MATHEMATICS OF COMPUTATION
Volume 74, Number 252, Pages 1637–1652
S 0025-5718(05)01761-8
Article electronically published on June 7, 2005

STABILIZED GALERKIN APPROXIMATION OF CONVECTION-DIFFUSION-REACTION EQUATIONS: DISCRETE MAXIMUM PRINCIPLE AND CONVERGENCE

ERIK BURMAN AND ALEXANDRE ERN

Abstract. We analyze a nonlinear shock-capturing scheme for H^1-conforming, piecewise-affine finite element approximations of linear elliptic problems. The meshes are assumed to satisfy two standard conditions: a local quasi-uniformity property and the Xu–Zikatanov condition ensuring that the stiffness matrix associated with the Poisson equation is an M-matrix. A discrete maximum principle is rigorously established in any space dimension for convection-diffusion-reaction problems. We prove that the shock-capturing finite element solution converges to that without shock-capturing if the cell Péclet numbers are sufficiently small. Moreover, in the diffusion-dominated regime, the difference between the two finite element solutions super-converges with respect to the actual approximation error. Numerical experiments on test problems with stiff layers confirm the sharpness of the a priori error estimates.

1. Introduction

In many applications, it is important to design approximation methods guaranteeing that the discrete solution satisfies some maximum principle. For instance, in the computation of chemically reacting flows, the species concentrations should remain nonnegative.

In this paper, we investigate the convection-diffusion-reaction problem

$$\beta \cdot \nabla u + \sigma u - \varepsilon \Delta u = f \quad \text{a.e. in } \Omega, \tag{1.1}$$

$$u = g \quad \text{a.e. on } \partial\Omega, \tag{1.2}$$

where Ω is an open bounded connected subset of $\mathbb{R}^d$ with a Lipschitz boundary $\partial\Omega$, β is the velocity field, d is the space dimension, $\varepsilon > 0$ is the (constant) diffusion coefficient, and σ is the reaction coefficient. Henceforth, it is assumed that β is in the Sobolev space $W^{1,\infty}(\Omega)$, σ is in $L^\infty(\Omega)$, f is in $L^2(\Omega)$, and g is in $H^{\frac{1}{2}}(\partial\Omega)$. It is also assumed that

$$\sigma \geq 0 \quad \text{a.e. in } \Omega \tag{1.3}$$

and that there exists a constant $\sigma_0 \geq 0$ such that

$$\sigma - \tfrac{1}{2}\nabla \cdot \beta \geq \sigma_0 \quad \text{a.e. in } \Omega. \tag{1.4}$$

Owing to the above assumptions, the Lax–Milgram Lemma implies that problem (1.1)–(1.2) is well posed. It also satisfies a maximum principle, i.e., under some

Received by the editor February 18, 2003 and, in revised form, August 16, 2004.
2000 *Mathematics Subject Classification.* Primary 65N12, 65N30, 76R99.

assumptions on the data f and g, the solution attains its maximum or minimum at the boundary.

An approximation method to (1.1)–(1.2) satisfies a so-called discrete maximum principle (DMP for short) if the relevant maximum principle is transferred to the discrete problem. For the Poisson equation, Ciarlet and Raviart established that the standard H^1-conforming, piecewise-affine finite element approximation satisfies a DMP on weakly acute meshes [6]. The weakly acute condition was sharpened by Xu and Zikatanov [20]; see also [17] for the Poisson problem in three dimensions and [8] for an analysis of the origin of DMP failure. In many practical problems, nonphysical oscillations violating the maximum principle are triggered by the lower-order terms in (1.1). When the convective term is dominant, standard Galerkin approximations become unstable, and stabilizing terms must be added in order to increase robustness and accuracy. Stabilized finite element methods for convection-diffusion problems include least-squares techniques [1, 13], Galerkin least-squares techniques such as the SUPG method or the streamline diffusion (SD) method [2, 9, 14], subgrid viscosity techniques [10], and continuous interior penalty (CIP) techniques based on gradient jumps across element interfaces [5]. In general, such stabilizations do not fully eliminate spurious oscillations near sharp layers, thereby prompting several authors to propose additional shock-capturing terms [3, 7, 11, 15, 16, 18, 19].

The goal of this paper is to derive a shock-capturing scheme for which a DMP can be rigorously established in the context of convection-diffusion-reaction equations. This is an important design criterion for a scheme that can tackle computationally-demanding problems such as chemically reactive flows. If the shock-capturing operator is linear, the only way to ensure a DMP is generally to employ first-order artificial viscosity, leading to excessive smearing of the solution. Therefore, this work focuses on a nonlinear shock-capturing scheme; this may seem a rather cumbersome approach for linear model problems, but the computational overheads resulting from nonlinearity are expected to influence marginally the overall cost of a flame simulation. To date, only two shock-capturing schemes satisfy rigorously a DMP, namely the first-order artificial viscosity of [7] and the nonlinear artificial viscosity of [3]. The main idea of this work is to express the nonlinear shock-capturing operator in terms of the jumps across element faces of the gradient of the finite element solution. Another important issue is to analyze the convergence of the shock-capturing operator when the cell Péclet numbers are small enough. This is important in practice when working with locally refined meshes; for instance, flame front resolution requires that the cells locally capture the reaction-diffusion layer at the flame front. To this purpose, we prove that the finite element solution with shock-capturing converges to that without shock-capturing if the mesh size is small enough. Moreover, in the diffusion-dominated regime, we establish that the difference between the two finite element solutions super-converges with respect to the actual approximation error.

The paper is organized as follows. The discrete setting is presented in Section 2. Section 3 analyzes the shock-capturing scheme; it addresses the DMP, the existence of discrete solutions, and the convergence properties of the scheme. Section 4 illustrates the theoretical results with numerical experiments. Conclusions are drawn in Section 5.

2. The discrete setting

2.1. **Notation.** Let $\{\mathcal{T}\}$ be a family of simplicial meshes of Ω. For simplicity, assume that all the meshes cover Ω exactly, i.e., Ω is a polygon or a polyhedron.

For $T \in \mathcal{T}$, let μ_T denote its d-dimensional measure, let h_T denote its diameter, and set $h = \max_{T\in\mathcal{T}} h_T$. Without loss of generality, assume that $h \leq 1$. Let $\mathcal{F}$ denote the set of interior faces in the mesh, i.e., the set of $(d-1)$-manifolds that are not included in the boundary $\partial\Omega$. For $F \in \mathcal{F}$, let h_F denote its diameter and let μ_F denote its $(d-1)$-dimensional measure. Moreover, let $\mathcal{T}(F)$ denote the set of elements sharing F as a face and let $\mathcal{E}(F)$ denote the set of edges, i.e., 1-dimensional manifolds, in F. Let $\mathcal{E}$ denote the set of edges in the mesh. For $E \in \mathcal{E}$, let h_E denote its diameter, let t_E denote a tangential unit vector to E (its orientation is irrelevant), and let $\mathcal{T}(E)$ denote the set of elements to which E belongs. For a piecewise-smooth function v and for an interior face $F \in \mathcal{F}$, $F = T_1 \cap T_2$ where T_1 and T_2 are two distinct elements of $\mathcal{T}$ with respective outer normals n_1 and n_2, introduce the (scalar-valued) jump $[\nabla v]_F = \nabla(v|_{T_1})\cdot n_1 + \nabla(v|_{T_2})\cdot n_2$. For $F \in \mathcal{F}$, set $\Delta_F = \bigcup_{T\in\mathcal{T};T\cap F\neq\emptyset} T$.

