Journal of Computational Physics 410 (2020) 109378

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

# Bound preserving and energy dissipative schemes for porous medium equation [☆]

Yiqi Gu, Jie Shen [*]

*Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, United States of America*

ARTICLE INFO

*Article history:*
Received 7 November 2019
Received in revised form 13 February 2020
Accepted 1 March 2020
Available online 5 March 2020

*Keywords:*
Porous medium equation
Wasserstein gradient flow
Bound preserving
Degenerate PDEs
Energy stability

ABSTRACT

A class of bound preserving and energy dissipative schemes for the porous medium equation are constructed in this paper. The schemes are based on a positivity preserving approach for Wasserstein gradient flow and a perturbation technique, and are shown to be uniquely solvable, bound preserving, and in the first-order case, also energy dissipative. Ample numerical results are presented to validate the theoretical results and demonstrate the effectiveness of the new schemes.

## 1. Introduction

The porous medium equation (PME), $f_t = \Delta(f^m)$ with $m > 1$, arises in the study of diffusion of gas through a porous medium under the action of the Darcy law [12]. It is widely used in many applications, such as flow through porous media, heat and mass transfer or combustion theory, population dynamics and tumor growth models, etc. (see [30,14,17] and the references therein). The PME has several distinct properties, including bound preserving, degeneracy and finite speed propagation. We refer to [25,26] for a summary of its analytic properties.

There are many challenges in solving the PME numerically. A good numerical scheme should preserve, as much as possible, essential properties satisfied by the PME, in particular: (i) bound preserving which implies in particular positivity preserving; and (ii) energy dissipation. It should also be able to accurately deal with degeneracy and finite speed propagation which means the solution remains to be zero in certain region for some time if it initially has a compact support. The degenerate region in which $f = 0$ will impede the usage of many traditional parabolic schemes. Besides, for fractional powers $m$, the numerical solution at any time should be non-negative to avoid producing complex values. Thus, it is critical that a numerical scheme for PME should preserve bound.

Many numerical methods have been proposed for the PME, such as Galerkin methods with finite element approximation [16,21,8,28,9,31,10,27,15], linearization schemes [24,18,20,13,11], perturbation approach [19,7], and particle schemes [29,6]. However, to the best of authors' knowledge, there is no numerical scheme which is provable bound preserving and energy dissipative.

It should be noted that the PME, along with Keller-Segel equation, Poisson-Nernst-Planck (PNP) equation and many other nonlinear parabolic equations, can be viewed as a Wasserstein gradient flow [1] whose nonlinear structure guarantees that

[☆] This work is partially supported by NSF DMS-1720442 and AFOSR FA9550-16-1-0102.

[*] Corresponding author.

*E-mail addresses:* gu129@purdue.edu (Y. Gu), shen7@purdue.edu (J. Shen).

https://doi.org/10.1016/j.jcp.2020.109378
0021-9991/© 2020 Elsevier Inc. All rights reserved.

(i) the solution will remain positive if it is initially so, and (ii) the solution is energy (entropy) dissipative. Therefore, a key ingredient in designing numerical schemes which are positivity preserving and energy dissipative is to preserve the Wasserstein structure in the discrete setting. This approach has been used recently to construct bound preserving and energy dissipative schemes for the Keller-Segel equation in [23] and for the PNP equation [22].

However, to deal with PME, there are two additional essential difficulties which are not present in the Keller-Segel and PNP equations: (i) bound preserving, and (ii) degeneracy. The fact that the solution may vanish in certain region prevents a direct application of the approach used in [23,22]. An effective approach to overcome the degeneracy is to introduce a perturbation [19,7]. More precisely, the initial value is elevated by a small perturbation $\varepsilon > 0$ such that the initial condition in the whole domain is positive, which will also guarantee the solution remain to be positive for all time. In this case, the approximation error will depend on the discretization errors as well as perturbation parameter. A fully implicit backward Euler scheme is considered in [19], and the $L^2$ error is estimated to be of order $(\tau + \varepsilon^2)^{1/2}$.

In this paper, we shall combine the approach developed in [23,22] for Wasserstein gradient flows and the perturbation approach in [19,7] to construct a class of nonlinear schemes, both semi-discrete in time and fully discrete with a finite-difference in space, which are uniquely solvable, bound preserving, and in the first-order case, also energy dissipative. Moreover, the schemes at each time step can be interpreted as Euler Lagrangian equations of convex functionals, so they can be efficiently solved by a Newton type iterative method with just a few iterations. We believe that our schemes are the first which is bound preserving as well as energy (entropy) dissipative.

The organization of this paper is as follows. In Section 2, we introduce the Dirichlet and Neumann problem of PME, and present the perturbation technique. In Section 3, we construct semi-discrete schemes and prove their solvability, uniform boundedness, mass decrease, energy dissipation and $H^1$ stability. In Section 4, we consider fully discrete schemes with finite-difference in space and establish their properties. In Section 5, we present ample numerical experiments in 1-D and 2-D to validate our schemes. We conclude with some remarks in the final section.

## 2. The porous medium equation

We describe below the Dirichlet problem of the PME that we shall consider in this paper and recall some of mathematical properties.

### 2.1. The Dirichlet problem

We consider the following PME

$$\frac{\partial f}{\partial t} = \Delta(f^m), \quad \text{in } Q_T := \Omega \times (0,T), \tag{2.1}$$

with initial condition

$$f(x,0) = f_0(x), \quad \text{in } \Omega, \tag{2.2}$$

and one of the boundary conditions

$$\begin{cases} \text{Dirichlet B.C.:} \quad f(x,t) = 0, \\ \text{Neumann B.C.:} \quad \frac{\partial}{\partial \nu} f(x,t) = 0, \end{cases} \qquad \text{in } \Sigma_T := \partial\Omega \times [0,T], \tag{2.3}$$

where $m > 1$ is a constant power, and $\nu$ is the outer normal to the boundary $\partial\Omega$. The following existence result is well known (see, e.g., [26]):

**Theorem 2.1.** *Assume $\Omega$ is a $C^{2,\alpha}$ domain, and $f_0$ is bounded in $C^\alpha(\overline{\Omega})$, then the PME (2.1)-(2.2) with Dirichlet B.C. admits a solution $f$ in $C^{2,1}(\overline{Q_T})$. Furthermore, if $f_0$ is $C^\infty$, then so is $f$. Same results apply to the Neumann problem.*

However, the uniqueness of solutions is not guaranteed, since $\Delta(f^m) = m\nabla \cdot (f^{m-1}\nabla f)$, implying possible degeneracy where $f_0$ vanishes. The issue can be solved by introducing the following non-degenerate assumption on the data

$$0 < m_1 \le f_0(x) \le m_0, \quad \forall x \in \Omega. \tag{2.4}$$

Under the above assumption, the Dirichlet or Neumann problem is uniquely solvable and the solution $f$ is uniformly bounded above and below as the data, i.e.

$$m_1 < f(x,t) < m_0, \quad \forall (x,t) \in Q_T. \tag{2.5}$$

However, in many applications of interest, we can only assume that the initial data is bounded and non-negative, i.e.

$$0 \le f_0(x) \le m_0, \quad \forall x \in \Omega. \tag{2.6}$$

In this paper, we always assume (2.6) holds, and, for simplicity, $\Omega$ is sufficiently smooth (e.g. a $C^{2,\alpha}$ domain).

### 2.2. Some properties

The PME (2.1)-(2.3) is energy dissipative under the hypothesis (2.6), and is mass decreasing for Dirichlet boundary condition and mass conservative for Neumann boundary condition.

- Taking the inner product of (2.1) with $f^m$ and integrating by parts, we obtain energy dissipation

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega f^{m+1}\mathrm{d}x = -(m+1)\int_\Omega |\nabla(f^m)|^2\mathrm{d}x \le 0. \tag{2.7}$$

- For Dirichlet B.C., integrating (2.1) over $\Omega$, noticing that

$$\oint_{\partial\Omega}\frac{\partial}{\partial \mathrm{n}}(f^m)\mathrm{d}s = \oint_{\partial\Omega}\lim_{|\mathrm{n}|\to 0}\frac{f^m(x_s)-f^m(x_s-\mathrm{n})}{|\mathrm{n}|}\mathrm{d}s = -\oint_{\partial\Omega}\lim_{|\mathrm{n}|\to 0}\frac{f^m(x_s-\mathrm{n})}{|\mathrm{n}|}\mathrm{d}s, \tag{2.8}$$

which is non-positive since $f>0$ in $\Omega$, we find that the mass is decreasing as

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega f\,\mathrm{d}x = \oint_{\partial\Omega}\frac{\partial}{\partial \mathrm{n}}(f^m)\mathrm{d}s \le 0, \tag{2.9}$$

where n is the outward unit normal of $\partial\Omega$.
For Neumann B.C., $\oint_{\partial\Omega}\frac{\partial}{\partial \mathrm{n}}(f^m)\mathrm{d}s = 0$ directly leads to the mass conservation,

$$\int_\Omega f(x,t)\mathrm{d}x = \int_\Omega f(x,0)\mathrm{d}x, \quad \forall t>0. \tag{2.10}$$

## 3. Semi-discrete (in time) schemes

We construct in this section bound preserving semi-discrete (in time) schemes for the PME with Dirichlet boundary condition. Note that all schemes that we construct can be used for PME with Neumann boundary conditions with simple modifications, so we only discuss the Dirichlet case in the following sections.

### 3.1. A perturbed problem

If $f$ is positive and differentiable in $\Omega$, then by noting $\nabla f = f\nabla \ln f$, (2.1) can be rewritten as

$$\frac{\partial f}{\partial t} = m\nabla\cdot(f^m\nabla\ln f), \quad \text{in } Q_T. \tag{3.1}$$

However, the existence of term $\ln f$ requires $f>0$ in $Q_T$, which is a little more strict than the assumption (2.6). Hence we introduce a small perturbation to the initial data to enforce $f$ to be strictly positive. That is, we solve the following perturbed problem

$$\frac{\partial f_\varepsilon}{\partial t} = m\nabla\cdot(f_\varepsilon^m\nabla\ln f_\varepsilon), \quad \text{in } Q_T, \tag{3.2}$$

with initial condition

$$f_\varepsilon(x,0) = f_0(x)+\varepsilon, \quad \text{in } \Omega, \tag{3.3}$$

and boundary condition

$$f_\varepsilon(x,t) = \varepsilon, \quad \text{in } \Sigma_T, \tag{3.4}$$

where $\varepsilon>0$ is a small positive number to make sure the transformed log-type PME (3.2) to be well-defined and non-degenerate. Without loss of generality, we assume $\varepsilon \ll m_0$.

### 3.2. A first-order scheme

We first study the scheme in the strong formulation, followed by the study of its weak formulation which allows us to establish uniform bounds for the numerical solution.

