Preprint, Institute of Mathematics, AS CR, Prague. 2008-10-27

Discrete maximum principle for parabolic problems solved by prismatic finite elements

Tomáš Vejchodský[1,*] Sergey Korotov[2,†] Antti Hannukainen[2,‡]

September 30, 2008

[1] Institute of Mathematics, Czech Academy of Sciences
Žitná 25, CZ–115 67 Prague 1, Czech Republic
e-mail: vejchod@math.cas.cz [2] Institute of Mathematics, Helsinki University of Technology
P.O. Box 1100, FIN–02015 TKK, Finland
e-mail: antti.hannukainen, sergey.korotov@hut.fi

Abstract: In this paper we analyze the discrete maximum principle (DMP) for a non-stationary diffusion-reaction problem solved by means of prismatic finite elements and θ-method. We derive geometric conditions on the shape parameters of prismatic partitions and time-steps which guarantee validity of the DMP. The presented numerical tests illustrate the sharpness of the obtained conditions.

MSC: 65M60, 65M20, 35B50, 35K20

Keywords: parabolic problem, maximum principle, prismatic finite elements, discrete maximum principle

1 Introduction

Several physical phenomena, like advection, diffusion, reaction, deposition, and emission, play an important role in the modeling of the air-pollution processes, see e.g. [13]. These processes can be modeled by coupled systems of various nonstationary partial differential equations. Numerical methods used to solve such coupled systems are naturally required to discretize all parts of the system properly – in accordance with the underlying physics.

In this work we concentrate on parabolic problems, which form a crucial part of the air-pollution systems, and we study the validity of the associated maximum principle on the

* The first author was supported by Grant no. 102/07/0496 of the Czech Science Foundation, by Grant no. IAA100760702 of the Grant Agency of the Academy of Sciences of the Czech Republic, and by the institutional research plan no. AV0Z10190503 of the Academy of Sciences of the Czech Republic.

† The second author was supported by the Academy Research Fellowship no. 208628 and Project no. 124619 from the Academy of Finland.

‡ The third author was supported by Project no. 124619 from the Academy of Finland.

discrete level. The discrete maximum principle for parabolic problems was first studied in [8] and then in many other publications, see e.g. [1, 2, 3, 6, 5, 4, 7, 10]. However, the case of the prismatic finite elements presented in Section 4 has not been analyzed, yet.

We consider a d^*-dimensional linear parabolic problem in the classical setting: Find function $u \in C^{1,2}((0,\tau)\times\Omega)\cap C([0,\tau)\times\overline{\Omega})$ such that

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

$$u = g \quad \text{on} \quad [0,\tau)\times\partial\Omega, \quad \text{and} \quad u|_{t=0} = u_0 \quad \text{in} \quad \Omega, \tag{2}$$

where Ω is a bounded domain in $\mathbb{R}^{d^*}$ with a boundary $\partial\Omega$, $\tau > 0$, $\varrho = \varrho(t,x) \geq \varrho_0 > 0$, and $c = c(t,x) \geq 0$ with $t \in (0,\tau)$ and $x \in \Omega$. In order to guarantee the existence and uniqueness of the classical solution $u = u(t,x)$, $t \in (0,\tau)$, $x \in \Omega$, we assume that the boundary $\partial\Omega$, the functions $\varrho : (0,\tau)\times\Omega \to \mathbb{R}$, $c : (0,\tau)\times\Omega \to \mathbb{R}$, $u_0 : \Omega \to \mathbb{R}$, $f : (0,\tau)\times\Omega \to \mathbb{R}$ and $g : [0,\tau)\times\partial\Omega \to \mathbb{R}$ are sufficiently smooth and that the initial data u_0 and the boundary data g are compatible for $t = 0$ and $x \in \partial\Omega$.

The above problem serves as the mathematical model of various physical, chemical, and ecological phenomena. An important example is the reaction-diffusion process, where $u(t,x)$ stands for the concentration (of a pollutant for example), u_0 represents the initial concentration, c is the reaction coefficient, f defines the sources, and g is the concentration on the boundary. It is known from the second law of thermodynamics that, in the absence of any source or sink, the concentration takes its maximum value either at the initial state or on the boundary of the body. This property for the classical solution of the above problem with $f = 0$ is preserved, see e.g. [12, p. 79].

Now, we formulate the maximum principle for a general case, when f is not zero. Let $Q_{\bar{t}}$ stand for the cylinder $(0,t]\times\Omega$, $t \in [0,\tau]$, and let $\Gamma_{\bar{t}} = S_{\bar{t}} \cup \Omega_0$ denote the union of its lateral surface $S_{\bar{t}} = [0,t]\times\partial\Omega$ and its bottom part $\Omega_0 = \{0\}\times\Omega$. We present a simple modification of Theorem 2.1 from [11].

Theorem 1.1 *Let $u(t,x)$ be the solution of problem (1)–(2). Then the estimate*

$$\sup_{\lambda > -c_\infty} \min\left\{0; \min_{\Gamma_{\bar{t}_1}} \psi\exp(\lambda(t_1-t)); \min_{Q_{\bar{t}_1}} \frac{f\exp(\lambda(t_1-t))}{c+\varrho\lambda}\right\}$$
$$\leq u(t_1,x) \leq \inf_{\lambda > -c_\infty} \max\left\{0; \max_{\Gamma_{\bar{t}_1}} \psi\exp(\lambda(t_1-t)); \max_{Q_{\bar{t}_1}} \frac{f\exp(\lambda(t_1-t))}{c+\varrho\lambda}\right\}$$

holds for any $t_1 \in [0,\tau]$, where the function ψ coincides with u_0 on Ω_0, and with g on S_τ and $c_\infty = \|c/\varrho\|_{\infty,Q_{\bar{t}_1}} = \sup_{Q_{\bar{t}_1}} |c/\varrho|$.

Let us further introduce the following functions ($t_1 \in [0,\tau]$)

$$\tilde{u}(t_1,x) = \max\{0; \max_{\Gamma_{\bar{t}_1}} \psi\} + t_1 \max\{0; \max_{Q_{\bar{t}_1}} \frac{f}{\varrho}\} - u(t_1,x),$$

and

$$\bar{u}(t_1,x) = u(t_1,x) - \min\{0; \min_{\Gamma_{\bar{t}_1}} \psi\} - t_1 \min\{0; \min_{Q_{\bar{t}_1}} \frac{f}{\varrho}\},$$

where u and ψ are defined above. Due to the initial boundary-value problems which the functions $\tilde{u}$ and $\bar{u}$ satisfy to, and Theorem 1.1, we immediately observe that

$$\tilde{u}(t_1, x) \geq 0 \quad \text{and} \quad \bar{u}(t_1, x) \geq 0,$$

which implies

$$\min\{0; \min_{\Gamma_{\bar{t}_1}} \psi\} + t_1 \min\{0; \min_{Q_{\bar{t}_1}} \frac{f}{\varrho}\} \leq u(t_1, x) \leq \max\{0; \max_{\Gamma_{\bar{t}_1}} \psi\} + t_1 \max\{0; \max_{Q_{\bar{t}_1}} \frac{f}{\varrho}\}. \quad (3)$$

Formula (3) presents the (continuous) maximum principle we shall deal with in the paper. However, there exist several other variants of the maximum principle – see [6] for their overview and for their relationship with the other qualitative properties.

To solve the problem (1)–(2) numerically, we use certain discretizations, both in spatial and in time coordinates. It is obvious that the validity of the discrete analogue of the maximum principle, the so-called *discrete maximum principle* (DMP), is a natural requirement for getting an adequate numerical solution.

