Check for updates

Structure Preserving Schemes for Fokker–Planck Equations of Irreversible Processes

Chen Liu[1] · Yuan Gao[1] · Xiangxiong Zhang[1]

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

In this paper, we construct structure preserving schemes for solving Fokker–Planck equations associated with irreversible processes. The proposed method is first order in time. We consider two structure-preserving spatial discretizations, which are second order and fourth order accurate finite difference schemes. They are derived via finite difference implementation of the classical Q^k ($k = 1, 2$) finite element methods on uniform meshes. Under mild mesh conditions and practical time step constraints, the schemes are proved monotone, thus are positivity-preserving and energy dissipative. In particular, our scheme is suitable for capturing steady state solutions in large final time simulations.

1 Introduction

Irreversible drift-diffusion processes are a class of important stochastic processes in physics and chemistry. For instance, an irreversible drift-diffusion process can model non-equilibrium biochemical reactions, which possess non-equilibrium steady states (NESS). The most important features for non-equilibrium reactions are nonzero fluxes and positive entropy production rate at NESS. These features in an irreversible biochemical reaction maintain a circulation at

✉ Yuan Gao
  gao662@purdue.edu

  Chen Liu
  liu3373@purdue.edu

  Xiangxiong Zhang
  zhan1966@purdue.edu

[1] Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA

NESS and we refer to pioneering studies by PRIGOGINE [31]. Let $\vec{b} : \mathbb{R}^d \to \mathbb{R}^d$ be a general drift field depending only on the state variable $\vec{x} \in \mathbb{R}^d$. Consider a stationary drift-diffusion process with white noise B_t that satisfies a stochastic differential equation (SDE) for $\vec{x}_t \in \mathbb{R}^d$

$$\mathrm{d}\vec{x}_t = \vec{b}(\vec{x}_t)\,\mathrm{d}t + \sqrt{2}\sigma\,\mathrm{d}B_t. \tag{1.1}$$

In general, σ is a noise matrix and $D := \sigma\sigma^T \in \mathbb{R}^{d\times d}$. For simplicity, in this paper we only discuss the simple case where $D > 0$ is a constant number. By Ito's formula, the corresponding Fokker–Planck equation for SDE (1.1), i.e., the Kolmogorov forward equation for density $\rho(\vec{x}, t)$, is given by

$$\partial_t \rho = \mathcal{L}^* \rho := -\nabla \cdot (\vec{b}\rho) + \nabla \cdot (D\nabla\rho). \tag{1.2}$$

In terms of (1.2), the irreversibility means that one cannot find an invariant measure π such that the generator $\mathcal{L}$ is symmetric in $L^2(\pi)$. Irreversibility has many equivalent characterizations. One is equivalent to that it is impossible to write the drift $\vec{b}$ in a potential form $\vec{b} = -D\nabla\varphi$ for any potential φ. See another equivalent irreversibility condition (1.7). Irreversible processes and the associated Fokker–Planck equations can be used to describe more general dynamic processes, such as the biochemical reactions with non-equilibrium steady state [16, 18, 32], they can also be used as sampling acceleration and variance reduction [10, 36].

In this paper, we focus on studying second and fourth order in space numerical schemes for (1.2) with a general drift field $\vec{b}$. We will prove the proposed high-order schemes preserve (*i*) the conservation of total mass, (*ii*) the positivity of ρ, (*iii*) the energy dissipation law with respect to ϕ-entropy, and (*iv*) the exponential convergence to equilibrium $\mathcal{M}$. To be more precise, we consider the equation (1.2) in a bounded domain $\Omega \subset \mathbb{R}^d$ with no-flux boundary condition

$$-\rho\vec{b} \cdot \vec{n} + D\nabla\rho \cdot \vec{n} = 0 \quad \text{on } \Gamma, \tag{1.3}$$

where $\vec{n}$ is the unit outer normal of the boundary $\Gamma = \partial\Omega$. Let $\rho^0(\vec{x})$ be the initial condition to (1.2). Under the no-flux boundary condition, it is easy to verify the conservation of total mass

$$\int_\Omega \rho\,\mathrm{d}x = \int_\Omega \rho^0\,\mathrm{d}x. \tag{1.4}$$

Designing structure preserving high order numerical schemes for the irreversible Fokker–Planck equation (1.2) is very important, not only because the irreversible processes are able to describe lots of fundamental non-equilibrium behaviors, such as circulations at NESS in an ecosystem, but also because of a general drift field is commonly used to construct acceleration or control for a given stochastic process or for a process constructed from discretization of irreversible Fokker–Planck equations; see for instance the optimally controlled transition path computations [11, 13, 17, 24, 37] and the accelerated sampling and optimization [10, 14, 36, 40].

1.1 Invariant Measure $\mathcal{M}$ and $\mathcal{M}$-Symmetric Decomposition

Assume that there exists a positive invariant measure $\mathcal{M} \in \mathcal{C}^1(\bar{\Omega})$, $\int \mathcal{M}\,\mathrm{d}x = 1$ and $\mathcal{M} \geq \epsilon_0 > 0$. $\mathcal{M}$ satisfies the static equation

$$\nabla \cdot F^s := \nabla \cdot \left(\vec{b}\mathcal{M} - D\nabla\mathcal{M}\right) = 0, \tag{1.5}$$

and the same no-flux boundary condition

$$- \mathcal{M}\vec{b} \cdot \vec{n} + D\nabla\mathcal{M} \cdot \vec{n} = 0 \quad \text{on } \Gamma. \tag{1.6}$$

For our compact domain $\bar{\Omega}$, such an invariant measure always exists, cf. [5]. The irreversibility is then equivalently characterized as that the steady flux is not pointwisely zero

$$F^s = \vec{b}\mathcal{M} - D\nabla\mathcal{M} \neq 0. \tag{1.7}$$

The special invariant measure such that $F^s = 0$ is also called a detailed balanced invariant measure.

Leveraging the existence of the positive invariant measure $\mathcal{M}$, one can utilize $\mathcal{M}$ to decompose the irreversible Fokker–Planck equation (1.2) into a dissipative part and a conservative part. The dissipative part is a gradient flow, which represents the decay from any initial density to the invariant measure. Meanwhile, the conservative part preserves the total energy and maintains a nonzero equilibrium flux [18, 32]. To this end, let us derive the $\mathcal{M}$-symmetric decomposition and the associated energy dissipation relation.

With a positive invariant measure $\mathcal{M}$, we decompose (1.2) into the sum of a gradient flow part and a Hamiltonian flow part

$$\begin{aligned}
\partial_t \rho &= \nabla \cdot \left(D\nabla\rho - \vec{b}\rho \right) = \nabla \cdot \left(D\mathcal{M}\nabla\frac{\rho}{\mathcal{M}} + \frac{\rho}{\mathcal{M}} \left(D\nabla\mathcal{M} - \mathcal{M}\vec{b} \right) \right) \\
&= \nabla \cdot \left(D\mathcal{M}\nabla\frac{\rho}{\mathcal{M}} \right) + \left(D\nabla\mathcal{M} - \mathcal{M}\vec{b} \right) \cdot \nabla\frac{\rho}{\mathcal{M}}.
\end{aligned} \tag{1.8}$$

From (1.5), we know

$$\vec{u} := D\nabla\mathcal{M} - \mathcal{M}\vec{b}, \quad \nabla \cdot \vec{u} = 0 \quad \text{in } \Omega. \tag{1.9}$$

The reversibility condition for the drift-diffusion process becomes $\vec{u} \equiv 0$, but we focus on more general case that $\vec{u} \neq 0$. Using the notation $\vec{u} = D\nabla\mathcal{M} - \mathcal{M}\vec{b}$ and (1.6), we know

$$\vec{u} \cdot \vec{n} = 0 \quad \text{on } \Gamma. \tag{1.10}$$

By the exactly same decomposition in (1.8), the no-flux boundary condition (1.3) becomes

$$\left(D\mathcal{M}\nabla\frac{\rho}{\mathcal{M}} + \frac{\rho}{\mathcal{M}} \left(D\nabla\mathcal{M} - \mathcal{M}\vec{b} \right) \right) \cdot \vec{n} = 0 \quad \text{on } \Gamma. \tag{1.11}$$

This, together with (1.10), implies $D\mathcal{M}\nabla\frac{\rho}{\mathcal{M}} \cdot \vec{n} = 0$ on Γ. Thus, we conclude that

$$D\mathcal{M}\nabla\frac{\rho}{\mathcal{M}} \cdot \vec{n} = 0, \quad \vec{u} \cdot \vec{n} = 0 \quad \text{on } \Gamma. \tag{1.12}$$

For certain applications, in the case of the invariant measure $\mathcal{M} > 0$ is given and satisfies (1.5) and (1.6), we can utilize it to construct a numerical scheme for the following equation, which stems from above decomposition, in conservative form:

$$\begin{aligned}
\partial_t \rho &= \nabla \cdot \left(D\mathcal{M}\nabla\frac{\rho}{\mathcal{M}} \right) + \nabla \cdot \left(\vec{u}\frac{\rho}{\mathcal{M}} \right) \quad \text{in } \{t > 0\} \times \Omega, \\
\rho &= \rho^0 \quad \text{in } \{t = 0\} \times \Omega, \\
D\mathcal{M}\left(\nabla\frac{\rho}{\mathcal{M}} \right) \cdot \vec{n} &= 0, \quad \vec{u} \cdot \vec{n} = 0 \quad \text{on } \{t > 0\} \times \partial\Omega.
\end{aligned} \tag{1.13}$$

Let us refer to (1.13) as **Model 1**. In this model, we highlight the positive invariant measure $\mathcal{M}$ is prescribed and the vector field $\vec{u}$ is a given time-independent continuously differentiable function, which satisfies $\nabla \cdot \vec{u} = 0$. In general, this $\vec{u}$ could be prescribed directly, or computed from the original drift $\vec{b}$ in (1.2).

Let us show the energy dissipation relation of **Model 1.** Define operators L^* and T as follows

$$L^* := \nabla \cdot (D\mathcal{M}\nabla), \quad T := (D\nabla\mathcal{M} - \mathcal{M}\vec{b}) \cdot \nabla = \vec{u} \cdot \nabla. \tag{1.14}$$

Using the boundary condition in (1.13), it is easy to check that L^* is a symmetric and nonnegative operator in $L^2(\Omega)$, namely

$$(f, L^*g) = (L^*f, g), \quad (f, L^*f) \le 0, \quad \forall f, g \in L^2(\Omega) \text{ satisfying } D\mathcal{M}\nabla f \cdot \vec{n}|_\Gamma = 0. \tag{1.15}$$

Here, the notation $(\cdot, \cdot)$ denotes the standard L^2-inner product. From the properties of $\vec{u}$, it is straightforward to see the operator T is asymmetric in $L^2(\Omega)$, namely

$$(f, Tg) = -(Tf, g), \quad (f, Tf) = 0, \quad \forall f, g \in L^2(\Omega). \tag{1.16}$$

Therefore, define free energy $E := \int \phi\left(\frac{\rho}{\mathcal{M}}\right)\mathcal{M}\,\mathrm{d}x$ for any convex function ϕ, we have the following energy dissipation relation.

$$
\begin{aligned}
\frac{\mathrm{d}E}{\mathrm{d}t} &= (\phi'\left(\frac{\rho}{\mathcal{M}}\right), \partial_t\rho) = (\phi'\left(\frac{\rho}{\mathcal{M}}\right), (L^* + T)\frac{\rho}{\mathcal{M}}) \\
&= -\int \phi''\left(\frac{\rho}{\mathcal{M}}\right)\mathcal{M}\nabla\frac{\rho}{\mathcal{M}} \cdot D\nabla\frac{\rho}{\mathcal{M}}\,\mathrm{d}x \le 0.
\end{aligned}
\tag{1.17}
$$

Notice, in above, we used the identity

$$(\phi'\left(\frac{\rho}{\mathcal{M}}\right), \vec{u} \cdot \nabla\frac{\rho}{\mathcal{M}}) = (\vec{u}, \nabla\phi\left(\frac{\rho}{\mathcal{M}}\right)) = 0,$$

which is due to the integration by parts and properties $\nabla \cdot \vec{u} = 0$ and $\vec{u} \cdot \vec{n}|_\Gamma = 0$. This energy dissipation law was first observed by [4]. In the case of $\phi(x) = (x - 1)^2$, the (1.17) reduces to the following energy dissipation law with respect to the Pearson χ^2-divergence

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \frac{(\rho - \mathcal{M})^2}{\mathcal{M}}\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \frac{\rho^2}{\mathcal{M}}\,\mathrm{d}x = -2D\int_\Omega \mathcal{M}\left|\nabla\frac{\rho}{\mathcal{M}}\right|^2\,\mathrm{d}x \le 0. \tag{1.18}$$

In addition, from the Poincare's inequality in $L^2(\Omega)$ with $u = \frac{\rho}{\mathcal{M}} - 1$, there exists a constant c, such that,

$$\int_\Omega |u|^2\mathcal{M}\,\mathrm{d}x \le c(\mathcal{M}\nabla u, D\nabla u). \tag{1.19}$$

We obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \frac{(\rho - \mathcal{M})^2}{\mathcal{M}}\,\mathrm{d}x = -2D\int_\Omega \mathcal{M}\left|\nabla\frac{\rho}{\mathcal{M}}\right|^2\,\mathrm{d}x \le -\frac{2D}{c}\int_\Omega \frac{(\rho - \mathcal{M})^2}{\mathcal{M}}\,\mathrm{d}x. \tag{1.20}$$

Then, by Gronwall's inequality, the above inequality gives the exponential decay from dynamic solution ρ to the equilibrium $\mathcal{M}$,

$$\int_\Omega (\frac{\rho}{\mathcal{M}} - 1)^2\mathcal{M}\,\mathrm{d}x = \int_\Omega \frac{(\rho - \mathcal{M})^2}{\mathcal{M}}\,\mathrm{d}x \le e^{-\frac{2D}{c}t}\int_\Omega \frac{(\rho^0 - \mathcal{M})^2}{\mathcal{M}}\,\mathrm{d}x. \tag{1.21}$$

The ergodicity result above implies that as long as a positive invariant measure $\mathcal{M}$ exists, although sometimes its explicit form may not be known, one still can solve it as an equilibrium solution to the original model with no-flux boundary condition (1.3), i.e., the invariant measure $\mathcal{M}$ can be obtained from computing the steaty state solution of the following problem

$$\partial_t \rho = \nabla \cdot (D \nabla \rho) - \nabla \cdot (\vec{b} \rho) \qquad \text{in } \{t > 0\} \times \Omega,$$

$$\rho = \rho^0 \qquad \text{in } \{t = 0\} \times \Omega,$$

$$-\rho \vec{b} \cdot \vec{n} + D \nabla \rho \cdot \vec{n} = 0 \qquad \text{on } \{t > 0\} \times \partial\Omega.$$

Let us refer the above equation as **Model 2**, or the conservative form without decomposition. The vector field $\vec{b}$ is a prescribed time-independent continuously differentiable function. Here, we emphasis $\vec{b}$ can be any general drift field. In order to produce a non-negative invariant measure, a positivity-preserving numerical scheme for solving **Model 2** is also preferred.

1.2 State of the Art

The computational methods for solving the Fokker–Planck equation (1.2) or general convection–diffusion equations have been extensively investigated. Without being exhaustive, we mention several pioneering studies. A well-known positivity-preserving finite volume scheme was proposed by Scharfetter and Gummel for some one-dimensional semiconductor device equations [33]. Extending the Scharfetter-Gummel finite volume scheme to a variety of drift terms, which include the irreversible expression (1.2), and to different boundary conditions, is studied in [2, 3, 7, 28, 29, 38]. These numerical schemes are first order accurate structure-preserving and enjoy many good properties, such as preserving positivity and dissipating energy; see also the underlying Markov process structure for upwind schemes in [9, 15]. Besides the accuracy and positivity, when numerically solving Fokker–Planck equations, the large time convergence to the invariant measure is also essential. The ergodicity for general irreversible process described by (1.2) and the equivalence with the corresponding reversible process (with a special drift in gradient form) were proved in [8]; see also reviews in [1] and [4] for the analysis including mixed boundary values. Designing schemes that also preserve the large time convergence to the invariant measure, particularly for the general irreversible process without gradient structure, has been attracting lots of attentions, e.g., see [23] for Fokker–Planck equation in the whole space, and see [6, 12] and the references therein for boundary-driven convection–diffusion problems. To the best knowledge of the authors, high order accurate schemes for solving Fokker–Planck equation (1.2) with a general drift field, which preserve all the desired properties, such as positivity, energy dissipation relation, and particularly the exponential convergence to the invariant measure, are still not available in literatures. High order schemes for Fokker–Planck equation with gradient flow structure was proved in [19] and for generalized Allen-Cahn equation was proved in [34]. Although, a comprehensive review on various applications of Fokker–Planck equation brought by irreversible stochastic processes is out of the scope of the present paper, we highlight the general irreversible processes and the processes constructed from numerical schemes for the corresponding Fokker–Planck equations have extensively important applications, which include but not limited to the transition path computations [11, 13, 24, 25, 35, 37] and the accelerated sampling and optimization [10, 14, 36, 39, 40].

1.3 Main Results, Methodology, and Contributions

In general, for the Fokker–Planck equations with generic drift terms, it is nontrivial to construct high-order accurate numerical schemes that can preserve all the following structures: (*i*) mass conservation law (1.4); (*ii*) energy dissipation relation (1.18); (*iii*) well-balancedness, i.e., numerical equilibrium recovers given invariant measure $\mathcal{M}$; and (*iv*) ergodicity/spectral

gap estimate (1.21). In this paper, we focus on constructing and analyzing second and fourth order in space numerical schemes via finite difference implementation of the finite element method for solving Fokker–Planck equations, mainly for **Model 1**. We will obtain all the good properties (i)–(iv) in the fully discrete second and fourth order schemes for **Model 1**.

