DISCRETE MAXIMUM PRINCIPLE OF
A HIGH ORDER FINITE DIFFERENCE SCHEME FOR
A GENERALIZED ALLEN-CAHN EQUATION*

JIE SHEN† AND XIANGXIONG ZHANG‡

Abstract. We consider solving a generalized Allen-Cahn equation coupled with a passive convection for a given incompressible velocity field. The numerical scheme consists of the first order accurate stabilized implicit explicit time discretization and a fourth order accurate finite difference scheme, which is obtained from the finite difference formulation of the Q^2 spectral element method. We prove that the discrete maximum principle holds under suitable mesh size and time step constraints. The same result also applies to the construcion of a bound-preserving scheme for any passive convection with an incompressible velocity field.

Keywords. Discrete maximum principle; high order accuracy; monotonicity; bound-preserving; phase field equations; incompressible flow.

AMS subject classifications. 65M06; 65M60; 65M12.

1. Introduction

In this paper we consider bound-preserving schemes for a generalized Allen-Cahn equation

$$\phi_t + u\phi_x + v\phi_y = \mu\Delta\phi - \frac{F'(\phi)}{\varepsilon}, \quad (x,y) \in \Omega, \tag{1.1}$$

where Ω is an open bounded domain in $\mathbb{R}^2$, $\mu, \varepsilon > 0$ are parameters, $F(\phi)$ is an energy function, and (u,v) is a given incompressible velocity field. The Allen-Cahn equation, i.e.(1.1) with $(u,v) \equiv 0$, plays an important role in materials science [1,2]. The generalized Allen-Cahn Equation (1.1), often with an extra Lagrange multiplier to conserve the volume fraction [22], is frequently encountered in modeling of multi-phase incompressible flows, e.g., [14].

The generalized Allen-Cahn Equation (1.1) usually satisfies a maximum principle, so it is desired to have its numerical solution to preserve the maximum principle or to remain in a prescribed bound. In particular, this becomes crucial when the energy function $F(\phi)$ is of the form

$$F(\phi) = \frac{\theta}{2}[(1+\phi)\ln(1+\phi) + (1-\phi)\ln(1-\phi)] - \frac{\theta_c}{2}\phi \tag{1.2}$$

where θ, θ_c are two positive constants.

To construct bound-preserving schemes for Equation (1.1), we can first consider bound-preserving schemes for a convection-diffusion equation, e.g., $F(\phi) \equiv 0$. In the literature, there are many fully explicit high order accurate bound-preserving schemes for a scalar convection-diffusion equation [3,8,11,18,20,21,23]. In these schemes, the time

†Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, USA (shen@math.purdue.edu).
‡Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, USA (zhan1966@purdue.edu).

discretizations are high order explicit time strong stability preserving (SSP) Runge-Kutta and multistep methods, which are convex combinations of forward Euler steps. Even though such an approach allows various high order accurate spatial discretizations, all these fully explicit schemes require a small time step $\Delta t = \mathcal{O}(\frac{1}{\mu}\Delta x^2)$, which is inpractical unless μ is very small.

To construct bound-preserving schemes without the parabolic type CFL constraint $\Delta t = \mathcal{O}(\frac{1}{\mu}\Delta x^2)$ for (1.1), the second order finite difference was used in [19] with an implicit explicit (IMEX) time discretization

$$\frac{\phi^{n+1} - \phi^n}{\Delta t} + u^{n+1}\phi_x^{n+1} + v^{n+1}\phi_y^{n+1} = \mu\Delta\phi^{n+1} - \frac{F'(\phi^n)}{\varepsilon}, \qquad (1.3)$$

and a stabilized scheme with a parameter $S \geq 0$:

$$\frac{\phi^{n+1} - \phi^n}{\Delta t} + S(\phi^{n+1} - \phi^n) + u^{n+1}\phi_x^{n+1} + v^{n+1}\phi_y^{n+1} = \mu\Delta\phi^{n+1} - \frac{F'(\phi^n)}{\varepsilon}. \qquad (1.4)$$

For spatial discretization, it is well-known that the second order finite difference for (1.3) and (1.4) forms an M-matrix [19], thus the matrix of the linear system in (1.3) and (1.4) is *monotone*, i.e., the inverse matrix is entrywise non-negative. Monotonicity is the key property which implies the discrete maximum principle. In general, high order accurate schemes do not form M-matrices, thus it is also quite challenging to extend the method in [19] to higher spatial accuracy. Nonetheless, recent progress in [12] shows that the finite element method with Q^2 polynomial (tensor product quadratic polynomial) on structured meshes is a product of two M-matrices for a diffusion operator and thus is still monotone.

The main purpose of this paper is to extend the results in [12] to the spatial discretization for (1.3) and (1.4). In particular, when Q^2 finite element method for a convection-diffusion operator in (1.3) and (1.4) is implemented with 3-point Gauss-Lobatto quadrature as a finite difference scheme, it can be rigorously proven that it is a fourth order accurate spatial discretization in the discrete l^2-norm [10,13]. In the literature, Q^k finite element method implemented by m-point Gauss-Lobatto quadrature with $m \geq k+1$ is also called spectral element method [16]. The fourth order finite difference scheme in this paper is also equivalent to Q^2 spectral element method with only 3-point Gauss-Lobatto quadrature. More precisely, we will prove that this fourth order finite difference spatial discretization for (1.3) and (1.4) satisfies the discrete maximum principle under certain mesh size and time step constraints. For the discrete maximum principle to hold for a convection-diffusion equation, the time step constraint in this paper is a lower bound condition on $\frac{\Delta t}{\Delta x^2}$, thus still practical.

For extensions to higher order time accuracy, in general it is quite difficult since there are no high order SSP implicit time discretizations without the constraint $\Delta t = \mathcal{O}(\frac{1}{\mu}\Delta x^2)$, see [7]. For a second order spatial discretization, one possible approach to obtain a second order accurate time scheme is to consider the exponential time differencing schemes [5,6,9], which heavily depend on the ℓ^∞ estimate of the matrix exponential e^{Δ_h}, with Δ_h denoting the discrete Laplacian matrix. For the second order finite difference, such an estimate can be established by the exact solution of the ordinary differential equations of the semi-discrete scheme for solving heat equation since the second order finite difference gives a diagonally dominant matrix Δ_h. Unfortunately, if using the fourth order accurate spatial discretization in this paper, the discrete Laplacian matrix Δ_h is no longer diagonally dominant.

The rest of the paper is organized as follows. In Section 2, we review the finite difference scheme obtained from Q^2 spectral element method. Its monotonicity is proved

in Section 3. In Section 4, we establish the discrete maximum principle for the generalized Allen-Cahn equation with both polynomial and logarithmic energy functions. The main results in Section 3 can also be used to construct a fourth order bound-preserving finite difference spatial discretization for any passive convection. As a demonstration, we apply it to the two-dimensional incompressible Navier-Stokes equation in stream function vorticity formulation in Section 5. We present in Section 6 several numerical tests to validate our scheme. Some concluding remarks are given in Section 7.

2. Finite difference implementation of Q^2 spectral element method

For simplicity, we derive the scheme for an elliptic equation with incompressible velocity field $\mathbf{u} = (u, v)$ and a given function f on a square domain $\Omega = (0, 1) \times (0, 1)$ and Dirichlet boundary conditions:

$$\phi + \mathbf{u} \cdot \nabla \phi - \nabla \cdot (\mu \nabla \phi) = f \text{ on } \Omega, \quad \phi(x, y) = g(x, y) \text{ on } \partial \Omega. \tag{2.1}$$

We only consider a constant scalar μ and Q^2 elements on a uniform mesh, even though the scheme can also be easily extended to general scenarios such as Neumann boundary conditions, and diffusion terms like a variable μ or $\nabla \cdot (A \nabla \phi)$ with a positive definite matrix function A, see [13].

Let Ω_h denote a uniform rectangular mesh as shown in Figure 2.1 (a). Let $Q^2(e)$ be the set of tensor products of quadratic polynomials on a rectangular cell e:

$$Q^2(e) = \left\{ p(x, y) = \sum_{i=0}^{2} \sum_{j=0}^{2} p_{ij} x^i y^j, (x, y) \in e \right\}.$$

Let V^h and V_0^h denote two continuous piecewise Q^2 finite element spaces on Ω_h:

$$V^h = \{ p(x, y) \in C^0(\Omega_h) : p|_e \in Q^2(e), \quad \forall e \in \Omega_h \},$$

$$V_0^h = \{ v_h \in V^h : v_h = 0 \quad \text{on} \quad \partial \Omega \}.$$
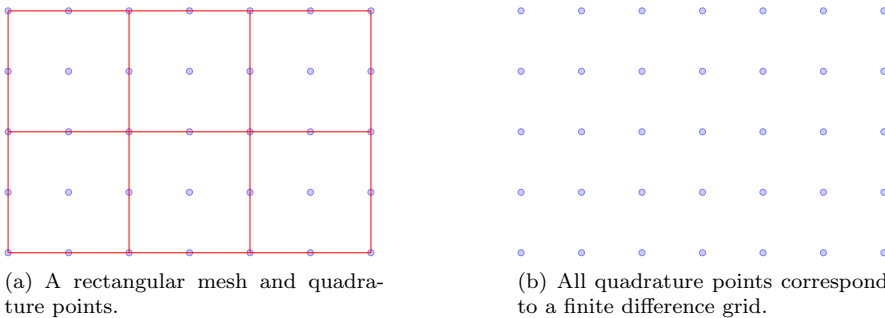

(a) A rectangular mesh and quadrature points.

(b) All quadrature points correspond to a finite difference grid.

FIG. 2.1. *An illustration of a uniform rectangular mesh for Q^2 elements and the 3×3 Gauss-Lobatto quadrature.*

2.1. Variational formulation. Assume there is a function $\bar{g} \in H^1(\Omega)$ as a smooth extension of g so that $\bar{g}|_{\partial\Omega} = g$. Introduce a bilinear form

$$B(\phi, \psi) := \langle \phi, \psi \rangle + \langle \mathbf{u} \cdot \nabla \phi, \psi \rangle + \mu \langle \nabla \phi, \nabla \psi \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the standard L^2 inner product on Ω, then the variational form of (2.1) is to find $\tilde{\phi} = \phi - \bar{g} \in H_0^1(\Omega)$ satisfying

$$B(\tilde{\phi}, \psi) = \langle f, \psi \rangle - B(\bar{g}, \psi), \quad \forall \psi \in H_0^1(\Omega). \tag{2.2}$$

In practice, $\bar{g}$ is not used explicitly. By abusing notation, the most convenient implementation is to consider

$$g(x,y) = \begin{cases} 0, & \text{if} \quad (x,y) \in (0,1) \times (0,1), \\ g(x,y), & \text{if} \quad (x,y) \in \partial\Omega, \end{cases}$$

and $g_I \in V^h$ which is defined as the Q^2 Lagrange interpolation of $g(x,y)$ at 3×3 Gauss-Lobatto points for each rectangular cell on Ω. Namely, $g_I \in V^h$ is the piecewise Q^k interpolation of g along the boundary grid points and $g_I = 0$ at the interior grid points.

The spectral element method, i.e., finite element method with suitable quadrature, is to find $\phi_h \in V_0^h$, s.t.

$$B_h(\phi_h + g_I, \psi_h) = \langle f, v_h \rangle_h, \quad \forall \psi_h \in V_0^h, \tag{2.3}$$

where $B_h(\phi_h, \psi_h)$ and $\langle f, v_h \rangle_h$ denote using 3×3 Gauss-Lobatto quadrature for integrals $B(\phi_h, \psi_h)$ and $\langle f, \psi_h \rangle$, respectively. Then $\phi_h + g_I$ will be our numerical solution to $\phi(x,y)$ for (2.1). Notice that (2.3) is not a straightforward approximation to (2.2) since $\bar{g}$ is never used. The scheme (2.3) is fourth order accurate, see [10, 13].

2.2. One-dimensional fourth order scheme. To derive an explicit expression of the scheme (2.3), we start with a one-dimensional steady state equation on $x \in (0,1)$ with homogeneous Neumann boundary conditions:

$$\phi(x) + u(x)\phi'(x) - (\mu\phi')' = f(x), \quad \phi'(0) = \phi'(1) = 0.$$

The variational form is to find $\phi(x) \in H^1([0,1])$ satisfying

$$B(\phi, \psi) := \langle \phi, \psi \rangle + \langle u\phi', \psi \rangle + \langle \mu\phi', \psi' \rangle = \langle f, \psi \rangle, \quad \forall \psi(x) \in H^1([0,1]),$$

where $\langle \cdot, \cdot \rangle$ denotes standard L^2 inner product. Consider a uniform mesh $x_i = ih$, $i = 0, 1, \ldots, n+1$, $h = \frac{1}{n+1}$. Assume n is odd and let $N = \frac{n+1}{2}$. Define a finite element mesh for P^2 basis with intervals $I_k = [x_{2k}, x_{2k+2}]$ for $k = 0, \ldots, N-1$. Define

$$V^h = \{\psi \in C^0([0,1]) : \psi|_{I_k} \in P^2(I_k), k = 0, \ldots, N-1\}.$$

Let $\{\psi_i\}_{i=0}^{n+1} \subset V^h$ be a basis of V^h such that $\psi_i(x_j) = \delta_{ij}$, $i, j = 0, 1, \ldots, n+1$. Then the continuous P^2 finite element method with 3-point Gauss-Lobatto quadrature is to seek $\phi_h(x) \in V^h$ satisfying

$$B_h(\phi_h, \psi_h) := \langle \phi_h, \psi_i \rangle_h + \langle u\phi_h', \psi_i \rangle_h + \langle \mu\phi_h', \psi_i' \rangle_h = \langle f, \psi_i \rangle_h, \quad i = 0, 1, \ldots, n+1, \tag{2.4}$$

where $\langle\cdot,\cdot\rangle_h$ denotes 3-point Gauss-Lobatto quadrature for approximating integration on each interval I_k. Let $\phi_j = \phi_h(x_j)$ and $u_j = u(x_j)$, then $\phi_h(x) = \sum_{j=0}^{n+1} \phi_j \psi_j(x)$. We have

$$\sum_{j=0}^{n+1} \phi_j \left(\langle\psi_j, \psi_i\rangle_h + \langle u\psi_j', \psi_i\rangle_h + \mu\langle\psi_j', \psi_i'\rangle_h \right) = \sum_{j=0}^{n+1} f_j \langle\psi_j, \psi_i\rangle_h, \quad i = 0, 1, \ldots, n+1.$$

The matrix form of this scheme is $\bar{M}\bar{\phi} + \bar{U}\bar{T}\bar{\phi} + \mu\bar{S}\bar{\phi} = \bar{M}\bar{\mathbf{f}}$, where

$$\bar{U} = \begin{bmatrix} u_0 & & & & \\ & u_1 & & & \\ & & \ddots & & \\ & & & u_n & \\ & & & & u_{n+1} \end{bmatrix}, \bar{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_n \\ \phi_{n+1} \end{bmatrix}, \quad \bar{\mathbf{f}} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \\ f_{n+1} \end{bmatrix}.$$

The stiffness matrix $\bar{S}$ has size $(n+2) \times (n+2)$ with (i,j)-th entry as $\langle\psi_i', \psi_j'\rangle_h$, the matrix $\bar{T}$ has size $(n+2) \times (n+2)$ with (i,j)-th entry as $\langle\psi_j', \psi_i\rangle_h$ and the lumped mass matrix $\bar{M}$ is a $(n+2) \times (n+2)$ diagonal matrix with diagonal entries $h\left(\frac{1}{3}, \frac{4}{3}, \frac{2}{3}, \frac{4}{3}, \frac{2}{3}, \ldots, \frac{2}{3}, \frac{4}{3}, \frac{1}{3}\right)$.

