IMA Journal of Numerical Analysis (2023) **43**, 1450–1484
https://doi.org/10.1093/imanum/drac014
Advance Access publication on 18 May 2022

Positivity-preserving and energy-dissipative finite difference schemes for the Fokker–Planck and Keller–Segel equations

JINGWEI HU
Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA

AND

XIANGXIONG ZHANG*
Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA
*Corresponding author: hujw@uw.edu
zhan1966@purdue.edu

[Received on 28 April 2021; revised on 28 November 2021]

In this work we introduce semi-implicit or implicit finite difference schemes for the continuity equation with a gradient flow structure. Examples of such equations include the linear Fokker–Planck equation and the Keller–Segel equations. The two proposed schemes are first-order accurate in time, explicitly solvable, and second-order and fourth-order accurate in space, which are obtained via finite difference implementation of the classical continuous finite element method. The fully discrete schemes are proved to be positivity preserving and energy dissipative: the second-order scheme can achieve so unconditionally while the fourth-order scheme only requires a mild time step and mesh size constraint. In particular, the fourth-order scheme is the *first* high order spatial discretization that can achieve both positivity and energy decay properties, which is suitable for long time simulation and to obtain accurate steady state solutions.

Keywords: Positivity; energy dissipation; Fokker-Planck; Keller-Segel; finite difference; high order accuracy; implicit.

1. Introduction

In this paper we are interested in the continuity equation of the form

$$\partial_t \rho = \nabla \cdot [\rho \nabla (\mathcal{H}'(\rho) + \mathcal{V} + \mathcal{W} * \rho)], \quad t > 0, \ \mathbf{x} \in \Omega \subset \mathbb{R}^d, \tag{1}$$

$$\rho(0, \mathbf{x}) = \rho_0(\mathbf{x}), \tag{2}$$

where $\rho = \rho(t, \mathbf{x}) \geq 0$ is the unknown density function, $\mathcal{H}(\rho)$ is the internal energy that is assumed to be convex, $\mathcal{V}(\mathbf{x})$ is the external potential and $\mathcal{W}(\mathbf{x})$ is the interaction potential. The typical boundary condition of (1) is the no-flux boundary:

$$\nabla (\mathcal{H}'(\rho) + \mathcal{V} + \mathcal{W} * \rho) \cdot \mathbf{n} = 0, \quad \mathbf{x} \in \partial\Omega, \tag{3}$$

where $\mathbf{n}$ is the outward normal. Therefore, the total mass is conserved

$$\int_\Omega \rho(t, \mathbf{x}) \, \mathrm{d}\mathbf{x} = \int_\Omega \rho_0(\mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

Equations of the form (1) appear in various contexts, for example, in modeling of porous medium Vazquez (2007), granular materials Carrillo *et al.* (2003), and collective behavior of biological and social systems Carrillo *et al.* (2019). In particular, we focus on the following two cases in this paper: the linear Fokker–Planck equation and the Keller–Segel model of chemotaxis. For both cases the internal energy function is given by

$$\mathcal{H}(\rho) = \rho \log \rho - \rho. \tag{4}$$

In the Fokker–Planck equation

$$\mathcal{V} = \mathcal{V}(\mathbf{x}), \quad \mathcal{W} \equiv 0,$$

where $\mathcal{V}(\mathbf{x})$ is some given function bounded from below in Ω. In this case (1) can also be written as a convection–diffusion equation,

$$\partial_t \rho = \Delta \rho + \nabla \cdot (\rho \nabla \mathcal{V}). \tag{5}$$

In the Keller–Segel model ρ is the density of some bacteria and

$$\mathcal{V} \equiv 0, \quad \mathcal{W} * \rho = -c,$$

where $c = c(t, \mathbf{x})$ is the density of chemical attractant satisfying an elliptic equation in Ω with a constant $\alpha \geq 0$:

$$-\Delta c + \alpha c = \rho. \tag{6}$$

In this case (1) can be written as

$$\partial_t \rho = \Delta \rho - \nabla \cdot (\rho \nabla c), \tag{7}$$

which is coupled with (6) to form a system. Note that if Ω is $\mathbb{R}^d$, $\mathcal{W}$ is the Newtonian potential when $\alpha = 0$ and the Bessel potential when $\alpha > 0$. By integrating (6) in Ω we obtain

$$-\nabla c \cdot \mathbf{n}\big|_{\partial\Omega} + \alpha \int_\Omega c \, \mathrm{d}\mathbf{x} = \int_\Omega \rho \, \mathrm{d}\mathbf{x}.$$

Therefore, the boundary condition of c must be compatible with the equation above. When $\alpha = 0$ the Neumann boundary condition must satisfy the compatibility condition

$$-\nabla c \cdot \mathbf{n}\big|_{\partial\Omega} = \int_\Omega \rho_0 \, \mathrm{d}\mathbf{x}.$$

When $\alpha > 0$, if we consider the homogeneous Neumann boundary $\nabla c \cdot \mathbf{n}\big|_{\partial\Omega} = 0$, then

$$\alpha \int_\Omega c \, \mathrm{d}\mathbf{x} = \int_\Omega \rho_0 \, \mathrm{d}\mathbf{x},$$

i.e., the mass of c is also conserved.

The equation (1) has a variational structure. It is the gradient flow, with respect to the 2-Wasserstein metric, of the free energy functional (Villani, 2003):

$$\mathcal{E}(\rho) = \int_\Omega \left(\mathcal{H}(\rho) + \mathcal{V}\rho + \frac{1}{2}(\mathcal{W} * \rho)\rho \right) \mathrm{d}\mathbf{x}. \tag{8}$$

Indeed

$$\frac{\delta \mathcal{E}}{\delta \rho} = \xi, \quad \xi := \mathcal{H}'(\rho) + \mathcal{V} + \mathcal{W} * \rho,$$

hence

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} = \int_\Omega \frac{\delta \mathcal{E}}{\delta \rho} \partial_t \rho \, \mathrm{d}\mathbf{x} = \int_\Omega \xi \nabla \cdot (\rho \nabla \xi) \, \mathrm{d}\mathbf{x} = -\int_\Omega \rho |\nabla \xi|^2 \, \mathrm{d}\mathbf{x} \leq 0. \tag{9}$$

Note that for $\mathcal{H}$ given in (4) we can define

$$\mathcal{M} = e^{\log \rho - \xi} = e^{-(\mathcal{V} + \mathcal{W} * \rho)}.$$

With this $\mathcal{M}$ the equation (1) can be written equivalently as

$$\partial_t \rho = \nabla \cdot \left(\mathcal{M} \nabla \left(\frac{\rho}{\mathcal{M}} \right) \right). \tag{10}$$

The boundary condition (3) becomes

$$\nabla \left(\frac{\rho}{\mathcal{M}} \right) \cdot \mathbf{n} = 0, \quad \mathbf{x} \in \partial \Omega. \tag{11}$$

Furthermore, the energy (8) can be written equivalently as

$$\mathcal{E}(\rho) = \int_\Omega \left(\rho \log \left(\frac{\rho}{\mathcal{M}} \right) - \rho - \frac{1}{2}(\mathcal{W} * \rho)\rho \right) \mathrm{d}\mathbf{x}. \tag{12}$$

When written in form (10), the original continuity equation (1) can be viewed as a 'variable coefficient' diffusion equation, for which we are able to construct efficient positivity-preserving and energy-dissipative schemes, i.e., the discrete analog of (12) is decreasing in time. In the literature there are many numerical schemes for the Fokker–Planck or Keller–Segel type equations. Recently, significant efforts have been devoted to structure-preserving discretizations to preserve, for instance, the positivity of the solution and energy decay at the semi-discrete or fully discrete level. We summarize some of the recent methods according to their types of time discretization. The first kind of methods are fully explicit schemes. For a scalar convection–diffusion equation such as (3) there are quite a few explicit positivity-preserving schemes (Zhang *et al.*, 2013; Li *et al.*, 2018; Srinivasan *et al.*, 2018; Qiu *et al.*, 2021), however, with a small time step constraint $\Delta t = \mathcal{O}(\Delta x^2)$, which is unacceptable in applications requiring long time simulation. Most importantly, it is usually quite difficult to establish energy dissipation in these positivity-preserving schemes. Some recent explicit schemes, including a finite volume method in Carrillo *et al.* (2015) and discontinuous Galerkin methods in Guo *et al.* (2019);

Sun *et al.* (2018), can indeed achieve energy dissipation, but only in the semi-discrete setting (i.e., the time is left as continuous). The second kind of methods are implicit or semi-implicit nonlinear schemes. For such schemes it is possible to preserve positivity and energy dissipation in the fully discrete setting without a small time step constraint (Almeida *et al.*, 2019; Bailo *et al.*, 2020; Shen & Xu, 2020), but they often involve nonlinear systems, for which robust nonlinear system solvers are needed. The third kind of methods are implicit or semi-implicit schemes that are explicitly solvable. By formulating the continuity equation as in (10) and treating $\mathcal{M}$ explicitly one can derive a semi-implicit scheme, in which only a linear system needs to be solved without small time-step constraint. Note that this approach is only possible for linear diffusions (for $\mathcal{H}$ given by (4)) and has been used in many previous works, for example, Jin & Yan (2011); Liu *et al.* (2018); Hu & Shu (2019); Hu & Huang (2020); Hu *et al.* (2021). Although details vary they all use the second-order central finite difference for spatial discretization. We use the third approach for the time discretization in this paper. However, the proposed spatial discretization can achieve fourth-order accuracy, which is one of the main novelties. Furthermore, we can prove the fully discrete positivity and energy decay property for the fourth-order spatial discretization under reasonable mesh size and time step constraints. We emphasize that the time step constraint in this paper is a lower bound, thus no small time-step constraint like $\Delta t = \mathcal{O}(\Delta x^2)$ is required. To the best of our knowledge, this is the first high order spatial discretization that can achieve these properties for the linear Fokker–Planck and Keller–Segel type equations.

The rest of this paper organized as follows. In Section 2 we introduce the finite difference schemes, which are obtained by finite difference implementation of continuous finite element method with the linear and quadratic polynomials. In Section 3 we show that both the second-order and fourth-order schemes are monotone. It is well known that the second-order central difference or linear finite element method for linear diffusion forms an M-matrix thus is monotone. The fourth-order accurate scheme or the finite element method with quadratic polynomial basis no longer gives an M-matrix, but monotonicity can still be proved under practical mesh size and time step constraints. In Section 4 we show that monotonicity implies positivity and fully discrete energy dissipation in these schemes. Section 5 includes numerical tests on the Fokker–Planck equation and Keller–Segel system. Concluding remarks are given in Section 6.

2. Finite difference schemes

In this section we introduce a simple numerical scheme for equation (10) with a first-order accurate semi-implicit time discretization. For the spatial discretization we use second-order and fourth-order accurate finite difference schemes, which are obtained from finite element method using linear and quadratic polynomial bases, respectively. It is well known that a finite element method with suitable quadrature is also a finite difference scheme. In particular, the fourth-order accurate finite difference scheme considered here is equivalent to the Lagrangian Q^2 (tensor product of polynomials of degree 2) finite element method with 3-point Gauss–Lobatto quadrature, which is also known as the Q^2 spectral element method (Maday & Rønquist, 1990). The main novelty here is that we can prove rigorous positivity-preserving and energy-dissipation properties for the fully discrete scheme, especially the fourth-order spatial discretization in one and two spatial dimensions.

In this section we mainly focus on how the finite difference schemes are defined. The explicit form of the schemes will be given in Section 3. We only consider one and two spatial dimensions in this paper, even though one can also derive these schemes in higher dimensions.

2.1 *Time discretization*

We propose the following semi-implicit discretization of (10):

$$\frac{\rho^{n+1}-\rho^n}{\Delta t} = \nabla\cdot\left(\mathcal{M}^n\nabla\left(\frac{\rho^{n+1}}{\mathcal{M}^n}\right)\right), \quad \mathbf{x}\in\Omega, \tag{13}$$

where

$$\mathcal{M}^n = e^{-(\mathcal{V}+\mathcal{W}*\rho^n)}.$$

The no-flux boundary condition (11) is imposed as

$$\nabla\left(\frac{\rho^{n+1}}{\mathcal{M}^n}\right)\cdot\mathbf{n} = 0, \quad \mathbf{x}\in\partial\Omega. \tag{14}$$

Note that (13) is equivalent to

$$\frac{\rho^{n+1}-\rho^n}{\Delta t} = \nabla\cdot(\rho^{n+1}\nabla(\log\rho^{n+1}+\mathcal{V}+\mathcal{W}*\rho^n))$$

for discretizing the original equation (1).

We then introduce the auxiliary variables defined as

$$\tilde{g}^{n+1} := \frac{\rho^{n+1}}{\mathcal{M}^n}, \quad g^n := \frac{\rho^n}{\mathcal{M}^n}, \tag{15}$$

and write the scheme (13) as

$$\mathcal{M}^n\tilde{g}^{n+1} - \Delta t\nabla\cdot\left(\mathcal{M}^n\nabla\tilde{g}^{n+1}\right) = \mathcal{M}^n g^n. \tag{16}$$

Accordingly, the boundary condition (14) becomes the homogeneous Neumann boundary for the auxiliary variable:

$$\nabla\tilde{g}^{n+1}\cdot\mathbf{n} = 0, \quad \mathbf{x}\in\partial\Omega.$$

After multiplying a test function $v\in H^1(\Omega)$ to (16) and integration by parts using the boundary condition for $\tilde{g}^{n+1}$, we obtain the variational form of (16): seek $\tilde{g}^{n+1}\in H^1(\Omega)$ that satisfies

$$(\mathcal{M}^n\tilde{g}^{n+1},v) + \Delta t(\mathcal{M}^n\nabla\tilde{g}^{n+1},\nabla v) = (\mathcal{M}^n g^n, v), \quad \forall v\in H^1(\Omega),$$

where $(v,w) := \int_\Omega vw\,\mathrm{d}\mathbf{x}$ denotes the L^2 inner product in Ω.

REMARK 2.1 In the Fokker–Planck equation $\mathcal{M} = \exp(-\mathcal{V}(\mathbf{x}))$ is a time-independent quantity and (13) simplifies to a fully implicit scheme. For brevity our following presentation will focus on the Keller–Segel equation for which $\mathcal{M}^n = \exp(c^n(\mathbf{x}))$. Reduction to the Fokker–Planck case will be commented whenever necessary.

2.2 *Spatial discretization*

We consider a uniform rectangular mesh Ω_h for the rectangular domain Ω. For any rectangle e in the mesh Ω_h let Q^k be the space of tensor product polynomials of degree k. For instance, in two dimensions,

$$Q^k(e) = \left\{ p(x,y) = \sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}x^i y^j, (x,y) \in e \right\}.$$

Let V^h be the continuous piecewise Q^k polynomial space defined on Ω_h:

$$V^h = \{v_h(\mathbf{x}) \in C(\Omega) : v_h\big|_e \in Q^k(e), \forall e \in \Omega_h\}.$$

The Q^k finite element method for (16) is to seek $\tilde{g}_h^{n+1} \in V^h$ satisfying

$$(\mathcal{M}^n \tilde{g}_h^{n+1}, v_h) + \Delta t(\mathcal{M}^n \nabla \tilde{g}_h^{n+1}, \nabla v_h) = (\mathcal{M}^n g_h^n, v_h), \quad \forall v_h \in V^h, \tag{17}$$

where $\mathcal{M}^n$ is regarded as a given variable coefficient at time step n.

The Q^k spectral element method is to replace all integrals in (17) by m-point Gauss–Lobatto quadrature with $m \geq k+1$ in each dimension. Standard finite element method error estimates still hold if $m \geq k+1$, i.e., the Q^k spectral element method is $(k+1)$th order accurate in L^2-norm and kth order accurate in H^1-norm for smooth solutions of an elliptic equation, see Maday & Rønquist (1990). We consider the simplest choice of quadrature, using $(k+1)$-point Gauss–Lobatto quadrature. Then the method is to find $\tilde{g}_h^{n+1} \in V^h$ satisfying

$$\langle \mathcal{M}^n \tilde{g}_h^{n+1}, v_h\rangle + \Delta t\langle \mathcal{M}^n \nabla \tilde{g}_h^{n+1}, \nabla v_h\rangle = \langle \mathcal{M}^n g_h^n, v_h\rangle, \quad \forall v_h \in V^h, \tag{18}$$

where $\langle\cdot,\cdot\rangle$ denotes that integrals are replaced by $(k+1)$-point Gauss–Lobatto quadrature.

