DIFFERENTIABLE MONOTONICITY-PRESERVING SCHEMES FOR DISCONTINUOUS GALERKIN METHODS ON ARBITRARY MESHES

SANTIAGO BADIA†‡, JESÚS BONILLA†‡, AND ALBA HIERRO†‡

ABSTRACT. This work is devoted to the design of interior penalty discontinuous Galerkin (dG) schemes that preserve maximum principles at the discrete level for the steady transport and convection-diffusion problems and the respective transient problems with implicit time integration. Monotonic schemes that combine explicit time stepping with dG space discretization are very common, but the design of such schemes for implicit time stepping is rare, and it had only been attained so far for 1D problems. The proposed scheme is based on a piecewise linear dG discretization supplemented with an artificial diffusion that linearly depends on a shock detector that identifies the troublesome areas. In order to define the new shock detector, we have introduced the concept of discrete local extrema. The diffusion operator is a graph-Laplacian, instead of the more common finite element discretization of the Laplacian operator, which is essential to keep monotonicity on general meshes and in multi-dimension. The resulting nonlinear stabilization is non-smooth and nonlinear solvers can fail to converge. As a result, we propose a smoothed (twice differentiable) version of the nonlinear stabilization, which allows us to use Newton with line search nonlinear solvers and dramatically improve nonlinear convergence. A theoretical numerical analysis of the proposed schemes show that they satisfy the desired monotonicity properties. Further, the resulting operator is Lipschitz continuous and there exists at least one solution of the discrete problem, even in the non-smooth version. We provide a set of numerical results to support our findings.

Keywords: Finite elements, discrete maximum principle, monotonicity, shock capturing, discontinuous Galerkin, local extrema diminishing

CONTENTS

1. Introduction 2
2. The Convection-diffusion problem and its discretization 3
2.1. Notation 3
2.2. Weak form and interior penalty dG approximation 4
2.3. Implicit time integration 5
3. Monotonicity Properties 5
4. The DMP-preserving artificial diffusion scheme 7
5. Lipschitz continuity and existence of solutions 12
6. Smoothing the shock detector 14
6.1. Parameters fine-tuning 15
7. Numerical experiments 18
7.1. Convergence to a smooth solution 18
7.2. DMP-preservation 20
7.3. Three body rotation 21
8. Conclusions 23
Acknowledgments 23
References 25

† Centre Internacional de Mètodes Numèrics en Enginyeria (CIMNE), Parc Mediterrani de la Tecnologia, UPC, Esteve Terradas 5, 08860 Castelldefels, Spain ({sbadia,jbonilla,ahierro}@cimne.upc.edu).
‡ Universitat Politècnica de Catalunya, Jordi Girona 1-3, Edifici C1, 08034 Barcelona, Spain.
1. Introduction

The transport problem is one of many problems that might satisfy a maximum principle (MP) or a positivity property. However, its numerical discretization may violate these properties at the discrete level. These violations arise in the form of local spurious oscillations near sharp layers of the solution. Such oscillations break the MP of the continuous problem. For steady problems with no source term, the MP implies that the extrema of the solution are on the boundary of the domain; they are bounded by the boundary and the initial solution extrema in the transient case.

Many authors have focused on developing accurate schemes that inherit the MP at the discrete level, i.e. discrete maximum principle (DMP) preserving schemes. To this end, several approaches have been used. In the case of explicit time integration combined with finite volumes or discontinuous Galerkin (dG) methods, the schemes are usually based on either slope or flux limiters, or special reconstruction algorithms. These methods are widely present in literature and already well understood (see, e.g., [24]).

For implicit time integration and continuous Galerkin (cG) finite element space discretization, methods attaining DMPs are not as well understood as the previous ones. However, several schemes have been developed to date. In this case, most of the approaches are based on adding an artificial diffusion operator. Then, in order to maintain high-order convergence rates in smooth regions, this operator is scaled such that it vanishes in smooth regions and it is active in the vicinity of sharp layers. Depending on how this activation is controlled, one may distinguish among residual-based, entropy-based, and fluctuation-based schemes. The first DMP-preserving schemes were residual-based (see e.g. [9, 25]). Afterwards, fluctuation based schemes were developed [3,10–13]. These schemes are based on computing a priori an artificial diffusion that ensure DMP preservation. The artificial diffusion is activated based on a so-called shock detector, usually based on the unknown gradient jumps across elements. Lately, Guermond and co-workers have proposed a similar approach for hyperbolic problems but using an alternative detector based on the entropy production [16, 17]. The more recent fluctuation-based schemes compute the amount of diffusion required to preserve the DMP in a way that resembles Algebraic Flux Correction (AFC) techniques [2, 5, 22, 23]. The reader might refer to [21] and the references therein for more insights about AFC.

In the case at hand, the dG space discretization of steady problems and transient problems, the situation is much less understood. An attempt to develop implicit DMP-preserving dG schemes has been proposed in [4]. but even though a DMP enjoying artificial diffusion method can be constructed for the 1D problem, the extension to the multi-dimensional case fails to enjoy such property. The objective of this work is to design a multidimensional DMP-preserving dG method on arbitrary meshes for both implicit time integration and steady problems. Furthermore, we propose a linearity preserving and differentiable method. This latter property is particularly important for improving the convergence of the nonlinear solver, as shown in [2].

In order to do so, we propose a stabilization method based on the following four key ingredients:

1. A shock detector that only activates the artificial diffusion in regions around shock. As previously said, a shock detector restricts the application of the stabilization to regions where the solution presents shocks or sharp layers, and is the key ingredient to obtain a high-order stabilization method;

2. The amount of diffusion added to ensure the DMP. We motivate it using similar ideas behind the AFC low-order scheme construction (see [19, 21]);

3. The discrete diffusion operator in order to keep the DMP on arbitrary meshes. Guermond and co-workers [16, 17] have proposed to use graph-theoretic artificial diffusion operators, instead of the classical PDE-based ones. This strategy has already been used in [2, 22, 23];

4. For transient problems, a perturbation of the mass matrix is required to obtain a local extremum diminishing (LED) scheme.

This work is structured as follows. In Sect. 2, we introduce the problem to solve, the notation, and the discretization of the problem in space using the interior penalty dG method. Then, in Sect. 3, we state a novel definition of the DMP property for dG methods, by introducing the concept of discrete
local extrema. In Sect. 4, we propose a scheme that fulfills such property. Lipschitz continuity and existence of solutions are proved in Sect 5. A discussion about the importance of smoothing the computation of the shock capturing terms and some tests to choose the optimal values of the smoothing parameters are developed in Sect. 6. Finally, numerical experiments show the performance of the method in Sect. 7, and some conclusions are drawn in Sect. 8.

2. The Convection-diffusion problem and its discretization

We consider a transient convection-diffusion problem with Dirichlet boundary conditions:

\[
\begin{align*}
\partial_t u + \nabla \cdot (\beta u) - \nabla \cdot (\mu \nabla u) &= g & \text{in } \Omega \times (0, T], \\
u(x, t) &= \overline{u}(x, t) & \text{on } \partial \Omega \times (0, T], \\
u(x, 0) &= u_0(x) & x \in \Omega.
\end{align*}
\]  

(1)

The domain \( \Omega \) is an open, bounded, connected subset of \( \mathbb{R}^d \) with a Lipschitz boundary \( \partial \Omega \), where \( d \) is the space dimension, \( \beta = \beta(x) \) is the convective velocity, which is assumed to be divergence-free, and \( \mu \geq 0 \) is a constant diffusion. Even though we have considered Dirichlet boundary conditions in the statement of problem (1), i.e., \( \partial \Omega = \partial \Omega_D \), Neumann boundary conditions can also be considered straightforwardly. In the case of pure convection \( (\mu = 0) \), boundary conditions are only imposed on the inflow boundary \( \partial \Omega^+ = \{ x \in \partial \Omega : \beta \cdot n_{\partial \Omega} < 0 \} \), where \( n_{\partial \Omega} \) is the outward-pointing unit normal. Further, we define the outflow boundary as \( \partial \Omega^+ = \partial \Omega \setminus \partial \Omega^- \). Below, we also consider the steady case, by eliminating the time derivative term.

2.1. Notation. Let \( T_h = \{ K \} \) be a partition of \( \Omega \) formed by elements \( K \) of characteristic length \( h_K \). For quasi-uniform meshes, we define a global characteristic length of the mesh \( h \). We denote by \( x_i \) the coordinates of vertex \( i \) and by \( \mathcal{V}_h \) the set of vertices of the partition. We also define \( \mathcal{V}_h(K) = \{ i \in \mathcal{V}_h : x_i \in K \} \).

