Appl. Numer. Math. 210 (2025) 222–244

Contents lists available at ScienceDirect

Applied Numerical Mathematics

journal homepage: www.elsevier.com/locate/apnum

Research Paper

Discrete maximum principles with computable mesh conditions for nonlinear elliptic finite element problems

M.T. Bahlibi [a,b], J. Karátson [c,d,*], S. Korotov [e]

[a] Department of Applied Analysis and Computational Mathematics, Eötvös Loránd University, Hungary
[b] Department of Mathematics, Mai Nefhi College of Science, National Higher Education and Research Institute, Asmara, Eritrea
[c] Department of Applied Analysis and Computational Mathematics & HUN-REN–ELTE Numerical Analysis and Large Networks Research Group, Eötvös Loránd University, Hungary
[d] Department of Analysis and Operations Research, Budapest University of Technology and Economics, Hungary
[e] Division of Mathematics and Physics, UKK, Mälardalen University, Västerås, Sweden

ARTICLE INFO

Keywords:
Nonlinear elliptic PDE
Finite elements
Discrete maximum principles

ABSTRACT

Discrete maximum principles are essential measures of the qualitative reliability of the given numerical method, therefore they have been in the focus of intense research, including nonlinear elliptic boundary value problems describing stationary states in many nonlinear processes. In this paper we consider a general class of nonlinear elliptic problems which covers various special cases and applications. We provide exactly computable conditions on the geometric characteristics of widely studied finite element shapes: triangles, tetrahedra, prisms and rectangles, and guarantee the validity of discrete maximum principles under these conditions.

1. Introduction

The validity of maximum principles is an important property of second-order elliptic equations with various boundary conditions [25,26]. Generally speaking, it provides a priory known bounds of the unknown solution, provided some conditions on the given data are imposed. Such maximum principles are now known for many nonlinear elliptic problems appearing in a number of important applications in physics and engineering.

Therefore, obviously, their suitable discrete analogues, commonly called discrete maximum principles, or DMPs, have drawn much attention. A DMP is a relevant measure of the qualitative reliability of the given numerical method, since otherwise one might get physically incorrect numerical solutions, such as negative concentrations in some parts of the space domain. The validity of DMPs for the most widespread numerical techniques, like finite difference and finite element methods, has thus also been in the focus of research work for the recent decades. Starting with the pioneering works by Ciarlet [5], a lot of generalizations (involving various nonlinearities in the equations, mixed boundary conditions, usage of quadrature rules, usage of different types of finite elements) have been addressed. For the finite element method (FEM), which is in the focus of the present paper, various results on DMPs under proper geometric conditions on the computational meshes have been given e.g. in [4,6,27,28] in the case of linear elliptic problems. Such ideas were also relevant for linear parabolic problems [9,10]. We have established a DMP for general nonlinear elliptic problems in [18], which has been applied and extended in various later works, see, e.g. [15,24] and also the authors' previous results, e.g. [16,19,20].

* Corresponding author.
E-mail addresses: menghis.teweldebrhan@gmail.com (M.T. Bahlibi), kajkaat@caesar.elte.hu (J. Karátson), smkorotov@gmail.com (S. Korotov).

https://doi.org/10.1016/j.apnum.2024.12.009
Received 5 February 2024; Received in revised form 2 December 2024; Accepted 19 December 2024

Available online 2 January 2025

In all the mentioned papers, the DMPs for nonlinear problems with lower order terms (typically arising from reaction type processes) have been derived under the condition that the mesh parameter is sufficiently small. However, no exact condition on the required mesh size has been given. The goal of the present paper is to fill this gap from the point of view of concrete applications. We provide exactly computable conditions on the geometric characteristics of widely studied finite element shapes: triangles, tetrahedra, prisms and rectangles, and guarantee the validity of discrete maximum principles under these conditions. The considered class of nonlinear elliptic problem is as general as possible within this scope, slightly extending the one used in [18], and also various special cases and applications are shown which are covered by our setting. Finally, we illustrate numerically that the restrictions on the mesh size are indeed necessary in practice.

2. The PDE problem and its FEM discretization

2.1. The general mathematical model and its properties

Let us consider the following nonlinear elliptic boundary value problem:

$$\begin{cases} -\operatorname{div}\Big(b(x,u,\nabla u)\,\nabla u\Big) + r(x,u,\nabla u)u = f(x) & \text{in } \Omega, \\ b(x,u,\nabla u)\frac{\partial u}{\partial \nu} = \gamma(x) & \text{on } \Gamma_N, \\ u = g(x) & \text{on } \Gamma_D, \end{cases} \tag{2.1}$$

where Ω is a bounded domain in $\mathbb{R}^d$ ($d = 2$ or 3) under the assumptions below:

Assumption 2.1.

(a) Ω has a piecewise smooth and Lipschitz continuous boundary $\partial\Omega$; $\Gamma_N, \Gamma_D \subset \partial\Omega$ are measurable open sets, such that $\Gamma_N \cap \Gamma_D = \emptyset$ and $\overline{\Gamma}_N \cup \overline{\Gamma}_D = \partial\Omega$, further $meas(\Gamma_D) > 0$.
(b) The scalar functions $b : \overline{\Omega}\times\mathbb{R}\times\mathbb{R}^d \to \mathbb{R}$ and $r : \overline{\Omega}\times\mathbb{R}\times\mathbb{R}^d \to \mathbb{R}$ are continuous. Further, $f \in L^2(\Omega)$, $\gamma \in L^2(\Gamma_N)$ and $g = g^*{}_{|\Gamma_D}$ for some $g^* \in H^1(\Omega)$.
(c) The functions b and r are bounded such that

$$0 < \mu_0 \le b(x,\xi,\eta) \le \mu_1, \quad 0 \le r(x,\xi,\eta) \le \beta \qquad \forall (x,\xi,\eta) \in \overline{\Omega}\times\mathbb{R}\times\mathbb{R}^d, \tag{2.2}$$

where μ_0, μ_1 and β are positive constants. The weak formulation of (2.1) is defined as follows: find $u \in H^1(\Omega)$ such that

$$u = g \quad \text{on } \Gamma_D \text{ in trace sense, and} \tag{2.3}$$

$$\int_\Omega \Big(b(x,u,\nabla u)\,\nabla u\cdot\nabla v + r(x,u,\nabla u)uv\Big)\,dx = \int_\Omega f v\,dx + \int_{\Gamma_N} \gamma v\,d\sigma \qquad \forall v \in H^1_D(\Omega), \tag{2.4}$$

where $H^1_D(\Omega) := \{u \in H^1(\Omega) : u|_{\Gamma_D} = 0 \text{ in trace sense}\}$ [1] with the following standard inner product and corresponding norm, respectively:

$$\langle u, v\rangle_1 = \int_\Omega \nabla u\cdot\nabla v\,dx, \qquad |v|_1 = \Big(\int_\Omega |\nabla v|^2\,dx\Big)^{\frac{1}{2}}. \tag{2.5}$$

Remark 2.1. The solvability of the problem can be ensured in two subclasses of (2.1).

(a) The existence and uniqueness of the weak solution can be proved for the following special case of (2.1) that leads to a uniformly monotone operator, see, e.g., [18,30]:

$$\begin{cases} -\operatorname{div}\Big(b(x,\nabla u)\,\nabla u\Big) + r(x,u)u = f(x) & \text{in } \Omega, \\ b(x,\nabla u)\frac{\partial u}{\partial \nu} = \gamma(x) & \text{on } \Gamma_N, \\ u = g(x) & \text{on } \Gamma_D, \end{cases} \tag{2.6}$$

under the following conditions:
(i) The mapping $f : \Omega\times\mathbb{R}^d \to \mathbb{R}^d$, defined as $f(x,\eta) := b(x,\eta)\eta$, is uniformly monotone and Lipschitz continuous with respect to η.
(ii) The scalar function $\xi \to r(x,\xi)\xi =: q(x,\xi)$ is monotone increasing and Lipschitz continuous with respect to ξ.
Some reaction-diffusion type equations are examples of this type of model, see subsections 2.2.1-2.2.3 below.

(b) The existence of a weak solution can be ensured for the following special case of (2.1):

$$\begin{cases} -\operatorname{div}\left(b(x,u)\nabla u\right) + r(x,u)u = f(x) & \text{in } \Omega, \\ b(x,u)\frac{\partial u}{\partial \nu} = \gamma(x) & \text{on } \Gamma_N, \\ u = g(x) & \text{on } \Gamma_D, \end{cases} \tag{2.7}$$

provided that b and r are Lipschitz continuous, $b \geq \mu_0 > 0$ and $r \geq 0$, see [17]. An example of this model is stationary heat conduction, see subsection 2.2.4 below.

A motivating property for the present paper is that problem (2.1) satisfies a maximum-minimum principle. First we formulate the maximum principle (sometimes called 'continuous maximum principle', CMP, in contrast to the later studied discrete version, DMP). The formulation and proof is similar to Theorem 5 in [18], now our problem is slightly more general.

Theorem 2.1. *Let Assumption 2.1 hold and the weak solution of (2.1) belong to $C^1(\Omega) \cap C(\bar{\Omega})$. If*

$$f(x) \leq 0 \quad (x \in \Omega) \quad \textit{and} \quad \gamma(x) \leq 0 \quad (x \in \Gamma_N), \tag{2.8}$$

then

$$\max_{\overline{\Omega}} u \leq \max\{0, \max_{\Gamma_D} g\}. \tag{2.9}$$

In particular, if $\max_{\Gamma_D} g \geq 0$ then

$$\max_{\overline{\Omega}} u = \max_{\Gamma_D} g, \tag{2.10}$$

and if $g \leq 0$ then we have the nonpositivity property

$$u \leq 0 \quad \text{on } \overline{\Omega}. \tag{2.11}$$

Proof. We follow [18]. First, let us define the functions $\tilde{a}(x) := b(x, u(x), \nabla u(x))$ and $\tilde{h}(x) := r(x, u(x), \nabla u(x))$ for $x \in \bar{\Omega}$. Then $\tilde{a} \in C(\Omega)$ and satisfies the uniform positivity $\mu_0 \leq \tilde{a} \leq \mu_1$. Besides, $\tilde{h} \in C(\bar{\Omega})$ and $\tilde{h} \geq 0$. Hence, the operator

$$\tilde{L}v \equiv -div(\tilde{a}(x)\nabla v) + \tilde{h}(x)v$$

satisfies the standard properties required in equation (2.1) of [18], and $\tilde{L}u = f$ coincides with our original PDE, from which the statements (2.8)-(2.10) follow exactly as in Theorem 5 therein. □

The corresponding continuous minimum principle for the problem (2.1) can be verified in the same way by reversing signs.

Theorem 2.2. *Let Assumption 2.1 hold and the weak solution of (2.1) belong to $C^1(\Omega) \cap C(\bar{\Omega})$. If*

$$f(x) \geq 0 \quad (x \in \Omega) \qquad \textit{and} \quad \gamma(x) \geq 0 \quad (x \in \Gamma_N), \tag{2.12}$$

then

$$\min_{\overline{\Omega}} u \geq \min\{0, \min_{\Gamma_D} g\}. \tag{2.13}$$

In particular, if $\min_{\Gamma_D} g \leq 0$ then

$$\min_{\overline{\Omega}} u = \min_{\Gamma_D} g, \tag{2.14}$$

and if $g \geq 0$ then we have nonnegativity property

$$u \geq 0 \quad \text{on } \overline{\Omega}. \tag{2.15}$$

2.2. Some examples and applications

2.2.1. Semilinear problems

A practically widely important special case of problem (2.1) is as follows, arising e.g. in various reaction-diffusion problems:

$$\begin{cases} -\operatorname{div}\big(b(x)\nabla u\big) + q(x,u) = f(x) & \text{in } \Omega, \\ b(x)\frac{\partial u}{\partial \nu} = \gamma(x) & \text{on } \Gamma_N, \\ u = g(x) & \text{on } \Gamma_D, \end{cases} \tag{2.16}$$

where $q \in C^1(\Omega \times \mathbb{R})$. We assume that there exists $\beta > 0$ such that

$$0 \le \frac{\partial q}{\partial \xi}(x,\xi) \le \beta \qquad \text{and} \qquad 0 < \mu_0 \le b(x) \le \mu_1, \qquad \forall (x,\xi) \in \Omega \times \mathbb{R}. \tag{2.17}$$

Then q itself is Lipschitz continuous w.r.t ξ with constant β, hence Remark 2.1 (a) implies the existence and uniqueness of the weak solution. Now let us first define a function r in terms of q:

$$r(x,\xi) := \begin{cases} \frac{q(x,\xi)-q(x,0)}{\xi}, & \text{if } \xi \neq 0, \\ \frac{\partial q}{\partial \xi}(x,0), & \text{if } \xi = 0, \end{cases} \tag{2.18}$$

then clearly r is continuous and $0 \le r(x,\xi) \le \beta$, further,

$$r(x,\xi)\xi = q(x,\xi) - q(x,0), \qquad \forall \xi \in \mathbb{R}. \tag{2.19}$$

The PDE in (2.16) can be written as:

$$-\operatorname{div}\big(b(x)\nabla u\big) + q(x,u) - q(x,0) = f(x) - q(x,0) \quad \text{in } \Omega, \tag{2.20}$$

that is,

$$-\operatorname{div}\big(b(x)\nabla u\big) + r(x,u)u = \tilde{f}(x) \quad \text{in } \Omega, \tag{2.21}$$

where $r(x,u)u = q(x,u) - q(x,0)$ from (2.19) and $\tilde{f}(x) = f(x) - q(x,0)$. Then we can see that (2.21) is a special case of the PDE in problem (2.1).

