BIT Numerical Mathematics (2023) 63:8
https://doi.org/10.1007/s10543-023-00957-z

BIT

Analysis of algebraic flux correction schemes for semi-discrete advection problems

Hennes Hajduk[1] · Andreas Rupp[2]

Received: 25 April 2022 / Accepted: 7 October 2022 / Published online: 30 January 2023

Abstract

Based on recent developments regarding the analysis of algebraic flux correction schemes, we consider a locally bound-preserving discretization of the time-dependent advection equation. Specifically, we analyze a monolithic convex limiting scheme based on piecewise (multi-)linear continuous finite elements in the semi-discrete formulation. To stabilize the discretization, we use low order time derivatives in the definition of raw antidiffusive fluxes. Our analytical investigation reveals that their limited counterparts should satisfy a certain compatibility condition. The conducted numerical experiments suggest that this prerequisite is satisfied unless the size of mesh elements is vastly different. We prove global-in-time existence of semi-discrete approximations and derive an a priori error estimate for finite time intervals with a worst-case convergence rate of $\frac{1}{2}$ w. r. t. the L^2 error. This rate is optimal in the setting under consideration because we allow all correction factors of the flux-corrected scheme to become zero. In this case, the algorithm reduces to the bound-preserving discrete upwinding method but the limited counterpart of this scheme converges much faster, in practice. Additional numerical experiments are performed to verify the provable convergence rate for a few variants of the scheme.

Keywords Algebraic flux correction · Time-dependent advection equation · Stability and a priori error estimates · Monolithic limiting · Semi-discrete analysis

Communicated by Gunilla Kreiss.

✉ Hennes Hajduk
hennes.hajduk@math.tu-dortmund.de

Andreas Rupp
andreas@rupp.ink

[1] Institute of Applied Mathematics (LS III), TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany

[2] School of Engineering Science, Lappeenranta–Lahti University of Technology (LUT), P.O. Box 20, 53851 Lappeenranta, Finland

Mathematics Subject Classification 65M12 · 65M60

1 Introduction

Algebraic flux correction (AFC) schemes were proposed in [25] and have since then become an active research area [3, 5, 18, 21, 28]. These methods provide a robust framework guaranteeing that discrete maximum principles hold [6, 27] and/or that entropy conditions are satisfied in the case of nonlinear equations [22, 23]. Nonlinear AFC approaches combine a high order baseline scheme, such as the Galerkin finite element discretization, with a provably bound-preserving low order approximation. In this manner, global and/or local constraints can be imposed on the values of AFC solutions resulting from discretizations of various partial differential equations.

The focus of most efforts dealing with AFC schemes was on the development of numerical algorithms, while theoretical aspects have only recently started to attract significant interest. Barrenechea et al. [5] were the first to show solvability of a nonlinear AFC system arising from discretization of stationary convection-diffusion-reaction equations. Moreover, they prove that the scheme convergences with a rate of at least $\frac{1}{2}$ in the AFC energy norm. In their subsequent work [6], they derived a sharper, first order error estimate under the assumption that the limiter is *linearity preserving*. Unfortunately, the proof technique that was used to obtain this superconvergence result relies on the presence of diffusive terms. Lohmann [27] extended the analysis of AFC schemes to the case without diffusive terms and obtained similar theoretical results for linear hyperbolic problems, again with a provable rate of $\frac{1}{2}$. Other theoretically investigated aspects of AFC procedures include their connection to edge-based diffusion [4], proofs of invariant domain preservation for the low order method [13], and a study of a posteriori error estimators [17]. The recent work of Jha and Ahmed [18] presents the first theoretical foundation of AFC schemes for parabolic convection-diffusion-reaction equations. The AFC schemes analyzed therein are based on *flux-corrected transport* algorithms that are fully discrete and employ implicit time stepping.

In contrast to [18], this manuscript presents semi-discrete stability and a priori error analysis of AFC schemes for finite element discretizations of the time-dependent linear advection equation. To cure the oscillatory behavior of continuous Galerkin methods, we stabilize the antidiffusive fluxes using *low order time derivatives* (defined by (2.14) below). Flux correction is performed using Kuzmin's [21] monolithic convex limiting (MCL) scheme. For analytical purposes, we make an assumption regarding compatibility of the semi-discrete approximations and corresponding time derivatives. As shown in [15, Sec. 3.3], it is possible to enforce this condition by adapting the limiting procedure of the standard MCL approach. The results obtained in this manner are slightly more diffusive but exhibit the same second-order convergence rates in practice. However, based on our experience, the additional fix only rarely *needs* to be activated because compatibility seems to be automatically satisfied for the standard MCL scheme in most cases. Evidence for this claim is provided in Sect. 5.4. Therefore, we do not discuss enforcement of the compatibility condition in this work and instead refer the interested reader to [15, Sec. 3.3].

We prove that the nonlinear semi-discrete scheme is stable and that for finite times its spatial accuracy w. r. t. the L^2 norm is at least of order $\frac{1}{2}$ for linear finite elements and generally unstructured meshes. In practice, second order superconvergence can be expected for smooth solutions and uniform meshes, as evidenced by the numerical examples of this paper and, for instance [3, 18, 28].

The structure of this article is as follows. We begin with the formulation of the continuous problem, and review the construction of the monolithic AFC schemes under discussion. Subsequently, in individual sections, we present the main theoretical outcomes of our work, an energy estimate and an a priori error analysis. Finally, we report the results of numerical experiments and draw conclusions. The contents of this paper are to a large degree based on [14, Ch. 5], which improves upon the analysis presented in our preprint [15]. In particular, we improved both the theoretical parts and the numerical examples by properly addressing the treatment of boundary conditions and presenting numerical results not just for simple 1D problems but also in the 2D case.

2 Discretization of the advection equation

In this section, we summarize the MCL strategy for linear transport problems [21]. Our presentation includes a brief discussion of the continuous model problem, a summary of the design principles of the low order method, as well as the formulation of the corresponding monolithic flux-correction schemes. Algebraic limiters of this kind have only recently been applied to different target discretizations such as high-order discontinuous Galerkin methods, e. g., [30]. These algorithms exploit ideas originally proposed in the context of continuous finite elements. It is therefore natural to perform our analysis in this framework as well.

2.1 Continuous model problem

Let $\Omega \subset \mathbb{R}^d, d \in \{1, 2, 3\}$ be a polyhedral domain, $\boldsymbol{v} \in \mathbf{C}(\overline{\Omega} \times \mathbb{R}_+)^d$ a known velocity field, and $\boldsymbol{n} \in \mathbb{R}^d$ the unit outward normal to $\partial\Omega$. We define the time-dependent in- and outflow boundaries of Ω as

$$\Gamma_-(t) := \{\boldsymbol{x} \in \partial\Omega : \boldsymbol{v}(\boldsymbol{x}, t) \cdot \boldsymbol{n}(\boldsymbol{x}) < 0\}, \qquad \Gamma_+(t) := \{\boldsymbol{x} \in \partial\Omega : \boldsymbol{v}(\boldsymbol{x}, t) \cdot \boldsymbol{n}(\boldsymbol{x}) > 0\}.$$

In what follows, we suppress the dependence of $\Gamma_\pm(t)$ on time t. The initial-boundary value problem for the linear advection equation reads

$$\partial_t u + \boldsymbol{v} \cdot \nabla u = 0 \quad \text{in } \Omega \times \mathbb{R}_+, \tag{2.1a}$$

$$u = \hat{u} \quad \text{on } \Gamma_- \times \mathbb{R}_+, \tag{2.1b}$$

$$u = u_0 \quad \text{in } \Omega, \tag{2.1c}$$

where $\hat{u}$ is a given inflow boundary profile and u_0 is an initial datum. For analytical purposes, we assume that the velocity field is *solenoidal*, i. e., $\nabla \cdot \boldsymbol{v} = 0$ in Ω, which allows us to interpret (2.1a) as a hyperbolic conservation law with flux function $\boldsymbol{f}(u, \boldsymbol{x}, t) = \boldsymbol{v}(\boldsymbol{x}, t)u$. Let us remark that the flux correction tools discussed in this section can also be applied to problem (2.1) in the case of more general velocities.

To derive the weak formulation of (2.1), we multiply (2.1a) by a test function w and perform integration by parts. Replacing the consistent flux $\boldsymbol{v} \cdot \boldsymbol{n}\, u$ appearing in the resulting boundary integral with the upwind flux

$$f_{\boldsymbol{n}}(u, \hat{u}) := \begin{cases} \boldsymbol{v} \cdot \boldsymbol{n}\, u & \text{on } \Gamma_+, \\ \boldsymbol{v} \cdot \boldsymbol{n}\, \hat{u} & \text{on } \Gamma_-, \end{cases} \tag{2.2}$$

to incorporate the boundary data $\hat{u}$, we obtain

$$\int_\Omega w \partial_t u \,\mathrm{d}\boldsymbol{x} - \int_\Omega \nabla \cdot (w\, \boldsymbol{v}) u \,\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} w\, f_{\boldsymbol{n}}(u, \hat{u}) \,\mathrm{d}s = 0. \tag{2.3}$$

With regard to the continuous weak formulation of (2.1), we follow Di Pietro and Ern [8, Chs. 2–3]. In particular, we introduce the *graph space* [8, Def. 2.1]

$$\mathrm{V} := \{ w \in \mathrm{L}^2(\Omega) : \boldsymbol{v} \cdot \nabla w \in \mathrm{L}^2(\Omega) \}$$

and define a weak solution to (2.1) as follows.

Definition 1 (Weak solutions to the linear advection equation) A function $u \in \mathrm{C}(\mathbb{R}_+; \mathrm{V}) \cap \mathrm{C}^1(\mathbb{R}_+; \mathrm{L}^2(\Omega))$ is a weak solution to (2.1) if $u(\cdot, 0) = u_0$ almost everywhere in Ω and

$$\int_\Omega w\, \partial_t u \,\mathrm{d}\boldsymbol{x} + a(u, w) = b(w) \qquad \forall w \in \mathrm{V},\ t \in \mathbb{R}_+, \tag{2.4}$$

where

$$a(\cdot, \cdot) : \mathrm{V} \times \mathrm{V} \to \mathbb{R}, \qquad a(u, w) := \int_\Omega w\, \boldsymbol{v} \cdot \nabla u \,\mathrm{d}\boldsymbol{x} - \int_{\Gamma_-} w\, u\, \boldsymbol{v} \cdot \boldsymbol{n} \,\mathrm{d}s, \tag{2.5}$$

$$b(\cdot) : \mathrm{V} \to \mathbb{R}, \qquad b(w) := -\int_{\Gamma_-} w\, \hat{u}\, \boldsymbol{v} \cdot \boldsymbol{n} \,\mathrm{d}s. \tag{2.6}$$

Formulation (2.4)–(2.6) is derived from (2.3) by performing integration by parts and using the definition of the upwind flux (2.2). Thus, only boundary integrals over the inlet Γ_- appear in (2.5) and (2.6).

Remark 1 In this work, we assume that a unique solution u in the sense of Definition 1 and [8] exists. For settings similar to ours, the validity of this assumption can be rigorously proven (see for instance [7]) but for general velocities this is not a trivial task. In principle, one can invoke the method of characteristics and use an energy estimate to show well-posedness. However, rigorous existence and uniqueness results regarding solutions of (2.4) are typically obtained under additional assumptions. For details on these issues, we refer the reader to [8, Sec. 3.1.1] and the references therein.

Remark 2 For the integrals in the weak formulation below to be well-defined, Di Pietro and Ern [8, Sec. 2.1.3] require that in- and outflow boundaries are well-separated, i. e.,

$$\inf_{(\boldsymbol{x},\boldsymbol{y})\in\Gamma_-\times\Gamma_+} \|\boldsymbol{x}-\boldsymbol{y}\| > 0. \tag{2.7}$$

Since our analysis is based on the same variational expression, we admit that the theory dictates this assumption on the model. However, we remark that some classical benchmarks for advection problems, such as LeVeque's solid body rotation [26] (see Sect. 5.3) do not satisfy (2.7).

2.2 Finite element discretization and low order method

The low order method that is employed in this work is the algebraic Lax–Friedrichs scheme [1, 13, 29, 33] adapted to linear advection problems. In the AFC literature, this linear version is called the *discrete upwinding* method because of its equivalence to the node-centered upwind finite volume scheme [25, Sec. 6]. Let us now review the main steps of deriving this low order method. First, we discretize (2.4) in space using continuous linear finite elements.

