Helsinki University of Technology, Institute of Mathematics, Research Reports
Teknillisen korkeakoulun matematiikan laitoksen tutkimusraporttisarja
Espoo 2006 A510

Discrete maximum principles for FEM solutions of some nonlinear elliptic interface problems

János Karátson **Sergey Korotov**

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Helsinki University of Technology, Institute of Mathematics, Research Reports
Teknillisen korkeakoulun matematiikan laitoksen tutkimusraporttisarja
Espoo 2006 A510

Discrete maximum principles for FEM solutions of some nonlinear elliptic interface problems

János Karátson **Sergey Korotov**

Helsinki University of Technology
Department of Engineering Physics and Mathematics
Institute of Mathematics

János Karátson and Sergey Korotov: *Discrete maximum principles for FEM solutions of some nonlinear elliptic interface problems*; Helsinki University of Technology, Institute of Mathematics, Research Reports A510 (2006).

Abstract: *Discrete maximum principles are proved for finite element discretizations of nonlinear elliptic interface problems with jumps of the normal derivatives. The geometric conditions in the case of simplicial meshes are suitable acuteness or nonobtuseness properties.*

AMS subject classifications: 35B50, 35J65, 65N30, 65N50.

Keywords: nonlinear elliptic problem, interface problem, maximum principle, discrete maximum principle, finite element method, simplicial mesh.

Correspondence

karatson@cs.elte.hu, sergey.korotov@hut.fi

ISBN-13 978-951-22-8445-0
ISBN-10 951-22-8445-6

Helsinki University of Technology
Department of Engineering Physics and Mathematics
Institute of Mathematics
P.O. Box 1100, 02015 HUT, Finland
email:math@hut.fi http://www.math.hut.fi/

1 Introduction

The maximum principle forms an important qualitative property of second order elliptic boundary value problems [10, 23, 27]. Consequently, the discrete analogues of the maximum principle (so-called discrete maximum principles, DMPs) have drawn much attention. Various DMPs have been formulated and proved including the case of finite difference, finite volume and finite element approximations, and corresponding geometric conditions on the computational meshes have been given, see, e.g., [2, 4, 6, 7, 11, 19, 28] for linear and [14, 15, 20] for nonlinear problems with standard (i.e., Dirichlet, and in [14, 15] mixed) boundary conditions.

In this paper we address interface problems, which arise in various branches of material science, biochemistry, multiphase flow etc., often when two distinct materials are involved with different conductivities or densities. Another (for our work, motivating) example is from localized reaction-diffusion problems [12, 13], see at the end of this paper. Many special numerical methods have been designed for interface problems, see e.g. [12, 25, 26, 24], but maximum principles have received less attention than for the case of standard boundary value problems. A continuous minimum principle for a related problem is given in [9]. The discrete maximum principle for suitable finite difference discretizations of linear interface problems has been proved in [25].

Our goal is to prove discrete maximum principles for nonlinear elliptic interface problems when finite element discretization is involved. The present paper is the extension of our paper [14] to a class of such problems, and relies on a similar technique using weak formulation and positivity conditions that ensure well-posedness. We consider matching conditions for the solution itself on the interface, i.e., the jump is allowed for the normal derivatives. Problems with jump of the solution or without well-posedness may be the subject of further research.

The paper is organized as follows. The formulation of the problem, together with the derivation of a continuous maximum principle, and the description of finite element discretization are given in Section 2. Discrete maximum principles are proved and examples are given in Section 3.

2 Nonlinear elliptic interface problems: basic properties and discretization

2.1 Formulation of the problem

We investigate nonlinear interface problems of the following form:

$$\begin{cases} -\operatorname{div}\Big(b(x,\nabla u)\,\nabla u\Big) + q(x,u) \;=\; f(x) & \text{in } \Omega\setminus\Gamma, \\ [\,u\,]_\Gamma \;=\; 0 & \text{on } \Gamma, \\ \big[\,b(x,\nabla u)\frac{\partial u}{\partial \nu}\big]_\Gamma + s(x,u) \;=\; \gamma(x) & \text{on } \Gamma, \\ u \;=\; g(x) & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $\partial\Omega$ denotes the boundary of the domain Ω and the interface Γ is a surface lying in Ω, further, $[\,u\,]_\Gamma$ and $\big[\,b(x,\nabla u)\frac{\partial u}{\partial \nu}\big]_\Gamma$ denote the jump (i.e., the difference of the limits from the two sides of the interface Γ) of u and $b(x,\nabla u)\frac{\partial u}{\partial \nu}$, respectively. We impose the following

Assumptions 2.1:

(A1) Ω is a bounded open domain in $\mathbf{R}^d$, the interface $\Gamma \subset \Omega$ and the boundary $\partial\Omega$ are piecewise smooth and Lipschitz continuous $(d-1)$-dimensional surfaces.

(A2) The scalar functions $b : \Omega \times \mathbf{R}^d \to \mathbf{R}$, $q : \Omega \times \mathbf{R} \to \mathbf{R}$ and $s : \Gamma \times \mathbf{R} \to \mathbf{R}$ are measurable and bounded w.r.t. their first variable $x \in \Omega$ (resp. $x \in \Gamma$) and continuously differentiable w.r.t. their second variable $\eta \in \mathbf{R}^d$ (resp. $\xi \in \mathbf{R}$). Further, $f \in L^2(\Omega)$, $\gamma \in L^2(\Gamma)$ and $g \in H^1(\Omega)$.

(A3) The function b satisfies

$$0 < \mu_0 \le b(x,\eta) \le \mu_1 \tag{2}$$

with positive constants μ_0 and μ_1 independent of (x,η), further, the diadic product matrix $\eta \cdot \frac{\partial b(x,\eta)}{\partial \eta}$ is symmetric positive semidefinite and bounded in matrix norm by some positive constant μ_2 independent of (x,η).

(A4) Let $2 \le p_1$ if $d = 2$, or $2 \le p_1 \le \frac{2d}{d-2}$ if $d > 2$, further, let $2 \le p_2$ if $d = 2$, or $2 \le p_2 \le \frac{2d-2}{d-2}$ if $d > 2$. There exist functions $\alpha_1 \in L^{d/2}(\Omega)$, $\alpha_2 \in L^{d-1}(\Gamma)$ and a constant $\beta \ge 0$ such that for any $x \in \Omega$ (or $x \in \Gamma$, resp.) and $\xi \in \mathbf{R}$

$$0 \le \frac{\partial q(x,\xi)}{\partial \xi} \le \alpha_1(x) + \beta|\xi|^{p_1-2}, \qquad 0 \le \frac{\partial s(x,\xi)}{\partial \xi} \le \alpha_2(x) + \beta|\xi|^{p_2-2}. \tag{3}$$

Remark 1. The role of assumption (A3) is to ensure that the Jacobian matrices $J(x,\eta) := \frac{\partial}{\partial \eta}\Big(b(x,\eta)\,\eta\Big)$ are symmetric and satisfy the uniform ellipticity property $\mu_0|\zeta|^2 \le J(x,\eta)\;\zeta\cdot\zeta \le \mu_3|\zeta|^2$, $\zeta \in \mathbf{R}^d$ (with $\mu_3 = \mu_1 + \mu_2$), which will be required for well-posedness. For instance, assumption (A3) holds for coefficients of the form $b(x,\eta) = a(x,|\eta|)$ (see [8, 21] for such nonlinearities), where the C^1 function $a : \Omega \times \mathbf{R}^+ \to \mathbf{R}$ satisfies

$0 < \mu_0 \le a(x,r) \le \frac{\partial}{\partial r}(a(x,r)\,r) \le \mu_3 \ \ (r > 0)$. More specially, one may have $b(x,\eta) = a(x)$ (i.e., linear principal part) with a measurable function a satisfying $0 < \mu_0 \le a(x) \le \mu_3$.

