Discrete maximum principles for finite element solutions of some mixed nonlinear elliptic problems using quadratures

J. Karátson\textsuperscript{a,*}, S. Korotov\textsuperscript{b,2}

\textsuperscript{a} Department of Applied Analysis, Eötvös Loránd University, Pf. 120, H-1518 Budapest, Hungary
\textsuperscript{b} Department of Mathematical Information Technology, University of Jyväskylä, P.O. Box 35, FIN-40014 Jyväskylä, Finland

Received 15 September 2004; received in revised form 15 February 2005

Abstract

The discrete maximum principles are proved for finite element solutions of some nonlinear elliptic problems with mixed boundary conditions. The effect of quadrature rules, used for the construction of the stiffness matrices, is taken into account.

© 2005 Elsevier B.V. All rights reserved.

MSC: 35B50; 35J65; 65N30; 65N50

Keywords: Nonlinear elliptic problem; Mixed boundary condition; Discrete maximum principle; Finite element method; Quadratures

1. Introduction

The maximum principle is a basic qualitative property of the solutions of elliptic boundary value problems (BVPs) [20,21], which is why the construction and validity of its discrete analogues, i.e., the discrete maximum principles (DMPs), have drawn much attention. Several DMPs had been formulated and proved in a number of papers, including the case of finite difference and finite element approximations,
and various convenient geometric conditions on the shape of (finite) elements providing DMPs have been proposed [4–6, 10–12, 22, 24]. In the above papers only linear problems with Dirichlet boundary conditions were studied, whereas nonlinear elliptic problems have been considered in [16] for Dirichlet and recently by the authors in [14, 13] for mixed boundary conditions.

In all the aforementioned papers that concern the finite element method (FEM), the stiffness matrices have been considered with exact evaluation except for [16] where the effect of quadrature rules on tetrahedral elements has also been taken into account. Our goal is to extend the results of [16] to mixed boundary conditions, i.e., to prove the DMP for mixed BVPs when quadrature rules are used to form the stiffness matrices. The considered class of mixed BVP falls into the type studied in [14, 13]; hence, the techniques used there will be relied on. As additional generalizations of the results of [16], we treat problems in any dimension, apply a more general case of basis functions (satisfying certain algebraic conditions as in [13]) instead of commonly used continuous piecewise linear elements, and introduce quadrature rules as linear functionals with suitable positivity properties.

More relevant material on various aspects of DMPs can be found in very recent works [3, 7, 23, 25]. For analysis of the comparison principle and its close relation to the DMP we refer to [17, 18].

After the formulation of the problem in Section 2 and some background in Section 3, we deal with the quadrature rules and the corresponding DMPs in Section 4.

2. The problem

2.1. Formulation of the continuous problem

We consider a nonlinear BVP of the following form

\[- \text{div}(b(x, \nabla u) \nabla u) = f(x) \quad \text{in } \Omega,\]

\[b(x, \nabla u) \frac{\partial u}{\partial n} = \gamma(x) \quad \text{on } \Gamma_N,\]

\[u = g(x) \quad \text{on } \Gamma_D,\]

where \(\Omega\) is a bounded domain in \(\mathbb{R}^d\), under the following assumptions:

(A1) \(\Omega\) has a piecewise smooth and Lipschitz continuous boundary \(\partial \Omega\); \(\Gamma_N, \Gamma_D \subset \partial \Omega\) are measurable open sets, such that \(\Gamma_N \cap \Gamma_D = \emptyset\) and \(\overline{\Gamma_N} \cup \overline{\Gamma_D} = \partial \Omega\).

(A2) \(b \in C^1(\overline{\Omega} \times \mathbb{R}^d), \ f \in L^2(\Omega), \ \gamma \in L^2(\Gamma_N)\) and \(g \in H^1(\Omega)\).

(A3) The function \(b\) satisfies

\[0 < \mu_0 \leq b(x, \eta) \leq \mu_1\]

with positive constants \(\mu_0\) and \(\mu_1\) independent of \((x, \eta)\); further, the diadic product matrix \(\eta \cdot \frac{\partial b(x, \eta)}{\partial \eta}\) is symmetric positive semidefinite and bounded in any matrix norm by some positive constant \(\mu_2\), independent of \((x, \eta)\).

