Unconditionally bound preserving and energy dissipative schemes for a class of Keller–Segel equations[*]

Jie Shen[†] and Jie Xu[‡]

Abstract

We propose numerical schemes for a class of Keller–Segel equations. The discretization is based on the gradient flow structure. The resulting first-order scheme is mass conservative, bound preserving, uniquely solvable and energy dissipative, and the second-order scheme satisfies the first three properties. For parabolic–elliptic equations, the schemes are decoupled. Numerical examples are presented to show that besides the above properties, the schemes are efficient and able to capture the spiky solutions for the aggregation in chemotaxis.

Keywords. Keller–Segel equations, chemotaxis, gradient flows, bound preserving, energy stability
AMS subject classification. 65M12, 35K61, 35K55, 65Z05, 92C17

1 Introduction

The Keller–Segel equations, proposed in [24, 20, 21], describe chemotaxis in biology. Chemotaxis refers to the motion of organisms according to signals (usually chemical, called chemoattractants) that can be generated by the organisms themselves. The chemotaxis system has two competing mechanisms: the diffusion from the Brownian motion, and the aggregation directed by the signals. This kind of competition can be found in various other systems. For each organism and signal, the evolution is described by a parabolic equation. In many cases, the signal responses to the concentration of organism much faster than the organism responses to the signal. So it is common to simplify the parabolic equation for the signal as an elliptic equation.

In the classical Keller–Segel system, the aggregation may lead to blow-up solutions. This property has drawn much attention in mathematical analysis. On the other hand, blow-up does not happen in real system. It actually implies that the concentration reaches several orders of magnitude larger and is beyond the range that the model can describe. Thus, many modifications of the classical model have been proposed to eliminate the blow-up. The modified models proved to exhibit spiky solutions but will not blow up. The literature on this topic is huge, of which we only mention the book [25] and some review articles [18, 1].

Compared with the analytical works, the numerical methods for Keller–Segel equations are far from well studied. The main difficulties in constructing suitable numerical schemes are to keep several essential properties of the Keller–Segel equations such as positivity, mass conservation and energy dissipation. Of these three properties, the positivity of the numerical solutions receives the most attention, but provable energy dissipation is rarely acheived. Most existing works focus on the classical Keller–Segel system. Some numerical schemes are developed with the discussion of positive-preserving conditions [16, 9, 15, 14, 13, 7].

[*] This work is supported in part by NSF grants DMS-1620262, DMS-1720442 and AFOSR FA9550-16-1-0102.

[†] Department of Mathematics, Purdue University, USA, Email: shen7@purdue.edu

[‡] Corresponding author. NCMIS & LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing (ICMSEC), Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences, Email: xujie@lsec.cc.ac.cn. The work of this author is partly supported by AMSS grants for Outstanding Youth Project, NCMIS, and ICMSEC director funds.

These schemes depend on particular spatial discretization and usually lead to strict CFL restrictions on the time step. There exist however some unconditionally positivity-preserving schemes. One is a linear finite-volume scheme proposed in [33] (see also [2]) where the upwind technique is utilized. The scheme is restricted to the parabolic–elliptic system, is only applicable to finite-volume spatial discretization, and is only a first-order approximation both in time and space. The other is a recent work which reformulates the equation to arrive at a positive-preserving scheme [23]. The reformulation depends on the particular form of the classical Keller–Segel equations, thus is not easily extended to modified systems. Modified systems are examined in [22], but without effort on keeping the desired properties. In particular, in some modified systems one can show that the concentration is bounded in certain interval. Very recently, a nonlinear finite-volume-based scheme, which adopts the upwind technique, is proposed in [10]. This scheme is able to bound the the solution in the desired interval and keep energy dissipation unconditionally, but it has the same restriction metioned above for the scheme in [33].

In this work, we construct a new class of numerical schemes for the classical and modified systems with a gradient flow structure. Generally, a gradient flow can be written as $\partial\rho/\partial t = \mathcal{G}(\rho) \circ (\delta E/\delta\rho)$, where the dissipation operator $\mathcal{G}$ is non-positive. For Keller–Segel equations, the dissipation operator is nonlinear, taking the form $\nabla\cdot(\eta(\rho)\nabla(\cdot))$ where ρ represents the concentration of the organism. To ensure this operator be non-positive, ρ needs to be constrained in the interval where $\eta(\rho) \geq 0$. A typical case is $\eta(\rho) = \rho$ that the classical Keller–Segel system takes. It leads to Wasserstein gradient flows, where ρ needs to be constrained nonnegative. So, special treatment is needed for preserving its bound in numerical schemes. Therefore, the existing approaches to gradient flows, such as the recently developed SAV approach [29, 30, 27], can not be easily applied.

The key ingredient in our new schemes is to rewrite the term $\Delta\rho$ as $\nabla\cdot(\frac{1}{f''(\rho)}\nabla f'(\rho))$ where $\frac{1}{f''(\rho)} = \eta(\rho)$, then to discretize using this interpretation by treating $f'(\rho)$ implicitly and $f''(\rho)$ explicitly. Then, with a proper treatment of other terms, we can show that the solution to our scheme is the unique minimizer of a strictly convex functional that confines the variable in the interval where $\eta(\rho) > 0$. Therefore, the solution can be efficiently obtained by Newton's iteration. Similar idea is adopted for Cahn-Hilliard equations with logarithmic potential [8] and Poisson–Planck–Nernst equations [28]. In particular, our schemes satisfy unconditionally four desired properties: (i) mass conservation; (ii) unique solvability; (iii) bound preserving; (iv) energy dissipation, without restriction to specific spatial discretizations.

The rest of paper is organized as follows. In the next section, we describe the class of Keller–Segel systems that will be considered in this paper and derive some basic properties which will be used later. Then we construct our numerical schemes in Section 3. The time discretization is proposed first, followed by a discussion on the spatial discretization. We will show that the first-order schemes satisfy the properties (i)-(iv), and the second-order schemes satisfy (i)-(iii). In Section 4, we provide numerical examples to verify our theoretical results, and compare the aggregation in different models. A conclusion is given in section 5.

2 Keller–Segel equations and their regularizations

We start from the simplest system where only one organism and one chemoattractant is present, and consider the following Keller–Segel system in a bounded domain Ω:

$$\frac{\partial\rho}{\partial t} = D\big(\gamma\Delta\rho - \chi\nabla\cdot(\eta(\rho)\nabla\phi)\big), \tag{2.1}$$

$$\tau\frac{\partial\phi}{\partial t} = \mu\Delta\phi - \alpha\phi + \chi\rho, \tag{2.2}$$

with either periodic boundary conditions, or no-flux boundary conditions on ρ and the Neumann boundary conditions on ϕ,

$$\gamma\frac{\partial\rho}{\partial\boldsymbol{n}} - \chi\eta(\rho)\frac{\partial\phi}{\partial\boldsymbol{n}} = 0, \quad \frac{\partial\phi}{\partial\boldsymbol{n}} = 0 \text{ on } \partial\Omega.$$

The boundary conditions on ρ ensure its mass conservation. In the above, the unknowns are ρ, the concentration of the organism, and ϕ, the concentration of the chemoattractant. The first equation describes the motion of the organism governed by the diffusion and the attraction by the chemoattractant. The second equation describes the distribution of the chemoattractant affected by the organism.

The parameters D, γ, χ, τ, μ, α are all positive, of which χ and τ are particularly important: χ is the chemotactic sensitivity, and τ describes how fast the chemoattractant concentration is reacted to the organism concentration. The model is called the parabolic–parabolic system when $\tau > 0$, and parabolic–elliptic system when $\tau = 0$ as an approximation of rapid reaction.

The function $\eta(\rho)$ describes the concentration-dependent mobility. It is the choice of $\eta(\rho)$ that characterizes different models. A basic assumption is that $\eta(s)$ is a smooth function with $\eta(0) = 0$. Moreover, one of the following conditions should hold:

(a) $\eta(s) > 0$ for $s > 0$.

(b) There exists a positive constant M such that $\eta(M) = 0$ and $\eta(s) > 0$ for $0 < s < M$.

Let us denote by $J = (0, +\infty)$ with $\bar{J} = [0, +\infty)$ for $\eta(s)$ with the condition (a), and $J = (0, M)$ with $\bar{J} = [0, M]$ for $\eta(s)$ with the condition (b). Some typical choices of η include:

(i) In the classical Keller–Segel system, the mobility $\eta(\rho) = \rho$;

(ii) Assume that the mobility is bounded, given in [31, 32] by

$$\eta(\rho) = \frac{\rho}{1 + \kappa\rho}, \tag{2.3}$$

with $\kappa > 0$;

(iii) Assume that the organism has a saturation concentration $M > 0$, and the mobility tends to zero when it is near saturation [17, 11],

$$\eta(\rho) = \rho(1 - \rho/M). \tag{2.4}$$

Then, solutions of (2.1)-(2.2) satisfy the following bound preserving properties:

- if $\eta(s)$ satisfies the condition (a), we have $\rho(\boldsymbol{x}, t) \geq 0$ if $\rho(\boldsymbol{x}, 0) \geq 0$;
- if $\eta(s)$ satisfies the condition (b), we have $0 \leq \rho(\boldsymbol{x}, t) \leq M$ if $0 \leq \rho(\boldsymbol{x}, 0) \leq M$.

Indeed, the above properties can be proved by splitting ρ into a positive part ρ_{in} and a negative part ρ_{out} if $\eta(s)$ satisfies the condition (a), or into two parts

$$\rho_{in} = \begin{cases} 0, & \rho \leq 0, \\ \rho, & 0 < \rho < M, \\ M, & \rho \geq M \end{cases} \qquad \rho_{out} = \begin{cases} \rho, & \rho \leq 0, \\ 0, & 0 < \rho < M, \\ \rho - M, & \rho \geq M \end{cases}$$

if $\eta(s)$ satisfies the condition (b), followed by considering an auxiliary problem (cf. for example, [26]),

$$\frac{\partial \rho}{\partial t} = D\big(\gamma \Delta \rho - \chi \nabla \cdot (\eta(\rho_{in}) \nabla \phi)\big).$$

It is easy to notice that $\eta(\rho_{in}) > 0$ for $\rho \in J$, and $\eta(\rho_{in}) = 0$ for $\rho \notin J$. Therefore, by taking the inner product with ρ_{out}, it can be deduced that $\rho_{out} = 0$, which implies that ρ_{in} also satisfies the original equation.

For the classical system ($\eta(\rho) = \rho$), many works on mathematical analysis have been done on the blow-up behavior. In particular, in the two dimensional case, whether the blow-up may appear depends on the total mass about ρ (a conserved quantity) $m = \int_\Omega \rho \mathrm{d}\boldsymbol{x}$. We do not attempt to summarize all existing results, but only state a typical result for the case $\gamma = \mu = 1$, $\tau = \alpha = 0$: there exists a critical value m^* for the total mass such that, for $m < m^*$, the solution exists globally in time, while for $m > m^*$, the solution blows up in finite time (see [3, 5, 6]).

The effect of mobility $\eta(s)$ on the blow-up is also discussed [19]. For η bounded by a power function $\eta(s) \leq cs^{\alpha}$ with sufficiently small $\alpha > 0$, it is guaranteed that the solution exists globally and is uniformly bounded. Note that η is bounded in the choices (2.3) and (2.4), thus for such η no blow-up can occur.

Next, we formulate the system (2.1)–(2.2) as a gradient flow about (ρ, ϕ). We will see that the choice of $\eta(\rho)$ leads to different free energy that also bounds the solution in the interval $\bar{J}$. Let $f''(s) = 1/\eta(s) > 0$ that is defined only in the open interval J. Thus, we could define by integration a strictly convex function $f(s)$ in J. We shall also state some other simple properties on the function f. Notice that $\eta(s)$ is smooth on $\bar{J}$ and $\eta(0) = 0$. So, under the condition (a), there exists a constant c such that $0 < \eta(s) < cs$ for sufficiently small $s > 0$, leading to $f''(s) > 1/cs$. Thus, we have the estimate $f'(s) < C_1 + (1/c)\log s$ for $s > 0$ sufficiently small, hence $\lim_{s\to 0^+} f'(s) = -\infty$. Similarly, under the condition (b), we also have $\lim_{s\to M^-} f'(s) = +\infty$ since $\eta(M) = 0$. Therefore, $f'(s)$ is a strictly monotonely increasing function that is defined only in the open interval J. However, it is possible that $f(s)$ can be continuously extended to the closed interval $\bar{J}$.