2.2.2. Diffusion-kinetics: Michaelis-Menten nonlinearity

A diffusion-kinetics equation governing the steady-state concentration u of some substrate in an enzyme-catalyzed reaction has the following form, see [21]:

$$\operatorname{div}(b(x)\nabla u) = q(x,u) \tag{2.22}$$

in a bounded domain Ω in $\mathbb{R}^2$. Here $b(x)$ is the uniformly positive molecular diffusion coefficient of the substrate in a medium containing some continuous distribution of bacteria, and q is the rate of the enzyme-substate reaction. In particular, the reaction rate is given by Michaelis-Menten theory:

$$q(x,\xi) \equiv q(\xi) = \frac{1}{\epsilon}\,\frac{\xi}{\xi + k} \qquad \text{for} \quad \xi \ge 0, \tag{2.23}$$

where $k > 0$ is the Michaelis constant. Here the more general model is $q(x,\xi) = \frac{1}{\epsilon(x)}\,\frac{\xi}{\xi+k}$, where $\epsilon(x) \ge \epsilon_0 > 0$, but typically ϵ is also constant [21], so we assume it from now on for simplicity. Let us impose mixed boundary conditions according to (2.16), where $u_0 \ge 0$ and the zero Neumann condition describes the insulated part of the boundary:

$$b(x)\frac{\partial u}{\partial n} = 0 \ \text{ on } \Gamma_N, \qquad u = u_0 \quad \text{on } \ \Gamma_D\,.$$

The rate (2.23) is defined only for the relevant concentrations $\xi \ge 0$. To obtain a proper operator, we can extend $q(\xi)$ from $\xi \ge 0$ to $\xi \in \mathbb{R}$ by the formula

$$q(\xi) = \frac{1}{\epsilon}\,\frac{\xi}{|\xi| + k} \qquad \text{for } \ \forall \xi \in \mathbb{R}.$$

Then $\xi \mapsto q(\xi)$ is monotone and Lipschitz continuous as required, i.e. the rearranged equation (2.22) with the mixed boundary conditions becomes a special case of (2.16) and has a unique weak solution. Furthermore, $r(\xi) := \frac{\epsilon^{-1}}{|\xi|+k}$ satisfies $0 \le r \le \frac{1}{\epsilon k}$, i.e. (2.2) holds. Then, by Theorem 2.2, the weak solution satisfies $u \ge 0$, hence u is the solution of the original problem (2.22) with nonlinearity (2.23).

2.2.3. Electrostatic potential equation

A semilinear model describing electrostatic potential is given by

$$\begin{cases} -\Delta u + e^u = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.24}$$

see [13,23]. It can be put in the above framework as follows. The equation is first written as $-\Delta u + e^u - 1 = -1$. Now let us denote

$$q(\xi) := \begin{cases} e^\xi - 1 & \text{if } \xi < 0, \\ \xi & \text{if } \xi \geq 0 \end{cases} \tag{2.25}$$

which is monotone and Lipschitz continuous, hence the problem

$$\begin{cases} -\Delta u + q(u) = -1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.26}$$

has a unique weak solution. Further, the function $r(\xi) = q(\xi)/\xi$ (defined as 1 for $\xi = 0$) satisfies $0 \leq r(\xi) \leq 1$ for all $\xi \in \mathbb{R}$, hence (2.2) holds and thus (2.26) becomes a special case of (2.1). Then Theorem 2.1 implies $u \leq 0$ for (2.26), hence $q(u)$ and $e^u - 1$ coincide and thus u is the solution of the original problem (2.24) as well.

2.2.4. Nonlinear stationary heat conduction

Problems of the form (2.7) appear in modelling heat conduction in nonlinear isotropic media. The problem

$$\begin{cases} -\operatorname{div}\left(a(x,u)\nabla u\right) + c(x,u)u = f(x) & \text{in } \Omega, \\ a(x,u)\frac{\partial u}{\partial \nu} = \gamma(x) & \text{on } \Gamma_N, \\ u = g(x) & \text{on } \Gamma_D, \end{cases}$$

is the isotropic case of the model described in [17]. The function u describes temperature distribution, where the heat conductivity is represented by the coefficient a, and the domain Ω is, for instance, a large transformer whose magnetic cores consist of iron tins.

2.2.5. Stefan–Boltzmann nonlinearity in radiation

The steady-state temperature distribution in various radiating bodies or gases lead to nonlinear PDEs of the form

$$\operatorname{div}(k(x)\nabla u) = \sigma(x)u^4 \quad \text{in } \Omega, \tag{2.27}$$

where the thermal conductivity $k(x)$ and the Boltzmann factor $\sigma(x)$ are uniformly positive, see [21].

In this example, similarly to subsection 2.2.2, the nonlinearity $q(x,\xi) = \sigma(x)\xi^4$ is defined only for $\xi \geq 0$ since we expect a solution $u \geq 0$, but we can define a monotone increasing extension to all $\xi \in \mathbb{R}$ by $q(x,\xi) := \sigma(x)|\xi|^3\xi$.

On the other hand, rewriting the above nonlinearity in the expected form $q(x,\xi) = r(x,\xi)\xi$, we find that r is not bounded as required in (2.2). This will lead us to discuss the extension of our results, allowing power order growth instead of the present bounds on r in (2.2), in subsection 4.5.1 at the end of the paper.

2.3. Finite element discretization

In the study of the discrete case, we assume that the domain Ω is a polytope, i.e. polygon or polyhedron in 2D or 3D, respectively. (If Ω has a curved boundary then it can be approximated with a polytope, see, e.g., [22].) To find the finite element solution for the model (2.1), consider a FEM subspace V_h of first-order elements. That is, the following general properties hold for the basis functions:

(B1) $0 \leq \phi_i \leq 1 \quad (\forall i = 1, \dots, n+m)$;
(B2) $\sum_{i=1}^{n+m} \phi_i \equiv 1$,
(B3) $\phi_i(P_j) = \delta_{ij}$ for proper nodes $P_1, \dots, P_n \in \Omega$ and $P_{n+1}, \dots, P_{n+m} \in \partial\Omega$.

In Section 4 we will consider Courant, tetrahedral, bilinear and prismatic elements, for all of which the conditions (B1)-(B3) hold.

We solve the following problem (which is the counterpart of (2.3) and (2.4) in V_h): find $u_h \in V_h$ such that

$$u_h = g_h \quad \text{on } \Gamma_D \qquad \text{and}$$
$$\int_\Omega \left[b(x,u_h,\nabla u_h)\,\nabla u_h \cdot \nabla v_h + r(x,u_h,\nabla u_h)u_h v_h\right] dx = \int_\Omega f_h v_h\, dx + \int_{\Gamma_N} \gamma_h v_h\, d\sigma \quad \forall v_h \in V_h^0. \tag{2.28}$$

To find the coefficient vector $\bar{\mathbf{c}}$ of u_h, following [18], the corresponding nonlinear algebraic system of equations is given by

$$\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} = \bar{\mathbf{b}}, \tag{2.29}$$

where the structure of the matrix is:

$$\bar{\mathbf{A}}(\bar{\mathbf{c}}) = \begin{pmatrix} \mathbf{A}(\overline{\mathbf{c}}) & \widetilde{\mathbf{A}}(\overline{\mathbf{c}}) \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{2.30}$$

where $\mathbf{I}$ is an $m\times m$ identity matrix and $\mathbf{0}$ is a $m\times n$ zero matrix, further, the entries of the matrix $\bar{\mathbf{A}}(\overline{\mathbf{c}})$ for $i = 1,\dots,n$ and $j = 1,\dots,n+m$ (i.e. both for $\mathbf{A}(\overline{\mathbf{c}})$ and $\widetilde{\mathbf{A}}(\overline{\mathbf{c}})$) are

$$a_{ij}(\bar{\mathbf{c}}) = \int_{\Omega_{ij}} \left[b(x,u_h,\nabla u_h)\,\nabla\phi_i\cdot\nabla\phi_j + r(x,u_h,\nabla u_h)\,\phi_i\phi_j \right] dx, \tag{2.31}$$

where ϕ_i and ϕ_j are corresponding basis functions and

$$\Omega_{ij} = supp\,\phi_i \cap supp\,\phi_j\,, \tag{2.32}$$

where *supp* refers to the support of a function (i.e. the closure of the set where it is nonvanishing). The vector $\bar{\mathbf{c}} = (c_1,\dots,c_{n+m})^T$ contains the values of the finite element solution u_h at all the nodal points, i.e. $c_i = u_h(P_i)$ and $u_h = \sum_{i=1}^{n+m} c_i\phi_i$, where $\phi_1,\dots.\phi_n$ are the interior basis functions and $\phi_{n+1},\dots,\phi_{n+m}$ are the boundary basis functions. Furthermore, $\bar{\mathbf{b}} = (b_1,\dots,b_n,g_1,\dots,g_m)^T$ and $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ is $(n+m)$ by $(n+m)$ matrix.

3. Background for DMP: matrix maximum principles

Consider a linear algebraic system $\overline{\mathbf{A}}\overline{\mathbf{c}} = \overline{\mathbf{b}}$ with a matrix with similar structure as in (2.30):

$$\overline{\mathbf{A}} = \begin{pmatrix} \mathbf{A} & \widetilde{\mathbf{A}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{3.1}$$

where the matrix $\bar{\mathbf{A}}$ has a dimension of $(n+m)$ by $(n+m)$.

Definition 3.1. (see [8]). The matrix $\bar{\mathbf{A}}$ satisfies

- the *discrete weak maximum principle (DwMP)* if for any vector $\bar{\mathbf{c}} = (c_1,\dots,c_{n+m})^T \in \mathbb{R}^{n+m}$ satisfying $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \le 0,\ i = 1,\dots,n$, one has

$$\max_{i=1,\dots,n+m} c_i \le \max\{0, \max_{i=n+1,\dots,n+m} c_i\};$$

- the *discrete strict weak maximum principle (DWMP)* if for any vector $\bar{\mathbf{c}} = (c_1,\dots,c_{n+m})^T \in \mathbb{R}^{n+m}$ satisfying $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \le 0,\ i = 1,\dots,n$, one has

$$\max_{i=1,\dots,n+m} c_i = \max_{i=n+1,\dots,n+m} c_i.$$

Theorem 3.1. *(see [18, Theorem 5]). Let the matrix $\bar{\mathbf{A}}$ in (3.1) satisfy the following conditions, where a_{ij} denote the entries of $\bar{\mathbf{A}}$:*

(i) $a_{ij} \le 0 \quad (\forall i = 1,\dots,n,\ j = 1,\dots,n+m;\ \ i \ne j)$,
(ii) $\sum_{j=1}^{n+m} a_{ij} \ge 0 \quad (\forall i = 1,\dots,n)$,
(iii) $\mathbf{A}$ *is positive definite.*

Then $\bar{\mathbf{A}}$ possesses the DwMP. If the inequality in condition (ii) is replaced by equality, then $\bar{\mathbf{A}}$ possesses the DWMP.

4. DMPs and mesh conditions for nonlinear elliptic problems

Now we can turn to our main goal, that is, to give sufficient conditions for the discrete analogues of Theorems 2.1–2.2 for the FE solution described in subsection 2.3. First we give some general properties, which show that the main task will be to ensure the nonpositivity of the offdiagonals of the FEM matrices. Then the latter will be presented for various first-order elements, where sufficient conditions are given for the mesh sizes under proper shape properties.

