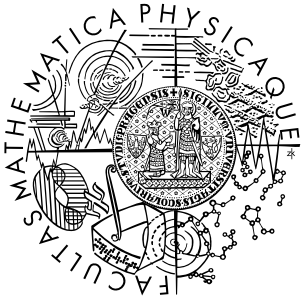
Charles University in Prague
Faculty of Mathematics and Physics

DIPLOMA THESIS



Bc. Martina Klejchová

The discrete maximum principle in the first order finite element method

Department of Numerical Mathematics

Supervisor: RNDr. Tomáš Vejchodský, Ph.D.

Study programme: Mathematics, Numerical and Computational Mathematics

2010

I would like to thank my supervisor RNDr. Tomáš Vejchodský, Ph.D. for careful leading and valuable help with writing this thesis.

I declare that I wrote this thesis by myself and with the exclusive use of the material I quote. I agree the thesis to be lent.

6th August 2010 in Prague Martina Klejchová

Contents

Název práce: Diskrétní princip maxima v metodě konečných prvků prvního řádu
Autor: Bc. Martina Klejchová
Katedra: Katedra numerické matematiky
Vedoucí diplomové práce: RNDr. Tomáš Vejchodský, Ph.D.
e-mail vedoucího: vejchod@math.cas.cz

Abstrakt:
Tato diplomová práce se zabývá studiem diskrétního principu maxima pro reakčně-difuzní úlohu, jež je řešena metodou konečných prvků s různými typy prvků. Práce obsahuje stručnou charakteristiku řešené úlohy a použité numerické metody, definici diskrétního pricipu maxima a obecné podmínky pro jeho platnost. Jádro práce tvoří analýza geometrických podmínek na tvar a případně i velikost jednotlivých konečných prvků tak, aby platil diskrétní princip maxima. Konkrétně analyzujeme podmínky pro intervaly, trojúhelníky, čtyřstěny, kvádry a trojboké hranoly.

Klíčová slova: reakčně difuzní úloha, diskrétní princip maxima, metoda konečných prvků, dělení na intervaly, trojúhelníková síť, síť obdelníků, síť čtyřstěnů, síť kvádrů, síť trojbokých hranolů

Title: The discrete maximum principle in the first order finite element method
Author: Bc. Martina Klejchová
Department: Department of Numerical Mathematics
Supervisor: RNDr. Tomáš Vejchodský, Ph.D.
Supervisor's e-mail address: vejchod@math.cas.cz

Abstract:
In this thesis we study the discrete maximum principle for a diffusion-reaction problem solved by means of various types of finite elements. The work includes the brief characterization of the problem and of the finite element method, definition of the discrete maximum principle and general conditions for its validity. The main focus of this work is in the analysis of such geometric conditions for the shape and possibly also for the size of the specific finite elements that guarantee the validity of the discrete maximum principle. Namely, we analyze conditions for intervals, triangles, rectangles, tetrahedra, blocks and triangular prisms.

Keywords: diffusion-reaction problem, discrete maximum principle, finite element method, interval partition, triangular partition, rectangular partition, tetrahedral partition, block partition, prismatic partition

Chapter 1

Introduction

The existence of maximum principles is an important aspect of many mathematical models based on elliptic and parabolic partial differential equations (see e.g. [18], [30], [31]). It is a widely used analytic tool in the study of second order linear and nonlinear equations. It can help to determine many qualitative properties about solutions of these equations without the need of knowing the solution in advance. It also forms an important part of several proofs of existence and uniqueness theorems.

The maximum principle is not only a mathematical feature. It finds important applications especially in real-life problems with physically non-negative quantities like the temperature, density, or concentration, because it adequately mirrors the real behavior of physical systems. Since the exact solutions of these problems satisfy the maximum principle, it is quite natural to require the preservation of similar property by the corresponding approximations. We speak about the discrete maximum principle.

Hence, it is not surprising that many authors concentrate their effort on the large field of construction of approximations satisfying the discrete maximum principle. The first work dealing with the validity of the discrete maximum principle for elliptic problems in the context of the finite difference method was probably [34] in 1966. Later on, many other works on this subject were presented (e.g. [2], [3], [7], [10], [19], [20], [26], [27] and many others). The validity of the discrete maximum principles was studied also for nonlinear elliptic equations, e.g. in [24] or [29], for parabolic equations, see e.g. [12], [13], [14], [15] or [17]. There are also results of analysis of the discrete maximum principle for higher-order finite elements, presented e.g. in [22], [35], [36], [37], [39].

Besides the previous results for the finite element approximations, several new methods were developed, that preserve the discrete maximum principle without any restriction on the underlined mesh. These schemes, however, are typically nonlinear even for linear problems – see e.g. [4] for Laplacian or [38] for a convection-diffusion problem.

The aim of this thesis is to summarize the known results for the reaction-diffusion problem numerically solved by the lowest order finite element approxi-

mations on different types of meshes. We systematically investigate 1D, 2D and 3D problems and analyze the corresponding finite elements: intervals, triangles, rectangles, tetrahedra, blocks and triangular prisms. We present sufficient conditions for the validity of the discrete maximum principle in these special cases and show their geometrical interpretations as restrictions on shapes and sizes of specific finite elements.

The thesis is organized as follows. Chapter 1 brings in six sections the required theoretical background. It includes the definitions of the reaction-diffusion problem and of the continuous maximum principle (Section 2.1), description of the Galerkin and finite element methods (Sections 2.2 and 2.3), definition of the discrete maximum principle and certain general results concerning its validity (Section 2.4), and finally the last two sections of this chapter are about M-matrices and local mass and stiffness matrices. In the following chapters results for the specific types of finite elements are presented. In Chapter 3 we analyze 1D meshes, in Chapter 4 triangles, in Chapter 5 rectangles, in Chapter 6 tetrahedra, in Chapter 7 blocks (i.e. cartesian products of intervals) and in Chapter 8 triangular prisms. Chapter 9 then gives a summary and points out possible generalizations and open problems.

Chapter 2

General framework

2.1 Definition of the problem

Assume an open bounded domain $\Omega \subset \mathbb{R}^d$, $d \geq 1$, with Lipschitz-continuous boundary $\partial\Omega$. We consider the following problem whose classical formulation reads:
Find $u \in \mathcal{C}(\overline{\Omega}) \cup \mathcal{C}^2(\Omega)$ such that

$$-\Delta u + \kappa^2 u = f \quad \text{in} \quad \Omega, \; u = 0 \quad \text{on} \quad \partial\Omega, \tag{2.1}$$

where $\kappa^2 \in \mathcal{C}(\overline{\Omega})$ and $f \in \mathcal{C}(\overline{\Omega})$ are given functions. To solve this problem by the finite element method we introduce its weak formulation first.

Weak formulation

To derive the weak formulation of problem (2.1), we multiply (2.1) by a test function $v \in \mathcal{C}_0^\infty(\Omega)$ and integrate it over Ω. Then by using Green's theorem and the fact that $v = 0$ on $\partial\Omega$ we get the following identity:

$$\int_\Omega \left(\nabla u \cdot \nabla v + \kappa^2 uv\right) \mathrm{d}x = \int_\Omega fv \, \mathrm{d}x$$

All these integrals remain finite even for $u, v \in V$, where $f \in \mathcal{L}^2(\Omega)$, $\kappa^2 \in \mathcal{L}^\infty(\Omega)$ and

$$V = H_0^1(\Omega) = \left\{ w \in H^1(\Omega), \; w = 0 \text{ on } \partial\Omega \text{ in the sense of traces} \right\}.$$

This motivates the following definition.

Definition 2.1 (Weak formulation)**.** *Given $f \in \mathcal{L}^2(\Omega)$ and $\kappa^2 \in \mathcal{L}^\infty(\Omega)$. Find $u \in V$ such that*

$$\int_\Omega \left(\nabla u \cdot \nabla v + \kappa^2 uv\right) \mathrm{d}x = \int_\Omega fv \, \mathrm{d}x \quad \forall v \in V. \tag{2.2}$$

Now we define a bilinear form $a(\cdot,\cdot) : V \times V \mapsto \mathbb{R}$,

$$a(u,v) = \int_\Omega \left(\nabla u \cdot \nabla v + \kappa^2 uv\right) \mathrm{d}x, \tag{2.3}$$

and a linear form $F : V \mapsto \mathbb{R}$,

$$F(v) = \int_\Omega fv \,\mathrm{d}x. \tag{2.4}$$

Thus, the weak formulation can be written as follows:
Find $u \in V$ such that

$$a(u,v) = F(v) \quad \forall v \in V. \tag{2.5}$$

Existence and uniqueness of the weak solution

The existence and uniqueness of the weak solution of problem (2.5) follows from the well known Lax-Milgram lemma. For completeness, we recall this lemma and verify its assumptions.

Lemma 2.1 (Lax-Milgram)**.** *Let V be a Hilbert space, $a(\cdot,\cdot) : V \times V \mapsto \mathbb{R}$ a bounded V-elliptic bilinear form and $F \in V'$ (V' is a dual space to V). Then there exists a unique weak solution to problem* (2.5)*.*

Proof.
See e.g. [32, p. 18].

□

A bilinear form $a(u,v)$ is said to be bounded if there exists a constant $C_a > 0$ such that $|a(u,v)| \le C_a \|u\|_V \|v\|_V$ for all $u, v \in V$, where the norm $\|\cdot\|_V = \|\cdot\|_{H^1(\Omega)}$. Similarly $a(u,v)$ is V-elliptic if there exists a constant $C_b > 0$ such that $a(u,u) \ge C_b \|u\|_V^2$ for all $u \in V$. The boundedness of bilinear form (2.3) follows from the Hölder inequality (see e.g. [32, p. 381]). Obviously,

$$|a(u,v)| \le \int_\Omega \left(|\nabla u \cdot \nabla v| + \kappa^2 |uv|\right) \mathrm{d}x \le \max\left\{1, \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\Omega)}\right\} \left(\int_\Omega |\nabla u \cdot \nabla v| \,\mathrm{d}x + \int_\Omega |uv| \,\mathrm{d}x\right).$$

Then the Hölder inequality yields

$$\int_\Omega |\nabla u \cdot \nabla v| \,\mathrm{d}x \le \left(\int_\Omega |\nabla u|^2 \,\mathrm{d}x\right)^{1/2} \left(\int_\Omega |\nabla v|^2 \,\mathrm{d}x\right)^{1/2} = |u|_{H^1(\Omega)} |v|_{H^1(\Omega)}$$

and

$$\int_\Omega |uv| \,\mathrm{d}x \le \left(\int_\Omega |u|^2 \,\mathrm{d}x\right)^{1/2} \left(\int_\Omega |v|^2 \,\mathrm{d}x\right)^{1/2} = \|u\|_{\mathcal{L}^2(\Omega)} \|v\|_{\mathcal{L}^2(\Omega)}.$$

Using the Cauchy-Schwarz inequality (see e.g. [32, p. 391]) we obtain

$$|u|_{H^1(\Omega)}\,|v|_{H^1(\Omega)} + \|u\|_{\mathcal{L}^2(\Omega)}\,\|v\|_{\mathcal{L}^2(\Omega)} \leq \left(|u|^2_{H^1(\Omega)} + \|u\|^2_{\mathcal{L}^2(\Omega)}\right)^{1/2} \left(|v|^2_{H^1(\Omega)} + \|v\|^2_{\mathcal{L}^2(\Omega)}\right)^{1/2} = \|u\|_V\,\|v\|_V\,.$$

Therefore

$$|a(u,v)| \leq C_a\,\|u\|_V\,\|v\|_V\,,$$

with $C_a = \max\left\{1, \|\kappa^2\|_{\mathcal{L}^\infty(\Omega)}\right\}$. To prove the ellipticity we use the Fridrichs inequality (see e.g [32, p. 422]):

$$|a(u,u)| = \int_\Omega \left(|\nabla u|^2 + \kappa^2 u^2\right) \mathrm{d}x \geq \int_\Omega |\nabla u|^2\,\mathrm{d}x = |u|^2_{H^1(\Omega)} \geq C_b\,\|u\|^2_V\,.$$

Hence, all the assumptions of Lax-Milgram lemma are fulfilled and problem (2.5) has a unique solution.

The maximum principle

Problem (2.1) is well known to satisfy the maximum principle. Here we present a version of this principle in the weak setting.

Definition 2.2 (The maximum principle (MaxP))**.** *Let $u \in V$ be the weak solution of problem (2.5) with $f \in \mathcal{L}^2(\Omega)$. This problem satisfies the maximum principle if*

$$\operatorname*{ess\,sup}_{\overline{\Omega}}\ u \leq \max\left\{0, \operatorname*{ess\,sup}_{\partial\Omega}\ u\right\} \ \textit{for all}\ \ f(x) \leq 0\ \ \textit{for a.e.}\ x \in \Omega,$$

where the essential supremum of u in $\overline{\Omega}$ is given by

$$\operatorname*{ess\,sup}_{\overline{\Omega}}\ u := \inf\left\{a \in \mathbb{R} : \mu\left(\left\{x \in \overline{\Omega} : f(x) > a\right\}\right) = 0\right\}$$

and μ denotes the Lebesgue measure on $\mathbb{R}^d$.

Remark 2.1. *Clearly, due to homogeneous Dirichlet boundary condition, we have*

$$\max\left\{0, \operatorname*{ess\,sup}_{\partial\Omega}\ u\right\} = 0$$

for problem (2.5). Hence, the discrete maximum principle is equivalent to the following conservation of non-positivity:

$$f(x) \leq 0\ \ \textit{for a.e.}\ x \in \Omega \implies u(x) \leq 0\ \ \textit{for a.e.}\ x \in \Omega\,.$$

The following lemma shows the validity of the maximum principle for problem (2.5), see e.g. [30].

Lemma 2.2 (The maximum principle)**.** *Problem (2.5) satisfies the maximum principle.*

Proof.

Define

$$v(x) := \frac{u(x) + |u(x)|}{2} = u(x)^+ \geq 0 \text{ for a.e. } x \in \Omega\,.$$

We can see that

$$v(x) = \begin{cases} 0 & \text{if } u(x) \leq 0 \\ u(x) & \text{if } u(x) > 0 \end{cases} \text{ for a.e. } x \in \Omega\,.$$

Since $u \in V$, we have $v \in V$ as well and we can use it as a test function in (2.2). As $v(x) \geq 0$ for a.e. $x \in \Omega$ and $f(x) \leq 0$ for a.e. $x \in \Omega$ we have $F(v) = \int_\Omega fv \,\mathrm{d}x \leq 0$. Hence,

$$0 \geq F(v) = a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + \kappa^2 uv)\,\mathrm{d}x = \int_{\Omega^+} (\nabla v \cdot \nabla v + \kappa^2 vv)\,\mathrm{d}x \geq 0,$$

where $\Omega^+ = \{x \in \Omega;\ v(x) \neq 0\}$.

Thus

$$\int_\Omega \nabla v \cdot \nabla v \,\mathrm{d}x = 0,$$

which implies

$$\nabla v(x) = 0 \text{ for a.e. } x \in \Omega\,.$$

Therefore $v(x) = \text{const.}$ for a.e. $x \in \Omega$. Further, since $v = 0$ a.e. on $\partial\Omega$ and $v = \text{const.}$ a.e. in Ω we have $v = 0$ a.e. in $\overline{\Omega}$. Consequently $u \leq 0$ a.e. in $\overline{\Omega}$.

□

There exist two other principles equivalent to the maximum principle:

Definition 2.3 (The comparison principle (ComP))**.** *Let $u_1 \in V$ be the weak solution of (2.5) with the right-hand side $f_1 \in \mathcal{L}^2(\Omega)$ and $u_2 \in V$ the weak solution of (2.5) with the right-hand side $f_2 \in \mathcal{L}^2(\Omega)$. We say that problem (2.5) satisfies the comparison principle if*

$$f_1(x) \leq f_2(x) \text{ for a.e. } x \in \Omega \implies u_1(x) \leq u_2(x) \text{ for a.e. } x \in \Omega\,.$$

Definition 2.4 (The minimum principle (MinP))**.** *Let $u \in V$ be the weak solution of (2.5) with $f \in \mathcal{L}^2(\Omega)$. We say that problem (2.5) satisfies the minimum principle if*

$$\operatorname*{ess\,inf}_{\overline{\Omega}} u \geq \min\left\{0, \operatorname*{ess\,inf}_{\partial\Omega} u\right\} \text{ for all } f(x) \geq 0 \text{ for a.e. } x \in \Omega,$$

where the essential minimum of u in $\overline{\Omega}$ is given by

$$\operatorname*{ess\,inf}_{\overline{\Omega}} u := \sup\left\{a \in \mathbb{R} : \mu\left(\left\{x \in \overline{\Omega} : f(x) < a\right\}\right) = 0\right\}.$$

Remark 2.2. *Similarly to the maximum principle, due to homogeneous Dirichlet boundary condition,*

$$\min\left\{0, \operatorname*{ess\,inf}_{\partial\Omega} \; u\right\} = 0$$

for problem (2.5). Hence, the discrete minimum principle is equivalent to the conservation of non-negativity:

$$f(x) \geq 0 \;\; \textit{for a.e. } x \in \Omega \implies u(x) \geq 0 \;\; \textit{for a.e. } x \in \Omega\,.$$

Theorem 2.1. *Assume problem (2.5). The following statements are equivalent:*

1. *The maximum principle holds.*
2. *The comparison principle holds.*
3. *The minimum principle holds.*

Proof.

