

# Subcell flux limiting for high-order Bernstein finite element discretizations of scalar hyperbolic conservation laws

Dmitri Kuzmin<sup>a,\*</sup>, Manuel Quezada de Luna<sup>b</sup>

<sup>a</sup>*Institute of Applied Mathematics (LS III), TU Dortmund University  
Vogelpothsweg 87, D-44227 Dortmund, Germany*

<sup>b</sup>*King Abdullah University of Science and Technology (KAUST)  
Thuwal 23955-6900, Saudi Arabia*

---

## Abstract

This work extends the concepts of algebraic flux correction and convex limiting to continuous high-order Bernstein finite element discretizations of scalar hyperbolic problems. Using an array of adjustable diffusive fluxes, the standard Galerkin approximation is transformed into a nonlinear high-resolution scheme which has the compact sparsity pattern of the piecewise-linear or multilinear subcell discretization. The representation of this scheme in terms of invariant domain preserving states makes it possible to prove the validity of local discrete maximum principles under CFL-like conditions. In contrast to predictor-corrector approaches based on the flux-corrected transport methodology, the proposed flux limiting strategy is monolithic, i.e., limited antidiffusive terms are incorporated into the well-defined residual of a nonlinear (semi-)discrete problem. A stabilized high-order Galerkin discretization is recovered if no limiting is performed. In the limited version, the compact stencil property prevents direct mass exchange between nodes that are not nearest neighbors. A formal proof of sparsity is provided for simplicial and box elements. The involved element contributions can be calculated efficiently making use of matrix-free algorithms and precomputed element matrices of the reference element. Numerical studies for  $\mathbb{Q}_2$  discretizations of linear and nonlinear two-dimensional test problems illustrate the virtues of monolithic convex limiting based on subcell flux decompositions.

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

---



---

\*Corresponding author

*Email addresses:* [kuzmin@math.uni-dortmund.de](mailto:kuzmin@math.uni-dortmund.de) (Dmitri Kuzmin), [manuel.quezada@kaust.edu.sa](mailto:manuel.quezada@kaust.edu.sa) (Manuel Quezada de Luna)

## 1. Introduction

Algebraic flux correction (AFC) [6, 7, 35, 41] is a general framework for the design of bound-preserving finite element schemes. Many representatives of nonlinear high-resolution AFC schemes are based on algebraic interpretations and generalizations of flux-based structured grid methods for hyperbolic conservation laws. Finite element AFC versions of upwinding techniques, flux-corrected transport (FCT) algorithms [9, 50], total variation diminishing (TVD) limiters [24, 25], and their local extremum diminishing (LED) counterparts [26, 27] have been used since the late 1980s [4, 35, 38, 40, 43, 44, 45, 47, 48]. In recent years, their further development was stimulated by major breakthroughs in theoretical analysis of the involved ‘variational crimes’. The work of Barrenea et al. [5, 6, 7] established a theoretical framework for proving convergence and well-posedness of AFC schemes for steady convection-diffusion equations. Lohmann [41] extended this framework to finite element discretizations of steady and unsteady linear advection problems. Guermond et al. [18, 17, 19, 20] introduced a family of explicit invariant domain preserving (IDP) schemes for nonlinear hyperbolic problems. Their analytical studies paved the way for the development of novel convex limiting techniques [17, 21, 34] based on generalizations of localized FCT schemes [10, 42] and monolithic AFC approaches [34].

As of this writing, the overwhelming majority of algebraic flux correction tools and the underlying theory are not readily applicable to finite element approximations of degree  $p > 1$ . Using the Bernstein basis representation, a few element-based high-order extensions of residual distribution methods [1, 22] and localized FCT schemes [3, 42] were developed for continuous and discontinuous Galerkin discretizations. A common drawback of the underlying limiting techniques for antidiffusive element contributions is the possibility of direct mass exchange between all nodes of a high-order Bernstein element. This lack of locality was found to be acceptable in applications to linear advection problems [3, 22, 42] but the design of high-resolution AFC schemes for nonlinear conservation laws calls for the use of flux-based subcell approximations with compact computational stencils.

The AFC methodology that we introduce in the present paper converts a high-order continuous Galerkin discretization into a nonlinear IDP scheme with the compact sparsity pattern of a piecewise  $\mathbb{P}_1/\mathbb{Q}_1$  subcell approximation. We begin in §2 with the description of the high-order Bernstein finite element discretization. Then, in §3, we derive a low-order IDP approximation which has a compact stencil and is less diffusive than the full stencil version using the same kind of algebraic residual correction (discrete upwinding [35, 38, 42] or Rusanov dissipation [19, 22, 34, 36]). Next, in §4, we present a monolithic convex limiting procedure for the antidiffusive correction terms corresponding to a (stabilized) high-order target. The compact stencil property is preserved using a decomposition of the antidiffusive element contributions into subcell fluxes between nearest neighbor nodes. This approach, which is described in §5, involves the solution of small sparse linear systems on each macroelement. The IDP property of the corresponding discrete problem is shown using the proof techniques developed in [17, 34]. In §6 and §7, we discuss the optional stabilization techniques for the high-order target flux and Laplacian-based smoothness indicators that preserve the high-order accuracy near smooth local extrema. Time integration is performed using an explicit (third order with three stages) strong stability preserving Runge-Kutta method [15, 49]. The possibility of using precomputed element matrices of the reference element and matrix-free solvers for the global system may be exploited in efficient

implementations of the proposed algorithms. The results of numerical studies for linear and nonlinear conservation laws are presented in §8. Finally, we close in §9 with conclusions.

## 2. High-order Bernstein finite element discretization

We restrict our presentation to the case of a scalar conservation law. An extension of the proposed methodology to nonlinear hyperbolic systems can be carried out as in [34] and will be presented elsewhere. Let  $u(\mathbf{x}, t)$  be a scalar quantity of interest depending on the space location  $\mathbf{x} \in \mathbb{R}^d$ ,  $d \in \{1, 2, 3\}$  and time instant  $t \geq 0$ . Consider an initial-boundary value problem of the form [17, 34]

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

$$u(\cdot, 0) = u_0 \quad \text{in } \Omega, \quad (1b)$$

$$(u - u_{\text{in}})\mathbf{f}'(u) \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_-, \quad (1c)$$

where  $\Omega \subset \mathbb{R}^d$  is a bounded domain,  $\mathbf{f} = (f_1, \dots, f_d)$  is a possibly nonlinear flux function,  $u_0$  is the initial data,  $u_{\text{in}}$  is the Dirichlet boundary data,  $\mathbf{n}$  is the unit outward normal to the Lipschitz boundary  $\Gamma = \partial\Omega$ , and  $\Gamma_- = \{\mathbf{x} \in \Gamma : \mathbf{f}'(u) \cdot \mathbf{n} < 0\}$  is the hyperbolic inlet.

Suppose that the exact solution  $u$  belongs to a convex set  $\mathcal{G} \subset \mathbb{R}$  for all  $t \geq 0$ . Then  $\mathcal{G}$  is called an invariant set of problem (1a)–(1c), and it is natural to require that numerical approximations belong to (a subset of)  $\mathcal{G}$  as well. Adopting the terminology of Guermond et al. [17, 19, 20], we will call a discretization of problem (1a)–(1c) invariant domain preserving (IDP) if the solution of the (semi-)discrete problem is guaranteed to stay in a convex invariant set.

To begin with, we discretize (1a) in space using a high-order continuous Galerkin method. Given a conforming mesh  $\mathcal{T}_h = \{K^1, \dots, K^{E_h}\}$ , we define a finite element approximation  $u_h \approx u$  in terms of globally continuous piecewise-polynomial basis functions  $\varphi_j$ , where  $j \in \{1, \dots, N_h\}$  is the global number of a nodal point  $\mathbf{x}_j$ . The local number  $j_e = \mathcal{I}^e(j)$  of node  $j$  in  $K^e$  is determined by a mapping  $\mathcal{I}^e : \{1, \dots, N_h\} \rightarrow \{1, \dots, N\}$ . The corresponding local basis function is denoted by  $\varphi_{j_e}^e$ . The global numbers of nodes  $\mathbf{x}_1^e, \dots, \mathbf{x}_N^e$  belonging to  $K^e$  are stored in the integer set  $\mathcal{N}^e \subset \{1, \dots, N_h\}$ .

The polynomial restriction of  $u_h = \sum_{j=1}^{N_h} u_j \varphi_j$  to element  $K^e$ ,  $e = 1, \dots, E_h$  is given by

$$u_h^e := u_h|_{K^e} = \sum_{j \in \mathcal{N}^e} u_j \varphi_j = \sum_{j \in \mathcal{N}^e} u_j \varphi_{j_e}^e = \sum_{i=1}^N u_i^e \varphi_i^e, \quad (2)$$

where  $u_i^e = u_{j_e}$  is the degree of freedom (DoF) associated with the nodal point  $\mathbf{x}_i^e = \mathbf{x}_{j_e}$ ,  $j \in \mathcal{N}^e$ .

To enforce the IDP property using algebraic flux correction [6, 7, 35, 41] in what follows, we will use the Bernstein basis representation of  $u_h$ . The Bernstein basis functions  $\varphi_j^e$ , the definition of which for simplicial and tensor product meshes can be found in the Appendix, are nonnegative and form a partition of unity, i.e.,  $\sum_{j=1}^N \varphi_j^e \equiv 1$ . It follows that for any  $\mathbf{x} \in K^e$ , the state  $u_h(\mathbf{x})$  is a convex combination of the nodal states  $u_1^e, \dots, u_N^e$ . Thus, we have

$$u_1^e, \dots, u_N^e \in \mathcal{G} \quad \Rightarrow \quad u_h(\mathbf{x}) \in \mathcal{G} \quad \forall \mathbf{x} \in K^e \quad (3)$$

for any convex invariant set  $\mathcal{G}$  of the hyperbolic initial-boundary value problem (1a)–(1c).

Integrating the weighted residuals of (1a) and (1c) over  $\Omega$  and  $\Gamma_-$ , respectively, we obtain a weak form of the problem at hand. The standard Galerkin discretization replaces it with

$$\sum_{e=1}^{E_h} \int_{\Omega} w_h \left( \frac{\partial u_h}{\partial t} + \nabla \cdot \mathbf{f}(u_h) \right) d\mathbf{x} = \sum_{e=1}^{E_h} \int_{\partial K^e \cap \Gamma_-} w_h (u_h - u_{\text{in}}) \mathbf{f}'(u_h) \cdot \mathbf{n} ds \quad \forall w_h \in W_h, \quad (4)$$

where  $W_h$  is the finite-dimensional space spanned by the Bernstein basis functions  $\varphi_1, \dots, \varphi_{N_h}$ .

Substitution of (2) into (4) with the test function  $w_h = \varphi_i$  produces the semi-discrete equation

$$\sum_{j \in \mathcal{N}_i} m_{ij} \frac{du_j}{dt} = b_i(u_h, u_{\text{in}}) - \sum_{e \in \mathcal{E}_i} \int_{K^e} \varphi_i \nabla \cdot \mathbf{f}(u_h) d\mathbf{x}, \quad (5)$$

where  $\mathcal{E}_i$  is the set of elements containing node  $i$  and  $\mathcal{N}_i$  is the set of nodes belonging to these elements. The entries  $m_{ij}$  of the global consistent mass matrix and the boundary term  $b_i$  are defined by

$$m_{ij} = \sum_{e \in \mathcal{E}_i \cap \mathcal{E}_j} m_{ij}^e, \quad m_{ij}^e = \int_{K^e} \varphi_i \varphi_j d\mathbf{x}, \quad (6)$$

$$b_i(u_h, u_{\text{in}}) = \sum_{e \in \mathcal{E}_i} \int_{\partial K^e \cap \Gamma_-} \varphi_i (u_h - u_{\text{in}}) \mathbf{f}'(u_h) \cdot \mathbf{n} ds. \quad (7)$$

In practice, only the  $N^2$  nonvanishing entries of element matrices like  $M_C^e = \{m_{ij}^e\}_{i,j=1}^{N_h}$  are calculated and inserted into global matrices. To avoid conversion between global and local indices, we will use the global index notation for element matrices and vectors in this paper.