For a two-dimensional problem, a Q^k polynomial on a rectangular element e can be represented as a Lagrangian interpolation polynomial at $(k+1)\times(k+1)$ Gauss–Lobatto points, thus all Gauss–Lobatto points in (18) are not only quadrature nodes but also nodes, representing all degrees of freedom. So the Q^k spectral element method (18) also becomes a finite difference scheme on all Gauss–Lobatto nodes. For $k \geq 3$ the Gauss–Lobatto points are not uniform in each element. For $k \leq 2$ all Gauss–Lobatto nodes on Ω_h correspond to a uniform grid, see Fig. 1 for an illustration of the Q^2 mesh. Moreover, for $k \geq 2$, such a finite difference scheme can be proved to be $(k+2)$-order accurate in discrete l^2-norm for elliptic equations (Li & Zhang, 2020b) and for parabolic equations (Li *et al.*, 2022), e.g., the Q^2 spectral element method can be regarded as a fourth-order accurate finite difference scheme.

In this paper we only consider the linear case $k = 1$ and quadratic case $k = 2$, because only in these two cases the schemes can be proved to be positivity preserving and energy dissipative. To derive an equivalent matrix form of the scheme (18) let $\phi_i(\mathbf{x})$ $(i = 1, \cdots, N)$ be the Q^k Lagrangian basis

at all Gauss–Lobatto points $\mathbf{x}_i$ ($i = 1, \cdots, N$) on Ω_h. For any piecewise polynomial $u_h(\mathbf{x}) \in V^h$ let $u_i = u_h(\mathbf{x}_i)$. Then $u_h(\mathbf{x}) = \sum_{i=1}^{N} u_i \phi_i(\mathbf{x})$. Let $\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix}$ and w_i be the quadrature weight at $\mathbf{x}_i$.

With the notation above we have

$$\langle \mathcal{M}^n \tilde{g}_h^{n+1}, v_h \rangle = \sum_{i=1}^{N} w_i \mathcal{M}_i^n \tilde{g}_i^{n+1} v_i = \mathbf{v}^T W M^n \mathbf{g}^{n+1}, \tag{19}$$

where $W = \mathrm{diag}\{w_1, \cdots, w_N\}$ and $M^n = \mathrm{diag}\{\mathcal{M}_1^n, \cdots, \mathcal{M}_N^n\}$ are diagonal matrices. We also have

$$\langle \mathcal{M}^n \nabla \tilde{g}_h^{n+1}, \nabla v_h \rangle = \mathbf{v}^T S \tilde{\mathbf{g}}^{n+1}, \tag{20}$$

where S is the stiffness matrix from the same spectral element method solving a Poisson equation $-\nabla \cdot (\mathcal{M}^n \nabla u) = f$ in Ω with homogeneous Neumann boundary condition $\nabla u \cdot \mathbf{n} = 0$ on $\partial\Omega$. In other words, S is the stiffness matrix in the scheme of seeking $u_h \in V^h$ satisfying

$$\langle \mathcal{M}^n \nabla u_h, \nabla v_h \rangle = \langle f, v_h \rangle, \quad \forall v_h \in V^h.$$

We emphasize that the stiffness matrix S depends on $\mathcal{M}_i^n > 0$. It is common knowledge in finite element theory that S satisfies two properties:

1. S is real symmetric and positive semi-definite.
2. Its null space is one-dimensional and the null vector is $\mathbf{1}$.

Here for brevity we do not give the explicit form of S. The complete scheme (18) in one and two dimensions will be given in Section 3.

Using (19) and (20) the finite difference scheme (18) can be written in the matrix form as: find $\tilde{\mathbf{g}}^{n+1}$ satisfying

$$\mathbf{v}^T W M^n \tilde{\mathbf{g}}^{n+1} + \Delta t \mathbf{v}^T S \tilde{\mathbf{g}}^{n+1} = \mathbf{v}^T W M^n \mathbf{g}^n, \quad \forall \mathbf{v} \in \mathbb{R}^N, \tag{21}$$

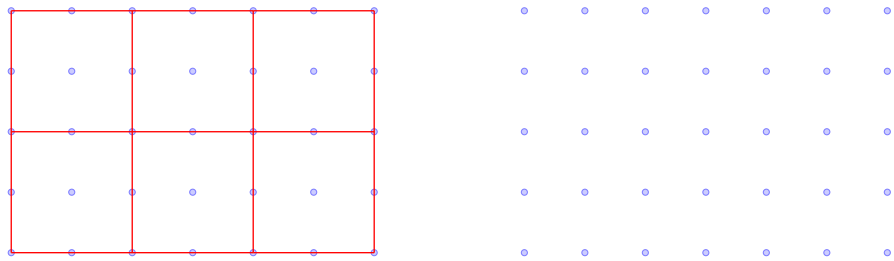
or equivalently

$$W M^n \tilde{\mathbf{g}}^{n+1} + \Delta t S \tilde{\mathbf{g}}^{n+1} = W M^n \mathbf{g}^n. \tag{22}$$

Noticing (15), (22) can also be written as

$$W \boldsymbol{\rho}^{n+1} + \Delta t S (M^n)^{-1} \boldsymbol{\rho}^{n+1} = W \boldsymbol{\rho}^n. \tag{23}$$

REMARK 2.2 Even though the scheme (23) for $\boldsymbol{\rho}$ does not involve any auxiliary variable $\mathbf{g}$, the division by $\mathcal{M}_i^n$ is still needed in (23). Moreover, (22) gives a symmetric positive definite linear system, but (23) does not. In practice both can be solved by preconditioned conjugate gradient methods with efficient inversion of Laplacian as a preconditioner, see Section 7 in Li & Zhang (2020b) for implementation details. In our numerical tests we solve the system (22) by preconditioned conjugate gradient.



(a) All quadrature points on Ω_h

(b) The corresponding finite difference grid on Ω_h

FIG. 1. The 3×3 Gauss–Lobatto quadrature points for Q^2 finite element method on a uniform mesh Ω_h naturally gives a uniform finite difference grid.

2.3 *The full scheme for the Keller–Segel system*

In the case of the Keller–Segel system, in addition to (22) (the discretization for (7)), one also needs to discretize the equation (6). Here we consider $\alpha > 0$ and the homogeneous Neumann boundary condition $\nabla c \cdot \mathbf{n}|_{\partial\Omega} = 0$. We use the same scheme as in (18): find $c_h^n \in V^h$ satisfying

$$\langle \nabla c_h^n, \nabla v_h \rangle + \alpha \langle c_h^n, v_h \rangle = \langle \rho^n, v_h \rangle, \quad \forall v_h \in V^h. \tag{24}$$

Similarly, as in the previous subsection (24) can be written equivalently in the finite difference or matrix form.

In one dimension the second order scheme ($k = 1$) can be written as

$$\frac{1}{h^2} K \mathbf{c}^n + \alpha \mathbf{c}^n = \boldsymbol{\rho}^n,$$

and the fourth-order scheme ($k = 2$) can be written as

$$\frac{1}{h^2} H \mathbf{c}^n + \alpha \mathbf{c}^n = \boldsymbol{\rho}^n,$$

where h is the grid spacing and

$$K = \begin{pmatrix} 2 & -2 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -2 & 2 & \end{pmatrix}_{N\times N}, \quad H = \begin{pmatrix} \frac{7}{2} & -4 & \frac{1}{2} & & & & & \\ -1 & 2 & -1 & & & & & \\ \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} & & & \\ & & -1 & 2 & -1 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} \\ & & & & & -1 & 2 & -1 \\ & & & & & \frac{1}{2} & -4 & \frac{7}{2} \end{pmatrix}_{N\times N}.$$

We emphasize that N must be odd in the matrix H for the fourth-order scheme because the grid points are from Gauss–Lobatto nodes, see Fig. 1.

In two dimensions let $\mathbf{c}$ be a two-dimensional array with $\mathbf{c}_{ij}$ denoting (i,j) point value. Let $vec(\mathbf{c})$ be a column vector obtained by rearranging entries in $\mathbf{c}$ column by column. Then the second-order and fourth-order schemes can be written, respectively, as

$$\frac{1}{h^2}(K \otimes K)vec(\mathbf{c}^n) + \alpha vec(\mathbf{c}^n) = \boldsymbol{\rho}^n,$$

and

$$\frac{1}{h^2}(H \otimes H)vec(\mathbf{c}^n) + \alpha vec(\mathbf{c}^n) = \boldsymbol{\rho}^n.$$

To summarize, the full finite difference scheme for the Keller–Segel system (6)–(7) is implemented as follows:

1. At time level n, given point values ρ_i^n at each node $\mathbf{x}_i$, solve (24) to obtain c_i^n, then compute point values of $\mathcal{M}_i^n = \exp(c_i^n)$. In multiple dimensions the linear system can be easily and efficiently inverted by eigenvalue decomposition of K and H, see Li & Zhang (2020b) for details.
2. With point values $g_i^n := \frac{\rho_i^n}{\mathcal{M}_i^n}$ obtain $\tilde{g}_i^{n+1}$ by solving (22).
3. Update ρ by $\rho_i^{n+1} := \mathcal{M}_i^n \tilde{g}_i^{n+1}$.

Remark 2.3 The finite difference scheme for the Fokker–Planck equation (3) is simpler: at each node $\mathbf{x}_i$, first define $\mathcal{M}_i = \exp(-\mathcal{V}_i)$.

1. At time level n, given point values ρ_i^n, compute $g_i^n := \frac{\rho_i^n}{\mathcal{M}_i}$, then obtain $\tilde{g}_i^{n+1}$ by solving (22).
2. Update ρ by $\rho_i^{n+1} := \mathcal{M}_i \tilde{g}_i^{n+1}$.

2.4 *Accuracy of the spatial discretization*

For the Q^2 finite element method with 3-point Gauss–Lobatto quadrature, it is well known that the standard L^2-norm error estimate is third order. However, when regarded as a finite difference scheme at Gauss–Lobatto points, it can be rigorously proved that it is a fourth-order accurate scheme in the discrete l^2-norm (Li & Zhang, 2020b; Li *et al.*, 2022). In particular, this has been proved for Dirichlet boundary conditions in Li & Zhang (2020b). Only $\mathcal{O}(h^{3.5})$ can be proved for Neumann boundary conditions for an operator like $-\nabla(A(\mathbf{x})\nabla u)$ where $A(\mathbf{x})$ is a positive definite matrix, and the one half order loss is purely due to the mixed second-order derivatives. Nonetheless, for the equations we are interested in here, i.e., an operator like $-\nabla \cdot (a(\mathbf{x})\nabla u)$ with a scalar coefficient $a(\mathbf{x})$, since there are no mixed second-order derivatives involved, the same proof in Li & Zhang (2020b); Li *et al.* (2022) applies to show that the fourth-order accuracy also holds for Neumann boundary conditions of elliptic equations, see Li (2021) for a detailed proof. So for both (22) and (24) we will refer to the Q^2 scheme as the fourth-order accurate spatial discretization, i.e., it is a fourth-order accurate scheme for solving a steady state problem.

For the Q^1 finite element method with quadrature, it is also well known that it gives the most popular second-order central finite difference scheme. However, for the Neumann boundary condition, there is still some subtle difference, which will be reviewed in Remark 3.3 of Section 3.

3. Monotonicity of the finite difference schemes

A matrix A is called *monotone* if its inverse has non-negative entries $A^{-1} \geq 0$. In this section we discuss the monotonicity of the matrix used in the second-order and fourth-order finite difference schemes (18), which is the key intrinsic property implying positivity and energy dissipation.

In particular, we consider the matrix form (22), which can also be written as

$$(M^n + \Delta t W^{-1} S)\tilde{\mathbf{g}}^{n+1} = M^n \mathbf{g}^n. \tag{25}$$

We will discuss the monotonicity of the matrix $M^n + \Delta t W^{-1} S$. For simplicity, we will drop superscript n in M in the rest of this section.

For the second-order scheme it is well known that it forms an M-matrix thus is monotone, which will be reviewed. For the fourth order scheme the monotonicity for Dirichlet boundary condition in two dimensions was proved in Li & Zhang (2020a). The same results in Li & Zhang (2020a) also hold for the Neumann boundary conditions. For completeness, in this section, we include a detailed proof for the monotonicity of the fourth-order scheme (25) with the homogeneous Neumann boundary condition for $\tilde{g}^{n+1}$, which is equivalent to the no-flux boundary condition for ρ^{n+1}.

3.1 *M-matrices*

The only viable tool in the literature to prove monotonicity is to use M-matrices. Nonsingular M-matrices are monotone matrices and there are many equivalent definitions or characterizations of M-matrices, see Plemmons (1977). By condition K_{35} in Plemmons (1977) a sufficient and necessary characterization is as follows:

THEOREM 3.1 For a real square matrix A with positive diagonal entries and nonpositive off-diagonal entries, A is a nonsingular M-matrix if and only if there exists a positive diagonal matrix D such that AD has all positive row sums.

The following is a convenient sufficient, but not necessary characterization, of nonsingular M-matrices Li & Zhang (2020a):

THEOREM 3.2 For a real square matrix A with positive diagonal entries and nonpositive off-diagonal entries A is a nonsingular M-matrix if all the row sums of A are nonnegative and at least one row sum is positive.

3.2 *The second-order scheme in one dimension*

In the one-dimensional case assume the domain is $\Omega = [-L, L]$ and the uniform grid points are $-L = x_1 < x_2 < \cdots < x_N = L$ with grid spacing h. Following derivations in Section 7 of Li & Zhang (2020b), it is straightforward to show that the linear finite element method (25) with a variable coefficient $\mathcal{M} > 0$

can be explicitly written as:

$$\begin{aligned}
&\mathcal{M}_1\tilde{g}_1^{n+1} + \Delta t\frac{(\mathcal{M}_1+\mathcal{M}_2)\tilde{g}_1^{n+1} - (\mathcal{M}_1+\mathcal{M}_2)\tilde{g}_2^{n+1}}{h^2} = \mathcal{M}_1 g_1^n;\\
&\mathcal{M}_i\tilde{g}_i^{n+1} + \Delta t\frac{-(\mathcal{M}_{i-1}+\mathcal{M}_i)\tilde{g}_{i-1}^{n+1} + (\mathcal{M}_{i-1}+2\mathcal{M}_i+\mathcal{M}_{i+1})\tilde{g}_i^{n+1} - (\mathcal{M}_i+\mathcal{M}_{i+1})\tilde{g}_{i+1}^{n+1}}{2h^2}\\
&= \mathcal{M}_i g_i^n, \quad i = 2,\cdots,N-1;\\
&\mathcal{M}_N\tilde{g}_N^{n+1} + \Delta t\frac{-(\mathcal{M}_{N-1}+\mathcal{M}_N)\tilde{g}_{N-1}^{n+1} + (\mathcal{M}_{N-1}+\mathcal{M}_N)\tilde{g}_N^{n+1}}{h^2} = \mathcal{M}_N g_N^n.
\end{aligned} \tag{26}$$

It is easy to see that $M^n + \Delta t W^{-1}S$ is a tridiagonal matrix satisfying Theorem 3.2, thus is a nonsingular M-matrix and monotone.

Now for the ease of presentation of the scheme we will abuse notation by introducing ghost point values as $\tilde{g}_0^{n+1} := \tilde{g}_2^{n+1}$, $\tilde{g}_{N+1}^{n+1} := \tilde{g}_{N-1}^{n+1}$ and $\mathcal{M}_0 := \mathcal{M}_2$, $\mathcal{M}_{N+1} := \mathcal{M}_{N-1}$. Then the scheme can be equivalently written as

$$\begin{aligned}
&\mathcal{M}_i\tilde{g}_i^{n+1} + \Delta t\frac{-(\mathcal{M}_{i-1}+\mathcal{M}_i)\tilde{g}_{i-1}^{n+1} + (\mathcal{M}_{i-1}+2\mathcal{M}_i+\mathcal{M}_{i+1})\tilde{g}_i^{n+1} - (\mathcal{M}_i+\mathcal{M}_{i+1})\tilde{g}_{i+1}^{n+1}}{2h^2}\\
&= \mathcal{M}_i g_i^n, \quad i = 1,\cdots,N.
\end{aligned} \tag{27}$$

We emphasize that the scheme still has a different structure at the boundary points, and here ghost points are used only for a uniform expression of the scheme. In actual implementation there are no ghost points.

