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

[†] 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 ensures 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:

$$\left\{ \begin{array}{rcll} \partial_t u + \boldsymbol{\nabla}\cdot(\boldsymbol{\beta} u) - \boldsymbol{\nabla}\cdot(\mu\boldsymbol{\nabla} u) & = & g & \text{in } \Omega\times[0,\mathrm{T}], \\ u(x,t) & = & \overline{u}(x,t) & \text{on } \partial\Omega\times[0,\mathrm{T}], \\ u(x,0) & = & u_0(x) & x\in\Omega. \end{array} \right. \tag{1}$$

The domain Ω is an open, bounded, connected subset of $\mathbb{R}^d$ with a Lipschitz boundary $\partial\Omega$, where d is the space dimension, $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{x})$ 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., $\partial\Omega \equiv \partial\Omega_D$, 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 $\partial\Omega^- \doteq \{\boldsymbol{x}\in\partial\Omega : \boldsymbol{\beta}\cdot\boldsymbol{n}_{\partial\Omega} < 0\}$, where $\boldsymbol{n}_{\partial\Omega}$ is the outward-pointing unit normal. Further, we define the outflow boundary as $\partial\Omega^+ \doteq \partial\Omega\backslash\partial\Omega^-$. Below, we also consider the steady case, by eliminating the time derivative term.

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

Given the mesh $\mathcal{T}_h$, the non-empty intersection $F = \partial K \cap \partial K'$ of two neighbor elements $K, K' \in \mathcal{T}_h$ is called an interior facet of $\mathcal{T}_h$ if it is a subdomain of dimension $d-1$. The set of all the interior facets is denoted by $\mathcal{E}_h^0$. On the other hand, the non-empty intersection $F = \partial K \cap \partial\Omega^-$ of an element $K \in \mathcal{T}_h$ on the boundary with 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 $\mathcal{E}_h^-$, and the set of outflow boundary facets is denoted by $\mathcal{E}_h^+$. (We assume that the finite element partition is conforming with the inflow and outflow boundaries.) The set of all the facets is denoted by $\mathcal{E}_h \doteq \mathcal{E}_h^0 \cup \mathcal{E}_h^+ \cup \mathcal{E}_h^-$. For any facet $F \in \mathcal{E}_h^0$, we represent with K_F^+ and K_F^- the only two neighbor 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 $V_h = \{v_h : v_h|_K \in \mathbb{P}_1(K)\ \forall K \in \mathcal{T}_h\}$, where $\mathbb{P}_1(K)$ is the space of linear polynomials in K. For hexahedral (or quadrilateral) meshes, $V_h = \{v_h : v_h|_K \in \mathbb{Q}_1(K)\ \forall K \in \mathcal{T}_h\}$, where $\mathbb{Q}_1(K)$ 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., $\mathcal{N}_h = \Pi_{K\in\mathcal{T}_h}\mathcal{V}_h(K)$. Thus, every node $a \in \mathcal{N}_h$ can also be represented as a pair (i, K), with $K \in \mathcal{T}_h$ 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 $\boldsymbol{x}_i$, there will be n nodes at $\boldsymbol{x}_i$, each one with its own degree of freedom. Given the node $a \in \mathcal{N}_h$, its coordinates are represented with $\boldsymbol{x}_a$, $\Omega_a \doteq \{K \in \mathcal{T}_h : \boldsymbol{x}_a \in K\}$ is its support, and $\mathcal{N}_h(a) = \{b \in \mathcal{N}_h : \boldsymbol{x}_b \in \Omega_a\}$ 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 $\mathcal{N}_h^\partial \doteq \{a \in \mathcal{N}_h : \boldsymbol{x}_a \in \partial\Omega\}$, and $\mathcal{N}_h^\partial(a) \doteq \mathcal{N}_h(a) \cap \mathcal{N}_h^\partial$.

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 φ_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 φ_a as the elementwise (bi)linear function such that $\varphi_a(\boldsymbol{x}_a)|_K = 1$ and $\varphi_a(\boldsymbol{x}_b)|_K = 0$ 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 $\mathcal{E}_h^0$ and single-valued on $\partial\Omega$. Thus, $v_h \in V_h$ can be expressed as $v_h = \sum_{a\in\mathcal{N}_h} v_a \varphi_a$. 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 $\{\!\{\cdot\}\!\}$ and jump $[\![\cdot]\!]$ on an interior point $\boldsymbol{x}$ of a facet $F \in \mathcal{E}_h^0$ as follows:

$$\{\!\{v_h\}\!\}(\boldsymbol{x}) = \frac{1}{2}\left(v_h^{K_F^+}(\boldsymbol{x}) + v_h^{K_F^-}(\boldsymbol{x})\right), \quad [\![v_h]\!](\boldsymbol{x}) = v_h^{K_F^+}(\boldsymbol{x})\boldsymbol{n}_F^+ + v_h^{K_F^-}(\boldsymbol{x})\boldsymbol{n}_F^-,$$

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

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 $\omega \equiv \Omega$. The $L^2(\Omega)$ norm is denoted by $\|\cdot\|$. We will denote by $\mathbb{1}$ the function that is equal to 1 in Ω; $\mathbb{1}(\boldsymbol{x}) = 1$ $\forall \boldsymbol{x} \in \Omega$.

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:

$$\text{Find} \quad u_h \in V_h \quad \text{such that} \quad (\partial_t u_h, v_h) + K_h(u_h, v_h) = G_h(v_h) + B_h(\overline{u}_h; v_h) \quad \forall v_h \in V_h, \tag{2}$$

with

$$\begin{aligned} K_h(u_h, v_h) \doteq & \sum_{K\in\mathcal{T}_h} \int_K (\mu \boldsymbol{\nabla} u_h \cdot \boldsymbol{\nabla} v_h - u_h \boldsymbol{\beta} \cdot \boldsymbol{\nabla} v_h) \\ & + \sum_{F\in\mathcal{E}_h} \int_F \mu \left(-[\![u_h]\!] \cdot \{\!\{\boldsymbol{\nabla} v_h\}\!\} - \{\!\{\boldsymbol{\nabla} u_h\}\!\} \cdot [\![v_h]\!] + c^{\mathrm{ip}} h_F^{-1} [\![u_h]\!] \cdot [\![v_h]\!]\right) \\ & + \sum_{F\in\mathcal{E}_h^+\cup\mathcal{E}_h^0} \int_F \{\!\{\boldsymbol{\beta} u_h\}\!\} \cdot [\![v_h]\!] + \sum_{F\in\mathcal{E}_h^0} \int_F \frac{|\boldsymbol{\beta}\cdot\boldsymbol{n}_F|}{2} [\![u_h]\!] \cdot [\![v_h]\!], \end{aligned}$$

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

$$G_h(v_h) \doteq \sum_{K\in\mathcal{T}_h} (g, v_h)_K,$$

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

$$B_h(w_h; v_h) \doteq -\sum_{F\in\mathcal{E}_h^-} \int_F \boldsymbol{\beta}\cdot\boldsymbol{n}_{\partial\Omega} w_h v_h - \sum_{F\in\mathcal{E}_h^-\cup\mathcal{E}_h^+} \int_F \mu w_h \{\!\{\boldsymbol{\nabla} v_h\}\!\} \cdot \boldsymbol{n}_{\partial\Omega} + \sum_{F\in\mathcal{E}_h^-\cup\mathcal{E}_h^+} \int_F c^{\mathrm{ip}} \mu h_F^{-1} w_h v_h.$$

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 $\mathcal{N}_h^\partial$, i.e., $\bar{u}_h = \sum_{a\in\mathcal{N}_h^\partial} \varphi_a \overline{u}(\boldsymbol{x}_a)$. 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

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

where $\mathsf{M}_{ab} \doteq (\varphi_b, \varphi_a)$ and $\mathsf{K}_{ab} \doteq K_h(\varphi_b, \varphi_a)$, for a, $b \in \mathcal{N}_h$, $\mathsf{G}_a \doteq G_h(\varphi_a)$, for $a \in \mathcal{N}_h$, and $\mathsf{B}_{ab} \doteq B_h(\varphi_b; \varphi_a)$, for $a \in \mathcal{N}_h$, $b \in \mathcal{N}_h^\partial$.

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 extrema diminishing (LED). In particular, we will use the θ-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 $t_n = n\Delta t$, $n = 0, \cdots, N^t$. The problem will be solved by computing an approximation of u in each of those time steps $u_h^n \approx u(\cdot, t_n)$. The discretization of (1) by means of the θ-method reads: Find $u_h^{n+1} \in V_h$ such that

$$\frac{1}{\Delta t}(u_h^{n+1} - u_h^n, v_h) + K_h(\theta u_h^{n+1} + (1-\theta)u_h^n, v_h) = G_h(v_h) + B_h(\overline{u}_h; v_h) \quad \forall v_h \in V_h. \tag{4}$$

We use a projection of the actual initial condition u at $t = 0$ as the initial discrete solution u_h^0, such that it inherits the DMP. The value of θ 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

$$\mathsf{M}\delta_t u_h + \mathsf{K}u_h^{n+1} = \mathsf{G} + \mathsf{B}\overline{u}_h,$$

where $\delta_t u_h = \Delta t^{-1}(u_h^{n+1} - u_h^n)$. Other choices of θ 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 possitive 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 local extremum, but not the opposite. We will see later on that this weaker definition is enough for 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 $a \in \mathcal{N}_h$ if $u_a \geq u_h(\boldsymbol{x})$ (resp. $u_a \leq u_h(\boldsymbol{x})$) $\forall \boldsymbol{x} \in \Omega_a$, and also $u_a \geq \overline{u}(\boldsymbol{x})$ (resp. $u_a \leq \overline{u}(\boldsymbol{x})$) $\forall \boldsymbol{x} \in \partial\Omega_a \cap \partial\Omega$.*

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 $a \in \mathcal{N}_h$, we have:*

$$u_a^{\min} \leq u_a \leq u_a^{\max}, \quad \text{where } u_a^{\max} \doteq \max\left\{ \max_{b \in \mathcal{N}_h(a)\setminus\{a\}} u_b, \max_{\boldsymbol{x} \in \partial\Omega_a \cap \partial\Omega_D} \overline{u}_h(\boldsymbol{x}) \right\},$$
$$\text{and } u_a^{\min} \doteq \min\left\{ \min_{b \in \mathcal{N}_h(a)\setminus\{a\}} u_b, \min_{\boldsymbol{x} \in \partial\Omega_a \cap \partial\Omega_D} \overline{u}_h(\boldsymbol{x}) \right\},$$

where $\overline{u}_h \in V_h|_{\partial\Omega}$ is the finite element interpolation of the boundary conditions on the Dirichlet boundary $\partial\Omega_D$. In the case of transient problems, $u_a^{\max}$ and $u_a^{\min}$ are defined as

$$u_a^{\max} \doteq \max\left\{ \max_{\mathcal{N}_h(a)\setminus\{a\}} u_b, \max_{\boldsymbol{x} \in \partial\Omega_a \cap \partial\Omega_D \times (0,T]} \overline{u}(\boldsymbol{x}), \max_{\boldsymbol{x} \in \Omega} u_h^0(\boldsymbol{x}) \right\}$$
$$u_a^{\min} \doteq \min\left\{ \min_{\mathcal{N}_h(a)\setminus\{a\}} u_b, \min_{\boldsymbol{x} \in \partial\Omega_a \cap \partial\Omega_D \times (0,T]} \overline{u}(\boldsymbol{x}), \min_{\boldsymbol{x} \in \Omega} u_h^0(\boldsymbol{x}) \right\},$$

where $u_h^0 \in V_h$ 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*

$$\mathrm{d}_t u_a \leq 0 \textit{ if } u_a \textit{ is a maximum and } \mathrm{d}_t u_a \geq 0 \textit{ if } u_a \textit{ is a minimum.}$$