(A4) \(\Gamma_D \neq \emptyset\).
**Remark 2.1.** Assumption (A3) ensures that the Jacobian matrices \( \frac{\partial}{\partial \eta} (b(x, \eta) \eta) \) are symmetric and satisfy the uniform ellipticity property

\[
\mu_0 |\zeta|^2 \leq \frac{\partial}{\partial \eta} (b(x, \eta) \eta) \cdot \zeta \leq \mu_3 |\zeta|^2, \quad \zeta \in \mathbb{R}^d
\]

with \( \mu_3 = \mu_1 + \mu_2 \). For instance, assumption (A3) holds for coefficients of the form

\[ b(x, \eta) = a(x, |\eta|), \]

where the \( C^1 \) function \( a : \Omega \times \mathbb{R}^+ \to \mathbb{R} \) satisfies

\[ 0 < \mu_0 \leq a(x, r) \leq \frac{\partial}{\partial r} (a(x, r) r) \leq \mu_3 \quad (r > 0). \]

Such nonlinearities arise in various applications (see e.g., [1,9,19]).

(We incidentally note that linear mixed BVPs are also included in (1) if \( b(x, \eta) = a(x) \) with \( 0 < \mu_0 \leq a(x) \leq \mu_3 \).)

In order to formulate the weak solution, we introduce the Sobolev space corresponding to the homogeneous Dirichlet boundary condition:

\[ H^1_D(\Omega) := \{ u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_D \} \]

with inner product and corresponding norm

\[
\langle u, v \rangle_{H^1_D} = \int_{\Omega} \nabla u \cdot \nabla v, \quad \| v \|_{H^1_D}^2 = \int_{\Omega} |\nabla v|^2.
\]

Assumptions (A1)–(A4) lead to a setting with a convex potential (see e.g., [9]). Accordingly, we obtain the following well-posedness result:

**Theorem 2.1 (Karátson and Korotov [14]).** Problem (1) has a unique weak solution \( u^* \in H^1(\Omega) \), defined as follows:

\[
u^* = g \quad \text{on } \Gamma_D \text{ in trace sense, and}
\]

\[
\int_{\Omega} b(x, \nabla u^*) \cdot \nabla v \, dx = \int_{\Omega} f v \, dx + \int_{\Gamma_N} \gamma v \, d\sigma \quad \forall v \in H^1_D(\Omega).
\]

**Remark 2.2.** The restriction of \( b \) to depend only on \( x \) and \( \nabla u \) serves to ensure that the above well-posedness results in a potential framework, and to have \( \| u^* - u_h \|_1 \to 0 \) for the FEM solutions in the next subsection. However, the results of this paper on the DMP also hold if \( b \) is also allowed to depend on \( u \); see Remark 4.2 later.

### 2.2. Finite element discretization

In what follows, we additionally assume that \( \Omega \) is a polytopic domain. We define the finite element discretization of our problem using some continuous piecewise polynomial basis functions
\( \phi_i \in H^1(\Omega) \) \((i = 1, \ldots, \tilde{n})\) and let \( V_h \) denote the finite element subspace spanned by the above basis functions:

\[
V_h = \text{span}\{\phi_1, \ldots, \phi_{\tilde{n}}\} \subset H^1(\Omega).
\]

Our main assumptions are as follows:

\[
\phi_i \geq 0 \quad (i = 1, \ldots, \tilde{n}),
\]

\[
\sum_{j=1}^{\tilde{n}} \phi_j = 1,
\]

\[
\lim_{h \to 0} \text{dist}(u, V_h) = 0 \quad (u \in H^1(\Omega),)
\]

where \( \text{dist}(u, V_h) = \inf_{v_h \in V_h} \|u - v_h\|_1 \).

Let us denote by \( \phi_1, \ldots, \phi_n \) the basis functions that satisfy the homogeneous Dirichlet boundary condition on \( \Gamma_D \), i.e., \( \phi_i \in H^1_D(\Omega) \), and by \( \phi_{n+1}, \ldots, \phi_{\tilde{n}} \) the basis functions that also have nonzero values on \( \Gamma_D \). We define

\[
V_0^h = \text{span}\{\phi_1, \ldots, \phi_n\} \subset H^1_D(\Omega).
\]

Further, let

\[
g_h = \sum_{j=n+1}^{\tilde{n}} g_j \phi_j \in V_h
\]

(with \( g_j \in \mathbb{R} \)) be the approximation of the function \( g \) on \( \Gamma_D \) (and on the neighbouring elements).

To find the approximate solution, we solve the following problem (which is the counterpart of (5)–(6) in \( V_h \)): find \( u_h \in V_h \) such that

\[
u_h = g_h \quad \text{on} \quad \Gamma_D \quad \text{and}
\]

\[
\int_\Omega b(x, \nabla u_h) \cdot \nabla v_h \, dx = \int_\Omega f v_h \, dx + \int_{\Gamma_N} \gamma v_h \, d\sigma \quad \forall v_h \in V_0^h.
\]

If (A1)–(A4) hold then problem (12) has a unique solution \( u_h \in V_h \), which follows similarly as for the weak solution of (1). Further, using (9), it follows similarly as in [14] that \( \|u^* - u_h\|_1 \to 0 \) as \( h \to 0 \).

