Hennes Hajduk*

Improvements of algebraic flux-correction schemes based on Bernstein finite elements

https://doi.org/10.1515/jnma-2024-0098
Received July 8, 2024; accepted December 9, 2024; published online June 26, 2025

Abstract: In Galerkin finite element schemes, the discrete first derivative operator for each spatial dimension is a square matrix that is skew-symmetric under restrictive assumptions for certain types of discretizations and boundary conditions. In most settings, however, this desirable property is violated, often only for a few pairs of nodes. These exceptions can invalidate certain design principles based on the skew-symmetry assumption made for these operators. This paper demonstrates that algebraic manipulations can be performed to make the discrete gradient operators of Bernstein polynomial-based finite element methods skew symmetric. Interest in such discretizations has recently been increasing because they represent natural extensions of second-order algebraic flux correction schemes to higher-order spaces. We employ the new operators in the context of such property-preserving methods, mostly based on discontinuous Galerkin discretizations of arbitrary order. Additional theoretical results for the schemes under investigation are derived, including local and global entropy inequalities, among others. Moreover, a discussion on the optimality of CFL-like time step restrictions arising in explicit Runge–Kutta methods shows that our new approach is superior to earlier representatives of operators employed in similar contexts. These techniques use the monolithic convex limiting paradigm and are applied to the compressible Euler equations.

Keywords: algebraic flux correction; hyperbolic conservation laws; skew-symmetric discrete gradients; entropy stability; Bernstein finite elements; monolithic convex limiting

MSC 2010 Classification: 65M60

1 Introduction

Hyperbolic problems are commonly solved numerically using high-resolution schemes. These encompass stabilization and limiting techniques designed for the purpose of obtaining physics-conforming approximations are, for instance, residual distribution schemes, slope and a posteriori limiters, entropy-based semi-discrete approaches, smoothness indicators, and/or weighted essentially non-oscillatory (WENO) techniques, see e.g., [1]–[6]. For an overview of property-preservation, we refer to the book [7]. Most of the techniques mentioned here work by blending a certain high-order target scheme with a low-order counterpart to be used in the vicinity of steep gradients [8]. The main question is how to choose the numerical viscosity locally to preserve accuracy in smooth regions but avoid formation of spurious numerical solutions. These features include violations of nonnegativity constraints, Gibbs phenomena, and/or nonentropic results.

In this work, we focus on algebraic flux correction (AFC) schemes, e.g., [9]–[14] because they are *failsafe*, in the sense that it is possible to guarantee the validity of various desirable constraints, e.g., the nonnegativity of certain quantities. Moreover, they are prone to numerical analysis, e.g., [7], [15]–[17] and can also be designed in a way that a multitude of desirable properties are guaranteed. Usual properties include discrete maximum

***Corresponding author: Hennes Hajduk**, Department of Geosciences, University of Oslo, Oslo, Norway,
E-mail: hennes.hajduk@geo.uio.no

principles [14], nonnegativity constraints [13], entropy stability conditions [18], and well-balancedness [12]. The latter reference presents a scheme that possesses all of these features at once. As such, it is unique among the families of high-resolution schemes. Because of such progress, additional improvements of AFC schemes are called for, and this work is aimed at that purpose.

The first work on algebraic flux correction schemes [19] has spawned many further developments and theoretical analyses [10], [11], [13]–[15], [17], [20], [21]. For instance, AFC techniques based on high-order Bernstein polynomials are presented in refs. [10], [14] for discontinuous Galerkin (DG) methods and classical finite elements, respectively. Combining the theoretical frameworks developed in the seminal papers by Hoff [22] and Harten et al. [23], Guermond and Popov [24] propose a new way of analyzing the low-order method, which provides sufficient numerical dissipation to guarantee the failsafe property of the scheme. In ref. [11] and related works, this approach is used in the predictor stage of the proposed *convex limiting* techniques. Kuzmin's monolithic convex limiting (MCL) paradigm [13] is no representative of such FCT approaches. Instead, it operates on the semi-discrete level of spatial discretizations, allowing for steady-state calculations [13], [25], [26], implicit time stepping [26]–[28], and semi-discrete entropy stabilization [18], [28], [29] based on Tadmor's entropy stability theory [30], [31]. MCL techniques have since been further developed, e.g., [18], [21], [25], [32] and analyzed [7], [17]. Following the work of Lohmann et al. [14] on high-order FCT schemes for finite elements based on Bernstein basis functions, Kuzmin and Quezada de Luna [32] first developed a high-order counterpart of MCL and later introduced entropy limiting based on Tadmor's criterion into these algorithms [18], [29]–[31]. Subsequently, we extended the approaches in refs. [13], [32] to Bernstein-DG discretizations [25]. Rueda-Ramírez et al. [21] use MCL techniques in combination with Legendre–Gauss–Lobatto (LGL) bases for DG. These authors were the first who enforced Tadmor's entropy stability criterion [30], [31] in the AFC-DG context to stabilize entropy-producing terms arising from volume integrals. In ref. [21], numerical fluxes across interior boundaries are of local Lax–Friedrichs (LLF) type. In contrast, here and in ref. [28, Ch. 6], an arbitrary flux is blended with such an LLF counterpart and entropy limiting can also be performed for blended interfacial fluxes.

The present article addresses the following issues: The employed discrete gradient operators in AFC schemes are generally not skew-symmetric as a result of the type of discretization and/or due to using other than periodic boundary conditions. Nevertheless, the assumption of skew-symmetry is often made for theoretical purposes [27, Rem. 4] and in the design of flux limiters [33]. The concept of summation-by-parts, often invoked in the LGL-DG context [20], [21], [34], [35], yields skew-symmetric discrete gradients seemingly for free, which suggests that, in this sense, LGL schemes are superior to their Bernstein polynomial-based counterparts [14], [25], [32]. However, the latter produce significantly less restrictive CFL conditions because the Bernstein nodes are uniformly distributed within the elements as opposed to the LGL nodes. Thus, it is worthwhile to adapt the LGL-summation-by-parts concept to high-order convex limiting schemes based on the Bernstein basis. A main aspect of this work is dedicated to this effort. Bernstein basis functions possess many further desirable properties (see [14], [25], [32] and the references therein) such as nonnegativity and boundedness of local approximations by the min- and max-values of nodal values within the elements, which can be exploited for mathematical purposes. We also demonstrate that a supposedly entropy-stable AFC limiter can actually produce entropy instead of dissipating it if the non-skew-symmetric version is used instead of our new one. Additionally, this work presents novel theoretical results for monolithic flux correction schemes applied to hyperbolic problems. In this regard, the present paper can be seen as a follow-up to ref. [25], again focusing on DG but second-order continuous Galerkin schemes are also discussed.

2 Preliminaries

2.1 Model problem

Let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, $\boldsymbol{x} \in \Omega$, $t \geqslant 0$, and let $u(\boldsymbol{x}, t) \in \mathbb{R}^M$, $M \in \mathbb{N}$, be the solution to the conservation law

$$\frac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{f}(u) = 0 \qquad \text{in } \Omega \times (0, \infty), \tag{1}$$

where $\nabla\cdot$ is the matrix-divergence operator applied to the inviscid flux $\boldsymbol{f} = \boldsymbol{f}(u) \in \mathbb{R}^{M\times d}$. To formulate a well-posed initial-boundary value problem, (1) has to be equipped with the initial condition $u(\cdot, 0) = u_0$ and suitable data $u_0 = u_0(\boldsymbol{x}) \in \mathbb{R}^M$ as well as appropriate boundary conditions. The latter are specified for the test problems in Section 6. For many specific applications (e.g., scalar problems, shallow water equations, gas dynamics), the exact solution to (1) is known to satisfy certain admissibility constraints that can be described by constraining u to lie in a some convex set $\mathcal{A} \subset \mathbb{R}^M$. For instance, in the scalar case, $M = 1$, $\mathcal{A}$ is an interval bounded by the extrema of initial and boundary conditions [36, Ch. 6]. Suppose there exists a convex function $U\colon \mathcal{A} \to \mathbb{R}$ and a corresponding flux $\boldsymbol{F}\colon \mathcal{A} \to \mathbb{R}^d$ such that $\boldsymbol{F}'(u) = U'(u)^\top \boldsymbol{f}'(u)$, then $(U, \boldsymbol{F})$ is called an entropy pair for the conservation law (1). By the chain rule, a conservation law for the *entropy* U can be derived under the assumption that u and U are smooth. In general however, weak solutions to the nonlinear problem (1) can develop discontinuities in finite time. Using, for instance, the concept of vanishing viscosity solutions [36, Sect. 6.3], one can show well-posedness of weak entropy solutions for the scalar case, if a weak form of the entropy inequality

$$\frac{\partial U(u)}{\partial t} + \nabla\cdot \boldsymbol{F}(u) \leqslant 0 \qquad \text{in } \Omega\times(0,\infty) \tag{2}$$

holds for all $(U, \boldsymbol{F})$ that are entropy pairs for (1). If discrete versions of (2) are satisfied, ideally in a localized manner, the occurrence of nonphysical weak solutions can be averted in numerical approximations, which motivates the use of entropy-stable approximations, e.g., [3], [4], [7], [12], [21].

2.2 Generic monolithic convex limiting discretization

We now summarize the concept of monolithic convex limiting (MCL), first proposed by Kuzmin [13], see also [7], [18], [21], [25], [32]. This algebraic flux correction (AFC) paradigm operates on the semi-discrete level and is capable of enforcing a variety of constraints. A generic MCL semi-discretization can be written as

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j\in\mathcal{N}_i\backslash\{i\}} 2d_{ij}\Big(\bar{u}_{ij}^* - u_i\Big). \tag{3}$$

Here the index $i \in \{1,\dots,N\}$ refers to a node, N is the number of unknowns for each variable, and thus the total number of unknowns is MN. Note that in the system case all components of the solution vector $\boldsymbol{u}$ are interpolated on the same set of nodes, which are associated with a certain *node* $\boldsymbol{x}_i \in \bar{\Omega}$, i.e., we do not consider staggered approaches. Similarly, $m_i > 0$ is associated with this node and typically represents an entry of a diagonalized mass matrix. The set $\mathcal{N}_i \subset \{1,\dots,N\}$ is the nodal stencil of $\boldsymbol{x}_i$, i.e., the set of nodes $\boldsymbol{x}_j$ that directly interact with $\boldsymbol{x}_i$. In piecewise linear continuous finite element methods, $\mathcal{N}_i$ is the set of mesh vertices that are nearest neighbors to node $\boldsymbol{x}_i$ in the sense that they are endpoints of a certain mesh edge. Entities featuring two distinct indices such as the symmetric artificial diffusion coefficient $d_{ij} = d_{ji} \geqslant 0$ refer to terms depending on the two nodal values u_i and u_j. In the MCL framework, the *flux-corrected bar states* $\bar{u}_{ij}^*$ [13] are algebraically enforced to satisfy user-defined constraints of convex nature. Upon temporal discretization of (3) with a strong-stability preserving (SSP) Runge–Kutta (RK) method [37]–[39], Shu–Osher updates can be written as a convex combination of forward Euler steps

$$u_i^{\mathrm{FE}} = \left(1 - \frac{\tau}{m_i}\sum_{j\in\mathcal{N}_i\backslash\{i\}} 2d_{ij}\right)u_i + \frac{\tau}{m_i}\sum_{j\in\mathcal{N}_i\backslash\{i\}} 2d_{ij}\bar{u}_{ij}^*, \tag{4}$$

where $\tau > 0$ is the time step. Suppose that $u_i \in \mathcal{A}_i$ for a certain convex subset $\mathcal{A}_i \subset \mathcal{A}$ of the largest admissible set. If $\bar{u}_{ij}^* \in \mathcal{A}_i$ can be guaranteed for $j \in \mathcal{N}_i\backslash\{i\}$ and the CFL-like time step restriction

$$\tau \leqslant \min_{i\in\{1,\dots,N\}} \frac{m_i}{\sum_{j\in\mathcal{N}_i\backslash\{i\}} 2d_{ij}} \tag{5}$$

holds, then u_i^{FE} is a convex combination of the solution at the previous iteration u_i and the flux-corrected bar states $\bar{u}_{ij}^*$. In conclusion, by convexity of $\mathcal{A}_i$, we then have $u_i^{\mathrm{FE}} \in \mathcal{A}_i$ as well. Recent works, e.g., [33], [40], also demonstrate how AFC and MCL can be performed using non-SSP time discretizations, including implicit methods

and applications to viscous problems. Although they appear promising, these issues shall not be addressed here since our focus lies on the spatial discretization. The particular choice of the sets $\mathcal{A}_i$ and how to enforce $\bar{u}_{ij}^* \in \mathcal{A}_i$ for all $j \in \mathcal{N}_i \backslash \{i\}$ sets different MCL techniques apart from each other. Before discussing high-order DG discretizations, we give a brief example of arguably the simplest MCL variant.

2.3 Case study: MCL based on classical finite elements for the Euler equations

We only summarize the steady case here. More information can be found in ref. [7] and the references therein. Let $\mathbf{I}$ be the $d \times d$-identity matrix and let $\rho = \rho(\boldsymbol{x}, t) \in \mathbb{R}$, $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t) \in \mathbb{R}^d$, $p = p(\boldsymbol{x}, t) \in \mathbb{R}$ denote density, velocity, and pressure of an ideal polytropic gas. We consider the compressible Euler equations

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0, \tag{6a}$$

$$\frac{\partial (\rho \boldsymbol{v})}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}\boldsymbol{v}^\top + p\, \mathbf{I}) = 0, \tag{6b}$$

$$\frac{\partial (\rho E)}{\partial t} + \nabla \cdot \big((\rho E + p)\boldsymbol{v}\big) = 0, \tag{6c}$$

where $u = [\rho, \rho \boldsymbol{v}^\top, \rho E]^\top$ is the vector of conserved unknowns (density, momentum, and total energy), and

$$p = (\gamma - 1)\left(\rho E - \frac{\rho |\boldsymbol{v}|^2}{2}\right) = (\gamma - 1)\rho e, \tag{7}$$

where $\gamma > 1$ is the adiabatic constant and e is the specific internal energy. The (physically motivated) largest admissible set for (6) and (7) is $\mathcal{A} = \{u = [\rho, \rho \boldsymbol{v}^\top, \rho E]^\top \in \mathbb{R}^{d+2} \colon \rho \geqslant 0,\ \ p \geqslant 0\}$. Note that due to (7), nonnegativity of pressure is equivalent to the physically motivated constraint that internal energy e may not become negative. The usual entropy pair for (6) and (7) is given by $(U, \boldsymbol{F})$ with $U = -\rho \log(p\rho^{-\gamma})$, $\boldsymbol{F} = \boldsymbol{v}U$ [41].

Let us discretize (1) using a conforming simplicial mesh and local $\mathbb{P}_1$-polynomials with corresponding Lagrange basis functions $\varphi_i \in C(\bar{\Omega})$, $i \in \{1, \dots, N\}$, defined implicitly by the interpolation property $\varphi_i(\boldsymbol{x}_j) = \delta_{ij}$ for all $i, j \in \{1, \dots, N\}$. We then set $m_i = \int_\Omega \varphi_i \, \mathrm{d}\boldsymbol{x}$, $\boldsymbol{c}_{ij} = \int_\Omega \varphi_i \nabla \varphi_j \, \mathrm{d}\boldsymbol{x}$ and

$$\lambda_{ij} = \lambda(u_i, u_j, \boldsymbol{c}_{ij}/|\boldsymbol{c}_{ij}|) := \max\left\{ \left|\boldsymbol{v}_i \cdot \boldsymbol{c}_{ij}/\left|\boldsymbol{c}_{ij}\right|\right| + \sqrt{\gamma\, p_i/\rho_i},\ \left|\boldsymbol{v}_j \cdot \boldsymbol{c}_{ij}/\left|\boldsymbol{c}_{ij}\right|\right| + \sqrt{\gamma\, p_j/\rho_j} \right\}. \tag{8}$$

For a unit vector $\boldsymbol{n}_{ij} = \boldsymbol{c}_{ij}/|\boldsymbol{c}_{ij}|$, (8) corresponds to the largest absolute eigenvalue of the Jacobian of $\boldsymbol{f}(u_i)$ or $\boldsymbol{f}(u_j)$ projected onto $\boldsymbol{n}_{ij}$. This estimate for the exact wave speed does not produce states outside of the set $\mathcal{A}$ [42, App.], which is why we use (8) to define the diffusion coefficients as follows [24], [43], [44]:

$$d_{ij} = \max\{\lambda_{ij}|\boldsymbol{c}_{ij}|, \lambda_{ji}|\boldsymbol{c}_{ji}|\} \quad \forall i, j \in \{1, \dots, N\}, \quad i \neq j. \tag{9}$$

Next, we define the low-order bar states $\bar{u}_{ij} = \frac{u_i + u_j}{2} + \frac{(\boldsymbol{f}_i - \boldsymbol{f}_j)\, \boldsymbol{c}_{ij}}{2d_{ij}}$ [23], [24] and their MCL counterparts [13]:

$$\bar{u}_{ij}^* = \bar{u}_{ij} + \frac{f_{ij}^*}{2d_{ij}}, \tag{10}$$

where $\boldsymbol{f}_i := \boldsymbol{f}(u_i)$ and $f_{ij}^* = -f_{ji}^*$ is a flux-corrected counterpart of the *target flux* f_{ij}, which for steady-state calculations can simply be set to $f_{ij} = d_{ij}(u_i - u_j)$. The low-order counterpart $\bar{u}_{ij}$ of $\bar{u}_{ij}^*$ initially appeared in Harten et al. [23]. Guermond and Popov [24] were the first to use it in the AFC context to analyze the low-order method that is the first-order predictor in FCT-type convex limiting schemes and, similarly, the low-order version of MCL approaches. In either case, the limited antidiffusive fluxes f_{ij}^* are set to zero.