Let V_h be any FEM subspace as described in the subsection 2.3. The entries of the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ for $i = 1, \dots, n$ and $j = 1, \dots, n+m$ are given by (2.31), where ϕ_i and ϕ_j are corresponding basis functions and $\Omega_{ij} = supp\,\phi_i \cap supp\,\phi_j$.

Proposition 4.1. *Let the general properties* (B1)-(B3) *in Section 2.3 hold. Then the matrix* (2.30)–(2.31) *satisfies*

(i) $\sum_{j=1}^{n+m} a_{ij}(\bar{\mathbf{c}}) \geq 0$ *(*$\forall i = 1, \dots, n$*);*

(ii) $\mathbf{A}(\bar{\mathbf{c}})$ *is positive definite.*

Proof. It is similar as in [18] for simplicial elements, for completeness we summarize the proof.

(i) For any $i = 1, ..., n$ in (2.31), using properties (B1)-(B2) and (2.2), we have

$$\sum_{j=1}^{n+m} a_{ij}(\bar{\mathbf{c}}) = \int_\Omega \Big[b(x, u_h, \nabla u_h)\, \nabla\phi_i \cdot \nabla\Big(\sum_{j=1}^{n+m} \phi_j\Big) + r(x, u_h, \nabla u_h)\, \phi_i \Big(\sum_{j=1}^{n+m} \phi_j\Big) \Big]\, dx$$
$$= \int_\Omega r(x, u_h, \nabla u_h)\, \phi_i\, dx \geq 0\,.$$

(iii) To verify that $\mathbf{A}(\bar{\mathbf{c}})$ is positive definite, let $\mathbf{d} \neq \mathbf{0}$ be an arbitrary vector in $\mathbb{R}^n$, formed by the coefficients d_i, and let

$$v_h = \sum_{j=1}^{n} d_j \phi_j \in V_h.$$

Then $v_h \not\equiv 0$. The vector $\bar{\mathbf{c}} \in \mathbb{R}^{n+m}$ contains the coefficients for u_h as given in (2.28)–(2.29). Then, using (2.28) and (2.2), we have

$$A(\bar{\mathbf{c}})\mathbf{d} \cdot \mathbf{d} = \sum_{i,j=1}^{n} a_{ij}(\bar{\mathbf{c}}) d_i d_j$$
$$= \int_\Omega \Big(b(x, u_h, \nabla u_h) \nabla\Big(\sum_{i=1}^{n} d_i \phi_i \Big) \cdot \nabla\Big(\sum_{j=1}^{n} d_j \phi_j \Big) + r(x, u_h, \nabla u_h) \sum_{i=1}^{n} d_i \phi_i \sum_{j=1}^{n} d_j \phi_j \Big)\, dx$$
$$= \int_\Omega \Big(b(x, u_h, \nabla u_h) |\nabla v_h|^2 + r(x, u_h, \nabla u_h) v_h^2 \Big)\, dx \geq \mu_0 \int_\Omega |\nabla v_h|^2 = \mu_0 |v_h|_1^2 > 0. \quad \square$$

Proposition 4.2. *Let the general properties* (B1)-(B3) *in subsection 2.3 hold. If* $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ *in* (2.30)–(2.31) *satisfies the DwMP, then* u_h *satisfies the DMP: that is, if*

$$f(x) \leq 0 \quad (x \in \Omega) \quad \textit{and} \quad \gamma(x) \leq 0 \quad (x \in \Gamma_N), \tag{4.1}$$

then

$$\max_{\overline{\Omega}} u_h \leq \max\{0, \max_{\Gamma_D} g_h\}. \tag{4.2}$$

In particular, if $\max_{\Gamma_D} g_h \geq 0$*, then*

$$\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} g_h, \tag{4.3}$$

and if $g_h \leq 0$*, then we have the nonpositivity property*

$$u_h \leq 0 \quad \text{on } \overline{\Omega}. \tag{4.4}$$

Proof. Let $\bar{\mathbf{c}} = (c_1, ..., c_{n+m})^T \in \mathbb{R}^{n+m}$ and $\bar{\mathbf{b}} = (b_1, ..., b_n, g_1, ..., g_m)^T \in \mathbb{R}^{n+m}$ be the vectors that appear in (2.29). Then

$$b_i = \int_\Omega f \phi_i\, dx + \int_{\Gamma_N} \gamma \phi_i\, d\sigma \leq 0 \qquad (i = 1, \dots, n) \tag{4.5}$$

owing to $f \leq 0$, $\gamma \leq 0$ and property (B1) of subsection of 2.3. Then (2.29) and (4.5) imply $\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} = \bar{\mathbf{b}} \leq \mathbf{0}$. Since it is assumed that the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ possesses the DwMP, we have by definition

$$\max_{i=1,\dots,n+m} c_i \leq \max\{0, \max_{i=n+1,\dots,n+m} c_i\} = \max\{0, \max_{i=n+1,\dots,n+m} g_i\} \tag{4.6}$$

since $c_i = g_i$ for all $i = n, \dots, n+m$. Using the fact that $0 \le \phi_i \le 1$ and $\sum_{i=1}^{n+m} \phi_i = 1$ from properties (B2)-(B3), the solution vectors u_h and g_h can be estimated respectively as

$$u_h = \sum_{i=1}^{n+m} c_i \phi_i \le \max_{i=1,\dots,n+m} c_i \sum_{i=1}^{n+m} \phi_i = \max_{i=1,\dots,n+m} c_i, \tag{4.7}$$

$$g_h = \sum_{i=n+1}^{n+m} g_i \phi_i \le \max_{i=n+1,\dots,n+m} g_i \sum_{i=n+1}^{n+m} \phi_i \;\Rightarrow\; \max_{\Gamma_D} g_h \le \max_{i=n+1,\dots,n+m} g_i. \tag{4.8}$$

Moreover, equality holds in (4.8) because of the following argument. Let $g_k := \max_{i=n+1,\dots,n+m} g_i$. Then, using (B3) of the general properties in subsection of 2.3, $g_h(P_k) = \sum_{i=n+1}^{n+m} g_i \phi_i(P_k) = g_k$ since $\phi_i(P_k) = 1$ and 0 in the other nodes. Hence, indeed,

$$\max_{\Gamma_D} g_h = \max_{i=n+1,\dots,n+m} g_i$$

and thus also

$$\max\{0, \max_{\Gamma_D} g_h\} = \max\{0, \max_{i=n+1,\dots,n+m} g_i\}. \tag{4.9}$$

Altogether, (4.7), (4.6) and (4.9) imply (4.2). The remaining two statements are direct consequences of (4.2). □

Now we come to the main point to be used in the sequel of this paper:

Corollary 4.1. *Let the general properties* (B1)-(B3) *in subsection 2.3 hold. If the matrix* (2.30)–(2.31) *satisfies*

$$a_{ij}(\bar{\mathbf{c}}) \le 0 \qquad \forall i = 1, \dots, n,\ j = 1, \dots, n+m;\ \ i \ne j, \tag{4.10}$$

then the DMP (4.2) holds as well as its consequences (4.3) and (4.4) under the proper sign conditions given in Proposition 4.2.

Indeed, if (4.10) holds then Proposition 4.1 and Theorem 3.1 imply that $\mathbf{A}(\bar{\mathbf{c}})$ satisfies the DwMP, and thus Proposition 4.2 yields that the DMP and its consequences hold.
Therefore, in what follows, our remaining task is to verify (4.10) (i.e. the nonpositivity of the off-diagonals of the FEM matrices) for the studied types of first order elements under proper mesh conditions.

4.1. Courant elements

In our recent article [2], we have determined the threshold mesh size h_0 to ensure the validity of DMPs by Courant FEM in 2D for suitable meshes with uniformly acute angle conditions for the nonlinear elliptic model (2.1). In our result, we have used the definition below. Here a *family of triangulations* means a collection $\mathcal{F} = \{\mathcal{T}_h\}_{h>0}$ for which the mesh widths h accumulate at 0.

Definition 4.1. A family $\mathcal{F}$ of triangulations of a bounded polygonal domain is said to be *uniformly acute* if there exists $\alpha_0 < \frac{\pi}{2}$ such that $\alpha_n \le \alpha_0$ for any angle α_n in all triangles T_k in all $\mathcal{T}_h$, where $\mathcal{T}_h \in \mathcal{F}$.

Theorem 4.1. *[2] Let Assumption 2.1 hold and Courant finite elements be used with triangulations satisfying Definition 4.1. Let the mesh size h satisfy*

$$0 < h \le h_0 = \Big(\frac{12 \mu_0 \cos \alpha_0}{\beta} \Big)^{\frac{1}{2}}, \tag{4.11}$$

where α_0 is the angle that obeys Definition 2, μ_0 and β are positive constants from (2.2). Then the matrix in (2.30) satisfies

$$a_{ij}(\bar{\mathbf{c}}) \le 0, \quad i = 1, \dots, n,\ j = 1, \dots, n+m \quad (i \ne j).$$

Consequently, the DMP (4.2) holds.
Now, we will give similar results to investigate the validity of DMPs using tetrahedral, bilinear, and prismatic elements.

4.2. Tetrahedral elements

Now let us consider a 3D problem of the type (2.1) and apply tetrahedral P1 elements. For the description we rely on [12]. For a given tetrahedron K we denote by α_{ij}^K the angle between the two dimensional facets F_i^K and F_j^K which are adjacent to the given edge (Fig. 1).

Definition 4.2. A family $\mathcal{F}$ of tetrahedral triangulations of a bounded polyhedral domain is said to be *uniformly acute* if there exists $\alpha_0 < \frac{\pi}{2}$ such that $\alpha_{ij}^K \le \alpha_0$ for any angle α_{ij}^K in all $K \in \mathcal{T}_h$, and $\mathcal{T}_h \in \mathcal{F}$.

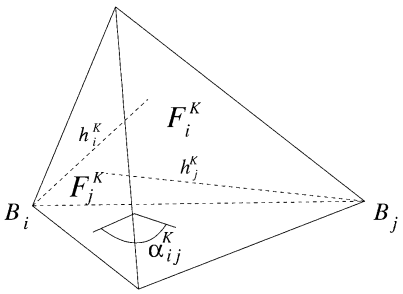

Fig. 1. A tetrahedral cell K from [12].

Theorem 4.2. *Let $d = 3$ and Assumption 2.1 hold, and let the tetrahedral finite element method be used with triangulations satisfying Definition 4.2. Let the mesh size h satisfy*

$$0 < h \le h_0 = \Big(\frac{20\mu_0 \cos\alpha_0}{\beta} \Big)^{\frac{1}{2}}, \tag{4.12}$$

where α_0 is the angle that obeys Definition 3, μ_0 and β are the positive constants from (2). Then the matrix in (2.30) satisfies

$$a_{ij}(\bar{\mathbf{c}}) \le 0, \quad i = 1, ..., n, \ j = 1, ..., n+m \quad (i \ne j).$$

Consequently, the DMP (4.2) holds.

