Helsinki University of Technology, Institute of Mathematics, Research Reports
Teknillisen korkeakoulun matematiikan laitoksen tutkimusraporttisarja
Espoo 2006 A504

On discrete maximum principles for nonlinear elliptic problems

János Karátson Sergey Korotov Michal Křížek

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 A504

On discrete maximum principles for nonlinear elliptic problems

János Karátson **Sergey Korotov** **Michal Křížek**

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

János Karátson, Sergey Korotov, Michal Křížek: *On discrete maximum principles for nonlinear elliptic problems* ; Helsinki University of Technology, Institute of Mathematics, Research Reports A504 (2006).

Abstract: *In order to have reliable numerical simulations it is very important to preserve basic qualitative properties of solutions of mathematical models by computed approximations. For scalar second-order elliptic equations, one of such properties is the maximum principle. In our work, we give a short review of the most important results devoted to discrete counterparts of the maximum principle (called discrete maximum principles, DMPs), mainly in the framework of the finite element method, and also present our own recent results on DMPs for a class of second-order nonlinear elliptic problems with mixed boundary conditions.*

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

Keywords: nonlinear elliptic problem, mixed boundary conditions, discrete maximum principle, finite element method, quadratures.

Correspondence

karatson@cs.elte.hu, sergey.korotov@hut.fi, krizek@math.cas.cz

ISBN 951-22-8314-X
ISSN 0784-3143

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 is an important feature of scalar second-order elliptic equations that distinguishes them from higher order equations and systems of equations (cf. [10, 40]). The principle, in its simplest form, was first discovered for harmonic functions: any nonconstant harmonic function u (i.e. $\Delta u = 0$) takes its minimum and maximum values only on the boundary $\partial\Omega$ of any bounded domain Ω in which $u \in C(\overline{\Omega})$,

$$\min_{s\in\partial\Omega} u(s) < u(x) < \max_{s\in\partial\Omega} u(s) \qquad \forall x \in \Omega. \tag{1}$$

The relation (1) gives, in fact, an *a priori estimate* for $u(x)$ in Ω via its values on $\partial\Omega$.

Later, (continuous) maximum principles (denoted by CMPs from now on) were formulated for various second order boundary value problems (BVPs) (see, e.g., [10, 24, 26, 27, 30]). For convenience of presentation, in what follows we introduce the following simple BVP of elliptic type

$$-\Delta u + au = f \quad \text{in} \quad \Omega \qquad \& \qquad u = g \quad \text{on} \quad \partial\Omega, \tag{2}$$

where the constant coefficient $a \geq 0$, f and g are given functions.

The paper [39] by R. Varga in 1966 was the first publication (to the authors' knowledge) devoted to the construction of discrete analogues of maximum principles, usually called *discrete maximum principles (DMPs)*. In short, that paper deals with the case $f \equiv 0$ in (2), and analyses CMP & DMP in the following forms

$$\max_{x\in\overline{\Omega}} |u(x)| \leq \max_{s\in\partial\Omega} |g(s)| \qquad \& \qquad \max_{i=1,\dots,N} |u_i| \leq \max_{j=1,\dots,N^\partial} |g_j|,$$

where u_i are values of the finite difference (FD) solution at interior nodes, g_j are values of g at boundary nodes, and N^∂ denotes the number of boundary nodes. *Sufficient conditions* for the validity of the above DMP were given in [39] in terms of the matrix appearing in the FD discretization.

However (cf. [24]), for problem (2), the corresponding CMP, in fact, takes a more sophisticated form

$$\max_{x\in\overline{\Omega}} u(x) \leq \max\{0, \max_{s\in\partial\Omega} g(s)\}, \text{ or } \max_{x\in\overline{\Omega}} u(x) = \max_{s\in\partial\Omega} g(s), \tag{3}$$

provided the *sign-condition* $f \leq 0$ holds. Therefore, later, in works by Ciarlet [6] and Ciarlet & Raviart [7] (of 1970 and 1973, respectively), a more adequate form of DMP adopted to the maximum principle (3) was proposed for FD and finite element (FE) approximations. In particular, for linear simplicial finite elements it states

$$\max_{x\in\overline{\Omega}} u_h(x) \leq \max\{0, \max_{s\in\partial\Omega} g_h(s)\}, \text{ or } \max_{x\in\overline{\Omega}} u_h(x) = \max_{s\in\partial\Omega} g_h(s), \tag{4}$$

where u_h is a FE solution and g_h is a continuous piecewise linear approximation of g. In both papers several sets of sufficient conditions providing the validity of DMP (4) were given. In particular, in [7] simplicial meshes and piecewise linear continuous FE approximations were used, and for the first time sufficient *geometric conditions of nonobtuseness or acuteness of triangular elements* (depending on the coefficient a) were obtained.

Later, various generalizations of the above mentioned results were done. Thus, Lorenz [28] in 1977 and Höhn & Mittelmann [12] in 1981 made attempts to derive similar geometrical conditions under which relevant DMPs hold for approximations obtained with the help of higher order finite elements. Unfortunately, even for the simplest case ($a \equiv 0$ & $f \equiv 0$ in (2)), their (sufficient) conditions on the triangular meshes turned out to be very stringent (only right or equilateral triangles are allowed) and, thus, hardly employed in real computations. However, in a very recent work by Šolín & Vejchodský [35] a new attempt to consider some *weaker form of DMP* for higher order finite elements is done for 1D case.