In general, the least-restrictive constraint that should be enforced for any hyperbolic problem is the admissibility of solutions, i.e., $u \in \mathcal{A}$. For the compressible Euler equations, we need to ensure that densities remain nonnegative, which can be achieved by enforcing positivity of the first component of $\bar{u}_{ij}^* = \left[\bar{\rho}_{ij}^*, \overline{(\rho \boldsymbol{v})}_{ij}^*, \overline{(\rho E)}_{ij}^*\right]$. Using definition (10) and skew symmetry of f_{ij}^*, this task is accomplished by setting the first component of f_{ij}^* to $\max\left\{-2d_{ij}\bar{\rho}_{ij}, \min\{f_{ij}^{\rho}, 2d_{ij}\bar{\rho}_{ji}\}\right\}$, where $\bar{\rho}_{ij}$ and f_{ij}^{ρ} correspond to the first component of the low-order bar state $\bar{u}_{ij}$ and the unlimited flux f_{ij}, respectively. Next, we can enforce additional constraints such as nonnegativity of pressure and Tadmor's entropy stability condition [30], [31] by further reducing the magnitude of f_{ij}^*. These issues are described in detail in refs. [13], [18], [27], respectively, see also [7]. Additionally, the quality of numerical solutions can be improved by enforcing discrete maximum principles in addition to nonnegativity, see, e.g., [7], [10], [16], [21], [28], [40]. All of these tasks (and more, for instance, well-balancedness in the case of the shallow water equations [12]) can be guaranteed simultaneously.

Since nonnegativity of pressure is a nonnegotiable constraint (as is nonnegativity of density), we briefly summarize the typical MCL pressure fix [13], [21], [25]. For a pair of nodes $i \neq j$ let f_{ij}^* be the prelimited antidiffusive flux. Setting $f_{ij}^{**} = \alpha_{ij} f_{ij}^*$, where the correction factor $\alpha_{ij} = \alpha_{ji} \in [0,1]$ is defined via [13]:

$$\alpha_{ij} = \begin{cases} \dfrac{Q_{ij}}{R_{ij}} & \text{if } R_{ij} > Q_{ij}, \\ 1 & \text{otherwise,} \end{cases} \tag{11}$$

where $\bar{w}_{ij} = 2d_{ij}\bar{u}_{ij} = \left[\bar{w}_{ij}^{\rho}, \left(\bar{\boldsymbol{w}}_{ij}^{\rho \boldsymbol{v}}\right)^{\top}, \bar{w}_{ij}^{\rho E}\right]^{\top}$, $f_{ij}^* = \left[f_{ij}^{\rho,*}, \left(\boldsymbol{f}_{ij}^{\rho \boldsymbol{v},*}\right)^{\top}, f_{ij}^{\rho E,*}\right]^{\top}$, and

$$\begin{aligned} Q_{ij} = Q_{ji} &= \min\left\{ \bar{w}_{ij}^{\rho}\bar{w}_{ij}^{\rho E} - \frac{1}{2}|\bar{\boldsymbol{w}}_{ij}^{\rho \boldsymbol{v}}|^2,\ \bar{w}_{ji}^{\rho}\bar{w}_{ji}^{\rho E} - \frac{1}{2}|\bar{\boldsymbol{w}}_{ji}^{\rho \boldsymbol{v}}|^2 \right\}, \\ R_{ij} = R_{ji} &= \max\left\{ |\bar{\boldsymbol{w}}_{ij}^{\rho \boldsymbol{v}}|,\ |\bar{\boldsymbol{w}}_{ji}^{\rho \boldsymbol{v}}| \right\} |\boldsymbol{f}_{ij}^{\rho \boldsymbol{v},*}| + \max\left\{ \bar{w}_{ij}^{\rho},\ \bar{w}_{ji}^{\rho} \right\} |f_{ij}^{\rho E,*}| \\ &\quad + \max\left\{ \bar{w}_{ij}^{\rho E},\ \bar{w}_{ji}^{\rho E} \right\} |f_{ij}^{\rho,*}| + \max\left\{ 0,\ \frac{1}{2}|\boldsymbol{f}_{ij}^{\rho \boldsymbol{v},*}|^2 - f_{ij}^{\rho,*} f_{ij}^{\rho E,*} \right\} \end{aligned}$$

guarantees that the pressure of the bar state $\bar{u}_{ij}^{**} = \bar{u}_{ij} + f_{ij}^{**}/(2d_{ij})$ remains nonnegative. The derivation of (11) can be found in refs. [13], [28]. The following result is an interesting observation made in the process of this work.

Lemma 1 (pressure fix yields nonnegative density). *Let $u_i,\ u_j \in \mathcal{A}$ be arbitrary, $\bar{u}_{ij}^{**} = \bar{u}_{ij} + f_{ij}^{**}/(2d_{ij})$, where d_{ij} is given by* (9)*, $f_{ij}^{**} = \alpha_{ij} f_{ij}^*$, and let α_{ij} be defined by* (11)*. Then the first component of $\bar{u}_{ij}^{**}$ is nonnegative.*

Proof. For α_{ij} given by [13, Eq. (92)] or (11), the pressures of the states $\bar{u}_{ij}^{**} = \bar{u}_{ij} + \alpha_{ij} f_{ij}^*/(2d_{ij})$ and $\bar{u}_{ji}^{**} = \bar{u}_{ji} - \alpha_{ij} f_{ij}^*/(2d_{ij})$ are nonnegative [13]. The density and total energy components of the low-order bar states $\bar{u}_{ij},\ \bar{u}_{ji}$ are also nonnegative [24, Lem. 2.1], see also [23]. Thus, only one of the two terms $\bar{w}_{ij}^{\rho} + \alpha_{ij} f_{ij}^{\rho,*}$ or $\bar{w}_{ji}^{\rho} - \alpha_{ij} f_{ij}^{\rho,*}$ can potentially become negative. Let us first consider the case $f_{ij}^{\rho,*} < 0$, where

$$-\alpha_{ij} f_{ij}^{\rho,*} \leqslant \frac{Q_{ij}}{R_{ij}} |f_{ij}^{\rho,*}| \leqslant \frac{\bar{w}_{ij}^{\rho}\bar{w}_{ij}^{\rho E} - \frac{1}{2}|\bar{\boldsymbol{w}}_{ij}^{\rho \boldsymbol{v}}|^2}{\bar{w}_{ij}^{\rho E}|f_{ij}^{\rho,*}|} |f_{ij}^{\rho,*}| = \bar{w}_{ij}^{\rho} - \frac{|\bar{\boldsymbol{w}}_{ij}^{\rho \boldsymbol{v}}|^2}{2\bar{w}_{ij}^{\rho E}} \leqslant \bar{w}_{ij}^{\rho}$$

since $\bar{w}_{ij}^{\rho E} \geqslant 0$. Thus, the first component of $\bar{w}_{ij}^{**} = 2d_{ij}\bar{u}_{ij}^{**}$, $\bar{w}_{ij}^{\rho} + \alpha_{ij} f_{ij}^{\rho} \geqslant 0$. The case $f_{ij}^{\rho} > 0$ is similar. □

3 Skew-symmetry of discrete gradient operators

3.1 General considerations and second-order continuous Galerkin schemes

Let us once more consider a $\mathbb{P}_1$ or $\mathbb{Q}_1$ continuous Galerkin discretization as described in Section 2.3. If the domain has only periodic boundaries, then each component C^k of the discrete gradient operator $\boldsymbol{C} = (C^1, \dots, C^d) = (\boldsymbol{c}_{ij})_{i,j=1}^N$ is a skew-symmetric $N \times N$ matrix due to the divergence theorem. Otherwise,

$$\boldsymbol{c}_{ij} = \int_\Omega \varphi_i \nabla \varphi_j \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \nabla\varphi_i \varphi_j \,\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} \varphi_i\varphi_j \boldsymbol{n} \,\mathrm{d}s =: -\boldsymbol{c}_{ji} + \boldsymbol{b}_{ij},$$

where $\boldsymbol{n} = \boldsymbol{n}(\boldsymbol{x}) \in \mathbb{R}^d$ denotes the unit outward normal to $\partial\Omega$. In this setting, each matrix C^k, $k \in \{1, \dots, d\}$, is almost skew-symmetric, with only a few pairs of entries where $\boldsymbol{b}_{ij} \neq 0$ being exceptions to this rule. For flux-correction schemes, the skew symmetry of $\boldsymbol{C}$ is often desirable to design new algorithms [33], simplify computations (note that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ implies $\lambda_{ij} = \lambda_{ji}$ in (8) (the computation of λ_{ij} is a significant cost factor in AFC schemes for systems), and to prove certain theoretical results such as entropy stability [18], see also Section 4.3.2. Therefore, it is unfortunate that a few boundary nodes invalidate skew symmetry.

We now explain how the issue raised at the beginning of this section can be resolved starting in the already introduced CG setting. Let us take a step back and remind ourselves how the matrices $\boldsymbol{C}$ arise in the first place. The strong form of the CG discretization of the conservation law (1) contains the integral

$$\int_\Omega \varphi_i \nabla \cdot \boldsymbol{f}(u_h) \,\mathrm{d}\boldsymbol{x}, \tag{12}$$

which can generally only be approximated via quadrature. The *group finite element* formulation [45], [46]:

$$\boldsymbol{f}_h := \sum_{j=1}^N \boldsymbol{f}_j \varphi_j \approx \boldsymbol{f}(u_h) \tag{13}$$

interpolates the inviscid flux $\boldsymbol{f}(u_h)$ in the finite element space. On simplicial meshes with periodic boundary conditions, this approach can be interpreted as nodal quadrature [46]. Inserting (13) into (12), we obtain

$$\int_\Omega \varphi_i \nabla \cdot \boldsymbol{f}(u) \,\mathrm{d}\boldsymbol{x} \approx \sum_{j=1}^N \boldsymbol{f}_j \cdot \int_\Omega \varphi_i \nabla \varphi_j \,\mathrm{d}\boldsymbol{x} = \sum_{j=1}^N \boldsymbol{f}_j \cdot \boldsymbol{c}_{ij} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} (\boldsymbol{f}_j - \boldsymbol{f}_i) \cdot \boldsymbol{c}_{ij}, \tag{14}$$

which is how the discrete gradient operators arise in AFC schemes. The last equality in (14) holds because the row sums of each C^k are zero due to the partition-of-unity property $\sum_{j=1}^N \varphi_j(\boldsymbol{x}) = 1$ for any $\boldsymbol{x} \in \bar{\Omega}$. Replacing $\boldsymbol{c}_{ij}$ in (14) by some $\tilde{\boldsymbol{c}}_{ij} \in \mathbb{R}^d$ such that $\tilde{\boldsymbol{c}}_{ij} = \boldsymbol{0}$ if $j \notin \mathcal{N}_i$ is equivalent to adding

$$\sum_{j \in \mathcal{N}_i \setminus \{i\}} (\boldsymbol{f}_j - \boldsymbol{f}_i) \cdot (\tilde{\boldsymbol{c}}_{ij} - \boldsymbol{c}_{ij}) \tag{15}$$

to (14). If for all $k \in \{1, \dots, d\}$ the row sums of each $\tilde{C}^k$ in $\tilde{\boldsymbol{C}} = (\tilde{C}^1, \dots, \tilde{C}^d) = (\tilde{\boldsymbol{c}}_{ij})_{i,j=1}^N$ sum to zero, and the columns of the matrices C^k and $\tilde{C}^k$ add up to the same values c_j^k stored in $\boldsymbol{c}_j = \left(c_j^1, \dots, c_j^d\right)$, then

$$\sum_{i=1}^N \sum_{j \in \mathcal{N}_i \setminus \{i\}} (\boldsymbol{f}_j - \boldsymbol{f}_i) \cdot (\tilde{\boldsymbol{c}}_{ij} - \boldsymbol{c}_{ij}) = \sum_{j=1}^N \boldsymbol{f}_j \cdot \left(\sum_{i=1}^N \tilde{\boldsymbol{c}}_{ij} - \sum_{i=1}^N \boldsymbol{c}_{ij} \right) = \sum_{j=1}^N \boldsymbol{f}_j \cdot (\boldsymbol{c}_j - \boldsymbol{c}_j) = 0.$$

Thus, the proposed modification does not lead to a change in the global mass balance. Note that by writing the correction term (15) using a sum over only indices $j \neq i$, we may choose the diagonal entries $\tilde{\boldsymbol{c}}_{ii}$ arbitrarily. This possibility is what will allow us to define the skew-symmetric discrete gradients below.

Remark 1. In this paper, we are referring to a matrix $A = (a_{ij})_{i,j=1}^N$ as skew symmetric, if $a_{ij} + a_{ji} = 0$ for all $i, j \in \{1, \dots, N\}, i \neq j$, which does not imply zero diagonal entries. Their values are unimportant since their use can be avoided in AFC schemes via formulations such as (15).

An additional constraint that is desirable in the context of high-order AFC schemes is the sparsity of $\tilde{\boldsymbol{C}}$ [10], [14], [20], [21], [25], [32], [47]. By that we refer to the property that every nonzero offdiagonal entry corresponds to two distinct nodes that are closest neighbors (in a certain metric) in the submesh obtained by using each Bernstein node as a degree of freedom of a continuous subcell $\mathbb{P}_1$ or $\mathbb{Q}_1$ (depending on the element geometry) discretization [13], [14], [25]. The corresponding stencil $\tilde{\mathcal{N}}_i$ will be further specified below. Each pair of nodes (i, j) with $j \in \tilde{\mathcal{N}}_i \backslash \{i\}$ corresponds to one antidiffusive flux that needs to be limited and incorporated. Thus, the fewer elements in the stencils $\tilde{\mathcal{N}}_i$ the better in terms of computational resources.

Thus, we may replace the discrete gradient operator $\boldsymbol{C}$ by matrices $\tilde{\boldsymbol{C}}$ of the same size satisfying

$$\tilde{\boldsymbol{c}}_{ij} = -\tilde{\boldsymbol{c}}_{ji} \quad \forall i, j \in \{1, \dots, N\}, \quad i \neq j, \tag{16a}$$

$$\sum_{j=1}^N \tilde{\boldsymbol{c}}_{ij} = \boldsymbol{0}, \tag{16b}$$

$$\sum_{i=1}^N (\boldsymbol{c}_{ij} - \tilde{\boldsymbol{c}}_{ij}) = \boldsymbol{0} \quad \forall j \in \{1 \dots, N\}, \tag{16c}$$

$$\tilde{c}_{ij}^k \neq 0 \quad \Rightarrow \quad i = j \text{ or } i \text{ and } j \text{ are closest neighbors in the stencil } \tilde{\mathcal{N}}_i. \tag{16d}$$

The remainder of this section is dedicated to constructing such matrices $\tilde{\boldsymbol{C}}$. For second-order continuous Galerkin schemes with weakly-enforced boundary conditions, $\boldsymbol{C}$ satisfies (16b)–(16d). Thus finding a $\tilde{\boldsymbol{C}}$ that also fulfills (16a) is simple (which is why no originality is claimed here, although we found no works using these matrices in a similar context). Given the original operators $\boldsymbol{C}$ and symmetric boundary matrices $\boldsymbol{B} = (B^1, \dots, B^d) = (\boldsymbol{b}_{ij})_{i,j=1}^N$, we only have to adjust a few entries of $\boldsymbol{C}$, which can be done as follows:

$$\tilde{\boldsymbol{c}}_{ij} = \begin{cases} \boldsymbol{c}_{ij} - \dfrac{1}{2}\boldsymbol{b}_{ij} & \text{if } i \neq j, \\ \boldsymbol{c}_{ii} + \dfrac{1}{2}\displaystyle\sum_{k \in \mathcal{N}_i \backslash \{i\}} \boldsymbol{b}_{ik} & \text{otherwise.} \end{cases}$$

Proving the validity of (16) for this operator is straightforward. So far, we have focused on global matrices. Below, we discuss elementwise operators because these become relevant in the high-order and/or DG cases.

