Some discrete maximum principles arising for nonlinear elliptic finite element problems

J. Karátson[1,*], S. Korotov[2,†]

June 29, 2015

Abstract: The discrete maximum principle (DMP) is an important measure of the qualitative reliability of the applied numerical scheme for elliptic problems. This paper starts with formulating simple sufficient conditions for the matrix case and for nonlinear forms in Banach spaces. Then a DMP is derived for finite element solutions for certain nonlinear partial differential equations: we address nonlinear elliptic problems with mixed boundary conditions and interface conditions, allowing possibly degenerate nonlinearities and thus extending our previous results.

1 Introduction

The maximum principle forms an important qualitative property of second order elliptic equations [22], therefore its discrete analogues, the so-called discrete maximum principles (DMPs) have drawn much attention. The DMP is in fact an important measure of the qualitative reliability of the numerical scheme, otherwise one could get unphysical numerical solutions like negative concentrations etc. Typical maximum principles arise either in the form

$$\max_{\overline{\Omega}} u = \max_{\partial\Omega} u \tag{1}$$

(i.e. the solution u attains its maximum on the boundary), which occurs for proper elliptic operators with only principal part, or in the form

$$\max_{\overline{\Omega}} u \le \max\{0, \max_{\partial\Omega} u\} \tag{2}$$

(i.e. the solution u can attain a nonnegative maximum only on the boundary), which occurs for proper elliptic operators including lower order terms as well. We are interested for DMPs in the context of the finite element method (FEM), in which case a DMP reproduces one of the above relations for the FEM solution u_h instead of u.

Various DMPs, including geometric conditions on the computational meshes for FEM solutions, have been given e.g. in [3, 6, 7, 9, 24, 27, 28]. The authors' previous work,

*Department of Applied Analysis & MTA-ELTE NumNet Research Group, ELTE University; Department of Analysis, Technical University; Budapest, Hungary; karatson@cs.elte.hu.

†Bergen University College, Bergen, Norway & IKERBASQUE, Basque Foundation for Science, Bilbao, Spain; sergey.korotov@hib.no .

e.g. [15, 16, 17, 18, 19], involves various types of linear and nonlinear equations and systems, also including the analogue of (1)-(2) for mixed boundary conditions such that only the Dirichlet boundary needs to be considered. Typical geometric conditions are nonobtuseness or acuteness in the case of simplicial meshes.

In this paper we first discuss the algebraic background, i.e. matrix maximum principles (MMPs). Since the early works [6, 7] such results usually involve some irreducibility condition on the matrix, which is a delicate issue and cannot be considered as granted or easily provable in FEM context [13]. Therefore, we formulate a simple sufficient condition for some MMP to hold, such that irreducibility is avoided. Then we derive a related result for nonlinear forms in Banach space. Finally, we apply the results to FEM problems for certain nonlinear partial differential equations (PDEs). Using the mentioned MMP, it becomes more straightforward to verify the conditions for the discretized PDE, allowing possibly degenerate nonlinearities. We address nonlinear elliptic problems with mixed boundary conditions and interface conditions, such that we do not assume lower and upper boundedness of the diffusion coefficients, thus extending our mentioned previous results. In particular, we can allow degeneracy in the equation, i.e. loss of ellipticity due to possible vanishing of the diffusion coefficient in some parts of the domain. This is illustrated by various examples at the end.

2 Matrix maximum principles

2.1 Classical results

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

$$\bar{\mathbf{A}}\bar{\mathbf{c}} = \bar{\mathbf{b}}, \tag{3}$$

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

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \tilde{\mathbf{A}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} . \tag{4}$$

In the above, $\mathbf{I}$ is the $m\times m$ identity matrix, $\mathbf{0}$ is the $m\times n$ zero matrix. In FEM problems such a partitioning arises corresponding to interior and boundary points.

We first recall some classical definitions and results, see, e.g., [6, 25]. We follow the terminology of [10]. Throughout, inequalities for matrices or vectors are understood elementwise, and the symbols $\mathbf{e}$, $\tilde{\mathbf{e}}$ and $\bar{\mathbf{e}}$ denote the vectors of all ones of length n, m or $n+m$, respectively.

Definition 1 The matrix $\bar{\mathbf{A}}$ in (4) satisfies

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

$$\max_{i=1,...,n+m} c_i \ \le\ \max\{0, \max_{i=n+1,...,n+m} c_i\}; \tag{5}$$

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

$$\max_{i=1,...,n+m} c_i \ =\ \max_{i=n+1,...,n+m} c_i. \tag{6}$$

(DMPs without the term 'weak' also assert that only constant vectors may attain a maximum for 'interior' indices, but we do not address this property here.)

