IMA Journal of Numerical Analysis (2024) **44**, 2198–2219
https://doi.org/10.1093/imanum/drad055
Advance Access publication on 26 August 2023

A nodally bound-preserving finite element method

Gabriel R. Barrenechea*
Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland
*Corresponding author: gabriel.barrenechea@strath.ac.uk

Emmanuil Georgoulis
School of Mathematical & Computer Sciences, Heriot Watt University, UK
Department of Mathematics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Greece
IACM-FORTH, Greece

Tristan Pryer
Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK

and

Andreas Veeser
Dipartimento di Matematica, Università degli Studi di Milano, Via C. Saldini 50, 20133 Milano, Italy

[Received on 11 June 2022; revised on 25 March 2023]

This work proposes a nonlinear finite element method whose nodal values preserve bounds known for the exact solution. The discrete problem involves a nonlinear projection operator mapping arbitrary nodal values into bound-preserving ones and seeks the numerical solution in the range of this projection. As the projection is not injective, a stabilisation based upon the complementary projection is added in order to restore well-posedness. Within the framework of elliptic problems, the discrete problem may be viewed as a reformulation of a discrete obstacle problem, incorporating the inequality constraints through Lipschitz projections. The derivation of the proposed method is exemplified for linear and nonlinear reaction-diffusion problems. Near-best approximation results in suitable norms are established. In particular, we prove that, in the linear case, the numerical solution is the best approximation in the energy norm among all nodally bound-preserving finite element functions. A series of numerical experiments for such problems showcase the good behaviour of the proposed bound-preserving finite element method.

1. Introduction

Structure-preserving numerical methods have been an overarching theme in computational partial differential equations (PDEs) over the years. By structure-preserving we mean methods that produce approximations satisfying certain desired properties of the underlying exact problem, e.g., local conservation, entropy inequalities, maximum principle, pointwise divergence-free constraints, or exactly symmetric stress tensor approximations, just to name a few.

Numerical methods satisfying *Discrete Maximum Principles (DMP)* and/or *monotonicity* properties have been studied extensively in the finite element literature; see Ciarlet (1970); Ciarlet & Raviart (1973); Kikuchi (1977); Mizukami & Hughes (1985); Burman & Ern (2005); Brandts *et al.* (2008); Barrenechea *et al.* (2017a,b) for a (very) nonexhaustive list, and Barrenechea *et al.* (2023) for a recent review and more related references. Methods satisfying the latter properties imply *positivity preservation* or, more generally, *bound preservation* of the resulting numerical solutions.

Crucially, however, bound-preservation appears to be a weaker structural requirement than the DMP. Nonetheless, bound-preserving numerical solutions are indispensable for the numerical stability of many complex nonlinear phenomena modelled by systems of PDEs. For instance, many such PDE models are valid only for positive solutions of the constituent equations. Examples include nonlinear reaction-diffusion systems modelling concentrations of reactants, or turbulence-inducing fields. Also, phase-field PDE models are usually characterised by solutions satisfying pointwise global maxima and minima. Although failure to preserve the same bounds for the numerical solutions is not typically catastrophic for scalar PDE problems, they may have a compounding effect when such PDEs are part of a more complex system of equations. There is thus an interest in bound-preserving methods which, when appropriately designed, may be less 'rigid' than current schemes striving to satisfy the DMP or monotonicity.

This work proposes a new finite element method in this spirit. To this end, assume that we are given a partial differential equation (with its associated weak form), and a Lagrange finite element space of arbitrary order. Moreover, for simplicity, assume that the exact solution is positive, i.e., the zero function is a lower bound. Then, the formal derivation of the new method can be summarised by the following three steps:

- Introduce a projection operator $v_h \mapsto v_h^+$ on the finite element space so that the nodal values of the projected function v_h^+ are positive.
- As usual, replace the test space in the weak formulation of the continuous problem with the finite element space, but, instead of looking for a finite element function as the numerical solution, seek for a projected one in the form u_h^+.
- Add a stabilisation based upon the complementary projection $v_h \mapsto v_h^-$ onto the finite element space.

Notice that the projection operator and so the proposed method are nonlinear. The latter is expected in light of the classical 'Godunov barrier' principle. The solution of the proposed method is sought in the range of the projection, where the projection acts as a parametrisation and the parameter domain is the finite element space. The fact that the projection is not injective entails the need of the stabilisation in the third step.

Interestingly, if its nodal values are positive, the classical finite element solution coincides with the one of the proposed method. More precisely, and more importantly, the solution to the present method turns out to be the orthogonal projection of the exact solution onto the closed and convex set of finite element functions with positive nodal values. This observation is instrumental in this work and yields the following two important consequences:

- The proposed method can be viewed as a reformulation of a discrete obstacle problem. In contrast to the latter, the reformulation consists of a discrete variational equality and encodes the crucial inequality constraints in Lipschitz-continuous projections. This difference may prepare the ground for solvers that are an alternative to the usually used constrained optimisation techniques; see, e.g., Evans *et al.* (2009). Here, we employ a simple approach based upon Richardson-like iterations.
- Near best approximation for suitable error notions follows as a corollary. In particular, for the linear reaction-diffusion problem, the solution of the proposed method is actually the best approximation to the exact solution in the energy norm from the aforementioned convex subset. For a more general, semilinear problem with a power-like nonlinear reaction, the proposed method is shown

to be quasi-optimal with respect to the sum of the H^1-seminorm and a quasinorm associated with the reaction term.

To illustrate both aspects, a series of numerical experiments is presented, showcasing the superior performance of the proposed method compared to standard finite element approximations.

It is worth mentioning that related approaches (albeit with distinct differences to the present one) have been advocated in the literature. In Bochev *et al.* (2020) the bounds on the continuous solution are imposed as a restriction in a constraint optimisation problem, and a link to nonlinear stabilisation is presented. Also, in Mudunuru & Nakshatrala (2016) a constrained optimisation problem involving a mixed weak formulation is employed to enforce bound-preservation. In Burman & Ern (2017) the link between positivity-preservation and the contact problem is used to motivate a nonlinear stabilised method that enforces the positivity of the solution in a weak way. In addition, the *cut-off* finite element method (Lu *et al.*, 2013) truncates the finite element function *after* it is computed at a given time step, so as to input the truncated function as approximation of the current time step; see also Yang *et al.* (2022) for an application of a related idea to the Allen-Cahn equation. In the steady state case, the idea of *truncating* the finite element solution to respect given bounds has been justified for reaction-diffusion equations in Kreuzer (2014) using energy arguments. In Lisa & Sashkov (2008) a conservative recovery strategy is proposed and tested numerically. Finally, in the context of the Joule heating problem a truncation of one of the variables is introduced in order to regularise a rough right-hand side in Jensen & Malquist (2013).

The remainder of the manuscript is organised as follows. The rest of introduction is devoted to setting up notations and main assumptions of this work. The finite element method is presented in Section 3, and its stability and optimal error estimates are proven. Then, in Section 4 we extend this framework to a nonlinear reaction-diffusion equation. Finally, Section 5 is devoted to presenting numerical experiments.

2. General setting and the linear model problem

We will use standard notation for Sobolev spaces, in line with, e.g., Ern & Guermond (2021). More precisely, a domain $D \subseteq \mathbb{R}^d$, $d = 1, 2, 3$, we denote by $\|\cdot\|_{0,p,D}$ the $L^p(D)$-norm, when $p = 2$ the subscript p will be often omitted, and we only write $\|\cdot\|_{0,D}$. In addition, for $s \geq 0, p \in (1,\infty)$, we denote by $\|\cdot\|_{s,p,D}$ ($|\cdot|_{s,p,D}$) the norm (seminorm) in $W^{s,p}(D)$; when $p = 2$, we will again often omit the subscript p. In addition, we denote by $H^{-1}(D)$ the dual of $H^1_0(D)$ while identifying $L^2(D)$ with its dual. Thus, writing $\langle\cdot,\cdot\rangle_D$ for the duality pairing, we have

$$\langle f, v\rangle_\Omega = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \forall\, v \in H^1_0(\Omega),$$

whenever $f \in H^{-1}(D)$ is regular enough. The $L^2(D)$-inner product is denoted by $(\cdot,\cdot)_D$. We will not distinguish between inner product and duality pairing for scalar and for vector-valued functions.

The boundary-value problems we will be concerned with are posed in an open, bounded domain Ω with polyhedral Lipschitz boundary $\partial\Omega$. Let $\mathcal{P}$ be a conforming, shape-regular partition of Ω into simplices (or quadrilaterals/hexahedra) K. Over $\mathcal{P}$, and for $k \geq 1$, we define the finite element space

$$V_{\mathcal{P}} := \{v_h \in C^0(\Omega) : v_h|_K \in \mathcal{R}(K)\ \forall\, K \in \mathcal{P}\} \cap H^1_0(\Omega), \tag{2.1}$$

where

$$\mathcal{R}(K) = \begin{cases} \mathbb{P}_k(K) & \text{if } K \text{ is a simplex,} \\ \mathbb{Q}_k(K) & \text{if } K \text{ is an affine quadrilateral/hexahedron,} \end{cases} \tag{2.2}$$

with $\mathbb{P}_k(K)$ denoting the polynomials of total degree k on K and $\mathbb{Q}_k(K)$ denoting the the affine pull-back of the polynomial of degree less than, or equal to, k in each variable. We denote by $\boldsymbol{x}_i, i = 1, \dots, N$ the set of internal nodes of $\mathcal{P}$, and by $\phi_1, \dots, \phi_N$ the set of usual Lagrangian basis functions spanning the space $V_{\mathcal{P}}$.

