Discrete nonnegativity for nonlinear cooperative parabolic PDE systems with non-monotone coupling

István Faragó[1,2], János Karátson[1,2], Sergey Korotov[3,4,5]

September 1, 2016

[1] ELTE University Budapest, Department of Applied Analysis and
MTA-ELTE Numerical Analysis and Large Networks Research Group; Budapest,
Hungary
e-mail: {farago, karatson}@cs.elte.hu

[2] Department of Analysis, Technical University; Budapest, Hungary

[3] BCAM – Basque Center for Applied Mathematics
Mazarredo, 14, E–48009 Bilbao, Basque Country, Spain
e-mail: korotov@bcamath.org
[4] IKERBASQUE, Basque Foundation for Science
E–48011, Bilbao, Basque Country, Spain
[5] Department of Computing, Mathematics and Physics, Bergen University College,
Inndalsveien 28, 5020 Bergen, Norway

Abstract: Discrete nonnegativity principles are established for finite element approximations of nonlinear parabolic PDE systems with mixed boundary conditions. Previous results of the authors are extended such that diagonal dominance (or essentially monotonicity) of the nonlinear coupling can be relaxed, allowing to include much more general situations in suitable models.

Keywords: nonlinear parabolic system, discrete maximum principle, finite element method, acute simplicial meshes

Mathematics Subject Classification: 65M60, 65M50, 35B50

1 Introduction

The numerical solution of parabolic partial differential equations or systems of equations is a widespread task in numerical analysis, see, e.g., [11, 12, 13]. The discrete solution is naturally required to reproduce the basic qualitative properties of the exact solution, e.g. the maximum principle.

In our recent paper [8], we proved the discrete analogue (discrete maximum principle, or DMP, in short) of the maximum principle for the case of finite element space discretizations for some nonlinear parabolic PDE systems. Besides standard general smoothness and growth conditions, we assumed cooperativity and diagonal dominance for the nonlinear coupling of the equations. Whereas cooperativity seems to be an inherent property behind the DMP, the diagonal dominance (which implies monotonicity of the coupling vector function) is a strong assumption which was only technical. In the present paper diagonal dominance is not assumed, we only require instead that the lower bound of the sums of Jacobians does not deteriorate as t or $|\xi|$ tends to infinity.

We prove the discrete nonnegativity of the solution, which is a special case of the discrete maximum principle. In fact, we consider a more special situation in some other respects as compared to the problem in [8]. First, we only include mixed boundary conditions but no interface conditions inside the domain. Second, we only use implicit time discretization instead of a more general θ-method. We note, however, that the results of [8] using θ-methods were more restrictive for $\theta < 1$ than for the implicit case $\theta = 1$: for the latter it sufficed to assume the lower estimate $\Delta t \geq ch^2$, whereas for the former one needed the two-sided estimate $\Delta t = O(h^2)$. Moreover, in the present paper we do not even require $\Delta t \geq ch^2$, instead, we assume $\Delta t \leq 1/\mu_0$, where $-\mu_0$ is the lower bound of the sums of Jacobians. This also shows that if especially $\mu_0 = 0$ (i.e. the coupling is diagonally dominant as in [8]) then no restriction remains on Δt, i.e. the new result improves the previous one in this respect, too.

The paper is organized as follows. We first summarize the problem and then its discretization, the latter built on [8]. Then we give some background on elliptic DMPs, and based on it, we derive the desired result for our parabolic system.

2 The class of problems

In this paper we consider the following type of nonlinear parabolic systems, involving cooperative and weakly diagonally dominant coupling, nonsymmetric terms and mixed boundary conditions. Find a vector function $u = u(x,t) = (u_1(x,t), \dots, u_s(x,t))$ such that for all $k = 1, \dots, s$,

$$\frac{\partial u_k}{\partial t} - \operatorname{div}\big(a_k(x,t,\nabla u)\nabla u_k\big) + \mathbf{w}_k(x,t)\cdot\nabla u_k + q_k(x,t,u) = f_k(x,t) \quad \text{for } (x,t) \in Q_T := \Omega\times(0,T), \tag{1}$$

where Ω is a bounded domain in $\mathbf{R}^d$ and $T > 0$, further, the boundary and initial conditions are as follows ($k = 1, \dots, s$):

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

$$a_k(x,t,\nabla u)\tfrac{\partial u_k}{\partial \nu} = \gamma_k(x,t) \quad \text{for} \quad (x,t) \in \Gamma_N \times [0,T], \tag{3}$$

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

respectively, where ν stands for the outer normal vector. We impose the following

Assumptions 2.1.

(A1) (Domain.) Ω is a bounded polytopic domain in $\mathbf{R}^d$; $\Gamma_N, \Gamma_D \subset \partial\Omega$ are disjoint open measurable subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$.

(A2) (Smoothness.) For all $k = 1, \dots, s$, we have scalar functions $a_k \in (C^1 \cap L^\infty)(Q_T \times \mathbf{R}^{d\times s})$ and $q_k \in C^1(Q_T \times \mathbf{R}^s)$. Further, $\mathbf{w}_k \in PC^1(Q_T)$, $f_k \in PC(Q_T)$, $\gamma_k \in PC(\Gamma_N \times [0,T])$, $g_k \in PC(\Gamma_D \times [0,T])$ and $u_k^{(0)} \in PC(\Omega)$.

(A3) (Coercivity.) There exists a constant μ_0 such that $a_k(x,t,\eta) \geq \mu_0 > 0$ for all $k = 1, \dots, s$ and all $(x,t,\eta) \in \Omega \times (0,T) \times \mathbf{R}^{d\times s}$, further, the Jacobian matrices $\frac{\partial}{\partial \eta}\Big(a_k(x,t,\eta)\eta\Big)$ are uniformly spectrally bounded from both below and above. Finally, for all $k = 1, \dots, s$, we have $\operatorname{div} \mathbf{w}_k \leq 0$ on Ω, $\mathbf{w}_k \cdot \nu \geq 0$ on Γ_N.

(A4) (Growth.) Let $2 \leq p$ if $d = 2$ and $2 \leq p < \frac{2d}{d-2}$ if $d > 2$. There exist constants $\alpha, \beta \geq 0$ such that for any $x \in \Omega$, $t \in (0,T)$, $\xi \in \mathbf{R}^s$, and any $k, l = 1, \dots, s$,

$$\left|\frac{\partial q_k}{\partial \xi_l}(x,t,\xi)\right| \leq \alpha + \beta|\xi|^{p-2}. \tag{5}$$

(A5) (Cooperativity.) For all $k, l = 1, \dots, s$, $x \in \Omega$, $t \in (0,T)$, $\xi \in \mathbf{R}^s$,

$$\frac{\partial q_k}{\partial \xi_l}(x,t,\xi) \leq 0 \qquad \text{whenever } k \neq l. \tag{6}$$

(A6) (Boundedness below for the Jacobians w.r.t. rows and columns.) There exists a number $\mu_0 \geq 0$ such that for all $k = 1, \dots, s$, $x \in \Omega$, $t \in (0,T)$, $\xi \in \mathbf{R}^s$,

$$\sum_{l=1}^{s} \frac{\partial q_k}{\partial \xi_l}(x,t,\xi) \geq -\mu_0, \qquad \sum_{l=1}^{s} \frac{\partial q_l}{\partial \xi_k}(x,t,\xi) \geq -\mu_0. \tag{7}$$

Remark 2.1 (i) In the previous paper [8] we assumed (7) with $\mu_0 := 0$, i.e. the diagonal dominance. Now that strong assumption is relaxed, and we only require that the lower bound of the sums of Jacobians does not deteriorate as t or $|\xi|$ tends to infinity.

(ii) Assumptions (A5)-(A6) imply for all $k = 1, \dots, s$, $x \in \Omega$, $t \in (0,T)$, $\xi \in \mathbf{R}^s$,

$$\frac{\partial q_k}{\partial \xi_k}(x,t,\xi) \geq -\mu_0. \tag{8}$$

We will define weak solutions in a usual way as follows. Let

$$H_D^1(\Omega) := \{u \in H^1(\Omega) : \ u_{|\Gamma_D} = 0\}.$$

A function $u : Q_T \to \mathbf{R}^s$ is called the weak solution of the problem (1)–(4) if for all $k = 1, \dots, s$, u_k are continuously differentiable with respect to t and $u_k(.,t) \in H_D^1(\Omega)$ for all $t \in (0,T)$, and satisfy the relation

$$\int_\Omega \sum_{k=1}^{s} \frac{\partial u_k}{\partial t} v_k \, dx + \int_\Omega \sum_{k=1}^{s} \Big(a_k(x,t,\nabla u)\nabla u_k \cdot \nabla v_k + (\mathbf{w}_k(x,t) \cdot \nabla u_k) v_k + q_k(x,t,u) v_k\Big)\, dx \tag{9}$$

$$= \int_\Omega \sum_{k=1}^{s} f_k v_k \, dx + \int_\Gamma \sum_{k=1}^{s} \gamma_k v_k \, d\sigma \qquad (\forall v \in H_D^1(\Omega)^s, \quad t \in (0,T)),$$

further,

$$u_k = g_k \quad \text{on} \quad [0,T] \times \Gamma_D, \qquad u_k|_{t=0} = u_k^{(0)} \quad \text{in} \quad \Omega. \tag{10}$$

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

3 Discretization scheme

The full discretization of problem (1)–(4) is built up in the same way as in [8]. It includes two standard steps in space and time; in addition, suitable vector basis functions are involved. In this section we summarize this process.