Given the mesh \( T_h \), the non-empty intersection \( F = \partial K \cap \partial K' \) of two neighboring elements \( K, K' \in T_h \) is called an interior facet of \( T_h \) if it is a subdomain of dimension \( d - 1 \). The set of all the interior facets is denoted by \( E_h^i \). On the other hand, the non-empty intersection \( F = \partial K \cap \partial \Omega \) of an element \( K \in T_h \) on the boundary of the domain is called an inflow boundary facet (analogously for \( \partial \Omega^+ \) and outflow boundary facets). The set of inflow boundary facets is denoted by \( E_h^i \), and the set of outflow boundary facets is denoted by \( E_h^o \). We assume that the finite element partition is conforming with the inflow and outflow boundaries. The set of all the facets is denoted by \( E_h = E_h^i \cup E_h^o \).

For any facet \( F \in E_h \), we represent with \( K_F^+ \) and \( K_F^- \) the only two neighboring elements such that \( \partial K_F^+ \cap \partial K_F^- = F \). In addition, we call \( n_F^+ \) and \( n_F^- \) the unitary normal to facet \( F \) outside \( K_F^+ \) and \( K_F^- \), respectively. Given a facet \( F \), we can also define the characteristic facet length \( h_F \).

On tetrahedral (or triangular) meshes, the discrete space considered henceforth is the discontinuous space of piecewise linear functions \( V_h = \{ v_h : v_h|_K \in P_1(K) \forall K \in T_h \} \), where \( P_1(K) \) is the space of linear polynomials in \( K \). For hexahedral (or quadrilateral) meshes, \( V_h = \{ v_h : v_h|_K \in P_2(K) \forall K \in T_h \} \), where \( P_2(K) \) is the tensor product space of piecewise linear 1D polynomials. In addition, we represent the space of traces of \( V_h \) on \( \partial \Omega \) as \( V_h|_{\partial \Omega} \).

In order to define dG spaces, we use the nodal set as the Cartesian product of element vertices, i.e., \( \mathcal{N}_h = \prod_{K \in T_h} \mathcal{V}_h(K) \). Thus, every node \( a \in \mathcal{N}_h \) can also be represented as a pair \( (i, K) \), with \( K \in T_h \) and \( i \in \mathcal{V}_h(K) \). Therefore, more than one interior node might be placed at the same coordinates. Indeed, if \( n \) elements have a vertex on \( x_i \), there will be \( n \) nodes at \( x_i \), each one with its own degree of freedom.

Given the node \( a \in \mathcal{N}_h \), its coordinates are represented with \( x_a \), \( \Omega_a = \{ K \in T_h : x_a \in K \} \) is its support, and \( N_h(a) = \{ b \in \mathcal{N}_h : x_b \in \Omega_a \} \) is the set of nodes connected to \( a \). Notice that \( a \) itself is included in \( \mathcal{N}_h(a) \). We define the set of boundary nodes \( \mathcal{N}_h^{\partial} = \{ a \in \mathcal{N}_h : x_a \in \partial \Omega \} \), and \( \mathcal{N}_h^{\partial, b} = \mathcal{N}_h(a) \cap \mathcal{N}_h^{\partial} \).

The functions \( v_h \in V_h \) can be expressed as a linear combination of the basis \( \{ \varphi_a \}_{a \in \mathcal{N}_h} \), where \( \varphi_a \) corresponds to the shape function of node \( a \). It is defined as follows. Given \( a \in \mathcal{N}_h \) and its corresponding vertex-element pair \( (i, K) \), we define \( \varphi_a \) as the elementwise (b)linear function such that \( \varphi_a(x_a)|_K = 1 \) and \( \varphi_a(x_b)|_K = 0 \) for \( b \neq a \), and \( \varphi_a|_{K'} \equiv 0 \) for \( K' \neq K \). Any function \( v_h \in V_h \) is
double-valued on $E_h^0$ and single-valued on $\partial \Omega$. Thus, $v_h \in V_h$ can be expressed as $v_h = \sum_{a \in N_h} v_a \varphi_a$. Moreover we consider $v_h^K$ as the restriction of $v_h$ into $K$.

Given $v_h \in V_h$, we can define the common concepts of average $\langle \cdot \rangle$ and jump $[\cdot]$ on an interior point $x$ of a facet $F \in E_h^0$ as follows:

\[
\langle v_h \rangle (x) = \frac{1}{2} \left( v_h^+ (x) + v_h^- (x) \right), \quad [v_h] (x) = v_h^+ (x) n_F^+ + v_h^- (x) n_F^-,
\]

where $n_F^+$ (resp. $n_F^-$) is the outward normal with respect to $K^+$ (resp. $K^-$) on $F$; we use $n_F$ on boundary facets and in places where the sign is not relevant. On boundary facet points $x \in F, F \subset \partial \Omega$, we define $\langle v_h \rangle (x) = v_h^K (x), [v_h] (x) = v_h^K (x) n_F^K (x)$.

We will use standard notation for Sobolev spaces (see, e.g., [7]). In particular, the $L^2(\omega)$ scalar product will be denoted by $(\cdot, \cdot)_\omega$ for some $\omega \subset \Omega$, but the domain subscript is omitted for $\omega = \Omega$. The $L^2(\Omega)$ norm is denoted by $\| \cdot \|$.

### 2.2. Weak form and interior penalty dG approximation.

The stabilized dG bilinear form for the transport problem proposed in [8] combined with the interior penalty (IP) method for the viscosity term reads as:

Find $u_h \in V_h$ such that

\[
(\partial_t u_h, v_h) + K_h(u_h, v_h) = G_h(v_h) + B_h(\overline{u}_h; v_h) \quad \forall v_h \in V_h, \tag{2}
\]

with

\[
K_h(u_h, v_h) \doteq \sum_{K \in T_h} \int_K (\mu \nabla u_h \cdot \nabla v_h - u_h \beta \cdot \nabla v_h) + \sum_{F \in \mathcal{E}_h^+} \int_F \mu (-[u_h] \cdot [\nabla v_h] - [\nabla u_h] \cdot [v_h]) + c^{ip} \mu n_F \cdot [v_h]
\]

\[
+ \sum_{F \in \mathcal{E}_h^+ \cup \mathcal{E}_h^-} \left( \int_F |\beta \cdot n_F| \|u_h\|_F \right) + \sum_{F \in \mathcal{E}_h^+ \cup \mathcal{E}_h^-} \left( \int_F c^{ip} \mu h^{-1} w_h v_h \right),
\]

where the right hand side (RHS) includes the terms corresponding to the source

\[
G_h(v_h) \doteq \sum_{K \in T_h} (g, v_h)_K,
\]

and weak boundary conditions $B_h(\overline{u}_h; v_h)$,

\[
B_h(w_h; v_h) \doteq - \sum_{F \in \mathcal{E}_h^+} \int_F \beta \cdot n_{\partial \Omega} w_h v_h - \sum_{F \in \mathcal{E}_h^+ \cup \mathcal{E}_h^-} \int_F \mu w_h \langle \nabla v_h \rangle \cdot n_{\partial \Omega} + \sum_{F \in \mathcal{E}_h^+ \cup \mathcal{E}_h^-} \int_F c^{ip} \mu h^{-1} w_h v_h.
\]

The parameter $c^{ip}$ is a constant set to 10, as suggested in [4]. The projection $\overline{u}_h \in V_h|_{\partial \Omega}$ is the facetwise linear polynomial function obtained, e.g., by the nodal interpolation of the given Dirichlet boundary data $\overline{u}$ on the nodes of the boundary $N_h^{\partial}$, i.e., $u_h = \sum_{a \in N_h^{\partial}} \varphi_a \overline{u}(x_a)$. When the nodal projector is not well-defined, other projections that preserve the DMP can be also used, e.g., the Scott-Zhang projection [26]. Notice that with this definition $\overline{u}_h$ is bounded by the maximum and minimum values of the function $\overline{u}$. Moreover, the semi-discrete problem (2) can be rewritten in algebraic form as

\[
M \partial_t u_h + K u_h = G + B \overline{u}_h, \tag{3}
\]

where $M_{ab} \doteq (\varphi_b, \varphi_a)$ and $K_{ab} \doteq K_h(\varphi_b, \varphi_a)$, for $a, b \in N_h$, $G_a \doteq G_h(\varphi_a)$, for $a \in N_h$, and $B_{ab} \doteq B_h(\varphi_b; \varphi_a)$, for $a \in N_h, b \in N_h^0$. 