Let $\mathcal{K}_h = \{K^1, \dots K^E\}$ be a simplicial mesh of $E = E(h) \in \mathbb{N}$ disjoint elements such that $\Omega = \bigcup_{e=1}^{E} K^e$. Furthermore, let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \overline{\Omega}$, $N = N(h) \in \mathbb{N}$, be the vertices of the mesh and $\varphi_1, \dots, \varphi_N$ be the corresponding piecewise linear Lagrange basis polynomials, satisfying $\varphi_i(\boldsymbol{x}_j) = \delta_{ij}$. For simplicity, we assume that the mesh has no hanging nodes. The corresponding finite element space shall be denoted as $\mathrm{V}_h := \{w_h \in \mathrm{C}(\Omega) : w_h|_K \in \mathbb{P}_1(K)\ \forall K \in \mathcal{K}_h\}$ and the semi-discrete numerical solution is expanded as follows

$$u_h(\boldsymbol{x}, t) := \sum_{i=1}^{N} u_i(t)\varphi_i(\boldsymbol{x}), \qquad u_i(t) = u_h(\boldsymbol{x}_i, t).$$

Testing (2.4) with φ_i, $i \in \{1, \dots, N\}$, we obtain the spatial semi-discretization

$$\sum_{j=1}^{N} m_{ij}\frac{\mathrm{d}u_j}{\mathrm{d}t} = -\sum_{j=1}^{N} a_{ij}\, u_j - \int_{\Gamma_-} \varphi_i \left(\hat{u} - \sum_{j=1}^{N} u_j\varphi_j\right) \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s, \tag{2.8}$$

where m_{ij} are scalar-valued entries of the *consistent mass matrix*

$$M = (m_{ij})_{i,j=1}^{N}, \qquad m_{ij} = \int_{\Omega} \varphi_i\, \varphi_j\, \mathrm{d}\boldsymbol{x}, \qquad i, j \in \{1, \dots, N\},$$

and

$$A = (a_{ij})_{i,j=1}^{N}, \qquad a_{ij} = \int_{\Omega} \varphi_i\, \boldsymbol{v}\cdot\nabla\varphi_j\, \mathrm{d}\boldsymbol{x}, \qquad i, j \in \{1, \dots, N\}.$$

Let us now briefly summarize the steps to construct the low order method used in [13, 21], among others. We perform row sum mass lumping, i. e., replace the entries of M in the left hand side of (2.8) with those of

$$M_{\mathrm{L}} = \mathrm{diag}(m_1, \dots, m_N), \qquad m_i := \sum_{j=1}^{N} m_{ij} = \int_{\Omega} \varphi_i \,\mathrm{d}\boldsymbol{x}, \qquad i \in \{1, \dots, N\}.$$

In addition, we use the partition of unity property of basis functions, i. e., the fact that they sum to one everywhere in Ω to rewrite

$$\sum_{j=1}^{N} a_{ij} u_j = \sum_{j \in \mathcal{N}_i \setminus \{i\}} a_{ij}(u_j - u_i).$$

Here $\mathcal{N}_i$ is the nodal stencil defined by

$$\mathcal{N}_i := \{j \in \{1, \dots, N\} : \mathrm{int}(\mathrm{supp}\,\varphi_i) \cap \mathrm{int}(\mathrm{supp}\,\varphi_j) \neq \emptyset\},$$

where int$(\cdot)$ denotes the interior of a set and supp is the support of a function. Moreover, we add diffusive fluxes of the form $d_{ij}(u_j - u_i)$, where

$$d_{ij} = \max\{|a_{ij}|, |a_{ji}|\}, \qquad i \in \{1, \dots, N\},\ j \in \mathcal{N}_i \setminus \{i\}. \tag{2.9}$$

As a final modification to (2.8), we employ a lumped approximation of boundary terms. This step involves a localization of boundary integrals to individual *faces* on the domain boundary.

Definition 2 (Nodal boundary faces, [14]) Let $\mathcal{F}_{\partial\Omega}$ denote the set of $(d-1)$-dimensional boundary faces of $\mathcal{K}_h$. Then the set $\mathcal{F}_i$ contains all boundary faces that meet at node $\boldsymbol{x}_i \in \partial\Omega$, $i \in \{1, \dots, N\}$. For $\Gamma_k \in \mathcal{F}_i$, we define $\hat{u}_i^k := \hat{u}(\boldsymbol{x}_i)$ as the value of $\hat{u}$ corresponding to $\Gamma_k \subseteq \Gamma_-$.

The above modifications made to (2.8) yield the low order method

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} (d_{ij} - a_{ij})(u_j - u_i) + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i), \qquad i \in \{1, \dots, N\}, \tag{2.10}$$

where

$$b_i^k := -\int_{\Gamma_k} \varphi_i \, \min\{0, \boldsymbol{v} \cdot \boldsymbol{n}\} \,\mathrm{d}s.$$

Note that $b_i^k \geq 0$. We may also write (2.10) in the *bar state form* [13, 16, 21]

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} 2d_{ij}(\bar{u}_{ij} - u_i) + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i), \qquad i \in \{1, \dots, N\}, \tag{2.11}$$

where the *bar states* $\bar{u}_{ij}$ are defined by

$$\bar{u}_{ij} = \begin{cases} \dfrac{u_i + u_j}{2} - \dfrac{a_{ij}(u_j - u_i)}{2d_{ij}} & \text{if } a_{ij} \neq 0, \\ \dfrac{u_i + u_j}{2} & \text{if } a_{ij} = 0, \end{cases} \qquad i \in \{1, \dots, N\},\ j \in \mathcal{N}_i \setminus \{i\}. \tag{2.12}$$

Remark 3 Definition (2.9) ensures that $\bar{u}_{ij}$ is a convex combination of u_i and u_j because

$$\min\{u_i, u_j\} \leq \bar{u}_{ij} \leq \max\{u_i, u_j\} \qquad \Leftrightarrow \qquad |a_{ij}| \leq d_{ij}.$$

Instead of (2.9), the classical version of the discrete upwinding method uses [25]

$$d_{ij} = \max\{a_{ij}, 0, a_{ji}\}, \qquad i \in \{1, \dots, N\},\ j \in \mathcal{N}_i \setminus \{i\}. \tag{2.13}$$

If $\nabla \cdot \boldsymbol{v} = 0$, this definition is equivalent to (2.9), unless both nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ lie on $\partial\Omega$. This fact follows from integration by parts and omission of the resulting boundary integral. The validity of discrete maximum principles for nodal values can be shown for (2.13) using alternative proof techniques [27, Sec. 4.3.2]. However, individual bar states $\bar{u}_{ij}$ of the discrete upwinding method based on (2.13) may violate the local maximum principle $\min\{u_i, u_j\} \leq \bar{u}_{ij} \leq \max\{u_i, u_j\}$.

2.3 Monolithic convex limiting

The low order method (2.10) produces very diffusive approximations. To recover the accuracy of the standard finite element discretization (2.8), we perform algebraic flux correction. First, we define raw antidiffusive fluxes $f_{ij} = m_{ij}(\dot{u}_i - \dot{u}_j) + d_{ij}(u_i - u_j)$ for $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ and their limited counterparts f_{ij}^*, which are specified below. Here $\dot{u}_h = \sum_{i=1}^N \dot{u}_i \varphi_i$ is a suitable approximation to the time derivative $\frac{\mathrm{d}u_h}{\mathrm{d}t}$. Following [21], we employ the low order nodal values

$$\dot{u}_i = \frac{1}{m_i} \sum_{j \in \mathcal{N}_i \setminus \{i\}} (d_{ij} - a_{ij})(u_j - u_i) + \frac{1}{m_i} \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i), \qquad i \in \{1, \dots, N\}, \tag{2.14}$$

to compute $\dot{u}_h$ in practice. This approach can be interpreted as a modification of the *target scheme* corresponding to the standard continuous Galerkin discretization that otherwise exhibits a suboptimal first order convergence rate [31, Sec. 14.3.1]. As illustrated in Sect. 5.2, the use of low order time derivatives $\dot{u}_i$ (instead of their consistent Galerkin counterparts defined by (5.2) below) also has a stabilizing effect on the overall approximation. This approach was proposed in the original publication on MCL schemes [21] and corresponds to a cheap and effective target scheme. Advanced stabilization techniques for higher-order methods applied to linear and nonlinear hyperbolic problems can be found in [28] and [23], respectively. By the above definition of raw antidiffusive fluxes, we have $f_{ij} = -f_{ji}$. To preserve the conservation property of the limited scheme, we enforce the corresponding constraint $f_{ij}^* = -f_{ji}^*$ for the limited antidiffusive fluxes as is common in the AFC methodology [25]. Using the equivalence of formulations (2.10) and (2.11), we obtain a similar bar state form for the semi-discrete flux correction scheme [21]

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} [(d_{ij} - a_{ij})(u_j - u_i) + f_{ij}^*] + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i) \tag{2.15a}$$

$$= \sum_{j \in \mathcal{N}_i \setminus \{i\}} 2d_{ij}(\bar{u}_{ij}^* - u_i) + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i), \qquad i \in \{1, \ldots, N\}, \tag{2.15b}$$

where the limited bar states are defined by [21]

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

Thus, a forward Euler update for (2.15) reads

$$\tilde{u}_i = \left[1 - \frac{\Delta t}{m_i}\left(\sum_{j \in \mathcal{N}_i \setminus \{i\}} 2d_{ij} + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k\right)\right] u_i + \frac{\Delta t}{m_i}\left(\sum_{j \in \mathcal{N}_i \setminus \{i\}} 2d_{ij}\bar{u}_{ij}^* + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k \hat{u}_i^k\right),$$

where Δt is the time step. Hence, the updated solution $\tilde{u}_i$ is a convex combination of u_i, the $\bar{u}_{ij}^*$, and the $\hat{u}_i^k$, provided that the Courant–Friedrichs–Lewy (CFL) condition

$$\Delta t \leq \min_{i \in \{1, \ldots, N\}} \frac{m_i}{\sum_{j \in \mathcal{N}_i \setminus \{i\}} 2d_{ij} + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k} \tag{2.16}$$

is satisfied. In other words, if (2.16) holds, the forward-Euler updated solution $\tilde{u}_i$ preserves all local bounds that these states are constrained by. This argument made here for a forward Euler step directly carries over to p-stage, pth-order accurate strong-stability-preserving Runge–Kutta (SSPp-RK) methods [12, 34], where $p \in \{1, 2, 3\}$.

In the process of flux correction, we enforce the local maximum principles

$$u_i^{\min} \leq \bar{u}_{ij}^* \leq u_i^{\max}, \qquad u_i^{\min} := \min_{j \in \mathcal{N}_i} u_j, \qquad u_i^{\max} := \max_{j \in \mathcal{N}_i} u_j \tag{2.17}$$

in addition to skew symmetry of antidiffusive fluxes. Rearranging these constraints, we obtain Kuzmin's formula for the limited antidiffusive fluxes of his monolithic convex limiter [21]

$$f_{ij}^* = \begin{cases} \min\{f_{ij}, 2d_{ij}u_i^{\max} - \bar{w}_{ij}, \bar{w}_{ji} - 2d_{ij}u_j^{\min}\} & \text{if } f_{ij} \geq 0, \\ \max\{f_{ij}, 2d_{ij}u_i^{\min} - \bar{w}_{ij}, \bar{w}_{ji} - 2d_{ij}u_j^{\max}\} & \text{if } f_{ij} \leq 0, \end{cases} \tag{2.18}$$

where $\bar{w}_{ij} := 2d_{ij}\bar{u}_{ij}$.

Lemma 1 (Conservation property of the MCL scheme, [21, 25, 33]) *The semi-discrete scheme (2.15) in which $f_{ij}^* = -f_{ji}^*$ is conservative in the following sense*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u_h \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \boldsymbol{v}\cdot\nabla u_h \,\mathrm{d}\boldsymbol{x} - \sum_{i=1}^N \sum_{\Gamma_k\in\mathcal{F}_i} \int_{\Gamma_k} \varphi_i\,(\hat{u}_i^k - u_i)\min\{0, \boldsymbol{v}\cdot\boldsymbol{n}\}\,\mathrm{d}s. \tag{2.19}$$

Proof Summing over all degrees of freedom, we exploit the symmetry of diffusion coefficients d_{ij}, skew symmetry of antidiffusive fluxes f_{ij}^*, and the zero row sum property of matrix A. □

Remark 4 Continuous weak solutions u defined by (2.4) satisfy the conservation relation

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \boldsymbol{v}\cdot\nabla u \,\mathrm{d}\boldsymbol{x} - \int_{\Gamma_-} (\hat{u} - u)\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s. \tag{2.20}$$

Thus, (2.19) is a semi-discrete counterpart of (2.20) that accounts for the flux-lumped quadrature rule used in the AFC setting.