2.2 Weak solutions

Theorem 1. *Under Assumptions 2.1, problem (1) has a unique weak solution $u^* \in H^1(\Omega)$ defined as follows:*

$$\int_\Omega \Big(b(x,\nabla u^*)\ \nabla u^* \cdot \nabla v + q(x,u^*)v \Big)\, dx \ + \int_\Gamma s(x,u^*)v\, d\sigma =$$

$$= \int_\Omega f v\, dx \ + \int_\Gamma \gamma v\, d\sigma \qquad \forall v \in H_0^1(\Omega) \tag{4}$$

$$\textit{and} \quad u^* \ = \ g \quad \text{on } \partial\Omega. \tag{5}$$

Proof. We first prove the theorem for homogeneous boundary condition, i.e. when $g = 0$. In this case the weak solution u^* can be obtained using monotone operators, in a similar way as in [8, Chap. 6], therefore we only indicate the main steps of the proof. First, we define

$$\langle F(u), v\rangle \ = \ \int_\Omega \Big(b(x,\nabla u)\ \nabla u \cdot \nabla v + q(x,u)v - fv \Big)\, dx +$$

$$+ \int_\Gamma \Big(s(x,u)v - \gamma v \Big)\, d\sigma \qquad (v \in H_0^1(\Omega)), \tag{6}$$

where the growth conditions in (A1)–(A4) ensure that the arising integrals are finite. Let $J(x,\eta) := \frac{\partial}{\partial \eta}\Big(b(x,\eta)\,\eta\Big)$ as in Remark 1. Then, from (A3)-(A4), we obtain that the Gateaux derivative $F'(u)$ exists, is self-adjoint for all u and satisfies

$$\langle F'(u)v, v\rangle \ = \ \int_\Omega \Big(J(x,\nabla u)\ \nabla v \cdot \nabla v + q'_u(x,u)v^2 \Big)\, dx +$$

$$+ \int_\Gamma s'_u(x,u)v^2\, d\sigma \ge \mu_0 \int_\Omega |\nabla v|^2\, dx \tag{7}$$

(for all $u, v \in H_0^1(\Omega)$), where q'_u, s'_u denote derivatives w.r.t. u. Using the standard Sobolev norm defined via

$$\|v\|_1^2 = \int_\Omega |\nabla v|^2\, dx \qquad (v \in H_0^1(\Omega)),$$

the uniform ellipticity (7) implies that the operator equation $F(u) = 0$ has a unique solution $u^* \in H_0^1(\Omega)$. Here $F(u^*) = 0$ is equivalent to (4), i.e., u^* is the weak solution.

For non-homogeneous boundary conditions the problem can be reduced to the homogeneous case using a usual translation. Let $g \in H^1(\Omega)$ be arbitrary and let us require (5) on the boundary. Then we look for u^* in the form $u^* = u + g$, in which case $u = 0$ on $\partial\Omega$. Substituting this sum into (4), we observe that u must satisfy the same problem with homogeneous boundary conditions and with coefficients

$$\hat{b}(x,\eta) = b(x,\eta+\nabla g(x)), \quad \hat{q}(x,\xi) = q(x,\xi+g(x)), \quad \hat{s}(x,\xi) = s(x,\xi+g(x)).$$

Here $g(x)$ is independent of ξ, η, hence these coefficients remain C^1 in their second variable and satisfy the same growth conditions as b, q, s. This implies existence and uniqueness for u, and then the same for u^* owing to the relation $u^* = u + g$. ∎

The above notion of weak solution is justified by showing that any classical (or strong) solution is also a weak solution. To define the classical solution, we assume in addition that the interface Γ is a closed surface, or more generally, it is any compact subset of an (also piecewise smooth and Lipschitz continuous) closed surface $\hat{\Gamma} \subset \Omega$ as illustrated in Figure 2.2. Let us denote by Ω_0 the interior of the surface $\hat{\Gamma}$.

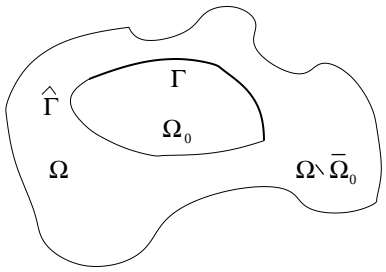

Figure 1: Interface in a domain

Definition 1. We call $u : \overline{\Omega} \to \mathbf{R}$ a *classical solution of problem (1)* if $u \in C^2(\Omega \setminus \Gamma)$, $u\big|_{\overline{\Omega}_0} \in C^1(\overline{\Omega}_0)$, $u\big|_{\overline{\Omega\setminus\Omega}_0} \in C^1(\overline{\Omega \setminus \Omega}_0)$ and u satisfies (1) pointwise.

Proposition 1. *A classical solution of problem (1) is also a weak solution.*

Proof. Let $x \in \hat{\Gamma}$, let ν denote the normal unit vector pointing out of Ω_0, and let $\hat{\nu} := -\nu$ (the normal unit vector pointing out of $\Omega \setminus \Omega_0$). The jump of $b(x,\nabla u)\frac{\partial u}{\partial \nu}$ at x is the difference of the limits of $b(\cdot,\nabla u)\frac{\partial u}{\partial \nu}$ at x from Ω_0 and from $\Omega \setminus \Omega_0$. Using the definition $\frac{\partial u}{\partial \nu}(x) := \lim\limits_{t\to 0^+} \frac{1}{t}\Big(u(x) - u(x-t\nu)\Big)$, we thus have

$$\left[b(x,\nabla u)\frac{\partial u}{\partial \nu}\right]_{x\in\Gamma} := b\Big(x, \nabla(u\big|_{\overline{\Omega}_0})(x)\Big) \lim_{t\to 0^+} \frac{1}{t}\Big(u(x) - u(x-t\nu)\Big) -$$

$$
\begin{aligned}
&- b\Big(x, \nabla(u|_{\overline{\Omega\setminus\Omega_0}})(x)\Big) \lim_{t\to 0^-} \frac{1}{t}\Big(u(x) - u(x - t\nu)\Big) = \\
&= b\Big(x, \nabla(u|_{\overline{\Omega}_0})(x)\Big) \lim_{t\to 0^+} \frac{1}{t}\Big(u(x) - u(x - t\nu)\Big) + \\
&+ b\Big(x, \nabla(u|_{\overline{\Omega\setminus\Omega_0}})(x)\Big) \lim_{s\to 0^+} \frac{1}{s}\Big(u(x) - u(x - s\hat{\nu})\Big) = \\
&= \left(b\Big(x, \nabla(u|_{\overline{\Omega}_0})(x)\Big)\, \frac{\partial u}{\partial \nu} + b\Big(x, \nabla(u|_{\overline{\Omega\setminus\Omega_0}})(x)\Big)\, \frac{\partial u}{\partial \hat{\nu}}\right)_{x\in\Gamma}. \qquad (8)
\end{aligned}
$$

Now let u be a classical solution. The assumptions imply that $u \in H^1(\Omega)$, and (5) holds trivially. For any $v \in H^1_0(\Omega)$, Green's formula for equation (1) on Ω_0 and $\Omega \setminus \Omega_0$, respectively, yields

$$
\int_{\Omega_0} fv\,dx = \int_{\Omega_0} \Big(b(x,\nabla u)\ \nabla u\cdot\nabla v + q(x,u)v\Big)\,dx - \int_{\hat{\Gamma}} b\Big(x, \nabla(u|_{\overline{\Omega}_0})\Big)\frac{\partial u}{\partial \nu}\, v\,d\sigma
$$

and

$$
\int_{\Omega\setminus\Omega_0} fv\,dx = \int_{\Omega\setminus\Omega_0} \Big(b(x,\nabla u)\ \nabla u\cdot\nabla v + q(x,u)v\Big)\,dx - \int_{\hat{\Gamma}} b\Big(x, \nabla(u|_{\overline{\Omega\setminus\Omega_0}})\Big)\frac{\partial u}{\partial \hat{\nu}}\, v\,d\sigma.
$$

Summing up, the integrand on $\hat{\Gamma}$ becomes the jump on Γ (using (8)) and zero on $\hat{\Gamma}\setminus\Gamma$ (since ∇u is continuous there and $\hat{\nu} = -\nu$). In virtue of the jump condition in (1), we altogether obtain

$$
\begin{aligned}
\int_{\Omega} fv\,dx &= \int_{\Omega} \Big(b(x,\nabla u)\ \nabla u\cdot\nabla v + q(x,u)v\Big)\,dx - \int_{\Gamma} \left[b(x,\nabla u)\frac{\partial u}{\partial \nu}\right]_{\Gamma} v\,d\sigma \\
&= \int_{\Omega} \Big(b(x,\nabla u)\ \nabla u\cdot\nabla v + q(x,u)v\Big)\,dx + \int_{\Gamma} \Big((s(x,u) - \gamma\Big)\,v\,d\sigma\,. \qquad \blacksquare
\end{aligned}
$$

2.3 Continuous maximum principles

We formulate and prove two continuous maximum principles for our PDE problem (1). These statements provide the properties whose discrete analogues can be expected for suitable FEM solutions.