Register for free at https://www.scipedia.com to download the version without the watermark
2.3. Implicit time integration. We consider the time discretization of (2) using the method of lines. In doing so, we are interested in schemes that ensure the DMP as the discrete solution evolves in time. This kind of methods are also known as local extremum diminishing (LED). In particular, we will use the \( \theta \)-method, even though the generalization of the following results to other schemes that preserve monotonicity properties is straightforward. We consider a partition of \( (0,T) \) into \( N_t \) time steps with equal time step length \( \Delta t = \frac{T}{N_t} \) in such a way that \( t_n = n \Delta t, \ n = 0, \ldots, N_t \). The problem will be solved by computing an approximation of \( u \) in each of those time steps \( u^n_h \approx u(\cdot, t_n) \). The discretization of (1) by means of the \( \theta \)-method reads: Find \( u^{n+1}_h \in V_h \) such that
\[
\frac{1}{\Delta t} (u^{n+1}_h - u^n_h, v_h) + K_h(\theta u^{n+1}_h + (1 - \theta) u^n_h, v_h) = G_h(v_h) + B_h(\pi_h; v_h) \quad \forall v_h \in V_h. \tag{4}
\]
We use a projection of the actual initial condition \( u_t = 0 \) as the initial discrete solution \( u^0_h \), such that it inherits the DMP. The value of \( \theta \) is to be chosen in the interval \([0, 1]\). Some common values are \( \theta = 0 \), which leads to the explicit forward Euler scheme, \( \theta = 0.5 \) for the Crank-Nicolson scheme, and \( \theta = 1 \), leading to the Backward-Euler (BE) scheme. Each of these methods has different features. In particular, in order to obtain an unconditionally LED scheme, it is necessary to use \( \theta = 1 \). Using BE, the discrete problem in compact form reads
\[
M\delta_t u_h + K_{h_{n+1}} = G + B\overline{u}_h,
\]
where \( \delta_t u_h = \Delta t^{-1} (u^{n+1}_h - u^n_h) \). Other choices of \( \theta \) lead to LED schemes under a CFL-like condition. This is specially important in the case of the Crank-Nicolson (CN) method, which is second-order accurate and non-dissipative. The requirements to obtain monotonicity-preserving schemes in these cases are commented in Sect. 4.

3. Monotonicity Properties

In this section we introduce the desired properties that we want our discrete problem to fulfill. The use of local extrema in dG is too restrictive for our purposes. In dG, due to the existence of jumps, local extrema that do not harm the MP may appear, e.g. a positive jump between two elements with negative gradients. Thus, we consider the concept of discrete local extrema, which is defined on nodes, and means that the nodal value is extremum in the support of the node. A discrete local extremum is a function that is locally constant, but not necessarily constant at the node. This is more restrictive than our purposes.

**Definition 3.1** (Local Discrete Extremum). The function \( u_h \in V_h \) has a local discrete maximum (resp. minimum) on a node \( a \in N_h \) if \( u_a \geq u_b(x) \) (resp. \( u_a \leq u_b(x) \)) \( \forall x \in \Omega_a \), and also \( u_a \geq \pi(x) \) (resp. \( u_a \leq \pi(x) \)) \( \forall x \in \partial\Omega_a \cap \partial\Omega \).

Therefore, the DMP can be defined as follows.

**Definition 3.2** (DMP). For steady problems a solution \( u_h \in V_h \) satisfies the local DMP if for every \( a \in N_h \), we have:
\[
\begin{align*}
\min u_{a,\text{min}} & \leq u_a \leq \max u_{a,\text{max}}, & \text{where } u_{a,\text{max}} & = \max \left\{ \max_{b \in N_h(a) \setminus \{a\}} u_b, \max_{x \in \partial\Omega_a \cap \partial\Omega_D} \pi_h(x) \right\}, \\
\text{and } u_{a,\text{min}} & = \min \left\{ \min_{b \in N_h(a) \setminus \{a\}} u_b, \min_{x \in \partial\Omega_a \cap \partial\Omega_D} \pi_h(x) \right\},
\end{align*}
\]
where \( \pi_h \in V_h|_{\partial\Omega} \) is the finite element interpolation of the boundary conditions on the Dirichlet boundary \( \partial\Omega_D \). In the case of transient problems, \( u_{a,\text{max}} \) and \( u_{a,\text{min}} \) are defined as
\[
\begin{align*}
\max u_{a,\text{max}} & = \max \left\{ \max_{\{a\} \cup \partial\Omega_D \times (0,T)} \pi(x), \max_{x \in \Omega} u_{0,h}(x) \right\}, \\
\text{and } u_{a,\text{min}} & = \min \left\{ \min_{\{a\} \cup \partial\Omega_D \times (0,T)} \pi(x), \min_{x \in \Omega} u_{0,h}(x) \right\},
\end{align*}
\]
where \( u_{0,h} \in V_h \) is the finite element projection of the initial condition \( u_0 \).
A scheme such that its solutions satisfy the DMP is called DMP-preserving. Instead, for transient problems, we define LED schemes.

**Definition 3.3 (LED).** A method is called LED if for $g = 0$ and any time in $t \in (0, T]$, the solution $u_h(t) \in V_h$ satisfies

$$d_t u_a \leq 0 \text{ if } u_a \text{ is a maximum and } d_t u_a \geq 0 \text{ if } u_a \text{ is a minimum.}$$

For time-discrete methods, the same definition applies, replacing $d_t$ by the time derivative discrete approximation $\delta_t$.

Let us assume now that we have system (3) plus a Lipschitz continuous nonlinear diffusion term (nonlinear stabilization):

$$\tilde{M}(u_h, \tilde{\varpi}_h) \partial_t u_h + \tilde{K}(u_h, \tilde{\varpi}_h) u_h = G + \tilde{B}(u_h, \tilde{\varpi}_h) \tilde{\varpi}_h. \tag{5}$$

The superscript, e.g., in $\tilde{K}$, denotes the fact that the operator $\tilde{K}$ is equal to $K$ plus stabilization terms. We have written, e.g., $\tilde{K}(u_h, \tilde{\varpi}_h)$, to explicitly denote the fact that the entries of the matrix $\tilde{K}$ are potentially nonlinear with respect to $u_h$ and $\tilde{\varpi}_h$. Problem (5) is LED under the following requirements.

**Theorem 3.4 (LED).** The semi-discrete problem (5) is LED (as defined in Def. 3.3) if $g = 0$ and for every $a \in N_h$ such that $u_a$ is a local extremum, it holds:

$$\tilde{M}_{ab}(u_h, \tilde{\varpi}_h) \delta_{ab} m_a \geq 0, \tag{6a}$$

$$\tilde{K}_{ab}(u_h, \tilde{\varpi}_h) \leq 0, \forall b \in N_h \setminus \{a\}, \tag{6b}$$

$$\tilde{B}_{ab}(u_h, \tilde{\varpi}_h) = 0, \tag{6c}$$

where $m_a = \int_{\Omega} \varphi_a d\Omega$, $\delta_{ab}$ is the Kronecker delta. Further, for $g \leq 0$ (resp. $g \geq 0$) in $\Omega$, solutions of (5) satisfy the DMP property in Def. 3.2. Moreover, the discrete problem (5) is positivity-preserving for $g \geq 0$ and $u_0 > 0$.

**Proof.** Assume $u_a$ is a discrete maximum. From the conditions in (6) and particularizing equation (5)

$$G_a = m_a d_t u_a + \sum_{b \in N_h(a)} \tilde{K}_{ab}(u_h, \tilde{\varpi}_h) u_b - \sum_{b \in N^g_h(a)} \tilde{B}_{ab}(u_h, \tilde{\varpi}_h) \tilde{\varpi}_b$$

$$\geq m_a d_t u_a + \left( \sum_{b \in N_h(a)} \tilde{K}_{ab}(u_h, \tilde{\varpi}_h) - \sum_{b \in N^g_h(a)} \tilde{B}_{ab}(u_h, \tilde{\varpi}_h) \right) u_a = m_a d_t u_a.$$ 

Therefore, $d_t u_a \leq G_a = 0$. Proceeding analogously for a minimum we can prove that the method is LED. The proof is equivalent 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 degree of freedom $a$ in which the minimum value is attained. Using the previous result for a minimum at the discrete level, we have that $\delta_t u_a \geq 0$ and thus $u_m^a \geq u_m^{a-1}$. It leads to a contradiction, since $u_a^{m-1} \geq 0$. Hence, the solution must remain positive.

**Corollary 3.5.** If the problem in (5) is discretized in time with BE and it meets the conditions in Th. 3.4, then it leads to solutions that satisfy the local DMP in Def. 3.2 at every time $t^n$, for $n = 1, \ldots N^t$.

**Proof.** By the LED property we know that a discrete maximum (resp. minimum) will be bounded above (resp. below) by the solution at the previous time step. Proceeding by induction, the solution will be bounded by the initial condition $u_0^a$ and the boundary conditions imposed at any previous time step.

