No. 550 **December 2016**

Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements

C. Lohmann, D. Kuzmin, J. N. Shadid, S. Mabuza

ISSN: 2190-1767

Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements

Christoph Lohmann[a], Dmitri Kuzmin[a], John N. Shadid[b,c], Sibusiso Mabuza[b]

[a] *Institute of Applied Mathematics (LS III), TU Dortmund University, Vogelpothsweg 87, D-44227 Dortmund, Germany*
[b] *Computational Mathematics Department, Sandia National Laboratories, PO Box 5800 MS 1321, Albuquerque, NM 87185-1321, USA*
[c] *Department of Mathematics and Statistics, University of New Mexico, MSC01 1115, Albuquerque, NM 87131, USA*

Abstract

This work extends the flux-corrected transport (FCT) methodology to arbitrary-order continuous finite element discretizations of scalar conservation laws on simplex meshes. Using Bernstein polynomials as local basis functions, we constrain the total variation of the numerical solution by imposing local discrete maximum principles on the Bézier net. The design of accuracy-preserving FCT schemes for high order Bernstein-Bézier finite elements requires the development of new algorithms and/or generalization of limiting techniques tailored for linear and multilinear Lagrange elements. In this paper, we propose (i) a new discrete upwinding strategy leading to local extremum bounded low order approximations with compact stencils, (ii) a high order stabilization operator based on gradient recovery, and (iii) new localized limiting techniques for antidiffusive element contributions. The optional use of a smoothness indicator, based on a second derivative test, makes it possible to potentially avoid unnecessary limiting at smooth extrema and achieve optimal convergence rates for problems with smooth solutions. The accuracy of the proposed schemes is assessed in numerical studies for the linear transport equation in 1D and 2D.

Keywords: Bernstein-Bézier finite elements, continuous Galerkin method, flux-corrected transport, artificial diffusion, local discrete maximum principles, total variation diminishing property

1. Introduction

The traditional approach to stabilization of finite element methods for convection-dominated transport problems is based on the use of modified test functions or additional terms in the variational form of the governing equations. For decades, the analysis and design of stabilized finite element schemes for flow problems was focused on various extensions of the streamline upwind Petrov-Galerkin (SUPG) method [11, 31]. More recent successful approaches include continuous interior penalty (CIP) methods [13], local projection stabilization (LPS) schemes [10], and variational multiscale (VMS) methods [33] to name just a few. In general, variational stabilization techniques are effective in localizing and dampening non-physical oscillations. However, small undershoots and overshoots may survive in the neighborhood of steep fronts even if nonlinear discontinuity-capturing terms [15, 18, 31, 33] are employed. The lack of strong nonlinear stability in these methods can therefore lead to positivity and maximum principle violations. Additionally, the robustness and accuracy of some methods are strongly influenced by the choice of free parameters.

Recent years have witnessed an increased interest of the finite element community in nonlinear high-resolution schemes satisfying local discrete maximum principles [3–5, 12, 30]. The oldest representative of such schemes is the flux-corrected transport (FCT) algorithm introduced by Boris and Book [9] and generalized by Zalesak [49] in the 1970s. Modern FCT-like schemes constrain the troublesome antidiffusive part of a linearly stable high order approximation using a conservative decomposition into total mass conserving components (fluxes or element contributions). In FCT algorithms of predictor-corrector type, the local bounds and raw antidiffusive components are

Email addresses: christoph.lohmann@math.tu-dortmund.de (Christoph Lohmann), dmitri.kuzmin@math.tu-dortmund.de (Dmitri Kuzmin), jnshadi@sandia.gov (John N. Shadid), smabuza@sandia.gov (Sibusiso Mabuza)

defined in terms of a low order solution which is guaranteed to satisfy the relevant inequality constraints. High-resolution finite element schemes based on generalizations of the FCT methodology to unstructured grids can be found in [17, 30, 35, 36, 40, 45, 46]. Some algorithms are designed to manipulate the global finite element matrix and limit antidiffusive fluxes associated with the edges of its sparsity graph, whereas others construct artificial diffusion operators on the element matrix level and constrain antidiffusive element contributions. In the present paper, we follow the latter approach.

So far, the use of FCT in the context of continuous Galerkin methods has been restricted to low order (linear or multilinear) finite elements. The stronger coupling between the local degrees of freedom in high order FEM produces larger computational stencils. Furthermore, numerical solutions are no longer bounded by their nodal values if standard Lagrange finite elements are employed. Thus, imposing local bounds on pointwise solution values at the nodes is generally insufficient to guarantee that no undershoots or overshoots will emerge elsewhere.

In this work, we consider high order finite element approximations based on the Bernstein-Bézier representation [1, 34] of polynomial shape functions on simplices. The Bernstein polynomials are all non-negative and form a partition of unity. Moreover, they possess variation diminishing and monotonicity preserving properties [7, 23, 24, 26, 27] which make them a perfect tool for the design of constrained high-resolution schemes. In contrast to Lagrange finite elements, only vertex-based degrees of freedom coincide with the pointwise solution values. However, the restriction of the numerical solution to each cell is uniformly bounded by the largest and smallest coefficient of the Bernstein basis functions. The piecewise-linear interpolant of the Bernstein coefficients is known as the *Bézier net* and its variation represents an upper bound for the variation of the high order polynomial shape function [26, 27]. Hence, the total variation may be constrained by imposing local discrete maximum principles on the Bernstein coefficients. The FCT approach has already been used for this purpose in the framework of explicit discontinuous Galerkin (DG) methods [2]. Promising numerical results were obtained for Bernstein polynomials of degree as high as $p = 23$ in 1D and $p = 5$ in 2D using localized bounds and localized limiting techniques for antidiffusive element contributions. This work was focused on explicit Eulerian advection and remap algorithms in which exact local bounds are provided by the solution at the beginning of the time step or at the end of a Lagrangian advection phase, respectively.

The present paper introduces some new tools for the design of full accuracy-preserving FCT algorithms for high order Bernstein finite element discretizations. In contrast to [2], the new features are introduced in the context of continuous Galerkin schemes but most of them are readily applicable in the DG setting as well. The first highlight is a new generalization of the discrete upwinding technique that we use to construct the artificial diffusion operator for the low order scheme. Using a local mass lumping operator, we obtain element matrices which have the same sparsity pattern as the piecewise-linear approximation on a submesh corresponding to the Bézier net, at least in the case of a constant velocity field. This approach prevents the low order scheme from becoming more diffusive as we increase the order of the Bernstein polynomials while keeping the total number of degrees of freedom fixed. The second highlight of the proposed methodology is the use of antidiffusive element contributions incorporating high order background dissipation. The dissipative component represents a weak form of the Laplacian operator acting on the difference between two gradient approximations. The stabilized high order scheme belongs to the family of VMS and LPS methods [10, 32]. Interestingly enough, the 1D version of this scheme turns out to be a natural extension of SUPG to unsteady transport problems. Further improvements to be discussed in the context of high order FCT schemes include new localized limiters based on a decomposition of element vectors into subcell contributions and smoothness indicators inspired by the non-clipping PPM limiter of Colella and Sekora [16]. The positive effect of the proposed upgrades is illustrated by numerical studies for linear convection of smooth and discontinuous data.

2. Bernstein basis functions

In this section, we introduce the Bernstein basis functions in the one- and two-dimensional space. An extension to the three-dimensional case follows directly. Furthermore, we summarize several important properties which are used in the following sections to prove some analytical results for the proposed FCT algorithms.

Bernstein polynomials are a special set of real-valued polynomials with integer coefficients and build a basis of the underlying polynomial vector space. In the one-dimensional space, the Bernstein basis functions are defined on the unit interval $x \in [0, 1]$ by

$$B_k^p(x) = \binom{p}{k} x^k (1-x)^{p-k}, \qquad 0 \le k \le p. \tag{1}$$

An extension to two dimensions is given by the tensor product ansatz

$$B^p_{i,j}(x,y) = B^p_i(x)B^p_j(y) \qquad 0 \le i,j \le p \tag{2}$$

for quadrilaterals (unit square: $x, y \in [0,1]$) and

$$B^p_{i,j,k}(u,v,w) = \frac{p!}{i!\,j!k!}u^i v^j w^k, \qquad 0 \le i,j,k, \quad i+j+k = p \tag{3}$$

for triangles using barycentric coordinates u, v, w such that [22]

$$(u,v,w) \in \Lambda \coloneqq \{(\lambda_1,\lambda_2,\lambda_3) \in \mathbb{R}^3 \,|\, 0 \le \lambda_1,\lambda_2,\lambda_3 \le 1,\ \lambda_1+\lambda_2+\lambda_3 = 1\}.$$

The Bernstein basis functions of an arbitrary order p form a partition of unity

$$\sum_{k=0}^{p} B^p_k(x) = \sum_{k=0}^{p}\binom{n}{k}x^k(1-x)^{p-k} = (x+(1-x))^p = 1 \qquad \text{for all } x, \tag{4a}$$

$$\begin{aligned}\sum_{i=0}^{p}\sum_{j=0}^{p-i} B^p_{i,j,p-i-j}(u,v,w) &= \sum_{i=0}^{p}\binom{p}{i}u^i \sum_{j=0}^{p-i}\binom{p-i}{j}v^j w^{p-i-j} \\ &= \sum_{i=0}^{p}\binom{p}{i}u^i(v+w)^{p-i} = (u+v+w)^p = 1 \qquad \text{for all } (u,v,w) \in \Lambda,\end{aligned} \tag{4b}$$

where we have used the binomial theorem

$$(x+y)^l = \sum_{i=0}^{l}\binom{l}{i}x^i y^{l-i} \qquad \text{for all } x,y \in \mathbb{R} \text{ and } l \in \mathbb{N}.$$

Furthermore, the basis functions are non-negative since $B^p_k \ge 0$ and $B^p_{i,j,k} \ge 0$ by definition of the Bernstein polynomials. By virtue of (4), it follows that

$$\begin{aligned} &0 \le B^p_k(x) \le 1 && \text{for all } 0 \le k \le p \text{ and } x \in [0,1], \\ &0 \le B^p_{i,j,k}(u,v,w) \le 1 && \text{for all } 0 \le i,j,k, \le p,\ i+j+k = p \text{ and } (u,v,w) \in \Lambda. \end{aligned}$$

Using the partition-of-unity property (4), we can also prove that a function defined as an arbitrary linear combination of B^p_k is uniformly bounded by the minimal and maximal coefficients

$$\min_{0\le k\le p} c_k = (\min_{0\le k\le p} c_k)\underbrace{\sum_{k=0}^{p} B^p_k(x)}_{=1} \le u_h(x) = \sum_{k=0}^{p} c_k B^p_k(x) \le (\max_{0\le k\le p} c_k)\underbrace{\sum_{k=0}^{p} B^p_k(x)}_{=1} = \max_{0\le k\le p} c_k \qquad \text{for all } x \in [0,1], \tag{5}$$

where $c_k \in \mathbb{R}$ are the Bernstein coefficients. Obviously, the same relationship is valid in the two-dimensional case. When it comes to the design of FCT algorithms in the next sections, this will guarantee the local boundedness of function values and also a total variation boundedness (TVB) property implied by the local extremum diminishing (LED) property of the antidiffusive correction for the Bernstein coefficients / Bézier nets.

In contrast to Lagrange basis polynomials, the values of the coefficients in the representation of a function as a linear combination of Bernstein basis functions do not necessarily coincide with the actual function values at some points. In other words, a potential disadvantage of Bernstein polynomials compared to Lagrange polynomials is the lack of an interpolation property. However, it is possible to assign the Bernstein coefficients to certain points by examining corresponding local extrema: Each Bernstein polynomial has exactly one local extremum, which is given by $x = \frac{k}{p}$ in the case of one-dimensional basis functions B^p_k and $(u,v,w) = (\frac{i}{p},\frac{j}{p},\frac{k}{p})$ for Bernstein functions $B^p_{i,j,k}$ defined on a triangle. Therefore, the degree of freedom c^p_k corresponding to the basis function B^p_k can be associated

with $x = \frac{k}{p}$ (see, e.g., [26] for relationships in the context of Bernstein approximations of continuous functions and [42] for estimates of the difference between the Bernstein coefficients and function values at nodal points). Only in the cases $k = 0$ and $k = p$ ($i = p$, $j = p$ or $k = p$ for triangles) the interpolation property is valid, i.e., function values and coefficients match. In the context of finite element approximations, these degrees of freedom will correspond to nodes of the mesh. Furthermore, in two dimensions $B^p_{i,j,k}$ vanishes on the boundary of the triangle if $0 < i, j, k < p$.

In the next sections, we will not distinguish between the one- and two-dimensional space and corresponding Bernstein polynomials B^p_k and $B^p_{i,j,k}$, respectively, any more. Henceforth, the global basis functions of the continuous finite element space of an arbitrary order p are denoted by B_j, while the corresponding coefficients/degrees of freedom are given by c_j. Furthermore, the total numbers of degrees of freedom and elements are denoted by N_{dof} and N_{elem}, respectively.

The integer set containing the global numbers of nodes belonging to an element K_e of the triangulation will be referred to as $\mathcal{N}^e$ and called the *element stencil*. The notation $\mathcal{N}_j$ is reserved for *nodal stencils*, i.e., integer sets containing the global numbers of a node and its neighbors. The above definitions imply that $i \in \mathcal{N}^e$ if and only if the corresponding node is in K_e, while $j \in \mathcal{N}_i$ if and only if there exists an index e such that $i, j \in \mathcal{N}^e$.

3. Finite element approximations

In this paper, we use Bernstein finite elements in continuous Galerkin approximations to the scalar transport equation in the one- and two-dimensional space. The initial-boundary value problem is given by

$$\begin{cases} \partial_t u + \operatorname{div}(\mathbf{v}u) = 0 & \text{in } \Omega, \quad (6a) \\ u(\cdot, t) = u_{\text{in}} & \text{on } \Gamma_{\text{in}} = \{\mathbf{x} \in \partial\Omega : \mathbf{n}(\mathbf{x}) \cdot \mathbf{v} < 0\}, \quad (6b) \\ u(\cdot, 0) = u_0 & \text{in } \Omega, \quad (6c) \end{cases}$$

where $\Omega \subset \mathbb{R}^d$, $d = 1, 2$ is a bounded domain, Γ_{in} is the inflow part of the boundary $\partial\Omega$, $\mathbf{n} : \partial\Omega \to \mathbb{R}^d$ is the unit outward normal vector, $\mathbf{v} : \mathbb{R}^d \times \mathbb{R}^+_0 \to \mathbb{R}^d$ is the velocity field, and $u : \mathbb{R}^d \times \mathbb{R}^+_0 \to \mathbb{R}$ is the unknown solution. The initial and boundary conditions are given by the functions $u_0 : \Omega \to \mathbb{R}$ and $u_{\text{in}} : \Gamma_{\text{in}} \times \mathbb{R}^+_0$.

3.1. Galerkin space discretization

To discretize the above problem using finite elements, we introduce the variational formulation

$$\int_\Omega w\Big(\frac{\partial u}{\partial t} + \operatorname{div}(\mathbf{v}u)\Big)\,\mathrm{d}\mathbf{x} = 0 \qquad \text{for all } w \in V, \tag{7}$$

where V is the space of admissible test functions. Integration by parts yields

$$\int_\Omega w\frac{\partial u}{\partial t}\,\mathrm{d}\mathbf{x} - \int_\Omega \operatorname{grad}(w) \cdot (\mathbf{v}u)\,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\text{out}}} wu\mathbf{v} \cdot \mathbf{n}\,\mathrm{d}\mathbf{s} + \int_{\Gamma_{\text{in}}} wu_{\text{in}}\mathbf{v} \cdot \mathbf{n}\,\mathrm{d}\mathbf{s} = 0. \tag{8}$$

The above splitting of the boundary integral into two parts corresponding to the inlet Γ_{in} and outlet $\Gamma_{\text{out}} \subseteq \partial\Omega\backslash\Gamma_{\text{in}}$ makes it possible to implement the Dirichlet boundary condition (6b) in a weak sense by substituting u_{in} into the integral over Γ_{in} instead of imposing the usual restrictions on the functional spaces for u and w.

Introducing the finite-dimensional space $V_h \coloneqq \operatorname{span}\{B_1, \ldots, B_{N_{\text{dof}}}\}$, approximating u by $u_h = \sum_{j=1}^{N_{\text{dof}}} c_j B_j \in V_h$, and testing with the Bernstein basis functions B_i, we obtain the (high order) semi-discrete Galerkin scheme

$$\sum_{j=1}^{N_{\text{dof}}} \underbrace{\sum_{e=1}^{N_{\text{elem}}} m^e_{ij}}_{=:m_{ij}} \frac{\mathrm{d}c_j}{\mathrm{d}t} = \sum_{j=1}^{N_{\text{dof}}} \underbrace{\sum_{e=1}^{N_{\text{elem}}} k^e_{ij}}_{=:k_{ij}} c_j + \underbrace{\sum_{e=1}^{N_{\text{elem}}} b^e_i}_{=:b_i} \qquad \text{for all } 1 \le i \le N_{\text{dof}}, \tag{9a}$$

$$m^e_{ij} = \int_{K_e} B_i B_j\,\mathrm{d}\mathbf{x}, \qquad k^e_{ij} = \int_{K_e} \operatorname{grad}(B_i) \cdot \mathbf{v}B_j\,\mathrm{d}\mathbf{x} - \int_{\Gamma_{\text{out}} \cap \partial K_e} B_i B_j \mathbf{v} \cdot \mathbf{n}\,\mathrm{d}\mathbf{s}, \qquad b^e_i = -\int_{\Gamma_{\text{in}} \cap \partial K_e} B_i u_{\text{in}} \mathbf{v} \cdot \mathbf{n}\,\mathrm{d}\mathbf{s}, \tag{9b}$$

where m_{ij}^e, k_{ij}^e, and b_i^e are element contributions to the mass matrix, convection matrix and load vector of the weakly imposed Dirichlet boundary condition, respectively.

The Galerkin approximation defined by (9) may produce disappointing results, especially if p is large. The resulting numerical solutions are frequently corrupted by spurious oscillations causing violations of local maximum principles that may apply to the exact solution of the continuous initial-boundary value problem (6) for the transport equation.

3.2. Low order upwind methods

To obtain a local extremum diminishing (LED) reference solution that provides local bounds for undershoot/overshoot detection, FCT algorithms add artificial diffusion to a high order target scheme. The resulting low order solution is used to define inequality constraints for a subsequent antidiffusive correction step. In the context of linear finite elements, a low order method with provable LED properties can be constructed as in [40] by using mass lumping on the left-hand side and adding an artificial diffusion operator on the right-hand side of (9a). If the diagonal entries of the modified mass matrix are positive and the off-diagonal entries of the resulting low order convection operator are non-negative, i.e., if each semi-discrete equation can be written as

$$m_i \frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{j=1}^{N_{\mathrm{dof}}} l_{ij} c_j + b_i \qquad \text{with} \quad m_i > 0, \quad l_{ij} \geq 0 \quad \text{for all } 1 \leq i, j \leq N_{\mathrm{dof}},\ j \neq i, \tag{10}$$

then the low order approximation satisfies a semi-discrete local maximum principle and proves total variation diminishing (TVD) in 1D [36].

In the case of high order Bernstein finite elements, such modifications of the semi-discrete problem (9a) do not rule out formation and/or growth of local extrema due to missing interpolation properties of Bernstein basis functions. Nevertheless, a low order method satisfying (10) guarantees that a maximum principle holds for the coefficients. It follows that pointwise values must satisfy a weak LED condition: If the degree of freedom c_i does not correspond to a node on the boundary of the domain Ω and its current value is greater/smaller than or equal to each common-element degree of freedom, then c_i cannot increase/decrease in time. In our stencil notation, a common-element neighbor is a node $j \neq i$ such that $j \in \mathcal{N}_i$. Expressed in formulas, the so-defined weak LED property means that

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} \leq 0 \qquad \text{if } c_i \geq \max_{j \neq i,\, j \in \mathcal{N}_i} c_j, \qquad\qquad \frac{\mathrm{d}c_i}{\mathrm{d}t} \geq 0 \qquad \text{if } c_i \leq \min_{j \neq i,\, j \in \mathcal{N}_i} c_j. \tag{11}$$

Under certain additional assumptions regarding the time integrator, the semi-discrete property (11) carries over to the fully discrete method. Therefore, the Bernstein coefficients of the numerical solution are bounded by the coefficients of the initial and (inflow) boundary condition. In addition to (5), the low order function values satisfy local constraints which prevent arbitrary growth of extrema. Furthermore, the total variation boundedness (TVB) property of 1D approximations can be shown using the following estimate in terms of bounded Bernstein coefficients:

$$\begin{aligned} \mathrm{TV}(u_h(\cdot,t)) &= \int_{-\infty}^{\infty} \left|\mathrm{grad}(u_h)\right| \mathrm{d}\mathbf{x} = \int_{-\infty}^{\infty} \Big| \sum_{j=1}^{N_{\mathrm{dof}}} c_j \mathrm{grad}(B_j) \Big| \,\mathrm{d}\mathbf{x} \leq \sum_{j=1}^{N_{\mathrm{dof}}} |c_j| \underbrace{\int_{-\infty}^{\infty} \left|\mathrm{grad}(B_j)\right| \mathrm{d}\mathbf{x}}_{\leq C} \\ &\leq C \sum_{j=1}^{N_{\mathrm{dof}}} |c_j| \leq C \cdot N_{\mathrm{dof}} \cdot \max\left\{|c^{\max}|, |c^{\min}|\right\} \leq \ldots \leq \tilde{C}(u_0, u_{\mathrm{in}}). \end{aligned} \tag{12}$$

Further estimates can presumably be derived using the variation diminishing [7, 26, 27] and monotonicity-preserving [23, 24] properties of Bernstein polynomials. We remark that the use of weakened LED constraints is essential for maintaining optimal accuracy in FCT algorithms based on high order Bernstein finite element approximations.

