Discrete maximum-minimum principle for a linearly implicit scheme for nonlinear parabolic FEM problems under weakened time restrictions

István Faragó*[1,2,3], Róbert Horváth†[2,3], János Karátson‡[1,2,3]

[1] *ELTE Eötvös Loránd University, Department of Applied Analysis and Computational Mathematics, Pázmány Péter sétány 1/C, 1111 Budapest, Hungary*
[2] *HUN-REN–ELTE Numerical Analysis and Large Networks Research Group, Budapest*
[3] *Department of Analysis and Operations Research, Budapest University of Technology and Economics, Műegyetem rakpart 3., 1111 Budapest, Hungary*

Abstract

In this paper, we extend our earlier results in [17] on the DMP for nonlinear parabolic systems of PDEs. We propose a linearly implicit scheme, where only linear problems have to be solved on the time layers. We obtain a DMP without the restrictive condition $\Delta t \leq O(h^2)$. We show that we only need the lower bound $\Delta t \geq O(h^2)$, further, depending on the Lipschitz condition of the given nonlinearity, the upper bound is just $\Delta t \leq C$ (for globally Lipschitz) or $\Delta t \leq O(h^\gamma)$ (for locally Lipschitz) for some constant $C > 0$ arising from the PDE, or some $\gamma < 2$, respectively. In most situations in practical models, the latter condition becomes $\Delta t \leq O(h^{2/3})$ in 2D and $\Delta t \leq O(h)$ in 3D. Various real-life examples are also presented where the results can be applied to obtain physically relevant numerical solutions.

Keywords: Nonlinear parabolic PDEs, finite element methods, linearly implicit schemes, maximum-minimum principles

1 Introduction

Parabolic partial differential equations or systems arise in abundant diffusion type models in applied mathematics, hence their numerical solution is a widespread task, see, e.g., [27, 31, 32]. The qualitative reliability of these numerical methods is naturally based on the need to reproduce the qualitative properties of the original solution. These properties

*Email: istvan.farago@ttk.elte.hu
†E-mail: rhorvath@math.bme.hu
‡Corresponding author, e-mail: kajkaat@caesar.elte.hu

for parabolic problems are essentially various forms of the (continuous) maximum principle (CMP), see e.g. [30] for its classical variants and [6, 13] for the authors' results on the nonlinear case. In the discrete case the so-called discrete maximum principles (DMPs) for parabolic problems were first derived for the linear case in [19, 25], and were later extended in several works, see e.g. [10, 12, 16, 17, 33] and the references therein. An important special case of the DMP is the discrete nonnegativity preservation, which ensures the physical relevance of the obtained numerical solution such as obeying the laws of thermodynamics.

In our earlier paper [17] in this journal, we have presented the DMP for nonlinear parabolic systems of PDEs, allowing a general framework including nonsymmetric terms and mixed boundary conditions, discretized in space by finite elements (FEM). This paper has been then used and extended in several directions, see, e.g., [3, 8, 18, 34]. However, in this and the other mentioned works, the usual requirement for the space and time discretization steps is generally $\Delta t = O(h^2)$. The essential restrictive part of this requirement is the upper bound $\Delta t \leq O(h^2)$, since this leads to very small time-steps on a fine space mesh. This condition is usually considered as inevitable, both to achieve convergence and to satisfy the DMP, unless one uses an implicit scheme. Then, however, in turn one has to solve nonlinear problems on each time layer.

In this paper we propose a linearly implicit scheme, where only linear problems have to be solved on the time layers. Still, extending the results of [17], we obtain a DMP without the restrictive condition $\Delta t \leq O(h^2)$. We only need the lower bound

$$\Delta t \geq O(h^2),$$

further, depending on the Lipschitz condition of the given nonlinearity, the upper bound is just

$$\Delta t \leq C \quad \text{(for globally Lipschitz)} \qquad \text{or} \qquad \Delta t \leq O(h^\gamma) \quad \text{(for locally Lipschitz)}$$

for some constant $C > 0$ arising from the PDE, or some $\gamma < 2$, respectively. In most situations in practical models the latter condition becomes $\Delta t \leq O(h^{2/3})$ in 2D and $\Delta t \leq O(h)$ in 3D. The space FEM mesh is constructed under similar geometric properties as in [17].

The paper is organized as follows. The basic nonlinear PDE model is described in Section 2, and the linearly implicit scheme is presented in Section 3. Then the main results on the DMP for this problem are derived in Section 4. Further, Section 5 provides generalization of these results to more general PDEs and systems, and finally, Section 6 presents various real-life examples where the results can be applied to obtain physically relevant numerical solutions.

2 The nonlinear problem

2.1 Formulation of the problem

Let us consider the following nonlinear parabolic problem. Find a function $u = u(x,t)$ such that

$$\frac{\partial u}{\partial t} - \operatorname{div}\big(a(x)\nabla u\big) + \mathbf{w}(x)\cdot\nabla u + q(x,t,u) = f(x,t) \quad \text{in} \;\; Q_T := \Omega\times(0,T), \qquad (2.1)$$

where Ω is a bounded domain in $\mathbb{R}^d$ $(d \geq 2)$ and $T > 0$ is fixed. The boundary and initial conditions are given as

$$u(x,t) = g(x,t) \quad \text{for} \quad (x,t)\in\Gamma_D\times[0,T], \qquad (2.2)$$

$$\tfrac{\partial u}{\partial \nu} = 0 \quad \text{for} \quad (x,t)\in\Gamma_N\times[0,T], \qquad (2.3)$$

$$u(x,0) = u_0(x) \quad \text{for} \quad x\in\Omega, \qquad (2.4)$$

respectively, where ν is the outward unit normal vector of the boundary surface. Notice that while we allow general nonhomogeneous Dirichlet boundary condition, the Neumann condition is assumed to be homogeneous, see Remark 2.1 below.

Assumptions 2.1.

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

(A2) The coefficients satisfy $a \in L^\infty(\Omega)$, $\mathbf{w} \in W^{1,\infty}(\Omega)^d$, further, the function $q : Q_T \times \mathbb{R} \to \mathbb{R}$ is bounded w.r.t x, t and is continuously differentiable w.r.t. its last variable. In addition, $f \in L^\infty(Q_T)$, $g \in L^\infty(\Gamma_D \times [0,T])$ and $u_0 \in L^\infty(\Omega)$.

(A3) There exist positive constants μ_0 and μ_1 such that

$$0 < \mu_0 \leq a(x) \leq \mu_1 \qquad (2.5)$$

for all $x \in \Omega$.

(A4) We have $\operatorname{div}\mathbf{w} \leq 0$ on Q_T, $\quad \mathbf{w}\cdot\nu \geq 0$ on $\Gamma_N \times [0,T]$.

(A5) There exists a constant $\alpha > 0$ such that for any $(x,t) \in Q_T$ and $\xi \in \mathbb{R}$,

$$0 \leq \frac{\partial q(x,t,\xi)}{\partial \xi} \leq \alpha. \qquad (2.6)$$

(A6) For any $(x,t) \in Q_T$

$$q(x,t,0) = 0. \qquad (2.7)$$

We define weak solutions in the usual way. Here and in the sequel, equality of functions in Lebesgue or Sobolev spaces is understood almost everywhere on domains and in trace sense on boundaries. Letting $H^1_D(\Omega) := \{u \in H^1(\Omega) : \; u|_{\Gamma_D} = 0\}$, we expect that $u(.,t) \in H^1(\Omega)$ for all $t \in (0,T)$ and satisfies the relation

$$\int_\Omega \frac{\partial u}{\partial t} v \, dx + \int_\Omega \Big(a(x)\nabla u \cdot \nabla v + (\mathbf{w}(x)\cdot \nabla u)v + q(x,t,u)v \Big)\, dx = \int_\Omega f v \, dx \tag{2.8}$$

$(\forall v \in H^1_D(\Omega), \quad t \in (0,T))$, further,

$$u = g \quad \text{on} \quad [0,T]\times \Gamma_D, \qquad u|_{t=0} = u_0 \quad \text{in} \quad \Omega. \tag{2.9}$$

Remark 2.1 (Discussion of the conditions.)

(i) Most conditions are standard to make the formulae meaningful and to ensure the monotonicity of the elliptic terms, this concerns (A1)-(A4) and the lower bound in (A5).

(ii) Condition (A6) can be achieved for any equation (2.1) by subtracting $q(x,t,0)$ from both sides and redefining q and f. On the other hand, in many real models (A6) already holds since zero values of u imply zero force.

(iii) The upper bound in (A5) yields Lipschitz continuity of q w.r.t. ξ. We will briefly discuss the analogue of the above problem for locally Lipschitz nonlinearities, and also for some systems of PDEs in Section 5.

(iv) The Neumann condition in (2.3) is considered to be homogeneous, which is natural since it is induced in this form by the weak formulation (2.8). We will also refer to nonhomogeneous Neumann condition in Section 5, but the corresponding DMP is not as directly analogous to the continuous case as for the homogeneous one.

2.2 Maximum-minimum principles in the continuous case

Now let us recall the studied qualitative properties on the level of the above nonlinear PDE. The classical formulations assume that $u \in C(\overline{\Omega})$, thus it has a maximum and minimum. (The latter are replaced by ess sup and ess inf in the general case.) Various maximum-minimum principles (MPs) have been studied in [13] as proper extensions of the linear case [12], and in particular, Corollary 13 therein implies the following MP for problem (2.1)–(2.4): for all $(x,t) \in \bar{Q}_T$

$$\min\{0, \min u|_{\Gamma_{par}}\} + t\cdot \min\{0, \inf_{Q_{\bar{T}}} f\} \le u(x,t) \le \max\{0, \max u|_{\Gamma_{par}}\} + t \cdot \max\{0, \sup_{Q_{\bar{T}}} f\}\,, \tag{2.10}$$

where the parabolic boundary for the studied mixed problem is defined as

$$\Gamma_{par} := (\Gamma_D \times [0,T]) \cup (\Omega \times \{0\}) \tag{2.11}$$

and

$$Q_{\bar{T}} = Q_T \cup (\Omega \times T). \tag{2.12}$$

The MP (2.10) can be rewritten in a simpler form using the negative and positive parts of a function as

$$\min(u|_{\Gamma_{par}})^{-} + t\cdot \inf_{Q_{\bar{T}}} f^{-} \le u(x,t) \le \max(u|_{\Gamma_{par}})^{+} + t\cdot \sup_{Q_{\bar{T}}} f^{+}\ , \tag{2.13}$$

where the negative (resp. positive) part of a function is defined as follows, e.g. for the function f such that f^- and f^+ are also in $L^\infty(Q_T)$:

$$f^{-}(x,t) := \min\{0, f(x,t)\} \quad (\text{resp.} f^{+}(x,t) := \max\{0, f(x,t)\}).$$

The general form covers various special cases where one can readily derive more specific bounds from the above. For instance, as simple consequences of (2.13), we may mention the more classical maximum principle:

$$\text{if } f \ \equiv\ 0 \quad \text{and } u|_{\Gamma_{par}} \ge 0, \qquad \text{then} \quad \max_{\bar{Q}_T} u = \max u|_{\Gamma_{par}};$$

and the nonnegativity preservation:

$$\text{if } f \ge 0, \quad g \ge 0 \ \text{ and } \ u_0 \ge 0, \qquad \text{then} \quad u \ge 0.$$

In what follows, our goal is to derive the analogue of the general MP (2.10) for the properly discretized problems. The main result will be developed in Section 4, where Theorem 4.2 reproduces the discrete form of (2.10) for problem (2.1)–(2.4). Further, similar results will be proved for some more general PDEs in Section 5.

3 Construction of the discretization scheme

The discretization of problem (2.1)–(2.4) is built up in a linearly implicit way. The presentation below is the modification of that in [16, 17] to this case.