Following [14, Th. 1], we can prove that the steady counterpart of problem (5) is DMP-preserving.
Theorem 3.6 (DMP). A steady solution of the semi-discrete problem (5) satisfies the DMP in Def. 3.2 if \( g = 0 \) in \( \Omega \) and, for every degree of freedom \( a \in N_h \) such that \( u_a \) is a local discrete extremum, conditions (6b)-(6c) hold.

Proof. Assume \( u_a \) is a discrete maximum, then the steady counterpart of problem (5) reads

\[
\sum_{b \in N_h(a)} \tilde{K}_{ab}(u_h, \overline{\nu}_h)u_b - \sum_{b \in N_h^b(a)} \tilde{B}_{ab}(u_h, \overline{\nu}_h)\overline{u}_b = 0,
\]

Therefore, \( u_a \) can be computed as

\[
u_a \delta = \sum_{b \in N_h^b(a)} \tilde{B}_{ab}(u_h, \overline{\nu}_h)\overline{u}_b - \sum_{b \in N_h^b(a)} \tilde{K}_{ab}(u_h, \overline{\nu}_h)u_b.
\]

From conditions (6b)-(6c), the coefficients that multiply \( u_b \) and \( \overline{u}_b \) are in \([0, 1]\), and the sum of all these coefficients add up to one. Therefore, \( u_a \) is a convex combination of its neighbors (including boundary conditions \( \overline{u}_b \)). Since \( u_a \) is a maximum and a convex combination of its neighbors, then \( u_b = u_a \) for some \( b \in N_h(a) \). Further, it can also be proved that \( u_a \) is a convex combination of all its neighbors but \( u_b \), and vice versa \( u_b \) is a convex combination of all its neighbors but \( u_a \). Hence, by induction, we know that extremum at any degree of freedom are bounded by the boundary conditions. Thus, the DMP is satisfied.

4. The DMP-preserving artificial diffusion scheme

In the previous section, we have stated the requirements to be fulfilled by our discrete scheme to be DMP-preserving and [29]. In this section, we build a nonlinear stabilization of the dG formulation (2) that satisfies all these conditions. The nonlinear stabilization will rely on an artificial graph-viscosity term. The graph-viscosity is supplemented with a shock detector, in order to obtain higher than linear convergence on smooth regions. Moreover, for transient methods we make use of the shock detector in order to perform the mass matrix lumping only where is required, which allows us to minimize the phase error of the method.

Let us start by defining the graph-viscosity \( \nu_{ab} \). For \( a \in N_h \) and \( b \in N_h^b(a) \) we define

\[
\nu_{ab} = \max\{\alpha_a R_a(\varphi_a, \varphi_b), 0\},
\]

Clearly, this viscosity is only non-zero when \( a \in N_h^b \). Next, for \( a \in N_h \) and \( b \in N_h^b(a) \), we define

\[
\nu_{ab} = \left\{ \begin{array}{ll}
\max\{\alpha_a K_{ab}(\varphi_a, \varphi_b), 0, \alpha_b K_{ba}(\varphi_b, \varphi_a)\} & b \neq a, \\
\sum_{b \in N_h(a) \neq a} \nu_{ab} + \sum_{b \in N_h^b(a)} \nu_{ab} & \text{otherwise},
\end{array} \right.
\]

where \( \alpha_a \) is a parameter that enjoys the following property.

Definition 4.1. Given \( a \in N_h \), we say that \( \alpha_a : V_h \to \mathbb{R} \) enjoys the shock detector property if it is such that \( \alpha_a(u_h, \overline{\nu}_h) \in [0, 1] \) \( \forall u_h \in \nabla_h \) and \( \alpha_a(u_h, \overline{\nu}_h) = 1 \) if \( u_h \) has a local discrete extremum on \( x_a \).

Next, we design a shock detector that satisfies this property. Given \( a \in N_h \) and \( b \in N_h(a) \) with \( x_b \neq x_a \), we define \( x_{ab}^{\text{sym}} \) as the intersection between \( \partial \Omega_a \) and the line that passes through \( x_b \) and \( x_a \), and it is not \( x_b \). Moreover, we define \( r_{ab} = x_b - x_a \), \( r_{ab}^{\text{sym}} = x_{ab}^{\text{sym}} - x_a \), and \( u_{ab}^{\text{sym}} = u_h(x_{ab}^{\text{sym}}) \) (see Fig. 1). Further, we denote by \( \hat{r}_{ab} \) the unit vector of \( r_{ab} \), and by \( h_a \) a characteristic length of \( \Omega_a \). Then, we define the jump and the mean of the unknown gradients as

\[
\left[ \nabla u_h \right]_{ab} = \left\{ \begin{array}{ll}
u_a u_a + \frac{h_a}{|r_{ab}|} u_b - u_a & \text{if } x_a = x_b, \\
\frac{h_a}{|r_{ab}|} u_b - u_a & \text{otherwise,}
\end{array} \right.
\]

\[
\left\langle \left[ \nabla u_h \cdot \hat{r}_{ab} \right] \right\rangle_{ab} = \left\{ \begin{array}{ll}
u_a u_a - u_a & \text{if } x_a = x_b, \\
\frac{h_a}{2} \left( \frac{u_b - u_a}{|r_{ab}|} + \frac{u_{ab}^{\text{sym}} - u_a}{|r_{ab}^{\text{sym}}|} \right) & \text{otherwise.}
\end{array} \right.
\]
Remark 4.2. These definitions may imply $x_{ab}^{\text{sym}} = x_a$ on some boundaries. In these cases, the value at the symmetric point of $x_a$ with respect to $x_b$ takes an extrapolation of the boundary condition value such that the method is linearly preserving, i.e., $u_{ab}^{\text{sym}} = \pi_a + (u_b - u_a) \text{sign}((\pi_a - u_a)(u_b - u_a))$, and the value $|r_{ab}^{\text{sym}}|$ is taken equal to $|r_{ab}|$. This extrapolation is not only important for linear preservation, but also for obtaining optimal convergence rates in convection-diffusion problems with boundary layers.

Remark 4.3. Notice that when $x_{ab}^{\text{sym}}$ coincides with a node the value $u_h(x_{ab}^{\text{sym}})$ is not unique. In this case, we compute $\langle \nabla u_h \rangle_{ab}$ and $\langle \|
abla u_h \cdot \hat{r}_{ab}\| \rangle_{ab}$ for all values of $u_h(x_{ab}^{\text{sym}})$ in $\Omega_a$.

Making use of the above definitions, the proposed shock detector reads

$$
\alpha_a(u_h, \vec{u}_h) \doteq \left\{ \begin{array}{ll}
\left( \frac{\sum_{b \in \mathcal{N}_h(a)} |\nabla u_h|_{ab}}{\sum_{b \in \mathcal{N}_h(a)} 2 \langle \|
abla u_h \cdot \hat{r}_{ab}\| \rangle_{ab}} \right)^q & \text{if } \sum_{b \in \mathcal{N}_h(a)} \langle \|
abla u_h \cdot \hat{r}_{ab}\| \rangle_{ab} \neq 0,
0 & \text{otherwise},
\end{array} \right. \quad (9)
$$

where $q \in \mathbb{R}$. Let us prove that the shock detector (9) satisfies the shock detector property in Def. 4.1. We note that the definition of $\alpha_a$ is motivated from [3]. Here, instead of using the maximum coefficient obtained, we use the sum over all gradient jumps divided by the sum of all gradient means. A similar formulation was used in [3], but with a slightly different definition.

Lemma 4.4. The function $\alpha_a(u_h, \vec{u}_h)$ defined in (9) satisfies the shock detector property in Def. 4.1. Furthermore, if $a \in \mathcal{N}_h$ is not an extremum and $q = \infty$, $\alpha_a(u_h, \vec{u}_h) = 0$.