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 in the following way. First, let $\bar{n}_0 \leq \bar{n}$ be positive integers and let us choose basis functions

$$\varphi_1, \dots, \varphi_{\bar{n}_0} \in H_D^1(\Omega), \qquad \varphi_{\bar{n}_0+1}, \dots, \varphi_{\bar{n}} \in H^1(\Omega) \setminus H_D^1(\Omega), \tag{11}$$

which correspond to homogeneous and inhomogeneous boundary conditions on Γ_D, respectively. These basis functions are assumed to be continuous and to satisfy

$$\varphi_p \geq 0 \quad (p = 1, \dots, \bar{n}), \qquad \sum_{p=1}^{\bar{n}} \varphi_p \equiv 1, \tag{12}$$

further, that there exist node points $B_p \in \Omega \cup \Gamma_N$ $(p = 1, \dots, \bar{n}_0)$ and $B_p \in \Gamma_D$ $(p = \bar{n}_0 + 1, \dots, \bar{n})$ such that

$$\varphi_p(B_q) = \delta_{pq} \tag{13}$$

where δ_{pq} is the Kronecker symbol; and finally, there exists a constant $c_{grad} > 0$ (independent of the basis functions) such that

$$\max |\nabla \varphi_p| \leq \frac{c_{grad}}{diam(\operatorname{supp} \varphi_p)} \qquad (p = 1, \dots, \bar{n}), \tag{14}$$

where supp denotes the support, i.e. the closure of the set where the function does not vanish, and diam stands for the diameter. These conditions hold e.g. for standard linear, bilinear, or prismatic finite elements.

We in fact need a basis in the corresponding product spaces, which we define by repeating the above functions in each of the s coordinates and setting zero in the other coordinates. That is, let $N_0 := s\bar{n}_0$ and $N := s\bar{n}$. First, for any $1 \leq i \leq N_0$,

$$\text{if } \ i = (k_0 - 1)\bar{n}_0 + p \quad \text{for some } 1 \leq k_0 \leq s \text{ and } 1 \leq p \leq \bar{n}_0, \ \text{ then}$$

$$\phi_i := (0, \dots, 0, \varphi_p, 0, \dots, 0) \qquad \text{where } \ \varphi_p \ \text{ stands at the } k_0\text{th entry}, \tag{15}$$

that is, the mth coordinate of ϕ_i satisfies $(\phi_i)_m = \varphi_p$ if $m = k_0$ and $(\phi_i)_m = 0$ if $m \neq k_0$. From these, we let

$$V_h^0 := \text{span}\{\phi_1, ..., \phi_{N_0}\} \subset H^1_D(\Omega)^s. \tag{16}$$

Similarly, for any $N_0 + 1 \leq i \leq N$,

$$\text{if } \ i = N_0 + (k_0 - 1)(\bar{n} - \bar{n}_0) + p - \bar{n}_0 \quad \text{for some } 1 \leq k_0 \leq s \text{ and } \bar{n}_0 + 1 \leq p \leq \bar{n}, \ \text{ then}$$

$$\phi_i := (0, \dots, 0, \varphi_p, 0, \dots, 0)^T \qquad \text{where } \ \varphi_p \ \text{ stands at the } k_0\text{th entry,} \tag{17}$$

that is, the mth coordinate of ϕ_i satisfies $(\phi_i)_m = \varphi_p$ if $m = k_0$ and $(\phi_i)_m = 0$ if $m \neq k_0$. From (16) and these, we let

$$V_h := \text{span}\{\phi_1, ..., \phi_N\} \subset H^1(\Omega)^s. \tag{18}$$

Using the above FEM subspaces, one can define the semidiscrete problem for (9) with initial-boundary conditions (10). We look for a vector function $u_h = u_h(x,t)$ that satisfies (9) for all vector functions $v_h = (v_1, \dots, v_s) \in V_h^0$, and the conditions

$$u^h(x,0) = u^{(0),h}(x) \quad (x \in \Omega), \qquad u^h(.,t) - g^h(.,t) \in V_h^0 \quad (t \in (0,T))$$

must hold. In the above formulae, the functions $u_k^{(0),h}$ and $g_k^h(.,t)$ (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 the kth coordinate g_k^h:

$$g_k^h(x,t) = \sum_{p=1}^{\bar{n}_\partial} g_p^{(k)}(t)\, \varphi_{\bar{n}_0+p}(x), \tag{19}$$

where $g_p^{(k)}(t) = g_k(B_{\bar{n}_0+p}, t)$ and

$$\bar{n}_\partial := \bar{n} - \bar{n}_0.$$

We seek the kth coordinate function u_k of the numerical solution in the form

$$u_k^h(x,t) = \sum_{p=1}^{\bar{n}} u_p^{(k)}(t)\, \varphi_p(x) + g_k(x,t) = \sum_{p=1}^{\bar{n}_0} u_p^{(k)}(t)\, \varphi_p(x) + \sum_{p=1}^{\bar{n}_\partial} g_p^{(k)}(t)\, \varphi_{\bar{n}_0+p}(x), \tag{20}$$

where the coefficients $u_p^{(k)}(t)$ $(p = 1, \dots, \bar{n}_0)$ are unknown. The set of all coefficient functions will be ordered in the following vector:

$$\begin{aligned}\mathbf{u}^h(t) = \big(\, & u_1^{(1)}(t), \dots, u_{\bar{n}_0}^{(1)}(t);\ u_1^{(2)}(t), \dots, u_{\bar{n}_0}^{(2)}(t); \dots;\ u_1^{(s)}(t), \dots, u_{\bar{n}_0}^{(s)}(t); \\ & g_1^{(1)}(t), \dots, g_{\bar{n}_\partial}^{(1)}(t);\ g_1^{(2)}(t), \dots, g_{\bar{n}_\partial}^{(2)}(t); \dots;\ g_1^{(s)}(t), \dots, g_{\bar{n}_\partial}^{(s)}(t)\big)^T\end{aligned} \tag{21}$$

(where ^{T} denotes the transposed of a vector), that is, $\mathbf{u}^h(t)$ has $N_0 = s\bar{n}_0$ coordinates from $u_1^{(1)}(t)$ to $u_{\bar{n}_0}^{(s)}(t)$ belonging to the points in the interior or on Γ_N, and then

$N - N_0 = s(\bar n - \bar n_0)$ coordinates from $g_1^{(1)}(t)$ to $g_{\bar n_\partial}^{(s)}(t)$ belonging to the boundary points on Γ_D, such that the upper index from 1 to s gives the number of coordinate in the parabolic system. For the second subvector of (21), we use the obvious notation $\mathbf{g}^h(t) = \big(g_1^{(1)}(t), \dots, g_{\bar n_\partial}^{(1)}(t);\ g_1^{(2)}(t), \dots, g_{\bar n_\partial}^{(2)}(t); \dots;\ g_1^{(s)}(t), \dots, g_{\bar n_\partial}^{(s)}(t)\big)^T$. We will also use the notations

$$\mathbf{u}^{(k_0)}(t) := \big(\,u_1^{(k_0)}(t), \dots, u_{\bar n_0}^{(k_0)}(t)\big), \qquad \mathbf{g}^{(k_0)}(t) := \big(\,g_1^{(k_0)}(t), \dots, g_{\bar n_\partial}^{(k_0)}(t)\big)$$

for any fixed $k_0 = 1, \dots, s$, to denote the corresponding sub-$\bar n_0$-tuples of $\mathbf{u}^h(t)$ and sub-$\bar n_\partial$-tuples of $\mathbf{g}^h(t)$, respectively.