Further, Christie & Hall [5] in 1984 considered the case of bilinear FE approximations for problem (2) with $a \equiv 0$ & $f \equiv 0$. In fact, the notion of *non-narrow rectangular element* was introduced there as a sufficient geometric condition for the corresponding DMP to hold.

Next efforts in the analysis of DMPs were done by Křížek & Qun Lin [20] in 1995. For $f \leq 0$ and a sufficiently smooth function $b \quad (0 < \mu_0 \leq b \leq \mu_1)$ they considered the following 3D nonlinear elliptic problem:

$$- \operatorname{div}\Big(b(x,u,\nabla u)\nabla u\Big) = f \quad \text{in} \quad \Omega \subset \mathbf{R}^3 \qquad \& \qquad u = 0 \quad \text{on} \quad \partial\Omega,$$

for which the corresponding CMP and DMP take the form $u \leq 0$ and $u_h \leq 0$ (for linear elements), respectively. In addition, the effect of quadrature rules was analysed there and DMP was proved under the condition of nonobtuseness of the tetrahedral meshes used.

In all the above mentioned papers, only the cases of linear problems (besides [20]) with pure Dirichlet boundary conditions were analysed. Then, Karátson & Korotov [14, 15] in 2005 considered a more general case of second-order nonlinear elliptic problems with *mixed boundary conditions* in arbitrary space dimension. They formulated and proved the corresponding CMPs and DMPs (also taking into account the effect of numerical integration).

The present paper is organized as follows. In Section 2 we introduce a nonlinear elliptic problem with mixed boundary conditions which is of more general type than those analysed in [15] and [20], and formulate the corresponding CMP. In Section 3 the relevant FE discretization scheme is shortly described. In Section 4 we formulate sufficient conditions under which the corresponding DMP holds. In Section 5 we consider examples of simplicial and block meshes and, in particular, show that Q_1-block elements in higher dimensions do not yield irreducibly diagonally dominant stiffness matrix even for the Poisson equation. In Section 6 we discuss various weakening conditions.

Other works devoted to various aspects of DMPs and related issues include [1], [8], [13], [21], [22], [31], [32], [33], and [34]. Several examples of real-life problems for which the validity of DMP is essential are given, e.g., in papers [3], [14], [25], and [37]. Finally, let us point out that if the DMP is not valid, then some pathological nonphysical situations may appear. For instance, the numerical heat could flow from colder parts of the body to hotter parts (see [22]).

2 The continuous problem and maximum principle

We consider the following nonlinear elliptic boundary value problem with mixed boundary conditions:

$$\begin{cases} -\operatorname{div}\Big(b(x,u,\nabla u)\,\nabla u\Big) = f & \text{in } \Omega, \\ b(x,u,\nabla u)\frac{\partial u}{\partial \nu} = \gamma & \text{on } \Gamma_N, \\ u = g & \text{on } \Gamma_D, \end{cases} \tag{5}$$

where Ω is a bounded polytopic domain in $\mathbf{R}^d$ with Lipschitz continuous boundary $\partial\overline{\Omega} = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, contained in a finite number of hyperplanes, Γ_D and Γ_N are relatively open sets in $\partial\overline{\Omega}$ such that $\operatorname{meas}_{d-1}\overline{\Gamma}_D \cap \overline{\Gamma}_N = 0$, $\Gamma_D \neq \emptyset$, ν denotes the outward unit normal, $f \in L^2(\Omega)$, $\gamma \in L^2(\Gamma_N)$, and $g = g^*{}_{|\Gamma_D}$ with $g^* \in H^1(\Omega)$. The function b is measurable and it satisfies $0 < \mu_0 \leq b(\cdot,\cdot,\cdot) \leq \mu_1$. Problem (5) is of more general type than those in [15] and [20].

We assume that (5) has a unique weak solution $u \in H^1(\Omega)$ such that

$$u = g \quad \text{on } \Gamma_D \ \text{ in the sense of traces, and} \tag{6}$$

$$\int_\Omega b(x,u,\nabla u)\ \nabla u \cdot \nabla v\, dx = \int_\Omega f v\, dx + \int_{\Gamma_N} \gamma v\, ds \qquad \forall v \in H^1_D(\Omega), \tag{7}$$

where $H^1_D(\Omega) := \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_D\}$. Sufficient conditions to guarantee a unique solvability of this problem are given, e.g., in [9], [10], [11], [24], and [29].

For the classical formulation of the CMP below, we assume hereafter that $u \in C^1(\Omega) \cap C(\overline{\Omega})$ (see [10], [24] for sufficient conditions). If this fails to hold, then the results remain true by replacing all max and min by ess sup and ess inf, respectively, provided that g is bounded on Γ_D.

The following CMP has been proposed and proved in [14] for the problem of the above type when b does not depend explicitly on u. If the sign-conditions

$$f(x) \leq 0 \text{ for a.e. } x \in \Omega \quad \text{and} \quad \gamma(s) \leq 0 \text{ for a.e. } s \in \Gamma_N, \tag{8}$$

hold, then

$$\max\nolimits_{\overline{\Omega}} u = \max\nolimits_{\Gamma_D} g. \tag{9}$$

In [10, p. 206] a similar relation is proved also for b dependent on u. A related nonnegativity result for linear problems is developed in [26].

