Helsinki University of Technology Institute of Mathematics Research Reports

Espoo 2008 A550

DISCRETE MAXIMUM PRINCIPLES FOR FE SOLUTIONS OF NONSTATIONARY DIFFUSION-REACTION PROBLEMS WITH MIXED BOUNDARY CONDITIONS

István Faragó Róbert Horváth Sergey Korotov

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Helsinki University of Technology Institute of Mathematics Research Reports
Espoo 2008 A550

DISCRETE MAXIMUM PRINCIPLES FOR FE SOLUTIONS OF NONSTATIONARY DIFFUSION-REACTION PROBLEMS WITH MIXED BOUNDARY CONDITIONS

István Faragó Róbert Horváth Sergey Korotov

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Mathematics and Systems Analysis

István Faragó, Róbert Horváth, Sergey Korotov : *Discrete maximum principles for FE solutions of nonstationary diffusion-reaction problems with mixed boundary conditions* ; Helsinki University of Technology Institute of Mathematics Research Reports A550 (2008).

Abstract: *In this paper we derive and discuss sufficient conditions that provide the validity of the discrete maximum principle for nonstationary diffusion-reaction problems with mixed boundary conditions solved by means of simplicial finite elements and the θ time discretization method. Theoretical analysis is supported by numerical experiments.*

AMS subject classifications: 65M60, 65M50, 35B50

Keywords: nonstationary diffusion-reaction problem, maximum principle, mixed boundary conditions, linear finite elements, discrete maximum principle, simplicial partition, angle condition

Correspondence

Department of Applied Analysis, Eötvös Loránd University
H–1518, Budapest, Pf. 120, Hungary

Institute of Mathematics and Statistics, University of West-Hungary
Erzsébet u. 9, H–9400, Sopron, Hungary

Institute of Mathematics, Helsinki University of Technology
P.O. Box 1100, FIN–02015 TKK, Finland

faragois@cs.elte.hu, rhorvath@ktk.nyme.hu, sergey.korotov@hut.fi

ISBN 978-951-22-9510-4 (print)
ISBN 978-951-22-9511-1 (PDF)
ISSN 0784-3143 (print)
ISSN 1797-5867 (PDF)

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Mathematics and Systems Analysis
P.O. Box 1100, FI-02015 TKK, Finland
email: math@tkk.fi http://math.tkk.fi/

1 Introduction

Besides the obligatory requirement of convergence of computed approximations to the exact solution of a model under investigation, the approximations are naturally required to mirror some basic qualitative properties of these exact solutions 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 (continuous) maximum principles (CMPs) [17, 18]. Its discrete analogues, the so-called discrete maximum principles (DMPs) were first presented and analysed in the papers [21, 4, 6], and [14, 11] for elliptic and parabolic cases, respectively. If the finite element method (FEM) is employed, the corresponding DMPs are normally ensured by imposing certain geometrical restrictions on the spatial meshes used: such as acuteness or nonobtuseness – for simplicial meshes [6, 11, 13, 16], non-narrowness - for rectangular meshes [3, 9, 12]. In [10], sufficient geometric conditions for DMPs are given for the case of planar hybrid meshes. The validity of DMPs for higher order finite elements and associated geometric restrictions on FE meshes are considered in [23]. In addition, in the case of DMPs for parabolic problems, the time-steps have to be often chosen inbetween certain lower and upper bounds. (For finite difference discretizations, there are however only upper bounds present for the time-step (see e.g., [8, 25]).) Another important discrete qualitative property of the numerical solutions of partial differential equations is the so-called nonnegativity preservation property. This property was investigated e.g. in [7, 8, 22]. The relation between the nonnegativity preservation and the DMPs were analysed e.g. in [8, 10].

In this work, for the first time with respect to parabolic problems and discrete maximum principles, the cases with the mixed boundary conditions and an additional reactive term presented in the governing equation are considered. We derive the relevant continuous maximum principle, and also give and prove its discrete analogue, when simplicial finite elements and the θ time discretization method are used. In addition, several numerical tests illustrating the sharpness of the proposed (sufficient) conditions for a validity of DMP are presented.

The paper is structured as follows. In Section 2, we describe the model parabolic problem with mixed boundary conditions and present the continuous maximum principle. The discretization scheme is given in detail in Section 3. The DMP and the algebraic conditions for its validity are presented in Section 4. In Section 5 we derive geometric conditions on simplicial meshes and two-sided bounds for the time-steps providing the DMP. Further, several illustrative numerical tests are given in Section 6.

2 Model problem and CMP

Consider the following nonstationary diffusion-reaction problem with mixed boundary conditions: Find a function $u = u(t, x)$ such that

$$\frac{\partial u}{\partial t} - b\Delta u + cu = f \quad \text{in} \quad Q_T := (0, T) \times \Omega, \tag{1}$$

$$u = g \quad \text{on} \quad S_T^D := (0, T) \times \partial\Omega_D, \tag{2}$$

$$b\nabla u \cdot \nu = q \quad \text{on} \quad S_T^N := (0, T) \times \partial\Omega_N, \tag{3}$$

$$b\nabla u \cdot \nu + \sigma u = r \quad \text{on} \quad S_T^R := (0, T) \times \partial\Omega_R, \tag{4}$$

$$u|_{t=0} = u^0 \quad \text{in} \quad \Omega, \tag{5}$$

where $\Omega \subset \mathbf{R}^d$, $d = 1, 2, 3, \ldots$, is a bounded polytopic domain with Lipschitz boundary $\partial\Omega$. Further, we assume that $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N \cup \partial\Omega_R$, where $\partial\Omega_D \neq \emptyset$ and is closed, $\partial\Omega_N$ and $\partial\Omega_R$ are mutually disjoint measurable open sets. The subscripts (or superscripts) D, N, and R always stand for Dirichlet, Neumann, and Robin types of boundary conditions, respectively, ν is the outward normal to $\partial\Omega$, $T > 0$, the problem coefficients are constant and such that

$$b > 0, \quad c \geq 0, \quad \sigma > 0, \tag{6}$$

and f, g, q, r, u^0 are given functions.

Let us introduce the following notations for any $t \in (0, T]$. Let Q_t stand for the cylinder $(0, t) \times \Omega$, and let $\Gamma_0 := \{0\} \times \Omega$ denote its bottom. Moreover, let us define $Q_{\overline{t}} := (0, t] \times \Omega$, $S_{\overline{t}}^D := [0, t] \times \partial\Omega_D$, $S_{\overline{t}}^N := [0, t] \times \partial\Omega_N$, and $S_{\overline{t}}^R := [0, t] \times \partial\Omega_R$.

Remark 1 *We assume that all the given functions are sufficiently smooth so that the classical solution of problem* (1)–(5) *exists in the space* $C^{1,2}(Q_{\overline{T}}) \cap C^{0,1}(Q_{\overline{T}} \cup S_{\overline{T}}^D \cup S_{\overline{T}}^N \cup S_{\overline{T}}^R \cup \Gamma_0)$ *and it is unique.*

First we shall derive the continuous maximum principle for the above type of problems. The following result holds (cf. Theorem 2.1 and Theorem 2.2 from [17]).