3.2 Skew-symmetric discrete gradient operators for Bernstein polynomials

In Section 3.1, we established that the group finite element formulation [45], [46] produces the discrete gradients

$$\boldsymbol{C} = (C^1, \dots, C^d) = (\boldsymbol{c}_{ij})_{i,j=1}^N, \qquad \boldsymbol{c}_{ij} = \int_\Omega \varphi_i \nabla \varphi_j \, \mathrm{d}\boldsymbol{x}, \tag{17}$$

where $\{\varphi_i\}_{i=1}^N$ are the Lagrange basis functions of $\mathbb{P}_1$ or $\mathbb{Q}_1$ continuous finite element spaces. Let us now consider elementwise operators $\boldsymbol{C}$ for a certain reference element $K \subset \mathbb{R}^d$ and the Bernstein basis. These matrices are obtained by replacing the number of rows and columns N and the integration domain Ω in (17) by the local number of unknowns and the reference element K, respectively. With some abuse of notation, we avoid making

this distinction, to avoid redefining all operators or carry additional indices. For each of the below element types, we construct matrices $\tilde{\boldsymbol{C}} = (\tilde{C}^1, \dots, \tilde{C}^d) = (\tilde{\boldsymbol{c}}_{ij})_{i,j=1}^N$, satisfying conditions (16).

3.2.1 1D elements

Denote the 1D reference element by $K = [0,1]$. The Bernstein polynomials of degree $p \in \mathbb{N}$ associated with the nodes $x_k = k/p$ are $\varphi_k^p(x) = \binom{p}{k}(1-x)^{p-k}x^k$, $k \in \{0, \dots, p\}$. The derivative operator $C = C^1 = (c_{ij})_{i,j=0}^p$, $c_{ij} = \int_0^1 \varphi_i^p(x)(\varphi_j^p)'(x)\,\mathrm{d}x$ is a dense matrix with zero row sums, and its column sums read

$$\sum_{i=0}^{p} \int_0^1 \varphi_i^p(x) \frac{\partial \varphi_j^p(x)}{\partial x}\,\mathrm{d}x = \int_0^1 \frac{\partial \varphi_j^p(x)}{\partial x}\,\mathrm{d}x = \varphi_j^p(1) - \varphi_j^p(0) = \begin{cases} -1 & \text{if } j = 0, \\ 1 & \text{if } j = p, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Lemma 2. *The tridiagonal matrix (originally proposed by Pazner [20, Eq. (21)] for an LGL-FCT scheme)*

$$\tilde{C} = (\tilde{c}_{i,j})_{i,j=0}^p, \quad \tilde{c}_{1,1} = \tilde{c}_{l,l-1} = -0.5, \quad \tilde{c}_{p,p} = \tilde{c}_{l-1,l} = 0.5 \tag{19}$$

for $l \in \{1, \dots, p\}$ and all other entries set to zero, satisfies conditions (16)*.*

Proof. Condition (16c) follows from (18). The rest is obvious, see also Pazner [20]. □

Remark 2. Applying the theory by Lohmann et al. [14, App. B] to the case of discrete gradients, we obtain alternative sparsified operators given by [14, Eq. (B.4)]. Note that this matrix satisfies all of the conditions (16) except for (16a). As we quantify in Sections 5 and 6, the use of [14, Eq. (B.4)] requires significantly smaller time steps for large p than its skew-symmetric counterpart defined in Lemma 2.

Lemma 3. *Let $\{\psi_k\}_{k=0}^p$ be the piecewise linear continuous Lagrange basis interpolating the Bernstein nodes $\{k/p\}_{k=0}^p$ of the reference element [0, 1]. Then the following matrix coincides with* (19)*:*

$$\bar{C} = (\bar{c}_{ij})_{i,j=0}^p, \qquad \bar{c}_{ij} = \int_0^1 \psi_i(x) \frac{\partial \psi_j(x)}{\partial x}\,\mathrm{d}x. \tag{20}$$

Proof. For $i \neq j$, the integral in (20) reduces to a single subcell connecting nodes x_i and x_j if they are closest neighbors. The subcell length is the reciprocal of the constant slope of ψ_j^p and so it remains to integrate ψ_i^p, which yields precisely $\bar{c}_{ij} = \pm 1/2$. The case $i = j$ follows from $\bar{c}_{ii} = \frac{1}{2}\int_0^1 \frac{\partial \psi_i(x)^2}{\partial x}\,\mathrm{d}x$. □

At first, it seems natural to extend the definition of subcell $\mathbb{P}_1/\mathbb{Q}_1$ Lagrange polynomials to the multidimensional case and use an analog of (20) to define sparse discrete gradients. However, similar to the case of dense discrete gradients $\boldsymbol{C}$, these matrices are not fully skew-symmetric because of entries corresponding to pairs of nodes that are neighbors on the same boundary segment of the element. Thus, we have to find other matrices to achieve skew symmetry, which we focus on next.

3.2.2 Box elements

Let us now discuss the case of $\mathbb{Q}_p$ elements, i.e., quadrilaterals in 2D, hexahedra in 3D, or higher-dimensional analogues. For generality, we allow varying polynomial degrees p_l in each spatial direction $l \in \{1, \dots, d\}$. The Bernstein basis functions $\{\varphi_i\}_{i=1}^N$ are simply tensor-products of the one-dimensional Bernstein polynomials and

$N = \prod_{l=1}^{d}(p_l+1)$. We choose a lexicographical ordering of basis functions and corresponding node indices $i = (i_1, \dots, i_d)$ within the reference element, where to obtain the element-global index i we first iterate over all basis functions with respect to the first spatial variable, i.e., going through all $i_1 \in \{0, \dots, p_1\}$ before incrementing the subsequent one i_2 once and so forth. This common approach enables us to mimic the 1D case by using Kronecker products $A \otimes B \in \mathbb{R}^{(kl)\times(kl)}$ of matrices $A \in \mathbb{R}^{k\times k}$, $B \in \mathbb{R}^{l\times l}$. By $\mathbf{I}_l \in \mathbb{R}^{l\times l}$, we denote the identity matrix and by $\mathbf{1}_l \in \mathbb{R}^l$ the vector of ones.

Lemma 4. *Let* $\tilde{C}_{p_k} \in \mathbb{R}^{(p_k+1)\times(p_k+1)}$ *be given by* (19) *for* $p_k = p$. *Then the matrices* $\tilde{\boldsymbol{C}} = (\tilde{\boldsymbol{c}}_{ij})_{i,j=1}^{N}$,

$$\tilde{\boldsymbol{C}} = (\tilde{C}^1, \dots, \tilde{C}^d), \qquad \tilde{C}^k = \frac{1}{\prod_{\substack{l=1\\ l\neq k}}^{d}(p_l+1)}\left[\mathbf{I}_{\Pi_{l=1}^{k-1}(p_l+1)} \otimes \tilde{C}_{p_k} \otimes \mathbf{I}_{\Pi_{l=k+1}^{d}(p_l+1)}\right] \in \mathbb{R}^{N\times N},$$

satisfy conditions (16) *for multidimensional box elements (quadrilaterals, hexahedra, etc.).*

Proof. If $\tilde{c}_{ij}^k \neq 0$ and $i = (i_1, \dots, i_d) \neq j = (j_1, \dots, j_d)$, then $i_l = j_l$ for all $l \in \{1, \dots, d\}\backslash\{k\}$ and $i_k = j_k \pm 1$. This property follows from the definition of $\tilde{\boldsymbol{C}}$ with the lexicographical numbering of nodes and implies sparsity (16d). Similarly, considering such a pair of nodes, skew symmetry (16a) can be shown by definition of the 1D matrices $\tilde{C}_{p_k}$. Furthermore, let A, B, C, D be matrices such that products AC and BD can be formed, then the Kronecker product satisfies the relationship $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. Thus,

$$\tilde{C}^k(\mathbf{1}_{p_1+1} \otimes \dots \otimes \mathbf{1}_{p_d+1}) = \frac{1}{\prod_{\substack{l=1\\ l\neq k}}^{d}(p_l+1)}\left[\mathbf{1}_{(p_1+1)\dots(p_{k-1}+1)} \otimes \mathbf{0}_{p_k+1} \otimes \mathbf{1}_{(p_{k+1}+1)\dots(p_d+1)}\right] = \mathbf{0}_N$$

because $\tilde{C}_{p_k}\mathbf{1}_{p_k+1} = \mathbf{0}_{p_k+1}$, where $\mathbf{0}_l \in \mathbb{R}^l$ is the zero vector. Hence, (16b) holds. Finally, to show (16c), we recall the definition $c_j^k = \int_K \frac{\partial\varphi_j}{\partial x_k}\,\mathrm{d}\boldsymbol{x} = \int_{\partial K}\varphi_j n_k\,\mathrm{d}s$, where $K \subset \mathbb{R}^d$ is the reference element and n_k is the kth component of the normal to ∂K. Note that for fixed $j \in \{1, \dots, N\}$, $k \in \{1, \dots, d\}$, ∂K in this integral can be replaced by a single boundary face Γ, on which φ_j is equal to a Bernstein polynomial in $(d-1)$ variables. Since Γ is a box element in $\mathbb{R}^{d-1}$ with volume $|\Gamma| = 1$ and all Bernstein polynomials possess equal mass, that is $\int_\Gamma \varphi_i\,\mathrm{d}s = |\Gamma|/N_k$, where $N_k = \prod_{\substack{l=1\\ l\neq k}}^{d}(p_l+1)$ is the number of nodes on Γ, it follows that

$$\sum_{i=1}^{N} c_{ij}^k = \sum_{i=1}^{N}\int_K \varphi_i\frac{\partial\varphi_j}{\partial x_k}\,\mathrm{d}\boldsymbol{x} = c_j^k = \frac{1}{\prod_{\substack{l=1\\ l\neq k}}^{d}(p_l+1)} \times \begin{cases} -1 & \text{if } j_k = 0,\\ 0 & \text{if } j_k \in \{1, \dots, p-1\},\\ 1 & \text{if } j_k = p. \end{cases}$$

This is equal to the corresponding row sum of $\tilde{C}^k$ because

$$(\mathbf{1}_{p_1+1} \otimes \dots \otimes \mathbf{1}_{p_d+1})^\top\tilde{C}^k = \frac{1}{\prod_{\substack{l=1\\ l\neq k}}^{d}(p_l+1)}\left[\mathbf{1}_{(p_1+1)\dots(p_{k-1}+1)}^\top \otimes \mathbf{1}_{p_k+1}^\top\tilde{C}_{p_k} \otimes \mathbf{1}_{(p_{k+1}+1)\dots(p_d+1)}^\top\right]$$

and $\mathbf{1}_{p_k+1}^\top\tilde{C}_{p_k} = [-1, 0, \dots, 0, 1]$, see (19). Thus, all properties in (16) hold. □

3.2.3 Simplices

We now move on to the reference simplex $K = \mathrm{conv}\{\mathbf{0}_d, \mathbf{e}_1, \dots, \mathbf{e}_d\} \subset \mathbb{R}^d$, where conv{ } denotes the convex hull and $\mathbf{e}_i$ is the ith unit vector in $\mathbb{R}^d$. The barycentric coordinates are $\Lambda_0(\boldsymbol{x}) = 1 - \sum_{i=1}^{d} x_i$, $\Lambda_i(\boldsymbol{x}) = x_i$ for $i \in$

$\{1, \dots, d\}$, and the Bernstein nodes shall be numbered using a multi-index notation $i = (i_0, \dots, i_d)$ similar to the case of box elements. The set of multi-indices representing nodes in K is

$$\mathcal{J} = \mathcal{J}(p,d) = \left\{ (i_0, \dots, i_d) \in \{0, \dots, p\}^{d+1} \colon \sum_{j=0}^{d} i_j = N := \binom{p+d}{p} \right\}$$

and the (isotropic) pth-degree Bernstein basis functions read $\varphi_k^p(\boldsymbol{x}) = \frac{p!}{\prod_{l=0}^d k_l!} \prod_{l=0}^d \Lambda_l(\boldsymbol{x})^{k_l}$, $k \in \mathcal{J}$.

3.2.3.1 The $\mathbb{P}_1$ triangle

Before studying the general simplex case, let us briefly focus on the $\mathbb{P}_1$-triangle in 2D. By (16a)–(16c), all matrices satisfying (16) can be expressed as

$$C^1 = \frac{1}{4}\begin{bmatrix} -1 & a & 1-a \\ -a & 1 & a-1 \\ a-1 & 1-a & 0 \end{bmatrix}, \qquad C^2 = \frac{1}{4}\begin{bmatrix} -1 & b & 1-b \\ -b & 0 & b \\ b-1 & -b & 1 \end{bmatrix}, \qquad a, b \in \mathbb{R}.$$

The magnitude of nonzero off-diagonal entries in these matrices affects the time step restriction of convex limiting schemes due to (9) and (5). Thus, all nonzero off-diagonal entries of C^1 and C^2 should generally have the same absolute value if possible (see also Section 5 for further discussions on this issue). By this argument, it makes sense to only choose $a, b \in \{0, 0.5, 1\}$. Figure 1 shows various choices including three (Figure 1b–d) out of the possible nine combinations of parameters a and b in addition to the usual, non-skew-symmetric matrices $\boldsymbol{C}$ (Figure 1a) along with a variant, where $a, b \notin \{0, 0.5, 1\}$ (Figure 1e).

The case $a = 1$, $b = 0$ leads to dimensional decoupling and inhibits optimal convergence rates, see the numerical examples in Section 6.2. Setting $a = 0$, $b = 1$ corresponds to an unusual gradient, where coupling occurs in a direction that is opposite to the one in the previous case. This scheme exhibits optimal convergence behavior but requires significantly more time steps than the choice $a = b = 0.5$, which seems to be optimal in terms of CFL conditions. The gradient in Figure 1e corresponds to a discrete gradient that satisfies conditions (16) but uses off-diagonal entries of varying magnitude. As a result, the resulting CFL condition is much more restrictive, however, this method also exhibits optimal convergence rates. We have not experimented with gradients other than the ones shown in Figure 1 because the two matrices C^1, C^2 corresponding to any other choice of $a, b \in \{0, 0.5, 1\}$ would be dimensionally inconsistent to each other.

In conclusion, uniqueness of simplicial discrete gradients satisfying conditions (16) does not hold. Note that for box elements the operators specified earlier are only unique because of their tensor-product structure (not possible here). Based on the summarized results obtained with each of the schemes in Figure 1 (detailed in Section 6.2), it makes sense to generalize the gradient in Figure 1d. In the remainder of this section, we demonstrate that this approach is feasible for general high-order simplices in arbitrary space dimensions.

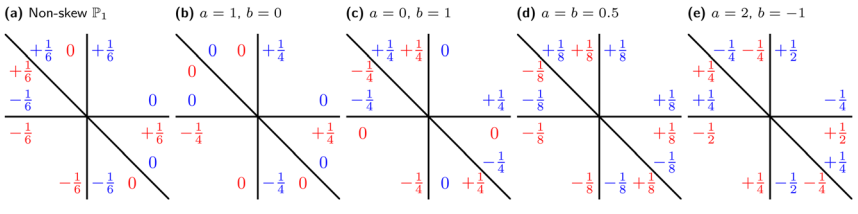

Figure 1: Five different options for how to choose $\tilde{\boldsymbol{C}}$ on triangles, red represents entries in C^1, blue represents entries in C^2.

3.2.3.2 General simplices

The realization that the gradient in Figure 1d appears to be optimal for $\mathbb{P}_1$-triangles implies the following structure for arbitrary $p, d \in \mathbb{N}$: Conditions (16a)–(16c) fully determine the values of diagonal entries since

$$c_j^k = \sum_{i=1}^N \tilde{c}_{ij}^k = \sum_{i=1}^N \tilde{c}_{ij}^k + \sum_{i=1}^N \tilde{c}_{ji}^k = \sum_{i=1}^N \left(\tilde{c}_{ij}^k + \tilde{c}_{ji}^k\right) = 2\tilde{c}_{jj}^k + \sum_{\substack{i=1 \\ i\neq j}}^N (\tilde{c}_{ij}^k - \tilde{c}_{ij}^k) = 2\tilde{c}_{jj}^k \qquad \Rightarrow \qquad \tilde{c}_{jj}^k = \frac{c_j^k}{2}.$$

In any matrix row i, the offdiagonal entry in the jth column is nonzero only if the multiindices i and j correspond to nodes that are nearest neighbors. The magnitude of offdiagonal entries should be constant, and their sign should be determined by the spatial direction. If the node $\boldsymbol{x}_j$ lies in positive direction w.r.t. any Cartesian coordinate, then $\tilde{c}_{ij}^k > 0$ and $\tilde{c}_{ij}^k < 0$ otherwise for all $k \in \{1, \dots, d\}$, i.e., for all matrices. For closest neighbors other than these nodes (diagonal neighbors), the sign of $\tilde{c}_{ij}^k$ depends on whether the diagonal direction is positive w.r.t. the coordinate x_k. If $d > 2$, there are diagonal neighbors with the same multiindex component for the x_k direction. The corresponding indices are set to zero in $\tilde{C}^k$.