3.2.1. Element-stencil upwinding

The easiest way to derive a low order method of the form (10) is by using discrete upwinding approaches. The classical one, as described in detail in [36] for linear finite elements, can be summarized as follows: The mass matrix is

		element upwinding		subcell upwinding		subcell Rusanov	
p	N_{dof}	L^1	L^∞	L^1	L^∞	L^1	L^∞
1	$(128+1)^2 = 16\,641$	1.019e−1	7.642e−1	1.019e−1	7.642e−1	1.097e−1	8.116e−1
2	$(128+1)^2 = 16\,641$	1.081e−1	7.915e−1	1.044e−1	7.742e−1	1.105e−1	8.185e−1
3	$(129+1)^2 = 16\,900$	1.109e−1	8.137e−1	1.048e−1	7.756e−1	1.111e−1	8.218e−1
4	$(128+1)^2 = 16\,641$	1.127e−1	8.302e−1	1.048e−1	7.765e−1	1.113e−1	8.243e−1
5	$(130+1)^2 = 17\,161$	1.139e−1	8.405e−1	1.044e−1	7.769e−1	1.113e−1	8.226e−1
6	$(126+1)^2 = 16\,129$	1.154e−1	8.449e−1	1.050e−1	7.731e−1	1.120e−1	8.191e−1
7	$(126+1)^2 = 16\,129$	1.162e−1	8.618e−1	1.049e−1	7.769e−1	1.121e−1	8.275e−1
8	$(128+1)^2 = 16\,641$	1.166e−1	8.664e−1	1.042e−1	7.780e−1	1.118e−1	8.260e−1
9	$(126+1)^2 = 16\,129$	1.171e−1	8.744e−1	1.045e−1	7.750e−1	1.121e−1	8.280e−1
10	$(130+1)^2 = 17\,161$	1.173e−1	8.655e−1	1.041e−1	7.672e−1	1.118e−1	8.137e−1

Table 1: Comparison of various low order methods in the context of the solid body rotation test, the discussion of which can be found in Section 6.2. The method corresponding to element upwinding is defined by (13). Both subcell upwinding techniques are defined in Section 3.2.2. The mesh size is adjusted to keep the number of global degrees of freedom approximately the same for all orders p of Bernstein finite elements.

replaced by its lumped counterpart and negative off-diagonal entries of the convection matrix are eliminated by adding a symmetric diffusion matrix with vanishing row and column sums. In an element-based version, the application of this procedure to the element matrices of the Galerkin scheme results in the following low order method:

$$\underbrace{\sum_{e=1}^{N_{\text{elem}}} m_i^e}_{=:m_i} \frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{j=1}^{N_{\text{dof}}} \underbrace{\sum_{e=1}^{N_{\text{elem}}} l_{ij}^e}_{=:l_{ij}} c_j + \sum_{e=1}^{N_{\text{elem}}} b_i^e \qquad \text{for all } 1 \le i \le N_{\text{dof}}, \tag{13a}$$

$$m_i^e = \sum_{j=1}^{N_{\text{dof}}} m_{ij}^e, \qquad l_{ij}^e = k_{ij}^e + d_{ij}^e, \quad d_{ij}^e = \begin{cases} \max\{-k_{ij}^e, 0, -k_{ji}^e\} & : j \ne i, \\ -\sum_{k \ne i} d_{ik}^e & : j = i. \end{cases} \tag{13b}$$

Anderson et al. [2] used this *element-stencil upwinding* approach also in the context of FCT based on high order discontinuous Galerkin methods. However, this kind of algebraic upwinding produces increasingly diffusive solutions as the order p of the Bernstein finite element approximation is increased while keeping the total number of degrees of freedom fixed (see Table 1). The reason for this behavior lies in the fact that the underlying element matrices are dense and more negative off-diagonal entries need to be eliminated to satisfy the inequality constraints imposed in (10) as p increases. Additional nonzero entries of the artificial diffusion operator give rise to additional numerical dissipation, the amount of which depends on the discrete upwinding strategy. In contrast to the case $p = 1$, the one based on (13b) does not generally produce the least diffusive approximation of the form (10) in the case $p > 1$.

3.2.2. Subcell-stencil upwinding

A high order Bernstein element has the same nodal points and, therefore, produces matrices of the same size as the Bézier net defined as the piecewise-linear interpolant of its Bernstein coefficients [26]. Since the high order basis functions are non-vanishing on the whole element, the resulting element matrices are full and some non-zero entries are associated with pairs of nodes i and j that do not belong to the same subcell of the triangulation associated with the Bézier net. Since the distance between such nodes is larger than that between nodes of the same subcell, the differences between the values of c_i and c_j are also likely to be more significant. Therefore, additional diffusive fluxes associated with such pairs of nodes tend to be large even if the corresponding coefficients d_{ij}^e are small.

To avoid this drawback, we introduce a less diffusive discrete upwinding strategy which modifies the convection element matrix in a manner consistent with (element-based) mass lumping before calculating the artificial diffusion coefficients. This preprocessing step reduces the magnitude of off-diagonal entries associated with pairs of nodes belonging to different subcells or even eliminates these entries completely (compare matrices shown in Fig. 1). In the special case of a constant velocity field, the modified convection matrix has the same sparsity pattern as the matrix

of the piecewise-linear subcell approximation. The remaining negative off-diagonal entries are eliminated as in (13b) producing smaller artificial diffusion coefficients d_{ij}^e for pairs of nodes i and j that do not belong to the same subcell. This approach to construction of the low order scheme will be called *subcell-stencil upwinding*.

To justify the proposed modification of the convection matrix, let us consider the local semi-discrete problem

$$\begin{aligned}
0 &= \sum_{j=1}^{N_{\text{dof}}} \Big(\int_{K_e} B_i B_j \,\mathrm{d}\mathbf{x} \Big) \frac{\mathrm{d}c_j}{\mathrm{d}t} + \sum_{j=1}^{N_{\text{dof}}} \Big(\int_{K_e} B_i \mathrm{div}(\mathbf{v} B_j) \,\mathrm{d}\mathbf{x} \Big) c_j \\
&= \sum_{j=1}^{N_{\text{dof}}} \Big(\int_{K_e} B_i B_j \,\mathrm{d}\mathbf{x} \Big) \frac{\mathrm{d}c_j}{\mathrm{d}t} + \sum_{j=1}^{N_{\text{dof}}} \Big(- \int_{K_e} \mathrm{grad}(B_i) \cdot \mathbf{v} B_j \,\mathrm{d}\mathbf{x} + \int_{\partial K_e} B_i B_j \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}\mathbf{s} \Big) c_j \\
&= \sum_{j=1}^{N_{\text{dof}}} m_{ij}^e \frac{\mathrm{d}c_j}{\mathrm{d}t} - \sum_{j=1}^{N_{\text{dof}}} (k_{ij}^e + h_{ij}^e) c_j = \sum_{j=1}^{N_{\text{dof}}} m_{ij}^e \frac{\mathrm{d}c_j}{\mathrm{d}t} - \sum_{j=1}^{N_{\text{dof}}} \varkappa_{ij}^e c_j \qquad \text{for all } 1 \le i \le N_{\text{dof}}
\end{aligned} \tag{14}$$

associated with a single element K_e. In the last line of (14), we define $\varkappa_{ij}^e = k_{ij}^e + h_{ij}^e$ as entries of the convection matrix corresponding to the local variational formulation without integration by parts. We have

$$h_{ij}^e = -\int_{\partial K_e \setminus \Gamma_{\text{out}}} B_i B_j \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}\mathbf{s} \qquad \Rightarrow \qquad \varkappa_{ij}^e = \int_{K_e} B_i \mathrm{div}(\mathbf{v} B_j) \,\mathrm{d}\mathbf{s}. \tag{15}$$

The representation in terms of $\varkappa_{ij}^e$ will be used to derive the local artificial diffusion operator for subcell-stencil upwinding in this section. However, the antidiffusive element contributions for FCT will be defined using the representation in terms of k_{ij}^e to guarantee local mass conservation properties when it comes to limiting.

In the context of preconditioning, multiplication of a linear system by a non-singular matrix is often used to obtain an equivalent system with improved properties. In the context of subcell-stencil upwinding, we opt to modify the element matrices with entries m_{ij}^e and $\varkappa_{ij}^e$ in a consistent manner. To that end, we multiply both element matrices by the same local 'preconditioner' with entries $m_i^e m_{ij}^{-e}$, where m_{ij}^{-e} are the coefficients of the inverse element mass matrix. The application of the so-defined local mass lumping operator transforms (14) into the equivalent system

$$0 = m_i^e \frac{\mathrm{d}c_i}{\mathrm{d}t} - \sum_{j=1}^{N_{\text{dof}}} \underbrace{\sum_{r=1}^{N_{\text{dof}}} m_i^e m_{ir}^{-e} \varkappa_{rj}^e}_{=: \tilde{\varkappa}_{ij}^e} c_j = m_i^e \frac{\mathrm{d}c_i}{\mathrm{d}t} - \sum_{j=1}^{N_{\text{dof}}} \tilde{\varkappa}_{ij}^e c_j \qquad \text{for all } 1 \le i \le N_{\text{dof}}. \tag{16}$$

In a similar vein, element contributions to equations (9a) of the global semi-discrete problem can be modified using the above local preconditioning to perform consistent element-based mass lumping in the process of global matrix assembly. For this purpose, we rewrite (9a) in terms of $\varkappa_{ij}^e$ as follows:

$$\begin{aligned}
\sum_{j=1}^{N_{\text{dof}}} \sum_{e=1}^{N_{\text{elem}}} m_{ij}^e \frac{\mathrm{d}c_j}{\mathrm{d}t} &= \sum_{j=1}^{N_{\text{dof}}} \sum_{e=1}^{N_{\text{elem}}} k_{ij}^e c_j + \sum_{e=1}^{N_{\text{elem}}} b_i^e = \sum_{j=1}^{N_{\text{dof}}} \sum_{e=1}^{N_{\text{elem}}} (k_{ij}^e + h_{ij}^e) c_j - \sum_{j=1}^{N_{\text{dof}}} \sum_{e=1}^{N_{\text{elem}}} h_{ij}^e c_j + \sum_{e=1}^{N_{\text{elem}}} b_i^e \\
&= \sum_{j=1}^{N_{\text{dof}}} \sum_{e=1}^{N_{\text{elem}}} \varkappa_{ij}^e c_j + \sum_{j=1}^{N_{\text{dof}}} \Big(\int_{\Gamma_{\text{in}}} B_i B_j \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}\mathbf{s} \Big) c_j + \sum_{e=1}^{N_{\text{elem}}} b_i^e \qquad \text{for all } 1 \le i \le N_{\text{dof}}.
\end{aligned} \tag{17}$$

The contribution of element K_e to the residual of this system is given by the right-hand side of (14). Replacing it by the right-hand side of the preconditioned local problem (16), we obtain the lumped-mass counterpart of (9a)

$$m_i \frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{e=1}^{N_{\text{elem}}} m_i^e \frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{j=1}^{N_{\text{dof}}} \sum_{e=1}^{N_{\text{elem}}} \tilde{\varkappa}_{ij}^e c_j + \sum_{j=1}^{N_{\text{dof}}} \Big(\int_{\Gamma_{\text{in}}} B_i B_j \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}\mathbf{s} \Big) c_j + \sum_{e=1}^{N_{\text{elem}}} b_i^e. \tag{18}$$

Due to the property that derivatives of Bernstein polynomials can be written as linear combinations of a few Bernstein polynomials of the same degree (see (B.3)), this lumping procedure results in a sparser convection matrix for constant

$$\varkappa_{ij}^e = 10^{-3} \cdot \begin{pmatrix} 1.63 & -2.63 & -1.16 & -0.33 & 2.31 & -0.50 & -0.17 & 0.84 & -0.17 & 0.17 \\ 0.98 & -1.00 & -1.51 & -1.00 & 1.50 & 0.34 & -0.16 & 0.68 & 0.00 & 0.17 \\ 0.49 & -0.02 & -1.03 & -2.02 & 0.84 & 0.85 & 0.03 & 0.51 & 0.18 & 0.17 \\ 0.16 & 0.32 & 0.28 & -3.39 & 0.34 & 1.02 & 0.40 & 0.34 & 0.36 & 0.18 \\ 0.99 & -1.50 & -0.84 & -0.34 & 2.01 & -1.18 & -0.51 & 1.53 & -0.68 & 0.52 \\ 0.50 & -0.34 & -0.85 & -1.02 & 1.18 & 0.00 & -0.84 & 1.20 & -0.33 & 0.52 \\ 0.17 & 0.16 & -0.03 & -2.05 & 0.51 & 0.84 & -1.00 & 0.86 & 0.02 & 0.53 \\ 0.50 & -0.68 & -0.51 & -0.34 & 1.53 & -1.20 & -0.86 & 2.06 & -1.55 & 1.04 \\ 0.17 & -0.00 & -0.18 & -1.03 & 0.68 & 0.33 & -1.55 & 1.55 & -1.03 & 1.06 \\ 0.17 & -0.17 & -0.17 & -0.35 & 0.86 & -0.52 & -1.22 & 2.44 & -2.80 & 1.76 \end{pmatrix},$$

$$\tilde{\varkappa}_{ij}^e = 10^{-3} \cdot \begin{pmatrix} 3.38 & -6.75 & -0.00 & -0.00 & 3.38 & 0.00 & -0.00 & -0.00 & 0.00 & 0.00 \\ 1.13 & -0.13 & -4.50 & -0.00 & 1.25 & 2.25 & 0.00 & -0.00 & -0.00 & 0.00 \\ 0.00 & 2.25 & -3.63 & -2.25 & 0.00 & 2.50 & 1.13 & -0.00 & -0.00 & 0.00 \\ -0.00 & 0.00 & 3.38 & -7.13 & -0.00 & 0.00 & 3.75 & -0.00 & -0.00 & 0.00 \\ 1.25 & -2.50 & 0.00 & 0.00 & 3.50 & -4.50 & -0.00 & 2.25 & 0.00 & -0.00 \\ 0.00 & 1.25 & -2.50 & -0.00 & 1.13 & 0.00 & -2.25 & 1.25 & 1.13 & -0.00 \\ 0.00 & 0.00 & 1.25 & -2.50 & 0.00 & 2.25 & -3.50 & 0.00 & 2.50 & -0.00 \\ 0.00 & -0.00 & -0.00 & -0.00 & 2.50 & -5.00 & 0.00 & 3.63 & -2.25 & 1.13 \\ 0.00 & 0.00 & -0.00 & -0.00 & 0.00 & 2.50 & -5.00 & 1.13 & 0.13 & 1.25 \\ -0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.00 & 0.00 & 3.75 & -7.50 & 3.75 \end{pmatrix}.$$

Figure 1: Example of element-based convection matrices for $p = 3$ in the context of the solid body rotation test, the discussion of which can be found in Section 6.2.

velocity fields. Even in the case of space-dependent velocity fields, the magnitude of off-diagonal entries that do not match the Bézier net sparsity pattern can be drastically reduced (see Fig. 1).

To construct an improved low order method satisfying conditions (10), we add the artificial diffusion coefficients

$$\tilde{d}_{ij}^e = \begin{cases} \max\{-\tilde{\varkappa}_{ij}^e, 0, -\tilde{\varkappa}_{ji}^e\} & : j \neq i, \\ -\sum_{k \neq i} \tilde{d}_{ik}^e & : j = i \end{cases} \tag{19}$$

to the entries $\tilde{\varkappa}_{ij}^e$ of the modified convection matrix and end up with the semi-discrete scheme

$$m_i \frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{j=1}^{N_{\mathrm{dof}}} l_{ij} c_j + b_i \qquad \text{for all } 1 \leq i \leq N_{\mathrm{dof}}, \qquad l_{ij} = \sum_{e=1}^{N_{\mathrm{elem}}} \tilde{\varkappa}_{ij}^e + \tilde{d}_{ij}^e + \Big(\int_{\Gamma_{\mathrm{in}}} B_i B_j \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}\mathbf{s} \Big). \tag{20}$$

To perform subcell-stencil upwinding for a convection matrix defined in terms of k_{ij}^e rather than $\varkappa_{ij}^e$, we use the following representation of the element contribution l_{ij}^e to the entry l_{ij} of the global matrix:

$$l_{ij}^e = k_{ij}^e + d_{ij}^e, \qquad \text{where} \qquad d_{ij}^e = \tilde{d}_{ij}^e - \varkappa_{ij}^e + \tilde{\varkappa}_{ij}^e = \tilde{d}_{ij}^e + \sum_{r=1}^{N_{\mathrm{dof}}} (m_i^e m_{ir}^{-e} - \delta_{ir}) \varkappa_{rj}^e \tag{21}$$

because

$$\sum_{e=1}^{N_{\mathrm{elem}}} l_{ij}^e = \sum_{e=1}^{N_{\mathrm{elem}}} \tilde{\varkappa}_{ij}^e + \tilde{d}_{ij}^e + k_{ij}^e - (k_{ij}^e + h_{ij}^e) = \sum_{e=1}^{N_{\mathrm{elem}}} \tilde{\varkappa}_{ij}^e + \tilde{d}_{ij}^e - \sum_{e=1}^{N_{\mathrm{elem}}} h_{ij}^e = \sum_{e=1}^{N_{\mathrm{elem}}} \tilde{\varkappa}_{ij}^e + \tilde{d}_{ij}^e + \int_{\Gamma_{\mathrm{in}}} B_i B_j \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}\mathbf{s}.$$

Hence, the coefficients l_{ij}^e of the low order convection matrix can be calculated directly by adding d_{ij}^e to k_{ij}^e.

Note that in contrast to the matrix of coefficients $\tilde{d}_{ij}^e$, the element-based artificial diffusion operator defined in

terms of d_{ij}^e does not need to be symmetric. However, it does possess the zero column sum property

$$\sum_{i=1}^{N_{\text{dof}}} d_{ij}^e = \underbrace{\sum_{i=1}^{N_{\text{dof}}} \tilde{d}_{ij}^e}_{=0} + \sum_{i,r=1}^{N_{\text{dof}}} (m_i^e m_{ir}^{-e} - \delta_{ir}) \varkappa_{rj}^e = \sum_{r,s=1}^{N_{\text{dof}}} \underbrace{\sum_{i=1}^{N_{\text{dof}}} (m_i^e \delta_{is} - m_{is}^e)}_{=0} m_{sr}^{-e} \varkappa_{rj}^e = 0, \tag{22}$$

which proves that the low order scheme inherits the mass conservation properties of the underlying Galerkin discretization. Additionally, the row sums of the element-based artificial diffusion operator satisfy

$$\sum_{j=1}^{N_{\text{dof}}} d_{ij}^e = \underbrace{\sum_{j=1}^{N_{\text{dof}}} \tilde{d}_{ij}^e}_{=0} + \sum_{r=1}^{N_{\text{dof}}} (m_i^e m_{ir}^{-e} - \delta_{ir}) \sum_{j=1}^{N_{\text{dof}}} \varkappa_{rj}^e = - \sum_{r=1}^{N_{\text{dof}}} (m_i^e m_{ir}^{-e} - \delta_{ir}) \int_{K_e} B_r \underbrace{\operatorname{div}\Big(\mathbf{v} \overbrace{\sum_{j=1}^{N_{\text{dof}}} B_j}^{=1}\Big)}_{=0 \text{ if } \operatorname{div}(\mathbf{v})=0} \mathrm{d}\mathbf{x} \tag{23}$$

and vanish in the case of a divergence-free velocity field, as required for preservation of consistency. Since the convective derivative of a constant is zero, the corresponding discrete operator must be a matrix with zero row sums for each internal node and any manipulations that may destroy this property are unacceptable.

An alternative interpretation of subcell-stencil upwinding as a modified projection scheme for the time derivatives of local degrees of freedom is given in AppendixA. The compact stencil property of the low order convection matrix is shown in AppendixB for the special case of a one-dimensional constant velocity field. Additionally, a remarkable relationship to discrete upwinding for linear Lagrange finite element approximations is discovered.

Numerical studies have shown that the low order method based on the revised upwinding strategy is less diffusive than the one based on (19), especially for high order Bernstein elements. In fact, the accuracy of low order solutions seems to be largely independent of the order p as it is varied while holding the number of degrees of freedom constant (see Table 1). On the negative side, the numerical examples to be presented in section 6 show that low order solutions produced by the subcell-stencil version of discrete upwinding are less smooth than those produced by the element-stencil version. This minor drawback can be remedied by using a Rusanov-type diffusion operator defined by

$$\tilde{d}_{ij}^e = \begin{cases} \max\{|\tilde{\varkappa}_{ij}^e|, |\tilde{\varkappa}_{ji}^e|\} & : j \neq i, \\ -\sum_{k \neq i} \tilde{d}_{ik}^e & : j = i \end{cases} \tag{24}$$

instead of definition (19). The resulting low order method produces smoother solutions than subcell-stencil upwinding based on (19) and its convergence behavior is also invariant to changes in p (see Table 1).

Detailed studies and a further discussion of different low order methods will be presented in section 6.

3.3. Time integration

To preserve the order of accuracy after the discretization in time and make it possible to capture rapid evolution of local features using high order time integrators, the finite element solution corresponding to the high order Galerkin space discretization (9) is advanced in time using an S-stage Runge-Kutta scheme

$$c^H = c^n + \Delta t \sum_{s=1}^{S} b_s \dot{c}^{n,s}, \tag{25a}$$

$$\sum_{j=1}^{N_{\text{dof}}} m_{ij} \dot{c}_j^{n,s} = \sum_{j=1}^{N_{\text{dof}}} k_{ij} \Big(c_j^n + \Delta t \sum_{t=1}^{S} a_{st} \dot{c}_j^{n,t} \Big) + b_j^{n+c_s} \qquad \text{for all } 1 \leq i \leq N_{\text{dof}}, \tag{25b}$$

where c^n and c^H are the Bernstein coefficients of u_h at t^n and of the high order solution at $t^{n+1} = t^n + \Delta t$, respectively. The time increments $\dot{c}^{n,s}$ are calculated by solving system (25b) and can be interpreted as the Bernstein coefficients of $\partial_t u_h$ at $t^{n+s} = t^n + \Delta t c_s$. The vector of inflow boundary conditions evaluated at t^{n+s} is denoted by b^{n+c_s}. The Runge-Kutta scheme is defined by the coefficients a_{st}, b_s, c_s and is at most Sth order accurate for S stages.

For the low order method (13), a strong stability preserving (SSP) time discretization [28, 29] should be used to ensure the LED property of the Bernstein coefficients at the fully discrete level. We use the backward Euler scheme because it is very cheap and provides the SSP property without imposing any restrictions on the time step Δt (in contrast to any other time integrator). The first order accuracy of the backward Euler scheme is not an issue as long as it is used as a time stepping method for the low order predictor in the context of FCT. However, it is worthwhile to integrate the vector of inflow boundary conditions b in time using the Runge-Kutta scheme of the high order method. Later on, the small additional effort invested in accurate time integration of the convective fluxes across the inflow boundary will be compensated by a simpler formula for the antidiffusive element contributions of the FCT algorithm. Therefore, the Bernstein coefficients c^L of the low order solution at t^{n+1} are calculated as follows:

$$m_i c_i^L - \Delta t \sum_{j=1}^{N_{\text{dof}}} l_{ij} c_j^L = m_i c_i^n + \Delta t \sum_{s=1}^{S} b_s b_i^{n+c_s} \qquad \text{for all } 1 \le i \le N_{\text{dof}}. \tag{26}$$

Note that the high order Runge-Kutta approximation of b is positivity-preserving only in the case of $b_s \ge 0$ for all $1 \le s \le S$. Otherwise, a negative coefficient b_s can result in $\sum_{s=1}^{S} b_s b_i^{n+c_s} < 0$ even if $b_i \ge 0$ for all $t^n \le t \le t^{n+1}$.

4. Phoenical element-based FCT algorithms

A well-designed flux-corrected transport (FCT) algorithm is supposed to preserve the high order of convergence (in space and time) in regions where the solution is sufficiently smooth, guarantee certain nonlinear stability properties, and be capable of resolving discontinuities in a crisp and non-oscillatory manner. The finite element discretizations presented so far do not satisfy (some of) these design criteria. On the one hand, the low order method produces non-oscillatory solutions which possess a weak LED property. However, the accuracy of this approximation is first order at best by the Godunov theorem [25]. On the other hand, the high order method produces spurious oscillations at discontinuities, which may result in violations of discrete maximum principles. To rectify these deficiencies, FCT blends a given high order scheme and its low order LED counterpart by adjusting the difference between the two approximations in a continuous and conservative manner while taking the local solution behavior into account.