#### 3.2.1. Strong formulation

Let $N$ be the total number of time steps, and $\tau := T/N$, we consider the following nonlinear scheme for (3.2):

$$\frac{f_\varepsilon^n - f_\varepsilon^{n-1}}{\tau} = m\nabla \cdot \left((f_\varepsilon^{n-1})^m \nabla \ln f_\varepsilon^n\right), \quad \text{in } Q_T, \tag{3.5}$$

with initial condition

$$f_\varepsilon^0(x) = f_0(x) + \varepsilon, \quad \text{in } \Omega, \tag{3.6}$$

and boundary condition

$$f_\varepsilon^n(x) = \varepsilon, \quad \text{on } \partial\Omega. \tag{3.7}$$

**Theorem 3.1.** *Let $f_\varepsilon^{n-1} > 0$ in the scheme (3.5), then $f_\varepsilon^n$ is uniquely solvable and positivity preserving with boundary condition (3.7), i.e., $f_\varepsilon^n > 0$.*

**Proof.** Let $\mathcal{L}_n$ be a linear operator that $\mathcal{L}_n g$ is the (unique) solution of

$$\begin{cases} -m\tau\nabla \cdot \left((f_\varepsilon^{n-1})^m \nabla u\right) = g, & \text{in } \Omega, \\ u = -\ln \varepsilon, & \text{on } \partial\Omega. \end{cases} \tag{3.8}$$

Note $\mathcal{L}_n$ is self-adjoint and positive definite. Consider the functional

$$F[f] := \int_\Omega f(\ln f - 1)\mathrm{d}x + \frac{1}{2}\int_\Omega (f - f_\varepsilon^{n-1})\mathcal{L}_n(f - f_\varepsilon^{n-1})\mathrm{d}x. \tag{3.9}$$

It is easy to check that $F$ is strictly convex in the admissible set

$$\mathcal{A} := \{f \in H^2(\Omega) : f > 0 \text{ in } \Omega,\ f|_{\partial\Omega} = \varepsilon\}. \tag{3.10}$$

Hence, there exists a unique $f_\varepsilon^n \in \mathcal{A}$ such that $\frac{\delta F}{\delta f_\varepsilon^n} = 0$ [4]. Since $\frac{\delta F}{\delta f} = \ln f + \mathcal{L}_n(f - f_\varepsilon^{n-1})$, we find that $\frac{\delta F}{\delta f_\varepsilon^n} = 0$ is exactly the scheme (3.6). □

**Remark 3.1.** If we integrate (3.5) over $\Omega$ with $f_\varepsilon^n - \varepsilon$, the following $L^2$ energy dissipation can be obtained by similar argument,

$$\int_\Omega |f_\varepsilon^n|^2 \mathrm{d}x - \int_\Omega |f_\varepsilon^{n-1}|^2 \mathrm{d}x \le -m\tau \int_\Omega \frac{(f_\varepsilon^{n-1})^m}{f_\varepsilon^n} |\nabla f_\varepsilon^n|^2 \mathrm{d}x.$$

#### 3.2.2. Weak formulation and uniform boundedness

It is shown above that the scheme (3.5) is positivity preserving, but does not show it is bound preserving. In order to prove a uniform bound and associated properties for the numerical solution, we shall consider a weak formulation for (3.5). Let

$$\hat{H}_\varepsilon^1(\Omega) := \{f \in H^1(\Omega), f \ge 0 \text{ in } \Omega,\ f|_{\partial\Omega} = \varepsilon\}. \tag{3.11}$$

Then a weak formulation of (3.5) is: Given $f_\varepsilon^{n-1} \in \hat{H}_\varepsilon^1(\Omega)$, find $f_\varepsilon^n \in \hat{H}_\varepsilon^1(\Omega)$ such that

$$\left(f_\varepsilon^n - f_\varepsilon^{n-1}, \phi\right) + m\tau\left((f_\varepsilon^{n-1})^m \nabla \ln f_\varepsilon^n, \nabla\phi\right) = 0, \quad \forall \phi \in H_0^1(\Omega). \tag{3.12}$$

**Theorem 3.2.** *Let $\varepsilon \le f_\varepsilon^{n-1} \le m_0 + \varepsilon$. Then, the problem (3.12) admits a unique solution such that*

$$\varepsilon \le f_\varepsilon^n \le m_0 + \varepsilon. \tag{3.13}$$

**Proof.** The existence of weak solutions directly follows Theorem 3.1, which implies the weak problem admits at least one solution. The uniqueness can be derived as follows. Let $f_1^n$ and $f_2^n$ be two solutions of (3.12), then it follows

$$\left(f_1^n - f_2^n, \phi\right) + m\tau\left((f_\varepsilon^{n-1})^m \nabla(\ln f_1^n - \ln f_2^n), \nabla\phi\right) = 0, \quad \forall \phi \in H_0^1(\Omega). \tag{3.14}$$

Taking $\phi = \ln f_1^n - \ln f_2^n \in H_0^1(\Omega)$ gives

$$\left(f_1^n - f_2^n, \ln f_1^n - \ln f_2^n\right) + m\tau \left((f_\varepsilon^{n-1})^m \nabla(\ln f_1^n - \ln f_2^n), \nabla(\ln f_1^n - \ln f_2^n)\right) = 0. \tag{3.15}$$

The first term on the left is non-negative because of the monotonicity of $\ln x$. This is also true for the second term since $(f_\varepsilon^{n-1})^m > 0$. Therefore both terms have to equal to zero, which implies $f_1^n = f_2^n$.

It remains to show the uniform bounds. Taking $\phi = [\varepsilon - f_\varepsilon^n]_+$ in (3.12), we find

$$\left(f_\varepsilon^n - \varepsilon, [\varepsilon - f_\varepsilon^n]_+\right) + m\tau \left((f_\varepsilon^{n-1})^m \nabla \ln f_\varepsilon^n, \nabla[\varepsilon - f_\varepsilon^n]_+\right) = \left(f_\varepsilon^{n-1} - \varepsilon, [\varepsilon - f_\varepsilon^n]_+\right). \tag{3.16}$$

Note the first term on the left is non-positive, and the second term equals to

$$-m\tau \int\limits_{\{f_\varepsilon^n < \varepsilon\}} \frac{(f_\varepsilon^{n-1})^m}{f_\varepsilon^n} |\nabla f_\varepsilon^n|^2 \mathrm{d}x \le 0,$$

so the left hand side of (3.16) is non-positive. But the right hand side of (3.16) is non-negative, hence all the three terms in (3.16) are equal to zero, which implies $\mathrm{meas}\{f_\varepsilon^n < \varepsilon\} = 0$, i.e. $f_\varepsilon^n \ge \varepsilon$.

Similarly, by taking $\phi = [f_\varepsilon^n - m_0 - \varepsilon]_+$, we can obtain $f_\varepsilon^n \le m_0 + \varepsilon$. $\square$

The uniform boundedness of the weak solution is consistent with that of classical PME equation (2.5), and plays an important role for our analysis below. In particular, we derive from (3.13) that

$$\int\limits_\Omega |\nabla \ln f_\varepsilon^n|^2 \mathrm{d}x = \int\limits_\Omega \frac{|\nabla f_\varepsilon^n|^2}{(f_\varepsilon^n)^2} \mathrm{d}x \le \frac{1}{\varepsilon^2} \int\limits_\Omega |\nabla f_\varepsilon^n|^2 \mathrm{d}x.$$

Moreover, since the classical solution of (3.5)-(3.7) solves the weak problem (3.12), the boundedness result (3.13) is also true for (3.5)-(3.7), by which we can show the classical solution is mass decreasing and energy dissipative. More precisely, we have

**Theorem 3.3.** *The solution of scheme (3.5) with conditions (3.6) and (3.7) is mass decreasing, i.e.,*

$$\int\limits_\Omega f_\varepsilon^n \mathrm{d}x \le \int\limits_\Omega f_\varepsilon^{n-1} \mathrm{d}x, \tag{3.17}$$

*and energy dissipative (as $\varepsilon \to 0$) in the sense that*

$$E[f_\varepsilon^n] - E[f_\varepsilon^{n-1}] \le -m\tau \left( \int\limits_\Omega (f_\varepsilon^{n-1})^m |\nabla \ln f_\varepsilon^n|^2 - \varepsilon^{m-1} \ln \varepsilon \oint\limits_{\partial\Omega} \frac{\partial}{\partial \mathrm{n}} f_\varepsilon^n \mathrm{d}s \right), \tag{3.18}$$

*where $E[f] := \int_\Omega f(\ln f - 1)\mathrm{d}x$.*

**Proof.** By integrating (3.5) over $\Omega$ and performing integration by parts, it follows that

$$\begin{aligned} \frac{1}{\tau}\left( \int\limits_\Omega f_\varepsilon^n \mathrm{d}x - \int\limits_\Omega f_\varepsilon^{n-1} \mathrm{d}x \right) &= m \int\limits_\Omega \nabla \cdot \left((f_\varepsilon^{n-1})^m \nabla \ln f_\varepsilon^n\right) \mathrm{d}x \\ &= m \oint\limits_{\partial\Omega} (f_\varepsilon^{n-1})^m \frac{\partial}{\partial \mathrm{n}} \ln f_\varepsilon^n \mathrm{d}s = m \oint\limits_{\partial\Omega} \frac{(f_\varepsilon^{n-1})^m}{f_\varepsilon^n} \frac{\partial}{\partial \mathrm{n}} f_\varepsilon^n \mathrm{d}s \le 0, \end{aligned} \tag{3.19}$$

where the last inequality is derived by using a similar argument as in the proof of (2.9) and the boundedness (3.13).

It remains to prove the energy dissipation law. By multiplying $\ln f_\varepsilon^n$ on both sides of (3.5), integrating over $\Omega$ and performing a similar calculation as in (3.19), we obtain

$$\begin{aligned} \frac{1}{\tau} \int\limits_\Omega (f_\varepsilon^n - f_\varepsilon^{n-1}) \ln f_\varepsilon^n \mathrm{d}x &= m \int\limits_\Omega \nabla \cdot \left((f_\varepsilon^{n-1})^m \nabla \ln f_\varepsilon^n\right) \ln f_\varepsilon^n \mathrm{d}x \\ &= -m \left( \int\limits_\Omega (f_\varepsilon^{n-1})^m |\nabla \ln f_\varepsilon^n|^2 - \oint\limits_{\partial\Omega} \frac{(f_\varepsilon^{n-1})^m \ln f_\varepsilon^n}{f_\varepsilon^n} \frac{\partial}{\partial \mathrm{n}} f_\varepsilon^n \mathrm{d}s \right) \\ &= -m \left( \int\limits_\Omega (f_\varepsilon^{n-1})^m |\nabla \ln f_\varepsilon^n|^2 - \varepsilon^{m-1} \ln \varepsilon \oint\limits_{\partial\Omega} \frac{\partial}{\partial \mathrm{n}} f_\varepsilon^n \mathrm{d}s \right). \end{aligned}$$

Given $a, b > 0$, by Taylor expansion, there exists $\xi$ between $a$ and $b$ such that

$$(a-b)\ln a = (a\ln a - a) - (b\ln b - b) + \frac{(a-b)^2}{2\xi}. \tag{3.20}$$

Hence,

$$\int_\Omega f_\varepsilon^n(\ln f_\varepsilon^n - 1)\mathrm{d}x - \int_\Omega f_\varepsilon^{n-1}(\ln f_\varepsilon^{n-1} - 1)\mathrm{d}x \le \int_\Omega (f_\varepsilon^n - f_\varepsilon^{n-1})\ln f_\varepsilon^n \mathrm{d}x \le$$
$$-m\tau\left(\int_\Omega (f_\varepsilon^{n-1})^m|\nabla \ln f_\varepsilon^n|^2 - \varepsilon^{m-1}\ln\varepsilon \oint_{\partial\Omega}\frac{\partial}{\partial \mathrm{n}} f_\varepsilon^n \mathrm{d}s\right),$$

which is the desired energy dissipation law. □

Next, we derive the $l^2(H^1)$ stability for the semi-discrete weak problem (3.12).

**Theorem 3.4.** *Under the hypothesis of* (2.6)*,* $\theta_\varepsilon^n := \ln f_\varepsilon^n$ *satisfies*

$$\tau\sum_{n=1}^p \|e^{m\theta_\varepsilon^{n-1}/2}\nabla\theta_\varepsilon^n\|^2 \le (1+\ln\varepsilon)\cdot C(m, m_0, \Omega), \tag{3.21}$$

*for any* $p = 1\cdots, N$*. Especially, it satisfies*

$$\tau\sum_{n=1}^N \|\nabla\theta_\varepsilon^n\|^2 \le (1+\ln\varepsilon)\cdot C(m, m_0, \Omega). \tag{3.22}$$

**Proof.** Rewrite the problem (3.12) as

$$\begin{cases} \text{find } \theta_\varepsilon^n \in H^1_{\ln\varepsilon}(\Omega) := \{\theta \in H^1(\Omega) : \mathrm{tr}(\theta) = \ln\varepsilon \text{ on } \partial\Omega\} \text{ such that} \\ \left(e^{\theta_\varepsilon^n} - e^{\theta_\varepsilon^{n-1}}, \phi\right) + m\tau\left(e^{m\theta_\varepsilon^{n-1}}\nabla\theta_\varepsilon^n, \nabla\phi\right) = 0, \quad \forall\phi \in H_0^1(\Omega), \end{cases} \tag{3.23}$$

with $\theta_\varepsilon^0(x) = \ln(f_0(x) + \varepsilon)$.