Let S_i be an interior vertex of $\mathcal{T}$. Let ω_i denote the associated P_1-nodal Lagrange basis function, i.e., the unique continuous, piecewise-affine function that takes the value 1 at S_i and vanishes at all the other vertices. Let Ω_i denote the support of ω_i. Let $\mathcal{F}(S_i)$ denote the set consisting of the faces in $\mathcal{F}$ to which S_i belongs.

For a region R consisting of a collection of mesh elements, $|\beta|_{\infty,R}$ denotes the norm $\|\beta\|_{[L^\infty(R)]^d}$ and $|\sigma|_{\infty,R}$ denotes the norm $\|\sigma\|_{L^\infty(R)}$. Finally, $\|\cdot\|$ denotes the $L^2(\Omega)$-norm.

2.2. **Basic assumptions on the mesh.** Henceforth, the following assumptions are made on the mesh family $\{\mathcal{T}\}$.

Hypothesis 2.1 (Local quasi-uniformity)**.** There exists a constant ρ such that for all $\mathcal{T}$ in $\{\mathcal{T}\}$ and for all vertices S_i in $\mathcal{T}$,

$$\max_{E\subset\Omega_i} h_E \leq \rho \min_{E\subset\Omega_i} h_E, \tag{2.1}$$

where $E \subset \Omega_i$ stands for all edges E in Ω_i.

Hypothesis 2.2 (Xu–Zikatanov)**.** For all $E \in \mathcal{E}$, the following inequality holds:

$$\frac{1}{d(d-1)} \sum_{T\in\mathcal{T}(E)} |\kappa_{E,T}| \cot\theta_{E,T} \geq 0, \tag{2.2}$$

where $|\kappa_{E,T}|$ is the $(d-2)$-dimensional measure of the simplex $F_{i,T} \cap F_{j,T}$ opposite to the edge E in T, $F_{i,T}$ (resp. $F_{j,T}$) is the face of T opposite to the vertex S_i (resp. S_j), S_i and S_j are the vertices connected by E, and $\theta_{E,T}$ is the angle between the faces $F_{i,T}$ and $F_{j,T}$.

Hypothesis 2.1 amounts to a local quasi-uniformity property of the mesh. It implies that there is a finite number, n_ρ, of elements in each macro-element Ω_i. Hypothesis 2.2 has been introduced by Xu and Zikatanov; it implies that the stiffness matrix associated with the Poisson equation discretized using piecewise linears on $\mathcal{T}$ is an M-matrix; see [20, Lemma 2.1]. In two dimensions, Hypothesis 2.2 means that the sum of the two angles facing any edge in the mesh is less than π, and this condition then implies that the triangulation $\mathcal{T}$ is a so-called Delaunay triangulation. A sufficient condition for Hypothesis 2.2 to hold is the so-called weakly acute

condition, meaning that all the angles in each triangle are less than or equal to $\frac{\pi}{2}$. This is a more stringent condition than Hypothesis 2.2; it is also more difficult to fulfill in practice.

2.3. **The strong and weak DMP properties.** Consider H^1-conforming, piecewise-affine finite elements. Set

$$V_h^g = \{\, v \in C^0(\overline{\Omega});\ \ \forall T \in \mathcal{T},\, v|_T \in P_1(T);\ \ v = \mathcal{P}g \text{ on } \partial\Omega \,\},$$

where $P_1(T)$ denotes the space spanned by linear polynomials on T and $\mathcal{P}$ denotes the L^2-projection onto the space of piecewise-affine functions on the boundary. Consider the abstract problem of finding $U \in V_h^g$ such that

$$\tilde{a}(U; v) = (f, v), \quad \forall v \in V_h^0. \tag{2.3}$$

Here, $\tilde{a}$ is a semilinear form (to be specified below) and $(\cdot,\cdot)$ denotes the $L^2(\Omega)$–scalar product. Semicolons are used for forms that are nonlinear with respect to their first argument. Owing to the nonlinearity of $\tilde{a}(\cdot;\cdot)$, a DMP for (2.3) cannot be proved by showing that the stiffness matrix is an M-matrix. A convenient framework to establish the DMP is briefly presented.

Definition 2.3 (Strong DMP property)**.** The semilinear form $\tilde{a}(\cdot;\cdot)$ is said to satisfy the *strong DMP property* if the following holds: For all $U \in V_h^g$ and for all interior vertices S_i, if U is locally minimal (resp. maximal) on the vertex S_i over the macro-element Ω_i, then there exist positive quantities $(\alpha_F)_{F\in\mathcal{F}(S_i)}$ such that

$$\tilde{a}(U; \omega_i) \le -\sum_{F\in\mathcal{F}(S_i)} \alpha_F\, |[\nabla U]_F|, \tag{2.4}$$

(resp., $\tilde{a}(U;\omega_i) \ge \sum_{F\in\mathcal{F}(S_i)} \alpha_F\, |[\nabla U]_F|$).

Definition 2.4 (Weak DMP property)**.** The semilinear form $\tilde{a}(\cdot;\cdot)$ is said to satisfy the *weak DMP property* if it satisfies the strong DMP property under the additional assumption that the local minimum is negative (resp., the maximum is positive).

Following the ideas of [3], the DMP property can be used to prove that the finite element solution U of (2.3) satisfies a DMP. For a vector U with nonnegative components, we write $U \ge 0$.

Proposition 2.5. *Assume that the semilinear form $\tilde{a}(\cdot;\cdot)$ satisfies the strong DMP property. Assume that $U \in V_h^g$ solves* (2.3) *and that $f \ge 0$. Then U reaches its minimum on the boundary $\partial\Omega$.*

Proof. Assume that U reaches its minimum at an interior vertex S_i. Since $\tilde{a}(\cdot;\cdot)$ satisfies the strong DMP property and $f \ge 0$, (2.4) implies that ∇U is constant over Ω_i. Therefore, the minimum is reached on a further vertex, whence it is eventually deduced that the minimum is reached on the boundary. □

Proposition 2.6. *Assume that the semilinear form $\tilde{a}(\cdot;\cdot)$ satisfies the weak DMP property. Assume that $U \in V_h^g$ solves* (2.3) *and that $f \ge 0$ and $g \ge 0$. Then $U \ge 0$.*

Proof. Similar to the previous one. □

2.4. **The shock-capturing scheme.** Consider the standard Galerkin formulation of (1.1)–(1.2) supplemented with one or several stabilizing terms, usually a least-squares term on the residual yielding L^2-norm control of the streamline derivative, and a shock-capturing term to quench the remaining spurious oscillations. The resulting discrete problem consists of finding $U \in V_h^g$ such that

$$a(U,v) + s(U,v) + j(U;v) = (f,v), \quad \forall v \in V_h^0, \tag{2.5}$$

where $a(U,v)$ denotes the standard Galerkin contribution

$$a(U,v) = (\varepsilon \nabla U, \nabla v) + (\beta \cdot \nabla U, v) + (\sigma U, v),$$

and $s(U,v)$ and $j(U;v)$ denote the least-squares and the shock-capturing terms, respectively.