To formalize these considerations mathematically, we require some notation. For consistency with previous works [25], [32], we define connectivity on simplices as follows.

Definition 1. Let $i, j \in \mathcal{J}$. If $\exists k, l \in \{0, \dots, p\}, k \neq l$, such that $j = i + \mathrm{e}_k - \mathrm{e}_l$, where $\mathrm{e}_{k,l}$ are the kth and lth unit vectors in $\mathbb{R}^{d+1}$, respectively, then we say $j \in \tilde{\mathcal{N}}_i$, i.e., j is in the local stencil $\tilde{\mathcal{N}}_i$.

In addition, let

$$s = \frac{1}{2(d-1)!\binom{p+d-1}{p}} = \frac{p!}{2(p+d-1)!} \tag{21}$$

denote half the mass of a pth-degree Bernstein polynomial on the $(d-1)$-dimensional reference simplex. Furthermore, let $c_j^k = \int_{\partial K} \varphi_j n_k \, \mathrm{d}s$ be as before, and for $i, j \in \mathcal{J}$, define

$$\tilde{c}_{ij}^k := \begin{cases} -s/d & \text{if } (j \in \tilde{\mathcal{N}}_i) \wedge [(j_0 = i_0 + 1) \vee (j_k = i_k - 1)], \\ s/d & \text{if } (j \in \tilde{\mathcal{N}}_i) \wedge [(j_0 = i_0 - 1) \vee (j_k = i_k + 1)], \\ -s & \text{if } (i = j) \wedge (i_0 > 0) \wedge (i_k = 0), \\ s & \text{if } (i = j) \wedge (i_0 = 0) \wedge (i_k > 0), \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Lemma 5. *The matrices* $\tilde{C}^k = (\tilde{c}_{ij}^k)_{i,j\in\mathcal{J}} = (\tilde{c}_{(i_0,\dots,i_d)(j_0,\dots,j_d)}^k)_{i,j\in\mathcal{J}}$ *given by* (22) *satisfy conditions* (16).

Proof. By Definition 1, $i \notin \tilde{\mathcal{N}}_i$ for all $i \in \mathcal{J}$. Therefore, all cases in (22) are exclusive, and $\tilde{c}_{ij}^k$ is well defined. Sparsity (16d) is built into Definition 1, and skew symmetry (16a) is proved by observing that for $i \neq j$,

$$\begin{aligned} \tilde{c}_{ji}^k &= \begin{cases} -s/d & \text{if } (i \in \tilde{\mathcal{N}}_j) \wedge [(i_0 = j_0 + 1) \vee (i_k = j_k - 1)], \\ s/d & \text{if } (i \in \tilde{\mathcal{N}}_j) \wedge [(i_0 = j_0 - 1) \vee (i_k = j_k + 1)], \\ 0 & \text{otherwise,} \end{cases} \\ &= \left\{ \begin{array}{ll} s/d & \text{if } (j \in \tilde{\mathcal{N}}_i) \wedge [(j_0 = i_0 + 1) \vee (j_k = i_k - 1)] \\ -s/d & \text{if } (j \in \tilde{\mathcal{N}}_i) \wedge [(j_0 = i_0 - 1) \vee (j_k = i_k + 1)] \\ 0 & \text{otherwise} \end{array} \right\} = -\tilde{c}_{ij}^k. \end{aligned}$$

We show (16c) by proving that

$$\tilde{c}^k_{jj} = \frac{1}{2} c^k_j \quad \forall j \in \{1, \dots, N\}, \quad k \in \{1, \dots, d\}, \tag{23}$$

which, together with (16a) and (16b) (to be proven last), implies (16c). To better illustrate how (23) can be shown, let us consider the example sketched in Figure 2. The red nodes are precisely those corresponding to the nonzero diagonal entries in (22). These are set in accordance with (23), as are the entries for black and blue nodes: Diagonal entries for the former are zero because these nodes are either within the element interior or lie on a boundary where the respective component of the normal appearing in c^k_j is zero. The blue nodes are degrees of freedom for which the Bernstein polynomial is nonzero on more than one boundary segment, with the corresponding normal component being nonzero on precisely two (also for $d > 2$) of these, which depends on k. These two integrals cancel, which is consistent with definition (22).

Let us now rigorously prove (23) for (22) by formalizing these exemplified considerations in all required cases: If $j_0 j_k > 0$ (black nodes), then $c^k_j = 0$ and (22) is consistent with (23) because $\tilde{c}^k_{jj} = 0$. For $j_0 > 0 = j_k$ (red nodes not opposite the origin), the integral $c^k_j = \int_{\partial K} \varphi_j n_k \, \mathrm{d}s$ can be reduced to a $(d-1)$-dimensional reference simplex, on which φ_j is a Bernstein polynomial in $d-1$ variables. The component n_k of the normal to this boundary is always -1, thus, $c^k_j = -2s$ by (21), and $\tilde{c}^k_{jj} = -s$. If $j_0 = 0$ and $j_k > 0$ (red nodes opposite the origin), the situation is exactly reversed, and by similar arguments (including transformation to the reference simplex), we obtain $c^k_j = 2s = 2\tilde{c}^k_{jj}$. In the case $j_0 = j_k = 0$ (blue nodes), contributions to the integral c^k_j arise from precisely two boundaries, one of which is always opposite the origin. These two integrals are clearly of opposite sign. Invoking transformation rules, one can show that their magnitudes are equal. Hence, $c^k_j = 0$, which is consistent with (22) because $\tilde{c}^k_{jj} = 0$. Thus, we have shown (23).

It remains to prove (16b), where, for clarity, we also distinguish between all relevant cases based on the row multiindex $i \in \mathcal{J}$: For $i_0 = p$, we have $\tilde{c}^k_{ii} = -s$, precisely d positive, and no negative entries in the row. Similarly, for $i_k = p$, $\tilde{c}^k_{ii} = s$, and the other d nonzero row entries are negative. For any other vertex, i.e., $\exists l \in \{1, \dots, d\} \backslash \{k\}$ with $i_l = p$ (for $d = 2$, these are blue nodes in Figure 2), we have $\tilde{c}^k_{ii} = 0$. The nodes with $j_0 = 1$, $j_l = p - 1$ and $j_k = 1$, $j_l = p - 1$ each contribute one negative and one positive row entry, of which there are no more nonzero ones. We have now dealt with all corners of the simplex. At this stage, it makes sense to restrict ourselves to the case $d > 1$ (in 1D, it can easily be shown that (22) coincides with (18)). We define $q := |\{l \in \{1, \dots, d\} \backslash \{k\} : i_l > 0\}|$, where $|\cdot|$ denotes the cardinality of a set. For $i_0 = i_k = 0$, we have $\tilde{c}^k_{ii} = 0$ and there are precisely q negative and q positive entries in row i, which correspond to $j_0 = i_0 + 1$ and $j_k = i_k + 1$, respectively. For $i_0 = 0 < i_k < p$, we have $\tilde{c}^k_{ii} = s$. The negative row entries correspond to either $j = i + \mathrm{e}_0 - \mathrm{e}_k$ (one entry), $j_0 = i_0 + 1$ (q additional entries), or $j_k = i_k - 1$ ($d - 1$ additional (diagonal) neighbors). Moreover, there are precisely q positive entries in that row, which correspond to $j_k = i_k + 1$. Thus, the row sum is $s - \frac{s}{d}(1 + q + d - 1) + \frac{s}{d} q = 0$. The case $i_k = 0 < i_0 < p$ is similar, and we obtain $-s - \frac{s}{d} q + \frac{s}{d}(1 + d - 1 + q) = 0$ for the row sum. Finally, for $0 < i_{0,k} < p$, we have $\tilde{c}^k_{ii} = 0$ and the row sum is $-\frac{s}{d}(1 + q + d - 1) + \frac{s}{d}(1 + d - 1 + q) = 0$. We have thus shown (16b), which together with (23) concludes the proof of (16c) as well as that of Lemma 5. □

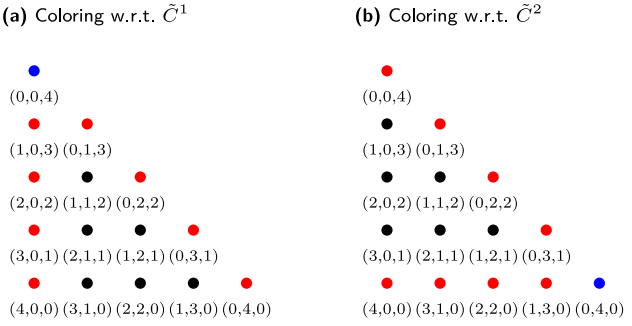

Figure 2: Bernstein nodes of the $\mathbb{P}_4$ triangle and corresponding multiindices.

3.2.4 Prisms

Finally, we study prismatic cells (also called wedges) in $\mathbb{R}^3$. These are tensor products of a triangular element and an interval, which is exploited here. For simplicity, we make the assumption that the spatial dimensions are numbered such that the reference element is $K = \triangle \times [0,1]$, where $\triangle = \mathrm{conv}\{(0,0)^\top, (1,0)^\top, (0,1)^\top\}$. We allow the Bernstein basis functions to be of different orders $p_{xy}, p_z \in \mathbb{N}$ on the triangle and the interval, respectively. The Bernstein basis functions are now tensor products of the triangular Bernstein polynomials and the 1D basis. We define $n = \binom{p_{xy}+2}{p_{xy}}$ as the number of nodes on $\triangle$.

Lemma 6. *Let $\tilde{C}_{p_z}$ be given by* (19) *for $p = p_z$ and $\tilde{C}^x, \tilde{C}^y \in \mathbb{R}^{n\times n}$ be given by* (22) *for $p = p_{xy}$ and let*

$$\tilde{C}^1 = \frac{1}{p_z+1}\mathbf{I}_{p_z+1} \otimes \tilde{C}^x, \qquad \tilde{C}^2 = \frac{1}{p_z+1}\mathbf{I}_{p_z+1} \otimes \tilde{C}^y, \qquad \tilde{C}^3 = \frac{1}{(p_{xy}+1)(p_{xy}+2)}\tilde{C}_{p_z} \otimes \mathbf{I}_n.$$

Then the matrices $\tilde{\boldsymbol{C}} = (\tilde{C}^1, \tilde{C}^2, \tilde{C}^3)$ satisfy conditions (16)*.*

Proof. The proof (omitted for brevity) follows those for quadrilaterals and simplices, see Lemmas 4 and 5. □

3.3 Implementation aspects

The GitHub repository [48] was published together with the first DG-MCL paper [25]. It provides codes for computing the discrete gradient operators on simplices following [32]. This repository has been updated to include both the old and new sparsification approaches for all element geometries considered in this work.

Having discussed various geometry reference elements, let us briefly summarize the required modifications to obtain local discrete gradients on actual mesh elements. Let $\hat{\boldsymbol{C}} = (\hat{C}^1, \dots, \hat{C}^d)$, $\hat{C}^k = (\hat{c}_{ij})_{i,j=1}^N$ for $k \in \{1, \dots, d\}$ be the reference element matrices and let $\mathrm{adj}(J)$ denote the adjugate of the Jacobian matrix J to the transformation for mapping the reference element to physical cells. Then all entries of the corresponding discrete gradient operators on physical cells are obtained via $\mathbb{R}^d \ni \boldsymbol{c}_{ij} = \mathrm{adj}(J)^\top[\hat{c}_{ij}^1, \dots, \hat{c}_{ij}^d]^\top$. Here we assumed linearity of the transformation to factor out $\mathrm{adj}(J)^\top$. Whether the discrete gradients defined in this manner are suitable for nonlinear mappings too is yet to be determined. We refer to ref. [21, App. B] for a discussion regarding curvature and nonlinear transformations in the very similar LGL-AFC context.

4 Theoretical results for Bernstein-DG-AFC schemes

4.1 Notation

The superscript $e \in \{1, \dots, E\}$ generally denotes the element index, where E is the number of mesh cells. The global solution u_h is obtained from local contributions u_h^e via $u_h = \sum_{e=1}^E u_h^e$. A single subscript index refers to local nodes within the element. Two subscripts indicate interplay between distinct degrees of freedom, e.g., volumetric diffusion coefficients d_{ij}^e (here i, j are local indices of nodes within K^e). If subscripts are separated by a comma, this indicates coupling across element boundaries, e.g., interfacial diffusion coefficients $d_{i,k}^e$. Let $\mathcal{F}_i^e$ denote the set of (interior) boundary segments Γ_k^e that the node $\boldsymbol{x}_i^e$ belongs to. For fixed $k \in \mathbb{N}$, there exists a unique node $\boldsymbol{x}_{i'}^{e'}$ at the same location as $\boldsymbol{x}_i^e \in \partial K^e$. To ease readability, these quantities are denoted using a hat symbol, e.g., the degrees of freedom associated with the same location as $\boldsymbol{x}_i^e$ but located within the neighbor element that shares face Γ_k^e with K^e is $\hat{u}_{i,k}^e$ (not $u_{i'}^{e'}$).

The semi-discrete AFC discretization for the Bernstein-DG discretization reads

$$m_i^e \frac{\mathrm{d}u_i^e}{\mathrm{d}t} = \sum_{j\in\mathcal{N}_i^e}\left[d_{ij}^e\left(u_j^e - u_i^e\right) - (\boldsymbol{f}(u_j^e) - \boldsymbol{f}(u_i^e))\cdot\tilde{\boldsymbol{c}}_{ij}^e + f_{ij}^{e,*}\right] + \sum_{k\in\mathcal{F}_i^e}\left[d_{i,k}^e\left(\hat{u}_{i,k}^e - u_i^e\right) - (\boldsymbol{f}(\hat{u}_{i,k}^e) - \boldsymbol{f}(u_i^e))\cdot\boldsymbol{c}_{i,k}^e + f_{i,k}^{e,*}\right], \tag{24}$$

where

$$m_i^e = \int_{K^e}\varphi_i^e\,\mathrm{d}\boldsymbol{x}, \quad d_{ij}^e = \max\left\{|\boldsymbol{c}_{ij}^e|\lambda_{ij}^e, |\boldsymbol{c}_{ji}^e|\lambda_{ji}^e\right\}, \quad \boldsymbol{c}_{i,k}^e = \frac{1}{2}\int_{\Gamma_k^e}\varphi_i^e\boldsymbol{n}_k^e\,\mathrm{d}s, \quad d_{i,k}^e = |\boldsymbol{c}_{i,k}^e|\lambda_{i,k}^e, \tag{25}$$

and $\boldsymbol{n}_k^e$ is the unit normal to Γ_k^e pointing outside of K^e. The volumetric and interfacial wave speeds λ_{ij}^e and $\lambda_{i,k}^e$ in the directions $\tilde{\boldsymbol{c}}_{ij}^e/|\tilde{\boldsymbol{c}}_{ij}^e|$ and $\boldsymbol{c}_{i,k}^e/|\boldsymbol{c}_{i,k}^e|$ depend on u_i^e and u_j^e or u_i^e and $\hat{u}_{i,k}^e$, respectively. In this section, we generally only assume (16b) and (16c) for entries of the discrete gradient operator. The volumetric and interfacial limited antidiffusive fluxes $f_{ij}^{e,*}$ and $\boldsymbol{f}_{i,k}^{e,*}$ are obtained from their unlimited counterparts f_{ij}^e and $\boldsymbol{f}_{i,k}^e$ by enforcing various constraints via MCL. These aspects [7], [25] need not be addressed here.

4.2 General results

For completeness, we reformulate an already established result [25, Lem. 1] in the context of quadrature.