Starting with c^L, the Bernstein coefficients of the high order approximation c^H can be recovered exactly thus:

$$m_i c_i^H = m_i c_i^L + m_i (c_i^H - c_i^L). \tag{27}$$

This relationship can be transformed using (26) and (25) as follows:

$$\begin{aligned}
m_i c_i^H &= m_i c_i^L + m_i c_i^H - m_i c_i^L \underbrace{+ m_i c_i^L - \Delta t \sum_{j=1}^{N_{\text{dof}}} l_{ij} c_j^L - \Delta t \sum_{s=1}^{S} b_s b_i^{n+c_s} - m_i c_i^n}_{=0} \\
&= m_i c_i^L - \Delta t \sum_{j=1}^{N_{\text{dof}}} l_{ij} c_j^L - \Delta t \sum_{s=1}^{S} b_s b_i^{n+c_s} + m_i (c_i^H - c_i^n) \underbrace{- \sum_{j=1}^{N_{\text{dof}}} m_{ij} (c_j^H - c_j^n) + \Delta t \sum_{s=1}^{S} b_s \sum_{j=1}^{N_{\text{dof}}} m_{ij} \dot{c}_j^{n,s}}_{=0} \\
&= m_i c_i^L + \sum_{j=1}^{N_{\text{dof}}} (m_i \delta_{ij} - m_{ij})(c_j^H - c_j^n) - \Delta t \sum_{j=1}^{N_{\text{dof}}} l_{ij} c_j^L + \Delta t \sum_{s=1}^{S} b_s \sum_{j=1}^{N_{\text{dof}}} k_{ij} \Big(c_j^n + \Delta t \sum_{t=1}^{S} a_{st} \dot{c}_j^{n,t} \Big) \\
&= m_i c_i^L + \sum_{j=1}^{N_{\text{dof}}} (m_i \delta_{ij} - m_{ij})(c_j^H - c_j^n) - \Delta t \sum_{j=1}^{N_{\text{dof}}} l_{ij} c_j^L + \Delta t \sum_{j=1}^{N_{\text{dof}}} k_{ij} c_j^n + (\Delta t)^2 \sum_{j=1}^{N_{\text{dof}}} k_{ij} \sum_{s,t=1}^{S} b_s a_{st} \dot{c}_j^{n,t} \\
&= m_i c_i^L + \sum_{j=1}^{N_{\text{dof}}} \Big((m_i \delta_{ij} - m_{ij})(c_j^H - c_j^n) - \Delta t l_{ij} c_j^L + \Delta t k_{ij} c_j^n + (\Delta t)^2 k_{ij} \sum_{s,t=1}^{S} b_s a_{st} \dot{c}_j^{n,t} \Big) \\
&=: m_i c_i^L + \sum_{j=1}^{N_{\text{dof}}} f_{ij} =: m_i c_i^L + f_i
\end{aligned} \tag{28}$$

because $\sum_{s=1}^{S} b_s = 1$ by definition. In the derivation of (28), we assumed the high order approximation of the boundary condition in (26) to remove the dependency of the antidiffusive fluxes f_{ij} on b. If a low order boundary approximation is employed in (26), dependencies on b should be included in the definition of f_{ij} to recover c_i^H.

In edge-based algebraic flux correction schemes of FCT type [35, 36, 44], the antidiffusive correction of the provisional low order solution is rendered local extremum diminishing by limiting the fluxes f_{ij} which describe bilateral mass exchange between pairs of nodes. In the present work, we go back to the roots of FEM-FCT [40, 45, 46] and adopt an element-based limiting strategy taking advantage of the fact that the antidiffusive term

$$f_i = \sum_{j=1}^{N_{\text{elem}}} f_i^e \qquad \text{for all } 1 \le i \le N_{\text{dof}} \tag{29}$$

admits a natural decomposition into element contributions f_i^e. By (28), we have

$$f_{ij} = \sum_{e=1}^{N_{\text{elem}}} f_{ij}^e \qquad \text{for all } 1 \le i, j \le N_{\text{dof}},\ j \ne i, \tag{30}$$

where

$$f_{ij}^e := (m_i^e \delta_{ij} - m_{ij}^e)(c_j^H - c_j^n) - \Delta t l_{ij}^e c_j^L + \Delta t k_{ij}^e c_j^n + (\Delta t)^2 k_{ij}^e \Big(\sum_{s,t=1}^{S} b_s a_{st} \dot{c}_j^{n,t} \Big). \tag{31}$$

The corresponding antidiffusive element contributions are given by

$$f_i^e := \sum_{j=1}^{N_{\text{dof}}} f_{ij}^e \qquad \text{for all } 1 \le i \le N_{\text{dof}}. \tag{32}$$

It is easy to verify that

$$\sum_{i=1}^{N_{\text{dof}}} f_i^e = 0 \quad \text{if } \partial K_e \cap \Gamma_{\text{out}} = \emptyset \tag{33}$$

using the definition of m_i^e and the fact that $\sum_i k_{ij}^e = 0$ by the partition of unity property of the Bernstein basis functions.

FCT algorithms based on such decompositions of $c^H - c^L$ are called *phoenical* [8] because c^H can be restored "like a phoenix" using unconstrained antidiffusive fluxes f_{ij} or element contributions f_i^e in (28) to correct c^L.

In the process of limiting, we will adjust the components of the antidiffusive element vectors $f^e = \{f_i^e\}_{i=1}^{N_{\text{dof}}}$ in a way which preserves the zero sum property (33). The element-based version of FEM-FCT offers better locality properties in the context of high order finite element approximations and enables us to use different time discretizations for the underlying high- and low order schemes without losing the mass conservation property (cf. [38], p. 256). Interestingly enough, it can be interpreted as element-by-element derivative correction for polynomial shape functions followed by a local maximum-preserving projection into the continuous finite element space (see AppendixC).

4.1. Local bounds and maximum principles

In classical element-based FEM-FCT algorithms [45, 46], all components of the element vector $f^e, 1 \le e \le N_{\text{elem}}$ are multiplied by the same solution-dependent correction factor $\alpha^e \in [0, 1]$. This adjustment does not violate the discrete conservation property. In the case $\partial K_e \cap \Gamma_{\text{out}} = \emptyset$, the components of the corrected element vector $\bar{f}^e := \alpha^e f^e$ sum to zero regardless of α^e. However, the value of α^e may influence the redistribution of mass on a local patch of elements consisting of K_e and its neighbors. As shown in [2], using the same correction factor α^e for all components of f^e may not be the best approach to enforcing local bounds in applications of FCT to high order Bernstein polynomials. Higher accuracy can often be attained by limiting the components of f^e or the corresponding subcell contributions in a segregated manner, while making sure that the corrected element vector $\bar{f}^e$ inherits the zero sum property

$$\sum_{i=1}^{N_{\text{dof}}} \bar{f}_i^e = 0 \quad \text{if } \partial K_e \cap \Gamma_{\text{out}} = \emptyset. \tag{34}$$

Examples of localized element-based FCT limiters based on such generalizations will be presented below.

In addition to the discrete conservation property (34), the FCT algorithm for calculating the limited antidiffusive element contributions $\bar{f}_i^e$ should guarantee that the final values of the Bernstein coefficients

$$c_i^{n+1} = c_i^L + \frac{1}{m_i} \sum_{e=1}^{N_{\text{elem}}} \bar{f}_i^e, \qquad 1 \le i \le N_{\text{dof}} \tag{35}$$

satisfy local discrete maximum principles of the form

$$c_i^{\min,n+1} \le c_i^{n+1} \le c_i^{\max,n+1} \qquad \text{for all } 1 \le i \le N_{\text{dof}}. \tag{36}$$

To ensure that these inequality constraints hold at least for $\bar{f}_i^e = 0$, the low order solution c_i^L must lie in the range of admissible values $[c_i^{\min,n+1}, c_i^{\max,n+1}]$. The local bounds $c_i^{\min,n+1}$ and $c_i^{\max,n+1}$ may also depend on the values of c_j^L at other nodes of elements containing node i and/or on the coefficients of the old solution c^n which are supposed to be physically admissible [2, 45, 49]. Whereas the use of c^n in the definition of $c_i^{\min,n+1}$ and $c_i^{\max,n+1}$ may alleviate the *peak clipping* effect [8] by making the bounds less restrictive, it may be inappropriate, e.g., in applications to conservation laws with source or sink terms. For this reason, we abstain from using c^n in constraints for c^{n+1}.

In the context of FCT for high order Bernstein finite elements, we consider local bounds of the form

$$c_i^{\min,n+1} \coloneqq \min_{j \in \mathcal{N}_i^*} c_j^L, \qquad c_i^{\max,n+1} \coloneqq \max_{j \in \mathcal{N}_i^*} c_j^L$$

and call them *element-stencil bounds* if $\mathcal{N}_i^* \coloneqq \mathcal{N}_i$ is the full nodal stencil of the pth order Bernstein approximation, i.e., the search for local maxima and minima of c^L involves all of the nodes belonging to the same element(s) as node i. Following Anderson et al. [2], we also explore the possibility of using just the low order values at node i and its nearest neighbors, i.e, nodes belonging to the same subcells of the Bézier net. As shown in [2], element-based FCT algorithms designed to enforce such *subcell-stencil bounds* do a better job in detecting spurious local extrema that do not violate the inequality constraints (36) formulated in terms of element-stencil bounds.

In the case of linear finite elements, the element-stencil and subcell-stencil bounds coincide and the local maximum principle (36) for the nodal values $c_i^{n+1} = u_i^{n+1}$ implies that no local extrema may arise or grow as a result of the antidiffusive correction. For a Bernstein finite element approximation of arbitrary order and $\mathcal{N}_i^* \subseteq \mathcal{N}_i$, we have

$$c_i^{\min,n+1} \ge \min_{j \in \mathcal{N}_i} c_j^L, \qquad c_i^{\max,n+1} \le \max_{j \in \mathcal{N}_i} c_j^L.$$

Hence, using (5), we can show that the finite element shape functions will stay bounded in terms of the Bernstein coefficients as follows:

$$\begin{aligned} \min_{1 \le j \le N_{\text{dof}}} c_j^L \le \min_{j \in \mathcal{N}^e} c_j^{\min,n+1} \le \min_{j \in \mathcal{N}^e} c_j^{n+1} \le \min_{\boldsymbol{\xi} \in K_e} u_h^{n+1}(\boldsymbol{\xi}) \\ \le u_h^{n+1}(\mathbf{x}) \le \max_{\boldsymbol{\xi} \in K_e} u_h^{n+1}(\boldsymbol{\xi}) \le \max_{j \in \mathcal{N}^e} c_j^{n+1} \le \max_{j \in \mathcal{N}^e} c_j^{\max,n+1} \le \max_{1 \le j \le N_{\text{dof}}} c_j^L \qquad \text{for all } \mathbf{x} \in \Omega. \end{aligned}$$

Additionally, the total variation of u_h^{n+1} will be bounded by the constrained variation of the piecewise-linear Bézier interpolant of c^{n+1} due to the variation diminishing property of Bernstein polynomials [7, 26, 27].

4.2. Global element-stencil limiter

Once the local bounds are defined, a way to compute the antidiffusive element contributions $\bar{f}_i^e$ satisfying the inequality constraints (36) for all nodes of K_e subject to the additional equality constraint (34) must be found. The trivial choice $\bar{f}_i^e = 0$ is clearly not the best option. To benefit from the use of high order finite elements, we must have $\bar{f}_i^e \to f_i^e$ in smooth regions. In particular, an optimal limiter for element contributions of the form

$$\bar{f}_i^e \coloneqq \alpha^e f_i^e \tag{37}$$

would choose the element-based correction factors $\alpha^e \in [0,1]$ to be as close to 1 as possible without violating

$$m_i(c_i^{\min,n+1} - c_i^L) \le \sum_{e=1}^{N_{\text{elem}}} \alpha^e f_i^e \le m_i(c_i^{\max,n+1} - c_i^L) \qquad \text{for all } 1 \le i \le N_{\text{dof}}. \tag{38}$$

To avoid solving constrained optimization problems, some worst-case assumptions are commonly made in practical algorithms for calculating α^e. For example, the following implementation of Zalesak's multidimensional FCT limiter [49] can be used to constrain the sums of positive and negative element contributions as in [38, 40, 45, 46]

$$\begin{aligned}
P_i^+ &= \sum_{e=1}^{N_{\text{elem}}} \max\{0, f_i^e\}, & P_i^- &= \sum_{e=1}^{N_{\text{elem}}} \min\{0, f_i^e\}, \\
Q_i^+ &= m_i(c_i^{\max,n+1} - c_i^L), & Q_i^- &= m_i(c_i^{\min,n+1} - c_i^L), \\
R_i^+ &= \min\Big\{1, \frac{Q_i^+}{P_i^+}\Big\}, & R_i^- &= \min\Big\{1, \frac{Q_i^-}{P_i^-}\Big\}, \\
\alpha^e &= \min_{1\le i\le N_{\text{dof}}} R_i^e, & R_i^e &= \begin{cases} R_i^+ & : f_i^e > 0, \\ 1 & : f_i^e = 0, \\ R_i^- & : f_i^e < 0. \end{cases}
\end{aligned} \tag{39}$$

Assuming that positive and negative contributions of different elements to node i do not cancel out, this FCT limiter imposes an upper bound Q_i^+ on the sum $R_i^+ P_i^+$ and a lower bound Q_i^- on the sum $R_i^- P_i^-$. To enforce these bounds for all nodes of K_e, the element-based correction factor α^e is defined as the minimum of the nodal correction factors R_i^e. Due to the fact that the bounds are imposed on the sums of element contributions and a synchronization of correction factors is performed for all nodes of element K_e, algorithm (39) represents a *global element-stencil limiter*.

4.3. Local element-stencil limiter

In applications of Zalesak's limiter to high order Bernstein finite elements, the number N_{elem} of element contributions to the sums $P_i^\pm$ can be as small as one if the Bernstein basis function B_i attains its maximum at an internal point of K_e and vanishes on the boundary ∂K_e. In this case, the value of the nodal correction factor R_i^e can be calculated without taking the contributions of other elements into account. However, the correction factors for some other nodes still depend on the data in other elements. To perform all computations in a single loop over elements, the FCT limiter may be localized by decomposing the bounds $Q_i^\pm$ into element contributions and constraining $R_i^e f_i^e$ to satisfy

$$m_i^e(c_i^{\min,n+1} - c_i^L) \le R_i^e f_i^e \le m_i^e(c_i^{\max,n+1} - c_i^L) \qquad \text{for all } 1 \le i \le N_{\text{dof}}, \tag{40}$$

where m_i^e denotes the contribution of element K_e to the diagonal entry m_i of the global lumped mass matrix. This decomposition leads to the following simple formula which was originally proposed in [17]:

$$\alpha^e = \min_{1\le i\le N_{\text{dof}}} R_i^e, \qquad R_i^e = \begin{cases} \min\Big\{1, \frac{m_i^e(c_i^{\max,n+1} - c_i^L)}{f_i^e}\Big\} & : f_i^e > 0, \\ 1 & : f_i^e = 0, \\ \min\Big\{1, \frac{m_i^e(c_i^{\min,n+1} - c_i^L)}{f_i^e}\Big\} & : f_i^e < 0. \end{cases} \tag{41}$$

Remarkably, the so-defined *local element-stencil limiter* has the same structure as the vertex-based version [35] of the Barth-Jespersen slope limiter [6] for finite volume and discontinuous Galerkin (DG) methods.

4.4. Nodal-point limiter

The element-based FCT algorithms presented so far apply the same correction factor α^e to all components f_i^e of the antidiffusive element vector f^e to obtain $\bar{f}^e$ defined by (37). This limiting strategy corresponds to simple scaling and automatically provides the (local) mass conservation property (34). However, the straightforward definition of α^e as the minimum of R_i^e may produce very diffusive solutions and/or spurious distortions, especially in applications to

high order finite elements. The reason for this lies in the large number of local degrees of freedom and, hence, many inequality constraints per element. Note that element-stencil limiters must set α^e to zero if a local maximum/minimum is attained at any node where f_i^e is positive/negative. Therefore, it is worthwhile to use different correction factors $\alpha_i^e \le R_i^e$ for different components of f^e but it is essential to guarantee that $\bar{f}_i^e := \alpha_i^e f_i^e$ will satisfy (34).

The choice $\alpha_i^e = R_i^e$ is generally non-conservative. However, the zero-sum property can be easily restored using additional limiting of positive or negative components. The simplest mass correction technique is based on uniform scaling, as proposed by Anderson et al. [2]. This approach leads to the following *nodal-point limiter*

$$\alpha_i^e = \begin{cases} -\frac{\sum_{j=1}^{N_{\text{dof}}} R_j^e \min\{0, f_j^e\}}{\sum_{j=1}^{N_{\text{dof}}} R_j^e \max\{0, f_j^e\}} R_i^e & : f_i^e > 0 \text{ and } \sum_{j=1}^{N_{\text{dof}}} R_j^e f_j^e > 0, \\ -\frac{\sum_{j=1}^{N_{\text{dof}}} R_j^e \max\{0, f_j^e\}}{\sum_{j=1}^{N_{\text{dof}}} R_j^e \min\{0, f_j^e\}} R_i^e & : f_i^e < 0 \text{ and } \sum_{j=1}^{N_{\text{dof}}} R_j^e f_j^e < 0, \\ R_i^e & : \text{otherwise}, \end{cases} \tag{42}$$

where the nodal correction factors R_i^e may be defined as in (39) and (41) to enforce the inequality constraints (36) in the framework of global and local element-based FCT algorithms, respectively.

Formula (42) is designed to produce an element vector $\bar{f}^e$ satisfying the zero sum condition $\sum_{j=1}^{N_{\text{dof}}} \bar{f}_i^e = 0$. According to (33), the components of the original element vector f^e do not necessarily sum to zero if $\partial K_e \cap \Gamma_{\text{out}} \neq \emptyset$. In this case, non-vanishing sums represent antidiffusive corrections of numerical fluxes across Γ_{out}. Even if the solution is smooth and $R_i^e = 1$ for all i, the nodal-point limiter based on (42) would produce $\alpha_i^e < 1$ and correct the fluxes to enforce the zero sum property instead of reproducing the high order solution. This drawback can be avoided using a generalized mass correction procedure which constrains the antidiffusive element contributions to satisfy

$$\min\Big\{0, \sum_{i=1}^{N_{\text{dof}}} f_i^e\Big\} \le \sum_{i=1}^{N_{\text{dof}}} \bar{f}_i^e \le \max\Big\{0, \sum_{i=1}^{N_{\text{dof}}} f_i^e\Big\}. \tag{43}$$

The corresponding generalization of the nodal-point limiter (42) is given by

$$\alpha_i^e = \begin{cases} \frac{q_+ - \sum_{j=1}^{N_{\text{dof}}} R_j^e \min\{0, f_j^e\}}{\sum_{j=1}^{N_{\text{dof}}} R_j^e \max\{0, f_j^e\}} R_i^e & : f_i^e > 0 \text{ and } \sum_{j=1}^{N_{\text{dof}}} R_j^e f_j^e > q_+, \\ \frac{q_- - \sum_{j=1}^{N_{\text{dof}}} R_j^e \max\{0, f_j^e\}}{\sum_{j=1}^{N_{\text{dof}}} R_j^e \min\{0, f_j^e\}} R_i^e & : f_i^e < 0 \text{ and } \sum_{j=1}^{N_{\text{dof}}} R_j^e f_j^e < q_-, \\ R_i^e & : \text{otherwise}, \end{cases} \qquad q_\pm = \begin{matrix} \max \\ \min \end{matrix} \Big\{0, \sum_{j=1}^{N_{\text{dof}}} R_j^e f_j^e\Big\}. \tag{44}$$

For any element K_e, $1 \le e \le N_{\text{elem}}$, this limiting strategy has the property that

$$R_i^e = 1 \qquad \text{for all } 1 \le i \le N_{\text{dof}} \qquad \Rightarrow \qquad \alpha_i^e = 1 \qquad \text{for all } 1 \le i \le N_{\text{dof}}. \tag{45}$$

Hence, the high order solution is recovered if the local maximum principles are satisfied at all nodes of K_e. For all interior elements, the nodal-point limiters based on (42) and (44) are equivalent.

More sophisticated mass correction procedures were developed in [2] using nonlinear penalization depending on $|u_j^L + R_j^e f_j^e - u_j^H|$ to localize subsequent corrections of the prelimited element contributions $R_j^e f_j^e$. Since these nonlinear corrections are also designed to enforce a single equality constraint for the sum of limited antidiffusive components $\alpha_j^e f_j^e$, they do not guarantee local mass conservation for subcell contributions in the case $p > 1$. Hence, non-local mass transfer effects cannot be ruled out, which makes it difficult to justify the high complexity and considerable overhead cost of nonlinear element-stencil mass correction procedures for nodal-point limiters.

4.5. Subcell-stencil limiter

As an alternative to pointwise limiting followed by (non-local) mass corrections, the locality of the limiting procedure can be improved using a decomposition of f^e into N_{sub} subcell contributions $f^{e,m}$ such that

$$\sum_{m=1}^{N_{\text{sub}}} f_i^{e,m} = f_i^e \qquad \text{for all } 1 \le i \le N_{\text{dof}} \tag{46}$$

and

$$\sum_{i=1}^{N_{\text{dof}}} f_i^{e,m} = 0 \quad \text{if } \partial K_{e,m} \cap \Gamma_{\text{out}} = \emptyset. \tag{47}$$

In the context of Bernstein finite elements, the decomposition of K_e into subcells $K_{e,m}$ is defined by the Bézier net triangulation [26]. Let $\mathcal{N}^{e,m}$ denote the set of node numbers associated with the $d+1$ vertices of the simplex $K_{e,m}$. To make sure that $f^{e,m}$ cannot create an undershoot or overshoot at any other node of K_e, we require that

$$f_i^{e,m} = 0 \qquad \text{for all } 1 \le i \le N_{\text{dof}},\ i \notin \mathcal{N}^{e,m}. \tag{48}$$

Given a subcell decomposition satisfying (46)–(48), we multiply each subcell contribution $f_i^{e,m}$ by

$$\alpha^{e,m} = \min_{j \in \mathcal{N}^{e,m}} R_j^e, \tag{49}$$

where the nodal correction factors R_j^e are calculated using a *subcell-stencil* version of Zalesak's limiter (39) to enforce global or local bounds for the sums of positive and negative subcell contributions. This limiting strategy yields

$$\bar{f}_i^e = \sum_{m=1}^{N_{\text{sub}}} \alpha^{e,m} f_i^{e,m} \qquad \text{for all } 1 \le i \le N_{\text{dof}}. \tag{50}$$

In one space dimension, each subcell contribution vector $f^{e,m}$ has at most two non-zero components of equal magnitude and opposite sign. Hence, the total number of subcell degrees of freedom equals N_{sub}. By the zero sum property (33), the original vector f^e has just N_{sub} independent components. Invoking condition (46), we obtain a system of N_{sub} equations for N_{sub} unknown parameters. The unique 1D subcell decomposition is given by

$$f_i^{e,m} = \begin{cases} \sum_{j=1}^{m} f_j^e & : i = m, \\ -\sum_{j=1}^{m} f_j^e & : i = m+1, \\ 0 & : \text{otherwise.} \end{cases} \tag{51}$$

In multidimensions, optimal subcell contributions can be determined by minimizing an objective function like

$$J(f^{e,m}) = \sum_{i=1}^{N_{\text{dof}}} \sum_{m=1}^{N_{\text{sub}}} (f_i^{e,m})^2$$

subject to the equality constraints (46)–(48). The effort invested in the computation of $f_i^{e,m}$ can be justified by the fact that the above definition is intended to minimize the magnitude of the antidiffusive terms to be limited.