To find the function $\mathbf{u}^h(t)$, first note that it is sufficient that u_h satisfies (9) for $v = \phi_i$ only $(i = 1, 2, \dots, N_0)$. Writing the index i in the following form as before:

$$i = (k_0 - 1)\bar n_0 + p \quad \text{for some } 1 \le k_0 \le s \text{ and } 1 \le p \le \bar n_0, \tag{22}$$

the function $v = \phi_i$ has kth coordinates $v_k = \delta_{k,k_0}\varphi_p$ (where δ_{k,k_0} is the Kronecker symbol) for $k = 1, \dots, s$, hence (9) yields

$$\int\limits_\Omega \frac{\partial u_{k_0}}{\partial t}\varphi_p\,dx + \int\limits_\Omega \Big(a_{k_0}(x,t,\nabla u)\nabla u_{k_0}\cdot\nabla\varphi_p + (\mathbf{w}_{k_0}(x,t)\cdot\nabla u_{k_0})\varphi_p + q_{k_0}(x,t,u)\varphi_p\Big)\,dx \tag{23}$$

$$= \int\limits_\Omega f_{k_0}\varphi_p\,dx + \int\limits_\Gamma \gamma_{k_0}\varphi_p\,d\sigma \qquad (1 \le k_0 \le s,\ 1 \le p \le \bar n_0).$$

For fixed k_0, using (20), the first integral in (23) becomes $\bar{\mathbf{M}}\,[\frac{\mathrm{d}\mathbf{u}^{(k_0)}}{\mathrm{d}t}, \frac{\mathrm{d}\mathbf{g}^{(k_0)}}{\mathrm{d}t}]$, where

$$\bar{\mathbf{M}} = [M_{pq}]_{\bar n_0 \times \bar n}, \quad M_{pq} = \int_\Omega \varphi_p\,\varphi_q\,dx.$$

We shall use the corresponding partition

$$\bar{\mathbf{M}} = [\bar{\mathbf{M}}_0|\bar{\mathbf{M}}_\partial], \quad \text{where} \quad \bar{\mathbf{M}}_0 \in \mathbf{R}^{\bar n_0 \times \bar n_0}, \ \ \bar{\mathbf{M}}_\partial \in \mathbf{R}^{\bar n_0 \times \bar n_\partial}$$

and here $\bar{\mathbf{M}}_0$ is the mass matrix corresponding to the interior of Ω. Let $k_0 = 1, \dots, s$ and let us define the partitioned block matrix

$$\mathbf{M} := \left(\begin{array}{cccc|cccc} \bar{\mathbf{M}}_0 & \mathbf{0} & \dots & \mathbf{0} & \bar{\mathbf{M}}_\partial & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{M}}_0 & \dots & \mathbf{0} & \mathbf{0} & \bar{\mathbf{M}}_\partial & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \bar{\mathbf{M}}_0 & \mathbf{0} & \mathbf{0} & \dots & \bar{\mathbf{M}}_\partial \end{array}\right) \in \mathbf{R}^{N_0 \times N}, \tag{24}$$

or briefly

$$\mathbf{M} := \Big[\mathit{blockdiag}_s(\bar{\mathbf{M}}_0, \bar{\mathbf{M}}_0, \dots, \bar{\mathbf{M}}_0)\ \Big|\ \mathit{blockdiag}_s(\bar{\mathbf{M}}_\partial, \bar{\mathbf{M}}_\partial, \dots, \bar{\mathbf{M}}_\partial)\Big] \in \mathbf{R}^{N_0 \times N}. \tag{25}$$

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

$$\mathbf{M}\frac{\mathrm{d}\mathbf{u}^h}{\mathrm{d}t} + \mathbf{G}(\mathbf{u}^h(t)) = \mathbf{f}(t), \tag{26}$$

$$\mathbf{u}^h(0) = \mathbf{u}_0^h, \tag{27}$$

where using the form of i as in (22),

$$\mathbf{G}(\mathbf{u}^h(t)) = [G(\mathbf{u}^h(t))_i]_{i=1,\dots,N_0},$$

$$G(\mathbf{u}^h(t))_i = \int\limits_\Omega \Big(a_{k_0}(x,t,\nabla u)\nabla u_{k_0} \cdot \nabla\varphi_p + (\mathbf{w}_{k_0}(x,t) \cdot \nabla u_{k_0})\varphi_p + q_{k_0}(x,t,u)\varphi_p \Big)\, dx,$$

$$\mathbf{f}(t) = [f_i(t)]_{i=1,\dots,N_0}, \quad f_i(t) = \int_\Omega f_{k_0}(x,t)\varphi_p(x)\, dx + \int_\Gamma \gamma_{k_0}(x,t)\varphi_p(x)\, d\sigma(x),$$

and finally, $\mathbf{u}_0^h$ is defined by setting $t = 0$ in (21) and using that $u_p^{(k)}(0) = u_k^{(0)}(B_p)$ for $k = 1, \dots, s$ and $p = 1, \dots, \bar{n}_0$.

The solution $\mathbf{u}^h = \mathbf{u}^h(t)$ of problem (26)–(27) is called the semidiscrete solution. Here the coefficients $g_p^{(k)}(t)$ are given, hence (26) can be reduced to a system where $\mathbf{M}$ is replaced by the nonsingular square matrix $\mathbf{M}_0 := \mathit{blockdiag}_s(\bar{\mathbf{M}}_0, \bar{\mathbf{M}}_0, \dots, \bar{\mathbf{M}}_0)$ only. Then existence and uniqueness for (26)–(27) 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 Δt 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 $n = 0, 1, 2, \dots, n_T$, where $n_T\Delta t = T$), respectively.

To discretize (26) in time, we apply the implicit method. We then obtain a system of nonlinear algebraic equations of the form

$$\mathbf{M}\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + \mathbf{G}(\mathbf{u}^{n+1}) = \mathbf{f}^{n+1}, \tag{28}$$

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

$$\mathbf{M}\mathbf{u}^{n+1} + \Delta t\mathbf{G}(\mathbf{u}^{n+1}) = \mathbf{M}\mathbf{u}^n + \Delta t\ \mathbf{f}^{n+1} \tag{29}$$

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

$$\mathbf{P}(\mathbf{u}^{n+1}) := \mathbf{M}\mathbf{u}^{n+1} + \Delta t\mathbf{G}(\mathbf{u}^{n+1}). \tag{30}$$

Then, the iteration procedure (29) can be also written as

$$\mathbf{P}(\mathbf{u}^{n+1}) = \mathbf{M}\mathbf{u}^n + \Delta t\ \mathbf{f}^{n+1}. \tag{31}$$

Finding $\mathbf{u}^{n+1}$ in (31) requires the solution of a nonlinear algebraic system. Similarly as mentioned before, (31) can be reduced to a system with the first N_0 coefficients, i.e. $\mathbf{M}$ is replaced by the nonsingular square matrix $\mathbf{M}_0 := \mathit{blockdiag}_s(\bar{\mathbf{M}}_0, \bar{\mathbf{M}}_0, \dots, \bar{\mathbf{M}}_0)$ only, since the other coefficients of $\mathbf{u}^{n+1}$ are given from the $g_p^{(k)}(t)$. Analogously, $\mathbf{P}$ is replaced by $\mathbf{P}_0$. The block mass matrix $\mathbf{M}_0$ is positive definite, and it follows from Assumptions 2.1 that $\mathbf{u} \mapsto \mathbf{G}(\mathbf{u})$ has positive semidefinite derivatives. hence by the definition in (30), the function $\mathbf{u} \mapsto \mathbf{P}_0(\mathbf{u})$ has regular derivatives. This ensures the unique solvability of (31) and, under standard local Lipschitz conditions on the coefficients, also the convergence of the damped Newton iteration, see e.g. [6].

4 Discrete nonnegativity for the nonlinear system

4.1 Reformulation of the problem

First we rewrite problem (9) to a problem with nonlinear coefficients. Let us define, for any $k,l = 1,\dots,s,\;\; x \in \Omega$ resp. $\Gamma,\;\; t > 0,\, \xi \in \mathbf{R}$,

$$r_{kl}(x,t,\xi) := \int_0^1 \frac{\partial q_k}{\partial \xi_l}(x,t,\alpha\xi)\,d\alpha, \qquad \hat{f}_k(x,t) := f_k(x,t) - q_k(x,t,0). \tag{32}$$

Then the Newton-Leibniz formula yields for all x,t,ξ that

$$q_k(x,t,\xi) - q_k(x,t,0) = \sum_{l=1}^{s} r_{kl}(x,t,\xi)\,\xi_l. \tag{33}$$

Subtracting $q_k(x,t,0)$ from (1), we thus obtain that problem (9) is equivalent to

$$\int\limits_\Omega \sum_{k=1}^{s} \frac{\partial u_k}{\partial t} v_k\,dx + B(u;u,v) = \langle \psi, v\rangle \qquad (\forall v \in H^1_D(\Omega)^s, \quad t \in (0,T)), \tag{34}$$

where

$$B(y;u,v) := \int\limits_\Omega \sum_{k=1}^{s} \Big(a_k(x,t,\nabla y)\nabla u_k \cdot \nabla v_k + (\mathbf{w}_k(x,t)\cdot\nabla u_k)v_k \tag{35}$$

$$+ \sum_{k,l=1}^{s} r_{kl}(x,t,y)\,u_l v_k \Big)\,dx$$

and

$$\langle \psi, v\rangle := \int\limits_\Omega \sum_{k=1}^{s} \hat{f}_k v_k\,dx + \int\limits_\Gamma \sum_{k=1}^{s} \gamma_k v_k\,d\sigma.$$