Our algorithms enjoy desired numerical properties. Benefiting from the inherent nature of finite element method, the discrete mass conservation law is satisfied naturally. If the matrix of the linear system in the backward Euler time discretization is a monotone matrix, i.e., its inverse matrix has non-negative entries, then we call such schemes *monotone* schemes. We show that the schemes for **Model 1** are monotone under practical mesh conditions and time step constraints in Sect. 3, thus both the positivity of the numerical solution and the discrete energy dissipation law for any convex function f hold. Define discrete energy as

$$E^n = \sum_i \omega_i \mathcal{M}_i f \left(\frac{\rho_i^n}{\mathcal{M}_i} \right). \tag{1.22}$$

Then the quantity E^n is non-increasing with respect to time step n. In particular, by selecting the convex function $f(x) = (x-1)^2$, we show discrete Pearson χ^2-divergence energy dissipation law. Finally, the invariant measure is also recovered with an exponential convergence rate. For the definition of notation in (1.22) and more details on related proofs are shown in Sect. 4.

When $\vec{u} \equiv 0$, the scheme in this paper for **Model 1** reduces to the scheme for the Fokker–Planck equation in [19], thus the monotonicity discussion of the fourth order scheme is similar to those in [19]. However, due to the extra term $\nabla \cdot \left(\vec{u} \frac{\rho}{\mathcal{M}} \right)$ in **Model 1**, the monotonicity discussion in Sect. 3 is not only necessary but also nontrivial. More importantly, the mesh size and time step constraints for monotonicity in Sect. 3 are simpler than those in [19], even for the case $\vec{u} \equiv 0$, which is another contribution of this paper.

1.4 Organization of the Paper

The rest of this paper is organized as follows. In Sect. 2, we introduce our numerical schemes, which are constructed by finite difference implementation of Q^1 and Q^2 continuous finite element methods. In Sect. 3, we show the monotonicity of our second-order and fourth-order schemes in one and two dimension. The system matrices from our fourth order schemes no longer hold the M-matrix structure, however, we still obtain the monotonicity under simple sufficient conditions. The structure-preserving properties are discussed in Sect. 4. Numerical experiment validations are in Sect. 5. Concluding remarks are given in Sect. 6.

2 The Numerical Schemes

Consider a rectangular computational domain $\Omega \subset \mathbb{R}^d$ ($d = 1$ or 2) with unit outward normal $\vec{n}$. Uniformly partition the time interval $[0, T]$ into N_{st} subintervals. Let $\Delta t = T/N_{st}$ denote the time step size. We discretize two model problems in the previous section with the only unknown density ρ.

2.1 A First Order Accurate Time Discretization

Let ρ^n denote the solution at time step n. For Model 1 with a given incompressible $\vec{u}$ satifying $\vec{u} \cdot \vec{n} = 0$ along the boundary $\partial\Omega$ and a given invariant measure $\mathcal{M}$, we consider the following backward Euler time discretization:

$$\rho^{n+1} = \rho^n + \Delta t\, \nabla \cdot \left(D\mathcal{M}\nabla \frac{\rho^{n+1}}{\mathcal{M}} \right) + \Delta t\, \nabla \cdot \left(\vec{u}\, \frac{\rho^{n+1}}{\mathcal{M}} \right). \tag{2.1}$$

For convenience, introduce an auxiliary variable $g := \frac{\rho}{\mathcal{M}}$, then g^{n+1} can be computed as follows:

$$\mathcal{M}g^{n+1} - \Delta t\, \nabla \cdot (D\mathcal{M}\nabla g^{n+1}) - \Delta t\, \nabla \cdot (\vec{u}g^{n+1}) = \mathcal{M}g^n \quad \text{in } \Omega, \tag{2.2a}$$

$$D\mathcal{M}\nabla g^{n+1} \cdot \vec{n} = 0, \quad \vec{u} \cdot \vec{n} = 0 \quad \text{on } \partial\Omega. \tag{2.2b}$$

For solving Model 2 with given $\vec{b}$, the same backward Euler time discretization is given as

$$\rho^{n+1} - \Delta t\, \nabla \cdot (D\nabla\rho^{n+1}) + \Delta t\, \nabla \cdot (\vec{b}\rho^{n+1}) = \rho^n \quad \text{in } \Omega, \tag{2.3a}$$

$$-\rho^{n+1}\vec{b} \cdot \vec{n} + D\nabla\rho^{n+1} \cdot \vec{n} = 0 \quad \text{on } \partial\Omega. \tag{2.3b}$$

Although we do not have explicit formula for $\mathcal{M}$ in the most general case Model 2, for compact domain, we still know the existence of such an invariant measure and thus we can use $\mathcal{M}$ to obtain the stability and energy dissipation relation. Similar to the derivation of (1.8), we recast (2.3a) as

$$\rho^{n+1} - \Delta t\, \nabla \cdot \left(D\mathcal{M}\nabla \frac{\rho^{n+1}}{\mathcal{M}} \right) - \Delta t\, \nabla \cdot \left(\vec{u}\frac{\rho^{n+1}}{\mathcal{M}} \right) = \rho^n \quad \text{in } \Omega, \tag{2.4}$$

which is exactly the same as (2.1). Therefore, thanks to the existence of $\mathcal{M}$ and the fact that $\mathcal{M}$-decomposition reduces Model 2 to Model 1, the stability analysis and energy dissipation law for the backward Euler scheme of **Model 2** can also be derived, with an unknown function $\mathcal{M}$. We have the following proposition for the time discretization (2.1) and (2.3).

Proposition 2.1 *Let the invariant measure satisfying (1.5) and (1.6) be $\mathcal{M}$, which has no explicit formula for Model 2. The backward Euler time discretization (2.1) for **Model 1** and (2.3) for **Model 2** satisfy the stability estimate*

$$\frac{1}{\Delta t} \int_\Omega \frac{(\rho^{n+1})^2}{\mathcal{M}} - \frac{1}{\Delta t} \int_\Omega \frac{(\rho^n)^2}{\mathcal{M}} \le -2D \int_\Omega \mathcal{M} \left| \nabla \frac{\rho^{n+1}}{\mathcal{M}} \right|^2, \tag{2.5}$$

and the exponential decay of energy

$$\int_\Omega \frac{(\rho^n - \mathcal{M})^2}{\mathcal{M}} \le (1-\beta)^n \int_\Omega \frac{(\rho^0 - \mathcal{M})^2}{\mathcal{M}} \le e^{-\beta n} \int_\Omega \frac{(\rho^0 - \mathcal{M})^2}{\mathcal{M}}, \tag{2.6}$$

where $\beta = \frac{2D\Delta t}{c + 2D\Delta t} > 0$.

Proof First, since the invariant measure $\mathcal{M}$ always exists due to the ergodicity of the continuous Fokker–Planck equation on a compact domain (2.3) can be recast as (2.1); as we mentioned above.

   Second, for (2.1), i.e., (2.2) in terms of $g^{n+1} = \frac{\rho^{n+1}}{\mathcal{M}}$, we prove the stability estimate (2.5). Multiplying (2.2a) by g^{n+1} and integrating in Ω, we have

$$\int_\Omega \mathcal{M}|g^{n+1}|^2 + \Delta t \left(D\mathcal{M}\nabla g^{n+1}, \nabla g^{n+1}\right)$$

$$= \int_\Omega \mathcal{M}g^n g^{n+1} \le \int_\Omega \mathcal{M}\left(\frac{|g^{n+1}|^2}{2} + \frac{|g^n|^2}{2}\right). \tag{2.7}$$

Here in the first equality, we used (1.16) and integration by parts with boundary condition (1.11). Therefore, we obtain

$$\int_\Omega \mathcal{M}|g^{n+1}|^2 + 2D\Delta t \int_\Omega \mathcal{M}|\nabla g^{n+1}|^2 \le \int_\Omega \mathcal{M}|g^n|^2, \tag{2.8}$$

and thus (2.5) holds.

Third, from the mass conservation $\int_\Omega \rho^0 = \int_\Omega \mathcal{M}$, we know $\int \frac{\rho^2}{\mathcal{M}}dx - 1 = \int \frac{(\rho-\mathcal{M})^2}{\mathcal{M}}dx$ and thus (2.5) becomes

$$\frac{1}{\Delta t}\int_\Omega \frac{(\rho^{n+1}-\mathcal{M})^2}{\mathcal{M}} - \frac{1}{\Delta t}\int_\Omega \frac{(\rho^n-\mathcal{M})^2}{\mathcal{M}} \le -2D\int_\Omega \mathcal{M}\left|\nabla\frac{\rho^{n+1}}{\mathcal{M}}\right|^2. \tag{2.9}$$

Then by Poincare's inequality (1.19) and the inequality in (1.20), we obtain

$$\frac{1}{\Delta t}\int_\Omega \frac{(\rho^{n+1}-\mathcal{M})^2}{\mathcal{M}} - \frac{1}{\Delta t}\int_\Omega \frac{(\rho^n-\mathcal{M})^2}{\mathcal{M}} \le -\frac{2D}{c}\int_\Omega \frac{(\rho^{n+1}-\mathcal{M})^2}{\mathcal{M}}. \tag{2.10}$$

Repeating $\int_\Omega \frac{(\rho^n-\mathcal{M})^2}{\mathcal{M}} \le \frac{1}{1+\frac{2D}{c}\Delta t}\int_\Omega \frac{(\rho^{n-1}-\mathcal{M})^2}{\mathcal{M}}$ for n times we obtain (2.6). $\square$

On the other hand, without the explicit expression of $\mathcal{M}$, we cannot obtain similar theoretical results for the full discretization of Model 2 beyond the positivity and mass conservation. Without having any explicit information on the invariant measure, it is challenging to capture a steady solution by directly solving the dynamic equation in Model 2. For the above two reasons, in the rest of the paper, we only focus on the analysis of the scheme (2.2) for Model 1.

2.2 The Continuous Q^k Finite Element Method for Spatial Derivatives

Given the function g^n, the semi-discrete scheme (2.2) is a variable coefficient elliptic equation for g^{n+1} with homogeneous Neumann boundary conditions. Recall that $(\cdot, \cdot)$ denote the L^2 inner product on Ω. After multiplying a test function $\phi \in H^1(\Omega)$ and integration by parts, the equivalent variational formulation for the unknown $g^{n+1} \in H^1(\Omega)$ to the equation (2.2) can be written as:

$$(\mathcal{M}g^{n+1}, \phi) + \Delta t\,(D\mathcal{M}\nabla g^{n+1}, \nabla\phi) + \Delta t\,(\vec{u}g^{n+1}, \nabla\phi) = (\mathcal{M}g^n, \phi), \quad \forall\phi \in H^1(\Omega).$$

Next we consider the finite element method for the spatial operators. Let $\mathcal{E}_h = \{E\}$ denote a uniform rectangular partition of the rectangular computational domain Ω. For any integer $k \ge 1$, let $\mathbb{Q}^k(E)$ be the space of tensor product polynomials of degree at most k. As an example, for two dimensions,

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

The continuous piecewise $\mathbb{Q}^k$ polynomial space $V^h \subset H^1(\Omega)$ is defined by

$$V^h = \{v_h \in C(\Omega): \; v_h|_E \in \mathbb{Q}^k(E), \; \forall E \in \mathcal{E}_h\}.$$

Given $g_h^n \in V^h$, the Q^k finite element method is to find $g_h^{n+1} \in V^h$ satisfying

$$(\mathcal{M} g_h^{n+1}, \phi_h) + \Delta t \, (D\mathcal{M}\nabla g_h^{n+1}, \nabla \phi_h) + \Delta t \, (\vec{u} g_h^{n+1}, \nabla \phi_h) = (\mathcal{M} g_h^n, \phi_h), \quad \forall \phi_h \in V^h. \tag{2.11}$$

Denote $\rho_h^n = \mathcal{M} g_h^n$. The finite element scheme (2.11) has the following properties.

Proposition 2.2 *For any $n \geq 0$, the total mass is conserved in the scheme* (2.11), *namely*

$$\int_\Omega \rho_h^n = \int_\Omega \rho_h^0. \tag{2.12}$$

Proof Take $\phi_h = 1$ in (2.11), then for any $n \geq 0$, the identity $(\mathcal{M} g_h^{n+1}, 1) = (\mathcal{M} g_h^n, 1)$ holds. □

Proposition 2.3 *For any $n \geq 0$, the following two inequalities hold in the scheme* (2.11),

$$\frac{1}{\Delta t} \int_\Omega \frac{(\rho_h^{n+1})^2}{\mathcal{M}} - \frac{1}{\Delta t} \int_\Omega \frac{(\rho_h^n)^2}{\mathcal{M}} \leq -2D \int_\Omega \mathcal{M} \left| \nabla \frac{\rho_h^{n+1}}{\mathcal{M}} \right|^2, \tag{2.13a}$$

$$\frac{1}{\Delta t} \int_\Omega \frac{(\rho_h^{n+1} - \mathcal{M})^2}{\mathcal{M}} - \frac{1}{\Delta t} \int_\Omega \frac{(\rho_h^n - \mathcal{M})^2}{\mathcal{M}} \leq -2D \int_\Omega \mathcal{M} \left| \nabla \frac{\rho_h^{n+1}}{\mathcal{M}} \right|^2. \tag{2.13b}$$

Proof Consider $\mathcal{M} > 0$ is time independent and $D > 0$ is constant. For any $\phi_h \in V^h$, let us rewrite (2.11) into the following form

$$(\sqrt{\mathcal{M}} g_h^{n+1} - \sqrt{\mathcal{M}} g_h^n, \sqrt{\mathcal{M}} \phi_h) + \Delta t \, (\vec{u} g_h^{n+1}, \nabla \phi_h) = -\Delta t D \, (\sqrt{\mathcal{M}} \nabla g_h^{n+1}, \sqrt{\mathcal{M}} \nabla \phi_h). \tag{2.14}$$

Take $\phi_h = g_h^{n+1} \in V^h$ in (2.14). By formula $(a - b)a = \frac{1}{2}a^2 - \frac{1}{2}b^2 + \frac{1}{2}(a - b)^2$, the first term on the left-hand side of above becomes

$$(\sqrt{\mathcal{M}} g_h^{n+1} - \sqrt{\mathcal{M}} g_h^n, \sqrt{\mathcal{M}} g_h^{n+1}) = \frac{1}{2}(\sqrt{\mathcal{M}} g_h^{n+1}, \sqrt{\mathcal{M}} g_h^{n+1}) - \frac{1}{2}(\sqrt{\mathcal{M}} g_h^n, \sqrt{\mathcal{M}} g_h^n)$$
$$+ \frac{1}{2}(\sqrt{\mathcal{M}} g_h^{n+1} - \sqrt{\mathcal{M}} g_h^n, \sqrt{\mathcal{M}} g_h^{n+1} - \sqrt{\mathcal{M}} g_h^n). \tag{2.15}$$

For the second term on the left-hand side of (2.14), we have

$$\Delta t \, (\vec{u} g_h^{n+1}, \nabla g_h^{n+1}) = \Delta t \int_\Omega \vec{u} \cdot g_h^{n+1} \nabla g_h^{n+1} = \frac{\Delta t}{2} \int_\Omega \vec{u} \cdot \nabla |g_h^{n+1}|^2.$$

Since the field $\vec{u}$ is incompressible, we have

$$\Delta t \, (\vec{u} g_h^{n+1}, \nabla g_h^{n+1}) = \frac{\Delta t}{2} \int_\Omega \left(\vec{u} \cdot \nabla |g_h^{n+1}|^2 + (\nabla \cdot \vec{u})|g_h^{n+1}|^2 \right) = \frac{\Delta t}{2} \int_\Omega \nabla \cdot (\vec{u} \, |g_h^{n+1}|^2).$$

By condition $\vec{u} \cdot \vec{n} = 0$, see the last equation in (1.13), we have

$$\Delta t \, (\vec{u} g_h^{n+1}, \nabla g_h^{n+1}) = \frac{\Delta t}{2} \int_{\partial\Omega} |g_h^{n+1}|^2 \, (\vec{u} \cdot \vec{n}) = 0. \tag{2.16}$$

Thus, select $\phi_h = g_h^{n+1} \in V^h$, substitute (2.15) and (2.16) into (2.14), we get the following inequality

$$
\frac{1}{2}(\sqrt{\mathcal{M}}g_h^{n+1}, \sqrt{\mathcal{M}}g_h^{n+1}) - \frac{1}{2}(\sqrt{\mathcal{M}}g_h^n, \sqrt{\mathcal{M}}g_h^n)
$$
$$
= -\Delta t\, D\,(\sqrt{\mathcal{M}}\nabla g_h^{n+1}, \sqrt{\mathcal{M}}\nabla g_h^{n+1}) - \frac{1}{2}(\sqrt{\mathcal{M}}g_h^{n+1} - \sqrt{\mathcal{M}}g_h^n, \sqrt{\mathcal{M}}g_h^{n+1} - \sqrt{\mathcal{M}}g_h^n)
$$
$$
\leq -\Delta t\, D\,(\sqrt{\mathcal{M}}\nabla g_h^{n+1}, \sqrt{\mathcal{M}}\nabla g_h^{n+1}). \tag{2.17}
$$

Recall that $\rho_h^{n+1} = \mathcal{M}g_h^{n+1}$, which is equivalent to $g_h^{n+1} = \rho_h^{n+1}/\mathcal{M}$. Multiply (2.17) by $2/\Delta t$ on both side, we obtain (2.13a). Finally, from mass conservation (2.12), we also have (2.13b) holds. $\qquad\square$

Proposition 2.4 *For any $n \geq 0$, the following inequality hold*

$$
\frac{1}{\Delta t}\int_{\Omega}\frac{(\rho_h^{n+1} - \mathcal{M})^2}{\mathcal{M}} - \frac{1}{\Delta t}\int_{\Omega}\frac{(\rho_h^n - \mathcal{M})^2}{\mathcal{M}} \leq -\frac{2D}{c}\int_{\Omega}\frac{(\rho_h^{n+1} - \mathcal{M})^2}{\mathcal{M}}. \tag{2.18}
$$

Consequently, we have the exponential decay of ρ_h^{n+1}

$$
\int_{\Omega}\frac{(\rho_h^n - \mathcal{M})^2}{\mathcal{M}} \leq (1 - \beta)^n \int_{\Omega}\frac{(\rho_h^0 - \mathcal{M})^2}{\mathcal{M}} \leq e^{-\beta n}\int_{\Omega}\frac{(\rho_h^0 - \mathcal{M})^2}{\mathcal{M}}, \tag{2.19}
$$

where $\beta = \frac{2D\Delta t}{c + 2D\Delta t} > 0$.