Notice that $\bar{U}$ and $\bar{M}$ are diagonal, thus they commute. Multiplying $\bar{M}^{-1}$ for both sides, we get a finite difference representation

$$\bar{\phi} + \bar{U}\bar{D}_1\bar{\phi} - \mu\bar{D}_2\bar{\phi} = \bar{\mathbf{f}},$$

with square difference matrices for approximating first order and second order derivatives as

$$\bar{D}_1 = \bar{M}^{-1}\bar{T} = \frac{1}{2h} \begin{pmatrix} -3 & 4 & -1 & & & & & & & \\ -1 & 0 & 1 & & & & & & & \\ \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & & & \\ & & -1 & 0 & 1 & & & & & \\ & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & \\ & & & & -1 & 0 & 1 & & & \\ & & & & & \ddots & \ddots & \ddots & & \\ & & & & & & -1 & 0 & 1 & \\ & & & & & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} \\ & & & & & & & -1 & 0 & 1 \\ & & & & & & & & 1 & -4 & 3 \end{pmatrix},$$

$$\bar{D}_2 = -\bar{M}^{-1}\bar{S} = -\frac{1}{h^2} \begin{pmatrix} \frac{7}{2} & -4 & \frac{1}{2} & & & & & & \\ -1 & 2 & -1 & & & & & & \\ \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} & & & & \\ & & -1 & 2 & -1 & & & & \\ & & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} & & \\ & & & -1 & 2 & -1 & & & \\ & & & & \ddots & \ddots & \ddots & & \\ & & & & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} \\ & & & & & -1 & 2 & -1 \\ & & & & & \frac{1}{2} & -4 & \frac{7}{2} \end{pmatrix}.$$

Now consider the one-dimensional Dirichlet boundary value problem:

$$\phi(x) + u(x)\phi'(x) - (\mu\phi')' = f(x) \text{ on } [0,1], \quad \phi(0) = \sigma_1, \quad \phi(1) = \sigma_2.$$

Consider the same mesh as above and define

$$V_0^h = \{\psi \in C^0([0,1]) : \psi|_{I_k} \in P^2(I_k), k = 0, \ldots, N-1; \psi(0) = \psi(1) = 0\}.$$

Then $\{\psi_i\}_{i=1}^n \subset V^h$ is a basis of V_0^h. Let $g_I(x) = \sigma_0 \psi_0(x) + \sigma_1 \psi_{n+1}(x)$, then the one-dimensional version of (2.3) is to seek $\phi_h \in V_0^h$ satisfying

$$B_h(\phi_h + g_I, \psi_i)_h = \langle f, \psi_i \rangle_h, \quad i = 1, 2, \ldots, n. \tag{2.5}$$

Notice that we can obtain (2.5) by simply setting $\phi_h(0) = \sigma_0$ and $\phi_h(1) = \sigma_1$ in (2.4). So the finite difference implementation of (2.5) is given as

$$\begin{aligned}
\phi + U D_1 \bar{\phi} - \mu D_2 \bar{\phi} &= f, \\
\phi_0 = \sigma_0, \phi_1 &= \sigma_1,
\end{aligned} \tag{2.6}$$

with

$$U = \begin{bmatrix} u_1 & & \\ & \ddots & \\ & & u_n \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}, \phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_n \end{bmatrix}, \bar{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_n \\ \phi_1 \end{bmatrix},$$

and difference matrices of size $n \times (n+2)$:

$$D_1 = \frac{1}{2h} \begin{pmatrix} -1 & 0 & 1 & & & & & & \\ \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & & \\ & -1 & 0 & 1 & & & & & \\ & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & \\ & & -1 & 0 & 1 & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & -1 & 0 & 1 & & \\ & & & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} \\ & & & & & & -1 & 0 & 1 \end{pmatrix}, D_2 = -\frac{1}{h^2} \begin{pmatrix} -1 & 2 & -1 & & & & & & \\ \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} & & & & \\ & -1 & 2 & -1 & & & & & \\ & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} & & & \\ & & -1 & 2 & -1 & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} \\ & & & & & & -1 & 2 & -1 \end{pmatrix}.$$

Let I_n denote the identity matrix of size $n \times n$, and define a restriction matrix

$$R = \begin{pmatrix} \mathbf{0} & I_n & \mathbf{0} \end{pmatrix}_{n \times (n+2)}.$$

Then the left-hand side of (2.6) for interior points can be regarded as a linear operator on $\bar{\phi}$: $\mathcal{L}(\bar{\phi}) = R\bar{\phi} + U D_1 \bar{\phi} - \mu D_2 \bar{\phi} = f$.

The scheme can also be explicitly written as

$$\phi_0 = \sigma_0, \quad \phi_{n+1} = \sigma_1; \tag{2.7a}$$

$$\phi_i + u_i \frac{\phi_{i+1} - \phi_{i-1}}{2h} + \mu \frac{-\phi_{i-1} + 2\phi_i - \phi_{i+1}}{h^2} = f_i, \text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center;} \tag{2.7b}$$

$$\phi_i + u_i \frac{\phi_{i-2} - 4\phi_{i-1} + 4\phi_{i+1} - \phi_{i+2}}{4h} + \mu \frac{\phi_{i-2} - 8\phi_{i-1} + 14\phi_i - 8\phi_{i+1} + \phi_{i+2}}{4h^2} = f_i, \tag{2.7c}$$

if i is even, i.e., x_i is a cell end.

REMARK 2.1. The difference matrices D_1 and D_2 are only second order accurate in truncation errors approximating derivatives. But they give a fourth order accurate scheme for second order PDEs such as elliptic equations [13], and wave and parabolic equations [10]. However, if only using D_1 for a pure convection equation, then the scheme can only be second order accurate.

2.3. Two-dimensional fourth order scheme. Consider a uniform grid (x_i, y_j) for a rectangular domain $\bar{\Omega} = [0,1] \times [0,1]$ where $x_i = i\Delta x$, $i = 0, 1, \ldots, n_x + 1$, $\Delta x = \frac{1}{n_x+1}$ and $y_j = j\Delta y$, $j = 0, 1, \ldots, n_y + 1$, $\Delta y = \frac{1}{n_y+1}$. Assume n_x and n_y are odd and let $N_x = \frac{n_x+1}{2}$ and $N_y = \frac{n_y+1}{2}$. We consider a Q^2 finite element mesh consisting of rectangular cells $e_{kl} = [x_{2k}, x_{2k+2}] \times [y_{2l}, y_{2l+2}]$ for $k = 0, \ldots, N_x - 1$ and $l = 0, \ldots, N_y - 1$. For given functions u, v, f and g, let indices denote point values at corresponding grid points, e.g., $u_{ij} = u(x_i, y_j)$. Let u, v, f denote matrices of size $n_y \times n_x$ consisting of point values of corresponding functions, e.g.,

$$u = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1,n_x} \\ u_{21} & u_{22} & \cdots & u_{2,n_x} \\ \vdots & \vdots & & \vdots \\ u_{n_y,1} & u_{n_y,2} & \cdots & u_{n_y,n_x} \end{pmatrix}_{n_y \times n_x} , f = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1,n_x} \\ f_{21} & f_{22} & \cdots & f_{2,n_x} \\ \vdots & \vdots & & \vdots \\ f_{n_y,1} & f_{n_y,2} & \cdots & f_{n_y,n_x} \end{pmatrix}_{n_y \times n_x} .$$

Let D_{ix} and D_{iy} denote the D_i $(i = 1, 2)$ matrices for x and y variables correspondingly, e.g.,

$$D_{1x} = \frac{1}{2\Delta x} \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} \\ & -1 & 0 & 1 \\ & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} \\ & & & -1 & 0 & 1 \\ & & & & \ddots & \ddots & \ddots \\ & & & & & -1 & 0 & 1 \\ & & & & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} \\ & & & & & & -1 & 0 & 1 \end{pmatrix}_{n_x \times (n_x+2)} ,$$

$$D_{2y} = -\frac{1}{\Delta y^2} \begin{pmatrix} -1 & 2 & -1 \\ \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} \\ & -1 & 2 & -1 \\ & & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} \\ & & & -1 & 2 & -1 \\ & & & & \ddots & \ddots & \ddots \\ & & & & & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} \\ & & & & & & -1 & 2 & -1 \end{pmatrix}_{n_y \times (n_y+2)} .$$

Let ϕ be a $n_y \times n_x$ matrix consisting of interior point values of ϕ_h:

$$\phi = \begin{pmatrix} \phi_{11} & \cdots & \phi_{1,n_x} \\ \vdots & & \vdots \\ \phi_{n_y,1} & \cdots & \phi_{n_y,n_x} \end{pmatrix}_{n_y \times n_x} .$$

Let $\bar{\phi}$ be a $(n_y + 2) \times (n_x + 2)$ matrix consisting of both interior and boundary point values $\phi_h + g_I$:

$$\bar{\phi} = \begin{pmatrix} \phi_{00} & \phi_{01} & \cdots & \phi_{0,n_x} & \phi_{0,n_x+1} \\ \phi_{10} & \phi_{11} & \cdots & \phi_{1,n_x} & \phi_{1,n_x+1} \\ \vdots & \vdots & & \vdots & \vdots \\ \phi_{n_y,0} & \phi_{n_y,1} & \cdots & \phi_{n_y,n_x} & \phi_{n_y,n_x+1} \\ \phi_{n_y+1,0} & \phi_{n_y+1,2} & \cdots & \phi_{n_y+1,n_x} & \phi_{n_y+1,n_x+1} \end{pmatrix}_{n_y \times n_x} .$$

Let R_x and R_y denote restriction matrices:

$$R_x = \begin{pmatrix} \mathbf{0} & I_{n_x} & \mathbf{0} \end{pmatrix}_{n_x \times (n_x+2)}, \quad R_y = \begin{pmatrix} \mathbf{0} & I_{n_y} & \mathbf{0} \end{pmatrix}_{n_y \times (n_y+2)} .$$

Then the scheme (2.3) for interior grid points is equivalent to the linear operator form

$$\mathcal{L}(\bar{\phi}) := \phi + u.*(R_y\bar{\phi}D_{1x}^T) + v.*(D_{1y}\bar{\phi}R_x^T) - \mu(R_y\bar{\phi}D_{2x}^T + D_{2y}\bar{\phi}R_x^T) = f, \qquad (2.8)$$

where .∗ denotes entrywise product of two matrices. For the boundary points, we simply have

$$\phi_{ij} = g_{ij}, \quad \text{if } (x_i, y_j) \in \partial\Omega.$$

Define the following operators:

- ⊗ denotes Kronecker product of two matrices;
- $vec(X)$ denotes the vectorization of the matrix X by rearranging X into a vector column by column;
- $diag(\mathbf{x})$ denotes a diagonal matrix with the vector $\mathbf{x}$ as diagonal entries.

Then (2.8) is also equivalent to an abstract matrix-vector form

$$[R_x \otimes R_x + diag(vec(u))D_{1x} \otimes R_y + diag(vec(v))R_x \otimes D_{1y} - \mu(D_{2x} \otimes R_y + R_x \otimes D_{2y})]vec(\bar{\phi}) = vec(f).$$
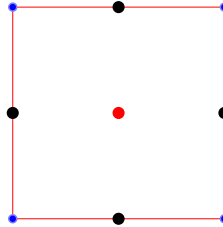
FIG. 2.2. *Three types of interior grid points: red cell center, blue knots and black edge centers for a finite element cell.*

For interior grid points, there are three types: cell center, edge center and knots. See Figure 2.2. The scheme can also be explicitly written as:

$$\phi_{ij} + \Delta t u_{ij}\frac{\phi_{i+1,j} - \phi_{i-1,j}}{2\Delta x} + \Delta t v_{ij}\frac{\phi_{i,j+1} - \phi_{i,j-1}}{2\Delta y}$$

$$+ \Delta t\mu\frac{-\phi_{i-1,j} + 2\phi_{ij} - \phi_{i+1,j}}{\Delta x^2} + \Delta t\mu\frac{-\phi_{i,j-1} + 2\phi_{ij} - \phi_{i,j+1}}{\Delta y^2} = f_{ij}, \quad \text{if } (x_i, y_j) \text{ is a cell center;}$$

$$\phi_{ij} + \Delta t u_{ij}\frac{\phi_{i+1,j} - \phi_{i-1,j}}{2\Delta x} + \Delta t v_{ij}\frac{\phi_{i,j-2} - 4\phi_{i,j-1} + 4\phi_{i,j+1} - \phi_{i,j+2}}{4\Delta y}$$

$$+ \Delta t\mu\frac{-\phi_{i-1,j} + 2\phi_{i,j} - \phi_{i+1,j}}{\Delta x^2} + \Delta t\mu\frac{\phi_{i,j-2} - 8\phi_{i,j-1} + 14\phi_{i,j} - 8\phi_{i,j+1} + \phi_{i,j+2}}{4\Delta y^2} = f_{ij}$$

if (x_i, y_j) is an interior edge center for an edge parallel to x-axis;

$$\phi_{ij} + \Delta t u_{ij}\frac{\phi_{i-2,j} - 4\phi_{i-1,j} + 4\phi_{i+1,j} - \phi_{i+2,j}}{4\Delta x} + \Delta t v_{ij}\frac{\phi_{i,j+1} - \phi_{i,j-1}}{2\Delta y}$$

$$\Delta t\mu\frac{\phi_{i-2,j} - 8\phi_{i-1,j} + 14\phi_{i,j} - 8\phi_{i+1,j} + \phi_{i+2,j}}{4\Delta x^2} + \Delta t\mu\frac{-\phi_{i,j-1} + 2\phi_{i,j} - \phi_{i,j+1}}{\Delta y^2} = f_{ij},$$

if (x_i, y_j) is an interior edge center for an edge parallel to y-axis;

$$\phi_{ij} + \Delta t u_{ij}\frac{\phi_{i-2,j} - 4\phi_{i-1,j} + 4\phi_{i+1,j} - \phi_{i+2,j}}{4\Delta x} + \Delta t v_{ij}\frac{\phi_{i,j-2} - 4\phi_{i,j-1} + 4\phi_{i,j+1} - \phi_{i,j+2}}{4\Delta y}$$

$$+ \Delta t\mu\frac{\phi_{i-2,j} - 8\phi_{i-1,j} + 14\phi_{i,j} - 8\phi_{i+1,j} + \phi_{i+2,j}}{4\Delta x^2} + \Delta t\mu\frac{\phi_{i,j-2} - 8\phi_{i,j-1} + 14\phi_{i,j} - 8\phi_{i,j+1} + \phi_{i,j+2}}{4\Delta y^2} = f_{ij}$$

if (x_i, y_j) is an interior knot.

