Monotone local projection stabilization schemes for continuous finite elements

Dmitri Kuzmin[a], Steffen Basting[a], John N. Shadid[b,c]

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

[b] *Computational Mathematics Department, Sandia National Laboratories P.O. 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 paper presents a new approach to enforcing discrete maximum principles and/or positivity preservation in continuous piecewise-linear finite element approximations to convection-dominated transport problems. Using a linear first-order advection equation as a model problem, we construct element-level bilinear forms associated with first-order artificial diffusion operators and their higher-order counterparts. The underlying design philosophy is similar to that behind local projection stabilization (LPS) techniques and variational multiscale (VMS) methods. The difference lies in the structure of the local stabilization operator and in the way in which the resolved scales are detected. The proposed stabilization term penalizes the difference between the nodal values and cell averages of the finite element solution in a manner which guarantees monotonicity and linearity preservation. The optimal value of the stabilization parameter is determined using a new multidimensional limiter function designed to prevent unresolvable fine scale effects from creating undershoots or overshoots. The result is a nonlinear high-resolution

Email addresses: kuzmin@math.math-dortmund.de (Dmitri Kuzmin), steffen.basting@math.tu-dortmund.de (Steffen Basting), jnshadi@sandia.gov (John N. Shadid)

scheme capable of resolving moving fronts and internal/boundary layers as sharp localized nonoscillatory features. The use of variational gradient recovery makes it possible to add high-order background dissipation leading to improved approximation properties in smooth regions. The numerical behavior of the constrained LPS schemes is illustrated by a grid convergence study for stationary and time-dependent test problems in two space dimensions.

Keywords: finite element methods, local projection stabilization, discrete maximum principles, artificial diffusion, limiters, linearity preservation

---

1. Introduction

The Galerkin finite element discretization of convection-dominated transport equations is known to produce numerical approximations that may violate the discrete maximum principle and/or the criterion of positivity preservation on meshes that are too coarse to resolve certain fine-scale features (moving fronts, interior and boundary layers). The most common approach to avoiding nonphysical undershoots and overshoots in finite element methods is based on the use of nonlinear shock-capturing terms within the framework of variationally consistent Petrov-Galerkin methods (see, e.g., [8, 22, 23] for a review and comparative study of existing schemes). Additionally entropy viscosity approaches that attempt to control oscillations by introducing artificial dissipation that is based on an auxiliary entropy production residual have been proposed [18]. The main drawback of many existing approaches is the presence of problem-dependent free parameters along with the lack of provable nonlinear stability properties such as positivity and monotonicity preservation on general meshes. While small spurious oscillations can be tolerated in some applications, many models are very sensitive to nonphysical values of the transported variable. For this reason, the use of physics-compatible finite element approximations may be appropriate or even indispensable.

Nonlinear shock-capturing operators backed by the theory of discrete maximum principles (DMP) were recently developed and analyzed in [1, 2, 6, 7, 14]. In the case of [1] and [7], the proof of the DMP property imposes restrictions of sufficient mesh regularity. Moreover, mass lumping is required in applications to transient problems [1]. The explicit second-order method

proposed in [14] satisfies a discrete maximum principle for arbitrary meshes and employs the consistent mass matrix. The way in which [14] enforces DMP constraints is closely related to the concept of *algebraic flux correction* [27]. This approach provides a general framework for the design of artificial diffusion operators that render a finite element discretization *local extremum diminishing* (LED) or positivity-preserving. In nonlinear high-resolution schemes based on algebraic flux correction, the antidiffusive part of a high-order discretization is constrained using a *limiter*.

The most prominent representative of algebraic flux correction schemes is the *flux-corrected transport* (FCT) algorithm introduced by Boris and Book [5] and Zalesak [48] in the context of explicit finite difference schemes. Unlike many other limiting techniques, FCT can be extended to finite element approximations using conservative decompositions of the antidiffusive term into internodal fluxes associated with edges of the sparsity graph [44, 9, 39, 33, 25] or element contributions associated with individual mesh cells [37, 38, 33]. The simplicity and efficiency of predictor-corrector FCT schemes make them very attractive in situations when the problem is time-dependent and the time steps are small [23]. Additionally, edge-based generalizations of *total variation diminishing* (TVD) schemes [15, 16] can be designed using reconstruction of 1D stencils [40, 43, 36] or algebraic flux correction schemes proposed in [26, 27]. In contrast to FCT, multidimensional extensions of TVD limiters are directly applicable to stationary transport equations and produce steady-state solutions independent of the time step.

In this paper, we bridge the gap between variational shock capturing methods and algebraic flux correction schemes of FCT and TVD type by introducing local bilinear forms that lead to monotone element-level corrections of the Galerkin variational formulation. Similarly to *local projection stabilization* (LPS) methods [4, 41, 42, 45], these operators are designed to introduce numerical dissipation acting on unresolvable fine scales. In contrast to the standard LPS approach, the proposed local bilinear forms achieve the desired effect by penalizing the difference between linear shape functions and their limited counterparts. The new element-based limiting strategy guarantees positivity and accuracy preservation on simplex meshes. The optional definition of the target shape function in terms of reconstructed nodal gradients makes it possible to introduce high-order LPS stabilization which results in better convergence rates for smooth solutions. In applications to time-

dependent problems, the time derivatives are stabilized using the same LPS bilinear form which is shown to be equivalent to selective mass lumping. Additionally, we describe a way to reduce phase errors by enforcing a local discrete maximum principle for the nodal time derivatives. The paper concludes with a numerical study for two-dimensional transport problems.

2. Continuous problem

As a linear model problem, consider the time-dependent linear first-order advection equation

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

where $u : \Omega \times \mathbb{R}_+ \mapsto \mathbb{R}$ is the conserved quantity, $\mathbf{v} : \Omega \times \mathbb{R}_+ \mapsto \mathbb{R}^d$ is a given velocity field, and Ω is a bounded domain in $\mathbb{R}^d,\ d \in \{1, 2, 3\}$.

The initial condition for the linear advection model is given by

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega. \tag{2}$$

At the inlet $\Gamma_{\text{in}} := \{\mathbf{x} \in \Gamma \,|\, \mathbf{v} \cdot \mathbf{n} < 0\}$, where $\mathbf{n}$ is the unit outward normal to the boundary $\Gamma := \partial\Omega$, a Dirichlet boundary condition is prescribed

$$u = u_{\text{in}} \quad \text{on } \Gamma_{\text{in}}. \tag{3}$$

Let $\Sigma := \{(\mathbf{x}, t) \,|\, \mathbf{x} \in \Gamma_{\text{in}} \vee t = 0\}$ denote the set of points and time instants such that $u(\mathbf{x}, t)$ is known from the data prescribed in (2) or (3).

The solution to problem (1)–(3) is known to be positivity-preserving, i.e.,

$$u(\mathbf{x}, t) \geq 0 \quad \forall (\mathbf{x}, t) \in \Sigma \quad \Rightarrow \quad u(\mathbf{x}, t) \geq 0 \quad \forall (\mathbf{x}, t) \in \bar{\Omega} \times \mathbb{R}_+, \tag{4}$$

which can be easily shown using the method of characteristics.

Moreover, the following maximum principle holds in the case $\nabla \cdot \mathbf{v} = 0$:

$$\min_{\Sigma} u \leq u(\mathbf{x}, t) \leq \max_{\Sigma} u. \tag{5}$$

If $\nabla \cdot \mathbf{v} \neq 0$, then the solution to (1)–(2) will satisfy (4) but may violate (5).

3. Galerkin discretization

Multiplying the governing equation (1) by a test function w, integrating by parts and invoking the Dirichlet boundary condition (3), we obtain

$$\int_\Omega \left[w\frac{\partial u}{\partial t} - \nabla w \cdot (\mathbf{v}u) \right] \mathrm{d}\mathbf{x} + \int_{\Gamma_{\text{out}}} wu(\mathbf{v}\cdot\mathbf{n})\mathrm{d}s = -\int_{\Gamma_{\text{in}}} wu_{\text{in}}(\mathbf{v}\cdot\mathbf{n})\mathrm{d}s, \quad (6)$$

where $\Gamma_{\text{out}} := \{\mathbf{x} \in \Gamma \,|\, \mathbf{v}\cdot\mathbf{n} > 0\}$ is the outflow boundary. The so-defined variational formulation with weakly imposed boundary conditions is globally conservative since it reduces to the integral form of (1) in the case $w \equiv 1$.

Let $\{\varphi_1, \ldots, \varphi_N\}$ be a set of global piecewise-linear basis functions associated with vertices of a (possibly unstructured) simplex mesh $\mathcal{T}_h$. Substituting

$$u_h(\mathbf{x}, t) = \sum_{j=1}^{N} u_j(t)\varphi_j(\mathbf{x}) \quad (7)$$

into (6), one obtains the Galerkin discretization which can be written as

$$(w_h, \dot{u}_h) + a(w_h, u_h) = b(w_h) \qquad \forall w_h \in \{\varphi_1, \ldots, \varphi_N\}, \quad (8)$$

where

$$\begin{aligned}
(w_h, \dot{u}_h) &= \int_\Omega w_h \dot{u}_h \,\mathrm{d}\mathbf{x}, \qquad \dot{u}_h = \frac{\partial u_h}{\partial t}, \\
a(w_h, u_h) &= -\int_\Omega \nabla w_h \cdot (\mathbf{v}u_h)\,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\text{out}}} w_h u_h(\mathbf{v}\cdot\mathbf{n})\mathrm{d}s, \\
b(w_h) &= -\int_{\Gamma_{\text{in}}} w_h u_{\text{in}}(\mathbf{v}\cdot\mathbf{n})\mathrm{d}s.
\end{aligned}$$

The semi-discrete Galerkin equation associated with $w_h = \varphi_i$ is given by

$$\sum_{j\in\mathcal{N}(i)} m_{ij}\frac{\mathrm{d}u_j}{\mathrm{d}t} = \sum_{j\in\mathcal{N}(i)} k_{ij}u_j + g_i, \quad (9)$$

where $\mathcal{N}(i) := \{j \,:\, m_{ij} \neq 0\}$ is the stencil of node i,

$$m_{ij} = (\varphi_i, \varphi_j), \quad k_{ij} = -a(\varphi_i, \varphi_j), \quad g_i = b(\varphi_i).$$

The global matrix form of the semi-discrete finite element scheme reads

$$M_C \frac{\mathrm{d}u}{\mathrm{d}t} = Ku + g, \tag{10}$$

where u is the vector of nodal values, M_C is the consistent mass matrix, K is the discrete convection operator and g is a vector of fluxes across Γ_{in}.

4. Low-order stabilization

The analysis of matrix properties in [27] reveals that the oscillatory behavior of the Galerkin discretization (10) is due to the fact that some off-diagonal entries of the consistent mass matrix M_C are nonzero ($\exists j \neq i \,:\, m_{ij} > 0$) and some off-diagonal entries of the matrix K are negative ($\exists j \neq i \,:\, k_{ij} < 0$).