Theorem 1 *[7, 10]. The matrix* $\bar{\mathbf{A}}$ *possesses*

(a) the DwMP if and only if the following three conditions hold:

(i) $\mathbf{A}^{-1} \geq 0$, *(ii)* $-\mathbf{A}^{-1}\tilde{\mathbf{A}} \geq 0$, *(iii)* $-\mathbf{A}^{-1}\tilde{\mathbf{A}}\tilde{\mathbf{e}} \leq \mathbf{e}$;

(b) the DWMP if and only if the following three conditions hold:

(i) $\mathbf{A}^{-1} \geq 0$, *(ii)* $-\mathbf{A}^{-1}\tilde{\mathbf{A}} \geq 0$, *(iii)* $-\mathbf{A}^{-1}\tilde{\mathbf{A}}\tilde{\mathbf{e}} = \mathbf{e}$.

Here the expression $\mathbf{A}^{-1} \geq 0$ means that $\mathbf{A}^{-1}$ exists (i.e. $\mathbf{A}$ is nonsingular) and $\mathbf{A}^{-1}$ has nonnegative entries. Such matrices are called monotone [25]. In view of the sign conditions, and since the upper block row of the matrix (4) satisfies $[\mathbf{A}\ \ \tilde{\mathbf{A}}]\,\bar{\mathbf{e}} = \mathbf{A}\mathbf{e} + \tilde{\mathbf{A}}\tilde{\mathbf{e}}$, we obtain

Corollary 1 *Sufficient conditions for the matrix* $\bar{\mathbf{A}}$ *to possess*

(a) the DwMP, are the following:

(i) $\mathbf{A}^{-1} \geq 0$, *(ii)* $\tilde{\mathbf{A}} \leq 0$, *(iii)* $[\mathbf{A}\ \ \tilde{\mathbf{A}}]\,\bar{\mathbf{e}} \geq \mathbf{0}$;

(b) the DWMP, are the following:

(i) $\mathbf{A}^{-1} \geq 0$, *(ii)* $\tilde{\mathbf{A}} \leq 0$, *(iii)* $[\mathbf{A}\ \ \tilde{\mathbf{A}}]\,\bar{\mathbf{e}} = \mathbf{0}$;

The hardest part is usually to ensure (i), i.e. the "monotonicity property" $\mathbf{A}^{-1} \geq 0$. Often this property is connected with irreducibility [25], then one even obtains a stronger result:

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

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

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

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

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

Definition 3 We say that $\operatorname{offdiag}(A) \leq 0$ if $a_{ij} \leq 0$ for all $i \neq j$.

Theorem 2 *[25] If a square* $n \times n$ *matrix* $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ *is irreducibly diagonally dominant and, further,* $\operatorname{offdiag}(A) \leq 0$ *and* $a_{ii} > 0$ *for all* $i = 1, ..., n$, *then* $\mathbf{A^{-1}} > 0$.

2.2 Some simple conditions

In practice it is often not straightforward to verify the irreducibility of the stiffness matrix, in particular, if one considers a linearized form of a nonlinear problem. A weakened form of irreducibility has been proposed and applied in [17]. However, it is most useful to have conditions that avoid irreducibility at all. Such conditions exist in the case of positive definiteness. The first classical result involves also symmetry, and it has been shown e.g. in [1, 13] that it helps to check easier the validity of the DMP for various linear FEM matrices.

Definition 4 The matrix $\mathbf{A}$ is called a *Stieltjes* matrix if $\mathrm{offdiag}\,(A) \leq 0$ and $\mathbf{A}$ is symmetric and positive definite.

Theorem 3 *[12, 25] If a square $n \times n$ matrix $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ is a Stieltjes matrix, then $\mathbf{A^{-1}} \geq 0$.*

On the other hand, one may need not assume symmetry:

Theorem 4 *[26] If a square $n \times n$ matrix $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ satisfies $\mathrm{offdiag}\,(A) \leq 0$ and $\mathbf{A}$ is positive definite, then $\mathbf{A^{-1}} \geq 0$.*

As a ready consequence, we obtain

Theorem 5 *Let the matrix $\bar{\mathbf{A}}$ in (4) satisfy the following conditions, where a_{ij} denote the entries of $\bar{\mathbf{A}}$:*

(i) $a_{ij} \leq 0 \qquad (\forall i = 1, \ldots, n,\ j = 1, \ldots, n+m;\quad i \neq j)$,

(ii) $\sum\limits_{j=1}^{n+m} a_{ij} \geq 0 \qquad (\forall i = 1, \ldots, n)$,

(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.