Taking $\phi = \theta_\varepsilon^n - \ln\varepsilon \in H_0^1(\Omega)$ in (3.23) leads to

$$\left(e^{\theta_\varepsilon^n} - e^{\theta_\varepsilon^{n-1}}, \theta_\varepsilon^n - \ln\varepsilon\right) + m\tau\left(e^{m\theta_\varepsilon^{n-1}}\nabla\theta_\varepsilon^n, \nabla\theta_\varepsilon^n\right) = 0. \tag{3.24}$$

Sum up (3.24) for $n = 1, \cdots, p$, we obtain

$$\sum_{n=1}^p\left(e^{\theta_\varepsilon^n} - e^{\theta_\varepsilon^{n-1}}, \theta_\varepsilon^n\right) + m\tau\sum_{n=1}^p\left(e^{m\theta_\varepsilon^{n-1}}\nabla\theta_\varepsilon^n, \nabla\theta_\varepsilon^n\right) = \ln\varepsilon\left(e^{\theta_\varepsilon^n} - e^{\theta_\varepsilon^{n-1}}, 1\right). \tag{3.25}$$

Since $\ln\varepsilon \le \theta_\varepsilon^n \le \ln(m_0 + \varepsilon)$ by (3.13), the first term in (3.25) can be bounded from below as follows.

$$\sum_{n=1}^p\int_\Omega (e^{\theta_\varepsilon^n} - e^{\theta_\varepsilon^{n-1}})\theta_\varepsilon^n \mathrm{d}x \ge \sum_{n=1}^p\int_\Omega\int_{\theta_\varepsilon^{n-1}(x)}^{\theta_\varepsilon^n(x)} se^s\mathrm{d}s\mathrm{d}x$$
$$= \int_\Omega\int_{\theta_\varepsilon^0(x)}^{\theta_\varepsilon^p(x)} se^s\mathrm{d}s\mathrm{d}x = \int_\Omega\left((\theta_\varepsilon^p - 1)e^{\theta_\varepsilon^p} - (\theta_\varepsilon^0 - 1)e^{\theta_\varepsilon^0}\right)\mathrm{d}x \ge -C(m_0, \Omega),$$

where $C(m_0, \Omega)$ is a constant only depending on $m_0$ and $\Omega$.

On the other hand, the right hand side in (3.25) satisfies

$$\ln\varepsilon\int_\Omega (e^{\theta_\varepsilon^p} - e^{\theta_\varepsilon^0})\mathrm{d}x \le \ln\varepsilon\cdot C(m_0, \Omega).$$

Combining the above relations into (3.25), we obtain the desired results. □

**Remark 3.2.** Introducing the transformation $\theta_\varepsilon^n := \ln f_\varepsilon^n$ is only for convenience in the statement and proof of Theorem 3.4. In numerical simulations, we solve the equation for $f_\varepsilon^n$ directly and do not compute $\theta_\varepsilon^n$.

### 3.3. A second-order scheme

Special care has to be taken when constructing second-order schemes since linear extrapolation of two positive functions is not guaranteed to be positive. Therefore, a nonlinear extrapolation has to be used. For example, we can construct the following second-order scheme based on Crank-Nicolson:

$$\frac{f_\varepsilon^n - f_\varepsilon^{n-1}}{\tau} = m\nabla\cdot\left((\widetilde{f_\varepsilon}^{n-\frac{1}{2}})^m \nabla \ln\frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2}\right), \quad \text{in } Q_T, \tag{3.26}$$

with initial condition (3.6) and boundary condition (3.7). Here $\widetilde{f_\varepsilon}^{n-\frac{1}{2}}$ is defined by

$$\widetilde{f_\varepsilon}^{n-\frac{1}{2}} := \begin{cases} \frac{1}{2}(3f_\varepsilon^{n-1} - f_\varepsilon^{n-2}), & \text{if } 3f_\varepsilon^{n-1} > f_\varepsilon^{n-2},\ \forall x\in\Omega, \\ (f_\varepsilon^{n-1})^{\frac{3}{2}}/(f_\varepsilon^{n-2})^{\frac{1}{2}}, & \text{if } 3f_\varepsilon^{n-1} \le f_\varepsilon^{n-2}, \text{ otherwise}, \end{cases} \tag{3.27}$$

which is a second-order approximation to $f_\varepsilon(x, t_{n-1/2})$ and always positive.

By using similar arguments as in the proof of Theorem 3.1 with (3.9) replaced by

$$F[f] := \int_\Omega (f + f_\varepsilon^n)(\ln\frac{f+f_\varepsilon^n}{2} - 1)\mathrm{d}x + \frac{1}{2}\int_\Omega (f - f_\varepsilon^n)\,\mathcal{L}_n(f - f_\varepsilon^n)\mathrm{d}x, \tag{3.28}$$

we can prove the following result:

**Theorem 3.5.** *Let $f_\varepsilon^{n-1}$, $f_\varepsilon^{n-2} > 0$. The scheme (3.26) with conditions (3.6) and (3.7) is uniquely solvable and bound preserving, i.e., $f_\varepsilon^n > 0$. Furthermore, the solution is mass decreasing, i.e.,*

$$\int_\Omega f_\varepsilon^n \mathrm{d}x \le \int_\Omega f_\varepsilon^{n-1}\mathrm{d}x.$$

Note that we are unable to prove that the second-order scheme is energy dissipative as the first-order scheme.

In order to establish a uniform bound for the numerical solution, we consider the following weak formulation of (3.26):

$$\begin{cases} \text{find } f_\varepsilon^n \in \hat{H}_\varepsilon^1(\Omega) \text{ such that} \\ (f_\varepsilon^n - f_\varepsilon^{n-1}, \phi) + m\tau\left((\widetilde{f_\varepsilon}^{n-\frac{1}{2}})^m\nabla\ln\frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2}, \nabla\phi\right) = 0, \quad \forall\phi\in H_0^1(\Omega). \end{cases} \tag{3.29}$$

It can be shown that the problem (3.29) admits a unique solution by using similar arguments for the first-order case. However, we can only derive a uniform bound for $(f_\varepsilon^n + f_\varepsilon^{n-1})/2$.

**Theorem 3.6.** *Assuming $\varepsilon \le f_\varepsilon^{n-1} \le m_0 + \varepsilon$, the solution of (3.29) satisfies $\varepsilon \le \frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2} \le m_0 + \varepsilon$.*

**Proof.** By taking $\phi = [\varepsilon - \frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2}]_+$ in (3.29), it follows

$$\left(f_\varepsilon^n - \varepsilon, [\varepsilon - \frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2}]_+\right) + m\tau\left((\widetilde{f_\varepsilon}^{n-\frac{1}{2}})^m\nabla\ln\frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2}, \nabla[\varepsilon - \frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2}]_+\right) \\ = \left(f_\varepsilon^{n-1} - \varepsilon, [\varepsilon - \frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2}]_+\right). \tag{3.30}$$

Note the second term on the left of (3.30) is equal to

$$-\frac{m\tau}{2}\int_{\{x:\frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2} < \varepsilon\}} \frac{(\widetilde{f_\varepsilon}^{n-\frac{1}{2}})^m}{f_\varepsilon^n + f_\varepsilon^{n-1}}|\nabla(f_\varepsilon^n + f_\varepsilon^{n-1})|^2\mathrm{d}x \le 0.$$

Also, note $\left\{x : \frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2} < \varepsilon\right\} \subset \left\{f_\varepsilon^n < \varepsilon\right\}$ since $f_\varepsilon^{n-1} \ge \varepsilon$, so the first term on the left of (3.30) is non-positive. But the right hand side of (3.30) is non-negative, hence all the three terms in (3.30) are equal to zero, which implies meas$\{\frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2} < \varepsilon\} = 0$, i.e. $\frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2} \ge \varepsilon$.

Similarly, by taking $\phi = [\frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2} - m_0 - \varepsilon]_+$, we can obtain $\frac{f_\varepsilon^n + f_\varepsilon^{n-1}}{2} \le m_0 + \varepsilon$. □

Since we can not apply the above results recursively to derive a uniform bound for $f_\varepsilon^n$, we construct below a modified scheme that will allow us to do so.

- Compute $\widehat{f_\varepsilon}^n$ from

$$\begin{aligned}&\frac{\widehat{f_\varepsilon}^n - f_\varepsilon^{n-1}}{\tau} = m\nabla\cdot\left((\widetilde{f_\varepsilon}^{n-\frac{1}{2}})^m \nabla \ln \frac{\widehat{f_\varepsilon}^n + f_\varepsilon^{n-1}}{2}\right), \quad \text{in } Q_T,\\ &\text{with } \widetilde{f_\varepsilon}^{n-\frac{1}{2}} := \begin{cases}\frac{1}{2}(3f_\varepsilon^{n-1} - f_\varepsilon^{n-2}), & \text{if } 3f_\varepsilon^{n-1} > f_\varepsilon^{n-2},\ \forall x\in\Omega,\\ (f_\varepsilon^{n-1})^{\frac{3}{2}}/(f_\varepsilon^{n-2})^{\frac{1}{2}}, & \text{otherwise.}\end{cases}\end{aligned} \tag{3.31}$$

  If $\widehat{f_\varepsilon}^n \geq \varepsilon\ \forall x\in\Omega$, set $f_\varepsilon^n = \widehat{f_\varepsilon}^n$ and go to the next time step.
- Otherwise, set

$$\widehat{f_\varepsilon}^{n-\frac{1}{2}} = \frac{\widehat{f_\varepsilon}^n + f_\varepsilon^{n-1}}{2},$$

  and compute $\widehat{f_\varepsilon}^{n+\frac{1}{2}}$ from

$$\begin{aligned}&\frac{\widehat{f_\varepsilon}^{n+\frac{1}{2}} - \widehat{f_\varepsilon}^{n-\frac{1}{2}}}{\tau} = m\nabla\cdot\left((\widetilde{f_\varepsilon}^n)^m \nabla \ln \frac{\widehat{f_\varepsilon}^{n+\frac{1}{2}} + \widehat{f_\varepsilon}^{n-\frac{1}{2}}}{2}\right), \quad \text{in } Q_T,\\ &\qquad \text{with } \widetilde{f_\varepsilon}^n := (\widehat{f_\varepsilon}^{n-\frac{1}{2}})^2/f_\varepsilon^{n-1},\end{aligned} \tag{3.32}$$

  set $f_\varepsilon^n = \frac{\widehat{f_\varepsilon}^{n+\frac{1}{2}} + \widehat{f_\varepsilon}^{n-\frac{1}{2}}}{2}$ with homogeneous Dirichlet boundary condition, and go to the next time step.

Following a similar procedure as in the proof of Theorem 3.6, we can establish the following result:

**Corollary 3.7.** *Assuming $\varepsilon \leq f_\varepsilon^{n-1} \leq m_0 + \varepsilon$, then the solution of (3.31)-(3.32) satisfies*

$$\varepsilon \leq f_\varepsilon^n \leq m_0 + \varepsilon.$$

## 4. Fully discrete schemes

We consider in this section fully discrete schemes for the PME using a second-order finite difference method for spatial discretization. To simplify the presentation, we shall only consider the 1-D case below, but the scheme and the associated results can be easily extended to multi-dimensional rectangular domains.