The equivalence 1. $\Leftrightarrow$ 2. is an easy consequence of the multiplication (2.5) by -1. It remains to prove 1. $\Leftrightarrow$ 3.:

Let the maximum principle be satisfied by problem (2.5). Let $u_1 \in V$ be the weak solution of (2.5) with the right-hand side $f_1 \in \mathcal{L}^2(\Omega)$ and $u_2 \in V$ the weak solution of (2.5) with the right-hand side $f_2 \in \mathcal{L}^2(\Omega)$, $f_1 \leq f_2$ a.e. in Ω. Then it is obvious that $f_1 - f_2 \leq 0$ a.e. in Ω and $u_1 - u_2$ is the weak solution of (2.5) with $f_1 - f_2$. Because the maximum principle holds we have $\operatorname{ess\,sup}_{\overline{\Omega}} \; u_1 - u_2 \leq 0$ so $u_1 \leq u_2$ a.e. in Ω and the comparison principle holds.

Analogically, we can show that if the comparison principle holds then also the maximum principle holds.

□

2.2 Galerkin method

The Galerkin method is based on a replacement of the infinite-dimensional space V by a finite dimensional one. Therefore, we consider a sequence of finite-dimensional subspaces $\{V_h\} \subset V,\ h > 0$. In analogy with problem (2.5) we solve in each V_h the following discrete problem.

Definition 2.5 (Discrete problem)**.** *Find $u_h \in V_h$ such that*

$$a(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h. \tag{2.6}$$

As every V_h is obviously a Hilbert space (because it is a finite-dimensional subspace of the Hilbert space V) and the form $a(\cdot,\cdot)$ restricted to $V_h \times V_h$ remains bilinear, V_h-elliptic and bounded, according to Lax-Milgram lemma there exists a unique solution $u_h \in V_h$ of the discrete problem (2.6) and under suitable assumptions the sequence of these approximate solutions converges to the exact solution of (2.5).

Lemma 2.3 (Céa). *Let V be a Hilbert space, $a(\cdot,\cdot) : V \times V \mapsto \mathbb{R}$ a bounded V-elliptic bilinear form and $F \in V'$. Let $u \in V$ be the solution of problem (2.5). Let V_h be a subspace of V and $u_h \in V_h$ the solution of the discrete problem (2.6). Let C_a, C_b be the continuity and V-ellipticity constants of the form $a(\cdot,\cdot)$. Then*

$$\|u - u_h\|_V \leq \frac{C_a}{C_b} \inf_{v_h \in V_h} \|u - v_h\|_V .$$

Proof.

First proved by Céa [5] in 1964, the proof can be found also e.g. in [32, p. 49].

□

Thus the sufficient condition for convergence to the exact solution of (2.5) is the existence of a system $\{V_h\}$ of subspaces of V such that for every $v \in V$

$$\lim_{h \to 0} \inf_{v_h \in V_h} \|v - v_h\|_V = 0.$$

This condition is equivalent to

$$\overline{\bigcup_{h>0} V_h}^{V} = V.$$

In the following we show that problem (2.6) is equivalent to a system of linear algebraic equations.

Every solution $u_h \in V_h$ can be written as a linear combination of basis functions of the finite-dimensional space V_h :

$$u_h = \sum_{j=1}^{N_h} c_j \varphi_j, \tag{2.7}$$

where N_h is the dimension of V_h, $\{\varphi_j\}_{j=1}^{N_h}$ is a basis of V_h and $c_j \in \mathbb{R}$, $j = 1, \ldots, N_h$, are unknown coefficients. Notice, that the discretization error $u - u_h$ depends only on the subspace V_h and does not depend on the choice of the basis. If we substitute (2.7) into (2.6), we get

$$a\left(\sum_{j=1}^{N_h} c_j \varphi_j, v_h\right) = F(v_h) \quad \forall v_h \in V_h.$$

Hence, thanks to the linearity of the form $a(\cdot,\cdot)$ we obtain

$$\sum_{j=1}^{N_h} c_j a(\varphi_j, v_h) = F(v_h) \quad \forall v_h \in V_h. \tag{2.8}$$

The validity of (2.8) for all test functions $v_h \in V_h$ is equivalent to the validity of (2.8) tested by all basis functions. This fact yields the following equivalent system of linear algebraic equations:

$$\sum_{j=1}^{N_h} c_j a(\varphi_j, \varphi_i) = F(\varphi_i) \quad \forall i = 1, \dots, N_h. \tag{2.9}$$

This motivates the definition of the finite element matrix.

Definition 2.6 (Finite element matrix (FE-matrix) A_h)**.**

$$A_h = \{a_{ij}\}_{i,j=1}^{N_h}, \quad a_{ij} = a(\varphi_j, \varphi_i) = \int_\Omega (\nabla\varphi_j \cdot \nabla\varphi_i + \kappa^2 \varphi_j \varphi_i)\,\mathrm{d}x. \tag{2.10}$$

Then (2.9) can be written in a matrix form

$$A_h c_h = F_h,$$

where $c_h = \{c_j\}_{j=1}^{N_h}$ is the vector of unknown coefficients and $F_h = \{F_j\}_{j=1}^{N_h}$, $F_j = F(\varphi_j)$, is called the load vector.

The FE-matrix A_h consists of two components, the stiffness and the mass matrices:

Definition 2.7 (Stiffness matrix S_h)**.**

$$S_h = \{s_{ij}\}_{i,j=1}^{N_h}, \quad s_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i\,\mathrm{d}x. \tag{2.11}$$

Definition 2.8 (Mass matrix M_h)**.**

$$M_h = \{m_{ij}\}_{i,j=1}^{N_h}, \quad m_{ij} = \int_\Omega \varphi_j \varphi_i\,\mathrm{d}x. \tag{2.12}$$

Clearly, if κ^2 is a constant a.e. in Ω then

$$A_h = S_h + \kappa^2 M_h, \tag{2.13}$$

otherwise

$$A_h \le S_h + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\Omega)} M_h. \tag{2.14}$$

The matrix A_h corresponding to problem (2.6) is symmetric and it can be proved that A_h is positive definite and invertible: We consider an arbitrary vector $0 \neq y = (y_1, y_2, \dots, y_{N_h})^T$. Then the function $v = \sum_{i=1}^{N_h} y_i \varphi_i$ is in V_h and by the V-ellipticity of the form $a(u, v)$ we obtain

$$\begin{aligned} y^T A_h y &= \sum_{i=1}^{N_h}\sum_{j=1}^{N_h} y_i a_{ij} y_j = \sum_{i=1}^{N_h}\sum_{j=1}^{N_h} y_i a(\varphi_j, \varphi_i) y_j = a\left(\sum_{j=1}^{N_h} \varphi_j y_j, \sum_{i=1}^{N_h} \varphi_i y_i\right) \\ &= a(v, v) \ge C_b \|v\|_V^2 > 0. \end{aligned}$$

Thus A_h is positive definite and hence invertible (Otherwise there would exist a vector $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{N_h})^T \neq 0$ such that $A_h\tilde{y} = 0$. But then $\tilde{y}^T A_h \tilde{y} = 0$ which would be in contradiction with the positive definiteness of A_h.). This allows to express the vector of coefficients c_h as

$$c_h = A_h^{-1} F_h. \tag{2.15}$$

2.3 Finite element method

The finite element method can be understood as a special case of the Galerkin method in which we take piecewise-polynomial functions as basis functions for every V_h and in which the FE-matrices A_h are sparse. We consider problem (2.5), where $a(u, v)$ and $F(v)$ are defined by (2.3) and (2.4). The finite-dimensional subspace $V_h \subset V$ is constructed by the following two steps:

1. We cover the domain $\overline{\Omega} \subset \mathbb{R}^d,\ d \geq 1$, by a mesh $\mathcal{T}_h$, which divide $\overline{\Omega}$ into the finite number of subdomains $K_1, K_2, \ldots, K_{D_h}$, called the finite elements. The discretization parameter is chosen as $h = \max\{\mathrm{Diam}\, K_i, i = 1, 2, \ldots, D_h\}$. The finite elements $K_i \in \mathcal{T}_h$ are usually intervals (for $d = 1$), triangles or quadrilaterals (for $d = 2$), tetrahedra, blocks (i.e. cartesian products of intervals), triangular prisms or pyramids (for $d = 3$). The finite element mesh is assumed to met the following conditions:

   - $\overline{\Omega} = \bigcup\limits_{K_i \in \mathcal{T}_h} K_i$
   - every $K_i \in \mathcal{T}_h$ is a closed set whose interior K^0 is not empty
   - $K_i^0 \cap K_j^0 = \emptyset$ holds for every $K_i, K_j \in \mathcal{T}_h,\ K_i \neq K_j$
   - every $K_i \in \mathcal{T}_h$ has a Lipschitz-continuous boundary
   - every side of $K_i \in \mathcal{T}_h$ is either a subset of $\partial\Omega$ or a side of some other element $K_j \in \mathcal{T}_h$

2. Second we create subspace $V_h \subset V$ such that

$$V_h = \left\{ v_h \in V \,\middle|\, v_h \in \mathcal{C}(\overline{\Omega}),\ v_h\big|_{K_i} \in P_{K_i},\ \forall K_i \in \mathcal{T}_h \right\}, \tag{2.16}$$

   where P_K is a suitable finite dimensional space of functions defined on K. For example if K is an interval, triangle or tetrahedron, we choose

$$\begin{aligned} P_K = P^k(K) := \{ v;\ v(x) = \sum_{i=1}^{k} \sum_{j=1}^{d} \alpha_{ij} x_j^i + \alpha_0,\ \alpha_{ij} \in \mathbb{R},\ i = 1, \ldots, k, \\ j = 1, \ldots, d;\ x = (x_1, \ldots, x_d) \in K \subset \mathbb{R}^d \}, \end{aligned} \tag{2.17}$$

in the case K is a quadrilateral or block we set

$$P_K = Q^k(K) := \{ v;\ v(x) = \sum_{l_1=0}^{k} \sum_{l_2=0}^{k} \cdots \sum_{l_d=0}^{k} \alpha_{l_1 l_2 \ldots l_d} x_1^{l_1} x_2^{l_2} \ldots x_d^{l_d};\ \alpha_{l_1 l_2 \ldots l_d} \in \mathbb{R} \\ \forall l_1 l_2 \ldots l_d;\ x = (x_1, \ldots, x_d) \in K \subset \mathbb{R}^d \}. \tag{2.18}$$

From now on, we take into consideration only the case $k = 1$. Thus the space V_h defined by (2.16) and (2.17) is a space of piecewise-linear functions, V_h defined by (2.16) and (2.18) is either a space of piecewise-bilinear functions (for $d = 2$) or a space of piecewise-trilinear functions (for $d = 3$).

Definition 2.9 (Vertex of a mesh)**.** *We say that $B \in \Omega$ is a vertex of the mesh $\mathcal{T}_h$ if B is a vertex of any element $K \in \mathcal{T}_h$.*

Definition 2.10 (Interior vertex)**.** *We say that a vertex $B \in \mathcal{T}_h$ is the interior vertex if $B \notin \partial\Omega$.*

Every $v_h \in V_h$ can be then uniquely defined by the so-called set of degrees of freedom $\sum_h$. These degrees of freedom are linear functionals giving the function values at the interior vertices $B_1,\ B_2, \ldots, B_{N_h}$ of the mesh $\mathcal{T}_h$:

$$\sum\nolimits_h = \{\phi_i \in V_h' : \phi_i(v_h) = v_h(B_i);\ B_i \text{ is an interior vertex},\ i = 1, \ldots, N_h\}.$$

In every vertex $B \in \partial\Omega$, there is $v_h(B) = 0$ because $V_h \subset V$ and all $v \in V$ vanish on $\partial\Omega$.

We can select a basis $\varphi_1, \varphi_2, \ldots, \varphi_{N_h}$ of V_h where every function φ_i has a small support: Let $B_1,\ B_2, \ldots, B_{N_h} \in \Omega$ be the interior vertices of the mesh $\mathcal{T}_h$. We define basis functions $\varphi_i \in V_h$ such that $\varphi_i \geq 0$ and

$$\varphi_i(B_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \qquad i, j = 1, 2, \ldots, N_h. \tag{2.19}$$

Figures 2.1 and 2.2 present examples of such basis functions in one dimensional and two dimensional cases.

2.4 Discrete maximum principle

Definition 2.11 (The discrete maximum principle)**.** *Let $u_h \in V_h$ be the weak solution of problem (2.6). We say that the discrete problem satisfies the discrete maximum principle if*

$$f(x) \leq 0 \ \textit{for a.e. } x \in \Omega \implies \max_{\bar{\Omega}} u_h \leq \max\left\{0, \max_{\partial\Omega} u_h\right\}. \tag{2.20}$$

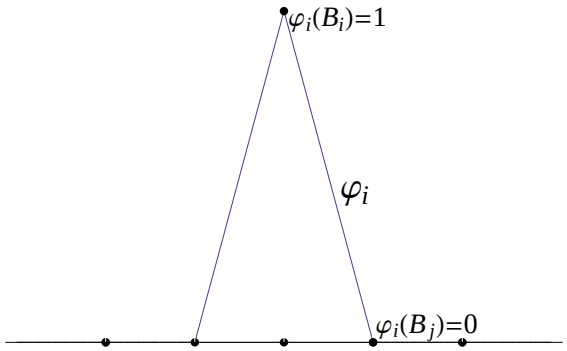


Figure 2.1: Example of basis function for $\Omega \subset \mathbb{R}^1$.

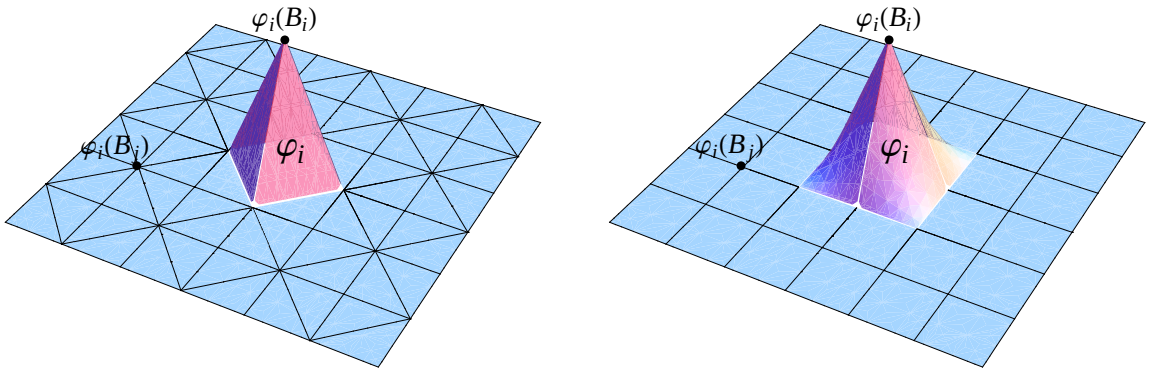


Figure 2.2: Examples of basis functions on meshes formed by triangles and squares for $\Omega \subset \mathbb{R}^2$.

Since in our case $u_h \in V_h \subset H_0^1(\Omega)$ we have $u_h = 0$ on $\partial\Omega$ and we can reformulate (2.20) as

$$f(x) \leq 0 \text{ for a.e. } x \in \Omega \implies \max_{\bar{\Omega}} u_h = 0.$$

We know that u_h can be written in the form (2.7), where $\{\varphi_j\}_{j=1}^{N_h}$ is the basis of V_h defined by (2.19). Hence, $u_h \leq 0$ everywhere in Ω if and only if all the coefficients c_j, $j = 1, \ldots, N_h$, are non-positive.

