Helsinki University of Technology Institute of Mathematics Research Reports
Espoo 2008 A551

# DISCRETE MAXIMUM PRINCIPLES FOR THE FEM SOLUTION OF SOME NONLINEAR PARABOLIC PROBLEMS

István Faragó János Karátson 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 A551

# DISCRETE MAXIMUM PRINCIPLES FOR THE FEM SOLUTION OF SOME NONLINEAR PARABOLIC PROBLEMS

István Faragó János Karátson Sergey Korotov

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

**István Faragó, János Karátson, Sergey Korotov** : *Discrete maximum principles for the FEM solution of some nonlinear parabolic problems*; Helsinki University of Technology Institute of Mathematics Research Reports A551 (2008).

**Abstract:** *Discrete maximum principles for nonlinear parabolic problems and also associated (geometric) conditions on the meshes and time-steps in the FEM-type schemes are discussed.*

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

**Keywords:** Nonlinear parabolic problems, discrete maximum principle, finite element method

**Correspondence**

Department of Applied Analysis and Computational Mathematics, Eötvös Loránd University; H-1518, Budapest, Pf. 120, Hungary

Institute of Mathematics, Helsinki University of Technology
P.O. Box 1100, FI-02015 TKK, Finland

{farago, karatson}@cs.elte.hu, sergey.korotov@hut.fi

ISBN 978-951-22-9512-8 (print)
ISBN 978-951-22-0513-5 (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

The numerical approximations of models described by partial differential equations are naturally required to mirror some basic qualitative properties of the exact solutions. For parabolic equations, such a basic qualitative property is the (continuous) maximum principle (CMP). Several variants of CMPs exist, see e.g. [16, 26]. Its discrete analogues, the so-called discrete maximum principles (DMPs) for parabolic problems were first presented and analysed in the papers [17, 22]. If the finite element method (FEM) is employed for the spatial discretization, then the corresponding DMPs are normally ensured by imposing certain geometrical restrictions on the spatial meshes used, see, e.g., [10, 12, 17, 19] and the references therein. In addition, the time-steps have to be often chosen between certain lower and upper bounds. A related important discrete qualitative property of the numerical solutions is the so-called nonnegativity preservation, investigated e.g. in [8, 9]. The connection of nonnegativity preservation to DMPs is analysed e.g. in [7, 9, 10, 12].

In this paper, we prove discrete maximum principles for nonlinear parabolic problems, which has never been considered so far according to the authors' knowledge. The results are natural extensions of those in [20] (for nonlinear elliptic problems) and [13] (for linear parabolic problems).

The paper is organized as follows. In Section 2, we formulate the nonlinear parabolic problem. The discretization scheme is given in detail in Section 3. Some preliminaries on linear problems and the maximum principle are given in Section 4. The DMP and related nonnegativity preservation, and the conditions for their validity are presented in Section 5: we consider two types of growth conditions for the reaction terms, then we also discuss sufficient geometric conditions on the FE meshes used and finally give two relevant real-life examples.

# 2 The problem

In the sequel, we consider the following mixed nonlinear parabolic problem. Find a function $u = u(x,t)$ such that

$$\frac{\partial u}{\partial t} - \operatorname{div}\Big(k(x,t,u,\nabla u)\nabla u\Big) + q(x,t,u) = f(x,t) \quad \text{in } Q_T := \Omega \times (0,T), \tag{1}$$

where $\Omega$ is a bounded domain in $\mathbf{R}^d$ and $T > 0$. The boundary and initial conditions are given as

$$u(x,t) = g(x,t) \quad \text{for} \quad (x,t) \in \Gamma_D \times [0,T], \tag{2}$$

$$k(x,t,u,\nabla u)\tfrac{\partial u}{\partial \nu} + s(x,t,u) = \gamma(x,t) \quad \text{for} \quad (x,t) \in \Gamma_N \times [0,T], \tag{3}$$

$$u(x,0) = u_0(x) \quad \text{for} \quad x \in \Omega, \tag{4}$$

respectively. We impose the following

**Assumptions 2.1.**

(A1) $\Omega$ is a bounded polytopic domain in $\mathbf{R}^d$ with a Lipschitz continuous boundary $\partial\Omega$; $\Gamma_N, \Gamma_D \subset \partial\Omega$ are open sets, such that $\Gamma_N \cap \Gamma_D = \emptyset$ and $\overline{\Gamma}_N \cup \overline{\Gamma}_D = \partial\Omega$.

(A2) The scalar functions $k : \overline{Q}_T \times \mathbf{R}^{d+1} \to \mathbf{R}$, $\quad q : \overline{Q}_T \times \mathbf{R} \to \mathbf{R}$ and $s : \overline{\Gamma}_N \times [0,T] \times \mathbf{R} \to \mathbf{R}$ are measurable and bounded, further, $q$ and $s$ are continuously differentiable w.r.t. $t$, on their domains of definition. Further, $f \in L^\infty(Q_T)$, $\gamma \in L^2(\Gamma_N \times [0,T])$, $g \in L^\infty(\Gamma_D \times [0,T])$ and $u_0 \in L^\infty(\Omega)$.

(A3) There exist positive constants $\mu_0$ and $\mu_1$ such that

$$0 < \mu_0 \le k(x,t,\xi,\eta) \le \mu_1 \tag{5}$$

for all $(x,t,\xi,\eta) \in \Omega \times (0,T) \times \mathbf{R} \times \mathbf{R}^d$.

(A4) Let $2 \le p_1$ if $d = 2$, or $2 \le p_1 < \frac{2d}{d-2}$ if $d > 2$, further, let $2 \le p_2 < 2.5$ if $d = 2$ or 3 and $p_2 = 2$ if $d > 3$. There exist constants $\alpha_1, \alpha_2, \beta_1, \beta_2 \ge 0$ such that for any $x \in \Omega$ (or $x \in \Gamma_N$, resp.), $t \in (0,T)$ and $\xi \in \mathbf{R}$,

$$0 \le \frac{\partial q(x,t,\xi)}{\partial \xi} \le \alpha_1 + \beta_1 |\xi|^{p_1-2}, \qquad 0 \le \frac{\partial s(x,t,\xi)}{\partial \xi} \le \alpha_2 + \beta_2 |\xi|^{p_2-2}. \tag{6}$$

We define weak solutions in the usual way as follows. Let $H_D^1(\Omega) := \{u \in H^1(\Omega) : u_{|\Gamma_D} = 0\}$. A function $u : Q_T \to \mathbf{R}$ is called the weak solution of the problem (1)–(4) if $u$ is continuously differentiable with respect to $t$ and $u(.,t) \in H_D^1(\Omega)$ for all $t \in (0,T)$ and satisfies the relation

$$\int_\Omega \frac{\partial u}{\partial t} v\,dx + \int_\Omega \Big( k(x,t,u,\nabla u)\nabla u \cdot \nabla v + q(x,t,u)v \Big)\,dx + \int_{\Gamma_N} s(x,t,u)v\,d\sigma \tag{7}$$

$$= \int_\Omega fv\,dx + \int_{\Gamma_N} \gamma v\,d\sigma \qquad (\forall v \in H_D^1(\Omega), \quad t \in (0,T)),$$

further,

$$u = g \quad \text{on} \quad [0,T] \times \Gamma_D, \qquad u|_{t=0} = u_0 \quad \text{in} \quad \Omega, \tag{8}$$

Here and in the sequel, equality of functions in Lebesgue or Sobolev spaces is understood almost everywhere.

# 3 Discretization scheme

The discretization of problem (1)–(4) is built up in a standard way. The presentation below is the modification of that in [12] to the nonlinear case.