Proof. Condition (i) for indices $j \leq n$ means that $\mathrm{offdiag}\,(A) \leq 0$, hence by condition (iii) and Theorem 4 we have $\mathbf{A^{-1}} \geq 0$. Condition (i) for indices $j > n$ means that $\tilde{\mathbf{A}} \leq 0$, hence conditions (i)-(ii) of Corollary 1 are satisfied. Finally, our condition (ii) and its variant with equality coincide with the two versions of condition (iii) of Corollary 1, hence the latter implies the DwMP resp. DWMP. ∎

3 A discrete maximum principle under nonlinear forms

Certain nonlinear problems, such as typically the weak formulations of some nonlinear elliptic PDEs, can be described in the following framework. Let X be a real Banach space and $X_0 \subset X$ be a given subspace. Let $a : X \times X \times X \to \mathbf{R}$ be a mapping such that for all fixed $u \in X$, the mapping $v, z \mapsto a(u; v, z)$ is bilinear. For a given bounded linear

functional $\ell : X \to \mathbf{R}$ and element $\tilde{g} \in X$, we consider the following problem: find $u \in X$ such that

$$\begin{cases} a(u;\ u, v) = \ell v & (\forall v \in X_0) \\ u - \tilde{g} \in X_0. \end{cases} \tag{7}$$

A usual property in this setting is positive definiteness (i.e. 'ellipticity') with respect to X_0 in the last two variables, which we only assume here on the solution $u \in X$ of (7):

$$\text{for all } v \in X_0,\ v \neq 0, \text{ we have } \ a(u;\ v, v) > 0. \tag{8}$$

Existence and uniqueness of the solution is normally ensured by suitably strengthening the above ellipticity property with proper monotonicity and continuity conditions, see, e.g., [11, 30]. We do not detail these here, since we are only interested in qualitative properties of such problems.

Let $V_h \subset X$ be a given finite dimensional subspace and

$$\phi_1, ..., \phi_{\bar{n}}$$

a basis in V_h, such that for some given index $1 < n < \bar{n}$ we have $\phi_1, ..., \phi_n \in X_0$ and $\phi_{n+1}, ..., \phi_{\bar{n}} \notin X_0$. Let

$$V_h^0 := \text{span}\{\phi_1, ..., \phi_n\} \ \subset X_0.$$

Further, let

$$g_h = \sum_{j=n+1}^{\bar{n}} g_j \phi_j \ \in V_h \tag{9}$$

(with $g_j \in \mathbf{R}$) be a suitable approximation of the component of $\tilde{g}$ in $X \setminus X_0$.

Then the Galerkin solution of (7) is defined as an element $u_h \in V_h$ such that

$$\begin{cases} a(u_h;\ u_h, v_h) = \ell v_h & (\forall v_h \in V_h^0) \\ u_h - g_h \in V_h^0. \end{cases} \tag{10}$$

Our first assumption is the discrete version of the positive definiteness (8) on the solution $u_h \in V_h$ of (10):

$$(\mathbf{A1}) \qquad \text{for all } v_h \in V_h^0,\ v_h \neq 0, \text{ we have } \ a(u_h;\ v_h, v_h) > 0. \tag{11}$$

We set

$$u_h = \sum_{j=1}^{\bar{n}} c_j \phi_j, \tag{12}$$

and look for the unknown coefficient vector $\bar{\mathbf{c}} = (c_1, \ldots, c_{\bar{n}})^T$ that represents u_h.

Now we can formulate a discrete maximum principle for the coefficient vector $\bar{\mathbf{c}}$, relating the coordinates $c_1, \ldots, c_n$ and $c_{n+1}, \ldots, c_{\bar{n}}$ in the vein of Section 2.

Theorem 6 *Let the form 'a' satisfy assumption (A1), let $u_h \in V_h$ be the solution of (10), and let the basis functions fulfil the following conditions:*

(A2) $\quad a(u_h;\ \phi_j, \phi_i) \le 0 \qquad (\forall i = 1, \ldots, n,\ j = 1, \ldots, \bar{n};\quad i \ne j);$

(A3) $\quad a(u_h;\ \sum_{j=1}^{\bar{n}} \phi_j,\ \phi_i) \ge 0 \qquad (\forall i = 1, \ldots, n);$

(A4) $\quad \ell\phi_i \le 0 \qquad (\forall i = 1, \ldots, n).$

Then the coefficient vector $\bar{\mathbf{c}} = (c_1, \ldots, c_{\bar{n}})^T$ *given in (12) satisfies*

$$\max_{i=1,\ldots,\bar{n}} c_i \ \le\ \max\{0, \max_{i=n+1,\ldots,\bar{n}} c_i\}; \tag{13}$$

If assumption (A3) is replaced by equality:

