Available online at www.sciencedirect.com

ScienceDirect

Comput. Methods Appl. Mech. Engrg. 391 (2022) 114585

Computer methods in applied mechanics and engineering

www.elsevier.com/locate/cma

A new Lagrange multiplier approach for constructing structure preserving schemes, I. Positivity preserving[☆]

Qing Cheng, Jie Shen

Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA

Received 4 August 2021; received in revised form 1 January 2022; accepted 3 January 2022
Available online xxxx

Abstract

We propose a new Lagrange multiplier approach to construct positivity preserving schemes for parabolic type equations. The new approach introduces a space–time Lagrange multiplier to enforce the positivity with the Karush–Kuhn–Tucker (KKT) conditions. We then use a predictor–corrector approach to construct a class of positivity schemes: with a generic semi-implicit or implicit scheme as the prediction step, and the correction step, which enforces the positivity, can be implemented with negligible cost. We also present a modification which allows us to construct schemes which, in addition to positivity preserving, is also mass conserving. This new approach is not restricted to any particular spatial discretization and can be combined with various time discretization schemes. We establish stability results for our first- and second-order schemes under a general setting, and present ample numerical results to validate the new approach.

MSC: 65M70; 65K15; 65N22

Keywords: Positivity preserving; Mass conserving; KKT conditions; Lagrange multiplier; Stability

1. Introduction

Solutions for a large class of partial differential equations (PDEs) arising from sciences and engineering applications, e.g., solutions for physical variables such as density, concentration, height, population, etc., are required to be positive. It is of critical importance that their numerical approximations preserve the positivity of these variables at the discrete level, as violation of the positivity may render the discrete problems ill posed, although the original problems are well posed.

In recent years, a large effort has been devoted to construct positivity preserving schemes for various problems. The existing approaches can be roughly classified into the following categories:

- Cut-off approach: an ad-hoc approach which simply cuts off the values outside of the desired range. This approach is perhaps used in many simulations without being explicitly mentioned, and it is recently analyzed in [1,2] for certain class of time discretization schemes. The main advantages of the cut-off approach is (i) simple to implement, and (ii) it is able to preserve the accuracy of the underlying numerical schemes for problems with smooth solutions (cf. [2]). A disadvantage is that it does not preserve mass.

[☆] This work is partially supported by NSF Grant DMS-2012585 and AFOSR Grant FA9550-20-1-0309.
E-mail addresses: cheng573@purdue.edu (Q. Cheng), shen7@purdue.edu (J. Shen).

https://doi.org/10.1016/j.cma.2022.114585

- Discrete maximum principle preserving schemes (see, for instance, [3] and the references therein): these schemes are usually based on second-order finite differences or piecewise linear finite elements so they are limited in accuracy, see however some recent work on fourth-order finite differences [4,5] applied to second-order elliptic or parabolic equations.
- Post-processing approach: sophisticated procedures are designed in [6,7] for hyperbolic systems: these are explicit schemes which are not quite suitable for parabolic systems.
- Convex splitting approach: for examples, see [8] for Cahn–Hilliard equations with logarithmic potential, [9,10] for Poisson–Nernst–Planck (PNP) and Keller–Segel equations. The drawback of this approach is that a nonlinear system has to be solved at each time step.
- Reformulation approach: reformulate the problem so that solution of the corresponding discrete problem is always positive, see, for instance, [9,11].

We observe that there is still a lack of more general and efficient numerical approach which can deal with a large class of positivity preserving parabolic systems. Recently, an interesting Lagrange multiplier approach was proposed in [12] and applied to solve positivity preserving parabolic systems. The key idea was to introduce a Lagrange multiplier and use the well-known Karush–Kuhn–Tucker (KKT) conditions [13–16] to enforce the positivity. At each time step, their approach reduces to solving a nonlinear constrained minimization problem. But since these constrained minimization problems are only semi-smooth, a delicate and costly iterative method has to be used. The main goal of this paper is to adopt a predictor–corrector approach to develop efficient and accurate schemes for a large class of positivity preserving parabolic systems without solving a constrained minimization problem at each time step.

More precisely, we consider in this paper a class of linear or nonlinear parabolic equations with positive solutions in the following form:

$$u_t + \mathcal{L}u = 0, \tag{1.1}$$

with suitable initial and boundary conditions, where $\mathcal{L}$ could be $\mathcal{L}u = Au + f(u)$ with A being a linear or nonlinear positive operator and $f(u)$ a semi-linear or quasi-linear operator. Consider a generic spatial discretization of (1.1):

$$\partial_t u_h + \mathcal{L}_h u_h = 0, \tag{1.2}$$

where u_h is in certain finite dimensional approximation space X_h and $\mathcal{L}_h$ is a certain approximation of $\mathcal{L}$. In general, u_h, if it exists, may not preserve positivity. Oftentimes, (1.2) may not be well posed if u_h cannot preserve positivity, e.g., a direct finite elements or spectral approximation to the porous media equation [17] $u_t - m\nabla \cdot (u^{m-1}\nabla u) = 0$ with ($m > 1$) cannot preserve positivity so it is not well posed. Hence, special efforts have to be devoted to construct spatial discretization such that (1.2) is positivity preserving.

Alternatively, we can introduce a Lagrange multiplier function $\lambda_h(x, t)$ and solve the following expanded system:

$$\begin{aligned} &\partial_t u_h + \mathcal{L}_h u_h = \lambda_h, \\ &\lambda_h \geq 0, \ u_h \geq 0, \ \lambda_h u_h = 0. \end{aligned} \tag{1.3}$$

Note that the second equation in the above represents the well-known Karush–Kuhn–Tucker (KKT) conditions [14–16,18] for constrained minimizations. The expanded system (1.3) is equivalent with system (1.2) with positivity constraint under suitable conditions. In the absence of time variable, the problem (1.3) has been well studied mathematically and numerically. However, how to efficiently solve the time dependent (1.3) is a completely different issue which has not received much attention. One can of course use an implicit discretization scheme such as backward-Euler or Backward Difference Formula (BDF) [19–21] schemes so that at each time step, the nonlinear system can still be interpreted as a constrained minimization and apply a suitable iterative procedure. But since these constrained minimization problems are only semi-smooth, a delicate and costly iterative method has to be used. We refer to [12] for such an attempt with a diagonally implicit Runge–Kutta discretization.