Proof. Let us assume that $u_h$ has a discrete maximum (resp. minimum) on $x_a$, then

$$
\begin{align*}
&u_b - u_a \leq 0 \quad \forall b \in \mathcal{N}_h(a) \quad \text{and} \quad u_{ab}^{\text{sym}} - u_a \leq 0 \quad \forall b \in \mathcal{N}_h(a), \\
&\text{resp. } u_b - u_a \geq 0 \quad \forall b \in \mathcal{N}_h(a) \quad \text{and} \quad u_{ab}^{\text{sym}} - u_a \geq 0 \quad \forall b \in \mathcal{N}_h(a).
\end{align*}
$$

Therefore,

$$
\begin{align*}
\sum_{b \in \mathcal{N}_h(a)} \|\nabla u_h\|_{ab} &= \sum_{b \in \mathcal{N}_h(a),} \|\nabla u_h\|_{ab} + \sum_{b \in \mathcal{N}_h(a), x_b \neq x_a} \|\nabla u_h\|_{ab} \\
&= \sum_{b \in \mathcal{N}_h(a), x_b = x_a} \frac{u_b - u_a}{h_a} + \sum_{b \in \mathcal{N}_h(a), x_b \neq x_a} \frac{u_b - u_a}{|r_{ab}|^q} + \frac{u_{ab}^{\text{sym}} - u_a}{|r_{ab}^{\text{sym}}|^q} \\
&= \sum_{b \in \mathcal{N}_h(a), x_b = x_a} \frac{|u_b - u_a|}{h_a} + \sum_{b \in \mathcal{N}_h(a), x_b \neq x_a} \frac{|u_b - u_a|}{|r_{ab}|^q} + \frac{|u_{ab}^{\text{sym}} - u_a|}{|r_{ab}^{\text{sym}}|^q} = \sum_{b \in \mathcal{N}_h(a)} 2 \langle \|
abla u_h \cdot \hat{r}_{ab}\| \rangle_{ab}.
\end{align*}
$$
Thus, $\alpha_a(u_h, \overline{u}_h) = 1$. Further, if $u_a$ is not an extremum, then (10) is no longer true. Hence,

$$
\left| \sum_{b \in \mathcal{N}_h(a)} [\nabla u_h]_{ab} \right| = \left| \sum_{b \in \mathcal{N}_h(a)} [\nabla u_h]_{ab} + \sum_{b \notin \mathcal{N}_h(a)} [\nabla u_h]_{ab} - \frac{u_b - u_a}{h_a} \right|
\leq \sum_{b \in \mathcal{N}_h(a)} \frac{|u_b - u_a|}{h_a} + \sum_{b \notin \mathcal{N}_h(a)} \frac{|u_b - u_a|}{|p_{ab}|} + \frac{|u_b^{\text{sym}} - u_a|}{|p_{ab}^{\text{sym}}|} = \sum_{b \in \mathcal{N}_h(a)} 2\|\nabla u_h : \hat{r}_{ab}\|_{ab}.
$$

Therefore, $\alpha_a(u_h, \overline{u}_h) < 1$. Moreover, when $q = \infty$, $\alpha_a(u_h, \overline{u}_h) = 0$ if $u_a$ is not an extremum.

In order to prove the DMP-preservation in the numerical analysis below we need to perturb both the weak boundary conditions and the bilinear form. The perturbed weak boundary conditions read:

$$
\tilde{B}_h(u_h, \overline{w}_h, \overline{v}_h; v_h) = B_h(u_h; v_h) + \sum_{a \in \mathcal{N}_h} \sum_{b \notin \mathcal{N}_h(a)} \nu_{ab} (u_h, v_h) u_a \overline{v}_b,
$$

where $v_h, w_h \in V_h$, and $\overline{w}_h, \overline{v}_h \in V_h|_{\partial \Omega}$. Furthermore, given $u_h, v_h, w_h \in V_h$ and $\overline{w}_h, \overline{v}_h \in V_h|_{\partial \Omega}$, we can define the perturbed bilinear form $\tilde{K}_h$ as:

$$
\tilde{K}_h(w_h, \overline{w}_h; u_h, v_h) = K_h(u_h, v_h) + \sum_{a \in \mathcal{N}_h} \sum_{b \notin \mathcal{N}_h(a)} \nu_{ab} (w_h, \overline{w}_h) u_a v_b.
$$

where $\ell(a, b) = 2\delta_{ab} - 1$ is the graph-Laplacian operator. It leads to the following stabilized steady discrete problem: Find $u_h \in V_h$ with $\overline{u}_h \in V_h|_{\partial \Omega}$ such that

$$
\tilde{K}_h(u_h, \overline{u}_h; u_h, v_h) = G_h(v_h) + \tilde{B}_h(u_h, \overline{u}_h, \overline{u}_h; v_h) \quad \forall v_h \in V_h.
$$

We are ready to prove the desired DMP property of this method. For this purpose, we define $\tilde{K}_{ab}(u_h, \overline{u}_h) = \tilde{K}_h(u_h, \overline{u}_h; \varphi_b, \varphi_a)$ for $a, b \in \mathcal{N}_h$, and $\tilde{B}_{ab}(u_h, \overline{u}_h) = \tilde{B}_h(u_h, \overline{u}_h; \varphi_b, \varphi_a)$ for $a \in \mathcal{N}_h, b \in \mathcal{N}_h^\circ$.

**Theorem 4.5.** The discrete problem (13) with the stabilized semilinear forms defined in (11) and (12) is DMP-preserving for $g = 0$.

**Proof.** As seen in Th. 3.6, the solution is DMP-preserving if conditions (6b)-(6c) are satisfied. Let us verify these two conditions. Let $x_a$ be an interior node and assume that $u_h$ has an extremum on $x_a$. 

Register for free at https://www.scipedia.com to download the version without the watermark
Given the set of all nodes \( b \in \mathcal{N}_h(a) \) coupled to node \( a \in \mathcal{N}_h \), we have:

\[
\sum_{b \in \mathcal{N}_h(a)} K_{ab} - B_{ab} = \sum_{b \in \mathcal{N}_h(a)} \left\{ \sum_{K \in T_h} \int_K (\mu \nabla \phi_b \cdot \nabla \phi_a - \varphi_b \beta \cdot \nabla \phi_a) + \sum_{F \in \mathcal{E}_h^b} \int_F \mu \left( \beta [\phi_b] \cdot [\phi_a] + c h^{-1} [\phi_b] \cdot [\phi_a] \right) \right. \\
+ \sum_{F \in \mathcal{E}_h^b \cup \mathcal{E}_h^a} \int_F \beta [\phi_b] \cdot [\phi_a] + \sum_{F \in \mathcal{E}_h^a} \int_F \left( \beta \cdot n_F \right) [\phi_b] \cdot [\phi_a] \\
+ \left. \sum_{F \in \mathcal{E}_h^b \cup \mathcal{E}_h^a} \int_F \beta \cdot n_F \varphi_b \varphi_a + \sum_{F \in \mathcal{E}_h^a} \int_F \mu \varphi_b \nabla \phi_a \cdot n_F \right) \\
- \sum_{F \in \mathcal{E}_h^b \cup \mathcal{E}_h^a} \int_F c h^{-1} \phi_b \varphi_a \right\}.
\]

(We note that \( B_{ab} \) has only been defined for nodes \( b \in \mathcal{N}_h(a) \). Here, we abuse of notation, and extend by zero the definition to all nodes, with \( B_{ab} = 0 \) when \( b \notin \mathcal{N}_h(a) \).) We use the fact that the shape functions are a partition of unity, i.e., \( \sum_{b \in \mathcal{N}_h(a)} \varphi_b \) is equal to one on \( \Omega_a \) and zero elsewhere. As a result, \( \sum_{b \in \mathcal{N}_h(a)} [\varphi_b] = 0 \) on facets \( F \subset \Omega_a \setminus \partial \Omega \), and \( \sum_{b \in \mathcal{N}_h(a)} \nabla \varphi_b = 0 \) in any \( K \in T_h \). On the other hand, \( \varphi_a \) vanishes on any \( F \in \partial \Omega_b \setminus \partial \Omega \) by construction. Using these properties, we get:

\[
\sum_{b \in \mathcal{N}_h(a)} \sum_{F \in \mathcal{E}_h^b} \int_F \left( \beta [\phi_b] \cdot [\phi_a] + \sum_{F \in \mathcal{E}_h^a} \int_F \mu [\phi_b] \cdot [\phi_a] - \sum_{F \in \mathcal{E}_h^b \cup \mathcal{E}_h^a} \int_F c h^{-1} \phi_b \varphi_a \right) = 0,
\]

\[
\sum_{b \in \mathcal{N}_h(a)} \sum_{F \in \mathcal{E}_h^b} \int_F \mu [\nabla \phi_b] \cdot [\phi_a] = 0, \quad \sum_{b \in \mathcal{N}_h(a)} \int_{\Omega} \mu [\nabla \varphi_b] \cdot [\varphi_a] = 0,
\]

and the following terms can be integrated by parts as

\[
\sum_{b \in \mathcal{N}_h(a)} \left\{ - \sum_{K \in T_h} \int_K \varphi_b \beta \cdot \nabla \varphi_a + \sum_{F \in \mathcal{E}_h^b \cup \mathcal{E}_h^a} \int_F \beta [\varphi_b] \cdot [\varphi_a] + \sum_{F \in \mathcal{E}_h^a} \int_F \beta \cdot n_F \varphi_b \varphi_a \right\} \\
= \sum_{b \in \mathcal{N}_h(a)} \left\{ - \sum_{K \in T_h} \left( \int_K \varphi_b \beta \cdot \nabla \varphi_a + \int_{\partial K} \beta \cdot n_K \varphi_a \right) \right\} \\
= \sum_{b \in \mathcal{N}_h(a)} \sum_{K \in T_h} \int_K \varphi_b \beta \cdot \nabla \varphi_a = 0.
\]

