A POSITIVITY-PRESERVING AND ENERGY STABLE SCHEME FOR A QUANTUM DIFFUSION EQUATION

XIAOKAI HUO AND HAILIANG LIU

ABSTRACT. We propose a new fully-discretized finite difference scheme for a quantum diffusion equation, in both one and two dimensions. This is the first fully-discretized scheme with proved positivity-preserving and energy stable properties using only standard finite difference discretization. The difficulty in proving the positivity-preserving property lies in the lack of a maximum principle for fourth order PDEs. To overcome this difficulty, we reformulate the scheme as a variational structure based optimization problem and use the singularity of the energy functional at zero to exclude the possibility of non-positive solutions. The scheme is also shown to be mass conservative and consistent.

1. Introduction

Nonlinear diffusion equations of fourth and higher order have since long been of interest in various fields of mathematical physics with diverse applications. However, few mathematical results and numerical tools are available for higher-order equations when compared to the theory of second-order diffusion equations. One most important property for many higher-order diffusion equations is the positiveness of their solutions. From the maximum principle it naturally follows the positiveness for second order equations. However, for higher order equations, the positiveness of solutions is rather subtle to prove due to the lack of a maximum principle. The purpose of this paper is to seek a novel numerical scheme for higher-order diffusion equations using standard finite difference discretization, yet with proven positivity of numerical solutions.

In this paper, we focus on the quantum diffusion equation on the torus $\mathbb{T}^d$ ($d \geq 1$ be any positive integer) with periodic boundary conditions:

$$\partial_t u = -2\nabla \cdot \left(u \nabla \frac{\Delta \sqrt{u}}{\sqrt{u}} \right), \quad u(\cdot, 0) = u_0 > 0. \tag{1}$$

Here $u = u(x,t) \in \mathbb{R}$ is a scalar unknown function.

Equation (1) has various physical backgrounds. In one dimensional case, it is the Derrida-Lebowitz-Speer-Spohn equation and was first deduced in [8] in the context of spin systems. In the multi-dimensional case, it appears in the context of semiconductor modeling. It can be viewed as the evolution of the density of electrons u with vanishing temperature of the simplified quantum drift-diffusion model [1, 11]:

$$\partial_t u = \operatorname{div}(T\nabla u + u\nabla V), \quad V = V_e - \frac{\epsilon^2}{6}\frac{\Delta\sqrt{u}}{\sqrt{u}}.$$

Key words and phrases. Finite difference, Higher-order parabolic equations, Positivity-preserving, Energy dissipation .

Here $T > 0$ is the temperature, ϵ the Planck constant, and V is the potential felt by the electrons, which splits into the classical electric potential V_e and the Bohm potential, describing quantum effects. The equation can also be derived from quantum hydrodynamics as the high friction limit of some quantum hydrodynamic equations, see [13, 14, 21].

1.1. **Related work.** Positivity of solutions plays an important role in the analysis of equation (1). The existence and uniqueness of solutions of the equation on the one dimensional torus were proved in the space $H^1(\mathbb{S}^1)$ for finite time in [3]. It was proved that the solution $u(x,t) > 0$ for any $x \in \mathbb{S}^1$, $t \in [0, T^*)$ holds if the initial condition $u_0(x) > 0$ holds almost everywhere and if the maximal existence time T^* is finite, the limit $\lim_{t\to T^*} u(\cdot, t)$ vanishes at some point. The global existence of non-negative weak solutions to the equation (1) was proved in [17] for the one dimensional case and in [16] and [12] for the multi-dimensional case. In all these works positivity of solutions are essential. We also refer the reader to [8, 5, 19] for further theoretical works on this equation.

The positiveness of equation (1) is seen as natural within the theory of gradient flows in a Wasserstein space. In fact, the equation (1) can be rewritten into the form

$$\partial_t u = \nabla \cdot \left(u \nabla \left(\frac{\delta F}{\delta u} \right) \right), \tag{2}$$

where $F = F(u)$ is a functional of u given by

$$F(u) = \frac{1}{2} \int_{\mathbb{T}^d} \frac{|\nabla u|^2}{u} dx. \tag{3}$$

The functional F is often called *Fisher information*, it will be also called *energy functional* in this paper. It was proved in [12] that the equation (1) is the gradient flow driven by the functional F of the probability measure u with respect to the Wasserstein distance

$$\mathrm{W}_2^2(u^0, u^1) = \inf_{\gamma \in \Pi(u^0, u^1)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 d\gamma(x, y). \tag{4}$$

where $\Pi(u^0, u^1)$ is the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with first marginal u^0 and second marginal u^1, and the symbol $|\cdot|$ denotes the usual Euclidean norm in $\mathbb{R}^d$. As a probability measure, u cannot be negative.

Therefore, positivity of numerical solutions is naturally desired for any numerical method to solve the equation (1). However, this cannot be implemented easily. Due to the maximum principle for second order heat equations, a central finite difference scheme can be readily shown positivity-preserving, but the lack of maximum principle for the equation (1) does pose a challenge on the design of a positivity-preserving scheme [8].

One way to obtain positivity is to introduce new variables and enforce the solution positivity through them. For example, one can introduce $v = \log u$ and take $u = e^v$ in the numerical scheme to get a non-negative solution, as done in [6, 18]. Another choice is $v = \sqrt{u}$ [4]. However, positivity of numerical solutions does not seem to follow naturally from the respective scheme, instead, additional equations need to be solved in these works.

Another way to obtain positivity is to use the Wasserstein gradient flow structure. For one dimensional case, a fully discrete Lagrangian scheme was developed in [28]. In such framework, the Wasserstein distance (4) can be expressed as the L^2 distance of two Lagrangian maps corresponding to measure u^0 and u^1, respectively. Positivity of u then follows from the fact that $u = 1/(\partial_\xi X(t,\xi))$ and monotonicity of the Lagrangian map $\xi \to X(\cdot,\xi)$, see [28]. Another scheme proposed in [22] utilized the Eulerian formulation of the Wasserstein gradient flow structure. More precisely, by the Benamou-Brenier formulation of the Wasserstein distance [2], which is defined by

$$\mathrm{W}(u^0,u^1) := \left\{ \inf_{u,m} \int_0^1 \int_{\mathbb{T}^d} \frac{m^2}{u} dxdt, \quad \text{s.t. } \partial_t u + \nabla \cdot m = 0, u(0,x) = u^0(x), u(1,x) = u^1(x) \right\},$$

their scheme for (1) takes the form

$$u^{n+1} = \left\{ \arg\inf_{u,m} \int_{\mathbb{T}^d} \frac{m^2}{2u} dx + \Delta t F(u) \quad \text{s.t. } u - u^n(x) + \nabla \cdot m(x) = 0 \right\}. \tag{5}$$

Here Δt is the time step, and u^n is the solution at the n-th time step. Positivity of u^n was proved since the objective function in the above optimization problem becomes infinite when u touches zero [22].

Some other numerical methods like finite volume approximations were also developed to solve this equation, see, e.g., [26]. However, the current known positivity-preserving numerical methods for equation (1) do not follow simply from a direct finite difference discretization. In [28], the authors even argued that there is no reason to expect the positivity-preserving property from a standard discretization approach.

1.2. **Our contributions.** The main objectives of this paper are to present a novel numerical scheme to approximate (1) using a standard finite difference discretization and prove how positivity of numerical solutions can be obtained from such a discretization. Our scheme, based on the gradient flow structure (2), reads as

$$\frac{u^{n+1} - u^n}{\Delta t} = \nabla \cdot (u^n \nabla H^{n+1}), \quad H^{n+1} = \delta_u F(u^{n+1}), \tag{6}$$

where H^{n+1} is obtained by a variational derivative of energy functional $F = F(u)$. To prove the positivity property, we rewrite (6) into an optimization problem

$$u^{n+1} = \arg\inf_{u} \left\{ \frac{1}{2} \|u - u^n\|^2_{\mathcal{L}^{-1}_{u^n}} + \Delta t F(u) \right\}, \tag{7}$$

with the norm defined by

$$\|f\|_{\mathcal{L}^{-1}_{u^n}} := \{ \|\nabla \phi\|_{L^2_{u^n}} : f = -\nabla \cdot (u^n \nabla \phi) \}.$$

Note that (7) when setting $u - u^n = -\nabla \cdot m$ with $m = u^n \nabla \phi$ can be reformulated as

$$u^{n+1} = \left\{ \arg\inf_{u,m} \int_{\mathbb{T}^d} \frac{|m|^2}{2u^n} dx + \Delta t F(u) \quad \text{s.t. } u - u^n(x) + \nabla \cdot m = 0 \right\}. \tag{8}$$

This is similar in structure to the scheme (5). Note that the optimal conditions for this optimization problem lead to precisely the scheme (6), but the optimal conditions for (5) links only to a first order approximation of (6) or (11). From the optimization problem (7) positivity of u can be deduced because the objective function therein develops singularity if u touches zero assuming that ∇u does not vanish simultaneously. The issue that u and ∇u vanish simultaneously was resolved in the spatial discretization of (6), together with the conservation property of the fully-discretized scheme.

Our approach in proving the positivity-preserving property using optimization formulations was motivated by the works [7] and [9], where the authors used the finite difference discretization to study the Cahn-Hillard equation

$$\partial_t \phi = \nabla \cdot (M(\phi)\nabla \mu),\ \mu = \delta_\phi E$$

with the energy functional E given by, for example

$$E(\phi) = \int_\Omega \left((1+\phi)\log\phi + (1-\phi)\log(1-\phi) - \frac{\theta_0}{2}\phi^2 + \frac{\varepsilon^2}{2}|\nabla\phi|^2 \right) dx$$

in [7]. Our problem here differs from the Cahn-Hillard equation in that the functional includes a higher order term which develops singularity at zero, and the possibility of ∇u and u being zero simultaneously also poses additional challenges.

The main results of this paper include a new fully discretized finite difference scheme for (1), in both one and higher dimensions, and rigorous proofs of scheme properties such as positivity-preserving, mass conservation, energy stability, and the scheme consistency. The most remarkable contribution of this work is to show that standard finite difference discretization of the quantum diffusion equation (1) can meet the requirement of both solution positivity and energy dissipation simultaneously.

The paper is organized as follows. In the next section we present four semidiscrete schemes with only time discretization for the equation (1) and briefly discuss their properties. Section 3 is devoted to the fully discretized scheme in one dimensional case, where the scheme is shown to be energy stable. The positivity-preserving property will be proved in section 4, which is the most important part of this paper. The consistency error will be calculated in section 5. In section 6, we present and analyze the scheme in higher dimensions, taking the two dimensional case as an example. Finally, some numerical examples are presented in section 7.

Notations. We use L^∞ to denote the space of bounded sequences and $C_{t,x}$ to denote the space of continuous functions depending on time and space. We use bold symbol $\mathbf{f} = (f^x, f^y)$ to denote vector in $\mathbb{R}^2$. We take h to be the mesh size and Δt to be the time step in discretization.

2. Time discretization

In this section we present four semidiscrete schemes with only time discretization, and comment on both pros and cons of each scheme.