To avoid solving a constrained minimization problem at each time step, we adopt in this paper a predictor–corrector approach to develop efficient and accurate schemes for (1.3). This approach enjoys the following advantages:

- It allows us to construct positivity preserving schemes for a large class of linear or nonlinear parabolic equations with positive solutions, and the schemes can also be made mass conservative if the PDE is mass conserving;

- It can be combined with most existing numerical schemes — particularly legacy codes which are not necessarily positivity preserving;
- It has essentially the same computational cost as the corresponding semi-implicit or implicit scheme with the same spatial discretization;
- It has good stability property: the first- and second-order versions of our scheme are proven to be unconditionally stable for a large class of problems.

Moreover, we show that schemes based on the ad-hoc cut-off approach can be interpreted as special cases of our approach. Thus, this approach allows us to construct mass conserving schemes based on the cut-off approach, and our analysis leads to new stability results for the cut-off approach. We shall apply our new schemes to a variety of problems with positive solutions, including the challenging porous media equation [17,22] and the very challenging Lubrication equation [23].

The rest of the paper is organized as follows. In Section 2, we introduce the positivity and mass preserving schemes with Lagrange multiplier. In Section 3, we carry out stability analysis for the proposed positivity preserving schemes. In Section 4, we present numerical results for a variety of problems to validate our schemes. Some concluding remarks are given in Section 5.

2. Positivity and mass preserving schemes with Lagrange multiplier

We start with a general description of the spatial discretization, followed by the construction of positivity preserving time discretization schemes without and with mass conservation.

2.1. Spatial discretization

We now give a more precise description on the generic spatial discretization in (1.3). Let $\bar{\Omega}$ be the domain including boundary and Σ_h be a set of mesh points or collocation points in $\bar{\Omega}$. Note that Σ_h should not include the points at the part of the boundary where a Dirichlet (or essential) boundary condition is prescribed, while it should include the points at the part of the boundary where a Neumann or mixed (or non-essential) boundary condition is prescribed.

We consider a Galerkin type discretization with finite-elements or spectral methods or finite-differences with summation-by-parts in a subspace $X_h \subset X$, and define a discrete inner product, i.e. numerical integration, on $\Sigma_h = \{z\}$ in $\bar{\Omega}$:

$$[u, v] = \sum_{z \in \Sigma_h} \omega_z u(z) v(z), \tag{2.1}$$

where we require that the weights $\omega_z > 0$. We also denote the induced norm by $\|u\| = [u, u]^{\frac{1}{2}}$. For finite element methods, the sum should be understood as $\sum_{K \subset \mathcal{T}} \sum_{z \in Z(K)}$ where $\mathcal{T}$ is a given triangulation. We assume that there is a unique function $\psi_z(\boldsymbol{x})$ in X_h satisfying $\psi_z(z') = \delta_{zz'}$ for $z, z' \in \Sigma_h$. Then, (1.3) is interpreted as follows: Find $u_h \in X_h$ such that

$$\begin{aligned}
&\partial_t u_h(z, t) + \mathcal{L}_h u_h(z, t) = \lambda_h(z, t), \quad \forall z \in \Sigma_h, \\
&\lambda_h(z, t) \geq 0,\ u_h(z, t) \geq 0,\ \lambda_h(z, t) u_h(z, t) = 0, \quad \forall z \in \Sigma_h, \\
&u_h(z, 0) = u_h^0(z),
\end{aligned} \tag{2.2}$$

with the Dirichlet boundary condition to be satisfied pointwise if the original problem includes Dirichlet boundary condition at part or all of boundary. u_h^0 is the initial condition.

2.2. Time discretization

Let $\mathcal{L}_h^n$ be an approximate operator of $\mathcal{L}$ at t_n. For examples, if $\mathcal{L}u = -\nabla \cdot (f(u)\nabla u)$, $\mathcal{L}_h^n$ could be a lagged linear approximation

$$\mathcal{L}_h^n \tilde{u}_h^{n+1} := -\nabla \cdot (f(u_h^n) \nabla \tilde{u}_h^{n+1}), \tag{2.3}$$

and for $\mathcal{L}u = Au + f(u)$, $\mathcal{L}_h^n$ could be a fully implicit discretization

$$\mathcal{L}_h^n \tilde{u}_h^{n+1} := A\tilde{u}_h^{n+1} + f(\tilde{u}_h^{n+1}), \tag{2.4}$$

or an implicit–explicit (IMEX) discretization,

$$\mathcal{L}_h^n \tilde{u}_h^{n+1} := A\tilde{u}_h^{n+1} + f(\tilde{u}_h^{n}), \tag{2.5}$$

or some other type of discretization such as the convex splitting [24] or the SAV approach [25,26].

2.2.1. First-order operator splitting scheme

Let $\lambda_h^0 \equiv 0$, for $n \geq 0$, we proceed as follows.

Step 1: solve $\tilde{u}_h^{n+1}$ from

$$\frac{\tilde{u}_h^{n+1}(z) - u_h^n(z)}{\delta t} + \mathcal{L}_h^n \tilde{u}_h^{n+1}(z) = 0, \quad \forall z \in \Sigma_h; \tag{2.6}$$

Step 2: solve $(u_h^{n+1}, \lambda_h^{n+1})$ from

$$\begin{aligned} &\frac{u_h^{n+1}(z) - \tilde{u}_h^{n+1}(z)}{\delta t} = \lambda_h^{n+1}(z), \quad \forall z \in \Sigma_h, \\ &\lambda_h^{n+1}(z) \geq 0,\ u_h^{n+1}(z) \geq 0,\ \lambda_h^{n+1}(z)u_h^{n+1}(z) = 0, \quad \forall z \in \Sigma_h. \end{aligned} \tag{2.7}$$

The above scheme can be viewed as an operator splitting method. The first step is just a usual time stepping scheme and can be implemented as usual. However, $\tilde{u}_h^{n+1}$ may not be positive. In the second step, we use the KKT conditions to enforce the positivity of u_h^{n+1}.