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

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

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

The superscript, e.g., in $\tilde{\mathbf{K}}$, denotes the fact that the operator $\tilde{\mathbf{K}}$ is equal to $\mathbf{K}$ plus stabilization terms. We have written, e.g., $\tilde{\mathbf{K}}(u_h, \overline{u}_h)$, to explicitly denote the fact that the entries of the matrix $\tilde{\mathbf{K}}$ are potentially nonlinear with respect to u_h and $\overline{u}_h$. 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 $a \in \mathcal{N}_h$ such that u_a is a local extremum, it holds:*

$$\tilde{\mathbf{M}}_{ab}(u_h, \overline{u}_h) \doteq \delta_{ab} m_a, \textit{ with } m_a > 0, \tag{6a}$$

$$\tilde{\mathbf{K}}_{ab}(u_h, \overline{u}_h) \leq 0, \, \forall\, b \in \mathcal{N}_h \,:\, b \neq a, \qquad \tilde{\mathbf{B}}_{ab}(u_h, \overline{u}_h) \geq 0, \quad \forall\, b \in \mathcal{N}_h^{\partial}, \tag{6b}$$

$$\textstyle\sum_{b\in\mathcal{N}_h(a)} \tilde{\mathbf{K}}_{ab}(u_h, \overline{u}_h) - \sum_{b\in\mathcal{N}_h^{\partial}(a)} \tilde{\mathbf{B}}_{ab}(u_h, \overline{u}_h) = 0, \tag{6c}$$

where $m_a \doteq \int_\Omega \varphi_a d\Omega$, δ_{ab} is the Kronecker delta. Further, for $g \leq 0$ (resp. $g \geq 0$) in Ω 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) to node a, we have that

$$\begin{aligned} \mathbf{G}_a &= m_a \mathrm{d}_t u_a + \sum_{b\in\mathcal{N}_h(a)} \tilde{\mathbf{K}}_{ab}(u_h, \overline{u}_h)u_b - \sum_{b\in\mathcal{N}_h^{\partial}(a)} \tilde{\mathbf{B}}_{ab}(u_h, \overline{u}_h)\overline{u}_b \\ &\geq m_a \mathrm{d}_t u_a + \left(\sum_{b\in\mathcal{N}_h(a)} \tilde{\mathbf{K}}_{ab}(u_h, \overline{u}_h) - \sum_{b\in\mathcal{N}_h^{\partial}(a)} \tilde{\mathbf{B}}_{ab}(u_h, \overline{u}_h) \right) u_a = m_a \mathrm{d}_t u_a. \end{aligned}$$

Therefore, $\mathrm{d}_t u_a \leq \mathbf{G}_a = 0$. 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 $u_a^m \geq u_a^{m-1}$. 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 $n = 1, ...N^t$.*

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_h^0 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 Ω and, for every degree of freedom $a \in \mathcal{N}_h$ 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

$$\sum_{b \in \mathcal{N}_h(a)} \tilde{\mathsf{K}}_{ab}(u_h, \overline{u}_h) u_b - \sum_{b \in \mathcal{N}_h^\partial(a)} \tilde{\mathsf{B}}_{ab}(u_h, \overline{u}_h) \overline{u}_b = 0,$$

Therefore, u_a can be computed as

$$u_a = \frac{\sum_{b \in \mathcal{N}_h^\partial(a)} \tilde{\mathsf{B}}_{ab}(u_h, \overline{u}_h)\overline{u}_b - \sum_{b \in \mathcal{N}_h(a)\setminus\{a\}} \tilde{\mathsf{K}}_{ab}(u_h, \overline{u}_h) u_b}{\tilde{\mathsf{K}}_{aa}(u_h, \overline{u}_h)}.$$

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 $\overline{u}_h$). Since u_a is a maximum and a convex combination of its neighbors, then $u_b = u_a$ for some $b \in \mathcal{N}_h(a)$. 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 extrema 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 LED. 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 ν_{ab}. For $a \in \mathcal{N}_h$ and $b \in \mathcal{N}_h^\partial(a)$ we define

$$\nu_{ab}^\partial \doteq \max\{-\alpha_a B_h(\varphi_b; \varphi_a), 0\}. \tag{7}$$

Clearly, this viscosity is only non-zero when $a \in \mathcal{N}_h^\partial$. Next, for $a \in \mathcal{N}_h$ and $b \in \mathcal{N}_h(a)$, we define

$$\nu_{ab} \doteq \begin{cases} \max\{\alpha_a K_h(\varphi_b, \varphi_a), 0, \alpha_b K_h(\varphi_a, \varphi_b)\} & b \neq a, \\ \sum_{b \in \mathcal{N}_h(a)\setminus a} \nu_{ab} + \sum_{b \in \mathcal{N}_h^\partial(a)} \nu_{ab}^\partial & \text{otherwise}, \end{cases} \tag{8}$$

where α_a is a parameter that enjoys the following property.

Definition 4.1. *Given $a \in \mathcal{N}_h$, we say that $\alpha_a : V_h \longrightarrow \mathbb{R}$ enjoys the* shock detector property *if it is such that $\alpha_a(u_h, \overline{u}_h) \in [0, 1]$ $\forall u_h \in V_h$ and $\alpha_a(u_h, \overline{u}_h) = 1$ if u_h has a local discrete extremum on $\boldsymbol{x}_a$.*

Next, we design a shock detector that satisfies this property. Given $a \in \mathcal{N}_h$ and $b \in \mathcal{N}_h(a)$ with $\boldsymbol{x}_b \neq \boldsymbol{x}_a$, we define $\boldsymbol{x}_{ab}^{\mathrm{sym}}$ as the intersection between $\partial\Omega_a$ and the line that passes through $\boldsymbol{x}_b$ and $\boldsymbol{x}_a$, and it is not $\boldsymbol{x}_b$. Moreover, we define $\boldsymbol{r}_{ab} \doteq \boldsymbol{x}_b - \boldsymbol{x}_a$, $\boldsymbol{r}_{ab}^{\mathrm{sym}} \doteq \boldsymbol{x}_{ab}^{\mathrm{sym}} - \boldsymbol{x}_a$, and $u_{ab}^{\mathrm{sym}} \doteq u_h(\boldsymbol{x}_{ab}^{\mathrm{sym}})$ (see Fig. 1). Further, we denote by $\hat{\boldsymbol{r}}_{ab}$ the unit vector of $\boldsymbol{r}_{ab}$, and by h_a a characteristic length of Ω_a. Then, we define the jump and the mean of the unknown gradients as

$$[\![\boldsymbol{\nabla} u_h]\!]_{ab} \doteq \begin{cases} \dfrac{u_b - u_a}{h_a} & \text{if } \boldsymbol{x}_a = \boldsymbol{x}_b, \\ \dfrac{u_b - u_a}{|\boldsymbol{r}_{ab}|} + \dfrac{u_{ab}^{\mathrm{sym}} - u_a}{|\boldsymbol{r}_{ab}^{\mathrm{sym}}|} & \text{otherwise}, \end{cases}$$

$$\{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ab}|\}\!\}_{ab} \doteq \begin{cases} \dfrac{|u_b - u_a|}{h_a} & \text{if } \boldsymbol{x}_a = \boldsymbol{x}_b, \\ \dfrac{1}{2}\left(\dfrac{|u_b - u_a|}{|\boldsymbol{r}_{ab}|} + \dfrac{|u_{ab}^{\mathrm{sym}} - u_a|}{|\boldsymbol{r}_{ab}^{\mathrm{sym}}|}\right) & \text{otherwise}. \end{cases}$$

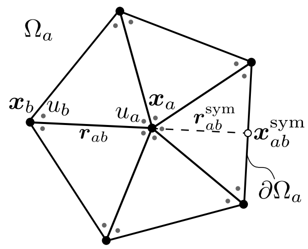


FIGURE 1. Representation of the symmetric node $\boldsymbol{x}^{\mathrm{sym}}_{ab}$ of $\boldsymbol{x}_b$ with respect to $\boldsymbol{x}_a$.

Remark 4.2. *These definitions may imply $\boldsymbol{x}^{\mathrm{sym}}_{ab} = \boldsymbol{x}_a$ on some boundaries. In these cases, the value at the symmetric point of $\boldsymbol{x}_b$ with respect to $\boldsymbol{x}_a$ takes an extrapolation of the boundary condition value such that the method is linearly preserving, i.e., $u^{\mathrm{sym}}_{ab} = \overline{u}_a + (u_b - u_a)\,\mathrm{sign}((\overline{u}_a - u_a)(u_b - u_a))$, and the value $|\boldsymbol{r}^{\mathrm{sym}}_{ab}|$ is taken equal to $|\boldsymbol{r}_{ab}|$. 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 $\boldsymbol{x}^{\mathrm{sym}}_{ab}$ coincides with a node the value $u_h(\boldsymbol{x}^{\mathrm{sym}}_{ab})$ is not unique. In this case, we compute $[\![\boldsymbol{\nabla} u_h]\!]_{ab}$ and $\{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ab}|\}\!\}_{ab}$ for all values of $u_h(\boldsymbol{x}^{\mathrm{sym}}_{ab})$ in Ω_a.*

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

$$\alpha_a(u_h, \overline{u}_h) \doteq \begin{cases} \left(\dfrac{\left|\sum_{b\in\mathcal{N}_h(a)} [\![\boldsymbol{\nabla} u_h]\!]_{ab}\right|}{\sum_{b\in\mathcal{N}_h(a)} 2\{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ab}|\}\!\}_{ab}} \right)^q & \text{if } \sum_{b\in\mathcal{N}_h(a)} \{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ab}|\}\!\}_{ab} \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

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 α_a is motivated from [3]. Here, instead of using the maximum coefficient obtained, we use the sum all gradient jumps divided by the sum of all gradient means. A similar (more involved) modification can be found in [13].

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

Proof. Let us assume that u_h has a discrete maximum (resp. minimum) on $\boldsymbol{x}_a$, then

$$\begin{array}{lll} u_b - u_a \leq 0 \quad \forall b \in \mathcal{N}_h(a) & \text{and} & u^{\mathrm{sym}}_{ab} - u_a \leq 0 \quad \forall b \in \mathcal{N}_h(a), \\ (\text{resp. } u_b - u_a \geq 0 \quad \forall b \in \mathcal{N}_h(a) & \text{and} & u^{\mathrm{sym}}_{ab} - u_a \geq 0 \quad \forall b \in \mathcal{N}_h(a)). \end{array} \tag{10}$$

Therefore,

$$\begin{aligned} \left| \sum_{b\in\mathcal{N}_h(a)} [\![\boldsymbol{\nabla} u_h]\!]_{ab} \right| &= \left| \sum_{\substack{b\in\mathcal{N}_h(a)\\ \boldsymbol{x}_b=\boldsymbol{x}_a}} [\![\boldsymbol{\nabla} u_h]\!]_{ab} + \sum_{\substack{b\in\mathcal{N}_h(a)\\ \boldsymbol{x}_b\neq\boldsymbol{x}_a}} [\![\boldsymbol{\nabla} u_h]\!]_{ab} \right| \\ &= \left| \sum_{\substack{b\in\mathcal{N}_h(a)\\ \boldsymbol{x}_b=\boldsymbol{x}_a}} \frac{u_b - u_a}{h_a} + \sum_{\substack{b\in\mathcal{N}_h(a)\\ \boldsymbol{x}_b\neq\boldsymbol{x}_a}} \frac{u_b - u_a}{|\boldsymbol{r}_{ab}|} + \frac{u^{\mathrm{sym}}_{ab} - u_a}{|\boldsymbol{r}^{\mathrm{sym}}_{ab}|} \right| \\ &= \sum_{\substack{b\in\mathcal{N}_h(a)\\ \boldsymbol{x}_b=\boldsymbol{x}_a}} \frac{|u_b - u_a|}{h_a} + \sum_{\substack{b\in\mathcal{N}_h(a)\\ \boldsymbol{x}_b\neq\boldsymbol{x}_a}} \frac{|u_b - u_a|}{|\boldsymbol{r}_{ab}|} + \frac{|u^{\mathrm{sym}}_{ab} - u_a|}{|\boldsymbol{r}^{\mathrm{sym}}_{ab}|} = \sum_{b\in\mathcal{N}_h(a)} 2\{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ab}|\}\!\}_{ab}. \end{aligned}$$