The proof of this proposition is identically same as (2.6) in Proposition 2.1.

2.3 The Finite Difference Implementation

For implementing the finite element method above, usually quadrature is used for computing the integrals. On the other hand, it is well known that a finite element method with suitable quadrature is also a finite difference scheme. When $(k+1)$-point Gauss-Lobatto quadrature is used for (2.11), the scheme is also referred to as Q^k spectral element method [27], and it can be regarded as a $(k+2)$th order accurate finite difference scheme with respect to the discrete ℓ^2-norm at quadrature points for $k \geq 2$. See [20, 22] for rigorous a priori error estimates.

Any $\mathbb{Q}^k$ polynomial on rectangular element can be represented as a Lagrangian interpolation polynomial at $(k+1)^d$ Gauss-Lobatto points. Thus, these points are not only quadrature nodes but also representing all degrees of freedom. In addition, for $k \leq 2$, all Gauss-Lobatto points form a uniform grid. Let $\langle\cdot, \cdot\rangle$ denote the inner product $(\cdot, \cdot)$ evaluated by Gauss-Lobatto quadrature. Then, after replacing all integrals in (2.11) by Gauss-Lobatto quadrature, the scheme becomes:

$$
\langle \mathcal{M}g_h^{n+1}, \phi_h\rangle + \Delta t\,\langle D\mathcal{M}\nabla g_h^{n+1}, \nabla\phi_h\rangle + \Delta t\,\langle \vec{u}g_h^{n+1}, \nabla\phi_h\rangle = \langle \mathcal{M}g_h^n, \phi_h\rangle, \quad \forall\phi_h \in V^h. \tag{2.20}
$$

In particular, we can obtain a fourth order accurate finite difference scheme when using Q^2 element with 3-point Gauss-Lobatto quadrature in (2.11). For the Q^1 element method with 2-point Gauss-Lobatto quadrature in (2.11), we get a second order accurate finite difference scheme, which is exactly the same as the centered difference at interior grid points. These two finite difference schemes can be proved monotone for convection–diffusion operators [19, 21, 34], thus positivity-preserving and energy decaying. In this paper, we only consider these two finite difference schemes.

2.4 The Second Order Scheme in One Dimension

For $\Omega = [-L, L]$, consider uniform grid points with spacing h, $-L = x_1 < x_2 < \cdots < x_N = L$. The mesh $\mathcal{E}_h$ consists of intervals $[x_i, x_{i+1}]$, $i = 1, \cdots, N-1$.

Following the discussion in [19], it is straightforward to verify that the scheme (2.20) can be equivalently written in the following finite difference form:

$$
\mathcal{M}_1 g_1^{n+1} \; -\Delta t \, \frac{u_1 g_1^{n+1} + u_2 g_2^{n+1}}{h} + \Delta t \, \frac{D(\mathcal{M}_1 + \mathcal{M}_2) g_1^{n+1} - D(\mathcal{M}_1 + \mathcal{M}_2) g_2^{n+1}}{h^2}
$$
$$
= \mathcal{M}_1 g_1^n;
$$

for $i = 2, \cdots, N-1$,

$$
\mathcal{M}_i g_i^{n+1} + \Delta t \, \frac{u_{i-1} g_{i-1}^{n+1} - u_{i+1} g_{i+1}^{n+1}}{2h}
$$
$$
+ \Delta t \, \frac{-D(\mathcal{M}_{i-1} + \mathcal{M}_i) g_{i-1}^{n+1} + D(\mathcal{M}_{i-1} + 2\mathcal{M}_i + \mathcal{M}_{i+1}) g_i^{n+1} - D(\mathcal{M}_i + \mathcal{M}_{i+1}) g_{i+1}^{n+1}}{2h^2}
$$
$$
= \mathcal{M}_i g_i^n,
$$

and

$$
\mathcal{M}_N g_N^{n+1} + \Delta t \, \frac{u_{N-1} g_{N-1}^{n+1} + u_N g_N^{n+1}}{h}
$$
$$
+ \Delta t \, \frac{-D(\mathcal{M}_{N-1} + \mathcal{M}_N) g_{N-1}^{n+1} + D(\mathcal{M}_{N-1} + \mathcal{M}_N) g_N^{n+1}}{h^2}
$$
$$
= \mathcal{M}_N g_N^n.
$$

For convenience, we introduce some ghost point values defined as

$$
g_0 := g_2, \quad g_{N+1} := g_{N-1}, \mathcal{M}_0 := \mathcal{M}_2, \mathcal{M}_{N+1} := \mathcal{M}_{N-1}
$$

and

$$
u_0 := -u_2, \quad u_{N+1} := -u_{N-1}.
$$

We emphasize that ghost point values are only used to simplify the representation of the scheme, and they are not needed or used in the implementation. Recall that the velocity field in Model 1 satisfies boundary condition $\vec{u} \cdot \vec{n} = 0$, thus $u_1 = u_N = 0$. With the ghost point value notation and the boundary condition $u_1 = u_N = 0$, it is straightforward to see that the scheme above can be written as

$$
\mathcal{M}_i g_i^{n+1} + \Delta t \, \frac{u_{i-1} g_{i-1}^{n+1} - u_{i+1} g_{i+1}^{n+1}}{2h}
$$
$$
+ \Delta t \, \frac{-D(\mathcal{M}_{i-1} + \mathcal{M}_i) g_{i-1}^{n+1} + D(\mathcal{M}_{i-1} + 2\mathcal{M}_i + \mathcal{M}_{i+1}) g_i^{n+1} - D(\mathcal{M}_i + \mathcal{M}_{i+1}) g_{i+1}^{n+1}}{2h^2}
$$
$$
= \mathcal{M}_i g_i^n, \quad \forall i = 1, \cdots, N. \tag{2.21}
$$

The finite difference scheme (2.21) is obtained from finite element method with quadrature, and its second order accuracy is trivially implied by standard finite element error estimates. At domain interior points, the scheme (2.4) is the same as the traditional second order centered difference scheme. However, the traditional centered difference scheme would give a different

boundary scheme, for which the second order accuracy is quite difficult to justify due to the first order truncation error at boundaries. See Remark 3.3 in [19] for more details.

2.5 The Second Order Scheme in Two Dimensions

For a square domain $\Omega = [-L, L]^2$, we consider a uniform grid with spacing h consisting of $-L = x_1 < x_2 < \cdots < x_N = L$ and $-L = y_1 < y_2 < \cdots < y_N = L$. The mesh $\mathcal{E}_h$ consists of $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$, $i, j = 1, \cdots, N - 1$. We employ the abbreviation $\mathcal{M}_{i,j} = \mathcal{M}(x_i, y_j)$ for a function. With similarly defined ghost point values and the boundary condition for velocity $\vec{u} \cdot \vec{n} = 0$, the scheme finite difference form of (2.20) with Q^1 element and 2×2 Gauss-Lobatto quadrature is given as

$$
\begin{aligned}
&\mathcal{M}_{i,j} g_{ij}^{n+1} + \Delta t \frac{u_{i-1,j} g_{i-1,j}^{n+1} - u_{i+1,j} g_{i+1,j}^{n+1}}{2h} + \Delta t \frac{v_{i,j-1} g_{i,j-1}^{n+1} - v_{i,j+1} g_{i,j+1}^{n+1}}{2h} \\
&+ \Delta t \frac{-D(\mathcal{M}_{i-1,j} + \mathcal{M}_{i,j}) g_{i-1,j}^{n+1} + D(\mathcal{M}_{i-1,j} + 2\mathcal{M}_{i,j} + \mathcal{M}_{i+1,j}) g_{ij}^{n+1} - D(\mathcal{M}_{i,j} + \mathcal{M}_{i+1,j}) g_{i+1,j}^{n+1}}{2h^2} \\
&+ \Delta t \frac{-D(\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j}) g_{i,j-1}^{n+1} + D(\mathcal{M}_{i,j-1} + 2\mathcal{M}_{i,j} + \mathcal{M}_{i,j+1}) g_{ij}^{n+1} - D(\mathcal{M}_{i,j} + \mathcal{M}_{i,j+1}) g_{i,j+1}^{n+1}}{2h^2} \\
&= \mathcal{M}_{i,j} g_{ij}^n, \quad \forall i, j = 1, \cdots, N.
\end{aligned}
\tag{2.22}
$$

2.6 The Fourth Order Scheme in One Dimension

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 an 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. The mesh $\mathcal{E}_h$ consists of intervals $[x_i, x_{i+2}]$, $i = 1, 3, \cdots, N - 2$.

Following the discussion in [19, 22, 34], it is straightforward to verify that the scheme (2.20) with P^2 element and 3-point Gauss-Lobatto quadrature can be equivalently written in the following finite difference form:

$$
\begin{aligned}
&\mathcal{M}_1 g_1^{n+1} + \Delta t \frac{-3u_1 g_1^{n+1} - 4u_2 g_2^{n+1} + u_3 g_3^{n+1}}{2h} \\
&+ \Delta t \frac{D(9\mathcal{M}_1 + 4\mathcal{M}_2 + \mathcal{M}_3) g_1^{n+1} - D(12\mathcal{M}_1 + 4\mathcal{M}_3) g_2^{n+1}}{4h^2} \\
&+ \Delta t \frac{D(3\mathcal{M}_1 - 4\mathcal{M}_2 + 3\mathcal{M}_3) g_3^{n+1}}{4h^2} = \mathcal{M}_1 g_1^n; \\
&\mathcal{M}_i g_i^{n+1} + \Delta t \frac{-u_{i-2} g_{i-2}^{n+1} + 4u_{i-1} g_{i-1}^{n+1} - 4u_{i+1} g_{i+1}^{n+1} + u_{i+2} g_{i+2}^{n+1}}{4h} \\
&+ \Delta t \frac{D(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i) g_{i-2}^{n+1} - D(4\mathcal{M}_{i-2} + 12\mathcal{M}_i) g_{i-1}^{n+1}}{8h^2} \\
&+ \Delta t \frac{D(\mathcal{M}_{i-2} + 4\mathcal{M}_{i-1} + 18\mathcal{M}_i + 4\mathcal{M}_{i+1} + \mathcal{M}_{i+2}) g_i^{n+1}}{8h^2} \\
&+ \Delta t \frac{-D(12\mathcal{M}_i + 4\mathcal{M}_{i+2}) g_{i+1}^{n+1} + D(3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2}) g_{i+2}^{n+1}}{8h^2} = \mathcal{M}_i g_i^n, \\
&\forall i = 3, 5, \cdots, N - 2 \quad \text{(corresponding to interior knots)}; \\
&\mathcal{M}_i g_i^{n+1} + \Delta t \frac{u_{i-1} g_{i-1}^{n+1} - u_{i+1} g_{i+1}^{n+1}}{2h}
\end{aligned}
$$

$$+\Delta t \, \frac{-D(3\mathcal{M}_{i-1} + \mathcal{M}_{i+1})g_{i-1}^{n+1} + 4D(\mathcal{M}_{i-1} + \mathcal{M}_{i+1})g_{i}^{n+1} - D(\mathcal{M}_{i-1} + 3\mathcal{M}_{i+1})g_{i+1}^{n+1}}{4h^2}$$

$$= \mathcal{M}_i g_i^n,$$

$\forall i = 2, 4, \cdots N - 1$　(corresponding to interval midpoints);

$$\mathcal{M}_N g_N^{n+1} + \Delta t \, \frac{-u_{N-2}g_{N-2}^{n+1} + 4u_{N-1}g_{N-1}^{n+1} + 3u_N g_N^{n+1}}{2h}$$

$$+\Delta t \, \frac{D(3\mathcal{M}_{N-2} - 4\mathcal{M}_{N-1} + 3\mathcal{M}_N)g_{N-2}^{n+1} - D(4\mathcal{M}_{N-2} + 12\mathcal{M}_N)g_{N-1}^{n+1}}{4h^2}$$

$$+\Delta t \, \frac{D(\mathcal{M}_{N-2} + 4\mathcal{M}_{N-1} + 9\mathcal{M}_N)g_N^{n+1}}{4h^2} = \mathcal{M}_N g_N^n.$$

For convenience, we introduce ghost point values defined as

$$g_{-1} := g_3, \, g_0 := g_2, \, g_{N+1} := g_{N-1}, \, g_{N+2} := g_{N-2},$$
$$\mathcal{M}_{-1} := \mathcal{M}_3, \, \mathcal{M}_0 := \mathcal{M}_2, \, \mathcal{M}_{N+1} := \mathcal{M}_{N-1}, \, \mathcal{M}_{N+2} := \mathcal{M}_{N-2},$$

and

$$u_{-1} := -u_3, \, u_0 := -u_2, \quad u_{N+1} := -u_{N-1}, \, u_{N+2} := -u_{N-2}.$$

With the ghost point value notation and the velocity boundary condition $u_1 = u_N = 0$, the fourth order finite difference scheme above can be written as
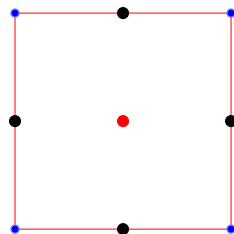
$$\mathcal{M}_i g_i^{n+1} + \Delta t \frac{-u_{i-2}g_{i-2}^{n+1} + 4u_{i-1}g_{i-1}^{n+1} - 4u_{i+1}g_{i+1}^{n+1} + u_{i+2}g_{i+2}^{n+1}}{4h}$$

$$+\Delta t \, \frac{D(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)g_{i-2}^{n+1} - D(4\mathcal{M}_{i-2} + 12\mathcal{M}_i)g_{i-1}^{n+1}}{8h^2}$$

$$+\Delta t \, \frac{D(\mathcal{M}_{i-2} + 4\mathcal{M}_{i-1} + 18\mathcal{M}_i + 4\mathcal{M}_{i+1} + \mathcal{M}_{i+2})g_i^{n+1}}{8h^2}$$

$$+\Delta t \, \frac{-D(12\mathcal{M}_i + 4\mathcal{M}_{i+2})g_{i+1}^{n+1} + D(3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2})g_{i+2}^{n+1}}{8h^2} = \mathcal{M}_i g_i^n,$$

$\forall i = 1, 5, \cdots, N$　(odd i, corresponding to knots);　　　　　　　　　　(2.23a)

$$\mathcal{M}_i g_i^{n+1} + \Delta t \, \frac{u_{i-1}g_{i-1}^{n+1} - u_{i+1}g_{i+1}^{n+1}}{2h}$$

$$+\Delta t \, \frac{-D(3\mathcal{M}_{i-1} + \mathcal{M}_{i+1})g_{i-1}^{n+1} + 4D(\mathcal{M}_{i-1} + \mathcal{M}_{i+1})g_i^{n+1} - D(\mathcal{M}_{i-1} + 3\mathcal{M}_{i+1})g_{i+1}^{n+1}}{4h^2}$$

$$= \mathcal{M}_i g_i^n,$$

$\forall i = 2, 4, \cdots N - 1$　(even i, corresponding to interval midpoints).　　　　(2.23b)

2.7 The Fourth Order Scheme in Two Dimensions

Assume the domain is $\Omega = [-L, L] \times [-L, L]$ with an 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 $N = 2k + 1$. For the Q^2 finite element method on uniform rectangular meshes, there are three types of grid point values as shown in Fig. 1:

Fig. 1 Three types of grid points: red cell center, blue knots and black edge centers for a Q^2 finite element cell (Color figure online)

$$
\begin{aligned}
\text{knot:} \quad & \text{both } i \text{ and } j \text{ are odd,} \\
\text{edge center:} \quad & i \text{ is even and } j \text{ is odd (parallel to x-axis)} \quad \text{or} \\
& i \text{ is odd and } j \text{ is even (parallel to y-axis),} \\
\text{cell center:} \quad & \text{both } i \text{ and } j \text{ are even.}
\end{aligned}
$$