In what follows, we shall analyse a natural discrete analogue to (9) of the form

$$\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} g_h. \tag{10}$$

Note that the corresponding minimum principle obviously holds if the sign conditions in (8) are reversed. Further, if f and γ are zero constants then both the the maximum and minimum principles are valid. That is, we have

Corollary 1. *Let the weak solution u of problem* (5) *satisfy $u \in C^1(\Omega) \cap C(\overline{\Omega})$. Then*

1) *If $f \geq 0$ and $\gamma \geq 0$, then* $\min\limits_{\overline{\Omega}} u = \min\limits_{\Gamma_D} g$.

2) *If $f = 0$ and $\gamma = 0$, then the ranges of u and g coincide, i.e., we have* $[\min\limits_{\overline{\Omega}} u, \max_{\overline{\Omega}} u] = [\min\limits_{\Gamma_D} g, \max_{\Gamma_D} g]$.

3 FE discretization with quadratures

We shortly present the finite element discretization scheme for problem (5). Let $\mathcal{T}_h$ denote a partition of $\overline{\Omega}$ into simplicial or rectangular finite elements. We assume that Γ_D and Γ_N have only a finite number of components and that the faces of elements lying on the boundary $\partial\overline{\Omega}$ are subsets of either $\overline{\Gamma}_D$ or $\overline{\Gamma}_N$. Further, by $\mathcal{S}_h$ we denote the set of those faces of elements from $\mathcal{T}_h$ which belong to $\overline{\Gamma}_N$.

Using continuous basis functions ϕ_i, $i = 1, \ldots, \bar{n}$, that are piecewise linear on each element, we define $V_h = \text{span}\{\phi_1, \ldots, \phi_{\bar{n}}\} \subset H^1(\Omega)$. The main properties of the basis functions are

$$\phi_i \geq 0, \quad i = 1, \ldots, \bar{n}, \qquad \sum_{j=1}^{\bar{n}} \phi_j \equiv 1, \tag{11}$$

and that there exist nodal points (the mesh vertices) $B_i \in \overline{\Omega}$, $i = 1, \ldots, \bar{n}$, such that

$$\phi_i(B_j) = \delta_{ij}, \tag{12}$$

where δ_{ij} is Kronecker's symbol.

Let for $i = 1, \ldots, n$ the basis functions $\phi_i \in H^1_D(\Omega)$, and let $\phi_{n+1}, \ldots, \phi_{\bar{n}}$ be those having nonzero values on Γ_D. Let the nodal points satisfy

$$B_{n+1}, \ldots, B_{\bar{n}} \in \overline{\Gamma}_D\,. \tag{13}$$

We define $V_h^0 = \text{span}\{\phi_1, \ldots, \phi_n\} \subset H^1_D(\Omega)$. Further, let $g_h = \sum\limits_{j=n+1}^{\bar{n}} g_j\phi_j \in V_h$ (with $g_j \in \mathbf{R}$) be the approximation of the function g^*.

The FE approximation u_h is defined as a function satisfying the conditions

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

$$\int_\Omega b(x, u_h, \nabla u_h) \, \nabla u_h \cdot \nabla v_h \, dx = \int_\Omega f v_h \, dx + \int_{\Gamma_N} \gamma v_h \, ds \quad \forall v_h \in V_h^0. \tag{14}$$

We look for u_h in the form

$$u_h = \sum_{j=1}^{\bar{n}} c_j \phi_j. \tag{15}$$

For any $\bar{\mathbf{c}} = [c_1, \ldots, c_{\bar{n}}]^T = [\mathbf{c}, \tilde{\mathbf{c}}]^T$, where $\mathbf{c} \in \mathbf{R}^n$ and $\tilde{\mathbf{c}} \in \mathbf{R}^{\bar{n}-n}$, $i = 1, \ldots, n$, and $j = 1, \ldots, \bar{n}$, we put

$$a_{ij}(\bar{\mathbf{c}}) = \int_\Omega b(x, \sum_{k=1}^{\bar{n}} c_k \phi_k, \sum_{k=1}^{\bar{n}} c_k \nabla \phi_k) \, \nabla \phi_j \cdot \nabla \phi_i \, dx, \quad d_i = \int_\Omega f \phi_i \, dx + \int_{\Gamma_N} \gamma \phi_i \, ds \,. \tag{16}$$

However, in practice the integrals in (16) have to be computed numerically using certain quadratures. Thus, we approximate the integral $\int_\Omega \psi \, dx$ by the sum

$$Q_1(\psi) := \sum_{T \in \mathcal{T}_h} \text{meas}_d(T) \sum_{k=1}^{K} \omega_{T,k} \, \psi(x_{T,k}) \,, \tag{17}$$

where for each element $T \in \mathcal{T}_h$ one chooses nodes $x_{T,k} \in T$ and weights $\omega_{T,k} \in \mathbf{R}$, $k = 1, \ldots, K$, such that $\omega_{T,k} > 0$ and $\sum_{k=1}^{K} \omega_{T,k} = 1$. We can similarly approximate the integral $\int_{\Gamma_N} \varphi \, ds$ by

$$Q_2(\varphi) := \sum_{S \in \mathcal{S}_h} \text{meas}_{d-1}(S) \sum_{\ell=1}^{L} \sigma_{S,\ell} \, \varphi(x_{S,\ell}) \,, \tag{18}$$

where for each face $S \in \mathcal{S}_h$ one chooses nodes $x_{S,\ell} \in S$ and weights $\sigma_{S,\ell} \in \mathbf{R}$, $\ell = 1, \ldots, L$, such that $\sigma_{S,\ell} > 0$ and $\sum_{\ell=1}^{L} \sigma_{S,\ell} = 1$.

