The discrete maximum principle for Galerkin solutions of elliptic problems

Tomáš Vejchodský

Institute of Mathematics, Academy of Sciences of the Czech Republic,
Žitná 25, 115 67, Praha 1, Czech Republic, vejchod@math.cas.cz

Abstract: This paper provides equivalent characterization of the discrete maximum principle for Galerkin solutions of general linear elliptic problems. The characterization is formulated in terms of the discrete Green's function and the elliptic projection of the boundary data. This general concept is applied to the analysis of the discrete maximum principle for the higher-order finite elements in one-dimension and to the lowest-order finite elements on arbitrarily dimensional simplices. The paper surveys the state of the art in the field of the discrete maximum principles and provide new generalizations of several results.

Keywords: discrete maximum principle, monotone methods, Galerkin solution, finite element method

MSC: 65N30, 35B50

1 Introduction

The maximum principle is a typical qualitative property of (not only) linear elliptic boundary value problems. A natural question in numerical analysis of these problems is whether the approximate solution possesses the property of the maximum principle or not. This problem is referred to as the discrete maximum principle (DMP) or as the problem of monotonicity of the numerical method.

In mathematical modeling of various physical phenomena, the maximum principle reflects natural nonnegativity of quantities like temperature, concentration, density, etc. The validity of the maximum principle on the discrete level is therefore critical for the reliability of numerical models. Approximations with negative concentrations, or heat fluxes going from colder to warmer places, or financial fluxes in the opposite direction than expected are definitely not desirable and considered as unreliable.

In this paper we concentrate on general linear elliptic partial differential equations with diffusion, convection, and reaction terms in arbitrary dimension supplemented with general mixed boundary conditions of Dirichlet and Newton (Robin) type. We study the Galerkin method, because the most popular numerical scheme for elliptic problems – the finite element method – can be regarded as its special case. We analyze the DMP for Galerkin approximations and provide an equivalent characterization of its validity in terms of the discrete Green's function (DGF) and the elliptic projection of the Dirichlet boundary data.

The DMP has been studied and analyzed for many decades. The first results dealing with the finite difference method appeared in 1960s [1, 2, 40, 7, 8], etc. They were later generalized to the finite element method, see for example [10, 18, 33, 38, 12], etc. The proofs of the DMP are based on monotone matrices and mostly on the theory of M-matrices. The monograph [39] is fundamental and pioneering in this field. However, the more modern book [19] can be recommended as well. Today's literature on the subject of the DMP is vast. The above mentioned publications as well as this contribution handle elliptic problems, nevertheless the other major branch of results concerns parabolic problems [20, 16, 15, 17, 23] etc.

The current paper can be regarded as a generalization of the works [7] and [8]. Straightforward generalization of these results to the Galerkin method (including the finite element method) yields conditions which can be satisfied in the case of homogeneous Dirichlet boundary conditions or in 1D problems only. A novel treatment of the Dirichlet boundary data enabling higher-dimensional applications together with broad generality of the considered problem are the main contribution of the current paper. In addition, we made an effort to present the results in a selfcontained way and enable understanding also to nonspecialists. This contribution can be used as a survey paper presenting the current state of the art in the field of the DMP.

Section 2 introduces the general linear elliptic problem with mixed boundary conditions and defines and proves four equivalent variants of the maximum principle. Section 3 presents the Galerkin method and defines the DMP. In Section 4 we introduce the discrete Green's function and provide an equivalent characterization of the DMP. Section 5 briefly discusses the finite element method – especially the stiffness matrices. Section 6 shows an application of the general concept to higher-order finite elements in one dimension, while Section 7 concentrates on the lowest-order finite elements and prepares the necessary notions for Section 8, where an application to the lowest-order simplicial finite elements is presented. Finally, Section 9 summarizes the paper and draws the conclusions.

2 Formulation of the problem and the maximum principle

Let us consider a liner second-order elliptic problem of finding $u \in C^1(\overline{\Omega}) \cap C^2(\Omega)$ such that

$$\begin{aligned} -\operatorname{div}(\mathcal{A}\boldsymbol{\nabla} u) + \boldsymbol{b}\cdot\boldsymbol{\nabla} u + cu &= f \quad \text{in } \Omega, \\ u &= g_{\mathrm{D}} \quad \text{on } \Gamma_{\mathrm{D}}, \\ \alpha u + (\mathcal{A}\boldsymbol{\nabla} u)\cdot\boldsymbol{n} &= g_{\mathrm{N}} \quad \text{on } \Gamma_{\mathrm{N}}, \end{aligned} \tag{1}$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,\dots\}$, is a bounded domain with Lipschitz boundary, $\boldsymbol{n}$ is the unit outer normal to the boundary $\partial\Omega$, the sets Γ_{D} and Γ_{N} are relatively open in $\partial\Omega$, disjoint, and $\overline{\Gamma}_{\mathrm{D}} \cup \overline{\Gamma}_{\mathrm{N}} = \partial\Omega$. The sets Γ_{D} and Γ_{N} are assumed to have finite number of components and Lipschitz boundary relative to $\partial\Omega$. The coefficients $\mathcal{A}(\boldsymbol{x}) \in \mathbb{R}^{d\times d}$, $\boldsymbol{b}(\boldsymbol{x}) \in \mathbb{R}^d$, $c(\boldsymbol{x}) \in \mathbb{R}$, and the right-hand side $f(\boldsymbol{x}) \in \mathbb{R}$ are in general functions of $\boldsymbol{x} \in \Omega$, $g_{\mathrm{D}}(\boldsymbol{s})$ is a function of $\boldsymbol{s} \in \Gamma_{\mathrm{D}}$, and $\alpha(\boldsymbol{s}) \in \mathbb{R}$, $\mathcal{A}(\boldsymbol{s})$, $g_{\mathrm{N}}(\boldsymbol{s}) \in \mathbb{R}$ are functions of $\boldsymbol{s} \in \Gamma_{\mathrm{N}}$. Further we assume that

$$c - \frac{1}{2}\operatorname{div}\boldsymbol{b} \ge 0 \quad \text{in } \Omega \quad \text{and} \quad \alpha + \frac{1}{2}\boldsymbol{b}\cdot\boldsymbol{n} \ge 0 \quad \text{on } \Gamma_{\mathrm{N}} \tag{2}$$

and that the matrix $\mathcal{A}$ is uniformly positive definite, i.e. there exists $\lambda_{\min} > 0$ such that

$$(\mathcal{A}(\boldsymbol{x})\boldsymbol{\xi})\cdot\boldsymbol{\xi} \ge \lambda_{\min}|\boldsymbol{\xi}|^2, \quad \forall \boldsymbol{\xi} \in \mathbb{R}^d,\ \forall \boldsymbol{x} \in \Omega, \tag{3}$$

where $|\boldsymbol{\xi}| = (\boldsymbol{\xi}\cdot\boldsymbol{\xi})^{1/2}$ stands for the Euclidean norm of $\boldsymbol{\xi} \in \mathbb{R}^d$. Problem (1) and conditions (2) are well-posed in the classical sense under additional smoothness assumptions on the data and on the domain. However, we will not specify these assumptions here, since we will concentrate on the concept of weak solutions.

In order to introduce the weak formulation of problem (1), we assume $\mathcal{A} \in [L^\infty(\Omega)]^{d\times d}$, $\boldsymbol{b} \in [L^\infty(\Omega)]^d$, $\operatorname{div}\boldsymbol{b} \in L^\infty(\Omega)$, $c \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, $g_{\mathrm{D}} \in L^2(\Gamma_{\mathrm{D}})$, $g_{\mathrm{N}} \in L^2(\Gamma_{\mathrm{N}})$, and $\alpha \in L^\infty(\Gamma_{\mathrm{N}})$. Further, we consider the so called Dirichlet lift $\widetilde{g}_{\mathrm{D}}$ of g_{D}. It is an arbitrary but fixed function $\widetilde{g}_{\mathrm{D}} \in H^1(\Omega)$ such that $\widetilde{g}_{\mathrm{D}} = g_{\mathrm{D}}$ on $\overline{\Gamma}_{\mathrm{D}}$ in the sense of traces. Here and below we denote by $H^1(\Omega)$ the usual Sobolev space $W^{1,2}(\Omega)$. Further, we assume conditions

(2) to be satisfied a.e. in Ω and a.e. on Γ_{N}, respectively, and the uniform positive definiteness (3) for a.e. $\boldsymbol{x} \in \Omega$. Finally, let

$$V = \{v \in H^1(\Omega) : v = 0 \text{ on } \overline{\Gamma}_{\mathrm{D}} \text{ in the sense of traces}\}$$

and let the bilinear form a and the linear functional $\mathcal{F}$ be given by

$$\begin{aligned} a(u,v) &= \int_\Omega [(\mathcal{A}\boldsymbol{\nabla} u)\cdot\boldsymbol{\nabla} v + (\boldsymbol{b}\cdot\boldsymbol{\nabla} u)v + cuv]\,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}} \alpha uv\,\mathrm{d}\boldsymbol{s} \quad \text{and} \\ \mathcal{F}(v) &= \int_\Omega fv\,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}} v\,\mathrm{d}\boldsymbol{s}. \end{aligned} \tag{4}$$

We say that $u \in H^1(\Omega)$ is a weak solution of (1) if $u = u^0 + \widetilde{g}_{\mathrm{D}}$, where $u^0 \in V$ and

$$a(u^0, v) = \mathcal{F}(v) - a(\widetilde{g}_{\mathrm{D}}, v) \quad \forall v \in V. \tag{5}$$

It is easy to verify that the boundedness of the coefficients $\mathcal{A}$, $\boldsymbol{b}$, c and the trace theorem (see e.g. [30]) imply the continuity of the bilinear form a, i.e., the existence of a contant $C > 0$ such that

$$a(u,v) \le C\,\|u\|_{1,\Omega}\,\|v\|_{1,\Omega} \quad \forall u, v \in V. \tag{6}$$

The crucial condition for the existence of the weak solution and also for the validity of the maximum principle (see Theorem 2.3 below) is the V-ellipticity of the bilinear form $a(\cdot,\cdot)$. We say that the bilinear form $a(\cdot,\cdot)$ is V-elliptic if there exists a contant $C > 0$ such that

$$a(v,v) \ge C\,\|v\|^2_{1,\Omega} \quad \forall v \in V. \tag{7}$$

Condition (7) follows from the Friedrichs inequality (see e.g. [30]) provided at least one of the following conditions is satisfied: (a) the set Γ_{D} is a relatively open subset of $\partial\Omega$, (b) there exists a constant c_0 and a ball $B \subset \Omega$ such that $c - \dfrac{1}{2}\operatorname{div}\boldsymbol{b} \ge c_0 > 0$ a.e. in B, (c) there exists a constant α_0 and a relatively open subset Γ^0_{N} of Γ_{N} such that $\alpha + \dfrac{1}{2}\boldsymbol{b}\cdot\boldsymbol{n} \ge \alpha_0 > 0$ a.e. on Γ^0_{N}.