Now we turn to the nonlinear algebraic system corresponding to (12). We set

\[
u_h = \sum_{j=1}^{\tilde{n}} c_j \phi_j
\]

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

\[
\begin{align*}
    a_{ij}(\overline{c}) &= \int_\Omega b \left( x, \sum_{k=1}^{\tilde{n}} c_k \nabla \phi_k \right) \nabla \phi_j \cdot \nabla \phi_i \, dx, \\
    d_i(\overline{c}) &= \int_\Omega f \phi_i \, dx + \int_{\Gamma_N} \gamma \phi_i \, d\sigma.
\end{align*}
\]
Inserting (13) and \(v_h = \phi_i\) into (12), we obtain the \(n \times \bar{n}\) system of algebraic equations

\[
\sum_{j=1}^{\bar{n}} a_{ij}(\bar{c}) \cdot c_j = d_i, \quad i = 1, \ldots, n.
\]

(15)

Using the notations

\[ A(\bar{c}) = \{a_{ij}(\bar{c})\}, \quad i, j = 1, \ldots, n, \quad \text{and} \quad \tilde{A}(\bar{c}) = \{a_{ij}(\bar{c})\}, \quad i = 1, \ldots, n; \quad j = n + 1, \ldots, \bar{n}, \]

\[ d = \{d_j\}, \quad c = \{c_j\}, \quad j = 1, \ldots, n, \quad \text{and} \quad \tilde{c} = \{c_j\}, \quad j = n + 1, \ldots, \bar{n}, \]

system (15) turns into

\[ A(\bar{c})c + \tilde{A}(\bar{c})\tilde{c} = d. \]

(17)

Defining further

\[ \tilde{A}(\bar{c}) = [A(\bar{c}) \quad \tilde{A}(\bar{c})], \quad \tilde{c} = \begin{bmatrix} c \\ \tilde{c} \end{bmatrix}, \]

we rewrite (17) as follows:

\[ \tilde{A}(\bar{c})\tilde{c} = d. \]

(19)

In order to obtain a system with a square matrix, we enlarge our system to an \(\bar{n} \times \bar{n}\) one. Namely, since \(u_h = g_h\) on \(\Gamma_D\), the coordinates \(c_i\) with \(n + 1 \leq i \leq \bar{n}\) satisfy automatically \(c_i = g_i\), i.e.,

\[ \tilde{c} = \tilde{g}, \]

where

\[ \tilde{g} = \{g_j\}, \quad j = n + 1, \ldots, \bar{n}. \]

That is, we can replace (17) by the equivalent system

\[
\begin{bmatrix}
A(\bar{c}) & \tilde{A}(\bar{c}) \\
0 & I
\end{bmatrix}
\begin{bmatrix}
c \\ \tilde{c}
\end{bmatrix}
= 
\begin{bmatrix}
d \\ \tilde{g}
\end{bmatrix}.
\]

(20)

3. Preliminaries on maximum principles for nonlinear elliptic problems

We summarize briefly some background on maximum principles, mainly including our previous results for problem (1) presented in [14].

3.1. Continuous maximum principles

The following continuous maximum principle has been verified in [14].

**Theorem 3.1.** Let problem (1) satisfy assumptions (A1)–(A4) and let the weak solution \(u\) belong to \(C^1(\Omega) \cap C(\overline{\Omega})\). If

\[ f(x) \leq 0, \quad x \in \Omega, \quad \text{and} \quad \gamma(x) \leq 0, \quad x \in \Gamma_N, \]

(21)
then

\[
\max_{\Omega} u = \max_{\Gamma_D} g.
\]  

(22)

Note that the corresponding minimum principle obviously holds if the sign conditions in (21) are reversed; further, if \( f \) and \( \gamma \) are constantly zero then both the maximum and minimum principles are valid (see also [14]). That is, we have

**Corollary 3.1.** Let assumptions (A1)–(A4) hold and let the weak solution \( u \) of problem (1) belong to \( C^1(\Omega) \cap C(\bar{\Omega}) \).

1. If \( f \geq 0 \) and \( \gamma \geq 0 \), then \( \min_{\Omega} u = \min_{\Gamma_D} g \).
2. If \( f = 0 \) and \( \gamma = 0 \), then the ranges of \( u \) and \( g \) coincide, i.e., we have \([\min_{\Omega} u, \max_{\Omega} u] = [\min_{\Gamma_D} g, \max_{\Gamma_D} g]\) for the corresponding intervals.