Finally, since \( \sum_{b \in \mathcal{N}_h(a)} \nu_{ab}(u_h)v_a v_b f(a, b) + \sum_{b \in \mathcal{N}_h(a)} \nu_{ab}^b(u_h)v_a v_b = 0 \) by construction (see (7) and (8)), then \( \sum_{\phi \in \mathcal{N}_h(a)} \tilde{K}_{ab}(u_h, \pi_h) - \tilde{B}_{ab}(u_h, \pi_h) = 0 \). Moreover, it is clear that \( \tilde{K}_{ab}(u_h, \pi_h) \leq 0 \) for any \( b \neq a \) and \( \tilde{B}_{ab}(u_h, \pi_h) \geq 0 \) in all cases, based on the definition of these operators in (11)-(12) and their respective graph-viscosities in (7)-(8). It finishes the proof. \( \square \)
Thus, by Th. 4.5, we can ensure that the extrema of the solution of (3), will be on the boundary of the domain when \( g = 0 \). Let us now define the mass matrix perturbation used in order to obtain a LED scheme. As its name reveals, this property ensures that the value of the discrete maximum and the minimum of a transient problem can be bounded by those in the initial solution \( u_0 = u(\cdot, t_0) \) and boundary conditions. It has been proved in [12, Lemma 3.2] that, if the steady problem, e.g., (13), enjoys the DMP property, its transient version enjoys the LED property if we replace the mass matrix \((\partial_t u_h, v_h)\) by its lumped version \((\partial_t u_h, v_h)_h\) corresponding to the Gauss-Lobatto sub-integration. The form \((\cdot, \cdot)_h\) is such that \((\partial_t u_h, \varphi_a)_h = \partial_t u_a (\mathbb{I}, \varphi_a)\) for any \( a \in \mathcal{N}_h \). In fact, as Kuzmin and co-workers have proved in [22, Th. 4], it is enough to lump only the terms associated to the degrees of freedom where \( u_h \) has an extremum. Following the same strategy as in [2], we can perform selective lumping using the shock detector. We define:

\[
M_h^L(u_h, \bar{u}_h; \partial_t u_h, v_h) \equiv \sum_{a \in \mathcal{N}_h} v_a (1 - \alpha^Q_a(u_h, \bar{u}_h)) (\partial_t u_h, \varphi_a) + \alpha^Q_a(u_h, \bar{u}_h) \partial_t u_a v_a (\mathbb{I}, \varphi_a).
\]  

(14)

The exponent \( Q > 0 \) is added in order to minimize the lumping perturbation, which leads to phase error in the discrete solution. In addition, we define \( \mathcal{M}_{ab}(u_h, \bar{u}_h) \equiv M_h^L(u_h, \bar{u}_h; \partial_t u_h, v_h) \), for \( a, b \in \mathcal{N}_h \).

If one considers the semi-discrete problem in space only, we have: Find \( u_h \in V_h \) such that

\[
M_h^L(u_h, \bar{u}_h; \partial_t u_h, v_h) + K_h(u_h, \bar{u}_h; u_h, v_h) = G_h(v_h) + B_h(u_h, \bar{u}_h; \bar{u}_h; v_h) \quad \forall v_h \in V_h.
\]  

(15)

**Lemma 4.6.** The scheme (15) with the semilinear forms defined in (14), (11), and (12) is LED for \( g = 0 \).

*Proof.* The conditions required on \( \mathcal{K}(u_h, \bar{u}_h) \) and \( \mathcal{B}(u_h, \bar{u}_h) \) in Th. 3.4 to obtain a LED scheme have already been proved in Th. 4.5. Further, if we assume that \( u_a \) is an extremum, then \( \alpha_a(u_h, \bar{u}_h) = 1 \) and \( \mathcal{M}_{ab}(u_h) \) becomes \((\mathbb{I}, \varphi_a) = \delta_a m_a \) with \( m_a = \int_{\Omega} \varphi_a \). Hence, the definition of the mass matrix in (14) satisfies (6a). As a result, we fulfill all conditions stated in Th. 3.4 and thus the scheme is LED.

Furthermore, the stabilized problem (15) is linearity preserving, i.e. linear solutions are solution of the original IP dG method (2).

**Lemma 4.7.** The stabilization terms in (11), (12), and (14), vanish for functions \( u \in P_1(\Omega) \), i.e.,

\[
\mathcal{B}_h(u, u_{\partial\Omega}; \bar{u}_h; v_h) = B_h(\bar{u}_h; v_h), \quad \mathcal{K}_h(u, u_{\partial\Omega}; u_h, v_h) = K_h(u_h, v_h), \quad M_h^L(u, u_{\partial\Omega}; u_h, v_h) = (u_h, v_h), \quad \text{for any } u_h \in V_h.
\]

*Proof.* If \( u \) is linear and continuous, then by definition \([\nabla u]_{ab} = 0 \) \( \forall a \) and \( b \in \mathcal{N}_h(a) \). Hence, for any \( a \) we have that \( \alpha_a \equiv 0 \). Thus, both \( v^0_{ab} \) and \( v_{ab} \) are equal to zero. Thus, all the stabilization terms vanish and we recover the original formulation.

The results in this section for the BE time discretization can be extended to any \( \theta \)-method. We refer the reader to the work by Kuzmin and co-workers [19,21] for the proofs of such properties. In particular, \( \theta \)-methods are positivity-preserving under the CFL-like condition (see [19, Th. 1])

\[
\Delta t \leq \min_{a \in \mathcal{N}_h} \frac{(1 - \theta) K_h(u_h^{n+1}, \bar{u}_h^{n+1}; \varphi_a; \varphi_a)}{(1 - \theta) K_h(u_h^n, \bar{u}_h^n; \varphi_a; \varphi_a)}.
\]

Furthermore, under certain conditions of the matrix and the RHS, it has been proved in [21, Th. 4] that the scheme is not only positivity-preserving but satisfies the DMP. This means that the discrete maximum and the minimum of the solution are bounded by the values of the initial solution and the boundary conditions for any \( \theta \)-method.

The authors in [19, Th. 1] take advantage of the mass lumping properties for all of these proofs, but the lumping only needs to be activated for the degrees of freedom where the discrete solution has an extrema. Thus, the scheme defined in (4) together with the definition of \( M_h^L(\cdot, \cdot; \cdot, \cdot) \) given by (14) leads to a DMP-preserving method under the above CFL-like condition.
5. Lipschitz continuity and existence of solutions

Let us define the Cartesian product space $\tilde{V}_h = V_h \times V_h|_{\partial \Omega}$. Thus, any function $\tilde{v} \in \tilde{V}_h$ can be expressed as $\tilde{v} = (v, \tilde{v})$, where the first component includes the values of the dG function $v \in V_h$ and the second component the projection of the Dirichlet values $\tilde{v} \in V_h|_{\partial \Omega}$. Analogously, we can define the set of nodes for $\tilde{V}_h$ as $M_h = N_h \times N_h^\partial = \{(a_1, 0), (0, a_2) : a_1 \in N_h, a_2 \in N_h^\partial\}$. We consider an extended graph-Laplacian operator over $\tilde{V}_h \times \tilde{V}_h$ as follows:

$$
\tilde{D}(\tilde{u}, \tilde{v}) = \tilde{D}((u, \tilde{u}), (v, \tilde{v})) \doteq \sum_{a \in M_h} \sum_{b \in M_h} \tilde{\nu}_{ab} \ell(a, b) \tilde{u}_a \tilde{v}_b
$$

$$
\doteq \sum_{a \in N_h} \sum_{b \in N_h} \nu_{ab} \ell(a, b) u_b v_a - \sum_{a \in N_h^\partial} \sum_{b \in N_h^\partial} \nu^\partial_{ab} u_b v_a - \sum_{a \in N_h} \sum_{b \in N_h} \nu^\partial_{ab} u_b \tilde{v}_a.
$$

Note that the boundary degrees of freedom are replicated. Based on this definition, we implicitly have:

$$
\tilde{\nu}_{ab} = \nu_{ab}, \quad \text{if} \quad a, b \in (N_h, 0),
$$

$$
\tilde{\nu}_{ab} = \nu^\partial_{ab}, \quad \text{if} \quad a \in (N_h^\partial, 0), \quad b \in (0, N_h^\partial),
$$

$$
\tilde{\nu}_{ab} = \nu^\partial_{ab}, \quad \text{if} \quad a \in (0, N_h^\partial), \quad b \in (N_h^\partial, 0),
$$

$$
\tilde{\nu}_{ab} = 0, \quad \text{if} \quad a, b \in (0, N_h^\partial).
$$

It is easy to check that this operator is symmetric and positive-semidefinite. In order to show the second property, we use the expression for $\nu_{ab}$ and $\nu^\partial_{ab}$ in (7) and (8), respectively, in order to get $\nu_{ab}, \nu^\partial_{ab} \geq 0$, and

$$
\tilde{\nu}_{aa} = \nu_{aa} = \sum_{b \in N_h(a) \setminus a} \nu_{ab} + \sum_{b \in N_h^\partial(a) \setminus a} \nu^\partial_{ab} = \sum_{b \in M_h(a) \setminus a} \tilde{\nu}_{ab}.
$$