Now let us write $\Delta\rho = \nabla \cdot \big(\frac{1}{f''(\rho)}\nabla f'(\rho)\big)$. Consider the free energy

$$F[\rho,\phi] = \int_{\Omega} \big(\gamma f(\rho) - \chi\rho\phi + \frac{\mu}{2}|\nabla\phi|^2 + \frac{\alpha}{2}\phi^2\big)\mathrm{d}\boldsymbol{x}. \tag{2.5}$$

Then, the two equations (2.1)-(2.2) can be rewritten as

$$\frac{\partial\rho}{\partial t} = D\nabla \cdot \big(\frac{1}{f''(\rho)}\nabla(\gamma f'(\rho) - \chi\phi)\big) = D\nabla \cdot \big(\frac{1}{f''(\rho)}\nabla\frac{\delta F}{\delta\rho}\big). \tag{2.6}$$

$$\tau\frac{\partial\phi}{\partial t} = \mu\Delta\phi - \alpha\phi + \chi\rho = -\frac{\delta F}{\delta\phi}. \tag{2.7}$$

Note that in the case of the no-flux boundary conditions, we can rewrite the one on ρ as

$$\gamma\frac{\partial\rho}{\partial\boldsymbol{n}} - \chi\eta(\rho)\frac{\partial\phi}{\partial\boldsymbol{n}} = \eta(\rho)\frac{\partial}{\partial\boldsymbol{n}}(\gamma f'(\rho) - \chi\phi) = \eta(\rho)\frac{\partial}{\partial\boldsymbol{n}}\left(\frac{\delta F}{\delta\rho}\right) = 0.$$

Taking the inner products of (2.6) with $\frac{\delta F}{\delta\rho}$, and of (2.7) with $\frac{\partial\phi}{\partial t}$, and summing up, we deduce the energy law,

$$\frac{\mathrm{d}F[\rho(t),\phi(t)]}{\mathrm{d}t} = -\int \big[D\frac{1}{f''(\rho)}\left(\nabla\frac{\delta F}{\delta\rho}\right)^2 + \tau\left(\frac{\partial\phi}{\partial t}\right)^2\big]\mathrm{d}\boldsymbol{x}. \tag{2.8}$$

For the above energy law to be dissipative, it is necessary that $\eta(\rho) = 1/f''(\rho) \geq 0$, which is indeed true if ρ is initially within the interval where $\eta(s) \geq 0$ thanks to the bound preserving property.

Let us we write down the function f for the three choices (i)–(iii). Note that when $f''(s)$ is known, $f(s)$ may differ by a linear function $as + b$. Since we have mass conservation, the integral of this linear function gives a constant, making no difference.

(i) In the classical system, $f''(s) = 1/s$, we choose $f(s) = s\log s - s$.

(ii) For $f''(s) = (1 + \kappa s)/s = 1/s + \kappa$, we choose $f(s) = s\log s - s + \kappa s^2/2$.

(iii) For $f''(s) = (s(1 - s/M))^{-1}$, we choose $f(s) = s\log s + (M - s)\log(1 - s/M)$.

For (i) and (ii), the function f is defined in $[0, +\infty)$, while f' and f'' are defined in $(0, +\infty)$. For (iii), the function f is defined in $[0, M]$, while f' and f'' are defined in $(0, M)$.

It has been noticed that the lower-boundedness of the free energy is directly related to whether the solution may blow-up. Actually, it has been proved that for the classical system $\eta(\rho) = \rho$, at least for some special cases (see [4]), there exists a critical mass m^* such that the free energy is lower bounded if $m < m^*$, and the solution exists for $t \in [0, +\infty)$. But $m > m^*$ leads to $\inf F = -\infty$. On the other hand, for $\eta(s)$ satisfying the condition (b), we can easily have the estimate

$$F[\rho,\phi] \geq \int_{\Omega} (\gamma f(\rho) - C\rho^2)\mathrm{d}\boldsymbol{x},$$

where the right-hand side is bounded from below by noticing that ρ is bounded in $[0, M]$ and that $f(s)$ is strictly convex.

The model (2.1)-(2.2) can be extended to describe multiple organisms. Below, we write down the equations for two organisms that both response to and generate the chemoattractant with different intensity, mobility, etc., in the gradient flow formulation,

$$\frac{\partial \rho_1}{\partial t} = D_1 \nabla \cdot \big(\frac{1}{f_1''(\rho_1)} \nabla(\gamma_1 f_1'(\rho_1) - \chi_1 \phi)\big), \tag{2.9}$$

$$\frac{\partial \rho_2}{\partial t} = D_2 \nabla \cdot \big(\frac{1}{f_2''(\rho)} \nabla(\gamma_2 f_2'(\rho_2) - \chi_2 \phi)\big), \tag{2.10}$$

$$\tau \frac{\partial \phi}{\partial t} = \mu \Delta \phi - \alpha \phi + \chi_1 \rho_1 + \chi_2 \rho_2. \tag{2.11}$$

3 Numerical methods

We construct in this section numerical schemes satisfying unconditionally the four properties mentioned in the introduction. We discuss parabolic–elliptic ($\tau = 0$) and parabolic–parabolic equations ($\tau > 0$) separately. For parabolic–parabolic equations, our schemes are coupled between ρ and ϕ, but for parabolic–elliptic equations, we are able to construct decoupled numerical schemes in which ρ and ϕ can be solved sequentially. We recall that the parabolic–elliptic equations are significant, since they give a good approximation when the chemoattractant diffuses much faster than organisms.

3.1 Parabolic–elliptic equations

3.1.1 First-order scheme

We consider first the time discretization. Let (ρ^n, ϕ^n) be the approximation of (ρ, ϕ) at t^n with $\rho^n \in J$. We solve for (ρ^{n+1}, ϕ^{n+1}) from

$$\frac{\rho^{n+1} - \rho^n}{\delta t} = D \nabla \cdot \Big[\frac{1}{f''(\rho^n)} \nabla \Big(\gamma f'(\rho^{n+1}) - \chi \phi^n \Big) \Big], \tag{3.1}$$

$$0 = \mu \Delta \phi^{n+1} - \alpha \phi^{n+1} + \chi \rho^{n+1}. \tag{3.2}$$

Note that we can solve ρ^{n+1} first from (3.1), then ϕ^{n+1} can be determined from (3.2). In the first equation, the mobility is treated explicitly so $f''(\rho^n)$ is well-defined since $\rho^n \in J$.

Recall that $f'(s)$ is defined only on the interval J, so we treat it implicitly to constrain $\rho^{n+1} \in J$. The price we pay is that (3.1) is a nonlinear equation for ρ^{n+1}, but as we shall prove below, it is equivalent to a convex minimization problem so that it is still efficient and easy to implement.

Theorem 3.1. *Assume that the initial value $\rho^0 \in J$. Any solution to the above scheme satisfies:*

1. *Mass conservation:*
$$\int_\Omega \rho^{n+1} \mathrm{d}\boldsymbol{x} = \int_\Omega \rho^n \mathrm{d}\boldsymbol{x}.$$

2. *Bound preserving: $\rho^{n+1}(\boldsymbol{x}) \in J$.*

3. *Energy dissipation:*
$$E^{n+1} - E^n \leq -\delta t \int \frac{D}{f''(\rho^n)} |\nabla \big(\gamma f'(\rho^{n+1}) - \chi \phi^n \big)|^2 \mathrm{d}\boldsymbol{x},$$
where $E^n = \int \big(\gamma f(\rho^n) - \chi \rho^n \phi^n + \frac{\mu}{2} |\nabla \phi^n|^2 + \frac{\alpha}{2} (\phi^n)^2 \big) \mathrm{d}\boldsymbol{x}$.

Proof. The mass conservation is obtained by integrating the first equation and applying the boundary conditions.

Bound preserving is due to the presence of $f'(\rho^{n+1})$, by noticing that f' is only defined in the interval J.

For the energy dissipation, we take the inner product of (3.2) with ϕ^{n+1} and change the superscript to n to arrive at

$$\mu\|\nabla\phi^n\|^2 + \alpha\|\phi^n\|^2 = \chi(\rho^n, \phi^n).$$

Using the above identity, we rewrite the energy E^n as

$$E^n = \int \gamma f(\rho^n) - \frac{\mu}{2}|\nabla\phi^n|^2 - \frac{\alpha}{2}(\phi^n)^2 \mathrm{d}\boldsymbol{x}.$$

Now we take the inner product of (3.1) with $\nu^{n+1} = \delta t(\gamma f'(\rho^{n+1}) - \chi\phi^n)$, yielding

$$\int_\Omega (\rho^{n+1} - \rho^n)(\gamma f'(\rho^{n+1}) - \chi\phi^n)\mathrm{d}\boldsymbol{x} = -\delta t \int_\Omega \frac{D}{f''(\rho^n)}|\nabla\nu^{n+1}|^2 \mathrm{d}\boldsymbol{x}.$$

we handle the left-hand side as follows. Using $f''(s) > 0$ for $s \in J$, we deduce that

$$(\rho^{n+1} - \rho^n, f'(\rho^{n+1})) = (f(\rho^{n+1}) - f(\rho^n) + \frac{f''(\xi)}{2}(\rho^{n+1} - \rho^n)^2, 1) \geq (f(\rho^{n+1}) - f(\rho^n), 1). \tag{3.3}$$

On the other hand, using (3.2) and the equality

$$2a(a-b) = a^2 - b^2 + (a-b)^2, \tag{3.4}$$

we find

$$\begin{aligned}(\rho^{n+1} - \rho^n, -\chi\phi^n) =& (-\mu\Delta(\phi^{n+1} - \phi^n) + \alpha(\phi^{n+1} - \phi^n), -\phi^n)\\ =& \mu(\nabla(\phi^{n+1} - \phi^n), -\nabla\phi^n) + \alpha(\phi^{n+1} - \phi^n, -\phi^n)\\ =& \frac{\mu}{2}\Big(- \|\nabla\phi^{n+1}\|^2 + \|\nabla\phi^n\|^2 + \|\nabla\phi^{n+1} - \nabla\phi^n\|^2\Big)\\ &+ \frac{\alpha}{2}\Big(- \|\phi^{n+1}\|^2 + \|\phi^n\|^2 + \|\phi^{n+1} - \phi^n\|^2\Big).\end{aligned}$$

Combining the above equations, we arrive at the energy dissipation. □

It is clear that if ρ^{n+1} is known, there exists a unique solution ϕ^{n+1} for (3.2). It remains to examine whether a solution exists for (3.1). We provide below (for time before blow-up if there will be a blow-up) a formal derivation in the spatially continuous case by formulating the scheme as the minimizer of a strictly convex functional.

Let the linear operator $\mathcal{L}^n$ be defined such that for any $g(\boldsymbol{x})$ satisfying $\int g\mathrm{d}\boldsymbol{x} = 0$, its image $\mathcal{L}^n g = \psi$ is the solution to the following elliptic equation under the periodic or Neumann boundary conditions,

$$-\nabla \cdot (\frac{1}{f''(\rho^n)}\nabla\psi) = g, \quad \int \psi \mathrm{d}\boldsymbol{x} = 0. \tag{3.5}$$

Consider the functional

$$F^n[\rho; \rho^n, \phi^n] = \frac{1}{2\delta t}\big(\rho - \rho^n, \mathcal{L}^n(\rho - \rho^n)\big) + \gamma(f(\rho), 1) - \chi(\rho, \phi^n). \tag{3.6}$$

The first two terms on the right-hand side are strictly convex and bounded from below. The third term is linear, and bounded if we assume that $\phi^n \in L^\infty$. In this sense, the whole functional is strictly convex and lower-bounded. One can check that under the mass conservation constraint $\int \rho \mathrm{d}\boldsymbol{x} = \int \rho^n \mathrm{d}\boldsymbol{x}$, the Euler–Lagrange equation is equivalent to (3.1). Since the functional possesses a term with $f(\rho)$, so it is only defined for ρ taking the value in the closed interval $\bar{J}$. Furthermore, the minimizer cannot take the value on the endpoint of $\bar{J}$ because f' goes to infinity. Thus, the functional has a unique minimizer with $\rho \in J$.