### 4.1. First-order in time schemes

Let $\Omega = (-L, L)$ with $L > 0$, given $I \in \mathbb{N}^+$, we denote $h := 2L/I$ be the grid width and $x_i := -L + ih$, $i = 0, \cdots, I$, be the collocation points. Let $f_i^n$ be the approximation to $f_\varepsilon(x_i, n\tau)$, then a fully discrete version of (3.5) is as follows

$$\frac{f_i^n - f_i^{n-1}}{\tau} = \frac{m}{h^2}\left((f_{i+\frac{1}{2}}^{n-1})^m(\ln f_{i+1}^n - \ln f_i^n) - (f_{i-\frac{1}{2}}^{n-1})^m(\ln f_i^n - \ln f_{i-1}^n)\right),\ i = 1, \cdots, I-1, \tag{4.1}$$

with initial condition

$$f_i^0 = f_0(x_i) + \varepsilon, \quad i = 1, \cdots, I-1, \tag{4.2}$$

and boundary condition

$$f_0^n = f_I^n = \varepsilon, \tag{4.3}$$

for $n = 1, \cdots, N$. It is clear that the above scheme is formally second-order in space.

The solvability of the fully discretized scheme can be described as follows.

**Theorem 4.1.** *Let $f_i^{n-1} > 0$ for $i = 1, \cdots, I-1$ in the scheme (4.1), then $\boldsymbol{f}^n := [f_1^n\ f_2^n\ \cdots\ f_{I-1}^n]^T$ is uniquely solvable and positivity preserving with condition (4.3), i.e., $\boldsymbol{f}^n > 0$.*

**Proof.** The scheme (4.1) leads to a system of nonlinear equations

$$F(\boldsymbol{f}^n)=0, \tag{4.4}$$

with

$$F(\boldsymbol{f}^n):=\frac{1}{\tau}(\boldsymbol{f}^n-\boldsymbol{f}^{n-1})+\boldsymbol{A}\ln\boldsymbol{f}^n+\boldsymbol{c}, \tag{4.5}$$

where

$$(\boldsymbol{A})_{ij}=\begin{cases}-\frac{m}{h^2}(f^{n-1}_{i-\frac12})^m, & j=i-1,\\ \frac{m}{h^2}\left((f^{n-1}_{i+\frac12})^m+(f^{n-1}_{i-\frac12})^m\right), & j=i,\\ -\frac{m}{h^2}(f^{n-1}_{i+\frac12})^m, & j=i+1,\\ 0, & |j-i|>1,\end{cases} \tag{4.6}$$

and $\boldsymbol{c}=\left[-\frac{m}{h^2}(f^{n-1}_{\frac12})^m\ln\varepsilon,0,\cdots,0,-\frac{m}{h^2}(f^{n-1}_{I-\frac12})^m\ln\varepsilon\right]^T$.

It can be easily seen that $\boldsymbol{A}$ is symmetric positive definite, hence so is $\boldsymbol{A}^{-1}$. Then the unique solvability can be directly obtained by the fact that (4.4) is equivalent to $\boldsymbol{A}^{-1}F(\boldsymbol{f}^n)=0$, which is the Euler-Lagrange equation of the strictly convex function

$$G(\boldsymbol{f}^n)=\frac{1}{2\tau}(\boldsymbol{f}^n-\boldsymbol{f}^{n-1})^T\boldsymbol{A}^{-1}(\boldsymbol{f}^n-\boldsymbol{f}^{n-1})+(\boldsymbol{f}^n)^T(\ln\boldsymbol{f}^n-1)+\boldsymbol{c}^T\boldsymbol{A}^{-1}\boldsymbol{f}^n. \tag{4.7}$$

The proof is complete. □

The above nonlinear system can be solved efficiently by using Newton's iteration (see the numerical examples in the next section) with Jacobian of $F$ given by

$$(\nabla F)_{ij}=\begin{cases}-\frac{m\tau}{h^2}(f^{n-1}_{i-\frac12})^m/f^n_{i-1}, & j=i-1,\\ 1+\frac{m\tau}{h^2}\left((f^{n-1}_{i+\frac12})^m+(f^{n-1}_{i-\frac12})^m\right)/f^n_i, & j=i,\\ -\frac{m\tau}{h^2}(f^{n-1}_{i+\frac12})^m/f^n_{i+1}, & j=i+1,\\ 0, & |j-i|>1.\end{cases} \tag{4.8}$$

*4.1.1. Uniform boundedness*

Similar to the semi-discrete cases, we have

**Theorem 4.2.** *Given $\varepsilon\le f^{n-1}_i\le m_0+\varepsilon$, $i=1,\cdots,I-1$ with $f^{n-1}_0=f^{n-1}_N=\varepsilon$, then the solution of (4.1) satisfies*

$$\varepsilon\le f^n_i\le m_0+\varepsilon\quad\text{for } i=1,\cdots,I-1 \text{ and } n=1,\cdots,N. \tag{4.9}$$

**Proof.** Multiplying $\phi_i\in\mathbb{R}$ to both sides of (4.1) with $\phi_0=\phi_N=0$, summing up for $i=1,\cdots,I-1$ and using summation by parts, we obtain

$$\sum_{i=1}^{I-1}(f^n_i-f^{n-1}_i)\phi_i+\frac{m\tau}{h^2}\sum_{i=0}^{I-1}(f^{n-1}_{i+\frac12})^m(\ln f^n_{i+1}-\ln f^n_i)(\phi_{i+1}-\phi_i)=0. \tag{4.10}$$

Taking $\phi_i=\max\{\varepsilon-f^n_i,0\}$ in (4.10) gives

$$\sum_{i=1}^{I-1}(f^n_i-\varepsilon)\max\{\varepsilon-f^n_i,0\}-\sum_{i=1}^{I-1}(f^{n-1}_i-\varepsilon)\max\{\varepsilon-f^n_i,0\}$$
$$+\frac{m\tau}{h^2}\sum_{i=0}^{I-1}(f^{n-1}_{i+\frac12})^m(\ln f^n_{i+1}-\ln f^n_i)\left(\max\{\varepsilon-f^n_{i+1},0\}-\max\{\varepsilon-f^n_i,0\}\right)=0.$$

Note all three terms (including the signs) on the left side are always non-positive, and the right side is zero, so the only possibility is that all terms are zero, in particular,

$$\sum_{i=1}^{I-1}(f_i^n-\varepsilon)\max\{\varepsilon-f_i^n,0\}=0,$$

which implies $\max\{\varepsilon-f_i^n,0\}=0$, i.e. $f_i^n\ge\varepsilon$. Similarly, we can show that $f_i^n\le m_0+\varepsilon$. □

After obtaining the boundedness, we can show

**Theorem 4.3.** *Given $\varepsilon\le f_i^{n-1}\le m_0+\varepsilon$, $i=1,\cdots,I-1$ with $f_0^{n-1}=f_N^{n-1}=\varepsilon$, then the solution $\{f_i^n\}$ of the fully discretized scheme (4.10) is mass decreasing, i.e.*

$$\frac{1}{\tau}\sum_{i=1}^{I-1}f_i^n\le\frac{1}{\tau}\sum_{i=1}^{I-1}f_i^{n-1}, \tag{4.11}$$

*and energy dissipative in two separated forms, i.e.*

$$E_h^n-E_h^{n-1}\le-\frac{m\tau\varepsilon^m}{h}\sum_{i=0}^{I-1}|\ln f_{i+1}^n-\ln f_i^n|^2, \tag{4.12}$$

*where $E_h^n:=h\sum_{i=1}^{I-1}f_i^n(\ln f_i^n-1)$, and*

$$\sum_{i=1}^{I-1}|f_i^n|^2-\sum_{i=1}^{I-1}|f_i^{n-1}|^2\le-\frac{2m\tau}{(m_0+\varepsilon)h^2}\sum_{i=1}^{I-2}(f_{i+\frac12}^{n-1})^m|f_{i+1}^n-f_i^n|^2. \tag{4.13}$$

**Proof.** By summing up (4.1) for $i=1,\cdots,I-1$, we obtain

$$\frac{1}{\tau}\left(\sum_{i=1}^{I-1}f_i^n-\sum_{i=1}^{I-1}f_i^{n-1}\right)=\frac{m}{h^2}\left((f_{I-\frac12}^{n-1})^m(\ln\varepsilon-\ln f_{I-1}^n)-(f_{\frac12}^{n-1})^m(\ln f_1^n-\ln\varepsilon)\right). \tag{4.14}$$

By (4.9), the right hand side of above is non-positive, which implies (4.11).

To prove (4.12), it suffices to take $\phi_i=\ln f_i^n$ in (4.10) then apply (4.9) and (3.20).

To prove (4.13), we take $\phi_i=f_i^n$ in (4.10), then it follows

$$\begin{aligned}\sum_{i=1}^{I-1}(f_i^{n+1}-f_i^n)f_i^{n+1}+\frac{m\tau}{h^2}\sum_{i=1}^{I-2}(f_{i+\frac12}^{n-1})^m(\ln f_{i+1}^n-\ln f_i^n)(f_{i+1}^n-f_i^n)\\+\frac{m\tau}{h^2}\left(-(f_{N-\frac12}^{n-1})^m(\ln\varepsilon-\ln f_{N-1}^n)f_{N-1}^n+(f_{\frac12}^{n-1})^m(\ln f_1^n-\ln\varepsilon)f_1^n\right)=0.\end{aligned} \tag{4.15}$$

By (4.9), the third term on the left side is non-negative. Hence by using the identity

$$(a-b)a=\frac12\left(a^2-b^2+(a-b)^2\right), \tag{4.16}$$

it follows that

$$\sum_{i=1}^{I-1}|f_i^n|^2-\sum_{i=1}^{I-1}|f_i^{n-1}|^2\le-\frac{2m\tau}{h^2}\sum_{i=1}^{I-2}\frac{(f_{i+\frac12}^{n-1})^m}{\xi_i^n}|f_{i+1}^n-f_i^n|^2,$$

where $\xi_i^n>0$ takes a value between $f_i^n$ and $f_{i+1}^n$. Then we can obtain (4.13) from the above and (4.9). □

### 4.2. Second-order in time schemes

The fully discrete version of the modified second-order Crank-Nicolson scheme (3.31)-(3.32) is as follows:

- Compute $\{\widehat{f}_i^n\}$ from

$$\frac{\widehat{f}_i^n - f_i^{n-1}}{\tau} = \frac{m}{h^2}(\widetilde{f}_{i+\frac{1}{2}}^{n-\frac{1}{2}})^m \left(\ln\frac{\widehat{f}_{i+1}^n + f_{i+1}^{n-1}}{2} - \ln\frac{\widehat{f}_i^n + f_i^{n-1}}{2}\right) - \frac{m}{h^2}(\widetilde{f}_{i-\frac{1}{2}}^{n-\frac{1}{2}})^m \left(\ln\frac{\widehat{f}_i^n + f_i^{n-1}}{2} - \ln\frac{\widehat{f}_{i-1}^n + f_{i-1}^{n-1}}{2}\right), \tag{4.17}$$

with

$$\widetilde{f}_i^{n-\frac{1}{2}} := \begin{cases} \frac{1}{2}(3f_i^{n-1} - f_i^{n-2}), & \text{if } 3f_j^{n-1} > f_j^{n-2}, \forall j, \\ (f_i^{n-1})^{\frac{3}{2}}/(f_i^{n-2})^{\frac{1}{2}}, & \text{otherwise.} \end{cases}$$

If $\widehat{f}_i^n \geq \varepsilon\ \forall i$, set $f_i^n = \widehat{f}_i^n\ \forall i$ and go to the next time step.

- Otherwise, set

$$\widehat{f}_i^{n-\frac{1}{2}} = \frac{\widehat{f}_i^n + f_i^{n-1}}{2}, \ \forall i$$

and compute $\{\widehat{f}_i^{n+\frac{1}{2}}\}$ from

$$\frac{\widehat{f}_i^{n+\frac{1}{2}} - \widehat{f}_i^{n-\frac{1}{2}}}{\tau} = \frac{m}{h^2}(\widetilde{f}_{i+\frac{1}{2}}^n)^m \left(\ln\frac{\widehat{f}_{i+1}^{n+\frac{1}{2}} + \widehat{f}_{i+1}^{n-\frac{1}{2}}}{2} - \ln\frac{\widehat{f}_i^{n+\frac{1}{2}} + \widehat{f}_i^{n-\frac{1}{2}}}{2}\right) - \frac{m}{h^2}(\widetilde{f}_{i-\frac{1}{2}}^n)^m \left(\ln\frac{\widehat{f}_i^{n+\frac{1}{2}} + \widehat{f}_i^{n-\frac{1}{2}}}{2} - \ln\frac{\widehat{f}_{i-1}^{n+\frac{1}{2}} + \widehat{f}_{i-1}^{n-\frac{1}{2}}}{2}\right), \tag{4.18}$$

with

$$\widetilde{f}_i^n := (\widehat{f}_i^{n-\frac{1}{2}})^2 / f_i^{n-1}.$$

Then, we set $f_i^n = \frac{\widehat{f}_i^{n+\frac{1}{2}} + \widehat{f}_i^{n-\frac{1}{2}}}{2}$ for $i = 1, \cdots, I-1$ with conditions (4.2) and (4.3), and go to the next time step.