Let us now rewrite the bar state form (2.15b) of the semi-discrete MCL scheme in a formulation that is more amenable to theoretical investigations. Despite the fact that using MCL, the fluxes f_{ij}^* can be calculated directly via (2.18), we introduce correction factors $\alpha_{ij}(u_h) = \alpha_{ji}(u_h) \in [0, 1]$ defined by $\alpha_{ij}(u_h) = f_{ij}^*/f_{ij}$ if $f_{ij} \neq 0$ and $\alpha_{ij}(u_h) = 1$ otherwise. The dependence of correction factors on the discrete solution makes AFC schemes nonlinear. Using the above definition of f_{ij}, the semi-discrete MCL scheme (2.15) reads

$$\begin{aligned} m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = &\sum_{j\in\mathcal{N}_i\setminus\{i\}} [(1-\alpha_{ij}(u_h))\,d_{ij}(u_j - u_i) - a_{ij}(u_j - u_i) + \alpha_{ij}(u_h)\,m_{ij}(\dot{u}_i - \dot{u}_j)] \\ &+ \sum_{\Gamma_k\in\mathcal{F}_i} b_i^k(\hat{u}_i^k - u_i), \qquad i \in \{1,\dots,N\}, \end{aligned} \tag{2.21}$$

and can equivalently be written as

$$\sum_{i=1}^N w_i m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} + a_h(u_h, w_h) + d_h(u_h; u_h, w_h) - m_h(u_h; \dot{u}_h, w_h) = b_h(w_h) \tag{2.22}$$

for all $w_h \in \mathrm{V}_h$ given by $w_h = \sum_{j=1}^{N} w_j \varphi_j$. The bilinear and linear forms

$$a_h(u_h, w_h) := \int_\Omega w_h\, \boldsymbol{v} \cdot \nabla u_h \,\mathrm{d}\boldsymbol{x} - \sum_{i=1}^{N} w_i\, u_i \int_{\Gamma_-} \varphi_i\, \boldsymbol{v} \cdot \boldsymbol{n} \,\mathrm{d}s,$$

$$b_h(w_h) := -\sum_{i=1}^{N} w_i \sum_{\Gamma_k \in \mathcal{F}_i} \hat{u}_i^k \int_{\Gamma_k} \varphi_i \, \min\{0, \boldsymbol{v} \cdot \boldsymbol{n}\} \,\mathrm{d}s$$

are associated with the (stabilized) Galerkin finite element discretization corresponding to $\alpha_{ij} = 1$ for all $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. The nonlinear forms [5, 18, 27]

$$d_h(u_h; v_h, w_h) = \sum_{i=1}^{N} w_i \sum_{j \in \mathcal{N}_i \setminus \{i\}} (1 - \alpha_{ij}(u_h))\, d_{ij}(v_i - v_j), \tag{2.23}$$

$$m_h(u_h; v_h, w_h) = \sum_{i=1}^{N} w_i \sum_{j \in \mathcal{N}_i \setminus \{i\}} \alpha_{ij}(u_h)\, m_{ij}(v_i - v_j) \tag{2.24}$$

in (2.22) are due to algebraic flux correction.

Lemma 2 (Scalar product properties of nonlinear forms, [5]) *For arbitrary $u_h, v_h, w_h \in \mathrm{V}_h$, the nonlinear forms (2.23) and (2.24) satisfy*

$$d_h(u_h; v_h, v_h) \geq 0, \qquad d_h(u_h; v_h, w_h)^2 \leq d_h(u_h; v_h, v_h)\, d_h(u_h; w_h, w_h),$$
$$m_h(u_h; v_h, v_h) \geq 0, \qquad m_h(u_h; v_h, w_h)^2 \leq m_h(u_h; v_h, v_h)\, m_h(u_h; w_h, w_h).$$

Proof Proofs of these statements for $d_h(\cdot; \cdot, \cdot)$ can be found in [27, p. 113], see also [5, Lem. 3.1 and Sec. 6]. The same arguments apply to $m_h(\cdot; \cdot, \cdot)$. □

3 Energy estimate

Let us now derive an energy estimate for approximations obtained via (2.22). In the proof of this stability result, we rely on the assumption that the following requirement is satisfied.

Definition 3 (Compatibility condition, [15, Ineq. (3.16)]) Let $\dot{u}_h, u_h \in \mathrm{V}_h$ be given functions and $\lambda := \|\boldsymbol{v}\|_{\mathbf{L}^\infty(\Omega \times \mathbb{R}_+)^d}$ be the maximum velocity. Define the nonlinear forms $d_h(\cdot; \cdot, \cdot)$ and $m_h(\cdot; \cdot, \cdot)$ as in (2.23) and (2.24), respectively. Suppose that there exists a constant $\gamma \in (0, 1)$ such that

$$\frac{\gamma h}{\lambda} m_h(u_h; \dot{u}_h, \dot{u}_h) \leq (1 - \gamma) d_h(u_h; u_h, u_h) - m_h(u_h; \dot{u}_h, u_h). \tag{3.1}$$

Then we say that $\dot{u}_h \in \mathrm{V}_h$ is compatible with $u_h \in \mathrm{V}_h$.

The ratio h/λ has physical units $[h]/[\lambda] = \mathrm{m}/(\mathrm{ms}^{-1}) = \mathrm{s}$. It is used in inequality 3.1 to ensure that all terms have the same units for $[\dot{u}_h] = \mathrm{s}^{-1}[u_h]$. Note that if we set $w_h = u_h$ in (2.22), the two nonlinear forms contained therein coincide with the right hand side of (3.1) plus the nonnegative remainder $\gamma\, d_h(u_h; u_h, u_h)$. Due to Lemma 2, we can bound these terms below by a positive number if $(u_h, \dot{u}_h)$ is a compatible pair. This argument is our main motivation for relying on (3.1) for theoretical purposes. Clearly, the pair $(u_h, 0)$ satisfies (3.1). Thus if we do not compensate the mass lumping error in the process of limiting, the scheme automatically satisfies (3.1). As illustrated in Sect. 5.4, the standard MCL scheme using low order time derivatives (2.14) for stabilization purposes is also prone to producing compatible pairs. In [15, Sec. 3.3] we present a modified MCL procedure with which (3.1) can be guaranteed. Due to the complicated nature of this approach we chose not to discuss it any further in this work.

Before presenting our energy estimate, we need to prove the following technical result.

Lemma 3 *Any function $v_h \in \mathrm{V}_h$ defined by $v_h = \sum_{i=1}^N v_i\varphi_i$ satisfies the identity*

$$v_h^2 - \sum_{i=1}^N v_i^2\varphi_i = - \sum_{i,j=1 i<j}^N (v_i - v_j)^2 \varphi_i\, \varphi_j.$$

Proof Invoking the partition of unity property of basis functions, we obtain

$$\begin{aligned} v_h^2 - \sum_{i=1}^N v_i^2\, \varphi_i &= \sum_{i=1}^N v_i^2 \varphi_i\, (\varphi_i - 1) + \sum_{i,j=1 i\neq j}^N v_i\, v_j\, \varphi_i\, \varphi_j = - \sum_{i,j=1 i\neq j}^N v_i^2 \varphi_i\, \varphi_j \\ &\quad + \sum_{i,j=1 i\neq j}^N v_i\, v_j \varphi_i\, \varphi_j \\ &= \sum_{i,j=1 i<j}^N v_i\, (v_j - v_i)\, \varphi_i\, \varphi_j + \sum_{i,j=1 j<i}^N v_i\, (v_j - v_i)\, \varphi_i\, \varphi_j \\ &= \sum_{i,j=1 i<j}^N (v_i - v_j)(v_j - v_i)\, \varphi_i\, \varphi_j. \end{aligned}$$

□

Proposition 1 (Semi-discrete energy estimate) *Assume that there is a finite time $T > 0$ such that $\boldsymbol{v}(\cdot, t) \in \mathbf{W}^{1,\infty}(\Omega)$ and $\nabla \cdot \boldsymbol{v}(\cdot, t) = 0$ in Ω for all $t \in (0, T)$. Let $u_h(t)$ and $\dot{u}_h(t)$ satisfy* (2.22) *and, additionally, the compatibility condition* (3.1) *with a constant $\gamma \in (0, 1)$ for all $t \in (0, T)$. Then the following estimate holds for the solution $u_h(T)$ of the semi-discrete problem* (2.22)

$$\sum_{i=1}^N m_i\, u_i(T)^2 + \int_0^T \int_{\Gamma_+} u_h^2\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s\,\mathrm{d}t - \int_0^T \sum_{i,j=1 i<j}^N (u_i - u_j)^2 \int_{\Gamma_-} \varphi_i\,\varphi_j\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s$$
$$- \frac{1}{2}\int_0^T \sum_{i=1}^N u_i^2 \int_{\Gamma_-} \varphi_i\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s + 2\gamma\int_0^T \big[\frac{h}{\lambda} m_h(u_h;\dot{u}_h,\dot{u}_h) + d_h(u_h;u_h,u_h)\big]\,\mathrm{d}t$$
$$\le \sum_{i=1}^N m_i\,u_i(0)^2 + 2\int_0^T \sum_{i=1}^N \sum_{\Gamma_k\in\mathcal{F}_i} \int_{\Gamma_k} \varphi_i\,(\hat{u}_i^k)^2\,\max\{0,-\boldsymbol{v}\cdot\boldsymbol{n}\}\,\mathrm{d}s\,\mathrm{d}t. \tag{3.2}$$

Proof Testing (2.22) with $w_h = u_h$, we use the compatibility condition (3.1), the identity $u_h\,\boldsymbol{v}\cdot\nabla u_h = \frac{1}{2}\nabla\cdot(\boldsymbol{v}\,u_h^2)$, the divergence theorem and Young's inequality to show that

$$\frac{1}{2}\sum_{i=1}^N m_i \frac{\mathrm{d}(u_i)^2}{\mathrm{d}t} + \frac{1}{2}\int_{\partial\Omega} u_h^2\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s - \sum_{i=1}^N u_i^2 \int_{\Gamma_-}\varphi_i\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s$$
$$+ \frac{\gamma h}{\lambda} m_h(u_h;\dot{u}_h,\dot{u}_h) + \gamma d_h(u_h;u_h,u_h)$$
$$\le \sum_{i=1}^N u_i\,m_i\frac{\mathrm{d}u_i}{\mathrm{d}t} + \int_\Omega u_h\,\boldsymbol{v}\cdot\nabla u_h\,\mathrm{d}\boldsymbol{x} - \sum_{i=1}^N u_i^2\int_{\Gamma_-}\varphi_i\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s$$
$$+ d_h(u_h;u_h,u_h) - m_h(u_h;\dot{u}_h,u_h)$$
$$= b_h(u_h) = -\sum_{i=1}^N \sum_{\Gamma_k\in\mathcal{F}_i}\int_{\Gamma_k}\varphi_i\,u_i\,\hat{u}_i^k\,\min\{0,\boldsymbol{v}\cdot\boldsymbol{n}\}\,\mathrm{d}s$$
$$\le -\sum_{i=1}^N \frac{u_i^2}{4}\int_{\Gamma_-}\varphi_i\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s - \sum_{i=1}^N\sum_{\Gamma_k\in\mathcal{F}_i}\int_{\Gamma_k}\varphi_i\,(\hat{u}_i^k)^2\,\min\{0,\boldsymbol{v}\cdot\boldsymbol{n}\}\,\mathrm{d}s.$$

Multiplying by factor 2 and combining the integrals over Γ_-, we write this inequality as

$$\sum_{i=1}^N m_i\frac{\mathrm{d}(u_i)^2}{\mathrm{d}t} + \int_{\Gamma_+} u_h^2\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s + \int_{\Gamma_-}\boldsymbol{v}\cdot\boldsymbol{n}\left(u_h^2 - \sum_{i=1}^N u_i^2\,\varphi_i\right)\mathrm{d}s$$
$$- \frac{1}{2}\sum_{i=1}^N u_i^2\int_{\Gamma_-}\varphi_i\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s + \frac{2\gamma h}{\lambda}m_h(u_h;\dot{u}_h,\dot{u}_h) + 2\gamma d_h(u_h;u_h,u_h)$$
$$\le 2\sum_{i=1}^N\sum_{\Gamma_k\in\mathcal{F}_i}\int_{\Gamma_k}\varphi_i\,(\hat{u}_i^k)^2\,\max\{0,-\boldsymbol{v}\cdot\boldsymbol{n}\}\,\mathrm{d}s.$$

Employing Lemma 3 and integrating in time produces (3.2). □

Note that as a consequence of Lemma 2 and of the nonnegativity of basis functions, all terms appearing on the left hand side of inequality (3.2) are nonnegative. To guarantee that the assumptions of Proposition 1 are satisfied in practice, one can use the scheme proposed in [15, Sec. 3.3], which enforces (3.1) for user defined values of γ. In our experience, failure to apply this limiter has no negative practical effects, however.