A remarkable property of (2.7) is that it can be solved pointwise as follows:

$$(u_h^{n+1}(z), \lambda_h^{n+1}(z)) = \left\{ \begin{array}{ll} (\tilde{u}_h^{n+1}(z), 0) & \text{if} \quad 0 < \tilde{u}_h^{n+1}(z) \\ (0, -\frac{\tilde{u}_h^{n+1}(z)}{\delta t}) & \text{otherwise} \end{array} \right. , \quad \forall z \in \Sigma_h. \tag{2.8}$$

Remark 2.1. The second step in the above scheme is equivalent to the simple cut-off approach [1,2]:

$$u_h^{n+1}(z) = \begin{cases} \tilde{u}_h^{n+1}(z) & \text{if } \ \tilde{u}_h^{n+1}(z) > 0 \\ 0 & \text{if } \ \tilde{u}_h^{n+1}(z) \leq 0, \end{cases} \quad \forall z \in \Sigma_h. \tag{2.9}$$

Hence, the cut-off approach can also be understood from an operator splitting point of view which opens new avenue for analysis and further algorithm improvement.

2.2.2. Higher-order schemes

We can construct higher-order schemes by using a predictor–corrector approach. More precisely, a kth-order IMEX scheme based on Backward Difference Formula (BDF) and Adam–Bashforth [27] can be constructed as follows:

Step 1 (prediction): solve $\tilde{u}_h^{n+1}$ from

$$\frac{\alpha_k \tilde{u}_h^{n+1} - A_k(u_h^n)}{\delta t} + \mathcal{L}_h^n \tilde{u}_h^{n+1} = B_{k-1}(\lambda_h^n); \tag{2.10}$$

Step 2 (correction): solve $(u_h^{n+1}, \lambda_h^{n+1})$ from

$$\begin{aligned} &\frac{\alpha_k (u_h^{n+1}(z) - \tilde{u}_h^{n+1}(z))}{\delta t} = \lambda_h^{n+1}(z) - B_{k-1}(\lambda_h^n), \quad \forall z \in \Sigma_h, \\ &\lambda_h^{n+1}(z) \geq 0,\ u_h^{n+1}(z) \geq 0,\ \lambda_h^{n+1}(z)u_h^{n+1}(z) = 0, \quad \forall z \in \Sigma_h, \end{aligned} \tag{2.11}$$

where α_k, the operators A_k and B_{k-1} ($k = 2, 3, 4$) are given by:

First-order:

$$\alpha_1 = 1, \quad A_1(u_h^n) = u_h^n, \quad B_0(\lambda_h^n) = 0; \tag{2.12}$$

Second-order:

$$\alpha_2=\frac{3}{2},\quad A_2(u_h^n)=2u_h^n-\frac{1}{2}u_h^{n-1},\quad B_1(\lambda_h^n)=\lambda_h^n; \tag{2.13}$$

Third-order:

$$\alpha_3=\frac{11}{6},\quad A_3(u_h^n)=3u_h^n-\frac{3}{2}u_h^{n-1}+\frac{1}{3}u_h^{n-2},\quad B_2(\lambda_h^n)=2\lambda_h^n-\lambda_h^{n-1}; \tag{2.14}$$

Fourth-order:

$$\alpha_4=\frac{25}{12},\; A_4(u_h^n)=4u_h^n-3u_h^{n-1}+\frac{4}{3}u_h^{n-2}-\frac{1}{4}u_h^{n-3},\; B_3(\lambda_h^n)=3\lambda_h^n-3\lambda_h^{n-1}+\lambda_h^{n-2}. \tag{2.15}$$

The formulae for $k=5,6$ can be derived similarly with Taylor expansions. For the sake of simplicity and with a slight abuse of notations, we used $A_k(u_h^n)$ and $B_k(u_h^n)$ to denote $A_k(u_h^n,\dots,u_h^{n-k+1})$ and $B_k(u_h^n,\dots,u_h^{n-k+1})$, respectively. Note that for $k=1$, the above scheme is exactly (2.6)–(2.7).

The first-step is a usual kth-order IMEX scheme. The second step (2.11) can be viewed as a correction step in which λ_h^{n+1} is introduced to enforce the pointwise positivity of u_h^{n+1}, and can be efficiently solved as follows:

$$(u_h^{n+1}(z),\lambda_h^{n+1}(z))=\begin{cases}(\tilde{u}_h^{n+1}(z)-\frac{\delta t}{\alpha_k}B_{k-1}(\lambda_h^n),0) & \text{if}\quad 0<\tilde{u}_h^{n+1}(z)-\frac{\delta t}{\alpha_k}B_{k-1}(\lambda_h^n)\\ (0,B_{k-1}(\lambda_h^n)-\frac{\alpha_k}{\delta t}\tilde{u}_h^{n+1}(z)) & \text{otherwise}\end{cases},\quad\forall z\in\Sigma_h. \tag{2.16}$$

Remark 2.2. Since λ_h^n is an approximation to λ_h which tends to zero as $h\to 0$, an alternative is to replace $B_{k-1}(\lambda_h^n)$ by zero, i.e., leading to the scheme:
Step 1: solve $\tilde{u}_h^{n+1}$ from

$$\frac{\alpha_k\tilde{u}_h^{n+1}-A_k(u_h^n)}{\delta t}+\mathcal{L}_h^n\tilde{u}_h^{n+1}=0; \tag{2.17}$$

Step 2: solve $(u_h^{n+1},\lambda_h^{n+1})$ from

$$\begin{aligned}&\frac{\alpha_k(u_h^{n+1}(z)-\tilde{u}_h^{n+1}(z))}{\delta t}=\lambda_h^{n+1}(z),\quad\forall z\in\Sigma_h,\\&\lambda_h^{n+1}(z)\geq 0,\; u_h^{n+1}(z)\geq 0,\; \lambda_h^{n+1}(z)u_h^{n+1}(z)=0,\quad\forall z\in\Sigma_h.\end{aligned} \tag{2.18}$$

Since the second step is once again equivalent to the cut-off approach (2.9), the above scheme can be interpreted as a kth-order cut-off scheme.

2.2.3. *Positivity preserving schemes with mass conservation*

A drawback of the scheme (2.10)–(2.11) is that it does not preserve mass if the original equation does. For clarity, we consider first the first-order scheme (2.6)–(2.7).

Let $\langle\mathcal{L}\cdot,\cdot\rangle$ (resp. $\langle\mathcal{L}_h^n\cdot,\cdot\rangle$) denote the continuous (resp. discrete) bilinear form after proper integration by parts, e.g., if $\mathcal{L}_h^n u_h:=-\nabla\cdot(f(u_h^n)\nabla u_h)$, then $\langle\mathcal{L}_h^n u_h,v_h\rangle:=[f(u_h^n)\nabla u_h,\nabla v_h]$. Assuming $\langle\mathcal{L}u,1\rangle=0$, we find from (1.1) that $\partial_t(u,1)=0$, i.e., the mass is conserved. But assuming $\langle\mathcal{L}_h^n v_h,1\rangle=0$ for any $v_h\in X_h$, we derive from (2.6)–(2.7) that

$$[u_h^{n+1},1]-[u_h^n,1]=\delta t[\lambda_h^{n+1},1].$$

Since $\lambda_h^{n+1}\geq 0$, we find that the mass is not conserved, in fact it is increasing with n.

