = \log_2 \left( \frac{E_1(2h)}{E_1(h)} \right). \tag{95}$$

In some examples, the values of global maxima and minima are reported in order to verify the invariant domain preservation properties of the AFC schemes under investigation and quantify the levels of numerical dissipation.

### *6.1. Steady circular advection*

In the first numerical experiment, we solve the steady advection equation

$$\nabla \cdot (\mathbf{v} u) = 0 \quad \text{in } \Omega = (0,1)^2 \tag{96}$$

using the divergence-free velocity field $\mathbf{v}(x,y) = (y, -x)$. The inflow boundary condition and the exact solution at any point in $\bar{\Omega}$ are given by

$$u(x,y) = \begin{cases} 1, & \text{if } \ 0.15 \le r(x,y) \le 0.45, \\ \cos^2\left(10\pi \frac{r(x,y)-0.7}{3}\right), & \text{if } \ 0.55 \le r(x,y) \le 0.85, \\ 0, & \text{otherwise}, \end{cases} \tag{97}$$

where $r(x,y) = \sqrt{x^2 + y^2}$ denotes the distance to the corner point $(0,0)$.

Numerical solutions are marched to the steady state using explicit SSP Runge-Kutta time stepping. Simulations are terminated when the Euclidean norm of the steady state residual becomes smaller than the prescribed tolerance. In contrast to the linearity-preserving local bounds of the NVL version,

the MCL constraints for the bar states are defined without using free parameters. Remarkably, the steady state residuals of MCL discrete problem converge to machine zero in a monotone fashion, while the behavior of NVL-like monolithic limiters is strongly affected by the choice of such parameters. Less restrictive bounds imply lower levels of numerical diffusion but the convergence behavior of fixed-point iterations becomes unsatisfactory and steady state residuals begin to stagnate. No such problems were observed in simulations with MCL.

In Figure 1, we present the results of steady-state computations on uniform meshes with $N_h = (129)^2 = 16,642$ nodes using $\mathbb{P}_1$ and $\mathbb{Q}_1$ elements. All

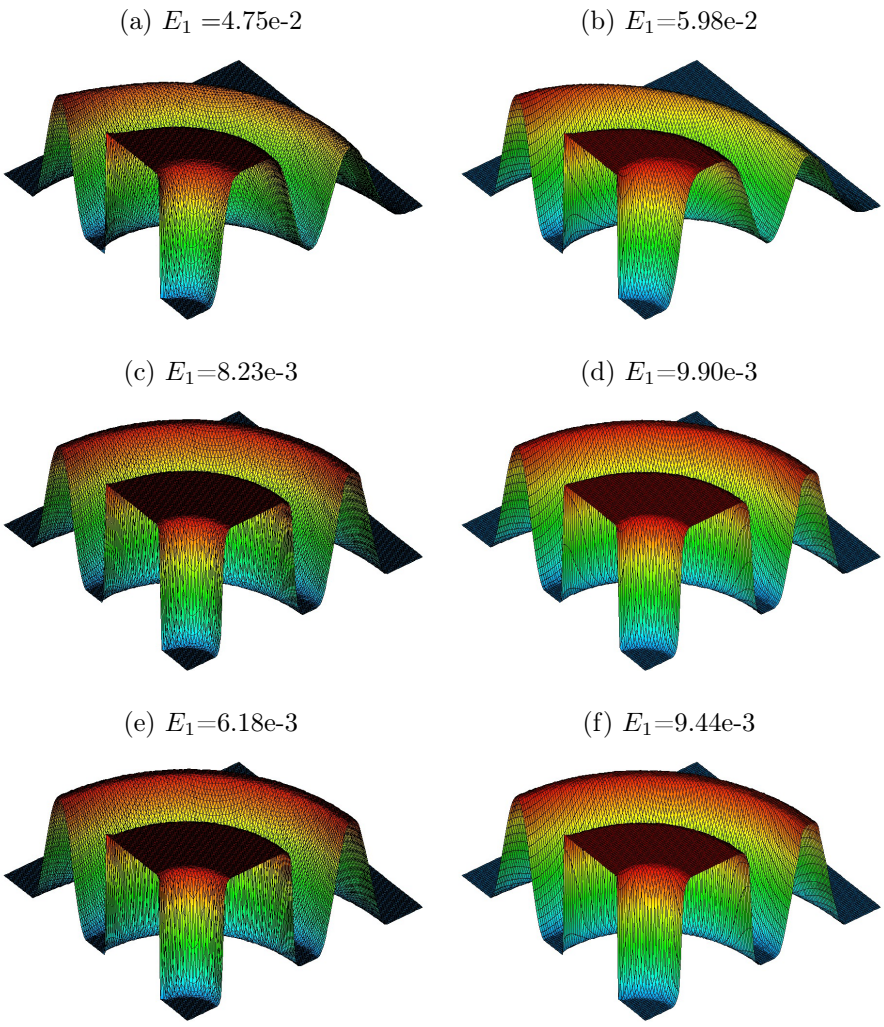



Figure 1: Steady circular advection. Solutions produced by the low-order scheme (top row), MCL (middle row), and NVL (bottom row) on uniform meshes with $N_h = (129)^2 = 16,642$ nodes. The results for $\mathbb{P}_1$ and $\mathbb{Q}_1$ elements are shown in the left and right column, respectively.

numerical approximations are bounded below by $u^{\min} = 0 = \min_{1\le i\le N_h} u_i$ and above by $u^{\max} = 1 = \max_{1\le i\le N_h} u_i$. For a better quantitative comparison, the $E_1$ errors are listed above each diagram. The flux-corrected solutions are far more accurate than the diffusive results produced by the low-order version of the IDP scheme. Both AFC schemes used the unstabilized Galerkin target flux $f_{ij} = d_{ij}(u_i - u_j)$. The NVL parameter $\gamma_i$ was set equal to $2\gamma_i^{\min}$, where $\gamma_i^{\min}$ is the linearity-preserving lower bound presented in [40].

