A nonlinear consistent penalty method weakly enforcing positivity in the finite element approximation of the transport equation

Erik Burman [a] and Alexandre Ern [b]

[a] *Department of Mathematics, University College London, London, UK–WC1E 6BT, UK*

[b] *University Paris-Est, CERMICS (ENPC), 77455 Marne la Vallée, France*

Abstract

We devise and analyze a new stabilized finite element method to solve the first-order transport (or advection-reaction) equation. The method combines the usual Galerkin/Least-Squares approach to achieve stability with a nonlinear consistent penalty term inspired by recent discretizations of contact problems to weakly enforce a positivity condition on the discrete solution. We prove the existence and the uniqueness of the discrete solution. Then we establish quasi-optimal error estimates for smooth solutions bounding the usual error terms in the Galerkin/Least-Squares error analysis together with the violation of the maximum principle by the discrete solution. Numerical examples are presented to illustrate the performances of the method.

Key words: stabilized finite element method, consistent penalty, positivity preserving, transport equation, discrete maximum principle

1 Introduction

The design of robust and accurate finite element methods for first-order transport (or advection-reaction) equations or for advection-dominated advection-diffusion equations remains an active field of research. Indeed, the task of designing a numerical scheme that is of higher order than one in the zone where the exact solution is smooth, but preserves the monotonicity properties of the exact solution on the discrete level, is nontrivial. Since it is known that such a scheme necessarily must be nonlinear even for linear equations, one typical strategy adopted when working with stabilized finite element methods is to add an additional nonlinear shock-capturing term, designed to make the method satisfy a discrete maximum principle; see, e.g., [1,2]. These methods, however, often result in ill-conditioned nonlinear equations and include parameters that may be difficult to tune and that depend on the mesh geometry. Another approach is the so-called flux-corrected finite element method [3,4]. In this scheme, the system matrix is manipulated so that it becomes a so called M-matrix, the inverse of which has positive coefficients which yields a maximum principle. Such a scheme is monotonicity preserving, but of first order. In order to improve the accuracy, anti-diffusive mechanisms, or flux-limiter techniques, have been proposed so as to reduce the amount of dissipation in the smooth regions by blending a low- and a high-order approximation [4,5].

In this paper, we consider a method that follows a completely different approach. The starting observation is that, if the problem satisfies a maximum principle of the form

$$u \geq 0, \tag{1}$$

then this constraint can be added to the problem without any perturbation. On the discrete level however, the condition (1) is not necessarily satisfied, unless enforced by the numerical method, which is the purpose of all the methods discussed above. One may argue that one can solve the problem with any method under the constraint (1). This was proposed in [6]. The resulting method, in the form of a variational inequality, is unwieldy, with the need of Lagrange multipliers to impose the constraint and associated stability and solver issues. In the present work, we instead draw on recent advances in the field of contact problems [7,8], where the variational inequality instead is discretized by means of a nonlinear consistent penalty method. Note however that in the present context the formulation cannot be associated with an augmented Lagrangian method, due to the lack of symmetry of the formulation.

Our method combines the well-known Galerkin/Least Squares (GaLS) discretization of the transport equation (see [9,10]) with a nonlinear switch inspired from [8] and that changes the equations in the zones where (1) is violated to a least-squares penalty on this inequality (more precisely, the negative part of the discrete solution) together with a least-squares penalty on the residual. Our method is not meant to enforce strictly a discrete maximum principle, but to blend asymptotically the satisfaction of the GaLS approximation of the PDE with the satisfaction of the discrete maximum principle, in the same spirit of the above-mentioned methods for contact problems. We first prove the existence and uniqueness of the discrete solution. Then, our main result is Theorem 3.1 where we establish a quasi-optimal error estimate bounding at the same time the error measured in the usual GaLS norm (combining the L^2-norm on the solution, its boundary values, and the advective derivative weighted by the local mesh size) and the violation of the positivity condition (1) measured by the weighted L^2-norm of the negative part of the discrete solution. These convergence results hold for all polynomial orders $k \geq 1$. Another salient feature of our method is its flexibility in incorporating a priori lower and upper bounds on the discrete solution by simply adding the corresponding consistent penalty term to the discrete formulation. Finally, we report some numerical experiments illustrating that accurate solutions with mild and asymtotically vanishing violations of the discrete maximum principle can be obtained at moderate computational costs.

2 Model problem and GaLS discretization

Let Ω be an open, bounded, Lipschitz set in $\mathbb{R}^d$, $d \in \{2,3\}$, let $\beta \in W^{1,\infty}(\Omega;\mathbb{R}^d)$ be a given advection velocity, and let $\sigma \in L^\infty(\Omega;\mathbb{R})$ be a given reaction coefficient. We assume that β and σ satisfy the following (classical) positivity assumption: There exists $\sigma_0 > 0$ such that

$$0 < \sigma_0 \leq \sigma - \frac{1}{2}\nabla\cdot\beta, \qquad \text{a.e. in } \Omega. \tag{2}$$

We split the boundary $\partial\Omega$ of Ω as $\partial\Omega = \partial\Omega^- \cup \partial\Omega^0 \cup \partial\Omega^+$ with $\partial\Omega^- = \{x \in \partial\Omega \mid (\beta\cdot n)(x) < 0\}$ (inflow boundary), $\partial\Omega^0 = \{x \in \partial\Omega \mid (\beta\cdot n)(x) = 0\}$ (characteristic boundary), and $\partial\Omega^+ = \{x \in \partial\Omega \mid (\beta\cdot n)(x) > 0\}$ (outflow boundary). On the boundary, we introduce the linear space composed of those functions $v : \partial\Omega \mapsto \mathbb{R}$ such that the weighted L^2-norm $\||\beta\cdot n|^{\frac{1}{2}}v\|_{\partial\Omega}$ is bounded, and we denote this space by $L^2_{|\beta\cdot n|}(\partial\Omega;\mathbb{R})$.