2 Discretization

2.1 Variational formulation

The semidiscrete analogue of problem (1)–(2) is based on the variational formulation described shortly below. Let $V_0 = H_0^1(\Omega)$, multiplying (1) for a given t by a function $v \in V_0$, integrating over Ω and using the Green formula, we get

$$\int_\Omega \varrho \frac{\partial u}{\partial t} v \, \mathrm{d}x + L(u, v) = \int_\Omega f v \, \mathrm{d}x,$$

where

$$L(u, v) = \int_\Omega \operatorname{grad} u \cdot \operatorname{grad} v \, \mathrm{d}x + \int_\Omega cuv \, \mathrm{d}x.$$

Under the assumptions providing the existence of the classical solution, it satisfies the following variational formulation:

$$\int_\Omega \varrho \frac{\partial u}{\partial t} v \, \mathrm{d}x + L(u, v) = \int_\Omega f v \, \mathrm{d}x \quad \forall v \in V_0, \quad t \in (0, \tau), \quad (4)$$

with

$$u(0, x) = u_0(x), \quad x \in \Omega, \quad (5)$$

and

$$u(t, x) = g(t, x), \quad x \in \partial\Omega, \quad t \in (0, \tau). \quad (6)$$

We remark that we may assume more general data in the context of the variational formulation. Namely $\varrho, c \in C([0, \tau], L^\infty(\Omega))$ and $f \in C([0, \tau], L^2(\Omega))$.

2.2 Semidiscretization in space

Let Ω be a solution domain covered by a finite element mesh $\mathcal{T}_h$, where h stands for the discretization parameter. Let $P_1, \ldots, P_N$ denote the interior vertices of elements from $\mathcal{T}_h$, and $P_{N+1}, \ldots, P_{\bar{N}}$ the boundary ones. We also define $N_\partial = \bar{N} - N$.

Let the finite element basis functions $\phi_1, \ldots, \phi_{\bar{N}}$ satisfy the delta property $\phi_i(P_j) = \delta_{ij}$, $i, j = 1, \ldots, \bar{N}$, with δ_{ij} being the Kronecker symbol. Further, let

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

Obviously the standard finite element basis functions like the piecewise linear functions on simplices, the piecewise multilinear functions on Cartesian product elements, and the piecewise multilinear functions on prismatic elements satisfy these requirements. In this paper we will focus on the basis functions for the three dimensional right triangular prismatic elements. Detailed construction will be given in Section 4.

We denote $V^h = \operatorname{span}\{\phi_i, i = 1, 2, \ldots, \bar{N}\}$ the space of all possible linear combinations of the basis functions and define its subspace $V_0^h = \{v \in V^h \mid v|_{\partial\Omega} = 0\}$. Then the semidiscrete problem for (4)–(6) (or, equivalently, for (1)–(2)) reads:

Find a function $u_h = u_h(t, x) \in C^1([0, \tau], V^h)$ such that

$$\int_\Omega \varrho \frac{\partial u_h}{\partial t} v_h \,\mathrm{d}x + L(u_h, v_h) = \int_\Omega f v_h \,\mathrm{d}x \quad \forall v_h \in V_0^h, \quad t \in (0, \tau), \tag{8}$$

and

$$u_h(0, x) = u_0^h(x), \quad x \in \Omega, \tag{9}$$

$$u_h(t, x) - g_h(t, x) \in V_0^h, \quad t \in (0, \tau). \tag{10}$$

In the above, $u_0^h(x)$ 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.

$$u_0^h(x) = \sum_{j=1}^{\bar{N}} u_0(P_j)\phi_j(x),$$

and, similarly,

$$g_h(t, x) = \sum_{j=1}^{N_\partial} g_j^h(t)\phi_{N+j}(x), \quad \text{where} \quad g_j^h(t) = g(t, P_{N+j}),\ j = 1, \ldots, N_\partial.$$

We notice that due to the consistency of the initial and the boundary conditions ($g(0, x) = u_0(x),\ x \in \partial\Omega$), we have $g_j^h(0) = u_0(P_{N+j}),\ j = 1, \ldots, N_\partial$.

We search for the semidiscrete solution in the form

$$u_h(t, x) = \sum_{j=1}^{N} u_j^h(t)\phi_j(x) + g_h(t, x),$$

and notice that it is sufficient that u_h satisfies (8) for $v_h = \phi_i$, $i = 1, \dots, N$, only.

Introducing the notation

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

we, thus, arrive at the Cauchy problem for the system of ordinary differential equations

$$\mathbf{M}\frac{\mathrm{d}\mathbf{v}^h}{\mathrm{d}t} + \mathbf{K}\mathbf{v}^h = \mathbf{f}, \quad \mathbf{v}^h(0) = [u_0(P_1), \dots, u_0(P_N), g_1^h(0), \dots, g_{N_\partial}^h(0)]^\top \tag{11}$$

for the solution of the semidiscrete problem, where

$$\begin{aligned}
\mathbf{M} = \mathbf{M}(t) = [M_{ij}(t)]_{N\times\bar{N}}, \quad & M_{ij}(t) = \int_\Omega \varrho(t,x)\phi_j(x)\phi_i(x)\,\mathrm{d}x,\\
\mathbf{K} = \mathbf{K}(t) = [K_{ij}(t)]_{N\times\bar{N}}, \quad & K_{ij}(t) = L(\phi_j, \phi_i),\\
\mathbf{f} = \mathbf{f}(t) = [f_i(t)]_{N\times 1}, \quad & f_i(t) = \int_\Omega f\phi_i\,\mathrm{d}x,
\end{aligned}$$

$i = 1, 2, \dots, N$, $j = 1, 2, \dots, \bar{N}$. The above defined matrices $\mathbf{M}$ and $\mathbf{K}$ are called *mass and stiffness matrices*, respectively.

We notice that, due to our choice of the projections into V^h in (9) and (10), we obtain

$$\max\{0, \max\{(\mathbf{v}^h(0))_i, i = 1, 2, \dots, \bar{N}\}\} \le \max\{0, \max\{u_0(x), x \in \overline{\Omega}\}\}$$

and

$$\min\{0, \min\{(\mathbf{v}^h(0))_i, i = 1, 2, \dots, \bar{N}\}\} \ge \min\{0, \min\{u_0(x), x \in \overline{\Omega}\}\},$$

respectively. Obviously, $g_i^h(t)$ can be estimated by $g(t,x)$ in the same manner. Hence, we have

$$\max\{0, \max_{\substack{t\in[0,\tau]\\ i=1,\dots,N_\partial}} g_i^h(t)\} \le \max\{0, \max_{\substack{t\in[0,\tau]\\ x\in\partial\Omega}} g(t,x)\}$$

and

$$\min\{0, \min_{\substack{t\in[0,\tau]\\ i=1,\dots,N_\partial}} g_i^h(t)\} \ge \min\{0, \min_{\substack{t\in[0,\tau]\\ x\in\partial\Omega}} g(t,x)\}.$$

2.3 Fully discretized problem

In order to get a fully discrete numerical scheme, we choose a time-step Δt and define the partition $t_n = n\Delta t$, $n = 0, 1, \dots, n_\tau$, of the time interval $[0, \tau]$, where $t_{n_\tau} = n_\tau \Delta t = \tau$. Let us remark that we consider constant the time step Δt for the simplicity only. All the subsequent analysis can be easily generalized to the variable time step.

We denote the approximations to $\mathbf{v}^h(t_n)$ and $\mathbf{f}(t_n)$ by $\mathbf{v}^n$ and $\mathbf{f}^n$, $n = 0, 1, \dots, n_\tau$, respectively. Similarly we denote $\mathbf{M}^n = \mathbf{M}(t_n)$ and $\mathbf{K}^n = \mathbf{K}(t_n)$ as well as the entries $M_{ij}^n = M_{ij}(t_n)$, $K_{ij}^n = K_{ij}(t_n)$. This notation is used further on also for other quantities. To discretize (11), we apply the θ-method ($\theta \in [0, 1]$ is a given parameter) and obtain the system of linear algebraic equations

$$\mathbf{M}^{(n,\theta)}\frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{\Delta t} + \theta\mathbf{K}^{n+1}\mathbf{v}^{n+1} + (1-\theta)\mathbf{K}^n\mathbf{v}^n = \mathbf{f}^{(n,\theta)}, \tag{12}$$

where $\mathbf{M}^{(n,\theta)} = \theta\mathbf{M}^{n+1} + (1-\theta)\mathbf{M}^n$ and $\mathbf{f}^{(n,\theta)} = \theta\mathbf{f}^{n+1} + (1-\theta)\mathbf{f}^n$, $n = 0, 1, \ldots, n_\tau - 1$.