Following similar arguments in the proofs of Theorem 4.2, we can also establish the following results:

**Theorem 4.4.** *Given $\varepsilon \leq f_i^{n-1} \leq m_0 + \varepsilon$, $i = 1, \cdots, I-1$ with $f_0^{n-1} = f_N^{n-1} = \varepsilon$, then the solution of the second-order scheme (4.17)-(4.18) satisfies*

$$\varepsilon \leq f_i^n \leq m_0 + \varepsilon \quad \text{for } i = 1, \cdots, I-1, \tag{4.19}$$

*and*

$$\frac{1}{\tau}\sum_{i=1}^{I-1} f_i^n \leq \frac{1}{\tau}\sum_{i=1}^{I-1} f_i^{n-1}. \tag{4.20}$$

We can also construct a second-order scheme based on the second-order backward difference formula (BDF2) as follows:

$$\frac{3f_i^n - 4f_i^{n-1} + f_i^{n-2}}{2\tau} = \frac{m}{h^2}\left((\widetilde{f}_{i+\frac{1}{2}}^n)^m(\ln f_{i+1}^n - \ln f_i^n) - (\widetilde{f}_{i-\frac{1}{2}}^n)^m(\ln f_i^n - \ln f_{i-1}^n)\right), \ i = 1, \cdots, I-1, \tag{4.21}$$

with

$$\widetilde{f}_i^n := \begin{cases} 2f_i^{n-1} - f_i^{n-2}, & \text{if } 2f_j^{n-1} > f_j^{n-2}, \forall j, \\ (f_i^{n-1})^2/f_i^{n-2}, & \text{otherwise.} \end{cases}$$

For multi-dimensional cases, fully discretized schemes of first- and second-order can be constructed by using similar finite difference approaches on (3.5) and (3.26) with corresponding initial and boundary conditions. For the second-order BDF scheme (4.21), we are unable to prove a uniform bound for the solutions.

## 5. Numerical experiments

We present several numerical examples in this section to validate our theoretical results and demonstrate the effectiveness of the proposed schemes. For all examples, the standard Newton's iteration is employed to solve the nonlinear algebraic equations in each time step with stopping tolerance $10^{-12}$.

### 5.1. Accuracy test

First we test the accuracy of schemes by solving the Dirichlet problem of PME with a source term $g$:

$$\frac{\partial f}{\partial t} = \Delta(f^m) + g(x,t), \quad \text{in } (0,T)\times(-1,1), \tag{5.1}$$

such that the exact solution is given by

$$f(x,t) = \exp(-t)\cos(\pi x/2), \tag{5.2}$$

which is $C^\infty$ and supported in $[-1,1]$.

To evaluate the accuracy, we use the maximum error and discrete $L^2$ error defined by

$$e_\infty^N := \max_i \left| f_i^N - f(x_i,T) \right|, \quad e_2^N := \left( h\sum_{i=1}^{I-1} (f_i^N)^2 \right)^{\frac{1}{2}}. \tag{5.3}$$

We discretize the space with a high resolution $I = 8000$ and test our schemes with various $\tau$ and $\varepsilon$. The final time is set to $T = 1$. The errors for $\tau = \tau = 2^{-7}, \cdots, \tau = 2^{-10}$ and $\varepsilon = 10^{-4}\cdots, 10^{-10}$ by BDF1, BDF2 and modified C-N schemes are listed in Table 5.1 and 5.2. Furthermore, we list the average number of Newton's iterations in each time step in Table 5.3. It is worth mentioning that the Newton iteration does not guarantee the positivity of the intermediate solutions so negative intermediate solution may occur, although we have not meet such issues. One may need to use a more robust nonlinear solver (e.g. downhill algorithm) if such situation occurs.

We observe from these two tables that first-order convergence rate is achieved by BDF1, and second-order convergence rate is achieved by BDF2 and modified C-N for maximum and $L^2$ errors, as long as the perturbation parameter $\varepsilon$ is small enough (e.g. $\varepsilon = 10^{-8}$ or $10^{-10}$) so that the errors are dominated by the discretization error. On the other hand, it is clearly seen larger $\varepsilon$ will cause slower or no error decay (the cases of $\varepsilon = 10^{-4}$ or $10^{-6}$ for BDF2 and C-N).

### 5.2. Barenblatt solution

We consider the well-known Barenblatt solution [26], whose explicit expression is given by

$$B_{m,d}(x,t) = (t+1)^{-\alpha}\left(1 - \frac{\alpha(m-1)}{2md}\frac{|x|^2}{(t+1)^{2\alpha/d}}\right)_+^{1/(m-1)}, \tag{5.4}$$

**Table 5.1**
Maximum error for various $\tau$ and $\varepsilon$ by BDF1, BDF2 and C-N.

| $\varepsilon$ | $\tau$ | BDF1 | | BDF2 | | C-N | |
|---|---|---|---|---|---|---|---|
| | | $e_\infty^N$ | Order | $e_\infty^N$ | Order | $e_\infty^N$ | Order |
| $10^{-4}$ | $2^{-7}$ | 2.68e-02 | ~ | 2.49e-03 | ~ | 1.92e-03 | ~ |
| | $2^{-8}$ | 1.49e-02 | 0.85 | 1.16e-03 | 1.11 | 1.02e-03 | 0.91 |
| | $2^{-9}$ | 8.12e-03 | 0.87 | 8.30e-04 | 0.48 | 7.99e-04 | 0.35 |
| | $2^{-10}$ | 4.51e-03 | 0.85 | 7.55e-04 | 0.14 | 7.48e-04 | 0.10 |
| $10^{-6}$ | $2^{-7}$ | 2.63e-02 | ~ | 1.85e-03 | ~ | 1.27e-03 | ~ |
| | $2^{-8}$ | 1.43e-02 | 0.88 | 4.86e-04 | 1.93 | 3.37e-04 | 1.91 |
| | $2^{-9}$ | 7.46e-03 | 0.94 | 1.29e-04 | 1.92 | 9.11e-05 | 1.89 |
| | $2^{-10}$ | 3.82e-03 | 0.97 | 3.75e-05 | 1.78 | 2.81e-05 | 1.70 |
| $10^{-8}$ | $2^{-7}$ | 2.63e-02 | ~ | 1.84e-03 | ~ | 1.26e-03 | ~ |
| | $2^{-8}$ | 1.43e-02 | 0.88 | 4.80e-14 | 1.94 | 3.30e-04 | 1.93 |
| | $2^{-9}$ | 7.45e-03 | 0.94 | 1.22e-04 | 1.97 | 8.46e-05 | 1.96 |
| | $2^{-10}$ | 3.81e-03 | 0.97 | 3.10e-05 | 1.98 | 2.15e-05 | 1.97 |
| $10^{-10}$ | $2^{-7}$ | 2.63e-02 | ~ | 1.84e-03 | ~ | 1.26e-03 | ~ |
| | $2^{-8}$ | 1.43e-02 | 0.88 | 4.80e-04 | 1.94 | 3.30e-04 | 1.93 |
| | $2^{-9}$ | 7.45e-03 | 0.94 | 1.22e-04 | 1.97 | 8.45e-05 | 1.97 |
| | $2^{-10}$ | 3.81e-03 | 0.97 | 3.09e-05 | 1.98 | 2.15e-05 | 1.98 |

**Table 5.2**
$L^2$ error for various $\tau$ and $\varepsilon$ by BDF1, BDF2 and C-N.

| $\varepsilon$ | $\tau$ | BDF1 | | BDF2 | | C-N | |
|---|---|---|---|---|---|---|---|
| | | $e_2^N$ | Order | $e_2^N$ | Order | $e_2^N$ | Order |
| $10^{-4}$ | $2^{-7}$ | 1.13e-02 | ∼ | 6.73e-04 | ∼ | 5.43e-04 | ∼ |
| | $2^{-8}$ | 5.97e-03 | 0.93 | 3.64e-04 | 0.89 | 3.39e-04 | 0.68 |
| | $2^{-9}$ | 3.16e-03 | 0.92 | 2.99e-04 | 0.28 | 2.94e-04 | 0.21 |
| | $2^{-10}$ | 1.72e-03 | 0.87 | 2.85e-04 | 0.07 | 2.84e-04 | 0.05 |
| $10^{-6}$ | $2^{-7}$ | 1.11e-02 | ∼ | 4.72e-04 | ∼ | 3.21e-04 | ∼ |
| | $2^{-8}$ | 5.72e-03 | 0.96 | 1.22e-04 | 1.95 | 8.48e-05 | 1.92 |
| | $2^{-9}$ | 2.91e-03 | 0.98 | 3.23e-05 | 1.92 | 2.31e-05 | 1.88 |
| | $2^{-10}$ | 1.47e-03 | 0.99 | 9.63e-06 | 1.75 | 7.48e-06 | 1.63 |
| $10^{-8}$ | $2^{-7}$ | 1.11e-02 | ∼ | 4.70e-04 | ∼ | 3.20e-04 | ∼ |
| | $2^{-8}$ | 5.72e-03 | 0.96 | 1.21e-04 | 1.96 | 8.30e-05 | 1.94 |
| | $2^{-9}$ | 2.90e-03 | 0.98 | 3.06e-05 | 1.98 | 2.12e-05 | 1.97 |
| | $2^{-10}$ | 1.46e-03 | 0.99 | 7.75e-06 | 1.98 | 5.40e-06 | 1.97 |
| $10^{-10}$ | $2^{-7}$ | 1.11e-02 | ∼ | 4.70e-04 | ∼ | 3.20e-04 | ∼ |
| | $2^{-8}$ | 5.72e-03 | 0.96 | 1.21e-04 | 1.96 | 8.30e-05 | 1.95 |
| | $2^{-9}$ | 2.90e-03 | 0.98 | 3.06e-05 | 1.98 | 2.12e-05 | 1.97 |
| | $2^{-10}$ | 1.46e-03 | 0.99 | 7.73e-06 | 1.99 | 5.38e-06 | 1.98 |

**Table 5.3**
Average number of Newton's iterations in each time step for various $\tau$ and $\varepsilon$ by BDF1, BDF2 and C-N.

| $\varepsilon$ | $\tau$ | BDF1 | BDF2 | C-N |
|---|---|---|---|---|
| $10^{-4}$ | $2^{-7}$ | 3.01 | 3.01 | 3.01 |
| | $2^{-8}$ | 3 | 3 | 3 |
| | $2^{-9}$ | 3 | 2.82 | 2.49 |
| | $2^{-10}$ | 3 | 2.29 | 2.17 |
| $10^{-6}$ | $2^{-7}$ | 3.01 | 3.01 | 3.01 |
| | $2^{-8}$ | 3 | 3 | 2.94 |
| | $2^{-9}$ | 3 | 2.47 | 2.29 |
| | $2^{-10}$ | 2.94 | 2.12 | 2.06 |
| $10^{-8}$ | $2^{-7}$ | 3.01 | 3.01 | 3.01 |
| | $2^{-8}$ | 3 | 3 | 2.94 |
| | $2^{-9}$ | 3 | 2.47 | 2.3 |
| | $2^{-10}$ | 2.94 | 2.12 | 2.06 |
| $10^{-10}$ | $2^{-7}$ | 3.01 | 3.01 | 3.01 |
| | $2^{-8}$ | 3 | 3 | 2.94 |
| | $2^{-9}$ | 3 | 2.47 | 2.29 |
| | $2^{-10}$ | 2.94 | 2.12 | 2.06 |

where $(s)_+ = \max(s, 0)$, $\alpha = \frac{d}{d(m-1)+2}$ with $d$ being the dimension of the problem. Note that the solution is weakly singular at $|x| = 0$ and at the moving front where $1 - \frac{\alpha(m-1)}{2md}\frac{|x|^2}{(t+1)^{2\alpha/d}} = 0$.