### 3. Low-order Bernstein finite element discretization

A space discretization of the form (5) can be transformed into a compact-stencil IDP scheme by using row-sum mass lumping and modifying the Galerkin element contributions

$$\int_{\partial K^e \cap \Gamma_-} \varphi_i (u_h - u_{\text{in}}) \mathbf{f}'(u_h) \cdot \mathbf{n} ds - \int_{K^e} \varphi_i \nabla \cdot \mathbf{f}(u_h) d\mathbf{x}. \quad (8)$$

Approximating the flux  $\mathbf{f}(u_h)$  by the *group finite element* interpolant [8, 13, 14, 47, 48]

$$\mathbf{f}_h^e = \sum_{j \in \mathcal{N}^e} \mathbf{f}_j \varphi_j, \quad \mathbf{f}_j = (f_{j,1}, \dots, f_{j,d}) = \mathbf{f}(u_j) \quad (9)$$

and using a lumped approximation of the boundary term, we replace (8) with

$$\int_{\partial K^e \cap \Gamma_-} \varphi_i (u_i - u_{\text{in}}) \mathbf{f}'(u_h) \cdot \mathbf{n} ds - \sum_{j \in \mathcal{N}^e} \mathbf{c}_{ij}^e \cdot \mathbf{f}_j. \quad (10)$$

The vector valued coefficients  $\mathbf{c}_{ij}^e = (c_{ij,1}^e, \dots, c_{ij,d}^e)$  of the discrete gradient operator are defined by

$$\mathbf{c}_{ij}^e = \int_{K^e} \varphi_i \nabla \varphi_j \, d\mathbf{x} = -\mathbf{c}_{ji}^e + \int_{\partial K^e} \varphi_i \varphi_j \mathbf{n} \, ds. \quad (11)$$

The transformation of the consistent element mass matrix  $M_C^e = \{m_{ij}^e\}_{i,j=1}^{N_h}$  into its lumped counterpart  $M_L^e = \{\delta_{ij} m_i^e\}_{i,j=1}^{N_h}$  with the diagonal entries

$$m_i^e = \sum_{j=1}^{N_h} m_{ij}^e = \sum_{j \in \mathcal{N}_i} m_{ij}^e = \int_{K^e} \varphi_i \, d\mathbf{x} = \frac{|K^e|}{N} > 0 \quad (12)$$

corresponds to multiplication by the local mass lumping operator

$$P^e = M_L^e (M_C^e)^{-1}. \quad (13)$$

Following the approach proposed in [42], we apply  $P^e$  to  $C_k^e = \{c_{ij,k}^e\}_{i,j=1}^{N_h}$ ,  $k = 1, \dots, d$  as well. As shown in [42] for the 1D case, this modification produces **sparse** element matrices

$$\tilde{C}_k^e = P^e C_k^e, \quad k = 1, \dots, d \quad (14)$$

such that  $c_{ij,k}^e = 0$  for  $j \notin \tilde{\mathcal{N}}_i^e$ , where  $\tilde{\mathcal{N}}_i^e \subseteq \mathcal{N}^e$  is the local stencil of the  $\mathbb{P}_1/\mathbb{Q}_1$  subcell discretization (see Fig. 1), i.e., the integer set containing the local numbers of the nearest neighbors of node  $i$  in  $K^e$ . In the Appendix, we show the compact-stencil property of the element contributions  $\tilde{\mathbf{C}}^e = (\tilde{C}_1^e, \dots, \tilde{C}_d^e)$  to the lumped discrete gradient operator for  $d$ -simplex and  $d$ -box Bernstein elements.

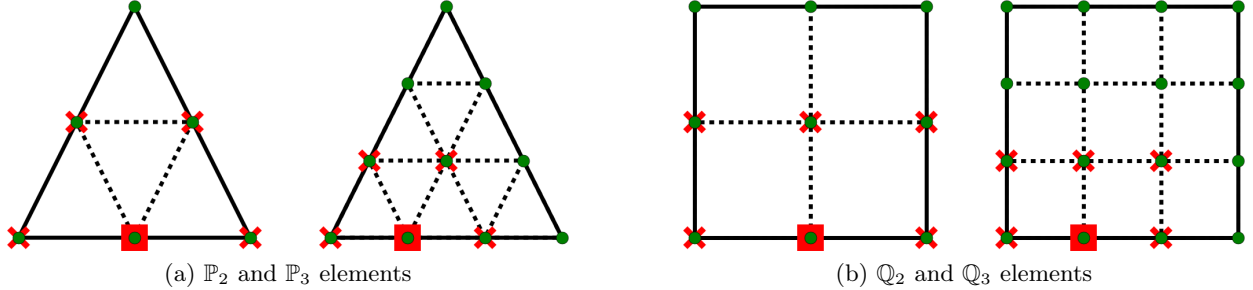


Figure 1: Nodes and subcells of typical high-order elements. The boundary of the macroelement  $K^e$  is marked with solid black lines. The internal boundaries of its subcells are marked with dashed black lines. All local DoFs are marked with green circles. The red crosses correspond to the nearest neighbors of the DoF marked by the red square.

**Remark 1.** Strict positivity of all lumped mass matrix entries  $m_i$  and the compact sparsity pattern of  $\tilde{\mathbf{C}}^e$  are due to the use of the Bernstein basis. High-order Lagrange finite elements do not provide these properties which will play an important role in the derivation of the proposed correction procedures.

**Remark 2.** In [42] and [22], the mass lumping operator  $P^e$  was applied to the element matrix of the advective term discretized without using the group finite element formulation (9) for the linear flux function  $\mathbf{f}(\mathbf{x}, u) = \mathbf{v}(\mathbf{x})u$ . This approach does not guarantee exact sparsity for general velocity fields  $\mathbf{v}(\mathbf{x})$ . As a consequence, the resulting schemes become less accurate as the polynomial degree  $p$  is increased while keeping the total number of DoFs  $N_h$  fixed [22].

The replacement of  $M_C^e$  and  $\mathbf{C}^e = (C_1^e, \dots, C_d^e)$  with the lumped element matrices  $M_L^e$  and  $\tilde{\mathbf{C}}^e$  is not enough to guarantee that the modified Galerkin scheme is IDP. To enforce the IDP property in a provable manner, we replace the element vector  $\tilde{\mathbf{C}}^e \cdot \mathbf{f}^e = \sum_{k=1}^d C_k^e \mathbf{f}_k^e$  by  $\tilde{\mathbf{C}}^e \cdot \mathbf{f}^e - \tilde{D}^e u^e$ , where  $\tilde{D}^e = \{\tilde{d}_{ij}^e\}_{i,j=1}^{N_h}$  is the element matrix of a graph Laplacian (discrete diffusion) operator.

The above manipulations convert (5) into the compact-stencil low-order approximation

$$m_i \frac{du_i}{dt} = \sum_{e \in \mathcal{E}_i} \left( \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} \tilde{d}_{ij}^e (u_j - u_i) - \sum_{j \in \tilde{\mathcal{N}}_i^e} \tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{f}_j \right) + \tilde{b}_i(u_h, u_{\text{in}}), \quad (15)$$

where  $m_i = \sum_{e \in \mathcal{E}_i} m_i^e$  is a diagonal entry of the global lumped mass matrix and

$$\tilde{b}_i(u_h, u_{\text{in}}) = \sum_{e \in \mathcal{E}_i} \int_{\partial K^e \cap \Gamma_-} \varphi_i(u_i - u_{\text{in}}) \mathbf{f}'(u_h) \cdot \mathbf{n} \, ds. \quad (16)$$

To define artificial diffusion coefficients  $\tilde{d}_{ij}^e$  that guarantee the IDP property for general hyperbolic problems, we write (15) in the equivalent form

$$m_i \frac{du_i}{dt} = \sum_{e \in \mathcal{E}_i} \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} 2\tilde{d}_{ij}^e (\bar{u}_{ij}^e - u_i) + \tilde{b}_i(u_h, u_{\text{in}}), \quad (17)$$

where

$$\bar{u}_{ij}^e = \frac{u_i + u_j}{2} - \frac{\tilde{\mathbf{c}}_{ij}^e \cdot (\mathbf{f}_j - \mathbf{f}_i)}{2\tilde{d}_{ij}^e}. \quad (18)$$

Guermond and Popov [19] were the first to recognize that representations of explicit schemes in terms of the bar states  $\bar{u}_{ij}$  lead to remarkably simple proofs of the IDP property. Indeed, (17) exhibits the structure of a discretized diffusion equation in which the nodal state  $u_j \in \mathcal{G}$  is replaced with  $\bar{u}_{ij}^e \in \mathcal{G}$ .

If time discretization is performed using an explicit SSP Runge-Kutta method [15], each stage is a forward Euler update of the form

$$m_i \bar{u}_i = m_i u_i + \Delta t \sum_{e \in \mathcal{E}_i} \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} 2\tilde{d}_{ij}^e (\bar{u}_{ij}^e - u_i) + \tilde{b}_i(u_h, u_{\text{in}}). \quad (19)$$

The result is IDP for time steps  $\Delta t$  satisfying the CFL-like condition

$$\Delta t \left( \sum_{e \in \mathcal{E}_i} \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} 2\tilde{d}_{ij}^e - \sum_{e \in \mathcal{E}_i} \int_{\partial K^e \cap \Gamma_-} \varphi_i \mathbf{f}'(u_h) \cdot \mathbf{n} \, ds \right) \leq m_i \quad (20)$$

provided that all  $\bar{u}_{ij}^e$  stay in  $\mathcal{G}$  for  $u_i, u_j \in \mathcal{G}$ . As explained in [19], this requirement can be satisfied by using the guaranteed maximum speed (GMS)

$$\lambda_{ij}^e = \max_{\omega \in [0,1]} |\mathbf{n}_{ij}^e \cdot \mathbf{f}'(\omega u_i + (1-\omega)u_j)|, \quad \tilde{\mathbf{n}}_{ij}^e = \frac{\tilde{\mathbf{c}}_{ij}^e}{|\tilde{\mathbf{c}}_{ij}^e|} \quad (21)$$

to define the Rusanov-type artificial viscosity coefficients

$$\tilde{d}_{ij}^e = \begin{cases} \max\{|\tilde{\mathbf{c}}_{ij}^e|, |\tilde{\mathbf{c}}_{ji}^e|\} \max\{\lambda_{ij}^e, \lambda_{ji}^e\} & \text{if } i \in \mathcal{N}^e, j \in \mathcal{N}^e \setminus \{i\}, \\ -\sum_{k \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} \tilde{d}_{ik}^e & \text{if } j = i \in \mathcal{N}^e, \\ 0 & \text{otherwise} \end{cases} \quad (22)$$

such that [34]

$$\min\{u_i, u_j\} \leq \bar{u}_{ij}^e \leq \max\{u_i, u_j\}. \quad (23)$$

Note that the element matrix  $\tilde{D}^e$  has the same compact sparsity pattern as  $\tilde{\mathbf{C}}^e$ .

For linear flux functions of the form  $\mathbf{f}(\mathbf{x}, u) = \mathbf{v}(\mathbf{x})u$ , where  $\mathbf{v}$  is a spatially variable velocity field, the validity of (23) cannot be guaranteed, e.g., in the case when  $u_i = u_j$  and  $\mathbf{v}_i \neq \mathbf{v}_j$  [34]. The edge contributions of the low-order scheme defined by (19) and (22) are given by

$$\begin{aligned} 2\tilde{d}_{ij}^e(\bar{u}_{ij}^e - u_i) &= \tilde{d}_{ij}^e(u_j - u_i) - \tilde{\mathbf{c}}_{ij}^e \cdot (\mathbf{v}_j u_j - \mathbf{v}_i u_i) \\ &= \underbrace{(\tilde{d}_{ij}^e - \tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{v}_j)}_{\in [0, 2\tilde{d}_{ij}^e]} u_j - \underbrace{(\tilde{d}_{ij}^e - \tilde{\mathbf{c}}_{ji}^e \cdot \mathbf{v}_i)}_{\in [0, 2\tilde{d}_{ij}^e]} u_i. \end{aligned}$$

Adapting the GMS formula (21) to the case of linear advection, the maximum speeds that appear in definition (22) of the Rusanov diffusion coefficient  $\tilde{d}_{ij}^e$  can be redefined as  $\lambda_{ij}^e = \max_{\mathbf{x} \in K^e} |\mathbf{v}(\mathbf{x})|$ . The resulting approximation is IDP w.r.t.  $\mathcal{G} = \{u \in \mathbb{R} \mid u \geq 0\}$  under the time step restriction (20).

