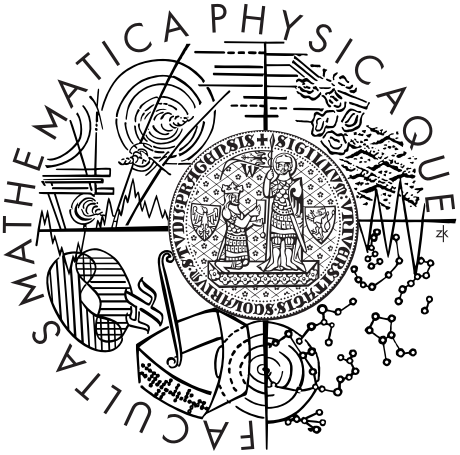
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

HABILITAČNÍ PRÁCE



Tomáš Vejchodský

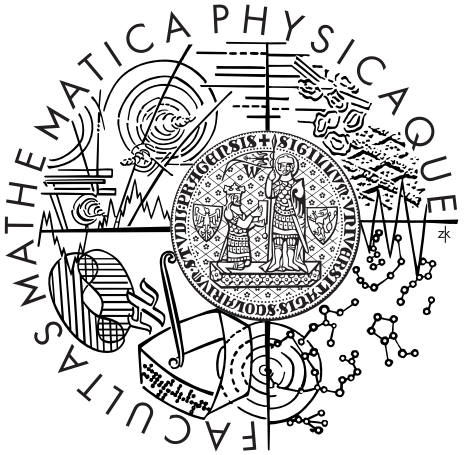
DISKRÉTNÍ PRINCIPY MAXIMA

Obor: matematika – přibližné a numerické metody

Praha 2011

Charles University in Prague
Faculty of Mathematics and Physics

HABILITATION THESIS



Tomáš Vejchodský

DISCRETE MAXIMUM PRINCIPLES

Branch: mathematics – approximate and numerical methods

Prague 2011

In spite of the fact I am the sole author of this thesis, many people have more or less indirect share on its appearance.

I wish to express my gratitude to my colleagues, co-authors, and friends Pavel Šolín, Sergey Korotov, Antti Hannukainen, Jan Brandts, Karel Segeth, and Michal Křížek for their willingness to work with me, for the inspirations, and clever ideas they share with me. Special thanks belong to my former teacher and Ph.D. supervisor Michal Křížek for his kind and endless support of all kinds. He and Karel Segeth carefully read and corrected this thesis in the last stage of its creation. Michal Křížek also organizes regular seminars "Current Problems in Numerical Analysis", where the results of this thesis were presented, discussed, and revised. In addition, I wish to thank to my other colleagues for inspiring discussions, interesting suggestions, and for creating friendly and motivating atmosphere in the Institute of Mathematics.

I appreciate the support of the Czech Science Foundation and the Grant Agency of the Academy of Sciences, because most of the presented original results were obtained during the solution of projects no. 102/07/0496 and IAA100760702.

Last but not least, I am grateful to my wife Eliška for her constant support.

Praha, August 25, 2011
Tomáš Vejchodský

Název práce: Diskrétní principy maxima

Autor: Tomáš Vejchodský

Matematický ústav AV ČR, v.v.i.

Abstrakt: Tato práce shrnuje současné znalosti z oblasti diskrétních principů maxima pro lineární eliptické parciální diferenciální rovnice řešené metodou konečných prvků. Zabývá se jak standardními metodami nejnižšího řádu, tak metodami vyšších řádů přesnosti. Cílem bylo podat ucelený přehled o této problematice. Důraz byl kladen na jednotný styl. Práce je budována hierarchicky. Nejdříve je zavedena jednotná obecná teorie, ze které se následně odvozují specifické výsledky. Kromě přehledu známých výsledků práce obsahuje také řadu nových zobecnění a několik původních výsledků autora.

Klíčová slova: eliptické parciální differenciální rovnice, metoda konečných prvků, diskrétní princip maxima

Title: Discrete maximum principles

Author: Tomáš Vejchodský

Institute of Mathematics, Academy of Sciences of the Czech Republic

Abstract: This thesis surveys the current knowledge from the field of the discrete maximum principles for linear elliptic partial differential equations solved by the finite element method. It deals with the standard lowest-order methods as well as with the methods of higher order of accuracy. The aim is to give a comprehensive overview of this topic. Emphasis is placed on a unified style. The thesis has a hierarchical structure. First, a general theory is introduced and subsequently specific results are derived. In addition to the survey of known results, the thesis also contains many new generalizations and several original results of the author.

Keywords: elliptic partial differential equations, finite element method, discrete maximum principle

CONTENTS

CHAPTER

ONE

Introduction

The maximum principle is a remarkable feature of various differential equations and inequalities. We can derive it for linear as well as for certain nonlinear problems. It is best known and analyzed in the context of the second order partial differential equations of elliptic and parabolic types. However, it is not only a mathematical statement, the maximum principle reflects in the nature through certain properties of specific physical fields. For example, quantities like temperature, concentration, pressure, density, etc. possess naturally nonnegative values. Interestingly, thanks to the validity of the maximum principle, solutions of the corresponding mathematical models possess nonnegative values, too.

The maximum principle is known and studied from the very beginning of the development of the differential equations. It was already known to Gauss in 1839. Modern studies of the maximum principle were initiated by the pioneering work of Eberhard Hopf [40] in 1927. Monographs [65] and [35] from the second half of the 20th century are now considered as classical. From the recent texts on this topic, we can recommend [33] and [66].

Concerning the numerics, it is natural to attempt a construction of such numerical methods that reflect the natural property of the maximum principle even on the discrete level. Thus, we speak about the *discrete maximum principle* (DMP) or equivalently about a *monotone* numerical method. The first numerical method, where the DMP was studied, was the finite difference method. Papers [4, 5, 15, 16, 79], etc. present the first DMP results in this context. These results were later generalized to the finite element method, see for example [18, 22, 70, 77], etc.

The proofs of the DMP are based on monotone matrices and mostly on the theory of M-matrices. Monograph [78] is fundamental and pioneering in this field. However, the more modern books [3, 32] can be recommended as well.

The goal of this thesis is to present a more or less complete survey of the DMP

results for the linear second-order elliptic partial differential equations discretized by the lowest-order and the higher-order finite element methods. The DMP for the *lowest-order* finite elements is quite well understood, it is studied for several decades and many results have been already published. Therefore, the aim of this thesis is to survey these results and present them in a unified way. Nevertheless, the style of the presentation, the general theoretical framework, as well as several results are original.

On the other hand, the situation for the *higher-order* finite elements is much less clear. The literature is scarce and the results are mostly negative [39]. However, at least for simple problems certain positive results are possible and this thesis presents mostly the author's original contributions to this field.

Certainly, the DMP for linear elliptic problems discretized by the finite element method is not the only topic of interest. There exists a variety of results for nonlinear problems [43, 44, 45], etc., and for the parabolic problems [27, 28, 29, 30, 31, 34, 41, 75, 81, 82], etc. Other numerical methods like the finite differences [4, 5, 15, 16, 79], the finite volumes [11, 62], mixed methods [26, 60], the collocation methods [85], etc. are analyzed as well. There are also approaches, where special methods are constructed such that the resulting numerical scheme always yields the DMP, see e.g. [13, 84]. In addition, the validity of the DMP is connected to the stability of the finite element method and to the L^∞ error estimates [2, 70]. The problem of the DMP can also be handled by the theory of reproducing kernels [1]. However, in this thesis we will not address these topics and we will concentrate on the linear elliptic problems discretized by the finite element method only.

The thesis is organized as follows. After this introductory chapter we proceed with Chapter 2, where we briefly discuss the maximum principle for linear second-order elliptic problems. We define the problem, state the maximum principle, its equivalent variants, and introduce the Green's function. Chapter 3 is devoted to the general theory of the DMP in the finite element method. First, the finite element method is recalled, the DMP and the discrete Green's function (DGF) are defined, and the relationship between the DMP and the DGF is proved. A special attention is paid to the treatment of nonhomogeneous Dirichlet boundary conditions. Chapters 4 and 5 form the core parts of the thesis. Chapter 4 surveys the DMP results for the lowest-order finite elements. We introduce the fundamental theoretical framework and treat the one-dimensional case separately. From the variety of possible geometric types of finite elements in higher dimension we treat the simplicial elements, the block finite elements (Cartesian products of intervals), and the right-triangular prisms. Chapter 5 concerns the higher-order finite elements. We present several theoretical one-dimensional results and a variety of two-dimensional numerical experiments. Most of the original results are reproduced from publications of the author and his co-authors. The most relevant

author's papers are attached as Appendices B–I, see also the list at the very end of the thesis. Herewith I declare that my share in these publications and the share of my co-authors is approximately equal.

1.1 List of notation

Within the thesis we use the standard mathematical notation. In order to increase the readability we denote the vectors in bold. For the reader's convenience we present a list of used mathematical symbols which are not explained within the thesis.

$\mathbb{R}$	set of real numbers
$\mathbb{R}^d$	d-dimensional Euclidean space
$\subset$	subset (or subspace)
$\cup$	union
$\cap$	intersection
$\overline{\Omega}$	closure of the set $\Omega \subset \mathbb{R}^d$
$\lvert x \rvert$	absolute value
$\boldsymbol{a} \cdot \boldsymbol{b}$	Euclidean scalar (dot) product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$
$A^\top$	transposed matrix (or vector)
$\Rightarrow$	implication
$\Leftrightarrow$	equivalence
div	divergence, $\operatorname{div} \boldsymbol{q} = \partial q_1/\partial x_1 + \cdots + \partial q_d/\partial x_d$
$\boldsymbol{\nabla}$	gradient, $\boldsymbol{\nabla} u = (\partial u/\partial x_1, \ldots, \partial u/\partial x_d)^\top$
Δ	Laplace operator, $\Delta u = \partial^2 u/\partial x_1^2 + \cdots + \partial^2 u/\partial x_d^2$
meas_d	d-dimensional Lebesgue measure (often abbreviated as meas)
a.a.	almost all (up to a set of zero measure)
a.e.	almost everywhere (up to a set of zero measure)
ess sup	essential supremum, $\operatorname{ess\,sup} u = \inf \{\varrho \in \mathbb{R} : \operatorname{meas}\{\boldsymbol{x} : u(\boldsymbol{x}) > \varrho\} = 0\}$
ess inf	essential infimum, $\operatorname{ess\,inf} u = \sup \{\varrho \in \mathbb{R} : \operatorname{meas}\{\boldsymbol{x} : u(\boldsymbol{x}) < \varrho\} = 0\}$
C	positive generic constant (it may have different values in different occurrences)

CHAPTER

TWO

Maximum principles for elliptic problems

The main purpose of this chapter is to present the studied elliptic problem and to introduce the notation, essential hypotheses, and definitions. Naturally, we stress the maximum principle. We point out its variants and relationships among them.

First, we formulate the general diffusion-convection-reaction linear elliptic problem with mixed boundary conditions of Dirichlet, Neumann, and Newton (often called Robin) type. We present the classical (strong) formulation of this problem and then we concentrate on its weak formulation. The emphasis on the weak formulation is clearly motivated by the subsequent discretization by the finite element method.

We have made an attempt to explicitly formulate all assumptions needed for the well posedness of the weak formulation. The reason is the self-consistency of the text and mainly the fact that these assumptions are also fundamental for the maximum principle.

Subsequently, we present the maximum principle and its variants – the minimum principle, the conservation of nonnegativity, and the comparison principle. The equivalence of all these variants is proved. Although the maximum principles in the classical and in the weak form are essentially the same, we present these two forms separately. The reason is the fundamental difference in the proofs of the maximum principle.

This chapter is concluded by a section devoted to the Green's function. We show the fundamental statement that the nonnegativity of the Green's function is equivalent with the validity of the maximum principle. A discrete analogy of this statement is a core result we build upon in the subsequent chapters, where the discrete maximum principle is analyzed.

The results presented in this chapter are mostly well known and they can be found in many textbooks. We used mainly the monographs [53, 54, 61] as a

source for the elliptic problems, the books [35, 65] for the maximum principles, and publications [19, 37, 55, 59, 73] for the Green's function.

2.1 Linear second-order elliptic problems

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

$$-\operatorname{div}(\mathcal{A}\boldsymbol{\nabla} u) + \boldsymbol{b} \cdot \boldsymbol{\nabla} u + cu = f \quad \text{in } \Omega, \tag{2.1}$$

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

$$\alpha u + (\mathcal{A}\boldsymbol{\nabla} u) \cdot \boldsymbol{n} = g_{\mathrm{N}} \quad \text{on } \Gamma_{\mathrm{N}}, \tag{2.3}$$

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 $\overline{\Gamma}_{\mathrm{D}} \cup \overline{\Gamma}_{\mathrm{N}} = \partial\Omega$. We remind that a subset $\Gamma \subset \partial\Omega$ is relatively open if for every $\boldsymbol{x} \in \Gamma$ there exists a ball B such that $\boldsymbol{x} \in B$ and $B \cap \partial\Omega \subset \Gamma$. The sets Γ_{D} and Γ_{N} are assumed to have a finite number of components and Lipschitz boundary relative to $\partial\Omega$. The matrix of diffusivities $\mathcal{A}(\boldsymbol{x}) \in \mathbb{R}^{d\times d}$, the vector of convection $\boldsymbol{b}(\boldsymbol{x}) \in \mathbb{R}^d$, the scalar reaction coefficient $c(\boldsymbol{x}) \in \mathbb{R}$, 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 $\alpha(\boldsymbol{s}) \in \mathbb{R}$, $g_{\mathrm{N}}(\boldsymbol{s}) \in \mathbb{R}$, $\mathcal{A}(\boldsymbol{s})$, $\boldsymbol{b}(\boldsymbol{s})$ are functions of $\boldsymbol{s} \in \Gamma_{\mathrm{N}}$. Further, we assume that

$$c - \frac{1}{2}\operatorname{div}\boldsymbol{b} \geq 0 \quad \text{in } \Omega \quad \text{and} \quad \alpha + \frac{1}{2}\boldsymbol{b}\cdot\boldsymbol{n} \geq 0 \quad \text{on } \Gamma_{\mathrm{N}} \tag{2.4}$$

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

$$(\mathcal{A}(\boldsymbol{x})\boldsymbol{\xi}) \cdot \boldsymbol{\xi} \geq \lambda_{\min}|\boldsymbol{\xi}|^2, \quad \forall \boldsymbol{\xi} \in \mathbb{R}^d,\ \forall \boldsymbol{x} \in \Omega, \tag{2.5}$$

where $\lambda_{\min} > 0$ and $|\boldsymbol{\xi}| = (\boldsymbol{\xi} \cdot \boldsymbol{\xi})^{1/2}$ stands for the Euclidean norm of $\boldsymbol{\xi} \in \mathbb{R}^d$. In general, we consider matrix $\mathcal{A}$ as nonsymmetric. Problem (2.1)–(2.3) is 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.

2.2 Weak solution

The weak formulation of problem (2.1)–(2.3) is naturally stated in the framework of the Sobolev space $H^1(\Omega)$ of square integrable functions whose distributional derivatives are square integrable as well. The norm in $H^1(\Omega)$ is denoted by $\|\cdot\|_{1,\Omega}$.

Similarly the norm in the Lebesgue space $L^2(\Omega)$ is denoted by $\|\cdot\|_{0,\Omega}$. Sometimes, if there is no danger of confusion, we write simply $\|\cdot\|_1$ or $\|\cdot\|_0$ for these norms.

Before we introduce the weak formulation, let us recall the fundamental properties of functions from the Sobolev space $H^1(\Omega)$. First, the trace theorem states that there exists unique linear continuous operator $\gamma : H^1(\Omega) \mapsto L^2(\partial\Omega)$ such that $\gamma z = z|_{\partial\Omega}$ for all $z \in C^\infty(\overline{\Omega})$. The function $\gamma v \in L^2(\partial\Omega)$ is called a *trace* of $v \in H^1(\Omega)$. For simplicity, we will write v instead of γv in what follows. In fact, the trace theorem states that there exists a constant $C > 0$ such that

$$\|v\|_{0,\partial\Omega} \leq C \|v\|_{1,\Omega} \quad \forall v \in H^1(\Omega). \tag{2.6}$$

The second fundamental property is the following Friedrichs' inequality

$$\|v\|^2_{1,\Omega} \leq C \left(\|\boldsymbol{\nabla} v\|^2_{0,\Omega} + \|v\|^2_{0,\Gamma} \right) \quad \forall v \in H^1(\Omega), \tag{2.7}$$

where C is a positive constant and $\Gamma \neq \emptyset$ is a relatively open subset of $\partial\Omega$. The final fundamental property we will need is the following variant of the Friedrichs' inequality

$$\|v\|^2_{1,\Omega} \leq C \left(\|\boldsymbol{\nabla} v\|^2_{0,\Omega} + \|v\|^2_{0,B} \right) \quad \forall v \in H^1(\Omega), \tag{2.8}$$

where $B \subset \Omega$ is a ball. Notice that we do not consider the empty set as a ball. The proofs of these properties can be found for example in [20] and [61].

In order to introduce the weak formulation of problem (2.1)–(2.3), 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 of g_D, i.e. let $\widetilde{g}_\mathrm{D} \in H^1(\Omega)$ be a function with traces on Γ_D equal to g_D. Without a danger of confusion we denote the Dirichlet lift $\widetilde{g}_\mathrm{D}$ also by g_D. Further, we assume conditions (2.4) to be satisfied a.e. in Ω and a.e. on Γ_N, respectively, and the uniform positive definiteness (2.5) for a.a. $\boldsymbol{x} \in \Omega$. Finally, let

$$V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_\mathrm{D} \text{ in the sense of traces}\}. \tag{2.9}$$

We say that $u \in H^1(\Omega)$ is a weak solution of (2.1)–(2.3) if $u = u^0 + g_\mathrm{D}$, where $u^0 \in V$ and

$$a(u^0, v) = \mathcal{F}(v) - a(g_\mathrm{D}, v) \quad \forall v \in V, \tag{2.10}$$

where

$$a(u, v) = \int_\Omega [(\mathcal{A}\boldsymbol{\nabla} u) \cdot \boldsymbol{\nabla} v + (\boldsymbol{b} \cdot \boldsymbol{\nabla} u)v + cuv] \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_\mathrm{N}} \alpha uv \,\mathrm{d}\boldsymbol{s}, \tag{2.11}$$

$$\mathcal{F}(v) = \int_\Omega fv \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_\mathrm{N}} g_\mathrm{N} v \,\mathrm{d}\boldsymbol{s}. \tag{2.12}$$

It is easy to verify that the boundedness of the coefficients $\mathcal{A}$, $\boldsymbol{b}$, c and the trace theorem (2.6) imply the continuity of the bilinear form a, i.e. there exists a contant $C > 0$ such that

$$|a(u,v)| \leq C \|u\|_{1,\Omega} \|v\|_{1,\Omega} \quad \forall u, v \in V. \tag{2.13}$$

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

$$a(v,v) \geq C \|v\|^2_{1,\Omega} \quad \forall v \in V. \tag{2.14}$$

The following lemma proves this V-ellipticity.

Lemma 2.1. *Let the matrix $\mathcal{A}$ be uniformly positive definite for a.a. $\boldsymbol{x} \in \Omega$, see (2.5), let coefficients $\boldsymbol{b}$, c, and α satisfy conditions (2.4) a.e. in Ω and a.e. on Γ_{N}, respectively, and let at least one of the following conditions be satisfied:*

(a) *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 $c - \frac{1}{2}\operatorname{div}\boldsymbol{b} \geq c_0 > 0$ a.e. in B,*

(c) *there exists a constant α_0 and a relatively open subset Γ^0_{N} of Γ_{N} such that $\alpha + \frac{1}{2}\boldsymbol{b}\cdot\boldsymbol{n} \geq \alpha_0 > 0$ a.e. on Γ^0_{N}.*

Then the bilinear form $a(\cdot,\cdot)$ is V-elliptic.

Proof. Let $v \in V$ be arbitrary. Since $2v\boldsymbol{\nabla}v = \boldsymbol{\nabla}(v^2)$, we can use the Green's theorem to obtain

$$\int_\Omega (\boldsymbol{b}\cdot\boldsymbol{\nabla}v)v \,\mathrm{d}\boldsymbol{x} = \frac{1}{2}\int_\Omega \boldsymbol{b}\cdot\boldsymbol{\nabla}(v^2)\,\mathrm{d}\boldsymbol{x} = -\frac{1}{2}\int_\Omega (\operatorname{div}\boldsymbol{b})v^2\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\Gamma_{\mathrm{N}}} \boldsymbol{b}\cdot\boldsymbol{n}v^2\,\mathrm{d}\boldsymbol{s}.$$

Consequently,

$$\begin{aligned} a(v,v) &= \int_\Omega \left[(\mathcal{A}\boldsymbol{\nabla}v)\cdot\boldsymbol{\nabla}v + \left(c - \frac{1}{2}\operatorname{div}\boldsymbol{b}\right)v^2\right]\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}}\left(\alpha + \frac{1}{2}\boldsymbol{b}\cdot\boldsymbol{n}\right)v^2\,\mathrm{d}\boldsymbol{s} \\ &\geq \lambda_{\min}\|\boldsymbol{\nabla}v\|^2_{0,\Omega} + \int_\Omega \left(c - \frac{1}{2}\operatorname{div}\boldsymbol{b}\right)v^2\,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}}\left(\alpha + \frac{1}{2}\boldsymbol{b}\cdot\boldsymbol{n}\right)v^2\,\mathrm{d}\boldsymbol{s}, \end{aligned} \tag{2.15}$$

where we use the uniform positive definiteness (2.5) of $\mathcal{A}$. Conditions (a)–(c), estimate (2.15), and the nonnegativity of all the three terms on the right-hand side of (2.15) can be used in the following way to prove the V-ellipticity of $a(\cdot,\cdot)$.

If condition (a) is satisfied, we can use the Friedrichs' inequality (2.7) with $\Gamma = \Gamma_{\mathrm{D}}$ to obtain

$$a(v,v) \geq \lambda_{\min} \|\boldsymbol{\nabla} v\|^2_{0,\Omega} \geq C \|v\|^2_{1,\Omega}\,.$$

If condition (b) is satisfied, we can use the inequality (2.8) to derive

$$a(v,v) \geq \lambda_{\min} \|\boldsymbol{\nabla} v\|^2_{0,\Omega} + c_0 \left\|v^2\right\|_{0,B} \geq C \|v\|^2_{1,\Omega}\,.$$

Finally, if condition (c) is satisfied, then inequality (2.7) with $\Gamma = \Gamma^0_{\mathrm{N}}$ yields

$$a(v,v) \geq \lambda_{\min} \|\boldsymbol{\nabla} v\|^2_{0,\Omega} + \alpha_0 \left\|v^2\right\|_{0,\Gamma^0_{\mathrm{N}}} \geq C \|v\|^2_{1,\Omega}\,.$$

□

Let us note that the continuity (2.13), the V-ellipticity (2.14) of the bilinear form $a(\cdot,\cdot)$, and the continuity of the linear functional $\mathcal{F}(\cdot)$ guarantee the existence of a unique weak solution to problem (2.10). This weak solution is independent of the particular choice of the Dirichlet lift g_{D}.

2.3 Maximum principles in the classical sense

In this section we consider the classical formulation (2.1)–(2.3). Throughout this section we assume $f \in C(\Omega)$, $g_{\mathrm{D}} \in C(\Gamma_{\mathrm{D}})$, $g_{\mathrm{N}} \in C(\Gamma_{\mathrm{N}})$, and for the corresponding classical solution we consider $u \in C^1(\overline{\Omega}) \cap C^2(\Omega)$. We will also use the positive and negative parts of a real function v, i.e. we define $v^+ = (|v| + v)/2$ and $v^- = (|v| - v)/2$. Clearly, $v^+ \geq 0$, $v^- \geq 0$, $v = v^+ - v^-$, and $v^+ = \max\{0, v\}$, $v^- = -\min\{0, v\}$.

Definition 2.1. Problem (2.1)–(2.3) satisfies the maximum principle if

$$f \leq 0 \text{ and } g_{\mathrm{N}} \leq 0 \quad \Rightarrow \quad \max_{\overline{\Omega}} u \leq \max_{\Gamma_{\mathrm{D}}} u^+.$$

Definition 2.2. Problem (2.1)–(2.3) satisfies the minimum principle if

$$f \geq 0 \text{ and } g_{\mathrm{N}} \geq 0 \quad \Rightarrow \quad \min_{\overline{\Omega}} u \geq \min_{\Gamma_{\mathrm{D}}} (-u^-).$$

Definition 2.3. Problem (2.1)–(2.3) conserves nonnegativity if

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

Definition 2.4. Let u_1 be the solution to problem (2.1)–(2.3) with right-hand side f_1 and boundary data $g_{\mathrm{D},1}$, $g_{\mathrm{N},1}$ and similarly let u_2 be the solution for f_2, $g_{\mathrm{D},2}$, and $g_{\mathrm{N},2}$. We say that problem (2.1)–(2.3) satisfies the comparison principle if

$$f_1 \geq f_2, \ g_{\mathrm{D},1} \geq g_{\mathrm{D},2}, \text{ and } g_{\mathrm{N},1} \geq g_{\mathrm{N},2} \quad \Rightarrow \quad u_1 \geq u_2.$$

Theorem 2.2. *Let the coefficients c and α be nonnegative. Then the following statements are equivalent*

(i) *Problem (2.1)–(2.3) satisfies the maximum principle.*
(ii) *Problem (2.1)–(2.3) satisfies the minimum principle.*
(iii) *Problem (2.1)–(2.3) conserves nonnegativity.*
(iv) *Problem (2.1)–(2.3) satisfies the comparison principle.*

Proof. The proof is essentially the same as the proof of Theorem 2.4 below which shows the equivalence of these principles in the weak sense. □

Theorem 2.3. *Let the coefficients c and α be nonnegative. Then problem (2.1)–(2.3) satisfies the maximum principle.*

The standard proof of the maximum principle for elliptic problems is based on the fact that $u \in C^2(\Omega)$ and it can be found at many places in both linear and nonlinear settings, see e.g. [35, 57, 65, 66].

2.4 Maximum principles in the weak sense

The situation with the maximum (and the other) principles in the weak setting is essentially the same as in the classical setting. The differences are of the technical character only. In the definitions we have to take into account the fact that the data are defined up to a set of zero measure. The proof of the equivalence of the considered principles is practically the same in both settings. However, the proofs of the maximum principles themselves differ. The standard proofs in the classical sense utilize substantially the C^2 continuity of the solution. This cannot be done in the weak setting. Therefore, we present at the end of this section a proof of the maximum principle for the weak solution. Similar proofs are given e.g. in [44, 51, 52, 80], but not for the general linear elliptic problem (2.10). In the following definitions we assume that $u \in H^1(\Omega)$ is a solution of problem (2.10) corresponding to $f \in L^2(\Omega)$, $g_{\mathrm{D}} \in L^2(\Gamma_{\mathrm{D}})$, and $g_{\mathrm{N}} \in L^2(\Gamma_{\mathrm{N}})$.

Definition 2.5. Problem (2.10) satisfies the maximum principle if

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

Definition 2.6. Problem (2.10) satisfies the minimum principle if

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

Definition 2.7. Problem (2.10) conserves nonnegativity if

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

Definition 2.8. Let $u_1 \in H^1(\Omega)$ be the solution to problem (2.10) with right-hand side f_1 and boundary data $g_{\mathrm{D},1}$, $g_{\mathrm{N},1}$ and similarly let $u_2 \in H^1(\Omega)$ be the solution for f_2, $g_{\mathrm{D},2}$, and $g_{\mathrm{N},2}$. We say that problem (2.10) satisfies the comparison principle if

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

Theorem 2.4. *Let $c \geq 0$ a.e. in Ω and $\alpha \geq 0$ a.e. on Γ_{N}. Then the following statements are equivalent*

(i) *Problem (2.10) satisfies the maximum principle.*
(ii) *Problem (2.10) satisfies the minimum principle.*
(iii) *Problem (2.10) conserves nonnegativity.*
(iv) *Problem (2.10) satisfies the comparison principle.*

Proof. First, we prove the implication (i) $\Rightarrow$ (ii). Let $f \geq 0$ a.e. in Ω and let $g_{\mathrm{N}} \geq 0$ a.e. on Γ_{N}. If $u \in H^1(\Omega)$ is the weak solution corresponding to f, g_{N}, and g_{D} then $-u$ is the weak solution of the same problem with $-f$, $-g_{\mathrm{N}}$, and $-g_{\mathrm{D}}$. Using the maximum principle with $-f \leq 0$ and $-g_{\mathrm{N}} \leq 0$, we conclude

$$\operatorname*{ess\,inf}_{\overline{\Omega}} u = -\operatorname*{ess\,sup}_{\overline{\Omega}}(-u) \geq -\operatorname*{ess\,sup}_{\Gamma_{\mathrm{D}}}(-u)^+ = \operatorname*{ess\,inf}_{\Gamma_{\mathrm{D}}}(-u^-).$$

In order to prove (ii) $\Rightarrow$ (iii), we consider the solution $u \in H^1(\Omega)$ of (2.10) corresponding to $f \geq 0$ a.e. in Ω, $g_{\mathrm{D}} \geq 0$ a.e. on Γ_{D}, and $g_{\mathrm{N}} \geq 0$ a.e. on Γ_{N}. The minimum principle immediately implies that

$$\operatorname*{ess\,inf}_{\overline{\Omega}} u \geq \operatorname*{ess\,inf}_{\Gamma_{\mathrm{D}}}(-u^-) = \operatorname*{ess\,inf}_{\Gamma_{\mathrm{D}}}(-g_{\mathrm{D}}^-) = 0.$$

The implication (iii) $\Rightarrow$ (iv) readily follows from the linearity of the problem. Indeed, considering u_i, f_i, $g_{\mathrm{D},i}$, and $g_{\mathrm{N},i}$, $i = 1, 2$, as Definition 2.8 requires, we observe that $u = u_1 - u_2$ is a solution of (2.10) corresponding to $f = f_1 - f_2$,

$g_{\mathrm{D}} = g_{\mathrm{D},1} - g_{\mathrm{D},2}$, and $g_{\mathrm{N}} = g_{\mathrm{N},1} - g_{\mathrm{N},2}$. Since $f \geq 0$ a.e. in Ω, $g_{\mathrm{D}} \geq 0$ a.e. on Γ_{D}, and $g_{\mathrm{N}} \geq 0$ a.e. on Γ_{N}, the conservation of nonnegativity implies $u \geq 0$ a.e. in Ω.

Finally, we prove the implication (iv) $\Rightarrow$ (i). Let $u \in H^1(\Omega)$ be the solution of (2.10) corresponding to $f \leq 0$ a.e. in Ω, $g_{\mathrm{N}} \leq 0$ a.e. on Γ_{N}, and a certain g_{D}, i.e. we have $u = g_{\mathrm{D}}$ a.e. on Γ_{D}. Let us set $\overline{g} = \operatorname{ess\,sup}_{\Gamma_{\mathrm{D}}} u = \operatorname{ess\,sup}_{\Gamma_{\mathrm{D}}} g_{\mathrm{D}}$. Now we distinguish two cases. First, if $\overline{g} \leq 0$ then necessarily $g_{\mathrm{D}} \leq 0$ a.e. on Γ_{D} and we use the comparison principle to obtain $u \leq 0$ a.e. in Ω. Thus, $u \leq 0 = \operatorname{ess\,sup}_{\Gamma_{\mathrm{D}}} u^+$ a.e. in Ω, which is the statement of the maximum principle. Second, if $\overline{g} > 0$ then we consider the constant solution $u_1 = \overline{g}$ of (2.10) with $f_1 = c\overline{g}$, $g_{\mathrm{D},1} = \overline{g}$, and $g_{\mathrm{N},1} = \alpha\overline{g}$. Since $f_1 \geq 0 \geq f$ a.e. in Ω, $g_{\mathrm{D},1} \geq g_{\mathrm{D}}$ a.e. on Γ_{D}, and $g_{\mathrm{N},1} \geq 0 \geq g_{\mathrm{N}}$ a.e. on Γ_{N}, we apply the comparison principle to conclude that $u_1 \geq u$ a.e. in Ω. Hence, $\operatorname{ess\,sup}_{\Gamma_{\mathrm{D}}} u^+ = \overline{g} = u_1 \geq u$ a.e. in Ω, which is again the statement of the maximum principle. □

Remark 2.1. If the assumptions $c \geq 0$ a.e. in Ω and $\alpha \geq 0$ a.e. on Γ_{N} in Theorem 2.4 are not satisfied, then we can still easily prove equivalences (i) $\Leftrightarrow$ (ii) and (iii) $\Leftrightarrow$ (iv) as well as the implication (ii) $\Rightarrow$ (iii). We can actually use the above proof. However, if the coefficients satisfy (2.4) only and if they are not nonnegative then the converse implication (ii) $\Leftarrow$ (iii) is, in general, not valid.

This remark can be stated also in the classical setting – see Theorem 2.2 – and also on the discrete level – see Theorem 3.1 below.

Theorem 2.5. *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 (2.14). Then problem (2.10) satisfies the maximum principle.*

Proof. Let us consider problem (2.10) 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 $\overline{g} = \operatorname{ess\,sup}_{\Gamma_{\mathrm{D}}} u^+$ and $v(\boldsymbol{x}) = (u(\boldsymbol{x}) - \overline{g})^+$. Since the positive part w^+ is a continuous mapping from $H^1(\Omega)$ into itself, see e.g. [36, p. 29], the function v lies in $H^1(\Omega)$. Further, clearly, $\overline{g} \geq 0$, $v \geq 0$ a.e. in Ω, $v = 0$ on Γ_{D} in the sense of traces, and $u = v + \overline{g}$ whenever v does not vanish. These facts together with assumptions (2.4) and with the V-ellipticity of $a(\cdot,\cdot)$ enable us to estimate

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

Hence $v = 0$ a.e. in Ω and thus $u \leq \overline{g}$ a.e. in Ω. □

2.5 Green's function

Let us consider the weak formulation (2.10) of problem (2.1)–(2.3). If the coefficients $\mathcal{A}$, $\boldsymbol{b}$, c, the domain $\Omega \subset \mathbb{R}^d$, and the parts Γ_{D} and Γ_{N} are fixed then any triplet $(f, g_{\mathrm{D}}, g_{\mathrm{N}}) \in L^2(\Omega) \times L^2(\Gamma_{\mathrm{D}}) \times L^2(\Gamma_{\mathrm{N}})$ yields the unique weak solution $u \in H^1(\Omega)$. This defines an operator $\widetilde{G} : L^2(\Omega) \times L^2(\Gamma_{\mathrm{D}}) \times L^2(\Gamma_{\mathrm{N}}) \mapsto H^1(\Omega)$ such that $\widetilde{G}(f, g_{\mathrm{D}}, g_{\mathrm{N}}) = u$. The operator $\widetilde{G}$ is inverse to the differential operator (2.10) and it is called the *Green's operator*. The existence, uniqueness, compactness, and other properties of this operator are well described and proved in [61].

The Green's operator for problem (2.10) can be often expressed as the following integral operator, see e.g. [65, p. 88],

$$u(\boldsymbol{y}) = \int_{\Omega} f(\boldsymbol{x}) G(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}}(\boldsymbol{s}) G(\boldsymbol{s}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{s} - \int_{\Gamma_{\mathrm{D}}} g_{\mathrm{D}}(\boldsymbol{s}) \boldsymbol{n}^{\top} \mathcal{A}(\boldsymbol{s}) \boldsymbol{\nabla}_{\boldsymbol{s}} G(\boldsymbol{s}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{s}, \tag{2.16}$$

where $\boldsymbol{y} \in \Omega$ and $\boldsymbol{\nabla}_{\boldsymbol{s}} G(\boldsymbol{s}, \boldsymbol{y})$ denotes the gradient of $G(\boldsymbol{s}, \boldsymbol{y})$ with respect to its first variable. This identity is known as the *Kirchhoff-Helmholtz representation formula* and the kernel $G(\boldsymbol{x}, \boldsymbol{y})$ of the involved integral operators is called the *Green's function*. It is often convenient to consider G as a function of one variable only and therefore, we use the notation $G_{\boldsymbol{y}}(\cdot) = G(\cdot, \boldsymbol{y})$ for $\boldsymbol{y} \in \Omega$.

The Green's function $G(\boldsymbol{x}, \boldsymbol{y})$ has a singularity for $\boldsymbol{x} = \boldsymbol{y}$. In addition, in order to enforce the finite values of integrals in (2.16), it is natural to consider

$$G_{\boldsymbol{y}} \in L^1(\Omega), \quad G_{\boldsymbol{y}} \in L^1(\Gamma_{\mathrm{N}}), \quad \text{and} \quad \boldsymbol{n}^{\top} \mathcal{A} \boldsymbol{\nabla} G_{\boldsymbol{y}} \in L^1(\Gamma_{\mathrm{D}}). \tag{2.17}$$

Under these regularity conditions the integrals in (2.16) are well defined for $f \in L^\infty(\Omega)$, $g_{\mathrm{N}} \in L^\infty(\Gamma_{\mathrm{N}})$, and $g_{\mathrm{D}} \in L^\infty(\Gamma_{\mathrm{D}})$.

Let us note that the requirement of the additional L^∞-regularity on the data f, g_{D}, and g_{N} is necessary. The natural L^2-regularity is not sufficient because $G_{\boldsymbol{y}} \notin L^2(\Omega)$ in general, and integrals in (2.16) might be infinite. For an example we refer to (2.21) below, where we present the fundamental solution and subsequently the Green's function for the Poisson problem.

If the Green's operator can be expressed in the integral form (2.16) and if the Green's function $G_{\boldsymbol{y}}(\cdot) = G(\cdot, \boldsymbol{y})$ is sufficiently regular, then it can be regarded as a solution to the differential equation adjoint to (2.1) with the Dirac distribution $\delta_{\boldsymbol{y}}$ on the right-hand side and with homogeneous boundary conditions:

$$-\operatorname{div}(\mathcal{A}^{\top} \boldsymbol{\nabla} G_{\boldsymbol{y}}) - \operatorname{div}(G_{\boldsymbol{y}} \boldsymbol{b}) + c G_{\boldsymbol{y}} = \delta_{\boldsymbol{y}} \quad \text{in } \Omega, \tag{2.18}$$

$$G_{\boldsymbol{y}} = 0 \quad \text{on } \Gamma_{\mathrm{D}}, \tag{2.19}$$

$$(\alpha + \boldsymbol{b} \cdot \boldsymbol{n}) G_{\boldsymbol{y}} + \boldsymbol{n}^{\top} \mathcal{A}^{\top} \boldsymbol{\nabla} G_{\boldsymbol{y}} = 0 \quad \text{on } \Gamma_{\mathrm{N}}. \tag{2.20}$$

The equality (2.18) is understood in the sense of distributions.

Anyway, the rigorous proof of the existence and uniqueness of the Green's function for the general problem (2.1)–(2.3) is a delicate mathematical problem and requires additional technical assumptions. In [55, Ch. 11] the C^2 regularity of the boundary $\partial\Omega$ is assumed, the H^2 regularity of the elliptic problem (2.10) is exploited, and subsequently the existence and uniqueness of the Green's function is established. The Green's function on a general domain Ω is then obtained as a limit of the Green's functions on an expanding sequence of C^2 domains Ω_n such that $\cup_n \Omega_n = \Omega$. Nevertheless, the homogeneous Dirichlet boundary conditions only are considered there.

Book [59] provides the existence and uniqueness of the Green's function for problem (2.1)–(2.3) with smooth coefficients, homogeneous Dirichlet boundary conditions only, and for piecewise smooth polyhedral-like domains $\Omega \subset \mathbb{R}^3$. Similarly, paper [37] proves the existence and uniqueness of the Green's function for $d \geq 3$, for coefficients $\mathcal{A} \in [L^\infty(\Omega)]^{d\times d}$, $\boldsymbol{b} = \boldsymbol{0}$, and $c = 0$ and again for homogeneous Dirichlet boundary conditions only. Completely different and quite general approach based on the theory of distributions is presented in [67].

On the other hand, in a special case of one dimension there are practically no technicalities and the Green's function can be defined very naturally, see e.g. [19]. This discrepancy between the one- and the higher-dimensional case can be explained by the fact that the Sobolev space $H^1(\Omega)$ is embedded in the space of continuous functions $C(\overline{\Omega})$ for dimension $d = 1$ and not for $d > 1$.

In addition, in certain special cases even the explicit formulas for the Green's function exist. These explicit formulas are very useful in physics and engineering, because they enable to gain a lot of information about the corresponding problem and its solutions. The engineering approach to Green's function is well described in [73] and in [23], where a variety of explicit expressions of Green's functions in special cases can be found.

As an example, let us present the well-known case of the Poisson problem. Equation (2.1) collapses to the Poisson equation for $\mathcal{A} = I$, $\boldsymbol{b} = \boldsymbol{0}$, and $c = 0$. Poisson equation possesses the following well-known fundamental solution:

$$F(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} \dfrac{1}{2\pi}\ln\dfrac{1}{|\boldsymbol{x}-\boldsymbol{y}|} & \text{for } d = 2, \\ \dfrac{1}{(d-2)\varkappa_d}\dfrac{1}{|\boldsymbol{x}-\boldsymbol{y}|^{d-2}} & \text{for } d \geq 3, \end{cases} \tag{2.21}$$

where $\varkappa_d$ stands for the $(d-1)$-dimensional measure of the unit sphere in $\mathbb{R}^d$ and $|\cdot|$ denotes the Euclidean norm. It can be easily shown that for all $\boldsymbol{y} \in \mathbb{R}^d$ the fundamental solution $F_{\boldsymbol{y}}(\cdot) = F(\cdot,\boldsymbol{y})$ is a harmonic function in any domain not

containing the point $\boldsymbol{y}$. In addition it satisfies the equality

$$-\Delta F_{\boldsymbol{y}} = \delta_{\boldsymbol{y}} \quad \text{in } \Omega$$

for all $\boldsymbol{y} \in \Omega$ in the sense of distributions. In order to incorporate the boundary conditions (2.2)–(2.3), we consider for all $\boldsymbol{y} \in \Omega$ the following problem with the homogeneous right-hand side:

$$\begin{aligned} -\Delta N_{\boldsymbol{y}} &= 0 && \text{in } \Omega, \\ N_{\boldsymbol{y}} &= F_{\boldsymbol{y}} && \text{on } \Gamma_{\mathrm{D}}, \\ \alpha N_{\boldsymbol{y}} + \boldsymbol{n} \cdot \boldsymbol{\nabla} N_{\boldsymbol{y}} &= \alpha F_{\boldsymbol{y}} + \boldsymbol{n} \cdot \boldsymbol{\nabla} F_{\boldsymbol{y}} && \text{on } \Gamma_{\mathrm{N}}. \end{aligned}$$

Since $\boldsymbol{y} \notin \partial\Omega$, there is no singularity in the boundary conditions, this problem is well-posed, and it possesses a unique solution $N_{\boldsymbol{y}}$. Thus, we finally put $G(\boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x}, \boldsymbol{y}) - N(\boldsymbol{x}, \boldsymbol{y})$, where $N(\boldsymbol{x}, \boldsymbol{y}) = N_{\boldsymbol{y}}(\boldsymbol{x})$. This $G_{\boldsymbol{y}}(\cdot) = G(\cdot, \boldsymbol{y})$ clearly satisfies (2.18)–(2.20) (with $\mathcal{A} = I$, $\boldsymbol{b} = \boldsymbol{0}$, and $c = 0$) and hence it is the Green's function yielding the representation formula (2.16).

We finish this section by the well-known equivalence between the nonnegativity of the Green's function and the validity of the maximum principle. This equivalence is mentioned e.g. in [15, 19, 48] for homogeneous Dirichlet boundary conditions and it is presented in [65, p. 88] for the general mixed boundary conditions. For the reader's convenience we state it and prove it again here.

Theorem 2.6. *Let us consider problem (2.10) with $f \in L^\infty(\Omega)$, $g_{\mathrm{N}} \in L^\infty(\Gamma_{\mathrm{N}})$, and $g_{\mathrm{D}} \in L^\infty(\Gamma_{\mathrm{D}})$. Let the Green's operator corresponding to problem (2.10) admit the integral form (2.16) and let the Green's function $G_{\boldsymbol{y}}(\cdot) = G(\cdot, \boldsymbol{y})$ satisfy the regularity (2.17) for a.a. $\boldsymbol{y} \in \Omega$. In addition, let $G_{\boldsymbol{y}}(\boldsymbol{s}) = 0$ for a.a. $\boldsymbol{s} \in \Gamma_{\mathrm{D}}$, see (2.19). Further, let for a.a. $\boldsymbol{y} \in \Omega$ an open set ω exist such that $\Gamma_{\mathrm{D}} \subset \omega$, $\Gamma_{\mathrm{N}} \cap \omega = \emptyset$, $\mathrm{meas}_d(\omega) > 0$, $\boldsymbol{y} \notin \omega$, and $G_{\boldsymbol{y}} \in C^1(\omega \cap \overline{\Omega})$. Then problem (2.10) satisfies the conservation of nonnegativity if and only if $G(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for a.a. $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega^2$ and for a.a. $(\boldsymbol{x}, \boldsymbol{y}) \in \Gamma_{\mathrm{N}} \times \Omega$.*

Proof. Let us first assume that problem (2.10) satisfies the conservation of nonnegativity. Then we can take $g_{\mathrm{D}} = 0$, $g_{\mathrm{N}} = 0$ in Definition 2.7, and by (2.16) we obtain

$$u(\boldsymbol{y}) = \int_\Omega f(\boldsymbol{x}) G(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \geq 0 \quad \forall f \in L^\infty(\Omega),\ f \geq 0,$$

for a.a. $\boldsymbol{y} \in \Omega$. Therefore, $G_{\boldsymbol{y}}(\boldsymbol{x}) \geq 0$ for a.a. $\boldsymbol{x} \in \Omega$. Similarly, the conservation of nonnegativity with $f = 0$, $g_{\mathrm{D}} = 0$, and arbitrary $g_{\mathrm{N}} \geq 0$ yields the nonnegativity of $G_{\boldsymbol{y}}(\boldsymbol{x})$ for a.a. $\boldsymbol{x} \in \Gamma_{\mathrm{N}}$.

The converse implication follows from (2.16) as well. If we assume the nonnegativity of $G(\boldsymbol{x}, \boldsymbol{y})$ for a.a. $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega^2$ and the nonnegativity of $G(\boldsymbol{x}, \boldsymbol{y})$ for

a.a. $(\boldsymbol{x}, \boldsymbol{y}) \in \Gamma_{\mathrm{N}} \times \Omega$, then inequalities

$$\int_{\Omega} f(\boldsymbol{x}) G(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \geq 0 \quad \text{and} \quad \int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}}(\boldsymbol{s}) G(\boldsymbol{s}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{s} \geq 0 \tag{2.22}$$

hold trivially for any $f \in L^{\infty}(\Omega)$, $f \geq 0$, and any $g_{\mathrm{N}} \in L^{\infty}(\Gamma_{\mathrm{N}})$, $g_{\mathrm{N}} \geq 0$.

Further, the assumption $G_{\boldsymbol{y}} \in C^1(\omega \cap \overline{\Omega})$ enables us to utilize the classical definition of the directional derivative

$$\boldsymbol{n}^{\top} \mathcal{A} \boldsymbol{\nabla} G_{\boldsymbol{y}}(\boldsymbol{s}) = \lim_{t \to 0-} \frac{G_{\boldsymbol{y}}(\boldsymbol{s} + t\boldsymbol{\mu}) - G_{\boldsymbol{y}}(\boldsymbol{s})}{t},$$

where $\boldsymbol{\mu} = \mathcal{A}^{\top} \boldsymbol{n}$. Now, we realize that $\boldsymbol{n}^{\top} \boldsymbol{\mu} = \boldsymbol{\mu}^{\top} \boldsymbol{n} = \boldsymbol{n}^{\top} \mathcal{A} \boldsymbol{n} > 0$ by (2.5). Thus, the angle between $\boldsymbol{\mu}$ and the normal vector $\boldsymbol{n}$ is acute and the point $\boldsymbol{s} + t\boldsymbol{\mu}$ lies inside the domain Ω for all sufficiently small $t < 0$. Therefore, $G_{\boldsymbol{y}}(\boldsymbol{s} + t\boldsymbol{\mu}) \geq 0$ and since $G_{\boldsymbol{y}}(\boldsymbol{s}) = 0$ for $\boldsymbol{s} \in \Gamma_{\mathrm{D}}$ we conclude that $\boldsymbol{n}^{\top} \mathcal{A} \boldsymbol{\nabla} G_{\boldsymbol{y}}(\boldsymbol{s}) \leq 0$ for $\boldsymbol{s} \in \Gamma_{\mathrm{D}}$. Thus, the inequality

$$- \int_{\Gamma_{\mathrm{D}}} g_{\mathrm{D}}(\boldsymbol{s}) \boldsymbol{n}^{\top} \mathcal{A}(\boldsymbol{s}) \boldsymbol{\nabla} G_{\boldsymbol{y}}(\boldsymbol{s}) \, \mathrm{d}\boldsymbol{s} \geq 0 \tag{2.23}$$

holds true for any $g_{\mathrm{D}} \in L^{\infty}(\Gamma_{\mathrm{D}})$, $g_{\mathrm{D}} \geq 0$. Finally, inequalities (2.22) and (2.23) used in (2.16) finish the proof. □

We note that especially the assumption $G_{\boldsymbol{y}} \in C^1(\omega \cap \overline{\Omega})$ of Theorem 2.6 is somewhat artificial. On the other hand, the Green's function $G_{\boldsymbol{y}}(\boldsymbol{x})$ is known to be smooth in any subdomain of Ω not-containing arbitrarily small neighborhood of the singular point $\boldsymbol{x} = \boldsymbol{y}$. Thus, the assumption $G_{\boldsymbol{y}} \in C^1(\omega \cap \overline{\Omega})$ is realistic. For an illustration we refer to the above example of the Poisson problem, where all the technical assumptions of Theorem 2.6 are satisfied. Finally, we note that conditions for the nonnegativity of the Green's function are studied in [48].

CHAPTER

THREE

Discrete maximum principles in the finite element method

The previous chapter described the qualitative properties of the solution of linear second-order partial differential equations like the maximum, minimum, and comparison principles. If the (continuous) problem is discretized then it is natural to consider the discrete counterparts of these (continuous) principles. This chapter defines the discrete maximum principles (DMP) and shows that they possess the same properties as the continuous principles including the relationship with the discrete Green's function.

Let us point out that all statements of this chapter are valid for a general continuous and V-elliptic bilinear form a on a Hilbert space V, see (2.13) and (2.14), and for a general linear and continuous operator $\mathcal{F}$ on V. In particular, it is not necessary to assume the particular form (2.11) and (2.12) of a and $\mathcal{F}$.

This chapter is organized as follows. Section 3.1 gives a brief summary of the finite element method. Section 3.2 defines the discrete qualitative properties of the finite element solution and presents their equivalence. In Section 3.3 the discrete Green's function is defined and the equivalence of its nonnegativity and of the DMP is proved.

3.1 Finite element method

The finite element method (FEM) is a standard method in the numerical analysis of partial differential equations. Its detailed description can be found in many textbooks, see e.g. [10, 17, 71]. In what follows, we summarize the FEM very briefly in order to introduce the necessary notation and properties needed in the subsequent analysis of the DMP.

On the continuous level, the weak solution is naturally defined in $H^1(\Omega)$ and the Dirichlet boundary conditions are represented by the subspace $V \subset H^1(\Omega)$, see (2.9). On the discrete level, we need a compatible discretization of these spaces. To construct such a compatible discretization it is usually necessary to approximate the domain Ω somehow. Most often a polytopic approximation of Ω is used. Nevertheless, here we will not describe this topic and we will just assume that there are finite dimensional spaces V_h and X_h such that

$$V_h \subset V, \quad X_h \subset H^1(\Omega), \quad V_h \subset X_h \subset C(\overline{\Omega}). \tag{3.1}$$

The space X_h is used for the approximation of the Dirichlet lift. Hence, let $g_{\mathrm{D},h} \in X_h$ be an approximation of the lift $g_{\mathrm{D}} \in H^1(\Omega)$. The values of $g_{\mathrm{D},h}$ on Γ_{D} are obtained in a suitable way (often 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 $g_{\mathrm{D},h}$ is not important at this point. For the purposes of this chapter, we consider arbitrary $g_{\mathrm{D},h} \in X_h$.

The FEM is a special case of the well-known Galerkin method. We say that $u_h = u_h^0 + g_{\mathrm{D},h}$ is a Galerkin solution of (2.10) if $u_h^0 \in V_h$ and

$$a(u_h^0, v_h) = \mathcal{F}(v_h) - a(g_{\mathrm{D},h}, v_h) \quad \forall v_h \in V_h, \tag{3.2}$$

where the bilinear form a and the linear functional $\mathcal{F}$ are given by (2.11) and (2.12), respectively. For a fixed discrete Dirichlet lift $g_{\mathrm{D},h}$ there exists the unique Galerkin solution u_h.

Considering a basis $\varphi_1, \varphi_2, \ldots, \varphi_{N^0}$, where $N^0 = \dim V_h$, and expressing the solution u_h^0 as a linear combination of the basis functions as

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

problem (3.2) is equivalent to a system of linear algebraic equations

$$A\boldsymbol{z} = \boldsymbol{F},$$

where $\boldsymbol{z} = (z_1, z_2, \ldots, z_{N^0})^\top$ and the stiffness matrix $A \in \mathbb{R}^{N^0 \times N^0}$ and the load vector $\boldsymbol{F} \in \mathbb{R}^{N^0}$ have entries

$$A_{ij} = a(\varphi_j, \varphi_i) \quad \text{and} \quad F_i = \mathcal{F}(\varphi_i) - a(g_{\mathrm{D},h}, \varphi_i), \quad i, j = 1, 2, \ldots, N^0. \tag{3.3}$$

Notice that the V-ellipticity of the bilinear form a implies the nonsingularity of A. Even more, it implies the positive definiteness

$$\boldsymbol{z}^\top A \boldsymbol{z} > 0 \quad \text{for all } \boldsymbol{z} \in \mathbb{R}^{N^0},\ \boldsymbol{z} \neq \boldsymbol{0}. \tag{3.4}$$

In order to handle the approximation of the Dirichlet lift $g_{\mathrm{D},h}$, we append the basis $\varphi_1, \varphi_2, \ldots, \varphi_{N^0}$ of V_h by functions $\varphi_{N^0+1}, \varphi_{N^0+2}, \ldots, \varphi_N$ 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 $\varphi_{N^0+1}, \varphi_{N^0+2}, \ldots, \varphi_N$. Hence, $X_h = V_h \oplus V_h^\partial$, where $\oplus$ denotes the direct sum, $\dim X_h = N$, $\dim V_h = N^0$, and $\dim V_h^\partial = N^\partial$. Clearly, $N = N^0 + N^\partial$. For further reference, we also set $\varphi_k^\partial = \varphi_{N^0+k}$ for $k = 1, 2, \ldots, N^\partial$. See Figure 3.1 for an illustration.

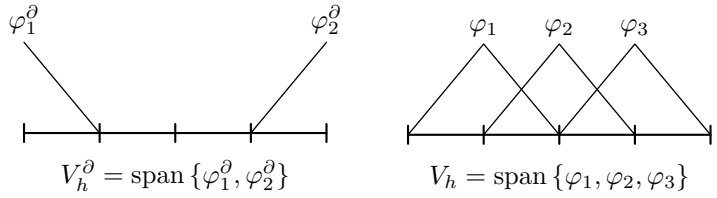


Figure 3.1: Notation for the interior and boundary basis functions – a 1D illustration.

For the subsequent analysis of the discrete Green's function and the DMP, we will utilize the matrix $A^\partial \in \mathbb{R}^{N^0 \times N^\partial}$ with entries

$$A_{ik}^\partial = a(\varphi_k^\partial, \varphi_i), \quad i = 1, 2, \ldots, N^0, \quad k = 1, 2, \ldots, N^\partial. \tag{3.5}$$

The FEM can be regarded as a special case of the Galerkin method, where the space X_h and the basis functions $\varphi_1, \varphi_2, \ldots, \varphi_N$ are constructed with the aid of a triangulation of the domain Ω in such a way that the corresponding stiffness matrices A and A^∂ are sparse.

The finite element triangulation (or partition or mesh) of Ω is a finite set $\mathcal{T}_h = \{K_i : i = 1, 2, \ldots, M\}$ of subdomains – elements – $K_i \subset \overline{\Omega}$ with the following properties, see e.g. [17, 53]:

($\mathcal{T}1$) $\bigcup_{i=1}^M K_i = \overline{\Omega}$,

($\mathcal{T}2$) each $K \in \mathcal{T}_h$ is a closed set and its interior K^0 is nonempty,

($\mathcal{T}3$) all pairs of distinct elements $K_1, K_2 \in \mathcal{T}_h$ satisfy $K_1^0 \cap K_2^0 = \emptyset$,

($\mathcal{T}4$) the boundary ∂K is Lipschitz for all $K \in \mathcal{T}_h$.

In the subsequent chapters, we will limit ourselves to polytopic domains $\Omega \subset \mathbb{R}^d$. In that case we will consider polytopic finite element meshes which are required to have the following additional properties:

($\mathcal{T}5$) all elements $K \in \mathcal{T}_h$ are polytopic and convex,

($\mathcal{T}6$) each face of any element $K \in \mathcal{T}_h$ lies either on the boundary $\partial\Omega$ or it is a face of another element $K^* \in \mathcal{T}_h$,

($\mathcal{T}7$) interiors of all faces of all elements in $\mathcal{T}_h$ are disjoint with $\overline{\Gamma}_{\mathrm{D}} \cap \overline{\Gamma}_{\mathrm{N}}$.

Anyway, having the finite element mesh $\mathcal{T}_h$ satisfying properties ($\mathcal{T}1$)–($\mathcal{T}4$), we can split the bilinear and linear forms a and $\mathcal{F}$ into local contributions:

$$a(u,v) = \sum_{K\in\mathcal{T}_h} a_K(u,v) \quad \text{and} \quad \mathcal{F}(v) = \sum_{K\in\mathcal{T}_h} \mathcal{F}_K(v) \quad \forall u, v \in H^1(\Omega), \tag{3.6}$$

where in accordance with (2.11) and (2.12) we put

$$\begin{aligned} a_K(u,v) &= \int_K [(\mathcal{A}\boldsymbol{\nabla} u)\cdot\boldsymbol{\nabla} v + (\boldsymbol{b}\cdot\boldsymbol{\nabla} u)v + cuv]\,\mathrm{d}\boldsymbol{x} + \int_{\partial K\cap\Gamma_{\mathrm{N}}} \alpha uv\,\mathrm{d}\boldsymbol{s}, \\ \mathcal{F}_K(v) &= \int_K fv\,\mathrm{d}\boldsymbol{x} + \int_{\partial K\cap\Gamma_{\mathrm{N}}} g_{\mathrm{N}} v\,\mathrm{d}\boldsymbol{s}. \end{aligned}$$

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) $\overline{A}^K \in \mathbb{R}^{N^0\times N^0}$ and $\overline{A}^{\partial,K} \in \mathbb{R}^{N^0\times N^\partial}$ as

$$\begin{aligned} \overline{A}^K_{ij} &= a_K(\varphi_j,\varphi_i), \quad i,j = 1,2,\ldots,N^0, \\ \overline{A}^{\partial,K}_{ik} &= a_K(\varphi^\partial_k,\varphi_i), \quad i = 1,2,\ldots,N^0, \quad k = 1,2,\ldots,N^\partial. \end{aligned}$$

However, if the basis functions are defined using the standard finite element machinery, then any given element K is contained in supports of a few basis functions only 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 *connectivity mappings*.

Let us define sets $I(K)$, $I^0(K)$, and $I^\partial(K)$ of indices of basis functions whose support contains the element K:

$$\begin{aligned} I(K) &= \{i \in \mathbb{N} : 1 \le i \le N,\ K \subset \operatorname{supp}\varphi_i\}, \\ I^0(K) &= \{j \in \mathbb{N} : 1 \le j \le N^0,\ K \subset \operatorname{supp}\varphi_j\}, \\ I^\partial(K) &= \{k \in \mathbb{N} : 1 \le k \le N^\partial,\ K \subset \operatorname{supp}\varphi^\partial_k\}. \end{aligned}$$

We denote by N_K, N^0_K, and N^∂_K the numbers of indices in the sets $I(K)$, $I^0(K)$, and $I^\partial(K)$, respectively. Clearly, $I^0(K) \subset I(K)$ and $N_K = N^0_K + N^\partial_K$. By *connectivity mappings* we understand arbitrary but fixed one-to-one mappings $\iota_K : \{1,2,\ldots,N_K\} \mapsto I(K)$, $\iota^0_K : \{1,2,\ldots,N^0_K\} \mapsto I^0(K)$, and $\iota^\partial_K : \{1,2,\ldots,N^\partial_K\} \mapsto$

$I^\partial(K)$ such that $\iota_K(m) = \iota^0_K(m)$ for $m = 1, 2, \dots, N^0_K$ and $\iota_K(N^0_K + m) = N^0 + \iota^\partial_K(m)$ 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. [21, 71, 72].

The connectivity mappings enable us to define the *shape functions* as $\varphi^K_m = \varphi_i|_K$ with $i = \iota_K(m)$, $K \in \mathcal{T}_h$, $m = 1, 2, \dots, N_K$. In particular, we set $\varphi^{K,\partial}_q = \varphi^K_{N^0_K+q} = \varphi^\partial_j|_K$ with $j = \iota^\partial_K(q)$, $K \in \mathcal{T}_h$, $q = 1, 2, \dots, N^\partial_K$. Afterall, we use the shape functions to define the entries of the condensed local stiffness matrices $A^K \in \mathbb{R}^{N^0_K \times N^0_K}$ and $A^{\partial,K} \in \mathbb{R}^{N^0_K \times N^\partial_K}$ as

$$A^K_{mn} = \overline{A}^K_{\iota_K(m),\iota_K(n)} = a_K\big(\varphi^K_n, \varphi^K_m\big), \qquad m, n = 1, 2, \dots, N^0_K, \tag{3.7}$$

$$A^{\partial,K}_{mq} = \overline{A}^{\partial,K}_{\iota_K(m),\iota_K(q)} = a_K\big(\varphi^{K,\partial}_q, \varphi^K_m\big), \quad m = 1, \dots, N^0_K,\ q = 1, \dots, N^\partial_K. \tag{3.8}$$

Using (3.6) and the above definitions, we can express the entries of the (global) matrices A and A^∂ as follows

$$A_{ij} = \sum_{K\in\mathcal{T}_h} a_K(\varphi_j, \varphi_i) = \sum_{K\in\mathcal{T}_h} \overline{A}^K_{ij} = \sum_{\{K\in\mathcal{T}_h : i,j\in I^0(K)\}} A^K_{\iota_K^{-1}(i),\iota_K^{-1}(j)}, \tag{3.9}$$

$$A^\partial_{ik} = \sum_{K\in\mathcal{T}_h} a_K(\varphi^\partial_k, \varphi_i) = \sum_{K\in\mathcal{T}_h} \overline{A}^{\partial,K}_{ij} = \sum_{\{K\in\mathcal{T}_h : i\in I^0(K),\ k\in I^\partial(K)\}} A^{\partial,K}_{\iota_K^{-1}(i),(\iota^\partial_K)^{-1}(k)}, \tag{3.10}$$

where $i, j = 1, 2, \dots, N^0$ and $k = 1, 2, \dots, N^\partial$. Further on we will solely use the condensed local stiffness matrices A^K and $A^{\partial,K}$ and we will call them simply local (stiffness) matrices.

Nevertheless, the subsequent results do not need any special information about the space V_h, its basis, and the local stiffness matrices. All the remaining results in this chapter concern the general Galerkin solution. However, the refined analysis of the DMP presented in the next chapters will be based on the computations of entries of the local stiffness matrices.

3.2 Discrete maximum principles

This section presents natural discrete analogues of the qualitative properties given in Definitions 2.1–2.4. Here and in the sequel we assume that V_h contains continuous functions only.

Definition 3.1. Let the spaces V_h and X_h be fixed. Problem (3.2) satisfies the discrete maximum principle if

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

Definition 3.2. Let the spaces V_h and X_h be fixed. Problem (3.2) satisfies the discrete minimum principle if

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

Definition 3.3. Let the spaces V_h and X_h be fixed. Problem (3.2) satisfies the discrete conservation of nonnegativity if

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

Definition 3.4. Let the spaces V_h and X_h be fixed. Let $u_{h,1} \in X_h$ be the solution to problem (3.2) with right-hand side f_1 and boundary data $g_{\mathrm{D},h,1}$, $g_{\mathrm{N},1}$ and similarly let $u_{h,2} \in X_h$ be the solution for f_2, $g_{\mathrm{D},h,2}$, and $g_{\mathrm{N},2}$. We say that problem (3.2) satisfies the discrete comparison principle if

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

Theorem 3.1. *Let the space X_h contain all constant functions. Let $c \geq 0$ a.e. in Ω and $\alpha \geq 0$ a.e. on Γ_{N}. Then the following statements are equivalent.*

(i) *Problem (3.2) satisfies the discrete maximum principle.*
(ii) *Problem (3.2) satisfies the discrete minimum principle.*
(iii) *Problem (3.2) satisfies the discrete conservation of nonnegativity.*
(iv) *Problem (3.2) satisfies the discrete comparison principle.*

Proof. The proof is analogous to the proof of Theorem 2.4 above. □

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 our case it is the finite element space V_h and consequently the underlined triangulation. The standard results about the DMP for the linear finite elements, see Chapter 4, usually define a class of spaces V_h (or equivalently a class of triangulations) for which the DMP is satisfied.

However, this is not the only possibility how to transfer the maximum principle to the discrete level. Another option, equally natural, is to consider the given nonnegative data f, g_{D}, and g_{N} to be fixed and seek a suitable class of spaces V_h (or triangulations) specific for the given data such that the corresponding solution $u_h \in V_h$ is nonnegative. The author of this thesis is not aware of any source, where this approach is mentioned or treated. This is an interesting open problem and a topic for further research.

3.3 Discrete Green's function

In the context of the FEM a natural discrete analog of the Green's function, the discrete Green's function (DGF), can be defined. The DGF has been introduced already in [16, 19]. We point out also the analysis [22] of the DGF for the lowest-order finite elements. The DGF possesses the analogous properties as the Green's function for the continuous problem including the equivalence of the DMP with the nonnegativity of the DGF. This section defines the DGF and proves its properties.

Let us recall the assumption that V_h is a finite dimensional space containing continuous functions, see (3.1).

Definition 3.5. Let $\boldsymbol{y} \in \Omega$ and let $G_{h,\boldsymbol{y}} \in V_h$ be the unique solution of the problem

$$a(v_h, G_{h,\boldsymbol{y}}) = v_h(\boldsymbol{y}) \quad \forall v_h \in V_h. \tag{3.11}$$

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}. In order to handle this action, we consider the elliptic projection $\Pi_h^0 : X_h \mapsto V_h$. The elliptic projection $\Pi_h^0 w_h \in V_h$ of a $w_h \in X_h$ is uniquely determined by the requirement

$$a(w_h - \Pi_h^0 w_h, v_h) = 0 \quad \forall v_h \in V_h. \tag{3.12}$$

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

Theorem 3.2. *The Galerkin solution $u_h \in X_h$ to problem (3.2) satisfies the following representation formula*

$$u_h(\boldsymbol{y}) = \mathcal{F}(G_{h,\boldsymbol{y}}) + g_{\mathrm{D},h}(\boldsymbol{y}) - (\Pi_h^0 g_{\mathrm{D},h})(\boldsymbol{y}). \tag{3.13}$$

Proof. By (3.11) with $v_h = u_h^0 + \Pi_h^0 g_{\mathrm{D},h}$, (3.12), and (3.2) we immediately obtain

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

Hence, the statement follows from the fact that $u_h = u_h^0 + g_{\mathrm{D},h}$. □

Remark 3.1. Using the particular form (2.12) of the linear functional $\mathcal{F}$ we can express the representation formula (3.13) as

$$u_h(\boldsymbol{y}) = \int_\Omega f(\boldsymbol{x}) G_h(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}}(\boldsymbol{s}) G_h(\boldsymbol{s}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{s} + g_{\mathrm{D},h}(\boldsymbol{y}) - (\Pi_h^0 g_{\mathrm{D},h})(\boldsymbol{y}). \tag{3.14}$$

Here, we clearly observe the explicit dependence of the solution u_h on the data f, $g_{\mathrm{D},h}$, and g_{N}. Furthermore, we can compare (3.14) with the Kirchhoff-Helmholtz representation formula (2.16) to see the difference between the continuous and discrete case.

The following theorem states the main result of this section: the equivalent conditions for the validity of the DMP.

Theorem 3.3. *Problem (3.2) satisfies the discrete conservation of nonnegativity if and only if*

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

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

Proof. The fact that conditions (a) and (b) imply the discrete conservation of nonnegativity is an immediate consequence of (3.14). Notice that the nonnegativity of $G_{h,\boldsymbol{y}}$ on Γ_{N} is guaranteed by the continuity of G_h in Ω^2. The converse implication follows from (3.14), too. Indeed, taking $\boldsymbol{y} \in \Omega$, $g_{\mathrm{N}} = 0$, and $g_{\mathrm{D},h} = 0$, 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 (3.14) with $f = 0$ and $g_{\mathrm{N}} = 0$. □

3.4 Expressing the discrete Green's function in a basis

The Green's function on the continuous level can be explicitly found in exceptional cases only. In contrast, on the discrete level, the DGF can always be computed – at least theoretically. Practically, we can compute it only if the size of the discrete problem (the dimension N^0) allows it. The following theorem shows an explicit expression for the DGF in terms of the inverse of the stiffness matrix A, see (3.3). We point out that a version of this result based on eigenfunctions of the discrete Laplacian was published already in 1970 in [16] and [19]. Anyway, for the reader's convenience we present its proof here, although it can be found in [83], too.

Theorem 3.4. *Let $\varphi_1, \varphi_2, \dots, \varphi_{N^0}$ be a basis in V_h and let A be the corresponding stiffness matrix given by (3.3). Then the DGF can be expressed as follows*

$$G_h(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N^0} \sum_{j=1}^{N^0} \varphi_i(\boldsymbol{y}) (A^{-1})_{ij} \varphi_j(\boldsymbol{x}). \tag{3.15}$$

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

$$G_{h,\boldsymbol{y}}(\boldsymbol{x}) = \sum_{j=1}^{N^0} d_j(\boldsymbol{y}) \varphi_j(\boldsymbol{x}). \tag{3.16}$$

Using this expansion in (3.11) tested by all the basis functions, we obtain

$$\varphi_i(\boldsymbol{y}) = a\left(\varphi_i, \sum_{j=1}^{N^0} d_j(\boldsymbol{y}) \varphi_j(\boldsymbol{x})\right) = \sum_{j=1}^{N^0} d_j(\boldsymbol{y}) A_{ji}, \quad i = 1, 2, \dots, N^0.$$

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

$$d_k(\boldsymbol{y}) = \sum_{i=1}^{N^0} \varphi_i(\boldsymbol{y}) (A^{-1})_{ik}, \quad k = 1, 2, \dots, N^0.$$

Inserting this into (3.16), we obtain (3.15). □

The error of the elliptic projection $\Pi_h^0 g_{\mathrm{D},h}$ needed in the representation formula (3.14) can be expressed in a similar way as the DGF, using the basis functions and the stiffness matrices.

Theorem 3.5. *Let $X_h = V_h \oplus V_h^\partial$, let $\varphi_1, \varphi_2, \dots, \varphi_{N^0}$ be a basis in V_h, let $\varphi_1^\partial, \varphi_2^\partial, \dots, \varphi_{N^\partial}^\partial$ be a basis in V_h^∂, and let the matrices A and A^∂ be given by (3.3) and (3.5), respectively. Let the approximation of the Dirichlet lift $g_{\mathrm{D},h} \in X_h$ be expressed as*

$$g_{\mathrm{D},h}(\boldsymbol{y}) = \sum_{\ell=1}^{N^\partial} c_\ell^\partial \varphi_\ell^\partial(\boldsymbol{y}) + \sum_{i=1}^{N^0} c_i^0 \varphi_i(\boldsymbol{y}) \quad \forall \boldsymbol{y} \in \Omega. \tag{3.17}$$

Then

$$g_{\mathrm{D},h}(\boldsymbol{y}) - \Pi_h^0 g_{\mathrm{D},h}(\boldsymbol{y}) = \sum_{\ell=1}^{N^\partial} c_\ell^\partial \left[\varphi_\ell^\partial(\boldsymbol{y}) - \Pi_h^0 \varphi_\ell^\partial(\boldsymbol{y})\right] \quad \forall \boldsymbol{y} \in \Omega, \tag{3.18}$$

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

$$\Pi_h^0 \varphi_\ell^\partial(\boldsymbol{y}) = \sum_{i=1}^{N^0} \sum_{j=1}^{N^0} \varphi_i(\boldsymbol{y}) (A^{-1})_{ij} A_{j\ell}^\partial \quad \forall \boldsymbol{y} \in \Omega,\ \ell = 1, 2, \ldots, N^\partial. \tag{3.19}$$

Proof. The equality (3.18) follows immediately from the linearity of the elliptic projection Π_h^0 and from the fact that $\Pi_h^0 \varphi_i = \varphi_i$, because $\varphi_i \in V_h$ for all $i = 1, 2, \ldots, N^0$. To prove (3.19), we express $\Pi_h^0 \varphi_\ell^\partial$ as

$$\Pi_h^0 \varphi_\ell^\partial = \sum_{i=1}^{N^0} d_{\ell i} \varphi_i. \tag{3.20}$$

This expansion in the definition of the elliptic projection (3.12) yields

$$\sum_{i=1}^{N^0} d_{\ell i} a(\varphi_i, \varphi_j) = a(\varphi_\ell^\partial, \varphi_j) \quad \forall j = 1, 2, \ldots, N^0.$$

Consequently, by (3.3) and (3.5) we can express the coefficients $d_{\ell i}$ in terms of the inverse matrix to the stiffness matrix A as follows

$$d_{\ell i} = \sum_{j=1}^{N^0} (A^{-1})_{ij} A_{j\ell}^\partial.$$

The statement (3.19) follows by substitution of this into (3.20). □

Remark 3.2. The statements (3.15) and (3.19) of Theorems 3.4 and 3.5 can be written in a more compact way using the matrix notation. If the basis functions are arranged into vectors $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_{N^0})^\top$ and $\boldsymbol{\varphi^\partial} = (\varphi_1^\partial, \varphi_2^\partial, \ldots, \varphi_{N^\partial}^\partial)^\top$, then (3.15) and (3.19) can be expressed as

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

Remark 3.3. Formula (3.15) implies that not only $G_{h,\boldsymbol{y}} = G_h(\cdot, \boldsymbol{y}) \in V_h$ for all $\boldsymbol{y} \in \Omega$ but also that $G_{h,\boldsymbol{x}} = G_h(\boldsymbol{x}, \cdot) \in V_h$ for all $\boldsymbol{x} \in \Omega$.

Theorems 3.3–3.5 represent the general concept for investigation of the DMP in the FEM. Theorem 3.3 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 3.4 and 3.5 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. This is treated in Chapter 4. In the case of the higher-order FEM, not only the matrices but also the basis functions play a crucial role as it will be presented in Chapter 5.

CHAPTER

FOUR

Survey of discrete maximum principles for the lowest-order finite elements

This chapter provides a survey of the discrete maximum principle (DMP) results for problem (2.1)–(2.3) discretized by the lowest-order finite elements. This case covers the most often used approximations of the solution u, namely the continuous and piecewise linear approximation on simplices and the continuous and multilinear approximation on blocks (Cartesian products of intervals). The nonnegativity of such an approximation in a domain $\Omega \subset \mathbb{R}^d$ is equivalent to the nonnegativity of its nodal values. This is a fundamental property which makes the analysis of the DMP much simpler for the lowest-order finite elements in comparison with the higher-order finite elements.

The DMP for the lowest-order finite elements is already studied for several decades. The first DMP results in the context of the FEM appeared in 1970s, see [18, 63]. Later, other publications appeared [22, 70, 77] etc. This chapter summarizes the known DMP results in a unified way, using the general concept developed above.

The DMP results in the case of linear finite elements are based on several statements from the matrix theory which are presented in Section 4.1. At first, general DMP results are summarized in Section 4.2. The subsequent statements are based on these general results. Section 4.3 completely characterizes the DMP for problem (2.1)–(2.3) in one dimension. Section 4.4 introduces the two- and higher-dimensional cases. Section 4.5 provides general conditions for the validity of the DMP on simplicial meshes in any dimension higher than one. Section 4.6 attempts the same for the case of block-meshes. In this case, however, the sufficient conditions for the DMP have to be investigated individually for dimension two, three, and higher. See Subsections 4.6.1, 4.6.2, and 4.6.3, respectively. Two artificial examples showing the validity of the DMP on block-meshes in extreme

cases are described in Subsection 4.6.4. Section 4.7 presents the result obtain for prismatic meshes. Finally, Section 4.8 mentions various generalization of the standard results.

4.1 Selected results from the matrix theory

As we will see in Theorem 4.4 below, the analysis of the DMP is based on the nonnegative and monotone matrices. We recall that a real matrix A is said to be *nonnegative* if all its entries are nonnegative and it is denoted by inequality $A \geq 0$, i.e. this inequality is understood componentwise. Similarly, we use $A \leq 0$ for nonpositive matrices. A 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.

Definition 4.1. 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).

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, see (3.4), 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 [32, Thm. 5.1, p. 114]. □

Let us note that Theorem 4.1 is a generalization of the well-known result of Varga [78, 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 [32], 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 $M \subset \{1, 2, \ldots, N\}$ 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, see (3.4). 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 were singular then there would exist a vector $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{x} \neq \mathbf{0}$ such that $A\boldsymbol{x} = \mathbf{0}$. Thus, $\boldsymbol{x}^\top A\boldsymbol{x} = 0$ contradicts the assumption of the lemma.

(b) Let us consider $\lambda \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{x} \neq \mathbf{0}$ 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 $\lambda_1, \lambda_2, \ldots, \lambda_N$ be all eigenvalues of A, (some of them may coincide, depending on their multiplicity). If the eigenvalue λ_i, $i = 1, 2, \ldots, N$, is real, then $\lambda_i > 0$ by (b). The complex eigenvalues appear in pairs with their complex conjugate, i.e. if λ_i is complex then there exists $j \in \{1, 2, \ldots, N\}$ such that $\lambda_j = \overline{\lambda}_i$. Hence, $\lambda_i\lambda_j \geq 0$. Since $\det A = \lambda_1\lambda_2 \ldots \lambda_N$, we obtain $\det A \geq 0$ and by (a) we have $\det A > 0$.

(d) Let $\emptyset \neq M \subset \{1, 2, \ldots, N\}$, 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 $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{y} \neq \mathbf{0}$ 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 $\boldsymbol{x} = A^{-1}\boldsymbol{y} \neq \mathbf{0}$. 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 Lemma 4.2 we see that any real eigenvalue of A is positive. Thus, 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 $A_{i,i+1}$ can be expressed as

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

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

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

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 $b_i \leq 0$. □

4.2 General framework

The general results about the DMP and the DGF described in Sections 3.2–3.4 can be well used for the lowest-order finite elements. Even more, the above mentioned advantageous property of the lowest-order finite elements enables to refine the general results presented above. To formalize the advantageous property, we assume the same notation as in Section 3.1. We consider the finite dimensional spaces $X_h = V_h \oplus V_h^\partial$, with $N^0 = \dim V_h$, $N^\partial = \dim V_h^\partial$ and with a basis $\varphi_1, \varphi_2, \ldots, \varphi_{N^0}$ of V_h and with a basis $\varphi_1^\partial, \varphi_2^\partial, \ldots, \varphi_{N^\partial}^\partial$ of V_h^∂. For these basis functions we assume the following properties

$$\sum_{i=1}^{N^0} c_i \varphi_i(\boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in \Omega \quad \Leftrightarrow \quad c_i \geq 0 \quad \forall i = 1, 2, \ldots, N^0, \tag{4.1}$$

$$\sum_{\ell=1}^{N^\partial} c_\ell^\partial \varphi_\ell^\partial(\boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in \Omega \quad \Leftrightarrow \quad c_\ell^\partial \geq 0 \quad \forall \ell = 1, 2, \ldots, N^\partial. \tag{4.2}$$

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.

The special properties (4.1) and (4.2) enable to reformulate the general result stated in Theorem 3.3. In the lowest-order case the role of the DGF is played by the inverse of the stiffness matrix A.

Theorem 4.4. *Let the finite dimensional spaces V_h and V_h^∂ possess basis functions $\varphi_1, \varphi_2, \ldots, \varphi_{N^0}$ and $\varphi_1^\partial, \varphi_2^\partial, \ldots, \varphi_{N^\partial}^\partial$ with properties (4.1) and (4.2). Then problem (3.2) 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 matrices A and A^∂ are defined in (3.3) and (3.5).

Proof. The proof follows from Theorems 3.3–3.5 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 $g_{\mathrm{D},h} - \Pi_h^0 g_{\mathrm{D},h}$ is nonnegative for all $g_{\mathrm{D},h} \geq 0$, $g_{\mathrm{D},h} \in V_h^\partial$ if and only if $-A^{-1}A^\partial \geq 0$.

The equivalence (i) follows from the expression (3.15) and from the property (4.1). Indeed, the DGF G_h can be expressed as a linear combination of basis functions as follows

$$G_h(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N^0} \gamma_i(\boldsymbol{x}) \varphi_i(\boldsymbol{y}), \quad \text{where } \gamma_i(\boldsymbol{x}) = \sum_{j=1}^{N^0} (A^{-1})_{ij} \varphi_j(\boldsymbol{x}).$$

Hence, property (4.1) 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 $i = 1, 2, \ldots, N^0$ and all $\boldsymbol{x} \in \Omega$. Using the property (4.1) again, we obtain that $\gamma_i(\boldsymbol{x}) \geq 0$ for all $i = 1, 2, \ldots, N^0$ and all $\boldsymbol{x} \in \Omega$ if and only if $(A^{-1})_{ij} \geq 0$ for all $i, j = 1, 2, \ldots, N^0$.

To prove the equivalence (ii) we proceed as follows. According to (4.2) and (3.18), the statement

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

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 (3.19) we can express the difference $\varphi_\ell^\partial - \Pi_h^0 \varphi_\ell^\partial$ as a linear combination $\varphi_\ell^\partial + \sum_{i=1}^{N^0} D_{i\ell}\varphi_i$ with $D_{i\ell} = -\sum_{j=1}^{N^0} (A^{-1})_{ij} A_{j\ell}^\partial$. Such a linear combination is nonnegative by (4.1) and (4.2) if and only if $D_{i\ell} \geq 0$ for all $i = 1, 2, \ldots, N^0$ and $\ell = 1, 2, \ldots, N^\partial$. □

The above theorem provides an equivalent characterization of the DMP by means of the global stiffness matrices. However, a 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 4.5. *Let the finite dimensional spaces V_h and V_h^∂ possess basis functions $\varphi_1, \varphi_2, \ldots, \varphi_{N^0}$ and $\varphi_1^\partial, \varphi_2^\partial, \ldots, \varphi_{N^\partial}^\partial$ with properties (4.1) and (4.2). Let A and A^∂ be the stiffness matrices given by (3.3) and (3.5). If*

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

then problem (3.2) satisfies the discrete conservation of nonnegativity.

Proof. The statement follows immediately from Theorems 4.4 and 4.1, because the stiffness matrix A is positive definite, see (3.4). □

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 formulates a sufficient condition for the DMP in terms of these local matrices.

Theorem 4.6. *Let the finite dimensional spaces V_h and V_h^∂ possess basis functions $\varphi_1, \varphi_2, \ldots, \varphi_{N^0}$ and $\varphi_1^\partial, \varphi_2^\partial, \ldots, \varphi_{N^\partial}^\partial$ with properties (4.1) and (4.2). 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 (3.7) and (3.8). If*

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

then problem (3.2) satisfies the discrete conservation of nonnegativity.

Proof. The statement follows directly from Theorem 4.5 and from (3.9) and (3.10). □

4.3 One dimension

This section concentrates on problem (2.1)–(2.3) in one spatial dimension. In this simple case we succeed to prove a sufficient and necessary condition for the validity of the DMP. Such a result is exceptional, because the usual DMP results provide sufficient conditions only. Furthermore, we are able to find such a condition for the general non-symmetric elliptic problem with general boundary conditions, which is again unusual in the field of the DMP.

Using the special one-dimensional notation, problem (2.1)–(2.3) can be rewritten as follows

$$-(\mathcal{A}u')' + bu' + cu = f \quad \text{in } \Omega, \tag{4.3}$$

$$u = g_\mathrm{D} \quad \text{on } \Gamma_\mathrm{D}, \tag{4.4}$$

$$\alpha u + \mathcal{A}u' n_{1\mathrm{D}} = g_\mathrm{N} \quad \text{on } \Gamma_\mathrm{N}, \tag{4.5}$$

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 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.

The general weak formulation of this problem is presented in Section 2.2, see (2.10). Using the one-dimensional notation, the bilinear form $a(\cdot,\cdot)$ and the right-hand side functional $\mathcal{F}$ can be expressed as

$$a(u,v) = \int_\Omega (\mathcal{A}u'v' + bu'v + cuv)\,\mathrm{d}x + \int_{\Gamma_\mathrm{N}} \alpha uv\,\mathrm{d}s, \tag{4.6}$$

$$\mathcal{F}(v) = \int_\Omega fv\,\mathrm{d}x + \int_{\Gamma_\mathrm{N}} g_\mathrm{N} v\,\mathrm{d}s. \tag{4.7}$$

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

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

Integral over an empty set is understood as zero.

To ensure the correctness of the above setting and also the unique solvability of the corresponding weak formulation, we assume the validity of the one-dimensional analogues of the general requirements introduced in Sections 2.1–2.2. Namely, we assume

$$\mathcal{A} \geq \lambda_{\min} > 0 \text{ in } \Omega, \quad c - \frac{1}{2}b' \geq 0 \text{ in } \Omega, \quad \alpha + \frac{1}{2}b \geq 0 \text{ on } \Gamma_{\rm N}. \tag{4.8}$$

We also assume the V-ellipticity of the bilinear form a, see Lemma 2.1.