**One-dimensional case.**

We solve the Dirichlet problem from $t = 0$ to $t = 1$ in the domain $\Omega = (-8, 8)$ using the first-order scheme (4.1), the second-order BDF scheme (4.21), and the second-order C-N scheme (4.17)-(4.18) with $m = 1.5, 3$, $\varepsilon = 10^{-8}, 10^{-10}$, and time step $\tau$ from $2^{-5}$ to $2^{-10}$. Since the exact solution is singular, we use a fine mesh ($I = 8000$) in space to resolve the singularity.

The error curves for all schemes are shown in Fig. 5.1. It can be observed the errors from second-order schemes (BDF2 and C-N), although not achieving second-order due to singularity, decay faster than the first-order scheme, both in maximum norm and $L^2$ norm. Also, similar results are obtained with $\varepsilon = 10^{-8}$ and $10^{-10}$, which implies that the range for the perturbation parameter between $10^{-10}$ and $10^{-8}$ is acceptable.

The exact and numerical solutions with $\varepsilon = 10^{-8}$ and $\tau = 2^{-10}$ at various time are plotted in Fig. 5.2, from which it is clearly observed that the support of solutions is expanding with finite speed. We also plot in Fig. 5.3 the time evolution of the positive interface point defined by

$$x_R(t) := \sup\{x : x > 0,\ f(x,t) > 10^{-2}\}. \tag{5.5}$$

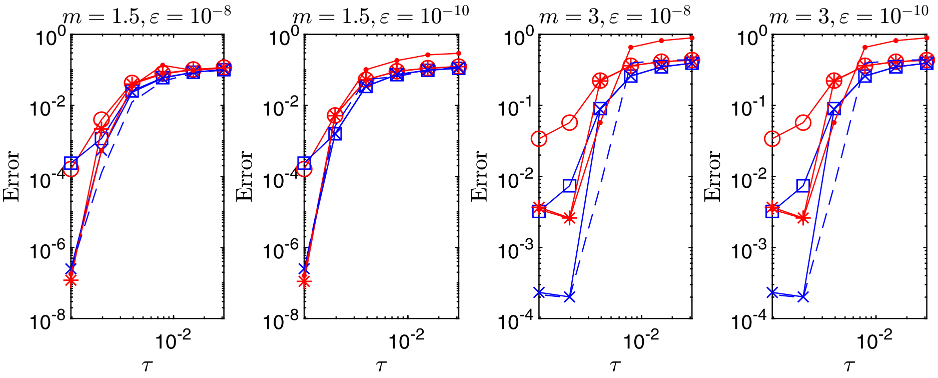


**Fig. 5.1.** Maximum error (red) by BDF1 (circle line), BDF2 (star line) and C-N (dotted line); $L^2$ error (blue) by BDF1 (square line), BDF2 (cross line) and C-N (dashed line). (For interpretation of the colors in the figures, the reader is referred to the web version of this article.)

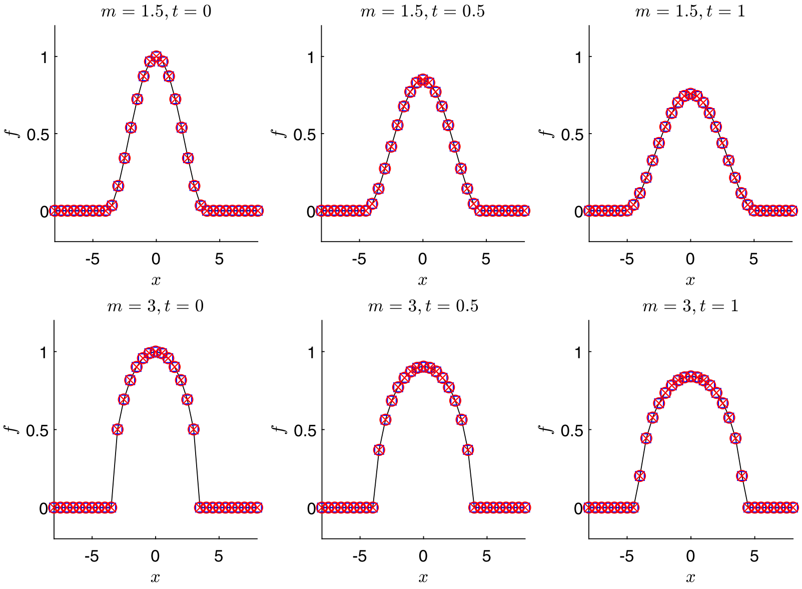


**Fig. 5.2.** The plot of Barenblatt solution for m=1.5 and 3 at t=0, 0.5 and 1 by various schemes (Exact: black solid curve; BDF1: blue square; BDF2: red circle; C-N: red cross).

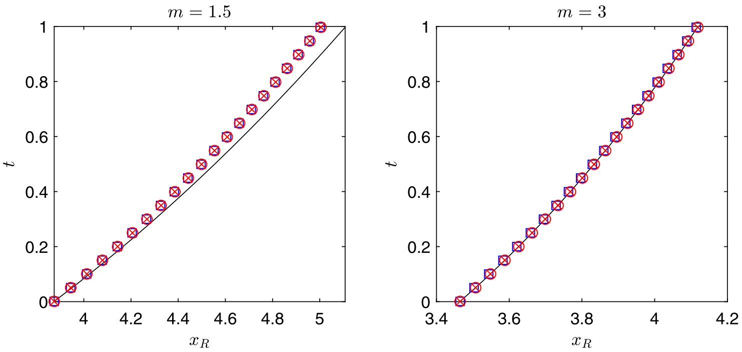


**Fig. 5.3.** Time evolution of the positive interface point $x_R$ for the Barenblatt solution for m=1.5 and 3 by various schemes (Exact: black solid curve; BDF1: blue square; BDF2: red circle; C-N: red cross).

**Table 5.4**
The difference between global minimum and $\varepsilon$ for various $m$, $\tau$ and $\varepsilon$ by BDF1, BDF2 and C-N.

| $\varepsilon$ | $\tau$ | $m = 1.5$ | | | $m = 3$ | | |
|---|---|---|---|---|---|---|---|
| | | BDF1 | BDF2 | C-N | BDF1 | BDF2 | C-N |
| $10^{-8}$ | $2^{-5}$ | -2.65e-22 | -2.90e-22 | -1.47e-22 | -6.46e-25 | -6.59e-25 | -2.58e-25 |
| | $2^{-6}$ | -2.42e-22 | -2.93e-22 | -9.26e-23 | -1.81e-25 | -2.58e-26 | -2.46e-25 |
| | $2^{-7}$ | -1.31e-22 | -2.40e-22 | -7.11e-23 | -6.46e-26 | -7.75e-25 | -1.29e-26 |
| | $2^{-8}$ | -1.04e-22 | -2.56e-22 | -3.80e-23 | -2.58e-26 | <1e-27 | -1.29e-26 |
| | $2^{-9}$ | -5.29e-23 | -2.91e-22 | -2.81e-23 | -1.29e-26 | -1.55e-25 | -2.58e-26 |
| | $2^{-10}$ | -1.82e-23 | -2.17e-22 | -9.93e-24 | <1e-27 | -7.75e-26 | 1e-27 |
| $10^{-10}$ | $2^{-5}$ | <1e-27 | <1e-27 | -5.13e-23 | <1e-27 | <1e-27 | <1e-27 |
| | $2^{-6}$ | <1e-27 | <1e-27 | <1e-27 | <1e-27 | -2.35e-24 | <1e-27 |
| | $2^{-7}$ | -2.10e-22 | -4.17e-22 | <1e-27 | -1.64e-24 | -1.29e-26 | <1e-27 |
| | $2^{-8}$ | <1e-27 | <1e-27 | <1e-27 | <1e-27 | <1e-27 | <1e-27 |
| | $2^{-9}$ | <1e-27 | <1e-27 | <1e-27 | -6.60e-24 | -3.88e-26 | -6.60e-24 |
| | $2^{-10}$ | <1e-27 | <1e-27 | <1e-27 | <1e-27 | -3.96e-23 | <1e-27 |

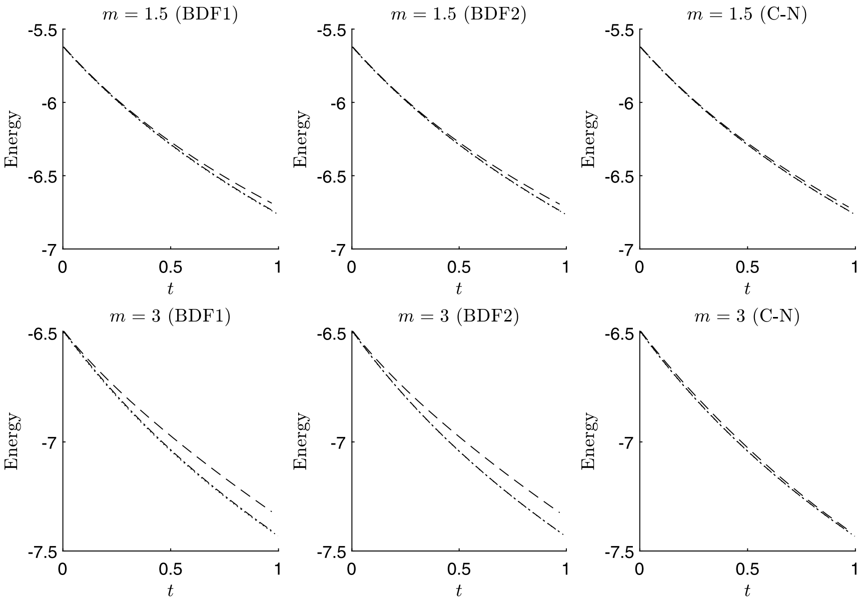


**Fig. 5.4.** The energy evolution of approximate Barenblatt solution for m=1.5 and 3 by various schemes ($\tau = 2^{-8}$: dashed curve; $\tau = 2^{-9}$: dotted dashed curve; $\tau = 2^{-10}$: dotted curve).

We observe that the accuracy when $m = 3$ is better than $m = 1.5$ since the solution is more regular as $m$ increases.

To verify the uniform boundedness of the numerical solution, we compute the global maximum and minimum by

$$M_f := \max\{f_i^n : 0 \le i \le I, 0 \le n \le N\}, \tag{5.6}$$

$$m_f := \min\{f_i^n : 0 \le i \le I, 0 \le n \le N\}. \tag{5.7}$$

It is shown in examples $M_f = 1$ for all $\tau$, $m$, $\varepsilon$ and all schemes. Actually, $\max\limits_i f_i^n$ is always decreasing over time, and hence $M_f$ is attained at the initial value. For the global minimum, we show the difference between $m_f$ and the theoretical lower bound $\varepsilon$ in Table 5.4, from which it can be observed $m_f - \varepsilon$ is numerically zero (within the machine precision), hence the uniform boundedness is indeed satisfied.

Finally, we track the discrete energy of the numerical solutions over time for various $\tau$. We define the discrete energy at $t = n\tau$ by

$$E_h^n := h\sum_{i=1}^{I-1} \left(f_i^n(\ln f_i^n - 1)\right). \tag{5.8}$$

The plot of energy evolution is presented in Fig. 5.4. We observe that the discrete energy of all schemes appears to be unconditionally dissipative, although the result is only proved for the first-order scheme.