2.4. The second order scheme. If using P^1 basis for one-dimensional case or Q^1 basis for two-dimensional case in continuous finite element method with 2-point Gauss-Lobatto quadrature in (2.3), we get exactly the classical second order centered difference scheme which can be written in the same abstract form (2.6) or (2.8) with difference matrices defined as

$$D_1 = \frac{1}{2h} \begin{pmatrix} -1 & 0 & 1 \\ & -1 & 0 & 1 \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 0 & 1 \end{pmatrix}, D_2 = -\frac{1}{h^2} \begin{pmatrix} -1 & 2 & -1 \\ & -1 & 2 & -1 \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \end{pmatrix}.$$

The scheme can also be written explicitly in one dimension as

$$\phi_i + u_i \frac{\phi_{i+1} - \phi_{i-1}}{2h} + \mu \frac{-\phi_{i-1} + 2\phi_i - \phi_{i+1}}{h^2} = f_i,$$

and in two dimensions as

$$\phi_{ij} + u_{ij} \frac{\phi_{i+1,j} - \phi_{i-1,j}}{2\Delta x} + v_{ij} \frac{\phi_{i,j+1} - \phi_{i,j-1}}{2\Delta y} + \mu \frac{-\phi_{i-1,j} + 2\phi_{ij} - \phi_{i+1,j}}{\Delta x^2} + \mu \frac{-\phi_{i,j-1} + 2\phi_{ij} - \phi_{i,j+1}}{\Delta y^2} = f_i. \tag{2.9}$$

3. Monotonicity and discrete maximum principle

3.1. Backward Euler time discretization. Now consider backward Euler time discretization for solving an initial value problem for a linear convection-diffusion equation on a square domain $\Omega = (0,1) \times (0,1)$ with Dirichlet boundary conditions:

$$\phi_t + u\phi_x + v\phi_y = \nabla \cdot (\mu \nabla \phi), \quad (x,y) \in \Omega,$$
$$\phi(x,y,0) = f(x,y), \quad (x,y) \in \Omega,$$
$$\phi(x,y,t) = g(x,y,t), \quad (x,y) \in \partial\Omega.$$

We get an elliptic equation for ϕ^{n+1}:

$$\phi^{n+1} + \Delta t u^{n+1} \phi_x^{n+1} + \Delta t v^{n+1} \phi_y^{n+1} - \Delta t \nabla \cdot (\mu \nabla \phi^{n+1}) = \phi^n, \tag{3.1}$$

with boundary condition $\phi(x,y,t_{n+1}) = g(x,y,t_{n+1}), (x,y) \in \partial\Omega$. With the same notations in Section 2, the variational difference scheme for (3.1) at interior grid points is

$$\mathcal{L}(\bar{\phi}^{n+1}) := \phi^{n+1} + \Delta t \left[u^{n+1}.*(R_y \bar{\phi}^{n+1} D_{1x}^T) + v^{n+1}.*(D_{1y} \bar{\phi}^{n+1} R_x^T) - \mu(R_y \bar{\phi}^{n+1} D_{2x}^T + D_{2y} \bar{\phi}^{n+1} R_x^T) \right] = \phi^n. \tag{3.2a}$$

Now define a linear operator $\bar{\mathcal{L}} : \mathbb{R}^{(n_y+2) \times (n_x+2)} \to \mathbb{R}^{(n_y+2) \times (n_x+2)}$:

$$\bar{\mathcal{L}}(\bar{\phi}^{n+1})_{ij} := \begin{cases} \mathcal{L}(\bar{\phi}^{n+1})_{ij}, & (x_i, y_j) \in \Omega, \\ \phi_{ij}^{n+1}, & (x_i, y_j) \in \partial\Omega. \end{cases}$$

Then the finite difference scheme can be written as

$$\begin{cases} \bar{\mathcal{L}}(\bar{\phi}^{n+1})_{ij} = \phi_{ij}^n, & (x_i, y_j) \in \Omega, \\ \bar{\mathcal{L}}(\bar{\phi}^{n+1})_{ij} = g_{ij}^{n+1}, & (x_i, y_j) \in \partial\Omega. \end{cases} \tag{3.2b}$$

We first have two straightforward results:

THEOREM 3.1. *Let $\bar{1}$ denote a matrix of the same size as $\bar{\phi}$ with all entries being 1; for the scheme operator in (3.2), $\bar{\mathcal{L}}(\bar{1}) = \bar{1}$.*

Proof.   Notice that row sums of D_1 and D_2 in second order and fourth order accurate schemes are all zeros, thus $D_1\mathbf{1} = D_2\mathbf{1} = \mathbf{0}$, which implies the result.   □

THEOREM 3.2.   *For the scheme* (3.2b), *let* $\bar{L}$ *be the matrix representation of the linear operator* $\bar{\mathcal{L}}$. *If the inverse matrix has non-negative entries, i.e.,* $\bar{L}^{-1} \geq 0$, *then the finite difference scheme satisfies a discrete maximum principle:*

$$\min\left\{\min_{(x_i,y_j)\in\Omega}\phi_{ij}^n, \min_{(x_i,y_j)\in\partial\Omega}g_{ij}^{n+1}\right\} \leq \phi_{ij}^{n+1} \leq \max\left\{\max_{(x_i,y_j)\in\Omega}\phi_{ij}^n, \max_{(x_i,y_j)\in\partial\Omega}g_{ij}^{n+1}\right\}. \quad (3.3)$$

Proof.   Let $\mathbf{1}$ denote the vector consisting of ones, then Theorem 3.1 implies $\bar{L}\mathbf{1} = \mathbf{1}$ thus $\bar{L}^{-1}\mathbf{1} = \mathbf{1}$. Since all entries in $\bar{L}^{-1}$ are non-negative, each row in $\bar{L}^{-1}$ forms a set of coefficients for a convex combination, which implies the maximum principle.   □

3.2. M-matrix and the second order scheme.   Nonsingular M-matrices are inverse-positive matrices, which is the main tool for proving inverse positivity. There are many equivalent definitions or characterizations of M-matrices, see [17]. One convenient sufficient but not necessary characterization of nonsingular M-matrices is as follows:

THEOREM 3.3.   *For a real square matrix A with positive diagonal entries and non-positive off-diagonal entries, A is a nonsingular M-matrix if all the row sums of A are non-negative and at least one row sum is positive.*

Proof.   By condition C_{10} in [17], A is a nonsingular M-matrix if and only if $A + aI$ is nonsingular for any $a \geq 0$. Since all the row sums of A are non-negative and at least one row sum is positive, the matrix A is irreducibly diagonally dominant, thus nonsingular, and $A + aI$ is strictly diagonally dominant, thus nonsingular for any $a > 0$.   □

By condition K_{35} in [17], a sufficient and necessary characterization of nonsingular M-matrices is the following:

THEOREM 3.4.   *For a real square matrix A with positive diagonal entries and non-positive 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.*

If using second order scheme (2.9) in (3.2b), then the scheme operator $\bar{\mathcal{L}}$ acting on $\bar{\phi}$ is given as

$$\begin{cases} \bar{\mathcal{L}}(\bar{\phi})_{ij} = & \phi_{ij} + \Delta t u_{ij}\frac{\phi_{i+1,j}-\phi_{i-1,j}}{2\Delta x} + \Delta t v_{ij}\frac{\phi_{i,j+1}-\phi_{i,j-1}}{2\Delta y} \\ & + \Delta t\mu\frac{-\phi_{i-1,j}+2\phi_{ij}-\phi_{i+1,j}}{\Delta x^2} + \Delta t\mu\frac{-\phi_{i,j-1}+2\phi_{ij}-\phi_{i,j+1}}{\Delta y^2}, & (x_i,y_j)\in\Omega, \\ \bar{\mathcal{L}}(\bar{\phi})_{ij} = & \phi_{ij}, \quad (x_i,y_j)\in\partial\Omega. \end{cases}$$

For interior points $(x_i,y_j)\in\Omega$, we have

$$\bar{\mathcal{L}}(\bar{\phi})_{ij} = \left(1+\frac{2\mu\Delta t}{\Delta x^2}+\frac{2\mu\Delta t}{\Delta y^2}\right)\phi_{ij} - \frac{\Delta t}{\Delta x}\left(\frac{\mu}{\Delta x}-\frac{u_{ij}}{2}\right)(\phi_{i+1,j}+\phi_{i-1,j}) - \frac{\Delta t}{\Delta y}\left(\frac{\mu}{\Delta y}-\frac{v_{ij}}{2}\right)(\phi_{i,j+1}+\phi_{i,j-1}).$$

Assume

$$\Delta x \max_{ij}|u_{ij}| \leq 2\mu, \quad \Delta y \max_{ij}|v_{ij}| \leq 2\mu, \quad (3.4)$$

then all off-diagonal entries of $\bar{L}$ will be non-positive, thus $\bar{L}$ is an M-matrix.

THEOREM 3.5.   *Under the mesh constraints* (3.4), *the second order accurate scheme* (2.9) *is monotone, i.e.,* $\bar{L}^{-1} \geq 0$, *and satisfies the discrete maximum principle* (3.3).

3.3. Lorenz's condition for the fourth order scheme. Unfortunately, almost all high order schemes will lead to positive off-diagonal entries in the system matrix, which can no longer be an M-matrix, see [4]. In [15], Lorenz proposed a convenient condition under which a matrix can be shown to be a product of M-matrices. We briefly review Lorenz's condition in this subsection.

DEFINITION 3.1. *Let $\mathcal{N} = \{1,2,\ldots,n\}$. 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*

$$\forall i_0 \in \mathcal{N}_1, \exists i_r \in \mathcal{N}_2, \exists i_1, \ldots, i_{r-1} \in \mathcal{N} \quad s.t. \quad a_{i_{k-1} i_k} \neq 0, \quad k = 1, \cdots, r. \tag{3.5}$$

If perceiving A as a directed graph adjacency matrix of vertices labeled by $\mathcal{N}$, then (3.5) 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, 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 result was proven in [15]. See also [12] for a detailed proof.

THEOREM 3.6 (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*

$$A_d + A^z \text{ is a nonsingular M-matrix}, \tag{3.6a}$$

$$A_a^+ \leq A^z A_d^{-1} A^s \text{ or equivalently } \forall a_{ij} > 0 \text{ with } i \neq j, a_{ij} \leq \sum_{k=1}^{n} a_{ik}^z a_{kk}^{-1} a_{kj}^s, \tag{3.6b}$$

$$\exists \mathbf{e} \in \mathbb{R}^n \setminus \{\mathbf{0}\}, \mathbf{e} \geq 0 \text{ with } A\mathbf{e} \geq 0 \text{ s.t. } A^z \text{ or } A^s \text{ connects } \mathcal{N}^0(A\mathbf{e}) \text{ with } \mathcal{N}^+(A\mathbf{e}). \tag{3.6c}$$

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

In general, the condition (3.6c) can be difficult to verify. But for the finite difference schemes, the vector $\mathbf{e}$ can be taken as $\mathbf{1}$ to simply (3.6c). In particular, for the fourth order accurate scheme (3.2), we have $\bar{\mathcal{L}}(\bar{\mathbf{1}}) = \bar{\mathbf{1}}$, thus $\bar{L}\mathbf{1} = \mathbf{1}$. Therefore, $\mathcal{N}^0(\bar{L}\mathbf{1}) = \emptyset$ implies that the condition (3.6c) is trivially satisfied. So we can state a simpler Lorenz's condition for the scheme considered in this paper:

THEOREM 3.7. *Let A denote the matrix representation of a new linear operator $\mathcal{A} := \frac{h^2}{\mu \Delta t} \bar{\mathcal{L}}$ for the scheme (3.2), with a corresponding matrix $A := \frac{h^2}{\mu \Delta t} \bar{L}$. Assume A_a^- has a decomposition $A_a^- = A^z + A^s$ with $A^s \leq 0$ and $A^z \leq 0$. Then $A^{-1} \geq 0$ if the following are satisfied:*

(1) *$A_d + A^z$ is a nonsingular M-matrix;*

(2) *$A_a^+ \leq A^z A_d^{-1} A^s$.*

3.4. Verification of Lorenz's condition: one-dimensional case. We first show how to verify Theorem 3.7 for the one-dimensional version of (3.2). For simplicity, let $c = \frac{h^2}{\mu \Delta t}$, then we have the linear operator $\mathcal{A} := \frac{h^2}{\mu \Delta t} \bar{\mathcal{L}} = c\bar{\mathcal{L}}$ with a corresponding matrix $A := c\bar{L}$. Applying (2.7) to (3.1), we get the explicit expression of $\mathcal{A}(\bar{\phi})$ as

$$\mathcal{A}(\bar{\phi})_0 = c\phi_0, \quad \mathcal{A}(\bar{\phi})_{n+1} = c\phi_{n+1};$$

$$\mathcal{A}(\bar{\phi})_i = c\phi_i + \frac{hu_i}{2\mu}(\phi_{i+1} - \phi_{i-1}) + (-\phi_{i-1} + 2\phi_i - \phi_{i+1}), \quad \text{if } x_i \text{ is a cell center};$$

$$\mathcal{A}(\bar{\phi})_i = c\phi_i + \frac{hu_i}{\mu} \frac{\phi_{i-2} - 4\phi_{i-1} + 4\phi_{i+1} - \phi_{i+2}}{4} + \frac{\phi_{i-2} - 8\phi_{i-1} + 14\phi_i - 8\phi_{i+1} + \phi_{i+2}}{4},$$

if x_i is a cell end.

