Survey of discrete maximum principles for linear elliptic and parabolic problems.

Brandts, J.H.; Korotov, S.; Krizek, M.

Publication date
2004

Published in
Proc. Conf. ECCOMAS 2004

European Congress on Computational Methods in Applied Sciences and Engineering
ECCOMAS 2004
P. Neittaanmäki, T. Rossi, S. Korotov, E. Oñate, J. Périaux, and D. Knörzer (eds.)
Jyväskylä, 24–28 July 2004

SURVEY OF DISCRETE MAXIMUM PRINCIPLES FOR LINEAR ELLIPTIC AND PARABOLIC PROBLEMS

J. Brandts[⋆], S. Korotov[†], and M. Křížek[‡]

[⋆]Korteweg-de Vries Institute
Faculty of Science, University of Amsterdam
Plantage Muidergracht 24, 1018 TV Amsterdam, Netherlands
e-mail: brandts@science.uva.nl

[†]Department of Mathematical Information Technology
P.O. Box 35 (Agora), FIN-400 14 University of Jyväskylä, Finland
e-mail: korotov@mit.jyu.fi

[‡] Mathematical Institute, Academy of Sciences
Žitná 25, CZ-115 67 Prague 1, Czech Republic
e-mail: krizek@math.cas.cz

Key words: maximum principle, discrete maximum principle, elliptic equation, parabolic equation, finite element method, tetrahedral mesh

Abstract. *We survey techniques for proving discrete maximum principles for finite element approximations of linear elliptic and parabolic problems. Special emphasis is laid on approximations built on tetrahedral meshes.*

1 INTRODUCTION

Besides an obligatory requirement of convergence of computed approximations to the exact solution of a model under investigation, the approximations are naturally required to mirror basic qualitative properties of this exact solution in order to be reliable and useful in computer simulation and visualization.

It is well known that solutions of mathematical models described by second order elliptic and parabolic equations satisfy the maximum principles (see, e.g., [15], [16]). Its discrete analogues (so-called the discrete maximum principles/DMPs) were first presented and analysed in the papers [2], [3], and [7] for elliptic and parabolic cases, respectively. Later, different aspects of DMPs for various types of problems were discussed in [1], [4], [5], [6], [8], [9], [11], [12], [13], [14], [17], [18], [19], [20].

In most of those papers, it was noticed that DMPs hold true if certain geometric restrictions are imposed on the computational meshes used. Thus, triangular meshes are required to be acute or nonobtuse [3], [7], rectangular meshes to be non-narrow [1], etc.

In this work, we present a general framework for proving discrete maximum principles for finite element approximations of linear elliptic and parabolic problems, and separately discuss in detail the case of approximations built on tetrahedral meshes.

2 ELLIPTIC EQUATIONS

2.1 Problem setting and the continuous maximum principle

The first model (of elliptic type) consists of the Poisson equation with Dirichlet boundary condition: Find a function u such that

$$-\Delta u = f \quad \text{in} \quad \Omega, \tag{1}$$

$$u = u^0 \quad \text{on} \quad \partial\Omega, \tag{2}$$

where Ω is a bounded polytopic domain with Lipschitz boundary $\partial\Omega$, and f, u^0 are given functions. In this paper, we assume that all given functions are sufficiently smooth and that a classical solution u of problem (1)–(2) exists and is unique. The continuous maximum principle for problem (1)–(2) is expressed by the following theorem (for the proof see [16]).

Theorem 2.1 *The classical solution u of problem* (1)–(2) *satisfies*

$$\max_{x\in\Omega} u(x) \le \max\{0; \max_{s\in\partial\Omega} u^0(s)\}, \tag{3}$$

provided $f \le 0$.

Remark 2.2 *In many papers dealing with maximum principles for elliptic problems (see, e.g.,* [13] *and references therein), the authors consider only homogeneous Dirichlet boundary conditions, i.e., $u^0 \equiv 0$ in* (2)*. The maximum principle then takes the simpler form*

$$\max_{x\in\Omega} u(x) \le 0, \tag{4}$$

provided $f \leq 0$.

2.2 Discretization

2.2.1 Variational formulation

The variational formulation for (1)–(2) reads as follows: Find $u \in H^1(\Omega)$ such that

$$a(u,v) = F(v) \quad \forall v \in H_0^1(\Omega) \qquad \text{and} \qquad u - u^0 \in H_0^1(\Omega), \tag{5}$$

where

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx \qquad \text{and} \qquad F(v) = \int_\Omega f\, v \, dx. \tag{6}$$

2.2.2 Computational scheme

We assume that a simplicial partition $\mathcal{T}_h$ of Ω is given, where h denotes the discretization parameter. Let $P_1, \dots, P_N$ denote interior nodes, and $P_{N+1}, \dots, P_{\bar{N}}$ boundary nodes in $\mathcal{T}_h$. We also let $N_\partial := \bar{N} - N$.

Let $\phi_1, \dots, \phi_{\bar{N}}$ be the continuous piecewise linear nodal basis functions. It is obvious that

$$\phi_i \geq 0, \quad i = 1, \dots, \bar{N}, \qquad \text{and} \qquad \sum_{i=1}^{\bar{N}} \phi_i \equiv 1 \;\text{ in }\; \overline{\Omega}. \tag{7}$$

We denote the span of the basis functions by $V^h \subset H^1(\Omega)$, and define its subspace $V_0^h = \{v \in V^h \mid v|_{\partial\Omega} = 0\} \subset H_0^1(\Omega)$.

The discrete formulation of (1)–(2) reads: Find $u_h \in V^h$ such that

$$a(u_h, v_h) = F(v_h) \quad \forall v_h \in V_0^h \qquad \text{and} \qquad u_h - u_h^0 \in V_0^h, \tag{8}$$

where u_h^0 is a given approximation of u^0. In particular, we can take

$$u_h^0 = \sum_{i=1}^{N_\partial} u^0(P_{N+i})\, \phi_{N+i}. \tag{9}$$

Then $u_h^0(P_{N+i}) = u^0(P_{N+i}),\ i = 1, \dots, N_\partial$. Looking for u_h in the form

$$u_h = \sum_{i=1}^{\bar{N}} u_i \phi_i = \sum_{i=1}^{N} u_i \phi_i + \sum_{i=1}^{N_\partial} u^0(P_{N+i}) \phi_{N+i}\,, \tag{10}$$

we arrive at the following system of linear algebraic equations

$$\bar{\mathbf{A}} \bar{\mathbf{u}} = \bar{\mathbf{b}}. \tag{11}$$

In the above,

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{A}_\partial \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{12}$$

$$\bar{\mathbf{u}} = [u_1, \ldots, u_{\bar{N}}]^T = [\mathbf{u} \mid \mathbf{u}_\partial]^T, \quad \mathbf{u} \in \mathbf{R}^N, \quad \mathbf{u}_\partial \in \mathbf{R}^{N_\partial}, \tag{13}$$

$$\bar{\mathbf{b}} = [b_1, \ldots, b_N, u^0(P_{N+1}), \ldots, u^0(P_{\bar{N}})]^T = [\mathbf{b} \mid \mathbf{u}_\partial]^T, \quad \mathbf{b} \in \mathbf{R}^N, \tag{14}$$

$$\mathbf{A} = (a_{ij})_{i=1,\ j=1}^{N,\ N}, \quad \text{where} \quad a_{ij} = a(\phi_i, \phi_j), \tag{15}$$

$$\mathbf{A}_\partial = (a_{ij}^\partial)_{i=1,\ j=1}^{N,\ N_\partial}, \quad \text{where} \quad a_{ij}^\partial = a(\phi_i, \phi_{N+j}), \tag{16}$$

$$b_i = \int_\Omega f\, \phi_i \, dx, \quad i = 1, \ldots, N, \tag{17}$$

the symbol $\mathbf{I}$ stands for the $N_\partial \times N_\partial$ identity matrix, and $\mathbf{0}$ denotes the $N_\partial \times N$ zero matrix. In what follows all matrix and vector inequalities are meant elementwise.