To introduce the finite element solution of the one-dimensional problem, 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 $\Gamma_{\rm D}$. It is natural to define the approximate Dirichlet lift $g_{{\rm D},h} \in X_h$ as a function which vanishes at all interior nodes x_i, $i = 1, 2, \ldots, M-1$, and on $\Gamma_{\rm N}$, and which is equal to $g_{\rm D}$ on $\Gamma_{\rm D}$. Thus, such a $g_{{\rm D},h}$ belongs to the complement V_h^∂ of V_h in X_h (the spaces satisfy $X_h = V_h \oplus V_h^\partial$). The general finite element formulation is presented in (3.2). For the reader's convenience, we present this formulation again, but now having in mind the special one-dimensional case. We seek $u_h \in X_h$ such that $u_h = u_h^0 + g_{{\rm D},h}$ and $u_h^0 \in V_h$ satisfies

$$a(u_h^0, v_h) = \mathcal{F}(v_h) - a(g_{{\rm D},h}, v_h) \quad \forall v_h \in V_h, \tag{4.9}$$

where a and $\mathcal{F}$ are given by (4.6)–(4.7).

For the subsequent considerations we introduce the standard finite element basis $\varphi_0, \varphi_1, \ldots, \varphi_M$ of X_h. This basis is uniquely determined by the δ-property $\varphi_j(x_i) = \delta_{ji}$ for $j, i = 0, 1, 2, \ldots, M$, where δ_{ji} stands for Kronecker's tensor. The basis function φ_0 and φ_M corresponding to the end-points a^∂ and b^∂ of Ω belong either to V_h or to V_h^∂ depending on the type of the prescribed boundary condition at the particular point.

Nevertheless, in order to formulate the sufficient and necessary condition for the validity of the DMP we first introduce the following constants on each element

K_k, $k = 1, 2, \ldots, M$:

$$\mathcal{A}_k = \frac{1}{h_k} \int_{K_k} \mathcal{A}(x)\,\mathrm{d}x, \tag{4.10}$$

$$b_k^L = \frac{\int_{K_k} b(x)\varphi_{k-1}(x)\,\mathrm{d}x}{\int_{K_k} \varphi_{k-1}(x)\,\mathrm{d}x} = \frac{2}{h_k} \int_{K_k} b(x)\varphi_{k-1}(x)\,\mathrm{d}x, \tag{4.11}$$

$$b_k^R = \frac{\int_{K_k} b(x)\varphi_k(x)\,\mathrm{d}x}{\int_{K_k} \varphi_k(x)\,\mathrm{d}x} = \frac{2}{h_k} \int_{K_k} b(x)\varphi_k(x)\,\mathrm{d}x, \tag{4.12}$$

$$c_k = \frac{\int_{K_k} c(x)\varphi_{k-1}(x)\varphi_k(x)\,\mathrm{d}x}{\int_{K_k} \varphi_{k-1}(x)\varphi_k(x)\,\mathrm{d}x} = \frac{6}{h_k} \int_{K_k} c(x)\varphi_{k-1}(x)\varphi_k(x)\,\mathrm{d}x. \tag{4.13}$$

Notice that we utilized the facts that

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

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 (4.10)–(4.13) can be used to express the integrals needed for evaluation of the off-diagonal entries of the stiffness matrix:

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

Consequently,

$$a(\varphi_k, \varphi_{k-1}) = -\frac{\mathcal{A}_k}{h_k} + \frac{b_k^L}{2} + c_k\frac{h_k}{6}, \tag{4.14}$$

$$a(\varphi_{k-1}, \varphi_k) = -\frac{\mathcal{A}_k}{h_k} - \frac{b_k^R}{2} + c_k\frac{h_k}{6}. \tag{4.15}$$

We clearly see that both $a(\varphi_k, \varphi_{k-1})$ and $a(\varphi_{k-1}, \varphi_k)$ 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 condition for the validity of the DMP. The precise statement is formulated in the following theorem.

Theorem 4.7. *Let the coefficients of problem (4.3)–(4.5) satisfy (4.8) and let the bilinear form (4.6) be V-elliptic. Then the lowest-order finite element discretization (4.9) satisfies the discrete conservation of nonnegativity if and only if the condition*

$$c_k h_k^2 + 3h_k \max\{b_k^L, -b_k^R\} \leq 6\mathcal{A}_k \tag{4.16}$$

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

Proof. Let $\varphi_1, \varphi_2, \ldots, \varphi_{N^0}$ be the finite element basis functions in V_h. Then the stiffness matrix $A \in \mathbb{R}^{N^0 \times N^0}$ has entries $A_{ij} = a(\varphi_j, \varphi_i)$, $i, j = 1, 2, \ldots, N^0$. Since the bilinear form (4.6) is V elliptic, the stiffness matrix is positive definite, see (3.4). In addition, the matrix A^∂ has the following form provided both end-points a^∂, b^∂ are on Γ_{D}

$$A^\partial = \begin{pmatrix} a(\varphi_1, \varphi_0) & 0 & \ldots & 0 \\ 0 & \ldots & 0 & a(\varphi_M, \varphi_{M-1}) \end{pmatrix}^\top \in \mathbb{R}^{N^0 \times 2}. \tag{4.17}$$

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 (4.16) holds for all $k = 1, 2, \ldots, M$ and if we recall that the off-diagonal entries of A are given by (4.14)–(4.15), then clearly off-diag$(A) \leq 0$. Furthermore, condition (4.16) is satisfied also for elements adjacent to Γ_{D} (for $k = 1$ and/or $k = M$) and, therefore, $A^\partial \leq 0$. Thus, Theorem 4.5 yields the discrete conservation of nonnegativity.

To prove the converse implication we use Theorem 4.4 to obtain that $A^{-1} \geq 0$ and $-A^{-1}A^\partial \geq 0$. Since the stiffness matrix is tridiagonal and it is positive definite (3.4), we conclude by Lemma 4.3 that off-diag$(A) \leq 0$. The nonpositivity of the off-diagonal entries of A yields the validity of the condition (4.16) at least for $k = 2, 3, \ldots, M-1$. If $a^\partial \notin \Gamma_{\mathrm{D}}$ then φ_0 is in V_h and condition (4.16) holds also for $k = 1$. Similarly, if $b^\partial \notin \Gamma_{\mathrm{D}}$ then (4.16) holds also for $k = M$.

However, if $a^\partial \in \Gamma_{\mathrm{D}}$ then $0 \leq (-A^{-1}A^\partial)_{11} = -(A^{-1})_{11} a(\varphi_0, \varphi_1)$, where we use the special structure (4.17) of A^∂. Since $(A^{-1})_{11} > 0$ (see Lemma 4.2), we obtain $a(\varphi_0, \varphi_1) \leq 0$ and consequently, the validity of the condition (4.16) for $k = 1$. Similarly, if $b^\partial \in \Gamma_{\mathrm{D}}$ we obtain (4.16) for $k = M$.

□

Theorem 4.7 states the main result of this section. It is exceptional among the results about the DMP, because it provides an equivalent condition for the DMP. The usual results about the DMP provide sufficient conditions only. In

addition, condition (4.16) is very easy to verify, especially if the coefficients $\mathcal{A}$, b, and c are piecewise constant.

Theorem 4.7 enables us to make several conclusions. For example, if the convection and reaction coefficients b and c vanish, then condition (4.16) 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 (4.16) 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.

Theorem 4.7 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 (4.16) 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 (4.16) enables us 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 (4.16) is trivial to check. However, we have to admit, that condition (4.16) 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 4.8. *Let us assume the hypothesis of Theorem 4.7. If*

$$\overline{c}_k = \operatorname*{ess\,sup}_{x\in K_k} c(x) \quad \textit{and} \quad \overline{b}_k = \operatorname*{ess\,sup}_{x\in K_k} |b(x)|, \quad k = 1, 2, \ldots, M,$$

then the lowest-order finite element discretization (4.9) satisfies the discrete conservation of nonnegativity provided the condition

$$\overline{c}_k h_k^2 + 3h_k \overline{b}_k \leq 6\mathcal{A}_k$$

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

Proof. The statement follows immediately from Theorem 4.7, because $c_k \leq \overline{c}_k$ and $\max\{b_k^L, -b_k^R\} \leq \overline{b}_k$ for all $k = 1, 2, \ldots, M$. □

Transformation to a problem without convection

Interestingly, the general problem (4.3)–(4.5) can be transformed to a problem with vanishing convection coefficient b. It is natural to present this transformation for the classical formulation.

Theorem 4.9. *Let us consider one-dimensional problem (4.3)–(4.5) with coefficients $\mathcal{A} \in C^1(\Omega)$, $b, c, f \in C(\Omega)$ and with $\mathcal{A} > 0$. Then $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ is a classical solution to problem (4.3)–(4.5) if and only if the function u is a classical solution to problem*

$$-(\widehat{\mathcal{A}}u')' + \widehat{c}u = \widehat{f} \quad \text{in } \Omega, \tag{4.18}$$

$$u = g_{\mathrm{D}} \quad \text{on } \Gamma_{\mathrm{D}}, \tag{4.19}$$

$$\widehat{\alpha}u + \widehat{\mathcal{A}}u' n_{1\mathrm{D}} = \widehat{g}_{\mathrm{N}} \quad \text{on } \Gamma_{\mathrm{N}}, \tag{4.20}$$

where

$$\widehat{\mathcal{A}}(x) = \exp\left(\int_0^x \frac{b(t) - \mathcal{A}'(t)}{\mathcal{A}(t)}\,\mathrm{d}t\right), \tag{4.21}$$

$\widehat{c} = c\widehat{\mathcal{A}}/\mathcal{A}$, $\widehat{f} = f\widehat{\mathcal{A}}/\mathcal{A}$, $\widehat{\alpha} = \alpha\widehat{\mathcal{A}}/\mathcal{A}$, *and* $\widehat{g}_{\mathrm{N}} = g_{\mathrm{N}}\widehat{\mathcal{A}}/\mathcal{A}$.

Proof. Differentiating the product $\mathcal{A}u'$ in (4.3) and dividing by the positive number $\mathcal{A}$, allows us to rewrite the equality (4.3) equivalently as

$$-u'' + \frac{b - \mathcal{A}'}{\mathcal{A}}u' + \frac{c}{\mathcal{A}}u = \frac{f}{\mathcal{A}} \quad \text{in } \Omega.$$

Differentiating (4.21) we find out that

$$-\frac{\widehat{\mathcal{A}}'}{\widehat{\mathcal{A}}} = \frac{b - \mathcal{A}'}{\mathcal{A}}.$$

Substituting this into the above equality and multiplying by the positive quantity $\widehat{\mathcal{A}}$, we obtain (4.18). The equivalence of (4.5) with (4.20) follows immediately by multiplication by $\widehat{\mathcal{A}}/\mathcal{A}$. □

Results of Theorem 4.7 can be applied to the transformed problem (4.18)–(4.20) to conclude that a finite element discretization of problem (4.18)–(4.20) satisfies the DMP if and only if

$$\widehat{c}_k h_k^2 \le 6\widehat{\mathcal{A}}_k, \quad \forall k = 1, 2, \ldots, M.$$

However, we point out that in general the finite element solution u_h of the original problem (4.3)–(4.5) differs from the finite element solution $\widehat{u}_h$ of the transformed problem (4.18)–(4.20) even if the same partition of the domain Ω is used.

4.4 Two- and higher-dimensional case

The investigation of the DMP for two- (and higher-) dimensional linear elliptic problems discretized by the lowest-order finite element method is based on Theorems 4.5 and 4.6 which provide sufficient conditions for the validity of the DMP. In contrast to the one-dimensional case, the stiffness matrix is no longer tridiagonal and there is no simple equivalent characterization of monotone stiffness matrices. Therefore, we cannot utilize Theorem 4.4 and we lose the equivalent conditions for the DMP. The monotonicity of the stiffness matrix is most often guaranteed by various sufficient conditions yielding nonpositivity of entries of matrices off-diag(A) and A^{∂}, see Theorem 4.5, or of the local matrices off-diag(A^K) and $A^{\partial,K}$, see Theorem 4.6. These sufficient conditions are usually of a geometrical nature and are specific for particular shapes of the used finite elements.

There are two natural shapes of elements which can be used in arbitrary dimension: simplices and blocks (Cartesian products of intervals). The case of the lowest-order (linear) finite elements on simplices is analyzed in Section 4.5, while the case of the lowest-order (multi-linear) finite elements on blocks is treated in Section 4.6. We will see that these two cases substantially differ from the perspective of the conditions for the discrete maximum principle. While for simplices there exists a universal condition which is valid in arbitrary dimension $d \geq 2$, the conditions for blocks depend substantially on the dimension. For $d = 2$ we have the nonnarrowness condition [14] for rectangles. For $d = 3$ it is possible to satisfy the DMP in exceptional cases, but for $d \geq 4$ it is practically never possible.

Besides simplices and blocks, there are other types of elements specific for the particular dimension. For $d = 3$ the right triangular prisms have certain practical relevance. We analyze the DMP for these prisms in Section 4.7. Another type of practically used elements are pyramids (one rectangular base and four triangular faces). Pyramids are important in hybrid three-dimensional meshes, where the tetrahedral and block meshes have to be joined together face-to-face. This cannot be done without pyramids and triangular prisms, in general. However, pyramidal elements are technically complicated and the DMP on them has not been analyzed yet.

In the sequel, we will analyze the following simplified version of problem (2.1)–(2.3):

$$-\operatorname{div}(\lambda \boldsymbol{\nabla} u) + cu = f \quad \text{in } \Omega, \tag{4.22}$$

$$u = g_{\mathrm{D}} \quad \text{on } \Gamma_{\mathrm{D}}, \tag{4.23}$$

$$\alpha u + \lambda \boldsymbol{\nabla} u \cdot \boldsymbol{n} = g_{\mathrm{N}} \quad \text{on } \Gamma_{\mathrm{N}}. \tag{4.24}$$

In comparison with the general diffusion-convection-reaction problem (2.1)–(2.3), we consider in (4.22)–(4.24) 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 Sections 2.1–2.2. Namely, the assumption (2.5) of the uniform positive definiteness of $\mathcal{A}$ turns into to boundedness of λ from below

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

and assumptions (2.4) simplify to $c \ge 0$ in Ω and $\alpha \ge 0$ on Γ_{N}.

Remark 4.1. Successful approximate solution of the general problem (2.1)–(2.3) with nonvanishing convection coefficient $\boldsymbol{b}$ by the finite element method is a subtle problem, because it requires special stabilization approaches [46, 68]. It is not the goal of this thesis to investigate this case and therefore, we consider $\boldsymbol{b} = \boldsymbol{0}$ in (4.22)–(4.24). The interested reader is referred to [84]. Similarly, the treatment of the general anisotropic tensor $\mathcal{A} \in \mathbb{R}^{d\times d}$ is complicated and we refer to [56] for details.

4.5 Simplicial finite elements

Let us consider the domain $\Omega \subset \mathbb{R}^d$, $d \ge 2$, to be polytopic and to be covered by a polytopic finite element mesh $\mathcal{T}_h$ consisting of d dimensional simplices, see Section 3.1.

We consider a set of all vertices of all simplices in $\mathcal{T}_h$ and we call it a set of nodal points. We distinguish the interior and Newton nodal points $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{N^0}$ lying in $\Omega \cup \Gamma_{\mathrm{N}}$ and the Dirichlet nodal points $\boldsymbol{x}_{N^0+1}, \boldsymbol{x}_{N^0+2}, \dots, \boldsymbol{x}_N$ lying on $\overline{\Gamma}_{\mathrm{D}}$. We recall that Γ_{D} and Γ_{N} are considered as relatively open in $\partial\Omega$. According to the notation of the basis functions, we also put $\boldsymbol{x}_k^\partial = \boldsymbol{x}_{N^0+k}$, $k = 1, 2, \dots, N^\partial$, for the Dirichlet nodal points.

The lowest-order finite element space X_h is 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. In accordance with Section 3.1, 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 $\varphi_1, \varphi_2, \dots, \varphi_{N^0}$ in V_h are uniquely determined by the δ-property

$$\varphi_i(\boldsymbol{x}_j) = \delta_{ij}, \quad i, j = 1, 2, \dots, N^0,$$

where δ_{ij} stands for the Kronecker's tensor and $\boldsymbol{x}_i$, $i = 1, 2, \dots, N^0$, are the interior and Newton nodal points of the mesh $\mathcal{T}_h$. Similarly, the standard finite

element basis functions $\varphi_1^\partial, \varphi_2^\partial, \ldots, \varphi_{N^\partial}^\partial$ in V_h^∂ are uniquely determined by the δ-property

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

where $\boldsymbol{x}_i^\partial$, $i = 1, 2, \ldots, N^\partial$, are the Dirichlet nodal points of $\mathcal{T}_h$.

The general finite element scheme described in Section 3.1 fits well also for the lowest-order case. In particular, the lowest-order finite element solution of problem (4.22)–(4.23) is given as $u_h = u_h^0 + g_{\mathrm{D},h}$ with $u_h^0 \in V_h$ determined by the requirement

$$a(u_h^0, v_h) = \mathcal{F}(v_h) - a(g_{\mathrm{D},h}, v_h) \quad \forall v_h \in V_h, \tag{4.25}$$

where the bilinear form a and the linear functional $\mathcal{F}$ are

$$\begin{aligned} a(u, v) &= \int_\Omega [(\lambda \boldsymbol{\nabla} u) \cdot \boldsymbol{\nabla} v + cuv] \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_\mathrm{N}} \alpha uv \,\mathrm{d}\boldsymbol{s}, \\ \mathcal{F}(v) &= \int_\Omega fv \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_\mathrm{N}} g_\mathrm{N} v \,\mathrm{d}\boldsymbol{s}. \end{aligned} \tag{4.26}$$

From the point of view of the DMP the simplicial finite elements have advantageous properties. Namely, there exist simple formulas for the key integrals used for computation of the entries of the local stiffness matrices. However, in order to present these formulas, we have to introduce certain notation.

Let $K \in \mathcal{T}_h$ be a simplex. We denote its vertices by $\boldsymbol{x}_\ell^K$, $\ell = 1, 2, \ldots, N_K$, $N_K = d+1$. The connection between the vertices of the simplex K and the nodes of the mesh $\mathcal{T}_h$ is provided by the connectivity mapping: $\boldsymbol{x}_\ell^K = \boldsymbol{x}_i$ for $i = \iota_K(\ell)$, $\ell = 1, 2, \ldots, N_K$. 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 [9], 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

$$\int_K \boldsymbol{\nabla} \varphi_\ell^K \cdot \boldsymbol{\nabla} \varphi_m^K \,\mathrm{d}\boldsymbol{x} = \begin{cases} \dfrac{1}{\eta_\ell^2}|K| & \text{for } \ell = m, \\ -\dfrac{\cos(F_\ell, F_m)}{\eta_\ell \eta_m}|K| & \text{for } \ell \neq m, \end{cases} \tag{4.27}$$

$$\int_K \varphi_\ell^K \varphi_m^K \,\mathrm{d}\boldsymbol{x} = \frac{1 + \delta_{\ell m}}{(d+1)(d+2)}|K|, \tag{4.28}$$

where $\ell, m = 1, 2, \ldots, N_K$ and the shape functions $\varphi_\ell^K = \varphi_{\iota_K(\ell)}$ are defined in the simplex K only, they are linear in K, and they vanish at all vertices of K except for $\boldsymbol{x}_\ell^K$, where they have the value 1.

The validity of formula (4.27) can be readily seen from the fact that $\boldsymbol{\nabla}\varphi_\ell^K = -\boldsymbol{n}_\ell/\eta_\ell$. Its proof is published in [7, 84]. The special cases of $d \leq 3$ are well known, see e.g. [50]. The formula (4.28) comes from [17, p. 201], see also [8]. In addition, the equality (4.28) is a special case of the quadrature formula for the barycentric monomials in simplices, see Lemma A.1 in Appendix A.

Now, we can present the basic result about the DMP for problem (4.22)–(4.24). For each element $K \in \mathcal{T}_h$ and for each pair of indices $\ell \neq m$, $\ell, m = 1, 2, \ldots, N_K$, we define the following quantities

$$\lambda^K = \frac{\int_K \lambda(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{|K|}, \quad c_{\ell m}^K = \frac{\int_K c(\boldsymbol{x})\varphi_m^K(\boldsymbol{x})\varphi_\ell^K(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\int_K \varphi_m^K(\boldsymbol{x})\varphi_\ell^K(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}} \tag{4.29}$$

and

$$\alpha_{\ell m}^K = \begin{cases} \dfrac{\int_{\partial K\cap\Gamma_\mathrm{N}} \alpha(\boldsymbol{s})\varphi_m^K(\boldsymbol{s})\varphi_\ell^K(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}}{\int_{\partial K\cap\Gamma_\mathrm{N}} \varphi_m^K(\boldsymbol{s})\varphi_\ell^K(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}} & \text{if } \operatorname{meas}_{d-1}(\partial K \cap \Gamma_\mathrm{N}) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.30}$$

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

Lemma 4.10. *Let $K \in \mathcal{T}_h$ be a d-dimensional simplicial element. Let the local stiffness matrix A^K be given by (3.7) with the bilinear form a defined by (4.26). Then* off-diag $A^K \leq 0$ *if and only if condition*

$$\frac{c_{\ell m}^K}{(d+1)(d+2)}\eta_\ell\eta_m + \frac{\alpha_{\ell m}^K}{d(d+1)}\frac{|\omega_{\ell m}^K|}{|K|}\eta_\ell\eta_m \leq \lambda^K \cos(F_\ell, F_m), \tag{4.31}$$

holds true for all $\ell \neq m$, $\ell, m = 1, 2, \ldots, N_K^0$.

Proof. From (4.27), (4.28), (4.29), and (4.30) we directly compute all the off-diagonal entries of the local stiffness matrix:

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

for all $\ell \neq m$, $\ell, m = 1, 2, \ldots, N_K^0$. □

Here, we recall that $N_K = N_K^0 + N_K^\partial$, where N_K^0 stands for the number of vertices of K lying in $\Omega \cup \Gamma_\mathrm{N}$ and N_K^∂ for the number of vertices of K lying on $\overline{\Gamma}_\mathrm{D}$. This corresponds exactly to the definitions given in Section 3.1.

Lemma 4.11. *Let $K \in \mathcal{T}_h$ be a d-dimensional simplicial element. Let the local stiffness matrix $A^{\partial,K}$ be given by (3.8) with the bilinear form a defined by (4.26). Then $A^{\partial,K} \leq 0$ if and only if condition (4.31) holds for all $\ell = 1, 2, \ldots, N_K^0$ and $m = N_K^0 + 1, N_K^0 + 2, \ldots, N_K$.*

Proof. The proof follows the same steps as the proof of Lemma 4.10. □

Corollary 4.12. *Let us consider the lowest-order simplicial finite element discretization (4.25) of problem (4.22)–(4.24) as described above. If the condition (4.31) is satisfied for all simplices $K \in \mathcal{T}_h$ and all indices $\ell \neq m$, $\ell = 1, 2, \ldots, N_K^0$ and $m = 1, 2, \ldots, N_K$, then problem (4.25) satisfies the discrete conservation of nonnegativity.*

Proof. The statement follows immediately from Theorem 4.6 and Lemmas 4.10 and 4.11. □

Corollary 4.12 represents the main result of this section. It gives a sufficient condition for the validity of the discrete conservation of nonnegativity and hence also for the validity of the DMP, see Theorem 3.1. This result generalizes the standard results and especially the result [9] 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 4.10 and 4.11 show both sufficient and necessary conditions for the proper sign properties of the local matrices, while in the literature usually sufficient conditions only are presented.

In case of the Poisson problem with mixed Dirichlet and Neumann boundary conditions ($c = 0$, $\alpha = 0$), the crucial condition (4.31) reduces to

$$\cos(F_\ell, F_m) \geq 0. \tag{4.32}$$

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 (4.31) simplifies to the condition derived in [9]. However, here we extend its validity also for Neumann type boundary conditions.

Practically, condition (4.31) 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 $\ell, m = 1, 2, \ldots, N_K$. 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 their subsequent utilization in (4.31) might not be

practical. If this is the case, we can recommend to compute the maximal value of c and α on each element $K \in \mathcal{T}_h$:

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

and use the following lemma.

Lemma 4.13. *Under the assumptions of Corollary 4.12, problem (4.25) satisfies the discrete conservation of nonnegativity if*

$$\frac{\overline{c}^K}{(d+1)(d+2)}\eta_\ell\eta_m + \frac{\overline{\alpha}^K}{d(d+1)}\frac{|\omega_{\ell m}^K|}{|K|}\eta_\ell\eta_m \le \lambda^K \cos(F_\ell, F_m) \tag{4.33}$$

holds true for all $\ell \neq m$, $\ell = 1, 2, \ldots, N_K^0$, $m = 1, 2, \ldots, N_K$.

Proof. The statement follows immediately from Corollary 4.12, because $c_{\ell m}^K \le \overline{c}^K$ and $\alpha_{\ell m}^K \le \overline{\alpha}^K$ for all $K \in \mathcal{T}_h$. □

Remark 4.2. The validity of the DMP on simplicial meshes requires at least the nonobtuseness conditions (4.32). However, construction of nonobtuse simplicial meshes might be complicated especially in higher dimensions.

If the Hadwiger conjecture is valid then any polytope in $\mathbb{R}^d$ can be partitioned into nonobtuse simplices (all dihedral angles are at most $\pi/2$) [6]. The Hadwiger conjecture is known to be valid for $d \le 5$ and, thus, for $d \le 5$ we have a guarantee of the existence of a nonobtuse simplicial partion of any polytope. However, this partition is not face-to-face in general. The existence of a face-to-face partition of any polytope into nonobtuse simplices is an open problem even in three dimensions.

Moreover, if c or α do not vanish then condition (4.31) 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 even more problematic. A face-to-face acute simplicial partition of the space $\mathbb{R}^d$ for $d \ge 5$ does not exists [49]. 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 [25] and a cube [76] was successfully constructed quite recently. On the other hand, an acute triangulation of any two-dimensional polygon can always be constructed [12, 58, 86].

4.6 Block finite elements

In this section, we analyze the discrete maximum principle for the finite element formulation (4.25) of problem (2.1)–(2.3) on block finite elements. To employ

the special Cartesian product structure of the used elements, we assume in this section that the coefficients λ, c, and α are in the following product form:

$$\lambda(\boldsymbol{x}) = \prod_{k=1}^{d} \lambda_k(x_k), \quad c(\boldsymbol{x}) = \prod_{k=1}^{d} c_k(x_k), \quad \text{and} \quad \alpha(\boldsymbol{x}) = \prod_{k=1}^{d} \alpha_k(x_k), \tag{4.34}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$.

Further, let the domain $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be partitioned into a finite element mesh $\mathcal{T}_h$ consisting of blocks K (Cartesian products of intervals). We assume the mesh $\mathcal{T}_h$ to satisfy the requirements $(\mathcal{T}1)$–$(\mathcal{T}7)$ from Section 3.1 and in the analogy with the previous section we consider the notion of nodes for the vertices of blocks in $\mathcal{T}_h$. Further, we consider the lowest-order finite element space

$$X_h = \{v_h \in H^1(\Omega) : v_h|_K \in \mathbb{Q}^1(K) \ \forall K \in \mathcal{T}_h\},$$

where $\mathbb{Q}^1(K)$ stands for the space of the multilinear functions on the block K. As above, we consider the standard finite element basis functions of X_h. These are uniquely determined by the requirement $\varphi_i(\boldsymbol{x}_j) = \delta_{ij}$, $i, j = 1, 2, \ldots, N$, where δ_{ij} stands for the Kronecker's tensor, $N = \dim X_h$, and $\boldsymbol{x}_j$ are the nodes of the mesh $\mathcal{T}_h$.

Each block $K \in \mathcal{T}_h$ is a Cartesian product of intervals, i.e. $K = I_1 \times I_2 \times \cdots \times I_d$ and $I_k = [z_k^0, z_k^1]$. We denote by $h_k = z_k^1 - z_k^0$ the length of I_k for all $k = 1, 2, \ldots, d$ and by $|K| = h_1 h_2 \cdots h_d$ the volume of K. On each interval I_k, $k = 1, 2, \ldots, d$, we consider a pair of linear functions

$$\ell_k^0(x) = \frac{z_k^1 - x}{h_k} \quad \text{and} \quad \ell_k^1(x) = \frac{x - z_k^0}{h_k}, \quad x \in I_k.$$

Clearly, $\ell_k^j(z_k^i) = \delta_{ij}$ for $i, j = 0, 1$.

The 2^d vertices of the d-dimensional block K are $\boldsymbol{z}_{\boldsymbol{j}}^K = (z_1^{j_1}, z_2^{j_2}, \ldots, z_d^{j_d})$, where the elements of the binary multiindex $\boldsymbol{j} = (j_1, j_2, \ldots, j_d)$ are zeros and ones only, i.e. $j_k \in \{0, 1\}$, $k = 1, 2, \ldots, d$. Each vertex $\boldsymbol{z}_{\boldsymbol{j}}^K$ of the block K corresponds to a shape function $\varphi_{\boldsymbol{j}}^K$ defined as

$$\varphi_{\boldsymbol{j}}^K(\boldsymbol{x}) = \prod_{k=1}^{d} \ell_k^{j_k}(x_k), \quad \text{where } \boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in K \tag{4.35}$$

and $\boldsymbol{j}$ is a binary multiindex. Clearly, $\varphi_{\boldsymbol{j}}^K(\boldsymbol{z}_{\boldsymbol{i}}^K) = \delta_{\boldsymbol{ij}}$, where $\boldsymbol{i}$ and $\boldsymbol{j}$ are binary multiindices and $\delta_{\boldsymbol{ij}} = 1$ if $\boldsymbol{i} = \boldsymbol{j}$ and $\delta_{\boldsymbol{ij}} = 0$ otherwise. The connection between the shape functions $\varphi_{\boldsymbol{j}}^K$ and the basis functions φ_i, $i = 1, 2, \ldots, N$, is straightforward: if the node $\boldsymbol{x}_i$ of the partition $\mathcal{T}_h$ is a vertex of K then K lies in the support of φ_i and $\varphi_i|_K = \varphi_{\boldsymbol{j}}^K$, where i and $\boldsymbol{j}$ are such that $\boldsymbol{x}_i = \boldsymbol{z}_{\boldsymbol{j}}^K$.

Further, we will compute the crucial integrals needed in the local stiffness matrices (3.7) and (3.8) in terms of the following moments of the coefficients λ, c, and α:

$$\overline{\lambda}^{K,k} = \frac{\int_{I_k} \lambda_k(x)\,\mathrm{d}x}{h_k}, \qquad \lambda_{ij}^{K,k} = \frac{\int_{I_k} \lambda_k(x)\ell_k^i(x)\ell_k^j(x)\,\mathrm{d}x}{\int_{I_k} \ell_k^i(x)\ell_k^j(x)\,\mathrm{d}x}, \tag{4.36}$$
$$c_{ij}^{K,k} = \frac{\int_{I_k} c_k(x)\ell_k^i(x)\ell_k^j(x)\,\mathrm{d}x}{\int_{I_k} \ell_k^i(x)\ell_k^j(x)\,\mathrm{d}x}, \quad \alpha_{ij}^{K,k} = \frac{\int_{I_k} \alpha_k(x)\ell_k^i(x)\ell_k^j(x)\,\mathrm{d}x}{\int_{I_k} \ell_k^i(x)\ell_k^j(x)\,\mathrm{d}x},$$

where $K \in \mathcal{T}_h$, $k = 1, 2, \ldots, d$, and $i, j = 0, 1$. Notice the symmetries $\lambda_{ij}^{K,k} = \lambda_{ji}^{K,k}$, $c_{ij}^{K,k} = c_{ji}^{K,k}$, and $\alpha_{ij}^{K,k} = \alpha_{ji}^{K,k}$ for all $k = 1, 2, \ldots, d$ and $i, j = 0, 1$.

We point out that by (4.28) the integrals in denominators are

$$\int_{I_k} \ell_k^i(x)\ell_k^j(x)\,\mathrm{d}x = \frac{h_k}{6}(1 + \delta_{ij}) \tag{4.37}$$

and, thus, formulas (4.36) can be simplified. For further reference we also define the following quantities

$$\widetilde{\lambda}_{\boldsymbol{ij}}^{K,n} = \overline{\lambda}^{K,n} \prod_{\substack{k=1 \\ k\neq n}}^{d} \lambda_{i_k j_k}^{K,k}, \quad \widetilde{c}_{\boldsymbol{ij}}^{K} = \prod_{k=1}^{d} c_{i_k j_k}^{K,k}, \quad \text{and} \quad \widetilde{\alpha}_{\boldsymbol{ij}}^{K,n,\ell} = \alpha_n(z_n^\ell) \prod_{\substack{k=1 \\ k\neq n}}^{d} \alpha_{i_k j_k}^{K,k}$$

for binary multiindices $\boldsymbol{i}$, $\boldsymbol{j}$, for $n = 1, 2, \ldots, d$, and for $\ell = 0, 1$. Further, we recall that the local bilinear form a_K corresponding to (4.26) is in the context of block finite elements given by

$$a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K) = \int_K \lambda \boldsymbol{\nabla}\varphi_{\boldsymbol{i}}^K \cdot \boldsymbol{\nabla}\varphi_{\boldsymbol{j}}^K\,\mathrm{d}\boldsymbol{x} + \int_K c\varphi_{\boldsymbol{i}}^K\varphi_{\boldsymbol{j}}^K\,\mathrm{d}\boldsymbol{x} + \int_{\partial K\cap\Gamma_{\mathrm{N}}} \alpha\varphi_{\boldsymbol{i}}^K\varphi_{\boldsymbol{j}}^K\,\mathrm{d}\boldsymbol{s}$$

for suitable binary multiindices $\boldsymbol{i}$ and $\boldsymbol{j}$. Since we trivially compute $(\ell_k^i)' = (-1)^i/h_k$ for all $i = 0, 1$ and $k = 1, 2, \ldots, d$, we can express the partial derivatives of $\varphi_{\boldsymbol{j}}^K$ as

$$\frac{\partial\varphi_{\boldsymbol{j}}^K}{\partial x_n}(x_1, x_2, \ldots, x_d) = \frac{(-1)^{j_n}}{h_n} \prod_{\substack{k=1 \\ k\neq n}}^{d} \ell_k^{j_k}(x_k), \quad n = 1, 2, \ldots, d.$$

Consequently,

$$\begin{aligned}\int_K \lambda \boldsymbol{\nabla}\varphi_{\boldsymbol{i}}^K \cdot \boldsymbol{\nabla}\varphi_{\boldsymbol{j}}^K \,\mathrm{d}\boldsymbol{x} &= \int_K \left(\prod_{k=1}^d \lambda_k(x_k)\right) \sum_{n=1}^d \frac{(-1)^{i_n}}{h_n}\frac{(-1)^{j_n}}{h_n} \prod_{\substack{k=1\\k\neq n}}^d \ell_k^{i_k}(x_k)\ell_k^{j_k}(x_k)\,\mathrm{d}\boldsymbol{x}\\ &= \sum_{n=1}^d \frac{(-1)^{i_n+j_n}}{h_n^2}\overline{\lambda}^{K,n} h_n \prod_{\substack{k=1\\k\neq n}}^d \int_{I_k} \lambda_k(x_k)\ell_k^{i_k}(x_k)\ell_k^{j_k}(x_k)\,\mathrm{d}x_k\\ &= \frac{|K|}{6^{d-1}} \sum_{n=1}^d \frac{(-1)^{i_n+j_n}}{h_n^2}\overline{\lambda}^{K,n} \prod_{\substack{k=1\\k\neq n}}^d \lambda_{i_k j_k}^{K,k}(1+\delta_{i_k j_k})\\ &= \frac{|K|}{6^{d-1}} \left(\prod_{k=1}^d (1+\delta_{i_k j_k})\right) \sum_{n=1}^d \frac{(-1)^{i_n+j_n}}{h_n^2}\widetilde{\lambda}_{\boldsymbol{ij}}^{K,n} \frac{1}{1+\delta_{i_n j_n}}, \quad (4.38)\end{aligned}$$

where we use (4.37). Similarly, we can compute

$$\int_K c\varphi_{\boldsymbol{i}}^K \varphi_{\boldsymbol{j}}^K \,\mathrm{d}\boldsymbol{x} = \prod_{k=1}^d \int_{I_k} c(x_k)\ell_k^{i_k}(x_k)\ell_k^{j_k}(x_k)\,\mathrm{d}x_k = \frac{|K|}{6^d}\widetilde{c}_{\boldsymbol{ij}}^K \prod_{k=1}^d (1+\delta_{i_k j_k}). \quad (4.39)$$

In order to express the integral coming from the Newton boundary condition, we have to introduce a suitable notation for the faces of the block $K \in \mathcal{T}_h$. If $K = I_1 \times I_2 \times \cdots \times I_d$ with $I_n = [z_n^0, z_n^1]$, $n = 1, 2, \ldots, d$, then its faces can be expressed as $F_{(n)}^\ell = I_1 \times \cdots \times I_{n-1} \times \{z_n^\ell\} \times I_{n+1} \times \cdots \times I_d$, where $n = 1, 2, \ldots, d$ and $\ell = 0, 1$. Further, we define the indicator of the Newton type boundary: $\omega_\ell^{K,n} = 1$ if the face $F_{(n)}^\ell$ lies on Γ_{N} and $\omega_\ell^{K,n} = 0$ otherwise. This helps us to express the boundary integral in a_K as follows

$$\begin{aligned}\int_{\partial K\cap\Gamma_{\mathrm{N}}} \alpha\varphi_{\boldsymbol{i}}^K \varphi_{\boldsymbol{j}}^K \,\mathrm{d}\boldsymbol{s} &= \sum_{n=1}^d \sum_{\ell=0}^1 \omega_\ell^{K,n} \int_{F_{(n)}^\ell} \alpha\varphi_{\boldsymbol{i}}^K \varphi_{\boldsymbol{j}}^K \,\mathrm{d}\boldsymbol{s}\\ &= \sum_{n=1}^d \sum_{\ell=0}^1 \omega_\ell^{K,n} \alpha_n(z_n^\ell)\ell_n^{i_n}(z_n^\ell)\ell_n^{j_n}(z_n^\ell) \prod_{\substack{k=1\\k\neq n}}^d \int_{I_k} \alpha_k(x_k)\ell_k^{i_k}(x_k)\ell_k^{j_k}(x_k)\,\mathrm{d}x_k\\ &= \sum_{n=1}^d \sum_{\ell=0}^1 \omega_\ell^{K,n} \alpha_n(z_n^\ell)\delta_{i_n\ell}\delta_{j_n\ell}\frac{1}{h_n}\frac{|K|}{6^{d-1}} \prod_{\substack{k=1\\k\neq n}}^d \alpha_{i_k j_k}^{K,k}(1+\delta_{i_k j_k})\\ &= \frac{|K|}{6^{d-1}} \left(\prod_{k=1}^d (1+\delta_{i_k j_k})\right) \sum_{n=1}^d \omega_{i_n}^{K,n} \frac{\delta_{i_n j_n}}{2h_n}\widetilde{\alpha}_{\boldsymbol{ij}}^{K,n,i_n}, \quad (4.40)\end{aligned}$$

where we used in the last step the fact that if $\delta_{i_n j_n} \neq 0$ then $1/(1+\delta_{i_n j_n}) = 1/2$.

Formulas (4.38)–(4.40) enable us to characterize the nonpositivity of the entries of the local stiffness matrix off-diag A^K and $A^{K,\partial}$, see (3.7) and (3.8). This characterization, however, depends strongly on the dimension d.

4.6.1 Dimension two

We first investigate the case $d = 2$. In order to formulate the following lemma, we introduce a suitable notation. For $d = 2$ and $n = 1, 2$ we define the quantity

$$B_n = \frac{1}{2\overline{\lambda}^K_{\overline{n}} \min\{\lambda^{K,n}_{00}, \lambda^{K,n}_{11}\}} \left[\overline{\lambda}^K_n \lambda^{K,\overline{n}}_{01} + \frac{1}{3} c^{K,\overline{n}}_{01} \max\{c^{K,n}_{00}, c^{K,n}_{11}\} h_n^2 \right. \\ \left. + \max\{\omega^{K,n}_0 \alpha_n(z^0_n), \omega^{K,n}_1 \alpha_n(z^1_n)\} \alpha^{K,\overline{n}}_{01} h_n \right], \quad (4.41)$$

where $\overline{n} = 3 - n$ has the opposite value than n, i.e. if $n = 1$ then $\overline{n} = 2$ and if $n = 2$ then $\overline{n} = 1$.

Lemma 4.14. *Let $d = 2$ and let the coefficients λ, c and α be in the form (4.34). Let $K \in \mathcal{T}_h$, $K = I_1 \times I_2$, be a rectangular element and let h_1 and h_2 stand for the lengths of I_1 and I_2, respectively. Then $a_K(\varphi^K_{\boldsymbol{j}}, \varphi^K_{\boldsymbol{i}}) \leq 0$ for all 2-dimensional binary multiindices $\boldsymbol{i} \neq \boldsymbol{j}$ if and only if*

$$B_1 \leq \frac{h_1^2}{h_2^2} \leq \frac{1}{B_2} \quad (4.42)$$

and

$$\frac{1}{6} c^{K,1}_{01} c^{K,2}_{01} \leq \frac{\overline{\lambda}^K_1 \lambda^{K,2}_{01}}{h_1^2} + \frac{\overline{\lambda}^K_2 \lambda^{K,1}_{01}}{h_2^2}. \quad (4.43)$$

Proof. Using (4.38)–(4.40) with $d = 2$, we express the value $a_K(\varphi^K_{\boldsymbol{j}}, \varphi^K_{\boldsymbol{i}})$ as

$$a_K(\varphi^K_{\boldsymbol{j}}, \varphi^K_{\boldsymbol{i}}) = \frac{|K|}{6}(1+\delta_{i_1 j_1})(1+\delta_{i_2 j_2}) \times \\ \left[\frac{(-1)^{i_1+j_1}}{(1+\delta_{i_1 j_1}) h_1^2} \overline{\lambda}^K_1 \lambda^{K,2}_{i_2 j_2} + \frac{(-1)^{i_2+j_2}}{(1+\delta_{i_2 j_2}) h_2^2} \overline{\lambda}^K_2 \lambda^{K,1}_{i_1 j_1} + \frac{1}{6} c^{K,1}_{i_1 j_1} c^{K,2}_{i_2 j_2} \right. \\ \left. + \omega^{K,1}_{i_1} \frac{\delta_{i_1 j_1}}{2h_1} \alpha_1(z^{i_1}_1) \alpha^{K,2}_{i_2 j_2} + \omega^{K,2}_{i_2} \frac{\delta_{i_2 j_2}}{2h_2} \alpha_2(z^{i_2}_2) \alpha^{K,1}_{i_1 j_1} \right].$$

By the direct examination, we obtain that the two values of $a_K(\varphi^K_{\boldsymbol{j}}, \varphi^K_{\boldsymbol{i}})$ given by $\boldsymbol{i} = (0,0)$, $\boldsymbol{j} = (1,0)$ and $\boldsymbol{i} = (1,1)$, $\boldsymbol{j} = (0,1)$ are nonpositive if and only if $B_1 \leq h_1^2/h_2^2$. Similarly, these values for $\boldsymbol{i} = (1,0)$, $\boldsymbol{j} = (1,1)$ and $\boldsymbol{i} = (0,0)$,

$\boldsymbol{j} = (0,1)$ are nonpositive if and only if $h_1^2/h_2^2 \le 1/B_2$. Finally, these values for $\boldsymbol{i} = (0,0)$, $\boldsymbol{j} = (1,1)$ and $\boldsymbol{i} = (1,1)$, $\boldsymbol{j} = (0,0)$ are identical and they are nonpositive if and only if condition (4.43) holds true. The other combination of indices $\boldsymbol{i}$ and $\boldsymbol{j}$ coincide with one of the previous cases, because of the symmetry $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K) = a_K(\varphi_{\boldsymbol{i}}^K, \varphi_{\boldsymbol{j}}^K)$.

Hence, we conclude that all values $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K)$ for $\boldsymbol{i} \neq \boldsymbol{j}$ are nonpositive if and only if conditions (4.42) and (4.43) hold true. □

Lemma 4.15. *Let us consider all the assumptions of Lemma 4.14. If the coefficients λ, c, and α are piecewise constant, i.e. if $\lambda(\boldsymbol{x}) = \lambda_K$, $c(\boldsymbol{x}) = c_K$, and $\alpha(\boldsymbol{x}) = \alpha_K$ for all $\boldsymbol{x} \in K$, then*

$$B_n = B_n^{const} = \frac{1}{2} + \frac{1}{6}\frac{c_K}{\lambda_K} h_n^2 + \frac{1}{2}\max\{\omega_0^{K,n}, \omega_1^{K,n}\}\frac{\alpha_K}{\lambda_K} h_n, \quad n = 1, 2, \tag{4.44}$$

and the values $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K)$ for all $\boldsymbol{i} \neq \boldsymbol{j}$ are nonpositive if and only if

$$B_1^{const} \le \frac{h_1^2}{h_2^2} \le \frac{1}{B_2^{const}}. \tag{4.45}$$

Moreover, if λ is piecewise constant and $c = 0$ and $\alpha = 0$, then the values $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K)$ for all $\boldsymbol{i} \neq \boldsymbol{j}$ are nonpositive if and only if

$$\frac{1}{2} \le \frac{h_1^2}{h_2^2} \le 2. \tag{4.46}$$

Proof. If the coefficients λ, c, and α are piecewise constant, then it is straightforward that formula (4.41) reduces to (4.44). The sufficiency and the necessity of conditions (4.45) comes from Lemma 4.14 and from the fact that (4.43) follows from (4.42) in the case of piecewise constant coefficients. Indeed, if $B_1^{\text{const}} \le h_1^2/h_2^2$ then

$$\frac{1}{6}\frac{c_K}{\lambda_K} h_1^2 \le \frac{h_1^2}{h_2^2}$$

and condition (4.43) immediately follows.

Finally, if λ is piecewise constant and $c = 0$ and $\alpha = 0$, then $B_n^{\text{const}} = 1/2$ and (4.45) simplifies to (4.46). □

Corollary 4.16. *Let us consider problem (4.22)–(4.24) for $d = 2$ discretized by the rectangular finite elements with the coefficients in the form (4.34). If conditions (4.42) and (4.43) are satisfied for all rectangles $K \in \mathcal{T}_h$, then the discretization (4.25) satisfies the discrete maximum principle.*

Proof. The statement follows immediately from Theorem 4.6, Lemma 4.14, the definition of the local stiffness matrices (3.7) and (3.8), and from Theorem 3.1. □

Corollary 4.17. *Let the assumptions of Corollary 4.16 hold true and let the coefficients λ, c, and α be piecewise constant. If condition (4.45) is satisfied for all rectangles $K \in \mathcal{T}_h$, then the discretization (4.25) satisfies the discrete maximum principle.*

In addition, if λ is piecewise constant and $c = 0$ and $\alpha = 0$ and if condition (4.46) is satisfied for all rectangles $K \in \mathcal{T}_h$ then the discretization (4.25) satisfies the discrete maximum principle.

Proof. The statement follows immediately from Theorem 4.6, Lemma 4.15, the definition of the local stiffness matrices (3.7) and (3.8), and from Theorem 3.1. □

Lemmas 4.14 and 4.15 provide sufficient and necessary conditions for the nonpositivity of the contributions to the off-diagonal entries of the local stiffness matrices A^K and $A^{K,\partial}$, while Corollaries 4.16 and 4.17 use them in a straightforward way to formulate sufficient conditions for the validity of the discrete maximum principle. We have to admit that conditions (4.42)–(4.43) are too complicated for any practical utilization. However, in the case of piecewise constant coefficients these conditions considerably simplify, see (4.45). Let us point out that the nonnarrowness condition (4.46) for the validity of the DMP for Poisson problem was derived already in [14].

Similarly as for simplices, conditions (4.42)–(4.43), (4.45), and (4.46) limit the shape (not the size) of the elements. In case of rectangles, these conditions limit the aspect ratio. The rectangles have to be close to the square. We can also clearly observe the general fact that if the coefficients c or α are nonzero, then their effect decreases as the size of the elements decreases.

4.6.2 Dimension three

Let us proceed with the three-dimensional case. Lemma 4.18 and Corollary 4.19 below state that the discrete maximum principle on 3D block finite elements is satisfied only if all the elements are cubes and c and α vanish. However, this statement is true for the piecewise constant coefficient λ, only. In general, if we admit variable λ it is possible under certain circumstances to obtain the nonpositivity of matrices off-diag A^K and $A^{K,\partial}$ and consequently the conservation of nonnegativity. Nevertheless, the special circumstances leading to the conservation of nonnegativity are very artificial with no practical use. Below in Subsection 4.6.4 we present Example 4.1 showing that for any block finite element mesh in any dimension $d \geq 2$ there exists a coefficient λ such that the discrete maximum principle is satisfied.

Lemma 4.18. *Let us consider problem (2.1)–(2.3) and its finite element discretization (4.25) with $d = 3$ and $\mathcal{T}_h$ being a block partition of the domain $\Omega \subset \mathbb{R}^3$.*

Let the coefficient λ be piecewise constant with respect to $\mathcal{T}_h$. Let $K \in \mathcal{T}_h$, $K = I_1 \times I_2 \times I_3$, be a block element and let h_1, h_2, h_3 stand for the lengths of I_1, I_2, I_3, respectively. Finally, let the shape functions $\varphi_{\boldsymbol{i}}^K$ on K be given by (4.35). Then $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K) \le 0$ for all three-dimensional binary multiindices $\boldsymbol{i} \neq \boldsymbol{j}$ if and only if

$$h_1 = h_2 = h_3 \quad \text{and} \quad c = 0 \text{ a.e. in } K \quad \text{and} \quad \alpha = 0 \text{ a.e. on } \partial K \cap \Gamma_{\mathrm{N}}. \tag{4.47}$$

Proof. Let λ^K be the constant value of λ on $K \in \mathcal{T}_h$. If conditions (4.47) are satisfied, then (4.38) implies

$$a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K) = \lambda^K \frac{|K|}{36} \left(\prod_{k=1}^{3} (1 + \delta_{i_k j_k}) \right) \frac{1}{h_1^2} \sum_{n=1}^{3} \frac{(-1)^{i_n + j_n}}{1 + \delta_{i_n j_n}}. \tag{4.48}$$

The term $(-1)^{i_n+j_n}/(1+\delta_{i_n j_n})$ is equal either to -1 if $i_n \neq j_n$ or to $1/2$ if $i_n = j_n$. Since $\boldsymbol{i} \neq \boldsymbol{j}$, there exists $\ell \in \{1, 2, 3\}$ such that $i_\ell \neq j_\ell$ and we can estimate

$$\sum_{n=1}^{3} \frac{(-1)^{i_n + j_n}}{1 + \delta_{i_n j_n}} \le -1 + \sum_{\substack{n=1 \\ n \neq \ell}}^{3} \frac{1}{2} = 0. \tag{4.49}$$

Combination of (4.48) and (4.49) proves nonpositivity of $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K)$ for all multiindices $\boldsymbol{i} \neq \boldsymbol{j}$.

To prove the converse implication we assume that $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K) \le 0$ for all multiindices $\boldsymbol{i} \neq \boldsymbol{j}$. Since the local bilinear form a_K is a sum of the integrals (4.38)–(4.40) and since the integrals (4.39) and (4.40) are nonnegative, the value of (4.38) must be nonpositive. For example, if $\boldsymbol{i} = (0, 0, 0)$ and $\boldsymbol{j} = (1, 0, 0)$ then

$$\int_K \lambda \boldsymbol{\nabla} \varphi_{\boldsymbol{i}}^K \cdot \boldsymbol{\nabla} \varphi_{\boldsymbol{j}}^K \, \mathrm{d}\boldsymbol{x} = \lambda^K \frac{|K|}{9} \left(-\frac{1}{h_1^2} + \frac{1}{2h_2^2} + \frac{1}{2h_3^2} \right). \tag{4.50}$$

Clearly, the integral (4.38) is nonpositive for $\boldsymbol{i} = (0, 0, 0)$ and $\boldsymbol{j} = (1, 0, 0)$, $\boldsymbol{j} = (0, 1, 0)$, $\boldsymbol{j} = (0, 0, 1)$, respectively, only if

$$\begin{aligned} -h_1^{-2} + h_2^{-2}/2 + h_3^{-2}/2 &\le 0, \\ h_1^{-2}/2 - h_2^{-2} + h_3^{-2}/2 &\le 0, \\ h_1^{-2}/2 + h_2^{-2}/2 - h_3^{-2} &\le 0. \end{aligned}$$

The first inequality together with the sum of the second and the third one yields $0 \le h_1^{-2} - h_2^{-2}/2 - h_3^{-2}/2 \le 0$ and, hence, $2h_1^{-2} = h_2^{-2} + h_3^{-2}$. Similarly, we obtain $2h_2^{-2} = h_1^{-2} + h_3^{-2}$ and $2h_3^{-2} = h_1^{-2} + h_2^{-2}$. These three equalities easily imply $h_1 = h_2 = h_3$.

However, if $h_1 = h_2 = h_3$ then the value of (4.50) is zero. Consequently, the nonpositivity of $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K)$ and the nonnegativity of (4.39) and (4.40) yields $c = 0$ a.e. in K and $\alpha = 0$ a.e. on $\partial K \cap \Gamma_{\mathrm{N}}$. The other possible values of multiindices $\boldsymbol{i}$ and $\boldsymbol{j}$ can be treated analogously. □

Corollary 4.19. *Let the coefficient λ be piecewise constant. The discretization (4.25) of problem (2.1)–(2.3) based on the lowest-order block finite elements in three dimensions satisfies the discrete maximum principle provided all elements $K \in \mathcal{T}_h$ are cubes and $c = 0$ a.e. in Ω and $\alpha = 0$ a.e. on Γ_{N}.*

Proof. Lemma 4.18 and Theorem 4.6 yields the conservation of nonnegativity, which is equivalent to the discrete maximum principle due to Theorem 3.1. □

Let us note that the result of Corollary 4.19 was derived already in [45].

4.6.3 Dimensions four and higher

For block elements, in dimensions higher than three, in the case of piecewise constant coefficient λ there are always positive entries in the local stiffness matrices off-diag A^K and $A^{K,\partial}$ – even on hypercubes. This observation was made already in [45]. Below, we formulate this observation in a rigorous way in the context of the general problem (2.1)–(2.3).

Lemma 4.20. *Let us consider problem (2.1)–(2.3) and its finite element discretization (4.25) with $d \geq 4$ and $\mathcal{T}_h$ being a block partition of the domain $\Omega \subset \mathbb{R}^d$. Let the coefficient λ be piecewise constant with respect to $\mathcal{T}_h$ and let the shape functions $\varphi_{\boldsymbol{i}}^K$ on K be given by (4.35). Then for any d-dimensional binary multiindex $\boldsymbol{i}$ there exists another d-dimensional binary multiindices $\boldsymbol{j}$ such that $\boldsymbol{i} \neq \boldsymbol{j}$ and $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K) > 0$.*

Proof. Let $K \in \mathcal{T}_h$, $K = I_1 \times I_2 \times \cdots \times I_d$, be a block element and let $h_1, h_2, \ldots, h_d$ stand for the lengths of $I_1, I_2, \ldots, I_d$, respectively. Without loss of generality we consider $h_1 \geq h_2 \geq \cdots \geq h_d$. Given a d-dimensional binary multiindices $\boldsymbol{i}$, we define the multiindex $\boldsymbol{j}$ as $\boldsymbol{j} = (\bar{i}_1, i_2, \ldots, i_d)$, where $\bar{i}_1 = 1 - i_1$ has the opposite value than i_1, i.e. if $i_1 = 0$ then $\bar{i}_1 = 1$ and if $i_1 = 1$ then $\bar{i}_1 = 0$. Since $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K)$ is a sum of integrals (4.38)–(4.40), we obtain for these $\boldsymbol{i}$ and $\boldsymbol{j}$ the following expression

$$a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K) \geq \int_K \lambda \boldsymbol{\nabla} \varphi_{\boldsymbol{i}}^K \cdot \boldsymbol{\nabla} \varphi_{\boldsymbol{j}}^K \,\mathrm{d}\boldsymbol{x} = \lambda^K \frac{|K|}{3^{d-1}} \left(-\frac{1}{h_1^2} + \sum_{n=2}^{d} \frac{1}{2h_n^2} \right).$$

The positivity of this expression is immediate from the following estimate

$$-\frac{1}{h_1^2} + \sum_{n=2}^{d} \frac{1}{2h_n^2} \geq -\frac{1}{h_1^2} + \frac{d-1}{2h_1^2} = \frac{d-3}{2h_1^2} > 0,$$

where the last inequality holds true for $d \geq 4$. □

A direct consequence of this lemma is that Theorem 4.6 cannot be used to prove the validity of the DMP. Moreover, applying Lemma 4.20 to all elements sharing the longest edge in the block partition $\mathcal{T}_h$, we end up with a positive off-diagonal entry in the global stiffness matrix A and, thus, Theorem 4.5 cannot be employed for the proof of the DMP as well. Subsequently, numerical experiments indicate that the global stiffness matrix A is not monotone for $d \geq 4$ not even on meshes consisting of hypercubes. In view of Theorem 4.4, it seems that the discrete maximum principle is not satisfied for $d \geq 4$ and for piecewise constant coefficient λ on any block finite element mesh.

The numerical experiments leading to this conclusion were published in [A2]. This paper is attached to this thesis as Appendix C.

4.6.4 Artificial examples

We conclude this section by a few examples showing that the discrete maximum principle on block finite elements can be satisfied in certain artificial case.

Example 4.1. Let us consider any block finite element mesh $\mathcal{T}_h$ in arbitrary dimension $d \geq 2$. For this mesh we construct the coefficient λ in such a way that for $c = 0$ and $\alpha = 0$ the discrete maximum principle is satisfied.

Let $K = I_1 \times I_2 \times \cdots \times I_d$ be any element in $\mathcal{T}_h$. On $I_k = [z_k^0, z_k^1]$ with $h_k = z_k^1 - z_k^0$, $k = 1, 2, \ldots, d$, we construct the function $\lambda_k(x)$ as shown in Figure 4.1. Its formal definition is as follows:

$$\lambda_k(x) = \begin{cases} U & \text{for } x \in [z_k^0, z_k^0 + \delta h_k] \cup [z_k^1 - \delta h_k, z_k^1], \\ L & \text{for } x \in (z_k^0 + \delta h_k, z_k^1 - \delta h_k), \end{cases} \tag{4.51}$$

where a sufficiently small value of $\delta \in (0, 1/4)$ will be fixed later and

$$U = \frac{1}{2\delta + \delta^2}, \quad L = \frac{1 - 2\delta U}{1 - 2\delta} = \frac{\delta}{(2 + \delta)(1 - 2\delta)}.$$

These values are chosen in such a way that $\int_{I_k} \lambda_k(x)\,\mathrm{d}x = h_k$ and $L \to 0$ for $\delta \to 0$.

From definition (4.36) we clearly see that

$$\overline{\lambda}^{K,k} = 1.$$

Furthermore, using the facts that

$$\frac{2}{3} < \delta U < \frac{1}{2} \quad \text{and} \quad 0 < L < \delta \quad \text{for } \delta \in (0, 1/4)$$

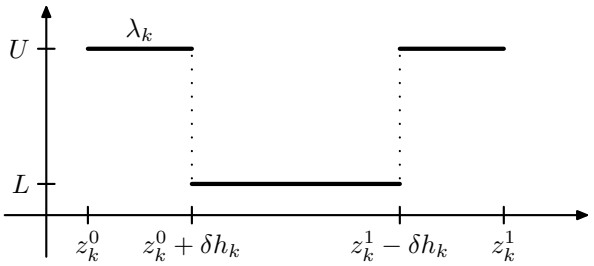


Figure 4.1: The graph of the piecewise constant function λ_k.

together with the values

$$\begin{aligned} \ell_k^0(z_k^0 + \delta h_k) = \ell_k^1(z_k^1 - \delta h_k) &= 1 - \delta, \\ \ell_k^0(z_k^1 - \delta h_k) = \ell_k^1(z_k^0 + \delta h_k) &= \delta, \\ \ell_k^0\left(\frac{z_k^1 + z_k^0}{2}\right) = \ell_k^1\left(\frac{z_k^1 + z_k^0}{2}\right) &= \frac{1}{2}, \end{aligned}$$

we obtain the following estimates

$$1 < \lambda_{00}^{K,k} = \lambda_{11}^{K,k} < 3 \quad \text{and} \quad 3\delta < \lambda_{01}^{K,k} < 8\delta. \tag{4.52}$$

Hence, let us consider two distinct d-dimensional binary multiindices $\boldsymbol{i}$ and $\boldsymbol{j}$ and the corresponding shape functions $\varphi_{\boldsymbol{i}}^K$ and $\varphi_{\boldsymbol{j}}^K$, see (4.35). If $c = 0$ and $\alpha = 0$ then the value of $a_K(\varphi_{\boldsymbol{j}}^K, \varphi_{\boldsymbol{i}}^K)$ is given by (4.38). We consider the sets of indices $\mathcal{P} = \{k : i_k = j_k,\ 1 \le k \le d\}$ and $\mathcal{N} = \{k : i_k \neq j_k,\ 1 \le k \le d\}$ and we denote by $\#\mathcal{P}$ and $\#\mathcal{N}$ the numbers of their elements, respectively. This enables us to express the integral (4.38) as follows

$$\begin{aligned} \int_K \lambda \boldsymbol{\nabla}\varphi_{\boldsymbol{i}}^K \cdot \boldsymbol{\nabla}\varphi_{\boldsymbol{j}}^K \,\mathrm{d}\boldsymbol{x} = \frac{|K|}{6^{d-1}} 2^{\#\mathcal{P}} \Bigg[&\sum_{n\in\mathcal{P}} \frac{1}{2h_n^2} \overline{\lambda}^{K,n} \left(\prod_{k\in\mathcal{P}\setminus\{n\}} \lambda_{00}^{K,k}\right) \left(\prod_{k\in\mathcal{N}} \lambda_{01}^{K,k}\right) \\ &- \sum_{n\in\mathcal{N}} \frac{1}{h_n^2} \overline{\lambda}^{K,n} \left(\prod_{k\in\mathcal{P}} \lambda_{00}^{K,k}\right) \left(\prod_{k\in\mathcal{N}\setminus\{n\}} \lambda_{01}^{K,k}\right) \Bigg], \end{aligned}$$

where we use the symmetric definition (4.51) of $\lambda_k(x)$ which yields $\lambda_{00}^{K,k} = \lambda_{11}^{K,k}$. The estimates (4.52) then lead to

$$\int_K \lambda \boldsymbol{\nabla}\varphi_{\boldsymbol{i}}^K \cdot \boldsymbol{\nabla}\varphi_{\boldsymbol{j}}^K \,\mathrm{d}\boldsymbol{x} \le \delta^{\#\mathcal{N}-1} \left[\frac{3^{\#\mathcal{P}-1} 8^{\#\mathcal{N}}}{2} \delta \sum_{n\in\mathcal{P}} \frac{1}{h_n^2} - 3^{\#\mathcal{N}-1} \sum_{n\in\mathcal{N}} \frac{1}{h_n^2} \right] < 0,$$

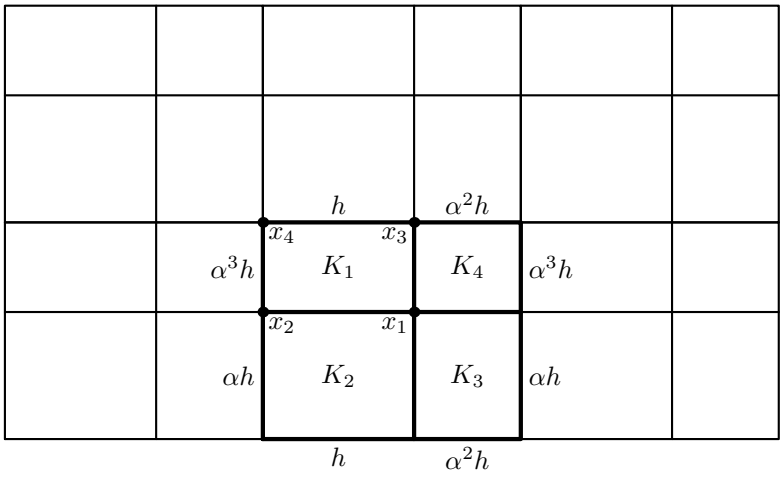


Figure 4.2: A pattern of four rectangles. It can be periodically repeated to produce more complex meshes.

where the last inequality holds true for sufficiently small δ. Hence, for sufficiently small δ we obtain negative entries of the local stiffness matrices off-diag A^K and $A^{K,\partial}$ and Theorems 4.6 and 3.1 yield the discrete maximum principle.

Let us note that this construction can be utilized even in the case of nonvanishing coefficients c and α. However, in this case we have to consider in addition a sufficiently fine uniform refinement of the given block finite element mesh $\mathcal{T}_h$ in order to satisfy the discrete maximum principle.

Example 4.2. In two dimensions, we showed that the discrete maximum principle is satisfied if the rectangular elements are nonnarrow, see Lemma 4.15 and condition (4.46). However, this result is based on the local stiffness matrices and on Theorem 4.6. Using Theorem 4.5 we can show that certain rectangles in the mesh can be more narrow than condition (4.46) admits and the discrete maximum principle still holds.

In order to construct such an example, we will consider the Poisson problem in a domain $\Omega \subset \mathbb{R}^2$, i.e. problem (4.22)–(4.24) with $\lambda = 1$, $c = 0$, $\Gamma_{\mathrm{N}} = \emptyset$, $\Gamma_{\mathrm{D}} = \partial\Omega$, and $g_{\mathrm{D}} = 0$. We assume that the finite element mesh $\mathcal{T}_h$ is constructed by a periodic repetition of the pattern shown in Figure 4.2. The aspect ratio (the ratio of the lengths of sides) of the top-left rectangle K_1 is α^3 and the aspect ratio of the other three rectangles K_2, K_3, K_4 is α. We assume the domain Ω such that it can be covered by this mesh.

The parameter α is considered in $[1/\sqrt{2}, 1]$. This choice guarantees that the matrices off-diag A^K and $A^{K,\partial}$ are nonpositive for elements $K = K_2, \ldots, K_4$, see Lemma 4.15. However, if α is sufficiently small (below $1/\sqrt[6]{2}$) then α^3 is below

$1/\sqrt{2}$ and there is a positive entry in the local stiffness matrices. On the other hand, if α is not too small then this positive contribution to the global stiffness matrix will be overcome by a negative contribution from a neighboring element, the global stiffness matrix will be M-matrix and Theorem 4.5 will guarantee the validity of the DMP.

To prove this, we consider the piecewise bilinear basis functions $\varphi_1, \dots, \varphi_4$ corresponding to vertices $\boldsymbol{x}_1, \dots, \boldsymbol{x}_4$, see Figure 4.2. By (4.38) we have

$$a(\varphi_1, \varphi_4) = \int_{K_1} \boldsymbol{\nabla}\varphi_1 \cdot \boldsymbol{\nabla}\varphi_4 \,\mathrm{d}\boldsymbol{x} = \frac{\alpha^3 h^2}{6}\left(\frac{-1}{h^2} - \frac{1}{\alpha^6 h^2}\right).$$

This is clearly nonpositive. Similarly, for another pair of basis functions we obtain a negative entry in the global stiffness matrix:

$$\begin{aligned} a(\varphi_1, \varphi_3) &= \int_{K_1 \cup K_4} \boldsymbol{\nabla}\varphi_1 \cdot \boldsymbol{\nabla}\varphi_3 \,\mathrm{d}\boldsymbol{x} \\ &= \frac{\alpha^3 h^2}{3}\left(\frac{-1}{\alpha^6 h^2} + \frac{1}{2h^2}\right) + \frac{\alpha^5 h^2}{3}\left(\frac{-1}{\alpha^6 h^2} + \frac{1}{2\alpha^4 h^2}\right) \\ &\le \frac{1}{3}(-1 + \frac{1}{2} - 1 + \frac{1}{2}) = -\frac{1}{3}, \end{aligned}$$

where we use the fact that $\alpha \le 1$. Finally,

$$\begin{aligned} a(\varphi_1, \varphi_2) &= \int_{K_1 \cup K_2} \boldsymbol{\nabla}\varphi_1 \cdot \boldsymbol{\nabla}\varphi_2 \,\mathrm{d}\boldsymbol{x} = \frac{\alpha^3 h^2}{3}\left(\frac{-1}{h^2} + \frac{1}{2\alpha^6 h^2}\right) + \frac{\alpha h^2}{3}\left(\frac{-1}{h^2} + \frac{1}{2\alpha^2 h^2}\right) \\ &= \frac{1}{6\alpha^3}\left(-2\alpha^6 - 2\alpha^4 + \alpha^2 + 1\right). \end{aligned} \tag{4.53}$$

This is nonpositive if and only if

$$-2\alpha^6 - 2\alpha^4 + \alpha^2 + 1 = -\left(\alpha^2 - 1/\sqrt{2}\right)\left(2\alpha^4 + \left(2 + \sqrt{2}\right)\alpha^2 + \sqrt{2}\right) \le 0.$$

The last inequality holds true if and only if $\alpha^2 \in [1/\sqrt{2}, 1]$. Thus, (4.53) is nonpositive for all $\alpha \in [1/\sqrt[4]{2}, 1]$.

The other pairs of basis functions lead to the same values of the already computed ones. Namely,

$$a(\varphi_2, \varphi_3) = a(\varphi_1, \varphi_4), \quad a(\varphi_2, \varphi_4) = a(\varphi_1, \varphi_3), \quad a(\varphi_3, \varphi_4) = a(\varphi_1, \varphi_2).$$

Hence, for $\alpha \in [1/\sqrt[4]{2}, 1]$, the global stiffness matrix is M-matrix and Theorem 4.5 guarantees the validity of the DMP.

To conclude, if $\alpha \in [1/\sqrt[4]{2}, 1/\sqrt[6]{2})$, i.e. approximately $\alpha \in [0.8409, 0.8909)$, then the element K_1 is more narrow than the nonnarrowness condition (4.46)

allows, but the DMP is satisfied. Finally, let us point out that this example is valid for arbitrary choice of the domain and homogeneous Dirichlet boundary conditions provided they are chosen compatibly with the periodic pattern shown in Figure 4.2

4.7 Right triangular prisms

The three-dimensional meshes consisting of right-triangular prisms are useful especially for cylindrical geometries of the computational domains. They are also needed (together with pyramids) in three-dimensional hybrid meshes to join face-to-face tetrahedra and blocks.

The validity of the DMP on prismatic meshes was analyzed in detail in [A1]. This paper is attached to this thesis as Appendix B. For the reader's convenience we present below the main results of this paper.

Let us consider problem (4.22)–(4.24) with $\Gamma_{\mathrm{N}} = \emptyset$, $\Gamma_{\mathrm{D}} = \partial\Omega$, $g_{\mathrm{D}} = 0$, and $\lambda = 1$:

$$-\Delta u + cu = f \quad \text{in } \Omega, \tag{4.54}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{4.55}$$

In order to discretize this problem by the lowest-order FEM, we further consider the domain Ω to be polytopic and such that it can be covered by a prismatic mesh $\mathcal{T}_h$. The prismatic mesh $\mathcal{T}_h$ satisfies requirements $(\mathcal{T}1)$–$(\mathcal{T}7)$ from Section 3.1 and it consists of right triangular prisms $P = T \times I$ with T being a triangle and I an interval. The corresponding finite element space consists of functions piecewise linear in both (x_1, x_2)-plane and in the x_3-direction:

$$V_h = \Big\{ \varphi \in H_0^1(\Omega) : \ \varphi(x_1,x_2,x_3)|_P = \sum_{i=1}^{3}\sum_{j=1}^{2} z_{i,j}\lambda_i(x_1,x_2)\ell_j(x_3), \text{ where } z_{i,j} \in \mathbb{R},\ \lambda_i(x_1,x_2) \in \mathbb{P}^1(T),\ \ell_j(x_3) \in \mathbb{P}^1(I),\ P \in \mathcal{T}_h,\ P = T \times I \Big\} \tag{4.56}$$

with $\mathbb{P}^1(T)$ and $\mathbb{P}^1(I)$ denoting the spaces of linear functions on the triangle T and on the interval I, respectively. The finite element formulation, see (4.25), reads: find $u_h \in V_h$ such that

$$a(u_h, v_h) = \mathcal{F}(v_h) \quad \forall v_h \in V_h, \tag{4.57}$$

where the bilinear form a and the linear functional $\mathcal{F}$ are given by (4.26).

In order to formulate the main result of [A1], we have to introduce for each prism $P \in \mathcal{T}_h$, $P = T \times I$, quantities

$$d_L^{(P)} = \left(\frac{2 \cot \alpha_{\max}^{(T)}}{|T|} - \frac{\|c\|_{\infty,P}}{3} \right)^{-1/2}, \tag{4.58}$$

$$d_U^{(P)} = \left(\frac{\|c\|_{\infty,P}}{6} + \frac{\cot \alpha_{\text{mid}}^{(T)} + \cot \alpha_{\min}^{(T)}}{2\,|T|} \right)^{-1/2},$$

where $\alpha_{\max}^{(T)}$, $\alpha_{\text{med}}^{(T)}$, and $\alpha_{\min}^{(T)}$ stand for the maximal, medium, and minimal angle in the triangular base T of the prism P, respectively. The value $d_U^{(P)}$ is well defined for any prism P, because $\alpha_{\text{mid}}^{(T)} < \pi/2$ and $\alpha_{\min}^{(T)} \leq \pi/3$ hold true for any triangle T. On the other hand, the value $d_L^{(P)}$ is well defined only if $\cot \alpha_{\max}^{(T)} > \|c\|_{\infty,P}\,|T|/6$.

Theorem 4.21. *Let us consider problem (4.54)–(4.55). Further, let us assume a prismatic partition $\mathcal{T}_h$ of Ω and the corresponding discretization by the lowest-order prismatic finite elements (4.56)–(4.57). Let $d_L^{(P)}$, $d_U^{(P)}$ be given by (4.58) and let $d^{(P)}$ stand for the altitude of the prism $P \in \mathcal{T}_h$. If*

$$d_L^{(P)} \leq d^{(P)} \leq d_U^{(P)} \quad \textit{for all } P \in \mathcal{T}_h, \tag{4.59}$$

then the discretization (4.57) satisfies the discrete maximum principle.

Proof. See [A1]. □

Paper [A1] further discusses the limitations on angles of the triangular bases of prisms such that condition (4.59) can be satisfied. Briefly, if $c = 0$ and prismatic partition $\mathcal{T}_h$ satisfies (4.59) then

$$\alpha_{\max}^{\mathcal{T}_h} \leq \arctan \sqrt{8} = \arccos 1/3 \approx 70.5288^\circ,$$
$$\alpha_{\min}^{\mathcal{T}_h} \geq \arctan(\sqrt{5}/2) = \arccos 2/3 \approx 48.1897^\circ,$$

and

$$\frac{|T_{\max}|}{|T_{\min}|} \leq 2,$$

where $\alpha_{\max}^{\mathcal{T}_h}$ and $\alpha_{\min}^{\mathcal{T}_h}$ denote the maximal and the minimal angle in the triangular bases over the whole partition $\mathcal{T}_h$ and similarly, $T_{\max}$ and $T_{\min}$ stand for the triangular bases with maximal and minimal area over the whole partition $\mathcal{T}_h$, respectively.

We observe that these limitations on both the angles and the areas of the triangular bases are quite severe. Nevertheless, suitable prismatic partitions providing the DMP exist and they can be used if the validity of the DMP is desired.

Finally, let us note that the methodology presented in Section 4.2 can be easily applied to generalize the statement of Theorem 4.21 to problems with nonhomogeneous Dirichlet boundary conditions.

Furthermore, the technique used above for simplicial and block finite elements, see Sections 4.5 and 4.6, can be well used also for prismatic elements to generalize condition (4.59) to problems with variable diffusion coefficient λ. Also, this technique enables us to replace the $L^\infty(K)$-norm in (4.58) by the corresponding moments of c similar to (4.29) or (4.36). This would provide even more general condition.

4.8 Generalizations of the standard approach

In the above Sections 4.5–4.7 we applied the standard approach described in Section 4.2 to obtain sufficient conditions for the validity of the DMP. This approach is based on the investigation of the local finite element matrices, see Theorem 4.6. Thinking about generalizations of this approach it is natural to investigate the global (assembled) finite element matrices and to employ Theorem 4.5.

However, the investigation of the global finite element matrix is more demanding and therefore people usually restrict themselves to simple problems. From this reason we consider in this section the Poisson equation with homogeneous Dirichlet boundary conditions.

The corresponding result in 2D for triangular meshes is quite well known. The global stiffness matrix has nonpositive off-diagonal entries essentially if and only if the underlined triangulation is of the Delaunay type, see [77, 84] and also [11, 60]. The point is to have the sum of the two angles opposite each edge of the triangulation at most π. This is the sufficient and necessary condition for the global stiffness matrix to have nonpositive off-diagonal entries. Moreover, this condition means that the triangulation is of the Delaunay type. Paper [69] shows that the DMP may hold in some cases even if there is an edge with both opposite angles obtuse. On the other hand, Jan Brandts showed that only one badly shaped triangle in a triangulation can destroy the validity of the DMP [6].

The result [47] is based on Theorem 4.4 and on a more general sufficient condition for monotonicity of a matrix [4]. They obtain a sufficient conditions on the dihedral angles of tetrahedral partitions yielding the DMP. Their condition allows for angles slightly greater then $\pi/2$. They present a numerical experiment, where the greatest dihedral angle in the tetrahedral mesh is slightly greater than 100° and the DMP still holds true.

The author of this thesis is not aware of a publication presenting similar generalizations for block finite elements. Application of Theorem 4.5 to rectangular finite elements yields a possibility of having the DMP even if certain narrow rect-

angles appear in the mesh. This possibility was already discussed in Section 4.6, Example 4.2. Similar generalization to 3D block finite elements brings no improvement of the "cube conditions" from Corollary 4.19. This fact is quite easy to see, because a contribution to the finite element matrix coming from two trilinear basis functions corresponding to two nodes connected by an edge can never be negative, see Lemma 4.18. In the best, it can be zero if all corresponding elements are cubes.

There are also other approaches how to guarantee the discrete maximum principle. In [13] the finite element method is modified in such a way that the resulting approximation satisfies the DMP on arbitrary meshes. However, a disadvantage of this approach is the nonlinearity of the resulting numerical scheme. The scheme is nonlinear even if the underlined partial differential equation is linear.

Let us note that it is possible to find in the literature even less general approaches than we describe in this thesis. Classical approach of Ph. Ciarlet [15, 18] essentially requires the corresponding finite element matrix to be irreducibly diagonally dominant in order to prove the DMP. This assumption is superfluous and we present this fact in detail in [38].

CHAPTER

FIVE

Discrete maximum principles for higher-order finite elements

The analysis of the DMP for higher-order finite elements substantially differs from the lowest-order case. The crucial point is that the higher-order basis functions do not satisfy conditions (4.1)–(4.2). Consequently, Theorems 4.4–4.6, the analysis of the lowest-order finite elements was based on, cannot be used.

The problem is fundamental. In principle, if we express a higher-order polynomial as a linear combination of certain basis functions, it is very complicated to find conditions on the coefficients which would be equivalent with the nonnegativity of the polynomial. Such equivalent conditions can be found for quadratic even cubic polynomials, but the higher the degree the more complicated the conditions are.

The nonnegative polynomials are connected to the 17th of the 23 famous Hilbert problems. Originally, people asked if any nonnegative multivariate polynomial can be represented as a sum of squares of polynomials. It turned out that this is not true. For example, the polynomial $z^6 + x^4y^2 + x^2y^43x^2y^2z^2$ is nonnegative in $\mathbb{R}^3$, but cannot be expressed as a sum of squares of polynomials. Therefore, David Hilbert included among his problems also the following question: *Given a multivariate polynomial that takes only non-negative values over the reals, can it be represented as a sum of squares of rational functions?* This problem was solved in 1927 by Emil Artin. The answer is affirmative. For more information we recommend the book [64].

Anyway, the analysis of the DMP for higher-order finite elements using the expansion coefficients seems to be untreatable. Therefore, we choose another approach and analyze directly the discrete Green's function (DGF). In particular, to prove the DMP we use Theorem 3.3 and to handle the DGF and the error of the elliptic projection of the Dirichlet lift we employ Theorems 3.4 and 3.5.

Nevertheless, both the DGF and the error of the elliptic projection of the Dirichlet lift are complicated objects and, therefore, we will concentrate first on very simple problems trying to characterize the validity of the DMP and then extending the results to more complex problems. The introductory Section 5.1 presents the discretization of the 1D Poisson equation by higher-order finite elements. Section 5.2 provides the discrete maximum principle results to 1D diffusion problems with piecewise constant coefficient and with general boundary conditions. This is a summary of results published in [A3], [A4], and [A5], see Appendices D–F. The more complicated case of 1D diffusion-reaction problem is analyzed in Section 5.3. It is a presentation of paper [A6], see Appendix G. In Section 5.4 we comment the two-dimensional case. Finally, Section 5.5 shows another approach how to handle the DMP in the higher-order case, see [A8] attached in Appendix I.

5.1 Higher-order finite elements in 1D

Let us consider interval $\Omega = (a^\partial, b^\partial)$ and a general 1D elliptic problem (4.3)–(4.5) with general boundary conditions. Its weak formulation reads: find $u \in H^1(\Omega)$ such that $u - g_\mathrm{D} \in V$ and

$$a(u, v) = \mathcal{F}(v) \quad \forall v \in V, \tag{5.1}$$

where $V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_\mathrm{D}\}$, the bilinear form $a(\cdot,\cdot)$ and the linear functional $\mathcal{F}(\cdot)$ are defined by (4.6) and (4.7), respectively, and g_D is the Dirichlet lift of the Dirichlet data, see Section 4.3 for the 1D case and Section 3.1 for the general case.

We will solve this problem by 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 Ω 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, $k = 1, 2, \ldots, M$. We set the higher-order finite element space

$$X_h = \{v_h \in H^1(\Omega) : v_h|_{K_k} \in \mathbb{P}^{p_k}(K_k),\ k = 1, 2, \ldots, M\}, \tag{5.2}$$

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 a subspace $V_h = X_h \cap V$. The higher-order finite element solution $u_h \in X_h$ is then determined by the requirements $u_h - g_{\mathrm{D},h} \in V_h$ and

$$a(u_h, v_h) = \mathcal{F}(v_h) \quad \forall v_h \in V_h, \tag{5.3}$$

where the approximate Dirichlet lift $g_{\mathrm{D},h}$ is defined in the same way as in Section 4.3.

We construct the basis of V_h in the standard finite element way transforming the *Lobatto shape functions* from the reference element $\widehat{K} = [-1, 1]$ to the physical elements K_k. The Lobatto shape functions, see e.g. [74], are more-less standard in the higher-order finite elements, but for the reader's convenience, we recall their definition and properties.

First, there are two linear shape functions $\ell_0(\xi) = (1 - \xi)/2$ and $\ell_1(\xi) = (1 + \xi)/2$, $\xi \in \widehat{K}$. The higher-order shape functions $\ell_2, \ell_3, \dots$ are defined as antiderivatives of the Legendre polynomials vanishing at both end-points of $\widehat{K}$, i.e.

$$\ell_i(\xi) = \sqrt{\frac{2i-1}{2}} \int_{-1}^{\xi} P_{i-1}(s)\,\mathrm{d}s, \quad i = 2, 3, \dots, \tag{5.4}$$

where $P_i(\xi)$ stands for the Legendre polynomial of degree i. Clearly, ℓ_i is a polynomial of degree i for $i = 2, 3, \dots$, it vanishes at both points ± 1, and these functions are orthonormal in the following sense

$$\int_{-1}^{1} \ell_i'(\xi)\ell_j'(\xi)\,\mathrm{d}\xi = \delta_{ij}, \quad i, j = 2, 3, \dots, \tag{5.5}$$

where δ_{ij} stands for the Kronecker's tensor. See Figure 5.1 (left) for an illustration. It is possible to factor out the root-factor $\ell_0(\xi)\ell_1(\xi) = (1 - \xi)(1 + \xi)/4$ for each ℓ_i, $i = 2, 3, \dots$, and express

$$\ell_{i+2}(\xi) = \ell_0(\xi)\ell_1(\xi)\kappa_i(\xi), \quad i = 0, 1, 2, \dots \tag{5.6}$$

with $\kappa_i(\xi)$ being a polynomial of degree i, see Figure 5.1 (right). In the sequel, we examine the properties of polynomials κ_i, $i = 0, 1, 2, \dots$.

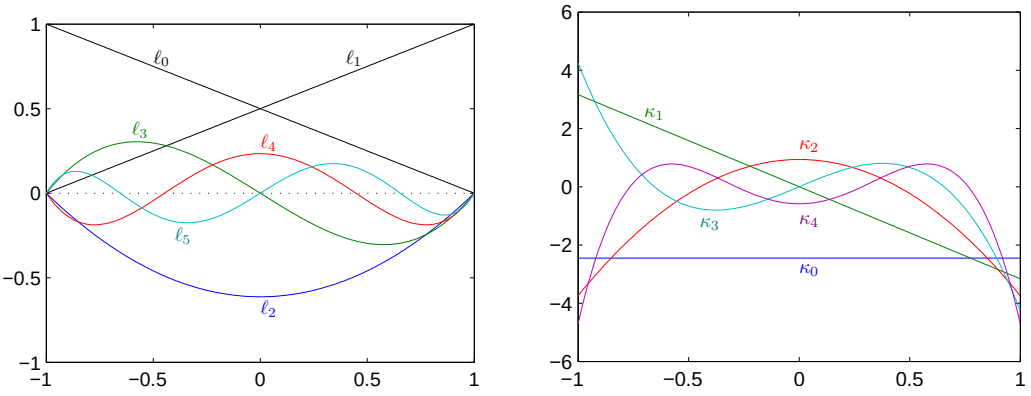


Figure 5.1: Lobatto shape functions ℓ_0, ℓ_1, …, ℓ_5 (left) and the corresponding polynomials κ_0, κ_1, …, κ_4 (right).