Proof. Let us consider the entries in (2.31). If $int(\Omega_{ij}) \ne \emptyset$, then to estimate (2.31) we should find the values of the following integrals:

$$\int\limits_{\Omega_{ij}} \nabla\phi_i \cdot \nabla\phi_j \, dx \qquad \text{and} \qquad \int\limits_{\Omega_{ij}} \phi_i\phi_j \, dx, \tag{4.13}$$

and the goal is to find proper upper bounds. To compute the entries of the stiffness matrix over a tetrahedron K, we shall use the formula stated in [12] for the inner product of the basis functions, and for the upper bound we let $\sigma_0 := cos(\alpha_0)$, where α_0 is the maximum angle from Definition 4.2, hence $\sigma_0 > 0$ and it is independent of i, j and h. Altogether, if $i \ne j$, then

$$\nabla\phi_i \cdot \nabla\phi_j | K = -\frac{cos(\alpha_{ij}^K)}{h_i^K h_j^K} \le \frac{-cos(\alpha_0)}{h^2}, \tag{4.14}$$

since all ${h_i}^K, {h_j}^K \le h$ and $\alpha_{ij}^K \le \alpha_0$. Therefore, we have a bound independently of K:

$$\nabla\phi_i \cdot \nabla\phi_j | K \le -\frac{\sigma_0}{h^2} < 0. \tag{4.15}$$

Hence the bounds of the off-diagonal entries of the stiffness matrix are given by

$$\int\limits_{\Omega_{ij}} \nabla\phi_i \cdot \nabla\phi_j \, dx = \sum_{K \subset \Omega_{ij}} \int\limits_K \nabla\phi_i \cdot \nabla\phi_j \, dx \le -\frac{\sigma_0}{h^2} meas(\Omega_{ij}). \tag{4.16}$$

The formula for the entries of the mass matrix M over K from [12] is

$$m_{ij} | K = \int\limits_K \phi_i\phi_j \, dx = (1 + \delta_{ij}) \frac{d!}{(d+1)!} meas_d K,$$

where δ_{ij} is Kronecker's symbol. Therefore, since $d = 3$ and $\delta_{ij} = 0$ for $i \ne j$, we get

$$\int\limits_{\Omega_{ij}} \phi_i\phi_j \, dx = \sum_{K \subset \Omega_{ij}} \int\limits_K \phi_i\phi_j \, dx = \sum_{K \subset \Omega_{ij}} \frac{1}{20} meas(K) = \frac{1}{20} meas(\Omega_{ij}), \tag{4.17}$$

where Ω_{ij} is as in (2.32). Using (2.2), (2.31), (4.16), (4.17) and (B1) of subsection 2.3, we have

$$a_{ij}(\bar{\mathbf{c}}) \le \mu_0 \int\limits_{\Omega_{ij}} \nabla\phi_i \cdot \nabla\phi_j \, dx + \beta \int\limits_{\Omega_{ij}} \phi_i\phi_j \, dx$$

$$\le -\frac{\sigma_0}{h^2} \mu_0 \, \text{meas}\,(\Omega_{ij}) + \frac{\beta}{20} \, meas(\Omega_{ij}) = meas(\Omega_{ij}) \Big(-\frac{\sigma_0}{h^2} \mu_0 + \frac{\beta}{20} \Big).$$

Let

$$b_{ij}(h) := -\frac{\sigma_0}{h^2}\mu_0 + \frac{\beta}{20}, \tag{4.18}$$

then

$$a_{ij}(\bar{\mathbf{c}}) \le b_{ij}(h)\, meas(\Omega_{ij}). \tag{4.19}$$

The sum of the terms in $b_{ij}(h)$ tends to $-\infty$ as $h \to 0$, which implies $a_{ij}(h) \le 0$ if h is small. The main task here is to find how small h should be to guarantee the nonpositivity of (2.31). To determine the threshold $h = h_0$, the following equation must hold:

$$-\frac{\sigma_0}{h^2}\mu_0 + \frac{\beta}{20} = 0.$$

This implies $h_0 = \left(\frac{20\sigma_0\mu_0}{\beta}\right)^{\frac{1}{2}}$. In summary, if $0 < h \le h_0 = \left(\frac{20\sigma_0\mu_0}{\beta}\right)^{\frac{1}{2}}$, then (4.19) yields $a_{ij}(\bar{\mathbf{c}}) \le 0$. □

Remark 4.1. A nice family of simple examples is when Ω is the union of acute tetrahedra forming a so-called TCP-structure. Various such space-tiling TCP-structures, motivated by crystallography, are described in [7]. For instance, the mesh called A15, constructed from a square tiling, consists of tetrahedra with maximal dihedral angle 78.46°, i.e. with cosine value 0.2. Then $h_0 = 2\left(\mu_0/\beta\right)^{1/2}$.

4.3. Bilinear elements

Consider a semilinear special case for the model problem (2.1) for $d = 2$:

$$\begin{cases} -\mu_0\,\Delta u + r(x,u,\nabla u)u = f(x) & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = \gamma(x) & \text{on } \Gamma_N, \\ u = g(x) & \text{on } \Gamma_D, \end{cases} \tag{4.20}$$

where $\mu_0 > 0$ is constant and $0 \le r(x,\xi,\eta) \le \beta$ from (2.2). We consider bilinear elements for a 2D rectangular mesh. The following definition and Theorem are crucial to investigate the validity of DMP for (4.20).

Definition 4.3. A family $\mathcal{F}$ of rectangular meshes on a given domain is said to be *uniformly non-narrow* if there exists $\rho_0 < \sqrt{2}$ such that for any rectangle we have $\frac{H}{h} \le \rho_0$ where H and h denote the longest and shortest side of the rectangle, respectively.

Theorem 4.3. *Let Assumption 2.1 hold and the bilinear finite element method be used with a mesh satisfying Definition 4.3. Let the mesh size h satisfy*

$$0 < h \le h_0 = \frac{\sqrt{3\mu_0(2-\rho_0^2)}}{\rho_0\sqrt{\beta}} \tag{4.21}$$

where ρ_0 obeys Definition 4.3, μ_0 and β are the positive constants from (2.2). Then the matrix in (2.30) satisfies

$$a_{ij}(\bar{\mathbf{c}}) \le 0, \quad i = 1,\dots,n,\ j = 1,\dots,n+m \quad (i \ne j).$$

Consequently, the DMP (4.2) holds.

Proof. We have two possible patches in a rectangular mesh when $i \ne j$: the nodes can be diagonal or edge neighbours. We use the results of [11] for the integrals on a rectangle R:

$$\int_R \nabla\phi_i \cdot \nabla\phi_j\, dx = \begin{cases} \frac{H^2-2h^2}{6hH} & \text{if } \phi_i \text{ and } \phi_j \text{ are edge neighbours,} \\ -\frac{h^2+H^2}{6hH} & \text{if } \phi_i \text{ and } \phi_j \text{ are diagonal neighbours,} \end{cases} \tag{4.22}$$

and

$$\int_R \phi_i\phi_j\, dx = \begin{cases} \frac{hH}{18} & \text{if } \phi_i \text{ and } \phi_i \text{ are edge neighbours,} \\ \frac{hH}{36} & \text{if } \phi_i \text{ and } \phi_i \text{ are diagonal neighbours.} \end{cases} \tag{4.23}$$

Let us first consider edge neighbours and let ϕ_i and ϕ_j be two such basis functions. Here Ω_{ij} consists of two rectangles. The entries of the matrix for one rectangle R are given as follows:

$$a_{ij}(\bar{\mathbf{c}})|_R = \mu_0 \int_R \nabla\phi_i \cdot \nabla\phi_j \, dx + \int_R r(x,u,\nabla u)\phi_i\phi_j \, dx$$

$$\le \mu_0 \int_R \nabla\phi_i \cdot \nabla\phi_j \, dx + \beta \int_R \phi_i\phi_j \, dx$$

$$= \frac{\mu_0(H^2-2h^2)}{6hH} + \frac{\beta hH}{18} = \frac{1}{18}\Big(3\mu_0\Big(\frac{H}{h} - \frac{2h}{H}\Big) + \beta Hh\Big). \tag{4.24}$$

Letting $\rho := \frac{H}{h} \le \rho_0$, i.e. $H = \rho h$, gives

$$a_{ij}(\bar{\mathbf{c}})|_R = \frac{1}{18}\Big(3\mu_0\Big(\rho - \frac{2}{\rho}\Big) + \beta\rho h^2\Big) \le \frac{1}{18}\Big(3\mu_0\Big(\rho_0 - \frac{2}{\rho_0}\Big) + \beta\rho_0 h^2\Big).$$

This holds on both rectangles, hence

$$a_{ij}(\bar{\mathbf{c}}) \le a_{ij}(h) := \frac{2}{18}\Big(3\mu_0\Big(\rho_0 - \frac{2}{\rho_0}\Big) + \beta\rho_0 h^2\Big).$$

Multiplying with ρ_0, we see we should satisfy

$$3\mu_0(\rho_0^2 - 2) + \beta\rho_0^2 h^2 \le 0.$$

Indeed, this holds for sufficiently small h using that $\rho_0 < \sqrt{2}$, hence $a_{ij}(\bar{\mathbf{c}}) \le 0$ as well. The threshold h_0 is obtained by expressing h from equality above. Altogether, for the edge neighbours the threshold is

$$h_0 = \frac{\sqrt{3\mu_0(2-\rho_0^2)}}{\rho_0\sqrt{\beta}}. \tag{4.25}$$

Now we consider diagonal neighbours, we use similar arguments as for the edge neighbours. Now Ω_{ij} is one rectangle R, and the entries of the matrix are calculated by combining equations (4.22) and (4.23):

$$a_{ij}(\bar{\mathbf{c}}) \le \mu_0 \int_R \nabla\phi_i \cdot \nabla\phi_j \, dx + \beta \int_R \phi_i\phi_j \, dx = -\frac{\mu_0(h^2+H^2)}{6hH} + \frac{\beta hH}{36} =: a_{ij}(h,H).$$

We want to determine h and H such that $a_{ij}(h,H) \le 0$, i.e.

$$6\mu_0\Big(-\frac{h}{H} - \frac{H}{h}\Big) + \beta Hh \le 0.$$

Letting $\rho = \frac{H}{h}$, i.e. $H = \rho h$ and multiply by ρ gives

$$-6\mu_0 - 6\mu_0\rho^2 + \beta\rho^2 h^2 \le 0.$$

This is true for sufficiently small h, namely, when

$$0 < h \le \frac{\sqrt{6\mu_0(1+\rho^2)}}{\rho\sqrt{\beta}}. \tag{4.26}$$

Note that $\sqrt{2} > \rho_0 \ge 1$, hence h_0 in (4.25) satisfies

$$h_0 \le \frac{\sqrt{3\mu_0}}{\sqrt{\beta}} < \frac{\sqrt{6\mu_0}}{\sqrt{\beta}}\sqrt{\rho^{-2}+1} = \frac{\sqrt{6\mu_0(1+\rho^2)}}{\rho\sqrt{\beta}} \qquad (\forall \rho > 0),$$

hence if $h \le h_0$, then (4.26) also holds. Altogether, h_0 in (4.25) is a suitable threshold for both diagonal and edge neighbours. □

Example. Let us apply a uniform square mesh on Ω for the following problem:

$$-\mu_0\Delta u + \frac{u}{\lambda + \epsilon u} = f \quad \text{in} \quad \Omega \tag{4.27}$$

(with proper boundary conditions), which involves the rewritten form of the Michaelis-Menten nonlinearity (2.23), i.e. $\lambda, \epsilon > 0$ are given constants. To compute h_0 in (4.21), we need to calculate the constants therein. Since $\beta = \frac{1}{\lambda}$ and $\rho_0 = 1$, we obtain

$$h_0 = \sqrt{3\mu_0\lambda}. \tag{4.28}$$

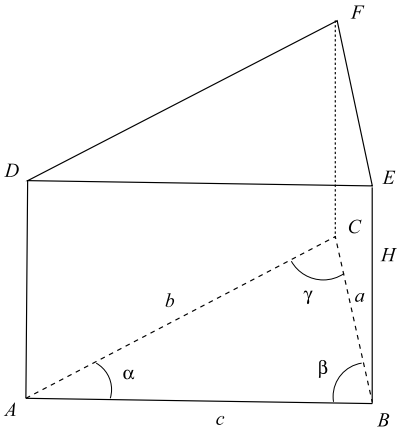

Fig. 2. Basic notations for prismatic elements, based on [14].

4.4. Prismatic elements

Consider the semilinear special case (4.20) of our model problem, where

$$\mu_0 > 0 \ \text{ is constant and } \quad 0 \le r(x,\xi,\eta) \le \beta_0. \tag{4.29}$$

(Here β_0 is written instead of the previously used β so as not to collide with the notation for angles α, β, γ of triangles that we will rely on in the mesh.) Now $d = 3$, and Ω is a given domain in $\mathbb{R}^3$ which can be partitioned face-to-face into triangular prisms. A prism is of the form $P = T \times I$ for a given triangle T and interval I, and the corresponding first-order elements have the vertices of the prism as degrees of freedom, see [14] for more details. We will rely on Fig. 2 for the notations to be used in the sequel. The parameter for the triangular mesh on which the prisms are based will be denoted by h. We also impose the following mesh regularity assumption:

Assumption 4.4. Let $h > 0$ be the triangular mesh parameter. There exist fixed angles $0 < \gamma_{min} \le \gamma_{max} < \frac{\pi}{2}$ such that the area $|T|$ of any triangle T satisfies

$$\frac{1}{2}h^2 \sin\gamma_{min} \le |T| \le \frac{1}{2}h^2 \sin\gamma_{max}\,.$$

Further, let γ_{med} denote a lower bound for the second largest degrees of the triangles T.