Remark 5 The reader may wonder what significance is attached to Proposition 1. Since the *fully discrete* MCL scheme produces locally bound-preserving approximations, it is stable by design. Preservation of global bounds in the *semi-discrete* setting can be shown as in [24] under the assumption that a solution exists. The semi-discrete MCL scheme represents a nonlinear system of ordinary differential equations. Well-posedness of such initial value problems can be shown by invoking the Picard–Lindelöf theorem, which guarantees the existence of solutions on finite time intervals. Once local existence is established, we exploit a global existence and uniqueness result for ordinary differential equations [2, Thm. 7.6]. According to this theorem, solutions that cannot be extended to arbitrary times must in fact blow up, which, in our case, is prevented by Proposition 1. It follows that the semi-discrete MCL scheme (2.22) possesses a unique solution that exists for all times $t \geq 0$.

4 Error analysis

Compared to the energy estimate derived in the previous section, our error analysis is rather involved. In particular, we need to make additional assumptions on the data of the continuous problem (2.4) as well as on the mesh sequences. These aspects are discussed in Sect. 4.1. Subsequently, in Sect. 4.2, we recall some auxiliary results from the literature on numerical analysis of finite element methods including AFC schemes. Finally, in Sect. 4.3, we state, prove, and discuss the main result of this work, which is a semi-discrete a priori error estimate for MCL approximations.

Throughout this section, the letter C (possibly with a subscript) denotes a generic positive constant that is independent of the mesh size h. Moreover, we assume that $h \leq 1$ and therefore $h^p \leq h^q$ for $p \geq q$.

4.1 Preliminaries

Recall that we only consider meshes that are affine and geometrically conforming triangulations of $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. Additionally, we restricted ourselves to simplicial meshes, which allows us to exploit the linearity of finite element approximations inside mesh cells.

The a priori error estimate that we present in Sect. 4.3 is valid only for *quasi-uniform* families of meshes, i. e., there has to exist $C > 0$ such that [8, Sec. 3.1.2]

$$h := \max_{K \in \mathcal{K}_h} h_K \leq C \min_{K \in \mathcal{K}_h} h_K,$$

where $h_K = \operatorname{diam}(K)$. As is standard in finite element analysis, we also assume *shape-regularity* of $(\mathcal{K}_h)_{h>0}$. For this requirement to be satisfied, there has to exist

$C > 0$ such that $Ch_K \leq r_K$, where r_K is the radius of the largest open ball that fits into K [8, Sec. 1.4.1]. Additionally, we assume that the mesh faces, which are simplices in $\mathbb{R}^{d-1}$, are also shape regular in this sense. Our final assumption regarding the mesh sequence is that there exists $C > 0$ such that $h \leq C\tilde{h}$, where $\tilde{h} = \min_{\Gamma\in\mathcal{F}_{\partial\Omega}} \operatorname{diam}(\Gamma)$ and $\mathcal{F}_{\partial\Omega}$ is the set of boundary faces (cf. Definition 2). We do not need to assume *contact regularity* of the mesh sequence [8, Def. 1.38] as Di Pietro and Ern do because this requirement is automatically satisfied for simplicial triangulations.

Following [5, 27], we assume $\mathrm{H}^2(\Omega)$ regularity of the exact solution $u(\cdot, t)$ for all $t \geq 0$. We also require the time derivative $\partial_t u$ to have this regularity. Specifically, we restrict our investigations to exact solutions of (2.4) that satisfy

$$u \in \mathrm{W}^{1,\infty}(\mathbb{R}_+; \mathrm{H}^2(\Omega)), \qquad u|_{\Gamma_-} \in \mathrm{L}^\infty(\mathbb{R}_+; \mathrm{H}^2(\Gamma_-)).$$

For simplicity, we set $u_h(\cdot, 0)$ equal to the continuous interpolant $\mathrm{I}_h u_0 \in \mathrm{V}_h$ of $u_0 \in \mathrm{C}(\overline{\Omega})$. The interpolation operator $\mathrm{I}_h : \mathrm{C}(\overline{\Omega}) \to \mathrm{V}_h$ is defined by

$$w \mapsto w_h := \sum_{i=1}^{N} w(\boldsymbol{x}_i)\,\varphi_i .$$

Also for simplicity, we assume that the boundary data $\hat{u}$ is linear on every boundary face $\Gamma \in \mathcal{F}_-$, where $\mathcal{F}_- = \mathcal{F}_-(t) := \{\Gamma \in \mathcal{F}_{\partial\Omega} : \Gamma \cap \Gamma_- \neq \emptyset\}$. This assumption corresponds to a particular choice of the quadrature rule for boundary integrals.

4.2 Auxiliary statements

To prepare the ground for the derivation of our error estimate, we first summarize a few important ingredients of its proof, beginning with some standard inequalities. Then we discuss aspects that are peculiar to algebraic flux correction schemes. Most of the AFC results were originally proven by Barrenechea et al. [5].

Lemma 4 (Interpolation error estimate for volume integrals) *Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. Then there exists $C > 0$ such that*

$$\|w - \mathrm{I}_h w\|_{\mathrm{L}^2(\Omega)} + h|w - \mathrm{I}_h w|_{\mathrm{H}^1(\Omega)} \leq Ch^2 |w|_{\mathrm{H}^2(\Omega)} \qquad \forall w \in \mathrm{H}^2(\Omega).$$

Proof See [10, Sec. 1.5.1, in particular Ex. 1.111]. □

Lemma 5 (Interpolation error estimate for face integrals) *Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$ and let $\Gamma \subset \partial K$ be a face of $K \in \mathcal{K}_h$. Then there exists $C > 0$ such that*

$$\|w - \mathrm{I}_h w\|_{\mathrm{L}^2(\Gamma)} \leq Ch_K^{3/2} |w|_{\mathrm{H}^2(K)} \qquad \forall w \in \mathrm{H}^2(K).$$

Proof The claim follows from the *continuous trace inequality* [8, Lem. 1.49] in combination with Lemma 4. □

Lemma 6 (Discrete trace inequality, [8] Lem. 1.46) *Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$ and let $\Gamma \subset \partial K$ be a face of $K \in \mathcal{K}_h$. Then there exists $C > 0$ such that*

$$\|v_h\|_{\mathrm{L}^2(\Gamma)} \leq C h_K^{-1/2} \|v_h\|_{\mathrm{L}^2(K)} \qquad \forall v_h \in \mathbb{P}_1(K).$$

Lemma 7 (Inverse inequality, [8] Lem. 1.44) *Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$ and let $K \in \mathcal{K}_h$. Then there exists $C > 0$ such that*

$$|v_h|_{\mathrm{H}^1(K)} \leq C h_K^{-1} \|v_h\|_{\mathrm{L}^2(K)} \qquad \forall v_h \in \mathbb{P}_1(K).$$

Lemma 8 ([5]) *Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$. Define $\Gamma_{ij} := \{\mu \boldsymbol{x}_i + (1-\mu)\boldsymbol{x}_j : \mu \in [0,1]\}$ for a pair of mesh vertices $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ $i \in \{1,\dots,N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. Let $K \in \mathcal{K}_h$ with $\Gamma_{ij} \subset \partial K$. Then there exists $C > 0$ such that*

$$|v_h(\boldsymbol{x}_i) - v_h(\boldsymbol{x}_j)| \leq C h_K^{1-d/2} |v_h|_{\mathrm{H}^1(K)} \qquad \forall v_h \in \mathbb{P}_1(K).$$

Proof The claim follows from a Taylor expansion, linearity, and shape regularity, see [5, Pf. of Lem. 7.3] or [27, Ineq. (4.90)] for details. □

Lemma 9 ([5, 18]) *Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$. Then there exist constants $C_1 = C_1(d) > 0$ and $C_2 = C_2(d, \boldsymbol{v}) > 0$ such that*

$$m_{ij} \leq C_1 h^d, \qquad d_{ij} \leq C_2 h^{d-1}, \qquad i \in \{1,\dots,N\},\ j \in \mathcal{N}_i \setminus \{i\}.$$

Proof Clearly, $\operatorname{supp}(\varphi_i \varphi_j) \subseteq \Omega_{ij} := \{\boldsymbol{x} \in \overline{\Omega} : \exists \mu \in [0,1] : |\boldsymbol{x} - (\mu \boldsymbol{x}_i + (1-\mu)\boldsymbol{x}_j)| \leq h\}$, and due to shape regularity, there exists $C = C(d) > 0$ such that $|\Omega_{ij}| \leq Ch^d$. Therefore

$$m_{ij} = \int_{\Omega_{ij}} \varphi_i \, \varphi_j \,\mathrm{d}\boldsymbol{x} \leq \|\varphi_i\|_{\mathrm{L}^2(\Omega_{ij})} \|\varphi_j\|_{\mathrm{L}^2(\Omega_{ij})} \leq \|1\|^2_{\mathrm{L}^2(\Omega_{ij})} = |\Omega_{ij}| \leq Ch^d.$$

The estimate for d_{ij} is obtained similarly by invoking (2.9), factoring out the maximum velocity λ and using the inverse inequality, i.e., Lemma 7, see [5, Pf. of Lem. 7.3] or [27, Pf. of Thm. 4.72] for details. □

Lemma 10 ([5]) *Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$. Then there exist constants $C_1 = C_1(d) > 0$ and $C_2 = C_2(d, \boldsymbol{v}) > 0$ such that*

$$m_h(v_h; \mathrm{I}_h w, \mathrm{I}_h w) \leq C_1 h^2 \|w\|^2_{\mathrm{H}^2(\Omega)}, \qquad d_h(v_h; \mathrm{I}_h w, \mathrm{I}_h w) \leq C_2 h \|w\|^2_{\mathrm{H}^2(\Omega)}$$

for all $v_h \in \mathrm{V}_h$, $w \in \mathrm{H}^2(\Omega)$.

Proof The estimate for $d_h(\cdot;\cdot,\cdot)$ is derived in [27, Ineq. (4.122)] by invoking Lemma 4, 8, and 9, see also [5, Lem. 3.1]. The estimate for $m_h(\cdot;\cdot,\cdot)$ is obtained similarly. □

4.3 A priori error estimate

To state our main result, we need to define some auxiliary quantities. For $t \geq 0$, let $\vartheta_h(t) = \sum_{i=1}^N \vartheta_i(t)\varphi_i \in \mathrm{V}_h$ be the *discrete error* $\vartheta_h(t) := \mathrm{I}_h u(t) - u_h(t)$ and define

$$
\begin{aligned}
q(T) :=& \sum_{i,j=1 i<j}^{N} m_{ij}(\vartheta_i(T) - \vartheta_j(T))^2 + \int_0^T \Bigg[\int_{\Gamma_+} \vartheta_h^2\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s - \sum_{i=1}^N \vartheta_i^2 \int_{\Gamma_-} \varphi_i\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s \\
&- \sum_{i,j=1 i<j}^{N} (\vartheta_i - \vartheta_j)^2 \int_{\Gamma_-} \varphi_i\,\varphi_j\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s + \gamma d_h(u_h; u_h, u_h) + \frac{\gamma h}{\lambda} m_h(u_h; \dot{u}_h, \dot{u}_h) \Bigg] \mathrm{d}t, \\
z(T) :=& \int_0^T \Big[\|\partial_t u\|^2_{\mathrm{H}^2(\Omega)} + \|u\|^2_{\mathrm{H}^2(\Omega)} + \|u\|^2_{\mathrm{H}^2(\Gamma_-)} + |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} \Big] \mathrm{d}t.
\end{aligned}
$$

Proposition 2 (Semi-discrete a priori error estimate) *Let the assumptions made in Sect. 4.1 be satisfied. Assume that there is a finite time $T > 0$ such that $\boldsymbol{v}(\cdot,t) \in \mathbf{W}^{1,\infty}(\Omega)$ and $\nabla\cdot\boldsymbol{v}(\cdot,t) = 0$ in Ω for all $t \in (0,T)$. Let $u_h(t)$ and $\dot{u}_h(t)$ satisfy (2.22) and, additionally, the compatibility condition (3.1) with a constant $\gamma \in (0,1)$ independent of h for all $t \in (0,T)$. Then there exist positive constants $C_1 = C_1(d,\boldsymbol{v})$, $C_2 = C_2(d,\boldsymbol{v},\gamma)$, and $C_3 = C_3(d)$ such that the estimate*

$$
\|u(T) - u_h(T)\|_{\mathrm{L}^2(\Omega)} \leq C_3 h^2 |u(T)|_{\mathrm{H}^2(\Omega)} + \sqrt{y(T) + C_1 \int_0^T \mathrm{e}^{C_1(T-t)}\, y(t)\,\mathrm{d}t} \tag{4.1}
$$

holds for the exact solution $u(T)$ of the continuous problem (2.4), the exact solution $u_h(T)$ of the semi-discrete problem (2.22), and $y(T) := C_2 h\, z(T) - q(T)$.