Thus, $\alpha_a(u_h, \overline{u}_h) = 1$. Further, if u_a is not an extremum, then (10) is no longer true. Hence,

$$
\begin{aligned}
\left| \sum_{b \in \mathcal{N}_h(a)} [\![\boldsymbol{\nabla} u_h]\!]_{ab} \right| &= \left| \sum_{\substack{b \in \mathcal{N}_h(a) \\ \boldsymbol{x}_b = \boldsymbol{x}_a}} [\![\boldsymbol{\nabla} u_h]\!]_{ab} + \sum_{\substack{b \in \mathcal{N}_h(a) \\ \boldsymbol{x}_b \neq \boldsymbol{x}_a}} [\![\boldsymbol{\nabla} u_h]\!]_{ab} \right| \\
&= \left| \sum_{\substack{b \in \mathcal{N}_h(a) \\ \boldsymbol{x}_b = \boldsymbol{x}_a}} \frac{u_b - u_a}{h_a} + \sum_{\substack{b \in \mathcal{N}_h(a) \\ \boldsymbol{x}_b \neq \boldsymbol{x}_a}} \frac{u_b - u_a}{|\boldsymbol{r}_{ab}|} + \frac{u_{ab}^{\mathrm{sym}} - u_a}{|\boldsymbol{r}_{ab}^{\mathrm{sym}}|} \right| \\
&< \sum_{\substack{b \in \mathcal{N}_h(a) \\ \boldsymbol{x}_b = \boldsymbol{x}_a}} \frac{|u_b - u_a|}{h_a} + \sum_{\substack{b \in \mathcal{N}_h(a) \\ \boldsymbol{x}_b \neq \boldsymbol{x}_a}} \frac{|u_b - u_a|}{|\boldsymbol{r}_{ab}|} + \frac{|u_{ab}^{\mathrm{sym}} - u_a|}{|\boldsymbol{r}_{ab}^{\mathrm{sym}}|} = \sum_{b \in \mathcal{N}_h(a)} 2 \{\!\{ |\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ab}| \}\!\}_{ab}.
\end{aligned}
$$

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:

$$
\tilde{B}_h(w_h, \overline{w}_h, \overline{u}_h; v_h) \doteq B_h(\overline{u}_h; v_h) + \sum_{a \in \mathcal{N}_h} \sum_{b \in \mathcal{N}_h^{\partial}(a)} \nu_{ab}^{\partial}(w_h, \overline{w}_h) v_a \overline{u}_b. \tag{11}
$$

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

$$
\tilde{K}_h(w_h, \overline{w}_h; u_h, v_h) \doteq K_h(u_h, v_h) + \sum_{a \in \mathcal{N}_h} \sum_{b \in \mathcal{N}_h(a)} \nu_{ab}(w_h, \overline{w}_h) v_a u_b \ell(a, b), \tag{12}
$$

where $\ell(a, b) \doteq 2\delta_{ab} - 1$ 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

$$
\tilde{K}_h(u_h, \overline{u}_h; u_h, v_h) = G_h(v_h) + \tilde{B}_h(u_h, \overline{u}_h, \overline{u}_h; v_h) \quad \forall v_h \in V_h. \tag{13}
$$

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

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 $\boldsymbol{x}_a$ be an interior node and assume that u_h has an extremum on $\boldsymbol{x}_a$.

Given the set of all nodes $b \in \mathcal{N}_h(a)$ coupled to node $a \in \mathcal{N}_h$, we have:

$$\begin{aligned}
\sum_{b\in\mathcal{N}_h(a)} \mathsf{K}_{ab} - \mathsf{B}_{ab} = \sum_{b\in\mathcal{N}_h(a)} \Bigg\{ & \sum_{K\in\mathcal{T}_h} \int_K (\mu \boldsymbol{\nabla}\varphi_b \cdot \boldsymbol{\nabla}\varphi_a - \varphi_b \boldsymbol{\beta}\cdot\boldsymbol{\nabla}\varphi_a) \\
& + \sum_{F\in\mathcal{E}_h} \int_F \mu \left(-[\![\varphi_b]\!]\cdot \{\!\{\boldsymbol{\nabla}\varphi_a\}\!\} - \{\!\{\boldsymbol{\nabla}\varphi_b\}\!\}\cdot[\![\varphi_a]\!] + c^{\mathrm{ip}} h_F^{-1} [\![\varphi_b]\!]\cdot[\![\varphi_a]\!] \right) \\
& + \sum_{F\in\mathcal{E}_h^+\cup\mathcal{E}_h^0} \int_F \boldsymbol{\beta}\{\!\{\varphi_b\}\!\}\cdot[\![\varphi_a]\!] + \sum_{F\in\mathcal{E}_h^0}\int_F \frac{|\boldsymbol{\beta}\cdot\boldsymbol{n}_F|}{2}\cdot[\![\varphi_b]\!]\cdot[\![\varphi_a]\!] \\
& + \sum_{F\in\mathcal{E}_h^-} \int_F \boldsymbol{\beta}\cdot\boldsymbol{n}_F \varphi_b\varphi_a + \sum_{F\in\mathcal{E}_h^-\cup\mathcal{E}_h^+} \int_F \mu\varphi_b \{\!\{\boldsymbol{\nabla}\varphi_a\}\!\}\cdot\boldsymbol{n}_F \\
& - \sum_{F\in\mathcal{E}_h^-\cup\mathcal{E}_h^+} \int_F c^{\mathrm{ip}}\mu h_F^{-1}\varphi_b\varphi_a \Bigg\}.
\end{aligned}$$

(We note that B_{ab} has only been defined for nodes $b \in \mathcal{N}_h^{\partial}$. Here, we abuse of notation, and extend by zero the definition to all nodes, with $\mathsf{B}_{ab} = 0$ when $b \notin \mathcal{N}_h$.) 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 Ω_a and zero elsewhere. As a result, $\sum_{b\in\mathcal{N}_h(a)}[\![\varphi_b]\!] = 0$ on facets $F \subset \Omega_a \setminus \partial\Omega_a$, and $\sum_{b\in\mathcal{N}_h(a)} \boldsymbol{\nabla}\varphi_b = 0$ in any $K \in \mathcal{T}_h$. On the other hand, φ_a vanishes on any $F \subset \partial\Omega_a \setminus \partial\Omega$ by construction. Using these properties, we get:

$$\begin{aligned}
& \sum_{b\in\mathcal{N}_h(a)} \sum_{F\in\mathcal{E}_h^0} \int_F \frac{|\boldsymbol{\beta}\cdot\boldsymbol{n}_F|}{2}[\![\varphi_b]\!]\cdot[\![\varphi_a]\!] = 0, \\
& \sum_{b\in\mathcal{N}_h(a)} \left\{ \sum_{F\in\mathcal{E}_h} \int_F \mu c^{\mathrm{ip}} h_F^{-1}[\![\varphi_b]\!]\cdot[\![\varphi_a]\!] - \sum_{F\in\mathcal{E}_h^-\cup\mathcal{E}_h^+}\int_F c^{\mathrm{ip}}\mu h_F^{-1}\varphi_b\varphi_a \right\} = 0, \\
& \sum_{b\in\mathcal{N}_h(a)} \left\{ \sum_{F\in\mathcal{E}_h} \int_F -\mu[\![\varphi_b]\!]\cdot\{\!\{\boldsymbol{\nabla}\varphi_a\}\!\} + \sum_{F\in\mathcal{E}_h^-\cup\mathcal{E}_h^+}\int_F \mu\varphi_b\{\!\{\boldsymbol{\nabla}\varphi_a\}\!\}\cdot\boldsymbol{n}_F \right\} = 0, \\
& \sum_{b\in\mathcal{N}_h(a)} \sum_{F\in\mathcal{E}_h} \int_F \mu\{\!\{\boldsymbol{\nabla}\varphi_b\}\!\}\cdot[\![\varphi_a]\!] = 0, \qquad \sum_{b\in\mathcal{N}_h(a)} \int_\Omega \mu\boldsymbol{\nabla}\varphi_b\cdot\boldsymbol{\nabla}\varphi_a = 0,
\end{aligned}$$

and the following terms can be integrated by parts as

$$\begin{aligned}
& \sum_{b\in\mathcal{N}_h(a)} \left\{ -\sum_{K\in\mathcal{T}_h}\int_K \varphi_b\boldsymbol{\beta}\cdot\boldsymbol{\nabla}\varphi_a + \sum_{F\in\mathcal{E}_h^+\cup\mathcal{E}_h^0}\int_F \boldsymbol{\beta}\{\!\{\varphi_b\}\!\}\cdot[\![\varphi_a]\!] + \sum_{F\in\mathcal{E}_h^-}\int_F \boldsymbol{\beta}\cdot\boldsymbol{n}_F\varphi_b\varphi_a \right\} \\
& = \sum_{b\in\mathcal{N}_h(a)} \left\{ -\sum_{K\in\mathcal{T}_h} \left(\int_K \varphi_b\boldsymbol{\beta}\cdot\boldsymbol{\nabla}\varphi_a + \int_{\partial K} \boldsymbol{\beta}\cdot\boldsymbol{n}\,\varphi_b\varphi_a \right) \right\} \\
& = \sum_{b\in\mathcal{N}_h(a)} \sum_{K\in\mathcal{T}_h} \int_K \varphi_a\boldsymbol{\beta}\cdot\boldsymbol{\nabla}\varphi_b = 0.
\end{aligned}$$

Finally, since $\sum_{b\in\mathcal{N}_h(a)} \nu_{ab}(u_h)v_a v_b \ell(a,b) + \sum_{b\in\mathcal{N}_h^{\partial}(a)} \nu_{ab}^{\partial}(u_h)v_a v_b = 0$ by construction (see (7) and (8)), then $\sum_{b\in\mathcal{N}_h(a)} \tilde{\mathsf{K}}_{ab}(u_h,\overline{u}_h) - \tilde{\mathsf{B}}_{ab}(u_h,\overline{u}_h) = 0$. Moreover, it is clear that $\tilde{\mathsf{K}}_{ab}(u_h,\overline{u}_h) \le 0$ for any $b \neq a$ and $\tilde{\mathsf{B}}_{ab}(u_h,\overline{u}_h) \ge 0$ 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 $(\partial_t u_h, \varphi_a)_h = \partial_t u_a(\mathbb{1}, \varphi_a)$ for any $a \in \mathcal{N}_h$. In fact, as Kuzmin and co-workers have proved in [22, 23], 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:

$$M_h^L(u_h, \overline{u}_h; \partial_t u_h, v_h) \doteq \sum_{a\in\mathcal{N}_h} v_a(1 - \alpha_a^Q(u_h, \overline{u}_h))(\partial_t u_h, \varphi_a) + \alpha_a^Q(u_h, \overline{u}_h)\partial_t u_a v_a(\mathbb{1}, \varphi_a). \tag{14}$$

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 $\tilde{\mathbf{M}}_{ab}(u_h, \overline{u}_h) \doteq M_h^L(u_h, \overline{u}_h; \varphi_b, \varphi_a)$, 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

$$M_h^L(u_h, \overline{u}_h; \partial_t u_h, v_h) + \tilde{K}_h(u_h, \overline{u}_h; u_h, v_h) = G_h(v_h) + \tilde{B}_h(u_h, \overline{u}_h, \overline{u}_h; v_h) \quad \forall v_h \in V_h. \tag{15}$$

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 $\tilde{\mathbf{K}}(u_h, \overline{u}_h)$ and $\tilde{\mathbf{B}}(u_h, \overline{u}_h)$ 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 $\alpha_a(u_h, \overline{u}_h) = 1$ and $\tilde{\mathbf{M}}_{ab}(u_h)$ becomes $(\mathbb{1}, \varphi_a) = \delta_{ab} m_a$ 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. $\square$

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.,*

$$\tilde{B}_h(u, u_{\partial\Omega}, \overline{u}_h; v_h) = B_h(\overline{u}_h; v_h), \qquad \tilde{K}_h(u, u_{\partial\Omega}; u_h, v_h) = K_h(u_h, v_h),$$
$$M_h^L(u, u_{\partial\Omega}; u_h, v_h) = (u_h, v_h), \qquad \textit{for any } u_h \in V_h.$$

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

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

$$\Delta t \leq \min_{a\in\mathcal{N}_h} \frac{(\mathbb{1}, \varphi_a)}{(1-\theta)\tilde{K}_h(u_h^{n+1}, \overline{u}_h^{n+1}; \varphi_a, \varphi_a)}.$$

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 θ-method.