The 1D example in Fig. 2 demonstrates that the use of localized (nodal-point or subcell-stencil) limiters may significantly improve the resolving power of FCT algorithms for high order Bernstein finite elements. The nomenclature for the interpretation of different versions denoted by FCT(·,·,·;·), as well as additional numerical examples illustrating the benefits of localized limiting techniques will be presented in Section 6.

4.6. Extremum-preserving limiting

Even localized FCT limiters cannot prevent a loss of accuracy at smooth peaks. As shown in the recent survey by Zhang and Shu [51], methods designed to satisfy local maximum principles are doomed to be at most second-order accurate at local extrema. In the context of DG-FCT algorithms based on high order Bernstein finite elements, the second-order barrier for L^1 errors was confirmed numerically in [2]. Higher accuracy can still be achieved outside of numerical layers even in the presence of discontinuities [14] but it is commonly believed that deactivation of limiters in 'smooth' cells is the only way to achieve optimal convergence for problems with peaks.

Many troubled cell detectors [47, 48] and derivative-based smoothness criteria [19–21, 39] for estimation of local regularity have been proposed in the literature to circumvent order barriers in high order discontinuous Galerkin and

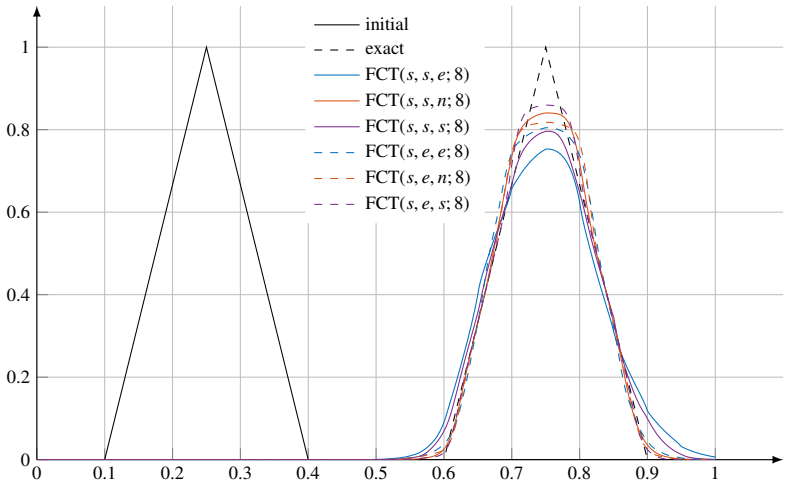


Figure 2: Linear convection in 1D: $\Omega = (0, 1)$, uniform mesh, 20 Bernstein elements of order $p = 8$, continuous Galerkin scheme stabilized by high order background dissipation ($\omega = 0.1$, see Section 5), subcell-stencil upwinding, local element-based FCT using element-stencil (FCT$(s, \cdot, e; 8)$), nodal-point (FCT$(s, \cdot, n; 8)$), and subcell-stencil (FCT$(s, \cdot, s; 8)$) limiters to enforce subcell-stencil (solid) and element-stencil (dashed) bounds.

finite volume methods. The simplest way to distinguish between harmless smooth extrema and dangerous sharp peaks is to look at the second derivatives which provide valuable information about the total variation [26] and nodal error [42] of Bernstein polynomial approximations. As an illustration of the potential benefit of applying these types of methods, we present below a numerical study for 1D linear convection problems, in which we use a second derivative test based on the same criterion as the PPM limiter of Colella and Sekora [16].

To introduce the basic idea in a simple setting, we consider a uniform 1D mesh of quadratic Bernstein finite elements ($p = 2$). To prevent peak clipping while preserving continuous dependence of constrained solutions on the data, the correction factors for the element-stencil and localized FCT limiters are modified as follows:

$$R_i^e = \max\{R_i^e, \gamma^e\}, \tag{52}$$

where γ^e is a smoothness indicator to be defined in terms of the piecewise-constant second derivatives $\eta_e \coloneqq u_e''$.

Following the work of Colella and Sekora [16], we treat an element K_e as smooth if the second derivatives have the same signs and their magnitudes are comparable on the patch $\mathcal{T}_e \coloneqq \bigcup_{e' \in \mathcal{E}_e} K_{e'}$, where the set $\mathcal{E}_e$ contains the numbers of elements that share a node with K_e. Our smoothness indicator is defined by

$$\gamma^e = \frac{\min\{\eta_e^2, C \max\{0, \min_{e' \in \mathcal{E}_e} \eta_e \eta_{e'}\}\}}{\eta_e^2 + \epsilon}, \tag{53}$$

where $C \geq 1$ is a bound for the jumps in the magnitudes of the second derivatives and ϵ is a small positive constant that we add to prevent division by zero. If any pair of second derivatives η_e and $\eta_{e'}$ have opposite signs, we obtain $\gamma^e = 0$. If all signs are the same and $|\eta_e| \leq C|\eta_{e'}|$ for all $e' \in \mathcal{E}_e$, we obtain $\gamma^e = 1$. In all other cases, γ^e is the largest non-negative number satisfying $\gamma^e|\eta_e| \leq C|\eta_{e'}|$ for all $e' \in \mathcal{E}_e$. The so-defined smoothness indicator has a lot in common with the one proposed in [39] and can also be interpreted as a limiter for the slopes of u'.

In the limit $\eta_e \to 0$, definition (53) yields $\gamma^e = 0$. If all second derivatives have the same sign and their magnitudes are comparable to the parameter ϵ, the value of γ^e may exhibit high sensitivity to small fluctuations in η and to round-off errors. However, small magnitudes of the second derivatives imply that the solution is almost linear or almost constant. In the former case, definition (52) yields $\alpha^e = 1$ for any value of γ^e. In the latter case, the magnitude of the antidiffusive components is very small and the value of γ^e hardly makes any difference. Hence, the proposed limiting strategy does preserve the continuous dependence of FCT-constrained solutions on the data.

We remark that Colella and Sekora [16] perform additional checks and use their second-derivative sensor only at extrema. Away from extrema, they use a standard monotonicity-preserving limiter. This approach makes the limiting

process more efficient but lacks continuous dependence on the data, and the underlying criteria may require careful revisions in the context of FCT schemes based on high order Bernstein finite elements.

The $u2$ *detection criteria* developed by Diot et al. [19, 20] for smoothness estimation in MOOD finite volume schemes are based on similar approaches to checking the signs and magnitudes of the (directional) second derivatives. In particular, the final simplified formula proposed in [20] identifies an element as smooth if all signs are the same and the ratio of the largest and smallest second derivative is bounded by 2. This corresponds to (53) with $C = 2$ and η_e replaced by the greatest-in-magnitude second derivative. Detailed numerical studies were performed in [19, 20] to demonstrate the usefulness of this second derivative test in the context of high order finite volume methods.

In the case of a higher-order ($p > 2$) finite element approximation, the second derivatives are generally not constant inside each element. Moreover, the shape function corresponding to a high order Bernstein element may have multiple internal peaks. Some of them may be smooth, whereas others may represent spurious local extrema. Hence, a local second derivative test may be devised for each node of the Bézier net to determine the optimal values of the nodal correction factors for smooth and sharp peaks. In the current implementation of our extrema-preserving FCT limiter, we use a straightforward generalization of formula (53). If the shape function is a polynomial of degree $p > 2$, we perform a local L^2-projection into the space of quadratic polynomials defined on K_e. The projection corresponds to 'inverting' the element mass matrix of the local P_2 approximation and yields the best least squares fit $\tilde{u}_h^e \in P_2(K_e)$. The constant second derivatives of $\tilde{u}_h^e$ are used in the smoothness indicator γ^e defined by (53).

In multidimensions, the second derivative test based on formula (53) may be performed for each coordinate direction, as in the work of Diot et al. [19, 20]. Alternatively, the local smoothness indicator γ^e may be defined in terms of the scalar quantity $\eta_e = \det H_e$, where H_e stands for the element Hessian matrix. This generalization is motivated by the fact that sufficient conditions for local extrema depend on the sign of the Hessian determinant. In our 2D numerical study, we use (53) with $\eta_e = \det H_e$ and choose $\mathcal{E}_e$ to be the set of all common-node neighbors for safety reasons.

5. High order dissipation

The element-based FCT algorithms presented so far use the high order Galerkin discretization of the transport equation as a target scheme and constrain its antidiffusive part so as to enforce local maximum principles formulated in terms of the non-oscillatory low order solution. The use of tight bounds makes it more difficult to avoid the peak clipping effect. On the other hand, if the range of admissible values is too wide, constrained solutions may exhibit another kind of spurious behavior which is called *terracing* in the FCT literature. The main reason for this side effect, which has plagued FCT since its inception [8], is poor phase accuracy of the high order scheme.

Terraces are spurious distortions that represent 'an integrated, nonlinear effect of residual phase errors' [50] and can be interpreted as 'ghosts of departed ripples' [8]. In FCT algorithms based on linear finite elements, terracing can be avoided by *prelimiting* the antidiffusive fluxes or element contributions [36, 37]. However, the only real cure is to use a high order scheme with better phase properties [50] and, hence, lower levels of numerical dispersion.

The standard Galerkin method is highly dispersive and tends to produce ripples even in smooth regions of the exact solution [37]. Ripples that do not violate local bounds may survive the limiting step and become terraces. In our experience, this is less likely to happen if high order Bernstein elements are employed. In contrast to linear finite elements, distortions caused by terracing are typically rather small and may look harmless. However, the lack of inherent numerical dissipation can give rise to non-local pollution errors and degrades the rates of convergence.

To achieve optimal convergence rates for problems with smooth solutions and reduce phase errors, a high order dissipative component must be included in the antidiffusive fluxes or element contributions [37, 44, 50]. In particular, the accuracy and convergence behavior of an FCT algorithm based on a Bernstein finite element approximation can be improved by adding high order background dissipation to the variational form of the Galerkin scheme or directly to the antidiffusive components to be limited. Then the deactivation of limiters will produce a high order solution corresponding to a stabilized finite element scheme with good phase properties.

For example, the streamline upwind Petrov-Galerkin (SUPG) method [11] is a better target scheme for FCT than

the standard Galerkin discretization. Adding the SUPG stabilization term to (8), we obtain

$$\int_\Omega w_h \frac{\partial u_h}{\partial t}\,\mathrm{d}\mathbf{x} - \int_\Omega \mathrm{grad}(w_h)\cdot(\mathbf{v}u_h)\,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\text{out}}} w_h u_h \mathbf{v}\cdot\mathbf{n}\,\mathrm{d}\mathbf{s} + \int_{\Gamma_{\text{in}}} w_h u_{\text{in}}\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}\mathbf{s} + \int_\Omega \tau\,(\mathbf{v}\cdot\mathrm{grad}(w_h))\Big(\frac{\partial u_h}{\partial t} + \mathrm{div}(\mathbf{v}u_h)\Big)\,\mathrm{d}\mathbf{x} = 0. \tag{54}$$

The value of the stabilization parameter $\tau > 0$ can be taken as a constant in each element and must be chosen carefully to generate the right amount of high order background dissipation. The residual-based SUPG approach introduces streamline diffusion in a manner which preserves the consistency of the variational formulation. In contrast to the standard Galerkin method, the resulting system matrix is non-symmetric even in the case of an explicit time discretization. Hence, the corresponding linear system is much more expensive to solve. Moreover, the modified mass matrix depends on the velocity $\mathbf{v}$ and must be updated after each iteration/time step if $\mathbf{v}$ is time-dependent.

The use of modified mass matrices can be avoided by using another consistent approximation of the time derivative $\partial_t u_h$ in the SUPG stabilization term. Following the approach proposed in [37] for handling time derivatives in antidiffusive element contributions, we approximate $\partial_t u_h$ by $g_h \approx -\mathrm{div}(\mathbf{v}u_h)$ using the fact that $\partial_t u = -\mathrm{div}(\mathbf{v}u)$ for the exact solution u of the transport equation. This approximation leads to

$$\begin{aligned}\int_\Omega w_h \frac{\partial u_h}{\partial t}\,\mathrm{d}\mathbf{x} - \int_\Omega \mathrm{grad}(w_h)\cdot(\mathbf{v}u_h)\,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\text{out}}} w_h u_h \mathbf{v}\cdot\mathbf{n}\,\mathrm{d}\mathbf{s} + \int_{\Gamma_{\text{in}}} w_h u_{\text{in}}\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}\mathbf{s} \\ + \int_\Omega \tau\,(\mathbf{v}\cdot\mathrm{grad}(w_h))\Big(g_h + \mathrm{div}(\mathbf{v}u_h)\Big)\,\mathrm{d}\mathbf{x} = 0, \\ \int_\Omega w_h\,(g_h + \mathrm{div}(\mathbf{v}u_h))\,\mathrm{d}\mathbf{x} = 0,\end{aligned} \tag{55}$$

where we have defined the approximate time derivative g_h as the L^2 projection of $-\mathrm{div}(\mathbf{v}u_h)$.

Even though discretizations (54) and (55) are based on the same continuous variational formulation, they are not equivalent due to different treatments of the time derivative in the stabilization term. The use of g_h in lieu of $\partial_t u_h$ preserves the symmetry of the mass matrix and facilitates the use of explicit time integrators. At each step of the forward Euler scheme or stage of an explicit SSP Runge-Kutta scheme [29], we have to solve two systems with the standard Galerkin mass matrix instead of one system with the non-symmetric SUPG mass matrix.

A closely related approach to high order stabilization follows the design philosophy of two-level variational multiscale (VMS) methods [32, 41]. Using a Laplacian-type stabilization term defined in terms of the consistent gradient $\mathrm{grad}(u_h)$ and a continuous L^2 projection $\mathbf{g}_H$ of $-\mathrm{grad}(u_h)$, we obtain the high order scheme

$$\begin{aligned}\int_\Omega w_h \frac{\partial u_h}{\partial t}\,\mathrm{d}\mathbf{x} - \int_\Omega \mathrm{grad}(w_h)\cdot(\mathbf{v}u_h)\,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\text{out}}} w_h u_h \mathbf{v}\cdot\mathbf{n}\,\mathrm{d}\mathbf{s} + \int_{\Gamma_{\text{in}}} w_h u_{\text{in}}\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}\mathbf{s} \\ + \int_\Omega \epsilon_{\text{add}}\mathrm{grad}(w_h)\cdot\mathrm{grad}(u_h)\,\mathrm{d}\mathbf{x} + \int_\Omega \epsilon_{\text{add}}\mathrm{grad}(w_h)\cdot\mathbf{g}_H\,\mathrm{d}\mathbf{x} = 0, \\ \int_\Omega w_H\,(\mathbf{g}_H + \mathrm{grad}(u_h))\,d\mathbf{x} = 0,\end{aligned} \tag{56}$$

where the amount of added dissipation depends on the residual $\mathbf{g}_H + \mathrm{grad}(u_h)$ and on the value of the parameter $\epsilon_{\text{add}} \geq 0$ which is again constant in each element. In the context of two-level VMS methods [32, 41], the test function w_H and the L^2 projection $\mathbf{g}_H$ are finite element functions defined on a coarser mesh or using a reduced order approximation. In this paper, we use the same finite element space for functions with subscripts h and H.

Compared to (55), we have to solve one additional system in the two-dimensional space. Note that (55) and (56) have the same structure but the former method is stabilized using $g_h \approx -\mathrm{div}(\mathbf{v}u_h)$, whereas the latter uses $\mathbf{g}_H \approx -\mathrm{grad}(u_h)$. In contrast to SUPG, the stabilization term of the high order scheme (56) adds diffusion not only along the streamlines but also in all crosswind directions. Then the coarse scale diffusivity is removed using the antidiffusive term depending on $\mathbf{g}_H$. Similar two-scale decompositions are used to design variationally consistent high order artificial diffusion operators in local projection stabilization (LPS) schemes [10, 37].

It is worth mentioning that $g_h = \mathbf{v} \cdot \mathbf{g}_h$ in the case of a constant velocity field $\mathbf{v}$. In the one-dimensional space, the stabilization terms of the two methods are related by

$$\int_\Omega \tau \left(\mathbf{v} \cdot \operatorname{grad}(w_h)\right)\Big(g_h + \operatorname{div}(\mathbf{v}u_h)\Big)\,\mathrm{d}\mathbf{x} \overset{\text{1D}}{=} \int_\Omega \tau(\mathbf{v} \cdot w_h')\Big(\mathbf{v} \cdot \mathbf{g}_h + \mathbf{v} \cdot u_h'\Big)\,\mathrm{d}x = \int_\Omega (\tau\|\mathbf{v}\|^2)w_h'\Big(\mathbf{g}_h + u_h'\Big)\,\mathrm{d}x. \tag{57}$$

Hence, the one-dimensional versions of (55) and (56) are equivalent for $\epsilon_{\text{add}} = \tau\|\mathbf{v}\|^2$.

In the following, we focus on the gradient-based high order dissipation defined in (56). Nevertheless, the relationship to SUPG methods may be exploited to define the parameter $\epsilon_{\text{add}} = \tau\|\mathbf{v}\|^2$ depending on τ. For example, the following formula based on the work of John and Schmeyer [33] seems to be a good choice for our purposes:

$$\epsilon_{\text{add}} = \omega\|\mathbf{v}\|^2_{L^\infty(K_e)}\tau = \omega\|\mathbf{v}\|^2_{L^\infty(K_e)}\frac{h_e}{2p\|\mathbf{v}\|_{L^\infty(K_e)}} = \omega\frac{\|\mathbf{v}\|_{L^\infty(K_e)}h_e}{2p}, \tag{58}$$

where h_e is an appropriate measure of the size of element K_e and $\omega \in [0, 1]$ is a constant scaling parameter. The choice $\omega = 0$ corresponds to the unstabilized high order Galerkin discretization.

For linear finite elements in the one-dimensional space, the numerical diffusion coefficient of the first order discrete upwinding scheme is $d = \frac{|v|h_e}{2}$ [36, 40]. This value corresponds to the above choice of ϵ_{add} with $p = 1$ and $\omega = 1$.

Discretizing the stabilized Galerkin method (56) in time using a Runge-Kutta scheme, we obtain

$$c^H = c^n + \Delta t \sum_{s=1}^{S} b_s \dot{c}^{n,s}, \tag{59a}$$

$$\sum_{j=1}^{N_{\text{dof}}} m_{ij}\dot{c}_j^{n,s} = \sum_{j=1}^{N_{\text{dof}}}(k_{ij} - s_{ij})\Big(c_j^n + \Delta t \sum_{t=1}^{S} a_{st}\dot{c}_j^{n,t}\Big) - \mathbf{k}_{ij} \cdot \Big(\mathbf{g}_j^n + \Delta t \sum_{t=1}^{S} a_{st}\dot{\mathbf{g}}_j^{n,t}\Big) + b_j^{n+c_s} \qquad \text{for all } 1 \le i \le N_{\text{dof}}, \tag{59b}$$

where $\mathbf{g}_j^n$ and $\dot{\mathbf{g}}_j^{n,t}$ are the coefficients of the involved L^2 projections. The coefficients s_{ij} and $\mathbf{k}_{ij}$ are defined by

$$s_{ij} = \sum_{e=1}^{N_{\text{elem}}} s_{ij}^e, \qquad s_{ij}^e = \epsilon_{\text{add}}^e \int_{K_e} \operatorname{grad}(B_i) \cdot \operatorname{grad}(B_j)\,\mathrm{d}\mathbf{x}, \tag{60a}$$

$$\mathbf{k}_{ij} = \sum_{e=1}^{N_{\text{elem}}} \mathbf{k}_{ij}^e, \qquad \mathbf{k}_{ij}^e = \epsilon_{\text{add}}^e \int_{K_e} \operatorname{grad}(B_i)B_j\,\mathrm{d}\mathbf{x}. \tag{60b}$$

A straightforward calculation shows that the difference between the stabilized high order solution c^H and the low order predictor c^L defined as before can be decomposed into the element-based antidiffusive fluxes

$$\begin{aligned} f_{ij}^e = {} & (m_i^e\delta_{ij} - m_{ij}^e)(c_j^H - c_j^n) - \Delta t l_{ij}^e c_j^L + \Delta t k_{ij}^e c_j^n + (\Delta t)^2 k_{ij}^e\Big(\sum_{s,t=1}^{S} b_s a_{st}\dot{c}_j^{n,t}\Big) \\ & - (\Delta t)^2 s_{ij}^e c_j^n - (\Delta t)^2 s_{ij}^e\Big(\sum_{s,t=1}^{S} b_s a_{st}\dot{c}_j^{n,t}\Big) - (\Delta t)^2 \mathbf{k}_{ij}^e \cdot \mathbf{g}_j^n - (\Delta t)^2 \mathbf{k}_{ij}^e \cdot \Big(\sum_{s,t=1}^{S} b_s a_{st}\dot{\mathbf{g}}_j^{n,t}\Big) \end{aligned} \tag{61}$$

and the corresponding antidiffusive element contributions $f_i^e = \sum_{j=1}^{N_{\text{dof}}} f_{ij}^e$ have zero net mass, i.e., $\sum_{i=1}^{N_{\text{dof}}} f_i^e = 0$.

6. Numerical experiments

FCT algorithms distinguish themselves from conventional stabilization techniques for finite elements in that optimal accuracy and order of convergence can potentially be achieved in a manner which rules out formation of undershoots and/or overshoots. To assess the ability of the presented algorithms to achieve this goal, we apply them to several one- and two-dimensional problems, which represent a rich variety of interesting applications and are commonly used for testing purposes. In the one-dimensional case, the below numerical results confirm that high order

of convergence can be preserved for linear advection of smooth initial data. Snapshots of numerical solutions are included to illustrate the reduction of terracing due to the use of high order dissipation, as well as the preservation of smooth extrema by FCT schemes equipped with smoothness indicators. In the two-dimensional case, the focus is on comparing different components of the algorithm in the context of the solid body rotation benchmark [43, 49]. The pros and cons of each approach are discussed and suggestions regarding the choice of FCT ingredients are made.

To shorten and unify the description of the methods under investigation, the FCT schemes to be considered in this numerical study will be referred to using abbreviations of the form

$$\text{FCT}(\text{up},\text{bo},\text{li};p), \qquad \text{where} \qquad \begin{cases} \text{up} \in \begin{cases} e & \text{: Element-stencil upwinding,} \\ s & \text{: Subcell-stencil upwinding,} \\ s_r & \text{: Subcell-stencil Rusanov upwinding,} \end{cases} \\ \text{bo} \in \begin{cases} s & \text{: Subcell-stencil bounds,} \\ e & \text{: Element-stencil bounds,} \end{cases} \\ \text{li} \in \begin{cases} n & \text{: Nodal limiting \& mass correction,} \\ e & \text{: Local element-stencil limiting,} \end{cases} \\ p \quad \text{: order of Bernstein finite elements.} \end{cases} \tag{62}$$

In numerical studies of FCT components, the underlying low and high order methods are abbreviated by

$$\text{LO}(\text{up};p) \qquad \text{and} \qquad \text{HI}(p), \tag{63}$$

respectively. All FCT solutions are obtained using limiters that constrain antidiffusive contributions of different elements to the same node independently using bounds proportional to diagonal entries of the corresponding lumped element mass matrices instead of constraining the sums of positive and negative contributions using bounds proportional to diagonal entries of the global lumped mass matrix as in the classical version (39) of Zalesak's limiter.