| $h$ | MCL | $p$ | NVL | $p$ |
|---|---|---|---|---|
| $\frac{1}{64}$ | 4.52e-3 | | 3.87e-3 | |
| $\frac{1}{128}$ | 1.16e-3 | 1.96 | 9.17e-4 | 2.08 |
| $\frac{1}{256}$ | 2.99e-4 | 1.96 | 2.10e-4 | 2.13 |

Table 1: Steady circular advection. $E_1$ convergence history of the MCL = MCL-LP and NVL schemes on uniform triangular meshes.

| $h$ | MCL | $p$ | NVL | $p$ | MCL-LP | $p$ |
|---|---|---|---|---|---|---|
| $\frac{1}{64}$ | 4.87e-3 | | 4.85e-3 | | 4.61e-3 | |
| $\frac{1}{128}$ | 1.46e-3 | 1.74 | 1.50e-3 | 1.69 | 1.37e-3 | 1.75 |
| $\frac{1}{256}$ | 5.14e-4 | 1.51 | 5.02e-4 | 1.58 | 4.14e-4 | 1.72 |

Table 2: Steady circular advection. $E_1$ convergence history of the MCL, NVL, and MCL-LP schemes on perturbed triangular meshes.

Tables 1 and 2 show the results of grid convergence studies on uniform and perturbed triangular meshes for the smooth exact solution [52]

$$u(x,y) = \exp\left(-100(r(x,y)-0.7)^2\right), \qquad 0 \le x, y \le 1. \tag{98}$$

In this numerical study of MCL and NVL, perturbed counterparts of uniform grids with spacing $h = \frac{1}{\sqrt{N_h}-1}$ were generated by adding random numbers $\xi_i, \eta_j \in [-0.25h, 0.25h]$ to the Cartesian coordinates $(x_i, y_j) \in \Omega$ of internal mesh nodes. By definition of local bounds, NVL is linearity-preserving (LP) on any mesh, while the basic definition (39) of the MCL bounds is LP only on symmetric meshes (as defined in [8]). There is strong numerical and some theoretical evidence indicating that the LP property is essential for achieving optimal convergence rates ($p = 1.5$ for linear finite elements and smooth solutions) on general meshes, see [10, 40, 42]. As an upgrade of the basic MCL scheme, we consider a linearity-preserving (MCL-LP) version in which the bounds

$$u_i^{\min} = \min_{j\in\mathcal{N}_i} \min\{u_j, \hat{u}_{k(i,j)}\}, \qquad u_i^{\max} = \max_{j\in\mathcal{N}_i} \max\{u_j, \hat{u}_{k(i,j)}\} \tag{99}$$

for the $\mathbb{P}_1$ discretization are defined using bound-preserving extrapolated values $\hat{u}_{k(i,j)} = u_i + (\nabla u)_i \cdot (\mathbf{x}_i - \mathbf{x}_j)$ at the dummy nodes $\mathbf{x}_{k(i,j)} = \mathbf{x}_j + 2(\mathbf{x}_i - \mathbf{x}_j)$.

If the nodal gradient $(\nabla u)_i$ is calculated using the basis functions of the first element crossed by the line connecting an internal node $\mathbf{x}_i$ to $\mathbf{x}_{k(i,j)}$, then $\hat{u}_{k(i,j)}$ is a convex combination of the solution values at the nodes of this element [35, 58, 59]. Moreover, we have $u_i = \frac{1}{2}(u_j + \hat{u}_{k(i,j)})$ if $\nabla u_h$ is continuous at node $i$. Hence, definition (99) is linearity-preserving on general simplex meshes. On symmetric meshes, the so-defined MCL-LP scheme reduces to MCL.

The $E_1$ convergence history presented in Table 1 demonstrates that both MCL(-LP) and NVL exhibit second-order convergence on uniform meshes. On perturbed meshes, the experimental orders of accuracy exceed the provable lower bound $p = 1.5$ for stabilized FEM. The use of linearity-preserving bounds (99) in MCL-LP leads to a marked improvement compared to the basic MCL scheme. Although even the MCL-LP bounds are relatively tight and no parameter tuning is involved, both the EOCs and the convergence rates w.r.t. steady-state residuals are higher than those of the carefully configured NVL scheme.

### *6.2. Solid body rotation*

The solid body rotation benchmark [38, 47] is an obligatory 2D test for numerical advection schemes. We use it in this numerical study to facilitate direct comparison with other algebraic flux correction schemes [40, 41, 42, 52] and variational shock capturing techniques for stabilized finite element methods [38]. In this experiment, we solve the unsteady linear advection equation

$$\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) = 0 \quad \text{in } \Omega = (0,1)^2 \tag{100}$$

using the velocity field $\mathbf{v}(x,y) = (0.5-y, x-0.5)^\top$ to rotate a slotted cylinder, a sharp cone, and a smooth hump around the center $(0.5, 0.5)$ of the domain $\Omega$. The initial condition, as defned by LeVeque [47], is given by

$$u_0(x,y) = \begin{cases} u_0^{\text{hump}}(x,y) & \text{if } \sqrt{(x-0.25)^2+(y-0.5)^2} \le 0.15, \\ u_0^{\text{cone}}(x,y) & \text{if } \sqrt{(x-0.5)^2+(y-0.25)^2} \le 0.15, \\ 1 & \text{if } \begin{cases} \left(\sqrt{(x-0.5)^2+(y-0.75)^2} \le 0.15\right) \wedge \\ \left(|x-0.5| \ge 0.025 \vee \ y \ge 0.85\right), \end{cases} \\ 0 & \text{otherwise,} \end{cases}$$

where

$$u_0^{\text{hump}}(x,y) = \frac{1}{4} + \frac{1}{4}\cos\left(\frac{\pi\sqrt{(x-0.25)^2+(y-0.5)^2}}{0.15}\right), \tag{101}$$

$$u_0^{\text{cone}}(x,y) = 1 - \frac{\sqrt{(x-0.5)^2+(y-0.25)^2}}{0.15}. \tag{102}$$

Homogeneous Dirichlet boundary conditions are prescribed at the inlets.