REMARK 3.3 One popular finite difference method to solve (13) is to apply the central finite difference as

$$\frac{\rho_i^{n+1} - \rho_i^n}{\Delta t} = \frac{F_{i+\frac{1}{2}}^{n+1} - F_{i-\frac{1}{2}}^{n+1}}{h},$$

with the flux term defined by

$$F_{i+\frac{1}{2}}^{n+1} = \frac{1}{h}\frac{\mathcal{M}_i+\mathcal{M}_{i+1}}{2}\left(\frac{\rho_{i+1}^{n+1}}{\mathcal{M}_{i+1}} - \frac{\rho_i^{n+1}}{\mathcal{M}_i}\right),$$

which is equivalent to

$$\tilde{g}_i^{n+1} - g_i^n = \frac{\Delta t}{h\mathcal{M}_i}(G_{i+\frac{1}{2}}^{n+1} - G_{i-\frac{1}{2}}^{n+1}), \quad G_{i+\frac{1}{2}}^{n+1} = \frac{1}{h}\frac{\mathcal{M}_i+\mathcal{M}_{i+1}}{2}\left(\tilde{g}_{i+1}^{n+1} - \tilde{g}_i^{n+1}\right).$$

For approximating no-flux boundary condition, if simply setting $G^{n+1}_{\frac{1}{2}} = G^{n+1}_{N+\frac{1}{2}} = 0$, then the scheme becomes

$$\begin{aligned}
&\mathcal{M}_1\tilde{g}_1^{n+1} + \Delta t\frac{(\mathcal{M}_1+\mathcal{M}_2)\tilde{g}_1^{n+1} - (\mathcal{M}_1+\mathcal{M}_2)\tilde{g}_2^{n+1}}{2h^2} = \mathcal{M}_1 g_1^n;\\
&\mathcal{M}_i\tilde{g}_i^{n+1} + \Delta t\frac{-(\mathcal{M}_{i-1}+\mathcal{M}_i)\tilde{g}_{i-1}^{n+1} + (\mathcal{M}_{i-1}+2\mathcal{M}_i+\mathcal{M}_{i+1})\tilde{g}_i^{n+1} - (\mathcal{M}_i+\mathcal{M}_{i+1})\tilde{g}_{i+1}^{n+1}}{2h^2}\\
&= \mathcal{M}_i g_i^n, \quad i = 2,\cdots,N-1;\\
&\mathcal{M}_N\tilde{g}_N^{n+1} + \Delta t\frac{-(\mathcal{M}_{N-1}+\mathcal{M}_N)\tilde{g}_{N-1}^{n+1} + (\mathcal{M}_{N-1}+\mathcal{M}_N)\tilde{g}_N^{n+1}}{2h^2} = \mathcal{M}_N g_N^n.
\end{aligned} \tag{28}$$

If using the same grid $-L = x_1 < x_2 < \cdots < x_N = L$ with grid spacing h the scheme (28) is the same as (26) at interior points. For boundary points (28) is only first-order accurate, which can be easily verified for constant coefficient case $\mathcal{M}_i \equiv 1$. If redefining g_i and $\mathcal{M}_i$ as point values at a staggered uniform grid $-L + \frac{h}{2} = x_1 < x_2 < \cdots < x_N = L - \frac{h}{2}$ with spacing h (as has been done in most papers in the past, e.g., Hu & Huang, 2020), the scheme (28) exhibits second-order accuracy in many numerical tests. However, even on the staggered grid, the local truncation error of (28) at $x_1 = -L + \frac{h}{2}$ and $x_N = L - \frac{h}{2}$ is only first order, thus it is quite difficult to rigorously prove the second-order accuracy of (28) by conventional finite difference analysis. On the other hand, it can be easily proved that (26) is second-order accurate by standard finite element analysis.

3.3 *The second-order scheme in multiple dimensions*

In the two-dimensional case assume the domain is $\Omega = [-L, L] \times [-L, L]$ with a uniform $N \times N$ grid point with spacing h, which is a tensor product of the grid $-L = x_1 < x_2 < \cdots < x_N = L$. Let $\mathbf{g}$ be an $N \times N$ matrix with g_{ij} denoting the point value at the (i, j) grid point.

We introduce the ghost values for $i, j = 1, \cdots, N$ as:

$$\begin{aligned}
&\tilde{g}_{0,j}^{n+1} := \tilde{g}_{2,j}^{n+1},\ \tilde{g}_{N+1,j}^{n+1} := \tilde{g}_{N-1,j}^{n+1},\ \tilde{g}_{i,0}^{n+1} := \tilde{g}_{i,2}^{n+1},\ \tilde{g}_{i,N+1}^{n+1} := \tilde{g}_{i,N-1}^{n+1},\\
&\mathcal{M}_{0,j} := \mathcal{M}_{2,j},\ \mathcal{M}_{N+1,j} := \mathcal{M}_{N-1,j},\ \mathcal{M}_{i,0} := \mathcal{M}_{i,2},\ \mathcal{M}_{i,N+1} := \mathcal{M}_{i,N-1}.
\end{aligned}$$

Then the Lagrangian Q^1 finite element method with 2-point Gauss–Lobatto quadrature (18) can be explicitly expressed as

$$\begin{aligned}
&\Delta t\frac{-(\mathcal{M}_{i-1,j}+\mathcal{M}_{ij})\tilde{g}_{i-1,j}^{n+1} + (\mathcal{M}_{i-1,j}+2\mathcal{M}_{ij}+\mathcal{M}_{i+1,j})\tilde{g}_{ij}^{n+1} - (\mathcal{M}_{ij}+\mathcal{M}_{i+1,j})\tilde{g}_{i+1,j}^{n+1}}{2h^2}\\
&\quad + \Delta t\frac{-(\mathcal{M}_{i,j-1}+\mathcal{M}_{ij})\tilde{g}_{i,j-1}^{n+1} + (\mathcal{M}_{i,j-1}+2\mathcal{M}_{ij}+\mathcal{M}_{i,j+1})\tilde{g}_{ij}^{n+1} - (\mathcal{M}_{ij}+\mathcal{M}_{i,j+1})\tilde{g}_{i,j+1}^{n+1}}{2h^2}\\
&\quad + \mathcal{M}_{ij}\tilde{g}_{ij}^{n+1} = \mathcal{M}_{ij} g_{ij}^n, \quad \forall i, j = 1, \cdots, N.
\end{aligned}$$

It is easy to see that $M^n + \Delta t W^{-1} S$ is a matrix satisfying Theorem 3.2, thus is a nonsingular M-matrix and monotone.

REMARK 3.4 The scheme in the three-dimensional case can be similarly written, and it is also straightforward to verify that $M^n + \Delta t W^{-1} S$ is a matrix satisfying Theorem 3.2, thus is a nonsingular M-matrix and monotone.

REMARK 3.5 We have seen that using the formulation (10) the second-order finite difference scheme with a semi-implicit time discretization is unconditionally monotone, thus always positivity-preserving and energy-dissipative (details to be given in Section 4). This is true even for blow-up solutions. As a comparison, for the Keller–Segel equation, one can also use the formulation (7), and apply the second–order finite difference for both convection and diffusion operators with a semi-implicit time discretization, but the monotonicity can only be proved under a mesh constraint $h\|\nabla c\|_\infty \le 2$. This is one of the key advantages of solving (10) instead of (7).

3.4 *Lorenz's condition for monotonicity*

For high order accurate schemes, especially for a variable coefficient problem, the stiffness matrices are no longer M-matrices. Yet, it is possible to show that the stiffness matrix is a product of two or more M-matrices thus still monotone (Cross & Zhang, 2020; Li & Zhang, 2020a) by using the Lorenz's Theorem in Lorenz (1977), which will be briefly reviewed in this subsection.

DEFINITION 1 Let $\mathcal{N} = \{1, 2, \dots, n\}$. For $\mathcal{N}_1, \mathcal{N}_2 \subset \mathcal{N}$, we say a matrix A of size $n \times n$ connects $\mathcal{N}_1$ with $\mathcal{N}_2$ if

$$\forall i_0 \in \mathcal{N}_1, \exists i_r \in \mathcal{N}_2, \exists i_1, \dots, i_{r-1} \in \mathcal{N} \quad \text{s.t.} \quad a_{i_{k-1} i_k} \neq 0, \quad k = 1, \cdots, r. \tag{29}$$

If perceiving A as a directed graph adjacency matrix of vertices labeled by $\mathcal{N}$, then (29) simply means that there exists a directed path from any vertex in $\mathcal{N}_1$ to at least one vertex in $\mathcal{N}_2$. In particular, if $\mathcal{N}_1 = \emptyset$, then any matrix A connects $\mathcal{N}_1$ with $\mathcal{N}_2$.

Given a square matrix A and a column vector $\mathbf{x}$ we define

$$\mathcal{N}^0(A\mathbf{x}) = \{i : (A\mathbf{x})_i = 0\}, \quad \mathcal{N}^+(A\mathbf{x}) = \{i : (A\mathbf{x})_i > 0\}.$$

Given a matrix $A = [a_{ij}] \in \mathbb{R}^{n \times n}$ define its diagonal, off-diagonal, positive and negative off-diagonal parts as $n \times n$ matrices A_d, A_a, A_a^+, A_a^-:

$$(A_d)_{ij} = \begin{cases} a_{ii}, & \text{if} \quad i = j \\ 0, & \text{if} \quad i \neq j \end{cases}, \quad A_a = A - A_d,$$

$$(A_a^+)_{ij} = \begin{cases} a_{ij}, & \text{if} \quad a_{ij} > 0, \quad i \neq j \\ 0, & \text{otherwise.} \end{cases}, \quad A_a^- = A_a - A_a^+.$$

The following two results were proved in Lorenz (1977). See also Li & Zhang (2020a) for a detailed proof.

THEOREM 3.6 If $A \le M_1 M_2 \cdots M_k L$ where $M_1, \cdots, M_k$ are nonsingular M-matrices and $L_a \le 0$, and there exists a nonzero vector $\mathbf{e} \ge 0$ such that one of the matrices $M_1, \cdots, M_k, L$ connects $\mathcal{N}^0(A\mathbf{e})$ with

$\mathcal{N}^+(A\mathbf{e})$. Then $M_k^{-1}M_{k-1}^{-1}\cdots M_1^{-1}A$ is an M-matrix, thus A is a product of $k+1$ nonsingular M-matrices and $A^{-1} \geq 0$.

THEOREM 3.7 (Lorenz's condition). If A_a^- has a decomposition: $A_a^- = A^z + A^s = (a_{ij}^z) + (a_{ij}^s)$ with $A^s \leq 0$ and $A^z \leq 0$, such that

$$A_d + A^z \text{ is a nonsingular M-matrix,} \tag{30a}$$

$$A_a^+ \leq A^z A_d^{-1} A^s \text{ or equivalently } \forall a_{ij} > 0 \text{ with } i \neq j, a_{ij} \leq \sum_{k=1}^{n} a_{ik}^z a_{kk}^{-1} a_{kj}^s, \tag{30b}$$

$$\exists \mathbf{e} \in \mathbb{R}^n \setminus \{\mathbf{0}\}, \mathbf{e} \geq 0 \text{ with } A\mathbf{e} \geq 0 \text{ s.t. } A^z \text{ or } A^s \text{ connects } \mathcal{N}^0(A\mathbf{e}) \text{ with } \mathcal{N}^+(A\mathbf{e}). \tag{30c}$$

Then A is a product of two nonsingular M-matrices, thus $A^{-1} \geq 0$.

It was proved in Cross & Zhang (2020) that

COROLLARY 3.8 The matrix L in Theorem 3.6 must be an M-matrix.

In practice, the condition (30c) can be difficult to verify. For the scheme we are interested in here, the vector $\mathbf{e}$ can be taken as $\mathbf{1}$ consisting of all ones, then the condition (30c) can be simplified. For the scheme (25), as long as $\mathcal{M}_i > 0$, we always have $A\mathbf{1} > 0$, thus $N^0(A\mathbf{1}) = \emptyset$ and (30c) is trivially satisfied. We summarize it as follows:

THEOREM 3.9 Let A denote the matrix representation of the fourth-order finite difference scheme obtained from Lagrangian Q^2 finite element method with 3-point Gauss–Lobatto quadrature solving $-\nabla\cdot(b\nabla)u + cu = f$ with variable coefficients $b > 0$ and $c > 0$, and homogeneous Neumann boundary condition in a rectangular domain. Assume A_a^- has a decomposition $A_a^- = A^z + A^s$ with $A^s \leq 0$ and $A^z \leq 0$. Then A is a product of two M-matrices, thus $A^{-1} \geq 0$, if the following are satisfied:

1. $(A_d + A^z)\mathbf{1} \neq \mathbf{0}$ and $(A_d + A^z)\mathbf{1} \geq 0$;
2. $A_a^+ \leq A^z A_d^{-1} A^s$.

3.5 *The fourth-order scheme in one dimension*

In the one-dimension case assume the domain $\Omega = [-L, L]$ is partitioned into k uniform intervals with cell length $2h$. Then all 3-point Gauss–Lobatto points for each small interval form a uniform grid $-L = x_1 < x_2 < \cdots < x_N = L$ with grid spacing h and $N = 2k + 1$. Thus, the number of grid points for this fourth-order scheme must be odd.

For convenience we consider an equivalent form of (25):

$$W^{-1}S\tilde{\mathbf{g}}^{n+1} + \frac{1}{\Delta t}M^n\tilde{\mathbf{g}}^{n+1} = \frac{1}{\Delta t}M^n\mathbf{g}^n. \tag{31}$$

Let $A = W^{-1}S + \frac{1}{\Delta t}M^n$ and $\mathcal{A} : \mathbb{R}^{N\times 1} \longrightarrow R^{N\times 1}$ be the scheme operator, i.e., (31) can be written as $\mathcal{A}(\tilde{\mathbf{g}}^{n+1})_i = \frac{1}{\Delta t}\mathcal{M}_i g_i^n$. Following the derivations in Li & Zhang (2020a,b), with the same *ghost point values* notation in Section 3.2, the finite element method with quadratic basis and 3-point Gauss–Lobatto

quadrature can be explicitly written as follows: for all $i = 1, \cdots, N$, if x_i is a cell end (i is odd),

$$\begin{aligned}\mathcal{A}(\tilde{\mathbf{g}}^{n+1})_i :=& \frac{(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)\tilde{g}_{i-2}^{n+1} - (4\mathcal{M}_{i-2} + 12\mathcal{M}_i)\tilde{g}_{i-1}^{n+1}}{8h^2} \\ &+ \frac{(\mathcal{M}_{i-2} + 4\mathcal{M}_{i-1} + 18\mathcal{M}_i + 4\mathcal{M}_{i+1} + \mathcal{M}_{i+2})\tilde{g}_i^{n+1}}{8h^2} \\ &\frac{-(12\mathcal{M}_i + 4\mathcal{M}_{i+2})\tilde{g}_{i+1}^{n+1} + (3\mathcal{M}_{i+2} - 4\mathcal{M}_{i+1} + 3\mathcal{M}_i)\tilde{g}_{i+2}^{n+1}}{8h^2} + \frac{\mathcal{M}_i}{\Delta t}\tilde{g}_i^{n+1} \\ =& \frac{\mathcal{M}_i}{\Delta t} g_i^n; \end{aligned} \tag{32}$$

and if x_i is a cell center (i is even),

$$\mathcal{A}(\tilde{\mathbf{g}}^{n+1})_i := \frac{-(3\mathcal{M}_{i-1} + \mathcal{M}_{i+1})\tilde{g}_{i-1}^{n+1} + 4(\mathcal{M}_{i-1} + \mathcal{M}_{i+1})\tilde{g}_i^{n+1} - (\mathcal{M}_{i-1} + 3\mathcal{M}_{i+1})\tilde{g}_{i+1}^{n+1}}{4h^2} + \frac{\mathcal{M}_i}{\Delta t}\tilde{g}_i^{n+1} = \frac{\mathcal{M}_i}{\Delta t} g_i^n. \tag{33}$$