We present below a simple modification which enables mass conservation. More precisely, we introduce another Lagrange multiplier ξ_h^{n+1}, which is independent of spatial variables, to enforce the mass conservation in the correction step.

Step 1: solve $\tilde{u}_h^{n+1}$ from

$$\frac{\tilde{u}_h^{n+1}(z)-u_h^n(z)}{\delta t}+\mathcal{L}_h^n\tilde{u}_h^{n+1}(z)=0,\quad\forall z\in\Sigma_h; \tag{2.19}$$

Step 2: solve $(u_h^{n+1},\lambda_h^{n+1})$ from

$$\frac{u_h^{n+1}(z)-\tilde{u}_h^{n+1}(z)}{\delta t}=\lambda_h^{n+1}(z)+\xi_h^{n+1},\quad\forall z\in\Sigma_h, \tag{2.20a}$$

$$\lambda_h^{n+1}(z) \ge 0,\ u_h^{n+1}(z) \ge 0,\ \lambda_h^{n+1}(z)u_h^{n+1}(z) = 0, \quad \forall z \in \Sigma_h, \tag{2.20b}$$

$$[u_h^{n+1}, 1] = [u_h^n, 1]. \tag{2.20c}$$

In order to solve (2.20), we rewrite (2.20a) in the following equivalent form

$$\frac{u_h^{n+1}(z) - (\tilde{u}_h^{n+1}(z) + \delta t \xi_h^{n+1})}{\delta t} = \lambda_h^{n+1}(z). \tag{2.21}$$

Hence, assuming ξ_h^{n+1} is known, (2.21) and (2.20b) can be solved pointwise as follows:

$$(u_h^{n+1}(z), \lambda_h^{n+1}(z)) = \begin{cases} (\tilde{u}_h^{n+1}(z) + \delta t \xi_h^{n+1}, 0) & \text{if} \quad 0 < \tilde{u}_h^{n+1}(z) + \delta t \xi_h^{n+1} \\ (0, -\frac{\tilde{u}_h^{n+1}(z)+\delta t \xi_h^{n+1}}{\delta t}) & \text{otherwise} \end{cases}, \quad \forall z \in \Sigma_h. \tag{2.22}$$

It remains to determine ξ_h^{n+1}. We find from (2.20c) and (2.21) that

$$[\tilde{u}_h^{n+1} + \delta t \xi_h^{n+1}, 1] = [u_h^n, 1] - \delta t[\lambda_h^{n+1}, 1],$$

which, thanks to (2.22), can be rewritten as

$$\sum_{z\in\Sigma_h\ s.t.\ 0<\tilde{u}_h^{n+1}(z)+\delta t\xi_h^{n+1}} (\tilde{u}_h^{n+1}(z) + \delta t \xi_h^{n+1})\omega_z = [u_h^n, 1].$$

Hence, ξ_h^{n+1} is a solution to the nonlinear algebraic equation

$$F(\xi) = \sum_{z\in\Sigma_h\ s.t.\ 0<\tilde{u}_h^{n+1}(z)+\delta t\xi} (\tilde{u}_h^{n+1}(z) + \delta t \xi)\omega_z - [u_h^n, 1] = 0. \tag{2.23}$$

Since $F'(\xi)$ may not exist and difficult to compute if it exists, instead of the Newton iteration, we can use the following secant method:

$$\xi_{k+1} = \xi_k - \frac{F(\xi_k)(\xi_k - \xi_{k-1})}{F(\xi_k) - F(\xi_{k-1})}. \tag{2.24}$$

Since ξ_h^{n+1} is an approximation to zero and it will be shown below that $\xi_h^{n+1} \le 0$, we can choose $\xi_0 = 0$ and $\xi_1 = -O(\delta t)$. In all our experiments, (2.24) converges in a few iterations so that the cost is negligible.

Once ξ_h^{n+1} is known, we can update $(u_h^{n+1}, \lambda_h^{n+1})$ with (2.22).

Similarly, the higher-order scheme (2.10)–(2.11) can be modified to preserve mass as follows:

Step 1 (prediction): solve $\tilde{u}_h^{n+1}$ from

$$\frac{\alpha_k \tilde{u}_h^{n+1} - A_k(u_h^n)}{\delta t} + \mathcal{L}_h^n \tilde{u}_h^{n+1} = B_{k-1}(\lambda_h^n) + B_{k-1}(\xi_h^n); \tag{2.25}$$

Step 2 (correction): solve $(u_h^{n+1}, \lambda_h^{n+1})$ from

$$\frac{\alpha_k(u_h^{n+1}(z) - \tilde{u}_h^{n+1}(z))}{\delta t} = \lambda_h^{n+1}(z) - B_{k-1}(\lambda_h^n) + \xi_h^{n+1} - B_{k-1}(\xi_h^n), \quad \forall z \in \Sigma_h, \tag{2.26a}$$

$$\lambda_h^{n+1}(z) \ge 0,\ u_h^{n+1}(z) \ge 0,\ \lambda_h^{n+1}(z)u_h^{n+1}(z) = 0, \quad \forall z \in \Sigma_h, \tag{2.26b}$$

$$[u_h^{n+1}, 1] = [u_h^n, 1]. \tag{2.26c}$$

In order to solve the above system, we denote $\eta_h^{n+1} := \frac{\delta t}{\alpha_k}(\xi_h^{n+1} - B_{k-1}(\xi_h^n) - B_{k-1}(\lambda_h^n))$ and rewrite (2.26a) as

$$\frac{\alpha_k(u_h^{n+1}(z) - (\tilde{u}_h^{n+1}(z) + \eta_h^{n+1}(z)))}{\delta t} = \lambda_h^{n+1}(z). \tag{2.27}$$

