MONOTONICITY-PRESERVING FINITE ELEMENT SCHEMES BASED ON DIFFERENTIABLE NONLINEAR STABILIZATION

SANTIAGO BADIA† AND JESÚS BONILLA‡

Abstract. In this work, we propose a nonlinear stabilization technique for scalar conservation laws with implicit time stepping. The method relies on an artificial diffusion method, based on a graph-Laplacian operator. It is nonlinear, since it depends on a shock detector. Further, the resulting method is linearity preserving. The same shock detector is used to gradually lump the mass matrix. The resulting method is LED, positivity preserving, and also satisfies a global DMP. Lipschitz continuity has also been proved. However, the resulting scheme is highly nonlinear, leading to very poor nonlinear convergence rates. We propose a smooth version of the scheme, which leads to twice differentiable nonlinear stabilization schemes. It allows one to straightforwardly use Newton's method and obtain quadratic convergence. In the numerical experiments, steady and transient linear transport, and transient Burgers' equation have been considered in 2D. Using the Newton method with a smooth version of the scheme we can reduce 10 to 20 times the number of iterations of Anderson acceleration with the original non-smooth scheme. In any case, these properties are only true for the converged solution, but not for iterates. In this sense, we have also proposed the concept of projected nonlinear solvers, where a projection step is performed at the end of every nonlinear iterations onto a FE space of admissible solutions. The space of admissible solutions is the one that satisfies the desired monotonic properties (maximum principle or positivity).

Keywords: Finite elements, discrete maximum principle, monotonicity, nonlinear solvers, shock capturing

Contents

1. Introduction 2
2. Preliminaries 3
2.1. The continuous problem 3
2.2. Finite element spaces and meshes 4
2.3. The semi-discrete problem 4
3. Nonlinear stabilization 5
4. Monotonicity properties 7
5. Symmetric mass matrix stabilization 9
6. Lipschitz continuity 9
7. Differentiable stabilization 10
8. Nonlinear Solvers 12
9. Numerical Experiments 13
9.1. Steady problems 13

Date: July 6, 2021.

† Universitat Politècnica de Catalunya, Jordi Girona1-3, Edifici C1, E-08034 Barcelona & Centre Internacional de Mètodes Numèrics en Enginyeria, Parc Mediterrani de la Tecnologia, Esteve Terrades 5, E-08860 Castelldefels, Spain E-mail: sbadia@cimne.upc.edu. SB was partially supported by the European Research Council under the FP7 Program Ideas through the Starting Grant No. 258443 - COMFUS: Computational Methods for Fusion Technology and the FP7 NUMEXAS project under grant agreement 611636. SB gratefully acknowledges the support received from the Catalan Government through the ICREA Acadèmia Research Program.

‡ Universitat Politècnica de Catalunya, Jordi Girona1-3, Edifici C1, E-08034 Barcelona & Centre Internacional de Mètodes Numèrics en Enginyeria, Parc Mediterrani de la Tecnologia, Esteve Terrades 5, E-08860 Castelldefels, Spain E-mail: jbonilla@cimne.upc.edu. JB gratefully acknowledges the support received from "la Caixa" Foundation through its PhD scholarship program.

9.2. Transient transport problems 16
9.3. Burgers' equation 18
10. Conclusions 19
Appendix A. Proof of Theorem 6.1 21
References 25

1. Introduction

Many partial differential equations (PDEs) satisfy some sort of maximum principle or positivity property. However, numerical discretizations usually violate these structural properties at the discrete level, with implications in terms of accuracy and stability, e.g., leading to non-physical local oscillations.

It is well-understood now how to build methods that satisfy some sort of discrete maximum principle (DMP) based on explicit time integration combined with finite volume or discontinuous Galerkin schemes [9, 27]. However, implicit time integration is preferred in problems with multiple scales in time when the fastest scales are not relevant. E.g., under-resolved simulations of multi-scale problems in time are essential in plasma physics [20]. Unfortunately, implicit DMP-preserving hyperbolic solvers are scarce and not so well developed.

In the frame of finite element (FE) discretizations, the local instabilities present in the solution of hyperbolic problems have motivated the use of so-called shock capturing schemes based on artificial diffusion (see, e.g., [18]). These methods introduce nonlinear stabilization, in contrast with classical SUPG-type linear stabilization techniques [16, 17]. Since linear schemes are at most first-order accurate and highly dissipative [11], recent research on FE techniques for conservation laws has focused on the development of less dissipative nonlinear schemes. Many of these ideas come from the numerical approximation of convection dominated convection-diffusion-reaction (CDR), where one encounters similar issues. The cornerstone of these methods is the design of a nonlinear artificial diffusion that vanishes in smooth regions and works on discontinuities or sharp layers. Many residual-based diffusion methods have been considered so far (see, e.g., [10] and references therein). Most of these approaches have failed to reach DMP-preserving methods. A salient exception is the method by Burman and Ern [7], which satisfies a DMP under mesh restrictions. Recently, due to some interesting novel approaches in the field, the state-of-the-art in nonlinear stabilization has certainly advanced [1–3, 6, 8, 23, 24].

Implicit FE schemes for hyperbolic problems rely on four key ingredients:

(1) The first ingredient is the definition of the *shock detector* that only activates the nonlinear diffusion around shocks/discontinuities. Recent nonlinear stabilization techniques have been developed based on shock detectors driven by gradient jumps [1, 5] or edge differences [3, 23, 24]. The use of such schemes was proposed in [5] for 1D problems and extended to multiple dimensions in [1]. A salient property of the scheme in [1] is that it is DMP-preserving, but it relies on the DMP of the Poisson operator, which is only true under stringent constraints on the mesh. Another salient feature of the gradient-jump diffusion approach in [1] is the fact that it leads to so-called linearity preserving methods, i.e., the artificial diffusion vanishes for first order polynomials. This property is related to high-order convergence on smooth regions [25]. A modification of the nonlinear diffusion in [23] that also satisfies this property is proposed in [24].
(2) The second ingredient is the *amount of diffusion* to be introduced on shocks, which is the amount of diffusion introduced in a first order linear scheme. In this sense, one can consider flux-corrected transport techniques [26].
(3) The third ingredient is the form of the *discrete viscous operator*. In order to keep the DMP on arbitrary meshes, Guermond and Nazarov have proposed to use graph-theoretic, instead of PDE-based, operators for the artificial diffusion terms. This approach has been used in [3, 29] (for the steady-state convection-diffusion-reaction problem) and in [12] (for linear conservation laws) combined with artificial diffusion definitions similar to the one in [13].

(4) The fourth ingredient is the *perturbation of the mass matrix*, in order to satisfy a DMP. Full mass lumping is one choice, but it introduces an unacceptable phase error. For continuous FE methods, improved techniques can be found in [14]. Alternatively, limiting-type strategies are used, e.g., in [23, 24].
(5) The method in [3] is Lipschitz continuous, which is needed for the well-posedness of the resulting nonlinear scheme. However, in practice, all the methods presented above are still highly nonlinear, and nonlinear convergence becomes very hard and expensive. It leads to a fifth additional ingredient that has not been considered so far in much detail. In order to reduce the computational cost of these schemes, we consider the *smoothing* of the nonlinear artificial diffusion, to make it differentiable up to some fixed order. The possibility to define smooth nonlinear schemes can improve the nonlinear convergence of the methods and make them practical for realistic applications. Further, the smoothing step enables advanced linearization strategies based on Newton's method. It also involves the development of efficient nonlinear solvers, e.g., based on the combination of Newton, line search, and/or Anderson acceleration techniques.

All the results commented above are restricted to linear (or bilinear) FEs. We are not aware of the existence of high-order implicit DMP-preserving FE schemes. For explicit time integration and limiters, second order methods can be found in [12]. The use of hp-adaptive schemes that keep first order schemes around shocks has been proposed in [15].

In this work, we propose a novel nonlinear stabilization method that satisfies a DMP, positivity, and local extremum diminishing (LED) properties at the discrete level. It combines: (1) a novel shock detector related to the one in [1], which is simple and linearity preserving; (2) the graph-Laplacian artificial viscous term proposed in [13]; (3) an edge FCT-type definition of the amount of diffusion (see [23]); (4) a novel gradual mass lumping technique that exploits the same shock detector used for the artificial diffusion. We prove that the resulting method ticks all the boxes, i.e., it is total variation diminishing (TVD), DMP, positivity-preserving, linearity preserving, Lipschitz continuous, and introduces low dissipation. With regard to the last point, we prove that the amount of diffusion is the minimum needed in our analysis to prove the DMP. Further, we consider a novel approach to design a smoothed version of the resulting scheme that is twice differentiable. We prove that linear preservation is weakly enforced in this case, but all the other properties remain unchanged. Finally, we analyze the effect of the smoothing in the computational cost, and observe a clear reduction in the CPU cost of the nonlinear solver when using the smooth version of the method proposed herein while keeping almost unchanged the sharp layers of the non-smooth version. Future work will be focused on the entropy stability analysis of these schemes for nonlinear scalar conservation laws. A partial result in this direction is the proof of entropy stability for a related method when applied to the 1D Burger's equations (see [5]).

This work is structured as follows. In Sect. 2 the continuous problem and its discretization using the FE method are presented. Sect. 3 contains the formulation of a novel nonlinear stabilization method. Sect. 4 is devoted to the monotonicity analysis of the proposed method. An alternative approach is presented in Sect. 5. Lipschitz continuity of the methods is proved in Sect. 6. A differentiable version the previous method is presented in Sect. 7. Sect. 8 is devoted to nonlinear solvers. Different numerical experiments are introduced in Sect. 9. Finally, in Sect. 10 we draw some conclusions.

2. Preliminaries

2.1. **The continuous problem.** Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, where d is the space dimension, and $(0,T]$ the time interval. The scalar conservation equation reads: find $u(\boldsymbol{x},t)$ such that

$$\partial_t u + \boldsymbol{\nabla}\cdot\boldsymbol{f}(u) = g, \quad \text{on } \Omega\times(0,T], \tag{1}$$

where $\boldsymbol{f}\in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$ is the flux. It is also subject to the initial condition $u(\boldsymbol{x},0)=u_0\in L^\infty(\Omega)$ and boundary condition $u(\boldsymbol{x},t)=u_D(\boldsymbol{x},t)$ on the inflow $\Gamma_{\text{in}} \doteq \{(\boldsymbol{x},t)\in\partial\Omega\times(0,T] \mid \boldsymbol{f}(\boldsymbol{x},t)\cdot\boldsymbol{n}(\boldsymbol{x},t)<0\}$. There exist a unique entropy solution u of the above problem that satisfies the entropy inequalities $\partial_t E(u)+\boldsymbol{\nabla}\cdot\boldsymbol{F}(u)\le 0$ for all convex entropies $E\in\mathrm{Lip}(\mathbb{R};\mathbb{R})$ with its associated entropy fluxes

$\boldsymbol{F}_i(u) = \int_0^u E'(v)\boldsymbol{f}_i'(v)\,\mathrm{d}v$, $1 \le i \le d$ (see Kružkov [21]). Let us consider the weak form of this problem consists in seeking u such that $u = u_D$ on $\Gamma_{\text{in}} \times (0,T]$ and

$$(\partial_t u, v) + (\boldsymbol{f}'(u)\cdot\boldsymbol{\nabla} u, v) = (g, v) \quad \forall v \in L^2(\Omega), \tag{2}$$

almost everywhere in $(0,T]$, with $g \in L^2(\Omega)$.

2.2. **Finite element spaces and meshes.** Let $\mathcal{T}_h$ be a conforming partition of Ω into elements, K. Elements can be triangles or quadrilaterals for $d=2$, or tetrahedrals or hexahedra for $d=3$. The set of interpolation nodes of $\mathcal{T}_h$ is represented by $\mathcal{N}_h$, whereas $\mathcal{N}_h(K)$ denotes the set of nodes belonging to element $K \in \mathcal{T}_h$. Moreover, Ω_i is the macroelement composed by the union of the elements $K \in \mathcal{T}_h$ such that $i \in \mathcal{N}_h(K)$. $\mathcal{N}_h(\Omega_i)$ denotes the set of nodes in that macroelement. The continuous linear FE space is defined as

$$V_h \doteq \left\{ v_h \in \mathcal{C}^0(\Omega) :\ v_h|_K \in P_k(K)\ \ \forall K \in \mathcal{T}_h \right\} \tag{3}$$

for triangular or tetrahedral elements (replacing $P_1(K)$ by $Q_1(K)$ for quadrilateral or hexahedral elements). $P_1(K)$ (resp., $Q_1(K)$) is the space of polynomials with total (resp., partial) degree less or equal to 1. The nodal basis of V_h is written $\{\varphi_i\}_{i\in\mathcal{N}_h}$, and the FE functions can be expressed as $v_h = \sum_{i\in\mathcal{N}_h}\varphi_i v_i$, where v_i is the value of v_h at node i.

2.3. **The semi-discrete problem.** The semi-discrete Galerkin FE approximation of (2) reads: find $u_h \in V_h$ such that $u_h(\Gamma_{\text{in}}, t) = \pi_h(u_D)$ and

$$(\partial_t u_h, v_h) + (\boldsymbol{f}'(u_h)\cdot\boldsymbol{\nabla} u_h, v_h) = (g, v_h) \quad \forall v_h \in V_h, \tag{4}$$

for $t \in (0,T]$, with initial conditions $u_h(\cdot,0) = \pi_h(u_0)$. π_h denotes a FE interpolation, e.g., the Scott-Zhang projector [28].

Using the notation $\mathsf{M}u_h \doteq (u_h, \cdot)$ and $\mathsf{F}(w_h)u_h \doteq (\boldsymbol{f}'(w_h)\cdot\boldsymbol{\nabla} u_h, \cdot)$ we can write problem (4) in compact form as

$$\mathsf{M}\partial_t u_h + \mathsf{F}(u_h)u_h = g \tag{5}$$

in V_h', i.e., the dual space of V_h. Further, we define $\mathsf{M}_{ij} \doteq (\varphi_j, \varphi_i)$, $\mathsf{F}_{ij}(u_h) \doteq (\boldsymbol{f}'(u_h)\cdot\boldsymbol{\nabla}\varphi_j, \varphi_i)$, and $g_i \doteq (g, \varphi_i)$.

In order to carry out the time discretization of (5), let us consider a partition of the time domain $(0,T]$ into sub-intervals $(t^n, t^{n+1}]$, with $0 \doteq t^0 < t^1 < \ldots < t^N \doteq T$. We consider the Backward-Euler (BE) implicit time integrator to keep at the time-discrete level the monotonicity properties of the semi-discrete problem, leading to the discrete problem: given $u_h^0 \doteq \pi_h(u_0) \in V_h$, compute for $n = 1, \ldots, N-1$

$$\mathsf{M}\delta_t u_h^{n+1} + \mathsf{F}(u_h^{n+1})u_h^{n+1} = g \quad \text{ in } V_h', \tag{6}$$

where $\delta_t u_h^{n+1} \doteq \Delta t_{n+1}^{-1}(u_h^{n+1} - u_h^n)$, and $\Delta t_{n+1} \doteq |t^{n+1} - t^n|$. Implicit strong stability preserving Runge-Kutta methods [19] also preserve the monotonic properties at the discrete level [19], under some restrictions on the time step size. For the sake of brevity we consider the BE scheme.