### 3.1 Semidiscretization in space

Let $\mathcal{T}_h$ be a finite element mesh over the solution domain $\Omega \subset \mathbf{R}^d$, where $h$ stands for the discretization parameter. We choose basis functions $\phi_1, ..., \phi_{\bar{m}}$, assumed to be continuous and to satisfy

$$\phi_i \geq 0 \quad (i = 1, \ldots, \bar{m}), \qquad \sum_{i=1}^{\bar{m}} \phi_i \equiv 1, \tag{9}$$

further, that there exist node points $P_i \in \overline{\Omega}$ $(i = 1, \ldots, \bar{m})$ such that

$$\phi_i(P_j) = \delta_{ij}, \tag{10}$$

where $\delta_{ij}$ is the Kronecker symbol. (These conditions hold e.g. for standard linear, bilinear or prismatic FEM.) Let $V_h$ denote the finite element subspace spanned by the above basis functions:

$$V_h = \text{span}\{\phi_1, ..., \phi_{\bar{m}}\} \ \subset H^1(\Omega).$$

Now, let $m < \bar{m}$ be such that

$$P_1, ..., P_m \tag{11}$$

are the vertices that lie in $\Omega$ or on $\Gamma_N$, and let

$$P_{m+1}, ..., P_{\bar{m}} \tag{12}$$

be the vertices that lie on $\overline{\Gamma}_D$. Then the basis functions $\phi_1, ..., \phi_m$ satisfy the homogeneous Dirichlet boundary condition on $\Gamma_D$, i.e., $\phi_i \in H^1_D(\Omega)$. We define

$$V_h^0 = \text{span}\{\phi_1, ..., \phi_m\} \ \subset H^1_D(\Omega).$$

Then the semidiscrete problem for (7) with initial-boundary conditions (8) reads as follows: find a function $u_h = u_h(x,t)$ such that

$$u_h(x,0) = u_0^h(x), \quad x \in \Omega,$$

$$u_h(.,t) - g_h(.,t) \in V_0^h, \quad t \in (0,T),$$

and

$$\int\limits_\Omega \frac{\partial u_h}{\partial t} v_h\, dx + \int\limits_\Omega \Big( k(x,t,u_h,\nabla u_h) \nabla u_h \cdot \nabla v_h + q(x,t,u_h) v_h \Big)\, dx + \int\limits_{\Gamma_N} s(x,t,u_h) v_h\, d\sigma$$

$$= \int\limits_\Omega f v_h\, dx + \int\limits_{\Gamma_N} \gamma v\, d\sigma \qquad (\forall v_h \in V_0^h, \quad t \in (0,T).) \tag{13}$$

In the above formulae, the functions $u_0^h$ and $g_h(.,t)$ (for any fixed $t$) are suitable approximations of the given functions $u_0$ and $g(.,t)$, respectively. In particular, we will use the following form to describe $g_h$:

$$g_h(x,t) = \sum_{i=1}^{m_\partial} g_i^h(t) \phi_{m+i}(x), \tag{14}$$

where

$$m_\partial := \bar{m} - m.$$

We note that, based on the consistency of the initial and boundary conditions ($g(s,0) = u_0(s), s \in \partial\Omega$), we obtain

$$g(P_{m+i}, 0) = u_0(P_{m+i}), \quad i = 1, \ldots, m_\partial.$$

We seek the numerical solution in the form

$$u_h(x,t) = \sum_{i=1}^{m} u_i^h(t)\phi_i(x) + g_h(x,t) \tag{15}$$

and notice that it is sufficient that $u_h$ satisfies (13) for $v_h = \phi_i$, $i = 1, 2, \ldots, m$, only. Then, introducing the notation

$$\mathbf{u}^h(t) = [u_1^h(t), \ldots, u_m^h(t), g_1^h(t), \ldots, g_{m_\partial}^h(t)]^T, \tag{16}$$

we are led to the following Cauchy problem for the system of ordinary differential equations:

$$\mathbf{M}\frac{\mathrm{d}\mathbf{u}^h}{\mathrm{d}t} + \mathbf{G}(\mathbf{u}^h(t)) = \mathbf{f}(t), \tag{17}$$

$$\mathbf{u}^h(0) = \mathbf{u}_0^h = [u_0(P_1), \ldots, u_0(P_m), g_1^h(0), \ldots, g_{m_\partial}^h(0)]^T, \tag{18}$$

where

$$\mathbf{M} = [M_{ij}]_{m\times\bar{m}}, \quad M_{ij} = \int_\Omega \phi_j(x)\phi_i(x)\,dx, \tag{19}$$

$$\mathbf{G}(\mathbf{u}^h(t)) = [G(\mathbf{u}^h(t))_i]_{i=1,\ldots,m},$$

$$G(\mathbf{u}^h(t))_i = \int\limits_\Omega \Big(k(x,t,u_h,\nabla u_h)\nabla u_h\cdot\nabla\phi_i + q(x,t,u_h)\phi_i\Big)\,dx + \int_{\Gamma_N} s(x,t,u_h)\phi_i\,d\sigma(x),$$

$$\mathbf{f}(t) = [f_i(t)]_{i=1,\ldots,m}, \quad f_i(t) = \int_\Omega f(x,t)\phi_i(x)\,dx + \int_{\Gamma_N} \gamma(x,t)\phi_i(x)\,d\sigma(x).$$

The solution $\mathbf{u}^h = \mathbf{u}^h(t)$ of problem (17)–(18) is called the semidiscrete solution. Its existence and uniqueness is ensured by Assumptions 2.1, since then $\mathbf{G}$ is locally Lipschitz continuous.

## 3.2 Full discretization

In order to get a fully discrete numerical scheme, we choose a time-step $\Delta t$ and denote the approximation to $\mathbf{u}^h(n\Delta t)$ and $\mathbf{f}(n\Delta t)$ by $\mathbf{u}^n$ and $\mathbf{f}^n$ (for $n = 0, 1, 2, \ldots, n_T$, where $n_T\Delta t = T$), respectively. To discretize (17) in time, we apply the so-called $\theta$-method with some given parameter

$$\theta \in (0,\,1].$$

We note that the case $\theta = 0$, which is otherwise also acceptable, will be excluded later by condition (53). This gives no strong difference, since the presence of $\mathbf{M}$ makes the scheme not explicit even for $\theta = 0$.

We then obtain a system of nonlinear algebraic equations of the form

$$\mathbf{M}\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + \theta \mathbf{G}(\mathbf{u}^{n+1}) + (1-\theta)\mathbf{G}(\mathbf{u}^n) = \mathbf{f}^{(n,\theta)} := \theta \mathbf{f}^{n+1} + (1-\theta)\mathbf{f}^n, \quad (20)$$

$n = 0, 1, \ldots, n_T - 1$, which can be rewritten as a recursion

$$\mathbf{M}\mathbf{u}^{n+1} + \theta \Delta t \mathbf{G}(\mathbf{u}^{n+1}) = \mathbf{M}\mathbf{u}^n - (1-\theta)\Delta t \mathbf{G}(\mathbf{u}^n) + \Delta t\, \mathbf{f}^{(n,\theta)} \quad (21)$$

with $\mathbf{u}^0 = \mathbf{u}^h(0)$. Furthermore, we will use notations

$$\mathbf{P}(\mathbf{u}^{n+1}) := \mathbf{M}\mathbf{u}^{n+1} + \theta \Delta t \mathbf{G}(\mathbf{u}^{n+1}), \qquad \mathbf{Q}(\mathbf{u}^n) := \mathbf{M}\mathbf{u}^n - (1-\theta)\Delta t \mathbf{G}(\mathbf{u}^n), \quad (22)$$

respectively. Then, the iteration procedure (21) can be also written as

$$\mathbf{P}(\mathbf{u}^{n+1}) = \mathbf{Q}(\mathbf{u}^n) + \Delta t\, \mathbf{f}^{(n,\theta)}. \quad (23)$$

We note that finding $\mathbf{u}^{n+1}$ in (23) requires the solution of a nonlinear algebraic system. The mass matrix $\mathbf{M}$ is positive definite, and it follows from Assumptions 2.1 that $\mathbf{u} \mapsto \mathbf{G}(\mathbf{u})$ has positive semidefinite derivatives. Therefore, by the definition in (22), the function $\mathbf{u} \mapsto \mathbf{P}(\mathbf{u})$ has regular derivatives. This ensures the unique solvability of (23) and, under standard local Lipschitz conditions on the coefficients, also the convergence of the damped Newton iteration, see e.g. [14].

# 4 Preliminaries: linear problems and the maximum principle

An important and widely studied special case of (1)–(4) is the linear problem with Dirichlet boundary conditions

$$\frac{\partial u}{\partial t} - k\Delta u + c(x,t)u = f(x,t), \quad (24)$$

$$u = g \quad \text{on} \quad [0,T] \times \partial\Omega, \qquad u|_{t=0} = u_0 \quad \text{in} \quad \Omega \quad (25)$$

where $k > 0$ is constant and $c \geq 0$. If the data and solution are assumed to be sufficiently smooth, then problem (24)–(25) is known to satisfy the *continuous maximum principle*, which important property is a starting point for our study:

$$\begin{aligned} &\min\{0; \min_{\Gamma_{t_1}} u\} + t_1 \min\{0; \min_{Q_{t_1}} f\} \leq u(x,t_1) \leq \\ &\leq \max\{0; \max_{\Gamma_{t_1}} u\} + t_1 \max\{0; \max_{Q_{t_1}} f\} \end{aligned} \quad (26)$$

for all $x \in \Omega$ and any fixed $t_1 \in (0, T)$, where $Q_{t_1} := \Omega \times [0, t_1]$, and $\Gamma_{t_1}$ denotes the parabolic boundary, i.e., $\Gamma_{t_1} := (\partial\Omega \times [0, t_1]) \cup (\Omega \times \{0\})$. A related property, which follows from the above [11], is the *continuous non-negativity preservation principle*: relations $f \geq 0$, $g \geq 0$ and $u_0 \geq 0$ imply

$$u(x,t) \geq 0 \tag{27}$$

for all $(x,t) \in Q_T$.