We consider the following model problem: Find $u : \Omega \to \mathbb{R}$ such that

$$A(u) := \beta\cdot\nabla u + \sigma u = f \quad \text{in } \Omega, \tag{3a}$$

$$u = g \quad \text{on } \partial\Omega^-. \tag{3b}$$

We assume that $f \in L^2(\Omega;\mathbb{R})$ and $g \in L^2_{|\beta\cdot n|}(\partial\Omega^-;\mathbb{R})$ and look for a weak solution in the graph space $V := \{v \in L^2(\Omega;\mathbb{R}) \mid \beta\cdot\nabla v \in L^2(\Omega;\mathbb{R})\}$. Assuming that $\operatorname{dist}(\partial\Omega^-, \partial\Omega^+) > 0$, one can show [11] that functions in the graph space admit a trace in the weighted space $L^2_{|\beta\cdot n|}(\partial\Omega;\mathbb{R})$ and that there exists one and only one weak solution in the graph space V to the model problem (3). In particular, we observe that the following positivity condition holds:

$$\sigma_0\|v\|^2_\Omega + \frac{1}{2}\||\beta\cdot n|^{\frac{1}{2}}v\|^2_{\partial\Omega} \leq (\beta\cdot\nabla v + \sigma v, v)_\Omega - \langle(\beta\cdot n)v, v\rangle_{\partial\Omega^-}, \qquad \forall v \in V. \tag{4}$$

Several stabilized H^1-conforming finite element methods are available in the literature to discretize (3). We focus on the Galerkin/Least-Squares method (GaLS). Let $\mathcal{T}_h$ be a mesh from a shape-regular mesh sequence. We assume for simplicity that Ω is a polytope (polygon/polyhedron) so that $\mathcal{T}_h$ can cover Ω exactly. Let $k \geq 1$ be the polynomial degree and consider the H^1-conforming finite element space

$$V_h^k := \{v_h \in C^0(\overline{\Omega};\mathbb{R}) \mid v_h|_T \in \mathbb{P}_k(T;\mathbb{R}),\ \forall T \in \mathcal{T}_h\}, \tag{5}$$

where $\mathbb{P}_k(T;\mathbb{R})$ denotes the space composed of $\mathbb{R}$-valued functions that are the restriction to T of d-variate polynomials of degree at most k. We consider the following discrete problem: Find $u_h \in V_h^k$ such that

$$a_h^\tau(u_h, w_h) = \ell_h^\tau(w_h), \qquad \forall w_h \in V_h^k, \tag{6}$$

with the following bilinear and linear forms:

$$a_h^\tau(v_h, w_h) := (A(v_h), w_h + \tau A(w_h))_\Omega - \langle(\beta{\cdot}n)v_h, w_h\rangle_{\partial\Omega^-}\,, \tag{7a}$$
$$\ell_h^\tau(w_h) := (f, w_h + \tau A(w_h))_\Omega - \langle(\beta{\cdot}n)g, w_h\rangle_{\partial\Omega^-}\,. \tag{7b}$$

The stabilization parameter τ is piecewise constant on $\mathcal{T}_h$ and is of the order of the local mesh size h_T for all $T \in \mathcal{T}_h$; more precisely, a suitable choice is $\tau|_T = \min(\sigma_0^{-1}, \beta_T^{-1}h_T)$ with $\beta_T = \|\beta\|_{L^\infty(T;\mathbb{R}^d)}$. By construction, the discrete bilinear form a_h^τ is coercive with respect to the norm:

$$a_h^\tau(v,v) \geq \|\!|v|\!\|^2 := \sigma_0\|v\|_\Omega^2 + \||\beta{\cdot}n|^{\frac12}v\|_{\partial\Omega}^2 + \|\tau^{\frac12}A(v)\|_\Omega^2, \qquad \forall v \in V. \tag{8}$$

Moreover, exact consistency holds, and the following quasi-optimal error estimates can be established [9,10]: There exists C, uniform, such that

$$\|\!|u-u_h|\!\| \leq C \inf_{v_h\in V_h^k}(\|\tau^{-\frac12}(u-v_h)\|_\Omega+\||\beta{\cdot}n|^{\frac12}(u-v_h)\|_{\partial\Omega}+\|\tau^{\frac12}A(u-v_h)\|_\Omega), \tag{9}$$

and if $u \in H^{k+1}(\Omega)$, $\|\!|u-u_h|\!\| \leq C(\sum_{T\in\mathcal{T}_h}\phi_T h_T^{2k+1}|u|^2_{H^{k+1}(T;\mathbb{R})})^{\frac12}$ with $\phi_T = \max(\beta_T, \sigma_0 h_T)$.

3 The consistent penalty method

The model problem (3) has a maximum principle; for instance, if $f \geq 0$ and $g \geq 0$, then $u \geq 0$ in Ω. Unfortunately, this property rarely carries over to finite element discretizations. Our goal is to modify the GaLS finite element approximation (6) by using a consistent penalty method.

Let $\gamma > 0$ be a penalty parameter. For any function $v \in V$, let us define the function $\xi^\gamma : \Omega \to \mathbb{R}$ such that

$$\xi^\gamma(v) := [v - \gamma(A(v) - f)]_-, \tag{10}$$

where $x_- = \frac12(x - |x|)$ denotes the negative part of the real number x. Note that $\xi^\gamma(u) = 0$ in Ω for the weak solution u since $A(u) = f$ and $u_- = 0$. Let us consider the following discrete problem: Find $u_h \in V_h^k$ such that

$$a_h^{\tau\gamma}(u_h; w_h) = \ell_h^\tau(w_h), \qquad \forall w_h \in V_h^k, \tag{11}$$

with

$$a_h^{\tau\gamma}(v_h; w_h) := a_h^\tau(v_h, w_h) + (\gamma^{-1}\xi^\gamma(v_h), w_h)_\Omega. \tag{12}$$

Since $\xi^\gamma(u)$ vanishes identically in Ω, exact consistency still holds for (11). The discrete problem (11) remains meaningful and exactly consistent if the penalty parameter γ is replaced by a function taking uniformly positive values in Ω. We will take γ to be piecewise constant on the mesh $\mathcal{T}_h$ since the error analysis below will reveal that quasi-optimal error estimates are obtained by taking γ to be locally of the order of h_T (on quasi-uniform mesh sequences, a constant function γ can be considered).

3.1 Rationale of the consistent penalty method

Before embarking on the analysis of the method, let us briefly discuss the design principle behind the approach. First, we observe that if $[u_h - \gamma(A(u_h) - f)]_- = 0$, then the formulation coincides with the standard GaLS discretization. Assume now that $[u_h - \gamma(A(u_h) - f)]_- \neq 0$ everywhere in the macroelement $\Omega_i := \text{supp}(\varphi_i)$ where φ_i is an interior nodal (or hat) basis function. Then, since $x_- = x$ if $x_- \neq 0$, we observe that the standard Galerkin part is eliminated by the second term in the penalty term, so that (11) with $w_h = \varphi_i$ becomes

$$(\gamma^{-1}u_h, \varphi_i)_{\Omega_i} + (\tau(A(u_h) - f), A(\varphi_i))_{\Omega_i} = 0. \tag{13}$$

This shows that the nonlinear penalty term changes the discrete equation locally to the sum of two least-squares contributions, one on the violation of positivity by the discrete solution and one on the PDE residual. By choosing γ small, one can expect that the violation of the maximum principle is reduced. This is indeed one of the main conclusions of the error analysis below, where we additionally prove that quasi-optimal error estimates of the form (9) also hold for the consistent penalty method.