Systems (5) and (6) will be supplemented with additional stabilization terms to minimize the oscillations generated by the Galerkin FE approximation. Of particular interest are methods which provide solutions that satisfy the following property for all nodes, for zero forcing terms.

Definition 2.1 (Local DMP)**.** *A solution $u \in V_h$ satisfies the local DMP if*

$$u_i^{\min} \le u_i \le u_i^{\max}, \quad \textit{where}\ \ u_i^{\max} \doteq \max_{j\in\mathcal{N}_h(\Omega_i)\setminus\{i\}} u_j, \quad u_i^{\min} \doteq \min_{j\in\mathcal{N}_h(\Omega_i)\setminus\{i\}} u_j. \tag{7}$$

Actually, for steady problems, if this is satisfied for all $i \in \mathcal{N}_h$, then the extrema will be at the boundary and there exist no local extrema.

Furthermore, it is useful to define *local extremum diminishing* (LED) methods for transient problems.

Definition 2.2 (LED)**.** *A method is called LED if for $g = 0$ and any time in $(0,T]$, the solution satisfies*

$$\mathrm{d}_t u_i \le 0 \textit{ if } u_i \textit{ is a maximum and } \ \mathrm{d}_t u_i \ge 0 \textit{ if } u_i \textit{ is a minimum.} \tag{8}$$

For time-discrete methods, the same definition applies, replacing d_t by δ_t.

3. Nonlinear stabilization

We want to design a linearity preserving LED method for stabilizing the scalar semi-discrete hyperbolic problem (5) (or the discrete problem (6)), described in the previous section. As written above, this method is based on a graph-theoretic approach. Let us consider a nonlinear stabilization operator $\mathsf{B}(u_h) : V_h \to V_h'$ and denote $\mathsf{B}_{ij}(u_h) \doteq \langle \mathsf{B}(u_h)\varphi_j, \varphi_i\rangle$. Particularly, we require that the stabilization term will satisfy the following properties (see also [13]):

(1) compact support: $\mathsf{B}_{ij}(u_h) = 0$ if $j \notin \mathcal{N}_h(\Omega_i)$ for any $u_h \in V_h$,
(2) symmetry: $\mathsf{B}_{ij}(u_h) = \mathsf{B}_{ji}(u_h)$ for any $u_h \in V_h$,
(3) conservation: $\sum_{j\neq i} \mathsf{B}_{ij}(u_h) = -\mathsf{B}_{ii}(u_h)$ for any $u_h \in V_h$,
(4) linear preservation: $\mathsf{B}(u_h) = 0$ for any $u_h \in P_1(\Omega)$.

To achieve this properties we define the nonlinear stabilization term

$$\langle \mathsf{B}(w_h)u_h, v_h\rangle \doteq \sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} \nu_{ij}(w_h)v_i u_j \ell(i,j), \qquad u_h,\ v_h \in V_h, \tag{9}$$

where the graph-theoretic Laplacian is defined as $\ell(i,j) \doteq 2\delta_{ij} - 1$, and the artificial diffusion computed as

$$\begin{aligned} \nu_{ij}(w_h) &\doteq \max\{\alpha_i(w_h)\mathsf{F}_{ij}(w_h), \alpha_j(w_h)\mathsf{F}_{ji}(w_h), 0\} \qquad \text{for } i \neq j, \\ \nu_{ii}(w_h) &\doteq \sum_{\substack{j\in\mathcal{N}_h(\Omega_i)\\ j\neq i}} \nu_{ij}(w_h), \end{aligned} \tag{10}$$

where $\alpha_i(\cdot)$ is the shock detector. We note that this choice leads to a symmetric stabilization operator $\mathsf{B}(w_h)$. In order to define the shock detector, let us introduce some notation. Let $i \in \mathcal{N}_h$ be a node of the mesh, $\boldsymbol{v}$ a vector field, and w a scalar field. Let $\boldsymbol{r}_{ij} = \boldsymbol{x}_j - \boldsymbol{x}_i$ be the vector pointing from nodes i to j in $\mathcal{N}_h$ and $\hat{\boldsymbol{r}}_{ij} \doteq \frac{\boldsymbol{r}_{ij}}{|\boldsymbol{r}_{ij}|}$. Let $\boldsymbol{x}_{ij}^{\mathrm{sym}}$ be the point at the intersection between the line that passes through $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and $\partial\Omega_i$ that is not $\boldsymbol{x}_j$ (see Fig. 1). The set of all symmetric nodes with respect to node i is represented with $\mathcal{N}_h^{\mathrm{sym}}(\Omega_i)$. We define $\boldsymbol{r}_{ij}^{\mathrm{sym}} \doteq \boldsymbol{x}_{ij}^{\mathrm{sym}} - \boldsymbol{x}_i$, and $u_j^{\mathrm{sym}} \doteq u_h(\boldsymbol{x}_{ij}^{\mathrm{sym}})$. Then, one can define the jump and the mean of the unknown gradient at node i in direction $\boldsymbol{r}_{ij}$ as

$$[\![\boldsymbol{\nabla} u_h]\!]_{ij} \doteq \frac{u_j - u_i}{|\boldsymbol{r}_{ij}|} + \frac{u_j^{\mathrm{sym}} - u_i}{|\boldsymbol{r}_{ij}^{\mathrm{sym}}|}, \tag{11}$$

$$\{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}|\}\!\}_{ij} \doteq \frac{1}{2}\left(\frac{|u_j - u_i|}{|\boldsymbol{r}_{ij}|} + \frac{|u_j^{\mathrm{sym}} - u_i|}{|\boldsymbol{r}_{ij}^{\mathrm{sym}}|}\right). \tag{12}$$

We note that the symmetric nodes and their corresponding values u_j^{sym} are used in the proof of the following results, Lemma 3.2, and Theorem 6.1, but *not required in the implementation* of (19). For triangular or tetrahedral meshes, since $\boldsymbol{\nabla} u_h$ is constant, u_j^{sym} can be computed easily as

$$u_j^{\mathrm{sym}} = u_h(\boldsymbol{x}_i) + \boldsymbol{\nabla} u_h(\boldsymbol{x}_i) \cdot \boldsymbol{r}_{ij}^{\mathrm{sym}}.$$

For quadrilateral or hexahedral structured (possibly adapted and nonconforming) meshes, u_j^{sym} is also easy to obtain since j^{sym} is already in $\mathcal{N}_h(\Omega_i)$. It also applies for symmetric meshes, when a mesh is said to be symmetric with respect to its internal nodes if for any $i \in \mathcal{N}_h$ all symmetric nodes $j^{\mathrm{sym}} \in \mathcal{N}_h^{\mathrm{sym}}(\Omega_i)$ already belong to $\mathcal{N}_h(\Omega_i)$.

Making use of these definitions, the proposed shock detector at node $i \in \mathcal{N}_h$ for a FE solution u_h reads:

$$\alpha_i(u_h) \doteq \begin{cases} \left[\dfrac{\left|\sum_{j\in\mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij}\right|}{\sum_{j\in\mathcal{N}_h(\Omega_i)} 2\{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}|\}\!\}_{ij}}\right]^q & \text{if } \sum_{j\in\mathcal{N}_h(\Omega_i)} \{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}|\}\!\}_{ij} \neq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

for some $q \in \mathbb{R}^+$. We note that this shock detector is motivated from [1], where the directional nodal-wise jumps and mean values are first used for such purposes. For triangular or tetrahedral meshes, the only difference strives in the fact that the supremum over all $j \in \mathcal{N}_h(\Omega_i)$ in both the numerator and

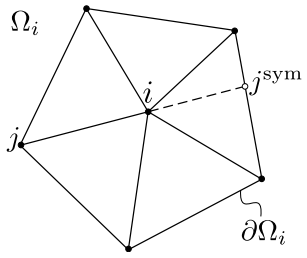


FIGURE 1. Representation of the symmetric node j^{sym} of j with respect to i.

denominator was used in [1] instead of the sum. In the next lemma we show that in fact (13) detects extrema.

Lemma 3.1. *The shock detector $\alpha_i(u_h)$ defined in* (13) *is equal to 1 if u_h has an extremum at point $\boldsymbol{x}_i$. Otherwise, $\alpha_i(u_h) < 1$ in general, and $\alpha_i(u_h) = 0$ for $q = \infty$.*

Proof. Using the fact that u_h has an extremum at $\boldsymbol{x}_i$,

$$\left|\sum_{j\in\mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij}\right| = \left|\sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{u_j-u_i}{|\boldsymbol{r}_{ij}|} + \frac{u_j^{\mathrm{sym}}-u_i}{|\boldsymbol{r}_{ij}^{\mathrm{sym}}|}\right| \tag{14}$$

$$= \sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{|u_j-u_i|}{|\boldsymbol{r}_{ij}|} + \frac{|u_j^{\mathrm{sym}}-u_i|}{|\boldsymbol{r}_{ij}^{\mathrm{sym}}|} = \sum_{j\in\mathcal{N}_h(\Omega_i)} 2\{\!\{|\boldsymbol{\nabla} u_h\cdot\hat{\boldsymbol{r}}_{ij}|\}\!\}, \tag{15}$$

since $u_j - u_i$ has the same sign (or it is equal to zero) in all directions. It proves that $\alpha_i(u_h) = 1$ on an extremum. In fact, if the solution does not have an extremum, these quantities neither can have the same sign nor be zero in all cases, and we only have

$$\left|\sum_{j\in\mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij}\right| < \sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{|u_j-u_i|}{|\boldsymbol{r}_{ij}|} + \frac{|u_j^{\mathrm{sym}}-u_i|}{|\boldsymbol{r}_{ij}^{\mathrm{sym}}|} = \sum_{j\in\mathcal{N}_h(\Omega_i)} 2\{\!\{|\boldsymbol{\nabla} u_h\cdot\hat{\boldsymbol{r}}_{ij}|\}\!\}. \tag{16}$$

Hence, $\alpha_i(u_h) < 1$ when there is no extremum at $\boldsymbol{x}_i$. Moreover, for $q = \infty$, the shock detector vanishes in all the nodes that are not extrema. □

In addition to the nonlinear stabilization term $\mathsf{B}(u_h)$, it is necessary to do a mass matrix lumping to prove that the LED property is satisfied. In the numerical analysis, it is enough to make this approximation when testing against the shape functions corresponding to nodes related to extrema, which is identified by the shock detector. Therefore, we propose the following stabilized semi-discrete version of (4):

$$\begin{aligned}(1-\alpha_i(u_h))(\partial_t u_h,\varphi_i) + \alpha_i(u_h)(\partial_t u_i,\varphi_i) + (\boldsymbol{f}'(u_h)\cdot\boldsymbol{\nabla} u_h,\varphi_i)\\ + \sum_{j\in\mathcal{N}_h(\Omega_i)} \nu_{ij}(u_h)v_i u_j l(i,j) = (g,\varphi_i) \quad \text{for any } i\in\mathcal{N}_h,\end{aligned} \tag{17}$$

with the definition of the shock detector (13) and the nonlinear artificial diffusion (10). Thus, the definition of the mass matrix is nonlinear

$$\mathsf{M}_{ij}(u_h) \doteq (1-\alpha_i(u_h))(\varphi_j,\varphi_i) + \alpha_i(u_h)(\delta_{ij},\varphi_i). \tag{18}$$

It can be understood as a mass matrix with gradual lumping. Full lumping is only attained at extrema. Denoting $\mathsf{K}(u_h) \doteq \mathsf{F}(u_h) + \mathsf{B}(u_h)$, the stabilized problem (17) can be expressed in compact form as

$$\mathsf{M}(u_h)\mathrm{d}_t u_h + \mathsf{K}(u_h)u_h = g \quad \text{ in } V_h'. \tag{19}$$

Analogously for the discrete problem (6),

$$\mathsf{M}(u_h^{n+1})\delta_t u_h^{n+1} + \mathsf{K}(u_h^{n+1})u_h^{n+1} = g^{n+1} \quad \text{ in } V_h'. \tag{20}$$

Finally, let us note that the shock detector (13) leads to the one of Barrenechea and co-workers [3],

$$\tilde{\alpha}_i \doteq \begin{cases} \left(\dfrac{\left| \sum_{j \in \mathcal{N}_h(\Omega_i)} u_i - u_j \right|}{\sum_{j \in \mathcal{N}_h(\Omega_i)} |u_i - u_j|} \right)^q & \text{if } \sum_{j \in \mathcal{N}_h(\Omega_i)} |u_i - u_j| \neq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

when restricted to symmetric meshes of equilateral triangles.

Lemma 3.2. *For a symmetric triangular mesh where all the edges have the same length, α_i in* (13) *is identical to $\tilde{\alpha}_i$ in* (21)*.*

Proof. For symmetric meshes, for every $j \in \mathcal{N}_h(\Omega_i)$, $j^{\text{sym}} \in \mathcal{N}_h(\Omega_i)$. So, we can group nodes in $\mathcal{N}_h(\Omega_i)$ in pairs, getting

$$2 \sum_{j \in \mathcal{N}_h(\Omega_i)} (u_i - u_j) = \sum_{j \in \mathcal{N}_h(\Omega_i)} (u_i - u_j + u_i - u_{ij}^{\text{sym}}).$$

We proceed analogously for the mean value. Further, since $\boldsymbol{r}_{ij}$ is identical for all $j \in \mathcal{N}_h(\Omega_i)$ by assumption, we get

$$\frac{\left| \sum_{j \in \mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij} \right|}{2 \sum_{j \in \mathcal{N}_h(\Omega_i)} \{\!\{ |\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}| \}\!\}_{ij}} = \frac{\left| \sum_{j \in \mathcal{N}_h(\Omega_i)} u_i - u_j \right|}{\sum_{j \in \mathcal{N}_h(\Omega_i)} |u_i - u_j|}.$$

□

For arbitrary symmetric meshes the methods only differ on the weights of the terms in the sums in (13) and all the required properties stated in (22) are readily satisfied for the use of the shock detector in (21). In general meshes, the shock detectors are different, and the one in (21) is not linearity preserving.

4. Monotonicity properties

In the sequel, we prove that the scheme (17) is LED. First, we define a set of necessary conditions on the nonlinear discrete operators that lead to LED schemes. They are the nonlinear extension of the ones for linear systems (see, e.g., [23]).