Theorem 2. *Let Assumptions 2.1 hold and*

$$
f(x) - q(x,0) \le 0,\ x\in\Omega, \quad \text{and} \quad \gamma(x) - s(x,0) \le 0,\ x\in\Gamma. \qquad (9)
$$

If the weak solution u of problem (1) belongs to $C^1(\Omega\setminus\Gamma)\cap C(\overline{\Omega})$, then

$$
\max_{\overline{\Omega}} u \le \max\{0, \max_{\partial\Omega} g\}. \qquad (10)
$$

In particular, if $g \ge 0$, then $\max_{\overline{\Omega}} u = \max_{\partial\Omega} g$, and if $g \le 0$, then we have the nonpositivity property $\max_{\overline{\Omega}} u \le 0$.

In general, if $u \in H^1(\Omega)$ only (without the above regularity assumption) and g is a.e. bounded on $\partial\Omega$, then the same statements hold if $\max u$ and $\max g$ are replaced by $\operatorname{ess\,sup} u$ and $\operatorname{ess\,sup} g$, respectively.

PROOF. We only prove the regular case, the general case is similar (if $\max u$ and $\max g$ are replaced by $\operatorname{ess\,sup} u$ and $\operatorname{ess\,sup} g$, respectively). Let

$$r(x,\xi) := \begin{cases} \frac{q(x,\xi)-q(x,0)}{\xi}, & \text{if } \xi \neq 0, \\ \frac{\partial q}{\partial \xi}(x,0), & \text{if } \xi = 0, \end{cases} \qquad z(x,\xi) := \begin{cases} \frac{s(x,\xi)-s(x,0)}{\xi}, & \text{if } \xi \neq 0, \\ \frac{\partial s}{\partial \xi}(x,0), & \text{if } \xi = 0. \end{cases} \tag{11}$$

Here, using (A2), the functions r and z are continuous in ξ. Further, in view of (A4), we have

$$r(x,\xi) \geq 0, \qquad z(x,\xi) \geq 0. \tag{12}$$

We define

$$\tilde{a}(x) := b(x, \nabla u(x)) \quad (x \in \Omega \setminus \Gamma), \qquad \tilde{h}(x) := r(x, u(x)) \quad (x \in \Omega), \tag{13}$$

$$\tilde{k}(x) := z(x, u(x)) \qquad (x \in \Gamma).$$

Using also the notations

$$\hat{f}(x) := f(x) - q(x,0) \qquad \text{and} \quad \hat{\gamma}(x) := \gamma(x) - s(x,0), \tag{14}$$

the weak formulation of problem (1) is rewritten as

$$\int_\Omega \left(\tilde{a}\ \nabla u \cdot \nabla v + \tilde{h} u v\right) dx\ + \int_\Gamma \tilde{k} u v\, d\sigma = \int_\Omega \hat{f} v\, dx\ + \int_\Gamma \hat{\gamma} v\, d\sigma \qquad \forall v \in H_0^1(\Omega). \tag{15}$$

Let $M := \max\{0, \max_{\partial\Omega} g\}$ and we introduce the piecewise C^1 function $v := \max\{u - M,\, 0\}$. Then we have $v \geq 0$ and $v_{|\partial\Omega} = 0$, further, $u(x) = v(x) + M$ for any $x \in \overline{\Omega}$ unless $v(x) = 0$. Hence, for this v the left-hand side of (15) satisfies

$$\int_\Omega \left(\tilde{a}\ \nabla u \cdot \nabla v + \tilde{h} u v\right) dx\ + \int_\Gamma \tilde{k} u v\, d\sigma =$$

$$= \int_\Omega \left(\tilde{a}\, |\nabla v|^2 + \tilde{h}(v+M)v\right) dx\ + \int_\Gamma \tilde{k}(v+M)v\, d\sigma \geq 0$$

since the functions $\tilde{a}, \tilde{h}, \tilde{k}, v$ and the constant M are nonnegative. On the other hand, the assumptions $\hat{f} \leq 0$, $\hat{\gamma} \leq 0$ imply that for this v the right-hand side of (15) satisfies

$$\int_\Omega \hat{f} v\, dx\ + \int_\Gamma \hat{\gamma} v\, d\sigma \leq 0,$$

which together imply the relation

$$\int_\Omega \left(\tilde{a}\, |\nabla v|^2 + \tilde{h}(v+M)v\right) dx\ + \int_\Gamma \tilde{k}(v+M)v\, d\sigma = 0.$$

By assumption (A3), here $\tilde{a}$ has a positive minimum, hence $|\nabla v| = 0$, i.e., v is constant. We have seen that $v_{|\partial\Omega} = 0$, hence we obtain that $v \equiv 0$, which just means that (10) holds. ∎

The following special case provides equality of maxima on $\partial\Omega$ without assuming $g \geq 0$:

Theorem 3. *Let $q \equiv 0$ and $s \equiv 0$ in problem (1). Let us impose the assumptions of Theorem 2, which now means that (A1)–(A3) are satisfied, $u \in C^1(\Omega \setminus \Gamma) \cap C(\overline{\Omega})$, and (9) takes the form*

$$f(x) \leq 0,\ x \in \Omega \quad \textit{and} \quad \gamma(x) \leq 0,\ x \in \Gamma. \tag{16}$$

Then

$$\max_{\overline{\Omega}} u = \max_{\partial\Omega} g. \tag{17}$$

(If $u \in H^1(\Omega)$ only and g is a.e. bounded on $\partial\Omega$, then $\operatorname{ess\,sup} u = \operatorname{ess\,sup} g$ on $\partial\Omega$.)

Proof. We only prove the regular case again. If $\max_{\partial\Omega} g \geq 0$ then (10) implies (17). Let $\max_{\partial\Omega} g < 0$, say, $\max_{\partial\Omega} g = -K$ with some $K > 0$. Then the function $w := u + K$ satisfies the same mixed problem with right-hand sides f, γ and $g + K$, respectively, hence Theorem 2 is valid for this problem as well, and (10) for w yields $\max_{\overline{\Omega}} w \leq \max\{0, \max_{\partial\Omega}(g + K)\} = 0$. Then $\max_{\overline{\Omega}} u \leq -K = \max_{\partial\Omega} g$. ∎

Remark 2. Analogously to Theorems 2 and 3, corresponding minimum principles and nonnegativity property hold if the sign conditions in (9) and (16) are reversed.

2.4 Finite element discretization

Our basic assumption in the sequel that Ω is a polytopic domain and the interface Γ is also polytopic. (We note that if $\partial\Omega$ or Γ are curved then the convergence of the discrete solution to the exact one is a much more difficult problem, out of the scope of this paper. Even for the simpler case of Dirichlet problems in 3D without interface, such an analysis has been given only recently in [16].)

We introduce a finite element discretization of our problem with simplicial elements and continuous piecewise linear basis functions. Let $\mathcal{T}_h$ be a a conforming triangulation of $\overline{\Omega}$ into tetrahedra, whose nodes are $B_1, ..., B_{\bar{n}}$. Denote by $\phi_1, ..., \phi_{\bar{n}}$ the piecewise linear continuous basis functions defined in a standard way, i.e., $\phi_i(B_j) = \delta_{ij}$ for $i, j = 1, ..., \bar{n}$, where δ_{ij} is the Kronecker symbol. Let V_h denote the finite element subspace spanned by the above basis functions:

$$V_h = \operatorname{span}\{\phi_1, ..., \phi_{\bar{n}}\} \subset H^1(\Omega).$$