The authors in [19, Th. 1] take advantage of the mass lumping properties for all of this 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 $\tilde{V}_h \doteq V_h \times V_h|_{\partial\Omega}$. Thus, any function $\tilde{v} \in \tilde{V}_h$ can be expressed as $\tilde{v} = (v, \bar{v})$, 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 $\bar{v} \in V_h|_{\partial\Omega}$. Analogously, we can define the set of nodes for $\tilde{V}_h$ as $\mathcal{M}_h \doteq \mathcal{N}_h \times \mathcal{N}_h^{\partial} \equiv \{(a_1, 0), (0, a_2) : a_1 \in \mathcal{N}_h, a_2 \in \mathcal{N}_h^{\partial}\}$. We consider an extended graph-Laplacian operator over $\tilde{V}_h \times \tilde{V}_h$ as follows:

$$
\begin{aligned}
\tilde{D}(\tilde{u}, \tilde{v}) = \tilde{D}((u, \bar{u}), (v, \bar{v})) &\doteq \sum_{a\in\mathcal{M}_h} \sum_{b\in\mathcal{M}_h} \tilde{\nu}_{ab}\ell(a,b)\tilde{u}_b\tilde{v}_a \\
&\doteq \sum_{a\in\mathcal{N}_h} \sum_{b\in\mathcal{N}_h} \nu_{ab}\ell(a,b)u_b v_a - \sum_{a\in\mathcal{N}_h^{\partial}} \sum_{b\in\mathcal{N}_h^{\partial}} \nu_{ab}^{\partial}\bar{u}_b v_a - \sum_{a\in\mathcal{N}_h^{\partial}} \sum_{b\in\mathcal{N}_h^{\partial}} \nu_{ba}^{\partial}u_b\bar{v}_a.
\end{aligned}
$$

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

$$
\begin{array}{ll}
\tilde{\nu}_{ab} = \nu_{ab}, & \text{if } a, b \in (\mathcal{N}_h, 0), \\
\tilde{\nu}_{ab} = \nu_{ab}^{\partial}, & \text{if } a \in (\mathcal{N}_h^{\partial}, 0),\, b \in (0, \mathcal{N}_h^{\partial}), \\
\tilde{\nu}_{ab} = \nu_{ba}^{\partial}, & \text{if } a \in (0, \mathcal{N}_h^{\partial}),\, b \in (\mathcal{N}_h^{\partial}, 0), \\
\tilde{\nu}_{ab} = 0, & \text{if } a, b \in (0, \mathcal{N}_h^{\partial}).
\end{array}
$$

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 ν_{ab} and ν_{ab}^{∂} in (7) and (8), respectively, in order to get $\nu_{ab}, \nu_{ab}^{\partial} \geq 0$, and

$$
\tilde{\nu}_{aa} = \nu_{aa} = \sum_{b\in\mathcal{N}_h(a)\backslash a} \nu_{ab} + \sum_{b\in\mathcal{N}_h^{\partial}(a)} \nu_{ab}^{\partial} = \sum_{b\in\mathcal{M}_h(a)\backslash a} \tilde{\nu}_{ab}.
$$

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

$$
\begin{aligned}
2\tilde{D}(\tilde{u}, \tilde{v}) &= \sum_{a\in\mathcal{M}_h} \sum_{b\in\mathcal{M}_h} \tilde{\nu}_{ab}\ell(a,b)\tilde{v}_a(\tilde{u}_b - \tilde{u}_a) + \sum_{a\in\mathcal{M}_h} \sum_{b\in\mathcal{M}_h} \tilde{\nu}_{ab}\ell(a,b)\tilde{u}_b(\tilde{v}_a - \tilde{v}_b) \\
&= \sum_{a\in\mathcal{M}_h} \sum_{b\in\mathcal{M}_h} \tilde{\nu}_{ab}\ell(a,b)(\tilde{u}_b - \tilde{u}_a)(\tilde{v}_a - \tilde{v}_b).
\end{aligned}
\tag{16}
$$

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

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

$$
\begin{aligned}
\langle \mathbf{T}(z), v\rangle &\doteq K_h(z, v) - B_h(\bar{u}, v) - G_h(v) + \tilde{D}((z, \bar{u}), (v, 0)) \\
&= K_h(z, v) - B_h(\bar{u}, v) - G_h(v) + \sum_{a\in\mathcal{N}_h} \sum_{b\in\mathcal{N}_h(a)} \nu_{ab}\ell(a,b)v_a z_b \\
&\quad - \sum_{a\in\mathcal{N}_h} \sum_{b\in\mathcal{N}_h^{\partial}(a)} \nu_{ab}^{\partial} v_a \bar{u}_b.
\end{aligned}
\tag{17}
$$

Clearly, to find $u_h \in V_h$ such that $\mathbf{T}(u_h) = 0$ 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 $\mathbf{T}^{n+1} : V_h \to V_h'$ at every time step as

$$
\langle \mathbf{T}^{n+1}(z), v\rangle \doteq M_h^L(z; z, v) - M_h^L(z; u_h^n, v) + \langle \mathbf{T}(z), v\rangle.
$$

System (15) can be stated in compact form as: find $u_h \in V_h$ such that $\mathbf{T}^{n+1}(u_h) = 0$. 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 $\mathbf{T}$ and $\mathbf{T}^{n+1}$ 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:

$$\begin{aligned}|\langle \mathbf{T}(u),w\rangle - \langle \mathbf{T}(v),w\rangle| \leq & |K_h(u-v,v)| + \left|\sum_{a\in\mathcal{N}_h}\sum_{b\in\mathcal{N}_h(a)} \nu_{ab}(v)\ell(a,b)w_a(u_b-v_b)\right| \\ & + \left|\sum_{a\in\mathcal{N}_h}\sum_{b\in\mathcal{N}_h(a)} (\nu_{ab}(u)-\nu_{ab}(v))\ell(a,b)w_a u_b\right| \\ & + \left|\sum_{a\in\mathcal{N}_h}\sum_{b\in\mathcal{N}_h^\partial(a)} (\nu_{ab}^\partial(u)-\nu_{ab}^\partial(v))w_a \bar{u}_b\right|.\end{aligned}$$

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:

$$\bar{K}_h(u_h,\bar{u}_h;\varphi_b,\varphi_a) - \bar{B}_h(u_h,\bar{u}_h;\varphi_b,\varphi_a) \leq Cq(h^{d-1}\|\boldsymbol{\beta}\|_{L^\infty(\Omega)} + \mu h^{d-2}). \tag{18}$$

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 $\boldsymbol{x}_a \neq \boldsymbol{x}_b$ are as in [2], using (18). The case $\boldsymbol{x}_a = \boldsymbol{x}_b$ 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:

$$K_h(z,z) \geq C\|z\|_h^2, \quad \text{with } \|z\|_h^2 \doteq \sum_{K\in\mathcal{T}_h} \mu|z|^2_{H^1(K)} + \sum_{F\in\mathcal{E}_h}\left(\mu c^{\text{ip}} h_F^{-1}\|[\![z]\!]\|^2_{L^2(F)} + \|c^{\frac12}_{\beta,F}[\![z]\!]\|^2_{L^2(F)}\right), \tag{19}$$

with c^{ip} big enough, and $c_{\beta,F}(\boldsymbol{x}) \doteq |\beta(\boldsymbol{x})\cdot\boldsymbol{n}_F(\boldsymbol{x})|$. On the other hand, using standard dG arguments (see [1] and [15, Prop. 3.55]), we have:

$$B_h(\bar{u},z) \leq Cc^{-1}\|c_e^{\frac12}\bar{u}\|^2_{L^2(\partial\Omega^-)} + Cc^{-1}h^{-1}\mu c^{\text{ip}}\|\bar{u}\|^2_{L^2(\partial\Omega)} + c\|z\|_h^2, \tag{20}$$

for c arbitrarily small.

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

$$\sum_{a\in\mathcal{N}_h}\sum_{b\in\mathcal{N}_h(a)} \nu_{ab}\ell(a,b)v_a z_b - \sum_{a\in\mathcal{N}_h}\sum_{b\in\mathcal{N}_h^\partial(a)} \nu^\partial_{ab} v_a \bar{u}_b = \tilde{D}((z,\bar{u}),(v,0)).$$

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

$$\tilde{D}((z,\bar{u}),(z,0)) \leq \frac{3}{2}|z|^2_D + \frac{1}{2}|\bar{u}|^2_{D^\partial}.$$

Combining (17), (19), and (20) (with c small enough), we finally obtain:

$$C\langle \mathbf{T}(z), z\rangle \geq \|z\|_h^2 + |z|_D^2 + |\bar{u}|_{D^\partial}^2 - \|c_e^{\frac{1}{2}}\bar{u}\|_{L^2(\partial\Omega^-)}^2 - h^{-1}\mu c^{\mathrm{ip}}\|\bar{u}\|_{L^2(\partial\Omega)}^2.$$

We can readily pick a $z \in V_h$ such that $\langle \mathbf{T}(z), z\rangle > 0$. Using the Brower's fixed point theorem, there exists $u_h \in V_h$ such that $\mathbf{T}(u_h) = 0$, 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 $\nu_{ab}^\partial(u_h,\bar{u}_h)$, $\nu_{ab}(u_h,\bar{u}_h)$, and $\alpha_a(u_h,\bar{u}_h)$ 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 (τ_h,γ_h,σ_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:

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

The value of τ_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 $\{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ab}|\}\!\}_{ab}$ as:

$$\{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ab}|_{2,\tau_h}\}\!\}_{ab} \doteq \begin{cases} \dfrac{|u_b-u_a|_{2,\tau_h}}{h_a} & \text{if } \boldsymbol{x}_a = \boldsymbol{x}_b, \\ \dfrac{1}{2}\left(\dfrac{|u_b-u_a|_{2,\tau_h}}{|\boldsymbol{r}_{ab}|} + \dfrac{|u_{ab}^{\mathrm{sym}}-u_a|_{2,\tau_h}}{|\boldsymbol{r}_{ab}^{\mathrm{sym}}|}\right) & \text{otherwise.} \end{cases}$$

The quotient associated to α_a would read:

$$\zeta_a = \frac{\left|\sum_{b\in\mathcal{N}_h(a)}[\![\boldsymbol{\nabla} u_h]\!]_{ab}\right|_{1,\tau_h} + \gamma_h}{\sum_{b\in\mathcal{N}_h(a)} 2\{\!\{|\boldsymbol{\nabla} u_h \cdot \hat{\boldsymbol{r}}_{ab}|_{2,\tau_h}\}\!\}_{ab} + \gamma_h}.$$

Here γ_h is another extra stability parameter added to ensure differentiability of ζ_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 α_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 α_a as $\tilde{\alpha}_a \doteq (Z(\zeta_a))^q$. 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 $\mathrm{sign}_{\tau_h}(x) \doteq x/\left|x\right|_{1,\tau_h}$. Then the smooth definition of ν_{ab} in (8) for $a \in \mathcal{N}_h$ and $b \in \mathcal{N}_h(a) \setminus \{a\}$ will read

$$\tilde{\nu}_{ab} = \max\nolimits_{\sigma_h}\left(0, \max\nolimits_{\sigma_h}\left(\tilde{\alpha}_a K_h\left(\varphi_b, \varphi_a\right), \tilde{\alpha}_b K_h\left(\varphi_a, \varphi_b\right)\right)\right),$$

and for $b \in \mathcal{N}_h^{\partial}(a)$

$$\nu_{ab}^{\partial} \doteq \max\nolimits_{\sigma_h}(-\tilde{\alpha}_a B_h(\varphi_b; \varphi_a), 0).$$