Theorem 4.1. *The semi-discrete problem* (19) *is LED if $g(\boldsymbol{x}) = 0$ in Ω and, for every node $i \in \mathcal{N}_h$ such that u_i is a local extremum, it holds:*

$$\mathsf{M}_{ij}(u_h) \doteq \delta_{ij} m_i, \ \textit{with } m_i > 0, \tag{22}$$

$$\mathsf{K}_{ij}(u_h) \leq 0 \ \ \forall i \neq j, \ \textit{and} \sum_{j \in \mathcal{N}_h(\Omega_i)} \mathsf{K}_{ij}(u_h) = 0. \tag{23}$$

Moreover, for $g(\boldsymbol{x}) \leq 0$ (resp. $g(\boldsymbol{x}) \geq 0$) in Ω and for all $i \in \mathcal{N}_h$ such that u_i is a local maximum (resp. minimum), if (22) *holds the maximum (resp. minimum) is diminishing (resp. increasing). These results are also true for the discrete problem* (20)*. Furthermore, the discrete problem* (20) *is positivity-preserving for $g = 0$ and $u_0 \geq 0$.*

Proof. Let us start proving the LED property. If u_i is a maximum, from (19), conditions in (22), and the fact that $\alpha_i(u_h) = 1$, we have:

$$g_i = m_i \mathrm{d}_t u_i + \sum_{j \in \mathcal{N}_h(\Omega_i)} \mathsf{K}_{ij}(u_h) u_j \geq m_i \mathrm{d}_t u_i + \sum_{j \in \mathcal{N}_h(\Omega_i)} \mathsf{K}_{ij}(u_h) u_i = m_i \mathrm{d}_t u_i,$$

for $m_i \doteq \int_\Omega \varphi_i \mathrm{d}\Omega$. As a result, $\mathrm{d}_t u_i \leq 0$ and thus LED. We proceed analogously for the minimum. The proof is analogous for the discrete problem with BE time integration.

Next, we prove positivity. Let us consider that at some time step m the solution becomes negative, and consider the node i in which the minimum value is attained. Using the previous result for a minimum at the discrete level, we have that $\delta_t u_i^m \geq 0$ and thus $u_i^m \geq u_i^{m-1}$. It leads to a contradiction, since $u_i^{m-1} \geq 0$. Thus, the solution must be positive at all times. □

Theorem 4.2 (LED)**.** *The semi-discrete (resp., discrete) problem* (19) *(resp.,* (20)*) leads to solutions* $u_h \in V_h$ *that enjoy the LED property in Def. 2.2 for any* $q \in \mathbb{R}^+$.

Proof. Assume u_h reaches an extremum on $i \in \mathcal{N}_h$. Then $\alpha_i(u_h) = 1$ and $\mathsf{M}_{ij}(u_h)\mathrm{d}_t u_j = m_i \mathrm{d}_t u_i$ with $m_i = \int_\Omega \varphi_i$. On the other hand, taking into account the definition of $\nu_{ij}(u_h)$ in (10), the convective term for $j \neq i$ reads

$$\mathsf{K}_{ij}(u_h) = \mathsf{F}_{ij}(u_h) - \max\{\mathsf{F}_{ij}(u_h), \alpha_j(u_h)\mathsf{F}_{ji}(u_h), 0\} \leq 0. \tag{24}$$

Using the fact that $\sum_{j\in\mathcal{N}_h(\Omega_i)} \mathsf{F}_{ij}(u_h) = (\boldsymbol{f}'(u_h)\cdot\boldsymbol{\nabla}1, \varphi_i) = 0$, the definition of $\nu_{ii}(u_h)$, and (9), we have

$$\begin{aligned}
\mathsf{K}_{ii}(u_h) &= \mathsf{F}_{ii}(u_h) + \sum_{j\in\mathcal{N}_h(\Omega_i)\setminus\{i\}} \max\{\mathsf{F}_{ij}(u_h), \alpha_j(u_h)\mathsf{F}_{ji}(u_h), 0\} && (25)\\
&= \sum_{j\in\mathcal{N}_h(\Omega_i)\setminus\{i\}} -\mathsf{F}_{ij}(u_h) + \sum_{j\in\mathcal{N}_h(\Omega_i)\setminus\{i\}} \max\{\mathsf{F}_{ij}(u_h), \alpha_j(u_h)\mathsf{F}_{ji}(u_h), 0\} && (26)\\
&= -\sum_{j\in\mathcal{N}_h(\Omega_i)\setminus\{i\}} \mathsf{K}_{ij}(u_h). && (27)
\end{aligned}$$

Therefore it is clear that the conditions stated in Theorem 4.1 hold, thus the method is LED. The discrete case is proved analogously. □

Corollary 4.3 (DMP)**.** *The discrete problem* (20) *leads to solutions that satisfy the local DMP property in Def. 2.1 at every* t^n*, for* $n = 1, \ldots, N$.

Proof. If the maximum (resp., minimum) at time t^n is on a node whose value is not on the Dirichlet boundary, it is known from the LED property in Theorem 4.2 that it is bounded above (resp., below) by the maximum (resp., minimum) at the previous time step value. By induction, it will be bounded by the maximum (resp., minimum) at $t = 0$. Alternatively, the maximum or minimum is on the Dirichlet boundary. It proves the result. □

Theorem 4.4. *The diffusion defined in* (10) *is the one that introduces the minimum amount of numerical dissipation* $\langle \mathsf{B}(u_h)u_h, u_h\rangle$ *required to satisfy* (22) *when* $q = \infty$.

Proof. Using the definition of the graph-Laplacian, the amount of dissipation introduced by the nonlinear stabilization is

$$\langle \mathsf{B}(u_h)u_h, u_h\rangle = \sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega)} \nu_{ij}(u_h)(u_i - u_j)^2.$$

Let us consider two connected nodes, i.e., $i, j \in \mathcal{N}_h$ and $j \in \mathcal{N}_h(\Omega_i)$. If neither i nor j are extrema, then $\alpha_i(u_h) = \alpha_j(u_h) = 0$ and $\nu_{ij} = 0$. Let us assume (without loss of generality) that u_h has an extremum at i. If $u_i = u_j$, the dissipation is independent of the expression for ν_{ij}. If $u_i > u_j$, $\alpha_j = 0$ (since $q = \infty$). Thus, $\nu_{ij} = -\max\{\mathsf{F}_{ij}(u_h), 0\}$. If $\mathsf{F}_{ij}(u_h) \leq 0$, no dissipation is introduced. If $\mathsf{F}_{ij}(u_h) > 0$, then the diffusion introduced by the method is $-\mathsf{F}_{ij}(u_h)$ and $\mathsf{K}_{ij}(u_h) = 0$.

Let us assume that we have a method that is less dissipative than the one proposed herein. Based on the previous analysis, there exists a pair of connected nodes such that $u_i > u_j$ and the dissipation introduced is smaller than $-\mathsf{F}_{ij}(u_h)$, for $\mathsf{F}_{ij}(u_h) > 0$. As a result, $\mathsf{K}_{ij}(u_h) > 0$. Thus, the properties in (4.1) do not hold. It proves the theorem. □

Furthermore, it can be proved that the above method (19) (also (20)) is linearly preserving. In addition, using (21) instead, the method is still linearly preserving for symmetric meshes.

Theorem 4.5 (Linearity preservation)**.** *Let* u_h *be a continuous first order FE approximation of* $u \in P_1(\Omega)$*, then the semi-discrete and discrete problems* (19) *and* (20)*, respectively, are linearity preserving, in the sense that the Galerkin problem and the stabilized one are identical.*

Proof. If $u_h \in P_1(\Omega)$, then it is obvious that $\boldsymbol{\nabla} u_h$ is constant. Thus, $[\![\boldsymbol{\nabla} u_h]\!]_{ij} = 0$ for any direction $\boldsymbol{r}_{ij}$, and $\alpha_i(u_h) = 0$ for any $i \in \mathcal{N}_h$. Therefore, recalling (10), it is easy to see that $\nu_{ij} = 0$ for any $i, j \in \mathcal{N}_h$. Thus, the nonlinear stabilization and gradual lumping terms vanish and the Galerkin scheme is recovered. □

5. Symmetric mass matrix stabilization

The nonlinear mass matrix that has been considered in (18) is nonsymmetric by construction. In any case, we can easily consider a symmetric version of the method.

Another alternative strategy to the nonlinear mass matrix definition in (18) is to consider the fully discrete problem (20), keeping the mass matrix at the current time step as a reaction term, leading to the following expression of the artificial diffusion

$$\begin{aligned} \tilde{\nu}_{ij}(w_h) &\doteq \nu_{ij}(w_h) + \tfrac{1}{\Delta t}\max\{\alpha_i \mathsf{M}_{ij}, 0, \alpha_j \mathsf{M}_{ji}\} \qquad \text{for } i \neq j, \\ \tilde{\nu}_{ii}(w_h) &\doteq \sum_{\substack{j \in \mathcal{N}_h(\Omega_i) \\ j \neq i}} \tilde{\nu}_{ij}. \end{aligned} \tag{28}$$

Let us consider another notion of DMP property.

Definition 5.1 (Global DMP)**.** *A solution satisfies the global DMP if given* $(\boldsymbol{x}, t)$ *in* $\Omega \times (0, T]$

$$\min_{(\boldsymbol{y},\bar{t})\in\Gamma} u(\boldsymbol{y}, \bar{t}) \leq u(\boldsymbol{x}, t) \leq \max_{(\boldsymbol{y},\bar{t})\in\Gamma} u(\boldsymbol{y}, \bar{t}) \tag{29}$$

where $\Gamma \doteq \Omega \times \{0\} \cup \Gamma_{\text{in}}$.

It is easy to check that the global DMP is a consequence of the local DMP and LED properties.

It is possible to prove that the modified method with BE time integration satisfies the global DMP in Def. 5.1. Linear preservation can also be easily checked.

Theorem 5.2 (Global DMP)**.** *Let* u_h *be a continuous first order FE approximation of* u*. Then, the BE time discretization of problem* (4) *with* $g = 0$*, stabilized with* (9)*, and using* (28) *as artificial diffusion, satisfies the global DMP property in Def.* (5.1) *for any* $q \in \mathbb{R}^+$.

Proof. Let us denote by $\mathsf{K}(u)$ and $\tilde{\mathsf{K}}(u)$ the stabilized matrix with the artificial diffusion computed with (10) and (28), respectively. Assume u_h reaches a maximum on $\boldsymbol{x}_i \in \Omega \backslash \Gamma_{\text{in}}$. Then $\alpha_i = 1$, and we have:

$$\mathsf{M}_{ij}(u_h)u_j + \tilde{\mathsf{K}}_{ij}(u_h)u_j = m_i u_i + \mathsf{K}_{ij}(u_h)u_j,$$

where we have used the fact that $\max\{\alpha_i \mathsf{M}_{ij}, 0, \alpha_j \mathsf{M}_{ji}\} = \mathsf{M}_{ij}$. Thus, the equation related to the test function φ_i leads to

$$\sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{\mathsf{M}_{ij}}{m_i} u_j^n = u_i^{n+1} + \sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{\mathsf{K}_{ij}(u_h)}{m_i} u_j^{n+1} \geq u_i^{n+1} + \sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{\mathsf{K}_{ij}(u_h)}{m_i} u_i^{n+1} = u_i^{n+1}. \tag{30}$$

Note that $\frac{\mathsf{M}_{ij}}{m_i} > 0$, and $\sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{\mathsf{M}_{ij}}{m_i} = 1$. Hence u_i^{n+1} is smaller or equal to a convex combination of u_j^n, for $j \in \mathcal{N}_h(\Omega_i)$, and thus it is bounded above by the largest of these values. As a result, $u_h^{n+1}(\boldsymbol{x}) \leq \max\{\max_{\boldsymbol{y}\in\Omega} u_h^n(\boldsymbol{y}), \max_{(\boldsymbol{y},t^{n+1})\in\Gamma_{\text{in}}} u_D(\boldsymbol{y}, t^{n+1})\}$. Using a recursion argument, we prove the upper bound. We proceed analogously for the case lower bound. It proves the theorem. □

6. Lipschitz continuity

In the next, we want to prove the Lipschitz continuity of the nonlinear operator at every time step, i.e., $\mathsf{T} : V_h \to V_h'$ defined as

$$\mathsf{T}(u_h) \doteq \Delta t_{n+1}^{-1} \mathsf{M}(u_h)u_h + \mathsf{K}(u_h)u_h - g - \Delta t_{n+1}^{-1} \mathsf{M}(u_h)u_h^n.$$

In order to prove the Lipschitz continuity of $\mathsf{T}(\cdot)$, we must deal with the nonlinear stabilization and gradual mass lumping terms. The Galerkin terms can be handled using the fact that $\boldsymbol{f} \in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$.

Let us introduce the following semi-norm generated by the graph-Laplacian operator

$$|w|_\ell \doteq \sqrt{\frac{1}{2}\sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)}(w_i-w_j)^2}.$$

Further, we define $|\beta|$ as the supremum of $|\boldsymbol{f}(v)|$ for $v \in V_h^{\mathrm{adm}}$, where $V_h^{\mathrm{adm}} \subset V_h$ is the subspace of functions that satisfy the global DMP in Def. 5.1.

Theorem 6.1. *Let us consider a non-degenerate partition $\mathcal{T}_h$. Given $u_h^n \in V_h$ and $g \in V_h'$, the nonlinear operators $\mathsf{B}(\cdot) : V_h \to V_h'$ and $\mathsf{M}(\cdot) : V_h \to V_h'$ are Lipschitz continuous in V_h^{adm} for $q \in \mathbb{N}^+$, since they satisfy*

$$\langle \mathsf{B}(u) - \mathsf{B}(v), z\rangle \le qh^{d-1}|\beta||u-v|_\ell |z|_\ell, \qquad \text{for any } z \in V_h,$$

$$\langle \mathsf{M}(u) - \mathsf{M}(v), z\rangle \le C(qh^{\frac{d}{2}}|u-v|_\ell + \|u-v\|)\|z\|, \qquad \text{for any } z \in V_h.$$

Proof. The proof of the theorem is included in Appendix A. □

7. Differentiable stabilization

The previous nonlinear system is Lipschitz continuous, which improves the convergence of the nonlinear iterations. In fact, assuming that we supplement (2) with a diffusive term, existence and uniqueness can be proved in the diffusive regime (see [3]). However, even using Anderson acceleration nonlinear convergence can be very hard (see [23, 24] and Sect. 9).

Based on these observations, we want to develop methods that lead to at least twice differentiable operators, i.e., $\frac{\partial^2 \mathsf{T}(u_h)}{\partial^2 u_h} \in \mathcal{C}^0$, using the previous framework. This allows the usage of the Newton method to linearize the system, and reduces the required number of nonlinear iterations. Smoothness is achieved by substituting the non-differentiable functions of the previous formulation with smooth approximations.