We investigate two possible choices for the least-squares term $s(U,v)$: the SD method [2, 14]

$$s(U,v) = (\beta \cdot \nabla U + \sigma U - f, \gamma_{\mathrm{sd}} \beta \cdot \nabla v) \tag{2.6}$$

and a recent CIP method derived in [5],

$$s(U,v) = \sum_{F \in \mathcal{F}} \gamma_{\mathrm{cip}} h_F^2 \mu_F [\nabla U]_F [\nabla v]_F. \tag{2.7}$$

The SD parameter γ_{sd} is set locally to $\gamma_{\mathrm{sd}}|_T = c_{\mathrm{sd}} h_T |\beta|_{\infty,T}^{-1}$ for all $T \in \mathcal{T}$ where c_{sd} is a user-specific constant, while the CIP parameter γ_{cip} depends only on ρ and $|\beta|_{\infty,\Omega}$. Although both (2.6) and (2.7) are well suited to simulations of convection-dominated flows, neither enjoys the DMP property so that a further stabilization by a shock-capturing term is needed.

The shock-capturing term $j(U;v)$ analyzed in this work takes the form

$$j(U;v) = c_\rho \sum_{F \in \mathcal{F}} \delta_F(U) \psi_F(U;v), \tag{2.8}$$

where c_ρ is a constant depending only on ρ, δ_F is a function of U evaluated on each face as

$$\forall F \in \mathcal{F}, \quad \delta_F(U) = (|\beta|_{\infty,\Delta_F} + \rho |\sigma|_{\infty,\Delta_F} h_F) h_F \mu_F |[\nabla U]_F|, \tag{2.9}$$

and

$$\psi_F(U;v) = \sum_{E \in \mathcal{E}(F)} h_E \mathrm{sign}(\nabla U \cdot t_E) \nabla v \cdot t_E. \tag{2.10}$$

In two dimensions, $\psi_F(U;v)$ consists of only one term since faces and edges coincide. Note that $\nabla v \cdot t_E$ is always single-valued since v is H^1-conforming.

The main motivation for introducing the jumps of the gradient in the shock-capturing operator $j(U;v)$ stems from the following result, which will be used repeatedly in the sequel.

Lemma 2.7. *If $U \in V_h^g$ has a local minimum in the vertex S_i, then*

$$\forall T \subset \Omega_i, \quad |(\nabla U|_T)| \le \sum_{F \in \mathcal{F}(S_i)} |[\nabla U]_F|. \tag{2.11}$$

Proof. Since U has a local minimum at the vertex S_i, it is clear that for all $T \subset \Omega_i$, there exists $\hat{T} \subset \Omega_i$ such that $\nabla U|_T \cdot \nabla U|_{\hat{T}} \leq 0$. Hence, $|(\nabla U|_T) - (\nabla U|_{\hat{T}})| \geq |(\nabla U|_T)|$. Let $\mathcal{F}(T,\hat{T}) \subset \mathcal{F}(S_i)$ be the set of faces crossed by a path connecting T to $\hat{T}$ in Ω_i. Then $\nabla U|_T = \nabla U|_{\hat{T}} + \sum_{F\in\mathcal{F}(T,\hat{T})} \varsigma_F [\nabla U]_F$ with $\varsigma_F = \pm 1$, yielding

$$|(\nabla U|_T)| \leq |(\nabla U|_T) - (\nabla U|_{\hat{T}})| \leq \sum_{F\in\mathcal{F}(T,\hat{T})} |[\nabla U]_F| \leq \sum_{F\in\mathcal{F}(S_i)} |[\nabla U]_F|.$$

The proof is complete. □

3. Analysis of the shock-capturing scheme

The analysis of the shock-capturing scheme addresses three aspects: the DMP, the existence of discrete solutions, and the (super-)convergence of the shock-capturing solution to the finite element solution with no shock-capturing in the diffusion-dominated regime.

3.1. **DMP.**

We first prove a DMP for the standard Galerkin approximation stabilized only with the shock-capturing term ($s = 0$) and then establish a DMP when the least-squares term s results from (2.6) or (2.7).

Theorem 3.1 (Standard Galerkin)**.** *Let $j(\cdot;\cdot)$ be defined in (2.8)–(2.10). Then, provided c_ρ is large enough, the semilinear form $a(\cdot,\cdot) + j(\cdot;\cdot)$ satisfies the weak DMP property.*

Proof. Assume that U has a local minimum with $U \leq 0$ at some interior vertex S_i. Owing to Hypothesis 2.2,

$$(\nabla U, \nabla \omega_i) \leq 0.$$

Consider now the reaction term and let $T \subset \Omega_i$. If U changes sign in T, then $\|U\|_{L^\infty(T)} \leq h_T \|\nabla U\|_{L^\infty(T)}$ and hence,

$$(\sigma U, \omega_i)_T \leq |\sigma|_{\infty,T} \|U\|_{L^\infty(T)} \|\omega_i\|_{L^1(T)} \leq \tfrac{1}{d+1} |\sigma|_{\infty,T} h_T \|\nabla U\|_{L^1(T)},$$

since ∇U is constant on T. If U is negative on T, the inequality $(\sigma U, \omega_i)_T \leq \frac{1}{d+1} |\sigma|_{\infty,T} h_T \|\nabla U\|_{L^1(T)}$ trivially holds since the left-hand side is negative. Therefore,

$$\begin{aligned}(\beta \cdot \nabla U + \sigma U, \omega_i) &\leq \tfrac{1}{d+1} \sum_{T\subset\Omega_i} (|\beta|_{\infty,T} + |\sigma|_{\infty,T} h_T) \|\nabla U\|_{L^1(T)} \\ &\leq \tfrac{n_\rho}{d+1} (|\beta|_{\infty,\Omega_i} + |\sigma|_{\infty,\Omega_i} h_i) \max_{T\subset\Omega_i} \|\nabla U\|_{L^1(T)},\end{aligned}$$

where $h_i = \max_{T\subset\Omega_i} h_T$. Lemma 2.7 yields

$$\begin{aligned}(\beta \cdot \nabla U + \sigma U, \omega_i) &\leq \tfrac{\rho^{d-1} n_\rho}{d+1} (|\beta|_{\infty,\Omega_i} + |\sigma|_{\infty,\Omega_i} h_i) h_i \sum_{F\in\mathcal{F}(S_i)} \mu_F |[\nabla U]_F| \\ &\leq \tfrac{\rho^d n_\rho}{d+1} \sum_{F\in\mathcal{F}(S_i)} (|\beta|_{\infty,\Delta_F} + \rho |\sigma|_{\infty,\Delta_F} h_F) h_F \mu_F |[\nabla U]_F|,\end{aligned}$$

since for all $F \in \mathcal{F}(S_i)$, $\mu_T \leq \rho^{d-1} h_T \mu_F \leq \rho^{d-1} h_i \mu_F$ and $h_i \leq \rho h_F$ owing to Hypothesis 2.1. Furthermore, since U is locally minimal at S_i over Ω_i and since

$\mathrm{card}(\mathcal{E}(F)) = \frac{1}{2}d(d-1)$, it is clear that for all $F \in \mathcal{F}(S_i)$, $\psi_F(U;\omega_i) \le -\frac{1}{2}d(d-1)$. As a result, taking $c_\rho > \frac{2\rho^2 n_\rho}{(d+1)d(d-1)}$ yields

$$(\varepsilon\nabla U, \nabla\omega_i) + (\beta\cdot\nabla U + \sigma U, \omega_i) + j(U;\omega_i) \le -\sum_{F\in\mathcal{F}(S_i)} \alpha_F\,|[\nabla U]_F|,$$

where all the quantities α_F are positive. Hence, $a(\cdot,\cdot) + j(\cdot;\cdot)$ satisfies the weak DMP property. □

Remark 3.2. If nodal quadrature is used, the standard Galerkin formulation does not need any stabilization of the source term to fulfill the weak DMP property. From an algebraic viewpoint, this is reflected by the fact that the lumped mass matrix is an M-matrix.