Theorem 1 *Let for problem* (1)–(5) *conditions* (6) *and* $q < 0$ *hold. Then the following upper estimate for the solution is valid for any* $t_1 \in (0, T)$*:*

$$u(t_1, x) \leq \inf_{\lambda > -c} \max\{0, \max_{\Gamma_0 \cup S_{\overline{t}_1}^D} u \exp(\lambda(t_1 - t)), \max_{Q_{\overline{t}_1}} \tfrac{f \exp(\lambda(t_1 - t))}{c + \lambda}, \max_{S_{\overline{t}_1}^R} \tfrac{r \exp(\lambda(t_1 - t))}{\sigma}\}. \tag{7}$$

P r o o f : Consider a function $v = v(t,x)$ defined as follows

$$v(t,x) = u(t,x)\exp(-\lambda t), \tag{8}$$

where λ is an arbitrary real number for the time being. This function obviously satisfies the equation

$$v_t - b\Delta v + (c+\lambda)v = f\exp(-\lambda t), \tag{9}$$

whenever u is the solution of (1)–(5). Let t_1 be arbitrary from $(0,T)$ and let us consider a closed cylinder $\overline{Q}_{t_1} := [0,t_1]\times\overline{\Omega}$. The following three cases are possible:

1. $v \le 0$ in $\overline{Q}_{t_1}$;
2. positive maximum of v is attained on $\Gamma_0 \cup S^D_{\bar{t}_1}$;
3. positive maximum of v is attained at some point (t^0,x^0) from $\Omega_{\bar{t}_1} \cup S^N_{\bar{t}_1} \cup S^R_{\bar{t}_1}$.

For the first situation we have

$$\max_{\overline{Q}_{t_1}} v(t,x) \le 0. \tag{10}$$

In the second case, it holds

$$0 < \max_{\overline{Q}_{t_1}} v(t,x) = \max_{\Gamma_0\cup S^D_{\bar{t}_1}} v. \tag{11}$$

Consider the third possibility when

$$0 < \max_{\overline{Q}_{t_1}} v(t,x) = v(t^0,x^0), \tag{12}$$

where $(t^0,x^0) \in Q_{\bar{t}_1} \cup S^N_{\bar{t}_1} \cup S^R_{\bar{t}_1}$. First, let $(t^0,x^0) \in Q_{\bar{t}_1}$. It is clear that in this situation we have

$$v_t(t^0,x^0) \ge 0, \quad \text{and} \quad v_{x_i,x_i}(t^0,x^0) \le 0, \quad (i = 1,\dots,d).$$

Then from (9) we get $(c+\lambda)v(t^0,x^0) \le f(t^0,x^0)\exp(-\lambda t^0)$, i.e., if the number λ is such that $c+\lambda > 0$, then

$$v(t^0,x^0) \le \max_{Q_{\bar{t}_1}} \frac{f\exp(-\lambda t)}{c+\lambda}. \tag{13}$$

Further, let $(t^0,x^0) \in S^R_{\bar{t}_1}$, then in a view of (4)

$$v(t^0,x^0) \le \max_{S^R_{\bar{t}_1}} \frac{r\,\exp(-\lambda t)}{\sigma}, \tag{14}$$

where the obvious property $\nabla v \cdot \nu|_{(t^0,x^0)} \ge 0$ was used. In addition, this property and condition $q|_{S^N_{\bar{t}_1}} < 0$ assure that the point (t^0,x^0) cannot belong to $S^N_{\bar{t}_1}$, which together with (8) proves the statement of the theorem. ■

Remark 2 *Theorem 1 implies, in particular, that* $u(t,x) \leq 0$ *provided* $f \leq 0$, $r \leq 0$, $u^0 \leq 0$, $g \leq 0$, *and* $q < 0$.

Let us further introduce the following function ($0 \leq t \leq t_1 < T$, t_1 is fixed):

$$\bar{u}(t,x) = u(t,x) - \max\{0; \max_{\Gamma_0 \cup S^D_{\bar{t}_1}} u\} - \max\{0; \max_{S^R_{\bar{t}_1}} \frac{r}{\sigma}\} - t \max\{0; \max_{Q_{\bar{t}_1}} f\}.$$

We immediately observe the non-positivity of the given data of the initial boundary-value problem to that the above defined function $\bar{v}$ satisfies. It is also clear that the function appearing in the RHS of the corresponding Neumann boundary condition is the same as for the function u, i.e., $q < 0$, which altogether, in a view of Remark 2, implies that

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

therefore for all $t_1 \in (0,T)$, we have

$$u(t_1,x) \leq \max\{0; \max_{\Gamma_0 \cup S^D_{\bar{t}_1}} u\} + \max\{0; \max_{S^R_{\bar{t}_1}} \frac{r}{\sigma}\} + t_1 \max\{0; \max_{Q_{\bar{t}_1}} f\}. \tag{15}$$

Inequality (15), together with the sign-condition $q < 0$ and (6), represents the form of the continuous maximum principle that we shall deal with for the above defined parabolic problem (1)–(5).

Remark 3 *In a similar way we can derive the relevant minimum principle under condition* $q > 0$.

3 Discretization scheme

Let

$$H^1_{\partial\Omega_D}(\Omega) = \{v \in H^1(\Omega) \mid v|_{\partial\Omega_D} = 0\}. \tag{16}$$

In what follows, we assume that a simplicial partition $\mathcal{T}_h$ of $\overline{\Omega}$ is given, where h denotes the standard discretization parameter (the maximal diameter of elements from $\mathcal{T}_h$), and that the partition is conforming and is such that any facet of any element is either a facet of the adjacent element or a part of the boundary. Let $B_1, \ldots, B_N$ denote all interior nodes and the nodes belonging to $\partial\Omega_N \cup \partial\Omega_R$, and let $B_{N+1}, \ldots, B_{\bar{N}}$ be the nodes lying on $\partial\Omega_D$. We also stand $N_\partial := \bar{N} - N$.

Let $\phi_1, \ldots, \phi_{\bar{N}}$ be the continuous piecewise linear nodal basis functions associated with nodes $B_1, \ldots, B_{\bar{N}}$, respectively. It is obvious that

$$\phi_i \geq 0, \quad i = 1, \ldots, \bar{N}, \qquad \text{and} \qquad \sum_{i=1}^{\bar{N}} \phi_i \equiv 1 \text{ in } \overline{\Omega}. \tag{17}$$

We denote the span of the basis functions by $V^h \subset H^1(\Omega)$, and define its subspace

$$V^h_{\partial\Omega_D} = \{v \in V^h \mid v|_{\partial\Omega_D} = 0\} \subset H^1_{\partial\Omega_D}(\Omega).$$

3.1 Weak formulation

The weak formulation for (1)–(5) reads as follows: Find $u = u(t,x) \in H^1(\Omega)$ for $t \in (0,T)$ such that

$$\int_\Omega \frac{\partial u}{\partial t} v \, dx + L(u,v) = \int_\Omega f v \, dx + \int_{\partial\Omega_N} q v \, ds + \int_{\partial\Omega_R} r v \, ds \qquad \forall v \in H^1_{\partial\Omega_D}(\Omega), \quad t \in (0,T), \tag{18}$$

and

$$u(0,x) = u^0(x), \; x \in \Omega, \quad \text{and} \quad u - g \in H^1_{\partial\Omega_D}(\Omega), \; t \in (0,T), \tag{19}$$

where

$$L(u,v) := b \int_\Omega \nabla u \cdot \nabla v \; dx + c \int_\Omega u \, v \; dx + \sigma \int_{\partial\Omega_R} u \, v \; ds. \tag{20}$$

3.2 Semidiscretization in space

The semidiscrete problem for (18)–(19) reads: Find a function $u_h = u_h(t,x)$ such that

$$\int_\Omega \frac{\partial u_h}{\partial t} v_h \, dx + L(u_h, v_h) = \int_\Omega f v_h \, dx + \int_{\partial\Omega_N} q v_h \, ds + \int_{\partial\Omega_R} r v_h \, ds \quad \forall v_h \in V^h_{\partial\Omega_D}, \; t \in (0,T), \tag{21}$$

and

$$u_h(0,x) = u^0_h(x), \; x \in \Omega, \qquad u_h(t,x) - g_h(t,x) \in V^h_{\partial\Omega_D}, \; t \in (0,T), \tag{22}$$

where $u^0_h(x)$ 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.,

$$u^0_h(x) = \sum_{i=1}^{\bar{N}} u^0(B_i)\phi_i(x), \tag{23}$$

and

$$g_h(t,x) = \sum_{i=1}^{N_\partial} g^h_i(t)\phi_{N+i}(x), \quad \text{where} \quad g^h_i(t) = g(t, B_{N+i}), \; i = 1, \ldots, N_\partial. \tag{24}$$

From the consistency of the initial and the boundary conditions $g(0,s) = u^0(s), \; s \in \partial\Omega_D$, we have $g^h_i(0) = u^0(B_{N+i}), \; i = 1, \ldots, N_\partial$.