In the discrete case, the ODE system (17) now becomes a linear system

$$\mathbf{M}\frac{\mathrm{d}\mathbf{u}^h}{\mathrm{d}t} + \mathbf{K}\mathbf{u}^h(t) = \mathbf{f}, \tag{28}$$

where $\mathbf{K} = \int\limits_{\Omega} \Big( k\,\nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j \Big)$. The full discretization is

$$\mathbf{M}\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + \theta\mathbf{K}\mathbf{u}^{n+1} + (1-\theta)\mathbf{K}\mathbf{u}^n = \mathbf{f}^{(n,\theta)} := \theta\mathbf{f}^{n+1} + (1-\theta)\mathbf{f}^n. \tag{29}$$

Then (22)–(23) can be simplified: introducing the matrices

$$\mathbf{A} := \mathbf{M} + \theta\Delta t\mathbf{K}, \qquad \mathbf{B} := \mathbf{M} - (1-\theta)\Delta t\mathbf{K}, \tag{30}$$

equation (29) can now be rewritten as

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

To formulate the discrete maximum principle, let us define the following values:

$$g^n_{min} = \min\{0, g^n_1, \ldots, g^n_{m_\partial}\}, \quad g^n_{max} = \max\{0, g^n_1, \ldots, g^n_{m_\partial}\}, \tag{32}$$

$$u^n_{min} = \min\{0, g^n_{min}, u^n_1, \ldots, u^n_m\}, \quad u^n_{max} = \max\{0, g^n_{max}, u^n_1, \ldots, u^n_m\}, \tag{33}$$

for $n = 0, 1, \ldots, n_T$, and

$$f^{(n,n+1)}_{min} := \inf_{\substack{x\in\Omega,\\ \tau\in(n\Delta t,(n+1)\Delta t)}} f(x,\tau), \qquad f^{(n,n+1)}_{max} := \sup_{\substack{x\in\Omega,\\ \tau\in(n\Delta t,(n+1)\Delta t)}} f(x,\tau), \tag{34}$$

for $n = 0, 1, \ldots, n_T - 1$. If $f$ is only in $L^\infty(\Omega)$, then the above infima and suprema will mean essential infima and suprema, respectively. Then the discrete analogue of the continuous maximum principle (26) can be formulated as follows:

$$\begin{aligned} &\min\{0, g^{(n+1)}_{min}, u^{(n)}_{min}\} + \Delta t \min\{0, f^{(n,n+1)}_{min}\} \leq \\ &u^{n+1}_i \leq \max\{0, g^{(n+1)}_{max}, u^{(n)}_{max}\} + \Delta t \max\{0, f^{(n,n+1)}_{max}\}. \end{aligned} \tag{35}$$

This will be denoted by DMP and it corresponds to the continuous maximum principle for one time-level, i.e., when $t_1 \in [n\Delta t, (n+1)\Delta t]$.

It has been proved that the full discretization of the linear problem satisfies the DMP (35) in the following case:

**Theorem 4.1** [17, 12]. *Let the basis functions satisfy (9)–(10), and let the following conditions hold for the matrices (30):*

*(i)* $A_{ij} \le 0 \quad (i \ne j,\ \ i = 1, ..., m,\ j = 1, ..., \bar{m})$*;*

*(ii)* $B_{ii} \ge 0 \quad (i = 1, ..., m)$*.*

*Then the Galerkin solution of the problem (24)–(25), combined with the $\theta$-method in the time discretization, satisfies the discrete maximum principle (35).*

We note that in the original form, see e.g. [12, Thm. 6], it is also assumed that $K_{ij} \le 0 \quad (i \ne j, \quad i = 1, ..., m,\ j = 1, ..., \bar{m})$. However, now by our assumption $\theta > 0$, using (9) and (19) we have $M_{ij} \ge 0$, hence it follows from assumption (i) and (30) that $K_{ij} = (1/\theta\Delta t)(A_{ij} - M_{ij}) \le 0$.

The above result has been extended recently to mixed boundary value problems [13]. Let the boundary conditions in (25) be replaced by

$$\begin{gathered} u = g \quad \text{on} \quad [0,T] \times \Gamma_D, \qquad k\,\nabla u \cdot \nu = q \quad \text{on} \quad [0,T] \times \Gamma_N^0, \\ k\,\nabla u \cdot \nu + \sigma u = \varrho \quad \text{on} \quad [0,T] \times \Gamma_N^1, \end{gathered} \tag{36}$$

where $\sigma > 0$ is constant. If the conditions of Theorem 4.1 hold and $q \le 0$, then

$$u_i^{n+1} \le \ \max\{0, g_{max}^{(n+1)}, u_{max}^{(n)}\} + \Delta t \max\{0, f_{max}^{(n,n+1)}\} + \frac{1}{\theta} \max\{0, \Big(\frac{\varrho}{\sigma}\Big)_{max}^{(n,n+1)}\}. \tag{37}$$

In [13] a constant $\sigma$ is considered for simplicity, in which case $\sigma$ is simply a constant factor above and $\varrho_{max}^{(n,n+1)}$ is defined analogously to (34). However, their proof can be rewritten exactly in the same way for a variable coefficient $\sigma = \sigma(x, \tau)$, simply estimating $\varrho/\sigma$ by its suprema, in which case we have the DMP (37) with

$$\Big(\frac{\varrho}{\sigma}\Big)_{max}^{(n,n+1)} := \sup_{\substack{x \in \Gamma_N^1, \\ \tau \in (n\Delta t, (n+1)\Delta t)}} \frac{\varrho(x,\tau)}{\sigma(x,\tau)}. \tag{38}$$

**Remark 4.1** The indices $1, ..., m$ that arise in (33) now correspond to node points in the interior of $\Omega$ or on $\Gamma_N$, as in (11), and accordingly, the other $m_\partial$ indices involved in $g_{max}^{(n+1)}$ in (37) correspond to the values on $\Gamma_D$. That is, whereas the DMP (35) involves the values of $g$ on $\partial\Omega$, the DMP (37) involves the values of $g$ on $\Gamma_D$ only.

# 5 The discrete maximum principle for the nonlinear problem

## 5.1 Reformulation of the problem

We can rewrite problem (7) as follows. Let

$$r(x,t,\xi) := \int_0^1 \frac{\partial q}{\partial \xi}(x,t,\alpha\xi)\,d\alpha, \quad z(x,t,\xi) := \int_0^1 \frac{\partial s}{\partial \xi}(x,t,\alpha\xi)\,d\alpha \tag{39}$$

(for any $x \in \Omega$, $t > 0$, $\xi \in \mathbf{R}$),

$$\hat{f}(x,t) := f(x,t) - q(x,t,0), \qquad \hat{\gamma}(x,t) := \gamma(x,t) - s(x,t,0) \qquad (x \in \Omega,\ t > 0).$$

Then the Newton-Leibniz formula yields for all $x, t, \xi$ that

$$q(x,t,\xi) - q(x,t,0) = r(x,t,\xi)\,\xi, \qquad s(x,t,\xi) - s(x,t,0) = z(x,t,\xi)\,\xi.$$

Subtracting $q(x,t,0)$ and $s(x,t,0)$ from (1) and (3), respectively, we thus obtain that problem (7) is equivalent to

$$\int_\Omega \frac{\partial u}{\partial t} v\,dx + B(u;u,v) = \int_\Omega \hat{f} v\,dx + \int_{\Gamma_N} \hat{\gamma} v\,d\sigma \qquad (\forall v \in H^1_D(\Omega), \quad t \in (0,T)), \tag{40}$$

where

$$B(w;u,v) := \int_\Omega \Big( k(x,t,w,\nabla w)\nabla u \cdot \nabla v + r(x,t,w)uv \Big)\,dx + \int_{\Gamma_N} z(x,t,w)uv\,d\sigma \tag{41}$$

$(w,u,v \in H^1_D(\Omega))$.