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:

$$\sigma_h = \sigma|\boldsymbol{\beta}|^2 L^{2(d-3)} h^4, \qquad \tau_h = \tau h^2 L^{-4}, \qquad \gamma_h = \gamma L^{-1},$$

where d is the space dimension of the problem, L a characteristic length of the problem, τ and γ have the same dimension as the unknown, and σ 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 $\boldsymbol{\beta} = (y, -x)$:

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\beta} u) = 0 \qquad \text{in } [0,1] \times [0,1]. \tag{21}$$

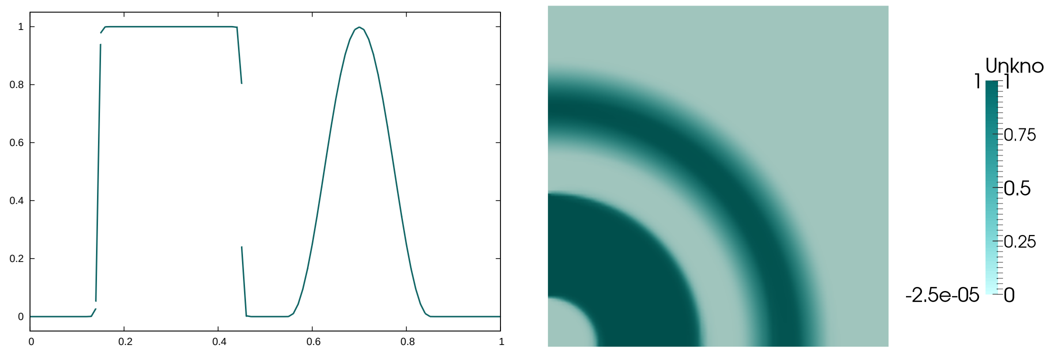
In the transport case, the Dirichlet boundary conditions are only imposed on the inflow boundaries, which, for this convection field, are the sides of the square $[0,1] \times [0,1]$ corresponding to $x = 0$ and $y = 1$. We will impose 0 all along the side $y = 1$ and the following function on the side $x = 0$:

$$\overline{u}(0,y) = \begin{cases} 1 & y \in [0.15, 0.45], \\ \cos^2\left(\frac{10}{3}\pi(y - 0.4)\right) & y \in [0.55, 0.85], \\ 0 & \text{elsewhere.} \end{cases}$$

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 $u(x,0) = \overline{u}(0,x)$. We solve this problem in a 100×100 Q_1 mesh and check the effect of the constants σ, τ and γ 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 γ 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 σ and τ. Nevertheless fixing $q = 10$, and tuning only σ and τ, 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 σ and τ. In the next test we fix $q = 10$, $\gamma = 10^{-2}$, and different values for τ and σ. In particular we will use $\tau = \{10^{-1}, 10^{-2}10^{-4}, 10^{-8}\}$



(a) Solution on inflow boundary $x = 0$. (b) Solution on the whole domain

FIGURE 2. Solution of the problem (21) used for parameter-tuning after 100 iterations with $q = 10$, and $\sigma = \tau = \gamma = 0$. A $100 \times 100\, Q_1$ mesh has been used.

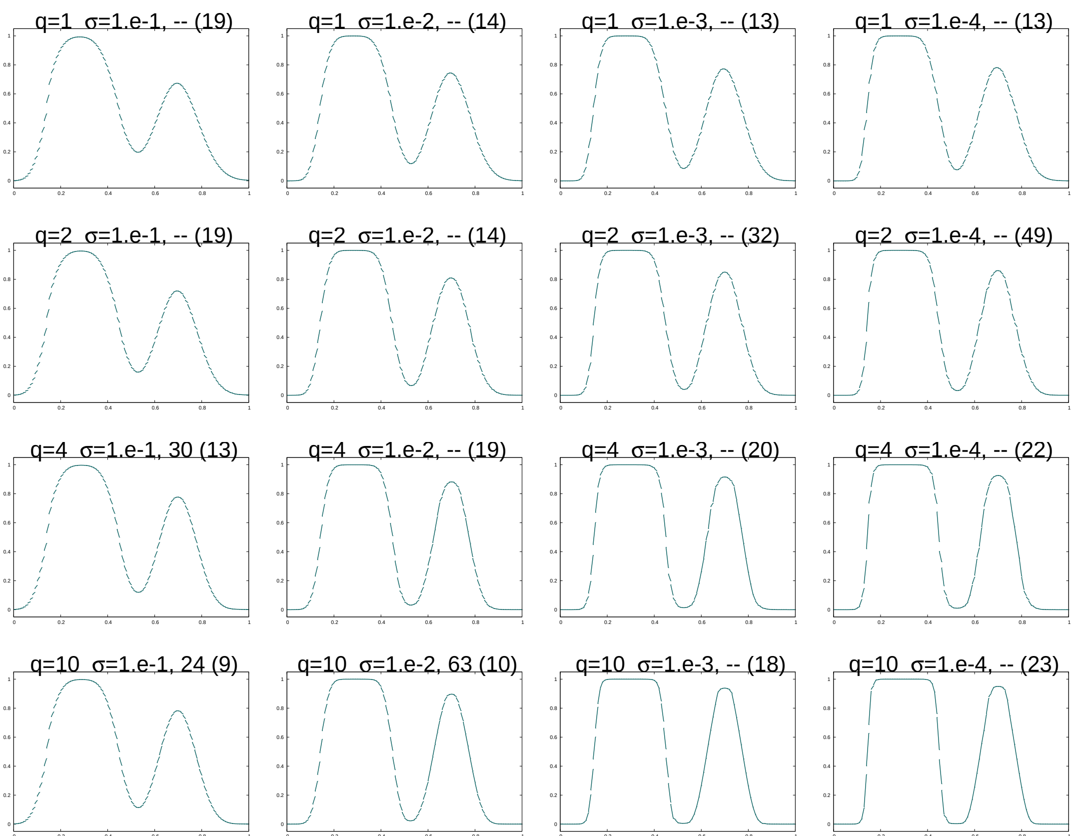


FIGURE 3. Profile of the solution of (21) on the outflow side ($y = 0$) for different values of q, σ, $\tau = \sigma^2$, and $\gamma = 10^{-2}$. Each figure title indicates the value of q, σ, the number of nonlinear iterations for Picard and the hybrid scheme (in brackets), (--) means "not converged". A 100×100 mesh has been used.

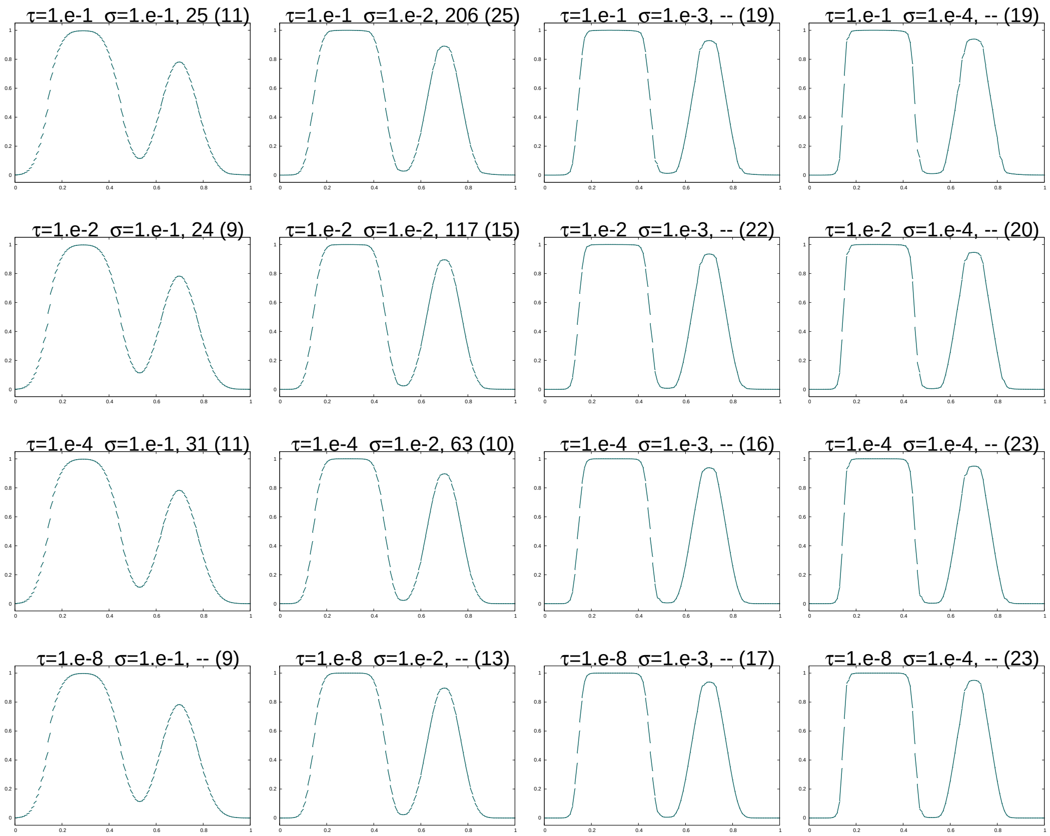


FIGURE 4. Profile of the solution of (21) on the outflow side ($y = 0$) for different values of σ, τ, $q = 10$, and $\gamma = 10^{-2}$. Each figure title indicates the value of τ, σ, the number of nonlinear iterations for Picard and the hybrid scheme (in brackets), (--) means "not converged". A 100×100 mesh has been used.

and $\sigma = \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. We will use the same nonlinear solvers as before. The obtained results are shown in Fig. 4. Values of σ around 10^{-2} and τ above 10^{-4} are needed to ensure Picard convergence. In the case of the hybrid scheme, we do not observe much difference in terms of the number of iterations required to converge for different values of τ. Nevertheless, as σ is reduced the method requires a slightly larger number of iterations.

Finally, we want to fine-tune γ. To do so, we will fix the values of $\sigma = 10^{-2}$ and $\tau = 10^{-4}$ and reduce the value of γ from 10^{-4} to 0. We can check in Fig. 5 how, even with $\gamma = 0$, the solution is able to converge, but the number of iteration is larger than between $\gamma = 10^{-4}$ and $\gamma = 10^{-12}$ whereas the solution is practically the same. In our numerical experiments we will work with the smooth and non-smooth version and with both nonlinear solvers. Thus, when comparing results we are interested in solutions that converge in a reasonable amount of time steps (we take $\gamma = 10^{-2}$).

In the transient examples, we will only use Picard linearization, allowing us take $\gamma = 0$ to see whether we can obtain a sharper solution. Additionally, we will let Q, the exponent of α_a for the perturbation of the mass matrix in (14), be $Q = +\infty$; meaning that the matrix is only perturbed when $\alpha_a = 1$. Nevertheless, we recall that we can do so because the nonlinear iterative method that we use does not need the stabilization to be differentiable; if that is not the case, γ must be greater than 0 and $Q < +\infty$. Summarizing, in view of the results obtained, we will use $q = 10$, $Q = \infty$, $\sigma = 10^{-2}$, $\tau \leq 10^{-6}$ and either $\gamma = 10^{-2}$ or $\gamma = 0$ in the oncoming transient numerical experiments.