The diameter of K is denoted by h_K, $h = \max\{h_K : K \in \mathcal{P}\}$, and we define the mesh function $\mathfrak{h}$ as a continuous, element-wise linear function defined as a local average. More precisely, for a node $\boldsymbol{x}_i$ of $\mathcal{P}$ we define the local neighbourhood $K_{\boldsymbol{x}_i} = \{K \in \mathcal{P} : \boldsymbol{x}_i \in K\}$, and define $\mathfrak{h}$ as the only element of $V_{\mathcal{P}}$ (with $k = 1$) given by the nodal values

$$\mathfrak{h}(\boldsymbol{x}_i) = \frac{\sum_{K \in K_{\boldsymbol{x}_i}} h_K}{\# K_{\boldsymbol{x}_i}}. \tag{2.3}$$

We recall the inverse inequality (see, e.g., Ern & Guermond, 2021, Lemma 12.1): for all $m, \ell \in \mathbb{N}, m \leq \ell$ and all $p, q \in [1, +\infty]$, there exists a constant C, independent of $K \in \mathcal{P}$ such that

$$|q|_{\ell,p,K} \leq C h_K^{m-\ell+d\left(\frac{1}{p}-\frac{1}{q}\right)} |q|_{m,q,K}\,, \tag{2.4}$$

for every polynomial function q defined on K.

In the space $V_{\mathcal{P}}$ we denote by $(\cdot,\cdot)_h$ the lumped $L^2(\Omega)$-inner product given by

$$(v_h, w_h)_h := \sum_{i=1}^{N} \mathfrak{h}^d(\boldsymbol{x}_i) v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i), \tag{2.5}$$

with associated norm $|v_h|_h = (v_h, v_h)_h^{\frac{1}{2}}$.

Remark 2.1 The following result, whose proof can be found in Proposition 12.5 of Ern & Guermond (2021), will be of importance for us in the analysis of the method proposed in Section 3: there exist $C, c > 0$, independent of $\mathfrak{h}$, such that

$$c\,|v_h|_h^2 \leq \|v_h\|_{0,\Omega}^2 \leq C\,|v_h|_h^2\,, \tag{2.6}$$

for all $v_h \in V_{\mathcal{P}}$.

2.1 *Linear model problem*

Given $f \in H^{-1}(\Omega)$ we consider the following reaction-diffusion equation: find $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -\mathrm{div}\left(\mathcal{D}\nabla u\right) + \mu u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{2.7}$$

with $\mathcal{D} = \big(d_{ij}\big)_{i,j=1}^{d} \in L^\infty(\Omega)^{d\times d}$ and $\mu \in L^\infty(\Omega)$ stand for the diffusion tensor and reaction coefficient, respectively. We assume that $\mu(\boldsymbol{x}) \geq \mu_0 \geq 0$ a.e. in Ω, and that the diffusion tensor $\mathcal{D}$ is symmetric and uniformly strictly positive-definite in Ω, viz., there exists a positive constant $d_0 > 0$ such that, for almost all $\boldsymbol{x} \in \Omega$, we have

$$\sum_{i,j=1}^{d} y_i d_{ij}(\boldsymbol{x}) y_j \geq d_0 \sum_{i=1}^{d} y_i^2 \qquad \forall\,(y_1,\dots,y_d) \in \mathbb{R}^d. \tag{2.8}$$

The weak formulation of (2.7) reads: find $u \in H_0^1(\Omega)$, such that

$$a(u,v) = \langle f, v\rangle_\Omega \qquad \forall\, v \in H_0^1(\Omega), \tag{2.9}$$

where $a(\cdot,\cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is the bilinear form defined by

$$a(v,w) := (\mathcal{D}\nabla v, \nabla w)_\Omega + (\mu v, w)_\Omega. \tag{2.10}$$

Defining the energy norm

$$\|v\|_a := \sqrt{a(v,v)}, \tag{2.11}$$

the well-posedness of (2.9) follows from the Lax-Milgram Lemma (see, e.g., Ern & Guermond, 2021).

REMARK 2.2 At the core of this work is the following property of the solution of (2.9): as a consequence of maximum and comparison principles (see Gilbarg & Trudinger, 2001, or Renardy & Rogers, 2004, Corollary 4.4) the following bounds can be proven: for almost all $\boldsymbol{x} \in \Omega$ the solution u of (2.9) satisfies

$$-\frac{\|f\|_{0,\infty,\Omega}}{\mu_0} \leq u(\boldsymbol{x}) \leq \frac{\|f\|_{0,\infty,\Omega}}{\mu_0}. \tag{2.12}$$

This last statement can be made more precise if $f \geq 0$ in Ω. In fact, in this case for almost all $\boldsymbol{x} \in \Omega$ the following inequality holds:

$$0 \leq u(\boldsymbol{x}) \leq \frac{\|f\|_{0,\infty,\Omega}}{\mu_0}. \tag{2.13}$$

The results given in the above remark motivate the introduction of the following assumption.
Assumption (A1): We will suppose that the solution of (2.7) satisfies

$$0 \leq u(x) \leq \mathfrak{a} \qquad \text{for almost all } x \in \Omega, \tag{2.14}$$

where $\mathfrak{a}$ is a known constant.

REMARK 2.3 The lower bound in Assumption (A1) is not required to be equal to zero and all the results proven below hold for a more general bounding box without major modifications. In particular, the value $\mathfrak{a}$ can be replaced by a non-negative, continuous function $\mathfrak{a}(\boldsymbol{x})$.

The standard Galerkin finite element method for (2.7) is given by: find $u_h^{\mathrm{FEM}} \in V_{\mathcal{P}}$ such that

$$a\big(u_h^{\mathrm{FEM}}, v_h\big) = \langle f, v_h\rangle_{\Omega} \qquad \forall\, v_h \in V_{\mathcal{P}}. \tag{2.15}$$

It is well-known that u_h^{FEM} is a best approximation in the following sense:

$$\big\|u - u_h^{\mathrm{FEM}}\big\|_a = \inf_{v_h \in V_{\mathcal{P}}} \|u - v_h\|_a. \tag{2.16}$$

However, this nice property does not prevent that, when the reaction μ dominates the diffusion $\mathcal{D}$, that is, e.g., if $\|\mathcal{D}\|_{0,\infty,\Omega} \ll \mu$, then u_h^{FEM} may exhibit spurious oscillations (see, e.g., Roos *et al.*, 2008) and, thus, u_h^{FEM} may fail to satisfy Assumption (A1). In fact, very stringent conditions need to be imposed on the mesh for the solution of a reaction-diffusion equation such as (2.9) to satisfy Assumption (A1) (for details, see, Brandts *et al.*, 2008), even in the case of scalar (isotropic) diffusion. For the case of a general diffusion $\mathcal{D}$, conditions on the mesh related to the weighted inner product $(\mathcal{D}\cdot, \cdot)$ in $\mathbb{R}^d$ need to be imposed, in addition to a mesh size restriction (for details, see, Huang, 2011).

3. The finite element method for the linear problem

The goal of this section is to derive a method that, on the one hand, essentially preserves the bounds in Assumption (A1) without restrictions on the mesh and, on the other hand, maintains the good approximation properties (2.16) of the classical finite element solution.

To this end, we define the following closed convex subset of $V_{\mathcal{P}}$:

$$V_{\mathcal{P}}^{+} := \{v_h \in V_{\mathcal{P}} : v_h(\boldsymbol{x}_i) \in [0, \mathfrak{a}] \text{ for all } \boldsymbol{x}_i = 1, \dots, N\}, \tag{3.1}$$

that is, the set of finite element functions that respect the bound (2.14) at their degrees of freedom. With this convex set in mind, every finite element function is split as $v_h = v_h^{+} + v_h^{-}$, where $v_h^{+} \in V_{\mathcal{P}}^{+}$ is defined by

$$v_h^{+} := \sum_{i=1}^{N} \max\big\{0, \min\{v_h(\boldsymbol{x}_i), \mathfrak{a}\}\big\}\phi_i, \tag{3.2}$$

and

$$v_h^{-} := v_h - v_h^{+}, \tag{3.3}$$

is the part of the function residing outside $V_{\mathcal{P}}^{+}$. From now on, we refer to the functions v_h^{+} and v_h^{-} as the *constrained* and *complementary* parts of v_h.

The finite element method proposed in this work reads: find $u_h \in V_{\mathcal{P}}$, such that

$$a_h(u_h; v_h) = \langle f, v_h\rangle_{\Omega} \qquad \forall\, v_h \in V_{\mathcal{P}}, \tag{3.4}$$

where $a_h(\cdot;\cdot)$ is the nonlinear form given by

$$a_h(u_h; v_h) := a(u_h^{+}, v_h) + s(u_h^{-}, v_h), \tag{3.5}$$

with $a(\cdot,\cdot)$ defined in (2.10), and $s(\cdot,\cdot) : C(\bar{\Omega}) \times C(\bar{\Omega}) \to \mathbb{R}$ is the stabilizing bilinear form defined by

$$s(v_h, w_h) = \alpha \sum_{i=1}^{N} \left(\|\mathcal{D}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \|\mu\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d} \right) v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i), \tag{3.6}$$

where $\alpha > 0$ is a nondimensional constant to be determined precisely in Theorem 3.2, and $\omega_i := \cup_{K\in\mathcal{P}:K\cap K_{\boldsymbol{x}_i}\neq\emptyset} K$ denotes an extended patch. The definition of ω_i will be exploited in establishing (3.8) below.