2.3 The discrete maximum principle

2.3.1 Formulation of DMP

A discrete analogue of maximum principle (3), in terms of the solution of system (11), reads as follows:

$$\max_{x \in \Omega} u_h(x) = \max_{i=1,\ldots,N} u_i \le \max\{0, \max_{j=N+1,\ldots,\bar{N}} u^0(P_j)\} = \max\{0, \max_{s \in \partial\Omega} u_h^0(s)\}, \tag{18}$$

provided $\mathbf{b} \le \mathbf{0}$.

2.3.2 Algebraic conditions guaranteeing the validity of DMP

Theorem 2.3 *The DMP* (18) *holds for all admissible* $\mathbf{b}$ *iff the matrix* $\bar{\mathbf{A}}$ *is monotone, i.e.,* $\bar{\mathbf{A}}$ *is nonsingular and*

$$\bar{\mathbf{A}}^{-1} \ge \mathbf{0}, \tag{A1}$$

and

$$-\mathbf{A}^{-1}\mathbf{A}_\partial \, \mathbf{e}_\partial \le \mathbf{e}_0, \tag{A2}$$

where

$$\mathbf{e}_0 = [1, \ldots, 1]^T \in \mathbf{R}^N \quad \text{and} \quad \mathbf{e}_\partial = [1, \ldots, 1]^T \in \mathbf{R}^{N_\partial}. \tag{19}$$

Proof. Let us denote $\bar{\mathbf{A}}^{-1}$ by $\bar{\mathbf{G}}$. Then, obviously,

$$\bar{\mathbf{G}} = \begin{bmatrix} \mathbf{G} & \mathbf{G}_\partial \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \text{where} \quad \mathbf{G} = \mathbf{A}^{-1} \qquad \text{and} \qquad \mathbf{G}_\partial = -\mathbf{A}^{-1}\mathbf{A}_\partial. \tag{20}$$

Assume first, that DMP (18) holds. We shall now show that the matrix $\bar{\mathbf{A}}$ is nonsingular, which is equivalent to $\mathbf{A}$ being nonsingular. Let the vector $\mathbf{u} = [u_1, \dots, u_N]^T$ be such that $\mathbf{Au} = \mathbf{0}$. Then $\bar{\mathbf{v}} = [\mathbf{u} \mid \mathbf{0}_\partial]^T \in \mathbf{R}^{\bar{N}}$, (where $\mathbf{0}_\partial$ is a zero vector of the length N_∂) is such that $\bar{\mathbf{A}}\bar{\mathbf{v}} = \mathbf{0}$. Thus, $\pm\mathbf{u} \leq \mathbf{0}$ in view of (18). Hence, $\mathbf{u} = \mathbf{0}$, which proves that both matrices $\mathbf{A}$ and $\bar{\mathbf{A}}$ are nonsingular.

Next, from representation (12), we have the identity

$$\mathbf{u} = \mathbf{Gb} + \mathbf{G}_\partial \mathbf{u}_\partial. \tag{21}$$

Let $\mathbf{G} = (g_{ij})_{i=1,\ j=1}^{N,\ N}$ and $\mathbf{G}_\partial = (g_{ij}^\partial)_{i=1,\ j=1}^{N,\ N_\partial}$. For each $j = 1, \dots, N$, the vector $\mathbf{g}_j = (g_{1j}, \dots, g_{Nj}) \in \mathbf{R}^N$ is the solution of (11) with the vector $\mathbf{b} = [0, \dots, 1, \dots, 0]^T$, where the 1 is in the j-th position and $\mathbf{u}_\partial = \mathbf{0}_\partial$. Thus, $\mathbf{g}_j \geq 0$ by the DMP (18), which proves that $\mathbf{G} \geq \mathbf{0}$.

Similarly, for each $j = 1, \dots, N_\partial$, $\mathbf{g}_j^\partial = (g_{1j}^\partial, \dots, g_{Nj}^\partial) \in \mathbf{R}^N$ is the solution of (11) for $\mathbf{b} = \mathbf{0}$, and $\mathbf{u}_\partial = [0, \dots, 1, \dots, 0]^T$, where the 1 is in the j-th position. Thus, by (18), it follows again that $\mathbf{g}_j^\partial \geq 0$, which proves that $\mathbf{G}_\partial \geq \mathbf{0}$. Thus, (A1) is proved.

Now, we prove (A2). The vector $\mathbf{g} = \sum_{j=1}^{N_\partial} g_j^\partial$ is the unique solution of (11) with $\mathbf{b} = \mathbf{0}$ and $\mathbf{u}_\partial = \mathbf{e}_\partial$. Thus, from (18), we get $(\mathbf{g})_i \ \leq \ 1\,, \quad i = 1, \dots, N$, which proves (A2), because $\mathbf{g} = \mathbf{G}_\partial \mathbf{e}_\partial = -\mathbf{A}^{-1}\mathbf{A}_\partial \mathbf{e}_\partial$.

Conversely, let us assume that (A1) and (A2) hold. Since $\bar{\mathbf{A}}$ is nonsingular in view of (A1), identity (21) holds, therefore,

$$u_i = \sum_{j=1}^{N} g_{ij} b_j + \sum_{j=1}^{N_\partial} g_{ij}^\partial u_j^\partial, \quad i = 1, \dots, N\,.$$

Then, obviously, for $\mathbf{b} \ \leq \ \mathbf{0}$, $\bar{\mathbf{u}}$ satisfies (18), as $g_{ij} \geq 0$ and $g_{ij}^\partial \geq 0$ (in view of (A1)), and $\sum_{j=1}^{N_\partial} g_{ij}^\partial \leq 1, \ i = 1, \dots, N$ (in view of (A2)). ■

Condition (A2) is not so easy to verify in practice. Therefore, in the following theorem we give a simpler (sufficient) condition instead of (A2).

Theorem 2.4 *The DMP* (18) *is valid if the following two conditions are satisfied:*

$$\bar{\mathbf{A}} \text{ is monotone}\,, \tag{A1*}$$

$$\mathbf{Ae}_0 + \mathbf{A}_\partial \mathbf{e}_\partial \ \geq \ 0\,. \tag{A2*}$$

PROOF. Since $\mathbf{A}^{-1} = \mathbf{G} \geq 0$ by (A1*), we observe from (A2*) that

$$\mathbf{e}_0 + \mathbf{A}^{-1}\mathbf{A}_\partial \mathbf{e}_\partial \geq 0\,, \qquad \text{or} \qquad -\mathbf{A}^{-1}\mathbf{A}_\partial \mathbf{e}_\partial \leq \mathbf{e}_0\,, \tag{22}$$

which proves (A2). Condition (A1) coincides with (A1*). ■

Condition (A2*) actually means that

$$\sum_{j=1}^{\bar{N}} a_{ij} \geq 0\,, \quad i = 1, \dots, N\,, \tag{23}$$

or, equivalently, that $\bar{\mathbf{A}}$ is diagonally dominant.