2.1. **Explicit scheme.** The simplest scheme is the explicit scheme:

$$\frac{u^{n+1}-u^n}{\Delta t} = -2\nabla \cdot \left(u^n \nabla \frac{\Delta\sqrt{u^n}}{\sqrt{u^n}} \right). \tag{9}$$

This scheme is easy to implement. However, the sign of the right hand side is not certain and it is difficult to show the positivity of solutions. The scheme can also be written in the following form

$$\frac{u^{n+1}-u^n}{\Delta t} = \nabla \cdot \left(u^n \nabla \frac{\delta F(u^n)}{\delta u} \right). \tag{10}$$

For the energy dissipation, we calculate the relative energy defined by

$$\begin{aligned}
F(u^{n+1}|u^n) :=& F(u^{n+1}) - F(u^n) - \int_{\mathbb{T}^d} \frac{\delta F}{\delta u}(u^n)(u^{n+1}-u^n)dx \\
=& \frac{1}{2}\int_{\mathbb{T}^d} \frac{|\nabla u^{n+1}|^2}{u^{n+1}}dx - \frac{1}{2}\int_{\mathbb{T}^d} \frac{|\nabla u^n|^2}{u^n}dx + \frac{1}{2}\int_{\mathbb{T}^d} \frac{|\nabla u^n|^2}{(u^n)^2}(u^{n+1}-u^n)dx \\
& - \frac{1}{2}\int_{\mathbb{T}^d} \frac{2(\nabla u^n \cdot (\nabla u^{n+1} - \nabla u^n))}{u^n}dx \\
=& \frac{1}{2}\int_{\mathbb{T}^d} u^{n+1} \left(\frac{\nabla u^{n+1}}{u^{n+1}} - \frac{\nabla u^n}{u^n} \right)^2 dx.
\end{aligned}$$

From the equation (10), we have

$$\begin{aligned}
\int_{\mathbb{T}^d} \frac{\delta F}{\delta u}(u^n)(u^{n+1}-u^n)dx =& \Delta t \int_{\mathbb{T}^d} \frac{\delta F}{\delta u}(u^n) \nabla \cdot \left(u^n \nabla \frac{\delta F}{\delta u}(u^n) \right) dx \\
=& -\Delta t \int_{\mathbb{T}^d} u^n \left| \nabla \frac{\delta F}{\delta u}(u^n) \right|^2 dx \le 0,
\end{aligned}$$

So the energy difference at each time step will change by

$$F(u^{n+1}) - F(u^n) = \frac{1}{2}\int_{\mathbb{T}^d} u^{n+1} \left(\frac{\nabla u^{n+1}}{u^{n+1}} - \frac{\nabla u^n}{u^n} \right)^2 dx - \Delta t \int_{\mathbb{T}^d} u^n \left| \nabla \frac{\delta F}{\delta u}(u^n) \right|^2 dx.$$

If we assume $u^n > 0$ and $u^{n+1} > 0$, then the second term on the right hand side will be negative while the first term be positive. The right hand side is not guaranteed to have a definite sign unless Δt is very small. So for the explicit scheme, the positivity-preserving and energy dissipation properties are not guaranteed.

2.2. **Fully implicit scheme.** For stability reasons, we prefer to use an implicit scheme instead of an explicit one. The implicit scheme for the equation (1) reads as

$$\frac{u^{n+1}-u^n}{\Delta t} = -\nabla \cdot \left(u^{n+1} \nabla \frac{\Delta \sqrt{u^{n+1}}}{\sqrt{u^{n+1}}} \right). \tag{11}$$

Such scheme was studied in [4], where the dissipation of Fisher information was shown. In other words, assuming $u^n > 0$ and $u^{n+1} > 0$, the scheme dissipates the energy. For the convenience of reference, we present this result and a proof as follows.

Lemma 2.1. *Suppose $u^n, u^{n+1} > 0$ everywhere in $\mathbb{T}^d$, then*

$$F(u^{n+1}) \le F(u^n) \tag{12}$$

for any $\Delta t > 0$.

Proof. We calculate the relative energy $F(u^n|u^{n+1})$ by

$$
\begin{aligned}
F(u^n|u^{n+1}) = & F(u^n) - F(u^{n+1}) - \int_{\mathbb{T}^d} \frac{\delta F}{\delta u}(u^{n+1})(u^n - u^{n+1})dx \\
=&\frac{1}{2}\int_{\mathbb{T}^d} \frac{|\nabla u^n|^2}{u^n}dx - \frac{1}{2}\int_{\mathbb{T}^d} \frac{|\nabla u^{n+1}|^2}{u^{n+1}}dx + \frac{1}{2}\int_{\mathbb{T}^d} \frac{|\nabla u^{n+1}|^2}{(u^{n+1})^2}(u^n - u^{n+1})dx \\
& - \frac{1}{2}\int_{\mathbb{T}^d} \frac{2\nabla u^{n+1}\cdot(\nabla u^n - \nabla u^{n+1})}{u^{n+1}}dx \\
=&\frac{1}{2}\int_{\mathbb{T}^d} u^n\left(\frac{|\nabla u^n|^2}{|u^n|^2} + \frac{|\nabla u^{n+1}|^2}{|u^{n+1}|^2} - \frac{2\nabla u^{n+1}\cdot\nabla u^n}{u^{n+1}u^n}\right)dx \\
=&\frac{1}{2}\int_{\mathbb{T}^d} u^n\left(\frac{\nabla u^n}{u^n} - \frac{\nabla u^{n+1}}{u^{n+1}}\right)^2 dx \geq 0. \qquad (13)
\end{aligned}
$$

We learned from the gradient structure that equation (11) can be written as

$$
\frac{u^{n+1} - u^n}{\Delta t} = \nabla\cdot\left(u^{n+1}\nabla\frac{\delta F}{\delta u}(u^{n+1})\right).
$$

And so

$$
\begin{aligned}
\int_{\mathbb{T}^d} \frac{\delta F}{\delta u}(u^{n+1})(u^n - u^{n+1})dx = & - \Delta t\int_{\mathbb{T}^d} \frac{\delta F}{\delta u}(u^{n+1})\nabla\cdot\left(u^{n+1}\nabla\frac{\delta F}{\delta u}(u^{n+1})\right)dx \\
=&\Delta t\int_{\mathbb{T}^d} u^{n+1}\left|\nabla\frac{\delta F}{\delta u}(u^{n+1})\right|^2 dx \geq 0, \qquad (14)
\end{aligned}
$$

due to the positivity of u^{n+1}. Substitute it into (13) gives

$$
F(u^n) - F(u^{n+1}) \geq F(u^n|u^{n+1}) \geq 0. \qquad (15)
$$

□

Notice that the positivity property is also crucial in establishing the energy stability, but it seems difficult to directly prove such property. Given u^n, the equation for u^{n+1} is a fourth order nonlinear elliptic equation, the positivity of u^{n+1} does not seem to be derivable from a maximum principle. Also the scheme is difficult to implement because we need to solve a fourth order nonlinear partial differential equations numerically each time step.

2.3. **A positivity-preserving scheme.** Drawing ideas from [23, 24, 25] in the design of unconditional positive schemes for second-order Fokker-Planck equations, we set

$$
H = -\frac{2\Delta\sqrt{u}}{\sqrt{u}}, \quad M = e^{lnu+H},
$$

system (1) can be rewritten in the form

$$
\partial_t u = \nabla\cdot\left(ue^H\nabla e^{-H}\right) = \nabla\cdot\left(M\nabla\left(\frac{u}{M}\right)\right),
$$

which can be approximated by

$$
\frac{u^{n+1} - u^n}{\Delta t} = \nabla\cdot\left(M^n\nabla\left(\frac{u^{n+1}}{M^n}\right)\right). \qquad (16)
$$

This scheme is linear in u^{n+1} and hence easy to code. Also it is unconditionally positivity-preserving.

Lemma 2.2. *If* $u^n > 0$*, then*

$$u^{n+1} > 0$$

for any $\Delta t > 0$.

Proof. Set $G^{n+1} = u^{n+1}/M^n$ so that

$$G^{n+1}M^n - \Delta t\nabla \cdot (M^n \nabla G^{n+1})) = u^n.$$

Note that both existence and regularity of the solution G^{n+1} are ensured by the classical elliptic theory. Assume G^{n+1} achieves its minimum at x_0, then $\nabla G^{n+1}(x_0) = 0$, and $\Delta G^{n+1}(x_0) \geq 0$. From the equation when evaluated at x_0 it follows that

$$G^{n+1}(x_0)M^n(x_0) = \Delta t M^n(x_0)\Delta G^{n+1}(x_0) + u^n(x_0) \geq u^n(x_0) > 0.$$

Hence $G^{n+1} > 0$ for all $x \in \mathbb{T}^d$, so is u^{n+1}. □

It seems less obvious to verify the energy dissipation property.

2.4. **An explicit-implicit scheme.** The fourth scheme is

$$\frac{u^{n+1} - u^n}{\Delta t} = \nabla \cdot \left(u^n \nabla H^{n+1}\right), \tag{17}$$

where

$$H^{n+1} = \frac{\delta F}{\delta u}(u^{n+1}) = -\frac{2\Delta\sqrt{u^{n+1}}}{\sqrt{u^{n+1}}}.$$

One can view this scheme as an intermediate one between (11) and (9) or (16). The scheme is unconditionally energy stable.

Lemma 2.3. *Suppose* $u^n > 0$*, then*

$$F(u^{n+1}) \leq F(u^n) \tag{18}$$

for any $\Delta t > 0$.

Proof. The proof is similar to the proof of Lemma 2.1, except that

$$\begin{aligned} \int_{\mathbb{T}^d} \frac{\delta F}{\delta u}(u^{n+1})(u^n - u^{n+1})dx =& - \Delta t \int_{\mathbb{T}^d} \frac{\delta F}{\delta u}(u^{n+1})\nabla \cdot \left(u^n \nabla \frac{\delta F}{\delta u}(u^{n+1})\right) dx \\ =& \Delta t \int_{\mathbb{T}^d} u^n \left|\nabla \frac{\delta F}{\delta u}(u^{n+1})\right|^2 dx \geq 0, \end{aligned} \tag{19}$$

so (15) holds, i.e., (18) holds. □

Notice that here the assumption $u^{n+1} > 0$ is not needed in getting the result, different from Lemma 2.1. In next section, we shall prove the positivity-preserving property of the fully discrete scheme corresponding to (17).