Let us remark that the continuity (6) and the V-ellipticity (7) of the bilinear form $a(\cdot,\cdot)$ guarantee the existence of a unique solution to problem (5). This weak solution is independent of the particular choice of the Dirichlet lift $\widetilde{g}_{\mathrm{D}}$.

Both the classical and the weak solutions of problem (1) satisfy the maximum principle under proper sign conditions. However, its rigorous formulation as well as its proof are different in the classical and in the weak setting due to technical reasons. For brevity, we formulate and prove the maximum principle in the weak setting only. The classical setting can be found for example in the well known monographs [21, 31].

Below we define four variants of the maximum principle. The following definition assumes that u is a solution of problem (5) corresponding to f, g_{D}, and g_{N} and that $u^{\pm} = (|u| \pm u)/2$ stands for the positive and negative part.

Definition 2.1 Problem (5) satisfies:

(a) the maximum principle if

$$f \le 0 \text{ a.e. in } \Omega \text{ and } g_{\mathrm{N}} \le 0 \text{ a.e. on } \Gamma_{\mathrm{N}} \quad \Rightarrow \quad \operatorname*{ess\,sup}_{\overline{\Omega}} u \le \operatorname*{ess\,sup}_{\Gamma_{\mathrm{D}}} u^{+},$$

(b) the minimum principle if

$$f \ge 0 \text{ a.e. in } \Omega \text{ and } g_{\mathrm{N}} \ge 0 \text{ a.e. on } \Gamma_{\mathrm{N}} \quad \Rightarrow \quad \operatorname*{ess\,inf}_{\overline{\Omega}} u \ge \operatorname*{ess\,inf}_{\Gamma_{\mathrm{D}}} -u^{-},$$

(c) the conservation of nonnegativity if

$$f \ge 0 \text{ a.e. in } \Omega,\ g_{\mathrm{D}} \ge 0 \text{ a.e. on } \Gamma_{\mathrm{D}},\ \text{and } g_{\mathrm{N}} \ge 0 \text{ a.e. on } \Gamma_{\mathrm{N}} \\ \Rightarrow \quad u \ge 0 \text{ a.e. in } \Omega,$$

(d) the comparison principle if

$$f_1 \ge f_2 \text{ a.e. in } \Omega,\ g_{\mathrm{D},1} \ge g_{\mathrm{D},2} \text{ a.e. on } \Gamma_{\mathrm{D}},\ \text{and } g_{\mathrm{N},1} \ge g_{\mathrm{N},2} \text{ a.e. on } \Gamma_{\mathrm{N}} \\ \Rightarrow \quad u_1 \ge u_2 \text{ a.e. in } \Omega,$$

where $u_i \in H^1(\Omega)$ is the solution to problem (5) with right-hand side f_i and boundary data $g_{\mathrm{D},i}$, $g_{\mathrm{N},i}$, respectively for $i = 1$ and 2.

Theorem 2.2 *Let $c \ge 0$ a.e. in Ω and $\alpha \ge 0$ a.e. on Γ_{N}. Then the principles* (a)–(d) *from Definition 2 are equivalent.*

It is not difficult to prove this theorem in a straightforward way. For this reason and to be brief, we skip the proof. The following theorem provides the validity of the maximum principle for problem (1). Although it is a well known result, we present its short proof for the reader's convenience. This proof is a variant of the proofs given e.g. in [21, 24, 27].

Theorem 2.3 *Let $c \geq 0$ a.e. in Ω, $\alpha \geq 0$ a.e. on Γ_{N}, and let the bilinear form $a(\cdot,\cdot)$ be V-elliptic, see* (7)*. Then problem* (5) *satisfies the maximum principle.*

Proof. Let us consider problem (5) with $f \leq 0$ a.e. in Ω, $g_{\mathrm{N}} \leq 0$ a.e. on Γ_{N} and with the corresponding solution $u \in H^1(\Omega)$. Let $M = \operatorname{ess\,sup}_{\Gamma_{\mathrm{D}}} u^+$ and $v(\boldsymbol{x}) = (u(\boldsymbol{x}) - M)^+$. Since the positive part w^+ is a continuous mapping from $H^1(\Omega)$ into itself, see e.g. [22, p. 29], the function v lies in $H^1(\Omega)$. Further, clearly, $M \geq 0$, $v \geq 0$ a.e. in Ω, $v = 0$ on Γ_{D} in the sense of traces, and $u = v + M$ whenever v does not vanish. These facts together with assumptions (2) and with the V-ellipticity of $a(\cdot,\cdot)$ enable to estimate

$$\begin{aligned}
0 &\geq \int_\Omega f v \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}} v \,\mathrm{d}\boldsymbol{s} \\
&= \int_\Omega [(\mathcal{A}\boldsymbol{\nabla} u)\cdot\boldsymbol{\nabla} v + \boldsymbol{b}\cdot\boldsymbol{\nabla} u\, v + cuv]\,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}} \alpha uv \,\mathrm{d}\boldsymbol{s} \\
&= \int_\Omega [(\mathcal{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{\nabla} v + \boldsymbol{b}\cdot\boldsymbol{\nabla} v\, v + c(v+M)v]\,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}} \alpha(v+M)v \,\mathrm{d}\boldsymbol{s} \\
&= a(v,v) + \int_\Omega cMv \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}} \alpha M v \,\mathrm{d}\boldsymbol{s} \geq a(v,v) \geq C\,\|v\|_{1,\Omega}^2 \geq 0.
\end{aligned}$$

Hence $v = 0$ a.e. in Ω and thus $u \leq M$ a.e. in Ω. □

3 Galerkin method and the discrete maximum principle

The idea of the Galerkin method is to project the infinite dimensional problem (5) to finite dimension. Therefore, we consider finite dimensional spaces V_h and X_h such that

$$X_h \subset H^1(\Omega), \quad V_h \subset V, \quad V_h \subset X_h \subset C(\overline{\Omega}),$$

where $C(\overline{\Omega})$ stands for the space of continuous functions in $\overline{\Omega}$. The particular choice of spaces V_h and X_h is not relevant at this point and we postpone their specifications to the subsequent sections.

The space X_h is used for the approximation of the Dirichlet lift $\widetilde{g}_{\mathrm{D}}$. Hence, let $\widetilde{g}_{\mathrm{D},h} \in X_h$ be such an approximation. The values of $\widetilde{g}_{\mathrm{D},h}$ on Γ_{D} are obtained in a suitable way (usually as the nodal interpolation or as the $L^2(\Gamma_{\mathrm{D}})$-projection of g_{D} into X_h) and the values in the interior nodes are often taken as zeros. However, the particular choice of $\widetilde{g}_{\mathrm{D},h} \in X_h$ is not important for the purposes of this paper.

The Galerkin solution $u_h \in X_h$ of problem (5) is uniquely defined as $u_h = u_h^0 + \widetilde{g}_{\mathrm{D},h}$, where $u_h^0 \in V_h$ satisfies

$$a(u_h^0, v_h) = \mathcal{F}(v_h) - a(\widetilde{g}_{\mathrm{D},h}, v_h) \quad \forall v_h \in V_h. \tag{8}$$

The bilinear form a and the linear functional $\mathcal{F}$ are given by (4). Its easy to see that for a fixed discrete Dirichlet lift $\widetilde{g}_{\mathrm{D},h}$ there exists a unique Galerkin solution u_h. However, it can be easily shown that the Galerkin solution u_h is unique independently of the choice of the Dirichlet lift $\widetilde{g}_{\mathrm{D},h} \in X_h$ provided the boundary values of $\widetilde{g}_{\mathrm{D},h}$ on Γ_{D} are fixed.

Problem (8) is equivalent to a system of linear algebraic equations. Indeed, if we consider a basis $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h, where $N = \dim V_h$, and if we express the solution u_h^0 as a linear combination of the basis functions as

$$u_h^0(\boldsymbol{x}) = \sum_{j=1}^{N} z_j \varphi_j(\boldsymbol{x})$$

then problem (8) is equivalent to a system of linear algebraic equations

$$Az = F, \tag{9}$$

where $z = (z_1, z_2, \ldots, z_N)^\top$ and the stiffness matrix $A \in \mathbb{R}^{N\times N}$ and the load vector $F \in \mathbb{R}^N$ have entries

$$A_{ij} = a(\varphi_j, \varphi_i) \quad \text{and} \quad F_i = \mathcal{F}(\varphi_i) - a(\widetilde{g}_{\mathrm{D},h}, \varphi_i), \quad i, j = 1, 2, \ldots, N. \tag{10}$$

Notice that the V-ellipticity of the bilinear form a implies the positive definiteness of A, i.e. the property

$$\boldsymbol{x}^T A \boldsymbol{x} > 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^N, \ \boldsymbol{x} \neq 0, \tag{11}$$

and hence the nonsingularity of A. We point out that A is nonsymmetric in general.

In order to handle the approximation $\widetilde{g}_{\mathrm{D},h}$ of the Dirichlet lift, we append the basis $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h by functions $\varphi_{N+1}, \varphi_{N+2}, \ldots, \varphi_{N+N^\partial}$ such that these functions all together form a basis in X_h. We define a space V_h^∂ as a linear span of the basis functions $\varphi_{N+1}, \varphi_{N+2}, \ldots, \varphi_{N+N^\partial}$. Hence, $X_h = V_h \oplus V_h^\partial$, where $\oplus$ denotes the direct sum, $\dim V_h^\partial = N^\partial$, and $\dim X_h = N + N^\partial$. For further references, we also set $\varphi_k^\partial = \varphi_{N+k}$ for $k = 1, 2, \ldots, N^\partial$.