System (12) can be rewritten as

$$\left(\mathbf{M}^{(n,\theta)} + \theta\Delta t\mathbf{K}^{n+1}\right)\mathbf{v}^{n+1} = \left(\mathbf{M}^{(n,\theta)} - (1-\theta)\Delta t\mathbf{K}^n\right)\mathbf{v}^n + \Delta t\ \mathbf{f}^{(n,\theta)}, \quad n = 0, 1, \ldots, n_\tau - 1, \tag{13}$$

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

To shorten the notation we put $\mathbf{A} = \mathbf{M}^{(n,\theta)} + \theta\Delta t\mathbf{K}^{n+1}$ and $\mathbf{B} = \mathbf{M}^{(n,\theta)} - (1-\theta)\Delta t\mathbf{K}^n$. In what follows, we shall use the following partitions of the matrices and vectors:

$$\mathbf{A} = [\mathbf{A}_0|\mathbf{A}_\partial], \quad \mathbf{B} = [\mathbf{B}_0|\mathbf{B}_\partial], \quad \mathbf{v}^n = \begin{bmatrix} \mathbf{u}^n \\ \mathbf{g}^n \end{bmatrix}, \tag{14}$$

where $\mathbf{A}_0$ and $\mathbf{B}_0$ are square matrices from $\mathbb{R}^{N\times N}$; $\mathbf{A}_\partial, \mathbf{B}_\partial \in \mathbb{R}^{N\times N_\partial}$, $\mathbf{u}^n = [u_1^n, \ldots, u_N^n]^\top \in \mathbb{R}^N$ and $\mathbf{g}^n = [g_1^n, \ldots, g_{N_\partial}^n]^\top \in \mathbb{R}^{N_\partial}$. Similarly, this partition is used for matrices $\mathbf{M}^n$ and $\mathbf{K}^n$. Then, the iteration (13) can be also written as

$$\mathbf{A}\mathbf{v}^{n+1} = \mathbf{B}\mathbf{v}^n + \Delta t\ \mathbf{f}^{(n,\theta)}, \tag{15}$$

or

$$[\mathbf{A}_0|\mathbf{A}_\partial]\begin{bmatrix} \mathbf{u}^{n+1} \\ \mathbf{g}^{n+1} \end{bmatrix} = [\mathbf{B}_0|\mathbf{B}_\partial]\begin{bmatrix} \mathbf{u}^n \\ \mathbf{g}^n \end{bmatrix} + \Delta t\ \mathbf{f}^{(n,\theta)} \tag{16}$$

with $n = 0, 1, \ldots, n_\tau - 1$.

3 The Discrete Maximum Principle

Let us define the following values

$$\begin{aligned} g_{\min}^n &= \min\{0, g_1^n, \ldots, g_{N_\partial}^n\}, & g_{\max}^n &= \max\{0, g_1^n, \ldots, g_{N_\partial}^n\}, \\ v_{\min}^n &= \min\{0, g_{\min}^n, u_1^n, \ldots, u_N^n\}, & v_{\max}^n &= \max\{0, g_{\max}^n, u_1^n, \ldots, u_N^n\}, \end{aligned}$$

for $n = 0, 1, \ldots, n_\tau$ and

$$f_{\min}^{(n,n+1)} = \min\left\{0, \min_{\substack{t\in[t_n,t_{n+1}]\\ x\in\overline{\Omega}}} \frac{f(t,x)}{\varrho(t,x)}\right\}, \quad f_{\max}^{(n,n+1)} = \max\left\{0, \max_{\substack{t\in[t_n,t_{n+1}]\\ x\in\overline{\Omega}}} \frac{f(t,x)}{\varrho(t,x)}\right\},$$

for $n = 0, 1, \ldots, n_\tau - 1$.

The discrete analogue (DMP) for the continuous maximum principle (3) can be represented as follows:

$$\begin{gathered} \min\{0, v_{\min}^0, \min\{g_{\min}^{k+1}, k = 0, \ldots, n\}\} + t_{n+1}\min\{0, \min\{f_{\min}^{(k,k+1)}, k = 0, \ldots, n\}\} \leq \\ \leq u_i^{n+1} \leq \\ \leq \max\{0, v_{\max}^0, \max\{g_{\max}^{k+1}, k = 0, \ldots, n\}\} + t_{n+1}\max\{0, \max\{f_{\max}^{(k,k+1)}, k = 0, \ldots, n\}\}, \end{gathered} \tag{17}$$

where $i = 1, 2, \ldots, N,\ n = 0, 1, \ldots, n_\tau - 1$.

The DMP (17) is equivalent to the following relation

$$\min\{0, v_{\min}^n, g_{\min}^{n+1}\} + \Delta t f_{\min}^{(n,n+1)} \le u_i^{n+1} \le \max\{0, v_{\max}^n, g_{\max}^{n+1}\} + \Delta t f_{\max}^{(n,n+1)}, \\ i = 1, \dots, N;\ n = 0, \dots, n_\tau - 1. \quad (18)$$

This DMP was presented in [8, p. 100], where it is proved in the case of $c = 0$ and simplicial finite elements.

Let us introduce the notation

$$\mathbf{e} = [1, \dots, 1]^\top \in \mathbb{R}^{\bar{N}}, \quad \mathbf{e}_0 = [1, \dots, 1]^\top \in \mathbb{R}^{N}, \quad \mathbf{e}_\partial = [1, \dots, 1]^\top \in \mathbb{R}^{N_\partial},$$
$$\mathbf{f}_{\max}^{(n,n+1)} = f_{\max}^{(n,n+1)} \mathbf{e} \in \mathbb{R}^{\bar{N}}, \quad \mathbf{v}_{\max}^{n} = v_{\max}^{n} \mathbf{e} \in \mathbb{R}^{\bar{N}},$$
$$\mathbf{f}_{0}^{(n,n+1)} = f_{\max}^{(n,n+1)} \mathbf{e}_0 \in \mathbb{R}^{N}, \quad \mathbf{v}_{0}^{n} = v_{\max}^{n} \mathbf{e}_0 \in \mathbb{R}^{N},$$
$$\mathbf{f}_{\partial}^{(n,n+1)} = f_{\max}^{(n,n+1)} \mathbf{e}_\partial \in \mathbb{R}^{N_\partial}, \quad \mathbf{v}_{\partial}^{n} = v_{\max}^{n} \mathbf{e}_\partial \in \mathbb{R}^{N_\partial}.$$

For simplicity, we denote zero matrices and zero vectors by the same symbol $\mathbf{0}$, whose size is always chosen according to the context. The ordering relations between vectors or matrices are meant elementwise.

Lemma 3.1 *Let the basis functions satisfy (7). Then the following relations hold*

$(P1)$ $\mathbf{K}(t)\mathbf{e} \ge \mathbf{0}, \quad t \in [0, \tau]$,
$(P2)$ $\mathbf{f}^{(n,\theta)} \le \mathbf{A}\mathbf{f}_{\max}^{(n,n+1)}, \quad n = 0, 1, \dots, n_\tau - 1$,
$(P3)$ *If* $\mathbf{A}_0^{-1} \ge \mathbf{0}$ *then* $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\, \mathbf{e}_\partial \le \mathbf{e}_0, \quad n = 0, 1, \dots, n_\tau - 1$.

Proof. (P1) For the ith coordinate of the vector $\mathbf{K}\mathbf{e} = \mathbf{K}(t)\mathbf{e}$, $t \in [0, \tau]$, we have

$$(\mathbf{K}\mathbf{e})_i = \sum_{j=1}^{\bar{N}} K_{ij} = \sum_{j=1}^{\bar{N}} L(\phi_j, \phi_i) = L\left(\sum_{j=1}^{\bar{N}} \phi_j, \phi_i\right) = L(1, \phi_i) = \\ = \int_\Omega \operatorname{grad} 1 \cdot \operatorname{grad} \phi_i \,\mathrm{d}x + \int_\Omega c\, 1\, \phi_i \,\mathrm{d}x = \int_\Omega c\, \phi_i \,\mathrm{d}x \ge 0,$$

which proves the statement.