3.1 Semidiscretization in space

Let $\mathcal{T}_h$ be a finite element mesh over the solution domain $\Omega \subset \mathbb{R}^d$, where

$$h = \max_{\mathcal{K}\in\mathcal{T}_h}\{\mathrm{diam}(\mathcal{K})\}$$

stands for the discretization parameter. We define the nodal points $P_1, ..., P_{\bar{m}}$ on the mesh such that $P_1, ..., P_{m_0}$ lie in Ω or on Γ_N, and $P_{m_0+1}, ..., P_{\bar{m}}$ lie on $\overline{\Gamma}_D$, respectively, where m and m_0 are given in $\mathbb{N}^+$. We choose continuous basis functions $\phi_1, ..., \phi_{\bar{m}}$, and assume that the basis functions satisfy

$$\phi_i \ge 0 \quad (i = 1, \dots, \bar{m}), \qquad \sum_{i=1}^{\bar{m}} \phi_i \equiv 1, \tag{3.1}$$

$$\phi_i(P_j) = \delta_{ij}, \tag{3.2}$$

where δ_{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:

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

The basis functions $\phi_1, ..., \phi_{m_0}$ satisfy the homogeneous Dirichlet boundary condition on Γ_D, i.e., $\phi_i \in H_D^1(\Omega)$. We define

$$V_{h,0} = \text{span}\{\phi_1, ..., \phi_{m_0}\} \ \subset H_D^1(\Omega).$$

Then the semidiscrete problem for (2.8) with initial-boundary conditions (2.9) reads as follows: find a function $u_h = u_h(x,t)$ such that

$$u_h(x,0) = u_0^h(x) \quad (\forall x \in \Omega), \qquad u_h(.,t) - g_h(.,t) \in V_{h,0} \quad (\forall t \in (0,T)),$$

and

$$\int_\Omega \frac{\partial u_h}{\partial t} v_h \, dx + \int_\Omega \Big(a(x)\nabla u_h \cdot \nabla v_h + (\mathbf{w}(x) \cdot \nabla u_h) v_h + q(x,t,u_h) v_h \Big) \, dx = \int_\Omega f v_h \, dx \quad (3.3)$$

$(\forall v_h \in V_{h,0}, \quad t \in (0,T))$. 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 the given basis. In particular, we will use the following form to describe g_h:

$$g_h(x,t) = \sum_{i=1}^{m_\partial} g_i^h(t) \phi_{m_0+i}(x), \quad (3.4)$$

where $m_\partial := \bar{m} - m_0$. We note that, based on the consistency of the initial and boundary conditions ($g(s,0) = u_0(s), s \in \Gamma_D$), we obtain

$$g(P_{m_0+i}, 0) = u_0(P_{m_0+i}), \quad i = 1, \ldots, m_\partial.$$

We seek the semidiscrete solution in the form

$$u_h(x,t) = \sum_{i=1}^{m_0} u_i^h(t) \phi_i(x) + g_h(x,t) \quad (3.5)$$

and notice that it is sufficient that u_h satisfies (3.3) for $v_h = \phi_i$, $i = 1, 2, \ldots, m_0$ only. Then, introducing the notation

$$\mathbf{u}^h(t) = [u_1^h(t), \ldots, u_{m_0}^h(t), g_1^h(t), \ldots, g_{m_\partial}^h(t)]^T, \quad (3.6)$$

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

$$\mathbf{M}\frac{\mathrm{d}\mathbf{u}^h}{\mathrm{d}t} + \mathbf{G}(\mathbf{u}^h(t)) = \mathbf{f}(t), \quad (3.7)$$

$$\mathbf{u}^h(0) = [u_0(P_1), \ldots, u_0(P_{m_0}), g_1^h(0), \ldots, g_{m_\partial}^h(0)]^T, \quad (3.8)$$

where

$$\mathbf{M} = [M_{ij}]_{m_0 \times \bar{m}}, \quad M_{ij} = \int_\Omega \phi_j(x)\phi_i(x)\,dx, \tag{3.9}$$

$$\mathbf{G}(\mathbf{u}^h(t)) = [G(\mathbf{u}^h(t))_i]_{i=1,\dots,m_0},$$

$$G(\mathbf{u}^h(t))_i = \int\limits_\Omega \Big(a(x)\nabla u_h \cdot \nabla\phi_i + (\mathbf{w}(x)\cdot\nabla u_h)\phi_i + q(x,t,u_h)\phi_i\Big)\,dx,$$

$$\mathbf{f}(t) = [f_i(t)]_{i=1,\dots,m_0}, \quad f_i(t) = \int_\Omega f(x,t)\phi_i(x)\,dx. \tag{3.10}$$

The solution $\mathbf{u}^h = \mathbf{u}^h(t)$ of problem (3.7)–(3.8) is also called the semidiscrete solution because it delivers the coefficient functions in (3.5). Notice that only the first m_0 components of $\mathbf{u}^h(t)$ are unknown. The other components are known from the boundary condition. The existence and uniqueness of $\mathbf{u}^h(t)$ is ensured by Assumptions 2.1, since then $\mathbf{G}$ is locally Lipschitz continuous.

3.2 Full discretization with a linearly implicit scheme

In order to get a fully discrete numerical scheme, we choose a time-step Δt and apply a linearly implicit scheme for discretizing (3.7) in time. Splitting the linear and nonlinear terms, we define the decomposition

$$\mathbf{G}(\mathbf{u}^h(t)) = \mathbf{K}\,\mathbf{u}^h(t) + \mathbf{Q}(\mathbf{u}^h(t)), \tag{3.11}$$

where, for $i = 1,\dots,m_0,\quad j = 1,\dots,\overline{m}$,

$$K_{ij} = \int\limits_\Omega \Big(a(x)\nabla\phi_j\cdot\nabla\phi_i + (\mathbf{w}(x)\cdot\nabla\phi_j)\phi_i\Big)\,dx, \qquad \mathbf{Q}(\mathbf{u}^h(t))_i = \int\limits_\Omega q(x,t,u_h)\phi_i\,dx\,. \tag{3.12}$$

We consider $n = 0,1,2,\dots,n_T$ timesteps, where $n_T\Delta t = T$, and denote the corresponding r.h.s as

$$\mathbf{f}^n := \mathbf{f}(n\Delta t) \qquad (n = 0,1,2,\dots,n_T).$$

We look for the approximations

$$\mathbf{u}^n \approx \mathbf{u}^h(n\Delta t).$$

We then define the scheme as the following system of nonlinear algebraic equations:

$$\mathbf{M}\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t} + \mathbf{K}\,\mathbf{u}^{n+1} + \mathbf{Q}(\mathbf{u}^n) = \mathbf{f}^n \tag{3.13}$$

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

$$(\mathbf{M} + \Delta t\mathbf{K})\,\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n - \Delta t\mathbf{Q}(\mathbf{u}^n) + \Delta t\,\mathbf{f}^n \tag{3.14}$$

with $\mathbf{u}^0 = \mathbf{u}^h(0)$. As can be seen from (3.9) and (3.12), here $\mathbf{K}$, $\mathbf{M}$ and $\mathbf{Q}(\mathbf{u}^n)$ are matrices from $\mathbb{R}^{m_0\times\overline{m}}$, further $\mathbf{u}^n \in \mathbb{R}^{\overline{m}}$. The elements of $\mathbf{Q}(\mathbf{u}^n)$ are meant as

$$\mathbf{Q}(\mathbf{u}^n)_i = \int\limits_\Omega q(x,t,u_h^n)\phi_i\,dx \quad (i = 1,\dots,m_0)\,, \tag{3.15}$$

where u_h^n is defined as follows. The semidiscrete solution u_h can be approximated with the fully discretized numerical solution in the form

$$u_h(x,t) \approx \tilde{u}_h(x,t) := \sum_{i=1}^{\bar{m}} (\mathbf{u}^n)_i \ \phi_i(x), \tag{3.16}$$

where $t = n\Delta t$, $x \in \Omega$ and $n = 0, 1, \ldots, n_T$. Then u_h^n is a shorthand for the function $\tilde{u}_h(\cdot, n\Delta t)$.

We note that finding $\mathbf{u}^{n+1}$ in (3.14) requires the solution of a linear algebraic system. However, in the typical cases, the matrix of this system has a simple structure, therefore its solution is relatively simple. The unique solvability of (3.14) will follow from our subsequent study of the scheme, see later (4.25) and property P2 of Theorem 4.1.

3.3 On the convergence of the fully discretized model

Here we justify the choice and convergence of the scheme (3.14), based on the paper [1]. Therein the following semidiscretized equation is considered:

$$\frac{\mathrm{d}\mathscr{U}^h(t)}{\mathrm{d}t} + \mathcal{L}_h \mathscr{U}^h(t) = \mathcal{N}_h(\mathscr{U}^h(t), t). \tag{3.17}$$

We use different notations here and unlike in the previous paragraphs we indicate the dependence of the mappings on h. Here $\mathcal{L}_h$ is a positive definite, self-adjoint, densely defined linear operator and $\mathcal{N}_h(\cdot, t)$ is a (possibly) nonlinear operator on a Hilbert space H. For stability purposes, the validity of the estimate

$$\|\mathcal{N}_h(\mathbf{v}, t) - \mathcal{N}_h(\mathbf{z}, t)\|_\star \le \lambda \|\mathbf{v} - \mathbf{z}\|_{\mathcal{L}_h} + \mu \|\mathbf{v} - \mathbf{z}\|_0 \tag{3.18}$$

is assumed uniformly in t in a tube around the solution $\mathscr{U}^h$, where $\|.\|_0$ is the norm in H, $\|\mathbf{v}\|_{\mathcal{L}_h} = \|\mathcal{L}_h^{1/2}\mathbf{v}\|_0$ is the energy norm, $\|.\|_\star$ is the dual norm w.r.t $\| \ .\|_{\mathcal{L}_h}$, further, $\lambda < 1$ is the stability constant and $\mu > 0$ is an appropriate constant.

In order to solve equation (3.17) numerically, a family of linear q-step methods is constructed in the form

$$\sum_{i=0}^{q} \rho_i(\Delta t \mathcal{L}_h)\mathbf{u}^{n+i} = \Delta t \sum_{i=0}^{q-1} \sigma_i(\Delta t \mathcal{L}_h)\mathcal{N}_h(\mathbf{u}^{n+i}, t^{n+i}), \tag{3.19}$$

where the functions $\rho_i, \sigma_i : [0, \infty] \to \mathbb{R}$ $(i = 0, \ldots, q)$ are rational functions that satisfy the properties $\rho_q = 1$, $\sigma_q = 0$ and $\sigma_i(\infty) = 0$. For $x \in [0, \infty]$ the polynomials $\rho(x, \cdot)$ and $\sigma(x, \cdot)$ are introduced by

$$\rho(x, \zeta) = \sum_{i=0}^{q} \rho_i(x)\zeta^i; \quad \sigma(x, \zeta) = \sum_{i=0}^{q-1} \sigma_i(x)\zeta^i. \tag{3.20}$$

The roots of $\zeta \mapsto \rho(x, \zeta)$ are denoted by $\zeta_j(x)$ (where $j = 1, 2, \ldots, q$). These roots are ordered such that they are continuous on $x \in [0, \infty]$, and $|\zeta_j(0)| = 1$ for $j = 1, \ldots, s$

($s \leq q$). These unimodular roots are called principal roots and are denoted by $\xi_j := \zeta_j(0)$. The growth factor of ξ_j is defined as

$$\lambda_j = \frac{\partial_1 \rho(0, \xi_j)}{\xi_j \partial_2 \rho(0, \xi_j)}.$$

For the numerical method (3.19), the following requirements are posed to guarantee the stability of the method.

1. $|\zeta_j(x)| \leq 1$ for all $j = 1, \dots, q$ and $0 < x \leq \infty$ (strong $A(0)$-stability).

2. The principal roots are simple and their growth factors have positive real parts.

3. If S_1 denotes the curve of the unit circle on the complex plane, then

$$\lambda < 1/(\sup_{x>0} \max_{\zeta \in S_1} |x \; \sigma(x, \zeta)/\rho(x, \zeta)|). \tag{3.21}$$

Then two types of order are defined for the method. The first one is the order of the consistency error

$$E_n(\Delta t) = \frac{1}{\Delta t} \sum_{i=0}^{q} \rho_i(\Delta t \mathcal{L}_h) \mathscr{U}^h(t_{n+i}) - \sum_{i=0}^{q-1} \sigma_i(\Delta t \mathcal{L}_h) \mathcal{N}_h(\mathscr{U}^h(t_{n+i}), t^{n+i}), \tag{3.22}$$

which is denoted by p when $E_n(\Delta t) = \mathcal{O}(\Delta t^p)$. This is proved to hold if $\varphi_l(x) = \mathcal{O}(x^{p+1-l})$ as $x \to 0+$ ($l = 1, \dots, p$), where

$$\begin{aligned} \varphi_l(x) &= \sum_{i=0}^{q} \left[i^l \rho_i(x) - (l i^{l-1} + x i^l) \sigma_i(x) \right], \quad l = 1, \dots, p-1, \\ \varphi_p(x) &= \sum_{i=0}^{q} \left[i^p \rho_i(x) - p i^{p-1} \sigma_i(x) \right]. \end{aligned} \tag{3.23}$$

The polynomial order of consistency can be defined as follows. Let us apply the method to the scalar equation $\left(\mathscr{U}^h\right)'(t) + a\mathscr{U}^h(t) = \mathscr{F}(t)$, where $a > 0$ is constant, and $\mathscr{F} \in C^p$. We say that the polynomial order of the scheme is $\tilde{p} \leq p$ if it integrates the above equation exactly, whenever the exact solution $\mathscr{U}(t)$ is a polynomial of degree at most $\tilde{p} - 1$.