Remark 6 We do not see any practical problems if (3.1) is invalid, only the theoretical results would no longer apply.

Corollary 1 (Convergence order of the semi-discrete MCL scheme) *Under the assumptions of Proposition 2, the a priori error estimate*

$$
\|u(T) - u_h(T)\|_{\mathrm{L}^2(\Omega)} \leq C_3 h^2 |u(T)|_{\mathrm{H}^2(\Omega)} + \sqrt{\mathrm{e}^{C_1 T}\, C_2 h\, \|z\|_{\mathrm{L}^\infty(0,T)}} \leq C_4\, h^{\frac{1}{2}} \tag{4.2}
$$

holds with a constant $C_4 = C_4(C_1, C_2, C_3, T, u, \hat{u}) > 0$, which behaves as $e^{C_1 T/2}$.

Proof (of Corollary 1) Since q and z are nonnegative functions, we may use the estimate $y(T) \leq C_2 h\, z(T)$ in (4.1). The claim follows by calculating the integral of the exponential function. □

Proof (of Proposition 2) This proof combines recent results on AFC schemes [5, 27] with a new way of proving a priori error estimates for nonconforming discretizations of the advection equation [32]. A particular similarity of the approach developed in [32] to our theory is that both apply to semi-discrete formulations.

We introduce the *interpolation error* $\Theta(t) = \Theta(u, h; t) := u(t) - \mathrm{I}_h u(t)$ and subtract (2.22) from (2.4). Setting $w = w_h = \vartheta_h$, we obtain the error equation

$$\overbrace{\int_\Omega \vartheta_h \frac{\partial u}{\partial t}\,\mathrm{d}\boldsymbol{x} - \sum_{i=1}^N \vartheta_i\, m_i \frac{\mathrm{d}u_i}{\mathrm{d}t}}^{\Xi_1} + \overbrace{a(u, \vartheta_h) - a_h(u_h, \vartheta_h)}^{\Xi_2}$$

$$= \underbrace{b(\vartheta_h) - b_h(\vartheta_h)}_{\Xi_3} + \underbrace{d_h(u_h; u_h, \vartheta_h) - m_h(u_h; \dot{u}_h, \vartheta_h)}_{\Xi_4}.$$

Recall that the identity $m_i = \sum_{j=1}^N m_{ij}$ holds for row sum mass lumping. Using this decomposition of m_i and the identities $u = \Theta + \vartheta_h + \mathrm{I}_h u - \vartheta_h,\ u_h = \mathrm{I}_h u - \vartheta_h$, we find that

$$\begin{aligned}
\Xi_1 &= \int_\Omega \vartheta_h \frac{\partial \Theta}{\partial t}\,\mathrm{d}\boldsymbol{x} + \int_\Omega \vartheta_h \frac{\mathrm{d}\vartheta_h}{\mathrm{d}t}\,\mathrm{d}\boldsymbol{x} + \sum_{i=1}^N \vartheta_i \frac{\mathrm{d}}{\mathrm{d}t}\Big(\sum_{j=1}^N m_{ij}[(\mathrm{I}_h u)_j - \vartheta_j] - m_i\,[(\mathrm{I}_h u)_i - \vartheta_i]\Big) \\
&= \int_\Omega \vartheta_h \frac{\partial \Theta}{\partial t}\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + \sum_{i,j=1}^N \vartheta_i\, m_{ij} \frac{\mathrm{d}}{\mathrm{d}t}\big[(\mathrm{I}_h u)_j - (\mathrm{I}_h u)_i - (\vartheta_j - \vartheta_i)\big] \\
&= \int_\Omega \vartheta_h \frac{\partial \Theta}{\partial t}\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} \\
&\quad + \sum_{i,j=1 i<j}^N (\vartheta_i - \vartheta_j)\, m_{ij} \frac{\mathrm{d}}{\mathrm{d}t}\big[(\mathrm{I}_h u)_j - (\mathrm{I}_h u)_i - (\vartheta_j - \vartheta_i)\big].
\end{aligned}$$

Arguing as in the proof of Proposition 1, we invoke the divergence theorem, Lemma 3 as well as the identities $u = \Theta + \mathrm{I}_h u$ and $u_h = \mathrm{I}_h u - \vartheta_h$, which yields

$$\begin{aligned}
\Xi_2 &= \int_\Omega \vartheta_h\, \boldsymbol{v} \cdot \nabla\Theta\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\partial\Omega} \vartheta_h^2\, \boldsymbol{v} \cdot \boldsymbol{n}\,\mathrm{d}s - \int_{\Gamma_-} \vartheta_h\, \Theta\, \boldsymbol{v} \cdot \boldsymbol{n}\,\mathrm{d}s \\
&\quad - \int_{\Gamma_-} \vartheta_h\, \mathrm{I}_h u\, \boldsymbol{v} \cdot \boldsymbol{n}\,\mathrm{d}s + \sum_{i=1}^N \vartheta_i\,(u(\boldsymbol{x}_i) - \vartheta_i) \int_{\Gamma_-} \varphi_i\, \boldsymbol{v} \cdot \boldsymbol{n}\,\mathrm{d}s \\
&= \int_\Omega \vartheta_h\, \boldsymbol{v} \cdot \nabla\Theta\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\Gamma_+} \vartheta_h^2\, \boldsymbol{v} \cdot \boldsymbol{n}\,\mathrm{d}s - \int_{\Gamma_-} \vartheta_h\, \Theta\, \boldsymbol{v} \cdot \boldsymbol{n}\,\mathrm{d}s \\
&\quad - \frac{1}{2}\sum_{i,j=1 i<j}^N (\vartheta_i - \vartheta_j)^2 \int_{\Gamma_-} \varphi_i\, \varphi_j\, \boldsymbol{v} \cdot \boldsymbol{n}\,\mathrm{d}s - \frac{1}{2}\sum_{i=1}^N \vartheta_i^2 \int_{\Gamma_-} \varphi_i\, \boldsymbol{v} \cdot \boldsymbol{n}\,\mathrm{d}s
\end{aligned}$$

$$+ \int_{\Gamma_-} \big(\sum_{i=1}^{N} \vartheta_i \, \varphi_i \, (u(\boldsymbol{x}_i) - \mathrm{I}_h u) \big) \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s.$$

As in [19, Thm. 3.43], we exploit transformation to the reference element, shape-regularity, and the equivalence of norms in finite dimensional spaces to show that

$$\sum_{i=1}^{N} \int_{\Gamma} (v_i \, \varphi_i)^2 \, \mathrm{d}s \le \sum_{i=1}^{N} \int_{\Gamma} v_i^2 \, \varphi_i \, \mathrm{d}s \le C \|v_h\|^2_{\mathrm{L}^2(\Gamma)} \qquad \forall v_h \in \mathrm{V}_h, \;\; \Gamma \in \mathcal{F}_{\partial\Omega}. \tag{4.3}$$

To derive an estimate for Ξ_3, we rewrite the boundary integrals as a sum of integrals over faces. On each face $\Gamma \in \mathcal{F}_-$, we use the estimate $|\hat{u}_i^k - \hat{u}| \le C h_\Gamma |\nabla \hat{u}|$, where $h_\Gamma = \mathrm{diam}(\Gamma)$. In addition, we invoke Young's inequality, estimate (4.3), Lemma 6, and incorporate $\lambda = \|\boldsymbol{v}\|_{\mathbf{L}^\infty(\Omega \times \mathbb{R}_+)^d}$ into the constant C, which yields

$$\begin{aligned} \Xi_3 &= \sum_{i=1}^{N} \sum_{\Gamma_k \in \mathcal{F}_i} \int_{\Gamma_k} \vartheta_i \varphi_i \, (\hat{u}_i^k - \hat{u}) \, \min\{0, \boldsymbol{v} \cdot \boldsymbol{n}\} \, \mathrm{d}s \\ &\le C \sum_{\Gamma \in \mathcal{F}_-} \sum_{i=1}^{N} \int_{\Gamma} \Big[h_\Gamma (\vartheta_i \varphi_i)^2 + \frac{1}{h_\Gamma} |\hat{u}_i^k - \hat{u}|^2 \Big] \, \mathrm{d}s \\ &\le C \sum_{\Gamma \in \mathcal{F}_-} h_\Gamma \left(\|\vartheta_h\|^2_{\mathrm{L}^2(\Gamma)} + |\hat{u}|^2_{\mathrm{H}^1(\Gamma)} \right) \le C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + C h |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)}. \end{aligned} \tag{4.4}$$

For the nonlinear terms in Ξ_4, we use Lemma 2, Young's inequality, the compatibility condition (3.1) with constant $\gamma \in (0, 1)$, and Lemma 10 to deduce

$$\begin{aligned} \Xi_4 &= d_h(u_h; u_h, \mathrm{I}_h u) - d_h(u_h; u_h, u_h) + m_h(u_h; \dot{u}_h, u_h) - m_h(u_h; \dot{u}_h, \mathrm{I}_h u) \\ &\le \frac{\gamma}{2} d_h(u_h; u_h, u_h) + \frac{1}{2\gamma} d_h(u_h; \mathrm{I}_h u, \mathrm{I}_h u) - \gamma d_h(u_h; u_h, u_h) \\ &\quad - \frac{\gamma h}{\lambda} m_h(u_h, \dot{u}_h, \dot{u}_h) + \frac{\gamma h}{2\lambda} m_h(u_h; \dot{u}_h, \dot{u}_h) + \frac{\lambda}{2\gamma h} m_h(u_h; \mathrm{I}_h u, \mathrm{I}_h u) \\ &\le -\frac{\gamma}{2} d_h(u_h; u_h, u_h) - \frac{\gamma h}{2\lambda} m_h(u_h, \dot{u}_h, \dot{u}_h) + C h \|u\|^2_{\mathrm{H}^2(\Omega)}, \end{aligned}$$

where the factor $1/\gamma$ was incorporated into the constant C. Combining the above identities for Ξ_1 and Ξ_2 with the inequalities for Ξ_3 and Ξ_4 produces the estimate

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} &+ \sum_{i,j=1 i<j}^{N} m_{ij} \frac{\mathrm{d}}{\mathrm{d}t} (\vartheta_i - \vartheta_j)^2 + \int_{\Gamma_+} \vartheta_h^2 \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s - \sum_{i=1}^{N} \vartheta_i^2 \int_{\Gamma_-} \varphi_i \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s \\ &- \sum_{i,j=1 i<j}^{N} (\vartheta_i - \vartheta_j)^2 \int_{\Gamma_-} \varphi_i \, \varphi_j \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s + \gamma d_h(u_h; u_h, u_h) \end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{\gamma h}{\lambda} m_h(u_h, \dot{u}_h, \dot{u}_h) \\
&\le -2 \int_\Omega \vartheta_h \frac{\partial \Theta}{\partial t} \,\mathrm{d}\boldsymbol{x} + 2 \sum_{i,j=1 i<j}^{N} (\vartheta_i - \vartheta_j)\, m_{ij} \left[(\mathrm{I}_h \partial_t u)_i - (\mathrm{I}_h \partial_t u)_j\right] \\
&\quad - 2 \int_\Omega \vartheta_h\, \boldsymbol{v} \cdot \nabla \Theta \,\mathrm{d}\boldsymbol{x} + 2 \int_{\Gamma_-} \vartheta_h\, \Theta\, \boldsymbol{v} \cdot \boldsymbol{n} \,\mathrm{d}s + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} \\
&\quad - 2 \int_{\Gamma_-} \Big(\sum_{i=1}^{N} \vartheta_i \varphi_i \, (u(\boldsymbol{x}_i) - \mathrm{I}_h u) \Big)\, \boldsymbol{v} \cdot \boldsymbol{n} \,\mathrm{d}s + Ch \|u\|^2_{\mathrm{H}^2(\Omega)} =: (\star).
\end{aligned}
\tag{4.5}
$$