The Legendre polynomials are well known [42] to satisfy the following differential equation of second order:

$$\big[(1-\xi^2)P_i'(\xi)\big]' = -i(i+1)P_i(\xi), \quad i = 0, 1, 2, \ldots.$$

Integrating this identity and taking into account definition (5.4), we easily find that

$$\ell_{i+2}(\xi) = -\frac{1-\xi^2}{4}\frac{\sqrt{8(2i+3)}}{(i+2)(i+1)}P_{i+1}'(\xi), \quad i = 0, 1, 2, \ldots.$$

Comparing this with (5.6) we conclude that

$$\kappa_i(\xi) = -\frac{\sqrt{8(2i+3)}}{(i+2)(i+1)}P_{i+1}'(\xi), \quad i = 0, 1, 2, \ldots.$$

Hence, the polynomials κ_i are just scaled derivatives of the Legendre polynomials.

Furthermore, the derivatives of the Legendre polynomials are known [42] to be proportional to Jacobi polynomials $P_i^{(1,1)}(\xi)$:

$$P_{i+1}'(\xi) = \frac{i+2}{2}P_i^{(1,1)}(\xi).$$

Thus, polynomials κ_i are proportional to Jacobi polynomials $P_i^{(1,1)}$ which are orthogonal on $[-1, 1]$ with respect to the weight $(1-\xi)(1+\xi)$. This orthogonality can be expressed as

$$\int_{-1}^{1} \ell_0(\xi)\ell_1(\xi)\kappa_i(\xi)\kappa_j(\xi)\,\mathrm{d}\xi = \frac{4}{(i+1)(i+2)}\delta_{ij}, \quad i, j = 0, 1, 2, \ldots.$$

Another consequence is that the polynomials κ_i can be generated by the three-term recurrence formula

$$\frac{i+4}{\sqrt{2i+7}}\kappa_{i+2}(\xi) = \sqrt{2i+5}\,\xi\,\kappa_{i+1}(\xi) - \frac{i+1}{\sqrt{2i+3}}\kappa_i(\xi), \quad i = 0, 1, 2, \ldots.$$

For illustration, this gives us the following identities (see also Figure 5.1)

$$\begin{aligned}
\kappa_0(\xi) &= -\sqrt{6},\\
\kappa_1(\xi) &= -\sqrt{10}\xi,\\
\kappa_2(\xi) &= -\frac{1}{4}\sqrt{14}(5\xi^2-1),\\
\kappa_3(\xi) &= -\frac{3}{4}\sqrt{2}(7\xi^2-3)\xi,\\
\kappa_4(\xi) &= -\frac{1}{8}\sqrt{22}(21\xi^4-14\xi^2+1),\\
\kappa_5(\xi) &= -\frac{1}{8}\sqrt{26}(33\xi^4-30\xi^2+5)\xi,\\
\kappa_6(\xi) &= -\frac{1}{64}\sqrt{30}(429\xi^6-495\xi^4+135\xi^2-5),\\
\kappa_7(\xi) &= -\frac{1}{64}\sqrt{34}(715\xi^6-1001\xi^4+385\xi^2-35)\xi,\\
\kappa_8(\xi) &= -\frac{1}{128}\sqrt{38}(2431\xi^8-4004\xi^6+2002\xi^4-308\xi^2+7).
\end{aligned}$$

Interestingly, the Lobatto shape functions possess certain orthogonality in the $L^2(-1,1)$ sense, too:

$$\int_{-1}^{1}\ell_i(\xi)\ell_j(\xi)\,\mathrm{d}\xi=\begin{cases}\dfrac{2}{(2i+1)(2i-3)} & \text{for } i=j,\\[2ex] \dfrac{-1}{(2i+1)\sqrt{(2i-1)(2i+3)}} & \text{for } i=j+2 \text{ and } j=i+2,\\[2ex] 0 & \text{otherwise.}\end{cases}$$

with $i,j=2,3,\ldots$. In addition, concerning ℓ_0 and ℓ_1, we have

$$\begin{aligned}
&\int_{-1}^{1}\ell_0(\xi)\ell_2(\xi)\,\mathrm{d}\xi=\int_{-1}^{1}\ell_1(\xi)\ell_2(\xi)\,\mathrm{d}\xi=-\frac{\sqrt{6}}{6},\\
&\int_{-1}^{1}\ell_0(\xi)\ell_3(\xi)\,\mathrm{d}\xi=-\int_{-1}^{1}\ell_1(\xi)\ell_3(\xi)\,\mathrm{d}\xi=\frac{\sqrt{10}}{30},\\
&\int_{-1}^{1}\ell_0(\xi)\ell_i(\xi)\,\mathrm{d}\xi=\int_{-1}^{1}\ell_1(\xi)\ell_i(\xi)\,\mathrm{d}\xi=0\quad\text{for } i=4,5,\ldots.
\end{aligned}$$

Now, let us return back to the finite element discretization. The basis functions of V_h are defined with the aid of the reference mapping

$$\chi_k(\xi)=\frac{h_k\xi+(x_k+x_{k-1})}{2},\quad k=1,2,\ldots,M. \tag{5.7}$$

We distinguish two types of basis functions – the vertex and the bubble functions. The vertex functions are composed of the transformed (linear) shape functions ℓ_0 and ℓ_1, while the bubble functions are just transformations of the higher-order shape functions ℓ_2, ℓ_3, etc. For example, vertex function $\varphi^{v,x_i}(x)$ corresponding to the node x_i of the partition is defined as

$$\varphi^{v,x_i}(x) = \begin{cases} \ell_1\left(\chi_i^{-1}(x)\right) & \text{for } x \in K_i, \\ \ell_0\left(\chi_{i+1}^{-1}(x)\right) & \text{for } x \in K_{i+1}, \end{cases} \qquad i = 1, 2, \ldots, M-1.$$

Similarly, $p_k - 1$ bubble functions supported in element K_k, $k = 1, 2, \ldots, M$, are defined as

$$\varphi_j^{b,K_k}(x) = \ell_{j+1}\left(\chi_k^{-1}(x)\right), \quad j = 1, 2, \ldots, p_k - 1. \tag{5.8}$$

The finite element space V_h naturally splits into two subspaces $V_h = V_h^v \oplus V_h^b$, where V_h^v is the span of the vertex basis functions and V_h^b is the span of the bubble functions. The dimension of V_h^b is always $N^b = \dim V_h^b = \sum_{k=1}^{M}(p_k - 1)$, while the dimension of V_h^v depends on the prescribed boundary conditions. The Dirichlet boundary conditions can be prescribed (i) at both end-points, (ii) at one of the two end-points, or (iii) nowhere. The corresponding dimension N^v of V_h^v is then (i) $M-1$, (ii) M, or (iii) $M+1$.

We denote the basis of V_h^v as $\varphi_1^v, \varphi_2^v, \ldots, \varphi_{N^v}^v$ and the basis of V_h^b as φ_1^b, $\varphi_2^b, \ldots, \varphi_{N^b}^b$. The basis functions φ_i, $i = 1, 2, \ldots, N^0$, $N^0 = N^v + N^b$, (denoted without any superscript) correspond to the basis of the entire finite element space V_h. We consider the first N^v functions to be vertex functions and the last N^b functions to be bubbles, i.e.

$$\varphi_1 = \varphi_1^v, \ \ldots, \ \varphi_{N^v} = \varphi_{N^v}^v, \ \varphi_{N^v+1} = \varphi_1^b, \ \ldots, \ \varphi_{N^v+N^b} = \varphi_{N^b}^b.$$

5.2 Discrete maximum principle for 1D diffusion problem

In this section we present a technique which can be successfully used to certain 1D elliptic problems. These results were already published by the author of this thesis and his co-authors. In particular, the case of Poisson problem with homogeneous Dirichlet boundary conditions appeared in [A3], the results for Poisson problem with mixed Dirichlet-Neumann boundary conditions are in [A4], and paper [A5] deals with elliptic problems with piecewise constant diffusion coefficients. These three papers are attached to this thesis as Appendices D–F.

Below we introduce these results in a unified way presenting the most general case. The framework built up in the previous sections enables us to handle even

the nonhomogeneous Dirichlet boundary conditions in the systematic manner, which is a slight novelty with respect to publications [A3], [A4], and [A5].

Let us consider 1D diffusion problem with piecewise constant diffusion coefficient and with general boundary conditions of Dirichlet and/or Neumann type:

$$-(\mathcal{A}u')' = f \quad \text{in } \Omega = (a^\partial, b^\partial), \tag{5.9}$$

$$u = g_{\mathrm{D}} \quad \text{on } \Gamma_{\mathrm{D}}, \tag{5.10}$$

$$\mathcal{A}u' n_{1\mathrm{D}} = g_{\mathrm{N}} \quad \text{on } \Gamma_{\mathrm{N}}. \tag{5.11}$$

Clearly, this is a special case of problem (4.3)–(4.5) introduced in Section 4.3. Therefore, we adopt the same setting and notation as in Section 4.3. Namely, concerning the types of boundary conditions: $\Gamma_{\mathrm{D}} \subset \partial\Omega$, $\Gamma_{\mathrm{N}} \subset \partial\Omega$, $\Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}} = \partial\Omega$, $\Gamma_{\mathrm{D}} \cap \Gamma_{\mathrm{N}} = \emptyset$, $\partial\Omega = \{a^\partial, b^\partial\}$. We point out that in the case of pure diffusion problem (5.9)–(5.11) we exclude the possibility of solely Neumann boundary conditions, because in this case (i.e. if $\Gamma_{\mathrm{D}} = \emptyset$ and $\Gamma_{\mathrm{N}} = \{a^\partial, b^\partial\}$) there is no uniqueness.

The existence and uniqueness is discussed in Section 4.3, too. Namely, we adopt the assumptions (4.8), which reduces in this case to

$$0 < \lambda_{\min} \leq \mathcal{A} \quad \text{in } \Omega. \tag{5.12}$$

Furthermore, the diffusion coefficient $\mathcal{A}$ is assumed to be piecewise constant and we denote by $\mathcal{A}_k$ the constant value of $\mathcal{A}$ in the element $K_k \in \mathcal{T}_h$, $k = 1, 2, \ldots, M$ – see Section 4.3 or 5.1 for the definition of the 1D partition $\mathcal{T}_h$ of the interval Ω.

We discretize problem (5.9)–(5.11) by higher order finite elements as described above in Section 5.1. The weak formulation is presented in (5.1), finite element formulation is in (5.3), and the corresponding bilinear form $a(\cdot,\cdot)$ and linear functional $\mathcal{F}(\cdot)$ are defined in (4.6)–(4.7) with $b = c = \alpha = 0$.

In particular, we utilize the splitting of the basis functions into the vertex and bubble functions. First of all, we show in the following lemma that in the pure diffusion case with piecewise constant diffusion coefficient the vertex and bubble functions are a-orthogonal.

Lemma 5.1. *Let X_h be given by (5.2). Let $a(\cdot,\cdot)$ be given by (4.6) with $b = c = \alpha = 0$. Let $\varphi^v \in X_h$ be any vertex function and let $\varphi^b \in X_h$ be any bubble function. Then*

$$a(\varphi^v, \varphi^b) = 0. \tag{5.13}$$

Proof. Any vertex function φ^v is supported in at most two elements. Any bubble function φ^b is supported in a single element only. If $\mathrm{meas}(\mathrm{supp}\,\varphi^v \cap \mathrm{supp}\,\varphi^b) = 0$ then clearly orthogonality (5.13) holds. Thus, the only remaining option is that

$\operatorname{supp}\varphi^v \cap \operatorname{supp}\varphi^b = K_k$ for some element $K_k = [x_{k-1}, x_k]$, $k \in \{1, 2, \ldots, M\}$. Orthogonality (5.13) now follows immediately from the integration by parts:

$$a(\varphi^v, \varphi^b) = \int_{K_k} \mathcal{A}(\varphi^v)'(\varphi^b)' \,\mathrm{d}x = \mathcal{A}_k \left[(\varphi^v)'\varphi^b\right]_{x_{k-1}}^{x_k} - \mathcal{A}_k \int_{x_{k-1}}^{x_k} (\varphi^v)''\varphi^b \,\mathrm{d}x = 0,$$

where $\mathcal{A}_k$ is constant, and the last equality holds true, because φ^b vanishes at both end-points x_{k-1} and x_k and because φ^v is linear in K_k. □

The splitting of the basis into the vertex and bubble part leads to a natural two-by-two block structure of the resulting stiffness matrix. Orthogonality (5.13) implies that the two off-diagonal blocks in this structure vanish. Thus, the stiffness matrix and its inverse can be expressed as

$$A = \begin{pmatrix} A^v & 0 \\ 0 & A^b \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} (A^v)^{-1} & 0 \\ 0 & (A^b)^{-1} \end{pmatrix}, \tag{5.14}$$

where $A^v \in \mathbb{R}^{N^v \times N^v}$ consists of the a-products of pairs of vertex functions and $A^b \in \mathbb{R}^{N^b \times N^b}$ consits of the a-products of pairs of bubbles. In addition, the matrix A^b is actually diagonal in this setting.

Lemma 5.2. *Let X_h be given by (5.2). Let $a(\cdot,\cdot)$ be given by (4.6) with $b = c = \alpha = 0$. Let $\varphi_i^b \in X_h$ and $\varphi_j^b \in X_h$ be arbitrary but distinct bubble functions. Then*

$$a(\varphi_i^b, \varphi_j^b) = 0 \quad \textit{and} \quad a(\varphi_i^b, \varphi_i^b) = 2\mathcal{A}_k/h_k,$$

where $\mathcal{A}_k$ is the constant value of $\mathcal{A}$ in K_k, h_k is the diameter of the element K_k, and K_k is the element the bubble function φ_i^b is supported in, $k \in \{1, 2, \ldots, M\}$.

Proof. If $\operatorname{meas}(\operatorname{supp}\varphi_i^b \cap \operatorname{supp}\varphi_j^b) = 0$ then clearly $a(\varphi_i^b, \varphi_j^b) = 0$. Thus, let us suppose that $\operatorname{supp}\varphi_i^b = \operatorname{supp}\varphi_j^b = K_k$ for some $k \in \{1, 2, \ldots, M\}$. Further, let $\varphi_i^b(x) = \ell_p(\chi_k^{-1}(x))$ and $\varphi_j^b(x) = \ell_q(\chi_k^{-1}(x))$ for $x \in K_k$ with $p \neq q$ being polynomial degrees of φ_i^b and φ_j^b, see (5.8). Then, we can compute

$$a(\varphi_i^b, \varphi_j^b) = \mathcal{A}_k \int_{K_k} (\varphi_i^b)'(\varphi_j^b)' \,\mathrm{d}x = \mathcal{A}_k \int_{\widehat{K}} \ell_p'\ell_q' \,\mathrm{d}\xi \, 2/h_K = 0,$$

where we use the substitution $x = \chi_k(\xi)$ mapping the reference element $\widehat{K} = [-1, 1]$ to the physical element K_k and the orthogonality (5.5) of the Lobatto shape functions.

Finally, using the same manipulations we obtain

$$a(\varphi_i^b, \varphi_i^b) = \mathcal{A}_k \int_{\widehat{K}} (\ell_p')^2 \,\mathrm{d}\xi \, 2/h_k = 2\mathcal{A}_k/h_k.$$

□

Thus, the bubble part A^b of the stiffness matrix A can be expressed as

$$A^b = \operatorname{diag}\left(\underbrace{\frac{2\mathcal{A}_1}{h_1}, \dots, \frac{2\mathcal{A}_1}{h_1}}_{(p_1-1) \text{ times}}, \underbrace{\frac{2\mathcal{A}_2}{h_2}, \dots, \frac{2\mathcal{A}_2}{h_2}}_{(p_2-1) \text{ times}}, \ \dots \ , \underbrace{\frac{2\mathcal{A}_M}{h_M}, \dots, \frac{2\mathcal{A}_M}{h_M}}_{(p_M-1) \text{ times}} \right).$$

Consequently, the inverse $(A^b)^{-1}$ is trivial and nonnegative. The other block A^v is tridiagonal and the nonnegativity of its inverse was investigated in Section 4.3. In (4.16) we have found a sufficient condition for the nonnegativity of $(A^v)^{-1}$. However, here we consider $c = b = \alpha = 0$ and therefore condition (4.16) reduces to $0 \le 6\mathcal{A}_k$, which is trivially satisfied for all $k = 1, 2, \dots, M$, see (5.12). Thus, we conclude that $(A^v)^{-1}$ is automatically nonnegative for the pure diffusion case with piecewise constant diffusion coefficient.

Thanks to the block structure (5.14), the discrete Green's function (3.15) can be split into the vertex and bubble part

$$G_h(x,y) = \sum_{i=1}^{N^0} \sum_{j=1}^{N^0} \varphi_i(y)(A^{-1})_{ij}\varphi_j(x) = G_h^v(x,y) + G_h^b(x,y),$$

where

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

$$G_h^b(x,y) = \sum_{k=1}^{M} \frac{h_k}{2\mathcal{A}_k} \sum_{j=2}^{p_k} \ell_j\left(\chi_k^{-1}(y)\right) \ell_j\left(\chi_k^{-1}(x)\right).$$

As we discussed above, the vertex part G_h^v is automatically nonnegative. Since $h_k/(2\mathcal{A}_k) > 0$, the nonnegativity of the bubble part depends solely on the nonnegativity of the polynomials

$$\mathcal{K}^p(\xi,\eta) = \sum_{j=0}^{p-2} \kappa_j(\eta)\kappa_j(\xi)$$

for $(\xi,\eta) \in \widehat{K}^2$ and $p = 2, 3, \dots$. Indeed, this is due to the fact that $G_h^b(x,y)$ is nonnegative if and only if the polynomial

$$\sum_{j=2}^{p_k} \ell_j\left(\chi_k^{-1}(y)\right) \ell_j\left(\chi_k^{-1}(x)\right) = \ell_0(\eta)\ell_1(\eta)\ell_0(\xi)\ell_1(\xi) \sum_{j=2}^{p_k} \kappa_{j-2}(\eta)\kappa_{j-2}(\xi)$$

is nonnegative for all $\xi = \chi_k^{-1}(x) \in [-1,1]$ and all $\eta = \chi_k^{-1}(y) \in [-1,1]$, see (5.6).

Unfortunately, polynomials $\mathcal{K}^p$ are nonnegative in $\widehat{K}^2$ for exceptional values of p. In particular, for $p = 2$, 4, and 6 only. For other values of p it is necessary to investigate under what conditions the nonnegative vertex part G_h^v overweights the possibly negative part G_h^b such that their sum $G_h = G_h^v + G_h^b$ is nonnegative. The analysis of this situation was done in [A3] for the Poisson problem with homogeneous Dirichlet boundary conditions, in [A4] for Poisson problem with mixed Dirichlet-Neumann boundary conditions, and in [A5] for the diffusion problem with piecewise constant diffusion coefficient and homogeneous Dirichlet boundary conditions. These papers are attached to this thesis as Appendices D–F.

The analysis of the nonnegativity of G_h is based on the explicit formula for the inverse of the vertex part A^v of the stiffness matrix. The particular shape of this formula depends on the boundary conditions. The formula for the pure Dirichlet boundary conditions is presented in [A3] and in [A5]. Slightly different formula for the mixed boundary conditions can be found in [A4].

In all three papers from Appendices D–F, we suppose primarily the homogeneous Dirichlet boundary conditions for the sake of simplicity. In what follows, we utilize the methodology presented above and generalize the results obtained in [A3], [A4], and [A5] to the case of nonhomogeneous Dirichlet boundary conditions.

Theorem 5.3. *Let us consider problem (5.9)–(5.11) with piecewise constant coefficient $\mathcal{A}$ and with the Dirichlet boundary condition prescribed at at least one of the end-points of Ω. Further, let us consider its higher-order finite element discretization (5.3). Then*

$$g_{\mathrm{D},h}(y) - (\Pi_h^0 g_{\mathrm{D},h})(y) \geq 0 \quad \textit{for all } g_{\mathrm{D},h} \in V_h^\partial,\ g_{\mathrm{D},h} \geq 0 \textit{ in } \Omega,\ y \in \Omega. \tag{5.15}$$

Proof. First, let us consider the Dirichlet boundary conditions prescribed at both end-points of Ω. Then $V_h^\partial = \mathrm{span}\{\varphi_1^\partial, \varphi_2^\partial\}$ with both φ_1^∂ and φ_2^∂ being the piecewise linear vertex functions corresponding to the end-points a^∂ and b^∂ of Ω, respectively.

Clearly, $g_{\mathrm{D},h} \in V_h^\partial$, $g_{\mathrm{D},h} = \sum_{m=1}^2 c_m^\partial \varphi_m^\partial$, is nonnegative if and only if both c_1^∂ and c_2^∂ are nonnegative. Therefore, condition (5.15) is equivalent to the condition

$$\sum_{m=1}^{2} c_m^\partial \left[\varphi_m^\partial(y) - \Pi_h^0 \varphi_m^\partial(y)\right] \geq 0 \quad \text{for all } c_m^\partial \geq 0,\ m = 1, 2, \text{ and all } y \in \Omega.$$

This condition is clearly satisfied if and only if

$$\varphi_m^\partial(y) - \Pi_h^0 \varphi_m^\partial(y) \geq 0 \quad \text{for all } y \in \Omega \text{ and for both } m = 1, 2.$$

Using Theorem 3.5, we express

$$\varphi_m^\partial(y) - \Pi_h^0 \varphi_m^\partial(y) = \varphi_m^\partial(y) - \sum_{i=1}^{N^0} \sum_{j=1}^{N^0} \varphi_i(y) (A^{-1})_{ij} A_{jm}^\partial.$$

Thanks to orthogonality (5.13), the matrix A^∂ has practically the same structure (4.17) as in the lowest-order case. Namely, the only nonzero entries of A^∂ are A^∂_{11} and $A^\partial_{N^v,2}$. In addition, these two nonzero entries are nonpositive, see (4.14)–(4.15).

Hence, for $m = 1$ we obtain

$$\varphi_1^\partial(y) - \Pi_h^0\varphi_1^\partial(y) = \varphi_1^\partial(y) - \sum_{i=1}^{N^v+N^b} \varphi_i(y)(A^{-1})_{i1}A^\partial_{11}$$
$$= \varphi_1^\partial(y) + \sum_{i=1}^{N^v} \varphi_i^v(y)(A^v)^{-1}_{i1}(-A^\partial_{11}) \geq 0,$$

where we utilize the fact that the vertex-bubble block of the stiffness matrix A vanishes – see (5.14), that the vertex functions as well as all entries of $(A^v)^{-1}$ are nonnegative, and that A^∂_{11} is negative.

For $m = 2$, we can proceed in practically the same way to obtain nonnegativity of $\varphi_2^\partial - \Pi_h^0\varphi_2^\partial$. This proves the theorem in the case of Dirichlet boundary conditions prescribed at both end-points of Ω. If the Dirichlet boundary condition is prescribed in one end-point only then the proof is analogous. □

Now, the statement of Theorem 5.3 can be combined with the nonnegativity of the discrete Green's function proved in Appendices D–F. Theorem 3.3 then provides the validity of the discrete maximum principle. The final conditions, however, differ for the pure Dirichlet boundary conditions (see Corollary 5.4 below) and for the mixed boundary conditions (see Corollary 5.5 below). Anyway, in order to present the results of papers from Appendices D–F, we have to introduce the critical relative element length

$$H^*_{\mathrm{rel}}(p) = 1 + \frac{1}{2} \min_{(\xi,\eta)\in\widehat{K}^2} \ell_0(\xi)\ell_0(\eta) \sum_{i=0}^{p-2} \kappa_i(\xi)\kappa_i(\eta). \tag{5.16}$$

This quantity depends on the polynomial degree p and it is given as a minimum of a bivariate polynomial over a compact set (a square). The computation of values $H^*_{\mathrm{rel}}(p)$ is a nontrivial task, in general. However, in this case, it is possible to find the values $H^*_{\mathrm{rel}}(p)$ analytically for $p = 2, 3, 4$. For values $p = 5, 6, \ldots, 100$, we did it numerically with high accuracy. The results are presented in Table 5.1 and in Figure 5.2. See [A3] and [A5] in Appendices D and F for more detail.

Corollary 5.4. *Let us consider problem (5.9)–(5.11) with piecewise constant coefficient $\mathcal{A}$ and with the Dirichlet boundary conditions prescribed at both end-points of Ω, i.e. $\Gamma_{\mathrm{D}} = \{a^\partial, b^\partial\}$ and $\Gamma_{\mathrm{N}} = \emptyset$. Further, let us consider higher-order*

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

Table 5.1: Critical relative element length $H^*_{\mathrm{rel}}(p)$ for $p = 1, 2, 3, \ldots, 20$.

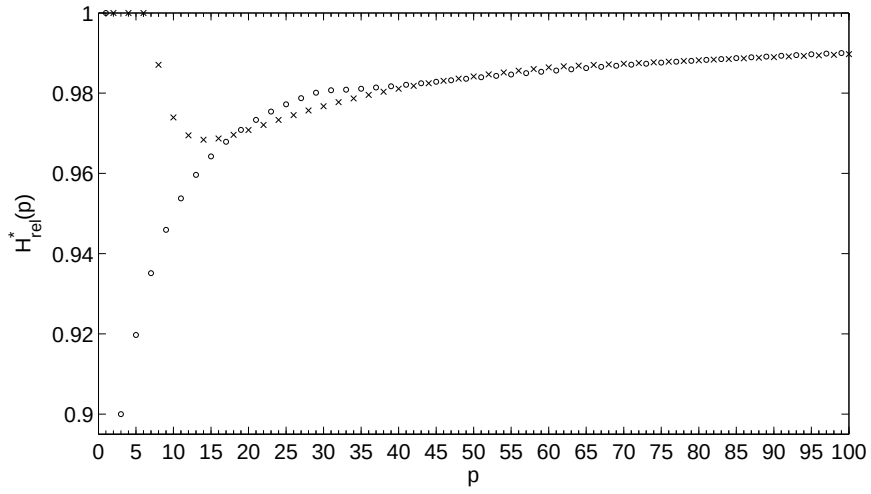


Figure 5.2: Critical relative element lengths $H^*_{\mathrm{rel}}(p)$ for $p = 1, 2, \ldots, 100$. Circles indicate the values for p odd and crosses for p even.

finite element discretization (5.3) of this problem. If

$$\frac{h_k}{\mathcal{A}_k} \leq H^*_{\mathrm{rel}}(p_k) h_{\Omega,\mathcal{A}} \quad \textit{for all } k = 1, 2, \ldots, M, \tag{5.17}$$

with $h_{\Omega,\mathcal{A}} = \sum_{j=1}^{M} h_j/\mathcal{A}_j$*, then discretization (5.3) satisfies the discrete conservation of nonnegativity.*

Proof. Results presented in paper [A5] (see Appendix F) imply the nonnegativity of the discrete Green's function:

$$G_h(x, y) \geq 0 \quad \forall (x, y) \in \Omega^2$$

provided condition (5.17) holds true (see the proof of Theorem 4.2 in [A5]). Theorem 5.3 provides nonnegativity of the error of the elliptic projection. Thus, both

assumptions (a) and (b) of Theorem 3.3 are fulfilled and the discrete conservation of nonnegativity is satisfied. □

Corollary 5.5. *Let us consider problem (5.9)–(5.11) with piecewise constant coefficient $\mathcal{A}$ and with the mixed boundary conditions, i.e. $\Gamma_{\mathrm{D}} = \{a^{\partial}\}$ and $\Gamma_{\mathrm{N}} = \{b^{\partial}\}$ or vice versa. Further, let us assume higher-order finite element discretization (5.3) of this problem. If*

$$H^*_{\mathrm{rel}}(p_k) \geq 0 \quad \textit{for all } k = 1, 2, \ldots, M, \tag{5.18}$$

then discretization (5.3) satisfies the discrete conservation of nonnegativity.

Proof. It goes through the same steps as the proof of Corollary 5.4 with the only difference that the nonnegativity of the discrete Green's function is guaranteed by condition (5.18) – see the proof of Theorem 6.1 in [A4] (Appendix E). □

First, notice that in the case of Poisson problem, i.e. for $\mathcal{A}_k = 1$, condition (5.17) reduces to $h_k \leq H^*_{\mathrm{rel}}(p_k) h_\Omega$ with $h_\Omega = b^{\partial} - a^{\partial}$ being the length of Ω. Hence, condition (5.17) for a general piecewise constant coefficient $\mathcal{A}$ can be regarded as a relation between distorted lengths of elements and the distorted length of the entire domain Ω. We also point out that this condition is rather weak. It seems that the smallest value of $H^*_{\mathrm{rel}}(p)$ is attained for $p = 3$ and it is $9/10$. Thus, in the case of Poisson problem the discrete maximum principle is satisfied if there is no element larger then $9/10$ of the length of Ω. Clearly, any "reasonable" mesh satisfies this condition.

Further notice the fundamental difference of conditions (5.17) and (5.18). The later one – corresponding to the mixed boundary conditions – is much weaker. It depends on polynomial degrees p_k only and not on the lengths h_k of elements. In addition, Figure 5.2 indicates that it is automatically satisfied for any polynomial degree p_k. In fact, if our computations of values of $H^*_{\mathrm{rel}}(p)$ involving root finding of high-order polynomials are correct and accurate then we actually verified the validity of (5.18) for all polynomial degrees up to $p = 100$, see [A3] (Appendix D) for details.

5.3 Discrete maximum principle for 1D diffusion–reaction problems

In this section we present a DMP result for 1D diffusion–reaction problem discretized by higher-order finite elements. In contrast to the pure diffusion case described above the presence of the reaction term complicates the analysis substantially. There are two principal difficulties: (i) The bubble functions are not

a-orthogonal to the vertex functions nor to themselves. (ii) There is no explicit formula for the inverse of the vertex part of the stiffness matrix. Difficulty (i) can be overcome by considering the generalized eigenfunctions of the 1D discrete Laplacian as the shape functions. Solution of difficulty (ii) requires suitable estimates of the entries of the inverse the stiffness matrix.

A detailed analysis of this case was published in [A6]. This paper is attached to this thesis as Appendix G. In this paper the reaction coefficient is supposed to be constant and the boundary conditions are considered to be homogeneous Dirichlet. In what follows, we briefly present the main result of this paper and generalize it to general non-homogeneous mixed boundary conditions of Dirichlet and Neumann type and to the piecewise constant reaction coefficient. However, we point out that these generalizations are quite simple. Especially, the generalization to the piecewise constant reaction coefficient is trivial.

Let us consider the following diffusion–reaction problem with general boundary conditions:

$$-u'' + cu = f \quad \text{in } \Omega, \tag{5.19}$$

$$u = g_{\mathrm{D}} \quad \text{on } \Gamma_{\mathrm{D}}, \tag{5.20}$$

$$u' n_{1\mathrm{D}} = g_{\mathrm{N}} \quad \text{on } \Gamma_{\mathrm{N}}. \tag{5.21}$$

Clearly, this is a special case of general problem (4.3)–(4.5) from Section 4.3. The only difference is that we suppose $\mathcal{A} = I$ and $b = \alpha = 0$, now. We assume the coefficient c to be piecewise constant and we denote by c_k the constant values of c in K_k. Technically, we adopt all the notation from Section 4.3. In particular, we consider the weak formulation (5.1) and the finite element formulation (5.3), where the bilinear form $a(\cdot,\cdot)$ and the linear functional $\mathcal{F}(\cdot)$ are given by (4.6)–(4.7) with $\mathcal{A} = I$ and $b = \alpha = 0$. Clearly, in the case $\Gamma_{\mathrm{N}} = \{a^\partial, b^\partial\}$ and $\Gamma_{\mathrm{D}} = \emptyset$ we assume $c > 0$ in at least one of the elements in order to preserve the unique solvability of (5.19)–(5.21).

We suppose the same higher-order finite element approximation as above, see Section 5.1. However, as we already mentioned, we consider the generalized eigenfunctions of the discrete 1D Laplacian for the bubble functions. Thus, we consider the reference element $\widehat{K} = [-1, 1]$ and a space $\mathbb{P}^p_0(\widehat{K})$ of polynomials of degree at most p on interval $\widehat{K}$ whose values at both end-points of $\widehat{K}$ vanish. Further, we suppose the eigenfunctions $\ell^p_i \in \mathbb{P}^p_0(\widehat{K})$ and the corresponding eigenvalues λ^p_i, $i = 2, 3, \ldots, p$, such that

$$\int_{-1}^{1} (\ell^p_i)' v' \,\mathrm{d}x = \lambda^p_i \int_{-1}^{1} \ell^p_i v \,\mathrm{d}x \quad \forall v \in \mathbb{P}^p_0(\widehat{K}).$$

The $p-1$ bubble functions supported in element K_k are then defined as trans-

formations of ℓ^p_{i+1}, $i = 1, 2, \ldots, p-1$, see (5.7) and (5.8):

$$\psi_i^{b,K_k}(x) = \ell^p_{i+1}\left(\chi_k^{-1}(x)\right), \quad i = 1, 2, \ldots, p_k - 1.$$

Hence, the new bubble functions ψ_1^b, ψ_2^b, …, $\psi_{N^b}^b$ can be used as a new basis in V^b.

Furthermore, we orthogonalize the vertex basis functions with respect to the space of bubbles V^b. The orthogonalized vertex function ψ_i^v, $i = 1, 2, \ldots, N^v$, can be expressed in the form

$$\psi_i^v(x) = \varphi_i^v(x) - \sum_{j=1}^{p_{i+1}-1} C_{1,j}^{K_{i+1}} \psi_{j+1}^{b,K_{i+1}}(x) - \sum_{j=1}^{p_i-1} C_{2,j}^{K_i} \psi_{j+1}^{b,K_i}(x). \tag{5.22}$$

Let us notice that both the original vertex function φ_i^v and the orthogonalized vertex function ψ_i^v are supported in elements K_{i+1} and K_i provided they correspond to an interior vertex. If the vertex function corresponds to a boundary vertex then it is supported in one element only and one of the two sums in (5.22) is missing. In addition, the coefficients $C_{m,j}^{K_i}$ in (5.22) can be computed as $C_{m,j}^{K_i} = \zeta \overline{B}_{mj}^{p_i}(1 + \zeta\mu_j^{p_i})^{-1}$, $\zeta = c_i h_i^2$, $\overline{B}_{mj}^{p_i} = \int_{-1}^{1} \ell_{m-1}\ell_{j+1}^{p_i}\,\mathrm{d}x/2$, and $\mu_j^{p_i} = 1/(4\lambda_{j+1}^{p_i})$, for $m = 1, 2$, $j = 1, \ldots, p_i$, $i = 1, 2, \ldots, M$. See Appendix G for details.

Thus, the new basis functions ψ_1^v, ψ_2^v, …, $\psi_{N^v}^v$ and ψ_1^b, ψ_2^b, …, $\psi_{N^b}^b$ are a-orthogonal in the same way as the standard basis functions in the case of Poisson problem, see Lemmas 5.1 and 5.2. In particular,

$$a(\psi_i^v, \psi_j^b) = 0 \quad \forall i = 1, 2, \ldots, N^v, \text{ and } j = 1, 2, \ldots, N^b,$$

and

$$a(\psi_i^b, \psi_j^b) = 0 \quad \forall i \neq j,\ i, j = 1, 2, \ldots, N^b.$$

Nevertheless, it turns out that the new vertex functions ψ_i^v do not remain nonnegative, in general. The nonnegativity of ψ_i^v depends on the size of the quantity $\zeta = c_k h_k^2$ for all values of k such that ψ_i^v is supported in K_k. Simply, ψ_i^v is nonnegative for small values of ζ only. In fact, there is a quantity α^{p_k} depending on the polynomial degree p_k only such that ψ_i^v is nonnegative in the corresponding element K_k for $\zeta \leq \alpha^{p_k}$. See [A6, Lemma 6.1] in Appendix G for details and proofs. See also Table 5.2 below for the approximate values of α^p.

Moreover, the nonpositivity of the off-diagonal entries of the vertex block of the stiffness matrix is not automatic in the reaction-diffusion case. Similarly, as the nonnegativity of the orthogonalized vertex functions, the nonpositivity of the off-diagonal entries follows for sufficiently small values of $\zeta = c_k h_k^2$. In particular, there exists a quantity β^{p_k} depending on the polynomial degree p_k only such

that $a(\psi_i^v, \psi_j^v) \leq 0$, $i \neq j$, for $\zeta \leq \beta^{p_k}$, where k corresponds to the element $K_k = \operatorname{supp} \psi_i^v \cap \operatorname{supp} \psi_j^v$. See [A6, Lemma 6.2] in Appendix G for details and proofs. See also Table 5.2 below for the approximate values of β^p.

These properties of the orthogonalized basis functions ψ_1^v, ψ_2^v, ..., $\psi_{N^v}^v$ and ψ_1^b, ψ_2^b, ..., $\psi_{N^b}^b$ enable to prove the following direct analogy of Theorem 5.3.

Theorem 5.6. *Let us consider problem (5.19)–(5.21) and its higher-order finite element discretization (5.3). Further let* $c_k h_k^2 \leq \min\{\alpha^{p_k}, \beta^{p_k}\}$ *for all* $k = 1, 2, \ldots, M$. *Then*

$$g_{\mathrm{D},h}(y) - (\Pi_h^0 g_{\mathrm{D},h})(y) \geq 0 \quad \textit{for all } g_{\mathrm{D},h} \in V_h^\partial,\ g_{\mathrm{D},h} \geq 0 \textit{ in } \Omega,\ y \in \Omega.$$

Proof. The proof goes through the same steps as the proof of Theorem 5.3. The only difference is that the standard basis functions φ_1^v, φ_2^v, ..., $\varphi_{N^v}^v$ and φ_1^b, φ_2^b, ..., $\varphi_{N^b}^b$ have to be replaced by the orthogonalized basis functions ψ_1^v, ψ_2^v, ..., $\psi_{N^v}^v$ and ψ_1^b, ψ_2^b, ..., $\psi_{N^b}^b$ as well as the boundary vertex functions φ_1^∂, φ_2^∂ have to be replaced by the corresponding orthogonalized boundary functions ψ_1^∂, ψ_2^∂. □

Theorem 5.6 together with the analysis of the nonnegativity of the discrete Green's function performed in [A6] enables us to formulate the main DMP result. For this purpose, we have to utilize a rational function $\widehat{\omega}^p(\theta, \zeta)$ and two auxiliary quantities γ^p and δ^p, which are defined through $\widehat{\omega}^p(\theta, \zeta)$ in certain way. For details see [A6, Section 7.4].

Theorem 5.7. *Let us consider problem (5.19)–(5.21) and its higher-order finite element discretization (5.3). Denote by* h_k *and* p_k *the lengths and the polynomial degrees of elements* K_k, $k = 1, 2, \ldots, M$. *Further, consider* $\theta^k = h_k/(h_\Omega - h_k)$ *with* $h_\Omega = b^\partial - a^\partial$ *being the length of interval* Ω. *Let us suppose that in case* $\delta^p < \infty$ *inequalities*

$$\gamma^p \geq 3/2 \quad \textit{and} \quad \widehat{\omega}^p(\theta, \gamma^p \theta + \delta^p) \geq 0 \textit{ for } \theta \in (0, 1/2] \tag{5.23}$$

hold true for all $p = p_k$, $k = 1, 2, \ldots, M$. *Furthermore, in case* $\delta^{p_k} < \infty$ *assume*

$$h_k \leq h_\Omega/3 \quad \textit{for all } k = 1, 2, \ldots, M. \tag{5.24}$$

If

$$c_k h_k^2 \leq \min\left\{\alpha^{p_k}, \beta^{p_k}, \gamma^{p_k}\theta^k + \delta^{p_k}\right\} \quad \textit{for all } k = 1, 2, \ldots, M, \tag{5.25}$$

then discretization (5.3) satisfies the discrete conservation of nonnegativity.

Proof. We apply Theorem 3.3. Assumption (a) of this theorem follows from the analysis performed in [A6] – see Corollary 7.1 and Lemma 7.4 in Appendix G. Assumption (b) is guaranteed by Theorem 5.6 proved above. □

The rather complicated assumptions of Theorem 5.7 deserve certain comments. The main point is that assumption (5.25) only is fundamental. The other assumptions (5.23) and (5.24) are technical. Especially, assumption (5.23) might be superfluous, because it can be a priori verified for all values of p. Its verification for $p = 1$ and 2 is easy. However, for higher values of p we have not succeeded to prove it theoretically. Therefore, we verified the validity of (5.23) computationally. We employed the interval arithmetic and confirmed its validity for p up to 10. See [A6] in Appendix G for details.

We note that the lowest-order case $p = 1$ was already analyzed in Section 4.3. We point out that condition (5.25) for $p = 1$ reduces to condition (4.16) of Theorem 4.7 for the considerted diffusion-reaction problem. The difference is that in the lowest-order case we proved sufficiency and necessity of condition (4.16), but Theorem 5.7 proves sufficiency only.

Table 5.2: The critical values α^p, β^p, γ^p, and δ^p.

p	α^p	β^p	γ^p	δ^p
1	∞	6	0	∞
2	20/3	∞	0	∞
3	38.61	25.89	5.608	0
4	18.91	∞	2.936	3.614
5	49.44	59.82	7.799	0
6	37.56	∞	7.247	0.887
7	72.82	107.81	9.791	0
8	62.62	∞	9.709	0
9	104.09	169.85	11.510	0
10	94.10	∞	10.644	0

For the reader's convenience we also reprint Table 5.2 from [A6]. The table contains numerically computed values of α^p, β^p, γ^p, δ^p for $p = 1, 2, \ldots, 10$. Observing these values yields to certain conclusions and hypothesis. First of all, condition (5.25) for $p = 1$ and $p = 2$ reduces to $ch_k^2 \leq 6$ and $ch_k^2 \leq 6 + 2/3$, respectively. (These values coming from Table 5.2 are exact!) Interestingly, the limitation for $p = 2$ is slightly weaker than for $p = 1$. Since the condition for $p = 1$ is also sufficient (see Theorem 4.7), we conclude that in this case the quadratic elements provide the DMP for slightly wider class of meshes than the linear elements. This is exceptional, however. In general the higher-order finite elements perform much worse with respect to the DMP than the lowest-order ones.

Further observations concern the values $p \geq 3$. It seems that the value of γ^p is the smallest for $p = 4$ and it is well above 3/2. Hence, the first inequality in

the assumption (5.23) seems to be satisfied for all values of p. Furthermore, the values in Table 5.2 show that $\gamma^p\theta + \delta^p \leq \min\{\alpha^p, \beta^p\}$ for $\theta \in (0, 1/2]$ and for $p = 3, 4, \ldots, 10$. (Notice that due to assumption (5.24) the quantity θ^k attains values in $(0, 1/2]$ only.) The observed trend allows to conjecture that this is the case for any $p > 10$, too. If this is true then condition (5.25) can be replaced for $p \geq 3$ by simpler one:

$$c_k h_k^2 \leq \gamma^{p_k}\theta^k + \delta^{p_k} \quad \text{for all } k = 1, 2, \ldots, M.$$

It can be shown, see [A6], that for the validity of this condition it suffices to have $c_k h_\Omega h_k \leq \gamma^{p_k} + \delta^{p_k}$. Hence, we can say that the DMP is satisfied provided the finite element mesh is sufficiently fine.

Finally, the data in Table 5.2 imply that the cubic elements yield the most strict limitations to the element sizes. Hence, if this is true also for $p > 10$ and if the above hypothesis are valid also for arbitrary $p > 10$, then the following conjecture holds true.

Conjecture 5.8. *Let us consider problem (5.19)–(5.21) and its higher-order finite element discretization (5.3) based on a finite element mesh consisting of M elements. Suppose arbitrary distribution of polynomial degrees p_k, $k = 1, 2, \ldots, M$. Denote by h_k the length of the element K_k and set $\theta^k = h_k/(h_\Omega - h_k)$, where $h_\Omega = b^\partial - a^\partial$ stands for the length of the interval Ω. If*

$$c_k h_k^2/\gamma^3 \leq \theta^k \leq 1/2 \quad \textit{for all } k = 1, 2, \ldots, M,$$

where $\gamma^3 \approx 5.608797$, then discretization (5.3) of problem (5.19)–(5.21) satisfies the discrete conservation of nonnegativity.

5.4 Higher-order discrete maximum principle in two-dimensions

The validity of the DMP for two (and higher) dimensions and for higher-order finite element approximations is a difficult problem. Up to the author's knowledge practically none positive result is known in this field. Therefore, we restrict ourselves in this section to the Poisson problem with homogeneous Dirichlet boundary conditions.

There is a result [39] from 1981 analyzing a local version of the DMP for higher-order finite elements. The authors of [39] require the validity of the DMP on each vertex-patch of elements $\omega(\boldsymbol{x}_i) = \{K \in \mathcal{T}_h : \boldsymbol{x}_i \in K\}$, where $\boldsymbol{x}_i$ is a vertex in the triangulation $\mathcal{T}_h$. This local version of the DMP is stronger than

the global version we analyze in this thesis, i.e. the validity of the local DMP implies the validity of the global one. Anyway, they show that the local DMP is not satisfied for quadratic triangular elements unless the mesh is very special – consisting of all equilateral triangles or the so-called quadratic mesh consisting of right-angle triangles. They also note that the local DMP fails for cubic elements even on the quadratic mesh.

Another result about the DMP for higher-order approximations is published in [85], but it concerns the collocation method which is not of interest in this thesis.

Certain numerical experiments about the validity of the DMP for the higher-order triangular finite elements were published by the author of this thesis in [A7]. This paper is attached as Appendix H. The experiments are based on the general theory presented in Chapter 3. In particular, the nonnegativity of the discrete Green's function is directly tested. The results indicate that the DMP is satisfied for quadratic finite elements on meshes consiting of nearly equilateral triangles and that the DMP is not satisfied at all for the polynomial degree three and higher.

In this section we review these numerical experiments and present more of them in order to draw a broader picture of the situation.

In the experiments we consider the Poisson problem in a polygon Ω with homogeneous Dirichlet boundary conditions:

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{in } \partial\Omega.$$

Its higher-order finite element discretization fits well to the general framework presented in Section 3.1, see (3.2). We consider triangular meshes and the corresponding finite element space is

$$V_h = \{v_h \in H_0^1(\Omega) : v_h|_K \in \mathbb{P}^{p_K}(K) \quad \forall K \in \mathcal{T}_h\},$$

where $\mathbb{P}^{p_K}(K)$ stands for a space of polynomials of degree at most p_K on the triangle K. The polynomial degrees p_K are assumed to be a priori given (sometimes they are considered as inseparable attributes of the mesh $\mathcal{T}_h$).

The sufficient and necessary conditions for the validity of the DMP are stated in Theorem 3.3. Since the discrete Dirichlet lift $g_{\mathrm{D},h}$ vanishes, the condition (a) of this theorem only applies, i.e. the DMP is satisfied if and only if the corresponding discrete Green's function (DGF) is nonnegative. The DGF G_h can be expressed by formula (3.15). This requires the basis functions of V_h and the inverse of the corresponding stiffness matrix. If the number of degrees of freedom is low then we can compute this inverse by standard numerical procedures. Subsequently, formula (3.15) can be used for the inspection of the nonnegativity of the DGF G_h.

Formula (3.15) should be used with care, however. It is advantageous to utilize the properties of the higher-order basis functions. As in the 1D case, we split the higher-order basis functions of V_h into two groups. The first consits of the standard piecewise linear and continuous nodal basis function. We denote them by $\varphi_1^v, \varphi_2^v, \ldots, \varphi_{N^v}^v$ and call them the vertex functions. In this case, the number N^v denotes the number of interior vertices in the triangulation $\mathcal{T}_h$. The vertex functions have the well-known delta-property

$$\varphi_i^v(\boldsymbol{x}_j) = \delta_{ij} \quad \forall i, j = 1, 2, \ldots, N^v, \tag{5.26}$$

where δ_{ij} stands for the Kronecker's tensor and $\boldsymbol{x}_j$ stand for the interior nodes (vertices) of the triangulation $\mathcal{T}_h$.

The other – non vertex – basis functions are of higher-order and they are denoted by $\varphi_1^n, \varphi_2^n, \ldots, \varphi_{N^n}^n$. Altogether, $N^v + N^n = \dim V_h$. These higher-order basis functions consist in 2D of *edge functions* and *bubble functions* [72]. The common property of all higher-order basis functions is the fact that they vanish at all vertices $\boldsymbol{x}_j$ of the triangulation $\mathcal{T}_h$.

The union of the vertex and higher-order functions forms a basis of V_h. In certain situations it will be convenient to denote this basis by $\varphi_1, \varphi_2, \ldots, \varphi_{N^v+N^n}$, where $\varphi_i = \varphi_i^v$ for $i = 1, 2, \ldots, N^v$ and $\varphi_{N^v+i} = \varphi_i^n$ for $i = 1, 2, \ldots, N^n$. As in the 1D case, the presence of these two groups of basis functions leads to a 2×2 block structure of the stiffness matrix. However, the orthogonalization of the vertex functions with respect to the higher-order functions, we performed in the 1D diffusion-reaction case, is a nonlocal operation in 2D. This is due to the presence of the edge-functions. Hence this orthogonalization is not practical to do and all the four blocks of the stiffness matrix remain nonzero. Thus, the stiffness matrix and its inverse have the following form

$$A = \begin{pmatrix} A^{vv} & A^{vn} \\ A^{nv} & A^{nn} \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} S^{-1} & -(A^{vv})^{-1}A^{vn}R^{-1} \\ -(A^{nn})^{-1}A^{nv}S^{-1} & R^{-1} \end{pmatrix},$$

where $A^{vv} \in \mathbb{R}^{N^v \times N^v}$, $A^{nn} \in \mathbb{R}^{N^n \times N^n}$, etc., $S = A^{vv} - A^{vn}(A^{nn})^{-1}A^{nv}$, and $R = A^{nn} - A^{nv}(A^{vv})^{-1}A^{vn}$.

Since all the higher-order basis functions vanish at the vertices $\boldsymbol{x}_j$ of the triangulation we easily obtain by (3.15) and (5.26) the indentity

$$G_h(\boldsymbol{x}_i, \boldsymbol{x}_j) = (A^{-1})_{ij}\varphi_i^v(\boldsymbol{x}_i)\varphi_j^v(\boldsymbol{x}_j) = (A^{-1})_{ij} = (S^{-1})_{ij}$$

for all pairs of interior vertices $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $i = 1, 2, \ldots, N^v$. This clearly shows that the vertex values of the DGF G_h coincide with the entries of the inverse of the Schur complement S.

Furthermore, the DGF has a natural structure given by the Cartesian product of the mesh $\mathcal{T}_h$ with itself. In particular, if K and L are two elements from $\mathcal{T}_h$

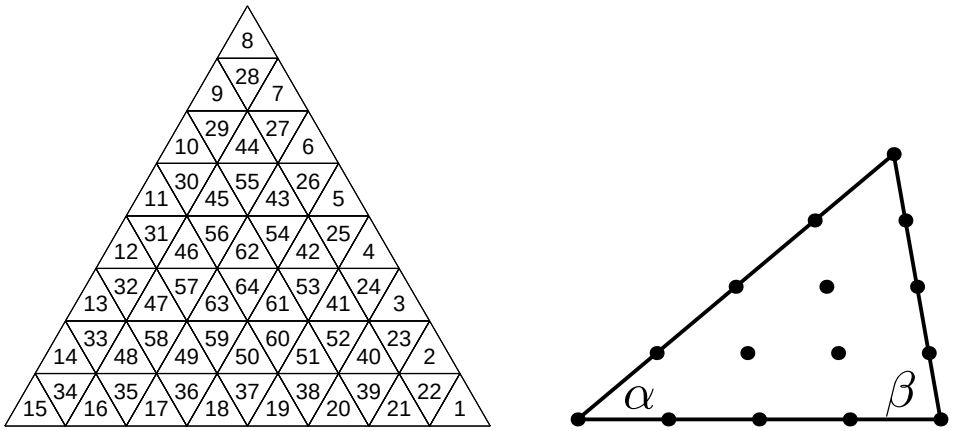


Figure 5.3: A uniform mesh with 64 triangles enumerated in a spiral way (left). A triangular element characterized by a pair of angles α and β (right). In addition, sample points for $N_{\mathrm{spl}} = 4$ are indicated.

and $I(K)$ and $I(L)$ denote the sets of indices of basis functions supported in K and L, respectively, i.e. $I(K) = \{i \in \mathbb{N} : 1 \leq i \leq N^v + N^n, K \subset \operatorname{supp} \varphi_i\}$ as in Section 3.1, then the DGF restricted to $K \times L$ is given by

$$G_h|_{K\times L}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i\in I(K)} \sum_{j\in I(L)} (A^{-1})_{ij} \varphi_i|_K(\boldsymbol{x}) \varphi_j|_L(\boldsymbol{y}), \quad (\boldsymbol{x}, \boldsymbol{y}) \in K \times L. \tag{5.27}$$

This formula contains a small number of basis functions and we use it for fast evaluation of the DGF at a given point.

Experiment 1: Nonnegativity of the DGF

In this experiment we try to reveal how the nonnegativity of the DGF depends on angles in the finite element triangulation. We consider a triangular domain Ω covered by the uniform triangular mesh $\mathcal{T}_h$ consiting of 64 triangles, see Figure 5.3 (left). The distribution of the polynomial degrees over $\mathcal{T}_h$ is assumed to be constant, i.e. $p_K = p$ for all $K \in \mathcal{T}_h$.

We denote by α and β two angles in the triangle Ω and below we test the dependence of the nonnegativity of the DGF G_h on these angles. The angles α and β completely determine the shape of Ω and we point out that the size of Ω is irrelevant for the nonnegativity of G_h. In Figure 5.4 (as well as in the subsequent figures) the horizontal and vertical axes correspond to the angles α and β. In each direction we consider 179 discrete values ranging uniformly from 1° to 179°. We consider points in these axis given by all pairs of these discrete values such

that $\alpha + \beta < 180°$. Each of these points correspond to a particular shape of the triangle Ω and we determine its color according to the properties of the DGF G_h on this domain Ω.

Nevertheless, checking the nonnegativity of the DGF is a difficult task. It is natural to check it element by element, but the DGF $G_h(\boldsymbol{x}, \boldsymbol{y})$ restricted to a pair $K, L \in \mathcal{T}_h$, $\boldsymbol{x} \in K$, $\boldsymbol{y} \in L$, is a multivariate polynomial and the verification of nonnegativity of a multivariate polynomial is connected to the 17th Hilbert problem, as we already mentioned at the beginning this chapter. For our purposes, it suffices to verify the nonnegativity in an approximate way only. For each triangle $K \in \mathcal{T}_h$ we introduce a set of sample points $\boldsymbol{s}_{k\ell}^K$ with barycentric coordinates $(k, \ell, N_{\text{spl}} - k - \ell)/N_{\text{spl}}$, $0 \leq k + \ell \leq N_{\text{spl}}$, see Figure 5.3 (right). The total number of these sample points in an element is $(N_{\text{spl}} + 1)(N_{\text{spl}} + 2)/2$. Finally, instead of checking the nonnegativity of G_h everywhere in $K \times L$, we check it for all sample points $(\boldsymbol{s}_{k\ell}^K, \boldsymbol{s}_{mn}^L)$ only.

Now, we can explain how do we color the points in Figures 5.4. If the DGF $G_h(\boldsymbol{x}, \boldsymbol{y})$ is nonnegative at all sample points over entire Ω^2 then the color is black. If there is a sample point, where G_h is negative, we color the corresponding point according to the vertex values given by the Schur complement S. If S is M-matrix then the color is green. If S is monotone but not M-matrix then the color is red. If S is not monotone and hence, the DGF G_h is negative in some vertex point, then the color is blue.

However, this coloring is valid for $p \geq 2$ only. The lowest-order case $p = 1$ is exceptional because there are no higher-order basis functions, we have $A = S$, and the DMP is satisfied if and only if A is monotone. Therefore, we use the following colors for $p = 1$. If A is M-matrix then the color is orange. If A is monotone but not M-matrix then the color is gray. If A is not monotone then the color is light blue. Clearly, the orange and gray colors mean the validity of the DMP for $p = 1$.

Thus, the black color in Figures 5.4 for $p \geq 2$ means that the DMP is probably satisfied for the corresponding angles α and β. We cannot assure this because the nonnegativity at all sample points does not guarantee nonnegativity everywhere. Nevertheless, in order to ensure that the number of sample points is sufficient, we always perform a series of computations starting with $N_{\text{spl}} = 8$ and doubling N_{spl} until the final results (the pictures) do not change.

Nevertheless, the colors other than black in Figures 5.4 for $p \geq 2$ mean that the DMP is definitely not satisfied. From these results we immediate observe that the DMP can be satisfied for polynomial degrees $p = 1$ and $p = 2$ only. The higher polynomial degrees do not lead to the DMP for any angles of Ω in this setting. For polynomial degrees higher than two the DMP is not satisfied even on the mesh consisting of all equilateral triangles. We also experimented with other than triangular domains. The resulting figures were similar to those shown

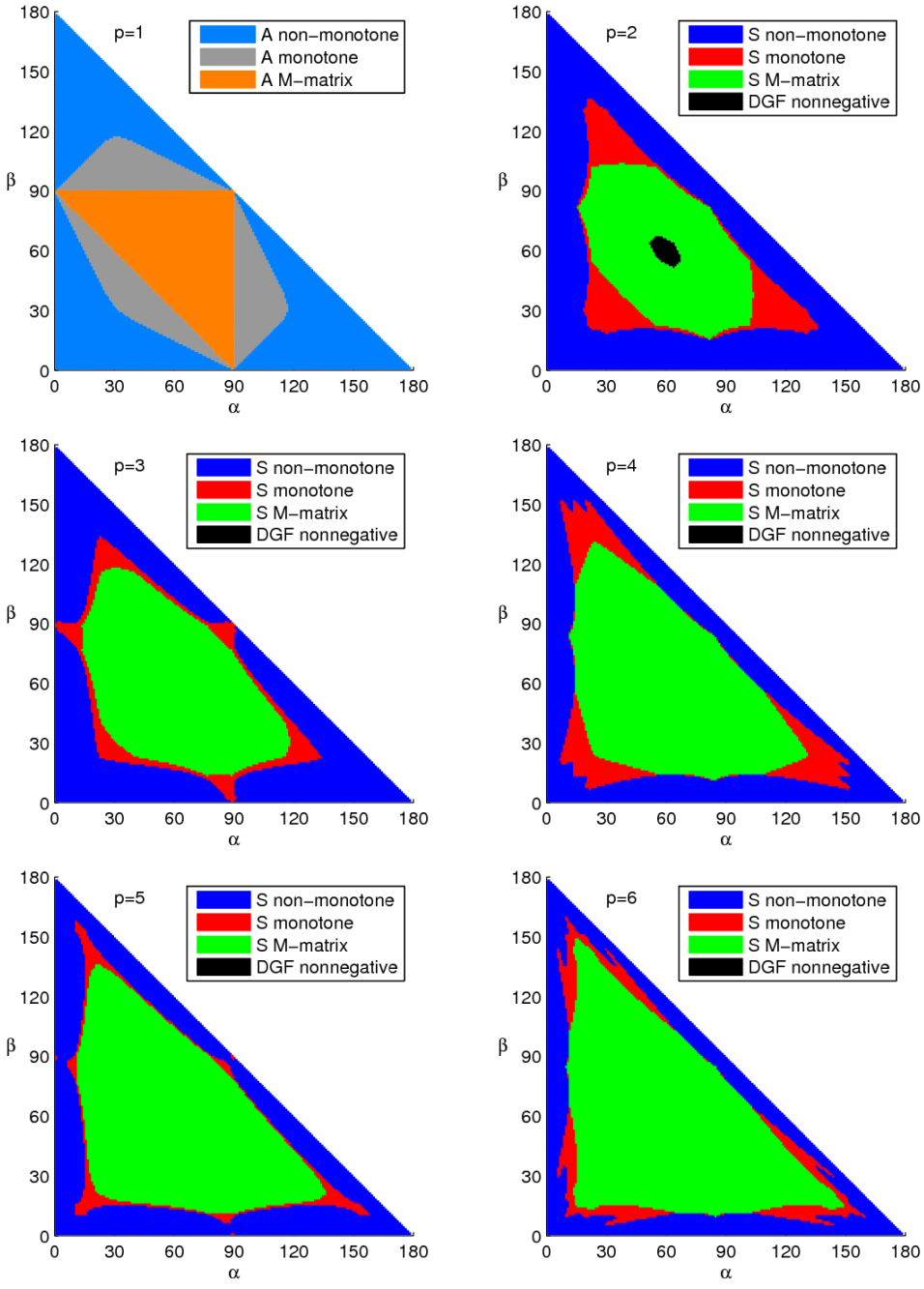


Figure 5.4: The nonnegativity of the DGF and its dependence on the angles in the triangulation for polynomial degrees $p = 1, 2, \ldots, 6$.

in Figure 5.4. Therefore, we dare to conjecture that the DMP is not satisfied in general for finite elements of order three and higher.

Even for $p = 2$ the DMP is (hopefully) satisfied in exceptional cases only. Our experiment indicates that it is satisfied on triangular meshes, where all elements are close to the equilateral triangle.

However, the results of this experiment do not mean that the higher-order finite elements are completely hopeless with respect to the DMP. In some sense they behave well. We observe that the higher polynomial degrees lead to larger sizes of the green and red areas in Figure 5.4. It means that the vertex values of the DGF G_h are nonnegative for wider range of angles if the polynomial degree increases. Hence, the negative values of the DGF are attained somewhere inside of the elements or on their edges.

Experiment 2: DGF in the boundary and interior regions

The results of the previous experiment are pessimistic in the sense that the DGF turned out to possess negative values for almost all triangulations in the higher-order cases. Nevertheless, a closer look to the DGF reveals that the negative values of the DGF appear close to the boundary of Ω (see Experiment 3 below). Therefore, we split the domain Ω into the boundary and interior regions. The boundary region $\Omega_{\mathcal{B}}$ contains a layer of elements adjacent to the boundary $\partial\Omega$ and the interior region $\Omega_{\mathcal{I}}$ contains the other (interior) elements. The rigorous definitions are

$$\Omega_{\mathcal{B}} = \bigcup\{K \in \mathcal{T}_h : K \cap \partial\Omega \neq \emptyset\}, \qquad \Omega_{\mathcal{I}} = \bigcup\{K \in \mathcal{T}_h : K \cap \partial\Omega = \emptyset\}.$$

In this experiment we try to investigate the nonnegativity of the DGF in the interior region and its dependence on the angles in the triangulation. More precisely, we investigate the nonnegativity of $G_h(\boldsymbol{x}, \boldsymbol{y})$ in $\Omega_{\mathcal{I}}^2$ and in $\Omega \times \Omega_{\mathcal{I}}$. The nonnegativity of the DGF in these two domains has certain consequences about the nonnegativity of the finite element solution in $\Omega_{\mathcal{I}}$. We formulate them in the following theorem.

Theorem 5.9. *Let us consider the general elliptic problem (2.10) with homogeneous Dirichlet and Neumann boundary conditions, i.e. with $g_{\mathrm{D}} = 0$ on Γ_{D} and $g_{\mathrm{N}} = 0$ on Γ_{N}. Further, let us consider the corresponding finite element approximation (3.2) and the DGF G_h given by (3.11). Then the property*

$$f \geq 0 \text{ a.e. in } \Omega \quad \Rightarrow \quad u_h \geq 0 \text{ in } \Omega_{\mathcal{I}} \tag{5.28}$$

is equivalent to the nonnegativity of G_h in $\Omega \times \Omega_{\mathcal{I}}$. Similarly, the property

$$f \geq 0 \text{ a.e. in } \Omega_{\mathcal{I}} \text{ and } f = 0 \text{ a.e. in } \Omega_{\mathcal{B}} \quad \Rightarrow \quad u_h \geq 0 \text{ in } \Omega_{\mathcal{I}} \tag{5.29}$$

is equivalent to the nonnegativity of G_h in $\Omega_{\mathcal{I}}^2$.

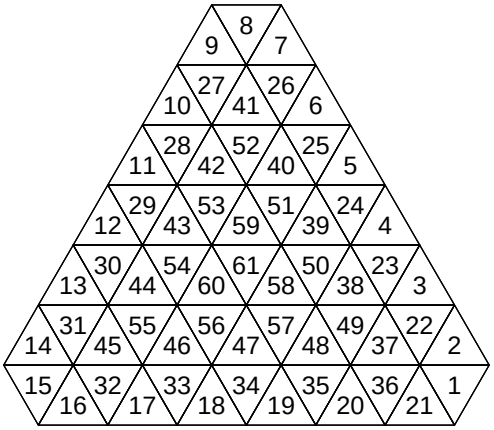


Figure 5.5: A triangle without corners, its triangulation, and the enumeration of elements in a spiral way.

Proof. This is a direct consequence of the representation formula (3.14). □

Properties (5.28) and (5.29) require nonnegativity of the approximate solution u_h in the interior region $\Omega_{\mathcal{I}}$ only. Therefore, they are clearly weaker variants of the conservation of nonnegativity (see Definition 3.3) provided the Dirichlet and Neumann boundary conditions are homogeneous. Under this assumption, the conservation of nonnegativity implies property (5.28) and property (5.28) implies (5.29).

Let us note that in this section we experiment with Laplacian only and hence we can assume the symmetry of the DGF, i.e. $G_h(\boldsymbol{x}, \boldsymbol{y}) = G_h(\boldsymbol{y}, \boldsymbol{x})$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega^2$. Thus, due to this symmetry, the nonnegativity of G_h in $\Omega \times \Omega_{\mathcal{I}}$ is equivalent to the nonnegativity of G_h in $\Omega^2 \setminus \Omega^2_{\mathcal{B}}$.

In Figure 5.6 we present the results of an experiment similar to Experiment 1. We consider the triangular domain Ω and construct the DGF G_h for many pairs of angles α and β in the same way as in Experiment 1. The difference is that now we color the corresponding points (α, β) according to the nonnegativity of the DGF in $\Omega \times \Omega_{\mathcal{I}}$ and in $\Omega^2_{\mathcal{I}}$. In particular, if $G_h \geq 0$ in Ω^2 then the color is black. If not and if $G_h \geq 0$ in $\Omega \times \Omega_{\mathcal{I}}$ then the color is yellow. Otherwise, if $G_h \geq 0$ in $\Omega^2_{\mathcal{I}}$ then the color is magenta. If none of these conditions is satisfied then $G_h < 0$ at some point of $\Omega^2_{\mathcal{I}}$ and the color is cyan. We point out that the boundary region $\Omega_{\mathcal{B}}$ is formed by the elements with indices $1, 2, \ldots, 39$ and the interior region $\Omega_{\mathcal{I}}$ consists of elements with indices $40, 41, \ldots, 64$.

In contrast to Experiment 1, the results of Experiment 2 depend on the domain Ω. More precisely, we identified quite strong dependence on the presence of *corner elements*. These elements have all their vertices on the boundary $\partial\Omega$

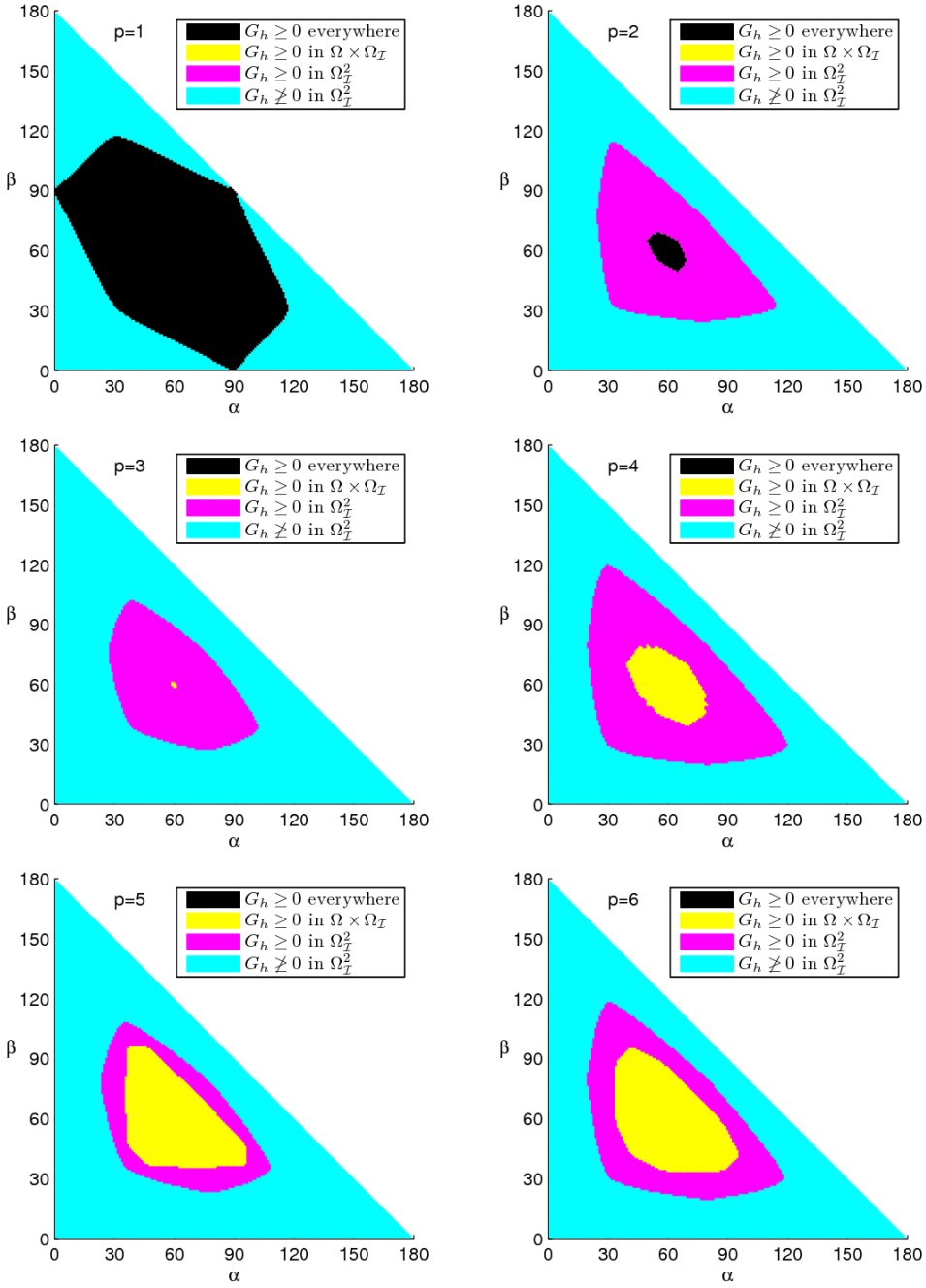


Figure 5.6: Nonnegativity of the DGF in the interior region for polynomial degrees $p = 1, 2, \ldots, 6$. The tested domain is a triangle, see Figure 5.3 (left). We observe the dependence of this nonnegativity on the angles in the triangulation.

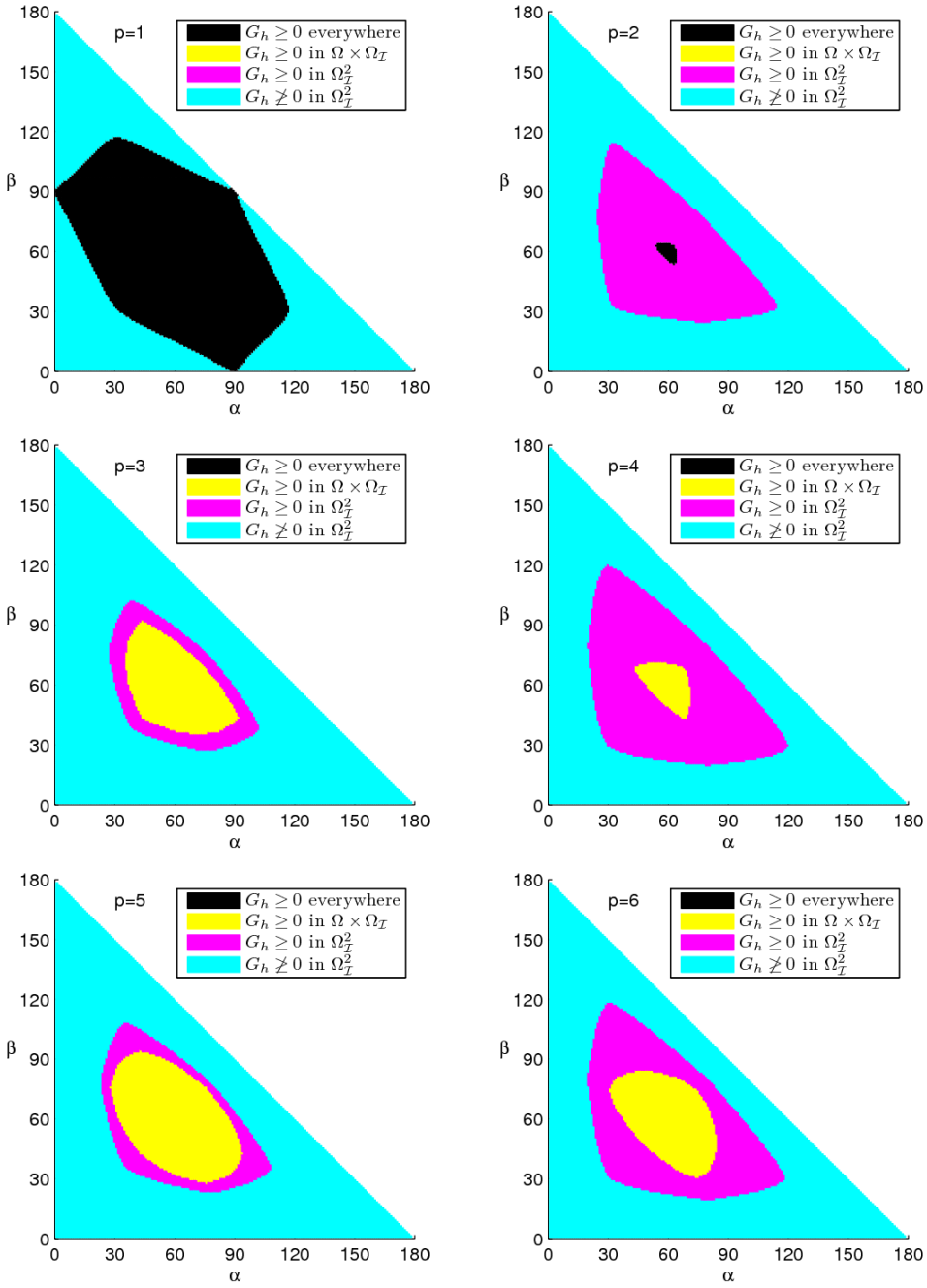


Figure 5.7: Nonnegativity of the DGF in the interior region for polynomial degrees $p = 1, 2, \ldots, 6$. The tested domain is a triangle without corners, see Figure 5.5. We observe the dependence of this nonnegativity on the angles in the triangulation.

– see elements 1, 8, and 15 in Figure 5.3 (left). Therefore, we present also the results obtained in a domain (and with a triangulation), where these elements are removed. The modified domain is a triangle without corners and it is depicted together with the used mesh and with the enumeration of the elements in Figure 5.5. The results in Figure 5.7 use the same color code as in Figure 5.6. In this case, the boundary region $\Omega_{\mathcal{B}}$ is formed by elements $1, 2, \ldots, 36$ and the interior region $\Omega_{\mathcal{I}}$ by elements $37, 38, \ldots, 61$.

Observing the results in Figures 5.6 and 5.7 we may conclude that the case $p = 1$ is again exceptional. The DGF for $p = 1$ is either nonnegative everywhere or it has negative values even in the interior region. In the case $p = 2$ we observe a small area of black points. Thus, if the angles in the triangulation are close to 60° then the DGF is nonnegative everywhere in Ω^2 as we already know from Experiment 1. Further, we observe no yellow points (the DGF nonnegative in $\Omega \times \Omega_{\mathcal{I}}$) and we see relatively large magenta area with the DGF nonnegative in $\Omega^2_{\mathcal{I}}$. The triangulations corresponding to this magenta area require the minimal angle to be greater then roughly 30° and allow for the maximal angle to be at most about 120°. For higher polynomial degrees ($p > 2$) we see no black points (this was already shown in Experiment 1). The yellow area is highly influenced by the polynomial degree p and by the presence of the corner elements – compare Figures 5.6 and 5.7. On the other hand, the magenta area is much more stable. Interestingly, if we concentrate on odd polynomial degrees only, the magenta area seems to be slightly growing when p is increasing. Similarly, it seems that it is growing for even polynomial degrees, too. We also observe that the magenta area for odd polynomial degrees is in general smaller than for even degrees.

From these results we may draw the following conclusions. The yellow area corresponds to the property (5.28) (any nonnegative f yields u_h nonnegative in the interior region $\Omega_{\mathcal{I}}$). This property does not seem to be suitable for further investigations, because its validity is dramatically changing with the used polynomial degree and with the presence of the corner elements. Nevertheless, for polynomial degrees 5 and higher our results indicate that this property is satisfied for reasonable wide range of triangulations – say for triangulations with minimal angle greater than roughly 35°–40° and with the maximal angle smaller then about 90°. On the other hand, the magenta area – corresponding to the property (5.29) – seems to be much more stable with respect to the varying polynomial degree. In addition, we have observed no change of this area when we removed the corner elements. Furthermore, the magenta area is quite large for all polynomial degrees $p \geq 2$. Therefore, we may conclude that property (5.29) (i.e. the nonnegativity of u_h in the interior region $\Omega_{\mathcal{I}}$ under the condition that f vanishes in the boundary region $\Omega_{\mathcal{B}}$ and it is nonnegative in the interior region $\Omega_{\mathcal{I}}$) seems to be satisfied for a fair range of triangulations and any polynomial degree. Based on our experiments we can quantify these triangulations as those with the

minimal angle greater than 30°–40° and with the maximal angle at most roughly 100°. In general, these angle conditions might be weaken for approximations of higher orders.

Experiment 3: Visualization of the DGF

To see the behavior of the DGF in more details, we tried to visualize it. It is quite a hard task, because the graph of the DGF G_h is a five-dimensional object. Nevertheless, we can consider pairs of elements $K_i, K_j \in \mathcal{T}_h$, $i, j = 1, 2, \ldots, M$, where M denotes the number of elements in $\mathcal{T}_h$. A pair (i, j) corresponds to a point in planar axis and we can color this point according to certain characteristic of the DGF G_h in $K_i \times K_j$. In Figures 5.8–5.19 we color the points according to the minimum of $G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and according to the fraction of the area, where G_h is negative, i.e. according to $\mathrm{meas}\{(\boldsymbol{x}, \boldsymbol{y}) \in K_i \times K_j : G_h(\boldsymbol{x}, \boldsymbol{y}) < 0\} / \mathrm{meas}(K_i \times K_j)$ (bottom-right). The top-left panel of these figures illustrates the corresponding domain Ω and the triangulation. As a benchmark we have chosen the polynomial degree $p = 3$ in all these figures. We point out that all these characteristics are not computed exactly. We use approximations based on the sample points – see Experiment 1.

The first six figures (5.8–5.13) correspond to a triangular domain Ω with various choices of angles. The subsequent six figures (5.14–5.19) correspond to the triangular domain without corners. Let us point out the enumeration of elements presented in top-left panels of these figures. For the triangular domain the elements adjacent to the boundary have indices 1–39 and the interior elements have indices 40–64. For the triangular domain without corners the elements adjacent to the boundary have indices 1–36 and the interior elements have indices 37–61.

A general observation from Figures 5.8–5.19 is that the meshes consisting of equilateral triangles provide the smallest magnitudes of the negative values and the smallest areas of negative values. The more the triangles differ from the equilateral one the more negative the values of G_h are and the larger the areas of negative values are.

The right panels of Figure 5.8 reveal that G_h is negative in small number of cases. Namely, the negative values only appear if the two elements in the pair are adjacent to the boundary and if they are neighboring to each other. This rule however applies in the case of equilateral triangle only. In the subsequent figures, we can observe how the area of negative values increases when the extremal angles differ more and more from 60°.

We also observe in the bottom-right panels that if the DGF G_h is negative somewhere in $K_i \times K_j$, $K_i, K_j \in \mathcal{T}_h$ then the area of the domain, where G_h is negative, is relatively small. If the angles are close to 60° then the fraction of this

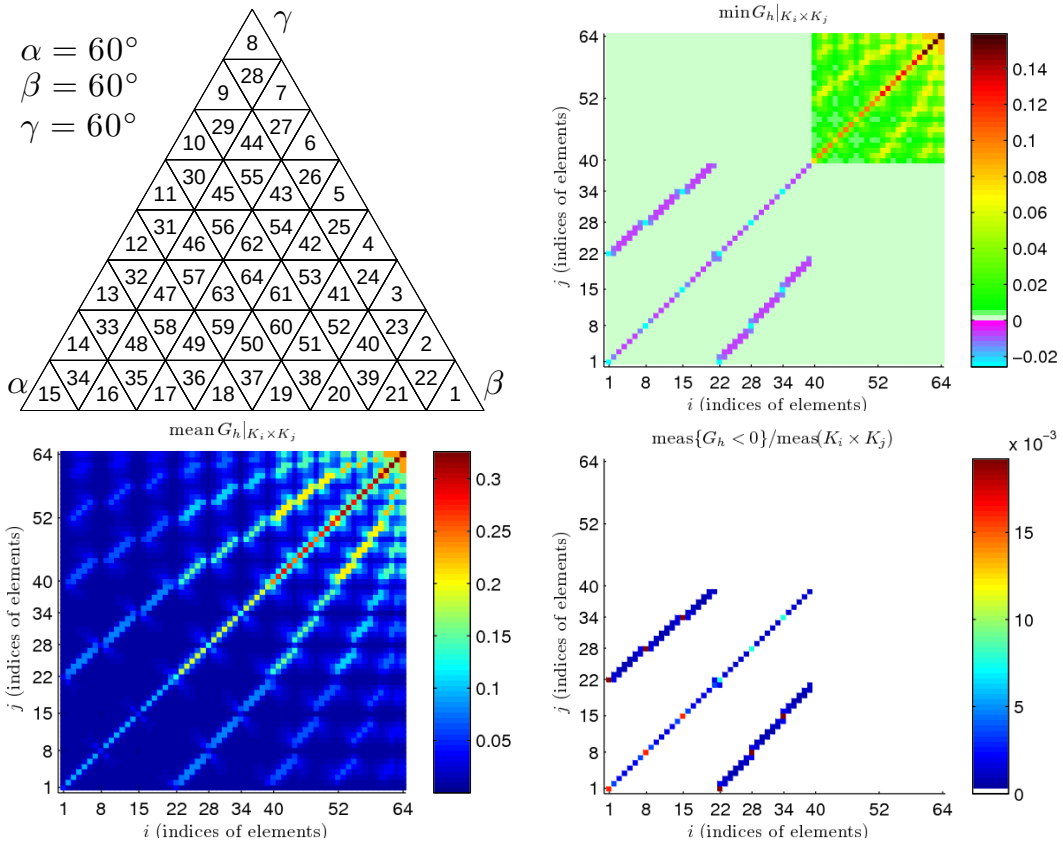


Figure 5.8: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 60°, 60°, 60°) and the triangulation with the enumeration of elements. The polynomial degree is $p = 3$.

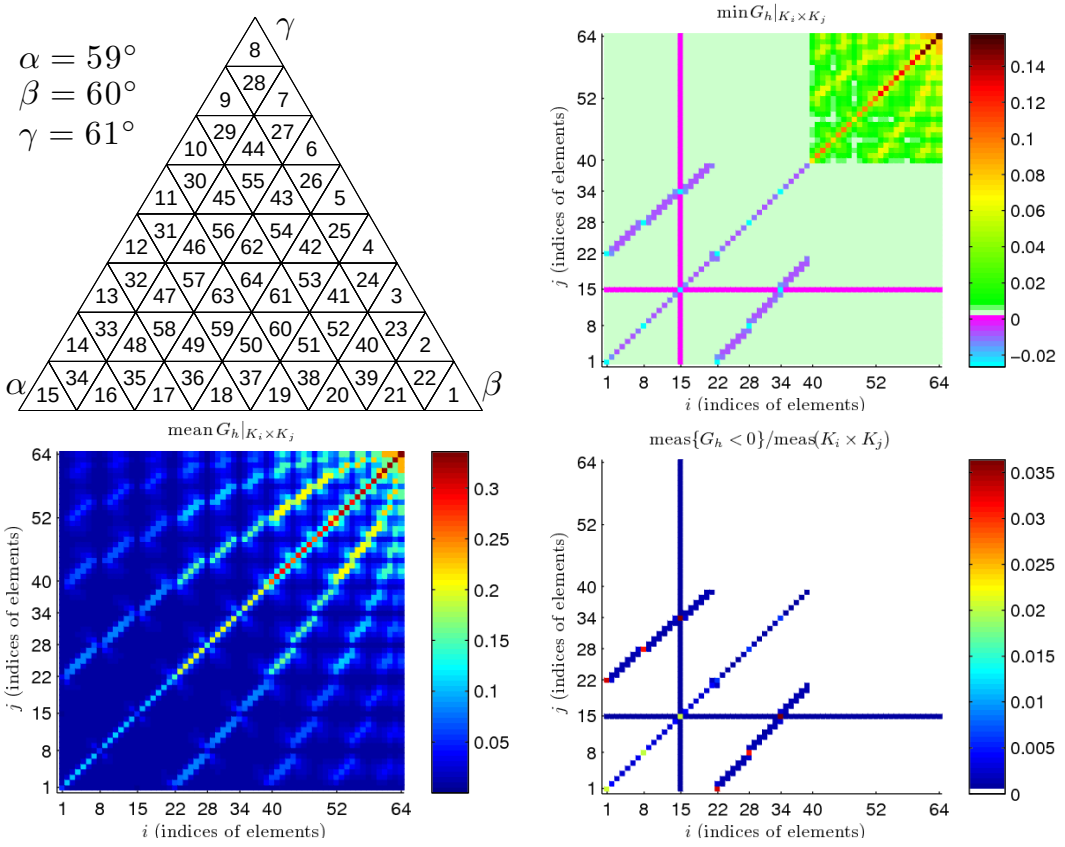


Figure 5.9: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 59°, 60°, 61°) and the triangulation with the enumeration of elements. The polynomial degree is $p = 3$.