Further, the analogues of the above results hold in the same way for the case \( u \in H^1(\Omega) \), i.e., with no regularity assumption on the weak solution, provided that \( g \) is bounded on \( \Gamma_D \). Then the maxima and minima are replaced by ess sup and ess inf, respectively.

### 3.2. Discrete maximum principles

First we quote a main theorem, on which various known results about DMPs are based (e.g., [5,6,16]). Let us consider a system of equations of order \( (n + m) \times (n + m) \):

\[
\bar{A}\bar{c} = \bar{d},
\]

(23)

where the matrix \( \bar{A} \) has the following structure:

\[
\bar{A} = \begin{bmatrix} A & \tilde{A} \\ 0 & I \end{bmatrix}.
\]

(24)

In the above, \( I \) is an \( m \times m \) identity matrix, \( 0 \) is an \( m \times n \) zero matrix. (The structure (24) is as in (20) and (35), where the unknown vector \( \bar{c} = (c_1, \ldots, c_{n+m}) \) contains all coefficients of the finite element solution \( u_h \).) Let us first recall the definition of irreducibly diagonally dominant matrices (cf. [26, p. 23]).

**Definition 3.1.** A square \( n \times n \) matrix \( M = (m_{ij})_{i,j=1}^n \) is called \textit{irreducibly diagonally dominant} if it satisfies the following conditions:

1. \( M \) is irreducible, i.e., for any \( i \neq j \) there exists a sequence of nonzero entries \( \{m_{i_1,i}, m_{i_1,i_2}, \ldots, m_{i_s,j}\} \) of \( M \), where \( i, i_1, i_2, \ldots, i_s, j \) are distinct indices,
2. \( M \) is diagonally dominant, i.e., \( |m_{ii}| \geq \sum_{j \neq i} |m_{ij}|, \quad i = 1, \ldots, n \),
3. for at least one index \( i_0 \in \{1, \ldots, n\} \) the above inequality is strict, i.e.,

\[
|m_{i_0,i_0}| > \sum_{\substack{j=1 \atop j \neq i_0}}^n |m_{i_0,j}|.
\]
**Theorem 3.2** (Ciarlet [5]). Let $\tilde{A}$ be an $(n + m) \times (n + m)$ matrix with the structure as in (24). Assume that

1. $a_{ii} > 0$, $i = 1, \ldots, n$,
2. $a_{ij} \leq 0$, $i = 1, \ldots, n$, $j = 1, \ldots, n + m$ $(i \neq j)$,
3. $\sum_{j=1}^{n+m} a_{ij} = 0$, $i = 1, \ldots, n$,
4. $A$ is irreducibly diagonally dominant.

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

$$\max_{i=1,\ldots,n+m} c_i = \max_{i=n+1,\ldots,n+m} c_i.$$  \hspace{1cm} (25)

Based on the above theorem, in [14] we have proved the discrete analogue of Theorem 3.1 for finite element discretization using an exact evaluation of the stiffness matrices, under the same condition (26) on the basis functions as used for homogeneous Dirichlet problems in [6,16]. Namely:

**Theorem 3.3.** Let problem (1) satisfy conditions (A1)–(A4), and let the subspace $V_h$ satisfy the following property:

for any $i = 1, \ldots, n$, $j = 1, \ldots, \tilde{n}$ $(i \neq j)$

$$\nabla \phi_i \cdot \nabla \phi_j \leq 0.$$  \hspace{1cm} (26)

Then the matrix $\tilde{\mathbf{A}}(\tilde{c})$ defined in (18) has the properties (i)–(iv) of Theorem 3.2. Accordingly, if the sign conditions (21) hold, then

$$\max_{N} u_h = \max_{I_D} g_h.$$  \hspace{1cm} (27)

Similarly (see also [14]), if the sign conditions (21) are replaced by those in Corollary 3.1, then the DMP (27) is replaced by the analogous minimum or maximum–minimum principles thereby.

4. Discrete maximum principles using quadratures

We deal with the quadrature rules in Section 4.1 and the corresponding DMP in Section 4.2. Finally, for the most important special cases, our results are summarized in Section 4.3.