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

$$u^h(x,0) = u^{(0),h}(x) \quad (x \in \Omega), \qquad u^h(.,t) - g^h(.,t) \in V_h^0 \quad (t \in (0,T))$$

and

$$\int\limits_\Omega \sum_{k=1}^{s} \frac{\partial u_k^h}{\partial t} v_k^h\,dx + B(u_h;u_h,v^h) = \langle \psi, v^h\rangle \qquad (\forall v^h \in V_h^0, \quad t \in (0,T)).$$

Proceeding as in (20)–(26), the Cauchy problem for the system of ordinary differential equations (26) takes the following form:

$$\mathbf{M}\frac{\mathrm{d}\mathbf{u}^h}{\mathrm{d}t} + \mathbf{K}(\mathbf{u}^h)\mathbf{u}^h = \hat{\mathbf{f}}, \tag{36}$$

$$\mathbf{u}^h(0) = \mathbf{u}^h_0, \tag{37}$$

where $\mathbf{M}$ is given in (24),

$$\mathbf{K}(\mathbf{u}^h) = \left[K(\mathbf{u}^h)_{ij}\right]_{N_0\times N}, \quad K(\mathbf{u}^h)_{ij} := B(u_h; \phi_j, \phi_i), \tag{38}$$

$$\hat{\mathbf{f}}(t) = [\hat{f}_i(t)]_{i=1,\dots,N_0}, \quad \hat{f}_i(t) = \int_\Omega \hat{f}_{k_0}(x,t)\varphi_p(x)\,dx + \int_\Gamma \gamma_{k_0}(x,t)\varphi_p(x)\,d\sigma(x). \tag{39}$$

The full discretization reads as

$$\mathbf{M}\mathbf{u}^{n+1} + \Delta t\mathbf{K}(\mathbf{u}^{n+1})\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n + \Delta t\,\hat{\mathbf{f}}^{n+1}. \tag{40}$$

Since we have set $\mathbf{G}(\mathbf{u}^h) = \mathbf{K}(\mathbf{u}^h)\mathbf{u}^h$ in (26), the expression (30) becomes

$$\mathbf{P}(\mathbf{u}^{n+1}) = \left(\mathbf{M} + \Delta t\mathbf{K}(\mathbf{u}^{n+1})\right)\mathbf{u}^{n+1}.$$

Then, letting

$$\mathbf{A}(\mathbf{u}^h) := \mathbf{M} + \Delta t\mathbf{K}(\mathbf{u}^h), \tag{41}$$

the iteration procedure (40) takes the form

$$\mathbf{A}(\mathbf{u}^{n+1})\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n + \Delta t\,\hat{\mathbf{f}}^{n+1}, \tag{42}$$

which is similar to (31), but now the nonlinear term arises through a coefficient matrix depending on $\mathbf{u}^{n+1}$.

4.2 The DMP: algebraic background

Some classical algebraic results, required in the sequel, are summarized first. We recall a basic definition in the study of DMP (cf. [15]):

Definition 4.1 A square $k \times k$ matrix $\mathbf{A} = (a_{ij})_{i,j=1}^k$ is called *irreducible* if for any $i \neq j$ there exists a sequence of nonzero entries $\{a_{i,i_1}, a_{i_1,i_2}, \dots, a_{i_s,j}\}$ of A, where $i, i_1, i_2, \dots, i_s, j$ are distinct indices.

Definition 4.2 Let $\mathbf{A}$ be an arbitrary $k \times k$ matrix. The *irreducible blocks* of $\mathbf{A}$ are the matrices $\mathbf{A}^{(l)}$ $(l = 1, \dots, q)$ defined as follows.

Let us call the indices $i, j \in \{1, \dots, k\}$ *connectible* if there exists a sequence of nonzero entries $\{a_{i,i_1}, a_{i_1,i_2}, \dots, a_{i_s,j}\}$ of $\mathbf{A}$, where $i, i_1, i_2, \dots, i_s, j \in \{1, \dots, k\}$ are distinct indices. Further, let us call the indices i, j mutually connectible if both i, j and j, i are connectible in the above sense. (Clearly, mutual connectibility is an equivalence relation.) Let $N_1, \dots, N_q$ be the equivalence classes, i.e. the maximal sets of mutually connectible indices. (Clearly, $\mathbf{A}$ is irreducible iff $q = 1$.) Letting $N_l = \{s_1^{(l)}, \dots, s_{k_l}^{(l)}\}$ for $l = 1, \dots, q$, we have $k_1 + \dots + k_q = k$. Then we define for all $l = 1, \dots, q$ the $k_l \times k_l$ matrix $\mathbf{A}^{(l)}$ by $\mathbf{A}^{(l)}_{pq} := a_{s_p^{(l)}, s_q^{(l)}} \qquad (p, q = 1, \dots, k_l)$.

Let us now consider a system of equations of order $(k+m)\times(k+m)$ with the following structure:

$$\bar{\mathbf{A}}\bar{\mathbf{c}} \equiv \begin{bmatrix} \mathbf{A} & \tilde{\mathbf{A}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \tilde{\mathbf{b}} \end{bmatrix} \equiv \bar{\mathbf{b}}, \tag{43}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix and $\mathbf{0}$ is the $m \times k$ zero matrix. Following [2], we may introduce

Definition 4.3 A $(k+m) \times (k+m)$ matrix $\bar{\mathbf{A}}$ with the structure (43) is said to be of *generalized nonnegative type* if the following properties hold:

(i) $a_{ii} > 0, \quad i = 1, ..., k,$

(ii) $a_{ij} \le 0, \quad i = 1, ..., k, \ j = 1, ..., k+m \quad (i \ne j),$

(iii) $\sum\limits_{j=1}^{k+m} a_{ij} \ge 0, \quad i = 1, ..., k,$

(iv) There exists an index $i_0 \in \{1, \dots, k\}$ for which $\sum\limits_{j=1}^{k} a_{i_0,j} > 0.$

(v) $\mathbf{A}$ is irreducible.

Many known results on various discrete maximum principles are based on the following theorem, considered as 'matrix maximum principle' [2, Th. 3]).

Theorem 4.1 *Let $\bar{\mathbf{A}}$ be a $(k+m) \times (k+m)$ matrix with the structure as in (43), and assume that $\bar{\mathbf{A}}$ is of generalized nonnegative type in the sense of Definition 4.3.*

If the vector $\bar{\mathbf{c}} = (c_1, ..., c_{k+m})^T \in \mathbf{R}^{k+m}$ (where $(\, . \,)^T$ denotes the transposed) is such that $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \le 0, \ i = 1, ..., k,$ then

$$\max_{i=1,...,k+m} c_i \le \max\{0, \max_{i=k+1,...,k+m} c_i\}. \tag{44}$$

However, the irreducibility of $\mathbf{A}$ is a technical condition which is sometimes difficult to check in applications, see e.g. [3]. We have shown in [9] that it can be omitted from the assumptions if (iv) is suitably strengthened. For convenient formulations, we will hence use the following

Definition 4.4 A $(k+m) \times (k+m)$ matrix $\bar{\mathbf{A}}$ with the structure as in (43) is said to be of *generalized nonnegative type with irreducible blocks* if properties (i)–(iii) of Definition 4.3 hold, further, property (iv) therein is replaced by the following stronger one:

(iv') For each irreducible component of $\mathbf{A}$ there exists an index $i_0 = i_0(l) \in N_l = \{s_1^{(l)}, \dots, s_{k_l}^{(l)}\}$ for which $\sum\limits_{j=1}^{k} a_{i_0,j} > 0.$

Theorem 4.2 [9]. *Let $\bar{\mathbf{A}}$ be a $(k+m) \times (k+m)$ matrix with the structure as in (43), and assume that $\bar{\mathbf{A}}$ is of generalized nonnegative type with irreducible blocks in the sense of Definition 4.4.*

If the vector $\bar{\mathbf{c}} = (c_1, ..., c_{k+m})^T \in \mathbf{R}^{k+m}$ is such that $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \le 0, \ i = 1, ..., k,$ then (44) holds.