Thus, the integrals in (16) are replaced by

$$\hat{a}_{ij}(\bar{\mathbf{c}}) = Q_1 \left(b(x, \sum_{k=1}^{\bar{n}} c_k \phi_k, \sum_{k=1}^{\bar{n}} c_k \nabla \phi_k) \, \nabla \phi_j \cdot \nabla \phi_i \right), \quad \hat{d}_i = Q_1 \left(f \phi_i \right) + Q_2 \left(\gamma \phi_i \right). \tag{19}$$

Now using the notations

$$\mathbf{Q}(\bar{\mathbf{c}}) = \{\hat{a}_{ij}(\bar{\mathbf{c}})\}, \; i, j = 1, \ldots, n, \quad \tilde{\mathbf{Q}}(\bar{\mathbf{c}}) = \{\hat{a}_{ij}(\bar{\mathbf{c}})\}, \; i = 1, \ldots, n, \; j = n+1, \ldots, \bar{n}, \tag{20}$$

$$\mathbf{q} = [\hat{d}_1, \ldots, \hat{d}_n]^T, \quad \tilde{\mathbf{q}} = [g_{n+1}, \ldots, g_{\bar{n}}]^T,$$

we come to the nonlinear system

$$\begin{bmatrix} \mathbf{Q}(\bar{\mathbf{c}}) & \tilde{\mathbf{Q}}(\bar{\mathbf{c}}) \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{q} \\ \tilde{\mathbf{q}} \end{bmatrix}, \tag{21}$$

where $\mathbf{I}$ denotes the $(\bar{n}-n)\times(\bar{n}-n)$ identity matrix and $\mathbf{0}$ is the $(\bar{n}-n)\times n$ zero matrix.

Remark 1. If only the homogeneous Dirichlet boundary condition is considered, i.e., $\tilde{\mathbf{q}} = \mathbf{0}$, then (21) reduces to a smaller system $\mathbf{Q}(\bar{\mathbf{c}}) = \mathbf{q}$. Such a system is considered in [20].

The vector-solution $\bar{\mathbf{c}}^* = [c_1^*, \ldots, c_{\bar{n}}^*]^T$ of system (21) defines the approximate solution

$$\hat{u}_h = \sum_{j=1}^{\bar{n}} c_j^* \phi_j. \tag{22}$$

In general, we can consider quadratures defined as functionals

$$Q_1 : PC(\Omega) \to \mathbf{R} \qquad \text{and} \qquad Q_2 : PC(\Gamma_N) \to \mathbf{R}, \tag{23}$$

where $PC(\cdot)$ denotes piecewise continuous functions on the corresponding set. We define the following properties for $r = 1, 2$:

(P1) Q_r is linear,

(P2) Q_r is monotone, i.e., if $q_1 \geq q_2$ then $Q_r(q_1) \geq Q_r(q_2)$,

(P3) Q_1 is strictly positive on the subspace V_h^0 in the sense that for any $v_h \in V_h^0$,

$$Q_1\left(|\nabla v_h|^2\right) = 0 \quad \text{implies} \quad v_h \equiv 0.$$

Remark 2. The exact integration satisfies the following well-known strict positivity property: if $f \geq 0$ is a Lebesgue integrable function and $\displaystyle\int_\Omega f\,dx = 0$, then $f \equiv 0$ a.e. in Ω. In particular, if $v \in H_D^1(\Omega)$ then $\displaystyle\int_\Omega |\nabla v|^2\,dx = 0$ implies $v \equiv \mathit{const}$ and from the condition $v|_{\Gamma_D} = 0$ we have $v \equiv 0$ a.e., that is, the analogue of (P3) holds. Clearly, one cannot require the previous strict positivity for all integrable f, since quadratures like (17) and (18) are zero for any function with support outside their nodes. Property (P3) is a natural requirement, since it ensures (together with (P1)) that the trace of the Sobolev norm in V_h under Q_1, i.e.,

$$\|v_h\|_{Q_1}^2 := Q_1\Big(|\nabla v_h|^2\Big),$$

defines a norm on V_h, induced by the inner product $\langle u_h, v_h\rangle_{Q_1} := Q_1\Big(\nabla u_h \cdot \nabla v_h\Big)$.

Proposition 1. *Quadratures* (17) *and* (18) *satisfy properties* (P1)–(P2). *Further, all those quadratures* (17) *that are exact for polynomials of degree* $2s-2$*, where* s *is the maximum degree of the piecewise polynomials in* V_h^0*, satisfy property* (P3).