Defining the stabilisation norm as

$$\|v_h\|_s := \sqrt{s(v_h, v_h)}, \tag{3.7}$$

and using (2.6), there exists a $C_{\text{equiv}} > 0$, depending only on the shape-regularity constant, such that

$$\|v_h\|_a^2 \le \frac{C_{\text{equiv}}}{\alpha} \|v_h\|_s^2 \qquad \forall\, v_h \in V_{\mathcal{P}}, \tag{3.8}$$

where $\alpha > 0$ is the stabilisation parameter appearing in the definition (3.6) of $s(\cdot,\cdot)$.

3.1 *Well-posedness and consistency*

In this section we will analyse the existence, uniqueness, and stability results for the proposed method (3.4). We start with the following monotonicity result for the stabilizing form $s(\cdot,\cdot)$.

LEMMA 3.1 The bilinear form $s(\cdot,\cdot)$ satisfies the following inequalities:

$$s(v_h^- - w_h^-, v_h^+ - w_h^+) \ge 0 \qquad \forall\, v_h, w_h \in V_{\mathcal{P}}, \tag{3.9}$$

$$s(v_h^-, w_h - v_h^+) \le 0 \qquad \forall\, v_h \in V_{\mathcal{P}}, w_h \in V_{\mathcal{P}}^+. \tag{3.10}$$

Proof. Let $\boldsymbol{x}_i \in \mathcal{P}$ be any internal node. If $v_h(\boldsymbol{x}_i) \ge w_h(\boldsymbol{x}_i)$, then $v_h^+(\boldsymbol{x}_i) \ge w_h^+(\boldsymbol{x}_i)$. Moreover, if $v_h(\boldsymbol{x}_i) \ge w_h(\boldsymbol{x}_i) > \mathfrak{a}$, then $v_h^-(\boldsymbol{x}_i) = v_h(\boldsymbol{x}_i) - \mathfrak{a} \ge w_h(\boldsymbol{x}_i) - \mathfrak{a} = w_h^-(\boldsymbol{x}_i)$. If $v_h(\boldsymbol{x}_i) > \mathfrak{a}$ and $w_h(\boldsymbol{x}_i) \le \mathfrak{a}$, then $v_h^-(\boldsymbol{x}_i) = v_h(\boldsymbol{x}_i) - \mathfrak{a} \ge 0 \ge w_h^-(\boldsymbol{x}_i)$. If $v_h(\boldsymbol{x}_i) \le \mathfrak{a}$, then either $v_h^-(\boldsymbol{x}_i) = 0 \ge w_h^-(\boldsymbol{x}_i)$, or $v_h^-(\boldsymbol{x}_i) = v_h(\boldsymbol{x}_i) \ge w_h(\boldsymbol{x}_i) = w_h^-(\boldsymbol{x}_i)$. Thus,

$$\begin{aligned} &s(v_h^- - w_h^-, v_h^+ - w_h^+) \\ &\quad = \alpha \sum_{i=1}^{N} \left(\|\mathcal{D}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-2} + \|\mu\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d} \right) (v_h^- - w_h^-)(\boldsymbol{x}_i)(v_h^+ - w_h^+)(\boldsymbol{x}_i) \\ &\quad \ge 0, \end{aligned} \tag{3.11}$$

which proves (3.9). Taking $w_h \in V_{\mathcal{P}}^+$ gives $w_h^- = 0$, and then (3.10) follows from (3.9). □

THEOREM 3.2 (Well-posedness). Let $T : V_{\mathcal{P}} \to [V_{\mathcal{P}}]'$ be the mapping defined by

$$[Tv_h, w_h] = a(v_h^+, w_h) + s(v_h^-, w_h). \tag{3.12}$$

Then, T is continuous and, if the nondimensional parameter α is chosen such that $\alpha \geq C_{\text{equiv}}$, it is also strongly monotone, since then T satisfies: there exists $\beta > 0$, independent of h, such that

$$[Tv_h - Tw_h, v_h - w_h] \geq \beta \,\|v_h - w_h\|_a^2, \tag{3.13}$$

for all $v_h, w_h \in V_{\mathcal{P}}$. As a consequence, (3.4) has a unique solution $u_h \in V_{\mathcal{P}}$.

Proof. We start defining the mesh-dependent norm $\|\cdot\|_{\mathcal{P}}$ by

$$\|v_h\|_{\mathcal{P}} := \left\{\|v_h\|_a^2 + \|v_h\|_s^2\right\}^{\frac{1}{2}}. \tag{3.14}$$

Then, for all $v_h, w_h, z_h \in V_{\mathcal{P}}$ using the Cauchy-Schwarz inequality and (3.8) we get to

$$\begin{aligned} [Tv_h - Tw_h, z_h] &= a\big(v_h^+ - w_h^+, z_h\big) + s\big(v_h^- - w_h^-, z_h\big) \\ &\leq \big(\|v_h^+ - w_h^+\|_a^2 + \|v_h^- - w_h^-\|_s^2\big)^{\frac{1}{2}} \,\|z_h\|_{\mathcal{P}}, \end{aligned} \tag{3.15}$$

which proves the continuity of T. To prove the monotonicity of T, let $v_h, w_h \in V_{\mathcal{P}}$. Using the Cauchy-Schwarz and Young inequalities, and (3.8), we obtain

$$\begin{aligned} [Tv_h - Tw_h, v_h - w_h] &= a(v_h^+ - w_h^+, v_h - w_h) + s(v_h^- - w_h^-, v_h - w_h) \\ &= \|v_h^+ - w_h^+\|_a^2 + a(v_h^+ - w_h^+, v_h^- - w_h^-) + \|v_h^- - w_h^-\|_s^2 + s(v_h^- - w_h^-, v_h^+ - w_h^+) \\ &\geq \frac{1}{2}\|v_h^+ - w_h^+\|_a^2 + \left(1 - \frac{C_{\text{equiv}}}{2\alpha}\right)\|v_h^- - w_h^-\|_s^2 + s(v_h^- - w_h^-, v_h^+ - w_h^+) \\ &\geq \frac{1}{2}\|v_h^+ - w_h^+\|_a^2 + \left(1 - \frac{C_{\text{equiv}}}{2\alpha}\right)\|v_h^- - w_h^-\|_s^2, \end{aligned} \tag{3.16}$$

where in the last inequality we used (3.9) in Lemma 3.1. Next, using that $\alpha \geq C_{\text{equiv}}$ and (3.8) we arrive at

$$\left(1 - \frac{C_{\text{equiv}}}{2\alpha}\right)\|v_h^- - w_h^-\|_s^2 \geq \frac{\alpha}{2C_{\text{equiv}}}\|v_h^- - w_h^-\|_a^2. \tag{3.17}$$

Hence, (3.13) follows replacing (3.17) in (3.16) and the fact that, since $a(\cdot,\cdot)$ is symmetric and elliptic, we have $\|v_h\|_a \leq \sqrt{2}\,(\|v_h^+\|_a + \|v_h^-\|_a)$.

Finally, the existence and uniqueness of solutions follows by using classical results in monotone operator theory (see, e.g., Renardy & Rogers, 2004, Theorem 10.49). □

For the next observation, it is useful to recall that, for Galerkin methods, consistency can be expressed as an invariance property of the operator mapping the exact solution to the Galerkin one.

LEMMA 3.3 (Consistency).Under Assumption (A1), the method (3.4) enjoys the following invariance property: if the exact solution u belongs to $V_{\mathcal{P}}$, then $u_h^+ = u_h = u$.

Proof. As $u \in V_{\mathcal{P}}$, the functions u^+ and u^- are defined and, due to Assumption (A1), we have $u^+ = u$ and $u^- = 0$. As a result, we get

$$a_h(u; v_h) = a(u^+, v_h) = a(u, v_h) = \langle f, v_h \rangle_{\Omega},$$

for all $v_h \in V_{\mathcal{P}}$. Thanks to Theorem 3.2, the solution of (3.4) is unique and therefore we obtain $u_h = u$. Using that $u_h^- = u^- = 0$, we conclude that $u_h^+ = u_h$. □

REMARK 3.4 Whenever $k = 1$ the method (3.4) is actually bound-preserving throughout the domain Ω, not only at the nodes. Furthermore, if $\mu > 0$, in addition it also respects the discrete maximum principle. That is, if $f \geq 0$ then u_h^+ cannot attain an interior negative minimum, and reaches its minimum at the boundary. If $\mu = 0$, it appears that the method by construction does not guarantee that u_h^+ cannot attain an interior minimum. Nonetheless, all our numerical experiments to date have failed to produce such a case.

3.2 *Characterisation of the constrained part and error estimates*

One of the salient features of the method (3.4) is that u_h^+ is characterised as being the unique solution of a variational inequality posed on the closed convex set $V_{\mathcal{P}}^+$. This is proven in the next result.