Theorem 4.4. *Let Assumption 4.4 hold, and let us fix a constant δ_1 such that*

$$0 < \delta_1 < \frac{4\cot\gamma_{max}}{\sin\gamma_{max}}. \tag{4.30}$$

If the mesh parameters satisfy the following conditions, where μ_0 and β_0 are from (4.29):

$$h^2 \le \frac{3\mu_0\delta_1}{\beta_0}\,, \tag{4.31}$$

$$\frac{\cot\gamma_{med} + \cot\gamma_{min}}{\sin\gamma_{min}} + \frac{1}{2}\,\delta_1 \ \le \left(\frac{h}{H}\right)^2 \le \frac{4\cot\gamma_{max}}{\sin\gamma_{max}} - \delta_1\,, \tag{4.32}$$

then the matrix in (2.30) satisfies

$$a_{ij}(\bar{\mathbf{c}}) \le 0, \quad i = 1,\dots,n,\ j = 1,\dots,n+m \quad (i \ne j)$$

Consequently, the DMP (4.2) holds.

Proof. Let ϕ_i and ϕ_j be basis functions, $i \ne j$. Then the entries of the FE matrix are given by

$$a_{ij}(\bar{\mathbf{c}}) = \sum_{P \subset \Omega_{ij}} a_{ij}^P(\bar{\mathbf{c}}),$$

where

$$a_{ij}^P(\bar{\mathbf{c}}) = \mu_0 \int_P \nabla\phi_i \cdot \nabla\phi_j \,dx + \int_P r(x,u,\nabla u)\phi_i\phi_j \,dx \le \mu_0 \int_P \nabla\phi_i \cdot \nabla\phi_j \,dx + \beta_0 \int_P \phi_i\phi_j \,dx$$

are the contributions on the individual prisms P. It suffices to ensure $a_{ij}^P(\bar{\mathbf{c}}) \le 0$ on each prism. Here, in the case of prismatic FEM, we have three cases as in [14], see (9)-(11) therein. We use the notations of Fig. 2.

Case (i):

$$a_{ij}^P(\bar{\mathbf{c}}) \le \mu_0 \int_P \nabla\phi_A \cdot \nabla\phi_B \, dx + \beta_0 \int_P \phi_A \phi_B \, dx = \frac{H}{12}\Big(\mu_0\Big(-2\cot\gamma + \frac{|T|}{H^2}\Big) + \frac{\beta_0|T|}{3}\Big).$$

Case (ii):

$$a_{ij}^P(\bar{\mathbf{c}}) \le \mu_0 \int_P \nabla\phi_A \cdot \nabla\phi_D \, dx + \beta_0 \int_P \phi_A \phi_D \, dx = \frac{H}{12}\Big(\mu_0\Big(\cot\beta + \cot\gamma - \frac{2|T|}{H^2}\Big) + \frac{\beta_0|T|}{3}\Big).$$

Case (iii):

$$a_{ij}^P(\bar{\mathbf{c}}) \le \mu_0 \int_P \nabla\phi_A \cdot \nabla\phi_E \, dx + \beta_0 \int_P \phi_A \phi_E \, dx = -\frac{H}{12}\Big(\mu_0\Big(\cot\gamma + \frac{|T|}{H^2}\Big) - \frac{\beta_0|T|}{6}\Big).$$

Now, our goal is to find conditions on h and H to satisfy $a_{ij}^P(\bar{\mathbf{c}}) \le 0$ in all the cases (i)-(iii).
Case (i). Using Assumption 4.4, we have

$$a_{ij}^P(\bar{\mathbf{c}}) \le \frac{H}{12}\Big(\mu_0\Big(-2\cot\gamma_{max} + \frac{h^2}{2H^2}\sin\gamma_{max}\Big) + \frac{\beta_0 h^2}{6}\sin\gamma_{max}\Big),$$

hence to achieve $a_{ij}^P(\bar{\mathbf{c}}) \le 0$, we need

$$\Big(\frac{h}{H}\Big)^2 \le \frac{4\cot\gamma_{max}}{\sin\gamma_{max}} - \frac{\beta_0 h^2}{3\mu_0}. \tag{4.33}$$

Assumption (4.31) implies

$$\frac{\beta_0 h^2}{3\mu_0} \le \delta_1. \tag{4.34}$$

This and the r.h.s. of (4.32) yield

$$\Big(\frac{h}{H}\Big)^2 \le \frac{4\cot\gamma_{max}}{\sin\gamma_{max}} - \delta_1 \le \frac{4\cot\gamma_{max}}{\sin\gamma_{max}} - \frac{\beta_0 h^2}{3\mu_0}, \tag{4.35}$$

i.e. the desired bound (4.33) is satisfied and hence $a_{ij}^P(\bar{\mathbf{c}}) \le 0$.
Case (ii). Using Assumption 4.4, we have

$$a_{ij}^P(\bar{\mathbf{c}}) \le \frac{H}{12}\Big(\mu_0\big(\cot\gamma_{med} + \cot\gamma_{min} - \frac{2|T|}{H^2}\big) + \frac{\beta_0|T|}{3}\Big), \tag{4.36}$$

hence to achieve $a_{ij}^P(\bar{\mathbf{c}}) \le 0$, we need

$$\frac{\cot\gamma_{med} + \cot\gamma_{min}}{|T|} \le \frac{2}{H^2} - \frac{\beta_0}{3\mu_0}. \tag{4.37}$$

Using Assumption 4.4 again, for (4.37) it suffices that

$$\frac{\cot\gamma_{med} + \cot\gamma_{min}}{\sin\gamma_{min}} \le \Big(\frac{h}{H}\Big)^2 - \frac{\beta_0 h^2}{6\mu_0}. \tag{4.38}$$

Since Assumption (4.31) implies (4.34), this and the l.h.s. of (4.32) yield

$$\frac{\cot\gamma_{med} + \cot\gamma_{min}}{\sin\gamma_{min}} + \frac{\beta_0 h^2}{6\mu_0} \le \frac{\cot\gamma_{med} + \cot\gamma_{min}}{\sin\gamma_{min}} + \frac{1}{2}\delta_1 \le \Big(\frac{h}{H}\Big)^2, \tag{4.39}$$

i.e. the desired bound (4.38) is satisfied and hence $a_{ij}^P(\bar{\mathbf{c}}) \le 0$.

Case (iii). Now, using the notation $r(T,H) := -\frac{\mu_0|T|}{H^2} + \frac{\beta_0|T|}{6}$, we have

$$a_{ij}^P(\bar{\mathbf{c}}) \le \frac{H}{12}\Big(-\mu_0 cot\gamma - \frac{\mu_0|T|}{H^2} + \frac{\beta_0|T|}{6}\Big) = \frac{H}{12}\Big(-\mu_0 cot\gamma + r(T,H)\Big) \le \frac{H}{12}\, r(T,H). \tag{4.40}$$

On the other hand, we have seen in case (ii) that the r.h.s. of (4.36) is nonpositive, i.e.

$$\mu_0\big(\cot\gamma_{med} + \cot\gamma_{min} - \frac{2|T|}{H^2}\big) + \frac{\beta_0|T|}{3} = \mu_0\big(\cot\gamma_{med} + \cot\gamma_{min}\big) + 2\,r(T,H) \le 0.$$

Hence also $r(T,H) \le 0$, and with (4.40), this implies that $a_{ij}^P(\bar{\mathbf{c}}) \le 0$ holds for case (iii) with no extra condition, given that it has already been ensured in case (ii). □

Remark 4.2. (i) In the case of a uniform equilateral triangular plane partition for the base of the prisms, the condition (4.32) becomes

$$\frac{4}{3} + \frac{1}{2}\,\delta_1 \;\le \left(\frac{h}{H}\right)^2 \le \frac{8}{3} - \delta_1\,. \tag{4.41}$$

This means that we can choose a fixed ratio of h and H to satisfy

$$\frac{4}{3} < \left(\frac{h}{H}\right)^2 < \frac{8}{3} \tag{4.42}$$

and then use the largest value of δ_1, expressed to satisfy (4.41), for the threshold (4.31). For example, if $(\frac{h}{H})^2 = \frac{16}{9}$, then $\delta_1 = \frac{8}{9}$ and the threshold in (4.31) is $h^2 \le \frac{8\mu_0}{3\beta_0}$. In particular, as a concrete illustration, if in the PDE we have $\beta_0 = 1$ and $\mu_0 = 10^{-3}$ (since the diffusion coefficient is often small), then the conditions altogether are

$$\frac{h}{H} = \frac{4}{3} \quad \text{and} \quad h \le 0.051. \tag{4.43}$$

We also note that the condition (4.42) is an analogue of [14, Remark 5], formulated for linear problems with vanishing reaction coefficient.

(ii) In the general case, if we can only access γ_{max} easily then we can use the estimates $\gamma_{min} \ge \pi - 2\gamma_{max}$ and $\gamma_{med} \ge \pi/4$, thus the l.h.s. of (4.32) can be replaced by the expression $\frac{1-\cot 2\gamma_{max}}{\sin 2\gamma_{max}} + \frac{1}{2}\,\delta_1$. However, this gives a more pessimistic sufficient condition.

(iii) To make (4.32) feasible for general partitions, the angles of the triangles must satisfy proper restrictions, analogously to the linear case in [14], wherein various examples are given for such so-called "well-shaped prismatic partitions".

4.5. Some extensions

The structure of our general problem (2.1) has enabled an organized study of the conditions in the above. Now we consider some extensions related to the nonlinearity in the PDE or to the boundary conditions.

4.5.1. Power order nonlinearities

Let $\Omega \subset \mathbb{R}^d$ for $d = 2$ or 3, and consider a semilinear elliptic Dirichlet problem

$$\begin{cases} -\mu_0\,\Delta u + r(x,u,\nabla u)u = f(x) & \text{in } \Omega, \\ \qquad\qquad\qquad\quad u = 0 \qquad\quad \text{on } \partial\Omega \end{cases} \tag{4.44}$$

with proper power order growth assumed for r w.r.t. u:

Assumption 4.5.1. $\mu_0 > 0$ is a given constant, further, there exists $\beta > 0$ and $q \ge 1$ such that the scalar function r satisfies

$$0 \le r(x,\xi,\eta) \le \beta|\xi|^q \tag{4.45}$$

for any $x \in \Omega$, $\xi \in \mathbb{R}$ and $\eta \in \mathbb{R}^d$. In addition, if $d = 3$ then $q \le 2$.

Remark 4.3. The estimations, required by the power growth, will use the following property. Let p be a real number satisfying $2 \le p$ (if $d = 2$) or $2 \le p \le \frac{2d}{d-2}$ (if $d \ge 2$). Then there hold the Sobolev embedding and corresponding estimate

$$H_0^1(\Omega) \subset L^p(\Omega), \quad \|u\|_{L^p(\Omega)} \le K_{p,\Omega}\,|u|_1 \tag{4.46}$$

for some constant $K_{p,\Omega} > 0$, see [13].

We note that this embedding is true for $H^1(\Omega)$ as well, but for $H_0^1(\Omega)$ there are more convenient estimations for the value of $K_{p,\Omega} > 0$, see Remark 4.5 below. Further, the strong restriction in Assumption 4.5.1 for $d = 3$ will arise from the stronger bound on p in (4.46).

The conditions for the used FEM, besides those in subsection 2.3, are as follows.

Assumptions 4.5.2. There exist constants $\sigma_0 > 0$ and $c_0 > 0$, independent of h, such that for any $i \ne j$ for which the interior of $\Omega_{ij} := supp\,\phi_i \cap supp\,\phi_j$ is nonempty,

(i) the basis functions satisfy

$$\int_{\Omega_{ij}} \nabla\phi_i \cdot \nabla\phi_j \le -\frac{\sigma_0}{h^2}\,meas(\Omega_{ij})\,; \tag{4.47}$$

(ii) we have

$$meas(\Omega_{ij}) \geq c_0 h^d \tag{4.48}$$

(where d is the space dimension), i.e. the mesh is uniformly shape regular.

Remark 4.4. Condition (4.47) is satisfied for Courant, tetrahedral, and bilinear elements as well under our conditions given in the previous subsections. Namely, for tetrahedral elements, it is shown in (4.16), and it is similar for Courant elements; for bilinear elements, it follows from (4.22) and the assumed non-narrowness.
Also, condition (4.48) is satisfied for Courant and tetrahedral elements under the minimum angle condition (which is used typically to ensure convergence), and follows again from the non-narrowness for bilinear elements.