We search for a semidiscrete solution of the form

$$u_h(t,x) = \sum_{j=1}^{N} u^h_j(t)\phi_j(x) + g_h(t,x) = \sum_{j=1}^{N} u^h_j(t)\phi_j(x) + \sum_{j=N+1}^{\bar{N}} g^h_{j-N}(t)\phi_j(x), \tag{25}$$

and notice that it is sufficient that u_h satisfies (21) only for $v_h = \phi_i$, $i = 1, \ldots, N$.

Introducing the notation

$$\mathbf{v}^h(t) = [u_1^h(t), \ldots, u_N^h(t), g_1^h(t), \ldots, g_{N_\partial}^h(t)]^T, \tag{26}$$

we get a Cauchy problem for the systems of ordinary differential equations

$$\mathbf{M}\frac{\mathrm{d}\mathbf{v}^h}{\mathrm{d}t} + \mathbf{K}\mathbf{v}^h = \mathbf{f} + \mathbf{q} + \mathbf{r}, \quad \mathbf{v}^h(0) = [u^0(B_1), \ldots, u^0(B_N), g_1^h(0), \ldots, g_{N_\partial}^h(0)]^T \tag{27}$$

for the solution of the semidiscrete problem, where

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

$$\mathbf{f} = [f_1, \ldots, f_N]^T, \quad f_i = \int\limits_{\Omega} f \phi_i \, dx,$$

$$\mathbf{q} = [q_1, \ldots, q_N]^T, \quad q_i = \int\limits_{\partial\Omega_N} q \phi_i \, ds,$$

and

$$\mathbf{r} = [r_1, \ldots, r_N]^T, \quad r_i = \int\limits_{\partial\Omega_R} r \phi_i \, ds.$$

3.3 Fully discretized problem

In order to get a fully discrete numerical scheme, we choose a time-step Δt and denote the approximations to $\mathbf{v}^h(n\Delta t)$, $\mathbf{f}(n\Delta t)$, $\mathbf{q}(n\Delta t)$, and $\mathbf{r}(n\Delta t)$ by $\mathbf{v}^n$, $\mathbf{f}^n$, $\mathbf{q}^n$, and $\mathbf{r}^n$, $n = 0, 1, \ldots, n_T$ ($n_T \Delta t = T$), respectively.

To discretize (27), we apply the θ-method ($\theta \in (0, 1]$ is a given parameter, the case $\theta = 0$, otherwise acceptable, is excluded in what follows due to the form of DMP and the condition of the lower bound for the time-step, see Sections 4 and 5) and obtain a system of linear algebraic equations

$$\mathbf{M}\frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{\Delta t} + \theta\mathbf{K}\mathbf{v}^{n+1} + (1-\theta)\mathbf{K}\mathbf{v}^n = \mathbf{f}^{(n,\theta)} + \mathbf{q}^{(n,\theta)} + \mathbf{r}^{(n,\theta)}, \tag{28}$$

where $\mathbf{f}^{(n,\theta)} := \theta\, \mathbf{f}^{n+1} + (1-\theta)\mathbf{f}^n$, $\mathbf{q}^{(n,\theta)} := \theta\, \mathbf{q}^{n+1} + (1-\theta)\mathbf{q}^n$, and $\mathbf{r}^{(n,\theta)} := \theta\, \mathbf{r}^{n+1} + (1-\theta)\mathbf{r}^n$.

Further, (28) can be rewritten as

$$(\mathbf{M} + \theta\Delta t\mathbf{K})\mathbf{v}^{n+1} = (\mathbf{M} - (1-\theta)\Delta t\mathbf{K})\mathbf{v}^n + \Delta t\, \mathbf{f}^{(n,\theta)} + \Delta t\, \mathbf{q}^{(n,\theta)} + \Delta t\, \mathbf{r}^{(n,\theta)}, \tag{29}$$

where $n = 0, 1, \ldots, n_T - 1$, and $\mathbf{v}^0 = \mathbf{v}^h(0)$.

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:

$$\mathbf{A} = [\mathbf{A}_0 | \mathbf{A}_\partial], \quad \mathbf{B} = [\mathbf{B}_0 | \mathbf{B}_\partial], \quad \mathbf{v}^n = [(\mathbf{u}^n)^T | (\mathbf{g}^n)^T]^T, \tag{30}$$

where $\mathbf{A}_0$ and $\mathbf{B}_0$ are $(N \times N)$ matrices, $\mathbf{A}_\partial, \mathbf{B}_\partial$ are of size $(N \times N_\partial)$, $\mathbf{u}^n = [u_1^n, \dots, u_N^n]^T \in \mathbf{R}^N$ and $\mathbf{g}^n = [g_1^n, \dots, g_{N_\partial}^n]^T \in \mathbf{R}^{N_\partial}$. Similar partitions are used for matrices $\mathbf{M}$ and $\mathbf{K}$. The iterative scheme (29) can now be rewritten as follows

$$\mathbf{A}\mathbf{v}^{n+1} = \mathbf{B}\mathbf{v}^n + \Delta t\, \mathbf{f}^{(n,\theta)} + \Delta t\, \mathbf{q}^{(n,\theta)} + \Delta t\, \mathbf{r}^{(n,\theta)}, \tag{31}$$

or

$$[\mathbf{A}_0|\mathbf{A}_\partial] \begin{bmatrix} \mathbf{u}^{n+1} \\ \mathbf{g}^{n+1} \end{bmatrix} = [\mathbf{B}_0|\mathbf{B}_\partial] \begin{bmatrix} \mathbf{u}^{n} \\ \mathbf{g}^{n} \end{bmatrix} + \Delta t\, \mathbf{f}^{(n,\theta)} + \Delta t\, \mathbf{q}^{(n,\theta)} + \Delta t\, \mathbf{r}^{(n,\theta)}. \tag{32}$$

4 The discrete maximum principle

4.1 Formulation of DMP

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

$$g_{max}^n = \max\{0, g_1^n, \dots, g_{N_\partial}^n\}, \tag{33}$$

$$v_{max}^n = \max\{0, g_{max}^n, u_1^n, \dots, u_N^n\}, \tag{34}$$

and

$$f_{max}^{(n,n+1)} = \max\{0, \max_{x\in\Omega, \tau\in(n\Delta t, (n+1)\Delta t)} f(\tau, x)\}, \tag{35}$$

$$r_{max}^{(n,n+1)} = \max\{0, \max_{x\in\partial\Omega_R, \tau\in(n\Delta t, (n+1)\Delta t)} r(\tau, x)\}, \tag{36}$$

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

The DMP corresponding to (15) (under condition $q < 0$) takes the following form (cf. [11, p. 100]):

$$u_i^{n+1} \le \max\{0, g_{max}^{n+1}, v_{max}^n\} + \frac{1}{\theta\sigma} r_{max}^{(n,n+1)} + \Delta t f_{max}^{(n,n+1)}, \tag{37}$$

for $i = 1, \dots, N;\ \ n = 0, \dots, n_T - 1$.