Now we return to our problem and guarantee the convergence of the scheme with Theorem 5.1 from [1]. Let us introduce the partitions

$$\mathbf{M} = [\mathbf{M}_0 | \mathbf{M}_\partial], \quad \mathbf{K} = [\mathbf{K}_0 | \mathbf{K}_\partial], \quad \begin{bmatrix} \mathbf{u}_0^h(t) \\ \mathbf{g}^h(t) \end{bmatrix}, \tag{3.24}$$

where $\mathbf{M}_0$ and $\mathbf{K}_0$ are in $\mathbb{R}^{m_0 \times m_0}$, $\mathbf{M}_\partial$ and $\mathbf{K}_\partial$ are in $\mathbb{R}^{m_0 \times m_\partial}$, $\mathbf{u}_0^h(t) \in \mathbb{R}^{m_0}$, and $\mathbf{g}^h(t) \in \mathbb{R}^{m_\partial}$, following (3.6). Because $\mathbf{M}_0$ is invertible, system (3.7) with (3.11) can be rewritten as

$$\frac{d\mathbf{u}_0^h}{dt} + \mathbf{M}_0^{-1} \mathbf{K}_0 \mathbf{u}_0^h(t) = \mathbf{M}_0^{-1} \mathbf{f}(t) - \mathbf{M}_0^{-1} \mathbf{Q}(\mathbf{u}^h(t)) - \mathbf{M}_0^{-1} \mathbf{K}_\partial \mathbf{g}^h(t) - \mathbf{M}_0^{-1} \mathbf{M}_\partial \frac{d\mathbf{g}^h}{dt}. \tag{3.25}$$

As one can easily see, (3.25) can be written in the form (3.17) with the choice

$$\mathcal{L}_h = \mathbf{M}_0^{-1}\mathbf{K}_0, \qquad \mathcal{N}_h(\mathbf{u}_0^h(t), t) = \mathbf{M}_0^{-1}\mathbf{f}(t) - \mathbf{M}_0^{-1}\mathbf{Q}(\mathbf{u}^h(t)) - \mathbf{M}_0^{-1}\mathbf{K}_\partial \mathbf{g}^h(t) - \mathbf{M}_0^{-1}\mathbf{M}_\partial \frac{d\mathbf{g}^h}{dt}. \tag{3.26}$$

With the partition

$$\mathbf{u}^n = \begin{bmatrix} \mathbf{u}_0^n \\ \mathbf{g}^n \end{bmatrix}, \quad \mathbf{u}_0^n = [u_1^n, \dots, u_{m_0}^n]^T \in \mathbb{R}^{m_0} \qquad \text{and} \qquad \mathbf{g}^n = [g_1^n, \dots, g_{m_\partial}^n]^T \in \mathbb{R}^{m_\partial}, \tag{3.27}$$

the scheme (3.14) can be written as

$$\begin{aligned} &\mathbf{u}_0^{n+1} - (\mathbf{I} + \Delta t \mathbf{M}_0^{-1}\mathbf{K}_0)^{-1}\mathbf{u}_0^n \\ =&(\mathbf{I} + \Delta t \mathbf{M}_0^{-1}\mathbf{K}_0)^{-1}\mathbf{M}_0^{-1}\left(\mathbf{M}_\partial \mathbf{g}^n - \Delta t \mathbf{Q}(\mathbf{u}^n) + \Delta t \mathbf{f}^n - (\mathbf{M}_\partial + \Delta t \mathbf{K}_\partial)\mathbf{g}^{n+1}\right), \end{aligned} \tag{3.28}$$

which corresponds to scheme (3.19) with the choice

$$q = 1; \quad \rho_0(x) = -\frac{1}{1+x}, \quad \rho_1(x) = 1, \quad \sigma_0(x) = \frac{1}{1+x}. \tag{3.29}$$

That is, we have

$$\rho(x,\zeta) = -\frac{1}{1+x} + \zeta \quad \text{and} \quad \sigma(x,\zeta) = \frac{1}{1+x}.$$

Let us notice that scheme (3.28) contains the approximation of $d\mathbf{g}^h/dt$ in the form $(\mathbf{g}^{n+1} - \mathbf{g}^n)/\Delta t$ that introduces a local approximation error of $\mathcal{O}(\Delta t)$. This does not modify the final order of the scheme.

We check the validity of the requirements for our problem and scheme.

First, it is easy to see that problem (3.25) satisfies the estimate (3.18) with the mapping $\mathcal{N}_h$ defined in (3.26).

We work in the space $\mathbb{R}^{m_0}$. If $\mathbf{z} = [z_1, \dots, z_{m_0}]^T$, then we define the following norms, which are related to the norms of the function $z := \sum_{i=1}^{m_0} z_i \phi_i$:

$$\|\mathbf{z}\|_0 := \left(\mathbf{M}_0\mathbf{z}\cdot\mathbf{z}\right)^{1/2} = \|z\|_{L^2(\Omega)}, \qquad \|\mathbf{z}\|_1 := \left(\mathbf{K}_0\mathbf{z}\cdot\mathbf{z}\right)^{1/2} = \|z\|_{H^1_D(\Omega),\mathcal{L}}$$

where $\|z\|^2_{H^1(\Omega),\mathcal{L}} = \int_\Omega a|\nabla z|^2$. Note that

$$\|\mathbf{z}\|^2_{\mathcal{L}_h} := \|\mathcal{L}_h^{1/2}\mathbf{z}\|_0 = \langle \mathcal{L}_h\mathbf{z}, \mathbf{z}\rangle_0 = \mathbf{M}_0\mathcal{L}_h\mathbf{z}\cdot\mathbf{z} = \mathbf{K}_0\mathbf{z}\cdot\mathbf{z} = \|\mathbf{z}\|_1^2\,.$$

The dual norm of $\mathcal{L}_h$ is

$$\|\mathbf{y}\|_* := \sup_{\mathbf{z}\neq 0} \frac{|\langle \mathbf{y}, \mathbf{z}\rangle_0|}{\|\mathbf{z}\|_{\mathcal{L}_h}} = \sup_{\mathbf{z}\neq 0} \frac{|\mathbf{M}_0\mathbf{y}\cdot\mathbf{z}|}{\|\mathbf{z}\|_1}\,. \tag{3.30}$$

Now let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^{m_0}$ and $\bar{\mathbf{v}}(t) := [\mathbf{v},\, \mathbf{g}^h(t)]^T$, $\bar{\mathbf{w}}(t) := [\mathbf{w},\, \mathbf{g}^h(t)]^T$. Then

$$\mathcal{N}_h(\mathbf{v}, t) - \mathcal{N}_h(\mathbf{w}, t) = -\mathbf{M}_0^{-1}\left(Q(\bar{\mathbf{v}}(t)) - Q(\bar{\mathbf{w}}(t))\right),$$

hence, letting v and w be the semidiscrete approximations in the form (3.5), and using (3.30) and the weighted Poincaré–Friedrichs inequality $\|z\|_{L^2(\Omega)} \leq C_{\Omega,\mathcal{L}}\|z\|_{H^1_D(\Omega),\mathcal{L}}$, respectively, we have

$$\|\mathcal{N}_h(\mathbf{v},t) - \mathcal{N}_h(\mathbf{w},t)\|_\star = \sup_{\mathbf{z}\neq 0} \frac{\left|\left(Q(\bar{\mathbf{v}}(t)) - Q(\bar{\mathbf{w}}(t))\right)\cdot \mathbf{z}\right|}{\|\mathbf{z}\|_1}$$

$$= \sup_{z\in H^1_D(\Omega)} \frac{\left|\int_\Omega (q(x,t,v) - q(x,t,w))\, z\right|}{\|z\|_{H^1_D(\Omega),\mathcal{L}}} \leq \alpha \sup_{z\in H^1_D(\Omega)} \frac{\|v-w\|_{L^2(\Omega)}\|z\|_{L^2(\Omega)}}{\|z\|_{H^1_D(\Omega),\mathcal{L}}}$$

$$\leq \alpha C_{\Omega,\mathcal{L}}\|v-w\|_{L^2(\Omega)} = \alpha C_{\Omega,\mathcal{L}}\|\mathbf{v}-\mathbf{w}\|_0 = \mu\|\mathbf{v}-\mathbf{w}\|_0,$$

that is, (3.18) holds with

$$\lambda = 0 \quad \text{and} \quad \mu = \alpha C_\Omega > 0. \tag{3.31}$$

Second, we check that our scheme satisfies the above requirements 1-3.

1. Since $q = 1$, we have only one root $\zeta_1(x) = \dfrac{1}{1+x}$, which is a continuous function in x. Moreover $|\zeta_1(x)| \leq 1$ if $x \in (0,\infty]$.

2. The only principal root is $\xi_1 = \zeta_1(0) = 1$. Since $\partial_1\rho(0,\xi_1) = 1$ and $\partial_2\rho(0,\xi_1) = 1$ therefore $\lambda_1 = 1$, which has positive real part.

3. We have seen in (3.31) that $\lambda = 0$, hence the estimate (3.21) is trivially satisfied.

Now, based on [1], we can already give an esimation for the fully discretized approximation. Clearly, the local approximation error of our (ρ,σ)-method is $p = 1$. The polynomial order of consistency can be determined as follows. An easy computation shows that the polynomial order of the method is also $\tilde{p} = 1$, since the method is exact on the scalar equation if its solution is constant. In order to estimate the fully dicretized solution, we use Theorem 5.1 in [1]. Let us denote the space discretization stepsize and the order of the semidiscretization by h and r, respectively, Then, according to the cited Theorem 5.1, there exists a constant $C > 0$, independent of Δt and h, such that the estimate

$$\max_n \|\mathbf{u}^h(t_n) - \mathbf{u}^n\|_{L^2(\Omega)} \leq C(\Delta t + h^r)$$

holds, provided that Δt and $h^{2r}\Delta t^{-1}$ are sufficiently small. We note that $2r \geq 2$, hence the latter restrictions can be ensured if there exist suitable constants $C_1, C_2 > 0$ such that

$$C_1\, h^2 \leq \Delta t \leq C_2\,. \tag{3.32}$$

4 The discrete maximum-minimum principle for the nonlinear problem

The goal of this section is to give conditions that guarantee the maximum-minimum principle for the fully discretized nonlinear problem. To achieve this, first we reformulate our problem, then we investigate the linear case, and based on the obtained results, we give the sufficient conditions in the nonlinear case. As we will see, the maximum-minimum principle has a linear algebraic formulation because we consider the totally discretized scheme.

4.1 Reformulation of the problem

We can rewrite problem (2.8) as follows. Let

$$r(x,t,\xi) := \int_0^1 \frac{\partial q}{\partial \xi}(x,t,s\xi)\, ds \qquad ((x,t) \in Q_T,\ \xi \in \mathbb{R}), \tag{4.1}$$

then the Newton-Leibniz formula yields (see assumption (A6) in Assumptions 2.1) for all x, t, ξ that

$$q(x,t,\xi) = r(x,t,\xi)\, \xi\, .$$

We thus obtain that problem (2.8) is equivalent to

$$\int\limits_\Omega \frac{\partial u}{\partial t} v\, dx + B(u;u,v)(t) = \int\limits_\Omega f v\, dx \qquad (\forall v \in H^1_D(\Omega), \quad t \in (0,T)),$$

where

$$B(z;u,v)(t) := \int\limits_\Omega \Big(a(x)\nabla u \cdot \nabla v + (\mathbf{w}(x) \cdot \nabla u)v + r(x,t,z)uv \Big)\, dx \tag{4.2}$$

$(z, u, v \in H^1_D(\Omega))$. Then, introducing the notations

$$\mathbf{R}(\mathbf{u}^n) = [R(\mathbf{u}^n)_{ij}]_{m_0 \times \bar{m}}\ , \quad \text{where} \quad R(\mathbf{u}^n)_{ij} = \int_\Omega r(x, t_n, u^n_h)\phi_i \phi_j\, dx \tag{4.3}$$

the full discretization (3.14) becomes

$$\mathbf{M}\mathbf{u}^{n+1} + \Delta t \mathbf{K}\, \mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n - \Delta t \mathbf{R}(\mathbf{u}^n)\mathbf{u}^n + \Delta t\ \mathbf{f}^n\, . \tag{4.4}$$

Letting

$$\mathbf{A} := \mathbf{M} + \Delta t \mathbf{K}, \qquad \mathbf{B}(\mathbf{u}^n) := \mathbf{M} - \Delta t \mathbf{R}(\mathbf{u}^n), \tag{4.5}$$

the iteration procedure (4.4) takes the form with a coefficient matrix $\mathbf{B}$ depending on $\mathbf{u}^n$:

$$\mathbf{A}\mathbf{u}^{n+1} = \mathbf{B}(\mathbf{u}^n)\mathbf{u}^n + \Delta t\ \mathbf{f}^n\, . \tag{4.6}$$

4.2 DMP in the linear case

The scheme (4.6) can be considered as an iteration with a varying matrix $\mathbf{B}$. Its stepwise qualitative properties can be investigated by considering the linear special case of the procedure:

$$\mathbf{A}\mathbf{u}^{n+1} = \mathbf{B}\mathbf{u}^n + \Delta t\ \mathbf{f}^n, \tag{4.7}$$

where now $\mathbf{B}$ is a fixed matrix. We recall some results from [17, Section 4].

We use the following partitions of the matrices and vectors:

$$\mathbf{A} = [\mathbf{A}_0 | \mathbf{A}_\partial], \quad \mathbf{B} = [\mathbf{B}_0 | \mathbf{B}_\partial], \quad \mathbf{u}^n = \begin{bmatrix} \mathbf{u}^n_0 \\ \mathbf{g}^n \end{bmatrix}, \tag{4.8}$$

where, obviously, $\mathbf{A}_0$ and $\mathbf{B}_0$ are square matrices from $\mathbb{R}^{m_0 \times m_0}$; $\mathbf{A}_\partial, \mathbf{B}_\partial \in \mathbb{R}^{m_0 \times m_\partial}$,

$$\mathbf{u}^n_0 = [u^n_1, \dots, u^n_{m_0}]^T \in \mathbb{R}^{m_0} \qquad \text{and} \qquad \mathbf{g}^n = [g^n_1, \dots, g^n_{m_\partial}]^T \in \mathbb{R}^{m_\partial}.$$

Then the iteration (4.7) can be also written as

$$[\mathbf{A}_0|\mathbf{A}_\partial]\left[\begin{array}{c}\mathbf{u}_0^{n+1}\\ \mathbf{g}^{n+1}\end{array}\right]=[\mathbf{B}_0|\mathbf{B}_\partial]\left[\begin{array}{c}\mathbf{u}_0^{n}\\ \mathbf{g}^{n}\end{array}\right]+\Delta t\,\mathbf{f}^n. \tag{4.9}$$

We will use the following notations:

$$\mathbf{e}_0=[1,\ldots,1]^T\in\mathbb{R}^{m_0},\quad \mathbf{e}_\partial=[1,\ldots,1]^T\in\mathbb{R}^{m_\partial},\quad \mathbf{e}=[1,\ldots,1]^T\in\mathbb{R}^{\bar{m}}. \tag{4.10}$$

Definition 4.1 The scheme (4.7) is called *admissible* if the following properties are valid:

P1. $\mathbf{Ae}>\mathbf{0}$;

P2. $\mathbf{A}_0$ is a monotone matrix, that is $\mathbf{A}_0$ is invertible and $\mathbf{A}_0^{-1}\geq\mathbf{0}$;

P3. $\mathbf{A}_0^{-1}\mathbf{A}_\partial\leq\mathbf{0}$; P4. $\mathbf{B}\geq\mathbf{0}$;

P5. $\mathbf{Ae}\geq\mathbf{Be}$.