Below, we convert the above formal derivation into a rigorous derivation by considering a fully discretized scheme with a Galerkin type discretization in space. More precisely, given a finite set of points $Z = \{\boldsymbol{z} \in \bar{\Omega}\}$, we define a discrete inner product

$$[u, v] = \sum_{\boldsymbol{z} \in Z} \beta_{\boldsymbol{z}} u(\boldsymbol{z}) v(\boldsymbol{z}),$$

with positive weights $\beta_{\boldsymbol{z}} > 0$. The inner product can be based on a finite element, spectral or even finite difference method. In finite element methods, the sum is calculated first on each element, then throughout all the elements, i.e. as $\sum_{K \subset \mathcal{T}} \sum_{z \in Z(K)}$. We denote the corresponding finite dimensional approximation space by X_N. For each $\boldsymbol{z} \in Z$, we assume that there exists a unique Lagrangian basis function $\psi_{\boldsymbol{z}}$ in X_N such that $\psi_{\boldsymbol{z}}(\boldsymbol{z'}) = \delta_{\boldsymbol{z}\boldsymbol{z'}}$ for any $\boldsymbol{z'} \in Z$, so they form a basis of X_N. Under this assumption, we can define an interpolation operator I_N as

$$(I_N g)(\boldsymbol{x}) = \sum_{\boldsymbol{z} \in Z} g(\boldsymbol{z}) \psi_{\boldsymbol{z}}(\boldsymbol{x}).$$

Then, our Galerkin method for the first-order scheme (3.1)-(3.2) is: to find $(\rho^{n+1}, \phi^{n+1}) \in X_N \times X_N$ such that

$$\left[\frac{\rho^{n+1} - \rho^n}{\delta t}, v\right] = -\left[\frac{D}{f''(\rho^n)} \nabla\Big(I_N\big(\gamma f'(\rho^{n+1}) - \chi\phi^n\big)\Big), \nabla v\right], \quad v \in X_N, \tag{3.7}$$

$$0 = -\mu(\nabla\phi^{n+1}, \nabla w) - \alpha(\phi^{n+1}, w) + \chi[\rho^{n+1}, w], \quad w \in X_N. \tag{3.8}$$

Here, the $(\cdot, \cdot)$ represents the usual L^2 inner product, and $[\cdot, \cdot]$ is the discrete inner product defined above.

Theorem 3.2. *The fully discretized scheme* (3.7)–(3.8) *satisfies the following properties:*

1. *Mass conservation:*
$$[\rho^{n+1}, 1] = [\rho^n, 1].$$
2. *Unique solvability: the scheme possesses a unique solution* $(\rho^{n+1}, \phi^{n+1}) \in X_N \times X_N$.
3. *Bound preserving: if* $\rho^n(\boldsymbol{z}) \in J$ *for all* $\boldsymbol{z} \in Z$, *then* $\rho^{n+1}(\boldsymbol{z}) \in J$ *for all* $\boldsymbol{z} \in Z$.
4. *Energy dissipation:*
$$\tilde{E}^{n+1} - \tilde{E}^n \leq -\delta t \left[\frac{D}{f''(\rho^n)} \nabla\nu^{n+1}, \nabla\nu^{n+1}\right],$$
where $\nu^{n+1} = I_N\big(\gamma f'(\rho^{n+1}) - \chi\phi^n\big)$, *and the discrete energy is given by*
$$E^n = [\gamma f(\rho^n), 1] - \chi[\rho^n, \phi^n] + \frac{\mu}{2}\|\nabla\phi^n\|^2 + \frac{\alpha}{2}\|\phi^n\|^2.$$

Proof. The mass conservation can be derived by taking $v = 1$ in (3.7).

Next, we prove the unique solvability for the first equation in the range $\rho^{n+1}(\boldsymbol{z}) \in J$ for all $\boldsymbol{z} \in Z$. Denote by $\tilde{\rho}^n$ and $\tilde{\phi}^n$ the two vectors $\big(\rho^n(\boldsymbol{z}), \boldsymbol{z} \in Z\big)$ and $\big(\phi^n(\boldsymbol{z}), \boldsymbol{z} \in Z\big)$. Define two matrices as

$$A^n = \left[\frac{D}{f''(\rho^n)} \nabla\psi_{\boldsymbol{z}}, \nabla\psi_{\boldsymbol{z'}}\right], B = [\psi_{\boldsymbol{z}}, \psi_{\boldsymbol{z'}}]. \tag{3.9}$$

The matrix B is diagonal with positive entries. If $\rho^n(\boldsymbol{z}) \in J$ for $\boldsymbol{z} \in Z$, the matrix A^n is symmetric positive semi-definite. Furthermore, $A^n\bar{x} = 0$ if and only if each component of $\bar{x}$ is equal. Thus, we can write down the pseudo-inverse $(A^n)^*$ by the eigen-decomposition. More precisely, assume $A^n = T^t\Lambda T$ where $\Lambda = \mathrm{diag}(0, \lambda_2, \ldots, \lambda_N)$, then $(A^n)^* = T^t\mathrm{diag}(0, \lambda_2^{-1}, \ldots, \lambda_N^{-1})T$. Then, we can write (3.7) in the matrix–vector form as

$$\frac{1}{\delta t} B(\tilde{\rho}^{n+1} - \tilde{\rho}^n) = -A^n\big(\gamma f'(\tilde{\rho}^{n+1}) - \chi\tilde{\phi}^n\big). \tag{3.10}$$

Multiplying the above from left by $B(A^n)^*$, noticing the null space of A^n, we obtain

$$\frac{1}{\delta t}B(A^n)^*B(\tilde{\rho}^{n+1}-\tilde{\rho}^n)+\gamma Bf'(\tilde{\rho}^{n+1})-\chi B\tilde{\phi}^n=\lambda B\mathbf{1}, \tag{3.11}$$

where $\mathbf{1}$ represents the vector with each component equal to one. It is easy to see that the above equation is the Euler–Lagrange equation of the function

$$G[\tilde{\rho}^{n+1}]=\frac{1}{2\delta t}(\tilde{\rho}^{n+1}-\tilde{\rho}^n)^tB(A^n)^*B(\tilde{\rho}^{n+1}-\tilde{\rho}^n)+\gamma\mathbf{1}^tBf(\tilde{\rho}^{n+1})-\chi(\rho^{n+1})^tB\tilde{\phi}^n,$$

under the mass conservation constraint $\mathbf{1}^tB\tilde{\rho}^{n+1}=\mathbf{1}^tB\tilde{\rho}^n$. Note that the first term on the right-hand side has a symmetric positive semi-definite coefficient matrix $B(A^n)^*B$. The second term is strictly convex about $\tilde{\rho}^{n+1}$ because f is strictly convex and B is diagonal with positive entries. The third term is linear and bounded because of mass conservation. Thus, the function $F[\tilde{\rho}^{n+1}]$ is strictly convex, lower-bounded in the domain

$$\{\tilde{\rho}^{n+1}:\tilde{\rho}^{n+1}(\boldsymbol{z})\in\bar{J},\ \mathbf{1}^tB\tilde{\rho}^{n+1}=\mathbf{1}^tB\tilde{\rho}^n\}.$$

Hence, there exists a unique minimizer in this domain. It remains to eliminate the possibility of the minimizer located on the boundary of the domain. In other words, we need to prove that the minimizer $\tilde{\rho}_0$ cannot have a component $\tilde{\rho}^{n+1}(\boldsymbol{z})$ taking the endpoint of the interval $\bar{J}$ (0 or M). We prove this by contradiction below.

Suppose the minimizer is such that $\tilde{\rho}_0(\boldsymbol{z})=0$ (for the case $J=(0,M)$, $\tilde{\rho}_0(\boldsymbol{z})=M$ leads to contradiction in the same way) for some $\boldsymbol{z}$. This can only occur when $f(0)$ is defined (finite), which we assume in the following. Let us choose another $\boldsymbol{z'}$ such that $\tilde{\rho}_0(\boldsymbol{z'})>0$. Consider another vector $\tilde{\rho}_1=\tilde{\rho}_0+\beta_{\boldsymbol{z'}}\epsilon\boldsymbol{e}_{\boldsymbol{z}}-\beta_{\boldsymbol{z}}\epsilon\boldsymbol{e}_{\boldsymbol{z'}}$. Here, we use $\boldsymbol{e}_{\boldsymbol{z}}$ to represent the vector with the entry one on the component $\boldsymbol{z}$ and zero for other components. We shall show that for ϵ small enough, we have $G[\tilde{\rho}_1]<G[\tilde{\rho}_0]$ which is a contradiction. To this end, we split the function G into two parts,

$$G_1=\frac{1}{2\delta t}(\tilde{\rho}^{n+1}-\tilde{\rho}^n)^tB(A^n)^*B(\tilde{\rho}^{n+1}-\tilde{\rho}^n)-\chi(\rho^{n+1})^tB\tilde{\phi}^n,$$

and

$$G_2=\gamma\mathbf{1}^tBf(\tilde{\rho}^{n+1}).$$

Since G_1 is a quadratic function, there exists a constant A_1 such that for sufficiently small ϵ,

$$|G_1[\tilde{\rho}_1]-G_1[\tilde{\rho}_0]|<A_1\epsilon.$$

On the other hand, denoting $a=\tilde{\rho}_0(\boldsymbol{z'})>0$, we can compute that

$$G_2[\tilde{\rho}_1]-G_2[\tilde{\rho}_0]=\beta_{\boldsymbol{z}}\big(f(\beta_{\boldsymbol{z'}}\epsilon)-f(0)\big)+\beta_{\boldsymbol{z'}}\big(f(a-\beta_{\boldsymbol{z}}\epsilon)-f(a)\big).$$

Since $f'(a)>-\infty$, we have another constant A_2, such that for sufficiently small ϵ,

$$f(a-\beta_{\boldsymbol{z}}\epsilon)-f(a)<A_2\epsilon.$$

Now we will make use of $\lim_{s\to0^+}f'(s)=-\infty$. It implies that for any $A>0$, there exists sufficiently small ϵ, such that

$$f(\beta_{\boldsymbol{z'}}\epsilon)-f(0)<-A\epsilon.$$

Choose $\beta_{\boldsymbol{z}}A>A_1+\beta_{\boldsymbol{z'}}A_2$ and ϵ small enough, we find that $G[\tilde{\rho}_1]-G[\tilde{\rho}_0]<0$.

With the unique solution ρ^{n+1} from (3.7), we can uniquely determine ϕ^{n+1} from (3.8) which is the Galerkin discretization of a linear elliptic equation.