Definition 2.12. *We say that a matrix $A \in \mathbb{R}^{N\times N}$ is non-negative ($A \geq 0$) if all entries of A are non-negative. Analogically, A is non-positive ($A \leq 0$) if all entries of A are non-positive. The same notation is used also for vectors.*

The following theorem provides the important sufficient and necessary condition for the validity of the discrete maximum principle.

Theorem 2.2. *Let $u_h \in V_h$ be the weak solution of problem (2.6). The discrete maximum principle holds if and only if $A_h^{-1} \geq 0$.*

Proof.

First, let A_h^{-1} be non-negative. It is obvious that $F_h \leq 0$ for an arbitrary $f \leq 0$. Hence we have $c_h = A_h^{-1}F_h \leq 0$ for $f \leq 0$. This implies $u_h \leq 0$ and therefore the discrete maximum principle is satisfied.

The other implication can be proved by contradiction: We suppose that the discrete maximum principle holds and $A_h^{-1} = \{\tilde{a}_{ij}\}_{i,j=1}^{N_h} \ngeq 0$, i.e. there exists $\tilde{a}_{kl} < 0$, $k, l \in \{1, \ldots, N_h\}$. Then if we choose $f \in \mathcal{L}^2(\Omega)$, $f \leq 0$, such that the corresponding $F_h = \{F_j\}_{j=1}^{N_h}$ has entries $F_l = -1$ and $F_i = 0$ for all $i \neq l$, $i \in \{1, \ldots, N_h\}$, we get

$$c_k = \tilde{a}_{kl} > 0.$$

This implies $u_h \nleq 0$ for $f \leq 0$ which is in contradiction with the assumption that the discrete maximum principle holds. A problem with the above reasoning is that the function f yielding the vector F_h with the needed properties would be the negative Dirac function δ_{x_l}, which is not in $\mathcal{L}^2(\Omega)$. However, a sequence of functions $\{f_m\}_{m=1}^{\infty}$ can be constructed , such that $0 \geq f_m \in \mathcal{L}^2(\Omega)$ and

$$f_m \xrightarrow{\mathcal{L}^2(\Omega)} -\delta_{x_l}.$$

For solutions $u_{h,m}$ of problem (2.6) with f_m we obtain

$$u_{h,m} \xrightarrow{V} u_h,$$

where $u_h \nleq 0$ is the solution of (2.6) with $f = -\delta_{x_l}$. Hence, there exists $m_0 \in \mathbb{N}$ such that $0 \geq f_{m_0} \in \mathcal{L}^2(\Omega)$ and $u_{h,m_0} \nleq 0$.

□

Thus, if we want the maximum principle to be satisfied we need the matrix A_h to be monotone, i.e. such that A_h^{-1} is non-negative. The question is: What are the conditions for the matrix A_h to have a non-negative inverse? This question will be the core of the following section.

2.5 M-matrices

In this section we show that sufficient conditions for matrix A_h to have a non-negative inverse is that A_h is symmetric, positive definite and all off-diagonal entries of A_h are non-positive.

Theorem 2.3 (Perron-Frobenius). *Let $A = \{a_{ij}\}_{i,j=1}^{N} \in \mathbb{R}^{N\times N}$ be a matrix such that $A \geq 0$. Then*

1. *the spectral radius $\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A) \text{ is an eigenvalue}\}$ is an eigenvalue of A and*
2. *there exists a non-negative eigenvector $v \in \mathbb{R}^N$ belonging to $\rho(A)$.*

Proof.
See, e.g., [16, p. 102].

□

Before we formulate another theorem we define the following class of matrices:

$$\mathcal{Z}_N = \left\{M \in \mathbb{R}^{N\times N}; m_{ij} \leq 0 \text{ for } i \neq j,\ i,j = 1,\ldots,N\right\}.$$

Theorem 2.4. *Let $A = \{a_{ij}\}_{i,j=1}^{N} \in \mathcal{Z}_N$ be a symmetric, positive definite matrix. Then $A^{-1} \geq 0$.*

Proof.
This is a consequence of [16, Theorem 5.1]. However for the readers convenience, we present a direct proof. We define $k := \max\{a_{ii}, i = 1,2,\ldots,N\}$ and the matrix $C := kI - A$, where $I \in \mathbb{R}^{N\times N}$ is an identity matrix. Then $C \geq 0$. Perron-Frobenius theorem implies that $\rho(C)$ is an eigenvalue of C and that there exists a non-negative eigenvector $w \in \mathbb{R}^N$ such that $Cw = \rho(C)w$. Since $A = kI - C$ we have

$$Aw = (kI - C)w = kw - Cw = kw - \rho(C)w = (k - \rho(C))w$$

and $k - \rho(C)$ is an eigenvalue of A. As A is symmetric and positive definite all eigenvalues of A are positive. Thus $k - \rho(C) > 0$ and

$$k > \rho(C) \geq 0 \quad \Longrightarrow \quad \frac{1}{k}\rho(C) = \rho\left(\frac{1}{k}C\right) < 1.$$

This enables us to express A^{-1} in terms of the Neumann series. Indeed

$$A = kI - C = k\left(I - \frac{1}{k}C\right)$$

and therefore

$$A^{-1} = \frac{1}{k}\left(I - \frac{1}{k}C\right)^{-1} = \frac{1}{k}\left(I + \frac{1}{k}C + \frac{1}{k^2}C^2 + \frac{1}{k^3}C^3 + \dots\right),$$

where we utilized the fact that $\rho\left(\frac{1}{k}C\right) < 1$. Finally, since $C \geq 0$ and $k \geq 0$ we conclude that

$$A^{-1} \geq 0.$$

□

Definition 2.13 (M-matrix)**.**
Nonsingular matrix $A \in \mathcal{Z}_N$ which is monotone (i.e. $A^{-1} \geq 0$) is called M-matrix.

The FE-matrix A_h associated to problem (2.6) is automatically symmetric and positive definite, therefore in order to have $A_h^{-1} \geq 0$ we only need to show that all off-diagonal entries of A_h are non-positive. These off-diagonal entries will be investigated using the fact that the FE-matrix A_h can be expressed as a sum of so-called local matrices.

2.6 Local matrices

The finite element method is characterized by a division $\mathcal{T}_h$ of the domain $\overline{\Omega}$ into subdomains $K \in \mathcal{T}_h$ (called finite elements). This enables us to replace integrals over the domain Ω by the sum of integrals over subdomains K in formulas (2.11), (2.12) and (2.10). Thus

$$s_{ij} = \sum_{K\in\mathcal{T}_h} \int_K \nabla\varphi_j \cdot \nabla\varphi_i \,\mathrm{d}x, \qquad m_{ij} = \sum_{K\in\mathcal{T}_h} \int_K \varphi_j\varphi_i \,\mathrm{d}x$$

and

$$a_{ij} = \sum_{K\in\mathcal{T}_h} \int_K \nabla\varphi_j \cdot \nabla\varphi_i \,\mathrm{d}x + \sum_{K\in\mathcal{T}_h} \int_K \kappa^2\varphi_j\varphi_i \,\mathrm{d}x.$$

This splitting motivates the following definitions of local matrices.

Definition 2.14 (Local stiffness matrix S_h^K)**.**

$$S_h^K = \left\{s_{ij}^K\right\}_{i,j=1}^{N_h}, \quad s_{ij}^K = \int_K \nabla\varphi_j \cdot \nabla\varphi_i \,\mathrm{d}x. \tag{2.21}$$

Definition 2.15 (Local mass matrix M_h^K)**.**

$$M_h^K = \left\{m_{ij}^K\right\}_{i,j=1}^{N_h}, \quad m_{ij}^K = \int_K \varphi_j\varphi_i \,\mathrm{d}x. \tag{2.22}$$

Definition 2.16 (Local FE-matrix A_h^K)**.**

$$A_h^K = \left\{a_{ij}^K\right\}_{i,j=1}^{N_h}, \quad a_{ij}^K = \int_K \nabla\varphi_j \cdot \nabla\varphi_i \,\mathrm{d}x + \int_K \kappa^2 \varphi_j \varphi_i \,\mathrm{d}x. \tag{2.23}$$

Analogously to conditions (2.13) and (2.14) for the matrices S_h, M_h and A_h we obtain the following relations for the local matrices S_h^K, M_h^K and A_h^K. If κ^2 is a constant a.e. in K then

$$A_h^K = S_h^K + \kappa^2 M_h^K, \tag{2.24}$$

otherwise

$$A_h^K \leq S_h^K + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} M_h^K. \tag{2.25}$$

Further,

$$s_{ij} = \sum_{K\in\mathcal{T}_h} s_{ij}^K, \quad m_{ij} = \sum_{K\in\mathcal{T}_h} m_{ij}^K \ \text{ and } \ a_{ij} = \sum_{K\in\mathcal{T}_h} a_{ij}^K.$$

Thus a sufficient (but not necessary) condition for A_h to have non-positive off-diagonal entries is the non-positivity of all off-diagonal entries of all local FE-matrices A_h^K, $K \in \mathcal{T}_h$:

Observation 2.1. *If*

$$a_{ij}^K \leq 0 \ \forall i \neq j, \ i,j = 1,\ldots,N_h, \ \forall K \in \mathcal{T}_h,$$

then

$$a_{ij} \leq 0 \ \forall i \neq j, \ i,j = 1,\ldots,N_h.$$

Furthermore since all basis functions are defined by (2.16) and (2.19) we can see that

$$\varphi_i\,|_K \equiv 0 \text{ for all } i \text{ such that } B_i \text{ is not a vertex of } K.$$

Therefore,

$$s_{jk}^K = m_{jk}^K = a_{jk}^K = 0 \text{ for all } j,k \text{ such that } B_j \text{ or } B_k \text{ is not a vertex of } K.$$

Hence, there are only few non-zero entries in every local matrix M_h^K, S_h^K and A_h^K and we can define condensed matrices $\bar{M}_h^K = \left\{\bar{m}_{ij}^K\right\}_{i,j=1}^{l}$, $\bar{S}_h^K = \left\{\bar{s}_{ij}^K\right\}_{i,j=1}^{l}$ and $\bar{A}_h^K = \left\{\bar{a}_{ij}^K\right\}_{i,j=1}^{l}$, where l is the number of vertices of the element K (i.e. $l = 2$ in case that elements are intervals, $l = 3$ for triangles, $l = 4$ for rectangles and tetrahedra, $l = 8$ for blocks and $l = 6$ for triangular prisms). These condensed matrices are such that contain all non-zero entries of corresponding local matrices and we can for simplicity operate with $\bar{M}_h^K$, $\bar{S}_h^K$ and $\bar{A}_h^K$ instead of M_h^K, A_h^K and S_h^K.

Observation 2.2. *If*

$$\bar{a}_{ij}^K \leq 0 \ \forall i \neq j, \ i,j = 1,\ldots,l, \ \forall K \in \mathcal{T}_h,$$

then

$$a_{ij} \leq 0 \ \forall i \neq j, \ i,j = 1,\ldots,N_h.$$

Finally we come to the following conclusion:

Theorem 2.5. *The discrete maximum principle holds if*

$$\bar{a}_{ij}^K \leq 0 \ \forall i < j, \ i,j = 1,\dots,l, \ \forall K \in \mathcal{T}_h.$$

Proof.

It follows from Theorem 2.2, Theorem 2.4 and Observation 2.2 and from the symmetry of the matrix $\bar{A}_h^K$.

□

Chapter 3

Finite elements in 1D – intervals

In this chapter we show the results of the validity of the discrete maximum principle for the meshes consisting of intervals. We consider problem (2.6) in a domain $\Omega = (\alpha, \beta) \in \mathbb{R}^1$ with a mesh $\mathcal{T}_h$, which divide $\overline{\Omega}$ into the finite number of intervals $K_i = [B_{i-1}, B_i]$, $i = 1, 2, \ldots, N_h + 1$, $-\infty < \alpha = B_0 < B_1 < B_2 < \cdots < B_{N_h+1} = \beta < \infty$. Let $h_{K_i} = B_i - B_{i-1}$ be the lengths of the intervals $K_i \in \mathcal{T}_h$ and $h = \max\{h_{K_i}, K_i \in \mathcal{T}_h\}$. The space V_h is defined by (2.16), where $P_K = P^1(K)$, thus V_h is a space of continuous and piecewise-linear functions. The basis of V_h are functions $\varphi_1, \varphi_2, \ldots, \varphi_{N_h} \in V_h$ defined by (2.19). Thus support of every φ_i is formed by two adjacent intervals that contain the vertex B_i (see Figure 2.1). It is $\varphi_i = 0$ elsewhere in Ω. Therefore

$$a_{ij} = \int_\Omega (\nabla\varphi_j \cdot \nabla\varphi_i + \kappa^2 \varphi_j \varphi_i)\,\mathrm{d}x = 0 \text{ whenever } |i-j| \geq 2$$

and A_h is a tridiagonal matrix.

We are dealing with the question what are the sufficient conditions for the validity of the discrete maximum principle of problem (2.6). Theorem 2.5 provides the following condition for the discrete maximum principle: the local condensed FE-matrices $\bar{A}_h^{K_i}$ of problem (2.6) should be such that

$$\bar{a}_{1,2}^{K_i} \leq 0 \quad \forall K_i \in \mathcal{T}_h. \tag{3.1}$$

If we take into consideration that

$$\varphi_j\,|_{K_i} = 0 \text{ for all } j \neq i, i-1,$$

then the condensed matrices are defined as follows:

$$\bar{S}_h^{K_i} = \begin{pmatrix} s_{i-1,i-1}^{K_i} & s_{i-1,i}^{K_i} \\ s_{i,i-1}^{K_i} & s_{i,i}^{K_i} \end{pmatrix}, \quad \bar{M}_h^{K_i} = \begin{pmatrix} m_{i-1,i-1}^{K_i} & m_{i-1,i}^{K_i} \\ m_{i,i-1}^{K_i} & m_{i,i}^{K_i} \end{pmatrix}$$

and

$$\bar{A}_h^{K_i} = \begin{pmatrix} a_{i-1,i-1}^{K_i} & a_{i-1,i}^{K_i} \\ a_{i,i-1}^{K_i} & a_{i,i}^{K_i} \end{pmatrix},$$

where $s_{ij}^{K_i}$, $m_{ij}^{K_i}$ and $a_{ij}^{K_i}$ are given by (2.21)–(2.23).

3.1 Reference element and reference mapping

In this section we define the reference element $\hat{K}$ which enable us to compute the entries of the matrices $\bar{M}_h^{K_i}$ and $\bar{S}_h^{K_i}$ in a simpler way. We choose the reference element

$$\hat{K} = [-1, 1].$$

The reference element can be transformed into an arbitrary $K_i \in \mathcal{T}_h$ by the reference mapping χ_{K_i}.

Definition 3.1 (The reference mapping)**.** *Mapping χ_{K_i}: $\hat{K} \mapsto K_i$, $K_i \in \mathcal{T}_h$, such that*

$$\chi_{K_i}(\xi) = x = B_{i-1} + \frac{1+\xi}{2}(B_i - B_{i-1}) = B_{i-1} + \frac{1+\xi}{2}h_{K_i}, \quad \xi \in [-1, 1], \tag{3.2}$$

is called the reference mapping.