Remark 4 *The maximum principle expresses the fact that the solution can be estimated from above using the solution at earlier time instants, the source function and the functions that are present in the boundary conditions. The definition of the DMP in* (37) *fulfills this requirement. However, we notice that, in* (37)*, the second term is multiplied by the reciprocal value of θ. This means that this estimation is somewhat weaker than the corresponding estimation in the continuous case unless $\theta = 1$.*

4.2 Algebraic conditions guaranteeing the validity of DMP

Write

$$\mathbf{e} = [1, \dots, 1]^T \in \mathbf{R}^{\bar{N}}, \quad \mathbf{e}_0 = [1, \dots, 1]^T \in \mathbf{R}^{N}, \quad \mathbf{e}_\partial = [1, \dots, 1]^T \in \mathbf{R}^{N_\partial}, \tag{38}$$

$$\begin{aligned}
&\mathbf{f}_{max}^{(n,n+1)} = f_{max}^{(n,n+1)}\mathbf{e} \in \mathbf{R}^{\bar{N}}, && \mathbf{r}_{max}^{(n,n+1)} = r_{max}^{(n,n+1)}\mathbf{e} \in \mathbf{R}^{\bar{N}}, && \mathbf{v}_{max}^{n} = v_{max}^{n}\mathbf{e} \in \mathbf{R}^{\bar{N}},\\
&\mathbf{f}_{0}^{(n,n+1)} = f_{max}^{(n,n+1)}\mathbf{e}_0 \in \mathbf{R}^{N}, && \mathbf{r}_{0}^{(n,n+1)} = r_{max}^{(n,n+1)}\mathbf{e}_0 \in \mathbf{R}^{N}, && \mathbf{v}_{0}^{n} = v_{max}^{n}\mathbf{e}_0 \in \mathbf{R}^{N},\\
&\mathbf{f}_{\partial}^{(n,n+1)} = f_{max}^{(n,n+1)}\mathbf{e}_\partial \in \mathbf{R}^{N_\partial}, && \mathbf{r}_{\partial}^{(n,n+1)} = r_{max}^{(n,n+1)}\mathbf{e}_\partial \in \mathbf{R}^{N_\partial}, && \mathbf{v}_{\partial}^{n} = v_{max}^{n}\mathbf{e}_\partial \in \mathbf{R}^{N_\partial}.
\end{aligned} \tag{39}$$

Lemma 1 *The following relations hold provided $q < 0$ (and that $\theta \neq 0$)*

$$\begin{aligned}
&(P1) \quad \mathbf{Ke} \geq \mathbf{0},\\
&(P2) \quad \mathbf{f}^{(n,\theta)} \leq \mathbf{A}\mathbf{f}_{max}^{(n,n+1)},\\
&(P2') \quad \mathbf{q}^{(n,\theta)} \leq \mathbf{0},\\
&(P2'') \quad \mathbf{r}^{(n,\theta)} \leq \tfrac{1}{\sigma\theta\Delta t}\mathbf{A}\mathbf{r}_{max}^{(n,n+1)},\\
&(P3) \quad \text{If } \mathbf{A}_0^{-1} \geq \mathbf{0}, \text{ then } -\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{e}_\partial \leq \mathbf{e}_0.
\end{aligned} \tag{40}$$

P r o o f : (P1) For the i-th coordinate of the vector $\mathbf{Ke}$ ($i = 1, \dots, N$), we have

$$(\mathbf{Ke})_i = \sum_{j=1}^{\bar{N}} k_{ij} = \sum_{j=1}^{\bar{N}} L(\phi_j, \phi_i) = L\left(\sum_{j=1}^{\bar{N}} \phi_j, \phi_i\right) = L(1, \phi_i) =$$

$$= b\int_\Omega \nabla 1 \cdot \nabla\phi_i\, dx + c\int_\Omega 1 \cdot \phi_i\, dx + \sigma \int_{\partial\Omega_R} 1 \cdot \phi_i\, ds \geq 0, \tag{41}$$

due to conditions (6) on the coefficients of the problem and properties (17) of the basis functions, which proves the statement.

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

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

$$\leq \int_\Omega f_{max}^{(n,n+1)}\phi_i(x)\, dx = f_{max}^{(n,n+1)}\int_\Omega \left(\sum_{j=1}^{\bar{N}} \phi_j(x)\right)\phi_i(x)\, dx = \tag{42}$$

$$= f_{max}^{(n,n+1)}\sum_{j=1}^{\bar{N}} m_{ij} = \left(\mathbf{M}\mathbf{f}_{max}^{(n,n+1)}\right)_i \leq \left((\mathbf{M} + \theta\Delta t\mathbf{K})\mathbf{f}_{max}^{(n,n+1)}\right)_i = \left(\mathbf{A}\mathbf{f}_{max}^{(n,n+1)}\right)_i,$$

where in the above, we used (P1).

$(P2')$ is trivial due to the condition $q < 0$.

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

$$(\mathbf{r}^{(n,\theta)})_i = \int\limits_{\partial\Omega_R} \Big((1-\theta)r(n\Delta t, x) + \theta r((n+1)\Delta t, x)\Big)\phi_i(x)\, ds \leq$$

$$\leq \int\limits_{\partial\Omega_R} r_{max}^{(n,n+1)}\phi_i(x)\, ds = \frac{r_{max}^{(n,n+1)}}{\theta\,\Delta t\,\sigma}\theta\Delta t\sigma \int\limits_{\partial\Omega_R} \left(\sum_{j=1}^{\bar{N}}\phi_j(x)\right)\phi_i(x)\, ds \leq \tag{43}$$

$$\leq \frac{1}{\sigma\,\theta\,\Delta t}\left(\theta\,\Delta t\,\mathbf{K}\mathbf{r}_{max}^{(n,n+1)}\right)_i \leq \frac{1}{\sigma\,\theta\,\Delta t}\left((\mathbf{M}+\theta\Delta t\mathbf{K})\mathbf{r}_{max}^{(n,n+1)}\right)_i = \frac{1}{\sigma\,\theta\,\Delta t}\left(\mathbf{A}\mathbf{r}_{max}^{(n,n+1)}\right)_i,$$

where in the above, we used the nonnegativity of the mass matrix $\mathbf{M}$ provided by the fact that the basis functions are nonnegative.