It remains to prove the energy dissipation. We take $v=\delta t\nu^{n+1}$ in (3.7), yielding

$$\begin{aligned}-\delta t\left[\frac{D}{f''(\rho^n)}\nabla\nu^{n+1},\nabla\nu^{n+1}\right]=&[\rho^{n+1}-\rho^n,I_N(\gamma f'(\rho^{n+1})-\chi\phi^n)]\\=&\gamma[\rho^{n+1}-\rho^n,f'(\rho^{n+1})]-[\chi(\rho^{n+1}-\rho^n),\phi^n].\end{aligned} \tag{3.12}$$

Then, taking $w = \phi^{n+1}$ in (3.8) and changing the superscript to n, we obtain

$$\chi[\rho^n, \phi^n] = \mu\|\nabla\phi^n\|^2 + \alpha\|\phi^n\|^2.$$

Taking $w = \phi^n$ in (3.7) and combining with the above, we find

$$\begin{aligned}
&2[\chi(\rho^{n+1} - \rho^n), -\phi^n] \\
=& -2\mu\big(\nabla(\phi^{n+1} - \phi^n), \nabla\phi^n\big) - 2\alpha\big((\phi^{n+1} - \phi^n), \phi^n\big) \\
=&\mu(-\|\nabla\phi^{n+1}\|^2 + \|\nabla\phi^n\|^2 + \|\nabla(\phi^{n+1} - \phi^n)\|^2) \\
&- \alpha(-\|\phi^{n+1}\|^2 + \|\phi^n\|^2 + \|\phi^{n+1} - \phi^n\|^2).
\end{aligned}$$

We can then we derive the energy dissipation by combining the above with (3.12) and (3.3). □

Remark 3.3. *The nonlinear equation* (3.7) *can be efficiently solved by using a Newton's iteration, with damping on step size to restrain the search in the interval J. From its matrix form* (3.10)*, we need to solve, for each Newton's step, a linear system in the form*

$$(B + A^n\Lambda)x = b,$$

where B and Λ are diagonal matrices with positive elements, and A^n is symmetric non-negative. We shall rewrite it as

$$\Lambda^{-1/2}(\Lambda^{1/2}B\Lambda^{-1/2} + \Lambda^{1/2}A^n\Lambda^{1/2})\Lambda^{1/2}x = b.$$

Thus, we can solve two diagonal systems and one symmetric positive definite system. For the symmetric positive definite linear system, we use the preconditioned conjugate gradient method. The choice of preconditioner is dependent on particular spatial discretization. For the Fourier spectral method that we will use in this paper, the preconditioner can be chosen as discretized from PDE with constant coefficients. In particular, when using Fourier spectral method, B is a multiple of the identity matrix I, and we could substitute Λ with a multiple of I. For A^n, we substitute the variable coefficients $D/f''(\rho^n)$ in (3.9) *with a constant. In this way, we arrive at a preconditioner that can be implemented by FFT.*

Remark 3.4. *In the fully discretized scheme, the mass conservation and bound preserving together imply that the l^1 norm is bounded for the solutions. Thus, even if there is a blow-up time in PDE, we will not see an actual blow-up in numerical solutions. Instead, as we will present in numerical examples, the mass accumulates on a very few discrete points.*

Remark 3.5. *Although we only discuss the Galerkin type spatial discretization in the theorem above, the results still hold for finite difference (finite volume) discretizations if the summation by parts is valid. We refer to [28] where a proof is provided.*

3.1.2 Second-order scheme

We can also construct second-order schemes with similar properties. For example, a fully discretized scheme based on the second-order BDF is: to find $\rho^{n+1}, \phi^{n+1} \in X_N$, such that for any $v, w \in X_N$,

$$\left[\frac{3\rho^{n+1} - 4\rho^n + \rho^{n-1}}{2\delta t}, v\right] = -\left[D\bar{a}^{n+1}\nabla I_N\Big(\gamma f'(\rho^{n+1}) - \chi(2\phi^n - \phi^{n-1})\Big), \nabla v\right], \tag{3.13}$$

$$0 = -\mu(\nabla\phi^{n+1}, \nabla w) - \alpha(\phi^{n+1}, w) + [\chi\rho^{n+1}, w]. \tag{3.14}$$

In the above, $\bar{a}^{n+1}$ is an $O(\delta t^2)$ explicit approximation of $1/f''(\rho(t^{n+1}))$. To ensure $\bar{a}^{n+1} > 0$, we choose

$$\bar{a}^{n+1} = \begin{cases} \frac{1}{2f''(\rho^n) - f''(\rho^{n-1})}, & \text{if } f''(\rho^n) \geq f''(\rho^{n-1}); \\ \frac{2}{f''(\rho^n)} - \frac{1}{f''(\rho^{n-1})}, & \text{if } f''(\rho^n) < f''(\rho^{n-1}). \end{cases} \tag{3.15}$$

The scheme is also decoupled, i.e. one can solve ρ^{n+1} first, followed by solving ϕ^{n+1}.

We could follow the proof of last theorem to establish the following:

Theorem 3.6. *The scheme* (3.13)-(3.14) *satisfies the following properties:*

1. *Mass conservation:*
$$[\rho^{n+1}, 1] = [\rho^n, 1].$$
2. *Unique solvability: the scheme possesses a unique solution* $(\rho^{n+1}, \phi^{n+1}) \in X_N \times X_N$.
3. *Bound preserving: if* $\rho^n(\boldsymbol{z}) \in J$ *for all* $\boldsymbol{z} \in Z$, *then* $\rho^{n+1}(\boldsymbol{z}) \in J$ *for all* $\boldsymbol{z} \in Z$.

Proof. With $\bar{a}^{n+1} > 0$ explicitly determined, we could define the symmetric positive semi-definite matrix A^n for the second-order scheme as in (3.9), by replacing $D/f''(\rho^n)$ with $\bar{a}^{n+1}$. Then, we could similarly write (3.13) in the matrix-vector form as follows,

$$\frac{3}{2\delta t}B(\tilde{\rho}^{n+1} - \tilde{b}) = A^n(\gamma f'(\tilde{\rho}^{n+1}) - \chi\tilde{c}), \tag{3.16}$$

where $\tilde{b} = (4\tilde{\rho}^n - \tilde{\rho}^{n-1})/3$ and $\tilde{c} = 2\tilde{\phi}^n - \tilde{\phi}^{n-1}$. The rest of the proof is the same as Theorem 3.2: we just replace $\tilde{\rho}^n$ with $\tilde{b}$, and $\tilde{\phi}^n$ with $\tilde{c}$ in (3.10) and follow the same steps afterwards. The components of the vector $\tilde{b}$ might not fall in the interval J, but it does not affect the definition of the strictly convex function $G[\tilde{\rho}^{n+1}]$ in the domain

$$\{\tilde{\rho}^{n+1} : \tilde{\rho}^{n+1}(\boldsymbol{z}) \in \bar{J},\ \mathbf{1}^t B\tilde{\rho}^{n+1} = \mathbf{1}^t B\tilde{b}\}.$$

□

Unfortunately, we are unable to prove the energy dissipation due to the lack of inequality similar to (3.3) for the second-order BDF.

3.1.3 Two species

We can construct similar schemes for the two-species system (2.9)-(2.11). For example, a fully discrete first order scheme is: to find ρ_1^{n+1}, ρ_2^{n+1}, $\phi^{n+1} \in X_N$, such that for any $v_1, v_2, w \in X_N$,

$$\left[\frac{\rho_1^{n+1} - \rho_1^n}{\delta t}, v_1\right] = -\left[\frac{D_1}{f_1''(\rho_1^n)}\nabla I_N\Big(\gamma_1 f_1'(\rho_1^{n+1}) - \chi_1\phi^n\Big), \nabla v_1\right], \tag{3.17}$$

$$\left[\frac{\rho_2^{n+1} - \rho_2^n}{\delta t}, v_2\right] = -\left[\frac{D_2}{f_2''(\rho^n)}\nabla I_N\Big(\gamma_2 f_2'(\rho_2^{n+1}) - \chi_2\phi^n\Big), \nabla v_2\right], \tag{3.18}$$

$$0 = -\mu(\nabla\phi^{n+1}, \nabla w) - \alpha(\phi^{n+1}, w) + [\chi_1\rho_1^{n+1} + \chi_2\rho_2^{n+1}, w]. \tag{3.19}$$

When applying the scheme to the above equations, we notice three equations are decoupled. One can solve ρ_1^{n+1} and ρ_2^{n+1} separately, then solve ϕ^{n+1}.

Theorem 3.7. *The above fully discretized scheme satisfies:*

1. *Mass conservation:*
$$[\rho_1^{n+1}, 1] = [\rho_1^n, 1], \quad [\rho_2^{n+1}, 1] = [\rho_2^n, 1].$$
2. *Unique solvability: the scheme possesses a unique solution* $(\rho_1^{n+1}, \rho_2^{n+1}, \phi^{n+1}) \in X_N \times X_N \times X_N$.
3. *Bound preserving: if* $\rho_i^n(\boldsymbol{z}) \in J_i$ *for all* $\boldsymbol{z} \in Z$, *then* $\rho_i^{n+1}(\boldsymbol{z}) \in J_i$ *for all* $\boldsymbol{z} \in Z$.
4. *Energy dissipation:*
$$\tilde{E}^{n+1} - \tilde{E}^n \le -\delta t\left(\left[\frac{D_1}{f_1''(\rho_1^n)}\nabla\nu_1^{n+1}, \nabla\nu_1^{n+1}\right] + \left[\frac{D_2}{f_2''(\rho_2^n)}\nabla\nu_2^{n+1}, \nabla\nu_2^{n+1}\right]\right),$$
where $\nu_i^{n+1} = I_N\big(\gamma_i f_i'(\rho_i^{n+1}) - \chi_i\phi^n\big)$, *and the discrete energy is given by*
$$E^n = [\gamma_1 f_1(\rho_1^n), 1] - \chi_1[\rho_1^n, \phi^n] + [\gamma_2 f_2(\rho_2^n), 1] - \chi_2[\rho_2^n, \phi^n] + \frac{\mu}{2}\|\nabla\phi^n\|^2 + \frac{\alpha}{2}\|\phi^n\|^2.$$

The proof follows the same arguments as in the case of one-species. The only point to be noticed is that since the scheme is decoupled, we could prove the unique solvability in J for ρ_1^{n+1} and ρ_2^{n+1} respectively.

3.2 Parabolic–parabolic equations

The discretization of parabolic–parabolic system is slightly different. We write the coupling term in the free energy as the difference of two squared terms, $-\rho\phi = \frac{1}{4}(\phi-\rho)^2 - \frac{1}{4}(\phi+\rho)^2$, and construct the following first-order scheme using an idea of convex splitting [12]:

$$\left[\frac{\rho^{n+1}-\rho^n}{\delta t}, v\right] = -\Big[\frac{D}{f''(\rho^n)}\nabla I_N\Big(\gamma f'(\rho^{n+1}) + \frac{\chi}{2}(\rho^{n+1}-\phi^{n+1}) - \frac{\chi}{2}(\rho^n+\phi^n)\Big), \nabla v\Big], \tag{3.20}$$

$$\left(\tau\frac{\phi^{n+1}-\phi^n}{\delta t}, w\right) = -\mu\Big(\nabla\phi^{n+1}, \nabla w\Big) - \alpha\Big(\phi^{n+1}, w\Big) - \Big[\frac{\chi}{2}(\phi^{n+1}-\rho^{n+1}) - \frac{\chi}{2}(\rho^n+\phi^n), w\Big]. \tag{3.21}$$

Theorem 3.8. *The scheme* (3.20)-(3.21) *for parabolic–parabolic equations satisfies the following properties:*

1. *Mass conservation:*
$$[\rho^{n+1}, 1] = [\rho^n, 1].$$

2. *Unique solvability: the scheme possesses a unique solution $(\rho^{n+1}, \phi^{n+1}) \in X_N \times X_N$*

3. *Bound preserving: if $\rho^n(\boldsymbol{z}) \in J$ for all $\boldsymbol{z} \in Z$, then $\rho^{n+1}(\boldsymbol{z}) \in J$ for all $\boldsymbol{z} \in Z$.*

4. *Energy dissipation:*
$$\tilde{E}^{n+1} - \tilde{E}^n \le -\delta t\left[\frac{D}{f''(\rho^n)}\nabla\nu^{n+1}, \nabla\nu^{n+1}\right] - \frac{\tau}{\delta t}\|\phi^{n+1}-\phi^n\|^2,$$
where $\nu^{n+1} = I_N\Big(\gamma f'(\rho^{n+1}) + \frac{\chi}{2}(\rho^{n+1}-\phi^{n+1}) - \frac{\chi}{2}(\rho^n+\phi^n)\Big)$, and the discrete energy is given by
$$E^n = \gamma[f(\rho^n), 1] - \chi[\rho^n, \phi^n] + \frac{\mu}{2}\|\nabla\phi^n\|^2 + \frac{\alpha}{2}\|\phi^n\|^2.$$