To enforce the discrete maximum principle, we perform row-sum mass lumping and modify the bilinear form $a(\cdot,\cdot)$ by adding the stabilization term

$$s(w_h, u_h) := \sum_T \nu_T s_T(w_h, u_h), \tag{11}$$

where ν_T is an artificial diffusion coefficient (to be defined below) and

$$s_T(w_h, u_h) = \int_T w_h(u_h - \bar{u}_T)\,\mathrm{d}\mathbf{x} \tag{12}$$

is designed to penalize the difference between u_h and the average

$$\bar{u}_T = \frac{\int_T u_h\,\mathrm{d}\mathbf{x}}{\int_T 1\,\mathrm{d}\mathbf{x}}. \tag{13}$$

Using a similar definition for the average $\bar{w}_T$ of the test function w_h, the local bilinear form (12) of the stabilization term (11) for the piecewise-linear finite element approximation can be written as

$$s_T(w_h, u_h) = \int_T (w_h - \bar{w}_T)(u_h - \bar{u}_T)\,\mathrm{d}\mathbf{x} \tag{14}$$

$$= \int_T (\mathbf{x} - \mathbf{x}_T) \cdot \nabla w_h (\mathbf{x} - \mathbf{x}_T) \cdot \nabla u_h, \tag{15}$$

where

$$\bar{\mathbf{x}}_T = \frac{\int_T \mathbf{x}\,\mathrm{d}\mathbf{x}}{\int_T 1\,\mathrm{d}\mathbf{x}} \tag{16}$$

denotes the center of mass of element T. It follows that (12) represents an anisotropic diffusion operator acting in the direction $\Delta\mathbf{x}_T := \mathbf{x} - \mathbf{x}_T$.

Remark. It is worth mentioning that on simplex meshes the local bilinear form $s_T(\cdot,\cdot)$ is proportional to the one considered by Guermond et al. [14].

In Appendix A, we analyze further properties of $s_T(\cdot,\cdot)$ and show that

$$s_T(\varphi_i,\varphi_j) = \frac{1}{d+1}\left(\int_T \varphi_i\,\mathrm{d}\mathbf{x} - \int_T \varphi_i\varphi_j\,\mathrm{d}\mathbf{x}\right) \tag{17}$$

in d space dimensions. That is, the discrete stabilization operator is proportional to the difference between the lumped and consistent mass matrices.

Let $M_C^{(e)} = \{m_{ij}^{(e)}\}$ and $K^{(e)} = \{k_{ij}^{(e)}\}$ denote the element matrices that represent the contribution of a given element $T_e \in \mathcal{T}_h$ to the global matrices M_C and K of the standard Galerkin discretization (10). By (17), the element matrix associated with the local bilinear form $s_{T_e}(\cdot,\cdot)$ is given by

$$S^{(e)} = \frac{1}{d+1}\left(M_L^{(e)} - M_C^{(e)}\right),$$

where $M_L^{(e)}$ denotes the lumped element mass matrix. That is,

$$M_L^{(e)} = \mathrm{diag}\left(\int_{T_e} \varphi_i\,\mathrm{d}\mathbf{x}\right) = \mathrm{diag}\left(\sum_j m_{ij}^{(e)}\right).$$

The stabilized counterpart of the element matrix $K^{(e)}$ is given by

$$L^{(e)} = K^{(e)} - \nu_{T_e} S^{(e)}. \tag{18}$$

By construction, the element matrix $S^{(e)}$ is symmetric with zero row and column sums. All off-diagonal entries of this matrix are negative. To eliminate all negative off-diagonal entries of $K^{(e)}$ we define ν_{T_e} as follows:

$$\nu_{T_e} = \max_{j\neq i} \frac{\max\{0, -k_{ij}^{(e)}\}}{|s_{ij}^{(e)}|}. \tag{19}$$

After the global matrix assembly, the semi-discrete problem becomes

$$M_L \frac{\mathrm{d}u}{\mathrm{d}t} = Lu + g, \tag{20}$$

where M_L is the global lumped mass matrix and L is the stabilized discrete transport operator assembled from element matrices $L^{(e)}$ defined by (18).

The equation associated with an interior node $\mathbf{x}_i \in \Omega$ can be written as

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}(i)} l_{ij} u_j, \tag{21}$$

where $m_i = \int_\Omega \varphi_i \, \mathrm{d}\mathbf{x}$ is a positive diagonal entry of the lumped mass matrix M_L. By definition of $L^{(e)}$, we have $l_{ij} \geq 0$ for all $j \neq i$. It follows that the modified semi-discrete scheme is *positivity-preserving*, i.e.,

$$u_i(0) \geq 0 \quad \forall i \quad \Rightarrow \quad u_i(t) \geq 0 \quad \forall i \quad \forall t > 0. \tag{22}$$

For a formal proof of this result we refer the reader to Theorem 7.1 in [19].

If additionally we have $\sum_{j \in \mathcal{N}(i)} l_{ij} = 0$, then equation (21) reduces to

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}(i) \backslash \{i\}} l_{ij}(u_j - u_i). \tag{23}$$

Due to the fact that $m_i > 0$ and $l_{ij} \geq 0$, we have

$$u_i \geq u_j \quad \forall j \in \mathcal{N}(i) \backslash \{i\} \qquad \Rightarrow \qquad \frac{\mathrm{d}u_i}{\mathrm{d}t} \leq 0, \tag{24}$$

$$u_i \leq u_j \quad \forall j \in \mathcal{N}(i) \backslash \{i\} \qquad \Rightarrow \qquad \frac{\mathrm{d}u_i}{\mathrm{d}t} \geq 0. \tag{25}$$

Hence, a local maximum cannot increase, and a local minimum cannot decrease. A discretization satisfying this semi-discrete maximum principle is called *local extremum diminishing* (LED) [20, 21, 27]. It can be easily verified that LED implies positivity preservation, but the converse is not true.

5. High-order stabilization

By the Godunov theorem [10], linear positivity-preserving and LED schemes can be at most first-order accurate. Following the design philosophy behind

local projection stabilization (LPS) [4, 41, 42, 45], two-level variational multiscale (VMS) methods [24] and slope-limited discontinuous Galerkin (DG) approximations [28, 29], we will now modify the stabilization term and the mass lumping operator so as to restrict their dissipative effect to unresolvable fine-scale components by antidiffusing the harmless resolvable scales.

Let $\tilde{u}_h|_T$ denote the resolvable component of the shape function $u_h|_T$. This will be our *target*. Consider the local projection stabilization operator

$$\tilde{s}_T(w_h, u_h) = \int_T w_h(u_h - \tilde{u}_h)\,\mathrm{d}\mathbf{x}. \tag{26}$$

Note that definition (13) corresponds to using $\tilde{u}_h|_T = \bar{u}_T$. The target for a high-order LPS stabilization operator is a shape function of the form

$$\tilde{u}_h(\mathbf{x}) = \bar{u}_T + \alpha_T \mathbf{g}_T(\mathbf{x}) \cdot (\mathbf{x} - \bar{\mathbf{x}}_T), \quad \mathbf{x} \in T, \tag{27}$$

where $\mathbf{g}_T$ is a suitable approximation to the gradient $\nabla u|_T$, and α_T is a solution-dependent correction factor to be defined in Section 7.

In particular, the choice $\alpha_T = 0$ corresponds to the positivity-preserving low-order approximation based on (13). The choice $\alpha_T = 1$, $\mathbf{g}_T = \nabla u_h|_T$ corresponds to the standard Galerkin approximation which exhibits suboptimal convergence behavior even for transport problems with smooth solutions. To retain a certain amount of high-order background dissipation in the case $\alpha_T = 1$, the target $\tilde{u}_h|_T$ may be defined using an averaged gradient $\mathbf{g}_h$ instead of ∇u_h. For example, John et al. [24] used this idea to design a linear stabilization operator for their two-scale variational multiscale method.

In this paper, we define (27) using a *target selector* $\omega \in [0, 1]$ to construct

$$\mathbf{g}_T = (1 - \omega)\nabla u_h|_T + \omega \mathbf{g}_h(\mathbf{x}_T), \tag{28}$$

where $\mathbf{g}_h = \sum_j \mathbf{g}_j \varphi_j$ is defined in terms of the nodal gradients [26, 27]

$$\mathbf{g}_i = \frac{1}{m_i} \sum_{j \neq i} \mathbf{c}_{ij}(u_j - u_i), \qquad \mathbf{c}_{ij} = \int_\Omega \varphi_i \nabla \varphi_j \,\mathrm{d}\mathbf{x}.$$

This gradient recovery technique corresponds to the lumped-mass L^2 projection of the piecewise-constant gradient ∇u_h. Importantly, it satisfies the discrete maximum principle [17] and is exact for linear functions [27].

The difference between the shape functions $u_h|_T$ and $\tilde{u}_h|_T$ is given by

$$u_h(\mathbf{x}) - \tilde{u}_h(\mathbf{x}) = (\nabla u_h|_T - \mathbf{g}_T) \cdot (\mathbf{x} - \bar{\mathbf{x}}_T), \qquad \mathbf{x} \in T. \tag{29}$$

Setting $\alpha_T = 1$, one obtains the high-order LPS stabilization operator

$$\begin{aligned} \tilde{s}_T(w_h, u_h) &= \int_T w_h (\nabla u_h|_T - \mathbf{g}_T) \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_T) \, \mathrm{d}\mathbf{x} \\ &= \omega \int_T w_h (\nabla u_h|_T - \mathbf{g}_h(\mathbf{x}_T)) \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_T) \, \mathrm{d}\mathbf{x}. \end{aligned}$$

In contrast to free parameters used in traditional shock-capturing schemes for finite elements, the blending factor ω represents a high-order scheme selector rather than an *ad hoc* stabilization parameter. Any value $\omega \in (0, 1]$ corresponds to adding consistent high-order LPS stabilization to the Galerkin scheme. If the numerical solution u_h is locally linear, then the nodal gradients are exact, whence $\mathbf{g}_T = \nabla u_h|_T$ and $\tilde{s}_T(w_h, u_h) = 0$ in the case $\alpha_T = 1$. In general, larger values of the parameter ω result in stronger smoothing without degrading the rate of convergence to the exact solution.

Clearly, the ability of the generalized stabilization term to detect and handle unresolvable fine-scale features depends on the choice of the element-based correction factors α_T for the local bilinear forms. As a rule of thumb, the choice $\alpha_T = 0$ is appropriate in elements containing a local extremum, whereas $\alpha_T = 1$ is appropriate in regions where u_h is smooth. The formula presented in Section 7 guarantees that the antidiffusive element contributions are constrained in a way which rules out any violations of the discrete maximum principle while maintaining low levels of numerical dissipation.

6. Antidiffusive element contributions

The element-by-element insertion of local stabilization terms into the Galerkin discretization leads to a constrained global system of the form

$$M_L \frac{\mathrm{d}u}{\mathrm{d}t} = Lu + f(u) + g. \tag{30}$$

The matrices M_L and L correspond to the positivity-preserving low-order approximation (20), whereas the limited antidiffusive correction term $f(u)$ is

assembled from element vectors associated with resolvable components

$$f^{(e)} = \min\{\alpha_{T_e}, \beta_{T_e}\}(M_L^{(e)} - M_C^{(e)})\dot{u}^{(e)} + \alpha_{T_e}\nu_{T_e}S^{(e)}u^e, \tag{31}$$

where u^e is the vector of local degrees of freedom

$$u_i^{(e)} = \bar{u}_{T_e} + \mathbf{g}_T \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_{T_e}),$$

and $\dot{u}^{(e)}$ is the vector of nodal time derivatives. The additional element-based correction factor β_{T_e} is introduced to switch off the contribution of the consistent mass matrix in steady state computations or reduce phase errors in applications to time-dependent transport problems (see Section 8).