Generally, (18) does not imply (A2*). The following matrix (cf. [2, p. 343])

$$\bar{\mathbf{A}}' = \begin{bmatrix} -1 & 2 & 0 \\ 2 & -3 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{24}$$

satisfies DMP, but (A2*) is, obviously, not fulfilled.

More suitable (sufficient) conditions providing the validity of DMP (18) are given in the following theorem.

Theorem 2.5 *The DMP* (18) *is valid if*

$$\textit{diagonal entries } a_{ii},\ i = 1, \dots, N,\ \textit{are positive}, \tag{A1$'$}$$

$$\textit{off-diagonal entries } a_{ij},\ i = 1, \dots, N,\ j = 1, \dots, \bar{N}\ (i \neq j),\ \textit{are nonpositive}, \tag{A2$'$}$$

$$\textit{matrix } \bar{\mathbf{A}} \textit{ is diagonally dominant (see [19, p. 23]), i.e., } \mathbf{A}\mathbf{e}_0 + \mathbf{A}_\partial \mathbf{e}_\partial \geq 0\,, \tag{A3$'$}$$

$$\textit{matrix } \mathbf{A} \textit{ is irreducibly diagonally dominant, see [19, p. 23]}. \tag{A4$'$}$$

PROOF. Condition (A3$'$) is equivalent to (A2*). From (A4$'$), (A3$'$), and (A1$'$), we have $\mathbf{A}^{-1} \geq 0$, (see [19, p. 85]). Further, since $\mathbf{A}_\partial \leq 0$ by (A2$'$), we observe that $G_\partial = -\mathbf{A}^{-1}\mathbf{A}_\partial \geq 0$, proving that $\bar{\mathbf{A}}$ is monotone, i.e., condition (A1*) holds. ■

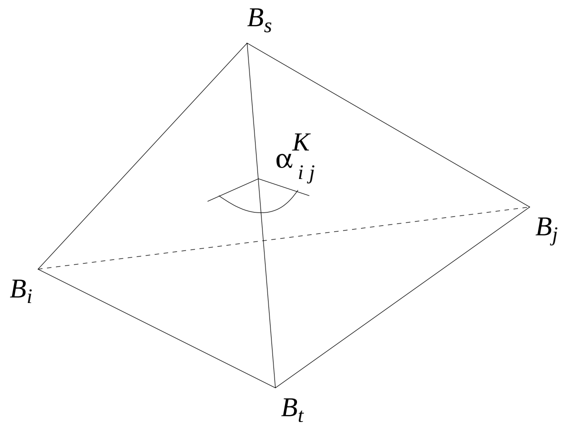

Figure 1: Tetrahedron K.

2.4 The validity of DMP on tetrahedral meshes

2.4.1 Entries of the finite element matrix

First, we recall the following result (see [13, p. 63]). Let $K = B_iB_jB_sB_t$ be an arbitrary tetrahedron. Let p and q be two linear functions such that

$$p(B_i) = 1,\ p(B_j) = p(B_s) = p(B_t) = 0, \quad \text{and} \quad q(B_j) = 1,\ q(B_i) = q(B_s) = q(B_t) = 0,$$

then

$$\nabla p \cdot \nabla q = -\frac{\text{meas}_2 B_iB_sB_t \cdot \text{meas}_2 B_jB_sB_t}{9\,(\text{meas}_3 K)^2} \cos \alpha_{ij}^K, \quad \|\nabla p\| = \frac{\text{meas}_2 B_jB_sB_t}{3\,\text{meas}_3 K}, \tag{25}$$

where α_{ij}^K is the angle between faces $B_iB_sB_t$ and $B_jB_sB_t$ (see Fig. 1) and the symbol meas_d stands for d-dimensional measure. Thus, each obtuse dihedral angle of the tetrahedron K_l gives a positive contribution to the corresponding off-diagonal entry of the element stiffness matrix, and each nonobtuse dihedral angle gives a nonpositive contribution. Further, we can write

$$a_{ij} = \sum_{K \in \mathcal{T}_{ij}^-} \text{meas}_3 K\, \nabla\phi_i \cdot \nabla\phi_j + \sum_{K \in \mathcal{T}_{ij}^+} \text{meas}_3 K\, \nabla\phi_i \cdot \nabla\phi_j, \tag{26}$$

where $\mathcal{T}_{ij}^- = \{K \in \mathcal{T}_h \,|\, \nabla\phi_i \cdot \nabla\phi_j < 0 \text{ on } K\}$, $\quad \mathcal{T}_{ij}^+ = \{K \in \mathcal{T}_h \,|\, \nabla\phi_i \cdot \nabla\phi_j > 0 \text{ on } K\}$.

2.4.2 Nonobtuse tetrahedral meshes

A tetrahedron is said to be *nonobtuse (acute)* if all its six dihedral angles between faces are nonobtuse (acute).

Theorem 2.6 *The DMP* (18) *holds for the linear finite element approximations of problem* (1)–(2) *if the employed tetrahedral mesh* $\mathcal{T}_h$ *is nonobtuse (i.e., consists of nonobtuse tetrahedra only).*

PROOF. Conditions (A1′) and (A2′) of Theorem 2.5 are, obviously, valid. Condition (A3′) can be rewritten in the following form:

$$\sum_{j=1}^{N} a_{ij} + \sum_{j=1}^{N_\partial} a_{ij}^{\partial} \;\geq\; 0\,, \quad i = 1, \ldots, N\,.$$

Actually,

$$\sum_{j=1}^{N} a_{ij} + \sum_{j=1}^{N_\partial} a_{ij}^{\partial} = \sum_{j=1}^{N} a(\phi_i, \phi_j) + \sum_{j=N+1}^{\bar{N}} a(\phi_i, \phi_j) = a(\phi_i, \sum_{j=1}^{\bar{N}} \phi_j) = a(\phi_i, 1) = 0\,,$$

for all $i = 1, \ldots, N$, i.e., condition (A3′) is satisfied as well.

Now, we check (A4′). Matrix $\mathbf{A}$ is irreducible, since its directed (oriented) graph is strongly connected (cf. [19, p. 20]). The fact that $\mathbf{A}$ is diagonally dominant follows immediately from (A2′) and (A3′) already proved above.

The property of $\mathbf{A}$ being irreducibly dominant now obviously follows from the fact that $\mathbf{A}$ is positive definite. To prove this, we observe that

$$\mathbf{A}\eta \cdot \eta = \sum_{i,j=1}^{N} a_{ij}\,\eta_i\,\eta_j = \sum_{i,j=1}^{N} a(\phi_i, \phi_j)\,\eta_i\,\eta_j =$$

$$= a(\sum_{i=1}^{N} \eta_i \phi_i, \sum_{j=1}^{N} \eta_j \phi_j) = a(v_h, v_h) = \int_{\Omega} |\nabla v_h|^2\,dx = \|\nabla v_h\|_0^2 \geq C\,\|v_h\|_1^2 > 0,$$

for all $\eta = [\eta_1, \ldots, \eta_N]^T \neq 0$, where $v_h = \sum_{i=1}^{N} \eta_i\,\phi_i$ and where the Friederichs' inequality was applied.