By reversing signs, we obtain

Corollary 4.1 *Let $\bar{\mathbf{A}}$ be a $(k+m)\times(k+m)$ matrix with the structure as in (43), and assume that $\bar{\mathbf{A}}$ is of generalized nonnegative type with irreducible blocks in the sense of Definition 4.4.*

If the vector $\bar{\mathbf{c}} = (c_1, ..., c_{k+m})^T \in \mathbf{R}^{k+m}$ is such that $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \geq 0,\ i = 1, ..., k,$ then

$$\min_{i=1,...,k+m} c_i \geq \min\{0, \min_{i=k+1,...,k+m} c_i\}.$$

In particular, if $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \geq 0,\ i = 1, ..., k,$ and $c_i \geq 0,\ i = k+1, ..., k+m,$ then

$$c_i \geq 0, \qquad i = 1, ..., k+m. \tag{45}$$

4.3 The DMP: preliminaries on elliptic problems

We briefly summarize our result on a special elliptic PDE system, presented in [10]. Consider the following elliptic system, which is similar to a steady-state problem corresponding to (1)–(4):

$$\left.\begin{aligned} -\mathrm{div}\,\Big(b_k(x,\nabla u)\,\nabla u_k\Big) + \mathbf{b}_k(x)\cdot\nabla u_k + \sigma_k(x,u_1,...,u_s) \;&=\; \omega_k(x) \quad \text{a.e. in } \Omega,\\ b_k(x,\nabla u)\tfrac{\partial u_k}{\partial\nu} \;&=\; \beta_k(x) \quad \text{a.e. on } \Gamma_N,\\ u_k \;&=\; \alpha_k(x) \quad \text{a.e. on } \Gamma_D \end{aligned}\right\} \tag{46}$$

$(k = 1, \ldots, s)$.

Assumptions 4.3.

(i) $\Omega \subset \mathbf{R}^d$ is a bounded piecewise C^1 domain; Γ_D, Γ_N are disjoint open measurable subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ and $\Gamma_D \neq \emptyset$.

(ii) (Smoothness and growth.) For all $k, l = 1, \ldots, s$ we have $b_k \in (C^1 \cap L^\infty)(\Omega\times\mathbf{R}^d)$, $\mathbf{b}_k \in W^{1,\infty}(\Omega)$ and $\sigma_k \in C^1(\Omega\times\mathbf{R}^s)$. Further, let

$$2 \leq p < p^*, \quad \text{where } p^* := \tfrac{2d}{d-2} \ \text{ if } d \geq 3 \ \text{ and } \ p^* := +\infty \text{ if } d = 2; \tag{47}$$

then there exist constants $\beta_1, \beta_2 \geq 0$ such that

$$\left|\frac{\partial\sigma_k}{\partial\xi_l}(x,\xi)\right| \leq \beta_1 + \beta_2|\xi|^{p-2} \qquad (k, l = 1, \ldots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \tag{48}$$

(iii) (Ellipticity.) There exists $m > 0$ such that $b_k \geq m$ holds pointwise for all $k = 1, \ldots, s$. Further, the Jacobian matrices $\frac{\partial}{\partial\eta}\Big(b_k(x,\eta)\eta\Big)$ are uniformly spectrally bounded from both below and above.

(iv) (Coercivity.) We have $\ \mathrm{div}\,\mathbf{b}_k \leq 0\ $ on Ω and $\ \mathbf{b}_k\cdot\nu \geq 0\ $ on Γ_N $\quad (k = 1, \ldots, s)$.

(v) (Cooperativity.) We have

$$\frac{\partial\sigma_k}{\partial\xi_l}(x,\xi) \leq 0 \qquad (k, l = 1, \ldots, s,\ k \neq l;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \tag{49}$$

(vi) (Weak diagonal dominance for the Jacobians w.r.t. rows and columns.) We have

$$\sum_{l=1}^{s} \frac{\partial \sigma_k}{\partial \xi_l}(x,\xi) \geq 0, \qquad \sum_{l=1}^{s} \frac{\partial \sigma_l}{\partial \xi_k}(x,\xi) \geq 0 \qquad (k=1,\dots,s;\ x\in\Omega,\ \xi\in\mathbf{R}^s). \quad (50)$$

(vii) For all $k=1,\dots,s$ we have $\omega_k \in L^2(\Omega)$, $\beta_k \in L^2(\Gamma_N)$, $\alpha_k = \alpha^*_{k|\Gamma_D}$ with $\alpha^*_k \in H^1(\Omega)$.

We use the following notion of the quasi-regularity of the mesh.

Definition 4.5 Let $\Omega \subset \mathbf{R}^d$ and let us consider a family of FEM subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ constructed as in subsection 3.1. The corresponding mesh will be called *quasi-regular* w.r.t. problem (46) if

$$c_1 h^\gamma \leq meas(\operatorname{supp}\varphi_p) \leq c_2 h^d\,, \quad (51)$$

where the positive real number γ satisfies

$$d \leq \gamma < \min\Big\{ 2d - \frac{(d-2)p}{2},\ \frac{d(d+2)}{d+1}.\Big\} \quad (52)$$

The FEM discretization of system (46), constructed similarly as in subsection 3.1, leads to a system in the form

$$\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} \equiv \begin{bmatrix} \mathbf{A}(\bar{\mathbf{c}}) & \tilde{\mathbf{A}}(\bar{\mathbf{c}}) \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \tilde{\mathbf{g}} \end{bmatrix} . \quad (53)$$

Here the entries of $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ are

$$a_{ij}(\bar{\mathbf{c}}) = \int_\Omega \Big(\sum_{k=1}^{s} b_k(x,\nabla u^h)\,(\nabla\phi_j)_k \cdot (\nabla\phi_i)_k + \sum_{k,l=1}^{s} V_{kl}(x,u^h)\,(\phi_j)_l\,(\phi_i)_k \Big), \quad (54)$$

where

$$V_{kl}(x,u^h(x)) = \int_0^1 \frac{\partial \sigma_k}{\partial \xi_l}(x,tu^h(x))\,dt \qquad (k,l=1,\dots,s;\ x\in\Omega). \quad (55)$$

Now we can formulate the desired nonnegativity result for the stiffness matrix.

Theorem 4.3 [10]. *Let system (46) satisfy Assumptions 4.3. Let us consider a family of finite element subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ as constructed in section 3, such that the corresponding family of meshes is quasi-regular according to Definition 4.5, further, for any $p = 1,...,\bar{n}_0,\ t = 1,...,\bar{n}\ (p \neq t)$, if $meas(\operatorname{supp}\varphi_p \cap \operatorname{supp}\varphi_t) > 0$ then*

$$\nabla\varphi_t \cdot \nabla\varphi_p \leq 0 \ \ \text{on } \Omega \quad \text{and} \quad \int_\Omega \nabla\varphi_t \cdot \nabla\varphi_p \leq -K_0\, h^{\gamma-2},$$

where γ is from (52) and $K_0 > 0$ is a constant independent of p, t and h.

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ defined in (53)–(54) is of generalized nonnegative type with irreducible blocks in the sense of Definition 4.4.

Then Corollary 4.1 can be used, in particular, (45) yields that the coefficients of $\bar{\mathbf{c}}$ in (53) are nonnegative whenever the coordinates of $\mathbf{d}$ and $\tilde{\mathbf{g}}$ are nonnegative. Then the assumptions (12)–(13) on the basis functions imply (in a similar vein as [9, Th. 4.5]) that this nonnegativity also holds for the discrete solutions corresponding to these coefficients. Thus we can derive the corresponding discrete nonnegativity principle:

Corollary 4.2 *Let problem (46) satisfy Assumptions 4.3, and let its FEM discretization satisfy the corresponding conditions of Theorem 4.3. If $\omega_k(x) \geq \sigma_k(x,0)$, $\beta_k(x) \geq 0$ and $\alpha_k(x) \geq 0$ $(k = 1,\dots,s,\ \ x \in \Omega$ resp. $x \in \Gamma_D)$, then for sufficiently small h,*

$$u_k^h \geq 0 \quad on\ \Omega \qquad (k = 1,\dots,s). \tag{56}$$

4.4 The main result on the parabolic system

Now we are in the position to derive the discrete nonnegativity principle for the parabolic system (1).

Theorem 4.4 *(Discrete nonegativity principle.) Let system (1) satisfy Assumptions 2.1, further,*

$$f_k(x,t) \geq q_k(x,t,0), \qquad \gamma_k(x) \geq 0, \qquad g_k(x) \geq 0 \quad and \quad u_k^{(0)}(x) \geq 0$$

(for all $k = 1,\dots,s,\ \ x \in \Omega$ resp. $x \in \Gamma_D$ and $t \in [0,T]$). Let us consider the full discretization as constructed in section 3, such that the corresponding family of space FE meshes is quasi-regular according to Definition 4.5, further, for any $p = 1,...,\bar{n}_0,\ t = 1,...,\bar{n}$ $(p \neq t)$, if $meas(\operatorname{supp}\varphi_p \cap \operatorname{supp}\varphi_t) > 0$ then

$$\nabla\varphi_t \cdot \nabla\varphi_p \leq 0 \quad on\ \Omega \quad and \quad \int_\Omega \nabla\varphi_t \cdot \nabla\varphi_p \leq -K_0\, h^{\gamma-2}, \tag{57}$$

where γ is from (52) and $K_0 > 0$ is a constant independent of p, t and h.