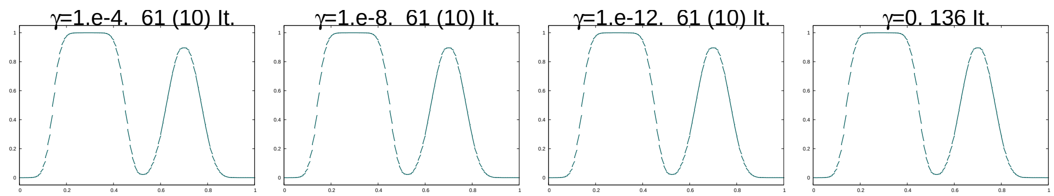


FIGURE 5. Profile of the solution of (21) on the outflow side ($y = 0$) for different values of γ, $q = 10$, $\sigma = 10^{-2}$, and $\tau = 10^{-4}$. Each figure title indicates the value of γ, and the number of nonlinear iterations for Picard scheme and the hybrid scheme (in brakets). For $\gamma = 0$ the number of iterations is not given for the hybrid scheme since the stabilization is not smooth and the Jacobian might be undefined at some points. A 100×100 mesh has been used.

7. Numerical experiments

In this section, we are interested in showing how the method previously introduced deals with a set of numerical experiments. We recall that, since we are using isotropic uniform meshes, the value of the characteristic length h_K of each element K can be computed as the length of the edges of the squares.

7.1. **Convergence to a smooth solution.** In this first experiment we want to determine the convergence of the method towards a smooth solution that has maxima and minima inside the domain (α_a reaches the value 1 in some regions in the domain). We compare the performance of the original interior penalty dG method against our dG method with artificial diffusion, both with and without smoothing (allowing a maximum of 100 iterations). The steady problems we solve to this end are the following:

$$\left\{\begin{array}{rcll} -\mu\Delta u + \boldsymbol{\nabla}\cdot(\boldsymbol{\beta} u) & = & -4\pi^2 \sin\left(2\pi\left(x - \frac{y}{\tan\theta}\right)\right)\left(1 + \frac{1}{\tan^2\theta}\right) & \text{in } \Omega = [0,1]\times[0,1], \\ u(x,y) & = & \sin\left(2\pi\left(x - \frac{y}{\tan\theta}\right)\right) & \text{on } \partial\Omega, \end{array}\right.$$

and

$$\left\{\begin{array}{rcll} \boldsymbol{\nabla}\cdot(\boldsymbol{\beta} u) & = & 0 & \text{in } \Omega = [0,1]\times[0,1], \\ u(x,y) & = & \sin\left(2\pi\left(x - \frac{y}{\tan\theta}\right)\right) & \text{on } \partial\Omega, \end{array}\right.$$

with $\boldsymbol{\beta} = (\cos(\theta), \sin(\theta))$ and $\theta = \pi/3$. In both cases the exact solution is $u(x,y) = \sin(2\pi(x - \frac{y}{\tan\theta}))$. We have second-order convergence of the non-stabilized method with (bi)linear finite elements and we would like to preserve it for the stabilized one when applied to smooth solutions. This is indeed what happens. When we consider the problem with diffusion $\mu = 1$ in which the contributions to the matrix are dominated by the diffusion term for the finest meshes, the error is almost the same for both the smoothed and the non-smoothed versions. It is important to note that the nonlinear solver is not able to converge for the finest meshes when using the non-smooth stabilization. This introduces an additional error responsible of the convergence degeneration observed at Fig. 6. In addition, it is worth noting that the requirement of the boundary condition extrapolation to achieve optimal convergence. Tests without the extrapolation in Remark 4.2, i.e., $u_{ab}^{\text{sym}} = \overline{u}_a$, show degenerated convergence rates. This correction is more important as the gradients on the boundary become larger and the jumps smaller. As opposed, when working with pure transport (see Fig. 7), the effect of this extrapolation becomes negligible. In any case, the order of convergence is maintained. For the pure convection test, the smoothing of the parameters add an extra error to the computed solution, as expected, since it implies more artificial diffusion.

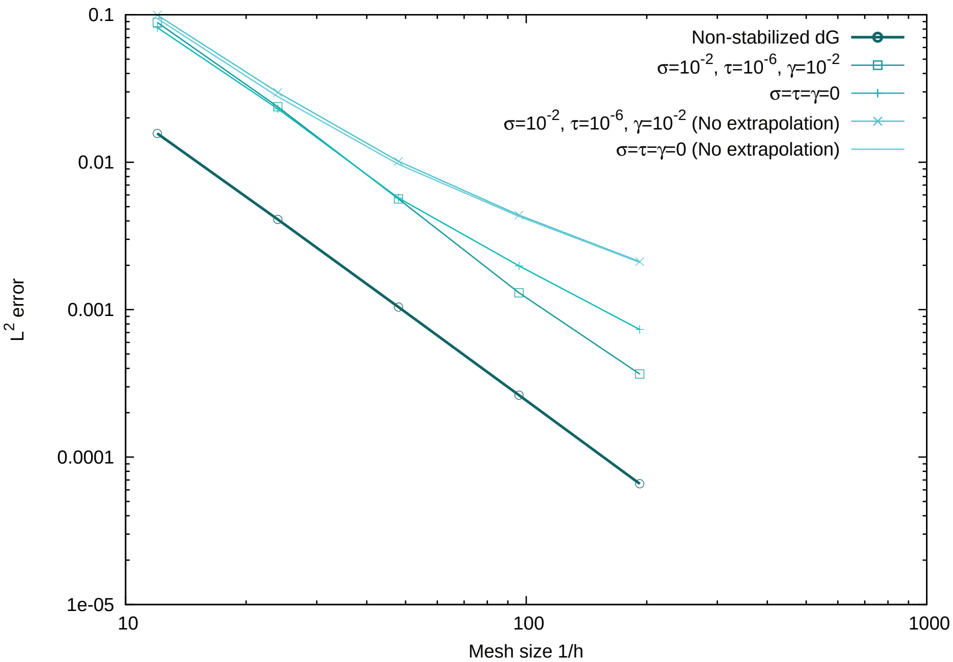


FIGURE 6. L^2 error convergence test for a convection-diffusion problem ($\mu = 1$) with a smooth solution. Different choices of stabilization parameters have been tested.

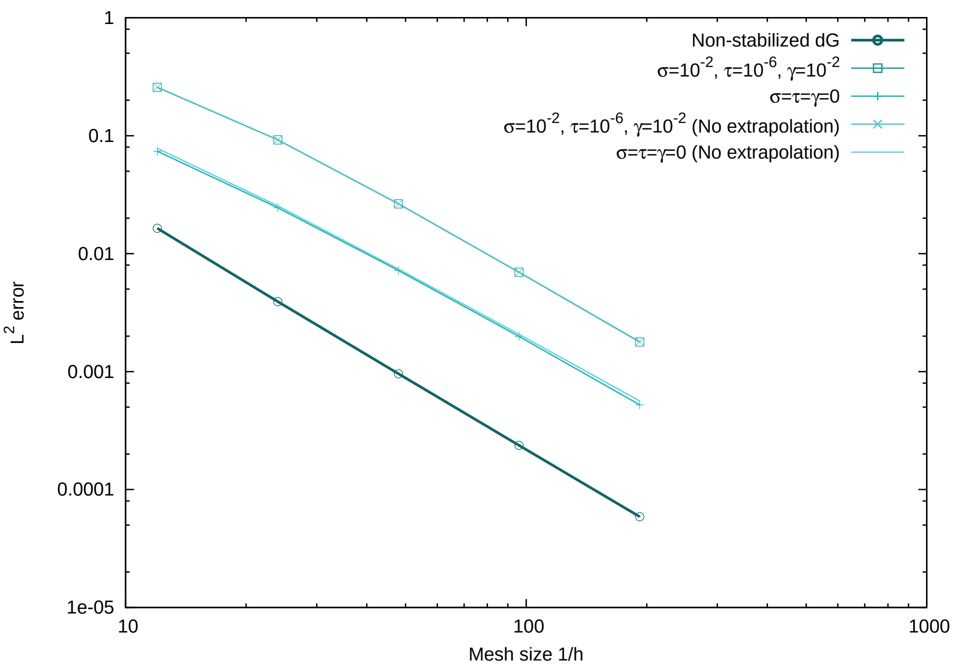


FIGURE 7. L^2 error convergence test for a pure convection problem ($\mu = 0$) with a smooth solution. Different choices of stabilization parameters have been tested.

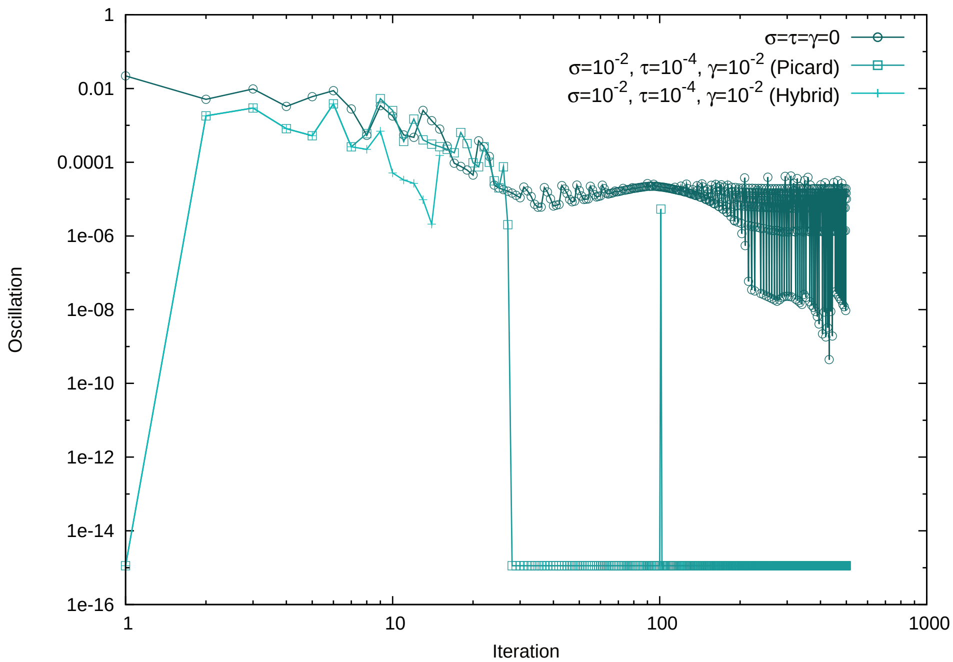


FIGURE 8. Maximum oscillation as defined in (23) for each nonlinear iteration performed for solving (22). The results for both a smooth and a non-smooth stabilization, as well as for different nonlinear schemes are depicted. A $100 \times 100\, Q_1$ mesh has been used for the tests.

7.2. **DMP-preservation.** In order to test the performance of the method in terms of DMP-preservation, we solve the steady problem

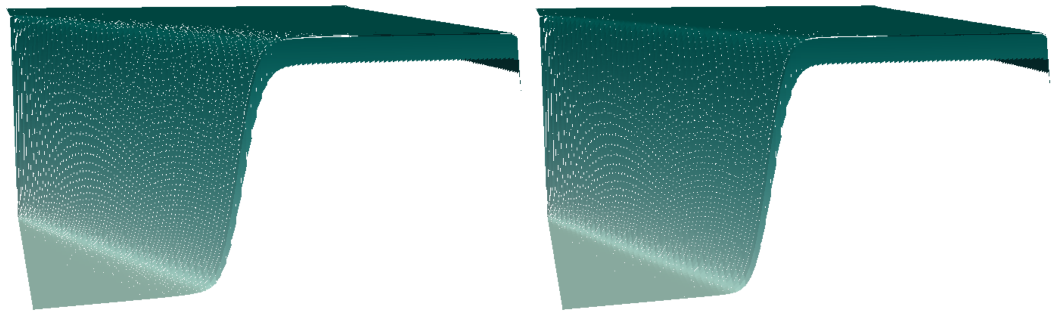
$$\left\{\begin{array}{rcll} -10^{-4}\Delta u + \boldsymbol{\nabla}\cdot(\boldsymbol{\beta} u) & = & 0 & \text{in } \Omega = [0,1]\times[0,1], \\ u(x,0) & = & 0 & x \in [0,1], \\ u(x,1) & = & 1 & x\in[0,1], \\ u(0,y) & = & \frac{1}{2} + \frac{1}{\pi}\arctan(10^4(y-0.7)) & y\in(0,1), \\ u(1,y) & = & 0 & y \in (0,1), \end{array}\right. \tag{22}$$

with $\boldsymbol{\beta} = (\cos(\pi/3), -\sin(\pi/3))$. As the problem is convection-dominated, the expected result is a propagation in the direction defined by $\boldsymbol{\beta}$ of the profile imposed in the inflow boundary $x = 0$. It is well known that when the method is not stabilized, it leads to a solution that has strong oscillations around the internal and boundary layers. We expect our method to control such spurious oscillations, as already proved in the numerical analysis.