Remark 1 (Variant) *Variants are possible in* (12) *for the penalty term, such as*

$$a_h^{\tau\gamma}(v_h; w_h) := a_h^\tau(v_h, w_h) + (\gamma^{-1}\xi^\gamma(v_h), w_h + \tau A(w_h))_\Omega. \tag{14}$$

This variant can be analyzed using the same arguments as below. In particular, considering an interior nodal basis function, we now obtain $(\gamma^{-1}u_h, \varphi_i + \tau A(\varphi_i))_{\Omega_i} = 0$*. Comparing with* (13)*, only the penalty on the violation of the positivity remains, but the term is no longer symmetric.*

Let us first establish that $a_h^{\tau\gamma}$ has reasonable monotonicity properties.

Lemma 3.1 (Monotonicity) *Assume that* $0 < \gamma \leq \tau$. *Then the following holds for all* $u_1, u_2 \in V$*:*

$$\frac{1}{2}(|\!|\!| u_1 - u_2 |\!|\!|^2 + \|\gamma^{-\frac{1}{2}}(\xi^\gamma(u_1) - \xi^\gamma(u_2))\|_\Omega^2) \leq a_h^{\tau\gamma}(u_1; u_1 - u_2) - a_h^{\tau\gamma}(u_2; u_1 - u_2), \tag{15a}$$

$$\frac{1}{4}(|\!|\!| u_1 |\!|\!|^2 + \|\gamma^{-\frac{1}{2}}\xi^\gamma(u_1)\|_\Omega^2) \leq a_h^{\tau\gamma}(u_1; u_1) + \|\gamma^{\frac{1}{2}} f\|_\Omega^2. \tag{15b}$$

Proof. Let us prove (15a). We observe that

$$\begin{aligned} a_h^{\tau\gamma}(u_1;&u_1 - u_2) - a_h^{\tau\gamma}(u_2; u_1 - u_2) \\ &= a_h^{\tau}(u_1 - u_2, u_1 - u_2) + (\gamma^{-1}(\xi^\gamma(u_1) - \xi^\gamma(u_2)), u_1 - u_2)_\Omega \\ &\geq |\!|\!| u_1 - u_2 |\!|\!|^2 + (\gamma^{-1}(\xi^\gamma(u_1) - \xi^\gamma(u_2)), u_1 - u_2)_\Omega, \end{aligned}$$

where we have used (8). Moreover, we have

$$\begin{aligned} &(\gamma^{-1}(\xi^\gamma(u_1) - \xi^\gamma(u_2)), u_1 - u_2)_\Omega \\ &\quad = (\gamma^{-1}(\xi^\gamma(u_1) - \xi^\gamma(u_2)), u_1 - \gamma(A(u_1) - f) - (u_2 - \gamma(A(u_2) - f)))_\Omega \\ &\qquad + (\xi^\gamma(u_1) - \xi^\gamma(u_2), A(u_1 - u_2))_\Omega \\ &\quad \geq \|\gamma^{-\frac{1}{2}}(\xi^\gamma(u_1) - \xi^\gamma(u_2))\|_\Omega^2 + (\xi^\gamma(u_1) - \xi^\gamma(u_2), A(u_1 - u_2))_\Omega, \end{aligned}$$

where we have used the fact that

$$|x_- - y_-|^2 \leq (x_- - y_-)(x - y), \qquad \forall x, y \in \mathbb{R}. \tag{16}$$

Using Young's inequality and the fact that $\gamma \leq \tau$, we infer that

$$((\xi^\gamma(u_1) - \xi^\gamma(u_2)), A(u_1 - u_2))_\Omega \geq -\frac{1}{2}\|\gamma^{-\frac{1}{2}}(\xi^\gamma(u_1) - \xi^\gamma(u_2))\|_\Omega^2 - \frac{1}{2}\|\tau^{\frac{1}{2}} A(u_1 - u_2)\|_\Omega^2.$$

Putting everything together shows that (15a) holds. Finally, the proof of (15b) follows from (15a) by taking $u_2 = 0$ and using the fact that $\frac{1}{2}\|\gamma^{-\frac{1}{2}}(\xi^\gamma(u_1) - \xi^\gamma(0))\|_\Omega^2 \leq \|\gamma^{-\frac{1}{2}}\xi^\gamma(u_1)\|_\Omega^2 + \|\gamma^{-\frac{1}{2}}\xi^\gamma(0)\|_\Omega^2 \leq \|\gamma^{-\frac{1}{2}}\xi^\gamma(u_1)\|_\Omega^2 + \|\gamma^{\frac{1}{2}} f\|_\Omega^2$.

We can now prove that the discrete problem (11) is well-posed.

Proposition 3.1 (Well-posedness) *Assume that* $0 < \gamma \leq \tau$. *Then the discrete problem* (11) *admits one and only one solution.*

Proof. Uniqueness follows from (15a). To prove existence, let $N := \dim V_h^k$ and let $G:\mathbb{R}^N\to\mathbb{R}^N$ be the map defined by $(G(U),V)_{\mathbb{R}^N}:=a_h^{\tau\gamma}(u_h;v_h)-\ell_h^\tau(v_h)$, where $U,V\in\mathbb{R}^N$ are the component vectors associated with the functions u_h,v_h in the Lagrange basis of V_h^k. It is readily seen that G is a continuous map (observe in particular that $|x_--y_-|\le|x-y|$ for all $x,y\in\mathbb{R}$). Furthermore, since Cauchy–Schwarz inequalities and $\tau\le\sigma_0^{-1}$ show that $|\ell_h^\tau(v_h)|\le K|\!|\!|v_h|\!|\!|$ with $K=(2\sigma_0^{-\frac{1}{2}}\|f\|_\Omega+\||\beta{\cdot}n|^{\frac{1}{2}}g\|_{\partial\Omega^-})$, we infer using (15b) that

$$\begin{aligned}(G(U),U)_{\mathbb{R}^N}&=a_h^{\tau\gamma}(u_h;u_h)-\ell_h^\tau(u_h)\\&\ge\tfrac{1}{4}(|\!|\!|u_h|\!|\!|^2+\|\gamma^{-\frac{1}{2}}\xi^\gamma(u_h)\|_\Omega^2)-\|\gamma^{\frac{1}{2}}f\|_\Omega^2-K|\!|\!|u_h|\!|\!|.\end{aligned}$$