A less dissipative low-order scheme for the linear advection equation can be constructed using

$$\tilde{d}_{ij}^e = \begin{cases} \max\{\tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{v}_j, 0, \tilde{\mathbf{c}}_{ji}^e \cdot \mathbf{v}_i\} & \text{if } i \in \mathcal{N}^e, j \in \mathcal{N}^e \setminus \{i\}, \\ -\sum_{k \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} \tilde{d}_{ik}^e & \text{if } j = i \in \mathcal{N}^e, \\ 0 & \text{otherwise.} \end{cases} \quad (24)$$

This alternative to (22) is known as *discrete upwinding* [35, 38, 42]. In view of the fact that

$$\begin{aligned} 2\tilde{d}_{ij}^e(\bar{u}_{ij}^e - u_i) &= \max\{\tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{v}_j, 0, \tilde{\mathbf{c}}_{ji}^e \cdot \mathbf{v}_i\} (u_j - u_i) - \tilde{\mathbf{c}}_{ij}^e \cdot (\mathbf{v}_j u_j - \mathbf{v}_i u_i) \\ &= \underbrace{(\max\{\tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{v}_j, 0, \tilde{\mathbf{c}}_{ji}^e \cdot \mathbf{v}_i\} - \tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{v}_j)}_{\geq 0} u_j \\ &\quad - (\max\{\tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{v}_j, 0, \tilde{\mathbf{c}}_{ji}^e \cdot \mathbf{v}_i\} - \tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{v}_i) u_i, \end{aligned}$$

the corresponding low-order scheme (19) is positivity-preserving for sufficiently small time steps  $\Delta t$ . It is at most as diffusive as the one based on (22) since  $|\tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{v}_j| \leq |\tilde{\mathbf{c}}_{ij}^e| \max_{\mathbf{x} \in K^e} |\mathbf{v}(\mathbf{x})| = |\tilde{\mathbf{c}}_{ij}^e| \lambda_{ij}^e$ .

In §8, we solve linear advection problems using (24). For nonlinear conservation laws, we use the GMS formula (22). As remarked by Guermond and Popov [19], the use of (24) with the nodal speeds  $\mathbf{v}_i := \mathbf{f}'(u_i)$  may result in entropy-violating weak solutions to nonlinear problems.

**Remark 3.** Instead of assembling the global graph Laplacian  $\tilde{D}$  from sparse element matrices  $\tilde{D}^e$  defined by (22) or (24), the global discrete gradient operator  $\tilde{\mathbf{C}}$  can be used to generate  $\tilde{D}$  after the element-by-element assembly from  $\tilde{\mathbf{C}}^e$ , cf. [17, 34].

**Remark 4.** The use of explicit SSP Runge-Kutta time discretizations is not a necessary condition for provable preservation of invariant domains. However, the verification of IDP properties for implicit and stationary versions of our low-order scheme requires more sophisticated analysis (cf. [6, 7, 41]).

As we show in the next section, the bar state form (17) of (15) is also ideally suited for the derivation of high-order extensions that preserve the IDP property using built-in flux limiters.

#### 4. Convex limiting for high-order subcell fluxes

Decomposing (5) into the low-order IDP part (15) and a remainder, we write it in the form

$$m_i \frac{du_i}{dt} = \sum_{e \in \mathcal{E}_i} \left( \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} \tilde{d}_{ij}^e (u_j - u_i) - \sum_{j \in \tilde{\mathcal{N}}_i} \tilde{\mathbf{c}}_{ij}^e \cdot \mathbf{f}_j + f_i^e + g_i^e \right) + \tilde{b}_i(u_h, u_{\text{in}}), \quad (25)$$

where

$$\begin{aligned} f_i^e &= \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} \tilde{d}_{ij}^e (u_i - u_j) + \sum_{j \in \mathcal{N}^e \setminus \{i\}} m_{ij}^e (\dot{u}_i - \dot{u}_j) + \sum_{j \in \mathcal{N}^e} (\tilde{\mathbf{c}}_{ij}^e - \mathbf{c}_{ij}^e) \cdot \mathbf{f}_j - \sum_{j \in \mathcal{N}^e} \mathbf{c}_{ji}^e \cdot \mathbf{f}_j \\ &\quad + \int_{K^e} \nabla \varphi_i \cdot \mathbf{f}(u_h) \, d\mathbf{x} + \int_{\partial K^e \cap \Gamma_-} \varphi_i (u_h - u_i) \mathbf{f}'(u_h) \cdot \mathbf{n} \, ds, \end{aligned} \quad (26)$$

$$g_i^e = \int_{\partial K^e \cap \Gamma} \varphi_i (\mathbf{f}_h - \mathbf{f}(u_h)) \cdot \mathbf{n} \, ds. \quad (27)$$

The time derivatives  $\dot{u}_i$  of the Bernstein coefficients corresponding to the standard Galerkin approximation (5) are given by the solution of the linear system

$$\sum_{j \in \mathcal{N}_i} m_{ij} \dot{u}_j = b_i(u_h, u_{\text{in}}) - \sum_{e \in \mathcal{E}_i} \int_{K^e} \varphi_i \nabla \cdot \mathbf{f}(u_h) \, d\mathbf{x}, \quad i = 1, \dots, N_h. \quad (28)$$

By definition (13) of the local mass lumping operator  $P^e$ , we have

$$\tilde{C}_k^e - C_k^e = P^e C_k^e - C_k^e = (M_L^e - M_C^e)(M_C^e)^{-1} C_k^e.$$



Using the global matrix/vector notation, the vector  $f^e = \{f_i^e\}_{i=1}^{N_h}$  of antidiffusive element contributions defined by (26) can be written as

$$f^e = (M_L^e - M_C^e)(\dot{u} + (M_C^e)^{-1} \mathbf{C}^e \cdot \mathbf{f}) - \tilde{D}^e u - (\mathbf{C}^e)^\top \cdot \mathbf{f} + r^e, \quad (29)$$

where  $r^e$  is an element vector containing the contributions

$$r_i^e = \int_{K^e} \nabla \varphi_i \cdot \mathbf{f}(u_h) \, d\mathbf{x} + \int_{\partial K^e \cap \Gamma_-} \varphi_i(u_h - u_i) \mathbf{f}'(u_h) \cdot \mathbf{n} \, ds.$$

For any element vector  $v^e \in \mathbb{R}^{N_h}$ , the components of the matrix-vector products  $(M_L^e - M_C^e)v^e$  and  $\tilde{D}^e v^e$  sum to zero. Moreover, the partition of unity property of the Bernstein basis functions  $\varphi_i$  implies that  $\sum_{i=1}^{N_h} \nabla \varphi_i = \mathbf{0}$  and, therefore,  $\sum_{i=1}^{N_h} \mathbf{c}_{ji}^e = \mathbf{0}$  by definition (11). It follows that

$$\sum_{i=1}^{N_h} f_i^e = \sum_{i \in \mathcal{N}^e} f_i^e = 0 \quad \forall e = 1, \dots, E_h. \quad (30)$$

The full element matrices  $M_C^e$  and  $\mathbf{C}^e$  can be calculated just once on the reference element and multiplied by element-dependent Jacobian data. A formula for  $\tilde{\mathbf{C}}^e$  is presented in the Appendix. Note that the involved integrals  $\sum_{j \in \mathcal{N}^e \setminus \{i\}} m_{ij}^e (\dot{u}_i - \dot{u}_j) = \int_{K^e} \varphi_i (\dot{u}_i - \dot{u}_h) \, d\mathbf{x}$ ,  $\sum_{j \in \mathcal{N}^e} \mathbf{c}_{ij}^e \cdot \mathbf{f}_j = \int_{K^e} \varphi_i \nabla \cdot \mathbf{f}_h \, d\mathbf{x}$ , and  $\sum_{j \in \mathcal{N}^e} \mathbf{c}_{ji}^e \cdot \mathbf{f}_j = \int_{K^e} \nabla \varphi_i \cdot \mathbf{f}_h \, d\mathbf{x}$  can also be calculated directly in a matrix-free manner.

In the next section, we decompose  $f_i^e$  into a sum of antidiffusive subcell fluxes  $f_{ij}^e$  such that

$$f_i^e = \sum_{j \in \tilde{\mathcal{N}}_i^e} f_{ij}^e, \quad f_{ji}^e = -f_{ij}^e \quad \forall j \in \tilde{\mathcal{N}}_i^e. \quad (31)$$

Restricting the monolithic convex limiting strategy proposed in [34] to  $f_{ij}^e$ , we will correct the bar states  $\bar{u}_{ij}^e$  of the low-order IDP scheme (17) in a bound-preserving manner. The limited counterpart  $f_{ij}^{e,*}$  of  $f_{ij}^e$  preserves the discrete conservation property and is local extremum diminishing if

$$f_{ij}^{e,*} = 0 \quad \forall j \notin \tilde{\mathcal{N}}_i^e, \quad f_{ji}^{e,*} = -f_{ij}^{e,*} \quad \forall j \in \tilde{\mathcal{N}}_i^e, \quad (32)$$

$$\bar{u}_{ij}^e \in \mathcal{G} \cap \mathcal{G}_i \quad \Rightarrow \quad \bar{u}_{ij}^{e,*} = \bar{u}_{ij}^e + \frac{f_{ij}^{e,*}}{2\tilde{d}_{ij}^e} \in \mathcal{G} \cap \mathcal{G}_i, \quad (33)$$

where  $\mathcal{G}_i$  is the set of states satisfying the local discrete maximum principle

$$\min_{j \in \tilde{\mathcal{N}}_i} u_j =: u_i^{\min} \leq \bar{u} \leq u_i^{\max} := \max_{j \in \tilde{\mathcal{N}}_i} u_j. \quad (34)$$

Note that we define the bounds  $u_i^{\min}$  and  $u_i^{\max}$  using the subcell stencil  $\tilde{\mathcal{N}}_i = \bigcup_{e \in \mathcal{E}_i} \tilde{\mathcal{N}}_i^e$  rather than the full element stencil  $\mathcal{N}_i$  of node  $i$ , unless mentioned otherwise. The pros and cons of using tight bounds are explained in [42] in the context of flux-corrected transport (FCT) algorithms.

A locally bound-preserving IDP approximation to a given target flux  $f_{ij}^e$  is given by [34]

$$f_{ij}^{e,*} = \begin{cases} \min \left\{ f_{ij}^e, \min \{ 2\tilde{d}_{ij}^e u_i^{\max} - \bar{w}_{ij}^e, \bar{w}_{ji}^e - 2\tilde{d}_{ij}^e u_j^{\min} \} \right\} & \text{if } f_{ij}^e > 0, \\ \max \left\{ f_{ij}^e, \max \{ 2\tilde{d}_{ij}^e u_i^{\min} - \bar{w}_{ij}^e, \bar{w}_{ji}^e - 2\tilde{d}_{ij}^e u_j^{\max} \} \right\} & \text{otherwise,} \end{cases} \quad (35)$$

where  $\bar{w}_{ij}^e = 2\tilde{d}_{ij}^e \frac{u_j + u_i}{2} - \tilde{\mathbf{c}}_{ij}^e \cdot (\mathbf{f}_j - \mathbf{f}_i)$ . In infinite-precision arithmetic, this product has the same value as  $2\tilde{d}_{ij}^e \bar{u}_{ij}^e$ , where  $\bar{u}_{ij}^e$  is the bar state defined by (17). In numerical implementations, we calculate  $\bar{w}_{ij}^e$  directly to avoid rounding errors due to division and multiplication by  $\tilde{d}_{ij}^e$ .

**Remark 5.** Guermond and Popov [19] proved the validity of a local entropy inequality for (17) using the fact that (see Theorem 4.7 in [19])

$$E(\bar{u}_{ij}^e) \leq \frac{E(u_i) + E(u_j)}{2} - \frac{\tilde{\mathbf{c}}_{ij}^e \cdot (\mathbf{F}(u_j) - \mathbf{F}(u_i))}{2\tilde{d}_{ij}^e}$$

for any entropy pair  $(E, \mathbf{F})$ . Our monolithic convex limiting strategy makes it possible to enforce such inequality constraints for  $E(\bar{u}_{ij}^{e,*})$  by reducing the magnitude of  $f_{ij}^{e,*}$  if necessary. That is, the set  $\mathcal{G}_i$  may be redefined so as to enforce local entropy conditions in addition to local maximum principles.