Definition 3.2 (Jacobian matrix of $f : \mathbb{R}^n \mapsto \mathbb{R}^n$)**.**

$$J(f) := \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_n} \end{pmatrix}. \tag{3.3}$$

Definition 3.3 (Jacobian of $f : \mathbb{R}^n \mapsto \mathbb{R}^n$)**.**

$$I(f) := \det J(f). \tag{3.4}$$

From (3.2), (3.3) and (3.4) we obtain

$$I(\chi_{K_i}) = J(\chi_{K_i}) = \frac{\partial \chi_{K_i}(\xi)}{\partial \xi} = \frac{h_{K_i}}{2}.$$

In $\hat{K}$ we define shape functions ℓ_0, ℓ_1 (see Figure 3.1) such that:

$$\ell_0(\xi) = \frac{1-\xi}{2}, \quad \ell_1(\xi) = \frac{1+\xi}{2}, \quad \xi \in [-1, 1].$$

The basis functions $\varphi_{i-1}(x)$ and $\varphi_i(x)$ restricted to the element $K_i \in \mathcal{T}_h$ (see Figure 3.1) are such that

$$\varphi_{i-1}|_{K_i}(x) = \frac{B_i - x}{h_{K_i}}, \quad \varphi_i|_{K_i}(x) = \frac{x - B_{i-1}}{h_{K_i}}, \quad x \in [B_{i-1}, B_i].$$

It can be shown that

$$\begin{aligned} \varphi_{i-1}|_{K_i}(\chi_{K_i}(\xi)) &= \frac{B_i - \chi_{K_i}(\xi)}{h_{K_i}} = \frac{B_i - B_{i-1} - \frac{1+\xi}{2}h_{K_i}}{h_{K_i}} = \frac{h_{K_i} - \frac{1+\xi}{2}h_{K_i}}{h_{K_i}} \\ &= \frac{1-\xi}{2} = \ell_0(\xi) \end{aligned} \tag{3.5}$$

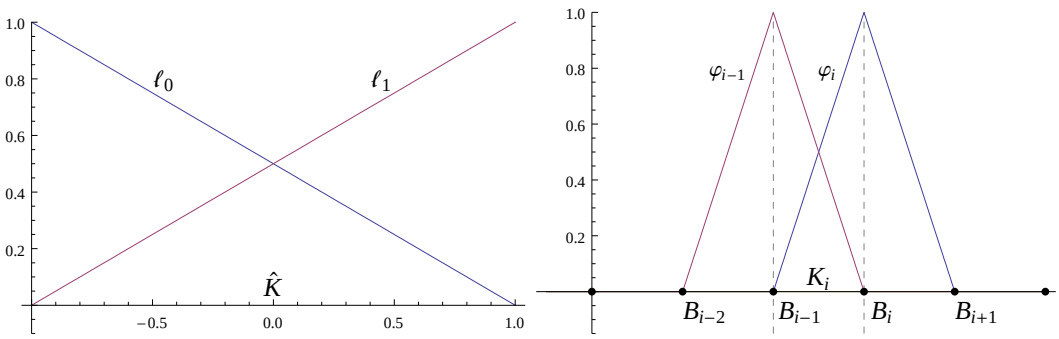


Figure 3.1: Shape functions ℓ_0, ℓ_1 in $\hat{K}$ and basis functions $\varphi_{i-1}(x)$, $\varphi_i(x)$.

and analogously

$$\varphi_i\,|_{K_i}(\chi_{K_i}(\xi)) = \frac{\chi_{K_i}(\xi) - B_{i-1}}{h_{K_i}} = \frac{1+\xi}{2} = \ell_1\,(\xi)\,,\ \text{for } \xi \in [-1,1], \tag{3.6}$$

for all $i = 1, 2, \ldots, N_h$. Then also

$$\frac{\partial \varphi_{i-1}\,|_{K_i}(x)}{\partial x} = \frac{\partial \ell_0(\chi_{K_i}^{-1}(x))}{\partial x} = \frac{\partial \ell_0(\chi_{K_i}^{-1}(x))}{\partial \xi}\frac{\partial \chi_{K_i}^{-1}(x)}{\partial x} = \frac{\partial \ell_0\,(\xi)}{\partial \xi} I(\chi_{K_i})^{-1} \tag{3.7}$$

and

$$\frac{\partial \varphi_i\,|_{K_i}(x)}{\partial x} = \frac{\partial \ell_1(\chi_{K_i}^{-1}(x))}{\partial x} = \frac{\partial \ell_1(\chi_{K_i}^{-1}(x))}{\partial \xi}\frac{\partial \chi_{K_i}^{-1}(x)}{\partial x} = \frac{\partial \ell_1\,(\xi)}{\partial \xi} I(\chi_{K_i})^{-1}, \tag{3.8}$$

for all $i = 1, 2, \ldots, N_h$.

3.2 Local mass and stiffness matrices

Using the theorem of substitution and relations (3.5), (3.6), (3.7), (3.8) we can transform the integral over K_i into the integral over $\hat{K}$ in the calculation of entries of local mass and stiffness matrices of every element K_i. Then

$$\begin{aligned}
\bar{m}_{jk}^{K_i} &= \int_{K_i} \varphi_{k+i-2}\,(x)\,\varphi_{j+i-2}\,(x)\,\mathrm{d}x \\
&= \int_{\hat{K}} \varphi_{k+i-2}\,(\chi_{K_i}(\xi))\,\varphi_{j+i-2}\,(\chi_{K_i}(\xi))\,I(\chi_{K_i})\,\mathrm{d}\xi \\
&= \int_{\hat{K}} \ell_{k-1}\,(\xi)\,\ell_{j-1}\,(\xi)\,\frac{h_{K_i}}{2}\,\mathrm{d}\xi = \frac{h_{K_i}}{2}\bar{m}_{jk}^{\hat{K}}, \quad j,k = 1,2,
\end{aligned}$$

and

$$\begin{aligned}\bar{s}_{jk}^{K_i} &= \int_{K_i} \frac{\partial \varphi_{k+i-2}(x)}{\partial x} \frac{\partial \varphi_{j+i-2}(x)}{\partial x} \,\mathrm{d}x \\ &= \int_{\hat{K}} \left(\frac{\partial \ell_{k-1}(\xi)}{\partial \xi} I(\chi_{K_i})^{-1} \right) \left(\frac{\partial \ell_{j-1}(\xi)}{\partial \xi} I(\chi_{K_i})^{-1} \right) I(\chi_{K_i}) \,\mathrm{d}\xi \\ &= \int_{\hat{K}} \frac{\partial \ell_{k-1}(\xi)}{\partial \xi} \frac{\partial \ell_{j-1}(\xi)}{\partial \xi} I(\chi_{K_i})^{-1} \,\mathrm{d}\xi \\ &= \int_{\hat{K}} \frac{\partial \ell_{k-1}(\xi)}{\partial \xi} \frac{\partial \ell_{j-1}(\xi)}{\partial \xi} \frac{2}{h_{K_i}} \,\mathrm{d}\xi = \frac{2}{h_{K_i}} \bar{s}_{jk}^{\hat{K}}, \quad j,k = 1,2\ .\end{aligned}$$

The entries

$$\bar{m}_{jk}^{\hat{K}} = \int_{-1}^{1} \ell_{k-1}(\xi)\, \ell_{j-1}(\xi) \,\mathrm{d}\xi$$

and

$$\bar{s}_{jk}^{\hat{K}} = \int_{-1}^{1} \frac{\partial \ell_{k-1}(\xi)}{\partial \xi} \frac{\partial \ell_{j-1}(\xi)}{\partial \xi} \,\mathrm{d}\xi, \quad j,k = 1,2,$$

of matrices $\bar{M}_h^{\hat{K}}$ and $\bar{S}_h^{\hat{K}}$ can be computed by a straightforward computation. We obtain:

$$\bar{S}_h^{\hat{K}} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \ \bar{M}^{\hat{K}} = \frac{1}{3} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \tag{3.9}$$

and finally we have

$$\bar{S}_h^{K_i} = \frac{2}{h_{K_i}} \bar{S}^{\hat{K}} = \frac{1}{h_{K_i}} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \ \bar{M}_h^{K_i} = \frac{h_{K_i}}{2} \bar{M}^{\hat{K}} = \frac{h_{K_i}}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

3.3 Discrete maximum principle

The following theorem summarizes the main result of this chapter.

Theorem 3.1. *Let us consider problem (2.6) in a domain $\Omega = (\alpha, \beta) \in \mathbb{R}^1$ with a mesh $\mathcal{T}_h$ consisting of intervals K_i. Let h_{K_i} be the lengths of the intervals $K_i \in \mathcal{T}_h$ and $h = \max\{h_{K_i}, K_i \in \mathcal{T}_h\}$. The space V_h is defined by (2.16), where $P_K = P^1(K)$. If $\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K_i})} h_{K_i}^2 \leq 6$ for all $K_i \in \mathcal{T}_h$, then the discrete maximum principle holds.*

Proof.

Since

$$\bar{a}_{1,2}^{K_i} = \bar{a}_{2,1}^{K_i} \leq \bar{s}_{1,2}^{K_i} + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K_i})} \bar{m}_{1,2}^{K_i} = \frac{-1}{h_{K_i}} + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K_i})} \frac{h_{K_i}}{6},$$

the sufficient condition (3.1) translates into

$$\frac{-1}{h_{K_i}} + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K_i})} \frac{h_{K_i}}{6} = \frac{-6 + \|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K_i})} h_{K_i}^2}{6 h_{K_i}} \leq 0,$$

which is satisfied if and only if

$$\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K_i})} h_{K_i}^2 \leq 6.$$

□

Remark 3.1. *Notice that in the case of the pure diffusion problem ($\kappa^2 = 0$ a.e. in Ω) the crucial condition $\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K_i})} h_{K_i}^2 \leq 6$ from Theorem 3.1 is trivially satisfied for all $K_i \in \mathcal{T}_h$ and the discrete maximum principle is valid on arbitrary meshes.*

Chapter 4

Finite elements in 2D – triangles

In this chapter we consider problem (2.6) in a polygonal domain $\Omega \subset \mathbb{R}^2$ with a mesh $\mathcal{T}_h$ consisting of a finite number of triangles. Let $K \in \mathcal{T}_h$ be a triangle given by vertices B_1, B_2 and B_3 and let α_1, α_2 and α_3 be the angles corresponding to the vertices B_1, B_2 and B_3, respectively. The space V_h is defined by (2.16), where $P_K = P^1(K)$, thus V_h is a space of continuous and piecewise-linear functions. The basis of V_h is formed by functions $\varphi_1, \varphi_2, \ldots, \varphi_{N_h} \in V_h$ defined by (2.19).

As in the previous chapter we are going to formulate sufficient conditions for the validity of the discrete maximum principle of problem (2.6). Theorem 2.5 gives the following condition for the local condensed FE-matrices A_h^K of problem (2.6):

$$\bar{a}_{jk}^K \leq 0 \ \forall j < k, \ j,k = 1,\ldots,3, \ \forall K \in \mathcal{T}_h. \tag{4.1}$$

The results for the triangular finite element meshes introduced in this chapter are well known for the pure diffusion case ($\kappa^2 = 0$ a.e. in Ω). The reaction-diffusion problem was analyzed quite recently in [3].

4.1 Reference element and reference mapping

For triangles we choose as the reference element $\hat{K}$ the right-angle triangle with vertices $v_1 = (-1,-1)$, $v_2 = (1,-1)$ and $v_3 = (-1,1)$ (see Figure 4.1). In $\hat{K}$ we then define shape functions ℓ_1, ℓ_2, ℓ_3 such that:

$$\ell_1(\xi) = \frac{-\xi_1 - \xi_2}{2}, \ \ell_2(\xi) = \frac{1+\xi_1}{2}, \ \ell_3(\xi) = \frac{1+\xi_2}{2}, \quad \xi = (\xi_1,\xi_2) \in \hat{K}.$$

For this case there is a reference mapping $\chi_K : \hat{K} \mapsto K$, $K \in \mathcal{T}_h$, such that

$$\chi_K(\xi) = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^3 x_k \ell_k(\xi) \\ \sum_{k=1}^3 y_k \ell_k(\xi) \end{pmatrix}, \quad \xi = (\xi_1,\xi_2) \in \hat{K}, \quad (x,y) \in K,$$

where (x_k, y_k) are coordinates of vertices B_k of the element K, $k = 1,\ldots,3$. For each element we can therefore consider a coordinate system such that vertex B_1

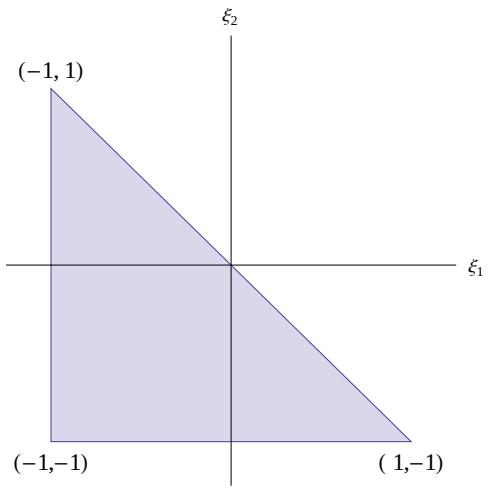


Figure 4.1: Reference element for triangles.

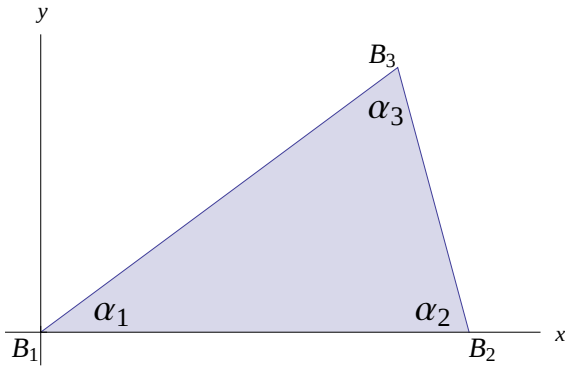


Figure 4.2: Element K.

lies at the origin and vertex B_2 on the x-axis, i.e. $B_1 = (0,0)$, $B_2 = (x_2, 0)$ and $B_3 = (x_3, y_3)$, with $x_2 \neq 0$ and $y_3 \neq 0$ (see Figure 4.2). Then the reference mapping can be written as follows:

$$\chi_K(\xi) = \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2} \begin{pmatrix} x_2(1+\xi_1) + x_3(1+\xi_2) \\ y_3(1+\xi_2) \end{pmatrix}, \quad \xi = (\xi_1, \xi_2) \in \hat{K}, \quad (x, y) \in K.$$

From (3.3) and (3.4) we obtain

$$J(\chi_K) = \frac{1}{2} \begin{pmatrix} x_2 & x_3 \\ 0 & y_3 \end{pmatrix}$$

and

$$I(\chi_K) = \frac{1}{4} x_2 y_3 = \frac{1}{2} |K|,$$

where $|K| = \dfrac{x_2 y_3}{2}$ is the area of the element K.

4.2 Local mass and stiffness matrices

Using the theorem of substitution we can transform the integral over K into the integral over $\hat{K}$ in the calculation of entries of local mass and stiffness matrices of every element K. We obtain the entries

$$\bar{m}_{jk}^K = I(\chi_K) \int_{-1}^{1} \int_{-1}^{-\xi_2} \ell_k(\xi)\, \ell_j(\xi)\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2$$

and

$$\bar{s}_{jk}^K = I(\chi_K) \int_{-1}^{1} \int_{-1}^{-\xi_2} \big(J(\chi_K)^{-1} \nabla \ell_k(\xi)\big) \cdot \big(J(\chi_K)^{-1} \nabla \ell_j(\xi)\big)\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2, \quad j,k = 1,\ldots,3,$$

of matrices $\bar{M}_h^K$ and $\bar{S}_h^K$ which can be computed by a straightforward computation. We get

$$\bar{M}_h^K = \frac{|K|}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \tag{4.2}$$

and

$$\bar{S}_h^K = \frac{1}{2} \begin{pmatrix} \dfrac{x_2}{y_3} + \dfrac{2x_3}{y_3} + \dfrac{x_3^2}{x_2 y_3} + \dfrac{y_3}{x_2} & \dfrac{x_3}{y_3} - \dfrac{x_3^2}{x_2 y_3} - \dfrac{y_3}{x_2} & -\dfrac{x_2 - x_3}{y_3} \\ \dfrac{x_3}{y_3} - \dfrac{x_3^2}{x_2 y_3} - \dfrac{y_3}{x_2} & \dfrac{x_3^2}{x_2 y_3} + \dfrac{y_3}{x_2} & -\dfrac{x_3}{y_3} \\ -\dfrac{x_2 - x_3}{y_3} & -\dfrac{x_3}{y_3} & \dfrac{x_2}{y_3} \end{pmatrix}.$$

Since

$$\frac{x_3}{y_3} = \cot\alpha_1,$$

$$\frac{x_2 - x_3}{y_3} = \cot\alpha_2,$$

$$\frac{y_3}{x_2} = \frac{1}{x_3/y_3 + (x_2 - x_3)/y_3} = \frac{1}{\cot\alpha_1 + \cot\alpha_2},$$

$$\frac{x_3}{x_2} = \frac{y_3}{x_2}\cot\alpha_1 = \frac{\cot\alpha_1}{\cot\alpha_1 + \cot\alpha_2}$$

and

$$\frac{x_3}{y_3} - \frac{x_3^2}{x_2 y_3} - \frac{y_3}{x_2} = \cot\alpha_1 - \cot\alpha_1 \frac{x_3}{x_2} - \frac{y_3}{x_2} = -\frac{1 - \cot\alpha_1 \cot\alpha_2}{\cot\alpha_1 + \cot\alpha_2} = \cot\alpha_3,$$

we have matrix $\bar{S}_h^K$ in the form

$$\bar{S}_h^K = \frac{1}{2} \begin{pmatrix} \cot\alpha_2 + \cot\alpha_3 & -\cot\alpha_3 & -\cot\alpha_2 \\ -\cot\alpha_3 & \cot\alpha_1 + \cot\alpha_3 & -\cot\alpha_1 \\ -\cot\alpha_2 & -\cot\alpha_1 & \cot\alpha_1 + \cot\alpha_2 \end{pmatrix}. \tag{4.3}$$

Both expressions for the local mass and stiffness matrices are well known, see e.g. [2] or [10].