4.1. Definitions of quadratures

4.1.1. Quadratures as weighted approximate sums

In practice the integrals in (14) are computed numerically using some quadratures. Thus, let $\mathcal{T}_h$ be a triangulation of a polytopic domain $\Omega$ into simplices. We approximate an integral $\int_{\Omega} g$ by

$$Q_1(g) := \sum_{T \in \mathcal{T}_h} \text{meas}_d(T) \sum_{k=1}^{K} \omega_{T,k} g(x_{T,k}),$$  \hspace{1cm} (28)
where for each given simplex $T \in \mathcal{T}_h$ one chooses nodes $x_{T,k} \in T$ and weights $\omega_{T,k} \in \mathbb{R}$ ($k = 1, \ldots, K$) such that for all $k$

$$\omega_{T,k} > 0, \quad \sum_{k=1}^{K} \omega_{T,k} = 1$$

(this type of quadrature has been used in [16] for the case of tetrahedral elements, i.e., $d = 3$).

The above quadratures can be analogously extended for the required (hyper)surface integrals. For convenience, we assume that the triangulation $\mathcal{T}_h$ is constructed so that $\Omega$ consists of a union of entire faces of simplices. Let $\mathcal{S}_h$ denote the collection of all such faces on $\Omega$. Then we approximate an integral $\int_{\Omega} \varphi$ by

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

(29)

where for each face $S \in \mathcal{S}_h$ one chooses nodes $x_{S,i} \in S$ and weights $\sigma_{S,i} \in \mathbb{R}$ ($i = 1, \ldots, I$) such that for all $i$

$$\sigma_{S,i} > 0, \quad \sum_{i=1}^{I} \sigma_{S,i} = 1.$$

4.1.2. General properties of quadratures

In general, we can consider quadratures defined as functionals

$$Q_1 : PC(\Omega) \to \mathbb{R} \quad \text{and} \quad Q_2 : PC(\Gamma_N) \to \mathbb{R}$$

(30)

(where $PC(S)$ denotes piecewise continuous functions on a set $S$). We define the following properties for $i = 1, 2$:

(Q1) $Q_i$ is linear.
(Q2) $Q_i$ is monotone, i.e., if $f \geq g$ then $Q_i(f) \geq Q_i(g)$.
(Q3) $Q_1$ is strictly positive on the subspace $V^0_h$, defined in (10), in the sense that for any $v_h \in V^0_h$,

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

Remark 4.1 (On condition Q3). Exact integration satisfies the following well-known strict positivity property: if $f \geq 0$ is a Lebesgue integrable function and $\int_{\Omega} f = 0$ then $f \equiv 0$ a.e. (almost everywhere) on $\Omega$. In particular, if $v \in H^1_0(\Omega)$ then $\int_{\Omega} |\nabla v|^2 = 0$ implies $v \equiv const$ and from $v|_{\Gamma_D} = 0$ we have $v \equiv 0$ a.e., that is, the analogue of (Q3) holds. Clearly, one cannot require the previous strict positivity for all integrable $f$, since quadratures like (28) and (29) are zero for any function with support outside their nodes. Property (Q3) is a natural requirement since it ensures (together with Q1) that the trace of the Sobolev norm (4) in $V_h$ under $Q_1$, i.e.,

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

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


Proposition 4.1. The quadratures (28) and (29) satisfy the properties (Q1)–(Q2). Further, all those quadratures (28) that are exact for polynomials of degree \(2s-2\), where \(s\) is the (maximal) degree of the piecewise polynomials in \(V^0_h\), satisfy property (Q3).

Proof. (Q1) and (Q2) are obvious. Further, if (28) is exact for polynomials of degree \(2s-2\), then \(Q_1(|\nabla v_h|^2) = \int_{\Omega} |\nabla v_h|^2\) because \(|\nabla v_h|^2\) has degree at most \(2s-2\). This being zero, as pointed out in Remark 4.1, implies \(v_h \equiv 0\). □

For instance, any quadrature of type (28) is exact for piecewise constants and hence Proposition 4.1 is valid for piecewise linear finite element subspaces; see also later in Section 4.3. Other examples of quadratures which are exact for higher degree piecewise polynomials can be easily built using quadratures e.g., from [8].

4.1.3. Formulation of the discretized system using quadratures

Using the quadratures \(Q_1\) and \(Q_2\) introduced in (30), the integrals in (14) are replaced by

\[
\hat{a}_{ij}(\bar{c}) = Q_1 \left( b \left( x, \sum_{k=1}^{n} c_k \nabla \phi_k \right) \nabla \phi_j \cdot \nabla \phi_i \right), \quad \hat{d}_i(\bar{c}) = Q_1(f \phi_i) + Q_2(g \phi_i),
\]

and system (15) is replaced by

\[
\sum_{j=1}^{n} \hat{a}_{ij}(\bar{c}) c_j = \hat{d}_i, \quad i = 1, \ldots, n. \tag{32}
\]