Below, we will utilize also the matrix $A^\partial \in \mathbb{R}^{N\times N^\partial}$ with entries

$$A_{ik}^\partial = a(\varphi_k^\partial, \varphi_i), \quad i = 1, 2, \ldots, N, \quad k = 1, 2, \ldots, N^\partial. \tag{12}$$

Now, let us concentrate on the discrete maximum principle for the Galerkin solutions. On the discrete level, there is a straightforward analogy of Definition 2. The following definition assumes the spaces V_h and X_h to be fixed.

Definition 3.1 Problem (8) satisfies

(a) the discrete maximum principle if

$$f \le 0 \text{ a.e. in } \Omega \text{ and } g_{\mathrm{N}} \le 0 \text{ a.e. on } \Gamma_{\mathrm{N}} \quad \Rightarrow \quad \max_{\overline{\Omega}} u_h \le \max_{\Gamma_{\mathrm{D}}} u_h^+ .$$

(b) the discrete minimum principle if

$$f \ge 0 \text{ a.e. in } \Omega \text{ and } g_{\mathrm{N}} \ge 0 \text{ a.e. on } \Gamma_{\mathrm{N}} \quad \Rightarrow \quad \min_{\overline{\Omega}} u_h \ge \min_{\Gamma_{\mathrm{D}}} -u_h^- .$$

(c) the discrete conservation of nonnegativity if

$$f \ge 0 \text{ a.e. in } \Omega,\ \widetilde{g}_{\mathrm{D},h} \ge 0 \text{ a.e. on } \Gamma_{\mathrm{D}},\ \text{and } g_{\mathrm{N}} \ge 0 \text{ a.e. on } \Gamma_{\mathrm{N}} \quad \Rightarrow \quad u_h \ge 0.$$

(d) the discrete comparison principle if

$$f_1 \ge f_2 \text{ a.e. in } \Omega,\ \widetilde{g}_{\mathrm{D},h,1} \ge \widetilde{g}_{\mathrm{D},h,2} \text{ a.e. on } \Gamma_{\mathrm{D}},\ g_{\mathrm{N},1} \ge g_{\mathrm{N},2} \text{ a.e. on } \Gamma_{\mathrm{N}} \quad \Rightarrow \quad u_{h,1} \ge u_{h,2},$$

where $u_{h,i} \in X_h$ is the solution to problem (8) with right-hand side f_i and boundary data $\widetilde{g}_{\mathrm{D},h,i}$, $g_{\mathrm{N},i}$, respectively for $i = 1$ and 2.

Theorem 3.2 *Let the space X_h contain constant functions. Let $c \ge 0$ a.e. in Ω and $\alpha \ge 0$ a.e. on Γ_{N}. Then the principles (a)–(d) from Definition 3 are equivalent.*

The proof is analogous to the proof of Theorem 2.2 and we skip it again.

The validity of the DMP is not automatic. It depends not only on the problem and its parameters but also on the used discretization method and its parameters. In the case of Galerkin solutions it is the finite dimensional space V_h. The standard results about the DMP for the linear finite elements usually define a class of spaces V_h (or equivalently a class of triangulations) for which the DMP is satisfied, see Sections 6 and 8 below.

4 Discrete Green's function

In the context of the Galerkin method a natural discrete analog of the Green's function (see e.g. [35, 13]), the so called discrete Green's function (DGF), can be defined. The DGF possesses the analogous properties as the Green's function for continuous problems including the equivalence of the DMP with the nonnegativity of the DGF. This section defines the DGF and proves its properties.

Definition 4.1 Let $\boldsymbol{y} \in \Omega$ and let $G_{h,\boldsymbol{y}} \in V_h$ be the unique solution of the problem

$$a(v_h, G_{h,\boldsymbol{y}}) = v_h(\boldsymbol{y}) \quad \forall v_h \in V_h. \tag{13}$$

The function $G_h(\boldsymbol{x}, \boldsymbol{y}) = G_{h,\boldsymbol{y}}(\boldsymbol{x})$, $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega^2$ is called the discrete Green's function (DGF).

The above definition does not handle the action of the Dirichlet data g_{D}. Therefore, we consider the elliptic projection $\Pi_h^0 : X_h \mapsto V_h$ onto the space V_h. The elliptic projection $\Pi_h^0 w_h \in V_h$ of an $w_h \in X_h$ is uniquely determined by the requirement

$$a(w_h - \Pi_h^0 w_h, v_h) = 0 \quad \forall v_h \in V_h. \tag{14}$$

The DGF G_h and the elliptic projection Π_h^0 enable the following characterization of the Galerkin solution.

Theorem 4.2 *The Galerkin solution $u_h \in X_h$ to problem* (8) *satisfies the following representation formula:*

$$u_h(\boldsymbol{y}) = \mathcal{F}(G_{h,\boldsymbol{y}}) + \widetilde{g}_{\mathrm{D},h}(\boldsymbol{y}) - (\Pi_h^0 \widetilde{g}_{\mathrm{D},h})(\boldsymbol{y}) \tag{15}$$

Proof. By (13), (14), and (8) we immediately obtain

$$u_h^0(\boldsymbol{y}) + (\Pi_h^0 \widetilde{g}_{\mathrm{D},h})(\boldsymbol{y}) = a(u_h^0 + \Pi_h^0 \widetilde{g}_{\mathrm{D},h}, G_{h,\boldsymbol{y}}) = \mathcal{F}(G_{h,\boldsymbol{y}}).$$

Hence, statement (15) follows from the fact that $u_h = u_h^0 + \widetilde{g}_{\mathrm{D},h}$. □

Let us note that using the particular form (4) of the linear functional $\mathcal{F}$ we can express the representation formula (15) as

$$u_h(\boldsymbol{y}) = \int_\Omega f(\boldsymbol{x}) G_h(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}}(\boldsymbol{s}) G_h(\boldsymbol{s}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{s} + \widetilde{g}_{\mathrm{D},h}(\boldsymbol{y}) - (\Pi_h^0 \widetilde{g}_{\mathrm{D},h})(\boldsymbol{y}). \tag{16}$$

Here, we clearly observe the explicit dependence of the solution u_h on the data f, $g_{\mathrm{D},h}$, and g_{N}.

Remark 4.3 *The error of the elliptic projection* Π_h^0 *can be expressed as*

$$\Phi_{\boldsymbol{y}}(w_h) = w_h(\boldsymbol{y}) - (\Pi_h^0 w_h)(\boldsymbol{y}) = w_h(\boldsymbol{y}) - a(w_h, G_{h,\boldsymbol{y}}) \quad \forall w_h \in X_h.$$

Since $\Phi_{\boldsymbol{y}}(w_h) = 0$ *for all* $w_h \in V_h$ *and* $X_h = V_h \oplus V_h^\partial$*, we can regard* $\Phi_{\boldsymbol{y}}$ *as a linear and continuous functional on* V_h^∂*. Its values are determined by the values of* w_h *on* Γ_{D} *and by the Riesz representation theorem there exists a function* $G_{h,\boldsymbol{y}}^\partial \in V_h^\partial$ *such that*

$$\Phi_{\boldsymbol{y}}(w_h) = \int_{\Gamma_{\mathrm{D}}} w_h(\boldsymbol{s}) G_{h,\boldsymbol{y}}^\partial(\boldsymbol{s}) \,\mathrm{d}\boldsymbol{s}.$$

The function $G_{h,\boldsymbol{y}}^\partial(\boldsymbol{s})$ *is analogous to the normal derivative of the (continuous) Green's function appearing in the Green's formula for the exact solution* u*, see e.g. [31, p. 88]. However, it is not convenient to work with* $G_{h,\boldsymbol{y}}^\partial$ *on the discrete level, because the nonnegativity of* $\Phi_{\boldsymbol{y}}(w_h)$ *for all nonnegative* $w_h \in V_h^\partial$ *does not imply the nonnegativity of* $G_{h,\boldsymbol{y}}^\partial$ *and we cannot prove the corresponding equivalent conditions for the DMP, cf. Theorem 4.4 below. Moreover, up to rare exceptions, the function* $G_{h,\boldsymbol{y}}^\partial$ *is practically never nonnegative. Therefore, we do not use* $G_{h,\boldsymbol{y}}^\partial$ *and analyze the error of the elliptic projection* $w_h - \Pi_h^0 w_h$ *instead.*

The following theorem shows the equivalent conditions for the validity of the DMP.

Theorem 4.4 *Problem* (8) *satisfies the discrete conservation of nonnegativity if and only if*

(a) $\quad G_h(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \Omega^2$*,*

(b) $\quad \widetilde{g}_{\mathrm{D},h}(\boldsymbol{y}) - (\Pi_h^0 \widetilde{g}_{\mathrm{D},h})(\boldsymbol{y}) \geq 0$ *for all* $\widetilde{g}_{\mathrm{D},h} \in V_h^\partial$*,* $\widetilde{g}_{\mathrm{D},h} \geq 0$ *in* Ω*,* $\boldsymbol{y} \in \Omega$*.*

Proof. The fact that conditions (a) and (b) imply the conservation of nonnegativity is immediate form (16). The opposite implication follows from (16), too. Indeed, taking $\boldsymbol{y} \in \Omega$, $g_{\mathrm{N}} = 0$, and $\widetilde{g}_{\mathrm{D},h} = 0$, the conservation of nonnegativity yields

$$u_h(\boldsymbol{y}) = \int_\Omega f(\boldsymbol{x}) G_h(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{x} \geq 0$$

for any $f \in L^2(\Omega)$ such that $f \geq 0$ a.e. in Ω. Thus, $G_{h,\boldsymbol{y}} \geq 0$ a.e. in Ω and since $G_{h,\boldsymbol{y}}$ is continuous, it is nonnegative everywhere in Ω. Condition (b) follows trivially from the conservation of nonnegativity and from (16) with $f = 0$ and $g_{\mathrm{N}} = 0$. □

The Green's function on the continuous level can be explicitly found in exceptional cases only. In contrast, the DGF can always be computed, at least theoretically. The following theorem shows an explicit expression for the DGF in terms of the inverse of the stiffness matrix A, see (10). We point out that a version of this result based on eigenfunctions of the discrete Laplacian was published already in 1970 in [8] and [11]. Anyway, for the reader's convenience we present its proof here, although it can be found in [43], too.

