Proceedings of ALGORITMY 2012
pp. 1–10

NECESSARY AND SUFFICIENT CONDITION FOR THE VALIDITY OF THE DISCRETE MAXIMUM PRINCIPLE [*]

TOMÁŠ VEJCHODSKÝ[†]

Abstract. In this contribution we present a necessary and sufficient condition for the validity of the discrete maximum principle for one-dimensional elliptic diffusion-convection-reaction problems discretized by the finite element method. The condition limits the size of individual mesh elements. This is an extraordinary result in this field, because similar mesh conditions known so far are sufficient only.

Key words. discrete maximum principle, finite element method, elliptic problems, diffusion-convection-reaction problem

AMS subject classifications. 65N30, 35B50

1. Introduction.

Maximum principle is a fundamental property of many linear and nonlinear second-order elliptic differential operators. The maximum principle is used to prove uniqueness and various qualitative properties of solutions of corresponding partial differential equations. In addition, maximum principle reflects natural property of many real physical systems, namely the nonnegativity of naturally nonnegative quantities like concentration, temperature, density, pressure, etc.

Natural question is, whether discretizations of these partial differential problems posses the maximum principle property as well. We speak about the discrete maximum principle (DMP). It turns out that the usual numerical methods do not satisfy the DMP in general. However, under special conditions or using various modifications, the DMP can be guaranteed.

The first approaches studied the finite difference method [2, 3, 6, 7, 21]. Later, the first DMP results in the context of the FEM appeared, see [1, 8, 16]. Modifications of the standard finite element method such that the resulting approximate solutions obey the maximum principle, are presented e.g. in [5, 26]. These approaches typically lead to systems of nonlinear algebraic equations even for linear partial differential problems. Recently, the problem of the DMP for the finite element method attracted a lot of attention, see [4, 9, 11, 13, 14, 15, 17, 19] etc.

All of these results concern the lowest-order finite element method. The higher-order finite element methods are much more complicated and the DMP results are limited, see [12, 18, 22, 23, 24, 25]. Moreover, all these results provide only *sufficient* conditions on the element shapes and sizes for the validity of the DMP.

In this contribution we present the first mesh condition for the DMP that is both *sufficient and necessary.* We are able to find such a condition for the general nonsymmetric linear elliptic problem with general boundary conditions, which is again unusual in the field of the DMP. On the other hand this result is limited to one-dimensional problems only.

[*] This work was supported by Project No.: IAA100190803 of the Grant Agency of the Academy of Sciences and by RVO 67985840.

[†] Institute of Mathematics, Academy of Sciences, Žitná 25, Prague 1, CZ-115 67, Czech Republic (vejchod@math.cas.cz).

The rest of the paper is organized as follows. Section 2 introduces the one-dimensional diffusion-convection-reaction problem with mixed boundary conditions and its discretization by the finite element method. Section 3 defines the DMP and provides a theorem the subsequent analysis is based on. Section 4 presents needed results from the matrix theory and proves the fundamental statement about the monotony of tridiagonal matrices. The main result – the sufficient and necessary mesh condition for the DMP – is stated and proved in Section 5. The final Section 6 draws the conclusions.

2. The problem and its discretization. We concentrate on a general second-order linear elliptic one-dimensional diffusion-convection-reaction problem with mixed boundary conditions of Dirichlet and Newton (Robin) type:

$$-(\mathcal{A}u')' + bu' + cu = f \qquad \text{in } \Omega, \tag{2.1}$$

$$u = g_{\mathrm{D}} \qquad \text{on } \Gamma_{\mathrm{D}}, \tag{2.2}$$

$$\alpha u + \mathcal{A}u' n_{1\mathrm{D}} = g_{\mathrm{N}} \qquad \text{on } \Gamma_{\mathrm{N}}, \tag{2.3}$$

where the prime denotes the derivative with respect to $x \in \Omega$, the domain is an open interval $\Omega = (a^\partial, b^\partial)$, and Γ_{D}, Γ_{N} are empty, or one-point, or two-point subsets of $\partial\Omega = \{a^\partial, b^\partial\}$ such that $\Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}} = \{a^\partial, b^\partial\}$ and $\Gamma_{\mathrm{D}} \cap \Gamma_{\mathrm{N}} = \emptyset$. We use the special symbol $n_{1\mathrm{D}}$ to cover all four possible combinations of the subsets Γ_{D} and Γ_{N} by a single notation. The meaning of this symbol is the following

$$n_{1\mathrm{D}}(x) = \begin{cases} -1 & \text{for } x = a^\partial, \\ 1 & \text{for } x = b^\partial. \end{cases}$$

The derivatives of u at the end-points of Ω are understood as onesided.