Since the element matrices $S^{(e)}$ and $M_L^{(e)} - M_C^{(e)} = (d+1)S^{(e)}$ have zero column sums, the components of the corresponding matrix-vector products sum to zero. Hence, the multiplication by an arbitrary correction factor α_{T_e} has the property of being a conservative correction to the local gradient.

For the nonlinear scheme (30) to be positivity-preserving, we must have $f_i(u) \le 0$ at a local maximum and $f_i(u) \ge 0$ at a local minimum [27, 34]. If u_i is not a local extremum, the LED property follows from the fact that the limited antidiffusive term can be written as $f_i(u) = c_i(u_i^{\min} - u_i)$ or $f_i(u) = c_i(u_i^{\max} - u_i)$, where c_i is a positive bounded coefficient and

$$u_i^{\min} = \min_{j\in\mathcal{N}(i)} u_j, \tag{32}$$

$$u_i^{\max} = \max_{j\in\mathcal{N}(i)} u_j \tag{33}$$

are the local minimum and maximum over the stencil $\mathcal{N}(i)$ of node i [34].

To enforce the LED constraint for all degrees of freedom, we introduce nodal correction factors Φ_i such that $\Phi_i = 0$ at a local extremum and define the element-based correction factor α_T as the generalized harmonic mean

$$\alpha_T = \frac{(d+1)\prod_{i\in\mathcal{V}(T)}\Phi_i}{\sum_{j\in\mathcal{V}(T)}\prod_{i\in\mathcal{V}(T)\backslash\{j\}}\Phi_i + \epsilon}, \tag{34}$$

where $\mathcal{V}(T)$ is the set of $d+1$ nodes of T. Here and below ϵ denotes a small positive constant added to prevent division by zero and ensure continuity. All numerical studies in Section 10 were performed using $\epsilon = 10^{-15}$.

In the case $d = 1$, we have

$$\alpha_T = \frac{2\Phi_i\Phi_j}{\Phi_i + \Phi_j + \epsilon} \tag{35}$$

for a 1D element with nodes i and j. In the 2D case, the harmonic mean limiter for a triangle T with nodes i, j, k is given by the formula

$$\alpha_T = \frac{3\Phi_i\Phi_j\Phi_k}{\Phi_i\Phi_j + \Phi_j\Phi_k + \Phi_k\Phi_i + \epsilon}. \tag{36}$$

In contrast to element-based FCT limiters [33, 37, 38], α_T is a differentiable function of nodal correction factors and does not depend on the signs of $f_i^{(e)}$. It can be interpreted as a generalized averaging operator of Van Leer type (see [20, 21] for a definition and discussion of LED limited averages).

The value of Φ_i may be determined using the following design principles:

- $\Phi_i \in [0, 1]$ depends continuously on the nodal values u_j, $j \in \mathcal{N}(i)$;
- $\Phi_i = 0$ at a local maximum ($u_i = u_i^{\max}$) or minimum ($u_i = u_i^{\min}$);
- $\Phi_i = 1$ if u_h is linear on the patch of elements $\Omega_i = \text{supp}(\varphi_i)$.

The first property is needed to make sure that the discrete problem is well-posed [2, 3]. The second property and definition (34) guarantee that the sum of antidiffusive element contributions to node i is local extremum diminishing. The third property is known as *linearity preservation* [2, 26, 27] and is a useful criterion for preserving second-order accuracy in smooth regions.

7. Design of the nodal limiter

Examples of nodal limiter functions satisfying (some of) the above design criteria can be found, e.g., in [1, 2, 34]. Adapting these limiters to the structure of our LPS stabilization operator (12), we consider

$$\Phi_i = 1 - \frac{\left|\sum_T \int_T \varphi_i (u_h - \bar{u}_T)\, \mathrm{d}\mathbf{x}\right| + \epsilon}{\sum_T \left|\int_T \varphi_i (u_h - \bar{u}_T)\, \mathrm{d}\mathbf{x}\right| + \epsilon}. \tag{37}$$

If u_i is a local extremum, then all integrals in (37) have the same sign (see Appendix A) and the absolute value of their sum equals the sum of the absolute values. It follows that $\Phi_i = 0$ at a local extremum in accordance with the LED principle. In Appendix B we also show that (37) is linearity-preserving on patches Ω_i satisfying certain geometric conditions. Note that the sum of integrals in the numerator is the residual of the L^2 projection

$$\sum_T \int_T \varphi_i u_h \,\mathrm{d}\mathbf{x} = \sum_T \int_T \varphi_i \bar{u}_T \,\mathrm{d}\mathbf{x}, \quad i = 1, \ldots, N.$$

The superconvergence property of piecewise-linear reconstructions from cell averages [45, 46] implies that the residual is small for smooth functions. However, the limiter based on (37) does not guarantee exact linearity preservation on general meshes. To rectify this, we generalize (37) as follows:

$$\Phi_i = 1 - \frac{\left|\sum_T \sigma_{i,T} \int_T \varphi_i (u_h - \bar{u}_T) \,\mathrm{d}\mathbf{x}\right| + \epsilon}{\sum_T \sigma_{i,T} \left|\int_T \varphi_i (u_h - \bar{u}_T) \,\mathrm{d}\mathbf{x}\right| + \epsilon}. \tag{38}$$

At each interior node $\mathbf{x}_i$, the positive weights $\sigma_{i,T} > 0$ are defined so that

$$\sum_T \sigma_{i,T} \int_T \varphi_i (u_h - \bar{u}_T) \,\mathrm{d}\mathbf{x} = 0 \tag{39}$$

whenever $\nabla u_h|_T = \mathbf{g}_i$ for all elements T of the patch Ω_i. The sums

$$S_i^+ = \sum_T \max\left\{0, \int_T \varphi_i \mathbf{g}_i \cdot (\mathbf{x} - \bar{\mathbf{x}}_T) \,\mathrm{d}\mathbf{x}\right\}, \tag{40}$$

$$S_i^- = \sum_T \min\left\{0, \int_T \varphi_i \mathbf{g}_i \cdot (\mathbf{x} - \bar{\mathbf{x}}_T) \,\mathrm{d}\mathbf{x}\right\} \tag{41}$$

can be balanced using the nodal correction factors

$$\sigma_i^+ = \min\left\{1, \frac{|S_i^-| + \epsilon}{S_i^+ + \epsilon}\right\}, \qquad \sigma_i^- = \min\left\{1, \frac{S_i^+ + \epsilon}{|S_i^-| + \epsilon}\right\} \tag{42}$$

to enforce the zero sum condition

$$\sigma_i^+ S_i^+ + \sigma_i^- S_i^- = 0.$$

Note that $S_i^+ > 0$ and $S_i^- < 0$ at any interior node $\mathbf{x}_i$ unless $\mathbf{g}_i = \mathbf{0}$.

If u_h is locally linear, then $\nabla u_h|_T = \mathbf{g}_i$ for all T and (39) holds for

$$\sigma_{i,T} = \begin{cases} \sigma_i^+ & \text{if } \int_T \varphi_i (u_h - \bar{u}_T)\, \mathrm{d}\mathbf{x} > 0, \\ \sigma_i^- & \text{if } \int_T \varphi_i (u_h - \bar{u}_T)\, \mathrm{d}\mathbf{x} < 0. \end{cases} \tag{43}$$

This is sufficient to guarantee that (38) yields $\Phi_i = 1$ for any $u_h \in P_1(\Omega_i)$.

We remark that any nodal limiter function that can be written in the form

$$\Phi_i = 1 - \frac{P_i}{Q_i}, \qquad 0 \le P_i \le Q_i > 0 \tag{44}$$

can be modified to produce $\Phi_i = 1$ not only in the case $P_i = 0$ but also for sufficiently small values of the numerator P_i. To that end, consider [34]

$$\Phi_i = 1 - \frac{\max\{0, P_i - \gamma Q_i\}}{(1-\gamma) Q_i}, \qquad \gamma \in [0, 1). \tag{45}$$

This modification preserves the property that $\Phi_i = 0$ for $P_i = Q_i$ and $\Phi_i = 1$ for $P_i = 0$. Choosing a larger value of γ makes the limiter less diffusive but increases the number of fixed-point iterations when it comes to solving the nonlinear algebraic system. In the numerical study below, we use $\gamma = \frac{3}{4}$.

Remark. Another way to make a limiter of the form (44) less diffusive is to replace the ratio P_i/Q_i by $(P_i/Q_i)^q$ with $q > 1$, see, e.g., [1, 2]. This modification has the same antidiffusive effect as scaling by γ in (45).

8. Correction of the time derivatives

In applications to unsteady transport equations, we use an additional correction factor β_{T_e} in (31) to limit the time derivatives in antidiffusive element contributions associated with mass lumping. This optional correction leads to cosmetic improvements in situations when an oscillatory target generates large phase errors leading to optically disturbing ripples in the constrained solutions (see Section 10.1). In the literature on flux-corrected transport algorithms this phenomenon is known as *terracing*. It can be cured by using a better target or suitable *prelimiting* of the antidiffusive components [27].

Let $\dot{u}^C$ denote the vector of constrained time derivatives that corresponds to

$$\dot{u}^C = M_L^{-1}(Lu + f^M + f^S + g), \tag{46}$$

where f^M and f^S are assembled from the antidiffusive element vectors

$$f^{(e),M} = \min\{\alpha_{T_e}, \beta_{T_e}\}(M_L^{(e)} - M_C^{(e)})\dot{u}^{(e)}$$

and

$$f^{(e),S} = \alpha_{T_e}\nu_{T_e}S^{(e)}u^e,$$

respectively. The corresponding lumped-mass approximation is given by

$$\dot{u}^L = M_L^{-1}(Lu + f^S + g). \tag{47}$$

We have

$$\dot{u}^C = \dot{u}^L + M_L^{-1}f^M. \tag{48}$$

Hence, the contribution of f^M can be interpreted as a high-order correction to $\dot{u}^L$. In order to constrain the changes of the time derivatives due to this correction, we choose β_{T_e} so as to enforce the inequality constraints

$$\dot{u}_i^{\min} \leq \dot{u}^C \leq \dot{u}_i^{\max},$$

where $\dot{u}_i^{\min}$ and $\dot{u}_i^{\max}$ denote the local maxima and minima of $\dot{u}^L$, i.e.,

$$\dot{u}_i^{\min} = \min_{j\in\mathcal{N}(i)} \dot{u}_j^L, \tag{49}$$

$$\dot{u}_i^{\max} = \max_{j\in\mathcal{N}(i)} \dot{u}_j^L. \tag{50}$$

Substituting $\dot{u}^L$ for the time derivative in the constrained element vector

$$f^{(e),M} \approx \min\{\alpha_{T_e}, \beta_{T_e}\}(M_C^{(e)} - M_L^{(e)})\dot{u}^L,$$

we use a local version of the element-based FCT algorithm [33] to calculate

$$\beta_{T_e} = \min_{i\in\mathcal{V}(e)} \Psi_i^{(e)}, \tag{51}$$

where $\Psi_i^{(e)}$ are the nodal correction factors defined by

$$\Psi_i^{(e)} = \begin{cases} \min\left\{1, \frac{m_i^{(e)}(\dot{u}_i^{\max} - \dot{u}_i^L)}{f_i^{(e),M}}\right\} & \text{if} \quad f_i^{(e),M} > 0, \\ 1 & \text{if} \quad f_i^{(e),M} = 0, \\ \min\left\{1, \frac{m_i^{(e)}(\dot{u}_i^{\min} - \dot{u}_i^L)}{f_i^{(e),M}}\right\} & \text{if} \quad f_i^{(e),M} < 0, \end{cases} \tag{52}$$

where $m_i^{(e)} = \int_{T_e} \varphi_i \, \mathrm{d}\mathbf{x}$ is the ith diagonal entry of the element matrix $M_L^{(e)}$.