Theorem 4.5 *Let $\varphi_1, \varphi_2, \ldots, \varphi_N$ be a basis in V_h and let A be the corresponding stiffness matrix given by* (10)*. Then the DGF can be expressed as follows*

$$G_h(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \varphi_i(\boldsymbol{y}) (A^{-1})_{ij} \varphi_j(\boldsymbol{x}). \tag{17}$$

Proof. The DGF $G_{h,\boldsymbol{y}}$ is defined as an element of V_h, hence, it can be expanded as a linear combination of the basis functions

$$G_{h,\boldsymbol{y}}(\boldsymbol{x}) = \sum_{j=1}^{N} d_j(\boldsymbol{y}) \varphi_j(\boldsymbol{x}). \tag{18}$$

Using this expansion in (13) tested by all the basis functions, we obtain

$$\varphi_i(\boldsymbol{y}) = a\left(\varphi_i, \sum_{j=1}^{N} d_j(\boldsymbol{y}) \varphi_j(\boldsymbol{x})\right) = \sum_{j=1}^{N} d_j(\boldsymbol{y}) A_{ji}, \quad i = 1, 2, \ldots, N.$$

Since the stiffness matrix is nonsingular, we can multiply this identity by the inverse matrix to express the coefficients $d_k(\boldsymbol{y})$:

$$d_k(\boldsymbol{y}) = \sum_{i=1}^{N} \varphi_i(\boldsymbol{y}) (A^{-1})_{ik}, \quad k = 1, 2, \ldots, N.$$

Inserting this into (18), we obtain (17). □

The error of the elliptic projection $\Pi_h^0 \widetilde{g}_{\mathrm{D},h}$ needed in the representation formula (16) can be expressed in a similar way as the DGF using the basis functions and the stiffness matrices A and A^{∂}.

Theorem 4.6 *Let $X_h = V_h \oplus V_h^\partial$, let $\varphi_1, \varphi_2, \dots, \varphi_N$ be a basis in V_h, let $\varphi_1^\partial, \varphi_2^\partial, \dots, \varphi_N^\partial$ be a basis in V_h^∂, and let the matrices A and A^∂ be given by* (10) *and* (12)*, respectively. Let the approximation of the Dirichlet lift $\widetilde{g}_{\mathrm{D},h} \in X_h$ be expressed as*

$$\widetilde{g}_{\mathrm{D},h}(\boldsymbol{y}) = \sum_{\ell=1}^{N^\partial} c_\ell^\partial \varphi_\ell^\partial(\boldsymbol{y}) + \sum_{i=1}^{N} c_i^0 \varphi_i(\boldsymbol{y}) \quad \forall \boldsymbol{y} \in \Omega.$$

Then

$$\widetilde{g}_{\mathrm{D},h}(\boldsymbol{y}) - \Pi_h^0 \widetilde{g}_{\mathrm{D},h}(\boldsymbol{y}) = \sum_{\ell=1}^{N^\partial} c_\ell^\partial \left[\varphi_\ell^\partial(\boldsymbol{y}) - \Pi_h^0 \varphi_\ell^\partial(\boldsymbol{y})\right] \quad \forall \boldsymbol{y} \in \Omega, \tag{19}$$

where the elliptic projection of the basis functions φ_ℓ^∂ can be expressed as

$$\Pi_h^0 \varphi_\ell^\partial(\boldsymbol{y}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \varphi_i(\boldsymbol{y}) (A^{-1})_{ij} A_{j\ell}^\partial \quad \forall \boldsymbol{y} \in \Omega. \tag{20}$$

Proof. The equality (19) is immediate from the linearity of the elliptic projection Π_h^0 and from the fact that $\Pi_h^0 \varphi_i = \varphi_i$ for all $i = 1, 2, \dots, N$. To prove (20), we express $\Pi_h^0 \varphi_\ell^\partial \in V_h$ as

$$\Pi_h^0 \varphi_\ell^\partial = \sum_{i=1}^{N} d_{\ell i} \varphi_i. \tag{21}$$

This expansion substituted to the definition of the elliptic projection (14) yields

$$\sum_{i=1}^{N} d_{\ell i} a(\varphi_i, \varphi_j) = a(\varphi_\ell^\partial, \varphi_j) \quad \forall j = 1, 2, \dots, N.$$

Consequently, by (10) and (12) we can express the coefficients $d_{\ell i}$ in terms of the inverse matrix to the stiffness matrix A as follows

$$d_{\ell i} = \sum_{j=1}^{N} (A^{-1})_{ij} A_{j\ell}^\partial.$$

The statement (20) follows by substitution of this into (21). □

Let us point out that statements (17) and (20) of Theorems 4.5 and 4.6 can be written in a more compact way using the matrix notation. If

$\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \dots, \varphi_N)^\top$ and $\boldsymbol{\varphi^\partial} = (\varphi_1^\partial, \varphi_2^\partial, \dots, \varphi_{N^\partial}^\partial)^\top$ stand for the vectors of basis functions then (17) and (20) can be expressed as

$$G_h(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\varphi}(\boldsymbol{x})^\top A^{-\top} \boldsymbol{\varphi}(\boldsymbol{y}) \quad \text{and} \quad \Pi_h^0 \boldsymbol{\varphi^\partial}(\boldsymbol{y}) = (A^\partial)^\top A^{-\top} \boldsymbol{\varphi}(\boldsymbol{y}).$$

In addition, notice that formula (17) implies that not only $G_{h,\boldsymbol{y}} = G_h(\cdot, \boldsymbol{y})$ but also $G_{h,\boldsymbol{x}} = G_h(\boldsymbol{x}, \cdot)$ belongs to V_h for all $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega^2$.

Theorems 4.4–4.6 represent a general concept for investigation of the DMP for the Galerkin solutions and in particular for the finite element method (FEM). Theorem 4.4 shows the equivalence of the DMP with the nonnegativity of the DGF and with the nonnegativity of the error of the elliptic projection of the discrete Dirichlet lift. Theorems 4.5 and 4.6 provide explicit formulas for the DGF and for the error of the elliptic projection. In certain cases these formulas enable to deduce certain sufficient conditions for the nonnegativity of the DGF and consequently for the validity of the DMP. In the case of the lowest-order FEM the investigation of the nonnegativity of the DGF is equivalent to the investigation of the monotonicity of the corresponding matrices, see Section 7 below. In the case of the higher-order FEM, not only the matrices but also the basis functions play a crucial role as we briefly indicate in Section 6.

5 Finite element method

The finite element method (FEM) can be seen as a special case of the Galerkin method, where the finite element spaces V_h and X_h are chosen in such a way that the corresponding stiffness matrices A and A^∂, see (10) and (12), are sparse and efficient linear algebraic solvers for system (9) can be employed. Practically, the spaces V_h and X_h are constructed using a finite element mesh $\mathcal{T}_h$. The mesh is a partition of the domain Ω into a finite number of geometrically simple subdomains – known as *elements*. The elements are typically simplices or blocks (Cartesian products of intervals). For the purposes of this paper, it suffices to consider the finite element mesh as a set $\mathcal{T}_h = \{K_i : i = 1, 2, \dots, M\}$ of elements $K_i \subset \overline{\Omega}$. The elements are traditionally considered as closed sets with nonzero measure, their interiors are pairwise disjoint, and their union is the entire $\overline{\Omega}$. For a rigorous definition of the finite element mesh and for a detailed treatment of the FEM see e.g. [9, 36].

Anyway, the partition of the domain Ω into the elements enables to split the bilinear and linear forms a and $\mathcal{F}$ into local (element) contributions:

$$a(u, v) = \sum_{K \in \mathcal{T}_h} a_K(u, v) \quad \text{and} \quad \mathcal{F}(v) = \sum_{K \in \mathcal{T}_h} \mathcal{F}_K(v) \quad \forall u, v \in H^1(\Omega), \quad (22)$$

where in accordance with (4) we put

$$a_K(u,v) = \int\limits_K [(\mathcal{A}\boldsymbol{\nabla} u)\cdot\boldsymbol{\nabla} v + (\boldsymbol{b}\cdot\boldsymbol{\nabla} u)v + cuv]\,\mathrm{d}\boldsymbol{x} + \int\limits_{\Gamma_\mathrm{N}\cap K} \alpha uv\,\mathrm{d}\boldsymbol{s}, \tag{23}$$

$$\mathcal{F}_K(v) = \int\limits_K fv\,\mathrm{d}\boldsymbol{x} + \int\limits_{\Gamma_\mathrm{N}\cap K} g_\mathrm{N} v\,\mathrm{d}\boldsymbol{s}.$$

These local bilinear forms a_K and the above introduced basis functions of V_h and V_h^∂ can be used to define the local stiffness matrices (some authors call them element stiffness matrices) $\overline{A}^K \in \mathbb{R}^{N\times N}$ and $\overline{A}^{\partial,K} \in \mathbb{R}^{N\times N^\partial}$ as

$$\begin{aligned} \overline{A}^K_{ij} &= a_K(\varphi_j,\varphi_i), \quad i,j=1,2,\dots,N,\\ \overline{A}^{\partial,K}_{ik} &= a_K(\varphi^\partial_k,\varphi_i), \quad i=1,2,\dots,N, \quad k=1,2,\dots,N^\partial. \end{aligned}$$

However, if the basis function are defined using the standard finite element machinery then only a few basis functions are supported in a single element and, therefore, the corresponding local stiffness matrices $\overline{A}^K$ and $\overline{A}^{\partial,K}$ have many zero entries. Thus, they can be condensed into matrices with smaller dimension by leaving out their zero entries. To perform formally this condensation, we have to introduce the so called connectivity mappings.