In order to introduce the weak formulation, we define the space $V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_{\mathrm{D}}\}$ and the lift $\widetilde{g}_{\mathrm{D}}$ of the Dirichlet data g_{D}. This lift is an arbitrary function $\widetilde{g}_{\mathrm{D}} \in H^1(\Omega)$ satisfying $\widetilde{g}_{\mathrm{D}} = g_{\mathrm{D}}$ on Γ_{D}. The weak solution $u \in H^1(\Omega)$ of problem (2.1)–(2.3) is determined by the requirements $u - \widetilde{g}_{\mathrm{D}} \in V$ and

$$a(u, v) = \mathcal{F}(v) \quad \forall v \in V, \tag{2.4}$$

where the bilinear form $a(\cdot,\cdot)$ and the right-hand side functional $\mathcal{F}$ are

$$a(u, v) = \int_\Omega (\mathcal{A}u'v' + bu'v + cuv)\,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} \alpha uv\,\mathrm{d}s, \tag{2.5}$$

$$\mathcal{F}(v) = \int_\Omega fv\,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}} v\,\mathrm{d}s. \tag{2.6}$$

We recall that the integral over a finite point-set is defined as a sum. Hence, for example if $\Gamma_{\mathrm{N}} = \{a^\partial, b^\partial\}$ then

$$\int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}} v\,\mathrm{d}s = g_{\mathrm{N}}(a^\partial)v(a^\partial) + g_{\mathrm{N}}(b^\partial)v(b^\partial).$$

Integral over an empty set is understood as zero.

To ensure the correctness of the above weak setting and also the unique solvability of the weak formulation, we assume that $\mathcal{A}, c \in L^\infty(\Omega)$, $b \in W^{1,\infty}(\Omega)$, $f \in L^2(\Omega)$, and

$$\mathcal{A} \geq \lambda_{\min} > 0 \text{ in } \Omega, \quad c - \frac{1}{2}b' \geq 0 \text{ in } \Omega, \quad \alpha + \frac{1}{2}bn_{1\mathrm{D}} \geq 0 \text{ on } \Gamma_{\mathrm{N}}. \tag{2.7}$$

We also assume that at least one of the following conditions is satisfied: (a) Γ_{D} is nonempty, (b) there exists a constant c_0 and an open interval $B \subset \Omega$ such that $c - \frac{1}{2}b' \geq c_0 > 0$ a.e. in B, (c) the inequality $\alpha + \frac{1}{2}bn_{1\mathrm{D}} > 0$ holds for at least one point of Γ_{N}. In that case, the bilinear form a is V-elliptic, namely there exists a constant $C > 0$ such that $a(v,v) \geq C\|v\|_V^2$ for all $v \in V$.

To introduce the finite element solution of problem (2.4), we consider a partition $a^\partial = x_0 < x_1 < \cdots < x_{M-1} < x_M = b^\partial$ of the interval Ω and define the finite elements $K_k = [x_{k-1}, x_k]$, $k = 1, 2, \ldots, M$, with $h_k = x_k - x_{k-1}$. The finite element solution u_h lies in the space of continuous and piecewise linear functions $X_h = \{v_h \in H^1(\Omega) : v_h|_{K_i} \in \mathbb{P}^1(K_i),\ i = 1, 2, \ldots, M\}$, where $\mathbb{P}^1(K_i)$ stands for the space of linear functions in the interval K_i. The Dirichlet boundary conditions are represented by a subspace $V_h \subset X_h$ which contains functions vanishing on Γ_{D}, i.e. $V_h = X_h \cap V$. It is natural to define the approximate Dirichlet lift $\widetilde{g}_{\mathrm{D},h} \in X_h$ as a function which vanishes at all interior nodes x_i, $i = 1, 2, \ldots, M-1$, and on Γ_{N}, and which is equal to g_{D} on Γ_{D}. Thus, such a $\widetilde{g}_{\mathrm{D},h}$ belongs to the complement V_h^∂ of V_h in X_h (the space X_h is a direct sum of the linear spaces V_h and V_h^∂, i.e. $X_h = V_h \oplus V_h^\partial$).

The general finite element formulation reads: find $u_h \in X_h$ such that $u_h = u_h^0 + \widetilde{g}_{\mathrm{D},h}$ and $u_h^0 \in V_h$ satisfies

$$a(u_h^0, v_h) = \mathcal{F}(v_h) - a(g_{\mathrm{D},h}, v_h) \quad \forall v_h \in V_h, \tag{2.8}$$

where a and $\mathcal{F}$ are given by (2.5)–(2.6).

We introduce the standard finite element basis functions $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h and $\varphi_1^\partial, \ldots, \varphi_{N^\partial}^\partial$ of V_h^∂. Note that the only possible values for N^∂ in one dimension are 0, 1, and 2. In case $N^\partial = 0$ there are no Dirichlet boundary conditions and the space $V_h^\partial = \{0\}$ is trivial.