The above choice of β_{T_e} implies that the limited element contributions satisfy

$$m_i^{(e)}(\dot{u}_i^{\min} - \dot{u}_i^L) \le f_i^{(e),M} \le m_i^{(e)}(\dot{u}_i^{\max} - \dot{u}_i^L).$$

Summing over all elements containing node i, it is easy to verify that the values of $\dot{u}^C$ are bounded by the local extrema $\dot{u}_i^{\max}$ and $\dot{u}_i^{\min}$.

9. Time discretization and positivity

For the fully discrete scheme to inherit the LED property of a given space discretization, the time-stepping method must be consistent with the discrete maximum principle, at least under certain time step restrictions. For example, consider (23) discretized in time using the two-level θ scheme

$$m_i \frac{u_i^{n+1} - u_i^n}{\Delta t} = \theta \sum_{j \in \mathcal{N}(i)\setminus\{i\}} l_{ij}(u_j^{n+1} - u_i^{n+1}) \tag{53}$$

$$+ (1-\theta) \sum_{j \in \mathcal{N}(i)\setminus\{i\}} l_{ij}(u_j^n - u_i^n), \tag{54}$$

where $u_i^n \approx u(\mathbf{x}_i, t^n)$ denotes an approximate solution value at the time level $t^n = n\Delta t$ and $\theta \in [0,1]$ is the implicitness parameter.

The solution of the fully discrete problem (54) satisfies a discrete maximum principle if u_i^{n+1} is bounded by the maximum and minimum of the other nodal values that appear in (54). For a LED space discretization, this will be the case if the time step Δt satisfies the CFL-like condition [33, 27]

$$\frac{1}{\Delta t} \ge (1-\theta) \sum_{j \in \mathcal{N}(i)\setminus\{i\}} l_{ij} \qquad \forall i = 1, \ldots, N. \tag{55}$$

The fully implicit backward Euler method ($\theta = 1$) preserves the LED property for arbitrary time steps. The Crank-Nicolson ($\theta = \frac{1}{2}$) and forward Euler ($\theta = 0$) time discretizations are LED for time steps satisfying (55).

The use of the θ scheme as a time stepping method for (30) leads to

$$M_L \frac{u^{n+1} - u^n}{\Delta t} = \theta(Lu^{n+1} + f^{n+1}) + (1-\theta)(Lu^n + f^n) + g. \tag{56}$$

In the case $\theta < 1$, this discretization is positivity-preserving provided that

$$\frac{1}{\Delta t} \geq (1-\theta)\left[\sum_{j \in \mathcal{N}(i)\backslash\{i\}} l_{ij} + c_i^n\right] \qquad \forall i = 1, \ldots, N, \tag{57}$$

where $c_i \geq 0$ is defined by the LED representation $f_i = c_i(u_i^{\min} - u_i)$ or $f_i = c_i(u_i^{\max} - u_i)$ of the limited antidiffusive term.

Remark. The semi-discrete nature of the proposed approach makes it possible to use a wide range of time discretizations including explicit and implicit strong stability preserving (SSP) Runge-Kutta schemes [11, 12, 13].

Due to the dependence of α_{T_e} and β_{T_e} on the unknown solution, the algebraic system (56) is nonlinear. It can be solved using the fixed-point iteration

$$u^{(m+1)} = u^{(m)} + \left[\frac{1}{\Delta t} M_L - \theta L\right]^{-1} r^{(m)}, \qquad m = 0, 1, 2, \ldots \tag{58}$$

$$r^{(m)} = \theta(Lu^{(m)} + \bar{f}^{(m)}) + (1-\theta)(Lu^n + f^n) + g - M_L \frac{u^{(m)} - u^n}{\Delta t}. \tag{59}$$

The rates of convergence to steady state solutions can be greatly improved using Anderson acceleration for fixed-point iterations [26, 27, 47].

10. Numerical examples

In this section, we apply the proposed methodology to two-dimensional test problems that have been used to study edge-based algebraic flux correction schemes in [23, 25, 26, 27]. Given the exact solution u, we use the following norms to assess the accuracy of a finite element approximation u_h

$$E_1(h) = \sum_i m_i |u(\mathbf{x}_i) - u_i| \approx \|u - u_h\|_1, \tag{60}$$

$$E_2(h) = \sqrt{\sum_i m_i |u(\mathbf{x}_i) - u_i|^2} \approx \|u - u_h\|_2, \tag{61}$$

where $m_i = \int_\Omega \varphi_i \, \mathrm{d}\mathbf{x}$ is a diagonal coefficient of the lumped mass matrix M_L.

To study the dependence of E_1 and E_2 on the mesh size h, the numerical solutions computed on two different meshes are used to estimate the experimental order of convergence (EOC) using the formula [35]

$$p = \log_2 \left(\frac{E(2h)}{E(h)} \right). \tag{62}$$

In grid convergence studies for time-dependent problems, the ratio of the time step and mesh size is held constant in the process of refinement.

10.1. Solid body rotation

The solid body rotation test [35, 48] is often used to evaluate numerical advection schemes. The problem to be solved is the continuity equation

$$\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) = 0 \qquad \text{in } \Omega = (0,1) \times (0,1). \tag{63}$$

The velocity $\mathbf{v}$ describes a counterclockwise rotation about the center

$$\mathbf{v}(x,y) = (0.5 - y, x - 0.5). \tag{64}$$

After each full revolution, the exact solution u coincides with the given initial data u_0. Hence, the challenge of this test is to preserve the shape of u_0.

Following LeVeque [35], we simulate solid body rotation of a profile that consists of a slotted cylinder, a sharp cone, and a smooth hump (see Fig. 2 (a)). The geometry of each body is described by a given function $G(x,y)$ defined on a circle of radius $r_0 = 0.15$ centered at some point (x_0, y_0). Let

$$r(x,y) = \frac{1}{r_0}\sqrt{(x-x_0)^2 + (y-y_0)^2}$$

be the normalized distance from (x_0, y_0). Then $r(x,y) \leq 1$ inside the circle.

The slotted cylinder is centered at the point $(x_0, y_0) = (0.5, 0.75)$ and

$$G(x,y) = \begin{cases} 1 & \text{if } |x - x_0| \geq 0.025 \ \text{ or } \ y \geq 0.85, \\ 0 & \text{otherwise.} \end{cases}$$

The cone is centered at $(x_0, y_0) = (0.5, 0.25)$, and its shape is given by

$$G(x,y) = 1 - r(x,y).$$

The hump is centered at $(x_0, y_0) = (0.25, 0.5)$, and the shape function is

$$G(x, y) = \frac{1 + \cos(\pi r(x, y))}{4}.$$

In the rest of the domain, the solution to (63) is initialized by zero, and homogeneous Dirichlet boundary conditions are prescribed at the inlets.

The initial profile shown in Fig. 1 coincides with the exact solution after each full rotation. The numerical solutions presented in Figs 2 and 3 correspond to the final time $T = 2\pi$. All computations were performed on a uniform mesh of $2 \times 128 \times 128$ $\mathcal{P}_1$ elements using the Crank-Nicolson time-stepping and the constant time step $\Delta t = 10^{-3}$. The diffusive approximation shown in Fig. 2 (a) was obtained using the low-order LPS operator ($\alpha_T = 0$). The results shown in Fig. 2 (b)-(d) were obtained using the constant correction factor $\alpha_T = 1$ and different values of the target selector ω. The magnitude of spurious undershoots and overshoots can be readily inferred from the range of solution values above each plot. The target $\omega = 0.0$ corresponds to the standard Galerkin discretization which produces global oscillations. The results for $\omega = 0.1$ and $\omega = 1.0$ show that linear high-order LPS stabilization localizes nonphysical oscillations to a neighborhood of steep gradients leading to a marked improvement compared to the standard Galerkin scheme.

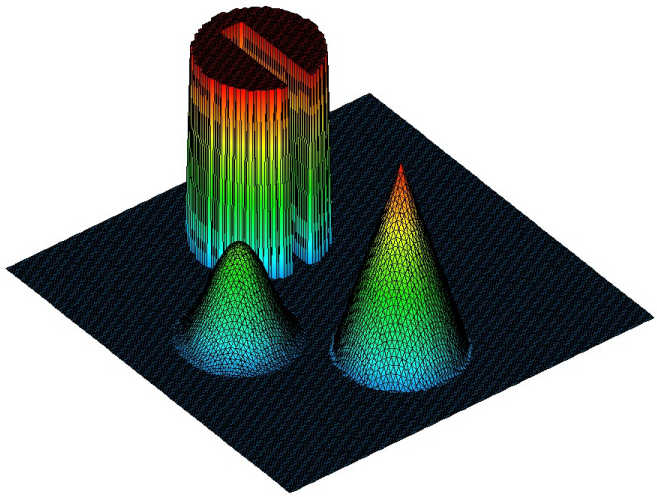


Figure 1: Solid body rotation: initial data/exact solution.