Then the semidiscretization of the problem reads as follows: find a function $u_h = u_h(x,t)$ such that

$$u_h(x,0) = u_0^h(x), \quad x \in \Omega,$$

$$u_h(.,t) - g_h(.,t) \in V_0^h, \quad t \in (0,T),$$

and

$$\int_\Omega \frac{\partial u_h}{\partial t} v_h\,dx + B(u_h;u_h,v_h) = \int_\Omega \hat{f} v_h\,dx + \int_{\Gamma_N} \hat{\gamma} v_h\,d\sigma \quad (\forall v_h \in V_0^h, \quad t \in (0,T)).$$

Proceeding as in (15)–(17), the Cauchy problem for the system of ordinary differential equations (17) takes the following form:

$$\mathbf{M}\frac{\mathrm{d}\mathbf{u}^h}{\mathrm{d}t} + \mathbf{K}(\mathbf{u}^h)\mathbf{u}^h = \hat{\mathbf{f}}, \tag{42}$$

$$\mathbf{u}^h(0) = \mathbf{u}_0^h = [u_0(P_1), \ldots, u_0(P_m), g_1^h(0), \ldots, g_{m_\partial}^h(0)]^T, \tag{43}$$

where $\mathbf{M}$ is as in (17),

$$\mathbf{K}(\mathbf{u}^h) = [K(\mathbf{u}^h)_{ij}]_{m \times \bar{m}}, \quad K(\mathbf{u}^h)_{ij} = B(u_h;\phi_j,\phi_i),$$

$$\hat{\mathbf{f}}(t) = [\hat{f}_i(t)]_{i=1,\ldots,m}, \quad \hat{f}_i(t) = \int_\Omega \hat{f}(x,t)\phi_i(x)\,dx + \int_{\Gamma_N} \hat{\gamma}(x,t)\phi_i(x)\,d\sigma(x). \tag{44}$$

The full discretization reads as

$$\mathbf{M}\mathbf{u}^{n+1} + \theta\Delta t\mathbf{K}(\mathbf{u}^{n+1})\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n - (1-\theta)\Delta t\mathbf{K}(\mathbf{u}^n)\mathbf{u}^n + \Delta t\,\hat{\mathbf{f}}^{(n,\theta)}. \tag{45}$$

Since we have set $\mathbf{G}(\mathbf{u}^h) = \mathbf{K}(\mathbf{u}^h)\mathbf{u}^h$ in (17), the expressions (22)–(23) become

$$\mathbf{P}(\mathbf{u}^{n+1}) = \big(\mathbf{M}+\theta\Delta t\mathbf{K}(\mathbf{u}^{n+1})\big)\,\mathbf{u}^{n+1}, \qquad \mathbf{Q}(\mathbf{u}^n) = \big(\mathbf{M}-(1-\theta)\Delta t\,\mathbf{K}(\mathbf{u}^n)\big)\,\mathbf{u}^n,$$

respectively. Then, letting

$$\mathbf{A}(\mathbf{u}^h) := \mathbf{M}+\theta\Delta t\mathbf{K}(\mathbf{u}^h), \qquad \mathbf{B}(\mathbf{u}^h) := \mathbf{M}-(1-\theta)\Delta t\mathbf{K}(\mathbf{u}^h) \qquad (\mathbf{u}^h \in \mathbf{R}^{\bar{m}}), \tag{46}$$

the iteration procedure (45) takes the form

$$\mathbf{A}(\mathbf{u}^{n+1})\mathbf{u}^{n+1} = \mathbf{B}(\mathbf{u}^n)\mathbf{u}^n + \Delta t\;\hat{\mathbf{f}}^{(n,\theta)}, \tag{47}$$

which is similar to (31), but now the coefficient matrices depend on $\mathbf{u}^{n+1}$ resp. $\mathbf{u}^n$.

## 5.2 The DMP: problems with sublinear growth

Let us consider Assumptions 2.1, where we let $p_1 = p_2 = 2$ in assumption (A4), i.e. we have

*Assumption (A4')*: there exist constants $\alpha_1, \alpha_2 \geq 0$ such that for any $x \in \Omega$ (or $x \in \Gamma_N$, resp.), $t \in (0,T)$ and $\xi \in \mathbf{R}$,

$$0 \leq \frac{\partial q(x,t,\xi)}{\partial \xi} \leq \alpha_1, \qquad 0 \leq \frac{\partial s(x,t,\xi)}{\partial \xi} \leq \alpha_2. \tag{48}$$

In what follows, we will need the standard notion of (patch-)regularity of the considered meshes (cf. [3]).

**Definition 5.1** Let $\Omega \subset \mathbf{R}^d$ and let us consider a family of FEM subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$. The corresponding family of FE meshes will be called *regular* if there exist constants $c_0, c_1 > 0$ such that for any $h > 0$ and basis function $\phi_p$,

$$c_1 h^d \leq meas(\operatorname{supp}\phi_p), \quad diam(\operatorname{supp}\phi_p) \leq c_0 h \tag{49}$$

(where *meas* denotes $d$-dimensional measure and *supp* denotes the support, i.e. the closure of the set where the function does not vanish).

Note that the upper bound in (49) implies the following estimates for the corresponding supports and their boundaries:

$$meas(\operatorname{supp}\phi_p) \leq c_2 h^d \quad \text{and} \quad meas\big(\partial\big(\operatorname{supp}\phi_p\big)\big) \leq c_2 h^{d-1}. \tag{50}$$

**Theorem 5.1** *Let problem (1)–(4) satisfy Assumptions 2.1, such that we let $p_1 = p_2 = 2$ in (6), i.e. (A4) reduces to assumption (A4') above. us consider a family of finite element subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ such that the basis functions satisfy (9)–(10), and the family of associated FE meshes is regular as in Definition 5.1. Let the following assumptions hold:*

*(i) for any $i = 1, ..., m,\ j = 1, ..., \bar{m}\ (i \neq j)$, if $meas(\mathrm{supp}\,\phi_i \cap \mathrm{supp}\,\phi_j) > 0$ then*

$$\nabla\phi_i \cdot \nabla\phi_j \leq 0 \quad \text{on } \Omega \quad \text{and} \quad \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \leq -K_0\, h^{d-2} \tag{51}$$

*with some constant $K_0 > 0$ independent of $i, j$ and $h$;*

*(ii) the mesh parameter $h$ satisfies*

$$h < h_0 := \frac{2\mu_0 K_0}{c_2\alpha_2 + \sqrt{c_2^2\alpha_2^2 + 4\mu_0 K_0 c_2 \alpha_1}}; \tag{52}$$

*(iii) we have*

$$\Delta t \geq \frac{c_2 h^2}{\theta\big(\mu_0 K_0 - \alpha_1 c_2 h^2 - \alpha_2 c_2 h\big)}; \tag{53}$$

*(iv) if $\theta < 1$ then*

$$\Delta t \leq \frac{1}{(1-\theta)\, R(h)}, \tag{54}$$

*where*

$$R(h) := \max_{i=1,\dots,m} \frac{\left[\int\limits_\Omega \Big(\mu_1 |\nabla\phi_i|^2 + \alpha_1 \phi_i^2\Big) + \int\limits_{\Gamma_N} \alpha_2\, \phi_i^2\right]}{\int_\Omega \phi_i^2}. \tag{55}$$

*Then for all $\mathbf{u}^h \in \mathbf{R}^{\bar{m}}$, the matrices $\mathbf{A}(\mathbf{u}^h)$ and $\mathbf{B}(\mathbf{u}^h)$, defined in (46), have the following properties:*

*(1) $A(\mathbf{u}^h)_{ij} \leq 0 \quad (i \neq j,\ \ i = 1, ..., m,\ j = 1, ..., \bar{m})$;*

*(2) $B(\mathbf{u}^h)_{ii} \geq 0 \quad (i = 1, ..., m)$.*

Proof. (1) We have

$$A(\mathbf{u}^h)_{ij} := M_{ij} + \theta\Delta t K(\mathbf{u}^h)_{ij} = \int_\Omega \phi_j\phi_i \ + \ \theta\Delta t B(u_h; \phi_j, \phi_i) \tag{56}$$

$$= \int_\Omega \phi_j\phi_i + \theta\Delta t \left[\int\limits_\Omega \Big(k(x,t,u_h,\nabla u_h)\nabla\phi_j \cdot \nabla\phi_i + r(x,t,u_h)\phi_j\phi_i\Big) + \int\limits_{\Gamma_N} z(x,t,u_h)\phi_j\phi_i\right].$$