3.4.1. A splitting $A_a^- = A^z + A^s$. In order to have fixed signs for all entries, we assume

$$h|u_i| \le 2\mu, \quad \forall i. \tag{3.7}$$

Then we have

$$\mathcal{A}_d(\bar{\phi})_0 = c\phi_0, \quad \mathcal{A}^d(\bar{\phi})_{n+1} = c\phi_{n+1},$$

$$\mathcal{A}_d(\bar{\phi})_i = c\phi_i + 2\phi_i, \quad \text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center};$$

$$\mathcal{A}_d(\bar{\phi})_i = c\phi_i + \frac{7}{2}\phi_i, \quad \text{if } i \text{ is even, i.e., } x_i \text{ is a cell end}.$$

$$\mathcal{A}_a^+(\bar{\phi})_0 = \mathcal{A}^+(\bar{\phi})_{n+1} = 0,$$

$$\mathcal{A}_a^+(\bar{\phi})_i = 0, \quad \text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center};$$

$$\mathcal{A}_a^+(\bar{\phi})_i = \left(1 + \frac{hu_i}{\mu}\right)\frac{1}{4}\phi_{i-2} + \left(1 - \frac{hu_i}{\mu}\right)\frac{1}{4}\phi_{i+2}, \quad \text{if } i \text{ is even, i.e., } x_i \text{ is a cell end}.$$

$$\mathcal{A}_a^-(\bar{\phi})_0 = \mathcal{A}^-(\bar{\phi})_{n+1} = 0,$$

$$\mathcal{A}_a^-(\bar{\phi})_i = -\left(1 + \frac{hu_i}{2\mu}\right)\phi_{i-1} - \left(1 - \frac{hu_i}{2\mu}\right)\phi_{i+1}, \quad \text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center};$$

$$\mathcal{A}_a^-(\bar{\phi})_i = -2\left(1 + \frac{hu_i}{2\mu}\right)\phi_{i-1} - 2\left(1 - \frac{hu_i}{2\mu}\right)\phi_{i+1} \quad \text{if } i \text{ is even, i.e., } x_i \text{ is a cell end}.$$

Next, we define a splitting $A_a^- = A^z + A^s$ as:

$$\mathcal{A}^z(\bar{\phi})_i = 0, \quad \text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center};$$

$$\mathcal{A}^z(\bar{\phi})_i = -2\left(1 - \frac{hu_i}{2\mu}\right)\phi_{i-1} - 2\left(1 + \frac{hu_i}{2\mu}\right)\phi_{i+1} \quad \text{if } i \text{ is even, i.e., } x_i \text{ is a cell end}.$$

$$\mathcal{A}^s(\bar{\phi})_i = -\left(1 - \frac{hu_i}{2\mu}\right)\phi_{i-1} - \left(1 + \frac{hu_i}{2\mu}\right)\phi_{i+1}, \quad \text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center};$$

$$\mathcal{A}^s(\bar{\phi})_i = 0, \quad \text{if } i \text{ is even, i.e., } x_i \text{ is a cell end}.$$

3.4.2. Verification of $A_d + A^z$ being an M-matrix. For simplicity, define $B = A_d + A^z$, then the corresponding linear operator $\mathcal{B} = \mathcal{A}_d + \mathcal{A}^z$:

$$\mathcal{B}(\bar{\phi})_0 = c\phi_0, \quad \mathcal{B}(\bar{\phi})_{n+1} = c\phi_{n+1},$$

$$\mathcal{B}(\bar{\phi})_i = c\phi_i + 2\phi_i, \quad \text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center;}$$

$$\mathcal{B}(\bar{\phi})_i = \left(c + \frac{7}{2}\right)\phi_i - 2\left(1 - \frac{hu_i}{2\mu}\right)\phi_{i-1} - 2\left(1 + \frac{hu_i}{2\mu}\right)\phi_{i+1},$$

if i is even, i.e., x_i is a cell end.

Since for even i, $\mathcal{B}(\mathbf{1}) = c - \frac{1}{2}$ which is for small c, thus Theorem 3.3 cannot be applied. Define the following linear operator $\mathcal{D}$:

$$\mathcal{D}(\bar{\phi})_0 = \phi_0, \quad \mathcal{D}(\bar{\phi})_{n+1} = \phi_{n+1},$$

$$\mathcal{D}(\bar{\phi})_i = \frac{1}{2}\phi_i, \quad \text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center;}$$

$$\mathcal{D}(\bar{\phi})_i = \phi_i, \quad \text{if } i \text{ is even, i.e., } x_i \text{ is a cell end.}$$

Let D be the matrix representing the operator $\mathcal{D}$, then D is a diagonal matrix with positive diagonal entries. And we have

$$\mathcal{B}[\mathcal{D}(\bar{\phi})]_0 = c\phi_0, \quad \mathcal{B}[\mathcal{D}(\bar{\phi})]_{n+1} = c\phi_{n+1},$$

$$\mathcal{B}[\mathcal{D}(\bar{\phi})]_i = (c+2)\mathcal{D}(\bar{\phi})_i = \frac{c+2}{2}\bar{\phi}_i, \quad \text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center;}$$

$$\mathcal{B}[\mathcal{D}(\bar{\phi})]_i = \left(c + \frac{7}{2}\right)D(\bar{\phi})_i - 2\left(1 - \frac{hu_i}{2\mu}\right)D(\bar{\phi})_{i-1} - 2\left(1 + \frac{hu_i}{2\mu}\right)D(\bar{\phi})_{i+1}$$

$$= \left(c + \frac{7}{2}\right)\phi_i - \left(1 - \frac{hu_i}{2\mu}\right)\phi_{i-1} - \left(1 + \frac{hu_i}{2\mu}\right)\phi_{i+1}$$

if i is even, i.e., x_i is a cell end.

Then it is easy to see that $\mathcal{B}[\mathcal{D}(\bar{\mathbf{1}})] > 0$, thus $BD\mathbf{1} > 0$ for any $c > 0$. By Theorem 3.4, BD has positive row sums, thus $A_d + A^z = B$ is a nonsingular M-matrix.

3.4.3. Verification of $A_a^+ \leq A^z A_d^{-1} A^s$. Since $A^z A_d^{-1} A^s \geq 0$, in order to verify $A_a^+ \leq A^z A_d^{-1} A^s$, we only need to compare $\mathcal{A}_a^+(\bar{\phi})_i$ with $\mathcal{A}^z[\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))]_i$ for even i. For a cell end x_i, we have

$$\mathcal{A}^z[\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))]_i = -2\left(1 - \frac{hu_i}{2\mu}\right)\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))_{i-1} - 2\left(1 + \frac{hu_i}{2\mu}\right)\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))_{i+1}$$

$$= -2\left(1 - \frac{hu_i}{2\mu}\right)\frac{1}{c+2}\mathcal{A}^s(\bar{\phi})_{i-1} - 2\left(1 + \frac{hu_i}{2\mu}\right)\frac{1}{c+2}\mathcal{A}^s(\bar{\phi})_{i+1}$$

$$= 2\frac{\left(1 - \frac{hu_i}{2\mu}\right)\left(1 - \frac{hu_{i-1}}{2\mu}\right)}{c+2}\phi_{i-2} + 2\frac{\left(1 - \frac{hu_i}{2\mu}\right)\left(1 + \frac{hu_{i-1}}{2\mu}\right)}{c+2}\phi_i$$

$$+ 2\frac{\left(1 + \frac{hu_i}{2\mu}\right)\left(1 - \frac{hu_{i+1}}{2\mu}\right)}{c+2}\phi_i + 2\frac{\left(1 + \frac{hu_i}{2\mu}\right)\left(1 + \frac{hu_{i+1}}{2\mu}\right)}{c+2}\phi_{i+2}.$$

It suffices to have

$$\left(1 + \frac{hu_i}{\mu}\right)\frac{1}{4} \leq 2\frac{\left(1 - \frac{hu_i}{2\mu}\right)\left(1 - \frac{hu_{i-1}}{2\mu}\right)}{c+2}, \quad \left(1 - \frac{hu_i}{\mu}\right)\frac{1}{4} \leq 2\frac{\left(1 + \frac{hu_i}{2\mu}\right)\left(1 + \frac{hu_{i+1}}{2\mu}\right)}{c+2},$$

which are equivalent to

$$(12 + 2c)\frac{hu_i}{2\mu} + 8\frac{hu_{i-1}}{2\mu} - 8\frac{hu_{i-1}}{2\mu}\frac{hu_i}{2\mu} \leq 6 - c, \quad -(12 + 2c)\frac{hu_i}{2\mu} - 8\frac{hu_{i+1}}{2\mu} - 8\frac{hu_{i-1}}{2\mu}\frac{hu_i}{2\mu} \leq 6 - c.$$

Let $a = \max_i |u_i| \frac{h}{2\mu}$, then it suffices to require

$$(12 + 2c)a + 8a + 8a^2 \le 6 - c \Longleftrightarrow 8a^2 + (20 + 2c)a - (6 - c) \le 0.$$

From the inequality above, we get $a \le \frac{\sqrt{(c+6)^2 + 112} - (c+10)}{8}$ for a fixed $c > 0$. Since $\frac{\sqrt{(c+6)^2 + 112} - (c+10)}{8} > 0$ implies $c < 6$, we have $c \in (0,6)$.

For a fixed $a > 0$, then we need $c \le \frac{-8a^2 - 20a + 6}{2a + 1}$. For $\frac{-8a^2 - 20a + 6}{2a + 1} > 0$, we must have $a < \frac{\sqrt{37} - 5}{4}$.

3.4.4. Sufficient conditions in 1-D. Now we can summarize all the constraints to apply Theorem 3.7.

THEOREM 3.8. *Let $\|u\|_\infty = \max_i |u_i|$. For the scheme (2.7) to be inverse positive, i.e., $\bar{L}^{-1} \ge 0$, the following conditions are sufficient:*

- *For a mesh size h satisfying $h\frac{\|u\|_\infty}{2\mu} = a < \frac{\sqrt{37}-5}{4} \approx 0.271$, time step Δt satisfies $\Delta t \frac{\mu}{h^2} \ge \frac{2a+1}{-8a^2 - 20a + 6}$.*

- *For a time step Δt satisfying $\Delta t \frac{\mu}{h^2} = \frac{1}{c} > \frac{1}{6}$, the mesh size h satisfies $h\frac{\|u\|_\infty}{\mu} \le \frac{\sqrt{(c+6)^2 + 112} - (c+10)}{4}$.*

In particular, the following are convenient explicit sufficient mesh constraints for the inverse positivity:

- *For a mesh size h satisfying $h\frac{\|u\|_\infty}{\mu} \le \frac{1}{2}$, time step satisfies $\Delta t \frac{\mu}{h^2} \ge 3$.*

- *For a time step Δt satisfying $\Delta t \frac{\mu}{h^2} \ge \frac{1}{2}$, the mesh size h satisfies $h\frac{\|u\|_\infty}{\mu} \le \frac{1}{2}$.*

REMARK 3.1. We emphasize that the time step and mesh size conditions in Theorem 3.8 are not sharp, but the constraint $\Delta t \ge \mathcal{O}(h^2)$ is necessary for monotonicity of the fourth order accurate scheme (2.7) in numerical tests. For achieving temporal convergence with monotonicity always satisfied, we can set $\Delta t \to 0$ with $h \to 0$ satisfying $\Delta t \frac{\mu}{h^2} \ge 3$. On the other hand, if setting $\Delta t \to 0$ for fixed h, monotonicity will be simply lost because the semi-discrete system of ordinary differential equations might not be as bound-preserving as the PDE it approximates. For example, consider solving the simple heat equation $u_t = \Delta u$, the semi-discrete scheme with space variable discretized by finite difference can be written as $\frac{d}{dt}\mathbf{u} = \Delta_h \mathbf{u}$. If using second order centered difference, then $\mathbf{u}(t) = e^{\Delta_h t}\mathbf{u}(0)$ will preserve the bound or positivity of $\mathbf{u}(0)$ because $-\Delta_h$ is a diagonally dominant matrix with all off-diagonal entries being non-positive. If using the Q^2 spectral element method, then $-\Delta_h$ is not diagonally dominant because of the positive off-diagonal entries, thus it is easy to construct a non-negative initial condition $\mathbf{u}(0)$ such that $\mathbf{u}(t) = e^{\Delta_h t}\mathbf{u}(0)$ contains negative values. This simply implies that we should not approximate the ODE system $\mathbf{u}(t) = e^{\Delta_h t}\mathbf{u}(0)$ too accurately if high order spatial discretization is used and positivity is a concern. On the other hand, one can still achieve convergence to the PDE solution with positivity enforced in the high order spatial discretization, if the constraints such as the ones in Theorem 3.8 are enforced.