The terms on the right hand side of inequality (4.5) are now bounded using standard arguments. Specifically, we make use of the assumptions on the mesh and of Young's inequality, apply Lemma 4 to $\Theta = u - \mathrm{I}_h u$ and $\partial_t \Theta$, invoke Lemma 5 through 9 and argue as in the derivation of (4.4) to obtain

$$
\begin{aligned}
(\star) &\le \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^4 |\partial_t u|^2_{\mathrm{H}^2(\Omega)} + Ch^2 \sum_{K \in \mathcal{K}_h} |\vartheta_h|_{\mathrm{H}^1(K)} |\mathrm{I}_h \partial_t u - \partial_t u + \partial_t u|_{\mathrm{H}^1(K)} \\
&\quad + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^2 |u|^2_{\mathrm{H}^2(\Omega)} + \lambda \sum_{\Gamma \in \mathcal{F}_-} \Big(h_\Gamma \|\vartheta_h\|^2_{\mathrm{L}^2(\Gamma)} + \frac{1}{h_\Gamma} \|\Theta\|^2_{\mathrm{L}^2(\Gamma)} \Big) \\
&\quad + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} \\
&\quad + C \sum_{\Gamma \in \mathcal{F}_-} h_\Gamma \int_\Gamma \Big[\sum_{i=1}^{N} (\vartheta_i \varphi_i)^2 + |\nabla(\mathrm{I}_h u - u + u)|^2 \Big] \,\mathrm{d}s + Ch \|u\|^2_{\mathrm{H}^2(\Omega)} \\
&\le \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^4 |\partial_t u|^2_{\mathrm{H}^2(\Omega)} + Ch \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^3 |\partial_t u|^2_{\mathrm{H}^2(\Omega)} + Ch |\partial_t u|^2_{\mathrm{H}^1(\Omega)} \\
&\quad + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^2 |u|^2_{\mathrm{H}^2(\Omega)} + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^2 |u|^2_{\mathrm{H}^2(\Omega)} + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} \\
&\quad + Ch |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} + C \sum_{\Gamma \in \mathcal{F}_-} h_\Gamma \Big(\|\vartheta_h\|^2_{\mathrm{L}^2(\Gamma)} + \|u\|^2_{\mathrm{H}^2(\Gamma)} \Big) + Ch \|u\|^2_{\mathrm{H}^2(\Omega)} \\
&\le C_1 \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + C_2 h \Big(\|\partial_t u\|^2_{\mathrm{H}^2(\Omega)} + \|u\|^2_{\mathrm{H}^2(\Omega)} + \|u\|^2_{\mathrm{H}^2(\Gamma_-)} + |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} \Big).
\end{aligned}
$$

We now integrate in time observing that, by our definition of the discrete initial data, we have $\vartheta_h(0) \equiv 0$. At this stage, we recall the previously given definitions of q and z, which enables us to write the resulting inequality as

$$
\|\vartheta_h(T)\|^2_{\mathrm{L}^2(\Omega)} \le C_1 \int_0^T \|\vartheta_h(t)\|^2_{\mathrm{L}^2(\Omega)} \,\mathrm{d}t + C_2 h\, z(T) - q(T).
$$

Using Grönwall's Lemma as in [9, Lem. 1.9], we obtain

$$
\|\vartheta_h(T)\|^2_{\mathrm{L}^2(\Omega)} \le C_1 \int_0^T \mathrm{e}^{C_1(T-t)} \left(C_2 h\, z(t) - q(t) \right) \mathrm{d}t + C_2 h\, z(T) - q(T).
$$

The triangle inequality applied to $u - u_h = \Theta + \vartheta_h$ then yields the error estimate (4.1) by Lemma 4. □

We conclude the theoretical discussion with a few remarks regarding the derived error estimate. Let us first point out that in the general setting with unspecified correction factors α_{ij} our result is indeed optimal (cf. Sect. 5.1). If we set all correction factors equal to zero, we obtain the low order method, which cannot be expected to be more than $\frac{1}{2}$ order accurate in general.

A drawback of our current approach is that the constant on the right hand side of the a priori error estimate (4.2) depends exponentially on the time T. Kučera and Shu [20] demonstrate that exponentially increasing constants can be avoided in some situations. They discretize the advection equation using discontinuous Galerkin methods and derive an error estimate without invoking Grönwall's inequality. It would be interesting to investigate the merit of their approach for the purposes of our analysis.

Let us briefly remark that we assumed all integrals appearing in the bilinear and linear forms $a_h(\cdot, \cdot)$ and $b_h(\cdot)$ to be evaluated exactly. In fact, even the energy estimate stated in Proposition 1 was derived under this assumption. For polynomial velocities one can indeed employ a quadrature rule of sufficiently high order to accurately compute all integrals. For general velocities, the theory we present needs to be adapted to include quadrature errors. As is common for linear finite elements [10, Thm. 8.5], we recommend to employ quadrature rules that are exact for polynomials in $\mathbb{P}_2$ and $\mathbb{P}_3$ for volume and boundary integrals, respectively.

Admittedly, a major limitation of Proposition 2 is the fact that the estimate is valid only for problems with exact solutions of very high regularity. In particular, the assumption that $\partial_t u$ is H^2 in space is restrictive. In our opinion, the adaptation of the proofs in [5, 27] to the time-dependent setting necessitates this regularity. One can argue that if the exact solution is smooth enough for Proposition 2 to be applicable, a limiter may not even be needed and we could instead employ a stabilized Galerkin method. Since this strategy does not guarantee the validity of discrete maximum principles, AFC schemes provide an appealing alternative. Therefore, theoretical investigations of these methods should be undertaken. It is hoped that our results may serve as a stepping stone for further efforts in this direction.

5 Numerical examples

Let us now corroborate the theoretical results of this work with numerical experiments. The following acronyms are used to distinguish the numerical methods under investigation

- LOW: low order method (2.10),
- MCL: monolithic convex limiting scheme (2.15),
- target: target scheme, i. e., (2.15) with $f_{ij}^* = f_{ij}$ for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \backslash \{i\}$.

Other methods used for comparative purposes are specified below. Recall that our stability and convergence proofs rely on the compatibility condition (3.1). In general, this condition is not fulfilled by the standard MCL approach. However, if $\dot{u}_h$ is set to zero, i. e., if the mass lumping error is not compensated, (3.1) holds due to Lemma 2. To distinguish between the standard MCL scheme employing *low order time derivatives* $\dot{u}_i$ given by (2.14) and the lumped-mass version, in which $\dot{u}_h \equiv 0$, we employ the

Table 1 Convergence history for the one-dimensional advection equation on a sequence of uniform periodic meshes. The $\|\cdot\|_{L^2(\Omega)}$ errors at $T = 1$ and the corresponding EOC for $u_0(x) = \exp(-100(x - 0.5)^2)$

$1/h$	LOW	EOC	MCL-L	EOC	MCL-0	EOC
32	2.21E−01		6.92E−02		9.93E−02	
64	1.75E−01	0.34	2.07E−02	1.74	4.46E−02	1.16
128	1.26E−01	0.47	4.65E−03	2.16	1.65E−02	1.44
256	8.18E−02	0.62	1.12E−03	2.06	5.29E−03	1.64
512	4.84E−02	0.76	2.76E−04	2.02	1.65E−03	1.68

acronyms MCL-L and MCL-0, respectively. Here the letter L stands for *low order time derivatives*, while 0 stands for *zero time derivatives*.

In all simulations, we choose the time step according to (2.16) by setting $\Delta t = \nu\Delta t_{\max}$, where $\nu \in (0, 1]$ is a user-defined value and $\Delta t_{\max}$ is the right hand side of inequality (2.16). Discrete initial conditions are obtained by interpolating the continuous initial datum in the discrete space.

In the following sections, we verify that approximations converge at least as fast as the provable rate of $\frac{1}{2}$. Moreover, we stress the need for stabilization by the use of low order time derivatives (2.14) and present a comparison of results obtained with various definitions of antidiffusive fluxes in the MCL scheme. Finally, we perform an a posteriori check to see for which values of the parameter γ the compatibility condition (3.1) is satisfied by the MCL-L scheme.

5.1 Experimental orders of convergence

In this section, we solve the one-dimensional advection equation with constant velocity $v = 1$. The spatial domain $\Omega = (0, 1)$ has periodic boundaries. Thus, at each time instant $T \in \mathbb{N}_0$, the exact solution coincides with the initial condition. In this example, we use $u_0(x) = \exp(-100(x - 0.5)^2)$.

We study the experimental orders of convergence for discrete upwinding (LOW), MCL-L, and MCL-0 schemes using SSP2-RK time stepping and CFL parameter $\nu = 0.5$. While values as large as $\nu = 1$ can safely be employed without causing violations of maximum principles, smaller values may be necessary to observe certain rates of convergence. Alternatively, SSP3-RK time stepping can be used to improve the temporal accuracy. In this study, we employ sequences of nested meshes with generally nonuniform mesh size h obtained by randomly perturbing the positions of the interior mesh vertices of the coarsest grid. The relative mesh sizes $\min_{K\in\mathcal{K}_h} h_K/h$ of the three sequences are 1 (uniform), ≈ 0.69 (mildly perturbed), and ≈ 0.087 (severely perturbed), respectively. We present the $L^2(\Omega)$ errors at the final time $T = 1$ and the corresponding experimental orders of convergence (EOC) in Tables 1, 2, 3.

The observed rates are in accordance with our expectations. As suggested by Corollary 1, discrete upwinding converges at least with the rate of $\frac{1}{2}$. Actually, the low order method becomes first order accurate on very fine uniform meshes. Our preferred MCL-L scheme produces second order accurate results in this test. If no correction of the mass lumping error is performed, the order of accuracy deteriorates, while still exceed-

Table 2 Convergence history for the one-dimensional advection equation on a sequence of mildly perturbed periodic meshes. The $\| \cdot \|_{L^2(\Omega)}$ errors at $T = 1$ and the corresponding EOC for $u_0(x) = \exp(-100(x - 0.5)^2)$

$1/h$	LOW	EOC	MCL-L	EOC	MCL-0	EOC
32	2.21E−01		7.06E−02		9.97E-02	
64	1.75E−01	0.34	2.22E−02	1.67	4.51E−02	1.15
128	1.26E−01	0.47	4.95E−03	2.17	1.70E−02	1.40
256	8.23E−02	0.62	1.16E−03	2.09	5.48E−03	1.64
512	4.88E−02	0.75	2.87E−04	2.01	1.71E−03	1.68

Table 3 Convergence history for the one-dimensional advection equation on a sequence of severely perturbed periodic meshes. The $\| \cdot \|_{L^2(\Omega)}$ errors at $T = 1$ and the corresponding EOC for $u_0(x) = \exp(-100(x - 0.5)^2)$

$1/h$	LOW	EOC	MCL-L	EOC	MCL-0	EOC
32	2.24E−01		1.01E−01		1.16E−01	
64	1.82E−01	0.30	4.85E−02	1.05	5.65E−02	1.03
128	1.36E−01	0.43	1.25E−02	1.96	2.33E−02	1.28
256	9.08E−02	0.58	2.98E−03	2.07	7.72E−03	1.59
512	5.51E−02	0.72	7.26E−04	2.04	2.46E−03	1.65

ing the provable rate of $\frac{1}{2}$. In this example, the influence of mesh perturbations on the results is insignificant. A decay in the convergence rate of the standard MCL scheme for a steady problem was observed on perturbed 2D meshes in [21, Sec. 6.1].