**Remark 6.** Since the bar states  $\bar{u}_{ij}$  of the low-order method for the linear advection equation with the flux function  $\mathbf{f}(\mathbf{x}, u) = \mathbf{v}(\mathbf{x})u$  may fail to satisfy (23), the generalized version

$$f_{ij}^{e,*} = \begin{cases} \min \left\{ f_{ij}^e, \max \{ 0, \min \{ 2\tilde{d}_{ij}^e u_i^{\max} - \bar{w}_{ij}^e, \bar{w}_{ji}^e - 2\tilde{d}_{ij}^e u_j^{\min} \} \} \right\} & \text{if } f_{ij}^e > 0, \\ \max \left\{ f_{ij}^e, \min \{ 0, \max \{ 2\tilde{d}_{ij}^e u_i^{\min} - \bar{w}_{ij}^e, \bar{w}_{ji}^e - 2\tilde{d}_{ij}^e u_j^{\max} \} \} \right\} & \text{otherwise} \end{cases} \quad (36)$$

of formula (35) should be used to ensure positivity preservation for such linear flux functions.

To correct possible errors in the approximation of boundary terms, we define

$$b_i^*(u_h, u_{\text{in}}) = \tilde{b}_i(u_h, u_{\text{in}}) + \sum_{e \in \mathcal{E}_i} g_i^{e,*} \quad (37)$$

using

$$g_i^{e,*} = \min \left\{ g_i^{e,\max}, \max \left\{ g_i^e, g_i^{e,\min} \right\} \right\}, \quad (38)$$

where the target  $g_i^e$  is defined by (27) and the bounds are given by

$$g_i^{e,\max} = (u_i^{\max} - u_i) \int_{\partial K^e \cap \Gamma} \varphi_i |\mathbf{f}'(u_h) \cdot \mathbf{n}| \, ds, \quad (39)$$

$$g_i^{e,\min} = (u_i^{\min} - u_i) \int_{\partial K^e \cap \Gamma} \varphi_i |\mathbf{f}'(u_h) \cdot \mathbf{n}| \, ds. \quad (40)$$

The semi-discrete version of the flux-corrected Galerkin scheme is given by

$$m_i \frac{du_i}{dt} = \sum_{e \in \mathcal{E}_i} \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} [\tilde{d}_{ij}^e(u_j - u_i) + f_{ij}^{e,*}] - \sum_{j \in \tilde{\mathcal{N}}_i} \tilde{\mathbf{c}}_{ij} \cdot \mathbf{f}_j + b_i^*(u_h, u_{\text{in}}). \quad (41)$$

The IDP property can be shown as before using the equivalent form

$$\begin{aligned} m_i \frac{du_i}{dt} &= \sum_{e \in \mathcal{E}_i} \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} 2\tilde{d}_{ij}^e(\bar{u}_{ij}^{e,*} - u_i) + b_i^*(u_h, u_{\text{in}}), \\ &= \sum_{e \in \mathcal{E}_i} \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} 2\tilde{d}_{ij}^e(\bar{u}_{ij}^{e,*} - u_i) + c_i(u_i^* - u_i) + \tilde{b}_i(u_h, u_{\text{in}}), \end{aligned}$$

where  $\bar{u}_{ij}^{e,*}$  is the flux-corrected bar state defined by (33),  $u_i^* \in \{u_i^{\min}, u_i^{\max}\}$  and

$$0 \leq c_i \leq \sum_{e \in \mathcal{E}_i} \int_{\partial K^e \cap \Gamma} \varphi_i |\mathbf{f}'(u_h) \cdot \mathbf{n}| ds \quad (42)$$

by definition of  $g_i^{e,*}$ . We remark that the representation of the flux-corrected scheme in terms of  $\bar{u}_{ij}^{e,*}$  and  $u_i^*$  is used for theoretical analysis only. Practical implementations should be based on (41).

**Remark 7.** In contrast to the element-based algorithms proposed in [3, 22, 23, 42], the above limiting strategy rules out direct mass exchange between nodes that are not nearest neighbors.

**Remark 8.** To avoid strong peak clipping effects and achieve optimal convergence rates for  $p > 1$ , the discrete maximum principle (34) needs to be replaced with less restrictive constraints in a neighborhood of smooth local extrema [42]. To that end, a subset  $\mathcal{G}_i$  of the invariant set  $\mathcal{G}$  can be defined, e.g., using the smoothness criteria presented in [11, 12, 17, 37, 42]. We explore this possibility further in §7.

## 5. Computation of subcell antidiffusive fluxes

Clearly, the accuracy of the flux-corrected Galerkin discretization (41) depends on the definition of the subcell fluxes  $f_{ij}^e$ ,  $j \in \tilde{\mathcal{N}}_i^e$  which we have left unspecified so far. The antidiffusive element contributions defined by (26) can be written as

$$f_i^e = \sum_{j \in \tilde{\mathcal{N}}_i^e \setminus \{i\}} \tilde{d}_{ij}^e(u_i - u_j) + q_i^e, \quad (43)$$

where

$$\begin{aligned} q_i^e &= \sum_{j \in \mathcal{N}^e \setminus \{i\}} m_{ij}^e(\dot{u}_i - \dot{u}_j) + \sum_{j \in \mathcal{N}^e} (\tilde{\mathbf{c}}_{ij}^e - \mathbf{c}_{ij}^e) \cdot \mathbf{f}_j - \sum_{j \in \mathcal{N}^e} \mathbf{c}_{ji}^e \cdot \mathbf{f}_j \\ &+ \int_{K^e} \nabla \varphi_i \cdot \mathbf{f}(u_h) d\mathbf{x} + \int_{\partial K^e \cap \Gamma_-} \varphi_i(u_h - u_i) \mathbf{f}'(u_h) \cdot \mathbf{n} ds \end{aligned} \quad (44)$$

is the vector of element contributions that require further decomposition into subcell fluxes.

The zero-sum property  $\sum_{i=1}^{N_h} q_i^e = 0$  of the element contributions  $q_i^e$  implies the existence of a (generally non-unique) representation in the flux form

$$q_i^e = \sum_{\substack{j=1 \\ i \neq j}}^{N_h} q_{ij}^e, \quad q_{ji}^e = -q_{ij}^e. \quad (45)$$

Let the auxiliary vector  $v^e \in \mathbb{R}^N$  be defined as a solution of the linear system

$$(\hat{M}_L^e - \hat{M}_C^e) \hat{v}^e = \hat{q}^e, \quad (46)$$

where  $\hat{q}_{i_e}^e := q_i^e$  for  $i \in \mathcal{N}^e$ . The sparse  $N \times N$  mass matrices

$$\hat{M}_C^e = \left\{ \int_{K^e} \psi_i^e \psi_j^e \, d\mathbf{x} \right\}_{i,j=1}^N, \quad \hat{M}_L^e = \left\{ \delta_{ij} \int_{K^e} \psi_i^e \, d\mathbf{x} \right\}_{i,j=1}^N$$

are defined using the local basis functions  $\psi_i^e$  of the piecewise  $\mathbb{P}_1/\mathbb{Q}_1$  Bézier net approximation on the macroelement  $K^e$ . The subcell fluxes defined by

$$q_{ij}^e = \hat{m}_{i_e j_e}^e (\hat{v}_{i_e}^e - \hat{v}_{j_e}^e) \quad (47)$$

satisfy (45) and vanish if nodes  $i$  and  $j$  are not nearest neighbors. The matrix  $\hat{M}_L^e - \hat{M}_C^e$  is symmetric with vanishing row sums. Hence, the solution  $\hat{v}^e$  of the auxiliary problem (46) is defined up to a constant. Since our definition of  $q_{ij}^e$  is independent of this constant, it can be chosen arbitrarily. In our implementation, we solve (46) subject to the linear equality constraint

$$\sum_{i=1}^N \hat{v}_i^e = 0.$$

In summary, the original Galerkin discretization (5) can be recovered using

$$f_{ij}^e = \tilde{d}_{ij}^e (u_i - u_j) + q_{ij}^e. \quad (48)$$

In contrast to algebraic flux correction schemes for  $\mathbb{P}_1$  and  $\mathbb{Q}_1$  discretizations of general conservation laws [17, 34], the error associated with the group finite element approximation (9) cannot be neglected in high-order versions. Our definition of the target fluxes  $f_{ij}^e$  corrects this error even for  $p = 1$ .

**Remark 9.** If the coefficients  $\tilde{d}_{ij}$  of the graph Laplacian operator are defined using the assembled global matrix  $\tilde{\mathbf{C}}$ , the corresponding fluxes  $f_{ij}$  should be calculated using the formula

$$f_{ij} = \tilde{d}_{ij} (u_i - u_j) + \sum_{e \in \mathcal{E}_i} q_{ij}^e \quad (49)$$

and limited using the low-order bar states  $\bar{u}_{ij} = \frac{u_j + u_i}{2} - \frac{\tilde{\mathbf{c}}_{ij} \cdot (\mathbf{f}_j - \mathbf{f}_i)}{2\tilde{d}_{ij}}$  of the global system.

**Remark 10.** The 1D version of the compact-stencil FCT limiter introduced in [42] is also based on a decomposition of generic element contributions into (uniquely defined) subcell fluxes. However, the multidimensional subcell decomposition proposed in Section 4.5 of [42] requires the computationally intensive solution of minimization problems and has not been tested in practice so far.

## 6. Stabilization of subcell antidiffusive fluxes

The continuous Galerkin method exhibits suboptimal  $\mathcal{O}(h^p)$  convergence behavior even for smooth solutions of linear advection problems on general meshes. To achieve optimal accuracy and prevent formation of spurious ripples within the local bounds of the limiting procedures, some high-order stabilization should be included in the target flux. In the numerical studies of Lohmann et al. [42], optimal convergence rates for high-order finite element discretizations of the linear advection equation were achieved using two-level Laplacian stabilization which can be added to the vector  $q^e$  before decomposing it into subcell fluxes  $q_{ij}^e$  in the manner described in Section 5. For nonlinear conservation laws, Guermond et al. [17, 18] recommend the use of entropy viscosity (EV) stabilization. Its ability to preserve the optimal order for  $p > 1$  is yet to be verified. The same is true for stabilization via low-order approximations to the nodal time derivatives  $\dot{u}_i$ , as proposed in [34] for  $p = 1$ .

The selection of genuinely high-order stabilization tools for Bernstein finite element approximations is beyond the scope of this work. In the numerical experiments of §8, we replace (48) with

$$f_{ij}^{e,\text{stab}} = (1 - C_E \max(R_i, R_j)) \tilde{d}_{ij}^e (u_i - u_j) + q_{ij}^e, \quad (50)$$

where  $R_i \in [0, 1]$  is a nodal sensor that determines the appropriate amount of nonlinear stabilization and  $C_E = \mathcal{O}(1)$  is a user-defined parameter (we use  $C_E = 1$ ).

Following Guermond et al. [17], we choose an entropy pair  $(E(u), \mathbf{F}(u))$  for (1a) and use

$$R_i = \frac{\left| \sum_{j \in \mathcal{N}_i} [\mathbf{F}(u_j) - E'(u_i) \mathbf{f}(u_j)] \cdot \mathbf{c}_{ij} \right|}{\left| \sum_{j \in \mathcal{N}_i} \mathbf{F}(u_j) \cdot \mathbf{c}_{ij} \right| + |E'(u_i)| \left| \sum_{j \in \mathcal{N}_i} \mathbf{f}(u_j) \cdot \mathbf{c}_{ij} \right| + \epsilon}, \quad (51)$$

where  $\epsilon$  is a positive constant which prevents division by zero (we use  $\epsilon = 10^{-10}$ ). The so-defined  $R_i$  measures the rate of entropy production at node  $i$ . Note that we use the coefficients  $\mathbf{c}_{ij}$  of the discrete gradient operator corresponding to the high-order space in (51). This definition of  $R_i$  extends the domain of dependence to the full stencil  $\mathcal{N}_i$  of node  $i$  to improve robustness. However, the stabilized subcell fluxes (50) preserve the compact stencil property of the nonlinear AFC scheme.

For all test problems in §8, we use  $E(u) = \frac{1}{2}u^2$  and  $\mathbf{F}(u) = \int_0^u E'(z) \mathbf{f}'(z) dz$ . For a detailed discussion of entropy viscosity stabilization, we refer the reader to Guermond et al. [17, 18].