Let $n < \bar{n}$ be such that

$$B_1, ..., B_n \tag{18}$$

are the nodes that lie in Ω and let

$$B_{n+1}, ..., B_{\bar{n}} \tag{19}$$

be the nodes that lie on $\partial\Omega$. Then the basis functions $\phi_1, ..., \phi_n$ satisfy homogeneous boundary condition on $\partial\Omega$, i.e., $\phi_i \in H_0^1(\Omega)$. We define

$$V_h^0 = \text{span}\{\phi_1, ..., \phi_n\} \ \subset H_0^1(\Omega).$$

Further, let

$$g_h = \sum_{j=n+1}^{\bar{n}} g_j \phi_j \ \in V_h \tag{20}$$

(with $g_j \in \mathbf{R}$) be the piecewise linear approximation of the function g on $\partial\Omega$ (and on the neighbouring elements). To find the FEM solution of (4)-(5) in V_h, we solve the following problem: find $u_h \in V_h$ such that

$$\int_\Omega \Big(b(x, \nabla u_h) \ \nabla u_h \cdot \nabla v_h + \ q(x, u_h) v_h \Big) \, dx \ + \int_\Gamma s(x, u_h) v_h \, d\sigma =$$

$$= \int_\Omega f v_h \, dx \ + \int_\Gamma \gamma v_h \, d\sigma \quad \forall v_h \in V_h^0 \tag{21}$$

$$\text{and} \qquad u_h \ = \ g_h \quad \text{on } \partial\Omega.$$

Theorem 4. *Under Assumptions 2.1, problem (21) has a unique solution $u_h \in V_h$, and $\|u^* - u_h\|_1 \to 0$ as $h \to 0$.*

Proof. The proof of Theorem 1 can be repeated to obtain u_h, just replacing $H^1(\Omega)$ by V_h. The convergence of u_h to u^* in H^1-norm follows in the standard way from the ellipticity of the equation and the fact that the finite-dimesional subspaces V_h satisfy the condition $\lim_{h\to 0} \text{dist}\,(u, V_h) \ = \ 0$ for all $u \in H^1(\Omega)$, where $\text{dist}\,(u, V_h) \ = \inf_{v_h \in V_h} \|u - v_h\|_1$ (see [5]). ∎

Let us now formulate the nonlinear algebraic system corresponding to (21). First we rewrite problem (21) with the notations (11) and (14):

$$\int_\Omega \Big(b(x, \nabla u_h) \ \nabla u_h \cdot \nabla v_h + \ r(x, u_h) u_h v_h \Big) \, dx +$$

$$+ \int_\Gamma z(x, u_h) u_h v_h \, d\sigma = \int_\Omega \hat{f} v_h \, dx \ + \int_\Gamma \hat{\gamma} v_h \, d\sigma \tag{22}$$

$\forall v_h \in V_h^0$. We set

$$u_h = \sum_{j=1}^{\bar{n}} c_j \phi_j, \tag{23}$$

and look for the coefficients $c_1, \ldots, c_{\bar{n}}$. For any $\bar{\mathbf{c}} = (c_1, ..., c_{\bar{n}}) \in \mathbf{R}^{\bar{n}}$, $i = 1, ..., n$ and $j = 1, ..., \bar{n}$, we set

$$b_{ij}(\bar{\mathbf{c}}) := \int_\Omega b(x, \sum_{k=1}^{\bar{n}} c_k \nabla\phi_k)\ \nabla\phi_j \cdot \nabla\phi_i \, dx,$$

$$r_{ij}(\bar{\mathbf{c}}) := \int_\Omega r(x, \sum_{k=1}^{\bar{n}} c_k \phi_k)\ \phi_j \phi_i \, dx,$$

$$z_{ij}(\bar{\mathbf{c}}) := \int_\Gamma z(x, \sum_{k=1}^{\bar{n}} c_k \phi_k)\ \phi_j \phi_i \, d\sigma, \qquad d_i(\bar{\mathbf{c}}) := \int_\Omega \hat{f} \phi_i \, dx + \int_\Gamma \hat{\gamma} \phi_i \, d\sigma\,,$$

$$a_{ij}(\bar{\mathbf{c}}) := b_{ij}(\bar{\mathbf{c}}) + r_{ij}(\bar{\mathbf{c}}) + z_{ij}(\bar{\mathbf{c}}). \tag{24}$$

Putting (23) and $v_h = \phi_i$ into (22), we obtain the $n \times \bar{n}$ system of algebraic equations

$$\sum_{j=1}^{\bar{n}} a_{ij}(\bar{\mathbf{c}})\, c_j = d_i, \quad i = 1, ..., n. \tag{25}$$

Using the notations

$$\begin{aligned} \mathbf{A}(\bar{\mathbf{c}}) :=& \{a_{ij}(\bar{\mathbf{c}})\},\ i, j = 1, ..., n, \\ \tilde{\mathbf{A}}(\bar{\mathbf{c}}) :=& \{a_{ij}(\mathbf{c})\},\ i = 1, ..., n;\ j = n+1, ..., \bar{n}, \\ \mathbf{d} :=& \{d_j\},\ \mathbf{c} := \{c_j\}, \quad j = 1, ..., n, \quad \text{and} \\ \tilde{\mathbf{c}} :=& \{c_j\}, \quad j = n+1, ..., \bar{n}, \end{aligned} \tag{26}$$

system (25) turns into

$$\mathbf{A}(\bar{\mathbf{c}})\mathbf{c} + \tilde{\mathbf{A}}(\bar{\mathbf{c}})\tilde{\mathbf{c}} = \mathbf{d}. \tag{27}$$

Defining further

$$\bar{\mathbf{A}}(\bar{\mathbf{c}}) := \begin{bmatrix} \mathbf{A}(\bar{\mathbf{c}}) & \tilde{\mathbf{A}}(\bar{\mathbf{c}}) \end{bmatrix}, \qquad \bar{\mathbf{c}} := \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{c}} \end{bmatrix}, \tag{28}$$

we rewrite (27) as follows

$$\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} = \mathbf{d}. \tag{29}$$

In order to obtain a system with a square matrix, we enlarge our system to an $\bar{n} \times \bar{n}$ one. Since $u_h = g_h$ on $\partial\Omega$, the coordinates c_i with $n+1 \le i \le \bar{n}$ satisfy automatically $c_i = g_i$, i.e.,

$$\tilde{\mathbf{c}} = \tilde{\mathbf{g}} := \{g_j\}, \quad j = n+1, ..., \bar{n},$$

hence we can replace (27) by the equivalent system

$$\begin{bmatrix} \mathbf{A}(\bar{\mathbf{c}}) & \tilde{\mathbf{A}}(\bar{\mathbf{c}}) \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \tilde{\mathbf{g}} \end{bmatrix}. \tag{30}$$

3 Maximum principle for the discretized problem

3.1 Background

First we recall a basic definition in the study of DMP (cf. [29, p. 23]):

Definition 2. A square $n \times n$ matrix $\mathbf{M} = (m_{ij})_{i,j=1}^{n}$ is called *irreducibly diagonally dominant* if it satisfies the following conditions:

(i) $\mathbf{M}$ is irreducible, i.e., for any $i \neq j$ there exists a sequence of nonzero entries $\{m_{i,i_1}, m_{i_1,i_2}, \ldots, m_{i_s,j}\}$ of M, where $i, i_1, i_2, \ldots, i_s, j$ are distinct indices,

(ii) $\mathbf{M}$ is diagonally dominant, i.e., $|m_{ii}| \geq \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} |m_{ij}|,\ i = 1, ..., n,$

(iii) for at least one index $i_0 \in \{1, ..., n\}$ the above inequality is strict, i.e.,

$$|m_{i_0,i_0}| > \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |m_{i_0,j}|.$$

Let us now consider a system of equations of order $(n + m) \times (n + m)$:

$$\bar{\mathbf{A}}\bar{\mathbf{c}} = \bar{\mathbf{b}},$$

where the matrix $\bar{\mathbf{A}}$ has the following structure:

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \tilde{\mathbf{A}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} . \tag{31}$$

Here $\mathbf{I}$ is the $m \times m$ identity matrix and $\mathbf{0}$ is the $m \times n$ zero matrix. Following [6], we introduce

Definition 3. An $(n + m) \times (n + m)$ matrix $\bar{\mathbf{A}}$ with the structure (31) is said to be of *generalized nonnegative type* if the following properties hold:

(i) $a_{ii} > 0, \quad i = 1, ..., n,$

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

(iii) $\sum\limits_{j=1}^{n+m} a_{ij} \geq 0, \quad i = 1, ..., n,$

(iv) There exists an index $i_0 \in \{1, \ldots, n\}$ for which $\sum\limits_{j=1}^{n} a_{i_0,j} > 0.$

(v) $\mathbf{A}$ is irreducible.