4.3 Discrete maximum principle

Condition (4.1), formula (2.25) and computed values of entries of $\bar{S}_h^K$ and $\bar{M}_h^K$ imply the following conditions for the discrete maximum principle:

$$-\cot\alpha_i + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{|K|}{6} \leq 0, \quad i = 1,2,3.$$

From this we obtain

$$\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{|K|}{6} \leq \cot\alpha_i, \quad i = 1,2,3. \tag{4.4}$$

As $\cot\alpha_i \geq 0$ for $0 < \alpha_i \leq \pi/2$, all angles in the triangle have to be at most $\pi/2$. Further cotangent is a decreasing function on $(0, \pi/2]$. Therefore if

$$0 < \alpha_1 \leq \alpha_2 \leq \alpha_3 \leq \frac{\pi}{2}$$

then

$$\cot\alpha_1 \geq \cot\alpha_2 \geq \cot\alpha_3$$

and we can state the following theorem:

Theorem 4.1. *Let $\mathcal{T}_h$ be a partition of $\overline{\Omega}$ into the finite number of triangles. Let $\alpha_1,\ \alpha_2,\ \alpha_3 > 0$ be the angles of a triangle $K \in \mathcal{T}_h$ and $\alpha_{max} = \max\{\alpha_1, \alpha_2, \alpha_3\}$. Then problem (2.6) satisfies the discrete maximum if*

$$\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{|K|}{6} \leq \cot\alpha_{max} \ \textit{ for all } \ K \in \mathcal{T}_h. \tag{4.5}$$

Proof.

It follows from the above analysis. If (4.4) is met for the greatest angle of every $K \in \mathcal{T}_h$, then

$$\bar{a}_{jk}^K \leq 0 \ \forall j \neq k,\ j,k = 1,\ldots,3$$

and the discrete maximum principle holds by Theorem 2.5.

□

Remark 4.1. *As α_{max} is the greatest angle in a triangle it must be $\alpha_{max} \geq \frac{\pi}{3}$. If condition (4.5) is satisfied then*

$$\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{|K|}{6} \leq \cot\alpha_{max} \leq \cot\frac{\pi}{3} = \frac{1}{\sqrt{3}}$$

and therefore

$$|K| \leq \frac{2\sqrt{3}}{\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}} \ \textit{ for all } \ K \in \mathcal{T}_h. \tag{4.6}$$

Thus, the bigger the reaction coefficient κ^2 is, the smaller the elements must be in order to guarantee the discrete maximum principle by Theorem 4.1.

Remark 4.2. *If $\kappa^2 = 0$ a.e. in Ω then problem (2.6) satisfies the discrete maximum if*

$$\alpha_{max} \leq \frac{\pi}{2} \text{ for all } K \in \mathcal{T}_h.$$

Hence, in the case $\kappa^2 \neq 0$ it is necessary to consider all angles in $\mathcal{T}_h$ to be acute in order to satisfy (4.5). In addition the mesh $\mathcal{T}_h$ must be sufficiently fine, such that the limitation (4.6) is met, see Remark 4.1. For $\kappa^2 = 0$ a.e. in Ω the non-obtuseness of all angles in the triangulation $\mathcal{T}_h$ is adequate for the validity of the discrete maximum principle.

Chapter 5

Finite elements in 2D – rectangles

In this chapter we deal with the discrete maximum principle for rectangular meshes and bilinear finite element approximation of problem (2.6). It was studied e.g. in [6].

Again, we consider problem (2.6) in a polygonal domain $\Omega \subset \mathbb{R}^2$, but now the corresponding mesh $\mathcal{T}_h$ consists of rectangles. Let h_1^K and h_2^K be the lengths of the two sides of an element $K \in \mathcal{T}_h$ and $h = \max\left\{h_j^K, K \in \mathcal{T}_h,\ j = 1, 2\right\}$. The space V_h is defined by (2.16), where $P_K = Q^1(K)$, thus V_h is a space of continuous and piecewise-bilinear functions. The functions $\varphi_1, \varphi_2, \ldots, \varphi_{N_h} \in V_h$ defined by (2.19) form a basis of V_h.

As in both previous chapters we are going to find conditions for validity of the discrete maximum principle for problem (2.6). Theorem 2.5 gives us the following condition for the local condensed FE-matrices A_h^K of problem (2.6):

$$\bar{a}_{jk}^K \leq 0 \ \forall j < k,\ j, k = 1, \ldots, 4,\ \forall K \in \mathcal{T}_h. \tag{5.1}$$

In the sequel we analyze this condition in the particular case of rectangular elements.

5.1 Reference element and reference mapping

For rectangles we choose the following reference element (see Figure 5.1):

$$\hat{K} = [-1, 1]^2.$$

In $\hat{K}$ we then define shape functions ℓ_1, ℓ_2, ℓ_3, ℓ_4 (see Figure 5.2) such that:

$$\begin{aligned}
\ell_1(\xi) &= \frac{(1-\xi_1)(1-\xi_2)}{4}, \quad \ell_2(\xi) = \frac{(1+\xi_1)(1-\xi_2)}{4}, \\
\ell_3(\xi) &= \frac{(1-\xi_1)(1+\xi_2)}{4}, \quad \ell_4(\xi) = \frac{(1+\xi_1)(1+\xi_2)}{4}, \quad \xi = (\xi_1, \xi_2) \in \hat{K},
\end{aligned}$$

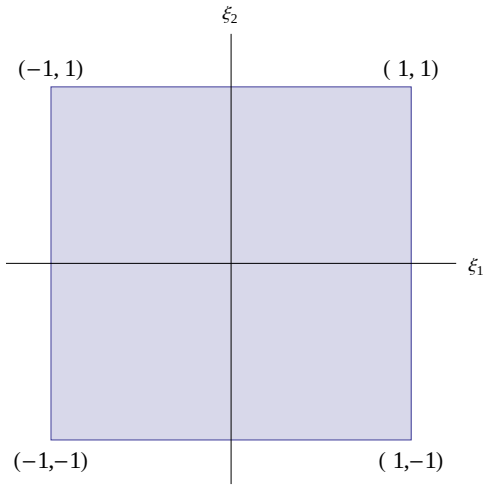


Figure 5.1: Reference element for rectangles.

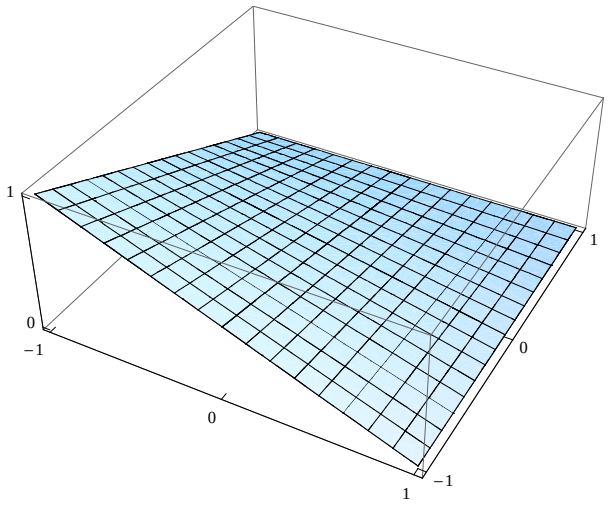


Figure 5.2: Shape function ℓ_1.

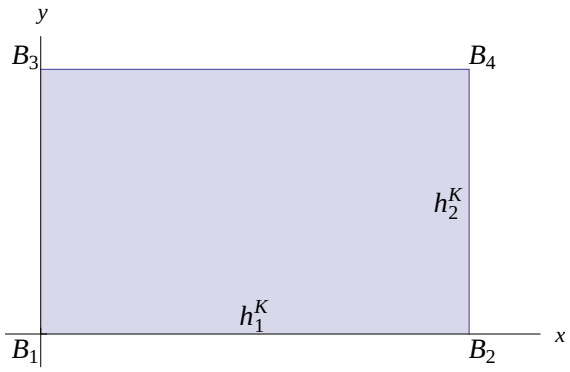


Figure 5.3: Element $K \in \mathcal{T}_h$.

For this case there is a reference mapping $\chi_K : \hat{K} \mapsto K$, $K \in \mathcal{T}_h$, given by

$$\chi_K(\xi) = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^4 x_k \ell_k(\xi) \\ \sum_{k=1}^4 y_k \ell_k(\xi) \end{pmatrix}, \quad \xi = (\xi_1, \xi_2) \in \hat{K}, \quad (x, y) \in K,$$

where (x_k, y_k) are coordinates of vertices B_k of the element K, $k = 1, \ldots, 4$. Since we apply this mapping for calculation of the integrals over elements, we are interested only in lengths h_1^K and h_2^K and not in the particular location of an element in Ω. Therefore, as in the previous chapter, we can for each element consider a coordinate system such that one vertex lies at the origin and one has coordinates (h_1^K, h_2^K) (see Figure 5.3). Then the reference mapping can be written as follows:

$$\chi_K(\xi) = \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2} \begin{pmatrix} h_1^K(1 + \xi_1) \\ h_2^K(1 + \xi_2) \end{pmatrix}, \quad \xi = (\xi_1, \xi_2) \in \hat{K}, \quad (x, y) \in K.$$

The Jacobian matrix of this mapping is

$$J(\chi_K) = \frac{1}{2} \begin{pmatrix} h_1^K & 0 \\ 0 & h_2^K \end{pmatrix}$$

and the Jacobian is

$$I(\chi_K) = \frac{1}{4} h_1^K h_2^K = \frac{|K|}{4},$$

where $|K|$ is the area of the rectangle K.

The basis functions $\varphi_1^K(x, y)$, $\varphi_2^K(x, y)$, $\varphi_3^K(x, y)$ and $\varphi_4^K(x, y)$ restricted to the element $K \in \mathcal{T}_h$ are such that

$$\begin{aligned} \varphi_1^K|_K(x, y) &= \frac{(h_1^K - x)(h_2^K - y)}{h_1^K h_2^K}, & \varphi_2^K|_K(x, y) &= \frac{x(h_2^K - y)}{h_1^K h_2^K}, \\ \varphi_3^K|_K(x, y) &= \frac{(h_1^K - x)y}{h_1^K h_2^K}, & \varphi_4^K|_K(x, y) &= \frac{xy}{h_1^K h_2^K}, \quad (x, y) \in K. \end{aligned} \tag{5.2}$$

5.2 Local mass and stiffness matrices

Using the theorem of substitution we can, similarly to the previous chapter, transform the integral over K into the integral over $\hat{K}$ in the calculation of entries of the local mass and stiffness matrices of every element K. Hence,

$$\begin{aligned}\bar{m}_{jk}^K &= \int_K \varphi_k^K(x)\,\varphi_j^K(x)\,\mathrm{d}x = \int_{\hat{K}} \varphi_k^K(\chi_K(\xi))\,\varphi_j^K(\chi_K(\xi))\,I(\chi_K)\,\mathrm{d}\xi \\ &= \int_{\hat{K}} \ell_k(\xi)\,\ell_j(\xi)\,I(\chi_K)\,\mathrm{d}\xi, \quad j,k=1,\ldots,4,\end{aligned}$$

and

$$\begin{aligned}\bar{s}_{jk}^K &= \int_K \nabla\varphi_k^K(x)\cdot\nabla\varphi_j^K(x)\,\mathrm{d}x \\ &= \int_{\hat{K}} \left(J(\chi_K)^{-1}\nabla\ell_k(\xi)\right)\cdot\left(J(\chi_K)^{-1}\nabla\ell_j(\xi)\right)I(\chi_K)\,\mathrm{d}\xi, \quad j,k=1,\ldots,4\ .\end{aligned}$$

Entries

$$\bar{m}_{jk}^K = \int_{-1}^1\int_{-1}^1 \ell_k(\xi)\,\ell_j(\xi)\,I(\chi_K)\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2$$

and

$$\bar{s}_{jk}^K = \int_{-1}^1\int_{-1}^1 \left(J(\chi_K)^{-1}\nabla\ell_k(\xi)\right)\cdot\left(J(\chi_K)^{-1}\nabla\ell_j(\xi)\right)I(\chi_K)\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2, \quad j,k=1,\ldots,4,$$

of matrices $\bar{M}_h^K$ and $\bar{S}_h^K$ can be then computed by a straightforward computation. We obtain:

$$\bar{S}_h^K = \frac{1}{6h_1^Kh_2^K}\begin{pmatrix} a & b & c & d\\ b & a & d & c\\ c & d & a & b\\ d & c & b & a\end{pmatrix}, \tag{5.3}$$

where

$$\begin{aligned}a &= 2(h_1^K)^2+2(h_2^K)^2, \quad b=(h_1^K)^2-2(h_2^K)^2,\\ c &= -(h_1^K)^2-(h_2^K)^2, \quad d=-2(h_1^K)^2+(h_2^K)^2,\end{aligned}$$

and

$$\bar{M}_h^K = \frac{h_1^Kh_2^K}{36}\begin{pmatrix} 4 & 2 & 1 & 2\\ 2 & 4 & 2 & 1\\ 1 & 2 & 4 & 2\\ 2 & 1 & 2 & 4\end{pmatrix}. \tag{5.4}$$

5.3 Discrete maximum principle

From (5.1), formula (2.25) and from computed values of entries of $\bar{S}_h^K$ and $\bar{M}_h^K$ we obtain the following three conditions for the discrete maximum principle:

$$\frac{b}{6h_1^K h_2^K} + 2\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{h_1^K h_2^K}{36} = \frac{(h_1^K)^2 - 2(h_2^K)^2}{6h_1^K h_2^K} + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{h_1^K h_2^K}{18} \leq 0, \quad (5.5)$$

$$\frac{c}{6h_1^K h_2^K} + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{h_1^K h_2^K}{36} = \frac{-(h_1^K)^2 - (h_2^K)^2}{6h_1^K h_2^K} + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{h_1^K h_2^K}{36} \leq 0, \quad (5.6)$$

$$\frac{d}{6h_1^K h_2^K} + 2\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{h_1^K h_2^K}{36} = \frac{-2(h_1^K)^2 + (h_2^K)^2}{6h_1^K h_2^K} + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{h_1^K h_2^K}{18} \leq 0. \quad (5.7)$$

After rearranging we have from (5.5)

$$3(h_1^K)^2 - 6(h_2^K)^2 + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2(h_2^K)^2 \leq 0, \quad (5.8)$$

from (5.6)

$$-6(h_1^K)^2 - 6(h_2^K)^2 + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2(h_2^K)^2 \leq 0 \quad (5.9)$$

and from (5.7)

$$-6(h_1^K)^2 + 3(h_2^K)^2 + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2(h_2^K)^2 \leq 0. \quad (5.10)$$

Without loss of generality we can assume $h_1^K \geq h_2^K$. (Otherwise it suffices to switch the notation for the lengths of the edges.) From (5.8) we have

$$\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2(h_2^K)^2 \leq -3(h_1^K)^2 + 6(h_2^K)^2,$$

thus

$$\begin{aligned}-6(h_1^K)^2 - 6(h_2^K)^2 + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2(h_2^K)^2 &\leq -6(h_1^K)^2 - 6(h_2^K)^2 - 3(h_1^K)^2 + 6(h_2^K)^2 \\ &= -9(h_1^K)^2 \leq 0\end{aligned}$$

and (5.9) holds. Similarly

$$\begin{aligned}-6(h_1^K)^2 + 3(h_2^K)^2 + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2(h_2^K)^2 &\leq -6(h_1^K)^2 + 3(h_2^K)^2 - 3(h_1^K)^2 + 6(h_2^K)^2 \\ &= -9(h_1^K)^2 + 9(h_2^K)^2.\end{aligned}$$

As $h_1^K \geq h_2^K$, (5.10) is also met. Hence conditions (5.9) and (5.10) are met automatically with (5.8) and we can state the following lemma:

Lemma 5.1. *Let $h_1^K \geq h_2^K$. If condition (5.8) is met, then*

$$\bar{a}_{jk}^K < 0 \ \forall j \neq k, \ j,k = 1,\ldots,4.$$

Proof.
It follows from the above analysis.