Assuming ξ_h^{n+1} is known, we find from (2.26a) and (2.27) that

$$(u_h^{n+1}(z), \lambda_h^{n+1}(z)) = \begin{cases} (\tilde{u}_h^{n+1}(z) + \eta_h^{n+1}, 0) & \text{if} \quad 0 < \tilde{u}_h^{n+1}(z) + \eta_h^{n+1}, \\ (0, -\frac{\alpha_k}{\delta t}(\tilde{u}_h^{n+1}(z) + \eta_h^{n+1}(z))) & \text{otherwise.} \end{cases} \tag{2.28}$$

Finally, we can determine ξ_h^{n+1} by solving the nonlinear algebraic equation

$$F(\xi_h^{n+1}) = \sum_{z\in\Sigma_h\ s.t.\ 0<\tilde{u}_h^{n+1}(z)+\eta_h^{n+1(z)}} (\tilde{u}_h^{n+1}(z) + \eta_h^{n+1}(z))\omega_z - [u_h^n, 1] = 0. \tag{2.29}$$

Remark 2.3. Replacing $B_{k-1}(\lambda_h^n)$ in (2.25)–(2.26) by zero, we obtain a mass conserved kth-order cut-off scheme.

3. Stability results

We prove in this section that the first- and second-order positivity preserving schemes with or without mass conservation are dissipative and unconditionally stable if $\langle \mathcal{L}_h^n v_h, v_h\rangle \geq 0\ \forall v_h \in X_h$ for all n.

3.1. First-order schemes

We consider first the scheme (2.6)–(2.7).

Theorem 3.1. *For the scheme (2.6)–(2.7), we have*

$$\|u_h^m\|^2 + \sum_{n=0}^{m-1}(\|\tilde{u}_h^{n+1} - u_h^n\|^2 + \delta t^2\|\lambda_h^{n+1}\|^2) + 2\delta t\sum_{n=0}^{m-1}\langle \mathcal{L}_h^n\tilde{u}_h^{n+1}, \tilde{u}_h^{n+1}\rangle = \|u_h^0\|^2, \quad \forall m \geq 1. \tag{3.1}$$

In particular, if for all n, $\langle \mathcal{L}_h^n v_h, v_h\rangle \geq 0\ \forall v_h \in X_h$, then the scheme (2.6)–(2.7) with $k = 1$ is dissipative and unconditionally stable.

Proof. Taking the discrete inner product of (2.6) with $2\delta t\tilde{u}_h^{n+1}$, we obtain

$$\|\tilde{u}_h^{n+1}\|^2 - \|u_h^n\|^2 + \|\tilde{u}_h^{n+1} - u_h^n\|^2 + 2\delta t\langle \mathcal{L}_h^n\tilde{u}_h^{n+1}, \tilde{u}_h^{n+1}\rangle = 0. \tag{3.2}$$

We rewrite (2.7) as

$$u_h^{n+1}(z) - \delta t\lambda_h^{n+1}(z) = \tilde{u}_h^{n+1}(z). \tag{3.3}$$

Taking the discrete inner product of each side of the above equation with itself, thanks to the last KKT condition in (2.7), we derive

$$\|u_h^{n+1}\|^2 + \delta t^2\|\lambda_h^{n+1}\|^2 = \|\tilde{u}_h^{n+1}\|^2.$$

Summing up the above with (3.2), we obtain

$$\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \delta t^2\|\lambda_h^{n+1}\|^2 + \|\tilde{u}_h^{n+1} - u_h^n\|^2 + 2\delta t\langle \mathcal{L}_h^n\tilde{u}_h^{n+1}, \tilde{u}_h^{n+1}\rangle = 0.$$

Summing up the above for n from 0 to $m-1$, we arrive at (3.1). In particular, if $\langle \mathcal{L}_h^n v_h, v_h\rangle \geq 0\ \forall v_h \in X_h$, then we have

$$\|u_h^{n+1}\|^2 + \delta t^2\|\lambda_h^{n+1}\|^2 + \|\tilde{u}_h^{n+1} - u_h^n\|^2 \leq \|u_h^n\|^2.$$

We derive from above that $\|u_h^{n+1}\|^2 \leq \|u_h^n\|^2$ for any n, so that the scheme (2.6)–(2.7) with $k = 1$ is unconditional stable and dissipative. □

Next, we consider the mass conserved scheme (2.19)–(2.20).

Theorem 3.2. *For the scheme (2.19)–(2.20), if $\langle \mathcal{L}_h^n v_h, 1\rangle = 0$ for any $v_h \in X_h$, we have*

$$\|u_h^m\|^2 + \sum_{n=0}^{m-1}(\|\tilde{u}_h^{n+1} - u_h^n\|^2 + \delta t^2\|\lambda_h^{n+1} + \xi_h^{n+1}\|^2) + 2\delta t\sum_{n=0}^{m-1}\langle \mathcal{L}_h^n\tilde{u}_h^{n+1}, \tilde{u}_h^{n+1}\rangle \leq \|u_h^0\|^2, \quad \forall m \geq 1. \tag{3.4}$$

In particular, if for all n, $\langle \mathcal{L}_h^n v_h, v_h\rangle \geq 0\ \forall v_h \in X_h$, then the scheme (2.19)–(2.20) is dissipative and unconditionally stable.

Proof. The proof follows the same procedure as that of Theorem 3.1. Indeed, we can replace (3.3) by

$$u_h^{n+1}(z) - \delta t(\lambda_h^{n+1}(z) + \xi_h^{n+1}) = \tilde{u}_h^{n+1}(z), \tag{3.5}$$

Taking the inner product of (3.5) with itself on both sides, we obtain

$$\|u_h^{n+1}(z)\|^2 + \delta t^2\|\lambda_h^{n+1}(z) + \xi_h^{n+1}\|^2 - 2\delta t[u_h^{n+1}(z), \xi_h^{n+1}] = \|\tilde{u}_h^{n+1}(z)\|^2. \tag{3.6}$$

Summing up (2.19) and (2.20a), we obtain

$$\frac{u_h^{n+1}(z)-u_h^n(z)}{\delta t}+\mathcal{L}_h^n\tilde{u}_h^{n+1}(z)=\lambda_h^{n+1}+\xi_h^{n+1},\quad \forall z\in\Sigma_h, \tag{3.7}$$