Remark 3. In the original definition in [6, p. 343], it is assumed instead of the above properties (iv)-(v) that the principal block $\mathbf{A}$ is irreducibly diagonally dominant. However, the latter follows directly from Definition 3 under the given sign conditions on a_{ij}.

We also note that a well-known theorem [29, p. 85] implies in this case that $\mathbf{A^{-1}} > 0$, i.e., the entries of the matrix $\mathbf{A^{-1}}$ are positive.

The known results on various discrete maximum principles (e.g., [6, 7, 14, 20]) are essentially based on the following theorem:

Theorem 5. *Let $\bar{\mathbf{A}}$ be a $(n+m)\times(n+m)$ matrix with the structure (31), and assume that $\bar{\mathbf{A}}$ is of generalized nonnegative type in the sense of Definition 3.*

If the vector $\bar{\mathbf{c}} = (c_1, ..., c_{n+m}) \in \mathbf{R}^{n+m}$ is such that $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \leq 0,\ i = 1, ..., n$, then

$$\max_{i=1,...,n+m} c_i \ \leq\ \max\{0, \max_{i=n+1,...,n+m} c_i\}. \tag{32}$$

If, in addition,

$$\sum_{j=1}^{n+m} a_{ij} = 0, \qquad i = 1, ..., n, \tag{33}$$

then

$$\max_{i=1,...,n+m} c_i \ =\ \max_{i=n+1,...,n+m} c_i. \tag{34}$$

Proof. As stated in Remark 3, $\mathbf{A}$ is irreducibly diagonally dominant. This, together with (i)-(iii), implies both statements (32) and (34), see [6, Th. 3] and [14, Th. 3], respectively. ∎

3.2 Algebraic conditions for the discrete maximum principle

The following theorem is the main result of the present paper since it will allow us to derive various forms of the discrete maximum principle. The sign condition (36) is similar to the one given in [7, 14].

Theorem 6. *Let Assumptions 2.1 hold and let*

$$f(x) - q(x, 0) \leq 0,\ x \in \Omega, \qquad \text{and} \qquad \gamma(x) - s(x, 0) \leq 0,\ x \in \Gamma. \tag{35}$$

Let us consider a family of simplicial triangulations $\mathcal{T}_h$ $(h > 0)$ satisfying the following property: for any $i = 1, ..., n,\ j = 1, ..., \bar{n}$ $(i \neq j)$

$$\nabla\phi_i \cdot \nabla\phi_j \leq -\frac{\sigma_0}{h^2} < 0 \tag{36}$$

on supp $\phi_i \cap$ supp ϕ_j with $\sigma_0 > 0$ independent of i, j and h.

(1) Let the triangulations $\mathcal{T}_h$ be regular, i.e., there exist constants $c_1, c_2 > 0$ such that for any $h > 0$ and any simplex $T \in \mathcal{T}_h$

$$c_1 h^d \leq \text{meas}(T) \leq c_2 h^d \tag{37}$$

(where meas *denotes d-dimensional measure). Then for sufficiently small h, the matrix* $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ *defined in (28) is of generalized nonnegative type in the sense of Definition 3.*

(2) More generally, for statement (1) to hold, it suffices to assume instead of (37) that the triangulations $\mathcal{T}_h$ *are only quasi-regular in the following sense: the left-hand side of (37) is replaced by*

$$c_1 h^\gamma \le \text{meas}(T) \tag{38}$$

with some $\gamma \ge d$ *satisfying*

$$2 \le \gamma < 3 \quad \text{if} \quad d = 2, \qquad 3 \le \gamma < \min\{\tfrac{12}{p_1 - 2},\ 5 - \tfrac{p_2}{2}\} \quad \text{if} \quad d = 3 \tag{39}$$

(or in general, $d \le \gamma < \min\{\frac{4d}{(p_1-2)(d-2)},\ 3 + \frac{(4-p_2)(d-2)}{2}\}$ *if* $d \ge 3$*) where* p_1 *and* p_2 *are defined in Assumptions 2.1, (A4).*

Proof. The coefficients of $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ satisfy

$$a_{ij}(\bar{\mathbf{c}}) = \int_\Omega \Big[b(x, \nabla u_h)\ \nabla\phi_i \cdot \nabla\phi_j\ + r(x, u_h)\ \phi_i\phi_j \Big]\, dx + \int_\Gamma z(x, u_h)\ \phi_i\phi_j\, d\sigma$$

($i = 1, ..., n,\ j = 1, ..., \bar{n}$). We now prove the properties (i)-(v) in the case (2); the conditions (38)-(39) are only used in part (ii).

(i) From our assumptions $b \ge \mu_0 > 0$, $r \ge 0$ and $z \ge 0$ we have

$$a_{ii}(\bar{\mathbf{c}}) \ge \mu_0 \int_\Omega |\nabla\phi_i|^2\, dx\ > 0.$$

(ii) Let $i = 1, ..., n,\ j = 1, ..., \bar{n}$ with $i \ne j$ and let Ω_{ij} denote the interior of *supp* $\phi_i \cap$ *supp* ϕ_j. If $\Omega_{ij} = \emptyset$ then

$$a_{ij}(\bar{\mathbf{c}}) = 0.$$

If $\Omega_{ij} \ne \emptyset$ then properties (12) and (36) and the fact $0 \le \phi_i \le 1, i = 1, ..., \bar{n}$, imply

$$a_{ij}(\bar{\mathbf{c}}) \le -\frac{\sigma_0}{h^2}\, \mu_0\ \text{meas}(\Omega_{ij}) + \int_{\Omega_{ij}} r(x, u_h)\, dx + \int_{\Gamma_{ij}} z(x, u_h)\, d\sigma, \tag{40}$$

using notation $\Gamma_{ij} = \Gamma \cap \overline{\Omega}_{ij}$. Here, from (11) and Assumption (A4),

$$\int_{\Omega_{ij}} r(x, u_h)\, dx = \int_{\Omega_{ij}} \frac{\partial q}{\partial \xi}(x, \theta u_h)\, dx \le \int_{\Omega_{ij}} \big(\alpha_1(x) + \beta|\theta u_h|^{p_1 - 2}\big)\, dx \le$$

$$\le \int_{\Omega_{ij}} \alpha_1(x)\, dx + \beta \int_{\Omega_{ij}} |u_h|^{p_1 - 2}\, dx$$

(where we had some $\theta = \theta(x) \in [0,1]$), and in just the same way we have

$$\int\limits_{\Gamma_{ij}} z(x,u_h)\,d\sigma \le \int\limits_{\Gamma_{ij}} \alpha_2(x)\,d\sigma + \beta \int\limits_{\Gamma_{ij}} |u_h|^{p_2-2}\,d\sigma.$$

Now we can estimate the integrals $\int\limits_{\Omega_{ij}} |u_h|^{p_1-2}\,dx$ and $\int\limits_{\Gamma_{ij}} |u_h|^{p_2-2}\,d\sigma$ as follows. We define $p^* := \frac{2d}{d-2}$ and $p^{**} := \frac{2(d-1)}{d-2}$ if $d \ge 3$, and $p^* := p^{**} := +\infty$ if $d = 2$. Then the Sobolev embedding estimates

$$\|v\|_{L^{p^*}(\Omega)} \le k_1 \|v\|_1, \quad \|v\|_{L^{p^{**}}(\Gamma)} \le k_2 \|v\|_1, \quad v \in H^1(\Omega), \tag{41}$$

hold with constants $k_1, k_2 > 0$, where $\|v\|_1 = \|v\|_{H^1(\Omega)}$ (see [1]). Assume for a while that $p_1, p_2 > 2$ and let us fix real numbers r and t satisfying

$$\frac{\gamma}{2} < r \le \frac{p^*}{p_1 - 2}\,, \qquad \frac{d-1}{d+1-\gamma} < t \le \frac{p^{**}}{p_2 - 2}\,. \tag{42}$$

Such numbers exist since for $d \ge 3$, by (39),

$$\gamma < \frac{2p^*}{p_1 - 2} \quad \text{and} \quad \gamma < 3 + \frac{(4-p_2)(d-2)}{2} = d + 1 + \frac{(2-p_2)(d-2)}{2} =$$

$$= d + 1 - \frac{(p_2 - 2)(d-1)}{p^{**}}\,.$$

Further, $\gamma \ge 2$ implies $r \ge 1$ and $t \ge 1$. If $\frac{1}{r} + \frac{1}{s} = \frac{1}{t} + \frac{1}{l} = 1$ then Hölder's inequality implies

$$\int\limits_{\Omega_{ij}} |u_h|^{p_1-2}\,dx \le \|1\|_{L^s(\Omega_{ij})} \left\| |u_h|^{p_1-2} \right\|_{L^r(\Omega_{ij})} = \operatorname{meas}(\Omega_{ij})^{1/s}\, \|u_h\|^{p_1-2}_{L^{(p_1-2)r}(\Omega_{ij})}\,. \tag{43}$$