The fourth order finite difference scheme is given as the following:

$$
\begin{aligned}
\mathcal{M}_{ij} g_{ij}^{n+1} &+ \Delta t \, \frac{-u_{i-2,j} g_{i-2,j}^{n+1} + 4 u_{i-1,j} g_{i-1,j}^{n+1} - 4 u_{i+1,j} g_{i+1,j}^{n+1} + u_{i+2,j} g_{i+2,j}^{n+1}}{4h} \\
&+ \Delta t \, \frac{-v_{i,j-2} g_{i,j-2}^{n+1} + 4 v_{i,j-1} g_{i,j-1}^{n+1} - 4 v_{i,j+1} g_{i,j+1}^{n+1} + v_{i,j+2} g_{i,j+2}^{n+1}}{4h} \\
&+ \Delta t \, \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j}) g_{i-2,j}^{n+1} - D(4\mathcal{M}_{i-2,j} + 12\mathcal{M}_{i,j}) g_{i-1,j}^{n+1}}{8h^2} \\
&+ \Delta t \, \frac{D(\mathcal{M}_{i-2,j} + 4\mathcal{M}_{i-1,j} + 18\mathcal{M}_{i,j} + 4\mathcal{M}_{i+1,j} + \mathcal{M}_{i+2,j}) g_{ij}^{n+1}}{8h^2} \\
&+ \Delta t \, \frac{-D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i+2,j}) g_{i+1,j}^{n+1} + D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j}) g_{i+2,j}^{n+1}}{8h^2} \\
&+ \Delta t \, \frac{D(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j}) g_{i,j-2}^{n+1} - D(4\mathcal{M}_{i,j-2} + 12\mathcal{M}_{i,j}) g_{i,j-1}^{n+1}}{8h^2} \\
&+ \Delta t \, \frac{D(\mathcal{M}_{i,j-2} + 4\mathcal{M}_{i,j-1} + 18\mathcal{M}_{i,j} + 4\mathcal{M}_{i,j+1} + \mathcal{M}_{i,j+2}) g_{ij}^{n+1}}{8h^2} \\
&+ \Delta t \, \frac{-D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i,j+2}) g_{i,j+1}^{n+1} + D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j+2}) g_{i,j+2}^{n+1}}{8h^2} \\
&= \mathcal{M}_{ij} g_{ij}^{n},
\end{aligned}
$$

if (x_i, y_j) is a knot;. $\hspace{4cm}$ (2.24a)

$$
\begin{aligned}
\mathcal{M}_{ij} g_{ij}^{n+1} &+ \Delta t \, \frac{v_{i,j-1} g_{i,j-1}^{n+1} - v_{i,j+1} g_{i,j+1}^{n+1}}{2h} \\
&+ \Delta t \, \frac{-u_{i-2,j} g_{i-2,j}^{n+1} + 4 u_{i-1,j} g_{i-1,j}^{n+1} - 4 u_{i+1,j} g_{i+1,j}^{n+1} + u_{i+2,j} g_{i+2,j}^{n+1}}{4h} \\
&+ \Delta t \, \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j}) g_{i-2,j}^{n+1} - D(4\mathcal{M}_{i-2,j} + 12\mathcal{M}_{i,j}) g_{i-1,j}^{n+1}}{8h^2} \\
&+ \Delta t \, \frac{D(\mathcal{M}_{i-2,j} + 4\mathcal{M}_{i-1,j} + 18\mathcal{M}_{ij} + 4\mathcal{M}_{i+1,j} + \mathcal{M}_{i+2,j}) g_{ij}^{n+1}}{8h^2}
\end{aligned}
$$

$$+ \Delta t \, \frac{-D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i+2,j})g_{i+1,j}^{n+1} + D(3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})g_{i+2,j}^{n+1}}{8h^2}$$

$$+ \Delta t \, \frac{-D(3\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})g_{i,j-1}^{n+1}}{4h^2}$$

$$+ \Delta t \, \frac{D(\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})g_{ij}^{n+1}}{h^2}$$

$$+ \Delta t \, \frac{-D(\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j+1})g_{i,j+1}^{n+1}}{4h^2} = \mathcal{M}_{ij}g_{ij}^n,$$

if (x_i, y_j) is an edge (parallel to y-axis) center;   (2.24b)

$$\mathcal{M}_{ij}g_{ij}^{n+1} + \Delta t \, \frac{u_{i-1,j}g_{i-1,j}^{n+1} - u_{i+1,j}g_{i+1,j}^{n+1}}{2h} + \Delta t \, \frac{v_{i,j-1}g_{i,j-1}^{n+1} - v_{i,j+1}g_{i,j+1}^{n+1}}{2h}$$

$$+ \Delta t \, \frac{-D(3\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j})g_{i-1,j}^{n+1} - D(\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i+1,j})g_{i+1,j}^{n+1}}{4h^2}$$

$$+ \Delta t \, \frac{D(\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j} + \mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})g_{ij}^{n+1}}{h^2}$$

$$+ \Delta t \, \frac{-D(3\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})g_{i,j-1}^{n+1} - D(\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j+1})g_{i,j+1}^{n+1}}{4h^2} = \mathcal{M}_{ij}g_{ij}^n,$$

if (x_i, y_j) is a cell center.   (2.24c)

For the grid point (x_i, y_j) which is an edge center for an edge parallel to x-axis, the scheme is very similar as above, thus omitted here.

3 Monotonicity of the Fully Discrete Schemes

In this section, we will prove the monotonicity of the two fully discrete schemes.

3.1 The M-Matrix Structure in the Second Order Scheme

A matrix A is called monotone if all entries of its inverse are nonnegative, namely, $A^{-1} \geq 0$. In this paper, all inequalities for matrices are entry-wise inequalities. A square matrix A is called an M-matrix if it can be expressed in the form $A = sI - B$, where $B \geq 0$ and s is greater than the spectral radius of B. There are many equivalent definitions or characterizations of M-matrix. A comprehensive review of M-matrix can be found in [30]. The nonsingular M-matrix is an inverse-positive matrix and it serves as a convenient tool for proving monotonicity. A sufficient and necessary condition to characteristic nonsingular M-matrix is stated as follows:

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

We also state a sufficient but not necessary condition to verify nonsingular M-matrix; c.f. [30].

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

3.1.1 The Second Order Scheme in One Dimension

We verify that the matrix in the scheme above satisfies Lemma 3.2. The following mesh constraint is sufficient for off-diagonal entries in the system matrix to be non-positive:

$$h|u_j| \leq D \min\{\mathcal{M}_{j-1}, \mathcal{M}_j, \mathcal{M}_{j+1}\}, \quad \forall j. \tag{3.1}$$

To guarantee the nonnegative row sums of the system matrix with at least one strictly positive row sum, the following constraints on time step size are sufficient:

$$\mathcal{M}_i + \Delta t \, \frac{u_{i-1} - u_{i+1}}{2h} > 0 \quad \forall i. \tag{3.2}$$

Recall that the velocity field $\vec{u}$ is incompressible in Model 1, thus $u(x) \equiv C$ in one dimension. So (3.2) is trivially satisfied for positive measure $\mathcal{M}$. So we have the following result.

Theorem 3.3 *Under the mesh and time step constraints* (3.1) *and* (3.2)*, the coefficient matrix for the unknown vector g^{n+1} in the second order finite difference scheme* (2.21) *forms an M-matrix and thus is monotone. In particular, in one dimension, discrete divergence free velocity field u is constant and the second order finite difference scheme* (2.21) *is monotone under the mesh constraint* (3.1).

3.1.2 The Second Order Scheme in Two Dimensions

Next we verify that the matrix in the scheme (2.22) satisfies Lemma 3.2. The following mesh constraint is sufficient for off-diagonal entries in the system matrix to be non-positive:

$$h|u_{i,j}| \leq D \min\{\mathcal{M}_{i-1,j}, \mathcal{M}_{i+1,j}, \mathcal{M}_{i,j}, \mathcal{M}_{i,j-1}, \mathcal{M}_{i,j+1}\}, \quad \forall i, j. \tag{3.3}$$

To guarantee the nonnegative row sums of the system matrix with at least one strictly positive row sum, the following constraints on time step size are sufficient:

$$\mathcal{M}_{i,j} + \Delta t \, \frac{u_{i-1,j} - u_{i+1,j}}{2h} + \Delta t \, \frac{v_{i,j-1} - v_{i,j+1}}{2h} > 0. \tag{3.4}$$

Notice that (3.4) is trivially satisfied for positive measure $\mathcal{M}$, if the following discrete divergence free constraint is satisfied

$$\frac{u_{i-1,j} - u_{i+1,j}}{2h} + \frac{v_{i,j-1} - v_{i,j+1}}{2h} = 0. \tag{3.5}$$

Recall that the velocity field $\vec{u}$ is incompressible in Model 1, thus one can preprocess the given $\vec{u}$ such that the velocity point values satisfies (3.5). So we have the following result.

Theorem 3.4 *Under the mesh and time step constraints* (3.3) *and* (3.4)*, the coefficient matrix for the unknown vector g^{n+1} in the second order finite difference scheme* (2.22) *forms an M-matrix thus is monotone. In particular, with a discrete divergence free velocity field satisfying* (3.5)*, the matrix in second order finite difference scheme* (2.22) *is monotone under the mesh constraint* (3.3).

3.2 Lorenz's Sufficient Condition for Monotonicity

In general, M-matrix structure is only a very conveneint condition for verifying monotonicity, rather than a necessary condition. Moreover, almost all high order schemes simply do not have any M-matrix structure due to positive off-diagonal entries. In [26], Lorenz proposed a convenient sufficient condition for a matrix to be a product of M-matrices. We review Lorenz's sufficient condition in this subsection. See also [21] for a review.

Let matrix A_d be a diagonal matrix denoting the diagonal part of $\mathsf{A} = [a_{ij}] \in \mathbb{R}^{n \times n}$ and $\mathsf{A}_a = \mathsf{A} - \mathsf{A}_d$. We further decompose A_a into positive and negative off-diagonal parts. More precisely, we define:

$$\mathsf{A}_d = \begin{cases} a_{ii}, & \text{if } i = j, \\ 0, & \text{if } i \neq j, \end{cases} \qquad \mathsf{A}_a^+ = \begin{cases} a_{ij}, & \text{if } a_{ij} > 0, \ i \neq j, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathsf{A}_a^- = \mathsf{A}_a - \mathsf{A}_a^+.$$

Definition 3.5 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 $\mathcal{N}_1$ and $\mathcal{N}_2$, if

$$\forall i_0 \in \mathcal{N}_1, \ \exists i_r \in \mathcal{N}_2, \ \exists i_1, \ldots, i_{r-1} \in \mathcal{N} \quad \text{s.t.} \quad a_{i_{k-1} i_k} \neq 0, \quad k = 1, \ldots, r. \tag{3.6}$$

If perceiving A as a directed graph adjacency matrix of vertices labeled by $\mathcal{N}$, then (3.6) 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 $\mathcal{N}_1$ and $\mathcal{N}_2$.

Definition 3.6 Given a square matrix A and a column vector $\vec{x}$, define

$$\mathcal{N}^0(\mathsf{A}\vec{x}) = \{i : (\mathsf{A}\vec{x})_i = 0\} \quad \text{and} \quad \mathcal{N}^+(\mathsf{A}\vec{x}) = \{i : (\mathsf{A}\vec{x})_i > 0\}.$$

The following theorem was proved in [26], see also [21] for a detailed proof.

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

1. *$\mathsf{A}_d + \mathsf{A}^z$ is a nonsingular M-matrix,*

2. *$\mathsf{A}_a^+ \leq \mathsf{A}^z \mathsf{A}_d^{-1} \mathsf{A}^s$ or equivalently $\forall a_{ij} > 0$ with $i \neq j$, $a_{ij} \leq \displaystyle\sum_{k=1}^{n} a_{ik}^z a_{kk}^{-1} a_{kj}^s$,*

3. *$\exists \vec{e} \in \mathbb{R}^n \backslash \{0\}, \vec{e} \geq 0$ with $\mathsf{A}\vec{e} \geq 0$ such that A^z or A^s connects $\mathcal{N}^0(\mathsf{A}\vec{e})$ with $\mathcal{N}^+(\mathsf{A}\vec{e})$.*

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

In the rest of this section, to obtain monotonicity, we will show that the fourth order scheme matrix satisfies the conditions in Theorem 3.7 under suitable mesh and time step constraints.

3.3 The Fourth Order Scheme in One Dimension

In general, the high order finite element methods do not have an M-matrix structure. But it is possible to show that they are products of M-matrices. Next we verify that the Lorenz's condition in Theorem 3.7 can be satisfied for the matrix in the scheme (2.23). For convenience of writing and similar to references [19, 34], we use operator notation. Let $\mathcal{A}$ be the linear operator corresponding the scheme matrix A. The linear operator $\mathcal{A}_d$ (associated with the diagonal matrix A_d) is:

If i is odd, $\quad \mathcal{A}_d(\vec{g}^{\,n+1})_i$

$$= \mathcal{M}_i g_i^{n+1} + \Delta t \frac{D(\mathcal{M}_{i-2} + 4\mathcal{M}_{i-1} + 18\mathcal{M}_i + 4\mathcal{M}_{i+1} + \mathcal{M}_{i+2})}{8h^2} g_i^{n+1};$$

if i is even, $\quad \mathcal{A}_d(\vec{g}^{n+1})_i = \mathcal{M}_i g_i^{n+1} + \Delta t \frac{D(\mathcal{M}_{i-1} + \mathcal{M}_{i+1})}{h^2} g_i^{n+1}.$

Let $a^+ = \max\{a, 0\}$ be the positive part and $a^- = -\min\{a, 0\}$ negative parts of a number a. The operator $\mathcal{A}_a^+$ (associated with the matrix A_a^+) is given by:

If i is odd,

$$\mathcal{A}_a^+(\vec{g}^{n+1})_i = \left(-\frac{\Delta t}{h}\frac{u_{i-2}}{4} + \Delta t \frac{D(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)}{8h^2} \right)^+ g_{i-2}^{n+1}$$
$$+ \left(\frac{\Delta t}{h}\frac{u_{i+2}}{4} + \Delta t \frac{D(3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2})}{8h^2} \right)^+ g_{i+2}^{n+1};$$

if i is even, $\quad \mathcal{A}_a^+(\vec{g}^{n+1})_i = 0.$

By definition of $(\cdot)^+$, it is straightforward to see the matrix A_a^+ is entry-wise non-negative. Let $\mathsf{A}_a^- = \mathsf{A} - \mathsf{A}_d - \mathsf{A}_a^+$ and we further split it by introducing the operator $\mathcal{A}^z$ (associated with the matrix A^z) as follows:

If i is odd,

$$\mathcal{A}^z(\vec{g}^{n+1})_i = -\left(-\frac{\Delta t}{h}u_{i-1} + \Delta t \frac{D(4\mathcal{M}_{i-2} + 12\mathcal{M}_i)}{8h^2} \right.$$
$$-\left(-\frac{\Delta t}{h}\frac{u_{i-2}}{4} + \Delta t \frac{D(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)}{8h^2} \right)^+ \right) g_{i-1}^{n+1}$$
$$-\left(\frac{\Delta t}{h}u_{i+1} + \Delta t \frac{D(12\mathcal{M}_i + 4\mathcal{M}_{i+2})}{8h^2} \right.$$
$$-\left(\frac{\Delta t}{h}\frac{u_{i+2}}{4} + \Delta t \frac{D(3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2})}{8h^2} \right)^+ \right) g_{i+1}^{n+1}$$
$$-\left(-\frac{\Delta t}{h}\frac{u_{i-2}}{4} + \Delta t \frac{D(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)}{8h^2} \right)^- g_{i-2}^{n+1}$$
$$-\left(\frac{\Delta t}{h}\frac{u_{i+2}}{4} + \Delta t \frac{D(3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2})}{8h^2} \right)^- g_{i+2}^{n+1};$$

if i is even, $\quad \mathcal{A}^z(\vec{g}^{n+1})_i = 0.$

Let $\mathsf{A}^s = \mathsf{A}_a^- - \mathsf{A}^z$. The operator $\mathcal{A}^s$ (associated with the matrix A^s) is as follows:

If i is odd,

$$\mathcal{A}^s(\vec{g}^{n+1})_i = -\left(-\frac{\Delta t}{h}\frac{u_{i-2}}{4} + \Delta t \frac{D(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)}{8h^2} \right)^+ g_{i-2}^{n+1}$$
$$-\left(\frac{\Delta t}{h}\frac{u_{i+2}}{4} + \Delta t \frac{D(3\mathcal{M}_i - 4\mathcal{M}_{i+1} + 3\mathcal{M}_{i+2})}{8h^2} \right)^+ g_{i+2}^{n+1};$$

if i is even,

$$\mathcal{A}^s(\vec{g}^{n+1})_i = -\left(-\frac{\Delta t}{h}\frac{u_{i-1}}{2} + \frac{\Delta t}{h^2}\frac{D(3\mathcal{M}_{i-1} + \mathcal{M}_{i+1})}{4} \right) g_{i-1}^{n+1}$$
$$-\left(\frac{\Delta t}{h}\frac{u_{i+1}}{2} + \frac{\Delta t}{h^2}\frac{D(\mathcal{M}_{i-1} + 3\mathcal{M}_{i+1})}{4} \right) g_{i+1}^{n+1}.$$

It is easy to verify that the following mesh constraint is sufficient for $A^z \leq 0$ and $A^s \leq 0$:

$$h \max\{|u_i|, |u_{i+1}|, |u_{i+2}|\} \leq D \min\{\mathcal{M}_i, \mathcal{M}_{i+1}, \mathcal{M}_{i+2}\}, \quad \text{for odd } i. \tag{3.7}$$

The matrix $A_d + A^z$ is an N-by-N real square matrix with positive diagonal entries and nonpositive off-diagonals. We use Lemma 3.1 to verify the first condition in Theorem 3.7. Let D equal to the identity matrix and let $\vec{1} = [1, \cdots, 1]^T$. Then, the row sum of $(A_d + A^z)D$ can be evaluated by $(A_d + A^z)\vec{1}$, namely

If i is odd, the sum of the ith row is: $\mathcal{M}_i + \dfrac{\Delta t}{4h}(-u_{i-2} + 4u_{i-1} - 4u_{i+1} + u_{i+2})$;

if i is even, the sum of the ith row is: $\mathcal{M}_i + \dfrac{\Delta t}{h^2}D(\mathcal{M}_{i-1} + \mathcal{M}_{i+1})$.