Taking the discrete inner product of (3.7) with 1 on both sides, using the fact that $\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},1\rangle=0$, we obtain

$$[\lambda_h^{n+1},1]+[\xi_h^{n+1},1]=0, \tag{3.8}$$

which implies that $\xi_h^{n+1}=-\frac{[\lambda_h^{n+1},1]}{|\Omega|}\le 0$ since $\lambda_h^{n+1}\ge 0$. Therefore,

$$-2\delta t[u_h^{n+1}(z),\xi_h^{n+1}]\ge 0.$$

Finally, summing up (3.6) with (3.2), we arrive at the following result

$$\|u_h^{n+1}\|^2-\|u_h^n\|^2+\delta t^2\|\lambda_h^{n+1}+\xi_h^{n+1}\|^2+\|\tilde{u}_h^{n+1}-u_h^n\|^2 \\ +\ 2\delta t\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle=2\delta t[u_h^{n+1}(z),\xi_h^{n+1}]\le 0.$$

Summing up the above for n from 0 to $m-1$, we arrive at the result (3.4). In particular, if $\langle\mathcal{L}_h^n v_h,v_h\rangle\ \ge 0\ \forall v_h\in X_h$, then we have

$$\|u_h^{n+1}\|^2+\delta t^2\|\lambda_h^{n+1}+\xi_h^{n+1}\|^2+\|\tilde{u}_h^{n+1}-u_h^n\|^2\le\|u_h^n\|^2.$$

We derive from above that $\|u_h^{n+1}\|^2\ \le\ \|u_h^n\|^2$ for any n, so that the scheme (2.19)–(2.20) is dissipative and unconditionally stable. □

3.2. Second-order schemes

We first consider the scheme (2.10)–(2.11) with $k=2$.

Theorem 3.3. *For the scheme (2.10)–(2.11) with $k\ =\ 2$, we assume that the first step is computed with the first-order scheme (2.6)–(2.7). Then, we have*

$$4\|u_h^m\|^2+\|2u_h^m-u_h^{m-1}\|^2+\frac{4}{3}\delta t^2\|\lambda_h^m\|^2+4\delta t\sum_{n=0}^{m-1}\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle\le\|2u_h^1-u_h^0\|^2+4\|u_h^0\|^2,\quad\forall m\ge 1. \tag{3.9}$$

In particular, if for all n, $\langle\mathcal{L}_h^n v_h,v_h\rangle\ \ge 0\ \forall v_h\in X_h$, then the scheme (2.10)–(2.11) with $k=2$ is dissipative and unconditionally stable.

Proof. Taking inner product of Eq. (2.10) (with $k=2$) with $4\delta t\tilde{u}_h^{n+1}$, we obtain

$$[3\tilde{u}_h^{n+1}-4u_h^n+u_h^{n-1},2\tilde{u}_h^{n+1}]+4\delta t\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle=4\delta t[\lambda_h^n,\tilde{u}_h^{n+1}]. \tag{3.10}$$

The term on the left can be written as

$$[3\tilde{u}_h^{n+1}-4u_h^n+u_h^{n-1},2\tilde{u}_h^{n+1}]=2[3u_h^{n+1}-4u_h^n+u_h^{n-1},u_h^{n+1}] \\ +\ 2[3u_h^{n+1}-4u_h^n+u_h^{n-1},\tilde{u}_h^{n+1}-u_h^{n+1}]+6[\tilde{u}_h^{n+1}-u_h^{n+1},\tilde{u}_h^{n+1}]. \tag{3.11}$$

For the first term on the righthand side of (3.11), we have

$$2[3u_h^{n+1}-4u_h^n+u_h^{n-1},u_h^{n+1}]=\|u_h^{n+1}\|^2-\|u_h^n\|^2+\|2u_h^{n+1}-u_h^n\|^2 \\ -\ \|2u_h^n-u_h^{n-1}\|^2+\|u_h^{n+1}-2u_h^n+u_h^{n-1}\|^2. \tag{3.12}$$

For the last term in (3.11), we have

$$6[\tilde{u}_h^{n+1}-u_h^{n+1},\tilde{u}_h^{n+1}]=3(\|\tilde{u}_h^{n+1}\|^2-\|u_h^{n+1}\|^2+\|\tilde{u}_h^{n+1}-u_h^{n+1}\|^2). \tag{3.13}$$

And for the second term on the righthand side of (3.11). Similarly, we have

$$2[3u_h^{n+1}-4u_h^n+u_h^{n-1},\tilde{u}_h^{n+1}-u_h^{n+1}]=2[u_h^{n+1}-2u_h^n+u_h^{n-1},\tilde{u}_h^{n+1}-u_h^{n+1}] \\ +\ 4[u_h^{n+1}-u_h^n,\tilde{u}_h^{n+1}-u_h^{n+1}]. \tag{3.14}$$

By Cauchy–Schwartz inequality, the first term on the righthand side of (3.14) can be bounded by

$$2[u_h^{n+1}-2u_h^n+u_h^{n-1},\tilde{u}_h^{n+1}-u_h^{n+1}] \le \|u_h^{n+1}-2u_h^n+u_h^{n-1}\|^2+\|\tilde{u}_h^{n+1}-u_h^{n+1}\|^2. \tag{3.15}$$

Thanks to the KKT conditions in (2.11), we have $[u_h^{n+1},\lambda_h^{n+1}]=[u_h^n,\lambda_h^n]=0$ for all n, so for the second term on the righthand side of (3.14), we have

$$\begin{aligned}4[u_h^{n+1}-u_h^n,\tilde{u}_h^{n+1}-u_h^{n+1}] &= -\frac{8\delta t}{3}[u_h^{n+1}-u_h^n,\lambda_h^{n+1}-\lambda_h^n]\\ &= \frac{8\delta t}{3}\{[u_h^{n+1},\lambda_h^n]+[u_h^n,\lambda_h^{n+1}]\} \ge 0.\end{aligned} \tag{3.16}$$

Next, we rewrite (2.11) with $k=2$ as

$$3u_h^{n+1}-2\delta t\lambda_h^{n+1}=3\tilde{u}_h^{n+1}-2\delta t\lambda_h^n. \tag{3.17}$$

Taking the discrete inner product of each side of the above equation with itself, since $[\lambda_h^{n+1},u_h^{n+1}]=0$, we derive

$$3\|u_h^{n+1}\|^2+\frac{4}{3}\delta t^2\|\lambda_h^{n+1}\|^2=3\|\tilde{u}_h^{n+1}\|^2-4\delta t[\tilde{u}_h^{n+1},\lambda_h^n]+\frac{4}{3}\delta t^2\|\lambda_h^n\|^2. \tag{3.18}$$