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

$$\begin{aligned}
(\mathbf{f}^{(n,\theta)})_i &= \int_\Omega \Big[(1-\theta)\frac{f(t_n,x)}{\varrho(t_n,x)}\varrho(t_n,x) + \theta\frac{f(t_{n+1},x)}{\varrho(t_{n+1},x)}\varrho(t_{n+1},x)\Big]\phi_i(x)\,\mathrm{d}x \\
&\le f_{\max}^{(n,n+1)}\left[\int_\Omega (1-\theta)\varrho(t_n,x)\phi_i(x)\,\mathrm{d}x + \int_\Omega \theta\varrho(t_{n+1},x)\phi_i(x)\,\mathrm{d}x\right] \\
&= f_{\max}^{(n,n+1)} \sum_{j=1}^{\bar{N}} \Big[(1-\theta)M_{ij}^n + \theta M_{ij}^{n+1}\Big] = \Big(\mathbf{M}^{(n,\theta)}\mathbf{f}_{\max}^{(n,n+1)}\Big)_i \\
&\le \Big((\mathbf{M}^{(n,\theta)} + \theta\Delta t\mathbf{K}^{n+1})\mathbf{f}_{\max}^{(n,n+1)}\Big)_i = \Big(\mathbf{A}\mathbf{f}_{\max}^{(n,n+1)}\Big)_i,
\end{aligned}$$

where we used (P1) and the following fact

$$\int_\Omega \varrho(t,x)\phi_i(x)\,\mathrm{d}x = \int_\Omega \varrho(t,x)\left(\sum_{j=1}^{\bar{N}}\phi_j(x)\right)\phi_i(x)\,\mathrm{d}x = \sum_{j=1}^{\bar{N}} M_{ij}(t).$$

(P3) Matrix $\mathbf{M}^{(n,\theta)}$ is nonnegative, because $\varrho > 0$ and because the basis functions are nonnegative. Thus, $\mathbf{0} \le \mathbf{M}^{(n,\theta)}\mathbf{e} \le (\mathbf{M}^{(n,\theta)} + \theta\Delta t\mathbf{K}^{n+1})\mathbf{e} = \mathbf{A}\mathbf{e} = \mathbf{A}_0\mathbf{e}_0 + \mathbf{A}_\partial\mathbf{e}_\partial$, where we utilized (P1). The statement (P3) is obtained by multiplying both sides by the non-negative matrix $\mathbf{A}_0^{-1}$. ■

Theorem 3.2 *Let the basis functions satisfy (7). Then the Galerkin solution of the problem (1)–(2), combined with the θ-method in the time discretization, satisfies (18) (and, therefore, the DMP (17)) if and only if the conditions*

$$\begin{array}{ll}(C1) & \mathbf{A}_0^{-1} \ge \mathbf{0},\\ (C2) & \mathbf{A}_0^{-1}\mathbf{A}_\partial \le \mathbf{0},\\ (C3) & \mathbf{A}_0^{-1}\mathbf{B} \ge \mathbf{0},\end{array}$$

hold for $n = 0, 1, \ldots, n_\tau - 1$.

Proof. First, we prove the sufficiency of the conditions verifying the inequality on the right-hand side in (18). From (15) and (P2), we have

$$\mathbf{A}_0\mathbf{u}^{n+1} + \mathbf{A}_\partial\mathbf{g}^{n+1} = \mathbf{A}\mathbf{v}^{n+1} = \mathbf{B}\mathbf{v}^n + \Delta t\ \mathbf{f}^{(n,\theta)} \le \mathbf{B}\mathbf{v}^n + \Delta t\mathbf{A}\mathbf{f}_{\max}^{(n,n+1)}. \tag{19}$$

From (P1), it follows that $\mathbf{B}\mathbf{v}_{\max}^n \le \mathbf{A}\mathbf{v}_{\max}^n$. Multiplying both sides of (19) by $\mathbf{A}_0^{-1} \ge \mathbf{0}$ (see (C1)), expressing $\mathbf{u}^{n+1}$ and using (C3), we obtain

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

Regrouping the above inequality, we get

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

Hence, for the ith coordinate of the both sides we obtain

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

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

The inequality on the left-hand side of (18) can be proved similarly. This completes the proof of the sufficiency of the conditions.

Now, let the DMP (18) be valid, then it is valid for any choice of the vectors $\mathbf{f}^{(n,\theta)}$, $\mathbf{g}^n$, $\mathbf{g}^{n+1}$, and $\mathbf{u}^n$. Below in this proof the symbol $\mathbf{e}_j$ stands for the jth unit vector with all entries equal to zero except for the jth entry which is one. With the choice $\mathbf{g}^{n+1} = 0$, $\mathbf{v}^n = \mathbf{0}$, $\varrho = 1$, $\mathbf{f}^{(n,\theta)} = \mathbf{e}_j$, we get the relation $\mathbf{A}_0^{-1} \geq \mathbf{0}$. Really, combining (16) and (18) 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 non-negative, i.e., $\mathbf{A}_0^{-1} \geq \mathbf{0}$. Thus, the necessity of (C1) is proved. [1]

Using again (16) and (18) with the choice $\mathbf{g}^{n+1} = \mathbf{e}_j$, $\mathbf{g}^n = \mathbf{0}$, $\varrho = 1$, $\mathbf{f}^{(n,\theta)} = \mathbf{0}$ and $\mathbf{u}^n = \mathbf{0}$, we obtain the necessity of (C2), and similarly, with $\mathbf{g}^{n+1} = \mathbf{0}$, $\mathbf{v}^n = \mathbf{e}_j$, $\varrho = 1$, $\mathbf{f}^{(n,\theta)} = \mathbf{0}$, we get the necessity of the condition (C3). ■

Remark 3.3 *It is easy to show that if the basis functions satisfy (7) then the conditions*

$$\begin{array}{ll} (C1^\star) & \mathbf{A}_0^{-1} \geq \mathbf{0}, \\ (C2^\star) & \mathbf{A}_\partial \leq \mathbf{0}, \\ (C3^\star) & \mathbf{B} \geq \mathbf{0}, \end{array}$$

(where $n = 0, 1, \ldots, n_\tau - 1$) ensure (C1)–(C3).

Theorem 3.4 *Let the basis functions satisfy (7). Then the Galerkin solution of the problem (1)–(2), combined with the θ-method in the time discretization, satisfies the discrete maximum principle (18) if the conditions*

$$\begin{array}{ll} (C1') & K_{ij}^n \leq 0, \quad i \neq j,\ i = 1, \ldots, N,\ j = 1, \ldots, \bar{N},\ n = 0, \ldots, n_\tau, \\ (C2') & M_{ij}^{(n,\theta)} + \theta \Delta t\ K_{ij}^{n+1} \leq 0, \quad i \neq j,\ i = 1, \ldots, N,\ j = 1, \ldots, \bar{N},\ n = 0, \ldots, n_\tau - 1, \\ (C3') & M_{ii}^{(n,\theta)} - (1-\theta)\Delta t\ K_{ii}^n \geq 0, \quad i = 1, \ldots, N,\ n = 0, \ldots, n_\tau - 1, \end{array}$$

hold. Here, $M_{ij}^{(n,\theta)}$ and K_{ij}^n stand for the entries of matrices $\mathbf{M}^{(n,\theta)}$ and $\mathbf{K}^n$.

Proof. It is enough to show that $(C1^\star)$–$(C3^\star)$ follow from the assumptions of the theorem. Relations $(C1')$ and $(C3')$ yield $(C3^\star)$, condition $(C2^\star)$ follows from $(C2')$, and $(C1^\star)$ can be shown proving that under the assumptions of the theorem $\mathbf{A}_0$ is a so-called M-matrix (which matrices have non-negative inverse). This follows from the facts that the off-diagonal elements of $\mathbf{A}_0$ are non-positive (see $(C2')$) and $\mathbf{A}_0$ is a positive definite matrix. ■

[1] We remark that the straightforward possibility how to obtain $\mathbf{f}^{(n,\theta)} = \mathbf{f}^n = \mathbf{e}_j$ is to choose $f(t,x) = \delta_{P_j}(x)$, where δ_{P_j} is the Dirac function centered at the vertex P_j. This choice, however, does not satisfy the smoothness requirements for f. To be rigorous, we have to consider a sequence f^ε which approximates the Dirac function δ_{P_j} and pass to the limit.