Next, for the matrix A, we will discuss a decomposition of its negative off-diagonal parts of $A_a^- = A^z + A^s$ such that Theorem 3.9 can be verified under suitable mesh and time step constraints. We will use operator notations to represent all matrices. With the positive and negative parts for a number f defined as:

$$f^+ = \frac{|f| + f}{2}, \quad f^- = \frac{|f| - f}{2},$$

the linear operators $\mathcal{A}_d$, $\mathcal{A}_a^{\pm}$ are:

If x_i is a cell end (i is odd),

$$\mathcal{A}_d(\tilde{\mathbf{g}}^{n+1})_i = \left(\frac{\mathcal{M}_{i-2} + 4\mathcal{M}_{i-1} + 18\mathcal{M}_i + 4\mathcal{M}_{i+1} + \mathcal{M}_{i+2}}{8h^2} + \frac{\mathcal{M}_i}{\Delta t}\right)\tilde{g}_i^{n+1};$$

if $x_i$is a cell center (i is even), $\quad \mathcal{A}_d(\tilde{\mathbf{g}}^{n+1})_i = \left(\frac{\mathcal{M}_{i-1} + \mathcal{M}_{i+1}}{h^2} + \frac{\mathcal{M}_i}{\Delta t}\right)\tilde{g}_i^{n+1}.$

If x_i is a cell end (i is odd),

$$\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_i = \frac{(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)^+\tilde{g}_{i-2}^{n+1} + (3\mathcal{M}_{i+2} - 4\mathcal{M}_{i+1} + 3\mathcal{M}_i)^+\tilde{g}_{i+2}^{n+1}}{8h^2};$$

if x_i is a cell center (i is even), $\quad \mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_i = 0.$

If x_i is a cell center, $\quad \mathcal{A}_a^-(\tilde{\mathbf{g}}^{n+1})_i = \frac{-(3\mathcal{M}_{i-1} + \mathcal{M}_{i+1})\tilde{g}_{i-1}^{n+1} - (\mathcal{M}_{i-1} + 3\mathcal{M}_{i+1})\tilde{g}_{i+1}^{n+1}}{4h^2};$

if x_i is a cell end, $\quad \mathcal{A}_a^-(\tilde{\mathbf{g}}^{n+1})_i = \frac{-(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)^-\tilde{g}_{i-2}^{n+1}}{8h^2}$

$$+ \frac{-(4\mathcal{M}_{i-2} + 12\mathcal{M}_i)\tilde{g}_{i-1}^{n+1} - (12\mathcal{M}_i + 4\mathcal{M}_{i+2})\tilde{g}_{i+1}^{n+1} - (3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2})^-\tilde{g}_{i+2}^{n+1}}{8h^2}.$$

We can easily verify that $(A_d + A^z)\mathbf{1} > 0$ for the following $\mathcal{A}^z$:

$$\text{if } x_i \text{ is a cell center}, \quad \mathcal{A}^z(\tilde{\mathbf{g}}^{n+1})_i = 0,$$

$$\text{if } x_i \text{ is an interior cell end}, \quad \mathcal{A}^z(\tilde{\mathbf{g}}^{n+1})_i$$

$$= \frac{-(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)^- \tilde{g}_{i-2}^{n+1} - [4\mathcal{M}_{i-2} + 12\mathcal{M}_i - (3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)^+]\tilde{g}_{i-1}^{n+1}}{8h^2}$$

$$+ \frac{-[12\mathcal{M}_i + 4\mathcal{M}_{i+2} - (3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2})^+]\tilde{g}_{i+1}^{n+1} - (3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2})^- \tilde{g}_{i+2}^{n+1}}{8h^2}.$$

We can also verify that $A^s := A_a^- - A^z \le 0$:

$$\text{If } x_i \text{ is a cell center}, \quad \mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_i = \frac{-(3\mathcal{M}_{i-1} + \mathcal{M}_{i+1})\tilde{g}_{i-1}^{n+1} - (\mathcal{M}_{i-1} + 3\mathcal{M}_{i+1})\tilde{g}_{i+1}^{n+1}}{4h^2},$$

If $x_i$is a cell end,

$$\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_i = \frac{-(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)^+ \tilde{g}_{i-1}^{n+1} - (3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2})^+ \tilde{g}_{i+1}^{n+1}}{8h^2}.$$

Now in order to verify $A^z A_d^{-1} A^s \ge A_a^+$ (entrywise inequality), we only need to compare nonzero coefficients in $\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_i$ and $\mathcal{A}^z\left(\mathcal{A}_d^{-1}[\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})]\right)_i$ for x_i being a cell end. When x_i is a cell end, $x_{i\pm1}$ are cell centers, and we have

$$\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_{i-1} = \frac{-(3\mathcal{M}_{i-2} + \mathcal{M}_i)\tilde{g}_{i-2}^{n+1} - (\mathcal{M}_{i-2} + 3\mathcal{M}_i)\tilde{g}_i^{n+1}}{4h^2},$$

$$A^s(\tilde{\mathbf{g}}^{n+1})_{i-2} = \frac{-(3\mathcal{M}_{i-4} - 4\mathcal{M}_{i-3} + 3\mathcal{M}_{i-2})^+ \tilde{g}_{i-3}^{n+1} - (3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)^+ \tilde{g}_{i-1}^{n+1}}{8h^2},$$

$$\mathcal{A}_d^{-1}[\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})]_{i-1} = \frac{h^2 \mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_{i-1}}{(\mathcal{M}_{i-2} + \mathcal{M}_i + h^2\mathcal{M}_{i-1}/\Delta t)} = \frac{-(3\mathcal{M}_{i-2} + \mathcal{M}_i)\tilde{g}_{i-2}^{n+1} - (\mathcal{M}_{i-2} + 3\mathcal{M}_i)\tilde{g}_i^{n+1}}{4(\mathcal{M}_{i-2} + \mathcal{M}_i + h^2\mathcal{M}_{i-1}/\Delta t)}.$$

It suffices to focus on the coefficient of $\tilde{g}_{i-2}^{n+1}$ in $\mathcal{A}^z(\mathcal{A}_d^{-1}[\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})])_i$ and the discussion for the coefficient of $\tilde{g}_{i+2}^{n+1}$ is similar. Notice that $\mathcal{A}_d^{-1}[\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})]_{i-2}$ will contribute nothing to the coefficient of $\tilde{g}_{i-2}^{n+1}$. So the coefficient of $\tilde{g}_{i-2}^{n+1}$ in $\mathcal{A}^z(\mathcal{A}_d^{-1}[\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})])_i$ is

$$\frac{(3\mathcal{M}_{i-2} + \mathcal{M}_i)(4\mathcal{M}_{i-2} + 12\mathcal{M}_i - (3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)^+)}{32h^2(\mathcal{M}_{i-2} + \mathcal{M}_i + h^2\mathcal{M}_{i-1}/\Delta t)}.$$

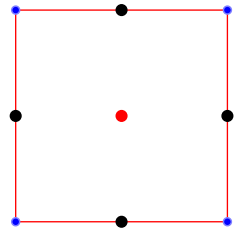


FIG. 2. Three types of grid points: red cell center, blue knots and black edge centers for a Q^2 finite element cell.

Thus to ensure $A_a^+ \leq A^z A_d^- A^s$, it suffices to have the following holds for any cell end x_i:

$$\frac{(3\mathcal{M}_{i-2}+\mathcal{M}_i)(4\mathcal{M}_{i-2}+12\mathcal{M}_i-(3\mathcal{M}_{i-2}-4\mathcal{M}_{i-1}+3\mathcal{M}_i)^+)}{32h^2(\mathcal{M}_{i-2}+\mathcal{M}_i+h^2\mathcal{M}_{i-1}/\Delta t)} \geq \frac{(3\mathcal{M}_{i-2}-4\mathcal{M}_{i-1}+3\mathcal{M}_i)^+}{8h^2}.$$

Equivalently, we need the following inequality holds for any cell center x_i:

$$\frac{(3\mathcal{M}_{i-1}+\mathcal{M}_{i+1})(4\mathcal{M}_{i-1}+12\mathcal{M}_{i+1}-(3\mathcal{M}_{i-1}-4\mathcal{M}_i+3\mathcal{M}_{i+1})^+)}{32h^2(\mathcal{M}_{i-1}+\mathcal{M}_{i+1}+h^2\mathcal{M}_i/\Delta t)} \geq \frac{(3\mathcal{M}_{i-1}-4\mathcal{M}_i+3\mathcal{M}_{i+1})^+}{8h^2}. \tag{34}$$

If $3\mathcal{M}_{i-1}-4\mathcal{M}_i+3\mathcal{M}_{i+1} \leq 0$ then (34) holds trivially. We only need to discuss the case $3\mathcal{M}_{i-1}-4\mathcal{M}_i+3\mathcal{M}_{i+1} > 0$, for which (34) becomes

$$(3\mathcal{M}_{i-1}+\mathcal{M}_{i+1})(\mathcal{M}_{i-1}+4\mathcal{M}_i+9\mathcal{M}_{i+1}) > 4\left(\mathcal{M}_{i-1}+\mathcal{M}_{i+1}+\frac{h^2}{\Delta t}\mathcal{M}_i\right)(3\mathcal{M}_{i-1}-4\mathcal{M}_i+3\mathcal{M}_{i+1}). \tag{35}$$

Let $a = \max\{\mathcal{M}_{i-1}, \mathcal{M}_i, \mathcal{M}_{i+1}\}$ and $b = \min\{\mathcal{M}_{i-1}, \mathcal{M}_i, \mathcal{M}_{i+1}\}$, a convenient sufficient condition to ensure (35) is

$$56b^2 > 4\left(2+\frac{h^2}{\Delta t}\right)a(6a-4b),$$

which is equivalent to $2+\frac{h^2}{\Delta t} < 14\frac{b^2}{6a^2-4ab}$.

So we have proven the first result for the variable coefficient case:

THEOREM 3.10 For the scheme (31) with $\mathcal{M}_i > 0$ its matrix representation A satisfies $A^{-1} \geq 0$ if (35) holds for any cell center x_i. A sufficient condition is to have the following constraints for each finite element cell $I_i = [x_{i-1}, x_{i+1}]$ (i is even):

$$2 + \frac{h^2}{\Delta t} < 7 \frac{1}{\max_{I_i} \mathcal{M}} \frac{\min_{I_i} \mathcal{M}^2}{3 \max_{I_i} \mathcal{M} - 2 \min_{I_i} \mathcal{M}}, \tag{36}$$

where

$$\max_{I_i} \mathcal{M} := \max\{\mathcal{M}_{i-1}, \mathcal{M}_i, \mathcal{M}_{i+1}\}, \quad \min_{I_i} \mathcal{M} := \min\{\mathcal{M}_{i-1}, \mathcal{M}_i, \mathcal{M}_{i+1}\}.$$

REMARK 3.11 Note that for a smooth function $\mathcal{M}$ the mesh and time step constraints (36) are possible to achieve because the right-hand side of (36) will converge to 7 as h goes to zero. Furthermore, for fixed h, the condition (36) gives a lower bound on Δt (not an upper bound).

3.6 *The fourth-order scheme in two dimensions*

Assume the domain is $\Omega = [-L, L] \times [-L, L]$ with a uniform $N \times N$ grid point with spacing h, obtained from all 3×3 Gauss–Lobatto points on a uniform rectangular mesh with $k \times k$ cells. Thus $N = 2k + 1$. Let $\mathbf{g}$ be an $N \times N$ matrix with g_{ij} denoting the point value at the (i,j) grid point. For the Q^2 finite element method on uniform rectangular meshes there are three types of grid point values, see Fig. 2.

Let $A = W^{-1}S + \frac{1}{\Delta t}M^n$ and $\mathcal{A} : \mathbb{R}^{N\times N} \longrightarrow R^{N\times N}$ be the scheme operator, i.e., (31) can be written as $\mathcal{A}(\tilde{\mathbf{g}}^{n+1})_{ij} = \frac{1}{\Delta t}\mathcal{M}_{ij} g^n_{ij}$. With the same *ghost point values* notation as in Section 3.3, following the derivations in Li & Zhang (2020a), the scheme can be explicitly written as:

if x_{ij} is a cell center, $\quad \mathcal{A}_d(\tilde{\mathbf{g}}^{n+1})_{ij} = \left(\dfrac{\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j} + \mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1}}{h^2} + \dfrac{1}{\Delta t}\mathcal{M}_{ij} \right) \tilde{g}^{n+1}_{ij}$;

if x_{ij} is an edge center for an edge parallel to y -axis,

$$\mathcal{A}_d(\tilde{\mathbf{g}}^{n+1})_{ij} = \left(\frac{(\mathcal{M}_{i-2,j} + 4\mathcal{M}_{i-1,j} + 18\mathcal{M}_{ij} + 4\mathcal{M}_{i+1,j} + \mathcal{M}_{i+2,j}) + 8(\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})}{8h^2} + \frac{1}{\Delta t}\mathcal{M}_{ij} \right) \tilde{g}^{n+1}_{ij};$$

if x_{ij} is an edge center for an edge parallel to x-axis,

$$\mathcal{A}_d(\tilde{\mathbf{g}}^{n+1})_{ij} = \left(\frac{(\mathcal{M}_{i,j-2} + 4\mathcal{M}_{i,j-1} + 18\mathcal{M}_{ij} + 4\mathcal{M}_{i,j+1} + \mathcal{M}_{i,j+2}) + 8(\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j})}{8h^2} + \frac{1}{\Delta t}\mathcal{M}_{ij} \right) \tilde{g}^{n+1}_{ij};$$

if x_{ij} is a knot,

$$\mathcal{A}_d(\tilde{\mathbf{g}}^{n+1})_{ij} = \left(\frac{\mathcal{M}_{i-2,j} + 4\mathcal{M}_{i-1,j} + 18\mathcal{M}_{ij} + 4\mathcal{M}_{i+1,j} + \mathcal{M}_{i+2,j}}{8h^2} \right.$$
$$\left. + \frac{(\mathcal{M}_{i,j-2} + 4\mathcal{M}_{i,j-1} + 18\mathcal{M}_{ij} + 4\mathcal{M}_{i,j+1} + \mathcal{M}_{i,j+2})}{8h^2} + \frac{1}{\Delta t}\mathcal{M}_{ij} \right) \tilde{g}^{n+1}_{ij}.$$

For the operator $\mathcal{A}_a^+$ it is given as

if x_{ij} is a cell center, $\quad \mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij} = 0$;

if x_{ij} is an edge center for an edge parallel to y-axis,

$$\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij} = \frac{(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})^+\tilde{g}_{i-2,j}^{n+1} + (3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})^+\tilde{g}_{i+2,j}^{n+1}}{8h^2};$$

if x_{ij} is an edge center for an edge parallel to x-axis,

$$\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij} = \frac{(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})^+\tilde{g}_{i,j-2}^{n+1} + (3\mathcal{M}_{i,j+2} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j})^+\tilde{g}_{i,j+2}^{n+1}}{8h^2};$$

if x_{ij} is a knot, $\quad \mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij}$

$$= \frac{(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})^+\tilde{g}_{i-2,j}^{n+1} + (3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})^+\tilde{g}_{i+2,j}^{n+1}}{8h^2}$$
$$+ \frac{(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})^+\tilde{g}_{i,j-2}^{n+1} + (3\mathcal{M}_{i,j+2} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j})^+\tilde{g}_{i,j+2}^{n+1}}{8h^2}.$$

