Computers and Mathematics with Applications 66 (2013) 356–375

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Nonnegativity of exact and numerical solutions of some chemotactic models[☆]

Patrick De Leenheer[a], Jay Gopalakrishnan[b,*], Erica Zuhr[c]

[a] University of Florida, PO Box 118105, Gainesville, FL 32611–8105, United States
[b] Portland State University, PO Box 751, Portland, OR 97207–0751, United States
[c] High Point University, 833 Montlieu Avenue, High Point, NC, 27262, United States

ARTICLE INFO

Article history:
Received 30 October 2012
Received in revised form 11 April 2013
Accepted 20 May 2013

Keywords:
Keller–Segel
Pattern formation
Monotone
Numerical methods
Spurious
Blow up

ABSTRACT

We investigate nonnegativity of exact and numerical solutions to a generalized Keller–Segel model. This model includes the so-called “minimal” Keller–Segel model, but can cover more general chemistry. We use maximum principles and invariant sets to prove that all components of the solution of the generalized model are nonnegative. We then derive numerical methods, using finite element techniques, for the generalized Keller–Segel model. Adapting the ideas in our proof of nonnegativity of exact solutions to the discrete setting, we are able to show nonnegativity of discrete solutions from the numerical methods under certain standard assumptions. One of the numerical methods is then applied to the minimal Keller–Segel model. Recalling known results on the qualitative behavior of this model, we are able to choose parameters that yield convergence to a nonhomogeneous stationary solution. While proceeding to exhibit these stationary patterns, we also demonstrate how naive choices of numerical methods can give physically unrealistic solutions, thereby justifying the need to study positivity preserving methods.

1. Introduction

Negative approximations of intrinsically nonnegative quantities, such as density, are erroneous. They reduce confidence in simulation techniques, even when the negative values are close to zero. They often generate instabilities in nonlinear iterations thwarting convergence to a solution. Accordingly, nonnegativity of simulated solutions has become the first sanity check for any computer simulation of biological cell densities or chemical concentrations. However, not all nonlinear systems of partial differential equations come with a guarantee that their exact solutions are nonnegative. Their numerical discretization introduces a further layer of difficulty before one can certify that the simulated solution will be nonnegative. In this work, we examine these difficulties in the context of a specific class of nonlinear systems, which extend the influential chemotactic model proposed by Patlak [1] and later studied by Keller and Segel [2]. We provide a numerical approximation technique and prove that both the exact and the numerical solutions of the model are nonnegative.

We begin by describing a generalization of the Keller–Segel model that we shall focus on in later sections. It involves a species of density u occupying a domain $\Omega \subseteq \mathbb{R}^n$ and N reacting chemicals of concentrations v_i, $i = 1, 2, \ldots, N$, represented as components of a vector function $\vec{v}$. We are interested in the situation where one of these chemicals is a chemoattractant for

[☆] This work was partially supported by the NSF under grant DMS-1211635 and the Flemish Academic Centre for Science and the Arts (VLAC).

[*] Corresponding author. Tel.: +1 503 725 2205.
E-mail addresses: deleenhe@ufl.edu (P. De Leenheer), gjay@pdx.edu (J. Gopalakrishnan), ezuhr@highpoint.edu (E. Zuhr).
URLs: http://www.math.ufl.edu/~deleenhe/ (P. De Leenheer), http://web.pdx.edu/~gjay/ (J. Gopalakrishnan).

0898-1221/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.camwa.2013.05.014

the species. This situation is modeled by the following system of equations, taken from [3], which we refer to as a *generalized Keller–Segel system* throughout this paper:

$$\partial_t u = \nabla \cdot (D\nabla u - \chi u \nabla v_N) \quad x \in \Omega,\ t > 0, \tag{1a}$$

$$\partial_t \vec{v} = \tilde{D}\Delta\vec{v} + \vec{\alpha}u + \vec{g}(\vec{v}) \quad x \in \Omega,\ t > 0, \tag{1b}$$

$$u(x,0) = u_0(x), \qquad \vec{v}(x,0) = \vec{v}_0(x) \quad x \in \Omega, \tag{1c}$$

$$\frac{\partial u(x,t)}{\partial n} = \frac{\partial v_i(x,t)}{\partial n} = 0 \quad x \in \partial\Omega,\ t \geq 0,\ 1 \leq i \leq N. \tag{1d}$$

The rate of change of $\vec{v}$ is determined by a chemical reaction network, represented by the (nonlinear) function $\vec{g}(\vec{v})$. The parameter χ describes the chemotactic sensitivity. The term $\vec{\alpha}u$, with a constant vector $\vec{\alpha} \in \mathbb{R}^N$ satisfying $\vec{\alpha} \geq 0$, indicates that some or all of the chemicals can be produced by the species. (Here and throughout, the notation $\vec{w} \geq 0$ signifies that each component of $\vec{w}$ is nonnegative.) Above, we have augmented the differential equations with no-flux boundary conditions and initial conditions. We assume that the initial data $u_0 \geq 0$ and $\vec{v}_0 \geq 0$ are nontrivial functions on Ω. Additionally, $\vec{n}$ generically denotes the unit outward normal on the boundary of any domain under consideration and $\partial/\partial n = \vec{n} \cdot \vec{\nabla}$. (E.g., in (1), the domain under consideration is Ω and $\vec{n}$ is the outward unit normal on the boundary $\partial\Omega$.) For now, we assume that $\partial\Omega$ is Lipschitz so that $\vec{n}$ is defined a.e. on $\partial\Omega$, but we will place further assumptions on Ω in later sections for theoretical reasons. Throughout, we also use the abbreviated notations ∂_t and ∂_i for $\partial/\partial t$ and $\partial/\partial x_i$, respectively. We note that other generalizations of the Keller–Segel system have recently been proposed in [4,5], but most of their analysis is limited to a particular number of chemicals, and moreover their focus is not on the nonnegativity questions we intend to study here.

Let us now consider a few examples that fit the generalized Keller–Segel model (1).

Example 1.1 (*Minimal Keller–Segel Model*)**.** This very well-known nonlinear system of two equations is obtained by choosing, in (1),

$$N = 1, \qquad \vec{\alpha} = \left[\alpha_1\right], \qquad \vec{g}(\vec{v}) = \left[-\gamma v_1\right],$$

with $\alpha_1 > 0$ and $\gamma > 0$. In this setting, u represents the density of the amoeba *Dictyostelium discoideum* and v the density of the chemoattractant cyclic adenosine monophosphate (cAMP). Keller and Segel proposed this model to understand aggregation by chemotaxis in [2]. The extensive reviews in [6,7] puts this model in perspective and points to many known results on the behavior of its solutions.

Example 1.2 (*Full Keller–Segel Model with Decay*)**.** The minimal Keller–Segel model from the previous example is a simplified version of a four-equation model, also originally derived in [2]. As in the case of the minimal model we let u denote the density of the amoeba, but the density of the cAMP is now denoted by v_3. Additionally, an enzyme which degrades the cAMP, and is also emitted by the amoeba, enters the model. We denote the density of this enzyme by v_1. As explained in [2] or [6], the cAMP and the enzyme undergo a reversible reaction to form a complex, with density v_2. (This complex may then degrade into the enzyme plus a degraded product which is not typically considered in this model.) The reactions just described can be briefly written as

$$v_1 + v_3 \longleftrightarrow v_2 \longrightarrow v_2 + \text{ degraded product.} \tag{2}$$

Moreover, we assume that the enzyme decays at some positive rate γ_1, as justified biologically in [8]. Assuming a linear (in u) chemotactic sensitivity function, as in the minimal model, as well as constant production rates, we can now express the four equation model in the setting of (1). If the forwards, backwards and decay reaction rate constants from (2) are given by k_i for $1 \leq i \leq 3$ respectively, using the laws of mass action, the reaction network is described by

$$\vec{g}(\vec{v}) = \begin{bmatrix} -k_1 v_1 v_3 + (k_2 + k_3)v_2 - \gamma_1 v_1 \\ -(k_2 + k_3)v_2 + k_1 v_1 v_3 \\ -k_1 v_1 v_3 + k_2 v_2 \end{bmatrix}.$$

Using this $\vec{g}$ in (1), and setting chemical production rates by

$$\alpha = \begin{bmatrix} \alpha_1 \\ 0 \\ \alpha_3 \end{bmatrix},$$

with $\alpha_1 > 0$ and $\alpha_3 > 0$, we obtain the "full Keller–Segel model". Except for the decay term, this is the same as the four-equation model of [2].

Example 1.3 (*Dimerization*)**.** The chemistry of dimerization involves two molecules of a chemical v_1 combining reversibly to form another chemical v_2. This reaction can be written as

$$2v_1 \longleftrightarrow v_2$$

with the forwards and backwards reaction rate constants given by $k_1 > 0$ and $k_2 \geq 0$ respectively. If we also assume that the chemicals decay with rate constants $\gamma_1 \geq 0$ and $\gamma_2 \geq 0$, then the laws of mass action yield

$$\vec{g}(\vec{v}) = \begin{bmatrix} -2k_1 v_1^2 + 2k_2 v_2 - \gamma_1 v_1 \\ k_1 v_1^2 - k_2 v_2 - \gamma_2 v_2 \end{bmatrix} \tag{3}$$

to describe the reaction kinetics of the network. Letting

$$\alpha = \begin{bmatrix} \alpha_1 \\ 0 \end{bmatrix},$$

the system (1) describes a scenario in which an organism produces a chemical, which undergoes dimerization to form another chemical, which is in turn a chemoattractant for the organism.

In the next section, we show that under a reasonable assumption on $\vec{g}$ (satisfied for all the above examples), solutions of the generalized Keller–Segel system (1) have nonnegative components. Nonnegativity of solutions of the minimal model of Example 1.1 seems to be widely recognized among researchers to follow from the maximum principle. However, we have found it difficult to locate a reference which details the application of the appropriate maximum principle to be used. While the result seems to be known for the minimal model, nonnegativity of solutions of the full Keller–Segel model and the generalized model (1) was not known previously. Accordingly, in Theorem 2.2, we prove a nonnegativity result for the general model. We include all the essential details (and include an Appendix collecting the more standard results we use in the proof). The arguments we present are elementary and self-contained. These arguments then serve as our motivation in the design and nonnegativity analyses of a numerical method we present in Section 3.

There are a few other factors driving the design of our numerical method. We aim for the simplest positivity preserving method that can be obtained using the lowest order Lagrange finite element approximation space, for two reasons: first, finite element spaces allow us to compute on unstructured meshes and visualize solutions on complicated domains, unlike the finite difference models which are limited to simple domains like the square or to one space dimension [9]. Second, the lowest order finite element space (which has been around since [10]) is now available in almost any computational package for partial differential equations. We hope to lower the overhead of implementing a new numerical method for chemotaxis by designing a method using standard tools. Indeed, the method we propose is a combination of various tools, all of which by itself, are standard, but their combination, as we shall show, is particularly suited for the chemotaxis application. Apart from the Lagrange finite element space, the other standard ingredients we use include a discretization of [11], a mass lumping technique [12], and a simple backward differencing in time. The numerical treatment of nonlinear reaction term is motivated by [13].

The need for good numerical methods in understanding chemotaxis have not been underestimated by other researchers. A comparison of the performance and efficiency of some of the various (not necessarily positivity preserving) numerical schemes for chemotaxis is presented in [14]. Among the other numerous previous works are upwind finite element and central upwind finite volume methods [15–17], mixed finite elements [18] (that borrows techniques from similar semiconductor models [19–21]), discontinuous Galerkin (DG) methods [22,23], and a fractional step method [24]. The mixed and DG methods are particularly interesting since they are extendable to higher order (unlike our method, which is essentially first order). These methods also possess mass conservation properties. However, most of these methods use techniques that are not as standard as the simple Lagrange finite element, and have an additional overhead in solution cost, primarily due to the increased system size. Furthermore, while conservation is of critical importance in many applications where total output is important (e.g., oil flow), it is currently unclear if it is important in chemotaxis, especially if it requires extra expense.

The finite element models of [16,17] seem to be the closest to this paper. Considering the minimal model and assuming no time dependence in chemical concentration, [16] solves for the chemical concentration using a discrete Greens function, and uses it to obtain a nonlocal approximation of (1a). The approach is extended in [17] to the case where both quantities vary in time. Mass lumping and mesh angle conditions are used in [17] (very similar to the our treatment later) to provide a convergence theory. There, stability is gained (see [17, Eq.(4)]) by an upwinding approximation of [25] applied only to the last term of (1a). In contrast, we use the approach of [11] to approximate the entire flux term in (1a). Another finite element approach is considered in [26]. There, the focus is on maintaining positivity of solutions and conservation properties via a flux correction, resulting in a different method from what we propose in this paper. To indicate other points of departure from the existing literature, many previous numerical studies have focused on solutions of the minimal model that blow up [15,23,26]. We instead focus on obtaining a bounded stationary pattern. Also, several studies like [18,16] consider models even further simplified from the minimal model. None of the above mentioned numerical studies consider a system of more than two partial differential equations. In contrast, our aim is to develop techniques that apply to the more general system (1).