3.5. Verification of Lorenz's condition: two-dimensional case. For simplicity, we only consider the case $\Delta x = \Delta y = h$. Let $c = \frac{h^2}{\mu \Delta t}$. For the linear operator $\mathcal{A} = \frac{h^2}{\mu \Delta t} \bar{\mathcal{L}} : \mathbb{R}^{(n_y+2) \times (n_x+2)} \to R^{(n_y+2) \times (n_x+2)}$, we have

$$\mathcal{A}(\bar{\phi})_{ij} = c\phi_{ij}, \quad (x_i, y_j) \text{ is a boundary point;}$$

$$\mathcal{A}(\bar{\phi})_{ij} = c\phi_{ij} + \frac{hu_{ij}}{2\mu}(\phi_{i+1,j} - \phi_{i-1,j}) + \frac{hv_{ij}}{2\mu}(\phi_{i,j+1} - \phi_{i,j-1})$$

$$+(-\phi_{i-1,j}+2\phi_{ij}-\phi_{i+1,j})+(-\phi_{i,j-1}+2\phi_{ij}-\phi_{i,j+1}), \text{ if } (x_i,y_j) \text{ is a cell center;}$$

$$\mathcal{A}(\bar{\phi})_{ij}=c\phi_{ij}+\frac{hu_{ij}}{2\mu}(\phi_{i+1,j}-\phi_{i-1,j})+\frac{hv_{ij}}{4\mu}(\phi_{i,j-2}-4\phi_{i,j-1}+4\phi_{i,j+1}-\phi_{i,j+2})$$

$$+(-\phi_{i-1,j}+2\phi_{i,j}-\phi_{i+1,j})+\frac{\phi_{i,j-2}-8\phi_{i,j-1}+14\phi_{i,j}-8\phi_{i,j+1}+\phi_{i,j+2}}{4}$$

if (x_i,y_j) is an interior edge center for an edge parallel to x-axis;

$$\mathcal{A}(\bar{\phi})_{ij}=c\phi_{ij}+\frac{hu_{ij}}{4\mu}(\phi_{i-2,j}-4\phi_{i-1,j}+4\phi_{i+1,j}-\phi_{i+2,j})+\frac{hv_{ij}}{2\mu}(\phi_{i,j+1}-\phi_{i,j-1})$$

$$+\frac{\phi_{i-2,j}-8\phi_{i-1,j}+14\phi_{i,j}-8\phi_{i+1,j}+\phi_{i+2,j}}{4}+(-\phi_{i,j-1}+2\phi_{i,j}-\phi_{i,j+1}),$$

if (x_i,y_j) is an interior edge center for an edge parallel to y-axis;

$$\mathcal{A}(\bar{\phi})_{ij}=c\phi_{ij}+\frac{hu_{ij}}{\mu}\frac{\phi_{i-2,j}-4\phi_{i-1,j}+4\phi_{i+1,j}-\phi_{i+2,j}}{4}+\frac{hv_{ij}}{\mu}\frac{\phi_{i,j-2}-4\phi_{i,j-1}+4\phi_{i,j+1}-\phi_{i,j+2}}{4}$$

$$+\frac{\phi_{i-2,j}-8\phi_{i-1,j}+14\phi_{i,j}-8\phi_{i+1,j}+\phi_{i+2,j}}{4}+\frac{\phi_{i,j-2}-8\phi_{i,j-1}+14\phi_{i,j}-8\phi_{i,j+1}+\phi_{i,j+2}}{4}$$

if (x_i,y_j) is an interior knot.

3.5.1. Splitting of negative off-diagonal entries.

In order to have fixed signs for all entries, we assume $h\max_{ij}\{|u_{ij}|,|v_{ij}|\}\le 2\mu$ for all i,j. Then we have

$$\mathcal{A}_d(\bar{\phi})_{ij}=c\phi_{ij},\quad (x_i,y_j) \text{ is a boundary point;}$$

$$\mathcal{A}_d(\bar{\phi})_{ij}=(c+4)\phi_{ij},(x_i,y_j) \text{ is a cell center;}$$

$$\mathcal{A}_d(\bar{\phi})_{ij}=\left(c+\frac{11}{2}\right)\phi_{ij},(x_i,y_j) \text{ is an interior edge center;}$$

$$\mathcal{A}_d(\bar{\phi})_{ij}=(c+7)\phi_{ij},(x_i,y_j) \text{ is an interior knot.}$$

For positive off-diagonal parts, we have:

$$\mathcal{A}_a^+(\bar{\phi})_{ij}=0,\quad (x_i,y_j) \text{ is a boundary point;}$$

$$\mathcal{A}_a^+(\bar{\phi})_{ij}=\frac{1}{4}\left(1+\frac{hv_{ij}}{\mu}\right)\phi_{i,j-2}+\frac{1}{4}\left(1-\frac{hv_{ij}}{\mu}\right)\phi_{i,j+2},$$

$$(x_i,y_j) \text{ is an interior edge center for an edge parallel to x-axis;}$$

$$\mathcal{A}_a^+(\bar{\phi})_{ij}=\frac{1}{4}\left(1+\frac{hu_{ij}}{\mu}\right)\phi_{i-2,j}+\frac{1}{4}\left(1-\frac{hu_{ij}}{\mu}\right)\phi_{i+2,j},$$

$$(x_i,y_j) \text{ is an interior edge center for an edge parallel to y-axis;}$$

$$\mathcal{A}_a^+(\bar{\phi})_{ij}=\frac{1}{4}\left(1+\frac{hv_{ij}}{\mu}\right)\phi_{i,j-2}+\frac{1}{4}\left(1-\frac{hv_{ij}}{\mu}\right)\phi_{i,j+2}$$

$$+\frac{1}{4}\left(1+\frac{hu_{ij}}{\mu}\right)\phi_{i-2,j}+\frac{1}{4}\left(1-\frac{hu_{ij}}{\mu}\right)\phi_{i+2,j},$$

(x_i,y_j) is an interior knot.

Then we defined a splitting $A_a^-=A^z+A^s$ as:

$$\mathcal{A}^z(\bar{\phi})_{ij}=0,\quad (x_i,y_j) \text{ is a boundary point;}$$

$$\mathcal{A}^z(\bar{\phi})_{ij}=0,\quad (x_i,y_j) \text{ is a cell center;}$$

$$\mathcal{A}^z(\bar{\phi})_{ij}=-2\left(1+\frac{hv_{ij}}{2\mu}\right)\phi_{i,j-1}-2\left(1-\frac{hv_{ij}}{2\mu}\right)\phi_{i,j+1}$$

(x_i, y_j) is an interior edge center for an edge parallel to x-axis;

$$\mathcal{A}^z(\bar{\phi})_{ij} = -2\left(1 + \frac{hu_{ij}}{2\mu}\right)\phi_{i-1,j} - 2\left(1 - \frac{hu_{ij}}{2\mu}\right)\phi_{i+1,j},$$

(x_i, y_j) is an interior edge center for an edge parallel to y-axis;

$$\mathcal{A}^z(\bar{\phi})_{ij} = -2\left(1 + \frac{hv_{ij}}{2\mu}\right)\phi_{i,j-1} - 2\left(1 - \frac{hv_{ij}}{2\mu}\right)\phi_{i,j+1}$$
$$-2\left(1 + \frac{hu_{ij}}{2\mu}\right)\phi_{i-1,j} - 2\left(1 - \frac{hu_{ij}}{2\mu}\right)\phi_{i+1,j},$$

(x_i, y_j) is an interior knot.

$$\mathcal{A}^s(\bar{\phi})_{ij} = 0, \quad (x_i, y_j) \text{ is a boundary point;}$$
$$\mathcal{A}^s(\bar{\phi})_{ij} = -\left(1 - \frac{hu_{ij}}{2\mu}\right)\phi_{i+1,j} - \left(1 + \frac{hu_{ij}}{2\mu}\right)\phi_{i-1,j}$$
$$-\left(1 - \frac{hv_{ij}}{2\mu}\right)\phi_{i,j+1} - \left(1 + \frac{hv_{ij}}{2\mu}\right)\phi_{i,j-1},$$

(x_i, y_j) is a cell center;

$$\mathcal{A}^s(\bar{\phi})_{ij} = -\left(1 - \frac{hu_{ij}}{2\mu}\right)\phi_{i+1,j} - \left(1 + \frac{hu_{ij}}{2\mu}\right)\phi_{i-1,j}$$

(x_i, y_j) is an interior edge center for an edge parallel to x-axis;

$$\mathcal{A}^s(\bar{\phi})_{ij} = -\left(1 - \frac{hv_{ij}}{2\mu}\right)\phi_{i,j+1} - \left(1 + \frac{hv_{ij}}{2\mu}\right)\phi_{i,j-1},$$

(x_i, y_j) is an interior edge center for an edge parallel to y-axis;

$$\mathcal{A}^s(\bar{\phi})_{ij} = 0, \quad (x_i, y_j) \text{ is an interior knot.}$$

3.5.2. Verification of $A_d + A^z$ being an M-matrix. Let $\mathcal{B} = \mathcal{A}_d + \mathcal{A}^z$, then we have

$$\mathcal{B}(\bar{\phi})_{ij} = c\phi_{ij}, \quad (x_i, y_j) \text{ is a boundary point;}$$
$$\mathcal{B}(\bar{\phi})_{ij} = (c+4)\phi_{ij}, \quad (x_i, y_j) \text{ is a cell center;}$$
$$\mathcal{B}(\bar{\phi})_{ij} = \left(c + \frac{11}{2}\right)\phi_{ij} - 2\left(1 + \frac{hv_{ij}}{2\mu}\right)\phi_{i,j-1} - 2\left(1 - \frac{hv_{ij}}{2\mu}\right)\phi_{i,j+1},$$

(x_i, y_j) is an interior edge center for an edge parallel to x-axis;

$$\mathcal{B}(\bar{\phi})_{ij} = \left(c + \frac{11}{2}\right)\phi_{ij} - 2\left(1 + \frac{hu_{ij}}{2\mu}\right)\phi_{i-1,j} - 2\left(1 - \frac{hu_{ij}}{2\mu}\right)\phi_{i+1,j},$$

(x_i, y_j) is an interior edge center for an edge parallel to y-axis;

$$\mathcal{B}(\bar{\phi})_{ij} = (c+7)\phi_{ij} - 2\left(1 + \frac{hv_{ij}}{2\mu}\right)\phi_{i,j-1} - 2\left(1 - \frac{hv_{ij}}{2\mu}\right)\phi_{i,j+1}$$
$$-2\left(1 + \frac{hu_{ij}}{2\mu}\right)\phi_{i-1,j} - 2\left(1 - \frac{hu_{ij}}{2\mu}\right)\phi_{i+1,j},$$

(x_i, y_j) is an interior knot.

Define a positive diagonal operator $\mathcal{D}$ as

$$\mathcal{D}(\bar{\phi})_{ij} = \phi_{ij}, \quad (x_i, y_j) \text{ is a boundary point, or a cell center, or an interior knot;}$$

$$\mathcal{D}(\bar{\phi})_{ij} = \frac{3}{4}\phi_{ij}, \quad (x_i, y_j) \text{ is an interior edge center.}$$

Then we have

$$\mathbf{B}[\mathcal{D}(\bar{\phi})]_{ij} = c\mathcal{D}(\bar{\phi})_{ij} = c\phi_{ij}, \quad (x_i, y_j) \text{ is a boundary point;}$$

$$\mathbf{B}[\mathcal{D}(\bar{\phi})]_{ij} = (c+4)\mathcal{D}(\bar{\phi})_{ij} = (c+4)\phi_{ij}, \quad (x_i, y_j) \text{ is a cell center;}$$

$$\mathbf{B}[\mathcal{D}(\bar{\phi})]_{ij} = \left(c + \frac{11}{2}\right)\mathcal{D}(\bar{\phi})_{ij} - 2\left(1 + \frac{hv_{ij}}{2\mu}\right)\mathcal{D}(\bar{\phi})_{i,j-1} - 2\left(1 - \frac{hv_{ij}}{2\mu}\right)\mathcal{D}(\bar{\phi})_{i,j+1}$$

$$= \left(\frac{3}{4}c + \frac{33}{8}\right)\phi_{ij} - 2\left(1 + \frac{hv_{ij}}{2\mu}\right)\phi_{i,j-1} - 2\left(1 - \frac{hv_{ij}}{2\mu}\right)\phi_{i,j+1},$$

$$(x_i, y_j) \text{ is an interior edge center for an edge parallel to x-axis;}$$

$$\mathbf{B}[\mathcal{D}(\bar{\phi})]_{ij} = \left(c + \frac{11}{2}\right)\mathcal{D}(\bar{\phi})_{ij} - 2\left(1 + \frac{hu_{ij}}{2\mu}\right)\mathcal{D}(\bar{\phi})_{i-1,j} - 2\left(1 - \frac{hu_{ij}}{2\mu}\right)\mathcal{D}(\bar{\phi})_{i+1,j}$$

$$= \left(\frac{3}{4}c + \frac{33}{8}\right)\phi_{ij} - 2\left(1 + \frac{hu_{ij}}{2\mu}\right)\phi_{i-1,j} - 2\left(1 - \frac{hu_{ij}}{2\mu}\right)\phi_{i+1,j},$$

$$(x_i, y_j) \text{ is an interior edge center for an edge parallel to y-axis;}$$

$$\mathbf{B}[\mathcal{D}(\bar{\phi})]_{ij} = (c+7)\mathcal{D}(\bar{\phi})_{ij} - 2\left(1 + \frac{hv_{ij}}{2\mu}\right)\mathcal{D}(\bar{\phi})_{i,j-1} - 2\left(1 - \frac{hv_{ij}}{2\mu}\right)\mathcal{D}(\bar{\phi})_{i,j+1}$$

$$- 2\left(1 + \frac{hu_{ij}}{2\mu}\right)\mathcal{D}(\bar{\phi})_{i-1,j} - 2\left(1 - \frac{hu_{ij}}{2\mu}\right)\mathcal{D}(\bar{\phi})_{i+1,j},$$

$$= (c+7)\phi_{ij} - \frac{3}{2}\left(1 + \frac{hv_{ij}}{2\mu}\right)\phi_{i,j-1} - \frac{3}{2}\left(1 - \frac{hv_{ij}}{2\mu}\right)\phi_{i,j+1}$$

$$- \frac{3}{2}\left(1 + \frac{hu_{ij}}{2\mu}\right)\phi_{i-1,j} - \frac{3}{2}\left(1 - \frac{hu_{ij}}{2\mu}\right)\phi_{i+1,j},$$

$$(x_i, y_j) \text{ is an interior knot.}$$

And the matrix BD has positive row sums for any $c > 0$ due to the following:

$$\mathbf{B}[\mathcal{D}(\bar{1})]_{ij} = c, \quad (x_i, y_j) \text{ is a boundary point;}$$

$$\mathbf{B}[\mathcal{D}(\bar{1})]_{ij} = c+4, \quad (x_i, y_j) \text{ is a cell center;}$$

$$\mathbf{B}[\mathcal{D}(\bar{1})]_{ij} = \frac{3}{4}c + \frac{1}{8}, \quad (x_i, y_j) \text{ is an interior edge center;}$$

$$\mathbf{B}[\mathcal{D}(\bar{1})]_{ij} = c+1, \quad (x_i, y_j) \text{ is an interior knot.}$$

By Theorem 3.4, $A_d + A^z = B$ is a nonsingular M-matrix.