We will use the following various notations:

$$g_{min}^n:=\min\mathbf{g}^n=\min\{g_1^n,\ldots,g_{m_\partial}^n\},\qquad g_{max}^n:=\max\mathbf{g}^n=\max\{g_1^n,\ldots,g_{m_\partial}^n\}, \tag{4.11}$$

$$u_{min}^n:=\min\mathbf{u}_0^n=\min\{u_1^n,\ldots,u_{m_0}^n\},\qquad u_{max}^n:=\max\mathbf{u}_0^n=\max\{u_1^n,\ldots,u_{m_0}^n\}, \tag{4.12}$$

$$v_{min}^n:=\min\mathbf{u}^n=\min\{g_{min}^n,u_{min}^n\},\qquad v_{max}^n:=\max\mathbf{u}^n=\max\{g_{max}^n,u_{max}^n\}, \tag{4.13}$$

$$f_{min}^n=\min\left\{\frac{f_1^n}{(\mathbf{Ae})_1},\ldots,\frac{f_{m_0}^n}{(\mathbf{Ae})_{m_0}}\right\},\quad f_{max}^n=\max\left\{\frac{f_1^n}{(\mathbf{Ae})_1},\ldots,\frac{f_{m_0}^n}{(\mathbf{Ae})_{m_0}}\right\}. \tag{4.14}$$

We formulate below a stepwise discrete maximum-minimum principle (DMP) for the discrete model (4.7), following [19, p. 100] and [17, Section 4.2]:

$$\begin{array}{c}\min\{0,g_{min}^n,g_{min}^{n+1},u_{min}^n\}+\Delta t\min\{0,f_{min}^n\}\leq\\ u_i^{n+1}\leq\max\{0,g_{max}^n,g_{max}^{n+1},u_{max}^n\}+\Delta t\max\{0,f_{max}^n\}\end{array} \tag{4.15}$$

$(i=1,\ldots,m_0;\ n=0,1\ldots,n_T-1)$.

Proposition 4.1 *If the scheme* (4.7) *is admissible in the sense of Definition 4.1, then the stepwise DMP of the form* (4.15) *holds.*

Proof. The statement follows from [17, Lemma 4.1 and Theorem 4.1]. For completeness, we summarize the proof, since [17] uses these properties in a combined way with auxiliary ones. For any $n=0,\ldots,n_T$ let

$$v_{min,-}^n=\min\{0,v_{min}^n\},\quad v_{max,+}^n=\max\{0,v_{max}^n\}, \tag{4.16}$$

$$f_{min,-}^n=\min\{0,f_{min}^n\},\quad f_{max,+}^n=\max\{0,f_{max}^n\}.$$

In view of P1, the values in (4.14) are well-defined. We define the positive diagonal matrix $\mathbf{D}=\operatorname{diag}((\mathbf{Ae})_1,\ldots,(\mathbf{Ae})_{m_0})\in\mathbb{R}^{m_0\times m_0}$, which is invertible. The inverse is also a positive diagonal matrix. Then $\mathbf{Ae}=\mathbf{De}_0$.

Note that (4.9) implies

$$\mathbf{A}_0\mathbf{u}_0^{n+1} + \mathbf{A}_\partial\mathbf{g}^{n+1} = \mathbf{B}\mathbf{u}^n + \Delta t\mathbf{f}^n. \tag{4.17}$$

Here (4.14) yields $\max \mathbf{D}^{-1}\mathbf{f}^n = f^n_{max} \le f^n_{max,+}$, hence $\mathbf{D}^{-1}\mathbf{f}^n \le f^n_{max,+}\,\mathbf{e}_0$, and from the positivity of $\mathbf{D}$ we get $\mathbf{f}^n \le f^n_{max,+}\,\mathbf{D}\mathbf{e}_0 = f^n_{max,+}\,\mathbf{A}\mathbf{e}$. This and (4.17) yield

$$\mathbf{A}_0\mathbf{u}_0^{n+1} + \mathbf{A}_\partial\mathbf{g}^{n+1} \le \mathbf{B}\mathbf{u}^n + \Delta t f^n_{max,+}\mathbf{A}\mathbf{e}.$$

Hence, using P2, and then P4 and P5, respectively, we get

$$\begin{aligned}
\mathbf{u}_0^{n+1} &= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\,\mathbf{g}^{n+1} + \mathbf{A}_0^{-1}\mathbf{B}\mathbf{u}^n + \Delta t f^n_{max,+}\mathbf{A}_0^{-1}\mathbf{A}\mathbf{e} \le \\
&\le -\mathbf{A}_0^{-1}\mathbf{A}_\partial\,\mathbf{g}^{n+1} + v^n_{max,+}\mathbf{A}_0^{-1}\mathbf{B}\mathbf{e} + \Delta t f^n_{max,+}\mathbf{A}_0^{-1}\mathbf{A}\mathbf{e} \le \\
&\le -\mathbf{A}_0^{-1}\mathbf{A}_\partial\,\mathbf{g}^{n+1} + v^n_{max,+}\mathbf{A}_0^{-1}\mathbf{A}\mathbf{e} + \Delta t f^n_{max,+}\mathbf{A}_0^{-1}\mathbf{A}\mathbf{e} = \\
&= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\,\mathbf{g}^{n+1} + (v^n_{max,+} + \Delta t f^n_{max,+})\mathbf{A}_0^{-1}(\mathbf{A}_0\mathbf{e}_0 + \mathbf{A}_\partial\,\mathbf{e}_\partial) \\
&= -\mathbf{A}_0^{-1}\mathbf{A}_\partial\,\mathbf{g}^{n+1} + (v^n_{max,+} + \Delta t f^n_{max,+})(\mathbf{e}_0 + \mathbf{A}_0^{-1}\mathbf{A}_\partial\,\mathbf{e}_\partial).
\end{aligned} \tag{4.18}$$

Regrouping the above inequality, we get

$$\mathbf{u}_0^{n+1} - (v^n_{max,+} + \Delta t f^n_{max,+})\mathbf{e}_0 \le -\mathbf{A}_0^{-1}\mathbf{A}_\partial\big(\mathbf{g}^{n+1} - (v^n_{max,+} + \Delta t f^n_{max,+})\mathbf{e}_\partial\big). \tag{4.19}$$

Now we observe that, due to P1 and P2, we have that $-\mathbf{A}_0^{-1}\mathbf{A}\mathbf{e} \le \mathbf{0}$, that is, taking into the account the partition, we obtain that $-\mathbf{A}_0^{-1}(\mathbf{A}_0\mathbf{e}_0 + \mathbf{A}_\partial\mathbf{e}_\partial) \le \mathbf{0}$. This can be transformed into the relation

$$-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e}_\partial \le \mathbf{e}_0\,. \tag{4.20}$$

From this, for the i-th coordinate of the both sides of (4.19), first using P3 and $f^n_{max,+} \ge 0$, we obtain

$$\begin{aligned}
&u_i^{n+1} - v^n_{max,+} - \Delta t f^n_{max,+} \le \textstyle\sum_{j=1}^{m_\partial}\left(-\mathbf{A}_0^{-1}\mathbf{A}_\partial\right)_{ij}(g_j^{n+1} - v^n_{max,+} - \Delta t\, f^n_{max,+}) \le \\
&\le \left(\textstyle\sum_{j=1}^{m_\partial}\left(-\mathbf{A}_0^{-1}\mathbf{A}_\partial\right)_{ij}\right)\cdot \max\{0, \max_j\{g_j^{n+1} - v^n_{max,+}\}\} \le \max\{0, \max_j\{g_j^{n+1} - v^n_{max,+}\}\}.
\end{aligned} \tag{4.21}$$

Finally, expressing u_i^{n+1}, we obtain the required inequality. The inequality on the left-hand side of (4.15) can be proved similarly. ■

4.3 The DMP for the nonlinear problem

Now we can study our main goal. Recall our full discretization scheme (4.6):

$$\mathbf{A}\mathbf{u}^{n+1} = \mathbf{B}(\mathbf{u}^n)\mathbf{u}^n + \Delta t\,\mathbf{f}^n, \tag{4.22}$$

where according to (4.5) we have

$$\mathbf{A} = \mathbf{M} + \Delta t\mathbf{K}, \qquad \mathbf{B}(\mathbf{u}^n) = \mathbf{M} - \Delta t\mathbf{R}(\mathbf{u}^n)\,. \tag{4.23}$$

We will rely on subsection 4.2, hence we use the following partitions of the matrices and vectors, analogous to (4.8):

$$\mathbf{A} = [\mathbf{A}_0|\mathbf{A}_\partial], \quad \mathbf{B}(\mathbf{u}^n) = [\mathbf{B}(\mathbf{u}^n)_0|\mathbf{B}(\mathbf{u}^n)_\partial], \quad \mathbf{u}^n = \begin{bmatrix} \mathbf{u}_0^n \\ \mathbf{g}^n \end{bmatrix}, \tag{4.24}$$

where $\mathbf{A}_0$ and $\mathbf{B}(\mathbf{u}^n)_0$ are from $\mathbb{R}^{m_0 \times m_0}$; $\mathbf{A}_\partial, \mathbf{B}(\mathbf{u}^n)_\partial \in \mathbb{R}^{m_0 \times m_\partial}$, $\mathbf{u}_0^n = [u_1^n, \ldots, u_{m_0}^n]^T \in \mathbb{R}^{m_0}$ and $\mathbf{g}^n = [g_1^n, \ldots, g_{m_\partial}^n]^T \in \mathbb{R}^{m_\partial}$. Then the iteration (4.22) can be written as

$$[\mathbf{A}_0|\mathbf{A}_\partial] \begin{bmatrix} \mathbf{u}^{n+1} \\ \mathbf{g}^{n+1} \end{bmatrix} = [\mathbf{B}(\mathbf{u}^n)_0|\mathbf{B}(\mathbf{u}^n)_\partial] \begin{bmatrix} \mathbf{u}^n \\ \mathbf{g}^n \end{bmatrix} + \Delta t\, \mathbf{f}^n. \tag{4.25}$$

In what follows, we will need a proper (patch-)regularity of the considered meshes.

Definition 4.2 Let $\Omega \subset \mathbf{R}^d$. A family of FEM subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ will be called *regular* if there exist constants $c_1, c_2, c_3 > 0$ such that for any $h > 0$ and basis function ϕ_j,

$$c_1 h^d \le \operatorname{meas}(\operatorname{supp}\phi_j) \le c_3 h^d, \qquad \max|\nabla\phi_j| \le \frac{c_2}{h} \tag{4.26}$$

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

Conditions (3.1)–(3.2) and the above regularity properties are satisfied for linear, bilinear or prismatic elements if the mesh is shape regular, i.e., satisfies the minimum angle condition. This follows from the standard formulae for the basis functions [29].