We express the finite element solution as

$$u_h = u_h^0 + \widetilde{g}_{\mathrm{D},h} = \sum_{j=1}^{N} y_j \varphi_j + \sum_{\ell=1}^{N^\partial} y_\ell^\partial \varphi_\ell^\partial. \tag{2.9}$$

The expansion coefficients y_i, $i = 1, 2, \ldots, N$ and y_ℓ^∂, $\ell = 1, \ldots, N^\partial$, are determined by the linear algebraic system

$$\overline{A}\overline{\boldsymbol{y}} = \overline{\boldsymbol{F}}, \quad \text{where} \quad \overline{A} = \begin{pmatrix} A & A^\partial \\ 0 & I \end{pmatrix}, \quad \overline{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{y}^\partial \end{pmatrix}, \quad \overline{\boldsymbol{F}} = \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{F}^\partial \end{pmatrix},$$

the entries of the vectors $\boldsymbol{y} \in \mathbb{R}^N$ and $\boldsymbol{y}^\partial \in \mathbb{R}^{N^\partial}$ are formed by the expansion coefficients y_i and y_ℓ^∂ and the entries of matrices $A \in \mathbb{R}^{N\times N}$, $A^\partial \in \mathbb{R}^{N\times N^\partial}$, and of vector $\boldsymbol{F} \in \mathbb{R}^N$ are given by $A_{ij} = a(\varphi_j, \varphi_i)$, $A_{i\ell}^\partial = a(\varphi_\ell^\partial, \varphi_i)$, $F_i = \mathcal{F}(\varphi_i)$, $i, j = 1, 2, \ldots, N$, $\ell = 1, \ldots, N^\partial$. The entries of the vector $\boldsymbol{F}^\partial \in \mathbb{R}^{N^\partial}$ are given by the values of $g_{\mathrm{D},h}$ at the corresponding end-points of Ω.

3. The discrete maximum principle. The standard maximum principle for the differential operator on the left-hand side of problem (2.1)–(2.3) is equivalent to the following conservation of nonnegativity:

$$f \geq 0,\ g_{\mathrm{D}} \geq 0,\ g_{\mathrm{N}} \geq 0 \quad \Rightarrow \quad u \geq 0.$$

This property can be naturally formulated in the discrete setting as

$$f \geq 0,\ g_{\mathrm{D}} \geq 0,\ g_{\mathrm{N}} \geq 0 \quad \Rightarrow \quad u_h \geq 0 \tag{3.1}$$

and we will call it the discrete maximum principle (DMP). To be more precise, if the finite element space $X_h = V_h \oplus V_h^\partial$ is fixed (consequently the mesh is fixed) then the discretization (2.8) satisfies the DMP if the implication (3.1) holds true for all right-hand side data $f \in L^2(\Omega)$, g_D, and g_N.

The following theorem is crucial for the subsequent analysis. Its variant for the finite difference method, can be found already in [6]. We note that the inequalities between matrices and vectors are understood entrywise, i.e. the statement $A^{-1} \geq 0$ means that all entries of the matrix A^{-1} are nonnegative.

Theorem 3.1. *Discretization (2.8) satisfies the DMP if and only if*

$$A^{-1} \geq 0 \quad \textit{and} \quad -A^{-1}A^\partial \geq 0. \tag{3.2}$$

Proof. First, let us assume that (3.2) holds true. If $f \geq 0$, $g_\mathrm{D} \geq 0$, and $g_\mathrm{N} \geq 0$ then clearly the entries of vectors $\boldsymbol{F}$ and $\boldsymbol{F}^\partial$ are nonnegative, because of the nonnegativity of the standard finite element basis functions. Now, we observe that

$$\overline{A}^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}A^\partial \\ 0 & I \end{pmatrix}.$$

Thus, $\overline{A}^{-1} \geq 0$ and the vector $\overline{\boldsymbol{y}} = \overline{A}^{-1}\overline{\boldsymbol{F}}$ has nonnegative entries only. From the expansion (2.9) we conclude that $u_h \geq 0$ and, hence, the DMP is satisfied.

To prove the converse implication, let us assume that the DMP is satisfied. Observe that the validity of conditions (3.2) is equivalent to the statement that the solution $\overline{\boldsymbol{y}}$ of the linear system $\overline{A}\overline{\boldsymbol{y}} = \overline{\boldsymbol{F}}$ is nonnegative for all vectors $\overline{\boldsymbol{F}} \geq 0$. Therefore, we consider arbitrary vectors $\boldsymbol{F} \geq 0$ and $\boldsymbol{F}^\partial \geq 0$ and construct suitable nonnegative right-hand side data f, g_D, and g_N corresponding to these vectors. Let us define the values of g_D to coincide with the corresponding values of $\boldsymbol{F}^\partial$. Trivially, $g_\mathrm{D} \geq 0$. Now, let us assume that there exists $f \in L^2(\Omega)$ such that

$$f \geq 0 \quad \text{and} \quad \int_\Omega f\varphi_i \,\mathrm{d}x = F_i, \quad i = 1, 2, \dots, N. \tag{3.3}$$

Taking $g_\mathrm{N} = 0$ we can use the DMP to infer that $u_h \geq 0$. Due to (2.9) and due to the standard properties of the finite element basis functions, we conclude that $\overline{\boldsymbol{y}} \geq 0$, which we wanted to prove.