We consider the following $A^z \le 0$ and it is straightforward to see $(A_d + A^z)\mathbf{1} > 0$:

if x_{ij} is a cell center, $\quad \mathcal{A}^z(\tilde{\mathbf{g}}^{n+1})_{ij} = 0$;

if x_{ij} is an edge center for an edge parallel to y-axis, $\quad \mathcal{A}^z(\tilde{\mathbf{g}}^{n+1})_{ij}$

$$= \frac{-(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})^-\tilde{g}_{i-2,j}^{n+1} - [4\mathcal{M}_{i-2,j} + 12\mathcal{M}_{i,j} - (3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})^+]\tilde{g}_{i-1,j}^{n+1}}{8h^2}$$
$$+ \frac{-[12\mathcal{M}_{i,j} + 4\mathcal{M}_{i+2,j} - (3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})^+]\tilde{g}_{i+1,j}^{n+1} - (3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})^-\tilde{g}_{i+2,j}^{n+1}}{8h^2};$$

if x_{ij} is an edge center for an edge parallel to x-axis,

$$\mathcal{A}^z(\tilde{\mathbf{g}}^{n+1})_{ij} = \frac{-(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})^-\tilde{g}_{i,j-2}^{n+1} - [4\mathcal{M}_{i,j-2} + 12\mathcal{M}_{i,j} - (3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})^+]\tilde{g}_{i,j-1}^{n+1}}{8h^2}$$
$$+ \frac{-[12\mathcal{M}_{i,j} + 4\mathcal{M}_{i,j+2} - (3\mathcal{M}_{i,j+2} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j})^+]\tilde{g}_{i,j+1}^{n+1} - (3\mathcal{M}_{i,j+2} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j})^-\tilde{g}_{i,j+2}^{n+1}}{8h^2};$$

if x_{ij} is a knot, $\quad \mathcal{A}^z(\tilde{\mathbf{g}}^{n+1})_{ij}$

$$= \frac{-(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})^-\tilde{g}_{i-2,j}^{n+1} - [4\mathcal{M}_{i-2,j} + 12\mathcal{M}_{i,j} - (3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})^+]\tilde{g}_{i-1,j}^{n+1}}{8h^2}$$
$$+ \frac{-[12\mathcal{M}_{i,j} + 4\mathcal{M}_{i+2,j} - (3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})^+]\tilde{g}_{i+1,j}^{n+1} - (3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})^-\tilde{g}_{i+2,j}^{n+1}}{8h^2}$$
$$+ \frac{-(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})^-\tilde{g}_{i,j-2}^{n+1} - [4\mathcal{M}_{i,j-2} + 12\mathcal{M}_{i,j} - (3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})^+]\tilde{g}_{i,j-1}^{n+1}}{8h^2}$$
$$+ \frac{-[12\mathcal{M}_{i,j} + 4\mathcal{M}_{i,j+2} - (3\mathcal{M}_{i,j+2} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j})^+]\tilde{g}_{i,j+1}^{n+1} - (3\mathcal{M}_{i,j+2} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j})^-\tilde{g}_{i,j+2}^{n+1}}{8h^2}.$$

Then $A^s = A_a^- - A^z$ is given as:

$$\text{if } x_i \text{ is a cell center,} \quad \mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_{ij} = -\frac{(3\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j})\tilde{g}^{n+1}_{i-1,j} + (\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i+1,j})\tilde{g}^{n+1}_{i+1,j}}{4h^2}$$

$$- \frac{(3\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})\tilde{g}^{n+1}_{i,j-1} + (\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j+1})\tilde{g}^{n+1}_{i,j+1}}{4h^2};$$

$$\text{if } x_{ij} \text{ is an edge center for an edge parallel to } y\text{-axis,} \quad \mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_{ij}$$

$$= \frac{-(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})^+\tilde{g}^{n+1}_{i-1,j} - (3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})^+\tilde{g}^{n+1}_{i+1,j}}{8h^2}$$

$$+ \frac{-(3\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})\tilde{g}^{n+1}_{i,j-1} - (\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j+1})\tilde{g}^{n+1}_{i,j+1}}{4h^2};$$

$$\text{if } x_{ij} \text{ is an edge center for an edge parallel to } x\text{-axis,} \quad \mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_{ij}$$

$$= \frac{-(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})^+\tilde{g}^{n+1}_{i,j-1} - (3\mathcal{M}_{i,j+2} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j})^+\tilde{g}^{n+1}_{i,j+1}}{8h^2}$$

$$+ \frac{-(3\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j})\tilde{g}^{n+1}_{i-1,j} - (\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i+1,j})\tilde{g}^{n+1}_{i+1,j}}{4h^2};$$

$$\text{if } x_{ij} \text{ is a knot,} \quad \mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_{ij}$$

$$= \frac{-(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})^+\tilde{g}^{n+1}_{i-1,j} - (3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})^+\tilde{g}^{n+1}_{i+1,j}}{8h^2}$$

$$+ \frac{-(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})^+\tilde{g}^{n+1}_{i,j-1} - (3\mathcal{M}_{i,j+2} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j})^+\tilde{g}^{n+1}_{i,j+1}}{8h^2}.$$

For the positive off-diagonal entries $\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij}$ is nonzero only for x_{ij} being an edge center or a cell center. Thus to verify $A_a^+ \le A^z A_d^{-1} A^s$ it suffices to compare $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})\right)\right]_{ij}$ with $\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij}$ for x_{ij} being an edge center or a cell center.

If x_{ij} is an edge center for an edge parallel to y-axis then $x_{i\pm1,j}$ are cell centers. Since everything here has a symmetric structure we only need to compare the coefficients of $\tilde{g}^{n+1}_{i-2,j}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})\right)\right]_{ij}$ and $\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij}$, and the comparison for the coefficients of $\tilde{g}^{n+1}_{i+2,j}$ will be similar.

$$\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_{i-1,j} = -\frac{(3\mathcal{M}_{i-2,j} + \mathcal{M}_{ij})\tilde{g}^{n+1}_{i-2,j} + (\mathcal{M}_{i-2,j} + 3\mathcal{M}_{ij})\tilde{g}^{n+1}_{ij}}{4h^2}$$

$$- \frac{(3\mathcal{M}_{i-1,j-1} + \mathcal{M}_{i-1,j+1})\tilde{g}^{n+1}_{i-1,j-1} + (\mathcal{M}_{i-1,j-1} + 3\mathcal{M}_{i-1,j+1})\tilde{g}^{n+1}_{i-1,j+1}}{4h^2},$$

$$\mathcal{A}_d^{-1}[\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})]_{i-1,j} = -\frac{(3\mathcal{M}_{i-2,j}+\mathcal{M}_{ij})\tilde{g}_{i-2,j}^{n+1}+(\mathcal{M}_{i-2,j}+3\mathcal{M}_{ij})\tilde{g}_{i,j}^{n+1}}{4(\mathcal{M}_{i-2,j}+\mathcal{M}_{ij}+\mathcal{M}_{i-1,j+1}+\mathcal{M}_{i-1,j-1}+h^2\frac{1}{\Delta t}\mathcal{M}_{i-1,j})}$$
$$-\frac{(3\mathcal{M}_{i-1,j-1}+\mathcal{M}_{i-1,j+1})\tilde{g}_{i-1,j-1}^{n+1}+(\mathcal{M}_{i-1,j-1}+3\mathcal{M}_{i-1,j+1})\tilde{g}_{i-1,j+1}^{n+1}}{4(\mathcal{M}_{i-2,j}+\mathcal{M}_{ij}+\mathcal{M}_{i-1,j+1}+\mathcal{M}_{i-1,j-1}+h^2\frac{1}{\Delta t}\mathcal{M}_{i-1,j})}.$$

Since the coefficient of $\tilde{g}_{i-2,j}^{n+1}$ in $\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij}$ is $(3\mathcal{M}_{i-2,j}-4\mathcal{M}_{i-1,j}+3\mathcal{M}_{ij})^+/(8h^2)$, we only need to discuss the case $3\mathcal{M}_{i-2,j}-4\mathcal{M}_{i-1,j}+3\mathcal{M}_{ij} > 0$, for which the coefficient of $\tilde{g}_{i-2,j}^{n+1}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})\right)\right]_{ij}$ becomes

$$\frac{\mathcal{M}_{i-2,j}+4\mathcal{M}_{i-1,j}+9\mathcal{M}_{ij}}{8h^2}\frac{(3\mathcal{M}_{i-2,j}+\mathcal{M}_{ij})}{4(\mathcal{M}_{i-2,j}+\mathcal{M}_{ij}+\mathcal{M}_{i-1,j+1}+\mathcal{M}_{i-1,j-1}+h^2\frac{1}{\Delta t}\mathcal{M}_{i-1,j})}.$$

To ensure the coefficient of $\tilde{g}_{i-2,j}^{n+1}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})\right)\right]_{ij}$ is no less than the coefficient of $\tilde{g}_{i-2,j}^{n+1}$ in $\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij}$, we need

$$\frac{(\mathcal{M}_{i-2,j}+4\mathcal{M}_{i-1,j}+9\mathcal{M}_{ij})(3\mathcal{M}_{i-2,j}+\mathcal{M}_{ij})}{32h^2(\mathcal{M}_{i-2,j}+\mathcal{M}_{ij}+\mathcal{M}_{i-1,j+1}+\mathcal{M}_{i-1,j-1}+h^2\frac{1}{\Delta t}\mathcal{M}_{i-1,j})} \geq \frac{3\mathcal{M}_{i-2,j}-4\mathcal{M}_{i-1,j}+3\mathcal{M}_{ij}}{8h^2}.$$

Similar to the one-dimensional case it suffices to require

$$\frac{(\mathcal{M}_{i-2,j}+4\mathcal{M}_{i-1,j}+9\mathcal{M}_{ij})(3\mathcal{M}_{i-2,j}+\mathcal{M}_{ij})}{4(\mathcal{M}_{i-2,j}+\mathcal{M}_{ij}+\mathcal{M}_{i-1,j+1}+\mathcal{M}_{i-1,j-1}+h^2\frac{1}{\Delta t}\mathcal{M}_{i-1,j})} > 3\mathcal{M}_{i-2,j}-4\mathcal{M}_{i-1,j}+3\mathcal{M}_{ij}.$$

Equivalently, we need the following inequality holds for any cell center x_{ij}:

$$\frac{(\mathcal{M}_{i-1,j}+4\mathcal{M}_{i,j}+9\mathcal{M}_{i+1,j})(3\mathcal{M}_{i-1,j}+\mathcal{M}_{i+1,j})}{4(\mathcal{M}_{i-1,j}+\mathcal{M}_{i+1,j}+\mathcal{M}_{i,j+1}+\mathcal{M}_{i,j-1}+h^2\frac{1}{\Delta t}\mathcal{M}_{i,j})} > 3\mathcal{M}_{i-1,j}-4\mathcal{M}_{i,j}+3\mathcal{M}_{i+1,j}. \tag{37a}$$

Notice that (37a) was derived for comparing $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})\right)\right]_{ij}$ and $\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij}$ for x_{ij} being an edge center of an edge parallel to y-axis. If x_{ij} is an edge center of an edge parallel to x-axis then we can derive a similar constraint:

$$\frac{(\mathcal{M}_{i,j-1}+4\mathcal{M}_{i,j}+9\mathcal{M}_{i,j+1})(3\mathcal{M}_{i,j-1}+\mathcal{M}_{i,j+1})}{4(\mathcal{M}_{i,j-1}+\mathcal{M}_{i,j+1}+\mathcal{M}_{i+1,j}+\mathcal{M}_{i-1,j}+h^2\frac{1}{\Delta t}\mathcal{M}_{i,j})} > 3\mathcal{M}_{i,j-1}-4\mathcal{M}_{i,j}+3\mathcal{M}_{i,j+1}. \tag{37b}$$

If x_{ij} is a knot then $x_{i\pm1,j}$ are edge centers for an edge parallel to x-axis. Since everything here has a symmetric structure we only need to compare the coefficients of $\tilde{g}_{i-2,j}^{n+1}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})\right)\right]_{ij}$ and

$\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij}$, and the comparison for the coefficients of $\tilde{g}_{i+2,j}^{n+1}$, $\tilde{g}_{i,j-2}^{n+1}$ and $\tilde{g}_{i,j+2}^{n+1}$ will be similar.

$$\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})_{i-1,j} = \frac{-(3\mathcal{M}_{i-2,j}+\mathcal{M}_{i,j})\tilde{g}_{i-2,j}^{n+1} - (\mathcal{M}_{i-2,j}+3\mathcal{M}_{i,j})\tilde{g}_{i,j}^{n+1}}{4h^2}$$
$$+ \frac{-(3\mathcal{M}_{i-1,j-2}-4\mathcal{M}_{i-1,j-1}+3\mathcal{M}_{i-1,j})^+\tilde{g}_{i-1,j-1}^{n+1} - (3\mathcal{M}_{i-1,j+2}-4\mathcal{M}_{i-1,j+1}+3\mathcal{M}_{i-1,j})^+\tilde{g}_{i-1,j+1}^{n+1}}{8h^2}$$

$$\mathcal{A}_d^{-1}[\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})]_{i-1,j}$$
$$= \frac{-(3\mathcal{M}_{i-2,j}+\mathcal{M}_{i,j})\tilde{g}_{i-2,j}^{n+1} - (\mathcal{M}_{i-2,j}+3\mathcal{M}_{i,j})\tilde{g}_{i,j}^{n+1}}{\frac{1}{2}(\mathcal{M}_{i-1,j-2}+4\mathcal{M}_{i-1,j-1}+18\mathcal{M}_{i-1,j}+4\mathcal{M}_{i-1,j+1}+\mathcal{M}_{i-1,j+2})+4(\mathcal{M}_{i-2,j}+\mathcal{M}_{i,j})+4h^2\frac{1}{\Delta t}\mathcal{M}_{i-1,j}}$$
$$+ \frac{-(3\mathcal{M}_{i-1,j-2}-4\mathcal{M}_{i-1,j-1}+3\mathcal{M}_{i-1,j})^+\tilde{g}_{i-1,j-1}^{n+1} - (3\mathcal{M}_{i-1,j+2}-4\mathcal{M}_{i-1,j+1}+3\mathcal{M}_{i-1,j})^+\tilde{g}_{i-1,j+1}^{n+1}}{(\mathcal{M}_{i-1,j-2}+4\mathcal{M}_{i-1,j-1}+18\mathcal{M}_{i-1,j}+4\mathcal{M}_{i-1,j+1}+\mathcal{M}_{i-1,j+2})+8(\mathcal{M}_{i-2,j}+\mathcal{M}_{i,j})+8h^2\frac{1}{\Delta t}\mathcal{M}_{i-1,j}}.$$

For the same reason as above we still only consider the case where $3\mathcal{M}_{i-2,j}-4\mathcal{M}_{i-1,j}+3\mathcal{M}_{ij}>0$. So the coefficient of $\tilde{g}_{i-2,j}^{n+1}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})\right)\right]_{ij}$ is

$$\frac{1}{4h^2}\frac{(\mathcal{M}_{i-2,j}+4\mathcal{M}_{i-1,j}+9\mathcal{M}_{ij})(3\mathcal{M}_{i-2,j}+\mathcal{M}_{i,j})}{(\mathcal{M}_{i-1,j-2}+4\mathcal{M}_{i-1,j-1}+18\mathcal{M}_{i-1,j}+4\mathcal{M}_{i-1,j+1}+\mathcal{M}_{i-1,j+2})+8(\mathcal{M}_{i-2,j}+\mathcal{M}_{i,j})+8\frac{1}{\Delta t}\mathcal{M}_{i-1,j}h^2}.$$

To ensure the coefficient of $\tilde{g}_{i-2,j}^{n+1}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\tilde{\mathbf{g}}^{n+1})\right)\right]_{ij}$ is no less than the coefficient of $\tilde{g}_{i-2,j}^{n+1}$ in $\mathcal{A}_a^+(\tilde{\mathbf{g}}^{n+1})_{ij}$, we only need

$$\frac{2(\mathcal{M}_{i-2,j}+4\mathcal{M}_{i-1,j}+9\mathcal{M}_{ij})(3\mathcal{M}_{i-2,j}+\mathcal{M}_{i,j})}{(\mathcal{M}_{i-1,j-2}+4\mathcal{M}_{i-1,j-1}+18\mathcal{M}_{i-1,j}+4\mathcal{M}_{i-1,j+1}+\mathcal{M}_{i-1,j+2})+8(\mathcal{M}_{i-2,j}+\mathcal{M}_{i,j})+8\frac{1}{\Delta t}\mathcal{M}_{i-1,j}h^2}$$
$$> 3\mathcal{M}_{i-2,j}-4\mathcal{M}_{i-1,j}+3\mathcal{M}_{ij}.$$

Equivalently, we need the following inequality holds for any edge center x_{ij} for an edge parallel to x-axis:

$$\frac{2(\mathcal{M}_{i-1,j}+4\mathcal{M}_{i,j}+9\mathcal{M}_{i+1,j})(3\mathcal{M}_{i-1,j}+\mathcal{M}_{i+1,j})}{(\mathcal{M}_{i,j-2}+4\mathcal{M}_{i,j-1}+18\mathcal{M}_{i,j}+4\mathcal{M}_{i,j+1}+\mathcal{M}_{i,j+2})+8(\mathcal{M}_{i-1,j}+\mathcal{M}_{i+1,j})+8c_{i,j}h^2}$$
$$> 3\mathcal{M}_{i-1,j}-4\mathcal{M}_{i,j}+3\mathcal{M}_{i+1,j}. \tag{38a}$$

We also need the following inequality holds for any edge center x_{ij} for an edge parallel to y-axis:

$$\frac{2(\mathcal{M}_{i,j-1}+4\mathcal{M}_{i,j}+9\mathcal{M}_{i,j+1})(3\mathcal{M}_{i,j-1}+\mathcal{M}_{i,j-1})}{(\mathcal{M}_{i-2,j}+4\mathcal{M}_{i-1,j}+18\mathcal{M}_{i,j}+4\mathcal{M}_{i+1,j}+\mathcal{M}_{i+2,j})+8(\mathcal{M}_{i,j-1}+\mathcal{M}_{i,j+1})+8c_{i,j}h^2}$$
$$> 3\mathcal{M}_{i,j-1}-4\mathcal{M}_{i,j}+3\mathcal{M}_{i,j+1}. \tag{38b}$$

We have a similar result to the one-dimensional case as follows:

THEOREM 3.12 For the scheme (31) its matrix representation A satisfies $A^{-1} \geq 0$ if (37) holds for any cell center x_{ij}, (38a) holds for x_{ij} being any edge center of an edge parallel to x-axis and (38b) holds for x_{ij} being any edge center of an edge parallel to y-axis.