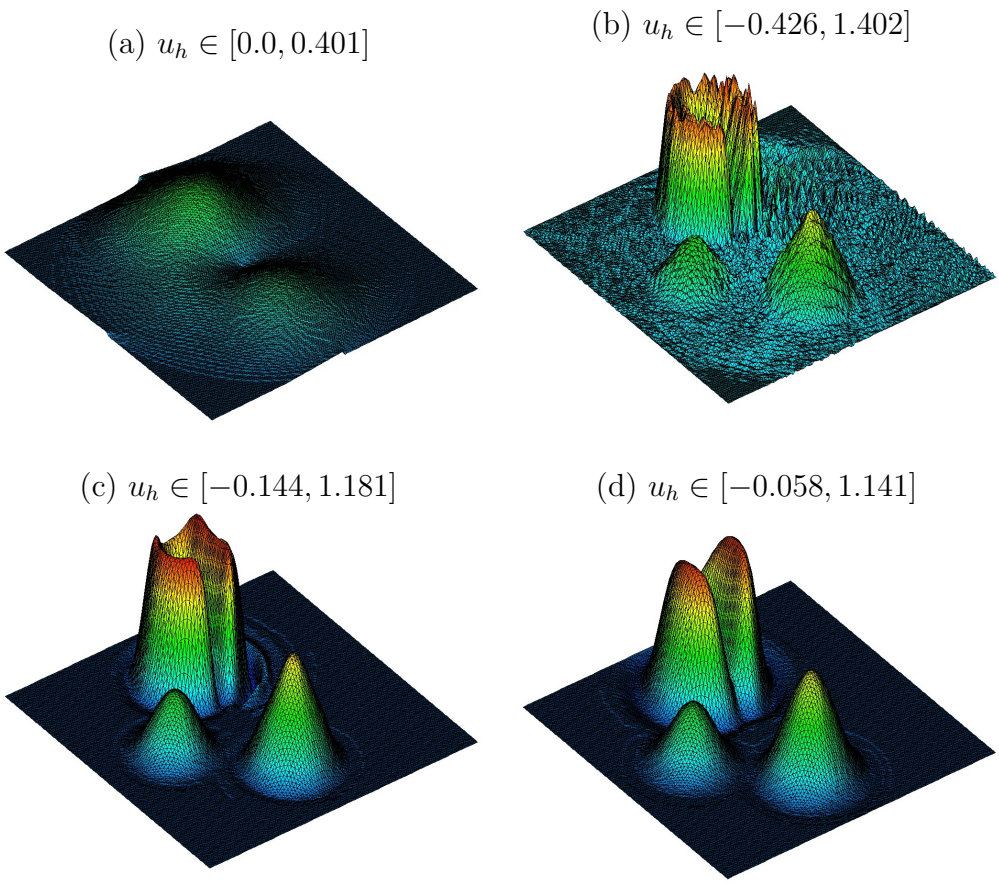


Figure 2: Solid body rotation: linear schemes, (a) $\alpha_T = 0$, (b) $\alpha_T = 1,\ \omega = 0.0$, (c) $\alpha_T = 1,\ \omega = 0.1$, (d) $\alpha_T = 1,\ \omega = 1.0$. Space discretization: $\mathcal{P}_1$ elements, $h = \frac{1}{128}$.

The numerical solutions produced by nonlinear LPS operators are shown in Fig. 3. The label LPS-α stands for using the correction factor $\beta_T = \alpha_T$ to limit the antidiffusive element contributions of the consistent mass matrix. The constrained solution satisfies the discrete maximum principle but the definition of the target $\tilde{u}_h$ in terms of unconstrained time derivatives gives rise to terracing at the top of the slotted cylinder due to element contributions which tend to flatten the solution profiles instead of steepening them. The activation of the optional time derivative limiter (51) in the LPS-β version results in a reduction of phase errors and a more accurate target leading to improved approximations. The need for using the safeguard β to con-

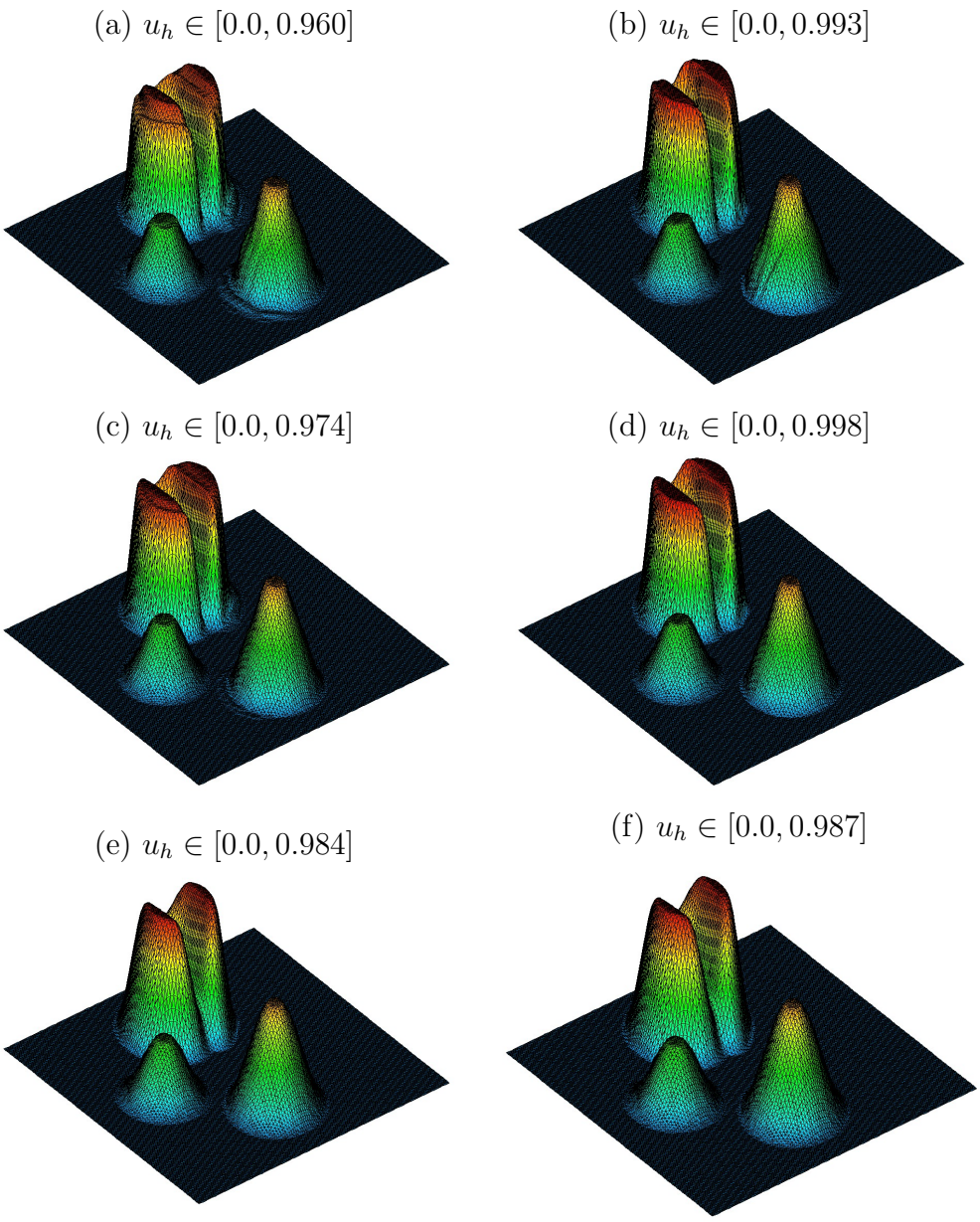


Figure 3: Solid body rotation: nonlinear schemes, (a) LPS-α limiter, $\omega = 0.0$, (b) LPS-β limiter, $\omega = 0.0$, (c) LPS-α limiter, $\omega = 0.1$, (d) LPS-β limiter, $\omega = 0.1$, (e) LPS-α limiter, $\omega = 1.0$, (f) LPS-β limiter, $\omega = 1.0$, Space discretization: $\mathcal{P}_1$ elements, $h = \frac{1}{128}$.

trol the time derivatives becomes less pronounced as the level of high-order background dissipation is increased leading to smaller phase errors.

The convergence history and EOCs for the low-order scheme and its high-order counterparts are listed in Tables 1–3. It can be seen that the use of LPS-β limiting does not degrade the EOC of the underlying high-order scheme ($\omega = 0.1$). The slow rates of grid convergence are caused by the presence of discontinuities in the exact solution to this problem. To show this, a grid convergence study was performed for the initial profile without the cylinder and cone. Tables 4–6 illustrate the convergence behavior of the three LPS methods in the absence of discontinuities and steep fronts.

h	E_1	EOC	E_2	EOC
1/32	0.108e+00		0.249e+00	
1/64	0.111e+00	-0.04	0.230e+00	0.12
1/128	0.104e+00	0.09	0.205e+00	0.17
1/256	0.919e-01	0.18	0.181e+00	0.18

Table 1: Solid body rotation: linear LPS, $\alpha_T = 0$, nonsmooth data.

h	E_1	EOC	E_2	EOC
1/32	0.621e-01		0.141e+00	
1/64	0.360e-01	0.79	0.103e+00	0.45
1/128	0.208e-01	0.79	0.738e-01	0.48
1/256	0.147e-01	0.50	0.615e-01	0.26

Table 2: Solid body rotation: linear LPS, $\alpha_T = 1,\ \omega = 0.1$, nonsmooth data.

h	E_1	EOC	E_2	EOC
1/32	0.666e-01		0.146e+00	
1/64	0.412e-01	0.69	0.122e+00	0.26
1/128	0.194e-01	1.09	0.770e-01	0.66
1/256	0.101e-01	0.94	0.557e-01	0.47

Table 3: Solid body rotation: LPS-β limiter, $\omega = 0.1$, nonsmooth data.

h	E_1	EOC	E_2	EOC
1/32	0.128e-01		0.523e-01	
1/64	0.131e-01	-0.03	0.486e-01	0.11
1/128	0.121e-01	0.11	0.429e-01	0.18
1/256	0.974e-02	0.31	0.350e-01	0.29

Table 4: Solid body rotation: linear LPS, $\alpha_T = 0$, smooth data.

h	E_1	EOC	E_2	EOC
1/32	0.563e-02		0.162e-01	
1/64	0.107e-02	2.40	0.312e-02	2.38
1/128	0.193e-03	2.47	0.647e-03	2.27
1/256	0.622e-04	1.63	0.243e-03	1.41

Table 5: Solid body rotation: linear LPS, $\alpha_T = 1,\ \omega = 0.1$, smooth data.

h	E_1	EOC	E_2	EOC
1/32	0.752e-02		0.311e-01	
1/64	0.133e-02	2.50	0.792e-02	1.97
1/128	0.187e-03	2.83	0.151e-02	2.39
1/256	0.244e-04	2.94	0.273e-03	2.47

Table 6: Solid body rotation: LPS-β limiter, $\omega = 0.1$, smooth data.

10.2. Circular convection

In the second test, we solve the steady advection equation

$$\nabla \cdot (\mathbf{v}u) = 0 \quad \text{in } \Omega = (0,1) \times (0,1). \tag{65}$$

The divergence-free velocity field is defined by

$$\mathbf{v}(x,y) = (y, -x). \tag{66}$$

The exact solution is constant along the circular streamlines. The inflow boundary condition and the exact solution at any point in $\bar{\Omega}$ are given by

$$u(x,y) = \begin{cases} 1, & \text{if } \ 0.15 \le r(x,y) \le 0.45, \\ \cos^2\left(10\pi \frac{r(x,y)-0.5}{3}\right), & \text{if } \ 0.55 \le r(x,y) \le 0.85, \\ 0, & \text{otherwise}, \end{cases} \tag{67}$$

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

Stationary numerical solutions calculated using the standard Galerkin scheme and its linear LPS counterpart with the target $\tilde{u}_h$ defined by $\alpha_T = 1,\ \omega = 0.1$ are shown in Fig. 4. As in the first example, even a small amount of linear high-order LPS stabilization is sufficient to localize nonphysical oscillations leaving just bounded undershoots and overshoots around the discontinuities. Solutions produced by the low-order and constrained high-order LPS schemes are presented in Fig. 5. Both solutions are bounded by 0 and 1 as required by the discrete maximum principle. However, the use of low-order LPS stabilization gives rise to unacceptably high levels of numerical diffusion. The convergence history of LPS schemes (low-order: $\alpha_T = 0$, high-order: $\alpha_T = 1,\ \omega = 0.1$, constrained: $\omega = 0.1$) is presented in Tables 7-9.