THEOREM 3.5 (Characterisation of the constrained part). Let $u_h \in V_{\mathcal{P}}$ be the unique solution of (3.4). Then, u_h^+ satisfies the following variational inequality: $u_h^+ \in V_{\mathcal{P}}^+$ and satisfies

$$a(u_h^+, v_h - u_h^+) \geq \langle f, v_h - u_h^+ \rangle_{\Omega} \qquad \forall\, v_h \in V_{\mathcal{P}}^+. \tag{3.18}$$

Proof. Using (3.4) with $v_h \in V_{\mathcal{P}}$, v_h^-, and u_h^+ as test functions we have the following equalities:

$$\begin{aligned} a(u_h^+, v_h) + s(u_h^-, v_h) &= \langle f, v_h \rangle_{\Omega}, \\ a(u_h^+, v_h^-) + s(u_h^-, v_h^-) &= \langle f, v_h^- \rangle_{\Omega}, \\ a(u_h^+, u_h^+) + s(u_h^-, u_h^+) &= \langle f, u_h^+ \rangle_{\Omega}. \end{aligned}$$

Subtracting the second and third equation from the first one, and using that $v_h^+ = v_h - v_h^-$, we get

$$a(u_h^+, v_h^+ - u_h^+) + s(u_h^-, v_h^+ - u_h^+) = \langle f, v_h^+ - u_h^+ \rangle_{\Omega}, \tag{3.19}$$

for all $v_h^+ \in V_{\mathcal{P}}^+$. Finally, using (3.10) in Lemma 3.1, we get that $u_h^+ \in V_{\mathcal{P}}^+$ satisfies (3.18), thus completing the proof. □

REMARK 3.6 Due to Stampacchia's Theorem, problem (3.18) can be proven directly to have a unique solution. This proves that both the stabilized method (3.4) and the variational inequality (3.18) are, in fact, equivalent.

The equivalence of the method as a variational inequality enables us to prove best approximation error estimates in a standard fashion.

THEOREM 3.7 (Abstract error analysis). Let u be the solution of (2.7) and $u_h \in V_{\mathcal{P}}$ be the unique solution of (3.4). Then,

$$\|u - u_h^+\|_a = \min_{v_h \in V_{\mathcal{P}}^+} \|u - v_h\|_a. \tag{3.20}$$

Moreover, let u_h^{FEM} be the solution of the (standard) finite element method (2.15). Then, the negative part u_h^- satisfies the following error estimate

$$s(u_h^-, u_h^-)^{\frac{1}{2}} \leq \sqrt{\frac{C_{\text{equiv}}}{\alpha}} \min\left\{\|u - u_h^+\|_a, \|u_h^{\text{FEM}} - u_h^+\|_a\right\}. \tag{3.21}$$

Proof. Since u_h^+ is the solution of (3.18), it satisfies in particular

$$a(u_h^+ - u, v_h - u_h^+) \geq 0 \qquad \forall\, v_h \in V_{\mathcal{P}}^+. \tag{3.22}$$

Thus, u_h^+ is the best approximation of u in $V_{\mathcal{P}}^+$ with respect to the norm induced by $a(\cdot,\cdot)$, thus proving the quasi-optimality for u_h^+ stated in (3.20).

To prove (3.21) we use the Cauchy-Schwarz inequality and (3.8) to get

$$s(u_h^-, u_h^-) = a(u - u_h^+, u_h^-) \leq \sqrt{\frac{C_{\text{equiv}}}{\alpha}}\, \|u - u_h^+\|_a \sqrt{s(u_h^-, u_h^-)}. \tag{3.23}$$

Alternatively, using that $a(u - u_h^{\text{FEM}}, u_h^-) = 0$, then, we also have

$$s(u_h^-, u_h^-) = a(u_h^{\text{FEM}} - u_h^+, u_h^-) \leq \sqrt{\frac{C_{\text{equiv}}}{\alpha}} \|u_h^{\text{FEM}} - u_h^+\|_a \sqrt{s(u_h^-, u_h^-)}, \tag{3.24}$$

which proves (3.21). □

REMARK 3.8 (Convergence of the complementary part). We may interpret the last result in two ways. First, u_h^- converges to zero at least at the same speed as u_h^+ converges to u. Moreover, (3.21) implies that in certain cases this convergence is much faster than the one for u_h^+. More precisely, focusing on the case of piecewise linear finite element functions, if the mesh satisfies the conditions for the plain Galerkin method to admit discrete maximum principle (see, e.g., Brandts *et al.*, 2008), then $u_h^{\text{FEM}} \in V_{\mathcal{P}}^+$, which implies that $u_h = u_h^{\text{FEM}}$ is also the solution to (3.4). Thus, thanks to (3.21), for certain meshes, and their regular refinements, we have that $u_h^- = 0$.

REMARK 3.9 (Best approximation of the constrained part). In view of the best approximation result (3.20) for the constrained part u_h^+, there is no better finite element function with bound-preserving nodal values in the energy norm. Thus, (3.20) is the counterpart of the classical best approximation result (2.16) for the proposed method.

We finish this section by discussing briefly the notion of numerical solution. Since our main interest is the part of the solution u_h that belongs to $V_{\mathcal{P}}^+$, namely u_h^+, then the latter will be considered to be the numerical solution in the remaining of the manuscript. The 'intermediate' solution u_h appears mostly as a tool to be able to replace the variational inequality by an equality posed over the whole vector space V_h.

4. A problem with nonlinear reaction

To showcase the potential generality of the proposed bound-preserving approach, we now extend (3.4) to a semilinear problem with monotone nonlinearity. Specifically, for $p \in (1, \infty)$ we consider the problem of finding u such that

$$\begin{aligned} -\text{div}\left(\mathcal{D}\,\nabla u\right) + |u|^{p-2}\,u = f \qquad & \text{in } \Omega, \\ u = 0 \qquad & \text{on } \partial\Omega, \end{aligned} \tag{4.1}$$

where $\mathcal{D}$ satisfies the same assumptions as above. To avoid technical diversions, we will only consider $p \geq 2$. This class of equation is sometimes referred to as the Lane–Emden–Fowler equation and is related to problems with critical exponents (Clément *et al.,* 1996). Furthermore, they arise in the theory of boundary layers of viscous fluids (Wong, 1975), among other application areas.

The weak form of this problem is given by: find $u \in \mathcal{X} := H^1_0(\Omega) \cap L^p(\Omega)$ such that

$$a(u, v) + b(u; u, v) = \langle f, v\rangle_\Omega \qquad \forall\, v \in \mathcal{X}, \tag{4.2}$$

where $a(\cdot,\cdot)$ is given by (2.10) (with $\mu = 0$ in this case), and the semilinear form b is given by

$$b(w; u, v) := (|w|^{p-2}\, u, v)_\Omega. \tag{4.3}$$

The space $\mathcal{X}$ is provided with the norm

$$\|v\|_{\mathcal{X}} := |v|_{1,\Omega} + \|v\|_{0,p,\Omega}, \tag{4.4}$$

thus making it a reflexive Banach space. So, using monotone operator theory (see, e.g., Renardy & Rogers, 2004, Chap 10), this problem can be proven to have a unique solution.

The error analysis of this type of problem has been carried out in several works, as early as Glowinski & Marrocco (1974, 1975) (in the context of the p-Laplacian). In there, the estimates are suboptimal for some values of the exponent p. So, later approaches (see, e.g., Barret & Liu, 1993) have made use of the concept of quasinorm in order to obtain optimal error estimates. As this is the approach we will follow in this work, we start recalling the definition of a quasinorm.

Definition 4.1 (Quasinorm). Let V be a real vector space. A *quasinorm* $\|\cdot\|_{(q)}$ in V is a mapping $\|\cdot\|_{(q)} : V \to \mathbb{R}$ that satisfies

$$\|v\|_{(q)} \geq 0\,, \text{and } \ \|v\|_{(q)} = 0 \iff v = 0, \tag{4.5}$$

for all $v \in V$. However, the usual triangle inequality is replaced by

$$\|v + w\|_{(q)} \leq C\big(\|v\|_{(q)} + \|w\|_{(q)}\big), \tag{4.6}$$

for all $v, w \in V$, where C may depend on the definition of $\|\cdot\|_{(q)}$, and the elements v and w themselves.

REMARK 4.2 Strictly speaking, we should also demand that the quasinorm is homogeneous, that is, $\|\alpha v\|_{(q)} = |\alpha| \, \|v\|_{(q)}$ for all $v \in V$ and all $\alpha \in \mathbb{R}$. The mapping we will use to measure the error does not satisfy this last property, but this will not affect the error estimates presented below.

In our analysis below we will make use of the following quasinorm in $L^p(\Omega)$: for a given $w \in L^p(\Omega)$ we define

$$\|v\|_{(w,p)}^2 := \int_\Omega |v|^2(|w| + |v|)^{p-2} \, \mathrm{d}\boldsymbol{x}, \tag{4.7}$$

for all $v \in L^p(\Omega)$. It has the following properties.