In order to end up with a twice differentiable method, we propose to use the following artificial diffusion:

$$\begin{aligned}
\nu_{ij} &\doteq \max\nolimits_\sigma\left\{\max\nolimits_\sigma\left\{\alpha_{\varepsilon,i}(\mathsf{F}_{ij}(w_h)), \alpha_{\varepsilon,j}\mathsf{F}_{ji}(w_h)\right\}, 0\right\}, \quad \text{for } i \neq j,\\
\nu_{ii} &\doteq \sum_{\substack{j\in\mathcal{N}_h(\Omega_i)\\ j\neq i}} \nu_{ij}.
\end{aligned} \tag{31}$$

The function $\max_\sigma(\cdot)$ is a regularized maximum function

$$\max\nolimits_\sigma\{x,y\} \doteq \frac{|x-y|_{1,\sigma}}{2} + \frac{x+y}{2}, \tag{32}$$

where $|x|_{1,\sigma} \doteq \sqrt{x^2+\sigma}$ is a smooth approximation of the absolute value. In order to keep dimensional consistency, σ should be a small parameter of order $\mathcal{O}\left(|\beta|^2\ell^{2(d-1)}\right)$, where ℓ is a characteristic length of the problem. Let us define the smooth limiter function $f(x) \in \mathcal{C}^2$ that will be used in the definition of α_ε,

$$f(x) \doteq \begin{cases} 2x^4 - 5x^3 + 3x^2 + x & \text{if} \quad x < 1\\ 1 & \text{if} \quad x \ge 1 \end{cases}. \tag{33}$$

This function is used to smoothly limit the value of x up to 1. Further, let us define another smooth approximation of the absolute value, namely

$$|x|_{2,\varepsilon} \doteq \frac{x^2}{\sqrt{x^2+\varepsilon}}.$$

Finally, the shock detector is defined as

$$\alpha_{\varepsilon,i}(u_h) \doteq \left[f\left(\frac{\left| \sum_{j\in\mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij} \right|_{1,\varepsilon} + \gamma}{\sum_{j\in\mathcal{N}_h(\Omega_i)} 2\left\{\!\!\left\{ |\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}|_{2,\varepsilon} \right\}\!\!\right\}_{ij} + \gamma} \right) \right]^q, \tag{34}$$

where γ is a small parameter that prevents division by zero.

It has been proved in Lemma 3.1 that α_i equals 1 when i is an extremum in Ω_i. Let us prove that this is still true for $\alpha_{\varepsilon,i}$.

Lemma 7.1. *If u_h has an extremum on $i \in \mathcal{N}_h$ then $\alpha_{\varepsilon,i}(u_h) = 1$.*

Proof. It is clear that $f(x)$ equals 1 for $x \geq 1$, then the proof reduces to check that

$$\left| \sum_{j\in\mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij} \right|_{1,\varepsilon} + \gamma \geq \sum_{j\in\mathcal{N}_h(\Omega_i)} 2\left\{\!\!\left\{ |\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}|_{2,\varepsilon} \right\}\!\!\right\}_{ij} + \gamma. \tag{35}$$

Taking into account that

$$\sqrt{x^2+\varepsilon} = |x|_{1,\varepsilon} > |x| \geq |x|_{2,\varepsilon} = \frac{x^2}{\sqrt{x^2+\varepsilon}}, \tag{36}$$

and the fact that $u_j - u_i$ has the same sign (or it is equal to zero) in all directions, it is easy to see that

$$\left| \sum_{j\in\mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij} \right|_{1,\varepsilon} = \left| \sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{u_j - u_i}{|\boldsymbol{r}_{ij}|} + \frac{u_j^{\mathrm{sym}} - u_i}{|\boldsymbol{r}_{ij}^{\mathrm{sym}}|} \right|_{1,\varepsilon} \tag{37}$$

$$\geq \left| \sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{u_j - u_i}{|\boldsymbol{r}_{ij}|} + \frac{u_j^{\mathrm{sym}} - u_i}{|\boldsymbol{r}_{ij}^{\mathrm{sym}}|} \right| \tag{38}$$

$$= \sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{|u_j - u_i|}{|\boldsymbol{r}_{ij}|} + \frac{\left|u_j^{\mathrm{sym}} - u_i\right|}{|\boldsymbol{r}_{ij}^{\mathrm{sym}}|} \geq \sum_{j\in\mathcal{N}_h(\Omega_i)} 2\{\!\{ |\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}| \}\!\} \tag{39}$$

$$\geq \sum_{j\in\mathcal{N}_h(\Omega_i)} 2\left\{\!\!\left\{ |\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}|_{2,\varepsilon} \right\}\!\!\right\}. \tag{40}$$

It proves that $\alpha_{\varepsilon,i}(u_h) = 1$ on an extremum. In fact, if the solution does not have an extremum, these quantities neither can have the same sign nor be zero in all cases. Since

$$\left| \sum_{j\in\mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij} \right| = \lim_{\varepsilon\to 0} \left| \sum_{j\in\mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij} \right|_{1,\varepsilon} \tag{41}$$

and

$$\sum_{j\in\mathcal{N}_h(\Omega_i)} 2\{\!\{ |\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}| \}\!\} = \lim_{\varepsilon\to 0} \sum_{j\in\mathcal{N}_h(\Omega_i)} 2\left\{\!\!\left\{ |\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ij}|_{2,\varepsilon} \right\}\!\!\right\}, \tag{42}$$

bound (16) leads to the fact that $\lim_{\varepsilon\to 0} \alpha_{\varepsilon,i}(u_h) < 1$ when there is no extremum on i. □

It is straightforward to check the following results.

Corollary 7.2. *System* (19) *with the definition of the shock detector* (34) *and artificial diffusion* (31) *is LED and satisfies the local DMP. The method tends to a linearly preserving scheme as $\gamma \to 0$.*

Proof. From lemma 7.1 and the definition of the regularized maximum (32) it is easy to see that artificial diffusion in (31) is greater or equal to the one in (10). Hence, Theorem 4.2 still holds. The linearity preservation is straighforward. □

Remark 7.3. *Note that the smoothed shock detector is not linearly preserving because $\alpha_{\varepsilon,i}$ will never be zero. However, for regions where u_h is constant the gradient is zero, thus the solution is not affected. In the case of $u_h \in P_1(\Omega)$, but not constant, $\alpha_{\varepsilon,i}$ goes to zero with γ. Values of γ of order 10^{-8} (or even smaller) have been considered in the numerical experiments section with good nonlinear convergence properties. Thus, the linearity preservation is virtually preserved in practice.*

As in the previous section, when restricted to symmetric meshes, the following approximation (similar to the one in Barrenechea et al. [3]) of (34) maintains the same properties

$$\tilde{\alpha}_{\varepsilon,i} \doteq \left[f\left(\frac{\left| \sum_{j \in \mathcal{N}_h(\Omega_i)} u_i - u_j \right|_{1,\varepsilon^*} + \gamma^*}{\sum_{j \in \mathcal{N}_h(\Omega_i)} |u_i - u_j|_{2,\varepsilon^*} + \gamma^*} \right) \right]^q , \tag{43}$$

with $\varepsilon^* \sim \mathcal{O}(h^2\varepsilon)$ and $\gamma^* \sim \mathcal{O}(h\gamma)$.

8. Nonlinear Solvers

In this section the methods used for solving the system of nonlinear equations resulting from the above formulation (20) with the artificial diffusion defined in (31) is discussed. Taking advantage of the differentiability of the stabilization described in Sect. 7, Newton's method is used for the smooth version of the method. In addition, we use fixed point iterations with Anderson acceleration to compare against Newton's method performance. In order to define the schemes, it is useful to write the time-discrete problem (20) as

$$\mathsf{A}(u_h^{n+1})u_h^{n+1} = \mathsf{G} \tag{44}$$

where G is the force vector. Let $\mathsf{J}(u_h^{n+1}) \doteq \frac{\partial \mathsf{T}(u_h^{n+1})}{\partial u_h^{n+1}}$ be the Jacobian.

Since the above problem is nonlinear we will solve it iteratively. We denote by $u_h^{k,n+1}$ the k-th iteration of u_h at time step $n+1$. Let us define some auxiliary variables used in the definition of the algorithms: m denotes the number of previous nonlinear iterations used in Anderson acceleration, s is the slope resulting form fitting the last m nonlinear errors, $s_{\min}$ is the minimum slope allowed before increasing the relaxation, ω is the relaxation parameter, $\omega_{\min}$ is its allowed minimum, $k_{\max}$ is the maximum nonlinear iterations allowed, *tol* is the nonlinear tolerance, and *nlerr* is the nonlinear error.

For the non-differentiable methods in Sect. 3 we use Picard linearization with Anderson acceleration (see Alg. 1). Our particular implementation also includes a simple convergence rate test, where it is decided if the relaxation parameter should be reduced or not. This improves the global convergence rate and the robustness of the method. Moreover, we add a projection onto V_h^{adm} to ensure that the global DMP in Def. 5.1 is satisfied at all nonlinear iterations. This step is of special interest in the case of solving for variables that cannot become negative, e.g., the density. In this case, the projection onto the space of admissible solutions is performed truncating the obtained solution. However, more sophisticated methodologies can be also applied but at a higher computational cost.

For the differentiable method, Newton's linearization is used (see Alg. 2). In addition, we supplement it with the line search method to improve robustness. We use numerical 1D minimization of the residual norm up to a tolerance of 10^{-4} for the line search method. Following the same approach in Alg. 1, a projection to the FE space of admissible solutions can be performed in Alg. 2. As said before, this step ensures that for all nonlinear iterations the solution satisfies the global DMP. The numerical experiments in the next section show that the modified method keeps quadratic convergence, even though we do not have a theoretical analysis.

Algorithm 1: Fixed point iterations with relaxed Anderson acceleration

Input: $u_h^{0,n+1}$, m, $s_{\min}$, $\omega_{\min}$, tol, A, G, $k_{\max}$
Output: $u_h^{k,n+1}$,k
$k=1$, $nlerr^1 = tol$
while *($nlerr^k \geq tol$) and ($k < k_{\max}$)* **do**
  Set $m^k = \min(k, m)$
  Solve $\mathsf{A}(u_h^{k,n+1})\tilde{u}_h^{k,n+1} = \mathsf{G}$
  Compute $r^{k,n+1} = \tilde{u}_h^{k,n+1} - u_h^{k,n+1}$
  Minimize $\| \sum_{i=1}^{m^k} \xi_i^k r^{k-m^k+i,n+1} \|$ with respect to ξ_i^k subject to $\sum_{i=1}^{m^k} \xi_i^k = 1$
  Set $u_h^{k+1,n+1} = (1-\omega_k) \sum_{i=1}^{m^k} \xi_i^k u_h^{k-m+i,n+1} + \omega_k \sum_{i=1}^{m^k} \xi_i^k \tilde{u}_h^{k-m^k+i,n+1}$
  Project $u_h^{k+1,n+1}$ to V_h^{adm}
  Set $nlerr^k = \frac{\|u_h^{k+1,n+1} - u_h^{k,n+1}\|}{\|u_h^{k+1,n+1}\|}$
  Compute the slope (s) of $\{nlerr^i\}$ with $k \geq i \geq k - m^k$
  if *($s < s_{\min}$) and ($\omega > \omega_{\min}$)* **then**
    Set $\omega_{k+1} = \omega_k - 0.1$
  else
    Set $\omega_{k+1} = \omega_k$
  Update $k = k+1$

Algorithm 2: Newton's method + Line search

Input: $u_h^{0,n+1}$,u_h^n, tol, J, R, $k_{\max}$
Output: $u_h^{k,n+1}$,k
$k=1$, $nlerr^1 = tol$
while *($nlerr^k \geq tol$) and ($k < k_{\max}$)* **do**
  Solve $\mathsf{J}(u_h^{k,n+1})\Delta u_h^{k,n+1} = -\mathsf{T}(u_h^{k,n+1})$
  Minimize $\|\mathsf{T}(u_h^{k,n+1} + \xi^k \Delta u_h^{k,n+1})\|$ with respect to $\xi \in [0,1]$
  Set $u_h^{k+1,n+1} = u_h^{k,n+1} + \xi^k \Delta u_h^{k,n+1}$
  Project $u_h^{k+1,n+1}$ to V_h^{adm}
  Set $nlerr^k = \frac{\|\xi^k \Delta u_h^{k,n+1}\|}{\|u_h^{k+1,n+1}\|}$
  Update $k = k+1$

9. Numerical Experiments

9.1. **Steady problems.** First, in order to test the previous formulation, the convergence to a smooth solution is analyzed. For this purpose, the following equation is solved

$$\begin{aligned} \boldsymbol{\nabla} \cdot (\mathbf{v}u) &= 0 \quad \text{in} \quad \Omega = [0,1] \times [0,1], \\ u &= u_D \quad \text{on} \quad \Gamma_{\text{in}}, \end{aligned} \tag{45}$$

with $\mathbf{v}(x,y) \doteq (1,0)$, and inflow boundary conditions $u_D = y - y^2$ on $\partial\Omega \backslash \{x=1\}$. This problem consists in the transport of the parabolic profile along the x direction, which has the analytical solution $u(x,y) = y - y^2$.

Fig. 2 shows the convergence rates using the previously defined formulation ((20) with (31)), and the Galerkin formulation. To perform this test, an initial mesh of $12 \times 12\, Q_1$ has been considered, then successive refinements have been performed up to a $96 \times 96\, Q_1$ mesh. Analogous meshes has been also used for P_1 FE. Newton's method has been used with $q=4$, $\varepsilon = 10^{-7}$, $\sigma = |\beta| h^4 10^{-8}$ and $\gamma = 10^{-10}$. In this case, σ has been scaled as $|\beta|^2 L^{2(d-3)} h^4$ in order to recover optimal convergence,

where L denotes a characteristic length of the physical domain Ω. As desired, the convergence rates are not affected by the stabilization, while (as expected) the stabilized solutions have higher errors.