In the next section, we prove a nonnegativity result for the solutions of the nonlinear boundary value problem (1). In Section 3, we consider a finite element discretization of (1), motivating the derivation of matrices involved and our treatment of the nonlinear term. In Section 4, we prove that, under certain mesh angle conditions, the numerical solutions are nonnegative. We then apply the numerical method to visualize a stationary solution of the Keller–Segel model.

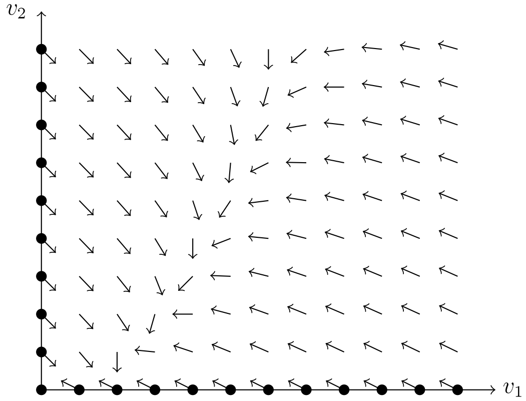


Fig. 1. The directions of the vector field $\vec{g}(\vec{v})$ of Example 1.3 – see (3) – point inward at the points marked on the axes. (The parameters in (3) were set to $k_1 = 0.1$, $k_2 = 0.2$, $\gamma_1 = \gamma_2 = 0$ for this illustration.)

2. Nonnegativity of exact solutions

In this section, we prove that solutions of (1) are nonnegative. Throughout this section we assume that Ω is a bounded connected set whose boundary $\partial\Omega$ is smooth and in particular, satisfies the interior ball condition. We also assume that the u_0 and $\vec{v}_0$ are smooth functions. Assuming that a smooth solution to (1) exists, we now proceed to determine the sign of the solution. Our analysis is under the following assumption on the reaction term $\vec{g}$.

Assumption 2.1. Assume that $\vec{g}$ is uniformly Lipschitz on compact subsets of $\mathbb{R}^N$ and that for any $\vec{v} \geq 0$, we have

$$g_j(\vec{v}\,) \geq 0 \quad \text{whenever } v_j = 0, \text{ for any } j = 1, 2, \ldots, N.$$

Loosely speaking, this assumption requires the vector field $\vec{g}$ to be "inward pointing" at the boundaries of the positive orthant in the $\vec{v}$-coordinate system (see Fig. 1). The following theorem is the main result of this section.

Theorem 2.2. *Suppose Assumption 2.1 holds. Assume that $(u, \vec{v})$ is a smooth solution to (1) with nontrivial $u_0 \geq 0$ and $\vec{v}_0 \geq 0$. Then $u > 0$ for all $x \in \bar{\Omega}$ and $t > 0$ and moreover, $\vec{v} \geq 0$ for all $x \in \bar{\Omega}$ and $t \geq 0$.*

The proof of this theorem appears at the end of this section. We will need a few standard, and some not so standard, maximum principles for parabolic problems. These results are collected in the Appendix in a form appropriate for use in the arguments of this section. In order to prove the theorem, we will need to develop a few intermediate lemmas below.

Before proceeding to these lemmas, let us first show that for a large class of practically important reaction networks, Assumption 2.1 holds. A general reaction network, with N chemicals v_i, is composed of many reactions of the type

$$v_{i_1} + v_{i_2} + \cdots + v_{i_k} \xrightarrow{\kappa} v_{j_1} + v_{j_2} + \cdots v_{j_l} \tag{4}$$

where i's and j's (are possibly repeated) indices contained in $\{1, 2, \ldots, N\}$. We consider any $\vec{g}(\vec{v})$ derived from the laws of mass action kinetics. Fix any index i_*, and consider a component reaction of the form (4), where i_* appears M_1 times on the left hand side of (4) and M_2 times on the right hand side of (4). When the kinetics of a reaction is determined by mass action, the rate at which the reaction occurs is directly proportional to the products of the concentrations of the reactants, and the constant of proportionality is called the rate constant, denoted by κ in (4). Thus, the i_*th component of $\vec{g}(\vec{v})$ satisfies

$$g_{i_*}(\vec{v}) = -M_1\kappa v_{i_1}v_{i_2}\cdots v_{i_k} + M_2\kappa v_{i_1}v_{i_2}\cdots v_{i_k} = (M_2 - M_1)\kappa v_{i_1}v_{i_2}\cdots v_{i_k}. \tag{5}$$

This gives the rate of production of v_{i_*} in the reaction network.

Next, we consider two cases: first, if $M_2 \geq M_1$, it is obvious from (5) that

$$g_{i_*}(\vec{v}) \geq 0 \quad \text{for all } \vec{v} \geq 0. \tag{6}$$

In the remaining case, if $M_2 < M_1$, then since both M_1 and M_2 are nonnegative integers, obviously $M_1 \geq 1$. In other words, v_{i_*} appears on the left hand side of (4) at least once, and hence appears as a factor in the product $v_{i_1}v_{i_2}\cdots v_{i_k}$. Therefore,

$$g_{i_*}(\vec{v}) \geq 0 \quad \text{whenever } v_{i_*} = 0, \tag{7}$$

for all $\vec{v} \geq 0$. Since i_* was an arbitrary index, we conclude that either (6) or (7) holds for every component of $\vec{g}$. Therefore, we have just established that Assumption 2.1 *holds for any $\vec{g}$ obtained from mass action kinetics.* (Note that any function derived by mass action, being a polynomial, obviously satisfies the local Lipschitz continuity required in Assumption 2.1.) Finally, observe that the functions $\vec{g}$ in Examples 1.2 and 1.3 were derived using the law of mass action, so the positivity result of Theorem 2.2 applies to those examples. That Assumption 2.1 is satisfied in the case of Example 1.1 is trivial to show.

Let us now develop the intermediate results required to prove Theorem 2.2. We begin with a result on a scalar equation and eventually proceed to generalize it to the system of equations defining $\vec{v}$. Although results like the next lemma seem to be known [27] in the community, we aim to give an elementary and self-contained presentation.

Lemma 2.3. *Suppose there is an $\varepsilon_0 > 0$ such that for all $0 < \varepsilon < \varepsilon_0$, the continuous function $f : \mathbb{R} \times \bar{\Omega} \times [0, T] \mapsto \mathbb{R}$ satisfies*

$$f(-\varepsilon, x, t) > 0, \quad (x, t) \in \bar{\Omega} \times [0, T]. \tag{8}$$

Let $w : \bar{\Omega} \times [0, T] \to \mathbb{R}$ be continuously differentiable with respect to t, twice continuously differentiable with respect to x, and satisfy

$$\partial_t w = \tilde{d}\Delta w + f(w, x, t), \quad x \in \bar{\Omega},\ t \in (0, T] \tag{9a}$$

$$w(x, t) \geq 0 \quad x \in \bar{\Omega},\ t = 0, \tag{9b}$$

for some number $\tilde{d} > 0$, together with the following boundary condition:

$$\text{Either } w(x, t) \geq 0, \quad x \in \partial\Omega,\ t \in (0, T], \tag{9c}$$

$$\text{or } \frac{\partial w}{\partial n} \geq 0, \quad x \in \partial\Omega,\ t \in (0, T]. \tag{9d}$$

Then $w(x, t) \geq 0$ for all $x \in \bar{\Omega}$ and $t \in [0, T]$.

Proof. If not, there is an $(\tilde{x}_0, \tilde{t}_0) \in \bar{\Omega} \times [0, T]$ such that $w(\tilde{x}_0, \tilde{t}_0) < 0$. Then $w(\tilde{x}_0, \tau)$ is negative at $\tau = \tilde{t}_0$ but nonnegative at $\tau = 0$ due to (9b). Hence, there are (small enough) values of $\varepsilon \in (0, \varepsilon_0)$ such that w attains the value of $-\varepsilon$ at one or more $\tau \in (0, \tilde{t}_0)$. Fix such an ε and let t_0 be the *first* time when w attains the value of $-\varepsilon$. Let x_0 be any point in $\bar{\Omega}$ such that $w(x_0, t_0) = -\varepsilon$. Then, as τ increases to t_0 and is sufficiently close to t_0, the function $w(x_0, \tau)$ cannot increase, so

$$\partial_t w(x_0, t_0) \leq 0. \tag{10}$$

Note also that

$$\min_{(x,t)\in\bar{\Omega}\times[0,t_0]} w(x, t) = w(x_0, t_0) = -\varepsilon \tag{11}$$

(because if w took a value lesser than $-\varepsilon$ in $\bar{\Omega} \times [0, t_0]$, then t_0 would not be the first time when w attains $-\varepsilon$).

The remainder of the proof is split in two parts. First, suppose (9c) holds. Then, $x_0 \notin \partial\Omega$. In view of (11), we therefore have

$$\Delta w(x_0, t_0) \geq 0. \tag{12}$$

Combining (10) and (12), we find that $0 \geq (\partial_t w - \tilde{d}\Delta w)|_{(x_0,t_0)} = f(-\varepsilon, x_0, t_0)$. This contradicts (8) and finishes the proof in the case of the boundary condition (9c).

To complete the proof for the case of the boundary condition (9d), first note that if x_0 is an interior point of Ω, then we obtain a contradiction using (10) and (12) as above, so we need only consider $x_0 \in \partial\Omega$. By (9a) and (8), the inequality $\partial_t w - \tilde{d}\Delta w = f > 0$ holds at (x_0, t_0), and so by continuity, it holds in $\bar{\Omega}_0 \times [t_1, t_0] \subseteq \Omega \times (0, t_0]$, where Ω_0 is a ball whose boundary contains x_0 (possible due to our assumption that the interior ball condition holds). By (11), we know that the minimum of w in $\bar{\Omega}_0 \times [t_1, t_0]$ is attained at (x_0, t_0). If this minimum is also attained at another point (x_0', t_0') in the same neighborhood, then x_0' is an interior point of Ω, so (10) and (12) finish the proof as before. Hence it only remains to consider the situation when $x_0 \in \partial\Omega_0$ is the sole point in $\bar{\Omega}_0 \times [t_1, t_0]$ where the minimum is attained. But in this situation, all conditions of Lemma A.1 are satisfied in a sufficiently small parabolic frustum contained in $\Omega_0 \times (t_1, t_0]$. Hence, (56) implies that $\partial w/\partial n < 0$ at x_0 which contradicts (9d). □

In order to strengthen this result to a more useful form for systems, we need a result on the existence and uniqueness of a PDE of the form (9). The needed results can be found in [28, Theorems 6 and 10 in Section 7.4]. To fit our application, we slightly modify and restate them in the theorem below. Recall that $f(w, x, t)$ is said to be "locally Hölder continuous" in (x, t) if there is a C and $0 < \alpha < 1$ such that

$$|f(w, x_1, t_1) - f(w, x_2, t_2)| \leq C|(x_1, t_1) - (x_2, t_2)|^\alpha$$

for all (x_1, t_1) and (x_2, t_2) in every closed bounded subset B of $\bar{\Omega}_T$.

Theorem 2.4. *Consider (9a) together with the initial and boundary condition*

$$w(x, t) = w_0(x, t), \quad (x, t) \in (\bar{\Omega} \times \{t = 0\}) \cup (\partial\Omega \times [0, T]) \tag{13}$$

for some smooth w_0. Assume that $f(w, x, t)$ is Lipschitz continuous in w uniformly with respect to bounded subsets of $\mathbb{R} \times \bar{\Omega} \times [0, T]$, and that $f(w, x, t)$ is locally Hölder continuous in (x, t). Suppose that the equation $\partial_t w_0 = \tilde{d}\Delta w_0 + f(w_0, x, 0)$ holds on $\partial\Omega$ for some number $\tilde{d} > 0$. Then there is a $0 < T_0 \leq T$ such that a unique solution to (9a), satisfying the initial and boundary condition (13), exists in Ω_{T_0}.