Proof. The proof of the first three properties is essentially the same as in the proof of Theorem 3.2, so we only describe the strictly convex function related to the scheme. Define

$$A_0(\boldsymbol{z}, \boldsymbol{z'}) = (\psi_{\boldsymbol{z}}, \psi_{\boldsymbol{z}}), \quad A_1(\boldsymbol{z}, \boldsymbol{z'}) = (\nabla\psi_{\boldsymbol{z}}, \nabla\psi_{\boldsymbol{z}}).$$

Together with the definition of A^n and B in (3.9), we can rewrite the scheme as

$$\begin{aligned}\frac{1}{\delta t}B(\tilde{\rho}^{n+1}-\tilde{\rho}^n) &= -A^n\Big(\gamma f'(\tilde{\rho}^{n+1}) + \frac{\chi}{2}(\tilde{\rho}^{n+1}-\tilde{\phi}^{n+1}) - \frac{\chi}{2}(\tilde{\rho}^n+\tilde{\phi}^n)\Big),\\ \frac{\tau}{\delta t}A_0(\tilde{\phi}^{n+1}-\tilde{\phi}^n) &= -\mu A_1\tilde{\phi}^{n+1} - \alpha A_0\tilde{\phi}^{n+1} - B\Big(\frac{\chi}{2}(\tilde{\phi}^{n+1}-\tilde{\rho}^{n+1}) - \frac{\chi}{2}(\tilde{\rho}^n+\tilde{\phi}^n)\Big).\end{aligned}$$

It can be checked that the above are the Euler–Lagrange equations of the strictly convex function

$$\begin{aligned}G[\tilde{\rho}, \tilde{\phi}; \tilde{\rho}^n, \tilde{\phi}^n] =&\frac{1}{2\delta t}(\tilde{\rho}-\tilde{\rho}^n)^t B(A^n)^* B(\tilde{\rho}-\tilde{\rho}^n) + \frac{\tau}{2\delta t}(\tilde{\phi}-\tilde{\phi}^n)^t A_0(\tilde{\phi}-\tilde{\phi}^n) + \gamma\mathbf{1}^t Bf(\tilde{\rho})\\ &+ \frac{\mu}{2}\tilde{\phi}^t A_1\tilde{\phi} + \frac{\alpha}{2}\tilde{\phi}^t A_0\tilde{\phi} + \frac{\chi}{4}(\tilde{\phi}-\tilde{\rho})^t B(\tilde{\phi}-\tilde{\rho})\\ &- \frac{\chi}{2}(\tilde{\phi}+\tilde{\rho})^t B(\tilde{\phi}^n+\tilde{\rho}^n),\end{aligned} \tag{3.22}$$

under the mass conservation constraint $\mathbf{1}^t B\tilde{\rho}^{n+1} = \mathbf{1}^t B\tilde{\rho}^n$. Then, we can eliminate the possibility of the minimizer taking the values 0 or M as in the proof of Theorem 3.2. Thus, the minimizer of the function G is the unique solution of the scheme.

For energy dissipation, we take $v = \delta t\nu^{n+1} = \delta t I_N\Big(\gamma f'(\rho^{n+1}) + \frac{\chi}{2}(\rho^{n+1}-\phi^{n+1}) - \frac{\chi}{2}(\rho^n+\phi^n)\Big)$ in (3.20), and $w = \phi^{n+1}-\phi^n$ in in (3.21), summing up the two equations to obtain

$$\Big[\rho^{n+1}-\rho^n, \gamma f'(\rho^{n+1}) + \frac{\chi}{2}(\rho^{n+1}-\phi^{n+1}) - \frac{\chi}{2}(\rho^n+\phi^n)\Big]$$

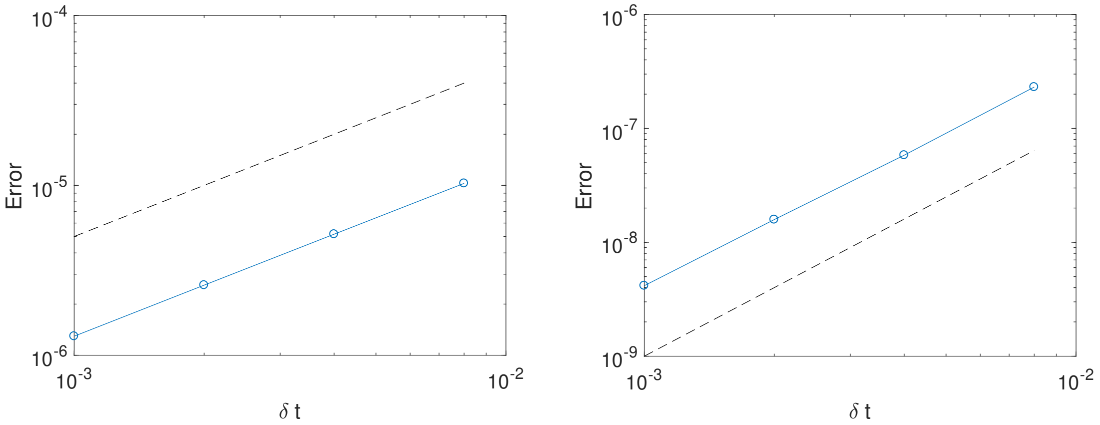


Fig. 1: Accuracy of the first-order (left) and second-order scheme (right). The dashed lines are reference for first-order and second-order convergence.

$$
\begin{aligned}
&+\mu\Big(\nabla\phi^{n+1},\nabla(\phi^{n+1}-\phi^n)\Big)+\alpha\Big(\phi^{n+1},(\phi^{n+1}-\phi^n)\Big)\\
&+\Big[\frac{\chi}{2}(\phi^{n+1}-\rho^{n+1})-\frac{\chi}{2}(\rho^n+\phi^n),\phi^{n+1}-\phi^n\Big]\\
&\qquad\qquad=-\delta t\Big[\frac{D}{f''(\rho^n)}\nabla\nu^{n+1},\nabla\nu^{n+1}\Big]-\frac{\tau}{\delta t}\|\phi^{n+1}-\phi^n\|^2.
\end{aligned}
$$

We rewrite the left-hand side as

$$
\begin{aligned}
&\frac{\chi}{2}\Big[(\rho^{n+1}-\phi^{n+1})-(\rho^n-\phi^n),\rho^{n+1}-\phi^{n+1}\Big]-\frac{\chi}{2}\Big[(\rho^{n+1}+\phi^{n+1})-(\rho^n+\phi^n),\rho^n+\phi^n\Big]\\
&+\mu\Big(\nabla\phi^{n+1},\nabla(\phi^{n+1}-\phi^n)\Big)+\alpha\Big(\phi^{n+1},(\phi^{n+1}-\phi^n)\Big)\\
&+[\rho^{n+1}-\rho^n,\gamma f'(\rho^{n+1})].
\end{aligned}
$$

Then, we can combine the above with the equalities (3.3) and (3.4) to arrive at the energy dissipation. □

4 Numerical results

We present in this section some numerical examples to validate our schemes. We will investigate the three choices of mobility (i)–(iii) stated in Section 2. We name the system with the mobility (ii) as bounded-mobility system, and the system with the mobility (iii) as saturation-concentration system. For all examples, the domain is chosen as $[0,L)^2$ where $L=2\pi$, and we adopt the periodic boundary conditions. The space is discretized using the Fourier spectral method with $N=64$ in each direction. We present several examples for one species, followed by an example for two species. If not specified separately, we fix the diffusion constant $D=\gamma=\mu=1$, and $\alpha=0.1$. The chemotactic sensitivity γ will be varied.

4.1 Accuracy and efficiency

First, we check the accuracy of the schemes. We use the first-order and second-order schemes for one-species parabolic-elliptic ($\tau=0$) saturation-concentration system. The initial condition is given by

$$
\rho(x,y,0)=2\exp\left(-\frac{(x-L/2)^2+(y-L/2)^2}{4}\right). \tag{4.1}
$$

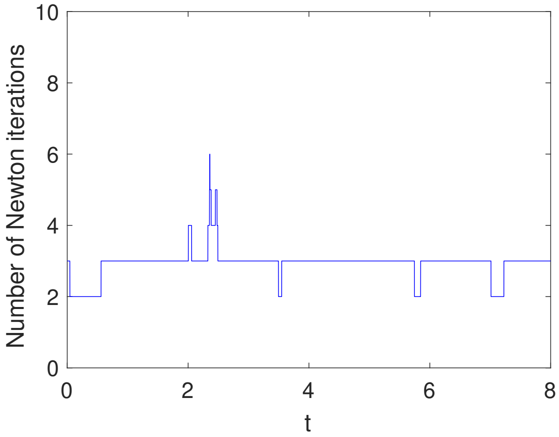


Fig. 2: Efficiency of the scheme: the number of Newton iteration in the spiky solution to saturation-concentration system.

The chemotactic sensitivity is chosen as $\gamma = 1$, and the largest concentration is $M = 100$. The error is computed at $t = 0.4$, with the reference solution computed using the time step $\delta t = 10^{-4}$. We plot the L^2 error in Fig. 1, where we can observe the first-order and second-order accuracy.

Next, we examine the efficiency of the scheme. In Fig. 2, we plot the number of Newton's iteration for a simulation of the saturation-concentration system with the spiky solution described below. We observe that, for most time steps, we only need to run at most three Newton's iterations, and the largest number is six. It indicates that our nonlinear scheme is competitive to linearly implicit schemes in efficiency, but enjoy the many advantages, such as unconditionally bound preserving and energy dissipative, that a linearly implicit scheme does not possess.

4.2 Comparison of the classical and modified systems

Next, we compare the classical and two modified parabolic–elliptic systems. We start from a case where the chemotaxis does not happen. In this case, the solution does not blow up in the classical system. The initial condition and γ are chosen as in the previous section. We use the first-order scheme with the time step $\delta t = 10^{-3}$. For the parameters in the mobility, in bounded-mobility system we choose $\kappa = 0.01$, in saturation-concentration system we choose $M = 100$. The three systems are evolved till $t = 4$. In Fig. 3, we draw the concentration ρ at $t = 4$, and plot the evolution of $\max \rho$ and the energy. The results for three systems are very close.

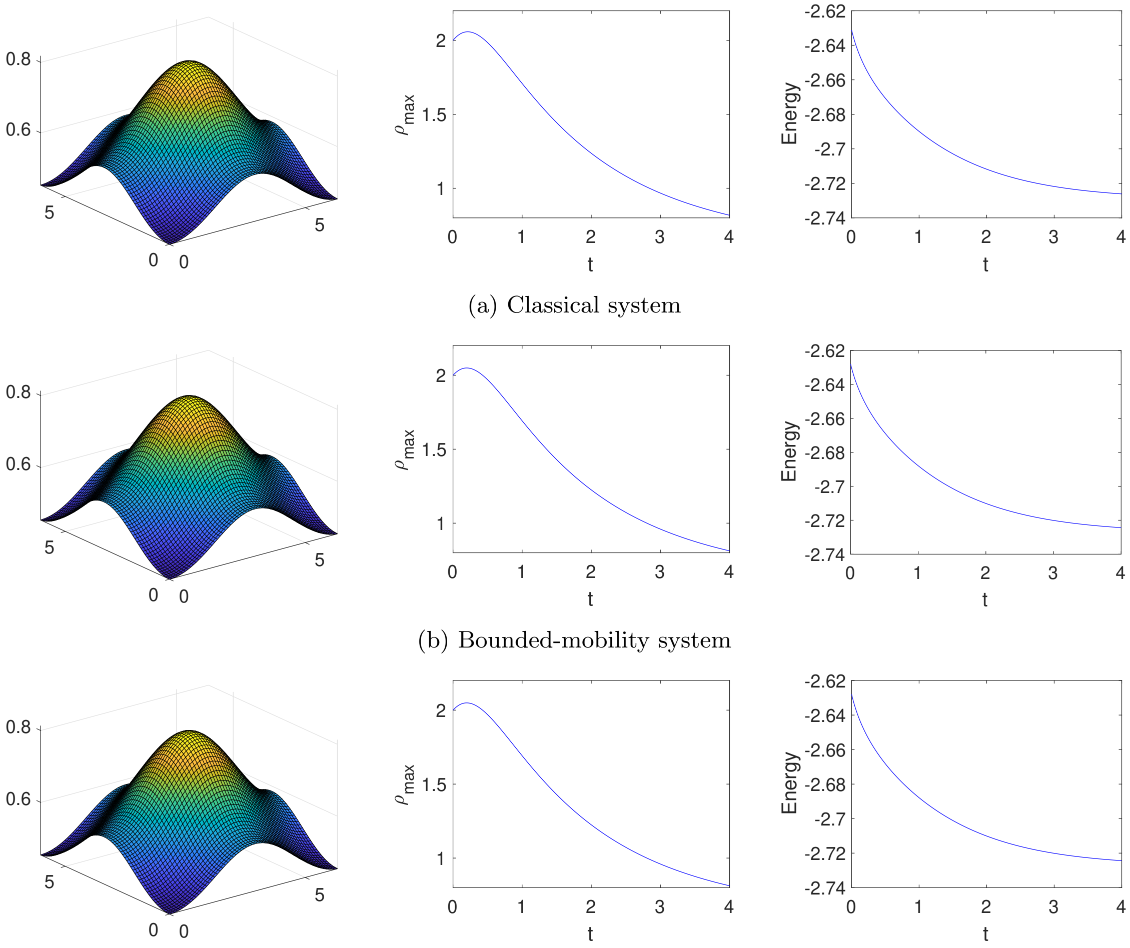
Then, we keep all the other settings, but consider another initial condition with larger total mass such that chemotaxis happens,

$$\rho(x, y, 0) = 4 \exp\left(-\frac{(x - L/2)^2 + (y - L/2)^2}{4}\right). \tag{4.2}$$