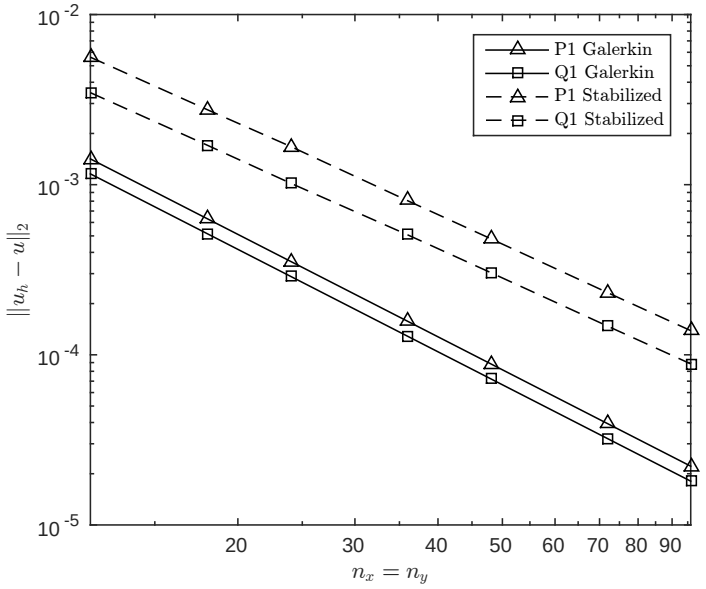


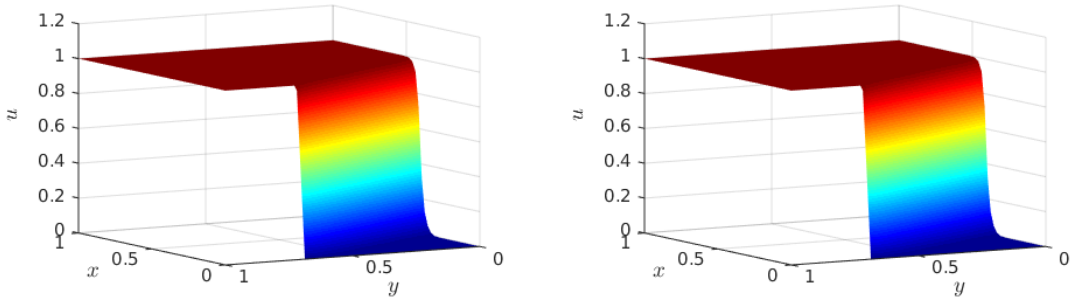
FIGURE 2. Convergence test, $L^2(\Omega)$ error versus size of the mesh. For P_1 and Q_1 FE meshes ranging from $h = 1/12$ to $h = 1/96$. Newton's method has been used with parameters $q = 4$, $\varepsilon = 10^{-7}$, $\sigma = |\beta|h^4 10^{-8}$ and $\gamma = 10^{-10}$.

A typical linear test to assess the performance of a shock capturing method is the propagation of a discontinuity. Consider now the previous hyperbolic PDE (45) with $\mathbf{v}(x, y) \doteq (1/2, \sin -\pi/3)$, and inflow boundary conditions $u_D = 1$ on $\{x = 0\} \cap \{y > 0.7\}$ and $y = 1$, while $u_D = 0$ at the rest of the inflow boundary. This problem has the following analytical solution

$$u(x, y) = \begin{cases} 1 & \text{if} \quad y > 0.7 + 2x \sin -\pi/3, \\ 0 & \text{otherwise.} \end{cases} \tag{46}$$

At Fig. 3(a), the numerical solution using the stabilization in (31) is shown. A $48 \times 48\, Q_1$ mesh have been used. The values chosen for the parameters in (31) are $q = 25$, $\varepsilon = 10^{-4}$, $\sigma = |\beta|10^{-9}$, and $\gamma = 10^{-10}$. This parameter choice makes the solution at the outflow sharp while the DMP is always satisfied. Furthermore, convergence is not jeopardized thanks to the smoothed stabilization. Particularly, it took 18 iterations for the Newton's method to converge to a nonlinear tolerance of 10^{-6}. The non-smooth version in Fig. 3(b) ((19) with (10)) did not converge using Anderson acceleration, adding a fixed relaxation parameter of $\omega = 0.5$ took 392 iterations, and 117 with Alg. 1. In any case, observing Fig. 4, where the outflow profile is depicted, no apparent improvement on accuracy is observed when using the non-smooth version.

Fig. 5 shows the solution for several combinations of q and ε, with $\sigma = |\beta|\varepsilon 10^{-5}$ and $\gamma = 10^{-10}$, solved with the two nonlinear solvers presented in the previous section over a $48 \times 48\, Q_1$ mesh. Furthermore, the $\|u - u_h\|_{L^1}$ and $\|u - u_h\|$ errors, computed at the whole domain and restricted to the outflow boundary, are listed in Table 1. These results show that, as expected, either increasing q or reducing ε the L^1 error diminishes. Nevertheless, the computational cost also increases at a higher rate. The same can be observed for the L^2 error. It is slightly reduced after increasing q or diminishing ε, while this makes nonlinear convergence much harder. Moreover, comparing both nonlinear solvers in Sect. 8, it is important to note that using Newton's method the number of nonlinear iterations is reduced between 10 to 15 times.



(A) Smooth stabilization (34), with parameters $q = 25$, $\varepsilon = 10^{-4}$, $\sigma = |\beta|10^{-9}$, and $\gamma = 10^{-10}$.

(B) Non-smooth version (13) with $q = 25$.

FIGURE 3. Stabilized solution of the straight propagation of a discontinuity test using the steady version of discrete problem (20) with two stabilization choices (34) or (13).

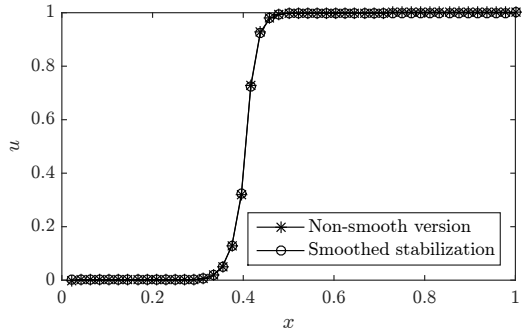


FIGURE 4. Stabilized solution of the straight propagation of a discontinuity test using the steady version of discrete problem (20) with two stabilization choices (34) and (13). The stabilization parameters used for the smoothed version are $q = 25$, $\varepsilon = 10^{-4}$, $\sigma = |\beta|10^{-9}$, and $\gamma = 10^{-10}$.

It is important to analyze the solution at each nonlinear iteration. If the projection to the space of admissible solutions is not performed, it is possible that the solution does neither satisfy the local nor the global DMP (Def. 2.1 or 5.1, resp.) at some nonlinear iterations. The DMP is only proved when convergence is attained. We denote by global DMP violation the difference between the global extremum of the analytical solution and the actual global extremum of the numerical solution. Fig. 6 shows the global DMP violation of the maximum and the minimum values produced at each nonlinear iteration for different values of q, ε, and σ. For $q = 25$, the global DMP is clearly not satisfied at the beginning of the iterative process. In this particular case, this does not destroy the nonlinear convergence, but this is not the case in some other problems, e.g. Euler's equations. Therefore, adding a projection step to V_h^{adm} is highly recommended. Further, it can be observed in Table 1 that in practice the projection step almost does not affect Newton convergence rate.

Finally, it is worth to test the nonlinear convergence of the method as the mesh is refined for a problem with a discontinuity. For this purpose, we have solved the previous benchmark with $q = 4$, $\varepsilon = 10^{-2}$, $\sigma = |\beta|h^4 10^{-6}$, and $\gamma = 10^{-10}$. The used meshes range form $12 \times 12\, Q_1$ to $96 \times 96\, Q_1$.

At Fig. 7, the number of nonlinear iterations for each mesh size is depicted. For Alg. 1 it can be observed that the number of iterations is increasing. On the contrary, this behavior is much less pronounced for Alg. 2; the number of iterations slightly increases and remains constant in the last interval.

TABLE 1. Straight propagation test errors and iterations, using the steady version of discrete problem (20) and nonlinear diffusion (31), for different values of q and ε, $\sigma = |\beta|\varepsilon 10^{-5}$, $\gamma = 10^{-10}$, and both nonlinear solvers in Sect. 8.

q	ε	Iterations				L_1 error	L_1 error at Γ_{out}	L_2 error	L_2 error at Γ_{out}
		A	Ap	N	Np				
1	10^{-1}	42	42	9	9	2.77e-02	5.57e-02	8.65e-02	1.23e-01
1	10^{-2}	43	42	8	8	2.61e-02	5.16e-02	8.40e-02	1.18e-01
1	10^{-3}	50	58	7	7	2.59e-02	5.09e-02	8.37e-02	1.17e-01
1	10^{-4}	50	57	7	7	2.58e-02	5.08e-02	8.37e-02	1.17e-01
1	0	56	47			2.59e-02	5.10e-02	8.37e-02	1.17e-01
4	10^{-1}	51	64	8	8	2.20e-02	4.43e-02	7.79e-02	1.12e-01
4	10^{-2}	58	61	11	11	1.83e-02	3.45e-02	6.97e-02	9.70e-02
4	10^{-3}	60	68	10	10	1.77e-02	3.28e-02	6.83e-02	9.44e-02
4	10^{-4}	66	85	11	11	1.76e-02	3.25e-02	6.82e-02	9.40e-02
4	0	70	73			1.76e-02	3.24e-02	6.81e-02	9.39e-02
8	10^{-1}	62	70	9	9	2.10e-02	4.27e-02	7.68e-02	1.11e-01
8	10^{-2}	71	63	11	11	1.62e-02	3.04e-02	6.63e-02	9.23e-02
8	10^{-3}	82	67	13	13	1.51e-02	2.75e-02	6.33e-02	8.74e-02
8	10^{-4}	70	77	12	12	1.49e-02	2.69e-02	6.27e-02	8.66e-02
8	0	94	60			1.48e-02	2.68e-02	6.26e-02	8.64e-02
25	10^{-1}	39	58	11	12	2.03e-02	4.18e-02	7.63e-02	1.11e-01
25	10^{-2}	57	62	19	20	1.46e-02	2.78e-02	6.39e-02	8.95e-02
25	10^{-3}	154	66	15	15	1.28e-02	2.35e-02	5.90e-02	8.24e-02
25	10^{-4}	116	82	17	18	1.25e-02	2.27e-02	5.79e-02	8.18e-02
25	0	86	163			1.23e-02	2.25e-02	5.75e-02	8.15e-02

A: Alg. 1 without projecting to V_h^{adm}, Ap: Alg. 1.
N: Alg. 2 without projecting to V_h^{adm}, Np: Alg. 2.

Consider now the hyperbolic PDE (45) on $\Omega = [0,1] \times [-1,1]$ with $\mathbf{v}(x,y) \doteq (y,-x)$, and inflow boundary conditions

$$u_D = \begin{cases} 1 & \text{if} \quad 0.35 < x < 0.65, \\ 0 & \text{otherwise.} \end{cases} \tag{47}$$

This particular configuration has the following analytical solution

$$u(x,y) = \begin{cases} 1 & \text{if} \quad 0.35 < \sqrt{x^2+y^2} < 0.65, \\ 0 & \text{otherwise.} \end{cases} \tag{48}$$

At Fig. 8 the solutions at the outflow boundary are depicted for several combinations of q and ε, with $\sigma = |\beta|\varepsilon 10^{-5}$ and $\gamma = 10^{-10}$. In all cases, we have considered the two schemes presented in Sect. 8 using a $64 \times 128\, Q_1$ FE mesh. As for the previous numerical experiment, we collect the number of iterations and the errors in Table 2. We observe that it is particularly difficult to converge to the solution for $q = 1$ and small values of ε. In any case, for q equal to 4 or greater, the number of iterations increase with q, as naturally expected. We also observe in this test that the number of nonlinear iterations can be highly reduced using Newton's method. Particularly, it reduces the number of nonlinear iterations up to 20 times. 3D plots of the smoothest and the sharpest solutions in Fig. 8 (respectively top-left and bottom-right subfigures) are shown in Fig. 9 .

Fig. 10 shows that in this second test, as in the previous one, if the projection step is not performed the global DMP (Def. 5.1) is not satisfied at all nonlinear iterations. This is specially evident for the combination shown in the figure, i.e., high values of q and low values of ε and σ.

9.2. **Transient transport problems.** Let us test the performance of the stabilization method in Sect. 7 for transient problems. For this purpose we will consider the 3 body rotation benchmark that reads as:

$$\begin{aligned} \partial_t u + \boldsymbol{\nabla} \cdot (\mathbf{v} u) &= 0 \quad \text{in} \quad \Omega = [0,1] \times [0,1], \\ u &= 0 \quad \text{on} \quad \Gamma_{\text{in}}, \\ u &= u_0 \quad \text{at} \quad t = 0, \end{aligned} \tag{49}$$

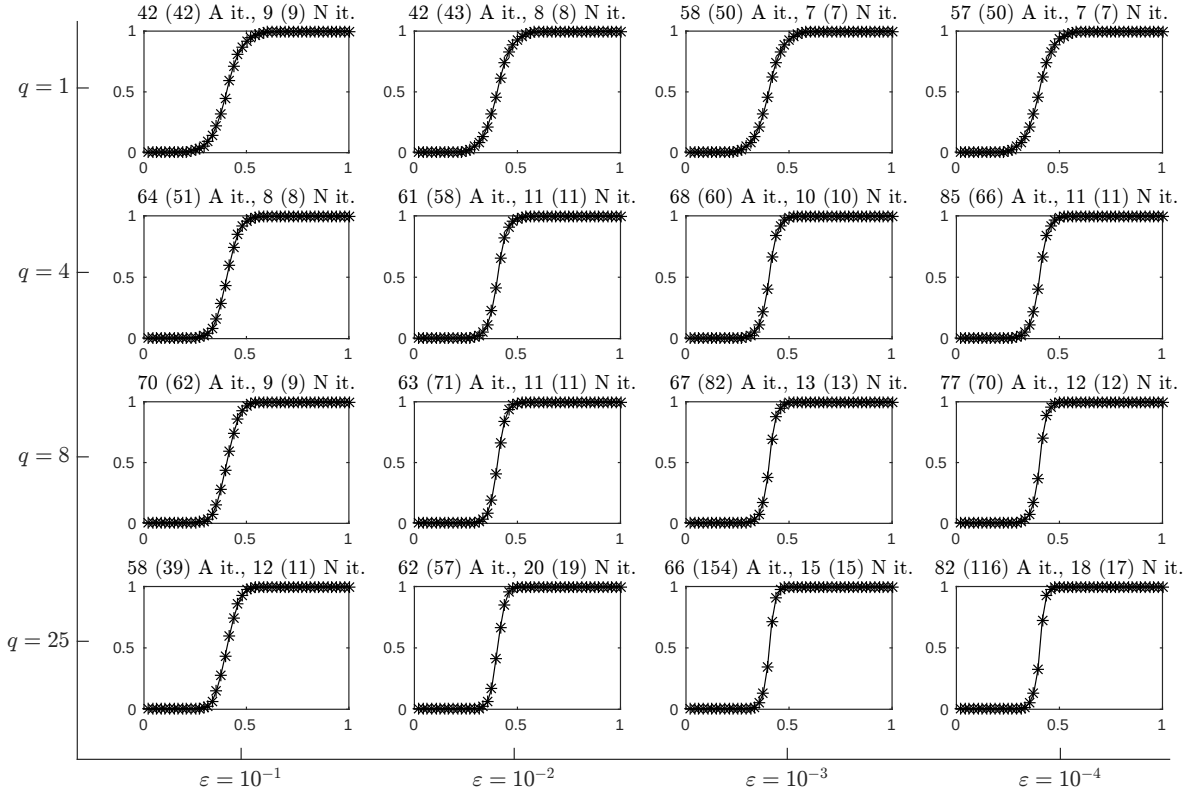


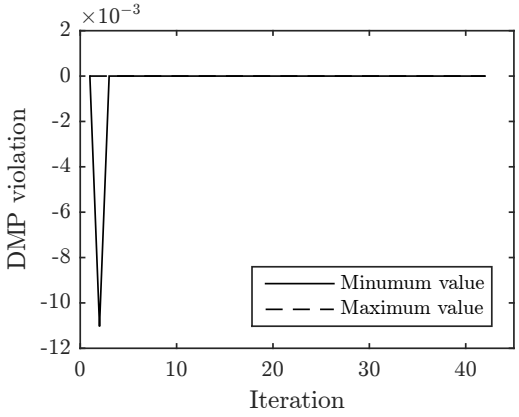
FIGURE 5. Straight propagation test solution at the outflow boundary $\partial\Omega\backslash\Gamma_{\text{in}}$. Using the steady version of discrete problem (20) and nonlinear diffusion (31), for different values of q and ε, $\sigma = |\beta|\varepsilon 10^{-5}$, $\gamma = 10^{-10}$, and both nonlinear solvers in Sect. 8. The result in brackets shows the number of iterations if no projection to V_h^{adm} is done.