Theorem 4.5. *Let Assumptions 4.5.1-4.5.2 hold. Let the mesh size h satisfy*

$$0 < h \leq h_0 = \frac{\mu_0 \sigma_0 c_0^{1/d}}{\beta K_{qd,\Omega}^q w_0^q} \tag{4.49}$$

where d is the space dimension, $\mu_0, \beta, \sigma_0, c_0$ are constants from Assumptions 4.5.1-4.5.2, respectively, $K_{qd,\Omega}$ is defined with (4.46), and

$$w_0 := \frac{diam(\Omega)\|f\|_{L^2(\Omega)}}{\mu_0 \pi \sqrt{d}} \tag{4.50}$$

(where $diam$ refers to the diameter). Then the matrix in (2.30) satisfies

$$a_{ij}(\bar{\mathbf{c}}) \leq 0, \quad i = 1, \ldots, n, \ j = 1, \ldots, n+m \quad (i \neq j).$$

Consequently, the DMP (4.2) holds.

Proof. Let ϕ_i and ϕ_j be basis functions, $i \neq j$. Using (4.45), (4.47) and the fact that $0 \leq \phi_i \leq 1$, we have

$$\begin{aligned} a_{ij}(\bar{\mathbf{c}}) &= \mu_0 \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, dx + \int_\Omega r(x, u_h, \nabla u_h)\phi_i\phi_j \, dx \\ &\leq -\frac{\sigma_0}{h^2}\mu_0 \, \mathrm{meas}\,(\Omega_{ij}) + \beta \int_{\Omega_{ij}} |u_h|^q \, dx. \end{aligned} \tag{4.51}$$

To estimate $\int_\Omega |u_h|^q \, dx$, we apply embeddings of $H^1(\Omega)$ to $L^p(\Omega)$ spaces from Remark 4.3. If $\frac{1}{r} + \frac{1}{s} = 1$, then similarly to (59) in [18],

$$\begin{aligned} \int_{\Omega_{ij}} |u_h|^q &\leq \|1\|_{L^s(\Omega_{ij})} \left\| |u_h|_1^q \right\|_{L^r(\Omega_{ij})} \leq meas(\Omega_{ij})^{1/s} \|u_h\|_{L^{qr}(\Omega_{ij})}^q \\ &\leq meas(\Omega_{ij})^{1/s} K_{qr,\Omega}^q |u_h|_1^q. \end{aligned} \tag{4.52}$$

Let $r = d$. If $\frac{1}{d} + \frac{1}{s} = 1$ then $\frac{1}{s} = \frac{d-1}{d}$ in (4.52), and then (4.51) becomes

$$a_{ij}(\bar{\mathbf{c}}) \leq -\frac{\sigma_0}{h^2}\mu_0 \, \mathrm{meas}\,(\Omega_{ij}) + \beta \, meas(\Omega_{ij})^{(d-1)/d} K_{qd,\Omega}^q |u_h|_1^q. \tag{4.53}$$

Let w_0 be as in (4.50), then we prove $|u_h|_1 \leq w_0$ by coercivity. Indeed, the weak formulation in V_h of (4.44) with test function $u_h = v_h$ gives the following:

$$\int_\Omega \mu_0 |\nabla u_h|^2 \, dx + \int_\Omega r(x, u_h, \nabla u_h) u_h^2 \, dx = \int_\Omega f u_h \, dx. \tag{4.54}$$

The l.h.s. of (4.54) gives

$$\int_\Omega \mu_0 |\nabla u_h|^2 \, dx + \int_\Omega r(x, u_h, \nabla u_h) u_h^2 \, dx \geq \mu_0 \int_\Omega |\nabla u_h|^2 \, dx = \mu_0 |u_h|_1^2,$$

since $r \geq 0$ and $u_h^2 \geq 0$. Hence, from equation (4.54),

$$\mu_0 |u_h|_1^2 \leq \int_\Omega f u_h \, dx \leq \|f\|_{L^2(\Omega)} \|u_h\|_{L^2(\Omega)} \leq C_\Omega \|f\|_{L^2(\Omega)} |u_h|_1,$$

by the Poincaré-Friedrichs inequality. From these argument, the upper bound of $|u_h|_1$ is summarized as follows:

$$|u_h|_1 \le \frac{C_\Omega \|f\|_{L^2(\Omega)}}{\mu_0} \le \frac{diam(\Omega)\|f\|_{L^2(\Omega)}}{\mu_0 \pi\sqrt{d}} = w_0 \tag{4.55}$$

since $C_\Omega \le \frac{diam(\Omega)}{\pi\sqrt{d}}$, see in [13]. Using Assumptions 4.5.1-4.5.2 and equation (4.53) and (4.55) we have

$$\begin{aligned} a_{ij}(\bar{\mathbf{c}}) &\le -\frac{\sigma_0}{h^2}\mu_0 \,\mathrm{meas}\,(\Omega_{ij}) + \beta meas(\Omega_{ij})^{d-1/d} K^q_{qd,\Omega} w_0^q \\ &= meas(\Omega_{ij})^{(d-1)/d}\Big(-\frac{\sigma_0}{h^2}\mu_0 \,\mathrm{meas}\,(\Omega_{ij})^{1/d} + \beta K^q_{qd,\Omega} w_0^q \Big) \\ &\le meas(\Omega_{ij})^{(d-1)/d}\Big(-\frac{\sigma_0}{h^2}\mu_0\, c_0^{1/d} h + \beta K^q_{qd,\Omega} w_0^q \Big) = meas(\Omega_{ij})^{(d-1)/d}\Big(-\frac{c_1}{h} + c_2 \Big), \end{aligned}$$

where $c_1 = \mu_0\sigma_0 c_0^{1/d}$ and $c_2 = \beta K^q_{qd,\Omega} w_0^q$. Let

$$b_{ij}(h) = -\frac{c_1}{h} + c_2.$$

Then, for small enough h, we have $b_{ij}(h) \le 0$ and hence $a_{ij}(\bar{\mathbf{c}}) \le 0$. The threshold for h is $h_0 = \frac{c_1}{c_2}$, i.e. $a_{ij}(\bar{\mathbf{c}}) \le 0$ holds for

$$0 < h \le h_0 = \frac{c_1}{c_2} = \frac{\mu_0\sigma_0 c_0^{1/d}}{\beta K^q_{qd,\Omega} w_0^q} \tag{4.56}$$

as was stated. □

Remark 4.5. To estimate $K^q_{qd,\Omega}$ recursively, we can use Lemma 11.2 from [13] which states the following:

$$K^{p_1+p_2}_{p_1+p_2,\Omega} \le \frac{p_1 p_2}{2} K^{p_1-1}_{2(p_1-1),\Omega} K^{p_2-1}_{2(p_2-1),\Omega}. \tag{4.57}$$

In particular, $K_{2,\Omega} = C_\Omega$ is the Poincaré-Friedrichs constant, hence $K_{2,\Omega} \le \frac{diam(\Omega)}{\pi\sqrt{d}}$ as seen in (4.55) above, and from this one can derive a bound for $K_{4,\Omega}$ given below in (4.60).

Example. An example for the problem (4.44) is as follows, involving Stefan-Boltzmann nonlinearity, see (2.27) as rewritten in subsection 2.2.5:

$$\begin{cases} -\Delta u + |u|^3 u = f(x) \quad \text{in } \Omega, \\ \qquad\qquad u = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{4.58}$$

where $\Omega = [0,1]^2$, and we use bilinear FEM on a square mesh. To estimate the mesh size h, we calculate

$$h_0 = \frac{\mu_0\sigma_0 c_0^{1/2}}{\beta K_6^3 w_0^3} \tag{4.59}$$

since $d = 2$ and $q = 3$ in (4.56). To find σ_0, note that for $h = H$ we get $-\frac{1}{6}$ and $-\frac{1}{3}$ in (4.22), hence

$$\int_R \nabla\phi_i \cdot \nabla\phi_j \, dx \le -\frac{1}{6} = -\frac{1}{6}\frac{meas(R)}{h^2}.$$

Then (4.47) is obtained by adding up the above for the rectangles $R \subset \Omega_{ij}$, so the factor is preserved, i.e. $\sigma_0 = \frac{1}{6}$. We have $\mu_0 = 1$, $\beta = 1$, and from (4.50), using $d = 2$ and $diam(\Omega) = \sqrt{2}$,

$$w_0 := \frac{\sqrt{2}\|f\|_{L^2(\Omega)}}{\pi\sqrt{2}} = \frac{\|f\|_{L^2(\Omega)}}{\pi}.$$

We have $c_0^{1/2} = 1$ in Assumption 4.5.2 (ii) since $meas(\Omega_{ij}) \ge meas(R) = h^2$. We can determine K_6^3 using the formula in (4.57). For $p = 4$, we have from (11.10) in [13] that

$$K^4_{4,\Omega} \le \frac{2diam(\Omega)^2}{2\pi^2} = \frac{2}{\pi^2} \tag{4.60}$$

and for $p_1 = p_2 = 3$ in (4.57), using (4.60), we have

$$K^6_{6,\Omega} \le \frac{9}{2}(K^2_{4,\Omega})^2 = \frac{9}{2}K^4_{4,\Omega} \le \frac{9}{\pi^2},$$

hence we get

$$K_{6,\Omega}^3 \le \frac{3}{\pi}.$$

Altogether, we have

$$\mu_0 = 1,\ \sigma_0 = \frac{1}{6},\ c_0^{1/2} = 1,\ \beta = 1,\ K_{6,\Omega}^3 \le \frac{3}{\pi},\ \text{and}\ w_0^3 = \frac{\|f\|_{L^2(\Omega)}^3}{\pi^3}.$$

Therefore the threshold for in (4.56) is given by

$$0 < h \le h_0 = \frac{\pi^4}{18\|f\|_{L^2(\Omega)}^3}.$$

For instance, for the constant source function $f \equiv 10$, we get

$$h_0 = \frac{\pi^4}{18 \cdot 10^3} \approx 0.05.$$

4.5.2. Pure Neumann boundary

Let us consider problem (2.1) when there is only a Neumann boundary, i.e. $\Gamma_N = \partial\Omega$:

$$\begin{cases} -\operatorname{div}\Big(b(x,u,\nabla u)\,\nabla u\Big) + r(x,u,\nabla u)u = f(x) & \text{in } \Omega, \\ b(x,u,\nabla u)\frac{\partial u}{\partial \nu} = \gamma(x) & \text{on } \partial\Omega. \end{cases} \tag{4.61}$$

Assumption 2.1 (a) required $meas(\Gamma_D) > 0$, hence the above case is not covered. The condition $meas(\Gamma_D) > 0$ was used to enable working in the space $H_D^1(\Omega)$ with norm (2.5), and to derive item (ii) of Proposition 4.1. Now we must estimate in $H^1(\Omega)$. To ensure that this proposition still holds, we strengthen Assumption 2.1 such that the condition on r in (2.2) is replaced by

$$0 < \alpha \le r(x,\xi,\eta) \le \beta \qquad \forall (x,\xi,\eta) \in \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^d \tag{4.62}$$

for some constant α. Then, letting $\tilde{\mu} := \min\{\mu_0,\ \alpha\}$, we can complete the proof of (ii) as

$$\begin{aligned} A(\bar{\mathbf{c}})\mathbf{d}\cdot\mathbf{d} &= \int_\Omega \Big(b(x,u_h,\nabla u_h)|\nabla v_h|^2 + r(x,u_h,\nabla u_h)v_h^2 \Big) \\ &\ge \mu_0 \int_\Omega |\nabla v_h|^2 + \alpha \int_\Omega v_h^2 \ge \tilde{\mu} \int_\Omega \big(|\nabla v_h|^2 + v_h^2\big) = \tilde{\mu}\|v_h\|_{H^1}^2 > 0. \end{aligned}$$

All the other results on h_0 for the specific elements then remain true, since the choice of norm does not affect those calculations.

4.5.3. Boundary nonlinearities

The earlier results can be also extended to boundary nonlinearities. We consider the case when the nonlinearity is assumed to appear only on the boundary, such models arise e.g. in localized chemical reactions or radiation problems. We illustrate this with the following 2D model problem, where $\Omega \subset \mathbb{R}^2$ is a bounded domain:

$$\begin{cases} -\mu_0 \Delta u = f(x) & \text{in } \Omega, \\ \mu_0 \frac{\partial u}{\partial \nu} + z(x,u)u = \gamma(x) & \text{on } \Gamma_N, \\ u = g(x) & \text{on } \Gamma_D. \end{cases} \tag{4.63}$$

Assumptions 4.5.3.

(i) $\mu_0 > 0$ is a given constant, $z : \Gamma_N \times \mathbb{R} \to \mathbb{R}$ is continuous and there exists a constant $\eta > 0$ such that

$$0 \le z(x,\xi) \le \eta \qquad \forall (x,\xi) \in \Gamma_N \times \mathbb{R}. \tag{4.64}$$

(ii) The Neumann boundary curve Γ_N is a connected union of some edges of the polygon $\partial\Omega$.