Here $(p_1 - 2)r \le p^*$ and (41) imply

$$\|u_h\|^{p_1-2}_{L^{(p_1-2)r}(\Omega_{ij})} \le \|u_h\|^{p_1-2}_{L^{(p_1-2)r}(\Omega)} \le \text{const.} \cdot \|u_h\|^{p_1-2}_{L^{p^*}(\Omega)} \le \text{const.} \cdot \|u_h\|^{p_1-2}_1\,.$$

Owing to the basic FEM convergence result, we have $\|u_h\|_1 \to \|u^*\|_1$, where u^* is the exact weak solution of our problem. Hence if h is less than some fixed h_0 then (43) finally turns into

$$\int\limits_{\Omega_{ij}} |u_h|^{p_1-2}\,dx \le K_1\ \operatorname{meas}(\Omega_{ij})^{1/s} \tag{44}$$

with some constant $K_1 > 0$ independent of h. In just the same way we obtain

$$\int\limits_{\Gamma_{ij}} |u_h|^{p_2-2}\,dx \le K_2\ \operatorname{meas}(\Gamma_{ij})^{1/l}\,. \tag{45}$$

Finally, if p_1 or p_2 equals 2 then the corresponding equality (44) or (45) holds with $s = 1$ or $l = 1$, respectively.

The integrals of $\alpha_1(x)$ and $\alpha_2(x)$ can be estimated with Hölder's inequality similarly to (43) by letting $\frac{2}{d} + \frac{1}{s'} = \frac{1}{d-1} + \frac{1}{l'} = 1$:

$$\int\limits_{\Omega_{ij}} \alpha_1(x)\,dx \le K_3\ \mathrm{meas}(\Omega_{ij})^{1/s'}, \qquad \int\limits_{\Gamma_{ij}} \alpha_2(x)\,d\sigma \le K_4\ \mathrm{meas}(\Gamma_{ij})^{1/l'}$$

with $K_3 = \|\alpha_1\|_{L^{d/2}(\Omega)}$ and $K_4 = \|\alpha_2\|_{L^{d-1}(\Gamma)}$.

Substituting all the estimates in (40), we obtain

$$\begin{aligned} a_{ij}(\bar{\mathbf{c}}) \le\ & -\ \frac{\sigma_0\mu_0}{h^2}\ \mathrm{meas}(\Omega_{ij}) + \beta K_1\ \mathrm{meas}(\Omega_{ij})^{1/s} +\ K_3\ \mathrm{meas}(\Omega_{ij})^{1/s'} \\ & +\ \beta K_2\ \mathrm{meas}(\Gamma_{ij})^{1/l} + K_4\ \mathrm{meas}(\Gamma_{ij})^{1/l'}\,. \end{aligned} \tag{46}$$

We can write

$$a_{ij}(\bar{\mathbf{c}}) \le\ A_1^{ij}(h) + A_2^{ij}(h) + A_3^{ij}(h) + A_4^{ij}(h)$$

where, with suitable constants C_0, C_1, C_2, C_3, $C_4 > 0$ independent of h and i, j,

$$\begin{aligned} A_1^{ij}(h) &:= -\frac{C_0}{h^2}\ \mathrm{meas}(\Omega_{ij}) +\ C_1\ \mathrm{meas}(\Omega_{ij})^{1/s}, \\ A_2^{ij}(h) &:= -\frac{C_0}{h^2}\ \mathrm{meas}(\Omega_{ij}) +\ C_2\ \mathrm{meas}(\Gamma_{ij})^{1/l}, \\ A_3^{ij}(h) &:= -\frac{C_0}{h^2}\ \mathrm{meas}(\Omega_{ij}) +\ C_3\ \mathrm{meas}(\Omega_{ij})^{1/s'}, \\ A_4^{ij}(h) &:= -\frac{C_0}{h^2}\ \mathrm{meas}(\Omega_{ij}) +\ C_4\ \mathrm{meas}(\Gamma_{ij})^{1/l'}\,. \end{aligned}$$

We verify that for small enough h we have $\ A_k^{ij}(h) < 0$ $(k = 1, 2, 3, 4)$.

Using $\frac{1}{r} + \frac{1}{s} = 1$ and (38), we have

$$A_1^{ij}(h) = \mathrm{meas}(\Omega_{ij})^{1/s}\left(-\frac{C_0}{h^2}\ \mathrm{meas}(\Omega_{ij})^{1/r}\ +\ C_1\right) \le$$

$$\le \mathrm{meas}(\Omega_{ij})^{1/s}\left(-\,C_5\,h^{-2+(\gamma/r)}\ +\ C_1\right).$$

Since (42) implies $\frac{\gamma}{r} < 2$, the term in brackets tends to $-\infty$ as $h \to 0$ and hence $A_1^{ij}(h) < 0$ for small h.

Using (38) again and the fact that $\mathrm{meas}(\Gamma_{ij}) \le \mathrm{const} \cdot h^{d-1}$ (since h is the diameter of the simplices and Γ_{ij} lies on the $(d-1)$-dimensional surface), we have

$$A_2^{ij}(h) \le -\,C_6\,h^{\gamma-2} +\ C_7\,h^{\frac{d-1}{l}}.$$

Since (42) implies $1 - \frac{1}{l} = \frac{1}{t} < \frac{d+1-\gamma}{d-1} = 1 - \frac{\gamma-2}{d-1}$, we obtain $\frac{d-1}{l} > \gamma - 2$, i.e., the second term tends to 0 faster and hence $A_2^{ij}(h) < 0$ for small h.

The terms $A_3^{ij}(h)$ and $A_4^{ij}(h)$ can be handled similarly, since s' and l' satisfy the same estimates as s and l. Namely, we have $\frac{d}{2} = \frac{p^*}{p^*-2}$ and $d-1 = \frac{p^{**}}{p^{**}-2}$, hence by substituting $\frac{d}{2}$ and $d-1$ for r and t, respectively, we obtain that (42) holds in the special case $p_1 = p^*$ and $p_2 = p^{**}$. Owing to the condition $\frac{2}{d} + \frac{1}{s'} = \frac{1}{d-1} + \frac{1}{l'} = 1$, the numbers s' and l' play the same role as s and l and therefore the above estimates on $A_1^{ij}(h)$ and $A_2^{ij}(h)$ can be repeated for $A_3^{ij}(h)$ and $A_4^{ij}(h)$.

Altogether, we obtain that for small enough h, $A_k^{ij}(h) < 0$ $(k = 1, 2, 3, 4)$, that is, there exists $h_0 > 0$ such that

$$a_{ij}(\bar{\mathbf{c}}) < 0 \tag{47}$$

for all $h \le h_0$ and all i, j.

(iii) For any $i = 1, ..., n$,

$$\begin{aligned}\sum_{j=1}^{\bar{n}} a_{ij}(\bar{\mathbf{c}}) \;&= \int_\Omega \Big[b(x, \nabla u_h)\; \nabla\phi_i \cdot \nabla(\sum_{j=1}^{\bar{n}} \phi_j) + r(x, u_h)\; \phi_i(\sum_{j=1}^{\bar{n}} \phi_j) \Big]\, dx \\ &\qquad + \int_\Gamma z(x, u_h)\; \phi_i(\sum_{j=1}^{\bar{n}} \phi_j)\, d\sigma \\ &= \int_\Omega r(x, u_h)\, \phi_i\, dx \;+ \int_\Gamma z(x, u_h)\, \phi_i\, d\sigma \quad \ge 0,\end{aligned} \tag{48}$$

using the fact that $\sum_{j=1}^{\bar{n}} \phi_j \equiv 1$ and r, z, ϕ_i are nonnegative.