This proves that there is a real number, say K', so that $(G(U),U)_{\mathbb{R}^N}>0$ for all $U\in\mathbb{R}^N$ with $\|U\|_{\mathbb{R}^N}\ge K'$. Indeed, using norm equivalence on discrete spaces, we infer that there exists $C_N>0$ such that $C_N\|U\|_{\mathbb{R}^N}\le|\!|\!|u_h|\!|\!|$ for all $U\in\mathbb{R}^N$ with associated discrete function $u_h\in V_h^k$. This leads to

$$(G(U),U)_{\mathbb{R}^N}\ge\tfrac{1}{8}|\!|\!|u_h|\!|\!|^2-\|\gamma^{\frac{1}{2}}f\|_\Omega^2-2K^2\ge\tfrac{1}{8}C_N^2\|U\|_{\mathbb{R}^N}^2-\|\gamma^{\frac{1}{2}}f\|_\Omega^2-2K^2,$$

and we conclude that the expected inequality holds with

$$K'=\frac{8}{C_N}\sqrt{\|\gamma^{\frac{1}{2}}f\|_\Omega^2+2K^2+1}.$$

Existence then follows using well-known arguments (see, for instance, [12, Lemma 1.4, Chapter 2]).

The next theorem is the main result of this paper. It shows that the GaLS finite element method with penalty has essentially the same behavior as that without penalty when approximating smooth solutions.

Theorem 3.1 (Error estimate) *Let $u\in V$ be the solution to* (3) *and let $u_h\in V_h^k$ be the solution to* (11). *Assume that $0<\gamma\le\tau$. Then there exists $C>0$, uniform, such that*

$$\begin{aligned}&|\!|\!|u-u_h|\!|\!|+\|\gamma^{-\frac{1}{2}}[u_h]_-\|_\Omega\le\\&\quad C\inf_{v_h\in V_h^k}(\|\tau^{\frac{1}{2}}A(u-v_h)\|_\Omega+\||\beta{\cdot}n|^{\frac{1}{2}}(u-v_h)\|_{\partial\Omega}+\|\gamma^{-\frac{1}{2}}(u-v_h)\|_\Omega).\end{aligned}\quad(17)$$

Moreover, if $u\in H^{k+1}(\Omega)$, τ is chosen as in the GaLS method as $\tau|_T=\min(\sigma_0^{-1},\beta_T^{-1}h_T)$, and $c\tau|_T\le\gamma|_T$ for all $T\in\mathcal{T}_h$ with c uniformly bounded from below away from zero, then

$$|\!|\!|u-u_h|\!|\!|+\|\gamma^{-\frac{1}{2}}[u_h]_-\|_\Omega\le C\left(\sum_{T\in\mathcal{T}_h}\phi_T h_T^{2k+1}\|u\|_{H^{k+1}(T;\mathbb{R})}^2\right)^{\frac{1}{2}}.\quad(18)$$

Proof. Let $e = u - u_h$. Then, using (8), we infer that

$$|\!|\!| e |\!|\!|^2 \leq a_h^\tau(e,e) = a_h^\tau(e, u - v_h) + a_h^\tau(e, v_h - u_h).$$

Moreover, the exact consistency of the GaLS approximation and the definition of the discrete problem (11) imply that

$$\begin{aligned} a_h^\tau(e, v_h - u_h) &= \ell_h^\tau(v_h - u_h) - \ell_h^\tau(v_h - u_h) + (\gamma^{-1}\xi^\gamma(u_h), v_h - u_h)_\Omega \\ &= (\gamma^{-1}\xi^\gamma(u_h), v_h - u_h)_\Omega \\ &= (\gamma^{-1}\xi^\gamma(u_h), v_h - u + \gamma A(e))_\Omega + (\gamma^{-1}\xi^\gamma(u_h), u - u_h - \gamma A(e))_\Omega. \end{aligned}$$

Since $\xi^\gamma(u) = 0$, using the monotonicity property (16), we infer that

$$(\gamma^{-1}\xi^\gamma(u_h), u - u_h - \gamma A(e))_\Omega \leq -\|\gamma^{-\frac{1}{2}}\xi^\gamma(u_h)\|_\Omega^2.$$

As a result, we obtain

$$|\!|\!| e |\!|\!|^2 + \|\gamma^{-\frac{1}{2}}\xi^\gamma(u_h)\|_\Omega^2 \leq a_h^\tau(e, u - v_h) + (\gamma^{-1}\xi^\gamma(u_h), v_h - u + \gamma A(e))_\Omega.$$

The boundedness properties of the GaLS approximation yield

$$a_h^\tau(e, u - v_h) \leq |\!|\!| e |\!|\!| (\|\tau^{\frac{1}{2}} A(u - v_h)\|_\Omega + \||\beta{\cdot}n|^{\frac{1}{2}}(u - v_h)\|_{\partial\Omega} + \|\tau^{-\frac{1}{2}}(u - v_h)\|_\Omega).$$

Moreover, we have

$$\begin{aligned} (\gamma^{-1}\xi^\gamma(u_h), v_h - u + \gamma A(e))_\Omega &\leq \|\gamma^{-\frac{1}{2}}\xi^\gamma(u_h)\|_\Omega (\|\gamma^{-\frac{1}{2}}(u - v_h)\|_\Omega + \|\gamma^{\frac{1}{2}} A(e)\|_\Omega) \\ &\leq \frac{3}{4}\|\gamma^{-\frac{1}{2}}\xi^\gamma(u_h)\|_\Omega^2 + \|\gamma^{-\frac{1}{2}}(u - v_h)\|_\Omega^2 + \frac{1}{2}\|\gamma^{\frac{1}{2}} A(e)\|_\Omega^2. \end{aligned}$$

Collecting these two bounds and using that $\gamma \leq \tau$, we infer that $|\!|\!| e |\!|\!| + \|\gamma^{-\frac{1}{2}}\xi^\gamma(u_h)\|_\Omega$ is bounded by the right-hand side of (17). To conclude that (17) holds, it suffices to establish that

$$\|\gamma^{-\frac{1}{2}}[u_h]_-\|_\Omega \leq \|\gamma^{-\frac{1}{2}}\xi^\gamma(u_h)\|_\Omega + |\!|\!| e |\!|\!|.$$

This inequality, in turn, follows from the elementary inequality $|[x + y]_-| \leq |x_-| + |y_-|$ for arbitrary real numbers x and y, leading to

$$\|\gamma^{-\frac{1}{2}}[u_h]_-\|_\Omega \leq \|\gamma^{-\frac{1}{2}}\xi^\gamma(u_h)\|_\Omega + \|\gamma^{\frac{1}{2}}[(A(u_h) - f)]_-\|_\Omega,$$

together with the fact that $A(u) = f$ and the assumption that $\gamma \leq \tau$ so that $\|\gamma^{\frac{1}{2}}[(A(u_h) - f)]_-\|_\Omega \leq |\!|\!| e |\!|\!|$. Finally, (18) results from the approximation properties of finite elements and the assumptions on τ and γ.