However, if the function $f \in L^2(\Omega)$ with properties (3.3) does not exist, then we can consider an approximation $f^\varepsilon \in L^2(\Omega)$, $f^\varepsilon \geq 0$, such that $|F_i - F_i^\varepsilon| \leq \varepsilon$, where $F_i^\varepsilon = \int_\Omega f^\epsilon \varphi_i \,\mathrm{d}x$, $i = 1, 2, \dots, N$. As above, the solution $\overline{\boldsymbol{y}}^\varepsilon$ of the linear system $\overline{A}\overline{\boldsymbol{y}}^\varepsilon = \overline{\boldsymbol{F}}^\varepsilon$, where $\overline{\boldsymbol{F}}^\varepsilon = (\boldsymbol{F}^\varepsilon, \boldsymbol{F}^\partial)^\top$, is nonnegative due to the DMP. Since $\overline{\boldsymbol{y}}^\varepsilon$ converges to $\overline{\boldsymbol{y}}$ as $\varepsilon \to 0$, we conclude that $\overline{\boldsymbol{y}} \geq 0$. □

4. Selected results from the matrix theory. Theorem 3.1 shows that the analysis of the DMP is based on the nonnegative and monotone matrices. We recall that a real square matrix $A \in \mathbb{R}^{N\times N}$ is said to be *monotone* if it is nonsingular and $A^{-1} \geq 0$. Further, we introduce a special notation for the off-diagonal part of a matrix. The *off-diagonal part* of $A \in \mathbb{R}^{N\times N}$ is a matrix $B \in \mathbb{R}^{N\times N}$ with entries $B_{ii} = 0$ for $i = 1, 2, \dots, N$ and $B_{ij} = A_{ij}$ for $i \neq j$, $i, j = 1, 2, \dots, N$. We denote the off-diagonal part of A by off-diag(A). Finally, we say that a matrix $A \in \mathbb{R}^{N\times N}$ is positive definite if it satisfies $\boldsymbol{x}^\top A\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{x} \neq 0$.

For the DMP, the crucial class of matices are the M-matrices. A matrix $A \in \mathbb{R}^{N\times N}$ is said to be *M-matrix* if off-diag$(A) \leq 0$ and if it is nonsingular and $A^{-1} \geq 0$.

Clearly, M-matrices form a subclass of the monotone matrices. Their significance for the DMP stems from the following well-known theorem.

THEOREM 4.1. *Let a matrix $A \in \mathbb{R}^{N\times N}$ be positive definite and let* off-diag$(A) \leq 0$. *Then A is M-matrix, i.e. $A^{-1} \geq 0$.*

Proof. Using Lemma 4.2 below, it follows from [10, Thm. 5.1, p. 114]. □

Let us note that Theorem 4.1 is a generalization of the well-known result of Varga [20, p. 85] to nonsymmetric matrices.

In the special case of tridiagonal matrices, we can prove even the equivalence in Theorem 4.1. This equivalence is proved in Lemma 4.3 below, but first we introduce Lemma 4.2 which summarizes important facts about the nonsymmetric and positive definite matrices. Although these facts are quite well known and they (or their modifications) can be found for example in [10], we present their proof for the reader's convenience. Further, let us recall a few definitions. Formally, we say that a matrix $A \in \mathbb{R}^{N\times N}$ is *tridiagonal* if all its entries A_{ij} with $|i-j| \geq 2$ vanish. We also remind that having a nonempty subset of indices $M \subset \{1,2,\dots,N\}$ then a principal submatrix $A(M,M)$ of a square matrix $A \in \mathbb{R}^{N\times N}$ contains only entries A_{ij} with $i \in M$ and $j \in M$. The determinant of $A(M,M)$ is called the principal minor of A.

LEMMA 4.2. *Let a matrix $A \in \mathbb{R}^{N\times N}$ be positive definite. Then*

(a) *A is nonsingular,*
(b) *any real eigenvalue of A is positive,*
(c) $\det A > 0$*,*
(d) *all principal minors of A are positive,*
(e) *all principal minors of $A^{-\top}$ are positive.*