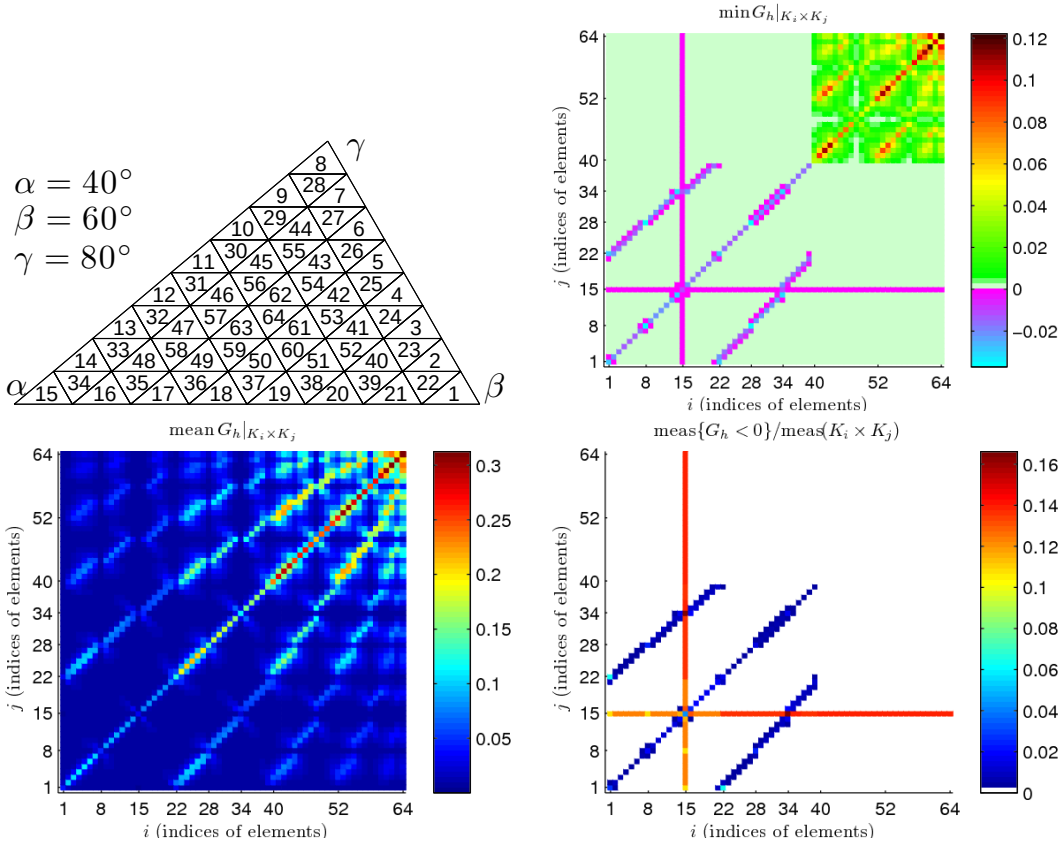


Figure 5.10: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 40°, 60°, 80°) and the triangulation with the enumeration of elements. The polynomial degree is $p = 3$.

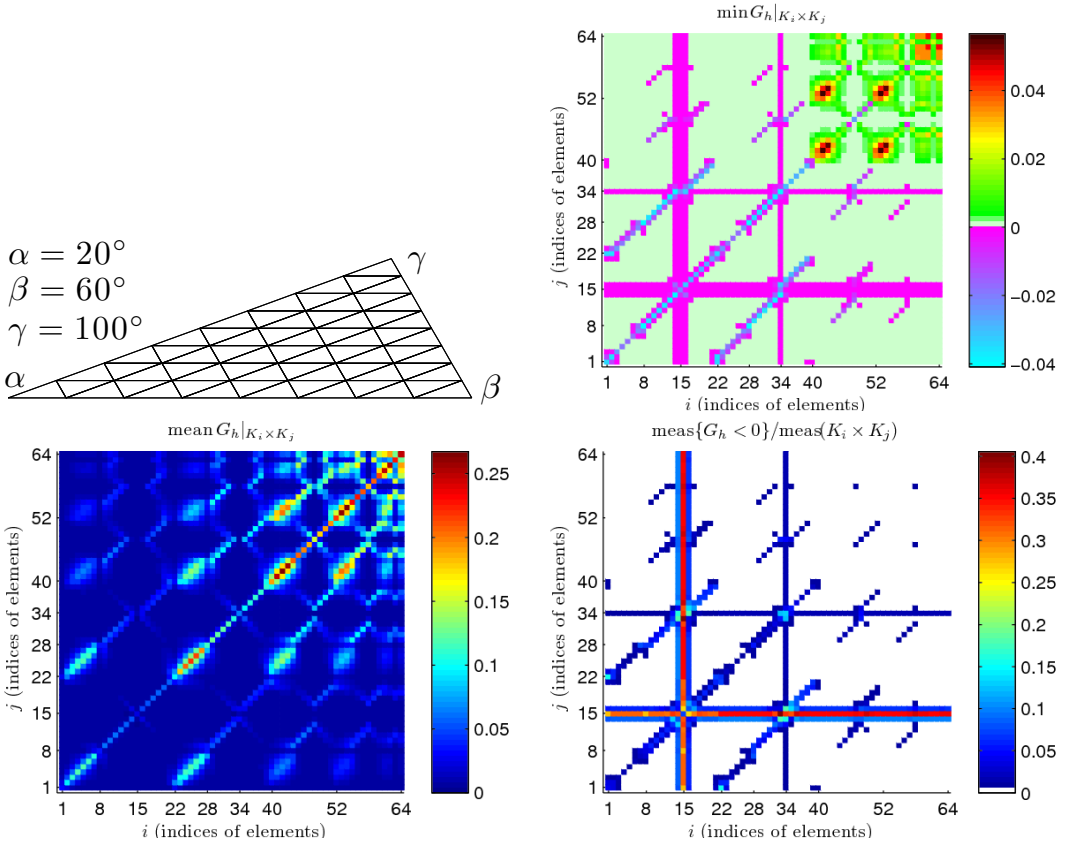


Figure 5.11: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 20°, 60°, 100°) and the triangulation. The enumeration of elements follows the same pattern as in Figures 5.8–5.10. The polynomial degree is $p = 3$.

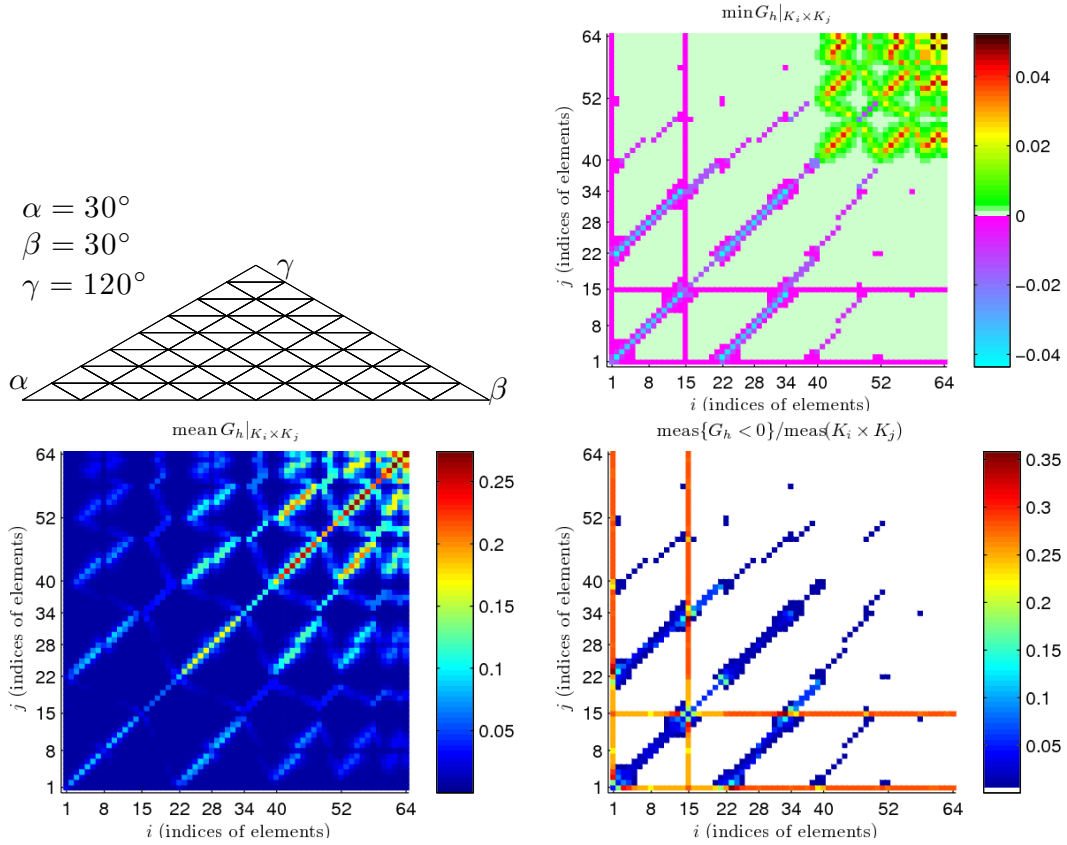


Figure 5.12: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 30°, 30°, 120°) and the triangulation. The enumeration of elements follows the same pattern as in Figures 5.8–5.10. The polynomial degree is $p = 3$.

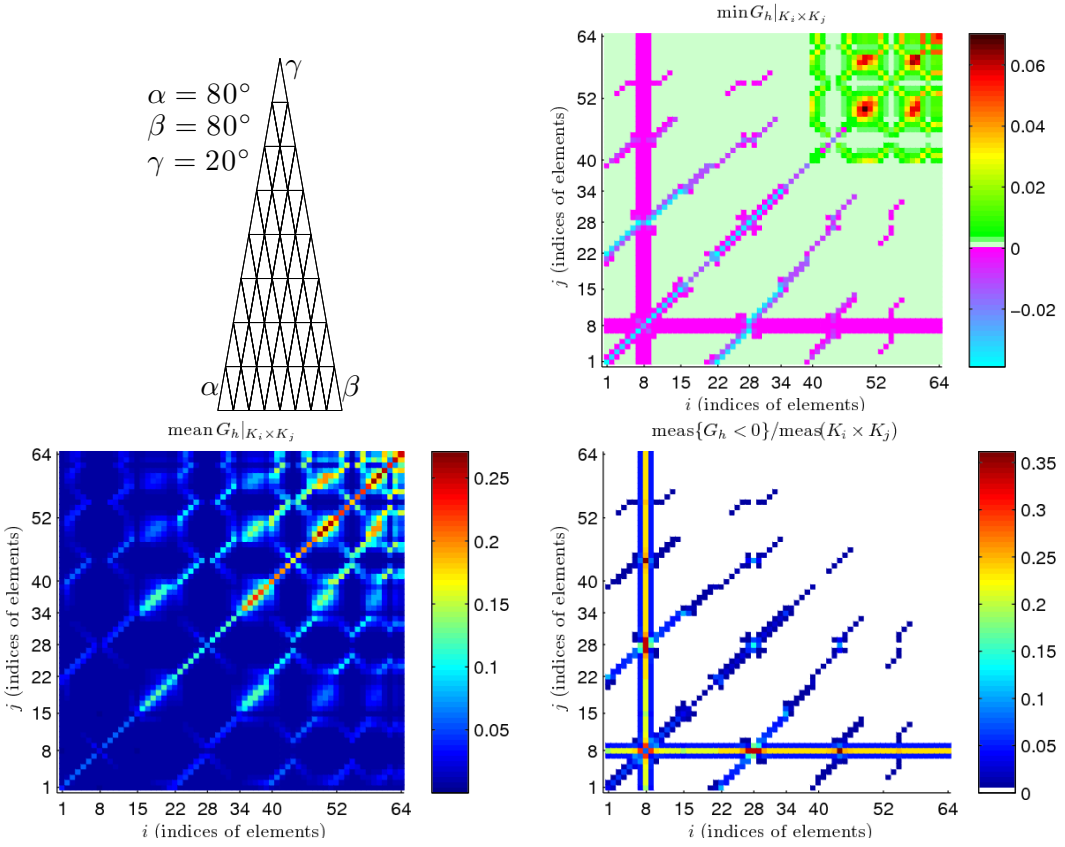


Figure 5.13: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 80°, 80°, 20°) and the triangulation. The enumeration of elements follows the same pattern as in Figures 5.8–5.10. The polynomial degree is $p = 3$.

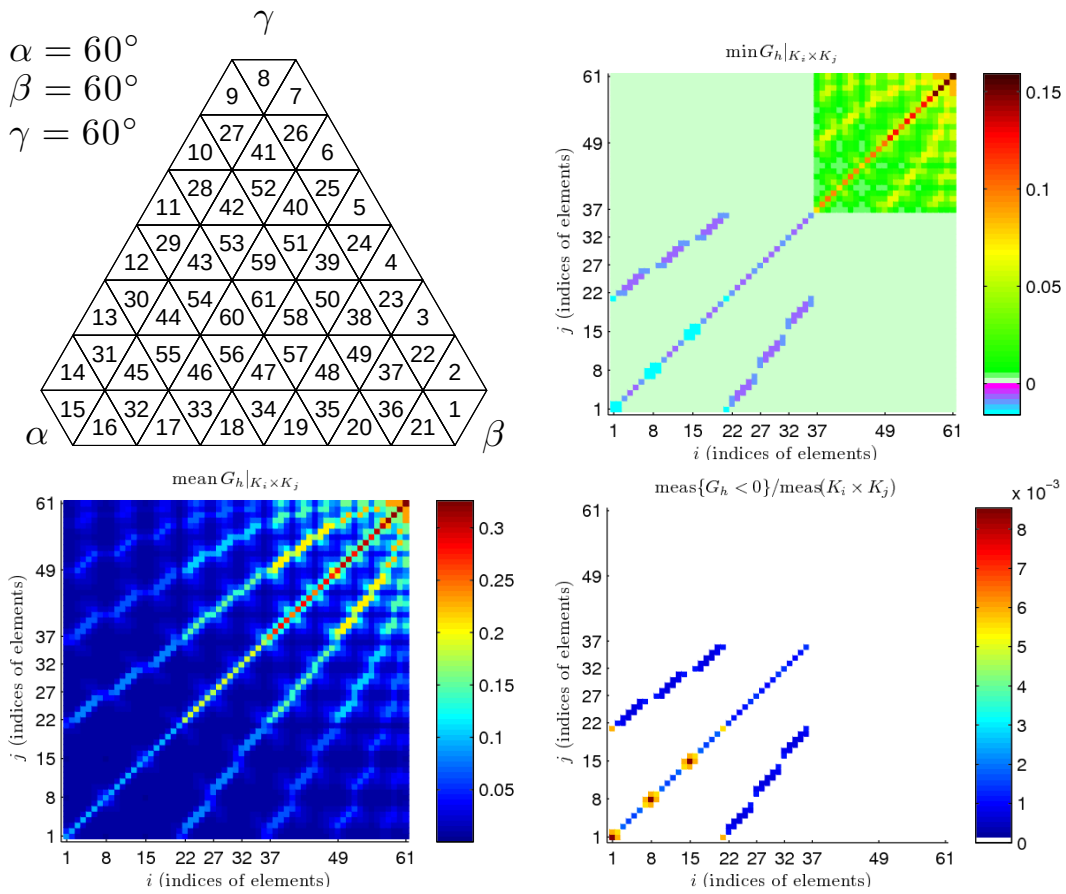


Figure 5.14: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 60°, 60°, 60° without corners) and the triangulation with the enumeration of elements. The polynomial degree is $p = 3$.

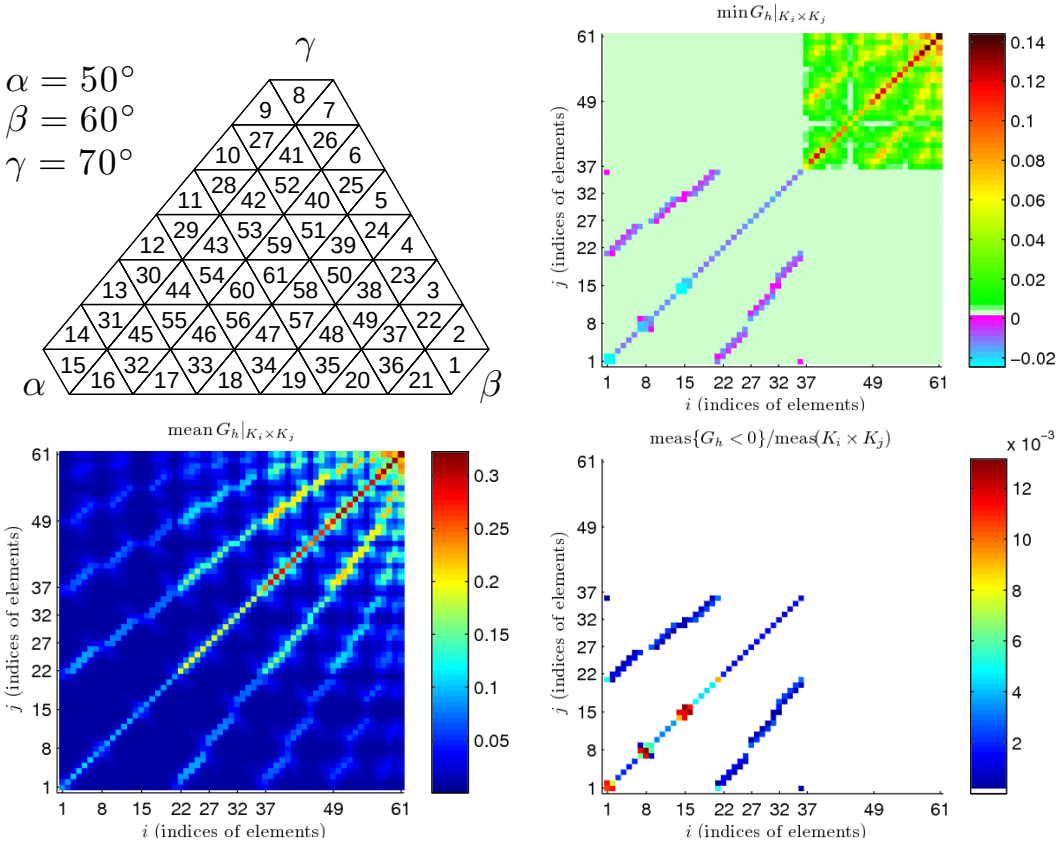


Figure 5.15: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 50°, 60°, 70° without corners) and the triangulation with the enumeration of elements. The polynomial degree is $p = 3$.

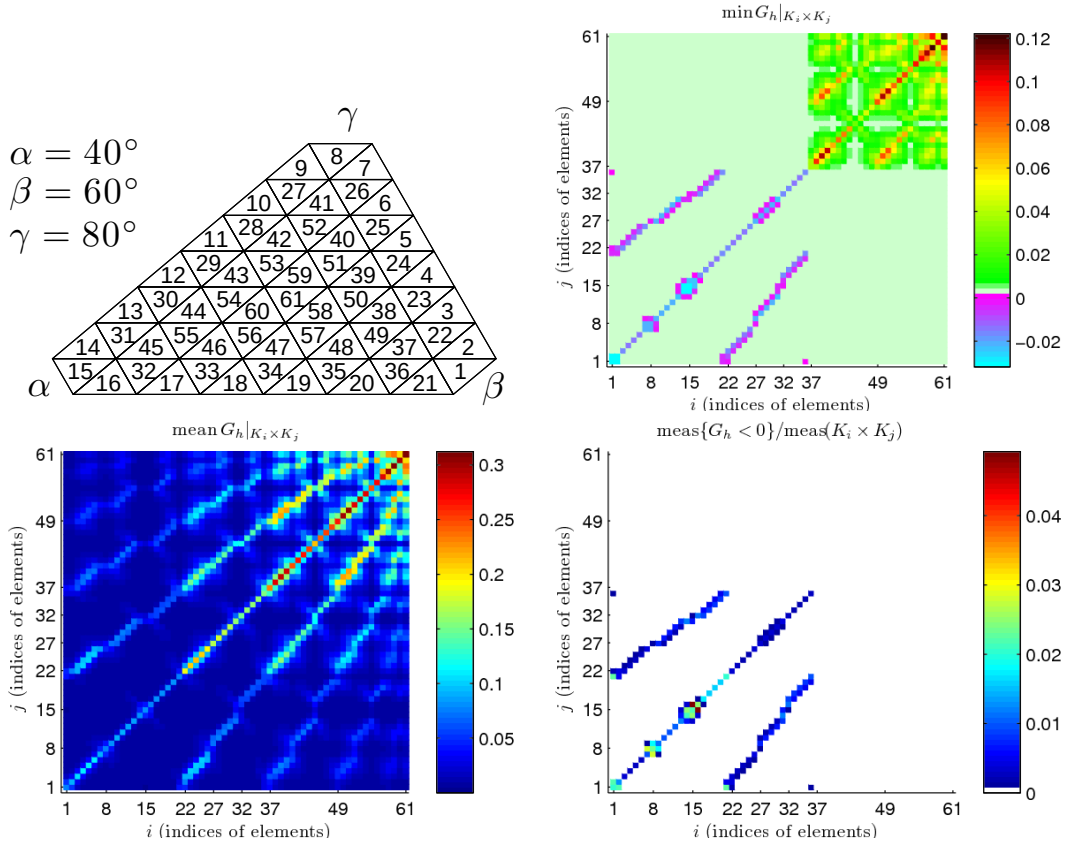


Figure 5.16: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 40°, 60°, 80° without corners) and the triangulation with the enumeration of elements. The polynomial degree is $p = 3$.

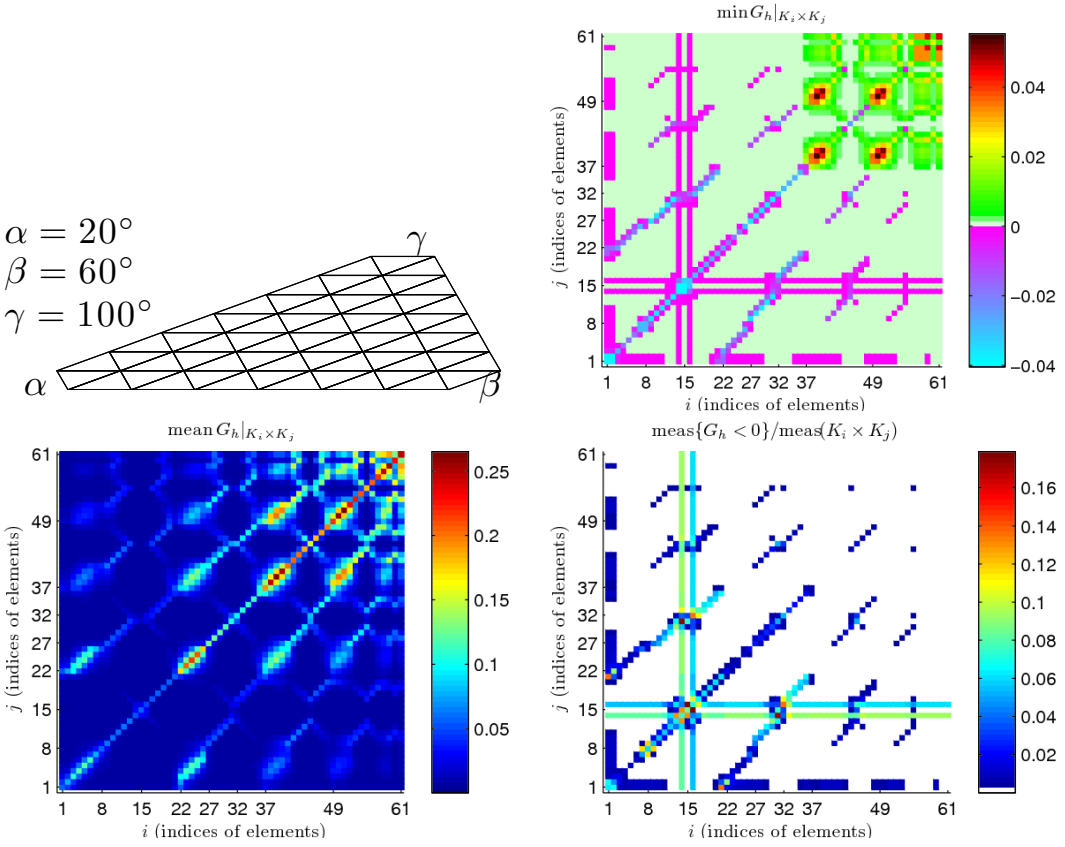


Figure 5.17: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 20°, 60°, 100° without corners) and the triangulation. The enumeration of elements follows the same pattern as in Figures 5.14–5.16. The polynomial degree is $p = 3$.

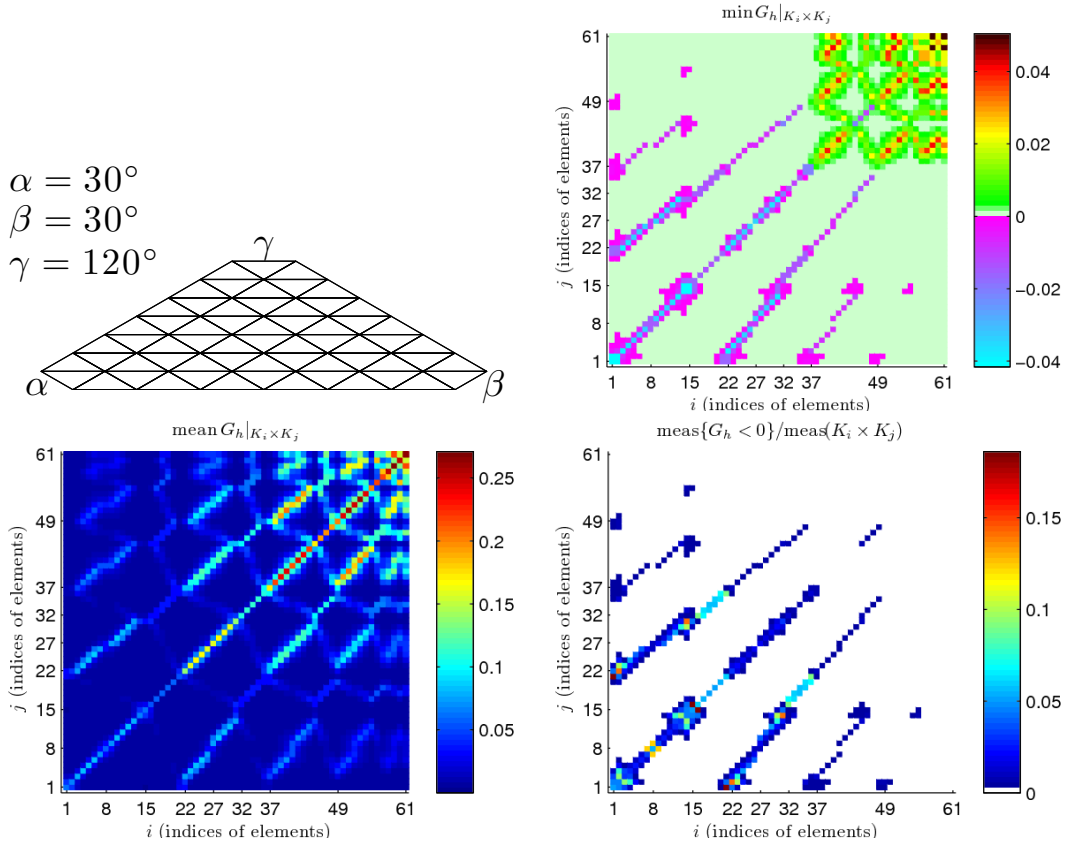


Figure 5.18: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 30°, 30°, 120° without corners) and the triangulation. The enumeration of elements follows the same pattern as in Figures 5.14–5.16. The polynomial degree is $p = 3$.

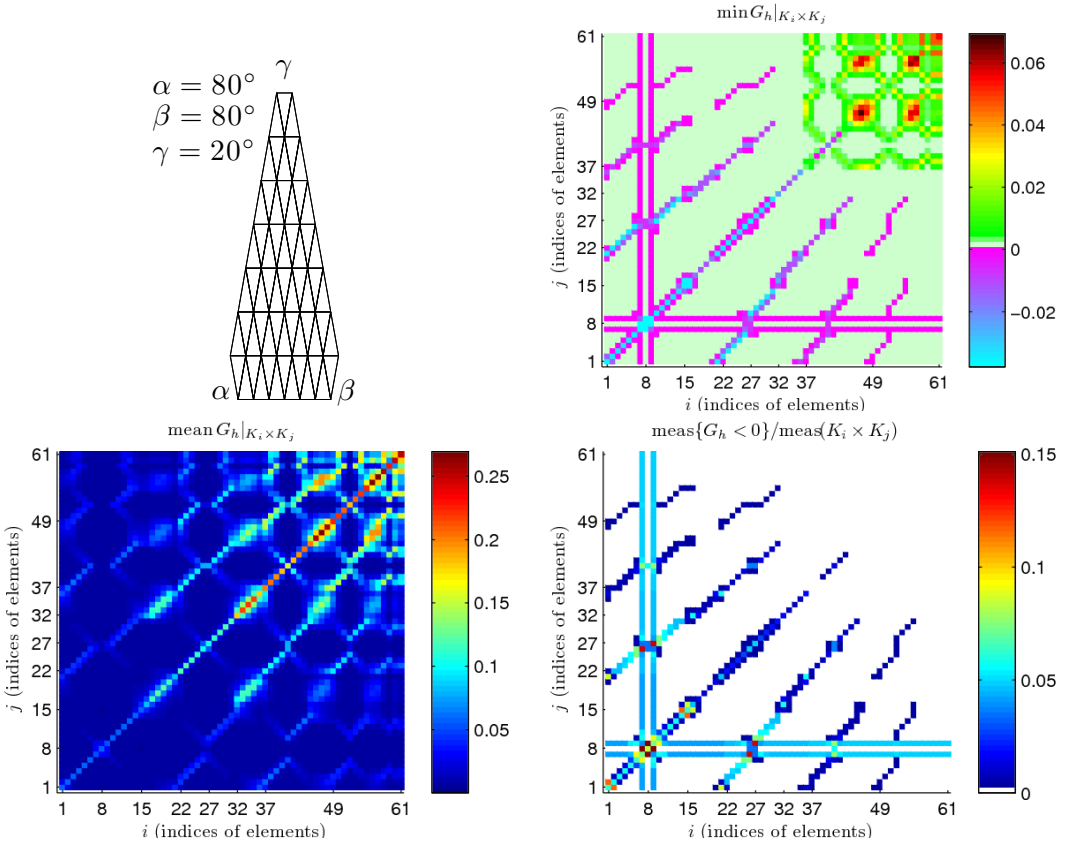


Figure 5.19: Visualization of $\min G_h|_{K_i \times K_j}$ (top-right), mean value of $G_h|_{K_i \times K_j}$ (bottom-left), and of the fraction of the area, where G_h is negative (bottom-right). The top-left panel shows the domain Ω (a triangle with angles 80°, 80°, 20° without corners) and the triangulation. The enumeration of elements follows the same pattern as in Figures 5.14–5.16. The polynomial degree is $p = 3$.

area to the area of $K_i \times K_j$ is in the order of percents. If there are small angles in the triangulation then this fraction is at most 40 %. In addition, the actual size of the minimum of G_h is relatively small with respect to the maximum of G_h.

A distinctive phenomenon in right panels of Figures 5.8–5.19 is the frequent presence of vertical and horizontal lines. For example, in Figure 5.9 there are these lines for $i = 15$ and for $j = 15$. This means that there exist negative values of G_h in $K_{15} \times K_j$ for all $j = 1, 2, \ldots, 64$ and symmetrically in $K_i \times K_{15}$ for all $i = 1, 2, \ldots, 64$. From the colors of these lines we judge that the negativity of G_h on these lines is quite tiny. Notice that the element K_{15} is in this case the corner element corresponding to the smallest angle $\alpha = 59°$.

From the performed experiments it seems that for $p = 3$, Ω being a triangle, and $\mathcal{T}_h$ being a uniform triangulation of Ω these lines always exist with the only exception of the equilateral triangle. This is in agreement with results in Figure 5.6 for $p = 3$.

In general, we found out that these lines are often caused by the corner elements corresponding to a vertex of Ω with small angles. If we remove the corner elements then these lines emerge for dramatically smaller angles only – see Figures 5.14–5.19.

Next, we can easily compare the visualizations of the DGF G_h in Figures 5.8–5.19 with the results of Experiment 2 presented in Figures 5.6 and 5.7. We clearly see the positive and negative values in regions $\Omega \times \Omega_{\mathcal{I}}$ and $\Omega^2_{\mathcal{I}}$.

The bottom-left panels of Figures 5.8–5.19 show the mean value of G_h in $K_i \times K_j$. It is of certain interest that these mean values are all nonnegative in all tested cases. The nonnegativity of the elementwise mean values of the DGF implies the property presented in the following theorem. This property can be understood as certain weak version of the discrete conservation of nonnegativity.

Theorem 5.10. *Let us consider the general elliptic problem (2.10) with homogeneous Dirichlet and Neumann boundary conditions, i.e. with $g_{\mathrm{D}} = 0$ on Γ_{D} and $g_{\mathrm{N}} = 0$ on Γ_{N}. Further, let us consider the corresponding finite element approximation (3.2) and the DGF G_h given by (3.11). Furthermore, let the mean values of the DGF G_h in $K_i \times K_j$ be nonnegative for all $K_i \in \mathcal{T}_h$ and $K_j \in \mathcal{T}_h$. If the right-hand side f is piecewise constant and nonnegative in Ω then the corresponding finite element solution $u_h \in V_h$ is nonnegative in Ω as well.*

Proof. This is a direct consequence of the representation formula (3.14) and the assumptions of the theorem. □

However, the mean values of $G_h|_{K_i \times K_j}$ are not always nonnegative. If the angles in the triangulation become sufficiently small (well below 30°) then even some of these mean values become negative and the weak variant of the discrete conservation of nonnegativity presented in Theorem 5.10 does not apply. The range

of angles yielding the nonnegative mean values of $G_h|_{K_i \times K_j}$ has been investigated in Experiment 4 below.

Experiment 4: Nonnegativity of mean values of $G_h|_{K_i \times K_j}$

Motivated by the nonnegativity of the mean values of $G_h|_{K_i \times K_j}$ obtained in Experiment 3, we performed thorough test of this property. We proceed in the same way as in Experiments 1 and 2. We test all possible combinations of angles α and β. A pair of angles α and β corresponds to a point in a plain and we color this point to black if the mean value of $G_h|_{K_i \times K_j}$ is nonnegative for all pairs $K_i, K_j \in \mathcal{T}_h$. If this mean value is negative for certain pair $K_i \times K_j$ then the color is gray. We again emphasize that the mean value is not computed exactly. It is an approximation obtained from the sample points as in Experiment 1.

The results for the triangular domain Ω (see Figure 5.3) and for $p = 1, 2, \ldots, 6$ are presented in Figure 5.20 and the results for the triangular domain without corners (see Figure 5.5) are in Figure 5.21.

The results for $p = 1$ in Figures 5.20 and 5.21 are not too surprising. They coincide with the monotony of the stiffness matrix, see results of Experiment 1 and 2 in Figures 5.4, 5.6, and 5.7.

The results for $p \geq 2$ are also similar to the previous results. The black areas in Figure 5.20 are similar to but different from the red areas in Figure 5.4, i.e. the cases of nonnegative mean values of the DGF and the cases of nonnegative vertex values of the DGF are similar but slightly different. The range of angles with the nonnegative mean value of $G_h|_{K_i \times K_j}$ is fairly large. The shape of the black areas in Figures 5.20 and 5.21 might suggest certain minimal angle condition for the nonnegativity of the mean values.

Experiment 5: Dependence of the DGF on the polynomial degree

In this experiment we investigate how the negative values of the DGF behave with respect to the polynomial degree p. We concentrate on two characteristics – on the global minimum of G_h over $\overline{\Omega}^2$ (briefly $\min G_h$) and on the measure of the domain, where the DGF G_h is negative. We express this measure relatively with respect to the measure of entire Ω^2, i.e. we investigate the ratio $r_{\text{neg}} = \text{meas}\{(\boldsymbol{x}, \boldsymbol{y}) \in \Omega^2 : G_h(\boldsymbol{x}, \boldsymbol{y}) < 0\} / \text{meas}\, \Omega^2$.

In Figures 5.22 and 5.23 we consider fixed domain Ω, fixed triangulation $\mathcal{T}_h$ and we present the dependence of $\min G_h$ and of r_{neg} on p. Figure 5.22 corresponds to the triangular domain Ω (see Figure 5.3), while Figure 5.23 shows the results for the triangular domain without corners (see Figure 5.5). Each graph in these figures corresponds to a specific choice of angles α and β – see the legends.

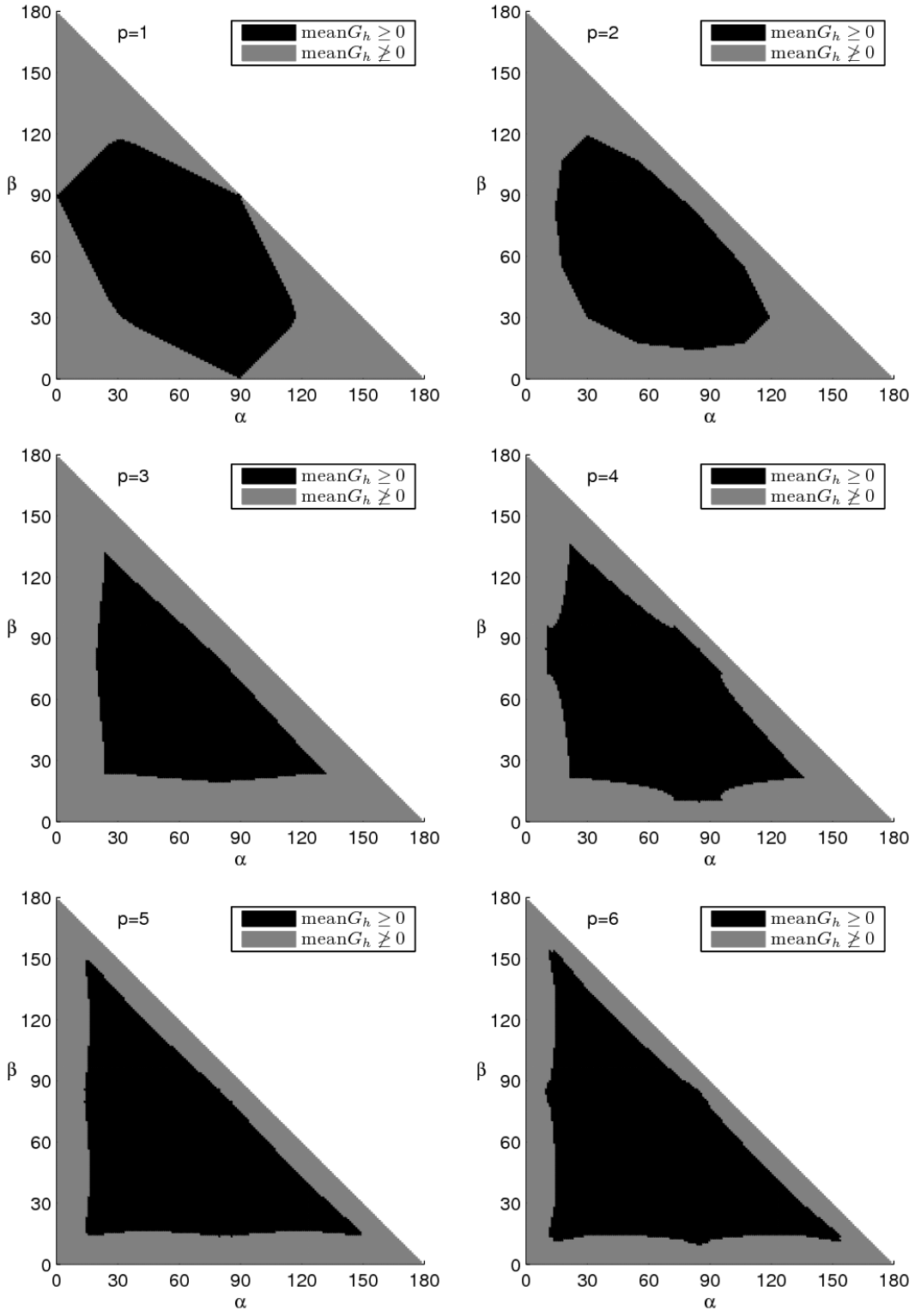


Figure 5.20: Nonnegativity of the mean values of $G_h|_{K_i \times K_j}$. The domain Ω is a triangle – see Figure 5.3.

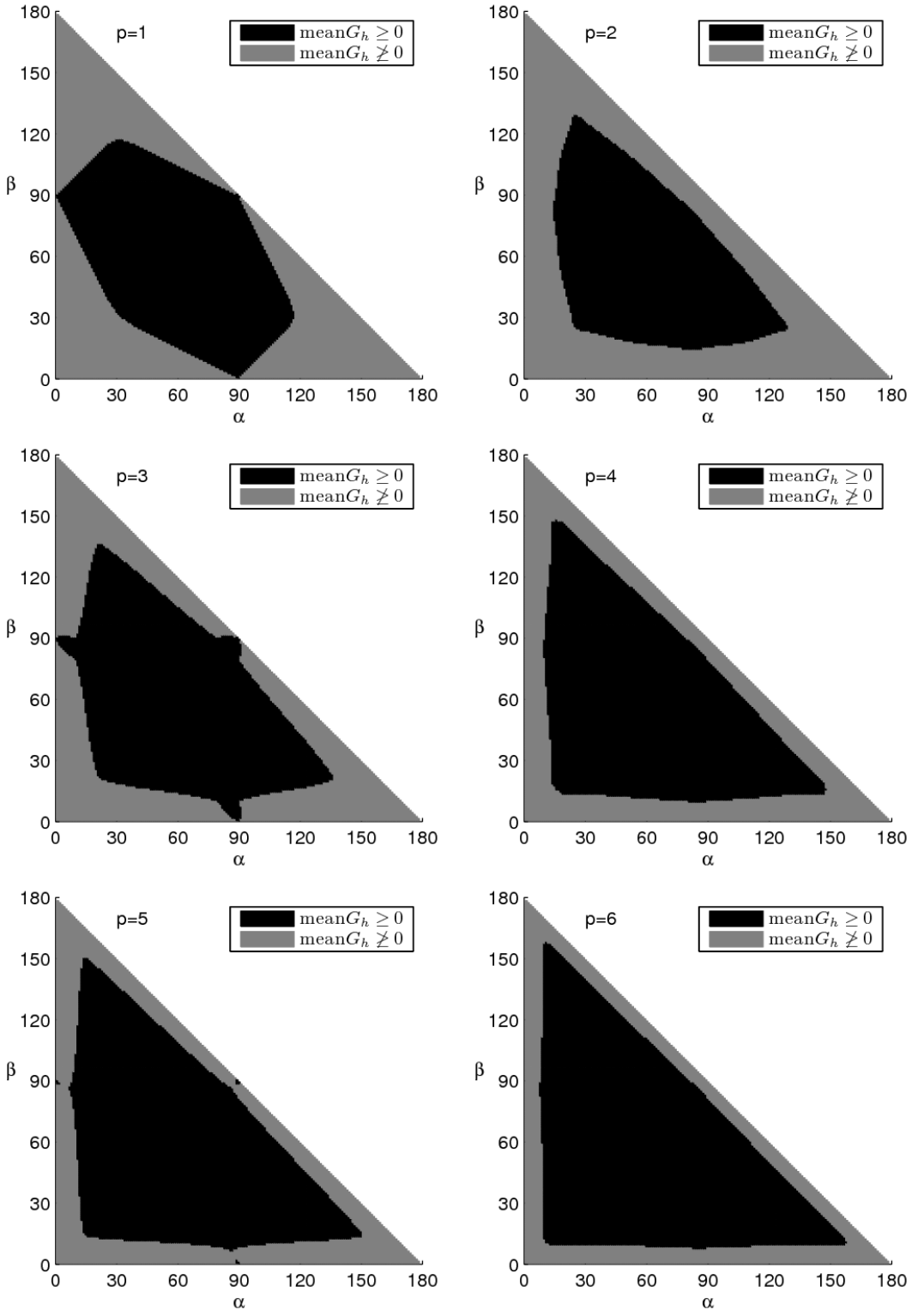


Figure 5.21: Nonnegativity of the mean values of $G_h|_{K_i \times K_j}$. The domain Ω is a triangle without corners – see Figure 5.5.

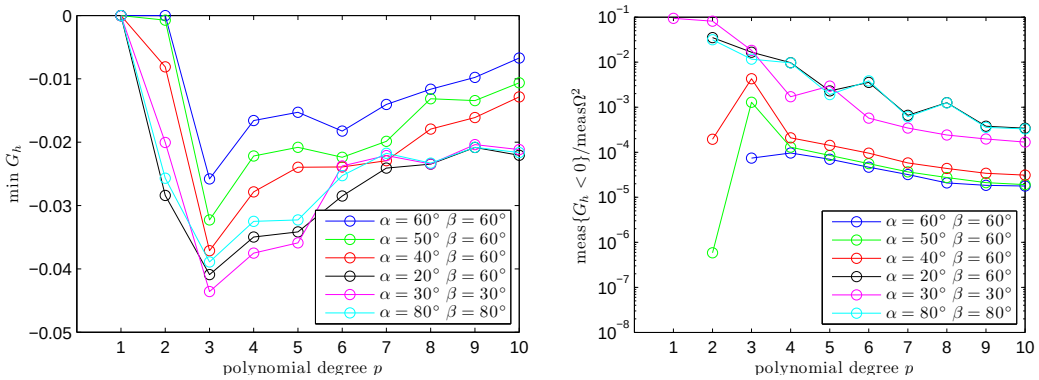


Figure 5.22: The dependence of the minimum of G_h in $\overline{\Omega}^2$ (left) and of the ratio $r_{\mathrm{neg}} = \mathrm{meas}\{G_h < 0\}/\,\mathrm{meas}\,\Omega^2$ (right) on the polynomial degree p. Results for the triangular domain (see Figure 5.3) and for various angles.

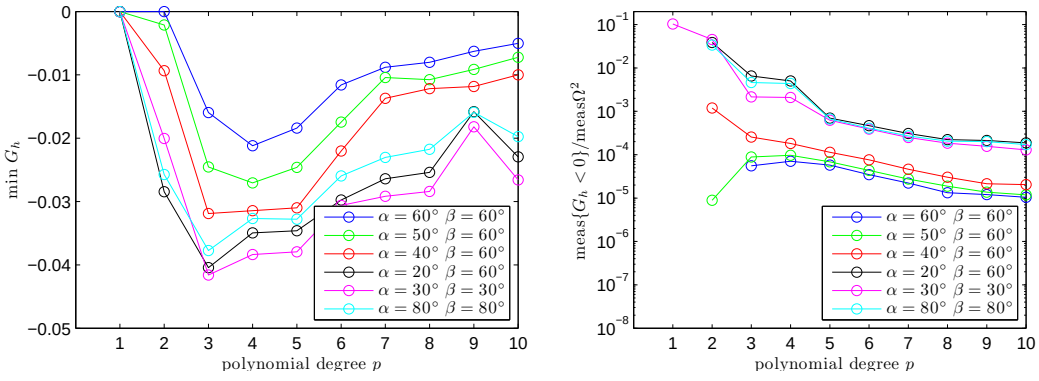


Figure 5.23: The dependence of the minimum of G_h in $\overline{\Omega}^2$ (left) and of the ratio $r_{\mathrm{neg}} = \mathrm{meas}\{G_h < 0\}/\,\mathrm{meas}\,\Omega^2$ (right) on the polynomial degree p. Results for the triangular domain without corners (see Figure 5.5) and for various angles.

The first observation from the results in Figures 5.22 and 5.23 is that the behavior of the $\min G_h$ and of the r_{neg} has the same character for both tested domains Ω. The actual values of these characteristics are also more-less the same. Further, we see that for sufficiently high p the minimum of G_h increases and the measure of the area, where G_h is negative, decreases. However, this increase and decrease are not monotone. The decrease of r_{neg} is faster than the increase of $\min G_h$ – notice the semi-logarithmic scale in the right panels.

Thus, it seems that most often the minimum of G_h is the deepest for $p = 3$. However, in certain cases the most negative values are obtained for $p = 4$ – see for example the case $\alpha = \beta = 60°$ in Figure 5.23 (left). For polynomial degrees $p = 3$ and $p = 4$ we also observe relatively large domains with negative values of G_h. However, if there are negative values of G_h for $p = 1$ or $p = 2$ (often tiny negative) then the domain with negative values is even larger.

In general, we can conclude, that the DGF looses its nonnegativity for polynomial degrees $p \geq 2$ in the vast majority of cases. However, the actual size of the negative values is relatively small (in the orders of percents of the positive maximum). The measure of the domain, where the DGF is negative, is often very small too. The ratio r_{neg} ranges from 10^{-5} to 10^{-1} in our experiments.

Thus, the DMP is not satisfied for absolute majority of cases for higher-order finite element meshes, but the right-hand side function f leading to the violation of the DMP have to look "obscure". The numerical experiments show that a nonnegative function f yielding a finite element solution u_h being negative at some point has to possess great values in relatively small region (often close to the boundary) and relatively small values in the rest of the domain Ω. These requirements on f are the stronger the higher polynomial degrees are considered.

5.5 A weaker type of the discrete maximum principle

The numerical experiments presented in the previous section clearly indicate that the higher-order finite element methods satisfy the DMP in exceptional cases only. Thus, it is natural to think about weaker concepts. One option how to weaken the DMP is presented in [A8]. This paper is attached to this thesis in Appendix I.

The idea is to replace the requirement of nonnegativity of f a.e. in Ω by the nonnegativity of the L^2 projection of f onto the corresponding piecewise polynomial space. The motivation is straightforward – the finite element solution $u_h \in V_h$ corresponding to the original right-hand size f is the same as the finite

element solution corresponding to the L^2 projection f_h, because

$$\int_\Omega f v_h \,\mathrm{d}\boldsymbol{x} = \int_\Omega f_h v_h \,\mathrm{d}\boldsymbol{x} \quad \forall v_h \in V_h.$$

In what follows we first briefly review the result [A8] attached in Appendix I and then we present its slight but practical generalization.

For simplicity we consider 1D Poisson problem with homogeneous Dirichlet boundary conditions:

$$-u'' = f \quad \text{in } \Omega = (a^\partial, b^\partial), \qquad u(a^\partial) = u(b^\partial) = 0.$$

We consider a finite element partition $\mathcal{T}_h$ of Ω and the standard piecewise polynomial finite element space

$$V_h = \{v_h \in H_0^1(\Omega) : v_h|_K \in \mathbb{P}^{p_K}(K),\ K \in \mathcal{T}_h\},$$

(see Section 5.1 for details) and we define the finite element solution $u_h \in V_h$ such that

$$\int_\Omega u_h' v_h' \,\mathrm{d}x = \int_\Omega f v_h \,\mathrm{d}x \quad \forall v_h \in V_h. \tag{5.30}$$

As we mentioned above, the idea is to introduce the $L^2(\Omega)$ projection of f. It is defined as the unique $f_h \in V_h$ such that

$$\int_\Omega (f - f_h) v_h \,\mathrm{d}x = 0 \quad \forall v_h \in V_h.$$

The discretization (5.30) then satisfies the *weak DMP* if

$$f_h \le 0 \text{ in } \Omega \quad \Rightarrow \quad \max_{\overline{\Omega}} u_h \le \max_{\partial\Omega} u_h = 0.$$

This is clearly equivalent to the *weak conservation of nonnegativity*:

$$f_h \ge 0 \text{ in } \Omega \quad \Rightarrow \quad u_h \ge 0 \text{ in } \Omega,$$

cf. Theorem 3.1.

The result in [A8] states that discretization (5.30) satisfies the weak DMP provided a certain special quadrature rule exists. This result is in principle general and it can be used in higher-dimension and for problems other than the Poisson problem. However, the construction of the special quadrature rule is demanding. In [A8] we present a numerical construction of these quadrature rules in 1D for polynomial degrees up to $p = 10$. Thus, for 1D problem (5.30) we have verified that the finite elements of order up to $p = 10$ satisfy the weak DMP on arbitrary meshes.

A disadvantage of the concept of this weak DMP is its globality. The L^2 projection f_h is considered in V_h and its construction requires to solve a global problem with the number of degrees of freedom equal to the dimension of V_h.

Possible remedy is to consider L^2 projections local to the elements $K \in \mathcal{T}_h$. Let the local L^2 projection $\bar{f}_K \in \mathbb{P}^{p_K}(K)$ be defined as

$$\int_K (f - \bar{f}_K)\varphi \,\mathrm{d}x = 0 \quad \forall \varphi \in \mathbb{P}^{p_K}(K).$$

Projections $\bar{f}_K$ defined in $K \in \mathcal{T}_h$ can be unified into a single function $\bar{f}$ defined in Ω such that $\bar{f}(x) = f_K(x)$ for all $x \in K$ and all $K \in \mathcal{T}_h$. Naturally, the function $\bar{f}$ is piecewise polynomial but in general discontinuous over the element interfaces.

This setting enables us to define another concept of the weak DMP. The discretization (5.30) satisfies the *weak DMP* if

$$\bar{f} \le 0 \text{ in } \Omega \quad \Rightarrow \quad \max_{\overline{\Omega}} u_h \le \max_{\partial\Omega} u_h = 0. \tag{5.31}$$

This is again equivalent to the corresponding *weak conservation of nonnegativity*:

$$\bar{f} \ge 0 \text{ in } \Omega \quad \Rightarrow \quad u_h \ge 0 \text{ in } \Omega.$$

The analysis of the proof presented in [A8] reveals that also the weak DMP (5.31) is satisfied for finite elements of orders up to $p = 10$ on arbitrary meshes. The proof in [A8] requires essentially no changes in order to be valid in this new setting. The quadrature rules constructed there can be taken exactly the same.

The advantage of this new concept is clear. The projection $\bar{f}$ can be constructed much faster than f_h by solving a small system on each element $K \in \mathcal{T}_h$. Thus, in contrast to the original setting the verification of the assumptions in the new setting is much more practical.

CHAPTER

SIX

Conclusions

This thesis is a result of an attempt to present more-less complete survey of the DMP results for the linear second-order elliptic partial differential equations discretized by the lowest-order and the higher-order finite element methods. Up to the author's knowledge another survey of this extend and detailness is not available.

On the other hand, there are other important areas in the field of the DMP, which attract a lot of attention. One of the biggest is the area of the DMP for parabolic problems. Parabolic problems provide the same variety of possible approaches and results as the elliptic problems. A similar survey targetted to the parabolic case is definitely possible and would be useful. The doctoral thesis [27] addresses this issue from the perspective of the finite difference method.

The unified and systematic treatment of the topic has enabled not only to present the issue of the DMP for elliptic problems discretized by the finite element method in a consistent and hopefully well understandable way but also to formulate and prove new and original DMP results. This concerns both the lowest- and the higher-order case.

The DMP for the lowest-order finite elements is already quite well understood. It is studied for several decades. Nevertheless, we were still able to formulate and prove new results especially concerning the variable equation coefficients and treating new types of finite elements. Various examples showing the validity and the failure of the DMP are original as well.

The higher-order case is understood far less. This thesis presents the author's original contributions to this field. There are several positive results about simple one-dimensional problems. For two- and higher-dimensional case the standard version of the DMP seems to be valid in exceptional situations only. The thesis presents many numerical experiments to support this hypothesis. A natural remedy is to develop weaker concepts of maximum principles for higher-order ap-

proximations. One attempt in this direction is presented as well, see Section 5.5.

A generalization of the described concept of the weak DMP to the two- and higher-dimensional cases would be an interesting topic for further research. Another topic deserving deeper analysis was mentioned in Section 3.2. The point is to develop a methodology how to construct the finite element meshes based on the given data of the problem (like the right-hand side f and the boundary data g_{D} and g_{N}) such that the resulting discretization satisfies the DMP.

APPENDIX

A

Integral of powers of barycentric coordinates

Lemma A.1. *Let $K_d \subset \mathbb{R}^d$, $d \geq 1$, be a d-dimensional simplex with barycentric coordinates $\lambda_1, \lambda_2, \ldots, \lambda_{d+1}$. Let Γ denote the gamma function and $|K_d|$ the d-dimensional Lebesgue measure of K_d. If $s_1, s_2, \ldots, s_{d+1}$ are complex numbers with real parts greater than -1, then*

$$\int_{K_d} \prod_{i=1}^{d+1} \lambda_i^{s_i}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \frac{d! \prod_{i=1}^{d+1} \Gamma(s_i+1)}{\Gamma\left(1+d+\sum_{i=1}^{d+1} s_i\right)} |K_d|. \tag{A.1}$$

Proof. We prove the statement by the mathematical induction. First, the validity of (A.1) for $d = 1$ can be verified easily. If $K_1 = [x_L, x_R] \subset \mathbb{R}$ then

$$\int_{K_1} \lambda_1^{s_1}(x)\lambda_2^{s_2}(x) \,\mathrm{d}x = \int_{x_L}^{x_R} \lambda_1^{s_1}(x)\big(1-\lambda_1(x)\big)^{s_2} \,\mathrm{d}x = (x_R - x_L) \int_0^1 t^{s_1}(1-t)^{s_2} \,\mathrm{d}t,$$

where we used the transformation $t = \lambda_1(x)$. This integral is in the form of the beta function

$$B(r,s) = \int_0^1 t^{r-1}(1-t)^{s-1} \,\mathrm{d}t = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} \tag{A.2}$$

and thus

$$\int_{K_1} \lambda_1^{s_1}(x)\lambda_2^{s_2}(x) \,\mathrm{d}x = \frac{\Gamma(s_1+1)\Gamma(s_2+1)}{\Gamma(s_1+s_2+2)} |K|,$$

Now, we assume the validity of the equality (A.1) for dimensions up to $d-1$

and we will prove it for d. To this end, we use the following identity

$$\int_{K_d} \prod_{i=1}^{d+1} \lambda_i^{s_i}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{K_d} \left(\prod_{i=1}^{d} \lambda_i^{s_i}(\boldsymbol{x})\right)\left(1-\sum_{i=1}^{d}\lambda_i(\boldsymbol{x})\right)^{s_{d+1}} \mathrm{d}\boldsymbol{x}$$

$$= d!|K_d| \int_0^1 \int_0^{1-\xi_1} \cdots \int_0^{1-\sum_{i=1}^{d-1}\xi_i} \left(\prod_{i=1}^{d}\xi_i^{s_i}\right)\left(1-\sum_{i=1}^{d}\xi_i\right)^{s_{d+1}} \mathrm{d}\xi_d \cdots \mathrm{d}\xi_2\,\mathrm{d}\xi_1, \quad \text{(A.3)}$$

where we employed the substitutions $\xi_j = \lambda_j(\boldsymbol{x})$, $j = 1, 2, \ldots, d$. Further, using the substitution $(1 - \sum_{i=1}^{d-1}\xi_i)t = \xi_d$ in the inner integral, we obtain that the above integral equals to

$$d!|K_d| \int_0^1 \int_0^{1-\xi_1} \cdots \int_0^1 \left(\prod_{i=1}^{d-1}\xi_i^{s_i}\right)\left(1-\sum_{i=1}^{d-1}\xi_i\right)^{s_d+s_{d+1}+1} t^{s_d}(1-t)^{s_{d+1}}\,\mathrm{d}t \cdots \mathrm{d}\xi_2\,\mathrm{d}\xi_1.$$

Rearanging this expression and using (A.3) for K_{d-1}, we find that

$$\int_{K_d} \prod_{i=1}^{d+1} \lambda_i^{s_i}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

$$= \frac{d!|K_d|}{(d-1)!|K_{d-1}|} \int_{K_{d-1}} \left(\prod_{i=1}^{d-1}\lambda_i^{s_i}(\boldsymbol{x})\right)\lambda_d^{s_d+s_{d+1}+1}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \int_0^1 t^{s_d}(1-t)^{s_{d+1}}\,\mathrm{d}t.$$

Hence, by (A.2) and (A.1) with K_{d-1} we obtain the claimed result. □

Since $\Gamma(m+1) = m!$ for a nonnegative integer m, we can rewrite (A.1) as

$$\int_{K_d} \prod_{i=1}^{d+1} \lambda_i^{s_i}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \frac{d!\prod_{i=1}^{d+1} s_i!}{\left(d+\sum_{i=1}^{d+1} s_i\right)!}|K_d|$$

provided $s_1, s_2, \ldots, s_{d+1}$ are nonnegative integers. We note that our interest in Lemma A.1 and its proof was motivated by paper [24], where formula (A.1) is proved for d up to 3 only.

APPENDIX

B

Discrete maximum principle for FE solutions of the diffusion-reaction problem on prismatic meshes

Below we attach a copy of the paper

[A1] 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), 275–287.

Journal of Computational and Applied Mathematics 226 (2009) 275–287

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Discrete maximum principle for FE solutions of the diffusion-reaction problem on prismatic meshes

Antti Hannukainen [a], Sergey Korotov [a], Tomáš Vejchodský [b,*]

[a] Institute of Mathematics, Helsinki University of Technology, P.O. Box 1100, FIN-02015 Espoo, Finland
[b] Institute of Mathematics, Czech Academy of Sciences, Žitná 25, CZ-115 67 Prague 1, Czech Republic

ARTICLE INFO

MSC:
65N30
65N50
35B50
35J25

Keywords:
Diffusion-reaction problem
Maximum principle
Prismatic finite elements
Discrete maximum principle

ABSTRACT

In this paper we analyse the discrete maximum principle (DMP) for a stationary diffusion-reaction problem solved by means of prismatic finite elements. We derive geometric conditions on the shape parameters of the prismatic partitions which guarantee validity of the DMP. The presented numerical tests show the sharpness of the obtained conditions.

1. Introduction

Mathematical models consisting of elliptic and parabolic partial differential equations with various boundary and initial conditions are useful tools in modeling and numerical simulations of various real-life problems (see e.g. [7,11]). Usually, the exact (classical) solutions of these models exhibit certain qualitative properties such as the maximum-minimum principle (or, as a particular case, the nonnegativity preservation) [24], the sign-stability (often called as a preservation of number of peaks) [14,15], the maximum norm contractivity, etc. For more details in the subject see recent reviews [10,18].

Among these, the maximum principle is the basic characteristic usually associated with the second order elliptic (and parabolic) boundary value problems [17,24,25]. It can be mathematically described as an *a priori estimate* of the magnitude of the solution (unknown in the whole domain) by the magnitude of the given (i.e. known), or easily computable, data. The maximum principle is not only a mathematical feature of the model but it also adequately describes the real behavior of physical systems.

It is quite natural to require a suitable imitation of this property from the computed approximations. This is the reason why the construction and validity of the corresponding discrete analogues (the so-called discrete maximum principles, or DMPs in short) have drawn much attention. To the authors' knowledge, papers [27] by R. Varga in 1966 and [13] by H. Fujii in 1973 were probably the very first works aimed at the construction of a reasonable DMP for elliptic and parabolic problems, respectively. These original papers as well as the presented work use special properties of the finite difference and finite element matrices to analyse the DMPs.

Later on, other types of the DMPs were formulated and proved in a number of papers, see e.g. [6,8,17,18,21,25,28,29]. They discuss various numerical methods for different problems and study the validity of the DMPs. Most of the attention was paid to the finite difference and finite element approximations of elliptic and parabolic problems and to various geometric conditions on the shape of the classical simplicial and block finite element partitions that provide the DMPs. Particularly

* Corresponding author.
E-mail addresses: antti.hannukainen@hut.fi (A. Hannukainen), sergey.korotov@hut.fi (S. Korotov), vejchod@math.cas.cz (T. Vejchodský).

0377-0427/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2008.08.029

challenging is the analysis of the DMPs for the less standard but more promising and economical higher order finite elements, see recent results [22,28]. However, the validity of the DMPs on prismatic meshes has not been considered so far in spite of the fact that the prismatic partitions can often be more natural and practically convenient compared to the standard tetrahedral or block partitions, especially for cylindrical 3D domains.

The paper is organized as follows. Section 2 describes the 3D diffusion-reaction model problem and Section 3 presents its finite element discretization by the lowest order prismatic elements with six degrees of freedom. The main theoretical result about the DMP is contained in Section 4. Section 5 provides practical geometric conditions for prismatic partitions to guarantee the validity of the DMP. The sharpness of the obtained geometric conditions is verified by numerical tests in Section 6. Finally, Section 7 points out possible generalizations and several open problems.

2. Model problem

Throughout the paper we shall use the standard Sobolev space notation (see e.g. [7,11]). We consider the following reaction-diffusion boundary value problem

$$-\Delta u + cu = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega, \tag{1}$$

where $\Omega \subset \mathbb{R}^3$ is a bounded domain with Lipschitz boundary $\partial\Omega$ and c is a nonnegative reaction coefficient. To define the weak solution of (1), we assume $f \in L^2(\Omega)$, $c \in L^\infty(\Omega)$, and

$$0 \le c \le \|c\|_{\infty,\Omega}, \tag{2}$$

where $\|c\|_{\infty,\Omega} = \|c\|_{L^\infty(\Omega)}$ stands for the L^∞-norm of the reaction coefficient c over the domain Ω.

The weak formulation of problem (1) reads: Find a function $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x + \int_\Omega cuv \,\mathrm{d}x = \int_\Omega fv \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega). \tag{3}$$

Under the above conditions the weak solution u exists and is unique.

The following theorem shows the continuous maximum principle (CMP) for problem (1), see [24] and also [17,18] for a more general case of nonlinear problems with mixed boundary conditions. In what follows, the equalities and inequalities between functions from Lebesgue spaces should be understood up to a set of zero measure, as usual.

Theorem 1. *Let u be a solution to (1). If $f \le 0$ and $u \in C(\overline{\Omega})$ then* $\max_{\overline{\Omega}} u = 0$.

A natural discrete analogue to the above implication is known as the discrete maximum principle (DMP). In what follows, we formulate the DMP precisely and we derive geometric conditions on the shape of prismatic finite elements guaranteeing its validity *a priori*.

3. FE discretization on prismatic meshes

In general, we could consider any domain Ω which can be partitioned (face-to-face) into triangular prisms. For instance, a union of cylindrical domains is acceptable. However, for the sake of simplicity, we assume $\Omega = \mathcal{G} \times \mathcal{I}$ to be a cylindrical domain, where $\mathcal{G} \subset \mathbb{R}^2$ is a polygon possibly with polygonal holes, $\mathcal{I} = (0, z_0)$, and z_0 is a positive number. We shall consider a face-to-face partition $\mathcal{T}_{h,\tau} = \mathcal{T}_h^{\mathcal{G}} \times \mathcal{T}_\tau^{\mathcal{I}}$ of $\overline{\Omega}$ into prisms (and call it *prismatic mesh* or *prismatic partition* of Ω), where $\mathcal{T}_h^{\mathcal{G}}$ is a triangulation of $\mathcal{G}$ and $\mathcal{T}_\tau^{\mathcal{I}}$ is a partition of $\mathcal{I}$ into segments (not necessarily with the same lengths). Prismatic elements of $\mathcal{T}_{h,\tau}$ will be denoted from now on with the symbol P possibly with certain indices. The elements of the triangulation $\mathcal{T}_h^{\mathcal{G}}$ (being, actually, the bases of the prismatic elements) will be denoted by T and the elements of $\mathcal{T}_\tau^{\mathcal{I}}$ will be denoted by I possibly with indices. Let B_i, $i = 1, \dots, N + N^\partial$, be the vertices of $\mathcal{T}_{h,\tau}$, where $B_1, \dots, B_N$ are the interior nodes and $B_{N+1}, \dots, B_{N+N^\partial}$ belong to the boundary $\partial\Omega$.

Let $V_{h,\tau} \subset H_0^1(\Omega)$ be the finite element space associated to $\mathcal{T}_{h,\tau}$ and defined as follows:

$$V_{h,\tau} = \Bigg\{ \varphi \in H_0^1(\Omega) :\ \varphi(x,y,z)|_P = \sum_{i=1}^{3}\sum_{j=1}^{2} b_{i,j}\lambda_i(x,y)\ell_j(z), \\ \text{where } P = T \times I, P \in \mathcal{T}_{h,\tau},\ T \in \mathcal{T}_h^{\mathcal{G}},\ I \in \mathcal{T}_\tau^{\mathcal{I}},\ b_{i,j} \in \mathbb{R},\ \lambda_i \in \mathbb{P}^1(T),\ \ell_j \in \mathbb{P}^1(I) \Bigg\}, \tag{4}$$

where $\mathbb{P}^1(T)$ and $\mathbb{P}^1(I)$ stand for the spaces of linear functions defined in the triangle T and in the interval I, respectively. Further, let $\phi_1, \dots, \phi_N$ denote the standard finite element basis functions of $V_{h,\tau}$ satisfying $\phi_i(B_j) = \delta_{ij}$, $i = 1, 2, \dots, N$, $j = 1, 2, \dots, N + N^\partial$, where δ_{ij} is the Kronecker symbol.

The finite element discretization based on the weak formulation (3) reads: Find a function $u_{h,\tau} \in V_{h,\tau}$ such that

$$\int_\Omega \nabla u_{h,\tau} \cdot \nabla v_{h,\tau} \,\mathrm{d}x + \int_\Omega cu_{h,\tau}v_{h,\tau} \,\mathrm{d}x = \int_\Omega fv_{h,\tau} \,\mathrm{d}x \quad \forall v_{h,\tau} \in V_{h,\tau}. \tag{5}$$

4. Discrete maximum principle

The discrete problem introduced above should, ideally, satisfy the following natural property (see [8,17,18,21,28]):

$$f \le 0 \quad \Longrightarrow \quad \max_{\overline{\Omega}} u_{h,\tau} = 0. \tag{6}$$

This implication, however, can lead to different interpretations. Therefore, we provide the following precise formulation of the DMP.

Definition 1. Let $\mathcal{T}_{h,\tau}$ be a partition of Ω and let $V_{h,\tau}$ given by (4) be the finite element space based on $\mathcal{T}_{h,\tau}$. We say that approximate problem (5) satisfies the *discrete maximum principle* (DMP) if

$$\max_{\overline{\Omega}} u_{h,\tau} = 0 \quad \text{for all } f \le 0. \tag{7}$$

Notice that this definition leads to a task to characterize a suitable class of meshes that guarantee (7). This is done in Theorem 2 below, where we present sufficient conditions for prismatic partitions guaranteeing (7).

Remark 1. Another possibility how to handle the DMP is to fix the right-hand side $f \le 0$ and construct a suitable partition $\mathcal{T}_{h,\tau}$ (according to this f) such that $\max_{\overline{\Omega}} u_{h,\tau} = 0$. However, this possibility is a completely different issue from the investigation of the DMP according to Definition 1 and it will not be treated here.

Remark 2. As all the basis functions are nonnegative, it is obvious that the FE approximation satisfies $u_{h,\tau} \le 0$ everywhere in Ω if and only if $u_{h,\tau}$ has nonpositive values at all nodal points B_i, $i = 1, \dots, N + N^{\partial}$.

Letting $u_{h,\tau} = \sum_{i=1}^{N} y_i \phi_i$, we come to the system of N linear equations

$$\mathbf{A}\mathbf{y} = \mathbf{F}, \tag{8}$$

where $\mathbf{A} = (a_{ij})_{i,j=1}^{N}$ is called the *FE matrix* (to distinguish it form the stiffness and mass matrices), the vector of unknowns $\mathbf{y} = (y_1, \dots, y_N)^{\top}$ consists of the values of $u_{h,\tau}$ at the interior nodes, and the vector $\mathbf{F} = (F_1, \dots, F_N)^{\top}$ is known as the *load vector*. The entries of the matrix $\mathbf{A}$ and of the vector $\mathbf{F}$ associated to problem (1) are

$$a_{ij} = \int_{\Omega} \nabla\phi_i \cdot \nabla\phi_j \, \mathrm{d}x + \int_{\Omega} c\phi_i\phi_j \, \mathrm{d}x \quad \text{and} \quad F_i = \int_{\Omega} f\phi_i \, \mathrm{d}x, \quad i, j = 1, \dots, N.$$

Various geometric conditions on the shape of the simplices in FE partitions come, in fact, from the set of algebraic requirements on the entries of $\mathbf{A}$ providing the validity of the DMPs, as is done for example in [6,8,17,21], where $\mathbf{A}$ is assumed to be irreducibly diagonally dominant.

However, we find that it is sufficient and more convenient to require the matrix $\mathbf{A}$ to be a Stieltjes matrix, i.e., symmetric, positive definite and having nonpositive off-diagonal entries. Notice that Stieltjes matrices form a subclass of M-matrices which are not required to be symmetric [26, p. 85] or [12, p. 121]. M-matrices have nonnegative inverse, which is a sufficient and necessary condition for the DMP in the sense of Definition 1. In the case of Stieltjes matrices we avoid checking the irreducibility of the finite element matrix which is not always true (cf. [9, p. 4]) and, moreover, it might be difficult to verify, in general.

Before we formulate the main result, we compute the element stiffness and mass matrices for an interval I of length d, for a triangle T, and for a prism $P = T \times I$. It is well known that if $\ell_0(z) = 1 - z/d$ and $\ell_1(z) = z/d$, $z \in I$, are the 1D shape functions then the corresponding local (element) stiffness and local (element) mass matrices $\mathbf{S}^{(I)}$ and $\mathbf{M}^{(I)}$ with entries $S_{ij}^{(I)} = \int_I \ell'_{i-1}\ell'_{j-1} \, \mathrm{d}z$ and $M_{ij}^{(I)} = \int_I \ell_{i-1}\ell_{j-1} \, \mathrm{d}z$, $i, j = 1, 2$, respectively, are

$$\mathbf{S}^{(I)} = \frac{1}{d}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{M}^{(I)} = \frac{d}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

The element matrices for the triangle are well known, too, see e.g. [2,7,9,16,29]. If we use the barycentric coordinates λ_A, λ_B, and λ_C as the shape functions and if we denote by α, β, and γ the corresponding angles, see Fig. 1 (left), then

$$\mathbf{S}^{(T)} = \frac{1}{2}\begin{pmatrix} \cot\beta + \cot\gamma & -\cot\gamma & -\cot\beta \\ -\cot\gamma & \cot\alpha + \cot\gamma & -\cot\alpha \\ -\cot\beta & -\cot\alpha & \cot\alpha + \cot\beta \end{pmatrix}, \quad \mathbf{M}^{(T)} = \frac{|T|}{12}\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

where $|T|$ stands for the area of the triangle T. Finally, it is an easy exercise to verify that the element stiffness and mass matrices for the prism $P = T \times I$, see Fig. 1 (right), are given by

$$\mathbf{S}^{(P)} = \frac{d}{6}\begin{pmatrix} 2\mathbf{S}^{(T)} & \mathbf{S}^{(T)} \\ \mathbf{S}^{(T)} & 2\mathbf{S}^{(T)} \end{pmatrix} + \frac{1}{d}\begin{pmatrix} \mathbf{M}^{(T)} & -\mathbf{M}^{(T)} \\ -\mathbf{M}^{(T)} & \mathbf{M}^{(T)} \end{pmatrix}, \quad \mathbf{M}^{(P)} = \frac{d}{6}\begin{pmatrix} 2\mathbf{M}^{(T)} & \mathbf{M}^{(T)} \\ \mathbf{M}^{(T)} & 2\mathbf{M}^{(T)} \end{pmatrix}.$$

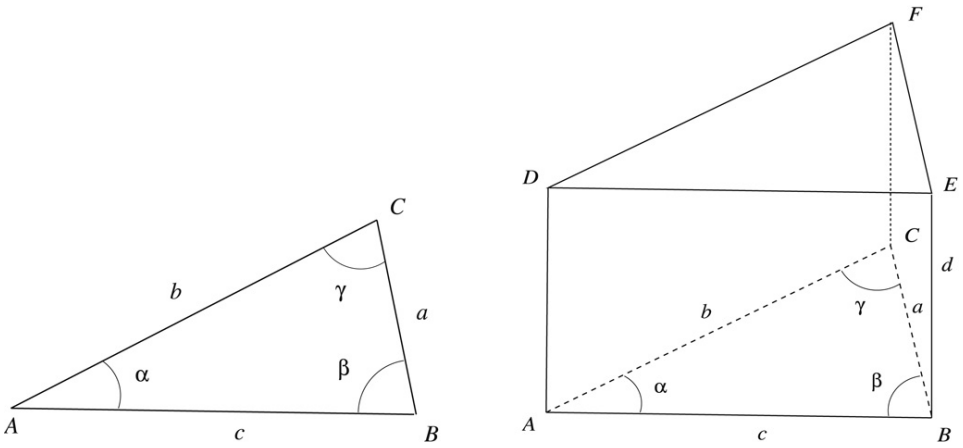


Fig. 1. Basic notation for the triangular and prismatic elements.

Notice the tensor (Kronecker) product structures $\mathbf{S}^{(P)} = \mathbf{M}^{(I)} \otimes \mathbf{S}^{(T)} + \mathbf{S}^{(I)} \otimes \mathbf{M}^{(T)}$ and $\mathbf{M}^{(P)} = \mathbf{M}^{(I)} \otimes \mathbf{M}^{(T)}$. For later reference, we introduce explicit expressions for certain entries of $\mathbf{S}^{(P)}$ and $\mathbf{M}^{(P)}$. If

$$\varphi_A(x,y,z) = \lambda_A(x,y)\ell_0(z), \qquad \varphi_B(x,y,z) = \lambda_B(x,y)\ell_0(z),$$
$$\varphi_D(x,y,z) = \lambda_A(x,y)\ell_1(z), \qquad \varphi_E(x,y,z) = \lambda_B(x,y)\ell_1(z),$$

then

$$\int_P \nabla\varphi_A \cdot \nabla\varphi_B \,\mathrm{d}P = -\frac{d}{12}\left(2\cot\gamma - \frac{|T|}{d^2}\right), \qquad \int_P \varphi_A\varphi_B \,\mathrm{d}P = \frac{d|T|}{36}, \tag{9}$$

$$\int_P \nabla\varphi_A \cdot \nabla\varphi_D \,\mathrm{d}P = \frac{d}{12}\left(\cot\beta + \cot\gamma - \frac{2|T|}{d^2}\right), \qquad \int_P \varphi_A\varphi_D \,\mathrm{d}P = \frac{d|T|}{36}, \tag{10}$$

$$\int_P \nabla\varphi_A \cdot \nabla\varphi_E \,\mathrm{d}P = -\frac{d}{12}\left(\cot\gamma + \frac{|T|}{d^2}\right), \qquad \int_P \varphi_A\varphi_E \,\mathrm{d}P = \frac{d|T|}{72}. \tag{11}$$

In what follows, all inequalities between matrices, vectors, and scalars are to be understood entrywise. For example, the symbol $\mathbf{A} \geq 0$ means that all entries of a matrix $\mathbf{A} = (a_{ij})_{i,j=1}^N$ are nonnegative, i.e., $a_{ij} \geq 0$ for all $i, j = 1, 2, \ldots, N$.

Definition 2. Let $P = T \times I$ be a prism and let $\alpha_{\max}^{(T)} \geq \alpha_{\mathrm{med}}^{(T)} \geq \alpha_{\min}^{(T)} > 0$ be the maximal, medium, and minimal angles of the triangular base T of the prism P, respectively. We define the lower and upper bounds for the altitude of the prism P as

$$d_L^{(P)} = \left(\frac{2\cot\alpha_{\max}^{(T)}}{|T|} - \frac{\|c\|_{\infty,P}}{3}\right)^{-\frac{1}{2}}, \qquad d_U^{(P)} = \left(\frac{\|c\|_{\infty,P}}{6} + \frac{\cot\alpha_{\mathrm{med}}^{(T)} + \cot\alpha_{\min}^{(T)}}{2|T|}\right)^{-\frac{1}{2}}. \tag{12}$$

The lower bound $d_L^{(P)}$ is well defined only if $\frac{2\cot\alpha_{\max}^{(T)}}{|T|} - \frac{\|c\|_{\infty,P}}{3} > 0$.

Notice that $\alpha_{\mathrm{med}}^{(T)} < \pi/2$ and $\alpha_{\min}^{(T)} \leq \pi/3$ for any triangle. Thus, $d_U^{(P)}$ is always well defined by (12).

Theorem 2. *Let $\mathcal{T}_{h,\tau}$ be a prismatic partition of Ω. For a prism $P \in \mathcal{T}_{h,\tau}$, let values $d_L^{(P)}$ and $d_U^{(P)}$ be defined by (12), and let $d^{(P)}$ denote the altitude of the prism P. If*

$$d_L^{(P)} \leq d^{(P)} \leq d_U^{(P)} \quad \text{for all } P \in \mathcal{T}_{h,\tau}, \tag{13}$$

then problem (5) satisfies the DMP according to Definition 1.

Proof. We have

$$a_{ij} = \sum_{P \subseteq \operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j} \int_P \left(\nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j\right) \mathrm{d}P = \sum_{P \subseteq \operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j} a_{ij}^{(P)}.$$

As the finite element matrix associated to our problem is obviously symmetric and positive definite, we only need to show that

$$a_{ij}^{(P)} \leq 0 \tag{14}$$

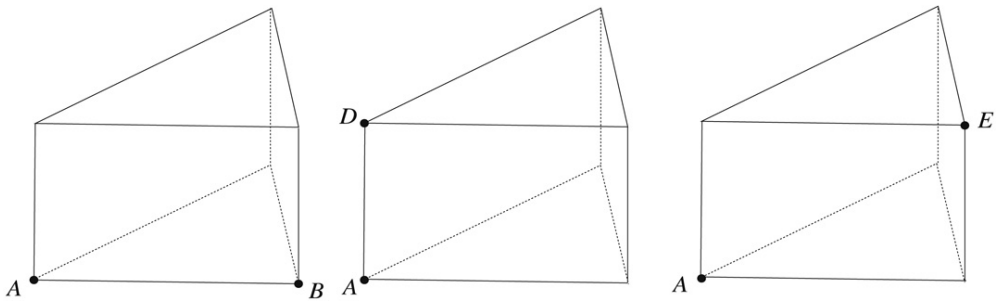


Fig. 2. Illustration of node positions in cases (i), (ii), and (iii).

for all $i \neq j$. Then the matrix $\mathbf{A}$ is a Stieltjes matrix, hence, $\mathbf{A}^{-1} \geq 0$, see [26, p. 85]. Further, because $\phi_i \geq 0$ and $f \leq 0$, we have $F_i \leq 0$ for all $i = 1, \ldots, N$. Thus, by (8), we obtain $\mathbf{y} \leq 0$ and the DMP (7) holds.

It remains to prove (14). Let us consider a prism $P \in \mathcal{T}_{h,\tau}$, $P = T \times I$. We adopt the notation from Fig. 1 and we use the short-hand notation $d = d^{(P)}$ for the altitude of the prism. Since we assume that $d_L^{(P)}$ is well defined, we can reformulate conditions (12) and (13) equivalently as

$$-2\cot\alpha_{\max}^{(T)} + \frac{|T|}{d^2} + \|c\|_{\infty,P}\frac{|T|}{3} \leq 0 \tag{15}$$

and

$$\|c\|_{\infty,P}\frac{|T|}{3} - \frac{2|T|}{d^2} + \cot\alpha_{\mathrm{med}}^{(T)} + \cot\alpha_{\min}^{(T)} \leq 0. \tag{16}$$

To compute all the entries $a_{ij}^{(P)}$ of the local finite element matrix it is enough to distinguish the following three different cases, see Fig. 2.

(i) Let A be any vertex of P and let B be one of the two remaining vertices in the same triangular base. If the basis functions ϕ_i and ϕ_j correspond to the vertices A and B, respectively, then by (9)

$$a_{ij}^{(P)} = \int_P \nabla\varphi_A \cdot \nabla\varphi_B \,\mathrm{d}P + \int_P c\varphi_A\varphi_B \,\mathrm{d}P \leq \frac{d}{12}\left(-2\cot\gamma + \frac{|T|}{d^2} + \|c\|_{\infty,P}\frac{|T|}{3}\right). \tag{17}$$

The nonpositivity of this value is guaranteed by (15), because the cotangent is a decreasing function, and hence $-\cot\gamma \leq -\cot\alpha_{\max}^{(T)}$.

(ii) Let A be any vertex of P and let D be the vertex in the opposite triangular base joined with A by an edge. If the basis functions ϕ_i and ϕ_j correspond to the vertices A and D, respectively, then by (10)

$$a_{ij}^{(P)} = \int_P \nabla\varphi_A \cdot \nabla\varphi_D \,\mathrm{d}P + \int_P c\varphi_A\varphi_D \,\mathrm{d}P \leq \frac{d}{12}\left(\cot\beta + \cot\gamma - \frac{2|T|}{d^2} + \|c\|_{\infty,P}\frac{|T|}{3}\right). \tag{18}$$

The nonpositivity of this value follows from (16), because $\cot\beta + \cot\gamma \leq \cot\alpha_{\mathrm{med}}^{(T)} + \cot\alpha_{\min}^{(T)}$.

(iii) Let A be any vertex of P and let E be the vertex in the opposite triangular base not joined with A by any edge. If the basis functions ϕ_i and ϕ_j correspond to the vertices A and E, respectively, then by (11)

$$\begin{aligned} a_{ij}^{(P)} &= \int_P \nabla\varphi_A \cdot \nabla\varphi_E \,\mathrm{d}P + \int_P c\varphi_A\varphi_E \,\mathrm{d}P \leq -\frac{d}{12}\left(\cot\gamma + \frac{|T|}{d^2} - \|c\|_{\infty,P}\frac{|T|}{6}\right) \\ &= \frac{d}{24}\left(-2\cot\gamma + \frac{|T|}{d^2} + \|c\|_{\infty,P}\frac{|T|}{3}\right) - \frac{3d}{24}\frac{|T|}{d^2}. \end{aligned} \tag{19}$$

This is clearly nonpositive due to case (i), see (17). □

5. Construction of meshes for the DMP

It is not immediately clear, how the prismatic partitions satisfying the crucial conditions (12) and (13) look like. In this section, we prove several results which characterize prismatic partitions with the desired properties (12) and (13). First of all, we present Lemma 1 which states that conditions (12) and (13) are sharp in the sense that their violation leads to positive entries in the local finite element matrices in certain situations.