To guarantee the positive row sums of the matrix $(A_d + A^z)D$, the following constraints on time step size are sufficient:

$$\mathcal{M}_i + \Delta t \frac{-u_{i-2} + 4u_{i-1} - 4u_{i+1} + u_{i+2}}{4h} > 0, \quad \text{for odd } i. \tag{3.8a}$$

Recall that the velocity field $u(x)$ is incompressible in Model 1, thus $u \equiv C$ in one dimension. So (3.8) is trivially satisfied for positive measure $\mathcal{M}$. Thus $A_d + A^z$ is a nonsingular M-matrix. Meanwhile, the divergence free velocity in one dimension also implies $\frac{u_{i-1}-u_{i+1}}{2h} = 0$, namely

$$\mathcal{M}_i + \Delta t \frac{u_{i-1} - u_{i+1}}{2h} > 0, \quad \text{for even } i. \tag{3.8b}$$

Thus, we have $A\vec{1} > 0$. Therefore, $\mathcal{N}^0(A\vec{1}) = \emptyset$ and the third condition in Theorem 3.7 is trivially satisfied. Our next goal is to seek a sufficient condition such that the second condition in Theorem 3.7 hold. By comparing $\mathcal{A}_a^+(\vec{g}^{n+1})_i$ with $\mathcal{A}^z \circ (A^d)^{-1} \circ \mathcal{A}^s(\vec{g}^{n+1})_i$, it is straightforward to verify that $A_a^+ \leq A^z A_d^{-1} A^s$ is equivalent to the following: for odd i,

$$\left(\mathcal{M}_{i-1} - \frac{\Delta t}{h}\frac{u_{i-2}}{2} + \frac{\Delta t}{h^2}\frac{D(7\mathcal{M}_{i-2} + 5\mathcal{M}_i)}{4}\right)$$
$$\left(-\frac{\Delta t}{h}\frac{u_{i-2}}{4} + \frac{\Delta t}{h^2}\frac{D(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)}{8}\right)$$
$$\leq \left(-\frac{\Delta t}{h}\frac{u_{i-2}}{2} + \frac{\Delta t}{h^2}\frac{D(3\mathcal{M}_{i-2} + \mathcal{M}_i)}{4}\right)\left(-\frac{\Delta t}{h}u_{i-1} + \frac{\Delta t}{h^2}\frac{D(4\mathcal{M}_{i-2} + 12\mathcal{M}_i)}{8}\right). \tag{3.9}$$

Multiply $32\left(\frac{h^2}{\Delta t}\right)^2$ on both side of above inequality, after some manipulation, we get:

$$\left(4\frac{h^2}{\Delta t}\mathcal{M}_{i-1} - 2hu_{i-2} + D(7\mathcal{M}_{i-2} + 5\mathcal{M}_i)\right)$$
$$\left(-2hu_{i-2} + D(3\mathcal{M}_{i-2} - 4\mathcal{M}_{i-1} + 3\mathcal{M}_i)\right)$$
$$\leq \left(-2hu_{i-2} + D(3\mathcal{M}_{i-2} + \mathcal{M}_i)\right)\left(-8hu_{i-1} + D(4\mathcal{M}_{i-2} + 12\mathcal{M}_i)\right).$$

Let $b = \max\{\mathcal{M}_{i-2}, \mathcal{M}_{i-1}, \mathcal{M}_i\}$ and $s = \min\{\mathcal{M}_{i-2}, \mathcal{M}_{i-1}, \mathcal{M}_i\}$, namely, the largest and smallest quadrature point values of $\mathcal{M}$ on an element $[x_{i-2}, x_i]$. Assume the finite difference grid spacing satisfies:

$$h \max\{|u_{i-2}|, |u_{i-1}|, |u_i|\} \leq \frac{1}{4}D \min\{\mathcal{M}_{i-2}, \mathcal{M}_{i-1}, \mathcal{M}_i\}, \quad \text{for odd } i. \tag{3.10}$$

Note, (3.10) implies (3.7). It is easy to verify that a sufficient condition for (3.9) is

$$\left(\frac{1}{4}Ds + 2(3D + \frac{h^2}{\Delta t})b\right)\left(\frac{1}{2}Ds + D(6b - 4s)\right) \leq (2Ds - \frac{1}{4}Ds)(16Ds - 2Ds).$$

Therefore, a sufficient condition is:

$$3D + \frac{h^2}{\Delta t} \leq \frac{49Ds^2}{2b(12b - 7s)} - \frac{Ds}{8b}.$$

Now, we simplify the sufficient condition above. The invariant measure $\mathcal{M} \geq \epsilon_0 > 0$, define $r = b/s$, then above inequality can be rewritten as

$$\frac{h^2}{\Delta t} \leq \frac{49D}{2r(12r - 7)} - \frac{D}{8r} - 3D = \frac{7}{2}D\left(\frac{1}{r - \frac{7}{12}} - \frac{1}{r}\right) - \frac{D}{8r} - 3D.$$

From the definition of r, we know $r \geq 1$. Thus, it is sufficient to employ the conditions $r \in [1, 1.15]$ and

$$\frac{h^2}{\Delta t} \leq 0.02D < \min_{r \in [1,1.15]}\left\{\frac{7}{2}D\left(\frac{1}{r - \frac{7}{12}} - \frac{1}{r}\right) - \frac{D}{8r} - 3D\right\}.$$

This indicates we only need to find a suitable upper bound on h such that $b \leq 1.15s$ (namely $r \in [1, 1.15]$) holds. Recall $\mathcal{M}$ is continuously differentiable. Assume $\mathcal{M}$ take its maximum at point x^* on cell $[x_{i-2}, x_i]$ and $\mathcal{M}$ take its minimum at point x_* on cell $[x_{i-2}, x_i]$. By mean value theorem, there exist a point $\xi \in [x_{i-2}, x_i]$ such that

$$\mathcal{M}(x^*) = \mathcal{M}(x_*) + (x^* - x_*)\mathcal{M}'(\xi).$$

Therefore

$$b \leq \mathcal{M}(x^*) = \mathcal{M}(x_*) + (x^* - x_*)\mathcal{M}'(\xi) \leq s + 2h\max_{[x_{i-2},x_i]}|\mathcal{M}'|,$$

which means in order to let $b \leq 1.15s$ hold, we can employ a sufficient condition as follows

$$s + 2h\max_{[x_{i-2},x_i]}|\mathcal{M}'| \leq 1.15s.$$

To this end, we obtain a constraint on h, as follows

$$h\max_{[x_{i-2},x_i]}|\mathcal{M}'| \leq 0.075\min\{\mathcal{M}_{i-2}, \mathcal{M}_{i-1}, \mathcal{M}_i\}, \quad \text{for odd } i. \tag{3.11}$$

As a summary, we have the following theorem:

Theorem 3.8 *Under the mesh and time step constraints (3.8), (3.10), (3.11) and $\frac{\Delta t}{h^2} \geq \frac{50}{D}$, the coefficient matrix for the unknown vector $\vec{g}^{n+1}$ in the fourth order finite difference scheme (2.23) satisfies the Lorenz's conditions, so it is a product of two M-matrices thus monotone. In particular, in one dimension, for a discrete divergence free velocity field u (which is constant), the matrix in fourth order finite difference scheme (2.23) is monotone under the following constraints: $\frac{\Delta t}{h^2} \geq \frac{50}{D}$ and, for odd i,*

$$h\max\{|u_i|, |u_{i+1}|, |u_{i+2}|\} \leq \frac{1}{4}D\min\{\mathcal{M}_i, \mathcal{M}_{i+1}, \mathcal{M}_{i+2}\},$$

$$h\max_{[x_i,x_{i+2}]}|\mathcal{M}'| \leq 0.075\min\{\mathcal{M}_i, \mathcal{M}_{i+1}, \mathcal{M}_{i+2}\}.$$

Remark 1 In practice, to realize the mesh size and time step satifying the constraints above, one can first choose a small enough mesh size h, then choose a large enough time step Δt. For instance, for a constant velocity case, for a small enough h, one can use $\Delta t \geq \frac{50}{D} h^2$. We emphasize that the sufficient conditions above are not sharp for monotonicity to hold, but with a fixed mesh size h monotonicity will be lost in the fourth order scheme when $\Delta t \to 0$.

3.4 The Fourth Order Scheme in Two Dimension

Next we derive a sufficient mesh size and time step conditions for the two-dimensional fourth order scheme to satisfy the Lorenz's conditions in Theorem 3.7. For convenience, we follow [19, 34], to use operator notation for all matrices.

Similar to the one dimensional discussion above, $\mathcal{A}$ denotes the linear operator for the scheme matrix. The linear operator $\mathcal{A}_d$ (associated with the diagonal matrix A_d) is:

If (x_i, y_j) is a knot, $\quad \mathcal{A}_d(\vec{g}^{n+1})_{ij}$

$$
= \mathcal{M}_{i,j} g_{ij}^{n+1} + \Delta t \, \frac{D(\mathcal{M}_{i-2,j} + 4\mathcal{M}_{i-1,j} + 18\mathcal{M}_{i,j} + 4\mathcal{M}_{i+1,j} + \mathcal{M}_{i+2,j})}{8h^2} g_{ij}^{n+1}
$$
$$
+ \Delta t \, \frac{D(\mathcal{M}_{i,j-2} + 4\mathcal{M}_{i,j-1} + 18\mathcal{M}_{i,j} + 4\mathcal{M}_{i,j+1} + \mathcal{M}_{i,j+2})}{8h^2} g_{ij}^{n+1};
$$

if (x_i, y_j) is an edge (parallel to y-axis) center, $\quad \mathcal{A}_d(\vec{g}^{n+1})_{ij}$

$$
= \mathcal{M}_{i,j} g_{ij}^{n+1} + \Delta t \, \frac{D(\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})}{h^2} g_{ij}^{n+1}
$$
$$
+ \Delta t \, \frac{D(\mathcal{M}_{i-2,j} + 4\mathcal{M}_{i-1,j} + 18\mathcal{M}_{ij} + 4\mathcal{M}_{i+1,j} + \mathcal{M}_{i+2,j})}{8h^2} g_{ij}^{n+1};
$$

if (x_i, y_j) is a cell center, $\quad \mathcal{A}_d(\vec{g}^{n+1})_{ij}$

$$
= \mathcal{M}_{i,j} g_{ij}^{n+1} + \Delta t \, \frac{D(\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j} + \mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})}{h^2} g_{ij}^{n+1}.
$$

For (x_i, y_j) is an edge (parallel to x-axis) center, this case is very similar to the case which (x_i, y_j) is an edge (parallel to y-axis) center, thus omitted. For the sake of brevity, we omit the case (x_i, y_j) is an edge (parallel to x-axis) center when defining operators. Recall that we use notation $(\cdot)^+ = \max\{\cdot, 0\}$ to denote the positive part and $(\cdot)^- = -\min\{\cdot, 0\}$ to denote the negative parts of a number $(\cdot)$. The operator $\mathcal{A}_a^+$ (associated with the matrix A_a^+) is given by:

If (x_i, y_j) is a knot,

$$
\mathcal{A}_a^+(\vec{g}^{n+1})_{ij} = \left(-\frac{\Delta t}{h} \frac{u_{i-2,j}}{4} + \Delta t \, \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2} \right)^+ g_{i-2,j}^{n+1}
$$
$$
+ \left(\frac{\Delta t}{h} \frac{u_{i+2,j}}{4} + \Delta t \, \frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j})}{8h^2} \right)^+ g_{i+2,j}^{n+1}
$$
$$
+ \left(-\frac{\Delta t}{h} \frac{v_{i,j-2}}{4} + \Delta t \, \frac{D(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})}{8h^2} \right)^+ g_{i,j-2}^{n+1}
$$
$$
+ \left(\frac{\Delta t}{h} \frac{v_{i,j+2}}{4} + \Delta t \, \frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j+2})}{8h^2} \right)^+ g_{i,j+2}^{n+1};
$$

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

$$\mathcal{A}_a^+(\vec{g}^{n+1})_{ij} = \left(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+ g_{i-2,j}^{n+1}$$

$$+ \left(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+ g_{i+2,j}^{n+1};$$

if (x_i, y_j) is a cell center, $\mathcal{A}_a^+(\vec{g}^{n+1})_{ij} = 0.$

It is straightforward to see the matrix A_a^+ is entry-wise non-negative. Let $A_a^- = A - A_d - A_a^+$ and we further split it by introducing the operator $\mathcal{A}^z$ (associated with the matrix A^z) as follows:

If (x_i, y_j) is a knot,

$$\mathcal{A}^z(\vec{g}^{n+1})_{ij} = -\left(-\frac{\Delta t}{h}u_{i-1,j} + \Delta t\,\frac{D(4\mathcal{M}_{i-2,j} + 12\mathcal{M}_{i,j})}{8h^2}\right.$$

$$-\left(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+\left.\right) g_{i-1,j}^{n+1}$$

$$-\left(\frac{\Delta t}{h}u_{i+1,j} + \Delta t\,\frac{D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i+2,j})}{8h^2} - \right.$$

$$\left(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j})}{8h^2}\right)^+\left.\right) g_{i+1,j}^{n+1}$$

$$-\left(-\frac{\Delta t}{h}v_{i,j-1} + \Delta t\,\frac{D(4\mathcal{M}_{i,j-2} + 12\mathcal{M}_{i,j})}{8h^2} - \right.$$

$$\left(-\frac{\Delta t}{h}\frac{v_{i,j-2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+\left.\right) g_{i,j-1}^{n+1}$$

$$-\left(\frac{\Delta t}{h}v_{i,j+1} + \Delta t\,\frac{D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i,j+2})}{8h^2} - \right.$$

$$\left(\frac{\Delta t}{h}\frac{v_{i,j+2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j+2})}{8h^2}\right)^+\left.\right) g_{i,j+1}^{n+1}$$

$$-\left(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2}\right)^- g_{i-2,j}^{n+1}$$

$$-\left(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j})}{8h^2}\right)^- g_{i+2,j}^{n+1}$$

$$-\left(-\frac{\Delta t}{h}\frac{v_{i,j-2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})}{8h^2}\right)^- g_{i,j-2}^{n+1}$$

$$-\left(\frac{\Delta t}{h}\frac{v_{i,j+2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j+2})}{8h^2}\right)^- g_{i,j+2}^{n+1};$$

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

$$\mathcal{A}^z(\vec{g}^{n+1})_{ij} = -\left(-\frac{\Delta t}{h}u_{i-1,j} + \Delta t\,\frac{D(4\mathcal{M}_{i-2,j} + 12\mathcal{M}_{i,j})}{8h^2} - \right.$$

$$\left(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+\left.\right) g_{i-1,j}^{n+1}$$

$$-\left(\frac{\Delta t}{h}u_{i+1,j} + \Delta t\,\frac{D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i+2,j})}{8h^2} - \right.$$

$$\left(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j})}{8h^2}\right)^+\right) g_{i+1,j}^{n+1}$$

$$-\left(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2}\right)^- g_{i-2,j}^{n+1}$$

$$-\left(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})}{8h^2}\right)^- g_{i+2,j}^{n+1};$$

If (x_i, y_j) is a cell center, $\mathcal{A}^z(\vec{g}^{n+1})_{ij} = 0$.

Similar to (3.7) in subsection 3.3, it is easy to verify that $\mathsf{A}^z \le 0$ under the following sufficient condition: for odd i and odd j,

$$h\max\{|u_{i,j}|, |u_{i+1,j}|, |u_{i+2,j}|\} \le D\min\{\mathcal{M}_{i,j}, \mathcal{M}_{i+1,j}, \mathcal{M}_{i+2,j}\}, \tag{3.12a}$$

$$\text{and} \quad h\max\{|v_{i,j}|, |v_{i,j+1}|, |v_{i,j+2}|\} \le D\min\{\mathcal{M}_{i,j}, \mathcal{M}_{i,j+1}, \mathcal{M}_{i,j+2}\}. \tag{3.12b}$$

The matrix $\mathsf{A}^s = \mathsf{A}_a^- - \mathsf{A}^z$. Therefore, the operator $\mathcal{A}^s$ (associated with the matrix A^s) is as follows:

If (x_i, y_j) is a knot,

$$\mathcal{A}^s(\vec{g}^{n+1})_{ij} = -\left(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+ g_{i-1,j}^{n+1}$$

$$-\left(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j})}{8h^2}\right)^+ g_{i+1,j}^{n+1}$$

$$-\left(-\frac{\Delta t}{h}\frac{v_{i,j-2}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+ g_{i,j-1}^{n+1}$$

$$-\left(\frac{\Delta t}{h}\frac{v_{i,j+2}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j+2})}{8h^2}\right)^+ g_{i,j+1}^{n+1};$$

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

$$\mathcal{A}^s(\vec{g}^{n+1})_{ij} = -\left(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+ g_{i-1,j}^{n+1}$$

$$-\left(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j})}{8h^2}\right)^+ g_{i+1,j}^{n+1}$$

$$-\left(-\frac{\Delta t}{h}\frac{v_{i,j-1}}{2} + \Delta t\, \frac{D(3\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})}{4h^2}\right) g_{i,j-1}^{n+1}$$

$$-\left(\frac{\Delta t}{h}\frac{v_{i,j+1}}{2} + \Delta t\, \frac{D(\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j+1})}{4h^2}\right) g_{i,j+1}^{n+1};$$

if (x_i, y_j) is a cell center,

$$\mathcal{A}^s(\vec{g}^{n+1})_{ij} = -\left(-\frac{\Delta t}{h}\frac{u_{i-1,j}}{2} + \Delta t\, \frac{D(3\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j})}{4h^2}\right) g_{i-1,j}^{n+1}$$

$$-\left(\frac{\Delta t}{h}\frac{u_{i+1,j}}{2} + \Delta t\, \frac{D(\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i+1,j})}{4h^2}\right) g_{i+1,j}^{n+1}$$

$$-\left(-\frac{\Delta t}{h}\frac{v_{i,j-1}}{2} + \Delta t\, \frac{D(3\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})}{4h^2}\right) g_{i,j-1}^{n+1}$$

$$-\left(\frac{\Delta t}{h}\frac{v_{i,j+1}}{2} + \Delta t\, \frac{D(\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j+1})}{4h^2}\right) g_{i,j+1}^{n+1}.$$

Obviously, under the sufficient condition (3.12), $A^s \le 0$ also holds. The matrix $A_d + A^z$ is a real squared matrix with positive diagonal entries and nonpositive off-diagonals. We use Lemma 3.1 to verify the first condition in Theorem 3.7. Let D equal to the identity matrix and applying the same argument as in subsection 3.3. Notice that the row sums of the matrix $A_d + A^z$ are the outputs of $[\mathcal{A}_d + \mathcal{A}^z](\vec{1})_{ij}$:

If (x_i, y_j) is a knot, $\quad \mathcal{M}_{i,j} + \Delta t \dfrac{-u_{i-2,j} + 4u_{i-1,j} - 4u_{i+1,j} + u_{i+2,j}}{4h}$

$\qquad + \Delta t \dfrac{-v_{i,j-2} + 4v_{i,j-1} - 4v_{i,j+1} + v_{i,j+2}}{4h};$

if (x_i, y_j) is an edge (parallel to y-axis) center, $\quad \mathcal{M}_{i,j} + \Delta t \dfrac{D(\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})}{h^2}$