Lemma 7 (recovery of the target scheme). *If no limiting is performed, i.e., if $f_{ij}^{e,*} = f_{ij}^e$ and $f_{i,k}^{e,*} = f_{i,k}^e$ for all unlimited antidiffusive fluxes defined as in ref. [25], then* (24) *is equivalent to the DG target scheme*

$$\int_{K^e}\varphi_i^e\frac{\partial u_h^e}{\partial t}\,\mathrm{d}\boldsymbol{x} - \sum_{j=1}^{l^e}\omega_j^e\boldsymbol{f}(u_h^e(\boldsymbol{q}_j^e))\cdot\nabla\varphi_i^e(\boldsymbol{q}_j^e) + \sum_{k\in\mathcal{F}_i^e}\sum_{j=1}^{l_k^e}\omega_{k,j}^e\varphi_i^e(\boldsymbol{q}_i^e)\boldsymbol{f}_{\boldsymbol{n}_k^e}(u_h^e(\boldsymbol{q}_i^e), \hat{u}_h^e(\boldsymbol{q}_i^e)) = 0$$

for all $e \in \{1,\dots,E\}$, $i \in \{1,\dots,N\}$, where $\omega_j^e, \boldsymbol{q}_j^e$ are the l^e quadrature weights and points used to approximate the nonlinear volume integral, and $\omega_{k,j}^e, \boldsymbol{x}_{k,j}^e$ are the l_k^e quadrature weights and points used to approximate the boundary integral containing the target numerical flux $\boldsymbol{f}_{\boldsymbol{n}_k^e}(u_h^e, \hat{u}_h^e)$ in direction $\boldsymbol{n}_k^e$.

Proof. The claim reformulates [25, Lem. 1], which assumes exact integration, using quadrature rules. □

As discussed in ref. [25], the similar structure of volumetric and interfacial terms in (24) allows for a combination of the two by adjusting the definition of local stencils and extending definitions of volumetric terms to include their interfacial counterparts. In this paper, we simply write sums over indices $j \neq i$ (with some abuse of notation) to indicate that both types of couplings are considered. Using this convention, the forward Euler time discretization reads

$$u_i^{e,\mathrm{FE}} = u_i^e + \frac{\tau}{m_i^e}\sum_{j\neq i}\left[d_{ij}^e\left(u_j^e - u_i^e\right) - \left(\boldsymbol{f}_j^e - \boldsymbol{f}_i^e\right)\cdot\boldsymbol{c}_{ij}^e + f_{ij}^{e,*}\right] = u_i^e + \frac{\tau}{m_i^e}\sum_{j\neq i}2d_{ij}^e\left(\bar{u}_{ij}^{e,*} - u_i^e\right),$$

where u_i^e and $u_i^{e,\mathrm{FE}}$ are the old and new nodal values. Next, we adapt [24, Thm. 4.7] to our setting.

Proposition 1 (fully discrete local entropy inequalities, [24]). *Consider an explicit SSP time discretization of* (24). *The low-order method, in which $f_{ij}^{e,*} = f_{i,k}^{e,*} = 0$, satisfies the local discrete entropy inequalities*

$$U_i^{e,\mathrm{FE}} \leqslant U_i^e + \frac{\tau}{m_i^e}\sum_{j\neq i}\left[d_{ij}^e\left(U_j^e - U_i^e\right) - \left(\boldsymbol{F}_j^e - \boldsymbol{F}_i^e\right)\cdot\tilde{\boldsymbol{c}}_{ij}^e\right] \tag{26}$$

w.r.t. all entropy pairs $(U, \boldsymbol{F})$ consistent with system (1) *if the CFL condition* (5) *holds.*

Proof. Our low-order method fits in the framework of [24, Thm. 4.7] and the proof directly carries over. □

Remark 3. Inequalities such as (26) cannot be derived for flux-limiters because (26) is based on properties that hold for $\bar{u}_{ij}$ but not for $\bar{u}_{ij}^*$ in the MCL case nor for FCT-like alternatives as in refs. [11], [44]. Thus, results akin to Proposition 1 are of limited practical relevance except for the analysis of first-order schemes. For comments on which entropy conditions should be enforced, we refer to refs. [7], [27] and the references therein.

Proposition 2 (local conservation property). *Let $\alpha_{i,k}^e \in [0,1]^m$ be the vectors of effective correction factors, with which each component of the interfacial antidiffusive flux $f_{i,k}^e$ is multiplied to obtain $f_{i,k}^{e,*}$ for (24), then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{K^e} u_h^e \,\mathrm{d}\boldsymbol{x} = -\sum_{i=1}^{N}\sum_{k\in\mathcal{F}_i^e}\int_{\Gamma_k^e}\varphi_i^e\left[\left(\mathbf{1}_m-\alpha_{i,k}^e\right)\circ f_{\boldsymbol{n}_k^e}^{LLF}\left(u_i^e,u_{i,k}^e\right)+\alpha_{i,k}^e\circ f_{\boldsymbol{n}_k^e}\left(u_h^e,u_{h,k}^e\right)\right]\mathrm{d}s,$$

where $f_{\boldsymbol{n}}^{\mathrm{LLF}}(u,v)=\frac{1}{2}[(\boldsymbol{f}(u)+\boldsymbol{f}(v))\cdot\boldsymbol{n}+\lambda_{\boldsymbol{n}}(u,v)(u-v)]$ is the local Lax–Friedrichs (LLF) flux and $\circ$ denotes componentwise multiplication of vectors of the same size.

Proof. Exploiting $d_{ij}^e=d_{ji}^e$, (16b) and (16c), skew-symmetry of antidiffusive fluxes, as well as the definitions of $\boldsymbol{c}_{i,k}^e$ and of $f_{i,k}^e$, see [25, Eq. (4.3)], we sum (24) over all local nodes, which yields the claim. □

The following two results were implied in ref. [25] but have not been formulated as such.

Proposition 3 (preservation of global bounds, [13], [24]). *Consider an SSP time discretization of (24) satisfying the CFL condition (5). Let $u_i^e, \bar{u}_{ij}^{e,*} \in \mathcal{A}$ for all $i \in \{1,\dots,N\}$, $j \neq i$, and all element indices e, where $\mathcal{A}$ is the largest admissible set. Then the solution at the next time step is also in $\mathcal{A}$.*

Proof. Again, it suffices to consider the forward Euler case, in which the updated solution $u_i^{e,\mathrm{FE}}$ is a convex combination of u_i^e and the $\bar{u}_{ij}^{e,*}$ under the CFL condition (5). The claim follows because $\mathcal{A}$ is convex. □

Proposition 3 implies that constraints such as global lower and upper bounds hold for numerical solutions to scalar conservation laws. For the Euler equations, limiting of $\bar{u}_{ij}^{e,*}$ w.r.t. density and pressure (see Section 2.3) can be used to guarantee nonnegativity of these two quantities (naturally only up to machine precision). For shock-capturing purposes, limiting w.r.t. local bounds may also be desirable.

Corollary 1 (preservation of local bounds, [13], [24]). *Consider an SSP time discretization of (24) satisfying the CFL condition (5). Let $u_i^e, \bar{u}_{ij}^{e,*} \in \mathcal{A}_i$ for all $j \neq i$ in the DG stencil (volumetric and interfacial neighbors, the former depending on the sparsity of $\tilde{\boldsymbol{C}}^e$), where $\mathcal{A}_i \subseteq \mathcal{A}$ is convex. Then $u_i^{e,\mathrm{FE}} \in \mathcal{A}_i$.*

Proof. The arguments used in the proof of Proposition 3 directly carry over to this setting. □

4.3 Results exploiting skew-symmetric discrete gradients

4.3.1 Two issues regarding sequential limiters

In the process of algebraic flux correction, it is not uncommon to perform more than one limiting step. For the Euler equations (and similar systems), we adopt the *sequential* approach [13], [21], [25], [49] of first adjusting density, followed by limiting of velocity components and specific total energy (the latter are limited independently of each other). Additionally, a pressure correction should always be performed and an entropy fix may also be desirable. These two steps each multiply all components of the antidiffusive flux by the same correction factors. However, due to the sequential approach used to limit specific unknowns rather than conserved ones [49],

these *synchronized* steps may violate the maximum principles enforced beforehand for specific unknowns. This issue may not be of critical importance because such constraints are used for shock-capturing purposes rather than for enforcing nonnegotiable solution properties. Nevertheless, it would be desirable to show that this issue can be avoided. In ref. [28, Lem. 3.16], a sufficient condition for such *compatibility* of sequential limiters with subsequent synchronized limiters is given, which reads

$$\bar{\rho}_{ij}^e \varphi_i^{e,\min} \leqslant \overline{(\rho\varphi)}_{ij}^e \leqslant \bar{\rho}_{ij}^e \varphi_i^{e,\max} \quad \forall j \neq i. \tag{27}$$

Here ρ is the main unknown (e.g., density), φ the specific unknown (e.g., velocity), and $(\rho\varphi)$ the conserved unknown (e.g., momentum) with corresponding bar states $\bar{\rho}_{ij}^e$,

$$\bar{\varphi}_{ij}^e = \frac{\overline{(\rho\varphi)}_{ij}^e + \overline{(\rho\varphi)}_{ji}^e}{\bar{\rho}_{ij}^e + \bar{\rho}_{ji}^e}, \tag{28}$$

and $\overline{(\rho\varphi)}_{ij}^e$, respectively [13]. If the discrete gradient is skew symmetric, we have $\bar{u}_{ij}^e = \bar{u}_{ji}^e$, hence (28) simplifies to $\bar{\varphi}_{ij}^e = \overline{(\rho\varphi)}_{ij}^e / \bar{\rho}_{ij}^e$ and (27) becomes $\varphi_i^{e,\min} \leqslant \bar{\varphi}_{ij}^e \leqslant \varphi_i^{e,\max}$ $\forall j \neq i$ (since $\rho \geqslant 0$). This simplified compatibility condition is nothing else than a design criterion for the definition of local bounds $\varphi_i^{e,\min}$ and $\varphi_i^{e,\max}$. It is automatically satisfied for the canonical choice [13, Eq. (78)] $\varphi_i^{\min} = \min_j \bar{\varphi}_{ij}$, $\varphi_i^{\max} = \max_j \bar{\varphi}_{ij}$, while for nonskew-symmetric discrete gradients, the validity of (27) is more difficult to guarantee.

A related issue is that if all correction factors of the sequential limiter are set to zero, the semi-discrete low-order method for product type variables (such as momentum or total energy) actually reads

$$m_i^e \frac{\mathrm{d}(\rho\varphi)_i^e}{\mathrm{d}t} = \sum_{j\neq i} 2 d_{ij}^e \left[\bar{\rho}_{ij}^e \bar{\varphi}_{ij}^e - (\rho\varphi)_i^e \right] \tag{29}$$

instead of

$$m_i^e \frac{\mathrm{d}(\rho\varphi)_i^e}{\mathrm{d}t} = \sum_{j\neq i} 2 d_{ij}^e \left[\overline{(\rho\varphi)}_{ij}^e - (\rho\varphi)_i^e \right]. \tag{30}$$

If $\bar{u}_{ij}^e = \bar{u}_{ji}^e$, (29) and (30) are equivalent but the symmetry of bar states generally requires skew symmetry of discrete gradients. The low-order method (29) is less theoretically justified than (30). For instance, the fully discrete entropy inequalities derived in Proposition 1 and [24, Thm. 4.7] can only be shown for (30). Skew symmetry of volumetric discrete gradients is therefore desirable for sequential limiters.

4.3.2 Discrete entropy stability

As discussed in Remark 3, high-resolution schemes cannot be expected to satisfy fully discrete local entropy inequalities w.r.t. all admissible entropy pairs as the low-order method does (see Proposition 1). A different concept to achieve entropy stability is to make use of Tadmor's semi-discrete theory [30], [31]. Kuzmin and Quezada de Luna [18] derived Tadmor's condition for AFC schemes and proposed a limiter that enforces local semi-discrete entropy inequalities. Further results on these techniques can be found in refs. [7], [27], [29]. Rueda-Ramírez et al. [21] were the first to use this limiter in the AFC-DG context, but no theoretical results are derived therein. Since their LGL framework uses skew-symmetric discrete gradients, one can derive results similar to the next two statements, which apply to our Bernstein-DG methods.

Proposition 4 (local semi-discrete entropy inequalities, [18], [30]). *Let $\boldsymbol{\psi}(u) = v(u)^\top \boldsymbol{f}(u) - \boldsymbol{F}(u)$, where $v = U'$, be the entropy potential and let Tadmor's condition [18], [30] for volumetric and interfacial fluxes:*

$$\frac{v_i^e - v_j^e}{2} \left[d_{ij}^e \left(u_j^e - u_i^e \right) - \left(\boldsymbol{f}_j^e + \boldsymbol{f}_i^e \right) \cdot \tilde{\boldsymbol{c}}_{ij}^e + f_{ij}^{e,*} \right] \leqslant \left(\boldsymbol{\psi}_j^e - \boldsymbol{\psi}_i^e \right) \cdot \tilde{\boldsymbol{c}}_{ij}^e \quad \forall j \in \tilde{\mathcal{N}}_i^e, \tag{31a}$$

$$\frac{v_i^e - \hat{v}_{i,k}^e}{2} \left[d_{i,k}^e \left(\hat{u}_{i,k}^e - u_i^e \right) - \left(\hat{\boldsymbol{f}}_{i,k}^e + \boldsymbol{f}_i^e \right) \cdot \boldsymbol{c}_{i,k}^e + f_{i,k}^{e,*} \right] \leqslant \left(\hat{\boldsymbol{\psi}}_{i,k}^e - \boldsymbol{\psi}_i^e \right) \cdot \boldsymbol{c}_{i,k}^e \quad \forall k \in \mathcal{F}_i^e \tag{31b}$$

hold. Then scheme (24) *satisfies the inequality:*

$$m_i^e \frac{\mathrm{d}U_i^e}{\mathrm{d}t} \leqslant \sum_{j\in\mathcal{N}_i\setminus\{i\}} \left[G_{ij}^e + \left(\boldsymbol{F}_i^e - \boldsymbol{F}_j^e\right)\cdot\tilde{\boldsymbol{c}}_{ij}^e \right] + \sum_{k\in\mathcal{F}_i^e} \left[G_{i,k}^e + \left(\boldsymbol{F}_i^e - \hat{\boldsymbol{F}}_{i,k}^e\right)\cdot\boldsymbol{c}_{i,k}^e \right], \tag{32}$$

where

$$G_{ij}^e = \frac{v_i^e + v_j^e}{2}\left[d_{ij}^e\left(u_j^e - u_i^e\right) + f_{ij}^{e,*}\right] + \frac{v_i^e - v_j^e}{2}\left(\boldsymbol{f}_i^e - \boldsymbol{f}_j^e\right)\cdot\tilde{\boldsymbol{c}}_{ij}^e,$$
$$G_{i,k}^e = \frac{v_i^e + \hat{v}_{i,k}^e}{2}\left[d_{i,k}^e\left(\hat{u}_{i,k}^e - u_i^e\right) + f_{i,k}^{e,*}\right] + \frac{v_i^e - \hat{v}_{i,k}^e}{2}\left(\boldsymbol{f}_i^e - \hat{\boldsymbol{f}}_{i,k}^e\right)\cdot\boldsymbol{c}_{i,k}^e.$$

Proof. Following [7], [18], [30], [31], we multiply (24) by v_i^e, which yields the left-hand side of (32). On the right, v_i^e is split into symmetric and antisymmetric parts. After exploiting Tadmor's condition, the rest of the proof follows from algebraic manipulations. We refer to ref. [27, Sect. 4.1] for a proof in the CG context that can directly be adapted to handle both the volumetric and interfacial terms appearing in (24). □

Corollary 2 (elementwise entropy inequalities, [18]). *Let the assumptions of Proposition* 4 *hold for all* $i \in \{1,\dots,N\}$*, and, additionally, let* $\tilde{\boldsymbol{c}}_{ij}^e = -\tilde{\boldsymbol{c}}_{ji}^e$ *for all* $i, j \in \{1,\dots,N\}$, $j \neq i$*, then* (24) *satisfies the inequality*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{K^e} U_h^e \,\mathrm{d}\boldsymbol{x} + \sum_{i=1}^N \sum_{k\in\mathcal{F}_i^e} R_{i,k}^e\left(u_h^e, \hat{u}_{h,k}^e\right) \leqslant 0, \qquad R_{i,k}^e\left(u_h^e, \hat{u}_{h,k}^e\right) := \left(\boldsymbol{F}_i^e + \hat{\boldsymbol{F}}_{i,k}^e\right)\cdot\boldsymbol{c}_{i,k}^e - G_{i,k}^e. \tag{33}$$