□

Theorem 5.1. *Let $\mathcal{T}_h$ be a partition of $\overline{\Omega}$ into rectangles. Then problem (2.6) satisfies the discrete maximum principle for $\kappa^2 = 0$ a.e. in Ω if*

$$0 \leq h_1^K \text{ and } \frac{h_1^K}{\sqrt{2}} \leq h_2^K \leq h_1^K, \text{ for all } K \in \mathcal{T}_h, \tag{5.11}$$

and for $\kappa^2 \neq 0$ if

$$0 \leq h_1^K \leq \sqrt{\frac{3}{\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})}}}, \text{ for all } K \in \mathcal{T}_h, \tag{5.12}$$

and

$$\frac{1}{\sqrt{2 - \frac{1}{3}\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2}} h_1^K \leq h_2^K \leq h_1^K, \text{ for all } K \in \mathcal{T}_h, \tag{5.13}$$

where h_1^K and h_2^K are lengths of sides of the rectangle K.

Proof.
First we consider the case $\kappa^2 = 0$ a.e. in Ω. If (5.11) is met, then

$$\begin{aligned} 3(h_1^K)^2 - 6(h_2^K)^2 + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2(h_2^K)^2 &= 3(h_1^K)^2 - 6(h_2^K)^2 \\ &\leq 3(h_1^K)^2 - 6\left(\frac{h_1^K}{\sqrt{2}}\right)^2 = 0 \end{aligned}$$

and (5.8) holds. Second, if $\kappa^2 \neq 0$ then (5.13) can be satisfied only if

$$\frac{1}{\sqrt{2 - \frac{1}{3}\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2}} h_1^K \leq h_1^K.$$

It is equivalent to

$$h_1^K \leq \sqrt{\frac{3}{\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})}}},$$

which is guaranteed by condition (5.12). Then condition (5.13) implies

$$3(h_1^K)^2 \leq 6(h_2^K)^2 - \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}(h_1^K)^2(h_2^K)^2$$

and hence (5.8) holds. Finally, Lemma 5.1 finishes the proof.

□

Remark 5.1. *If we give up the technical assumption* $h_1^K \geq h_2^K$ *then condition (5.11) can be reformulated as*

$$\frac{1}{\sqrt{2}} \leq \frac{h_1^K}{h_2^K} \leq \sqrt{2}, \text{ for all } K \in \mathcal{T}_h.$$

This give us a conclusion that the discrete maximum principle is satisfied for problem (2.6) on the rectangular finite element mesh if all rectangles $K \in \mathcal{T}_h$ *are non-narrow. It means that the lengths of both sides do not differ too much – the aspect ratio can vary from* $1/\sqrt{2}$ *to* $\sqrt{2}$.

Chapter 6

Finite elements in 3D – tetrahedra

The goal of this chapter is to describe the results for the discrete maximum principle of problem (2.6) for tetrahedral finite element meshes, mentioned e.g. in [2], [3], [26] or [27].

We consider problem (2.6) in a polytopic domain $\Omega \subset \mathbb{R}^3$ with a mesh $\mathcal{T}_h$ consisting of tetrahedra. For each tetrahedron $K \in \mathcal{T}_h$ we consider the following notation (see Figure 6.1): Let $B_1,\ B_2,\ B_3,\ B_4$ be the vertices of K. For $i = 1, \ldots, 4$, let F_i be the face of the tetrahedron opposite to vertex B_i, $|F_i|$ be the area of this face and h_i^K be the altitude of K over the face F_i. Furthermore, we denote θ_{ij} the dihedral angle between faces F_i and F_j, $i, j = 1, \ldots, 4$, which is defined as

$$\theta_{ij} = \pi - \rho_{ij},$$

where ρ_{ij} is the angle between the unit outward normals n_i and n_j to F_i and F_j. The space V_h is defined by (2.16), where $P_K = P^1(K)$, thus V_h is a space of continuous and piecewise-linear functions. The basis of V_h are functions $\varphi_1, \varphi_2, \ldots, \varphi_{N_h} \in V_h$ defined by (2.19).

As in the previous chapters we are going to formulate sufficient conditions for the validity of the discrete maximum principle of problem (2.6). Theorem 2.5 gives us the following condition for the local condensed FE-matrices A_h^K of problem (2.6):

$$\bar{a}_{jk}^K \leq 0 \ \forall j < k,\ j, k = 1, \ldots, 4,\ \forall K \in \mathcal{T}_h. \tag{6.1}$$

6.1 Local mass and stiffness matrices

For computing local mass and stiffness matrices we define barycentric coordinates λ_i^K as ratios of the two distances:

$$\lambda_i^K(x, y, z) = \frac{\mathrm{Dist}((x, y, z), F_i)}{\mathrm{Dist}(B_i, F_i)} = \frac{\mathrm{Dist}((x, y, z), F_i)}{h_i^K} = \frac{((x, y, z) - B_j) \cdot (-n_i)}{h_i^K},$$

where $(x, y, z) \in K$, h_i^K is the altitude of the tetrahedron K over its face F_i, B_j is an arbitrary vertex of F_i and $i = 1, \ldots, 4$. Clearly, λ_i^K is equal to 1 for the vertex

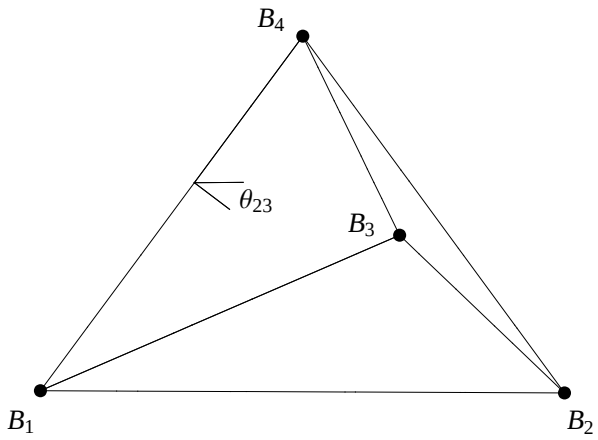


Figure 6.1: Notation for tetrahedra.

B_i and it vanishes for points on the face F_i. For points inside the tetrahedron it attains values between 0 and 1. This enable us to use barycentric coordinates as shape functions on K. Thus

$$\begin{aligned}\bar{s}_{ij}^K &= \int_K \nabla\varphi_j(x,y,z)\cdot\nabla\varphi_i(x,y,z)\,\mathrm{d}(x,y,z)\\ &= \int_K \nabla\lambda_j^K(x,y,z)\cdot\nabla\lambda_i^K(x,y,z)\,\mathrm{d}(x,y,z)\end{aligned}$$

and

$$\bar{m}_{ij}^K = \int_K \varphi_j(x,y,z)\,\varphi_i(x,y,z)\,\mathrm{d}(x,y,z) = \int_K \lambda_j^K(x,y,z)\,\lambda_i^K(x,y,z)\,\mathrm{d}(x,y,z).$$

Since

$$\nabla\lambda_i(x,y,z) = \nabla\frac{((x,y,z)-B_j)\cdot(-n_i)}{h_i^K} = -\frac{1}{h_i^K}n_i$$

we can compute

$$\begin{aligned}\bar{s}_{ij}^K &= \int_K \nabla\lambda_j^K(x,y,z)\cdot\nabla\lambda_i^K(x,y,z)\,\mathrm{d}(x,y,z) = \int_K \frac{1}{h_i^K h_j^K}n_i\cdot n_j\,\mathrm{d}(x,y,z)\\ &= \frac{|K|}{h_i^K h_j^K}n_i\cdot n_j,\end{aligned}$$

where $|K|$ is the volume of K. Furthermore,

$$n_i\cdot n_j = \cos\rho_{ij} = \cos(\pi-\theta_{ij}) = -\cos\theta_{ij},$$

where ρ_{ij} is the angle between n_i and n_j and θ_{ij} is the dihedral angle between faces F_i and F_j. Then

$$\bar{s}_{ij}^K = -\frac{|K|}{h_i^K h_j^K}\cos\theta_{ij}.$$

Therefore

$$\bar{S}_h^K = \left\{ -\frac{|K|}{h_i^K h_j^K} \cos\theta_{ij} \right\}_{i,j=1}^{4}.$$

As $|K| = h_i^K \left|F_i^K\right|/3$, $i = 1, \ldots, 4$, the local stiffness matrix can be written equivalently in the following form:

$$\bar{S}_h^K = \left\{ -\frac{|F_i|\,|F_j|}{9\,|K|} \cos\theta_{ij} \right\}_{i,j=1}^{4},$$

which was already derived e.g. in [2].

For computing entries of matrix $\bar{M}_h^K$ we use the following general formula for integrals of products of barycentric coordinates in the three-dimensional space proved in [23, p. 84]:

$$\int_K \lambda_1^{q_1}(x,y,z)\lambda_2^{q_2}(x,y,z)\lambda_3^{q_3}(x,y,z)\lambda_4^{q_4}(x,y,z)\,\mathrm{d}(x,y,z)$$

$$= \frac{q_1!q_2!q_3!q_4!3!}{(q_1+q_2+q_3+q_4+3)!}\,|K|\,.$$

This formula gives us for $q_i, q_j = 1$ and $q_k, q_l = 0$, $i \neq j \neq k \neq l$, $i,j,k,l = 1,\ldots,4$,

$$\bar{m}_{ij}^K = \int_K \lambda_j(x,y,z)\lambda_i(x,y,z)\,\mathrm{d}(x,y,z) = \frac{1!1!0!0!3!}{5!}\,|K| = \frac{|K|}{20}$$

and for $q_i = 2$ and $q_j, q_k, q_l = 0$, $i \neq j \neq k \neq l$, $i,j,k,l = 1,\ldots,4$,

$$\bar{m}_{ii}^K = \int_K \lambda_i^2(x,y,z)\,\mathrm{d}(x,y,z) = \frac{2!0!0!0!3!}{5!}\,|K| = \frac{2\,|K|}{20}.$$

Thus

$$\bar{M}_h^K = \frac{|K|}{20}\begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}.$$

6.2 Discrete maximum principle

The main result of this chapter is presented as the following two theorems.

Theorem 6.1. *For every $K \in \mathcal{T}_h$ let F_i^K and F_j^K be its faces, θ_{ij} be the dihedral angle between faces F_i^K and F_j^K and h_i^K, h_j^K be the altitudes of K above F_i^K, F_j^K. Matrix $\bar{A}_h^K$ has non-positive off-diagonal entries if*

$$\frac{\cos\theta_{ij}}{h_i^K h_j^K} \geq \frac{\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})}}{20}. \tag{6.2}$$

Proof.

Formula (2.25) and computed values of entries of $\bar{S}_h^K$ and $\bar{M}_h^K$ imply

$$\bar{a}_{ij}^K \leq -\frac{|K|}{h_i^K h_j^K} \cos\theta_{ij} + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{|K|}{20},$$

for $i \neq j$, $i, j = 1, \ldots, 4$. Hence, $\bar{a}_{ij}^K$ is non-negative if (6.2) is satisfied.

□

Theorem 6.2. *Let $\mathcal{T}_h$ be a partition of $\overline{\Omega}$ into the finite number of tetrahedra K. If condition (6.2) holds for every $K \in \mathcal{T}_h$ then problem (2.6) satisfies the discrete maximum principle.*

Proof.

It follows from Theorem 6.1 and from condition (6.1).

□

Remark 6.1. *From (6.2) we obtain for $\kappa^2 = 0$ a.e. in Ω:*

$$\cos\theta_{ij} \geq \frac{\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})} h_i^K h_j^K}{20} = 0$$

and thus

$$\theta_{ij} \leq \frac{\pi}{2}, \text{ for all } i, j = 1, \ldots 4, i \neq j.$$

Consequently, the discrete maximum principle is satisfied provided all angles in the tetrahedral mesh $\mathcal{T}_h$ are non-obtuse.

Theorem 6.2 bounds the dihedral angles and altitudes of tetrahedra in the finite element mesh in terms of the the reaction coefficient κ^2. As a result, it can be computed how small the acute tetrahedra should be for the discrete maximum principle to be valid. However, it was shown that these non-obtuseness (for $\kappa^2 = 0$ a.e. in Ω, see Remark 6.1) or acuteness (for $\kappa^2 \neq 0$) conditions are sufficient but not necessary. The discrete maximum principle may also hold under various weaker conditions on the simplicial meshes and some obtuse interior angles can be acceptable, see e.g. [25].

Restricting to acute tetrahedral meshes brings surprising problem: how to tile a three-dimensional domain using only acute tetrahedra. This problem turns out to be considerably harder than the non-obtuse triangulation problem, see e.g. [11], [33]. It is for example known that Euclidian space cannot be filled with regular tetrahedra, which restricts their possible use in practice. Also it is known that an acute simplicial partition of $\mathbb{R}^d$ does not exist for $d \geq 5$, see e.g. [28].

Chapter 7

Finite elements in 3D – blocks

Now we are going to show the results for the discrete maximum principle for block partitions and trilinear finite elements approximation. This topic was already treated e.g. in [26].

As above, we consider problem (2.6) in a polytopic domain $\Omega \subset \mathbb{R}^3$. Next, we consider a mesh $\mathcal{T}_h$ which forms a partition of $\overline{\Omega}$ into blocks. By blocks we understand rectangular cuboids, i.e. cartesian products of three intervals. Every block can be written as $K = R \times I$, where $I \in \mathbb{R}^1$ is an interval of the length h_3^K, which forms the altitude of the block K, and $R \in \mathbb{R}^2$, $|R| = h_1^K h_2^K$, is a rectangular base of K (see Figure 7.1). We can define the space V_h associated to $\mathcal{T}_h$ as

$$V_h = \{\, \varphi \in H_0^1(\Omega) : \varphi(x,y,z)\,|_K = \sum_{i=1}^{4}\sum_{j=1}^{2} b_{ij}\varphi_i^{(R)}(x,y)\varphi_j^{(I)}(z), \quad b_{ij} \in \mathbb{R},$$
$$\varphi_i^{(R)} \in Q^1(R), \quad \varphi_j^{(I)} \in P^1(I), \quad K \in \mathcal{T}_h, \quad K = R \times I \,\},$$

where $Q^1(R)$ is a space of bilinear functions defined on the rectangle R and $P^1(I)$ is a space of linear functions defined on the interval I. The functions $\varphi_1, \varphi_2, \ldots, \varphi_{N_h} \in V_h$ given by (2.19) form a basis of V_h.