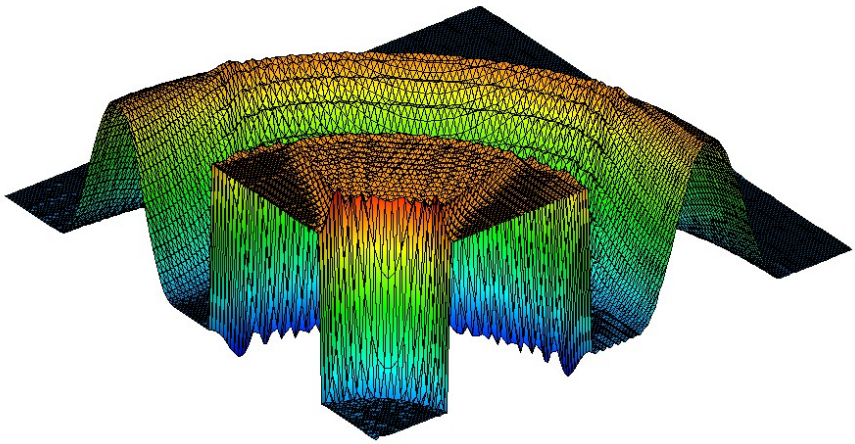
To study the effect of the linearity-preserving correction (38) to the basic nodal limiter (37), a comparison of the two versions of the constrained LPS scheme was performed on nonuniform triangular meshes. Given a uniform grid with spacing h, its distorted counterpart was generated by applying random perturbations to the Cartesian coordinates of internal nodes

$$x_i := x_i + \alpha h \xi_i \qquad y_i := y_i + \alpha h \eta_i, \tag{68}$$

where $\xi_i, \eta_i \in [-0.5, 0.5]$ are random numbers. The parameter $\alpha \in [0,1]$ quantifies the degree of distortion. In this numerical study, we use $\alpha = 0.5$ to generate grid deformations strong enough to violate the patch conditions under which (38) proves linearity-preserving (see Appendix B).

The numerical solutions obtained with the basic limiter (LPS-B) and its linearity-preserving counterpart (LPS-L) on a perturbed mesh of $2\times128\times128$ $\mathcal{P}_1$ elements are displayed in Fig. 6. The errors listed in Tables 10 and 11 illustrate the convergence behavior of the two constrained LPS methods as

(a) $u_h \in [-0.239, 1.227]$



(b) $u_h \in [-0.068, 1.067]$

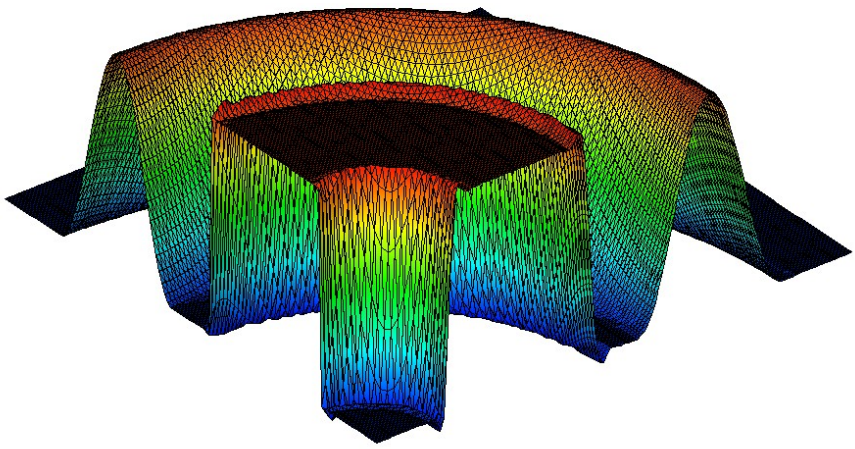


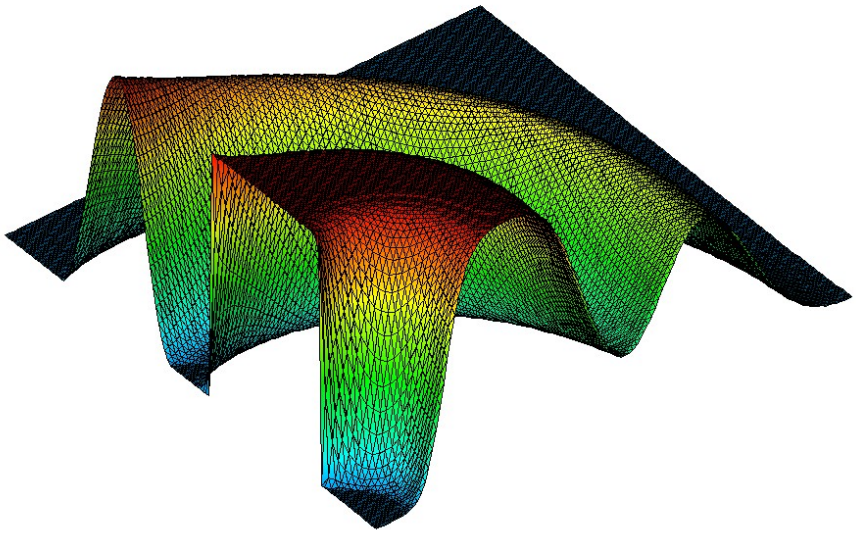
Figure 4: Circular convection: linear high-order schemes, (a) $\alpha_T = 1,\ \omega = 0.0$, (b) $\alpha_T = 1,\ \omega = 0.1$. Discretization: $2 \times 128 \times 128$ $\mathcal{P}_1$ elements.

applied to the circular convection problem with the modified inflow profile

$$u(x,y) = \begin{cases} \cos^2\left(10\pi \frac{r(x,y)-0.5}{3}\right), & \text{if } \ 0.55 \le r(x,y) \le 0.85, \\ 0, & \text{otherwise} \end{cases} \tag{69}$$

constructed by removing the discontinuous component of (67). It can be seen

(a) $u_h \in [0.0, 1.0]$



(b) $u_h \in [0.0, 1.0]$

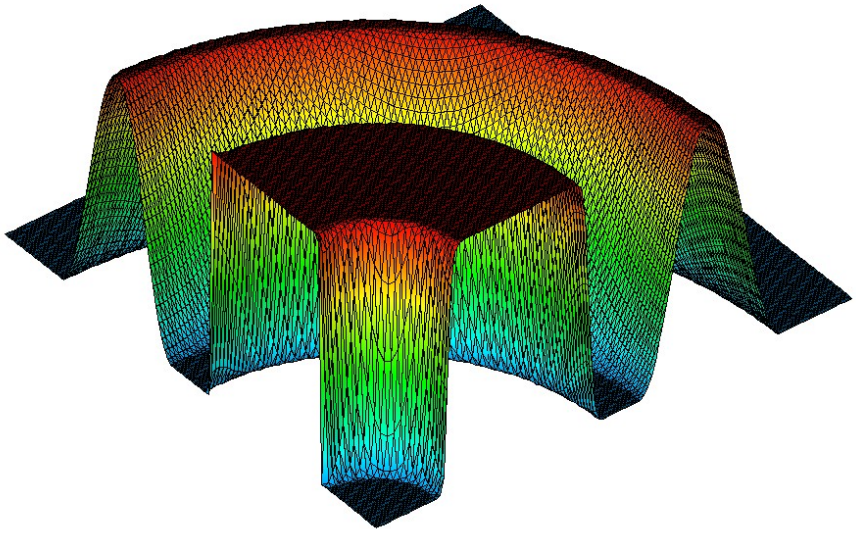
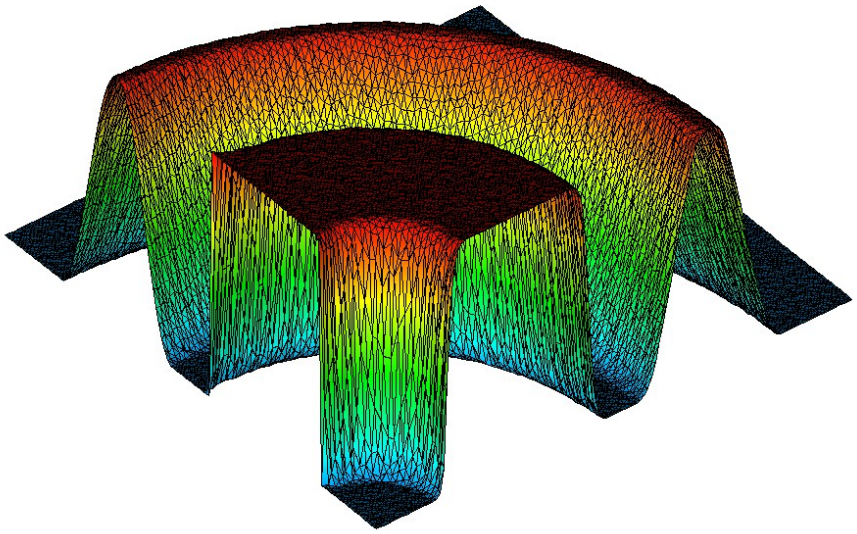


Figure 5: Circular convection: monotone LPS schemes, (a) low-order ($\alpha_T = 0$), (b) limited high-order, $\omega = 0.1$. Discretization: uniform mesh, $2 \times 128 \times 128$ $\mathcal{P}_1$ elements.

that the linearity-preserving version produces smaller errors on nonuniform meshes and delivers optimal convergence rates in this example.

(a) $u_h \in [0.0, 1.0]$



(b) $u_h \in [0.0, 1.0]$

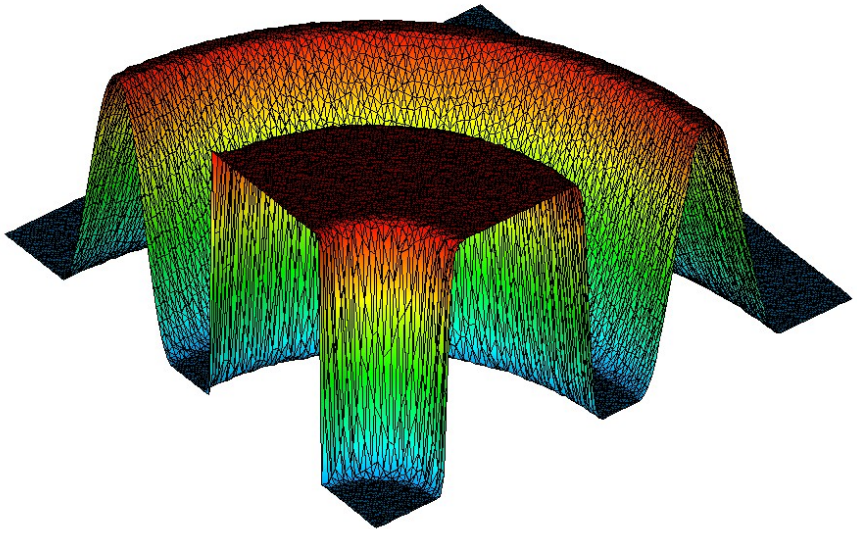


Figure 6: Circular convection: limited high-order schemes, (a) LPS-B, $\omega = 0.1$, (b) LPS-L, $\omega = 0.1$. Discretization: perturbed mesh, $2 \times 128 \times 128$ $\mathcal{P}_1$ elements.