Theorem 3.3 (SD)**.** *Let $j(\cdot;\cdot)$ be defined in* (2.8)–(2.10) *and let $s(\cdot,\cdot)$ be defined in* (2.6)*. Assume that $f = 0$ and $\sigma = 0$. Then, provided c_ρ is large enough, the semilinear form $a(\cdot,\cdot) + s(\cdot,\cdot) + j(\cdot;\cdot)$ satisfies the weak DMP property.*

Proof. Owing to Theorem 3.1, it suffices to prove that $s(\cdot,\cdot)+j(\cdot;\cdot)$ satisfies the weak DMP property. Using (2.6) with $f = \sigma = 0$, the fact that $\gamma_{\mathrm{sd}}|_T = c_{\mathrm{sd}}h_T|\beta|_{\infty,T}^{-1}$, and Hypothesis 2.1 yields

$$\begin{aligned}(\beta\cdot\nabla U, \gamma_{\mathrm{sd}}\beta\cdot\nabla\omega_i) &\le \sum_{T\subset\Omega_i} |\beta|_{\infty,T}\|\nabla U\|_{L^1(T)}(\gamma_{\mathrm{sd}}|_T)\|\beta\cdot\nabla\omega_i\|_{L^\infty(T)}\\ &\le \sum_{T\subset\Omega_i} c_{\mathrm{sd}}|\beta|_{\infty,T}\|\nabla U\|_{L^1(T)}h_T\|\nabla\omega_i\|_{L^\infty(T)}\\ &\le \sum_{T\subset\Omega_i} c_{\mathrm{sd}}\rho|\beta|_{\infty,T}\|\nabla U\|_{L^1(T)}.\end{aligned}$$

Proceed as in the proof of Theorem 3.1 to conclude. □

Remark 3.4. Another approach to designing a shock-capturing operator for the standard Galerkin method together with the SD stabilization is to use residual-based artificial diffusion; see the design in [16] for well-posedness and convergence and the design in [3] for DMP.

We now turn to a more recent CIP technique to stabilize convection-diffusion-reaction problems by using the jumps of the gradient across element interfaces. We prove that this stabilized method can also be used in conjunction with a shock-capturing operator to satisfy a DMP. The advantage compared to SD is that the DMP is proved also for the nonhomogeneous problem and also for nonzero reaction. Furthermore, the mass matrix can be lumped and, in this case, the source terms or any terms arising from time discretization do not need additional stabilization.

Theorem 3.5 (CIP)**.** *Let $j(\cdot;\cdot)$ be defined in* (2.8) *with $\psi_F(U;v)$ defined in* (2.10) *and for all $F \in \mathcal{F}$,*

$$\delta_F(U) = (|\beta|_{\infty,\Delta_F} + \rho|\sigma|_{\infty,\Delta_F}h_F)h_F\mu_F|[\nabla U]_F| + \gamma_{\mathrm{cip}}h_F\mu_F m_F(U), \tag{3.1}$$

where

$$m_F(U) = \max_{\substack{F'\in\mathcal{F}\\ F'\subset\partial T';T'\in\mathcal{T}(F)}} |[\nabla U]_{F'}|. \tag{3.2}$$

Let $s(\cdot,\cdot)$ be defined in (2.7)*. Then, provided c_ρ is large enough, the semilinear form $a(\cdot,\cdot) + s(\cdot,\cdot) + j(\cdot;\cdot)$ satisfies the weak DMP property.*

Proof. (1) Let S_i be an interior vertex of the mesh. Owing to Hypothesis 2.1,

$$\begin{aligned} s(U,\omega_i) &= \sum_{F\in\mathcal{F}'(S_i)} \gamma_{\rm cip} h_F^2 \mu_F [\nabla U]_F [\nabla \omega_i]_F \\ &\le \sum_{F\in\mathcal{F}'(S_i)} 2\rho\gamma_{\rm cip} h_F \mu_F |[\nabla U]_F|, \end{aligned}$$

where $\mathcal{F}'(S_i)$ denotes the set of faces F in $\mathcal{F}$ such that $F\cap\Omega_i\neq\emptyset$. Hence,

$$s(U,\omega_i) \le \sum_{F\in\mathcal{F}(S_i)} 2\rho\gamma_{\rm cip} h_F \mu_F m_F(U) + \sum_{F'\in\mathcal{F}'(S_i)\setminus\mathcal{F}(S_i)} 2\rho\gamma_{\rm cip} h_{F'} \mu_{F'} |[\nabla U]_{F'}|.$$

For all $F'\in\mathcal{F}'(S_i)\setminus\mathcal{F}(S_i)$, there is $F\in\mathcal{F}(S_i)$, belonging to the same element as F', such that $|[\nabla U]_{F'}|\le m_F(U)$; notice that a given $F\in\mathcal{F}(S_i)$ arises at most twice when F' sweeps $\mathcal{F}'(S_i)\setminus\mathcal{F}(S_i)$. Moreover, owing to Hypothesis 2.1, $h_{F'}\mu_{F'}\le\rho^2 h_F\mu_F$ since F' and F belong to the same element. Hence,

$$s(U,\omega_i) \le \sum_{F\in\mathcal{F}(S_i)} 2\rho(1+2\rho^2)\gamma_{\rm cip} h_F \mu_F m_F(U).$$

As a result, taking $c_\rho > \frac{4\rho(1+2\rho^2)}{d(d-1)}$ yields

$$s(U,\omega_i) + c_\rho \sum_{F\in\mathcal{F}(S_i)} \gamma_{\rm cip} h_F \mu_F m_F(U)\psi_F(U;v) \le 0.$$

(2) Assume that U has a minimal value at the vertex S_i with $U(S_i)\le 0$. Take

$$c_\rho > \max\left(\frac{2\rho^2 n_\rho}{(d+1)d(d-1)}, \frac{4\rho(1+2\rho^2)}{d(d-1)}\right).$$

Then it is readily inferred from the proof of Theorem 3.1 and the first step of this proof that

$$a(U,\omega_i) + s(U,\omega_i) + j(U;\omega_i) \le -\sum_{F\in\mathcal{F}(S_i)} \alpha_F\, |[\nabla U]_F|,$$

where all the quantities α_F are positive. The proof is complete. □

Corollary 3.6. *If $\sigma = 0$, the semilinear form $a(\cdot,\cdot)+s(\cdot,\cdot)+j(\cdot;\cdot)$ satisfies the strong DMP property in Theorems* 3.1*,* 3.3*, and* 3.5*.*

Proof. Follows immediately by setting $\sigma = 0$ in the above proofs. □

Remark 3.7. Hypothesis 2.2 is made only to ensure that the discrete Laplacian satisfies a DMP, since the shock-capturing term controls the convection-reaction terms on any type of mesh that satisfies Hypothesis 2.1 alone. For a shock-capturing operator yielding a Laplacian with DMP on any type of mesh satisfying Hypothesis 2.1 alone, see [4].

3.2. **Existence.** The existence of solutions to (2.5) is a nontrivial problem since the shock-capturing term is nonlinear. The analysis is presented for homogeneous Dirichlet boundary data.