We calculate until $t = 8$ so that the three system reach steady state. The concentration at $t = 8$, the evolution of $\max \rho$ and energy are shown in Fig. 4, where in line graphs we plot results from first-order schemes and second-order schemes.

In the classical system, the mass is concentrated at four grid points, taking about 98.6% of the total mass. Note that in the classical Keller–Segel equations, the solution may blow-up in finite time. However, since our fully discretized schemes preserve positivity and conserve mass, therefore, instead of blowing up, the mass will accumulate at a few grid points. So we observe indeed a blow-up-like behavior in this case.

On the other hand, in the modified systems we obtain spiky solutions, but the accumulation still occupies some area, so they are not viewed as blow-up-like behavior. With the spiky solutions, it can be seen that

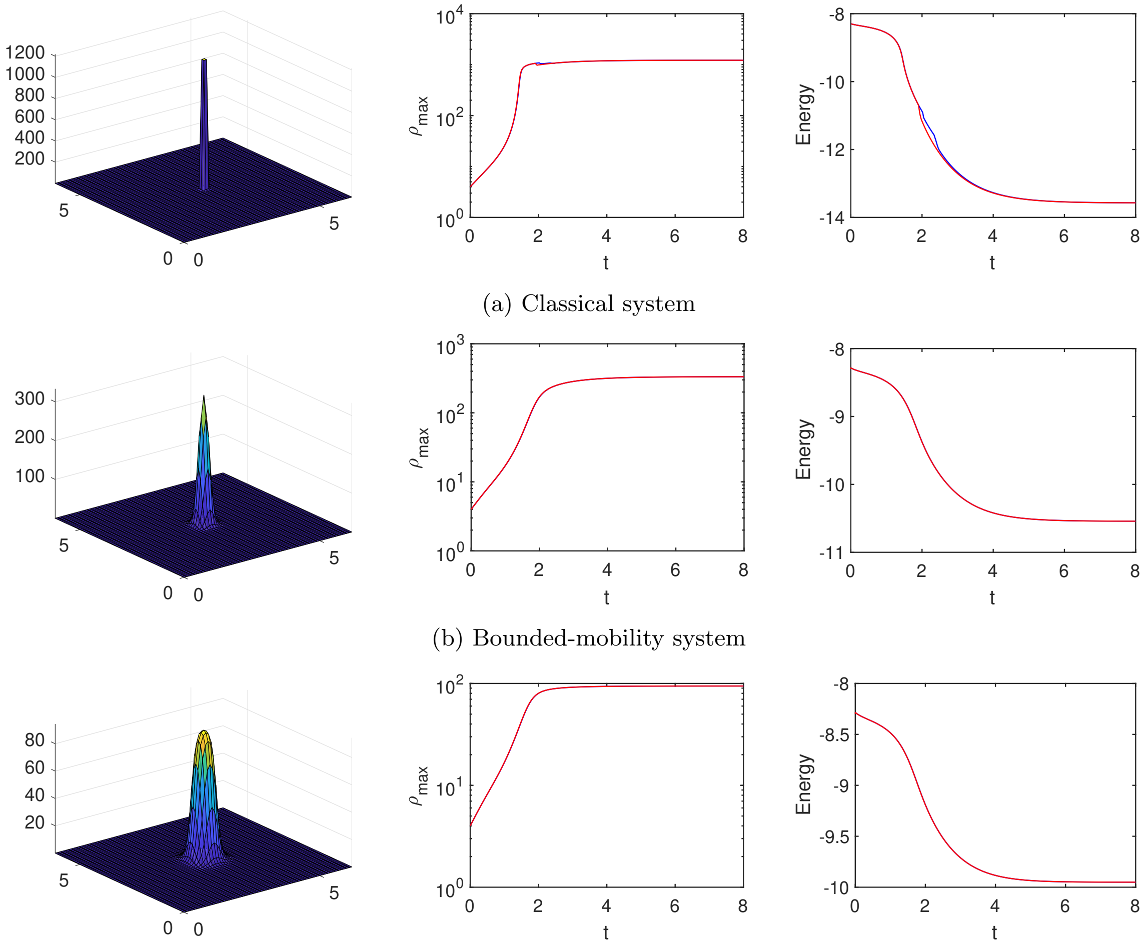


(a) Classical system

(b) Bounded-mobility system

(c) Saturation-concentration system

Fig. 3: Comparison for the classical and two modified systems: non-accumulating solutions.



(a) Classical system

(b) Bounded-mobility system

(c) Saturation-concentration system

Fig. 4: Comparison for the classical and two modified systems: chemotaxis solutions. In the graphs on energy and maximum concentration, we plot results from the first-order scheme (blue) and the second-order scheme (red).

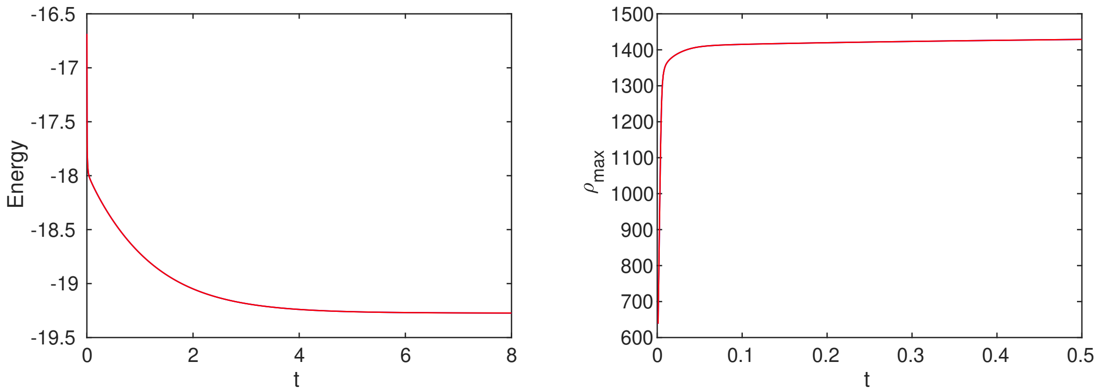


Fig. 5: Blow-up with spiky initial condition (4.3) in the classical system. The curves from the first-order scheme (blue) and second-order scheme (red) are not distinguishable.

the energy dissipation is still maintained in our scheme. Besides, from the maximum value or ρ, we observe that the accumulation is slightly slower in the modified systems than in the classical system.

It shall be clarified that the blow-up time could weight heavily on the initial conditions and possibly the parameters. We would like to provide one example, in the classical system, with a spiky initial condition,

$$\rho(x,y,0) = 500\exp\left(-30\big((x-L/2)^2+(y-L/2)^2\big)\right)+0.1, \tag{4.3}$$

without changing the parameters. We plot the evolution of energy and maximum concentration in Fig. 5, where we find that blow-up occurs rapidly in comparison with the previous example.

From the above results, it can be seen that the modified systems can successfully describe chemotaxis. Moreover, it does not lead to blow-up that is unrealistic. In what follows, we will focus on the saturation-concentration system.

Let us first look at the role of different saturation value M. Using the initial condition (4.2) with $\rho_{max} = 4$, we study three cases $M = 5, 10, 15$. The energy, maximum concentration, and steady states are given in Fig. 6. The behaviors are different for three M. For $M = 5$, the maximum concentration decreases with time, showing no accumulation. For $M = 15$, we could identify it as a typical chemotaxis. For $M = 10$, the maximum concentration does increase but still far from the saturation, and ρ is not quite close to zero for the positions away from the center, which can be regarded as an intermediate between spreading and chamotaxis. The example implies that the saturation value might affect whether the chamotaxis will happen, and that a low saturation value can inhibit the chamotaxis.

In the last example for one species parabolic–elliptic saturation-concentration system, we consider an initial condition with two bulges, given by

$$\rho(x,y,0) = 3\exp\left(-\frac{(x-L/4)^2+(y-L/4)^2}{4}\right)+3\exp\left(-\frac{(x-3L/4)^2+(y-3L/4)^2}{4}\right). \tag{4.4}$$

We choose $M = 100$, and keep the other settings the same. We plot the evolution of energy, maximum concentration for first- and second-order schemes, and four snapshots in Fig. 7. Although the total mass is larger than the previous example, we observe slower accumulation. It takes some time before two bulges merge into one, followed by accumulation. The energy dissipation is observed, slower in the merging stage and faster in the accumulating stage.

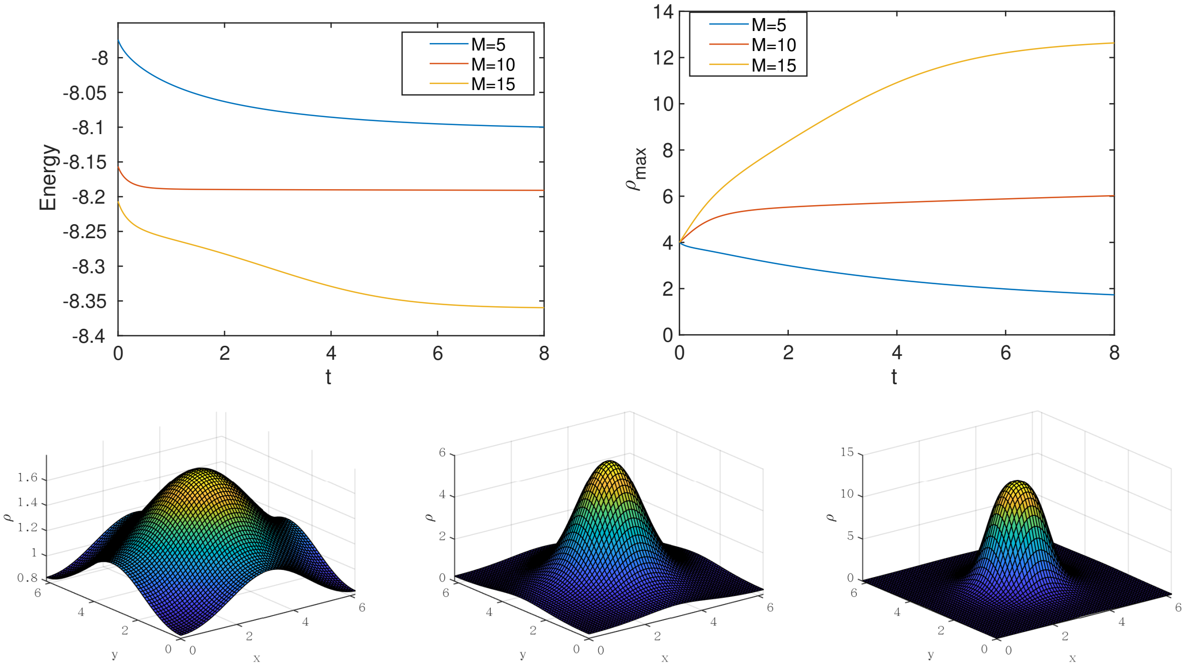


Fig. 6: Comparison of saturation-concentration system for different M. The second-order scheme is adopted. The evolution of energy and maximum concentration are given in the first row. The snapshots at $t = 8$ are given in the second row, from left to right $M = 5, 10, 15$, respectively.