We use a $100 \times 100\, Q_1$ Cartesian mesh, set the tolerance of the nonlinear iterations to 10^{-4}, and allow a maximum of 500 iterations. We plot the maximum oscillation, defined as

$$\mathrm{OSC} = \max\{0, -\min_{x\in\Omega} u_h(x), \max_{x\in\Omega}(u_h(x)-1), \}, \tag{23}$$

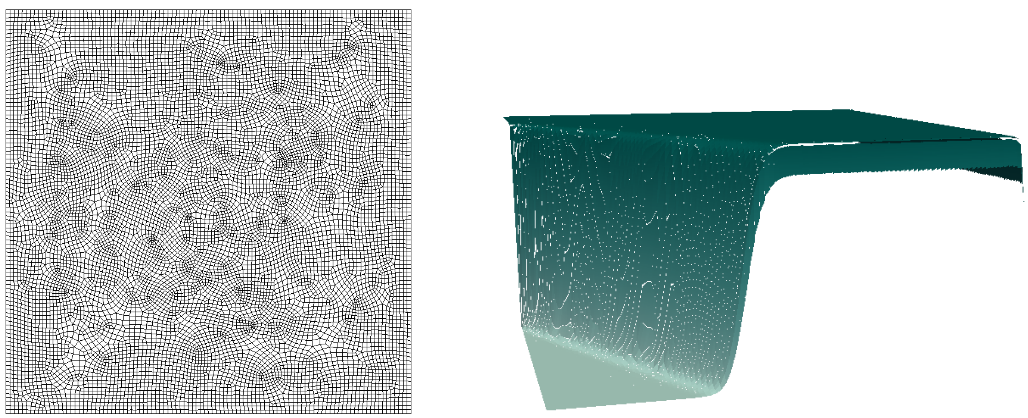
at each iteration. Both the smooth and the non-smooth methods are tested and both nonlinear solvers above are used for the smooth version. The results in Fig. 8 show that only the hybrid method is able to converge. Let us remark that Picard's method has not reeached convergence (having a maximum of 500 iterations) in any case, even though the non-converged solution with smooth stabilization satisfies the DMP up to machine precision. On the contrary, the hybrid method satisfies the DMP up to the tolerance, but it only needs 15 iterations to converge. The results obtained are plotted in Fig. 9 and are as sharp as the non-smooth ones.



(a) $\sigma_h = 10^{-2}$; $\tau_h = 10^{-4}$; $\gamma_h = 10^{-2}$. 13 Iterations.

(b) $\sigma = \tau = \gamma = 0$. NC.

Figure 9. Solution of problem (22) using both the smooth and the non-smooth version of the stabilization. A 100×100 mesh have been used.



(a) Unstructured mesh used

(b) $\sigma_h = 10^{-2}$; $\tau_h = 10^{-4}$; $\gamma_h = 10^{-2}$. 17 Iterations.

Figure 10. Solution of problem (22) using the smooth version of the stabilization. An unstructured mesh of $h \approx 10^{-2}$ have been used.

We know from the numerical analysis that the method preserves the DMP on unstructured meshes. In the following test we are checking this result and analyzing if it affects to the accuracy of the method. To this end, we solve again the previous problem using an unstructured mesh with an element size $h \approx 10^{-2}$ (see Fig. 10 (a)). The solution is depicted in Fig. 10(b). It can be seen that the accuracy is not affected by reasonable mesh perturbations. Moreover, we compare the effect of the smoothing in the case of unstructured meshes. It can be observed in Fig. 11 that the stabilization method is minimally affected by the mesh topology. Nevertheless, it is worth noting that the number of iterations slightly increases for the hybrid method, from 15 to 17. In any case, we achieve the same conclusions as when using the structured mesh. The hybrid method is able to converge and it satisfies de DMP up to the tolerance. The relaxed Picard it is not able to converge, even though it leads to non-converged solutions that satisfy the DMP for the smooth case at the lasts iterations.

7.3. **Three body rotation.** Finally, we want to test the DMP-preservation and LED property of the method for transient problems. To this end, we use the classical three body rotation test. We solve the 2D transport equation (1) in $\Omega = [0,1] \times [0,1]$ with $\mu = 0$, $\boldsymbol{\beta} = (-2\pi(y-0.5), 2\pi(x-0.5))$. The initial solution is given in [20] and its interpolation in a mesh of 200×200 bilinear elements is displayed in Fig. 12.

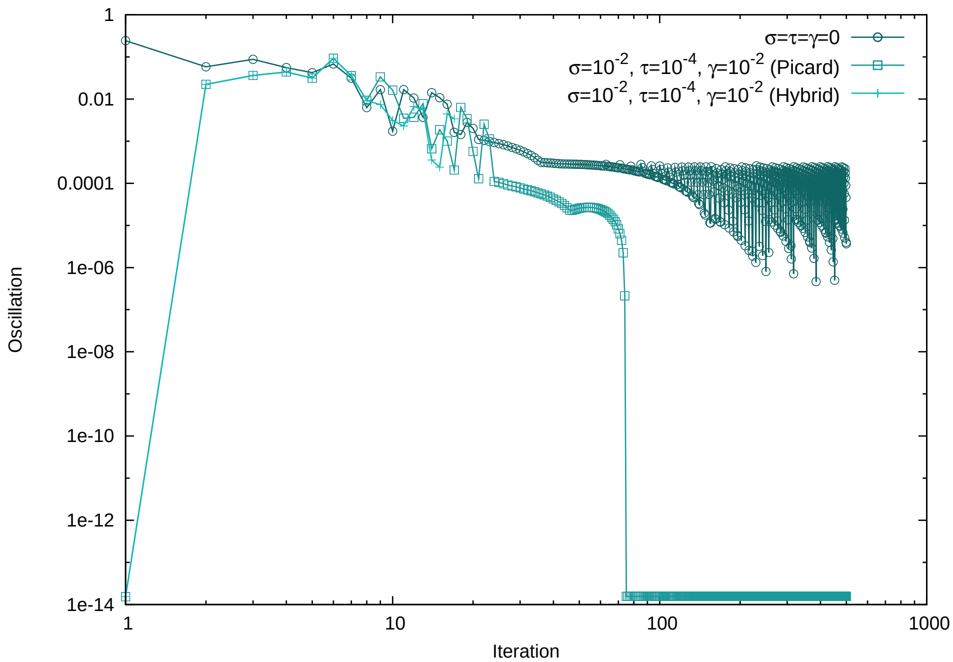
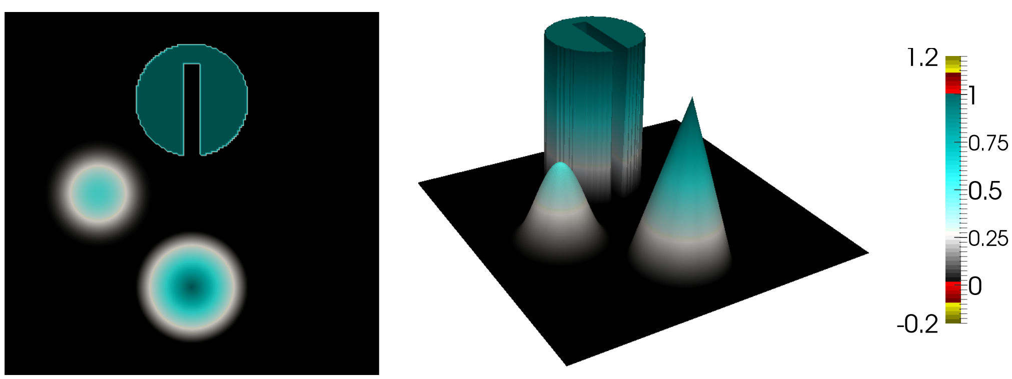


FIGURE 11. Maximum oscillation as defined in (23) for each nonlinear iteration performed for solving (22). The results for both a smooth and a non-smooth stabilization, as well as for different nonlinear schemes are depicted. An unstructured mesh of $h \approx 10^{-2}$ has been used for the tests.



(a) 2D (b) 3D (c) Legend

FIGURE 12. Initial solution of three body rotation test problem discretized in a $200 \times 200\, Q_1$ mesh.

The solution to the transport problem is simply a translation in the direction of the convection. In this case, the initial solution rotates counterclockwise and the final solution is computed at $T = 1$ after one complete round. The idea is to compare the initial and the final values to see how dissipative the method is and, at the same time, check the values of the maximum oscillation in each time step to see if there is any violation of the DMP. In order to do so, we have used the color map plotted in Fig. 12(c), which takes colors from black to green in the interval $[0, 1]$ but it uses shades of red in

$[-0.1, 0) \cup (1, 1.1]$ and shades of yellow in $[-0.2, 0.1) \cup (1.1, 1.2]$. This way, it is easy to identify the violations of the global DMP of the problem.

The solution is computed with the dG method without any stabilization, the smoothed stabilized method (in which we consider both $\gamma = 0$ and $\gamma = 10^{-2}$), and the stabilized method without smoothing ($\sigma = \tau = \gamma = 0$). The smoothed case with $\gamma = 0$ has been considered to reduce the mass lumping activation without spoiling the convergence of the nonlinear iterations. As we will see in the numerical results, although the method is less dissipative, the final results do not differ much (see Fig. 13(d) and Fig. 13(f)). We recall that, for the integration in time, a weighted mass lumping is used and the parameter Q is set to $Q = 10$ in the smoothed case and $Q = +\infty$ in the non-smoothed, to minimize the phase error induced by the lumping of the matrix. We have run the test in a mesh of 200×200 bilinear elements and used $N^t = 2000$ time steps, with nonlinear tolerance of $5 \cdot 10^{-4}$ and a maximum of 50 iterations. The results are plotted in Fig. 13. We have also plotted the maximum oscillation in time in Fig. 14. Actually, instead of printing the maximum oscillation at each time step we depict the mean value of the maximum in each bunch of 10 time steps, for improving the result visualization and analysis.

We can observe in Fig. 14 how, if the method is not stabilized, the oscillations appear from the first iterations and do not decrease in time, being of order 10^{-1}. On the other hand, one can observe that the stabilized version of the method gives oscillatory results on the first iterations if the method is not smoothed. This is due to the fact that the nonlinear solver is not able to converge in the first time steps and what is plotted is the result after 50 nonlinear iterations. Instead, when smoothing the stabilization, the nonlinear solver converges and the violation of the DMP is of the order of the tolerance for $\gamma = 0$. When using $\gamma = 10^{-2}$, the method is even more dissipative (as expected since the greater the value of γ the more the shock detector is activated) and the DMP is only violated up to machine precision. We want to point out that the appearance of jumps on the smooth part of the figures, such as the cone, is due to the lumping of the mass matrix. When avoiding the mass lumping, those jumps disappear, but the DMP is then violated.