Theorem 3.8. *Assume that*

$$\forall U\in V_h^0,\quad a(U,U)+s(U,U) \ge c_\varepsilon \|U\|_{H^1(\Omega)}^2, \tag{3.3}$$

for some positive constant c_ε. Then the nonlinear scheme (2.5) *admits at least one solution.*

Proof. Let $\epsilon > 0$. For a face $F \in \mathcal{F}$ and an edge $E \subset \partial F$, define the functional

$$\phi_{F,E} : V_h^0 \ni U \longmapsto h_E \frac{\delta_F(U)}{|\nabla U \cdot t_E| + \epsilon} \in \mathbb{R}.$$

(1) For all $\tilde{U} \in V_h^0$, the problem of finding $U \in V_h^0$ such that, for all $v \in V_h^0$,

$$a(U,v) + s(U,v) + c_\rho \sum_{F\in\mathcal{F}} \sum_{E\in\mathcal{E}(F)} \phi_{F,E}(\tilde{U})(\nabla U \cdot t_E, \nabla v \cdot t_E)_E = (f,v), \tag{3.4}$$

where $(\cdot,\cdot)_E$ denotes the $L^2(E)$-scalar product, has a unique solution owing to the Lax–Milgram Lemma. Hence, the operator $T_\epsilon : V_h^0 \ni \tilde{U} \longmapsto U \in V_h^0$ is well defined.

(2) Taking $v = U$ in (3.4) and using (3.3) yield

$$c_\varepsilon \|U\|_{H^1(\Omega)}^2 + c_\rho \sum_{F\in\mathcal{F}} \sum_{E\in\mathcal{E}(F)} \phi_{F,E}(\tilde{U}) \|\nabla U \cdot t_E\|_{L^2(E)}^2 \le c_\Omega \|f\| \|U\|_{H^1(\Omega)},$$

where c_Ω is the constant associated with the Poincaré inequality, namely for all $u \in H_0^1(\Omega)$, $\|u\| \le c_\Omega \|u\|_{H^1(\Omega)}$. Since the second term in the left-hand side is nonnegative, the above inequality readily implies the a priori estimate $\|U\|_{H^1(\Omega)} \le \frac{c_\Omega}{c_\varepsilon}\|f\|$. Therefore, if $\tilde{U}$ is such that $\|\tilde{U}\|_{H^1(\Omega)} \le \frac{c_\Omega}{c_\varepsilon}\|f\|$, $U = T_\epsilon(\tilde{U})$ will also be in this ball.

(3) Let $\tilde{U}_1$ and $\tilde{U}_2$ be in V_h^0. Set $U_1 = T_\epsilon(\tilde{U}_1)$ and $U_2 = T_\epsilon(\tilde{U}_2)$. Subtracting (3.4) for U_2 from (3.4) for U_1 and testing with $v = U_1 - U_2$ yield

$$\begin{aligned}
&a(U_1 - U_2, U_1 - U_2) + s(U_1 - U_2, U_1 - U_2) \\
&+ c_\rho \sum_{F\in\mathcal{F}} \sum_{E\in\mathcal{E}(F)} \phi_{F,E}(\tilde{U}_1) \|\nabla(U_1 - U_2) \cdot t_E\|_{L^2(E)}^2 \\
&= c_\rho \sum_{F\in\mathcal{F}} \sum_{E\in\mathcal{E}(F)} (\phi_{F,E}(\tilde{U}_2) - \phi_{F,E}(\tilde{U}_1))(\nabla U_2 \cdot t_E, \nabla(U_1 - U_2) \cdot t_E)_E.
\end{aligned}$$

The left-hand side of this equation is lower-bounded by $c_\varepsilon \|U_1 - U_2\|_{H^1(\Omega)}^2$. Furthermore, the right-hand side of the equation can be upper-bounded in the form $c_{\beta,\sigma,\mathcal{T}} \|\nabla(\tilde{U}_1 - \tilde{U}_2)\| \|\nabla U_2\| \|\nabla(U_1 - U_2)\|$ where the constant $c_{\beta,\sigma,\mathcal{T}}$ depends on β, σ, and $\mathcal{T}$. This yields

$$\|U_1 - U_2\|_{H^1(\Omega)} \le \left(c_{\beta,\sigma,\mathcal{T}} \frac{c_\Omega}{c_\varepsilon} \|f\| \right) \|\tilde{U}_1 - \tilde{U}_2\|_{H^1(\Omega)}.$$

Therefore, the operator T_ϵ is continuous.

(4) Owing to Brouwer's Theorem, steps (2) and (3) imply that the operator T_ϵ admits a fixed point, say U_ϵ, in the ball of radius $\frac{c_\Omega}{c_\varepsilon}\|f\|$ in V_h^0. Since the sequence $(U_\epsilon)_\epsilon$ is in a finite-dimensional ball, there is a subsequence, still denoted by $(U_\epsilon)_\epsilon$, such that $U_\epsilon \to U$ in $H_0^1(\Omega)$ as $\epsilon \to 0$. Passing to the limit $\epsilon \to 0$ in (3.4), it is inferred that U solves (2.5). □

Remark 3.9. Assumption (3.3) holds for the standard Galerkin method alone and for the standard Galerkin method with SD (and $f = 0$) or with CIP.

3.3. **Convergence.** The goal of this section is to prove that the shock-capturing solution endowed with a DMP converges to the finite element solution without shock-capturing in the diffusion-dominated regime. This result is of practical importance in the context of locally refined meshes since the cell Péclet numbers can undergo significant variations in the domain. In Theorem 3.10 below, we address this issue; in particular, we prove a stronger result, namely that the difference between the

shock-capturing solution and that without shock-capturing super-converges with respect to the actual approximation error.

Consider the nonlinear scheme (2.5) with CIP operator $s(\cdot,\cdot)$ defined in (2.7) and shock-capturing operator $j(\cdot;\cdot)$ defined in (2.8) with coefficient $\delta_F(U)$ defined in (3.1). Owing to Theorem 3.5, this scheme satisfies a DMP property. Assuming for simplicity that $g = 0$, the finite element method with CIP, but without shock-capturing, consists of finding $\tilde{U} \in V_h^0$ such that

$$a(\tilde{U}, v) + s(\tilde{U}, v) = (f, v), \quad \forall v \in V_h^0. \tag{3.5}$$

The discrete problem (3.5) has been analyzed in [5] where the following stability and a priori estimates were proved: there exists a stability constant $c_s > 0$, independent of h and ε, such that for all $w \in V_h^0$,

$$c_s|||w||| \le \sup_{v \in V_h^0} \frac{a(w,v) + s(w,v)}{|||v|||}, \tag{3.6}$$

with the triple norm

$$|||w|||^2 = \|\varepsilon^{\frac{1}{2}} \nabla w\|^2 + \|\sigma_0^{\frac{1}{2}} w\|^2 + \|h^{\frac{1}{2}} \beta \cdot \nabla w\|^2 + s(w,w). \tag{3.7}$$

Moreover, assuming that the exact solution u of (1.1)–(1.2) is in $H^2(\Omega)$ and denoting by $\|u\|_{H^2(\Omega)}$ its corresponding norm, there exists a constant c_1, independent of h and ε, such that the following a priori estimate holds:

$$|||u - \tilde{U}||| \le c_1(\varepsilon^{\frac{1}{2}} h + |\beta|_{\infty,\Omega}^{\frac{1}{2}} h^{\frac{3}{2}} + |\sigma|_{\infty,\Omega}^{\frac{1}{2}} h^2)\|u\|_{H^2(\Omega)}. \tag{3.8}$$