Thus, in view of Theorem 2.5, we have now proved the statement. ∎

2.4.3 Weakened nonobtuse type conditions

From the representation (26), we immediately observe that the requirement for the tetrahedral elements of the mesh to be nonobtuse in order to get the validity of DMP is too strong. In fact, some tetrahedra in the mesh can be nonobtuse, thus giving positive contributions to the corresponding entry, but the total sum of the contributions can still be nonpositive.

Condition $(A2')$ on off-diagonal entries (to be nonpositive) is only sufficient. We can observe the validity of DMP even in case when some off-diagonal entries are positive (in this case also some tetrahedra may have obtuse dihedral angles).

An analysis of the above cases gives rise to the so-called "weakened nonobtuse type conditions" [10, 17, 18].

3 PARABOLIC MODEL

3.1 Problem setting and the continuous maximum principle

Consider the second model (of parabolic type): Find a function $u = u(t, x)$ such that

$$\frac{\partial u}{\partial t} - c\Delta u = f \quad \text{in} \quad (0,T)\times\Omega, \tag{27}$$

$$u = g \quad \text{on} \quad [0,T)\times\partial\Omega, \quad \text{and} \quad u|_{t=0} = u_0 \quad \text{in} \quad \Omega, \tag{28}$$

where Ω is, as before, a bounded polytopic domain with Lipschitz boundary $\partial\Omega$, $T > 0$, c is a positive constant, and f, g, u_0 are given functions. We assume that all the given functions are sufficiently smooth and that the classical solution of the above problem exists and is unique.

We formulate now the continuous maximum principle for this type of problems. Let Q_t stand for the cylinder $(0,t)\times\Omega$, and Γ_t – for the union of its lateral surface S_t and its bottom Γ_0, $t \in [0,T]$. The following theorem holds (cf. Theorem 2.1 from [15]).

Theorem 3.1 *Let $u(t,x)$ be the classical solution of problem* (27)–(28) *in Q_T. Then*

$$\begin{aligned} &\sup_{\lambda>0} \min\{0; \min_{\Gamma_{t_1}}(\exp(\lambda(t_1-t))\psi); \tfrac{1}{\lambda}\min_{Q_{t_1}}(\exp(\lambda(t_1-t))f)\} \;\le\; u(t_1,x) \;\le \\ &\le\; \inf_{\lambda>0} \max\{0; \max_{\Gamma_{t_1}}(\exp(\lambda(t_1-t))\psi); \tfrac{1}{\lambda}\max_{Q_{t_1}}(\exp(\lambda(t_1-t))f)\} \end{aligned} \tag{29}$$

holds for any $t_1 \in [0,T]$, where the function ψ coincides with u_0 on Γ_0, and with g on S_T. In particular, $u(t,x) \ge 0$ provided $f \ge 0$, $u_0 \ge 0$, and $g \ge 0$.

Let us further introduce the following functions ($0 \le t \le t_1 \le T$, t_1 is fixed):

$$\tilde{v}(t,x) = \max\{0; \max_{\Gamma_{t_1}}\psi\} + t\max\{0; \max_{Q_{t_1}} f\} - u(t,x),$$

and

$$\bar{v}(t,x) = u(t,x) - \min\{0; \min_{\Gamma_{t_1}}\psi\} - t\min\{0; \min_{Q_{t_1}} f\},$$

where u and ψ are defined above. Due to positivity of the given data of the initial boundary-value problems to that the functions $\tilde{v}$ and $\bar{v}$ satisfy, and Theorem 3.1, we immediately observe that

$$\tilde{v}(t,x),\ \bar{v}(t,x) \geq 0,\ \text{for } t \in [0,t_1],\ \text{i.e.,}\quad \tilde{v}(t_1,x),\ \bar{v}(t_1,x) \geq 0,$$

which implies that for all $t_1 \in [0,T]$,

$$\min\{0; \min_{\Gamma_{t_1}} \psi\} + t_1 \min\{0; \min_{Q_{t_1}} f\} \leq u(t_1,x) \leq \max\{0; \max_{\Gamma_{t_1}} \psi\} + t_1 \max\{0; \max_{Q_{t_1}} f\}. \quad (30)$$

Formula (30) represents the form of the continuous maximum (and also minimum) principle that we shall deal with for the above parabolic problem.

3.2 Discretization

3.2.1 Weak formulation

The weak formulation for (27)–(28) reads as follows: Find $u = u(t,x) \in H^1(\Omega)$ for $t \in (0,T)$ such that

$$\int_\Omega \frac{\partial u}{\partial t} v \, dx + L(u,v) = \int_\Omega f v \, dx \quad \forall v \in H_0^1(\Omega), \quad t \in (0,T), \quad (31)$$

and

$$u(0,x) = u^0(x),\ x \in \Omega, \quad \text{and} \quad u - g \in H_0^1(\Omega), \quad t \in (0,T), \quad (32)$$

where $L(u,v) = c \int_\Omega \nabla u \cdot \nabla v \, dx$.

3.2.2 Semidiscretization in space

The semidiscrete problem for (31)–(32) reads: Find a function $u_h = u_h(t,x)$ such that

$$\int_\Omega \frac{\partial u_h}{\partial t} v_h \, dx + L(u_h,v_h) = \int_\Omega f v_h \, dx \quad \forall v_h \in V_0^h, \quad t \in (0,T), \quad (33)$$

and

$$u_h(0,x) = u_h^0(x), \quad x \in \Omega,\ u_h(t,x) - g_h(t,x) \in V_0^h, \quad t \in (0,T), \quad (34)$$

where $u_h^0(x)$ and $g_h(t,x)$ (for any fixed t) are suitable approximations of $u_0(x)$ and $g(t,x)$, respectively. In what follows, we assume that they are linear interpolants in V^h, i.e.,

$$u_h^0(x) = \sum_{i=1}^{\bar{N}} u_0(P_i)\phi_i(x), \quad (35)$$

and

$$g_h(t,x) = \sum_{i=1}^{N_\partial} g_i^h(t)\phi_{N+i}(x), \quad \text{where} \quad g_i^h(t) = g(t, P_{N+i}),\ i = 1,\dots,N_\partial, \tag{36}$$

(see Subsection 2.2.2 for the notation and definitions).

From the consistency of the initial and the boundary conditions $g(0,s) = u_0(s),\ s \in \partial\Omega$, we have $g_i^h(0) = u_0(P_{N+i}),\ i = 1,\dots,N_\partial$.

We search for a semidiscrete solution of the form

$$u_h(t,x) = \sum_{i=1}^{N} u_i^h(t)\phi_i(x) + g_h(t,x), \tag{37}$$

and notice that it is sufficient that u_h satisfies (33) only for $v_h = \phi_i$, $i = 1,\dots,N$.