After each full rotation, the exact solution $u(\cdot, 2\pi k),\ k \in \mathbb{N}$ coincides with the initial data $u_0$. In Figure 2, we present the $\mathbb{Q}_1$ interpolant $u_h(\cdot,0) = I_h u_0$ of

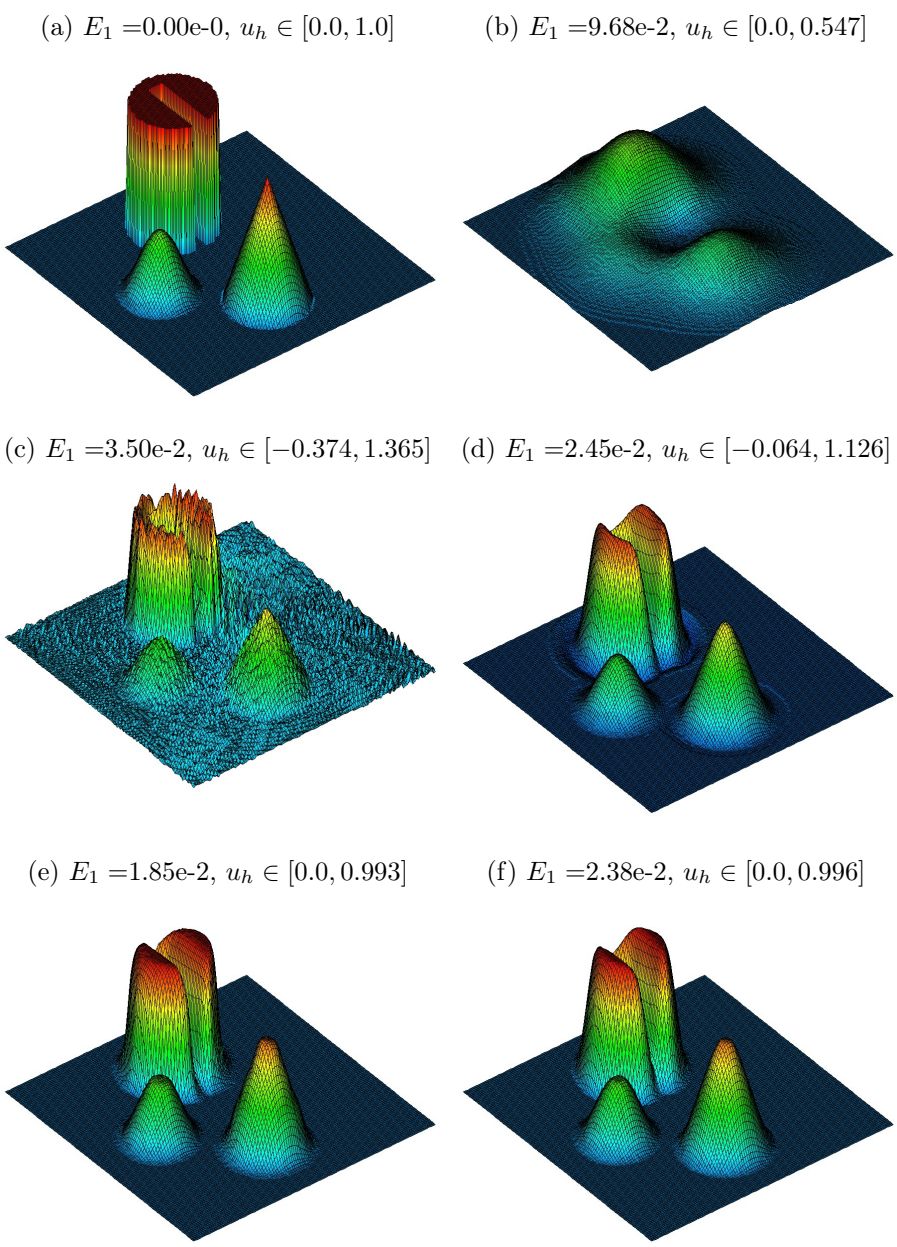



Figure 2: Solid body rotation [47]. The plots show (a) projected initial data $u_h(\cdot, 0)$ and $\mathbb{Q}_1$ approximations $u_h(\cdot, 2\pi)$ produced by (b) the low-order scheme ($f_{ij} = 0$) and flux-corrected high-order schemes using (c) unlimited fluxes $f_{ij}^H$, (d) unlimited fluxes $f_{ij}$, (e) MCL for $f_{ij}^H$, and (f) MCL for $f_{ij}$ on a uniform mesh with $h = \frac{1}{128}$ and $\Delta t = 10^{-3}$.

$u_0$ and numerical solutions $u_h(\cdot, 2\pi)$ calculated on a uniform mesh of $128 \times 128$ bilinear elements using the time step $\Delta t = 10^{-3}$. In addition to the values of the $E_1$ error w.r.t. $u_h(\cdot, 0) = I_h u(\cdot, 2\pi)$, the global maxima and minima of the $\mathbb{Q}_1$ approximations are shown above each plot to assess the amounts of numerical diffusion and magnitudes of spurious undershoots/overshoots (if any).

The low-order solution presented in Fig. 2(a) is bound-preserving but very diffusive. After one full rotation, the global maximum decreases to 0.547, and the shapes of the rotating objects are smeared so strongly that they are hardly recognizable. In Figs 2(c,d), we show the results of adding unlimited antidiffusive fluxes $f_{ij}^H$ and $f_{ij}$, as defined by (26) and (29), respectively. By definition of $f_{ij}^H$, this choice recovers the highly oscillatory standard Galerkin approximation (25). The fluxes $f_{ij}$ can be calculated more efficiently and correspond to a stabilized high-order target. The discrete maximum principle is still violated but undershoots and overshoots stay in a neighborhood of steep gradients, and their magnitude is smaller than the case of the AFC scheme using the $f_{ij}^H$ target. The MCL-constrained counterparts of these solutions are presented in Figs 2(e,f). The difference between the flux-corrected approximations is marginal. We conclude that $f_{ij}$ is a better alternative to $f_{ij}^H$ in terms of stability and efficiency. For that reason, the remaining experiments of this section are performed using the $f_{ij}$ version of MCL, i.e., definition (29) of the target flux.