Let us now proceed to systems of partial differential equations. We say that $\vec{f} : \mathbb{R}^N \times \bar{\Omega} \times [0, T]$ is "locally Lipschitz continuous in $\vec{y}$, uniformly in (x, t)" if for every bounded subset $D \subset \mathbb{R}^N$, there is a constant $M_f > 0$ such that

$$\max_i |f_i(\vec{y}, x, t) - f_i(\vec{z}, x, t)| \leq M_f \max_i |y_i - z_i| \tag{14}$$

for all $\vec{y}, \vec{z} \in D$ and all $(x, t) \in \bar{\Omega} \times [0, T]$. Note that M_f does not depend on (x, t). Extensions of results like Lemma 2.3 for systems can be found in [29]. However, there is a long list of assumptions in [29], to be verified before one can conclude positivity. Some of those assumptions do not apply to (1), yet his techniques, rooted on the "positive invariance of the positive cone", appear to be powerful enough to extend to (1). Nonetheless, instead of adapting his technique, we have chosen to present another elementary proof, inspired by Weinberger [13], one of the original architects of such techniques. Furthermore, the details of the ensuing argument will serve as motivation to construct a discrete version of the same argument to prove that numerical solutions are also nonnegative (in Section 4).

Lemma 2.5. *Suppose that $\vec{f} : \mathbb{R}^N \times \bar{\Omega} \times [0, T]$ is locally Lipschitz continuous in $\vec{y}$, uniformly in (x, t), and locally Hölder continuous in (x, t). Assume that for all $(x, t) \in \bar{\Omega}_T$,*

$$f_i(y_1, y_2, \ldots, y_{i-1}, 0, y_{i+1}, \ldots, y_N, x, t) \geq 0, \quad i = 1, 2, \ldots, N \tag{15}$$

whenever $y_j \geq 0$ for all $j \neq i$. Let $\vec{w} : \mathbb{R}^N \times \bar{\Omega} \times [0, T] \to \mathbb{R}^N$ be a smooth solution of

$$\partial_t w_i = \tilde{D}_i \Delta w_i + f_i(\vec{w}, x, t), \quad x \in \bar{\Omega},\ t \in (0, T],\ i = 1, 2, \ldots, N, \tag{16a}$$

$$\vec{w}(x, t) \geq 0 \quad x \in \bar{\Omega},\ t = 0, \tag{16b}$$

together with the following boundary condition:

$$\text{Either } \vec{w}(x, t) \geq 0, \quad x \in \partial\Omega,\ t \in (0, T], \tag{16c}$$

$$\text{or } \quad \frac{\partial \vec{w}}{\partial n} \geq 0, \quad x \in \partial\Omega,\ t \in (0, T]. \tag{16d}$$

Then $\vec{w}(x, t) \geq 0$ for all $x \in \bar{\Omega}$ and $t \in [0, T]$.

Proof. The idea is to construct a new function F_i from f_i such that Lemma 2.3 can be applied. To this end, first let $a^+ = \max(a, 0)$. It can be easily verified, case by case, that for any two numbers a and b,

$$|a^+ - b^+| \leq |a - b|. \tag{17}$$

Next, define $F_i : \mathbb{R} \times \bar{\Omega} \times [0, T] \to \mathbb{R}$ by

$$F_i(v, x, t) = f_i(w_1(x, t), \ldots, w_{i-1}(x, t),\ v^+, w_{i+1}(x, t), \ldots, w_N(x, t), x, t) + v^+ - v. \tag{18}$$

By (17) and (14), F_i is locally Lipschitz continuous in v, uniformly in (x, t).

We now prove that, for an arbitrary $\varepsilon > 0$,

$$F_i(-\varepsilon, x, t) > 0, \quad \forall\ (x, t) \in K_i^{-\varepsilon_1} \tag{19}$$

for any $\varepsilon_1 < \varepsilon/M_f$. Here $K_i^\alpha = \{(x, t) \in \bar{\Omega} \times [0, T] : w_j(x, t) \geq \alpha \text{ for all } j \neq i\}$. Before proving this, note that although (15) immediately implies the inequality $F_i(-\varepsilon, x, t) \geq \varepsilon$, this inequality holds in general only for $(x, t) \in K_i^0$. To obtain a similar inequality in a larger set, we use (14). Whence, for any $\varepsilon_1 > 0$ and any $(x, t) \in K_i^{-\varepsilon_1}$,

$$f_i(\vec{0}, x, t) - f_i(w_1, \ldots, w_{i-1}, 0, w_{i+1}, \ldots, w_N, x, t) \leq M_f \varepsilon_1.$$

Since the first term above is nonnegative due to (15), this implies that

$$F_i(-\varepsilon, x, t) = f_i(w_1, \ldots, w_{i-1}, 0, w_{i+1}, \ldots, w_N, x, t) + \varepsilon \geq -M_f \varepsilon_1 + \varepsilon$$

and (19) follows. Accordingly, we fix $\varepsilon_1 < \varepsilon/M_f$ and proceed.

To prove the lemma by way of contradiction, suppose $\vec{w} \not\geq 0$. Then there exists some ε_2, sufficiently small, and chosen so that $0 < \varepsilon_2 \leq \varepsilon_1$, such that at least one of the components of $\vec{w}$ attains the value $-\varepsilon_2$. Let $t_1 > 0$ be the first time that any of the components of $\vec{w}$ attain the value $-\varepsilon_2$, and let i^* and $x_1 \in \bar{\Omega}$ be such that $w_{i^*}(x_1, t_1) = -\varepsilon_2$. Then (cf. (11))

$$\min_i \min_{(x,t) \in \bar{\Omega} \times [0, t_1]} w_i(x, t) = w_{i^*}(x_1, t_1) = -\varepsilon_2. \tag{20}$$

Clearly, this implies that

$$\bar{\Omega} \times [0, t_1] \subseteq K_{i^*}^{-\varepsilon_2} \subseteq K_{i^*}^{-\varepsilon_1}. \tag{21}$$

Now, let v_{i^*} be the solution to

$$\partial_t v_{i^*} = \tilde{D}_{i^*} \Delta v_{i^*} + F_{i^*}(v_{i^*}, x, t), \quad x \in \bar{\Omega},\ t > 0, \tag{22a}$$

$$v_{i^*} = w_{i^*}, \quad (x, t) \in (\bar{\Omega} \times \{t = 0\}) \cup (\partial\Omega \times \{t > 0\}). \tag{22b}$$

By Theorem 2.4 and the aforementioned continuity properties of each F_i, v_{i^*} exists in an interval $[0, t_2]$, where $t_2 > 0$ is the maximal time of existence of the solution. The remainder of the proof is split into two cases: $t_2 \geq t_1$ and $t_2 < t_1$.

Consider the first case $t_2 \geq t_1$. Because of (21), we find that the inequality in (19) holds for all $x \in \bar{\Omega}$ and all $0 \leq t \leq t_1$. So we can apply Lemma 2.3 to conclude that $v_{i^*} \geq 0$ on $[0, t_1)$, which in turn implies that

$$F_{i^*}(v_{i^*}, x, t) \equiv f_{i^*}(w_1, \dots, w_{i-1}, v_{i^*}, w_{i+1}, \dots, w_N, x, t).$$

But then, the (i^*th) equations in (16a)–(16c) show that w_{i^*} also solves (22) on $[0, t_1)$. By uniqueness (Theorem 2.4) we conclude that $w_{i^*} = v_{i^*} \geq 0$ on $[0, t_1)$, a contradiction to (20).

The other case, $t_2 < t_1$, also leads to a contradiction, as we now show. As above, we conclude that $w_{i^*} = v_{i^*} \geq 0$ on $[0, t_2)$. Now, due to the smoothness assumptions on w_{i^*} on $[0, T]$ (and noting that $t_2 < t_1 \leq T$) the solution v_{i^*} is smooth at $t = t_2$. But then we can extend the solution v_{i^*} to some interval $[0, t_3)$ with $t_3 > t_2$ by again invoking Theorem 2.4. This is a contradiction to the maximality of t_2 and finishes the proof. □

Finally, we are in a position to prove our main result.

Proof of Theorem 2.2. First, let us prove that $u > 0$. From (1a) and the product rule,

$$\partial_t u - D\Delta u + \chi \nabla u \cdot \nabla v_N + \chi u \Delta v_N = 0.$$

Therefore, by Lemma A.4 in the Appendix, applied with $Lu := -D\Delta u + \chi \nabla u \cdot \nabla v_N + \chi u \Delta v_N$, we obtain that $u > 0$ on Ω_T. (Note that the lowest order term $c = \chi \Delta v_N$, being a continuous function over a compact set, is bounded below.)

Next, we prove that $\vec{v} \geq 0$. Due to (1b), we may apply Lemma 2.5 with $\vec{w} = \vec{v}$ and

$$\vec{f} = \vec{\alpha} u + \vec{g}(\vec{v}).$$

Condition (15) is satisfied because whenever $v_i = 0$, we have $f_i = \alpha_i u + g_i(\vec{v}) \geq 0$ by virtue of the assumption on $\vec{g}$ and the already proved positivity of u. The remaining assumptions of the lemma are easily verified, so its conclusion $\vec{v} \geq 0$ holds on $\bar{\Omega}_T$. □

3. Numerical method

In this section, we describe two computational schemes for solving (1). Under certain conditions, the resulting numerical solutions are nonnegative, as proved in the succeeding section. To focus on the numerical approximation, from now on, we assume that Ω is a bounded connected polygon in $\mathbb{R}^2$, which is partitioned into triangles. The collection of these triangles (or "elements") is denoted by $\mathcal{T}_h$ and we assume that they satisfy the standard assumptions (see e.g., [12]) of a geometrically conforming finite element mesh.

3.1. Description of the method

The solution functions u and v_i are approximated in the lowest order Lagrange finite element space $S_h = \{w \in C(\Omega) : w|_K \text{ is linear on } K \text{ for every mesh triangle } K\}$. For any $\ell = 1, 2, \dots, P$, let $\phi_\ell \in S_h$ denote the function that equals one at the ℓth mesh vertex and equals 0 at all other mesh vertices. We expand the approximating functions, at time $t_n = kn$, in the basis $\{\phi_\ell\}$

$$u(x, t_n) \approx u_h^n = \sum_{\ell=1}^{P} U_\ell^n \phi_\ell, \qquad v_i(x, t_n) \approx v_{h,i}^n = \sum_{\ell=1}^{P} V_{i,\ell}^n \phi_\ell.$$

Given the vectors $\vec{U}^n$ and $\vec{V}_i^n$, whose ℓth components are U_ℓ^n and $V_{i,\ell}^n$, resp., we propose to compute the approximations at time $k(n+1)$ by solving the following system:

$$\frac{1}{k} M(\vec{U}^{n+1} - \vec{U}^n) = -DA(v_{h,N}^{n+1})\vec{U}^{n+1}, \tag{23a}$$

$$\frac{1}{k} M(\vec{V}_i^{n+1} - \vec{V}_i^n) = -\tilde{D}_i L \vec{V}_i^{n+1} + \alpha_i M \vec{U}^n + M\vec{G}_i^+(\vec{v}_h^{n+1}). \tag{23b}$$

The $P \times P$ matrices $A(\cdot)$, L, and M are defined below. The vector $\vec{G}_i^+(\vec{v}_h^n)$ is the vector whose ℓth component equals

$$g_{i,\ell}^{+,n} \equiv g_i\left((V_{1,\ell}^n)^+, (V_{2,\ell}^n)^+, \dots, (V_{N,\ell}^n)^+\right) + (V_{i,\ell}^n)^+ - V_{i,\ell}^n \tag{24}$$

and for any number a, as before, $a^+ = \max(a, 0)$. This construction is motivated by the adaption of Weinberger's technique described in the previous section — cf. (18). Note that the method (23) is an implicit method: Given $\vec{U}^n$ and $\vec{V}_i^n$, we first solve the (second) nonlinear equation (23b) for $\vec{V}_{i,\ell}^{n+1}$. Then we solve a $P \times P$ linear system for $\vec{U}^{n+1}$ given by (23a). Note that the existence of numerical solutions for (23b) is not as straightforward and is postponed for future studies. In the remainder, we take the approach of the previous section, whereby we study the nonnegativity properties of solutions, assuming they exist.

Let us now describe the matrices featured above in both the methods. The $P \times P$ matrix M is a diagonal matrix obtained after "lumping the masses" [12], i.e.,

$$M_{jj} = \sum_{k=1}^{P} \tilde{M}_{jk}, \quad \text{where } \tilde{M}_{jk} = \int_\Omega \phi_j \phi_k.$$

Clearly, $\tilde{M}$ is the standard mass matrix while M is the so-called lumped mass matrix. The remaining matrices are "stiffness" matrices, given by

$$\begin{aligned} L_{jk} &= \sum_{K \in \mathfrak{T}_h} L^K_{l(j),l(k)}, \quad L^K_{l(j),l(k)} = \int_K \vec{\nabla}\phi_{l(j)} \cdot \vec{\nabla}\phi_{l(k)}, \\ A_{jk}(z) &= \sum_{K \in \mathfrak{T}_h} A^{K,z}_{l(j),l(k)} \end{aligned} \tag{25}$$

where $A^{K,z}_{l(j),l(k)}$ is defined below and $l(j) \in \{0, 1, 2\}$ denotes the local vertex number of the lth global vertex.