Theorem 3.10. *Let $\tilde{U}$ be the solution of (3.5) with $s(\cdot,\cdot)$ defined in (2.7) and let U be the solution of (2.5) with shock-capturing term $j(\cdot;\cdot)$ defined in (2.8) and coefficient $\delta_F(U)$ defined in (3.1). Assume $g = 0$ and that the exact solution u of (1.1)–(1.2) is in $H^2(\Omega)$. Then there exists a constant c_2, independent of h and ε, such that, if $c_2 h < \varepsilon$,*

$$|||\tilde{U} - U||| \le c_1(\varepsilon^{\frac{1}{2}} h + |\beta|_{\infty,\Omega}^{\frac{1}{2}} h^{\frac{3}{2}} + |\sigma|_{\infty,\Omega}^{\frac{1}{2}} h^2)\|u\|_{H^2(\Omega)}. \tag{3.9}$$

Moreover, in the asymptotic case $c_2 h \ll \varepsilon$,

$$|||\tilde{U} - U||| \le c_3 h^{\frac{3}{2}} \|u\|_{H^2(\Omega)}, \tag{3.10}$$

with $c_3 = c_1(\frac{c_2}{2})^{\frac{1}{2}}$.

Proof. (1) Let v and w be in V_h^0. Owing to the definition of $j(\cdot;\cdot)$,

$$j(w;v) \le c_\rho \tfrac{d(d-1)}{2} \sum_{F \in \mathcal{F}} \delta_F(w) h_F \|\nabla v\|_{L^\infty(T(F))},$$

where $T(F)$ is any element of $\mathcal{T}$ containing F. Since

$$h_F \mu_F \|\nabla v\|_{L^\infty(T(F))}^2 \le \rho \mu_{T(F)} \|\nabla v\|_{L^\infty(T(F))}^2 = \rho \|\nabla v\|_{L^2(T(F))}^2,$$

owing to Hypothesis 2.1, it is clear that $\sum_{F \in \mathcal{F}} h_F \mu_F \|\nabla v\|_{L^\infty(F)}^2 \le (d+1)\rho \|\nabla v\|^2$. As a result, using the Cauchy–Schwarz inequality yields

$$j(w;v) \le c_\rho \tfrac{d(d-1)}{2} h^{\frac{1}{2}} \left(\sum_{F \in \mathcal{F}} \mu_F^{-1} \delta_F(w)^2 \right)^{\frac{1}{2}} ((d+1)\rho)^{\frac{1}{2}} \|\nabla v\|.$$

Furthermore, since $h \leq 1$, it is inferred that

$$\sum_{F\in\mathcal{F}} \mu_F^{-1}\delta_F(w)^2 \leq 2\sum_{F\in\mathcal{F}} (|\beta|_{\infty,\Omega} + \rho|\sigma|_{\infty,\Omega})^2 h_F^2 \mu_F [\nabla w]_F^2 + 2\sum_{F\in\mathcal{F}} \gamma_{\text{cip}}^2 h_F^2 \mu_F m_F(w)^2.$$

For $F \in \mathcal{F}$, there exists $F' \in \mathcal{F}$, belonging to the same element as F, such that $m_F(w) = |[\nabla w]_{F'}|$. Since F and F' belong to the same element, $h_F^2\mu_F \leq \rho^2 h_{F'}\mu_{F'}$. Since a given face $F' \in \mathcal{F}$ can be attained at most twice when F sweeps $\mathcal{F}$, it is inferred that

$$\sum_{F\in\mathcal{F}} \mu_F^{-1}\delta_F(w)^2 \leq c_4 \sum_{F\in\mathcal{F}} \gamma_{\text{cip}} h_F^2 \mu_F [\nabla w]_F^2 = c_4 s(w,w),$$

with

$$c_4 = \frac{2}{\gamma_{\text{cip}}}((|\beta|_{\infty,\Omega} + \rho|\sigma|_{\infty,\Omega})^2 + 2\gamma_{\text{cip}}^2\rho^3).$$

Setting $c_5 = c_\rho \frac{d(d-1)}{2}((d+1)\rho c_4)^{\frac{1}{2}}$ then yields the following boundedness result:

$$j(w;v) \leq c_5 h^{\frac{1}{2}} s(w,w)^{\frac{1}{2}} \|\nabla v\|. \tag{3.11}$$

(2) Owing to (3.6),

$$c_s |||\tilde{U} - U||| \leq \sup_{v\in V_h^0} \frac{a(\tilde{U}-U,v) + s(\tilde{U}-U,v)}{|||v|||} = \sup_{v\in V_h^0} \frac{j(U;v)}{|||v|||}.$$

Using (3.11) yields

$$|||\tilde{U} - U||| \leq \left(\frac{\xi}{2}\right)^{\frac{1}{2}} h^{\frac{1}{2}} s(U,U)^{\frac{1}{2}} \frac{\|\nabla v\|}{|||v|||} \leq \left(\frac{\xi}{2}\right)^{\frac{1}{2}} \left(\frac{h}{\varepsilon}\right)^{\frac{1}{2}} s(U,U)^{\frac{1}{2}}, \tag{3.12}$$

with $\xi = 2c_5^2 c_s^{-2}$.

(3) Use the inequality

$$s(U,U) - 2s(\tilde{U},\tilde{U}) \leq 2s(U-\tilde{U}, U-\tilde{U}) \leq 2|||\tilde{U} - U|||^2$$

together with (3.12) to infer

$$s(U,U) \leq \left(\frac{\xi h}{\varepsilon}\right) s(U,U) + 2s(\tilde{U},\tilde{U}).$$

Since for $u \in H^2(\Omega)$, $s(\tilde{U},\tilde{U}) = s(\tilde{U}-u, \tilde{U}-u) \leq |||\tilde{U}-u|||^2$, this implies

$$s(U,U) \leq \left(\frac{\xi h}{\varepsilon}\right) s(U,U) + 2|||\tilde{U} - u|||^2.$$

The above estimate, together with (3.12), yields

$$|||\tilde{U} - U||| \leq \left(\frac{\xi h}{\varepsilon - \xi h}\right)^{\frac{1}{2}} |||\tilde{U} - u|||. \tag{3.13}$$

Set $c_2 = 2\xi$. Then, if $c_2 h < \varepsilon$, (3.13), together with the a priori estimate (3.8), leads to (3.9).

(4) In the asymptotic case $c_2 h \ll \varepsilon$, (3.8) and (3.13) readily yield (3.10). □

Remark 3.11. An interesting consequence of (3.10) is the fact that when the diffusion is dominant, the shock-capturing finite element solution, U, super-converges to the solution without shock-capturing, $\tilde{U}$, a phenomenon which is also observed numerically.