The following main theorem uses the constants $c_1, c_2, c_3 > 0$ from the above definition, μ_0 from (2.5), and we let

$$\omega := \sup_\Omega |\mathbf{w}|\,.$$

Theorem 4.1 *Let problem (2.1)–(2.4) satisfy Assumptions 2.1. Let us consider a family of finite element subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ that is regular in the sense of Definition 4.2 and the basis functions satisfy (3.1)–(3.2). Let the following assumptions hold:*

(i) (acuteness/non-narrowness of the space mesh) *for any $i = 1, ..., m_0,\ j = 1, ..., \bar{m}$ $(i \ne j)$, if* $\;meas(\operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j) > 0$ *then*

$$\nabla\phi_i \cdot \nabla\phi_j \le 0 \quad \text{on } \Omega \quad \text{and} \quad \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \le -K_0\, h^{d-2} \tag{4.27}$$

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

(ii) (upper bound of the space mesh size) *the mesh parameter h satisfies*

$$h < h_0 := \frac{\mu_0 K_0}{\omega c_2 c_3}; \tag{4.28}$$

(iii) (lower bound of the time step) *the time-step Δt satisfies*

$$\frac{c_3 h^2}{\mu_0 K_0 - \omega c_2 c_3\, h} \le \Delta t\,; \tag{4.29}$$

(iv) (upper bound of the time step) *the time-step Δt satisfies, with α from (2.6),*

$$\Delta t \le \frac{1}{\alpha}\,. \tag{4.30}$$

Then the scheme (3.14) satisfies the following properties, where the notations $\mathbf{e}_0$, $\mathbf{e}_\partial$ and $\mathbf{e}$ are from (4.10)*:*

P1. $\mathbf{Ae} > \mathbf{0}$*;* P2. $\mathbf{A}_0$ *is a monotone matrix;*

P3. $\mathbf{A}_0^{-1}\mathbf{A}_\partial \le \mathbf{0}$*;* P4. $\mathbf{B}(\mathbf{u}^n) \ge \mathbf{0}$*;*

P5. $\mathbf{Ae} \ge \mathbf{B}(\mathbf{u}^n)\mathbf{e}$*.*

Proof. We have, from (3.12), (4.3) and (4.23), for $i = 1,\dots,m_0,\quad j = 1,\dots,\overline{m}$,

$$A_{ij} = M_{ij} + \Delta t K_{ij} = \int_\Omega \phi_j\phi_i + \ \Delta t \int_\Omega \Big(a\,\nabla\phi_j\cdot\nabla\phi_i + (\mathbf{w}\cdot\nabla\phi_j)\phi_i\Big) \tag{4.31}$$

and

$$B(\mathbf{u}^n)_{ij} = M_{ij} - \Delta t R(\mathbf{u}^n)_{ij} = \int_\Omega \phi_j\phi_i - \ \Delta t \int_\Omega r(x,t_n,u_h^n)\phi_i\phi_j\,. \tag{4.32}$$

(1) First we prove that

$$A_{ij} \le 0 \qquad (\forall i = 1,\dots,m_0,\ j = 1,\dots,\overline{m};\quad i \ne j). \tag{4.33}$$

Let

$$\Omega_{ij} := \operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j. \tag{4.34}$$

Here, by (3.1) and (4.26),

$$\int_\Omega \phi_j\phi_i \le meas(\Omega_{ij}) \le c_3 h^d \tag{4.35}$$

By (2.5) and (4.27),

$$\int_\Omega a\,\nabla\phi_j\cdot\nabla\phi_i \le -\mu_0 K_0\, h^{d-2} \tag{4.36}$$

and by (3.1), (4.26)

$$\int_\Omega (\mathbf{w}\cdot\nabla\phi_j)\phi_i \le \frac{\omega c_2}{h}\, meas(\Omega_{ij}) \le \omega c_2 c_3 h^{d-1}\,. \tag{4.37}$$

Altogether, we obtain

$$A_{ij} \le h^d\left(c_3 + \ \Delta t\Big(-\frac{\mu_0 K_0}{h^2} + \frac{\omega c_2 c_3}{h}\Big)\right) = h^d\left(c_3 + \ \Delta t\ \frac{-\mu_0 K_0 + \omega c_2 c_3\, h}{h^2}\right).$$

Since $h < h_0$ for h_0 defined in (4.28), it follows that we have a negative coefficient of Δt above, and from (4.29) we obtain that the expression in the large parentheses is nonpositive, hence $A_{ij} \le 0$.

(2) Now we can verify properties P1–P5.

P1., Let $i = 1, ..., m_0$. From (3.1) and (3.12),

$$(\mathbf{K}\mathbf{e})_i = \sum_{j=1}^{\overline{m}} K_{ij} = \int\limits_{\Omega} \left(\left(a\,\nabla \left(\sum_{j=1}^{\overline{m}} \phi_j \right) \right) \cdot \nabla\phi_i + \left(\mathbf{w} \cdot \left(\sum_{j=1}^{\overline{m}} \phi_j \right) \right) \phi_i \right) = 0\,. \tag{4.38}$$

Hence by (4.31)

$$(\mathbf{A}\mathbf{e})_i = \sum_{j=1}^{\overline{m}} A_{ij} = \sum_{j=1}^{\overline{m}} M_{ij} = \int_{\Omega} \left(\sum_{j=1}^{\overline{m}} \phi_j \right) \phi_i = \int_{\Omega} \phi_i > 0\,. \tag{4.39}$$

P2., As seen in (4.33), we have offdiag$(\mathbf{A}) \le \mathbf{0}$. Hence, in particular, $\mathbf{A}_\partial \le \mathbf{0}$, and based on P1, we obtain

$$\mathbf{A}_0\mathbf{e}_0 \ge \mathbf{A}_0\mathbf{e}_0 + \mathbf{A}_\partial\mathbf{e}_\partial = \mathbf{A}\mathbf{e} > \mathbf{0} \tag{4.40}$$

for the vector $\mathbf{e}_0 > \mathbf{0}$. Therefore $\mathbf{A}_0$ is an M-matrix, which implies P2.

P3., This is now obvious, since we have seen $\mathbf{A}_0^{-1} \ge \mathbf{0}$ and $\mathbf{A}_\partial \le \mathbf{0}$.

P4., By its definition (4.1) the function r inherits (2.6), i.e.,

$$0 \le r \le \alpha. \tag{4.41}$$

Hence, from (4.30) and (4.32), also using (3.1), for any $i = 1, \ldots, m_0,\ \ j = 1, \ldots, \overline{m}$ we have

$$B(\mathbf{u}^n)_{ij} = \int_{\Omega} \big(1 - \Delta t\; r(x, t_n, u_h^n)\big)\, \phi_i\phi_j \ge \big(1 - \alpha\,\Delta t\big) \int_{\Omega} \phi_i\phi_j \ge 0.$$

P5., By (4.23) we have $\mathbf{A} - \mathbf{B}(\mathbf{u}^n) = \Delta t\,\big(\mathbf{K} + \mathbf{R}(\mathbf{u}^n)\big)$. Hence, using (4.3), (4.38) and (4.41),

$$\big((\mathbf{A} - \mathbf{B}(\mathbf{u}^n))\mathbf{e}\big)_i = \Delta t\,\big((\mathbf{K} + \mathbf{R}(\mathbf{u}^n))\mathbf{e}\big)_i = \Delta t \left(\sum_{j=1}^{\overline{m}} K_{ij} + \sum_{j=1}^{\overline{m}} R(\mathbf{u}^n)_{ij} \right) \tag{4.42}$$

$$= \Delta t \int\limits_{\Omega} r(x, t_n, u_h^n) \left(\sum_{j=1}^{\overline{m}} \phi_j \right) \phi_i = \Delta t \int\limits_{\Omega} r(x, t_n, u_h^n)\, \phi_i \ge 0\,,$$

hence P5 holds. ∎

Now we are in the position to derive the corresponding *discrete maximum principles*. First, Proposition 4.1 implies that the stepwise DMP of the form (4.15) holds:

Corollary 4.1 *If the conditions of Theorem 4.1 hold, then the scheme (3.14) satisfies*

$$\begin{gathered} \min\{0, g_{min}^n, g_{min}^{n+1}, u_{min}^n\} + \Delta t\, \min\{0, f_{min}^n\} \le \\ u_i^{n+1} \le \max\{0, g_{max}^n, g_{max}^{n+1}, u_{max}^n\} + \Delta t\, \max\{0, f_{max}^n\} \end{gathered} \tag{4.43}$$

($i = 1, \ldots, m_0$; $n = 0, 1, \ldots, n_T - 1$).

The above result reduces the values of u on a given time level to those on the previous time level. By induction, we obtain a global DMP that reduces the values of u only to the input data until the considered time level. For this we introduce the notations

$$\begin{aligned} g_{min}^{(n)} &:= \min\Big\{ g_{min}^{0}, \ldots, g_{min}^{n} \Big\}, \quad f_{min}^{(n)} := \min\Big\{ f_{min}^{0}, \ldots, f_{min}^{n} \Big\}, \\ g_{max}^{(n)} &:= \max\Big\{ g_{max}^{0}, \ldots, g_{max}^{n} \Big\}, \quad f_{max}^{(n)} := \max\Big\{ f_{max}^{0}, \ldots, f_{max}^{n} \Big\}. \end{aligned} \tag{4.44}$$

Corollary 4.2 *If the conditions of Theorem 4.1 hold, then the scheme (3.14) satisfies*

$$\begin{aligned} &\min\{0, g_{min}^{(n)}, u_{min}^{0}\} + n\Delta t \min\{0, f_{min}^{(n+1)}\} \\ &\le u_i^n \le \max\{0, g_{max}^{(n)}, u_{max}^{0}\} + n\Delta t \max\{0, f_{max}^{(n-1)}\} \end{aligned} \tag{4.45}$$

($i = 1, \ldots, m_0$; $n = 0, \ldots, n_T$).

Our final goal is to reproduce the discrete form of the maximum-minimum principle (2.10) for the approximation of the semidiscrete solution (see (3.16)). For this, we introduce the discrete parabolic boundary Γ_{par}^h as an analogue of (2.11) such that we only involve the discrete time levels: letting

$$\tau := \{n\Delta t\}_{n=0,1,\ldots,n_T},$$

we define

$$\Gamma_{par}^h := (\Gamma_D \times \tau) \cup (\Omega \times \{0\}). \tag{4.46}$$

Theorem 4.2 *If the conditions of Theorem 4.1 hold, then the scheme (3.14) satisfies*

$$\min\{0, \min \tilde{u}_{h\,|\Gamma_{par}^h}\} + t \cdot \min\{0, \inf_{Q_T} f\} \le \tilde{u}_h(x,t) \le \max\{0, \max \tilde{u}_{h\,|\Gamma_{par}^h}\} + t \cdot \max\{0, \sup_{Q_T} f\} \tag{4.47}$$

for all $x \in \Omega$ and $t = n\Delta t$, where $n = 0, \ldots, n_T$.

Proof. Firstly, to replace the terms with g and u^0 in (4.45), we just have to reformulate them and take the partition-of-unity property (3.1)–(3.2) into account. Indeed, then the definitions (4.11) and (4.44) for g_h and the definitions (3.8) and (4.13) for u_0^h, respectively, imply

$$g_{min}^{(n)} \ge \min g_{h\,|\Gamma_D \times \tau}, \qquad u_{min}^0 = \min_{\Omega} u_0^h, \qquad g_{max}^{(n)} \le \max g_{h\,|\Gamma_D \times \tau}, \qquad u_{max}^0 = \max_{\Omega} u_0^h\ .$$

Then, from (4.46),

$$\min\{0, g_{min}^{(n)}, u_{min}^0\} \ge \min\{0, \min \tilde{u}_{h\,|\Gamma_{par}^h}\} \quad \text{and} \quad \max\{0, g_{max}^{(n)}, u_{max}^0\} \le \max\{0, \max \tilde{u}_{h\,|\Gamma_{par}^h}\}\,.$$

Secondly, let us estimate the terms with f. Using (3.10) and (4.39), for any $i = 1, \ldots, m_0$ and $j = 0, 1, \ldots, n$ we have

$$f_i^j := f_i(t_j) = \int_{\Omega} f(x, t_j)\phi_i(x)\, dx \le \sup_{Q_T} f \cdot \int_{\Omega} \phi_i = \sup_{Q_T} f \cdot (\mathbf{Ae})_i\,,$$

hence, by (4.14) and (4.44),

$$f_{max}^{(n)} = \max_{\substack{i=1,\dots,m_0 \\ j=0,\dots,n}} \frac{f_i^j}{(\mathbf{Ae})_i} \le \sup_{Q_T} f\,.$$

Similarly $f_{min}^{(n)} \ge \inf_{Q_T} f$, hence in (4.45)

$$\min\{0, f_{min}^{(n)}\} \ge \min\{0, \inf_{Q_T} f\} \qquad \text{and} \qquad \max\{0, f_{max}^{(n)}\} \le \max\{0, \sup_{Q_T} f\}.$$

Altogether, we have obtained that (4.45) implies (4.47) for all u_i^n and thus, taking again the partition-of-unity property (3.1)–(3.2) into account, for $u_h(x,t)$ for all $x \in \Omega$ and $t = n\Delta t$. ■

We note that we can rewrite (4.47) similarly to (2.13):

$$\min(\tilde{u}_{h\,|\Gamma_{par}^h})^- + t \cdot \inf_{Q_T} f^- \le \tilde{u}_h(.,t) \le \max(\tilde{u}_{h\,|\Gamma_{par}^h})^+ + t \cdot \sup_{Q_T} f^+\,. \tag{4.48}$$

for $t = n\Delta t$, where $n = 0, 1, \dots, n_T$.

Remark 4.1 *(Discussion of the assumptions in Theorem 4.1.)*

(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 (4.29) is asymptotically

$$\Delta t \ge O(h^2) \tag{4.49}$$

as $h \to 0$, and the constants are similarly computable. The upper bound $1/\alpha$ is also computable.

An important special case is the discrete *nonnegativity preservation principle*:

Corollary 4.3 *Let the conditions of Theorem 4.1 hold, and let $f \ge 0$, $g \ge 0$ and $u_0 \ge 0$. Then the discrete solution satisfies*

$$\tilde{u}_h(.,t_n) \ge 0 \qquad (n = 0, 1, ..., n_T).$$

This means that $\tilde{u}^h$ is 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

$$\tilde{u}^h \ge 0 \quad \text{on} \quad Q_T. \tag{4.50}$$

5 Extensions to some more general problems

5.1 DMP for equations with superlinear nonlinearity

Now let us allow stronger growth of the nonlinearity q than in the previous subsection. Namely, a polynomial growth can be prescribed such that the limitation is the embedding of $H^1(\Omega)$ to $L^p(\Omega)$ that still allows the weak form of the PDE to make sense.

Assumptions 5.1.1. Let the item (A5) in Assumptions 2.1 be replaced by the following:

(A5)' Let $2 \le p$ if $d = 2$, or $2 \le p < \frac{2d}{d-2}$ if $d > 2$. There exist constants $\alpha, \beta \ge 0$ such that for any $(x,t) \in Q_T$ and $\xi \in \mathbb{R}$

$$0 \le \frac{\partial q(x,t,\xi)}{\partial \xi} \le \alpha + \beta |\xi|^{p-2}. \tag{5.1}$$

Further, let the other items of Assumptions 2.1 hold as they are.