3.5.3. Verification of $A_a^+ \leq A^z A_d^{-1} A^s$ for edge centers. Next, we verify $A_a^+ \leq A^z A_d^{-1} A^s$. We first compare $\mathcal{A}^z[\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))]_{ij}$ with $\mathcal{A}_a^+(\bar{\phi})_{ij}$ for the case that (x_i, y_j) is an interior edge center for an edge parallel to x-axis.

$$\mathcal{A}^z[\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))]_{ij}$$

$$= -2\left(1 + \frac{hv_{ij}}{2\mu}\right)\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))_{i,j-1} - 2\left(1 - \frac{hv_{ij}}{2\mu}\right)\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))_{i,j+1}$$

$$= -2\left(1 + \frac{hv_{ij}}{2\mu}\right)(c+4)^{-1}\mathcal{A}^s(\bar{\phi})_{i,j-1} - 2\left(1 - \frac{hv_{ij}}{2\mu}\right)(c+4)^{-1}\mathcal{A}^s(\bar{\phi})_{i,j+1}$$

$$= 2\frac{1 + \frac{hv_{ij}}{2\mu}}{c+4}\left[\left(1 - \frac{hu_{i,j-1}}{2\mu}\right)\phi_{i+1,j-1} + \left(1 + \frac{hu_{i,j-1}}{2\mu}\right)\phi_{i-1,j-1} + \left(1 - \frac{hv_{i,j-1}}{2\mu}\right)\phi_{i,j} + \left(1 + \frac{hv_{i,j-1}}{2\mu}\right)\phi_{i,j-2}\right]$$

$$+2\frac{1-\frac{hv_{ij}}{2\mu}}{c+4}\left[\left(1-\frac{hu_{i,j+1}}{2\mu}\right)\phi_{i+1,j+1}+\left(1+\frac{hu_{i,j+1}}{2\mu}\right)\phi_{i-1,j+1}+\left(1-\frac{hv_{i,j+1}}{2\mu}\right)\phi_{i,j+2}+\left(1+\frac{hv_{i,j+1}}{2\mu}\right)\phi_{i,j}\right].$$

For $A_a^+ \leq A^z A_d^{-1} A^s$ to hold, we need

$$2\frac{1+\frac{hv_{ij}}{2\mu}}{c+4}\left(1+\frac{hv_{i,j-1}}{2\mu}\right)\geq\frac{1}{4}\left(1+\frac{hv_{ij}}{\mu}\right),\quad 2\frac{1-\frac{hv_{ij}}{2\mu}}{c+4}\left(1-\frac{hv_{i,j+1}}{2\mu}\right)\geq\frac{1}{4}\left(1-\frac{hv_{ij}}{\mu}\right),$$

which are equivalent to

$$2c\frac{hv_{ij}}{2\mu}-8\frac{hv_{i,j-1}}{2\mu}-8\frac{hv_{ij}}{2\mu}\frac{hv_{i,j-1}}{2\mu}\leq4-c,$$

$$-2c\frac{hv_{ij}}{2\mu}+8\frac{hv_{i,j+1}}{2\mu}-8\frac{hv_{ij}}{2\mu}\frac{hv_{i,j+1}}{2\mu}\leq4-c.$$

Let $a=\max_{i,j}|v_{ij}|\frac{h}{2\mu}>0$, then it suffices to have

$$2ca+8a+8a^2\leq4-c\Longleftrightarrow8a^2+(8+2c)a-(4-c)\geq0\Longleftrightarrow0<a\leq\frac{\sqrt{c^2+48}-(c+4)}{8}.$$

For $\frac{\sqrt{c^2+48}-(c+4)}{8}>0$ to hold, we need $c<4$.

For fixed $a>0$, we have

$$2ca+8a+8a^2\leq4-c\Longleftrightarrow c\leq\frac{4-8a^2-8a}{1+2a}=\frac{-8(a+\frac{1}{2})^2+6}{2a+1}.$$

For $\frac{-8(a+\frac{1}{2})^2+6}{2a+1}$ to be positive, we need $a<\frac{\sqrt{3}-1}{2}$.

For the case that (x_i,y_j) is an interior edge center for an edge parallel to y-axis, the discussion will be similar and the same mesh constraints apply due to the symmetry.

We summarize the constraints obtained so far as the following:

- For any $c\in(0,4)$, $0<a\leq\frac{\sqrt{c^2+48}-(c+4)}{8}$;
- For any $a\in(0,\frac{\sqrt{3}-1}{2})$, $0<c\leq c\leq\frac{4-8a^2-8a}{1+2a}$.

3.5.4. Verification of $A_a^+ \leq A^z A_d^{-1} A^s$ for knots.   Next, consider the case (x_i,y_j) being an interior knot.

$$\mathcal{A}^z[\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))]_{ij}$$

$$=-2\left(1+\frac{hv_{ij}}{2\mu}\right)\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))_{i,j-1}-2\left(1-\frac{hv_{ij}}{2\mu}\right)\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))_{i,j+1}$$

$$-2\left(1+\frac{hu_{ij}}{2\mu}\right)\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))_{i-1,j}-2\left(1-\frac{hu_{ij}}{2\mu}\right)\mathcal{A}_d^{-1}(\mathcal{A}^s(\bar{\phi}))_{i+1,j},$$

$$=-2\frac{1+\frac{hv_{ij}}{2\mu}}{c+\frac{11}{2}}\mathcal{A}^s(\bar{\phi})_{i,j-1}-2\frac{1-\frac{hv_{ij}}{2\mu}}{c+\frac{11}{2}}\mathcal{A}^s(\bar{\phi})_{i,j+1}-2\frac{1+\frac{hu_{ij}}{2\mu}}{c+\frac{11}{2}}\mathcal{A}^s(\bar{\phi})_{i-1,j}-2\frac{1-\frac{hu_{ij}}{2\mu}}{c+\frac{11}{2}}\mathcal{A}^s(\bar{\phi})_{i+1,j},$$

$$=2\frac{1+\frac{hv_{ij}}{2\mu}}{c+\frac{11}{2}}\left[\left(1-\frac{hv_{i,j-1}}{2\mu}\right)\phi_{i,j}+\left(1+\frac{hv_{i,j-1}}{2\mu}\right)\phi_{i,j-2}\right]$$

$$+2\frac{1-\frac{hv_{ij}}{2\mu}}{c+\frac{11}{2}}\left[\left(1-\frac{hv_{i,j+1}}{2\mu}\right)\phi_{i,j+2}+\left(1+\frac{hv_{i,j+1}}{2\mu}\right)\phi_{i,j}\right]$$

$$+2\frac{1+\frac{hu_{ij}}{2\mu}}{c+\frac{11}{2}}\left[\left(1-\frac{hu_{i-1,j}}{2\mu}\right)\phi_{i,j}+\left(1+\frac{hu_{i-1,j}}{2\mu}\right)\phi_{i-2,j}\right]$$

$$+2\frac{1-\frac{hu_{ij}}{2\mu}}{c+\frac{11}{2}}\left[\left(1-\frac{hu_{i+1,j}}{2\mu}\right)\phi_{i+2,j}+\left(1+\frac{hu_{i+1,j}}{2\mu}\right)\phi_{i,j}\right].$$

For $A_a^+ \le A^z A_d^{-1} A^s$ to hold, we need

$$2\frac{1+\frac{hv_{ij}}{2\mu}}{c+\frac{11}{2}}\left(1+\frac{hv_{i,j-1}}{2\mu}\right)\ge\frac{1}{4}\left(1+\frac{hv_{ij}}{\mu}\right),\quad 2\frac{1-\frac{hv_{ij}}{2\mu}}{c+\frac{11}{2}}\left(1-\frac{hv_{i,j+1}}{2\mu}\right)\ge\frac{1}{4}\left(1-\frac{hv_{ij}}{\mu}\right),$$

$$2\frac{1+\frac{hu_{ij}}{2\mu}}{c+\frac{11}{2}}\left(1+\frac{hu_{i-1,j}}{2\mu}\right)\ge\frac{1}{4}\left(1+\frac{hu_{ij}}{\mu}\right),\quad 2\frac{1-\frac{hu_{ij}}{2\mu}}{c+\frac{11}{2}}\left(1-\frac{hu_{i+1,j}}{2\mu}\right)\ge\frac{1}{4}\left(1-\frac{hu_{ij}}{\mu}\right),$$

which are equivalent to

$$(3+2c)\frac{hv_{ij}}{2\mu}-8\frac{hv_{i,j-1}}{2\mu}-8\frac{hv_{ij}}{2\mu}\frac{hv_{i,j-1}}{2\mu}\le\frac{5}{2}-c,$$

$$-(3+2c)\frac{hv_{ij}}{2\mu}+8\frac{hv_{i,j-1}}{2\mu}-8\frac{hv_{ij}}{2\mu}\frac{hv_{i,j-1}}{2\mu}\le\frac{5}{2}-c,$$

$$(3+2c)\frac{hu_{ij}}{2\mu}-8\frac{hu_{i-1,j}}{2\mu}-8\frac{hu_{ij}}{2\mu}\frac{hu_{i-1,j}}{2\mu}\le\frac{5}{2}-c,$$

$$-(3+2c)\frac{hu_{ij}}{2\mu}+8\frac{hu_{i-1,j}}{2\mu}-8\frac{hu_{ij}}{2\mu}\frac{hu_{i-1,j}}{2\mu}\le\frac{5}{2}-c.$$

Let $a=\max\{\max_{i,j}|v_{ij}|,\max_{i,j}|u_{ij}|\}\frac{h}{2\mu}>0$, then it suffices to require

$$(3+2c)a+8a+8a^2\le\frac{5}{2}-c\Longleftrightarrow 16a^2+(11+2c)2a-5+2c\le 0\Longleftarrow 0\le a\le\frac{\sqrt{(c+\frac{3}{2})^2+48}-(c+\frac{11}{2})}{8}.$$

To ensure $\frac{\sqrt{(c+\frac{3}{2})^2+48}-(c+\frac{11}{2})}{8}>0$, we need $c<\frac{5}{2}$, with which we have $\frac{\sqrt{(c+\frac{3}{2})^2+48}-(c+\frac{11}{2})}{8}<\frac{\sqrt{c^2+48}-(c+4)}{8}$. For a fixed $a>0$, we also have

$$(3+2c)a+8a+8a^2\le\frac{5}{2}-c\Longleftrightarrow 0<c\le\frac{-8a^2-11a+\frac{5}{2}}{2a+1}.$$

For $\frac{-8a^2-11a+\frac{5}{2}}{2a+1}>0$, we need $a<\frac{\sqrt{201}-11}{16}$ which is smaller than $\frac{\sqrt{3}-1}{2}$. When $\frac{-8a^2-11a+\frac{5}{2}}{2a+1}>0$, we also have $\frac{-8a^2-11a+\frac{5}{2}}{2a+1}<\frac{4-8a^2-8a}{1+2a}$.

Now we can summarize all mesh constraints for $A_a^+\le A^z A_d^{-1} A^s$ at both edge centers and knots:

- For any $c\in(0,\frac{5}{2})$, $0<a\le\frac{\sqrt{(c+\frac{3}{2})^2+48}-(c+\frac{11}{2})}{8}$;
- For any $a\in(0,\frac{\sqrt{201}-11}{16})$, $0<c\le\frac{-8a^2-11a+\frac{5}{2}}{2a+1}$.

3.5.5. Sufficient conditions in 2-D. Since $A^z A_d^{-1} A^s\ge 0$ and $\mathcal{A}_a^+(\phi)_{ij}$ are nonzero only at interior knots and interior edge centers, we have already found all constraints to ensure $A_a^+\le A^z A_d^{-1} A^s$. By applying Theorem 3.7, we get the monotonicity result for the high order scheme:

THEOREM 3.9. *Let* $\|\mathbf{u}\|_\infty=\max_{ij}\{|u_{ij}|,|v_{ij}|\}|$. *For the fourth order accurate scheme* (3.2) *to be inverse positive, i.e.,* $\bar{L}^{-1}\ge 0$, *the following conditions are sufficient:*

- *For a mesh size h satisfying $h\frac{\|\mathbf{u}\|_\infty}{2\mu}=a<\frac{\sqrt{201}-11}{16}\approx 0.199$, time step Δt satisfies $\Delta t\frac{\mu}{h^2}\geq \frac{2a+1}{-8a^2-11a+\frac{5}{2}}$.*

- *For a time step Δt satisfying $\Delta t\frac{\mu}{h^2}=\frac{1}{c}>\frac{2}{5}$, the mesh size h satisfies $h\frac{\|\mathbf{u}\|_\infty}{\mu}\leq \frac{\sqrt{(c+\frac{3}{2})^2+48}-(c+\frac{11}{2})}{8}$.*

In particular, the following are convenient explicit sufficient mesh constraints for the inverse positivity:

- *For a mesh size h satisfying $h\frac{\|\mathbf{u}\|_\infty}{\mu}\leq \frac{1}{3}$, time step satisfies $\Delta t\frac{\mu}{h^2}\geq 3$.*

- *For a time step Δt satisfying $\Delta t\frac{\mu}{h^2}\geq 1$, the mesh size h satisfies $h\frac{\|\mathbf{u}\|_\infty}{\mu}\leq \frac{\sqrt{217}-13}{8}\approx 0.216$.*

4. The generalized Allen-Cahn equation

We now consider (1.1). We shall assume that the free energy functional $F(\phi)$ has a double well form with minima at $\pm\beta$, where β satisfies

$$F'(\beta)=F'(-\beta)=0, \tag{4.1a}$$

$F'(\beta)$ satisfies the monotone conditions away from $(-\beta,\beta)$:

$$F'(\phi)<0,\forall\phi<-\beta;\quad F'(\beta)>0,\forall\phi>\beta. \tag{4.1b}$$

Such energy functionals include the polynomial energy $F(\phi)=\frac{1}{4}(\phi^2-1)^2$ with $\beta=1$, and the logarithmic energy (1.2) with β given by $\frac{1}{2\beta}\ln\frac{1+\beta}{1-\beta}=\frac{\theta_c}{\theta}$, see [19].