As in the previous chapters we are solving the question of the validity of the discrete maximum principle for problem (2.6). Theorem 2.5 gives the following condition for the local condensed FE-matrices A_h^K of problem (2.6):

$$\bar{a}_{jk}^K \le 0 \ \forall j < k, \ j,k = 1,\ldots,8, \ \forall K \in \mathcal{T}_h. \tag{7.1}$$

7.1 Local mass and stiffness matrices

In computation of local mass and stiffness matrices of the block $K = R \times I$ we can use formulas for mass and stiffness matrices for intervals and rectangles obtained in Chapters 3 and 5: We consider a block $K = R \times I$ and the same notation as in Figure 7.1. We denote φ_A the basis function of V_h corresponding

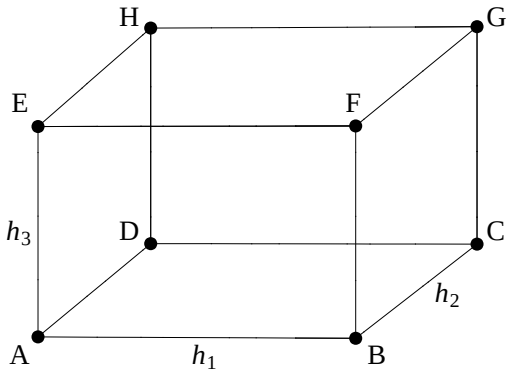


Figure 7.1: Notation for elements.

to the vertex A, φ_B the basis function corresponding to the vertex B, etc. The restriction of φ_A to the block K (and analogically restrictions of all other basis functions) can be written as a product of functions $\varphi_1^{(R)}(x,y)$ and $\varphi_1^{(I)}(z)$ where $\varphi_1^{(R)}(x,y)$ is the shape function on the rectangle R and $\varphi_1^{(I)}(z)$ is the shape function in the interval I, both corresponding to vertex A. Then we have

$$\bar{s}_{1,2} = \int_K \nabla\varphi_A \cdot \nabla\varphi_B \,\mathrm{d}(x,y,z) = \int_R \nabla\varphi_1^{(R)} \cdot \nabla\varphi_2^{(R)} \,\mathrm{d}(x,y) \int_I (\varphi_1^{(I)})^2 \,\mathrm{d}z + \int_R \varphi_1^{(R)}\varphi_2^{(R)} \,\mathrm{d}(x,y) \int_I \left[(\varphi_1^{(I)})'\right]^2 \mathrm{d}z,$$

$$\bar{m}_{1,2} = \int_K \varphi_A\varphi_B \,\mathrm{d}(x,y,z) = \int_R \varphi_1^{(R)}\varphi_2^{(R)} \,\mathrm{d}(x,y) \int_I (\varphi_1^{(I)})^2 \,\mathrm{d}z$$

and analogically we can compute all the other entries of matrices $\bar{S}_h^K$ and $\bar{M}_h^K$. Finally, we find that

$$\bar{S}_h^K = \frac{h_3^K}{6}\begin{pmatrix} 2S^{(R)} & S^{(R)} \\ S^{(R)} & 2S^{(R)} \end{pmatrix} + \frac{1}{h_3^K}\begin{pmatrix} M^{(R)} & -M^{(R)} \\ -M^{(R)} & M^{(R)} \end{pmatrix}$$

and

$$\bar{M}_h^K = \frac{h_3^K}{6}\begin{pmatrix} 2M^{(R)} & M^{(R)} \\ M^{(R)} & 2M^{(R)} \end{pmatrix},$$

where $S^{(R)}$, $M^{(R)}$ are stiffness and mass matrices on the rectangle R defined by (5.3) and (5.4) . Thus

$$\bar{S}_h^K = \frac{1}{36 h_1^K h_2^K h_1^K} \begin{pmatrix} a & b & c & d & e & f & g & h \\ b & a & d & c & f & e & h & g \\ c & d & a & b & g & h & e & f \\ d & c & b & a & h & g & f & e \\ e & f & g & h & a & b & c & d \\ f & e & h & g & b & a & d & c \\ g & h & e & f & c & d & a & b \\ h & g & f & e & d & c & b & a \end{pmatrix}, \tag{7.2}$$

where

$$\begin{aligned} a &= 4(h_1^K)^2(h_2^K)^2 + 4(h_1^K)^2(h_3^K)^2 + 4(h_2^K)^2(h_3^K)^2, \\ b &= 2(h_1^K)^2(h_2^K)^2 + 2(h_1^K)^2(h_3^K)^2 - 4(h_2^K)^2(h_3^K)^2, \\ c &= 2(h_1^K)^2(h_2^K)^2 - 4(h_1^K)^2(h_3^K)^2 + 2(h_2^K)^2(h_3^K)^2, \\ d &= (h_1^K)^2(h_2^K)^2 - 2(h_1^K)^2(h_3^K)^2 - 2(h_2^K)^2(h_3^K)^2, \\ e &= -4(h_1^K)^2(h_2^K)^2 + 2(h_1^K)^2(h_3^K)^2 + 2(h_2^K)^2(h_3^K)^2, \\ f &= -2(h_1^K)^2(h_2^K)^2 + (h_1^K)^2(h_3^K)^2 - 2(h_2^K)^2(h_3^K)^2, \\ g &= -2(h_1^K)^2(h_2^K)^2 - 2(h_1^K)^2(h_3^K)^2 + (h_2^K)^2(h_3^K)^2, \\ h &= -(h_1^K)^2(h_2^K)^2 - (h_1^K)^2(h_3^K)^2 - (h_2^K)^2(h_3^K)^2. \end{aligned}$$

The mass matrix can be expressed as

$$\bar{M}_h^K = \frac{h_1^K h_2^K h_3^K}{216} \begin{pmatrix} 8 & 4 & 4 & 2 & 4 & 2 & 2 & 1 \\ 4 & 8 & 2 & 4 & 2 & 4 & 1 & 2 \\ 4 & 2 & 8 & 4 & 2 & 1 & 4 & 2 \\ 2 & 4 & 4 & 8 & 1 & 2 & 2 & 4 \\ 4 & 2 & 2 & 1 & 8 & 4 & 4 & 2 \\ 2 & 4 & 1 & 2 & 4 & 8 & 2 & 4 \\ 2 & 1 & 4 & 2 & 4 & 2 & 8 & 4 \\ 1 & 2 & 2 & 4 & 2 & 4 & 4 & 8 \end{pmatrix}. \tag{7.3}$$

7.2 Discrete maximum principle

First of all, we introduce the following notation:

$$H_1 := (h_1^K)^2, H_2 := (h_2^K)^2, H_3 := (h_3^K)^2 \text{ and } S := \frac{(h_1^K)^2(h_2^K)^2(h_3^K)^2}{6}.$$

From (2.25), condition (7.1) and from computed values of entries of $\bar{S}_h^K$ and $\bar{M}_h^K$ we derive the following sufficient conditions for the discrete maximum principle:

$$2H_1H_2 + 2H_1H_3 - 4H_2H_3 + 4 \left\| \kappa^2 \right\|_{\mathcal{L}^\infty(\mathrm{K})} S \leq 0, \tag{7.4}$$

$$2H_1H_2 - 4H_1H_3 + 2H_2H_3 + 4\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} S \le 0, \tag{7.5}$$

$$H_1H_2 - 2H_1H_3 - 2H_2H_3 + 2\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} S \le 0, \tag{7.6}$$

$$-4H_1H_2 + 2H_1H_3 + 2H_2H_3 + 4\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} S \le 0, \tag{7.7}$$

$$-2H_1H_2 + H_1H_3 - 2H_2H_3 + 2\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} S \le 0, \tag{7.8}$$

$$-2H_1H_2 - 2H_1H_3 + H_2H_3 + 2\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} S \le 0, \tag{7.9}$$

$$-H_1H_2 - H_1H_3 - H_2H_3 + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} S \le 0, \tag{7.10}$$

Without loss of generality we can assume $h_1^K \ge h_2^K \ge h_3^K > 0$ and thus $H_1 \ge H_2 \ge H_3 > 0$. Then we have

$$\begin{aligned} 2H_1H_2 + 2H_1H_3 - 4H_2H_3 + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} S &\ge 2H_2H_3 + 2H_2H_3 - 4H_2H_3 \\ &\quad + 4\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} S = 4\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} S. \end{aligned}$$

Thus, condition (7.4) cannot be satisfied for $\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})} > 0$. If $\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})} = 0$ then (7.4) holds only for $H_1 = H_2 = H_3$ and it is obvious that also conditions (7.5)–(7.10) are satisfied in that case. Therefore, we can state the following lemma and theorem.

Lemma 7.1. *Let $\mathcal{T}_h$ be a partition of $\overline{\Omega}$ into the finite number of blocks. Then all off-diagonal entries of the FE-matrix associated to problem (2.6) are non-positive if $\kappa^2 = 0$ a.e. in Ω and*

$$h_1^K = h_2^K = h_3^K, \text{ for all } K \in \mathcal{T}_h.$$

Proof.

It follows from the above analysis.

□

Theorem 7.1. *Let $\mathcal{T}_h$ be a partition of $\overline{\Omega}$ into the finite number of blocks. Problem (2.6) satisfies the discrete maximum principle if $\kappa^2 = 0$ a.e. in Ω and*

$$h_1^K = h_2^K = h_3^K, \text{ for all } K \in \mathcal{T}_h,$$

where h_1^K, h_2^K, h_3^K are the lengths of sides of the block K.

Proof.

It follows from Lemma 7.1 and condition (7.1).

□

Hence, condition (7.1) for the discrete maximum principle of the diffusion-reaction problem is too restrictive in the case of block finite elements. It can never be satisfied unless the reaction coefficient κ^2 vanishes. Even in the pure diffusion case ($\kappa^2 = 0$ a.e. in Ω) condition (7.1) is satisfied only if all elements are cubes. Further, it was shown e.g. in [26] that condition (7.1) can never be satisfied for higher dimensional blocks ($d \ge 4$). Moreover, numerical experiments published in [26] indicate that the discrete maximum principle is not satisfied even on hypercubes for $d \ge 4$, because the corresponding stiffness matrix is not monotone.

Chapter 8

Finite elements in 3D – triangular prisms

The validity of the discrete maximum principle for prismatic elements considered in this chapter was not studied until quite recently, see [19], in spite of the fact that for certain types of domains prismatic meshes could be more natural (e.g. for cylindrical domains). In addition it is necessary to use prisms together with pyramids in hybrid finite element meshes as connecting elements between tetrahedra and blocks.

In this chapter we consider problem (2.6) in a polytopic domain $\Omega \subset \mathbb{R}^3$. Furthermore, we consider a mesh $\mathcal{T}_h$ which creates a partition of $\overline{\Omega}$ into triangular prisms. Every prism can be written as $K = T \times I$, where $I = [0, h_3^K] \in \mathbb{R}^1$ forms the altitude of the prism K and $T \in \mathbb{R}^2$ is its triangular base, α, β and γ are angles of T (see Figure 8.1). We can define the space V_h associated to $\mathcal{T}_h$ as

$$V_h = \{\, \varphi \in H_0^1(\Omega) : \varphi(x,y,z)\,|_K = \sum_{i=1}^{3}\sum_{j=1}^{2} b_{ij}\varphi_i^{(T)}(x,y)\varphi_j^{(I)}(z), \quad b_{ij} \in \mathbb{R},$$
$$\varphi_i^{(T)} \in P^1(T), \quad \varphi_j^{(I)} \in P^1(I), \quad K \in \mathcal{T}_h, \quad K = T \times I \,\},$$

where $P^1(T)$ and $P^1(I)$ are spaces of linear functions defined on the triangle T and in the interval I, respectively. Functions $\varphi_1, \varphi_2, \ldots, \varphi_{N_h}$ given by (2.19) form the basis of V_h. As in the previous chapters, Theorem 2.5 gives us the following sufficient conditions for the validity of the discrete maximum principle for the local condensed FE-matrices A_h^K of problem (2.6):

$$\bar{a}_{jk}^K \leq 0 \ \forall j < k, \ j,k = 1,\ldots,6, \ \forall K \in \mathcal{T}_h. \tag{8.1}$$

8.1 Local mass and stiffness matrices

In computation of local mass and stiffness matrices of the prism $K = T \times I$ we can use the formulas for mass and stiffness matrices for triangles and intervals

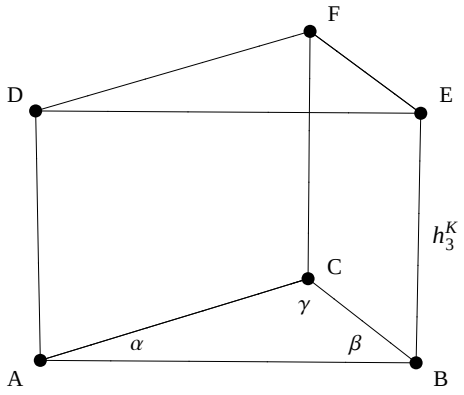


Figure 8.1: Notation for prismatic elements.

obtained in Chapters 3 and 4. We consider a prism $K = T \times I \in \mathcal{T}_h$ and we use the same notation as in Figure 8.1. We denote by φ_A the basis function of V_h corresponding to the vertex A of the prism, φ_B is the basis function corresponding to the vertex B, etc. Similarly to the previous chapter, a restriction of φ_A to the prism K (and analogically restrictions of all other basis functions) can be written as a product of functions $\varphi_1^{(T)}(x,y)$ and $\varphi_1^{(I)}(z)$, where $\varphi_1^{(T)}(x,y)$ is the shape function on the triangle T and $\varphi_1^{(I)}(z)$ is the shape function in the interval I, both corresponding to vertex A. Basis functions corresponding to the other vertices of $K \in \mathcal{T}_h$ can be treated analogically. The integrals needed for the evaluation of the entries of the local stiffness and mass matrices can be computed as follows:

$$\begin{aligned}\bar{s}_{1,2}^K &= \int_K \nabla\varphi_A \cdot \nabla\varphi_B \,\mathrm{d}(x,y,z) = \int_T \nabla\varphi_1^{(T)} \cdot \nabla\varphi_2^{(T)} \,\mathrm{d}(x,y) \int_I (\varphi_1^{(I)})^2 \,\mathrm{d}z \\ &\quad + \int_T \varphi_1^{(T)}\varphi_2^{(T)} \,\mathrm{d}(x,y) \int_I \left[(\varphi_1^{(I)})'\right]^2 \mathrm{d}z = -\frac{h_3^K}{12}\left(2\cot\gamma - \frac{|T|}{(h_3^K)^2}\right),\end{aligned}$$

$$\begin{aligned}\bar{s}_{1,4}^K &= \int_K \nabla\varphi_A \cdot \nabla\varphi_D \,\mathrm{d}(x,y,z) = \int_T \left|\nabla\varphi_1^{(T)}\right|^2 \mathrm{d}(x,y) \int_I \varphi_1^{(I)}\varphi_2^{(I)} \,\mathrm{d}z \\ &\quad + \int_T (\varphi_1^{(T)})^2 \,\mathrm{d}(x,y) \int_I (\varphi_1^{(I)})'(\varphi_2^{(I)})' \,\mathrm{d}z = \frac{h_3^K}{12}\left(\cot\beta + \cot\gamma - \frac{2\,|T|}{(h_3^K)^2}\right),\end{aligned}$$

$$\begin{aligned}\bar{s}_{1,5}^K &= \int_K \nabla\varphi_A \cdot \nabla\varphi_E \,\mathrm{d}(x,y,z) = \int_T \nabla\varphi_1^{(T)} \cdot \nabla\varphi_2^{(T)} \,\mathrm{d}(x,y) \int_I \varphi_1^{(I)}\varphi_2^{(I)} \,\mathrm{d}z \\ &\quad + \int_T \varphi_1^{(T)}\varphi_2^{(T)} \,\mathrm{d}(x,y) \int_I (\varphi_1^{(I)})'(\varphi_2^{(I)})' \,\mathrm{d}z = -\frac{h_3^K}{12}\left(\cot\gamma + \frac{|T|}{(h_3^K)^2}\right),\end{aligned}$$

$$\bar{m}_{1,2}^K = \int_K \varphi_A\varphi_B \,\mathrm{d}(x,y,z) = \int_T \varphi_1^{(T)}\varphi_2^{(T)} \,\mathrm{d}(x,y) \int_I (\varphi_1^{(I)})^2 \,\mathrm{d}z = \frac{h_3^K\,|T|}{36},$$

$$\bar{m}^K_{1,4} = \int_K \varphi_A \varphi_D \,\mathrm{d}(x,y,z) = \int_T (\varphi_1^{(T)})^2 \,\mathrm{d}(x,y) \int_I \varphi_1^{(I)} \varphi_2^{(I)} \,\mathrm{d}z = \frac{h_3^K\,|T|}{36},$$
$$\bar{m}^K_{1,5} = \int_K \varphi_A \varphi_E \,\mathrm{d}(x,y,z) = \int_T \varphi_1^{(T)} \varphi_2^{(T)} \,\mathrm{d}(x,y) \int_I \varphi_1^{(I)} \varphi_2^{(I)} \,\mathrm{d}z = \frac{h_3^K\,|T|}{72},$$

and analogically all the other entries of matrices $\bar{S}_h^K$ and $\bar{M}_h^K$. Similarly to the previous chapter we find that

$$\bar{S}_h^K = \frac{h_3^K}{6} \begin{pmatrix} 2S^{(T)} & S^{(T)} \\ S^{(T)} & 2S^{(T)} \end{pmatrix} + \frac{1}{h_3^K} \begin{pmatrix} M^{(T)} & -M^{(T)} \\ -M^{(T)} & M^{(T)} \end{pmatrix}$$

and

$$\bar{M}_h^K = \frac{h_3^K}{6} \begin{pmatrix} 2M^{(T)} & M^{(T)} \\ M^{(T)} & 2M^{(T)} \end{pmatrix} = \frac{h_3^K\,|T|}{72} \begin{pmatrix} 4 & 2 & 2 & 2 & 1 & 1 \\ 2 & 4 & 2 & 1 & 2 & 1 \\ 2 & 2 & 4 & 1 & 1 & 2 \\ 2 & 1 & 1 & 4 & 2 & 2 \\ 1 & 2 & 1 & 2 & 4 & 2 \\ 1 & 1 & 2 & 2 & 2 & 4 \end{pmatrix},$$

where $S^{(T)}$ and $M^{(T)}$ are defined by (4.3) and (4.2). In other words matrices $\bar{S}_h^K$ and $\bar{M}_h^K$ can be written as tensor products of matrices $S^{(T)}$, $M^{(T)}$, $S^{(I)}$ and $M^{(I)}$:

$$\bar{S}_h^K = M^{(I)} \otimes S^{(T)} + S^{(I)} \otimes M^{(T)} \text{ and } \bar{M}_h^K = M^{(I)} \otimes M^{(T)},$$

where $S^{(I)}$ and $M^{(I)}$ are given by (3.9).