PROOF. (P1) and (P2) are obvious. Further, if (17) is exact for polynomials up to degree $2s-2$, then $Q_1(|\nabla v_h|^2) = \int_\Omega |\nabla v_h|^2\,dx$, because $|\nabla v_h|^2$ has degree at most $2s-2$. This being zero, as pointed out above in Remark 2, implies $v_h \equiv 0$. ■

Remark 3. Any quadrature (17) is exact for piecewise constant functions and hence, Proposition 1 is valid for piecewise linear finite element functions.

4 Discrete maximum principle

Definition 1. 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 [38]:

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

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

3) for at least one subscript $i_0 \in \{1, \dots, n\}$ the above inequality is strict, i.e.,

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

Now, we present a main theorem, on which various known results about discrete maximum principles are based (e.g., [6, 7, 14, 15, 20]). Let us consider a system of linear algebraic equations of order $(n+m)\times(n+m)$

$$\bar{\mathbf{A}}\bar{\mathbf{c}} = \bar{\mathbf{d}}, \quad \text{where} \quad \bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \tilde{\mathbf{A}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{24}$$

where $\mathbf{I}$ is an $m \times m$ identity matrix, and $\mathbf{0}$ is a $m \times n$ zero matrix (cf. (21)).

Theorem 1. *Let* $\bar{\mathbf{A}}$ *be a* $(n+m)\times(n+m)$ *matrix with structure as in* (24). *Assume that*

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

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

(iii) $\sum_{j=1}^{n+m} a_{ij} = 0, \quad i = 1, \ldots, n,$

(iv) $\mathbf{A}$ *is irreducibly diagonally dominant.*

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

$$\max\nolimits_{i=1,\ldots,n+m} c_i \; = \; \max\nolimits_{j=n+1,\ldots,n+m} c_j. \tag{25}$$

For the proof see [6].

Theorem 2. *Let the basis functions satisfy the following property:*

$$\nabla\phi_i \cdot \nabla\phi_j \leq 0 \quad \text{for any} \quad i = 1, \ldots, n, \; j = 1, \ldots, \bar{n}, \; i \neq j. \tag{26}$$

Then the matrix defined in (20)–(21)*, where quadrature* Q_1 *satisfies properties* (P1) − (P3)*, has the following properties:*

(i) $\hat{a}_{ii}(\bar{\mathbf{c}}) > 0, \quad i = 1, \ldots, n,$

(ii) $\hat{a}_{ij}(\bar{\mathbf{c}}) \leq 0, \quad i = 1, \ldots, n, \; j = 1, \ldots, \bar{n}, \; i \neq j,$

(iii) $\sum_{j=1}^{\bar{n}} \hat{a}_{ij}(\bar{\mathbf{c}}) = 0, \quad i = 1, \ldots, n,$

(iv) *there exists an index* $i_0 \in \{1, \ldots, n\}$ *for which* $\sum_{j=1}^{n} \hat{a}_{i_0,j}(\bar{\mathbf{c}}) > 0,$

(v) $\mathbf{Q}(\bar{\mathbf{c}})$ *is irreducible.*

Proof. (i) From the assumption

$$0 < \mu_0 \leq b \leq \mu_1, \tag{27}$$

properties (P2)–(P3) of the quadrature rules and positivity of the basis functions over their supports, we have

$$\hat{a}_{ii}(\bar{\mathbf{c}}) \geq \mu_0 \; Q_1 \left(|\nabla\phi_i|^2 \right) = \mu_0 \, \|\phi_i\|_{Q_1}^2 \; > 0.$$

(ii) Let $i = 1, \ldots, n, \; j = 1, \ldots, \bar{n}$ with $i \neq j$. Then inequalities (26) and (27) imply $b(x, u_h, \nabla u_h) \; \nabla\phi_i \cdot \nabla\phi_j \leq 0$, hence by property (P2) we have $\hat{a}_{ij}(\bar{\mathbf{c}}) \leq 0$.

(iii) For any $i = 1, \ldots, n$,

$$\sum_{j=1}^{\bar{n}} \hat{a}_{ij}(\bar{\mathbf{c}}) \; = \; Q_1 \left(b(x, u_h, \nabla u_h) \; \nabla\phi_i \cdot \nabla \Big(\sum_{j=1}^{\bar{n}} \phi_j \Big) \right) = 0, \tag{28}$$

using (11) and property (P1).

(iv) We first verify that $\mathbf{Q}(\bar{\mathbf{c}})$ is positive definite. Let $\mathbf{p} = (p_1, \ldots, p_n) \in \mathbf{R}^n$ and $v_h = \sum_{i=1}^{n} p_i \phi_i$. Then, by all three properties (P1)–(P3), we observe

$$\mathbf{Q}(\bar{\mathbf{c}})\mathbf{p} \cdot \mathbf{p} = \sum_{i,j=1}^{n} \hat{a}_{ij}(\bar{\mathbf{c}}) p_i p_j = Q_1\Big(b(x, u_h, \nabla u_h)\ |\nabla v_h|^2\Big) \geq$$

$$\geq \mu_0\ Q_1\left(|\nabla v_h|^2\right)\ = \mu_0 \, \|v_h\|_{Q_1}^2 > 0$$

unless $\mathbf{p} = \mathbf{0}$. Assume to the contrary that $\sum_{j=1}^{n} \hat{a}_{ij}(\bar{\mathbf{c}}) = 0$ for all $i = 1, \ldots, n$. This means that $\mathbf{Q}(\bar{\mathbf{c}})$ carries the n-tuple of ones $[1, \ldots, 1]^T$ into the zero vector. This is impossible, since $\mathbf{Q}(\bar{\mathbf{c}})$ is positive definite and, hence, nonsingular.