Proof. Summing (32) over $i \in \{1,\dots,N\}$, exploiting (16a)–(16c) (thus $G_{ij}^e = -G_{ji}^e$), and (25) we obtain:

$$\begin{aligned}
\sum_{i=1}^N m_i^e \frac{\mathrm{d}U_i^e}{\mathrm{d}t} &\leqslant \sum_{i=1}^N \sum_{j\in\mathcal{N}_i\setminus\{i\}} \left[G_{ij}^e + \left(\boldsymbol{F}_i^e - \boldsymbol{F}_j^e\right)\cdot\tilde{\boldsymbol{c}}_{ij}^e \right] + \sum_{i=1}^N \sum_{k\in\mathcal{F}_i^e} \left[G_{i,k}^e + \left(\boldsymbol{F}_i^e - \boldsymbol{F}_{i,k}^e\right)\cdot\boldsymbol{c}_{i,k}^e \right] \\
&= -\sum_{j=1}^N \boldsymbol{F}_j^e \cdot \int_{K^e} \nabla\varphi_j^e \,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\sum_{i=1}^N \boldsymbol{F}_i^e \cdot \sum_{k\in\mathcal{F}_i^e} \int_{\Gamma_k^e} \varphi_i^e \boldsymbol{n}_k^e \,\mathrm{d}s + \sum_{i=1}^N \sum_{k\in\mathcal{F}_i^e} \left[G_{i,k}^e - \boldsymbol{F}_{i,k}^e \cdot \boldsymbol{c}_{i,k}^e \right] \\
&= -\sum_{i=1}^N \boldsymbol{F}_i^e \cdot \sum_{k\in\mathcal{F}_i^e} \boldsymbol{c}_{i,k}^e + \sum_{i=1}^N \sum_{k\in\mathcal{F}_i^e} \left[G_{i,k}^e - \boldsymbol{F}_{i,k}^e \cdot \boldsymbol{c}_{i,k}^e \right] = -\sum_{i=1}^N \sum_{k\in\mathcal{F}_i^e} R_{i,k}^e\left(u_h^e, u_{h,k}^e\right).
\end{aligned}$$

Extracting the time derivative, we find that the left-hand side of this inequality is equal to $\frac{\mathrm{d}}{\mathrm{d}t}\int_{K^e} U_h^e \,\mathrm{d}\boldsymbol{x}$. □

Corollary 3 (global entropy inequality, [18]). *Let* $\hat{u}_h$ *be the input for the Riemann solver at every domain boundary segment. Under the assumptions of Corollary* 2*, scheme* (24) *satisfies the global entropy inequality:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega U_h \,\mathrm{d}\boldsymbol{x} + \sum_{e=1}^E \sum_{\substack{i=1 \\ \boldsymbol{x}_i^e\in\partial\Omega}}^N \sum_{k\in\mathcal{F}_i^e} R_{i,k}^e(u_h, \hat{u}_h) \leqslant 0. \tag{34}$$

Proof. Summing (33) over all elements, we rewrite the sum of fluxes as a sum over faces, which we split into interior and boundary faces. As is common in the DG setting, all interior fluxes cancel, yielding (34). □

Tadmor's conditions (31) is enforced as in refs. [18], [27]. In the DG context, we compute the entropy-adjusted fluxes for both, volumetric and interfacial contributions w.r.t. the same entropy pairs.

5 CFL-like time step restrictions for AFC schemes with SSP-RK

The fact that the SSP update (4) is a convex combination of admissible states is essential for many algebraic limiting techniques, e.g., [11], [13], [20], [21], [24], [25], which imposes the CFL time step restriction (5). In simple settings, the right-hand side of (5) can be evaluated, allowing for an *a priori* comparison of different baseline discretizations in terms of their efficiency. We consider three different DG discretizations of the 1D advection equation $\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0$ with constant velocity $v \in \mathbb{R}\backslash\{0\}$ on a periodic domain. Let the 1D mesh be tessellated using intervals of uniform length $h > 0$. For the new Bernstein sparsification (19), we have

$$m_i^e = \frac{h}{p+1}, \qquad d_{ij}^e = |v| \max\Big\{ |c_{ij}^e|, |c_{ji}^e| \Big\} = \frac{|v|}{2}, \qquad d_{i,k}^e = |v|\ |c_{i,k}^e| = \frac{|v|}{2}. \tag{35}$$

Due to sparsity (16d), every node has exactly two neighbors. Interior nodes within the element possess two neighbor nodes within the same element, while nodes on the element boundaries have exactly one neighbor in that element and one outside of it. Thus, $\sum_{j\in\tilde{N}_i\backslash\{i\}} d_{ij} = |v|(1/2 + 1/2) = |v|$, and (5) reduces to

$$\tau \leqslant \frac{h}{2(p+1)|v|}, \tag{36}$$

which is only slightly more restrictive than the common CFL condition $\tau \leqslant \frac{h}{(2p+1)|v|}$ [50, Eq. (2.35)] for a $(p+1)$-order DG discretization combined with a $(p+1)$-order Runge–Kutta method.

For the non-skew-symmetric Bernstein sparsification [14], [25], [32], the value of

$$\sum_{j\in\tilde{N}_i\backslash\{i\}} d_{ij}^e = |v|\Big(\max\Big\{ |\tilde{c}_{i,i-1}^e|, |\tilde{c}_{i-1,i}^e| \Big\} + \max\Big\{ |\tilde{c}_{i,i+1}^e|, |\tilde{c}_{i+1,i}^e| \Big\} \Big) \tag{37}$$

varies while m_i^e and $d_{i,k}^e$ remain unchanged compared to (35). Using [14, Eq. (B.4)], it is possible to evaluate (37) for each $i \in \{0, \dots, p\}$. To avoid tedious calculations, we only show that for $p > 1$, this scheme requires time steps τ smaller than (36). Indeed, for $i = 0$, we have

$$d_{i,1}^e + \sum_{j\in\tilde{N}_i\backslash\{i\}} d_{ij}^e = \frac{|v|}{2} + |v| \max\Big\{ |c_{0,1}^e|, |c_{1,0}^e| \Big\} = |v| \left[\frac{1}{2} + \frac{1}{p+1} \max\{ p, |-1| \} \right] = |v| \frac{(p+1)+2p}{2(p+1)},$$

which is larger than $|v|$ if $p > 1$. Thus, the use of (19) instead of our old approach [14], [25], [32] leads to less restrictive CFL conditions for Bernstein elements.

Finally, we study the case of LGL discretizations [20], [21], in which $m_i^e = h\omega_i$, $|c_{ij}^e| = \frac{1}{2}\ \forall j \in \tilde{\mathcal{N}}_i\backslash\{i\}$, $|c_{ik}^e| \equiv \frac{1}{2}$, where ω_i are the LGL quadrature weights. For $p > 2$, the LGL nodes are not uniformly distributed within the elements but are concentrated around the element boundaries. Generally, the weights of the two boundary nodes for the unit interval $[0, 1]$ are given by $\omega_0 = \omega_p = \frac{1}{p(p+1)}$. Thus, the largest possible time step satisfying the CFL condition (5) is bounded by $\frac{h}{2p(p+1)|v|}$, which is more restrictive than the optimal Bernstein-CFL (36) by a factor of p. This unfortunate drawback of LGL discretizations could potentially be cured by replacing elementwise discrete gradient operators [20], [21] (which are identical to the ones we use here for the Bernstein ansatz) with matrices that account for variations in the LGL quadrature weights.

In conclusion, for $p > 1$, our new sparsification can use somewhat larger time steps (depending on p) than the previous one [14], [25], [32], while improvements over time steps for LGL are significant. For $p = 1$, Bernstein polynomials are simply the Lagrange basis functions, which makes the scheme equivalent to the LGL space discretization. In 1D and on box elements, the old and new sparsifications are also identical.

Remark 4. The comparisons made here can be easily adapted to discretizations of nonlinear systems in multiple space dimensions, where the same conclusions can be drawn. For instance, the multidimensional version of the optimal Bernstein-CFL (36) reads

$$\tau \leqslant \frac{m_i}{2\sum_{j\neq i} d_{ij}} = \frac{[h/(p+1)]^d}{2\sum_{j\neq i}\frac{1}{2}|\boldsymbol{v}\cdot\boldsymbol{n}_{ij}|[h/(p+1)]^{d-1}} = \frac{h}{(p+1)\sum_{j\neq i}|\boldsymbol{v}\cdot\boldsymbol{n}_{ij}|}, \tag{38}$$

where the $\boldsymbol{n}_{ij} \in \mathbb{R}^d$ are either the volumetric or interfacial integrals in the stencil.

Remark 5. The CFL restrictions derived here apply to low-order and flux-limited approximations. CFL numbers w.r.t. the target scheme may be more restrictive, see e.g., [51], [52] for further analysis on this topic. Violations of CFL conditions typically lead to severe blowups, in which case the limiter would adapt the unstable target solution to satisfy all constraints that are being enforced. In this manner, one can, for instance, guarantee nonnegativity. However, it is usually preferable to limit an already stable target scheme rather than to rely on limiting for that purpose as well. The numerical example in Section 6.1 provides an illustration of how our limiters work if applied to such an unstable baseline discretization. CFL conditions of standard DG methods combined with appropriate time stepping schemes can be relaxed by modifying the numerical fluxes [53]. While this may lead to a loss of the superconvergence property, all other advantages of standard DG remain valid. In practice, we have not observed any cases, where the DG target discretization combined with SSP-RK schemes produces more restrictive CFL conditions than the limiter does.

6 Numerical experiments

We now apply the proposed AFC schemes to common test problems for the compressible Euler equations, where $\gamma = 1.4$ by default. The application to other conservation laws is straightforward. Time discretization is performed using the SSP Runge–Kutta paradigm [37]–[39]. The methods are implemented in the open-source C++ library `mfem` [54], [55] and snapshots are visualized with the accompanying GLVis toolbox [56].

To distinguish between different variants of MCL schemes, we use the following conventions: If all antidiffusive fluxes are set to zero, the *low order* method is employed. The sequential limiter enforcing discrete maximum principles following [25], shall be labelled 'seq'. All MCL-type schemes enforcing global bounds are referred to by a tuple or triple of quantities that are being limited, e.g., (ρ, p) first enforces nonnegativity of density followed by the pressure fix, while (ρ, p, U) subsequently also enforces Tadmor's entropy stability conditions (31). Often, we append a suffix indicating which numerical flux (LLF, HLL, etc.) is used. To distinguish between the nonskew-symmetric and skew-symmetric sparsifications, we simply use the terminology *old* versus *new*. Unless stated otherwise, we use adaptive SSP time stepping of the same order as the polynomial space, e.g., SSP2 for $p = 1$. The time step is set equal to $\nu \in (0, 1]$ times the right-hand side of (5), where the constant ν is chosen to be 0.5 in all tests for the Euler equations.

6.1 1D blast wave

We begin our experiments with a challenging 1D test problem studied first in ref. [57]. The spatial domain $\Omega = (0, 1)$ is equipped with reflecting wall boundaries, and the initial flow configuration is as follows:

$$\rho_0 \equiv 1, \qquad v_0 \equiv 0, \qquad p_0(x) = \begin{cases} 1,000 & \text{if } x < 0.1, \\ 0.01 & \text{if } 0.1 < x < 0.9, \\ 100 & \text{if } 0.9 < x. \end{cases}$$

The solution exhibits strong shocks due to the vastly different pressure values, which makes this benchmark an extreme test for preservation of positive internal energies. For small times, the exact solution can be obtained by

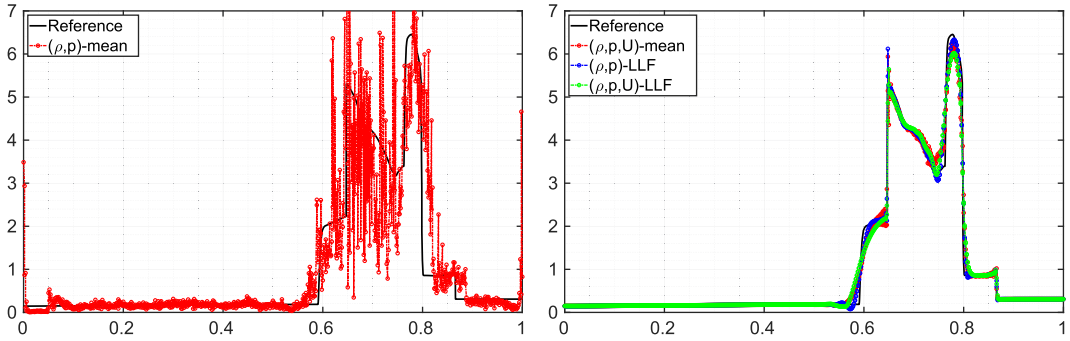

Figure 3: Density profiles of the 1D blast wave solved on 500 $\mathbb{P}_1$ elements.

solving two Riemann problems [58, Ch. 4] but at the end time $t = 0.038$, interactions with the domain boundaries and shock collisions have occurred. Therefore, only reference solutions are available.

Since these do not contain any highly oscillatory features, the most economical choice in terms of accuracy versus computational cost is to use second-order schemes on rather fine meshes. We use our $\mathbb{P}_1$-DG discretization to solve this problem on a grid consisting of 500 uniformly spaced elements. As our MCL approach is capable of using any numerical flux and performing limiting to guarantee various constraints, we begin by using *mean value fluxes* $f_{\boldsymbol{n}}(u,v) = \frac{1}{2}(\boldsymbol{f}(u)+\boldsymbol{f}(v))\cdot\boldsymbol{n}$ and ensure only nonnegativity of density and pressure. Since limiting is carried out w.r.t. global bounds only, the result in the left panel of Figure 3 is extremely poor. Performing entropy limiting subsequently to the other steps significantly enhances the solution quality, as seen in the red curve in the right panel of Figure 3 but some spurious oscillations remain.

A more reasonable approach than using mean value fluxes is to employ the local Lax–Friedrichs Riemann solver $f_{\boldsymbol{n}}(u,v) = \frac{1}{2}[(\boldsymbol{f}(u)+\boldsymbol{f}(v))\cdot\boldsymbol{n} + \lambda(u-v)]$. We combine this scheme with (ρ,p)-limiting for the volume terms (the interfacial LLF fluxes do not require limiting to ensure any properties in the 1D case [25]). Comparing this result with the (ρ,p,U)-limited profile, we observe less smearing of the contact discontinuity at $x \approx 0.6$, which occurs if entropy limiting is enabled. We investigate this issue further in the following section. At this stage, it is safe to say that entropy limiting in our context can drastically improve unstable schemes by adding sufficient amounts of dissipation. The question is, are these amounts also necessary?

6.2 2D isentropic vortex

Let us now numerically assess the convergence order of schemes by considering an example with a smooth solution [59, Sect. 5.1]. Here, the doubly-periodic domain is $\Omega = (-5,5)^2$ and at every time $t = 10k$, $k \in \mathbb{N}$, the solution coincides with the initial condition, which reads

$$\rho_0(x,y) = \theta_0(x,y)^{1/(\gamma-1)}, \qquad \boldsymbol{v}_0(x,y) = \begin{bmatrix}1\\1\end{bmatrix} + \frac{\varepsilon}{2\pi}\,\mathrm{e}^{0.5(1-(x^2+y^2))}\begin{bmatrix}-y\\x\end{bmatrix},$$

$$p_0(x,y) = \theta_0(x,y)^{\gamma/(\gamma-1)}, \qquad \theta_0(x,y) = 1 - \frac{(\gamma-1)\,\varepsilon^2}{8\gamma\pi^2}\mathrm{e}^{1-(x^2+y^2)}, \qquad \varepsilon = 5.$$

We use structured, uniform meshes and perform the tests that were alluded to in Section 3. In addition, the convergence rate with entropy limiting enabled is checked for the $\mathbb{Q}_1$ space. Other than entropy stability, only nonnegativity of density and pressure are enforced. Since the solution is smooth, we do not rely on limiters enforcing local bounds (these would deteriorate optimal rates). In Tables 1–3, we present the $\mathrm{L}^1(\Omega)$ errors in density at the final time $t = 10$ and the corresponding experimental orders of convergence (EOC).

First, we observe that the entropy fix leads to a catastrophic reduction of the convergence order. This serves as an explanation for our earlier results, in particular, it explains why the entropy fix is capable of fixing the oscillatory profile on the left of Figure 3. It is still shocking that the order is reduced by that much. Since the

Table 1: Isentropic vortex, $\mathbb{Q}_1$-HLL on square cells.

$\frac{10}{h}$	(ρ, p, U)	EOC	(ρ, p)	EOC
32	1.76E-02	—	9.32E-04	—
64	1.23E-02	0.52	1.58E-04	2.56
128	7.56E-03	0.70	2.90E-05	2.44
256	4.29E-03	0.82	6.28E-06	2.21
Avg. EOC		**0.68**		**2.40**

Table 2: Isentropic vortex, $\mathbb{Q}_2$-(ρ, p) on square cells.