The error estimates derived in Theorem 3.1 show that the present consistent penalty method delivers similar bounds on the error $|\!|\!| u - u_h |\!|\!|$ to those obtained with the usual GaLS discretization, while additionally controlling the violation of positivity by means of the measure $\|\gamma^{-\frac{1}{2}}[u_h]_-\|_\Omega$ (note that γ scales as the mesh size, so that the factor $\gamma^{-\frac{1}{2}}$ in front of $[u_h]_-$ makes the bound even stronger.

4 Numerical example

In this section, we assess the proposed method on two test cases: the first one features a solution with inner layer, and the second one a solution with discontinuity.

4.1 Test case 1: solution with inner layer

We consider problem (3) in the domain $\Omega \subset \mathbb{R}^2$ shown in the left panel of Figure 1, with

$$\beta = \frac{1}{(x^2+y^2)^{\frac{1}{2}}}(y, -x)^{\mathrm{T}}, \quad \sigma = 0, \quad f = 0.$$

The advection field β rotates clockwise, and the inflow boundary corresponds to the part of $\partial\Omega$ where $x = 0$. The exact solution given by $u = \frac{1}{2}(\tanh(((x^2 + y^2)^{\frac{1}{2}} - 0.5)/\epsilon) + 1.0)$ is a consequence of inflow data imposed on the inflow boundary (see the contourlines in the right panel of Figure 1). The boundary data has a sharp layer of width ϵ at $y = 0.5$. This creates spurious under- and overshoots that are transported downstream throughout the domain.

We compute approximate solutions for both a mild layer ($\epsilon = 0.1$) and a sharp layer ($\epsilon = 0.01$) using either the nonlinear method (11) with the bilinear form $a_h^{\tau\gamma}$ defined by (12) or the standard (linear) GaLS method obtained by dropping the nonlinear term. Computations not reported here indicate that using the definition (14) leads to similar results. We consider affine ($k = 1$) and quadratic ($k = 2$) finite elements on a sequence of quasi-uniform unstructured meshes characterized by the mesh sizes $h = 0.09 \times 2^{-l}$ with $l \in \{0, 1, 2, 3, 4\}$.

The nonlinear penalty term is evaluated using nodal quadrature when $k = 1$ and a quadrature that is exact for polynomials of order up to five when $k = 2$. Since the integrand of the nonlinear term is only H^1 in the interior of the elements, an exact quadrature requires a careful local analysis of where the nonlinearity is active. To assess the effect of possible under-integration of the nonlinear term, we also consider, in the case where $k = 2$, evaluating the nonlinear term with a hybrid quadrature rule obtained as a linear combination of two low-order terms, one using nodal quadrature and the other midpoint

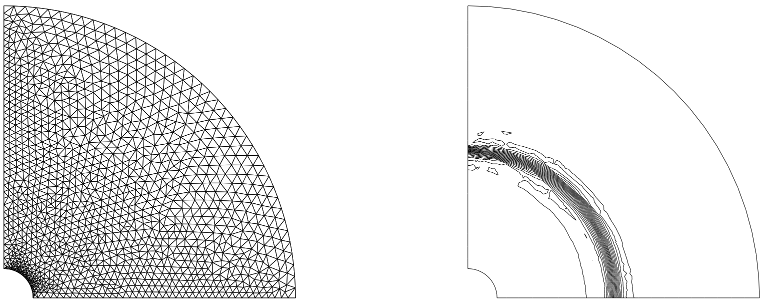

Figure 1. Test case 1. Left: computational mesh. Right: contourlines of discrete solution for $\epsilon = 0.01$

l	$\|e\|_\Omega$	$\|\beta\cdot\nabla e\|_\Omega$	$e_{\min}$	$e_{\max}$	$\Phi(u_h)$
0	1.13e-2 (–)	9.13e-2 (–)	–	1.09e-2	4.59-3
1	2.33e-3 (2.3)	2.65e-2 (1.8)	–	1.93e-3	1.49e-4
2	9.57e-4 (1.3)	1.94e-2 (0.4)	–	9.20e-4	2.72e-5
3	2.54e-4 (1.9)	9.92e-3 (1.0)	–	3.90e-4	8.59e-6
4	6.35e-5 (2.0)	4.64e-3 (1.1)	–	1.6e-4	2.17e-6

Table 1
Test case 1, mild layer ($\epsilon = 0.1$). Nonlinear method with consistent penalty, $k = 1$, $\gamma = 0.0001h$, $\tau = h/2$

quadrature. The hybrid quadrature is insufficient to resolve the integration of the nonlinearity, but it gives control on the degrees of freedom of the quadratic polynomials.

We set the penalty parameter to $\gamma = 0.0001h$ for $k = 1$ and to $\gamma = 0.005h$ for $k = 2$ when consistent quadrature is used and to $\gamma = 0.0001h$ for the low-order quadrature. Our results show that these choices are sufficient to reduce undershoots to less than one percent in all cases. Strengthening the penalty in the case of quadratic approximation does not improve the positivity, but increases the stiffness of the nonlinear problem. We present in Tables 1 to 5 the results for (i) the error $e = u - u_h$ in the L^2-norm and in the streamline derivative (experimental convergence orders are given between parenthesis), (ii) the violations of the maximum principle evaluated as $e_{\min} := -\min_{x_i \in \mathcal{N}}(u_h(x_i))_-$ and $e_{\max} := \max_{x_i \in \mathcal{N}} u(x_i) - 1$, where $\mathcal{N}$ denotes the set of nodes for the Lagrange basis functions (the symbol '–' means that the discrete maximum principle is actually satisfied), and (iii) the error on the global conservation property $\Phi(u_h) := |\int_{\partial\Omega}(\beta\cdot n)u_h \, \mathrm{d}s| = 0$. Note that the lack of exact global conservation for the linear problem is due to quadrature errors. Note also that we only impose weakly by the consistent penalty method the lower bound on the discrete solution; the upper bound could be imposed similarly.

l	$\|e\|_\Omega$	$\|\beta\cdot\nabla e\|_\Omega$	$e_{\min}$	$e_{\max}$	$\Phi(u_h)$
0	1.18e-2 (–)	9.69e-2 (–)	2.37e-2	1.09e-2	3.44e-4
1	2.33e-3 (2.3)	2.65e-2 (1.9)	–	1.93e-3	1.45e-4
2	9.57e-4 (1.3)	1.94e-2 (0.4)	–	9.20e-4	2.72e-5
3	2.54e-4 (1.9)	9.92e-3 (1.0)	–	3.90e-4	8.59e-6
4	6.35e-5 (2.0)	4.64e-3 (1.1)	–	1.60e-4	2.17e-6