THEOREM 3.13 For the scheme (31) its matrix representation A satisfies $A^{-1} \geq 0$ if the following mesh constraint is achieved for all edge centers x_{ij}:

$$\frac{11}{2} + \frac{h^2}{\Delta t} < 7 \frac{1}{\max_{J_{ij}} \mathcal{M}} \frac{\min_{J_{ij}} \mathcal{M}^2}{3 \max_{J_{ij}} \mathcal{M} - 2 \min_{J_{ij}} \mathcal{M}}, \tag{39}$$

where J_{ij} is the union of two finite element cells: if x_{ij} is an edge center of an edge parallel to x-axis, then $J_{ij} = [x_{i-1}, x_{i+1}] \times [y_{j-2}, y_{j+2}]$; if x_{ij} is an edge center of an edge parallel to y-axis, then $J_{ij} = [x_{i-2}, x_{i+2}] \times [y_{j-1}, y_{j+1}]$. Here the maximum and minimum of $\mathcal{M}$ are those of grid point values of $\mathcal{M}$ in J_{ij}.

REMARK 3.14 Similarly as the one-dimensional case, for smooth $\mathcal{M}$, the constraint (39) can be satisfied for small h.

4. Positivity and energy dissipation

In this section we prove a few properties of the proposed scheme (22), among which positivity and energy dissipation are the most important ones. First of all we rewrite (22) as

$$A\tilde{\mathbf{g}}^{n+1} = \mathbf{g}^n, \quad A := I + \Delta t (M^n)^{-1} W^{-1} S. \tag{40}$$

From the previous section we know that the matrix A is invertible and $A^{-1} \geq 0$ under suitable mesh size and time step constraints. Specifically, the second-order scheme is always monotone $A^{-1} \geq 0$ (entrywise inequality) for any mesh size and time step. For the fourth-order scheme, assume that the mesh size and time step satisfy the constraints (36) and (39) in one and two dimensions, respectively, we also have $A^{-1} \geq 0$.

4.1 *Conservation, steady state and positivity*

It is straightforward to verify the following properties:

1. *Mass conservation of* ρ. Multiplying $\mathbf{1}^T W M^n$ from the left on both sides of (40) and using $\mathbf{1}^T S = \mathbf{0}^T$ gives

$$\mathbf{1}^T W M^n \tilde{\mathbf{g}}^{n+1} = \mathbf{1}^T W M^n \mathbf{g}^n,$$

which is

$$\mathbf{1}^T W \boldsymbol{\rho}^{n+1} = \mathbf{1}^T W \boldsymbol{\rho}^n,$$

or equivalently,

$$\sum_i w_i \rho_i^{n+1} = \sum_i w_i \rho_i^n.$$

2. *Mass conservation of c*. By setting $v_h \equiv 1$ in (24) we get $\alpha \langle c_h^n, 1\rangle = \langle \rho_h^n, 1\rangle$, thus

$$\alpha \sum_i w_i c_i^n = \sum_i w_i \rho_i^n.$$

3. *Steady state preserving*. If $\mathbf{g}^n = C\mathbf{1}$ for some constant C then using $S\mathbf{1} = \mathbf{0}$ it can be easily seen that $\tilde{\mathbf{g}}^{n+1} = C\mathbf{1}$ is the unique solution to (40). In terms of the ρ variable this implies that

$$\rho_i^n = C\mathcal{M}_i^n, \forall i \Longrightarrow \rho_i^{n+1} = C\mathcal{M}_i^n, \forall i.$$

4. *Positivity of* ρ. If $\rho_i^n > 0$ for every i then $g_i^n = \rho_i^n / \mathcal{M}_i^n > 0$ for every i. When $A^{-1} \geq 0$ holds we have $\tilde{g}_i^{n+1} > 0$, consequently $\rho_i^{n+1} = \mathcal{M}_i^n \tilde{g}_i^{n+1} > 0$ for every i.

5. *Positivity of c*. All discussion in Section 3 applies to the scheme (24) with $\alpha > 0$ and suitable boundary conditions. Even though we only consider Neumann-type boundary condition in this paper the results hold also for Dirichlet-type boundary conditions. In particular, the second-order scheme is monotone. By setting $\mathcal{M} \equiv 1$ and $\Delta t = \frac{1}{\alpha}$ in Theorem 3.10 and Theorem 3.13, the fourth-order scheme is also monotone if $\alpha h^2 \leq 5$ in one dimension and $\alpha h^2 \leq \frac{3}{2}$ in two dimensions. When monotonicity in (24) holds positivity of c is implied by positivity of ρ.

4.2 *Energy dissipation*

In this subsection we show that the fully discrete scheme (40) decays energy. Following the continuous counterpart (12) we define the discrete energy as

$$E^n := \left\langle \rho^n \log \frac{\rho^n}{\mathcal{M}^n} - \rho^n + \frac{1}{2} c^n \rho^n, 1 \right\rangle = \sum_i w_i \left(\rho_i^n \log \frac{\rho_i^n}{\mathcal{M}_i^n} - \rho_i^n + \frac{1}{2} c_i^n \rho_i^n \right). \tag{41}$$

Note that by using c_i^n we consider the Keller–Segel equation directly. In the Fokker–Planck case the last term $\frac{1}{2} c_i^n \rho_i^n$ in E^n is zero.

THEOREM 4.1 Assume monotonicity holds for scheme (40), i.e., $A^{-1} \geq 0$, for the energy defined in (41) we have $E^{n+1} \leq E^n$.

Proof. First of all,

$$\begin{aligned}
&E^{n+1} - E^n \\
&= \sum_i w_i \left(\rho_i^{n+1} \log \frac{\rho_i^{n+1}}{\mathcal{M}_i^{n+1}} - \rho_i^{n+1} + \frac{1}{2} c_i^{n+1} \rho_i^{n+1} \right) - \sum_i w_i \left(\rho_i^n \log \frac{\rho_i^n}{\mathcal{M}_i^n} - \rho_i^n + \frac{1}{2} c_i^n \rho_i^n \right) \\
&= \sum_i w_i \left(\rho_i^{n+1} \log \frac{\rho_i^{n+1}}{\mathcal{M}_i^{n+1}} + \frac{1}{2} c_i^{n+1} \rho_i^{n+1} \right) - \sum_i w_i \left(\rho_i^n \log \frac{\rho_i^n}{\mathcal{M}_i^n} + \frac{1}{2} c_i^n \rho_i^n \right) \\
&= I + II,
\end{aligned}$$

where we used mass conservation in the second equality and

$$\begin{aligned}
I &:= \sum_i w_i \rho_i^{n+1} \log \frac{\rho_i^{n+1}}{\mathcal{M}_i^n} - \sum_i w_i \rho_i^n \log \frac{\rho_i^n}{\mathcal{M}_i^n}, \\
II &:= \sum_i w_i \left(\rho_i^{n+1} c_i^n - \frac{1}{2} \rho_i^{n+1} c_i^{n+1} - \frac{1}{2} \rho_i^n c_i^n \right).
\end{aligned}$$

On the other hand, it is easy to see $A^{-1}\mathbf{1} = \mathbf{1}$, since $A\mathbf{1} = \mathbf{1}$. Let a^{ij} be the entries of A^{-1}, then $\sum_j a^{ij} = 1$ and $a^{ij} \geq 0$ for all i, j if the monotonicity holds. Furthermore, since M^n and W are diagonal matrices, $M^n W = W M^n$, thus $\mathbf{1}^T M^n W A = \mathbf{1}^T M^n W (I + \Delta t (M^n)^{-1} W^{-1} S) = \mathbf{1}^T M^n W$. So we have $\mathbf{1}^T M^n W A^{-1} = \mathbf{1}^T M^n W$, which is $\sum_i \mathcal{M}_i^n w_i a^{ij} = \mathcal{M}_j^n w_j$ componentwise.

The above discussion implies that $\tilde{g}_i^{n+1} = \sum_j a^{ij} g_j^n$ is a convex combination. The function $x \log x$ is convex, so by Jensen's inequality,

$$\tilde{g}_i^{n+1} \log(\tilde{g}_i^{n+1}) \leq \sum_j a^{ij} g_j^n \log(g_j^n).$$

Then

$$\begin{aligned}
&\sum_i w_i \rho_i^{n+1} \log(\rho_i^{n+1}/\mathcal{M}_i^n) = \sum_i w_i \mathcal{M}_i^n \tilde{g}_i^{n+1} \log(\tilde{g}_i^{n+1}) \leq \sum_i w_i \mathcal{M}_i^n \sum_j a^{ij} g_j^n \log(g_j^n) \\
&= \sum_j \left(\sum_i a^{ij} w_i \mathcal{M}_i^n \right) g_j^n \log(g_j^n) = \sum_j w_j \mathcal{M}_j^n g_j^n \log(g_j^n) = \sum_i w_i \rho_i^n \log(\rho_i^n/\mathcal{M}_i^n).
\end{aligned}$$

We thus proved $I \leq 0$. The proof is done if it is the Fokker–Planck equation.

If it is the Keller–Segel equation we still need to show $II \leq 0$. Recall that we use the scheme (24) for c:

$$\langle \nabla c_h^n, \nabla v_h \rangle + \alpha \langle c_h^n, v_h \rangle = \langle \rho^n, v_h \rangle, \quad \forall v_h \in V^h. \tag{42}$$

At t^{n+1} this is

$$\langle \nabla c_h^{n+1}, \nabla v_h \rangle + \alpha \langle c_h^{n+1}, v_h \rangle = \langle \rho^{n+1}, v_h \rangle, \quad \forall v_h \in V^h. \tag{43}$$

Subtracting (42) from (43) gives

$$\langle \nabla(c_h^{n+1} - c_h^n), \nabla v_h \rangle + \alpha \langle c_h^{n+1} - c_h^n, v_h \rangle = \langle \rho^{n+1} - \rho^n, v_h \rangle, \quad \forall v_h \in V^h.$$

By setting $v_h = -(c_h^{n+1} - c_h^n) \in V^h$ we obtain

$$-\langle \rho^{n+1} - \rho^n, c^{n+1} - c^n \rangle = -\langle \nabla(c^{n+1} - c^n), \nabla(c^{n+1} - c^n) \rangle - \alpha \langle c^{n+1} - c^n, c^{n+1} - c^n \rangle \le 0.$$

On the other hand, choosing $v_h = c_h^{n+1}$ in (42) and $v_h = c_h^n$ in (43) and subtracting both, we obtain

$$\langle \rho^n, c_h^{n+1} \rangle = \langle \rho^{n+1}, c_h^n \rangle.$$

Therefore,

$$II = \langle \rho^{n+1}, c_h^n \rangle - \frac{1}{2}\langle \rho^n, c_h^n \rangle - \frac{1}{2}\langle \rho^{n+1}, c_h^{n+1} \rangle = -\frac{1}{2}\langle \rho^{n+1} - \rho^n, c_h^{n+1} - c_h^n \rangle \le 0.$$

□

5. Numerical tests

In this section we provide numerical examples to demonstrate the performance of the proposed schemes. We will mainly focus on the Keller–Segel equation as it is more challenging than the Fokker–Planck equation. But one example about the Fokker–Planck equation will be included.

We consider the Keller–Segel system in a square domain Ω with a source term:

$$\begin{cases} \partial_t \rho = \Delta \rho - \nabla \cdot (\rho \nabla c) + f(x, y), \\ -\Delta c + c = \rho, \end{cases}$$

with homogeneous Neumann boundary conditions $\nabla \rho \cdot \mathbf{n}|_{\partial\Omega} = \nabla c \cdot \mathbf{n}|_{\partial\Omega} = 0$. It is straightforward to verify that the system above is equivalent to

$$\begin{cases} \partial_t \rho = \nabla \cdot (\mathcal{M} \nabla \frac{\rho}{\mathcal{M}}) + f(x, y), \quad \mathcal{M} := e^c, \\ -\Delta c + c = \rho, \end{cases} \tag{44}$$

with boundary conditions $\nabla c \cdot \mathbf{n}|_{\partial\Omega} = 0$ and $\nabla \frac{\rho}{\mathcal{M}} \cdot \mathbf{n}|_{\partial\Omega} = 0$. We test the second-order and fourth-order semi-implicit finite difference schemes for solving (44).

5.1 *Accuracy test for the Keller–Segel system with a source term*

The proposed semi-implicit schemes can be at most first order accurate in time. For testing the spatial accuracy we consider an initial condition $\rho(0, x, y) = 3 \cos x \cos y + 3$, $c(0, x, y) = \cos x \cos y + 3$ on $\Omega = (0, \pi) \times (0, \pi)$ and a source term $f(x, y) = -3 \cos(2x) \cos^2 y - 3 \cos^2 x \cos(2y)$, so that the exact solution is a steady state solution. The time step is set as $\Delta t = \Delta x$ and errors at $T = 1$ are given in

TABLE 1 *Accuracy test for the Keller–Segel system with a source term*

FD Grid	The second-order scheme				The fourth-order scheme			
	l^2 error	Order	l^∞ error	Order	l^2 error	Order	l^∞ error	Order
9×9	2.09E−1	–	2.51E−1	–	1.37E−2	–	1.08E−2	–
17×17	4.11E−2	2.34	6.82E−2	1.89	7.70E−4	4.16	1.32E−3	3.03
33×33	8.19E−3	2.33	1.70E−2	2.00	4.52E−5	4.09	9.72E−5	3.76
65×65	1.77E−3	2.21	4.29E−3	1.99	2.76E−6	4.03	6.41E−6	3.92
129×129	4.04E−4	2.13	1.08E−3	1.99	1.71E−7	4.01	4.09E−7	3.97

Table 1 where l^2 error is defined as

$$\sqrt{\Delta x \Delta y \sum_i \sum_j |u_{ij} - u(x_i, y_j)|^2}$$

with u_{ij} and $u(x, y)$ denoting the numerical and exact solutions, respectively. We observe the expected order of spatial accuracy.