*6.3. Burgers equation*

As a nonlinear scalar test problem, we consider two-dimensional inviscid Burgers equation defined as in Section 4.2 using the constant vector $\mathbf{v} = (1, 1)$ in the flux function $\mathbf{f}(u) = \mathbf{v}\frac{u^2}{2}$. The exact solution $u_{\text{ex}}$ corresponding to

$$u_0(x, y) = \begin{cases} -0.2 & \text{if} \quad x < 0.5 \ \wedge y > 0.5, \\ -1.0 & \text{if} \quad x > 0.5 \ \wedge y > 0.5, \\ \ \ 0.5 & \text{if} \quad x < 0.5 \ \wedge y < 0.5, \\ \ \ 0.8 & \text{if} \quad x > 0.5 \ \wedge y < 0.5 \end{cases} \tag{103}$$

can be found in [24]. This solution is defined in $\mathbb{R}^2$ and stays in the invariant set $\mathcal{G} = [-1.0, 0.8]$. We use $u_{\text{ex}}$ to impose flux boundary conditions in the manner described in Section 4.2. In our numerical experiment, the computational domain $\Omega$ is a unit square and simulations are terminated at the final time $T = 0.5$. In Figure 3, we present the low-order solution and the MCL solution calculated using a $\mathbb{P}_1$ discretization on a uniform triangular mesh. The mesh size $h = \frac{1}{128}$ and time step $\Delta t = 10^{-3}$ are the same as in the solid body rotation test. The accuracy of the low-order solution shown in Fig. 3(a) is not as poor as in the linear advection examples. Since shocks are self-steepening, the low-order scheme resolves them fairly well but the rarefaction is strongly smeared and the corners are rounded. The MCL scheme performs much better in the rarefaction wave region (see Fig. 3(b)) and exhibits higher overall accuracy in terms of the $E_1$ error. The invariant domain preservation capability of both schemes is illustrated by the fact that $u_i \in \mathcal{G}$ for all $i = 1, \ldots, N_h$.

(a) $E_1$ =1.13e-2, $u_h \in [-1.0, 0.8]$ (b) $E_1$ =7.75e-3, $u_h \in [-1.0, 0.8]$

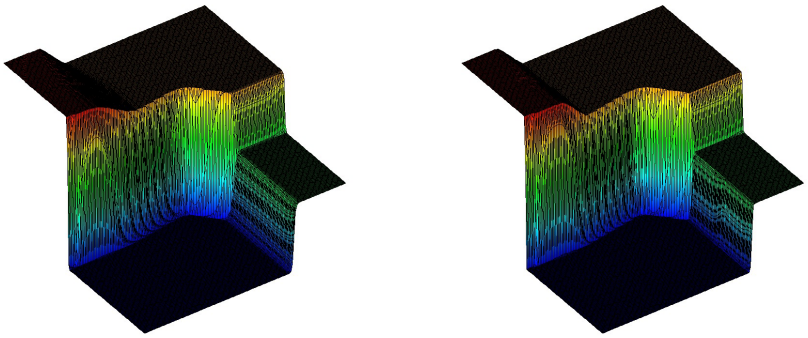


Figure 3: Burgers equation [24]. Snapshots $u_h(\cdot, 0.5)$ of the $\mathbb{P}_1$ approximations produced by (a) the low-order scheme and (b) MCL on a uniform mesh with $h = \frac{1}{128}$ and $\Delta t = 10^{-3}$.

*6.4. Sod's shock tube*

Sod's shock tube problem [67] is a well-known 1D benchmark for the Euler equations. The computational domain $\Omega = (0, 1)$ has reflective walls and is initially separated by a membrane into two sections. When the membrane is removed, the gas begins to flow into the region of lower pressure. The initial condition for the nonlinear Riemann problem is given by

$$\begin{bmatrix} \rho_L \\ v_L \\ p_L \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \end{bmatrix}, \qquad \begin{bmatrix} \rho_R \\ v_R \\ p_R \end{bmatrix} = \begin{bmatrix} 0.125 \\ 0.0 \\ 0.1 \end{bmatrix}. \tag{104}$$

The removal of the membrane at the time $t = 0$ releases a shock wave that propagates to the right with velocity satisfying the Rankine-Hugoniot conditions. All of the primitive variables are discontinuous across the shock which is followed by a contact discontinuity. The latter represents a moving interface between the regions of different densities but constant velocity and pressure. The rarefaction wave propagates in the opposite direction providing a smooth transition to the original values of the state variables in the left part of the domain. Hence, the one-dimensional flow pattern in the shock tube is characterized by three waves traveling at different speeds.

Since our MCL scheme for the Euler equations (as presented in Section 5.1) is closely related to the sequential FCT algorithm developed in [18], our preliminary evaluation of MCL involves a comparison to the FCT approach. In Figure 4, we show the density (blue), velocity (green), and pressure (red) distributions corresponding to the final time $T = 0.231$. The analytical solution of Sod's shock tube problem is shown by the solid lines without markers. The corresponding finite element approximations are shown as solid lines with bullet markers. As in [18], we use a uniform mesh of 128 linear elements and the time step $\Delta t = 10^{-3}$. Although the MCL and FCT schemes are based on the same design principles, the monolithic bar state version of the sequential $(\rho, v, E)$ limiting strategy achieves visibly higher resolution of the contact discontinuity and

produces smaller $E_1$ errors for all primitive variables. If the correction factor $\alpha_{ij}$ is calculated using (90) instead of (88) to ensure continuous dependence of $\alpha_{ij} f_{ij}^*$ on the input, the MCL solution remains unchanged in this example.