(iv) Assume for contradiction that $\sum_{j=1}^{n} a_{ij}(\bar{\mathbf{c}}) = 0$ for all $i = 1, ..., n$. This means that $\mathbf{A}(\bar{\mathbf{c}})$ carries the n-tuple of ones $\{1, ..., 1\}$ into the zero vector. This is impossible since $\mathbf{A}(\bar{\mathbf{c}})$ is symmetric and positive definite, and hence one-to-one.

(v) For any $i, j = 1, ..., n$ with $i \ne j$, let us pick a sequence of neighbouring vertices B_{i_k} $(k = 1, ..., s)$ in Ω that connect B_i with B_j (i.e., $i_0 = i$ and $i_s = j$). Here (47) shows that $a_{i_k, i_{k+1}}(\bar{\mathbf{c}}) < 0$, hence by Definition 2, $\mathbf{A}(\bar{\mathbf{c}})$ is irreducible. ∎

Theorem 6 enables us to derive the discrete the maximum principle for system (27):

Theorem 7. *Under the conditions of Theorem 6, we have*

$$\max_{\overline{\Omega}} u_h \le \max\{0, \max_{\partial\Omega} g_h\}. \tag{49}$$

In particular, if $g \ge 0$, then $\max_{\overline{\Omega}} u_h = \max_{\partial\Omega} g_h$, and if $g \le 0$, then we have the nonpositivity property $\max_{\overline{\Omega}} u_h \le 0$.

PROOF. Theorem 6 states that the condition of Theorem 5 is satisfied with $\bar{\mathbf{A}}(\mathbf{c})$ and $\bar{n}$ substituted for $\bar{\mathbf{A}}$ and $n+m$, respectively. Hence (32) yields

$$\max_{i=1,...,\bar{n}} c_i \;\le\; \max\{0, \max_{i=n+1,...,\bar{n}} c_i\}. \tag{50}$$

Since $c_i = g_i$ for all $i = n+1, ..., \bar{n}$, estimate (50) is equivalent to (49). ∎

The analogous minimum principle for system (27) can be verified in the same way.

Theorem 8. *Let the conditions of Theorem 6 hold, except for (35) which is now replaced by*

$$f(x) - q(x,0) \ge 0 \quad (x \in \Omega) \qquad \textit{and} \quad \gamma(x) - s(x,0) \ge 0 \quad (x \in \Gamma). \tag{51}$$

Then we have

$$\min_{\overline{\Omega}} u_h \ge \min\{0, \min_{\partial\Omega} g_h\}. \tag{52}$$

In particular, if $g \le 0$, then $\min\limits_{\overline{\Omega}} u_h = \min\limits_{\partial\Omega} g_h$, and if $g \ge 0$, then we have the nonnegativity property $\min\limits_{\overline{\Omega}} u_h \ge 0$.

Let us now consider the special case $q \equiv 0$ and $s \equiv 0$. Then the counterpart of Theorem 3 is valid, which we now formulate for both the maximum and minimum principles. Moreover, the strict negativity in (36) can be replaced by the weaker nonnegativity property, regularity conditions on the mesh like (38)-(39) are not required, and the result for a proper mesh holds for all parameters h instead of only sufficiently small h.

Theorem 9. *Let us consider the following special case of problem (1):*

$$\begin{cases} -\operatorname{div}\Big(b(x,\nabla u)\,\nabla u\Big) \;=\; f(x) & \text{in } \Omega\setminus\Gamma, \\ [\,u]_\Gamma \;=\; 0 & \text{on } \Gamma, \\ \big[\,b(x,\nabla u)\frac{\partial u}{\partial \nu}\big]_\Gamma \;=\; \gamma(x) & \text{on } \Gamma, \\ u \;=\; g(x) & \text{on } \partial\Omega, \end{cases} \tag{53}$$

Let (A1)–(A3) of Assumptions 2.1 hold and let the triangulation $\mathcal{T}_h$ satisfy the following property: for any $i = 1, ..., n,\ j = 1, ..., \bar{n}$ $(i \ne j)$

$$\nabla\phi_i \cdot \nabla\phi_j \le 0. \tag{54}$$

Then the following results hold:

(1) If $f \le 0$ and $\gamma \le 0$, then $\max\limits_{\overline{\Omega}} u_h = \max\limits_{\partial\Omega} g_h$.
(2) If $f \ge 0$ and $\gamma \ge 0$, then $\min\limits_{\overline{\Omega}} u_h = \min\limits_{\partial\Omega} g_h$.
(3) If $f = 0$ and $\gamma = 0$, then the ranges of u_h and g_h coincide, i.e., we have $[\min\limits_{\overline{\Omega}} u_h, \max\limits_{\overline{\Omega}} u_h] = [\min\limits_{\partial\Omega} g_h, \max\limits_{\partial\Omega} g_h]$ for the corresponding intervals.

PROOF. (1) The conditions of Theorem 5 follow similarly as in Theorem 6. The difference arises in proving property (ii), i.e., $a_{ij}(\mathbf{c}) \leq 0$, where only (54) is sufficient, since the assumptions $q \equiv 0$ and $s \equiv 0$ imply $r \equiv 0$ and $z \equiv 0$. In order to apply statement (34) of Theorem 5, it remains to verify that $\sum_{j=1}^{\bar{n}} a_{ij}(\mathbf{c}) = 0,\ i = 1, ..., n$. Since $r \equiv 0$ and $z \equiv 0$, the argument used in (48) yields that this holds indeed. Statement (2) follows from (1) by replacing u by $-u$, and (3) is a direct consequence of (1) and (2). ∎

Remark 4. Conditions (36) and (54) can be in fact relaxed such that $\nabla\phi_i \cdot \nabla\phi_j$ need not be negative resp. nonpositive on each element, see [14, Remark 6] for details.

3.3 Geometric conditions on the mesh

The conditions in the preceding subsection that guarantee the DMP have apparent geometric interpretations for our simplicial meshes. This relies on the fact that the values $\nabla\phi_i \cdot \nabla\phi_j$ are constant on each simplicial element, hence conditions (36) and (54) are not very difficult to check. Indeed, it is shown in [3] that

$$\nabla\phi_i \cdot \nabla\phi_j|_T = -\frac{\mathrm{meas}_{d-1}(S_i) \cdot \mathrm{meas}_{d-1}(S_j)}{d^2(\mathrm{meas}_d(T))^2} \cos(S_i, S_j) \quad \text{for } i \neq j, \tag{55}$$

where T is a d-dimensional simplex with vertices $P_1, \ldots, P_{d+1}$, S_i is the face of T opposite to P_i, and $\cos(S_i, S_j)$ is the cosine of the interior angle between faces S_i and S_j.

Thus, in order to satisfy condition (36) or (54), it is sufficient if the employed simplicial mesh is acute or nonobtuse, respectively (see [17, 18, 22], where also mesh refinement procedures preserving the above-mentioned geometrical properties are presented). We note that the conditions of acuteness or nonobtuseness are sufficient but not necessary: as referred to in Remark 4, the DMP may still hold if some obtuse interior angles occur in the simplices of the meshes. This is analogous to the case of linear problems [19, 28].

We note that the results can be easily extended to the case of meshes consisting of block elements, treated as in [15, Sect. 5.2].

3.4 Some applications to model problems

We quote two examples of problems where suitable discrete nonnegativity or nonpositivity properties are valid.