A further gain in accuracy could be achieved by deactivating the limiter at smooth extrema to avoid the peak clipping effect which is clearly seen in

h	E_1	EOC	E_2	EOC
1/32	0.157e+00		0.228e+00	
1/64	0.118e+00	0.41	0.181e+00	0.33
1/128	0.798e-01	0.56	0.133e+00	0.45
1/256	0.507e-01	0.65	0.967e-01	0.46

Table 7: Circular convection: linear LPS, $\alpha_T = 0$, uniform mesh.

h	E_1	EOC	E_2	EOC
1/32	0.218e-01		0.662e-01	
1/64	0.112e-01	0.96	0.545e-01	0.28
1/128	0.635e-02	0.82	0.411e-01	0.41
1/256	0.372e-02	0.77	0.319e-01	0.37

Table 8: Circular convection: linear LPS, $\alpha_T = 1,\ \omega = 0.1$, uniform mesh.

h	E_1	EOC	E_2	EOC
1/32	0.330e-01		0.863e-01	
1/64	0.141e-01	1.23	0.680e-01	0.34
1/128	0.633e-02	1.16	0.444e-01	0.61
1/256	0.324e-02	0.97	0.309e-01	0.52

Table 9: Circular convection: limited LPS, $\omega = 0.1$, uniform mesh.

the presented results. Since the reconstructed gradient provides information about the second derivatives of the approximate solution, the parameter-free smoothness sensor developed in [31] can be used for this purpose.

11. Summary

In this paper, we explored an element-based approach to constraining the consistent mass matrix and the discrete transport operators in continuous Galerkin methods. The proposed LPS-type corrections of Galerkin bilinear forms lead to a new class of variational high-resolution finite element schemes satisfying discrete maximum principles. Since the limiter for the

h	E_1	EOC	E_2	EOC
1/32	0.176e-01		0.399e-01	
1/64	0.542e-02	1.70	0.130e-01	1.62
1/128	0.125e-02	2.12	0.338e-02	1.94
1/256	0.440e-03	1.51	0.112e-02	1.59

Table 10: Circular convection: LPS-B, smooth data, perturbed mesh.

h	E_1	EOC	E_2	EOC
1/32	0.156e-01		0.366e-01	
1/64	0.396e-02	1.98	0.107e-01	1.77
1/128	0.803e-03	2.30	0.275e-02	1.96
1/256	0.174e-03	2.21	0.722e-03	1.93

Table 11: Circular convection: LPS-L, smooth data, perturbed mesh.

antidiffusive element contributions is designed at the semi-discrete level, the limiting procedure is applicable to stationary and time-dependent problems. Further work is under way to provide additional theoretical justification of the new approach building on existing analysis of local projection stabilization/variational multiscale methods and algebraic flux correction schemes.

Acknowledgments

This research of D. Kuzmin and S. Basting was supported by the German Research Association (DFG) under grant KU 1530/12-1. The work of J. N. Shadid was partially supported by the DOE Office of Science Applied Mathematics Program at Sandia National Laboratories under contract DE-AC04-94AL85000.

References

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

[2] G. Barrenechea, E. Burman, F. Karakatsani, Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. Preprint `arXiv:1509.08636v1 [math.NA]` 29 Sep 2015.

[3] G. Barrenechea, V. John, P. Knobloch, Analysis of algebraic flux correction schemes. WIAS Preprint No. **2107** (2015).

[4] 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** (2006) 2544–2566.

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

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

[7] E. Burman and A. Ern, Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. *Math. Comp.* **74** (2005) 1637–1652.

[8] J. Donea and A. Huerta, *Finite Element Methods for Flow Problems*. John Wiley & Sons, Chichester, 2003.

[9] J. Donea, V. Selmin, L. Quartapelle, Recent developments of the Taylor-Galerkin method for the numerical solution of hyperbolic problems. *Numerical Methods for Fluid Dynamics III*, Oxford, 1988, 171–185.

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

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

[12] S. Gottlieb and C. W. Shu, Total Variation Diminishing Runge-Kutta schemes. *Math. Comp.* **67** (1998) 73–85.

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

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

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

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

[17] P.E. Farrell, M.D. Piggott, C.C. Pain, G.J. Gorman, C.R. Wilson, Conservative interpolation between unstructured meshes via supermesh construction. *Comput. Methods Appl. Mech. Eng.* **198** (2009) 2632–2642.

[18] J-l. Guermond, R. Pasquetti, B. Popov, Entropy viscosity method for nonlinear conservation laws. *SIAM J. Sci. Comput.* **230** (2011) 2484267.

[19] W. Hundsdorfer and J.G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer, 2003.

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

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

[22] V. John and P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations: Part I - A review. *Comput. Methods Appl. Mech. Engrg.* 196:17–20 (2007) 2197–2215.

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

[24] V. John, S. Kaya, W. Layton, A two-level variational multiscale method for convection-dominated convectiondiffusion equations. *Comput. Methods Appl. Mech. Engrg.* **195** (2006) 4594–4603.

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

[26] D. Kuzmin, Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. *J. Comput. Appl. Math.* **236** (2012) 2317–2337.

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

[28] D. Kuzmin, A vertex-based hierarchical slope limiter for p-adaptive discontinuous Galerkin methods. *J. Comput. Appl. Math.* **233** (2010) 3077–3085.

[29] D. Kuzmin, Hierarchical slope limiting in explicit and implicit discontinuous Galerkin methods. *J. Comput. Phys.* **257** (2014) 1140–1162.

[30] D. Kuzmin, Algebraic flux correction for finite element discretizations of coupled systems. In: E. Oñate, M. Papadrakakis, B. Schrefler (eds) *Computational Methods for Coupled Problems in Science and Engineering II*, CIMNE, Barcelona, 2007, 653–656.

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

[32] D. Kuzmin and S. Turek, High-resolution FEM-TVD schemes based on a fully multidimensional flux limiter. *J. Comput. Phys.* **198** (2004) 131-158.

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

[34] D. Kuzmin and J.N. Shadid, Gradient-based nodal limiters for finite element schemes. Preprint: *Ergebnisber. Inst. Angew. Math.* **536** Department of Mathematics, TU Dortmund University, 2015. Submitted to *Int. J. Numer Methods Fluids*.

[35] R.J. LeVeque, High-resolution conservative algorithms for advection in incompressible flow. *SIAM J. Numer. Anal.* **33** (1996) 627–665.

[36] R. Löhner, *Applied CFD Techniques: An Introduction Based on Finite Element Methods*. John Wiley & Sons, 2nd edition, 2008.

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

[38] R. Löhner, K. Morgan, M. Vahdati, J.P. Boris, D.L. Book, FEM-FCT: combining unstructured grids with high resolution. *Commun. Appl. Numer. Methods* **4** (1988) 717–729.

[39] H. Luo, J.D. Baum, R. Löhner, J. Cabello, Adaptive edge-based finite element schemes for the Euler and Navier-Stokes equations; AIAA-93-0336, 1993.

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

[41] G. Matthies, P. Skrzypacz, L. Tobiska, A unified convergence analysis for local projection stabilisations applied to the Oseen problem. *M2AN Math. Model. Numer. Anal.* **41** (2007) 713–742.

[42] G. Matthies, P. Skrzypacz, L. Tobiska, Stabilization of local projection type applied to convection-diffusion problems with mixed boundary conditions. *Electron. Trans. Numer. Anal. 32* (2008) 90–105.

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

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

[45] F. Schieweck and P. Skrzypacz, A local projection stabilization method with shock capturing and diagonal mass matrix for solving non-stationary transport dominated problems. *Comput. Methods Appl. Math.* **12** (2012) 221–240.

[46] F. Schieweck and P. Skrzypacz, Local Projection Stabilization method with shock capturing and diagonal mass-matrix for solving non-stationary transport dominated problems. Slides of the presentation at the "Martin-60 workshop", Dresden, 16-18th Dec. 2011. `http://www.math.tu-dresden.de/wnaspp/slides/Th12_Schieweck/main.pdf`

[47] H.W. Walker, P. Ni, Anderson acceleration for fixed-point iterations. *SIAM J. Numer. Anal.* **49** (2011) 1715–1735.

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

Appendix A: Properties of the LPS operator

Let $\hat{T} \subset \mathbb{R}$ be the unit simplex in d dimensions with vertices $\{\hat{\mathbf{x}}_i\}_{i=1,\dots,d+1}$ and $\{\hat{\varphi}_i\}_{i=1,\dots,d+1}$ be the corresponding linear Lagrange basis functions

$$\hat{\varphi}_i(\hat{\mathbf{x}}_j) = \delta_{ij}, \quad i,j = 1,\dots,d+1. \tag{70}$$

Let $\hat{M}$ denote the reference element mass matrix on $\hat{T}$ with entries

$$\hat{m}_{ij} = \int_{\hat{T}} \hat{\varphi}_i \hat{\varphi}_j \, \mathrm{d}\hat{\mathbf{x}} \tag{71}$$

and $\hat{M}_L$ denote the lumped mass matrix on $\hat{T}$ with diagonal entries

$$\hat{m}_{L,ii} = \int_{\hat{T}} \hat{\varphi}_i \, \mathrm{d}\hat{\mathbf{x}}, \quad i = 1,\dots,d+1. \tag{72}$$

A simple calculation reveals that the entries of $\hat{M}$ are

$$\hat{m}_{ij} = \frac{1}{(d+2)!} \cdot \begin{cases} 2 & \text{if} \quad i = j, \\ 1 & \text{otherwise,} \end{cases} \tag{73}$$

and thus the entries of $\hat{M}_L$ are

$$\hat{m}_{L,ij} = \delta_{ij} \frac{1}{(d+2)!} \cdot (2 + d \cdot 1) = \delta_{ij} \frac{1}{(d+1)!}. \tag{74}$$

Let $F_e(\hat{\mathbf{x}})$ be the unique affine-linear mapping that maps $\hat{T}$ onto T_e, i.e.,

$$F_e(\hat{\mathbf{x}}) = A\hat{\mathbf{x}} + \mathbf{b} \tag{75}$$

for some $A \in \mathbb{R}^{d\times d}$, $\mathbf{b} \in \mathbb{R}^d$. Then the entries of the consistent element mass matrix $M_C^{(e)}$ and diagonal lumped mass matrix $M_L^{(e)}$ associated with T_e are given by the transformation rule

$$m_{ij}^{(e)} = \int_T \varphi_i\varphi_j \,\mathrm{d}\mathbf{x} = \int_{\hat{T}} \hat{\varphi}_i\hat{\varphi}_j |\det A| \,\mathrm{d}\hat{\mathbf{x}} = |\det A|\hat{m}_{ij}, \tag{76}$$

$$m_{L,ij}^{(e)} = \delta_{ij}\int_T \varphi_i \,\mathrm{d}\mathbf{x} = \delta_{ij}\int_{\hat{T}} \hat{\varphi}_i |\det A| \,\mathrm{d}\hat{\mathbf{x}} = |\det A|\hat{m}_{L,ij}. \tag{77}$$