LEMMA 4.3 The mapping $\|\cdot\|_{(w,p)}$ is a quasinorm in $L^p(\Omega)$. Moreover, for all $v, w \in L^p(\Omega)$, the following equivalence holds

$$\|v\|_{0,p,\Omega}^p \le \|v\|_{(w,p)}^2 \le \| |v| + |w| \|_{0,p,\Omega}^{p-2} \, \|v\|_{0,p,\Omega}^2 \, . \tag{4.8}$$

Proof. The proof that $\|\cdot\|_{(w,p)}$ is a quasinorm is completely analogous to that of Proposition 2.1 in Ebmeyer & Liu (2005). The equivalence (4.8) is proven following exactly the same steps as in the proof of Equation (10) in Ebmeyer & Liu (2005) (see also Liu, 2000, Proposition 3.1). □

In addition, the following monotonicity and continuity results can be proven for the nonlinear form $b(\cdot;\cdot,\cdot)$.

LEMMA 4.4 The nonlinear form $b(\cdot;\cdot,\cdot)$ is strongly monotone with respect to the quasinorm (4.7). More precisely, there exists a constant $C_C > 0$ such that

$$b(u; u, u - v) - b(v; v, u - v) \ge C_C \, \|u - v\|_{(u,p)}^2. \tag{4.9}$$

Moreover, for any $\theta \in (0, 1]$ the following holds:

$$|b(u; u, w) - b(v; v, w)| \le C_B\big(\theta \, \|u - v\|_{(u,p)}^2 + \theta^{-p+1} \, \|w\|_{(u,p)}^2\big) \qquad \forall \, u, v, w \in L^p(\Omega), \tag{4.10}$$

where C_B depends only on p.

Proof. In Lemma 2.1 of Barret & Liu (1993) the following bounds are proven:

$$\big(|x|^{p-2}x - |y|^{p-2}y\big)\,(x - y) \ge C_2\big(|x| + |y|\big)^{p-2}|x - y|^2, \tag{4.11}$$

$$\big||x|^{p-2}x - |y|^{p-2}y\big| \le C_1\big(|x| + |y|\big)^{p-2}|x - y|, \tag{4.12}$$

for all $x, y \in \mathbb{R}$. Next, (4.9) follows by from (4.11) and noticing that $|x| + |y| \ge \big(|x| + |x - y|\big)/2$. To prove (4.10) we combine (4.12) with the result from Lemma 2.2 in Liu (2000). □

4.1 *The finite element method*

The finite element method we consider is the natural extension of (3.4), that is, find $u_h \in V_{\mathcal{P}}$ such that

$$a\big(u_h^+, v_h\big) + b\big(u_h^+; u_h^+, v_h\big) + s\big(u_h^-, v_h\big) = \langle f, v_h\rangle_\Omega \qquad \forall\, v_h \in V_{\mathcal{P}}. \tag{4.13}$$

Here, the stabilisation term is given by

$$s(v_h, w_h) := \alpha \sum_{i=1}^{N} \|\mathcal{D}\|_{0,\infty,\omega_i} \mathfrak{h}(\boldsymbol{x}_i)^{d-2}\, v_h(\boldsymbol{x}_i) w_h(\boldsymbol{x}_i), \tag{4.14}$$

where, once again, $\alpha > 0$ is an dimensionless constant. Following very similar steps to those from the proof of Theorem 3.2, Method (4.13) can be proven to have a unique solution $u_h \in V_{\mathcal{P}}$. We note that there appears to be no traceable numerical advantage in including a linearised reaction term in the stabilisation (4.14), at least for modestly large values of p, so we omit it for simplicity.

As in the linear case, (4.13) can be linked to a variational inequality, as the following result (whose proof is totally analogous to that of Theorem 3.5) shows.

THEOREM 4.5 Let u_h be the unique solution of (4.13). Then, u_h^+ satisfies the variational inequality

$$a\big(u_h^+, v_h - u_h^+\big) + b\big(u_h^+; u_h^+, v_h - u_h^+\big) \geq \big\langle f, v_h - u_h^+\big\rangle_\Omega \qquad \forall\, v_h \in V_{\mathcal{P}}^+. \tag{4.15}$$

The following result is the main reason for the use of a quasinorm instead of the norm induced by the problem. In fact, starting from the last result a Céa type estimate can be obtained, but, analogously to what is reported in Ciarlet (2002), that would lead to a suboptimal estimate for certain values of p. So, in the next result we provide an optimal estimate with respect to the quasinorm (4.7).

THEOREM 4.6 Let u solve (4.1) and $u_h \in V_{\mathcal{P}}$ be the solution of (4.13). Then,

$$\|u - u_h^+\|_a^2 + \|u - u_h^+\|_{(u,p)}^2 \leq C \inf_{v_h \in V_{\mathcal{P}}^+} \big(\|u - v_h\|_a^2 + \|u - v_h\|_{(u,p)}^2\big). \tag{4.16}$$

Proof. Using the coercivity (4.9) of b followed by Theorem 4.5, we get

$$\begin{aligned} C_C\Big(\big\|u - u_h^+\big\|_a^2 + \big\|u - u_h^+\big\|_{(u,p)}^2\Big) &\leq a\big(u - u_h^+, u - u_h^+\big) + b\big(u; u, u - u_h^+\big) - b\big(u_h^+; u_h^+, u - u_h^+\big) \\ &\leq a\big(u - u_h^+, u - v_h\big) + b\big(u; u, u - v_h\big) - b\big(u_h^+; u_h^+, u - v_h\big), \end{aligned} \tag{4.17}$$

for every $v_h \in V_{\mathcal{P}}^+$. Next, using Young's inequality and (4.10) we have, for all $\theta \in (0, 1]$,

$$\begin{aligned} C_C\Big(\big\|u - u_h^+\big\|_a^2 + \big\|u - u_h^+\big\|_{(u,p)}^2\Big) &\leq \frac{C_C}{4}\big\|u - u_h^+\big\|_a^2 + C_C\big\|u - v_h\big\|_a^2 \\ &\quad + C_B\Big(\theta\big\|u - u_h^+\big\|_{(u,p)}^2 + \theta^{-p+1}\big\|u - v_h\big\|_{(u,p)}^2\Big). \end{aligned} \tag{4.18}$$

Choosing $\theta = \min\Big(\frac{C_C}{4C_B}, \frac{1}{2}\Big)$ and rearranging the inequality yields the desired result. □

5. Numerical tests

In the following tests we detail some aspects of our implementation and showcase the methodology looking at the symmetric problem (2.7) and its nonlinear counterpart (4.1) in two space dimensions. We recall once again that all references to the numerical solution refer to the function $u_h^+ \in V_{\mathcal{P}}^+$, and not the function u_h.

To linearise the problem, we pose the following Richardson-like iterative approximation for (3.4): Given u^0 and $\omega \in (0, 1]$, for each $n = 0, 1, \ldots$ find u^{n+1} such that

$$a(u^{n+1}, v) = a(u^n, v) + \omega\big(\langle f, v\rangle_\Omega - a\big(\left(u^n\right)^+, v\big) - s\big(\left(u^n\right)^-, v\big)\big). \tag{5.1}$$

We initialise the finite element approximation of (5.1) by the Galerkin approximation, that is we set $u_h^0 \in V_{\mathcal{P}}$ such that, for all $v_h \in V_{\mathcal{P}}$

$$a(u_h^0, v_h) = \langle f, v_h\rangle_\Omega. \tag{5.2}$$

Then, the approximation to the iteration (5.1) becomes for each $n = 0, \ldots, N-1$, find $u_h^{n+1} \in V_{\mathcal{P}}$ such that for all $v_h \in V_{\mathcal{P}}$

$$a(u_h^{n+1}, v_h) = a(u_h^n, v_h) + \omega\big(\langle f, v_h\rangle_\Omega - a\big(\left(u_h^n\right)^+, v_h\big) - s\big(\left(u_h^n\right)^-, v_h\big)\big). \tag{5.3}$$

In each experiment we take $\alpha = 1$ within the stabilisation. The linear systems arising in (5.3) are solved using an LU decomposition within the Eigen library. The linearisation was terminated when $\left\| u_h^{n+1} - u_h^n \right\|_{0,\Omega} \le 10^{-12}$.

5.1 *Convergence on a regular grid with a smooth solution*

We first consider $\mathcal{D} = \epsilon\mathcal{I}$, where $\mathcal{I}$ denotes the 2×2 identity matrix, with $\epsilon = 10^{-5}$, $\mu = 1$, and set f such that the function

$$u(x, y) = \sin(\pi x)\sin(\pi y) \tag{5.4}$$

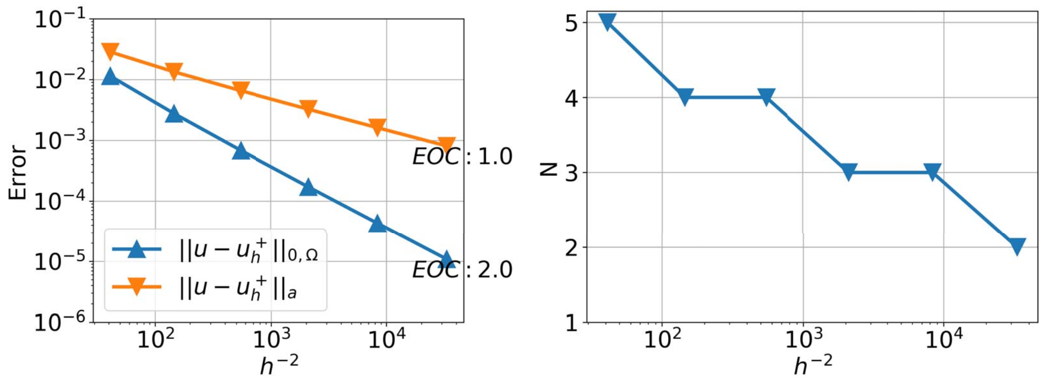
solves the problem (2.7) over the unit square.