Introducing the notation

$$\mathbf{v}^h(t) = [u_1^h(t),\dots,u_N^h(t), g_1^h(t),\dots,g_{N_\partial}^h(t)]^T, \tag{38}$$

we arrive at a Cauchy problem for the systems of ordinary differential equations,

$$\mathbf{M}\frac{\mathrm{d}\mathbf{v}^h}{\mathrm{d}t} + \mathbf{K}\mathbf{v}^h = \mathbf{f}, \quad \mathbf{v}^h(0) = [u^0(P_1),\dots,u^0(P_N), g_1^h(0),\dots,g_{N_\partial}^h(0)]^T \tag{39}$$

for the solution of the semidiscrete problem, where

$$\mathbf{M} = (m_{ij})_{i=1,\,j=1}^{N,\ \bar{N}}, \quad m_{ij} = \int_\Omega \phi_j\phi_i\,dx, \quad \mathbf{K} = (k_{ij})_{i=1,\,j=1}^{N,\ \bar{N}}, \quad k_{ij} = L(\phi_j,\phi_i),$$

$$\mathbf{f} = [f_1,\dots,f_N]^T, \quad f_i = \int_\Omega f\phi_i\,dx.$$

3.2.3 Fully discretized problem

In order to get a fully discrete numerical scheme, we choose a time-step Δt and denote the approximations to $\mathbf{v}^h(n\Delta t)$ and $\mathbf{f}(n\Delta t)$ by $\mathbf{v}^n$ and $\mathbf{f}^n$, $n = 0,1,\dots,n_T$ ($n_T\Delta t = T$), respectively. To discretize (39), we apply the θ-method ($\theta \in [0,\ 1]$ is a given parameter) and obtain a system of linear algebraic equations

$$\mathbf{M}\frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{\Delta t} + \theta\mathbf{K}\mathbf{v}^{n+1} + (1-\theta)\mathbf{K}\mathbf{v}^n = \mathbf{f}^{(n,\theta)} \quad (:= \theta\,\mathbf{f}^{n+1} + (1-\theta)\mathbf{f}^n\,), \tag{40}$$

which can be rewritten as

$$(\mathbf{M} + \theta\Delta t\mathbf{K})\mathbf{v}^{n+1} = (\mathbf{M} - (1-\theta)\Delta t\mathbf{K})\mathbf{v}^n + \Delta t\,\mathbf{f}^{(n,\theta)}, \quad n = 0,1,\dots,n_T - 1, \tag{41}$$

where $\mathbf{v}^0 = \mathbf{v}^h(0)$.

Further, let $\mathbf{A} = \mathbf{M} + \theta \Delta t \mathbf{K}$ and $\mathbf{B} = \mathbf{M} - (1-\theta)\Delta t \mathbf{K}$. We shall use the following partitions of the matrices and vectors:

$$\mathbf{A} = [\mathbf{A}_0|\mathbf{A}_\partial], \quad \mathbf{B} = [\mathbf{B}_0|\mathbf{B}_\partial], \quad \mathbf{v}^n = [\mathbf{u}^n|\mathbf{g}^n]^T, \tag{42}$$

where $\mathbf{A}_0$ and $\mathbf{B}_0$ are $(N \times N)$ matrices, $\mathbf{A}_\partial, \mathbf{B}_\partial$ are of size $(N \times N_\partial)$, $\mathbf{u}^n = [u_1^n, \ldots, u_N^n]^T \in \mathbf{R}^N$ and $\mathbf{g}^n = [g_1^n, \ldots, g_{N_\partial}^n]^T \in \mathbf{R}^{N_\partial}$. Similar partitions are used for matrices $\mathbf{M}$ and $\mathbf{K}$. Iteration (41) can now also be written as

$$\mathbf{A}\mathbf{v}^{n+1} = \mathbf{B}\mathbf{v}^n + \Delta t\, \mathbf{f}^{(n,\theta)}, \tag{43}$$

or

$$[\mathbf{A}_0|\mathbf{A}_\partial] \begin{bmatrix} \mathbf{u}^{n+1} \\ \mathbf{g}^{n+1} \end{bmatrix} = [\mathbf{B}_0|\mathbf{B}_\partial] \begin{bmatrix} \mathbf{u}^{n} \\ \mathbf{g}^{n} \end{bmatrix} + \Delta t\, \mathbf{f}^{(n,\theta)}. \tag{44}$$

3.3 The discrete maximum principle

3.3.1 Formulation of DMP

Let us define the following values for $n = 0, \ldots, n_T$:

$$g_{min}^n = \min\{0, g_1^n, \ldots, g_{N_\partial}^n\}, \quad g_{max}^n = \max\{0, g_1^n, \ldots, g_{N_\partial}^n\}, \tag{45}$$

$$v_{min}^n = \min\{0, g_{min}^n, u_1^n, \ldots, u_N^n\}, \quad v_{max}^n = \max\{0, g_{max}^n, u_1^n, \ldots, u_N^n\}, \tag{46}$$

and

$$f_{min}^{(n,n+1)} = \min\{0, \min_{x\in\Omega, \tau\in(n\Delta t,(n+1)\Delta t)} f(\tau,x)\}, \; f_{max}^{(n,n+1)} = \max\{0, \max_{x\in\Omega, \tau\in(n\Delta t,(n+1)\Delta t)} f(\tau,x)\} \tag{47}$$

for $n = 0, \ldots, n_T - 1$.

The discrete analogue (DMP) for the continuous maximum principle (30) can be represented as follows:

$$\min\{0, v_{min}^0, \min\{g_{min}^k, k=1,\ldots,n\}\} + n\Delta t \min\{0, \min\{f_{min}^{(k,k+1)}, k=0,\ldots,n-1\}\} \leq$$

$$\leq u_i^n \leq \tag{48}$$

$$\leq \max\{0, v_{max}^0, \max\{g_{max}^k, k=1,\ldots,n\}\} + n\Delta t \max\{0, \max\{f_{max}^{(k,k+1)}, k=0,\ldots,n-1\}\},$$

where $i = 1, \ldots, N, \; n = 1, \ldots, n_T$.

The DMP (48) easily follows from the following relation (cf. [7, p. 100]):

$$\min\{0, g_{min}^{n+1}, v_{min}^n\} + \Delta t f_{min}^{(n,n+1)} \leq u_i^{n+1} \leq \max\{0, g_{max}^{n+1}, v_{max}^n\} + \Delta t f_{max}^{(n,n+1)}, \tag{49}$$

$$i = 1, \ldots, N; \quad n = 0, \ldots, n_T - 1.$$

3.3.2 Algebraic conditions guaranteeing the validity of DMP

Write

$$\mathbf{e} = [1, \dots, 1]^T \in \mathbf{R}^{\bar{N}}, \quad \mathbf{e}_0 = [1, \dots, 1]^T \in \mathbf{R}^{N}, \quad \mathbf{e}_\partial = [1, \dots, 1]^T \in \mathbf{R}^{N_\partial}, \tag{50}$$

$$\begin{aligned} \mathbf{f}^{(n,n+1)}_{max} &= f^{(n,n+1)}_{max}\mathbf{e} \in \mathbf{R}^{\bar{N}}, & \mathbf{v}^n_{max} &= v^n_{max}\mathbf{e} \in \mathbf{R}^{\bar{N}}, \\ \mathbf{f}^{(n,n+1)}_0 &= f^{(n,n+1)}_{max}\mathbf{e}_0 \in \mathbf{R}^{N}, & \mathbf{v}^n_0 &= v^n_{max}\mathbf{e}_0 \in \mathbf{R}^{N}, \\ \mathbf{f}^{(n,n+1)}_\partial &= f^{(n,n+1)}_{max}\mathbf{e}_\partial \in \mathbf{R}^{N_\partial}, & \mathbf{v}^n_\partial &= v^n_{max}\mathbf{e}_\partial \in \mathbf{R}^{N_\partial}. \end{aligned} \tag{51}$$

Lemma 3.2 *We have*

$$\begin{array}{ll} (P1) & \mathbf{Ke} = \mathbf{0}, \\ (P2) & \mathbf{f}^{(n,\theta)} \leq \mathbf{Af}^{(n,n+1)}_{max}, \\ (P3) & \text{If } \mathbf{A}_0^{-1} \geq \mathbf{0}, \text{ then } -\mathbf{A}_0^{-1}\mathbf{A}_\partial\,\mathbf{e}_\partial \leq \mathbf{e}_0. \end{array} \tag{52}$$

Proof. (P1) For the i-th coordinate of the vector $\mathbf{Ke}$, we have

$$(\mathbf{Ke})_i = \sum_{j=1}^{\bar{N}} k_{ij} = \sum_{j=1}^{\bar{N}} L(\phi_j, \phi_i) = L\left(\sum_{j=1}^{\bar{N}} \phi_j, \phi_i\right) = L(1, \phi_i) = c\int_\Omega \nabla 1 \cdot \nabla\phi_i\, dx = 0, \tag{53}$$

which proves the statement.