By Lemma 2.1 in [19], we have

LEMMA 4.1. *For an energy function satisfying (4.1) and $f(x)=x-\frac{\Delta t}{\varepsilon}F'(x)$, the following bounds of $f(x)$ hold:*

$$f(x)\in[-\beta,\beta],\quad \forall x\in[-\beta,\beta],$$

under the time step constraint

$$\Delta t\max_{x\in[-\beta,\beta]}F''(x)\leq\varepsilon.$$

Consider a first order implicit explicit time (IMEX) discretization with the fourth order difference scheme:

$$\phi^{n+1}+\Delta t\left[u^{n+1}.*(R_y\bar\phi^{n+1}D_{1x}^T)+v^{n+1}.*(D_{1y}\bar\phi^{n+1}R_x^T)-\mu(R_y\bar\phi^{n+1}D_{2x}^T+D_{2y}\bar\phi^{n+1}R_x^T)\right]$$
$$=\phi^n-\frac{\Delta t}{\varepsilon}F'(\phi^n), \tag{4.2}$$

where $F'(\phi^n)$ is a matrix with entries $F'(\phi_{ij}^n)$.

Notice that the time step has a lower bound $\Delta t>\frac{2}{5}\frac{h^2}{\mu}$ in Theorem 3.9 and an upper bound $\Delta t\leq\frac{\varepsilon}{\max\limits_{x\in[-\beta,\beta]}F''(x)}$ in Lemma 4.1, thus we need an upper bound on the mesh size $h<\sqrt{\frac{5}{2}\frac{\mu\varepsilon}{\max\limits_{x\in[-\beta,\beta]}F''(x)}}$ so that the time step interval is not empty. Combined with Theorem 3.9, we have:

THEOREM 4.1. *The scheme (4.2) for homogeneous Dirichlet boundary condition satisfies the discrete maximum principle:*

$$\min_{i,j}\phi^n\leq\phi_{ij}^{n+1}\leq\max_{i,j}\phi^n,$$

under the following convenient mesh and time step constraints:

(1)

$$h \leq \min\left\{\frac{1}{3}\frac{\mu}{\|\mathbf{u}\|_\infty}, \sqrt{\frac{3\mu\varepsilon}{\max\limits_{x\in[-\beta,\beta]} F''(x)}}\right\}, \quad 3\frac{h^2}{\mu} \leq \Delta t \leq \frac{\varepsilon}{\max\limits_{x\in[-\beta,\beta]} F''(x)}.$$

(2)

$$h \leq \min\left\{0.216\frac{\mu}{\|\mathbf{u}\|_\infty}, \sqrt{\frac{\mu\varepsilon}{\max\limits_{x\in[-\beta,\beta]} F''(x)}}\right\}, \quad \frac{h^2}{\mu} \leq \Delta t \leq \frac{\varepsilon}{\max\limits_{x\in[-\beta,\beta]} F''(x)}.$$

REMARK 4.1. As a comparison, for the second order scheme with IMEX time discretization to satisfy the discrete maximum, there is no lower bound on the time step. By Theorem 3.5, we have the mesh constraints needed for second order scheme to be bound-preserving:

$$h \leq \min 2\frac{\mu}{\|\mathbf{u}\|_\infty}, \quad \Delta t \leq \frac{\varepsilon}{\max\limits_{x\in[-\beta,\beta]} F''(x)}.$$

Next, we consider the stabilized IMEX time discretization with the fourth order difference scheme with $S \geq 0$:

$$\phi^{n+1} + \Delta t\left[S\phi^{n+1} + u^{n+1}.*(R_y\bar{\phi}^{n+1}D_{1x}^T) + v^{n+1}.*(D_{1y}\bar{\phi}^{n+1}R_x^T) - \mu(R_y\bar{\phi}^{n+1}D_{2x}^T + D_{2y}\bar{\phi}^{n+1}R_x^T)\right]$$
$$= \phi^n - \frac{\Delta t}{\varepsilon}F'(\phi^n) + S\Delta t\phi^n. \tag{4.3}$$

Notice that the stabilized IMEX time discretization (1.4) can be written as

$$\frac{\phi^{n+1} - \phi^n}{\widetilde{\Delta t}} + u^{n+1}\phi_x^{n+1} + v^{n+1}\phi_y^{n+1} = \mu\Delta\phi^{n+1} - \frac{F'(\phi^n)}{\varepsilon}$$

with $\widetilde{\Delta t} = \frac{\Delta t}{1+\Delta tS}$. Replacing Δt by $\widetilde{\Delta t}$ in Theorem 4.1, we can easily get the result for the stabilized IMEX time discretization:

THEOREM 4.2. *The scheme* (4.3) *with* $S \geq 0$ *for homogeneous Dirichlet boundary condition satisfies the discrete maximum principle:*

$$\min_{i,j}\phi^n \leq \phi_{ij}^{n+1} \leq \max_{i,j}\phi^n,$$

under the following convenient mesh and time step constraints:

(1)

$$h \leq \min\left\{\frac{1}{3}\frac{\mu}{\|\mathbf{u}\|_\infty}, \sqrt{\frac{3\mu\varepsilon}{\max\limits_{x\in[-\beta,\beta]} F''(x)}}\right\}, 3\frac{h^2}{\mu} \leq \frac{\Delta t}{1+\Delta tS} \leq \frac{\varepsilon}{\max\limits_{x\in[-\beta,\beta]} F''(x)};$$

(2)

$$h \leq \min\left\{0.216\frac{\mu}{\|\mathbf{u}\|_\infty}, \sqrt{\frac{\mu\varepsilon}{\max\limits_{x\in[-\beta,\beta]} F''(x)}}\right\}, \frac{h^2}{\mu} \leq \frac{\Delta t}{1+\Delta tS} \leq \frac{\varepsilon}{\max\limits_{x\in[-\beta,\beta]} F''(x)}.$$

REMARK 4.2.   As a comparison, for the spatially second order scheme with the stabilized IMEX time discretization to satisfy the discrete maximum, there is no lower bound on the time step, and by Theorem 3.5, the mesh constraints for this scheme to be bound-preserving are:

$$h \leq \min 2\frac{\mu}{\|\mathbf{u}\|_\infty}, \quad \frac{1}{\Delta t} + S \geq \frac{\max\limits_{x\in[-\beta,\beta]} F''(x)}{\varepsilon}.$$

5. Stream function vorticity formulation of 2D incompressible flow

The results in Section 3 can also be used to construct a bound-preserving scheme for any passive convection-diffusion with an incompressible velocity field. As an example, we consider the two-dimensional incompressible Navier-Stokes equation in stream function vorticity form:

$$\omega_t + u\omega_x + v\omega_y = \mu\Delta\omega$$
$$\Delta\psi = \omega, \quad (u,v) = (-\psi_y, \psi_x)$$
$$\omega(x,y,0) = \omega_0(x,y), \quad (x,y) \in \Omega,$$

where ω is the vorticity, and ψ is the stream function. For simplicity, we only consider homogeneous Dirichlet boundary conditions, and extensions to periodic boundary conditions are straightforward. We consider a first order time discretization:

$$\Delta\psi^{n+1} = \omega^n, \quad (u^{n+1}, v^{n+1}) = (-\psi_y^{n+1}, \psi_x^{n+1}),$$
$$\frac{\omega^{n+1} - \omega^n}{\Delta t} + u^{n+1}\omega_x^{n+1} + v^{n+1}\omega_y^{n+1} = \mu\Delta\omega^{n+1},$$

with second order or fourth order finite difference spatial discretization as described in Section 2.

When using the fourth order scheme for ω, the same fourth order scheme can also be used to solve the Poisson equation $\Delta\psi = \omega$. See [13] for an efficient inversion of the discrete Laplacian via an eigenvector method. Once ψ is obtained from the Poisson equation, the velocity field can be computed by taking finite difference of ψ. However, for a fourth order scheme, the difference matrix D_1 cannot be used because it is only a second order finite difference approximating first order derivatives. Instead, a conventional fourth order finite difference operator should be used to compute the numerical differentiation.

Notice that the scheme for ω here is the same as (3.2). With Theorem 3.2, the fully discrete scheme with homogeneous Dirichlet boundary condition satisfies the discrete maximum principle

$$\min_{i,j}\omega^n \leq \omega_{ij}^{n+1} \leq \max_{i,j}\omega^n,$$

if the mesh size and time step constraints in Theorem 3.5, Theorem 3.8 and Theorem 3.9 are satisfied.

6. Numerical tests

For implementation of the scheme, biconjugate gradient stabilized (BiCGSTAB) method is used for solving the linear system with variable coefficients at each time, with the discrete Laplacian as a preconditioner which can be efficiently inverted, see [13] for details.

6.1. Accuracy test. We first test accuracy for the generalized Allen-Cahn equation with $F(\phi) = \frac{1}{4}(\phi^2 - 1)^2$, parameters $\mu = 0.1$ and $\varepsilon = 0.05$ and a given velocity field

$$u = v = \sin(y - x).$$

A source term is added so that the exact solution is

$$\phi = (0.75 + 0.25\sin(t))\sin y \sin^2 x.$$

Homogeneous Dirichlet boundary conditions are used on the domain $[0, 2\pi] \times [0, 2\pi]$. In order to test the designed spatial accuracy, a third order accurate IMEX backward differentiation formula (BDF) time discretization is used: the nonlinear term is treated explicitly in time, and the convection diffusion terms are treated implicitly. Errors at $T = 0.2$ are listed in Table 6.1, in which we can observe the expected spatial order of accuracy.

Finite Difference Grid	second order scheme				fourth order scheme			
	l^1 error	order	l^∞ error	order	l^1 error	order	l^∞ error	order
9×9	6.58E-2	-	2.38E-1	-	6.63E-2	-	2.66E-1	-
19×19	1.75E-2	1.91	8.80E-2	1.61	1.36E-2	2.28	5.23E-2	2.35
79×79	1.04E-3	2.02	4.75E-3	2.00	1.92E-5	4.85	1.21E-4	4.22
159×159	2.56E-4	2.02	1.19E-3	2.00	1.13E-6	4.09	7.15E-6	4.08

TABLE 6.1. *Accuracy test on uniform meshes for an Allen-Cahn equation. Third order IMEX BDF time discretization is used for time discretization.*

Next, we test accuracy of two schemes solving the stream function vorticity equations of incompressible flow on the domain $[0, 2\pi] \times [0, 2\pi]$ with periodic boundary conditions. An exact solution $\omega = -2e^{-2\mu t}\sin x \sin y$ with $\mu = 0.1$ is considered. In order to test the designed spatial accuracy, a third order accurate BDF time discretization is used. Errors at $T = 0.2$ are listed in Table 6.2, in which we can observe the expected spatial order of accuracy.

Finite Difference Grid	second order scheme				fourth order scheme			
	l^1 error	order	l^∞ error	order	l^1 error	order	l^∞ error	order
40×40	4.82E-5	-	1.14E-4	-	5.69E-5	-	2.30E-4	-
80×80	1.34E-5	1.84	3.23E-5	1.81	3.67E-6	3.96	1.51E-5	3.93
160×160	3.78E-6	1.83	9.21E-6	1.81	2.27E-7	4.01	9.47E-7	4.00
320×320	9.64E-7	1.97	2.36E-6	1.96	1.41E-8	4.00	5.91E-8	4.00

TABLE 6.2. *Accuracy test on uniform meshes for stream function vorticity equations with periodic boundary conditions. Third order BDF time discretization is used for time discretization.*

6.2. The generalized Allen-Cahn equation. Next, we take a given velocity field $u = v = \sin(y - x)$ in (1.1) with a logarithmic energy function

$$F(\phi) = \frac{\theta}{2}[(1 + \phi)\ln(1 + \phi) + (1 - \phi)\ln(1 - \phi)] - \frac{\theta_c}{2}\phi^2,$$

and parameters, $\theta = 1$, $\theta_c = 0.5$, $\mu = 0.01$ and $\varepsilon = 0.03$. The initial condition is $\phi_0(x, y) = 0.99\sin y \sin^2 x$. The stability parameter $S = 0$ is used, i.e., the time discretization is first
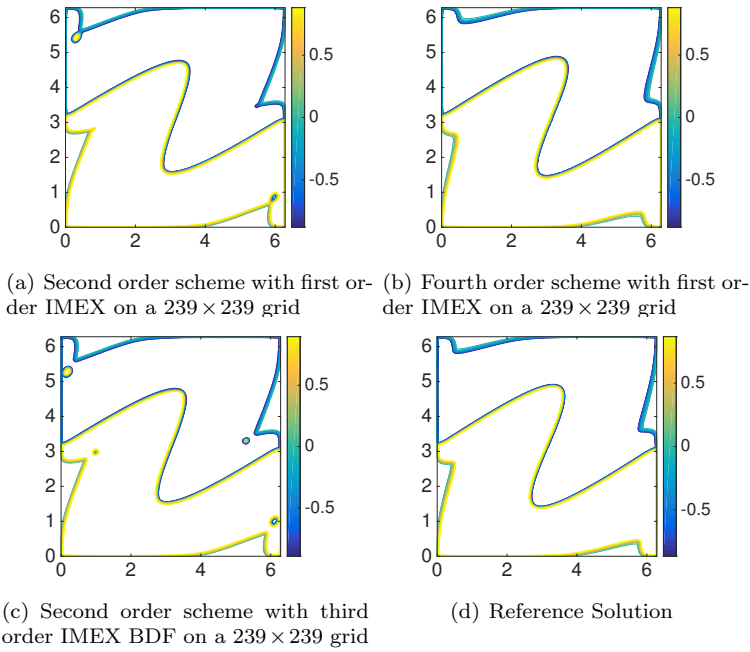
(a) Second order scheme with first or- (b) Fourth order scheme with first or-
der IMEX on a 239×239 grid   der IMEX on a 239×239 grid

(c) Second order scheme with third   (d) Reference Solution
order IMEX BDF on a 239×239 grid