To describe the 3×3 matrix $A^{K,z}$ for any triangle K and any linear function z on K, we use the following notations associated to K. Let $\vec{a}_l$, $l \in \{0, 1, 2\}$ denote the vertices of K, E_{lm} denote the edge connecting $\vec{a}_l$ and $\vec{a}_m$, and θ_{lm} denote the interior angle opposite to the edge E_{lm}. Then define

$$c_{lm}(z) = \frac{1}{|E_{lm}|} \int_{E_{lm}} e^{-(\chi/D)z}.$$

The local indices $l, m \in \{0, 1, 2\}$ are always calculated mod 3 (so, e.g., $A^K_{l,l+3} = A^K_{ll}$), and thus all nine entries of $A^{K,z}$ are defined by

$$A_{ll}^{K,z} = -L^K_{l,l+1} \frac{p_l(z)}{c_{l,l+1}(z)} - L^K_{l,l+2} \frac{p_l(z)}{c_{l,l+2}(z)} \tag{26a}$$

$$A_{l,l+1}^{K,z} = L^K_{l,l+1} \frac{p_{l+1}(z)}{c_{l,l+1}(z)} \tag{26b}$$

$$A_{l,l+2}^{K,z} = L^K_{l,l+2} \frac{p_{l+2}(z)}{c_{l+2,l}(z)} \tag{26c}$$

where $p_l(z) = e^{-(\chi/D)z(\vec{a}_l)}$. This matrix arises from a spatial discretization proposed in, as we will clarify in Section 3.2. These definitions complete the prescription of both the methods (23) and (33).

In the next section, we will prove that both the above proposed methods have monotonicity properties under certain assumptions.

Remark 3.1. In a computer implementation, we have to be careful while computing the ratios $p_l(z)/c_{lm}(z)$, so as not to divide by small numbers. The problem is evident when considering

$$\frac{p_l(z)}{c_{lm}(z)} = \frac{e^{-(\chi/D)z(\vec{a}_l)}}{\frac{1}{|E_{lm}|}\int_{E_{lm}} e^{-(\chi/D)z}}$$

where the denominator can be very small even when z takes moderately large positive values on E_{lm}. Nonetheless, observe that $\vec{a}_l$ is an endpoint of the edge E_{lm} and z is linear on E_{lm}. Hence the quantity $z - z(\vec{a}_l)$ is better suited for exponentiation. It is therefore better to compute the above ratio by implementing the following equivalent formula:

$$\frac{p_l(z)}{c_{lm}(z)} = \left(\frac{1}{|E_{lm}|}\int_{E_{lm}} e^{-(\chi/D)(z - z(\vec{a}_l))}\right)^{-1}.$$

3.2. *Derivation of the method*

We begin the derivation by obtaining a variational form of (1). Multiply the equations of (1) by a test function $\phi \in H^1(\Omega)$ and integrate by parts. Then, defining the flux

$$\vec{J} = D\vec{\nabla}u - \chi u \vec{\nabla} v_N,$$

we find that u and v_i satisfy

$$\int_\Omega (\partial_t u)\phi = -\int_\Omega \vec{J} \cdot \vec{\nabla}\phi, \tag{27a}$$

$$\int_\Omega (\partial_t v_i)\phi = -\tilde{D}_i \int_\Omega \vec{\nabla} v_i \cdot \vec{\nabla}\phi + \alpha_i \int_\Omega u\phi + \int_\Omega g_i(\vec{v})\phi, \tag{27b}$$

for all $i = 1, 2, \ldots, N$ and all $\phi \in H^1(\Omega)$. The discrete solution $\{u_h^n, v_{h,i}^n\}$ is obtained by using a suitably approximated version of the above equations for all ϕ in the previously defined finite element subspace $S_h \subseteq H^1(\Omega)$.

We now describe how each term in (27) is approximated, beginning with the last term. Let $\vec{G}_i(\vec{v}_h^{\,n})$ denote the vector whose ℓth component equals $g_{i,\ell}^n \equiv g_i(V_{1,\ell}^n, \ldots, V_{N,\ell}^n)$. At time $t = k(n+1)$, setting ϕ equal to a basis function ϕ_m,

$$\begin{aligned}\int_\Omega g_i(\vec{v})\phi_m &\approx \int_\Omega \left(\sum_{\ell=1}^{P} g_i(V_{1,\ell}^{n+1}, \ldots, V_{N,\ell}^{n+1})\phi_\ell \right) \phi_m = \sum_{\ell=1}^{P} \tilde{M}_{m\ell}\, g_{i,\ell}^{n+1} = \tilde{M}\vec{G}_i(\vec{v}_h^{\,n+1}) \\ &\approx M\vec{G}_i(\vec{v}_h^{\,n+1}).\end{aligned}$$

In the last step above, we have approximated the action of $\tilde{M}$ by that of M. This so-called process of "lumping the mass" is a well-known first order approximation [12]. Note also that in the first step above, we have approximated g_i by its Lagrange interpolant on the finite element mesh.

However, comparing with the last term in (23b), we find that in place of the term $\vec{G}_i(\vec{v}_h^{\,n+1})$ derived above, we have $\vec{G}_i^+(\vec{v}_h^{\,n+1})$ instead. Nonetheless, we shall prove in the next section, under a condition on the mesh, that any solution of (23) is nonnegative. Since

$$\vec{G}_i^+(\vec{v}_h^{\,n+1}) = \vec{G}_i(\vec{v}_h^{\,n+1}) \quad \text{whenever } \vec{v}_h^{\,n+1} \geq 0,$$

we conclude, *a posteriori*, that $\vec{G}_i^+(\vec{v}_h^{\,n+1})$ can be replaced by $\vec{G}_i(\vec{v}_h^{\,n+1})$, whenever the mesh conditions hold. The other two terms on the right hand side of (27b) are approximated in a straightforward manner using the standard finite element method, so those derivations are omitted.

To describe how the right hand side of (27a) is approximated, we recall a succinct presentation of [11] and modify it to account for our different boundary conditions and nonlinear terms. The main idea, going back to [30], is to approximate the term involving flux

$$\vec{J} = D\vec{\nabla}u - \chi u \vec{\nabla} v_N \tag{28}$$

as if the flux were spatially constant on each mesh element. Consider a mesh element K in $\mathcal{T}_h$. Let $\vec{a}_l$, $l \in \{0, 1, 2\}$ denote the vertices of K and let θ_{lm} denote the interior angle opposite to the edge connecting vertices $\vec{a}_l$ and $\vec{a}_m$. Let λ_l denote the linear function on K whose value at the vertex $\vec{a}_m$ is δ_{lm}. It is easy to prove that

$$L_{lm}^K \equiv \int_K \vec{\nabla}\lambda_m \cdot \vec{\nabla}\lambda_l = -\frac{1}{2}\cot\theta_{lm} \tag{29}$$

for all $l \neq m$ in $\{0, 1, 2\}$. It is also easy to show, using the symmetry and zero-row-sum property of the 3×3 matrix L^K, that for any two linear functions z and w,

$$\int_K \vec{\nabla}z \cdot \vec{\nabla}w = -\sum_{m<l} L_{ml}^K\, d_{lm}(z)\, d_{lm}(w), \tag{30}$$

where $d_{lm}(w) = w(\vec{a}_l) - w(\vec{a}_m)$. Let $\vec{c}$ be any constant vector. Then, combining (30) and (29), we obtain

$$\int_K \vec{c} \cdot \vec{\nabla}w = \sum_{m<l} \frac{1}{2}\vec{c} \cdot (\vec{a}_m - \vec{a}_l)\, d_{ml}(w)\, \cot\theta_{ml}. \tag{31}$$

This motivates the following approximation of the $\vec{J}$-term in (27a) when $\phi \in S_h$. The contribution to the integral on right hand side of (27a) from an element $K \in \mathcal{T}_h$ is

$$\int_K \vec{J} \cdot \vec{\nabla}\phi \approx \sum_{m<l} \frac{1}{2}\vec{J} \cdot (\vec{a}_m - \vec{a}_l)\, d_{ml}(\phi)\, \cot\theta_{ml}. \tag{32}$$

In view of (31), we expect this to be a reasonable approximation whenever ($\vec{J}$ is smooth enough and K is small enough so that) $\vec{J}$ is almost constant on each mesh element K.

Next, we describe how the term $\vec{J} \cdot (\vec{a}_m - \vec{a}_l)$ in (32) is further approximated. Letting $\vec{t}_{lm} = (\vec{a}_l - \vec{a}_m)/|\vec{a}_l - \vec{a}_m|$, it is easy to see from (28) that the following identity holds on the edge E_{lm} connecting $\vec{a}_l$ to $\vec{a}_m$:

$$\int_{E_{lm}} De^{-(\chi/D)v_N}\vec{J} \cdot \vec{t}_{ml} = \int_{E_{lm}} \vec{\nabla}(e^{-(\chi/D)v_N}u) \cdot \vec{t}_{ml} = d_{ml}(e^{-(\chi/D)v_N}u).$$

This motivates the approximation

$$\vec{J} \cdot (\vec{a}_m - \vec{a}_l) = \vec{J} \cdot \vec{t}_{ml}\,|E_{lm}| \approx \frac{|E_{lm}|\, d_{ml}(e^{-(\chi/D)v_N}u)}{\int_{E_{lm}} \frac{1}{D}e^{-(\chi/D)v_N}}.$$

Substituting this into (32), and using the approximate solution components $v^n_{h,N}$ at the nth time-step in place of v_N, and the approximate solution component u_h^{n+1} in place of u,

$$\begin{aligned}\int_K \vec{J} \cdot \vec{\nabla}\phi &\approx \sum_{m<l} \frac{1}{2}\, \frac{|E_{lm}|\, d_{ml}(e^{-(\chi/D)v^n_{h,N}}u_h^{n+1})}{\int_{E_{lm}} \frac{1}{D}e^{-(\chi/D)v^n_{h,N}}}\, d_{ml}(\phi)\ \cot\theta_{ml}\\ &= \sum_{m<l} -\frac{DL^K_{ml}}{c_{lm}(v^n_{h,N})}\, d_{ml}(e^{-(\chi/D)v^n_{h,N}}u_h^{n+1})\ d_{ml}(\phi)\\ &= \sum_{l,m} DA^{K,v^n_{h,N}}_{lm}\ u_h^{n+1}(\vec{a}_m)\,\phi(\vec{a}_l)\end{aligned}$$

for any $\phi \in S_h$. This completes the derivation of (23a), once we also approximate the time derivative in (27a) by a standard backward finite difference and lump the masses into the diagonal matrix M. The derivation of (23b) from (27b) is similar and easier, so we omit the details.

3.3. A variant of the method

The method (23) requires the solution of a nonlinear system at every time step and consequently can be expensive. A simple semi-implicit variant that does not require nonlinear solvers at each time step is the following: given U^n_ℓ and $V^n_{i,\ell}$, let U^{n+1}_ℓ and $V^{n+1}_{i,\ell}$ satisfy

$$\frac{1}{k}M(\vec{U}^{n+1} - \vec{U}^n) = -DA(v^n_{h,N})\vec{U}^{n+1} \tag{33a}$$

$$\frac{1}{k}M(\vec{V}_i^{n+1} - \vec{V}_i^n) = -\tilde{D}_iL\vec{V}_i^{n+1} + \alpha_iM\vec{U}^n + M\vec{G}_i(\vec{v}_h^n). \tag{33b}$$

In this method, in contrast to (23), we first solve the linear system (33a) to obtain $\vec{U}^{n+1}$, and then solve another linear system given by (33b) to obtain $\vec{V}_i^{n+1}$. Hence the existence of numerical solutions to (33) follows immediately from the invertibility of the two linear systems. (This invertibility, under certain mesh assumptions, is a simple consequence of the arguments in the next section.)