6.1. One-dimensional case

The below one-dimensional studies are mainly presented to verify the expected order of convergence in the case of smooth solutions. To avoid disturbing influences of the time discretization, the time integrator of the high order method is given by a six-stage diagonally implicit SSP Runge-Kutta scheme, which is of sixth order in time. The coefficients of this scheme are listed in the following Butcher tableau:

0						
0.61341879	0.30670939	0.30670939				
0.20080549	0.10040277	0.00000001	0.10040270			
0.42625809	0.00707081	0.00074821	0.39021074	0.02822831		
0.70805109	0.10486545	0.00007338	0.16278800	0.33300199	0.10732225	
0.95080507	0.01535263	0.00000764	0.41584884	0.04023172	0.43016927	0.04919492
	0.06494390	0.00222959	0.27072380	0.20524931	0.31965775	0.13719562

(64)

The experimental order of convergence (EOC) is determined using the formula [43]

$$p = \log\Big(\frac{E_1(h_2)}{E_1(h_1)}\Big)\log\Big(\frac{h_2}{h_1}\Big)^{-1}, \tag{65}$$

where h_1, h_2 and $E_1(h_1), E_1(h_2)$ are the mesh sizes and L^1-errors on two different meshes, respectively. Here, the L^1-error is based on the difference of the numerical solution and the exact solution. In the two-dimensional space, the L^∞-error is abbreviated by E_∞.

In the first test case, the function $u_t(x,t) = \tanh\big(10(x - vt - \frac{1}{4})\big)$ is convected on the interval $\Omega = (0,1)$ with the constant velocity $v = 1$. This benchmark is characterized by the absence of local extrema and, hence, the bounds should not constrain the choice of the correction factors and no peak clipping effects should occur. The L^1-errors at

N_{elem}	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC
19	1.41e−1		9.50e−2		7.10e−2		5.65e−2		4.68e−2	
27	1.12e−1	0.66	7.32e−2	0.74	5.37e−2	0.79	4.22e−2	0.83	3.47e−2	0.85
38	8.77e−2	0.70	5.59e−2	0.79	4.04e−2	0.84	3.14e−2	0.87	2.56e−2	0.89
53	6.83e−2	0.75	4.24e−2	0.83	3.02e−2	0.87	2.33e−2	0.90	1.89e−2	0.91
74	5.24e−2	0.80	3.19e−2	0.86	2.25e−2	0.89	1.72e−2	0.91	1.39e−2	0.92

(a) LO$(s;\cdot)$.

N_{elem}	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC
19	5.13e−3		4.23e−4		9.14e−5		1.54e−6		1.27e−6	
27	1.60e−3	3.32	1.80e−4	2.44	6.26e−6	7.63	3.49e−7	4.23	6.59e−8	8.42
38	5.53e−4	3.10	8.66e−5	2.13	8.54e−7	5.83	8.73e−8	4.05	1.85e−9	10.45
53	2.51e−4	2.38	4.20e−5	2.18	1.93e−7	4.47	2.25e−8	4.08	2.00e−10	6.69
74	1.22e−4	2.16	1.98e−5	2.25	4.56e−8	4.32	5.24e−9	4.36	2.47e−11	6.26

(b) FCT$(s, s, n;\cdot)$, $\omega = 0.0$.

N_{elem}	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC
19	4.92e−3		5.03e−4		2.64e−5		1.26e−6		1.86e−7	
27	1.59e−3	3.20	1.93e−4	2.73	3.05e−6	6.14	2.70e−7	4.38	1.12e−8	7.99
38	6.01e−4	2.86	7.35e−5	2.82	7.00e−7	4.31	5.95e−8	4.42	1.23e−9	6.49
53	2.68e−4	2.42	3.00e−5	2.69	1.75e−7	4.16	1.36e−8	4.45	1.57e−10	6.18
74	1.28e−4	2.21	1.27e−5	2.58	4.49e−8	4.08	2.91e−9	4.61	2.00e−11	6.17

(c) FCT$(s, s, n;\cdot)$, $\omega = 0.1$.

N_{elem}	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC
19	1.48e−2		4.58e−4		2.71e−5		1.60e−6		1.04e−7	
27	6.27e−3	2.45	1.11e−4	4.04	6.39e−6	4.11	2.22e−7	5.61	1.52e−8	5.48
38	2.45e−3	2.75	3.54e−5	3.35	1.55e−6	4.16	3.74e−8	5.22	2.19e−9	5.66
53	9.39e−4	2.88	1.24e−5	3.15	3.91e−7	4.13	6.61e−9	5.21	3.27e−10	5.72
74	3.53e−4	2.93	4.45e−6	3.07	9.97e−8	4.09	1.19e−9	5.13	4.58e−11	5.89

(d) FCT$(s, s, n;\cdot)$, $\omega = 1.0$.

Table 2: Linear convection in 1D: $u_t(x,t) = \tanh\left(10(x - vt - \frac{1}{4})\right)$, $v = 1$, $T = \frac{1}{2}$, $\Delta t = 10^{-2}(2p+1)^{-1}h$, convergence behavior of LO$(s;\cdot)$ and of FCT$(s, s, n;\cdot)$ with different amounts of high order dissipation.

N_{elem}	$p=2$		$p=3$		$p=4$		$p=5$		$p=6$	
	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC
10	6.07e−2		4.69e−2		4.48e−2		3.19e−2		2.70e−2	
14	3.29e−2	1.83	2.73e−2	1.61	2.35e−2	1.92	1.83e−2	1.66	1.51e−2	1.73
19	2.03e−2	1.57	1.81e−2	1.34	1.44e−2	1.60	1.03e−2	1.86	8.43e−3	1.91
27	1.09e−2	1.78	9.38e−3	1.87	6.87e−3	2.11	5.35e−3	1.87	3.98e−3	2.14
38	5.30e−3	2.10	4.84e−3	1.94	3.29e−3	2.15	2.65e−3	2.06	1.84e−3	2.25
53	2.60e−3	2.15	2.51e−3	1.98	1.66e−3	2.05	1.26e−3	2.22	8.65e−4	2.27
74	1.28e−3	2.13	1.26e−3	2.05	8.48e−4	2.01	6.03e−4	2.22	4.01e−4	2.30

(a) FCT$(s, s, n; \cdot)$, $\omega = 0.1$ without SI.

N_{elem}	$p=2$		$p=3$		$p=4$		$p=5$		$p=6$	
	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC	L^1	EOC
10	5.71e−2		4.54e−2		4.28e−2		3.19e−2		2.70e−2	
14	2.33e−2	2.67	1.76e−2	2.81	1.62e−2	2.89	1.42e−2	2.40	1.11e−2	2.64
19	8.93e−3	3.13	6.24e−3	3.40	6.23e−3	3.13	5.58e−3	3.05	3.96e−3	3.38
27	8.17e−4	6.81	9.57e−4	5.34	5.96e−4	6.68	5.37e−4	6.66	4.16e−4	6.42
38	1.38e−4	5.21	2.12e−6	17.89	1.90e−7	23.55	3.50e−9	34.94	2.16e−10	42.34
53	6.29e−5	2.36	5.25e−7	4.19	4.26e−8	4.49	4.56e−10	6.12	2.45e−11	6.54
74	2.79e−5	2.44	1.35e−7	4.07	9.14e−9	4.61	5.91e−11	6.12	2.65e−12	6.66

(b) FCT$(s, s, n; \cdot)$, $\omega = 0.1$ with SI.

Table 3: Linear convection in 1D: $u_g(x,t) = \exp(-100 \cdot (x - vt - \frac{1}{4})^2)$, $v = 1$, $T = \frac{1}{2}$, $\Delta t = 10^{-2}(2p+1)^{-1}h$, convergence behavior of FCT$(s, s, n; \cdot)$ with and without the optional use of the extremum-preserving smoothness indicator (SI) based on the second derivative test (53) with $C = 2$.

the final time $T = \frac{1}{2}$ and the experimental orders of convergence of different numerical approaches are presented in Table 2. While the low order method using subcell-stencil upwinding converges with an order of approximately 0.9 (see Table 2a), the FCT methods are able to reproduce the high experimental order of convergence (see Tables 2b, 2c, and 2d). If no background dissipation is added to stabilize the high order scheme, the EOC is approximately p if p is even and $p + 1$ otherwise (see Table 2b). This convergence behavior is consistent with the EOC of the unconstrained Galerkin method in the case of pure convection equations. Hence, the FCT algorithm is able to reproduce the expected order of convergence of the high order Galerkin discretization. By applying stabilization terms (see Tables 2c and 2d), the order can be increased to $p + 1$ even in the case of an even Bernstein order p.

If u_t is replaced by a Gaussian function $u_g(x,t) = \exp(-100 \cdot (x - vt - \frac{1}{4})^2)$, then a smooth local extremum occurs inside the domain $\Omega = (0, 1)$ and the nodal correction factor for the degree of freedom corresponding to the local extremum is set to zero. Therefore, the FCT algorithm fails to reproduce the smooth high order solution and the experimental order of convergence reduces to 2 (see Table 3a). Furthermore, peak clipping effects can be observed in the numerical solution. Zhang and Shu [51] have shown that this reduction of the order of convergence cannot be avoided while preserving the LED property. The use of the smoothness indicator introduced in Section 4.6 makes it possible to disable the FCT limiter in smooth regions, where unnecessary limiting of the antidiffusive element contributions would occur otherwise. Table 3b presents the corresponding L^1-errors depending on the mesh size h and the Bernstein order p. As expected, the EOC increases as soon as the smoothness indicator identifies the local extremum as a smooth maximum ($h \leq 0.037$). Therefore, in the case of the Gaussian function the smoothness indicator performs quite well and detects the smooth extremum if the mesh size is small enough. Although an FCT algorithm equipped with such a smoothness indicator does not guarantee that the antidiffusive correction is LED and/or variation

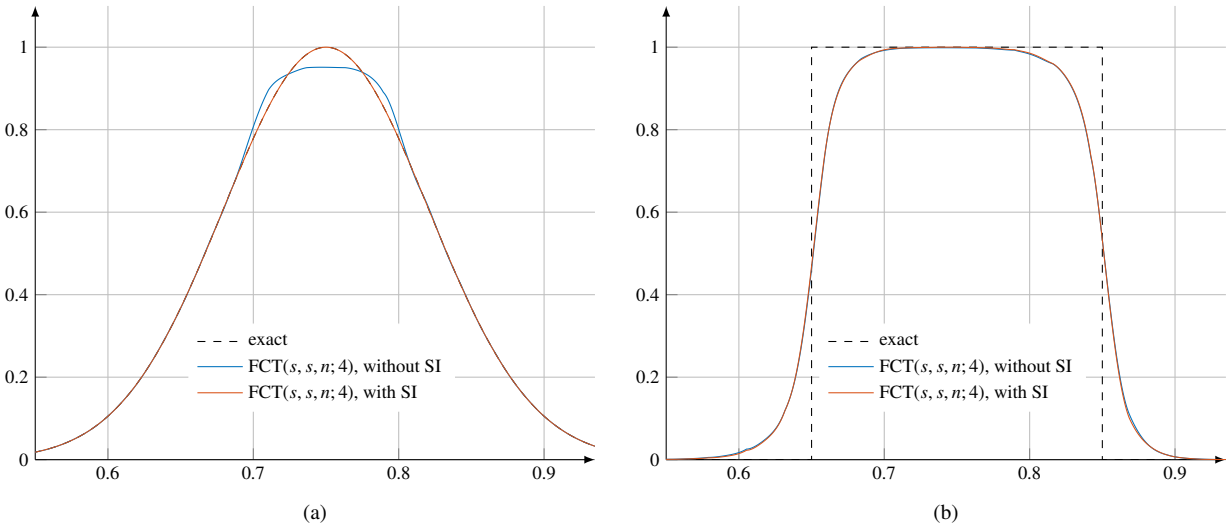


Figure 3: Linear convection in 1D: $\Omega = (0, 1)$, uniform mesh, 38 Bernstein elements, comparison of the FCT$(s, s, n; 4)$ results for (a) smooth and (b) discontinuous data obtained without and with the optional use of the extremum-preserving smoothness indicator (SI) based on the second derivative test (53) with $C = 2$.

diminishing, the proposed second derivative test activates the limiter in regions where oscillations are detected. The numerical solutions displayed in Fig. 3 indicate that selective limiting based on (52) and (53) with $C = 2$ does an excellent job in preserving smooth peaks while enforcing local maximum principles at discontinuities.

The snapshots presented in Fig. 4 indicate that not only a good limiting strategy but also the use of high order dissipation may reduce numerical errors caused by the combined effect of clipping and terracing which is particularly strong if the element-stencil limiting strategy is adopted. The nodal-point version produces just a moderate amount of terracing, and the amount of peak clipping can be significantly reduced by adding high order stabilization. In this one-dimensional example, the best results are obtained using $\omega = 1.0$ (see the right diagram in Fig. 4).

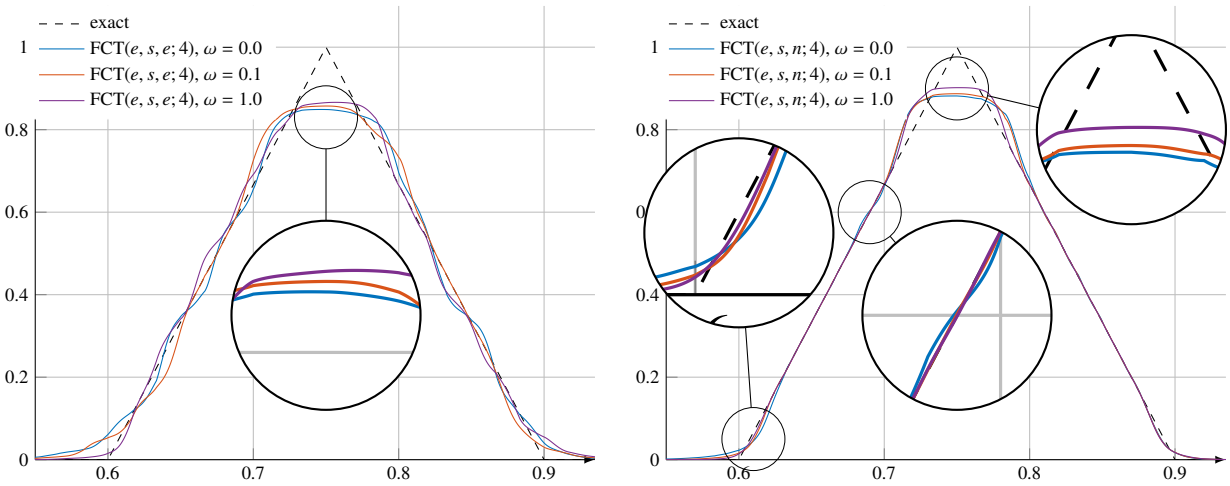
6.2. Two-dimensional case

The above numerical study of FCT algorithms in 1D was focused on investigations of the experimental order of convergence for different schemes. Additionally, the influence of high order dissipation and the use of smoothness indicators were illustrated by numerical examples. In two dimensions, no grid convergence studies are performed and the implications of using high order background dissipation are discussed just briefly. The main focus of this section is on a comparison of different components of the proposed FCT schemes in 2D. In the two-dimensional case, the differences between the performance of different upwinding and limiting techniques become more remarkable due to the increasing number of degrees of freedom and their arrangement on a high order Bernstein finite element.

Since no grid convergence studies are performed in this section, the cost of calculating the high order solution is reduced by using the classical four-stage explicit Runge-Kutta scheme to discretize in time. The Butcher tableau of this fourth order accurate time stepping method is as follows:

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{66}$$

Computations are performed on uniform triangular meshes using approximately the same number of degrees of freedom ($16129 \leq N_{\text{dof}} \leq 17161$) for all Bernstein orders p and the time step 10^{-3} which is sufficiently small to keep temporal errors negligible compared to space discretization errors in numerical solutions for the above range of N_{dof}.



(a) (b)

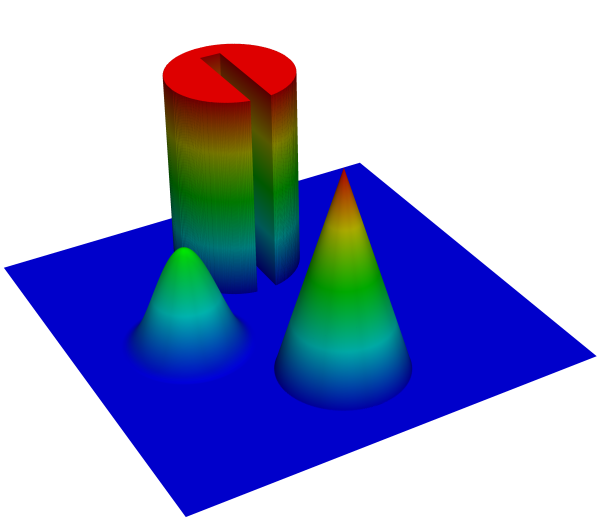
Figure 4: Linear convection in 1D: $\Omega = (0, 1)$, uniform mesh, 60 Bernstein elements, comparison of the FCT$(e, s, \cdot; 4)$ results obtained using the local (a) element-stencil and (b) nodal-point limiters to constrain antidiffusive element contributions with and without high order stabilization.

To discuss the influence of different components on the overall accuracy of FCT algorithms, we consider the solid body rotation benchmark [43, 49]. In this two-dimensional test, a smooth hump, a sharp cone, and a slotted cylinder are rotated around the center of the domain $\Omega = (0, 1)^2$ counterclockwise with the velocity $\mathbf{v} = (\frac{1}{2} - y, x - \frac{1}{2})^\top$. The initial condition is given by

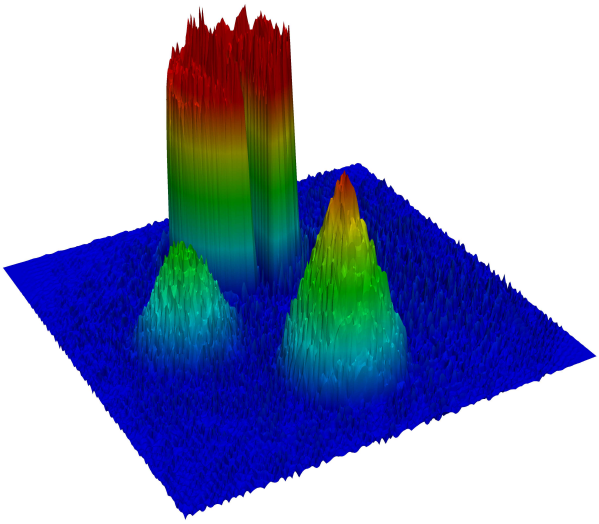
$$u_0(x,y) = \begin{cases} \frac{1}{4} + \frac{1}{4}\cos\left(\frac{\pi\sqrt{(x-0.25)^2+(y-0.5)^2}}{0.15}\right) & \text{if } \sqrt{(x-0.25)^2+(y-0.5)^2} \le 0.15, \\ 1 - \frac{\sqrt{(x-0.5)^2+(y-0.25)^2}}{0.15} & \text{if } \sqrt{(x-0.5)^2+(y-0.25)^2} \le 0.15, \\ 1 & \text{if } \begin{cases} \sqrt{(x-0.5)^2+(y-0.75)^2} \le 0.15, \\ 0.55 < x < 0.45,\ y > 0.85, \end{cases} \\ 0 & \text{otherwise.} \end{cases}$$

The errors listed in the tables and snapshots shown in the figures correspond to the results after one full rotation. The exact solution at the final time $T = 2\pi$ coincides with the initial condition (see Fig. 5a). The high order Galerkin method without stabilization (see Fig. 5b) produces large overshoots/undershoots in the neighborhood of the slotted cylinder and smaller high-frequency oscillations in the rest of the domain. After adding a moderate amount of high order dissipation ($\omega = 0.1$) to the standard Galerkin scheme, these spurious oscillations decrease in magnitude and frequency. Furthermore, undershoots and overshoots are localized to small layers around discontinuities, which has a positive effect on the smoothness and accuracy of numerical solutions (see Fig. 5c and 5d). Adding more high order dissipation than necessary to eliminate spurious oscillations ($\omega = 1.0$) results in stronger erosion at the top of the cone. Later on, this solution behavior can also be observed in the case of FCT approximations.

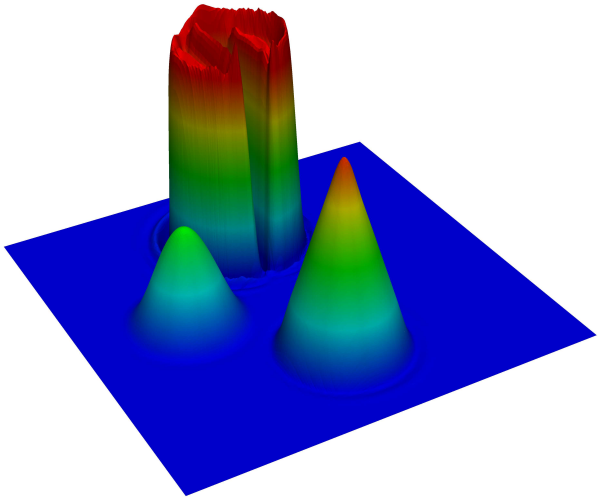
Figure 6 compares the results produced by different low order methods for Bernstein orders $p = 2, 4$. All of the low order methods under investigation add so much artificial diffusion that the slot of the cylinder disappears and the smooth hump is almost leveled to the ground. However, subcell-stencil upwinding is less diffusive than element-stencil upwinding and preserves the maxima of the cylinder and cone much better, especially in the case $p = 4$. Unfortunately, the subcell-stencil approximation is less smooth and exhibits some kinks on the boundaries of mesh cells. The lack of smoothness can be cured using the Rusanov-type subcell-stencil upwinding which smoothes the ripples away by introducing additional artificial diffusion. The resulting low order solution looks even smoother than the one obtained using element-stencil upwinding. Furthermore, the errors listed in Table 1 show that the accuracy of subcell-stencil algorithms is independent of the Bernstein order p, whereas the use of element-stencil upwinding results in stronger smearing of the solution profiles as p is increased while keeping N_{dof} (almost) fixed.



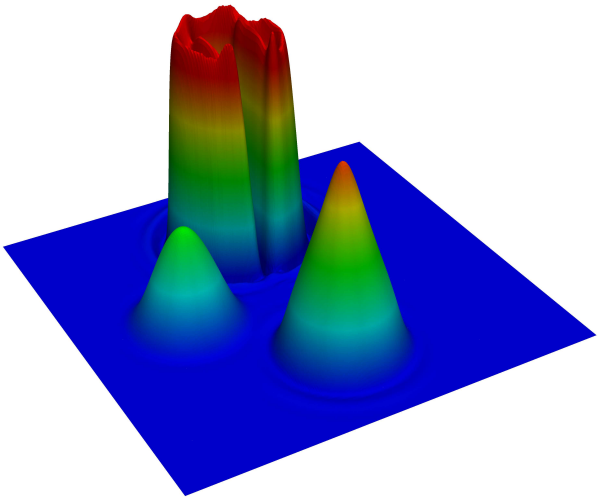
(a) Exact solution.



(b) HI(2), $\omega = 0.0$: $E_1 = 2.31\text{e}{-2}$, $E_\infty = 8.54\text{e}{-1}$.