(A3)' $\quad a(u_h;\ \sum_{j=1}^{\bar{n}} \phi_j,\ \phi_i) = 0 \qquad (\forall i = 1, \ldots, n),$

then the coefficient vector of (12) satisfies

$$\max_{i=1,\ldots,\bar{n}} c_i \ =\ \max_{i=n+1,\ldots,\bar{n}} c_i. \tag{14}$$

Proof. Let us decompose $u_h - g_h$ as

$$u_h - g_h = \sum_{j=1}^{n} c_j\phi_j + \sum_{j=n+1}^{\bar{n}} (c_j - g_j)\,\phi_j,$$

i.e. into components spanned by basis functions in V_h^0 and in $V_h \setminus V_h^0$, respectively. Since, by (10), the second sum must vanish, therefore

$$c_j = g_j \qquad (\forall j = n+1, \ldots, \bar{n})$$

and thus we only look for the coefficients of

$$u_h - g_h = \sum_{j=1}^{n} c_j\phi_j \in V_h^0.$$

Here problem (10) is equivalent to demanding only $v_h = \phi_i$ $(i = 1, ..., n)$:

$$a(u_h;\ u_h, \phi_i) = \ell\phi_i \qquad (\forall i = 1, ..., n). \tag{15}$$

Let

$$a_{ij}(\bar{\mathbf{c}}) := a(u_h;\ \phi_j, \phi_i) \qquad (i = 1, \ldots, n,\ j = 1, \ldots, \bar{n}).$$

Then (15) is equivalent to

$$\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} = \bar{\mathbf{b}}, \tag{16}$$

where the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ has the structure as in (4):

$$\bar{\mathbf{A}}(\bar{\mathbf{c}}) = \begin{bmatrix} \mathbf{A}(\bar{\mathbf{c}}) & \tilde{\mathbf{A}}(\bar{\mathbf{c}}) \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{17}$$

such that now

$$\bar{n} = n + m.$$

Clearly, assumptions (A2)-(A3) mean that the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ satisfies conditions (i)-(ii) of Theorem 5, and (A3)' leads to equality in (ii). Further, by assumption (A1), for any vector $\mathbf{d} \in \mathbf{R}^n \neq 0$ and corresponding element $v_h = \sum_{j=1}^n d_j \phi_j \neq 0$, we have

$$0 < a(u_h;\ v_h, v_h) = \sum_{i,j=1}^{n} a(u_h;\ \phi_j, \phi_i) d_j d_i = \mathbf{A}(\bar{\mathbf{c}})\mathbf{d} \cdot \mathbf{d},$$

i.e. $\mathbf{A}(\bar{\mathbf{c}})$ is positive definite. Hence altogether Theorem 5 yields that $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ possesses the DwMP resp. DWMP. Finally, assumption (A4) implies that $b_i := \ell\phi_i \leq 0$ $(i = 1, ..., n)$, hence by Definition 1, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ provides the desired results for $\bar{\mathbf{c}}$. ∎

4 Discrete maximum principles for some nonlinear elliptic problems

We consider a nonlinear boundary value problem of the following type, involving possibly both (mixed) boundary and interface conditions:

$$\begin{cases} -\operatorname{div}\big(b(x,u,\nabla u)\,\nabla u\big) = f(x) & \text{in } \Omega, \\ u = g(x) & \text{on } \Gamma_D, \\ b(x,u,\nabla u)\frac{\partial u}{\partial \nu} = \gamma_N(x) & \text{on } \Gamma_N, \\ \Big[b(x,u,\nabla u)\frac{\partial u}{\partial \nu}\Big]_{\Gamma_{int}} = \gamma_{int}(x) \quad \text{and} \quad [\,u\,]_{\Gamma_{int}} = 0 & \text{on } \Gamma_{int}, \end{cases} \tag{18}$$

where Ω is a bounded polyhedral domain in $\mathbf{R}^d$. The explanation of the other notations and the assumed conditions are as follows:

Assumptions 4.1:

(i) The domain Ω has a 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, Γ_D has positive measure relative to $\partial\Omega$.

(ii) The interface Γ_{int} is a piecewise smooth Lipschitz curve (when $d = 2$) or hypersurface (when $d \geq 3$) lying in Ω, further, $[\,u\,]_{\Gamma_{int}}$ and $\Big[b(x,u,\nabla u)\frac{\partial u}{\partial \nu}\Big]_{\Gamma_{int}}$ denote the jump (i.e., the difference of the limits from the two sides of the interface Γ_{int}) of u and $b(x,u,\nabla u)\frac{\partial u}{\partial \nu}$, respectively.

(iii) The function $b : \Omega \times \mathbf{R} \times \mathbf{R}^d \to \mathbf{R}$ is continuous and satisfies

$$b(x,\xi,\eta) > 0 \qquad (\forall x \in \Omega,\ \xi \in \mathbf{R}, \eta \in \mathbf{R}^d;\ \xi \neq 0, \eta \neq 0). \tag{19}$$

(Then, due to continuity, $b \geq 0$ everywhere on its domain.)

(iv) We have $f \in L^2(\Omega)$, $\gamma_N \in L^2(\Gamma_N)$ and $\gamma_{int} \in L^2(\Gamma_{int})$. Further, $g = g^*{}_{|\Gamma_D}$ for some $g^* \in H$, where the space H is introduced below.