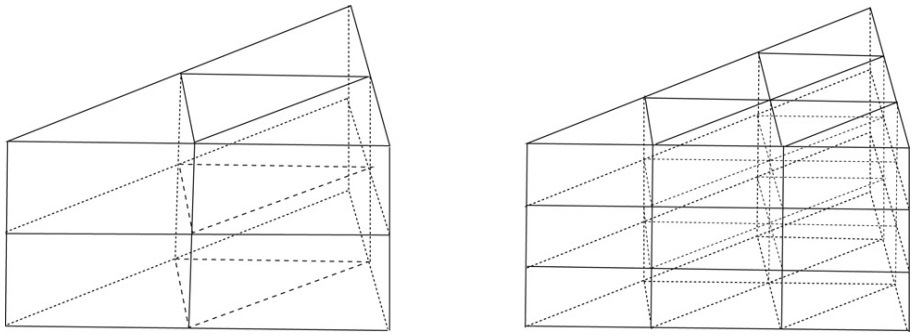


Fig. 3. Illustrations of 2-fold and 3-fold uniform refinements of a prism.

Lemma 1. *Let $\mathcal{T}_{h,\tau}$ be a prismatic partition of Ω and let the reaction coefficient c be piecewise constant so that $c|_P = \text{const}$. for each prism P in $\mathcal{T}_{h,\tau}$. Then all off-diagonal entries $a_{ij}^{(P)}$ of the local finite element matrices are nonnegative if and only if conditions (12) and (13) are satisfied.*

Proof. The "if" part is a special case of Theorem 2. The "only if" part follows from the fact that (17)–(19) hold in our case as equalities, because $\|c\|_{\infty,P} = c|_P$. Thus, if (12) and (13) were not valid then at least one of entries (17) and (18) would be positive. □

In the following proofs we implicitly assume that $d_L^{(P)}$ is well defined and we use an equivalent reformulation of conditions (12) and (13)

$$\frac{\|c\|_{\infty,P}}{6}|T| + \frac{\cot\alpha_{\text{med}}^{(T)} + \cot\alpha_{\text{min}}^{(T)}}{2} \le \frac{|T|}{(d^{(P)})^2} \le 2\cot\alpha_{\text{max}}^{(T)} - \frac{\|c\|_{\infty,P}}{3}|T|. \tag{20}$$

Below, Lemma 2 shows an important observation about the uniform (global) refinement of the prismatic partitions satisfying (12) and (13).

Definition 3. Let m be a positive integer and $\mathcal{T}_{h,\tau}$ be a prismatic partition of Ω. First, we refine each edge in $\mathcal{T}_{h,\tau}$ into m subedges. Further, for each prism $P \in \mathcal{T}_{h,\tau}$, $P = T \times I$, we refine the triangular base T into m^2 similar triangles $\widetilde{T}_i \subset T$, $i = 1, 2, \ldots, m^2$, each segment I into m equal segments $\widetilde{I}_j \subset I$, $j = 1, 2, \ldots, m$, and we obtain m^3 prisms $\widetilde{P}_{i,j} = \widetilde{T}_i \times \widetilde{I}_j$, $\widetilde{P}_{i,j} \subset P$. These prisms $\widetilde{P}_{i,j}$ form a new face-to-face prismatic partition $\widetilde{\mathcal{T}}_{h,\tau}$ of Ω which we call *m-fold uniform refinement* of $\mathcal{T}_{h,\tau}$. If $m = 1$ then $\widetilde{\mathcal{T}}_{h,\tau} = \mathcal{T}_{h,\tau}$. See Fig. 3 for an illustration.

Lemma 2. *If a prismatic partition $\mathcal{T}_{h,\tau}$ satisfies (12) and (13) then its any m-fold uniform refinement $\widetilde{\mathcal{T}}_{h,\tau}$ with $m \ge 1$ satisfies (12) and (13) as well.*

Proof. Let $P \in \mathcal{T}_{h,\tau}$, $P = T \times I$, and $\widetilde{P} \in \widetilde{\mathcal{T}}_{h,\tau}$, $\widetilde{P} = \widetilde{T} \times \widetilde{I}$, be such that $\widetilde{P} \subset P$. Then $m^2|\widetilde{T}| = |T|$ and $m\tilde{d} = d$, where d and $\tilde{d}$ stand for the altitudes of prisms P and $\widetilde{P}$, respectively. In addition, the triangles T and $\widetilde{T}$ are similar, and therefore, the corresponding maximal, medium, and minimal angles $\alpha \ge \beta \ge \gamma$ in T and $\tilde{\alpha} \ge \tilde{\beta} \ge \tilde{\gamma}$ in $\widetilde{T}$ are equal.

Since conditions (12) and (13) and, equivalently, (20) are valid for P, we estimate

$$\begin{aligned}\frac{\|c\|_{\infty,\widetilde{P}}}{6}|\widetilde{T}| + \frac{\cot\tilde{\beta} + \cot\tilde{\gamma}}{2} &\le \frac{\|c\|_{\infty,P}}{6}\frac{|T|}{m^2} + \frac{\cot\beta + \cot\gamma}{2} \\ &\le \frac{|T|}{d^2} \le 2\cot\alpha - \frac{\|c\|_{\infty,P}}{3}\frac{|T|}{m^2} \le 2\cot\tilde{\alpha} - \frac{\|c\|_{\infty,\widetilde{P}}}{3}|\widetilde{T}|,\end{aligned} \tag{21}$$

where we use the facts that $\|c\|_{\infty,\widetilde{P}} \le \|c\|_{\infty,P}$ and $m \ge 1$. To finish the proof we realize that inequalities (21) actually prove conditions (20) for the prism $\widetilde{P}$, because $|T|/d^2 = |\widetilde{T}|/\tilde{d}^2$. □

The following definition and the subsequent theorems provide easily verifiable sufficient conditions for prismatic partitions that yield the DMP. Furthermore, they give practical hints on how to construct such partitions.

Definition 4. Let $\mathcal{T}_{h,\tau} = \mathcal{T}_h^{\mathcal{G}} \times \mathcal{T}_\tau^{\mathcal{I}}$ be a prismatic partition. We denote by d_i, $i = 1, 2, \ldots, M$, the lengths of the M segments in $\mathcal{T}_\tau^{\mathcal{I}}$, by $T_{\max}$ and $T_{\min}$ the triangles in $\mathcal{T}_h^{\mathcal{G}}$ with the largest and smallest areas, respectively, and by $\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}}$ and $\alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}}$ the maximal and minimal angles in the whole triangulation $\mathcal{T}_h^{\mathcal{G}}$, respectively.

We say that the prismatic partition $\mathcal{T}_{h,\tau}$ is *well-shaped* for the DMP if $\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} < \pi/2$ and if

$$\frac{1}{2}|T_{\max}| \tan \alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} \le d_i^2 \le |T_{\min}| \tan \alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}} \quad \forall i = 1, 2, \ldots, M. \tag{22}$$

In addition, if $\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} < \pi/2$ and if

$$\frac{1}{2}|T_{\max}| \tan \alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} < d_i^2 < |T_{\min}| \tan \alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}} \quad \forall i = 1, 2, \ldots, M, \tag{23}$$

then the prismatic partition $\mathcal{T}_{h,\tau}$ is called *strictly well-shaped* for the DMP.

Furthermore, it is easy to see that any m-fold uniform refinement of a (strictly) well-shaped prismatic partition is again (strictly) well-shaped. Hence, we can say that conditions (22) and (23) only limit the shape of the prisms and not their actual sizes. Before we introduce theorems stating that well-shaped partitions guarantee the DMP we present Lemma 3 which discusses geometric properties of the well-shaped prismatic partitions. In particular, it demonstrates that the maximal angle in the base triangulation should be much smaller than the technical assumption $\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} < \pi/2$ requires.

Lemma 3. *Let $\mathcal{T}_{h,\tau} = \mathcal{T}_h^{\mathcal{G}} \times \mathcal{T}_\tau^{\mathcal{I}}$ be a well-shaped prismatic partition of a cylindrical domain $\Omega = \mathcal{G} \times \mathcal{I}$. Let $T_{\max}$, $T_{\min}$, $\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}}$, and $\alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}}$ have the same meaning as in Definition 4. Then*

$$\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} \le \arctan\sqrt{8} \approx 70.5288^\circ, \tag{24}$$

$$\alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}} \ge \arctan(\sqrt{5}/2) \approx 48.1897^\circ, \tag{25}$$

and

$$\frac{|T_{\max}|}{|T_{\min}|} \le 2. \tag{26}$$

Proof. We prove this lemma by contradiction. If a prismatic partition $\mathcal{T}_{h,\tau} = \mathcal{T}_h^{\mathcal{G}} \times \mathcal{T}_\tau^{\mathcal{I}}$ is well-shaped then

$$\frac{1}{2}|T_{\max}| \tan \alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} \le |T_{\min}| \tan \alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}} \tag{27}$$

independently of the particular partition $\mathcal{T}_\tau^{\mathcal{I}}$ of $\mathcal{I}$.

Let us suppose that (24) is not valid and let us consider the triangle $T \in \mathcal{T}_h^{\mathcal{G}}$ such that its greatest angle $\alpha = \alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} > \arctan\sqrt{8} = 2\arctan(\sqrt{2}/2)$. The smallest angle γ in this T satisfies $\gamma \le \pi/2 - \alpha/2$ which is equivalent to $\cot\gamma \ge \cot(\pi/2 - \alpha/2)$. It can be easily verified that the inequality $\alpha > 2\arctan(\sqrt{2}/2)$ is equivalent to the inequality $2\cot\alpha < \cot(\pi/2 - \alpha/2)$. Thus, $2\cot\alpha < \cot\gamma$. From (27) and from the technical assumption $\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} < \pi/2$ we conclude that

$$1 \le \frac{|T_{\max}|}{|T_{\min}|} \le \frac{2\cot\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}}}{\cot\alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}}} \le \frac{2\cot\alpha}{\cot\gamma} < 1, \tag{28}$$

which is a contradiction and (24) is proved.

To prove (25) by contradiction, we consider the triangle $T \in \mathcal{T}_h^{\mathcal{G}}$ such that its smallest angle $\gamma = \alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}} < \arctan(\sqrt{5}/2) = 2\arctan(1/\sqrt{5})$. The greatest angle α in this T satisfies $\alpha \ge \pi/2 - \gamma/2$ which is equivalent to $\cot\alpha \le \cot(\pi/2 - \gamma/2)$. It can easily be verified that the inequality $\gamma < 2\arctan(1/\sqrt{5})$ is equivalent to the inequality $2\cot(\pi/2 - \gamma/2) < \cot\gamma$. Thus, $2\cot\alpha < \cot\gamma$ which is a contradiction due to (28).

Finally, if (26) was not true then (27) together with the inequality $\tan\alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}} \le \tan\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}}$ would imply

$$2 < \frac{|T_{\max}|}{|T_{\min}|} \le \frac{2\tan\alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}}}{\tan\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}}} \le 2, \tag{29}$$

which is a contradiction, again. □

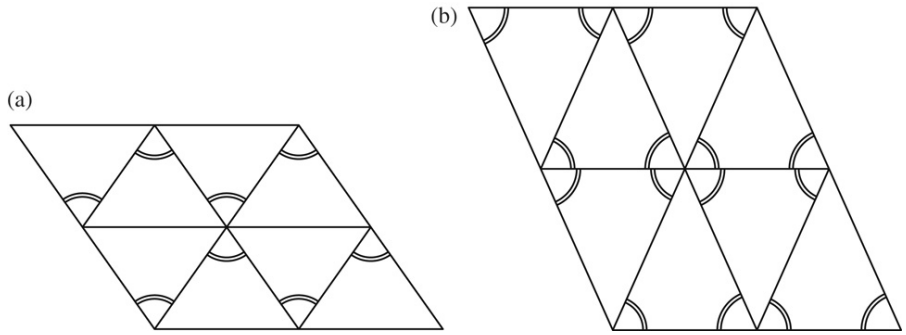


Fig. 4. Two examples of isosceles triangulations. (a) The greater angles ($\approx 70.5288°$) are marked by double arcs and the smaller angles ($\approx 54.7356°$) have no mark. (b) The greater angles ($\approx 65.9052°$) are marked by double arcs and the smaller angles ($\approx 48.1897°$) have no mark.

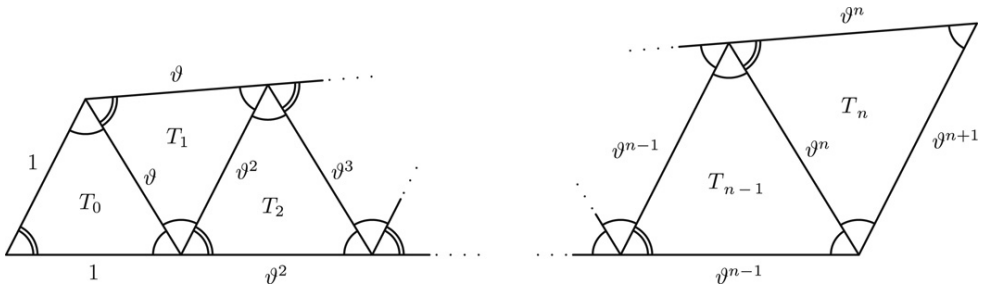


Fig. 5. Construction of a triangulation consisting of isosceles triangles which are close to equilateral triangles and whose areas grow slowly such that $|T_n|/|T_0|$ is close to 2. The angles marked by double arcs are equal to $\pi/3+2\omega$ and the ones marked by single arcs are $\pi/3-\omega$, where ω is a small positive angle. If a stands for the lengths of two sides of the isosceles triangle with angle $\pi/3+2\omega$ in between them then the third side has length ϑa, where $\vartheta = 2\sin(\pi/3+\omega)$.

Notice that the strictly well-shaped prismatic partitions satisfy (24)–(26) with strict inequalities. Further notice that for an arbitrary polygon, a triangulation satisfying (24) and (25) need not exist.

We would also like to emphasize that conditions (24) and (25) are sharp in the sense that there exist well-shaped prismatic partitions with the maximal and minimal angles equal to $\arctan\sqrt{8}$ and $\arctan(\sqrt{5}/2)$, respectively. Let us construct two examples of such well-shaped prismatic partitions.

(a) Let $\mathcal{T}_{h,\tau}^{(a)}$ consist of copies of a prism $P = T \times I$ whose base T is an isosceles triangle with angles $\alpha = \arctan\sqrt{8} \approx 70.5288°$ and $\beta = \gamma = \pi/2 - \alpha/2 \approx 54.7356°$. If the altitudes of all these prisms are set by (12) to be $d^2 = \left(d_L^{(P)}\right)^2 = \left(d_U^{(P)}\right)^2 = \sqrt{2}|T|$, then this prismatic partition $\mathcal{T}_{h,\tau}^{(a)}$ is well-shaped. See Fig. 4(a).

(b) Similarly, to show that (25) is sharp, we construct a prismatic partition $\mathcal{T}_{h,\tau}^{(b)}$ consisting of prisms with bases T being isosceles triangles with angles $\gamma = \arctan(\sqrt{5}/2) \approx 48.1897°$ and $\alpha = \beta = \pi/2 - \gamma/2 \approx 65.9052°$. If the altitudes of these prisms are chosen in agreement with (12) in between

$$\frac{1}{2}\sqrt{5}|T| = \left(d_L^{(P)}\right)^2 \le d^2 \le \left(d_U^{(P)}\right)^2 = \frac{2}{3}\sqrt{5}|T|,$$

then such a prismatic partition is well-shaped. See Fig. 4(b) for an illustration. Notice that the whole plane $\mathbb{R}^2$ can be tiled by copies of any triangle.

On the other hand, condition (26) is not sharp in this sense. A well-shaped prismatic partition such that $|T_{\max}|/|T_{\min}| = 2$ does not exist. Indeed, if $|T_{\max}|/|T_{\min}| = 2$ then (29) implies that $\alpha_{\min}^{\mathcal{T}_h^\vartheta} = \alpha_{\max}^{\mathcal{T}_h^\vartheta}$, hence all triangles in the triangulation $\mathcal{T}_h^\vartheta$ are equilateral and consequently all of them have equal areas. This obviously contradicts the fact that $|T_{\max}|/|T_{\min}| = 2$. Nevertheless, for any $\varepsilon > 0$, it is possible to construct a well-shaped prismatic partition such that $|T_{\max}|/|T_{\min}| = 2-\varepsilon$. Fig. 5 illustrates the construction of the base triangulation for such prismatic partitions. For example, to have $1.99 < |T_{\max}|/|T_{\min}| < 2$ it is enough to set $\omega = 0.03°$ and construct 381 ($n = 380$) triangles according to Fig. 5. If the altitudes of the prisms satisfy $0.749029 < d^2 < 0.749546$ then the resulting prismatic partition is strictly well-shaped. There are no interior points in Fig. 5. In order to obtain some we can uniformly refine the indicated partition or we can mirror the triangulation with respect to the (almost) horizontal lines.

The practical significance of Lemma 3 lies in the fact that it gives necessary conditions for a partition to be well-shaped. If at least one condition of (24)–(26) is not satisfied then the corresponding prismatic partition is not well-shaped. The following theorem says that well-shaped prismatic partitions yield the DMP in the pure diffusion case, i.e., for $c = 0$ in Ω.

Theorem 3. *Let $\Omega = \mathcal{G} \times \mathcal{I} \subset \mathbb{R}^3$ be a cylindrical domain and let $\mathcal{T}_{h,\tau} = \mathcal{T}_h^{\mathcal{G}} \times \mathcal{T}_\tau^{\mathcal{I}}$ be its well-shaped prismatic partition. If $c = 0$ in Ω, then discretization (5) based on the prismatic partition $\mathcal{T}_{h,\tau}$ satisfies the DMP according to Definition 1.*

Proof. Lemma 3, statement (24), implies that all angles in the triangulation $\mathcal{T}_h^{\mathcal{G}}$ are well below $\pi/2$. Hence, tangents and cotangents of all angles in $\mathcal{T}_h^{\mathcal{G}}$ are positive.

Let us consider a prism $P = T \times I$ in $\mathcal{T}_{h,\tau}$. Further, let $\alpha \geq \beta \geq \gamma > 0$ be the angles in the triangle T, and let d stand for the altitude of the prism P. Assumption (22) implies

$$\frac{\cot\beta + \cot\gamma}{2} \leq \frac{|T|}{|T_{\min}|} \cot\alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}} \leq \frac{|T|}{d^2} \leq \frac{|T|}{|T_{\max}|} 2\cot\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} \leq 2\cot\alpha.$$

Thus, conditions (20) and, equivalently, (12) and (13) are satisfied for all prisms $P \in \mathcal{T}_{h,\tau}$ and Theorem 2 concludes the proof. □

Theorem 4 below characterizes a class of prismatic partitions which provide the DMP for the general diffusion-reaction case $c \geq 0$ and $c \not\equiv 0$ in Ω. Such partitions must be *strictly* well-shaped and fine enough. Moreover, Theorem 4 quantifies how fine the suitable partitions have to be.

Theorem 4. *Let $\Omega = \mathcal{G} \times \mathcal{I} \subset \mathbb{R}^3$ be a cylindrical domain and let $\mathcal{T}_{h,\tau} = \mathcal{T}_h^{\mathcal{G}} \times \mathcal{T}_\tau^{\mathcal{I}}$ be its strictly well-shaped prismatic partition. Furthermore, let $m \geq 1$ be an integer such that*

$$m^2 \geq \max_{P \in \mathcal{T}_{h,\tau}} \frac{\|c\|_{\infty,P}\,|T|}{M_P}, \tag{30}$$

where $P = T \times I$ is a prism and

$$M_P = \min\left\{ 6\left(\frac{|T|}{d^2} - \frac{\cot\beta + \cot\gamma}{2}\right),\ 3\left(2\cot\alpha - \frac{|T|}{d^2}\right)\right\}, \tag{31}$$

with $\alpha \geq \beta \geq \gamma$ being the angles in the triangle T and d standing for the altitude of the prism P. Then discretization (5) based on the m-fold uniform refinement $\widetilde{\mathcal{T}}_{h,\tau}$ of $\mathcal{T}_{h,\tau}$ satisfies the DMP according to Definition 1.

Proof. Let us consider the m-fold uniform refinement $\widetilde{\mathcal{T}}_{h,\tau}$ of the strictly well-shaped prismatic partition $\mathcal{T}_{h,\tau}$ with $m \geq 1$ given by (30). Let $\widetilde{P} = \widetilde{T} \times \tilde{I}$ be a prism in $\widetilde{\mathcal{T}}_{h,\tau}$ and let $P \in \mathcal{T}_{h,\tau}$, $P = T \times I$, be such a prism that $\widetilde{P} \subset P$. Denote by $\tilde{d}$ and d the altitudes of prisms $\widetilde{P}$ and P, respectively. Clearly, $m^2|\widetilde{T}| = |T|$, $m\tilde{d} = d$, and the triangles $\widetilde{T}$ and T are similar, hence the corresponding angles $\tilde{\alpha} \geq \tilde{\beta} \geq \tilde{\gamma} > 0$ in $\widetilde{T}$ and $\alpha \geq \beta \geq \gamma > 0$ in T are equal. Notice that all angles in both $\mathcal{T}_{h,\tau}$ and $\widetilde{\mathcal{T}}_{h,\tau}$ are acute by Lemma 3.

Since the prismatic partition $\mathcal{T}_{h,\tau}$ is strictly well-shaped, we have $M_P > 0$ and assumption (30) implies

$$\|c\|_{\infty,\widetilde{P}}\,|\widetilde{T}| \leq \|c\|_{\infty,P}\,\frac{|T|}{m^2} \leq M_P,$$

where we used the inequality $\|c\|_{\infty,\widetilde{P}} \leq \|c\|_{\infty,P}$. Hence, from definition (31) we obtain

$$\frac{\|c\|_{\infty,\widetilde{P}}\,|\widetilde{T}|}{6} + \frac{\cot\tilde{\beta} + \cot\tilde{\gamma}}{2} \leq \frac{|\widetilde{T}|}{\tilde{d}^2} \leq 2\cot\tilde{\alpha} - \frac{\|c\|_{\infty,\widetilde{P}}\,|\widetilde{T}|}{3},$$

where we utilize the facts that $\tilde{\alpha} = \alpha$, $\tilde{\beta} = \beta$, $\tilde{\gamma} = \gamma$, and $|\widetilde{T}|/\tilde{d}^2 = |T|/d^2$. Thus we verified the validity of conditions (20) and, equivalently, (12) and (13) for all prisms $P \in \mathcal{T}_{h,\tau}$. Theorem 2 finishes the proof. □

Remark 3. In the pure diffusion case, i.e., $c = 0$ in Ω, the conditions for validity of the DMP limit the shape and not the size of elements, see (20). Indeed, condition (20) limits the ratio of the area of the base triangle and the square of the altitude of the prism by the angles in the base triangle, but the size (volume) of the prism can be made arbitrarily large or small while keeping this ratio constant. On the other hand, in the general case, if the reaction coefficient c does not vanish then the partition has to be, in addition, fine enough in order to obtain the DMP, see Theorem 4. This is a typical behavior of the diffusion-reaction problem and it is in agreement with the previous DMP results for elliptic problems with the reaction term, see e.g. [3,17] for simplicial finite elements.

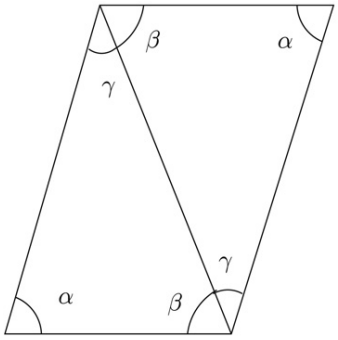


Fig. 6. The original partition.

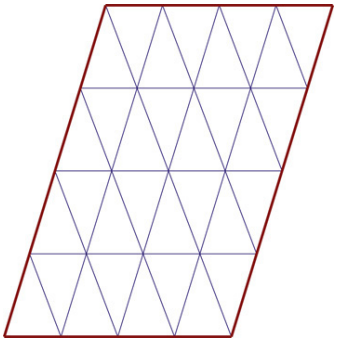


Fig. 7. The applied computational mesh.

Remark 4. Conditions (22) and (23) for the well-shaped and the strictly well-shaped prismatic partitions bound the altitudes of the prisms from two sides. Therefore, it could be troublesome or even impossible to divide an arbitrary cylindrical domain $\Omega = \mathcal{G} \times \mathcal{I}$ into layers with suitable altitudes. However, if there exists a triangulation of $\mathcal{G}$ satisfying (27) then for any altitude of Ω there exists a sequence of domains $\Omega_k = \mathcal{G} \times \mathcal{I}_k$, such that $\Omega_k \to \Omega$ as $k \to \infty$ and that a (strictly) well-shaped prismatic partition of Ω_k exists. Notice that the domains Ω_k and their (strictly) well-shaped prismatic partitions need not be necessarily nested.

Remark 5. For illustration let us consider the most favorable triangulation $\mathcal{T}_h^{\mathcal{G}}$ consisting of equilateral triangles with the same area. Let s stand for the length of each side of these triangles. Further, let the reaction coefficient c vanish. In order to satisfy conditions (12) and (13) and, hence, to obtain the DMP, the altitudes d of the prisms in the prismatic partition $\mathcal{T}_{h,\tau} = \mathcal{T}_h^{\mathcal{G}} \times \mathcal{T}_\tau^{\mathcal{I}}$ are to be limited by

$$\frac{3}{8}s^2 \le d^2 \le \frac{3}{4}s^2.$$

6. Numerical tests

In this section, we illustrate the theoretical results by numerical computations. The numerical tests also show that the DMP is actually valid for much wider class of meshes than the proposed theory predicts.

First, we construct a well-shaped triangulation for the DMP according to Definition 4. This triangulation will be used to demonstrate the usage of Lemmas 2 and 3 as well as Theorems 2–4. However, the construction of the well-shaped triangulation for the DMP requires some care. Lemma 3 gives necessary conditions on the shape of the well-shaped triangulations, but the question of finding the necessary and sufficient conditions is still open.

In order to construct a strictly well-shaped prismatic partition we consider a uniform triangulation consisting of congruent triangles as presented in Fig. 6. All computations are performed using two times refined original partition (4-fold refinement), presented in Fig. 7. The prismatic partition is constructed from this triangulation by creating four layers of prismatic elements with equal altitudes d. For these kinds of partitions, the well-shapedness condition (22) reduces to a simple inequality

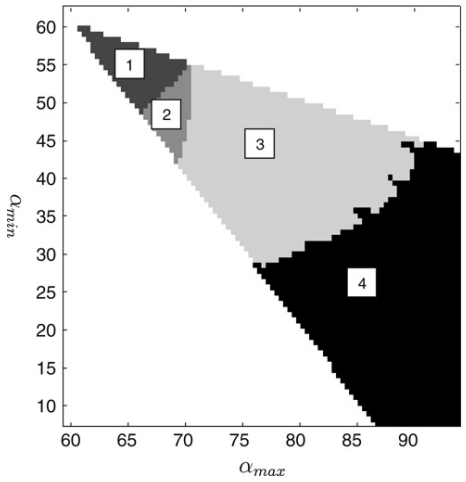


Fig. 8. Characterization of the applied partitions according to $\alpha_{\max}$ and $\alpha_{\min}$. In domains 1 and 2, the DMP is guaranteed by Theorems 2 and 3, respectively. Domain 3 is not covered by the theory but the DMP is valid there. Partitions corresponding to domain 4 do not yield the DMP at all.

$$\frac{1}{2}\tan\alpha_{\max} \leq \tan\alpha_{\min}. \tag{32}$$

We stress that in agreement with (1) we use zero Dirichlet boundary conditions in all computations.

In the first test, we study inequality (32) and its relation to the existence of a suitable altitude d which would yield the DMP for $c = 0$. We compare altitudes predicted by Theorems 2 and 3 with the altitudes computed numerically. Since the shape of the applied partition (see Fig. 6) is determined by the values of $\alpha_{\max}$ and $\alpha_{\min}$, we can visualize the results as a function of these two parameters. This is done in Fig. 8. Domain 1 illustrates the set of the well-shaped triangulations, according to Definition 4. Triangulations from this set satisfy the DMP by Theorem 3. In our case, domain 1 is determined by (32). Domain 2 is the set of the non-well-shaped triangulations, which satisfy the DMP with a suitable altitude d according to Theorem 2. Domain 3 corresponds to the set of triangulations for which we can computationally verify the DMP for a certain altitude d. All other triangulations (domain 4) do not satisfy the DMP for any altitude. We remark that the graining of the image is due to the finite resolution applied in computations. Still, we can verify the sharpness of the necessary bounds for $\alpha_{\min}$ and $\alpha_{\max}$ given by Lemma 3.

In Fig. 8, we can compare the set of triangulations, where the DMP is guaranteed by our theoretical results (domains 1 and 2), with the set of all triangulations yielding the DMP (domain 3). We observe that the theory covers considerable part of the triangulations yielding the DMP. On the other hand, this numerical experiment reveals that the set of triangulations yielding the DMP seems to be much wider than the theory predicts.

In the second test, we demonstrate the theoretical bounds (12) and (13) for the altitude d in the case $c = 0$, see Theorem 2. For this purpose, we construct a sequence of prismatic partitions. All these partitions are based on the same triangulation and have four layers of prisms with the altitude d varying from 0 to 1 with step 0.002. Based on the first test, we choose as the base triangulation a strictly well-shaped triangulation shown in Fig. 7 with angles 65, 60, and 55 degrees. This base triangulation is used also for all the subsequent tests.

For each prismatic partition in the sequence, we find the smallest entry $A^{-1}_{\min}$ of the inverse of the finite element system matrix, $A^{-1}_{\min} = \min_{ij} A^{-1}_{ij}$. As the DMP according to Definition 1 is valid if and only if $A^{-1}_{\min} \geq 0$, this value indicates whether the DMP property is satisfied. The results are visualized in Fig. 9. As one can observe, the computationally obtained bounds for the DMP are only little wider compared to the theoretically predicted bounds (12) and (13).

In the third test, we study the behavior of the bounds (12) and (13) for the altitude d, when the coefficient c is a constant greater than zero. We use the same prismatic partitions as in the previous case, but we vary the coefficient c from 1 to 30 with step 1. Theoretically calculated and computationally verified bounds for the altitude d yielding the DMP are visualized as functions of c in Fig. 10. In this figure, we observe that the DMP is lost for sufficiently large values of c, as predicted by bounds (12) and (13) presented in Theorem 2. The computational bounds for the DMP behave in a similar manner as the theoretical ones.

Finally, in the fourth test, we study if the DMP can be recovered for $c = 100$ by the m-fold uniform refinement, according to Theorem 4. In this case, the theoretical bounds for the altitude d with $c = 0$ are $d_L = 0.1792$ and $d_U = 0.2165$. The initial altitude was chosen between these bounds as $d_0 = 0.1930$. Fig. 11 presents the behavior of the computational and theoretical bounds for d as the refinements proceed. For the chosen value of the reaction coefficient c, the initial partition does not yield the DMP for any altitude. As the partition is strictly well-shaped, Theorem 4 states that a 3-fold ($M_P = 0.38595$ and $m = 3$) refinement should restore the DMP. This phenomenon is indeed observed in our computations. Nevertheless, the results show the existence of a suitable altitude d yielding the DMP even for $m = 2$. This test confirms that the DMP is

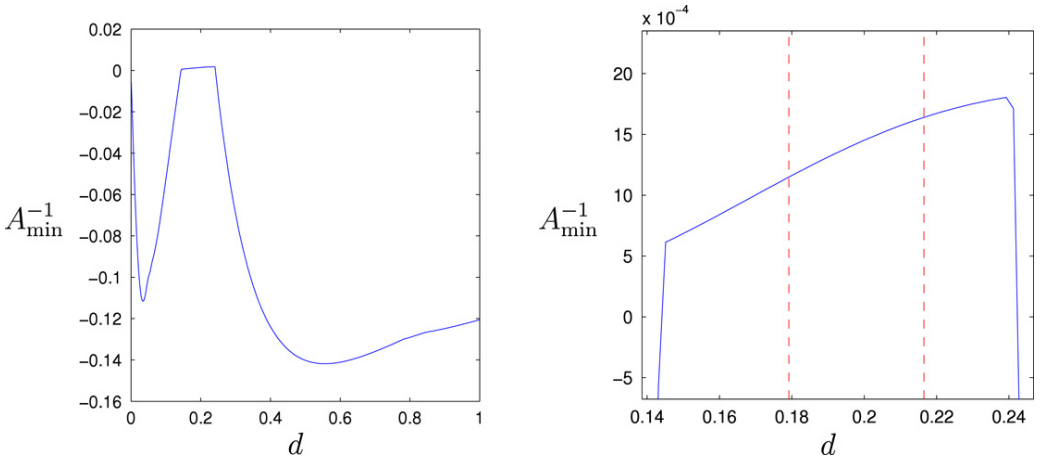


Fig. 9. The smallest entry $A_{\min}^{-1} = \min_{ij} A_{ij}^{-1}$ of the inverse of the finite element matrix as a function of the altitude d for $c = 0$ (left). Theoretical bounds (12) and (13) are plotted as the dashed lines (right). The right panel is a zoom from the left panel.

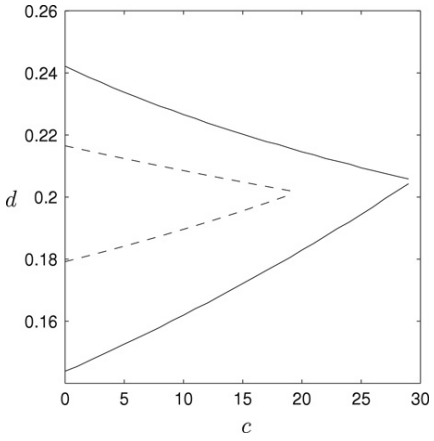


Fig. 10. Behavior of the theoretical (dashed lines) and the computational (solid lines) bounds for the altitude d as a function of the (constant) coefficient c.

valid for any m-fold uniform refinement with sufficiently large m, as predicted by Lemma 2 and Theorem 4. The theoretically predicted value of m could be, however, greater then it is necessary, in certain situations.

7. Conclusions, generalizations, and open problems

The crucial result of this paper is formulated in Theorem 2, where we present an easily verifiable condition (12) and (13) which guarantees the DMP. This theorem, however, does not provide any guidelines on how to construct suitable prismatic partitions for the validity of the DMP. Therefore, we developed the concept of the (strictly) well-shaped prismatic partitions to characterize the base triangulations which guarantee the existence of suitable altitudes of the layers of prisms. The corresponding DMP on the (strictly) well-shaped prismatic partitions is formulated and proven in Theorems 3 and 4.

In Section 6, we present various numerical tests to assess the sharpness of the theoretically obtained conditions. The first test (see Fig. 8) is of particular interest, because it indicates that the class of partitions which provide the DMP is much wider than one would expect from the theoretical results.

Let us conclude this paper by the following list of possible generalizations and open problems.

- To prove the DMP, we actually require the FE matrix $\mathbf{A}$ to have the nonnegative inverse, i.e., $\mathbf{A}^{-1} \geq 0$. It is well known that some off-diagonal entries can be positive and still one has $\mathbf{A}^{-1} \geq 0$ (see e.g. a very recent work [1] for a discussion and literature on this subject). This observation was actually used in [20] to weaken the standard condition of nonobtuseness (see [4,19]) for tetrahedral elements. A similar approach can be, obviously, applied to the case of prismatic meshes and conditions (12) and (13) can be thus weakened.
- The proofs of the DMPs for parabolic problems usually utilize the geometric conditions derived in the elliptic case, cf. [13] for the simplicial finite elements. The above presented concept of the (strictly) well-shaped prismatic partitions can be used to prove the DMP for parabolic problems discretized in space variables by prismatic finite elements.

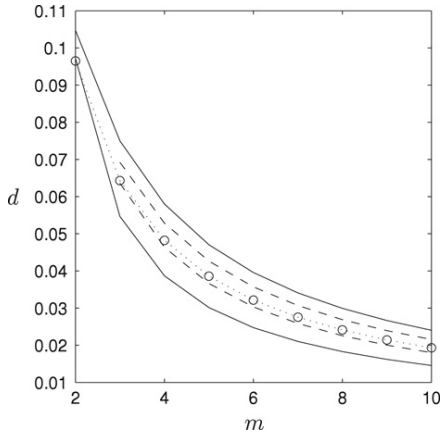


Fig. 11. Behavior of the theoretical (dashed lines) and the computational (solid lines) bounds for the altitude d with respect to the m-fold uniform refinement. The dotted line denotes the original altitude $d_0 = 0.1930$ and its refinement. The reaction coefficient is chosen as $c = 100$.

- Similarly, our concept of the (strictly) well-shaped prismatic partitions can be used to treat the DMPs for nonlinear elliptic problems. It is possible to follow the ideas introduced in [17,18].
- In recent works [5,22,23,29] the authors try to preserve the DMPs by nonlinear computational schemes which allow avoiding or considerably weakening the geometric limitations on the meshes. These techniques can be generalized to the prismatic finite elements as well.

Acknowledgements

The first author was supported by Project no. 124619 from the Academy of Finland. The second author was supported by the Academy Research Fellowship no. 208628 and Grant no. 121283 from the Academy of Finland. The third author was supported by Grant no. IAA100760702 of the Grant Agency of the Czech Academy of Sciences and by the institutional research plan no. AV0Z10190503 of the Czech Academy of Sciences.

References

[1] F. Bouchon, Monotonicity of some perturbations of irreducibly diagonally dominant M-matrices, Numer. Math. 105 (2007) 591–601.
[2] J. Brandts, S. Korotov, M. Křížek, Dissection of the path-simplex in $\mathbb{R}^n$ into n path-subsimplices, Linear Algebra Appl. 421 (2007) 382–393.
[3] J. Brandts, S. Korotov, M. Křížek, The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem, Linear Algebra Appl., in press (doi:10.1016/j.laa.2008.06.011). Available online 26 July 2008.
[4] J. Brandts, S. Korotov, M. Křížek, J. Šolc, On nonobtuse simplicial partitions, SIAM Rev. 1–20 (in press).
[5] E. Burman, A. Ern, Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes, C. R. Math. Acad. Sci. Paris 338 (2004) 641–646.
[6] P.G. Ciarlet, Discrete maximum principle for finite-difference operators, Aequationes Math. 4 (1970) 338–352.
[7] P.G. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, 1978.
[8] P.G. Ciarlet, P.-A. Raviart, Maximum principle and uniform convergence for the finite element method, Comput. Methods Appl. Mech. Engrg. 2 (1973) 17–31.
[9] A. Drăgănescu, T. Dupont, L.R. Scott, Failure of the discrete maximum principle for an elliptic finite element problem, Math. Comp. 74 (2005) 1–23.
[10] I. Faragó, R. Horváth, Discrete maximum principle and adequate discretizations of linear parabolic problems, SIAM J. Sci. Comput. 28 (2006) 2313–2336.
[11] I. Faragó, J. Karátson, Numerical Solution of Nonlinear Elliptic Problems via Preconditioning Operators, in: Theory and Applications, Advances in Computation, vol. 11, NOVA Science Publishers, New York, 2002.
[12] M. Fiedler, Special Matrices and Their Applications in Numerical Mathematics, Martinus Nijhoff Publishers, Dordrecht, 1986.
[13] H. Fujii, Some remarks on finite element analysis of time-dependent field problems, in: Theory and Practice in Finite element Structural Analysis, Univ. Tokyo Press, Tokyo, 1973, pp. 91–106.
[14] R. Horváth, On the sign-stability of the numerical solutions of the heat equation, Pure Math. Appl. 11 (2000) 281–291.
[15] R. Horváth, On the sign-stability of the numerical solutions of one-dimensional parabolic problems, Appl. Math. Modelling 32 (2008) 1570–1578.
[16] I. Holand, K. Bell, Finite Element Methods in Stress Analysis, Tapir, Trondheim, 1969.
[17] J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, Numer. Math. 99 (2005) 669–698.
[18] J. Karátson, S. Korotov, M. Křížek, On discrete maximum principles for nonlinear elliptic problems, Math. Comput. Simulation 76 (2007) 99–108.
[19] S. Korotov, M. Křížek, Acute type refinements of tetrahedral partitions of polyhedral domains, SIAM J. Numer. Anal. 39 (2001) 724–733.
[20] S. Korotov, M. Křížek, P. Neittaanmäki, Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle, Math. Comp. 70 (2001) 107–119.
[21] M. Křížek, Lin Qun, On diagonal dominance of stiffness matrices in 3D, East-West J. Numer. Math. 3 (1995) 59–69.
[22] D. Kuzmin, On the design of algebraic flux correction schemes for quadratic finite elements, J. Comp. Appl. Math. 218 (2008) 79–87.
[23] K. Lipnikov, M. Shashkov, D. Svyatskiy, Yu. Vassilevski, Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes, J. Comput. Phys. 227 (2007) 492–512.
[24] M. Protter, H. Weinberger, Maximum Principles in Differential Equations, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.
[25] A. Unterreiter, A. Juengel, Discrete minimum and maximum principles for finite element approximations of non-monotone elliptic equations, Numer. Math. 99 (2005) 485–508.
[26] R. Varga, Matrix Iterative Analysis, Prentice Hall, New Jersey, 1962.
[27] R. Varga, On discrete maximum principle, J. SIAM Numer. Anal. 3 (1966) 355–359.
[28] T. Vejchodský, P. Šolín, Discrete maximum principle for higher-order finite elements in 1D, Math. Comp. 76 (2007) 1833–1846.
[29] J. Xu, L. Zikatanov, A monotone finite element scheme for convection-diffusion equations, Math. Comp. 68 (1999) 1429–1446.

APPENDIX

C

A comparison of simplicial and block finite elements

Below we attach a copy of the paper

[A2] S. Korotov and T. Vejchodský: A comparison of simplicial and block finite elements. In: G. Kreiss, P. Lötstedt, A. Målqvist, and M. Neytcheva (eds.), *Numerical mathematics and advanced applications ENUMATH 2009*, Springer, Berlin, 2010, pp. 533–541.

A comparison of simplicial and block finite elements

Sergey Korotov and Tomáš Vejchodský

Abstract In this note we discuss and compare a performance of the finite element method (FEM) on two popular types of meshes – simplicial and block ones. A special emphasis is put on the validity of discrete maximum principles and on associated (geometric) mesh generation/refinement issues in higher dimensions. As a result, we would recommend to carefully reconsider the common belief that the simplicial finite elements are very convenient to describe complicated geometries (which appear in real-life problems), and also that the block finite elements, due to their simplicity, should be used if the geometry of the solution domain allows that.

1 Introduction

Geometrically, there are two types of finite elements (FEs) which can be naturally generalized to any dimension – simplices and blocks, where by blocks we mean Cartesian products of intervals. In what follows, we shall only consider the lowest-order finite elements, i.e., linear functions on simplices and multilinear functions on blocks. In 1D, the only reasonable element is an interval which can be understood both as a simplex and a block. Therefore, we shall make comparison for the case of two and more dimensions. Namely, we concentrate on validity of discrete maximum principles and on associated geometrical issues for mesh generation and adaptivity.

Sergey Korotov
Institute of Mathematics, Tampere University of Technology, P.O. Box 553, FI–33101 Tampere, Finland, e-mail: sergey.korotov@tut.fi

Tomáš Vejchodský
Institute of Mathematics, Academy of Sciences, Žitná 25, CZ–115 67 Prague 1, Czech Republic
e-mail: vejchod@math.cas.cz

2 Model problem at its finite element discretization

We consider the following test problem: Find a function u such that

$$-\Delta u + cu = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega, \tag{1}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded polytopic domain with Lipschitz boundary $\partial\Omega$ and $c \geq 0$. The classical solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$ of (1) satisfies the maximum principle:

$$f \leq 0 \qquad \Longrightarrow \qquad \max_{x\in\overline{\Omega}} u(x) \leq \max\{0\,,\, \max_{s\in\partial\Omega} g(s)\}. \tag{2}$$

Most of FE schemes are based on the weak formulation: Find $u \in H^1(\Omega)$ such that the boundary condition $u = g$ is satisfied in the sense of traces on $\partial\Omega$ and

$$a(u,v) = \mathscr{F}(v) \quad \forall v \in H^1_0(\Omega),$$

where $a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + cuv)\,\mathrm{d}x$, $\mathscr{F}(v) = \int_\Omega fv\,\mathrm{d}x$, $c \in L^\infty(\Omega)$, and $f \in L^2(\Omega)$.

Let $\mathscr{T}_h$ be a conforming FE mesh on $\overline{\Omega}$ with interior nodes $B_1,\ldots,B_N$ lying in Ω and boundary nodes $B_{N+1},\ldots,B_{N+N^\partial}$ lying on $\partial\Omega$. Further, let V_h be a finite-dimensional subspace of $H^1(\Omega)$, associated with $\mathscr{T}_h$ and its nodes, being spanned by the basis functions $\phi_1,\phi_2,\ldots,\phi_{N+N^\partial}$ with the following properties: $\phi_i \geq 0$ in $\overline{\Omega}$ (nonnegativity), $\phi_i(B_j) = \delta_{ij}$ (delta property), $i,j = 1,\ldots,N+N^\partial$, and $\sum_{i=1}^{N+N^\partial} \phi_i \equiv 1$ in $\overline{\Omega}$ (partition of unity). Notice that the lowest-order finite elements on simplices and on blocks meet these requirements. We also assume that the basis functions $\phi_1,\phi_2,\ldots,\phi_N$ vanish on the boundary $\partial\Omega$. Thus, they span a finite-dimensional subspace V_h^0 of $H^1_0(\Omega)$. Let, in addition, $g_h = \sum_{i=1}^{N^\partial} g_{N+i}\phi_{N+i} \in V_h$ be a suitable approximation of the function g, for example its nodal interpolant.

The FE approximation is a function $u_h = u_h^0 + g_h$ such that $u_h^0 \in V_h^0$ and

$$a(u_h, v_h) = \mathscr{F}(v_h) \quad \forall v_h \in V_h^0, \tag{3}$$

whose existence and uniqueness is also provided by the Lax-Milgram lemma.

Algorithmically, $u_h = \sum_{i=1}^{N+N^\partial} y_i\phi_i$, where y_i are the entries of the solution $\overline{\mathbf{y}} = [y_1,\ldots,y_{N+N^\partial}]^\top$ of the square system of $N+N^\partial$ linear algebraic equations

$$\overline{\mathbf{A}}\,\overline{\mathbf{y}} = \overline{\mathbf{F}}, \quad \text{where} \quad \overline{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{A}^\partial \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \text{and} \quad \overline{\mathbf{F}} = \begin{bmatrix} \mathbf{F} \\ \mathbf{F}^\partial \end{bmatrix}. \tag{4}$$

In the above, $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{A}^\partial \in \mathbb{R}^{N\times N^\partial}$, $\mathbf{0}$ and $\mathbf{I}$ stand for the zero and unit matrices of appropriate sizes. The entries of $\overline{\mathbf{A}}$ are $a_{ij} = a(\phi_j,\phi_i)$, $i = 1,\ldots,N$, $j = 1,\ldots,N+N^\partial$. The block $\mathbf{F}$ consists of entries $f_i = \mathscr{F}(\phi_i)$, $i = 1,\ldots,N$, and the block-vector $\mathbf{F}^\partial$ has entries $f_i^\partial = f_{N+i} = g_{N+i}$, $i = 1,\ldots,N^\partial$, given by the boundary data.

3 Discrete maximum principles for FEM

In this section we compare simplicial and block finite elements with respect to the so-called discrete maximum principle (DMP). For a fixed mesh $\mathscr{T}_h$, we say that the discretization (3) satisfies the DMP if

$$f \leq 0 \qquad \Longrightarrow \qquad \max_{x \in \overline{\Omega}} u_h(x) \leq \max\{0\,,\, \max_{s \in \partial\Omega} g_h(s)\}. \tag{5}$$

In the case of the lowest-order finite elements, it is well known [4] that the DMP is satisfied if (i) the stiffness matrix $\overline{\mathbf{A}}$ is monotone and if (ii) the row sums of $\overline{\mathbf{A}}$ are nonnegative. Condition (ii) is satisfied, because the basis functions form the partition of unity and the coefficient c is nonnegative. Sufficient conditions for (i) can be obtained from the theory of M-matrices [7]. This, in particular, requires the nonpositivity of the off-diagonal entries in the FE matrix $\overline{\mathbf{A}}$. Matrix $\overline{\mathbf{A}}$ is assembled from the local (element) FE matrices, $\overline{\mathbf{A}} = \sum_{K \in \mathscr{T}_h} \overline{\mathbf{A}}^K$, and hence it suffices to guarantee the nonpositivity of the off-diagonal entries of each $\overline{\mathbf{A}}^K$. This observation yields various geometric limitations for the finite elements which we discuss in what follows.

3.1 On entries of FE matrices for simplices

For simplicity, let us consider the Laplace operator only, i.e., $c \equiv 0$. In this case the off-diagonal entries $a_{ij}^K\,(i \neq j)$ of the local stiffness matrices $\overline{\mathbf{A}}^K$ for simplicial elements can be expressed in any dimension by the following formula [1]

$$a_{ij}^K = \int_K \nabla\phi_j \cdot \nabla\phi_i \,\mathrm{d}x = -\frac{\mathrm{meas}_{d-1}(F_i)\,\mathrm{meas}_{d-1}(F_j)}{d^2\,\mathrm{meas}_d(K)} \cos\alpha_{ij},$$

where α_{ij} stands for the dihedral angle between the facets F_i and F_j of the simplex $K \in \mathscr{T}_h$, see Fig. 1 (left).

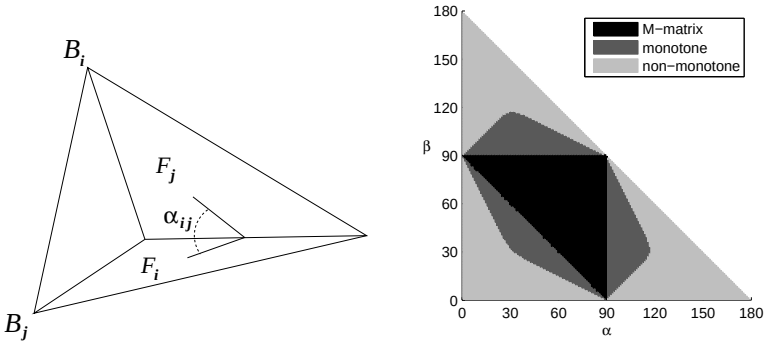


Fig. 1 The dihedral angle α_{ij} between faces F_i and F_j of a tetrahedron K (left). Results of the experiment for triangles (right).

Clearly, $a_{ij}^K \le 0$ if and only if $\alpha_{ij} \le \pi/2$. This nonobtuseness condition is well known for triangles and for tetrahedra, and it is crucial for the validity of DMPs [2]. For the case of general coefficients the conditions on meshes for DMP are stricter. Thus, if e.g. $c > 0$ then all dihedral angles in meshes have to be acute and, in addition, the meshes themselves have to be sufficiently fine due to the positive terms

$$\int_K \phi_j \phi_i \,\mathrm{d}x = \frac{d!}{(d+2)!}\,\mathrm{meas}_d(K), \quad i \neq j,$$

additionally appearing in computations, see e.g. [5, 2] for details.

Further, generalization can be obtained by requiring the stiffness matrix not to be M-matrix but to be monotone only. Theoretical handling of monotone matrices is difficult, but it can be checked numerically. Fig. 1 (right) shows results of an experiment, where we consider the Poisson problem with homogeneous Dirichlet boundary conditions. Hence, the block $\mathbf{A}$ of $\mathbf{A}$ only is relevant. The domain Ω is a triangle. The axis in Fig. 1 (right) correspond to two angles of Ω. For each pair of angles α and β, we construct a triangulation by three steps of uniform red refinement of Ω. Then we assemble the stiffness matrix $\mathbf{A}$, and color the corresponding point according to its properties. If $\mathbf{A}$ is M-matrix (has off-diagonal entries nonpositive) then the point is black. If $\mathbf{A}$ is monotone and not M-matrix then the point is dark gray. If $\mathbf{A}$ is not monotone then the point is light gray. We clearly see that in this case the stiffness matrix is M-matrix if and only if all angles are nonobtuse (black area). Further we observe that the DMP is satisfied under favorable circumstances even for angles up to 117° (dark gray area), see also [12] for a similar 3D test.

3.2 On entries of FE matrices for blocks

The analysis of the DMP for block FE partitions can be done in the same fashion as for the simplices. The results, however, strongly depend on the dimension. For simplicity we again consider the Laplacian with homogeneous Dirichlet boundary condition. Let K be an element of a d-dimensional block mesh with edges of lengths $b_1, b_2, \ldots, b_d$. If B_i and B_j are its two vertices connected by the edge of length b_1 then the corresponding entry of the local stiffness matrix $\bar{\mathbf{A}}^K$ is

$$a_{ij}^K = \frac{b_1 b_2 \ldots b_d}{3^{d-1}} \left(\sum_{k=2}^{d} \frac{1}{2b_k^2} - \frac{1}{b_1^2} \right), \quad i \neq j. \tag{6}$$

In 2D we immediately see that $a_{ij}^K \le 0$ if and only if $b_1/b_2 \le \sqrt{2}$. This yields the well-known nonnarrow condition for the DMP. A rectangle K is nonnarrow if $1/\sqrt{2} \le b_1/b_2 \le \sqrt{2}$, where b_1 and b_2 stand for the lengths of its sides of K. It can be shown [9] that the DMP is satisfied if all rectangles in the mesh $\mathscr{T}_h$ are nonnarrow.

The nonnarrow condition guarantees that the corresponding stiffness matrix is M-matrix. A similar experiment as before reveals that this condition can be weakened

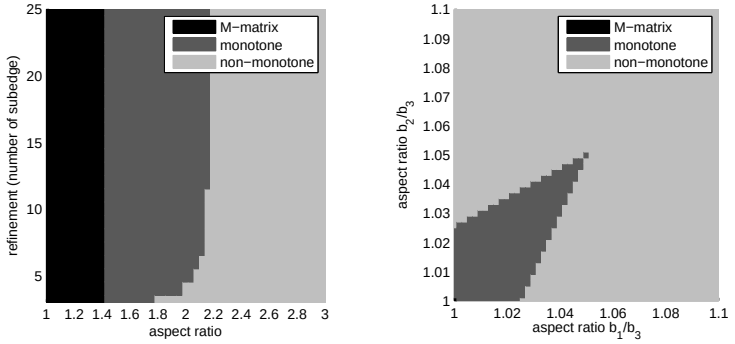


Fig. 2 The influence of the aspect ratio to the properties of the stiffness matrix $\mathbf{A}$. Left: Ω is a rectangle $(0,b_1)\times(0,b_2)$. Right: Ω is a rectangular cuboid $(0,b_1)\times(0,b_2)\times(0,b_3)$.

if the stiffness matrix is required to be monotone only. In this experiment, we again consider $c\equiv 0$ and $g=0$. The domain is a rectangle $\Omega=(0,b_1)\times(0,b_2)$. The finite element mesh is obtained by the uniform refinement of Ω into N_{sub}^2 elements, where N_{sub} is the number of subedges induced on each edge of Ω. The axes in Fig. 2 (left) correspond to the aspect ratio b_1/b_2 of the rectangle Ω (and of all elements) and to the value N_{sub}. The results in Fig. 2 (left) indicate that the value $\sqrt{2}$ in the nonnarrow condition can be increased up to about 2.16 provided the mesh is sufficiently fine.

The 3D analysis of the trilinear elements on rectangular cuboids based on (6) gives a bit pessimistic conclusion. The stiffness matrix is M-matrix (and the DMP is satisfied) if all the elements are cubes [9]. Similar experiment as before, see Fig. 2 (right), indicates that the cubes cannot be distorted much in order to retain the stiffness matrix monotone and to satisfy the DMP. The two possible aspect ratios we have in rectangular cuboids can be at most around 1.05.

In dimensions 4 and higher, certain contributions form the local stiffness matrices are always positive. Indeed, without loss of generality we may assume that $b_1\geq b_2\geq\cdots\geq b_d$. If a_{ij}^K was nonpositive then (6) would yield

$$\frac{1}{b_1^2}\geq\sum_{k=2}^{d}\frac{1}{2b_k^2}\geq\frac{d-1}{2b_2^2}>\frac{1}{b_2^2},$$

where the last inequality holds true for $d\geq 4$. This inequality, however, contradicts the fact that $b_1\geq b_2$. Furthermore, considering the longest edge in the mesh, we see that all the contributions from all the elements surrounding this edge are positive and, hence, the corresponding off-diagonal entry in the stiffness matrix $\mathbf{A}$ is positive. Consequently, $\mathbf{A}$ is not an M-matrix. Similar experiments as before reveal that the stiffness matrix is neither monotone even on hyper-cubes. Thus, from the point of the DMP, the block finite elements are less advantageous than the simplicial elements especially for 3D and higher dimensional problems.

4 On mesh generation and adaptivity

Modern FE computations require treatment of issues like generation of a mesh with desired geometric properties and its global and local refinements preserving those properties. In the following two subsections we shall discuss these issues for both, simplices and blocks, with respect to geometric limitations imposed by the DMP.

4.1 Simplicial FE meshes (acuteness and nonobtuseness)

The practical realization of angle conditions (nonobtuseness and acuteness) is not easy. Even in 2D, an initial generation of reasonable nonobtuse and acute triangulations, especially for complicated domains, is algorithmically a hard task, see e.g. [3] for examples and literature on the subject. In 3D it is becoming even more difficult. Some results on generation and proper refinements of nonobtuse tetrahedral meshes are reported e.g. in [11] (see also [3]). But the only known positive (and very recent results) on acute meshes are the acute face-to-face tetrahedralization of the whole 3D Euclidean space [16], an infinite slab [6], some types of tetrahedra and a regular octahedron [10], and a cube [10, 17]. It is worth to mention that the last two works (the only relevant for real-life computations which are mostly done in bounded domains) are published just in summer of 2009 ! Moreover, very many acute tetrahedra are required to fill the cube by their constructions. In addition, the generated tetrahedra are very densely placed in the interior of the cube which is not so good for real computations as meshes used in practice should be dense mainly in vertices and along edges. Concerning higher dimensions, the situation with acute simplices is getting even more pessimistic. For example, it was shown in [10, 13] that the space $\mathbb{R}^d$ $(d \geq 4)$ cannot (surprisingly !) be filled face-to-face by acute simplices at all, which means that, in general, it is not possible to generate (reasonable fine) acute simplicial meshes for most of domains in higher dimensions, even for such simple as hypercubes.

In order to get more accurate FE approximations one needs to make various (global and local refinements) of the meshes preserving the desired geometric properties. For example, a triangle can be split into four similar triangles using midlines (2D red refinement) (and thus acuteness or nonobtuseness are preserved), but a tetrahedron cannot be, in general, partitioned face-to-face into several similar tetrahedrons by similar technique. After cutting four vertices of the tetrahedron off (and thus producing four similar tetrahedra), an interior octahedron remains, which can be split into four tetrahedra in three different ways. And in most of cases the resulting tetrahedra are not similar to the original one, moreover, the acuteness property cannot be preserved in any case. In addition, all further refinements should be done with a special care in order to avoid producing degenerating subtetrahedra, see [19] for details. An alternative can be to use one of bisection algorithms, see e.g. [14] and references therein, but just bisecting as such cannot obviously produce acute angles.

As far it concerns local refinements, the only results in dimension 3 and higher are known for nonobtuse simplicial partitions, see [1].

4.2 Block FE meshes (preserving the aspect ratio)

In the case of block elements global refinement is obvious. Further, one can perform local refinements with or without hanging nodes [15]. However, local refinements without hanging nodes require forced refinements far from the targeted area and, moreover, elements with high aspect ratios are actually forming. Hanging nodes are practically more demanding to use, but they overcome these difficulties. The advantage is that the resulting meshes are nested and that the aspect ratio of subelements remains unchanged. Let us remark that the sufficient geometric conditions for the DMP are the same for meshes both with and without hanging nodes.

5 Conclusions

In 2D both triangular and rectangular meshes seem to be comparable in the sense that generation and refinement of meshes yielding the DMP is well treatable in both cases. Anyway, the triangles provide more flexibility for complicated domains (e.g. for those having non-right corners). In higher dimension, block elements can be recommended if the geometry of the domain allows them and if the DMP is not an issue. In the opposite case, the simplices should be used, but then we face the above described problems with mesh generation and local refinements constrained by the dihedral angle conditions. These problems are sometimes treatable by path-simplicial meshes, which guarantee the DMP at least for the Poisson problems. In addition, the practical implementation of simplicial meshes is technically more demanding than the implementation of the blocks. This fact must be weighted as well. Let us remark that it is geometrically advantageous to use simplices and blocks together in the hybrid meshes. However, from the point of the DMP the hybrid meshes inherit the discussed disadvantages of all used types of elements. Moreover, the practical implementation of hybrid meshes is technically very demanding. For example, a 3D hybrid mesh with tetrahedra and rectangular cuboids requires also right triangular prisms and pyramids to join the elements face-to-face [18]. The DMP on prismatic meshes has been analyzed in [8]. However, up to the authors' knowledge the DMP for pyramidal elements (and therefore on hybrid 3D meshes) has not been analyzed yet.

Finally, it is interesting to mention that angle and aspect ratio conditions similar to those we discussed above also appear in the analysis of the convergence of FE approximations [5].

Acknowledgements The first author has been supported by Project no. 124619 from the Academy of Finland. The second author has been supported by Grant no. IAA100760702 of the Grant Agency of the Czech Academy of Sciences, Grant no. 102/07/0496 of the Czech Science Foundation, and by the institutional research plan no. AV0Z10190503 of the Czech Academy of Sciences.

References

1. Brandts, J., Korotov, S., Křížek, M.: Dissection of the path-simplex in $\mathbf{R}^n$ into n path-subsimplices. Linear Algebra Appl. **421**, 382–393 (2007)
2. Brandts, J., Korotov, S., Křížek, M.: The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem. Linear Algebra Appl. **429**, 2344–2357 (2008)
3. Brandts, J., Korotov, S., Křížek, M., Šolc, J.: On nonobtuse simplicial partitions. SIAM Rev. **51**, 317–335 (2009)
4. Ciarlet, P.G.: Discrete maximum principle for finite-difference operators. Aequationes Math. **4**, 338–352 (1970)
5. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978)
6. Eppstein, D., Sullivan, J. M., Üngör, A.: Tiling space and slabs with acute tetrahedra. Comput. Geom.: Theory and Appl. **27**, 237–255 (2004)
7. Fiedler, M.: Special Matrices and Their Applications in Numerical Mathematics. Martinus Nijhoff Publishers, Dordrecht (1986)
8. Hannukainen, A., Korotov, S., Vejchodský, T.: Discrete maximum principle for FE-solutions of the diffusion-reaction problem on prismatic meshes. J. Comput. Appl. Math. **226**, 275–287 (2009)
9. Karátson, J., Korotov, S., Křížek, M.: On discrete maximum principles for nonlinear elliptic problems. Math. Comput. Simulation **76**, 99–108 (2007)
10. Kopczyński, E., Pak, I., Przytycki, P.: Acute triangulations of polyhedra and $\mathbf{R}^n$. arXiv:0909.3706 (2009)
11. Korotov, S., Křížek, M.: Acute type refinements of tetrahedral partitions of polyhedral domains. SIAM J. Numer. Anal. **39**, 724–733 (2001)
12. Korotov, S., Křížek, M., Neittaanmäki, P.: Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. Math. Comp. **70**, 107–119 (2001)
13. Křížek, M.: There is no face-to-face partition of R^5 into acute simplices. Discrete Comput. Geom. **36**, 381–390 (2006)
14. Rivara M.-C.: Lepp-bisection algorithms, applications and mathematical properties. Appl. Numer. Math. **59**, 2218–2235 (2009)
15. Šolín, P., Červený, J., Doležel, I.: Arbitrary-level hanging nodes and automatic adaptivity in the *hp*-FEM. Math. Comput. Simulation **77**, 117–132 (2008)
16. Üngör, A.: Tiling 3D Euclidean space with acute tetrahedra. In: Proc. Canadian Conf. Comput. Geom., Waterloo, 169–172 (2001)
17. VanderZee, E., Hirani, A. N., Zharnitsky, V., Guoy, D.: A dihedral acute triangulation of the cube. arXiv:0905.3715 (2009)
18. Wieners, C.: Conforming discretizations on tetrahedrons, pyramids, prisms and hexahedrons. Univ. Stuttgart, Bericht 97/15, 1–9 (1997)
19. Zhang, S.: Successive subdivisions of tetrahedra and multigrid methods on tetrahedral meshes. Houston J. Math. **21**, 541–556 (1995)

APPENDIX

D

Discrete maximum principle for higher-order finite elements in 1D

Below we attach a copy of the paper

[A3] T. Vejchodský and P. Šolín: Discrete maximum principle for higher-order finite elements in 1D. *Math. Comp.* **76** (2007), 1833–1846.

MATHEMATICS OF COMPUTATION
Volume 76, Number 260, October 2007, Pages 1833–1846
S 0025-5718(07)02022-4
Article electronically published on April 30, 2007

DISCRETE MAXIMUM PRINCIPLE FOR HIGHER-ORDER FINITE ELEMENTS IN 1D

TOMÁŠ VEJCHODSKÝ AND PAVEL ŠOLÍN

Abstract. We formulate a sufficient condition on the mesh under which we prove the discrete maximum principle (DMP) for the one-dimensional Poisson equation with Dirichlet boundary conditions discretized by the *hp*-FEM. The DMP holds if a relative length of every element K in the mesh is bounded by a value $H^*_{\text{rel}}(p) \in [0.9, 1]$, where $p \geq 1$ is the polynomial degree of the element K. The values $H^*_{\text{rel}}(p)$ are calculated for $1 \leq p \leq 100$.

1. Introduction

Classical (continuous) maximum principles belong to the most important results in the theory of second-order partial differential equations (PDEs). Their discrete counterparts, discrete maximum principles (DMP), appeared in the early 1970s. They were used by various authors to prove the convergence of the lowest-order finite difference and finite element methods (see, e.g., [3, 4] and the references therein). DMP have been studied intensively during the past decades in the context of linear PDEs [2, 8, 10, 17, 18, 20] and more recently also nonlinear equations [9]. Most of these results have two points in common:

- they are limited to lowest-order approximations,
- they are based on M-matrices [6, 16].

Much less is known about the DMP for methods of higher orders of accuracy such as higher-order finite difference methods, spectral FEM, or *hp*-FEM. Let us mention, e.g., a result [21] on higher-order collocation methods. Particularly noteworthy is a negative result [7] from 1981 stating that a stronger DMP is not valid for cubic and higher-order Lagrange elements in 2D. In the quadratic case, the stronger DMP is valid under extremely restrictive assumptions on the mesh, which almost never could be satisfied in practice. In light of this negative result, a few attempts were made to formulate and prove weakened forms of the DMP (see, e.g., [11, 14]). The present result is based on the analysis of the discrete Green's function (DGF) for higher-order elements. A similar concept was used in the piecewise-linear case in [5].

The paper is organized as follows. In Section 2 we introduce the one-dimensional Poisson problem, its *hp*-FEM discretization, and the discrete maximum principle.

Received by the editor January 31, 2006 and, in revised form, July 25, 2006.
2000 *Mathematics Subject Classification.* Primary 65N30; Secondary 35B50.
Key words and phrases. Discrete maximum principle, discrete Green's function, higher-order elements, *hp*-FEM, Poisson equation.

The discrete Green's function along with its basic properties is discussed in Section 3. In Section 4 we derive an explicit formula for the DGF for the Poisson problem discretized by *hp*-FEM, which is used to find sufficient conditions for its nonnegativity in Section 5. This leads to the notion of critical relative element length H^*_{rel}. The main result is presented in Section 6.

2. Model problem and its discretization

We consider the one-dimensional Poisson equation with homogeneous Dirichlet boundary conditions in an open bounded interval $\Omega = (\alpha, \beta)$. The standard weak formulation reads: Find $u \in V = H^1_0(\Omega)$ such that

$$a(u, v) = (f, v) \quad \forall v \in V, \tag{2.1}$$

where $f \in L^2(\Omega)$, the symbol $(\cdot, \cdot)$ stands for the inner product in $L^2(\Omega)$, $H^1_0(\Omega)$ is the standard Sobolev space, and $a(u, v) = (u', v')$.

We create a partition $\alpha = x_0 < x_1 < \ldots < x_M = \beta$ of the domain Ω consisting of M elements $K_i = [x_{i-1}, x_i]$, $i = 1, 2, \ldots, M$. Every element K_i is assigned an arbitrary polynomial degree $p_i \geq 1$. The corresponding finite element space of piecewise-polynomial continuous functions $V_{hp} \subset V$ has the form

$$V_{hp} = \{v_{hp} \in V;\ v_{hp}|_{K_i} \in P^{p_i}(K_i),\ i = 1, 2, \ldots, M\},$$

where $P^{p_i}(K_i)$ stands for the space of polynomials of degree at most p_i on the element K_i. The space V_{hp} has the dimension $N = -1 + \sum_{i=1}^{M} p_i$. There exists a unique function $u_{hp} \in V_{hp}$ satisfying

$$a(u_{hp}, v_{hp}) = (f, v_{hp}) \quad \forall v_{hp} \in V_{hp}. \tag{2.2}$$

Definition 2.1. We say that problem (2.2) satisfies the *discrete maximum principle* (DMP) if for any right-hand side $f \in L^2(\Omega)$ it holds that

$$f \geq 0 \text{ a.e. in } \Omega \quad \Rightarrow \quad u_{hp} \geq 0 \text{ in } \Omega.$$

Remark 2.2. The above implication is equivalent to

$$f \geq 0 \text{ a.e. in } \Omega \quad \Rightarrow \quad \min_{x \in \overline{\Omega}} u_{hp}(x) = \min_{x \in \partial\Omega} u_{hp}(x)$$

for homogeneous Dirichlet boundary conditions. This is further equivalent to

$$f \leq 0 \text{ a.e. in } \Omega \quad \Rightarrow \quad \max_{x \in \overline{\Omega}} u_{hp}(x) = \max_{x \in \partial\Omega} u_{hp}(x).$$

Remark 2.3. In problem (2.2), homogeneous Dirichlet conditions are considered without loss of generality. This follows immediately from the fact that every solution $\hat{u}_{hp}$ to a problem with nonhomogeneous Dirichlet boundary conditions can be written as $\hat{u}_{hp} = u^L_{hp} + u_{hp}$, where u^L_{hp} is a linear function satisfying the nonhomogeneous conditions and u_{hp} vanishes at Ω-endpoints.

3. Discrete Green's function

The discrete Green's function (DGF) is defined in analogy with the standard (continuous) Green's function:

Definition 3.1. For an arbitrary $z \in \Omega$, the unique solution $G_{hp,z} \in V_{hp}$ to the problem

$$a(v_{hp}, G_{hp,z}) = v_{hp}(z) \quad \forall v_{hp} \in V_{hp} \tag{3.1}$$

is called the *discrete Green's function* (DGF) corresponding to the point z.

In the following, we will use the notation $G_{hp}(x,z) = G_{hp,z}(x)$. A combination of (2.2) and (3.1) yields an important consequence:

$$u_{hp}(z) = \int_\Omega G_{hp}(x,z) f(x)\,\mathrm{d}x \quad \forall z \in \Omega. \tag{3.2}$$

The following lemma shows that the DGF can easily be expressed using any basis of V_{hp}; cf. [5]. We use the Kronecker symbol

$$\delta_{ik} = \begin{cases} 1 & \text{for } i = k, \\ 0 & \text{for } i \neq k. \end{cases}$$

Lemma 3.2. *Let $\{\varphi_1, \varphi_2, \dots, \varphi_N\}$ be any basis of V_{hp}. If the stiffness matrix $A_{ij} = a(\varphi_j, \varphi_i)$, $1 \le i, j \le N$, is nonsingular, then*

$$G_{hp}(x,z) = \sum_{j=1}^{N} \sum_{k=1}^{N} A^{-1}_{jk} \varphi_k(x) \varphi_j(z). \tag{3.3}$$

Here A^{-1}_{jk} are the entries of the inverse stiffness matrix, i.e., $\sum\limits_{j=1}^{N} A_{ij} A^{-1}_{jk} = \delta_{ik}$, $1 \le i, k \le N$.

Proof. Substitute

$$G_{hp}(x,z) = \sum_{i=1}^{N} c_i(z) \varphi_i(x) \tag{3.4}$$

into (3.1) with $v_{hp} = \varphi_j$. It follows that

$$\sum_{i=1}^{N} c_i(z) \underbrace{a(\varphi_j, \varphi_i)}_{A_{ij}} = \varphi_j(z).$$

The coefficients $c_i(z)$ can be expressed in terms of the inverse matrix as $c_k(z) = \sum_{j=1}^{N} \varphi_j(z) A^{-1}_{jk}$, and they can be substituted back into (3.4). □

Corollary 3.3. *Let $\{l_1, l_2, \cdots, l_N\}$ be a basis of V_{hp} such that $a(l_i, l_j) = \delta_{ij}$. Then*

$$G_{hp}(x,z) = \sum_{i=1}^{N} l_i(x) l_i(z).$$

Lemma 3.4. *If there exists a basis $\{l_1, l_2, \dots, l_N\}$ of V_{hp} such that $a(l_i, l_j) = \delta_{ij}$, $1 \le i, j \le N$, then $G_{hp}(x,x) > 0$ for all $x \in \Omega$.*

Proof. Let $x \in \Omega$. Since $\{l_1, l_2, \dots, l_N\}$ is a basis, there exists at least one $k \in \{1, 2, \dots, N\}$ such that $l_k(x) \neq 0$. Hence, by Corollary 3.3

$$G_{hp}(x,x) = \sum_{i=1}^{N} l_i^2(x) > 0.$$ □

Theorem 3.5. *Problem* (2.2) *satisfies the discrete maximum principle if and only if the corresponding discrete Green's function $G_{hp}(x,z) = G_{hp,z}(x)$ defined by* (3.1) *is nonnegative in Ω^2.*

Proof. Immediate consequence of (3.2). □

Remark 3.6. Results presented in this section are valid for any second-order elliptic problem of the form (2.1) as well as in higher spatial dimensions.

4. DGF for Poisson problem in 1D

4.1. **Lowest-order case.** Consider the case $p_1 = p_2 = \ldots = p_M = 1$ first. Let $\mathcal{B}^L = \{\phi_1, \phi_2, \ldots, \phi_{M-1}\}$ be the standard lowest-order basis consisting of the piecewise-linear "hat functions" such that $\phi_j(x_i) = \delta_{ij}$, $1 \le i, j \le M-1$. In this case the stiffness matrix $A^L \in \mathbb{R}^{(M-1)\times(M-1)}$ is tridiagonal,

$$A^L_{ij} = \begin{cases} 1/h_i + 1/h_{i+1} & \text{for } i = j, \\ -1/h_{i+1} & \text{for } i = j-1, \\ -1/h_{i-1} & \text{for } i = j+1, \\ 0 & \text{otherwise,} \end{cases}$$

with $h_i = x_i - x_{i-1}$.

Lemma 4.1. *The inverse matrix $(A^L)^{-1} \in \mathbb{R}^{(M-1)\times(M-1)}$ has the form*

$$(A^L)^{-1} = \frac{1}{\beta-\alpha} \begin{pmatrix} (x_1-\alpha)(\beta-x_1) & (x_1-\alpha)(\beta-x_2) & (x_1-\alpha)(\beta-x_3) & \ldots \\ (x_1-\alpha)(\beta-x_2) & (x_2-\alpha)(\beta-x_2) & (x_2-\alpha)(\beta-x_3) & \ldots \\ (x_1-\alpha)(\beta-x_3) & (x_2-\alpha)(\beta-x_3) & (x_3-\alpha)(\beta-x_3) & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

i.e., $(A^L)^{-1}_{ij} = (x_i-\alpha)(\beta-x_j)/(\beta-\alpha)$ *for* $1 \le i \le j \le M-1$ *and* $(A^L)^{-1}_{ij} = (x_j-\alpha)(\beta-x_i)/(\beta-\alpha)$ *for* $1 \le j < i \le M-1$.

Proof [1]. We need to show that $z_{ij} = \delta_{ij}$, where

$$z_{ij} = \sum_{k=1}^{M-1} (A^L)^{-1}_{ik} A^L_{kj} = \sum_{k=1}^{M-1} (A^L)^{-1}_{ik} a(\phi_j, \phi_k),$$

for all $i, j = 1, 2, \ldots, M-1$. Let us fix i and j and consider the bilinear forms

$$a_1(u,v) = \int_\alpha^{x_i} u'v' \,\mathrm{d}x \quad \text{and} \quad a_2(u,v) = \int_{x_i}^{\beta} u'v' \,\mathrm{d}x.$$

The explicit formulae for $(A^L)^{-1}_{ik}$ yield

$$(\beta-\alpha) z_{ij} = (\beta-x_i) a\Big(\phi_j, \sum_{k=1}^{i-1} (x_k-\alpha)\phi_k\Big) + (x_i-\alpha)(\beta-x_i) a(\phi_j,\phi_i) + (x_i-\alpha) a\Big(\phi_j, \sum_{k=i+1}^{M-1} (\beta-x_k)\phi_k\Big).$$

Now, we split the term $a(\phi_j,\phi_i) = a_1(\phi_j,\phi_i) + a_2(\phi_j,\phi_i)$ to obtain

$$(\beta-\alpha) z_{ij} = (\beta-x_i) a_1(\phi_j, x-\alpha) + (x_i-\alpha) a_2(\phi_j, \beta-x).$$

The fact that $a_1(\phi_j, \beta-x) = a_2(\phi_j, x-\alpha) = \delta_{ij}$ finishes the proof. □

[1] The authors thank an anonymous referee for simplifying their original proof.

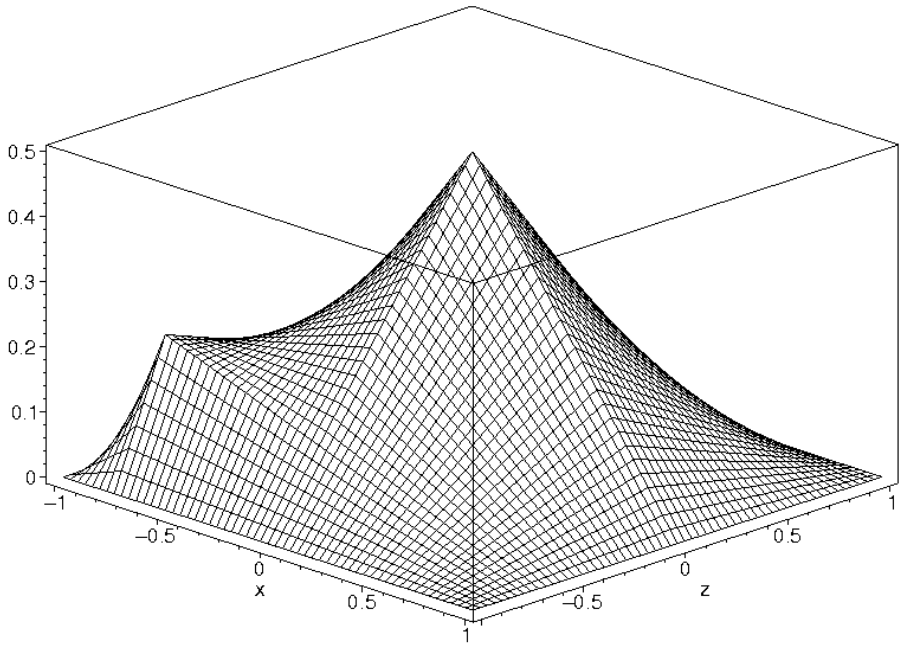


FIGURE 1. The lowest-order part $G^L_{hp}(x,z)$ of the discrete Green's function $G_{hp}(x,z)$ for the Poisson equation in $\Omega = (-1,1)$, on a mesh with three elements $[-1,-3/4]$, $[-3/4,0]$, and $[0,1]$.

Using Lemma 4.1 and identity (3.3), we can write the DGF in the form

$$G^L_{hp}(x,z) = \frac{1}{\beta-\alpha}\left(\sum_{i=1}^{M-1}(x_i-\alpha)(\beta-x_i)\phi_i(x)\phi_i(z) + \sum_{i=1}^{M-2}\sum_{j=i+1}^{M-1}(x_i-\alpha)(\beta-x_j)[\phi_i(x)\phi_j(z)+\phi_j(x)\phi_i(z)]\right). \tag{4.1}$$

In particular, we see immediately that

$$G^L_{hp}(x,z) \geq 0 \quad \forall [x,z] \in \Omega^2. \tag{4.2}$$

The situation is illustrated in Figure 1.

4.2. **Higher-order case.** In this paragraph we return to the original setting with arbitrary polynomial degrees $p_i \geq 1$. In order to facilitate the construction of higher-order basis functions of the space V_{hp}, let us introduce the Lobatto shape functions $l_0, l_1, l_2, \ldots$ on a reference interval $\hat{K} = [-1,1]$ (see, e.g., [12, 15]).

The lowest-order Lobatto shape functions l_0 and l_1 have the form $l_0(\xi) = (1-\xi)/2$, $l_1(\xi) = (1+\xi)/2$, $\xi \in \hat{K}$. The higher-order shape functions $l_2, l_3, \ldots$ are defined as antiderivatives to the Legendre polynomials. Therefore, they satisfy

$$\int_{-1}^{1} l_i'(\xi) l_j'(\xi)\,\mathrm{d}\xi = \delta_{ij}, \quad i,j = 2,3,\ldots.$$

Every Lobatto shape function l_i, $i = 2, 3, \ldots$, is a polynomial of degree i and it vanishes at ± 1. Thus it can be expressed as

$$l_i(\xi) = l_0(\xi) l_1(\xi) \kappa_i(\xi), \quad i = 2, 3, \ldots,$$

where κ_i is a polynomial of degree $i-2$. For reference, the first few kernels κ_i are listed in Appendix.

The basis $\mathcal{B} = \{\phi_1, \phi_2, \ldots, \phi_N\}$ of V_{hp} can be written as $\mathcal{B} = \mathcal{B}^L \cup \mathcal{B}^B$, where $\mathcal{B}^L$ was defined above and $\mathcal{B}^B$ is the higher-order part of the basis comprising functions $\phi_M, \phi_{M+1}, \ldots, \phi_N$. These are defined as follows.

Consider the standard linear transformations from $\hat{K}$ to K_i,

$$\chi_{K_i}(\xi) = \frac{(x_i - x_{i-1})\xi + (x_i + x_{i-1})}{2}. \tag{4.3}$$

On an element K_i of the polynomial degree p_i, there are $p_i - 1$ higher-order basis functions. These vanish outside of K_i and in K_i they are defined as the Lobatto shape functions $l_2, l_3, \ldots, l_{p_i}$ composed with the inverse map $\chi_{K_i}^{-1}(x)$.

Proposition 4.2. *We have the following orthogonality relations:*

$$a(\phi^L, \phi^B) = 0 \quad \forall \phi^L \in \mathcal{B}^L,\ \forall \phi^B \in \mathcal{B}^B,$$
$$a(\phi^B, \psi^B) = 0 \quad \forall \phi^B \in \mathcal{B}^B,\ \forall \psi^B \in \mathcal{B}^B,\ \phi^B \neq \psi^B.$$

Proof. The proof is straightforward, based on the L^2-orthogonality of the Legendre polynomials. □

By Proposition 4.2, both the stiffness matrix A and its inverse have the following block structure:

$$A = \begin{pmatrix} A^L & 0 \\ 0 & D \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} (A^L)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix}$$

with

$$D = \operatorname{diag}\Big(\underbrace{\frac{2}{h_1}, \ldots, \frac{2}{h_1}}_{(p_1-1)\text{ times}}, \underbrace{\frac{2}{h_2}, \ldots, \frac{2}{h_2}}_{(p_2-1)\text{ times}}, \ \cdots\ , \underbrace{\frac{2}{h_M}, \ldots, \frac{2}{h_M}}_{(p_M-1)\text{ times}} \Big). \tag{4.4}$$

By (3.3), the DGF can be written as

$$G_{hp}(x, z) = G_{hp}^L(x, z) + G_{hp}^B(x, z), \tag{4.5}$$

where $G_{hp}^L(x, z)$ corresponds to (4.1) and

$$G_{hp}^B(x, z) = \sum_{k=M}^{N} D_{kk}^{-1} \phi_k(x) \phi_k(z) \quad \forall [x, z] \in \Omega^2. \tag{4.6}$$

Unfortunately, $G_{hp}^B(x, z)$ defined by (4.6) is not nonnegative in the entire Ω^2 in general. For instance, in the example shown in Figure 2, there are small regions near the points $[1, 0]$ and $[0, 1]$, where the function $G_{hp}^B(x, z)$ is negative.

Notice that any partition of Ω produces a rectangular grid on Ω^2 and that $G_{hp}^B(x, z)$ can be nonzero within the diagonal squares of this grid only. In other words,

$$\operatorname{supp} G_{hp}^B \subset \bigcup_{i=1}^{M} K_i^2. \tag{4.7}$$

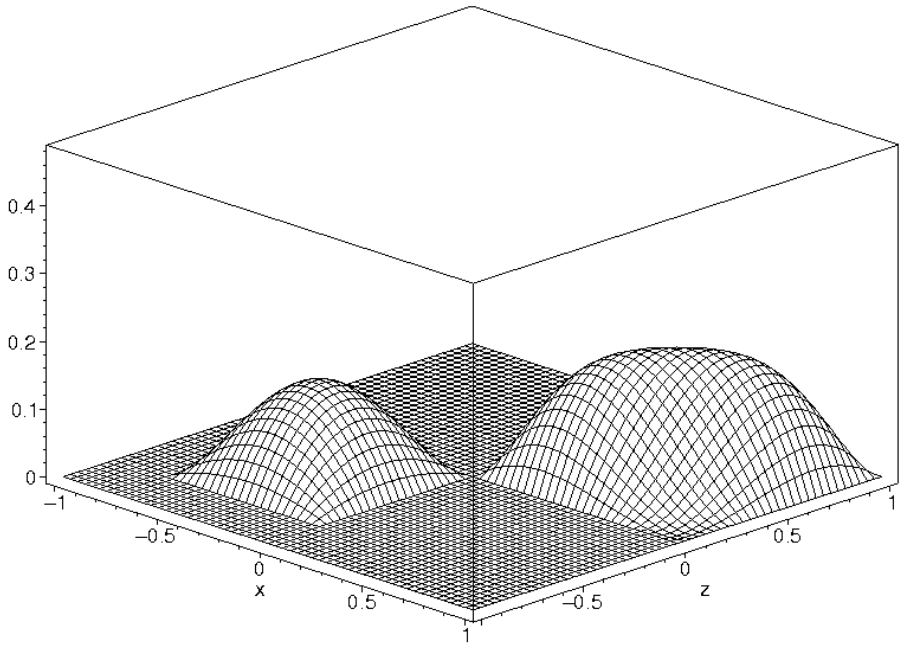


FIGURE 2. The higher-order part $G^B_{hp}(x,z)$ of the discrete Green's function $G_{hp}(x,z)$ for the Poisson equation in $\Omega = (-1,1)$, on a mesh with three elements $[-1,-3/4]$, $[-3/4,0]$, and $[0,1]$ of the polynomial degrees $p_1 = 1$, $p_2 = 2$, $p_3 = 3$.