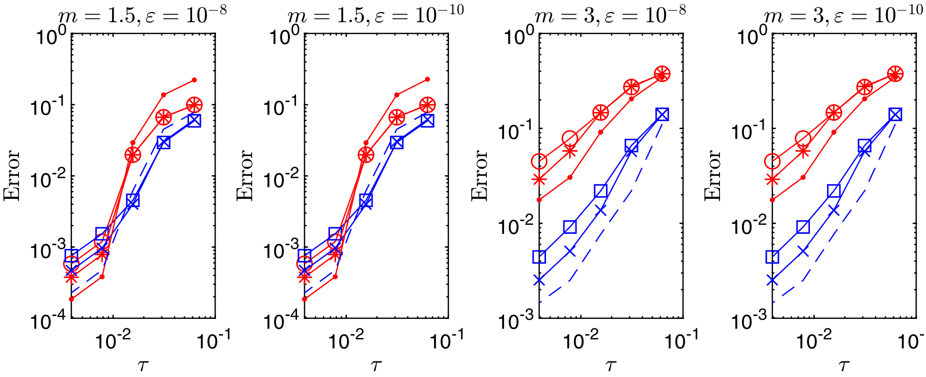


**Fig. 5.5.** Maximum error (red) by BDF1 (circle line), BDF2 (star line) and C-N (dotted line); $L^2$ error (blue) by BDF1 (square line), BDF2 (cross line) and C-N (dashed line).

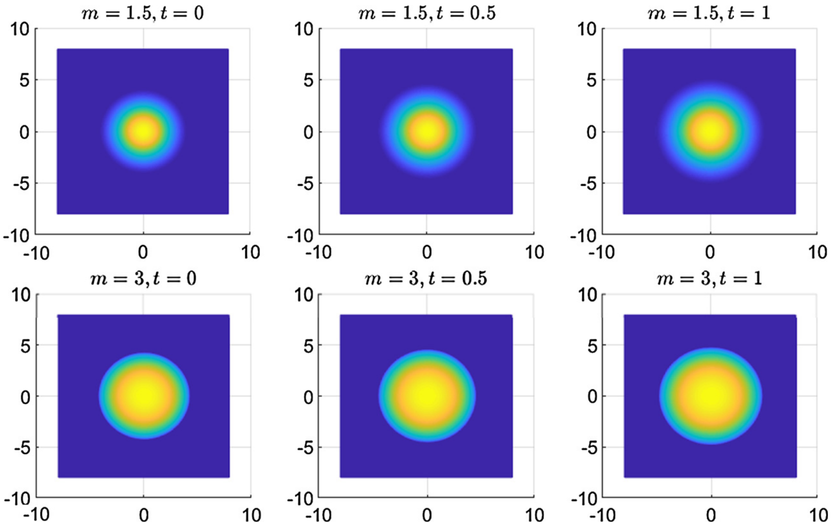


**Fig. 5.6.** The profile of Barenblatt solution computed by C-N for m=1.5 and 3 at t=0, 0.5 and 1.

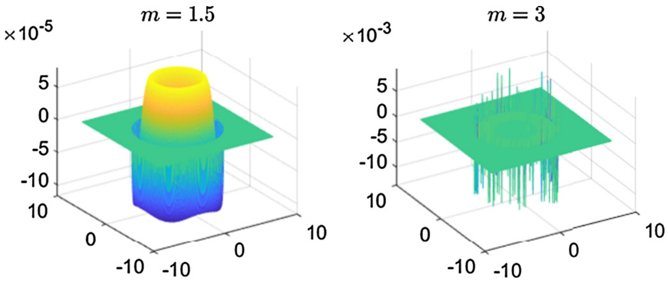


**Fig. 5.7.** The error profiles of Barenblatt solution computed by C-N for m=1.5 and 3 at t=1.

**Two-dimensional case.**

We consider the PME with the Barenblatt solution (5.4) for $m=1.5$ and 3 in the two-dimensional domain $\Omega=(-8,8)\times(-8,8)$ and the final time $T=1$. We test the effect of various time step $\tau$ from $2^{-4}$ to $2^{-8}$ with a high spatial resolution $I=1000$ in each direction. Also, we test the perturbation parameter $\varepsilon=10^{-8}$ and $10^{-10}$. Same as the 1-D case, the maximum error and the discrete $L^2$ error are taken as the metric. The errors obtained from various schemes are shown in Fig. 5.5.

The profiles of 2-D solution computed by modified C-N scheme with $\varepsilon=10^{-8}$ and $\tau=2^{-8}$ are shown in Fig. 5.6. We observe that the support is expanding with finite speed. The error profiles at the final time are plotted in Fig. 5.7. It can be observed the error is mainly distributed near the interface.

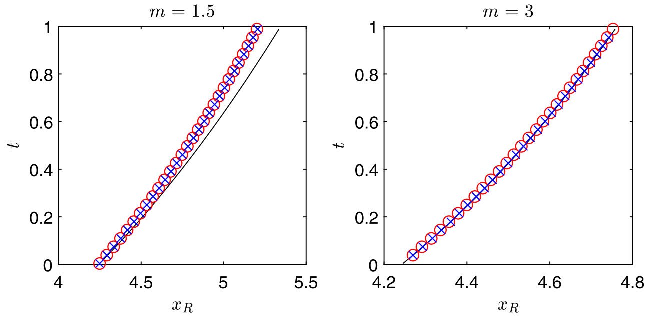


**Fig. 5.8.** Time evolution of the positive interface point $x_R$ for the Barenblatt solution for m=1.5 and 3 by various schemes (Exact: black solid curve; BDF1: blue square; BDF2: red circle; C-N: red cross).

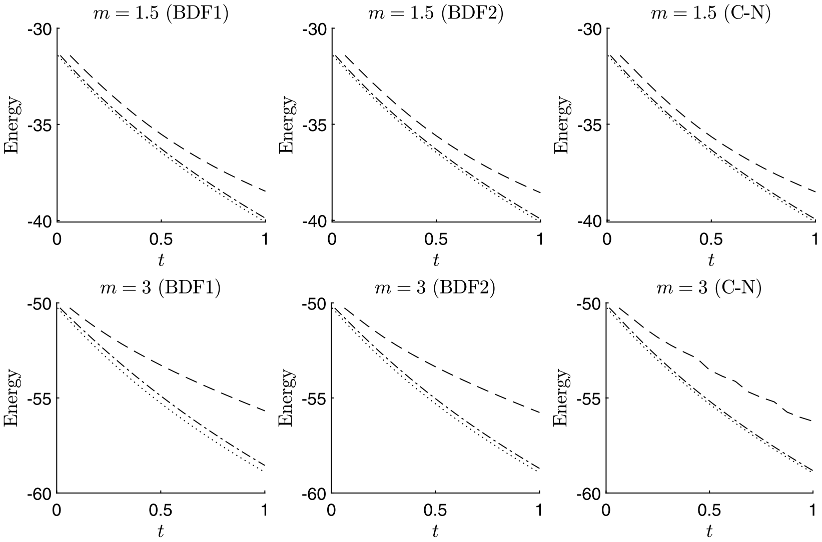


**Fig. 5.9.** The energy evolution of approximate Barenblatt solution for m=1.5 and 3 by various schemes ($\tau = 2^{-4}$: dashed curve; $\tau = 2^{-6}$: dotted dashed curve; $\tau = 2^{-8}$: dotted curve).

We also track and plot in Fig. 5.8 the interface point in the positive $x$-axis defined by

$$x_R(t) := \sup\{x : x > 0, f(x, 0, t) > 10^{-2}\}. \tag{5.9}$$

Finally, the plot of energy evolution is presented in Fig. 5.9. We observe again that the computed energy appears unconditionally dissipative in all cases.

### 5.3. Solution with waiting time

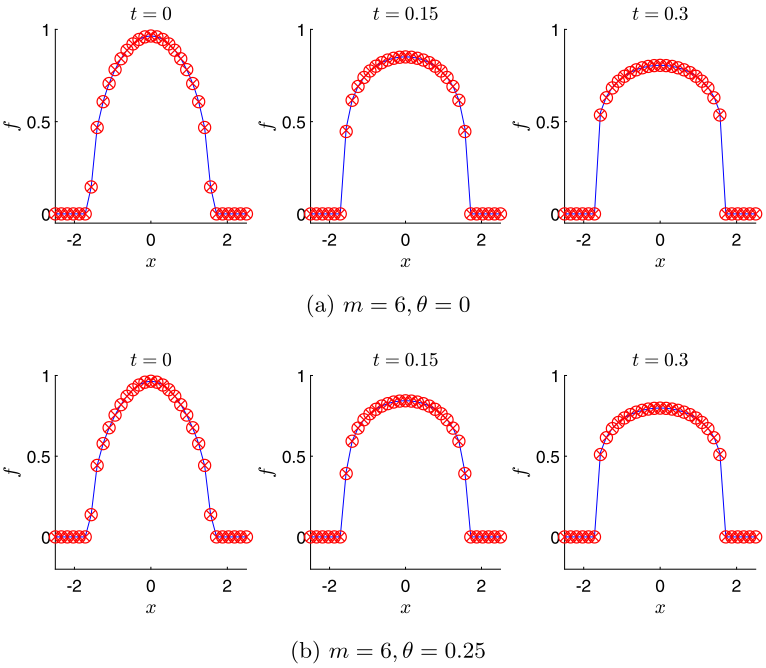
We now consider the 1-D PME with the following initial function

$$\rho_0(x) = \begin{cases} \left(\frac{m-1}{m}\left((1-\theta)\cos^2 x + \theta\cos^4 x\right)\right)^{\frac{1}{m-1}}, & \text{if } -\frac{\pi}{2} \le x \le \frac{\pi}{2}, \\ 0, \quad \text{otherwise.} \end{cases} \tag{5.10}$$

It is known that the solution with initial data (5.10) has the waiting time phenomenon for $0 \le \theta \le 1$, i.e., the interface will only move after a certain period of time from the start. Specifically, the theoretical waiting time for $0 \le \theta \le 1/4$ is given by [2]

$$t_{\text{waiting}} = \frac{1}{2(m+1)(1-\theta)}. \tag{5.11}$$

We solve the PME for $m = 6$ with initial data (5.10) for $\theta = 0$ or 0.25 by the BDF1, BDF2 and C-N schemes with $\varepsilon = 10^{-8}$, $\tau = 10^{-3}$ and $I = 8000$ as the parameters. The approximate solutions are plotted in Fig. 5.10. We observe that the support of profile remains unchanged while the shape of the profile is changing with time.



(a) $m=6, \theta=0$

(b) $m=6, \theta=0.25$

**Fig. 5.10.** The approximate solution with waiting time for m=6 and $\theta$=0, 0.25 at t=0, 0.15 and 0.3 by various schemes (BDF1: blue curve; BDF2: red circle; C-N: red cross).

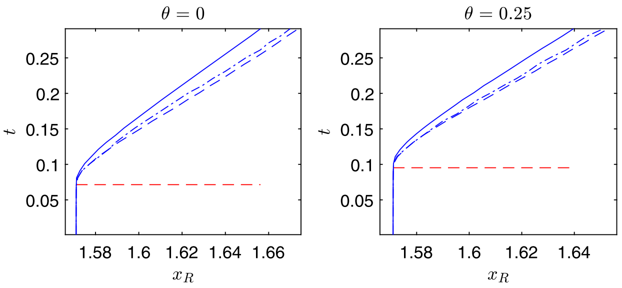


**Fig. 5.11.** Evolution of the right interface point $x_R$ for the 1-D approximate solution with waiting time for m=6 and $\theta$=0, 0.25 by various schemes (BDF1: solid curve; BDF2: dashed curve; C-N: dashed dotted curve). The red dashed line represents the theoretical waiting time.

We also plot the evolution of right interface point in Fig. 5.11 which clearly shows the waiting time phenomenon: the interface point stagnates until the waiting time.

Next, we consider the 2-D PME with waiting time for $m=3$ and 5 with the following initial function

$$\rho_0(x_1,x_2)=\begin{cases}\cos\left(\frac{\pi}{2}\sqrt{x_1^2+x_2^2}\right), & \text{if } x_1^2+x_2^2\le 1,\\ 0, & \text{otherwise.}\end{cases} \tag{5.12}$$

The evolution of the outer interface point at the $x$-axis are shown in Fig. 5.12. Similar to the 1-D case, the phenomenon of stagnation when $m=5$ is more significant than the case of $m=3$, and the waiting time of the former also appears longer.