(P2) For the i-th element of $\mathbf{f}^{(n,\theta)}$, we observe that

$$(\mathbf{f}^{(n,\theta)})_i = \int_\Omega ((1-\theta)f(n\Delta t, x) + \theta f((n+1)\Delta t, x)\phi_i(x))\, dx \leq \int_\Omega f^{(n,n+1)}_{max}\phi_i(x)\, dx =$$

$$= f^{(n,n+1)}_{max}\int_\Omega \left(\sum_{j=1}^{\bar{N}} \phi_j(x)\right)\phi_i(x)\, dx = f^{(n,n+1)}_{max}\sum_{j=1}^{\bar{N}} m_{ij} = \left(\mathbf{Mf}^{(n,n+1)}_{max}\right)_i = \tag{54}$$

$$= \left((\mathbf{M} + \theta\Delta t\mathbf{K})\mathbf{f}^{(n,n+1)}_{max}\right)_i = \left(\mathbf{Af}^{(n,n+1)}_{max}\right)_i,$$

where in the above, we used (P1).

(P3) Matrix $\mathbf{M}$ is non-negative, because the basis functions are nonnegative. Thus, $\mathbf{0} \leq \mathbf{Me} = (\mathbf{M} + \theta\Delta t\mathbf{K})\mathbf{e} = \mathbf{Ae} = \mathbf{A}_0\mathbf{e}_0 + \mathbf{A}_\partial\mathbf{e}_\partial$, and $(P3)$ is obtained by multiplying both sides by the non-negative matrix $\mathbf{A}_0^{-1}$. ■

Theorem 3.3 *Galerkin approximation for the solution of problem* (27)–(28)*, combined with the* θ*-method for time discretization, satisfies* (49) *if and only if*

$$\mathbf{A}_0^{-1} \geq \mathbf{0}, \qquad (C1)$$

$$\mathbf{A}_0^{-1}\mathbf{A}_\partial \leq \mathbf{0}, \qquad (C2)$$

$$\mathbf{A}_0^{-1}\mathbf{B} \geq \mathbf{0}. \qquad (C3)$$

Proof. First, we prove that (C1)–(C3) are sufficient by verifying the inequality on the right-hand side in (49). From (43) and (P2), we have

$$\mathbf{A}_0\mathbf{u}^{n+1} + \mathbf{A}_\partial\mathbf{g}^{n+1} = \mathbf{A}\mathbf{v}^{n+1} = \mathbf{B}\mathbf{v}^n + \Delta t\ \mathbf{f}^{(n,\theta)} \leq \mathbf{B}\mathbf{v}^n + \Delta t\mathbf{A}\mathbf{f}_{max}^{(n,n+1)}. \qquad (55)$$

From (P1), we find $\mathbf{B}\mathbf{v}_{max}^n = \mathbf{A}\mathbf{v}_{max}^n$. Multiplying both sides of (55) by $\mathbf{A}_0^{-1} \geq \mathbf{0}$ (see (C1)), expressing $\mathbf{u}^{n+1}$ and using (C3), we obtain

$$\begin{aligned}
\mathbf{u}^{n+1} &\leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial\ \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{B}\mathbf{v}^n + \Delta t\mathbf{A}_0^{-1}\mathbf{A}\mathbf{f}_{max}^{(n,n+1)} \leq \qquad (56)\\
&\leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial\ \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{B}\mathbf{v}_{max}^n + \Delta t\mathbf{A}_0^{-1}\mathbf{A}\mathbf{f}_{max}^{(n,n+1)} = \\
&= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\ \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{A}\mathbf{v}_{max}^n + \Delta t\mathbf{A}_0^{-1}\mathbf{A}\mathbf{f}_{max}^{(n,n+1)} = \\
&= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\ \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}[\mathbf{A}_0|\ \mathbf{A}_\partial]\mathbf{v}_{max}^n + \Delta t\mathbf{A}_0^{-1}[\mathbf{A}_0|\ \mathbf{A}_\partial]\ \mathbf{f}_{max}^{(n,n+1)} = \\
&= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\ \mathbf{g}^{n+1} + \mathbf{v}_0^n + \mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{v}_\partial^n + \Delta t\mathbf{f}_0^{(n,n+1)} + \Delta t\mathbf{A}_0^{-1}\mathbf{A}_\partial\ \mathbf{f}_\partial^{(n,n+1)}.
\end{aligned}$$

Regrouping the above inequality, we get

$$\mathbf{u}^{n+1} - \mathbf{v}_0^n - \Delta t\mathbf{f}_0^{(n,n+1)} \leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial(\mathbf{g}^{n+1} - \mathbf{v}_\partial^n - \Delta t\mathbf{f}_\partial^{(n,n+1)}). \qquad (57)$$

Hence, for the i-th coordinate of both sides we obtain

$$\begin{aligned}
u_i^{n+1} - v_{max}^n - \Delta t\ f_{max}^{(n,n+1)} &\leq \sum_{j=1}^{N_\partial} \left(-\mathbf{A}_0^{-1}\mathbf{A}_\partial\right)_{ij} (g_j^{n+1} - v_{max}^n - \Delta t\ f_{max}^{(n,n+1)}) \leq \qquad (58)\\
&\leq \left(\sum_{j=1}^{N_\partial} \left(-\mathbf{A}_0^{-1}\mathbf{A}_\partial\right)_{ij}\right) \cdot \max\{0, \max_j\{g_j^{n+1} - v_{max}^n\}\} \leq \\
&\leq \max\{0, \max_j\{g_j^{n+1} - v_{max}^n\}\} \leq \max\{0, g_{max}^{n+1}\},
\end{aligned}$$

where we applied (C2) and (P3). Finally, isolating u_i^{n+1}, we obtain the required inequality. The inequality on the left-hand side of (49) can be proved similarly. Hence, (C1)–(C3) are sufficient.