Further, we assume that the used FEM satisfies Assumptions 4.5.2.

Theorem 4.6. *Let Assumptions 4.5.2 and 4.5.3 hold, and let the mesh size h satisfy*

$$0 < h \le h_0 = \frac{6\sigma_0\mu_0 c_0}{\eta}.$$

Then

$$a_{ij}(\bar{\mathbf{c}}) \le 0, \quad i = 1, \dots, n,\ j = 1, \dots, n+m \quad (i \ne j).$$

Consequently, the DMP (4.2) holds.

Proof. Let ϕ_i and ϕ_j be basis functions, $i \ne j$, and denote

$$\Gamma_{N,ij} := \Omega_{ij} \cap \Gamma_N .$$

The entries of the matrix obtained from (4.63) can be estimated using (4.47) and (4.64):

$$a_{ij}(\bar{\mathbf{c}}) = \mu_0 \int_{\Omega_{ij}} \nabla\phi_i \cdot \nabla\phi_j \, dx + \int_{\Gamma_{N,ij}} z(x,u)\phi_i\phi_j \, d\sigma \le -\frac{\sigma_0\mu_0}{h^2} meas(\Omega_{ij}) + \eta \int_{\Gamma_{N,ij}} \phi_i\phi_j \, d\sigma .$$

Here $\Gamma_{N,ij}$ is the segment connecting the nodes P_i and P_j, for which $\phi_i(P_j) = \delta_{ij}$ (cf. condition (B3) in section 2.3). Hence, denoting by h_{ij} the length of $\Gamma_{N,ij}$,

$$\int_{\Gamma_{N,ij}} \phi_i\phi_j \, d\sigma = \int_0^{h_{ij}} \frac{s}{h_{ij}} \frac{h_{ij}-s}{h_{ij}} \, ds = \frac{h_{ij}}{6} \le \frac{h}{6},$$

since $h_{ij} \le h$ where h is the mesh parameter. From this and (4.48) for $d = 2$,

$$a_{ij}(\bar{\mathbf{c}}) \le -\sigma_0\mu_0 c_0 + \frac{\eta}{6}h =: a_{ij}(h).$$

It suffices to ensure $a_{ij}(h) \le 0$, which holds if

$$h \le h_0 := \frac{6\sigma_0\mu_0 c_0}{\eta}. \quad \square$$

Finally, we note that one may combine these results, i.e. similar calculations can be carried out if there is a coefficient r in the PDE and z in the boundary condition as well, furthermore, one may allow $\Gamma_N = \partial\Omega$ if either r or z has a positive lower bound as in (4.62). A simple example of the latter is the Newton law of cooling on the boundary: $\mu_0 \frac{\partial u}{\partial \nu} + \alpha(u - u_0) = 0$, which can be just rewritten as in (4.63) with $z \equiv \alpha$ and $\gamma = \alpha u_0$. This boundary condition coupled with the PDE (4.58) corresponds to the cooling model mentioned in [21].

4.6. Numerical illustration

We illustrate the above theoretical results with numerical experiments, with focus on the nonnegativity, which is the practically most important case of the minimum/maximum principle.

We will observe in the model problems that the nonnegativity can be indeed violated for too coarse meshes, as suggested by our theoretical results and also observed earlier for 1D linear problems in [3], and moreover, our bounds for the threshold mesh size h_0 are not far from the experimental thresholds.

For each of the following four model problems (after their brief description) we give the relevant minimal values in tables. Finally we give a graphical illustration of the violation phenomena.

4.6.1. Michaelis-Menten nonlinearity with bilinear FEM: homogeneous case

First, we consider the 2D reaction-diffusion problem (4.27), containing the rewritten form of the Michaelis-Menten nonlinearity, in the following concrete case:

$$\begin{cases} -\mu_0\Delta u + \frac{u}{1+\epsilon u} = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.65}$$

where $\Omega := [0,1]^2$, $\mu_0 = 10^{-5}$ and $\epsilon = 10^{-3}$ are constants given by Murray, see [21], and $f(x,y) := (2x-1)^6 \ge 0$ describes a source function mostly concentrated near two edges of the square domain.

Table 1
Minima of the FE solutions of (4.65) for some values of h.

h	0.25	0.01	0.0075	0.0074	0.005	0.001
$\min u_h$	-0.0170	-8.3e-11	-8.8e-14	0	0	0

Table 2
Minima of the FE solutions of (4.67) for some values of h.

h	0.25	0.01	0.0078	0.0077	0.005	0.001
$\min u_h$	-0.6157	-0.2689	-9.0e-05	+7.6e-14	+1.6e-11	+4.4e-11

We employ bilinear FEM a uniform square mesh, where the mesh parameter is the edge length h. Since the boundary function is $g = 0$, from (2.14) we expect the discrete nonegativity, in fact

$$\min_{\overline{\Omega}} u_h = 0 \tag{4.66}$$

should hold.

Table 1 illustrates the minima of some FE solutions u_h. It was experienced in the tests that the discrete nonnegativity (4.66) fails (that is, $\min u_h < 0$) when $h \geq 0.0075$, i.e. $n \leq 132$ nodes are used on the edges, whereas (4.66) already holds when $h \leq 0.0074$, i.e. $n \geq 133$ nodes are used.

The results confirm that the numerical solution satisfies the nonnegativity $u_h \geq 0$ only for sufficiently small mesh sizes h. Moreover, let us compare the obtained results with the theoretical sufficient condition. From (4.28) we have $h \leq h_0 = 0.0054$, and in the runs we obtained nonnegative minima for $h \leq 0.0074$, hence the magnitude of the estimation was reasonable. (We note that the theoretical value of h_0 is the same for all right-hand sides f and does not just concern the given one, hence h_0 can be pessimistic for a given problem.)

Qualitatively, the right-hand side of (4.65), which is concentrated near two edges of the square domain, forces the solution to a similar behaviour. The arising jumps seem to induce the numerical solution to violate nonnegativity.

4.6.2. Michaelis-Menten nonlinearity with bilinear FEM: non-homogeneous case

Let us modify the boundary condition of (4.65) to an inhomogeneous one with $g \equiv 1$:

$$\begin{cases} -\mu_0 \Delta u + \frac{u}{1+\epsilon u} = f & \text{in } \Omega := [0,1]^2, \\ u = 1 & \text{on } \partial\Omega, \end{cases} \tag{4.67}$$

and use the same uniform bilinear FEM as above. Since $f \geq 0$ and $g \equiv 1$, we expect the discrete analogue of (2.13), i.e.

$$\min_{\overline{\Omega}} u_h \geq \min\{0, \min_{\partial\Omega} 1\} = 0, \tag{4.68}$$

i.e. discrete nonegativity should hold again. Now we do not expect the equality (4.66), however, away from the two edges we expect $u_h \approx 0$ since $f \approx 0$ there.

Table 2 shows a similar behaviour to Table 1, now the discrete nonnegativity fails above almost the same threshold as before, and it is satisfied below this threshold, i.e. for $h \leq 0.0077$. (The theoretical sufficient condition is the same, $h \leq h_0 = 0.0054$, as in the homogeneous case.)

4.6.3. Exponential nonlinearity for hexagonal bars: prismatic elements

Now we consider a 3D domain discretized by prismatic elements. Our goal is to reproduce the behaviour of the previous two tests, i.e. to define a problem where the experimental threshold is close to the theoretical one h_0. For this we use a uniform mesh where we can easily control the parameters. Further, h_0 does not depend on the r.h.s, but the previous two tests suggest that discrete nonnegativity can fail if f is close or equal to zero an a large part of Ω and has sudden jumps on some portion, enforcing a similar behaviour of the solution if the diffusion is small and thus has no considerable smoothing effect.

Table 3
Minima of the FE solutions of (4.69) for some values of h.

h	0.3333	0.1666	0.1111	0.0833
k	1	2	3	4
DOF_k	57	637	2387	5955
$\min u_h$	-0.0048	-0.0028	-2.1e-11	0

Table 4
Minima of the FE solutions of the 2D version of (4.69) for some values of h.

h	0.3333	0.1666	0.1111	0.0833
k	1	2	3	4
DOF	19	91	217	397
$\min u_h$	-0.0099	-0.0113	-0.0057	0

Based on the above, we consider an analogue of the electrostatic problem (2.26) with proper data:

$$\begin{cases} -\mu_0 \Delta u + e^u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.69}$$

where Ω is a hexagonal bar, i.e. $\Omega = X \times [0, L]$, where X is a regular hexagon with edge length a. We note that hexagonal bars arise in certain applications for their favourable shapes, e.g. such a so-called honeycomb geometry can represent electrostatic precipitators [29]. We set $\mu_0 = 1.85 \cdot 10^{-3}$ and $a = L = 1$. We let f be a source function of the form

$$f = 1 + \varrho, \tag{4.70}$$

where ϱ is concentrated along a portion of the boundary, namely, $\varrho \equiv 0.1$ on $\Omega_\varepsilon = X \times [0, \varepsilon]$ for $\varepsilon = 1/16$, and $\varrho \equiv 0$ on $\Omega \setminus \Omega_\varepsilon$. Now a constant 1 r.h.s would lead to the constant 0 solution, hence $\varrho \geq 0$ implies that the discrete nonnegativity

$$\min_{\overline{\Omega}} u_h = 0$$

from (4.66) is expected again, and the small support of ϱ should induce the behaviour described above.

We employ prismatic elements, using the case described in Remark 4.2 (i). We define an equilateral triangular plane partition for the hexagonal base X of the prisms, denoting by h the edge length of the triangles. The height of the prismatic cells is $H = 3h/4$ based on (4.43). To enable a partition of the given Ω with these mesh parameters, we define an integer $k \in \mathbf{N}^+$ and let $h = 1/(3k)$, $H = 1/(4k)$. The number of unknowns is then

$$DOF_k = \big((3-1)9k + 1\big)(4k - 1).$$

Table 3 shows that discrete nonnegativity fails again until a threshold is attained by h. The computed threshold is obtained from the inequality $h^2 \leq \frac{8\mu_0}{3\beta_0}$, as seen in Remark 4.2 (i), where now $\beta_0 = 1$, hence we get $h_0 = 0.0702$. This is again slightly pessimistic, since we see nonnegativity already for $h = 0.0833$.

4.6.4. *Exponential nonlinearity for hexagonal cross-sections: Courant elements*

Finally we consider a domain being a 2D cross-section of the bar discussed in the previous subsection, and solve the same type of PDE as (4.69)–(4.70). Now ϱ is concentrated in a thin layer along an edge of the boundary. The mesh is also identical to the one used in the hexagonal base above, i.e. an equilateral triangular partition is defined, and Courant elements are used.

Table 4 shows similar conclusions as above. We find that discrete nonnegativity fails again until a threshold is attained by h. The computed threshold from (4.11) is $h_0 = \sqrt{6\mu_0}$: now we set $\mu_0 = 10^{-3}$, hence $h_0 = 0.0774$, which is not far from the experimental value where nonnegativity already holds.

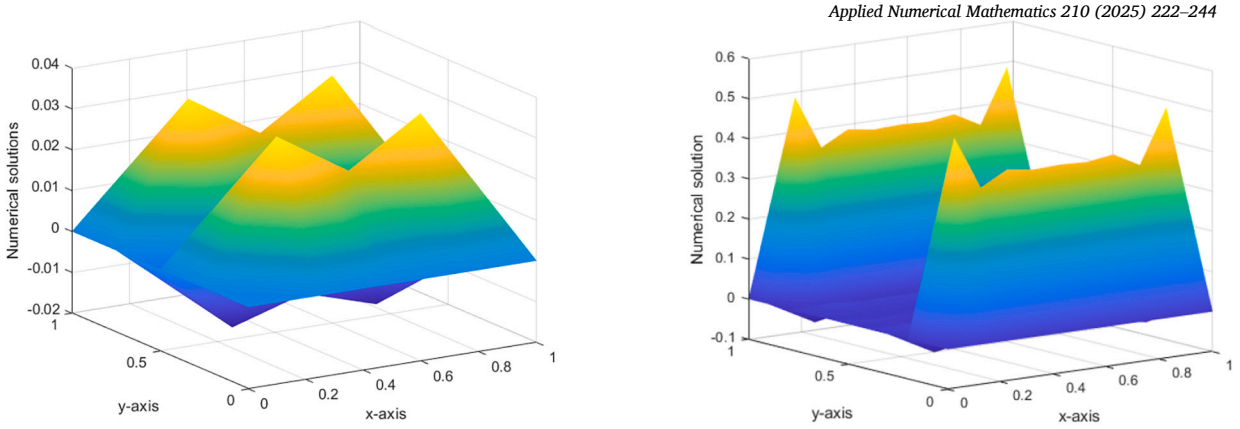

Fig. 3. Left: $h = 0.25$, $\min u_h = -0.0170$. Right: $h = 0.01$, $\min u_h = -0.0421$.