3. The full discrete scheme in one dimension

3.1. **Notations.** We use notations from [30]. We define the following two grids on the torus $\mathbb{T} = [0, L]$ with spacing $h = L/N$, where N is the number of mesh intervals:

$$\mathcal{C} := \left\{\frac{L}{N}, \frac{2L}{N}, \ldots, L\right\}, \quad \mathcal{E} := \left\{\frac{L}{2N}, \frac{3L}{2N}, \ldots, \frac{2N-1}{2N}L\right\}. \tag{20}$$

We treat $\mathcal{C}$ and $\mathcal{E}$ as periodic. For example, we write x_i as the i-th element in $\mathcal{C}$, then $x_N = L$ and $x_{N+1} = x_1 = L/N$. The elements in $\mathcal{E}$ then can be written as $x_{i+\frac{1}{2}}$ with $i \in \mathbb{Z}$. We define the discrete N-periodic function space as

$$\mathcal{C}_{\text{per}} := \{f : \mathcal{C} \to \mathbb{R}\}, \quad \mathcal{E}_{\text{per}} := \{f : \mathcal{E} \to \mathbb{R}\}.$$

Here we call $\mathcal{C}_{\text{per}}$ the space of *cell centered functions* and $\mathcal{E}_{\text{per}}$ the space of *edge centered functions*. We also define the homogeneous cell centered function $\mathring{\mathcal{C}}_{\text{per}}$ as the subspace of $\mathcal{C}_{\text{per}}$ with summation zero:

$$\mathring{\mathcal{C}}_{\text{per}} := \left\{ f : f \in \mathcal{C}_{\text{per}}, \sum_{i=1}^{N} f_i = 0 \right\}.$$

The discrete gradient D_h and d_h are defined to be

$$(D_h f)_{i+\frac{1}{2}} := \frac{f_{i+1} - f_i}{h}, \quad (d_h f)_i := \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{h}. \tag{21}$$

We define the average of the function values of nearby points by

$$\hat{f}_{i+\frac{1}{2}} = \frac{f_i + f_{i+1}}{2}, \text{ if } f \in \mathcal{C}_{\text{per}}, \quad \text{and} \quad \hat{f}_i = \frac{f_{i+\frac{1}{2}} + f_{i-\frac{1}{2}}}{2}, \text{ if } f \in \mathcal{E}_{\text{per}}. \tag{22}$$

The inner products on the grids are defined by

$$\langle f, g\rangle := h\sum_{i=1}^{N} f_i g_i, \ \forall f, g \in \mathcal{C}_{\text{per}}, \quad [f, g] := \frac{1}{2}h\sum_{i=1}^{N}(f_{i-\frac{1}{2}}g_{i-\frac{1}{2}} + f_{i+\frac{1}{2}}g_{i+\frac{1}{2}}), \ \forall f, g \in \mathcal{E}_{\text{per}}.$$

The corresponding norms in each space are defined by

$$\|f\|_{\mathcal{C}_{\text{per}}} = \sqrt{h\sum_{i=1}^{N} f_i^2}, \text{ for } f \in \mathcal{C}_{\text{per}}, \quad \|f\|_{\mathcal{E}_{\text{per}}} = \sqrt{h\sum_{i=1}^{N}\frac{1}{2}(f_{i+\frac{1}{2}}^2 + f_{i-\frac{1}{2}}^2)}, \text{ for } f \in \mathcal{E}_{\text{per}}.$$

Suppose $f, g \in \mathcal{C}_{\text{per}}$ and $\phi \in \mathcal{E}_{\text{per}}$, the following summation-by-parts formulas hold:

$$\langle f, d_h\phi\rangle = -[D_h f, \phi], \quad \langle f, d_h(\phi D_h g)\rangle = -[D_h f, \phi D_h g]. \tag{23}$$

We then introduce a discrete analogue of the space H^{-1}. Let $\phi \in \mathcal{E}_{\text{per}}, g \in \mathring{\mathcal{C}}_{\text{per}}$, and $f \in \mathring{\mathcal{C}}_{\text{per}}$ solve

$$\mathcal{L}_\phi(f) = -d_h(\phi D_h f) = g, \tag{24}$$

We define the following bilinear form on $\mathring{\mathcal{C}}_{\text{per}}$:

$$\langle g_1, g_2\rangle_{\mathcal{L}_\phi^{-1}} := [\phi D_h f_1, D_h f_2], \ \forall g_1, g_2 \in \mathring{\mathcal{C}}_{\text{per}},$$

where $f_1, f_2 \in \mathcal{C}_{\text{per}}$ are solutions of

$$\mathcal{L}_\phi(f_i) = -d_h(\phi D_h f_i) = g_i, \ \forall i = 1, 2. \tag{25}$$

The inner product $\langle\cdot\rangle_{\mathcal{L}_\phi^{-1}}$ satisfies the summation-by-parts formula

$$\langle g_1, g_2\rangle_{\mathcal{L}_\phi^{-1}} = \langle g_1, \mathcal{L}_\phi^{-1} g_2\rangle = \langle \mathcal{L}_\phi^{-1} g_1, g_2\rangle.$$

The corresponding norm on $\mathring{\mathcal{C}}_{\text{per}}$ is defined to be

$$\|g\|_{\mathcal{L}_\phi^{-1}} = \sqrt{\langle g, g\rangle_{\mathcal{L}_\phi^{-1}}}. \tag{26}$$

3.2. **The scheme.** We proceed to study the time discretization scheme (17). We adopt the following fully discrete scheme as

$$\frac{u^{n+1}-u^n}{\Delta t} = d_h\left(\widehat{u}^n D_h H^{n+1}\right), \tag{27}$$

where D_h, d_h are the forward and backward difference operators defined in (21), and $\hat{u}$ is the average operator defined in (22). The scheme can be written explicitly as

$$\begin{aligned}\frac{u_i^{n+1}-u_i^n}{\Delta t} &= (d_h(\widehat{u}^n D_h H^{n+1}))_i = \frac{\hat{u}^n_{i+\frac{1}{2}} D_h H^{n+1}_{i+\frac{1}{2}} - \hat{u}^n_{i-\frac{1}{2}} D_h H^{n+1}_{i-\frac{1}{2}}}{h} \\ &= \frac{(u^n_{i+1}+u^n_i)(H^{n+1}_{i+1}-H^{n+1}_i) - (u^n_i + u^n_{i-1})(H^{n+1}_i - H^{n+1}_{i-1})}{2h^2}.\end{aligned}$$

From the definition of H, we have

$$H = \frac{\delta F(u)}{\delta u} = -\frac{|\nabla u|^2}{2u^2} - \nabla\cdot\left(\frac{\nabla u}{u}\right).$$

To discretize H^{n+1}, we take

$$H^{n+1} = -\frac{|D_h u^{n+1}|^2}{2(u^{n+1})^2} - d_h\left(\frac{D_h u^{n+1}}{u^{n+1}}\right)$$

in the following sense

$$H_i^{n+1} = -\frac{1}{2h^2}\frac{(u^{n+1}_{i+1}-u^{n+1}_i)^2}{(u^{n+1}_i)^2} - \frac{1}{h^2}\left(\frac{u^{n+1}_{i+1}-u^{n+1}_i}{u^{n+1}_i} - \frac{u^{n+1}_i - u^{n+1}_{i-1}}{u^{n+1}_{i-1}}\right). \tag{28}$$

Actually, this is a discretized version of the variational formula. We define the discretized energy functional

$$F_h(u) := \frac{1}{h}\sum_{i=1}^{N}\frac{(u_{i+1}-u_i)^2}{2u_i}. \tag{29}$$

The formula (28) can be recovered by

$$H^{n+1} = \frac{1}{h}\frac{\partial F_h}{\partial u}(u^{n+1}). \tag{30}$$

The coefficient $1/h$ is due to here we define F to be the integral (3). For consistency, we also define the discrete energy as the discrete version of this integral and include an h before the discretization of the functional inside the integral.

The mass is preserved by the scheme (27)-(28).

Lemma 3.1. *The numerical scheme* (27)-(28) *satisfies the mass conservation property, i.e.*

$$\sum_{i=1}^{N} u_i^{n+1} = \sum_{i=1}^{N} u_i^n.$$

Proof. We sum (27) for $i = 1, \ldots, N$ against Δt and obtain

$$
\begin{aligned}
&\sum_{i=1}^{N} u_i^{n+1} - \sum_{i=1}^{N} u_i^n \\
=&\frac{\Delta t}{2h^2} \sum_{i=1}^{N} \left((u_{i+1}^n + u_i^n)(H_{i+1}^{n+1} - H_i^{n+1}) - (u_i^n + u_{i-1}^n)(H_i^{n+1} - H_{i-1}^{n+1}) \right) \\
=&\frac{\Delta t}{2h^2} \sum_{i=1}^{N} \Bigg(u_i^n (H_i^{n+1} - H_{i-1}^{n+1} + H_{i+1}^{n+1} - H_i^{n+1}) \\
&\qquad - u_i^n (H_i^{n+1} - H_{i-1}^{n+1} + H_{i+1}^{n+1} - H_i^{n+1}) \Bigg) \\
=&0.
\end{aligned}
$$

□

3.3. **Energy stability of the scheme.** In this section we will prove the dissipation of the discrete energy F_h. The proof is based on the variational structure of the discrete scheme. Our main result reads as

Theorem 3.2. (Unconditional energy stability) *Suppose* $\mathbb{T} = [0, 1]$, $u^n \in \mathcal{C}_{per}$ *is a periodic function on the grid* $\mathcal{C}$. *Suppose* $u_i^n > 0$ *for any* $i = 1, \ldots, N$ *and* u^{n+1} *is determined by the numerical scheme* (27) *and* (28), *then the discrete energy defined by* (29) *satisfies*

$$
F_h(u^{n+1}) \le F_h(u^n) \tag{31}
$$

for any $\Delta t > 0$.

Proof. The proof follows that of Lemma 2.3. Since the scheme (27) and (28) also have a variational structure

$$
\frac{u^{n+1} - u^n}{\Delta t} = \frac{1}{h} d_h \left(\widehat{u}^n D_h \frac{\partial F_h(u^{n+1})}{\partial u} \right),
$$

the corresponding discrete version of the equation (19) is

$$
\begin{aligned}
\langle H^{n+1}, u^n - u^{n+1} \rangle &= \langle H^{n+1}, -\Delta t d_h(\widehat{u}^n D_h H^{n+1}) \rangle \\
&= \Delta t \langle \widehat{u}^n D_h H^{n+1}, D_h H^{n+1} \rangle \ge 0.
\end{aligned} \tag{32}
$$

We also define the relative energy as

$$
F_h(u^n | u^{n+1}) = F_h(u^n) - F_h(u^{n+1}) - \langle H^{n+1}, (u^n - u^{n+1}) \rangle.
$$

Using (32), we have

$$
F_h(u^n | u^{n+1}) \le F_h(u^n) - F_h(u^{n+1}). \tag{33}
$$

We proceed to calculate the relative energy as

$$
\begin{aligned}
&hF_h(u^n|u^{n+1})\\
=&\sum_{i=1}^{N}\frac{(u_{i+1}^n-u_i^n)^2}{2u_i^n}-\sum_{i=1}^{N}\frac{(u_{i+1}^{n+1}-u_i^{n+1})^2}{2u_i^{n+1}}-h^2\sum_{i=1}^{N}H_i^{n+1}(u_i^n-u_i^{n+1})\\
=&\sum_{i=1}^{N}\Bigg(\frac{(u_{i+1}^n-u_i^n)^2}{2u_i^n}-\frac{(u_{i+1}^{n+1}-u_i^{n+1})^2}{2u_i^{n+1}}\\
&-\left(-\frac{(u_{i+1}^{n+1}-u_i^{n+1})^2}{2(u_i^{n+1})^2}-\left(\frac{u_{i+1}^{n+1}-u_i^{n+1}}{u_i^{n+1}}-\frac{u_i^{n+1}-u_{i-1}^{n+1}}{u_{i-1}^{n+1}}\right)\right)(u_i^n-u_i^{n+1})\\
=&\sum_{i=1}^{N}\Bigg(\frac{(u_{i+1}^n-u_i^n)^2}{2u_i^n}-\frac{(u_{i+1}^{n+1}-u_i^{n+1})^2}{u_i^{n+1}}+\frac{(u_{i+1}^{n+1}-u_i^{n+1})^2}{2(u_i^{n+1})^2}u_i^n\\
&+\frac{(u_{i+1}^{n+1}-u_i^{n+1})}{u_i^{n+1}}(u_i^n-u_i^{n+1})-\frac{u_{i+1}^{n+1}-u_i^{n+1}}{u_i^{n+1}}(u_{i+1}^n-u_{i+1}^{n+1})\Bigg)\\
=&\sum_{i=1}^{N}\left(\frac{(u_{i+1}^n-u_i^n)^2}{2u_i^n}+\frac{(u_{i+1}^{n+1}-u_i^{n+1})^2}{2(u_i^{n+1})^2}u_i^n-\frac{(u_{i+1}^{n+1}-u_i^{n+1})}{u_i^{n+1}}(u_{i+1}^n-u_i^n)\right)\\
=&\frac{1}{2}\sum_{i=1}^{N}u_i^n\left(\frac{u_{i+1}^n-u_i^n}{u_i^n}-\frac{u_{i+1}^{n+1}-u_i^{n+1}}{u_i^{n+1}}\right)^2\geq 0.
\end{aligned}
$$