Table 2
Test case 1, mild layer ($\epsilon = 0.1$). Linear GaLS method, $k = 1$, $\tau = h/2$

l	$\|e\|_\Omega$	$\|\beta\cdot\nabla e\|_\Omega$	$e_{\min}$	$e_{\max}$	$\Phi(u_h)$
0	7.24e-2 (–)	5.63e-1 (–)	8.37e-6	8.20e-2	2.11e-3
1	5.26e-2 (0.5)	4.90e-1 (0.2)	–	1.14e-1	8.34e-4
2	2.56e-2 (1.0)	6.53e-1 (-0.4)	–	6.70e-2	1.40e-3
3	1.54e-2 (0.7)	6.27e-1 (0.1)	–	5.97e-2	1.40e-3
4	5.20e-3 (1.6)	2.97e-1 (1.0)	–	2.04e-2	3.05e-4

Table 3
Test case 1, sharp layer ($\epsilon = 0.01$). Nonlinear method with consistent penalty, $k = 1$, $\gamma = 0.0001h$, $\tau = h/2$

l	$\|e\|_\Omega$	$\|\beta\cdot\nabla e\|_\Omega$	$e_{\min}$	$e_{\max}$	$\Phi(u_h)$
0	7.43e-2 (–)	5.64e-1 (–)	9.94e-2	8.86e-2	5.68e-4
1	5.23e-2 (0.5)	5.04e-1 (0.2)	7.01e-2	1.13e-1	1.41e-4
2	2.69e-2 (1.0)	6.67e-1 (-0.4)	6.41e-2	6.77e-2	1.84e-5
3	1.59e-2 (0.8)	6.47e-1 (0.0)	6.01e-2	5.88e-2	1.20e-5
4	5.31e-3 (1.6)	2.95e-1 (1.1)	1.94e-2	2.0e-2	2.01e-6

Table 4
Test case 1, sharp layer ($\epsilon = 0.01$). Linear GaLS method, $k = 1$, $\tau = h/2$

For the above results on the nonlinear method with consistent penalty, the nonlinear system is solved using fixed-point iteration to an accuracy of $TOL = 10^{-6}$ on the increment $\|u_h^k - u_h^{k+1}\|_\Omega$ where k is the index of the fixed-point iteration. This convergence criterion has been used for the sole purpose of numerical illustrations. In practice, a computationally-effective possibility is to prescribe the fixed-point convergence tolerance so that the error induced by the stopping criterion of the iteration is comparable to that of the a priori estimate.

l	$\|e\|_\Omega$	$\|\beta\cdot\nabla e\|_\Omega$	$e_{\min}$	$e_{\max}$	$\Phi(u_h)$
0	5.60e-2 (–)	6.47e-1 (–)	4.36e-3	7.58e-2	5.82e-3
1	2.77e-2 (1.0)	7.17e-1 (-0.1)	5.25e-3	6.27e-2	2.47e-3
2	1.05e-2 (1.4)	4.39e-1 (0.7)	1.81e-4	3.91e-2	1.06e-3
3	2.86e-3 (1.9)	1.90e-1 (1.2)	4.07e-4	8.81e-3	1.10e-4
4	3.96e-4 (2.9)	4.07e-2 (2.2)	2.31-19	–	3.78e-9

Table 5
Test case 1, sharp layer ($\epsilon = 0.01$). Nonlinear method with consistent penalty, fifth-order quadrature, $k = 2$, $\gamma = 0.005h$, $\tau = h/2$

l	$\|e\|_\Omega$	$\|\beta\cdot\nabla e\|_\Omega$	$e_{\min}$	$e_{\max}$	$\Phi(u_h)$
0	5.40e-2 (–)	6.27e-1 (–)	9.35e-6	7.66e-2	7.47e-3
1	2.60e-2 (1.1)	6.68e-1 (-0.1)	8.87e-6	6.21e-2	2.83e-3
2	1.03e-2 (1.3)	4.50e-1 (0.6)	–	3.80e-2	1.37e-3
3	2.86e-3 (1.8)	1.90e-1 (1.2)	–	8.82e-3	1.20e-4
4	3.96e-4 (2.9)	4.07e-2 (2.2)	–	–	3.78e-9

Table 6
Test case 1, sharp layer ($\epsilon = 0.01$). Nonlinear method with consistent penalty, hybrid quadrature, $k = 2$, $\gamma = 0.0001h$, $\tau = h/2$

l	$\|e\|_\Omega$	$\|\beta\cdot\nabla e\|_\Omega$	$e_{\min}$	$e_{\max}$	$\Phi(u_h)$
0	5.33e-2 (–)	6.58e-1 (–)	9.26e-2	7.68e-2	3.34e-4
1	2.56e-2 (1.1)	7.29e-1 (-0.1)	7.82e-2	6.22e-2	5.84e-5
2	1.05e-2 (1.3)	4.70e-1 (0.6)	4.17e-2	3.83e-2	2.07e-6
3	2.93e-3 (1.8)	1.92e-1 (1.3)	1.03e-2	8.82e-3	5.31e-8
4	3.96e-4 (2.9)	4.07e-2 (2.2)	4.76e-18	–	3.78e-9

Table 7
Test case 1, sharp layer ($\epsilon = 0.01$). Linear GaLS method, $k = 2$, $\tau = h/2$

Assume that the fixed-point iteration is a contraction so that for some $0 < \delta < 1$, the approximation at iteration k of u_h, say u_h^k, satisfies

$$\|u_h - u_h^k\|_\Omega \leq \delta \|u_h - u_h^{k-1}\|_\Omega.$$

Since

$$\|u_h - u_h^k\|_\Omega \leq \|u_h - u_h^{k+1}\|_\Omega + \|u_h^k - u_h^{k+1}\|_\Omega \leq \delta\|u_h - u_h^k\|_\Omega + \|u_h^k - u_h^{k+1}\|_\Omega,$$

l	$\|e\|_\Omega$	$\|\beta\cdot\nabla e\|_\Omega$	$e_{\min}$	$e_{\max}$	$\Phi(u_h)$
0	7.13e-2 (–)	5.57e-1 (–)	–	8.17e-2	6.04e-4
1	5.23e-2 (0.4)	4.90e-1 (0.2)	–	1.13e-1	5.50e-4
2	2.56e-2 (1.0)	6.53e-1 (-0.4)	–	6.70e-2	1.40e-3
3	1.54e-2 (0.7)	6.27e-1 (0.1)	–	5.97e-2	1.40e-3
4	5.20e-3 (1.6)	2.97e-1 (1.0)	–	2.04e-2	3.05e-4