Let us define sets $\bar{I}(K)$, $I(K)$, and $I^\partial(K)$ of indices of basis functions whose support contains an element K:

$$\begin{aligned} \bar{I}(K) &= \{i\in\mathbb{N} : 1\le i\le N+N^\partial,\ K\subset \operatorname{supp}\varphi_i\},\\ I(K) &= \{j\in\mathbb{N} : 1\le j\le N,\ K\subset \operatorname{supp}\varphi_j\},\\ I^\partial(K) &= \{k\in\mathbb{N} : 1\le k\le N^\partial,\ K\subset \operatorname{supp}\varphi^\partial_k\}. \end{aligned}$$

We denote by $\bar{N}_K$, N_K, and N_K^∂ the number of indices in the sets $\bar{I}(K)$, $I(K)$, and $I^\partial(K)$, respectively. Clearly, $I(K)\subset\bar{I}(K)$ and $\bar{N}_K = N_K + N_K^\partial$. By *connectivity mappings* we understand arbitrary but fixed one-to-one mappings $\bar{\iota}_K : \{1,2,\dots,\bar{N}_K\}\mapsto\bar{I}(K)$, $\iota_K : \{1,2,\dots,N_K\}\mapsto I(K)$, and $\iota_K^\partial : \{1,2,\dots,N_K^\partial\}\mapsto I^\partial(K)$ such that $\bar{\iota}_K(m)=\iota_K(m)$ for $m=1,2,\dots,N_K$ and $\bar{\iota}_K(N_K+m) = N+\iota^\partial_K(m)$ for $m=1,2,\dots,N^\partial_K$. These connectivity mappings are of a practical significance and they play an important role in many finite element codes, see e.g. [34].

The concept of elements and connectivity mappings enables to understand each basis function φ_i as a composition of so-called *shape functions* φ^K_m defined on elements $K\in\mathcal{T}_h$. The relation is $\varphi^K_m = \varphi_i|_K$ with $i=\bar{\iota}_K(m)$, $i=1,2,\dots,N+N^\partial$ and $m=1,2,\dots,\bar{N}_K$. In particular, we set $\varphi^{K,\partial}_q = \varphi^K_{N_K+q} = \varphi^\partial_j|_K$ with $j=\iota^\partial_K(q)$, $j=1,2,\dots,N^\partial$ and $q=1,2,\dots,N^\partial_K$.

Now, we determine the condensed local stiffness matrices $A^K \in \mathbb{R}^{N_K \times N_K}$ and $A^{\partial,K} \in \mathbb{R}^{N_K \times N_K^\partial}$ by their entries

$$A^K_{mn} = \overline{A}^K_{\iota_K(m),\iota_K(n)} = a_K\big(\varphi^K_n, \varphi^K_m\big), \qquad m,n = 1,2,\ldots,N_K, \tag{24}$$

$$A^{\partial,K}_{mq} = \overline{A}^{\partial,K}_{\iota_K(m),\iota^\partial_K(q)} = a_K\big(\varphi^{K,\partial}_q, \varphi^K_m\big), \quad m = 1,2,\ldots,N_K,\ q = 1,2,\ldots,N^\partial_K. \tag{25}$$

Using (22) and the above definitions, we can express the entries of the (global) matrices A and A^∂ as follows

$$A_{ij} = \sum_{K\in\mathcal{T}_h} a_K(\varphi_j, \varphi_i) = \sum_{K\in\mathcal{T}_h} \overline{A}^K_{ij} = \sum_{\{K\in\mathcal{T}_h : i,j\in I(K)\}} A^K_{\iota_K^{-1}(i),\iota_K^{-1}(j)}, \tag{26}$$

$$A^\partial_{ik} = \sum_{K\in\mathcal{T}_h} a_K(\varphi^\partial_k, \varphi_i) = \sum_{K\in\mathcal{T}_h} \overline{A}^{\partial,K}_{ij} = \sum_{\{K\in\mathcal{T}_h : i\in I(K),\ k\in I^\partial(K)\}} A^{\partial,K}_{\iota_K^{-1}(i),(\iota^\partial_K)^{-1}(k)}, \tag{27}$$

where $i,j = 1,2,\ldots,N$ and $k = 1,2,\ldots,N^\partial$. These formulas show how to assemble the global stiffness matrices A and A^∂ from the local matrices A^K and $A^{\partial,K}$. This process is known as the *assembling* in the finite element community. Below, we will solely use the condensed local stiffness matrices A^K and $A^{\partial,K}$ and we will call them simply local (stiffness) matrices.

6 Applications to higher-order FEM

The above described general concept can be successfully applied to the analysis of the DMP for higher-order FEM. It is especially useful for simple 1D problems. As an example, we present results published in [45, 44]. Let us consider the following 1D diffusion problem with Dirichlet boundary conditions:

$$-(\mathcal{A}u')' = f \quad \text{in } (a^\partial, b^\partial), \quad u(a^\partial) = g_{\mathrm{D}}(a^\partial), \quad u(b^\partial) = g_{\mathrm{D}}(b^\partial). \tag{28}$$

We discretize this problem by a higher-order finite element method. Therefore, we introduce a partition $a^\partial = x_0 < x_1 < \cdots < x_{M-1} < x_M = b^\partial$ of the interval (a^∂, b^∂) and define elements $K_k = [x_{k-1}, x_k]$, $k = 1,2,\ldots,M$, with $h_k = x_k - x_{k-1}$. For each element K_k we assign a polynomial degree p_k, $k = 1,2,\ldots,M$, and set the higher-order finite element space

$$X_h = \{v_h \in H^1(\Omega) : v_h|_{K_k} \in \mathbb{P}^{p_k}(K_k),\ k = 1,2,\ldots,M\},$$

where $\mathbb{P}^p(K)$ stands for the space of polynomials of degree at most p on interval K. To incorporate the Dirichlet boundary conditions we introduce

Table 1: Critical relative element length $H^*_{\mathrm{rel}}(p)$ for $p = 1, 2, \ldots, 20$.

p	$H^*_{\mathrm{rel}}(p)$
1	1
2	1
3	9/10
4	1
5	0.919731
6	1
7	0.935127
8	0.987060
9	0.945933
10	0.973952
11	0.953759
12	0.969485
13	0.959646
14	0.968378
15	0.964221
16	0.968695
17	0.967874
18	0.969629
19	0.970855
20	0.970814

a subspace $V_h \subset X_h$ of functions vanishing at both end-points a^∂ and b^∂. The higher-order finite element solution $u_h \in X_h$ is then determined by the requirements $u_h - \widetilde{g}_{\mathrm{D},h} \in V_h$ and

$$a(u_h, v_h) = \mathcal{F}(v_h) \quad \forall v_h \in V_h, \tag{29}$$

where the approximate Dirichlet lift $\widetilde{g}_{\mathrm{D},h}$ can be taken as the linear function having the prescribed values $g_{\mathrm{D}}(a^\partial)$ and $g_{\mathrm{D}}(b^\partial)$ at the end-points a^∂ and b^∂. The bilinear form a and the linear functional $\mathcal{F}$ have the particular form

$$a(u, v) = \int_{a^\partial}^{b^\partial} \mathcal{A} u' v' \,\mathrm{d}x \quad \text{and} \quad \mathcal{F}(v) = \int_{a^\partial}^{b^\partial} f v \,\mathrm{d}x.$$

This higher-order finite element discretization has several special features. The standard basis functions in X_h are of two types – the piecewise linear functions and the higher-order (bubble) functions. The higher-order functions are orthogonal (in the energy sense) to the piecewise linear ones. This orthogonality enables to split the DGF into the piecewise linear part and the higher-order part. In addition an explicit formula for the inverse of the stiffness matrix A exists. Therefore, the nonnegativity of the DGF can be straightforwardly analyzed using formula (13). Afterall, we obtain a sufficient condition for the validity of the DMP in terms of the lengths of the elements.

For the reader's convenience, we present Theorem 6.1 showing the main statement. In order to formulate it, we introdude so called critical relative element length $H^*_{\mathrm{rel}}(p)$. For a given polynomial degree p, it is defined as a minimum of a certain polynomial of two variables. For the purposes of this paper we present the values of $H^*_{\mathrm{rel}}(p)$ for $p = 1, 2, \ldots, 20$ in Table 1.

Theorem 6.1 *Let us consider problem* (28) *with piecewise constant coefficient $\mathcal{A}$. Further, let us consider its higher-order finite element discretization*

(29). *If*

$$\frac{h_k}{\mathcal{A}_k} \leq H^*_{\mathrm{rel}}(p_k) h_{\Omega,\mathcal{A}} \quad \textit{for all } k = 1, 2, \ldots, M,$$

with $h_{\Omega,\mathcal{A}} = \sum_{j=1}^{M} h_j/\mathcal{A}_j$ *and* $\mathcal{A}_k$ *denoting the constant value of* $\mathcal{A}$ *on the* k*-th element, then discretization* (29) *satisfies the discrete conservation of nonnegativity.*

The detailed analysis of this case including the proof of this theorem can be found in [45] for $\mathcal{A} = 1$ and in [44] for the general piecewise constant coefficient $\mathcal{A}$. Here we only point out the necessity of the above described concept (see Theorem 4.4) for the analysis of the DMP for higher-order finite element methods.

The result of Theorem 6.1 can be also generalized to the mixed Dirichlet-Neumann boundary conditions [46]. Generalization to the diffusion-reaction problem $-u'' + \kappa^2 u = f$ is however much more demanding, because the explicit formula for the inverse of the stiffness matrix A is no longer available. Nevertheless, technically complicated estimates of entries A^{-1} can be done and the resulting sufficient conditions for the DMP were published in [42].

The general concept of the DGF can be well used for higher dimensional problems, too. However, a straightforward analysis based on Theorem 4.4 is too demanding already in 2D. Moreover, the numerical experiments performed in [41] indicate that the DMP is satisfied for higher-order finite elements in two and more dimensions in exceptional cases only.

7 The lowest-order FEM

Since the DMP results for the lowest-order FEM are based on matrix theory, we first recall several notions and statements from this field. A real matrix A is said to be *nonnegative* if all its entries are nonnegative and it is denoted by inequality $A \geq 0$. A matrix $A \in \mathbb{R}^{N\times N}$ is said to be *monotone* if it is nonsingular and $A^{-1} \geq 0$. In the following definition, we introduce a special notation for the off-diagonal part of a matrix.