For this case we need some extra technical results and assumptions. Regarding the exact solution, [16, Lemma 5.12] states that for all $t \in [0,T]$

$$\|u(.,t)\|_{L^p(\Omega)} \le \; M_T := \; \|u^0\|_{L^p(\Omega)} + T|\Omega|^{\frac{1}{p}} \|f\|_{L^\infty(Q_T)} \, . \tag{5.2}$$

Now let us summarize the additional assumptions. We let $\Omega_{ij} := \operatorname{supp} \phi_i \cap \operatorname{supp} \phi_j$ as in (4.34), and let us use notation

$$|\Omega_{ij}| := \operatorname{meas}(\Omega_{ij}).$$

Assumptions 5.1.2.

(B1) There exists $\varepsilon_0 > 0$ such that for any basis functions ϕ_i and ϕ_j

$$\int_{\Omega_{ij}} \phi_i \phi_j \; \ge \varepsilon_0 |\Omega_{ij}| \, . \tag{5.3}$$

(This is satisfied for linear, bilinear or prismatic elements if the mesh is shape regular.)

Further, one of the following conditions is satisfied: either

(B2a) the capacity of the model is finite, i.e. there exists a constant $\bar{U} > 0$ such that any solution u_h satisfies $\sup_{Q_T} |u_h| < \bar{U}$; or:

(B2b) the discretization satisfies $M_p := \sup_{t \in [0,T]} \|u(.,t) - u_h(.,t)\|_{L^p(\Omega)} < \infty$.

Theorem 5.1 *Let problem (2.1)–(2.4) satisfy Assumptions 5.1.1. Let us consider a family of finite element subspaces $\mathcal{V} = \{V_h\}_{h \to 0}$ that is regular in the sense of Definition 4.2 and the basis functions satisfy (3.1)–(3.2), further, Assumptions 5.1.2 hold. Let assumptions (i)–(iii) of Theorem 4.1 hold, further, let its assumption (iv) be replaced by the following one:*

(iv') (upper bound of the time step) *the time-step Δt satisfies the following: if (B2a) holds, then*

$$\Delta t \le \tilde{\varepsilon}_0 := \frac{\varepsilon_0}{\varepsilon_0 \alpha + \bar{U}^{p-2}\beta}\,; \tag{5.4}$$

and if (B2b) holds, then with α from (2.6),

$$\Delta t \le \frac{1}{2\alpha} \quad \textit{and} \quad \Delta t \le M_0\, h^{\frac{d(p-2)}{p}} \tag{5.5}$$

where $M_0 := \frac{\varepsilon_0}{2\beta\ (M_T+M_p)^{p-2}}\, c_1^{\frac{2-p}{p}}$.

Then the scheme (3.14) satisfies the properties P1–P5 stated in Theorem 4.1.

Proof. Compared to Theorem 4.1, the only difference in the proof is in condition P4, since that was the only part where the restricton $r \le \alpha$ was used. Now, instead,

$$0 \le r(x, t, \xi) \le \alpha + \beta |\xi|^{p-2}. \tag{5.6}$$

Hence, by (3.1),

$$B(\mathbf{u}^n)_{ij} = \int_{\Omega_{ij}} \big(1 - \Delta t\; r(x, t_n, u_h^n)\big)\, \phi_i \phi_j \ge \big(1 - \alpha\,\Delta t\big) \int_{\Omega_{ij}} \phi_i \phi_j - \beta\,\Delta t \int_{\Omega_{ij}} |u_h^n|^{p-2} \phi_i \phi_j\,. \tag{5.7}$$

For the first term above, from (5.3),

$$\big(1 - \alpha\,\Delta t\big) \int_{\Omega_{ij}} \phi_i \phi_j \ge \varepsilon_0 \big(1 - \alpha\,\Delta t\big)\, |\Omega_{ij}|\,. \tag{5.8}$$

To involve the second term, let us use Assumptions 5.2 (B2). If (B2a) holds, then

$$\int_{\Omega_{ij}} |u_h^n|^{p-2} \phi_i \phi_j \le \int_{\Omega_{ij}} |u_h^n|^{p-2} \le \bar{U}^{p-2} |\Omega_{ij}|\,.$$

This and (5.7)–(5.8) yield

$$B(\mathbf{u}^n)_{ij} \ge \Big(\varepsilon_0 - \Delta t\, \big(\varepsilon_0 \alpha + \bar{U}^{p-2}\beta\big)\Big)\, |\Omega_{ij}| \ge 0$$

by assumption (5.4). In turn, if (B2b) holds, then first we use (3.1) and the general Hölder's inequality:

$$\int_{\Omega_{ij}} |u_h^n|^{p-2} \phi_i \phi_j \le \int_{\Omega_{ij}} |u_h^n|^{p-2} \cdot 1^2 \le \|u_h^n\|^{p-2}_{L^p(\Omega_{ij})} \|1\|^2_{L^p(\Omega_{ij})}$$

$$= \|u_h^n\|^{p-2}_{L^p(\Omega_{ij})}\ |\Omega_{ij}|^{2/p} \le \|u_h^n\|^{p-2}_{L^p(\Omega)}\ |\Omega_{ij}|^{2/p}\,.$$

Here (5.2) and condition (B2b) imply

$$\|u_h^n\|_{L^p(\Omega)} \le M_T + M_p =: \bar{M}$$

and thus

$$\int_{\Omega_{ij}} |u_h^n|^{p-2}\phi_i\phi_j \leq \bar{M}^{p-2}|\Omega_{ij}|^{2/p}\,. \tag{5.9}$$

This and (5.7)–(5.8) yield

$$B(\mathbf{u}^n)_{ij} \geq \varepsilon_0\big(1-\alpha\,\Delta t\big)\,|\Omega_{ij}| \,-\beta\,\Delta t\,\bar{M}^{p-2}\,|\Omega_{ij}|^{2/p}\,.$$

Using (5.5),

$$B(\mathbf{u}^n)_{ij} \geq |\Omega_{ij}|\,\Big(\frac{\varepsilon_0}{2}\,-\beta\,\Delta t\,\bar{M}^{p-2}\,|\Omega_{ij}|^{-(1-\frac{2}{p})}\Big).$$

Here, from (4.26) and (5.5),

$$\frac{\varepsilon_0}{2}\,-\beta\,\Delta t\,\bar{M}^{p-2}\,|\Omega_{ij}|^{-(1-\frac{2}{p})} \geq \frac{\varepsilon_0}{2}\,-\beta\,\Delta t\,\bar{M}^{p-2}\,c_1^{1-\frac{2}{p}}\,h^{-d(1-\frac{2}{p})} \geq 0,$$

hence $B(\mathbf{u}^n)_{ij} \geq 0$. ∎

Remark 5.1 *(Discussion of the assumptions in Theorem 5.1.)* The conditions on the space mesh are similar as discussed in Remark 4.1, as well as the lower bound on Δt.

Compared to Remark 4.1, depending on assumptions (B2), the upper bound on Δt is either constant again, or satisfies a restriction proportional to a power of h. Assumptions 2.1 item (A5) implies

$$\tfrac{d(p-2)}{p} < 2\,,$$

that is, the exponent in (5.5) is less than 2. In fact, in various practical models we have $p \leq 3$, see section 6, in which case $\frac{d(p-2)}{p} \leq \frac{d}{3}$, thus (5.5) becomes

$$\Delta t \leq O(h^{\frac{2}{3}}) \quad \text{in}\, 2D \quad \text{and} \quad \Delta t \leq O(h) \quad \text{in}\, 3D.$$

Similarly to the sublinear case, we can derive the corresponding maximum-minimum principle as in Theorem 4.2:

Corollary 5.1 *If the conditions of Theorem 5.1 hold, then the scheme (3.14) satisfies (4.47).*

Further, in particular, the nonnegativity preservation principle holds:

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

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

5.2 DMP for cooperative systems

Now let us consider systems of parabolic PDEs. The following system is analogous to (2.1) and is a special case of system (1)-(5) in [17]. The coupling arises in the zeroth order terms in a cooperative way. We sketch how the DMP can be extended to such systems for the linearly implicit scheme. Also a main point for our scheme, discussed later in (6.8), is that it leads to decoupled and hence parallelizable problems. We refer to [17] for more technical details on the FEM solution of such problems.

We look for $u_k = u_k(x,t) \quad (k = 1, ..., s)$ such that

$$\frac{\partial u_k}{\partial t} - \operatorname{div}\big(a(x)\,\nabla u_k\big) + \mathbf{w}(x)\cdot\nabla u_k + q_k(x,t,u_1,\dots,u_s) = f_k(x,t) \quad \text{in} \quad Q_T, \tag{5.10}$$

$$u_k(x,t) = g_k(x,t) \quad \text{for} \quad (x,t) \in \Gamma_D \times [0,T], \tag{5.11}$$

$$\tfrac{\partial u}{\partial \nu} = 0 \quad \text{for} \quad (x,t) \in \Gamma_N \times [0,T], \tag{5.12}$$

$$u_k(x,0) = u_k^{(0)}(x) \quad \text{for} \quad x \in \Omega, \tag{5.13}$$

Assumptions 5.2.

(S1) The domain and the coefficients a and $\mathbf{w}$ satisfy Assumptions 2.1 (A1)-(A4).

(S2) The nonlinearities q_k have the following properties:

(a) (Balance.) Analogously to (2.7), for any $(x,t) \in Q_T$ and $k = 1, ..., s$

$$q_k(x,t,\underline{0}) = \underline{0}. \tag{5.14}$$

(b) (Boundedness.) There exists a constant $\alpha > 0$ such that for all $(x,t) \in Q_T$, $\underline{\xi} \in \mathbb{R}^s$ and $k, \ell = 1, ..., s$,

$$\left|\frac{\partial q_k}{\partial \xi_\ell}(x,t,\underline{\xi})\right| \le \alpha. \tag{5.15}$$

(c) (Cooperativity.) For all $(x,t) \in Q_T$, $\underline{\xi} \in \mathbb{R}^s$ and $k, \ell = 1, ..., s$,

$$\frac{\partial q_k}{\partial \xi_\ell}(x,t,\underline{\xi}) \le 0 \qquad \text{whenever} \quad k \ne \ell. \tag{5.16}$$

(d) (Weak diagonal dominance.) For all $(x,t) \in Q_T$, $\underline{\xi} \in \mathbb{R}^s$ and $k = 1, ..., s$,

$$\sum_{\ell=1}^{s} \frac{\partial q_k}{\partial \xi_\ell}(x,t,\underline{\xi}) \ge 0. \tag{5.17}$$

The *FEM discretization in space* for such systems is described in detail in [17, Sec. 3]. We indicate the main points compared to the scalar case. Let us start from the FEM subspaces and basis functions that had been defined in our subsection 3.1 for the scalar case. Then the natural choice for subspaces for the PDE system are the direct products of these subspaces with themselves s times. Keeping the notation V_h and $V_{h,0}$ from subsection 3.1 for the scalar subspaces, we have $(V_h)^s$ and $(V_{h,0})^s$ as product subspaces.

The basis functions in these product spaces are defined by repeating the scalar basis functions in each of the s coordinates and setting zero in the other coordinates. That is, let

$$N_0 := sm_0 \quad \text{and} \quad N := s\overline{m}\,.$$

The interior basis functions are indexed as $1 \le i \le N_0$, rewritten as

$$i = (k_0 - 1)m_0 + p \quad \text{for some } 1 \le k_0 \le s \text{ and } 1 \le p \le m_0, \tag{5.18}$$

and the boundary basis functions are indexed as $N_0 + 1 \le i \le N$, rewritten as

$$i = N_0 + (k_0 - 1)(\overline{m} - m_0) + p - m_0 \quad \text{for some } 1 \le k_0 \le s \text{ and } m_0 + 1 \le p \le \overline{m}. \tag{5.19}$$

In both cases we define

$$\phi_i := [0, \dots, 0, \varphi_p, 0, \dots, 0]^T \qquad \text{where } \varphi_p \text{ stands at the } k_0\text{th entry}, \tag{5.20}$$

(denoting again the transposed of a vector by ^{T}), that is, the mth coordinate of ϕ_i satisfies $(\phi_i)_m = \varphi_p$ if $m = k_0$ and $(\phi_i)_m = 0$ if $m \ne k_0$.

The *full discretization* is defined analogously to (3.14). Here we denote the corresponding matrices by an upper index s to indicate that their size is s times larger than in the scalar case. Thus the scheme reads as

$$(\mathbf{M}^{(s)} + \Delta t \mathbf{K}^{(s)})\, \mathbf{u}^{n+1} = \mathbf{M}^{(s)}\mathbf{u}^n - \Delta t \mathbf{Q}^{(s)}(\mathbf{u}^n) + \Delta t\, \mathbf{f}^n \tag{5.21}$$

for $n = 0, 1, \dots, n_T - 1$. Note that the mass matrix $\mathbf{M}^{(s)}$ and the linear stiffness matrix $\mathbf{K}^{(s)}$ consist of identical scalar diagonal blocks, which follows from the form of (5.10). That is,

$$\mathbf{M}^{(s)} = \text{blockdiag}(\mathbf{M}, \dots, \mathbf{M}) \quad \text{and} \quad \mathbf{K}^{(s)} = \text{blockdiag}(\mathbf{K}, \dots, \mathbf{K}) \tag{5.22}$$

where $\mathbf{M}$ and $\mathbf{K}$ are the scalar mass matrix and the linear stiffness matrix appearing in (3.14), with entries (3.9) and (3.12), respectively.