4 Prismatic Meshes

4.1 Preliminaries

Let us assume that the domain Ω is three-dimensional and that it can be partitioned (face-to-face) to right triangular prisms. Let us denote such a partition $\mathcal{T}_h$ and call it prismatic mesh or prismatic partition. Each element of $\mathcal{T}_h$ is a right triangular prism $\mathcal{P} = T \times I$, where T is a triangle and I a line segment. A typical domain for which a prismatic partition exists is a cylindrical domain $\Omega = \mathcal{G} \times \mathcal{I}$ where $\mathcal{G} \subset \mathbb{R}^2$ is a polygon and $\mathcal{I} \subset \mathbb{R}$ a line segment. The prismatic partition $\mathcal{T}_h$ of Ω can be then constructed as the Cartesian product of a triangulation of $\mathcal{G}$ and a partition of $\mathcal{I}$. However, we remark that the domain Ω can be much more complicated. In general it can be a finite union of cylindrical domains.

The finite element space $V_0^h \subset H_0^1(\Omega)$ associated to $\mathcal{T}_h$ is defined in the case of right triangular prismatic elements as follows:

$$V_0^h = \Big\{ \varphi \in H_0^1(\Omega) :\ \varphi(x,y,z)|_{\mathcal{P}} = \sum_{i=1}^{3}\sum_{j=1}^{2} \sigma_{ij}\lambda_i(x,y)\ell_j(z),\ \text{where } \mathcal{P} \in \mathcal{T}_h, \\ \mathcal{P} = T \times I,\ \sigma_{ij} \in \mathbb{R},\ \lambda_i \in \mathbb{P}^1(T),\ \ell_j \in \mathbb{P}^1(I) \Big\},$$

where $\mathbb{P}^1(T)$ and $\mathbb{P}^1(I)$ stand for the spaces of linear functions defined in the triangle T and in the interval I, respectively. In agreement with the previous notation $\phi_1, \dots, \phi_N$ stand for the standard finite element basis functions in V_0^h. These basis functions are uniquely determined by the requirement $\phi_i(P_j) = \delta_{ij}$, $i = 1, 2, \dots, N$, $j = 1, 2, \dots, N + N^\partial$, where δ_{ij} is the Kronecker symbol and $P_i,\ i = 1, \dots, N + N^\partial$, stand for the vertices of $\mathcal{T}_h$.

The corresponding discrete solution is then given by (13). Our goal is to provide conditions on the prismatic partition $\mathcal{T}_h$ which would guarantee the DMP (17). We obtain these conditions by inspection of requirements $(C1')$–$(C3')$.

4.2 Element mass and stiffness matrices on prisms

To analyze requirements $(C1')$–$(C3')$ we have to investigate the matrices $\mathbf{K}(t)$ and $\mathbf{M}^{(n,\theta)}$. The matrix $\mathbf{K}(t)$ consists of two parts, $\mathbf{K}(t) = \mathbf{S} + \mathbf{C}(t)$, where

$$\mathbf{C}(t) = [C_{ij}(t)]_{N\times\bar{N}}, \quad C_{ij}(t) = \int_\Omega c(t,x)\phi_j(x)\phi_i(x)\,\mathrm{d}x,$$
$$\mathbf{S} = [S_{ij}]_{N\times\bar{N}}, \qquad S_{ij} = \int_\Omega \operatorname{grad}\phi_j \cdot \operatorname{grad}\phi_i\,\mathrm{d}x, \quad i = 1,2,\dots,N,\ j = 1,2,\dots,\bar{N}.$$

We consider also the element matrices $\mathbf{M}^{(n,\theta),(\mathcal{P})}$, $\mathbf{K}^{(\mathcal{P})}(t) = \mathbf{S}^{(\mathcal{P})} + \mathbf{C}^{(\mathcal{P})}(t)$ which are defined as follows

$$\mathbf{M}^{(n,\theta),(\mathcal{P})} = [M_{ij}^{(n,\theta),(\mathcal{P})}]_{N\times\bar{N}}, \quad M_{ij}^{(n,\theta),(\mathcal{P})} = \int_{\mathcal{P}} \varrho^{(n,\theta)}(x)\phi_j(x)\phi_i(x)\,\mathrm{d}x,$$
$$\mathbf{C}^{(\mathcal{P})}(t) = [C_{ij}^{(\mathcal{P})}(t)]_{N\times\bar{N}}, \qquad C_{ij}^{(\mathcal{P})}(t) = \int_{\mathcal{P}} c(t,x)\phi_j(x)\phi_i(x)\,\mathrm{d}x,$$
$$\mathbf{S}^{(\mathcal{P})} = [S_{ij}^{(\mathcal{P})}]_{N\times\bar{N}}, \qquad S_{ij}^{(\mathcal{P})} = \int_{\mathcal{P}} \operatorname{grad}\phi_j \cdot \operatorname{grad}\phi_i\,\mathrm{d}x \tag{20}$$

with $\varrho^{(n,\theta)}(x) = (1-\theta)\varrho(t_n, x) + \theta\varrho(t_{n+1}, x)$. Here $\mathcal{P} \in \mathcal{T}_h$, $\mathcal{P} = T \times I$ is a right triangular prism. We will use the notation for its vertices, edges, and angles shown in Figure 1.

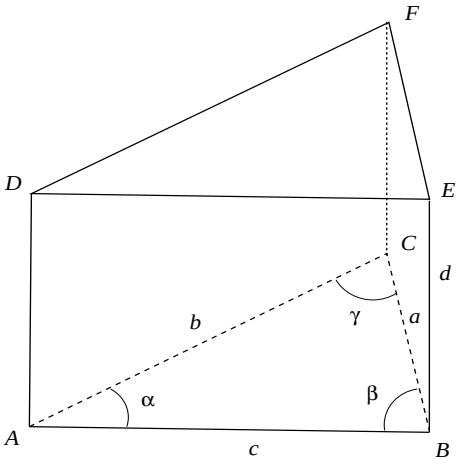

Figure 1: Basic notation for the prismatic element.

The analysis of conditions $(C1')$–$(C3')$ is further based on the investigation of element matrices $\mathbf{M}^{(n,\theta),(\mathcal{P})}$, $\mathbf{S}^{(\mathcal{P})}$ and $\mathbf{C}^{(\mathcal{P})}(t)$. Recently we proved the DMP for an elliptic problem discretized by prismatic elements, see [9], where we already computed the entries of the matrix $\mathbf{S}^{(\mathcal{P})}$ as well as the entries of the matrices $\mathbf{M}^{(n,\theta),(\mathcal{P})}$ and $\mathbf{C}^{(\mathcal{P})}(t)$ in case $\varrho^{(n,\theta)}(x) = c(t,x) = 1$ in $[0,\tau] \times \overline{\Omega}$. In [9] we also present explicit expressions for the off-diagonal entries of these matrices. In the parabolic case we will need, in addition, explicit expressions for the diagonal entries.

Let A denote a vertex of a prism $\mathcal{P} = T \times I$. Without loss of generality we assume that T lies in the xy-plane and that $I = [0,d]$. Let $\varphi_A(x,y,z) = \lambda_A(x,y)\ell_0(z)$ be the shape function corresponding to the vertex A of the prism $\mathcal{P}$, where λ_A is the barycentric coordinate corresponding to the vertex A of the base triangle T and $\ell_0(z) = 1 - z/d$, $z \in I$, is the 1D shape function. We may easily compute, see [9], the following integrals

$$\int_{\mathcal{P}} |\mathrm{grad}\ \varphi_A|^2 \,\mathrm{d}\mathcal{P} = \frac{d}{6}\left(\cot\beta + \cot\gamma + \frac{|T|}{d^2}\right), \qquad \int_{\mathcal{P}} \varphi_A^2 \,\mathrm{d}\mathcal{P} = \frac{d\,|T|}{18}. \tag{21}$$

These are the desired explicit expressions for the (nonzero) diagonal entries of the element matrices $\mathbf{S}^{(\mathcal{P})}$ and also for $\mathbf{M}^{(n,\theta),(\mathcal{P})}$ and $\mathbf{C}^{(\mathcal{P})}(t)$ provided $\varrho^{(n,\theta)}(x) = c(t,x) = 1$ in $[0,\tau]\times\overline{\Omega}$.