4.3 Parabolic–parabolic system

Next, we consider the parabolic–parabolic ($\tau > 0$) saturation-concentration system. The chemotactic sensitivity is chosen as $\chi = 1$. The initial condition for ρ and ϕ is given by

$$\rho(x, y, 0) = \phi(x, y, 0) = 4 \exp\left(-\frac{(x - L/2)^2 + (y - L/2)^2}{4}\right), \tag{4.5}$$

where $(-\mu\Delta + \alpha)\phi(x, y, 0) \neq \chi\rho(x, y, 0)$. We use the first-order scheme (3.20)–(3.21) with the time step $\delta t = 10^{-3}$. We choose three different $\tau = 1, 10^{-2}, 10^{-4}$ and compare the results with the parabolic–elliptic system ($\tau = 0$). The evolution of energy and $\max \rho$ is plotted in Fig. 8. It shows that as τ goes to zero, the curves are converging to the curve of $\tau = 0$, which is consistent with the formal derivation in [23].

4.4 Two species

As the last example, we consider the parabolic-elliptic saturation-concentration system for two species. The parameters are chosen as $D_1 = D_2 = \gamma_1 = \gamma_2 = \mu = \chi_1 = 1$, and $\alpha = 0.1$. The initial condition is

$$\rho_1(x, y, 0) = \rho_2(x, y, 0) = 4 \exp\left(-\frac{(x - L/2)^2 + (y - L/2)^2}{4}\right). \tag{4.6}$$

We consider two different chemotactic sensitivities of the second species χ_2. First, we set $\chi_2 = 0.1$. In Fig. 9 we plot ρ_1, ρ_2 and ϕ at $t = 8$. We also plot the evolution of energy and the maximum value of ρ_1, ρ_2, where we find that curves from the first- and second-order schemes are not distinguishable. At $t = 8$, ρ_1 shows accumulation, while ρ_2 is to some extent accumulated but does not exhibit typical chemotaxis. Actually, from the evolution of its maximum values, we can see that ρ_1 keeps accumulating, but ρ_2 diffuses at first, followed by accumulation after ρ_1 has accumulated for a while.

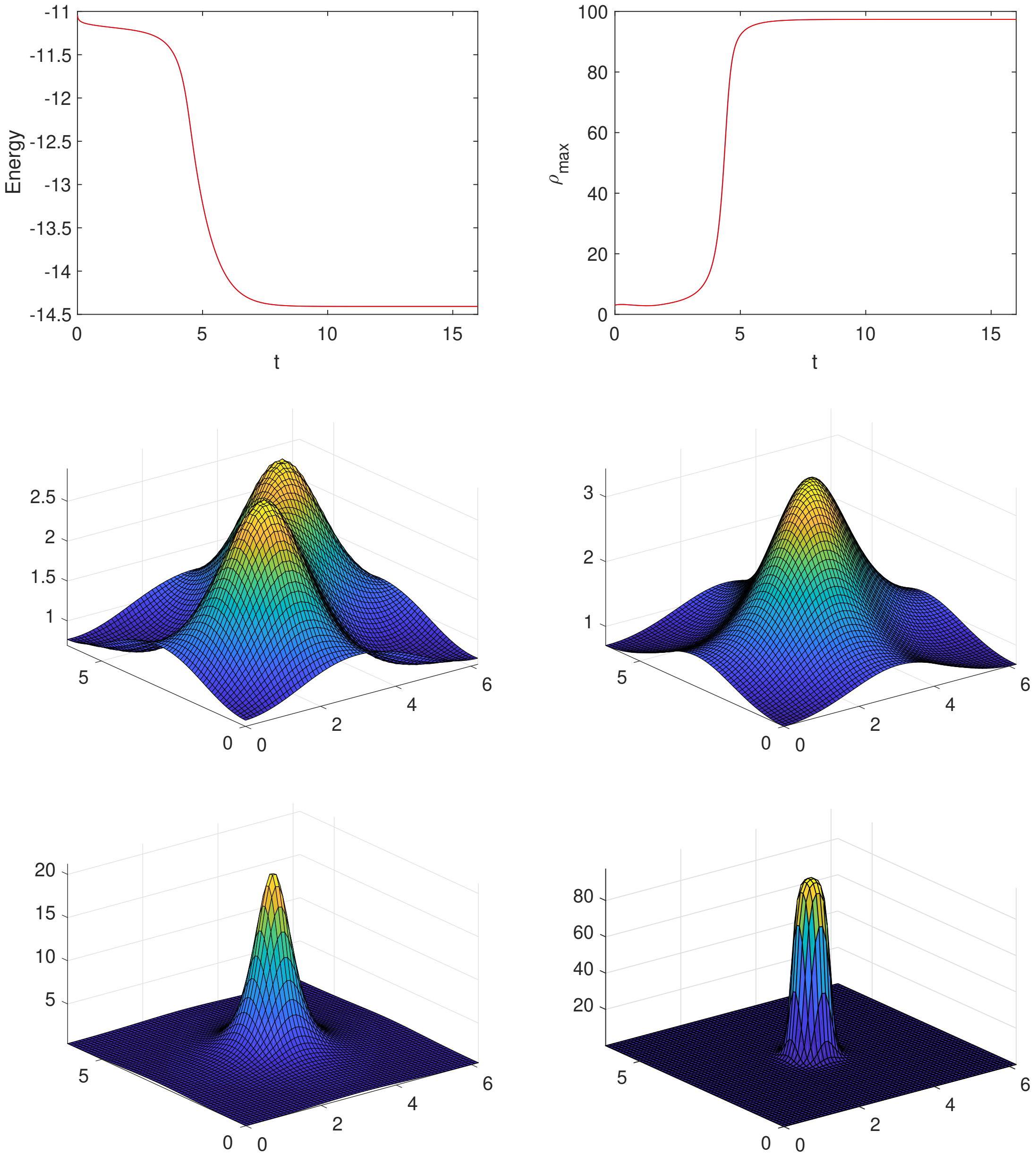


Fig. 7: Saturation-concentration system, with initial value having two bulges. The four snapshots are given at $t = 1, 2, 4, 16$. The curves from the first-order scheme (blue) and second-order scheme (red) are not distinguishable.

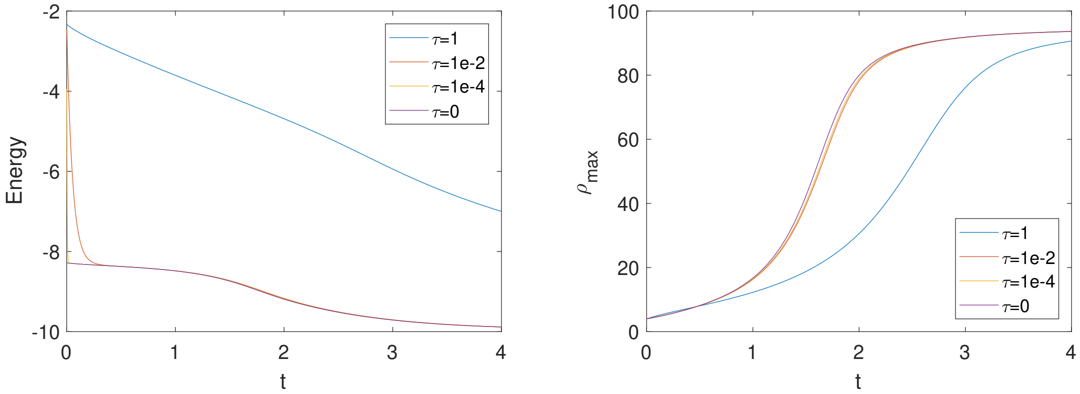


Fig. 8: Parabolic–parabolic system with different τ.

Then we look at the results with $\chi_2 = 0.01$ (Fig. 10). In this case, ρ_1 still shows accumulation, and the chemoattractant ϕ is similar to the previous case. However, ρ_2 is not actively responding to the chemoattractant due to the small χ_2. Actually, while ρ_1 keeps accumulating, ρ_2 keeps diffusing.

5 Conclusion

We proposed new numerical schemes for a class of Keller–Segel equations which possess a gradient flow structure. The main difficulties are to keep several essential properties of the Keller–Segel equations such as bound preserving, mass conservation and energy dissipation. By rewriting the dissipative operator into a form which can implicitly enforce the bound, we are able to construct a class of numerical schemes which satisfy desired properties. More precisely, our first-order schemes are mass conservative, bound preserving, uniquely solvable and energy dissipative, and our second-order schemes satisfy the first three properties but we can not prove that they are energy dissipative.

Although the schemes are nonlinear in nature, their solution can be efficiently obtained by Newton's iteration because it is the minimizer of a strictly convex functional. Furthermore, for parabolic-elliptic equations, the schemes are decoupled.

We presented numerical results to validate the theoretical results, as well as numerical simulations to show that our schemes are able to describe essential features of chemotaxis organisms such as mass accumulation, which, in the classical Keller–Segel system, is the numerical version of blow-up phenomenon.

References

[1] Nicola Bellomo, Abdelghani Bellouquid, Youshan Tao, and Michael Winkler. Toward a mathematical theory of Keller–Segel models of pattern formation in biological tissues. *Mathematical Models and Methods in Applied Sciences*, 25(09):1663–1763, 2015.

[2] Marianne Bessemoulin-Chatard and Ansgar Jüngel. A finite volume scheme for a Keller–Segel model with additional cross-diffusion. *IMA Journal of Numerical Analysis*, 34(1):96–122, 2013.

[3] Piotr Biler and Tadeusz Nadzieja. Existence and nonexistence of solutions for a model of gravitational interaction of particles, I. In *Colloquium Mathematicae*, volume 66(2), pages 319–334, 1993.

[4] Adrien Blanchet, José Antonio Carrillo, David Kinderlehrer, Michał Kowalczyk, Philippe Laurençot, and Stefano Lisini. A hybrid variational principle for the Keller–Segel system in $\mathbb{R}^2$. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1553–1576, 2015.