4. Nonnegativity of numerical solutions

In this section, we prove that the numerical solutions obtained by solving either (23) or (33) are nonnegative. We will need to assume certain conditions on the angles of mesh elements to proceed with this proof. The edges of all triangles in $\mathcal{T}_h$ form the collection of mesh edges $\mathcal{E}_h$. When an edge $e \in \mathcal{E}_h$ is shared by two triangles in $\mathcal{T}_h$, we denote by $\theta_e^\pm$ the two angles subtended by e at the two vertices opposite to e. We assume that

$$\theta_e^+ + \theta_e^- \le \pi. \tag{34a}$$

When $e \in \mathcal{E}_h$ is on the boundary $\partial\Omega$, there is only triangle adjacent to it and the angle subtended by e at its sole opposite vertex is denoted by θ_e. We assume that

$$\theta_e \le \pi/2 \tag{34b}$$

for all such edges $e \subseteq \partial\Omega$. Several software packages exist that generate meshes satisfying (34). E.g., Delaunay triangulations generated by [31] satisfy (34a). Other examples include software for generating acute triangulations [32] having elements with interior angles less than $\pi/2$ — obviously, such meshes satisfy (34a) and (34b). It is well known that some such angle condition is necessary for obtaining monotonicity properties in the finite element context (see, e.g. [12]).

4.1. Nonnegativity of the numerical cell density

We now show that the approximations u_h^n given by methods (23) and (33) are nonnegative provided the initial cell density is nonnegative.

Theorem 4.1. *Suppose that (34) holds. Fix some natural number $m > 0$ and assume that for every $0 \le n \le m$, solutions u_h^n and $v_{h,i}^n$ to (33) exist at time $t = kn > 0$. Then (for any time step size $k > 0$) a solution u_h^{m+1} to (33a) exists and we have*

$$u_h^{m+1} \ge 0 \quad \text{on } \Omega,$$

whenever $u_h^0 \ge 0$ on Ω. Similarly, if for every $0 \le n \le m$, solutions u_h^n and $v_{h,i}^{n+1}$ to (23) exist, then a solution u_h^{m+1} to (23a) exists and we have

$$u_h^{m+1} \ge 0 \quad \text{on } \Omega,$$

whenever $u_h^0 \ge 0$ on Ω.

Proof. Fix any $0 \le n \le m$. Consider (33) first. From (33a),

$$\vec{U}^{n+1} = (M + kDA)^{-1} M \vec{U}^n \tag{35}$$

where, for convenience, we have abbreviated $A = A(v_{h,N}^n)$. We will now prove that the inverse of $B = M + kDA$, appearing above, is a nonnegative matrix for any $v_{h,N}^n \in S_h$.

To this end, consider an off-diagonal entry A_{jk}, which is nonzero only if there is a mesh edge $e \in \mathcal{E}_h$ connecting the jth and kth mesh vertices. First, consider the case of boundary edges $e \subseteq \partial\Omega$. Then by (26b) or (26c), we conclude that the sign of A_{jk} is the same as the sign of an off-diagonal entry of L^K. But, this entry cannot be positive in view of (29) and (34b). Second, consider the case when e is an interior mesh edge. Then, again using (26), we find that the sign of A_{jk} is the same as the sign of

$$-(\cot\theta_e^+ + \cot\theta_e^-) = -\frac{\sin(\theta_e^+ + \theta_e^-)}{\sin\theta_e^+ \sin\theta_e^-} \le 0 \tag{36}$$

where we have used (34a). Thus

$$A_{jk} \le 0 \tag{37}$$

for all $j \ne k$.

Next, observe that (26) implies the column sum of the element matrices are zero — indeed, for any $l \in \{0, 1, 2\}$, moving indices mod 3 in (26),

$$A_{ll}^{K,z} = -L_{l,l+1}^K \frac{p_l(z)}{c_{l,l+1}(z)} - L_{l,l+2}^K \frac{p_l(z)}{c_{l,l+2}(z)}, \qquad A_{l+1,l}^K = L_{l,l+1}^K \frac{p_l(z)}{c_{l,l+1}(z)}, \qquad A_{l+2,k}^K = L_{l,l+2}^K \frac{p_l(z)}{c_{l,l+2}(z)}$$

so their sum vanishes for any function z for which the above quantities are well defined. Therefore, the matrix A obtained by summing $A^{K,z}$ over all K as in (25), has the property that

$$A_{jj} = -\sum_{j \ne k} A_{jk}. \tag{38}$$

In particular, this implies that the diagonal entries $A_{jj} \ge 0$.

Returning to B, we now find that all its diagonal entries $B_{jj} = M_{jj} + kDA_{jj}$ are positive as $M_{jj} > 0$. Moreover,

$$\begin{aligned} B_{jj} &= M_{jj} + kD \sum_{k \ne j} (-A_{kj}) \quad \text{(by (38))} \\ &= M_{jj} + kD \sum_{k \ne j} |A_{kj}| \quad \text{(by (37))} \\ &> kD \sum_{k \ne j} |A_{kj}| \quad (\text{as } M_{jj} > 0) \\ &= \sum_{k \ne j} |B_{kj}| \quad (\text{as } M_{kj} = 0). \end{aligned}$$

Thus, B is strictly diagonally dominant with positive diagonal entries. Hence [33, $M_{35} \Longleftrightarrow N_{38}$], we conclude that $B^{-1} \ge 0$. Since $M \ge 0$, by (35), we find that $\vec{U}^{n+1} \ge 0$ whenever $\vec{U}^n \ge 0$. Since this holds for any $0 \le n \le m$, then recalling $u_h^0 \ge 0$ and using a simple induction step prove the statement of the theorem for method (33).

Finally, observe that the above argument holds verbatim if we instead set $A = A(v_{h,N}^{n+1})$. Hence, we conclude that the theorem also holds for method (23). □

4.2. Nonnegativity of numerical signal concentrations

To prove that $v_{h,i}^n \geq 0$ for the first method (23), we develop a discrete version of the arguments used to prove Lemmas 2.3 and 2.5.

Theorem 4.2. *Suppose that Assumption 2.1 and (34) hold and fix some natural number $m > 0$. Suppose that for any $0 \leq n \leq m$ solutions u_h^n and $v_{h,i}^{n+1}$ solve (23) at any time $t = kn > 0$. Then for $i = 1, 2, \ldots, N$, and for any time step size $k > 0$, we have*

$$v_{h,i}^n \geq 0 \quad \text{on } \Omega,$$

for all $0 < n \leq m + 1$ whenever $v_{h,i}^0 \geq 0$ on Ω.

Proof. Supposing the result is not true, we proceed to find a contradiction. Let $n_* \geq 0$ be an integer such that there is an i_* for which $v_{h,i_*}^{n_*+1} \not\geq 0$, but $v_{h,i}^n \geq 0$ for all $n \leq n_*$ and all i, i.e., $k(n_* + 1)$ is the very first time when a numerical solution component becomes negative. (Cf. proof of Lemma 2.5 to see the analogy between the discrete and exact arguments.) Now, set ℓ_* such that

$$\min_{\ell} V_{i_*,\ell}^{n_*+1} = V_{i_*,\ell_*}^{n_*+1} < 0 \tag{39}$$

and let $\varepsilon \equiv -V_{i_*,\ell_*}^{n_*+1}$. By (23b),

$$\frac{1}{k} M_{\ell_*\ell_*} \left(V_{i_*,\ell_*}^{n_*+1} - V_{i_*,\ell_*}^{n_*} \right) = -\tilde{D}_{i_*} [L\vec{V}_{i_*}^{n_*+1}]_{\ell_*} + \alpha_{i_*} M_{\ell_*\ell_*} \vec{U}_{\ell_*}^{n_*} + M_{\ell_*\ell_*} g_{i_*,\ell_*}^{+,n_*+1}. \tag{40}$$

We proceed to establish a contradiction by examining the sign of each term above.

Beginning with the last term, we observe, in view of (24) and Assumption 2.1, that $g_{i_*,\ell_*}^{n_*+1} = g_i((V_{1,\ell_*}^{n_*+1})^+, \ldots, (V_{i_*-1,\ell_*}^{n_*+1})^+, 0, (V_{i_*+1,\ell_*}^{n_*+1})^+, \ldots, (V_{N,\ell}^{n_*+1})^+) + (0 - (-\varepsilon)) \geq \varepsilon > 0$. Since $M_{\ell_*\ell_*}$ is also positive, the last term in (40) is positive. Next, by Theorem 4.1, we know that $\vec{U}^{n_*} \geq 0$, so the penultimate term in (40) is nonnegative. To study the remaining term on the right hand side of (40), we begin with

$$-[L\vec{V}_{i_*}^{n_*+1}]_{\ell_*} = -\sum_{\ell=1}^{P} L_{\ell_*\ell} V_{i_*,\ell}^{n_*+1} = -L_{\ell_*\ell_*} V_{i_*,\ell_*}^{n_*+1} - \sum_{\ell \neq \ell_*} L_{\ell_*\ell} V_{i_*,\ell}^{n_*+1}. \tag{41}$$

Now, due to (39), $V_{i_*,\ell}^{n_*+1} \geq V_{i_*,\ell_*}^{n_*+1}$. Furthermore, due to (34), the off-diagonal entries of $-L$, being a cotangent or a sum of cotangents of interior angles, are nonnegative (see (29) and (36)). Hence

$$-L_{\ell_*\ell} V_{i_*,\ell}^{n_*+1} \geq -L_{\ell_*\ell} V_{i_*,\ell_*}^{n_*+1}$$

and (41) yields

$$-[L\vec{V}_{i_*}^{n_*+1}]_{\ell_*} \geq V_{i_*,\ell_*}^{n_*+1} \left(\sum_{\ell=1}^{P} L_{\ell_*\ell} \right) = 0$$

where we have also used the fact that the row sums of L vanish. Combining these observations, we find that the right hand side of (40) is (strictly) positive.

But the left hand side of (40) is (strictly) negative: Indeed, by the choice of n_* and ℓ_*, we have $V_{i_*,\ell_*}^{n_*+1} < 0$ and $V_{i_*,\ell_*}^{n_*} \geq 0$. Since $M_{\ell_*\ell_*} > 0$, this implies that

$$\frac{1}{k} M_{\ell_*\ell_*} \left(V_{i_*,\ell_*}^{n_*+1} - V_{i_*,\ell_*}^{n_*} \right) < 0.$$

In view of (40), this is a contradiction and it completes the proof. □

For the variant (33), in contrast to the fully implicit method (23), we are not able to prove an unconditional nonnegativity result that holds for any time step size $k > 0$. However, it is possible to obtain a result under a constraint on k. At the nth step, let

$$k_n = \min_{i=1,\ldots,N} \min_{\ell=1,\ldots,P} \left(\frac{V_{i,\ell}^n}{-\min\limits_{i,\ell}(g_i(V_{1,\ell}^n, \ldots, V_{N,\ell}^n) + \alpha_i U_\ell^n, 0)} \right).$$

Note that, this number can be computationally evaluated at any given time step. If the denominator above is zero, then k_n is defined to be $+\infty$. The nonnegativity of the next time iterate can be ensured if the next time step k is chosen so that $k \leq k_n$, as we show below. If $k_n = +\infty$, then any k would satisfy the constraint $k \leq k_n$. This is the case if $g_i(V_{1,\ell}^n, \ldots, V_{N,\ell}^n) \geq 0$ and the conditions of Theorem 4.1 hold (so $U_\ell^n \geq 0$). Note also that the constraint $k \leq k_n$ is weaker than the usual conditions for explicit methods (such as $k \leq O(h^2)$), which can be much more stringent depending on the spatial mesh size h.

Proposition 4.3. *Suppose that (34) holds. Given $u_h^n \geq 0$ and $v_{h,i}^n \geq 0$ (for all $i = 1, 2, \ldots, N$), suppose that we calculate $v_{h,i}^{n+1}$ using (33b) with any $0 < k \leq k_n$. Then*

$$v_{h,i}^{n+1} \geq 0 \quad \text{on } \Omega,$$

for all $i = 1, 2, \ldots, N$.

Proof. Clearly, (33b) implies that

$$(M + k\tilde{D}_i L)\vec{V}_i^{n+1} = M\vec{V}_i^n + k\alpha_i M\vec{U}^n + kM\vec{G}_i(\vec{v}_h^n). \tag{42}$$

An argument similar to that in the proof of Theorem 4.1 shows that $M + k\tilde{D}_i L$ has a nonnegative inverse. Hence, it suffices to prove that the right hand side of (42) is nonnegative.