Convergence results for piecewise linear, $k = 1$, and quadratic, $k = 2$, elements over a sequence of uniformly refined criss-cross meshes are shown in Figures 1 and 2, respectively. In line with the error estimates from Section 3.2, the method converges with optimal rate in the function approximation sense for both $k = 1$ and $k = 2$. Notice also that the number of iterations of the linearisation decreases as a function of h. Here, and thereafter, EOC stands for *estimated order of convergence*.

5.2 *Convergence on an obtuse grid with a smooth solution*

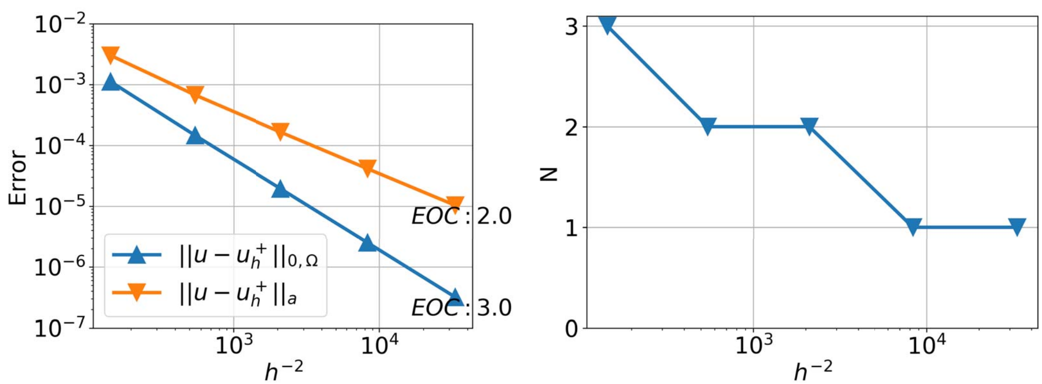
We again consider $\mathcal{D} = \epsilon\mathcal{I}$, with $\epsilon = 10^{-5}$, $\mu = 1$. The analytical solution of (2.7) over $(-1, 1) \times (0, 1)$ is taken as

$$u(x, y) = \sin(\pi(x + 1)/2)\sin(\pi y), \tag{5.5}$$



(A) Convergence in the L^2 and energy norms. (B) Number of iterations to achieve convergence as a function of the meshsize.

FIG. 1. We test the convergence of a piecewise linear conforming approximation given by (5.3) on a sequence of concurrently refined criss-cross meshes. The Richardson iteration converges for $\omega = 1$.

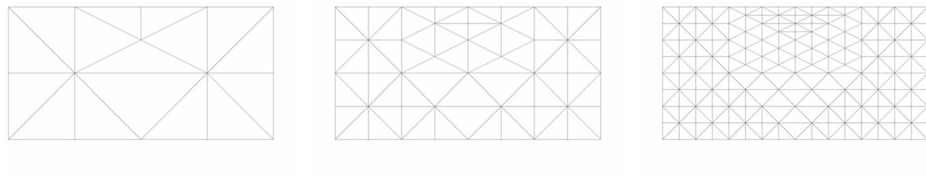


(A) Convergence in the L^2 and energy norms. (B) Number of iterations to achieve convergence as a function of the meshsize.

FIG. 2. We test convergence of a piecewise quadratic conforming approximation (5.3) on a sequence of concurrently refined criss-cross meshes. The Richardson iteration converges for $\omega = 1$.

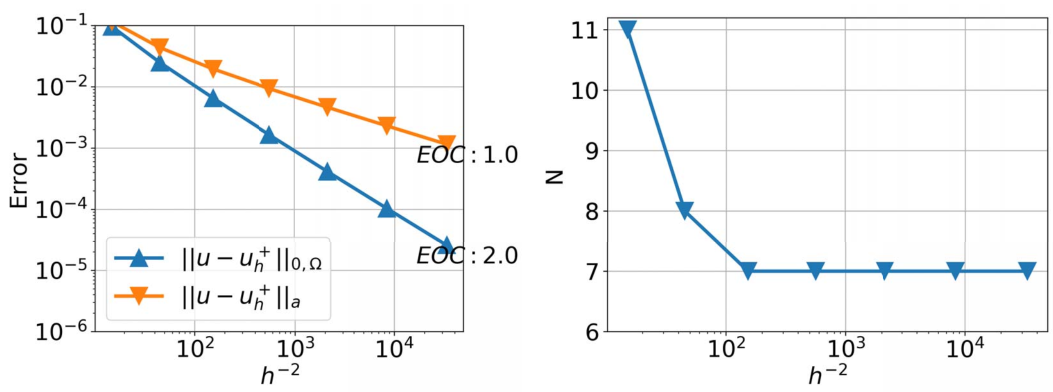
and we compute f accordingly. Notice that $u(\boldsymbol{x}) \in [0, 1]$ for all $\boldsymbol{x} \in \Omega$. We pose the problem over a sequence of triangulations with obtuse elements as described in Brandts *et al.* (2009) illustrated in Figure 3. This mesh was used in Brandts *et al.* (2009) as an example of triangulations for which the finite element method does not satisfy the discrete maximum principle, even for the Poisson equation; so, it poses a challenge to the finite element method as solutions do not in general satisfy DMP even if $\epsilon \gg 1$.

Convergence results for piecewise linear elements over this mesh are given in Figure 4. The method converges optimally and similar results are observed for higher order elements. Notice that the linearisation takes more iterations to achieve convergence, which is to be expected, as the Galerkin



(A) Iteration 1. (B) Iteration 2. (C) Iteration 3.

FIG. 3. Three refinements of the mesh with a single obtuse triangle from Brandts *et al.* (2009). The result is a layer of obtuse triangles.



(A) Convergence in the L^2 and energy norms. (B) Number of iterations to achieve convergence as a function of the meshsize.

FIG. 4. We test convergence of a piecewise linear conforming approximation (5.3) on a sequence of concurrently refined meshes with obtuse elements as illustrated in Figure 3. Here the Richardson linearisation converges for $\omega = 1$.

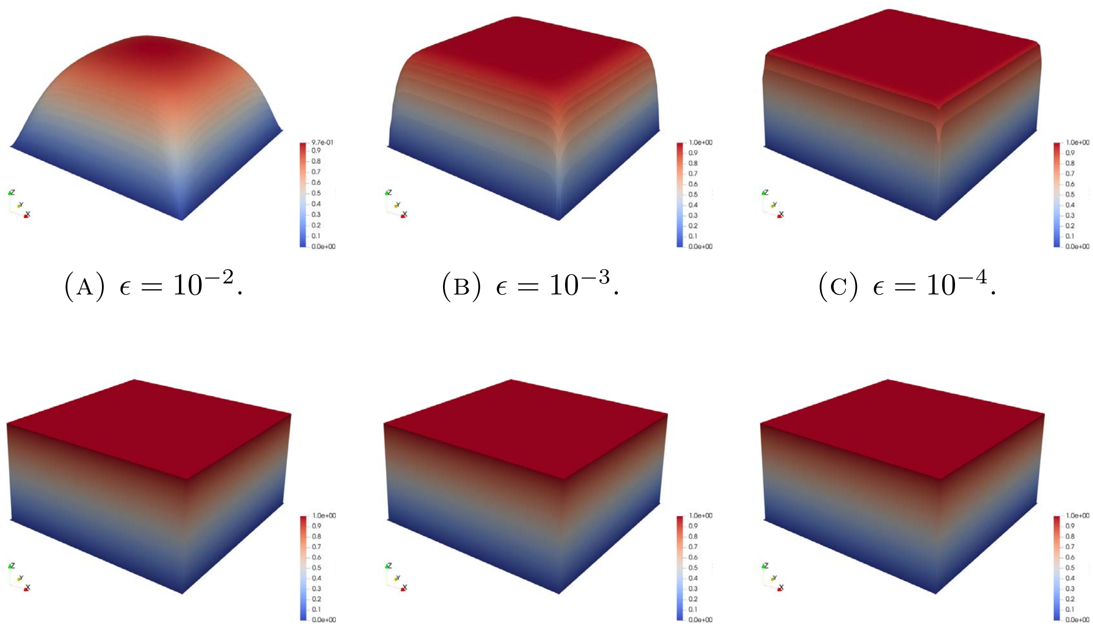
solution will, very likely, never respect the bounds for the problem, regardless of how fine the mesh is (the iteration count remains, nevertheless, low).