where $\mathbf{v} \doteq (1/2 - y, x - 1/2)$ and

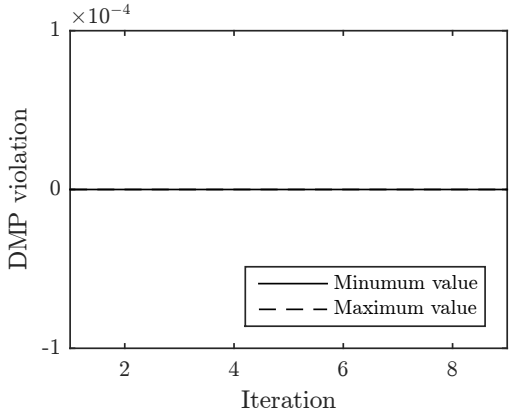
$$u_0(x,y) \doteq \begin{cases} \frac{1}{4} + \cos\left(\frac{\pi\sqrt{(x-0.25)^2+(y-0.5)^2}}{0.15}\right)/4 & \text{if} \quad \sqrt{(x-0.25)^2+(y-0.5)^2}/0.15 \leq 1 \\ 1 - \sqrt{(x-0.5)^2+(y-0.25)^2}/0.15 & \text{if} \quad \sqrt{(x-0.5)^2+(y-0.25)^2}/0.15 \leq 1 \\ 1 & \text{if} \quad \begin{cases} \sqrt{(x-0.5)^2+(y-0.75)^2}/0.15 \leq 1 \\ 0.55 < x < 0.45,\ y > 0.85 \end{cases} \end{cases} . \tag{50}$$

The above problem is solved in a 150×150 Q_1 FE mesh, with solver parameters $q = 25$, $\gamma = 10^{-8}$, $\sigma = |\beta|10^{-10}$, and $\varepsilon = 10^{-4}$. The discretization in time is performed using the BE method with a time step of 10^{-3}. At Fig. 11(a), the initial solution is depicted. Figs. 11(b) to 11(d) show the solution after one revolution (at time $t = 2\pi$).

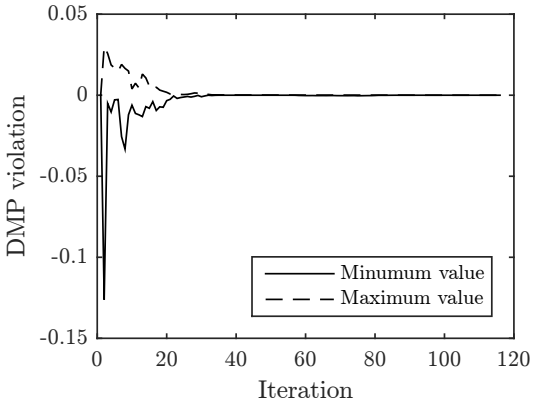
The solution obtained with the stabilization in (31), (28), and (10) are depicted in Figs. 11(b), 11(c), and 11(d), respectively. It is observed that the symmetric mass matrix method yields slightly more diffusive solutions than the LED method. This can be better observed in Fig. 12, where a cross-section of each of the figures rotated is depicted at $t = 0$ and after one revolution ($t = 2\pi$) for all three methods. As naturally expected, regularizing the stabilization makes the method faster to converge but the solution becomes smoother. Nevertheless, the regularization parameters (σ and ε) allow one to take the choice that better fits the requirements, either a faster but smoother method or the opposite. In any case, all schemes satisfy the DMP at all time steps.



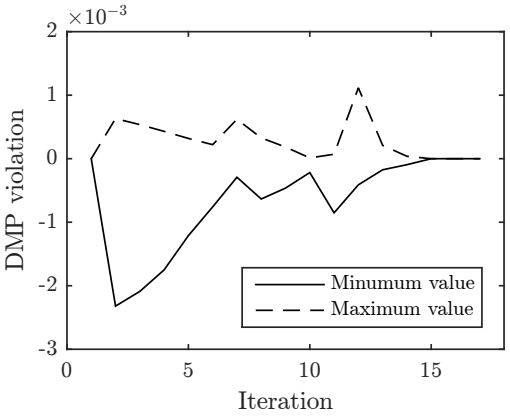
(A) Using Alg. 1 with $q = 1$, $\varepsilon = 10^{-1}$, $\sigma = |\beta|10^{-6}$, and $\gamma = 10^{-10}$.



(B) Using Alg. 2 with $q = 1$, $\varepsilon = 10^{-1}$, $\sigma = |\beta|10^{-6}$, and $\gamma = 10^{-10}$.



(C) Using Alg. 1 with $q = 25$, $\varepsilon = 10^{-4}$, $\sigma = |\beta|10^{-9}$, and $\gamma = 10^{-10}$.



(D) Using Alg. 2 with $q = 25$, $\varepsilon = 10^{-4}$, $\sigma = |\beta|10^{-9}$, and $\gamma = 10^{-10}$.

FIGURE 6. Evolution of global DMP violation during nonlinear iterations when avoiding the projection step in Algs. 1 and 2 for the straight propagation of a discontinuity test.

9.3. **Burgers' equation.** Finally, let us test our stabilization with a nonlinear transient problem. Particularly the 2D Burgers' equation, i.e. equation (49) with $\mathbf{v} \doteq (1,1)^u/2$, is solved on $\Omega = [0,1]\times[0,1]$ using a $150 \times 150\, Q_1$ mesh. The discretization in time is performed using the BE method with a time step of 10^{-2}. The initial conditions at $t = 0$ are

$$u_0(x,y) \doteq \begin{cases} -0.2 & \text{if} \quad x < 0.5 \quad \text{and} \quad y > 0.5 \\ -1 & \text{if} \quad x > 0.5 \quad \text{and} \quad y > 0.5 \\ 0.5 & \text{if} \quad x < 0.5 \quad \text{and} \quad y < 0.5 \\ 0.8 & \text{if} \quad x > 0.5 \quad \text{and} \quad y < 0.5 \end{cases}, \tag{51}$$

and the solution is advanced until $t = 0.5$.

The following stabilization parameters have been used for obtaining the results in Fig. 13(a): $q = 1$, $\varepsilon = 10^{-3}$, $\sigma = |\beta|10^{-6}$, and $\gamma = 10^{-8}$. Although the parameters used are not enforcing a particularly sharp solution (see Figs. 5 and 8), Fig. 13(a) shows properly transported and minimally smeared shocks. Only in the lower right region the method appears to be more diffusive than desired. Notice that in that region the gradient in the x direction spreads as y increases, while it should not. Nevertheless, in Fig. 13(b), that shows the solution for $q = 4$, $\varepsilon = 10^{-4}$, $\sigma = |\beta|10^{-7}$, and $\gamma = 10^{-8}$.

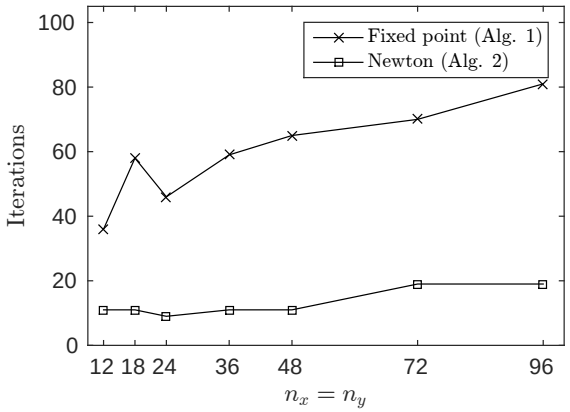


FIGURE 7. Straight propagation test nonlinear iterations as mesh refined from $12 \times 12\,Q_1$ to $96 \times 96\,Q_1$, for both Alg. 1 and Alg. 2. The shock capturing parameters used are $q = 4$, $\varepsilon = 10^{-2}$, $\sigma = |\beta| h^4 10^{-6}$, and $\gamma = 10^{-10}$.

TABLE 2. Circular propagation test errors and iterations, using the steady version of discrete problem (20) and nonlinear diffusion (31), for different values of q and ε, $\sigma = |\beta|\varepsilon 10^{-5}$, $\gamma = 10^{-10}$, and both nonlinear solvers in Sect. 8.

q	ε	Iterations				L_1 error	L_1 error at Γ_{out}	L_2 error	L_2 error at Γ_{out}
		A	Ap	N	Np				
1	10^{-1}	30	30	9	9	1.42e-01	1.93e-01	2.01e-01	2.36e-01
1	10^{-2}	–	54	10	10	1.11e-01	1.50e-01	1.74e-01	2.05e-01
1	10^{-3}	–	–	11	11	1.05e-01	1.42e-01	1.68e-01	1.99e-01
1	10^{-4}	196	–	19	19	1.04e-01	1.40e-01	1.68e-01	1.98e-01
1	0	–	–			–	–	–	–
4	10^{-1}	23	23	10	10	1.33e-01	1.82e-01	1.97e-01	2.31e-01
4	10^{-2}	64	64	15	15	8.47e-02	1.15e-01	1.55e-01	1.84e-01
4	10^{-3}	105	111	22	22	6.74e-02	9.31e-02	1.34e-01	1.64e-01
4	10^{-4}	–	139	24	24	6.38e-02	8.88e-02	1.29e-01	1.60e-01
4	0	198	194			6.31e-02	8.80e-02	1.28e-01	1.59e-01
8	10^{-1}	23	22	11	11	1.32e-01	1.81e-01	1.97e-01	2.31e-01
8	10^{-2}	73	68	15	15	8.10e-02	1.10e-01	1.53e-01	1.82e-01
8	10^{-3}	95	96	19	19	5.91e-02	8.18e-02	1.28e-01	1.57e-01
8	10^{-4}	100	109	22	22	5.28e-02	7.46e-02	1.18e-01	1.50e-01
8	0	256	231			5.12e-02	7.28e-02	1.16e-01	1.48e-01
25	10^{-1}	22	22	14	14	1.32e-01	1.80e-01	1.97e-01	2.31e-01
25	10^{-2}	45	49	16	15	7.82e-02	1.07e-01	1.51e-01	1.80e-01
25	10^{-3}	77	70	20	20	5.37e-02	7.50e-02	1.24e-01	1.54e-01
25	10^{-4}	131	109	23	24	4.51e-02	6.49e-02	1.11e-01	1.44e-01
25	0	180	289			4.22e-02	6.14e-02	1.06e-01	1.39e-01

A: Alg. 1 without projecting to V_h^{adm}, Ap: Alg. 1.
N: Alg. 2 without projecting to V_h^{adm}, Np: Alg. 2.

the method is less diffusive and the obtained shocks are even sharper. In any case, both choices satisfy the DMP for all time steps.

10. CONCLUSIONS

In this work, we have considered a nonlinear stabilization technique for the FE approximation of scalar conservation laws with implicit time stepping. The method relies on an artificial diffusion method, based on a graph-Laplacian operator. The artificial diffusion is judiciously chosen in order to satisfy a local DMP for steady problems. It is nonlinear, since it depends on a shock detector. Further,

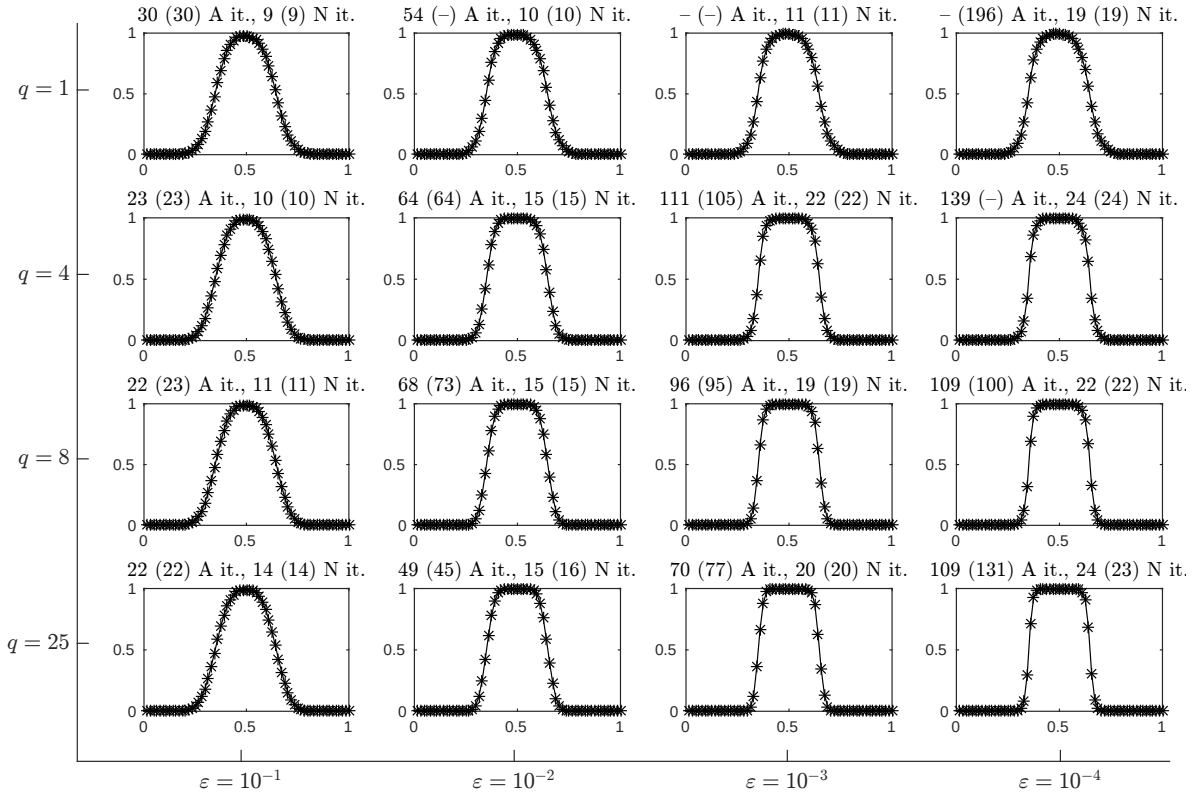


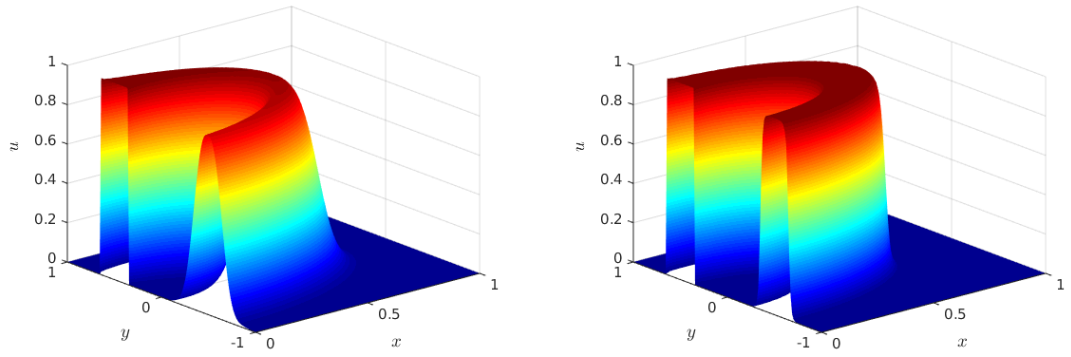
FIGURE 8. Circular propagation test solution at the outflow boundary $\partial\Omega\backslash\Gamma_{\text{in}}$. Using the steady version of discrete problem (20) and nonlinear diffusion (31), for different values of q and ε, $\sigma = |\beta|\varepsilon 10^{-5}$, $\gamma = 10^{-10}$ and both nonlinear solvers in Sect. 8. The result in brackets shows the number of iterations if no projection to V_h^{adm} is done.

the resulting method is linearity preserving. The same shock detector is used to gradually lump the mass matrix. The resulting method is LED, positivity preserving, and also satisfies a global DMP. Lipschitz continuity has also been proved.