Proof. (a) If A was singular then there would exist a vector $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{x} \neq \boldsymbol{0}$ such that $A\boldsymbol{x} = \boldsymbol{0}$. Thus, $\boldsymbol{x}^\top A\boldsymbol{x} = 0$ contradicts the assumption of the lemma.
(b) Let us consider $\lambda \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{x} \neq \boldsymbol{0}$ such that $A\boldsymbol{x} = \lambda\boldsymbol{x}$. Then $0 < \boldsymbol{x}^\top A\boldsymbol{x} = \lambda\boldsymbol{x}^\top\boldsymbol{x}$. Since $\boldsymbol{x}^\top\boldsymbol{x} > 0$, we conclude that $\lambda > 0$.
(c) Let $\lambda_1, \lambda_2, \dots, \lambda_N$ be all eigenvalues of A, (some of them may coincide, depending on their multiplicity). If the eigenvalue λ_i, $i = 1,2,\dots,N$, is real, then $\lambda_i > 0$ by the property (b). The complex eigenvalues appear in pairs with their complex conjugate, i.e. if λ_i is complex then there exists $j \in \{1,2,\dots,N\}$ such that $\lambda_j = \overline{\lambda_i}$. Hence, $\lambda_i\lambda_j \geq 0$. Since $\det A = \lambda_1\lambda_2\dots\lambda_N$, we obtain $\det A \geq 0$ and by (a) we have $\det A > 0$.
(d) Let $\emptyset \neq M \subset \{1,2,\dots,N\}$, let $\#M$ be the number of elements of M, let $\boldsymbol{x}(M) \in \mathbb{R}^{\#M}$ be arbitrary nonzero vector, and let $\boldsymbol{x} \in \mathbb{R}^N$ be the vector $\boldsymbol{x}(M)$ augmented by zeros, i.e. its entries x_i, $i \in M$ coincide with entries of $\boldsymbol{x}(M)$ and its other entries are zero. Clearly, $\boldsymbol{x}$ is nonzero and $0 < \boldsymbol{x}^\top A\boldsymbol{x} = \boldsymbol{x}(M)^\top A(M,M)\boldsymbol{x}(M)$. Thus, the principal submatrix $A(M,M)$ has the same positive definiteness property as the matrix A and all statements (a)–(c) apply to $A(M,M)$ as well.
(e) Let $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{y} \neq \boldsymbol{0}$ be arbitrary. Then $\boldsymbol{y}^\top A^{-\top}\boldsymbol{y} = \boldsymbol{y}^\top A^{-\top}AA^{-1}\boldsymbol{y} = \boldsymbol{x}^\top A\boldsymbol{x} > 0$, where $\boldsymbol{x} = A^{-1}\boldsymbol{y} \neq \boldsymbol{0}$. Thus, we can use the statement (d) for $A^{-\top}$. □

LEMMA 4.3. *Let a matrix $A \in \mathbb{R}^{N\times N}$ be tridiagonal and positive definite. Then A is monotone if and only if* off-diag$(A) \leq 0$.

Proof. First, consider the case off-diag$(A) \leq 0$. By Theorem 4.1 the matrix A is M-matrix and hence monotone.

To prove the converse implication, we introduce the following notation for the entries of the tridiagonal matrix A

$$A = \begin{pmatrix} a_1 & b_1 & & 0 \\ c_1 & a_2 & \ddots & \\ & \ddots & \ddots & b_{N-1} \\ 0 & & c_{N-1} & a_N \end{pmatrix}.$$

The minor $C_{i,i+1}$ of the entry $A_{i,i+1} = b_i$ can be expressed as

$$C_{i,i+1} = \det \left(\begin{array}{c|c|c} & 0 & \\ L_{i-1} & \vdots & \\ & b_{i-1} & \\ \hline 0 \quad \dots \quad 0 & c_i & b_{i+1} \quad \dots \quad 0 \\ \hline & 0 & \\ & \vdots & R_{i+2} \\ & 0 & \end{array} \right),$$

where

$$L_{i-1} = \begin{pmatrix} a_1 & b_1 & & \\ c_1 & a_2 & \ddots & \\ & \ddots & \ddots & b_{i-2} \\ & & c_{i-2} & a_{i-1} \end{pmatrix}, \quad R_{i+2} = \begin{pmatrix} a_{i+2} & b_{i+2} & & \\ c_{i+2} & a_{i+3} & \ddots & \\ & \ddots & \ddots & b_{N-1} \\ & & c_{N-1} & a_N \end{pmatrix}.$$

Expanding the determinant $C_{i,i+1}$ with respect to its i-th row gives

$$C_{i,i+1} = c_i \det \begin{pmatrix} L_{i-1} & 0 \\ 0 & R_{i+2} \end{pmatrix} - b_{i+1} \det D,$$

where

$$D = \left(\begin{array}{c|c|c} & 0 & \\ L_{i-1} & \vdots & \\ & b_{i-1} & \\ \hline 0 \quad \dots \quad 0 & 0 & b_{i+2} \quad \dots \quad 0 \\ \hline & 0 & \\ & \vdots & R_{i+3} \\ & 0 & \end{array} \right).$$

The first i columns of D are linearly dependent, because they have nonzero entries in the first $i-1$ positions only. Therefore, $\det D = 0$.