Remark 3.12. Assume $\sigma_0 > 0$. Then taking $v = U$ in (2.5) yields the a priori estimate

$$\|\varepsilon^{\frac{1}{2}}\nabla U\|^2 + \|\sigma_0^{\frac{1}{2}}U\|^2 + s(U,U) + j(U;U) \leq \|f\|\|U\|,$$

and hence, $s(U,U) \leq \frac{\|f\|^2}{\sigma_0}$. As a result, for all h and ε,

$$|||\tilde{U} - U||| \leq c_6 \left(\frac{h}{\varepsilon}\right)^{\frac{1}{2}}, \tag{3.14}$$

with $c_6 = (\frac{\xi}{2})^{\frac{1}{2}}\|f\|\sigma_0^{-\frac{1}{2}}$. A similar estimate holds for homogeneous convection-diffusion equations with $\sigma = 0$ and Dirichlet boundary data g since $s(U,U) \leq \frac{1}{2}\int_{\partial\Omega}|\beta\cdot\mathbf{n}|g^2\mathrm{d}s$ in this case. Since the shock-capturing term is large enough to guarantee a DMP, the a priori error estimate (3.14) involves a constant that scales as $\varepsilon^{-\frac{1}{2}}$ in the convection-dominated regime.

4. Numerical results

This section presents some numerical illustrations of the theoretical results. The nonlinear system of discrete equations, $F(X) = 0$, is solved approximately using a damped Newton method. Given an initial guess X^0, a sequence of iterates X^n is generated according to

$$J(X^n)(X^{n+1} - X^n) = -\lambda^n F(X^n), \tag{4.1}$$

where $J(X^n)$ denotes the Jacobian matrix of the nonlinear residual F at X^n and λ^n denotes the damping parameter. The linear system (4.1) is solved approximately using a preconditioned Krylov iterative method (e.g., BiCGStab and ILU). Convergence of Newton's method is achieved when the normalized Euclidean norm of the update vector $X^{n+1} - X^n$ is less than a prescribed tolerance (e.g., 10^{-5}). As noted in [3], the convergence behavior of Newton's method is improved by regularizing the sign operator in the shock-capturing term and replacing it by $\mathrm{sign}_\epsilon(x) = \tanh(x/\epsilon)$. The results reported below are obtained with $\epsilon = 1$, a choice for which Newton's method remains well behaved and spurious oscillations are essentially eliminated.

4.1. **Smooth solutions.** Consider the model problem (1.1)–(1.2) with $\Omega =]0,1[\times]0,1[$, $\varepsilon = 10^{-5}$, $\beta = (1,0)^t$, and $\sigma = 1$. Choose f and g so that the exact solution is

$$u = \exp\left(-\frac{(x-0.5)^2}{a_w} - \frac{3(y-0.5)^2}{a_w}\right),$$

with parameter $a_w = 0.2$. This problem is approximated using the CIP method with $s(\cdot,\cdot)$ defined in (2.7), shock-capturing term $j(\cdot;\cdot)$ defined in (2.8), and parameter $\delta_F(U)$ defined in (3.1) with constant c_ρ set to 5. Meshes are constructed from uniform tensor meshes with squares cut along a randomly chosen diagonal. Let N denote the number of mesh cells on each side of Ω.

Table 1 reports the errors between the exact solution u, the discrete solution $\tilde{U}$ without shock-capturing, and the discrete solution U with shock-capturing in the L^2, H^1, and L^∞-norms. All the errors exhibit h^2-order convergence in the L^2-norm (although the theoretical bound is $h^{\frac{3}{2}}$) and h-order convergence in the H^1-norm. Note that the shock-capturing scheme exhibits optimal order convergence although the flow is in the convection-dominated regime on all meshes considered. This results from the fact that the solution of the present test problem is smooth.

TABLE 1. Smooth solution problem: convergence results

	$\|u-U\|$			$\|u-\tilde{U}\|$			$\|U-\tilde{U}\|$		
N	L^2	H^1	L^∞	L^2	H^1	L^∞	L^2	H^1	L^∞
20	2.8e-2	4.2e-1	1.1e-1	3.0e-3	2.0e-1	9.1e-3	2.6e-2	3.7e-1	1.0e-1
40	2.3e-3	1.2e-1	1.8e-2	6.7e-4	9.9e-2	3.4e-3	2.1e-3	6.2e-2	1.7e-2
80	2.2e-4	4.9e-2	1.2e-3	1.4e-4	4.9e-2	9.3e-4	1.6e-4	7.8e-3	1.2e-3
160	3.7e-5	2.4e-2	2.1e-4	3.6e-5	2.4 e-2	2.5e-4	1.6e-5	1.1e-3	1.2e-4

4.2. **An outflow layer problem.** The second test case is a convection-diffusion problem with $\sigma = 0$ and β constant of norm 1. Experiments are performed for different values of the diffusion coefficient ($\varepsilon = 0.1$, $\varepsilon = 0.01$, and $\varepsilon = 0.001$) on a series of uniformly refined meshes. Nonhomogeneous Dirichlet boundary conditions are imposed as shown in Figure 1, yielding an outflow boundary layer whose thickness is of order ε. The discrete solution with CIP but no shock-capturing, $\tilde{U}$, exhibits strong oscillations; this is not the case for the discrete solution obtained with shock-capturing, U (see Figure 1).

Consider the error $\tilde{e} = U - \tilde{U}$. Our results show that in all cases, the H^1-norm contribution to the error dominates the L^2-norm contribution by two orders of magnitude and that the L^2-norm of the error converges much faster than the H^1-norm. Therefore, the H^1-norm of the error is compared with the bound resulting from the a priori estimates of Section 3.3. This bound, say $\eta(h,\varepsilon)$, is evaluated as follows: For $h > \varepsilon$, $\eta(h,\varepsilon)$ is evaluated from (3.14) with $c_6 = 1$, and for $h < \varepsilon$, $\eta(h,\varepsilon)$ is set to the minimal value of (3.9) with $c_1\|u\|_{H^2(\Omega)} = \varepsilon^{-1}$ and that of (3.14) with $c_6 = 1$. Results are reported in Table 2. For all values of ε, whenever the layer is sufficiently resolved, $\eta(h,\varepsilon)$ stays within a factor of three with respect to the actual error. Hence, the a priori bound $\eta(h,\varepsilon)$ accurately describes the triple norm behavior when passing from the convection-dominated regime to the diffusion-dominated regime. To illustrate, Figure 2 displays the H^1–norm of the error $\|\nabla(u-U)\|$ for $\varepsilon = 0.02$ and with u evaluated from a reference solution computed on a uniformly refined grid containing $131,585$ nodes. The change in the behavior of the error when $h \simeq \varepsilon$ is clearly visible.

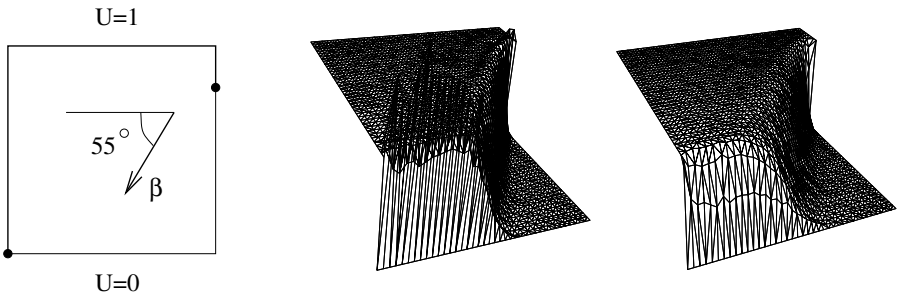

FIGURE 1. Outflow layer problem. Left: test case setup; middle: discrete solution without shock-capturing; right: discrete solution with shock-capturing