## 7. Extremum-preserving flux limiting

As mentioned in Remark 8, the local discrete maximum principle (34) may need to be relaxed to achieve high-order convergence and alleviate peak clipping at smooth local extrema. In this work,

we use one of the subcell smoothness indicators introduced by Hajduk et al. [23]. The underlying smoothness criterion is based on variations of the approximate nodal Laplacians

$$\tilde{\eta}_i = (\Delta_h \tilde{u}_h)_i := \frac{1}{\tilde{m}_i} \int_{\Omega} \nabla \tilde{u}_h \cdot \nabla \psi_i \, d\mathbf{x} \quad (52)$$

calculated using the piecewise  $\mathbb{P}_1/\mathbb{Q}_1$  basis functions  $\psi_1, \dots, \psi_{N_h}$ , the diagonal entries  $\tilde{m}_i := \int_{\Omega} \psi_i \, d\mathbf{x}$  of the corresponding lumped mass matrix, and the  $\mathbb{P}_1/\mathbb{Q}_1$  interpolant

$$\tilde{u}_h(\mathbf{x}) = \sum_{j=1}^{N_h} u_h(\mathbf{x}_j) \psi_j(\mathbf{x}) \quad (53)$$

of  $u_h(\mathbf{x}_j) = \sum_{k=1}^{N_h} \varphi_k(\mathbf{x}_j) u_k$ , where  $u_1, \dots, u_{N_h}$  are the Bernstein degrees of freedom. Given the Laplacian reconstruction (52), we calculate the nodal smoothness sensors [23]

$$\gamma_i = \begin{cases} \min \left\{ 1, \frac{C \max\{0, \eta_i^{\min} \eta_i^{\max}\} + \epsilon}{\max\{(\eta_i^{\min})^2, (\eta_i^{\max})^2\} + \epsilon} \right\} & \text{if } \mathbf{x}_i \in \Omega, \\ 1 & \text{if } \mathbf{x}_i \in \Gamma, \end{cases} \quad (54)$$

where  $\epsilon > 0$  is again a small positive number and  $C \geq 1$  is a sensitivity parameter. The maximum  $\eta_i^{\max} = \max_{j \in \tilde{\mathcal{N}}_i} \eta_j$  and minimum  $\eta_i^{\min} = \min_{j \in \tilde{\mathcal{N}}_i} \eta_j$  are taken over the set  $\tilde{\mathcal{N}}_i$  of nodes that share a subcell with node  $i$ . Formula (54) produces  $\gamma_i = 0$  if the signs of  $\eta_i^{\max}$  and  $\eta_i^{\min}$  differ. The maximal value  $\gamma_i = 1$  is attained if the signs of the two extremal values are the same and their magnitudes do not differ by more than a factor of  $C$ . In the numerical studies below, we use  $C = 3$ .

To prevent unnecessary flux limiting at smooth peaks, we modify formula (36) as follows:

$$f_{ij}^{e,*} = \begin{cases} \min \{ f_{ij}^e, \min \{ \gamma_i f_{ij}^e + (1 - \gamma_i) \max \{ 0, 2\tilde{d}_{ij}^e u_i^{\max} - \bar{w}_{ij}^e \}, \\ \gamma_j f_{ij}^e + (1 - \gamma_j) \max \{ 0, \bar{w}_{ji}^e - 2\tilde{d}_{ij}^e u_j^{\min} \} \} \} & \text{if } f_{ij}^e > 0, \\ \max \{ f_{ij}^e, \max \{ \gamma_i f_{ij}^e + (1 - \gamma_i) \min \{ 0, 2\tilde{d}_{ij}^e u_i^{\min} - \bar{w}_{ij}^e \}, \\ \gamma_j f_{ij}^e + (1 - \gamma_j) \min \{ 0, \bar{w}_{ji}^e - 2\tilde{d}_{ij}^e u_j^{\max} \} \} \} & \text{otherwise.} \end{cases} \quad (55)$$

This modification relaxes the bounds of the flux constraints associated with nodes  $i$  and  $j$  using the corresponding nodal smoothness indicators. The IDP property w.r.t. the invariant set  $\mathcal{G} = [u^{\min}, u^{\max}]$  can be enforced by using the relaxed bounds  $\gamma_i u^{\max} + (1 - \gamma_i) u_i^{\max}$  and  $\gamma_i u^{\min} + (1 - \gamma_i) u_i^{\min}$  in the limiting formula (36) instead of replacing it with (55), see [23] for details.

## 8. Numerical examples

In this section, we apply the subcell flux limiting procedure to (stabilized) Galerkin discretizations of scalar test problems. The main purpose of this numerical study is to show that the proposed low-order scheme and subcell flux decomposition are well suited for algebraic flux correction purposes. More detailed studies of stabilization approaches and smoothness indicators will be presented elsewhere.

All computations are performed using PROTEUS (<https://proteustoolkit.org>), an open-source Python toolkit for numerical simulations. We consider the following low-order methods:

- LO {full stencil}. In this version, we do not apply the mass lumping operator  $P^e$  to the element matrices  $C_k^e$  of the discrete gradient operator for Bernstein elements of degree  $p = 1, 2$ . The element matrix  $\tilde{D}^e$  of the resulting discrete diffusion operator has  $N^2$  nonvanishing entries.
- LO {compact stencil}. This is the low-order method defined by (15). In this section, it is used for  $p = 2$  only. The element matrix  $\tilde{D}^e$  of the discrete diffusion operator has the compact sparsity pattern of the piecewise  $\mathbb{Q}_1$  discretization on the 4-element submesh depicted in Fig. 1.

The high-order methods under investigation are abbreviated as follows:

- HO {Galerkin, L}. No stabilization of the Galerkin target (26), limiting via (35) or (36) for the nonlinear and the linear problems, respectively.
- HO {EV}. Stabilized EV target (50), no limiting.
- HO {EV, L}. Stabilized EV target (50), limiting via (35) or (36) for the nonlinear and the linear problems, respectively.
- HO {EV, L, SI}. Stabilized EV target (50), limiting using the smoothness indicator (54).

In the rest of this section, we proceed as follows. We first consider linear advection problems which we solve using the full and compact stencil versions of LO, as well as different versions of HO. The objective is to assess the quality of the low-order method and to study the convergence behavior of the high-order method in situations when the exact solution is smooth. Thereafter, we solve two nonlinear problems using LO {compact stencil}, HO {Galerkin, L}, and HO {EV, L}. The results of these numerical experiments illustrate the IDP property of the low-order method and the importance of using high-order stabilization for the target fluxes.

## 8.1. Linear advection

### 8.1.1. One-dimensional advection

The first linear problem that we consider in this study is the one-dimensional advection equation

$$\frac{\partial u}{\partial t} + v \frac{\partial u}{\partial x} = 0 \quad \text{in } \Omega = (0, 1) \quad (56)$$

with the constant velocity  $v = 1$ . The smooth initial condition is given by

$$u_0(x) = \exp[-100(x - 0.25)^2]. \quad (57)$$

We solve this problem up to the final time  $T = 0.5$  and measure numerical errors w.r.t. the  $L^1$  norm.

The grid convergence history for the low-order methods under investigation are reported in Table 1. The experimental orders of convergence (EOC) for pairs of uniform 1D meshes are calculated using the formula presented in [42]. We observe that the accuracy of the full stencil version deteriorates significantly as we switch from the subcell  $\mathbb{Q}_1$  discretization to the  $\mathbb{Q}_2$  approximation with the same number of DoFs. The compact-stencil  $\mathbb{Q}_2$  scheme produces more accurate results than its full-stencil

counterpart. The numerical studies presented in [42] indicate that more dramatic improvements can be expected for high-order Bernstein elements. At least for constant velocities, the convergence behavior of the compact-stencil version is largely independent of  $p$ , as shown in [42].

In Table 2, we present the results of grid convergence studies for the high-order stabilized  $\mathbb{Q}_2$  approximations. In the limited versions of the HO {EV} method, we use the full stencil bounds  $u_i^{\max} = \max_{j \in \mathcal{N}_i} u_j$  and  $u_i^{\min} = \min_{j \in \mathcal{N}_i} u_j$ . It can be seen that the SI relaxation based on (54) and (55) results in smaller global  $L^1$  errors and faster convergence on coarse meshes. However, the EOCs of flux-limited approximations are not as high as those of HO {EV} in this example.

$N_h$	LO {full stencil}, $\mathbb{Q}_1$		LO {full stencil}, $\mathbb{Q}_2$		LO {compact stencil}, $\mathbb{Q}_2$	
	$\ u_h - u_{\text{exact}}\ _{L^1}$	EOC	$\ u_h - u_{\text{exact}}\ _{L^1}$	EOC	$\ u_h - u_{\text{exact}}\ _{L^1}$	EOC
11	1.45E-2	–	1.61E-2	–	1.51E-2	–
15	1.25E-2	0.44	1.42E-2	0.39	1.31E-2	0.42
20	1.07E-2	0.49	1.24E-2	0.44	1.13E-2	0.47
28	8.79E-3	0.56	1.03E-2	0.50	9.34E-3	0.54
39	7.10E-3	0.62	8.51E-3	0.57	7.59E-3	0.60
54	5.65E-3	0.68	6.89E-3	0.63	6.08E-3	0.66
75	4.40E-3	0.74	5.46E-3	0.69	4.77E-3	0.72
105	3.36E-3	0.79	4.23E-3	0.75	3.66E-3	0.78
147	2.52E-3	0.84	3.22E-3	0.80	2.76E-3	0.82

Table 1: Linear advection in 1D, grid convergence history for the low-order methods.

$N_h$	HO {EV}, $\mathbb{Q}_2$		HO {EV,L}, $\mathbb{Q}_2$		HO {EV,L,SI}, $\mathbb{Q}_2$	
	$\ u_h - u_{\text{exact}}\ _{L^1}$	EOC	$\ u_h - u_{\text{exact}}\ _{L^1}$	EOC	$\ u_h - u_{\text{exact}}\ _{L^1}$	EOC
11	5.09E-3	–	7.77E-3	–	6.50E-3	–
15	3.05E-3	1.51	4.84E-3	1.40	3.63E-3	1.73
20	1.69E-3	1.94	2.66E-3	1.96	1.76E-3	2.37
28	7.39E-4	2.35	1.20E-3	2.27	7.89E-4	2.27
39	2.99E-4	2.64	6.33E-4	1.85	3.43E-4	2.43
54	1.25E-4	2.62	3.20E-4	2.05	1.47E-4	2.55
75	4.97E-5	2.76	1.54E-4	2.18	6.24E-5	2.56
105	1.87E-5	2.87	7.04E-5	2.30	2.63E-5	2.54
147	6.96E-6	2.91	3.35E-5	2.18	1.14E-5	2.46

Table 2: Linear advection in 1D, grid convergence history for the high-order methods.

**Remark 11.** To avoid errors due to inaccurate initialization, we  $L^2$ -project the smooth initial data of this test problem into the  $\mathbb{Q}_2$  finite element space by solving a linear system with the consistent mass matrix. For all other test problems, we define the Bernstein coefficients  $u_i(0) = u_0(\mathbf{x}_i)$  using the (generally inaccurate but bound-preserving, cf. [46]) interpolation at the control points  $\mathbf{x}_i$ .

### 8.1.2. Solid body rotation

To facilitate a direct comparison with the  $\mathbb{P}_1/\mathbb{Q}_1$  version of algebraic flux correction schemes and variational approaches to shock capturing, let us now consider the solid body rotation benchmark



[28, 34, 35, 39]. In this 2D experiment, we solve the unsteady linear advection equation

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

using the divergence-free velocity field  $\mathbf{v}(x, y) = 2\pi(0.5 - y, x - 0.5)^\top$  to rotate a slotted cylinder, a sharp cone, and a smooth hump around the center  $(0.5, 0.5)$  of the domain  $\Omega$ . Homogeneous boundary conditions are prescribed on  $\Gamma_-$ . The initial condition, as defined by LeVeque [39], is given by

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

where

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

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

After each complete revolution, the exact solution coincides with the initial condition.

In Figure 2, we show the low-order  $\mathbb{Q}_1$  and  $\mathbb{Q}_2$  approximations at the final time  $T = 1$  (one full rotation). The diagrams of the first and second row were obtained using  $N_h = 129^2$  and  $N_h = 257^2$  DoFs, respectively. For a better quantitative comparison, the  $L^1$  errors  $E_1 = \|u_h - u_{\text{exact}}\|_{L^1}$  and the global maxima  $u^{\text{max}} = \max_{i=1, \dots, N_h} u_i$  of the Bernstein coefficients are listed above each plot. As expected, the approximation calculated with the full stencil  $\mathbb{Q}_2$  scheme proves more dissipative than the compact-stencil  $\mathbb{Q}_1$  and  $\mathbb{Q}_2$  approximations. In contrast to the subcell upwinding strategy employed in [22, 42], the low-order scheme defined by (15) preserves the  $\mathbb{Q}_1$  sparsity pattern exactly even for nonuniform velocity fields and nonlinear flux functions. This remarkable property eliminates a major bottleneck to achieving high performance and  $p$ -independent convergence behavior with matrix-based algebraic flux correction schemes. In our numerical experiment, the low-order  $\mathbb{Q}_2$  solution obtained with (15) is as accurate as the subcell  $\mathbb{Q}_1$  approximation with the same number of DoFs.