The weak formulation is done in a proper Sobolev space H. This is usually $H^1(\Omega)$ or $W^{1,p}(\Omega)$, as required by the actual growth of the nonlinearity, see Remark 1 below. We also involve its subspace corresponding to the homogeneous Dirichlet boundary condition:

$$H_D(\Omega) := \{u \in H : \ u = 0 \text{ on } \Gamma_D\},$$

where the boundary condition is in the sense of traces. Further, we let $\Gamma := \Gamma_N \cup \Gamma_{int}$ and define the function $\gamma : \Gamma \to \mathbf{R}$ as equal to γ_N and γ_{int} on Γ_N and Γ_{int}, respectively. Then a weak solution of problem (18) is a function $u \in H$ satisfying

$$\int_\Omega b(x, u, \nabla u)\ \nabla u \cdot \nabla v = \int_\Omega f v\, dx + \int_\Gamma \gamma v\, d\sigma \qquad \forall v \in H_D(\Omega) \tag{20}$$

$$\text{and} \quad u = g \quad \text{on } \Gamma_D \text{ in the sense of traces,} \tag{21}$$

where

$$\int_\Gamma \gamma v\, d\sigma = \int_{\Gamma_N} \gamma_N v\, d\sigma + \int_{\Gamma_{int}} \gamma_{int} v\, d\sigma.$$

(For the treatment of the interface condition in the weak form above, we refer to [16]. In particular, a classical solution of problem (18) is also a weak solution, and the converse holds for sufficiently regular weak solutions.)

Remark 1 To ensure existence and uniqueness of the weak solution, one has again to strengthen the simple positivity property (19) with proper monotonicity and continuity conditions, see, e.g., [11, 23], which usually include certain growth conditions on $b(x, \eta)$ proportional to some power of $|\eta|$. As in Section 3, we do not detail these here, since we are only interested on the qualitative properties of such problems, which will simply follow using (19); thus we always simply assume the existence of the solution.

For the numerical solution of our problem, we define the finite element discretization using simplicial elements and continuous piecewise linear basis functions. The symbol $\mathcal{T}_h$ stands for a conforming partition of $\overline{\Omega}$ into triangles (when $d = 2$) or simplices (when $d \geq 3$), whose vertices are $B_1, ..., B_{\bar{n}}$. Here $\mathcal{T}_h$ need not necessarily conform the interface Γ_{int}, but we may allow this additional property to simplify numerical integration on Γ_{int}. We 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:

$$V_h = \text{span}\{\phi_1, ..., \phi_{\bar{n}}\} \ \subset H.$$

Now, let $n < \bar{n}$ be such that

$$B_1, ..., B_n \tag{22}$$

are the nodal points that lie in Ω or on Γ_N, and let

$$B_{n+1}, ..., B_{\bar{n}} \tag{23}$$

be the nodal points that lie on $\bar{\Gamma}_D$. Then the basis functions $\phi_1, ..., \phi_n$ satisfy the homogeneous Dirichlet boundary condition on Γ_D, i.e., $\phi_i \in H_D(\Omega)$. We define

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

Further, let

$$g_h = \sum_{j=n+1}^{\bar{n}} g_j \phi_j \ \in V_h \tag{24}$$

(with $g_j \in \mathbf{R}$) be the piecewise linear approximation of the function g on Γ_D (and on the neighbouring elements).

To find the FEM solution, we solve the counterpart of (20)–(21) in V_h: find $u_h \in V_h$ such that

$$\int_\Omega b(x, u_h, \nabla u_h)\ \nabla u_h \cdot \nabla v_h \, dx \ = \int_\Omega f v_h \, dx \ + \int_\Gamma \gamma v_h \, d\sigma \qquad (\forall v_h \in V_h^0), \tag{25}$$

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

We set

$$u_h = \sum_{j=1}^{\bar{n}} c_j \phi_j, \tag{26}$$

and look for the coefficients $c_1, \ldots, c_{\bar{n}}$.

Now we can formulate a discrete maximum principle for the FEM solution u_h in analogy with (1).

Theorem 7 *Let Assumptions 4.1 hold, and let the FEM problem (25) have a solution $u_h \in V_h$. Let us consider a simplicial triangulation which is nonobtuse, i.e.*

$$\nabla \phi_i \cdot \nabla \phi_j \le 0 \qquad (\forall i = 1, ..., n, \ j = 1, ..., \bar{n}; \quad i \ne j). \tag{27}$$

If

$$f < 0 \quad \textit{a.e. on } \Omega, \qquad \gamma_N \le 0 \quad \textit{a.e. on } \Gamma_N, \quad \gamma_{int} \le 0 \quad \textit{a.e. on } \Gamma_{int},$$

then u_h satisfies a 'discrete strict weak maximum principle' related to the mixed boundary condition, i.e.

$$\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} g_h. \tag{28}$$

Proof. Problem (25) can be written in the form (10), where

$$a(u_h;\ w_h, v_h) := \int_\Omega b(x, u_h, \nabla u_h)\ \nabla w_h \cdot \nabla v_h \, dx, \qquad \ell v_h := \int_\Omega f v_h \, dx \ + \int_\Gamma \gamma v_h \, d\sigma$$

($w_h, v_h \in V_h^0$). First we verify that this form and the basis functions satisfy assumptions (A1), (A2), (A3)' and (A4) of Theorem 6.