Let $\Omega_{ij} := \mathrm{supp}\,\phi_i \cap \mathrm{supp}\,\phi_j$ and $\Gamma_{ij} := \partial\Omega_{ij}$. Here, by (9) and (50),

$$\int_\Omega \phi_j\phi_i \leq meas(\Omega_{ij}) \leq c_2 h^d \quad \text{and} \quad \int_{\Gamma_N} \phi_j\phi_i \leq meas(\Gamma_{ij}) \leq c_2 h^{d-1}, \tag{57}$$

and similarly,

$$\int_\Omega r(x,t,u_h)\phi_j\phi_i \leq \alpha_1 c_2 h^d, \qquad \int\limits_{\Gamma_N} z(x,t,u_h)\phi_j\phi_i \leq \alpha_2 c_2 h^{d-1} \tag{58}$$

since by (39), $r$ and $z$ inherit (48). By (5) and (51),

$$\int_\Omega k(x,t,u_h,\nabla u_h)\nabla\phi_j\cdot\nabla\phi_i \le -\mu_0 K_0\, h^{d-2}\,. \tag{59}$$

Altogether, we obtain

$$A(\mathbf{u}^h)_{ij} \le c_2 h^d \left[1+\ \theta\Delta t\left(-\frac{\mu_0 K_0}{c_2}\,\frac{1}{h^2}\ +\alpha_1+\frac{\alpha_2}{h}\right)\right]\,.$$

Since $h < h_0$ for $h_0$ defined in (52), it readily follows that we have a negative coefficient of $\theta\Delta t$ above, and from (53) we obtain that the expression in the large brackets is nonpositive, hence $A(\mathbf{u}^h)_{ij} \le 0$.

(2) Analogously to (56), we have

$$B(\mathbf{u}^h)_{ii} := M_{ii} - (1-\theta)\Delta t K(\mathbf{u}^h)_{ii} \ge 0$$

if and only if

$$\int_\Omega \phi_i^2 \ge\ (1-\theta)\Delta t\left[\int_\Omega\Big(k(x,t,u_h,\nabla u_h)|\nabla\phi_i|^2 + r(x,t,u_h)\phi_i^2\Big) + \int_{\Gamma_N} z(x,t,u_h)\,\phi_i^2\right]\,.$$

The latter holds for all $\Delta t$ if

$$\theta = 1$$

(i.e. the scheme is implicit), and for all $\Delta t$ that satisfies (55) if $\theta < 1$. ■

Now we can derive the corresponding *discrete maximum principle:*

**Corollary 5.1** *Let the conditions of Theorem 5.1 hold, and let* $\hat\gamma(x,t) := \gamma(x,t) - s(x,t,0) \le 0$. *Then*

$$u_i^{n+1} \le\ \max\{0, g_{max}^{(n+1)}, u_{max}^{(n)}\} + \Delta t\max\{0, \hat f_{max}^{(n,n+1)}\}. \tag{60}$$

Proof. Our reformulated problem has the right-hand side $\hat f(x,t) := f(x,t) - q(x,t,0)$, which is in $L^\infty(Q_T)$ by Assumption 2.1 (A2). Further, by (40)–(41), we have the Neumann boundary condition

$$k(x,t,u,\nabla u)\,\nabla u\cdot\nu + z(x,t,u)u = \hat\gamma(x,t) \qquad \text{on}\ \ \Gamma_N,$$

where $z \ge 0$ and $\hat\gamma \le 0$. We can rewrite our boundary conditions to match (36): let $\Gamma_N^0$ and $\Gamma_N^1$ be the portions where $z \equiv 0$ and $z > 0$, respectively. Then, by assumption, $q := \hat\gamma_{|\Gamma_N^0} \le 0$ and $\varrho := \hat\gamma_{|\Gamma_N^1} \le 0$. Therefore (37) can be applied (with $\hat f$) and its last term can be dropped, whence we obtain (60).
■

**Remark 5.1** Note that the DMP (60) involves the values of $g$ on $\Gamma_D$, see also Remark 4.1. Besides that, (60) is formally identical to the upper part of (35), and could in fact be derived from it directly as an alternate proof. Namely, one can apply Theorem 4.1 as an algebraic result for the ODE system (42). Here $\mathbf{f}$ is replaced by $\hat{\mathbf{f}}$ that also involves the values of $\hat{\gamma}$, see (44). However, by our assumption $\hat{\gamma} \leq 0$, we obtain a further upper bound by dropping the integrals with $\hat{\gamma}$, and we are thus led to (60).

**Remark 5.2** *(Discussion of the assumptions in Theorem 5.1.)*

(i) Assumption (i) can be ensured by suitable geometric properties of the space mesh, see subsection 5.4 below.

(ii) The value of $h_0$ contains given or computable constants from the assumptions on the coefficients, the mesh regularity and geometry.

(iii) The lower bound in (53) is asymptotically

$$\Delta t \geq O(h^2) \tag{61}$$

as $h \to 0$, and the constants are similarly computable.

(iv) If $\theta = 1$, i.e. the scheme is implicit, then there is no upper restriction on $\Delta t$. If $\theta < 1$, then it can be often proved (e.g. for popular simplicial, bilinear and prismatic elements) that $R(h) = O(h^{-2})$ in (55), hence $\Delta t \geq O(h^2)$ as $h \to 0$, which yields with (61) the usual condition

$$\Delta t = O(h^2) \tag{62}$$

(as $h \to 0$) for the space and time discretizations. In addition, the lower bound in (53) must be smaller than the upper bound in (54): in view of the factor $1 - \theta$ in the latter, this gives a restriction on $\theta$ to be close enough to 1.

**Remark 5.3** Let us consider problem (1)–(4) with principal parts only, i.e. when $q \equiv s \equiv 0$:

$$\frac{\partial u}{\partial t} - \operatorname{div}\Big(k(x,t,u,\nabla u)\nabla u\Big) = f(x,t) \quad \text{in} \quad Q_T := \Omega \times (0,T),$$

$$u(x,t) = g(x,t) \quad \text{for} \quad (x,t) \in \Gamma_D \times [0,T],$$
$$k(x,t,u,\nabla u)\tfrac{\partial u}{\partial \nu} = \gamma(x,t) \quad \text{for} \quad (x,t) \in \Gamma_N \times [0,T],$$
$$u(x,0) = u_0(x) \quad \text{for} \quad x \in \Omega,$$

Then Assumptions (ii)-(iv) of Theorem 5.1 become much simplified, since $\alpha_1 = \alpha_2 = 0$. Namely, assumption (ii) is dropped since formally $h_0 = \infty$, i.e. there is no upper bound on $h$. Assumptions (iii)-(iv) read as follows:

$$\Delta t \geq \frac{c_2}{\theta \mu_0 K_0} h^2; \qquad \text{if} \ \ \theta < 1 \ \ \text{then} \ \ \Delta t \leq \frac{1}{\mu_1 (1-\theta)} \min_{i=1,\dots,m} \frac{\int_\Omega \phi_i^2}{\int\limits_\Omega |\nabla \phi_i|^2} \, . \tag{63}$$

Let us now return to the statement (60). By reversing signs in Corollary 5.1, we obtain the corresponding discrete minimum principle:

**Corollary 5.2** *Let the conditions of Theorem 5.1 hold, and let* $\hat{\gamma}(x,t) := \gamma(x,t) - s(x,t,0) \geq 0$. *Then*

$$u_i^{n+1} \geq \min\{0, g_{min}^{(n+1)}, u_{min}^{(n)}\} + \Delta t \min\{0, \hat{f}_{min}^{(n,n+1)}\}. \tag{64}$$

An important special case is the discrete *nonnegativity preservation principle*, the discrete analogue of (27):

**Theorem 5.2** *Let the conditions of Theorem 5.1 hold, and let* $\hat{f} \geq 0$, $g \geq 0$, $\hat{\gamma} \geq 0$ *and* $u_0 \geq 0$. *Then the discrete solution satisfies*

$$u_i^n \geq 0 \qquad (n = 0, 1, ..., n_T, \ i = 1, ..., m).$$

PROOF. Assumptions $\hat{f} \geq 0$, $g \geq 0$ and $\hat{\gamma} \geq 0$ imply $g_{min}^{(n+1)} \geq 0$ and $\hat{f}_{min}^{(n,n+1)}$ for all $n$ and $i$, hence (64) becomes

$$u_i^{n+1} \geq \min\{0, u_{min}^{(n)}\}.$$

Here assumption $u_0 \geq 0$ implies $u_{min}^{(0)} \geq 0$, hence we obtain by induction that $u_{min}^{(n)} \geq 0$ for all $n$. ■

By Theorem 5.2, $u^h$ is nonnegative in each node point. Properties (9)–(10) of the basis functions imply that the FEM solution $u^h(., n\Delta t)$ is also nonnegative for all time levels $n\Delta t$. If, in addition, we extend the solutions to $Q_T$ with values between those on the neighbouring time levels, e.g. with the method of lines, then we obtain that the discrete solution satisfies

$$u^h \geq 0 \quad \text{on} \quad Q_T.$$

## 5.3 The DMP: problems with superlinear growth

In this subsection we allow stronger growth of the nonlinearities $q$ and $s$ than in the above, i.e. we return to Assumption 2.1 (A4). For this we need some extra technical assumptions and results.