(v) This follows in the same way as in [20]. Namely, suitable intersections of the supports of the basis functions ϕ_i define a usual partition of $\overline{\Omega}$ into subdomains (here (11) ensures that the union of these subdomains is $\overline{\Omega}$ indeed). For such triangulations the directed graph of the corresponding matrix $\mathbf{Q}(\bar{\mathbf{c}})$ is strongly connected, and, hence, the matrix is irreducible. ∎

Theorem 3. *Let conditions* (i)–(v) *of Theorem* 2 *hold and let*

$$f(x) \leq 0,\ x \in \Omega, \quad \textit{and} \quad \gamma(s) \leq 0,\ s \in \Gamma_N. \tag{29}$$

Then for the numerical solution $\hat{u}_h$ *defined by* (22) *with quadrature rule* Q_2 *satisfying* (P1) – (P2)*, it is valid that*

$$\max_{\overline{\Omega}} \hat{u_h} = \max_{\Gamma_D} g_h. \tag{30}$$

PROOF. We verify that the conditions of Theorem 1 are satisfied for the system (21). Namely, conditions (i)–(iii) of Theorem 1 coincide with the statements (i)–(iii) of Theorem 2. Further, the three criteria of Definition 1 are also fulfilled, namely, $\mathbf{Q}(\mathbf{c})$ is irreducible due to statement (v), and the two other criteria follow from statements (ii)–(iv) under the signs obtained in statements (i)–(ii). Finally, using (11) and (29), property (P2) of the quadratures and (19) imply that $\hat{d}_i \leq 0$ for all i, i.e., $\mathbf{q} \leq 0$. Hence, (21) yields $\mathbf{Q}(\bar{\mathbf{c}})\mathbf{c} + \tilde{\mathbf{Q}}(\bar{\mathbf{c}})\tilde{\mathbf{c}} \leq 0$ and thus, Theorem 1 states that

$$\max_{i=1,\ldots,\bar{n}} c_i^* \ = \ \max_{j=n+1,\ldots,\bar{n}} c_j^*. \tag{31}$$

Since $c_j^* = g_j$ for all $j = n+1, \ldots, \bar{n}$, we obtain

$$\max_{i=1,\ldots,\bar{n}} c_i^* \ = \ \max_{j=n+1,\ldots,\bar{n}} g_j. \tag{32}$$

Then, using (22), we can prove (30). Namely, let j_0 be the subscript such that

$$c_{j_0}^* = \max_{j=1,\ldots,\bar{n}} c_j^* \ . \tag{33}$$

By (31) j_0 can be chosen so that $n+1 \le j_0 \le \bar{n}$. Relations (12) and (22) imply

$$\hat{u_h}(B_{j_0}) = \sum_{j=1}^{\bar{n}} c_j^* \phi_j(B_{j_0}) = c_{j_0}^* \,. \tag{34}$$

Hence, (13) yields $B_{j_0} \in \overline{\Gamma}_D$. Further, (11), (22), and (33) yield

$$\hat{u_h} = \sum_{j=1}^{\bar{n}} c_j^* \phi_j \le c_{j_0}^* \sum_{j=1}^{\bar{n}} \phi_j = c_{j_0}^* \tag{35}$$

pointwise on $\overline{\Omega}$. Altogether, we have

$$\max_{\overline{\Omega}} \hat{u_h} = \hat{u_h}(B_{j_0}) = c_{j_0}^* = \max_{\Gamma_D} u_h = \max_{\Gamma_D} g_h. \qquad \blacksquare$$

Corollary 2. *Under conditions* (i)–(v) *of Theorem* 2 *the following results hold:*

(1) *If $f \ge 0$ and $\gamma \ge 0$, then* $\min\limits_{\overline{\Omega}} \hat{u_h} = \min\limits_{\Gamma_D} g_h$.

(2) *If $f = 0$ and $\gamma = 0$, then the ranges of $\hat{u}_h$ and g_h coincide, i.e., we have* $[\min\limits_{\overline{\Omega}} \hat{u_h}, \max_{\overline{\Omega}} \hat{u_h}] = [\min\limits_{\Gamma_D} g_h, \max_{\Gamma_D} g_h]$.

5 Examples of FE meshes yielding DMP

5.1 Linear simplicial elements

In the case of linear finite elements, condition (26) allows a nice geometric interpretation. First, we can show that (see [2] for the proof)

$$\nabla\phi_i \cdot \nabla\phi_j|_T = -\frac{\mathrm{meas}_{d-1}(S_i) \cdot \mathrm{meas}_{d-1}(S_j)}{d^2(\mathrm{meas}_d(T))^2} \cos(S_i, S_j) \quad \text{for } i \ne j, \tag{36}$$

where T is a d-dimensional simplex with vertices $P_1, \dots, P_{d+1}$, 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 (26) it is sufficient if the employed simplicial mesh is nonobtuse (cf. [2, 7, 16, 17, 18, 20]).