(A1) We have obviously

$$a(u_h;\ v_h, v_h) = \int_\Omega b(x, u_h, \nabla u_h)\ |\nabla v_h|^2 \, dx \ge 0$$

for all $u_h \in V_h$, $v_h \in V_h^0$. It remains to prove that

$$\text{if} \quad a(u_h;\ v_h, v_h) = 0 \qquad \text{then} \quad v_h \equiv 0. \tag{29}$$

Assume for contradiction that there exists $v_h \in V_h^0$ such that $a(u_h;\ v_h, v_h) = 0$ but

$$v_h \not\equiv 0.$$

Letting

$$\Omega_0 := \{x \in \Omega :\ \nabla v_h(x) = 0\}, \qquad \Omega_1 := int(\Omega \setminus \Omega_0),$$

we obtain that

$$\Omega_0 \neq \Omega,$$

since otherwise v_h would be a constant that attains zero values on Γ_D due to $v_h \in V_h^0$, i.e. we would have $v_h \equiv 0$. Hence Ω_1 is the interior of the union of some simplices. Moreover, Ω_1 contains at least one nodal point B_i, otherwise it would consist of isolated elements such that $v_h \equiv 0$ on their boundary, which would imply $v_h \equiv 0$ in these elements as well due to the linearity of v_h, and we would obtain that $\Omega = \Omega_0$. Let $B_{i_1}, \ldots, B_{i_k}$ be all nodal points in Ω_1, and define

$$w_h := \sum_{j=1}^{k} \phi_{i_j}.$$

Then

$$w_h > 0 \quad \text{in } \Omega_1 \qquad \text{and} \quad w_h = 0 \quad \text{in } \Omega_0$$

by definition. Further, the assumption

$$0 = a(u_h;\ v_h, v_h) = \int_\Omega b(x, u_h, \nabla u_h)\ |\nabla v_h|^2\, dx = \int_{\Omega_1} b(x, u_h, \nabla u_h)\ |\nabla v_h|^2\, dx$$

implies that $b(x, u_h, \nabla u_h) = 0$ on Ω_1, further, $w_h = 0$ in Ω_0, and by assumption $\gamma \leq 0$ on Γ and $f < 0$ a.e. in Ω. These properties can combined with setting $v_h := w_h$ in (25) to obtain a contradiction:

$$0 = \int_{\Omega_0} b(x, u_h, \nabla u_h)\ \nabla u_h \cdot \nabla w_h\, dx + \int_{\Omega_1} b(x, u_h, \nabla u_h)\ \nabla u_h \cdot \nabla w_h\, dx$$

$$= \int_\Omega b(x, u_h, \nabla u_h)\ \nabla u_h \cdot \nabla w_h\, dx = \int_\Omega f w_h\, dx + \int_\Gamma \gamma w_h\, d\sigma \leq \int_{\Omega_1} f w_h\, dx < 0.$$

Hence (29) holds.

(A2) For all $i = 1, \ldots, n$, $j = 1, \ldots, \bar{n}$; $i \neq j$ we have

$$a(u_h;\ \phi_j, \phi_i) = \int_\Omega b(x, u_h, \nabla u_h)\, \nabla \phi_j \cdot \nabla \phi_i\, dx \leq 0$$

since $b(x, u_h, \nabla u_h) \geq 0$ and $\nabla \phi_j \cdot \nabla \phi_i \leq 0$.

(A3)' The linear basis functions corresponding to all nodal points satisfy

$$\sum_{j=1}^{\bar{n}} \phi_j = 1, \tag{30}$$

hence for all $i = 1, \ldots, n$ we have

$$a(u_h;\ \sum_{j=1}^{\bar{n}} \phi_j,\ \phi_i) = \int_\Omega b(x, u_h, \nabla u_h)\ \nabla(\sum_{j=1}^{\bar{n}} \phi_j) \cdot \nabla \phi_i\, dx = 0.$$

(A4) Since each of $f, \gamma_N, \gamma_{int} \leq 0$, and the linear basis functions satisfy

$$\phi_i \geq 0 \qquad (\forall i = 1, \ldots, n),$$

we have

$$\ell\phi_i = \int\limits_{\Omega} f\phi_i \, dx \, + \int\limits_{\Gamma} \gamma\phi_i \, d\sigma \equiv \int\limits_{\Omega} f\phi_i \, dx \, + \int\limits_{\Gamma_N} \gamma_N\phi_i \, d\sigma + \int\limits_{\Gamma_{int}} \gamma_{int}\phi_i \, d\sigma \leq 0.$$