Using the last property and the definition of $\ell(a, b)$, we get:

$$
2\tilde{D}(\tilde{u}, \tilde{v}) = \sum_{a \in M_h} \sum_{b \in M_h} \tilde{\nu}_{ab} \ell(a, b) \tilde{u}_a (\tilde{u}_b - \tilde{u}_a) + \sum_{a \in N_h} \sum_{b \in N_h^\partial} \tilde{\nu}_{ab} \ell(a, b) \tilde{u}_a (\tilde{v}_b - \tilde{v}_a).
$$

Thus, we have $|\tilde{u}|^2_{D^2} = \tilde{D}(\tilde{u}, \tilde{u}) \geq 0$. Further, we define the restriction operators $D(u, v) = \tilde{D}((u, 0), (v, 0))$ and $D^\partial(\tilde{u}, \tilde{v}) = \tilde{D}((0, \tilde{u}), (0, \tilde{v}))$, and their corresponding semi-norms $|u|_D = D(u, u)$ and $|\tilde{u}|_{D^\partial} = D^\partial(\tilde{u}, \tilde{u})$.

Given the source $g \in V'_h$ and $\tilde{u} \in V_h|_{\partial \Omega}$, we define the operator $T : V_h \to V'_h$ for the steady problem as:

$$
(T(z), v) = K_h(z, v) - B_h(\tilde{u}, v) - G_h(v) + \tilde{D}((z, \tilde{u}), (v, 0))
$$

$$
= K_h(z, v) - B_h(\tilde{u}, v) - G_h(v) + \sum_{a \in N_h} \sum_{b \in N_h(a)} \nu_{ab} \ell(a, b) v_a z_b
$$

$$
- \sum_{a \in N_h} \sum_{b \in N_h^\partial(a)} \nu^\partial_{ab} v_a \tilde{u}_b.
$$

Clearly, to find $u_h \in V_h$ such that $(T(u_h)) = 0$ is equivalent to the stabilized problem (13). For transient problems, given also the previous time step solution $u_h^n \in V_h$, we define the operator $T^{n+1} : V_h \to V'_h$ at every time step as

$$
(T^{n+1}(z), v) = M_h^f(z; z, v) - M_h^f(z; u_h^n, v) + (T(z), v).
$$

System (15) can be stated in compact form as: find $u_h \in V_h$ such that $T^{n+1}(u_h) = 0$. In the next theorem, we prove that both operators are Lipschitz continuous. We provide a sketch of the proof, since it follows the same lines as in [2, Th. 6.1].

**Theorem 5.1.** The nonlinear operators $T$ and $T^{n+1}$ are Lipschitz continuous in $V_h$ for $q \in \mathbb{N}^+$. 
Proof. In order to prove the Lipschitz continuity, we proceed as in [2]. After some manipulation, we get:

\[ |\langle \mathbf{T}(u), w \rangle - \langle \mathbf{T}(v), w \rangle| \leq |K_h(u - v, v)| + \left| \sum_{a \in N_h} \sum_{b \in N_h(a)} \nu_{ab}(v)\ell(a, b)w_a(u_b - v_b) \right| \\
+ \left| \sum_{a \in N_h} \sum_{b \in N_h(a)} (\nu_{ab}(u) - \nu_{ab}(v))\ell(a, b)w_a v_b \right| \\
+ \left| \sum_{a \in N_h} \sum_{b \in N_h(a)} (\nu_{ab}^0(u) - \nu_{ab}^0(v))w_a u_b \right| . \]

The first term is linear and continuous (see, e.g., [1]). We have to prove Lipschitz continuity for the rest of terms. We use the inverse inequalities \( \| \nabla \varphi_a \|_K \leq Ch^{-1} \| \varphi_a \|_K \) and \( \| \nabla \varphi_a \|_{F} \leq Ch^{-1} \| \varphi_a \|_{F} \) (see [6]) and the fact that shape functions are a partition of unity (\( \| \varphi_a \|_K \leq Ch^{d/2} \) and \( \| \varphi_a \|_{F} \leq Ch^{(d-1)/2} \)), to get:

\[ K_h(u_h, \bar{u}_h; \varphi_b, \varphi_a) - B_h(u_h, \bar{u}_h; \varphi_b, \varphi_a) \leq C q(h^{d-1} \| \beta \|_{L^\infty(\Omega)} + \mu h^{d-2}). \]  

(18)

The rest of the proof follows the same lines as in [2, Th. 6.1] and is not included for the sake of conciseness. The graph-Laplacian edges for pairs \( a, b \) such that \( x_a \neq x_b \) are as in [2], using (18). The case \( x_a = x_b \) is simpler.

Lipschitz continuity for the transient problem is a consequence of the Lipschitz continuity of \( \mathbf{T} \) and of the mass matrix with the selective mass lumping. The last property can be proved using again the analysis in [2, Th. 6.1].

Next, we show that the proposed schemes have at least one solution. Uniqueness results could also be obtained for the diffusion-dominated regime following the ideas in [5]. In the following, we will use \( C \) as a general constant that can take different values at different appearances.

**Theorem 5.2.** There is at least one solution \( u_h \in V_h \) of the steady problem \( \mathbf{T}(u_h) = 0 \), and one solution of every time step of the transient problem, i.e., \( \mathbf{T}^{n+1}(u_h) = 0 \).

Proof. In order to prove existence of solutions, we rely on the approach in [5], based on fixed point arguments. First, we combine the stability analysis in [8] (for first-order hyperbolic problems) with the stability analysis for the interior penalty discretization of the Laplacian operator (see, e.g., [1] for details), getting:

\[ K_h(z, z) \geq C \| z \|_h^2 , \quad \text{with} \quad \| z \|_h^2 \equiv \sum_{K \in T_h} \mu |z|_{H^1(K)}^2 + \sum_{F \in F_h} \left( \mu \epsilon \beta^{-1} \| \nabla z \|_{L^2(F)}^2 + \| c_{\beta,F}^2 \|_{H^1(F)}^2 \right) , \]

(19)

with \( \epsilon \) big enough, and \( c_{\beta,F}(x) \equiv |\beta(x) \cdot n_F(x)| \). On the other hand, using standard dG arguments (see [1] and [15, Prop. 3.55]), we have:

\[ B_h(\bar{u}, z) \leq C c^{-1} \| \bar{u} \|_{L^2(\partial\Omega)}^2 + C c^{-1} h^{-1} \mu c \| \bar{u} \|_{L^2(\partial\Omega)}^2 + C \| z \|_h^2 , \]

(20)

for \( c \) arbitrarily small.

We note that the nonlinear stabilization terms can be written in terms of the extended graph-Laplacian operator as:

\[ \sum_{a \in N_h} \sum_{b \in N_h(a)} \nu_{ab}\ell(a, b)v_a z_b - \sum_{a \in N_h} \sum_{b \in N_h^2(a)} \nu_{ab}^0 v_a \bar{u}_b = \tilde{D}(z, \bar{u}), (v, 0)) . \]

Taking \( v = z \) and using (16) and the Cauchy-Schwarz inequality, we get:

\[ \tilde{D}(z, \bar{u}), (z, 0)) \leq \frac{3}{2} |z|_{D}^2 + \frac{1}{2} |\bar{u}|_{D,0}^2 . \]
Combining (17), (19), and (20) (with $c$ small enough), we finally obtain:

$$C(T(z), z) \geq \|z\|^2 + \|u_x\|_{D^2}^2 - \|\epsilon \hat{u}\|^2_{L^2(\partial \Omega)} - h^{-1} \mu_c \|\hat{u}\|^2_{L^2(\partial \Omega)}.$$ 

We can readily pick a $z \in V_h$ such that $(T(z), z) > 0$. Using the Brower’s fixed point theorem, there exists $u_h \in V_h$ such that $T(u_h) = 0$, and thus, solves the steady version of (15) (see [5] for details). Existence is straightforward for the transient problem, combining the previous results with the coercivity of the mass matrix operator.

**Remark 5.3.** As a result of the previous theorem and Lemma 4.7, the method is linearly preserving and Lipschitz continuous. Using the ideas in [5, Theorem 4], one could prove optimal convergence in diffusion-dominated regimes.