Lemma 4.3. *The discrete Green's function G_{hp} defined by* (4.5) *is nonnegative in* $\Omega^2 \setminus \bigcup_{i=1}^M K_i^2$.

Proof. Consider (4.7) together with (4.2). □

5. The DGF on K_i^2

As justified by Lemma 4.3, we only need to continue with the study of the discrete Green's function $G_{hp}(x,z)$ in the union of the diagonal squares $\bigcup_{i=1}^M K_i^2$. Without loss of generality, let us restrict ourselves to only one square K_i^2, $1 \le i \le M$. Let $p = p_i$ be the polynomial degree assigned to K_i. Notice that only a few terms in (4.1) and (4.6) are nonzero in K_i^2. Hence, by (4.1), (4.4), and (4.6) we obtain

$$\begin{aligned} G_{hp}(x,z)\big|_{K_i^2} &= \frac{(x_i-\alpha)(\beta-x_i)}{\beta-\alpha}\phi_i(x)\phi_i(z) \\ &\quad + \frac{(x_{i-1}-\alpha)(\beta-x_{i-1})}{\beta-\alpha}\phi_{i-1}(x)\phi_{i-1}(z) \\ &\quad + \frac{(x_{i-1}-\alpha)(\beta-x_i)}{\beta-\alpha}\left[\phi_i(x)\phi_{i-1}(z)+\phi_{i-1}(x)\phi_i(z)\right] \\ &\quad + \frac{x_i-x_{i-1}}{2}G^B_{hp}(x,z), \end{aligned} \tag{5.1}$$

$[x,z] \in K_i^2$, $1 \le i \le M$. It is convenient to introduce the notation $K_i = [x_{i-1}, x_i] = [L,R]$.

We transform the function G_{hp} from K_i^2 to the reference square $\hat{K}^2 = [-1,1]^2$ using the linear transformation (4.3) with $x = \chi_{K_i}(\xi)$ and $z = \chi_{K_i}(\eta)$,

$$\begin{aligned} G_{hp}(x,z)\big|_{K_i^2} = \hat{G}_{hp}(\xi,\eta) &= \frac{(R-\alpha)(\beta-R)}{\beta-\alpha} l_1(\xi) l_1(\eta) \\ &+ \frac{(L-\alpha)(\beta-L)}{\beta-\alpha} l_0(\xi) l_0(\eta) \\ &+ \frac{(L-\alpha)(\beta-R)}{\beta-\alpha} \left[l_1(\xi) l_0(\eta) + l_0(\xi) l_1(\eta) \right] \\ &+ \frac{R-L}{2} \hat{G}_{hp}^{p,B}(\xi,\eta), \end{aligned} \tag{5.2}$$

$[\xi,\eta] \in \hat{K}^2$. Here $l_0(\xi)$ and $l_1(\xi)$ are the above-defined lowest-order shape functions on $\hat{K}$ and

$$\hat{G}_{hp}^{p,B}(\xi,\eta) = \sum_{k=2}^{p} l_k(\xi) l_k(\eta) = l_0(\xi) l_0(\eta) l_1(\xi) l_1(\eta) \sum_{k=2}^{p} \kappa_k(\xi) \kappa_k(\eta) \tag{5.3}$$

is the higher-order part.

Let us modify formula (5.2) in the following way: Divide (5.2) by $R-L > 0$ and use the identities

$$\begin{aligned} \frac{(L-\alpha)(\beta-L)}{(\beta-\alpha)(R-L)} &= \frac{(L-\alpha)(\beta-R)}{(\beta-\alpha)(R-L)} + \frac{L-\alpha}{\beta-\alpha}, \\ \frac{(R-\alpha)(\beta-R)}{(\beta-\alpha)(R-L)} &= \frac{(L-\alpha)(\beta-R)}{(\beta-\alpha)(R-L)} + \frac{\beta-R}{\beta-\alpha}, \end{aligned}$$

and

$$l_0(\xi) l_0(\eta) + l_1(\xi) l_1(\eta) + l_0(\xi) l_1(\eta) + l_1(\xi) l_0(\eta) = 1 \quad \forall [\xi,\eta] \in \hat{K}^2.$$

We obtain

$$\begin{aligned} \frac{\hat{G}_{hp}(\xi,\eta)}{R-L} &= \frac{(L-\alpha)(\beta-R)}{(\beta-\alpha)(R-L)} + \frac{L-\alpha}{\beta-\alpha} l_0(\xi) l_0(\eta) \\ &+ \frac{\beta-R}{\beta-\alpha} l_1(\xi) l_1(\eta) + \frac{1}{2} \hat{G}_{hp}^{p,B}(\xi,\eta). \end{aligned} \tag{5.4}$$

The endpoints of K_i can be parameterized using the element length $H = R - L$ and a real parameter $0 \le t \le 1$, so that $L = \alpha$ for $t = 0$ and $R = \beta$ for $t = 1$:

$$L = (1-t)\alpha + t(\beta - H), \tag{5.5}$$

$$R = (1-t)(\alpha + H) + t\beta. \tag{5.6}$$

Use (5.5) and (5.6), define relative element length H_{rel} by

$$H_{\text{rel}} = \frac{H}{\beta - \alpha},$$

and compute

$$\frac{L-\alpha}{\beta-\alpha} = \frac{t(\beta-\alpha-H)}{\beta-\alpha} = t(1-H_{\text{rel}}), \tag{5.7}$$

$$\frac{\beta-R}{\beta-\alpha} = \frac{(1-t)(\beta-\alpha-H)}{\beta-\alpha} = (1-t)(1-H_{\text{rel}}), \tag{5.8}$$

$$\frac{(L-\alpha)(\beta-R)}{(\beta-\alpha)(R-L)} = \frac{t(1-t)(\beta-\alpha-H)^2}{(\beta-\alpha)H} = t(1-t)\frac{(1-H_{\text{rel}})^2}{H_{\text{rel}}}. \tag{5.9}$$

Substitute (5.7)–(5.9) into (5.4) to obtain

$$\frac{\hat{G}_{hp}(\xi,\eta)}{H} = t(1-t)\frac{(1-H_{\mathrm{rel}})^2}{H_{\mathrm{rel}}} + t(1-H_{\mathrm{rel}})l_0(\xi)l_0(\eta) + (1-t)(1-H_{\mathrm{rel}})l_1(\xi)l_1(\eta) + \frac{1}{2}\hat{G}_{hp}^{p,B}(\xi,\eta). \tag{5.10}$$

Finally, use the identity

$$\hat{G}_{hp}^{p,B}(\xi,\eta) = t\hat{G}_{hp}^{p,B}(\xi,\eta) + (1-t)\hat{G}_{hp}^{p,B}(\xi,\eta),$$

substitute (5.3) into (5.10), and factor out $l_0(\xi)l_0(\eta)$ and $l_1(\xi)l_1(\eta)$:

$$\begin{aligned}\frac{\hat{G}_{hp}(\xi,\eta)}{H} &= t(1-t)\frac{(1-H_{\mathrm{rel}})^2}{H_{\mathrm{rel}}} \\ &\quad + tl_0(\xi)l_0(\eta)\left[1-H_{\mathrm{rel}}+\frac{1}{2}l_1(\xi)l_1(\eta)\sum_{k=2}^{p}\kappa_k(\xi)\kappa_k(\eta)\right] \\ &\quad + (1-t)l_1(\xi)l_1(\eta)\left[1-H_{\mathrm{rel}}+\frac{1}{2}l_0(\xi)l_0(\eta)\sum_{k=2}^{p}\kappa_k(\xi)\kappa_k(\eta)\right].\end{aligned} \tag{5.11}$$

Indeed, the value $t(1-t)(1-H_{\mathrm{rel}})^2/H_{\mathrm{rel}}$ is nonnegative for all $t \in [0,1]$ as well as the values $tl_0(\xi)l_0(\eta)$ and $(1-t)l_1(\xi)l_1(\eta)$, for all $[\xi,\eta] \in \hat{K}^2$. Hence, the discrete Green's function G_{hp} is nonnegative in K_i^2 if both expressions in the square brackets in (5.11) are nonnegative. To see that they impose the same restriction on the relative element length H_{rel}, let us introduce Lemma 5.1:

Lemma 5.1. *It is true that*

$$\min_{[\xi,\eta]\in\hat{K}^2} l_0(\xi)l_0(\eta)\sum_{k=2}^{p}\kappa_k(\xi)\kappa_k(\eta) = \min_{[\xi,\eta]\in\hat{K}^2} l_1(\xi)l_1(\eta)\sum_{k=2}^{p}\kappa_k(\xi)\kappa_k(\eta).$$

Proof. Using the definition of the functions κ_i, it is easy to see that $\kappa_k(\xi) = \kappa_k(-\xi)$ for k even and $\kappa_k(\xi) = -\kappa_k(-\xi)$ for k odd. Therefore, $\kappa_k(\xi)\kappa_k(\eta) = \kappa_k(-\xi)\kappa_k(-\eta)$ for every $k = 2, 3, \ldots$. Moreover, $l_0(\xi) = l_1(-\xi)$, which yields

$$\begin{aligned}\min_{[\xi,\eta]\in\hat{K}^2} l_0(\xi)l_0(\eta)\sum_{k=2}^{p}\kappa_k(\xi)\kappa_k(\eta) &= \min_{[\xi,\eta]\in\hat{K}^2} l_1(-\xi)l_1(-\eta)\sum_{k=2}^{p}\kappa_k(-\xi)\kappa_k(-\eta) \\ &= \min_{[\xi,\eta]\in\hat{K}^2} l_1(\xi)l_1(\eta)\sum_{k=2}^{p}\kappa_k(\xi)\kappa_k(\eta).\end{aligned}$$

□

Relation (5.11) and Lemma 5.1 motivate the following definition:

Definition 5.2. By critical relative element length H_{rel}^* corresponding to a polynomial degree $p \geq 2$ we mean the value

$$\begin{aligned}H_{\mathrm{rel}}^*(p) &= 1 + \frac{1}{2}\min_{(\xi,\eta)\in\hat{K}^2} l_0(\xi)l_0(\eta)\sum_{k=2}^{p}\kappa_k(\xi)\kappa_k(\eta) \\ &= 1 + \frac{1}{2}\min_{(\xi,\eta)\in\hat{K}^2} l_1(\xi)l_1(\eta)\sum_{k=2}^{p}\kappa_k(\xi)\kappa_k(\eta).\end{aligned} \tag{5.12}$$

For $p = 1$ we define $H_{\mathrm{rel}}^* = 1$.

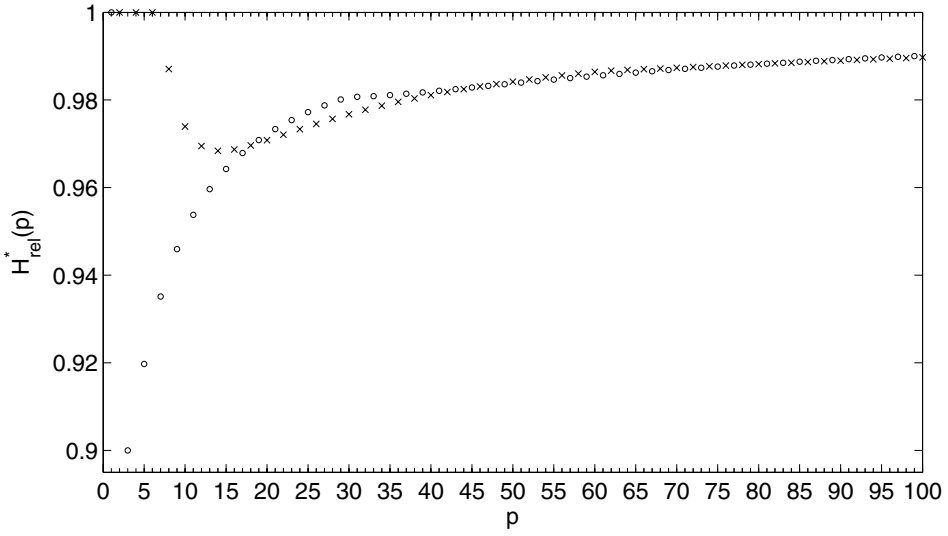


FIGURE 3. Critical relative element lengths $H^*_{\mathrm{rel}}(p)$ for $p = 1, 2, \ldots, 100$. Circles indicate the values for p odd and crosses indicate the value for p even.

Theorem 5.3. *If* $\alpha \le L < R \le \beta$ *and*

$$\frac{R-L}{\beta-\alpha} \le H^*_{\mathrm{rel}}(p), \tag{5.13}$$

then the function $\hat{G}_{hp}(\xi,\eta)$ *defined by* (5.2) *is nonnegative for all* $[\xi,\eta] \in \hat{K}^2 = [-1,1]^2$.

Proof. Apply (5.13) and the definition of $H^*_{\mathrm{rel}}(p)$ to infer

$$1 - H_{\mathrm{rel}} + \frac{1}{2} l_1(\xi) l_1(\eta) \sum_{k=2}^{p} \kappa_k(\xi)\kappa_k(\eta)$$
$$\ge 1 - H^*_{\mathrm{rel}}(p) + \frac{1}{2} l_1(\xi) l_1(\eta) \sum_{k=2}^{p} \kappa_k(\xi)\kappa_k(\eta) \ge 0 \quad \forall [\xi,\eta] \in \hat{K}^2.$$

Similarly,

$$1 - H_{\mathrm{rel}} + \frac{1}{2} l_0(\xi) l_0(\eta) \sum_{k=2}^{p} \kappa_k(\xi)\kappa_k(\eta) \ge 0 \quad \forall [\xi,\eta] \in \hat{K}^2.$$

Thus, all terms in (5.11) are nonnegative and we can conclude that

$$\hat{G}_{hp}(\xi,\eta) \ge 0 \quad \text{for all } [\xi,\eta] \in \hat{K}^2. \qquad \square$$

Computation of $H^*_{\mathrm{rel}}(p)$. In Table 1 we list the values of $H^*_{\mathrm{rel}}(p)$ for $p = 1, 2, \ldots, 20$.

The values of $H^*_{\mathrm{rel}}(p)$ for $p = 1, 2, \ldots, 100$ are plotted in Figure 3. While the values $H^*_{\mathrm{rel}}(p)$ for $p = 1, 2, 3, 4$ could be calculated analytically, results for $p \ge 5$ are numerical, obtained with high accuracy.

TABLE 1. Critical relative element length $H^*_{\text{rel}}(p)$ for $p = 1, 2, 3, \ldots, 20$.

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

6. Main results

Let us summarize the conclusions of the previous analysis:

Theorem 6.1. *If the partition $\alpha = x_0 < x_1 < \ldots < x_M = \beta$ of the domain $\Omega = (\alpha, \beta)$ satisfies the condition*

$$\frac{x_i - x_{i-1}}{\beta - \alpha} \le H^*_{\text{rel}}(p_i) \quad \text{for all } i = 1, 2, \ldots, M, \tag{6.1}$$

where $p_i \ge 1$ is the polynomial degree assigned to the element $K_i = [x_{i-1}, x_i]$, and $H^*_{\text{rel}}(p_i)$ is defined by (5.12)*, then the problem* (2.2) *satisfies the discrete maximum principle (i.e., $u_{hp} \ge 0$ in Ω for arbitrary $f \in L^2(\Omega)$ which is nonnegative a.e. in Ω).*

Proof. Let K_i be any element. By (5.2), condition (6.1), and Theorem 5.3 it holds that

$$G_{hp}(x, z)|_{K_i^2} = G_{hp}(\xi, \eta) \ge 0 \quad \text{for all } [x, z] \in K_i^2.$$

Thus, $G_{hp}(x, z) \ge 0$ in $\bigcup_{i=1}^M K_i^2$. Lemma 4.3 implies that $G_{hp}(x, z) \ge 0$ also in $\Omega^2 \setminus \bigcup_{i=1}^M K_i^2$. Theorem 3.5 finishes the proof. □

Table 1 indicates that the restriction on the relative element length $(x_i - x_{i-1})/(\beta - \alpha)$ is strongest in the cubic case where $H^*_{\text{rel}} = 9/10$. Moreover, Figure 3 shows a steadily growing trend in H^*_{rel} for $p \ge 50$. These observations motivate the following conjecture:

Conjecture 6.2. If the partition $\alpha = x_0 < x_1 < \ldots < x_M = \beta$ of the domain $\Omega = (\alpha, \beta)$ satisfies the condition

$$\frac{x_i - x_{i-1}}{\beta - \alpha} \le \frac{9}{10} \quad \text{for all } i = 1, 2, \ldots, M,$$

then problem (2.2) satisfies the discrete maximum principle (i.e., $u_{hp} \ge 0$ in Ω for arbitrary $f \in L^2(\Omega)$ which is nonnegative a.e. in Ω).

7. Possible generalizations

An analogous technique can be used to study problem (2.1) with mixed Dirichlet-Neumann boundary conditions. Of course, the structure of the stiffness matrix and the structure of the DGF are different, but analysis reveals that the quantity $H^*_{\text{rel}}(p)$ plays a central role again. Since $H^*_{\text{rel}}(p)$ is nonnegative in this case (at least for $p \le 100$), the DMP for problem (2.1) with mixed boundary conditions is valid with no restricting conditions on the mesh or polynomial degrees of elements. More details can be found in a recent report [19].

Generalization of these results to problems with variable coefficients and to higher-dimensional problems, however, will be more involved. In both of these cases, higher-order shape functions are no longer orthogonal, which yields a nontrivial cross term in the expression for the DGF. An analysis of this term will be crucial to achieve any progress in this direction. The goal of the analysis is to infer possibly simple conditions on the mesh and polynomial degrees of elements so that the DMP is valid. To achieve this goal, new techniques for the analysis of the DGF have to be developed.

The negative result from [7] does not imply that generalizations to 2D are impossible. This paper dealt with a stronger version of the DMP which required the maximum principle to be valid in all subdomains. Basically, the paper showed that the DMP for higher-order elements was not valid on vertex patches (patches of elements surrounding mesh vertices). It seems that vertex patches simply are too coarse for the DMP to be valid.

Another possibility would be to employ an idea from [1][2] to treat a class of 1D problems with a variable coefficient

$$-(\varrho(x)u')' = f, \quad u(\alpha) = u(\beta) = 0.$$

The idea would be to define new vertex functions to be piecewise-harmonic, such that each ϕ_i, $i = 1, 2, \dots, M-1$, solves

$$-(\varrho(x)\phi_i')' = 0 \quad \text{on } (x_{i-1}, x_i), \quad u(x_{i-1}) = 0, \qquad u(x_i) = 1, \tag{7.1}$$

$$-(\varrho(x)\phi_i')' = 0 \quad \text{on } (x_i, x_{i+1}), \qquad u(x_i) = 1, \quad u(x_{i+1}) = 0. \tag{7.2}$$

Such vertex functions, interestingly, would be orthogonal to bubble functions. However, the definition of the corresponding bubble functions and formulation of the condition for the DMP to be valid need further research.

Appendix

The Lobatto shape functions are defined by

$$l_j(\xi) = \sqrt{\frac{2j-1}{2}} \int_{-1}^{\xi} P_{j-1}(x)\,\mathrm{d}x, \quad j = 2, 3, \dots,$$

where $P_j(x) = \mathrm{d}^j/\mathrm{d}x^j (x^2-1)^j/(2^j j!)$ stands for the jth-degree Legendre polynomial. The kernels are defined by $\kappa_j(\xi) = l_j(\xi)/(l_0(\xi)l_1(\xi))$, where $l_0(\xi) = (1-\xi)/2$, $l_1(\xi) = (1+\xi)/2$, and $\xi \in [-1, 1]$. These kernels can be generated by the recurrence

$$\kappa_{j+2}(\xi) = \frac{\sqrt{2j+1}\sqrt{2j+3}}{j+2}\xi\kappa_{j+1}(\xi) - \frac{j-1}{j+2}\sqrt{\frac{2j+3}{2j-1}}\kappa_j(\xi), \quad j = 2, 3, \dots.$$

[2]We thank an anonymous referee for pointing this out.

For reference, we list several kernel functions κ_i (see, e.g., Section 3.1 in [15] or Section 1.2 in [13]):

$$\begin{aligned}
\kappa_2(\xi) &= -\sqrt{6},\\
\kappa_3(\xi) &= -\sqrt{10}\xi,\\
\kappa_4(\xi) &= -\frac{1}{4}\sqrt{14}(5\xi^2-1),\\
\kappa_5(\xi) &= -\frac{3}{4}\sqrt{2}(7\xi^2-3)\xi,\\
\kappa_6(\xi) &= -\frac{1}{8}\sqrt{22}(21\xi^4-14\xi^2+1),\\
\kappa_7(\xi) &= -\frac{1}{8}\sqrt{26}(33\xi^4-30\xi^2+5)\xi,\\
\kappa_8(\xi) &= -\frac{1}{64}\sqrt{30}(429\xi^6-495\xi^4+135\xi^2-5),\\
\kappa_9(\xi) &= -\frac{1}{64}\sqrt{34}(715\xi^6-1001\xi^4+385\xi^2-35)\xi,\\
\kappa_{10}(\xi) &= -\frac{1}{128}\sqrt{38}(2431\xi^8-4004\xi^6+2002\xi^4-308\xi^2+7).
\end{aligned}$$

Acknowledgments

The first author has been supported by the Grant Agency of the Czech Republic, project No. 201/04/P021 and by the Academy of Sciences of the Czech Republic, Institutional Research Plan No. AV0Z10190503. The second author has been supported in part by the U.S. Department of Defense under the Grant No. 05PR07548-00, by the NSF Grant No. DMS-0532645, and by the Grant Agency of the Czech Republic, project No. 102-05-0629. This support is gratefully acknowledged.

References

1. I. Babuška, G. Caloz, J. Osborn, Special finite element methods for a class of second order elliptic problems with rough coefficients, SIAM J. Numer. Anal., 31 (1994), pp. 945–981. MR1286212 (95g:65146)
2. E. Burman, A. Ern, Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes, C. R. Math. Acad. Sci. Paris 338 (2004), 641–646. MR2056474
3. P.G. Ciarlet, Discrete maximum principle for finite difference operators, Aequationes Math. 4 (1970), 338–352. MR0292317 (45:1404)
4. P.G. Ciarlet, P.A. Raviart, Maximum principle and uniform convergence for the finite element method, Computer Methods Appl. Mech. Engrg. 2 (1973), 17–31. MR0375802 (51:11992)
5. A. Drăgănescu, T.F. Dupont, L.R. Scott, Failure of the discrete maximum principle for an elliptic finite element problem, Math. Comp. 74 (2005), 1–23 (electronic). MR2085400 (2005f:65148)
6. M. Fiedler, Special matrices and their applications in numerical mathematics, Martinus Nijhoff Publishers, Dordrecht, 1986. MR1105955 (92b:15003)
7. W. Höhn, H.D. Mittelmann, Some remarks on the discrete maximum principle for finite elements of higher-order, Computing 27 (1981), 145–154. MR632125 (83a:65109)
8. A. Jüngel, A. Unterreiter, Discrete minimum and maximum principles for finite element approximations of non-monotone elliptic equations, Numer. Math. 99 (2005), 485–508. MR2117736 (2005m:65269)
9. J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, Numer. Math. 99 (2005), 669–698. MR2121074 (2005k:65253)

10. S. Korotov, M. Křížek, P. Neittaanmäki, Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle, Math. Comp. 70 (2000), 107–119. MR1803125 (2001i:65126)
11. A.H. Schatz, A weak discrete maximum principle and stability of the finite element method in L_∞ on plane polygonal domains. I, Math. Comp. 34 (1980), 77–91. MR551291 (81e:65063)
12. P. Šolín, Partial differential equations and the finite element method, J. Wiley & Sons, 2005. MR2180081 (2006f:35004)
13. P. Šolín, K. Segeth, I. Doležel, Higher-order finite element methods, Chapman & Hall/CRC Press, Boca Raton, 2003.
14. P. Šolín, T. Vejchodský, A weak discrete maximum principle for *hp*-FEM, J. Comput. Appl. Math., 2006 (to appear).
15. B. Szabó, I. Babuška, Finite element analysis, John Wiley & Sons, New York, 1991. MR1164869 (93f:73001)
16. R.S. Varga, Matrix iterative analysis, Englewood Cliffs, New Jersey, Prentice-Hall, 1962. MR0158502 (28:1725)
17. T. Vejchodský, On the nonnegativity conservation in semidiscrete parabolic problems. In: M. Křížek, P. Neittaanmäki, R. Glowinski, S. Korotov (Eds.), Conjugate gradients algorithms and finite element methods, Berlin, Springer-Verlag, 2004, pp. 197–210. MR2082563 (2005i:65135)
18. T. Vejchodský, Method of lines and conservation of nonnegativity. In: Proc. of the European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2004), Jyväskylä, Finland, 2004.
19. T. Vejchodský, P. Šolín, Discrete Maximum Principle for Mixed Boundary Conditions in 1D, Research Report No. 2006-09, Department of Math. Sciences, University of Texas at El Paso, July 2006.
20. J. Xu, L. Zikatanov, A monotone finite element scheme for convection-diffusion equations, Math. Comp. 68 (1999), 1429–1446. MR1654022 (99m:65225)
21. E.G. Yanik, Sufficient conditions for a discrete maximum principle for high-order collocation methods, Comput. Math. Appl. 17 (1989), 1431–1434. MR999250 (90c:65106)

Mathematical Institute, Academy of Sciences, Žitná 25, Praha 1, CZ-115 67, Czech Republic

E-mail address: vejchod@math.cas.cz

Institute of Thermomechanics, Academy of Sciences, Dolejškova 5, Praha 8, CZ-182 00, Czech Republic

Current address: Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas 79968-0514

E-mail address: solin@utep.edu

APPENDIX

E

Discrete maximum principle for Poisson equation with mixed boundary conditions solved by *hp*-FEM

Below we attach a copy of the paper

[A4] 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), 201–214.

Advances in Applied Mathematics and Mechanics
Adv. Appl. Math. Mech., Vol. **1**, No. 2, pp. 201-214 (2009)

March 2009

Discrete Maximum Principle for Poisson Equation with Mixed Boundary Conditions Solved by *hp*-FEM

Tomáš Vejchodský[1,*] and Pavel Šolín[1,2]

[1] *Institute of Mathematics, Czech Academy of Sciences, Žitná 25, CZ-115 67 Praha 1, Czech Republic,*

[2] *Department of Mathematics and Statistics, University of Nevada, Reno, USA, Institute of Thermomechanics, Czech Academy of Sciences, Dolejškova 5, Praha 8, CZ-18200, Czech Republic.*

Received 10 December 2008; Accepted (in revised version) 19 January 2009

Available online 17 March 2009

Abstract. We present a proof of the discrete maximum principle (DMP) for the 1D Poisson equation $-u''=f$ equipped with mixed Dirichlet-Neumann boundary conditions. The problem is discretized using finite elements of arbitrary lengths and polynomial degrees (*hp*-FEM). We show that the DMP holds on all meshes with no limitations to the sizes and polynomial degrees of the elements.

AMS subject classifications: 65N30

Key words: Discrete maximum principle, *hp*-FEM, Poisson equation, mixed boundary conditions

1 Introduction

It is well known that the finite element solutions to elliptic and parabolic PDEs sometimes exhibit behavior which is incompatible with the corresponding maximum principles and, consequently, incompatible with the underlying physics. Most frequently this happens when a finite element mesh contains large dihedral angles, but also in other situations. Discrete maximum principles (DMP) provide additional restrictions on finite element meshes under which the maximum principles are preserved on the discrete level.

Up to our knowledge the first DMP were introduced in the 1960s [16]. In the 1970s

*Corresponding author.
URL: `http://spilka.math.unr.edu/people/pavel/`
Email: vejchod@math.cas.cz (T. Vejchodský), solin@utep.edu (P. Šolín)

DMP were used to prove the convergence of finite differences and lowest-order finite element methods (see, e.g., [3,4]). Nowadays the DMP play an important role in computational PDEs by guaranteeing that approximation of physically nonnegative quantities such as the density, temperature, concentration, or electric charge remains nonnegative. Due to the difficulty of the topic, current research in the area of DMP almost exclusively deals with lowest-order elements (see, e.g., [2,7–10,17,18,20]). However, in the last decades, significant progress has been made in the development of the hp-FEM (finite element methods with variable size and polynomial degree of elements) and their applications to challenging large-scale problems in computational science and engineering (see, e.g., [1,11,12,15]). These methods are substantially more efficient compared to standard lowest-order schemes, and an increasing demand for them implies a need for the corresponding generalizations of the DMP.

However, the generalization of the DMP to higher-order approximations is quite demanding and there only are a few known results in this direction. We mention paper [21] concerning the high-order collocation method and a negative result [6] showing that a nonstandard version of DMP is not valid for quadratic and higher-order FEM in 2D.

It was shown in [14] that the DMP cannot be extended from the lowest-order FEM to hp-FEM in a straightforward manner, and a weak DMP was introduced. Recently, a maximum principle for one-dimensional Poisson equation equipped with Dirichlet boundary conditions and discretized by hp-FEM was presented in [19]. The result was proved under a mild sufficient condition stating that the length of the longest element in the mesh must be less than 90% of the length of the entire domain. In this paper we investigate the case of mixed Neumann-Dirichlet boundary conditions. using different analytical methods. Interestingly, it turns out that in this case, the DMP holds true with no restrictions.

In general, the analysis of the DMP for mixed boundary conditions follows the same steps as the analysis for the Dirichlet conditions presented in [19]. Nevertheless, the stiffness matrices in both cases differ. Fortunately, even in the case of the mixed boundary conditions there exists an explicit formula for entries of the inverse stiffness matrix, see Lemma 4.1. Naturally, this formula differs from the case of the pure Dirichlet conditions. Consequently, the corresponding discrete Green's functions differ and, hence, we had to develop a new proof of its nonnegativity in the case of the mixed boundary conditions, see Section 5. Interestingly, the same quantity $H^*_{\text{rel}}(p)$, where p stands for the polynomial degree, plays the crucial role in both cases. However, this role differs. While in the case of Dirichlet conditions the DMP is satisfied if the relative length of all elements is at most $H^*_{\text{rel}}(p)$, in the case of mixed conditions it suffices for the validity of DMP to have $H^*_{\text{rel}}(p)\geq 0$.

Furthermore, the nature of the maximum principle for the Dirichlet and for the mixed boundary conditions differs. In both cases the maximum principle is equivalent to the conservation of nonnegativity, see Definitions 2.1-2.3. However, in the case of Dirichlet conditions this equivalence is trivial and in the case of the mixed conditions the maximum principle implies the conservation of nonnegativity in a nontrivial way.

2 The model problem and its discretization

We solve the one dimensional Poisson equation with mixed Dirichlet-Neumann boundary conditions,

$$\begin{aligned} -u'' &= f, \qquad \text{in } \Omega, \\ u(\alpha) &= 0, \qquad u'(\beta) = g(\beta). \end{aligned}$$

Here, $\Omega{=}(\alpha,\beta){\subset}\mathbb{R}$ is an interval.

The corresponding weak formulation reads: Find $u{\in}V$ such that

$$a(u,v) = (f,v) + g(\beta)v(\beta), \qquad \forall v \in V, \tag{2.1}$$

where $V{=}\{v{\in}H^1(\Omega); v(\alpha){=}0\}$, $f{\in}L^2(\Omega)$ is a right-hand side, $g(\beta){\in}\mathbb{R}$, $(\cdot,\cdot)$ stands for an $L^2(\Omega)$ inner product, and $a(u,v){=}(u',v')$.

In a standard way we create a partition $\alpha{=}x_0{<}x_1{<}\ldots{<}x_M{=}\beta$ of the domain Ω consisting of M elements $K_i{=}[x_{i-1},x_i]$, $i{=}1,2,\ldots,M$. Every element K_i is assigned an arbitrary polynomial degree $p_i{\geq}1$. The corresponding finite element space $V_h{\subset}V$ of piecewise-polynomial and continuous functions has the form

$$V_{hp} = \left\{ v_{hp} \in V;\ v_{hp}|_{K_i} \in P^{p_i}(K_i),\ i = 1,2,\ldots,M \right\}.$$

Here $P^{p_i}(K_i)$ stands for the space of polynomials of degree at most p_i on the element K_i. The space V_{hp} has the dimension $N = \sum_{i=1}^{M} p_i$. There exists a unique finite element solution $u_{hp}{\in}V_{hp}$ satisfying

$$a(u_{hp},v_{hp}) = (f,v_{hp}) + g(\beta)v(\beta), \qquad \forall v_{hp} \in V_{hp}. \tag{2.2}$$

Definition 2.1. *Problem* (2.2) *satisfies the* discrete maximum principle (DMP) *if*

$$f \leq 0 \text{ a.e. in } \Omega \text{ and } g(\beta) \leq 0, \quad \Rightarrow \quad \max_{\overline{\Omega}} u_{hp} = \max_{\partial\Omega} u_{hp},$$

where $\partial\Omega$ *is the boundary of the domain* Ω.

Definition 2.2. *Problem* (2.2) *satisfies the* discrete minimum principle *if*

$$f \geq 0 \text{ a.e. in } \Omega \text{ and } g(\beta) \geq 0, \quad \Rightarrow \quad \min_{\overline{\Omega}} u_{hp} = \min_{\partial\Omega} u_{hp}.$$

Definition 2.3. *Problem* (2.2) conserves nonnegativity *if*

$$f \geq 0 \text{ a.e. in } \Omega \text{ and } g(\beta) \geq 0, \quad \Rightarrow \quad u_{hp} \geq 0 \text{ in } \Omega.$$

Clearly, the discrete maximum and minimum principles are equivalent for problem (2.2). We will use this equivalence and the following lemma to prove the DMP via conservation of nonnegativity.

Lemma 2.1. *If problem* (2.2) *conserves nonnegativity then it satisfies the discrete minimum principle.*

Proof. Since $u_{hp} \geq 0$ in Ω and $u_{hp}(\alpha)=0$, we conclude $\min_{\partial\Omega} u_{hp}=0=\min_{\overline{\Omega}} u_{hp}$. □

Remark 2.1. For the sake of simplicity, we formulated problem (2.2) with a homogeneous Dirichlet boundary condition $u(\alpha)=0$. However, all results of this study hold for a nonhomogeneous condition of the form $u(\alpha)=u_\alpha$. Indeed, the Dirichlet lift is constant in this case and every solution $\hat{u}_{hp}$ to problem (2.2) with nonhomogeneous condition $u(\alpha)=u_\alpha$ can be decomposed to

$$\hat{u}_{hp} = u_\alpha + u_{hp},$$

where u_{hp} vanishes at the endpoint α.

Remark 2.2. The Neumann boundary condition at the point β can be replaced by the more general Robin's boundary condition

$$u'(\beta) + \gamma u(\beta) = g(\beta), \quad \text{with } \gamma \geq 0.$$

The presented analysis can be generalized to this case as well.[†]

3 Discrete Green's function

The discrete Green's function (DGF) is defined in analogy to the standard Green's function:

Definition 3.1. *For an arbitrary $z \in \overline{\Omega}$, the unique solution $G_{hp,z} \in V_{hp}$ to the problem*

$$a(v_{hp}, G_{hp,z}) = v_{hp}(z), \qquad \forall v_{hp} \in V_{hp}, \tag{3.1}$$

is called the discrete Green's function (DGF) *corresponding to the point z.*

In the following, we will use the notation

$$G_{hp}(x,z) = G_{hp,z}(x), \quad \text{for } (x,z) \in \overline{\Omega}^2,$$

where $\overline{\Omega}^2 = \overline{\Omega} \times \overline{\Omega}$. A combination of (2.2) and (3.1) yields the so-called Kirchhoff-Helmholtz representation

$$u_{hp}(z) = \int_\Omega G_{hp}(x,z) f(x)\,\mathrm{d}x + g(\beta) G_{hp}(\beta,z), \qquad \forall z \in \overline{\Omega}. \tag{3.2}$$

The following lemma shows that the DGF can easily be expressed using any basis of V_{hp}, cf. [5]. We use the Kronecker symbol

$$\delta_{ik} = \begin{cases} 1 & \text{for } i = k, \\ 0 & \text{for } i \neq k. \end{cases}$$

[†]We thank Sergey Korotov from the Helsinki University of Technology for pointing this out.

Lemma 3.1. *Let* $\{\varphi_1, \varphi_2, \ldots, \varphi_N\}$ *be a basis of* V_{hp}*. If the stiffness matrix* $A_{ij}=a(\varphi_j, \varphi_i)$*,* $1\leq i, j\leq N$ *is nonsingular, then*

$$G_{hp}(x,z) = \sum_{j=1}^{N}\sum_{k=1}^{N} A^{-1}_{jk}\varphi_k(x)\varphi_j(z). \tag{3.3}$$

Here, A^{-1}_{jk} *are the entries of the inverse stiffness matrix, i.e.,* $\sum_{j=1}^{N} A_{ij}A^{-1}_{jk} = \delta_{ik}$*,* $1 \leq i, k \leq N$*.*

Proof. Substitute

$$G_{hp}(x,z) = \sum_{i=1}^{N} c_i(z)\varphi_i(x), \tag{3.4}$$

into (3.1) with $v_{hp}=\varphi_j$. It follows that

$$\sum_{i=1}^{N} c_i(z) \underbrace{a(\varphi_j, \varphi_i)}_{A_{ij}} = \varphi_j(z).$$

The coefficients $c_i(z)$ are expressed as $c_k(z)=\sum_{j=1}^{N} \varphi_j(z)A^{-1}_{jk}$ in terms of the inverse matrix, and they are substituted back into (3.4). This finishes the proof. □

Theorem 3.1. *Problem* (2.2) *conserves nonnegativity if and only if the corresponding discrete Green's function* $G_{hp}(x,z)=G_{hp,z}(x)$ *defined by* (3.1) *is nonnegative in* $\overline{\Omega}^2$*.*

Proof. By (3.3), the discrete Green's function $G_{hp}(x,z)$ is continuous up to the boundary of Ω. The rest follows immediately from (3.2). □

This theorem is a useful tool for the analysis of discrete maximum principles. In the rest of this paper we will show that the discrete Green's function corresponding to the problem (2.2) is nonnegative.

4 DGF for the model problem

4.1 Lowest-order case

In this section we will construct the DGF for problem (2.2). We begin with the case $p_1=p_2=\ldots=p_M=1$. Let us define $h_i=x_i - x_{i-1}$. By $\mathcal{B}^L=\{\phi_1, \phi_2, \ldots, \phi_M\}$ we denote the standard lowest-order basis consisting of the piecewise-linear "hat functions" such that $\phi_j(x_i)=\delta_{ij}$, $1 \leq i, j \leq M$. In this case the stiffness matrix $A^L\in\mathbb{R}^{M\times M}$ is tridiagonal,

$$A^L_{ij} = \begin{cases} 1/h_i + 1/h_{i+1}, & \text{for } i = j < M, \\ 1/h_M, & \text{for } i = j = M, \\ -1/h_{i+1}, & \text{for } i = j-1, \\ -1/h_{i-1}, & \text{for } i = j+1, \\ 0, & \text{otherwise,} \end{cases}$$

for $i, j = 1, 2, \ldots, M$.

Lemma 4.1. *The inverse matrix* $(A^L)^{-1} \in \mathbb{R}^{M\times M}$ *has the form*

$$(A^L)^{-1} = \begin{pmatrix} x_1-\alpha & x_1-\alpha & x_1-\alpha & \dots & x_1-\alpha \\ x_1-\alpha & x_2-\alpha & x_2-\alpha & \dots & x_2-\alpha \\ x_1-\alpha & x_2-\alpha & x_3-\alpha & \dots & x_3-\alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1-\alpha & x_2-\alpha & x_3-\alpha & \dots & x_M-\alpha \end{pmatrix},$$

i.e., $(A^L)^{-1}_{ij}=x_i-\alpha$ *for* $1\le i\le j\le M$ *and* $(A^L)^{-1}_{ij}=x_j-\alpha$ *for* $1\le j<i\le M$.

Proof. We want to show that $z_{ij}=\delta_{ij}$, where

$$z_{ij} = \sum_{k=1}^{M} (A^L)^{-1}_{ik} A^L_{kj} = \sum_{k=1}^{M} (A^L)^{-1}_{ik} a(\phi_j,\phi_k),$$

for all $i,j=1,2,\dots,M$. We fix i and j, and consider the bilinear forms

$$a_1(u,v) = \int_\alpha^{x_i} u'v'\,\mathrm{d}x \quad \text{and} \quad a_2(u,v) = \int_{x_i}^{\beta} u'v'\,\mathrm{d}x.$$

We use the explicit formulae for $(A^L)^{-1}_{ik}$ to get

$$z_{ij} = a\Big(\phi_j, \sum_{k=1}^{i-1}(x_k-\alpha)\phi_k\Big) + (x_i-\alpha)a(\phi_j,\phi_i) + (x_i-\alpha)a\Big(\phi_j, \sum_{k=i+1}^{M}\phi_k\Big).$$

Now, we split the term $a(\phi_j,\phi_i)=a_1(\phi_j,\phi_i)+a_2(\phi_j,\phi_i)$ to obtain

$$z_{ij} = a_1(\phi_j, x-\alpha) + (x_i-\alpha)a_2(\phi_j,1) = a_1(\phi_j,x-\alpha) = \delta_{ij},$$

where the last equality follows from a straightforward simple computation. □

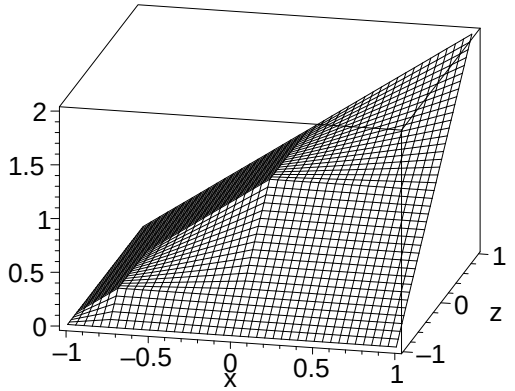


Figure 1: The lowest-order part $G^L_{hp}(x,z)$ of the discrete Green's function $G_{hp}(x,z)$ for the Poisson equation with homogeneous mixed boundary conditions in $\Omega=(-1,1)$ on a mesh with three elements $[-1,-3/4]$, $[-3/4,0]$, and $[0,1]$.

Using Lemma 4.1 and identity (3.3), we can write the DGF in the form

$$G_{hp}^L(x,z) = \sum_{i=1}^{M}(x_i-\alpha)\phi_i(x)\phi_i(z) + \sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(x_i-\alpha)[\phi_i(x)\phi_j(z)+\phi_j(x)\phi_i(z)]. \tag{4.1}$$

In particular, we see immediately that

$$G_{hp}^L(x,z) \geq 0, \quad \forall (x,z) \in \overline{\Omega}^2. \tag{4.2}$$

The situation is illustrated in Fig. 1.

4.2 Higher-order case

In this paragraph we return to the original setting with arbitrary polynomial degrees $p_i \geq 1$. In order to facilitate the construction of higher-order basis functions of the space V_{hp}, let us introduce the Lobatto shape functions $l_0, l_1, l_2, \ldots$ on a reference interval $\hat{K}=[-1,1]$, see, e.g., [12,15] and (7.1) in Appendix.

The lowest-order Lobatto shape functions l_0 and l_1 have the form $l_0(\xi)=(1-\xi)/2$, $l_1(\xi)=(1+\xi)/2$, $\xi \in \hat{K}$. The higher-order shape functions $l_2, l_3, \ldots$ are defined as antiderivatives to the Legendre polynomials. Therefore, they satisfy

$$\int_{-1}^{1} l_k'(\xi) l_m'(\xi)\,\mathrm{d}\xi = \delta_{km}, \quad k,m = 2,3,\ldots.$$

Every Lobatto shape function l_k, $k=2,3,\ldots$, is a polynomial of degree k and it vanishes at ± 1. Thus it can be expressed as

$$l_{k+2}(\xi) = l_0(\xi) l_1(\xi) \kappa_k(\xi), \quad k = 0,1,2,\ldots,$$

where κ_k is a polynomial of degree k. For reference, a first few kernels κ_k are listed in Appendix.

The basis $\mathcal{B}=\{\phi_1, \phi_2, \ldots, \phi_N\}$ of V_{hp} can be written as $\mathcal{B}=\mathcal{B}^L \cup \mathcal{B}^B$, where $\mathcal{B}^L$ was defined above and $\mathcal{B}^B$ is the higher-order part of the basis comprising functions ϕ_M, $\phi_{M+1}, \ldots, \phi_N$. These are defined in a standard way as follows:

Consider the standard affine transformations of the reference element $\hat{K}$ to an element $K_i=[x_{i-1}, x_i]$, $i=1,2,\ldots,M$,

$$\chi_{K_i}(\xi) = \frac{(x_i - x_{i-1})\xi + (x_i + x_{i-1})}{2}. \tag{4.3}$$

On an element K_i of the polynomial degree p_i, there are $p_i - 1$ higher-order basis functions. These vanish outside of K_i and in K_i they are defined as the Lobatto shape functions $l_2, l_3, \ldots, l_{p_i}$ composed with the inverse map $\chi_{K_i}^{-1}(x)$.

Lemma 4.2. *We have the following orthogonality relations:*

$$a(\phi^L, \phi^B) = 0, \qquad \forall \phi^L \in \mathcal{B}^L,\ \forall \phi^B \in \mathcal{B}^B,$$
$$a(\phi^B, \psi^B) = 0, \qquad \forall \phi^B \in \mathcal{B}^B,\ \forall \psi^B \in \mathcal{B}^B,\ \phi^B \neq \psi^B.$$

Proof. The proof is straightforward, based on the L^2-orthogonality of the Legendre polynomials. □

By Lemma 4.2, both the stiffness matrix A and its inverse have the following block structure:

$$A = \begin{pmatrix} A^L & 0 \\ 0 & D \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} (A^L)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix},$$

with

$$D = \operatorname{diag}\Big(\underbrace{\frac{2}{h_1}, \ldots, \frac{2}{h_1}}_{(p_1-1)\text{ times}}, \underbrace{\frac{2}{h_2}, \ldots, \frac{2}{h_2}}_{(p_2-1)\text{ times}}, \ \ldots\ , \underbrace{\frac{2}{h_M}, \ldots, \frac{2}{h_M}}_{(p_M-1)\text{ times}}\Big). \tag{4.4}$$

By (3.3), the DGF can be written as

$$G_{hp}(x,z) = G^L_{hp}(x,z) + G^B_{hp}(x,z), \tag{4.5}$$

where $G^L_{hp}(x,z)$ corresponds to (4.1) and

$$G^B_{hp}(x,z) = \sum_{k=M}^{N} D^{-1}_{jj} \phi_j(x) \phi_j(z), \qquad \forall (x,z) \in \overline{\Omega}^2. \tag{4.6}$$

Unfortunately, $G^B_{hp}(x,z)$ defined by (4.6) is not nonnegative in the entire $\overline{\Omega}^2$ in general. For instance, in the example shown in Fig. 2, there are small regions near the points $(1,0)$ and $(0,1)$, where the function $G^B_{hp}(x,z)$ is negative.

Notice that any partition of $\overline{\Omega}$ produces a rectangular grid on $\overline{\Omega}^2$, and that $G^B_{hp}(x,z)$ can be nonzero within the diagonal squares of this grid only. In other words,

$$\operatorname{supp} G^B_{hp} \subset \bigcup_{i=1}^{M} K_i^2. \tag{4.7}$$

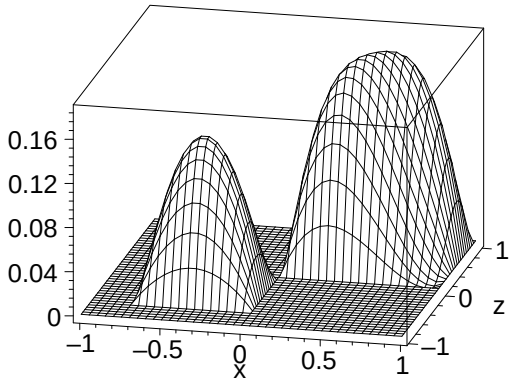


Figure 2: The higher-order part $G^B_{hp}(x,z)$ of the discrete Green's function $G_{hp}(x,z)$ for the Poisson equation with homogeneous mixed boundary conditions in $\Omega = (-1,1)$, on a mesh with three elements $[-1,-3/4]$, $[-3/4,0]$, and $[0,1]$ of the polynomial degrees p_1=1, p_2=2, p_3=3.

Lemma 4.3. *The discrete Green's function G_{hp} defined by* (4.5) *is nonnegative in* $\overline{\Omega}^2 \setminus \bigcup_{i=1}^M K_i^2$.

Proof. Considering (4.7) together with (4.2) leads to the conclusion. □

5 The DGF on K_i^2

As justified by Lemma 4.3, we only need to continue with the study of the discrete Green's function $G_{hp}(x,z)$ in the union of the diagonal squares $\bigcup_{i=1}^M K_i^2$. Without loss of generality, let us restrict ourselves to only one square K_i^2, $1 \le i \le M$. Let $p=p_i$ be the polynomial degree assigned to K_i. Notice that only a few terms in (4.1) and (4.6) are nonzero in K_i^2. Hence, by (4.1), (4.4), and (4.6) we obtain

$$\begin{aligned} G_{hp}(x,z)\big|_{K_i^2} = {} & (x_i-\alpha)\phi_i(x)\phi_i(z) + (x_{i-1}-\alpha)\phi_{i-1}(x)\phi_{i-1}(z) \\ & +(x_{i-1}-\alpha)\left[\phi_i(x)\phi_{i-1}(z) + \phi_{i-1}(x)\phi_i(z)\right] \\ & +\frac{x_i - x_{i-1}}{2} G_{hp}^B(x,z)|_{K_i^2}, \end{aligned} \tag{5.1}$$

for $(x,z) \in K_i^2$, $1 \le i \le M$. It is convenient to introduce the notation $K_i = [x_{i-1}, x_i] = [L, R]$.

We transform the function G_{hp} from K_i^2 to the reference square $\hat{K}^2 = [-1,1]^2$ using the linear transformation (4.3) with $x = \chi_{K_i}(\xi)$ and $z = \chi_{K_i}(\eta)$,

$$\begin{aligned} & G_{hp}(x,z)\big|_{K_i^2} = \hat{G}_{hp}(\xi,\eta) \\ = {} & (R-\alpha) l_1(\xi) l_1(\eta) + (L-\alpha) l_0(\xi) l_0(\eta) \\ & +(L-\alpha)\left[l_1(\xi) l_0(\eta) + l_0(\xi) l_1(\eta)\right] + \frac{R-L}{2} \hat{G}_{hp}^{p,B}(\xi,\eta), \end{aligned} \tag{5.2}$$

for $(\xi,\eta) \in \hat{K}^2$. Here $l_0(\xi)$ and $l_1(\xi)$ are the above-defined lowest-order shape functions on $\hat{K}$ and

$$\hat{G}_{hp}^{p,B}(\xi,\eta) = \sum_{m=2}^{p} l_m(\xi) l_m(\eta) = l_0(\xi) l_0(\eta) l_1(\xi) l_1(\eta) \sum_{k=0}^{p-2} \kappa_k(\xi) \kappa_k(\eta), \tag{5.3}$$

is the higher-order part.

Let us modify formula (5.2) in the following way: Divide (5.2) by $R-L>0$ and use the identities

$$\frac{R-\alpha}{R-L} = \frac{L-\alpha}{R-L} + 1,$$

and

$$l_0(\xi) l_0(\eta) + l_1(\xi) l_1(\eta) + l_0(\xi) l_1(\eta) + l_1(\xi) l_0(\eta) = 1, \qquad \forall (\xi,\eta) \in \hat{K}^2.$$

We obtain

$$\frac{\hat{G}_{hp}(\xi,\eta)}{R-L} = \frac{L-\alpha}{R-L} + l_1(\xi) l_1(\eta) + \frac{1}{2} \hat{G}_{hp}^{p,B}(\xi,\eta). \tag{5.4}$$

Table 1: The quantity $H^*_{\rm rel}(p)$ for $p=1,2,3,\dots,20$.

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

Using (5.3), this formula can be reshaped into

$$\frac{\hat{G}_{hp}(\xi,\eta)}{R-L}=\frac{L-\alpha}{R-L}+l_1(\xi)l_1(\eta)\Big[1+\frac{1}{2}l_0(\xi)l_0(\eta)\sum_{k=0}^{p-2}\kappa_k(\xi)\kappa_k(\eta)\Big]. \tag{5.5}$$

Clearly, $(L-\alpha)/(R-L)\geq 0$ and $l_1(\xi)l_1(\eta)\geq 0$ in $\hat{K}^2$. It remains to verify nonnegativity of the expression in the square brackets. For this reason we define

$$H^*_{\rm rel}(p)=1, \qquad \text{for } p=1,$$
$$H^*_{\rm rel}(p)=1+\frac{1}{2}\min_{(\xi,\eta)\in\hat{K}^2} l_0(\xi)l_0(\eta)\sum_{k=0}^{p-2}\kappa_k(\xi)\kappa_k(\eta), \quad \text{for } p\geq 2.$$

Hence, if $H^*_{\rm rel}(p)\geq 0$ then $\hat{G}_{hp}(\xi,\eta)\geq 0$ in $\hat{K}^2$ by (5.5). Transforming (ξ,η) back to (x,z) by (4.3), we obtain nonnegativity of $G_{hp}(x,z)$ in K_i^2, cf. (5.2), for all $i=1,2,\dots,M$. Thus, in view of Lemma 4.3 we showed that the discrete Green's function $G_{hp}(x,z)\geq 0$ in $\overline{\Omega}^2$, provided $H^*_{\rm rel}(p_i)\geq 0$ for all $i=1,2,\dots,M$.

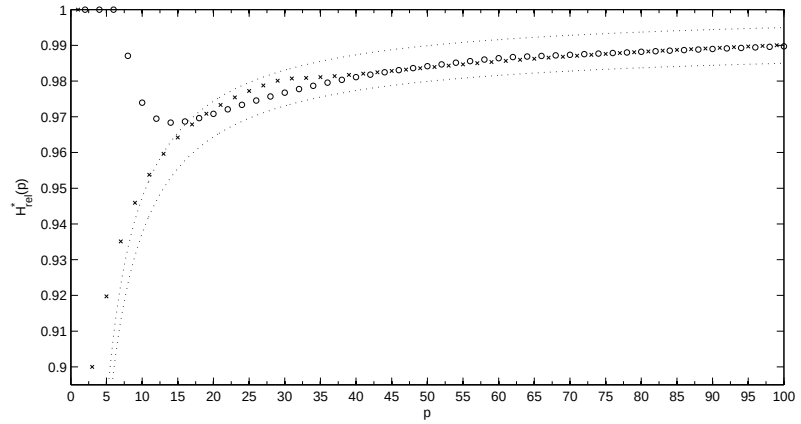


Figure 3: The values $H^*_{\rm rel}(p)$ for $p=1,2,\dots,104$. Circles indicate the values for p odd and crosses for p even. The upper dotted line is a graph of $1+0.5\ln(1-1/x)$ and the bottom line is a shift of this graph by -0.01.

In [19] it was verified that $H^*_{\rm rel}(p)\geq 0$ for $1\leq p\leq 100$. More precisely, the value of $H^*_{\rm rel}(p)$ can be found analytically for $2\leq p\leq 4$. For $5\leq p\leq 100$, it needs to be computed

numerically. As it is seen from Table 1 and Fig. 3, the smallest value of $H^*_{\rm rel}(p)$ is for $p=3$ and it equals to $9/10$. Thus, the crucial quantity $H^*_{\rm rel}(p)$ was checked to be nonnegative for $1\leq p\leq 100$. This, consequently, shows nonnegativity of the discrete Green's function in $\overline{\Omega}^2$ and validity of the discrete maximum principle.

6 Main result

Let us summarize the conclusions of the previous analysis:

Theorem 6.1. *Let $\alpha=x_0<x_1<\ldots<x_M=\beta$ be a partition of the domain $\Omega=(\alpha,\beta)$ and let $p_i\geq 1$ be a polynomial degree assigned to the element $K_i=[x_{i-1},x_i]$, $i=1,2,\ldots,M$. If*

$$H^*_{\rm rel}(p_i)\geq 0 \quad \text{for all } i=1,2,\ldots,M, \tag{6.1}$$

then problem (2.2) satisfies the discrete maximum principle

Proof. Let K_i be an element. By (5.2), (5.5), and (6.1) it holds

$$G_{hp}(x,z)|_{K_i^2}=\hat{G}_{hp}(\xi,\eta)\geq 0$$

for all $(x,z)\in K_i^2$ with $\xi=\chi_{K_i}^{-1}(x)$ and $\eta=\chi_{K_i}^{-1}(z)$. Thus, $G_{hp}(x,z)\geq 0$ in $\bigcup_{i=1}^M K_i^2$. Lemma 4.3 implies that $G_{hp}(x,z)\geq 0$ also in $\overline{\Omega}^2\setminus\bigcup_{i=1}^M K_i^2$. Theorem 3.1 and Lemma 2.1 finish the proof. □

The crucial condition (6.1) was verified analytically for $p\leq 4$, therefore Theorem 6.1 proves the discrete maximum principle for problem (2.2) for all meshes and arbitrary polynomial degrees not exceeding 4. However, numerical calculations of $H^*_{\rm rel}(p)$ show that the condition (6.1) is satisfied for $5\leq p\leq 100$ as well. Moreover, the steadily growing trend in $H^*_{\rm rel}$ for $p\geq 50$ observed in Fig. 3 motivates the following conjecture:

Conjecture 1. *The problem (2.2) satisfies the discrete maximum principle for arbitrary partition of the domain $\Omega=(\alpha,\beta)$ and for arbitrary distribution of polynomial degrees.*

7 Conclusions and further generalizations

We proved the DMP for the 1D Poisson problem solved by the sophisticated *hp*-version of the FEM. The next natural step is to generalize this result for more general problems in two (or more) dimensions.

Since the key ingredients (Lemma 3.1 and Theorem 3.1) are valid for arbitrary elliptic operator in arbitrary dimension, the presented approach can be, in principle, extended to prove the DMP even in more general settings. However, the conditions for the mesh and polynomial degrees which would guarantee the DMP are then more difficult to find.

More general operators, for example the diffusion-reaction operator, bring difficulties such as (i) the non-existence of a simple formula for the inverse of the stiffness matrix, cf. Lemma 4.1, and (ii) non-orthogonality of the bubble functions to the vertex ones, cf. Lemma 4.2. These difficulties can be treated for instance in the following way. In case (i) we have to find suitable lower bounds for the entries of the inverse stiffness matrix. This can be done by analysing simplified meshes with a few elements and showing that their refinement leads to an increase of nodal values of the discrete Green's function. Difficulty (ii) is not fundamental and it can be treated by orthogonalization of the vertex functions with respect to bubbles (the concept of the discrete minimum energy extensions).

With no doubts, the significance of the *hp*-FEM lies in 2D and 3D problems. When extending the DMP results to higher-order methods in higher spatial dimensions, one has to overcome not only the two difficulties mentioned above but also (iii) the presence of the edge (and face) basis functions. These basis functions make the process of orthogonalization of the vertex functions to the other basis functions non-local which makes the analysis more demanding but treatable.

The search for suitable conditions for more general and higher dimensional problems is a challenging task of high practical significance. Generalizations of the presented results are desirable because conditions guaranteeing the physical admissibility of *hp*-FEM approximations are valuable from the practical point of view, and they are demanded from the engineering community.

Appendix

The Lobatto shape functions are defined by

$$l_m(\xi) = \sqrt{\frac{2m-1}{2}} \int_{-1}^{\xi} P_{m-1}(x)\,\mathrm{d}x, \qquad m = 2, 3, \ldots, \tag{7.1}$$

where

$$P_m(x) = \mathrm{d}^m/\mathrm{d}x^m (x^2-1)^m/(2^m m!),$$

stands for the mth-degree Legendre polynomial. The kernels are defined by

$$\kappa_k(\xi) = l_{k+2}(\xi)/(l_0(\xi) l_1(\xi)),\ k = 0, 1, 2, \ldots,$$

where

$$l_0(\xi) = (1-\xi)/2, l_1(\xi) = (1+\xi)/2, \xi \in [-1, 1].$$

These kernels can be generated by the recurrence

$$\frac{k+4}{\sqrt{2k+7}}\kappa_{k+2}(\xi) = \sqrt{2k+5}\xi\kappa_{k+1}(\xi) - \frac{k+1}{\sqrt{2k+3}}\kappa_k(\xi), \qquad k = 0, 1, 2, \ldots.$$

Interesting observation is that these kernels are scaled derivatives of Legendre polynomials

$$\kappa_k(\xi) = -\frac{\sqrt{8(2k+3)}}{(k+2)(k+1)} P'_{k+1}(\xi), \qquad k = 0, 1, 2, \ldots.$$

Hence, they form a system of orthogonal polynomials with weight $1-\xi^2 = 4l_0(\xi)l_1(\xi)$. For reference, we list several kernel functions κ_k (see, e.g., Section 3.1 in [15] or Section 1.2 in [13]):

$$\begin{aligned}
\kappa_0(\xi) &= -\sqrt{6}, \quad \kappa_1(\xi) = -\sqrt{10}\xi, \\
\kappa_2(\xi) &= -\frac{1}{4}\sqrt{14}(5\xi^2 - 1), \\
\kappa_3(\xi) &= -\frac{3}{4}\sqrt{2}(7\xi^2 - 3)\xi, \\
\kappa_4(\xi) &= -\frac{1}{8}\sqrt{22}(21\xi^4 - 14\xi^2 + 1), \\
\kappa_5(\xi) &= -\frac{1}{8}\sqrt{26}(33\xi^4 - 30\xi^2 + 5)\xi, \\
\kappa_6(\xi) &= -\frac{1}{64}\sqrt{30}(429\xi^6 - 495\xi^4 + 135\xi^2 - 5), \\
\kappa_7(\xi) &= -\frac{1}{64}\sqrt{34}(715\xi^6 - 1001\xi^4 + 385\xi^2 - 35)\xi, \\
\kappa_8(\xi) &= -\frac{1}{128}\sqrt{38}(2431\xi^8 - 4004\xi^6 + 2002\xi^4 - 308\xi^2 + 7).
\end{aligned}$$

Acknowledgments

The authors gratefully acknowledg the support of the Czech Science Foundation, projects No. 102/07/0496 and 102/05/0629, the Grant Agency of the Academy of Sciences of the Czech Republic, project No. IAA100760702, and the Academy of Sciences of the Czech Republic, Institutional Research Plan No. AV0Z10190503.

References

[1] I. Babuška and B. Q. Guo, *Approximation properties of the hp version of the finite element method*, Comput. Methods Appl. Mech. Engrg., 133 (1996), pp. 319-346.

[2] 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.

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

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

[5] 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.

[6] 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.
[7] A. JÜNGEL AND A. UNTERREITER, *Discrete minimum and maximum principles for finite element approximations of non-monotone elliptic equations*, Numer. Math., 99 (2005), pp. 485-508.
[8] 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.
[9] S. KOROTOV, M. KŘÍŽEK AND P. NEITTAANMÄKI, *Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle*, Math. Comp., 70 (2000), pp. 107-119.
[10] M. KŘÍŽEK AND L. LIU, *On the maximum and comparison principles for a steady-state nonlinear heat conduction problem*, ZAMM Z. Angew. Math. Mech., 83 (2003), pp. 559-563.
[11] C. SCHWAB, *p*- and *hp*-Finite Element Methods, Clarendon Press, Oxford, 1998.
[12] P. ŠOLÍN, Partial Differential Equations and the Finite Element Method, J. Wiley & Sons, Hoboken, NJ, 2006.
[13] P. ŠOLÍN, K. SEGETH AND I. DOLEŽEL, Higher-Order Finite Element Methods, Chapman & Hall/CRC Press, Boca Raton, 2003.
[14] P. ŠOLÍN AND T. VEJCHODSKÝ, *A weak discrete maximum principle for hp-FEM*, J. Comput. Appl. Math., 209 (2007), pp. 54-65.
[15] B. SZABÓ AND I. BABUŠKA, Finite Element Analysis, John Wiley & Sons, New York, 1991.
[16] R. S. VARGA, *On a discrete maximum principle*, SIAM J. Numer. Anal., 3 (1966), pp. 355-359.
[17] T. VEJCHODSKÝ, *On the nonnegativity conservation in semidiscrete parabolic problems*, in: M. Křížek, P. Neittaanmäki, R. Glowinski, S. Korotov (Eds.), Conjugate Gradients Algorithms and Finite Element Methods, Springer-Verlag, Berlin, 2004, pp. 197-210.
[18] T. VEJCHODSKÝ, *Method of lines and conservation of nonnegativity*, in: P. Neittaanmäki, T. Rossi, S. Korotov, E. Oñate, J. Périaux, and D. Knörzer (Eds.), European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS 2004, Jyväskylä, 24-28 July 2004, 18 (electronic).
[19] T. VEJCHODSKÝ AND P. ŠOLÍN, *Discrete maximum principle for higher-order finite elements in 1D*, Math. Comp., 76 (2007), pp. 1833-1846.
[20] J. XU AND L. ZIKATANOV, *A monotone finite element scheme for convection-diffusion equations*, Math. Comp., 68 (1999), pp. 1429-1446.
[21] E. G. YANIK, *Sufficient conditions for a discrete maximum principle for high-order collocation methods*, Comput. Math. Appl., 17 (1989), pp. 1431-1434.

APPENDIX

F

Discrete maximum principle for a 1D problem with piecewise-constant coefficients solved by *hp*-FEM

Below we attach a copy of the paper

[A5] 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), 233–243.

J. Numer. Math., Vol. 15, No. 3, pp. 233–243 (2007)
DOI 10.1515/ JNUM.2007.011

Discrete maximum principle for a 1D problem with piecewise-constant coefficients solved by *hp*-FEM

T. VEJCHODSKÝ* and P. ŠOLÍN†

Received August 31, 2006
Received in revised form May 30, 2007

Abstract — In this paper we prove the discrete maximum principle for a one-dimensional equation of the form $-(au')' = f$ with piecewise-constant coefficient $a(x)$, discretized by the *hp*-FEM. The discrete problem is transformed in such a way that the discontinuity of the coefficient $a(x)$ disappears. Existing results are then applied to obtain a condition on the mesh which guarantees the satisfaction of the discrete maximum principle. Both Dirichlet and mixed Dirichlet–Neumann boundary conditions are discussed.

Keywords: discrete maximum principle, *hp*-FEM, Poisson equation, piecewise-constant coefficients

1. Introduction

Discrete maximum principles (DMP) have been studied since the 1970s (see, e.g., [3,4] and the references therein). In particular, the maximum and comparison principles belong to the most important qualitative properties of numerical schemes. They guarantee, for example, the nonnegativity of approximations of naturally nonnegative quantities such as temperature, density, concentration, etc. When a numerical method does not satisfy the DMP, it can happen that the resulting numerical solution contradicts the physics. Therefore, the study of numerical methods equipped with discrete maximum principles became very popular during the last years.

Absolute majority of results on DMP concern lowest-order, such as piecewise-linear approximations [2,5,7–10,16,17,22]. Recent rapid development of higher-

*Institute of Mathematics, Czech Academy of Sciences, Žitná 25, Praha 1, CZ-115 67, Czech Republic, e-mail: vejchod@math.cas.cz

†Institute of Thermomechanics, Czech Academy of Sciences, Dolejškova 5, Praha 8, CZ-182 00, Czech Republic, *Current address*: Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas 79968-0514, USA, e-mail: solin@utep.edu

The first author has been supported by the Grant Agency of the Czech Republic, project No. 201/04/P021 and by the Academy of Sciences of the Czech Republic, Institutional Research Plan No. AV0Z10190503. The second author has been supported in part by the U.S. Department of Defense under Grant No. 05PR07548-00, by the NSF Grant No. DMS-0532645, and by Grant Agency of the Czech Republic projects No. 102-05-0629.

order methods, especially *hp*-FEM [1,11–13,15], leads to a question whether and under which conditions the DMP can be extended to this type of approximations. The answer to this question is difficult, since there is no simple condition for polynomials to be nonnegative – in contrast to the lowest-order case. There are only a few papers addressing DMP for higher-order approximations (see [6,21] and recent results of the authors [14,18–20].

A particularly discouraging 2D result (see [6]) shows that a *stronger* DMP for quadratic elements only is valid under prohibitively restrictive conditions on the mesh and that for higher-order elements the stronger DMP is not valid at all. The point is that the stronger DMP requires the maximum principle to be fulfilled on all subdomains, particularly on patches sharing a common vertex. This requirement, however, is too strong and its relaxation leaves space for investigation of a condition for the mesh for the DMP on the whole domain.

Recently, discrete maximum principles for the Poisson equation in 1D, discretized by *hp*-FEM, were studied in [14,18–20]. The present paper generalizes these results to the equation $-(au')' = f$, where the coefficient $a(x)$ is assumed to be discontinuous and piecewise-constant. We derive a sufficient condition on the mesh that guarantees the DMP in the case of Dirichlet boundary conditions. This condition involves the coefficient a and it can be easily verified in an element-by-element fashion. The case of mixed Dirichlet–Neumann boundary conditions is studied as well, and it is shown that the DMP is valid on all meshes with arbitrary distribution of polynomial degrees.

The scope of the paper is as follows: A model problem with piecewise-constant coefficient $a(x)$ and homogeneous Dirichlet boundary conditions is formulated in Section 2. In Section 3 we transform this problem to a new one with a constant coefficient $\tilde{a}(x) = 1$, so that its solution exactly coincides with the solution to the original model problem. In Section 4 we infer a condition for DMP for the original model problem. Finally, Section 5 discusses the case of mixed Dirichlet–Neumann boundary conditions.

2. Model problem and its discretization

We solve the one-dimensional equation with homogeneous Dirichlet boundary conditions

$$\begin{aligned} -(a(x)u(x)')' &= f(x) \qquad \text{in } \Omega \\ u(\alpha) &= u(\beta) = 0 \end{aligned}$$

in an interval $\Omega = (\alpha, \beta) \subset \mathbb{R}$. The corresponding weak formulation reads: Find $u \in V$ such that

$$\mathscr{B}(u,v) = (f,v)_\Omega \qquad \forall v \in V \tag{2.1}$$

where $V = H_0^1(\Omega) = \{v \in H^1(\Omega) : v(\alpha) = v(\beta) = 0\}$, $a \in L^\infty(\Omega)$ is piecewise-constant, $f \in L^2(\Omega)$ is a right-hand side, and

$$(f,v)_\Omega = \int_\Omega f(x)v(x)\,\mathrm{d}x, \qquad \mathscr{B}(u,v) = \int_\Omega a(x)u'(x)v'(x)\,\mathrm{d}x.$$

As usual, we create a partition $\alpha = x_0 < x_1 < \cdots < x_M = \beta$ of the domain Ω consisting of M elements $K_i = [x_{i-1}, x_i]$, $i = 1,2,\ldots,M$. Every element K_i is assigned an arbitrary polynomial degree $p_i \geqslant 1$. Moreover, we assume the piecewise-constant coefficient a to be aligned with this partition. The corresponding finite element space of continuous and piecewise-polynomial functions $V_{hp} \subset V$ has the form

$$V_{hp} = \left\{ v_{hp} \in V;\ v_{hp}|_{K_i} \in P^{p_i}(K_i),\ i = 1,2,\ldots,M \right\}.$$

Here $P^{p_i}(K_i)$ stands for the space of polynomials of degree at most p_i on the element K_i. The space V_{hp} has the dimension $N = -1 + \sum_{i=1}^{M} p_i$. There exists a unique finite element function $u_{hp} \in V_{hp}$ satisfying

$$\mathscr{B}(u_{hp}, v_{hp}) = (f, v_{hp}) \qquad \forall v_{hp} \in V_{hp}. \tag{2.2}$$

3. Transformed problem

The coefficient $a = a(x)$ is considered to be piecewise-constant with respect to the partion of Ω, i.e., there exist constants a_i such that

$$a|_{K_i} = a_i, \qquad i = 1,2,\ldots,M.$$

We will transform the model problem (2.1) to a standard Poisson equation in a different domain $\tilde{\Omega} = (\tilde{\alpha}, \tilde{\beta})$ and with a different right-hand side. The right-hand side $\tilde{f}$, the domain $\tilde{\Omega}$, and the partition of $\tilde{\Omega}$ will be determined later. The Poisson equation has the form

$$\begin{aligned} -\tilde{u}''(\tilde{x}) &= \tilde{f}(\tilde{x}) \quad \text{in } \tilde{\Omega} \\ \tilde{u}(\tilde{\alpha}) = \tilde{u}(\tilde{\beta}) &= 0. \end{aligned}$$

The weak formulation reads: Find $\tilde{u} \in \tilde{V}$ such that

$$\tilde{\mathscr{B}}(\tilde{u}, \tilde{v}) = (\tilde{f}, \tilde{v})_{\tilde{\Omega}} \qquad \forall \tilde{v} \in \tilde{V} \tag{3.1}$$

where $\tilde{V} = H_0^1(\tilde{\Omega}) = \{\tilde{v} \in H^1(\tilde{\Omega}) : \tilde{v}(\tilde{\alpha}) = \tilde{v}(\tilde{\beta}) = 0\}$, $\tilde{f} \in L^2(\tilde{\Omega})$,

$$(\tilde{f}, \tilde{v})_{\tilde{\Omega}} = \int_{\tilde{\Omega}} \tilde{f}(\tilde{x})\tilde{v}(\tilde{x})\,\mathrm{d}\tilde{x}, \qquad \tilde{\mathscr{B}}(\tilde{u}, \tilde{v}) = \int_{\tilde{\Omega}} \tilde{u}'(\tilde{x})\tilde{v}'(\tilde{x})\,\mathrm{d}\tilde{x}.$$

We construct a partition $\tilde{\alpha} = \tilde{x}_0 < \tilde{x}_1 < \cdots < \tilde{x}_M = \tilde{\beta}$ of the domain $\tilde{\Omega}$ consisting of M elements $\tilde{K}_i = [\tilde{x}_{i-1}, \tilde{x}_i]$, $i = 1,2,\ldots,M$. Every element $\tilde{K}_i$ is assigned a polynomial

degree $p_i \geqslant 1$. Notice that the number of elements as well as the polynomial degrees are exactly the same as for the original problem (2.2).

The corresponding finite element space $\tilde{V}_{hp} \subset \tilde{V}$ is given by

$$\tilde{V}_{hp} = \left\{ \tilde{v}_{hp} \in \tilde{V};\ \tilde{v}_{hp}|_{\tilde{K}_i} \in P^{p_i}(\tilde{K}_i),\ i = 1,2,\ldots,M \right\}.$$

Clearly, $\dim \tilde{V}_{hp} = \dim V_{hp} = N$. Finally, the finite element solution $\tilde{u}_{hp} \in \tilde{V}_{hp}$ to problem (3.1) is uniquely given by requirement

$$\tilde{\mathscr{B}}(\tilde{u}_{hp}, \tilde{v}_{hp}) = (\tilde{f}, \tilde{v}_{hp})_{\tilde{\Omega}} \qquad \forall \tilde{v}_{hp} \in \tilde{V}_{hp}. \tag{3.2}$$

Now let us link the discrete problem (2.2) to the discrete Poisson problem (3.2). There is a single degree of freedom which is the left endpoint $\tilde{\alpha} \in \mathbb{R}$, all remaining data to problem (3.2) are uniquely determined by the data to the original problem (2.2). First let us define the lengths of the new elements,

$$\tilde{h}_i = h_i / a_i, \qquad i = 1,2,\ldots,M \tag{3.3}$$

where $h_i = x_i - x_{i-1}$. The points of the new partition of $\tilde{\Omega}$ are given by

$$\tilde{x}_i = \tilde{\alpha} + \sum_{k=1}^{i} \tilde{h}_k, \qquad i = 1,2,\ldots,M.$$

Moreover, we put $\tilde{x}_0 = \tilde{\alpha}$ and $\tilde{\beta} = \tilde{x}_M$. For future reference let us define affine transformations of elements K_i to elements $\tilde{K}_i$ by

$$\eta_i(x) = \frac{\tilde{h}_i}{h_i}(x - x_{i-1}) + \tilde{x}_{i-1}, \qquad i = 1,2,\ldots,M. \tag{3.4}$$

Finally, the right-hand side to the transformed problem (3.2) is defined in an element-by-element fashion as

$$\tilde{f}(\tilde{x})|_{\tilde{K}_i} = a_i f(x)|_{K_i}$$

where $x = \eta_i^{-1}(\tilde{x})$, $i = 1,2,\ldots,M$.

Our subsequent results are based on the following Lemma 3.1 and Theorem 3.1.

Lemma 3.1. *Let $u, v \in H^1(\Omega)$ and $\tilde{u}, \tilde{v} \in H^1(\tilde{\Omega})$ be functions satisfying*

$$u(x)|_{K_i} = \tilde{u}(\tilde{x})|_{\tilde{K}_i}, \qquad v(x)|_{K_i} = \tilde{v}(\tilde{x})|_{\tilde{K}_i}$$

where $\tilde{x} = \eta_i(x)$ and $i = 1,2,\ldots,M$. Then

$$\mathscr{B}(u,v) = \tilde{\mathscr{B}}(\tilde{u},\tilde{v}), \qquad (f,v)_{\Omega} = (\tilde{f},\tilde{v})_{\tilde{\Omega}}.$$

Proof. Let us calculate

$$\mathscr{B}(u,v)=\sum_{i=1}^{M}a_i\int_{K_i}u'(x)v'(x)\,\mathrm{d}x=\sum_{i=1}^{M}a_i\frac{\tilde{h}_i}{h_i}\int_{\tilde{K}_i}\tilde{u}'(\tilde{x})\tilde{v}'(\tilde{x})\,\mathrm{d}\tilde{x}=\tilde{\mathscr{B}}(\tilde{u},\tilde{v})$$

where we have used (3.4) for the substitution in the integral. Similarly,

$$(f,v)_\Omega=\sum_{i=1}^{M}\int_{K_i}f(x)v(x)\,\mathrm{d}x=\sum_{i=1}^{M}\frac{h_i}{\tilde{h}_i}\frac{1}{a_i}\int_{\tilde{K}_i}\tilde{f}(\tilde{x})\tilde{v}(\tilde{x})\,\mathrm{d}\tilde{x}=(\tilde{f},\tilde{v})_{\tilde{\Omega}}.\qquad\square$$

Theorem 3.1. *Let u_{hp} and $\tilde{u}_{hp}$ be solutions to problems* (2.2) *and* (3.2)*, respectively. Then*

$$u_{hp}(x)|_{K_i}=\tilde{u}_{hp}(\tilde{x})|_{\tilde{K}_i}\tag{3.5}$$

where $\tilde{x}=\eta_i(x)$ and $i=1,2,\ldots,M$.

Proof. Let $u_{hp}\in V_{hp}$ be the unique solution to (2.2). Let us use the transformations (3.4) to define $\tilde{u}^*_{hp}\in\tilde{V}_{hp}$ as

$$\tilde{u}^*_{hp}(\tilde{x})|_{\tilde{K}_i}=u_{hp}(x)|_{K_i}$$

where $x=\eta_i^{-1}(\tilde{x})$ and $i=1,2,\ldots,M$. Further, let $\tilde{v}_{hp}\in\tilde{V}_{hp}$ be arbitrary. Similarly, we define $v_{hp}\in V_{hp}$ by

$$v_{hp}(x)|_{K_i}=\tilde{v}_{hp}(\tilde{x})|_{\tilde{K}_i}$$

where $\tilde{x}=\eta_i(x)$ and $i=1,2,\ldots,M$.

Now Lemma 3.1 and equality (2.2) imply

$$\tilde{\mathscr{B}}(\tilde{u}^*_{hp},\tilde{v}_{hp})=\mathscr{B}(u_{hp},v_{hp})=(f,v_{hp})_\Omega=(\tilde{f},\tilde{v}_{hp})_{\tilde{\Omega}}.$$

Thus, $\tilde{u}^*_{hp}\in\tilde{V}_{hp}$ satisfies (3.2) for all $\tilde{v}_{hp}\in\tilde{V}_{hp}$. Since the solution $\tilde{u}_{hp}\in\tilde{V}_{hp}$ to problem (3.2) is unique, we have $\tilde{u}^*_{hp}=\tilde{u}_{hp}$ and the proof is finished. $\square$

4. Discrete maximum principle

In this section we will use existing results for the Poisson equation to infer a condition for the discrete maximum principle for the original problem (2.2). First let us recall the definition of the discrete maximum principle:

Definition 4.1. Problem (2.2) satisfies the *Discrete Maximum Principle* (DMP) if

$$f\leqslant 0\ \text{ a.e. in }\Omega\quad\Longrightarrow\quad\max_{\overline{\Omega}}u_{hp}=\max_{\partial\Omega}u_{hp}$$

where $\partial\Omega$ is the boundary of the domain Ω.

Table 1.
The values of $H^*_{\mathrm{rel}}(p)$ for $p = 1,2,3,\ldots,20$.

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

Let us define a fundamental quantity $H^*_{\mathrm{rel}}(p)$:

$$H^*_{\mathrm{rel}}(p) = 1 \qquad \text{for } p = 1$$

$$H^*_{\mathrm{rel}}(p) = 1 + \frac{1}{2} \min_{(\xi,\eta)\in[-1,1]^2} l_0(\xi) l_0(\eta) \sum_{k=2}^{p} \varkappa_k(\xi)\varkappa_k(\eta) \qquad \text{for } p \geqslant 2.$$

Here, $l_0(\xi) = (1-\xi)/2$ and

$$\varkappa_k(\xi) = \sqrt{\frac{2k-1}{2}}\, \frac{4}{k(1-k)}\, P'_{k-1}(\xi)$$

where $P_k(\xi)$ stand for the Legendre polynomials of degree k. See Table 1 for values of $H^*_{\mathrm{rel}}(p)$ for $1 \leqslant p \leqslant 20$. The values of $H^*_{\mathrm{rel}}(p)$ up to $p = 100$ are depicted in Fig. 1. Notice that the smallest value for $p \leqslant 100$ is $9/10$. This value can be calculated analytically.

Theorem 4.1. *Let $\tilde{\alpha} = \tilde{x}_0 < \tilde{x}_1 < \cdots < \tilde{x}_M = \tilde{\beta}$ be a partition of the domain $\tilde{\Omega} = (\tilde{\alpha}, \tilde{\beta})$ and let $p_i \geqslant 1$ be polynomial degrees assigned to the elements $\tilde{K}_i = [\tilde{x}_{i-1}, \tilde{x}_i]$, $i = 1,2,\ldots,M$. If*

$$\frac{\tilde{x}_i - \tilde{x}_{i-1}}{\tilde{\beta} - \tilde{\alpha}} \leqslant H^*_{\mathrm{rel}}(p_i), \qquad i = 1,2,\ldots,M \tag{4.1}$$

then problem (3.2) *satisfies the discrete maximum principle.*

Proof. This result is completely proven in [18]. For the readers' convenience we sketch the main ideas.

The discrete Green's function $\tilde{G}_{hp,\tilde{y}} \in \tilde{V}_{hp}$, $\tilde{y} \in \overline{\tilde{\Omega}}$, corresponding to problem (3.2) is defined in a standard way as the unique solution to

$$\tilde{\mathscr{B}}(\tilde{v}_{hp}, \tilde{G}_{hp,\tilde{y}}) = \tilde{v}_{hp}(\tilde{y}) \qquad \forall\, \tilde{v}_{hp} \in \tilde{V}_{hp}.$$

We use notation $\tilde{G}_{hp}(\tilde{x},\tilde{y}) = \tilde{G}_{hp,\tilde{y}}(\tilde{x})$, where $(\tilde{x},\tilde{y}) \in \overline{\tilde{\Omega}}^2$. The classical representation formula

$$\tilde{u}_{hp}(\tilde{y}) = \int_{\tilde{\Omega}} \tilde{G}_{hp}(\tilde{x},\tilde{y})\, \tilde{f}(\tilde{y})\, \mathrm{d}\tilde{x}$$

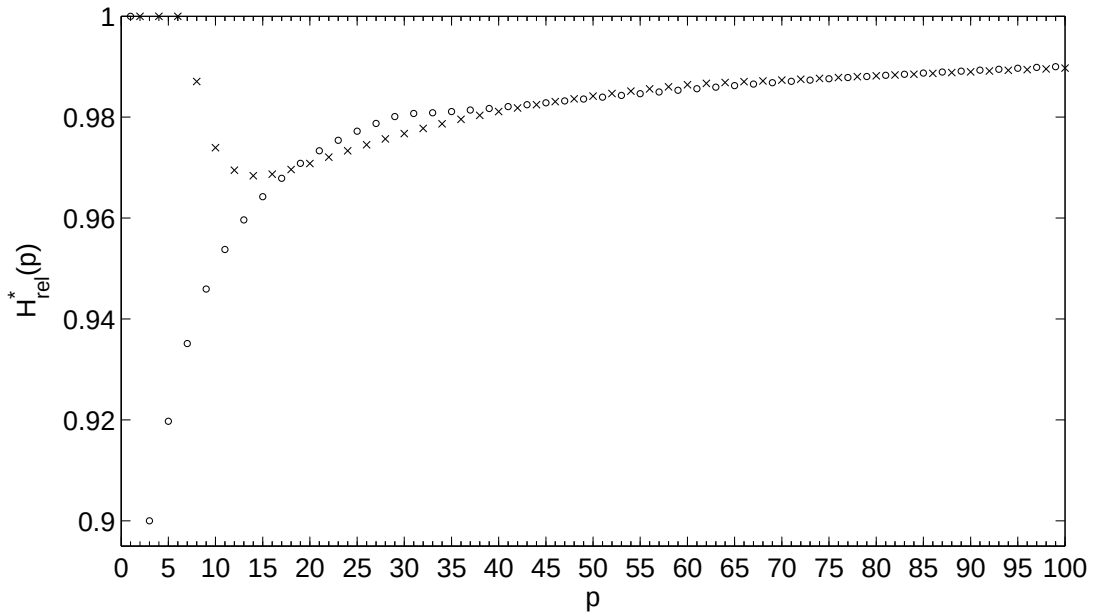


Figure 1. The values of $H^*_{\mathrm{rel}}(p)$ for $p = 1,2,\ldots,100$. Circles and crosses indicate the odd and even values of p, respectively.

proves that the nonnegativity of $\tilde{G}_{hp}$ in $\overline{\tilde{\Omega}}^2$ is equivalent to the discrete comparison principle which is further equivalent to the DMP. We recall that the discrete comparison principle for problem (3.2) states that any $\tilde{f} \geqslant 0$ implies $\tilde{u}_{hp} \geqslant 0$.