This when combined with (33) leads to (31) and finishes the proof. □

4. Positivity-preserving of the scheme

We now proceed to prove the positivity-preserving property of the scheme (27)-(28). As was described in the introduction, we will rewrite the finite difference scheme (27)-(28) into an optimization problem with the objective function including the discretized energy F_h. Positivity of the scheme will be derived by a contradiction argument using the singularity of F_h at zero. From the definition of F_h in (29), if u_i^{n+1} vanishes at some point i, the discrete energy F_h is also expected to become infinite if $D_h u$ does not vanish at this point. But due to Theorem 3.2, F_h cannot go to infinity as long as $u_i^n > 0$ for any $i = 1, \dots, N$. Therefore, if $u_i^n > 0$ holds for any $i = 1, \dots, N$, $F_h(u^{n+1})$ will be bounded and thus excludes the possibility for the existence of vanishing points for u^{n+1}. However, this is only true when $u_{i+1}^{n+1} - u_i^{n+1} \neq 0$. To overcome this difficulty, we notice that if $u_{i+1}^{n+1} = u_i^{n+1}$ then we can take u_{i+1}^{n+1} as the vanishing point, and we can move along the grids to find $u_{i+1}^{n+1} \neq u_i^{n+1}$ unless $u_{i+1}^{n+1} = u_i^{n+1}$ for all $i = 1, \dots, N$. In the first case, the dissipation property of F_h is contradictory to the singularity of F_h at this point. In the second case, since the mass is conserved, the same value of u_i^{n+1} will be positive since $u^n > 0$. In both cases, we can conclude $u^{n+1} > 0$ as long as $u^n > 0$. To make the above statement rigorously, we will show that the optimal solution of the optimization problem reformulated from our scheme cannot vanish at some point $i \in \{1, \dots, N\}$.

The main result of this section is the following theorem:

Theorem 4.1. (Positivity-preserving) *Given $u^n \in \mathcal{C}_{\rm per}$, with $u^n > 0$, there exists a unique solution $u^{n+1} \in \mathcal{C}_{\rm per}$ to the numerical scheme* (27)-(28)*, with $u^{n+1} > 0$.*

Notice that the scheme (27) can be expressed using the operator $\mathcal{L}_{\widehat{u}^n}$ as

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = - \left(\mathcal{L}_{\hat{u}^n}(H^{n+1})\right)_i . \tag{34}$$

To proceed, we first consider the operator $\mathcal{L}_\phi$ defined by (24). For each $\phi \in \mathcal{C}_{\text{per}}$, there is a unique solution to the discrete elliptic problem ([7])

$$\mathcal{L}_\phi(f) := -d_h(\phi D_h f) = g.$$

It induces the following norm (see (26)):

$$\|g\|_{\mathcal{L}_\phi^{-1}} = \sqrt{\langle g, g\rangle_{\mathcal{L}_\phi^{-1}}} =: [\phi D_h f, D_h f]^{\frac{1}{2}}.$$

The following lemma, due to [7], will be useful in our subsequent analysis.

Lemma 4.2. *Suppose $\phi \in \mathcal{E}_{\text{per}}$ has a positive minimum. Let $g \in \mathring{\mathcal{C}}_{\text{per}}$ be bounded in L^∞, then the following estimate holds*

$$\|\mathcal{L}_\phi^{-1} g\|_{L^\infty} \le \frac{C}{\min \phi} h^{-\frac{1}{2}} L \|g\|_{L^\infty}, \tag{35}$$

where $\min \phi$ is the minimum of ϕ over the grid points and C does not depend on h.

Remark 4.1. *This lemma is a restatement of Lemma 3.2 in* [7] *and Lemma 3 in* [9].

Proof. We adapt the proof of Lemma 3.2 in [7]. Define $f := \mathcal{L}_\phi^{-1} g$, then it follows from the definition (24) that

$$(\min \phi)\|D_h f\|_{L^2}^2 \le [\phi D_h f, D_h f] = \langle -d_h(\phi D_h f), f\rangle = \langle g, f\rangle \le \|g\|_{L^2}\|f\|_{L^2}.$$

Since $g \in \mathring{\mathcal{C}}_{\text{per}}$ and so is f. The use of the discrete Poincaré inequality

$$\|f\|_{L^2} \le C\|D_h f\|_{L^2}$$

in the previous inequality leads to

$$\|D_h f\|_{L^2} \le C \frac{1}{\min \phi} \|g\|_{L^2}.$$

Using an inverse inequality leads to

$$\|f\|_{L^\infty} \le C h^{-\frac{1}{2}} \|D_h f\|_{L^2} \le \frac{C}{\min \phi} h^{-\frac{1}{2}} \|g\|_{L^2} \le \frac{C}{\min \phi} h^{-\frac{1}{2}} L \|g\|_{L^\infty},$$

where $C > 0$ is a constant indepedent of h. □

This lemma tells us that as long as $\widehat{u}^n > \delta$ for fixed small δ, $\mathcal{L}_{\widehat{u}^n}^{-1} H^{n+1}$ will be bounded if H^{n+1} is bounded.

Now we are ready to prove Theorem 4.1.

Proof of Theorem 4.1. We can invert the equation (34) to get

$$H^{n+1} = -\mathcal{L}_{\widehat{u}^n}^{-1} \left(\frac{u^{n+1} - u^n}{\Delta t}\right).$$

Therefore, the numerical solution of (27)-(28) is equivalent to the following minimization problem

$$u^{n+1} = \underset{u \in \mathcal{A}_h}{\arg\min}\, \mathcal{J}[u] \tag{36}$$

over the set

$$\mathcal{A}_h = \left\{ u \in \mathcal{C}_{\text{per}} \ : \ u > 0, \quad \sum_{i=1}^{N} u_i = \sum_{i=1}^{N} u_i^0 \right\}.$$

Here

$$\mathcal{J}[u] = \frac{1}{2\Delta t}\|u - u^n\|^2_{\mathcal{L}^{-1}_{\hat{u}^n}} + F_h(u) = \frac{1}{2\Delta t}\|u - u^n\|^2_{\mathcal{L}^{-1}_{\hat{u}^n}} + \frac{1}{2h}\sum_{i=1}^{N} \frac{(u_{i+1} - u_i)^2}{u_i}.$$

Actually, if u^* is a minimizer of the above problem, we have

$$\frac{1}{h}\frac{\partial \mathcal{J}(u, u^n)}{\partial u_j}\bigg|_{u=u^*} = \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\widehat{u}^n}(u^* - u^n))_j + H^* = 0.$$

This is equivalent to the scheme (27) with H^* given by (28) with u^{n+1} replaced by u^*.

Next we will prove the existence of the optimization problem. We consider the following domain for δ sufficiently small,

$$\mathcal{A}_{h,\delta} = \left\{ u \in \mathcal{C}_{\text{per}} : \ u \geq \delta, \quad \sum_{i=1}^{N} u_i = \sum_{i=1}^{N} u_i^0 \right\} \subset \mathbb{R}^N.$$

We can take $K = \sum_{i=1}^{N} u_i^0 > 0$ and solve the above minimization problem on $u_i \in [0, K]$. So if the optimal solution touches the boundary $u_i = K$ then the value at the other points will be less than zero, which means that the solution also touches the boundary $u = \delta$. Therefore, we only need to prove that the optimal solution does not touch the boundary $u = \delta$ for $\delta > 0$ sufficiently small.

We introduce $M = \frac{1}{N}\sum_{i=1}^{N} u_i^0$ and $w = u - M$, and reformulate the minimization problem as

$$\min \Phi(w) := \mathcal{J}(w + M) = \frac{1}{2\Delta t}\|w + M - u^n\|^2_{\mathcal{L}^{-1}_{\hat{u}^n}} + \frac{1}{2h}\sum_{i=1}^{N} \frac{(w_{i+1} - w_i)^2}{w_i + M}, \tag{37}$$

over the set

$$\mathring{K}_{h,\delta} := \{w \in \mathring{\mathcal{C}}_{per} \mid K \geq w \geq \delta - M\} \subset \mathbb{R}^N.$$

Due to the set $\mathring{K}_{h,\delta}$ being bounded, compact and convex in $\mathring{\mathcal{C}}_{\text{per}}$, the optimization problem (37) is solvable over the domain $\mathring{K}_{h,\delta}$, i.e., there exists a unique solution w to the above minimization problem. This is equivalent to the existence and uniqueness of the solutions to the numerical scheme (27)-(28). To prove the positivity-preserving property, we need to show that the minimizer cannot occur on the boundary.

We prove the above statement by contradiction. We suppose that a minimizer occurs at the boundary and then calculate the directional derivative of Φ from the minimum point. If there is a direction that the objective function Φ decreases along this direction, then we will have a contradiction and hence can conclude that the minimizer never occurs at the boundary.