(P3) Matrix $\mathbf{M}$ is non-negative, thus, $\mathbf{0} \leq \mathbf{M}\mathbf{e} \leq (\mathbf{M}+\theta\Delta t\mathbf{K})\mathbf{e} = \mathbf{A}\mathbf{e} = \mathbf{A}_0\mathbf{e}_0 + \mathbf{A}_\partial\mathbf{e}_\partial$, and $(P3)$ is obtained by multiplying both sides by the non-negative matrix $\mathbf{A}_0^{-1}$. $\mathbf{A}_0$ is always regular because it is a Gram-matrix. ■

Theorem 2 *Galerkin approximation for the solution of problem* (1)–(5), *combined with the θ-method for time discretization (where $\theta \in (0,1]$), satisfies the discrete maximum principle* (37) *under condition $q < 0$ if*

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

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

$$\mathbf{A}_0^{-1}\mathbf{B} \geq \mathbf{0}. \tag{C3}$$

P r o o f :

From (31), (P2) and $(P2')$, we have

$$\mathbf{A}_0\mathbf{u}^{n+1} + \mathbf{A}_\partial\mathbf{g}^{n+1} = \mathbf{A}\mathbf{v}^{n+1} = \mathbf{B}\mathbf{v}^n + \Delta t\ \mathbf{f}^{(n,\theta)} + \Delta t\ \mathbf{q}^{(n,\theta)} + \Delta t\ \mathbf{r}^{(n,\theta)} \leq \tag{44}$$

$$\leq \mathbf{B}\mathbf{v}^n + \Delta t\mathbf{A}\mathbf{f}_{max}^{(n,n+1)} + \frac{1}{\theta\sigma}\mathbf{A}\mathbf{r}_{max}^{(n,n+1)}.$$

From (P1), we find $\mathbf{B}\mathbf{v}_{max}^n \leq \mathbf{A}\mathbf{v}_{max}^n$. Multiplying both sides of (44) by $\mathbf{A}_0^{-1} \geq \mathbf{0}$ (see (C1)), expressing $\mathbf{u}^{n+1}$ and using (C3), we obtain

$$\mathbf{u}^{n+1} \leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial\ \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{B}\mathbf{v}^n + \Delta t\mathbf{A}_0^{-1}\mathbf{A}\mathbf{f}_{max}^{(n,n+1)} + \frac{1}{\theta\sigma}\mathbf{A}_0^{-1}\mathbf{A}\mathbf{r}_{max}^{(n,n+1)} \leq \tag{45}$$

$$\leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial\ \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{B}\mathbf{v}_{max}^n + \Delta t\mathbf{A}_0^{-1}\mathbf{A}\mathbf{f}_{max}^{(n,n+1)} + \frac{1}{\theta\sigma}\mathbf{A}_0^{-1}\mathbf{A}\mathbf{r}_{max}^{(n,n+1)} \leq$$

$$\leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{A}\mathbf{v}^n_{max} + \Delta t\mathbf{A}_0^{-1}\mathbf{A}\mathbf{f}^{(n,n+1)}_{max} + \frac{1}{\theta\sigma}\mathbf{A}_0^{-1}\mathbf{A}\mathbf{r}^{(n,n+1)}_{max} =$$

$$= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{g}^{n+1} + \mathbf{A}_0^{-1}[\mathbf{A}_0|\ \mathbf{A}_\partial]\mathbf{v}^n_{max} + \Delta t\mathbf{A}_0^{-1}[\mathbf{A}_0|\ \mathbf{A}_\partial]\ \mathbf{f}^{(n,n+1)}_{max} + \frac{1}{\theta\sigma}\mathbf{A}_0^{-1}[\mathbf{A}_0|\ \mathbf{A}_\partial]\ \mathbf{r}^{(n,n+1)}_{max} =$$

$$= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{g}^{n+1} + \mathbf{v}^n_0 + \mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{v}^n_\partial + \Delta t\mathbf{f}^{(n,n+1)}_0 +$$

$$+\Delta t\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{f}^{(n,n+1)}_\partial + \frac{1}{\theta\sigma}\mathbf{r}^{(n,n+1)}_0 + \frac{1}{\theta\sigma}\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{r}^{(n,n+1)}_\partial.$$

Regrouping the above inequality, we get

$$\mathbf{u}^{n+1} - \mathbf{v}^n_0 - \Delta t\mathbf{f}^{(n,n+1)}_0 - \frac{1}{\theta\sigma}\mathbf{r}^{(n,n+1)}_0 \leq -\mathbf{A}_0^{-1}\mathbf{A}_\partial(\mathbf{g}^{n+1} - \mathbf{v}^n_\partial - \Delta t\mathbf{f}^{(n,n+1)}_\partial - \frac{1}{\theta\sigma}\mathbf{r}^{(n,n+1)}_\partial). \quad (46)$$

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

$$u_i^{n+1} - v^n_{max} - \Delta t\ f^{(n,n+1)}_{max} - \frac{1}{\theta\sigma}r^{(n,n+1)}_{max} \leq$$

$$\leq \sum_{j=1}^{N_\partial} \left(-\mathbf{A}_0^{-1}\mathbf{A}_\partial\right)_{ij} (g_j^{n+1} - v^n_{max} - \Delta t\ f^{(n,n+1)}_{max} - \frac{1}{\theta\sigma}r^{(n,n+1)}_{max}) \leq \quad (47)$$

$$\leq \left(\sum_{j=1}^{N_\partial} \left(-\mathbf{A}_0^{-1}\mathbf{A}_\partial\right)_{ij}\right) \cdot \max\{0, \max_j\{g_j^{n+1} - v^n_{max}\}\} \leq$$

$$\leq \max\{0, \max_j\{g_j^{n+1} - v^n_{max}\}\} \leq \max\{0, g^{n+1}_{max}\},$$

where we applied (C2) and (P3). Finally, isolating u_i^{n+1}, we obtain the required inequality. ■

Remark 5 *Conditions* $(C1)$–$(C3)$ *are ensured by the following simpler assumptions*

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

$$\mathbf{A}_\partial \leq \mathbf{0}, \quad (C2^\star)$$

$$\mathbf{B} \geq \mathbf{0}. \quad (C3^\star)$$

Theorem 3 *Galerkin approximation for the solution of problem* (1)–(5), *combined with the θ-method for time discretization (where $\theta \in (0,1]$), satisfies the discrete maximum principle* (37) *if*

$$k_{ij} \leq 0, \quad i = 1,\dots,N,\ j = 1,\dots,\bar{N},\ i \neq j, \quad (C1')$$

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

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

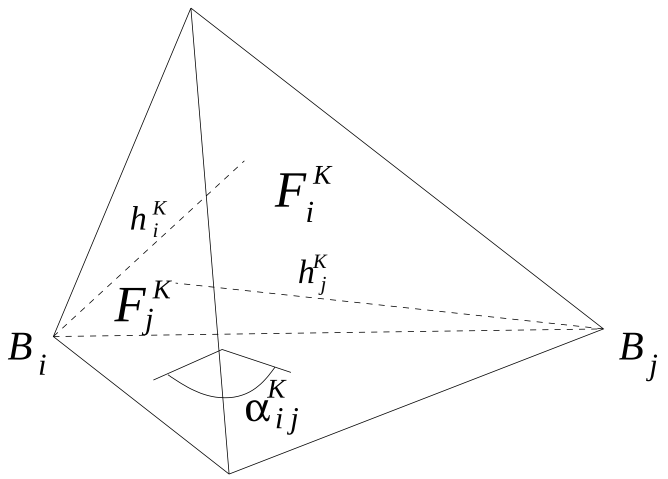

Figure 1: Tetrahedron K with denotations.

P r o o f : 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 in a view of [20, Corollary 3, p. 85]), because $\mathbf{A}_0$ is, obviously, positive definite and symmetric matrix with nonpositive off-diagonal entries provided by $(C2')$ (i.e., $\mathbf{A}_0$ is Stieltjes matrix). ■

Remark 6 *In a view of properties* (17) *of the basis functions and the assumption $\theta > 0$, condition ($C2'$), in fact, implies ($C1'$).*

5 The validity of DMP on simplicial meshes

5.1 The entries of matrices M and K

From now on, we denote a simplex from $\mathcal{T}_h$ by the symbol K and also use denotation α_{ij}^K for the angle between $(d-1)$-dimensional facets F_i^K and F_j^K of K which is opposite to the edge connecting vertices B_i and B_j, and let $h_i^K (h_j^K)$ be the height of K from $B_i (B_j)$ onto $F_i^K (F_j^K)$, see Figure 1.