Let

$$\Delta t \leq \frac{1}{\mu_0}, \tag{58}$$

where μ_0 is from (7), and let us extend the solutions $u(.,t_n)$ (on time levels $t_n := n\Delta t$) to the whole Q_T such that its values are between those on the neighbouring time levels, e.g. with the method of lines. Then, for sufficiently small h, the coordinates of the discrete solution satisfy

$$u_k^h \geq 0 \quad on\ Q_T \qquad (k = 1,\dots,s). \tag{59}$$

Proof. As seen in (42), the full discretization leads to the iteration

$$\mathbf{A}(\mathbf{u}^{n+1})\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n + \Delta t\, \hat{\mathbf{f}}^{n+1}, \tag{60}$$

where

$$\mathbf{A}(\mathbf{u}^h) := \mathbf{M} + \Delta t\, \mathbf{K}(\mathbf{u}^h), \qquad K(\mathbf{u}^h)_{ij} := B(u_h; \phi_j, \phi_i), \qquad \hat{\mathbf{f}}^{n+1} := [\hat{f}_i(t_{n+1})]_{i=1,\dots,N_0},$$

$\mathbf{M}$ and $\hat{f}_i$ are defined in (24) and (39), respectively, and $t_{n+1} := (n+1)\Delta t$. Let us rewrite (60) as

$$\frac{1}{\Delta t}\mathbf{A}(\mathbf{u}^{n+1})\mathbf{u}^{n+1} = \frac{1}{\Delta t}\mathbf{M}\mathbf{u}^n + \hat{\mathbf{f}}^{n+1}. \tag{61}$$

Here

$$\frac{1}{\Delta t}\mathbf{A}(\mathbf{u}^{n+1})_{ij} = \frac{1}{\Delta t}\mathbf{M}_{ij} + B(u_h^{n+1}; \phi_j, \phi_i),$$

where $\mathbf{M}$ is the system mass matrix from (24) and the form B was defined in (35). Therefore, by definition, $\frac{1}{\Delta t}\mathbf{A}(\mathbf{u}^{n+1})$ is the stiffness matrix corresponding to the s-tuple elliptic operator

$$\frac{1}{\Delta t}I + L,$$

where I is the identity on s-dimensional vectors and if $z = (z_1, \dots, z_s)$ then

$$Lz := \Big\{-\mathrm{div}\,\Big(a_k(x, t_{n+1}, \nabla u^{n+1})\nabla z_k\Big) + \mathbf{w}_k(x, t_{n+1})\cdot\nabla z_k + \sum_{l=1}^{s} r_{kl}(x, t_{n+1}, u^{n+1})\, z_l\Big\}_{k=1,\dots,s}.$$

Further, by definition, the vector $\bar{\mathbf{f}}^{n+1} := \frac{1}{\Delta t}\mathbf{M}\mathbf{u}^n + \hat{\mathbf{f}}^{n+1}$ comes from the discretization of the vector function

$$\{\frac{1}{\Delta t}u_k^n + \hat{f}^{n+1}\}_{k=1,\dots,s} \quad \text{with Neumann data} \quad \{\gamma_k^{n+1})\}_{k=1,\dots,s} \quad \text{on } \Gamma_N.$$

Therefore the algebraic system (61) is the FE discretization of the following nonlinear elliptic problem in V_h:

$$\left\{\begin{aligned} \tfrac{1}{\Delta t}u_k^{n+1} - \mathrm{div}\,\Big(a_k(x, t_{n+1}, \nabla u^{n+1})\,\nabla u_k^{n+1}\Big) + \mathbf{w}_k(x, t_{n+1})\cdot\nabla u_k^{n+1} + \sum_{l=1}^{s} r_{kl}(x, t_{n+1}, u^{n+1})\, u_l^{n+1}& \\ = \tfrac{1}{\Delta t}u_k^n + \hat{f}_k(x, t_{n+1}) \quad &\text{a.e. in } \Omega, \\ a_k(x, t_{n+1}, \nabla u^{n+1})\tfrac{\partial u_k^{n+1}}{\partial \nu} = \gamma_k(x, t_{n+1}) \quad &\text{a.e. on } \Gamma_N, \\ u_k = g_k(x, t_{n+1}) \quad &\text{a.e. on } \Gamma_D \end{aligned}\right. \tag{62}$$

$(k = 1, \dots, s)$. Using formulae (32)–(33), system (62) is equivalent to

$$\left\{\begin{aligned} -\mathrm{div}\,\Big(a_k(x, t_{n+1}, \nabla u^{n+1})\,\nabla u_k^{n+1}\Big) + \mathbf{w}_k(x, t_{n+1})\cdot\nabla u_k^{n+1} + \Big(q_k(x, t_{n+1}, u^{n+1}) + \tfrac{1}{\Delta t}u_k^{n+1}\Big)& \\ = \tfrac{1}{\Delta t}u_k^n + f_k(x, t_{n+1}) \quad &\text{a.e. in } \Omega, \\ a_k(x, t_{n+1}, \nabla u^{n+1})\tfrac{\partial u_k^{n+1}}{\partial \nu} = \gamma_k(x, t_{n+1}) \quad &\text{a.e. on } \Gamma_N, \\ u_k = g_k(x, t_{n+1}) \quad &\text{a.e. on } \Gamma_D \end{aligned}\right. \tag{63}$$

$(k = 1, \dots, s)$. This falls into the type (46) (for the unknown function u^{n+1}) if

$$b_k(x, \eta) := a_k(x, t_{n+1}, \eta), \qquad \mathbf{b}_k(x) := \mathbf{w}_k(x, t_{n+1}), \qquad \sigma_k(x, \xi) := q_k(x, t_{n+1}, \xi) + \frac{1}{\Delta t}\xi_k,$$

$$\omega_k(x) := \frac{1}{\Delta t} u_k^n(x) + f_k(x, t_{n+1}), \qquad \beta_k(x) := \gamma_k(x, t_{n+1}), \qquad \alpha_k(x) := g_k(x, t_{n+1}).$$

Now we verify that the above functions satisfy Assumptions 4.3, using that the original coefficients satisfy Assumptions 2.1. First, all domain and smoothness properties in Assumptions 4.3 follow from Assumptions 2.1. Then, assumption (ii) follows from (A4), and assumptions (iii)-(iv) follow from (A3). Since for all $k \neq l$

$$\frac{\partial \sigma_k}{\partial \xi_l}(x, \xi) = \frac{\partial q_k}{\partial \xi_l}(x, t_{n+1}, \xi),$$

assumption (v) follows from (A5). Finally, using (7) and the assumption $\Delta t \leq \frac{1}{\mu_0}$, the row sums satisfy

$$\sum_{l=1}^{s} \frac{\partial \sigma_k}{\partial \xi_l}(x, \xi) = \sum_{l=1}^{s} \frac{\partial q_k}{\partial \xi_l}(x, t_{n+1}, \xi) + \frac{1}{\Delta t} \geq -\mu_0 + \frac{1}{\Delta t} \geq 0,$$

and just similarly the column sums are also nonnegative, hence assumptions (vi) holds. That is, Assumptions 4.3 hold for system (63).

Our goal is to apply Corollary 4.2 to system (63). We have just seen that Assumptions 4.3 hold, and we have assumed in the theorem that the FEM discretization satisfies the corresponding conditions of Theorem 4.3. It remains to check that $\omega_k(x) \geq \sigma_k(x, 0)$, $\beta_k \geq 0$ and $\alpha_k \geq 0$ $(k = 1, \ldots, s)$. The last two are obvious from the assumptions $\gamma_k \geq 0$ and $g_k \geq 0$. Finally, the assumption $f_k(x, t) \geq q_k(x, t, 0)$ implies

$$\omega_k(x) \geq \frac{1}{\Delta t} u_k^n(x) + q_k(x, t, 0) \geq \frac{1}{\Delta t} u_k^n(x) + \sigma_k(x, 0),$$

which shows that if $u_k^n \geq 0$ then assumption $\omega_k(x) \geq \sigma_k(x, 0)$ also holds, i.e. all assumptions of Corollary 4.2 are satisfied for system (63) and thus $u_k^{n+1} \geq 0$. Altogether, under our assumptions, we have seen that the extra property $u_k^n \geq 0$ implies $u_k^{n+1} \geq 0$. Now we can carry out induction: since $u_k^0 \geq 0$, we obtain that $u_k^n \geq 0$ on Ω for all $n \in \mathbf{N}$. Since we extend the solutions $u(., t_n)$ (on time levels $t_n := n\Delta t$) to the whole Q_T such that its values are between those on the neighbouring time levels, we obtain that $u_k^h \geq 0$ on Q_T $(k = 1, \ldots, s)$. ■

Remark 4.1 *The validity of condition (57) can be guaranteed e.g. on acute simplicial meshes if linear finite elements are used in the space discretization [8]. The issue of generation of such meshes is considered in [1, 4, 14] and references therein.*

5 Examples

We give some examples of problems where the above DNP theorem yields new results. Let us recall here that the main conditions of the applied theorems are the relation $\Delta t \leq \frac{1}{\mu_0}$ for the time step and the "acuteness" property (57) for the space mesh. We will then derive nonnegativity for the discrete solution.