Definition 7.1 Let $A \in \mathbb{R}^{N\times N}$ be a real square matrix. The *off-diagonal part* of A is a matrix $B \in \mathbb{R}^{N\times N}$ with entries $B_{ii} = 0$ for $i = 1, 2, \ldots, N$ and $B_{ij} = A_{ij}$ for $i \neq j$, $i, j = 1, 2, \ldots, N$. We denote the off-diagonal part of A by off-diag(A).

The crucial class of matrices for our purposes are the M-matrices. A matrix $A \in \mathbb{R}^{N\times N}$ is said to be an *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 7.2 *Let a matrix $A \in \mathbb{R}^{N \times N}$ be positive definite, see* (11)*, and let* off-diag$(A) \leq 0$*. Then A is M-matrix, i.e., $A^{-1} \geq 0$.*

Proof. Clearly, the positive definiteness (11) implies that all real eigenvalues of A are positive. The rest follows from [19, Thm. 5.1,p. 114]. □

Let us remark that Theorem 7.2 is a generalization of the well known result of Varga [39, p. 85] to nonsymmetric matrices.

Now, we can describe the general concept of investigation of the DMP for the lowest-order FEM. It is based on Theorem 4.4 and on the simple characterization of nonnegativity of the lowest-order approximations. For example, a piecewise linear function is nonnegative in a domain Ω if and only if it is nonnegative in all nodal points. This advantageous property of the lowest-order finite elements enable to refine the general result from Theorem 4.4.

To formalize the idea, we consider (as above) the finite dimensional spaces $X_h = V_h \oplus V_h^\partial$, with $N = \dim V_h$, $N^\partial = \dim V_h^\partial$, with a basis $\varphi_1, \varphi_2, \dots, \varphi_N$ of V_h, and with a basis $\varphi_1^\partial, \varphi_2^\partial, \dots, \varphi_{N^\partial}^\partial$ of V_h^∂. For these basis functions we assume the following properties:

$$\sum_{i=1}^{N} c_i \varphi_i(\boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in \Omega \quad \Leftrightarrow \quad c_i \geq 0 \quad \forall i = 1, 2, \dots, N, \tag{30}$$

$$\sum_{\ell=1}^{N^\partial} c_\ell^\partial \varphi_\ell^\partial(\boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in \Omega \quad \Leftrightarrow \quad c_\ell^\partial \geq 0 \quad \forall \ell = 1, 2, \dots, N^\partial. \tag{31}$$

Let us notice that the standard (Lagrangian) lowest-order finite element basis functions, like piecewise linear functions on simplices or piecewise multi-linear functions on blocks, satisfy these properties.

Theorem 7.3 *Let the finite dimensional spaces V_h and V_h^∂ posses basis functions $\varphi_1, \varphi_2, \dots, \varphi_N$ and $\varphi_1^\partial, \varphi_2^\partial, \dots, \varphi_{N^\partial}^\partial$ with properties* (30) *and* (31)*. Then problem* (8) *satisfies the discrete conservation of nonnegativity if and only if*

$$A^{-1} \geq 0 \quad \textit{and} \quad -A^{-1}A^\partial \geq 0,$$

where A and A^∂ stand for the matrices (10) *and* (12)*.*

Proof. The proof follows from Theorems 4.4–4.6 and from the facts that in the lowest-order case (i) the DGF G_h is nonnegative if and only if $A^{-1} \geq 0$ and (ii) the error of the elliptic projection $\widetilde{g}_{\mathrm{D},h} - \Pi_h^0 \widetilde{g}_{\mathrm{D},h}$ is nonnegative for all $\widetilde{g}_{\mathrm{D},h} \geq 0$, $\widetilde{g}_{\mathrm{D},h} \in V_h^\partial$ if and only if $-A^{-1}A^\partial \geq 0$.

The equivalence (i) follows from the expression (17) and from the property (30). Indeed, the DGF G_h can be expressed as a linear combination of basis functions as follows:

$$G_h(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^N \gamma_i(\boldsymbol{x})\varphi_i(\boldsymbol{y}), \quad \text{where } \gamma_i(\boldsymbol{x}) = \sum_{j=1}^N (A^{-1})_{ij}\varphi_j(\boldsymbol{x}).$$

Hence, property (30) yields that $G_h(\boldsymbol{x},\boldsymbol{y}) \geq 0$ for all $(\boldsymbol{x},\boldsymbol{y}) \in \Omega^2$ if and only if $\gamma_i(\boldsymbol{x}) \geq 0$ for all $i = 1,2,\ldots,N$ and all $\boldsymbol{x} \in \Omega$. Using the property (30) again we obtain that $\gamma_i(\boldsymbol{x}) \geq 0$ for all $i = 1,2,\ldots,N$ and all $\boldsymbol{x} \in \Omega$ if and only if $(A^{-1})_{ij} \geq 0$ for all $i,j = 1,2,\ldots,N$.

To prove the equivalence (ii) we proceed as follows. According to (31) and (19), the statement

$$\widetilde{g}_{\mathrm{D},h} - \Pi_h^0 \widetilde{g}_{\mathrm{D},h} \geq 0 \quad \forall \widetilde{g}_{\mathrm{D},h} \geq 0,\ \widetilde{g}_{\mathrm{D},h} \in V_h^\partial$$

is equivalent to

$$\sum_{\ell=1}^{N^\partial} c_\ell^\partial \left[\varphi_\ell^\partial - \Pi_h^0 \varphi_\ell^\partial\right] \geq 0 \quad \forall c_\ell^\partial \geq 0,\ \ell = 1,2,\ldots,N^\partial.$$

This is further equivalent to

$$\varphi_\ell^\partial - \Pi_h^0 \varphi_\ell^\partial \geq 0 \quad \forall \ell = 1,2,\ldots,N^\partial.$$

However, by (20) we can express the difference $\varphi_\ell^\partial - \Pi_h^0 \varphi_\ell^\partial$ as a linear combination $\varphi_\ell^\partial + \sum_{i=1}^N D_{i\ell}\varphi_i$ with $D_{i\ell} = -\sum_{j=1}^N (A^{-1})_{ij} A_{j\ell}^\partial$. Such a linear combination is nonnegative by (30) and (31) if and only if $D_{i\ell} \geq 0$ for all $i = 1,2,\ldots,N$ and $\ell = 1,2,\ldots,N^\partial$. □

The above theorem provides equivalent characterization of the DMP by means of the global stiffness matrices. However, detailed investigation of the inverse A^{-1} and of the product $A^{-1}A^\partial$ might be complicated. This can be avoided for the price of losing the necessity of the obtained conditions. The following theorem provides a sufficient condition formulated in terms of entries of A and A^∂ only.

Theorem 7.4 *Let the finite dimensional spaces V_h and V_h^∂ posses basis functions $\varphi_1, \varphi_2, \dots, \varphi_N$ and $\varphi_1^\partial, \varphi_2^\partial, \dots, \varphi_{N^\partial}^\partial$ with properties* (30) *and* (31)*. Let A and A^∂ be the stiffness matrices given by* (10) *and* (12)*. If*

$$\text{off-diag}\, A \leq 0 \quad \textit{and} \quad A^\partial \leq 0$$

then problem (8) *satisfies the discrete conservation of nonnegativity.*

Proof. The statement follows immediately from Theorems 7.3 and 7.2. □

The verification of the nonpositivity of the entries of the (global) matrices A and A^∂ can be made even more convenient by checking the local matrices A^K and $A^{\partial,K}$ only. The next theorem formulate a sufficient condition for the DMP in terms of these local matrices.

Theorem 7.5 *Let the finite dimensional spaces V_h and V_h^∂ posses basis functions $\varphi_1, \varphi_2, \dots, \varphi_N$ and $\varphi_1^\partial, \varphi_2^\partial, \dots, \varphi_{N^\partial}^\partial$ with properties* (30) *and* (31)*. Let $\mathcal{T}_h$ be a finite element mesh and A^K and $A^{\partial,K}$, $K \in \mathcal{T}_h$, be the local stiffness matrices introduced in* (24) *and* (25)*. If*

$$\text{off-diag}\, A^K \leq 0 \quad \textit{and} \quad A^{\partial,K} \leq 0 \quad \forall K \in \mathcal{T}_h$$

then problem (8) *satisfies the discrete conservation of nonnegativity.*

Proof. The statement follows directly from Theorem 7.4 and from (26) and (27). □

8 Applications to the lowest-order simplicial elements

In the sequel, we will analyze the following simplified version of problem (1)

$$-\operatorname{div}(\lambda \boldsymbol{\nabla} u) + cu = f \quad \text{in } \Omega, \quad u = g_\mathrm{D} \quad \text{on } \Gamma_\mathrm{D}, \quad \alpha u + \lambda \boldsymbol{\nabla} u \cdot \boldsymbol{n} = g_\mathrm{N} \quad \text{on } \Gamma_\mathrm{N}. \tag{32}$$

In comparison with the general diffusion-convection-reaction problem (1) we consider in (32) no convection ($\boldsymbol{b} = \boldsymbol{0}$) and the general anisotropic tensor $\mathcal{A}$ in the diffusion term is replaced by an isotropic coefficient λ, i.e., we have set $\mathcal{A}(x) = \lambda(x) I$. We continue to assume the general requirements described in Section 2. Namely, the assumption (3) of the uniform positive definiteness of $\mathcal{A}$ turns into to the boundedness of λ from below

$$0 < \lambda_{\min} \leq \lambda(x) \quad \text{for } x \in \Omega$$

and assumptions (2) simplify to $c \geq 0$ in Ω and $\alpha \geq 0$ on Γ_{N}.

Remark 8.1 *Successful approximate solution of the general problem* (1) *with nonvanishing convection coefficient* $\boldsymbol{b}$ *by the finite element method is a delicate problem, because it requires special stabilization approaches [32, 25]. Detailed investigation of this case is out of scope of this paper and therefore we consider* $\boldsymbol{b} = \mathbf{0}$ *in* (32)*. Similarly, the treatment of the general anisotropic tensor* $\mathcal{A} \in \mathbb{R}^{d \times d}$ *is complicated and we reffer to [29] for details.*

Now, let us describe the lowest-order simplicial finite elements for discretization of problem (32). For simplicity, we consider the domain $\Omega \subset \mathbb{R}^d$, $d \geq 2$, to be polytopic. The corresponding finite element mesh $\mathcal{T}_h$ consists of d dimensional simplices K, which form a face-to-face partition of $\overline{\Omega}$. The vertices of all simplices from $\mathcal{T}_h$ are referred as nodes or nodal points.