TABLE 2. Outflow layer problem: convergence results

ε	0.1		0.01		0.001	
N	$\|\varepsilon^{\frac{1}{2}}\nabla\tilde{e}\|$	$\eta(h,\varepsilon)$	$\|\varepsilon^{\frac{1}{2}}\nabla\tilde{e}\|$	$\eta(h,\varepsilon)$	$\|\varepsilon^{\frac{1}{2}}\nabla\tilde{e}\|$	$\eta(h,\varepsilon)$
20	1.0e-1	2.0e-1	8.5e-1	2.2	1.2e-1	7.0
40	3.0e-2	5.0e-2	1.3	1.6	1.1	5.0
80	1.0e-2	1.6e-2	9.3e-1	1.1	1.7	3.5
160	2.4e-3	5.3e-3	2.8e-1	1.0e-1	1.6	2.5

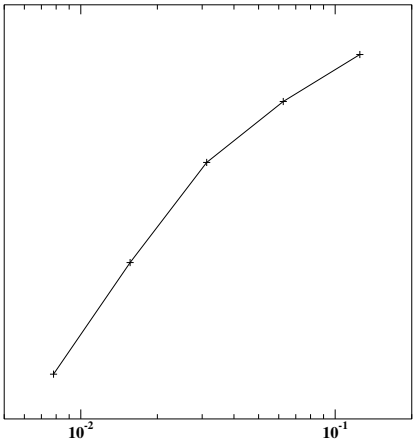

FIGURE 2. Outflow layer problem: H^1-norm of the error versus mesh size

Table 3 reports DMP violations for the discrete solutions $\tilde{U}$ and U in percentage of $\|u\|_{L^\infty(\Omega)}(=1)$. When the diffusion coefficient is large enough ($\varepsilon = 0.1$), no DMP violations are observed, regardless of the use of a shock-capturing term. For lower values of ε, DMP violations without shock-capturing can be as high as 86%. When $\varepsilon = 0.01$, the regularized shock-capturing term manages to quench spurious oscillations on the finer meshes while overshoots lower than 1% remain on the coarsest mesh. For sharper layers ($\varepsilon = 0.001$), the regularization of the absolute value is too loose, thereby leaving overshoots ranging between 2% and 4%.

TABLE 3. Outflow layer problem: DMP violations

ε	0.1		0.01		0.001	
N	$\tilde{U}$	U	$\tilde{U}$	U	$\tilde{U}$	U
20	0	0	36	0.5	42	2.2
40	0	0	33	0.0	86	2.1
80	0	0	0	0.0	78	2.1
160	0	0	0	0.0	58	4.2

5. Conclusion

In this paper, we have analyzed a new nonlinear shock-capturing scheme based on the jump in the gradient between adjacent elements. A discrete maximum principle has been proved rigorously for H^1-conforming, piecewise-affine finite element approximations of convection-diffusion-reaction equations. The shock-capturing finite element solution (super-)converges to the finite element solution without shock-capturing if the cell Péclet numbers are small enough. The sharpness of the a priori estimates has been verified numerically on test problems with internal and outflow layers. In particular, the behavior of the error in the energy norm is well captured when passing from the convection-dominated to the diffusion-dominated regime. The present finite element schemes can now be used for more complex problems, such as chemically reactive flows, by controlling the jumps of temperature and chemical species gradients.

Acknowledgment

The authors are thankful to an anonymous referee for valuable comments.

References

[1] J.H. Bramble, R. Lazarov, and J. Pasciak. A least-squares approach based on a discrete minus one inner product for first-order systems. *Math. Comp.*, **66**(219):935–955, 1997. MR1415797 (97m:65202)

[2] A.N. Brooks and T.J.R. Hughes. Streamline Upwind/Petrov–Galerkin formulations for convective dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, **32**:199–259, 1982. MR0679322 (83k:76005)

[3] E. Burman and A. Ern. Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the advection-diffusion-reaction equation. *Comput. Methods Appl. Mech. Engrg.*, **191**:3822–3855, 2002. MR1912655 (2003e:65211)

[4] E. Burman and A. Ern. Discrete maximum principle for Galerkin approximations of the Lapalce operator on arbitrary meshes. *C. R. Acad. Sci. Paris, Sér. I*, **338**:641–646, 2004. MR2056474

[5] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comp. Methods Appl. Mech. Engrg.*, **193**:1437–1453, 2004. MR2068903 (2005d:65186)

[6] P.G. Ciarlet and P.-A. Raviart. Maximum principle and uniform convergence for the finite element method. *Comput. Methods Appl. Mech. Engrg.*, **2**:17–31, 1973. MR0375802 (51:11992)

[7] R. Codina. A discontinuity-capturing crosswind dissipation for the finite element solution of the convection-diffusion equation. *Comput. Methods Appl. Mech. Engrg.*, **110**:325–342, 1993. MR1256324 (94m:76074)

[8] A. Drăgănescu, T.F. Dupont, and L.R. Scott. Failure of the discrete maximum principle for an elliptic finite element problem. *Math. Comp.*, **74**:1–23, 2005. MR2085400

[9] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*. Vol. 159 of Applied Mathematical Series, Springer-Verlag, New York, 2004. MR2050138 (2005d:65002)

[10] J.-L. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *Math. Model. Numer. Anal. (M2AN)*, **33**(6):1293–1316, 1999. MR1736900 (2000m:65114)

[11] T.J.R. Hughes, M. Mallet, and A. Mizukami. A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. *Comput. Methods Appl. Mech. Engrg.*, **54**:341–355, 1986. MR0836189 (87f:76010b)

[12] S. Idelsohn, N. Nigro, M. Storti, and G. Buscaglia. A Petrov-Galerkin formulation for advection-reaction-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, **136**:27–46, 1996. MR1409694 (97f:76059)

[13] B. Jiang. *The Least-Squares Finite element Method*. Scientific Computation, Springer-Verlag, New York, 1989. MR1639101 (99f:65141)

[14] C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic equations. *Comput. Methods Appl. Mech. Engrg.*, **45**:285–312, 1984. MR0759811 (86a:65103)
[15] C. Johnson, A. Schatz, and L. Wahlbin. Crosswind smear and pointwise error in streamline diffusion finite element methods. *Math. Comp.*, **49**:25–38, 1987. MR0890252 (88i:65130)
[16] T. Knopp; G. Lube; G. Rapin, Stabilized finite element methods with shock capturing for advection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, **191**:2997–3013, 2002. MR1903196 (2003c:65125)
[17] S. Korotov, M. Křížek, and P. Neittaanmäki. Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. *Math. Comp.*, **70**(233):107–119, 2000. MR1803125 (2001i:65126)
[18] Y.T. Shih and H.C. Elman. Modified streamline diffusion schemes for convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, **174**:137–151, 1999. MR1686684 (2000c:76052)
[19] T.E. Tezduyar and Y.J. Park. Discontinuity-capturing finite element formulations for nonlinear convection-diffusion-reaction equations. *Comput. Methods Appl. Mech. Engrg.*, **59**:307–325, 1986.
[20] J. Xu and L. Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Math. Comp.*, **68**(228):1429–1446, 1999. MR1654022 (99m:65225)

Ecole Polytechnique Federale de Lausanne, Institute of Analysis and Scientific Computing, 1015 Lausanne, Switzerland
E-mail address: Erik.Burman@epfl.ch

CERMICS, Ecole nationale des ponts et chaussées, 6 et 8, avenue B. Pascal, 77455 Marne la Vallée cedex 2, France
E-mail address: ern@cermics.enpc.fr