5.3 *Resolution of boundary layers*

Consider the problem

$$\begin{aligned} -\epsilon\Delta u + u &= 1 \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{5.6}$$

We fix $h \approx 0.02$ on a criss-cross and vary $\epsilon \in [10^{-2}, 10^{-7}]$. For particularly small ϵ the Richardson iteration required dampening for convergence. With $\epsilon > 10^{-5}$, we use $\omega = 1$ and convergence was achieved within 4 iterations. When $\epsilon \leq 10^{-5}$, $\omega = \frac{1}{2}$ is sufficient for convergence with fewer than 46 iterations in each case. The most challenging case being the smallest value of ϵ. Computed solutions for different values of ϵ are shown in Figure 5.



(A) $\epsilon = 10^{-2}$. (B) $\epsilon = 10^{-3}$. (C) $\epsilon = 10^{-4}$.

(D) $\epsilon = 10^{-5}$. (E) $\epsilon = 10^{-6}$. (F) $\epsilon = 10^{-7}$.

FIG. 5. Elevations of the approximation to (5.6) for fixed h and different values of ϵ. We notice the absence of oscillations even for particularly small values of ϵ.

5.4 *Resolution of boundary layers with discontinuous Dirichlet conditions*

Consider the problem

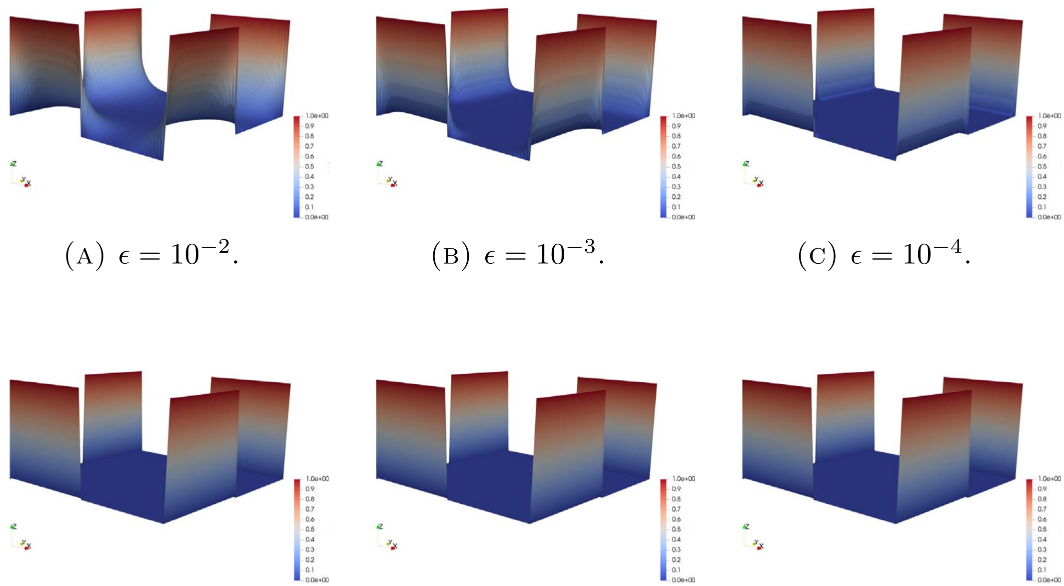
$$\begin{aligned} -\epsilon \Delta u + u &= 0 \text{ in } \Omega, \\ u &= g_D \text{ on } \partial\Omega, \end{aligned} \tag{5.7}$$

where $g_D = 1$ on $[0, \frac{1}{2}] \times \{0\}$, $g_D = 0$ on $(\frac{1}{2}, 1] \times \{0\}$ and periodically follows the same pattern counter-clockwise. We fix $h \approx 0.02$ on a criss-cross mesh and vary $\epsilon \in [10^{-2}, 10^{-7}]$. For particularly small ϵ the Richardson iteration required dampening for convergence. With $\epsilon > 10^{-5}$ we used $\omega = 1$ and convergence was achieved within 5 iterations. When $\epsilon \leq 10^{-5}$, then $\omega = 0.5$ provided a convergent algorithm and took fewer than 40 iterations in each case. Figure 6 shows some solutions. Once again, we notice the lack of oscillations in the discrete solution for the whole range of reaction and diffusion coefficients.

5.5 *A solution with an interior layer*

Consider the problem

$$\begin{aligned} -\epsilon \Delta u + u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{5.8}$$



(A) $\epsilon = 10^{-2}$. (B) $\epsilon = 10^{-3}$. (C) $\epsilon = 10^{-4}$.

(D) $\epsilon = 10^{-5}$. (E) $\epsilon = 10^{-6}$. (F) $\epsilon = 10^{-7}$.

FIG. 6. Elevations of the approximation to (5.7) for fixed h and varying ϵ. We notice there are no apparent oscillations even for particularly small values of ϵ.

with

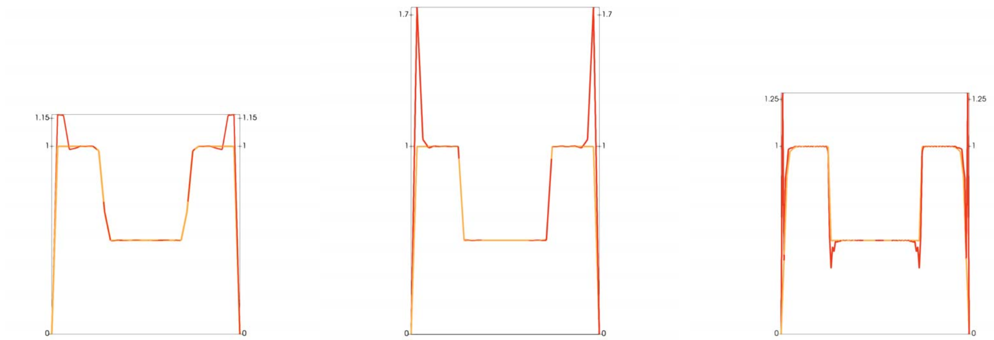
$$f = \begin{cases} \frac{1}{2} \text{ in } \left[\frac{1}{4}, \frac{3}{4}\right]^2 \\ 1 \text{ otherwise.} \end{cases} \tag{5.9}$$

In this case the solution is expected to achieve a local minimum in the interior. We fix $h \approx 0.02$ and examine the solution for $\epsilon = \{10^{-4}, 10^{-7}\}$. We compare the standard finite element solution and the approximation given by (5.3) in Figure 7. Notice that the plain Galerkin solution has oscillations near the boundary layer that become extreme for $\epsilon \ll 1$, which are totally removed by the current method. In addition, for $k = 2$, there are noticeable undershoots around the interior layer, which are totally removed by the current method.

5.6 *Anisotropic diffusion with nonlinear reaction*

Consider the domain $\Omega = \Omega_1 \backslash \Omega_2$ with $\Omega_1 = (0, 1)^2$, $\Omega_2 = [\frac{4}{9}, \frac{5}{9}]^2$ and the problem

$$\begin{aligned} -\text{div}(\epsilon \mathcal{D} \nabla u) + u^3 &= f \text{ in } \Omega, \\ u &= \begin{cases} 0 \text{ on } \partial\Omega_1 \\ 2 \text{ on } \partial\Omega_2 \end{cases}, \end{aligned} \tag{5.10}$$



(A) $\epsilon = 10^{-4}$, $k = 1$. (B) $\epsilon = 10^{-7}$, $k = 1$. (C) $\epsilon = 10^{-7}$, $k = 2$.

FIG. 7. A cross section taken about the $x = y$ plane of the finite element solution, coloured red presenting oscillations at the boundary layer, and the approximation given by (5.3), coloured green with no oscillations at the boundary. Panels (A) and (B) are computed using $k = 1$, while (C) depicts the approximations obtained using $k = 2$.