To utilize the analysis of the elliptic case from [9] as much as possible, we introduce the following theorem.

Theorem 4.1 *Let* $\widetilde{\mathbf{K}}^{(\mathcal{P})} = \mathbf{S}^{(\mathcal{P})} + \widetilde{\mathbf{M}}^{(\mathcal{P})}$ *with* $\mathbf{S}^{(\mathcal{P})}$ *given by* (20) *and with*

$$\widetilde{\mathbf{M}}^{(\mathcal{P})} = \big[\widetilde{M}_{ij}^{(\mathcal{P})}\big]_{N\times\bar{N}}, \quad \widetilde{M}_{ij}^{(\mathcal{P})} = \int_{\mathcal{P}} \widetilde{c}(x)\phi_j(x)\phi_i(x)\,\mathrm{d}x,$$

be the element matrix for the prismatic element $\mathcal{P} = T\times I$ *with the reaction coefficient* $\widetilde{c} \geq 0$. *Let* d *be the altitude of* $\mathcal{P}$ *and let* $\alpha_{\min}^{(T)} \leq \alpha_{\mathrm{med}}^{(T)} \leq \alpha_{\max}^{(T)}$ *be the angles in the triangular base*

T. If

$$\|\widetilde{c}\|_{\infty,\mathcal{P}}\,\frac{|T|}{6}+\frac{\cot\alpha_{\mathrm{med}}^{(T)}+\cot\alpha_{\mathrm{min}}^{(T)}}{2}\le\frac{|T|}{d^2}\le 2\cot\alpha_{\mathrm{max}}^{(T)}-\|\widetilde{c}\|_{\infty,\mathcal{P}}\,\frac{|T|}{3}\tag{22}$$

then the off-diagonal entries of $\widetilde{\mathbf{K}}^{(\mathcal{P})}$ *are nonpositive, i.e.,* $\widetilde{K}_{ij}^{(\mathcal{P})}\le 0$ *for* $i\neq j$.

Proof. This is just a reformulation of Theorem 2 and relation (20) from [9]. ■

4.3 Conditions on meshes and time-steps

Definition 4.2. Let $\mathcal{P}\in\mathcal{T}_h$, $\mathcal{P}=T\times I$ be a prism. For $n=0,1,\ldots,n_\tau-1$ we define

$$\delta_L^{n,(\mathcal{P})}=\frac{1}{\varrho_{\mathrm{min}}^{(n,\theta),(\mathcal{P})}}\left[\frac{3}{|T|}\left(\cot\alpha_{\mathrm{med}}^{(T)}+\cot\alpha_{\mathrm{min}}^{(T)}\right)+\|c^n\|_{\infty,\mathcal{P}}+\frac{3}{d^2}\right],$$

$$\delta_U^{n,(\mathcal{P})}=\frac{1}{\varrho_{\mathrm{max}}^{(n,\theta),(\mathcal{P})}}\min\left\{-\frac{3}{|T|}\left(\cot\alpha_{\mathrm{med}}^{(T)}+\cot\alpha_{\mathrm{min}}^{(T)}\right)-c_{\mathrm{max}}^{n,n+1,(\mathcal{P})}+\frac{6}{d^2},\right.$$
$$\left.\frac{6}{|T|}\cot\alpha_{\mathrm{max}}^{(T)}-c_{\mathrm{max}}^{n,n+1,(\mathcal{P})}-\frac{3}{d^2}\right\},$$

where $\varrho_{\mathrm{max}}^{(n,\theta),(\mathcal{P})}=\sup_{\mathcal{P}}\varrho^{(n,\theta)}(x)$, $\varrho_{\mathrm{min}}^{(n,\theta),(\mathcal{P})}=\inf_{\mathcal{P}}\varrho^{(n,\theta)}(x)$, $c_{\mathrm{max}}^{n,n+1,(\mathcal{P})}=\max\left\{\|c^n\|_{\infty,\mathcal{P}},\|c^{n+1}\|_{\infty,\mathcal{P}}\right\}$, $\varrho\ge\varrho_0>0$ and $c\ge 0$ are the coefficient from the equation (1), d stands for the altitude of the prism $\mathcal{P}$, $|T|$ denotes the area of the triangle T, and $\alpha_{\mathrm{min}}^{(T)}\le\alpha_{\mathrm{med}}^{(T)}\le\alpha_{\mathrm{max}}^{(T)}$ are the (ordered) angles in T.

Theorem 4.3 *Let* $\mathcal{T}_h$ *be a prismatic partition of* Ω. *Then the Galerkin solution of the problem (1)–(2), combined with the* θ*-method in the time discretization, satisfies the DMP (17) provided the following condition holds for all prisms* $\mathcal{P}\in\mathcal{T}_h$ *and* $n=0,1,\ldots,n_\tau-1$

$$(1-\theta)\delta_L^{n,(\mathcal{P})}\le\frac{1}{\Delta t}\le\theta\delta_U^{n,(\mathcal{P})}.\tag{23}$$

Proof. It suffices to verify conditions $(C1')$–$(C3')$. First, let us notice that if $\theta=0$ then (23) implies $\delta_L^{n,(\mathcal{P})}\le 0$ but by Definition 4.2 we have $\delta_L^{n,(\mathcal{P})}>0$. Therefore if (23) holds true then necessarily we have $\theta>0$.

Thus, we can reformulate inequalities (23) equivalently as follows

$$\frac{|T|}{d^2}\ge\left(\frac{\varrho_{\mathrm{max}}^{(n,\theta),(\mathcal{P})}}{\theta\Delta t}+c_{\mathrm{max}}^{n,n+1,(\mathcal{P})}\right)\frac{|T|}{6}+\frac{\cot\alpha_{\mathrm{med}}^{(T)}+\cot\alpha_{\mathrm{min}}^{(T)}}{2},\tag{24}$$

$$\frac{|T|}{d^2}\le 2\cot\alpha_{\mathrm{max}}^{(T)}-\left(\frac{\varrho_{\mathrm{max}}^{(n,\theta),(\mathcal{P})}}{\theta\Delta t}+c_{\mathrm{max}}^{n,n+1,(\mathcal{P})}\right)\frac{|T|}{3},\tag{25}$$

$$\frac{|T|}{d^2}\le\left(\frac{\varrho_{\mathrm{min}}^{(n,\theta),(\mathcal{P})}}{(1-\theta)\Delta t}-\|c^n\|_{\infty,\mathcal{P}}\right)\frac{|T|}{3}-\cot\alpha_{\mathrm{med}}^{(T)}-\cot\alpha_{\mathrm{min}}^{(T)},\tag{26}$$

where the right-hand side of (26) is understood as infinity if $\theta = 1$.

Conditions $(C1')$ and $(C2')$ are equivalent to

$$S_{ij} + C_{ij}^n \le 0 \quad \text{and} \quad S_{ij} + \frac{1}{\theta\Delta t} M_{ij}^{(n,\theta)} + C_{ij}^{n+1} \le 0, \quad i \ne j.$$

The validity of both these inequalities follows from (24), (25), and Theorem 4.1 with $\widetilde{c} = c(t_n, x)$ and $\widetilde{c} = \varrho^{(n,\theta)}(x)/(\theta\Delta t) + c(t_{n+1}, x)$, respectively. Here we use the inequality $\|\widetilde{c}\|_{\infty,\mathcal{P}} \le \varrho_{\max}^{(n,\theta),(\mathcal{P})}/(\theta\Delta t) + c_{\max}^{n,n+1,(\mathcal{P})}$ which holds in both cases.