8.2 Discrete maximum principle

Condition (8.1), formula (2.25) and computed values of entries of $\bar{S}_h^K$ and $\bar{M}_h^K$ yield the following conditions for the discrete maximum principle:

$$-\frac{h_3^K}{12}\left(2\cot\delta_1 - \frac{|T|}{(h_3^K)^2}\right) + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{h_3^K\,|T|}{36} \le 0, \tag{8.2}$$

$$\frac{h_3^K}{12}\left(\cot\delta_1 + \cot\delta_2 - \frac{2\,|T|}{(h_3^K)^2}\right) + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{h_3^K\,|T|}{36} \le 0, \tag{8.3}$$

$$-\frac{h_3^K}{12}\left(\cot\delta_1 + \frac{|T|}{(h_3^K)^2}\right) + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{h_3^K\,|T|}{72} \le 0, \quad \delta_1, \delta_2 \in \{\alpha, \beta, \gamma\}, \quad \delta_1 \ne \delta_2. \tag{8.4}$$

Definition 8.1 (Lower and upper bound). *Let $K = T \times I$ be a prism and $0 < \alpha_{min} \le \alpha_{med} \le \alpha_{max}$ be the minimal, medium and maximal angles of the triangle*

T and let $|T|$ be the area of T. Then the lower bound h_L and the upper bound h_U for the altitude h_3^K of K are defined by

$$h_L = \left(\frac{2 \cot \alpha_{max}^{(T)}}{|T|} - \frac{\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})}}{3} \right)^{-1/2} \tag{8.5}$$

and

$$h_U = \left(\frac{\cot \alpha_{min}^{(T)} + \cot \alpha_{med}^{(T)}}{2\,|T|} + \frac{\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})}}{6} \right)^{-1/2}. \tag{8.6}$$

Remark 8.1. *As $\alpha_{min}^{(T)} < \pi/2$ and $\alpha_{med}^{(T)} < \pi/2$, the upper bound h_U is always well defined. The lower bound h_L is well defined only for*

$$\frac{2 \cot \alpha_{max}^{(T)}}{|T|} - \frac{\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})}}{3} > 0.$$

Now we can state the following theorem (from [19]):

Theorem 8.1. *Let $\mathcal{T}_h$ be a prismatic partition of $\overline{\Omega}$ such that h_L is well defined. Problem (2.6) satisfies the discrete maximum principle if*

$$h_L \le h_3^K \le h_U, \ \textit{for all} \ \ K \in \mathcal{T}_h, \tag{8.7}$$

where h_3^K is the altitude of the triangular prism $K \in \mathcal{T}_h$ and h_L and h_D are the lower and upper bounds defined by (8.5) and (8.6).

Proof.

As the cotangent is a decreasing function, it is $-\cot \delta_1 \le -\cot \alpha_{max}^{(T)}$ and $\cot \delta_1 + \cot \delta_2 \le \cot \alpha_{min}^{(T)} + \cot \alpha_{med}^{(T)}$ for $\delta_1, \delta_2 \in \{\alpha, \beta, \gamma\}$, $\delta_1 \ne \delta_2$. Therefore it is possible to consider conditions

$$-\frac{h_3^K}{12} \left(2 \cot \alpha_{max}^{(T)} - \frac{|T|}{(h_3^K)^2} \right) + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{h_3^K\,|T|}{36} \le 0,$$

$$\frac{h_3^K}{12} \left(\cot \alpha_{min}^{(T)} + \cot \alpha_{med}^{(T)} - \frac{2\,|T|}{(h_3^K)^2} \right) + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{h_3^K\,|T|}{36} \le 0,$$

$$-\frac{h_3^K}{12} \left(\cot \alpha_{max}^{(T)} + \frac{|T|}{(h_3^K)^2} \right) + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{h_3^K\,|T|}{72} \le 0$$

instead of (8.2)–(8.4). Hence if (8.7) holds we have

$$-\frac{h_3^K}{12} \left(2 \cot \alpha_{max}^{(T)} - \frac{|T|}{(h_3^K)^2} \right) + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{h_3^K\,|T|}{36}$$

$$\le \frac{h_3^K}{12} \left(-2 \cot \alpha_{max}^{(T)} + \frac{|T|}{h_L^2} + \left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})} \frac{|T|}{3} \right) = 0,$$

$$\frac{h_3^K}{12}\left(\cot\alpha_{min}^{(T)}+\cot\alpha_{med}^{(T)}-\frac{2\,|T|}{(h_3^K)^2}\right)+\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{h_3^K\,|T|}{36}$$
$$\leq\frac{h_3^K}{12}\left(\cot\alpha_{min}^{(T)}+\cot\alpha_{med}^{(T)}-\frac{2\,|T|}{h_U^2}+\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{|T|}{3}\right)=0$$

and

$$-\frac{h_3^K}{12}\left(\cot\alpha_{max}^{(T)}+\frac{|T|}{(h_3^K)^2}\right)+\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{h_3^K\,|T|}{72}\leq\frac{h_3^K}{12}\left(-\cot\alpha_{max}^{(T)}-\frac{|T|}{h_U^2}\right.$$
$$\left.+\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}\frac{|T|}{6}\right)=\frac{h_3^K}{24}\left(-2\cot\alpha_{max}^{(T)}-\cot\alpha_{min}^{(T)}-\cot\alpha_{med}^{(T)}\right)\leq 0.$$

Finally, condition (8.1) finishes the proof.

□

Remark 8.2. *Notice that both h_L and h_U are given by the angles of the triangular base of the prism, but not for all combinations of angles it is $h_L \leq h_U$. Thus, the altitude h_3^K satisfying (8.7) does not have to exist in general and some conditions for angles of the triangular base have to be formulated.*

For the existence of an altitude of the prism that satisfy (8.7) it is necessary to have $h_L \leq h_U$. Thus, from (8.5) and (8.6) we get the following condition:

$$\left(\frac{2\cot\alpha_{max}^{(T)}}{|T|}-\frac{\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})}}{3}\right)^{-1/2}\leq\left(\frac{\cot\alpha_{min}^{(T)}+\cot\alpha_{med}^{(T)}}{2\,|T|}+\frac{\|\kappa^2\|_{\mathcal{L}^\infty(\mathrm{K})}}{6}\right)^{-1/2}.$$

After reformulation, we have:

$$4\cot\alpha_{max}^{(T)}-\left\|\kappa^2\right\|_{\mathcal{L}^\infty(\mathrm{K})}|T|\geq\cot\alpha_{min}^{(T)}+\cot\alpha_{med}^{(T)}.$$

For simplicity, we further consider only the case $\kappa^2 = 0$ a.e. in Ω. Thus we have to find such conditions for angles that

$$4\cot\alpha_{max}^{(T)}\geq\cot\alpha_{min}^{(T)}+\cot\alpha_{med}^{(T)}.$$

As $\alpha_{med}^{(T)}=\pi-\alpha_{max}^{(T)}-\alpha_{min}^{(T)}$ and

$$\cot(\pi-\alpha-\beta)=\frac{1-\cot\alpha\cot\beta}{\cot\alpha+\cot\beta},\tag{8.8}$$

we have after this substitution

$$4\cot\alpha_{max}^{(T)}-\cot\alpha_{min}^{(T)}-\frac{1-\cot\alpha_{min}^{(T)}\cot\alpha_{max}^{(T)}}{\cot\alpha_{min}^{(T)}+\cot\alpha_{max}^{(T)}}\geq 0,$$

which is equivalent to

$$4\cot^2\alpha_{max}^{(T)} - \cot^2\alpha_{min}^{(T)} - 1 + 4\cot\alpha_{min}^{(T)}\cot\alpha_{max}^{(T)} \geq 0.$$

Furthermore,

$$\alpha_{min}^{(T)} = \pi - \alpha_{max}^{(T)} - \alpha_{med}^{(T)} \geq \pi - 2\alpha_{max}^{(T)}$$

and therefore using again formula (8.8) we get

$$\cot(\alpha_{max}^{(T)}) \leq \cot(\alpha_{min}^{(T)}) \leq \frac{1-\cot^2\alpha_{max}^{(T)}}{2\cot\alpha_{max}^{(T)}}.$$

Hence, the sufficient condition for $\alpha_{max}^{(T)}$ is

$$4\cot^2\alpha_{max}^{(T)} - \left(\frac{1-\cot^2\alpha_{max}^{(T)}}{2\cot\alpha_{max}^{(T)}}\right)^2 - 1 + 4\cot^2\alpha_{max}^{(T)} \geq 0.$$

It is, after denoting $Z = \cot^2\alpha_{max}^{(T)}$,

$$4Z - \frac{1-2Z+Z^2}{4Z} - 1 + 4Z \geq 0,$$

which simplifies to

$$31Z^2 - 1 - 2Z \geq 0.$$

From this we get for $\cot\alpha_{max}^{(T)}$:

$$\cot\alpha_{max}^{(T)} \geq \sqrt{\frac{1+4\sqrt{2}}{31}},$$

and thus

$$\alpha_{max}^{(T)} \leq \arctan\sqrt{\frac{31}{1+4\sqrt{2}}} \approx 65,14°.$$

Lemma 8.1. *Let $0 < \alpha_{min}^{(T)} \leq \alpha_{med}^{(T)} \leq \alpha_{max}^{(T)}$ be the angles of the triangular base T of a prism $K = T \times I$. If*

$$\alpha_{max}^{(T)} \leq \arctan\sqrt{\frac{31}{1+4\sqrt{2}}} \approx 65,14°$$

then there exists an altitude h_3^K of K that satisfy condition (8.7) with $\kappa^2 = 0$ a.e. in Ω.

Proof.

It follows from the above analysis.

□

The condition from the previous lemma is obviously highly limiting and the case $\kappa^2 \neq 0$ yields even more restrictive conditions. A more favorable limitations for the angles of the triangular basis of prisms of the mesh were made in [19] (Lemma 3 and Theorem 3). However, as shown by numerical computations, also these estimates are only sufficient, not necessary.

Theorem 8.1 gives the conditions that limit not only the shape of every prism $K \in \mathcal{T}_h$ (i.e. the angles of the triangular base of every K) but also its altitude. Furthermore, it bounds the altitude from both sides, therefore it can be problematic to divide the given domain into the prisms suitable for the validity of the discrete maximum principle.

Chapter 9

Conclusions

The aim of this work was to give a summary of the known results for the discrete maximum principle in the first order finite element method for the diffusion-reaction problem. The most common types of finite elements in 1D, 2D and 3D were considered. The results were shown for the meshes consisting of intervals, triangles, rectangles, tetrahedra, blocks and triangular prisms. For every type of above mentioned finite elements the corresponding mass and stiffness matrices were computed and then the sufficient conditions for the validity of the discrete maximum principle were formulated.

The results obtained for the meshes consisting of triangles and rectangles are also applicable for hybrid meshes in 2D where triangular and rectangular elements are used together. One could expect the same possibility of application in 3D hybrid meshes. However, the hybrid meshes consisting only of tetrahedra and blocks cannot exist in 3D and pyramids together with right triangular prisms have to be used to join the elements face-to-face. Nonetheless, no analysis of conditions for the validity of the discrete maximum principle for pyramidal finite elements has been carried out so far.

Conditions for the validity of the discrete maximum principle presented in this thesis are based on the assumption, that the FE-matrix should be an M-matrix to have the non-negative inverse. They are not sharp and there is a space for further improvements. Certain generalization can be obtained by requiring the FE-matrix not to be an M-matrix but to be monotone only. It is well known that some off-diagonal entries can be positive and still inverse of the FE-matrix A_h^{-1} can be non-negative. Theoretical handling of monotone matrices is more difficult, but possible – see e.g. [1] or [25].

The study of the discrete maximum principle is a wide-ranging and interesting field, therefore there is a lot of further possible generalizations, improvements and open problems. For instance, it would be interesting to look at the validity of the discrete maximum principle for different and more sophisticated problems, e.g. elliptic problems with other types of boundary conditions, parabolic problems (see e.g. [12], [13], [14], [15], [17]), nonlinear problems (e.g. in [24], [29]) or problems in higher dimensions (see e.g. [3]).

Another area of investigation might be the discrete maximum principle for higher-order finite element methods, see e.g [22], [35], [36], [37] or [39]. Further possibility can be to design a new method which maintains the discrete maximum principle automatically (similarly as e.g. in [4] or [38]).

Numerical methods for second-order partial differential equations satisfying the discrete maximum principle have important applications in real-life problems where physically non-negative quantities are computed. Further possible extensions in analysis of this phenomenon could help practitioners to design their computations in harmony with the underlying physics.

Bibliography

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

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

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

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

[5] Céa J.: *Approximation variationelle des problémes aux limites*, Ann. Inst. Fourier (Grenoble) **14** (1964) 345–444.

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

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

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

[9] Ciarlet P.G., Varga R.S.: *Discrete variational Green's Function*, Numer. Math. **16** (1970) 115–128.

[10] Drăgănescu A., Dupont T.F., Scott L.R.: *Failure of the discrete maximum principle for an elliptic finite element problem*, Math. Comp. **74** (2004) 1–23.

[11] Eppstein D., Sullivan J.M., Üngör A.: *Tiling space and slabs with acute tetrahedra*, Computational Geometry **27** (2004) 237–255.

[12] Faragó I.: *Positivity of the finite difference scheme for the linear parabolic problem*, Colloquia Mathematica Societatis János Bolyai, 62. Differencial equations, Budapest (Hungary), 1991.

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

[14] Faragó I., Komáromi N.: *Nonnegativity of the numerical solution of parabolic problems*, Colloquia Mathematica Societatis János Bolyai, 59. Numerical Methods, Miskolc (Hungary), 1990.

[15] Faragó I., Tarvainen P.: *Quantitative analysis of matrix splitting methods*, Computers Math. Applic. **42** (2001) 1055–1067.

[16] Fiedler M.: *Special matrices and their applications in numerical mathematics*, Martinus Nijhoff Publishers, Dordrecht/Boston/Lancaster, 1986.

[17] Fujii H.: *Some remarks on finite element analysis of time-dependent field problems*, Theory and practice in finite element structural analysis, Proceedings of the Seminar, Tokyo, Japan, 5-10 Nov. 1973, pp. 91–106.

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

[19] 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** (2009) 257–287.

[20] Hannukainen A., Korotov S., Vejchodský T.: *On weakening conditions for discrete maximum principles for linear finite element schemes*, in S. Margenov, L.G. Vulkov, J. Wasniewski (eds), Numerical Analysis and Its Applications, Lecture Notes in Computer Science **5434**, Springer-Verlag, Berlin, 2009, pp. 297–304.

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

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

[23] Holand I., Bell K.: *Finite element methods in stress analysis*, Tapir, Trondheim, 1969.

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

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

[26] Korotov S., Vejchodský T.: *A comparison of simplicial and block finite elements*, to appear in Numerical Mathematics and Advanced Applications ENUMATH 2009 , Springer, 2010.

[27] Křížek M.: *On the maximum angle condition for linear tetrahedral elements*, SIAM J. Numer Anal. **29** (1992) 513–520.

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

[29] Křížek M., Liu L.: *On a comparison principle for a quasilinear elliptic boundary value problem of a nonmonotone type*, Appl. math. (Warsaw) **24,1** (1996) 97–107.

[30] Protter H.M., Weinberger H. F.: *Maximum principles in differential equations*, Springer, 1984.

[31] Pucci P., Serrin J.: *The maximum principle*, Birkhauser-Verlag, 2007.

[32] Šolín P.: *Partial differential equations and the finite element method*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

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

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

[35] Vejchodský T.: *Angle conditions for discrete maximum principles in higher-order FEM*, to appear in Numerical Mathematics and Advanced Applications ENUMATH 2009 , Springer, 2010.

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

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

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

[39] Yanik E.G.: *Sufficient conditions for a discrete maximum principle for high order collocation method*, Computers Math. Applic. **17** (1989) 1431–1434.