Now using the notations

\[
Q(\bar{c}) = \{\hat{a}_{ij}(\bar{c})\}, \quad i, j = 1, \ldots, n, \quad \bar{Q}(\bar{c}) = \{\hat{d}_i(\bar{c})\}, \quad i = 1, \ldots, n; \quad j = n+1, \ldots, \tilde{n},
\]

\[
\hat{d} = \{\hat{d}_i(\bar{c})\}, \quad i = 1, \ldots, n \quad \text{and} \quad \bar{Q}(\bar{c}) = [Q(\bar{c}) \quad \bar{Q}(\bar{c})],
\]

system (32) becomes

\[
Q(\bar{c}) c + \bar{Q}(\bar{c}) \bar{c} = \hat{d} \quad \text{or} \quad \bar{Q}(\bar{c}) \bar{c} = \hat{d}. \tag{34}
\]

Similar to (20), system (34) is equivalent to the enlarged system

\[
\begin{bmatrix}
Q(\bar{c}) & \bar{Q}(\bar{c})
\end{bmatrix}
\begin{bmatrix}
c \\
\bar{c}
\end{bmatrix}
=
\begin{bmatrix}
\hat{d} \\
\bar{g}
\end{bmatrix}. \tag{35}
\]

Then, our ultimate numerical solution is defined as

\[
u_h = \sum_{j=1}^{\tilde{n}} c_j \phi_j. \tag{36}
\]
4.2. General conditions for the validity of the discrete maximum principles for nonlinear elliptic problems

Our main results state that the DMP in Theorem 3.3 remains valid (together with its analogues for minima) under the same condition (26) if we apply any quadrature with the general properties (Q1)–(Q3) given in Section 4.1.2. We formulate this in two steps.

**Theorem 4.1.** Let problem (1) satisfy conditions (A1)–(A4), and let the subspace $V_h$ satisfy the following property:

For any $i = 1, \ldots, n$, $j = 1, \ldots, \bar{n}$ ($i \neq j$)

$$\nabla \phi_i \cdot \nabla \phi_j \leq 0.$$  \hspace{1cm} (37)

Further, let us apply some quadrature $Q_1$ on $\Omega$ from (30) satisfying the properties (Q1)–(Q3). Then the matrix $\hat{Q}(\bar{c})$ defined in (33) has the following properties:

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

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

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

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

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

**Proof.** Let us recall (31) and (33):

$$\bar{Q}(\bar{c}) = \{\hat{a}_{ij}(\bar{c})\}, \ i = 1, \ldots, n, \ j = 1, \ldots, \bar{n}, \quad Q(\bar{c}) = \{\hat{a}_{ij}(\bar{c})\}, \ i, j = 1, \ldots, n,$$

where

$$\hat{a}_{ij}(\bar{c}) = Q_1(b(x, \nabla u_h) \nabla \phi_i \cdot \nabla \phi_j).$$

We verify conditions (i)–(iv).

(i) From our assumption $b \geq \mu_0 > 0$ \hspace{1cm} (38)

in (2) and from properties (Q2)–(Q3) of the quadrature, we have

$$\hat{a}_{ii}(\bar{c}) \geq \mu_0 Q_1(\|\nabla \phi_i\|^2) = \mu_0 \|\phi_i\|^2_{Q_1} > 0.$$  \hspace{1cm} (39)

(ii) Let $i = 1, \ldots, n$, $j = 1, \ldots, \bar{n}$ with $i \neq j$. Then properties (37) and (38) imply

$$b(x, \nabla u_h) \nabla \phi_i \cdot \nabla \phi_j \leq 0,$$

hence by property (Q2) of the quadrature

$$\hat{a}_{ij}(\bar{c}) \leq 0.$$  \hspace{1cm} (39)
(iii) For any \(i = 1, \ldots, n\),
\[
\sum_{j=1}^{\tilde{n}} \hat{a}_{ij}(\bar{c}) = Q_1 \left( b(x, \nabla u_h) \nabla \phi_i \cdot \nabla \left( \sum_{j=1}^{\tilde{n}} \phi_j \right) \right) = 0,
\]
using (8) and property (Q1).

(iv) We first verify that \(Q(\bar{c})\) is positive definite. Let \(p = (p_1, \ldots, p_n) \in \mathbb{R}^n\) and \(v_h = \sum_{i=1}^{n} p_i \phi_i\). Then
\[
Q(\bar{c})p \cdot p = \sum_{i,j=1}^{n} \hat{a}_{ij}(\bar{c}) p_i p_j = Q_1(b(x, \nabla u_h) |\nabla v_h|^2) \geq \mu_0 Q_1(|\nabla v_h|^2) = \mu_0 \|v_h\|^2_{Q_1} > 0
\]
unless \(p = 0\), using all three properties (Q1)–(Q3).