6. Smoothing the shock detector

In Sect. 4, we have defined $u_{\alpha_{ab}}(u_h, \tau_h)$, $\nu_{ab}(u_h, \pi_h)$, and $\alpha_{ab}(u_h, \pi_h)$ in (7), (8), and (9), respectively, using non-smooth functions. The problem of using this raw definitions is that, since they are not smooth, it is difficult for the nonlinear solvers to converge. Thus, following the ideas in [2], we add some parameters $(\tau_h, \gamma_h, \sigma_h)$ and regularize the definition of non-smooth functions such as the absolute value and the maximum. In this section we will proceed to unfold all the smooth definitions to facilitate the reproducibility of the method. The resulting formulation is not only Lipschitz continuous but twice differentiable by construction. Furthermore, the smoothing involves slightly more diffusion, and it is easy to check that we keep the DMP and LED properties above. Linearity-preservation is only satisfied weakly (see [2, Remark 7.3]). We do not prove these results for the sake of conciseness, since the proofs are similar to the ones in [2, Lemma 7.1].

We will start by introducing a couple of smoothed versions of the absolute value:

$$|x|_{1, \tau_h} = \sqrt{x^2 + \tau_h}, \quad |x|_{2, \tau_h} = \frac{x^2}{\sqrt{x^2 + \tau_h}}.$$ 

The value of $\tau_h$ is assumed to be small and is going to be specified in Sect. 7. For values of $x \gg \tau_h$, we have $|x|_{1, \tau_h} \approx |x| \approx |x|_{2, \tau_h}$ but always $|x|_{2, \tau_h} \leq |x| \leq |x|_{1, \tau_h}$. Now, we can redefine $\|\nabla u_h \cdot \hat{r}_{ab}\|_{ab}$ as:

$$\|\nabla u_h \cdot \hat{r}_{ab}\|_{2, \tau_h} \triangleq \begin{cases} \frac{1}{2} \left( |u_b - u_a|_{2, \tau_h} + \left| u_{ab}^{\text{sym}} - u_a \right|_{2, \tau_h} \right) & \text{if } x_a = x_b, \\ 1 & \text{otherwise}. \end{cases}$$

The quotient associated to $\alpha_a$ would read:

$$\zeta_a = \frac{\sum_{b \in N(h)(a)} \|\nabla u_h\|_{1, \tau_h} \nabla u_h \|_{ab}^{\text{sym}} + \gamma_h}{\sum_{b \in N(h)(a)} 2 \|\nabla u_h \cdot \hat{r}_{ab}\|_{2, \tau_h} \|_{ab} + \gamma_h}.$$ 

Here $\gamma_h$ is another extra stability parameter added to ensure differentiability of $\zeta_a$ for values of $u_h$ such that the denominator is nullified. By the definition and the properties of $|\cdot|_{1, \tau_h}$ and $|\cdot|_{2, \tau_h}$, it is easy to prove that in the case that $u_h$ has a local discrete extremum on $a$, $\zeta_a > 1$. So, since we want $\alpha_a$ to enjoy the shock detector property stated in Def. 4.1, we need to construct a twice differentiable function $Z$ such that $Z(x) = 1$ when $x \geq 1$. To this end, we define

$$Z(x) = \begin{cases} 2x^4 - 5x^3 + 3x^2 + x & x < 1, \\ 1 & x \geq 1. \end{cases}$$

Now we are able to define the smooth value of $\alpha_a$ as $\hat{\alpha}_a = (Z(\zeta_a))^{\frac{3}{2}}$. Moreover, we have also modified the computation of the maximum in the following way:

$$\max_{\sigma_h}(x, y) = \frac{1}{2} \sqrt{(x - y)^2 + \sigma_h} + \frac{1}{2}(x + y).$$
Furthermore, at boundaries \( u_{ab}^{sym} \) is computed using the sign function which needs to be regularized too. In particular we use \( \text{sign}_{\sigma_h}(x) \equiv x/|x|_{1,\gamma} \). Then the smooth definition of \( \nu_{ab} \) in \((8)\) for \( a \in \mathcal{N}_h \) and \( b \in \mathcal{N}_h(a) \setminus \{a\} \) will read
\[
\nu_{ab} = \max_{\sigma_h}(0, \max_{\sigma_h}(\bar{\sigma}_a K_h (\varphi_b, \varphi_a), \bar{\sigma}_b K_h (\varphi_a, \varphi_b))),
\]
and for \( b \in \mathcal{N}_h^\rho(a) \)
\[
\nu_{ab}^\rho = \max_{\sigma_h}(-\bar{\sigma}_a B_h (\varphi_b; \varphi_a), 0).
\]
The objective of these modifications is twofold. On the one hand, they smooth the function improving the convergence of the nonlinear iterations. On the other hand, they make the method differentiable with respect to \( u_h \), and the Jacobian matrix is defined everywhere; some nonlinear iteration methods, such as Newton’s method, which need to compute the Jacobian matrix of the problem, can be used. Further, the method is twice differentiable which is required to get quadratic nonlinear convergence rates with Newton’s method.

In order to keep a dimensionally correct method and, at the same time, do not affect the convergence of the non-stabilized method, the parameters should scale as follows:
\[
\sigma_h = \sigma|\beta|^2 L^{2(d-3)}h^4, \quad \tau_h = \tau h^2 L^{-4}, \quad \gamma_h = \gamma L^{-1},
\]
where \( d \) is the space dimension of the problem, \( L \) a characteristic length of the problem, \( \tau \) and \( \gamma \) have the same dimension as the unknown, and \( \sigma \) is dimensionless.

6.1. Parameters fine-tuning. In order to find the appropriate values for all the parameters introduced before, we will check how these values affect the performance of the method. To this end, we will consider the steady \((\partial_t u = 0)\) transport \((\mu = 0)\) problem with no force \((g = 0)\) and rotational convection \( \beta = (y, -x) \):
\[
\nabla \cdot (\beta u) = 0 \quad \text{in} \ [0,1] \times [0,1].
\]
In the transport case, the Dirichlet boundary conditions are only imposed on the inflow boundaries, which, for this convection field, are the sides of the square \([0,1] \times [0,1]\) corresponding to \( x = 0 \) and \( y = 1 \). We will impose 0 all along the side \( y = 1 \) and the following function on the side \( x = 0 \):
\[
\pi(0,y) = \begin{cases} 
1 & \text{if } y \in [0.15,0.45], \\
\cos^2\left(\frac{10}{\pi}(y-0.4)\right) & \text{if } y \in [0.55,0.85], \\
0 & \text{elsewhere}.
\end{cases}
\]
We know that the exact solution of this problem consists of a translation of this function in the direction of the convection in such a way that on the outflow boundary corresponding to \( y = 0 \) the solution is \( u(x,0) = \pi(0,x) \). We solve this problem in a \( 100 \times 100 \) \( Q_1 \) mesh and check the effect of the constants \( \sigma, \tau \) and \( \gamma \) on the resulting outflow profile with respect to the value in the inflow boundary \( x = 0 \), plotted in Fig. 2(a).

First of all, we can observe the dissipative effect of the parameters on the final solution. We set values of \( q = \{1,2,4,10\} \), \( \sigma = \{10^{-1},10^{-2},10^{-3},10^{-4}\} \), and \( \tau = \sigma^2 \), and fix the value of \( \gamma \) to \( 10^{-2} \). We use Picard linearization and the nonlinear iterative scheme with the relaxation parameters proposed in [18], using the same parameter values therein. In addition, we also solve all tests using a hybrid Newton-Picard method; first we use Picard to get a better starting point for Newton, particularly when the nonlinear error is lower than \( 10^{-2} \) we change to Newton method with line search. Note that for the hybrid scheme the total number of iterations used for comparison also include the first iterations performed with Picard method. For both nonlinear solvers the tolerance is set to be \( 10^{-4} \) and we allow a maximum of 500 iterations. Whenever the solver exceeds 500 iterations we define the scheme to be not converged (NC). For both schemes the linearized system of equations is solved with a direct solver. The results are shown in Fig. 3. It can be observed that, in order to obtain sharp solutions, it is important to use both high values of \( q \) and low values of \( \sigma \) and \( \tau \). Nevertheless fixing \( q = 10 \), and tuning only \( \sigma \) and \( \tau \), we can either obtain a method that is easy to converge, but quite dissipative, or a method that is harder to converge, but much more accurate.

For the moment, we have fixed the relation between \( \sigma \) and \( \tau \). In the next test we fix \( q = 10 \), \( \gamma = 10^{-2} \), and different values for \( \tau \) and \( \sigma \). In particular we will use \( \tau \in \{10^{-1},10^{-2},10^{-4},10^{-8}\} \)