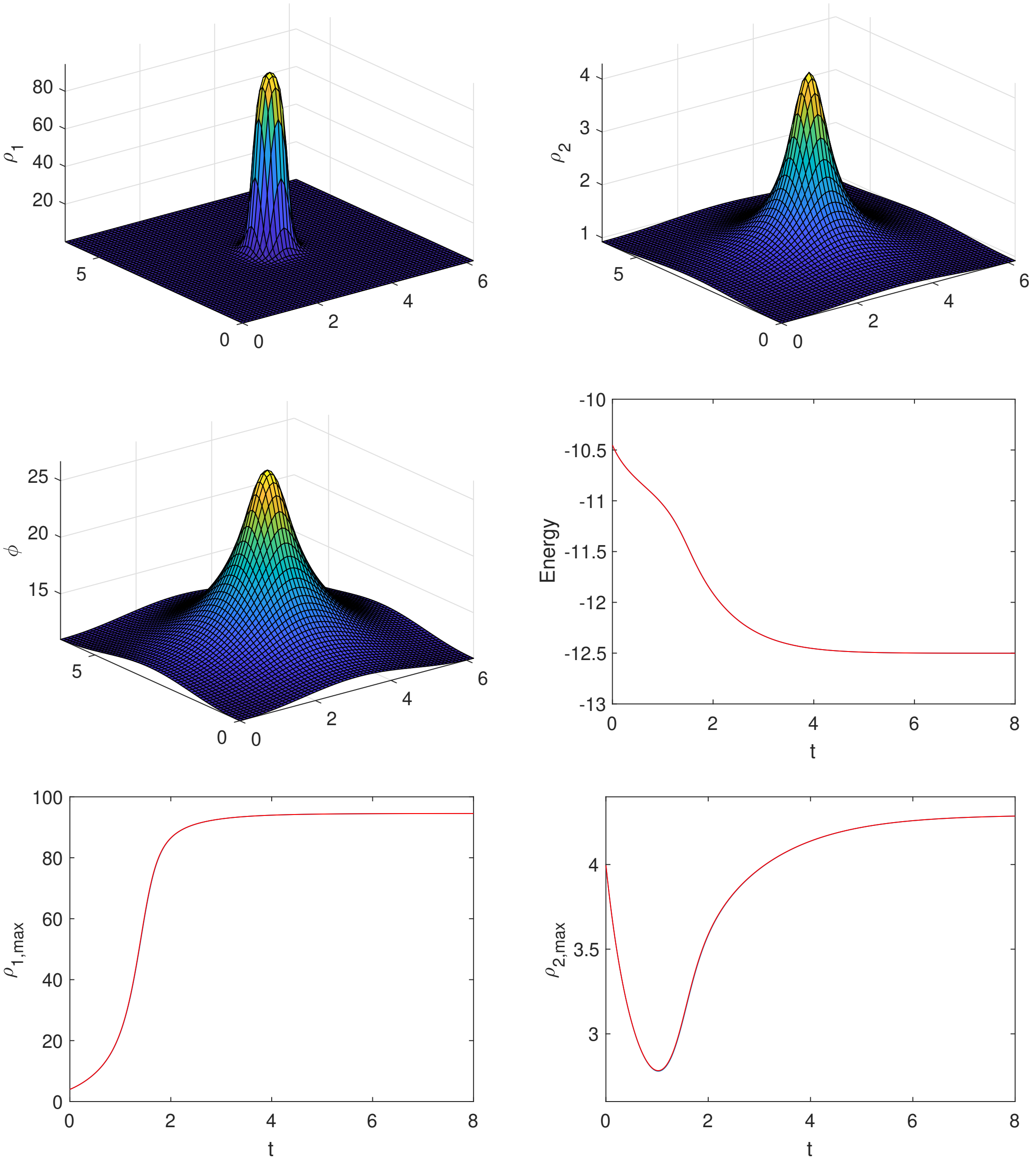


Fig. 9: Two species, $\chi_2 = 0.1$. The curves from the first-order scheme (blue) and second-order scheme (red) are not distinguishable. The snapshots are at $t = 8$.

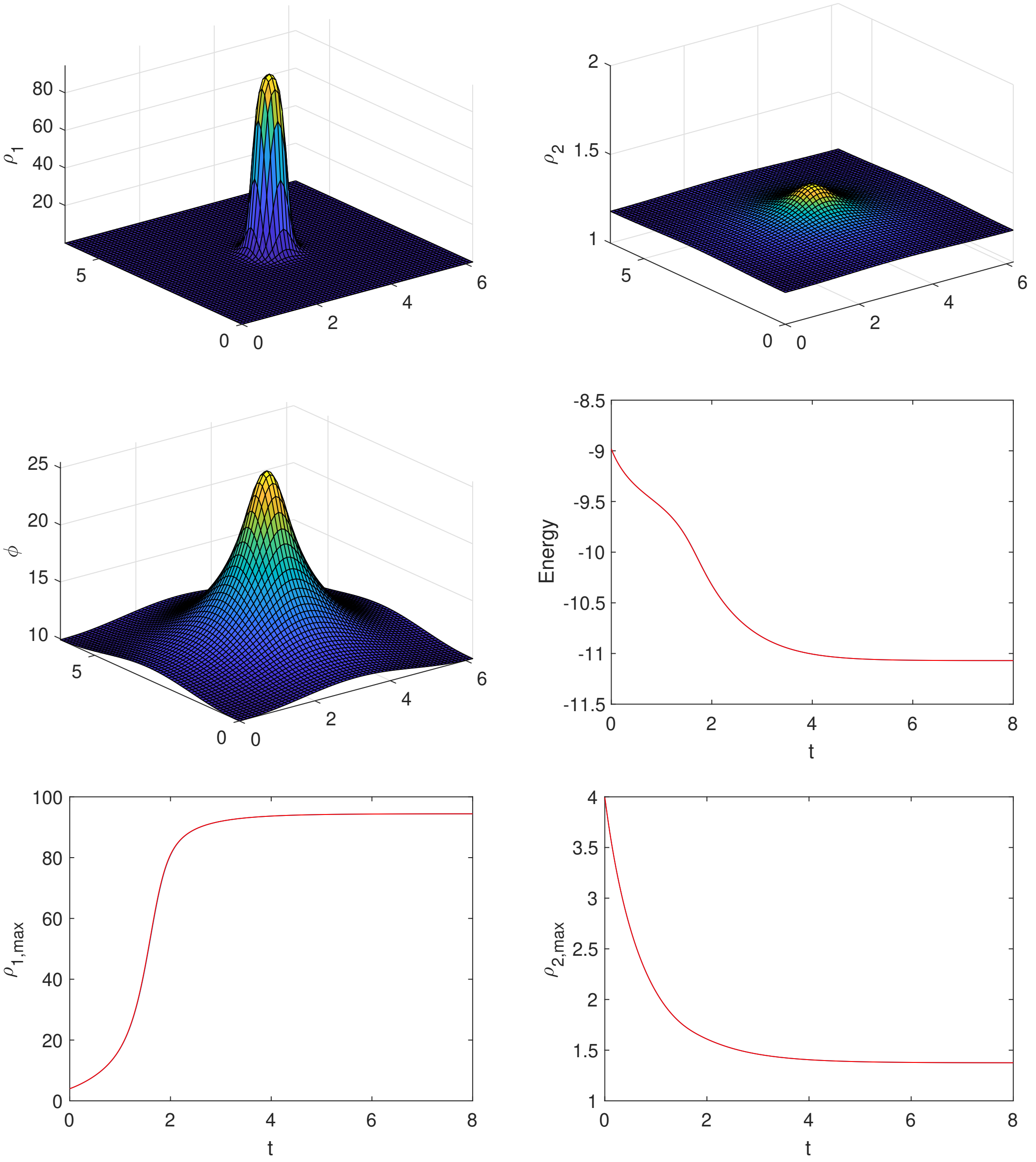


Fig. 10: Two species, $\chi_2 = 0.01$. The curves from the first-order scheme (blue) and second-order scheme (red) are not distinguishable. The snapshots are at $t = 8$.

[5] Adrien Blanchet, Jean Dolbeault, and Benoît Perthame. Two-dimensional Keller–Segel model: Optimal critical mass and qualitative properties of the solutions. *Electronic Journal of Differential Equations (EJDE)[electronic only]*, 2006.

[6] Vincent Calvez, Lucilla Corrias, et al. The parabolic–parabolic Keller–Segel model in $\mathbb{R}^2$. *Communications in Mathematical Sciences*, 6(2):417–447, 2008.

[7] José A Carrillo, Alina Chertock, and Yanghong Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17(1):233–258, 2015.

[8] Wenbin Chen, Cheng Wang, Xiaoming Wang, and Steven M Wise. A positivity-preserving, energy stable numerical scheme for the Cahn-Hilliard equation with logarithmic potential. *arXiv:1712.03225*, 2017.

[9] Alina Chertock and Alexander Kurganov. A second-order positivity preserving central-upwind scheme for chemotaxis and haptotaxis models. *Numerische Mathematik*, 111(2):169, 2008.

[10] Luís Neves de Almeida, Federica Bubba, Benoît Perthame, and Camille Pouchol. Energy and implicit discretization of the fokker-planck and keller-segel type equations. *arXiv preprint arXiv:1803.10629*, 2018.

[11] Yasmin Dolak and Christian Schmeiser. The Keller–Segel model with logistic sensitivity function and small diffusivity. *SIAM Journal on Applied Mathematics*, 66(1):286–308, 2005.

[12] C. M. Elliott and A. M. Stuart. The global dynamics of discrete semilinear parabolic equations. *SIAM J. Numer. Anal.*, 30(6):1622–1663, 1993.

[13] Yekaterina Epshteyn. Upwind-difference potentials method for Patlak–Keller–Segel chemotaxis model. *Journal of Scientific Computing*, 53(3):689–713, 2012.

[14] Yekaterina Epshteyn and Ahmet Izmirlioglu. Fully discrete analysis of a discontinuous finite element method for the Keller–Segel chemotaxis model. *Journal of Scientific Computing*, 40(1-3):211–256, 2009.

[15] Yekaterina Epshteyn and Alexander Kurganov. New interior penalty discontinuous Galerkin methods for the Keller–Segel chemotaxis model. *SIAM Journal on Numerical Analysis*, 47(1):386–408, 2008.

[16] Francis Filbet. A finite volume scheme for the Patlak–Keller–Segel chemotaxis model. *Numerische Mathematik*, 104(4):457–488, 2006.

[17] Thomas Hillen and Kevin Painter. Global existence for a parabolic chemotaxis model with prevention of overcrowding. *Advances in Applied Mathematics*, 26(4):280–301, 2001.

[18] Thomas Hillen and Kevin J Painter. A user's guide to PDE models for chemotaxis. *Journal of mathematical biology*, 58(1-2):183, 2009.

[19] Dirk Horstmann and Michael Winkler. Boundedness vs. blow-up in a chemotaxis system. *Journal of Differential Equations*, 215(1):52–107, 2005.

[20] Evelyn F Keller and Lee A Segel. Initiation of slime mold aggregation viewed as an instability. *Journal of theoretical biology*, 26(3):399–415, 1970.

[21] Evelyn F Keller and Lee A Segel. Model for chemotaxis. *Journal of theoretical biology*, 30(2):225–234, 1971.

[22] Alexander Kurganov and Maria Lukacova-Medvidova. Numerical study of two-species chemotaxis models. *Discrete Contin. Dyn. Syst. Ser. B*, 19(1):131–152, 2014.

[23] Jian-Guo Liu, Li Wang, and Zhennan Zhou. Positivity-preserving and asymptotic preserving method for 2D Keller–Segal equations. *Mathematics of Computation*, 87(311):1165–1189, 2018.

[24] Clifford S Patlak. Random walk with persistence and external bias. *The bulletin of mathematical biophysics*, 15(3):311–338, 1953.

[25] Benoît Perthame. *Transport equations in biology.* Frontiers in Mathematics, Birkhauser Verlag, Basel, 2007.

[26] Markus Schmuck. Analysis of the Navier–Stokes–Nernst–Planck–Poisson system. *Mathematical Models and Methods in Applied Sciences*, 19(06):993–1014, 2009.

[27] Jie Shen and Jie Xu. Convergence and error analysis for the scalar auxiliary variable (SAV) schemes to gradient flows. *SIAM Journal on Numerical Analysis*, 56(5):2895–2912, 2018.

[28] Jie Shen and Jie Xu. Unconditionally positivity preserving and energy dissipative schemes for Poisson–Nernst–Planck equations. *submitted*, 2019.

[29] Jie Shen, Jie Xu, and Jiang Yang. The scalar auxiliary variable (SAV) approach for gradient flows. *J. Comput. Phys.*, 353:407–416, 2018.

[30] Jie Shen, Jie Xu, and Jiang Yang. A new class of efficient and robust energy stable schemes for gradient flows. *SIAM Review*, 61(3):474–506, 2019.

[31] Juan JL Velázquez. Point dynamics in a singular limit of the Keller–Segel model 1: Motion of the concentration regions. *SIAM Journal on Applied Mathematics*, 64(4):1198–1223, 2004.

[32] Juan JL Velázquez. Point dynamics in a singular limit of the Keller–Segel model 2: Formation of the concentration regions. *SIAM Journal on Applied Mathematics*, 64(4):1224–1248, 2004.

[33] Guanyu Zhou and Norikazu Saito. Finite volume methods for a Keller–Segel system: discrete energy, error estimates and numerical blow-up analysis. *Numerische Mathematik*, 135(1):265–311, 2017.