To prove the nonnegativity of the discrete Green's function we reveal that $\tilde{G}_{hp}$ can be expressed as

$$\tilde{G}_{hp}(\tilde{x},\tilde{y}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \tilde{A}_{ij}^{-1}\tilde{\varphi}_i(\tilde{x})\tilde{\varphi}_j(\tilde{y})$$

where $\tilde{\varphi}_i$, $i = 1,2,\ldots,N$, form a basis in $\tilde{V}_{hp}$, $\tilde{A}_{ij} = \tilde{\mathscr{B}}(\tilde{\varphi}_j,\tilde{\varphi}_i)$ denote entries of the stiffness matrix and $\tilde{A}_{ij}^{-1}$ stand for entries of its inverse. In the case of 1D Poisson equation, the entries of the inverse stiffness matrix can be expressed explicitly. Moreover the orthogonality (in the energy sense) of the piecewise linear basis functions to the higher order basis functions enables to split the discrete Green's function $\tilde{G}_{hp}$ into the linear and higher order parts $\tilde{G}_{hp} = \tilde{G}^L_{hp} + \tilde{G}^B_{hp}$. While the linear part $\tilde{G}^L_{hp}$ is nonnegative by a standard argument, the higher order part $\tilde{G}^B_{hp}$ is not nonnegative for most polynomial degrees. Finally, a detailed and technically demanding analysis of $\tilde{G}^L_{hp}$ and $\tilde{G}^B_{hp}$ reveals that under condition (4.1) the linear part overcomes the higher order part and the discrete Green's function is nonnegative. □

We propose to generalize condition (4.1) to problems with a piecewise-constant coefficient $a(x)$ in the following way.

Theorem 4.2. *Let $\alpha = x_0 < x_1 < \cdots < x_M = \beta$ be a partition of the domain $\Omega = (\alpha, \beta)$, let $p_i \geqslant 1$ be polynomial degrees assigned to the elements $K_i = [x_{i-1}, x_i]$, let $h_i = x_i - x_{i-1}$, and let a_i stand for the constant values of a in K_i, $i = 1, 2, \ldots, M$. If*

$$\frac{\dfrac{h_i}{a_i}}{\displaystyle\sum_{k=1}^{M} \frac{h_k}{a_k}} \leqslant H^*_{\mathrm{rel}}(p_i), \qquad i = 1, 2, \ldots, M \tag{4.2}$$

then problem (2.2) *satisfies the discrete maximum principle.*

Proof. This is a simple consequence of (3.3) and Theorems 3.1 and 4.1. □

Note that condition (4.2) can easily be verified in an element-by-element fashion and that it can be written in a simple way using the notation from Section 3:

$$\frac{\tilde{h}_i}{\tilde{\beta} - \tilde{\alpha}} \leqslant H^*_{\mathrm{rel}}(p_i), \qquad i = 1, 2, \ldots, M.$$

5. Mixed boundary conditions

Next let us consider the problem from Section 2, equipped with a Neumann boundary condition at β:

$$\begin{aligned} -(a(x)u(x)')' &= f(x) \qquad \text{in } \Omega \\ u(\alpha) &= 0 \\ u'(\beta) &= g(\beta). \end{aligned}$$

The weak formulation reads: Find $u \in V$ such that

$$\mathscr{B}(u, v) = (f, v)_\Omega + g(\beta)v(\beta) \qquad \forall v \in V \tag{5.1}$$

where $V = \{v \in H^1(\Omega) : v(\alpha) = 0\}$, $a \in L^\infty(\Omega)$ is piecewise-constant, and $f \in L^2(\Omega)$.

We proceed analogously to Section 2 to obtain the finite element solution $u_{hp} \in V_{hp}$,

$$\mathscr{B}(u_{hp}, v_{hp}) = (f, v_{hp})_\Omega + g(\beta)v(\beta) \qquad \forall v_{hp} \in V_{hp}. \tag{5.2}$$

Now the dimension of the space $V_{hp} \subset V$ is greater by one compared to the space V_{hp} which was defined in Section 2. The original problem is transformed similarly to what was done in Section 3. The Neumann boundary data for the new problem are the same as for the original problem, i.e.,

$$\tilde{g}(\tilde{\beta}) = g(\beta).$$

Clearly, it follows from Lemma 3.1 that

$$(f,v)_\Omega + g(\beta)v(\beta) = (\tilde{f},\tilde{v})_{\tilde{\Omega}} + \tilde{g}(\tilde{\beta})\tilde{v}(\tilde{\beta})$$

for $v(x)|_{K_i} = \tilde{v}(\tilde{x})|_{\tilde{K}_i}$, where $\tilde{x} = \eta_k(x)$ and $k = 1,2,\ldots,M$. Thus, analogously to Theorem 3.1 we obtain

$$u_{hp}(x)|_{K_i} = \tilde{u}_{hp}(\tilde{x})|_{\tilde{K}_i}.$$

The Neumann data $g(\beta)$ as well as the right-hand side f enter the definition of the discrete maximum principle for problem (5.2).

Definition 5.1. Problem (5.2) satisfies the *discrete maximum principle* if

$$f \leqslant 0 \ \text{ a.e. in } \Omega \text{ and } g(\beta) \leqslant 0 \quad \Longrightarrow \quad \max_{\overline{\Omega}} u_{hp} = \max_{\partial\Omega} u_{hp}.$$

It was proven in [19] that the Poisson equation with mixed boundary conditions satisfies the DMP if

$$H^*_{\mathrm{rel}}(p_i) \geqslant 0, \qquad i = 1,2,\ldots,M. \tag{5.3}$$

This condition is problem-independent. Since the discrete solution to the transformed problem is equal to the discrete solution to the problem with piecewise-constant coefficient, we conclude that the DMP is valid for problem (5.2).

Theorem 5.1. *If condition* (5.3) *is satisfied, then problem* (5.2) *with mixed boundary conditions and piecewise-constant coefficient* $a(x)$ *satisfies the discrete maximum principle.*

Condition (5.3) is satisfied at least for all $p \leqslant 100$ (see Table 1 and Fig. 1). Hence, the discrete maximum principle for 1D problems of type (5.1) with mixed boundary conditions and with piecewise-constant coefficient $a(x)$ is satisfied on arbitrary meshes and with arbitrary distribution of polynomial degrees (not exceeding 100).

Remark 5.1. Above, homogeneous Dirichlet boundary conditions were considered for simplicity only. The result on the DMP holds for nonhomogeneous conditions as well. Indeed, we always can consider a harmonic Dirichlet lift γ satisfying general Dirichlet boundary conditions and $-(a(x)\gamma'(x))' = 0$. To obtain the solution $\hat{u}_{hp}$ to the problem with general Dirichlet boundary conditions, we just add this lift to the solution u_{hp} with homogeneous Dirichlet conditions, i.e., $\hat{u}_{hp} = \gamma + u_{hp}$. Notice that the lift γ satisfies the classical maximum principle. Thus, in the case of general Dirichlet boundary conditions, we have $u_{hp} = 0$ on $\partial\Omega$ and

$$\max_{\Omega}(u_{hp}+\gamma) \leqslant \max_{\Omega} u_{hp} + \max_{\Omega}\gamma = \max_{\partial\Omega} u_{hp} + \max_{\partial\Omega}\gamma = \max_{\partial\Omega}(u_{hp}+\gamma).$$

Hence, the solution $\hat{u}_{hp}$ satisfies the DMP.

Moreover, in the piecewise-constant coefficient case, the lift γ is piecewise-linear and continuous, and it can be expressed explicitly as

$$\begin{aligned}\gamma(x) &= C_1 \int_\alpha^x 1/a(s)\,\mathrm{d}s + C_2 \\ &= C_1 \left[\sum_{k=1}^{i-1} \frac{h_k}{a_k} + \frac{x - x_{i-1}}{a_i} \right] + C_2, \quad x \in K_i, \quad i = 1,2,\ldots,M\end{aligned}$$

where C_1 and C_2 are integration constants to be determined from the boundary conditions. If a Dirichlet condition is prescribed at the left endpoint, i.e., $\hat{u}_{hp}(\alpha) = g_\alpha$, then $C_2 = g_\alpha$. Similarly, condition $\hat{u}_{hp}(\beta) = g_\beta$ implies $C_1 = (g_\beta - g_\alpha)/\left(\sum_{k=1}^M h_k/a_k\right)$. For the case of mixed boundary conditions, the Dirichlet lift is constant, i.e., $\gamma = g_\alpha$. This follows from the requirement $\gamma'(\beta) = 0$ which implies $C_1 = 0$. Since the lift γ is constant, we conclude that the DMP is valid even for the case of mixed and nonhomogeneous boundary conditions.

Remark 5.2. The long term goal is to generalize the 1D results to higher space dimension. This is, however, a very demanding task. According to the authors' knowledge there is no 2D nor 3D positive result for DMP for higher order FEM available, yet. It is surprising, since the problematics of DMP already is studied for several decades and numerical results indicate that the DMP is satisfied provided the mesh is reasonably fine. Nevertheless, the general framework of the discrete Green's function suggested in this paper can well be used in higher space dimension, too.

We identify the following two points of the 1D proof that are not available in higher space dimension. First, the vertex basis functions are not orthogonal (in the energy sense) to the higher order basis functions in general. Second, there is no hope to compute explicitly the entries of the inverse stiffness.

To overcome the fist point, we can modify the vertex functions by suitable combination of the higher order basis functions in such a way that the modified vertex functions are orthogonal to the higher order functions. These modified vertex functions are supported in the whole domain Ω and under suitable conditions on the mesh they stay nonnegative. Interestingly enough, under further conditions on the mesh the stiffness matrix corresponding to the modified vertex functions will be M-matrix and hence it will have nonnegative inverse. Moreover the theory of M-matrices enables to estimate the entries of the inverse to an M-matrix from below which is sufficient to overcome the second point.

References

1. I. Babuška and B. Q. Guo, Approximation properties of the *hp* version of the finite element method. *Comp. Meth. Appl. Mech. Engrg.* (1996) **133**, 319 – 346.

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

3. P. G. Ciarlet, Discrete maximum principle for finite difference operators. *Aequationes Math.* (1970) **4**, 338 – 352.
4. P. G. Ciarlet and P. A. Raviart, Maximum principle and uniform convergence for the finite element method. *Comp. Meth. Appl. Mech. Engrg.* (1973) **2**, 17 – 31.
5. A. Drăgănescu, T. F. Dupont, and L. R. Scott, Failure of the discrete maximum principle for an elliptic finite element problem. *Math. Comp.* (2005) **74**, 1 – 23 (electronic).
6. W. Höhn and H. D. Mittelmann, Some remarks on the discrete maximum principle for finite elements of higher-order. *Computing* (1981) **27**, 145 – 154.
7. A. Jüngel and A. Unterreiter, Discrete minimum and maximum principles for finite element approximations of non-monotone elliptic equations. *Numer. Math.* (2005) **99**, 485 – 508.
8. J. Karátson and S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions. *Numer. Math.* (2005) **99**, 669 – 698.
9. S. Korotov, M. Křížek, and P. Neittaanmäki, Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. *Math. Comp.* (2000) **70**, 107 – 119.
10. A. H. Schatz, 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.
11. C. Schwab, *p- and hp-Finite Element Methods*. Clarendon Press, Oxford, 1998.
12. P. Šolín, *Partial Differential Equations and the Finite Element Method*. J. Wiley & Sons, 2005.
13. P. Šolín, K. Segeth, and I. Doležel, *Higher-Order Finite Element Methods*. Chapman & Hall/CRC Press, Boca Raton, 2003.
14. P. Šolín and T. Vejchodský, A weak discrete maximum principle for *hp*-FEM. *J. Comp. Appl. Math.* (2006) (to appear).
15. B. Szabó and I. Babuška, *Finite Element Analysis*. John Wiley & Sons, New York, 1991.
16. T. Vejchodský, On the nonnegativity conservation in semidiscrete parabolic problems. In: *Conjugate Gradients Algorithms and Finite Element Methods* (Eds. M. Křížek, P. Neittaanmäki, R. Glowinski and S. Korotov), Berlin, Springer-Verlag, 2004, pp. 197 – 210.
17. T. Vejchodský, Method of lines and conservation of nonnegativity. In: *European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS 2004* (Eds. P. Neittaanmäki, T. Rossi, S. Korotov, E. Onate, J. Périaux, and D. Knörzer), Jyväskylä, 24 – 28 July 2004. http://www.mit.jyu.fi/eccomas2004/
18. T. Vejchodský and P. Šolín, Discrete Maximum Principle for Higher-Order Finite Elements in 1D. *Math. Comp.* (2007) **76**, 1833 – 1846.
19. T. Vejchodský and P. Šolín, Discrete maximum principle for Poisson equation with mixed boundary conditions Solved by *hp*-FEM. *Research Report No. 2006-09*, Department of Math. Sciences, University of Texas at El Paso, April 2006.
20. T. Vejchodský and P. Šolín, Discrete Green's function and maximum principles. In: *Programs and Algorithms of Numerical Mathematics 13* (Eds. J. Chleboun, K. Segeth, and T. Vejchodský), Mathematical Institute ASCR, Prague, 2006, pp. 247 – 252.
21. E. G. Yanik, Sufficient conditions for a discrete maximum principle for high-order collocation methods. *Comp. Math. Appl.* (1989) **17**, 1431 – 1434.
22. J. Xu and L. Zikatanov, A monotone finite element scheme for convection–diffusion equations. *Math. Comp.* (1999) **68**, 1429 – 1446.

APPENDIX

G

Higher-order discrete maximum principle for 1D diffusion-reaction problems

Below we attach a copy of the paper

[A6] T. Vejchodský: Higher-order discrete maximum principle for 1D diffusion-reaction problems. *Appl. Numer. Math.* **60** (2010), 486–500.

Applied Numerical Mathematics 60 (2010) 486–500

Contents lists available at ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

Higher-order discrete maximum principle for 1D diffusion–reaction problems

Tomáš Vejchodský [1]

Institute of Mathematics, Czech Academy of Sciences, Žitná 25, CZ-115 67 Prague 1, Czech Republic

ARTICLE INFO

Article history:
Received 8 December 2008
Received in revised form 27 September 2009
Accepted 28 October 2009
Available online 4 November 2009

MSC:
65N30
65N50

Keywords:
Discrete maximum principle
Discrete Green's function
Diffusion–reaction problem
Higher-order finite element method
hp-FEM
M-matrix

ABSTRACT

Sufficient conditions for the validity of the discrete maximum principle (DMP) for a 1D diffusion–reaction problem $-u'' + \kappa^2 u = f$ with homogeneous Dirichlet boundary conditions discretized by the higher-order finite element method are presented. It is proved that the DMP is satisfied if the lengths h of all elements are shorter then one-third of the length of the entire domain and if $\kappa^2 h^2$ is small enough for all elements. In general, the bounds for $\kappa^2 h^2$ depend on the polynomial degree of the elements, on h, and on the size of the domain. The obtained conditions are simple and easy to verify. A technical assumption (nonnegativity of certain rational functions) was verified by computer for polynomial degrees up to 10. The paper contains an analysis of the discrete Green's function which can be of independent interest.

1. Introduction

The standard (continuous) maximum principles for elliptic and parabolic problems, in particular, guarantee the nonnegativity of the solution provided that the data are nonnegative. This is especially important if naturally nonnegative quantities like temperature, concentration, density, etc., are modelled. It is an important question whether the discretization of these problems satisfies the discrete maximum principle (DMP) as well, or, equivalently, if the resulting discrete solution is guaranteed to be nonnegative provided the data are nonnegative.

Unfortunately, the standard methods, e.g., the finite element methods, do not satisfy the DMP in general. Therefore, additional conditions for the validity of the DMP are proposed and studied. Up to the author's knowledge the paper of Varga [16] from 1966 was the first paper about the DMP. Since then many other papers about the DMP for various problems and various discretizations were published [3,4,7,8,12,21].

Interestingly, the majority of the published works deal with the lowest-order approximations only. The results about the DMP for higher-order approximations are scarce, see [1,11,22] and the recent works of the author and his coauthors [17,19]. In [17] we deal with the 1D Poisson problem. The current paper generalizes this result to the 1D diffusion–reaction problem

E-mail address: vejchod@math.cas.cz.

[1] The author was supported by Grants No. IAA100760702 and No. IAA100190803 of the Grant Agency of the Czech Academy of Sciences, and by the institutional research plan No. AV0Z10190503 of the Czech Academy of Sciences.

0168-9274/$30.00
doi:10.1016/j.apnum.2009.10.009

discretized by higher-order finite elements. In particular, these results are suitable for the *hp*-version of the finite element method (*hp*-FEM), see e.g. [18], because various polynomial degrees in different elements are allowed.

The generalization of the higher-order DMP from the Poisson problem to the diffusion–reaction problem is not straightforward. Many technical problems have to be overcome and new approaches introduced. For illustration let us mention that in contrast to the Poisson problem the bubble (interior) basis functions are not orthogonal to the vertex functions in the diffusion–reaction case, there is no explicit formula for the inverse of the stiffness matrix, the reaction coefficient κ^2 is a new free parameter, etc. Even for the lowest-order approximations, DMPs for diffusion–reaction problems were only treated very recently [2,9].

The paper is organized as follows. Section 2 introduces the diffusion–reaction problem and briefly describes its discretization by the *hp*-FEM. In Section 3 the discrete maximum principle is defined and its relation to the discrete Green's function [5,6] is explained. The useful concept of discrete minimum energy extensions is introduced in Section 4 and it is used in Section 5 to define suitable basis functions for the higher-order finite element space. The splitting of the discrete Green's function to the vertex and bubble part is shown in Section 6 together with the proof of the nonnegativity of the vertex part. Section 7 analyzes the influence of the bubble part to the nonnegativity of the discrete Green's function in several steps. Sufficient conditions for the DMP are presented here. Section 8 comments the technical assumptions and their verification. The computer was used to verify nonnegativity of certain rational functions on an interval. The final conclusions are drawn in Section 9.

2. The problem and its discretization

Let us consider an open interval $\Omega \subset \mathbb{R}$, $\Omega = (a_\Omega, b_\Omega)$, and the 1D reaction–diffusion problem with the homogeneous Dirichlet boundary conditions

$$-u'' + \kappa^2 u = f \quad \text{in } \Omega, \qquad u(a_\Omega) = u(b_\Omega) = 0, \tag{1}$$

where the reaction coefficient $\kappa \geqslant 0$ is assumed to be constant and the right-hand side f to be square integrable. The standard maximum principle for this problem is equivalent to the so-called conservation of nonnegativity

$$f \geqslant 0 \quad \Rightarrow \quad u \geqslant 0.$$

In what follows, we will study an analogue of this implication for the discrete solution obtained by the *hp*-FEM.

Let $a_\Omega = x_0 < x_1 < \cdots < x_{M+1} = b_\Omega$ be a partition of the interval $\Omega = (a_\Omega, b_\Omega)$. Consider $M+1 \geqslant 2$ finite elements $K_k = [x_{k-1}, x_k]$ with lengths $h_{K_k} = x_k - x_{k-1}$, $k = 1, 2, \ldots, M+1$. In the sequel, we will often omit the subscript K_k and we will simply use h for h_{K_k}. The set $\mathcal{T}_{hp} = \{K_k,\ k = 1, 2, \ldots, M+1\}$ is referred as the (finite element) mesh. Further, we consider an arbitrary distribution of polynomial degrees p_K assigned to the elements $K \in \mathcal{T}_{hp}$. The corresponding *hp*-FEM space V_{hp} is defined as follows

$$V_{hp} = \left\{ v_{hp} \in H_0^1(\Omega):\ v_{hp}|_K \in P^{p_K}(K),\ K \in \mathcal{T}_{hp} \right\}, \tag{2}$$

where $H_0^1(\Omega)$ is the standard Sobolev space of functions from $L^2(\Omega)$ with the generalized derivatives in $L^2(\Omega)$ which vanish on the boundary. The space $P^{p_K}(K)$ contains polynomials of degree at most $p_K \geqslant 1$ in the interval K. The *hp*-FEM solution $u_{hp} \in V_{hp}$ of problem (1) is defined by

$$a(u_{hp}, v_{hp}) = F(v_{hp}) \quad \forall v_{hp} \in V_{hp}, \tag{3}$$

where $a(u, v) = (u', v')_\Omega + \kappa^2 (u, v)_\Omega$, $F(v) = (f, v)_\Omega$, and $(u, v)_\Omega = \int_\Omega uv \, dx$ denotes the $L^2(\Omega)$ inner product. Notice that there exists a unique solution $u_{hp} \in V_{hp}$ to problem (3).

3. Discrete maximum principle and the discrete Green's function

Definition 3.1. Let V_{hp} given by (2) be the *hp*-FEM space based on the mesh $\mathcal{T}_{hp}$ and on the polynomial degrees p_K, $K \in \mathcal{T}_{hp}$. We say that approximate problem (3) satisfies the discrete maximum principle (DMP) if

$$\max_{\overline{\Omega}} u_{hp} = \max_{\partial\Omega} u_{hp} = 0 \quad \text{for all } f \in L^2(\Omega),\ f \leqslant 0 \text{ a.e. in } \Omega. \tag{4}$$

Notice that requirement (4) is equivalent to

$$u_{hp} \geqslant 0 \quad \text{for all } f \in L^2(\Omega),\ f \geqslant 0 \text{ a.e. in } \Omega. \tag{5}$$

This DMP is also equivalent to the nonnegativity of the discrete Green's function G_{hp}, see Theorem 3.2 below.

Definition 3.2. Let $y \in \Omega$ and let $G_{hp,y} \in V_{hp}$ be the unique solution of the problem

$$a(w_{hp}, G_{hp,y}) = \delta_y(w_{hp}) = w_{hp}(y) \quad \forall w_{hp} \in V_{hp}. \tag{6}$$

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

A combination of (3) and (6) yields the discrete Kirchhoff–Helmholtz representation formula

$$u_{hp}(y) = \int_\Omega G_{hp}(x, y) f(x)\,\mathrm{d}x, \quad y \in \Omega. \tag{7}$$

Interestingly, the DGF can be explicitly expressed in terms of a basis of V_{hp}.

Theorem 3.1. *Let $\varphi_1, \varphi_2, \ldots, \varphi_N$ be a basis in V_{hp} and let $\mathbb{A} \in \mathbb{R}^{N\times N}$ be the stiffness matrix with entries $\mathbb{A}_{ij} = a(\varphi_i, \varphi_j)$, $i, j = 1, 2, \ldots, N$. Then*

$$G_{hp}(x, y) = \sum_{i=1}^{N}\sum_{j=1}^{N}\big(\mathbb{A}^{-1}\big)_{ij}\varphi_i(x)\varphi_j(y), \tag{8}$$

where $(\mathbb{A}^{-1})_{ij}$ are the entries of the inverse matrix to $\mathbb{A}$.

Proof. See [17]. □

Consequently, Theorem 3.1 and the symmetry of the bilinear form $a(\cdot,\cdot)$ imply $G_{hp}(x, y) = G_{hp}(y, x)$. Notice that $G_{hp,x} = G_{hp}(x, \cdot) \in V_{hp}$.

Theorem 3.2. *Problem* (3) *satisfies the DMP if and only if $G_{hp}(x, y) \geqslant 0$ for all $(x, y) \in \Omega^2$.*

Proof. Immediate consequence of (7). See [17]. □

Thus, our goal is to prove the nonnegativity of G_{hp} in Ω^2. To this end, we will use (8). First, in Section 5, a suitable basis of V_{hp} will be constructed. For this purpose we will utilize the concept of the discrete minimum energy extensions which will be described in Section 4. The analysis of the nonnegativity of G_{hp} will be postponed to the subsequent sections.

4. Discrete minimum energy extensions

Let us consider a splitting of the space V_{hp} into a direct sum of two nontrivial subspaces $V_{hp} = V^*_{hp} \oplus V^\#_{hp}$. The discrete minimum energy extension $\psi^{\mathrm{me}} \in V_{hp}$ of a function $\psi^* \in V^*_{hp}$ with respect to $V^\#_{hp}$ is uniquely defined as

$$\psi^{\mathrm{me}} = \psi^* - \psi^\#,$$

where $\psi^\# \in V^\#_{hp}$ is the elliptic projection of ψ^* into $V^\#_{hp}$, i.e.,

$$0 = a\big(\psi^{\mathrm{me}}, v^\#\big) = a\big(\psi^* - \psi^\#, v^\#\big) \quad \text{for all } v^\# \in V^\#_{hp}. \tag{9}$$

Equally well, ψ^{me} is the component of ψ^* a-orthogonal to $V^\#_{hp}$. Definition (9) implies inequality $\|\psi^{\mathrm{me}}\| \leqslant \|\psi^* + v^\#\|$ for all $v^\# \in V^\#_{hp}$, where $\|v\|^2 = a(v, v)$ stands for the energy norm. This explains why we call ψ^{me} discrete minimum energy extension. Further, due to the symmetry of $a(\cdot,\cdot)$ and due to (9) we have

$$a\big(\psi^{\mathrm{me}}, \psi^{\mathrm{me}}\big) = a\big(\psi^{\mathrm{me}}, \psi^*\big) = a\big(\psi^*, \psi^*\big) - a\big(\psi^\#, \psi^*\big) = a\big(\psi^*, \psi^*\big) - a\big(\psi^\#, \psi^\#\big).$$

Hence, $\|\psi^{\mathrm{me}}\|^2 + \|\psi^\#\|^2 = \|\psi^*\|^2$. Consequently,

$$\big\|\psi^{\mathrm{me}}\big\| \leqslant \big\|\psi^*\big\| \quad \text{and} \quad \big\|\psi^\#\big\| \leqslant \big\|\psi^*\big\|. \tag{10}$$

Now, let us compute the discrete minimum energy extensions of basis functions from V^*_{hp}. Let $\mathcal{B}^* = \{\varphi^*_1, \varphi^*_2, \ldots, \varphi^*_{N^*}\}$ be a basis in V^*_{hp} and let $\mathcal{B}^\# = \{\varphi^\#_1, \varphi^\#_2, \ldots, \varphi^\#_{N^\#}\}$ be a basis in $V^\#_{hp}$. The stiffness matrix $\overline{\mathbb{A}}$ corresponding to the basis $\mathcal{B}^* \cup \mathcal{B}^\#$ of V_{hp} has the following 2-by-2 block structure

$$\overline{\mathbb{A}} = \begin{pmatrix} \overline{A} & \overline{B} \\ \overline{B}^T & \overline{D} \end{pmatrix},$$

where $\overline{A}_{ij} = a(\varphi_i^*, \varphi_j^*)$, $i, j = 1, 2, \ldots, N^*$, $\overline{B}_{ij} = a(\varphi_i^*, \varphi_j^\#)$, $i = 1, 2, \ldots, N^*$, $j = 1, 2, \ldots, N^\#$, and $\overline{D}_{ij} = a(\varphi_i^\#, \varphi_j^\#)$, $i, j = 1, 2, \ldots, N^\#$.

The discrete minimum energy extensions $\varphi_i^{\mathrm{me}} \in V_{hp}$ of $\varphi_i^* \in V_{hp}^*$ with respect to $V_{hp}^\#$ can be computed as

$$\varphi_i^{\mathrm{me}} = \varphi_i^* - \sum_{j=1}^{N^\#} \overline{C}_{ij} \varphi_j^\#, \quad i = 1, 2, \ldots, N^*. \tag{11}$$

The requirement (9) uniquely determines coefficients $\overline{C}_{ij}$ as follows:

$$0 = a\big(\varphi_i^*, \varphi_k^\#\big) - \sum_{j=1}^{N^\#} \overline{C}_{ij} a\big(\varphi_j^\#, \varphi_k^\#\big) \quad \forall i = 1, 2, \ldots, N^*,\ k = 1, 2, \ldots, N^\#. \tag{12}$$

This can be formulated in a matrix form as $0 = \overline{B} - \overline{C}\,\overline{D}$, where the matrix $\overline{C} \in \mathbb{R}^{N^* \times N^\#}$ consists of entries $\overline{C}_{ij}$. Hence,

$$\overline{C} = \overline{B}\,\overline{D}^{-1}. \tag{13}$$

The discrete minimum energy extensions $\varphi_i^{\mathrm{me}} \in V_{hp}$ can be used as an alternative basis $\mathcal{B}^{\mathrm{me}} = \{\varphi_1^{\mathrm{me}}, \varphi_2^{\mathrm{me}}, \ldots, \varphi_{N^*}^{\mathrm{me}}\}$ in V_{hp}^*. It can be easily verified that the corresponding stiffness matrix $\overline{S} \in \mathbb{R}^{N^* \times N^*}$ with entries $\overline{S}_{ij} = a(\varphi_i^{\mathrm{me}}, \varphi_j^{\mathrm{me}})$, $i, j = 1, 2, \ldots, N^*$ is just the Schur complement

$$\overline{S} = \overline{A} - \overline{B}\,\overline{D}^{-1}\overline{B}^T. \tag{14}$$

Finally, the well-known formula for the inversion of a 2-by-2 block matrix implies that the upper-left block of $\overline{\mathbb{A}}^{-1}$ is equal to the inverse of the Schur complement, i.e.,

$$\big(\overline{\mathbb{A}}^{-1}\big)_{ij} = \big(\overline{S}^{-1}\big)_{ij} \quad \forall i, j = 1, 2, \ldots, N^*. \tag{15}$$

5. Construction of the *hp*-FEM bases

As usual, we construct the finite element basis functions on elements $K_k \in \mathcal{T}_{hp}$ as images of the shape functions defined on the reference element $K_{\mathrm{ref}} = [-1, 1]$ under the reference maps

$$\chi_{K_k}(\xi) = \frac{h_{K_k}}{2}\xi + \frac{x_k + x_{k-1}}{2}, \quad \xi \in K_{\mathrm{ref}},\ k = 1, 2, \ldots, M+1. \tag{16}$$

The *hp*-FEM shape functions comprise two vertex functions

$$\ell_0(\xi) = (1 - \xi)/2, \qquad \ell_1(\xi) = (1 + \xi)/2, \quad \xi \in K_{\mathrm{ref}},$$

and $p - 1$ bubble functions ℓ_i^p, $i = 2, \ldots, p$, for each polynomial degree p. For the analysis of the DMP it is convenient to construct the bubble functions as the generalized eigenfunctions of the discrete Laplacian [20]. Hence, for given $p \geqslant 2$ we define $\ell_i^p \in \mathbb{P}_0^p(K_{\mathrm{ref}})$ by requirement

$$\big((\ell_i^p)', v'\big)_{K_{\mathrm{ref}}} = \lambda_i^p \big(\ell_i^p, v\big)_{K_{\mathrm{ref}}} \quad \forall v \in \mathbb{P}_0^p(K_{\mathrm{ref}}),$$

where $\mathbb{P}_0^p(K_{\mathrm{ref}})$ stands for the space of polynomials of degree at most p which vanish at the endpoints of the interval K_{ref}. For every polynomial degree $p \geqslant 2$ there exists $p - 1$ distinct positive eigenvalues $\lambda_2^p < \lambda_3^p < \cdots < \lambda_p^p$. The corresponding eigenfunctions ℓ_i^p, $i = 2, \ldots, p$, are orthogonal in both $H_0^1(K_{\mathrm{ref}})$- and $L^2(K_{\mathrm{ref}})$-inner products and they are normalized such that

$$\big((\ell_i^p)', (\ell_i^p)'\big)_{K_{\mathrm{ref}}} = 1/2 \quad \text{and} \quad \big(\ell_i^p, \ell_i^p\big)_{K_{\mathrm{ref}}} = 1/\big(2\lambda_i^p\big), \quad i = 2, 3, \ldots, p. \tag{17}$$

Finally, since each polynomial ℓ_i^p has roots ± 1, we can factor out these root factors and define the corresponding kernels $\mathcal{K}_i^p$ as follows:

$$\ell_i^p(\xi) = \ell_0(\xi)\ell_1(\xi)\mathcal{K}_i^p(\xi), \quad i = 2, \ldots, p,\ p \geqslant 2. \tag{18}$$

To define the basis of V_{hp} we transform the shape functions from the reference element K_{ref} to the physical elements $K \in \mathcal{T}_{hp}$ using the reference mapping (16). The standard piecewise linear vertex functions φ_k are constructed for $k = 1, 2, \ldots, M$ as follows:

$$\varphi_k(x) = \begin{cases} \ell_1(\chi_{K_k}^{-1}(x)), & \text{for } x \in K_k, \\ \ell_0(\chi_{K_{k+1}}^{-1}(x)), & \text{for } x \in K_{k+1}, \\ 0, & \text{otherwise.} \end{cases}$$

The $N-M$ bubble functions $\varphi_{M+1},\dots,\varphi_N$, where $N=-1+\sum_{K\in\mathcal{T}_{hp}} p_K$ is the dimension of V_{hp}, are defined in a similar way. The p_K-1 bubble functions $\varphi_2^{b,K},\varphi_3^{b,K},\dots,\varphi_{p_K}^{b,K}$ in an element K are constructed as

$$\varphi_i^{b,K}(x) = \begin{cases} \ell_i^{p_K}(\chi_K^{-1}(x)), & \text{for } x \in K, \\ 0, & \text{otherwise,} \end{cases} \quad i=2,3,\dots,p_K. \tag{19}$$

As usual, we assemble the stiffness matrix $\mathbb{A}\in\mathbb{R}^{N\times N}$, $\mathbb{A}_{ij}=a(\varphi_i,\varphi_j)$, $i,j=1,2,\dots,N$, from the local stiffness matrices $\mathbb{A}^K\in\mathbb{R}^{(p_K+1)\times(p_K+1)}$, $K\in\mathcal{T}_{hp}$. The entries of $\mathbb{A}^K$ can be computed as follows:

$$\mathbb{A}_{ij}^K = \frac{2}{h_K}\big(\ell'_{i-1},\ell'_{j-1}\big)_{K_{\mathrm{ref}}} + \frac{h_K}{2}\kappa^2(\ell_{i-1},\ell_{j-1})_{K_{\mathrm{ref}}}, \quad i,j=1,\dots,p_K+1,$$

where $\ell_i=\ell_i^{p_K}$ for $i=2,3,\dots,p_K$.

Due to the existence of the vertex and bubble functions, the matrices $\mathbb{A}$ and $\mathbb{A}^K$ have a natural 2-by-2 block structure

$$\mathbb{A}=\begin{pmatrix} A & B \\ B^T & D \end{pmatrix} \quad\text{and}\quad \mathbb{A}^K=\begin{pmatrix} A^K & B^K \\ (B^K)^T & D^K \end{pmatrix},$$

where $A\in\mathbb{R}^{M\times M}$, $B\in\mathbb{R}^{M\times(N-M)}$, and $D\in\mathbb{R}^{(N-M)\times(N-M)}$, $A^K\in\mathbb{R}^{2\times 2}$, $B^K\in\mathbb{R}^{2\times(p_K-1)}$, and $D^K\in\mathbb{R}^{(p_K-1)\times(p_K-1)}$. The entries of the local stiffness matrix $\mathbb{A}^K$ can be easily computed. If the length of the element K is denoted by h then

$$hA^K=\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}+\frac{\kappa^2h^2}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{20}$$

The entries of B^K depend on the polynomial degree p_K in a nontrivial way. There are no explicit formulas for them, but they can be computed easily as follows:

$$hB_{ij}^K=\kappa^2h^2\overline{B}_{ij}^{p_K}, \quad \text{where } \overline{B}_{ij}^{p_K}=\frac{1}{2}\big(\ell_{i-1},\ell_{j+1}^{p_K}\big)_{K_{\mathrm{ref}}}, \tag{21}$$

$i=1,2$, $j=1,\dots,p_K-1$. Notice that the values $\overline{B}_{ij}^{p_K}$ are independent from h and κ^2. The final block D^K is diagonal with entries

$$hD_{ii}^K=2\big(\big(\ell_{i+1}^{p_K}\big)',\big(\ell_{i+1}^{p_K}\big)'\big)_{K_{\mathrm{ref}}}+\kappa^2h^2\frac{1}{2}\big(\ell_{i+1}^{p_K},\ell_{i+1}^{p_K}\big)_{K_{\mathrm{ref}}}=1+\kappa^2h^2\mu_i^{p_K}, \tag{22}$$

where $\mu_i^{p_K}=1/(4\lambda_{i+1}^{p_K})$ is independent of h and κ^2, see (17), and $i=1,2,\dots,p_K-1$.

We remark that it is convenient to multiply the formulas for A^K, B^K, and D^K by h because then the entries of matrices hA^K, hB^K, and hD^K are functions of a single parameter $\zeta=\kappa^2h^2$.

To prove the DMP it is convenient to introduce the discrete minimum energy extensions $\psi_1,\psi_2,\dots,\psi_M$ of the vertex functions $\varphi_1,\varphi_2,\dots,\varphi_M$ with respect to the space of all bubbles $V_{hp}^b=\operatorname{span}\{\varphi_i^{b,K},\ i=2,3,\dots,p_K,\ K\in\mathcal{T}_{hp}\}$.

In analogy with (11) we express

$$\psi_i=\varphi_i-\sum_{\substack{K\in\mathcal{T}_{hp}\\ K\subset\operatorname{supp}\varphi_i}}\sum_{j=1}^{p_K-1}C_{\iota_K(i),j}^K\varphi_{j+1}^{b,K}, \quad i=1,2,\dots,M, \tag{23}$$

where $\iota_K(i)$ is the standard connectivity mapping, see, e.g., [18]. In our case $\iota_K(i)=1$ if ψ_i corresponds to the left endpoint of K and $\iota_K(i)=2$ if ψ_i corresponds to the right endpoint of K. The matrix C^K of coefficients $C_{\iota_K(i),j}^K$ is given by (13) as $C^K=B^K(D^K)^{-1}$ and hence, putting $\zeta=\kappa^2h_K^2$, the entries of C^K can be expressed by (21) and (22) as

$$C_{mj}^K=\zeta\,\overline{B}_{mj}^{p_K}\big(1+\zeta\,\mu_j^{p_K}\big)^{-1}, \quad m=1,2,\ j=1,\dots,p_K. \tag{24}$$

Thus, if the vertex function ψ_i is supported in an element $K\in\mathcal{T}_{hp}$ then we can transform it to the reference element K_{ref} as follows:

$$\overline{\psi}_{m-1}(\xi)=\psi_i\big(\chi_K(\xi)\big)=\ell_{m-1}(\xi)-\sum_{j=1}^{p_K-1}C_{mj}^K\ell_{j+1}^{p_K}(\xi)=\ell_{m-1}(\xi)\Psi_{m-1}^{p_K}(\zeta,\xi), \tag{25}$$

where $m = \iota_K(i) \in \{1,2\}$, $\xi \in K_{\mathrm{ref}}$, $\zeta = \kappa^2 h_K^2$, and by (18) and (24) we obtain

$$\Psi_{m-1}^{p_K}(\zeta,\xi) = 1 - \ell_{2-m}(\xi)\zeta \sum_{j=1}^{p_K-1} \overline{B}_{mj}^{p_K}\left(1+\zeta\mu_j^{p_K}\right)^{-1} \mathcal{K}_{j+1}^{p_K}(\xi). \tag{26}$$

Notice that $\Psi_0^p(\zeta,\xi) = \Psi_1^p(\zeta,-\xi)$ for $\zeta \geqslant 0$ and $\xi \in K_{\mathrm{ref}}$, because each generalized eigenfunction $\ell_j^p(\xi)$ is either odd or even.

Further, by (9) the discrete minimum energy extensions $\psi_1, \psi_2, \ldots, \psi_M$ are orthogonal to all bubbles $\varphi_i^{b,K}$, $i = 2,3,\ldots,p_K$, $K \in \mathcal{T}_{hp}$, where the orthogonality is understood in the energy inner product $a(\cdot,\cdot)$. Hence, by (14) the stiffness matrix $\mathbb{S} \in \mathbb{R}^{N\times N}$ formed from the discrete minimum energy extensions $\psi_1, \psi_2, \ldots, \psi_M$ and from the eigenfunctions $\varphi_i^{b,K}$, $i = 2,3,\ldots,p_K$, $K \in \mathcal{T}_{hp}$, has the following structure

$$\mathbb{S} = \begin{pmatrix} S & 0 \\ 0 & D \end{pmatrix}, \tag{27}$$

where $S = A - BD^{-1}B^T$ stands for the Schur complement and D is diagonal.

6. Nonnegativity of the discrete Green's function

The DGF corresponding to problem (3) can be expressed by (8) using the discrete minimum energy extensions ψ_i, $i = 1,2,\ldots,M$, see (23), and the bubble functions $\varphi_i^{b,K}$, $i = 2,3,\ldots,p_K$, $K \in \mathcal{T}_{hp}$, see (19). Thanks to the structure of the stiffness matrix $\mathbb{S}$, see (27), we can express the DGF as a sum of the vertex and bubble parts

$$G_{hp}(x,y) = G_{hp}^v(x,y) + G_{hp}^b(x,y), \quad (x,y) \in \Omega^2, \tag{28}$$

where

$$G_{hp}^v(x,y) = \sum_{i=1}^{M}\sum_{j=1}^{M} S_{ij}^{-1}\psi_i(x)\psi_j(y), \quad (x,y) \in \Omega^2, \tag{29}$$

$$G_{hp}^b(x,y) = \sum_{K\in\mathcal{T}_{hp}} \sum_{i=1}^{p_K-1} \left(D_{ii}^K\right)^{-1} \varphi_{i+1}^{b,K}(x)\varphi_{i+1}^{b,K}(y), \quad (x,y) \in \Omega^2, \tag{30}$$

and the entries D_{ii}^K are given by (22).

The following theorem introduces three sufficient conditions for the nonnegativity of the DGF G_{hp}.

Theorem 6.1. *Let ψ_i, $i = 1,2,\ldots,M$, $S \in \mathbb{R}^{M\times M}$, and G_{hp} be given by* (23), (27), *and* (28)–(30), *respectively. If*

(a) $\psi_i(x) \geqslant 0$ *for all* $i = 1,2,\ldots,M$ *and* $x \in \Omega$,
(b) $S_{ij} \leqslant 0$ *for all* $i \neq j$, $i, j = 1,2,\ldots,M$,
(c) $G_{hp}^v + G_{hp}^b \geqslant 0$ *in* K^2 *for all* $K \in \mathcal{T}_{hp}$,

then $G_{hp}(x,y) \geqslant 0$ *for all* $(x,y) \in \Omega^2$.

Proof. By the theory of M-matrices, see, e.g., [15], if all offdiagonal entries of S are nonpositive and if S is symmetric and positive definite then S^{-1} consists of nonnegative entries, i.e, $(S^{-1})_{ij} \geqslant 0$ for all $i, j = 1,2,\ldots,M$. Hence, conditions (a) and (b) imply the nonnegativity of the vertex part G_{hp}^v in Ω^2, see (29). Since the support of any bubble function consists of a single element, we find that

$$G_{hp}^b(x,y) = 0 \quad \text{for } (x,y) \in K \times K^*,\ K \neq K^*,\ K, K^* \in \mathcal{T}_{hp}.$$

This together with (c) proves the nonnegativity of $G_{hp} = G_{hp}^v + G_{hp}^b$ in the entire square Ω^2. □

6.1. Nonnegativity of the vertex DGF

We present two lemmas which show the validity of conditions (a) and (b) from Theorem 6.1 provided that the products $\kappa^2 h_K^2$ are bounded from above by values α^{p_K} and β^{p_K} for all elements $K \in \mathcal{T}_{hp}$. The bounds α^{p_K} and β^{p_K} are given by

$$\alpha^p = \sup\left\{\overline{\zeta}\colon\ \Psi_1^p(\zeta,\xi) \geqslant 0 \text{ for all } \xi \in K_{\mathrm{ref}} \text{ and all } 0 \leqslant \zeta \leqslant \overline{\zeta}\right\}, \tag{31}$$

$$\beta^p = \sup\left\{\overline{\zeta}\colon\ q^p(\zeta) \leqslant 0 \text{ for all } 0 \leqslant \zeta \leqslant \overline{\zeta}\right\}, \tag{32}$$

where

$$q^p(\zeta) = -1 + \zeta/6 - \zeta^2 \sum_{i=1}^{p-1} \overline{B}^p_{1i} \overline{B}^p_{2i} \big(1 + \zeta \mu_i^p\big)^{-1}. \tag{33}$$

Notice that $q^p(\kappa^2 h^2) = h S^K_{12}$, where $S^K = A^K - B^K (D^K)^{-1} (B^K)^T \in \mathbb{R}^{2\times 2}$, see (20)–(22). Further notice that both α^p and β^p are positive due to the continuity of Ψ^p_1 and q^p and due to the fact that $\Psi^p_1(0,\xi) = 1$ and $q^p(0) = -1$.

Lemma 6.1. *Let h_K and p_K stand for the length and polynomial degree of the element $K \in \mathcal{T}_{hp}$. Further, let ψ_i, $i = 1, 2, \ldots, M$, be given by* (23)*. If*

$$\kappa^2 h_K^2 \leqslant \alpha^{p_K} \quad \text{for all } K \in \mathcal{T}_{hp},$$

then $\psi_i(x) \geqslant 0$ for all $x \in \Omega$, i.e., condition (a) *from Theorem* 6.1 *is satisfied.*

Proof. Let the vertex function ψ_i correspond to a nodal point x_i, $i = 1, 2, \ldots, M$. The nonnegativity of ψ_i in $K_i = [x_{i-1}, x_i]$ follows immediately from (25) and (26) with $m = 1$ and then from (31). The nonnegativity of ψ_i in $K_{i+1} = [x_i, x_{i+1}]$ follows symmetrically, because $\Psi^p_0(\zeta,\xi) = \Psi^p_1(\zeta,-\xi)$. □

Lemma 6.2. *Let h_K and p_K denote the length and the polynomial degree of the element $K \in \mathcal{T}_{hp}$. Further, let S be the Schur complement from* (27)*. If*

$$\kappa^2 h_K^2 \leqslant \beta^{p_K} \quad \text{for all } K \in \mathcal{T}_{hp},$$

then $S_{ij} \leqslant 0$ for all $i \neq j$, $i, j = 1, 2, \ldots, M$, i.e., condition (b) *from Theorem* 6.1 *is satisfied.*

Proof. Clearly, the matrix S is tridiagonal, hence, the only nonzero off-diagonal entries are

$$S_{k,k-1} = S_{k-1,k} = a_K(\psi_{k-1}, \psi_k) = S^K_{12}, \quad k = 2, 3, \ldots, M,$$

where ψ_{k-1} and ψ_k are the vertex functions corresponding to the endpoints of $K \in \mathcal{T}_{hp}$, S is given by (27), and $S^K = A^K - B^K (D^K)^{-1} (B^K)^T$ is the local Schur complement. The nonpositivity of the entry S^K_{12} follows immediately from (33) and (32). □

We remark that β^p can be computed as the smallest positive root of the polynomial $q^p(\zeta) \prod_{i=1}^{p-1} (1 + \zeta \mu_i^p)$. Similarly, the computation of the values α^p requires root finding of certain polynomials. The proof of Theorem 6.1 shows that conditions (a) and (b) imply nonnegativity of the vertex part G^v_{hp}. Thus, Lemmas 6.1 and 6.2 yield the following corollary.

Corollary 6.1. *Let $\mathcal{T}_{hp}$ be a finite element mesh and let h_K and p_K denote the length and the polynomial degree of the element $K \in \mathcal{T}_{hp}$. If*

$$\kappa^2 h_K^2 \leqslant \min\big\{\alpha^{p_K}, \beta^{p_K}\big\} \quad \text{for all } K \in \mathcal{T}_{hp},$$

then $G^v_{hp}(x, y) \geqslant 0$ for all $(x, y) \in \Omega^2$.

7. The bubble part of the DGF

In this section we study the validity of condition (c) from Theorem 6.1. The bubble part G^b_{hp} defined by (30) is not nonnegative, in general. For $p = 3$, $p = 5$, and $p \geqslant 7$, there always are regions, where G^b_{hp} is negative. In these regions, the negative bubble part G^b_{hp} has to be compensated by the positive vertex part G^v_{hp} in order to obtain the nonnegativity of $G_{hp} = G^v_{hp} + G^b_{hp}$ and consequently the DMP. Fortunately, condition (c) can be investigated for each element $K \in \mathcal{T}_{hp}$ independently.

Therefore, throughout this section, we consider an arbitrary but fixed element $K = [x_{k-1}, x_k]$ in $\mathcal{T}_{hp}$. The length and the polynomial degree of this K are denoted by h and p.

For this K we will define two auxiliary DGFs $\widetilde{G}_{hp}$ and $\widehat{G}_{hp}$, see Fig. 1 for an illustration. We will show that $\widehat{G}_{hp} \leqslant \widetilde{G}_{hp} \leqslant G_{hp}$ in K^2. The nonnegativity of the second auxiliary DGF $\widehat{G}_{hp}$ is investigated below in Section 7.4.

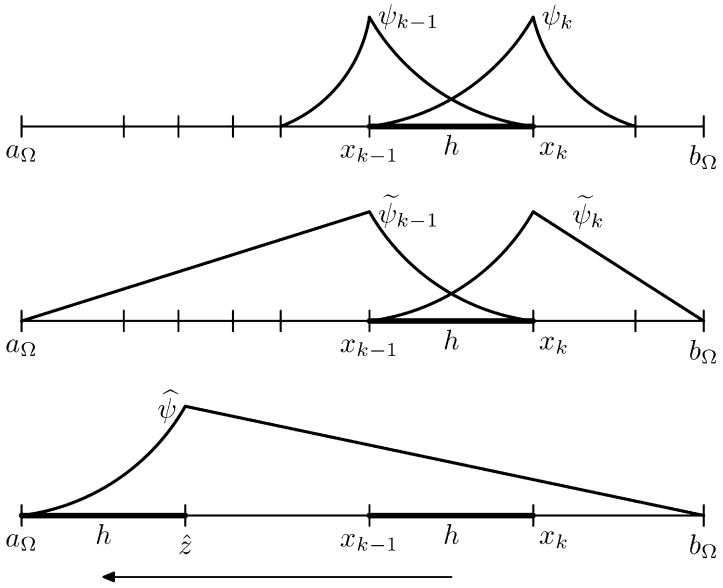


Fig. 1. An illustration of the basis functions used for the construction of G^v_{hp} (top), $\widetilde{G}^v_{hp}$ (middle), and $\widehat{G}^v_{hp}$ (bottom) corresponding to the element $K = [x_{k-1}, x_k]$ of length h.

7.1. The first auxiliary DGF

An element $K \in \mathcal{T}_{hp}$ is called interior if it is not adjacent to the boundary of Ω, i.e., if $K \subset \Omega$. We define the first auxiliary DGF $\widetilde{G}_{hp}$ for interior elements only. Thus, let $K = [x_{k-1}, x_k]$ be an interior element. We consider a partition $a_\Omega < x_{k-1} < x_k < b_\Omega$ which defines a mesh $\widetilde{\mathcal{T}}_{hp}$ consisting of three elements. The polynomial degree assigned to the element $K = [x_{k-1}, x_k] \in \widetilde{\mathcal{T}}_{hp}$ is p while the degree of the other two elements in $\widetilde{\mathcal{T}}_{hp}$ is set to 1. These polynomial degrees and the mesh $\widetilde{\mathcal{T}}_{hp}$ lead to an hp-FEM space $\widetilde{V}_{hp}$ defined in analogy with (2). In $\widetilde{V}_{hp}$ we consider two piecewise linear vertex functions $\widetilde{\varphi}_{k-1}$ and $\widetilde{\varphi}_k$ and $p-1$ bubble functions $\varphi_2^{b,K}, \varphi_3^{b,K}, \ldots, \varphi_p^{b,K}$, see (19). Notice that these bubble functions (generalized eigenfunctions of the Laplacian) coincide with the bubbles defined on the original mesh $\mathcal{T}_{hp}$.

Further, we consider the discrete minimum energy extensions $\widetilde{\psi}_{k-1}$ and $\widetilde{\psi}_k$ of $\widetilde{\varphi}_{k-1}$ and $\widetilde{\varphi}_k$ with respect to the space $V_{hp}^{b,K} = \operatorname{span}\{\varphi_2^{b,K}, \varphi_3^{b,K}, \ldots, \varphi_p^{b,K}\}$. Hence, $\widetilde{\psi}_{k-1}$ is linear in $[a_\Omega, x_{k-1}]$, $\widetilde{\psi}_{k-1} = \psi_{k-1}$ in $K = [x_{k-1}, x_k]$, and $\widetilde{\psi}_{k-1} = 0$ in $[x_k, b_\Omega]$. Similarly, $\widetilde{\psi}_k = 0$ in $[a_\Omega, x_{k-1}]$, $\widetilde{\psi}_k = \psi_k$ in $K = [x_{k-1}, x_k]$, and $\widetilde{\psi}_k$ is linear in $[x_k, b_\Omega]$. See the middle panel of Fig. 1.

We construct a stiffness matrix $\widetilde{A} \in \mathbb{R}^{2\times 2}$ from $\widetilde{\psi}_{k-1}$ and $\widetilde{\psi}_k$ as follows:

$$\widetilde{A}_{ij} = a(\widetilde{\psi}_{k-2+i}, \widetilde{\psi}_{k-2+j}), \quad i, j = 1, 2. \tag{34}$$

In agreement with (28)–(30), we define the first auxiliary DGF

$$\widetilde{G}_{hp}(x, y) = \widetilde{G}^v_{hp}(x, y) + \widetilde{G}^b_{hp}(x, y), \quad (x, y) \in \Omega^2, \tag{35}$$

where

$$\widetilde{G}^v_{hp}(x, y) = \sum_{i=1}^{2}\sum_{j=1}^{2} \widetilde{A}^{-1}_{ij} \widetilde{\psi}_{k-2+i}(x) \widetilde{\psi}_{k-2+j}(y), \quad (x, y) \in \Omega^2, \tag{36}$$

and $\widetilde{G}^b_{hp}(x, y) = G^b_{hp}(x, y)$, see (30).

The main result about $\widetilde{G}_{hp}(x, y)$ is formulated in the following lemma.

Lemma 7.1. *Let condition* (b) *from Theorem* 6.1 *be satisfied. For an interior element $K \in \mathcal{T}_{hp}$, $K = [x_{k-1}, x_k] \subset \Omega$, $k = 2, 3, \ldots, M$, consider the first auxiliary DGF $\widetilde{G}_{hp}$ defined by* (35)–(36) *and the DGF G_{hp} given by* (28)–(30)*. Then*

$$G_{hp}(x, y) \geqslant \widetilde{G}_{hp}(x, y) \quad \textit{for all } (x, y) \in K^2.$$

Proof. Clearly, it suffices to prove $G^{\nu}_{hp}(x,y) \geqslant \widetilde{G}^{\nu}_{hp}(x,y)$ for all $(x,y) \in K^2$. Let $K \in \mathcal{T}_{hp}$, $K = [x_{k-1}, x_k] \subset \Omega$, be an arbitrary but fixed interior element. First, we consider the original vertex functions $\psi_1, \psi_2, \ldots, \psi_M$. Let ψ^{me}_{k-1} and ψ^{me}_{k} be the discrete minimum energy extensions of ψ_{k-1} and ψ_k with respect to $V^{\nu\#}_{hp} = \operatorname{span}\{\psi_1, \ldots, \psi_{k-2}, \psi_{k+1}, \ldots, \psi_M\}$. Definition (11) yields $\psi^{\mathrm{me}}_{k-1}(x) = \psi_{k-1}(x)$ and $\psi^{\mathrm{me}}_{k}(x) = \psi_k(x)$ for all $x \in K$, because $\psi_j(x) = 0$ for all $x \in K$ and all $\psi_j \in V^{\nu\#}_{hp}$. Using the definition of $\widetilde{\psi}_{k-1}$ and $\widetilde{\psi}_k$, we summarize

$$\psi_{k-1} = \widetilde{\psi}_{k-1} = \psi^{\mathrm{me}}_{k-1} \quad \text{and} \quad \psi_k = \widetilde{\psi}_k = \psi^{\mathrm{me}}_{k} \quad \text{in } K. \tag{37}$$

The stiffness matrix $S^{\mathrm{me}} \in \mathbb{R}^{2\times 2}$ corresponding to the basis functions ψ^{me}_{k-1} and ψ^{me}_{k} can be computed as a suitable Schur complement, cf. (15).

Now, let us concentrate on $\widetilde{\psi}_{k-1}$ and $\widetilde{\psi}_k$. We remark that the discrete minimum energy extensions $\widetilde{\psi}^{\mathrm{me}}_{k-1}$ and $\widetilde{\psi}^{\mathrm{me}}_{k}$ of $\widetilde{\psi}_{k-1}$ and $\widetilde{\psi}_k$ with respect to $V^{\nu\#}_{hp}$ are equal to the already defined discrete minimum energy extensions ψ^{me}_{k-1} and ψ^{me}_{k}, respectively. Indeed, see (9), if $0 = a(\psi^{\mathrm{me}}_k, v^{\#}) = a(\widetilde{\psi}^{\mathrm{me}}_k, v^{\#})$ for all $v^{\#} \in V^{\nu\#}_{hp}$ then $0 = a(\psi^{\mathrm{me}}_k - \widetilde{\psi}^{\mathrm{me}}_k, v^{\#})$ for all $v^{\#} \in V^{\nu\#}_{hp}$ and since $\psi^{\mathrm{me}}_k - \widetilde{\psi}^{\mathrm{me}}_k \in V^{\nu\#}_{hp}$ then $\psi^{\mathrm{me}}_k = \widetilde{\psi}^{\mathrm{me}}_k$. The same steps can be repeated to show that $\psi^{\mathrm{me}}_{k-1} = \widetilde{\psi}^{\mathrm{me}}_{k-1}$.

From (37) we conclude that

$$\widetilde{A}_{12} = a_K(\widetilde{\psi}_{k-1}, \widetilde{\psi}_k) = a_K\big(\psi^{\mathrm{me}}_{k-1}, \psi^{\mathrm{me}}_{k}\big) = S^{\mathrm{me}}_{12}.$$

Similarly, from (10) we infer the inequalities

$$\begin{aligned}
\widetilde{A}_{11} &= a(\widetilde{\psi}_{k-1}, \widetilde{\psi}_{k-1}) \geqslant a\big(\widetilde{\psi}^{\mathrm{me}}_{k-1}, \widetilde{\psi}^{\mathrm{me}}_{k-1}\big) = a\big(\psi^{\mathrm{me}}_{k-1}, \psi^{\mathrm{me}}_{k-1}\big) = S^{\mathrm{me}}_{11},\\
\widetilde{A}_{22} &= a(\widetilde{\psi}_{k}, \widetilde{\psi}_{k}) \geqslant a\big(\widetilde{\psi}^{\mathrm{me}}_{k}, \widetilde{\psi}^{\mathrm{me}}_{k}\big) = a\big(\psi^{\mathrm{me}}_{k}, \psi^{\mathrm{me}}_{k}\big) = S^{\mathrm{me}}_{22}.
\end{aligned}$$

Hence, all entries of $\widetilde{A}$ are greater or equal to the corresponding entries of S^{me} and we write $\widetilde{A} \geqslant S^{\mathrm{me}}$. Condition (b) from Theorem 6.1 implies that both $\widetilde{A}$ and S^{me} are M-matrices. In particular, they have the nonnegative inverse and therefore $(S^{\mathrm{me}})^{-1} \geqslant \widetilde{A}^{-1}$. By this fact and by (29), (15), (37), we conclude

$$\begin{aligned}
G^{\nu}_{hp}(x,y) &= \sum_{i=1}^{M}\sum_{j=1}^{M}\big(S^{-1}\big)_{ij}\psi_i(x)\psi_j(y) = \sum_{i=1}^{2}\sum_{j=1}^{2}\big(S^{\mathrm{me}}\big)^{-1}_{ij}\psi^{\mathrm{me}}_{k-2+i}(x)\psi^{\mathrm{me}}_{k-2+j}(y)\\
&\geqslant \sum_{i=1}^{2}\sum_{j=1}^{2}\big(\widetilde{A}^{-1}\big)_{ij}\widetilde{\psi}_{k-2+i}(x)\widetilde{\psi}_{k-2+j}(y) = \widetilde{G}^{\nu}_{hp}(x,y)
\end{aligned}$$

for all $(x,y) \in K^2$. □

7.2. Analysis of the first auxiliary DGF

Let us analyze the auxiliary DGF $\widetilde{G}_{hp}$ in more detail. For arbitrary element $K \in \mathcal{T}_{hp}$, $K = [x_{k-1}, x_k]$, with the polynomial degree p and with the length h we introduce a parameter $t \in [0,1]$ such that

$$\begin{aligned}
x_{k-1} &= (1-t)a_\Omega + t(b_\Omega - h),\\
x_k &= (1-t)(a_\Omega + h) + t b_\Omega.
\end{aligned} \tag{38}$$

Clearly, the parameter t determines the position of K in $\Omega = (a_\Omega, b_\Omega)$. In addition, we define an auxiliary parameter $\theta \in (0,\infty]$ as

$$\theta = \frac{h}{|\Omega| - h}, \tag{39}$$

where $|\Omega| = b_\Omega - a_\Omega$ stands for the length of Ω.

Here, we restrict ourselves to the interior elements only, i.e., we assume $t \in (0,1)$. To express the stiffness matrix $\widetilde{A} \in \mathbb{R}^{2\times 2}$ assembled from $\widetilde{\psi}_{k-1}$ and $\widetilde{\psi}_k$ we introduce functions

$$r^p(\zeta) = hS^K_{11} = hS^K_{22} = 1 + \zeta/3 - \zeta^2\sum_{i=1}^{p-1}\big(\overline{B}^p_{1i}\big)^2\big(1+\zeta\mu^p_i\big)^{-1}, \tag{40}$$

where $\zeta = \kappa^2 h^2$, $S^K = A^K - B^K(D^K)^{-1}(B^K)^T$, cf. (27), and matrices A^K, B^K, D^K are given by (20)–(22). We remark that $(\overline{B}^p_{1i})^2 = (\overline{B}^p_{2i})^2$, because the generalized eigenfunctions of the Laplacian are either odd or even. Further we stress that $r^p(\zeta) = r^p(\kappa^2h^2) = a_K(\psi_{k-1}, \psi_{k-1}) = a_K(\psi_k, \psi_k) > 0$ for $h > 0$.

Using (33), (40), and the parameters t and θ, we can express $\widetilde{A}$ as

$$h\widetilde{A} = \begin{pmatrix} r^p(\kappa^2h^2) + \dfrac{\theta}{t} + \dfrac{\kappa^2h^2}{3}\dfrac{t}{\theta} & q^p(\kappa^2h^2) \\ q^p(\kappa^2h^2) & r^p(\kappa^2h^2) + \dfrac{\theta}{1-t} + \dfrac{\kappa^2h^2}{3}\dfrac{1-t}{\theta} \end{pmatrix}.$$

Our goal is to study the limit of $h\widetilde{A}$ for $t \to 0$. The entry $(h\widetilde{A})^{-1}_{11} \to 0$ for $t \to 0$ and, therefore, we concentrate on

$$s(t,\theta,\zeta) = (h\widetilde{A})^{-1}_{22} = \left(r^p(\zeta) + \frac{\theta}{1-t} + \frac{\zeta}{3}\frac{1-t}{\theta} - \frac{(q^p)^2(\zeta)}{r^p(\zeta) + \frac{\theta}{t} + \frac{\zeta}{3}\frac{t}{\theta}} \right)^{-1}, \tag{41}$$

which is well defined for $t \in (0,1)$, $\theta \in (0,\infty)$, and $\zeta = \kappa^2h^2 \in [0,\infty)$. For $t = 0$ we define $s(t,\theta,\zeta)$ by the following limit:

$$s(0,\theta,\zeta) = \lim_{t\to 0_+} s(t,\theta,\zeta) = \left(r^p(\zeta) + \theta + \frac{\zeta}{3\theta} \right)^{-1}. \tag{42}$$

Lemma 7.2. *If $s(t,\theta,\zeta)$ is defined by (41) and (42) then*

$$s(0,\theta,\zeta) \leqslant s(t,\theta,\zeta) \quad \text{for all } \theta \in (0,1/2],\ t \in [0,1/2],\ \zeta \in [0,\infty). \tag{43}$$

Proof. For $t > 0$ the inequality (43) is equivalent to

$$s^*(t,\theta,\zeta) = \frac{[s(t,\theta,\zeta)]^{-1} - [s(0,\theta,\zeta)]^{-1}}{t} = \frac{\theta}{1-t} - \frac{\zeta}{3\theta} - \frac{(q^p)^2(\zeta)}{r^p(\zeta)t + \theta + \frac{\zeta}{3}\frac{t^2}{\theta}} \leqslant 0. \tag{44}$$

Clearly, since $r^p(\zeta) > 0$, the function $s^*(t,\theta,\zeta)$ is increasing in the variable t. Hence,

$$s^*(t,\theta,\zeta) \leqslant s^*(1/2,\theta,\zeta) = 2\theta - \frac{\zeta}{3\theta} - \frac{(q^p)^2(\zeta)}{\frac{1}{2}r^p(\zeta) + \theta + \frac{\zeta}{12\theta}}. \tag{45}$$

Differentiating $s^*(1/2,\theta,\zeta)$ with respect to θ and using the fact that $\det(hS^K) = (r^p)^2(\zeta) - (q^p)^2(\zeta) > 0$, we find out that $s^*(1/2,\theta,\zeta)$ is increasing in θ. Thus,

$$s^*(1/2,\theta,\zeta) \leqslant s^*(1/2,1/2,\zeta) = 1 - \frac{2}{3}\zeta - \frac{2(q^p)^2(\zeta)}{r^p(\zeta) + 1 + \frac{\zeta}{3}}. \tag{46}$$

Similarly, it can be verified that $s^*(1/2,1/2,\zeta)$ is decreasing in ζ and, therefore,

$$s^*(1/2,1/2,\zeta) \leqslant s^*(1/2,1/2,0) = 0, \tag{47}$$

because $r^p(0) = 1$ and $q^p(0) = -1$. The combination of (44)–(47) finishes the proof. □

7.3. The second auxiliary DGF

In general, there are two second auxiliary DGF. The first one is adjacent to the left endpoint a_Ω and the second one is adjacent to the right endpoint b_Ω. For an interior element $K = [x_{k-1}, x_k] \in \mathcal{T}_{hp}$ we consider the parameter t given by (38). If $t \leqslant 1/2$ we define the second auxiliary DGF $\widehat{G}_{hp}$ as a limit of the first auxiliary DGF $\widetilde{G}_{hp}$ for $t \to 0_+$. If $t > 1/2$ then $\widehat{G}_{hp}$ is a limit of $\widetilde{G}_{hp}$ for $t \to 1_-$. However, the situation is symmetric and we can concentrate on the first case only without loss of generality.

If h and p stand for the length and polynomial degree of K then we set $\hat{z} = a_\Omega + h$ and consider two-element-mesh $\widehat{\mathcal{T}}_{hp}$ consisting of elements $\widehat{K} = [a_\Omega, \hat{z}]$ and $[\hat{z}, b_\Omega]$ with polynomial degrees p and 1, respectively. The hp-FEM basis on $\widehat{\mathcal{T}}_{hp}$ comprises one piecewise linear vertex function $\widehat{\varphi}$ and $p-1$ bubble functions $\widehat{\varphi}_2^{b,\widehat{K}}, \widehat{\varphi}_3^{b,\widehat{K}}, \ldots, \widehat{\varphi}_p^{b,\widehat{K}}$ supported in $\widehat{K}$.

As before, we define $\widehat{\psi}$ as the discrete minimum energy extension of the vertex function $\widehat{\varphi}$ with respect to the space of the bubbles $V_{hp}^{b,\widehat{K}} = \operatorname{span}\{\widehat{\varphi}_2^{b,\widehat{K}}, \widehat{\varphi}_3^{b,\widehat{K}}, \ldots, \widehat{\varphi}_p^{b,\widehat{K}}\}$, see the bottom panel of Fig. 1. Notice that $\widehat{\psi}$ is a linear function in $[\hat{z}, b_\Omega]$ and that $\widehat{\psi}$ restricted to $\widehat{K}$ is just the shifted function $\widetilde{\psi}_k = \psi_k$ restricted to K, i.e.,

$$\widehat{\psi}(x - x_{k-1} + a_\Omega) = \widetilde{\psi}_k(x) = \psi_k(x) \quad \text{for all } x \in K. \tag{48}$$

Furthermore, we can easily compute $a(\widehat{\psi},\widehat{\psi}) = [hs(0,\theta,\kappa^2h^2)]^{-1}$. Hence, in agreement with (28)–(30) we define

$$\widehat{G}_{hp}(x,y) = \widehat{G}^{v}_{hp}(x,y) + \widehat{G}^{b}_{hp}(x,y), \quad (x,y) \in \Omega^2, \tag{49}$$

$$\widehat{G}^{v}_{hp}(x,y) = hs\big(0,\theta,\kappa^2 h^2\big)\widehat{\psi}(x)\widehat{\psi}(y), \quad (x,y) \in \Omega^2, \tag{50}$$

$$\widehat{G}^{b}_{hp}(x,y) = \sum_{i=1}^{p-1} \big(D^{\widehat{K}}_{ii}\big)^{-1} \widehat{\varphi}^{b,\widehat{K}}_{i+1}(x)\widehat{\varphi}^{b,\widehat{K}}_{i+1}(y), \quad (x,y) \in \Omega^2. \tag{51}$$

We recall that the entries $D^{\widehat{K}}_{ii} = D^{K}_{ii}$ are given by (22).

For completeness, we also introduce the auxiliary DGFs for the elements adjacent to the boundary of Ω. For $K = [a_\Omega, x_1] \in \mathcal{T}_{hp}$ we define

$$\widetilde{G}_{hp}(x,y) = \widehat{G}_{hp}(x,y) \quad \text{for all } (x,y) \in \Omega^2, \tag{52}$$

where $\widehat{G}_{hp}(x,y)$ is given by (49)–(51) with $\widehat{K} = K$. For $K = [x_M, b_\Omega] \in \mathcal{T}_{hp}$ we define $\widehat{G}_{hp}(x,y) = \widetilde{G}_{hp}(x,y)$ symmetrically. The relation of the first and of the second auxiliary DGF explains the following lemma.

Lemma 7.3. *Let conditions* (a) *and* (b) *from Theorem* 6.1 *be satisfied. Further, let $K \in \mathcal{T}_{hp}$ be such that $t \leqslant 1/2$, see* (38)*, and let $\theta \leqslant 1/2$, see* (39)*. If $\widehat{G}_{hp}(x,y)$ and $\widetilde{G}_{hp}(x,y)$ are given by* (49)–(51) *and* (35)–(36) *with* (52)*, respectively, then*

$$\widehat{G}_{hp}(\hat{x},\hat{y}) \leqslant \widetilde{G}_{hp}(x,y) \quad \textit{for all } (x,y) \in K^2,$$

where $\hat{x} = x - x_{k-1} + a_\Omega$ and $\hat{y} = y - x_{k-1} + a_\Omega$.

Proof. First, if $K = [a_\Omega, x_1]$ then there is nothing to prove due to (52). The element K is not adjacent to the right endpoint due to the assumptions $t \leqslant 1/2$ and $\theta \leqslant 1/2$. Thus, it remains to consider the interior elements $K \in \mathcal{T}_{hp}$.

The bubble functions $\widehat{\varphi}^{b,\widehat{K}}_2, \widehat{\varphi}^{b,\widehat{K}}_3, \ldots, \widehat{\varphi}^{b,\widehat{K}}_p$ in $\widehat{K}$ are just shifted bubble functions $\varphi^{b,K}_2, \varphi^{b,K}_3, \ldots, \varphi^{b,K}_p$ from K, see (48). Therefore,

$$\widehat{G}^{b}_{hp}(\hat{x},\hat{y}) = \widetilde{G}^{b}_{hp}(x,y) \quad \text{for all } (x,y) \in K^2,$$

where $\hat{x} = x - x_{k-1} + a_\Omega$ and $\hat{y} = y - x_{k-1} + a_\Omega$. By (48)–(51), Lemma 7.2, the facts that $\widetilde{A}^{-1} \geqslant 0$, see (34), $\widetilde{\psi}_{k-1} \geqslant 0$ and $\widetilde{\psi}_k \geqslant 0$ in K, and by (36) we obtain

$$\begin{aligned}\widehat{G}^{v}_{hp}(\hat{x},\hat{y}) &= hs\big(0,\theta,\kappa^2 h^2\big)\widehat{\psi}(\hat{x})\widehat{\psi}(\hat{y}) \leqslant hs\big(t,\theta,\kappa^2 h^2\big)\widetilde{\psi}_k(x)\widetilde{\psi}_k(y) \\ &\leqslant \sum_{i=1}^{2}\sum_{j=1}^{2} \widetilde{A}^{-1}_{ij}\widetilde{\psi}_{k-2+i}(x)\widetilde{\psi}_{k-2+j}(y) = \widetilde{G}^{v}_{hp}(x,y)\end{aligned}$$

for all $(x,y) \in K^2$ with $\hat{x} = x - x_{k-1} + a_\Omega$ and $\hat{y} = y - x_{k-1} + a_\Omega$. □

Corollary 7.1. *Let G_{hp} be given by* (28)–(30)*. Further, let $\widehat{G}_{hp}$ given by* (49)–(52) *be the second auxiliary DGF corresponding to an element $K \in \mathcal{T}_{hp}$ and let $\theta \leqslant 1/2$, see* (39)*. If*

$$\widehat{G}_{hp}(\hat{x},\hat{y}) \geqslant 0 \quad \textit{for all } (\hat{x},\hat{y}) \in \hat{K}^2, \tag{53}$$

then

$$G_{hp}(x,y) \geqslant 0 \quad \textit{for all } (x,y) \in K^2,$$

i.e., the condition (c) *from Theorem* 6.1 *is satisfied.*

Proof. Let $K \in \mathcal{T}_{hp}$ be arbitrary. If $t \leqslant 1/2$, see (38), then assumption (53) and Lemmas 7.3 and 7.1 imply

$$0 \leqslant \widehat{G}_{hp}(\hat{x},\hat{y}) \leqslant \widetilde{G}_{hp}(x,y) \leqslant G_{hp}(x,y) \quad \forall (x,y) \in K^2,$$

where $\hat{x} = x - x_{k-1} + a_\Omega$ and $\hat{y} = y - x_{k-1} + a_\Omega$. The same conclusion is valid also for $t > 1/2$ due to the symmetry. □

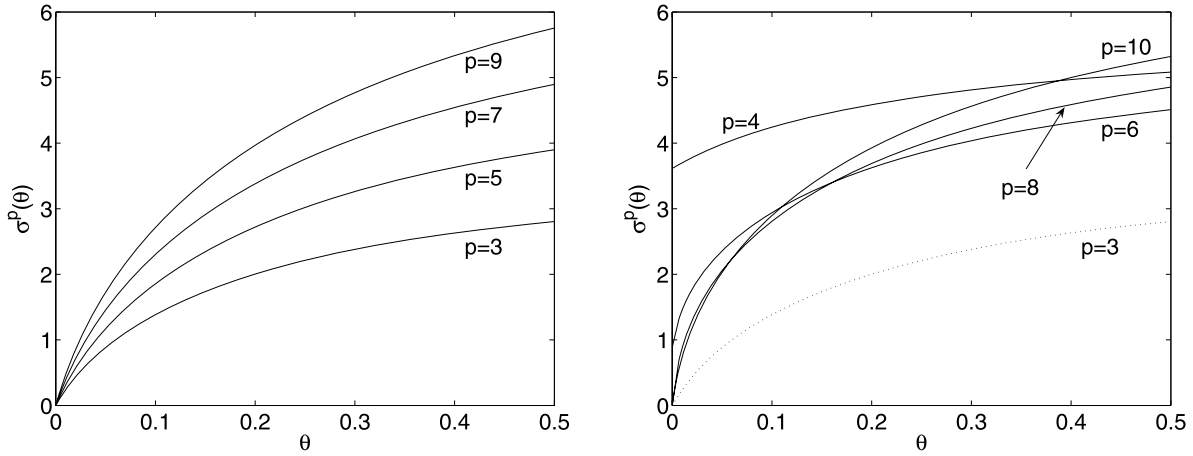


Fig. 2. The graphs of $\sigma^p(\theta)$ for $p = 3, 5, 7, 9$ (left) and for $p = 4, 6, 8, 10$ (right). The dotted line in the right panel shows $\sigma^3(\theta)$ to indicate that $\sigma^3(\theta) \leqslant \sigma^p(\theta)$ for all $p = 3, 4, \ldots, 10$, $\theta \in (0, 1/2]$.

7.4. Nonnegativity of the second auxiliary DGF

Finally, we will seek conditions for nonnegativity of $\widehat{G}_{hp}(\hat{x}, \hat{y})$ in $\widehat{K}^2$. It is convenient to transform $\widehat{G}_{hp}$ from $\widehat{K}^2 = [a_\Omega, \hat{z}]$ to $K^2_{\mathrm{ref}} = [-1, 1]^2$ using $\hat{x} = \chi_{\widehat{K}}(\xi)$ and $\hat{y} = \chi_{\widehat{K}}(\eta)$, where the reference map $\chi_{\widehat{K}}$ is given by (16). From (25) we have

$$\widehat{\psi}|_{\widehat{K}}\big(\chi_{\widehat{K}}(\xi)\big) = \overline{\psi}_1(\xi) = \ell_1(\xi)\Psi_1^p(\zeta, \xi), \tag{54}$$

where $\zeta = \kappa^2 h^2$, h is the length of K, p is the polynomial degree of K, and Ψ_1^p is given by (26).

With the help of (49)–(51), (54) and (42), the transformed DGF $\widehat{G}_{hp}$ can be expressed as follows

$$\begin{aligned} G^{\mathrm{ref}}_{hp}(\xi, \eta) &= \widehat{G}_{hp}\big(\chi_{\widehat{K}}(\xi), \chi_{\widehat{K}}(\eta)\big) = hs\big(0, \theta, \kappa^2 h^2\big)\widehat{\psi}\big(\chi_{\widehat{K}}(\xi)\big)\widehat{\psi}\big(\chi_{\widehat{K}}(\eta)\big) + h\sum_{i=1}^{p-1}\big(hD^K_{ii}\big)^{-1}\ell^p_{i+1}(\xi)\ell^p_{i+1}(\eta) \\ &= h\ell_1(\xi)\ell_1(\eta)\omega^p\big(\theta, \kappa^2 h^2, \xi, \eta\big), \end{aligned} \tag{55}$$

where, see (18), (22), and (25),

$$\omega^p(\theta, \zeta, \xi, \eta) = s(0, \theta, \zeta)\Psi_1^p(\zeta, \xi)\Psi_1^p(\zeta, \eta) + \ell_0(\xi)\ell_0(\eta)\mathrm{Ker}^{b,p}(\zeta, \xi, \eta), \tag{56}$$

$$\mathrm{Ker}^{b,p}(\zeta, \xi, \eta) = \sum_{i=1}^{p-1}\big(1 + \zeta\mu_i^p\big)^{-1}\mathcal{K}^p_{i+1}(\xi)\mathcal{K}^p_{j+1}(\eta), \tag{57}$$

and $\zeta = \kappa^2 h^2$. Finally, we define

$$\widehat{\omega}^p(\theta, \zeta) = \min_{(\xi,\eta)\in K^2_{\mathrm{ref}}} \omega^p(\theta, \zeta, \xi, \eta). \tag{58}$$

The motivation for this definition is clear. The second auxiliary DGF $\widehat{G}_{hp}(\hat{x}, \hat{y})$ is nonnegative in $\widehat{K}$ if and only if $\widehat{\omega}^p(\theta, \zeta) \geqslant 0$.

To analyze the nonnegativity of $\widehat{\omega}^p(\theta, \zeta)$, we set

$$\sigma^p(\theta) = \sup\big\{\bar{\zeta}\colon \widehat{\omega}^p(\theta, \zeta) \geqslant 0 \text{ for all } 0 \leqslant \zeta \leqslant \bar{\zeta}\big\}, \quad \text{where } \theta \in (0, 1/2].$$

By this definition, we immediately conclude that $\widehat{G}_{hp}$ is nonnegative provided $0 < \theta \leqslant 1/2$ and $0 \leqslant \zeta \leqslant \sigma^p(\theta)$.

For given p and θ we can approximately compute the value of $\sigma^p(\theta)$ by halving intervals. The results of these computations for $p = 3, 4, \ldots, 10$ are presented in Fig. 2. This figure suggests that functions $\sigma^p(\theta)$ are concave. Hence, in order to derive treatable conditions for the DMP, we estimate $\sigma^p(\theta)$ from below by a line

$$\gamma^p\theta + \delta^p \leqslant \sigma^p(\theta) \quad \text{for } p = 3, 4, \ldots, 10 \text{ and } \theta \in (0, 1/2]. \tag{59}$$

The constants γ^p and δ^p are defined as

Table 1
The critical values α^p, β^p, γ^p, and δ^p.

p	α^p	β^p	γ^p	δ^p
1	∞	6	0	∞
2	20/3	∞	0	∞
3	38.61	25.89	5.608	0
4	18.91	∞	2.936	3.614
5	49.44	59.82	7.799	0
6	37.56	∞	7.247	0.887
7	72.82	107.81	9.791	0
8	62.62	∞	9.709	0
9	104.09	169.85	11.510	0
10	94.10	∞	10.644	0

$$\delta^p = \sup\{0\} \cup \left\{\bar{\zeta}\colon \operatorname{Ker}^{b,p}(\zeta,\xi,\eta) \geqslant 0 \text{ for all } 0 \leqslant \zeta \leqslant \bar{\zeta}\right\}, \tag{60}$$

$$\gamma^p = \begin{cases} 2(\sigma^p(1/2) - \delta^p) & \text{for } \delta^p < \infty, \\ 0 & \text{for } \delta^p = \infty. \end{cases} \tag{61}$$

The computed values of γ^p and δ^p are presented in Table 1 for $p = 1, 2, \ldots, 10$. We remark that $\sigma^p(\theta) \to \delta^p$ for $\theta \to 0$. Further, we remark that the constant δ^p characterizes the nonnegativity of the bubble part of the DGF. If $\zeta \leqslant \delta^p$ then $\widehat{G}^b_{hp} \geqslant 0$ in $\widehat{K}^2$, see (60). However, the bubble part of the DGF is nonnegative in exceptional cases only. It is trivially nonnegative for $p = 1$ and $p = 2$. Therefore, $\delta^p = \infty$ for $p = 1, 2$. The other two exceptional cases are $p = 4$ and $p = 6$. These are the only cases when $\widehat{G}^b_{hp}$ is nonnegative for the Laplacian ($\kappa = 0$), see [17].

The nonnegativity result for the second auxiliary DGF $\widehat{G}_{hp}$ is based on the following assumption.

$$\text{In case } \delta^p < \infty \quad \text{assume} \quad \gamma^p \geqslant 3/2 \quad \text{and} \quad \widehat{\omega}^p\big(\theta, \gamma^p\theta + \delta^p\big) \geqslant 0 \quad \text{for all } \theta \in (0, 1/2]. \tag{62}$$

This assumption is verified in Section 8 for p up to 10.

Lemma 7.4. *Let $K \in \mathcal{T}_{hp}$, let $\widehat{G}_{hp}$ be defined by* (49)–(52)*, let h and p be the length and the polynomial degree of K, let $\theta = h/(|\Omega| - h)$, let $\zeta = \kappa^2 h^2$, and let γ^p and δ^p be given by* (61) *and* (60)*. In case $\delta^p < \infty$ assume $\theta \leqslant 1/2$. If the inequality $\zeta \leqslant \gamma^p\theta + \delta^p$, the condition* (a) *from Theorem* 6.1*, and the assumption* (62) *are satisfied then $\widehat{G}_{hp}(\hat{x}, \hat{y}) \geqslant 0$ for all $(\hat{x}, \hat{y}) \in \widehat{K}$.*

Proof. Due to (55)–(58) it suffices to prove the nonnegativity of $\widehat{\omega}^p(\theta, \zeta)$. Since $0 \leqslant \delta^p \leqslant \gamma^p\theta + \delta^p$, we can split the proof into three cases.

(i) If $\zeta \in (\delta^p, \gamma^p\theta + \delta^p]$ then $\delta^p < \infty$ and we set $\theta^* = (\zeta - \delta^p)/\gamma^p$. Clearly, $\zeta = \gamma^p\theta^* + \delta^p$ and $0 < \theta^* \leqslant \theta \leqslant 1/2$. Furthermore, $\zeta - 3\theta^*\theta \geqslant \zeta - 3\theta^*/2 = (\gamma^p - 3/2)\theta^* + \delta^p \geqslant 0$, where we use assumption $\gamma^p \geqslant 3/2$. From these inequalities and from (42) we infer

$$0 \leqslant \frac{1}{3\theta^*\theta}\big(\theta - \theta^*\big)\big(\zeta - 3\theta^*\theta\big) = \theta^* - \theta + \left(\frac{1}{3\theta^*} - \frac{1}{3\theta}\right)\zeta = \big[s\big(0,\theta^*,\zeta\big)\big]^{-1} - \big[s(0,\theta,\zeta)\big]^{-1}.$$

Hence, $s(0,\theta,\zeta) \geqslant s(0,\theta^*,\zeta) = s(0,\theta^*,\gamma^p\theta^* + \delta^p)$ and consequently $\omega^p(\theta,\zeta,\xi,\eta) \geqslant \omega^p(\theta^*, \gamma^p\theta^* + \delta^p, \xi, \eta) \geqslant \widehat{\omega}^p(\theta^*, \gamma^p\theta^* + \delta^p) \geqslant 0$ for all $(\xi,\eta) \in K^2_{\mathrm{ref}}$, where we use assumption (62).