The lowest-order finite element space X_h is then defined as

$$X_h = \{w_h \in H^1(\Omega) : w_h|_K \in \mathbb{P}^1(K) \quad \text{for all simplices } K \in \mathcal{T}_h\},$$

where $\mathbb{P}^1(K)$ stands for the space of linear functions on the simplex K. The functions in X_h are necessarily continuous and each of them is uniquely determined by its values in the nodal points. As above we consider the subspace $V_h \subset X_h$ of functions vanishing on $\overline{\Gamma}_{\mathrm{D}}$ and the space V_h^∂ such that $X_h = V_h \oplus V_h^\partial$. The standard lowest-order finite element basis functions $\varphi_1, \varphi_2, \dots, \varphi_N$ in V_h are uniquely determined by the δ-property

$$\varphi_i(\boldsymbol{x}_j) = \delta_{ij}, \quad i, j = 1, 2, \dots, N,$$

where δ_{ij} stands for Kronecker's delta and $\boldsymbol{x}_i$, $i = 1, 2, \dots, N$, are the nodal points not lying on $\overline{\Gamma}_{\mathrm{D}}$. Similarly, the standard finite element basis functions $\varphi_1^\partial, \varphi_2^\partial, \dots, \varphi_N^\partial$ in V_h^∂ are uniquely determined by the δ-property

$$\varphi_k^\partial(\boldsymbol{x}_\ell^\partial) = \delta_{k\ell}, \quad k, \ell = 1, 2, \dots, N^\partial,$$

where $\boldsymbol{x}_i^\partial$, $i = 1, 2, \dots, N^\partial$, are the nodal points lying on $\overline{\Gamma}_{\mathrm{D}}$.

The lowest-order finite element solution $u_h = u_h^0 + \widetilde{g}_{\mathrm{D},h}$ is now given by the Galerkin formulation (8) with the simplicial finite element spaces V_h and X_h described above.

From the point of view of the DMP the simplicial finite elements have advantageous properties. Namely, there exist simple formulas for the key integrals present in relations (23) for the entries of the local stiffness matrices. However, in order to introduce these formulas, we have to set certain notations.

Let $K \in \mathcal{T}_h$ be a simplex. We denote its vertices by $\boldsymbol{x}_\ell^K$, $\ell = 1, 2, \dots, \bar{N}_K$, $\bar{N}_K = d + 1$. The connection between the vertices of the simplex K and the nodes of the mesh $\mathcal{T}_h$ is provided by the connectivity mapping: $\boldsymbol{x}_\ell^K = \boldsymbol{x}_i$, where $i = \bar{\imath}_K(\ell)$, $\ell = 1, 2, \dots, \bar{N}_K$. We denote by F_ℓ and F_m the two facets of the simplex K opposite the vertices $\boldsymbol{x}_\ell^K$ and $\boldsymbol{x}_m^K$, respectively. We define the interior dihedral angle $\alpha_{\ell m}$ between F_ℓ and F_m as $\alpha_{\ell m} = \pi - \alpha_{\ell m}^*$, where $\alpha_{\ell m}^*$ is the angle between the outward normals $\boldsymbol{n}_\ell$ and $\boldsymbol{n}_m$ to facets F_ℓ and F_m. Following [6] we write $\cos(F_\ell, F_m)$ for $\cos \alpha_{\ell m}$. By $|K|$, $|F_\ell|$, and $|F_m|$ we understand the d-dimensional volume of the simplex K and the $(d-1)$-dimensional volumes of its facets F_ℓ and F_m. Further, the altitudes of the simplex K over its facets F_ℓ and F_m are denoted by η_ℓ and η_m. Clearly, $\eta_\ell = d|K|/|F_\ell|$. With this notation we can express the key integrals as follows

$$\int_K \boldsymbol{\nabla}\varphi_m^K \cdot \boldsymbol{\nabla}\varphi_\ell^K \, \mathrm{d}\boldsymbol{x} = \begin{cases} \dfrac{1}{\eta_\ell^2}|K| & \text{for } \ell = m, \\ -\dfrac{\cos(F_\ell, F_m)}{\eta_\ell \eta_m}|K| & \text{for } \ell \neq m, \end{cases} \tag{33}$$

$$\int_K \varphi_m^K \varphi_\ell^K \, \mathrm{d}\boldsymbol{x} = \frac{1 + \delta_{\ell m}}{(d+1)(d+2)}|K|,$$

where $\ell, m = 1, 2, \dots, \bar{N}_K$. Let us recall that $\varphi_\ell^K = \varphi_i$ and $\varphi_m^K = \varphi_j$ for $i = \bar{\imath}_K(\ell)$ and $j = \bar{\imath}_K(m)$, $\ell, m = 1, 2, \dots, \bar{N}_K$.

The first formula in (33) comes from [9, p. 201], see also [5]. The validity of the second formula in (33) can be readily seen from the fact that $\boldsymbol{\nabla}\varphi_\ell^K = -\boldsymbol{n}_\ell/\eta_\ell$. Its proof is published in [4, 47]. The special cases of $d \leq 3$ are well known, see e.g. [28].

Now, we can present the basic result about the DMP for problem (32). For each element $K \in \mathcal{T}_h$ and for each pair of indices $\ell \neq m$, $\ell, m = 1, 2, \dots, \bar{N}_K$, we define the following quantities

$$\lambda^K = \frac{\int_K \lambda(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}}{|K|}, \quad c_{\ell m}^K = \frac{\int_K c(\boldsymbol{x})\varphi_m^K(\boldsymbol{x})\varphi_\ell^K(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}}{\int_K \varphi_m^K(\boldsymbol{x})\varphi_\ell^K(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}}, \tag{34}$$

and

$$\alpha^K_{\ell m} = \begin{cases} \dfrac{\int_{\partial K \cap \Gamma_{\mathrm{N}}} \alpha(\boldsymbol{s})\varphi^K_m(\boldsymbol{s})\varphi^K_\ell(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}}{\int_{\partial K \cap \Gamma_{\mathrm{N}}} \varphi^K_m(\boldsymbol{s})\varphi^K_\ell(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}} & \text{if } \operatorname{meas}_{d-1}(\partial K \cap \Gamma_{\mathrm{N}}) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

In order to formulate the following lemma, we introduce further notations. Let $\gamma^K_{\ell m} = \overline{\boldsymbol{x}^K_\ell \boldsymbol{x}^K_m}$ be the edge (the line segment) between the vertices $\boldsymbol{x}^K_\ell$ and $\boldsymbol{x}^K_m$ of a simplex $K \in \mathcal{T}_h$. Let $\omega^K_{\ell m} = \{F : F \subset \partial K,\ F \subset \Gamma_{\mathrm{N}},\ \gamma^K_{\ell m} \subset F\}$ be the set of those facets of the element K that lie on Γ_{N} and share the common edge $\gamma^K_{\ell m}$. Finally, let us put $|\omega^K_{\ell m}| = \sum_{F \in \omega^K_{\ell m}} |F|$. If $\omega^K_{\ell m} = \emptyset$ we set $|\omega^K_{\ell m}| = 0$.

Lemma 8.2 *Let $K \in \mathcal{T}_h$ be a d-dimensional simplicial element. Let the local stiffness matrix A^K be given by* (24)*. Then* off-diag $A^K \le 0$ *if and only if condition*

$$\frac{c^K_{\ell m}}{(d+1)(d+2)}\eta_\ell\eta_m + \frac{\alpha^K_{\ell m}}{d(d+1)}\frac{|\omega^K_{\ell m}|}{|K|}\eta_\ell\eta_m \le \lambda^K \cos(F_\ell, F_m), \tag{36}$$

holds true for all $\ell \ne m$, $\ell, m = 1, 2, \ldots, N_K$.

Proof. From (33), (34), and (35) we directly compute all the offdiagonal entries of the local stiffness matrix:

$$\begin{aligned} A^K_{\ell m} &= \int_K \lambda \boldsymbol{\nabla}\varphi_j \cdot \boldsymbol{\nabla}\varphi_i \,\mathrm{d}\boldsymbol{x} + \int_K c\varphi_j\varphi_i \,\mathrm{d}\boldsymbol{x} + \int_{\partial K \cap \Gamma_{\mathrm{N}}} \alpha\varphi_j\varphi_i \,\mathrm{d}\boldsymbol{s} \\ &= -\lambda^K \frac{\cos(F_\ell, F_m)}{\eta_\ell\eta_m}|K| + c^K_{\ell m}\frac{1}{(d+1)(d+2)}|K| + \alpha^K_{\ell m}\frac{1}{d(d+1)}\sum_{F \in \omega^K_{\ell m}} |F| \end{aligned}$$

for all $\ell \ne m$, $\ell, m = 1, 2, \ldots, N_K$ with $i = \iota_K(\ell)$, $j = \iota_K(m)$. □

Lemma 8.3 *Let $K \in \mathcal{T}_h$ be a d-dimensional simplicial element. Let the local stiffness matrix $A^{\partial,K}$ be given by* (25)*. Then $A^{\partial,K} \le 0$ if and only if condition* (36) *holds for all $\ell = 1, 2, \ldots, N_K$ and $m = N_K + 1, N_K + 2, \ldots, N_K + N^\partial_K$.*

Proof. The proof follows the same steps as the proof of Lemma 8.2. □

Corollary 8.4 *Let us consider the lowest-order simplicial finite element discretization of problem* (32) *as described above. If the condition* (36) *is satisfied for all simplices $K \in \mathcal{T}_h$ and all indices $\ell = 1, 2, \dots, N_K$ and $m = 1, 2, \dots, N_K + N_K^\partial$ then the lowest-order simplicial finite element discretization of problem* (1) *satisfies the discrete conservation of nonnegativity.*

Proof. The statement follows immediately from Theorem 7.5 and Lemmas 8.2 and 8.3. □

Corollary 8.4 represents the main result of this section. It gives sufficient condition for the validity of the discrete conservation of nonnegativity and hence also for the validity of the DMP, see Theorem 3.2. This result generalizes the known results (see e.g. [6, 5]) in several respects. In contrast to the standard results we consider general mixed Dirichlet/Newton boundary conditions, general variable coefficient λ, and the general variable coefficient α. In addition, Lemmas 8.2 and 8.3 show both sufficient and necessary condition for the proper sign properties of the local matrices, while in the literature usually sufficient conditions only are presented.