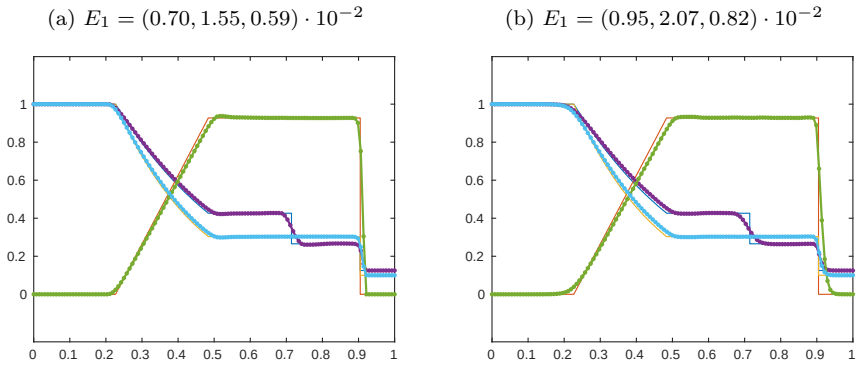



Figure 4: Sod's shock tube [67]. Exact solution for $(\rho, v, p)$ at $T = 0.231$ vs. $\mathbb{P}_1$ approximations obtained with (a) MCL and (b) sequential FCT using $h = \frac{1}{128}$ and $\Delta t = 10^{-3}$.

### *6.5. Blast wave problem*

The blast wave problem of Woodward and Colella [69] models the flow of a $\gamma = 1.4$ ideal gas between reflecting walls. The three constant states

$$\begin{bmatrix} \rho_L \\ v_L \\ p_L \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 1000.0 \end{bmatrix}, \qquad \begin{bmatrix} \rho_M \\ v_M \\ p_M \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 0.01 \end{bmatrix}, \qquad \begin{bmatrix} \rho_R \\ v_R \\ p_R \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 100.0 \end{bmatrix} \tag{105}$$

of the initial condition for the primitive variables are associated with the subdomains $\Omega_L = (0, 0.1)$, $\Omega_M = (0.1, 0.9)$, and $\Omega_R = (0.9, 1)$.

The above initial conditions give rise to two strong blast waves which eventually collide. The flow evolution involves complex interactions of shocks, rarefactions, and contact discontinuities in a small region of space. These interactions impose more stringent requirements on the robustness of numerical solution methods than Sod's shock tube problem. Limiters that do not guarantee positivity preservation for the pressure are likely to fail in this test. The MCL and FCT results for the density at $T = 0.038$ are shown in in Fig. 5. The numerical solutions obtained using 1000 linear elements and the time step $10^{-6}$ are shown as red circles. The blue solid line depicts the reference solution. It can be seen that MCL produces a smaller $E_1$ density error than the original FCT algorithm. This error increases marginally from $5.46 \cdot 10^{-2}$ to $5.60 \cdot 10^{-2}$ if the internal energy correction factor $\alpha_{ij}$ is calculated using (90) instead of (88). The shape of the density profile remains virtually unchanged and is not shown here.

### *6.6. Double Mach reflection*

In the last example, we consider the double Mach reflection benchmark [69] for the two-dimensional Euler equations. The computational domain for this

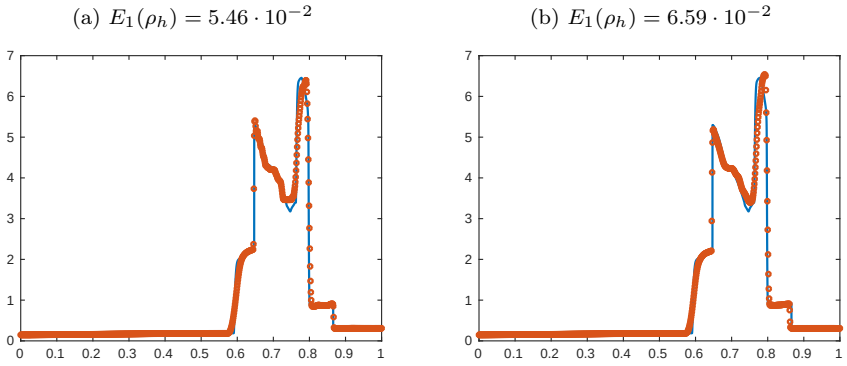



Figure 5: Blast wave problem [69]. Reference solution for $\rho$ at $T = 0.038$ vs. $\mathbb{P}_1$ approximations obtained with (a) MCL and (b) sequential FCT using $h = 10^{-3}$ and $\Delta t = 10^{-6}$.

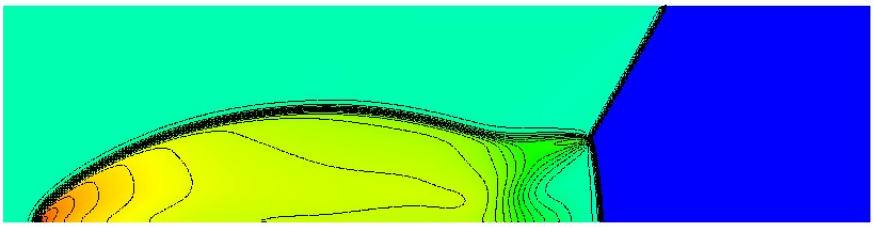
test is the rectangle $\Omega = (0, 4) \times (0, 1)$. The flow pattern features a propagating Mach 10 shock in air ($\gamma = 1.4$) which initially makes a 60° angle with a reflecting wall. The following pre-shock and post-shock values of the flow variables are used to define the initial and boundary conditions

$$\begin{bmatrix} \rho_L \\ u_L \\ v_L \\ p_L \end{bmatrix} = \begin{bmatrix} 8.0 \\ 8.25\cos(30°) \\ -8.25\sin(30°) \\ 116.5 \end{bmatrix}, \qquad \begin{bmatrix} \rho_R \\ u_R \\ v_R \\ p_R \end{bmatrix} = \begin{bmatrix} 1.4 \\ 0.0 \\ 0.0 \\ 1.0 \end{bmatrix}. \tag{106}$$

Initially, the post-shock values (subscript $L$) are prescribed in the subdomain $\Omega_L = \{(x, y) \mid x < 1/6 + y/\sqrt{3}\}$ and the pre-shock values (subscript $R$) in $\Omega_R = \Omega \backslash \Omega_L$. The reflecting wall corresponds to $1/6 \leq x \leq 4$ and $y = 0$. No boundary conditions are required along the line $x = 4$. On the rest of the boundary, the post-shock conditions are assigned for $x < 1/6 + (1 + 20t)/\sqrt{3}$ and the pre-shock conditions elsewhere. The so-defined values along the top boundary describe the exact motion of the initial Mach 10 shock.