Now, summing up (3.10) with (3.18), and using (3.11) to (3.16), we obtain that for $n \ge 1$,

$$\begin{aligned}&4(\|u_h^{n+1}\|^2-\|u_h^n\|^2)+\|2u_h^{n+1}-u_h^n\|^2-\|2u_h^n-u_h^{n-1}\|^2+2\|\tilde{u}_h^{n+1}-u_h^{n+1}\|^2\\ &+\frac{4}{3}\delta t^2(\|\lambda_h^{n+1}\|^2-\|\lambda_h^n\|^2)+4\delta t\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle=-\frac{8\delta t}{3}\{[u_h^{n+1},\lambda_h^n]+[u_h^n,\lambda_h^{n+1}]\}.\end{aligned} \tag{3.19}$$

From (3.19) and using (3.16), we derive easily

$$\begin{aligned}&4(\|u_h^{n+1}\|^2-\|u_h^n\|^2)+\|2u_h^{n+1}-u_h^n\|^2-\|2u_h^n-u_h^{n-1}\|^2\\ &+2\|\tilde{u}_h^{n+1}-u_h^{n+1}\|^2+\frac{4}{3}\delta t^2(\|\lambda_h^{n+1}\|^2-\|\lambda_h^n\|^2)+4\delta t\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle \le 0.\end{aligned} \tag{3.20}$$

On the other hand, since the first step is computed by using the first-order scheme, we take $n=1$ in (3.1) to obtain

$$\|u_h^1\|^2+\|\tilde{u}_h^1-u_h^0\|^2+2\delta t\langle\mathcal{L}_h^0\tilde{u}_h^1,\tilde{u}_h^1\rangle+\delta t^2\|\lambda_h^1\|^2=\|u_h^0\|^2. \tag{3.21}$$

Finally, summing up (3.20) from $n=1$ to $n=m-1$ with (3.21) multiplied by 4, we obtain, after dropping some unnecessary terms,

$$4\|u_h^m\|^2+\|2u_h^m-u_h^{m-1}\|^2+\frac{4}{3}\delta t^2\|\lambda_h^m\|^2+4\delta t\sum_{n=0}^{m-1}\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle \le \|2u_h^1-u_h^0\|^2+4\|u_h^0\|^2, \tag{3.22}$$

which implies the (3.9). In particular, if $\langle\mathcal{L}_h^n v_h,v_h\rangle \ge 0\ \forall v_h \in X_h$, the above indicates that $4\|u_h^m\|^2+\|2u_h^m-u_h^{m-1}\|^2 \le \|2u_h^1-u_h^0\|^2+4\|u_h^0\|^2$ for all m, so that the scheme (2.10)–(2.11) with $k=2$ is unconditional stable and dissipative. □

Next, we consider the mass conserved scheme (2.25)–(2.26) with $k=2$.

Theorem 3.4. *For the scheme (2.25)–(2.26) with $k=2$, we assume that the first step is computed with the first-order scheme (2.19)–(2.20). Then, if $\langle\mathcal{L}_h^n v_h,1\rangle=0$ for any $v_h \in X_h$, we have*

$$\begin{aligned}&4\|u_h^m\|^2+\|2u_h^m-u_h^{m-1}\|^2+\frac{4}{3}\delta t^2\|\lambda_h^m+\xi_h^m\|^2+4\delta t\sum_{n=0}^{m-1}\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle\\ &\le \|2u_h^1-u_h^0\|^2+4\|u_h^0\|^2, \quad \forall m \ge 1.\end{aligned} \tag{3.23}$$

In particular, if for all n, $\langle\mathcal{L}_h^n v_h,v_h\rangle \ge 0\ \forall v_h \in X_h$, then the scheme (2.25)–(2.26) with $k=2$ is dissipative and unconditionally stable.

Proof. The proof is again similar to that of Theorem 3.3 so we just point out the differences below.

First, (3.10) should be replaced by

$$[3\tilde{u}_h^{n+1}-4u_h^n+u_h^{n-1},2\tilde{u}_h^{n+1}]+4\delta t\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle=4\delta t[\lambda_h^n+\xi_h^n,\tilde{u}_h^{n+1}]. \tag{3.24}$$

Then (3.16) should be replaced by

$$\begin{aligned}4[u_h^{n+1}-u_h^n,\tilde{u}_h^{n+1}-u_h^{n+1}]&=-\frac{8\delta t}{3}[u_h^{n+1}-u_h^n,\lambda_h^{n+1}+\xi_h^{n+1}-(\lambda_h^n+\xi_h^n)]\\&=-\frac{8\delta t}{3}[u_h^{n+1}-u_h^n,\lambda_h^{n+1}-\lambda_h^n]-\frac{8\delta t}{3}[u_h^{n+1}-u_h^n,\xi_h^{n+1}-\xi_h^n]\\&=-\frac{8\delta t}{3}[u_h^{n+1}-u_h^n,\lambda_h^{n+1}-\lambda_h^n]\\&=\frac{8\delta t}{3}\{[u_h^{n+1},\lambda_h^n]+[u_h^n,\lambda_h^{n+1}]\}\geq 0,\end{aligned} \tag{3.25}$$

where we used the fact that

$$-\frac{8\delta t}{3}[u_h^{n+1}-u_h^n,\xi_h^{n+1}-\xi_h^n]=-\frac{8\delta t}{3}(\xi_h^{n+1}-\xi_h^n)\big([u_h^{n+1},1]-[u_h^n,1]\big)=0.$$

Next, (3.17) should be replaced by

$$3u_h^{n+1}(z)-2\delta t(\lambda_h^{n+1}(z)+\xi_h^{n+1})=3\tilde{u}_h^{n+1}(z)-2\delta t(\lambda_h^n(z)+\xi_h^n). \tag{3.26}$$

Taking the discrete inner product of (3.26) with itself on both sides, we obtain

$$\begin{aligned}&3\|u_h^{n+1}(z)\|^2+\frac{4}{3}\delta t^2\|\lambda_h^{n+1}(z)+\xi_h^{n+1}\|^2-2\delta t[u_h^{n+1},\lambda_h^{n+1}+\xi_h^{n+1}]\\&=3\|\tilde{u}_h^{n+1}(z)\|^2+\frac{4}{3}\delta t^2\|\lambda_h^n(z)+\xi_h^n\|^2-2\delta t[\tilde{u}_h^{n+1},\lambda_h^n+\xi_h^n].\end{aligned} \tag{3.27}$$