First we suppose that there exists a minimizer of $\Phi(w)$ denoted w^* on the boundary so that at least at one grid point i_0,

$$w^*_{i_0} + M = \delta,$$

at which w^* also reaches its minimum. We calculate the directional derivative along $v = (v_1, \dots, v_N)$ as

$$
\begin{aligned}
&\frac{d}{ds}\Phi(w^*+sv)\Big|_{s=0} \\
&\quad= \frac{d}{ds}\left(\frac{1}{2\Delta t}\|w^*+sv+M-u^n\|^2_{\mathcal{L}^{-1}_{\widehat{u}^n}} + \frac{1}{2h}\sum_{i=1}^N \frac{(w^*_{i+1}-w^*_i+sv_{i+1}-sv_i)^2}{w^*_i+sv_i+M}\right)\Big|_{s=0} \\
&\quad= \left(\frac{1}{\Delta t}\langle \mathcal{L}^{-1}_{\widehat{u}^n}(w^*+sv+M-u^n), v\rangle + \frac{1}{h}\sum_{i=1}^N \frac{(w^*_{i+1}-w^*_i+sv_{i+1}-sv_i)(v_{i+1}-v_i)}{w^*_i+sv_i+M}\right. \\
&\qquad \left. - \frac{1}{2h}\sum_{i=1}^N \frac{(w^*_{i+1}-w^*_i+sv_{i+1}-sv_i)^2}{(w^*_i+sv_i+M)^2}v_i\right)\Big|_{s=0} \\
&\quad= \frac{1}{\Delta t}\langle \mathcal{L}^{-1}_{\widehat{u}^n}(w^*+M-u^n), v\rangle + \frac{1}{h}\sum_{i=1}^N \frac{(w^*_{i+1}-w^*_i)(v_{i+1}-v_i)}{w^*_i+M} \\
&\qquad - \frac{1}{2h}\sum_{i=1}^N \frac{(w^*_{i+1}-w^*_i)^2 v_i}{(w^*_i+M)^2}.
\end{aligned}
$$

In the above, we choose the direction v to be

$$
v_i = \begin{cases} 1, & \text{for } i = i_0, \\ 0, & \text{otherwise,} \end{cases}
$$

so that

$$
\begin{aligned}
\frac{1}{h}\frac{d}{ds}\Phi(w^*+sv)\Big|_{s=0} =& \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\widehat{u}^n}(w^*+M-u^n))_{i_0} - \frac{w^*_{i_0+1}-w^*_{i_0}}{h^2(w^*_{i_0}+M)} + \frac{w^*_{i_0}-w^*_{i_0-1}}{h^2(w^*_{i_0-1}+M)} \\
& - \frac{(w^*_{i_0+1}-w^*_{i_0})^2}{2h^2(w^*_{i_0}+M)^2}.
\end{aligned}
$$

Since $w^*_{i_0}$ is the minimum point, so the last three terms are all non-positive. Using inequality (35) from Lemma 4.2, we have

$$
\begin{aligned}
\frac{1}{h}\frac{d}{ds}\Phi(w^*+sv)\Big|_{s=0} \le& \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\widehat{u}^n}(w^*+M-u^n))_{i_0} - \frac{w^*_{i_0+1}-w^*_{i_0}}{h^2(w^*_{i_0}+M)} \\
\le& \frac{C}{\Delta t \min \widehat{u}^n} h^{-\frac{1}{2}} L\|w^*+M-u^n\|_{L^\infty} - \frac{1}{h^2}\frac{w^*_{i_0+1}-w^*_{i_0}}{\delta}. \qquad (38)
\end{aligned}
$$

We take $w^*_{i_0+1} \neq w^*_{i_0}$. If $w^*_{i_0+1} = w^*_{i_0}$, we can take i_0+1 instead of i_0 and check whether $w_{i_0+2} \neq w_{i_0+1}$ holds. We can move along the grids and find an index i_0 such that $w^*_{i_0+1} \neq w^*_{i_0}$. Otherwise, all the values $w^*_i = \delta - M$ for all $i = 1, \dots, N$ are equal, but this is not possible for δ sufficiently small, since the average of $\{w^*_i,\, i = 1, \dots, N\}$ is 0 and $M > 0$. Using the fact $\widehat{u}^n > 0$ and taking

$$
\delta^{-1} > \frac{Ch^{\frac{3}{2}} L\|w^*+M-u^n\|_{L^\infty}}{\Delta t \min \widehat{u}^n} \cdot \frac{1}{w^*_{i_0+1}-w^*_{i_0}},
$$

then leads to

$$
\frac{d}{ds}\Phi(w^*+sv)\Big|_{s=0} < 0, \qquad (39)
$$

which is contradictory to the fact that w_* is a minimizer. So the minimizer of the functional (36) cannot be on the boundary of $\mathcal{A}_{h,\delta}$ and therefore $u^{n+1} > 0$ holds if $u^n > 0$. □

Remark 4.2. *Notice that for δ being constant, (39) may not hold when $h, \Delta t$ become small. For example, if $\Delta t \sim h^2$, then the first term in (38) will dominate as h tends to zero, thus making the contradiction argument a failure. Thus we cannot prove if $u^n > \delta$ then $u^{n+1} > \delta$ the same way as we did above. Due to this reason, it may not be possible to pass to the limit $h \to 0$ in order to prove that the solution to the continuous equation (1) is positive.*

Remark 4.3. *Note that a more natural scheme may be the central discretization by*

$$H = -\frac{|D_h u|^2}{2\hat{u}^2} - d_h\left(\frac{D_h u}{\hat{u}}\right),$$

which corresponds to a discretized energy

$$F_h(u) := \frac{1}{2}\left[\frac{D_h u}{\hat{u}}, D_h u\right] = \frac{h}{2}\sum_{i=1}^{N}\frac{|(D_h u)_{i+\frac{1}{2}}|^2}{\hat{u}_{i+\frac{1}{2}}} = \frac{1}{h}\sum_{i=1}^{N}\frac{(u_{i+1}-u_i)^2}{(u_{i+1}+u_i)}.$$

However, we cannot use the same contradiction argument to prove the positivity-preserving property for this scheme, since the denominator $u_{i+1}+u_i$ may not be 0 if $u_i = 0$ when the minimizer touches the boundary and (39) cannot be established.

Remark 4.4. *Other forms of discretization are also feasible. For example, we can take*

$$F_h(u) = \frac{1}{h}\sum_{i=1}^{N}\frac{(u_i - u_{i-1})^2}{2u_i}, \tag{40}$$

and H to be

$$H(u) = -\frac{|d_h u|^2}{2u^2} - D_h\left(\frac{d_h u}{u}\right) = -\frac{1}{2h^2}\frac{(u_i-u_{i-1})^2}{u_i^2} - \frac{1}{h^2}\left(\frac{u_{i+1}-u_i}{u_{i+1}} - \frac{u_i - u_{i-1}}{u_i}\right),$$

then the resulting scheme will also have properties stated in Theorem 3.2 and Theorem 4.1. Furthermore, we can also take F_h to be a combination of (40) and (29) to get a symmetric scheme. That is

$$F_h(u) = \frac{1}{4h}\sum_{i=1}^{N}\left(\frac{(u_{i+1}-u_i)^2 + (u_i - u_{i-1})^2}{u_i}\right)$$

with

$$\begin{aligned} H(u) = &-\frac{1}{4h^2}\left(\frac{(u_{i+1}-u_i)^2 + (u_i-u_{i-1})^2}{u_i^2}\right) \\ &-\frac{1}{2h^2}\left(\frac{u_{i+1}-u_i}{u_i} + \frac{u_{i+1}-u_i}{u_{i+1}} - \frac{u_i-u_{i-1}}{u_i} - \frac{u_i-u_{i-1}}{u_{i-1}}\right). \end{aligned}$$

Theorem 3.2 and Theorem 4.1 also hold true.

Remark 4.5. *The minimization problem* (36) *can be viewed as the discretization of the optimization problem in* (8)*, which is*

$$u^{n+1} = \left\{ \underset{u,\phi}{\arg\inf} \int_{\mathbb{T}^d} \frac{1}{2} u^n |\nabla \phi|^2 dx + \Delta t F(u) \right.$$
$$\left. s.t.\ u - u^n(x) + \nabla \cdot (u^n \nabla \phi) = 0 \right\}.$$

It can be discretized into

$$u^{n+1} = \min_{u \in \mathcal{C}_{\text{per}}} \left([\hat{u}^n D_h \phi, D_h \phi] + \Delta t F_h(u) \right),$$

where ϕ is defined to be the numerical solution of

$$u - u^n + d_h(\hat{u}^n D_h \phi) = 0.$$

From the definition of the operator $\mathcal{L}_{\hat{u}^n}^{-1}$ by (24)*, the above discretize minimization problem is the same as* (36)*.*

5. Consistency of the scheme

Since the semi-implicit scheme (27)-(28) is a nonlinear equation, the convergence of the scheme is beyond the scope of this paper. However, here we show the consistency of the scheme. Suppose $u = u(x,t)$ is a smooth solution to the equation (1) and u_j^n is a finite difference solution to (27)-(28), we will show that the local truncation error defined by

$$\tau_j^n := \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t} - d_h \left(\hat{u}(x, t_n) D_h H^{n+1}(u(x, t_{n+1})) \right)_{x_j} \tag{41}$$

converges to 0 as $\Delta t, h \to 0$.

The local truncation error τ_j^n can be computed using Taylor's expansion. We will not give the details of the calculations here. The result is

$$\begin{aligned}\tau_j^n = & -h \left(-\frac{3{u_x}^5}{u^4} + \frac{6u_{xx}u_x^3}{u^3} - \frac{3u_{xxx}u_x^2}{2u^2} - \frac{3u_{xx}^2 u_x}{2u^2} \right)\Bigg|_{(x_j,t_n)} - k \Bigg(\frac{3u_t u_x^4}{u^4} - \frac{3u_{xt}u_x^3}{u^3} \\ & + \frac{2u_{xxt}u_x^2}{u^2} - \frac{4u_t u_{xx} u_x^2}{u^3} + \frac{2u_{xt}u_{xx}u_x}{u^2} + \frac{u_t u_{xxx} u_x}{u^2} - \frac{u_{xxxt}u_x}{u} \Bigg)\Bigg|_{(x_j,t_n)} + o(h+k).\end{aligned}$$

We have the following theorem.

Theorem 5.1. *Suppose that the solution $u = u(x,t)$ to the equation* (1) *is positive and smooth, then the consistency error for the numerical scheme* (27)-(28)*, defined by* (41)*, satisfies*

$$\|\tau^n\|_{L^\infty} \le C \min\{h, k\} |u|_{\mathcal{C}^4_{x,t}}$$

6. The scheme in higher dimensions

The formulation of the numerical scheme (27)-(28) depends on the gradient flow structure of the equation (1) and can be generalized to higher dimensions. The key to obtain such a scheme is to use the explicit-implicit scheme (17) and discretize

H^{n+1} by using a discrete gradient defined by (30). For example, in the case of two dimensions, if we define

$$F_h(u) = \frac{1}{2}\sum_{i,j=1}^{N}\left(\frac{(u_{i+1,j}-u_{i,j})^2}{u_{i,j}} + \frac{(u_{i,j+1}-u_{i,j})^2}{u_{i,j}}\right), \tag{42}$$

we can get H^{n+1} using

$$\begin{aligned} H_{i,j} =& \frac{1}{h^2}\frac{\partial F_h(u)}{\partial u_{i,j}} \\ =& -\frac{1}{2h^2}\left(\frac{(u_{i+1,j}-u_{i,j})^2}{u_{i,j}^2} + \frac{(u_{i,j+1}-u_{i,j})^2}{u_{i,j}^2}\right) \\ & -\frac{1}{h^2}\left(\frac{u_{i+1,j}-u_{i,j}}{u_{i,j}} - \frac{u_{i,j}-u_{i-1,j}}{u_{i-1,j}} + \frac{u_{i,j+1}-u_{i,j}}{u_{i,j}} - \frac{u_{i,j}-u_{i,j-1}}{u_{i,j-1}}\right). \end{aligned} \tag{43}$$

Then the resulting scheme will satisfy all the properties proved for the one dimensional case above, including mass conservation, energy stability, positivity-preserving, and the consistency. We will present detailed proofs of these properties for the two dimensional case only; for higher dimensions, it works in the same way. Let use first introduce some notations.