Let us first summarize the additional conditions.

**Assumptions 5.3.**

(B1) We restrict ourselves to the case of implicit scheme:

$$\theta = 1.$$

(B2) $V_h$ is made by linear , bilinear or prismatic elements.

(B3) The coefficient on $\Gamma_N$ satisfies $\hat{\gamma}(x,t) := \gamma(x,t) - s(x,t,0) \equiv 0$, further, $\Gamma_D \neq \emptyset$.

(B4) The exact solution satisfies $u(.,t) \in W^{1,q}(\Omega)$ for some $q > 2$ (if $d = 2$) or some $q \geq 2d/(d-(d-2)(p_1-2))$ (if $d \geq 3$) for all $t \in [0,T]$.

(B5) The discretization satisfies $M_{p_1} := \sup_{t\in[0,T]} \|u(.,t) - u_h(.,t)\|_{L^{p_1}(\Omega)} < \infty$.

Now, by [1], under Assumption 2.1 (A4), we recall the Sobolev embedding estimates

$$\|v\|_{L^{p_1}(\Omega)} \leq C_{\Omega,p_1}\|v\|_{H^1_D}, \quad \|v\|_{L^{p_2}(\Gamma_N)} \leq C_{\Gamma_N,p_2}\|v\|_{H^1_D} \quad (\forall v \in H^1_D(\Omega)) \tag{65}$$

with some constants $C_{\Omega,p_1}, C_{\Gamma_N,p_2} > 0$ independent of $v$.

**Lemma 5.1** *Let $V_h$ be made by linear , bilinear or prismatic elements. Then there exists a constant $c_{p_2} > 0$ such that*

$$\|v\|_{L^{p_2}(\Gamma_N)} \leq c_{p_2} h^{-1}\|v\|_{L^2(\Omega)} \qquad (v \in V_h). \tag{66}$$

Proof. We have

$$\|v\|^2_{H^1_D} := \int_\Omega |\nabla v|^2 \leq \int_\Omega v^2 \max_{v\in V_h} \frac{\int_\Omega |\nabla v|^2}{\int_\Omega v^2} \leq const. \cdot R(h) \int_\Omega v^2\,,$$

where $R(h)$ comes from (55) and, as seen before, satisfies $R(h) = O(h^{-2})$. This, combined with (65), yields the required estimate. ■

Now we consider the full discretization (45) for $\theta = 1$:

$$\mathbf{M}\mathbf{u}^{n+1} + \Delta t\mathbf{K}(\mathbf{u}^{n+1})\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n + \Delta t\, \hat{\mathbf{f}}^{(n)}. \tag{67}$$

Let $u^{n+1} \in V_h$ denote the function with coefficient vector $\mathbf{u}^{n+1}$, and let $f^n(x) := f(x, n\Delta t)$. Then, by the definition of the mass and stiffness matrices, (67) implies

$$\int_\Omega u^{n+1}v + \Delta t\, B(u^{n+1}; u^{n+1}, v) = \int_\Omega u^n v + \Delta t\Big(\int_\Omega \hat{f}^n v + \int_{\Gamma_N} \hat{\gamma}^n v\Big) \qquad (v \in V_h). \tag{68}$$

Here, by assumption (B3), the integral on $\Gamma_N$ vanishes, further, recall that $\hat{f} \in L^\infty(Q_T)$ by Assumption 2.1 (A2).

**Lemma 5.2** *If Assumptions 5.3 hold, then for all $t \in [0,T]$*

$$\|u(.,t)\|_{L^{p_1}(\Omega)} \leq \|u^0\|_{L^{p_1}(\Omega)} + T(meas(\Omega))^{\frac{1}{p_1}}\|\hat{f}\|_{L^\infty(Q_T)}\,,$$

*wherein the r.h.s. is independent of $t$.*

Proof. Let $v = |u|^{p_1-2}u$, which satisfies $\nabla v = (p_1-1)|u|^{p_1-2}\nabla u$. By assumption (B4), $|\nabla u| \in L^q(\Omega)$, and it is easy to see from the condition on $q$ that $|u|^{p_1-2} \in L^{q'}(\Omega)$ where $(1/q)+(1/q') = 1/2$; these imply by Hölder's inequality that $|\nabla v| \in L^2(\Omega)$. That is, for all fixed $t$ we have $v(.,t) \in H^1_D(\Omega)$, hence we can set it in (40):

$$\int_\Omega \frac{\partial u}{\partial t}\Big(|u|^{p_1-2}u\Big)\,dx + B(u;u,|u|^{p_1-2}u) = \int_\Omega \hat{f}|u|^{p_1-2}u\,dx \qquad (\forall v \in H^1_D(\Omega), \quad t \in (0,T)), \tag{69}$$

where we have used $\hat{\gamma} \equiv 0$. Let

$$N(t) := \|u(.,t)\|^{p_1}_{L^{p_1}(\Omega)} = \int_\Omega |u(x,t)|^{p_1}\,dx\,,$$

then $N'(t) = \int_\Omega p_1\,|u|^{p_1-2}u\,\frac{\partial u}{\partial t}\,dx$. Further, using (41) and that $\nabla v = (p_1 - 1)|u|^{p_1-2}\nabla u$, we obtain

$$B(u;u,|u|^{p_1-2}u) = \int_\Omega \Big(k(x,t,u,\nabla u)(p_1-1)|u|^{p_1-2}|\nabla u|^2$$

$$+r(x,t,u)|u|^{p_1}\Big)\,dx + \int_{\Gamma_N} z(x,t,u)|u|^{p_1}\,d\sigma \geq 0$$

hence the left-hand side of (69) is estimated below by $N'(t)/p_1$. Using Hölder's inequality for the right-hand side of (69), we then obtain

$$\frac{1}{p_1}N'(t) \leq \|\hat{f}(.,t)\|_{L^{p_1}(\Omega)}\|u(.,t)\|^{p_1-1}_{L^{p_1}(\Omega)} \leq (meas(\Omega))^{\frac{1}{p_1}}\|\hat{f}\|_{L^\infty(Q_T)}\,N(t)^{\frac{p_1-1}{p_1}}\,.$$

Excluding the trivial case $u \equiv 0$, we can divide by $N(t)^{\frac{p_1-1}{p_1}}$ and integrate from 0 to $t$ to obtain

$$N(t)^{\frac{1}{p_1}} - N(0)^{\frac{1}{p_1}} \leq T(meas(\Omega))^{\frac{1}{p_1}}\|\hat{f}\|_{L^\infty(Q_T)},$$

which is the desired estimate. ■

**Lemma 5.3** *(1) If Assumptions 5.3 (B1) and (B3) hold, then the norms* $\|u^n\|_{L^2(\Omega)}$ *are bounded, independently of* $n$ *and* $V_h$*, by the constant* $K_{L_2} := \|u^0\|_{L^2(\Omega)} + T(meas(\Omega))^{\frac{1}{2}}\|\hat{f}\|_{L^\infty(Q_T)}$.

*(2) If all Assumptions 5.3 hold, then the norms* $\|u^n\|_{L^{p_1}(\Omega)}$ *are bounded, independently of* $n$ *and* $V_h$*, by the constant* $K_{p_1,\Omega} := M_{p_1} + \|u^0\|_{L^{p_1}(\Omega)} + T(meas(\Omega))^{\frac{1}{p_1}}\|\hat{f}\|_{L^\infty(Q_T)}$.