$\qquad + \Delta t \dfrac{-u_{i-2,j} + 4u_{i-1,j} - 4u_{i+1,j} + u_{i+2,j}}{4h};$

if (x_i, y_j) is an edge (parallel to x-axis) center, $\quad \mathcal{M}_{i,j} + \Delta t \dfrac{D(\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j})}{h^2}$

$\qquad + \Delta t \dfrac{-v_{i,j-2} + 4v_{i,j-1} - 4v_{i,j+1} + v_{i,j+2}}{4h};$

if (x_i, y_j) is a cell center, $\quad \mathcal{M}_{i,j}$

$\qquad + \Delta t \dfrac{D(\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j} + \mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})}{h^2}.$

To guarantee the positive row sums, the following constraints on time step size are sufficient:

For odd i and odd j,

$$\mathcal{M}_{i,j} + \Delta t \dfrac{-u_{i-2,j} + 4u_{i-1,j} - 4u_{i+1,j} + u_{i+2,j}}{4h}$$
$$+ \Delta t \dfrac{-v_{i,j-2} + 4v_{i,j-1} - 4v_{i,j+1} + v_{i,j+2}}{4h} > 0; \tag{3.13a}$$

for odd i and even j,

$$\mathcal{M}_{i,j} + \Delta t \dfrac{D(\mathcal{M}_{i,j-1} + \mathcal{M}_{i,j+1})}{h^2} + \Delta t \dfrac{-u_{i-2,j} + 4u_{i-1,j} - 4u_{i+1,j} + u_{i+2,j}}{4h} > 0; \tag{3.13b}$$

for even i and odd j,

$$\mathcal{M}_{i,j} + \Delta t \dfrac{D(\mathcal{M}_{i-1,j} + \mathcal{M}_{i+1,j})}{h^2} + \Delta t \dfrac{-v_{i,j-2} + 4v_{i,j-1} - 4v_{i,j+1} + v_{i,j+2}}{4h} > 0. \tag{3.13c}$$

Recall that the velocity field $\vec{u}$ is incompressible in Model 1, thus one can preprocess the given $\vec{u}$ such that the velocity point values satisfy the following discrete divergence free constraint:

For odd i and odd j,

$$\dfrac{-u_{i-2,j} + 4u_{i-1,j} - 4u_{i+1,j} + u_{i+2,j}}{4h} + \dfrac{-v_{i,j-2} + 4v_{i,j-1} - 4v_{i,j+1} + v_{i,j+2}}{4h} = 0; \tag{3.14a}$$

for odd i and even j,

$$\frac{-u_{i-2,j} + 4u_{i-1,j} - 4u_{i+1,j} + u_{i+2,j}}{4h} + \frac{v_{i,j-1} - v_{i,j+1}}{2h} = 0; \tag{3.14b}$$

for even i and odd j,

$$\frac{u_{i-1,j} - u_{i+1,j}}{2h} + \frac{-v_{i,j-2} + 4v_{i,j-1} - 4v_{i,j+1} + v_{i,j+2}}{4h} = 0; \tag{3.14c}$$

for even i and even j,

$$\frac{u_{i-1,j} - u_{i+1,j}}{2h} + \frac{v_{i,j-1} - v_{i,j+1}}{2h} = 0. \tag{3.14d}$$

Then, for any incompressible velocity satisfying the discrete divergence free constraint (3.14), we know the (3.13) is satisfied for positive measure $\mathcal{M}$ under the following sufficient condition:

$$h \max\{|v_{i,j-1}|, |v_{i,j+1}|\} \le 2D \min\{\mathcal{M}_{i,j-1}, \mathcal{M}_{i,j+1}\} \quad \text{for odd } i \text{ and even } j \tag{3.15a}$$

$$h \max\{|u_{i-1,j}|, |u_{i+1,j}|\} \le 2D \min\{\mathcal{M}_{i-1,j}, \mathcal{M}_{i+1,j}\} \quad \text{for even } i \text{ and odd } j. \tag{3.15b}$$

Notice that (3.12) implies (3.15). Thus, under the condition (3.13) (in particular, under the condition (3.12) for a discrete divergence free velocity field), the matrix $A_d + A^z$ is a nonsingular M-matrix. Meanwhile, under the same sufficient condition, we have $A\vec{1} > 0$, which indicates $\mathcal{N}^0(A\vec{1}) = \emptyset$, namely, the third condition in Theorem 3.7 is trivially satisfied.

Finally, to verify $A_a^+ \le A^z A_d^{-1} A^s$ in Theorem 3.7, we only need to compare the outputs of $\mathcal{A}_a^+(\vec{g}^{n+1})$ with $\mathcal{A}^z \circ (\mathcal{A}_d)^{-1} \circ \mathcal{A}^s(\vec{g}^{n+1})$. If x_{ij} is a knot, we only need the following inequalities hold:

- For the entry in A_a^+ associated with the coefficient of $g_{i-2,j}^{n+1}$ in $\mathcal{A}_a^+(\vec{g}^{n+1})_{ij}$.

$$\left(-\frac{\Delta t}{h} \frac{u_{i-2,j}}{4} + \Delta t \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2} \right)^+$$
$$\times \left(\mathcal{M}_{i-1,j} + \Delta t \frac{D(\mathcal{M}_{i-2,j} + \mathcal{M}_{i,j})}{h^2} \right.$$
$$\left. + \Delta t \frac{D(\mathcal{M}_{i-1,j-2} + 4\mathcal{M}_{i-1,j-1} + 18\mathcal{M}_{i-1,j} + 4\mathcal{M}_{i-1,j+1} + \mathcal{M}_{i-1,j+2})}{8h^2} \right)$$
$$\le \left(-\frac{\Delta t}{h} u_{i-1,j} + \Delta t \frac{D(4\mathcal{M}_{i-2,j} + 12\mathcal{M}_{i,j})}{8h^2} \right.$$
$$\left. - \left(-\frac{\Delta t}{h} \frac{u_{i-2,j}}{4} + \Delta t \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2} \right)^+ \right)$$
$$\times \left(-\frac{\Delta t}{h} \frac{u_{i-2,j}}{2} + \Delta t \frac{D(3\mathcal{M}_{i-2,j} + \mathcal{M}_{i,j})}{4h^2} \right) \tag{3.16}$$

- For the entry in A_a^+ associated with the coefficient of $g_{i+2,j}^{n+1}$ in $\mathcal{A}_a^+(\vec{g}^{n+1})_{ij}$.

$$\left(\frac{\Delta t}{h} \frac{u_{i+2,j}}{4} + \Delta t \frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j})}{8h^2} \right)^+$$
$$\times \left(\mathcal{M}_{i+1,j} + \Delta t \frac{D(\mathcal{M}_{i,j} + \mathcal{M}_{i+2,j})}{h^2} \right.$$
$$\left. + \Delta t \frac{D(\mathcal{M}_{i+1,j-2} + 4\mathcal{M}_{i+1,j-1} + 18\mathcal{M}_{i+1,j} + 4\mathcal{M}_{i+1,j+1} + \mathcal{M}_{i+1,j+2})}{8h^2} \right)$$
$$\le \left(\frac{\Delta t}{h} u_{i+1,j} + \Delta t \frac{D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i+2,j})}{8h^2} \right.$$

$$
-\left(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j})}{8h^2}\right)^{+}\right)
$$

$$
\times\left(\frac{\Delta t}{h}\frac{u_{i+2,j}}{2} + \Delta t\,\frac{D(\mathcal{M}_{i,j} + 3\mathcal{M}_{i+2,j})}{4h^2}\right)
\tag{3.17}
$$

- For the entry in A_a^{+} associated with the coefficient of $g_{i,j-2}^{n+1}$ in $\mathcal{A}_a^{+}(\vec{g}^{\,n+1})_{ij}$.

$$
\left(-\frac{\Delta t}{h}\frac{v_{i,j-2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})}{8h^2}\right)^{+}
$$

$$
\times\left(\mathcal{M}_{i,j-1} + \Delta t\,\frac{D(\mathcal{M}_{i,j-2} + \mathcal{M}_{i,j})}{h^2}\right.
$$

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

$$
\leq\left(-\frac{\Delta t}{h}v_{i,j-1} + \Delta t\,\frac{D(4\mathcal{M}_{i,j-2} + 12\mathcal{M}_{i,j})}{8h^2}\right.
$$

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

$$
\times\left(-\frac{\Delta t}{h}\frac{v_{i,j-2}}{2} + \Delta t\,\frac{D(3\mathcal{M}_{i,j-2} + \mathcal{M}_{i,j})}{4h^2}\right)
\tag{3.18}
$$

- For the entry in A_a^{+} associated with the coefficient of $g_{i,j+2}^{n+1}$ in $\mathcal{A}_a^{+}(\vec{g}^{\,n+1})_{ij}$.

$$
\left(\frac{\Delta t}{h}\frac{v_{i,j+2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j+2})}{8h^2}\right)^{+}
$$

$$
\times\left(\mathcal{M}_{i,j+1} + \Delta t\,\frac{D(\mathcal{M}_{i,j} + \mathcal{M}_{i,j+2})}{h^2}\right.
$$

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

$$
\leq\left(\frac{\Delta t}{h}v_{i,j+1} + \Delta t\,\frac{D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i,j+2})}{8h^2}\right.
$$

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

$$
\times\left(\frac{\Delta t}{h}\frac{v_{i,j+2}}{2} + \Delta t\,\frac{D(\mathcal{M}_{i,j} + 3\mathcal{M}_{i,j+2})}{4h^2}\right)
\tag{3.19}
$$

The above inequalities hold trivially, if the positive part in each inequalities is zero. For seeking a sufficient condition, we only need to consider the case that the positive parts are larger than zero. Let us use (3.16) as an example to derive a sufficient condition. The (3.17)–(3.19) are processed in the same way. Multiply $64(\frac{h^2}{\Delta t})^2$ on both side, after some manipulation, we have:

$$
\left(8\frac{h^2}{\Delta t}\mathcal{M}_{i-1,j} + 8D(\mathcal{M}_{i-2,j} + \mathcal{M}_{i,j})\right.
$$

$$
\left. + D(\mathcal{M}_{i-1,j-2} + 4\mathcal{M}_{i-1,j-1} + 18\mathcal{M}_{i-1,j} + 4\mathcal{M}_{i-1,j+1} + \mathcal{M}_{i-1,j+2})\right)
$$

$$
\times\left(-2hu_{i-2,j} + D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})\right)
$$

$$\leq \Big(-4hu_{i-2,j} + 2D(3\mathcal{M}_{i-2,j} + \mathcal{M}_{i,j}) \Big)$$
$$\times \Big(2hu_{i-2,j} - 8hu_{i-1,j} + D(\mathcal{M}_{i-2,j} + 4\mathcal{M}_{i-1,j} + 9\mathcal{M}_{i,j}) \Big).$$

Let $E_\Delta = [i-2, i+2] \times [j-2, j+2]$. Denote the largest and smallest values of the invariant measure on E_Δ by $b = \max_{E_\Delta}\{\mathcal{M}_{i,j}\}$ and $s = \min_{E_\Delta}\{\mathcal{M}_{i,j}\}$. Assume the finite difference grid spacing satisfies:

$$h \max_{E_\Delta} |\vec{u}_{i,j}| \leq \frac{1}{20} D \min_{E_\Delta}\{\mathcal{M}_{i,j}\}. \tag{3.20}$$

Note that (3.20) implies the condition (3.12). Then we only need

$$4\Big(11D + 2\frac{h^2}{\Delta t}\Big)b\Big(\frac{1}{10}Ds + D(6b - 4s)\Big) \leq (8Ds - \frac{1}{5}Ds)(14Ds - \frac{1}{2}Ds).$$

Therefore, a sufficient condition is:

$$11D + \frac{h^2}{\Delta t} \leq \frac{52.65Ds^2}{b(12b - 7.8s)}. \tag{3.21}$$

If x_{ij} is an edge center (either parallel to y-axis or parallel to x-axis), we only need the following inequalities hold:

- For the entry in A_a^+ associated with the coefficient of $g_{i-2,j}^{n+1}$ in $\mathcal{A}_a^+(\vec{g}^{\,n+1})_{ij}$.

$$\Big(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2} \Big)^+$$
$$\times \Big(\mathcal{M}_{i-1,j} + \Delta t\, \frac{D(\mathcal{M}_{i-2,j} + \mathcal{M}_{i,j} + \mathcal{M}_{i-1,j-1} + \mathcal{M}_{i-1,j+1})}{h^2} \Big)$$
$$\leq \Big(-\frac{\Delta t}{h}u_{i-1,j} + \Delta t\, \frac{D(4\mathcal{M}_{i-2,j} + 12\mathcal{M}_{i,j})}{8h^2}$$
$$-\Big(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})}{8h^2} \Big)^+ \Big)$$
$$\times \Big(-\frac{\Delta t}{h}\frac{u_{i-2,j}}{2} + \Delta t\, \frac{D(3\mathcal{M}_{i-2,j} + \mathcal{M}_{i,j})}{4h^2} \Big) \tag{3.22}$$

- For the entry in A_a^+ associated with the coefficient of $g_{i+2,j}^{n+1}$ in $\mathcal{A}_a^+(\vec{g}^{\,n+1})_{ij}$.

$$\Big(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i+2,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i,j})}{8h^2} \Big)^+$$
$$\times \Big(\mathcal{M}_{i+1,j} + \Delta t\, \frac{D(\mathcal{M}_{i,j} + \mathcal{M}_{i+2,j} + \mathcal{M}_{i+1,j-1} + \mathcal{M}_{i+1,j+1})}{h^2} \Big)$$
$$\leq \Big(\frac{\Delta t}{h}u_{i+1,j} + \Delta t\, \frac{D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i+2,j})}{8h^2}$$
$$-\Big(\frac{\Delta t}{h}\frac{u_{i+2,j}}{4} + \Delta t\, \frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i+1,j} + 3\mathcal{M}_{i+2,j})}{8h^2} \Big)^+ \Big)$$
$$\times \Big(\frac{\Delta t}{h}\frac{u_{i+2,j}}{2} + \Delta t\, \frac{D(\mathcal{M}_{i,j} + 3\mathcal{M}_{i+2,j})}{4h^2} \Big) \tag{3.23}$$

- For the entry in A_a^+ associated with the coefficient of $g_{i,j-2}^{n+1}$ in $A_a^+(\vec{g}^{\,n+1})_{ij}$.

$$\left(-\frac{\Delta t}{h}\frac{v_{i,j-2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+$$
$$\times\left(\mathcal{M}_{i,j-1} + \Delta t\,\frac{D(\mathcal{M}_{i-1,j-1} + \mathcal{M}_{i+1,j-1} + \mathcal{M}_{i,j-2} + \mathcal{M}_{i,j})}{h^2}\right)$$
$$\le\left(-\frac{\Delta t}{h}v_{i,j-1} + \Delta t\,\frac{D(4\mathcal{M}_{i,j-2} + 12\mathcal{M}_{i,j})}{8h^2}\right.$$
$$-\left(-\frac{\Delta t}{h}\frac{v_{i,j-2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j-2} - 4\mathcal{M}_{i,j-1} + 3\mathcal{M}_{i,j})}{8h^2}\right)^+\right)$$
$$\times\left(-\frac{\Delta t}{h}\frac{v_{i,j-2}}{2} + \Delta t\,\frac{D(3\mathcal{M}_{i,j-2} + \mathcal{M}_{i,j})}{4h^2}\right) \tag{3.24}$$

- For the entry in A_a^+ associated with the coefficient of $g_{i,j+2}^{n+1}$ in $A_a^+(\vec{g}^{\,n+1})_{ij}$.

$$\left(\frac{\Delta t}{h}\frac{v_{i,j+2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j+2})}{8h^2}\right)^+$$
$$\times\left(\mathcal{M}_{i,j+1} + \Delta t\,\frac{D(\mathcal{M}_{i-1,j+1} + \mathcal{M}_{i+1,j+1} + \mathcal{M}_{i,j} + \mathcal{M}_{i,j+2})}{h^2}\right)$$
$$\le\left(\frac{\Delta t}{h}v_{i,j+1} + \Delta t\,\frac{D(12\mathcal{M}_{i,j} + 4\mathcal{M}_{i,j+2})}{8h^2}\right.$$
$$-\left(\frac{\Delta t}{h}\frac{v_{i,j+2}}{4} + \Delta t\,\frac{D(3\mathcal{M}_{i,j} - 4\mathcal{M}_{i,j+1} + 3\mathcal{M}_{i,j+2})}{8h^2}\right)^+\right)$$
$$\times\left(\frac{\Delta t}{h}\frac{v_{i,j+2}}{2} + \Delta t\,\frac{D(\mathcal{M}_{i,j} + 3\mathcal{M}_{i,j+2})}{4h^2}\right) \tag{3.25}$$

Again, we only need to consider the case that the positive parts in above are larger than zero. Let us use (3.22) as an example to derive a sufficient condition. The (3.23)–(3.25) are processed in the same way. Multiply $64(\frac{h^2}{\Delta t})^2$ on both side, after some manipulation, we have:

$$\left(4\frac{h^2}{\Delta t}\mathcal{M}_{i-1,j} + 4D(\mathcal{M}_{i-2,j} + \mathcal{M}_{i,j} + \mathcal{M}_{i-1,j-1} + \mathcal{M}_{i-1,j+1})\right)$$
$$\times\left(-2hu_{i-2,j} + D(3\mathcal{M}_{i-2,j} - 4\mathcal{M}_{i-1,j} + 3\mathcal{M}_{i,j})\right) \le \left(-2hu_{i-2,j} + D(3\mathcal{M}_{i-2,j} + \mathcal{M}_{i,j})\right)$$
$$\times\left(2hu_{i-2,j} - 8hu_{i-1,j} + D(\mathcal{M}_{i-2,j} + 4\mathcal{M}_{i-1,j} + 9\mathcal{M}_{i,j})\right).$$