In Figure 6, we present snapshots of the density distribution at $T = 0.2$ calculated using $\mathbb{Q}_1$ approximation on a uniform mesh with $h = \frac{1}{128}$. In this example, time integration was performed using the implicit Crank-Nicolson scheme and the time step $\Delta t = 10^{-4}$. As already mentioned, monolithic AFC schemes like MCL support the use of general time integrators (although formal proofs of positivity preservation are usually restricted to the basic two-level $\theta$ scheme and SSP Runge-Kutta methods). The MCL solution of the double Mach reflection problem exhibits higher resolution than its low-order counterpart and is also free of spurious oscillations. Numerical results obtained with predictor-corrector approaches can be found in [18, 53]. Although no reference solutions are available for this test, the MCL scheme seems to resolve the fine-scale features at least as well as the best limiters we have tested so far. We conclude that convex limiting techniques of this kind are a useful tool for enforcing invariant domain preservation in finite element methods for computational gas dynamics.

(a)



(b)

Figure 6: Double Mach reflection [69]. Density distribution at $T = 0.2$ obtained with (a) the low-order scheme and (b) MCL using $h = \frac{1}{128}$ and $\Delta t = 10^{-4}$.

## 7. Conclusions

The main result of this paper is the finding that the representation of modified Galerkin schemes in terms of edge-based bar states is very useful not only for the derivation of GMS artificial viscosities that lead to provably invariant domain preserving low-order approximations but also for the design of monolithic algebraic flux correction schemes in which the limited antidiffusive terms are incorporated into the bar states. The new approach to edge-based convex limiting bridges the gap between predictor-corrector algorithms of FCT type and monolithic AFC approaches that enforce inequality constraints for nodal states using parameter-dependent local bounds. Any FCT scheme or the equivalent slope limiting procedure for the gradients of a piecewise-linear approximation (see [18, 71] for a unified presentation of algebraic and geometric limiting techniques) can be readily converted into a parameter-free limiter for the well-posed nonlinear discrete problem of the monolithically constrained version.

Our presentation of new bar state limiters for IDP high-resolution schemes was focused on continuous edge-based $\mathbb{P}_1/\mathbb{Q}_1$ finite element methods for scalar conservation laws and the Euler equations of gas dynamics. However, the generality of the proposed methodology paves the way for major upgrades of property-preserving limiters for symmetric tensor fields [49, 51, 52], shallow water equations, [4, 30, 31], element-based correction procedures [3, 18, 54], discontinuous Galerkin (DG) methods [3, 7, 18, 31], high-order Bernstein polynomial approximations [3, 54], residual distribution schemes [2, 1, 32], and *hp*-adaptive finite el-

ement methods that construct a continuous limiter-controlled partition of unity using the basis functions of high-order and low-order spaces [45]. Moreover, rigorous theoretical studies of the presented schemes and their extensions can be performed using the AFC analysis framework developed in [9, 11, 52].

**Acknowledgments.** This research was supported by the German Research Association (DFG) under grant KU 1530/23-1. The author would like to thank Christoph Lohmann (TU Dortmund University) for insightful remarks regarding the properties of bar state limiters.

## References

[1] R. Abgrall, P. Bacigaluppi, and S. Tokareva, High-order residual distribution scheme for the time-dependent Euler equations of fluid dynamics. *Computers & Mathematics with Applications* **78** (2019) 274–297.

[2] R. Abgrall and S. Tokareva, Staggered grid residual distribution scheme for Lagrangian hydrodynamics. *SIAM J. Sci. Comput.* **39** (2017) A2317–A2344.

[3] R. Anderson, V. Dobrev, Tz. Kolev, D. Kuzmin, M. Quezada de Luna, R. Rieben, and V. Tomov, High-order local maximum principle preserving (MPP) discontinuous Galerkin finite element method for the transport equation. *J. Comput. Phys.* **334** (2017) 102–124.

[4] P. Azerad, J.-L. Guermond, and B. Popov, Well-balanced second-order approximation of the shallow water equation with continuous finite elements. *SIAM J. Numer. Anal.* **55** (2017) 3203–3224.

[5] P. Bacigaluppi, R. Abgrall, and S. Tokareva, "A posteriori" limited high order and robust residual distribution schemes for transient simulations of fluid flows in gas dynamics. Preprint `arXiv:1902.07773 [math.NA]`, 2019.

[6] S. Badia and J. Bonilla, Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. *Computer Methods Appl. Mech. Engrg.* **313** (2017) 133-158.

[7] S. Badia, J. Bonilla, and A. Hierro, Differentiable monotonicity-preserving schemes for discontinuous Galerkin methods on arbitrary meshes. *Computer Methods Appl. Mech. Engrg.* **320** (2017 582-605.

[8] G. Barrenechea, E. Burman, and F. Karakatsani, Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. *Numer. Math.* **135** (2017) 521–545.

[9] G. Barrenechea, V. John, and P. Knobloch, Analysis of algebraic flux correction schemes. *SIAM J. Numer. Anal.* **54** (2016) 2427–2451.

[10] G. Barrenechea, V. John, and P. Knobloch, A linearity preserving algebraic flux correction scheme satisfying the discrete maximum principle on general meshes. *Mathematical Models and Methods in Applied Sciences (M3AS)* **27** (2017) 525–548.