By definition (13) of the local stabilization operator $s_T(\cdot,\cdot)$, we have

$$\begin{aligned} s_T(\varphi_i, u_h) &= \int_T \varphi_i \left(u_h - \overline{u}_T\right) \,\mathrm{d}\mathbf{x} = \int_T \varphi_i \left(\sum_j u_j\varphi_j - \frac{\int_T \sum_l u_l\varphi_l \,\mathrm{d}\mathbf{x}}{\int_T 1 \,\mathrm{d}\mathbf{x}}\right) \,\mathrm{d}\mathbf{x} \\ &= \sum_j u_j \int_T \varphi_i\varphi_j \,\mathrm{d}\mathbf{x} - \sum_l u_l \int_T \varphi_i \left(\frac{\int_T \varphi_l \,\mathrm{d}\mathbf{x}}{\int_T 1 \,\mathrm{d}\mathbf{x}}\right) \,\mathrm{d}\mathbf{x} \\ &= \sum_j u_j \left[\int_T \varphi_i\varphi_j \,\mathrm{d}\mathbf{x} - \frac{\int_T \varphi_i \,\mathrm{d}\mathbf{x} \int_T \varphi_j \,\mathrm{d}\mathbf{x}}{\int_T 1 \,\mathrm{d}\mathbf{x}}\right]. \end{aligned}$$

Notice that the integral $\int_T \varphi_i \,\mathrm{d}\mathbf{x}$ corresponds to the i-th diagonal entry $m_{L,ii}^{(e)}$ of the lumped element mass matrix $M_L^{(e)}$. It follows that the entries of the element matrix $S^{(e)}$ induced by the local LPS bilinear form are given by

$$\begin{aligned} s_{ij}^{(e)} &= \int_T \varphi_i\varphi_j \,\mathrm{d}\mathbf{x} - \frac{\int_T \varphi_i \,\mathrm{d}\mathbf{x} \int_T \varphi_j \,\mathrm{d}\mathbf{x}}{\int_T 1 \,\mathrm{d}\mathbf{x}} \\ &= |\det A|\hat{m}_{ij} - |\det A| \frac{\int_{\hat{T}} \hat{\varphi}_i \,\mathrm{d}\hat{\mathbf{x}} \int_T \hat{\varphi}_j \,\mathrm{d}\hat{\mathbf{x}}}{\int_{\hat{T}} 1 \,\mathrm{d}\hat{\mathbf{x}}} \\ &= |\det A| \left[\hat{m}_{ij} - d! \cdot \hat{m}_{L,i} \cdot \hat{m}_{L,j}\right] \\ &= |\det A| \left[\frac{1+\delta_{ij}}{(d+2)!} - \frac{d!}{(d+1)!(d+1)!}\right] \\ &= |\det A| \left[\frac{1}{(d+1)!}\left(\frac{1+\delta_{ij}}{d+2} - \frac{1}{d+1}\right)\right] \end{aligned}$$

$$
\begin{aligned}
&= |\det A| \left[\frac{1}{(d+2)!} \left(\frac{(1+\delta_{ij})(d+1) - (d+2)}{d+1} \right) \right] \\
&= |\det A| \left[\frac{1}{(d+2)!} \left(\frac{\delta_{ij}(d+1) - 1}{d+1} \right) \right] \\
&= |\det A| \left[\frac{1}{(d+2)!} \left(\frac{\delta_{ij}(d+2) - 1 - \delta_{ij}}{d+1} \right) \right] \\
&= \frac{1}{d+1} |\det A| \left[\hat{m}_{L,ij} - \hat{m}_{ij} \right] = \frac{1}{d+1} \left[m_{L,ij}^{(e)} - m_{ij}^{(e)} \right].
\end{aligned}
$$

The relation between $M_L^{(e)} - M_C^{(e)}$ and $S^{(e)}$ is thus given by

$$
S^{(e)} = \frac{1}{d+1} \left(M_L^{(e)} - M_C^{(e)} \right). \tag{78}
$$

It follows that

$$
s_T(\varphi_i, u_h) = \sum_j s_{ij}^{(e)} u_j = \frac{1}{d+1} \sum_j m_{ij}^{(e)} (u_i - u_j),
$$

which proves that $s_T(\varphi_i, u_h) \geq 0$ if u_i is a local maximum ($u_i \geq u_j \ \forall j \neq i$) and $s_T(\varphi_i, u_h) < 0$ if u_i is a local minimum ($u_i \leq u_j \ \forall j \neq i$). This property makes the nodal limiter function (37) local extremum diminishing.

The global stabilization operator $s(\cdot, \cdot)$ is defined by (11). We have

$$
\begin{aligned}
s(u_h, u_h) &= \sum_T \nu_T \int_T u_h \left(u_h - \overline{u}_T \right) \mathrm{d}\mathbf{x} \\
&= \sum_T \nu_T \int_T (u_h - \overline{u}_T)^2 \, \mathrm{d}\mathbf{x} \\
&= \sum_T \nu_T \| u_h - \overline{u}_T \|_{0,T}^2.
\end{aligned}
$$

The following estimate holds:

$$
\begin{aligned}
\| u_h - u_T \|_{0,T}^2 &= \int_T |\nabla u_h \cdot (\mathbf{x} - \bar{\mathbf{x}})|^2 \, \mathrm{d}\mathbf{x} \\
&\leq \int_T (h_T |\nabla u_h|)^2 \, \mathrm{d}\mathbf{x} = h_T^2 |u_h|_{1,T}^2.
\end{aligned}
$$

This result makes it possible to prove $\mathcal{O}(h^{1/2})$ consistency of the low-order scheme following the analysis of algebraic flux correction schemes in [2].

Appendix B: Properties of the nodal limiter

To derive sufficient conditions of linearity preservation and an error estimate for smooth data, we perform the following transformations:

$$\begin{aligned}\sum_T \int_T \varphi_i(u_h - \bar{u}_T)\,\mathrm{d}\mathbf{x} &= \sum_T \int_T (\varphi_i - \bar{\varphi}_i)(u_h - \bar{u}_T)\,\mathrm{d}\mathbf{x} \\ &= \sum_T \int_T (\varphi_i - \bar{\varphi}_i)\nabla u_h|_T \cdot (\mathbf{x} - \mathbf{x}_T)\,\mathrm{d}\mathbf{x} \\ &= \sum_T \int_T (\varphi_i - \bar{\varphi}_i)\nabla u_h|_T \cdot \mathbf{x}\,\mathrm{d}\mathbf{x} \\ &= \sum_T \int_T \varphi_i \nabla u_h|_T \cdot \mathbf{x}\,\mathrm{d}\mathbf{x} - \sum_T \int_T \bar{\varphi}_i \nabla u_h|_T \cdot \mathbf{x}\,\mathrm{d}\mathbf{x}.\end{aligned}$$

Using the midpoint rule on a triangle T, we find that

$$\sum_T \int_T \bar{\varphi}_i \nabla u_h|_T \cdot \mathbf{x}\,\mathrm{d}\mathbf{x} = \sum_T \frac{|T|}{3} \nabla u_h|_T \cdot \mathbf{x}_T.$$

Let $\{\mathbf{x}_P := \mathbf{x}_i,\ \mathbf{x}_{T,1},\ \mathbf{x}_{T,2}\}$ denote the vertices of T. The quadrature rule

$$\int_T f(\mathbf{x})\,\mathrm{d}\mathbf{x} \approx \frac{|T|}{3}\left[f\left(\frac{\mathbf{x}_P + \mathbf{x}_{T,1}}{2}\right) + f\left(\frac{\mathbf{x}_P + \mathbf{x}_{T,2}}{2}\right) + f\left(\frac{\mathbf{x}_{T,1} + \mathbf{x}_{T,2}}{2}\right)\right]$$

is exact for quadratic polynomials on T. Thus we have

$$\int_T \varphi_i \nabla u_h|_T \cdot \mathbf{x}\,\mathrm{d}\mathbf{x} = \sum_T \frac{|T|}{3}\frac{1}{2}\nabla u_h|_T \cdot \left[\frac{1}{2}\left(2\mathbf{x}_P + \mathbf{x}_{T,1} + \mathbf{x}_{T,2}\right)\right],$$

where we have used the fact that $\varphi_i\left(\frac{\mathbf{x}_{T,1}+\mathbf{x}_{T,2}}{2}\right) = 0$. It follows that

$$\sum_T \int_T \varphi_i(u_h - \bar{u}_T)\,\mathrm{d}\mathbf{x} = \sum_T \frac{|T|}{3}\nabla u_h|_T \cdot \left[\frac{1}{6}\mathbf{x}_P - \frac{1}{12}\mathbf{x}_{T,1} - \frac{1}{12}\mathbf{x}_{T,2}\right]. \quad (79)$$

Suppose that the vertex coordinates satisfy the zero sum condition

$$\sum_T \frac{|T|}{3}\mathbf{b} \cdot \left[\frac{1}{6}\mathbf{x}_P - \frac{1}{12}\mathbf{x}_{T,1} - \frac{1}{12}\mathbf{x}_{T,2}\right] = 0 \quad (80)$$

for an arbitrary constant vector $\mathbf{b} \in \mathbb{R}^2$. We also assume that there exists a constant vector $\mathbf{G}_P \in \mathbb{R}^2$ such that for each element T of the patch Ω_i the local gradient of $u_h|_T$ admits the decomposition

$$\nabla u_h|_T = \mathbf{G}_P + h\mathbf{G}(T)$$

with a vector $\mathbf{G}(T) \in \mathbb{R}^2$ such that $\|\mathbf{G}(T)\| \leq C$ for a constant $C \geq 0$.

Substituting this representation of $\nabla u_h|_T$ into (79) and using the zero sum property (80) to eliminate the contribution of $\mathbf{G}_P$, we obtain

$$\sum_T \int_T \varphi_i(u_h - \bar{u}_T)\,\mathrm{d}\mathbf{x} = \sum_T \frac{|T|}{3} h\mathbf{G}(T) \cdot \left[\frac{1}{6}\mathbf{x}_P - \frac{1}{12}\mathbf{x}_{T,1} - \frac{1}{12}\mathbf{x}_{T,2}\right] = \mathcal{O}(h^4),$$

whereas

$$\sum_T \left| \int_T \varphi_i(u_h - \bar{u}_T)\,\mathrm{d}\mathbf{x} \right| = \mathcal{O}(h^3).$$

Consequently, the ratio of these two sums behaves as $\mathcal{O}(h)$ and therefore

$$\Phi_i = 1 - \frac{\left|\sum_T \int_T \varphi_i(u_h - \bar{u}_T)\,\mathrm{d}\mathbf{x}\right|}{\sum_T \left|\int_T \varphi_i(u_h - \bar{u}_T)\,\mathrm{d}\mathbf{x}\right|} = 1 - \mathcal{O}(h).$$