5.2 Multi-linear block elements

Let $B = (0, b_1) \times (0, b_2) \times ... \times (0, b_d)$ be a d-dimensional block element with volume $b = \Pi_{i=1}^d b_i$. The Q_1 - block finite element has 2^d local basis functions. For our purposes we shall deal only with some of them, namely, we set

$$\phi^0(x_1, \dots, x_d) = \frac{1}{b}\,\Pi_{i=1}^d x_i,\ \phi^j(x_1, \dots, x_d) = \frac{1}{b}(b_j - x_j)\,\Pi_{i\ne j}^d x_i \quad \text{for } j = 1, \dots, d.$$

It is easy to calculate the following entries of the element stiffness matrix

$$a_{0j} = \int_B \nabla\phi^0 \cdot \nabla\phi^j \, dx = \frac{b}{3^{d-1}} \Big(\sum_{i \neq j}^{d} \frac{1}{2b_i^2} - \frac{1}{b_j^2} \Big). \tag{37}$$

Example 1 Let $d = 2$. Then

$$a_{01} = \frac{b_1}{6b_2} - \frac{b_2}{6b_1}, \quad a_{02} = \frac{b_2}{6b_1} - \frac{b_1}{6b_2}.$$

From this we find that $a_{01} \leq 0$ and $a_{02} \leq 0$ if $\frac{1}{\sqrt{2}}b_2 \leq b_1 \leq \sqrt{2}b_2$. It is well-known [5] that this is a sufficient condition for the validity of DMP for bilinear rectangular elements, since the global stiffness matrix is monotone.

Proposition 2. *If $d = 3$ and $a_{0j} \leq 0$ for $j = 1, 2, 3$, then B is a cube.*

Proof: By (37), we see that $a_{01} = \frac{b}{9}\Big(\frac{1}{2b_2^2} + \frac{1}{2b_3^2} - \frac{1}{b_1^2}\Big)$, etc. From this and the inequalities $a_{0j} \leq 0,\ j = 1, 2, 3$, we get relations

$$b_1^2\, b_3^2 + b_1^2\, b_2^2 \leq 2\, b_2^2\, b_3^2, \quad b_1^2\, b_2^2 + b_2^2\, b_3^2 \leq 2\, b_1^2\, b_3^2, \quad b_1^2\, b_3^2 + b_2^2\, b_3^2 \leq 2\, b_1^2\, b_2^2. \tag{38}$$

Without loss of generality we may assume that $b_1 \leq b_2 \leq b_3$. Then from the last inequality of (38), we find that $2b_1^2\, b_2^2 \geq b_3^2(b_1^2 + b_2^2) \geq 2b_3^2\, b_1\, b_2$. Hence, $b_1 b_2 \geq b_3^2$. From this and the assumptions on b_i, we obtain $b_1 = b_2 = b_3$. ■

It is easy to calculate that the global stiffness matrix will be monotone for trilinear elements on cubes. However, for general block elements some off-diagonal entries are positive, which may destroy the validity of DMP.

Remark 4. For $d \geq 4$, we can never get $a_{0j} \leq 0$ for all $j = 1, \ldots, d$ due to (37). However, using Kuhn's decomposition (see [23]), each block can be subdivided into $d!$ nonobtuse simplices and by (36) and Theorems 2 and 3 we get the validity of DMP. Another possibility would be to use a weakened form of DMP (cf. Section 6).

6 On weakening geometric conditions

DMP may also hold for continuous piecewise linear finite element approximations for elliptic problems under various weaker conditions on the simplicial meshes used. In particular, in certain situations, not too large obtuse interior angles in the simplices of the meshes are acceptable, see [19], [31], and [36, p. 78]. Then DMP holds even though some off-diagonal entries of the corresponding stiffness matrix are positive.

An interesting trick that helps to avoid any geometric restrictions on the mesh, in order to provide the validity of the DMP, is proposed in [4]. However, even for simple linear problem the discretization scheme becomes nonlinear, and moreover, it is not clear if the approximate solution obtained due to this scheme is unique.

Acknowledgements: The first author was supported by the Hungarian Research Fund OTKA under grant no. T 043765. The second author was supported by the Academy Research Fellowship no. 208628 from the Academy of Finland. The third author was supported by the Grant no. A1019201 of the Grant Agency of the Academy of Sciences of the Czech Republic and the Institutional Research Plan AV0Z 10190503.

References

[1] V.S. Borisov, On discrete maximum principles for linear equation systems and monotonicity of difference schemes, SIAM J. Matrix Anal. Appl. 24 (2003) 1110–1135.

[2] J. Brandts, S. Korotov, M. Křížek, Dissection of the path-simplex in $\mathbf{R}^n$ into n path-simplices, Preprint A496, Helsinki University of Technology, 2006.

[3] E. Burman, A. Ern, Nonlinear diffusion and the discrete maximum principle for stabilized Galerkin approximations of the convection-diffusion-reaction equation, Comput. Methods Appl. Mech. Engrg. 191 (2002) 3833–3855.