In all these examples, similarly as before, Ω stands for a bounded domain in $\mathbf{R}^d$ and $T > 0$ is a given number, and we denote $Q_T := (\Omega \setminus \Gamma_I)$.

As a main point, we will point out for each example that the lack of demanding diagonal dominance allows to cover much more general situations than before, without imposing any artifical extra conditions in the model.

5.1 A single equation: the Chaffee-Infante problem

Let us consider the so-called Chaffee-Infante equation (see e.g. [5]):

$$\frac{\partial u}{\partial t} - \Delta u + u^3 - u = 0 \quad \text{in} \;\; Q_T, \tag{64}$$

with the following boundary and initial conditions:

$$u(x,t) = g(x,t) \quad \text{for} \quad (x,t) \in \Gamma_D \times [0,T], \tag{65}$$

$$\tfrac{\partial u}{\partial \nu} = \gamma(x,t) \quad \text{for} \quad (x,t) \in \Gamma_N \times [0,T], \tag{66}$$

$$u(x,0) = u^{(0)}(x) \quad \text{for} \quad x \in \Omega, \tag{67}$$

respectively. We impose the corresponding additional items from Assumptions 2.1, which now reduce to the following simpler requirements:

Assumptions 5.1.

(A1) Ω is a bounded polytopic domain in $\mathbf{R}^d$; $\Gamma_N, \Gamma_D \subset \partial\Omega$ are are disjoint open measurable subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$.

(A2) $\gamma \in PC(\Gamma_N \times [0,T])$, $g \in PC(\Gamma_D \times [0,T])$ and $u^{(0)} \in PC(\Omega)$.

The other items from Assumptions 2.1 are trivially satisfied for the case $s = 1$: (A3) holds for $a(x,t,\eta) \equiv 1$ and $\mathbf{w} \equiv 0$, (A4) holds with $p = 4$, (A5) is a void condition for a single equation, and (A6) holds since $\lim_{\xi \to \pm\infty} (\xi^3 - \xi) = \pm\infty$, in fact, we have $\mu = 1$.

Then Theorem 4.4 implies that if

$$\gamma \geq 0, \qquad g \geq 0 \quad \text{and} \quad u^{(0)} \geq 0$$

on Ω resp. Γ_D, and the full discretization satisfies the conditions of Theorem 4.4 (including (57)–(58)), then

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

Note that the nonlinearity

$$q(x,\xi) := \xi^3 - \xi \tag{68}$$

is not monotone, hence this result is not covered by [7, 8].

5.2 Cross-catalytic reactions in chemistry

Certain reaction-diffusion processes in chemistry in a domain $\Omega \subset \mathbf{R}^d$, $d = 2$ or 3, are described by systems of the following form:

$$\frac{\partial u_k}{\partial t} - b_k \Delta u_k + \ P_k(x, u_1, \dots, u_s) = f_k(x,t) \quad \text{in} \quad Q_T, \tag{69}$$

with boundary and initial conditions

$$u_k(x,t) = g_k(x,t) \quad \text{for} \quad (x,t) \in \Gamma_D \times [0,T], \tag{70}$$

$$b_k \tfrac{\partial u_k}{\partial \nu} = 0 \quad \text{for} \quad (x,t) \in \Gamma_N \times [0,T], \tag{71}$$

$$u_k(x,0) = u_k^{(0)}(x) \quad \text{for} \quad x \in \Omega, \tag{72}$$

for all $k = 1, \dots, s$. Here, for all k, the quantity $u_k \geq 0$ describes the concentration of the kth species, and P_k is a polynomial which characterizes the rate of the reactions involving the k-th species, and satisfies $P_k(x,0) \equiv 0$ on Ω. The function $f_k \geq 0$ describes a source independent of concentrations.

We consider system (69)–(72) under the following conditions. The cooperativity means that such chemical models describe processes with cross-catalysis.

Assumptions 5.2.A.

(i) Ω is a bounded polytopic domain in $\mathbf{R}^d$, where $d = 2$ or 3, and $\Gamma_N, \Gamma_D \subset \partial\Omega$ are are disjoint open measurable subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$.

(ii) (Smoothness and growth.) For all $k, l = 1, \dots, s$, the functions P_k are polynomials of arbitrary degree if $d = 2$ or of degree at most 4 if $d = 3$, and we have $P_k(x,0) \equiv 0$ on Ω. Further, $f_k \in L^\infty(Q_T)$, $g_k \in L^\infty(\Gamma_D \times [0,T])$ and $u_k^{(0)} \in L^\infty(\Omega)$.

(iii) (Ellipticity for the principal space term.) $b_k > 0$ $(k = 1, \dots, s)$ are given numbers.

(iv) (Cooperativity.) We have

$$\frac{\partial P_k}{\partial \xi_l}(x, \xi) \leq 0 \qquad (k, l = 1, \dots, s, \ k \neq l; \ x \in \Omega, \ \xi \in \mathbf{R}^s). \tag{73}$$

By definition, the concentrations u_k are nonnegative, therefore a proper numerical model must produce such numerical solutions. Our topic is to give sufficient conditions to ensure this.

5.2.1 Monotone coupling

A DMP for such systems has been established in [8] for diagonally dominant nonlinearities. That is, the following additional condition was imposed together with Assumptions 5.2.A:

Assumption 5.2.B. (Diagonal dominance for the Jacobians w.r.t. rows and columns.)

$$\sum_{l=1}^{s} \frac{\partial P_k}{\partial \xi_l}(x,\xi) \geq 0, \qquad \sum_{l=1}^{s} \frac{\partial P_l}{\partial \xi_k}(x,\xi) \geq 0 \qquad (k=1,\dots,s;\ x\in\Omega,\ \xi\in\mathbf{R}^s). \tag{74}$$

If both Assumptions 5.2.A and Assumption 5.2.B hold, then it was proved in [8] that system (69) satisfies the DMP and, in particular, the discrete nonnegativity

$$u_k^h \geq 0 \quad \text{on } Q_T \qquad (k=1,\dots,s)$$

under similar mesh conditions as in Theorem 4.4.

However, Assumption 5.2.B is a very severe restriction. It implies self-inhibition for each chemical species:

$$\frac{\partial P_k}{\partial \xi_k}(x,\xi) \geq 0 \qquad (k=1,\dots,s)$$

and, moreover, the diagonal dominance (74) requires that this self-inhibition must be so strong that it compensates the total rate of cross-catalysis, i.e. although each cross-derivative is negative by (73), the sum of derivatives must be nonnegative as in (74). This set of properties is quite rarely valid for given chemical reactions.

5.2.2 Non-monotone coupling

In order to apply our new result, let us replace the above Assumption 5.2.B by the following:

Assumption 5.2.C. (Boundedness below for the Jacobians w.r.t. rows and columns.)

There exists a number $\mu_0 \geq 0$ such that

$$\sum_{l=1}^{s} \frac{\partial P_k}{\partial \xi_l}(x,\xi) \geq -\mu_0, \qquad \sum_{l=1}^{s} \frac{\partial P_l}{\partial \xi_k}(x,\xi) \geq -\mu_0 \qquad (k=1,\dots,s;\ x\in\Omega,\ \xi\in\mathbf{R}^s). \tag{75}$$

Together with Assumptions 5.2.A, our chemical system becomes a special case of system (1)–(4).

Let us study what restriction is imposed by demanding condition (75). It means that the directional derivatives $\frac{\partial P_k}{\partial \mathbf{e}}$ are bounded from below, where $\mathbf{e} := (1,\dots,1)$. As a first example, this is satisfied if, similarly to (68), the leading terms of the polynomials in each variable are odd and have positive coefficients. The oddity can always be ensured: since P_k are originally only defined for $\xi_k \geq 0$, we can redefine any factor ξ_k^j by the odd term $|\xi_k|^{j-1}\xi_k$ in order to achieve the above property.