[11] G. Barrenechea, V. John, P. Knobloch, and R. Rankin, A unified analysis of algebraic flux correction schemes for convection-diffusion equations. *SeMA* **75** (2018) 655–685.

[12] G. Barrenechea and P. Knobloch, Analysis of a group finite element formulation. *Applied Numerical Mathematics* **118** (2017) 238–248.

[13] T. Barth and D.C. Jespersen, The design and application of upwind schemes on unstructured meshes. *AIAA Paper*, 89-0366, 1989.

[14] P. Bochev, D. Ridzal, M. D'Elia, M. Perego, and K. Peterson, Optimization-based, property-preserving finite element methods for scalar advection equations and their connection to Algebraic Flux Correction. Preprint, 2019, `https://doi.org/10.13140/RG.2.2.20942.72000`.

[15] J.P. Boris and D.L. Book, Flux-Corrected Transport: I. SHASTA, a fluid transport algorithm that works. *J. Comput. Phys.* **11** (1973) 38–69.

[16] C.J. Cotter and D. Kuzmin, Embedded discontinuous Galerkin transport schemes with localised limiters. *J. Comput. Phys.* **311** (2016) 363–373.

[17] S. Diot, S. Clain, and R. Loubére, Improved detection criteria for the Multi-dimensional Optimal Order Detection (MOOD) on unstructured meshes with very high-order polynomials. *Computers & Fluids* **64** (2012) 43–63.

[18] V. Dobrev, Tz. Kolev, D. Kuzmin, R. Rieben, and V. Tomov, Sequential limiting in continuous and discontinuous Galerkin methods for the Euler equations. *J. Comput. Phys.* **356** (2018) 372–390.

[19] M. Dumbser, O. Zanotti, R. Loubère, and S. Diot, A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *J. Comput. Phys.* **278** (2014) 47–75.

[20] C.A.J. Fletcher, The group finite element formulation, *Comput. Methods Appl. Mech. Engrg.* **37** (1983) 225–243.

[21] C.A.J. Fletcher, A comparison of finite element and finite difference solutions of the one- and two-dimensional Burgers' equations. *J. Comput. Phys.* **51** (1983) 159–188.

[22] S.K. Godunov, Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mat. Sb.* **47** (1959) 271–306.

[23] S. Gottlieb, C.-W. Shu, and E. Tadmor, Strong stability-preserving high-order time discretization methods. *SIAM Review* **43** (2001) 89–112.

[24] J.-L. Guermond and M. Nazarov, A maximum-principle preserving $C^0$ finite element method for scalar conservation equations. *Computer Methods Appl. Mech. Engrg.* **272** (2014) 198–213.

[25] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas, Second-order invariant domain preserving approximation of the Euler equations using convex limiting. *SIAM J. Sci. Computing* **40** (2018) A3211-A3239.

[26] J.-L. Guermond, M. Nazarov, B. Popov, and Y. Yang, A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. *SIAM J. Numer. Anal.* **52** (2014) 2163–2182.

[27] J.-L. Guermond, M. Nazarov, and I. Tomas, Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems. *Computer Methods Appl. Mech. Engrg.* **347** (2019) 143–175.

[28] J.-L. Guermond and B. Popov, Invariant domains and first-order continuous finite element approximation for hyperbolic systems. *SIAM J. Numer. Anal.* **54** (2016) 2466–2489.

[29] J.-L. Guermond and B. Popov, Invariant domains and second-order continuous finite element approximation for scalar conservation equations. *SIAM J. Numer. Anal.* **55** (2017) 3120–3146.

[30] J.-L. Guermond, M. Quezada de Luna, B. Popov, C.E. Kees, and M.W. Farthing, Well-balanced second-order finite element approximation of the shallow water equations with friction. *SIAM J. Sci. Comput.* **40** (2018) A3873-A3901.

[31] H. Hajduk, D. Kuzmin, and V. Aizinger, New directional vector limiters for discontinuous Galerkin methods. *J. Comput. Phys.* **384** (2019) 308–325.

[32] H. Hajduk, D. Kuzmin, Tz. Kolev, and R. Abgrall, Matrix-free subcell residual distribution for Bernstein finite elements: Low-order schemes and FCT. Sbumitted to *Computer Methods Appl. Mech. Engrg.* Preprint: *Ergebnisber. Inst. Angew. Math.* **598** TU Dortmund, 2019.

[33] A. Harten, High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.* **49** (1983) 357–393.

[34] A. Harten, On a class of high resolution total-variation-stable finite-difference-schemes. *SIAM J. Numer. Anal.* **21** (1984) 1-23.

[35] A. Jameson, Computational algorithms for aerodynamic analysis and design. *Appl. Numer. Math.* **13** (1993) 383–422.

[36] A. Jameson, Positive schemes and shock modelling for compressible flows. *Int. J. Numer. Methods Fluids* **20** (1995) 743-776.

[37] A. Jameson, Analysis and design of numerical schemes for gas dynamics 1. Artificial diffusion, upwind biasing, limiters and their effect on accuracy and multigrid convergence. *Int. Journal of CFD* **4** (1995) 171–218.

[38] V. John and E. Schmeyer, On finite element methods for 3D time-dependent convection-diffusion-reaction equations with small diffusion. *Comput. Meth. Appl. Mech. Engrg.* **198** (2008) 475–494.

[39] P. Knobloch, On the discrete maximum principle for algebraic flux correction schemes with limiters of upwind type. In: Z. Huang, M. Stynes, and Z. Zhang (eds), *Boundary and Interior Layers, Computational and Asymptotic Methods* (Proceedings of the BAIL 2016 conference). Springer International Publishing, 2017, pp. 129–139.

[40] D. Kuzmin, Gradient-based limiting and stabilization of continuous Galerkin methods. *Ergebnisber. Angew. Math.* **589**, TU Dortmund University, 2018. To appear in: G. Rozza et al. (eds), LNCSE Springer FEF 2017 Special Volume (2019).