To verify $(C3')$ we show the nonnegativity of all element contributions

$$M_{ii}^{(n,\theta),(\mathcal{P})} - (1-\theta)\Delta t\, K_{ii}^{n,(\mathcal{P})} \ge 0 \quad \forall \mathcal{P} \in \mathcal{T}_h. \tag{27}$$

This inequality is trivially satisfied if $\theta = 1$. For $\theta < 1$ we rewrite (27) equivalently as

$$-S_{ii}^{(\mathcal{P})} + \frac{1}{(1-\theta)\Delta t} M_{ii}^{(n,\theta),(\mathcal{P})} - C_{ii}^{n,(\mathcal{P})} \ge 0 \quad \forall \mathcal{P} \in \mathcal{T}_h. \tag{28}$$

Now we show that (26) implies (28) and consequently $(C3')$.

Let P_i, $1 \le i \le N$, be an arbitrary interior node in the prismatic partition $\mathcal{T}_h$. Let P_i be a vertex of a prism $\mathcal{P} \in \mathcal{T}_h$ and let $\varphi_A = \phi_i|_{\mathcal{P}}$. Then we can verify the validity of (28) as follows

$$\begin{aligned}
&- S_{ii}^{(\mathcal{P})} + \frac{1}{(1-\theta)\Delta t} M_{ii}^{(n,\theta),(\mathcal{P})} - C_{ii}^{n,(\mathcal{P})} \\
&= -\int_{\mathcal{P}} |\operatorname{grad} \varphi_A|^2 \,\mathrm{d}\mathcal{P} + \int_{\mathcal{P}} \left(\frac{1}{(1-\theta)\Delta t} \varrho^{(n,\theta)}(x) - c(t_n, x) \right) \varphi_A^2 \,\mathrm{d}\mathcal{P} \\
&\ge -\int_{\mathcal{P}} |\operatorname{grad} \varphi_A|^2 \,\mathrm{d}\mathcal{P} + \left(\frac{\varrho_{\min}^{(n,\theta),(\mathcal{P})}}{(1-\theta)\Delta t} - \|c(t_n, x)\|_{\infty,\mathcal{P}} \right) \int_{\mathcal{P}} \varphi_A^2 \,\mathrm{d}\mathcal{P} \\
&= \frac{d}{6} \left[-\cot\beta - \cot\gamma - \frac{|T|}{d^2} + \left(\frac{\varrho_{\min}^{(n,\theta),(\mathcal{P})}}{(1-\theta)\Delta t} - \|c^n\|_{\infty,\mathcal{P}} \right) \frac{|T|}{3} \right] \ge 0,
\end{aligned}$$

where we used (21) with the angles β and γ being opposite to the vertex $A \equiv P_i$, see Figure 1. The final inequality follows from (26) because cotangent is a decreasing function and we have $\cot\alpha_{\mathrm{med}}^{(T)} + \cot\alpha_{\min}^{(T)} \ge \cot\beta + \cot\gamma$. ■

The crucial condition (23) for the validity of the DMP deserves certain discussion. The first observation is that $\delta_L^{n,(\mathcal{P})} > 0$. Hence, if $\delta_U^{n,(\mathcal{P})} > 0$ then inequalities (23) can always be satisfied by suitable choice of θ (close enough to 1).

Unfortunately, $\delta_U^{n,(\mathcal{P})}$ can be negative or zero in general. The requirement $\delta_U^{n,(\mathcal{P})} > 0$ is equivalent to

$$\frac{c_{\max}^{n,n+1,(\mathcal{P})}}{6}|T| + \frac{\cot\alpha_{\mathrm{mid}}^{(T)} + \cot\alpha_{\min}^{(T)}}{2} < \frac{|T|}{d^2} < 2\cot\alpha_{\max}^{(T)} - \frac{c_{\max}^{n,n+1,(\mathcal{P})}}{3}|T|. \tag{29}$$

This condition guarantees the DMP in the elliptic case, cf. (22) and [9, formula (20)]. Thus, we can conclude that condition (29), which was obtained in the elliptic case, guarantees the DMP also in the parabolic case, provided θ and Δt are chosen to satisfy (23).

The analysis of conditions (29) in the elliptic case yields the notion of the *strictly well-shaped prismatic partitions* of cylindrical domains, see [9].

DEFINITION 4.4. Let $\mathcal{T}_h = \mathcal{T}_h^{\mathcal{G}} \times \mathcal{T}_d^{\mathcal{I}}$ be a prismatic partition of a cylindrical domain $\Omega = \mathcal{G} \times \mathcal{I} \subset \mathbb{R}^3$, where $\mathcal{T}_h^{\mathcal{G}}$ is a triangulation of a polygon $\mathcal{G} \subset \mathbb{R}^2$ and $\mathcal{T}_d^{\mathcal{I}}$ is a partition of an interval $\mathcal{I} \subset \mathbb{R}$. Further we denote by d_i, $i = 1, 2, \ldots, M$, the lengths of the M segments in $\mathcal{T}_d^{\mathcal{I}}$, by $T_{\max}$ and $T_{\min}$ the triangles in $\mathcal{T}_h^{\mathcal{G}}$ with the largest and smallest areas, respectively, and by $\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}}$ and $\alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}}$ the maximal and minimal angles in the triangulation $\mathcal{T}_h^{\mathcal{G}}$, respectively. We say that the prismatic partition $\mathcal{T}_h$ is *strictly well-shaped* for the DMP if $\alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} < \pi/2$ and if

$$\frac{1}{2}|T_{\max}| \tan \alpha_{\max}^{\mathcal{T}_h^{\mathcal{G}}} < d_i^2 < |T_{\min}| \tan \alpha_{\min}^{\mathcal{T}_h^{\mathcal{G}}} \quad \forall i = 1, 2, \ldots, M,$$

In the case $c = 0$ in $[0, \tau] \times \overline{\Omega}$, it can be easily shown that if the prismatic partition is strictly well-shaped then $\delta_U^{n,(\mathcal{P})}$ is positive and, thus, the crucial condition (23) can be satisfied by a suitable choice of θ and Δt. In the case if c does not vanish then a fine enough uniform refinement of any strictly well-shaped prismatic partition exists such that $\delta_U^{n,(\mathcal{P})} > 0$, see [9, Theorem 4].

As we already mentioned $\delta_L^{n,(\mathcal{P})} > 0$ and therefore the case $\theta = 0$ (the explicit Euler's method) is a priori excluded by (23). Interestingly, the following theorem shows that condition (23) limits the choice of θ even much more.

Theorem 4.5 *If condition (23) is satisfied then* $\max\limits_{\substack{n=0,\ldots,n_\tau-1 \\ \mathcal{P}\in\mathcal{T}_h}} \dfrac{5}{5 + \dfrac{\varrho_{\min}^{(n,\theta),(\mathcal{P})}}{\varrho_{\max}^{(n,\theta),(\mathcal{P})}}} \le \theta \le 1.$

PROOF. Let us fix a prism $\mathcal{P} \in \mathcal{T}_h$ and a time level $n \in \{0, 1, \ldots, n_\tau - 1\}$. For $\theta \neq 0$, inequalities (23) are equivalent to (24)–(26). Since $c \ge 0$, conditions (24)–(26) imply

$$\frac{|T|}{d^2} \ge \frac{\varrho_{\max}^{(n,\theta),(\mathcal{P})}}{\theta \Delta t} \frac{|T|}{6} + \frac{\cot \alpha_{\text{med}}^{(T)} + \cot \alpha_{\min}^{(T)}}{2}, \tag{30}$$

$$\frac{|T|}{d^2} \le 2 \cot \alpha_{\max}^{(T)} - \frac{\varrho_{\max}^{(n,\theta),(\mathcal{P})}}{\theta \Delta t} \frac{|T|}{3}, \tag{31}$$

$$\frac{|T|}{d^2} \le \frac{\varrho_{\min}^{(n,\theta),(\mathcal{P})}}{(1-\theta)\Delta t} \frac{|T|}{3} - \cot \alpha_{\text{med}}^{(T)} - \cot \alpha_{\min}^{(T)}. \tag{32}$$