By the definition of k_n, we have $V_{i,\ell}^n \geq -k_n \min(g_i(V_{1,\ell}^n, \ldots, V_{N,\ell}^n) + \alpha_i U_\ell^n, 0)$ for any i and ℓ, so

$$V_{i,\ell}^n + k\alpha_i U_\ell^n + kg_i(V_{1,\ell}^n, \ldots, V_{N,\ell}^n) \geq (-k_n + k)\min(g_i(V_{1,\ell}^n, \ldots, V_{N,\ell}^n) + \alpha_i U_\ell^n, 0) \geq 0.$$

Thus the right hand side of (42) is nonnegative. □

5. Application to the minimal Keller–Segel model

In this section, we apply the previously developed numerical method to the minimal Keller–Segel system (of Example 1.1). This will lead us to exhibit a nonhomogeneous stationary state of the system, which to the best of our knowledge, that has not been seen before.

This system of two coupled partial differential equations was considered in the original work of Keller and Segel [2] to model the spontaneous aggregation of slime mold *Dictyostelium discoideum*. They obtained this system from a larger system by certain simplifying assumptions and they stress that "these simplifications, while reasonable, are made principally to avoid obscuring essential features with heavy calculations". Ever since, the qualitative behavior of this system and its solutions have been extensively studied [6,34]. Considering the wealth of known results, this model forms a good test bed for numerical methods for chemotaxis. To simplify our discussion, we set all parameters to unity and consider the following minimal system:

$$u_t = \nabla \cdot (\nabla u - u\nabla v), \quad x \in \Omega,\ t > 0 \tag{43a}$$

$$v_t = \Delta v - v + u, \quad x \in \Omega,\ t > 0 \tag{43b}$$

$$\frac{\partial u}{\partial n} = \frac{\partial v}{\partial n} = 0, \quad x \in \partial\Omega,\ t > 0 \tag{43c}$$

$$u(0, x) = u_0(x), \qquad v(0, x) = v_0(x), \quad x \in \Omega. \tag{43d}$$

Even with these simplifications, the system is sensitive to the size of the domain Ω and the initial conditions, as we shall now see.

Let us fix Ω to be a disk of radius R centered at the origin, and review three known results on how the behavior of solutions of (43) depends on R and u_0. First, it is known that if

$$\int_\Omega u_0(x)dx < 4\pi, \tag{44}$$

then the solution to (43) exists and is uniformly bounded globally in time [6,35]. Let $m(w)$ denote the mean value of a function w on Ω. Inequality (44), in terms of the mean value of the initial iterate $m(u_0)$, is the same as

$$m(u_0) < 4/R^2. \tag{45}$$

Second, if

$$\frac{4\pi}{\text{meas}(\Omega)} < m(u_0) < \frac{8\pi}{\text{meas}(\Omega)}, \tag{46}$$

then there exist initial data $\{u_0, v_0\}$ for which the solution of (43) blows up at the boundary of Ω in finite or infinite time (see [6, Table 5], [36] or [37]). This reinforces the need to satisfy (45) if we want to find the smallest value of $m(u_0)$ that yields (bounded) pattern formation. Third, an analysis of [34] (or its extension in [3]) shows that any constant solution u_* of (43) is unstable if

$$u_* > 1 + \mu_1(\Omega)$$

where $\mu_1(\Omega)$ is the smallest nonzero eigenvalue of the Laplacian with Neumann boundary conditions on Ω. These results suggest that to avoid convergence to the biologically uninteresting constant stationary states, we should choose the initial iterate to have mean value $m(u_0) > 1 + \mu_0$. Since the mean value of u remains unchanged (equals $m(u_0)$) at all times, the

Table 1
Indicators of convergence.

h	$\\|v_h^n - v_{2h}^n\\|_0$	$\\|u_h^n - u_h^{n-1}\\|_1$	$\\|v_h^n - v_h^{n-1}\\|_1$	$\|m(u_h^n) - m(u_0)\|$	$\\|u_h^n - U_h\\|_0$	$\dim(S_h)$
0.02454	0.03108	1.1×10^{-12}	1.6×10^{-13}	1.4×10^{-11}	7.9×10^{-13}	2,113
0.01227	0.29949	5.2×10^{-07}	6.1×10^{-10}	1.2×10^{-10}	9.1×10^{-09}	8,321
0.00614	0.02185	4.7×10^{-10}	9.5×10^{-13}	2.0×10^{-10}	2.0×10^{-11}	33,025
0.00307	0.00486	8.7×10^{-10}	1.9×10^{-12}	7.0×10^{-10}	6.8×10^{-11}	131,585
0.00153	0.00119	4.2×10^{-09}	6.8×10^{-12}	7.4×10^{-09}	7.3×10^{-10}	525,313

instability of the constant solution with the same mean value suggests that u will not approach this constant solution as $t \to \infty$.

Guided by these three results, we are led to perform numerical computations with initial data u_0 satisfying $1 + \mu_1(\Omega) < m(u_0) < 4/R^2$. The eigenvalue $\mu_1(\Omega)$, in our case of the disk Ω, can be found by a simple calculation involving a Bessel root: Namely, $\mu_1(\Omega) = \hat{\mu}_1/R^2$, where $\hat{\mu}_1 \approx 3.38996$ is the eigenvalue $\mu_1(\hat{D})$ on the unit disk $\hat{D}$. Thus, we choose u_0 satisfying

$$1 + \frac{\hat{\mu}_1}{R^2} < m(u_0) < \frac{4}{R^2}. \tag{47}$$

Note that, the unit disk does not satisfy (47). We set $R = 0.5$ so as to satisfy (47).

Next, we come to the question of selection of initial condition for the time iteration. Since our goal in this section is to compute a stationary state, we are now motivated by an analysis of [36,6], that establishes the existence of nonhomogeneous stationary states using a Mountain Pass argument. The argument proves the existence by locating the critical point of a functional using a Palais–Smale sequence of functions

$$z_\varepsilon = w_\varepsilon - m(w_\varepsilon), \tag{48}$$

as ε goes to 0, where

$$w_\varepsilon = \log\left(\frac{\varepsilon^2}{(\varepsilon^2 + \pi(x - x_0)^2)^2}\right) \tag{49}$$

and x_0 is a point in $\partial\Omega$. Motivated by this construction, we set our initial iterates by

$$u_0 = v_0 = z_\varepsilon + c \tag{50}$$

with $x_0 = (-R, 0)$, $\varepsilon = 10^{-10}$, and set the constant c so that $m(u_0) = 15$, thus satisfying (47).

With the above described settings, we use the implicit method (23). We use an almost uniform spatial mesh consisting of triangles with non-obtuse interior angles. (One of the meshes used is visible in Fig. 3.) All triangles of the mesh have an approximate diameter h obtained by dividing the perimeter of Ω by number of mesh vertices on $\partial\Omega$. We then set the time step size by $k = h$ and compute approximations u_h^n and v_h^n at time kn, for large n, by solving (23). Note that, (23b) does not require us to solve a nonlinear system because g is linear for the minimal Keller–Segel model.

We observed that the u and v iterates at every time step were nonnegative, as predicted by the theory in Section 4. We performed enough time iterations to reach $t = 20$, on a sequence of spatial meshes with decreasing h. The values of h for the meshes we used, and the corresponding size of the discrete linear systems to be inverted at each time step, are displayed in the first and the last columns of Table 1. To describe the results in the remainder of the table, first recall the definitions of the standard $L^2(\Omega)$ and $H^1(\Omega)$ norms,

$$\|w\|_0 = \left(\int_\Omega w^2\right)^{1/2} \quad \text{and} \quad \|w\|_1 = \left(\int_\Omega w^2 + |\vec{\nabla} w|^2\right)^{1/2}.$$

We report, in the second and third column of Table 1, the $H^1(\Omega)$ norms of the differences between the last two successive iterates, namely $\|u_h^n - u_h^{n-1}\|_1$ and $\|v_h^n - v_h^{n-1}\|_1$, where n denotes the final time iteration number. Since these differences are small, it appears that we are close to a stationary state. We also examine how these approximately computed stationary states differ as the mesh size h decreases. Note that h is halved as we go down a row in Table 1, which corresponds to a uniform mesh refinement with the boundary points adjusted to lie on $\partial\Omega$. The first column of the table suggests that the difference, in $L^2(\Omega)$, between the stationary v computed on two successive mesh refinements decreases like $O(h^2)$, for small enough values of h.

Next, we describe two further tests we performed to check the simulation. First, it is easy to see, by integration by parts applied to (43a), that $\partial_t m(u) = 0$. Thus, the mean value of $u(\cdot, t)$ at any time t must equal the initial mean $m(u_0)$, i.e.,

$$m(u) - m(u_0) = 0. \tag{51}$$

We verify that this property is preserved to good accuracy by our discrete approximation of u in the fourth column of Table 1. To describe the second test, let us begin by noting that an application of the maximum principle to the steady state

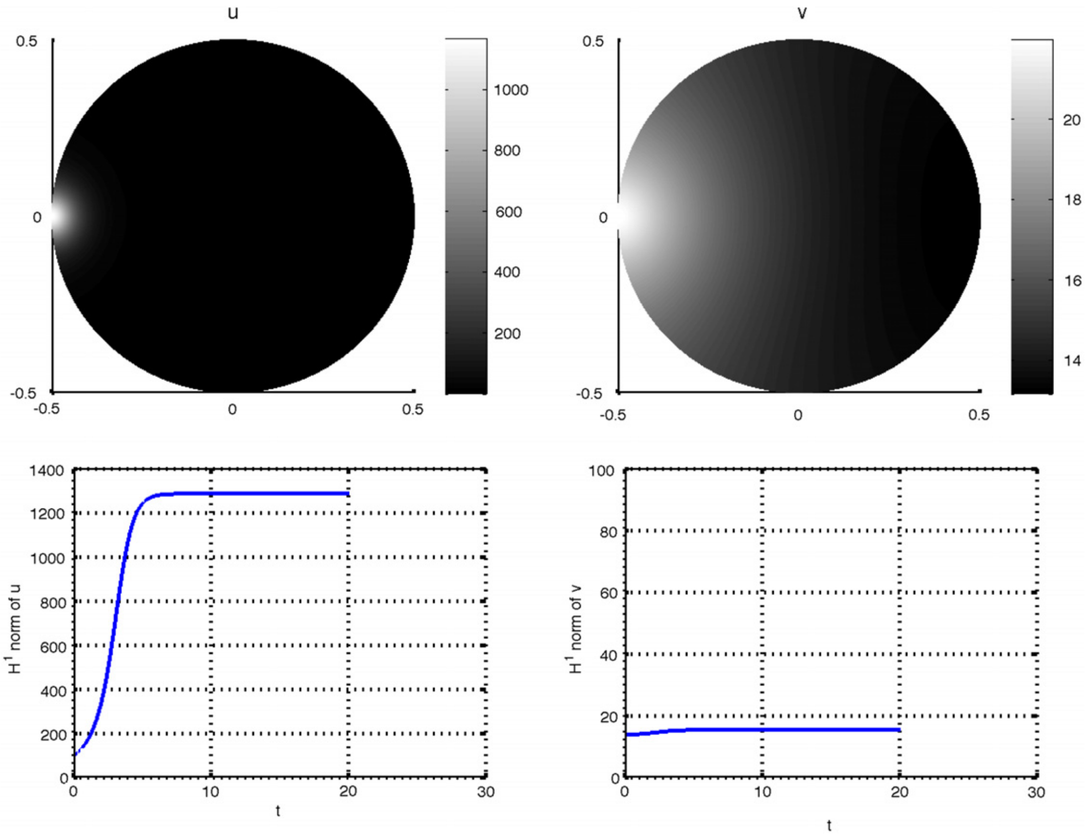


Fig. 2. Plot of final u and v and history of their norms.