Thus, if A is monotone then $A^{-1} \geq 0$, the entry $(A^{-1})_{i+1,i}$ of A^{-1} is nonnegative and we have

$$0 \leq (A^{-1})_{i+1,i} = -\frac{C_{i,i+1}}{\det A} = -\frac{c_i}{\det A} \det \begin{pmatrix} L_{i-1} & 0 \\ 0 & R_{i+2} \end{pmatrix}.$$

By Lemma 4.2 the determinants of A and of its principal submatrices L_{i-1} and R_{i+2} are positive and, thus, $c_i \leq 0$. Similar analysis of the minor $C_{i+1,i}$ of the entry $A_{i+1,i}$ shows that $b_i \leq 0$. □

5. Mesh conditions for the discrete maximum principle. In this section, we formulate the sufficient and necessary mesh condition for the validity of the DMP. We will use the notation introduced in Section 3. However, it is advantageous to denote the standard finite-element basis functions also in another way. The basis function corresponding to the node x_k of the mesh is denoted $\bar{\varphi}_k$, $k = 0, 1, \ldots, M$. Clearly, $\bar{\varphi}_k(x_j) = \delta_{kj}$, $k, j = 0, 1, \ldots, M$, where δ_{kj} stands for Kronecker's tensor. The sufficient and necessary mesh condition for the validity of the DMP is formulated in terms of the following constants on each element K_k, $k = 1, 2, \ldots, M$:

$$\mathcal{A}_k = \frac{1}{h_k} \int_{K_k} \mathcal{A}(x) \,\mathrm{d}x, \tag{5.1}$$

$$b_k^L = \frac{\int_{K_k} b(x)\bar{\varphi}_{k-1}(x)\,\mathrm{d}x}{\int_{K_k} \bar{\varphi}_{k-1}(x)\,\mathrm{d}x} = \frac{2}{h_k} \int_{K_k} b(x)\bar{\varphi}_{k-1}(x)\,\mathrm{d}x, \tag{5.2}$$

$$b_k^R = \frac{\int_{K_k} b(x)\bar{\varphi}_{k}(x)\,\mathrm{d}x}{\int_{K_k} \bar{\varphi}_{k}(x)\,\mathrm{d}x} = \frac{2}{h_k} \int_{K_k} b(x)\bar{\varphi}_{k}(x)\,\mathrm{d}x, \tag{5.3}$$

$$c_k = \frac{\int_{K_k} c(x)\bar{\varphi}_{k-1}(x)\bar{\varphi}_k(x)\,\mathrm{d}x}{\int_{K_k} \bar{\varphi}_{k-1}(x)\bar{\varphi}_k(x)\,\mathrm{d}x} = \frac{6}{h_k} \int_{K_k} c(x)\bar{\varphi}_{k-1}(x)\bar{\varphi}_k(x)\,\mathrm{d}x. \tag{5.4}$$

Notice that we utilized the facts that

$$\int_{K_k} \bar{\varphi}_{k-1}(x)\bar{\varphi}_k(x)\,\mathrm{d}x = \frac{h_k}{6} \quad \text{and} \quad \int_{K_k} \bar{\varphi}_{k-1}(x)\,\mathrm{d}x = \int_{K_k} \bar{\varphi}_k(x)\,\mathrm{d}x = \frac{h_k}{2}.$$

Notice also, that if the coefficients $\mathcal{A}$, b, and c are piecewise constant with respect to the considered partition then $\mathcal{A}_k$, $b_k^L = b_k^R$, and c_k equal to the constant values of the respective coefficients on the element K_k.

The constants (5.1)–(5.4) can be used to express the integrals needed for evaluation of the entries of the matrices off-diag(A) and A^∂:

$$\int_{K_k} \mathcal{A}(x)\bar{\varphi}'_{k-1}(x)\bar{\varphi}'_k(x)\,\mathrm{d}x = -\frac{\mathcal{A}_k}{h_k}, \qquad \int_{K_k} b(x)\bar{\varphi}'_{k-1}(x)\bar{\varphi}_k(x)\,\mathrm{d}x = -\frac{b_k^R}{2},$$
$$\int_{K_k} c(x)\bar{\varphi}_{k-1}(x)\bar{\varphi}_k(x)\,\mathrm{d}x = c_k\frac{h_k}{6}, \qquad \int_{K_k} b(x)\bar{\varphi}'_{k}(x)\bar{\varphi}_{k-1}(x)\,\mathrm{d}x = \frac{b_k^L}{2}.$$

Consequently,

$$a(\bar{\varphi}_k, \bar{\varphi}_{k-1}) = -\frac{\mathcal{A}_k}{h_k} + \frac{b_k^L}{2} + c_k\frac{h_k}{6}, \qquad a(\bar{\varphi}_{k-1}, \bar{\varphi}_k) = -\frac{\mathcal{A}_k}{h_k} - \frac{b_k^R}{2} + c_k\frac{h_k}{6}. \tag{5.5}$$

We clearly see that both $a(\bar{\varphi}_k, \bar{\varphi}_{k-1})$ and $a(\bar{\varphi}_{k-1}, \bar{\varphi}_k)$ are nonpositive if and only if

$$c_k h_k^2 + 3h_k \max\{b_k^L, -b_k^R\} \le 6\mathcal{A}_k.$$

This is the sufficient and necessary mesh condition for the validity of the DMP. The precise statement is formulated in the following theorem.