$\frac{10}{h}$	New-LLF	EOC	New-HLL	EOC	Old-HLL	EOC
32	5.07E-05	—	2.60E-05	—	2.60E-05	—
64	7.84E-06	2.69	2.35E-06	3.47	2.35E-06	3.47
128	1.22E-06	2.68	2.70E-07	3.12	2.70E-07	3.12
256	1.79E-07	2.78	3.31E-08	3.03	3.31E-08	3.03
Avg. EOC		**2.72**		**3.21**		**3.21**

Table 3: Isentropic vortex, (ρ, p)-HLL on triangles with different gradients, compare Figure 1.

$\frac{\sqrt{200}}{h}$	$\mathbb{P}_1$, Figure 1a	EOC	$\mathbb{P}_1$, Figure 1b	EOC	$\mathbb{P}_1$, Figure 1d	EOC	$\mathbb{P}_2$, Figure 1a	EOC	$\mathbb{P}_2$, Figure 1d	EOC
32	1.07E-03	—	3.24E-03	—	9.78E-04	—	4.08E-05	—	4.08E-05	—
64	1.66E-04	2.69	1.13E-03	1.52	1.66E-04	2.56	3.60E-06	3.50	3.60E-06	3.50
128	3.08E-05	2.43	4.16E-04	1.44	3.08E-05	2.43	3.96E-07	3.18	3.96E-07	3.18
256	6.70E-06	2.20	1.88E-04	1.15	6.70E-06	2.20	4.85E-08	3.03	4.85E-08	3.03
Avg. EOC		**2.44**		**1.37**		**2.40**		**3.24**		**3.24**

solution to this test is smooth, its exact entropy integrated over the domain remains constant in time. Numerical schemes may fail to mirror this behavior due to quadrature errors, numerical entropy production, or dissipation. It seems that Tadmor's entropy fix designed to make discretizations entropy *conservative* (not dissipative) does not work well in the DG case even for $\mathbb{Q}_1$ spaces (where Bernstein polynomials are simply the nodal Lagrange basis functions). In principle, an unnecessarily high rate of entropy dissipation can be attributed to too diffusive numerical fluxes such as LLF or HLL. However, the culprit here is clearly the volumetric entropy limiting, as we obtain second-order rates by disabling this fix. A comparison with the CG case and similar entropy-limited AFC schemes is in order to gain further understanding of this issue and how to resolve it. Second, we notice that LLF fluxes are too diffusive to exhibit third-order convergence for $\mathbb{Q}_2$ discretizations [28, Sect. 6.3]. This problem can be cured using more accurate numerical fluxes such as HLL. Finally, we confirm the results regarding the different gradients. The old and new sparsification approaches are almost identical in terms of error values and convergence orders. In particular, the fact that three leading digits in the error values are identical for third-order schemes is striking given that the new sparsification requires noticeably fewer time steps, cf. Table 5. Out of the five triangular gradients in Figure 1 only (b) does not converge with optimal accuracy (rates for Figures 1c and 1e not printed for brevity) but the gradient in Figure 1d generally requires the fewest time steps, cf. Tables 4 and 5.

Table 4: Isentropic vortex: number of time steps required by $\mathbb{P}_1$-HLL on triangles with different gradients, compare Figure 1.

$\frac{\sqrt{200}}{h}$	Figure 1a (old)	Figure 1b	Figure 1c	Figure 1d (new)	Figure 1e
32	1771	2,121	2,552	1742	4,130
64	3,557	4,254	5,080	3,453	8,272
128	7,068	8,496	10,119	6,866	16,535
256	14,126	16,966	20,193	13,688	33,062

Table 5: Isentropic vortex: number of time steps required by $\mathbb{P}_2$-HLL on triangles with different gradients, compare Figure 1, and $\mathbb{Q}_2$-HLL on square cells with old and new sparsification approaches.

$\frac{\sqrt{200}}{h}$	$\mathbb{P}_2$, Figure 1a (old)	$\mathbb{P}_2$, Figure 1b	$\mathbb{P}_2$, Figure 1c	$\mathbb{P}_2$, Figure 1d (new)	$\frac{10}{h}$	$\mathbb{Q}_2$ (old)	$\mathbb{Q}_2$ (new)
32	2,833	2,856	4,613	2,497	32	2,800	2,100
64	5,659	5,657	9,216	4,954	64	5,557	4,168
128	11,304	11,298	18,419	9,863	128	11,076	8,307
256	22,595	22,587	36,820	19,681	256	22,119	16,589

6.3 Modified Sod shock tube

Sod's shock tube is a common test case that we solved with Bernstein-DG-MCL and the old sparsification in ref. [25]. A popular variant of this benchmark [58, Sect. 8.5.1] slightly changes the classical setup such that a sonic point appears in the exact solution. Many numerical schemes fail in the proximity of this location, where one of the eigenvalues of the flux changes its sign, by producing a nonphysical entropy shock (see Figure 4a). The setup of this test is as follows: the domain $\Omega = (0, 1)$ has an inlet boundary on the left and an outlet on the right. The inlet boundary profile is identical to u_L for all times, where

$$u_0(x) = \begin{cases} u_L & \text{if } x < 0.25, \\ u_R & \text{if } x > 0.25, \end{cases} \qquad u_L = (1, 0.75, 2.78125)^\top, \qquad u_R = (0.125, 0, 0.25)^\top,$$

are the initial conditions. This Riemann problem admits an exact entropy solution that can be determined with arbitrary accuracy by solving a nonlinear equation for a single pressure state [58, Ch. 4]. It consists of a left rarefaction wave, a right shock wave, and contains a contact discontinuity in between.

First, we solve this problem up to $t = 0.2$ with continuous $\mathbb{P}_1$ elements and Roe [60] target fluxes (see [27] for details) employing uniform meshes of increased resolution and fixed time steps determined by the most restrictive CFL condition. The profiles in Figure 4a illustrate the appearance of the entropy shock.

Next, we use a coarser uniform mesh with only 16 uniform DG elements and polynomial degree $p = 7$. The result of the (ρ, p, U)-limited solution is displayed in Figure 4b. These profiles are obtained with the new sparsification approach but the old one looks quite similar. The availability of the exact solution allows us to compute the exact entropy evolution for this test. Luckily, in this example, the entropy flux $\boldsymbol{F}(u)$ at both domain boundaries is zero. Thus, entropy is only changed via dissipation at nonsmooth solution features. We see from Figure 4c and d that entropy decays linearly. Since we use entropy fixes, one would hope that this property holds true for all discrete entropy evolutions as well but we see that this is only the case for the new sparsification. Indeed, all three considered Riemann solvers produce entropy with the old sparsification at the beginning of the simulation. Since the old sparsification does not use skew-symmetric discrete gradients, this result does not contradict Corollary 3. Besides provable semi-discrete entropy stability, the new sparsification also has the advantage of needing significantly fewer time steps than the old one.

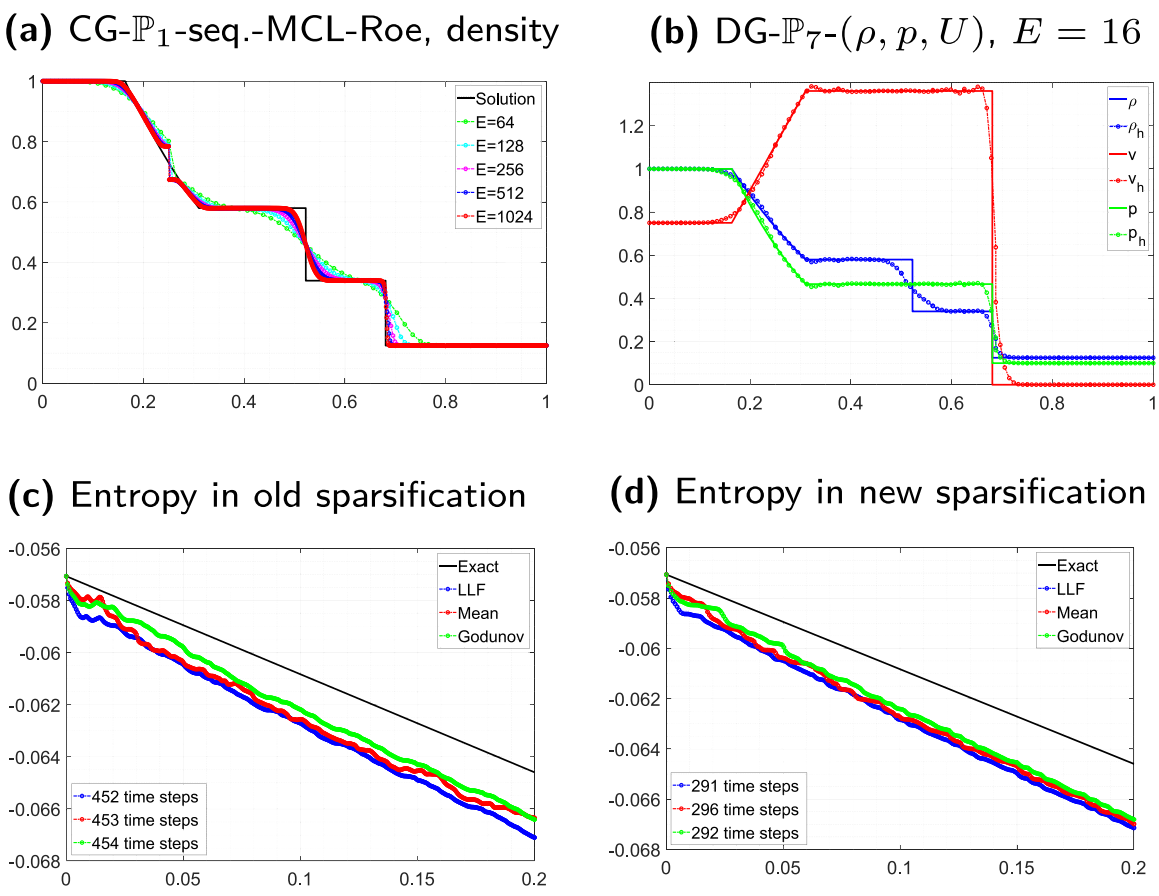

Figure 4: Modified Sod shock tube: solution profiles (a)–(b) and entropy evolutions (c)–(d). DG solutions use adaptive SSP3.

A few further remarks are in order. First, the red curve in Figure 4d is actually not monotone. This is also not in contradiction to Corollary 3 because our entropy fixes guarantee semi-discrete entropy stability rather than in the fully discrete setting. It is known that explicit time stepping schemes may result in entropy production proportional to a certain power of the time step [61]. Time step reductions can be used to make the red curve in Figure 4d monotonically decreasing. However, this approach does not remove the wiggles observed in Figure 4c, which are therefore indeed due to the spatial discretization, which fails to dissipate entropy even though an entropy limiter is used. Second, despite these small entropy productions, we do not actually observe the formation of entropy shocks. This is consistent with our previous assessment [27] that entropy limiting does not seem to be necessary for this system. Instead, preservation of nonnegotiable constraints and shock-capturing should be implemented. Finally, if we use the sequential approach in combination with either sparsification, all entropy evolutions become monotone. This observation shows that enforcing constraints such as maximum principles via AFC leads to dissipation of entropy even though entropy does not play a role in the sequential limiter. Since sufficient entropy dissipation cannot be guaranteed by this scheme, our above arguments regarding entropy evolutions remain valid and we recommend using the new sparsification rather than the old one.

6.4 High-Mach number astrophysical jet

Finally, we apply our schemes to a hypersonic test problem, in which the Mach number is more than 2,000 [62]. The spatial domain is $\Omega = (0, 1)^2$, the adiabatic constant is $\gamma = 5/3$, and the initial conditions for the primitive

unknowns are $\rho_0 \equiv 0.5$, $\boldsymbol{v}_0 \equiv 0$, $p_0 \equiv 0.4127$. A jet enters at the left boundary $\{0\} \times [0,1]$, where the inflow profile is set according to

$$(\rho, \boldsymbol{v}^\top, p)(0, y, t) = \begin{cases} (5, [800, 0], 0.4127) & \text{if } |y - 0.5| \leqslant 0.05, \\ (0.5, [0, 0], 0.4127) & \text{otherwise,} \end{cases} \qquad y \in [0,1].$$

For the end time $t = 10^{-3}$, the other three boundaries can be set somewhat arbitrarily (we use free outflow).

We solve this problem numerically on a uniform quadrilateral mesh with 512^2 cells and an unstructured triangular mesh with 543,744 elements. Generally, we compare results of the (ρ, p)-limiter with the sequential approach using $\mathbb{Q}_p$ and $\mathbb{P}_p$ spaces with $p \in \{1, 2\}$ and contrast the old and new sparsification approaches (not for $\mathbb{Q}_1$, as old and new are identical for these). The results are shown in Figures 5–8; Figure 5 additionally displays the low order solution. Note that the color bar (which is the same for all plots) is logarithmic. The number of required time steps (t.s.) is shown in the figure captions for each snapshot. These were computed using HLL fluxes [23] with wave speed estimates

$$s_- = \min\{\boldsymbol{v}_L \cdot \boldsymbol{n} - a_L, \boldsymbol{v}_R \cdot \boldsymbol{n} - a_R\}, \qquad s_+ = \max\{\boldsymbol{v}_L \cdot \boldsymbol{n} + a_L, \boldsymbol{v}_R \cdot \boldsymbol{n} + a_R\},$$

where L and R denote the two input states of the flux, $\boldsymbol{n}$ is the normal, and $a = \sqrt{\gamma p/\rho}$ the sound speed.

Overall, all numerical approximations agree well with each other and no unsurprising results are observed. The two sparsification approaches produce essentially indistinguishable results but the reduced number of time steps makes us favor the new technique. As before, sequential limiting usually requires fewer time steps but in this example, there are some exceptions ($\mathbb{P}_1$ with either sparsification, $\mathbb{P}_2$ with the new one). The sequential limiter introduces some additional streaks along the transition between the head of the jet and the background. It seems that these features are spurious and can be attributed to a choice of bounds for numerical admissibility that are not well-suited for this example. On the other hand, only enforcing nonnegativity of density and pressure

(a) Low order method, 5 696 t.s. **(b)** (ρ, p), 6 327 t.s. **(c)** seq., 6 141 t.s.

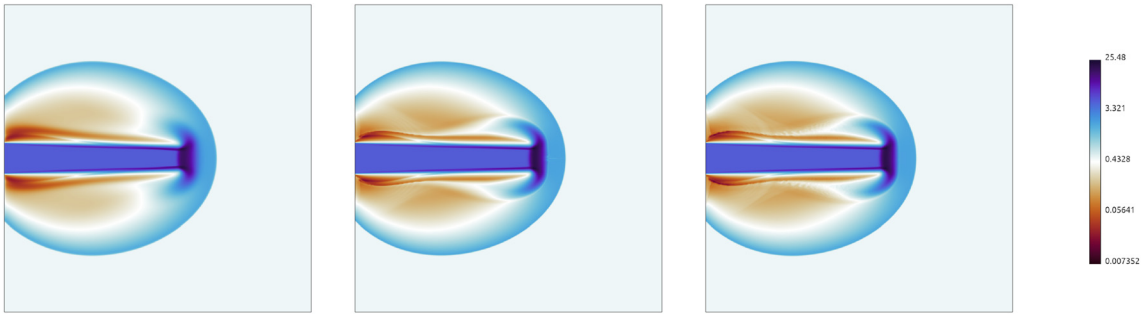

Figure 5: Density profiles of the astrophysical jet on a uniform quadrilateral mesh, $\mathbb{Q}_1$ solutions.

(a) Old-(ρ, p), 13 086 t.s. **(b)** New-(ρ, p), 9 638 t.s. **(c)** seq.-old, 12 471 t.s. **(d)** seq.-new, 9 276 t.s.

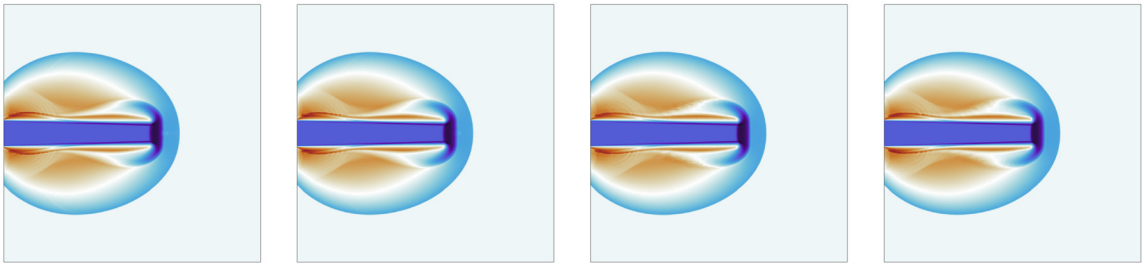

Figure 6: Density profiles of the astrophysical jet on a uniform quadrilateral mesh, $\mathbb{Q}_2$ solutions.