[4] E. Burman, A. Ern, Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes, C. R. Math. Acad. Sci. Paris 338 (2004) 641–646.

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

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

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

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

[9] I. Faragó, J. Karátson, Numerical Solution of Nonlinear Elliptic Problems via Preconditioning Operators. Theory and Applications, Advances in Computation, Volume 11, NOVA Science Publishers, New York, 2002.

[10] D. Gilbarg, N.S. Trudinger, Elliptic Partial Differential Equations of Second Order, Grundlehren der Mathematischen Wissenschaften 224, Springer, Berlin, 1977.

[11] I. Hlaváček, M. Křížek, J. Malý, On Galerkin approximations of a quasilinear nonpotential elliptic problem of a nonmonotone type, J. Math. Anal. Appl. 184 (1994) 168–189.

[12] W. Höhn, H.-D. Mittelmann, Some remarks on the discrete maximum-principle for finite elements of higher order, Computing 27 (1981) 145–154.

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

[14] 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.

[15] J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of some mixed nonlinear elliptic problems using quadratures, J. Comput. Appl. Math. 192 (2006) 75–88.

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

[17] S. Korotov, M. Křížek, Local nonobtuse tetrahedral refinements of a cube, Appl. Math. Lett. 16 (2003) 1101–1104.

[18] S. Korotov, M. Křížek, Global and local refinement techniques yielding nonobtuse tetrahedral partitions, Comput. Math. Appl. 50 (2005) 193–207.

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

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

[21] M. Křížek, L. Liu, On a comparison principle for a quasilinear elliptic boundary value problem of a nonmonotone type, Appl. Math. (Warsaw) 24 (1996) 97–107.

[22] M. Křížek, L. Liu, On the maximum and comparison principles for a steady-state nonlinear heat conduction problem, ZAMM Z. Angew. Math. Mech. 83 (2003) 559–563.

[23] H.W. Kuhn, Some combinatorial lemmas in topology, IBM Journal of Research and Development 45 (1960) 518–524.

[24] O.A. Ladyzhenskaya, N.N. Ural'tseva, Linear and Quasilinear Elliptic Equations, Leon Ehrenpreis Academic Press, New York-London, 1968.

[25] M. Lobo, A.F. Emery, The discrete maximum principle in finite-element thermal radiation analysis, Numer. Heat Transfer 24 (1993) 209–227.

[26] J. López-Gómez, Classifying smooth supersolutions for a general class of elliptic boundary value problems, Adv. Differential Equations 8 (2003) 1025–1042.

[27] J. López-Gómez, M. Molina-Meyer, The maximum principle for cooperative weakly coupled elliptic systems and some applications, Differential Integral Equations 7 (1994) 383–398.

[28] J. Lorenz, Zur Inversmonotonie diskreter Probleme, Numer. Math. 27 (1977) 227–238.

[29] J. Nečas, Introduction to the Theory of Nonlinear Elliptic Equations, Teubner, Leipzig, 1983.

[30] M.H. Protter, H.F. Weinberger, Maximum Principles in Differential Equations, Springer-Verlag, New York, 1984.

[31] V. Ruas Santos, 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.

[32] S.A. Sander, The maximum principle for elliptic problems in the finite element method, Russian Acad. Sci. Dokl. Math. 49 (1994) 386–390.

[33] A.H. Schatz, A weak discrete maximum principle and stability of the finite element method in L_∞ on plane polygonal domains. I., Math. Comp. 34 (1980) 77–91.

[34] V.V. Smelov, Algebraic aspect of the maximum principle, Russ. J. Numer. Anal. Math. Modelling 16 (2001) 175–190.

[35] P. Šolín, T. Vejchodský, On a weak discrete maximum principle for hp-FEM, Preprint 2006-01 (2006), University of Texas at El Paso (submitted).

[36] G. Strang, G.J. Fix, An Analysis of the Finite Element Method, Prentice Hall, Englewood Cliffs, N. J., 1973.

[37] A. Unterreiter, A. Juengel, Discrete minimum and maximum principles for finite element approximations of non-monotone elliptic equations, Numer. Math. 99 (2005) 485–508.

[38] R. Varga, Matrix Iterative Analysis, Prentice Hall, New Jersey, 1962.

[39] R. Varga, On discrete maximum principle, J. SIAM Numer. Anal. 3 (1966) 355–359.

[40] L.T. Wheeler, Maximum principles in classical elasticity. Mathematical problems in elasticity, in: Ser. Adv. Math. Appl. Sci., 38, 1996, pp. 157–185.

(continued from the back cover)

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

A497 Outi Elina Maasalo
Gehring Lemma in Metric Spaces
March 2006

A496 Jan Brandts , Sergey Korotov , Michal Krizek
Dissection of the path-simplex in $\mathbf{R}^n$ into n path-subsimplices
March 2006

A495 Sergey Korotov
A posteriori error estimation for linear elliptic problems with mixed boundary conditions
March 2006

A494 Antti Hannukainen , Sergey Korotov
Computational Technologies for Reliable Control of Global and Local Errors for Linear Elliptic Type Boundary Value Problems
February 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/* .

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

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

ISBN 951-22-8314-X
ISSN 0784-3143