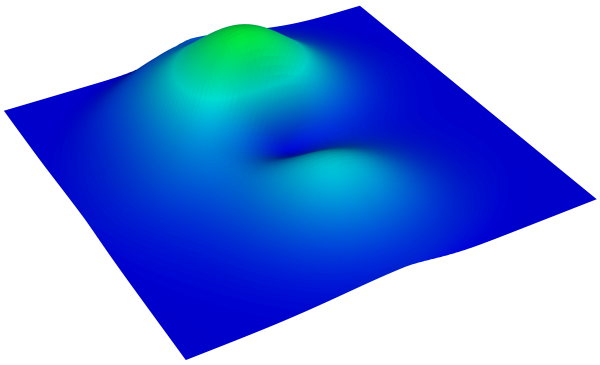


(c) HI(2), $\omega = 0.1$: $E_1 = 1.25\text{e}{-2}$, $E_\infty = 7.89\text{e}{-1}$.

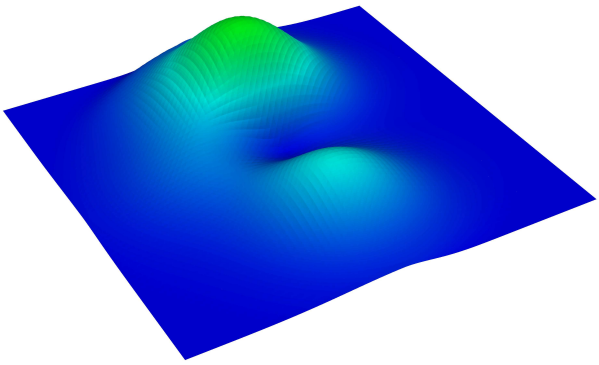


(d) HI(2), $\omega = 1.0$: $E_1 = 1.63\text{e}{-2}$, $E_\infty = 7.73\text{e}{-1}$.

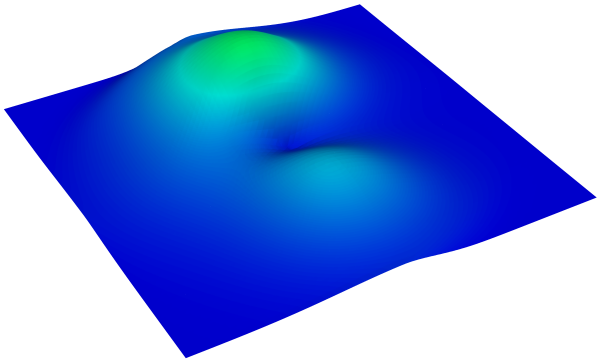
Figure 5: Solid body rotation: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128 + 1)^2$, initial data / exact solution vs. unconstrained high order solutions obtained using different values of the stabilization parameter ω.



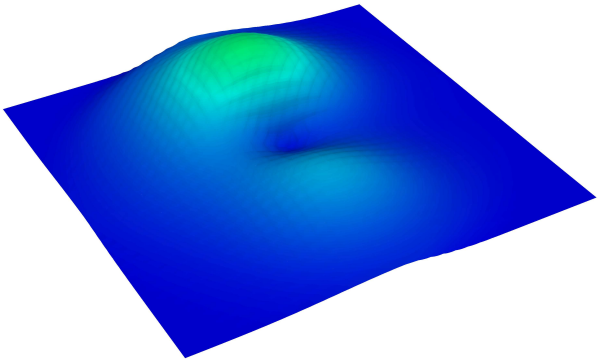
(a) LO(e; 2): $E_1 = 1.08\text{e}{-1}$, $E_\infty = 7.92\text{e}{-1}$.



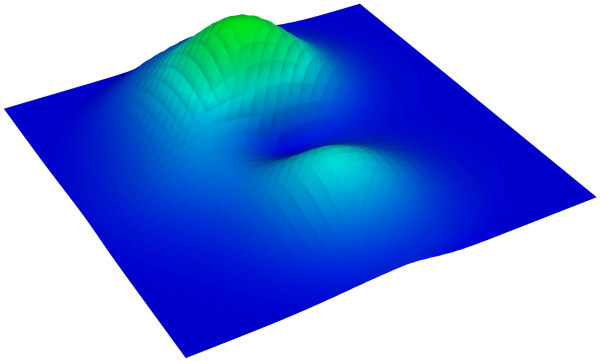
(b) LO(s; 2): $E_1 = 1.04\text{e}{-1}$, $E_\infty = 7.74\text{e}{-1}$.



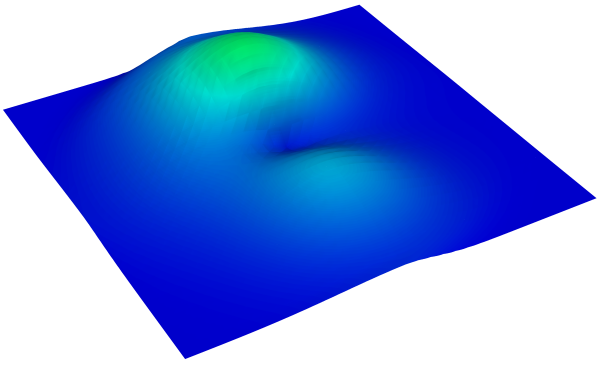
(c) LO(s_r; 2): $E_1 = 1.10\text{e}{-1}$, $E_\infty = 8.18\text{e}{-1}$.



(d) LO(e; 4): $E_1 = 1.13\text{e}{-1}$, $E_\infty = 8.30\text{e}{-1}$.



(e) LO(s; 4): $E_1 = 1.05\text{e}{-1}$, $E_\infty = 7.77\text{e}{-1}$.



(f) LO(s_r; 4): $E_1 = 1.11\text{e}{-1}$, $E_\infty = 8.24\text{e}{-1}$.

Figure 6: Solid body rotation problem: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128+1)^2$, comparison of low order solutions obtained using different upwinding techniques.

Different amounts of artificial diffusion in the three versions of the low order method are reflected in the accuracy of the corresponding FCT algorithms. Since element-stencil upwinding produces the most restrictive bounds, their use in inequality constraints for FCT gives rise to particularly strong peak clipping (see Fig. 7a, 7b, and 7c). As expected, the most accurate solution is produced by FCT based on subcell-stencil upwinding. Remarkably, the small ripples of the low order solution vanish after the antidiffusive correction. Therefore, the use of subcell-stencil Rusanov upwinding does not seem to offer any benefit in the context of FCT algorithms and subcell-stencil upwinding based on the minimal diffusion formula seems to be the best choice for the low order method. This observation is supported by numerical studies for increasing Bernstein orders (see Tables 4, 5, and 6): Subcell-stencil upwinding produces the most accurate results even in the case $p = 1$ in which subcell-stencil Rusanov upwinding is inferior to element-stencil upwinding. Due to the dependence of the subcell-stencil Rusanov method on the order p, the differences in the accuracy of FCT solutions based on these low order approximations diminish as p is increased.

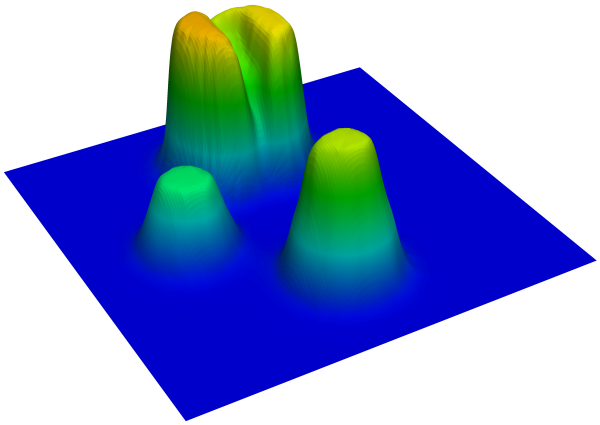
Figures 7d, 7e, and 7f show the FCT results for varying amounts of high order dissipation. So far, the impact of ω was discussed in the case of the unconstrained high order method. While the high order solution without stabilization may exhibit large oscillations in the whole domain, the FCT algorithm eliminates these oscillations. Nevertheless, the constrained solution has small ripples next to jumps, e.g. at the top of the cylinder. If high order dissipation is added, these artifacts disappear. On top of that, a stable high order approximation produces smaller antidiffusive fluxes which may be constrained using larger correction factors. The reduction in clipping and terracing due to the use of high order dissipation depends on the value of the stabilization parameter ω. In the case of the solid body rotation, $\omega = 0.1$ is a reasonable choice. If the amount of stabilization is increased by using $\omega = 1.0$, stronger peak clipping is observed and the L^1-error increases. Therefore, the parameter value $\omega = 0.1$ is used in our numerical examples.

The choice of the limiting strategy (element-based or nodal) and the definition of the bounds (element-stencil or subcell-stencil) for the underlying local maximum principles also have a strong influence on the quality of an FCT approximation (see Fig. 8). Regardless of the employed limiter, element-stencil bounds are less restrictive, allow larger correction factors and, hence, lead to approximations which are closer to the high order solution. As p increases while N_{dof} is held constant, element-stencil bounds become even less restrictive. This results in better preservation of peaks but may lead to formation of spurious ripples. To detect and remove these ripples, localized stencils should be used to sample low order values when it comes to calculating the bounds for FCT [2]. The smallest local stencil consists of a given node and its nearest neighbors, i.e. nodes belonging to the same subcells of the Bézier net. However, the use of element-stencil limiters to enforce tight subcell-stencil bounds may result in solutions with extremely strong clipping and terracing effects (see Fig. 8a and 8b). The multiplication of all components of the antidiffusive element contribution f^e by the same correction factor α^e penalizes all components even if inequality constraints (36) are violated at just one node. As the number of nodes (and constraints) per element increases, the synchronous limiting strategy tends to perform very poorly, especially in elements containing local extrema. For this reason, localized bounds should be used in conjunction with adaptive stencils and/or localized limiters. In particular, the FCT algorithm based on componentwise limiting with subsequent mass correction produces more accurate results and does a better job in preserving peaks than the element-stencil version (see Fig. 8c-8f).

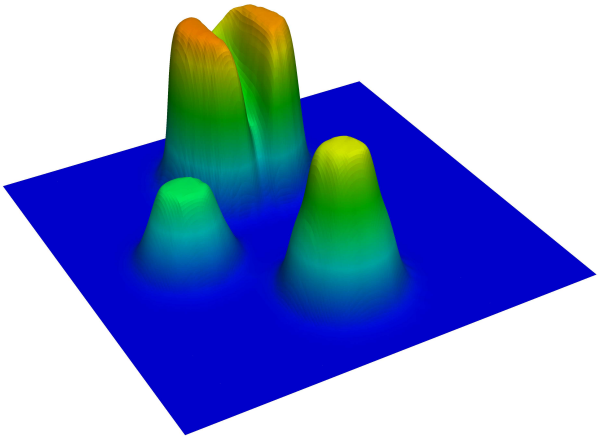
In the case of subcell-stencil upwinding and nodal limiting, the impact of the bounds is less crucial (see Fig. 8e and 8f) because the use of a less diffusive low order scheme results in smaller antidiffusive fluxes and, hence, less restrictive constraints.

For a better visual comparison, the best and the least accurate FCT approximations of order $p = 2, 3, 4$ are presented side by side in Fig. 9. While the difference is marginal in the case $p = 2$, it becomes more pronounced with increasing Bernstein order. In the case $p = 4$, the slot of the cylinder disappears completely and the height of the two peaks is drastically reduced. Note that the amount of peak clipping increases even in the case of the most accurate FCT algorithm. At the final time $T = 2\pi$, the cone and hump are 'depressed' and a plateau can be observed at the top. For that reason, the deactivation of limiters in elements identified as smooth by a robust and reliable smoothness indicator seems to be essential for preservation of accuracy in high order FCT algorithms.

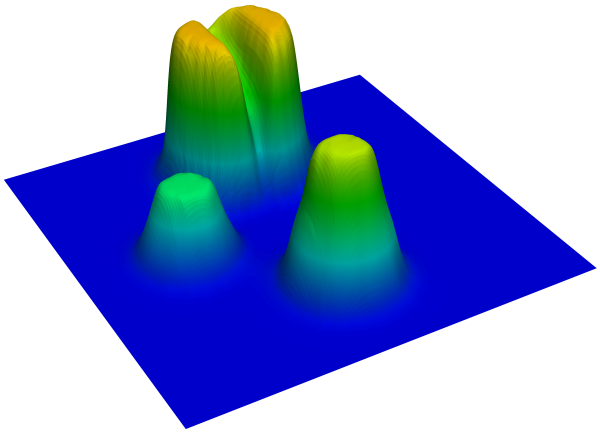
Figure 10 presents a comparison of FCT solutions obtained with and without the optional use of the smoothness indicator based on the two-dimensional Hessian determinant test proposed at the end of Section 4.6. The deactivation of limiting in smooth cells eliminates the peak clipping effect at the top of the smooth hump. Peak clipping at the top of the cone continues until the sharp peak becomes sufficiently rounded to be identified as a smooth maximum. The stronger smearing of this peak in the examples for $p = 3, 4$ is due to the fact that all numerical solutions correspond to (almost) the same number N_{dof} of global degrees of freedom. Since the current implementation of the Hessian



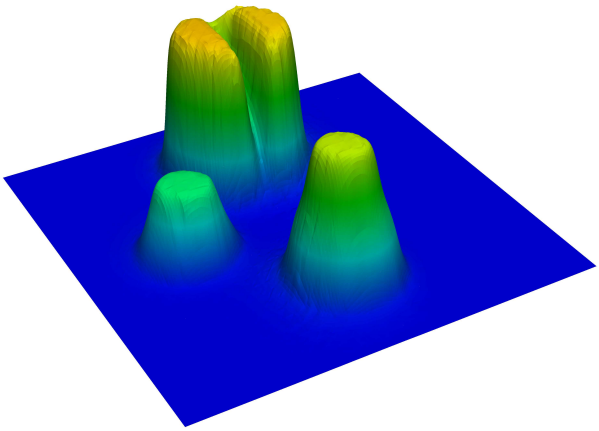
(a) FCT$(e, s, n; 4)$, $\omega = 0.1$: $E_1 = 3.77\text{e}{-2}$, $E_\infty = 8.39\text{e}{-1}$.



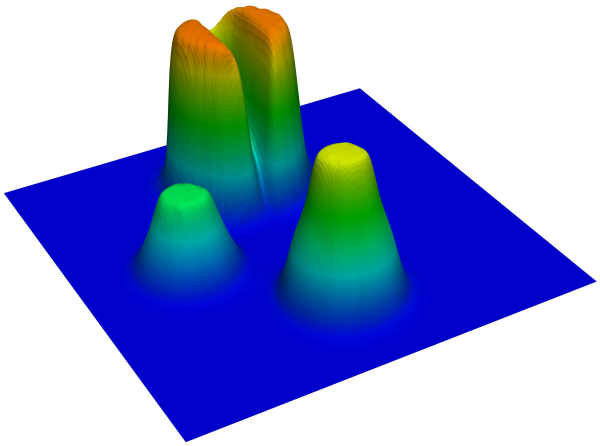
(b) FCT$(s, s, n; 4)$, $\omega = 0.1$: $E_1 = 3.28\text{e}{-2}$, $E_\infty = 8.36\text{e}{-1}$.



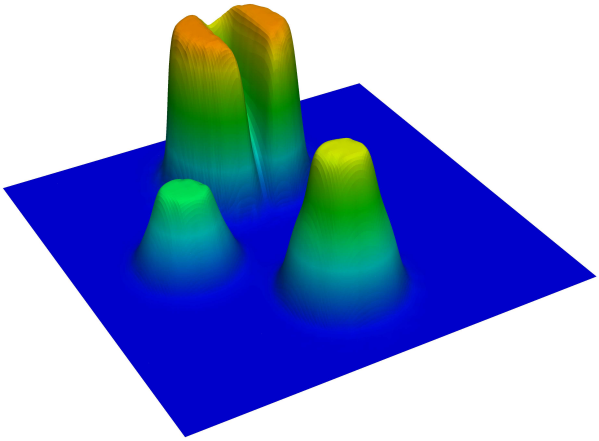
(c) FCT$(s_r, s, n; 4)$, $\omega = 0.1$: $E_1 = 3.60\text{e}{-2}$, $E_\infty = 8.38\text{e}{-1}$.



(d) FCT$(s, e, n; 4)$, $\omega = 0.0$: $E_1 = 3.61\text{e}{-2}$, $E_\infty = 8.20\text{e}{-1}$.

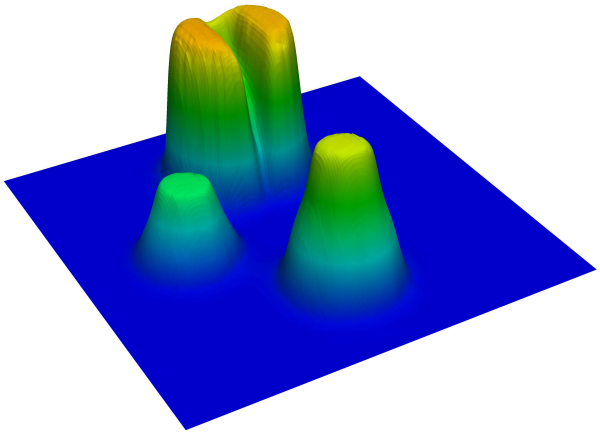


(e) FCT$(s, e, n; 4)$, $\omega = 0.1$: $E_1 = 2.89\text{e}{-2}$, $E_\infty = 8.28\text{e}{-1}$.

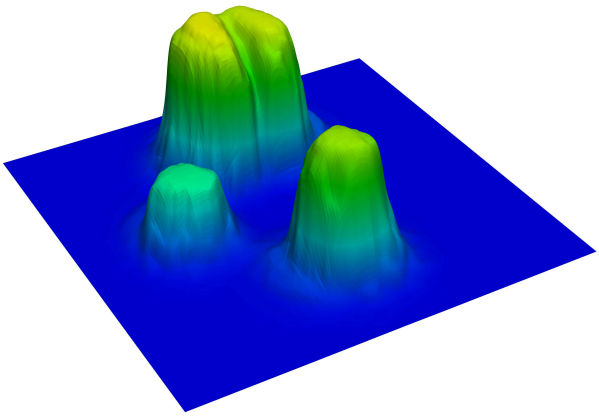


(f) FCT$(s, e, n; 4)$, $\omega = 1.0$: $E_1 = 3.00\text{e}{-2}$, $E_\infty = 8.34\text{e}{-1}$.

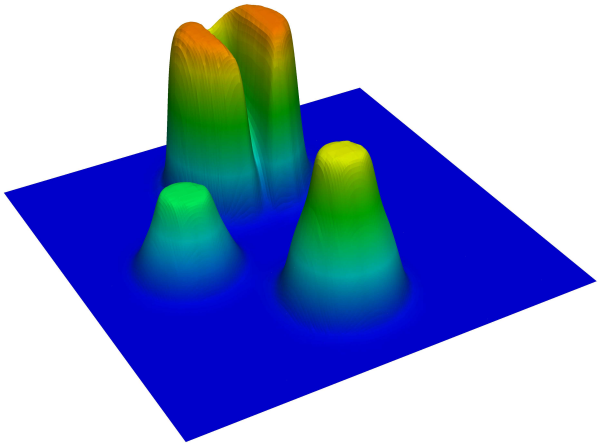
Figure 7: Solid body rotation problem: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128 + 1)^2$, comparison of FCT solutions obtained using different upwinding techniques and different amounts of high order dissipation.



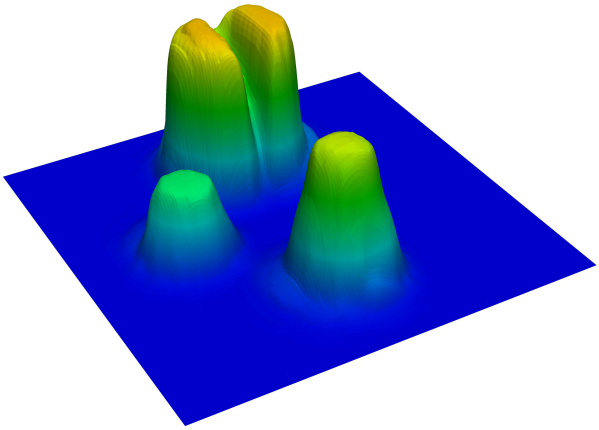
(a) FCT$(e, e, e; 3)$, $\omega = 0.1$: $E_1 = 3.30\text{e}{-2}$, $E_\infty = 8.36\text{e}{-1}$.



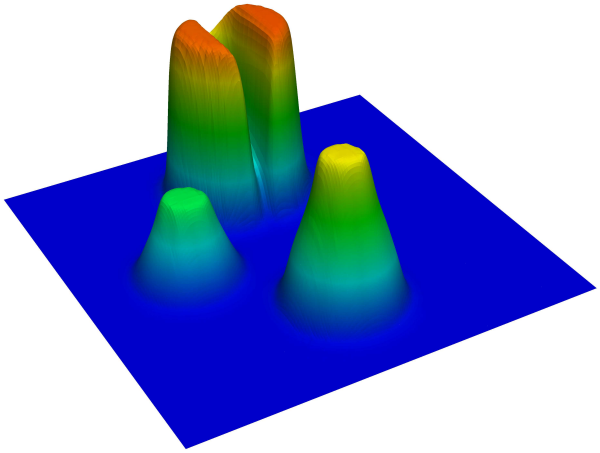
(b) FCT$(e, s, e; 3)$, $\omega = 0.1$: $E_1 = 4.71\text{e}{-2}$, $E_\infty = 8.47\text{e}{-1}$.



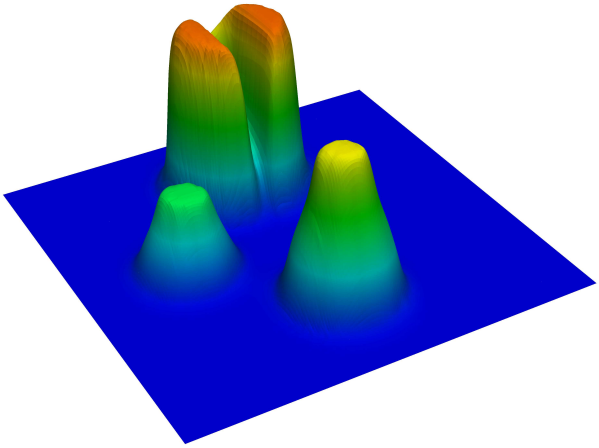
(c) FCT$(s, e, e; 3)$, $\omega = 0.1$: $E_1 = 2.85\text{e}{-2}$, $E_\infty = 8.47\text{e}{-1}$.



(d) FCT$(s, s, e; 3)$, $\omega = 0.1$: $E_1 = 3.83\text{e}{-2}$, $E_\infty = 8.52\text{e}{-1}$.

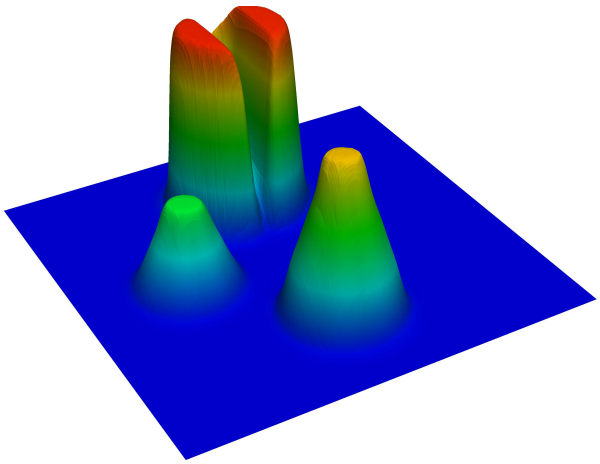


(e) FCT$(s, e, n; 3)$, $\omega = 0.1$: $E_1 = 2.52\text{e}{-2}$, $E_\infty = 8.43\text{e}{-1}$.