The results obtained with the high-order extensions and zooms of the flux-limited solutions are shown in Figs 3 and 4, respectively. The activation of subcell flux correction eliminates small under-shoots and overshoots at the edges of the slotted cylinder but smears the bound-preserving peaks of the hump and cone significantly. The stabilization of the target flux via entropy viscosity increases the  $L^1$  error without having any positive impact on the quality of the flux-corrected  $\mathbb{Q}_2$  approximations in this particular example. The Laplacian-based smoothness indicator  $\gamma_i$  defined by (54) was used to relax the bounds in formula (55). The SI version recognizes the top of the hump as a smooth extremum and resolves it very well even on the coarser mesh. Flux limiting at the top of the cone is

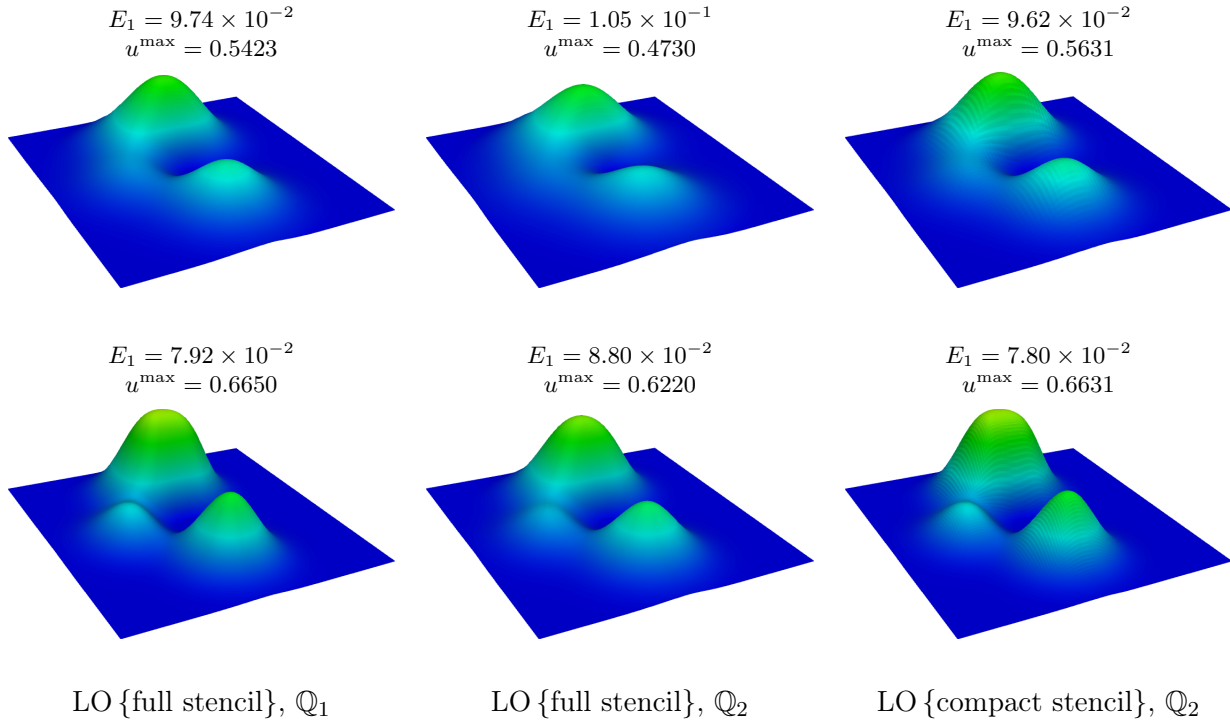


Figure 2: Solid body rotation [39]. Low-order solutions after one full rotation (final time  $T = 1$ ). The total number of DoFs is  $N_h = 129^2$  in the diagrams of the first row and  $N_h = 257^2$  in the diagrams of the second row.

deactivated as soon as the peak becomes rounded enough for (54) to produce  $\gamma_i^e = 1$ . At the same time, no violation of discrete maximum principles occurs in the neighborhood of discontinuities, where the second derivatives exhibit abrupt changes and (54) produces  $\gamma_i^e = 0$ .

### 8.1.3. Steady circular advection

In contrast to the FCT algorithms employed in [3, 17, 21, 22, 42], the monolithic convex limiting strategy is well suited for calculating steady-state solutions. To show this, we solve

$$\nabla \cdot (\mathbf{v}u) = 0 \quad \text{in } \Omega = (0, 1)^2 \quad (58)$$

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

$$u(x, y) = \begin{cases} 1, & \text{if } 0.15 \leq r(x, y) \leq 0.45, \\ \cos^2 \left( 10\pi \frac{r(x, y) - 0.7}{3} \right), & \text{if } 0.55 \leq r(x, y) \leq 0.85, \\ 0, & \text{otherwise,} \end{cases} \quad (59)$$

where  $r(x, y) = \sqrt{x^2 + y^2}$  denotes the distance to the corner point  $(0, 0)$ . The stationary  $Q_2$  solutions obtained with  $N_h = 65^2$  and  $N_h = 129^2$  are shown in Fig. 5. These numerical solutions were marched

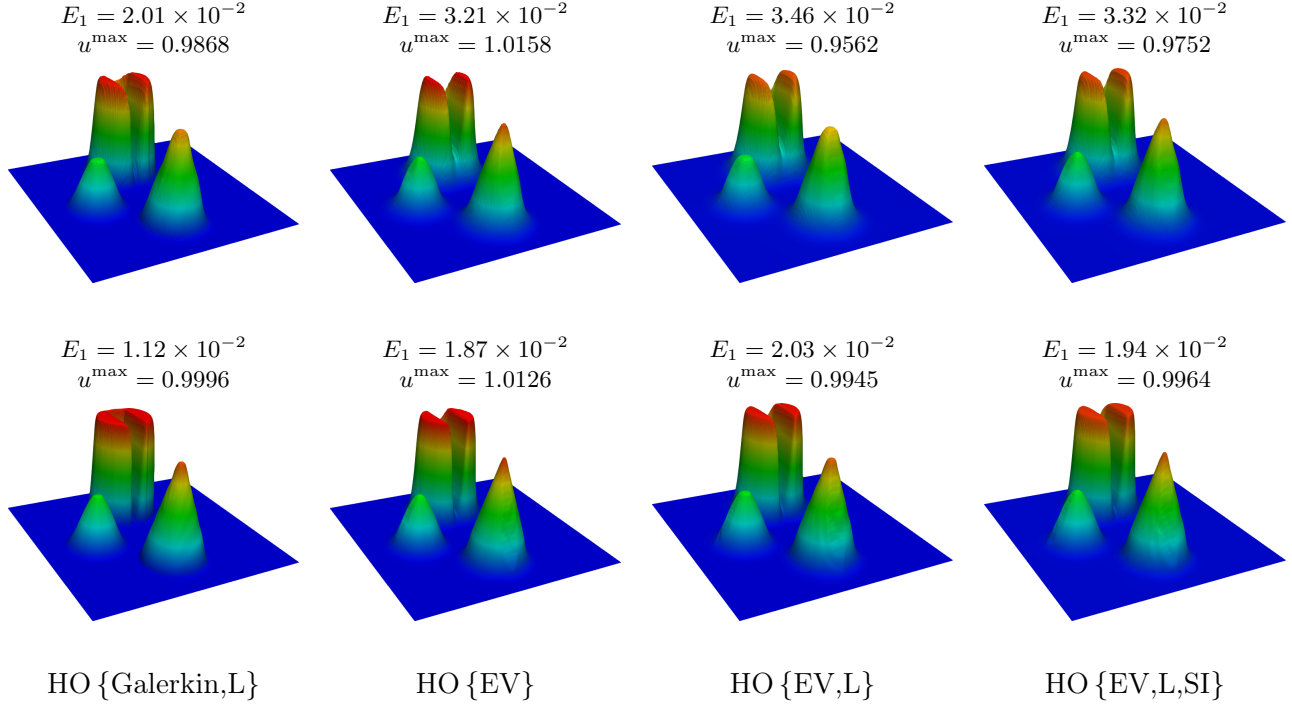


Figure 3: Solid body rotation problem [39]. High-order solutions after one full rotation (final time  $T = 1$ ). The total number of DoFs is  $N_h = 129^2$  in the diagrams of the first row and  $N_h = 257^2$  in the diagrams of the second row.

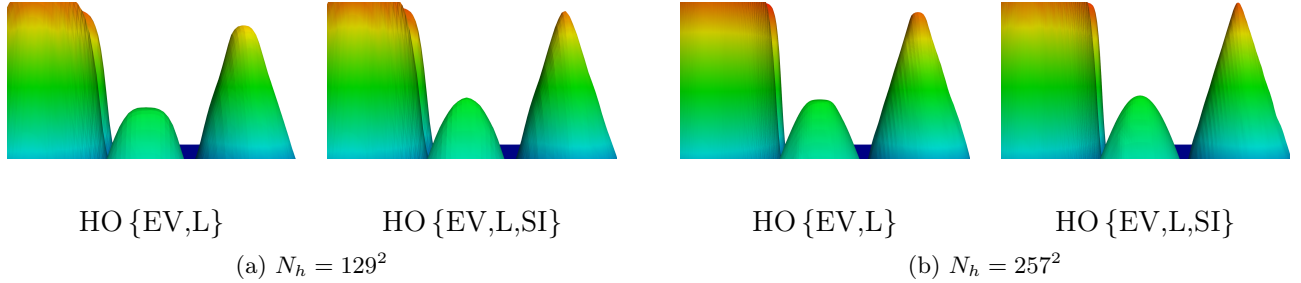


Figure 4: Solid body rotation problem [39]. Zooms of the limited high-order  $\mathbb{Q}_2$  solutions at  $T = 1$  obtained without and with using the smoothness indicator defined by (54) to reduce peak clipping effects.

to the steady state by solving the lumped-mass version of the  $\mathbb{Q}_2$  approximation to the time-dependent advection problem until the prescribed tolerance was reached for the steady-state residuals.

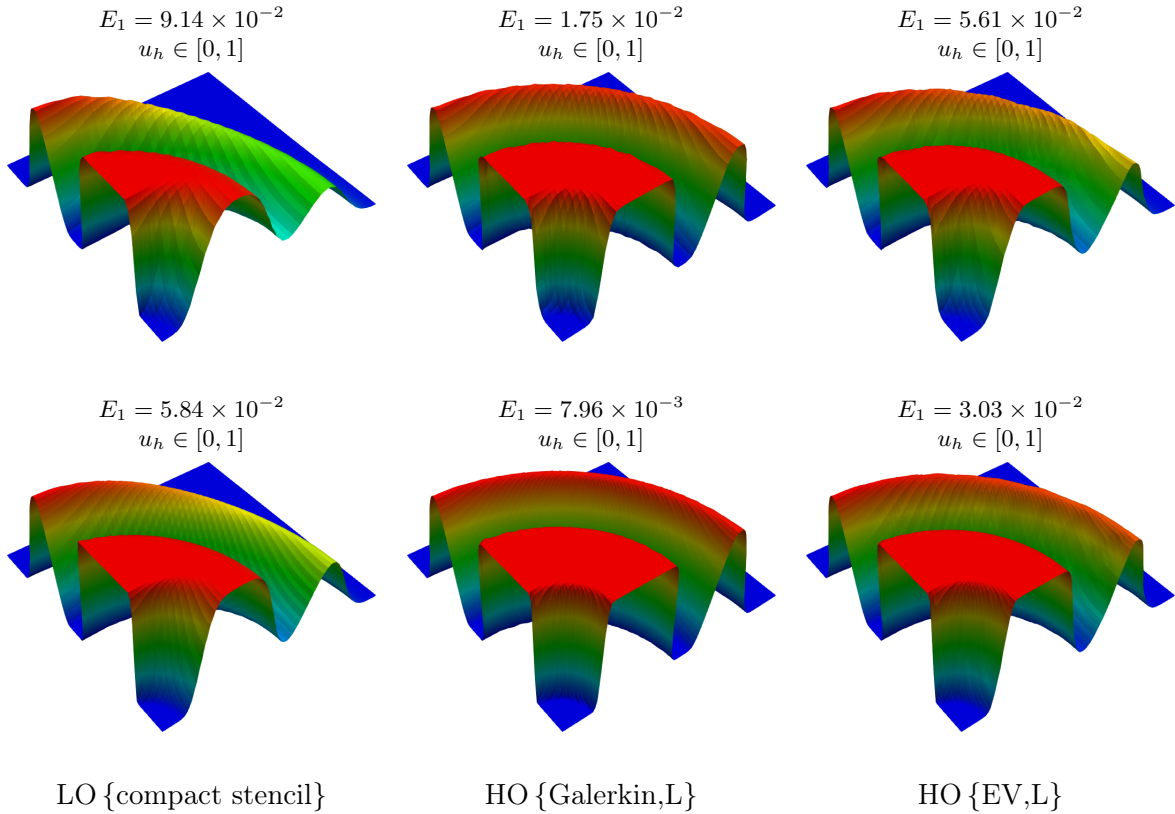


Figure 5: Steady circular advection. Stationary  $\mathbb{Q}_2$  solutions calculated using time marching. The total number of DoFs is  $N_h = 65^2$  in the diagrams of the first row and  $N_h = 129^2$  in the diagrams of the second row.