THEOREM 5.1. *Let the coefficients of problem (2.1)–(2.3) satisfy (2.7) and let the bilinear form (2.5) be V-elliptic. Then the lowest-order finite element discretization (2.8) satisfies the discrete maximum principle if and only if the condition*

$$c_k h_k^2 + 3h_k \max\{b_k^L, -b_k^R\} \le 6\mathcal{A}_k \tag{5.6}$$

holds for all $k = 1, 2, \dots, M$.

Proof. Let $\varphi_1, \varphi_2, \dots, \varphi_N$ be the finite element basis functions in V_h, see Section 2. Then the stiffness matrix $A \in \mathbb{R}^{N\times N}$ has entries $A_{ij} = a(\varphi_j, \varphi_i)$, $i, j = 1, 2, \dots, N$. Since the bilinear form (2.5) is V elliptic, the stiffness matrix is positive definite. In addition, the matrix A^∂ has the following form provided both end-points a^∂, b^∂ are on Γ_D

$$A^\partial = \begin{pmatrix} a(\bar\varphi_1, \bar\varphi_0) & 0 & \dots & 0 \\ 0 & \dots & 0 & a(\bar\varphi_M, \bar\varphi_{M-1}) \end{pmatrix}^\top \in \mathbb{R}^{N\times 2}. \tag{5.7}$$

If the end-point a^∂ or b^∂ (or both) is not on Γ_D then the corresponding row is missing in A^∂.

Hence, if condition (5.6) holds for all $k = 1, 2, \dots, M$ and if we recall that the off-diagonal entries of A are given by (5.5), then clearly off-diag$(A) \le 0$. Thus, $A^{-1} \ge 0$ by Theorem 4.1. Furthermore, condition (5.6) is satisfied also for elements adjacent to Γ_D (for $k = 1$ and/or $k = M$) and, therefore, $A^\partial \le 0$. Thus, Theorem 3.1 yields the DMP.

Now, we prove the converse implication. Assuming the validity of the DMP, Theorem 3.1 implies $A^{-1} \ge 0$ and $-A^{-1}A^\partial \ge 0$. Since the stiffness matrix A is tridiagonal and positive definite, we conclude by Lemma 4.3 that off-diag$(A) \le 0$. The nonpositivity of the off-diagonal entries of A yields the validity of the condition (5.6) at least for $k = 2, 3, \dots, M-1$. If $a^\partial \notin \Gamma_\mathrm{D}$ then $\bar\varphi_0$ is in V_h and condition (5.6) holds also for $k = 1$. Similarly, if $b^\partial \notin \Gamma_\mathrm{D}$ then (5.6) holds also for $k = M$.

However, if $a^\partial \in \Gamma_\mathrm{D}$ then $0 \le (-A^{-1}A^\partial)_{11} = -(A^{-1})_{11} a(\bar\varphi_0, \bar\varphi_1)$, where we use the special structure (5.7) of A^∂. Since $(A^{-1})_{11} > 0$ (see Lemma 4.2), we obtain $a(\bar\varphi_0, \bar\varphi_1) \le 0$ and consequently, the validity of the condition (5.6) for $k = 1$. Similarly, if $b^\partial \in \Gamma_\mathrm{D}$ we obtain (5.6) for $k = M$. □

Theorem 5.1 presents the complete characterization of the DMP for linear elliptic problems in one dimension discretized by the lowest-order finite element method. For given coefficients $\mathcal{A}$, b, and c, condition (5.6) determines the finite element meshes yielding the DMP. Let us point out that this condition is universal for any type of boundary conditions considered.

Practically, condition (5.6) enables to design sufficiently fine finite element meshes such that the DMP is satisfied. In addition, if the coefficients b and c are constant (or piecewise constant), then condition (5.6) is trivial to check. However, we have to admit, that condition (5.6) might be not practical to check in the case of general variable coefficients b and c. In this case we can recommend to use the following lemma.

Lemma 5.2. *Let us assume the hypothesis of Theorem 5.1. If*

$$\bar c_k = \operatorname*{ess\,sup}_{x\in K_k} c(x) \quad \textit{and} \quad \bar b_k = \operatorname*{ess\,sup}_{x\in K_k} |b(x)|, \quad k = 1, 2, \dots, M,$$

then the lowest-order finite element discretization (2.8) satisfies the discrete maximum principle provided the condition

$$\bar c_k h_k^2 + 3h_k \bar b_k \le 6\mathcal{A}_k \tag{5.8}$$

holds for all $k = 1, 2, \dots, M$.