5.2 On the stabilizing effect of low order time derivatives

Let us now compare the standard Galerkin approach to methods that are stabilized by incorporating low order time derivatives (2.14) via antidiffusive fluxes. We consider the same setup as in the previous section with the exception that the initial condition u_0 is replaced by [15]

$$u_0(x) = \begin{cases} 1 & \text{if } 0.2 \leq x \leq 0.4, \\ \exp(10)\exp(\frac{1}{0.5-x})\exp(\frac{1}{x-0.9}) & \text{if } 0.5 < x < 0.9, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

This profile features discontinuities as well as a C^∞ region. In Fig. 1 we display standard continuous Galerkin approximations obtained with four different combinations of time stepping schemes and CFL parameters on a uniform, a mildly perturbed and a severely perturbed mesh with 128 elements in each case. Spurious ripples that are not local to the vicinity of the discontinuities can be observed in all profiles. Although limiters can remove these oscillations, the quality of approximations obtained in this

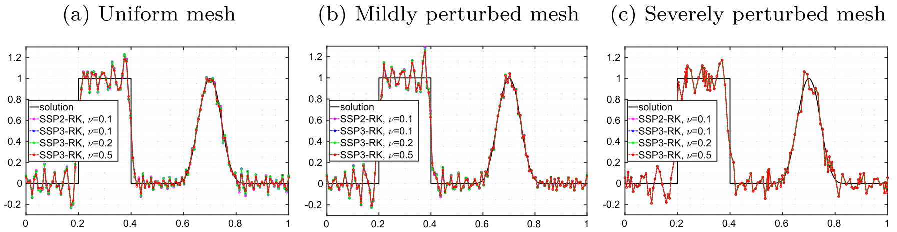

(a) Uniform mesh (b) Mildly perturbed mesh (c) Severely perturbed mesh

Fig. 1 One-dimensional advection equation with initial condition (5.1). Consistent Galerkin approximations at $T = 1$ obtained with SSP RK time stepping on periodic meshes consisting of 128 elements

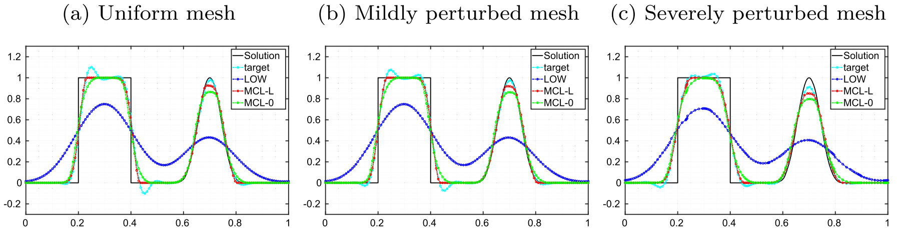

(a) Uniform mesh (b) Mildly perturbed mesh (c) Severely perturbed mesh

Fig. 2 One-dimensional advection equation with initial condition (5.1). Stabilized Galerkin approximations at $T = 1$ obtained with SSP2-RK time stepping and $\nu = 1$ on periodic meshes consisting of 128 elements

fashion is usually poor compared to solutions obtained with flux limiters applied to a stabilized target discretization (cf. Sect. 5.3).

Next, we compute approximations of the stabilized target scheme, i. e., (2.21) with $\alpha_{ij} = 1$ for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. These are compared to the profiles obtained with discrete upwinding (LOW), MCL-L, and MCL-0 schemes. SSP2-RK time stepping with CFL parameter $\nu = 1$ is employed in combination with all spatial semi-discretizations. The results of this study are displayed in Fig. 2.

We observe significant improvements in the solution quality for the unlimited target scheme compared to the consistent Galerkin approximations displayed in Fig. 1. Numerical results obtained with LOW, MCL-L and MCL-0 exhibit behavior similar to that observed in Sect. 5.1 In particular, the MCL-0 scheme produces a nonsymmetric profile in the left part of the domain, which can be attributed to dispersive errors that occur commonly if the mass lumping error is not compensated [35].

5.3 Solid body rotation

Let us now apply the MCL scheme to a 2D solid body rotation benchmark [26] in which $\Omega = (0, 1)^2$, $\boldsymbol{v}(x, y) = 2\pi\,(0.5 - y, x - 0.5)^T$, $\hat{u} = 0$ and

$$u_0(x, y) = \begin{cases} u_0^{\text{cone}}(x, y) & \text{if } r(x, y; 0.5, 0.25) \le r_0, \\ u_0^{\text{bump}}(x, y) & \text{if } r(x, y; 0.25, 0.5) \le r_0, \\ 1 & \text{if } r(x, y; 0.5, 0.75) \le r_0 \ \wedge\ (|x - 0.5| \ge 0.025 \ \vee\ y \ge 1 - r_0)\,, \\ 0 & \text{otherwise,} \end{cases}$$

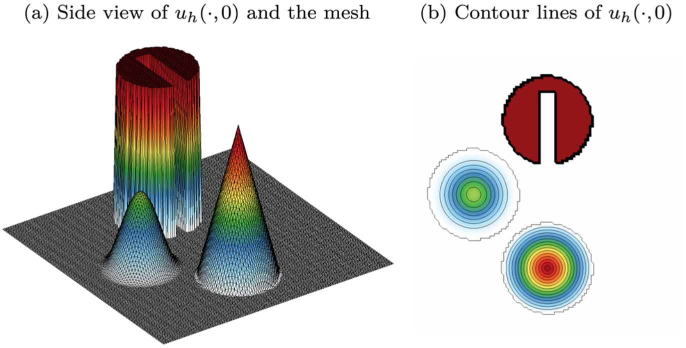


Fig. 3 Exact initial condition of the solid body rotation [26] interpolated in V_h for a uniform triangular mesh consisting of $2 \cdot 128^2$ elements

where $r(x, y; x_0, y_0) := \sqrt{(x-x_0)^2 + (y-y_0)^2}$, $r_0 = 0.15$, and

$$u_0^{\text{cone}}(x, y) = 1 - \frac{r(x, y; 0.5, 0.25)}{r_0},$$
$$u_0^{\text{bump}}(x, y) = \frac{1}{4}\left(1 + \cos\left(\frac{\pi\, r(x, y; 0.25, 0.5)}{r_0}\right)\right).$$

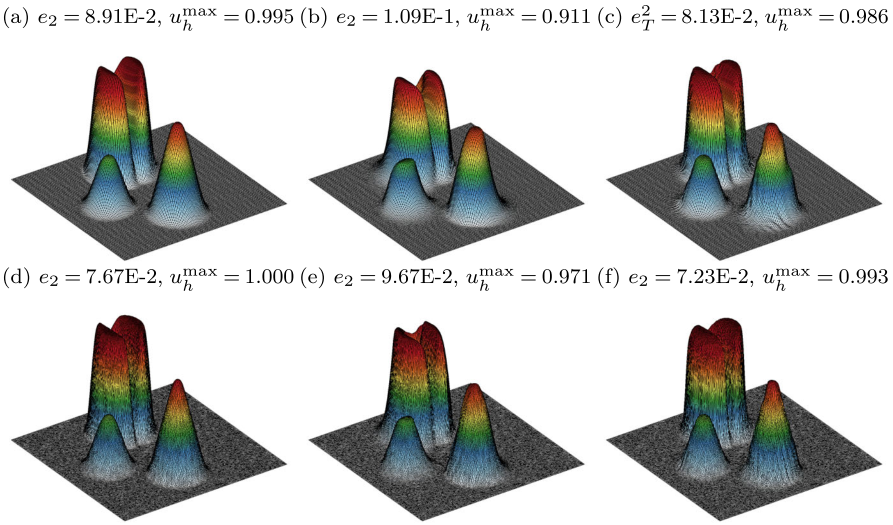
In this example, a cone, a smooth bump and a slotted cylinder rotate around the domain center. At each time instant $T \in \mathbb{N}_0$, the exact solution is equal to the initial condition, which is shown in Fig. 3. The numerical results displayed in this section are visualized with the open source C++ software GLVis [11].

We solve this problem numerically using triangular meshes and $h = c/128$, where $c = \sqrt{2}$ for uniform grids and $c = 1$ for unstructured ones. For time stepping we employ the SSP2-RK method with constant time steps $\Delta t = 5 \cdot 10^{-4}$ and $\Delta t = 3.125 \cdot 10^{-4}$, respectively. In addition to MCL-L and MCL-0 schemes, we test the MCL-G approach, in which the consistent Galerkin time derivative $\dot{u}_h^{\mathrm{G}} = \sum_{j=1}^{N} \dot{u}_j^{\mathrm{G}} \varphi_j$ defined by

$$\sum_{j=1}^{N} m_{ij} \dot{u}_j^{\mathrm{G}} = - \sum_{j \in \mathcal{N}_i \setminus \{i\}} a_{ij}(u_j - u_i) + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i), \qquad i \in \{1, \dots, N\} \tag{5.2}$$

is employed to correct the mass lumping error through the antidiffusive fluxes $f_{ij} = m_{ij}(\dot{u}_i - \dot{u}_j) + d_{ij}(u_i - u_j)$. The numerical results of this study are displayed in Fig. 4 and the approximate $\mathrm{L}^2(\Omega)$ errors e_2 at the final time $T = 1$ are presented in the captions along with the maximum solution value $u_h^{\max}$ for each approximation. The minimum value of each approximation is zero up to machine precision.

Although all obtained profiles appear to be similar to each other, we can make out some differences by closely examining the results. First, we observe that on the uniform mesh MCL-0 is noticeably more diffusive than MCL-L and MCL-G. In particular, the smooth hump is not well resolved in this approach due to dispersive errors that arise



(a) $e_2 = 8.91\text{E-}2$, $u_h^{\max} = 0.995$ (b) $e_2 = 1.09\text{E-}1$, $u_h^{\max} = 0.911$ (c) $e_T^2 = 8.13\text{E-}2$, $u_h^{\max} = 0.986$

(d) $e_2 = 7.67\text{E-}2$, $u_h^{\max} = 1.000$ (e) $e_2 = 9.67\text{E-}2$, $u_h^{\max} = 0.971$ (f) $e_2 = 7.23\text{E-}2$, $u_h^{\max} = 0.993$

Fig. 4 Solid body rotation for the 2D advection equation [26]. MCL-L (left), MCL-0 (center), and MCL-G (right) approximations at $T = 1$. Solutions obtained on uniform (top row) and unstructured (bottom row) triangular meshes with SSP2-RK time stepping using $\Delta t = 5 \cdot 10^{-4}$ and $\Delta t = 3.125 \cdot 10^{-4}$, respectively

in lumped Galerkin approximations [35]. The unstabilized MCL-G scheme does not suffer from this deficiency and, in fact, produces the smallest approximate error values among the three approaches. However, the lack of stabilization leads to a distortion in the shape of the sharp cone and, on the uniform mesh, a similar feature can be spotted in the region of the slotted cylinder. In the case of the advection equation, this issue does not seem to have a dominating effect on the obtained profiles but, in our experience [14, Sec. 3.4.3.2], the MCL-G scheme produces more pronounced spurious oscillations if applied to more involved problems like the Euler equations of gas dynamics. The quality of approximations obtained in this manner can be improved by employing smaller time steps or higher order SSP-RK methods [21, Fig. 2 (e)]. Nevertheless, some form of stabilization should be used in combination with continuous Galerkin discretizations of hyperbolic problems. Therefore, MCL-L is the preferable option among the three schemes under investigation. Alternative stabilization techniques for standard finite elements can be found in [23, 27] for linear and nonlinear problems, respectively.