### 8.2. Burgers equation

As a first nonlinear test problem, we consider the 2D inviscid Burgers equation [16, 34]

$$\frac{\partial u}{\partial t} + \nabla \cdot \left( \mathbf{v} \frac{u^2}{2} \right) = 0 \quad \text{in } \Omega = (0, 1)^2, \quad (60)$$

where  $\mathbf{v} = (1, 1)^\top$  is a constant vector. The piecewise-constant initial data is given by

$$u_0(x, y) = \begin{cases} -0.2 & \text{if } x < 0.5 \wedge y > 0.5, \\ -1.0 & \text{if } x > 0.5 \wedge y > 0.5, \\ 0.5 & \text{if } x < 0.5 \wedge y < 0.5, \\ 0.8 & \text{if } x > 0.5 \wedge y < 0.5. \end{cases} \quad (61)$$

The inflow boundary conditions are defined using the exact solution of the pure initial value problem in  $\mathbb{R}^2$ . This solution can be found in [16] and stays in the invariant set  $\mathcal{G} = [-1.0, 0.8]$ .

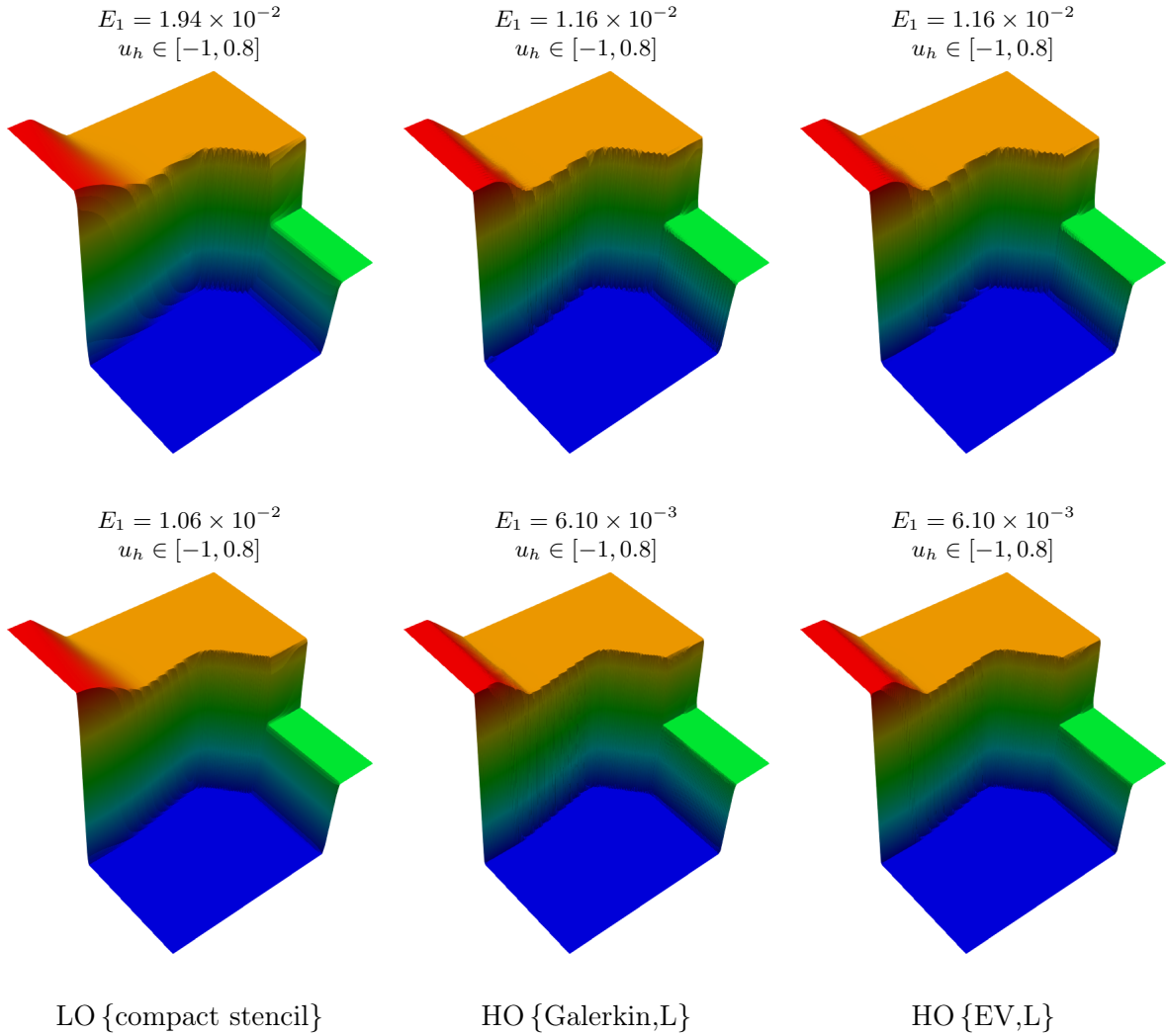


Figure 6: Burgers equation, bound-preserving  $\mathbb{Q}_2$  approximations at  $T = 0.5$ . The total number of DoFs is  $N_h = 129^2$  in the diagrams of the first row and  $N_h = 257^2$  in the diagrams of the second row.

The numerical solutions obtained at  $T = 0.5$  using  $\mathbb{Q}_2$  elements with  $N_h = 129^2$  and  $N_h = 257^2$  DoFs are shown in Fig. 6. The presented  $L^1$  errors indicate that considerable amounts of numerical diffusion can be safely removed in the process of subcell flux correction. The use of EV stabilization has no significant impact on the accuracy of the flux-corrected HO solutions in this example.

### 8.3. KPP problem

The KPP problem [18, 19, 20, 33] is a more challenging nonlinear test. In this final 2D experiment, we solve the scalar conservation law (1a) with the nonlinear and nonconvex flux function

$$\mathbf{f}(u) = (\sin(u), \cos(u)) \quad (62)$$

in the domain  $\Omega = (-2, 2) \times (-2.5, 1.5)$  using the initial condition

$$u_0(x, y) = \begin{cases} \frac{14\pi}{4} & \text{if } \sqrt{x^2 + y^2} \leq 1, \\ \frac{\pi}{4} & \text{otherwise.} \end{cases} \quad (63)$$

A simple (but rather pessimistic) upper bound for the maximum speed is  $\lambda = 1$ . More accurate GMS bounds can be found in [20]. The exact solution exhibits a two-dimensional rotating wave structure. The main challenge of this test is to prevent possible convergence to wrong weak solutions.

The numerical solutions obtained at  $T = 1$  using  $N_h = 257^2$  DoFs are displayed in Fig. 7. The plot shown in the middle demonstrates that the lack of nonlinear stabilization in the target flux of the AFC scheme may, indeed, cause convergence to an entropy-violating solution. This example confirms the findings of Guermond et al. [18] who observed such nonphysical behavior of flux-limited Galerkin methods in the context of predictor-corrector FCT algorithms. The use of entropy viscosity stabilization in the EV target of the monolithic AFC discretization cures this drawback without introducing inordinately large amounts of numerical dissipation (compare the well-resolved solution on the right of Fig. 7 to the diffusive and distorted approximations shown in the other two diagrams).

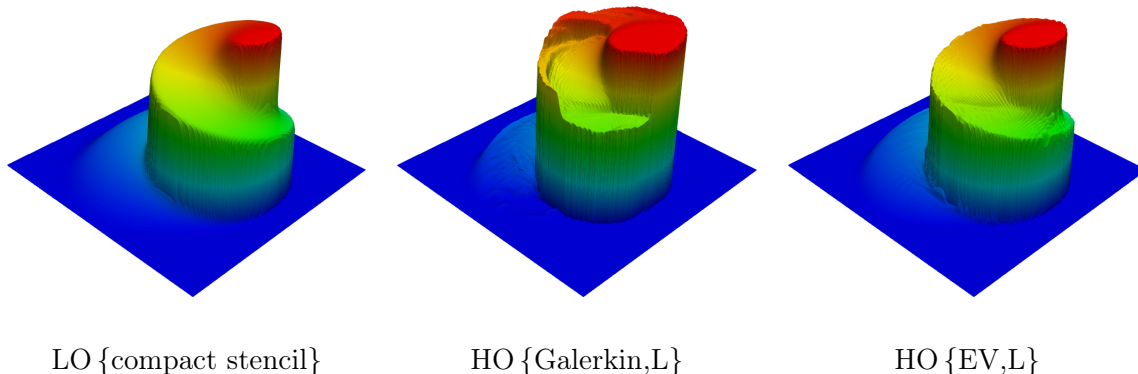


Figure 7: KPP problem [33], bound-preserving  $\mathbb{Q}_2$  approximations at  $T = 1$ . The total number of DoFs is  $N_h = 257^2$ .

## 9. Conclusions

The main result of this work is the development of a novel subcell flux correction procedure for high-order finite elements. The proposed definitions of the low-order scheme and of the antidiffusive fluxes lead to compact-stencil approximations which can be implemented efficiently. The monolithic

convex limiting strategy ensures well-posedness of nonlinear discrete problems and opens new avenues for theoretical analysis of high-order AFC schemes. Since the  $\mathbb{P}_1$  and  $\mathbb{Q}_1$  versions of the presented methodology have already been successfully applied to the Euler equations of gas dynamics in [34], it is hoped that extensions of subcell flux limiting to high-order Bernstein finite element discretizations of nonlinear hyperbolic systems will be relatively straightforward.

**Acknowledgments.** The work of Dmitri Kuzmin was supported by the German Research Association (DFG) under grant KU 1530/23-1. The work of Manuel Quezada de Luna was supported by King Abdullah University of Science and Technology (KAUST) in Thuwal, Saudi Arabia. The authors would like to thank Prof. David I. Ketcheson (KAUST) and Christoph Lohmann (TU Dortmund University) for helpful discussions.

## References

- [1] R. Abgrall and J. Treflík, An example of high order residual distribution scheme using non-Lagrange elements. *J. Sci. Comput.* **45** (2010) 3–25.
- [2] M. Ainsworth, G. Andriamaro, and O. Davydov, Bernstein-Bézier finite elements of arbitrary order and optimal assembly procedures. *SIAM J. Sci. Comput.* **33** (2011) 3087–3109.
- [3] R. Anderson, V. Dobrev, Tz. Kolev, D. Kuzmin, M. Quezada de Luna, R. Rieben, and V. Tomov, High-order local maximum principle preserving (MPP) discontinuous Galerkin finite element method for the transport equation. *J. Comput. Phys.* **334** (2017) 102–124.
- [4] S. Badia and J. Bonilla, Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. *Computer Methods Appl. Mech. Engrg.* **313** (2017) 133–158.
- [5] G. Barrenechea, E. Burman, and F. Karakatsani, Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. *Numer. Math.* **135** (2017) 521–545.
- [6] G. Barrenechea, V. John, and P. Knobloch, Analysis of algebraic flux correction schemes. *SIAM J. Numer. Anal.* **54** (2016) 2427–2451.
- [7] G. Barrenechea, V. John, P. Knobloch, and R. Rankin, A unified analysis of algebraic flux correction schemes for convection-diffusion equations. *SeMA* **75** (2018) 655–685.
- [8] G. Barrenechea and P. Knobloch, Analysis of a group finite element formulation. *Applied Numerical Mathematics* **118** (2017) 238–248.
- [9] J.P. Boris and D.L. Book, Flux-Corrected Transport: I. SHASTA, a fluid transport algorithm that works. *J. Comput. Phys.* **11** (1973) 38–69.