[41] D. Kuzmin, Explicit and implicit FEM-FCT algorithms with flux linearization. *J. Comput. Phys.* **228** (2009) 2517–2534.

[42] D. Kuzmin, Algebraic flux correction I. Scalar conservation laws. In: D. Kuzmin, R. Löhner and S. Turek (eds.) *Flux-Corrected Transport: Principles, Algorithms, and Applications*. Springer, 2nd edition: 145–192 (2012).

[43] D. Kuzmin, M. Möller, and M. Gurris, Algebraic flux correction II. Compressible flow problems. In: D. Kuzmin, R. Löhner, S. Turek (eds), *Flux-Corrected Transport: Principles, Algorithms, and Applications*. Springer, 2nd edition, 2012, pp. 193–238.

[44] D. Kuzmin and N. Klyushnev, Limiting and divergence cleaning for continuous finite element discretizations of the MHD equations. Submitted to *J. Comput. Phys.* Preprint *Ergebnisber. Inst. Angew. Math.* **608**, TU Dortmund, 2019.

[45] D. Kuzmin, M. Quezada de Luna and C. Kees, A partition of unity approach to adaptivity and limiting in continuous finite element methods. *Computers & Mathematics with Applications* **78** (2019) 944–957.

[46] D. Kuzmin and S. Turek, Flux correction tools for finite elements. *J. Comput. Phys.* **175** (2002) 525–558.

[47] R.J. LeVeque, High-resolution conservative algorithms for advection in incompressible flow. *SIAM Journal on Numerical Analysis* **33**, (1996) 627–665.

[48] R. Löhner, K. Morgan, J. Peraire, and M. Vahdati, Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations. *Int. J. Numer. Meth. Fluids* **7** (1987) 1093–1109.

[49] C. Lohmann, Flux-corrected transport algorithms preserving the eigenvalue range of symmetric tensor quantities. *J. Comput. Phys.* **350** (2017) 907–926.

[50] C. Lohmann, Eigenvalue range limiters for tensors in flux-corrected transport algorithms. Presentation given at the MultiMat 2017 Conference on September 20, 2017 in Santa Fe, USA. Slides available online at `https://custom.cvent.com/F6288ADDEF3C4A6CBA5358DAE922C966/files/e4c3aedf74394eb1a33e141f57f33b2e.pdf`

[51] C. Lohmann, Algebraic flux correction schemes preserving the eigenvalue range of symmetric tensor fields. *ESAIM: M2AN* **53** (2019) 833-867.

[52] C. Lohmann, *Physics-Compatible Finite Element Methods for Scalar and Tensorial Advection Problems.* PhD thesis, TU Dortmund University, 2019.

[53] C. Lohmann and D. Kuzmin, Synchronized flux limiting for gas dynamics variables. *J. Comput. Phys.* **326** (2016) 973–990.

[54] C. Lohmann, D. Kuzmin, J.N. Shadid, and S. Mabuza, Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements. *J. Comput. Phys.* **344** (2017) 151-186.

[55] H. Luo, J.D. Baum, and R. Löhner, Edge-based finite element scheme for the Euler equations. *AIAA Journal* **32** (1994) 1183–1190.

[56] G. Luttwak and J. Falcovitz, Slope limiting for vectors: A novel vector limiting algorithm. *Int. J. Numer. Methods Fluids* **65** (2011) 1365–1375.

[57] G. Luttwak and J. Falcovitz, VIP (Vector Image Polygon) multidimensional slope limiters for scalar variables. *Computers & Fluids* **83** (2013) 90-97.

[58] P.R.M. Lyra, *Unstructured Grid Adaptive Algorithms for Fluid Dynamics and Heat Conduction.* PhD thesis, University of Wales, Swansea, 1994.

[59] P.R.M. Lyra and K. Morgan, A review and comparative study of upwind biased schemes for compressible flow computation. III: Multidimensional extension on unstructured grids. *Arch. Comput. Methods Eng.* **9**:3 (2002) 207–256.

[60] P.R.M. Lyra, K. Morgan, J. Peraire, and J. Peiro, TVD algorithms for the solution of the compressible Euler equations on unstructured meshes. *Int. J. Numer. Methods Fluids* **19** (1994) 827–847.

[61] S. Mabuza, J.N. Shadid, and D. Kuzmin, Local bounds preserving stabilization for continuous Galerkin discretization of hyperbolic systems. *J. Comput. Phys.* **361** (2018) 82-110.

[62] J. Peraire, M. Vahdati, J. Peiro, and K. Morgan, The construction and behaviour of some unstructured grid algorithms for compressible flows. *Numerical Methods for Fluid Dynamics* **IV**, Oxford University Press, 1993, 221-239.

[63] V. Selmin, Finite element solution of hyperbolic equations. I. One-dimensional case. *INRIA Research Report* **655**, 1987.

[64] V. Selmin, Finite element solution of hyperbolic equations. II. Two-dimensional case. *INRIA Research Report* **708**, 1987.

[65] V. Selmin, The node-centred finite volume approach: bridge between finite differences and finite elements. *Comput. Methods Appl. Mech. Engrg.* **102** (1993) 107–138.

[66] V. Selmin and L. Formaggia, Unified construction of finite element and finite volume discretizations for compressible flows. *Int. J. Numer. Methods Engrg.* **39** (1996) 1–32.

[67] G. Sod, A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *J. Comput. Phys.* **27** (1978) 1–31.

[68] F. Vilar, A posteriori correction of high-order discontinuous Galerkin scheme through subcell finite volume formulation and flux reconstruction. *J. Comput. Phys.* **387** (2019) 245–279.

[69] P.R. Woodward and P. Colella, The numerical simulation of two-dimensional fluid flow with strong shocks. *J. Comput. Phys.* **54** (1984) 115–173.

[70] S.T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.* **31** (1979) 335–362.

[71] S.T. Zalesak, A preliminary comparison of modern shock-capturing schemes: linear advection. In: R. Vichnevetsky and R. Stepleman (eds), *Advances in Computer Methods for PDEs.* Publ. IMACS, 1987, 15–22.