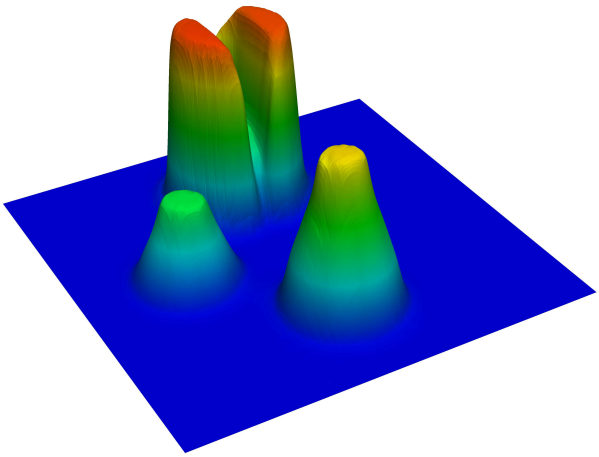


(f) FCT$(s, s, n; 3)$, $\omega = 0.1$: $E_1 = 2.77\text{e}{-2}$, $E_\infty = 8.47\text{e}{-1}$.

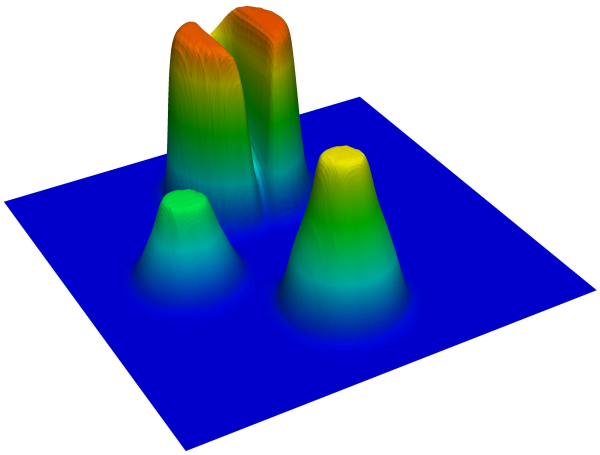
Figure 8: Solid body rotation: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (129 + 1)^2$, comparison of FCT solutions obtained using different versions of the element-stencil and nodal-point limiters.



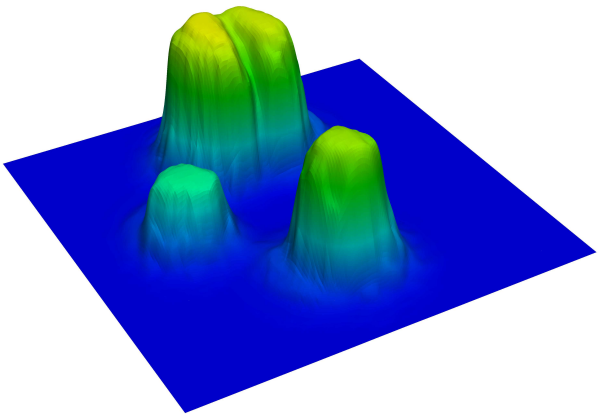
(a) FCT$(s, e, n; 2)$, $\omega = 0.1$: $E_1 = 2.02\text{e}{-2}$, $E_\infty = 8.45\text{e}{-1}$.



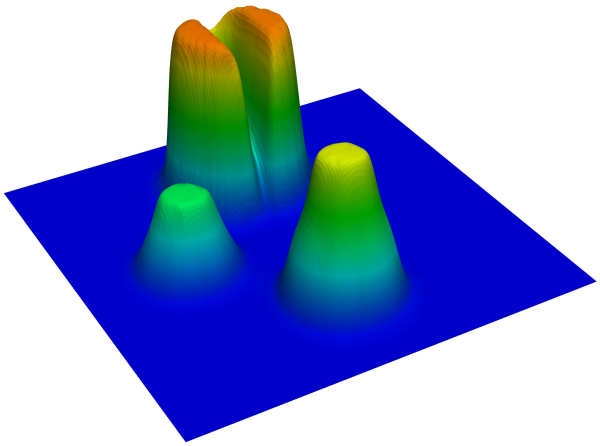
(b) FCT$(e, s, e; 2)$, $\omega = 0.1$: $E_1 = 2.45\text{e}{-2}$, $E_\infty = 8.44\text{e}{-1}$.



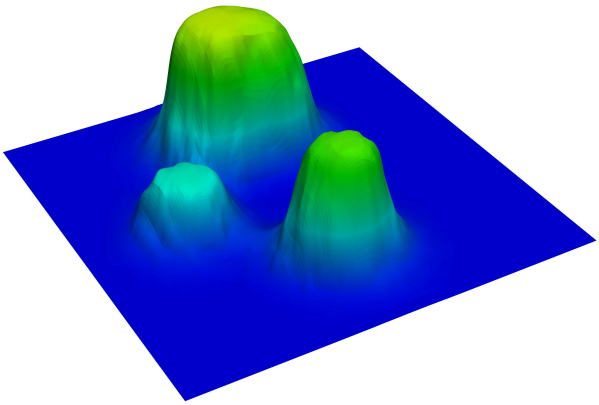
(c) FCT$(s, e, n; 3)$, $\omega = 0.1$: $E_1 = 2.52\text{e}{-2}$, $E_\infty = 8.43\text{e}{-1}$.



(d) FCT$(e, s, e; 3)$, $\omega = 0.1$: $E_1 = 4.71\text{e}{-2}$, $E_\infty = 8.47\text{e}{-1}$.



(e) FCT$(s, e, n; 4)$, $\omega = 0.1$: $E_1 = 2.89\text{e}{-2}$, $E_\infty = 8.28\text{e}{-1}$.



(f) FCT$(e, s, e; 4)$, $\omega = 0.1$: $E_1 = 5.60\text{e}{-2}$, $E_\infty = 8.27\text{e}{-1}$.

Figure 9: Solid body rotation: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128 + 1)^2$ for $p = 2, 4$, $N_{\text{dof}} = (129 + 1)^2$ for $p = 3$, comparison of the most accurate and least accurate FCT solutions.

determinant test for $p > 2$ is based on elementwise L^2 projections (see Section 4.6), Bernstein approximations of higher order correspond to smoothness estimation for larger elements using just cell-averaged values of the second derivatives. To use more information about the solution behavior inside these elements, the Hessian test can be applied to subcell-averaged second derivatives in the case $p > 2$. The further development and evaluation of localized smoothness criteria for local limiters is beyond the scope of this work and requires further research. However, based on this initial illustration of the benefit of the methods, we believe these approaches have promise and should be considered further in the context of the high order Bernstein finite elements we have developed here. Even though virtually no undershoots or overshoots are observed in numerical solutions to the solid body rotation problem, the ability of FCT algorithms equipped with smoothness indicators to identify troubled cells must be verified in more detailed numerical studies. Since the use of correction factors depending on the value of a smoothness indicator rules out the possibility of proving local extremum diminishing properties, a postprocessing procedure [21, 50] may be added to enforce extremum-preserving 'safeguard' bounds based on the data from the previous time steps and/or physical constraints like positivity preservation.

The norms of errors for several combinations of the FCT components are summarized in Tables 4, 5, and 6. To quantify the damage caused by clipping and terracing, the L^1- and L^∞-errors in numerical solutions for solid body rotation of the cone and hump are measured separately. The isolation of the two peaks in this numerical study is dictated by the fact that the total error of FCT solutions for the three-body benchmark configuration is dominated by numerical diffusion generated at the discontinuous slopes of the slotted cylinder. The errors in the resolution of the cone and hump are much smaller, which makes it impossible to quantify them without excluding the other bodies when it comes to calculation and comparison of errors. The errors of different FCT schemes behave similarly for the two peaks. The algorithm based on subcell-stencil upwinding with element-stencil bounds and nodal limiting seems to produce the most accurate solutions. The optional use of the smoothness indicator is seen to have the expected positive effect on the accuracy of the FCT results for the smooth hump.

7. Conclusions

The findings presented in this paper pave the way for designing FCT-like algorithms on the basis of high order Bernstein-Bézier finite elements. The non-negativity of local basis functions, the fact that the shape functions are bounded by their Bernstein coefficients, and the variation diminishing properties make them a very attractive alternative to standard Lagrange elements. The proposed approach to constraining the antidiffusive part of the discrete transport is readily applicable to Bernstein bases of arbitrary order. The new discrete upwinding strategy produces low order schemes with compact stencils. The accuracy of the resulting low order solutions depends just on the number of global degrees of freedom but not on the order of the underlying finite element approximation. The new approaches to localization of FCT limiters make it possible to perform all computations in a single loop over elements and constrain the vectors of antidiffusive element contributions using different correction factors for different components. Order barriers caused by the peak clipping effect can be circumvented using smoothness indicators based on the proposed second derivative test. The optional use of high order background dissipation offers better convergence rates for smooth solutions and eliminates the need for *ad hoc* prelimiting which is frequently used to reduce numerical dispersion in FCT algorithms. The numerical results confirm that the above revisions lead to significant improvements but further efforts are required to achieve optimal convergence behavior in a cost-effective manner.

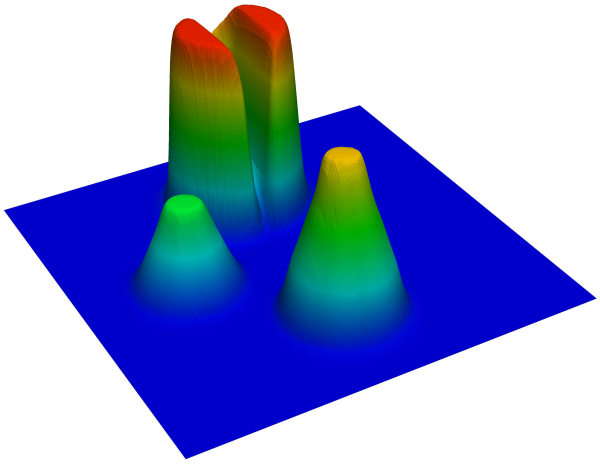
Acknowledgements

The research of Christoph Lohmann and Dmitri Kuzmin was sponsored by the German Research Association (DFG) under grant KU 1530/13-1. The research of John Shadid and Sibusiso Mabuza was supported by the DOE Office of Science AMR Program at Sandia National Laboratories under contract DE-AC04-94AL85000.

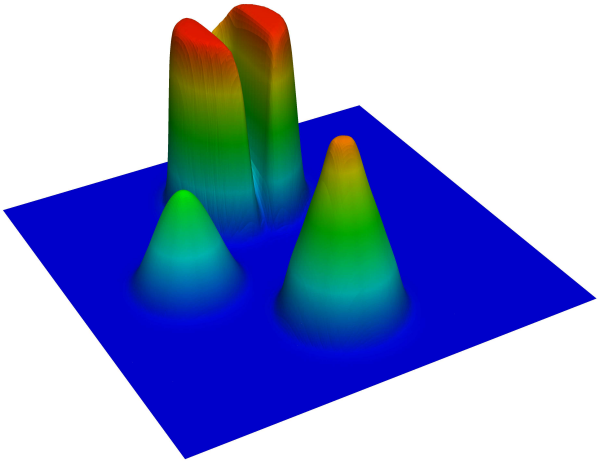
References

[1] M. Ainsworth, G. Andriamaro, and O. Davydov. Bernstein-Bèzier finite elements of arbitrary order and optimal assembly procedures. *SIAM J. Sci. Comput.*, 33:3087–3109, 2011.

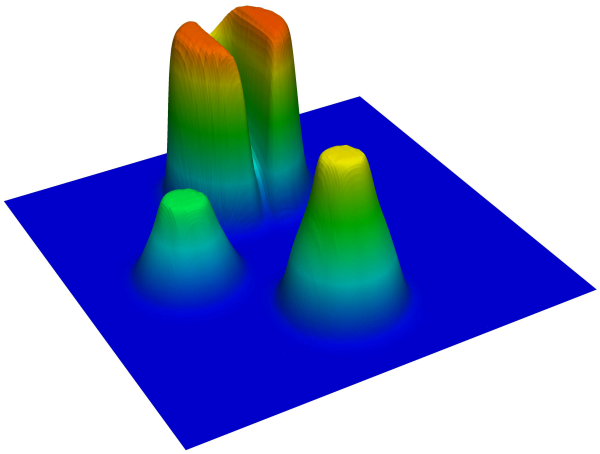
[2] R. Anderson, V. Dobrev, Tz. Kolev, D. Kuzmin, M. Quezada de Luna, R. Rieben, and V. Tomov. High-order local maximum principle preserving (MPP) discontinuous Galerkin finite element method for the transport equation. *Journal of Computational Physics*, in review, 2016.



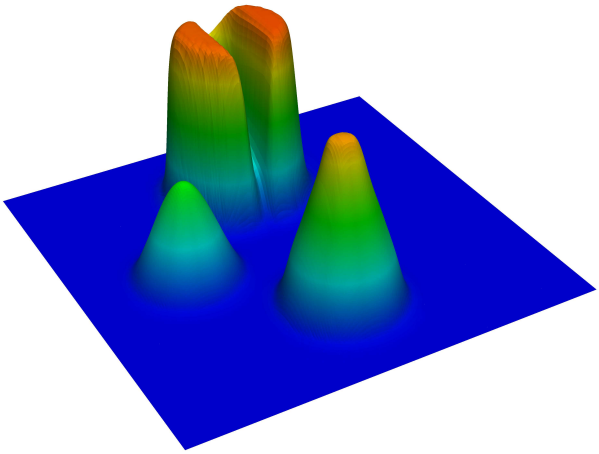
(a) FCT$(s, e, n; 2)$ without SI, $\omega = 0.1$: $E_1 = 2.02\text{e}{-2}$, $E_\infty = 8.45\text{e}{-1}$, $u_{\min} = -1.10\text{e}{-15}, u_{\max} = 9.63\text{e}{-1}$.



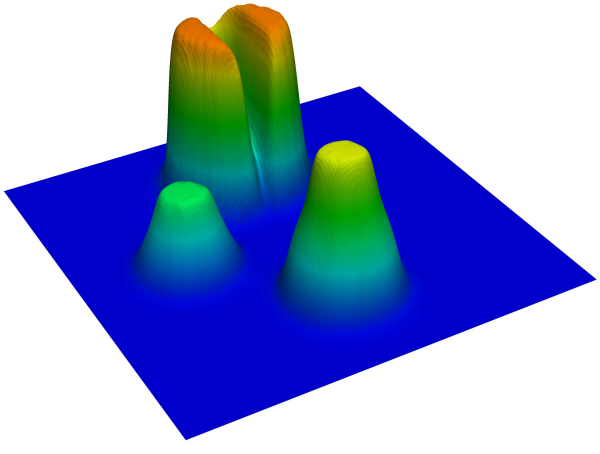
(b) FCT$(s, e, n; 2)$ with SI, $\omega = 0.1$: $E_1 = 1.96\text{e}{-2}$, $E_\infty = 8.45\text{e}{-1}$, $u_{\min} = -1.59\text{e}{-7}, u_{\max} = 9.71\text{e}{-1}$.



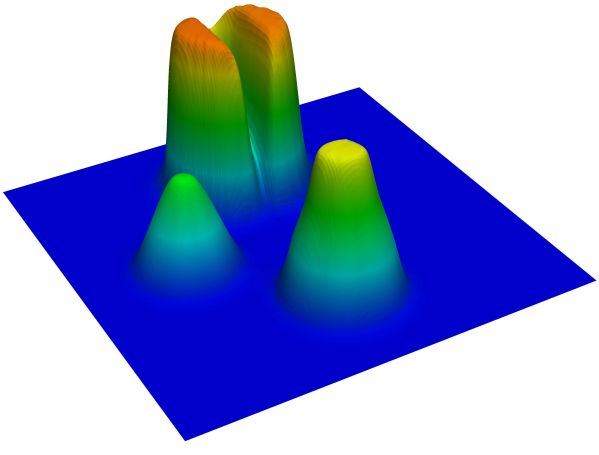
(c) FCT$(s, e, n; 3)$ without SI, $\omega = 0.1$: $E_1 = 2.52\text{e}{-2}$, $E_\infty = 8.43\text{e}{-1}$, $u_{\min} = -4.57\text{e}{-14}, u_{\max} = 9.08\text{e}{-1}$.



(d) FCT$(s, e, n; 3)$ with SI, $\omega = 0.1$: $E_1 = 2.43\text{e}{-2}$, $E_\infty = 8.43\text{e}{-1}$, $u_{\min} = -4.83\text{e}{-6}, u_{\max} = 9.12\text{e}{-1}$.



(e) FCT$(s, e, n; 4)$ without SI, $\omega = 0.1$: $E_1 = 2.89\text{e}{-2}$, $E_\infty = 8.28\text{e}{-1}$, $u_{\min} = -1.24\text{e}{-8}, u_{\max} = 8.75\text{e}{-1}$.



(f) FCT$(s, e, n; 4)$ with SI, $\omega = 0.1$: $E_1 = 2.81\text{e}{-2}$, $E_\infty = 8.28\text{e}{-1}$, $u_{\min} = -7.41\text{e}{-7}, u_{\max} = 8.82\text{e}{-1}$.

Figure 10: Solid body rotation: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128 + 1)^2$ for $p = 2, 4$, $N_{\text{dof}} = (129 + 1)^2$ for $p = 3$, comparison of FCT solutions with and without the optional use of the extremum-preserving smoothness indicator (SI) based on the second derivative test (53) with $C = 2$.

	element upwinding		subcell upwinding		subcell Rusanov	
	L^1	L^∞	L^1	L^∞	L^1	L^∞
LO$(\cdot;1)$	2.297e−2	7.402e−1	2.297e−2	7.403e−1	2.623e−2	8.135e−1
LO$(\cdot;2)$	2.566e−2	7.938e−1	2.404e−2	7.621e−1	2.665e−2	8.207e−1
LO$(\cdot;3)$	2.686e−2	8.168e−1	2.415e−2	7.650e−1	2.685e−2	8.243e−1
LO$(\cdot;4)$	2.777e−2	8.342e−1	2.420e−2	7.655e−1	2.704e−2	8.269e−1
FCT$(\cdot,s,e;1)$	6.654e−4	1.593e−1	6.480e−4	1.575e−1	7.011e−4	1.675e−1
FCT$(\cdot,s,e;2)$	1.310e−3	2.243e−1	1.165e−3	2.135e−1	1.131e−3	2.130e−1
FCT$(\cdot,s,e;3)$	4.522e−3	3.386e−1	3.472e−3	2.991e−1	2.848e−3	3.115e−1
FCT$(\cdot,s,e;4)$	5.743e−3	4.258e−1	4.007e−3	3.431e−1	3.925e−3	3.554e−1
FCT$(\cdot,e,e;1)$	6.654e−4	1.593e−1	6.480e−4	1.575e−1	7.011e−4	1.675e−1
FCT$(\cdot,e,e;2)$	1.244e−3	2.191e−1	1.150e−3	2.124e−1	1.124e−3	2.129e−1
FCT$(\cdot,e,e;3)$	2.036e−3	2.867e−1	1.716e−3	2.673e−1	1.958e−3	2.881e−1
FCT$(\cdot,e,e;4)$	3.030e−3	3.291e−1	2.355e−3	2.937e−1	2.846e−3	3.244e−1
FCT$(\cdot,s,n;1)$	6.792e−4	1.615e−1	6.614e−4	1.608e−1	7.292e−4	1.723e−1
FCT$(\cdot,s,n;2)$	9.974e−4	2.002e−1	9.987e−4	1.977e−1	1.077e−3	2.074e−1
FCT$(\cdot,s,n;3)$	2.014e−3	2.720e−1	1.688e−3	2.503e−1	2.048e−3	2.742e−1
FCT$(\cdot,s,n;4)$	2.639e−3	3.084e−1	2.088e−3	2.739e−1	2.481e−3	3.025e−1
FCT$(\cdot,e,n;1)$	6.792e−4	1.615e−1	6.614e−4	1.608e−1	7.292e−4	1.723e−1
FCT$(\cdot,e,n;2)$	9.918e−4	2.006e−1	9.955e−4	1.989e−1	1.069e−3	2.072e−1
FCT$(\cdot,e,n;3)$	1.599e−3	2.613e−1	1.410e−3	2.442e−1	1.682e−3	2.700e−1
FCT$(\cdot,e,n;4)$	2.084e−3	2.925e−1	1.826e−3	2.741e−1	2.143e−3	2.979e−1
FCT$(\cdot,e,n;1)$ with SI	6.792e−4	1.615e−1	6.614e−4	1.608e−1	7.292e−4	1.723e−1
FCT$(\cdot,e,n;2)$ with SI	7.040e−4	1.298e−1	7.102e−4	1.276e−1	7.556e−4	1.369e−1
FCT$(\cdot,e,n;3)$ with SI	9.884e−4	1.680e−1	9.155e−4	1.581e−1	1.117e−3	2.008e−1
FCT$(\cdot,e,n;4)$ with SI	1.833e−3	2.693e−1	1.711e−3	2.626e−1	2.026e−3	2.884e−1

Table 4: Rotation of cone: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128+1)^2$ for $p = 1, 2, 4$, $N_{\text{dof}} = (129+1)^2$ for $p = 3$, convergence behavior of LO and FCT.

	element upwinding		subcell upwinding		subcell Rusanov	
	L^1	L^∞	L^1	L^∞	L^1	L^∞
LO$(\cdot;1)$	1.135e−2	3.770e−1	1.135e−2	3.770e−1	1.267e−2	4.133e−1
LO$(\cdot;2)$	1.245e−2	4.036e−1	1.180e−2	3.881e−1	1.284e−2	4.168e−1
LO$(\cdot;3)$	1.293e−2	4.149e−1	1.184e−2	3.890e−1	1.292e−2	4.184e−1
LO$(\cdot;4)$	1.328e−2	4.233e−1	1.186e−2	3.900e−1	1.299e−2	4.198e−1
FCT$(\cdot,s,e;1)$	1.408e−4	3.535e−2	1.329e−4	3.455e−2	1.583e−4	3.886e−2
FCT$(\cdot,s,e;2)$	5.468e−4	7.074e−2	4.679e−4	6.424e−2	4.497e−4	6.335e−2
FCT$(\cdot,s,e;3)$	2.163e−3	1.498e−1	1.657e−3	1.151e−1	1.342e−3	1.203e−1
FCT$(\cdot,s,e;4)$	3.568e−3	2.128e−1	1.911e−3	1.495e−1	1.934e−3	1.575e−1
FCT$(\cdot,e,e;1)$	1.408e−4	3.535e−2	1.329e−4	3.455e−2	1.583e−4	3.886e−2
FCT$(\cdot,e,e;2)$	5.067e−4	6.753e−2	4.578e−4	6.358e−2	4.447e−4	6.318e−2
FCT$(\cdot,e,e;3)$	9.280e−4	1.048e−1	7.521e−4	9.367e−2	9.023e−4	1.054e−1
FCT$(\cdot,e,e;4)$	1.573e−3	1.359e−1	1.144e−3	1.095e−1	1.454e−3	1.298e−1
FCT$(\cdot,s,n;1)$	1.484e−4	3.655e−2	1.417e−4	3.634e−2	1.720e−4	4.111e−2
FCT$(\cdot,s,n;2)$	3.596e−4	5.584e−2	3.624e−4	5.537e−2	4.082e−4	6.013e−2
FCT$(\cdot,s,n;3)$	8.851e−4	9.687e−2	6.962e−4	8.368e−2	8.872e−4	9.722e−2
FCT$(\cdot,s,n;4)$	1.299e−3	1.200e−1	9.505e−4	9.889e−2	1.210e−3	1.158e−1
FCT$(\cdot,e,n;1)$	1.484e−4	3.655e−2	1.417e−4	3.634e−2	1.720e−4	4.111e−2
FCT$(\cdot,e,n;2)$	3.574e−4	5.626e−2	3.629e−4	5.625e−2	4.032e−4	6.011e−2
FCT$(\cdot,e,n;3)$	6.775e−4	8.918e−2	5.807e−4	8.089e−2	7.370e−4	9.468e−2
FCT$(\cdot,e,n;4)$	9.886e−4	1.087e−1	8.081e−4	9.641e−2	1.044e−3	1.127e−1
FCT$(\cdot,e,n;1)$ with SI	1.484e−4	3.655e−2	1.417e−4	3.634e−2	1.720e−4	4.111e−2
FCT$(\cdot,e,n;2)$ with SI	1.390e−4	6.222e−3	1.452e−4	6.431e−3	1.542e−4	6.815e−3
FCT$(\cdot,e,n;3)$ with SI	1.754e−4	6.353e−3	1.651e−4	6.534e−3	1.749e−4	7.200e−3
FCT$(\cdot,e,n;4)$ with SI	2.936e−4	2.826e−2	2.433e−4	1.613e−2	3.011e−4	3.037e−2