(ii) If $\zeta \in (0, \delta^p]$ then condition (a) from Theorem 6.1 guarantees $\Psi^p_1(\zeta,\xi) \geqslant 0$. From definition (60) we obtain $\operatorname{Ker}^{b,p}(\zeta,\xi,\eta) \geqslant 0$ for all $(\xi,\eta) \in K^2_{\mathrm{ref}}$. Since $s(0,\theta,\zeta) \geqslant 0$ by (42), we conclude that $\widehat{\omega}^p(\theta,\zeta) \geqslant 0$, see (56)–(58).

(iii) If $\zeta = 0$, i.e. if $\kappa = 0$, then we consider a sequence ζ_i, $i = 1, 2, \ldots$, such that $\zeta_i \to 0$ for $i \to \infty$ and $0 < \zeta_i \leqslant \gamma^p\theta + \delta^p$. For each ζ_i we may use either (i) or (ii) to conclude that $\omega^p(\theta,\zeta_i,\xi,\eta) \geqslant 0$ for all $(\xi,\eta) \in K^2_{\mathrm{ref}}$. Since $\omega^p(\theta,\zeta,\xi,\eta)$ is a continuous function for $\theta > 0$, $\zeta \geqslant 0$, and $(\xi,\eta) \in K^2_{\mathrm{ref}}$, we conclude that $\omega^p(\theta,\zeta_i,\xi,\eta) \to \omega^p(\theta,0,\xi,\eta) \geqslant 0$ as $i \to \infty$ for all $(\xi,\eta) \in K^2_{\mathrm{ref}}$. □

The following theorem concludes our analysis and summarizes the sufficient conditions for the DMP.

Theorem 7.1. *Let us consider the hp-FEM problem* (3) *discretized on a mesh $\mathcal{T}_{hp}$. Denote by h_K and p_K the lengths and the polynomial degrees of elements $K \in \mathcal{T}_{hp}$. Further, consider $\theta_K = h_K/(|\Omega| - h_K)$ and constants α^p, β^p, γ^p, and δ^p introduced in* (31)*,* (32)*,* (61)*, and* (60)*, respectively. Let assumption* (62) *be satisfied for all $p \in \{p_K\colon K \in \mathcal{T}_{hp}\}$. In case $\delta^{p_K} < \infty$ assume*

$$h_K \leqslant |\Omega|/3. \tag{63}$$

If

$$\kappa^2 h_K^2 \leqslant \min\big\{\alpha^{p_K}, \beta^{p_K}, \gamma^{p_K}\theta_K + \delta^{p_K}\big\} \quad \textit{for all } K \in \mathcal{T}_{hp}, \tag{64}$$

then the approximate problem (3) *satisfies the DMP.*

Proof. First notice that (63) is equivalent to $\theta_K \leqslant 1/2$. The DMP then follows from Theorems 6.1 and 3.2. The assumptions (a) and (b) of Theorem 6.1 are guaranteed by Lemmas 6.1 and 6.2 and hypothesis (64). The assumption (c) of Theorem 6.1 follows from (63)–(64), Lemma 7.4 and Corollary 7.1. □

Let us emphasize that both sides of the crucial condition (64) depend on h_K. The inequality $\kappa^2 h_K^2 \leqslant \gamma^{p_K}\theta_K + \delta^{p_K}$ is cubic in h_K and cannot be simplified, in general. However, we can infer a simple sufficient condition for its validity. Indeed, it suffices to have $\kappa^2 h_K|\Omega| \leqslant \gamma^{p_K} + \delta^{p_K}$, because $h_K/|\Omega| \leqslant 1$ and $h_K/|\Omega| \leqslant \theta_K$.

8. Verification of the assumptions

The computation of sharp lower estimates of constants α^p, β^p, γ^p, and δ^p is not very demanding. For $p = 1$ and $p = 2$ we can easily compute even the exact values. The results up to $p = 10$ are presented in Table 1. First, we observe the two exceptional cases $p = 1$ and $p = 2$. In these cases assumption (62) does not apply. We also verified that $\gamma^p \geqslant 3/2$ for $p = 3, 4, \ldots, 10$. A closer look shows that $\gamma^p\theta + \delta^p \leqslant \min\{\alpha^p, \beta^p\}$ for $p = 3, 4, \ldots, 10$. Hence, condition (64) can be replaced for $p_K = 3, 4, \ldots, 10$ by simpler condition

$$\kappa^2 h_K^2 \leqslant \gamma^{p_K}\theta_K + \delta^{p_K} \quad \text{for all } K \in \mathcal{T}_{hp}. \tag{65}$$

As we mentioned above, this can be further simplified to $\kappa^2 h_K|\Omega| \leqslant \gamma^{p_K} + \delta^{p_K}$, which clearly shows that the DMP is satisfied for sufficiently small h_K.

The crucial assumption (62) can be reformulated as nonnegativity of a polynomial in variables θ, ξ, η in a domain $(0, 1/2] \times K_{\text{ref}} \times K_{\text{ref}}$. Indeed, $\omega^p(\theta, \zeta, \xi, \eta)$ is a rational function with a positive denominator. The nonnegativity of a polynomial on an interval can be further reformulated as nonnegativity of a polynomial on entire $\mathbb{R}$. The verification of nonnegativity of a polynomial is connected with the 17th Hilbert problem [14]. There exist (NP-hard) algorithms for verification of nonnegativity of a polynomials, see e.g. [13]. These algorithms, however, are difficult to implement and lead to reasonable solution for small number of variables and for small polynomial degrees, only.

Another possibility is the usage of interval arithmetic. The idea is to compute an interval $R = f(I)$ containing all possible outputs of a function f on an interval I. If R is nonnegative (contains nonnegative numbers only) then nonnegativity of f in I is verified. If not, we split I into two (or more) subintervals and repeat the process for all these subintervals. If this algorithm terminates after a finite number of steps, the nonnegativity of f in I is verified.

Assumption (62) was verified by this algorithm for $p = 3, 4, \ldots, 10$. The `matlab` codes can be downloaded from `http://www.math.cas.cz/vejchod/DMPabs.html`. These codes utilize the interval arithmetic package `intlab` [10], where the interval operations provide guaranteed results even in the floating-point arithmetic.

9. Conclusions

The DMP for the diffusion–reaction problem discretized by *hp*-FEM, see (3), is essentially satisfied if the *hp*-mesh $\mathcal{T}_{hp}$ satisfies conditions (63)–(64) from Theorem 7.1. The other assumptions of this theorem are technical and were verified by computer for polynomial degrees up to 10.

The presented analysis applies to all polynomial degrees $p \geqslant 1$, but it is mainly relevant for $p \geqslant 3$. The cases $p = 1$ and $p = 2$ are exceptional, because the bubble part of the DGF is zero (for $p = 1$) or trivially nonnegative (for $p = 2$). Hence, the difficult analysis from Section 7 including assumption (62) is not needed for $p = 1$ and $p = 2$. In case $p = 1$ we can even show that the obtained condition is also necessary, i.e. in case of linear FEM with $M \geqslant 2$, the DMP for problem (3) is satisfied if and only if $\kappa^2 h_K^2 \leqslant 6$ for all $K \in \mathcal{T}_{hp}$.

Finally, let us notice that $\sigma^3(\theta) \leqslant \sigma^p(\theta)$ for all $p = 3, 4, \ldots, 10$, see Fig. 2. This (or more precisely the values of γ^p and δ^p in Table 1) implies that condition (64) in Theorem 7.1 or its simplified version (65) is the most strict for $p = 3$. This observation is in agreement with the previous results for the Poisson problem, see [17]. The growing trend of values $\sigma^p(\theta)$ for increasing p observed in Fig. 2 allows us to conclude this paper by the following conjecture.

Conjecture 9.1. *Let us consider a finite element mesh $\mathcal{T}_{hp}$ with an arbitrary distribution of polynomial degrees. Denote by h_K the length of the element K and set $\theta_K = h_K/(|\Omega| - h_K)$. If*

$$\kappa^2 h_K^2/\gamma^3 \leqslant \theta_K \leqslant 1/2 \quad \textit{for all } K \in \mathcal{T}_{hp},$$

where $\gamma^3 \approx 5.608797$, then the approximate problem (3) *satisfies the DMP.*

References

[1] M. Berzins, Preserving positivity for hyperbolic PDEs using variable-order finite elements with bounded polynomials, Appl. Numer. Math. 52 (2005) 197–217.

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

[3] E. Burman, A. Ern, Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes, C. R. Math. Acad. Sci. Paris 338 (8) (2004) 641–646.
[4] P.G. Ciarlet, Discrete maximum principle for finite-difference operators, Aequationes Math. 4 (1970) 338–352.
[5] P.G. Ciarlet, Discrete variational Green's function. I, Aequationes Math. 4 (1970) 74–82.
[6] P.G. Ciarlet, R.S. Varga, Discrete variational Green's function. II. One dimensional problem, Numer. Math. 16 (1970) 115–128.
[7] A. Drăgănescu, T. Dupont, L. Scott, Failure of the discrete maximum principle for an elliptic finite element problem, Math. Comp. 74 (249) (2004) 1–23.
[8] I. Faragó, R. Horváth, S. Korotov, Discrete maximum principle for linear parabolic problems solved on hybrid meshes, Appl. Numer. Math. 53 (2–4) (2005) 249–264.
[9] A. Hannukainen, S. Korotov, T. Vejchodský, Discrete maximum principle for FE solutions of the diffusion–reaction problem on prismatic meshes, J. Comput. Appl. Math. 226 (2) (2009) 275–287.
[10] G. Hargreaves, Interval analysis in MATLAB, Numerical analysis report No. 416, The University of Manchester, 2002.
[11] W. Höhn, H. Mittelmann, Some remarks on the discrete maximum-principle for finite elements of higher order, Computing 27 (2) (1981) 145–154.
[12] J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, Numer. Math. 99 (4) (2005) 669–698.
[13] V. Powers, T. Wörmann, An algorithm for sums of squares of real polynomials, J. Pure Appl. Algebra 127 (1) (1998) 99–104.
[14] A. Prestel, C.N. Delzell, Positive Polynomials: From Hilbert's 17th Problem to Real Algebra, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2001.
[15] R. Varga, Matrix Iterative Analysis, Prentice–Hall, Englewood Cliffs, NJ, 1962.
[16] R. Varga, On a discrete maximum principle, SIAM J. Numer. Anal. 3 (1966) 355–359.
[17] T. Vejchodský, P. Šolín, Discrete maximum principle for higher-order finite elements in 1D, Math. Comp. 76 (260) (2007) 1833–1846.
[18] P. Šolín, K. Segeth, I. Doležel, Higher-Order Finite Element Methods, Chapman & Hall/CRC, Boca Raton, FL, 2004.
[19] P. Šolín, T. Vejchodský, A weak discrete maximum principle for *hp*-FEM, J. Comput. Appl. Math. 209 (2007) 54–65.
[20] P. Šolín, T. Vejchodský, Higher-order finite elements based on generalized eigenfunctions of the Laplacian, J. Numer. Meth. Engrg. 73 (2008) 1374–1394.
[21] J. Xu, L. Zikatanov, A monotone finite element scheme for convection–diffusion equations, Math. Comp. 68 (228) (1999) 1429–1446.
[22] E. Yanik, Sufficient conditions for a discrete maximum principle for high order collocation methods, Comput. Math. Appl. 17 (11) (1989) 1431–1434.

APPENDIX

H

Angle conditions for discrete maximum principles in higher-order FEM

Below we attach a copy of the paper

[A7] T. Vejchodský: Angle conditions for discrete maximum principles in higher-order FEM. In: G. Kreiss, P. Lötstedt, A. Målqvist, and M. Neytcheva (eds.), *Numerical mathematics and advanced applications ENUMATH 2009*, Springer, Berlin, 2010, pp. 901–909.

Angle Conditions for Discrete Maximum Principles in Higher-Order FEM

Tomáš Vejchodský

Abstract This contribution reviews the general theory of the discrete Green's function and presents a numerical experiment indicating that the discrete maximum principle (DMP) fails to hold in the case of Poisson problem on any uniform triangulation of a triangular domain for orders of approximation three and higher. This extends the result [8] that the Laplace equation discretized by the higher-order FEM satisfies the DMP on a patch of triangular elements in exceptional cases only.

1 Introduction

The discrete maximum principle (DMP) is important in practice, because it guarantees nonnegativity of approximations of naturally nonnegative quantities like temperature, concentration, density, etc. Its theoretical significance lies in its connection with the uniform convergence of the finite element approximations [4]. In contrast to the lowest-order finite element method (FEM), the DMP for the higher-order FEM in dimension two and higher is not well understood, yet.

A stronger version of the DMP for the Laplace equation discretized by higher-order finite elements was studied by Höhn and Mittelman in [8]. This stronger version requires the validity of the DMP on all vertex patches (union of elements sharing a vertex) in the triangulation. They find that the quadratic elements do not satisfy the stronger DMP unless the triangulation is very special (e.g. all equilateral triangles) and that the restrictions for cubic elements are even more severe.

In the present contribution we briefly review the general theory about the discrete Green's function (DGF) and the standard DMP for the Poisson problem. Then we present a numerical experiment indicating that the standard DMP is not satisfied on any uniform triangulation for the finite elements of order three and higher.

Tomáš Vejchodský
Institute of Mathematics, Academy of Sciences, Žitná 25, CZ–115 67 Prague 1, Czech Republic
e-mail: vejchod@math.cas.cz

2 Model problem and its FEM discretization

First, we briefly introduce the Poisson problem and its discretization by the FEM. The main purpose of this section is to settle down the notation.

Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. The classical and the weak formulations of the Poisson problem reads as follows:

$$\text{Find } u \in C^2(\Omega) \cup C(\overline{\Omega}) \text{ such that } \quad -\Delta u = f \quad \text{in } \Omega, \text{ and } \quad u = 0 \text{ on } \partial\Omega. \tag{1}$$

$$\text{Find } u \in H^1_0(\Omega) \text{ such that } \quad a(u,v) = \mathscr{F}(v) \quad \forall v \in H^1_0(\Omega), \tag{2}$$

where $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x$ and $\mathscr{F}(v) = \int_\Omega f v \,\mathrm{d}x$. We require $f \in C(\Omega)$ for the classical formulation and $f \in L^2(\Omega)$ for the weak one.

In order to discretize problem (2) by the Galerkin method, we introduce a finite dimensional subspace V_h of $H^1_0(\Omega)$. We assume that $V_h \subset C(\overline{\Omega})$. The Galerkin solution $u_h \in V_h$ is given by the requirement

$$a(u_h, v_h) = \mathscr{F}(v_h) \quad \forall v_h \in V_h. \tag{3}$$

Considering a basis $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h, we can express $u_h = \sum_{i=1}^N z_i \varphi_i$ and verify that problem (3) is equivalent to the system $Az = F$ of linear algebraic equations, where the stiffness matrix $A \in \mathbb{R}^{N\times N}$ has entries $a_{ij} = a(\varphi_j, \varphi_i)$, the load vector $F \in \mathbb{R}^N$ has entries $F_i = \mathscr{F}(\varphi_i)$, and $z = (z_1, z_2, \ldots, z_N)^\top$.

The FEM can be seen as a special case of the Galerkin method, where the space V_h is chosen in a special way such that the stiffness matrix A is sparse. The particular choice of V_h is not important at this point and it will be specified later on.

3 Discrete maximum principle

Theorem 1 below is an equivalent formulation of the standard maximum principle due to E. Hopf [9] applied to problem (1). Similarly, Theorem 2 presents the same principle for the weak solution.

Theorem 1. *Let u be a classical solution to (1). If $f \geq 0$ in Ω then $u \geq 0$ in Ω.*

Theorem 2. *Let u be a weak solution to (2). If $f \geq 0$ a.e. in Ω then $u \geq 0$ a.e. in Ω.*

The same result for the the Galerkin solution $u_h \in V_h$ is known as the DMP. Unfortunately, it is not valid in general and various conditions for its validity are studied.

Definition 1. Let the finite dimensional space V_h be fixed. We say that discretization (3) satisfies the discrete maximum principle (DMP) if the solution $u_h \in V_h$ is nonnegative in Ω for any $f \in L^2(\Omega)$, $f \geq 0$ a.e. in Ω.

A usefull tool for investigation of the DMP especially for the higher-order FEM is the so-called discrete Green's function (DGF) which was already introduced in [2, 5]. For any $y \in \Omega$ let us define the DGF $G_{h,y} \in V_h$ as the unique function satisfying

$$a(v_h, G_{h,y}) = v_h(y) \quad \forall v_h \in V_h. \tag{4}$$

This definition together with (3) implies the representation formula

$$u_h(y) = \mathscr{F}(G_{h,y}) = \int_\Omega f(x) G_h(x,y)\,\mathrm{d}x \quad \forall y \in \Omega,$$

where we use the usual notation $G_h(x,y) = G_{h,y}(x)$. This representation formula immediately proves the following theorem.

Theorem 3. *The discretization* (3) *satisfies the DMP if and only if* $G_h(x,y) \geq 0$ *for all* $(x,y) \in \Omega^2$.

Interestingly, the DGF G_h can be expressed in terms of a basis of V_h [12]:

$$G_h(x,y) = \sum_{i=1}^{N} \sum_{j=1}^{N} (A^{-1})_{ij} \varphi_i(x) \varphi_j(y) \quad \forall (x,y) \in \Omega^2, \tag{5}$$

where $(A^{-1})_{ij}$ stand for entries of the inverse of the stiffness matrix A. Let us remark that a special case of this formula, where the basis is formed by the eigenvectors of the discrete Laplacian was already presented in [2]. Further, we remark that the concept of the DGF is relevant even for more general problems. However, in the case of nonhomogeneous Dirichlet boundary conditions the boundary Green's function has to be introduced [3]. General formula (5) is used below to analyze the nonnegativity of the DGF and consequently the validity of the DMP.

4 Nonnegativity of the DGF for the lowest-order FEM

The analysis of nonnegativity of expression (5) simplifies if the basis functions $\varphi_1, \varphi_2, \ldots, \varphi_N$ of V_h have the following property

$$\sum_{i=1}^{N} z_i \varphi_i \geq 0 \quad \text{in } \Omega \quad \Leftrightarrow \quad z_i \geq 0 \quad \forall i = 1,2,\ldots,N. \tag{6}$$

This property is typically satisfied for the lowest-order finite elements such as linear functions on simplices and multilinear functions on blocks (Cartesian products of intervals). Before we state the following well-known theorem, we recall that a square matrix A is monotone if it is nonsingular and $A^{-1} \geq 0$ (i.e. all entries of A^{-1} are nonnegative).

Theorem 4. *Let the basis functions* $\varphi_1, \varphi_2, \ldots, \varphi_N$ *of* V_h *have property* (6). *Then the discretization* (3) *satisfies the DMP if and only if the stiffness matrix* A *is monotone.*

Proof. It follows immediately from assumption (6), formula (5), and Theorem 3.

If the off-diagonal entries of the stiffness matrix A are nonpositive then A is M-matrix and, hence, monotone. The nonpositivity of the off-diagonal entries can

be guaranteed by various geometric conditions on finite element meshes like the nonobtuseness condition for simplicial meshes [1] or the nonnarrowness condition for rectangular finite elements [6]. However, these conditions could be too restrictive, because it suffices to have the stiffness matrix monotone and not M-matrix. An experiment indicating how much the nonobtuseness condition for triangles can be weaken is described in Section 6 and its results are presented in Fig. 2 (top-left).

5 Nonnegativity of the DGF for the higher-order FEM

Let us investigate the case of the higher-order FEM in more details. For simplicity let us consider two dimensional Poisson problem (1) in a polygonal domain Ω. We define the finite element space as $V_h = \{v \in H_0^1(\Omega) : v|_K \in \mathbb{P}^p(K) \quad \forall K \in \mathscr{T}_h\}$, where $\mathscr{T}_h$ is a face-to-face triangulation of Ω and $\mathbb{P}^p(K)$ stands for the space of polynomials of degree at most p on the triangle K.

The standard basis of V_h consists of N^{V} vertex (piecewise linear) functions $\varphi_1, \varphi_2, \ldots, \varphi_{N^{\mathrm{V}}}$ and of $N - N^{\mathrm{V}}$ higher-order basis functions $\varphi_{N^{\mathrm{V}}+1}, \varphi_{N^{\mathrm{V}}+2}, \ldots, \varphi_N$, see e.g. [11]. The vertex functions are the usual piecewise linear "hat" functions. Thus, if B_j, $j = 1,2,\ldots,N^{\mathrm{V}}$, denote the interior vertices of the triangulation $\mathscr{T}_h$ then the vertex functions satisfy $\varphi_i(B_j) = \delta_{ij}$, $i,j = 1,2,\ldots,N^{\mathrm{V}}$.

The vertex and the higher-order (non-vertex) basis functions yield a natural 2×2 block structure of the stiffness matrix and its inverse

$$A = \begin{pmatrix} A^{\mathrm{VV}} & A^{\mathrm{VN}} \\ A^{\mathrm{NV}} & A^{\mathrm{NN}} \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} S^{-1} & -(A^{\mathrm{VV}})^{-1}A^{\mathrm{VN}}R^{-1} \\ -(A^{\mathrm{NN}})^{-1}A^{\mathrm{NV}}S^{-1} & R^{-1} \end{pmatrix},$$

where $A^{\mathrm{VV}} \in \mathbb{R}^{N^{\mathrm{V}}\times N^{\mathrm{V}}}$, $A^{\mathrm{NN}} \in \mathbb{R}^{(N-N^{\mathrm{V}})\times(N-N^{\mathrm{V}})}$, etc., $S = A^{\mathrm{VV}} - A^{\mathrm{VN}}(A^{\mathrm{NN}})^{-1}A^{\mathrm{NV}}$, and $R = A^{\mathrm{NN}} - A^{\mathrm{NV}}(A^{\mathrm{VV}})^{-1}A^{\mathrm{VN}}$.

The Schur complement S has the following interesting property. Let B_i and B_j, $i,j = 1,2,\ldots,N^{\mathrm{V}}$, be two interior vertices of the triangulation $\mathscr{T}_h$. Since $\varphi_i(B_j) = \delta_{ij}$ and due to (5) we obtain

$$G_h(B_i,B_j) = (A^{-1})_{ij}\varphi_i(B_i)\varphi_j(B_j) = (A^{-1})_{ij} = (S^{-1})_{ij}. \tag{7}$$

Hence, the values of the DGF at the vertices of $\mathscr{T}_h$ coincide with the entries of S^{-1}. Furthermore, the DGF has a natural structure given by the Cartesian product of the mesh $\mathscr{T}_h$ with itself. In particular, if K and L are two elements from $\mathscr{T}_h$ and ι_K and ι_L denote the sets of indices of basis functions supported in K and L, respectively, i.e., $\iota_K = \{i : \mathrm{meas}(K \cap \mathrm{supp}\,\varphi_i) > 0\}$, then the DGF restricted to $K\times L$ is given by

$$G_h|_{K\times L}(x,y) = \sum_{i\in\iota_K}\sum_{j\in\iota_L}(A^{-1})_{ij}\varphi_i|_K(x)\varphi_j|_L(y), \quad (x,y)\in K\times L. \tag{8}$$

This formula contains a small number of basis functions and we use it for fast evaluation of the DGF at a given point.

6 Numerical experiment

In this experiment we test nonnegativity of the DGF on uniform meshes. We consider Poisson problem (1) on a triangle Ω. The finite element mesh is constructed by three successive uniform (red) refinements of Ω, see Fig. 1 (left).

To speed up the test of the nonnegativity of the DGF, we first check the values at vertices, using the Schur complement S, see (7). If S is monotone, it remains to verify the nonnegativity at the other points. We proceed by inspection of all pairs of elements $K, L \in \mathcal{T}_h$ using formula (8). Function $G_h|_{K\times L}$ is a polynomial. The test of nonnegativity of a multivariate polynomial is a complicated task (connected with the 17th Hilbert's problem [10]). Therefore, we sample the values of $G_h|_{K\times L}$ in a number of points $(x^K_{k\ell}, x^L_{mn}) \in K \times L$, where the sample point $x^K_{k\ell}$ has barycentric coordinates $(k,\ell,M-k-\ell)/M$, $0 \le k+\ell \le M$, see Fig. 1 (right). The total number of sample points in an element is $(M+1)(M+2)/2$. To ensure that the number of sample points is sufficient, we always perform a series of computations starting with $M = 8$ and doubling M until the results do not change.

Fig. 2 presents the results. Each point in a panel corresponds to a pair of angles α and β, which represent the vertex angles of the triangle Ω. The color of this point is given by the properties of the DGF. If the DGF is nonnegative at all vertices and at all sample points then the color is black. This is the only case when the DMP is hopefully satisfied. If the DGF is not nonnegative then we distinguish three more cases. (i) The DGF is negative in a sample point and S is M-matrix (dark gray). (ii) The DGF is negative in a sample point and S is monotone but not M-matrix (lighter gray). (iii) The DGF is negative in a vertex, i.e., S is nonmonotone (lightest gray).

The above description, however, applies for higher-order elements only ($p \ge 2$). The case of linear elements ($p = 1$) is exceptional, because just the vertex values of the DGF are relevant for its nonnegativity. Due to Theorem 4, we distinguish in the top-left panel of Fig. 2 the cases (a) A is nonmonotone, (b) A is monotone but not M-matrix, (c) A is M-matrix. Notice that the DMP is satisfied in cases (b) and (c).

Clear conclusion from Fig. 2 is that the DGF has negative values for all tested pairs of angles for orders $p \ge 3$. However, if we look on vertex values of the DGF only, we observe that the area of this region increases with p. The increase is not monotone but in principle the higher polynomial degree p we use the wider range of angles can be used in order to keep the vertex values of the DGF nonnegative.

The only polynomial degrees allowing the DMP on uniform meshes are $p = 1$ and $p = 2$. For the case $p = 1$ (see Section 4 above) the black area in the top-left panel of Fig. 2 clearly shows that the stiffness matrix A is M-matrix provided the maximal angle is at most 90°. In addition, we observe that the stiffness matrix can be monotone even if the maximal angle is about 117°. In the case $p = 2$ the DMP is satisfied only if all the angles are close to 60°. We also check the nonnegativity of the DGF for meshes finer than the mesh sketched in Fig. 1 (left). The results on meshes one and two times refined are exactly the same as those presented in Fig. 2.

It might be of further interest to see how the DGF really looks like. For illustration we choose $p = 3$ and $\alpha = \beta = 60^\circ$. For these values the DGF is nonnegative in the vertices and negative somewhere in between. The graph of the function $G_h(x,y)$,

$(x,y) \in \Omega^2$, is difficult to visualize, because it is a five dimensional object. However, each pair of elements $K_i \in \mathcal{T}_h$ and $K_j \in \mathcal{T}_h$ corresponds to a point in a plain and the color of this point can be chosen according to some characteristic of the DGF restricted to the polytop $K_i \times K_j$. The left panel of Fig. 3 presents the mean values of G_h over $K_i \times K_j$. The right panel illustrates the negative part of the minimum of G_h in $K_i \times K_j$, i.e., $(\min_{K_i \times K_j} G_h)^-$, where $\chi^- = (|\chi| - \chi)/2$. Both these quantities are approximated using the sample points as described above. The used triangulation together with indices of elements is shown in Fig. 1 (left). Notice that the elements with indices 1–39 are adjacent to the boundary of Ω while the elements 40–64 are interior. The right panel of Fig. 3 clearly shows that the DGF is negative in polytops $K_i \times K_j$, where K_i and K_j are both adjacent to the boundary and they are neighbors to each other including the case $K_i = K_j$. Another choice of angles α and β leads, however, to the negativity of the DGF for more pairs K_i, K_j.

7 Conclusions

We discussed the nonnegativity of the DGF and equivalently the validity of the DMP for Galerkin solutions of Poisson problem (1) with homogeneous Dirichlet boundary conditions. Results of the performed experiment indicate that the DGF is not nonnegative on uniform meshes for all shapes of triangular elements for the order three and higher. The quadratic elements yield nonnegative DGF for triangles close to equilateral ones.

The results also indicate that the DGF is negative in the areas close to the boundary. In accordance with [7] we could speculate that the nonnegativity of the DGF is not primarily determined by the angles in the triangulation but by the way how the boundary is resolved. In addition, the domain, where the DGF is negative, is relatively small with respect to the entire Ω^2 and it lies close to the boundary. This means that a nonnegative f corrupting the DMP (Definition 1) must have great values in an element close to the boundary and small values in the interior of Ω (like an approximation of the Dirac delta function). Such data are rare in practice, however. This leads us to another generalization of the (continuous) maximum principle from Theorem 2. If $f \geq 0$ is given, we may ask how must the mesh look like in order to obtain the nonnegative finite element solution. Up to the author's knowledge, this question was not considered in the literature, yet.

A possible remedy of the failure of the DMP for higher-order elements could be a modification of the higher-order basis functions based on the exact eigenfunctions of the Laplacian. This approach was successfully applied in [5] for 1D elliptic problems. A generalization to higher dimension is still an unsolved problem.

Acknowledgements The author acknowledges the support of the Czech Science Foundation, Grant no. 102/07/0496, and of the Czech Academy of Sciences, Grant no. IAA100760702, and Institutional Research Plan no. AV0Z10190503.

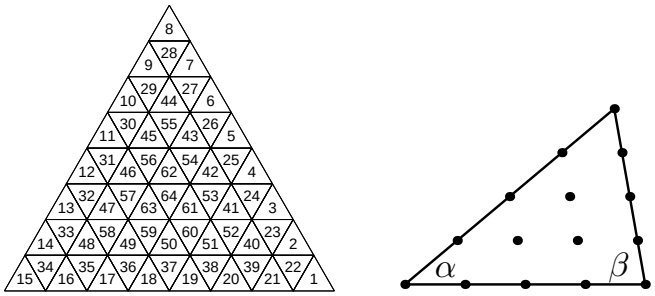


Fig. 1 A uniform mesh with 64 triangles enumerated in a spiral way (left). A triangular element characterized by a pair of of angles α and β with sample points for $M = 4$ (right).

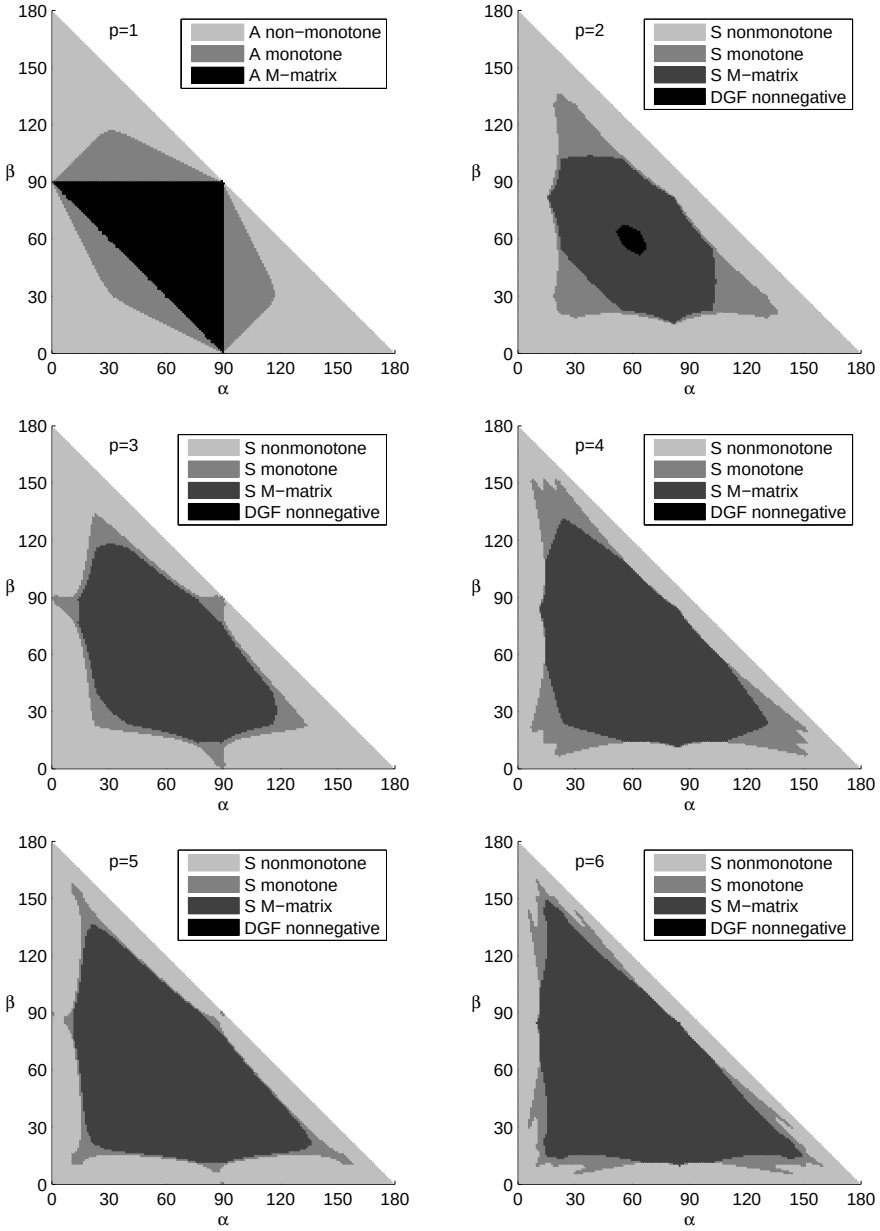


Fig. 2 The nonnegativity of the DGF and its dependence on the angles in the triangulation for orders $p = 1, 2, \ldots, 6$.

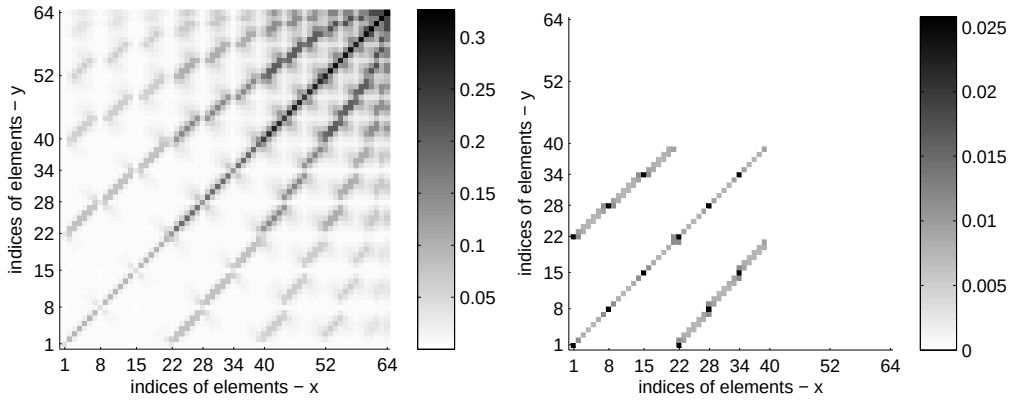


Fig. 3 A visualization of the entire DGF. A point with coordinates i, j corresponds to a pair of elements K_i, K_j. The color of this point represents the mean value (left) and the negative part of the minimum (right) of G_h in $K_i \times K_j$.

References

1. Brandts, J., Korotov, S., Křížek, M.: Dissection of the path-simplex in $\mathbf{R}^n$ into n path-subsimplices. Linear Algebra Appl. **421**, 382–393 (2007)
2. Ciarlet, P.G.: Discrete variational Green's function. I. Aequationes Math. **4**, 74–82 (1970)
3. Ciarlet, P.G.: Discrete maximum principle for finite-difference operators. Aequationes Math. **4**, 338–352 (1970)
4. Ciarlet, P.G., Raviart, P.A.: Maximum principle and uniform convergence for the finite element method. Comput. Methods Appl. Mech. Engrg. **2**, 17–31 (1973)
5. Ciarlet, P.G., Varga, R.S.: Discrete variational Green's function. II. One dimensional problem. Numer. Math. **16**, 115–128 (1970)
6. Christie, I., Hall, C.: The maximum principle for bilinear elements. Internat. J. Numer. Methods Engrg. **20**, 549–553 (1984)
7. Drăgănescu, A., Dupont, T.F., Scott, L.R.: Failure of the discrete maximum principle for an elliptic finite element problem. Math. Comp. **74**, 1–23 (2005)
8. Höhn, W., Mittelmann, H.-D.: Some remarks on the discrete maximum-principle for finite elements of higher order. Computing **27**, 145–154 (1981)
9. Hopf, E.: Elementäre Bemerkungen über die Lösungen partieller Differentialgleichungen zweiter Ordnung vom elliptischen Typus. Sitzungsberichte Preussische Akademie der Wissenschaften, Berlin, 147–152 (1927)
10. Prestel, A., Delzell, C. N.: Positive polynomials: From Hilbert's 17th problem to real algebra. Springer-Verlag, Berlin (2001)
11. Šolín, P., Segeth, K., Doležel, I.: Higher-order finite element methods. Chapman & Hall/CRC, Boca Raton, FL (2004)
12. Vejchodský, T., Šolín, P.: Discrete maximum principle for higher-order finite elements in 1D. Math. Comp. **76**, 1833–1846 (2007)
13. Vejchodský, T., Šolín, P.: Discrete maximum principle for a 1D problem with piecewise-constant coefficients solved by *hp*-FEM. J. Numer. Math. **15**, 233–243 (2007)
14. Vejchodský, T., Šolín, P.: Discrete maximum principle for Poisson equation with mixed boundary conditions solved by *hp*-FEM. Adv. Appl. Math. Mech. **1**, 201–214 (2009)

APPENDIX

I

A weak discrete maximum principle for *hp*-FEM

Below we attach a copy of the paper

[A8] P. Šolín and T. Vejchodský: A weak discrete maximum principle for *hp*-FEM. *J. Comput. Appl. Math.* **209** (2007), 54–65.

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Journal of Computational and Applied Mathematics 209 (2007) 54–65
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS
www.elsevier.com/locate/cam

A weak discrete maximum principle for *hp*-FEM

Pavel Šolín[a, c, *], Tomáš Vejchodský[b]

[a] *Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968-0514, USA*
[b] *Mathematical Institute, Academy of Sciences, Žitná 25, 11567 Praha 1, Czech Republic*
[c] *Institute of Thermomechanics, Academy of Sciences of the Czech Republic, Dolejskova 5, 18200 Praha 8, Czech Republic*

Received 23 May 2006; received in revised form 14 September 2006

Abstract

In this paper, we prove a new discrete maximum principle (DMP) for the one-dimensional Poisson equation discretized by the *hp*-FEM. While the DMP for piecewise-linear elements is a classical result from the 1970s, no extensions to *hp*-FEM are available to the present day. Due to a negative result by Höhn and Mittelmann from 1981, related to quadratic Lagrange elements, it was long assumed that higher-order finite elements do not satisfy discrete maximum principles. In this paper we explain why it is not possible to make a straightforward extension of the classical DMP to the higher-order case, and we propose stronger assumptions on the right-hand side under which an extension is possible.

MSC: 35B50; 65N60

Keywords: Discrete maximum principle; Poisson equation; *hp*-FEM; Higher-order elements

1. Introduction

Discrete maximum principles (DMP) are the numerical counterparts of the (continuous) maximum principles for elliptic and parabolic PDEs. In the 1970s, these results were used to prove the convergence of finite differences and lowest-order finite element methods (see, e.g., [3,4]). Nowadays, DMP still play an important role in computational PDEs by providing restrictions on the mesh under which the approximation of physically nonnegative quantities such as the density, temperature, concentration, or electric charge remains nonnegative. In the early 1980s, Höhn and Mittelmann [7] showed that a straightforward generalization of the standard DMP for piecewise-linear approximations to quadratic Lagrange elements did not hold but under unrealistic restrictions on the triangulation, and since then no new results on DMP for higher-order elements have been obtained. Also the current research on DMP deals exclusively with lowest-order elements (see, e.g., [8–10,18,19]).

In the last decades, significant progress has been made in the development of the *hp*-FEM and its applications to challenging large-scale problems in computational science and engineering (see, e.g., [1,2,5,11,12,14,16]). An increasing demand for these methods naturally implies a need for the generalization of the DMP from lowest-order to higher-order elements.

* Corresponding author.
E-mail addresses: solin@utep.edu (P. Šolín), vejchod@math.cas.cz (T. Vejchodský).
URL: http://hpfem.math.utep.edu/ (P. Šolín).

0377-0427/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2006.10.028

The outline of this paper is as follows: The *hp*-FEM discretization of one-dimensional Poisson equation is recalled briefly in Section 2. An alternative proof of the classical DMP for piecewise-linear FEM (which does not use *M*-matrices) is given in Section 3. In Section 4 we provide a counter example which demonstrates that a straightforward extension of the standard DMP to higher-order elements is not possible. In Section 5 we formulate and prove a new DMP for higher-order elements which instead of nonnegativity of the right-hand side assumes the non-negativity of its L^2-projection to the finite element subspace (we call this a *weak DMP*). The assumptions of the main theorem are verified in Section 6 for sufficiently high polynomial degrees.

2. Model problem and its discretization

Consider an open bounded interval $\Omega = (a, b) \subset \mathbb{R}$ and the Poisson equation $-u'' = f$ in Ω equipped with homogeneous Dirichlet boundary conditions $u(a) = u(b) = 0$. The standard variational formulation of this problem reads: Given a right-hand side $f \in L^2(\Omega)$, find a function $u \in H_0^1(\Omega)$ such that the identity

$$\int_a^b u'(x)v'(x)\,\mathrm{d}x = \int_a^b f(x)v(x)\,\mathrm{d}x \tag{1}$$

holds for all test functions $v \in V = H_0^1(\Omega)$. We can restrict ourselves to the homogeneous Dirichlet boundary conditions since the case of nonhomogeneous Dirichlet boundary conditions does not cause any difficulty nor it involves special considerations.

Consider a partition $a = x_0 < x_1 < x_2 \cdots < x_M = b$ that splits $\overline{\Omega}$ into $M \geqslant 1$ finite elements $K_1, K_2, \ldots, K_M$. Each element K_i is equipped with a polynomial degree $p_i = p(K_i) \geqslant 1$. The elements $K_1, K_2, \ldots, K_M$, equipped with the polynomial degrees $p_1, p_2, \ldots, p_M$, form a finite element mesh $\mathscr{T}_{hp}$. The finite element space $V_{hp} \subset V$ on the mesh $\mathscr{T}_{hp}$ has the form

$$V_{hp} = \{v \in V;\ v(a) = v(b) = 0;\ v|_{K_i} \in P^{p_i}(K_i),\ 1 \leqslant i \leqslant M\}. \tag{2}$$

Here the symbol $P^{p_i}(K_i)$ stands for the space of polynomials of degree less than or equal to p_i in the interval K_i. The dimension of this space is

$$\dim(V_{hp}) = -1 + \sum_{i=1}^{M} p_i.$$

The discrete problem reads: find a function $u_{hp} \in V_{hp}$ such that the identity

$$\int_a^b u'_{hp}(x)v'_{hp}(x)\,\mathrm{d}x = \int_a^b f(x)v_{hp}(x)\,\mathrm{d}x \tag{3}$$

holds for every test function $v_{hp} \in V_{hp}$. Obviously, there exist unique solutions to both the continuous problem (1) and the discrete problem (3) (see, e.g., [12]).

The classical DMP for the discrete problem (3) can be stated in several equivalent ways, from which we may choose, e.g., the following:

Definition 1. The discrete problem (3) satisfies the discrete maximum principle (DMP) if the approximation u_{hp} attains its minimum on the boundary $\partial\Omega$ for every right-hand side f which is nonnegative a.e. in Ω.

3. Classical DMP for piecewise-linear FEM

The analysis of the DMP for higher-order elements is quite different from the analysis of the piecewise-linear case. In particular, the nonnegativity of the right-hand side no longer implies the nonnegativity of the load vector, and therefore the application of *M*-matrices becomes useless. In this section we begin by re-doing the proof for the piecewise-linear case without *M*-matrices.

Remark 2. The Poisson equation in 1D is an exceptional case, where the stiffness matrix is an *M*-matrix even for higher degree approximations. It is a consequence of the orthogonality (in the energy sense) of vertex and bubble

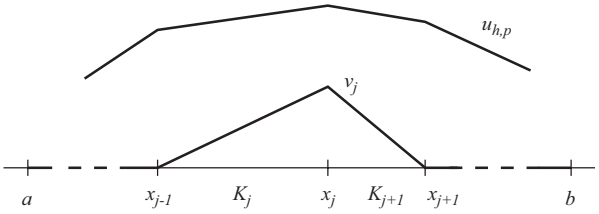


Fig. 1. Classical DMP in the piecewise-linear case.

functions and of orthogonality of bubbles themselves. However, as we will see below in Section 4, the fact that the stiffness matrix is an M-matrix is not enough to guarantee the DMP for hp-FEM.

Lemma 3. *If* $p_1 = p_2 = \cdots = p_M = 1$ *then problem* (3) *satisfies the DMP.*

Proof. For the standard proof based on M-matrices see, e.g., the fundamental book [17] or a more recent publication [6].

For our alternative proof let us consider a pair of adjacent elements $K_j = [x_{j-1}, x_j]$, $K_{j+1} = [x_j, x_{j+1}]$, and the piecewise-linear "hat function" $v_j \in V_{hp}$ associated with the grid point x_j, as shown in Fig. 1.

By substituting v_j for v_{hp} in the discrete problem (3) and using the nonnegativity of both f and v_j, we obtain

$$\int_{x_{j-1}}^{x_{j+1}} u'_{hp}(x)v'_j(x)\,\mathrm{d}x = \int_{x_{j-1}}^{x_{j+1}} f(x)v_j(x)\,\mathrm{d}x \geqslant 0. \tag{4}$$

By $Du_{hp}^{(j)}$ and $Du_{hp}^{(j+1)}$ let us denote the constant slopes of the piecewise-linear function u_{hp} in the elements K_j and K_{j+1}, respectively. Using the fact that the slopes of the test function v_j in the elements K_j and K_{j+1} are $1/(x_j - x_{j-1})$ and $-1/(x_{j+1} - x_j)$, respectively, from the inequality (4) we immediately obtain

$$0 \leqslant Du_{hp}^{(j)} \frac{x_j - x_{j-1}}{x_j - x_{j-1}} - Du_{hp}^{(j+1)} \frac{x_{j+1} - x_j}{x_{j+1} - x_j} = Du_{hp}^{(j)} - Du_{hp}^{(j+1)}.$$

Therefore, $Du_{hp}^{(j+1)} \leqslant Du_{hp}^{(j)}$ for every internal grid point x_j, $1 \leqslant j \leqslant M-1$. Thus, the function u_{hp} is concave in Ω. Taking into account its zero values at Ω-endpoints, we conclude that u_{hp} attains its minimum on the boundary of Ω. □

4. Attempt of straightforward extension to *hp*-FEM

Next let us show that a straightforward extension of Lemma 3 to higher-order elements fails already in the cubic case. For this, we need to recall the integrated Legendre polynomials (Lobatto shape functions) [12,13]. These polynomials are defined in the interval $[-1, 1]$ as

$$l_k(x) = \int_{-1}^{x} L_{k-1}(\xi)\,\mathrm{d}\xi, \quad 2 \leqslant k, \tag{5}$$

where L_{k-1} stands for the normalized Legendre polynomial of degree $p-1$. It follows from (5) that the functions $l_2, l_3, \ldots$ vanish at ± 1 and that they are orthonormal in the H_0^1-product,

$$(l_i, l_j)_{H_0^1(-1,1)} = \int_{-1}^{1} l'_i(x)l'_j(x)\,\mathrm{d}x = \delta_{ij}, \quad 2 \leqslant i, j. \tag{6}$$

The functions $l_2, l_3, \ldots, l_{10}$ are well-known, see, e.g., [14,16].

Example. Failure of the standard DMP for a cubic element.

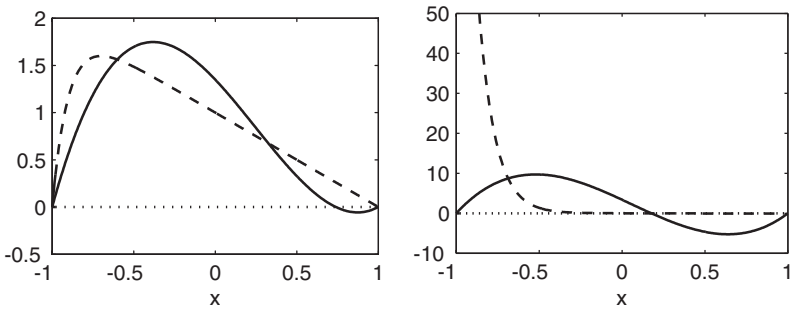


Fig. 2. Left: the *hp*-FEM solution u_{hp} (solid line) and the exact solution u (dashed line). Right: the right-hand side (7) (dashed line) and its L^2-projection (11) to the space V_{hp} (solid line).

Let $\Omega = (a, b) = (-1, 1)$,

$$f(x) = 200\mathrm{e}^{-10(x+1)}, \tag{7}$$

and consider a finite element mesh $\mathscr{T}_{hp}$ consisting of a single cubic element $K_1 = [a, b]$. The basis of the corresponding finite element space V_{hp} comprises the quadratic and cubic Lobatto shape functions l_2 and l_3, and the solution u_{hp} has the form $u_{hp}(x) = y_1 l_2(x) + y_2 l_3(x)$. By (6), the stiffness matrix is the identity matrix, and the unknown coefficients y_1, y_2 have the form

$$y_i = \int_{-1}^{1} \sum_{j=1}^{2} y_j l'_{j+1}(x) l'_{i+1}(x)\,\mathrm{d}x = \int_{-1}^{1} f(x) l_{i+1}(x)\,\mathrm{d}x, \quad i = 1, 2. \tag{8}$$

Using the right-hand side (7), we obtain that $y_1 = -\sqrt{6}(9 + 11\mathrm{e}^{-20})/10$ and $y_2 = \sqrt{10}(73 - 133\mathrm{e}^{-20})/100$. Thus the solution u_{hp} has the form

$$u_{hp}(x) = \tfrac{1}{40}(1 - x^2)(54 + 66\mathrm{e}^{-20} - (73 - 133\mathrm{e}^{-20})x). \tag{9}$$

Fig. 2 (left) shows that u_{hp} is not nonnegative in Ω, which means that the finite element approximation does not inherit the maximum principle from the continuous equation, i.e., the DMP does not hold.

To understand what happened, let us introduce the L^2-projection $f_{hp} \in V_{hp}$ of the right-hand side $f \in L^2(\Omega)$ to the space V_{hp} such that

$$\int_a^b (f_{hp}(x) - f(x)) v_{hp}(x)\,\mathrm{d}x = 0 \quad \text{for all } v_{hp} \in V_{hp}. \tag{10}$$

When expressing f_{hp} as a linear combination of the basis functions of V_{hp}, (10) yields a system of linear algebraic equations. The L^2-projection of the right-hand side (7),

$$f_{hp}(x) = \tfrac{3}{80}(1 - x^2)(110\mathrm{e}^{-20} + 90 + (931\mathrm{e}^{-20} - 511)x), \tag{11}$$

is negative in a subset of Ω, as illustrated in the right part of Fig. 2. Notice that (10) implies

$$\int_a^b f_{hp}(x) v_{hp}(x)\,\mathrm{d}x = \int_a^b f(x) v_{hp}(x)\,\mathrm{d}x \quad \text{for all } v_{hp} \in V_{hp},$$

and therefore it does not matter whether f or f_{hp} stands on the right-hand side of the discrete problem (3). In other words, the hp-FEM solution (9) depicted in Fig. 2 (left) corresponds to the right-hand side (11) which is not nonnegative in Ω.

5. Weak discrete maximum principle for hp-FEM

The above example motivates us to work with the L^2-projection f_{hp} of the right-hand side f onto V_{hp} rather than with f itself:

Definition 4. Let $f_{hp} \in V_{hp}$ be the L^2-projection of $f \in L^2(\Omega)$ to V_{hp} defined by (10). We say that problem (3) satisfies the weak discrete maximum principle (weak DMP) if the approximation $u_{hp} \in V_{hp}$ attains its minimum on the boundary of Ω whenever f_{hp} is nonnegative.

Notice that the classical DMP implies the weak DMP.

In the following Theorem 8 we prove the weak DMP for problem (3) under a technical assumption on existence of certain quadrature rules with nonnegative weights. This assumption is verified in Section 6.

Definition 5. Let $l_k(x)$, $k \geqslant 2$, be the Lobatto polynomials (5). For $(x, z) \in [-1, 1]^2$ and $p \geqslant 2$ we define the function

$$\Phi_p(x, z) = \sum_{k=1}^{p-1} l_{k+1}(x) l_{k+1}(z). \tag{12}$$

For $p = 1$ we define $\Phi_1(x, z) = 0$.

For $p \geqslant 1$, Φ_p is the discrete Green's function for problem (3) corresponding to a one-element mesh $K_1 = [-1, 1]$. Since $l_{i+1}(\pm 1) = 0$ for all $i \geqslant 1$, it is

$$\Phi_p(x, z) = 0 \quad \text{for all } (x, z) \in \Gamma, \tag{13}$$

where $\Gamma = \overline{(-1, 1)^2} \backslash (-1, 1)^2$.

Definition 6. Let $\mathcal{K}_p^+ \subset [-1, 1]^2$, $p \geqslant 1$, be a set of points, where $\Phi_p(x, z)$ is nonnegative, i.e.,

$$\mathcal{K}_p^+ = \{(x, z) \in [-1, 1]^2 : \Phi_p(x, z) \geqslant 0\}.$$

Finally, let $\mathcal{K}_p^+(x) \subset [-1, 1]$ be the cut of $\mathcal{K}_p^+$ at $x \in [-1, 1]$, i.e.,

$$\mathcal{K}_p^+(x) = \{z \in [-1, 1] : (x, z) \in \mathcal{K}_p^+\}.$$

Lemma 7. *We have symmetry relations* $\Phi_p(x, z) = \Phi_p(-x, -z) = \Phi_p(z, x)$ *for all* $p \geqslant 1$ *and* $(x, z) \in [-1, 1]^2$.

Proof. The identity $\Phi_p(x, z) = \Phi_p(z, x)$ follows immediately from (12). It follows from the fact that Legendre polynomials of odd/even degrees are odd/even functions, and from (5), that $l_k(x)$ is even for k even and that $l_k(x)$ is odd for k odd. Hence we have

$$\Phi_p(x, z) = \sum_{k=1}^{p-1} l_{k+1}(x) l_{k+1}(z) = \sum_{k=1}^{p-1} l_{k+1}(-x) l_{k+1}(-z) = \Phi_p(-x, -z)$$

for all $p \geqslant 1$. □

The main result of this paper is stated in Theorem 8:

Theorem 8. *Let $\Omega = (a, b) \subset \mathbb{R}$. Consider the discrete problem* (3) *on a mesh $\mathscr{T}_{hp}$ consisting of M finite elements $K_1, K_2, \ldots, K_M$ of polynomial degrees $p_1, p_2, \ldots, p_M$. If for every $p \in \{p_1, p_2, \ldots, p_M\}$ and every $x \in (-1, 1)$ there exists a quadrature rule $\mathscr{Q}_{2p}(x)$ such that*:

(i) *$\mathscr{Q}_{2p}(x)$ is exact for polynomials of degree $2p$ on $[-1, 1]$;*
(ii) *$\mathscr{Q}_{2p}(x)$ only has nonnegative weights*;
(iii) *$\mathscr{Q}_{2p}(x)$ has all points in $\mathscr{K}_p^+(x)$*;

then problem (3) *satisfies the weak DMP.*

Proof. Let us consider the exact solution $u \in H_0^1(\Omega)$ to the continuous problem (1) with a right-hand side $f \in L^2(\Omega)$. Let $0 \leqslant f_{hp} \in V_{hp}$ be the L^2-projection defined by (10). Then the approximation $u_{hp} \in V_{hp}$ is given by

$$\int_a^b u'_{hp}(x) v'_{hp}(x)\,\mathrm{d}x = \int_a^b f(x) v_{hp}(x)\,\mathrm{d}x = \int_a^b f_{hp}(x) v_{hp}(x)\,\mathrm{d}x \quad \forall v_{hp} \in V_{hp}. \tag{14}$$

In addition we introduce an auxiliary continuous problem: find $\tilde{u} \in H_0^1(\Omega)$ such that:

$$\int_a^b \tilde{u}'(x) v'(x)\,\mathrm{d}x = \int_a^b f_{hp}(x) v(x)\,\mathrm{d}x \quad \forall v \in H_0^1(\Omega).$$

It is well-known that when discretizing the Laplace operator in one spatial dimension by piecewise-linear finite elements, the approximation is exact at all grid vertices. The same holds for higher-order elements, which can be seen easily by using the orthogonality of higher-order basis functions (transformed Lobatto shape functions $l_2, l_3, \ldots$) to the lowest-order (piecewise linear) basis functions, see, e.g., [14]. In other words, we know that $u_{hp}(x_i) = u(x_i) = \tilde{u}(x_i)$ for all $i = 0, 1, \ldots, M$. Moreover, taking into account the (continuous) maximum principle, we have $\tilde{u} \geqslant 0$ in Ω and thus $u_{hp}(x_i) \geqslant 0$ for all $i = 0, 1, \ldots, M$. Therefore, it is sufficient to prove Theorem 8 for a single element $K_1 = \overline{\Omega}$, $\Omega = (-1, 1)$.

The solution u_{hp} is sought in the form

$$u_{hp}(x) = \sum_{i=1}^{p-1} y_i l_{i+1}(x). \tag{15}$$

By (6), relation (8) yields

$$y_i = \int_{-1}^{1} f_{hp}(z) l_{i+1}(z)\,\mathrm{d}z, \quad i = 1, 2, \ldots, p-1. \tag{16}$$

Putting (16) into (15), we obtain

$$u_{hp}(x) = \sum_{i=1}^{p-1} \left(\int_{-1}^{1} f_{hp}(z) l_{i+1}(z)\,\mathrm{d}z \right) l_{i+1}(x) = \int_{-1}^{1} f_{hp}(z) \Phi_p(x, z)\,\mathrm{d}z, \tag{17}$$

where $\Phi_p(x, z)$ is given by (12).

Let us now fix an arbitrary $x \in (-1, 1)$ and assume that there exists a quadrature rule $\mathscr{Q}_{2p}(x)$ with points $z_0, z_1, \ldots, z_{2p}$ in $\mathscr{K}_p^+(x)$ and nonnegative weights $w_0, w_1, \ldots, w_{2p}$. By (17) we have

$$u_{hp}(x) = \int_{-1}^{1} f_{hp}(z) \Phi_p(x, z)\,\mathrm{d}z = \sum_{i=0}^{2p} w_i f_{hp}(z_i) \Phi_p(x, z_i). \tag{18}$$

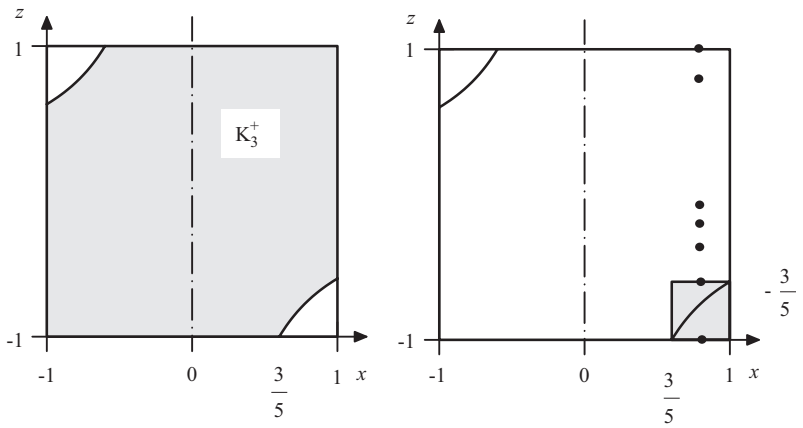


Fig. 3. Zero level set of the function $\Phi_3(x, z)$. Observe the set $\mathscr{K}_3^+$ (left) and the square $\mathscr{S}_3^- = (\frac{3}{5}, 1) \times (-1, -\frac{3}{5})$ (right). We also show the integration points (to be listed in Table 1).

Table 1
Case $p = 3$; 6th-order quadrature rule in $[-1, 1]$ with nonnegative weights and points outside of $(-1, -\frac{3}{5})$

Point	Weight	Point	Weight
-1	$\frac{271}{2268}$	$-\frac{3}{5}$	$\frac{2125}{3528}$
$-\frac{1}{5}$	$\frac{25}{252}$	0	$\frac{4}{63}$
$\frac{1}{5}$	$\frac{125}{189}$	$\frac{4}{5}$	$\frac{1700}{3969}$
1	$\frac{13}{504}$		

We have taken into account that $f_{hp}(z)\Phi_p(x, z)$ is a polynomial of degree at most $2p$ in z for fixed x and that $\mathscr{Q}_{2p}(x)$ is exact for all polynomials of this degree. By assumption, $w_i \geqslant 0$ for $i = 0, 1, \ldots, 2p$ and $f_{hp} \geqslant 0$ in $\Omega = (-1, 1)$. Since $z_i \in \mathscr{K}_p^+(x)$ then also $\Phi_p(x, z_i) \geqslant 0$. Hence, it follows from (18) that $u_{hp}(x) \geqslant 0$ for any $x \in (-1, 1)$, and thus the minimum of $u_{hp}(x)$ is attained on the boundary of Ω. □

6. Verification of assumptions to Theorem 8

It is easy to see that for $p = 2, 4, 6$ the function $\Phi_p(x, z)$ is nonnegative in $[-1, 1]^2$ and therefore even the classical DMP holds. For every other polynomial degree $p \geqslant 2$ one has to find a quadrature rule $\mathscr{Q}_{2p}(x)$ with nonnegative weights and points in $\mathscr{K}_p^+(x)$. By symmetry (see Lemma 7) it is enough to find such quadrature rule for $x \in [0, 1)$ only. The construction of the quadrature rules is not difficult. For spatial limitations, let us illustrate the procedure for $p \leqslant 10$ only.

Odd polynomial degrees: Let us start with $p = 3$. In this case we have

$$\Phi_3(x, z) = \tfrac{1}{8}(1 - x^2)(1 - z^2)(3 + 5xz). \tag{19}$$

Clearly, $\Phi_3(x, z) = 0$ on the curves $x = \pm 1$, $z = \pm 1$, and $xz = -\frac{3}{5}$ (see Fig. 3). The domain $\mathscr{K}_p^+$ is bounded by these curves.

Thus it is enough to find a quadrature rule with nonnegative weights and points in $[-1, 1]$ but outside $(-1, -\frac{3}{5})$, which is exact for all polynomials of degree at most 6. An example of such quadrature rule is given in Table 1.

The situation for $p = 5, 7, 9$ is similar to the case of $p = 3$. The sets $\mathscr{K}_p^+$ have similar shapes with the only difference that the regions of negativity become smaller with growing p. The regions of negativity in $[0, 1) \times (-1, 1)$ can be

Table 2
Case $p = 5$; 10th-order quadrature rule in $[-1, 1]$ with nonnegative weights and points outside of $(-1, -0.811)$

Point	Weight	Point	Weight
−1	0.0534286192	−0.811	0.3054087580
−0.59	0.0030544353	−0.42	0.4473230113
−0.2	0.0066984041	0	0.2760767276
0.2	0.2939694773	0.43	0.0149245373
0.6	0.3805105712	0.9	0.1999066353
1	0.0186988234		

Table 3
Case $p = 7$; 14th-order quadrature rule in $[-1, 1]$ with nonnegative weights and points outside of $(-1, -0.89)$

Point	Weight	Point	Weight
−1	0.0306200311	−0.89	0.1806438688
−0.75	0.0016558668	−0.65	0.2862680475
−0.45	0.0379885258	−0.31	0.2988638595
−0.16	0.0833146476	0.1	0.3554921618
0.16	0.0113639321	0.35	0.0204292124
0.47	0.3218682171	0.734	0.1289561668
0.80	0.1314089188	0.955	0.1093567805
1	0.0017697634		

Table 4
Case $p = 9$; 18th-order quadrature rule in $[-1, 1]$ with nonnegative weights and points outside of $(-1, -0.93)$

Point	Weight	Point	Weight
−1	0.01937406240	−0.93	0.1153128270
−0.885	0.00157968340	−0.772	0.1947443595
−0.65	0.00126499680	−0.55	0.2341166464
−0.4	0.06286669339	−0.25	0.2438572426
−0.08	0.08588496537	0.08	0.2395820916
0.19	0.04691799156	0.38	0.2665159766
0.6	0.00216030838	0.625	0.2029738760
0.73	0.04687189997	0.83	0.1072052560
0.89	0.06009091818	0.97	0.0648680095
1	0.00381219535		

safely enclosed in squares

$$\mathscr{S}_3^- = (\tfrac{3}{5}, 1) \times (-1, -\tfrac{3}{5}),$$
$$\mathscr{S}_5^- = (0.811, 1) \times (-1, -0.811),$$
$$\mathscr{S}_7^- = (0.89, 1) \times (-1, -0.89),$$
$$\mathscr{S}_9^- = (0.93, 1) \times (-1, -0.93).$$

These squares (as well as the domains for even polynomial degrees) were defined by investigation of the zero level sets of $\Phi_p(x, z)$. We used Maple to locate approximately the zero level sets of the discrete Green's functions Φ_p. After that, rigorous proof of their nonnegativity in $(-1, 1)^2$ minus these areas was performed using an adaptive interval computation technique in integer arithmetics. More details on this step can be found in [15]. Examples of quadrature rules required by Theorem 8 are shown in Tables 2–4.

Even polynomial degrees: For $p = 8$, there are four areas where the function $\Phi_p(x, z)$ is negative (see Fig. 4).

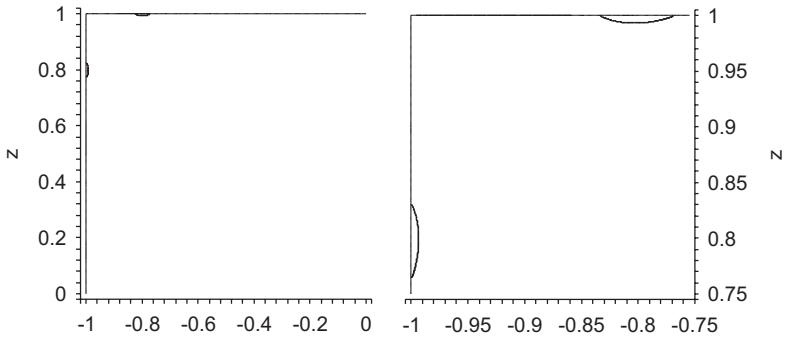


Fig. 4. Zero level set of the kernel $\Phi_8(x, z)$ in the second quadrant (left) and a detail of the upper left corner.

Table 5
Case $p = 8$; 16th-order quadrature rule in $[-1, 1]$ with nonnegative weights and points outside of (0.75, 0.85)

Point	Weight	Point	Weight
−1	0.0137599529	−0.9564181650	0.0618586932
−0.8854980347	0.0892150513	−0.7582972896	0.1646935265
−0.5719162652	0.1875234174	−0.4628139806	0.0729252387
−0.2917166274	0.2435469772	−0.0811621291	0.0841621866
−0.0061521460	0.1800939083	0.1655560030	0.1320371771
0.3391628868	0.2286184297	0.5726348225	0.2184036287
0.75	0.1285378345	0.85	0.0908051678
0.9230637084	0.0427456544	0.9648584341	0.0509010934
1	0.0101720626		

Table 6
Case $p = 8$; 16th-order quadrature rule in $[-1, 1]$ with nonnegative weights and points outside of (0.98, 1)

Point	Weight	Point	Weight
−1	0.0097495069	−0.9548248562	0.0857520162
−0.8409569422	0.1018591390	−0.7825414112	0.0149475627
−0.7708636219	0.0926211201	−0.5747624113	0.2476049720
−0.3937499257	0.0549434125	−0.3273530867	0.0276562411
−0.2532942335	0.2543287199	0.0382371812	0.2892622856
0.2837396038	0.1910189889	0.4501581170	0.1560300966
0.5808907063	0.1246581226	0.7443822112	0.1842879621
0.8927849373	0.0841645246	0.9421667341	0.0612885001
1	0.0198268291		

Two of these areas lie inside the rectangles $(-1, -0.98) \times (0.75, 0.85)$ and $(-0.85, -0.75) \times (0.98, 1)$, and the other two are located symmetrically at the opposite corner of $[-1, 1]^2$. The points and weights of the corresponding quadrature rules are listed in Tables 5 and 6. Thus, we have $u_{hp} \geqslant 0$ for $x \in (-1, 0]$. The nonnegativity of $u_{hp}(x)$ for $x \in (0, 1)$ follows from symmetry again.

The case $p = 10$ is similar to $p = 8$. There are four areas where the function $\Phi_{10}(x, z)$ is negative, analogously to the 8th-order case (see Fig. 5).

Two of these areas are inside the rectangles $(-1, -0.986) \times (0.82, 0.91)$ and $(-0.91, -0.82) \times (-0.986, 1)$ and the other two are located symmetrically at the opposite corner of $[-1, 1]^2$. The points and weights of the corresponding quadrature rules are listed in Tables 7 and 8, respectively. By symmetry, $u_{hp}(x) \geqslant 0$ also for $x \in (0, 1)$.

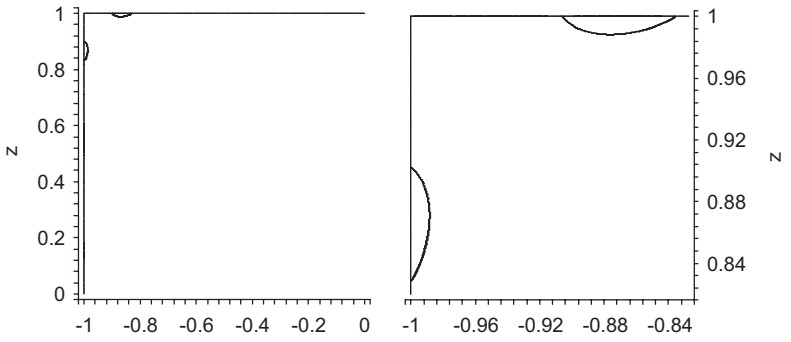


Fig. 5. Zero level set of the kernel $\Phi_{10}(x, z)$ in the second quadrant (left) and a detail of the upper left corner.

Table 7
Case $p = 10$; 20th-order quadrature rule in $[-1, 1]$ with nonnegative weights and points outside of (0.82, 0.91)

Point	Weight	Point	Weight
−1	0.0127411726	−0.9569019461	0.0603200758
−0.9344466123	0.0183508422	−0.8574545411	0.1032513172
−0.7530104489	0.1106942630	−0.6362178184	0.0412386636
−0.6061244531	0.1295220930	−0.4275824090	0.1937516842
−0.2340018112	0.1916905139	−0.0454114485	0.1774661870
0.0754465671	0.0755419308	0.1672504233	0.0745275871
0.2516247645	0.1488965177	0.3707975798	0.0207086237
0.4366736344	0.1397170181	0.5306011976	0.0924918512
0.6745457042	0.1639628301	0.82	0.1200387168
0.91	0.0649445615	0.9667274132	0.0502362251
1	0.0099073255		

Table 8
Case $p = 10$; 20th-order quadrature rule in $[-1, 1]$ with nonnegative weights and points outside of (0.986, 1)

Point	Weight	Point	Weight
Points: −1	0.0129961117	−0.9609467424	0.0393058650
−0.9366001558	0.0472129994	−0.8686571459	0.0307704321
−0.8222969304	0.1127110155	−0.6830858117	0.1442049485
−0.5515874908	0.1263749495	−0.4070028385	0.1615584597
−0.2391731402	0.1767071143	−0.0805321378	0.0223802647
−0.0404112041	0.1755155830	0.0382998004	0.0409103698
0.2054285570	0.2302298514	0.4168373782	0.1495405342
0.4862170553	0.0877842194	0.6284448676	0.0980645550
0.6932595712	0.1047143177	0.83041757281	0.1311485592
0.93562906418	0.0774056021	0.986	0.0267375743
1	0.0037266735		

Remark 9. It is worth mentioning that the points in Tables 1–8 were chosen to be rational numbers. The corresponding weights were obtained via the formula

$$w_i = \int_{-1}^{1} \mathscr{L}_i(x)\,\mathrm{d}x,$$

where $\mathscr{L}_i \in P^p(-1, 1)$ is the elementary Lagrange interpolation polynomial, $\mathscr{L}_i(z_j) = \delta_{ij}$, $0 \leqslant i, j \leqslant p$. In particular, if follows from here that the weights also are rational numbers. We have used Maple and its integer arithmetics to find all the weights listed in Tables 1–8; they are shown as decimals for printing purposes only.

7. Conclusions and future work

Virtually all existing results related to the analysis of discrete maximum principles are based on *M*-matrices and thus limited to lowest-degree approximations, such as finite differences or piecewise-linear finite elements. In this paper, we presented a new methodology which is based on the analysis of the discrete Green's function. The main advantage of this alternative approach is that it works in the same way both for piecewise-linear and higher-order finite element approximations.

It was demonstrated in Section 4 that the standard discrete maximum principle, as it is known for lowest-order approximations, did not work for higher-order elements. As a remedy we proposed that one should look at the L^2-projection of the load function f onto the finite element space instead of working with f itself.

The computation of the L^2-projection of f onto the finite element space involves the solution of a large system of linear algebraic equations. The linear system is much better conditioned (i.e., less stiff) compared to the discrete problem itself, but still the test is CPU demanding. Therefore, it is among our priorities to improve the practical usefulness of the criterion by finding alternative conditions which would be easier to verify. At the same time, the analysis of the discrete Green's function for two-dimensional elements (which is defined in $\mathbb{R}^4$) is in progress.

Acknowledgements

The first author gratefully acknowledges the financial support of the U.S. Department of Defense under Grant no. 05PR07548-00, of the NSF under Grant no. DMS-0532645, and of the Grant Agency of the Czech Republic under Grant no. 102-05-0629. The second author was supported in part by the Grant Agency of the Czech Republic, project no. 201/04/P021 and by the Academy of Sciences of the Czech Republic, Institutional Research Plan no. AV0Z10190503. The authors also wish to thank anonymous referees for their valuable suggestions which helped to improve the quality of the paper.

References

[1] I. Babuška, B.Q. Guo, Approximation properties of the *hp* version of the finite element method, Comput. Methods Appl. Mech. Eng. 133 (1996) 319–346.
[2] I. Babuška, T. Strouboulis, Finite Element Method and its Reliability, Clarendon Press, Oxford, 2001.
[3] P.G. Ciarlet, Discrete maximum principle for finite difference operators, Aequationes Math. 4 (1970) 338–352.
[4] P.G. Ciarlet, P.A. Raviart, Maximum principle and uniform convergence for the finite element method, Comput. Methods Appl. Mech. Eng. 2 (1973) 17–31.
[5] L. Demkowicz, et al., Toward a universal *hp*-adaptive finite element strategy, part 1: constrained approximation and data structure, Comput. Methods Appl. Math. Eng. 77 (1989) 79–112.
[6] M. Fiedler, Special Matrices and their Applications in Numerical Mathematics, Martinus Nijhoff Publishers, Dordrecht, 1986.
[7] W. Höhn, H.D. Mittelmann, Some remarks on the discrete maximum principle for finite elements of higher-order, Computing 27 (1981) 145–154.
[8] J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, Numer. Math. 99 (2005) 669–698.
[9] S. Korotov, M. Křížek, P. Neittaanmäki, Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle, Math. Comput. 70 (2000) 107–119.
[10] M. Křížek, L. Liu, On the maximum and comparison principles for a steady-state nonlinear heat conduction problem, ZAMM Z. Angew. Math. Mech. 83 (2003) 559–563.
[11] J.T. Oden, S. Prudhomme, Goal-oriented error estimation and adaptivity for the finite element method, Comput. Math. Appl. 41 (2001) 735–756.
[12] C. Schwab, *p*- and *hp*-Finite Element Methods, Clarendon Press, Oxford, 1998.
[13] P. Šolín, Partial Differential Equations and the Finite Element Method, Wiley, New York, 2005.
[14] P. Šolín, K. Segeth, I. Doležel, Higher-order Finite Element Methods, Chapman & Hall/CRC Press, Boca Raton, 2003.
[15] P. Šolín, T. Vejchodský, R. Araiza, Discrete conservation of nonnegativity for elliptic problems solved by the *hp*-FEM, Math. Comput. Simulat., (2006), accepted for publication.

[16] B. Szabó, I. Babuška, Finite Element Analysis, Wiley, New York, 1991.
[17] R.S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[18] T. Vejchodský, On the nonnegativity conservation in semidiscrete parabolic problems, in: M. Křížek, P. Neittaanmäki, R. Glowinski, S. Korotov (Eds.), Conjugate Gradients Algorithms and Finite Element Methods, Springer, Berlin, 2004, pp. 197–210.
[19] T. Vejchodský, Method of lines and conservation of nonnegativity, in: Proceedings of the European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2004), Jyväskylä, Finland, 2004, 18pp.

BIBLIOGRAPHY

[1] N. Aronszajn and K. T. Smith: Characterization of positive reproducing kernels. Applications to Green's functions. *Amer. J. Math.* **79** (1957), 611–622.

[2] O. Axelsson and L. Kolotilina: Monotonicity and discretization error estimates. *SIAM J. Numer. Anal.* **27** (1990), 1591–1611.

[3] A. Berman and R. J. Plemmons: *Nonnegative matrices in the mathematical sciences*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.

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

[5] 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), 117–132.

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

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

[8] J. H. Brandts, S. Korotov, and M. Křížek: Simplicial finite elements in higher dimensions. *Appl. Math.* **52** (2007), 251–265.

[9] 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), 2344–2357.

[10] S. C. Brenner and L. R. Scott: *The mathematical theory of finite element methods*, Springer, New York, third edn., 2008.

[11] M. Breuß, V. Dolejší, and A. Meister: Anisotropic adaptive resolution of boundary layers for heat conduction problems. *ZAMM Z. Angew. Math. Mech.* **86** (2006), 450–463.

[12] J. D. Burago and V. A. Zalgaller: Polyhedral embedding of a net. *Vestnik Leningrad. Univ.* **15** (1960), 66–80.

[13] 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), 641–646.

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

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

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

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

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

[19] P. G. Ciarlet and R. S. Varga: Discrete variational Green's function. II. One dimensional problem. *Numer. Math.* **16** (1970), 115–128.

[20] R. Dautray and J.-L. Lions: *Mathematical analysis and numerical methods for science and technology. Vol. 2*, Springer-Verlag, Berlin, 1988.

[21] L. Demkowicz: *Computing with hp-adaptive finite elements. Vol. 1*, Chapman & Hall/CRC, Boca Raton, FL, 2007.

[22] 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), 1–23 (electronic).

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

[24] M. A. Eisenberg and L. E. Malvern: On finite element integration in natural co-ordinates. *Int. J. Numer. Methods Eng.* **7** (1973), 574–575.

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

[26] R. Eymard, D. Hilhorst, and M. Vohralík: A combined finite volume–nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems. *Numer. Math.* **105** (2006), 73–131.

[27] I. Faragó: *Numerical treatment of linear parabolic problems*, Thesis for the Doctor of Hungarian Academy of Sciences, Eötvös Loránd University, Budapest, 2008.

[28] I. Faragó: Discrete maximum principle for finite element parabolic models in higher dimensions. *Math. Comput. Simulation* **80** (2010), 1601–1611.

[29] I. Faragó and R. Horváth: Discrete maximum principle and adequate discretizations of linear parabolic problems. *SIAM J. Sci. Comput.* **28** (2006), 2313–2336 (electronic).

[30] I. Faragó and R. Horváth: Continuous and discrete parabolic operators and their qualitative properties. *IMA J. Numer. Anal.* **29** (2009), 606–631.

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

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

[33] L. E. Fraenkel: *An introduction to maximum principles and symmetry in elliptic problems*, Cambridge University Press, Cambridge, 2000.

[34] H. Fujii: Some remarks on finite element analysis of time-dependent field problems. In: *Theory and practice in finite element structural analysis*, Univ. Tokyo Press, Tokyo, 1973, pp. 91–106.

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

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

[37] M. Grüter and K.-O. Widman: The Green function for uniformly elliptic equations. *Manuscripta Math.* **37** (1982), 303–342.

[38] A. Hannukainen, S. Korotov, and T. Vejchodský: On weakening conditions for discrete maximum principles for linear finite element schemes. In: S. Margenov, L. Vulkov, and J. Wasniewski (eds.), *Numerical Analysis and Its Applications, Lecture Notes in Computer Science 5434*, Springer-Verlag, Berlin, 2009, pp. 297–304.

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

[40] E. Hopf: Elementare Bemerkungen über die Lösungen partieller Differentialgleichungen zweiter Ordnung vom elliptischen Typus. *Sitzungsberichte Akad. Berlin* **1927** (1927), 147–152.

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

[42] D. Jackson: *Fourier series and orthogonal polynomials*, Dover Publications Inc., Mineola, NY, 2004.

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

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

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

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

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

[48] K. Kreith: Criteria for positive Green's functions. *Illinois J. Math.* **12** (1968), 475–478.

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

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

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

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

[53] M. Křížek and P. Neittaanmäki: *Finite element approximation of variational problems and applications*, Longman Scientific & Technical, Harlow, 1990.

[54] M. Křížek and P. Neittaanmäki: *Mathematical and numerical modelling in electrical engineering*, Kluwer Academic Publishers, Dordrecht, 1996.

[55] N. V. Krylov: *Lectures on elliptic and parabolic equations in Sobolev spaces*, American Mathematical Society, Providence, RI, 2008.

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

[57] O. A. Ladyzhenskaya and N. N. Ural′tseva: *Linear and quasilinear elliptic equations*, Academic Press, New York, 1968.

[58] H. Maehara: Acute triangulations of polygons. *European J. Combin.* **23** (2002), 45–55.

[59] V. Maz'ya and J. Rossmann: *Elliptic equations in polyhedral domains*, American Mathematical Society, Providence, RI, 2010.

[60] A. Mazzia: An analysis of monotonicity conditions in the mixed hybrid finite element method on unstructured triangulations. *Internat. J. Numer. Methods Engrg.* **76** (2008), 351–375.

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

[62] J. M. Nordbotten, I. Aavatsmark, and G. T. Eigestad: Monotonicity of control volume methods. *Numer. Math.* **106** (2007), 255–288.

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

[64] A. Prestel and C. N. Delzell: *Positive polynomials*, Springer-Verlag, Berlin, 2001.

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

[66] P. Pucci and J. Serrin: *The maximum principle*, Birkhäuser Verlag, Basel, 2007.

[67] Y. Roitberg: *Elliptic boundary value problems in the spaces of distributions*, Kluwer Academic Publishers Group, Dordrecht, 1996.

[68] H.-G. Roos, M. Stynes, and L. Tobiska: *Robust numerical methods for singularly perturbed differential equations*, Springer-Verlag, Berlin, second edn., 2008.

[69] V. Ruas Santos: On the strong maximum principle for some piecewise linear finite element approximate problems of nonpositive type. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **29** (1982), 473–491.

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

[71] P. Šolín: *Partial differential equations and the finite element method*, John Wiley & Sons, Hoboken, NJ, 2006.

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

[73] I. Stakgold: *Green's functions and boundary value problems*, John Wiley & Sons Inc., New York, second edn., 1998.

[74] B. Szabó and I. Babuška: *Finite element analysis*, John Wiley & Sons Inc., New York, 1991.

[75] V. Thomée and L. B. Wahlbin: On the existence of maximum principles in parabolic finite element equations. *Math. Comp.* **77** (2008), 11–19 (electronic).

[76] E. VanderZee, A. N. Hirani, V. Zharnitsky, and D. Guoy: A dihedral acute triangulation of the cube. *Comput. Geom.* **43** (2010), 445–452.

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

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

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

[80] T. Vejchodský: Comparison principle for a nonlinear parabolic problem of a nonmonotone type. *Appl. Math. (Warsaw)* **29** (2002), 65–73.

[81] T. Vejchodský: On the nonnegativity conservation in semidiscrete parabolic problems. In: M. Křížek, P. Neittaanmäki, R. Glowinski, and S. Korotov (eds.), *Conjugate gradient algorithms and finite element methods*, Springer, Berlin, 2004, pp. 197–210.

[82] T. Vejchodský, S. Korotov, and A. Hannukainen: Discrete maximum principle for parabolic problems solved by prismatic finite elements. *Math. Comput. Simulation* **80** (2010), 1758–1770.

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

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

[85] E. G. Yanik: Sufficient conditions for a discrete maximum principle for high order collocation methods. *Comput. Math. Appl.* **17** (1989), 1431–1434.

[86] L. Yuan: Acute triangulations of polygons. *Discrete Comput. Geom.* **34** (2005), 697–706.

LIST OF ATTACHED PAPERS

[A1] 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), 275–287.

[A2] S. Korotov and T. Vejchodský: A comparison of simplicial and block finite elements. In: G. Kreiss, P. Lötstedt, A. Målqvist, and M. Neytcheva (eds.), *Numerical mathematics and advanced applications ENUMATH 2009*, Springer, Berlin, 2010, pp. 533–541.

[A3] T. Vejchodský and P. Šolín: Discrete maximum principle for higher-order finite elements in 1D. *Math. Comp.* **76** (2007), 1833–1846.

[A4] 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), 201–214.

[A5] 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), 233–243.

[A6] T. Vejchodský: Higher-order discrete maximum principle for 1D diffusion-reaction problems. *Appl. Numer. Math.* **60** (2010), 486–500.

[A7] T. Vejchodský: Angle conditions for discrete maximum principles in higher-order FEM. In: G. Kreiss, P. Lötstedt, A. Målqvist, and M. Neytcheva (eds.), *Numerical mathematics and advanced applications ENUMATH 2009*, Springer, Berlin, 2010, pp. 901–909.

[A8] P. Šolín and T. Vejchodský: A weak discrete maximum principle for *hp*-FEM. *J. Comput. Appl. Math.* **209** (2007), 54–65.