Proof. (1) Setting $v = u^{n+1}$ in (68), we obtain

$$\int_\Omega (u^{n+1})^2 + \Delta t\, B(u^{n+1}; u^{n+1}, u^{n+1}) = \int_\Omega u^n u^{n+1} + \Delta t \int_\Omega \hat{f}^n u^{n+1}. \qquad (70)$$

To estimate below, the bilinear form can be dropped from the l.h.s. since it is coercive, and also using Cauchy-Schwarz inequalities, we have

$$\|u^{n+1}\|^2_{L^2(\Omega)} \le \|u^n\|_{L^2(\Omega)} \|u^{n+1}\|_{L^2(\Omega)} + \Delta t \|\hat{f}^n\|_{L^2(\Omega)} \|u^{n+1}\|_{L^2(\Omega)}.$$

Dividing by $\|u^{n+1}\|_{L^2(\Omega)}$ and repeating the argument $n$ times, we obtain

$$\|u^{n+1}\|_{L^2(\Omega)} \le \|u^0\|_{L^2(\Omega)} + (n+1)\Delta t \|\hat{f}^n\|_{L^2(\Omega)},$$

where the r.h.s. is bounded since $(n+1)\Delta t \le T$ and $\|\hat{f}^n\|_{L^2(\Omega)} \le (meas(\Omega))^{\frac{1}{2}} \|\hat{f}\|_{L^\infty(Q_T)}$.

(2) It follows directly from Lemma 5.2 and assumption (B5). ∎

Lemmas 5.1 and 5.3 imply

**Corollary 5.3** *We have*

$$\|u^n\|_{L^{p_2}(\Gamma_N)} \le K_{p_2,\Gamma_N} h^{-1}$$

*where the constant $K_{p_2,\Gamma_N} > 0$ is bounded independently of $n$ and $V_h$.*

**Theorem 5.3** *Let problem (1)–(4) satisfy Assumptions 2.1 and Assumptions 5.3. Let us consider a family of finite element subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ such that the family of associated FE meshes is regular as in Definition 5.1. Let the following assumptions hold:*

*(i) for any $i = 1, ..., m,\ j = 1, ..., \bar{m}$ $(i \ne j)$, if $meas(\operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j) > 0$ then*

$$\nabla\phi_i \cdot \nabla\phi_j \le 0 \quad on\ \Omega \quad and \quad \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \le -K_0\, h^{d-2} \qquad (71)$$

*with some constant $K_0 > 0$ independent of $i, j$ and $h$;*

*(ii) the mesh parameter $h$ satisfies $h < h_0$, where $h_0 > 0$ is the first positive root of the equation*

$$-\frac{\mu_0 K_0}{c_2}\frac{1}{h^2} + \alpha_1 + \frac{\alpha_2}{h} + \frac{\beta_1 K^{p_1-2}_{p_1,\Omega}}{h^{\gamma_1}} + \frac{\beta_2 K^{p_2-2}_{p_2,\Gamma_N}}{h^{\gamma_2}} = 0, \qquad (72)$$

*where the numbers $0 < \gamma_1, \gamma_2 < 2$ are defined below in (74), (75), respectively;*

*(iii) we have*

$$\Delta t \ge \frac{c_2 h^2}{\theta\big(\mu_0 K_0 2 - c_2\alpha_1 h^2 + c_2\alpha_2 h - c_2\beta_1 K^{p_1-2}_{p_1,\Omega} h^{2-\gamma_1} - c_2\beta_2 K^{p_2-2}_{p_2,\Gamma_N} h^{2-\gamma_2}\big)}. \qquad (73)$$

*Then the matrices* $\mathbf{A}(\mathbf{u}^{n+1})$ *and* $\mathbf{B}(\mathbf{u}^n)$*, defined in (46)–(47), have the following properties:*

*(1)* $A(\mathbf{u}^{n+1})_{ij} \le 0 \quad (i \neq j,\ \ i = 1, ..., m,\ j = 1, ..., \bar{m})$*;*

*(2)* $B(\mathbf{u}^n)_{ii} \ge 0 \quad (i = 1, ..., m)$*.*

Proof. We follow the proof of Theorem 5.1. As a first difference, instead of $u_h$ in the arguments, we must consider the functions $u^{n+1}$ (for $\mathbf{A}$) and $u^n$ (for $\mathbf{B}$) that have the coefficient vectors $\mathbf{u}^{n+1}$ and $\mathbf{u}^n$, respectively.

(1) Since we now have (6) instead of (48), the first estimate in (58) is replaced by

$$\int_\Omega r(x,t,u^{n+1})\phi_j\phi_i \le \int_\Omega \big(\alpha_1+\beta_1|u^{n+1}|^{p_1-2}\big)\phi_j\phi_i \le \alpha_1 meas(\Omega_{ij})+\beta_1 \int_{\Omega_{ij}} |u^{n+1}|^{p_1-2}\,.$$

Here the first term is bounded by $\alpha_1 c_2 h^d$ as before. To estimate the second term, we use Hölder's inequality:

$$\int_{\Omega_{ij}} |u^{n+1}|^{p_1-2} \le \|u^{n+1}\|^{p_1-2}_{L^{p_1}(\Omega_{ij})} \|1\|^2_{L^{p_1}(\Omega_{ij})}.$$

For the first factor, we use Lemma 5.3 (2) to find that

$$\|u^{n+1}\|^{p_1-2}_{L^{p_1}(\Omega_{ij})} \le \|u^{n+1}\|^{p_1-2}_{L^{p_1}(\Omega)} \le K^{p_1-2}_{p_1,\Omega}\,.$$

The second factor satisfies, by (57),

$$\|1\|^2_{L^{p_1}(\Omega_{ij})} = \big(meas(\Omega_{ij})\big)^{2/p_1} \le c_2 h^{\frac{2d}{p_1}} \equiv c_2 h^{d-\gamma_1}$$

with

$$\gamma_1 := d - \frac{2d}{p_1} < 2, \tag{74}$$

since from Assumption 2.1 (A4) we have $\frac{2d}{p_1} > d-2$. Hence

$$\int_{\Omega_{ij}} |u^{n+1}|^{p_1-2} \le K^{p_1-2}_{p_1,\Omega} c_2 h^{d-\gamma_1}$$

and altogether,

$$\int_\Omega r(x,t,u^{n+1})\phi_j\phi_i \le \alpha_1 c_2 h^d + \beta_1 K^{p_1-2}_{p_1,\Omega} c_2 h^{d-\gamma_1}\,.$$

Similarly,

$$\int_{\Gamma_N} z(x,t,u^{n+1})\phi_j\phi_i \le \alpha_2 c_2 h^{d-1} + \beta_2 \int_{\Gamma_{ij}} |u^{n+1}|^{p_2-2}$$

and here, for $d = 2, 3$ we use Corollary 5.3 and (50) to have

$$\int_{\Gamma_{ij}} |u^{n+1}|^{p_2-2} \le \|u^{n+1}\|^{p_2-2}_{L^{p_2}(\Gamma_{ij})} \|1\|^2_{L^{p_2}(\Gamma_{ij})} \le \|u^{n+1}\|^{p_2-2}_{L^{p_2}(\Gamma_N)} \big(meas(\Gamma_{ij})\big)^{2/p_2}$$

$$\le K^{p_2-2}_{p_2,\Gamma_N} c_2 h^{2-p_2+\frac{2(d-1)}{p_2}} \equiv K^{p_2-2}_{p_2,\Gamma_N} c_2 h^{d-\gamma_2},$$

where

$$\gamma_2 := d - 2 + p_2 - \frac{2(d-1)}{p_2} < 2 \qquad (75)$$

from assumption $p_2 \le 2.5$. Summing up, using the above and (59), we obtain

$$A(\mathbf{u}^h)_{ij} \le c_2 h^d \left[ 1 + \ \theta \Delta t \left( -\frac{\mu_0 K_0}{c_2} \frac{1}{h^2} + \alpha_1 + \frac{\alpha_2}{h} + \frac{\beta_1 K^{p_1-2}_{p_1,\Omega}}{h^{\gamma_1}} + \frac{\beta_2 K^{p_2-2}_{p_2,\Gamma_N}}{h^{\gamma_2}} \right) \right].$$

Since $h < h_0$ for $h_0$ defined in (72), it follows that we have a negative coefficient of $\theta\Delta t$ above, and from (73) we obtain that the expression in the large brackets is nonpositive, hence $A(\mathbf{u}^h)_{ij} \le 0$.