The contributions to the mass matrix $\mathbf{M}$ over the simplex K are (cf. [5, p. 201])

$$m_{ij}|_K = \int_K \phi_i \, \phi_j \, dx = (1 + \delta_{ij}) \frac{d!}{(d+2)!} \mathrm{meas}_d \, K, \tag{48}$$

where δ_{ij} is Kronecker's symbol.

In order to compute the entries of the matrix $\mathbf{K}$, we shall use the following formulae presented e.g. in [1, 24]:

$$\nabla \phi_i \cdot \nabla \phi_j|_K = -\frac{\mathrm{meas}_{d-1} F_i^K \cdot \mathrm{meas}_{d-1} F_j^K}{(d \, \mathrm{meas}_d \, K)^2} \cos \alpha_{ij}^K = -\frac{\cos \alpha_{ij}^K}{h_i^K \, h_j^K} \quad (i \neq j), \tag{49}$$

$$\nabla\phi_i \cdot \nabla\phi_i|_K = \frac{(\mathrm{meas}_{d-1} F_i^K)^2}{(d\,\mathrm{meas}_d\, K)^2} = \frac{1}{(h_i^K)^2}. \tag{50}$$

5.2 Conditions on simplicial partitions and time-step

Lemma 2 *Let the simplicial partition $\mathcal{T}_h$ of $\overline{\Omega}$ be such that for any pair of distinct $(d-1)$-dimensional facets F_i^K and F_j^K of any simplex K from $\mathcal{T}_h$, we have*

$$\frac{\cos\alpha_{ij}^K}{h_i^K\, h_j^K} \geq \frac{c}{b\,(d+1)(d+2)} + \frac{\sigma}{b\,d(d+1)} \frac{\mathrm{meas}_{d-1}\,(\partial K \cap \partial\Omega_R)}{\mathrm{meas}_d\, K}, \tag{51}$$

where ∂K is the boundary of K. Then

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

P r o o f : Let $S_{ij} := \mathrm{supp}\,\phi_i \cap \mathrm{supp}\,\phi_j$. Then for $k_{ij},\ i \neq j$, we have

$$k_{ij} = L(\phi_j, \phi_i) = b \int\limits_{\Omega} \nabla\phi_j \cdot \nabla\phi_i \, dx + c \int\limits_{\Omega} \phi_j\phi_i \, dx + \sigma \int\limits_{\partial\Omega_R} \phi_j\phi_i \, ds =$$

$$= b \sum_{K \subseteq S_{ij}} \int\limits_K \nabla\phi_j \cdot \nabla\phi_i \, dx + c \sum_{K \subseteq S_{ij}} \int\limits_K \phi_j\phi_i \, dx + \sigma \sum_{K \subseteq S_{ij}} \int\limits_{\partial K \cap \partial\Omega_R} \phi_j\phi_i \, ds =$$

$$= \sum_{K \subseteq S_{ij}} \Big(b \int\limits_K \nabla\phi_j \cdot \nabla\phi_i \, dx + c \int\limits_K \phi_j\phi_i \, dx + \sigma \int\limits_{\partial K \cap \partial\Omega_R} \phi_j\phi_i \, ds \Big) =$$

$$= \sum_{K \subseteq S_{ij}} \Big(-\frac{b\,\cos\alpha_{ij}^K}{h_i^K\, h_j^K}\,\mathrm{meas}_d\, K + \frac{c}{(d+1)(d+2)} \mathrm{meas}_d\, K + \frac{\sigma}{d(d+1)} \mathrm{meas}_{d-1}\,(\partial K \cap \partial\Omega_R) \Big).$$

In a view of (51) we immediately prove the lemma. ■

Remark 7 *For the case $c \equiv 0$ and when the Robin boundary condition is absent, the above requirement on the mesh is nothing else, but just the nonobtuseness property, which is only related to the shape of elements. However, in the general case we need more stringent (strict acuteness) condition on the shape of simplicial elements, and also the restriction on the size (value of h) of the mesh.*

Lemma 3 *Let the simplicial partition $\mathcal{T}_h$ of $\overline{\Omega}$ and the time-step Δt be such that for any pair of distinct $(d-1)$-dimensional facets F_i^K and F_j^K of any simplex K from $\mathcal{T}_h$, we have*

$$\frac{\cos\alpha_{ij}^K}{h_i^K\, h_j^K} \geq \frac{c + 1/(\theta\,\Delta t)}{b\,(d+1)(d+2)} + \frac{\sigma}{b\,d(d+1)} \frac{\mathrm{meas}_{d-1}\,(\partial K \cap \partial\Omega_R)}{\mathrm{meas}_d\, K}. \tag{52}$$

where ∂K is the boundary of K. Then

$$m_{ij} + \theta\Delta t\, k_{ij} \le 0, \text{ for } \quad i = 1,\dots,N,\ j = 1,\dots,\bar{N},\ i \ne j.$$

P r o o f : We observe that

$$m_{ij}+\theta\Delta t k_{ij} = \theta\,\Delta t\, b \int\limits_{\Omega} \nabla\phi_j\cdot\nabla\phi_i\, dx + (1+\theta\,\Delta t\, c)\int\limits_{\Omega} \phi_j\phi_i\, dx + \theta\,\Delta t\,\sigma \int\limits_{\partial\Omega_R} \phi_j\phi_i\, ds =$$

$$= \theta\,\Delta t\, b \sum_{K\subseteq S_{ij}} \int\limits_{K} \nabla\phi_j\cdot\nabla\phi_i\, dx + (1+\theta\,\Delta t\, c)\sum_{K\subseteq S_{ij}}\int\limits_{K} \phi_j\phi_i\, dx + \theta\,\Delta t\,\sigma \sum_{K\subseteq S_{ij}} \int\limits_{\partial K\cap\partial\Omega_R} \phi_j\phi_i\, ds =$$

$$= \sum_{K\subseteq S_{ij}} \Big(\theta\,\Delta t\, b \int\limits_{K} \nabla\phi_j\cdot\nabla\phi_i\, dx + (1+\theta\,\Delta t\, c)\int\limits_{K} \phi_j\phi_i\, dx + \theta\,\Delta t\,\sigma \int\limits_{\partial K\cap\partial\Omega_R} \phi_j\phi_i\, ds\Big) =$$

$$= \sum_{K\subseteq S_{ij}} \Big(-\theta\,\Delta t\, b\, \frac{\cos\alpha_{ij}^K}{h_i^K\, h_j^K}\, \mathrm{meas}_d\, K + \frac{1+\theta\,\Delta t\, c}{(d+1)(d+2)} \mathrm{meas}_d\, K + \sigma\, \frac{\theta\,\Delta t}{d(d+1)} \mathrm{meas}_{d-1}\,(\partial K\cap\partial\Omega_R)\Big).$$