- [10] C.J. Cotter and D. Kuzmin, Embedded discontinuous Galerkin transport schemes with localised limiters. *J. Comput. Phys.* **311** (2016) 363–373.
- [11] S. Diot, S. Clain, and R. Loubère, Improved detection criteria for the Multi-dimensional Optimal Order Detection (MOOD) on unstructured meshes with very high-order polynomials. *Computers & Fluids* **64** (2012) 43–63.
- [12] M. Dumbser, O. Zanotti, R. Loubère, and S. Diot, A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *J. Comput. Phys.* **278** (2014) 47–75.
- [13] C.A.J. Fletcher, The group finite element formulation, *Comput. Methods Appl. Mech. Engrg.* **37** (1983) 225–243.
- [14] C.A.J. Fletcher, A comparison of finite element and finite difference solutions of the one- and two-dimensional Burgers’ equations. *J. Comput. Phys.* **51** (1983) 159–188.
- [15] S. Gottlieb, C.-W. Shu, and E. Tadmor, Strong stability-preserving high-order time discretization methods. *SIAM Review* **43** (2001) 89–112.
- [16] J.-L. Guermond and M. Nazarov, A maximum-principle preserving  $C^0$  finite element method for scalar conservation equations. *Computer Methods Appl. Mech. Engrg.* **272** (2014) 198–213.
- [17] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas, Second-order invariant domain preserving approximation of the Euler equations using convex limiting. *SIAM J. Sci. Computing* **40** (2018) A3211-A3239.
- [18] J.-L. Guermond, M. Nazarov, B. Popov, and Y. Yang, A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. *SIAM J. Numer. Anal.* **52** (2014) 2163–2182.
- [19] J.-L. Guermond and B. Popov, Invariant domains and first-order continuous finite element approximation for hyperbolic systems. *SIAM J. Numer. Anal.* **54** (2016) 2466–2489.
- [20] J.-L. Guermond and B. Popov, Invariant domains and second-order continuous finite element approximation for scalar conservation equations. *SIAM J. Numer. Anal.* **55** (2017) 3120–3146.
- [21] J.-L. Guermond, B. Popov, and I. Tomas, Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems. *Computer Methods Appl. Mech. Engrg.* **347** (2019) 143–175.
- [22] H. Hajduk, D. Kuzmin, Tz. Kolev, and R. Abgrall, Matrix-free subcell residual distribution for Bernstein finite. element discretizations of linear advection problems. *Computer Methods Appl. Mech. Engrg.* **359** (2020) 112658.



- [23] H. Hajduk, D. Kuzmin, Tz. Kolev, V. Tomov, I. Tomas, J.N. Shadid, Matrix-free subcell residual distribution for Bernstein finite elements: Monolithic limiting. *Computers and Fluids*, Available online 22 January 2020, 104451, <https://doi.org/10.1016/j.compfluid.2020.104451>
- [24] A. Harten, High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.* **49** (1983) 357–393.
- [25] A. Harten, On a class of high resolution total-variation-stable finite-difference-schemes. *SIAM J. Numer. Anal.* **21** (1984) 1-23.
- [26] A. Jameson, Computational algorithms for aerodynamic analysis and design. *Appl. Numer. Math.* **13** (1993) 383–422.
- [27] A. Jameson, Positive schemes and shock modelling for compressible flows. *Int. J. Numer. Methods Fluids* **20** (1995) 743-776.
- [28] V. John and E. Schmeyer, On finite element methods for 3D time-dependent convection-diffusion-reaction equations with small diffusion. *Comput. Meth. Appl. Mech. Engrg.* **198** (2008) 475–494.
- [29] R.C. Kirby, Efficient discontinuous Galerkin finite element methods via Bernstein polynomials. [arXiv:1504.03990](https://arxiv.org/abs/1504.03990) [math.NA]
- [30] R.C. Kirby, Fast simplicial finite element algorithms using Bernstein polynomials. *Numer. Math.* **117** (2011) 631–652.
- [31] R.C. Kirby, Low-complexity finite element algorithms for the de Rham complex on simplices. *SIAM J. Sci. Comput.* **36** (2014) A846–A868.
- [32] R.C. Kirby, Fast inversion of the simplicial Bernstein mass matrix. *Numer. Math.* **135** (2017) 73–95.
- [33] A. Kurganov, G. Petrova, and B. Popov, Adaptive semidiscrete central-upwind schemes for non-convex hyperbolic conservation laws. *SIAM J. Sci. Comput.* **29** (2007) 2381–2401.
- [34] D. Kuzmin, Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws. *Comput. Methods Appl. Mech. Engrg.* **361** (2020) 112804.
- [35] D. Kuzmin, Algebraic flux correction I. Scalar conservation laws. In: D. Kuzmin, R. Löhner and S. Turek (eds.) *Flux-Corrected Transport: Principles, Algorithms, and Applications*. Springer, 2nd edition: 145–192 (2012).
- [36] D. Kuzmin, M. Möller, and M. Gurrus, Algebraic flux correction II. Compressible flow problems. In: D. Kuzmin, R. Löhner, S. Turek (eds), *Flux-Corrected Transport: Principles, Algorithms, and Applications*. Springer, 2nd edition, 2012, pp. 193–238.

- [37] D. Kuzmin, M. Quezada de Luna, and C. Kees, A partition of unity approach to adaptivity and limiting in continuous finite element methods. *Computers & Mathematics with Applications* **78** (2019) 944–957.
- [38] D. Kuzmin and S. Turek, Flux correction tools for finite elements. *J. Comput. Phys.* **175** (2002) 525–558.
- [39] R.J. LeVeque, High-resolution conservative algorithms for advection in incompressible flow. *SIAM Journal on Numerical Analysis* **33**, (1996) 627–665.
- [40] R. Löhner, K. Morgan, J. Peraire, and M. Vahdati, Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations. *Int. J. Numer. Meth. Fluids* **7** (1987) 1093–1109.
- [41] C. Lohmann, *Physics-Compatible Finite Element Methods for Scalar and Tensorial Advection Problems*. Springer Spektrum, 2019.
- [42] C. Lohmann, D. Kuzmin, J.N. Shadid, and S. Mabuza, Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements. *J. Comput. Phys.* **344** (2017) 151–186.
- [43] H. Luo, J.D. Baum, and R. Löhner, Edge-based finite element scheme for the Euler equations. *AIAA Journal* **32** (1994) 1183–1190.
- [44] P.R.M. Lyra, K. Morgan, J. Peraire, and J. Peiro, TVD algorithms for the solution of the compressible Euler equations on unstructured meshes. *Int. J. Numer. Methods Fluids* **19** (1994) 827–847.
- [45] J. Peraire, M. Vahdati, J. Peiro, and K. Morgan, The construction and behaviour of some unstructured grid algorithms for compressible flows. *Numerical Methods for Fluid Dynamics IV*, Oxford University Press, 1993, 221–239.
- [46] G.M. Phillips, *Interpolation and Approximation by Polynomials*. Springer, New York, 2003.
- [47] V. Selmin, The node-centred finite volume approach: bridge between finite differences and finite elements. *Comput. Methods Appl. Mech. Engrg.* **102** (1993) 107–138.
- [48] V. Selmin and L. Formaggia, Unified construction of finite element and finite volume discretizations for compressible flows. *Int. J. Numer. Methods Engrg.* **39** (1996) 1–32.
- [49] C.-W. Shu and S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comput. Phys.* **77** (1988) 439–471.
- [50] S.T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.* **31** (1979) 335–362.

## Appendix: Sparsity of the lumped discrete gradient

Let  $\lambda_1, \dots, \lambda_{d+1} : \hat{K} \rightarrow [0, 1]$  denote the barycentric coordinates (i.e.,  $\mathbb{P}_1$  basis functions) associated with the vertices of a  $d$ -dimensional reference element  $\hat{K}$ . The 1D Bernstein basis functions of degree  $p \in \mathbb{N}$  are defined on the unit interval  $\hat{K} = [0, 1]^d$  thus:

$$B_\alpha^p(x) = \binom{p}{\alpha} \lambda_1^\alpha(x) \lambda_2^{p-\alpha}(x), \quad 0 \leq \alpha \leq p.$$

The corresponding tensor product basis for  $\mathbb{Q}_p(\hat{K})$  on a unit  $d$ -box  $\hat{K} = [0, 1]^d$  is defined as follows:

$$B_\alpha^p(x_1, \dots, x_d) = B_{\alpha_1}^p(x_1) \dots B_{\alpha_d}^p(x_d), \quad 0 \leq \alpha_1, \dots, \alpha_d \leq p.$$

The Bernstein basis of  $\mathbb{P}_p(K^e)$  a  $d$ -simplex  $K^e = \text{conv}\{\mathbf{x}_1^e, \dots, \mathbf{x}_{d+1}^e\}$  is given by

$$B_\alpha^p(\lambda_1, \dots, \lambda_{d+1}) = \frac{p!}{\alpha_1! \dots \alpha_{d+1}!} \lambda_1^{\alpha_1} \dots \lambda_{d+1}^{\alpha_{d+1}},$$

where  $\alpha = (\alpha_1, \dots, \alpha_{d+1})$  is a multiindex such that

$$|\alpha| := \alpha_1 + \dots + \alpha_{d+1} = p.$$

Consider the  $N \times N$  element matrices  $P^e = M_L^e (M_C^e)^{-1}$  and  $C_k^e$ ,  $k = 1, \dots, d$  of the polynomial space spanned by  $\varphi_i^e = B_{\alpha(i)}^p$ ,  $i = 1, \dots, N$ . By definition (14), the  $j$ -th column of the element matrix  $\tilde{C}_k^e$  contains the Bernstein coefficients of  $\frac{\partial \varphi_j^e}{\partial x_k}$  multiplied by the diagonal entries  $m_i^e = \frac{|K^e|}{N}$  of  $M_L^e$  [42]. Indeed, the Bernstein polynomial  $B_h = \sum_{i=1}^N \tilde{c}_{ij,k}^e \varphi_i^e$  is the unique solution of

$$\int_{K^e} \varphi_h^e B_h \, d\mathbf{x} = \frac{|K^e|}{N} \int_{\hat{K}} \varphi_h^e \frac{\partial \varphi_j^e}{\partial x_k} \, d\hat{\mathbf{x}}, \quad \varphi_h^e \in \{\varphi_1^e, \dots, \varphi_N^e\}.$$

The solution of this linear system yields the Bernstein coefficients of the local  $L^2$  projection which is exact for polynomials of degree up to  $p$ . It follows that  $B_h = \frac{|K^e|}{N} \frac{\partial \varphi_j^e}{\partial x_k}$ .

Using the product rule, the gradient of the Bernstein basis function  $B_\alpha^p$  on a  $d$ -dimensional simplex element  $\hat{K}$  can be written as [29, 31]

$$\nabla B_\alpha^p = p \sum_{k=1}^{d+1} B_{\alpha - e_k}^{p-1} \nabla \lambda_k.$$

The degree elevation formula for simplicial Bernstein elements yields

$$B_{\alpha - e_k}^{p-1} = \frac{1}{p} \sum_{l=1}^{d+1} (\alpha_k - e_k + e_l + 1) B_\alpha^p$$

and the compact sparsity pattern follows from the fact that

$$\nabla B_\alpha^p = \sum_{k=1}^{d+1} \sum_{l=1}^{d+1} (\alpha_k - e_k + e_l + 1) \nabla \lambda_k B_{\alpha - e_k + e_l}^p = \sum_{|\beta|=p} \tilde{c}_\beta B_\beta^p,$$

where  $\tilde{c}_\beta = 0$  if  $\beta \neq \alpha - e_k + e_l$  for some  $k, l \in \{1, \dots, d\}$ . Hence, the coefficient  $\tilde{c}_{ij,k}^e$  is nonvanishing only if  $j = i$  or  $i$  and  $j$  are nearest neighbors belonging to the same grid line of the Bézier net.

To verify the sparsity of the element matrix  $\tilde{C}^e$  for a multidimensional  $d$ -box  $K^e$ , note that

$$\frac{\partial B_\alpha^p}{\partial x_k} = \frac{\partial B_{\alpha_k}^p}{\partial x_k} \prod_{\substack{l=1 \\ l \neq k}}^d B_{\alpha_l}^p, \quad k = 1, \dots, d.$$

The desired result follows from the proof of sparsity for the one-dimensional simplex element.

We remark that the above formulas can also be used for practical calculation of the lumped discrete gradient operator. Efficient algorithms for calculating and inverting the element matrices of high-order Bernstein finite element spaces can be found in [2, 30, 31, 32].