6.1. **Notations.** We take the domain to be $\mathbb{T}^2 = [0,L]\times[0,L]$, and the grid spacing $h = L/N$ for both x and y. We label the grid points with i,j for $i = 1,\ldots,N$ and $j = 1,\ldots,N$. Since we are taking a periodic grid, we extend the label to whole $\mathbb{Z}\times\mathbb{Z}$ by using $\mathcal{C}\times\mathcal{C}$ with $\mathcal{C}$ defined in (20) and also label the mid-point grids using $\mathcal{E}\times\mathcal{E}$ with $\mathcal{E}$ defined in (20). We define the periodic function spaces:

$$\mathcal{C}_{\text{per}} := \{f : \mathcal{C}\times\mathcal{C}\to\mathbb{R}\},\quad \mathcal{E}_{\text{per}}^X := \{f : \mathcal{E}\times\mathcal{C}\to\mathbb{R}\},\quad \mathcal{E}_{\text{per}}^Y := \{f : \mathcal{C}\times\mathcal{R}\to\mathbb{R}\},$$

and also the function space with zero mean:

$$\mathring{\mathcal{C}}_{\text{per}} := \left\{ f\in\mathcal{C}_{\text{per}} : \sum_{i,j=1}^{N} f_{i,j} = 0\right\}.$$

We define the difference operator D_x, D_y and d_x, d_y as

$$\begin{aligned} (D_x f)_{i+\frac{1}{2},j} &= \frac{f_{i+1,j}-f_{i,j}}{h}, \quad (D_y f)_{i,j+\frac{1}{2}} = \frac{f_{i,j+1}-f_{i,j}}{h}, \\ (d_x f)_{i,j} &= \frac{f_{i+\frac{1}{2},j}-f_{i-\frac{1}{2},j}}{h}, \quad (d_y f)_{i,j} = \frac{f_{i,j+\frac{1}{2}}-f_{i,j-\frac{1}{2}}}{h}. \end{aligned}$$

We also define the discrete gradient D_h as

$$(D_h f)_{i,j} = (D_x f_{i+\frac{1}{2},j}, D_y f_{i,j+\frac{1}{2}}),$$

and the discrete divergence d_h as

$$d_h\cdot\mathbf{f}_{i,j} = d_x f^x_{i,j} + d_y f^y_{i,j},\quad \mathbf{f}_{i,j} = (f^x, f^y)\in\mathcal{E}_{\text{per}}^X\times\mathcal{E}_{\text{per}}^Y.$$

The inner products on the grids are defined by

$$\langle f,g\rangle := h^2\sum_{i,j=1}^{N} f_{i,j}g_{i,j},\ \forall f,g\in\mathcal{C}_{\rm per},\quad [f,g]_x := \frac{1}{2}h^2\sum_{i,j=1}^{N}\left(f_{i+\frac12,j}g_{i+\frac12,j}+f_{i-\frac12,j}g_{i-\frac12,j}\right),$$

$$[f,g]_y := \frac{1}{2}h^2\sum_{i,j=1}^{N}\left(f_{i,j+\frac12}g_{i,j+\frac12}+f_{i,j-\frac12}g_{i,j-\frac12}\right),$$

and for vector functions:

$$[\mathbf{f},\mathbf{g}] := [f^x,g^x]_x+[f^y,g^y]_y.$$

And the corresponding norms in these spaces are defined accordingly. Suppose $f,g\in\mathcal{C}_{\rm per}$, and ϕ is defined on all the edge-center points $\mathcal{C}\times\mathcal{E}\cup\mathcal{E}\times\mathcal{C}$, then the following summation-by-parts formulas hold:

$$\langle f, d_h\cdot\mathbf{g}\rangle = -[D_hf,\mathbf{g}],\quad \langle f, d_h\cdot(\phi D_hg)\rangle = -[D_hf,\phi D_hg].$$

We introduce the following operator $\mathcal{L}$ on $\mathring{\mathcal{C}}_{\rm per}$:

$$\mathcal{L}_\phi f = -d_h\cdot(\phi D_h f) = g, \tag{44}$$

with $f,g\in\mathring{\mathcal{C}}_{\rm per}$ and ϕ defined on all edge-center points. With this operator, we can define the bilinear form as before:

$$\langle g_1,g_2\rangle_{\mathcal{L}_\phi^{-1}} := \langle g_1,\mathcal{L}_\phi^{-1}g_2\rangle = \langle\mathcal{L}_\phi^{-1}g_1,g_2\rangle = [\phi D_hf_1,D_hf_2],$$

and also the norm (26).

6.2. **The scheme.** With the notations above, we can now write our scheme similar as (27)-(28):

$$\frac{u^{n+1}-u^n}{\Delta t} = d_h\cdot(\hat{u}^nD_hH^{n+1}), \tag{45}$$

where $\hat{u}^n$ is defined on all the edge-center points with

$$\hat{u}^n_{i+\frac12,j} = \frac{u^n_{i,j}+u^n_{i+1,j}}{2},\quad \hat{u}_{i,j+\frac12} = \frac{u^n_{i,j}+u^n_{i,j+1}}{2},$$

and H^{n+1} is the same as (43):

$$\begin{aligned} H^{n+1}_{i,j} =& -\frac{1}{2h^2}\left(\frac{(u^{n+1}_{i+1,j}-u^{n+1}_{i,j})^2}{(u^{n+1}_{i,j})^2}+\frac{(u^{n+1}_{i,j+1}-u^{n+1}_{i,j})^2}{(u^{n+1}_{i,j})^2}\right)\\ &-\frac{1}{h^2}\left(\frac{u^{n+1}_{i+1,j}-u^{n+1}_{i,j}}{u^{n+1}_{i,j}}-\frac{u^{n+1}_{i,j}-u^{n+1}_{i-1,j}}{u^{n+1}_{i-1,j}}+\frac{u^{n+1}_{i,j+1}-u^{n+1}_{i,j}}{u^{n+1}_{i,j}}-\frac{u^{n+1}_{i,j}-u^{n+1}_{i,j-1}}{u^{n+1}_{i,j-1}}\right). \end{aligned} \tag{46}$$

We can also write (45) in the coordinate form as

$$\begin{aligned} \frac{u^{n+1}_{i,j}-u^n_{i,j}}{\Delta t} =& \frac{1}{2h^2}\left((u^n_{i+1,j}+u^n_{i,j})(H^{n+1}_{i+1,j}-H^n_{i,j})-(u^n_{i,j}+u^n_{i-1,j})(H^{n+1}_{i,j}-H^n_{i-1,j})\right)\\ &+\frac{1}{2h^2}\left((u^n_{i,j+1}+u^n_{i,j})(H^{n+1}_{i,j+1}-H^n_{i,j})-(u^n_{i,j}+u^n_{i,j-1})(H^{n+1}_{i,j}-H^n_{i,j-1})\right). \end{aligned}$$

6.3. **Properties of the scheme.** Since the scheme (45)-(46) follows the same structure of the scheme (27)-(28) in one dimension case, all the properties proved in the previous sections hold true. That is we have the following theorem.

Theorem 6.1. *The numerical scheme* (45)-(46) *satisfies:*

(1) Mass conservation: for any $n \in \mathbb{N}$*,*

$$\sum_{i,j=1}^{N} u_{i,j}^{n} = \sum_{i,j=1}^{N} u_{i,j}^{0}. \tag{47}$$

(2) Energy stability: suppose $u^n > 0$*, then*

$$F_h(u^{n+1}) \leq F_h(u^n) \tag{48}$$

holds with F_h *defined by* (42)*.*

(3) Positivity-preserving: if $u^n > 0$*, then there exists a unique positive solution* u^{n+1} *to the numerical scheme* (45)-(46) *with* $u^{n+1} > 0$*.*

(4) Consistency: when $h \to 0$*, the consistency error goes to zero, with*

$$\|\tau^n\|_{L^\infty} \leq C \min\{h, \Delta t\} |u|_{\mathcal{C}^4_{x,y,t}},$$

where τ *is the local truncation error defined by*

$$\begin{aligned}\tau_{i,j}^n := & \frac{u(x_i, y_j, t_{n+1}) - u(x_i, y_j, t_n)}{\Delta t} \\ & - d_h \cdot \left(\hat{u}(x, y, t_n) D_h H^{n+1}(u(x, y, t_{n+1}))\right)_{x_i, y_j}.\end{aligned}$$

Proof. We only need to slightly modify the previous proofs to prove this theorem.

(1) The conservation of mass can be verified by a direct verification as done in the proof of Lemma 3.1:

$$\begin{aligned}&\sum_{i,j=1}^{N} u_{i,j}^{n+1} - \sum_{i,j=1}^{N} u_{i,j}^{n} \\ =&\frac{\Delta t}{2h^2} \sum_{j=1}^{N}\sum_{i=1}^{N} \left((u_{i+1,j}^n + u_{i,j}^n)(H_{i+1,j}^{n+1} - H_{i,j}^n) - (u_{i,j}^n + u_{i-1,j}^n)(H_{i,j}^{n+1} - H_{i-1,j}^n)\right) \\ &+ \frac{\Delta t}{2h^2} \sum_{i=1}^{N}\sum_{j=1}^{N} \left((u_{i,j+1}^n + u_{i,j}^n)(H_{i,j+1}^{n+1} - H_{i,j}^n) - (u_{i,j}^n + u_{i,j-1}^n)(H_{i,j}^{n+1} - H_{i,j-1}^n)\right) \\ =&0.\end{aligned}$$

(2) The second property follows from the variational structure. Due to

$$\langle H^{n+1}, u^n - u^{n+1} \rangle = \langle H^{n+1}, -\Delta t d_h \cdot (\widehat{u}^n D_h H^{n+1}) \rangle = \Delta t \left[\widehat{u}^n D_h H^{n+1}, D_h H^{n+1}\right] \geq 0,$$

the inequality (48) follows if we can prove

$$F_h(u^n | u^{n+1}) = F(u^n) - F(u^{n+1}) - \langle H^{n+1}, u^n - u^{n+1} \rangle \geq 0,$$

which is (we omit the calculations here)

$$F_h(u^n|u^{n+1}) = \frac{1}{2}\sum_{i,j=1}^{N} u^n_{i,j}\left[\left(\frac{u^n_{i+1,j}-u^n_{i,j}}{u^n_{i,j}} - \frac{u^{n+1}_{i+1,j}-u^{n+1}_{i,j}}{u^{n+1}_{i,j}}\right)^2 + \left(\frac{u^n_{i,j+1}-u^n_{i,j}}{u^n_{i,j}} - \frac{u^{n+1}_{i,j+1}-u^{n+1}_{i,j}}{u^{n+1}_{i,j}}\right)^2\right]$$

being nonnegative if $u^n > 0$.