Proof. The statement follows immediately from Theorem 5.1, because $c_k \le \bar c_k$ and $\max\{b_k^L, -b_k^R\} \le \bar b_k$ for all $k = 1, 2, \dots, M$. □

6. Conclusions. Theorem 5.1 states the main result of this paper. It is exceptional among the results about the DMP, because it provides an equivalent mesh condition for the DMP. The usual results about the DMP provide sufficient conditions only. In addition, condition (5.6) is very easy to verify, especially if the coefficients $\mathcal{A}$, b, and c are piecewise constant.

Theorem 5.1 enables to make several conclusions. For example, if the convection and reaction coefficients b and c vanish, then condition (5.6) is automatically satisfied and the DMP holds true on any mesh. If coefficients b or c are nonzero, then the mesh must be sufficiently fine in order to satisfy the DMP. The bigger coefficients b or c and the smaller $\mathcal{A}$ the finer mesh must be considered. Further interesting property of the condition (5.6) is its locality. If the values of b or c are high with respect to $\mathcal{A}$ in certain subdomain of Ω then the mesh must be correspondingly fine in this subdomain. On the other hand, if b and c are small with respect to $\mathcal{A}$ elsewhere, then the mesh can be coarse there.

Let us note that in case of vanishing coefficients b and c and piecewise constant coefficient $\mathcal{A}$, the finite element solution u_h coincides with the exact solution u at the nodal points x_k, $k = 0, 1, \dots, M$. In that case the DMP is satisfied on arbitrary meshes due to the validity of the maximum principle for the continuous problem (2.1)–(2.3) and the analysis of the DMP based on M-matrices can be omitted.

In case of variable coefficients b and c, it might be impractical to calculate b_k^L, b_k^R, and c_k by (5.2)–(5.4). Then we recommend to use Lemma 5.2. However, the condition (5.8) is no longer necessary.

The sufficient and necessary condition (5.6) completely characterizes the lowest-order finite element meshes yielding the DMP for the general one-dimensional linear elliptic problem (2.1)–(2.3) with arbitrarily mixed boundary conditions of Dirichlet and Newton (Robin) type. This is a special result for the one-dimensional setting. The crucial feature is the tridiagonality of the corresponding stiffness matrix and the usage of Lemma 4.3. Since the higher-dimensional problems do not yield tridiagonal stiffness matrices, this result cannot be easily generalized to higher dimension. So far, no sufficient and necessary mesh condition for the validity of the DMP for two (or higher) dimensional problems is known.

REFERENCES

[1] K. Baba, M. Tabata, *On a conservative upwind finite element scheme for convective diffusion equations*, RAIRO Anal. Numér., 15 (1981), pp. 3–25.

[2] J. H. Bramble and B. E. Hubbard, *New monotone type approximations for elliptic problems*, Math. Comp., 18 (1964), pp. 349–367.

[3] J. H. Bramble and B. E. Hubbard, *On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type*, J. Math. and Phys., 43 (1964), pp. 117–132.

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

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

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

[7] P. G. Ciarlet, *Discrete variational Green's function. I*, Aequationes Math., 4 (1970), pp. 74–82.

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

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

[10] M. Fiedler, *Special matrices and their applications in numerical mathematics*, Martinus Nijhoff Publishers, Dordrecht, 1986.

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

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

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

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

[15] P. Knobloch and L. Tobiska, *On the stability of finite-element discretizations of convection-diffusion-reaction equations*, IMA J. Numer. Anal., 31 (2011), pp. 147–164.

[16] K. Ohmori, *The discrete maximum principle for nonconforming finite element approximations to stationary convective diffusion equations*, Math. Rep. Toyama Univ., 2 (1979), pp. 33–52.

[17] 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), pp. 77–91.

[18] P. Šolín and T. Vejchodský, *A weak discrete maximum principle for hp-FEM*, J. Comput. Appl. Math., 209 (2007), pp. 54–65.

[19] R. Vanselow, *About Delaunay triangulations and discrete maximum principles for the linear conforming FEM applied to the Poisson equation*, Appl. Math., 46 (2001), pp. 13–28.

[20] R. S. Varga, *Matrix iterative analysis*, Prentice-Hall Inc., Englewood Cliffs, N.J., 1962.

[21] R. S. Varga, *On a discrete maximum principle*, SIAM J. Numer. Anal., 3 (1966), pp. 355–359.

[22] T. Vejchodský, *Higher-order discrete maximum principle for 1D diffusion-reaction problems*, Appl. Numer. Math., 60 (2010), pp. 486–500.

[23] T. Vejchodský and P. Šolín, *Discrete maximum principle for higher-order finite elements in 1D*, Math. Comp., 76 (2007), pp. 1833–1846.

[24] T. Vejchodský and P. Šolín, *Discrete maximum principle for a 1D problem with piecewise-constant coefficients solved by hp-FEM*, J. Numer. Math., 15 (2007), pp. 233–243.

[25] T. Vejchodský and P. Šolín, *Discrete maximum principle for Poisson equation with mixed boundary conditions solved by hp-FEM*, Adv. Appl. Math. Mech., 1 (2009), pp. 201–214.

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