(2) For the implicit scheme, $B(\mathbf{u}^n)$ coincides with the mass matrix $M$, whose diagonal entries are positive. ∎

Similarly to the sublinear case, we can derive the corresponding discrete maximum, minimum and nonnegativity preservation principles. We only formulate here the latter:

**Corollary 5.4** *Let the conditions of Theorem 5.3 hold, and let $\hat{f} \ge 0$, $g \ge 0$, $\hat{\gamma} \ge 0$ and $u_0 \ge 0$. Then the discrete solution satisfies*

$$u_i^n \ge 0 \qquad (n = 0, 1, ..., n_T, \ i = 1, ..., m).$$

## 5.4 Geometric properties of the space mesh

In order to satisfy condition (71), the most direct way is to require

$$\nabla\phi_i \cdot \nabla\phi_j \le -K_0\, h^{-2} \qquad (76)$$

pointwise on the common support of these basis functions. In view of well-known formulae (see e.g. [2, 5, 25, 27]), the above condition has a nice geometric interpretation: in the case of simplicial meshes, it is sufficient if the employed mesh is uniformly acute [4, 25]. In the case of bilinear elements, condition (76) is equivalent to the so-called strict non-narrowness of the meshes, see [12, 19]. The case of prismatic finite elements is treated in [18].

These conditions are sufficient but not necessary. For instance, for linear elements, some obtuse interior angles may occur in the simplices of the meshes, just as for linear problems (see e.g. [24]), or one can require (76) only on a proper subpart of each intersection of supports with asymptotically nonvanishing measure, see more details in [21]. These weaker conditions may allow in general easier refinement procedures.

## 5.5 Examples

We give two real-life examples where discrete nonnegativity can be derived for suitable discretizations.

**(a) Nonlinear heat conduction.**

Heat conduction in a body $\Omega \subset \mathbf{R}^3$ with nonlinear diffusion coefficient is often described by the model

$$\frac{\partial u}{\partial t} - \operatorname{div}\Big(k(x,t,u)\nabla u\Big) = f(x,t) \tag{77}$$

in $Q_T := \Omega \times (0,T)$, where $T > 0$ is the time interval considered; see, e.g., [15]. The usual boundary and initial conditions are

$$u(x,t) = g(x,t) \quad \text{for} \quad (x,t) \in \Gamma_D \times [0,T], \tag{78}$$

$$k(x,t,u)\tfrac{\partial u}{\partial \nu} = \gamma(x,t) \quad \text{for} \quad (x,t) \in \Gamma_N \times [0,T], \tag{79}$$

$$u(x,0) = u_0(x) \quad \text{for} \quad x \in \Omega, \tag{80}$$

where all coefficients are bounded nonnegative measurable functions and $k$ has a positive lower bound. The function $u$ describes the temperature, hence $u \geq 0$.

**(b) Reaction-diffusion problems.**

A reaction-diffusion process in a body $\Omega \subset \mathbf{R}^d$, $d = 2$ or 3, is often described by the model

$$\frac{\partial u}{\partial t} - \operatorname{div}\Big(k(x,t)\nabla u\Big) + q(x,u) = f(x,t) \tag{81}$$

in $Q_T := \Omega \times (0,T)$. The boundary and initial conditions are

$$u(x,t) = g(x,t) \quad \text{for} \quad (x,t) \in \Gamma_D \times [0,T], \tag{82}$$

$$k(x,t)\tfrac{\partial u}{\partial \nu} + s(x,u) = \gamma(x,t) \quad \text{for} \quad (x,t) \in \Gamma_N \times [0,T], \tag{83}$$

$$u(x,0) = u_0(x) \quad \text{for} \quad x \in \Omega, \tag{84}$$

The function $u$ describes the temperature, hence

$$u \geq 0.$$

Here the coefficients $k$, $f$, $g$, $\gamma$ and $u_0$ are bounded nonnegative measurable functions and $k$ has a positive lower bound. Further, $q$ and $s$ desribe the rate of reaction in the body and on the transmission boundary, respectively, hence $q(x,0) = s(x,0) = 0$ for all $x$. In various examples the reaction process is such that $q$ and $s$ grow along with $u$, further, the rate is at most polynomial, i.e. we may assume that the growth conditions (6) are satisfied. For instance, $q(x,u) = u^\sigma$ for some $\sigma > 1$ in some autocatalytic chemical reactions, or

$q(x,u) = \frac{1}{\varepsilon}\frac{u}{u+\kappa}$ describes the Michaelis-Menten reaction in enzyme kynetics [6, 23].

In both examples, we have $\hat{f} = f \geq 0, \quad g \geq 0, \quad \hat{\gamma} = \gamma \geq 0$ and $u_0 \geq 0$. Therefore we can use Theorem 5.2 and Corollary 5.4, respectively, to derive the discrete nonnegativity principle:

**Theorem 5.4** *Let the full discretization satisfy the conditions of Theorem 5.1 for problem (77)–(80), or the conditions of Theorem 5.3 for problem (81)–(84). Then the discrete solution satisfies*

$$u_i^n \geq 0 \qquad (n = 0, 1, ..., n_T,\ i = 1, ..., m).$$

In particular, for problem (77)–(80) we can use the simplified assumptions (63) for Theorem 5.1, as given in Remark 5.3.

Consequently, as pointed out after Theorem 5.2, if we extend the solutions to $Q_T$ with values between those on the neighbouring time levels, e.g. with the method of lines, then the discrete solution satisfies

$$u^h \geq 0 \quad \text{on} \quad Q_T.$$

# References


[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York-London, 1975.

[2] J. Brandts, S. Korotov, M. Krízek, *Dissection of the Path-Simplex in* $\mathbf{R}^n$ *into n Path-Subsimplices*, Linear Algebra Appl. 421 (2007), pp. 382–393.

[3] J. Brandts, S. Korotov, M. Krízek, *On the Equivalence of Regularity Criteria for Triangular and Tetrahedral Finite Element Partitions*, Comput. Math. Appl. 55 (2008), pp. 2227–2233.

[4] J. Brandts, S. Korotov, M. Krízek, J. Solc, *On Nonobtuse Simplicial Partitions*, SIAM Rev. (in press).

[5] 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.

[6] J. I. Díaz, *Applications of Symmetric Rearrangement to Certain Nonlinear Elliptic Equations with a Free Boundary*, In: *Nonlinear differential equations* (Granada, 1984), 155–181, Res. Notes in Math., 132, Pitman, Boston, MA, 1985.

[7] M. Elshebli, *Maximum Principle and Non-Negativity Preservation in Linear Parabolic Problems*, Annales Univ. Sci. Sec. Math. 47 (2006), pp. 643–653.

[8] 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.

[9] 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.

[10] I. Faragó, R. Horváth, *Discrete Maximum Principle and Adequate Discretizations of Linear Parabolic Problems*, SIAM J. Sci. Comput. 28 (2006), pp. 2313–2336.

[11] I. Faragó, R. Horváth, *Continuous and Discrete Parabolic Operators and their Qualitative Properties*, IMA J. Numer. Anal. (2008) (in press).

[12] 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.

[13] I. Faragó, R. Horváth, S. Korotov, *Discrete Maximum Principles for FE Solutions of Nonstationary Diffusion-Reaction Problems with Mixed Boundary Conditions*, Technical Report, Helsinki University of Technology, 2008 (submitted).

[14] I. Faragó, J. Karátson, *Numerical Solution of Nonlinear Elliptic Problems via Preconditioning Operators. Theory and Applications.* Advances in Computation, Volume 11, NOVA Science Publishers, New York, 2002.

[15] J. Francu, *Monotone Operators. A Survey Directed to Applications to Differential Equations*, Apl. Mat. 35 (1990), pp. 257–301.

[16] A. Friedmann, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, 1964.

[17] 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.

[18] A. Hannukainen, S. Korotov, T. Vejchodský, *Discrete Maximum Principles for FE Solutions of the Diffusion-Reaction Problem on Prismatic Meshes*, J. Comput. Appl. Math. (in press), 1–16. Preprint

105(2008), Institute of Mathematics of the Academy of Sciences of the Czech Republic.

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

[20] 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.

[21] J. Karátson, S. Korotov, *A Discrete Maximum Principle in Hilbert Space with Applications to Nonlinear Cooperative Elliptic Systems*, Technical Report, Helsinki University of Technology, 2008 (submitted).

[22] 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.

[23] H. B. Keller, *Elliptic Boundary Value Problems Suggested by Nonlinear Diffusion Processes*, Arch. Rational Mech. Anal. 35 (1969), pp. 363–381.

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

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

[26] 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.

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

(continued from the back cover)

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

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

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.

A550 István Faragó, Róbert Horváth, Sergey Korotov
Discrete maximum principles for FE solutions of nonstationary diffusion-reaction problems with mixed boundary conditions
August 2008

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

ISBN 978-951-22-9512-8 (print)
ISBN 978-951-22-0513-5 (PDF)
ISSN 0784-3143 (print)
ISSN 1797-5867 (PDF)