Altogether, we may apply Theorem 6, and we obtain that the coefficient vector of (26) satisfies

$$\max_{i=1,\ldots,\bar{n}} c_i \; = \; \max_{i=n+1,\ldots,\bar{n}} c_i. \tag{31}$$

For linear basis functions the coefficients c_i coincide with the nodal values $u_h(B_i)$, and u_h are linear interpolants between these values, further, the indices $n+1, ..., \bar{n}$ correspond to the nodal points on Γ_D. It readily follows now from (31) that

$$\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} u_h.$$

Since $u_h = g_h$ on Γ_D, this means that (28) holds. ∎

By reversing signs, we obtain the corresponding discrete minimum principle:

Corollary 2 *Let Assumptions 4.1 hold and let the FEM problem (25) have a solution $u_h \in V_h$. Let us consider a nonobtuse simplicial triangulation, i.e. for which (27) holds. If*

$$f > 0 \;\; a.e. \; on \; \Omega, \qquad \gamma_N \geq 0 \;\; a.e. \; on \; \Gamma_N, \quad \gamma_{int} \geq 0 \;\; a.e. \; on \; \Gamma_{int},$$

then u_h satisfies

$$\min_{\overline{\Omega}} u_h = \min_{\Gamma_D} g_h. \tag{32}$$

In particular, if $g_h \geq 0$ then we obtain the discrete nonnegativity principle: $u_h \geq 0$ in Ω.

Remark 2 (i) Various practical and theoretical results related to generation of nonobtuse simplicial partitions are presented e.g. in the survey work [2].

(ii) It is easy to see that the same results hold for bilinear elements in 2D, trilinear elements in 3D and prismatic elements in 3D. In fact, the proof only uses the properties that the basis functions are nonnegative and satisfy (30), and that they are defined via nodal points in the standard way. The crucial condition here is (27), which can be guaranteed by particular "well-shapedness" type geometric conditions for bilinear, trilinear and prismatic elements, see [14, 18, 20].

(iii) (Technical remarks.) The proof shows that it suffices to assume $b \geq 0$ instead of (19), since the equation itself with the assumed property $f > 0$ forces b to attain a.e. positive values on the solution. On the other hand, if we only assumed $f \geq 0$ as in our earlier papers [15, 18], then the proof of condition (A1) would have failed.

Examples. Equations of the form

$$-\operatorname{div}\big(b(x,u,\nabla u)\,\nabla u\big) = f(x)$$

with proper nonlinearities $b(x,u,\nabla u)$ appear in various applications, usually either depending only on (x,u) or on $(x,\nabla u)$ (or even not on x). Nonlinear heat equations generally involve coefficients of the form

$$b(x,u,\nabla u) = k(x,u) > 0.$$

For gradient-dependent nonlinearities most often the dependence on ∇u is via its modulus, i.e.

$$b(x,u,\nabla u) := k(x,\,|\nabla u|) \qquad \text{or simply} \qquad b(x,u,\nabla u) := k(|\nabla u|),$$

depending whether the models involves fields inhomogeneous or homogeneous in space, respectively. In gas dynamics, the density of mass depends on the modulus of speed, i.e. the nonlinearity is

$$b(x,\nabla u) := \varrho(|\nabla u|^2)$$

where the function ϱ is determined by one of Bernoulli's laws, e.g.

$$\varrho(|\eta|^2) \;=\; \varrho_0\,\Big(1+\frac{1}{5}(M^2-|\eta|^2)\Big)^{5/2} \qquad \text{or} \qquad \varrho(|\eta|^2) \;=\; \varrho_0\exp\Big(\frac{|\eta|^2}{2\bar{c}^2}\Big)$$

for adiabatic or isothermal flows, respectively [5], where $\varrho_0, M, \bar{c} > 0$ are physical constants. For the subsonic case we have $\varrho(|\eta|^2) > 0$. A coefficient of the form

$$k(x,|\nabla u|) := k_0(x) + k_1(x)|\nabla u|^2 \qquad (\text{where } k_0(x), k_1(x) > 0)$$

describes dielectric susceptibility in electrorheological fluids [4]. The nonlinearity

$$k(|\nabla u|) := \frac{1}{1+|\nabla u|^2}$$

arises in mean curvature and minimal surface equations, e.g. describing capillary surfaces, see, e.g., [23]; further,

$$k(|\nabla u|) := |\nabla u|^{p-2}$$

(for a given constant $p > 2$) leads to the p-Laplacian, which is a widespread model of nonlinear diffusion operator, arising e.g. for a compressible fluid in a homogeneous isotropic porous medium [29], and is degenerate, i.e. the coefficient $|\nabla u|^{p-2}$ may vanish inside the domain Ω.

We note that interface problems for an elliptic equation typically arise when two distinct materials are involved in subparts of the domain, e.g. in material science or multiphase flow, see [21].