Table 5: Rotation of a hump: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128+1)^2$ for $p = 1,2,4$, $N_{\text{dof}} = (129+1)^2$ if $p = 3$, convergence behavior of LO and FCT.

	element upwinding		subcell upwinding		subcell Rusanov	
	L^1	L^∞	L^1	L^∞	L^1	L^∞
LO$(\cdot;1)$	6.972e−2	7.643e−1	6.972e−2	7.643e−1	7.391e−2	7.988e−1
LO$(\cdot;2)$	7.332e−2	7.914e−1	7.110e−2	7.743e−1	7.442e−2	8.042e−1
LO$(\cdot;3)$	7.511e−2	8.048e−1	7.136e−2	7.757e−1	7.488e−2	8.059e−1
LO$(\cdot;4)$	7.630e−2	8.160e−1	7.132e−2	7.767e−1	7.505e−2	8.085e−1
FCT$(\cdot,s,e;1)$	1.577e−2	8.309e−1	1.571e−2	8.253e−1	1.625e−2	8.267e−1
FCT$(\cdot,s,e;2)$	2.265e−2	8.440e−1	2.090e−2	8.472e−1	2.031e−2	8.467e−1
FCT$(\cdot,s,e;3)$	4.019e−2	8.518e−1	3.310e−2	8.434e−1	3.315e−2	8.493e−1
FCT$(\cdot,s,e;4)$	4.665e−2	8.301e−1	3.999e−2	8.143e−1	3.978e−2	8.235e−1
FCT$(\cdot,e,e;1)$	1.577e−2	8.309e−1	1.571e−2	8.253e−1	1.625e−2	8.267e−1
FCT$(\cdot,e,e;2)$	2.160e−2	8.431e−1	2.040e−2	8.462e−1	1.980e−2	8.447e−1
FCT$(\cdot,e,e;3)$	2.980e−2	8.367e−1	2.599e−2	8.472e−1	2.832e−2	8.462e−1
FCT$(\cdot,e,e;4)$	3.622e−2	8.323e−1	3.126e−2	8.348e−1	3.399e−2	8.332e−1
FCT$(\cdot,s,n;1)$	1.589e−2	8.313e−1	1.568e−2	8.201e−1	1.639e−2	8.286e−1
FCT$(\cdot,s,n;2)$	1.888e−2	8.384e−1	1.910e−2	8.477e−1	1.943e−2	8.460e−1
FCT$(\cdot,s,n;3)$	2.827e−2	8.445e−1	2.522e−2	8.476e−1	2.751e−2	8.479e−1
FCT$(\cdot,s,n;4)$	3.366e−2	8.373e−1	2.962e−2	8.353e−1	3.213e−2	8.367e−1
FCT$(\cdot,e,n;1)$	1.589e−2	8.313e−1	1.568e−2	8.201e−1	1.639e−2	8.286e−1
FCT$(\cdot,e,n;2)$	1.858e−2	8.362e−1	1.882e−2	8.452e−1	1.911e−2	8.440e−1
FCT$(\cdot,e,n;3)$	2.510e−2	8.360e−1	2.323e−2	8.431e−1	2.510e−2	8.390e−1
FCT$(\cdot,e,n;4)$	2.873e−2	8.296e−1	2.623e−2	8.283e−1	2.860e−2	8.306e−1
FCT$(\cdot,e,n;1)$ with SI	1.589e−2	8.313e−1	1.568e−2	8.201e−1	1.639e−2	8.286e−1
FCT$(\cdot,e,n;2)$ with SI	1.850e−2	8.362e−1	1.872e−2	8.452e−1	1.902e−2	8.440e−1
FCT$(\cdot,e,n;3)$ with SI	2.493e−2	8.360e−1	2.316e−2	8.431e−1	2.507e−2	8.390e−1
FCT$(\cdot,e,n;4)$ with SI	2.858e−2	8.298e−1	2.605e−2	8.284e−1	2.840e−2	8.308e−1

Table 6: Rotation of a cylinder: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128+1)^2$ for $p = 1, 2, 4$, $N_{\text{dof}} = (129+1)^2$ for $p = 3$, convergence behavior of LO and FCT.

[3] S. Badia and J. Bonilla. Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. *arXiv:1606.08743*, 2016.

[4] S. Badia and A. Hierro. On monotonicity-preserving stabilized finite element approximations of transport problems. *SIAM J. Sci. Comput.*, 36:A2673–A2697, 2014.

[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.*, 2016.

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

[7] A. Bhatt and A. Ojha. Variation diminishing properties of Bernstein polynomials on a tetrahedron. *Journal of Approximation Theory*, 79:180–189, 1994.

[8] D. L. Book. The conception, gestation, birth, and infancy of FCT. In D. Kuzmin, R. Löhner, and S. Turek, editors, *Flux-Corrected Transport*, Scientific Computation, pages 1–22. Springer Netherlands, 2012.

[9] J.P. Boris and D.L. Book. I. SHASTA, a fluid transport algorithm that works. *J. Comput. Phys.*, 11:38–69, 1973.

[10] M. Braack and E. Burman. Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. *SIAM J. Numer. Anal.*, 43:2544–2566, 2006.

[11] A.N. Brooks and T.J.R. Hughes. Streamline upwind Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32:199–259, 1982.

[12] E. Burman. A monotonicity preserving, nonlinear, finite element upwind method for the transport equation. *Applied Mathematics Letters*, 49:141–146, 2015.

[13] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193:1437–1453, 2004.

[14] B. Cockburn and J. Guzmán. Error estimates for the Runge-Kutta discontinuous Galerkin method for the transport equation with discontinuous initial data. *SIAM J. Numer. Anal.*, 46(3):1364–1398, 2008.

[15] R. Codina. A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation. *Comput. Methods Appl. Mech. Engrg.*, 110:325–342, 1993.

[16] P. Colella and M. Sekora. A limiter for PPM that preserves accuracy at smooth extrema. *J. Comput. Phys.*, 227:7069–7076, 2008.

[17] C.J. Cotter and D. Kuzmin. Embedded discontinuous Galerkin transport schemes with localised limiters. *Journal of Computational Physics*, 311:363 – 373, 2016.

[18] P.A.B. de Sampaio and A.L.G.A. Coutinho. A natural derivation of discontinuity capturing operator for convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 190:6291–6308, 2001.

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

[20] S. Diot, R. Loubére, and S. Clain. The Multidimensional Optimal Order Detection method in the three-dimensional case: very high-order finite volume method for hyperbolic systems. *Int. J. Numer. Methods Fluids*, 73:362–392, 2013.

[21] 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.*, 278:47–75, 2014.

[22] G. Farin. Triangular Bernstein-Bézier patches. *Computer Aided Geometric Design*, 3(2):83 – 127, 1986.

[23] M. S. Floater and J. M. Pena. Tensor-product monotonicity preservation. *Advances in Computational Mathematics*, 9:353–362, 1998.

[24] M. S. Floater and J. M. Pena. Monotonicity preservation on triangles. *Mathematics of Computation*, 69:1505–1519, 2000.

[25] S.K. Godunov. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Mat. Sb.*, 47(89)(3):271–306, 1959.

[26] T. N. T. Goodman. Further variation diminishing properties of Bernstein polynomials on triangles. *Constructive Approximation*, 3(1):297–305, 1987.

[27] T. N. T. Goodman. Variation diminishing properties of Bernstein polynomials on triangles. *Journal of Approximation Theory*, 50:111–126, 1987.

[28] S. Gottlieb, D. Ketcheson, and C.-W. Shu. *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*. World Scientific, 2011.

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

[30] 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.*, 52:2163–2182, 1989.

[31] T.J.R. Hughes, M. Mallet, and A. Mizukami. A new finite element formulation for computational fluid dynamics. II. Beyond SUPG. *Comput. Methods Appl. Mech. Engrg.*, 54(3):341–355, 1986.

[32] V. John, S. Kaya, and W. Layton. A two-level variational multiscale method for convection-dominated convection-diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 195(3336):4594 – 4603, 2006.

[33] V. John and E. Schmeyer. Finite element methods for time-dependent convection-diffusion-reaction equations with small diffusion. *Computer Methods in Applied Mechanics and Engineering*, 198(34):475 – 494, 2008.

[34] R. C. Kirby. Fast simplicial finite element algorithms using Bernstein polynomials. *Numer. Math.*, 117:631–652, 2011.

[35] D. Kuzmin. Explicit and implicit FEM-FCT algorithms with flux linearization. *J. Comput. Phys.*, 228(7):2517–2534, 2009.

[36] D. Kuzmin. Algebraic flux Correction I. Scalar conservation laws. In D. Kuzmin, R. Löhner, and S. Turek, editors, *Flux-Corrected Transport*, Scientific Computation, pages 145–192. Springer Netherlands, 2012.

[37] D. Kuzmin, S. Basting, and J. Shadid. Monotone local projection stabilization schemes for continuous finite elements. Technical Report 539, Ergebnisber. Inst. Angew. Math., TU Dortmund University, 2016.

[38] D. Kuzmin and J. Hämäläinen. *Finite Element Methods for Computational Fluid Dynamics: A Practical Guide*. SIAM, 2014.

[39] D. Kuzmin and F. Schieweck. A parameter-free smoothness indicator for high-resolution finite element schemes. *Central European Journal of Mathematics*, 11:1478–1488, 2013.

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

[41] W. Layton. A connection between subgrid scale eddy viscosity and mixed methods. *Appl. Math. Comput.*, 133(1):147–157, November 2002.
[42] R. Leroy. Certificates of positivity in the simplicial Bernstein basis. Technical report, HAL Id: hal-00589945, 2011.
[43] Randall J. LeVeque. High-resolution conservative algorithms for advection in incompressible flow. *SIAM Journal on Numerical Analysis*, 33(2):627–665, 1996.
[44] R. Löhner. *Applied CFD Techniques: An Introduction Based on Finite Element Methods*. John Wiley & Sons, 2008.
[45] 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*, 7:1093–1109, 1987.
[46] R. Löhner, K. Morgan, M. Vahdati, J.P. Boris, and D.L. Book. FEM-FCT: combining unstructured grids with high resolution. *Commun. Appl. Numer. Methods*, 4:717–729, 1988.
[47] J. Qiu and C.-W. Shu. A comparison of troubled-cell indicators for Runge-Kutta discontinuous Galerkin methods using weighted essentially nonoscillatory limiters. *SIAM J. Sci. Comput.*, 27:995–1013, 2005.
[48] C.-W. Shu. High order weighted essentially nonoscillatory schemes for convection dominated problems. *SIAM Review*, 51:82–126, 2009.
[49] S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31:335–362, June 1979.
[50] S.T. Zalesak. The design of Flux-Corrected Transport (FCT) algorithms for structured grids. In D. Kuzmin, R. Löhner, and S. Turek, editors, *Flux-Corrected Transport*, Scientific Computation, pages 23–65. Springer Netherlands, 2012.
[51] X. Zhang and C.-W. Shu. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. *Proc. R. Soc. A*, 467:2752–2776, 2011.

AppendixA. Properties of the mass lumping operator

In Section 3.2.2, we justified the approximation of (17) by (18) by considering preconditioned local problems of the form (16) and using the corresponding element matrices in (18). This modification of the high order scheme can also be interpreted as a modified projection of shape functions $\dot{u}_h^e$ given in terms of their Bernstein coefficients c_j^e into the continuous finite element space V_h. Let the local time derivatives $\dot{c}_j^e$ be defined by

$$\dot{c}_i^e = \sum_{r,s=1}^{N_{\text{dof}}} m_{ir}^{-e} \varkappa_{rs}^e c_s. \tag{A.1}$$

The Bernstein coefficients of a continuous function $\partial_t u_h \in V_h$ can be determined using a suitable averaging procedure to piece together the non-matching solutions $\dot{u}_h^e$ to N_{elem} local problems of the form (A.1).

Invoking (17), we find that the high order Galerkin scheme (9a) can be written in terms of $\dot{c}_j^e$ as follows:

$$\sum_{j=1}^{N_{\text{dof}}} \sum_{e=1}^{N_{\text{elem}}} m_{ij}^e \frac{\mathrm{d}c_j}{\mathrm{d}t} = \sum_{e=1}^{N_{\text{elem}}} m_{ij}^e \dot{c}_j^e + \sum_{j=1}^{N_{\text{dof}}} \Big(\int_{\Gamma_{\text{in}}} B_i B_j \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}\mathbf{s} \Big) c_j + \sum_{e=1}^{N_{\text{elem}}} b_i^e \qquad \text{for all } 1 \le i \le N_{\text{dof}}. \tag{A.2}$$

This averaging procedure corresponds to the consistent L^2-projection of $\dot{u}_h^e$ into the finite element space V_h.

The representation of the lumped-mass preconditioned Galerkin scheme (18) in terms of $\dot{c}_j^e$ is given by

$$m_i \frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{e=1}^{N_{\text{elem}}} m_i^e \frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{e=1}^{N_{\text{elem}}} m_i^e \dot{c}_i^e + \sum_{j=1}^{N_{\text{dof}}} \Big(\int_{\Gamma_{\text{in}}} B_i B_j \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}\mathbf{s} \Big) c_j + \sum_{e=1}^{N_{\text{elem}}} b_i^e \qquad \text{for all } 1 \le i \le N_{\text{dof}} \tag{A.3}$$

and can be interpreted as a modified local-maximum-principle-satisfying projection scheme. For any node i such that the global basis function B_i vanishes on the inflow boundary Γ_{in}, we have

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{e=1}^{N_{\text{elem}}} \omega_i^e \dot{c}_i^e, \qquad \text{where} \quad \omega_i^e = \frac{m_i^e}{m_i} = \frac{m_i^e}{\sum_{e=1}^{N_{\text{elem}}} m_i^e} \in [0,1], \qquad \sum_{i=1}^{N_{\text{elem}}} \omega_i^e = 1. \tag{A.4}$$

That is, the time derivative of c_i is a convex average of the one-sided element-based approximations $\dot{c}_i^e$.

We remark that it is also possible to incorporate boundary terms into the residuals of local problems (A.1) to absorb them into the definition of $\dot{c}_i^e$ and obtain a more compact representation in terms of local time derivatives.

AppendixB. Properties of subcell-stencil upwinding

The one-dimensional Bernstein polynomials satisfy the following relationships:

$$B_j^{p-1}(x) = \frac{p-j}{p} B_j^p(x) + \frac{j+1}{p} B_{j+1}^p(x), \tag{B.1}$$

$$\frac{\mathrm{d}}{\mathrm{d}x} B_j^p(x) = p B_{j-1}^{p-1}(x) - p B_j^{p-1}(x). \tag{B.2}$$

After inserting (B.1) into (B.2), the derivative can be represented as follows:

$$\frac{\mathrm{d}}{\mathrm{d}x} B_j^p(x) = (p-j+1) B_{j-1}^p(x) + (2j-p) B_j^p(x) - (j+1) B_{j+1}^p(x) \tag{B.3}$$

and the matrix entries $\tilde{\varkappa}_{ij}^e = \sum_r m_i^e m_{ir}^{-e} \varkappa_{rj}^e$ for a constant velocity field v are given by

$$\tilde{\varkappa}_{ij}^e = \frac{v}{p+1} \begin{cases} j-1-p & : i = j-1, \\ p-2j & : i = j, \\ j+1 & : i = j+1, \\ 0 & : \text{otherwise.} \end{cases} \tag{B.4}$$

For a positive constant velocity $v \geq 0$, algebraic upwinding based on (19) yields

$$\begin{aligned} \tilde{d}_{ij}^e = \max\{-\tilde{\varkappa}_{ij}^e, 0, -\tilde{\varkappa}_{ji}^e\} &= \begin{cases} -\tilde{\varkappa}_{j-1,j}^e & : i = j-1, \\ -\tilde{\varkappa}_{j,j+1}^e & : i = j+1, \\ 0 & : \text{otherwise} \end{cases} \\ &= \frac{v}{p+1} \begin{cases} p-j+1 & : i = j-1, \\ p-j & : i = j+1, \\ 0 & : \text{otherwise} \end{cases} \qquad \text{for all } i \neq j. \end{aligned} \tag{B.5}$$

Therefore, the off-diagonal entries of the element-based low order convection matrix can be written as

$$\begin{aligned} \tilde{l}_{ij}^e = \tilde{\varkappa}_{ij}^e + \tilde{d}_{ij}^e &= \frac{v}{p+1} \begin{cases} j-1-p+p-j+1 = 0 & : i = j-1, \\ j+1+p-j = p+1 & : i = j+1, \\ 0 & : \text{otherwise} \end{cases} \\ &= \begin{cases} 0 & : i = j-1, \\ v & : i = j+1, \\ 0 & : \text{otherwise} \end{cases} \end{aligned} \tag{B.6}$$

and, hence, the diagonal entries are

$$\tilde{l}_{ii}^e = -v \qquad \text{for all } 1 \leq i \leq p \tag{B.7}$$

due to the vanishing row sum property of the element-based convection and artificial diffusion matrices.

The element matrix $\tilde{L}^e$ of the 1D upwind-biased convection operator has the tridiagonal lower triangular form

$$\tilde{L}^e = \begin{pmatrix} 0 & & & \\ v & -v & & \\ & \ddots & \ddots & \\ & & v & -v \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}. \tag{B.8}$$

The same matrix is obtained if algebraic upwinding is applied to the discrete convection operator of the piecewise-linear finite element approximation on a submesh defined by the same $p+1$ nodal points. This striking result implies

that the subcell-stencil approach to discrete upwinding produces the same global matrix L for Bernstein elements of any order p as long as the same mesh of nodal points is employed. The uniqueness of the low order convection operator L for 1D Bernstein elements can also be shown in the case of a negative constant velocity field $v < 0$. It is worth mentioning that the diagonal entries m_i of the global lumped mass do depend on p. That is why the low order scheme defined by (20) does not produce the same solutions for Bernstein polynomials of different orders.

AppendixC. Properties of element-based FCT limiters

The antidiffusive correction step (35) of an element-based FCT algorithm can be written as

$$m_i c_i^{n+1} = m_i c_i^L + \sum_{e=1}^{N_{\text{elem}}} \bar{f}_i^e = \sum_{e=1}^{N_{\text{elem}}} (m_i^e c_i^L + \bar{f}_i^e) = \sum_{e=1}^{N_{\text{elem}}} m_i^e c_i^{e,n+1}, \tag{C.1}$$

$$m_i^e c_i^{e,n+1} = m_i^e c_i^L + \bar{f}_i^e, \qquad m_i^e = \frac{|K_e|}{N_{\text{dof}}^e} \qquad \text{for all } i \in \mathcal{N}^e, \tag{C.2}$$

where N_{dof}^e is the number of local degrees of freedom. By constraining $\bar{f}_i^e$ to satisfy

$$c_i^{\min,n+1} \le c_i^{e,n+1} = c_i^L + \frac{\bar{f}_i^e}{m_i} \le c_i^{\max,n+1} \tag{C.3}$$

for each element containing node i, a local FCT limiter like (41) guarantees that

$$c_i^{\min,n+1} \le c_i^{n+1} = \sum_{e=1}^{N_{\text{elem}}} \left(\frac{m_i^e}{m_i}\right) c_i^{e,n+1} \le c_i^{\max,n+1}. \tag{C.4}$$

In other words, local maximum principles can be enforced by limiting the Bernstein coefficients element-by-element and taking advantage of the fact that c_i^{n+1} is a convex combination of $c_i^{e,n+1}$.

The finite element shape functions defined in terms of the Bernstein coefficients $c_i^{e,n+1}$ are given by

$$u_h^{e,n+1}(\mathbf{x}) \coloneqq u_h^{e,L}(\mathbf{x}) + \frac{N_{\text{dof}}^e}{|K_e|} \sum_{j=1}^{N_{\text{dof}}} \bar{f}_j^e B_j(\mathbf{x}). \tag{C.5}$$

The equivalent Taylor basis representation [35] of the shape function $u_h^{e,n+1}$ in terms of the cell average $\bar{u}_h^{e,n+1}$ and partial derivatives $\mathrm{D}^{\mathbf{q}} u_h^{e,n+1}(\bar{\mathbf{x}}_e)$, $\mathbf{q} = (q_1, \ldots, q_d)$ at the center of mass $\bar{\mathbf{x}}_e$ reads

$$u_h^{e,n+1}(\mathbf{x}) = \bar{u}_h^{e,n+1} + \sum_{1 \le |\mathbf{q}| \le p} \mathrm{D}^{\mathbf{q}} u_h^{e,n+1}(\bar{\mathbf{x}}_e) \frac{(\mathbf{x} - \bar{\mathbf{x}}_e)^{\mathbf{q}} - \overline{(\mathbf{x} - \bar{\mathbf{x}}_e)^{\mathbf{q}}}}{\mathbf{q}!}. \tag{C.6}$$

In this formula, the overline denotes averaging over K_e and the multiindex notation is employed. We have

$$\int_{K_e} B_j(\mathbf{x}) \mathrm{d}\mathbf{x} = m_j^e = \frac{|K_e|}{N_{\text{dof}}^e}, \qquad \int_{K_e} u_h^{e,n+1}(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{K_e} u_h^{e,L}(\mathbf{x}) \mathrm{d}\mathbf{x} + \sum_{j=1}^{N_{\text{dof}}} \bar{f}_j^e. \tag{C.7}$$

By the zero sum property of $\bar{f}^e$, the cell averages of $u_h^{e,n+1}$ and $u_h^{e,L}$ are equal. It follows that element-based FCT is equivalent to element-by-element correction of derivatives followed by a projection of the discontinuous solution into the continuous finite element space V_h. This is not the way in which FCT algorithms should be implemented in practice but the above interpretation reveals an interesting relationship to slope-limited DG methods [17, 35].