In a view of (52) we immediately prove the lemma. ■

Lemma 4 *Let the simplicial partition $\mathcal{T}_h$ of $\overline{\Omega}$ and the time-step Δt be such that for any simplex K from $\mathcal{T}_h$, we have*

$$0 \le -\frac{1}{(h_i^K)^2} + \frac{2\Big(\frac{1}{(1-\theta)\Delta t} - c\Big)}{b(d+1)(d+2)} - \frac{2\sigma}{bd(d+1)} \frac{\mathrm{meas}_{d-1}\,(\partial K\cap\partial\Omega_R)}{\mathrm{meas}_d\, K}, \qquad (53)$$

where ∂K is the boundary of K. Then

$$m_{ii} - (1-\theta)\Delta t\, k_{ii} \ge 0, \quad i = 1,\dots,N.$$

P r o o f : Let $S_{ii} := \mathrm{supp}\,\phi_i$. We get

$$m_{ii} - (1-\theta)\Delta t\, k_{ii} =$$

$$= -(1-\theta)\,\Delta t\, b \int\limits_{\Omega} \nabla\phi_i\cdot\nabla\phi_i\, dx + (1-(1-\theta)\,\Delta t\, c)\int\limits_{\Omega} \phi_i\phi_i\, dx - (1-\theta)\,\Delta t\,\sigma \int\limits_{\partial\Omega_R} \phi_i\phi_i\, ds =$$

$$= \sum_{K\subseteq S_{ii}} \Big(-(1-\theta)\Delta t\, b \int\limits_{K} \nabla\phi_i\cdot\nabla\phi_i\, dx + (1-(1-\theta)\Delta t\, c)\int\limits_{K} \phi_i\phi_i\, dx - (1-\theta)\Delta t\sigma \int\limits_{\partial K\cap\partial\Omega_R} \phi_i\phi_i\, ds\Big) =$$

$$= \sum_{K\subseteq S_{ii}} \Big(\frac{-(1-\theta)\,\Delta t\, b}{(h_i^K)^2} \mathrm{meas}_d\, K + \frac{2(1-(1-\theta)\Delta t\, c)}{(d+1)(d+2)} \mathrm{meas}_d\, K - \frac{2(1-\theta)\sigma\,\Delta t}{d(d+1)} \mathrm{meas}_{d-1}\,(\partial K\cap\partial\Omega_R)\Big).$$

In a view of (53) we immediately prove the lemma. ■

Summarizing the above results we can formulate the main

Theorem 4 *Galerkin approximation for the solution of problem* (1)–(5)*, combined with the θ-method for time discretization (where $\theta \in (0, 1]$), satisfies the discrete maximum principle* (37) *if an acute simplicial mesh is used and the time-step satisfies the following (lower and upper) estimates:*

$$\frac{1}{\Delta t} \le \theta\Big(\frac{\cos\alpha_{ij}^K b(d+1)(d+2)}{h_i^K\, h_j^K} - \frac{\sigma(d+2)\mathrm{meas}_{d-1}\,(\partial K \cap \partial\Omega_R)}{d\,\mathrm{meas}_d\, K} - c\Big), \quad (54)$$

and

$$\Delta t \le \frac{1}{(1-\theta)\Big(\frac{b(d+1)(d+2)}{2(h_i^K)^2} + \frac{\sigma(d+2)\mathrm{meas}_{d-1}\,(\partial K\cap\partial\Omega_R)}{d\,\mathrm{meas}_d\, K} + c\Big)}. \quad (55)$$

5.3 Comments on conditions

- As was already mentioned earlier (see Remark 6), condition (52) implies, in fact, condition (51) for any (positive) value of Δt.
- The second terms in the RHS's of (51) and (52) are zeros for all simplices K which are not adjacent to the part $\partial\Omega_R$ of the solution domain boundary.
- Angle conditions (51) and (52) impose, in general, more severe condition on acuteness and size for those simplices K which are adjacent to $\partial\Omega_R$ than for those in the other parts of the solution domain.
- For the particular case $c = 0$ and $\sigma = 0$, the final conditions (54) and (55) reduce to some known [11, 10] requirements for DMPs.
- In the case $\theta = 1$, we have no upper estimate for the time-step Δt. It is also clear that in a view of (54) we should exclude the case $\theta = 0$. Also, the value of θ should, in general, be taken sufficiently close to 1 to provide the existence of Δt satisfying conditions (54) and (55).
- For details and literature on how to construct acute (and also nonobtuse) simplicial meshes, see [2].
- Our conditions are only sufficient, it is still possible to guarantee the DMP without using the concept of Stieltjes matrix, see [15, 19] for the relevant work in the elliptic case.

6 Numerical tests

In this section, we verify our theoretical results on a two-dimensional test problem. We solve problem (1)–(5) with constant coefficient functions $b = 1$, $c = 100$, and $f = 0$ in a trapezoidal spatial domain depicted in Figure 2. This trapezoid is obtained by cutting off one of the corners of a regular triangle with unit edge lengths.

On the top and on the bottom of the trapezoid the homogeneous Dirichlet boundary condition is prescribed, that is $g \equiv 0$. On the left-hand and on the right-hand sides we apply, respectively, Robin, and Neumann boundary conditions. The coefficient functions in the boundary conditions are chosen to be constant with $q = -1$, $r = -2$ and $\sigma = 100$. Naturally, for this problem, the continuous maximum principle (15) holds. In particular, the maximum principle implies that the solution u is non-positive provided that the initial function u^0 is a non-positive function (see Remark 2).

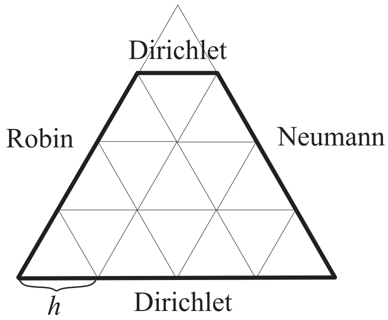

Figure 2: The spatial solution domain with a uniform triangular spatial mesh.

We are going to solve the continuous problem numerically with the finite element method given in Section 3. We use piecewise linear basis functions on uniform triangular meshes (Figure 2). The mesh size is denoted by h. With these basis functions, the elements of the mass and stiffness matrices can be computed relatively easily and the one-step iteration (32) can be constructed fast.

In order to guarantee the maximum principle for the numerical solution it is enough to choose the spatial and temporal discretization parameters according to the conditions (51)–(53). Condition (51) simplifies to

$$25h^2 + \frac{200h}{\sqrt{3}} \leq 2 \tag{56}$$

and results in an upper bound for the mesh size: $h \leq -4/\sqrt{3} + \sqrt{1218}/15 \approx 0.0173$. Thus let $h = 1/64 = 0.015625$. Then, conditions (52) and (53) give the requirement

$$\frac{1}{\theta\left(8/h^2 - 800/(\sqrt{3}h) - 100\right)} \leq \Delta t \leq \frac{1}{(1-\theta)\left(8/h^2 + 800/(\sqrt{3}h) + 100\right)} \tag{57}$$

for the time-step. This condition can be satisfied only for θ values not less than $\theta_{\min} \approx 0.9526$. Let us set $\theta = 0.99$. Then, based on condition (57) and the assumption that $h = 1/64$, the time-step should be chosen from the interval

$$[0.0003250, 0.001603]. \tag{58}$$

Let us suppose that the initial approximation presented in Figure 3 is a finite element approximation of a sufficiently smooth non-positive initial function u^0 that has a spike near the bottom edge of the trapezoid. With this initial approximation the numerical solution should be non-positive. Choosing, however, the time-step outside the interval (58), say $\Delta t = 0.00001$

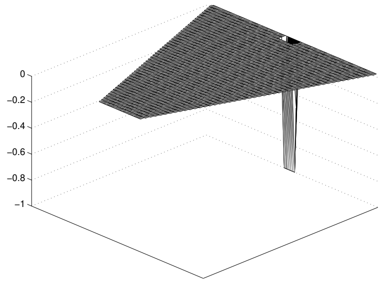

Figure 3: The approximation of a non-positive initial function on the given grid with $h = 1/64$.