Assume now for contradiction that \(\sum_{j=1}^{\tilde{n}} \hat{a}_{ij}(\bar{c}) = 0\) for all \(i = 1, \ldots, n\). This means that \(Q(\bar{c})\) carries the \(n\)-tuple of ones \(\{1, \ldots, 1\}\) into the zero vector. This is impossible since \(Q(\bar{c})\) is positive definite and hence one to one.

(v) This follows in the same way as in [16]. Namely, the suitable intersections of the supports of the basis functions \(\phi_i\) define a usual triangulation of \(\Omega\) into subdomains (here (8) ensures that the union of these subdomains is indeed \(\Omega\)). For such triangulations the directed graph of the corresponding matrix \(Q(\bar{c})\) is strongly connected; hence, the matrix is irreducible. \(\square\)

Now it is easy to verify the analogues of Theorem 3.1 and Corollary 3.1 for system (35), i.e., the discrete maximum and minimum principles for finite elements with quadratures.

**Theorem 4.2.** Under the conditions of Theorem 4.1, let
\[
f(x) \leq 0, \quad x \in \Omega, \quad \text{and} \quad \gamma(x) \leq 0, \quad x \in \Gamma_N,
\]
and let the quadrature \(Q_2\) have properties (Q1)–(Q2). Then the numerical solution (36) satisfies
\[
\max_{B} u_h = \max_{\Gamma_D} g_h.
\]

**Proof.** We verify that the conditions of Theorem 3.2 are satisfied with \(\bar{Q}(c)\) and \(\tilde{n}\) substituted for \(\bar{A}\) and \(n + m\), respectively. Namely, conditions (i)–(iii) of Theorem 3.2 coincide with the statements (i)–(iii) of Theorem 4.1. Further, the three criteria of Definition 3.1 are also fulfilled, namely, \(Q(c)\) is irreducible due to statement (v), and the other two criteria follow from statements (ii)–(iv) under the signs obtained in statements (i)–(ii). Finally, using (7) and (41), property (Q2) of the quadratures and (31) imply that \(\hat{d}_i(\bar{c}) \leq 0\) for all \(i\), i.e., \(d \leq 0\). Hence, (34) yields \(Q(c)\bar{c} \leq 0\) and hence Theorem 3.2 yields
\[
\max_{i=1, \ldots, \tilde{n}} c_i = \max_{i=n+1, \ldots, \tilde{n}} c_i.
\]
Since \(c_i = g_i\) for all \(i = n + 1, \ldots, \tilde{n}\), we obtain
\[
\max_{i=1, \ldots, \tilde{n}} c_i = \max_{i=n+1, \ldots, \tilde{n}} g_i,
\]
which implies (42). \(\square\)
Corollary 4.1. Under the conditions of Theorem 4.1, the following results hold:

1. If \( f \geq 0 \) and \( \gamma \geq 0 \), then \( \min_{\Omega} u_h = \min_{\Gamma_D} g_h \).
2. If \( f = 0 \) and \( \gamma = 0 \), then the ranges of \( u_h \) and \( g_h \) coincide, i.e., we have \( [\min_{\Omega} u_h, \max_{\Omega} u_h] = [\min_{\Gamma_D} g_h, \max_{\Gamma_D} g_h] \) for the corresponding intervals.

Proof. Statement (1) follows from Theorem 4.2 by changing the signs, and statement (2) follows from Theorem 4.2 and statement (1) using the continuity of \( u_h \) and \( g_h \).

Remark 4.2. Theorems 4.1–4.2 are also valid for problems

\[- \text{div} (b(x, u, \nabla u) \nabla u) = f(x) \quad \text{in} \ \Omega,\]

\[ b(x, u, \nabla u) \frac{\partial u}{\partial v} = \gamma(x) \quad \text{on} \ \Gamma_N, \]

\[ u = g(x) \quad \text{on} \ \Gamma_D, \]  \( (45) \)

i.e., allowing \( b \) to depend on \( u \) as well, if condition (2) is preserved:

\[ 0 < \mu_0 \leq b(x, \xi, \eta) \leq \mu_1 \]  \( (46) \)

for all \((x, \xi, \eta)\). Namely, the proofs of Theorems 4.1–4.2 do not use what \( b \) depends on. In (1) the restriction of \( b \) to depend only on \( x \) and \( \nabla u \) served to ensure the general well-posedness result in a potential framework and to have \( \|u^* - u_h\|_1 \to 0 \) for the FEM solutions (which are not available for (45) on the continuous level). Problem (45) is an extension of the one considered in [6,16] with homogeneous Dirichlet boundary conditions.