Now we can verify the analogue of Theorem 4.1 for problem (5.10)–(5.13).

Theorem 5.2 *Let the system* (5.10)–(5.13) *satisfy Assumptions 5.2. Consider a family of finite element subspaces such that V_h, h and Δt satisfy the conditions of Theorem 4.1. Then the scheme (5.21) satisfies the properties P1–P5 listed in Theorem 4.1.*

Proof. First we rewrite our problem analogously to (4.1)–(4.2): defining

$$r_{kl}(x, t, \xi) := \int_0^1 \frac{\partial q_k}{\partial \xi_l}(x, t, z\xi)\, dz \qquad (k, l = 1, \dots, s, \;\; (x, t) \in Q_T,\, \xi \in \mathbb{R}^s), \tag{5.23}$$

we obtain the weak form

$$\int_\Omega \sum_{k=1}^{s} \frac{\partial u_k}{\partial t} v_k \, dx + B(u;u,v)(t) = \langle \psi(t), v\rangle \qquad (\forall v \in H^1_D(\Omega)^s, \quad t \in (0,T)),$$

where

$$B(y;u,v)(t) := \int_\Omega \sum_{k=1}^{s} \Big(a(x)\nabla u_k \cdot \nabla v_k + (\mathbf{w}(x)\cdot \nabla u_k) v_k + \sum_{k,l=1}^{s} r_{kl}(x,t,y)\, u_l v_k \Big)\, dx \quad (5.24)$$

and $\langle \psi(t), v\rangle := \int_\Omega \sum_{k=1}^{s} f_k v_k \, dx$. We also reformulate the scheme (5.21) analogously to (4.22)–(4.23):

$$\mathbf{A}^{(s)} \mathbf{u}^{n+1} = \mathbf{B}^{(s)}(\mathbf{u}^n)\mathbf{u}^n + \Delta t \, \mathbf{f}^n, \tag{5.25}$$

where

$$\mathbf{A}^{(s)} := \mathbf{M}^{(s)} + \Delta t \mathbf{K}^{(s)}, \qquad \mathbf{B}^{(s)}(\mathbf{u}^n) := \mathbf{M}^{(s)} - \Delta t \mathbf{R}^{(\mathbf{s})}(\mathbf{u}^n). \tag{5.26}$$

Note that

$$\mathbf{A}^{(s)} = \mathrm{blockdiag}(\mathbf{A}, \ldots, \mathbf{A}). \tag{5.27}$$

Now let us write the indices $i = 1, \ldots, N_0$ and $j = 1, \ldots, N$ according to the notations (5.18)–(5.19):

$$i = (k_0 - 1)m_0 + p \quad \text{for some } 1 \le k_0 \le s \text{ and } 1 \le p \le m_0, \tag{5.28}$$

$$\begin{gathered} j = (l_0 - 1)m_0 + q \quad \text{for some } 1 \le l_0 \le s \text{ and } 1 \le q \le m_0 \text{ or} \\ j = N_0 + (l_0 - 1)(\overline{m} - m_0) + q - m_0 \quad \text{for some } 1 \le l_0 \le s \text{ and } m_0 + 1 \le q \le \overline{m}. \end{gathered} \tag{5.29}$$

Then the entries of $\mathbf{R}^{(\mathbf{s})}(\mathbf{u}^n)$ are

$$\mathbf{R}^{(\mathbf{s})}(\mathbf{u}^n)_{ij} = \int_\Omega r_{k_0\, l_0}(x, t_n, u_h^n)\, \varphi_p\, \varphi_q \, dx\,. \tag{5.30}$$

Similarly, from (5.22),

$$\mathbf{M}^{(\mathbf{s})}{}_{ij} = \begin{cases} 0 & \text{if } k_0 \ne l_0; \\ M_{pq} = \int_\Omega \varphi_p\, \varphi_q & \text{if } k_0 = l_0\,, \end{cases} \tag{5.31}$$

$$\mathbf{K}^{(\mathbf{s})}{}_{ij} = \begin{cases} 0 & \text{if } k_0 \ne l_0; \\ K_{pq} = \int_\Omega \Big(a\, \nabla\varphi_q \cdot \nabla\varphi_p + (\mathbf{w}\cdot\nabla\phi_q)\varphi_p \Big) & \text{if } k_0 = l_0\,. \end{cases} \tag{5.32}$$

Now we can verify properties P1–P5 for $\mathbf{A}^{(\mathbf{s})}$, relying on Theorem 4.1.

- P1–P3 for $\mathbf{A}^{(\mathbf{s})}$ follow readily from (5.27), since its blocks satisfy P1–P3 by Theorem 4.1, and (in P2–P3) also using that

$$\mathbf{A}_0^{(s)} = \text{blockdiag}(\mathbf{A}_0, \dots, \mathbf{A}_0).$$

- P4: Using (5.30) and (5.31), we have

$$\mathbf{B}^{(s)}(\mathbf{u}^n)_{ij} = \mathbf{M}_{ij}^{(s)} - \Delta t \mathbf{R}^{(\mathbf{s})}(\mathbf{u}^n)_{ij} = \begin{cases} -\Delta t \displaystyle\int_\Omega r_{k_0\, l_0}(x, t_n, u_h^n)\, \varphi_p\, \varphi_q & \text{if } k_0 \neq l_0; \\ \displaystyle\int_\Omega \big(1 - \Delta t\ r_{k_0\, k_0}(x, t_n, u_h^n)\big)\, \varphi_p\, \varphi_q & \text{if } k_0 = l_0\,. \end{cases}$$

Here, similarly to (4.41), the functions $r_{k\ell}$ inherit the properties of $\frac{\partial q_k}{\partial \xi_\ell}$. In particular, for all $(x,t) \in Q_T$, $\underline{\xi} \in \mathbb{R}^s$ and $k, \ell = 1, ..., s$, (5.15)–(5.16) yield

$$r_{k\ell}(x,t,\underline{\xi}) \leq 0 \quad (\text{if } k \neq \ell) \qquad \text{and} \qquad \big| r_{k\ell}(x,t,\underline{\xi}) \big| \leq \alpha.$$

Hence:

$$\text{if } k_0 \neq l_0, \quad \text{then} \qquad \mathbf{B}^{(s)}(\mathbf{u}^n)_{ij} = -\Delta t \int_\Omega r_{k_0\, l_0}(x, t_n, u_h^n)\, \varphi_p\, \varphi_q \geq 0\,;$$

$$\text{if } k_0 = l_0, \quad \text{then} \qquad \mathbf{B}^{(s)}(\mathbf{u}^n)_{ij} \geq \big(1 - \alpha\, \Delta t\big) \int_\Omega \varphi_p\, \varphi_q \geq 0.$$

- P5: From (5.17) we have for all $(x,t) \in Q_T$, $\underline{\xi} \in \mathbb{R}^s$ and $k = 1, ..., s$,

$$\sum_{\ell=1}^{s} r_{k\ell}(x,t,\underline{\xi}) \geq 0.$$

Let $i = 1, \dots, N$. First, using (4.38) and (5.22), we have

$$(\mathbf{K}^{(s)} \mathbf{e}^{(s)})_i = 0\,.$$

From the above, using (5.26), analogously to (4.42),

$$\Big(\big(\mathbf{A}^{(s)} - \mathbf{B}^{(s)}(\mathbf{u}^n)\big) \mathbf{e}^{(s)} \Big)_i = \Delta t \Big(\big(\mathbf{K}^{(s)} + \mathbf{R}^{(\mathbf{s})}(\mathbf{u}^n)\big) \mathbf{e} \Big)_i = \Delta t \left(\mathbf{R}^{(\mathbf{s})}(\mathbf{u}^n)) \mathbf{e} \right)_i =$$

$$= \Delta t \int_\Omega \sum_{l_0=1}^{s} r_{k_0\, l_0}(x,t,u_h)\, \varphi_p \,\Big(\sum_{q=1}^{\overline{m}} \varphi_q\Big) \ = \Delta t \int_\Omega \sum_{l_0=1}^{s} r_{k_0\, l_0}(x,t,u_h)\, \varphi_p \ \geq 0\,,$$

hence P5 holds. ∎

Now, from Theorem 5.2 and Proposition 4.1, we can derive the DMP, extending the scalar case in Theorem 4.2. Note that the estimations take into account the maxima/minima of the coordinate functions in analogy with [17, Thm 5.2].

Theorem 5.3 *If the conditions of Theorem 5.2 hold, then the scheme (5.21) satisfies*

$$\min\{0, \min_{k=1,\dots,s} \min \tilde{u}^h_{k\,|\Gamma^h_{par}}\} + t\cdot\min\{0, \min_{k=1,\dots,s} \inf_{Q_T} f_k\}$$
$$\le \tilde{u}^h_k(x,t) \le \max\{0, \max_{k=1,\dots,s} \max \tilde{u}^h_{k\,|\Gamma^h_{par}}\} + t\cdot\max\{0, \max_{k=1,\dots,s} \sup_{Q_T} f_k\}$$

for $x\in\Omega$ *and* $t = n\Delta t$, *where* $n = 0, 1, \dots, n_T$.

Remark 5.2 Finally we note that some parts of Assumptions 5.2 can be relaxed.

First, one may allow power growth of the nonlinear terms, analogously to (5.1) in subsection 5.1. For the same restrictions on p as in Assumptions 5.1.1, we may only require that

$$\left|\frac{\partial q_k}{\partial \xi_l}(x,t,\xi)\right| \le \alpha + \beta|\xi|^{p-2} \tag{5.33}$$

for all $(x,t)\in Q_T$, $\underline{\xi}\in\mathbb{R}^s$ and $k,\ell = 1,...,s$. In this case an analogue of Theorem 5.1 can be formulated.

Further, we may allow different coefficients a_k, $\mathbf{w}_k$ in the linear part of the equations in (5.10), moreover, their dependence on t can be also allowed. With uniform bounds one may ensure the DMP under the same conditions as in the scalar case.

5.3 Boundary source terms

As we have seen, the homogeneous Neumann condition $\frac{\partial u}{\partial \nu} = 0$ on $\Gamma_N\times[0,T]$ in (5.12) corresponds to a natural b.c. induced by the weak form (2.8). In the presence of a boundary source term $\gamma\in L^\infty(\Gamma_N\times[0,T])$, we have an inhomogeneous condition

$$\frac{\partial u}{\partial \nu} = \gamma \quad \text{on} \quad \Gamma_N\times[0,T].$$

The DMP (4.45) on discrete level is still valid under the conditions of Theorem 4.1. In turn, the r.h.s estimate of (4.47) is modified in the same way as for the fully implicit scheme, see [17, Lemma 5.1]. In the scalar case the factor $t\cdot\max\{0, \sup_{Q_T} f\}$ is replaced by

$$t\cdot\max\{0, \sup_{Q_T} f, \quad D(h)\cdot \sup_{\Gamma_N\times[0,T]} \gamma\}, \qquad \text{where} \quad D(h) := \max_{p=1,\dots,\overline{m}} \frac{\int_{\Gamma_N}\varphi_p\,d\sigma}{\int_\Omega \varphi_p\,dx},$$

and in the case of systems the additional term involves also $\sup_{k=1,\dots,s} \gamma_k$, furthermore, the same modification applies to the terms with minimum.

5.4 Geometric properties of the space mesh

In order to satisfy condition (4.27), the most direct way is to require

$$\nabla\phi_i\cdot\nabla\phi_j \le -\frac{\sigma}{h^2} \tag{5.34}$$

pointwise on the common support $\Omega_{ij} := \operatorname{supp}\phi_i \cap \operatorname{supp}\phi_j$ of these basis functions. This condition means a straightforward geometric restriction in the case of simplicial meshes: (5.34) requires that the family of meshes is uniformly acute [4, 5]. In the case of bilinear elements, condition (4.27) is equivalent to the "strict non-narrowness" of the meshes [14, 22]. Constructions of such meshes are given e.g. in [4, 9].

The above pointwise condition can be weakened [24, 28]. For instance, one can require (5.34) only on a proper subpart of each intersection of supports with asymptotically nonvanishing measure, that is, let there exist subsets $\Omega_{ij}^+ \subset \Omega_{ij}$ for all i and j such that the basis functions satisfy

$$\nabla\phi_i \cdot \nabla\phi_j \leq -\frac{\sigma}{h^2} < 0 \quad \text{on} \quad \Omega_{ij}^+, \qquad \nabla\phi_i \cdot \nabla\phi_j \leq 0 \quad \text{on} \quad \Omega_{ij} \setminus \Omega_{ij}^+, \tag{5.35}$$

and we have

$$\frac{\operatorname{meas}\left(\Omega_{ij}^+\right)}{\operatorname{meas}(\Omega_{ij})} \geq c_3 > 0. \tag{5.36}$$

Then (4.27) still holds. This condition for simplicial meshes means that some possible right angles are allowed, which makes the refinement easier.

6 Examples

We give some real-life examples where discrete maximum-minimum principles, and in particular, discrete nonnegativity can be derived for suitable discretizations.

6.1 Scalar equations

(a) Reaction-diffusion problems

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

$$\frac{\partial u}{\partial t} - \operatorname{div}\left(a(x)\nabla u\right) + q(x,u) = f(x,t) \tag{6.1}$$

in $Q_T := \Omega \times (0,T)$ with boundary and initial conditions (2.2)–(2.4). The function u describes the concentration of the considered chemical material or pollutant, hence

$$u \geq 0.$$

Here the coefficients a, f, g and u_0 are bounded nonnegative measurable functions and a has a positive lower bound. Further, q desribes the rate of reaction in the body, hence $q(x,0) = 0$ for all x. In various examples the reaction process is such that q grows along with u, further, the rate is at most polynomial, i.e. the growth condition (5.1) is satisfied.