and $\Delta t = 0.01$, we obtain that the solution function has also positive values (Figure 4). In the first case, positive values occur in the first 137 iteration steps, while in the second case only in the first step. The numerical tests

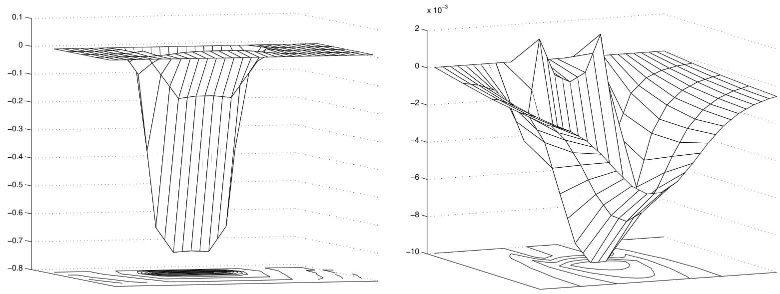

Figure 4: Left: the approximation at the second time level with $\Delta t = 0.00001$. Right: the approximation at the first time level with $\Delta t = 0.01$. The figures show only the critical region near the bottom edge of the trapezoid.

show that with the given initial function the choice of the time-step from the interval $[0.00005645, 0.00825]$ results in a qualitatively adequate solution. The length of this interval is approximately twice bigger than that of the interval (58) in the sufficient condition. The above two time steps, which gave false numerical results, were chosen outside of this interval. A qualitatively adequate solution with the time step $\Delta t = 0.00006$ at the 11-th time level can be seen in Figure 5.

Remark 8 *Estimates* (57) *also show that our sufficient conditions cannot guarantee the numerical maximum principle for the Crank-Nicolson time discretization scheme not even for sufficiently small mesh sizes.*

References

[1] J. Brandts, S. Korotov, M. Křížek, *Dissection of the Path-Simplex in* $\mathbf{R}^n$ *into* n *Path-Subsimplices*, Linear Algebra Appl. 421

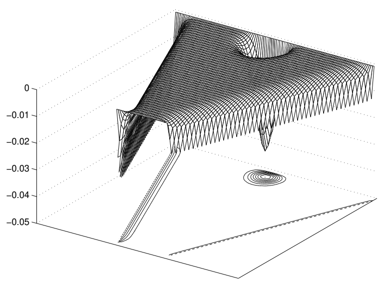

Figure 5: The numerical solution computed with the time step $\Delta t = 0.00006$ at the 11-th time level.

(2007), pp. 382–393.

[2] J. Brandts, S. Korotov, M. Křížek, J. Šolc, *On Nonobtuse Simplicial Partitions*, SIAM Rev. (in press).

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

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

[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[6] 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.

[7] 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.

[8] I. Faragó, R. Horváth, *A Review of Reliable Numerical Methods for Three-Dimensional Parabolic Problems*, Inter. J. Numer. Meth. Engrg. 70 (2006), pp. 25–45.

[9] I. Faragó, R. Horváth, S. Korotov, *Discrete Maximum Principle for Galerkin Finite Element Solutions to Parabolic Problems on Rectangular Meshes* , In: Proc. ENUMATH–2003, Numerical Mathematics and Advanced Applications (eds. M. Feistauer et al.), Prague (Czech Republic), 2003, pp. 298–307.

[10] I. Faragó, R. Horváth, S. Korotov, *Discrete Maximum Principle for Linear Parabolic Problems Solved on Hybrid Meshes*, Appl. Numer. Math. 53 (2005), pp. 249–264.

[11] 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.

[12] R. Horváth, *Sufficient Conditions of the Discrete Maximum-Minimum Principle for Parabolic Problems on Rectangular Meshes*, Comput. Math. Appl. 55 (2008), pp. 2306–2317.

[13] J. Karátson, S. Korotov, *Discrete Maximum Principles for Finite Element Solutions of Nonlinear Elliptic Problems with Mixed Boundary Conditions*, Numer. Math. 99 (2005), pp. 669–698.

[14] H. B. Keller, *The Numerical Solution of Parabolic Partial Differential Equations*, In: Mathematical Methods for Digital Computers (eds. A. Ralston, H.S. Wilf), New York, 1960, pp. 135–143.

[15] 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.

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

[17] 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.

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

[19] 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.

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

[21] R. Varga, *On Discrete Maximum Principle*, J. SIAM Numer. Anal. 3 (1966), pp. 355–359.

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

[23] T. Vejchodský, P. Šolín, *Discrete Maximum Principle for Higher-Order Finite Elements in 1D*, Math. Comp. 76 (2007), pp. 1833–1846.

[24] J. Xu, L. Zikatanov, *A Monotone Finite Element Scheme for Convection-Diffusion Equations*, Math. Comp. 68 (1999), pp. 1429–1446.

[25] C. Yang, Y. Gu, *Minimum Time-Step Criteria for the Galerkin Finite Element Methods Applied to the One-Dimensional Parabolic Partial Differential Equations*, Numer. Methods Partial Differential Equations 22 (2006), 259-273.

(continued from the back cover)

A544 Stig-Olof Londen, Hana Petzeltová
Convergence of solutions of a non-local phase-field system
March 2008

A543 Outi Elina Maasalo
Self-improving phenomena in the calculus of variations on metric spaces
February 2008

A542 Vladimir M. Miklyukov, Antti Rasila, Matti Vuorinen
Stagnation zones for A-harmonic functions on canonical domains
February 2008

A541 Teemu Lukkari
Nonlinear potential theory of elliptic equations with nonstandard growth
February 2008

A540 Riikka Korte
Geometric properties of metric measure spaces and Sobolev-type inequalities
January 2008

A539 Aly A. El-Sabbagh, F.A. Abd El Salam, K. El Nagaar
On the Spectrum of the Symmetric Relations for The Canonical Systems of Differential Equations in Hilbert Space
December 2007

A538 Aly A. El-Sabbagh, F.A. Abd El Salam, K. El Nagaar
On the Existence of the selfadjoint Extension of the Symmetric Relation in Hilbert Space
December 2007

A537 Teijo Arponen, Samuli Piipponen, Jukka Tuomela
Kinematic analysis of Bricard's mechanism
November 2007

A536 Toni Lassila
Optimal damping set of a membrane and topology discovering shape optimization
November 2007

HELSINKI UNIVERSITY OF TECHNOLOGY INSTITUTE OF MATHEMATICS
RESEARCH REPORTS

The reports are available at *http://math.tkk.fi/reports/* .
The list of reports is continued inside the back cover.

A549 Antti Hannukainen, Sergey Korotov, Tomás Vejchodský
On weakening conditions for discrete maximum principles for linear finite element schemes
August 2008

A548 Kalle Mikkola
Weakly coprime factorization, continuous-time systems, and strong-H^p and Nevanlinna fractions
August 2008

A547 Wolfgang Desch, Stig-Olof Londen
A generalization of an inequality by N. V. Krylov
June 2008

A546 Olavi Nevanlinna
Resolvent and polynomial numerical hull
May 2008

A545 Ruth Kaila
The integrated volatility implied by option prices, a Bayesian approach
April 2008

ISBN 978-951-22-9510-4 (print)
ISBN 978-951-22-9511-1 (PDF)
ISSN 0784-3143 (print)
ISSN 1797-5867 (PDF)