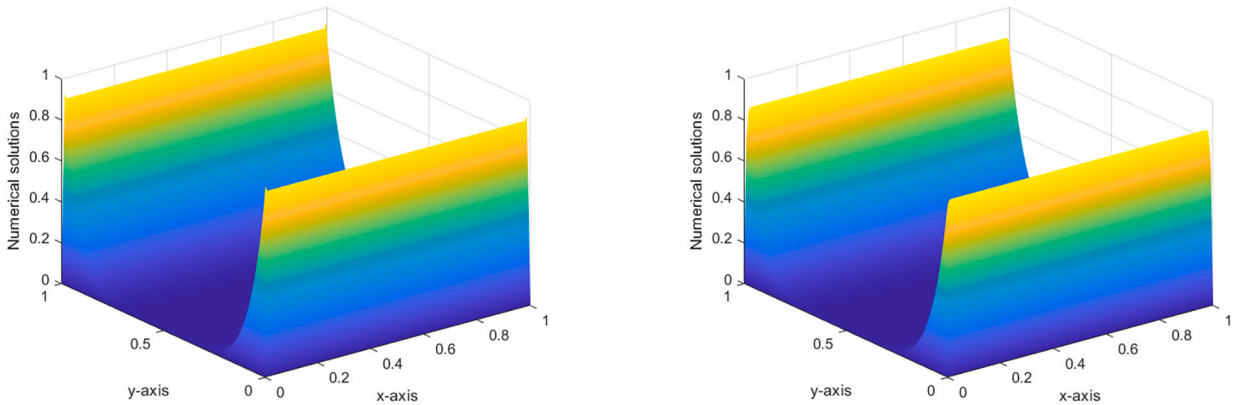

Fig. 4. Left: $h = 0.0075$, $\min u_h = -8.8e-14$. Right: $h = 0.0074$, $\min u_h = 0$.

4.6.5. *Graphical illustration of the mesh sensitivity*

For an illustration we enclose the graphs of the numerical solutions for the Michaelis-Menten problems with bilinear FEM from subsections 4.6.1–4.6.2 for some mesh sizes.

(a) Homogeneous case. The solution has large values in layers concentrated near two edges of the square domain, but it is close to zero in the remaining main part of Ω. The arising jumps induce that the approximation u_h on a coarse mesh violates nonnegativity with an oscillatory behaviour, since the average of u_h should be close to 0 (Figs. 3, 4).

(b) Inhomogeneous case. The solution has large values on the whole boundary and on neighbouring layers, but again it is close to zero in the remaining main part of Ω and thus provides jumps. Hence the violation of nonnegativity on coarse meshes can have the same explanation as in item (a) (Figs. 5, 6).

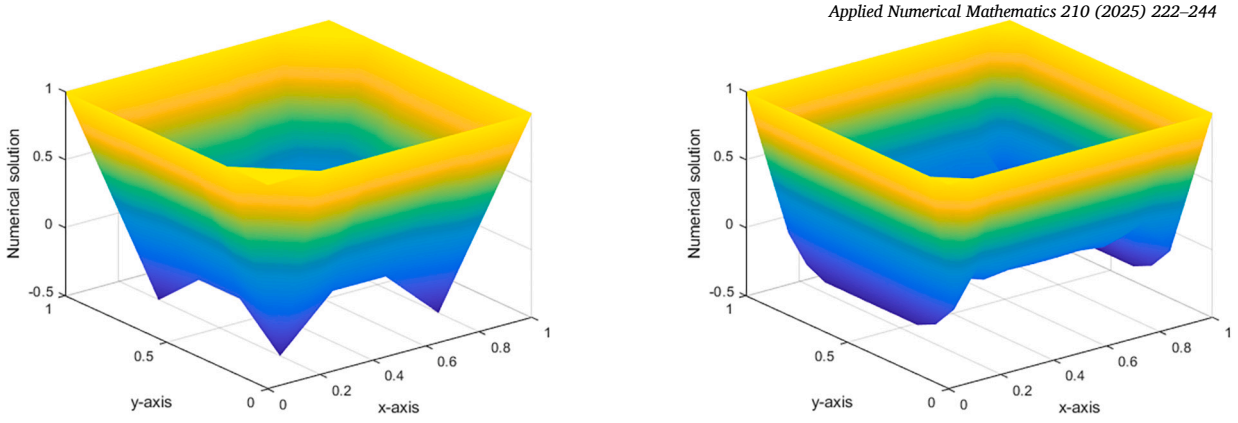

Fig. 5. Left: $h = 0.25$, $\min u_h = -0.6157$. Right: $h = 0.01$, $\min u_h = -0.2689$.

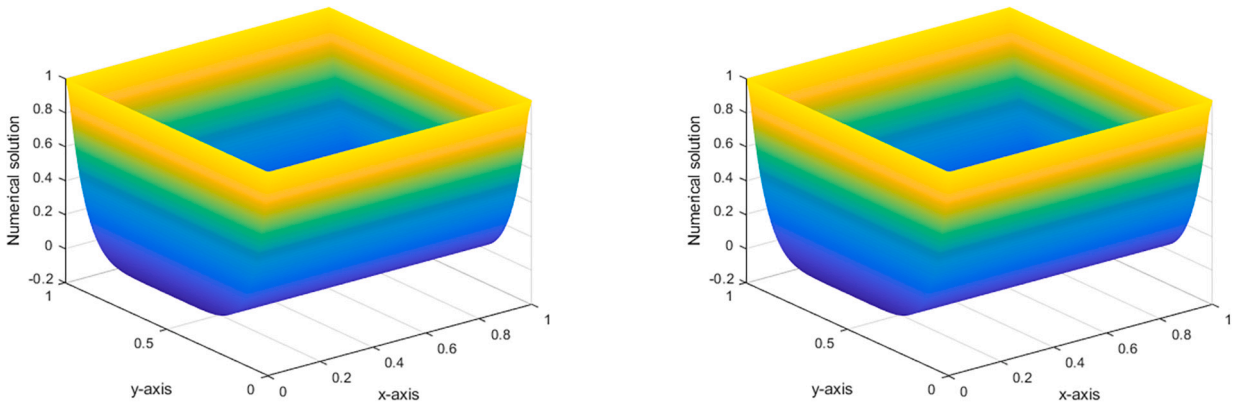

Fig. 6. Left: $h = 0.0078$, $\min u_h = -2.5e-04$. Right: $h = 0.0077$, $\min u_h = +7.1e-14$.

CRediT authorship contribution statement

M.T. Bahlibi: Formal analysis, Software. **J. Karátson:** Conceptualization, Formal analysis, Validation. **S. Korotov:** Investigation, Validation.

Acknowledgements

The research of J. Karátson is supported by the Hungarian National Research, Development and Innovation Fund (NKFIH), under the funding scheme ELTE TKP 2021-NKTA-62 and the grant no. K137699.

References

[1] R.A. Adams, J.F. Fournier, Sobolev Spaces, Elsevier–Academic Press, 2003.
[2] M.T. Bahlibi, Discrete maximum principles for the finite element solution of some nonlinear elliptic problems, Ann. Univ. Sci. Bp. Rolando Eötvös Nomin., Sect. Comput. 57 (2024) 55–68.
[3] J. Brandts, S. Korotov, M. Křížek, The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem, Linear Algebra Appl. 429 (2008) 2344–2357.
[4] E. Burman, A. Ern, Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence, Math. Comput. 74 (2005) 1637–1652.
[5] P.G. Ciarlet, Discrete maximum principle for finite-difference operators, Aequ. Math. 4 (1970) 338–352.
[6] A. Drăgănescu, T.F. Dupont, L.R. Scott, Failure of the discrete maximum principle for an elliptic finite element problem, Math. Comput. 74 (249) (2005) 1–23.
[7] D. Eppstein, J.M. Sullivan, A. Üngör, Tiling space and slabs with acute tetrahedra, Comput. Geom. 27 (3) (2004) 237–255.
[8] I. Faragó, Matrix and discrete maximum principles, in: I. Lirkov, S. Margenov, J. Waśniewski (Eds.), Large-Scale Scientific Computing, LSSC 2009, in: Lecture Notes in Computer Science, vol. 5910, Springer, 2010, pp. 563–570.
[9] I. Faragó, R. Horváth, Continuous and discrete parabolic operators and their qualitative properties, IMA J. Numer. Anal. 29 (3) (2009) 606–631.
[10] I. Faragó, R. Horváth, A review of reliable numerical models for three-dimensional linear parabolic problems, Int. J. Numer. Methods Eng. 70 (1) (2007) 25–45.
[11] I. Faragó, R. Horváth, S. Korotov, Discrete maximum principle for linear parabolic problems solved on hybrid meshes, Appl. Numer. Math. 53 (2–4) (2005) 249–264.

[12] I. Faragó, R. Horváth, S. Korotov, Discrete maximum principles for FE solutions of nonstationary diffusion-reaction problems with mixed boundary conditions, Numer. Methods Partial Differ. Equ. 27 (3) (2011) 702–720.
[13] I. Faragó, J. Karátson, Numerical Solution of Nonlinear Elliptic Problems via Preconditioning Operators: Theory and Applications, Adv. Comput. Theory Pract., vol. 11, Nova Science Publishers, 2002.
[14] A. Hannukainen, S. Korotov, T. Vejchodský, Discrete maximum principle for FE-solutions of the diffusion-reaction problem on prismatic meshes, J. Comput. Appl. Math. 226 (2009) 275–287.
[15] A. Harbi, Maximum norm analysis of a nonmatching grids method for a class of variational inequalities with nonlinear source terms, J. Inequal. Appl. (2016) 181.
[16] B. Hingyi, J. Karátson, Detection of dead cores for reaction-diffusion equations with a non-smooth nonlinearity, Appl. Numer. Math. 177 (2022) 111–122.
[17] I. Hlaváček, M. Křížek, J. Malý, On Galerkin approximations of a quasilinear nonpotential elliptic problem of a nonmonotone type, J. Math. Anal. Appl. 184 (1) (1994) 168–189.
[18] J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, Numer. Math. 99 (2005) 669–698.
[19] J. Karátson, S. Korotov, Discrete maximum principles for FEM solutions of some nonlinear elliptic interface problems, Int. J. Numer. Anal. Model. 6 (1) (2009) 1–16.
[20] J. Karátson, S. Korotov, Some discrete maximum principles arising for nonlinear elliptic finite element problems, Comput. Math. Appl. 70 (2015) 2732–2741.
[21] H.B. Keller, Elliptic boundary value problems suggested by nonlinear diffusion processes, Arch. Ration. Mech. Anal. 35 (1969) 363–381.
[22] S. Korotov, M. Křižek, Finite element analysis of variational crimes for a quasilinear elliptic problem in 3D, Numer. Math. 84 (2000) 549–576.
[23] M. Křižek, P. Neittaanmäki, Mathematical and Numerical Modelling in Electrical Engineering: Theory and Applications, Kluwer Academic Publishers, 1996.
[24] S. Pollock, Y. Zhu, Discrete comparison principles for quasilinear elliptic PDE, Appl. Numer. Math. 156 (2020) 106–124.
[25] M.H. Protter, H.F. Weinberger, Maximum Principles in Differential Equations, Springer-Verlag, New York, 1984.
[26] P. Pucci, J.B. Serrin, The Maximum Principle, Springer, 2007.
[27] T. Vejchodsky, The discrete maximum principle for Galerkin solutions of elliptic problems, Cent. Eur. J. Math. 10 (1) (2012) 25–43.
[28] T. Vejchodský, P. Solin, Discrete maximum principle for higher-order finite elements in 1D, Math. Comput. 76 (260) (2007) 1833–1846.
[29] Y. Wang, et al., Insights into the role of ionic wind in honeycomb electrostatic precipitators, J. Aerosol Sci. 133 (2019) 83–95.
[30] E. Zeidler, Nonlinear Functional Analysis and Its Applications, II/B: Nonlinear Monotone Operators, Springer-Verlag, New York, 1990.