However, the resulting scheme is highly nonlinear, leading to very poor nonlinear convergence rates, even using Anderson acceleration techniques. It is due to the fact that the nonlinear operator to be inverted at every time step is non-differentiable. The critical problem of nonlinear convergence of implicit monotonic methods based on nonlinear artificial diffusion have already been previously reported in the literature (see [22]). As a result, we propose a smooth version of the scheme. It leads to twice differentiable nonlinear stabilization schemes, which allows one to straightforwardly use Newton's method using the exact Jacobian. Twice differentiability ensures quadratic convergence.

We have considered two nonlinear solvers, namely Anderson acceleration and Newton's method. We have observed numerically that the effect of the smoothness has a positive impact in the reduction of the computational cost. The impact of using Newton's method versus Anderson acceleration is also very positive. In general, using the Newton method with a smooth version of the method we can reduce 10 to 20 times the number of iterations of Anderson acceleration with the original non-smooth algorithms.

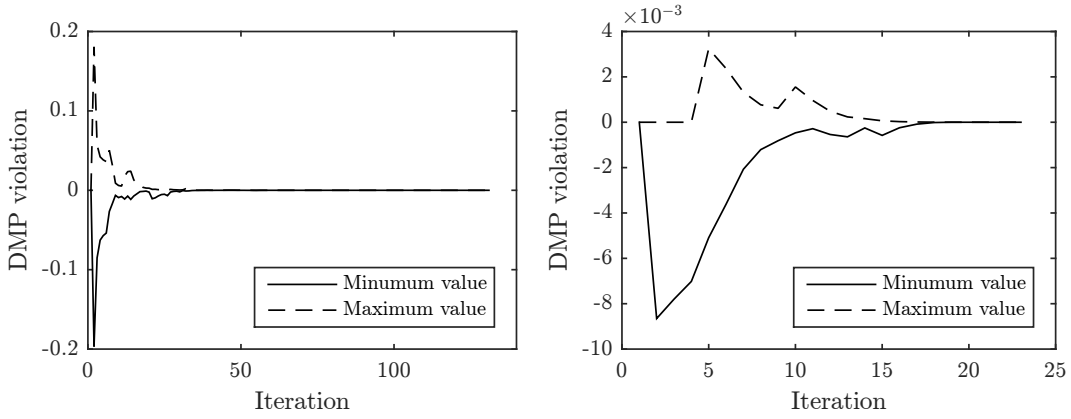
All the monotonic properties are satisfied (as theoretically proved) in the numerical experiments. Steady and transient linear transport, and transient Burgers' equation have been considered in 2D. In any case, these properties are only true for the converged solution, but not for iterates. In this sense, we have also proposed the concept of projected nonlinear solvers, where a projection step is



(A) Smoothest solution with parameters: $q = 1$, $\varepsilon = 10^{-1}$, $\sigma = |\beta|10^{-6}$, and $\gamma = 10^{-10}$.

(B) Sharpest solution with parameters: $q = 25$ and $\varepsilon = 10^{-4}$, $\sigma = |\beta|10^{-9}$, and $\gamma = 10^{-10}$.

FIGURE 9. Stabilized solution of the circular convection test using the steady version of the discrete problem (20) and the nonlinear diffusion (31) for two different parameter choices.



(A) Using Alg. 1

(B) Using Alg. 2

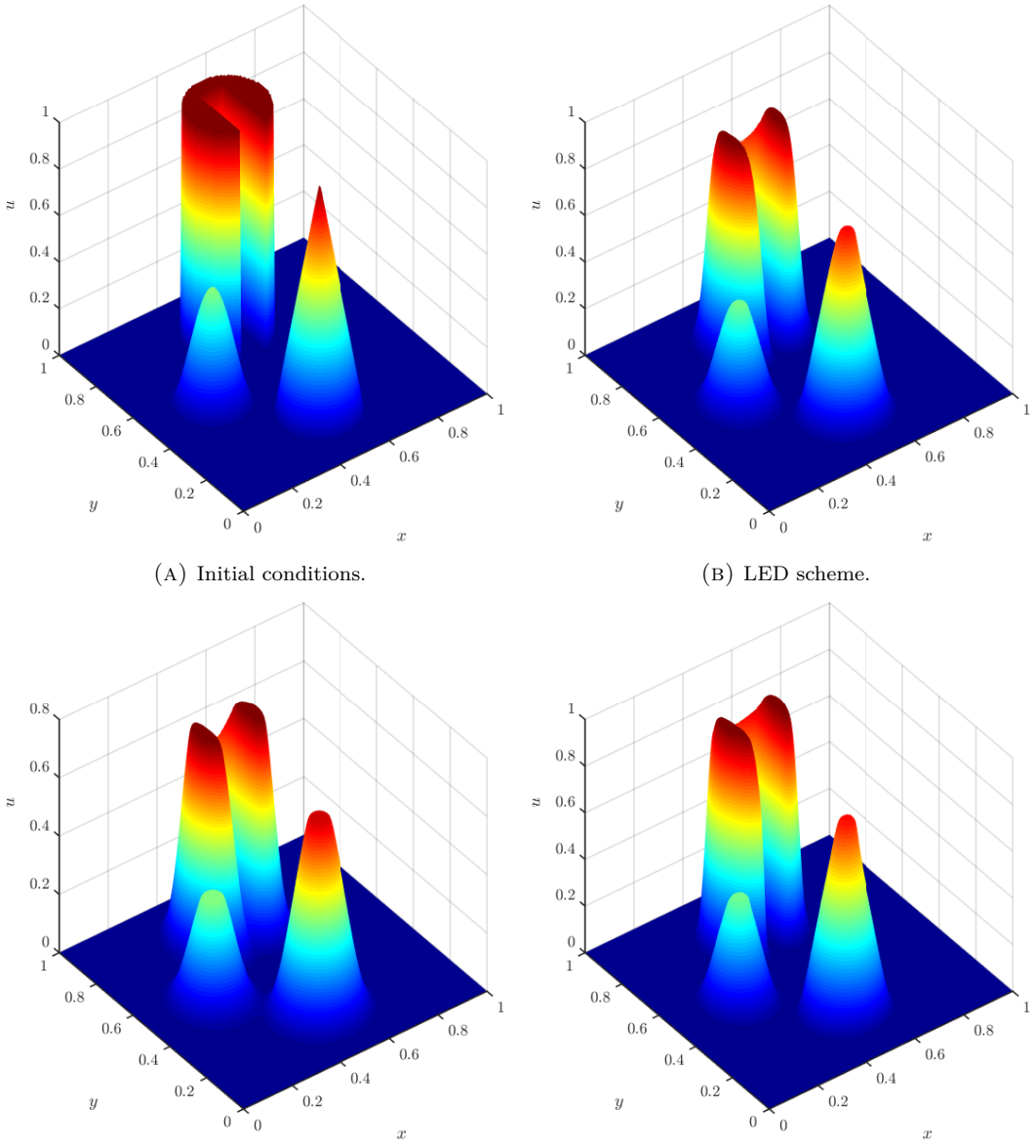
FIGURE 10. Evolution of global DMP violation during nonlinear iterations when avoiding the projection step in Algs. 1 and 2 for the circular propagation of a discontinuity. Using $q = 25$, $\varepsilon = 10^{-4}$, $\sigma = |\beta|10^{-9}$, $\gamma = 10^{-10}$.

performed at the end of every nonlinear iterations onto a FE space of admissible solutions. The space of admissible solutions is the one that satisfies the desired monotonic properties (maximum principle or positivity). The projection has no effect on the quality of the nonlinear convergence.

Future work should tackle the entropy stability analysis of the resulting schemes when applied to nonlinear problems. Some initial results in this direction can be found in [5]. The extension to systems of conservation laws and higher order methods in space and time is another interesting line of research.

APPENDIX A. PROOF OF THEOREM 6.1

Let us proof Theorem 6.1. We assume that the FE mesh is quasi-uniform in order to reduce technicalities. However, the proof for Lipschitz continuity can be extended to more general meshes. We denote $A = cB$ as $A \eqsim B$ and $A < cB$ as $A \lesssim B$, for any positive constant c that does not depend on the numerical or physical parameters.



(A) Initial conditions.

(B) LED scheme.

(C) Global DMP scheme.

(D) LED DMP nonsmooth stabilization.

FIGURE 11. 3 Body rotation test results using discrete problem (20) and two different artificial diffusions ((31) leading an LED scheme, and (28) with (34) leading a global DMP scheme). Using a 150×150 Q_1 element mesh, and parameters: $q = 25$, $\gamma = 10^{-8}$, $\sigma = |\beta|10^{-10}$, $\varepsilon = 10^{-4}$, and $\Delta t = 10^{-3}$.

From the definition of the nonlinear stabilization in (9), we get

$$|\langle \mathsf{B}(u)u, z\rangle - \langle \mathsf{B}(v)v, z\rangle| \leq \left| \sum_{i\in\mathcal{N}_h} \sum_{j\in\mathcal{N}_h(\Omega_i)} \nu_{ij}(v)\ell(i,j)(u_j - v_j)z_i \right| \tag{52}$$

$$+ \left| \sum_{i\in\mathcal{N}_h} \sum_{j\in\mathcal{N}_h(\Omega_i)} (\nu_{ij}(u) - \nu_{ij}(v))\ell(i,j)u_j z_i \right|. \tag{53}$$

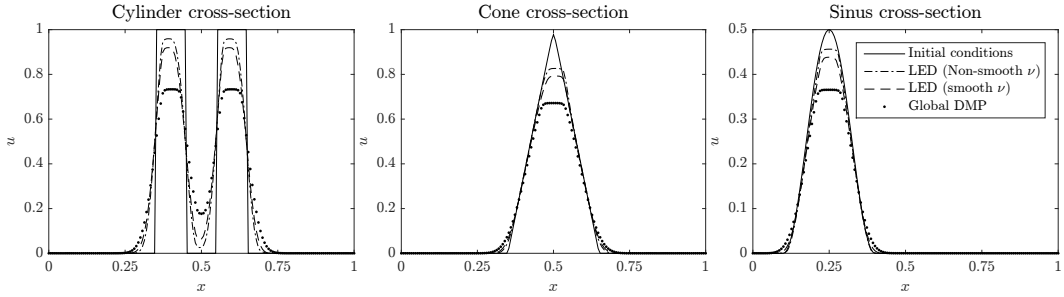
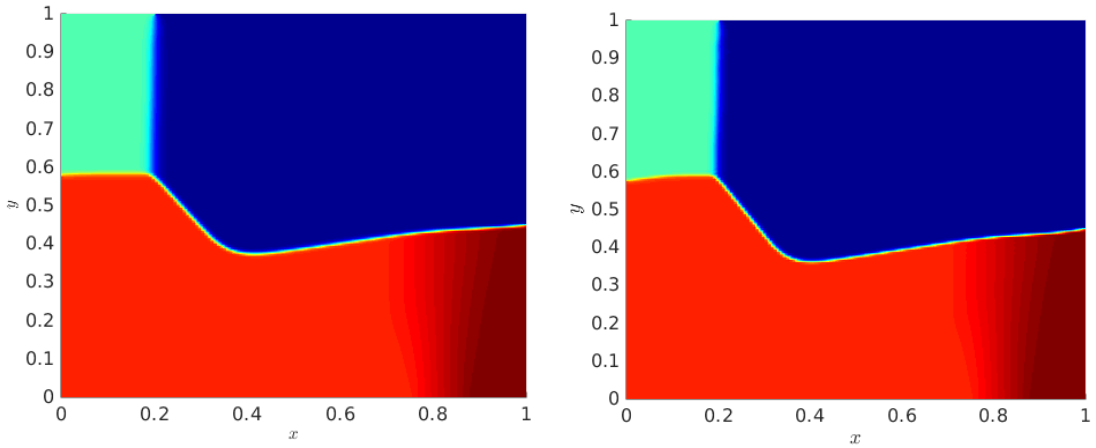


FIGURE 12. Cross-sections of each for the figures rotated in the three body rotation benchmark. The parameters used are $q = 25$, $\gamma = 10^{-8}$, $\sigma = |\beta|10^{-10}$, $\varepsilon = 10^{-4}$, and $\Delta t = 10^{-3}$, in a 150×150 Q_1 element mesh. The discrete problem (20) is used in combination with three different artificial diffusions (31) and (10) leading to a LED scheme, and (28) leading to a global DMP scheme.



(A) Solution for:$q = 1$, $\varepsilon = 10^{-2}$, $\sigma = |\beta|10^{-6}$, and $\gamma = 10^{-8}$.

(B) Solution for: $q = 4$, $\varepsilon = 10^{-3}$, $\sigma = |\beta|10^{-7}$, and $\gamma = 10^{-8}$.

FIGURE 13. Burger's equation solutions at $t = 0.5$ using discrete problem (20) and (10) with (31). Using a 150×150 Q_1 element mesh, $\Delta t = 10^{-2}$, and two sets of parameters q, γ, σ, and ε.