Recall that $E_\Delta = [i-2, i+2]\times[j-2, j+2]$ and $b = \max_{E_\Delta}\{\mathcal{M}_{i,j}\}$ and $s = \min_{E_\Delta}\{\mathcal{M}_{i,j}\}$. Assume the finite difference grid spacing satisfies (3.20). Then we only need

$$4\left(8D + 2\frac{h^2}{\Delta t}\right)b\left(\frac{1}{10}Ds + D(6b - 4s)\right) \le (8Ds - \frac{1}{5}Ds)(14Ds - \frac{1}{2}Ds).$$

Therefore, (3.21) still serves as a sufficient condition. Now, let us try to simplify above sufficient condition. The invariant measure $\mathcal{M} \ge \epsilon_0 > 0$, define $r = b/s$, then (3.21) can be rewritten as

$$\frac{h^2}{\Delta t} \le \frac{52.65D}{r(12r - 7.8)} - 11D = 6.75D\left(\frac{1}{r - 0.65} - \frac{1}{r}\right) - 11D.$$

From the definition of r, we know $r \geq 1$. Thus, it is sufficient to employ the conditions $r \in [1, 1.025]$ and

$$\frac{h^2}{\Delta t} \leq \sqrt{2}D < \min_{r \in [1, 1.025]} \left\{ 6.75D\left(\frac{1}{r - 0.65} - \frac{1}{r}\right) - 11D \right\}.$$

This indicates we only need to find a suitable upper bound on h such that $b \leq 1.025s$ (namely $r \in [1, 1.025]$) holds. Recall $\mathcal{M}$ is continuously differentiable. Assume $\mathcal{M}$ take its maximum at point $\vec{x}^*$ on E_Δ and $\mathcal{M}$ take its minimum at point $\vec{x}_*$ on E_Δ. By mean value theorem, there exist a point $\vec{\xi} \in E_\Delta$ such that

$$\mathcal{M}(\vec{x}^*) = \mathcal{M}(\vec{x}_*) + (\vec{x}^* - \vec{x}_*) \cdot \nabla \mathcal{M}(\vec{\xi}).$$

Therefore

$$b \leq s + 4\sqrt{2}\, h \max_{E_\Delta} |(\mathcal{M}'_x, \mathcal{M}'_y)|,$$

which means in order to let $b \leq 1.025s$ hold, we can employ a sufficient condition as follows

$$s + 4\sqrt{2}\, h \max_{E_\Delta} |(\mathcal{M}'_x, \mathcal{M}'_y)| \leq 1.025s. \tag{3.26}$$

To this end, we obtain a constraint on h, as follows

$$h \max_{E_\Delta} |(\mathcal{M}'_x, \mathcal{M}'_y)| \leq \frac{\sqrt{2}}{320} \min_{E_\Delta} \{\mathcal{M}_{i,j}\}. \tag{3.27}$$

As a summary, we have the following theorem:

Theorem 3.9 *Under the mesh and time step constraints (3.13), (3.20), (3.27) and $\frac{\Delta t}{h^2} \geq \frac{1}{\sqrt{2}D}$, the coefficient matrix for the unknown vector $\vec{g}$ in the fourth order finite difference scheme (2.24) satisfies the Lorenz's conditions, so it is a product of two M-matrices thus monotone. In particular, let $E_\Delta = [i-2, i+2] \times [j-2, j+2]$, with a two dimensional discrete divergence free velocity field satisfying (3.14), then the matrix in fourth order finite difference scheme (2.24) is monotone under the following constraints:*

$$h \max_{E_\Delta} |\vec{u}_{i,j}| \leq \frac{1}{20}D \min_{E_\Delta}\{\mathcal{M}_{i,j}\},$$

$$h \max_{E_\Delta} |(\mathcal{M}'_x, \mathcal{M}'_y)| \leq \frac{\sqrt{2}}{320} \min_{E_\Delta}\{\mathcal{M}_{i,j}\},$$

$$\frac{\Delta t}{h^2} \geq \frac{1}{\sqrt{2}D}.$$

Remark 2 We emphasize that the conditions above are only convenient sufficient conditions for monotonicity, rather than sharp necessary conditions. However, the monotonicity in the fourth order finite difference scheme (2.24) will be lost in numerical tests if Δt approaches 0. So certain lower bound on $\frac{\Delta t}{h^2}$ is a necessary condition for monotonicity.

4 Properties of the Fully Discrete Numerical Schemes

We only discuss the two dimensional case since all the results can be easily reduced to the one dimensional case. The discussion in this section holds for both the second order scheme (2.22) and the fourth order scheme (2.24). For convenience, we use $\vec{g}_h$ to denote

the numerical solution vector in two dimensions with entries g_i ($i = 1, \cdots, N^2$). The finite element space V^h is N^2-dimensional with Lagrangian Q^k basis $\{\phi_i(\mathbf{x})\}_{i=1}^{N^2}$ defined at the $(k+1) \times (k+1)$-point Gauss-Lobatto points.

4.1 Natural Properties of the Finite Element Method

Since the finite difference schemes in Sect. 2 are derived from a finite element method, they inherit many good properties from the original finite element method, which will be used for deriving energy dissipation. We can express the numerical scheme (2.20) in a matrix–vector form. We introduce the following matrices:

$$[\mathsf{W}]_{i,j} = \langle \phi_j, \phi_i \rangle, \qquad\qquad [\mathsf{M}]_{i,j} = \text{diag}(\mathcal{M}_1, \cdots, \mathcal{M}_{N^2}),$$
$$[\mathsf{A}^{\text{diff}}]_{i,j} = \langle D\mathcal{M}\nabla\phi_j, \nabla\phi_i \rangle, \qquad [\mathsf{A}^{\text{adv}}]_{i,j} = \langle u\phi_j, \nabla\phi_i \rangle.$$

Since we use the Gauss-Lobatto quadrature, the lumped mass matrix W is a diagonal matrix, with quadrature weights $\omega_i > 0$ on the diagonal. Then the matrix–vector form of (2.20) is

$$(\mathsf{WM} + \Delta t \mathsf{A}^{\text{diff}} + \Delta t \mathsf{A}^{\text{adv}})\vec{g}_h^{n+1} = \mathsf{WM}\vec{g}_h^n. \tag{4.1}$$

For simplicity, we define

$$\mathsf{A} := \mathsf{WM} + \Delta t \mathsf{A}^{\text{diff}} + \Delta t \mathsf{A}^{\text{adv}},$$

$$\mathsf{B} := \mathsf{M}^{-1}\mathsf{W}^{-1}\mathsf{A} = \mathsf{I} + \Delta t \mathsf{M}^{-1}\mathsf{W}^{-1}\mathsf{A}^{\text{diff}} + \Delta t \mathsf{M}^{-1}\mathsf{W}^{-1}\mathsf{A}^{\text{adv}}.$$

Thus $\vec{g}_h^{n+1} = \mathsf{A}^{-1}\mathsf{WM}\vec{g}_h^n = \mathsf{B}^{-1}\vec{g}_h^n$.

Consider an arbitrary test function $\phi_h \in V^h$ with point values ϕ_i ($i = 1, \cdots N^2$). Let $\vec{\phi}_h$ be the vector with entries ϕ_i. Then the scheme (2.20) is equivalent to the following matrix–vector form:

$$\vec{\phi}_h^T \mathsf{A}\vec{g}_h^{n+1} = \vec{\phi}_h^T \mathsf{WM}\vec{g}_h^n, \quad \forall \vec{\phi}_h \in \mathbb{R}^{N^2}.$$

By considering the test function $\phi_h \equiv 1$, we get

$$\forall g_h \in V^h, \quad \langle u g_h, \nabla 1 \rangle = 0 \implies \vec{1}^T \mathsf{A}^{\text{adv}}\vec{g}_h = 0, \forall \vec{g}_h \in \mathbb{R}^{N^2} \implies \vec{1}^T \mathsf{A}^{\text{adv}} = 0,$$
$$\forall g_h \in V^h, \quad \langle D\mathcal{M}\nabla g_h, \nabla 1 \rangle = 0 \implies \vec{1}^T \mathsf{A}^{\text{diff}}\vec{g}_h = 0, \forall \vec{g}_h \in \mathbb{R}^{N^2} \implies \vec{1}^T \mathsf{A}^{\text{diff}} = \mp$$

Thus we have

$$\vec{1}^T \mathsf{A} = \vec{1}^T (\mathsf{WM} + \Delta t \mathsf{A}^{\text{diff}} + \Delta t \mathsf{A}^{\text{adv}}) = \vec{1}^T \mathsf{WM}. \tag{4.2}$$

The next natural property of the finite element method (2.20) is $\mathsf{M}^{-1}\mathsf{W}^{-1}\mathsf{A}\vec{1} = \vec{1}$, under the assumption that the velocity field satisfies the following discrete divergence free constraint:

$$\forall \phi_h \in V^h, \quad \langle u, \nabla \phi_h \rangle = 0. \tag{4.3}$$

It is straightforward to verify that (4.3) is equivalent to (3.5) in the second order scheme, and equivalent to (3.14) in the fourth order scheme.

Notice that we first have

$$g_h \equiv 1 \implies \langle D\mathcal{M}\nabla g_h, \nabla \phi_h \rangle = 0, \forall \phi_h \in V^h \implies \vec{\phi}_h^T \mathsf{A}^{\text{diff}}\vec{1} = 0, \forall \vec{\phi}_h \in \mathbb{R}^{N^2} \implies \mathsf{A}^{\text{diff}}\vec{1} = \vec{0}.$$

With the discrete divergence free condition (4.3), we have

$$g_h \equiv 1 \Longrightarrow \langle ug_h, \nabla \phi_h \rangle = 0, \forall \phi_h \in V^h \Longrightarrow \vec{\phi}_h^T A^{\text{adv}} \vec{1} = 0, \forall \vec{\phi}_h \in \mathbb{R}^{N^2} \Longrightarrow A^{\text{adv}} \vec{1} = \vec{0}.$$

Therefore

$$M^{-1}W^{-1}A\vec{1} = (I + \Delta t M^{-1}W^{-1}A^{\text{diff}} + \Delta t M^{-1}W^{-1}A^{\text{adv}})\vec{1} = \vec{1}. \tag{4.4}$$

4.2 Mass Conservation

By plugging in the test function $\phi_h \equiv 1$ in (2.20), we get $\langle \mathcal{M}g_h^{n+1}, 1 \rangle = \langle \mathcal{M}g_h^n, 1 \rangle$, thus

$$\langle \rho_h^{n+1}, 1 \rangle = \langle \rho_h^n, 1 \rangle,$$

which can also be written as

$$\sum_{i=1}^{N^2} \omega_i \rho_i^{n+1} = \sum_{i=1}^{N^2} \omega_i \rho_i^n. \tag{4.5}$$

4.3 Steady State Preserving

If $\vec{g}_h^n = K\vec{1}$ for some constant K, then multiply $M^{-1}W^{-1}$ on both side of (4.1), we have

$$M^{-1}W^{-1}A\vec{g}_h^{n+1} = \vec{g}_h^n. \tag{4.6}$$

It is a well known fact that the stiffness matrix A in the finite element method (2.20) is nonsingular, which is implied by the coercivity of the bilinear form in (2.11) for an incompressible velocity field. When the linear system above is nonsingular, it is straightforward to verify that the unique solution is $\vec{g}_h^{n+1} = K\vec{1}$. Therefore, in terms of density, we have $\rho_i^n = K\mathcal{M}_i$, $\forall i$ implies $\rho_i^{n+1} = K\mathcal{M}_i$, $\forall i$.

4.4 Positivity

At time step n, assume $\rho_i^n > 0$ for every i, then $g_i^n = \rho_i^n / \mathcal{M}_i \geq 0$ for every i, since invariant measure is positive. If all suitable mesh and time step constraints hold so that all the monotonicity results in Sect. 3 hold, then $A^{-1} \geq 0$ holds. Since $\mathcal{M}_i > 0$ and $\omega_i > 0$, we have

$$\vec{g}_h^{n+1} = A^{-1}WM\vec{g}_h^n \geq 0$$

thus $\rho_i^{n+1} = \mathcal{M}_i g_i^{n+1} \geq 0$.

4.5 Energy Dissipation

For any convex function $f(x)$, define the discrete energy at time step n as

$$E(\rho_h^n) = \langle \mathcal{M}f(\frac{\rho_h^n}{\mathcal{M}}), 1 \rangle = \sum_{i=1}^{N^2} \omega_i \mathcal{M}_i f\left(\frac{\rho_i^n}{\mathcal{M}_i}\right) = \langle \mathcal{M}f(g_h^n), 1 \rangle = \sum_{i=1}^{N^2} \omega_i \mathcal{M}_i f(g_i^n).$$

Theorem 4.1 *Assume the velocity field satisfies the discrete divergence free constraint (4.3), e.g, (3.5) in the second order scheme, and (3.14) in the fourth order scheme. If the scheme (2.20) is monotone, then for any convec function $f(x)$, it dissipates the discrete energy:*

$$\sum_{i=1}^{N^2} \omega_i \mathcal{M}_i f\left(\frac{\rho_i^{n+1}}{\mathcal{M}_i}\right) \le \sum_{i=1}^{N^2} \omega_i \mathcal{M}_i f\left(\frac{\rho_i^n}{\mathcal{M}_i}\right).$$

Proof Let a^{ij} be the entries of A^{-1}. Then $\vec{g}_h^{n+1} = \mathsf{A}^{-1}\mathsf{W}\mathsf{M}\vec{g}_h^n$ gives

$$g_i^{n+1} = \sum_j a^{ij} \omega_j \mathcal{M}_j g_j^n. \tag{4.7}$$

Next we show that (4.7) is a convex combination due to monotonicity and natural properties of the finite element method. The monotonicity implies $a^{ij} \ge 0$. The property $\mathsf{M}^{-1}\mathsf{W}^{-1}\mathsf{A}\vec{1} = \vec{1}$ gives

$$\vec{1} = \mathsf{A}^{-1}\mathsf{W}\mathsf{M}\vec{1} \implies \sum_j a^{ij} \omega_j \mathcal{M}_j = 1.$$

Thus (4.7) is a convex combination. For a convex function f, Jensen's inequality gives

$$f(g_i^{n+1}) \le \sum_j a^{ij} \omega_j \mathcal{M}_j f(g_j^n).$$

On the other hand, the property (4.2) implies

$$\vec{1}^T \mathsf{A} = \vec{1}^T \mathsf{W}\mathsf{M} \implies \vec{1}^T = \vec{1}^T \mathsf{W}\mathsf{M}\mathsf{A}^{-1} \implies \sum_i a^{ij} \omega_i \mathcal{M}_i = 1$$

So we have

$$
\begin{aligned}
E^{n+1} &= \sum_i \omega_i \mathcal{M}_i f(\frac{\rho_i^{n+1}}{\mathcal{M}_i}) = \sum_i \omega_i \mathcal{M}_i f(g_i^{n+1}) \\
&\le \sum_i \omega_i \mathcal{M}_i \left[\sum_j a^{ij} \omega_j \mathcal{M}_j f(g_j^n)\right] \\
&= \sum_j \omega_j \mathcal{M}_j \left[\sum_i a^{ij} \omega_i \mathcal{M}_i\right] f(g_j^n) = \sum_j \omega_j \mathcal{M}_j f(g_j^n) = E^n. \quad (4.8)
\end{aligned}
$$

□

Remark 3 As a special case, by choosing the convex function $f(x) = (x-1)^2$ and using the discrete mass conservation (4.5), the discrete energy dissipation law (4.8) reduces to the following form

$$\sum_i \omega_i \left(\frac{\rho_i^{n+1}}{\mathcal{M}_i} - 1\right)^2 \mathcal{M}_i < \sum_i \omega_i \left(\frac{\rho_i^n}{\mathcal{M}_i} - 1\right)^2 \mathcal{M}_i, \tag{4.9}$$

which is viewed as a discrete energy dissipation law w.r.t the Pearson χ^2-divergence.

Table 1 Accuracy test for a steady state solution (5.2) to the Fokker–Planck equation (5.1) with a source term

FD grid	The second order scheme (2.22)				The fourth order scheme (2.24)			
	l^2 error	Order	l^∞ error	Order	l^2 error	Order	l^∞ error	Order
9×9	2.99E−1	–	2.93E−1	–	1.66E−2	–	1.17E−2	–
17×17	6.00E−2	2.32	8.38E−2	1.81	9.98E−4	4.05	8.15E−4	3.84
33×33	1.21E−2	2.31	2.21E−2	1.92	6.14E−5	4.02	5.31E−5	3.94
65×65	2.59E−3	2.23	5.67E−3	1.96	3.81E−6	4.01	3.31E−6	4.00
129×129	5.85E−4	2.15	1.44E−3	1.98	2.37E−7	4.01	2.07E−7	4.00

5 Numerical Tests

5.1 Accuracy Test

We consider the scheme (2.20) solving

$$\rho_t = \nabla \cdot \left(\mathcal{M} \nabla \frac{\rho}{\mathcal{M}} \right) + \vec{u} \cdot \nabla \frac{\rho}{\mathcal{M}} + f \tag{5.1}$$

on $\Omega = (0, \pi) \times (0, \pi)$ with no flux boundary condition, i.e., $\nabla \frac{\rho}{\mathcal{M}} \cdot \vec{n} = 0$. We test the second order and fourth order spatial accuracy on a steady state solution

$$\rho(x, y, t) = (3 \cos x \cos y + 3)(2 + \sin x \sin y), \tag{5.2}$$

with $\mathcal{M} = 2 + \sin x \sin y$, $\vec{u} = \langle \sin x \cos y, \cos x \sin y \rangle$. The source $f(x, y)$ is chosen such that $\rho(x, y, t)$ above is the exact solution to (5.1).