5.2 *A steady state solution of the Fokker–Planck equation*

We now test the second-order and fourth-order schemes for solving the following two-dimensional linear Fokker–Planck equation on $\Omega = (-3, 3) \times (-3, 3)$:

$$\partial_t \rho = \Delta \rho + \nabla \cdot (\rho \nabla \mathcal{V}), \quad \mathcal{V} = \frac{x^2 + y^2}{2}. \tag{45}$$

It is equivalent to

$$\partial_t \rho = \nabla \cdot \left(\mathcal{M} \nabla \frac{\rho}{\mathcal{M}} \right), \quad \mathcal{M} := e^{-\frac{x^2+y^2}{2}},$$

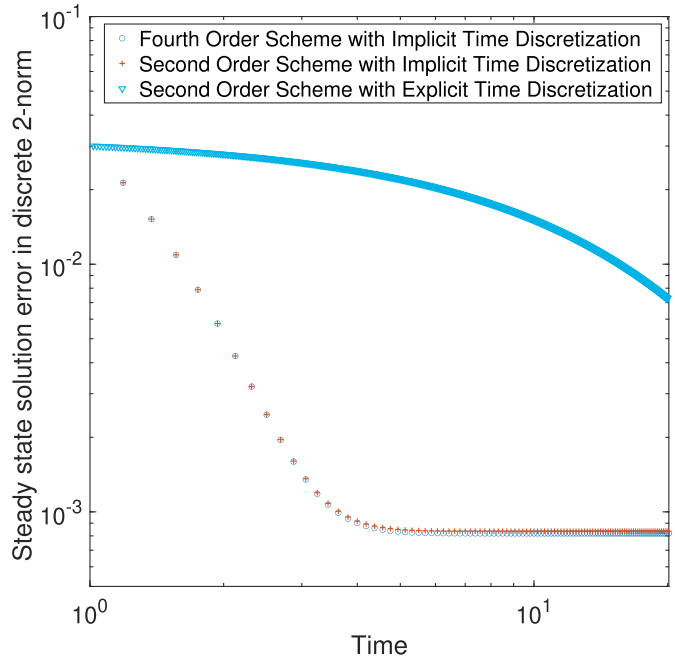
with the boundary condition $\nabla \frac{\rho}{\mathcal{M}} \cdot \mathbf{n}|_{\partial\Omega} = 0$. This equation admits an exact solution:

$$\rho(t, x, y) = \frac{1}{2\pi(1 - e^{-2t})} e^{-\frac{x^2+y^2}{2(1-e^{-2t})}}.$$

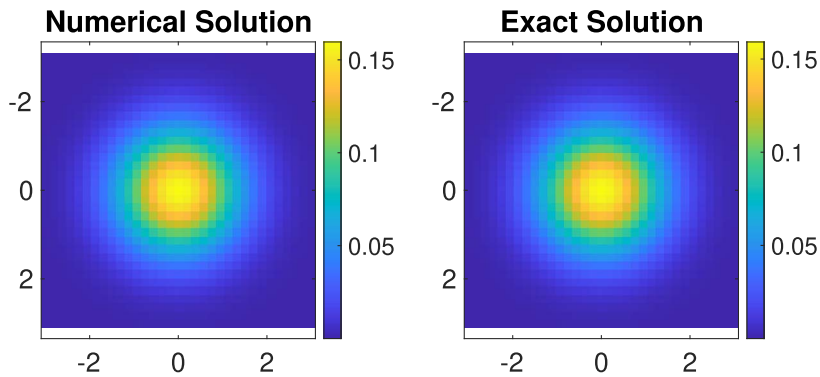
We use $\rho(1, x, y)$ as an initial condition and march to time $T = 20$ for approximating the steady state

$$\rho_\infty(x, y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}.$$

To demonstrate the advantages of our schemes we also compare them to the second-order spatial discretization with fully explicit forward Euler time discretization, which can also be proven positivity preserving and energy dissipative, but under a small time step constraint $\Delta t = \mathcal{O}(\Delta x^2)$. In Fig. 3 we can see that the convergence of the explicit scheme to the steady state solution is much slower. Moreover, the small time step $\Delta t = \mathcal{O}(\Delta x^2)$ is usually not desired in applications. The convergence to numerical



(a) Three schemes are used on the same 33×33 grid. The implicit schemes use a time step $\Delta t = \mathcal{O}(\Delta x)$ and the explicit scheme uses a time step $\Delta t = \mathcal{O}(\Delta x^2)$.

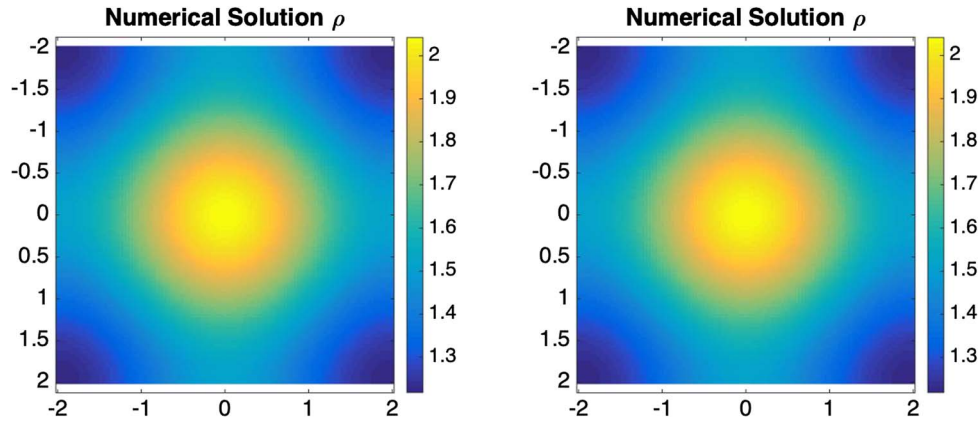


(b) The steady state solution. Numerical solution was generated by the fourth order scheme on a 33×33 grid.

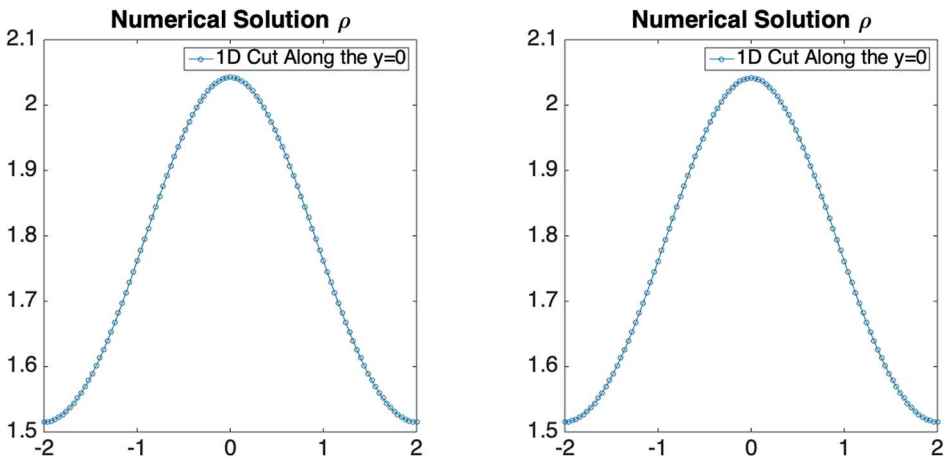
FIG. 3. Linear Fokker–Planck equation on $\Omega = (-3, 3) \times (-3, 3)$.

steady state solution of two implicit schemes is similar. On the other hand, the fourth-order scheme produces slightly smaller errors in the numerical steady state solution.

In Fig. 3, after $T = 10$, steady state solution errors of both implicit schemes stay flat, and in each time step $\|\rho^{n+1} - \rho^n\|_\infty$ is less than 10^{-10}, which is the accuracy tolerance of preconditioned conjugate



(a) The second order scheme. (b) The fourth order scheme.



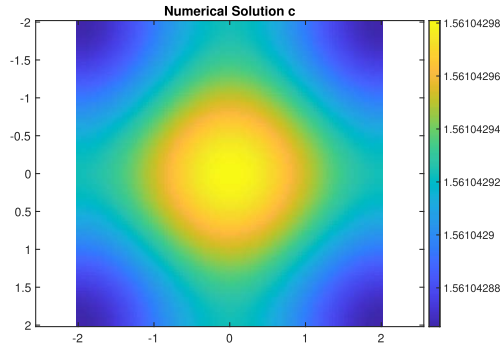
(c) The second order scheme. (d) The fourth order scheme.

FIG. 4. Keller–Segel system with an initial condition below critical mass $\rho(x, y, 0) = \frac{60}{1+40(x^2+y^2)}$ on $\Omega = (-2, 2) \times (-2, 2)$. The solutions at $T = 2$ are plotted. Both schemes are computed on a 101×101 grid.

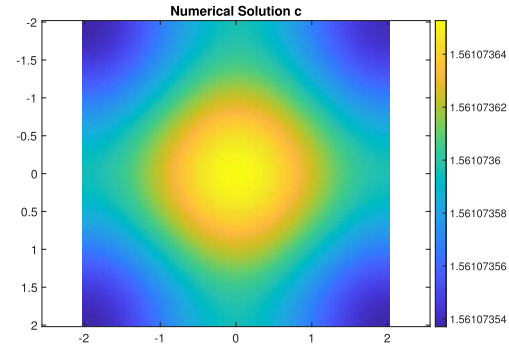
gradient linear system solver. At $T = 20$, compared to the exact steady state, the fourth-order scheme with implicit time stepping produces error in discrete 2-norm as 8.18×10^{-4}, and the second-order scheme with implicit time stepping produces error in discrete 2-norm 8.35×10^{-4}. We emphasize both implicit schemes are used on the same grid and the difference in computational cost is marginal, thus this is a clear advantage of using a high order accurate spatial discretization, even if the time accuracy is only first order.

5.3 *A smooth solution of the Keller–Segel system*

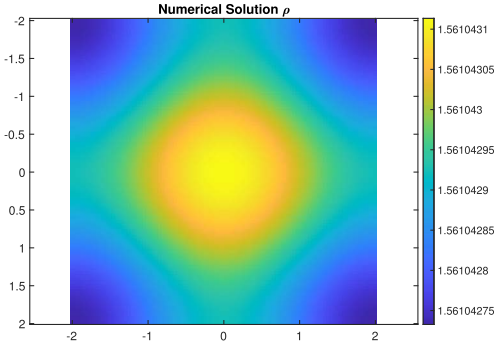
For the Keller–Segel system it is well known that there is a critical value for total mass in initial conditions, below which a globally well-posed solution exists (Dolbeault & Perthame, 2004;



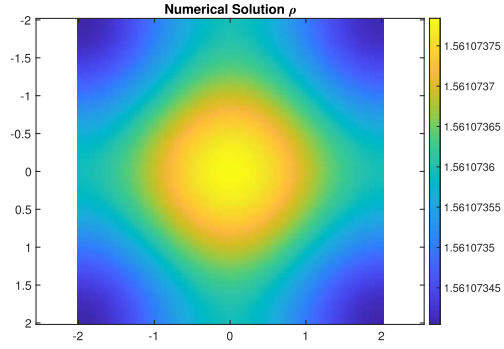
(a) The second order scheme.



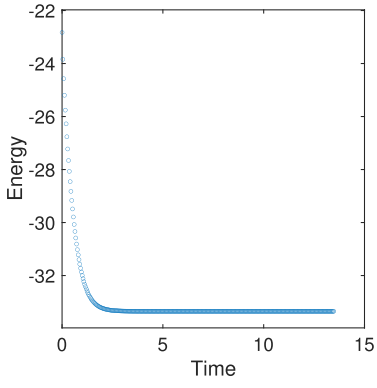
(b) The fourth order scheme.



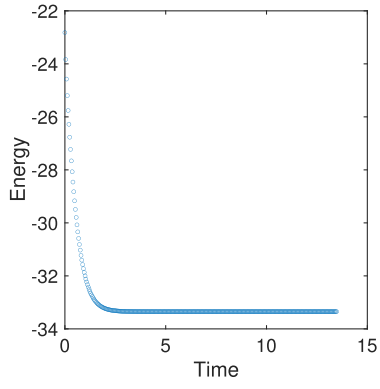
(c) The second order scheme.



(d) The fourth order scheme.

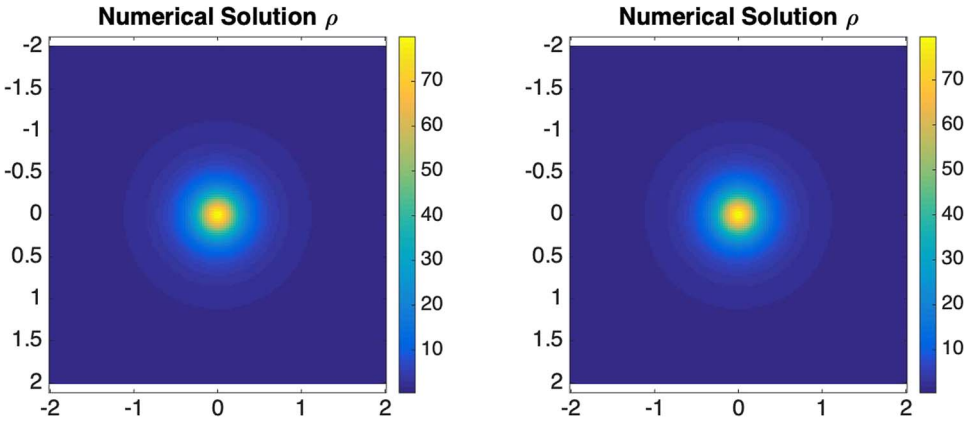


(e) The second order scheme.

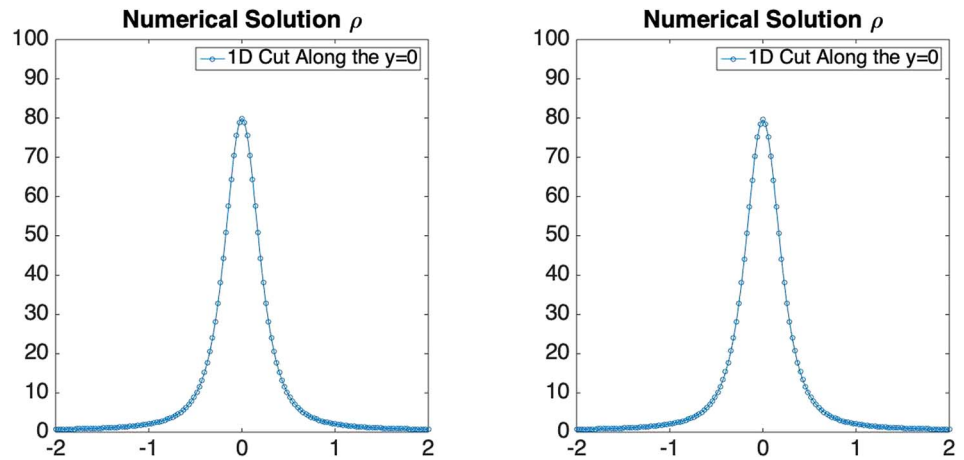


(f) The fourth order scheme.

FIG. 5. Keller–Segel system with an initial condition below critical mass $\rho(x, y, 0) = \frac{60}{1+40(x^2+y^2)}$. The plotted numerical solutions are around the time $T = 13.52$ when $\|\rho^{n+1} - \rho^n\|_\infty \leq 10^{-8}$. Both schemes are computed on a 101×101 grid.



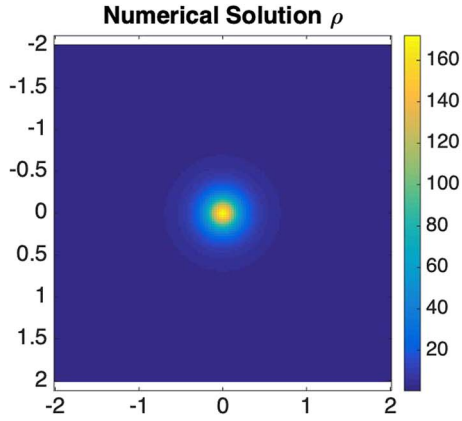
(a) The second order scheme at $T = 0.11$. (b) The fourth order scheme at $T = 0.11$.