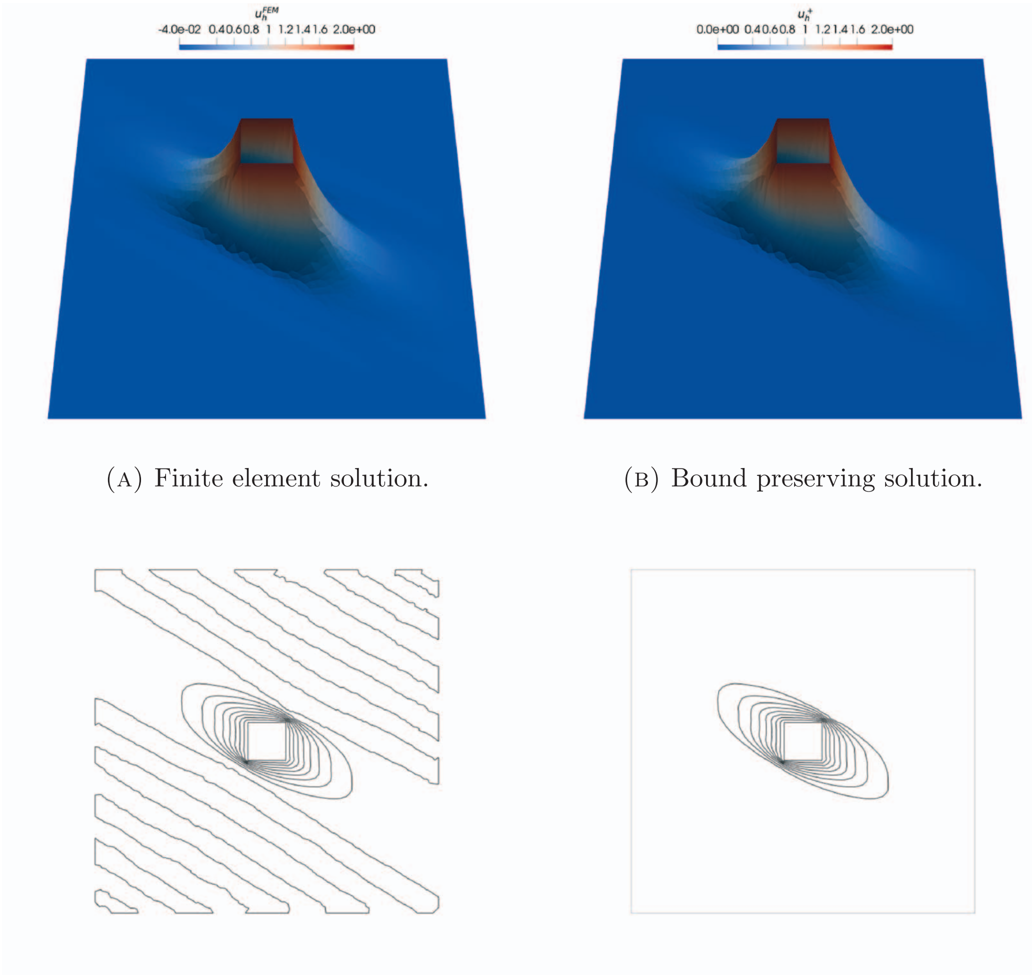
with $\epsilon = 10^{-5}$,

$$\mathcal{D} = \mathcal{R}\mathcal{A}\mathcal{R}^T, \quad \mathcal{R} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix}, \tag{5.11}$$

and $\theta = -\frac{\pi}{6}$. This is a challenging realisation of (4.1) with an anisotropic diffusion coefficient and nonlinear reaction term, with $p = 4$, posed over a nonconvex domain. In Figure 8 we show the finite element solution, the bound-preserving solution and contour plots highlighting the oscillatory nature of the standard finite element solution for this problem.

6. Concluding remarks

We proposed an inexpensive and simple way to impose hard bounds on the range of a finite element solution. This is achieved through the definition of a nonlinear stabilised Galerkin approach, which is designed to provide the orthogonal projection into the closed convex set of physically admissible solutions satisfying hard bounds. In an effort to highlight the key ideas, we have confined the presentation to linear and monotone semilinear reaction-diffusion equations. We stress, however, that the framework is general enough to allow other model problems, as well as bound-preserving variants of other known finite element methods, to be constructed following the methodology presented here. In particular, an interesting extension of the proposed methodology to convection-dominated problems is both relevant and, we believe, within reach. Moreover, the extension to discontinuous Galerkin methods posed on general polygonal/polyhedral meshes is also conceivable. Both these, and other, topics are ongoing, and will be discussed elsewhere.



(A) Finite element solution.

(B) Bound preserving solution.

(C) Contour plot of the finite element solution.

(D) Contour plot of the bound preserving solution.

FIG. 8. Approximations to the problem (5.10) with $h \approx 0.02$ over a Delaunay mesh. Notice the finite element approximation exhibits oscillatory behaviour in contrast to the bound preserving method.

Funding

The work of G.R.B. has been funded by the Leverhulme Trust through the Research Fellowship No. RF-2019-510. The work of A.V. is supported by the Italian GNCS and the MIUR PRIN 2017 NA-FROM-PDEs. The research work of EG was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant (Project Number: 3270). Also, EG wishes to acknowledge the financial support of EPSRC (grant number EP/W005840/1). TP is grateful for partial support from EPSRC (EP/X017206/1, EP/X030067/1 and EP/W026899/1).

References

BARRENECHEA, G. R., BURMAN, E. & KARAKATSANI, F. (2017a) Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. *Numer. Math.*, **135**, 521–545.

BARRENECHEA, G. R., JOHN, V. & KNOBLOCH, P. (2017b) An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Math. Models Methods Appl. Sci.*, **27**, 525–548.

BARRENECHEA, G. R., JOHN, V. & KNOBLOCH, P. (2023) Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. *SIAM Review*, to appear.

BARRETT, J. W. & LIU, W. (1993) Finite element approximation of the p-Laplacian. *Math. Comput.*, **61**, 523–537.

BOCHEV, P., RIDZAL, D., D'ELIA, M., PEREGO, M. & PETERSON, K. (2020) Optimization-based, property-preserving finite element methods for scalar advection equations and their connection to algebraic flux correction. *Comp. Methods Appl. Mech. Eng.*, **367**, 112982.

BRANDTS, J. H., KOROTOV, S. & KŘÍŽEK, M. (2008) The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem. *Linear Algebra Appl.*, **429**, 2344–2357.

BRANDTS, J., KOROTOV, S., KŘÍŽEK, M. & ŠOLC, J. (2009) On nonobtuse simplicial partitions. *SIAM Rev.*, **51**, 317–335.

BURMAN, E. & ERN, A. (2005) Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. *Math. Comp.*, **74**, 1637–1652 (electronic).

BURMAN, E. & ERN, A. (2017) A nonlinear consistent penalty method weakly enforcing positivity in the finite element approximation of the transport equation. *Comp. Methods Appl. Mech. Eng.*, **320**, 122–132.

CIARLET, P. G. (2002) The Finite Element Method for Elliptic Problems. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

CIARLET, P. G. (1970) Discrete maximum principle for finite-difference operators. *Aequationes Math.*, **4**, 338–352.

CIARLET, P. G. & RAVIART, P.-A. (1973) Maximum principle and uniform convergence for the finite element method. *Comput. Methods Appl. Mech. Eng.*, **2**, 17–31.

CLÉMENT, P., DE FIGUEIREDO, D. G. & MITIDIERI, E. (1996) Quasilinear elliptic equations with critical exponents. *Topol. Methods Nonlinear Anal.*, **7**, 133–170.

EBMEYER, C. & LIU, W. (2005) Quasi-norm interpolation error estimates for the piecewise linear finite element approximation of p-Laplacian problems. *Numerische Mathematik*, **100**, 233–258.

ERN, A. & GUERMOND, J.-L. (2021) *Finite Elements I. Approximation and Interpolation*. Volume 72 of Texts in Applied Mathematics. Springer-Verlag.

EVANS, J. A., HUGHES, T. J. & SANGALLI, G. (2009) Enforcement of constraints and maximum principles in the variational multiscale method. *Comp. Methods Appl. Mechanics Eng.*, **199**, 61–76.

GILBARG, D. & TRUDINGER, N. S. (2001) *Elliptic Partial Differential Equations of Second Order*. Springer. Reprint of the 1998 edition.

GLOWINSKI, R. & MARROCCO, A. (1974) Sur l'approximation, par éléments finis d'ordre 1, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires. *C. R. Acad. Sci., Ser. A*, **278**, 1649–1652.

GLOWINSKI, R. & MARROCCO, A. (1975) Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires. *R.A.I.R.O.*, **9**, 41–76.

HUANG, W. (2011) Discrete maximum principle and a Delaunay-type mesh condition for linear finite element approximations of two-dimensional anisotropic diffusion problems. *Numer. Math. Theory Methods Appl.*, **4**, 319–334.

JENSEN, M. & MALQUIST, A. (2013) Finite element convergence for the Joule heating problem with mixed boundary conditions. *BIT Numer. Math.*, **53**, 475–496.

KIKUCHI, F. (1977) Discrete maximum principle and artificial viscosity in finite element approximations to convective diffusion equations. *Institute of Space and Aeronautical Science*. University of Tokyo, 550.

KREUZER, C. (2014) A note on why enforcing discrete maximum principles by a simple a posteriori cutoff is a good idea. *Numer. Methods Partial Differential Equations*, **30**, 994–1002.

LISA, R. & SHASHKOV, M. (2008) Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems. *Commun. Comput. Phys.*, **3**, 852–877.

LIU, W. (2000) Finite element approximation of a nonlinear elliptic equation arising from bimaterial problems in elastic-plastic mechanics. *Numerische Mathematik*, **86**, 491–506.

LU, C., HUANG, W. & VLECK, E. S. V. (2013) The cutoff method for the numerical computation of nonnegative solutions of parabolic pdes with application to anisotropic diffusion and lubrication-type equations. *J. Comput. Phys.*, **242**, 24–36.

MIZUKAMI, A. & HUGHES, T. J. R. (1985) 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**, 181–193.

MUDUNURU, M. & NAKSHATRALA, K. (2016) On enforcing maximum principles and achieving element–wise species balance for advection–diffusion–reaction equations under the finite element method. *Journal of Computational Physics*, **305**, 448–493.

RENARDY, M. & ROGERS, R. C. (2004) An introduction to partial differential equations. *volume 13 of Texts in Applied Mathematics*, second edn. New York: Springer.

ROOS, H.-G., STYNES, M. & TOBISKA, L. (2008) *Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems*, 2nd edn. Berlin: Springer.

WONG, J. S. (1975) On the generalized Emden–Fowler equation. *SIAM Rev.*, **17**, 339–360.

YANG, J., YUAN, Z. & ZHOU, Z. (2022) Arbitrarily high–order maximum bound preserving schemes with cut-off postprocessing for Allen–Cahn equations. *J. Sci. Comp.*, **90**, 76.