3.4.1 Semilinear equations: reaction-diffusion problems with localized autocatalytic chemical reactions

The problem

$$\begin{cases} -\Delta u = f(x) & \text{in } \Omega \setminus \Gamma, \\ [\,u]_\Gamma = 0 & \text{on } \Gamma, \\ \left[\,\frac{\partial u}{\partial \nu}\right]_\Gamma + s(x,u) = 0 & \text{on } \Gamma, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{56}$$

in a planar domain $\Omega \subset \mathbf{R}^2$ describes a chemical reaction-diffusion process where the reaction is localized at the curve Γ, further, the reaction is autocatalytic, i.e., the growth of the concentration $u \geq 0$ speeds up the rate of the reaction, that is $\frac{\partial s(x,u)}{\partial u} \geq 0$ (see, e.g., [12, 13]). The reaction function s grows at most polynomially in u, hence Assumptions 2.1 hold. The fact that there is no reaction without material is expressed by $s(x,0) = 0$, further, we may assume that the source term f is nonnegative. These conditions imply that the requirement $u \geq 0$ is satisfied, see subsection 2.3, moreover, the boundary conditions yield $\min_{\overline{\Omega}} u = 0$. As a special case of Theorem 8, we obtain the corresponding discrete minimum principle:

Corollary 1. *Let u_h be the FEM solution to problem (56) under a FEM discretization with the acuteness property (36). If h is sufficiently small then*

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

3.4.2 Linear equations

The following linear interface model arises in many applications such as biochemistry or multiphase flow, see, e.g., [26]:

$$\begin{cases} -\operatorname{div}\Big(k(x)\,\nabla u\Big) = f(x) & \text{in } \Omega \setminus \Gamma, \\ [\,u]_\Gamma = 0 & \text{on } \Gamma, \\ \left[\,k(x)\frac{\partial u}{\partial \nu}\right]_\Gamma = \gamma(x) & \text{on } \Gamma, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{57}$$

where the bounded measurable function k is discontinuous on Γ. In addition, it suffices to assume that k has a positive lower bound and $f \in L^2(\Omega)$, $\gamma \in L^2(\Gamma)$. Then, as a special case of Theorem 9, we obtain the corresponding discrete maximum and minimum principles:

Corollary 2. *Let u_h be the FEM solution to problem (57) under a FEM discretization with the nonobtuseness property (54).*

If $f \leq 0$ and $\gamma \leq 0$ then $\max_{\overline{\Omega}} u_h = 0$, and if $f \geq 0$ and $\gamma \geq 0$ then $\min_{\overline{\Omega}} u_h = 0$.

References

[1] Adams, R.A., *Sobolev spaces*, Academic Press, New York-London, 1975.

[2] Bertolazzi, E., Manzini, G., A second-order maximum principle preserving finite volume method for steady convection-diffusion problems, *SIAM J. Numer. Anal.* 43 (2005), no. 5, 2172–2199 (electronic).

[3] Brandts, J., Korotov, S., Křížek, M., Dissection of the path-simplex in R^n into n path-subsimplices, Research Report A496, Helsinki University of Technology, 2006; to appear in *Linear Algebra Appl.*

[4] Christie, I., Hall, C., The maximum principle for bilinear elements, *Internat. J. Numer. Methods Engrg.* 20 (1984), 549–553.

[5] Ciarlet, P. G., *The finite element method for elliptic problems,* North-Holland, Amsterdam, 1978

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

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

[8] Faragó, I., Karátson, J., *Numerical solution of nonlinear elliptic problems via preconditioning operators. Theory and applications.* Advances in Computation, Volume 11, NOVA Science Publishers, New York, 2002.

[9] Fleishman B. A., Mahar T. J., A minimum principle for superharmonic functions subject to interface conditions, *J. Math. Anal. Appl.*, 80 (1981), 46-56.

[10] Gilbarg, D., Trudinger, N. S., *Elliptic partial differential equations of second order* (2nd edition), Grundlehren der Mathematischen Wissenschaften 224, Springer, 1983.

[11] Ishihara, K., Strong and weak discrete maximum principles for matrices associated with elliptic problems, *Linear Algebra Appl.* 88/89 (1987), 431–448.

[12] Kandilarov, J. D., A monotone iterative method for numerical solution of diffusion equations with nonlinear localized chemical reactions, to appear in *Lecture Notes in Comput. Sci.*

[13] Kandilarov, J. D., Vulkov, L. G.., Analysis of immersed interface difference schemes for reaction-diffusion problems with singular own sources, *Comput. Methods Appl. Math.* 3 (2003), no. 2, 253–273 (electronic).

[14] Karátson, J., Korotov, S., Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, *Numer. Math.* 99 (2005), 669–698.

[15] Karátson J., Korotov, S., Křížek, M., On discrete maximum principles for nonlinear elliptic problems, Research Report A504, Helsinki University of Technology, 2006, to appear in *Math. Comput. Simul.*

[16] Korotov, S., Křížek, M., Finite element analysis of variational crimes for a quasilinear elliptic problem in 3D, *Numer. Math.* 84 (2000), 549–576.

[17] Korotov, S., Křížek, M., Acute type refinements of tetrahedral partitions of polyhedral domains, *SIAM J. Numer. Anal.* 39 (2001), 724–733.

[18] Korotov, S., Křížek, M., Global and local refinement techniques yielding nonobtuse tetrahedral partitions, *Comput. Math. Appl.* 50 (2005), no. 7, 1105–1113.

[19] Korotov, S., Křížek, M., Neittaanmäki, P., Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle, *Math. Comp.* 70 (2001), 107–119.

[20] Křížek, M., Lin Qun, On diagonal dominance of stiffness matrices in 3D, *East-West J. Numer. Math.* 3 (1995), 59–69.

[21] Křížek, M., Neittaanmäki, P., *Mathematical and numerical modelling in electrical engineering: theory and applications*, Kluwer Academic Publishers, 1996.

[22] Křížek, M., Šolc, J., Acute versus nonobtuse tetrahedralizations, in: *Conjugate gradient algorithms and finite element methods*, 161–170, Sci. Comput., Springer, Berlin, 2004.

[23] Ladyzhenskaya, O. A., Ural'tseva, N. N., *Linear and quasilinear elliptic equations*, Leon Ehrenpreis Academic Press, New York-London, 1968.

[24] LeVeque, R. J., Li, Zh., The immersed interface method for elliptic equations with discontinuous coefficients and singular sources, *SIAM J. Numer. Anal.* 31 (1994), no. 4, 1019–1044.

[25] Li, Zh., A fast iterative algorithm for elliptic interface problems, *SIAM J. Numer. Anal.* 35 (1998), no. 1, 230–254.

[26] Li, Zh., Ito, K., Maximum principle preserving schemes for interface problems with discontinuous coefficients, *SIAM J. Sci. Comput.* 23 (2001), no. 1, 339–361 (electronic).

[27] Protter, M. H., Weinberger, H. F., *Maximum principles in differential equations*, Springer-Verlag, New York, 1984.

[28] Ruas Santos, V., On the strong maximum principle for some piecewise linear finite element approximate problems of non-positive type, *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* 29 (1982), 473–491.

[29] Varga, R., *Matrix iterative analysis*, Prentice Hall, New Jersey, 1962.

(continued from the back cover)

A506 Sergey Korotov
Error control in terms of linear functionals based on gradient averaging techniques
July 2006

A505 Jan Brandts , Sergey Korotov , Michal Krizek
On the equivalence of regularity criteria for triangular and tetrahedral finite element partitions
July 2006

A504 Janos Karatson , Sergey Korotov , Michal Krizek
On discrete maximum principles for nonlinear elliptic problems
July 2006

A503 Jan Brandts , Sergey Korotov , Michal Krizek , Jakub Solc
On acute and nonobtuse simplicial partitions
July 2006

A502 Vladimir M. Miklyukov , Antti Rasila , Matti Vuorinen
Three sphres theorem for p-harmonic functions
June 2006

A501 Marina Sirviö
On an inverse subordinator storage
June 2006

A500 Outi Elina Maasalo , Anna Zatorska-Goldstein
Stability of quasiminimizers of the p–Dirichlet integral with varying p on metric spaces
April 2006

A499 Mikko Parviainen
Global higher integrability for parabolic quasiminimizers in nonsmooth domains
April 2005

A498 Marcus Ruter , Sergey Korotov , Christian Steenbock
Goal-oriented Error Estimates based on Different FE-Spaces for the Primal and the Dual Problem with Applications to Fracture Mechanics
March 2006

HELSINKI UNIVERSITY OF TECHNOLOGY INSTITUTE OF MATHEMATICS
RESEARCH REPORTS

The list of reports is continued inside. Electronical versions of the reports are available at *http://www.math.hut.fi/reports/* .

A513 Wolfgang Desch , Stig-Olof Londen
On a Stochastic Parabolic Integral Equation
October 2006

A512 Joachim Schöberl , Rolf Stenberg
Multigrid methods for a stabilized Reissner-Mindlin plate formulation
October 2006

A509 Jukka Tuomela , Teijo Arponen , Villesamuli Normi
On the simulation of multibody systems with holonomic constraints
September 2006

A508 Teijo Arponen , Samuli Piipponen , Jukka Tuomela
Analysing singularities of a benchmark problem
September 2006

A507 Pekka Alestalo , Dmitry A. Trotsenko
Bilipschitz extendability in the plane
August 2006

ISBN-13 978-951-22-8445-0
ISBN-10 951-22-8445-6