version of (43a) shows that the stationary u and v satisfy $u = Ke^v$ for some constant K. Moreover, integrating we find that $K = m(u_0)/m(e^v)$. Thus, if u and v are the exact stationary states, then defining

$$U = \frac{m(u_0)}{m(e^v)} e^v, \tag{52}$$

we have

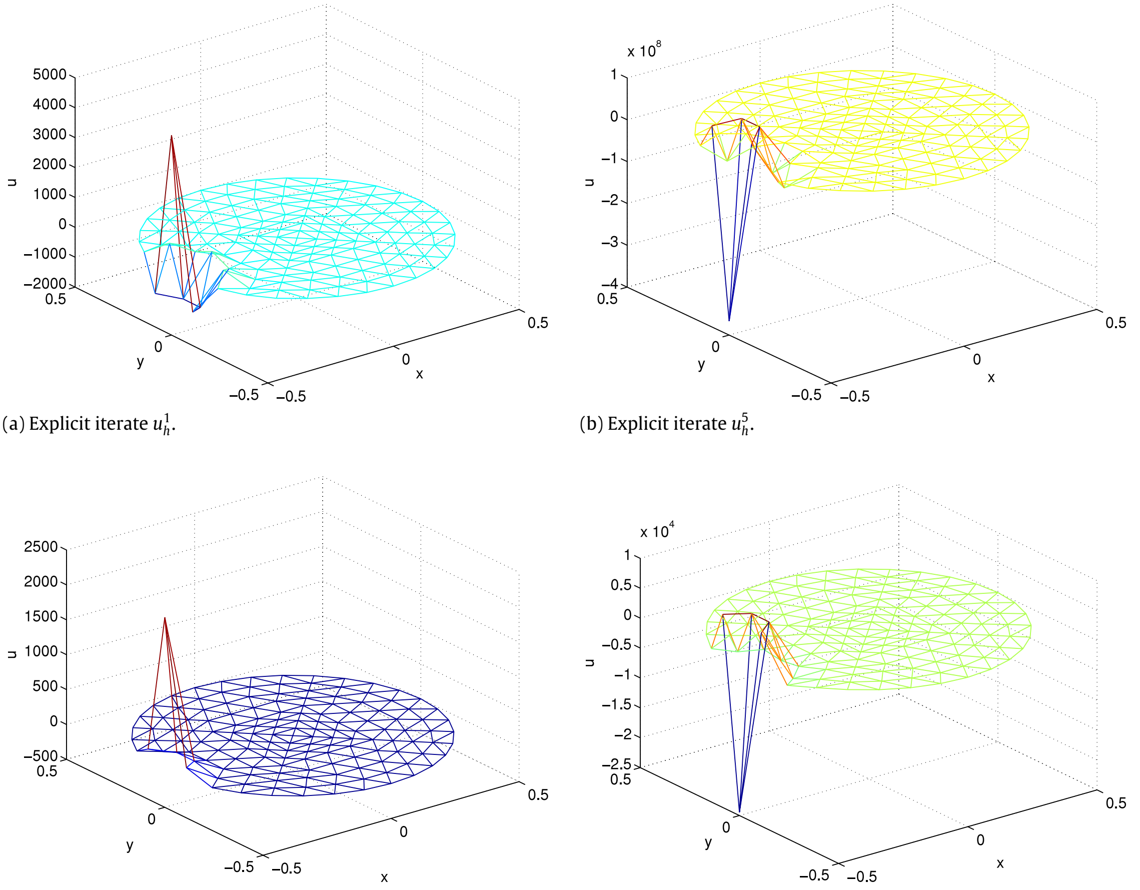
$$u = U. \tag{53}$$

Define U_h by replacing e^v in (52) by the finite element function whose nodal values coincide with those of $e^{v_h^n}$. Our method possesses a highly accurate discrete analogue of the property (53), namely $u_h^n \approx U_h$, as can be seen from the sixth column of Table 1.

The computed solution components u and v at the final time step are shown in Fig. 2. (This is obtained using the finest mesh size in Table 1.) The population u seems to aggregate near the point $(-R, 0)$. The signal concentration v also peaks at the same point. The history of the $H^1(\Omega)$ norms of the time iterates, also shown in Fig. 2, indicate that the norms have stabilized after a short transient region.

Note that by the spatial symmetry inherent in (43a)–(43c), if one rotates a stationary solution by any angle, one obtains another stationary solution. In Fig. 2, we have exhibited an approximation of just *one stationary solution from an infinite family of solutions*. We observed that it is possible to find other solutions in this family by our numerical method, simply by changing the initial iterate: More specifically, changing the choice of $x_0 = (-R, 0)$ in (49) is enough. Finally, we note that our (unreported) computational experience suggests that there are further families of stationary solutions of the minimal Keller–Segel model with multiple peaks. However, more elaborate numerical methods are needed to capture all the stationary states in a systematic way. Design of such methods seems to be a subject worthy of future research.

Although we are primarily concerned with convergence near stationary solutions in this work, let us briefly examine a blow-up case. We apply the method in the parameter range given by (46) where initial conditions which lead to blow-up of the exact solution have been proven to exist. (Blow-up refers to the situation where the $L^\infty(\Omega)$-norm of the exact solution can be proved to approach $+\infty$ in finite or infinite time.) It is not known how to construct a u_0 that guarantees blow-up. In the absence of such information, we experiment with u_0 as in (50), except that we now set the constant c there so that $m(u_0) = 17$, thus satisfying (46). As with the previous simulations, we chose $x_0 = (-R, 0)$, $\varepsilon = 10^{-10}$, and $k = h$, and



(a) Explicit iterate u_h^1.

(b) Explicit iterate u_h^5.

(c) Implicit iterate u_h^{181}.

(d) Implicit iterate u_h^{186}.

Fig. 3. False blow ups and positivity violations.

Table 2
Indicators of convergence vs. blow up.

h	$m(u_0) = 15$		$m(u_0) = 17$	
	$\|u_h^n - u_{2h}^n\|_1$	$\|v_h^n - v_{2h}^n\|_1$	$\|u_h^n - u_{2h}^n\|_1$	$\|v_h^n - v_{2h}^n\|_1$
0.02454	5.8×10^4	5.42866	6.6×10^4	6.19
0.01227	6.4×10^4	10.4524	2.6×10^5	6.22
0.00614	2.5×10^2	0.31918	1.1×10^6	6.23
0.00307	6.8×10^1	0.09772	4.2×10^6	6.23
0.00153	2.7×10^1	0.04114	1.7×10^7	6.24

performed enough iterations to reach $t = 20$. We observed that even in this parameter range which allows for potential blow up, iterates of our method do remain positive as predicted by our theory.

It is known that blow-up in finite time is typically exhibited as the formation of one or more δ-singularities at the boundary [6,38]. A numerical manifestation of this was observed in the behavior of the iterates in the $m(u_0) = 17$ case. Namely, although the iterates appeared to tend to a "stationary solution" on each mesh, as the mesh is refined, the apparent "stationary solution" begins to look more and more like a δ singularity in the $m(u_0) = 17$ case. This did not happen in the $m(u_0) = 15$ case. The contrast between the two cases is further clarified using the indicators reported in Table 2: As $h \to 0$, while $\|u_h^n - u_{2h}^n\|_1$ decreases in the $m(u_0) = 15$ case, it increases in the $m(u_0) = 17$ case.

To summarize these numerical findings, we observed that:

(1) the time iterates are nonnegative,
(2) the total species mass is conserved at each time up to round-off errors,
(3) for each fixed h, by the time $t = 20$, the iterates seemed to have converged in time to a stationary pattern when $m(u_0) = 15$,
(4) the apparent stationary u and v iterates satisfy a discrete analogue of (53), an exponential relation satisfied by the exact stationary solution,

(5) as the spatial mesh is refined ($h \to 0$), the solution at $t = 20$ appears to converge with h to a fixed bounded spatial pattern when $m(u_0) = 15$,
(6) when $m(u_0) = 17$, we continued to observe (1)–(4), but instead of (5), we observed blow-up as $h \to 0$.

Before concluding this section, a few words of caveat are in order, to show what can go wrong with naive choices for the numerical method. Suppose we use the standard Lagrange finite element method with explicit time stepping to discretize (43). It is well known [12] that in the presence of a parabolic term, the explicit time step size k must be chosen of the order of h^2. Returning to the $m(u_0) = 15$ case and choosing $k = h^2/10$, on a mesh with $h = 0.0625$, we find that the very first iterate u_h^1 exhibits negative values — see Fig. 3(a). These negative values subsequently get amplified in the next iterations (the fifth u iterate, as shown in Fig. 3(b), is orders of magnitude larger). The norms of the iterates quickly blow up forcing the computations to exit. The reason for this instability is that k was not adequately small. The difficulty with this type of (explicit) methods is that one must choose $k = \kappa h^2$, where κ depends on the previous iterates. A conservative estimate of κ saves the computation from blow up, but leads to extremely small time steps, and consequently expensive long computations.

Since $m(u_0) = 15$ in the above experiment, we are in the regime where (44) is satisfied, so exact solutions are known to be bounded. Hence the numerical blow up observed in Fig. 3 fails to predict the behavior of the exact solution (which does not blow up). This shows the fallacy of concluding that a model admits chemotactic collapse merely because solutions from a poorly chosen numerical method blow up.

We also note that, the use of an implicit method need not alleviate this problem. To show this, we consider an analogue of our implicit method (23) with $k = h$, but using *standard* finite element matrices (instead of (26)). The first iterate exhibiting negative values is shown in Fig. 3(c) (and the situation gets worse in five more iterations as seen in Fig. 3(d)). Unlike the explicit scheme, the reason for the failure of this implicit scheme, lies in its inability to handle the convective chemotactic term. It is well-known that the standard finite element method is not monotone, and moreover, produces spurious oscillations in convection-dominated problems. In chemotactic models, the convection strength is determined by the solution iterates, and varies from point to point in both space and time, often reaching high values.

In view of these complications, the advantages of a monotone scheme like (23) are clear. It appears to remain stable, and yield provably nonnegative solutions, irrespective of the time step size, spatial mesh size, and the strength and location of convective or diffusive regions. A faster simulation technique can be obtained by incorporating adaptive meshing using smaller elements where the solution changes more rapidly. A method like (23) with good pre-asymptotic stability behavior would then be critical.

Appendix. Minimum principles

In this appendix, we collect a few minimum/maximum principles [39] for general parabolic equations that we needed to use in the preceding analysis. Less standard results are given with proofs, while the more standard ones are accompanied by references.

Let L denote a uniformly elliptic differential operator of the form

$$Lw = -\sum_{i,j} a^{ij}(x,t) w_{x_i x_j}(x,t) + \sum_i b^i(x,t) w_{x_i}(x,t) + c(x,t)w(x,t) \tag{54}$$

with (everywhere) continuous and bounded coefficients a^{ij}, b^i and c. The following version of Hopf's lemma for a function z can be obtained by applying [40, Lemma 2.8] to $-z$. The notations are illustrated in Fig. 4. Loosely speaking, the lemma shows that at a boundary minimum (x', s) of z, its outward normal derivate must be strictly negative.

Lemma A.1 (*Boundary Point Lemma*). *Let $\eta > 0$ and $R > 0$ be some fixed constants, and fix $(y, s) \in \mathbb{R}^{N+1}$. Suppose that the operator $\partial_t + L$ is uniformly parabolic in the lower parabolic frustum*

$$F_R = \{(x,t) \in \mathbb{R}^{N+1} : |x-y|^2 + \eta^2(s-t) < R^2, t < s\}.$$

Suppose that $z(x,t)$ has continuous derivatives $\partial_t z$, $\partial_i z$, $\partial_i\partial_j z$ for all i, j, and let $(x', s) \in \mathbb{R}^{N+1}$ with $|x'-y| = R$. Let $C_{R/2} = \{(x,t) \in F_R : |x-y| \le R/2\}$. If

$$\partial_t z + Lz \ge 0 \quad \textit{for all } (x,t) \in F_R, \tag{55a}$$

$$z(x', s) \le z(x,t) \quad \textit{for all } (x,t) \in F_R, \tag{55b}$$

$$z(x', s) < z(x,t) \quad \textit{for all } (x,t) \in C_{R/2}, \tag{55c}$$

$$c(x,t)z(x', s) \le 0 \quad \textit{for all } (x,t) \in F_R, \tag{55d}$$

and if $n = (x'-y)/|x'-y|$, then

$$\frac{\partial z}{\partial n}(x', s) < 0 \tag{56}$$

and $z(x,t) > z(x', s)$ for all $(x,t) \in F_R$.

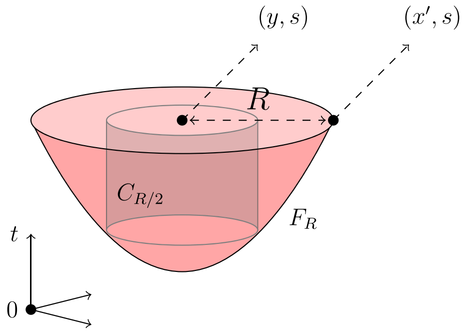


Fig. 4. The parabolic frustum in Lemma A.1.

Next, we recall the strong minimum (maximum) principle, which can be found e.g. [41, Theorem 12 in Section 7.1]. Let $\Omega_T = \Omega \times (0, T]$.

Theorem A.2 (*Minimum Principle*). *Let $c \geq 0$ in Ω_T. Assume that w, $\partial_t w$, $\partial_i w$ and $\partial_i \partial_j w$ for all i, j, are continuous on $\bar{\Omega}_T$ and $\partial_t + L$ is uniformly parabolic in Ω_T. If*

$$w_t + Lw \geq 0 \quad \text{in } \Omega_T$$

and w attains a nonpositive minimum over $\bar{\Omega}_T$ at point $(x_0, t_0) \in \Omega_T$ then w is a constant on $\bar{\Omega}_{t_0}$.