The time step is set as $\Delta t = \Delta x$ and errors at $T = 1$ are given in 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 Numerical Examples with a Given Sampling Target $\mathcal{M}$

We consider examples with a given sampling target π. On a 2D domain $\Omega = [-4.5, 4.5] \times [-4.5, 4.5]$, the stream function for a 2D sinusoidal cellular flow is given as

$$\psi(x, y) := A \sin(k\pi x) \sin(k\pi y), \tag{5.3}$$

where A represents the amplitude of the mixture velocity $\vec{u}$ and k is the normalized wave number of the mixture. Then the incompressible velocity field is given as $\vec{u} = \begin{pmatrix} -\partial_y \psi \\ \partial_x \psi \end{pmatrix}$.

The target density is taken to be a smiling triple-banana image:

$$\mathcal{M}(x, y) = e^{-20\left[\left(x - \frac{6}{5}\right)^2 + \left(y - \frac{6}{5}\right)^2 - \frac{1}{2}\right]^2 + \log\left(e^{-10(y-2)^2}\right)}$$
$$+ e^{-20\left[\left(x + \frac{6}{5}\right)^2 + \left(y - \frac{6}{5}\right)^2 - \frac{1}{2}\right]^2 + \log\left(e^{-10(y-2)^2}\right)}$$

$$+ e^{-20(x^2+y^2-2)^2+\log\left(e^{-10(y+1)^2}\right)} + 0.1. \tag{5.4}$$

Then we take a Gaussian mixture as the initial density

$$\rho^0(x, y) = e^{-16(x+3)^2-4y^2} + e^{-16(x-3)^2-4y^2} + e^{-4x^2-16(y+3)^2} + e^{-4x^2-16(y-3)^2} + 0.1. \tag{5.5}$$

The numerical solutions for both second order and fourth order schemes, as well as the energy evolution for $E = \int_\Omega \frac{\rho^2}{\mathcal{M}} d\mathbf{x}$, are given in Fig. 2. From the color contour, no visual difference can be observed. The positivity-preserving and energy decay can be proved for both schemes.

Next we consider a different example, in which the fourth order spatial discretization can produce visually better results than the second order one. The computational domain is $[-3, 3] \times [-3, 3]$, the diffusion constant is $D = 0.5$, and the velocity filed is defined by derivatives of the stream function (5.3) with $A = 0.2$ and $k = 1$.

Now the initial data is chosen to be

$$\rho^0(x, y) = \frac{1}{2}e^{-16(x+1)^2-4y^2} + \frac{1}{2}e^{-16(x-1)^2-4y^2} + e^{-x^2/4-(y+3)^2} + e^{-x^2/4-(y-3)^2} + 0.1, \tag{5.6}$$

while the target density is

$$\mathcal{M}(x, y) = e^{-(x+3)^2-y^2/4} + e^{-(x-3)^2-y^2/4} + \frac{1}{2}e^{-4x^2-16(y+1)^2} + \frac{1}{2}e^{-4x^2-16(y-1)^2} + 0.1. \tag{5.7}$$

See Fig. 3 for the numerical solutions of the second example. The second order scheme on the finest mesh 301×301 grid with time step $\Delta t = 0.005$ can be regarded as the reference solution. On the same coarse 101×101 grid with the same time step $\Delta t = 0.02$, we observe that the second order scheme produces a wrong solution, while the fourth order scheme produces a better solution. A preconditioned conjugate gradient method is used to solve the linear systems in the semi-implicit schemes, and the cost for both second order and fourth order schemes on the same grid is about the same. Thus the fourth order scheme has clear advantages, even though the time discretization is only first order.

6 Concluding Remarks

In this paper, we have constructed second order and fourth order space discretization via finite difference implementation of the finite element method for solving Fokker–Planck equations associated with irreversible processes. Under mild mesh conditions and time step constraints for smooth solutions, the high order schemes are proved to be monotone, thus are positivity-preserving and energy dissipative. Even though the time discretization is only first order, numerical tests suggest that the fourth order spatial scheme produces better solutions than the second order one on the same grid. The high order schemes proposed in this paper preserve all the good properties just as the classical first order upwind schemes, such as (i) the conservation of total mass, (ii) the positivity of ρ, (iii) the energy dissipation law with respect to ϕ-entropy, and (iv) exponential convergence to equilibrium $\mathcal{M}$. Those properties are important but difficult to obtain for high order space discretizations, particularly for irreversible drift-diffusion processes. This also enables the future studies for sampling
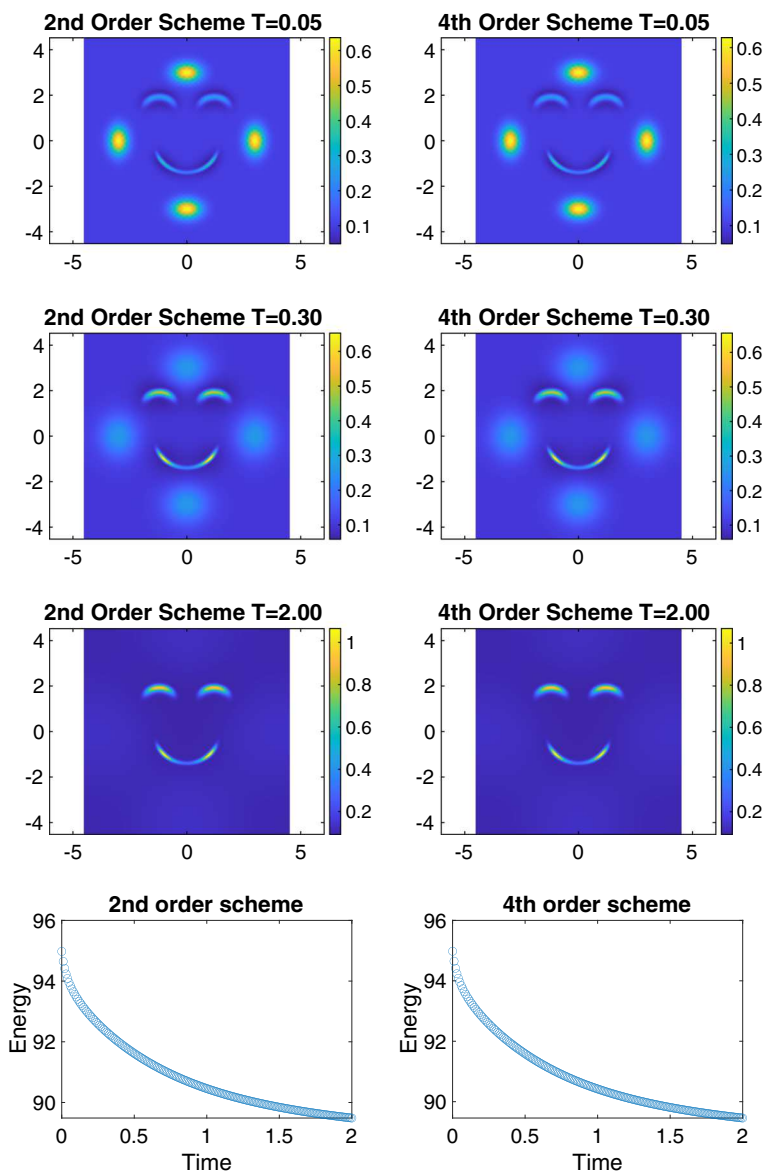
Fig. 2 Numerical solutions to the Fokker–Planck equation with a smiling triple-banana target density (5.4) and the initial density (5.5). Both the second order scheme (2.22) and the fourth order scheme (2.24) are used on the same 201×201 grid with $\Delta t = 0.01$
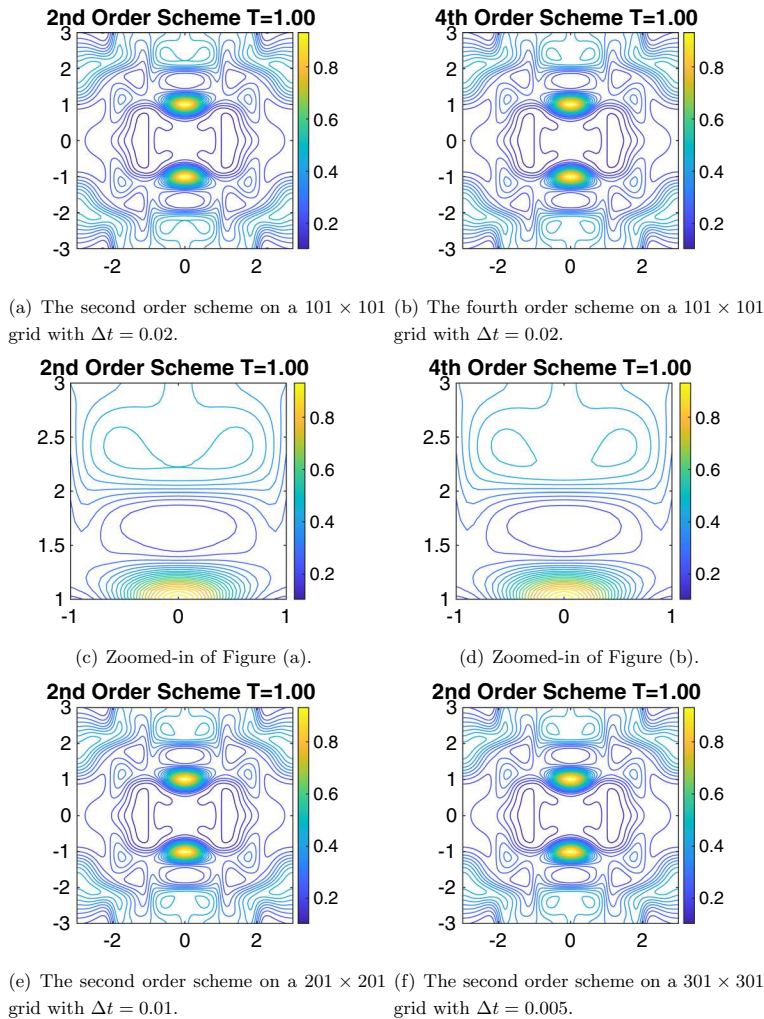
(a) The second order scheme on a 101×101 grid with $\Delta t = 0.02$.

(b) The fourth order scheme on a 101×101 grid with $\Delta t = 0.02$.

(c) Zoomed-in of Figure (a).

(d) Zoomed-in of Figure (b).

(e) The second order scheme on a 201×201 grid with $\Delta t = 0.01$.

(f) The second order scheme on a 301×301 grid with $\Delta t = 0.005$.

Fig. 3 Numerical solutions to the Fokker–Planck equation with the target density (5.7) and the initial density (5.6). The solution on the finest grid in **f** can be regarded as the reference solution. By comparing **a**, **b**, **c** and **d** with the reference solution in **f**, we can observe that the fourth order scheme (2.24) produces better results than the second order scheme (2.22) on a coarse 101×101 grid

acceleration and variance reduction using irreversible processes with high order numerical schemes.

Data Availibility All data generated or analysed during this study are included in this article.

Declarations

Conflict of interest The authors have no conflict of interest.

References

1. Arnold, A., Markowich, P., Toscani, G., Unterreiter, A.: On Convex Sobolev Inequalities and the Rate of Convergence to Equilibrium for Fokker–Planck Type Equations (2001)
2. Bank, R.E., Coughran, W.M., Jr., Cowsar, L.C.: The finite volume Scharfetter–Gummel method for steady convection–diffusion equations. Comput. Vis. Sci. **1**(3), 123–136 (1998)
3. Bessemoulin-Chatard, M.: A finite volume scheme for convection–diffusion equations with nonlinear diffusion derived from the Scharfetter–Gummel scheme. Numerische Mathematik **121**(4), 637–670 (2012)
4. Bodineau, T., Lebowitz, J., Mouhot, C., Villani, C.: Lyapunov functionals for boundary-driven nonlinear drift-diffusion equations. Nonlinearity **27**(9), 2111 (2014)
5. Cattiaux, P.: Stochastic calculus and degenerate boundary value problems. Annales de l'institut Fourier **42**(3), 541–624 (1992)
6. Chainais-Hillairet, C., Herda, M.: Large-time behaviour of a family of finite volume schemes for boundary-driven convection–diffusion equations. IMA J. Numer. Anal. **40**(4), 2473–2504 (2020)
7. Chainais-Hillairet, C., Liu, J.-G., Peng, Y.-J.: Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. ESAIM Math. Model. Numer. Anal.-Modélisation Mathématique et Analyse Numérique **37**(2), 319–338 (2003)
8. Chen, M.-F.: Equivalence of exponential ergodicity and L2-exponential convergence for Markov chains. Stoch. Process. Appl. **87**(2), 281–297 (2000)
9. Delarue, F., Lagoutiére, F.: Probabilistic analysis of the upwind scheme for transport equations. Arch. Ration. Mech. Anal. **199**(1), 229–268 (2011)
10. Duncan, A.B., Lelievre, T., Pavliotis, G.A.: Variance reduction using non reversible Langevin samplers. J. Stat. Phys. **163**(3), 457–491 (2016)
11. E, W., Vanden-Eijnden, E.: Towards a theory of transition paths. J. Stat. Phys. **123**(3), 503 (2006)
12. Filbet, F., Herda, M.: A finite volume scheme for boundary-driven convection–diffusion equations with relative entropy structure. Numerische Mathematik **137**(3), 535–577 (2017)
13. Gao, Y., Li, T., Li, X., Liu, J.-G.: Transition path theory for Langevin dynamics on manifold: optimal control and data-driven solver. Multiscale Model. Simul. **21**(1), 1–33 (2023)
14. Gao, Y., Liu, J.-G.: A note on parametric Bayesian inference via gradient flows. Ann. Math. Sci. Appl. **5**(2), 261–282 (2020)
15. Gao, Y., Liu, J.-G.: Random walk approximation for irreversible drift-diffusion process on manifold: ergodicity, unconditional stability and convergence. Commun. Comput. Phys. **6**, 66 (2023)
16. Gao, Y., Liu, J.-G.: Revisit of macroscopic dynamics for some non-equilibrium chemical reactions from a Hamiltonian viewpoint. J. Stat. Phys. **189**(2), 22 (2022)
17. Gao, Y., Liu, J.-G., Wu, N.: Data-driven efficient solvers for Langevin dynamics on manifold in high dimensions. Appl. Comput. Harmon. Anal. **62**, 261–309 (2023)
18. Hill, T.L.: Free Energy Transduction and Biochemical Cycle Kinetics. Dover Books on Chemistry, Dover Publications, New York (2005)
19. Hu, J., Zhang, X.: Positivity-preserving and energy-dissipative finite difference schemes for the Fokker–Planck and Keller–Segel equations. IMA J. Numer. Anal. **6**, 66 (2022)
20. Li, H., Appelö, D., Zhang, X.: Accuracy of spectral element method for wave, parabolic, and Schrödinger equations. SIAM J. Numer. Anal. **60**(1), 339–363 (2022)
21. Li, H., Zhang, X.: On the monotonicity and discrete maximum principle of the finite difference implementation of C0–Q2 finite element method. Numerische Mathematik **145**(2), 437–472 (2020)
22. Li, H., Zhang, X.: Superconvergence of high order finite difference schemes based on variational formulation for elliptic equations. J. Sci. Comput. **82**(2), 1–39 (2020)
23. Li, L., Liu, J.-G.: Large time behaviors of upwind schemes and B-schemes for Fokker–Planck equations on R by jump processes. Math. Comput. **89**(325), 2283–2320 (2020)
24. Li, T., Li, X., Zhou, X.: Finding transition pathways on manifolds. Multiscale Model. Simul. **14**(1), 173–206 (2016)
25. Li, T., Lin, F.: Large deviations for two-scale chemical kinetic processes. Commun. Math. Sci. **15**(1), 123–163 (2017)
26. Lorenz, J.: Zur inversmonotonie diskreter probleme. Numerische Mathematik **27**(2), 227–238 (1977)
27. Maday, Y., Ronquist, E.M.: 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)
28. Markowich, P.A.: The Stationary Semiconductor Device Equations. Springer, Berlin (1985)
29. Markowich, P.A., Zlámal, M.A.: Inverse-average-type finite element discretizations of selfadjoint second-order elliptic problems. Math. Comput. **51**(184), 431–449 (1988)
30. Plemmons, R.J.: M-matrix characterizations. I-nonsingular M-matrices. Linear Algebra Appl. **18**(2), 175–188 (1977)

31. Prigogine, I.: Introduction to Thermodynamics of Irreversible Processes. Wiley, New York (1968)
32. Qian, H., Ge, H.: Stochastic Chemical Reaction Systems in Biology. Lecture Notes on Mathematical Modelling in the Life Sciences. Springer, Berlin (2021). https://doi.org/10.1007/978-3-030-86252-7
33. Scharfetter, D.L., Gummel, H.K.: Large-signal analysis of a silicon read diode oscillator. IEEE Trans Electron Devices **16**(1), 64–77 (1969)
34. Shen, J., Zhang, X.: Discrete maximum principle of a high order finite difference scheme for a generalized Allen–Cahn equation. Commun. Math. Sci. **20**(5), 1409–1436 (2022)
35. Tao, M.: Hyperbolic periodic orbits in nongradient systems and small-noise-induced metastable transitions. Phys. D Nonlinear Phenom. **363**, 1–17 (2018)
36. Turitsyn, K.S., Chertkov, M., Vucelja, M.: Irreversible Monte Carlo algorithms for efficient sampling. Phys. D Nonlinear Phenom. **240**(4–5), 410–414 (2011)
37. Wei, W., Gao, T., Chen, X., Duan, J.: An optimal control method to compute the most likely transition path for stochastic dynamical systems with jumps'. Chaos Interdiscip. J. Nonlinear Sci. **32**(5), 051102 (2022)
38. Xu, J., Zikatanov, L.: A monotone finite element scheme for convection–diffusion equations. Math. Comput. **68**(228), 1429–1446 (1999)
39. Ye, X., Zhou, Z.: Efficient sampling of thermal averages of interacting quantum particle systems with random batches. J. Chem. Phys. **154**(20), 204106 (2021)
40. Zhang, L., Wang, H., Weinan, E.: Reinforced dynamics for enhanced sampling in large atomic and molecular systems. J. Chem. Phys. **148**, 124113 (2018)