Now, let the DMP be valid, then it is valid for any choice of $\mathbf{f}^{(n,\theta)}$, $\mathbf{g}^n$, $\mathbf{g}^{n+1}$, and $\mathbf{u}^n$. With choice $\mathbf{g}^{n+1} = 0$, $\mathbf{v}^n = \mathbf{0}$, $\mathbf{f}^{(n,\theta)} = \mathbf{e}_j$, we get $\mathbf{A}_0^{-1} \geq \mathbf{0}$. Indeed, combining (44) and (49) we observe $\mathbf{0} \leq \mathbf{u}^{n+1} = \Delta t \mathbf{A}_0^{-1} \mathbf{e}_j$, which means that each column of $\mathbf{A}_0^{-1}$ is nonnegative, and thus, $\mathbf{A}_0^{-1} \geq \mathbf{0}$. So the necessity of (C1) is proved. Using again (44) and (49) with the choice $\mathbf{g}^{n+1} = \mathbf{e}_j$, $\mathbf{g}^n = \mathbf{0}$, $\mathbf{f}^{(n,\theta)} = \mathbf{0}$ and $\mathbf{u}^n = \mathbf{0}$, we obtain the necessity of (C2), and similarly, with $\mathbf{g}^{n+1} = \mathbf{0}$, $\mathbf{v}^n = \mathbf{e}_j$, $\mathbf{f}^{(n,\theta)} = \mathbf{0}$, we get the necessity of condition (C3). ∎

Remark 3.4 *It is easy to see that conditions* $(C1)$–$(C3)$ *are ensured by*

$$\mathbf{A}_0^{-1} \geq \mathbf{0}, \tag{$C1^\star$}$$

$$\mathbf{A}_\partial \leq \mathbf{0}, \tag{$C2^\star$}$$

$$\mathbf{B} \geq \mathbf{0}. \tag{$C3^\star$}$$

Theorem 3.5 *Galerkin approximation for the solution of problem* (27)–(28)*, combined with the θ-method for time discretization, satisfies the discrete maximum principle* (49) *if*

$$k_{ij} \leq 0, \quad i \neq j,\ i = 1, \ldots, N,\ j = 1, \ldots, \bar{N}, \tag{$C1'$}$$

$$m_{ij} + \theta \Delta t\, k_{ij} \leq 0, \quad i \neq j,\ i = 1, \ldots, N,\ j = 1, \ldots, \bar{N}, \tag{$C2'$}$$

$$m_{ii} - (1-\theta)\Delta t\, k_{ii} \geq 0, \quad i = 1, \ldots, N. \tag{$C3'$}$$

Proof. It is enough to show that $(C1^\star)$–$(C3^\star)$ follow from the conditions of the theorem. The relations $(C1')$ and $(C3')$ yield $(C3^\star)$, whereas $(C2^\star)$ follows from $(C2')$. Condition $(C1^\star)$ is valid if we prove that, under the assumptions of the theorem, $\mathbf{A}_0$ is irreducibly diagonally dominant with positive diagonal and nonpositive off-diagonal entries (cf. [19, Corollary 1, p. 85]). The sign conditions on the entries of $\mathbf{A}_0$ follow from $(C2')$ and the structure of $\mathbf{A}_0$. Further, we observe that $\mathbf{A}_0$ is irreducible, since its directed (oriented) graph is strongly connected (cf. [19, p. 20]), also, we see that $\mathbf{K}_0\mathbf{e}_0 = -\mathbf{K}_\partial \mathbf{e}_\partial$ as follows from $(P1)$, and

$$\mathbf{A}_0\mathbf{e}_0 = (\mathbf{M}_0 + \theta\Delta t\, \mathbf{K}_0)\mathbf{e}_0 = \mathbf{M}_0\mathbf{e}_0 + \theta\Delta t\, \mathbf{K}_0\mathbf{e}_0 = \mathbf{M}_0\mathbf{e}_0 - \theta\Delta t\, \mathbf{K}_\partial\mathbf{e}_\partial. \tag{59}$$

Obviously, $\mathbf{M}_0\mathbf{e}_0 > \mathbf{0}$, and due to $(C1')$, the vector $-\theta\Delta t\, \mathbf{K}_\partial\mathbf{e}_\partial$ is nonnegative. Hence, $\mathbf{A}_0$ is irreducibly diagonally dominant. ∎

3.4 The validity of DMP on tetrahedral meshes

3.4.1 The entries of mass and stiffness matrices

The contributions to the mass matrix $\mathbf{M}$ over the tetrahedron K are

$$m_{ij}|_K = \frac{\text{meas}_3\, K}{20}\ (i \neq j), \quad m_{ii}|_K = \frac{\text{meas}_3\, K}{10}. \tag{60}$$

The contribution to the stiffness matrix $\mathbf{K}$ over K (with vertices B_i, B_j, B_s, B_t) is equal to (cf. Section 2.5)

$$k_{ij}|_K = -c\,\frac{\text{meas}_2 B_i B_s B_t \cdot \text{meas}_2 B_j B_s B_t}{9\,\text{meas}_3\, K} \cos\alpha_{ij}^K,\ (i \neq j),\ k_{ii}|_K = c\,\frac{(\text{meas}_2 B_j B_s B_t)^2}{9\,\text{meas}_3 K}. \tag{61}$$

We assume also that the mesh $\mathcal{T}_h$ we deal with is regular, i.e., there exist positive constants $C_{2,min}, C_{2,max}, C_{3,min}, C_{3,max}$, independent of h, such that for any tetrahedral element K from the mesh and any face S_{ijk} of any element of this mesh, we have

$$C_{2,min} h^2 \leq \text{meas}_2 S_{ijk} \leq C_{2,max} h^2, \quad C_{3,min} h^3 \leq \text{meas}_3 K \leq C_{3,max} h^3. \tag{62}$$

3.4.2 Conditions on mesh and time-step

Lemma 3.6 *Let the employed tetrahedral mesh $\mathcal{T}_h$ be nonobtuse, then*

$$k_{ij} \leq 0,\ \text{for } i \neq j,\ i = 1, \ldots, N,\ j = 1, \ldots, \bar{N}.$$

PROOF. We denote $\text{supp}\,\phi_i \cap \text{supp}\,\phi_j$ by S, then for $k_{ij},\ i \neq j$, we have

$$k_{ij} = L(\phi_j, \phi_i) = c \int_\Omega \nabla\phi_j \cdot \nabla\phi_i\, dx = c \sum_{K \subseteq S} \int_K \nabla\phi_j \cdot \nabla\phi_i\, dx = \sum_{K \subseteq S} k_{ij}|_K \leq 0.$$

since the value $k_{ij}|_K$ is nonpositive for any nonobtuse tetrahedron K. ■

A system of acute tetrahedral meshes is called *acute*, if there exists a constant α_0 such that all dihedral angles of all tetrahedra are less than $\pi/2 - \alpha_0$.