An even more general statement is valid. Namely, for chemical reactions the rates normally include the products of the different concentrations, i.e. the polynomial in (69) is of the form

$$P_k(x, \xi_1, \ldots, \xi_s) = \sum_{\substack{j=1 \\ j \neq k}}^{s} a_{jk} \xi_j \xi_k + r_k(\xi_k) \tag{76}$$

(where r_k is a proper function describing the self-action of the kth species), and the assumed cooperativity means that each $a_{jk} \leq 0$ $(j \neq k)$. Now, firstly, since P_k are originally only defined for $\xi_k \geq 0$, we can redefine it as

$$P_k(x, \xi_1, \ldots, \xi_s) = \sum_{\substack{j=1 \\ j \neq k}}^{s} a_{jk} \xi_j |\xi_k| + r_k(\xi_k),$$

hence

$$\frac{\partial P_k}{\partial \xi_l}(x, \xi) = a_{lk} |\xi_k| \leq 0 \qquad (k \neq l)$$

as demanded in (73). Moreover, the possible concentrations $u_1, \ldots, u_s$ are limited by proper constants, depending on the capacities in the described reaction-diffusion process. Therefore the sums arising in (75) are continuous functions defined on compact subsets of $\Omega \times \mathbf{R}^s$, which are always bounded (in particular from below). To define the operator in (69) for all possible values, one simply has to define all P_k as zero outside a larger compact subset, which does not influence the boundedness from below. Altogether, under the assumed cooperativity $a_{jk} \leq 0$ $(j \neq k)$, Assumption 5.2.C for (76) does not require any further restriction on the model.

We can now use Theorem 4.4 to obtain the required nonnegativity: if $f_k \geq 0$, $g_k^h \geq 0$ and $u_k^{(0)} \geq 0$ for all $k = 1, \ldots, s$, and the full discretization satisfies the conditions of Theorem 4.4 (including (57)–(58)), then the coordinates of the discrete solution satisfy

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

It is now clear from the above discussion that this result has a much wider scope that the similar statement in [8] under diagonal dominance.

Other models arise as suitable modifications of the above system, also described in [8, sec. 6]. Chemical reactions can be sometimes localized on an interface, or a convection (advection) term can be present to describe a transport process. For both models the corresponding system in [8, sec. 6] can be modified such that diagonal dominance is replaced by (74). Thus, following the above line of discussion, we can weaken those results for much more general reactions,

5.3 Symbiotic population systems in biology

In population dynamics one sometimes encounters systems in the form

$$\begin{cases} \dfrac{\partial u_1}{\partial t} - b_1 \Delta u_1 = u_1\, M_1(u_1, u_2) \\[2ex] \dfrac{\partial u_2}{\partial t} - b_2 \Delta u_2 = u_2\, M_2(u_1, u_2), \end{cases} \tag{77}$$

where u_1, u_2 denote the amounts of two species distributed continuously in a plane region Ω, see e.g. [5]. The simple boundary and initial conditions

$$u_k = g_k \quad \text{on } \partial\Omega \times [0,T], \qquad u_k(.,0) = u_k^{(0)} \quad \text{on } \Omega \qquad (k = 1,2) \tag{78}$$

are imposed. Such a system can also describe a chemical reaction as in subsection 5.2 if the reaction rates are proportional to the quantity of the species. Here we will use the population terminology. If the species live in symbiosis, then

$$\partial_2 M_1 \geq 0 \qquad \text{and} \qquad \partial_1 M_2 \geq 0. \tag{79}$$

System (77) falls into the type (1) where

$$q_1(\xi_1, \xi_2) = -\xi_1\, M_1(\xi_1, \xi_2) \qquad \text{and} \qquad q_2(\xi_1, \xi_2) = -\xi_2\, M_2(\xi_1, \xi_2)\,, \tag{80}$$

and $f_1 \equiv f_2 \equiv 0$. Most of Assumptions 2.1 are trivially satisfied in a natural way, namely, let us impose

Assumptions 5.3. Ω is a bounded polygonal domain in $\mathbf{R}^2$ and $b_1, b_2 > 0$ are given numbers. Further, $g_1, g_2 \in C(\partial\Omega \times [0,T])$, $u_1^{(0)}, u_2^{(0)} \in C(\overline{\Omega})$, $M_1, M_2 \in C^1(\mathbf{R}^2)$ and they can grow at most with polynomial rate with ξ_1, ξ_2.

These assumptions imply that (A1)-(A4) of Assumptions 2.1 are satisfied. Now let us examine the remaining conditions. We briefly compare the previous result of [8] with the new result of this paper, in a similar vein as in the previous subsection 5.2.

5.3.1 Monotone coupling

In [8] we had to impose diagonal dominance, which led to very strong restrictions on the growth of M_1, M_2. It was shown that they must satisfy

$$\partial_i \Big(\xi_i \; M_i(\xi_1, \xi_2) \Big) \leq -\xi_j \; \partial_k M_j(\xi_1, \xi_2) \qquad (j \neq k).$$

A realistic example was studied, using functions in the form

$$q_i(\xi_1, \xi_2) = G_i \xi_i - \xi_i \xi_j\, h_i(\xi_1, \xi_2), \quad \text{then} \quad M_i(\xi_1, \xi_2) = -G_i + \xi_j\, h_i(\xi_1, \xi_2)$$

$(i = 1, 2,\ i \neq j)$, where $G_i > 0$ is the birth-death rate and h_i is a factor for the coexistence of the species (for instance, some Lotka-Volterra type systems can fall into this type). In this case one must assume that the rates h_i are small for large populations, in particular, that $|\partial_k h_i(\xi_1, \xi_2)| \leq \frac{c_1}{1+\xi_1^2+\xi_2^2}$. Moreover, c_1 must be so small that $c_1(1 + 2\sqrt{2}) \leq \min(G_1, G_2)$, in order to provide diagonal dominance. Clearly, these are very strong restrictions and allow only a very small deviation from the trivial linear uncoupled case.

5.3.2 Non-monotone coupling

In order to apply our new result, we show that (A5)-(A6) of Assumptions 2.1 can be satisfied in a very general case compared to the above, essentially with no extra restriction. This can be done in a similar vein as in the previous subsection 5.2. First we observe that by definition the model gives natural bounds on u_1, u_2: the amounts of the species are positive, and cannot exceed a limit determined by the capacity of the area. Hence M_1, M_2 are only considered on a compact set $D \subset (\mathbf{R}^+)^2$. Thus, in order to define $M_1, M_2 \in C^1(\mathbf{R}^2)$ in the differential operator, we can extend them from D in an alternative way such that they equal zero outside a larger compact set $\widetilde{D} \subset (\mathbf{R}^+)^2$.

Then the cooperativity (A5) follows from (79), since the latter yields (6) for $\xi_k \geq 0$, and the derivative vanishes with M_k for $\xi_k \leq 0$ (i.e. outside $\widetilde{D}$). Further, condition (A6) follows similarly: since $M_1, M_2 \in C^1(\widetilde{D})$, the l.h.s. of (7) has a minimum on $\widetilde{D}$, and thus it has the same lower bound on $\mathbf{R}^2$ since it vanishes outside $\widetilde{D}$.

Altogether, all Assumptions 2.1 are satisfied. Now we can use Theorem 4.4 to obtain the required nonnegativity for the numerically computed populations, using that (by the definition of the model) $g_k^h \geq 0$ and $u_k^{(0)} \geq 0$ for $k = 1, 2$. If the full discretization satisfies the conditions of Theorem 4.4 (including (57)–(58)), then

$$u_1^h,\, u_2^h \geq 0 \quad \text{on } Q_T.$$

To sum up, in the corresponding result of [8] we required very strong growths restrictions to ensure diagonal dominance instead of (74), whereas now we did not impose any artifical extra condition in the model.

Acknowledgements: The first and the second authors are supported by the Hungarian Research Grant OTKA No.K 112157. The third author acknowledges the Grant MTM2011-24766 of MICINN (Spain).

References

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

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

[3] A. Draganescu, T. F. Dupont, L. R. Scott, *Failure of the Discrete Maximum Principle for an Elliptic Finite Element Problem*, Math. Comp. 74 (2005), pp. 1–23.

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

[5] 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.

[6] 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.

[7] I. Faragó, J. Karátson, S. Korotov, *Discrete maximum principles for the FEM solution of some nonlinear parabolic problems*, Electr. Trans. Numer. Anal. 36 (2010), pp. 149-167.

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

[9] 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.

[10] J. Karátson, S. Korotov, Discrete maximum principles for FEM solutions of nonlinear elliptic systems, in: *Computational Mathematics: Theory, Methods and Applications*, ed. Peter G. Chareton, Computational Mathematics and Analysis Series, NOVA Science Publishers, New York, 2010; pp. 213-260.

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

[12] C. V. Pao, *Numerical analysis of coupled systems of nonlinear parabolic equations*, SIAM J. Numer. Anal. 36 (1999), no. 2, 393–416

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

[14] E. VanderZee, A. N. Hirani, V. Zharnitsky, D. Guoy, *A dihedral acute triangulation of the cube*, Comput. Geom. 43 (2010), 445–452.

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