FIG. 6.1. *Allen-Cahn with log energy at* $T = 1.8$. *The reference solution is generated by second order scheme with third order IMEX BDF time discretization on a* 479×479 *grid.*
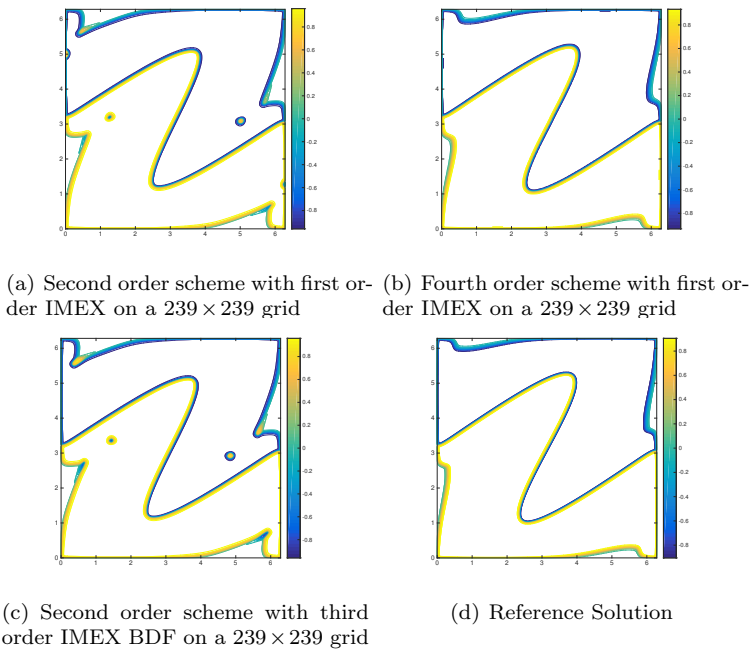

(a) Second order scheme with first or- (b) Fourth order scheme with first or-
der IMEX on a 239×239 grid   der IMEX on a 239×239 grid

(c) Second order scheme with third   (d) Reference Solution
order IMEX BDF on a 239×239 grid

FIG. 6.2. *Allen-Cahn with polynomial energy at* $T = 2.2$. *The reference solution is generated by second order scheme with third order IMEX BDF time discretization on a* 479×479 *grid.*

order IMEX method. See Figure 6.1 for performance of the schemes. We observe that the second order scheme with first order IMEX time discretization produces erroneous numerical artifacts on a relatively coarse 239×239 grid, and higher order time discretization does not help reducing such an error. On the other hand, the fourth order scheme with first order IMEX method produces a satisfying solution on the 239×239 grid. For both second order and fourth order schemes, the time step is taken as $\Delta t = \frac{1}{7}\Delta x$, and iterations needed for convergence in BiCGSTAB are almost the same for two schemes in each time step, thus the computational cost of both second order and fourth order schemes is almost the same on the same grid. Therefore, the fourth order scheme is obviously superior.
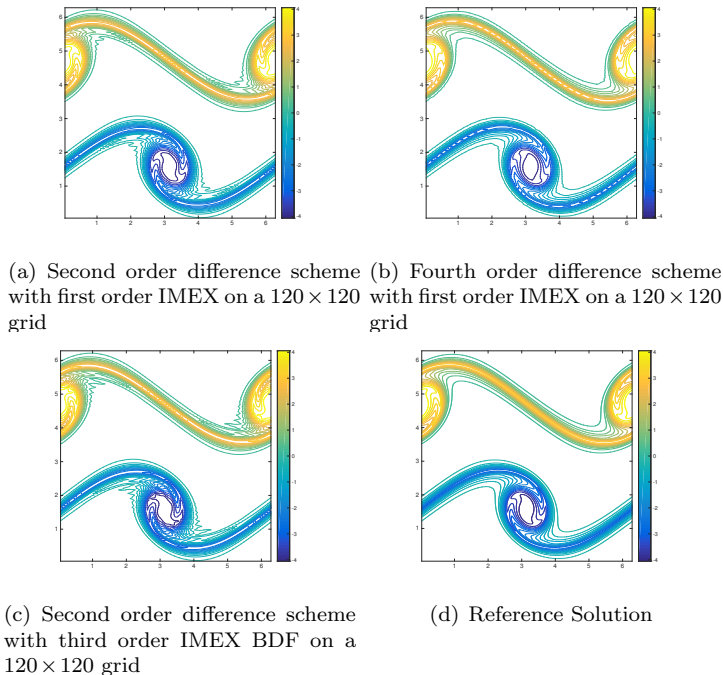
Next, we test a given velocity field $u = v = \sin(y - x)$ for (1.1) with a polynomial energy function $F(\phi) = \frac{1}{4}(\phi^2 - 1)^2$, parameters $\mu = 0.01$ and $\varepsilon = 0.05$. The initial condition is $\phi_0(x, y) = 0.75 \sin y \sin^2 x$. The stability parameter $S = 0$ is used, i.e., the time discretization is first order IMEX method. See Figure 6.2 for performance of the schemes. We observe that the second order scheme with first order IMEX time discretization produces erroneous numerical artifacts on a relatively coarse 239×239 grid, and higher order time discretization does not help reducing such an error. On the other hand, the fourth order scheme with first order IMEX method produces a satisfying solution on the 239×239 grid. For both second order and fourth order schemes, the time step is taken as $\Delta t = \frac{1}{6}\Delta x$, and iterations needed for convergence in BiCGSTAB are almost the same for two schemes in each time step, thus the computational cost of both second order and fourth order schemes is almost the same on the same grid. Therefore, the fourth order scheme is obviously superior.

(a) Second order difference scheme with first order IMEX on a 120×120 grid

(b) Fourth order difference scheme with first order IMEX on a 120×120 grid

(c) Second order difference scheme with third order IMEX BDF on a 120×120 grid

(d) Reference Solution

FIG. 6.3. *The 2D incompressible Navier-Stokes in vorticity form at $T = 6$ with $\mu = 0.001$. The reference solution is generated by second order difference scheme with third order IMEX BDF time discretization on a 240×240 grid.*
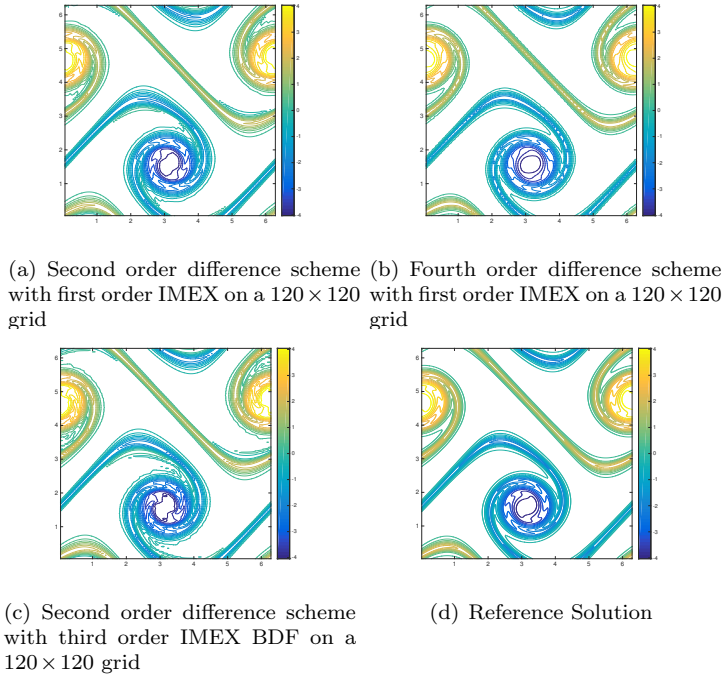
(a) Second order difference scheme with first order IMEX on a 120×120 grid

(b) Fourth order difference scheme with first order IMEX on a 120×120 grid

(c) Second order difference scheme with third order IMEX BDF on a 120×120 grid

(d) Reference Solution

FIG. 6.4. *The 2D incompressible Navier-Stokes in vorticity form at $T=8$ with $\mu=0.001$. The reference solution is generated by second order difference scheme with third order IMEX BDF time discretization on a 240×240 grid.*

6.3. Incompressible flow: double shear layer. We test the same schemes for solving the incompressible Navier-Stokes system with $\mu=0.001$ consisting of a scalar convection-diffusion for vorticity and a Poisson equation for stream function, as described in Section 5. We consider the following initial condition with the periodic boundary conditions on $[0, 2\pi] \times [0, 2\pi]$:

$$\omega(x, y, 0) = \begin{cases} \delta \cos x - \frac{1}{\rho} sech^2 \frac{y - \frac{\pi}{2}}{\rho}, & y \leq \pi \\ \delta \cos x + \frac{1}{\rho} sech^2 \frac{\frac{3\pi}{2} - y}{\rho}, & y > \pi \end{cases}$$

with $\rho = \frac{\pi}{15}$ and $\delta = 0.05$. This is a classical test for 2D incompressible Navier Stokes in vorticity form. See Figure 6.3 and Figure 6.4 for the performance of the schemes. For both schemes, the time step is taken as $\Delta t = \frac{1}{6 \|\mathbf{u}\|_\infty} \Delta x$ and iterations needed for convergence in BiCGSTAB are almost the same, thus the computational cost is almost the same. Similar to observations for the Allen-Cahn equation, we can see that the second order scheme with first order IMEX time discretization produces erroneous numerical oscillations on a relatively coarse 120×120 grid, and higher order time discretization does not help reducing such an error. The fourth order scheme produces much better solutions on the 120×120 grid.

7. Concluding remarks

In this paper we have proven the monotonicity of the finite difference implementation of the Q^2 spectral element method for a linear convection-diffusion operator with a given incompressible velocity field. Thanks to the monotonicity, we obtained a fourth order accurate finite difference spatial discretization satisfying the discrete max-

imum principle, and used it to construct bound-preserving schemes for the generalized Allen-Cahn equation. To the best of our knowledge, this is the first time that a high order spatial discretization with an IMEX discretization in time is proven to satisfy a discrete maximum principle for a linear convection-diffusion operator. We presented several numerical tests which showed superiority of higher order spatial accuracy compared to the most popular bound-preserving second order scheme. Even though we only discussed the monotonicity for the fourth order finite difference scheme solving the two-dimensional problem, it is straightforward to extend the discussion of monotonicity to a three-dimensional linear convection-diffusion problem with an incompressible velocity field.

REFERENCES

[1] S.M. Allen and J.W. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metall., 27(6):1085–1095, 1979. 1

[2] L.-Q. Chen, *Phase-field models for microstructure evolution*, Annu. Rev. Mater. Res., 32(1):113–140, 2002. 1

[3] Z. Chen, H. Huang, and J. Yan, *Third order maximum-principle-satisfying direct discontinuous Galerkin methods for time dependent convection diffusion equations on unstructured triangular meshes*, J. Comput. Phys., 308:198–217, 2016. 1

[4] L.J. Cross and X. Zhang, *On the monotonicity of high order discrete Laplacian*, arXiv preprint, arXiv:2010.07282, 2020. 3.3

[5] Q. Du, L. Ju, X. Li, and Z. Qiao, *Maximum principle preserving exponential time differencing schemes for the nonlocal Allen–Cahn equation*, SIAM J. Numer. Anal., 57(2):875–898, 2019. 1

[6] Q. Du, L. Ju, X. Li, and Z. Qiao, *Maximum bound principles for a class of semilinear parabolic equations and exponential time-differencing schemes*, SIAM Rev., 63(2):317–359, 2021. 1

[7] S. Gottlieb, D.I. Ketcheson, and C.-W. Shu, *High order strong stability preserving time discretizations*, J. Sci. Comput., 38(3):251–289, 2009. 1

[8] L. Guo, X. Li, and Y. Yang, *Energy dissipative local discontinuous Galerkin methods for Keller-Segel chemotaxis model*, J. Sci. Comput., 78:1387–1404, 2019. 1

[9] L. Ju, X. Li, Z. Qiao, and J. Yang, *Maximum bound principle preserving integrating factor Runge–Kutta methods for semilinear parabolic equations*, J. Comput. Phys., 439:110405, 2021. 1

[10] H. Li, D. Appelö, and X. Zhang, *Accuracy of spectral element method for wave, parabolic and Schrödinger equations*, SIAM J. Numer. Anal., 60(1):339–363, 2021. 1, 2.1, 2.1

[11] H. Li, S. Xie, and X. Zhang, *A high order accurate bound-preserving compact finite difference scheme for scalar convection diffusion equations*, SIAM J. Numer. Anal., 56(6):3308–3345, 2018. 1

[12] H. Li and X. Zhang, *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, 2020. 1, 3.3

[13] H. Li and X. Zhang, *Superconvergence of high order finite difference schemes based on variational formulation for elliptic equations*, J. Sci. Comput., 82(2):36, 2020. 1, 2, 2.1, 2.1, 5, 6

[14] C. Liu and J. Shen, *A phase field model for the mixture of two incompressible fluids and its approximation by a Fourier-spectral method*, Phys. D, 179(3-4):211–228, 2003. 1

[15] J. Lorenz, *Zur inversmonotonie diskreter probleme*, Numer. Math., 27(2):227–238, 1977. 3.3, 3.3

[16] Y. Maday and E.M. Rønquist, *Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries*, Comput. Methods Appl. Mech. Eng., 80(1-3):91–115, 1990. 1

[17] R.J. Plemmons, *M-matrix characterizations. I—nonsingular M-matrices*, Linear Algebra Appl., 18(2):175–188, 1977. 3.2, 3.2

[18] C. Qiu, Q. Liu, and J. Yan, *Third order positivity-preserving direct discontinuous Galerkin method with interface correction for chemotaxis Keller-Segel equations*, J. Comput. Phys., 433:110191, 2021. 1

[19] J. Shen, T. Tang, and J. Yang, *On the maximum principle preserving schemes for the generalized Allen–Cahn equation*, Commun. Math. Sci., 14(6):1517–1534, 2016. 1, 1, 4

[20] S. Srinivasan, J. Poggie, and X. Zhang, *A positivity-preserving high order discontinuous Galerkin scheme for convection–diffusion equations*, J. Comput. Phys., 366:120–143, 2018. 1

[21] Z. Sun, J.A. Carrillo, and C.-W. Shu, *A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials*, J. Comput. Phys., 352:76–

104, 2018. 1

[22] X. Yang, J.J. Feng, C. Liu, and J. Shen, *Numerical simulations of jet pinching-off and drop formation using an energetic variational phase-field method*, J. Comput. Phys., 218(1):417–428, 2006. 1

[23] X. Zhang, Y. Liu, and C.-W. Shu, *Maximum-principle-satisfying high order finite volume weighted essentially nonoscillatory schemes for convection-diffusion equations*, SIAM J. Sci. Comput., 34(2):A627–A658, 2012. 1