Lemma 3.7 *Let the tetrahedral mesh $\mathcal{T}_h$ be acute. Then*

$$m_{ij} + \theta\Delta t\, k_{ij} \leq 0, \quad i \neq j,\ i = 1, \ldots, N,\ j = 1, \ldots, \bar{N},$$

provided

$$\Delta t \geq \frac{9\, C_{3,max}^2}{20\,\theta\, c \sin\alpha_0\, C_{2,min}^2} h^2.$$

PROOF. We denote $\operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j$ by S, then

$$m_{ij} + \theta \Delta t k_{ij} = \sum_{K \subseteq S} (m_{ij}|_K + \theta\, \Delta t\, k_{ij}|_K) \leq \sum_{K \subseteq S} \left(\frac{\operatorname{meas}_3 K}{20} - \theta\, \Delta t\, c \sin\alpha_0 \frac{C_{2,min}^2\, h}{9\, C_{3,max}} \right) \leq$$

$$\leq \sum_{K \subseteq S} \left(\frac{\operatorname{meas}_3 K}{20} - \theta\, \Delta t\, c \sin\alpha_0 \frac{C_{2,min}^2\, h}{9\, C_{3,max}} \frac{\operatorname{meas}_3 K}{\operatorname{meas}_3 K} \right) \leq$$

$$\leq \sum_{K \subseteq S} \left(\frac{\operatorname{meas}_3 K}{20} - \theta\, \Delta t\, c \sin\alpha_0 \frac{C_{2,min}^2\, h}{9\, C_{3,max}} \frac{\operatorname{meas}_3 K}{C_{3,max}\; h^3} \right) \leq$$

$$\leq \operatorname{meas}_3\, S \left(\frac{1}{20} - \theta\, \Delta t\, c \sin\alpha_0 \frac{C_{2,min}^2}{9\, h^2\, C_{3,max}^2} \right) \leq 0. \qquad \blacksquare$$

Lemma 3.8 *Let $\mathcal{T}_h$ be nonobtuse, then*

$$m_{ii} - (1-\theta)\Delta t\; k_{ii} \geq 0, \quad i = 1, \ldots, N,$$

provided

$$\Delta t \leq \frac{9\, C_{3,min}^2}{10\, (1-\theta)\, c\, C_{2,max}^2}\, h^2.$$

PROOF. We denote $\operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j$ by S, then

$$m_{ii} - (1-\theta)\Delta t k_{ii} = \sum_{K \subseteq S} (m_{ii}|_K - (1-\theta)\Delta t k_{ii}|_K) \geq \sum_{K \subseteq S} \left(\frac{\operatorname{meas}_3 K}{10} - (1-\theta)\Delta t c \frac{C_{2,max}^2 h^4}{9 C_{3,min} h^3} \right) \geq$$

$$\geq \sum_{K \subseteq S} \left(\frac{\operatorname{meas}_3 K}{10} - (1-\theta)\Delta t c \frac{C_{2,max}^2 h \operatorname{meas}_3 K}{9 C_{3,min} \operatorname{meas}_3 K} \right) \geq \operatorname{meas}_3 S \left(\frac{1}{10} - (1-\theta)\Delta t c \frac{C_{2,max}^2}{9 C_{3,min}^2 h^2} \right) \geq 0.$$

$\blacksquare$

Theorem 3.9 *The piecewise linear finite element solution of* (27)–(28) *on a tetrahedral mesh of acute type satisfies the discrete maximum principle* (48) *if the conditions*

$$\frac{9\, C_{3,min}^2}{10\, (1-\theta)\, c\, C_{2,max}^2}\, h^2 \geq \Delta t \geq \frac{9\, C_{3,max}^2}{20\, \theta\, c \sin\alpha_0\, C_{2,min}^2} h^2.$$

are fulfilled.

The proof immediately follows from Theorem 3.5 and Lemmas 3.6–3.8.

4 Acknowledgements:

J. Brandts was supported by Fellowship of the Royal Netherlands Academy of Arts and Sciences, S. Korotov – by Grant INBCT of TEKES, and M. Křížek – by Grant A1019201 of the Academy of Sciences of the Czech Republic.

REFERENCES

[1] I. Christie, C. Hall, *The Maximum Principle for Bilinear Elements*, Internat. J. Numer. Methods Engrg. 20 (1984), pp. 549–553.

[2] P. G. Ciarlet, *Discrete Maximum Principle for Finite Difference Operators*, Aequationes Math. 4 (1970), pp. 338–352.

[3] P. G. Ciarlet, P. A. Raviart, *Maximum Principle and Uniform Convergence for the Finite Element Method*, Comput. Methods Appl. Mech. Engrg. 2 (1973), pp. 17–31.

[4] A. Drăgănescu, T. F. Dupont, L. R. Scott, *Failure of the Discrete Maximum Principle for an Elliptic Finite Element Problem*, Math. Comp. (to appear), pp. 1–23.

[5] I. Faragó, R. Horváth, *On the Nonnegativity Conservation of Finite Element Solutions of Parabolic Problems*, in Proc. Conf. Finite Element Methods: Three-dimensional Problems, Univ. of Jyväskylä, GAKUTO Internat. Ser. Math. Sci. Appl., vol. 15, Gakkotosho, Tokyo 2001, pp. 76–84.

[6] I. Faragó, R. Horváth, S. Korotov, *Discrete Maximum Principle for Galerkin Finite Element Solutions to Parabolic Problems on Rectangular Meshes*, in Proc. of the Fifth European Conf. on Numerical Mathematics and Advanced Applications (ENUMATH 2003), Prague, Czech Republic (to appear), pp. 1–8.

[7] H. Fujii, *Some Remarks on Finite Element Analysis of Time-Dependent Field Problems*, Theory and Practice in Finite Element Structural Analysis, Univ. Tokyo Press, Tokyo (1973), pp. 91–106.

[8] R. Horváth, *On the Positivity of Iterative Methods*, Annales Univ. Budapest. Sect. Comp., Budapest. 19 (2000), pp. 93–102.

[9] J. Karátson, S. Korotov, *Discrete Maximum Principles for Finite Element Solutions of Nonlinear Elliptic Problems with Mixed Boundary Conditions*, Preprint 2003-1, Dept. of Appl. Analysis, Eötvös Loránd University, Hungary (submitted to Numer. Math. in 2003).

[10] S. Korotov, M. Křížek, P. Neittaanmäki, *Weakened Acute Type Condition for Tetrahedral Triangulations and the Discrete Maximum Principle*, Math. Comp. 70 (2001), pp. 107–119.

[11] M. Křížek, P. Neittaanmäki, *Finite Element Approximation of Variational Problems and Applications*, Longman Scientific & Technical, Harlow, 1990.

[12] M. Křížek, P. Neittaanmäki, *Mathematical and Numerical Modelling in Electrical Engineering: Theory and Applications*, Kluwer, Dordrecht, 1996.

[13] M. Křížek, Qun Lin, *On Diagonal Dominance of Stiffness Matrices in 3D*, East-West J. Numer. Math. 3 (1995), pp. 59–69.

[14] M. Křížek, L. Liu, *On the Maximum and Comparison Principles for a Steady-State Nonlinear Heat Conduction Problem*, Z. Angew. Math. Mech. 83 (2003), pp. 559–563.

[15] O. A. Ladyzenskaja, V. A. Solonnikov, N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Translations of Mathematical Monographs, Vol. 23, Amer. Math. Soc., Providence, R.I., 1968.

[16] O. A. Ladyzenskaja, N. N. Ural'ceva, *Linear and Quasilinear Equations of Elliptic Type*, Academic Press, New York, 1968.

[17] V. Ruas Santos, *On the Strong Maximum Principle for Some Piecewise Linear Finite Element Approximate Problems of Non-Positive Type*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. 29 (1982), pp. 473–491.

[18] G. Strang, G. J. Fix, *An Analysis of the Finite Element Method*, Prentice Hall, Englewood Cliffs, N. J., 1973.

[19] R. Varga, *Matrix Iterative Analysis*, Prentice Hall, New Jersey, 1962.

[20] T. Vejchodský, *On the Nonnegativity Conservation in Semidiscrete Parabolic Problems*, In: *Conjugate Gradient Algorithms and Finite Element Methods*, Springer-Verlag, Berlin, 2004, 197–210.