The nonlinearity in such problems is only defined for nonnegative arguments u, hence one has to extend it for $u \leq 0$ to obtain an operator to be used in the numerical process. In the case $q(x,0) = 0$, this is done, e.g., by the formula

$$q(x,-u) = -q(x,u). \tag{6.2}$$

However, finally, using Corollary 4.3 or 5.2, we will obtain that the discrete solution satisfies $\min_{\overline{\Omega}} \tilde{u}_h \geq 0$. That is, $\tilde{u}_h$ is the solution of the problem with the original nonlinearity and preserves the physical meaning.

Some concrete situations are as follows:

- Michaelis-Menten reaction in enzyme kynetics [7, 26]:

$$q(x,u) = \frac{1}{\varepsilon}\frac{u}{u+\kappa} \qquad (u \geq 0)$$

  describes the rate of reaction, where $\kappa, \varepsilon > 0$ are constants. Using (6.2), we can rewrite the rate as

$$q(x,u) = \frac{1}{\varepsilon}\frac{u}{|u|+\kappa} \qquad (u \in \mathbb{R}).$$

  Then

$$r(x,u) = \frac{1}{\varepsilon(|u|+\kappa)} \leq \frac{1}{\varepsilon\kappa}$$

  hence Theorem 4.2 and Corollary 4.3 can be used. In particular, the condition on the time-step is

$$O(h^2) \leq \Delta t \leq O(1).$$

- A generalized case of the above is

$$q(x,u) = \frac{1}{\varepsilon}\frac{u}{(u+\kappa)^\beta}$$

  where $0 < \beta < 1$. Then, for large u, letting $\sigma = 1-\beta$, we have the order

$$q(x,u) \sim const. \cdot u^\sigma$$

  where $0 < \sigma < 1$, which is typical for some autocatalytic chemical reactions [7]. Then, after rewriting as before,

$$r(x,u) = \frac{1}{\varepsilon(|u|+\kappa)^\beta} \leq \frac{1}{\varepsilon\kappa^\beta}$$

  hence Theorem 4.2 and Corollary 4.3 can be used. In particular, the condition on the time-step is

$$O(h^2) \leq \Delta t \leq O(1).$$

- In general, the rate of reaction is limited by the given quantity up to some rate constant, hence (5.1) becomes

$$0 \leq \frac{\partial q(x,\xi)}{\partial \xi} \leq \alpha + \beta|\xi|. \tag{6.3}$$

  For instance, the decay of hypochlorite ions leads exactly to the reaction rate $q(x,u) = ku^2$ for $u \geq 0$ with some constant $k > 0$ (see [2]), i.e., using (6.2),

$$q(x,\xi) = k|\xi|\xi \qquad (\xi \in \mathbb{R}), \quad \text{and thus} \quad \frac{\partial q(x,\xi)}{\partial \xi} = 2k|\xi|.$$

The bound (6.3) corresponds to the case $p = 3$, hence and Corollaries 5.1 and 5.2 can be used. As mentioned in Remark 5.1, the upper bound for the time-step is

$$\Delta t \leq O(h^{\frac{2}{3}}) \ \text{ in } 2D \quad \text{and} \quad \Delta t \leq O(h) \ \text{ in } 3D.$$

(b) Transport problems

Several transport processes involve a convection field and also a nonlinear reaction term for the considered material. Then one has a model similar to (6.4) but with a first-order convection term:

$$\frac{\partial u}{\partial t} - \operatorname{div}\big(a(x)\nabla u\big) + \mathbf{w}(x)\cdot\nabla u + q(x,u) = f(x,t) \tag{6.4}$$

where $\mathbf{w}$ describes the convection field satisfying Assumptions 2.1.

Then, depending on the growth of the reaction term q, Theorem 4.2 and Corollary 4.3 or Corollaries 5.1 and 5.2, respectively, can be applied.

(c) On the obtained qualitative properties

Some consequences of the above results are as follows.

In the given examples we have $g \geq 0$ and $u_0 \geq 0$ for physical reasons. Therefore, under the imposed conditions, based on (4.50), the discrete nonnegativity principle holds:

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

In particular, if the r.h.s $f \equiv 0$, i.e. there is no outer source, then (4.47) also yields the upper bound

$$\tilde{u}_h \leq \ \max \tilde{u}_{h\,|\Gamma^h_{par}} \equiv \max\{\max u^h_0,\ \max g^h\} \quad \text{on} \quad Q_T.$$

6.2 Systems

Here we sketch two types of models for systems of parabolic PDEs that were described in more detail in [17] in the context of fully implicit schemes. It was seen therein that they satisfy Assumptions 5.2 (allowing (5.33) when applicable).

(a) Population dynamics

Some model systems in population dynamics can be given in the form

$$\begin{cases} \dfrac{\partial u_1}{\partial t} - \Delta u_1 = u_1\, M_1(u_1,u_2), \\[2ex] \dfrac{\partial u_2}{\partial t} - \Delta u_2 = u_2\, M_2(u_1,u_2), \end{cases} \tag{6.5}$$

where u_1, u_2 describe the quantity of two species modeled as distributed continuously in a plane region Ω, see e.g. [10]. Proper boundary and initial conditions hold. The species are considered to live in symbiosis, which means that

$$\partial_2 M_1 \geq 0 \qquad \text{and} \qquad \partial_1 M_2 \geq 0. \tag{6.6}$$

This implies cooperativity of the system. Conditions on weak diagonal dominance are discussed in [17, Sec 6.4].

(b) Reaction-diffusion(-convection) systems

Reaction-diffusion-convection processes are often described by systems of the following form:

$$\frac{\partial u_k}{\partial t} - \Delta u_k + \mathbf{w}(x) \cdot \nabla u_k + \ P_k(x, u_1, \ldots, u_s) = f_k(x,t) \quad \text{in} \ \ Q_T \qquad (k = 1, \ldots, s) \tag{6.7}$$

in a domain $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3. Such systems also model air pollution transport. In reaction-diffusion models in chemistry sometimes no convection is considered, in which case $\mathbf{w} \equiv 0$.

Here P_k (typically a polynomial) characterizes the rate of the reactions involving the k-th species. As discussed in [17, Sec 6.2], most items of Assumptions 5.2 hold in a natural way. In particular, the so-called mass action kinetics [20] implies that $P_k(x, 0) \equiv 0$ on Ω for all k. In turn, the cooperativity and weak diagonal dominance conditions mean that we consider chemical processes with cross-catalysis and strong autoinhibiton.

(c) Qualitative properties and efficiency of the proposed scheme

Similarly as mentioned in the scalar case, in the given examples the initial-boundary data and the source terms are all nonnegative for physical reasons. Therefore, under the imposed conditions, the discrete nonnegativity principle holds: all coordinate functions satisfy

$$\tilde{u}_k^h \geq 0 \quad (k = 1, \ldots, s) \quad \text{on} \quad Q_T.$$

We also underline the efficiency of the linearly implicit scheme for systems of PDEs. Some of these systems have a large number of components, e.g. in air pollution models (6.7) one may have around 30 or even 100 species, depending on the involved materials [36]. Now recall the scheme (5.21):

$$\left(\mathbf{M}^{(s)} + \Delta t \mathbf{K}^{(s)}\right) \mathbf{u}^{n+1} = \mathbf{M}^{(s)} \mathbf{u}^n - \Delta t \mathbf{Q}^{(s)}(\mathbf{u}^n) + \Delta t \ \mathbf{f}^n \tag{6.8}$$

for $n = 0, 1, \ldots, n_T - 1$. Thanks to (5.22), our scheme yields that $\mathbf{M}^{(s)} + \Delta t \mathbf{K}^{(s)}$ is a block diagonal matrix, hence the linear system (6.8) is decoupled to s independent systems of s times smaller size: we have to solve

$$\left(\mathbf{M} + \Delta t \mathbf{K}\right) \mathbf{u}_k^{n+1} = \mathbf{M} \mathbf{u}_k^n - \Delta t \left(\mathbf{Q}^{(s)}(\mathbf{u}^n)\right)_k + \Delta t \ \mathbf{f}_k^n \qquad (k = 1, \ldots, s),$$

which can be executed in parallel. This is much favourable, compared to the fully implicit scheme [17], for problems like the above-mentioned ones where s is large.

Acknowledgements: This research has been supported by the Hungarian National Research, Development and Innovation Fund (NKFIH) at the Ministry of Innovation and Technology, under the funding scheme ELTE TKP 2021-NKTA-62 and the grant no. K137699.

References

[1] G. Akrivis, M. Crouzeix, *Linearly Implicit Methods for Nonlinear Parabolic Equations*, Mathematics of Computation 73:(246), pp. 613–635 (2004).

[2] P. Atkins, J. de Paula, *Atkins' Physical Chemistry*, Oxford University Press, 2002.

[3] G. R. Barrenechea, V. John, P. Knobloch, *Finite element methods respecting the discrete maximum principle for convection-diffusion equations*, SIAM Rev. 66 (2024), no. 1, 388.

[4] J. Brandts, S. Korotov, M. Křížek, J. Šolc, *On Nonobtuse Simplicial Partitions*, SIAM Rev. 51 (2009), No. 2, pp. 317-335.

[5] J. Brandts, S. Korotov, M. Křížek, *Simplicial partitions with applications to the finite element method*, Springer, Cham, 2020

[6] J. Csóka, I. Faragó, R. Horváth, J. Karátson, S. Korotov, *Qualitative properties of nonlinear parabolic operators II: the case of PDE systems*, J. Math. Anal. Appl. 468 (2018), no. 1, 64–86.

[7] 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.

[8] Q. Du, L. Ju, X. Li, Zh. Qiao, *Maximum bound principles for a class of semilinear parabolic equations and exponential time-differencing schemes*, SIAM Rev. 63 (2021), no. 2, 317–359.

[9] D. Eppstein, J. M. Sullivan, A. Üngör, *Tiling space and slabs with acute tetrahedra*, Comput. Geom. 27 (2004), no. 3, 237–255.

[10] D. J. Estep, M. G. Larson, R. D. Williams, *Estimating the error of numerical solutions of systems of reaction-diffusion equations*, Mem. Amer. Math. Soc. 146 (2000), no. 696, viii+109 pp.

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

[12] I. Faragó, R. Horváth, *Continuous and Discrete Parabolic Operators and their Qualitative Properties*, IMA J. Numer. Anal. 29 (2009), pp. 606-631.

[13] I. Faragó, R. Horváth, J. Karátson, S. Korotov, *Qualitative properties of nonlinear parabolic operators*, J. Math. Anal. Appl. 448 (2017) 473–497.

[14] 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.

[15] 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.

[16] I. Faragó, J. Karátson, S. Korotov, Discrete maximum principles for the FEM solution of some nonlinear parabolic problems, *ETNA Electronic Transactions on Numerical Analysis* 36. (2009).

[17] I. Faragó, J. Karátson, S. Korotov, Discrete maximum principles for nonlinear parabolic PDE systems, *IMA J. Numer. Anal.* 32:(4) pp. 1541-1573. (2012)

[18] M. Frittelli, A. Madzvamuse, I. Sgura, Ch. Venkataraman, Preserving invariance properties of reaction-diffusion systems on stationary surfaces, *IMA J. Numer. Anal.* 39 (2019), no. 1, 235270.

[19] 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.

[20] Hárs, V., Tóth, J., On the inverse problem of reaction kinetics, In: *Qualitative Theory of Differential Equations* (Szeged, Hungary, 1979), Coll. Math. Soc. János Bolyai 30, ed. M. Farkas, North-Holland - János Bolyai Mathematical Society, Budapest, 1981, pp. 363-379.

[21] A. Hannukainen, S. Korotov, T. Vejchodský, *Discrete Maximum Principles for FE Solutions of the Diffusion-Reaction Problem on Prismatic Meshes*, J. Comput. Appl. Math. 226 (2009), pp. 275–287.

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

[23] 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.

[24] J. Karátson, S. Korotov, *A Discrete Maximum Principle in Hilbert Space with Applications to Nonlinear Cooperative Elliptic Systems*, SIAM Numer. Anal. 47 (2009), No. 4, pp. 2518-2549.

[25] 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.

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

[27] P. Knabner, L. Angermann, *Numerical methods for elliptic and parabolic partial differential equations*, second extended edition, Springer, Cham, 2021.

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

[29] S. Korotov, M. Křížek, *Discrete maximum principles in finite element modelling*, Springer-Verlag, Berlin, 2004,

[30] 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.

[31] C. V. Pao, *Nonlinear parabolic and elliptic equations*, Plenum Press, New York, 1992.

[32] Thomée, V. , *Galerkin Finite Element Methods for Parabolic Problems*, Springer, Berlin, 1997.

[33] T. Stys, T. Motsumi, O. Daman, *On the maximum principle and its application to diffusion equations*, Numer. Methods Part. Diff. Eqns. 23 (2007), pp. 60-72.

[34] K. Wu, Ch-W. Shu, *Geometric quasilinearization framework for analysis and design of bound-preserving schemes*, SIAM Rev. 65 (2023), no. 4, 10311073.

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

[36] Z. Zlatev, *Computer treatment of large air pollution models*, Kluwer Academic Publishers, Dordrecht-Boston-London, 1995.