4.3. An example: linear simplicial elements

In the case of linear finite elements, condition (37) allows a good geometric interpretation; further, one can use any quadratures in the form of weighted approximate sums (28)–(29). We summarize this briefly in what follows.

Let \( T_h \) be a triangulation of a polytopic domain \( \Omega \) into simplices. We define the finite element discretization of our problem using continuous piecewise linear basis functions \( \phi_i \in H^1(\Omega) \) \((i = 1, \ldots, \bar{n})\), i.e.,

\[ \phi_i |_T \in P_1(T) \quad (i = 1, \ldots, \bar{n}), \quad V_h = \text{span}\{\phi_1, \ldots, \phi_{\bar{n}}\} \subset H^1(\Omega). \]  \( (47) \)

Further, we let \( Q_1 \) and \( Q_2 \) be some quadratures of the form (28)–(29), i.e.,

\[ Q_1(g) := \sum_{T \in T_h} \text{meas}_d(T) \sum_{k=1}^{K} \omega_{T,k} g(x_{T,k}), \]

\[ Q_2(\varphi) := \sum_{S \in S_h} \text{meas}_{d-1}(S) \sum_{i=1}^{I} \sigma_{S,i} \varphi(x_{S,i}), \]  \( (48) \)
where for all simplices $T$, faces $S$, and indices $k$ and $i$, we have
\[
\omega_{T,k} > 0, \quad \sum_{k=1}^{K} \omega_{T,k} = 1; \quad \sigma_{S,i} > 0, \quad \sum_{i=1}^{I} \sigma_{S,i} = 1.
\]  
(49)

For piecewise linear functions (47), the gradients of the basis functions are constant on each element, which has two important advantages. First, we can show that (see [2] for the proof)
\[
\nabla \phi_i \cdot \nabla \phi_j = -\frac{\text{meas}_{d-1}(S_i) \cdot \text{meas}_{d-1}(S_j)}{d^2 \left( \text{meas}_d(T) \right)^2} \cos(S_i, S_j) \quad (i \neq j),
\]
(50)
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 (37) it is sufficient if the simplicial mesh employed is nonobtuse (cf. [6,16] for the two- and three-dimensional cases, respectively). Second, any quadrature $Q_1$ in the form of the weighted approximate sum (48) is obviously exact for piecewise constants: namely, if $g \equiv c_T$ on the element $T$, then (48)–(49) imply
\[
Q_1(g) = \sum_{T \in \mathcal{T}_h} \text{meas}_d(T) c_T \sum_{k=1}^{K} \omega_{T,k} = \sum_{T \in \mathcal{T}_h} \text{meas}_d(T) c_T = \sum_{T \in \mathcal{T}_h} \int_T c_T = \int_\Omega g,
\]
that is, Proposition 4.1 holds. Summing up, we can apply Theorems 4.1–4.2 and obtain the following.

**Corollary 4.2.** Let problem (1) satisfy conditions (A1)–(A4); let $V_h$ be the FEM subspace of continuous piecewise linear functions (47) on a simplicial triangulation $\mathcal{T}_h$ of a polytopic domain $\Omega$. If $\mathcal{T}_h$ consists of simplices with nonobtuse interior angles, then for any quadratures (48) satisfying property (49) the following results are valid:

1. If $f \leq 0$ and $\gamma \leq 0$, then $\max_{\Omega \setminus \Gamma} u_h = \max_{\Gamma_D} g_h$.
2. If $f \geq 0$ and $\gamma \geq 0$, then $\min_{\Omega \setminus \Gamma} u_h = \min_{\Gamma_D} g_h$.
3. If $f = 0$ and $\gamma = 0$, then the ranges of $u_h$ and $g_h$ coincide, i.e., we have $[\min_{\Omega \setminus \Gamma} u_h, \max_{\Omega \setminus \Gamma} u_h] = [\min_{\Gamma_D} g_h, \max_{\Gamma_D} g_h]$ for the corresponding intervals.

**Remark 4.3.** (i) We note that the nonobtuseness condition is sufficient but not necessary. Thus, the DMP may still hold in situations when certain obtuse interior angles occur [15,22].

(ii) Condition (37) allows a good geometric interpretation for not only simplicial but other meshes as well if linear finite elements are used. For instance, in the case of a rectangular mesh, condition (37) is satisfied if so-called nonnarrow rectangles are employed [4].

**References**