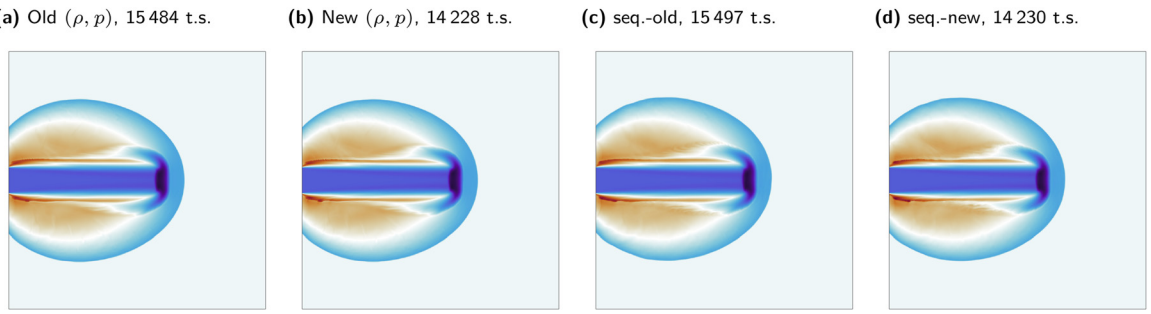

Figure 7: Density profiles of the astrophysical jet on an unstructured triangular mesh, $\mathbb{P}_1$ solutions.

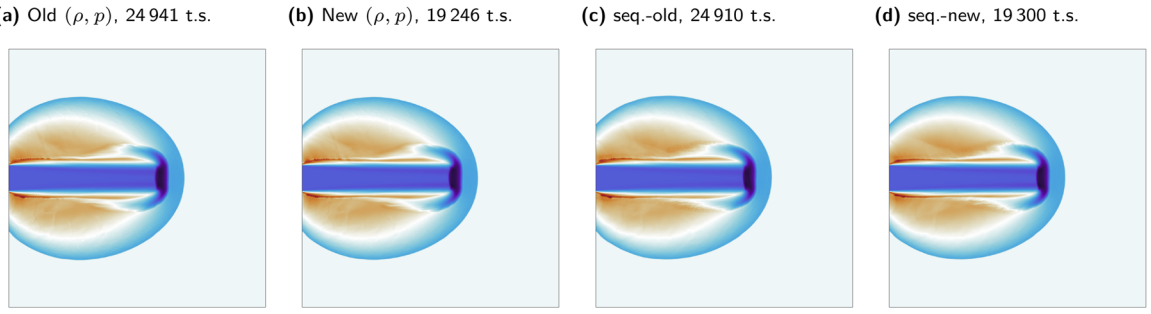

Figure 8: Density profiles of the astrophysical jet on an unstructured triangular mesh, $\mathbb{P}_2$ solutions.

seems to work quite well here despite the severity of this test case. Naturally, there exist examples where the oscillations arising because the numerical viscosity along shocks is too low will reverse the situation. Finding a framework that introduces precisely the right amount of diffusion in high-order methods, which also converge with optimal orders of accuracy is an open problem.

7 Conclusions

A new, skew-symmetric variant of discrete gradient operators, commonly used in finite element methods, and in particular algebraic flux correction schemes, was proposed and analyzed. For continuous and discontinuous Galerkin discretizations using Bernstein basis functions, the feasibility of the novel techniques was demonstrated in various geometries. We used the new operators in the context of our monolithic convex limiting scheme and compared it with existing techniques, based on non-skew-symmetric discrete gradients and LGL approaches. In the context of semi-discrete entropy stabilizations via MCL, the necessity of having skew-symmetric off-diagonal entries was demonstrated numerically. Moreover, additional theoretical results for the new and old versions of this approach were derived. Finally, we showed the optimality of the novel technique in terms of explicit time step restrictions. Numerical tests demonstrated that the new version of the scheme performs well in terms of shock capturing and also preserves optimal orders of accuracy.

The discrete gradients developed in this work are by no means restricted to flux-correction schemes nor monolithic limiting. In fact, they can readily be employed in the context of FCT algorithms for both continuous and discontinuous ansatz spaces using elementwise operators. Potential future studies include their use for other types of equations, such as incompressible flows [63]. Moreover, we plan to further develop our flux-correction techniques by adding additional features to the described methodology. These include smoothness indicators [47], flux limiting using general Runge–Kutta time discretizations [33], as well as the inclusion of diffusive [40] and source [12] terms. These topics shall be addressed in future publications.

Acknowledgments: The author would also like to thank the two anonymous reviewers whose comments helped to significantly improve this manuscript.
Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The author has accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: The author states no conflict of interest.
Research funding: This work was partially supported by The Rough Ocean Project, funded by the Research Council of Norway under the Klimaforsk-programme, Project 302743.
Data availability: Not applicable.

References

[1] R. Abgrall and J. Trefilík, "An example of high order residual distribution scheme using non-Lagrange elements," *J. Sci. Comput.*, vol. 45, pp. 3—25, 2010.
[2] T. J. Barth and D. C. Jespersen, "The design and application of upwind schemes on unstructured meshes," in *27th Aerospace Sciences Meeting*, American Institute of Aeronautics and Astronautics, 1989.
[3] E. Gaburro, P. Öffner, M. Ricchiuto, and D. Torlo, "High order entropy preserving ADER-DG schemes," *Appl. Math. Comput.*, vol. 440, p. 127644, 2023.
[4] Y. Lin and J. Chan, "High order entropy stable discontinuous Galerkin spectral element methods through subcell limiting," *J. Comput. Phys.*, vol. 498, p. 112677, 2024.
[5] S. Faghih-Naini and V. Aizinger, "*p*-adaptive discontinuous Galerkin method for the shallow water equations with a parameter-free error indicator," *Int. J. Geomath.*, vol. 13, p. 18, 2022.
[6] J. Vedral, "Dissipative weno stabilization of high-order discontinuous Galerkin methods for hyperbolic problems," *Preprint, arXiv: 2309.12019 [math.NA]*, 2023.
[7] D. Kuzmin and H. Hajduk, *Property-preserving Numerical Schemes for Conservation Laws*, London, UK, World Scientific Europe, 2023.
[8] L. Micalizzi, D. Torlo, and W. Boscheri, "Efficient iterative arbitrary high-order methods: an adaptive bridge between low and high order," *Commun. Appl. Math. Comput.*, vol. 7, pp. 40—77, 2023.
[9] D. Kuzmin, R. Löhner, and S. Turek, *Flux-Corrected Transport*, 2nd ed. Springer, 2012.
[10] R. Anderson, *et al*., "High-order local maximum principle preserving (MPP) discontinuous Galerkin finite element method for the transport equation," *J. Comput. Phys.*, vol. 334, pp. 102—124, 2017.
[11] 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. Comput.*, vol. 40, pp. A3211—A3239, 2018.
[12] H. Hajduk and D. Kuzmin, "Bound-preserving and entropy-stable algebraic flux correction schemes for the shallow water equations with topography," in *Eleventh International Conference on Computational Fluid Dynamics*, ICCFD11 Proceedings, 2022 [Online]. Available at: https://www.iccfd.org/iccfd11/assets/pdf/papers/ICCFD11_Paper-3003.pdf.
[13] D. Kuzmin, "Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws," *Comput. Methods Appl. Mech. Eng.*, vol. 361, p. 112804, 2020.
[14] 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.*, vol. 344, pp. 151—186, 2017.
[15] G. R. Barrenechea, V. John, and P. Knobloch, "Finite element methods respecting the discrete maximum principle for convection—diffusion equations," *SIAM Rev.*, vol. 66, pp. 3—88, 2024.
[16] C. Lohmann, *Physics-Compatible Finite Element Methods for Scalar and Tensorial Advection Problems*, Springer Spektrum, 2019.
[17] H. Hajduk and A. Rupp, "Analysis of algebraic flux correction for semi-discrete advection problems," *BIT Numer. Math.*, vol. 63, p. 8, 2023.
[18] D. Kuzmin and M. Quezada de Luna, "Algebraic entropy fixes and convex limiting for continuous finite element discretizations of scalar hyperbolic conservation laws," *Comput. Methods Appl. Mech. Eng.*, vol. 372, p. 113370, 2020.
[19] D. Kuzmin and S. Turek, "Flux correction tools for finite elements," *J. Comput. Phys.*, vol. 175, no. 2, pp. 525—558, 2002.
[20] W. Pazner, "Sparse invariant domain preserving discontinuous Galerkin methods with subcell convex limiting," *Comput. Methods Appl. Mech. Eng.*, vol. 382, p. 113876, 2021.
[21] A. M. Rueda-Ramírez, B. Bolm, D. Kuzmin, and G. J. Gassner, "Monolithic convex limiting for Legendre—Gauss—Lobatto discontinuous Galerkin spectral-element methods," *Commun. Appl. Math. Comput.*, vol. 6, pp. 1860—1898, 2024.
[22] D. Hoff, "A finite difference scheme for a system of two conservation laws with artificial viscosity," *Math. Comput.*, vol. 33, pp. 1171—1193, 1979.

[23] A. Harten, P. D. Lax, and B. van Leer, “On upstream differencing and Godunov-type schemes for hyperbolic conservation laws,” *SIAM Rev.*, vol. 25, no. 1, pp. 35—61, 1983.
[24] J.-L. Guermond and B. Popov, “Invariant domains and first-order continuous finite element approximation for hyperbolic systems,” *SIAM J. Numer. Anal.*, vol. 54, no. 4, pp. 2466—2489, 2016.
[25] H. Hajduk, “Monolithic convex limiting in discontinuous Galerkin discretizations of hyperbolic conservation laws,” *Comput. Math. Appl.*, vol. 87, pp. 120—138, 2021.
[26] P. Moujaes and D. Kuzmin, “Monolithic convex limiting and implicit pseudo-time stepping for calculating steady-state solutions of the Euler equations,” *Preprint, arXiv: 2407.03746 [math.NA]*, 2024.
[27] D. Kuzmin, H. Hajduk, and A. Rupp, “Limiter-based entropy stabilization of semi-discrete and fully discrete schemes for nonlinear hyperbolic problems,” *Comput. Methods Appl. Mech. Eng.*, vol. 389, p. 114428, 2022.
[28] H. Hajduk, “Algebraically constrained finite element methods for hyperbolic problems with applications in geophysics and gas dynamics,” Ph.D. dissertation, TU Dortmund University, 2022.
[29] D. Kuzmin and M. Quezada de Luna, “Entropy conservation property and entropy stabilization of high-order continuous Galerkin approximations to scalar conservation laws,” *Comput. Fluids*, vol. 213, p. 104742, 2020.
[30] E. Tadmor, “The numerical viscosity of entropy stable schemes for systems of conservation laws, I,” *Math. Comput.*, vol. 49, pp. 91—103, 1987.
[31] E. Tadmor, “Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems,” *Acta Numer.*, vol. 12, pp. 451—512, 2003.
[32] D. Kuzmin and M. Quezada de Luna, “Subcell flux limiting for high-order Bernstein finite element discretizations of scalar hyperbolic conservation laws,” *J. Comput. Phys.*, vol. 411, p. 109411, 2020.
[33] D. Kuzmin, M. Quezada de Luna, D. I. Ketcheson, and J. Grüll, “Bound-preserving flux limiting for high-order explicit Runge—Kutta time discretizations of hyperbolic conservation laws,” *J. Sci. Comput.*, vol. 91, 2022, Art. no. 21.
[34] T. C. Fisher and M. H. Carpenter, “High-order entropy stable finite difference schemes for nonlinear conservation laws: finite domains,” *J. Comput. Phys.*, vol. 252, pp. 518—557, 2013.
[35] G. J. Gassner, “A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to sbp-sat finite difference methods,” *SIAM J. Sci. Comput.*, vol. 35, no. 3, pp. A1233—A1253, 2013.
[36] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer, 2000.
[37] C.-W. Shu and S. Osher, “Efficient implementation of essentially non-oscillatory shock-capturing schemes,” *J. Comput. Phys.*, vol. 77, no. 2, pp. 439—471, 1988.
[38] S. Gottlieb, C.-W. Shu, and E. Tadmor, “Strong stability-preserving high-order time discretization methods,” *SIAM Rev.*, vol. 43, no. 1, pp. 89—112, 2001.
[39] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu, *Strong Stability Preserving Runge—Kutta and Multistep Time Discretizations*, World Scientific, 2011.
[40] M. Quezada de Luna and D. I. Ketcheson, “Maximum principle preserving space and time flux limiting for diagonally implicit Runge—Kutta discretizations of scalar convection—diffusion equations,” *J. Sci. Comput.*, vol. 92, p. 102, 2022.
[41] A. Harten, “On the symmetric form of systems of conservation laws with entropy,” *J. Comput. Phys.*, vol. 49, no. 1, pp. 151—164, 1983.
[42] X. Zhang, “On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier—Stokes equations,” *J. Comput. Phys.*, vol. 328, pp. 301—343, 2017.
[43] R. Abgrall, “Essentially non-oscillatory residual distribution schemes for hyperbolic problems,” *J. Comput. Phys.*, vol. 214, no. 2, pp. 773—808, 2006.
[44] D. Kuzmin, M. Möller, J. N. Shadid, and M. Shashkov, “Failsafe flux limiting and constrained data projections for equations of gas dynamics,” *J. Comput. Phys.*, vol. 229, no. 23, pp. 8766—8779, 2010.
[45] C. A. J. Fletcher, “The group finite element formulation,” *Comput. Methods Appl. Mech. Eng.*, vol. 37, no. 2, pp. 225—244, 1983.
[46] G. R. Barrenechea and P. Knobloch, “Analysis of a group finite element formulation,” *Appl. Numer. Math.*, vol. 118, pp. 238—248, 2017.
[47] A. M. Rueda-Ramírez, W. Pazner, and G. J. Gassner, “Subcell limiting strategies for discontinuous Galerkin spectral element methods,” *Comput. Fluids*, vol. 247, p. 105627, 2022.
[48] H. Hajduk, “Preconditioned gradient matrix on the reference simplex,” 2020. Available at: https://github.com/HennesHajduk/PrecMatSimplex.
[49] V. Dobrev, T. Kolev, D. Kuzmin, R. Rieben, and V. Tomov, “Sequential limiting in continuous and discontinuous Galerkin methods for the Euler equations,” *J. Comput. Phys.*, vol. 356, pp. 372—390, 2018.
[50] B. Cockburn and C.-W. Shu, “TVB Runge—Kutta local projection discontinuous Galerkin finite element method for conservation laws, II. General framework,” *Math. Comput.*, vol. 52, pp. 411—435, 1989.
[51] N. Chalmers and L. Krivodonova, “A robust CFL condition for the discontinuous Galerkin method on triangular meshes,” *J. Comput. Phys.*, vol. 403, p. 109095, 2020.
[52] L. Krivodonova and R. Qin, “An analysis of the spectrum of the discontinuous Galerkin method,” *Appl. Numer. Math.*, vol. 64, pp. 1—18, 2013.

[53] N. Chalmers, L. Krivodonova, and R. Qin, "Relaxing the CFL number of the discontinuous Galerkin method," *SIAM J. Sci. Comput.*, vol. 36, no. 4, pp. A2047—A2075, 2014.
[54] R. Anderson, *et al*., "MFEM: a modular finite element methods library," *Comput. Math. Appl.*, vol. 81, pp. 42—74, 2021.
[55] "MFEM: modular finite element methods [software]," Available at: https://mfem.org.
[56] "GLVis: OpenGL finite element visualization tool," Available at: https://glvis.org.
[57] P. Woodward and P. Colella, "The numerical simulation of two-dimensional fluid flow with strong shocks," *J. Comput. Phys.*, vol. 54, no. 1, pp. 115—173, 1984.
[58] E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, 3rd ed. Springer, 2009.
[59] C.-W. Shu, "Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws," in *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, ser. Lecture Notes in Mathematics, Springer, 1998, pp. 325—432.
[60] P. L. Roe, "Approximate Riemann solvers, parameter vectors, and difference schemes," *J. Comput. Phys.*, vol. 43, no. 2, pp. 357—372, 1981.
[61] C. Lozano, "Entropy production by explicit Runge—Kutta schemes," *J. Sci. Comput.*, vol. 76, pp. 521—564, 2018.
[62] Y. Ha, C. L. Gardner, A. Gelb, and C.-W. Shu, "Numerical simulation of high Mach number astrophysical jets with radiative cooling," *J. Sci. Comput.*, vol. 24, pp. 29—44, 2005.
[63] H. Hajduk, D. Kuzmin, G. Lube, and P. Öffner, "Locally energy-stable finite element schemes for incompressible flow problems: design and analysis for equal-order interpolations," *Comput. Fluids*, vol. 294, pp. 106622, 2025.