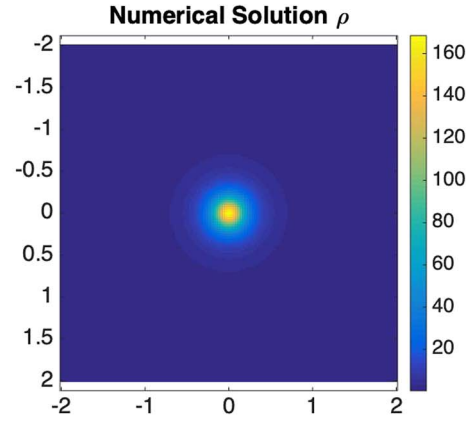
(c) The second order scheme at $T = 0.11$. (d) The fourth order scheme at $T = 0.11$.

FIG. 6. Keller–Segel system with an initial condition above critical mass $\rho(x, y, 0) = \frac{100}{1+40(x^2+y^2)}$ on $\Omega = (-2, 2) \times (-2, 2)$. Both schemes are computed on a 141×141 grid.

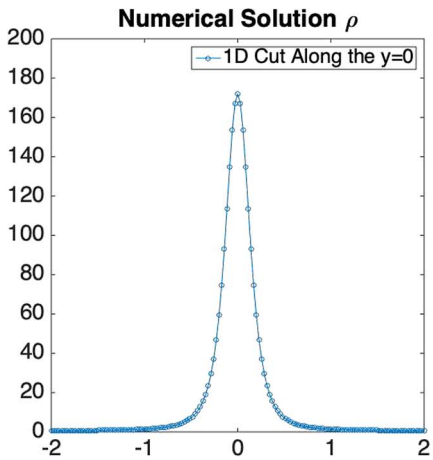
Blanchet *et al.*, 2006). We solve the system (44) with $f(x, y) \equiv 0$ on $\Omega = (-2, 2) \times (-2, 2)$ with an initial condition $\rho(0, x, y) = \frac{60}{1+40(x^2+y^2)}$ and its mass is below the critical value. See both schemes on the same grid of 101×101 points at $T = 2$ in Fig. 4. For both schemes $\Delta t = \Delta x$ is used. Then we run two schemes for longer time until $\|\rho^{n+1} - \rho^n\|_\infty \leq 10^{-8}$ is satisfied. Both schemes reach $\|\rho^{n+1} - \rho^n\|_\infty \leq 10^{-8}$ around $T = 13.52$. See numerical solutions at $T = 13.52$ in Fig. 5. Note that in this case the energy as defined in (41) reaches a constant value, which is an indicator that the system has already reached the steady state.



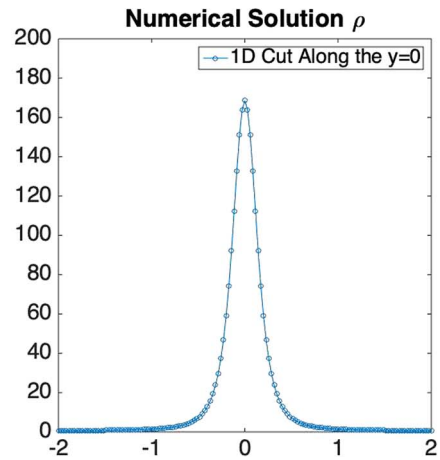
(a) The second order scheme at $T = 0.2$.



(b) The fourth order scheme at $T = 0.2$.



(c) The second order scheme at $T = 0.2$.



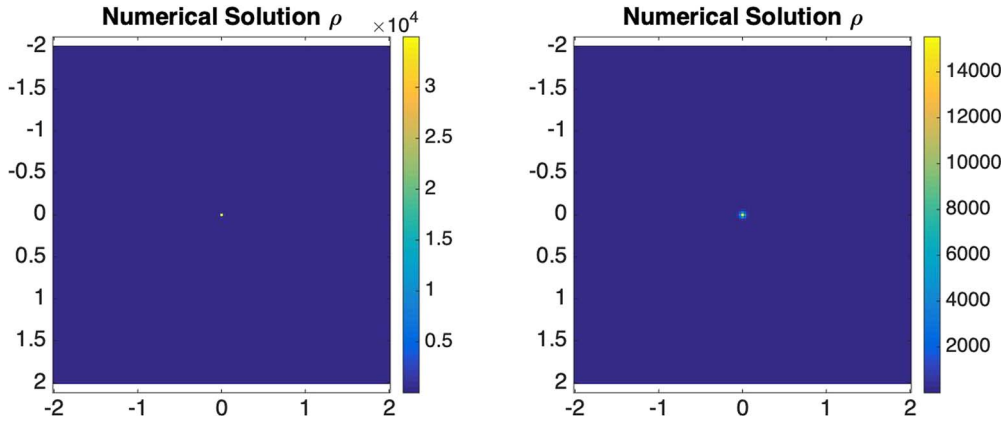
(d) The fourth order scheme at $T = 0.2$.

FIG. 7. Keller–Segel system with an initial condition above critical mass $\rho(x, y, 0) = \frac{100}{1+40(x^2+y^2)}$ on $\Omega = (-2, 2) \times (-2, 2)$. Both schemes are computed on a 141×141 grid.

5.4 *A blow-up solution of the Keller–Segel system*

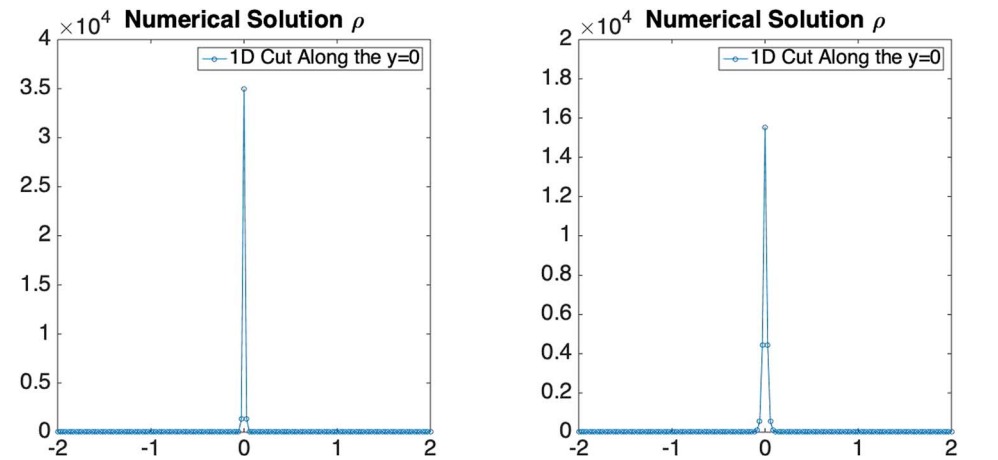
For an initial condition with total mass above the critical mass, a blow-up will emerge in finite time for the Keller–Segel system (Dolbeault & Perthame, 2004; Blanchet *et al.*, 2006), see also Carrillo *et al.* (2019); Guo *et al.* (2019) for computational examples.

We test both schemes for an initial condition $\rho(0, x, y) = \frac{100}{1+40(x^2+y^2)}$ with total mass above the critical value. See solutions at $T = 0.11$ in Fig. 6, at $T = 0.2$ in Fig. 7 and at $T = 0.8$ in Fig. 8. For both schemes $\Delta t = \Delta x$ is used. Note that at $T = 0.8$, the solution in the fourth-order scheme is significantly different from the second-order one, while the former is certainly more faithful due to its higher accuracy.



(a) The second order scheme at $T = 0.8$.

(b) The fourth order scheme at $T = 0.8$.



(c) The second order scheme at $T = 0.8$.

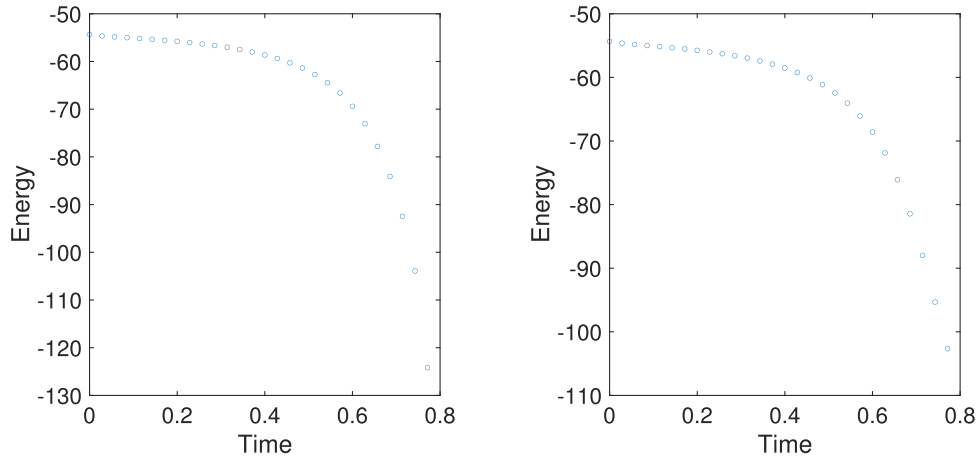
(d) The fourth order scheme at $T = 0.8$.

FIG. 8. Keller–Segel system with an initial condition above critical mass $\rho(x, y, 0) = \frac{100}{1+40(x^2+y^2)}$ on $\Omega = (-2, 2) \times (-2, 2)$. Both schemes are computed on a 141×141 grid.

The energy evolution of numerical solutions is shown in Fig. 9, where the discrete energy is defined as in (41). It should be mentioned that the mesh constraints in Section 3 for achieving monotonicity in the fourth-order scheme will be eventually impossible to be satisfied for a blow-up solution, yet these mesh constraints are only sufficient conditions for monotonicity. In our fourth-order numerical solutions, it has been checked that ρ is always positive even after blow up. Therefore, the energy dissipation is still in good faith.

6. Concluding remarks

We have constructed two finite difference schemes that are proved be positivity preserving and energy dissipative for the Fokker–Planck and Keller–Segel type equations. The time discretization is a first-



(a) The energy evolution of the second order scheme.

(b) The energy evolution of the fourth order scheme.

FIG. 9. Keller–Segel system with an initial condition above critical mass $\rho(x, y, 0) = \frac{100}{1+40(x^2+y^2)}$ on $\Omega = (-2, 2) \times (-2, 2)$. Both schemes are computed on a 141×141 grid.

order semi-implicit or implicit scheme. The spatial discretizations include a second-order and a fourth-order finite difference scheme, obtained via finite difference implementation of the finite element method with linear and quadratic polynomials on uniform meshes. Under mild mesh size and time step constraints for smooth solutions (a lower bound on time step rather than upper bound) the fourth-order scheme is proved to be monotone, thus is positivity-preserving and decays energy, which is the first high order spatial discretization with these properties. Numerical tests on both the Fokker–Planck equation and Keller–Segel system are performed to verify the performance of the proposed schemes.

Funding

US National Science Foundation CAREER (DMS-2153208 to J.H.); US Air Force Office of Scientific Research (FA9550-21-1-0358 to J.H.); US National Science Foundation (DMS-1913120 to X.Z.).

REFERENCES

ALMEIDA, L., BUBBA, F., PERTHAME, B. & POUCHOL, C. (2019) Energy and implicit discretization of the Fokker–Planck and Keller–Segel type equations. *Netw. Heterog. Media*, **14**, 23–41.

BAILO, R., CARRILLO, J. A. & HU, J. (2020) Fully discrete positivity-preserving and energy-dissipating schemes for aggregation–diffusion equations with a gradient flow structure. *Commun. Math. Sci.*, **18**, 1259–1303.

BLANCHET, A., DOLBEAULT, J. & PERTHAME, B. (2006) Two-dimensional Keller–Segel model: optimal critical mass and qualitative properties of the solutions. *Electron. J. Differ. Equ. (EJDE) [electronic only]*, Paper–No, 2006.

CARRILLO, J. A., CHERTOCK, A. & HUANG, Y. (2015) A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Commun. Comput. Phys.*, **17**, 233–258.

CARRILLO, J. A., CRAIG, K. & YAO, Y. (2019) Aggregation–diffusion equations: dynamics, asymptotics, and singular limits. *Active Particles, Volume 2: Advances in Theory, Models, and Applications*, vol. **2** (N. Bellomo, P. Degond & E. Tadmor eds). Switzerland: Springer, pp. 65–108.

Carrillo, J. A., McCann, R. & Villani, C. (2003) Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Rev. Mat. Iberoam*, **19**, 971–1018.

Cross, L. J. & Zhang, X. (2020) On the monotonicity of high order discrete Laplacian. arXiv preprint arXiv:2010.07282.

Dolbeault, J. & Perthame, B. (2004) Optimal critical mass in the two dimensional Keller–Segel model in R2. *Comp. Rend. Math.*, **339**, 611–616.

Guo, L., Li, X. & Yang, Y. (2019) Energy dissipative local discontinuous Galerkin methods for Keller–Segel chemotaxis model. *J. Sci. Comput.*, **78**, 1387–1404.

Hu, J. & Huang, X. (2020) A fully discrete positivity-preserving and energy-dissipative finite difference scheme for Poisson–Nernst–Planck equations. *Numer. Math.*, **145**, 77–115.

Hu, J., Liu, J.-G., Xie, Y. & Zhou, Z. (2021) A structure preserving numerical scheme for Fokker–Planck equations of neuron networks: numerical analysis and exploration. *J. Comput. Phys.*, **433**, 110195.

Hu, J. & Shu, R. (2019) A second-order asymptotic-preserving and positivity-preserving exponential Runge–Kutta method for a class of stiff kinetic equations. *Multiscale Model. Simul.*, **17**, 1123–1146.

Jin, S. & Yan, B. (2011) A class of asymptotic-preserving schemes for the Fokker–Planck–Landau equation. *J. Comput. Phys.*, **230**, 6420–6437.

Li, H. (2021) Accuracy and monotonicity of spectral element method on structured meshes. *Ph.D. Thesis*. West Lafayette, IN: Purdue University.

Li, H., Appelö, D. & Zhang, X. (2022) Accuracy of spectral element method for wave, parabolic and Schrödinger equations. *SIAM J. Numer. Anal.*, **60**, 339–363.

Li, H., Xie, S. & Zhang, X. (2018) A high order accurate bound-preserving compact finite difference scheme for scalar convection–diffusion equations. *SIAM J. Numer. Anal.*, **56**, 3308–3345.

Li, H. & Zhang, X. (2020a) On the monotonicity and discrete maximum principle of the finite difference implementation of C^0-Q^2 finite element method. *Numer. Math.*, **145**, 437–472.

Li, H. & Zhang, X. (2020b) Superconvergence of high order finite difference schemes based on variational formulation for elliptic equations. *J. Sci. Comput.*, **82**, 36.

Liu, J.-G., Wang, L. & Zhou, Z. (2018) Positivity-preserving and asymptotic-preserving method for 2D Keller–Segel equations. *Math. Comp.*, **87**, 1165–1189.

Lorenz, J. (1977) Zur inversmonotonie diskreter probleme. *Numer. Math.*, **27**, 227–238.

Maday, Y. & Rønquist, E. M. (1990) Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries. *Comput. Methods Appl. Mech. Eng.*, **80**, 91–115.

Plemmons, R. J. (1977) M-matrix characterizations. I—nonsingular M-matrices. *Linear Algebra Appl.*, **18**, 175–188.

Qiu, C., Liu, Q. & Yan, J. (2021) Third order positivity-preserving direct discontinuous Galerkin method with interface correction for chemotaxis Keller–Segel equations. *J. Comput. Phys.*, 110191.

Shen, J. & Xu, J. (2020) Unconditionally bound preserving and energy dissipative schemes for a class of Keller–Segel equations. *SIAM J. Numer. Anal.*, **58**, 1674–1695.

Srinivasan, S., Poggie, J. & Zhang, X. (2018) A positivity-preserving high order discontinuous Galerkin scheme for convection–diffusion equations. *J. Comput. Phys.*, **366**, 120–143.

Sun, Z., Carrillo, J. A. & Shu, C.-W. (2018) A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials. *J. Comput. Phys.*, **352**, 76–104.

Vazquez, J. (2007) *The Porous Medium Equation*. Oxford: Oxford University Press.

Villani, C. (2003) *Topics in Optimal Transportation*. Graduate Studies in Mathematics, vol. 58. Providence, RI: American Mathematical Society.

Zhang, Y., Zhang, X. & Shu, C.-W. (2013) Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection–diffusion equations on triangular meshes. *J. Comput. Phys.*, **234**, 295–316.