(3) Now we prove the positivity-preserving property. Similar as (36), the scheme (45)-(46) is equivalent to the minimization problem (36), with the norms replaced by the corresponding norms in two dimensions and F_h defined by (42).

$$u^{n+1} = \arg\min_{u\in\mathcal{A}_h} \mathcal{J}[u]$$

over the set

$$\mathcal{A}_h = \left\{ u \in \mathcal{C}_{\text{per}} \;:\; u > 0, \quad \sum_{i,j=1}^{N} u_{i,j} = \sum_{i,j=1}^{N} u^0_{i,j} \right\}.$$

Here again

$$\mathcal{J}[u] = \frac{1}{2\Delta t}\|u-u^n\|^2_{\mathcal{L}^{-1}_{\hat{u}^n}} + F_h(u).$$

To proceed, we introduce $M = \frac{1}{N^2}\sum_{i,j=1}^{N} u^0_{i,j}$ and $w = u - M$ and reformulate the minimization problem similar to (37) as

$$\begin{aligned}\min \Phi(w) :=& \mathcal{J}(w+M)\\ =& \frac{1}{2\Delta t}\|w+M-u^n\|^2_{\mathcal{L}^{-1}_{\hat{u}^n}} + \frac{1}{2}\sum_{i,j=1}^{N}\frac{(w_{i+1,j}-w_{i,j})^2 + (w_{i,j+1}-w_{i,j})^2}{w_{i,j}+M},\end{aligned}$$

over the set

$$\mathring{K}_{h,\delta} := \{w \in \mathring{\mathcal{C}}_{\text{per}} : K \geq w \geq \delta - M\} \subset \mathbb{R}^{N^2}.$$

The existence of the minimization problem follows from the fact that $\mathring{K}_{h,\delta}$ is a bounded, compact and convex subset of $\mathbb{R}^{N^2}$. The positivity-preserving property will follow if we can prove that the minimizer to the above optimization problem does not occur on the boundary.

We prove by contradiction. If the minimizer w^* touches the boundary, at least at one point x_0, j_0,

$$w^*_{i_0,j_0} + M = \delta.$$

We calculate the directional derivative along the direction $v = (v_{i,j})_{N\times N}$ as

$$\begin{aligned}
&\frac{d}{ds}\Phi(w^* + sv)\Big|_{s=0} \\
=&\frac{d}{ds}\Bigg(\frac{1}{2\Delta t}\|w^* + sv + M - u^n\|^2_{\mathcal{L}^{-1}_{\widehat{u}^n},h} \\
&+ \frac{1}{2}\sum_{i=1}^{N}\frac{(w^*_{i+1,j} - w^*_{i,j} + sv_{i+1,j} - sv_{i,j})^2 + (w^*_{i,j+1} - w^*_{i,j} + sv_{i,j+1} - sv_{i,j})^2}{w^*_{i,j} + sv_{i,j} + M}\Bigg)\Bigg|_{s=0} \\
=&\Bigg(\frac{1}{\Delta t}\langle\mathcal{L}^{-1}_{\widehat{u}^n}(w^* + M - u^n), v\rangle + \sum_{i=1}^{N}\frac{(w^*_{i+1,j} - w^*_{i,j})(v_{i+1,j} - v_{i,j})}{w^*_{i,j} + M} \\
&+ \sum_{i=1}^{N}\frac{(w^*_{i,j+1} - w^*_{i,j})(v_{i,j+1} - v_{i,j})}{w^*_{i,j} + M} - \sum_{i=1}^{N}\frac{(w^*_{i+1,j} - w^*_{i,j})^2 + (w^*_{i,j+1} - w^*_{i,j})^2}{2(w^*_{i,j} + M)^2}v_{i,j}\Bigg).
\end{aligned}$$

We take

$$v_{i,j} = \begin{cases} 1, & \text{for } i = i_0, j = j_0, \\ 0, & \text{otherwise,}\end{cases}$$

and the above equation becomes

$$\begin{aligned}
&\frac{1}{h^2}\frac{d}{ds}\Phi(w^* + sv)\Big|_{s=0} \\
&=\frac{1}{\Delta t}(\mathcal{L}^{-1}_{\widehat{u}^n}(w^* + M - u^n))_{i_0,j_0} - \frac{1}{h^2}\frac{w^*_{i_0+1,j_0} - w^*_{i_0,j_0}}{w^*_{i_0,j_0} + M} + \frac{1}{h^2}\frac{w^*_{i_0,j_0} - w^*_{i_0-1,j_0}}{w^*_{i_0-1,j_0} + M} \\
&\quad - \frac{1}{h^2}\frac{w^*_{i_0,j_0+1} - w^*_{i_0,j_0}}{w^*_{i_0,j_0} + M} + \frac{1}{h^2}\frac{w^*_{i_0,j_0} - w^*_{i_0,j_0-1}}{w^*_{i_0,j_0-1} + M} \\
&\quad - \frac{1}{h^2}\frac{(w^*_{i_0+1,j_0} - w^*_{i_0,j_0})^2 + (w^*_{i_0,j_0+1} - w^*_{i_0,j_0})^2}{2(w^*_{i_0,j_0} + M)^2}. \qquad (49)
\end{aligned}$$

For w^*, we can find an index i_0, j_0 that at least one of the following conditions holds

$$\text{(i) } w^*_{i_0+1,j_0} \neq w^*_{i_0,j_0} \quad \text{or} \quad \text{(ii) } w^*_{i_0,j_0+1} \neq w^*_{i_0,j_0},$$

otherwise, for example if $w^*_{i_0+1,j_0} = w^*_{i_0,j_0+1} = w^*_{i_0,j_0}$, we can replace $w^*_{i_0,j_0}$ by $w^*_{i_0+1,j_0}$ or $w^*_{i_0,j_0+1}$ and check the two inequalities. It is possible to move along the grids to find an index i_0, j_0 satisfies the above conditions. If such an index does not exist, then w^* has a uniform value across all grid points and its sum will be negative for sufficiently small δ, which contradicts the fact that the sum of $w^*_{i,j}$ for all $i, j = 1, \ldots, N$ is zero. Notice that Lemma 4.2 holds for any dimensions (see [7, 9]) and the fact that $w^*_{i_0,j_0}$ is the minimization point, we can estimate (49) by

$$\begin{aligned}
\frac{1}{h^2}\frac{d}{ds}\Phi(w^* + sv)\Big|_{s=0} &\leq \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\widehat{u}^n}(w^* + M - u^n))_{i_0,j_0} - \frac{1}{h^2}\frac{w^*_{i_0+1,j_0} - w^*_{i_0,j_0}}{w^*_{i_0,j_0} + M} \\
&\leq \frac{C}{\Delta t \min \widehat{u}^n}h^{-\frac{1}{2}}L\|w^* + M - u^n\|_{L^\infty} - \frac{1}{h^2}\frac{w^*_{i_0+1,j_0} - w^*_{i_0,j_0}}{\delta}, \qquad (50)
\end{aligned}$$

in the case (i), and by

$$\frac{1}{h^2}\frac{d}{ds}\Phi(w^*+sv)\Big|_{s=0} \leq \frac{1}{\Delta t}(\mathcal{L}_{\widehat{u}^n}^{-1}(w^*+M-u^n))_{i_0,j_0} - \frac{1}{h^2}\frac{w^*_{i_0,j_0+1}-w^*_{i_0,j_0}}{w^*_{i_0,j_0}+M}$$

$$\leq \frac{C}{\Delta t \min \widehat{u}^n} h^{-\frac{1}{2}} L\|w^*+M-u^n\|_{L^\infty} - \frac{1}{h^2}\frac{w^*_{i_0,j_0+1}-w^*_{i_0,j_0}}{\delta}. \tag{51}$$

in the case (ii). We can choose

$$\delta^{-1} = \frac{2Ch^{\frac{3}{2}}L\|w^*+M-u^n\|_{L^\infty}}{\Delta t \min \widehat{u}^n} \cdot \frac{1}{w^*_{i_0+1,j_0}-w^*_{i_0,j_0}}$$

in (50) and

$$\delta^{-1} = \frac{2Ch^{\frac{3}{2}}L\|w^*+M-u^n\|_{L^\infty}}{2\Delta t \min \widehat{u}^n} \cdot \frac{1}{w^*_{i_0,j_0+1}-w^*_{i_0,j_0}}$$

in (51) and in both cases

$$\frac{d}{ds}\Phi(w^*+sv)\Big|_{s=0} < 0,$$

so w^* cannot be a minimizer, which contradicts our assumption. Therefore, following the same arguement of the proof of Theorem 4.1, we conclude that the numerical scheme (45)-(46) is positivity-preserving.

(4) The consistency error can be computed by using the Taylor's expansion directly. We omit the details here. □

7. Numerical examples

We will present some numerical examples for the schemes above, in both one dimension and two dimensions. For efficiently numerical computation of nonlinear systems in each step, we use the DF-SANE method developed in [20]. The mass conservation, energy dissipation, positivity-preserving and convergence rates are verified numerically.

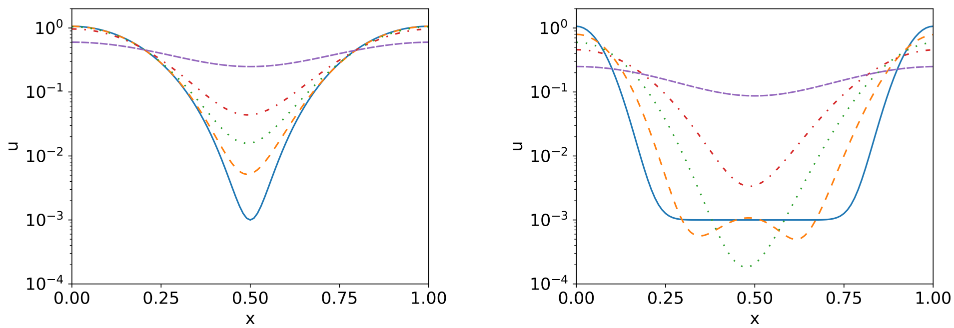
7.1. **One dimension example.**

First we consider a numerical example used in the paper [3]. We take $\mathbb{T} = [0,1]$ and the initial condition

$$u_0(x) = \left(\varepsilon^{\frac{1}{2}} + \left[\frac{1+\cos 2\pi x}{2}\right]^m\right)^2. \tag{52}$$

Here ε is taken to be 0.001 and $m = 1$ and 8. Plots are made for $t = 0.0$, $t = 8\times 10^{-6}$, $t = 3.2\times 10^{-5}$, $t = 1\times 10^{-4}$ and $t = 7.2\times 10^{-4}$ with solid, dashed, dots, dash-dot and dash-dash lines, respectively. The solution u is plotted in Figure 1. Mass variations, energy and minimum values are plotted in Figure 2 over time. These three figures verify the mass conservation, energy stability, and positivity of solutions, respectively. It can be seen that although in some cases the minimum value may be decreasing in some time interval and the maximum principle fails, the minimum value remains positive.