Expressing $\dfrac{|T|}{\Delta t}$ from inequalities (31) and (32), we obtain

$$\frac{|T|}{d^2} \le -\frac{Q}{Q+R}(\cot \alpha_{\text{med}}^{(T)} + \cot \alpha_{\min}^{(T)}) + \frac{R}{Q+R} 2 \cot \alpha_{\max}^{(T)}, \tag{33}$$

where $Q = (1-\theta)/\varrho_{\min}^{(n,\theta),(\mathcal{P})}$ and $R = \theta/\varrho_{\max}^{(n,\theta),(\mathcal{P})}$. Similarly, a combination of (30) and (32) yields

$$(2R - Q)\frac{|T|}{d^2} \ge (Q+R)(\cot \alpha_{\text{med}}^{(T)} + \cot \alpha_{\min}^{(T)}). \tag{34}$$

If $2R - Q \leq 0$ held true then inequality (34) would imply that $\cot \alpha_{\text{med}}^{(T)} + \cot \alpha_{\text{min}}^{(T)} \leq 0$. Since cotangent is decreasing we would have $2 \cot \alpha_{\text{max}}^{(T)} \leq \cot \alpha_{\text{med}}^{(T)} + \cot \alpha_{\text{min}}^{(T)} \leq 0$ which is in contradiction with (31). Hence, $2R - Q > 0$ and (34) is equivalent to

$$\frac{|T|}{d^2} \geq \frac{Q+R}{2R-Q}(\cot \alpha_{\text{med}}^{(T)} + \cot \alpha_{\text{min}}^{(T)}). \tag{35}$$

Thus, (33) and (35) imply

$$\left(\frac{Q+R}{2R-Q} + \frac{Q}{Q+R}\right)(\cot \alpha_{\text{med}}^{(T)} + \cot \alpha_{\text{min}}^{(T)}) \leq \frac{R}{Q+R} 2 \cot \alpha_{\text{max}}^{(T)}.$$

Since cotangent is a decreasing function, we utilize the inequality $2 \cot \alpha_{\text{max}}^{(T)} \leq \cot \alpha_{\text{med}}^{(T)} + \cot \alpha_{\text{min}}^{(T)}$ to infer

$$\frac{Q+R}{2R-Q} + \frac{Q}{Q+R} \leq \frac{R}{Q+R}$$

which simplifies to $5Q \leq R$. The statement of the theorem now follows from the definition of Q and R. ■

Notice that in the most favorable case $\varrho = \text{const.}$ the smallest possible value of θ allowed by Theorem 4.5 is 5/6.

Remark 4.6 *Theorem 4.5 is sharp in the following sense. If $\theta = 5/6$ then there exists a prismatic partition and values of ϱ and c such that condition (23) is satisfied. Indeed, if $\theta = 5/6$, $\varrho = 1$, $c = 0$, $\cot \alpha_{\text{max}}^{(T)} = \cot \alpha_{\text{med}}^{(T)} = \cot \alpha_{\text{min}}^{(T)} = \pi/3$, $d^2 = \frac{3}{4}\sqrt{3}|T|$, $\Delta t = \frac{3}{5}\sqrt{3}|T|$ (the base triangulation consists of equilateral triangles with the same area $|T|$) then inequalities (23) hold as equalities. This is the only possibility how to satisfy condition (23) in the case $\theta = 5/6$.*

5 Numerical experiments

In this section, we will consider model problem (1)–(2) in case $c = 0$, $\varrho = 1$. For constant coefficients, the validity of DMP can be tested by studying one time-step of the discretization. DMP will be valid, if and only if the conditions stated in Theorem 3.2 are satisfied. For small systems these conditions can be tested by explicit construction of the inverse matrix $\mathbf{A}_0^{-1}$.

As Theorem 4.3 is the main contribution of this paper, our interest is to numerically verify this result. For this purpose, we consider a prismatic partition, with a base mesh consisting out of triangles with angles 65, 60, and 55 degrees, see Figure 2(a). The solution domain Ω has altitude 0.5 and its prismatic partition consists of five layers of prismatic elements, such that the altitude of each layer is $d = 0.1$. This partition is strictly well-shaped according to Definition 4.4 and it yields the DMP also for elliptic problems.

For such a partition, we have

$$\delta_L^{n,(\mathcal{P})} = 812 \quad \text{and} \quad \delta_U^{n,(\mathcal{P})} \doteq 73.759. \tag{36}$$

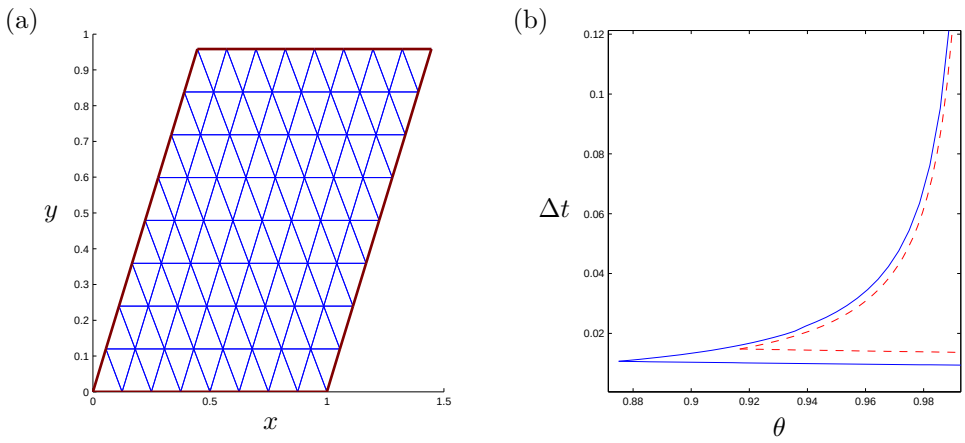

Figure 2: (a) The base mesh for the applied prismatic partition. (b) The upper and lower bounds for Δt as a function of θ. We verified computationally that the DMP is valid if and only if a point $(\Delta t, \theta)$ lies between the solid lines. Theorem 4.3 produces the dashed lines.

These two values define the smallest possible value for the parameter θ,

$$\theta \geq \frac{\delta_L^{n,(\mathcal{P})}}{\delta_L^{n,(\mathcal{P})} + \delta_U^{n,(\mathcal{P})}} \doteq 0.91673. \tag{37}$$

In Figure 2(b), we plot the (upper and lower) bounds for Δt as a function of θ. The dashed lines indicate the theoretical bounds obtained in Theorem 4.3 while the solid lines correspond to the computationally obtained bounds. Clearly, the theoretical bounds are more restrictive than the computational ones and the computed smallest acceptable value of θ is smaller than the value predicted by Theorem 4.3. An interesting phenomenon is, that when $\theta \to 1$, the longest acceptable timestep grows to infinity. This is clear in the light of formula (23). The lower bound for Δt stays very similar for each θ.

6 Conclusions

We analyzed the DMP for the linear parabolic problem (1)–(2) discretized by the lowest-order prismatic finite elements in space and by the θ-method in time. In Theorem 4.3 we obtained easily verifyable sufficient condition for the validity of the DMP. These conditions also limit the smallest possible value of the parameter θ, see Theorem 4.5. The performed numerical tests illustrate the sharpness of the sufficient conditions.

References

[1] M. Berzins, *Modified mass matrices and positivity preservation for hyperbolic and parabolic PDE's*, Commun. Numer. Meth. Engng. 17 (2001), pp. 659–666.

[2] I. Faragó, *Nonnegativity of the difference schemes*, Pure Math. Appl. 6 (1996), pp. 38–50.

[3] 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.

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

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

[6] I. Faragó, R. Horváth, *Continuous and discrete parabolic operators and their qualitative properties*, to appear in IMA J. Numer. Anal., 2008, 23pp.

[7] 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.

[8] 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.

[9] A. Hannukainen, S. Korotov, T. Vejchodský, *Discrete maximum principle for FE solutions of the diffusion-reaction problem on prismatic meshes*, in press, J. Comput. Appl. Math. (2008), doi:10.1016/j.cam.2008.08.02.

[10] R. Horváth, *On the positivity of iterative methods*, Annales Univ. Budapest. Sect. Comp., Budapest. 19 (2000), pp. 93–102.

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

[12] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer Verlag, 1981.

[13] Z. Zlatev, *Computer Treatment of Large Air Pollution Models*, Kluwer Academic Publishers, Dordrecht, 1995.