In general, condition (36) is satisfied provided all dihedral angles are acute and the mesh is sufficiently fine. In the case of the Poisson problem with mixed Dirichlet and Neumann boundary conditions ($c = 0$, $\alpha = 0$), the crucial condition (36) reduces to

$$\cos(F_\ell, F_m) \geq 0.$$

This corresponds to the well-known requirement of nonobtuseness of all dihedral angles in the simplicial partition $\mathcal{T}_h$. If $c \neq 0$ and $\alpha = 0$ then condition (36) simplifies to the condition derived in [6]. However, here we extend its validity also for Neumann type boundary conditions.

Practically, condition (36) is very easy to verify provided the coefficients c and α are piecewise constant. Indeed, in this case the values $c_{\ell m}^K$ and $\alpha_{\ell m}^K$ coincide with the constant value of the respective coefficient for all $\ell, m = 1, 2, \dots, \bar{N}_K$. Nevertheless, in the general case of variable coefficients c and α the computation of the values $c_{\ell m}^K$ and $\alpha_{\ell m}^K$ and theire subsequent utilization in (36) might not be practical. If this is the case, we can recommend to find suitable upper bounds on c and α on each element $K \in \mathcal{T}_h$:

$$\operatorname*{ess\,sup}_{x \in K} c(x) \leq \overline{c}^K \quad \text{and} \quad \operatorname*{ess\,sup}_{s \in \partial K \cap \Gamma_{\mathrm{N}}} \alpha(s) \leq \overline{\alpha}^K$$

and use the following lemma.

Lemma 8.5 *Under the assumptions of Corollary 8.4, the lowest-order simplicial finite element discretization of problem* (1) *satisfies the discrete conservation of nonnegativity if*

$$\frac{\overline{c}^K}{(d+1)(d+2)}\eta_\ell\eta_m + \frac{\overline{\alpha}^K}{d(d+1)}\frac{|\omega^K_{\ell m}|}{|K|}\eta_\ell\eta_m \leq \lambda^K \cos(F_\ell, F_m),$$

holds true for all $\ell \neq m$, $\ell = 1, 2, \ldots, N_K$, $m = 1, 2, \ldots, N_K + N_K^\partial$.

Proof. It follows immediately from Corollary 8.4, because $c^K_{\ell m} \leq \overline{c}^K$ and $\alpha^K_{\ell m} \leq \overline{\alpha}^K$ for all $K \in \mathcal{T}_h$. □

Remark 8.6 *We observe that the validity of the DMP is connected with the dihedral angles of used simplices and hence it translates into geometric issues. As stated in [3]: if the Hadwiger conjecture is valid then any polytope in $\mathbb{R}^d$ can be partitioned into a nonobtuse simplicial mesh (all dihedral angles are at most $\pi/2$). The Hadwiger conjecture is know to be true for $d \leq 6$. Thus, at least for $d \leq 6$ and the pure diffusion problem ($c = 0$ and $\alpha = 0$) on a polytopic domain, we can always construct a simplicial finite element mesh such that the discrete maximum principle is satisfied.*

On the other hand, if c or α do not vanish then condition (36) *requires the dihedral angles to be acute in order to satisfy the discrete conservation of nonnegativity. However, division of a space (or certain polytopes) in $\mathbb{R}^d$ into acute simplices is problematic. A face-to-face simplicial partition of the space $\mathbb{R}^d$ for $d \geq 5$ does not exists [26]. Existence of such a partition in $\mathbb{R}^4$ is still an open problem. Even in $\mathbb{R}^3$ this is not a simple problem. For example, a face-to-face acute simplicial partition of a slab [14] and a cube [37] was successfully constructed quite recently.*

9 Conclusions

This contribution surveys a general concept of the DGF and the elliptic projection of the Dirichlet lift for the analysis of the DMP for the Galerkin method in general and for the finite element method in particular. We recall a successful application of this concept to the analysis of the DMP for higher-order finite elements in one-dimension. Nevertheless, this concept applies to the lowest-order finite elements as well. Simple characterization of nonnegativity of the lowest-order approximations enables to reformulate the general characterization of the DMP in terms of global and local stiffness matrices. As a particular application, we analyze the lowest-order simplicial finite elements and obtain sufficient conditions for the validity of the DMP. In contrast to the well known results, we consider a quite general diffusion-reaction problem and the conditions for the DMP include the mixed Dirichlet-Newton (Robin) boundary conditions and nonhomogeneous diffusion and reaction coefficients. In addition, we provide sufficient and necessary conditions for the corresponding local stiffness matrices to be M-matrices.

The described general concept can be used in other variants of the finite element method as well. For example, the case of block finite elements can be analyzed in this way. For blocks, however, the conditions for the validity of the DMP are specific for particular dimensions and we cannot expect a universal condition as for simplices.

Acknowledgements

The author gratefully acknowledges the support of grant no. IAA100760702 of the Grant Agency of the Academy of Sciences of the Czech Republic, and of the institutional research plan no. AV0Z10190503 of the Academy of Sciences of the Czech Republic.

References

[1] Bramble J.H., Hubbard B.E., New monotone type approximations for elliptic problems, Math. Comp., 1964, 18, 349–367

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

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

[4] Brandts J.H., Korotov S., Křížek M., Dissection of the path-simplex in $\mathbb{R}^n$ into n path-subsimplices, Linear Algebra Appl., 2007, 421, 382–393

[5] Brandts J.H., Korotov S., Křížek M., Simplicial finite elements in higher dimensions, Appl. Math., 2007, 52, 251–265

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

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

[8] Ciarlet P.G., Discrete variational Green's function. I., Aequationes Math., 1970, 4, 74–82

[9] Ciarlet P.G., The finite element method for elliptic problems, North-Holland Publishing Co., Amsterdam, 1978

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

[11] Ciarlet P.G., Varga R.S., Discrete variational Green's function. II. One dimensional problem, Numer. Math., 1970, 16, 115–128

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

[13] Duffy D.G., Green's functions with applications, Chapman & Hall/CRC, Boca Raton, FL, 2001

[14] Eppstein D., Sullivan J.M., Üngör A., Tiling space and slabs with acute tetrahedra, Comput. Geom., 2004, 27, 237–255

[15] Faragó I., Horváth R., Discrete maximum principle and adequate discretizations of linear parabolic problems, SIAM J. Sci. Comput., 2006, 28, 2313–2336

[16] Faragó I., Horváth R., A review of reliable numerical models for three-dimensional linear parabolic problems, Internat. J. Numer. Methods Engrg., 2007, 70, 25–45

[17] Faragó I., Horváth R., Korotov S., Discrete maximum principle for linear parabolic problems solved on hybrid meshes, Appl. Numer. Math., 2005, 53, 249–264

[18] Faragó I., Korotov S., Szabó T., On modifications of continuous and discrete maximum principles for reaction-diffusion problems, Adv. Appl. Math. Mech., 2011, 3, 109–120.

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

[20] Fujii H., Some remarks on finite element analysis of time-dependent field problems, In: Theory and Practice in Finite Element Structural Analysis, Univ. Tokyo Press, 1973, 91–106

[21] Gilbarg D., Trudinger N.S., Elliptic partial differential equations of second order, Springer-Verlag, Berlin, 1977

[22] Glowinski R., Numerical methods for nonlinear variational problems, Springer-Verlag, New York, 1984

[23] Ikeda T., Maximum principle in finite element models for convection-diffusion phenomena, Kinokuniya Book Store Co., Ltd., Tokyo, 1983

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

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

[26] Křížek M., There is no face-to-face partition of $\mathbf{R}^5$ into acute simplices, Discrete Comput. Geom., 2006, 36, 381–390

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

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

[29] Kuzmin D., Shashkov M.J., Svyatskiy D., A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems, J. Comput. Phys., 2009, 228, 3448–3463

[30] Nečas J., Les méthodes directes en théorie des équations elliptiques, Masson et Cie, Éditeurs, Paris, 1967

[31] Protter M.H., Weinberger H.F., Maximum principles in differential equations, Prentice-Hall Inc., Englewood Cliffs, N.J., 1967

[32] Roos H.G., M. Stynes, L. Tobiska, Robust Numerical Methods for Singularly Perturbed Differential Equations, 2nd ed., Springer-Verlag, Berlin, 2008

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

[34] Šolín, P., Segeth, K., Doležel, I., Higher-order finite element methods, Chapman & Hall/CRC, Boca Raton, FL, 2004

[35] Stakgold I., Green's functions and boundary value problems, 2nd ed., Wiley, New York, 1998

[36] Szabó B., Babuška I., Finite element analysis, Wiley, New York, 1991

[37] VanderZee E., Hirani A.N., Zharnitsky V., Guoy D., A dihedral acute triangulation of the cube, Comput. Geom., 2010, 43, 445–452

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

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

[40] Varga R.S., On a discrete maximum principle, SIAM J. Numer. Anal., 1966, 3, 355–359

[41] Vejchodský T., Angle conditions for discrete maximum principles in higher-order FEM, In: Kreiss G., Lötstedt P., Målqvist A., Neytcheva M. (Eds.), Numerical mathematics and advanced applications ENUMATH 2009, Springer, 2010, 901–909

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

[43] Vejchodský T., Šolín P., Discrete Green's function and maximum principles, In: Chleboun J., Segeth K., Vejchodský T. (Eds.), Programs and Algorithms of Numerical Mathematics 13, Institute of Mathematics, Academy of Sciences, Czech Republic, 2006, 247–252

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

[45] Vejchodský T., Šolín P., Discrete maximum principle for higher-order finite elements in 1D, Math. Comp., 2007, 76, 1833–1846

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

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