The following lemma is similar to [42, Corollary 2.3 in Chapter 7], but the proof can be simplified significantly using Lemma A.1, as we show below.

Lemma A.3. *Let $\partial_t + L$ be a uniformly parabolic operator with $c \geq 0$. If v is continuous on $\bar{\Omega}_T$ and satisfies*

$$\partial_t v + Lv \geq 0 \quad \text{in } \Omega_T \tag{57}$$

$$\beta(x)v + \frac{\partial v}{\partial n} = 0 \quad \partial\Omega \times (0, T] \tag{58}$$

$$v = \gamma \quad \text{on } \Omega \times \{t = 0\} \tag{59}$$

with a $\gamma(x) \geq 0$ that does not vanish everywhere on $\bar{\Omega}$, and a continuous $\beta(x) \geq 0$ on $\bar{\Omega}$, then $v(x, t) > 0$ for all $(x, t) \in \Omega_T$.

Proof. We first claim that $v \geq 0$ on $\bar{\Omega}_T$. Indeed, if not, there is an $(x_0, t_0) \in \bar{\Omega}_T$ such that

$$v(x_0, t_0) = \min_{(x,t)\in\bar{\Omega}_T} v(x, t) < 0 \tag{60}$$

and $t_0 > 0$ as $\gamma \geq 0$. Now, if $x_0 \in \Omega$, then by the strong minimum principle (Theorem A.2) $v \equiv$ constant <0 on $\bar{\Omega}_{t_0}$ which contradicts $v(x, 0) = \gamma \geq 0$ at $t = 0$.

Hence the claim will be proved if we also find a contradiction when $x_0 \in \partial\Omega$. We can assume, without loss of generality that the minimum in (60) is achieved only on the boundary $\partial\Omega$, because if it is also achieved anywhere in the interior Ω, then the argument in the previous paragraph applies. Therefore (55c) holds with $(x', s) = (x_0, t_0)$ and we can apply the boundary point lemma, Lemma A.1 (whose remaining assumptions are easily verified), to conclude that $\partial v/\partial n < 0$ at (x_0, t_0). But then,

$$\beta(x_0)v(x_0, t_0) + \frac{\partial v}{\partial n}(x_0, t_0) < 0$$

which contradicts (58). This proves that $v \geq 0$ on $\bar{\Omega}_T$.

It only remains to prove that strict inequality $v > 0$ on Ω_T. Again, proceeding by contradiction, suppose not. Then, there is an $x' \in \bar{\Omega}$ and $d > 0$ such that $v(x', s) = 0$ (since we have already shown that $v \geq 0$). Therefore, the minimum of v over $\bar{\Omega}_T$ is achieved at (x', s) and so if $x' \in \Omega$, the strong minimum principle (Theorem A.2) gives that $v \equiv 0$ in Ω_{t_0}. By the continuity of v, this implies that $v(x, 0) = \gamma(x) \equiv 0$, which contradicts the given assumption on γ. If, on the other hand, $x' \in \partial\Omega$, then as in the previous paragraph, we can assume without loss of generality that the minimum is obtained only on the boundary and apply Lemma A.1 to find a contradiction to (58). □

The next lemma states a result for more general c. Although it is stated for a finite time interval $(0, T]$, the proof holds verbatim on the infinite time interval $(0, \infty)$ if c is bounded.

Lemma A.4. *Suppose w is a smooth solution of*

$$\partial_t w + Lw = 0 \quad \text{in } \Omega_T, \tag{61a}$$

$$\frac{\partial w}{\partial n} = 0 \quad \text{on } \partial\Omega \times (0, T], \tag{61b}$$

$$w = \gamma \quad \text{on } \Omega \times \{t = 0\}. \tag{61c}$$

where $\partial_t + L$ is a uniformly parabolic operator. If $\gamma(x) \geq 0$ does not vanish everywhere and if the function c is bounded below, i.e.,

$$m = \inf_{\bar{\Omega}\times(0,T]} c(x, t) > -\infty,$$

then $w > 0$ on Ω_T.

Proof. Let $v(x, t) := e^{mt}w(x, t)$. Then for any i and j, we have $\partial_i v = e^{mt}\partial_i w$ and $\partial_i\partial_j v = e^{mt}\partial_i\partial_j w$. Hence, $Lv = e^{mt}Lw$. Together with (61a), this implies

$$\partial_t v = e^{mt}\partial_t w + me^{mt}w = -e^{mt}Lw + mv = -Lv + mv.$$

Define $Kv := Lv - mv$. We have just proved that $\partial_t v + Kv = 0$. Moreover, (61b) implies that $\partial v/\partial n = 0$ on $\partial\Omega \times (0, T]$ and (61c) implies the initial condition $v(x, 0) = \gamma(x)$. Noting that $\partial_t + K$ is a uniformly parabolic operator with a nonnegative lowest order term (as $c - m \geq 0$), we can apply Lemma A.3 with $\beta \equiv 0$, and w and L replaced by v and K, respectively. We conclude that $v > 0$ on Ω_T and thus $w = e^{-mt}v > 0$ as well. □

References

[1] C.S. Patlak, Random walk with persistence and external bias, Bull. Math. Biophys. 15 (3) (1953) 311–338.

[2] E.F. Keller, L.A. Segel, Initiation of slime mold aggregation viewed as an instability, J. Theoret. Biol. 26 (1970) 399–415.

[3] P. De Leenheer, J. Gopalakrishnan, E. Zuhr, Instability in a generalized Keller–Segel model, J. Biol. Dyn. 6 (2012) 974–991.

[4] A. Fasano, A. Mancini, M. Primicerio, Equilibrium of two populations subject to chemotaxis, Math. Models Methods Appl. Sci. 14 (2004) 503–533.

[5] D. Horstmann, Generalizing the Keller–Segel model: Lyapunov functionals, steady state analysis, and blow-up results for multi-species chemotaxis models in the presence of attraction and repulsion between competitive interacting species, J. Nonlinear Sci. 21 (2011) 231–270.

[6] D. Horstmann, From 1970 until present: the Keller–Segel model in chemotaxis and its consequences, I, Jahresber. Dtsch. Math.-Ver. 105 (2003) 103–165.

[7] B. Perthame, Transport Equations in Biology, in: Frontiers in Mathematics, Birkhäuser-Verlag, Basel, 2007.

[8] C. Klein, Induction of phosphodiesterase by cyclic Adenosine 3′ : 5′-monophosphate in differentiating dictyostelium discoideum Amoebae, J. Biol. Chem. 250 (1975) 7134–7138.

[9] S. Childress, J.K. Percus, Nonlinear aspects of chemotaxis, Math. Biosci. 56 (1981) 217–237.

[10] R. Courant, Variational methods for the solution of problems of equilibrium and vibrations, Bull. Amer. Math. Soc. 49 (1943) 1–23.

[11] J. Xu, L. Zikatanov, A monotone finite element scheme for convection-diffusion equations, Math. Comp. 68 (1999) 1429–1446.

[12] V. Thomée, Galerkin Finite Element Methods for Parabolic Problems, second ed., in: Springer Series in Computational Mathematics, vol. 25, Springer-Verlag, Berlin, 2006.

[13] H.F. Weinberger, Invariant sets for weakly coupled parabolic and elliptic systems, Rend. Mat. Appl. (6) 8 (1975) 295–310. Collection of articles dedicated to Mauro Picone on the occasion of his ninetieth birthday.

[14] R. Strehl, A. Sokolov, S. Turek, Efficient, accurate and flexible finite element solvers for chemotaxis problems, Comput. Math. Appl. 64 (3) (2012) 175.

[15] A. Chertock, A. Kurganov, A second-order positivity preserving central-upwind scheme for chemotaxis and haptotaxis models, Numer. Math. 111 (2008) 169–205.

[16] N. Saito, Conservative upwind finite-element method for a simplified Keller–Segel system modelling chemotaxis, IMA J. Numer. Anal. 27 (2) (2007) 332–365.

[17] N. Saito, Error analysis of a conservative finite-element approximation for the Keller–Segel system of chemotaxis, Commun. Pure Appl. Anal. 11 (1) (2012) 339–364.

[18] A. Marrocco, Numerical simulation of chemotactic bacteria aggregation via mixed finite elements, M2AN Math. Model. Numer. Anal. 37 (2003) 617–630.

[19] M. Brera, J.W. Jerome, Y. Mori, R. Sacco, A conservative and monotone mixed-hybridized finite element approximation of transport problems in heterogeneous domains, Comput. Methods Appl. Mech. Engrg. 199 (2010) 2709–2770.

[20] F. Brezzi, L.D. Marini, P. Pietra, Two-dimensional exponential fitting and applications to drift-diffusion models, SIAM J. Numer. Anal. 26 (1989) 1342–1355.

[21] S. Holst, An a priori error estimate for a monotone mixed finite-element discretization of a convection-diffusion problem, Numer. Math. 109 (2008) 101–119.

[22] Y. Epshteyn, A. Izmirlioglu, Fully discrete analysis of a discontinuous finite element method for the Keller–Segel chemotaxis model, J. Sci. Comput. 40 (2009) 211–256.

[23] Y. Epshteyn, A. Kurganov, New interior penalty discontinuous Galerkin methods for the Keller–Segel chemotaxis model, SIAM J. Numer. Anal. 47 (2008–2009) 386–408.

[24] R. Tyson, L.G. Stern, R.J. LeVeque, Fractional step methods applied to a chemotaxis model, J. Math. Biol. 41 (2000) 455–475.

[25] K. Baba, M. Tabata, On a conservative upwind finite element scheme for convective diffusion equations, RAIRO Anal. Numer. 15 (1981) 3–25.

[26] R. Strehl, A. Sokolov, D. Kuzmin, S. Turek, A flux-corrected finite element method for chemotaxis problems, Comput. Methods Appl. Math. 10 (2) (2010) 219–232.

[27] C. Cosner, Analogues of maximum principles for systems, Lecture Notes, Private communication, unpublished.

[28] A. Friedman, Partial Differential Equations of Parabolic Type, Prentice-Hall Inc., Englewood Cliffs, NJ, 1964.

[29] H. Amann, Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems, in: Function Spaces, Differential Operators and Nonlinear Analysis (Friedrichroda, 1992), in: Teubner-Texte Math., vol. 133, Teubner, Stuttgart, 1993, pp. 9–126.

[30] P.A. Markowich, M.A. Zlámal, Inverse-average-type finite element discretizations of selfadjoint second-order elliptic problems, Math. Comp. 51 (1988) 431–449.

[31] J.R. Shewchuk, Triangle: engineering a 2D quality mesh generator and delaunay triangulator, in: M.C. Lin, D. Manocha (Eds.), Applied Computational Geometry: Towards Geometric Engineering, in: Lecture Notes in Computer Science, vol. 1148, Springer-Verlag, New York, 1996, pp. 203–222.

[32] H. Erten, A. Üngör, Computing acute and non-obtuse triangulations, in: Canadian Conference on Computational Geometry, CCCG, 2007, pp. 205–208.

[33] A. Berman, R.J. Plemmons, Nonnegative Matrices in The Mathematical Sciences, in: Classics in Applied Mathematics, vol. 9, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994, Revised reprint of the 1979 original.

[34] R. Schaaf, Stationary solutions of chemotaxis systems, Trans. Amer. Math. Soc. 292 (1985) 531–556.

[35] T. Nagai, T. Senba, K. Yoshida, Application of the Trudinger–Moser inequality to a parabolic system of chemotaxis, Funkcial. Ekvac. 40 (1997) 411–433.

[36] D. Horstmann, The nonsymmetric case of the Keller–Segel model in chemotaxis: some recent results, NoDEA Nonlinear Differential Equations Appl. 8 (2001) 399–423.

[37] T. Senba, T. Suzuki, Local and norm behavior of blowup solutions to a parabolic system of chemotaxis, J. Korean Math. Soc. 37 (2000) 929–941. International Conference on Differential Equations and Related Topics (Pusan, 1999).

[38] V. Nanjundiah, Chemotaxis, signal relaying and aggregation morphology, J. Theoret. Biol. 42 (1) (1973) 63–105.

[39] M.H. Protter, H.F. Weinberger, Maximum Principles in Differential Equations, Prentice-Hall Inc., Englewood Cliffs, NJ, 1967.

[40] G.M. Lieberman, Second Order Parabolic Differential Equations, World Scientific Publishing Co. Inc., River Edge, NJ, 1996.

[41] L.C. Evans, Partial Differential Equations, in: Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 1998.

[42] H. Smith, Monotone Dynamical Systems, Amer. Math. Soc., 1995.