Table 8
Test case 1, sharp layer ($\epsilon = 0.01$). Nonlinear method, balanced fixed-point iteration, $k = 1$, $\gamma = 0.0001h$, $\tau = h/2$

we infer that
$$\|u_h - u_h^k\|_\Omega \le (1-\delta)^{-1}\|u_h^k - u_h^{k+1}\|_\Omega.$$
Then, we obtain the following error estimate for the error at the iteration k:
$$\|u_h^k - u\|_\Omega \le \|u_h - u\|_\Omega + \|u_h - u_h^k\| \le Ch^{k+\frac{1}{2}} + (1-\delta)^{-1}\|u_h^k - u_h^{k+1}\|_\Omega.$$

Hence, if δ stays uniformly bounded away from 1 during the fixed-point iterations, we can reasonably stop the iterations whenever $\|u_h^k - u_h^{k+1}\|_\Omega \sim Ch^{k+\frac{1}{2}} \sim e_0/(2^l)^{k+\frac{1}{2}}$, where e_0 denotes the L^2-norm error on the coarsest mesh and l denotes the level of mesh refinement. We call the resulting iterative method the balanced fixed-point iteration. We present the results using balanced fixed-point iteration with $TOL = 0.01/(2^l)^{k+\frac{1}{2}}$ in Tables 8 and 9. We observe that the results are of comparable or even better quality when the balanced fixed-point iterations are used.

In this context, it is also interesting to compare the behaviour of the present consistent penalty method with a method employing diffusion-based shock-capturing terms with artificial viscosity depending on the residual. We observe that for the present method, the computational cost is reduced as the accuracy of the solution improves and the violation of the DMP is isolated close to layers, whereas no such reduction is observed for nonlinear diffusion where the nonlinearity appears to have a much more global character. Indeed, it is known that the effects of the nonlinearity for shock-capturing can propagate into the zones where the solution is smooth in the form of gradient oscillations (this is sometimes called a terracing phenomenon). No such spurious gradient fluctuations were observed for the present consistent penalty method.

4.2 Test case 2: solution with discontinuity

In this section we apply the present consistent penalty method to approximate a solution with a discontinuity. The example that we consider is taken from

l	$\|e\|_\Omega$	$\|\beta\cdot\nabla e\|_\Omega$	$e_{\min}$	$e_{\max}$	$\Phi(u_h)$
0	5.30e-2 (–)	6.56e-1 (–)	–	7.68e-2	9.08e-3
1	2.60e-2 (1.0)	7.30e-1 (-0.15)	–	6.21e-2	1.83e-3
2	1.03e-2 (1.3)	4.63e-1 (0.7)	–	3.80e-2	1.46e-3
3	2.87e-3 (1.8)	1.98e-1 (1.2)	–	8.82e-3	1.20e-4
4	3.96e-4 (2.9)	4.07e-2 (2.3)	–	–	3.78e-9

Table 9
Test case 1, sharp layer ($\epsilon = 0.01$). Nonlinear method, hybrid quadrature, balanced fixed-point iteration, $k = 2$, $\gamma = 0.0001h$, $\tau = h/2$

[13]. The equation is similar to in the previous example, but this time we set $\Omega = (-1,1)\times(0,1)$, take $\beta = (y,-x)^{\mathrm{T}}$ (so that $\partial\Omega^- = (-1,0)\times\{0\}\cup\{-1\}\times(0,1)\cup(0,1)\times\{1\}$) and use the inflow data

$$g = \begin{cases} 1 & \text{on } (-0.65,-0.35)\times\{0\},\\ 0 & \text{elsewhere on } \partial\Omega^-. \end{cases}$$

The corresponding exact solution reads

$$u = \begin{cases} 1 & \text{if } 0.35 \le \sqrt{x^2+y^2} \le 0.65,\\ 0 & \text{elsewhere in } \Omega. \end{cases}$$

We compute the approximate solutions to this problem on structured meshes with mesh-size $h = 0.1 \times 2^{-l}$, $l \in \{0,\dots,4\}$. First, using the linear GaLS method and piecewise linear or quadratic finite elements, we recorded violations of positivity of more than 14% for the former case and more than 11% in the latter on all meshes. Using the present consistent penalty method with lumped quadrature for linear elements and the hybrid low-order quadrature (using vertices and midpoints) for quadratic elements resulted in strictly nodally positive solutions in all cases. Balanced fixed-point iterations were also used and we observed convergence after two iterations, yielding an accuracy similar to that of the linear method. In Figure 2, we report an illustration for the piecewise linear case with $h = 1/20$. In Figure 3 we report the contour plots of the nodal interpolant of the negative part of the solution. The magnitude of the DMP violation in the linear case is of the order 15% and in the nonlinear case of the order $4\cdot 10^{-3}$%. The elevation of the linear GaLS method is presented in the left panel and that of the present method in the right one, here a term eliminating local overshoots has been added as well. In Figure 4 and 5, we report the same results, but this time for piecewise quadratic approximation and $h = 1/10$. The magnitude of the DMP violation in the linear case is of the order 20% and in the nonlinear case of the order

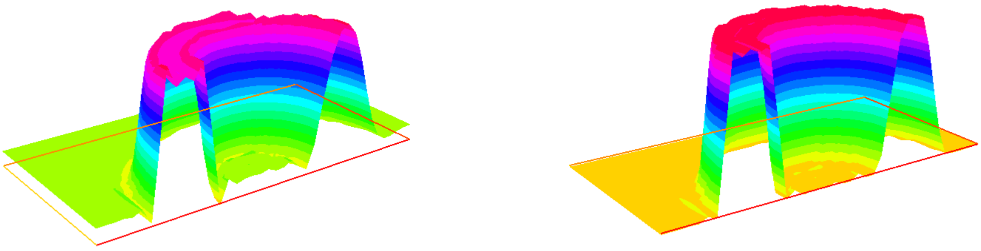

Figure 2. Test case 2. Elevations of solutions using piecewise affine elements. Left: Linear GaLS, violation DMP 15%. Right: consistent penalty method with lumped quadrature and balanced fixed-point iteration, nodal violation DMP less than $4 \cdot 10^{-3}\%$