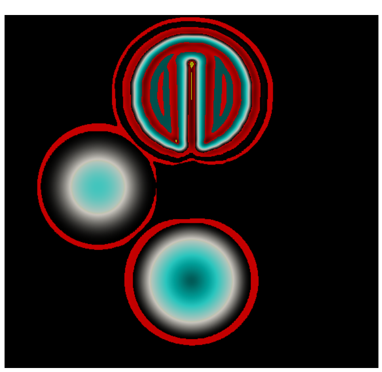
8. Conclusions

In this work we have designed a method that fulfills the DMP property for the steady multidimensional transport problem when a dG space discretization is used. In the transient case, together with implicit time stepping, the method enjoys the LED property. The original scheme is stabilized with an artificial diffusion graph-Laplacian operator. The edge diffusion is only activated on troublesome regions based on a shock detector that relies on the jumps of the gradient of the solution around the nodes of the mesh, in order to minimize the smearing of the solution and improve the solution in smooth regions. We provide a set of conditions to be satisfied by the stabilized formulation to enjoy the DMP (and LED) properties, and designed an artificial edge viscosity to fulfill these conditions. The results hold for arbitrary meshes and space dimensions. The resulting method is proved to be DMP-preserving, LED, linearity-preserving, and Lipschitz continuous. We have also proved the existence of solutions. However, the method is stil highly nonlinear and it is hard to attain nonlinear solver convergence. Thus, we propose a smooth version of the scheme that is twice differentiable and still enjoys DMP and LED properties. We provide a set of numerical experiments to check the features proved in the theoretical analysis, and to show the improvement in terms of computational cost due to the combination of the smooth version of the scheme with a Newton nonlinear solver with line search.

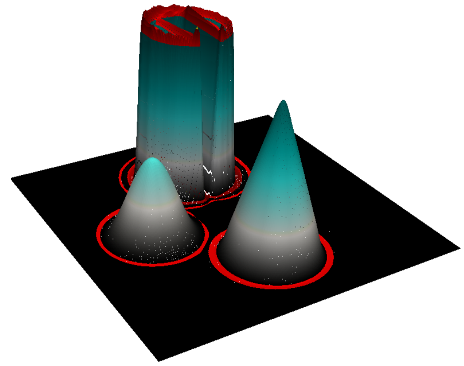
Some interesting future work might be to extend these features to approximations of much more involved equations such as Euler or compressible Navier-Stokes problems.

Acknowledgments

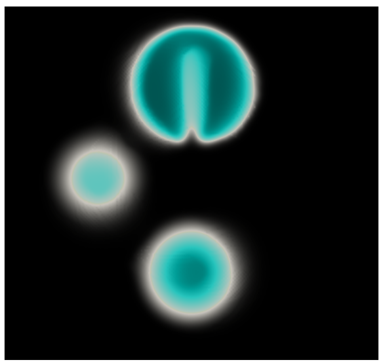
SB gratefully acknowledges the support received from the Catalan Government through the ICREA Acadèmia Research Program. JB gratefully acknowledges the support received from "la Caixa" Foundation through its PhD scholarship program. AH gratefully acknowledges the support received from



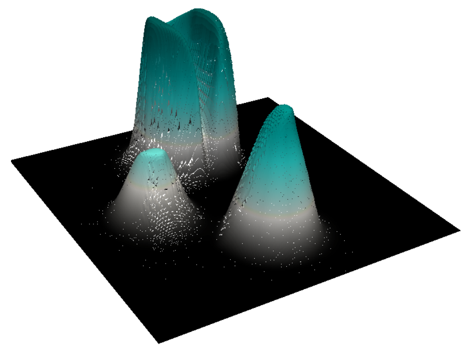
(a) 2D. No stabilization



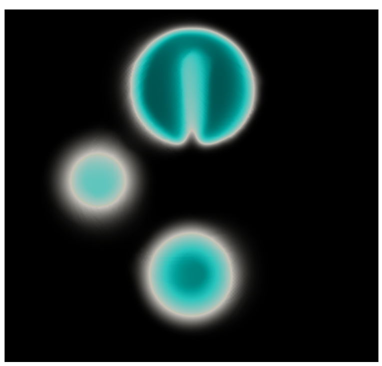
(b) 3D. No stabilization



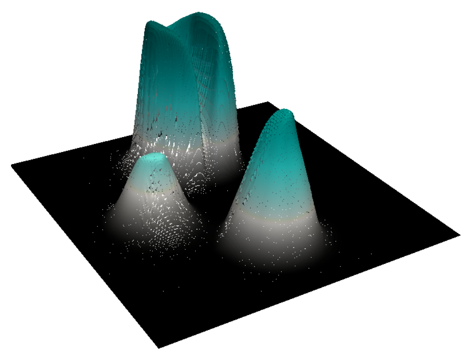
(c) 2D. $\sigma = 10^{-2}$; $\tau = 10^{-4}$; $\gamma = 10^{-2}$; $Q = 10$.



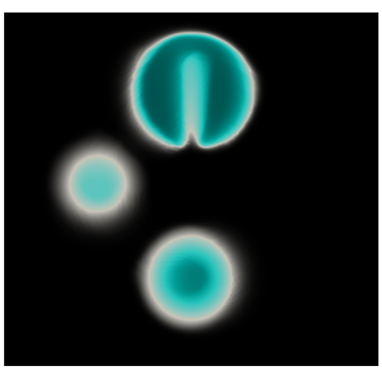
(d) 3D. $\sigma = 10^{-2}$; $\tau = 10^{-4}$; $\gamma = 10^{-2}$; $Q = 10$.



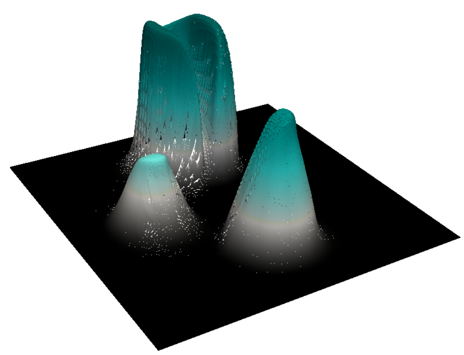
(e) 2D. $\sigma = 10^{-2}$; $\tau = 10^{-4}$; $\gamma = 0$; $Q = 10$.



(f) 3D. $\sigma = 10^{-2}$; $\tau = 10^{-4}$; $\gamma = 0$; $Q = 10$.



(g) 2D. $\sigma = \tau = \gamma = 0$. NC.



(h) 3D. $\sigma = \tau = \gamma = 0$. NC.

FIGURE 13. Solution of three body rotation test for different combinations of the stabilization parameters. A $200 \times 200\, Q_1$ mesh and Crank-Nicolson time integration with $N^t = 2000$ have been used.

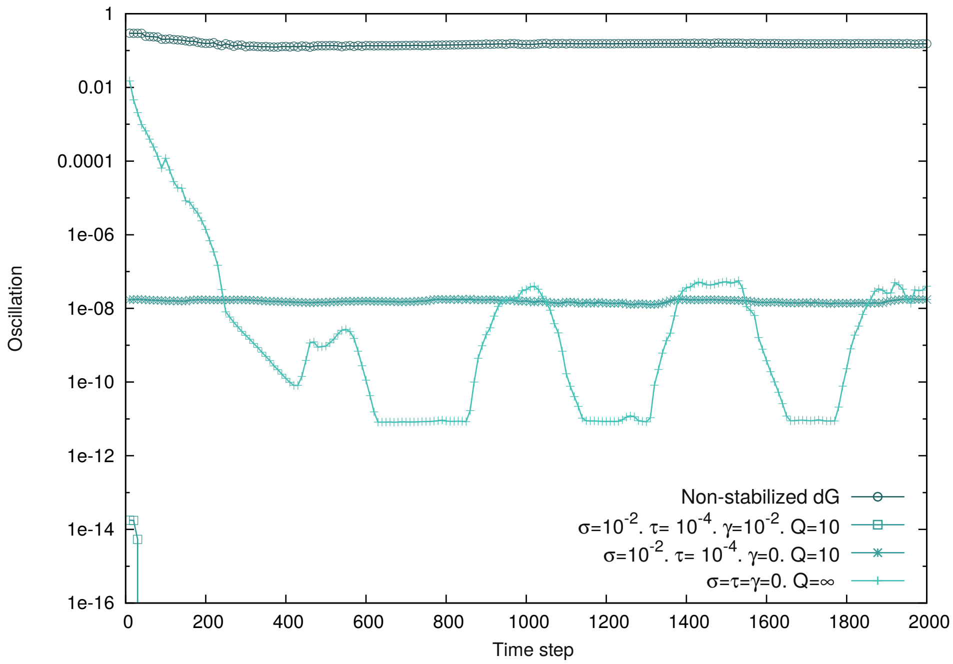


FIGURE 14. Oscillation values at each time step for the three body rotation test using different choices for the stabilization parameters. A 200×200 mesh and Crank-Nicolson time integration with $N^t = 2000$ have been used.

the Catalan Government through a FI fellowship. We acknowledge the financial support to CIMNE via the CERCA Programme / Generalitat de Catalunya.

References

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous galerkin methods for elliptic problems* **39** (2002), no. 5, 1749–1779.

[2] S. Badia and J. Bonilla, *Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization*, Computer Methods in Applied Mechanics and Engineering **313** (2017), 133–158.

[3] S. Badia and A. Hierro, *On Monotonicity-Preserving Stabilized Finite Element Approximations of Transport Problems*, SIAM Journal on Scientific Computing **36** (2014), no. 6, A2673–A2697.

[4] ———, *On discrete maximum principles for discontinuous Galerkin methods*, Computer Methods in Applied Mechanics and Engineering **286** (2015), 107–122.

[5] G. R. Barrenechea, E. Burman, and F. Karakatsani, *Edge-based nonlinear diffusion for finite element approximations of convectiondiffusion equations and its relation to algebraic flux-correction schemes*, Numerische Mathematik (2016), 1–25.

[6] S. C. Brenner and R. Scott, *The mathematical theory of finite element methods*, 3rd ed., Springer, 2008.

[7] H. Brezis, *Functional analysis, sobolev spaces and partial differential equations*, 1st Edition., Springer, 2010.

[8] F. Brezzi, L. D. Marini, and E. Süli, *Discontinuous Galerkin methods for first-order hyperbolic problems*, Mathematical Models and Methods in Applied Sciences **14** (2004), no. 12, 1893–1903.

[9] 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), no. 35, 3833–3855.

[10] ———, *Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence*, Mathematics of Computation **74** (2005), no. 252, 1637–1652.

[11] E. Burman and P. Hansbo, *Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems*, Computer Methods in Applied Mechanics and Engineering **193** (2004), no. 15–16, 1437–1453.

[12] E. Burman, *On nonlinear artificial viscosity, discrete maximum principle and hyperbolic conservation laws*, BIT Numerical Mathematics **47** (2007), no. 4, 715–733.

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

[14] R. Codina, *A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation*, Computer Methods in Applied Mechanics and Engineering **110** (1993), no. 3–4, 325–342.
[15] A. Ern and J.-L. Guermond, *Theory and practice of finite elements*, Springer, 2004.
[16] 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), no. 4, 2163–2182.
[17] 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), 198–213.
[18] V. John and P. Knobloch, *On spurious oscillations at layers diminishing (SOLD) methods for convectiondiffusion equations: Part II Analysis for and finite elements*, Computer Methods in Applied Mechanics and Engineering **197** (2008), no. 2124, 1997–2014.
[19] D. Kuzmin and S. Turek, *Flux Correction Tools for Finite Elements*, Journal of Computational Physics **175** (2002), no. 2, 525–558.
[20] D. Kuzmin, *A guide to numerical methods for transport equations*, unpublished, 2010.
[21] D. Kuzmin and M. Möller, *Algebraic Flux Correction I. Scalar Conservation Laws*, Flux-Corrected Transport, 2005. DOI: 10.1007/3-540-27206-2_6.
[22] D. Kuzmin and J. N. Shadid, *A new approach to enforcing discrete maximum principles in continuous Galerkin methods for convection-dominated transport equations*, submitted (2015).
[23] ———, *Gradient-based nodal limiters for artificial diffusion operators in finite element schemes for transport equations*, submitted (2016).
[24] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, 1st ed., 2002.
[25] A. Mizukami and T. J. R. Hughes, *A Petrov-Galerkin finite element method for convection-dominated flows: An accurate upwinding technique for satisfying the maximum principle*, Computer Methods in Applied Mechanics and Engineering **50** (1985), no. 2, 181–193.
[26] L. R. Scott and S. Zhang, *Finite Element Interpolation of Nonsmooth Functions Satisfying Boundary Conditions*, Mathematics of Computation **54** (1990), no. 190, 483–493.