We also preform error computations for different mesh sizes. We take $\Delta t = 1.6\times 10^{-8}h$ and take the number of mesh intervals to be $10, 20, 40, 80, 160$, with spacing $h = 0.1, 0.05, 0.025, 0.0125, 0.00625$, respectively. We compare the solutions at time $t = 7.2\times 10^{-4}$. Since we do not have a reference solution, we take the last



$m = 1$ $\qquad$ $m = 8$

FIGURE 1. Numerical solutions to the equation (1) in 1D

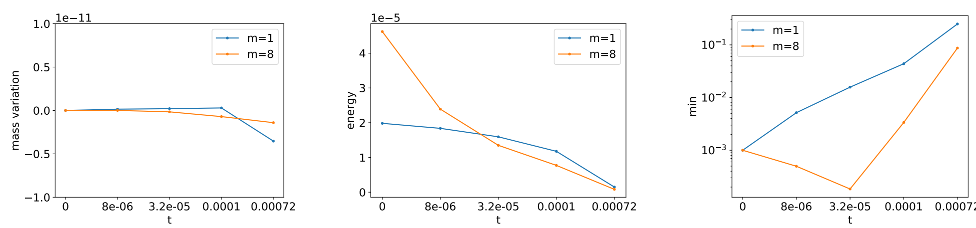


FIGURE 2. Mass variations, energy and minimum values of u

result $\widetilde{u}$ ($h = 0.00625$) as a reference solution. We compute the errors based on the fact for a numerical scheme of p-th order,

$$\frac{u_h - \widetilde{u}}{u_{h/2} - \widetilde{u}} = 2^p + O(h).$$

We interpolate u_h using the nearest neighbor method in space to find its difference with $\widetilde{u}$ and calculate the l^2 error. The result is plotted in Figure 3. The solid line is the error with respect to the reference solution and the dashed line is the line $1/h$. Thus the numerical scheme is approximately of order 1 in h.

7.2. **Two dimension example.** We compute the 2D problem on the torus $\mathbb{T}^2 = [0,1] \times [0,1]$ with the following initial conditions

$$u_0(x,y) = \left(0.001^{\frac{1}{2}} + \left[\frac{1 + \cos 2\pi x \cdot \cos 2\pi y}{2}\right]^8\right)^2.$$

This initial condition is chosen resembling the corresponding initial condition (52) in 1D with $m = 8$. However, there is one peak inside the domain $\mathbb{T}^2$ at point $x = y = \frac{1}{2}$. So the diffusion of this inner point and the four peaks at the boundary dominate the behavior of the solutions.

The solutions and their energy functions as well as minimum values are plotted over time in Figure 4. The minimum value decreases first and then increases over time but remains positive. We also perform error computations for different mesh

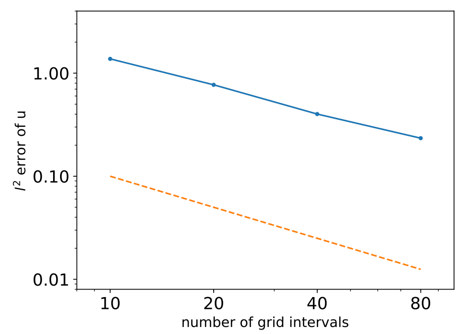
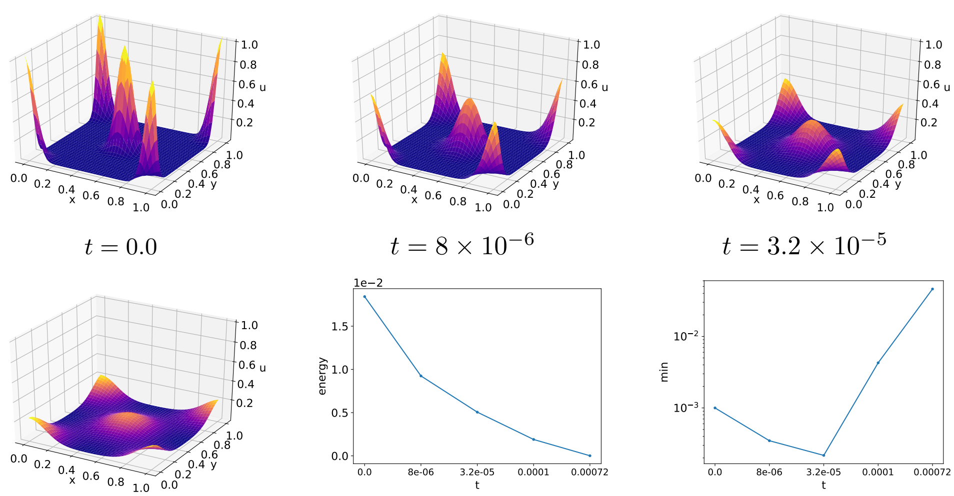


FIGURE 3. Error and convergence order



$t = 0.0$ $t = 8 \times 10^{-6}$ $t = 3.2 \times 10^{-5}$

$t = 7.2 \times 10^{-4}$ Energy Minimum value

FIGURE 4. Numerical results to the equation (1) in 2D

sizes in 2D. The result is plotted in Figure 5. It is similar with the 1D example and the numerical convergence order is approximately one in h.

ACKNOWLEDGEMENT

The first author is funded by KAUST. The second author is funded by the National Science Foundation under Grant DMS1812666. The authors are grateful to Athanasios E. Tzavaras for valuable suggestions and comments.

REFERENCES

[1] M. ANCONA, *Diffusion-drift modeling of strong inversion layers*, COMPEL, 6 (1987), pp. 11–18.

[2] J.D. BENAMOU, AND Y. BRENIER, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math., 84(3) (2000), pp. 375–393.

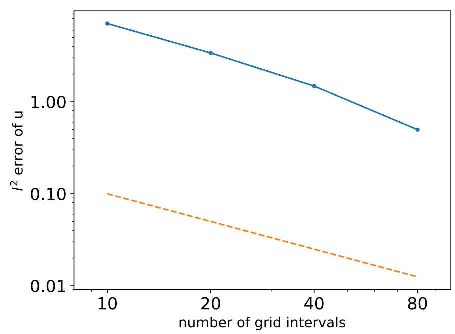


FIGURE 5. Error and convergence order in 2D

[3] P. M. Bleher, J. L. Lebowitz, and E. R. Speer, *Existence and positivity of solutions of a fourth-order nonlinear PDE describing interface fluctuations*, Commun. Pure Appl. Math., 47 (1994), pp. 923–942.

[4] M. Bukal, E. Emmrich, and A. Jüngel, *Entropy-stable and entropy-dissipative approximations of a fourth-order quantum diffusion equation*, Numer. Math., 127 (2014), pp. 365–396.

[5] M. Cáceres, J. Carrillo, and G. Toscani, *Long-time behavior for a nonlinear fourth-order parabolic equation*, Trans. Amer. Math. Soc., 357 (2005), pp. 1161–1175.

[6] J. A. Carrillo, A. Jüngel, and S. Tang, *Positive entropic schemes for a nonlinear fourth-order parabolic equation*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 1–20.

[7] W. Chen, C. Wang, X. Wang, and S. M. Wise, *Positivity-preserving, energy stable numerical schemes for the Cahn-Hilliard equation with logarithmic potential*, J. Comput. Phys. X, 3 (2019), 100031.

[8] B. Derrida, J. L. Lebowitz, E. R. Speer, and H. Spohn, *Fluctuations of a stationary nonequilibrium interface*, Phys. Rev. Lett., 67 (1991), pp. 165–168.

[9] L. Dong, C. Wang, H. Zhang, and Z. Zhang, *A positivity-preserving, energy stable and convergent numerical scheme for the Cahn–Hilliard equation with a Flory–Huggins–Degennes energy*, Commun. Math. Sci., (2018).

[10] B. Düring, D. Matthes, and J. P. Milišic, *A gradient flow scheme for nonlinear fourth order equations*, Discrete Contin. Dyn. Syst. Ser. B, 14 (2010), pp. 935–959.

[11] I. Gasser, P. Markowich, D. Schmidt, and A. Unterreiter, *Macroscopic theory of charged quantum fluids*, Longman, 1995.

[12] U. Gianazza, G. Savaré, and G. Toscani, *The Wasserstein gradient flow of the Fisher information and the quantum drift-diffusion equation*, Arch. Ration. Mech. Anal., 194 (2009), pp. 133–220.

[13] J. Giesselmann, C. Lattanzio, and A. E. Tzavaras, *Relative energy for the Korteweg theory and related Hamiltonian flows in gas dynamics*, Arch. Ration. Mech. Anal., 223 (2017), pp. 1427–1484.

[14] J. Giesselmann and A. E. Tzavaras, *Stability properties of the Euler–Korteweg system with nonmonotone pressures*, Appl. Anal., 96 (2017), pp. 1528–1546.

[15] A. Jüngel and D. Matthes, *A review on results for the Derrida-Lebowitz-Speer-Spohn equation*, Proc. of Equa. Diff07, (2007).

[16] ———, *The Derrida–Lebowitz–Speer–Spohn equation: Existence, nonuniqueness, and decay rates of the solutions*, SIAM J. Math. Anal., 39 (2008), pp. 1996–2015.

[17] A. Jüngel and R. Pinnau, *Global nonnegative solutions of a nonlinear fourth-order parabolic equation for quantum systems*, SIAM J. Math. Anal., 32 (2000), pp. 760–777.

[18] ———, *A positivity-preserving numerical scheme for a nonlinear fourth order parabolic system*, SIAM J. Numer. Anal., 39 (2001), pp. 385–406.

[19] A. Jüngel and G. Toscani, *Exponential time decay of solutions to a nonlinear fourth-order parabolic equation*, Z. Angew. Math. Phys., 54 (2003), pp. 377–386.

[20] W. La Cruz, J. Martínez, and M. Raydan, *Spectral residual method without gradient information for solving large-scale nonlinear systems of equations*, Math. Comput., 75 (2006), pp. 1429–1448.

[21] C. Lattanzio and A. E. Tzavaras, *From gas dynamics with large friction to gradient flows describing diffusion theories*, Commun. Part. Diff. Eq., 42 (2017), pp. 261–290.

[22] W. Li, J. Lu, and L. Wang, *Fisher information regularization schemes for Wasserstein gradient flows*,arXiv:1907.02152, (2019).

[23] H. Liu and H. Yu, *An entropy satisfying conservative method for the Fokker-Planck equation of the finitely extensible nonlinear elastic dumbbell model*, SIAM J. Numer. Anal., 50(3) (2012), pp. 1207–1239.

[24] H. Liu and W. Maimaitiyiming, *Unconditional positivity-preserving and energy stable schemes for a reduced Poisson-Nernst-Planck system*, Commun. Comput. Phys. (in press); arXiv:1909.13161, (2019).

[25] H. Liu and W. Maimaitiyiming, *Positive and free energy satisfying schemes for diffusion with interaction potentials*, arXiv:1910.00151, (2019).

[26] J. Maas and D. Matthes, *Long-time behavior of a finite volume discretization for a fourth order diffusion equation*, Nonlinearity, 29 (2016), pp. 1992–2023.

[27] D. Matthes, R. J. McCann, and G. Savaré, *A family of nonlinear fourth order equations of gradient flow type*, Commun. Part. Diff. Eq., 34 (2009), pp. 1352–1397.

[28] D. Matthes and H. Osberger, *A convergent Lagrangian discretization for a nonlinear fourth-order equation*, Found. Comput. Math., 17 (2017), pp. 73–126.

[29] H. Osberger, *Long-time behavior of a fully discrete Lagrangian scheme for a family of fourth order equations*, Discrete Contin. Dyn. Syst. Ser. A, 37 (2017), pp. 405–434.

[30] S. M. Wise, C. Wang, and J. S. Lowengrub, *An energy-stable and convergent finite-difference scheme for the phase field crystal equation*, SIAM J. Numer. Anal., 47 (2009), pp. 2269–2288.

Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST),Thuwal 23955-6900, Saudi Arabia

E-mail address: xiaokai.huo@kaust.edu.sa

Iowa State University, Department of Mathematics, Ames, IA 50011

E-mail address: hliu@iastate.edu