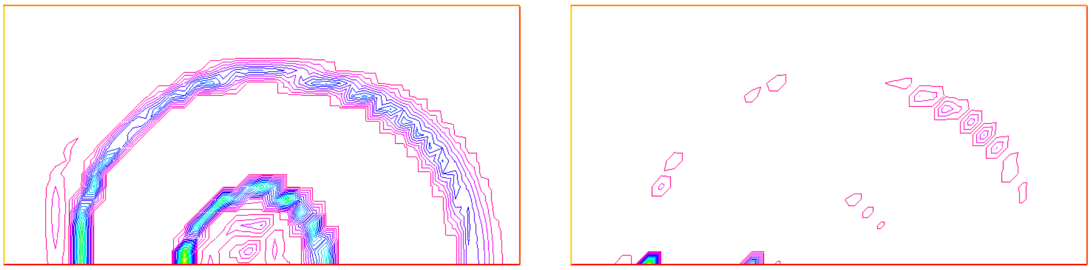

Figure 3. Test case 2. Contour plots of the negative part of the interpolant of solutions using piecewise affine elements. Left: Linear GaLS, violation DMP 15%. Right: consistent penalty method with lumped quadrature and balanced fixed-point iteration, nodal violation DMP less than $4 \cdot 10^{-3}\%$

$4 \cdot 10^{-3}\%$. Observe that in spite of the small nodal DMP violation, the piecewise quadratic approximation is observed to violate the DMP with up to 10% due to the non-positivity of the quadratic basis functions.

References

[1] A. Mizukami, T. J. R. Hughes, A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle, Comput. Methods Appl. Mech. Engrg. 50 (2) (1985) 181–193.

[2] E. Burman, A. Ern, Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence, Math. Comp. 74 (252) (2005) 1637–1652 (electronic).

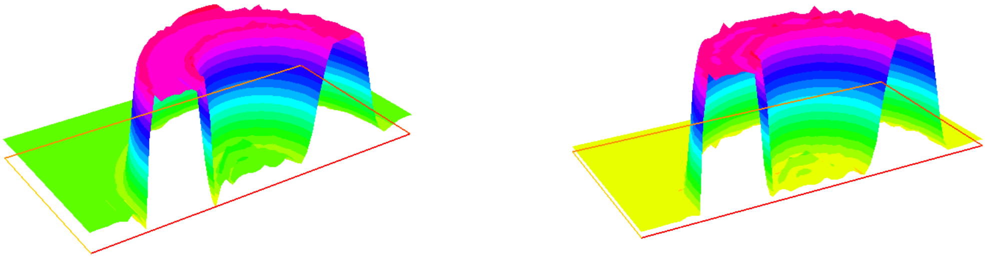

Figure 4. Test case 2. Elevations of solutions using piecewise quadratic elements. Left: Linear GaLS, nodal violation DMP 21%. Right: consistent penalty method with lumped quadrature and balanced fixed-point iteration, nodal violation DMP less than $4 \cdot 10^{-3}\%$

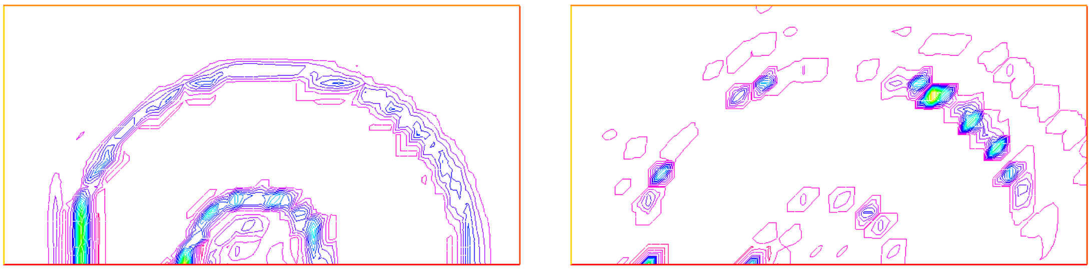

Figure 5. Test case 2. Contour plots of the negative part of the interpolant of solutions using piecewise quadratic elements. Left: Linear GaLS, nodal violation DMP 21%. Right: consistent penalty method with lumped quadrature and balanced fixed-point iteration, nodal violation DMP less than $4 \cdot 10^{-3}\%$

[3] R. Löhner, K. Morgan, J. Peraire, M. Vahdati, Finite element flux-corrected transport (FEM–FCT) for the Euler and Navier–Stokes equations, Int. J. Numer. Meths. Fluids 7 (10) (1987) 1093–1109.

[4] D. Kuzmin, S. Turek, Flux correction tools for finite elements, J. Comput. Phys. 175 (2) (2002) 525–558.

[5] J.-L. Guermond, M. Nazarov, B. Popov, Y. Yang, A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations, SIAM J. Numer. Anal. 52 (4) (2014) 2163–2182.

[6] J. A. Evans, T. J. R. Hughes, G. Sangalli, Enforcement of constraints and maximum principles in the variational multiscale method, Comput. Methods Appl. Mech. Engrg. 199 (1-4) (2009) 61–76. doi:10.1016/j.cma.2009.09.019. URL `http://dx.doi.org/10.1016/j.cma.2009.09.019`

[7] F. Chouly, P. Hild, A Nitsche-based method for unilateral contact problems: numerical analysis, SIAM J. Numer. Anal. 51 (2) (2013) 1295–1307.

doi:10.1137/12088344X.
URL `http://dx.doi.org/10.1137/12088344X`

[8] E. Burman, P. Hansbo, M. G. Larson, R. Stenberg, Galerkin least squares finite element method for the obstacle problem, Comput. Methods Appl. Mech. Engrg. 313 (2017) 362–374. doi:10.1016/j.cma.2016.09.025.
URL `http://dx.doi.org/10.1016/j.cma.2016.09.025`

[9] C. Johnson, U. Nävert, J. Pitkäranta, Finite element methods for linear hyperbolic problems, Comput. Methods Appl. Mech. Engrg. 45 (1-3) (1984) 285–312. doi:10.1016/0045-7825(84)90158-0.
URL `http://dx.doi.org/10.1016/0045-7825(84)90158-0`

[10] A. Ern, J.-L. Guermond, Theory and Practice of Finite Elements, Vol. 159 of Applied Mathematical Sciences, Springer-Verlag, New York, NY, 2004.

[11] A. Ern, J.-L. Guermond, Discontinuous Galerkin methods for Friedrichs' systems. I. General theory, SIAM J. Numer. Anal. 44 (2) (2006) 753–778.

[12] R. Temam, Navier-Stokes equations, AMS Chelsea Publishing, Providence, RI, 2001, theory and numerical analysis, Reprint of the 1984 edition. doi:10.1090/chel/343.
URL `http://dx.doi.org/10.1090/chel/343`

[13] D. Kuzmin, M. Möller, Goal-oriented mesh adaptation for flux-limited approximations to steady hyperbolic problems, J. Comput. Appl. Math. 233 (12) (2010) 3113–3120. doi:10.1016/j.cam.2009.07.026.
URL `http://dx.doi.org/10.1016/j.cam.2009.07.026`