Remark 3 Discrete maximum principles for similar problems have been considered in the earlier papers [8, 15, 16, 18, 20] as well. When compared to these, the main novelties in our work are as follows:

(i) For nonlinearities depending on $(x, |\nabla u|)$ and with potential structure, a generalization of the DMP to the so-called convex hull property is proved in [8]. Further, for similar nonlinearities as ours, without allowing degeneracy, a DMP in 3D is verified in [20]. In both papers, in addition to the above restrictions, the results only concern Dirichlet boundary conditions and do not include interface problems.

(ii) We have proved DMPs for nonlinear problems with mixed boundary conditions in [15, 18] and for interface problems in [16]. Those results assume lower and upper boundedness of the diffusion coefficients, hence do not allow degeneracy and thus are not applicable to such problems, illustrated above in the 'Examples' item.

Acknowledgements. The first author was supported by the by the Hungarian Scientific Research Fund OTKA, No. 112157. The second author was supported by MINECO under Grant MTM2011–24766.

References

[1] Brandts, J., Korotov, S., Křížek, M., The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem, *Linear Algebra Appl.* 429 (2008), 2344–2357.

[2] Brandts, J., Korotov, S., Křížek, M., Šolc, J., On nonobtuse simplicial partitions, *SIAM Rev.* 51(2) (2009), 317–335.

[3] Burman, E., Ern, A., Stabilized Galerkin approximation of convection-diffusion-reaction equations: Discrete maximum principle and convergence, *Math. Comput.* 74 (2005), 1637–1652.

[4] Busuioc, V., Cioranescu, D., On a class of electrorheological fluids. Contributions in honor of the memory of Ennio De Giorgi, *Ricerche Mat.* 49 (2000), suppl., 29–60.

[5] Chen, G. Q., Torres, M., Ziemer, W. P., Measure-Theoretic Analysis and Nonlinear Conservation Laws, *Pure and Applied Mathematics Quarterly*, Volume 3, Number 3, 841–879, 2007.

[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] Diening, L., Kreuzer, C., Schwarzacher, S., Convex hull property and maximum principles for finite element minimizers of general convex functionals, *Numer. Math.*, 124(4), 685–700 (2013).

[9] Drăgănescu, A., Dupont, T. F., Scott, L. R., Failure of the discrete maximum principle for an elliptic finite element problem, *Math. Comp.* 74 (2005), no. 249, 1–23.

[10] Faragó, I., Matrix and Discrete Maximum Principles, in: *LSSC 2009*, LNCS 5910, 563–570 (2010).

[11] 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.

[12] Fiedler, M., *Special Matrices and Their Applications in Numerical Mathematics*: Second Edition, Dover Books on Mathematics, 2008.

[13] Hannukainen, A., Korotov, S., Vejchodský, T., On Weakening Conditions for Discrete Maximum Principles for Linear Finite Element Schemes, in: S. Margenov, L.G. Vulkov, and J. Wasniewski (eds.), *NAA 2008*, LNCS 5434, pp. 297–304, 2009.

[14] Hannukainen, A., Korotov, S., Vejchodský, T., Discrete maximum principles for FE solutions of the diffusion-reaction problem on prismatic meshes, *J. Comput. Appl. Math.* 226 (2009), 275–287.

[15] 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.

[16] Karátson J., Korotov, S., Discrete maximum principles for FEM solutions of some nonlinear elliptic interface problems, *Int. J. Numer. Anal. Modelling.*, 6 (1), 1–16 (2008).

[17] Karátson, J., Korotov, S., A discrete maximum principle in Hilbert space with applications to nonlinear cooperative elliptic systems, *SIAM J. Numer. Anal.* 47 (4), 2518–2549 (2009).

[18] Karátson J., Korotov, S., Křížek, M., On discrete maximum principles for nonlinear elliptic problems, *Math. Comput. Simul.* 76 (2007) pp. 99–108.

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

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

[23] P. Pucci, J. Serrin, *The Strong Maximum Principle*, "Progress in Nonlinear Differential Equations and their Applications", Vol. 73, Birkhauser, Switzerland, 2007.

[24] 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.

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

[26] Vejchodsky, T., The discrete maximum principle for Galerkin solutions of elliptic problems, *Cent. Eur. J. Math.* 10 (1), 2012, 25–43.

[27] Vejchodský, T., Solin, P., Discrete maximum principle for higher-order finite elements in 1D, *Math. Comp.* 76 (2007), no. 260, 1833–1846.

[28] Xu, J., Zikatanov, L., A monotone finite element scheme for convection-diffusion equations, *Math. Comp.* 68 (1999), 1429–1446.

[29] Wu Z., Zhao J., Yin J. and Li H., *Nonlinear diffusion equations*, World Scientific, 2001.

[30] Zeidler, E., *Nonlinear functional analysis and its applications,* Springer, 1986.