Using the definition of $|\beta|$, the Cauchy-Schwarz inequality, the fact that $\|\varphi_i\| \leq Ch^{d/2}$, and the inverse inequality $\|\nabla v_h\| \lesssim h^{-1}\|v_h\|$ for $v_h \in V_h$ (see [4]), we get:

$$\mathsf{F}_{ij}(w) \leq |\beta|\|\boldsymbol{\nabla}\varphi_i\|_{L^2}\|\varphi_j\|_{L^2} \lesssim h^{d-1}|\beta|, \tag{54}$$

for any $w \in V_h^{\text{adm}}$. Using (54), the first term in the RHS of (52) is bounded as follows:

$$\left|\sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} \nu_{ij}(v)\ell(i,j)(u_j - v_j)z_i\right| \lesssim h^{d-1}|\beta||u-v|_\ell|z|_\ell.$$

The second term is bounded using the Cauchy-Schwarz inequality:

$$\sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} (\nu_{ij}(u) - \nu_{ij}(v))\ell(i,j)u_j z_i \lesssim \left|\sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{1}{2}(\nu_{ij}(u) - \nu_{ij}(v))^2(u_i - u_j)^2\right|^{\frac{1}{2}} \times |z|_\ell. \tag{55}$$

Using (54), we have:

$$\nu_{ij}(u) - \nu_{ij}(v) \tag{56}$$
$$= \max\{\alpha_i(u)\mathsf{F}_{ij}(u), \alpha_j(u)\mathsf{F}_{ji}(u), 0\} - \max\{\alpha_i(v)\mathsf{F}_{ij}(v), \alpha_j(v)\mathsf{F}_{ji}(v), 0\} \tag{57}$$
$$\leq \max\{(\alpha_i(u)\mathsf{F}_{ij}(u) - \alpha_i(v)\mathsf{F}_{ij}(v), \alpha_j(u)\mathsf{F}_{ji}(u) - \alpha_j(v)\mathsf{F}_{ji}(v), 0\} \tag{58}$$
$$\lesssim h^{d-1}|\beta| \max\{|\alpha_i(u) - \alpha_i(v)|, |\alpha_j(u) - \alpha_j(v)|\}. \tag{59}$$

Let us assume that $\sum_{j\in\mathcal{N}_h(\Omega_i)} \{\!\{|\boldsymbol{\nabla} u_h \cdot \boldsymbol{r}_{ij}|\}\!\}_{ij} \neq 0$. (The other case is straightforward.) On one hand, for a non-degenerate FE mesh, we have that $ch \leq \boldsymbol{r}_{ij} \leq Ch$, $j \in \mathcal{N}_h^{\text{sym}}(\Omega_i)$, for positive constants c, C that do not depend on h. Using this fact in the definition of the shock detector (13), we get:

$$\alpha_i(u)^{\frac{1}{q}} = \frac{\left|\sum_{j\in\mathcal{N}_h(\Omega_i)} [\![\boldsymbol{\nabla} u_h]\!]_{ij}\right|}{\sum_{j\in\mathcal{N}_h(\Omega_i)} 2\{\!\{|\boldsymbol{\nabla} u_h \cdot \boldsymbol{r}_{ij}|\}\!\}_{ij}} = \frac{\left|\sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{u_i-u_j}{|\boldsymbol{r}_{ij}|} + \frac{u_i-u_j^{\text{sym}}}{|\boldsymbol{r}_{ij}^{\text{sym}}|}\right|}{\sum_{j\in\mathcal{N}_h(\Omega_i)} \frac{|u_i-u_j|}{|\boldsymbol{r}_{ij}|} + \frac{|u_i-u_j^{\text{sym}}|}{|\boldsymbol{r}_{ij}^{\text{sym}}|}} \tag{60}$$
$$\eqsim \frac{\left|\sum_{j\in\mathcal{N}_h(\Omega_i)} (u_i - u_j) + (u_i - u_j^{\text{sym}})\right|}{\sum_{j\in\mathcal{N}_h(\Omega_i)} |u_i - u_j| + |u_i - u_j^{\text{sym}}|}. \tag{61}$$

Now, we use the following result for two sequences $\{a_i\}_{i=1}^n$ $\{b\}_{i=1}^n$ (see [3] for further details):

$$\frac{|\sum_{i=1}^n a_i|}{\sum_{i=1}^n |a_i|} - \frac{|\sum_{i=1}^n b_i|}{\sum_{i=1}^n |b_i|} = \frac{|\sum_{i=1}^n a_i| - |\sum_{i=1}^n b_i|}{\sum_{i=1}^n |a_i|} + \sum_{i=1}^n |b_i| \left(\frac{1}{\sum_{i=1}^n |a_i|} - \frac{1}{\sum_{i=1}^n |b_i|}\right) \tag{62}$$
$$\leq \frac{|\sum_{i=1}^n a_i - b_i|}{\sum_{i=1}^n |a_i|} + \frac{\sum_{i=1}^n |b_i| - \sum_{i=1}^n |a_i|}{\sum_{i=1}^n |a_i|} \leq \frac{|\sum_{i=1}^n a_i - b_i| + \sum_{i=1}^n |a_i - b_i|}{\sum_{i=1}^n |a_i|} \tag{63}$$
$$\leq 2\frac{\sum_{i=1}^n |a_i - b_i|}{\sum_{i=1}^n |a_i|}. \tag{64}$$

Using simple algebraic manipulation, we have $a^q - b^q = (a-b)\sum_{k=0}^{q-1} a^k b^{q-i} k$ for $q \in \mathbb{N}^+$. For $a, b \in [0,1]$, it leads to $|a^q - b^q| \leq q|a-b|$ (see [3]). This inequality, together with (60) and (64), leads to:

$$\frac{1}{q}|\alpha_i(u) - \alpha_i(v)| \lesssim \frac{\left|\sum_{j\in\mathcal{N}_h(\Omega_i)} ((u-v)i - (u-v)_j) + ((u-v)_i - (u-v)_j^{\text{sym}})\right|}{\sum_{j\in\mathcal{N}_h(\Omega_i)} |u_i - u_j| + |u_i - u_j^{\text{sym}}|}. \tag{65}$$

On the other hand, the bounds

$$|u_i - u_j| \leq \sum_{k\in\mathcal{N}_h(\Omega_i)} |u_i - u_k| \quad \text{and} \quad |u_i - u_j| \leq \sum_{k\in\mathcal{N}_h(\Omega_j)} |u_j - u_k|,$$

(57), and (65), yield

$$(\nu_{ij}(u) - \nu_{ij}(v))(u_i - u_j) \lesssim qh^{d-1}|\beta| \sum_{k\in\mathcal{N}_h^{\text{sym}}(\Omega_i)} |(u-v)_i - (u-v)_k| \tag{66}$$
$$+ qh^{d-1}|\beta| \sum_{k\in\mathcal{N}_h^{\text{sym}}(\Omega_j)} |(u-v)_j - (u-v)_k|. \tag{67}$$

The second term is bounded by combining (55), (66), and the fact that the number of elements surrounding a node is bounded above independently of h:

$$\sum_{i\in\mathcal{N}_h} \sum_{j\in\mathcal{N}_h(\Omega_i)} (\nu_{ij}(u) - \nu_{ij}(v))\ell(i,j)u_j z_i \lesssim qh^{d-1}|\beta||u-v|_\ell |z|_\ell. \tag{68}$$

Next, we have to prove that the nonlinear mass matrix is also Lipschitz continuous. First, we note that

$$\sum_{j\in\mathcal{N}_h(\Omega_i)} (1-\alpha_i(u_h))(\varphi_j,\varphi_i)u_j + \alpha_i(u_h)(1,\varphi_i)u_i \tag{69}$$

$$= \sum_{j\in\mathcal{N}_h(\Omega_i)} (\varphi_j,\varphi_i)u_j + \alpha_i(u_h)(\varphi_j,\varphi_i)(u_i-u_j). \tag{70}$$

Thus

$$\langle \mathsf{M}(u)u, z\rangle - \langle \mathsf{M}(v)v, z\rangle \leq \sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} (\varphi_i,\varphi_j)(u_j-v_j)z_i \tag{71}$$

$$+ \sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} (\varphi_i,\varphi_j)(u_i-u_j)(\alpha_i(u_h)-\alpha_i(v_h))z_i \tag{72}$$

$$+ \sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} (\varphi_i,\varphi_j)((u+v)_i-(u+v)_j)\alpha_i(v_h)z_i. \tag{73}$$

Bounds for the second and third term follow the same lines as above. For the second term, we proceed as in (55), getting:

$$\sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} (\varphi_i,\varphi_j)(u_i-u_j)(\alpha_i(u_h)-\alpha_i(v_h))z_i \tag{74}$$

$$\lesssim \left|\sum_{i\in\mathcal{N}_h}\frac{1}{2}\sum_{j\in\mathcal{N}_h(\Omega_i)} (\varphi_i,\varphi_j)(\alpha_i(u_h)-\alpha_i(v_h))^2(u_i-u_j)^2\right|^{\frac{1}{2}} \times \|z\| \tag{75}$$

$$\lesssim qh^{\frac{d}{2}}|u-v|_\ell\|z\|. \tag{76}$$

where we have used the spectral equivalence of the consistent and lumped mass matrices in the last inequality. The first and third term are easily bounded as

$$\sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} (\varphi_i,\varphi_j)(u_j-v_j)z_i \leq \|u-v\|\|z\|, \tag{77}$$

$$\sum_{i\in\mathcal{N}_h}\sum_{j\in\mathcal{N}_h(\Omega_i)} (\varphi_i,\varphi_j)((u+v)_i-(u+v)_j)\alpha_i(v_h)z_i \leq qh^{\frac{d}{2}}|u-v|_\ell\|z\|. \tag{78}$$

It proves the theorem.

References

[1] S. Badia and A. Hierro, *On Monotonicity-Preserving Stabilized Finite Element Approximations of Transport Problems*, SIAM Journal on Scientific Computing, 36 (2014), pp. A2673–A2697.

[2] G. Barrenechea, V. John, and P. Knobloch, *Analysis of Algebraic Flux Correction Schemes*, SIAM Journal on Numerical Analysis, 54 (2016), pp. 2427–2451.

[3] R. G. Barrenechea, E. Burman, and F. Karakatsani, *Edge-based nonlinear diffusion for finite element approximation of convection-diffusion equations and its relation to algebraic flux-correction schemes*, (2016).

[4] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, third ed., 2008.

[5] E. Burman, *On nonlinear artificial viscosity, discrete maximum principle and hyperbolic conservation laws*, BIT Numerical Mathematics, 47 (2007), pp. 715–733.

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

[7] E. Burman and A. Ern, *Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion-reaction equation*, Computer Methods in Applied Mechanics and Engineering, 191 (2002), pp. 3833–3855.

[8] ——, *Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence*, Mathematics of Computation, 74 (2005), pp. 1637–1652.
[9] B. Cockburn and C.-W. Shu, *Runge–Kutta Discontinuous Galerkin Methods for Convection-Dominated Problems*, Journal of Scientific Computing, 16 (2001), pp. 173–261.
[10] J. Donea and A. Huerta, *Finite Element Methods for Flow Problems*, Finite Element Methods for Flow Problems, John Wiley & Sons, 2003.
[11] S. K. Godunov, *A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics*, Mat. Sb. (NS), 47 (1959), pp. 271–306.
[12] J. Guermond, M. Nazarov, B. Popov, and Y. Yang, *A Second-Order Maximum Principle Preserving Lagrange Finite Element Technique for Nonlinear Scalar Conservation Equations*, SIAM Journal on Numerical Analysis, 52 (2014), pp. 2163–2182.
[13] J.-L. Guermond and M. Nazarov, *A maximum-principle preserving $\mathcal{C}^0$ finite element method for scalar conservation equations*, Computer Methods in Applied Mechanics and Engineering, 272 (2014), pp. 198–213.
[14] J.-L. Guermond and R. Pasquetti, *A correction technique for the dispersive effects of mass lumping for transport problems*, Computer Methods in Applied Mechanics and Engineering, 253 (2013), pp. 186–198.
[15] A. Hierro, S. Badia, and P. Kus, *Shock capturing techniques for hp-adaptive finite elements*, Computer Methods in Applied Mechanics and Engineering, (in press).
[16] T. Hughes and A. Brooks, *A multi-dimensioal upwind scheme with no crosswind diffusion.*, in: T.J.R. Hughes ed. Finite Element Methods for Convection Dominated Flows, (ASME, New York), 34 (1979), pp. 19–35.
[17] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert, *A new finite element formulation for computational fluid dynamics: VIII. The galerkin/least-squares method for advective-diffusive equations*, Computer Methods in Applied Mechanics and Engineering, 73 (1989), pp. 173–189.
[18] T. J. R. Hughes, M. Mallet, and M. Akira, *A new finite element formulation for computational fluid dynamics: II. Beyond SUPG*, Computer Methods in Applied Mechanics and Engineering, 54 (1986), pp. 341–355.
[19] D. I. Ketcheson, C. B. Macdonald, and S. Gottlieb, *Optimal implicit strong stability preserving Runge–Kutta methods*, Applied Numerical Mathematics, 59 (2009), pp. 373–392.
[20] A. Kritz and D. Keyes, *Fusion Simulation Project Workshop Report*, Journal of Fusion Energy, 28 (2008), pp. 1–59.
[21] S. N. Kružkov, *First order quasilinear equations in several independent variables*, Mathematics of the USSR-Sbornik, 10 (1970), pp. 217–243.
[22] D. Kuzmin, *Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes*, Journal of Computational and Applied Mathematics, 236 (2012), pp. 2317–2337.
[23] D. Kuzmin and J. N. Shadid, *A new approach to enforcing discrete maximum principles in continuous Galerkin methods for convection-dominated transport equations*, Journal of Computational Physics, (2015).
[24] D. Kuzmin and J. N. Shadid, *Gradient-based nodal limiters for artificial diffusion operators in finite element schemes for transport equations*, (2016).
[25] D. Kuzmin, M. J. Shashkov, and D. Svyatskiy, *A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems*, Journal of Computational Physics, 228 (2009), pp. 3448–3463.
[26] D. Kuzmin and S. Turek, *Flux Correction Tools for Finite Elements*, Journal of Computational Physics, 175 (2002), pp. 525–558.
[27] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, 1 ed., 2002.
[28] L. R. Scott and S. Zhang, *Finite Element Interpolation of Nonsmooth Functions Satisfying Boundary Conditions*, Mathematics of Computation, 54 (1990), pp. 483–493. ArticleType: research-article / Full publication date: Apr., 1990 / Copyright © 1990 American Mathematical Society.

[29] J. Xu and L. Zikatanov, *A Monotone Finite Element Scheme for Convection-Diffusion Equations*, Mathematics of Computation, 68 (1999), pp. 1429–1446.