### 5.3.1. *Solution having complex support*

Finally we perform two simulations of PME with complex initial data. First, we set the initial solution to be

$$\rho_0(x)=\begin{cases}\left(25(0.25^2-(|x|-0.75)^2)^{\frac{3}{2}}\right)^{\frac{1}{m-1}}, & \text{if } 0.5\le|x|\le 1, x_1<0, x_2<0,\\ \left(25(0.25^2-|x-(0,0.75)|^2)^{\frac{3}{2}}\right)^{\frac{1}{m-1}}, & \text{if } |x-(0,0.75)|\le 0.25, x_1\ge 0,\\ \left(25(0.25^2-|x-(0.75,0)|^2)^{\frac{3}{2}}\right)^{\frac{1}{m-1}}, & \text{if } |x-(0.75,0)|\le 0.25, x_2\ge 0,\\ 0, & \text{otherwise,}\end{cases} \tag{5.13}$$

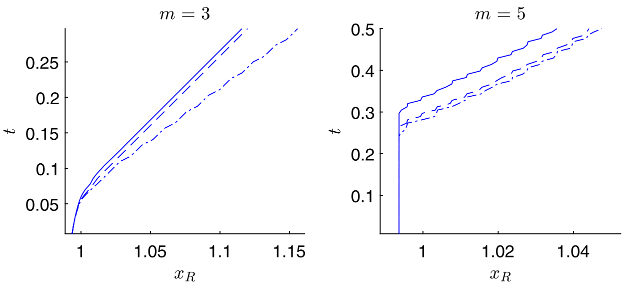


**Fig. 5.12.** Evolution of the outer interface point at the $x$-axis $x_R$ for the 2-D approximate solution with waiting time for m=3 and 5 by various schemes (BDF1: solid curve; BDF2: dashed curve; C-N: dashed dotted curve).

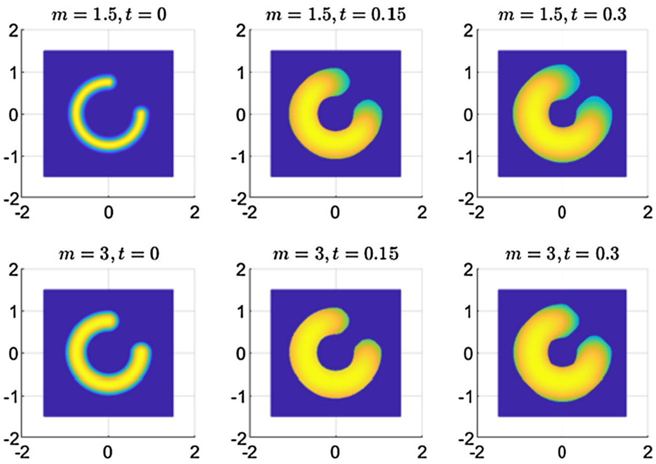


**Fig. 5.13.** Evolution of the approximate solution with compactly supported initial condition for m=1.5 and 3 at t=0, 0.15 and 0.3.

which has a partial donut shape as depicted in the first figure of Fig. 5.13, and was used in [3].

The second-order C-N scheme is utilized to solve the PME for $m = 1.5$ and 3 with initial data (5.13) in the 2-D domain $(-1.5, 1.5) \times (-1.5, 1.5)$ from $t = 0$ to $t = 0.3$. We set $\varepsilon = 10^{-8}$, $\tau = 2^{-7}$ and $I = 1000$. The numerical solutions at various time are shown in Fig. 5.13. It can be seen the support of the solution is spreading over time with a finite speed.

Second, we set the initial condition to be

$$\rho_0(x) = \begin{cases} \cos(\frac{\pi}{0.8}(x_1 - 0.4))\cos(\frac{\pi}{0.8}(x_2 - 0.4)), & \text{if } 0 \le x_1, x_2 \le 0.8, \\ \cos(\frac{\pi}{0.8}(x_1 + 0.4))\cos(\frac{\pi}{0.8}(x_2 + 0.4)), & \text{if } -0.8 \le x_1, x_2 \le 0, \\ 0, \quad \text{otherwise}, \end{cases} \tag{5.14}$$

whose support is two disconnected squares as shown in the first figure of Fig. 5.14, and was used in [15,5]. We set the domain to be $(-1.5, 1.5) \times (-1.5, 1.5)$ and the final time $T = 1$. We still apply the C-N scheme with the same parameters as in the previous example. The numerical solutions are presented in Fig. 5.14. We observe that the two disconnected squares move closer and eventually merge together. We also observe that the case with $m = 1.5$ spreads faster than that with $m = 3$.

## 6. Conclusion

We presented in this paper a class of bound preserving and energy dissipative schemes for the porous medium equation. Our schemes are based on a general approach for Wasserstein gradient flow developed in [23,22] and a perturbation technique [19,7]. We proved that both semi-discrete in time and fully discrete with finite difference schemes are uniquely solvable, bound preserving, and in the first-order case, also energy dissipative. We believe that these schemes are the first which is bound preserving as well as energy (entropy) dissipative.

Moreover, the schemes at each time step can be interpreted as Euler Lagrangian equations of convex functionals, so they can be efficiently solved by a Newton type iterative method with just a few iterations. Ample numerical results for well-known benchmark problems are presented to validate the theoretical results and demonstrate the effectiveness of the new schemes.

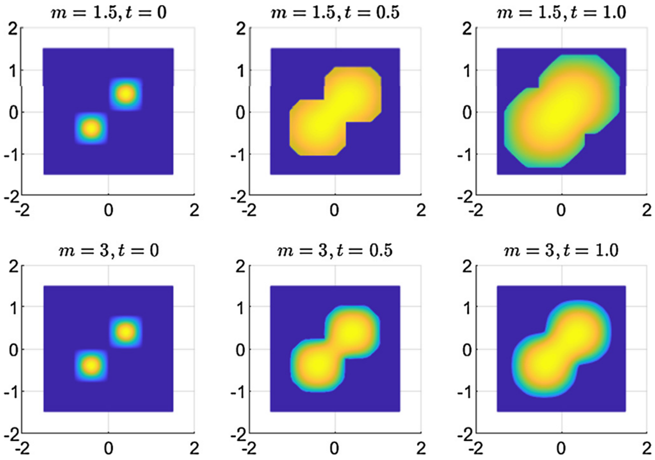


**Fig. 5.14.** Evolution of two peaks for m=1.5 and 3 at t=0, 0.5 and 1.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] Luigi Ambrosio, Nicola Gigli, Giuseppe Savaré, Gradient Flows in Metric Spaces and in the Space of Probability Measures, second edition, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2008.

[2] D. Aronson, L. Caffarelli, S. Kamin, How an initially stationary interface begins to move in porous medium flow, SIAM J. Math. Anal. 14 (4) (1983) 639–658.

[3] M.J. Baines, M.E. Hubbard, P.K. Jimack, A moving mesh finite element algorithm for the adaptive solution of time-dependent partial differential equations with moving boundaries, Appl. Numer. Math. 54 (3) (2005) 450–469.

[4] Stephen Boyd, Lieven Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[5] José A. Carrillo, Bertram Düring, Daniel Matthes, David S. McCormick, A Lagrangian scheme for the solution of nonlinear diffusion equations using moving simplex meshes, J. Sci. Comput. 75 (3) (Jun 2018) 1463–1499.

[6] José Antonio Carrillo, Yanghong Huang, Francesco Saverio Patacchini, Gershon Wolansky, Numerical study of a particle method for gradient flows, arXiv preprint, arXiv:1512.03029, 2015.

[7] J. Duque, R. Almeida, S. Antontsev, Convergence of the finite element method for the porous media equation with variable exponent, SIAM J. Numer. Anal. 51 (6) (2013) 3483–3504.

[8] Carsten Ebmeyer, Error estimates for a class of degenerate parabolic equations, SIAM J. Numer. Anal. 35 (3) (June 1998) 1095–1112.

[9] Carsten Ebmeyer, W.B. Liu, Finite element approximation of the fast diffusion and the porous medium equations, SIAM J. Numer. Anal. 46 (5) (June 2008) 2393–2410.

[10] E. Emmrich, D. Šiška, Full discretization of the porous medium/fast diffusion equation based on its very weak formulation, Commun. Math. Sci. 10 (2012) 1055–1080.

[11] Stefan Karpinski, Iuliu Sorin Pop, Florin Adrian Radu, Analysis of a linearization scheme for an interior penalty discontinuous Galerkin method for two-phase flow in porous media with dynamic capillarity effects, Int. J. Numer. Methods Eng. 112 (6) (2017) 553–577.

[12] A.A. Lacey, J.R. Ockendon, A.B. Tayler, "Waiting-time" solutions of a nonlinear diffusion equation, SIAM J. Appl. Math. 42 (6) (1982) 1252–1264.

[13] Florian List, Florin A. Radu, A study on iterative methods for solving Richards' equation, Comput. Geosci. 20 (2) (Apr 2016) 341–353.

[14] James D. Murray, Mathematical Biology, Springer-Verlag, New York, 2002.

[15] Cuong Ngo, Weizhang Huang, A study on moving mesh finite element solution of the porous medium equation, J. Comput. Phys. 331 (2017) 357–380.

[16] Ricardo H. Nochetto, Claudio Verdi, Approximation of degenerate parabolic problems using a numerical integration, SIAM J. Numer. Anal. 25 (1988) 784–814.

[17] Benoît Perthame, Fernando Quirós, Juan-Luis Vázquez, The Hele-Shaw asymptotics for mechanical models of tumor growth, Arch. Ration. Mech. Anal. 212 (2014) 93–127.

[18] I.S. Pop, F. Radu, P. Knabner, Mixed finite elements for the Richards' equation: linearization procedure, J. Comput. Appl. Math. 168 (1–2) (July 2004) 365–373.

[19] Iuliu Sorin Pop, Wen-An Yong, A numerical approach to degenerate parabolic equations, Numer. Math. 92 (2) (Aug 2002) 357–381.

[20] Florin Adrian Radu, Jan Martin Nordbotten, Iuliu Sorin Pop, Kundan Kumar, A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media, J. Comput. Appl. Math. 289 (Complete) (2015) 134–141.

[21] Jim Rulla, Noel J. Walkington, Optimal rates of convergence for degenerate parabolic problems in two dimensions, SIAM J. Numer. Anal. 33 (1) (February 1996) 56–67.

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

[23] Jie Shen, Jie Xu, Unconditionally bound preserving and energy dissipative schemes for a class of Keller–Segel equations, submitted for publication, 2019.

[24] Marián Slodicka, A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media, SIAM J. Sci. Comput. 23 (5) (May 2001) 1593–1614.

[25] Juan Luis Vázquez, An introduction to the mathematical theory of the porous medium equation, in: Shape Optimization and Free Boundaries, Springer, 1992, pp. 347–389.

[26] Juan Luis Vazquez, The Porous Medium Equation: Mathematical Theory, Clarendon Press, 2007.

[27] Lin Wang, Haijun Yu, Convergence analysis of an unconditionally energy stable linear Crank-Nicolson scheme for the Cahn-Hilliard equation, J. Math. Study 51 (2018) 89–114.

[28] Dongming Wei, Lew Lefton, A priori $L^p$ error estimates for Galerkin approximations to porous medium and fast diffusion equations, Math. Comput. 68 (227) (July 1999) 971–989.

[29] Michael Westdickenberg, Jon Wilkening, Variational particle schemes for the porous medium equation and for the system of isentropic Euler equations, ESAIM: Math. Model. Numer. Anal. 44 (1) (2010) 133–166.

[30] T.P. Witelski, Segregation and mixing in degenerate diffusion in population dynamics, J. Math. Biol. 35 (1997) 695–712.

[31] Qiang Zhang, Zi-Long Wu, Numerical simulation for porous medium equation by local discontinuous Galerkin finite element method, J. Sci. Comput. 38 (2) (Feb 2009) 127–148.