5.4 A posteriori compatibility check

The compatibility condition (3.1) turned out to be an invaluable tool for our theoretical investigations. Unfortunately, we are unable to prove that the MCL-L scheme automatically produces compatible pairs $(u_h, \dot{u}_h)$ under suitable assumptions on the mesh. However, it is easy to check for which values of $\gamma \in (0, 1)$ condition (3.1) is fulfilled *a posteriori*. Indeed, (3.1) is equivalent to

$$\gamma \leq \frac{d_h(u_h; u_h, u_h) - m_h(u_h; \dot{u}_h, u_h)}{d_h(u_h; u_h, u_h) + \frac{h}{\lambda} m_h(u_h; \dot{u}_h, \dot{u}_h)} \tag{5.3}$$

Table 4 Maximum values of γ over all time steps for four 1D examples. Results obtained on uniform meshes with SSP2-RK time stepping and CFL parameters $\nu = 1$ (upper half) and $\nu = 0.1$ (lower half)

$1/h$	Test 1	Test 2	Test 3	Test 4
32	0.58	0.65	0.62	0.67
64	0.54	0.54	0.56	0.56
128	0.52	0.53	0.52	0.55
256	0.51	0.52	0.51	0.53
512	0.50	0.52	0.51	0.52
32	0.62	0.66	0.64	0.66
64	0.56	0.58	0.58	0.60
128	0.53	0.55	0.54	0.57
256	0.52	0.53	0.52	0.57
512	0.51	0.52	0.51	0.55

if the denominator in the right hand side of (5.3) is nonzero (it is nonnegative due to Lemma 2). If the numerator in (5.3) is also nonnegative, this criterion yields an upper bound on γ. Having calculated these a posteriori bounds via (5.3), one can check how they behave upon mesh refinement. Two issues that lead to a violation of compatibility can occur in practice. First, (5.3) can produce a negative upper bound on γ, a case that is not covered by our theory. Secondly, γ may approach zero upon mesh refinement, which would cause the constant in the leading order term of our error estimate to approach infinity. We found the former concern to be valid on perturbed one-dimensional meshes. Using smaller time steps does not resolve incompatibility issues, which seem to be caused by triangulations of bad quality. In such instances, our stability and error estimates are not applicable to MCL-L but remain valid for the LOW and MCL-0 schemes, as well as for the method proposed in [15, Sec. 3.3] that enforces compatibility.

Having performed the described a posteriori check for various test problems, we conjecture that compatibility of $(u_h, \dot{u}_h)$ holds for the MCL-L scheme on uniform meshes. Below we report the results of our experiments in which we compute the values of the right hand side of (5.3). First, we consider four one-dimensional test problems. In each case, the spatial domain $\Omega = (0, 1)$ is equipped with periodic boundaries and the velocity is $v = 1$. The first and second tests are the same as in Sects. 5.1 and 5.2. In the third and fourth tests, the final time is $T = 0.5$ and the initial conditions read

$$u_0(x) = \begin{cases} 0.5\left(1 + \cos\left(\frac{\pi}{0.15}(x - 0.25)\right)\right) & \text{if } |x - 0.25| < 0.15, \\ 0 & \text{otherwise,} \end{cases}$$

and $u_0(x) = \max\{0, 1 - 10|x - 0.2|\}$, respectively.

We solve each of these problems on a hierarchy of uniform meshes using SSP2-RK time stepping with CFL parameters $\nu \in \{1, 0.1\}$. The largest value of γ for which (3.1) is satisfied during the whole simulation is presented in Table 4.

Additionally, we repeat the solid body rotation test [26] from Sect. 5.3 on sequences of uniform ($c = \sqrt{2}$) and unstructured ($c = 1$) triangular meshes and compute a pos-

Table 5 Maximum values of γ over all time steps for the solid body rotation [26]. Results obtained on uniform and unstructured meshes with SSP2-RK time stepping and constant time steps

c/h	Uniform meshes	#TS	Unstructured meshes	$\min_{K\in\mathcal{K}_h} h_K/h$	#TS
32	$\gamma = 0.59$	500	$\gamma = 0.48$	0.32	625
64	$\gamma = 0.54$	1000	$\gamma = 0.45$	0.31	1250
128	$\gamma = 0.49$	2000	$\gamma = 0.46$	0.29	3200
256	$\gamma = 0.49$	4000	$\gamma = 0.47$	0.28	6400
512	$\gamma = 0.48$	8000	$\gamma = 0.47$	0.26	12500

teriori values for γ via (5.3). The results of this study are summarized in Table 5, where #TS refers to the total number of employed time steps.

We observe slightly larger maximum values of γ on coarse girds than on fine meshes. The use of smaller time steps seems to have marginal influence on the results. In all cases, we have $\gamma > 0.4$, which is consistent to the value that we used in [15] to enforce (3.1). Contrary to the 1D case, (5.3) does not produce negative values for γ even on unstructured meshes in 2D. This observation leads us to believe that the low order time derivative $\dot{u}_h$ given by (2.14) is compatible to u_h even for a certain class of unstructured meshes. Further theoretical and numerical studies are required to pinpoint, exactly which conditions a sequence of unstructured meshes has to satisfy in order to produce compatible pairs $(u_h, \dot{u}_h)$ via the standard MCL approach. The opposite point of view is that the compatibility condition (3.1) can actually be used to determine the mesh quality for mesh optimization purposes. Feasibility and benefits of this approach are yet to be determined.

6 Conclusions

We performed numerical analysis for a discretization of the linear advection equation based on an algebraic flux correction scheme. The employed monolithic convex limiting technique is a semi-discrete approach that enforces discrete maximum principles in fully discrete discretizations based on strong stability preserving Runge–Kutta methods. Outcomes of the conducted research include a stability and an a priori error estimate in the semi-discrete setting. To prove that the scheme converges with a rate of at least $\frac{1}{2}$, we formulated a compatibility condition for the discrete solution and corresponding approximate time derivatives. It is possible to enforce such constraints via additional limiting. However, our numerical examples indicate that the original MCL scheme produces essentially compatible functions if the antidiffusive fluxes are stabilized, e. g., using a low order approximation (2.14) to the nodal time derivatives.

It is hoped that the ideas presented in this work can be used for analysis of fully discrete problems and extended to nonlinear conservation laws, hopefully even systems like the Euler equations of gas dynamics. Other interesting avenues to explore in future studies include analysis of AFC schemes for other target discretizations, such as discontinuous Galerkin methods and/or higher order finite elements. Moreover, the aspects of inexact numerical integration may need to be taken into account. We invite the interested reader to participate in these research endeavors.

Acknowledgements The authors would like to thank Prof. Dmitri Kuzmin for many fruitful discussions and his contributions to an earlier version of this manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Abgrall, R.: Essentially non-oscillatory residual distribution schemes for hyperbolic problems. J. Comput. Phys. **214**, 773–808 (2006). https://doi.org/10.1016/j.jcp.2005.10.034
2. Amann, H.: Ordinary Differential Equations (De Gruyter) (1990). https://doi.org/10.1515/9783110853698
3. Anderson, R., Dobrev, V., Kolev, T., Kuzmin, D., Quezada de Luna, M., Rieben, R., Tomov, V.: High-order local maximum principle preserving (MPP) discontinuous Galerkin finite element method for the transport equation. J. Comput. Phys. **334**, 102–124 (2017). https://doi.org/10.1016/j.jcp.2016.12.031
4. Barrenechea, G.R., Burman, E., Karakatsani, F.: Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. Numer. Math. **135**, 521–545 (2017). https://doi.org/10.1007/s00211-016-0808-z
5. Barrenechea, G.R., John, V., Knobloch, P.: Analysis of algebraic flux correction schemes. SIAM J. Numer. Anal. **54**, 2427–2451 (2016). https://doi.org/10.1137/15M1018216
6. Barrenechea, G.R., John, V., Knobloch, P., Rankin, R.: A unified analysis of algebraic flux correction schemes for convection-diffusion equations. SeMA J. **75**, 655–685 (2018). https://doi.org/10.1007/s40324-018-0160-6
7. Dafermos, C.M.: Hyperbolic Conservation Laws in Continuum Physics (Springer) 1st ed. (2000)https://doi.org/10.1007/978-3-662-22019-1
8. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods (Springer) (2012). https://doi.org/10.1007/978-3-642-22980-0
9. Dolejší, V., Feistauer, M.: Discontinuous Galerkin Method (Springer) (2015). https://doi.org/10.1007/978-3-319-19267-3
10. Ern, A., Guermond, J.-L.: Theory and Practice of Finite Elements (Springer) (2004). https://doi.org/10.1007/978-1-4757-4355-5
11. GLVis: OpenGL Finite Element Visualization Tool https://glvis.org
12. Gottlieb, S., Shu, C.-W., Tadmor, E.: Strong stability-preserving high-order time discretization methods. SIAM Rev. **43**, 89–112 (2001). https://doi.org/10.1137/S003614450036757X
13. Guermond, J.-L., Popov, B.: Invariant domains and first-order continuous finite element approximation for hyperbolic systems. SIAM J. Numer. Anal. **54**, 2466–2489 (2016). https://doi.org/10.1137/16M1074291

14. Hajduk, H.: Algebraically constrained finite element methods for hyperbolic problems with applications in geophysics and gas dynamics Ph.D. thesis TU Dortmund University (2022) https://doi.org/10.17877/DE290R-22850
15. Hajduk, H., Rupp, A., Kuzmin, D.: Analysis of algebraic flux correction for semi-discrete advection problems (2021) arXiv:2104.05639math.NA
16. Harten, A., Lax, P.D., van Leer, B.: On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**, 35–61 (1983). https://doi.org/10.1137/1025002
17. Jha, A.: A residual based a posteriori error estimators for AFC schemes for convection-diffusion equations. Comput. Math. Appl. **97**, 86–99 (2021). https://doi.org/10.1016/j.camwa.2021.05.031
18. Jha, A., Ahmed, N.: Analysis of flux corrected transport schemes for evolutionary convection-diffusion-reaction equations (2021) arXiv:2103.04776math.NA
19. Knabner, P., Angermann, L.: Numerical methods for elliptic and parabolic partial differential equations (Springer) (2003). https://doi.org/10.1007/b97419
20. Kučera, V., Shu, C.-W.: On the time growth of the error of the DG method for advective problems. IMA J. Numer. Anal. **39**, 687–712 (2018). https://doi.org/10.1093/imanum/dry013
21. Kuzmin, D.: Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws. Comput. Method. Appl. M. **361**, 112804 (2020). https://doi.org/10.1016/j.cma.2019.112804
22. Kuzmin, D., Hajduk, H., Rupp, A.: Limiter-based entropy stabilization of semi-discrete and fully discrete schemes for nonlinear hyperbolic problems. Comput. Method. Appl. M. **389**, 114428 (2022). https://doi.org/10.1016/j.cma.2021.114428
23. Kuzmin, D., Quezada de Luna, M.: Entropy conservation property and entropy stabilization of high-order continuous Galerkin approximations to scalar conservation laws. Comput. Fluids **213**, 104742 (2020). https://doi.org/10.1016/j.compfluid.2020.104742
24. Kuzmin, D., Quezada de Luna, M., Ketcheson, D.I., Grüll, J.: Bound-preserving flux limiting for high-order explicit Runge-Kutta time discretizations of hyperbolic conservation laws. J. Sci. Comput. **91**, 21 (2022). https://doi.org/10.1007/s10915-022-01784-0
25. Kuzmin, D., Turek, S.: Flux correction tools for finite elements. J. Comput. Phys. **175**, 525–558 (2002). https://doi.org/10.1006/jcph.2001.6955
26. LeVeque, R.J.: High-resolution conservative algorithms for advection in incompressible flow. SIAM J. Numer. Anal. **33**, 627–665 (1996). https://doi.org/10.1137/0733033
27. Lohmann, C.: Physics-Compatible Finite Element Methods for Scalar and Tensorial Advection Problems (Springer Spektrum) (2019). https://doi.org/10.1007/978-3-658-27737-6
28. Lohmann, C., Kuzmin, D., Shadid, J.N., Mabuza, S.: Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements. J. Comput. Phys. **344**, 151–186 (2017). https://doi.org/10.1016/j.jcp.2017.04.059
29. Löhner, R.: Applied computational fluid dynamics techniques: an introduction based on finite element methods (John Wiley & Sons) (2008) https://doi.org/10.1002/9780470989746
30. Pazner, W.: Sparse invariant domain preserving discontinuous Galerkin methods with subcell convex limiting. Comput. Method. Appl. M. **382**, 113876 (2021). https://doi.org/10.1016/j.cma.2021.113876
31. Quarteroni, A., Valli, A.: Numerical Approximation of Partial Differential Equations (Springer) (1994). https://doi.org/10.1007/978-3-540-85268-1
32. Rupp, A., Hauck, M., Aizinger, V.: A subcell-enriched Galerkin method for advection problems. Comput. Math. Appl. **93**, 120–129 (2021). https://doi.org/10.1016/j.camwa.2021.04.010
33. Selmin, V.: The node-centred finite volume approach: bridge between finite differences and finite elements. Comput. Methods Appl. Mech. Engrg. **102**, 107–138 (1993). https://doi.org/10.1016/0045-7825(93)90143-L
34. Shu, C.-W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988). https://doi.org/10.1016/0021-9991(88)90177-5
35. Thompson, T.: A discrete commutator theory for the consistency and phase error analysis of semi-discrete C^0 finite element approximations to the linear transport equation. J. Comput. Appl. Math. **303**, 229–248 (2016). https://doi.org/10.1016/j.cam.2016.02.042