Summing up (2.25) and (2.26a), we obtain

$$\frac{\alpha_k u_h^{n+1}-A_k(u_h^n)}{\delta t}+\mathcal{L}_h^n\tilde{u}_h^{n+1}=\lambda_h^{n+1}+\xi_h^{n+1}. \tag{3.28}$$

Taking the discrete inner product of (3.28) with 1 on both sides, using (2.26c), we obtain

$$[\lambda_h^{n+1}+\xi_h^{n+1},1]=[\lambda_h^{n+1},1]+[\xi_h^{n+1},1]=0, \tag{3.29}$$

which implies $\xi_h^{n+1}\leq 0$ since $\lambda_h^{n+1}\geq 0$. Therefore,

$$-2\delta t[u_h^{n+1},\lambda_h^{n+1}+\xi_h^{n+1}]=-2\delta t[u_h^{n+1},\xi_h^{n+1}]\geq 0. \tag{3.30}$$

Then, summing up (3.24) with (3.27), and using (3.25) and (3.30), we arrive at

$$\begin{aligned}&4(\|u_h^{n+1}\|^2-\|u_h^n\|^2)+\|2u_h^{n+1}-u_h^n\|^2-\|2u_h^n-u_h^{n-1}\|^2+2\|\tilde{u}_h^{n+1}-u_h^{n+1}\|^2\\&+\frac{4}{3}\delta t^2(\|\lambda_h^{n+1}+\xi_h^{n+1}\|^2-\|\lambda_h^n+\xi_h^n\|^2)+4\delta t\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle\\&=-\frac{8\delta t}{3}\{[u_h^{n+1},\lambda_h^n]+[u_h^n,\lambda_h^{n+1}]\}+2\delta t[u_h^{n+1},\xi_h^{n+1}].\end{aligned} \tag{3.31}$$

From (3.31) and using (3.25) and (3.30), we obtain

$$\begin{aligned}&4(\|u_h^{n+1}\|^2-\|u_h^n\|^2)+\|2u_h^{n+1}-u_h^n\|^2-\|2u_h^n-u_h^{n-1}\|^2\\&+2\|\tilde{u}_h^{n+1}-u_h^{n+1}\|^2+\frac{4}{3}\delta t^2(\|\lambda_h^{n+1}+\xi_h^{n+1}\|^2-\|\lambda_h^n+\xi_h^n\|^2)+4\delta t\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle\leq 0.\end{aligned} \tag{3.32}$$

For the first step, we take $m=1$ in (3.4) to obtain

$$\|u_h^1\|^2+\|\tilde{u}_h^1-u_h^0\|^2+2\delta t\langle\mathcal{L}_h^0\tilde{u}_h^1,\tilde{u}_h^1\rangle+\delta t^2\|\lambda_h^1+\xi_h^1\|^2\leq\|u_h^0\|^2. \tag{3.33}$$

Finally, summing up (3.32) from $n=1$ to $n=m-1$ with (3.33) multiplied by 4, we obtain,

$$4\|u_h^m\|^2+\|2u_h^m-u_h^{m-1}\|^2+\frac{4}{3}\delta t^2\|\lambda_h^m+\xi_h^m\|^2+4\delta t\sum_{n=0}^{m-1}\langle\mathcal{L}_h^n\tilde{u}_h^{n+1},\tilde{u}_h^{n+1}\rangle\leq\|2u_h^1-u_h^0\|^2+4\|u_h^0\|^2,$$

which implies the (3.23). In particular, if $\langle \mathcal{L}_h^n v_h, v_h\rangle \geq 0\ \forall v_h \in X_h$, the above indicates that $4\|u_h^m\|^2 + \|2u_h^m - u_h^{m-1}\|^2 \leq \|2u_h^1 - u_h^0\|^2 + 4\|u_h^0\|^2$ for all m, so that the scheme (2.25)–(2.26) with $k = 2$ is unconditional stable and dissipative. □

Remark 3.1. The results in the previous theorems are derived for a general approximate operator $\mathcal{L}_h^n$. They imply in particular:

- If $\mathcal{L}_h^n$ is non-negative, e.g., as in (2.3) with application to Porous Media equation, then the first- and second-order positivity preserving schemes with Lagrange multiplier are unconditionally energy stable.
- If one can show, perhaps under certain condition $\Delta t \leq c_0 h^\alpha$ with a semi-implicit discretization (where $\alpha > 0$ depending on the problem and discretization), that for the usual schemes, i.e., by setting $\lambda_h^n \equiv 0$ for all n, we have
$$\delta t \sum_{n=1}^{m-1} \langle \mathcal{L}_h^n \tilde{u}_h^{n+1}, \tilde{u}_h^{n+1}\rangle \;\geq\; \beta\delta t \sum_{n=1}^{m-1} a_h(\tilde{u}_h^{n+1}, \tilde{u}_h^{n+1}) - C_1,\ \forall m \leq T/\Delta t - 1,$$
where β is some positive constant in (0, 1] and T is the final time, then we derive from the above and (3.1) that the solutions of the corresponding schemes (2.10)–(2.11) and (2.25)–(2.26) with $k = 1, 2$ are bounded in the sense that
$$\|u_h^m\|^2 + 2\beta\delta t \sum_{n=1}^{m-1} a_h(\tilde{u}_h^{n+1}, \tilde{u}_h^{n+1}) \leq \|u_h^0\|^2 + C_1,\ \forall m \leq T/\Delta t - 1.$$

Remark 3.2.

We are unable to prove similar results for the schemes (2.10)–(2.11) and (2.25)–(2.26) with $k \geq 3$. The situation is similar to the pressure-correction schemes for the Navier–Stokes equations [28].

4. Numerical experiments

In this section, we carry out various numerical experiments to demonstrate the performance of proposed positivity preserving schemes. We use spectral Galerkin methods with numerical integration for all cases, namely, Fourier-spectral method [29] is used for problems with periodic boundary conditions, while a Legendre-spectral method [29] is used for problems with Dirichlet or Neumann boundary conditions. Note that in general it is much more difficult to preserve positivity with a spectral method than with a lower-order finite element or finite difference method. Below, $h = 1/N$ where N is the number of collocation points in each direction.

4.1. Convergence rate

We first test the convergence rates in time for the positivity preserving schemes using the Allen–Cahn equation [30]

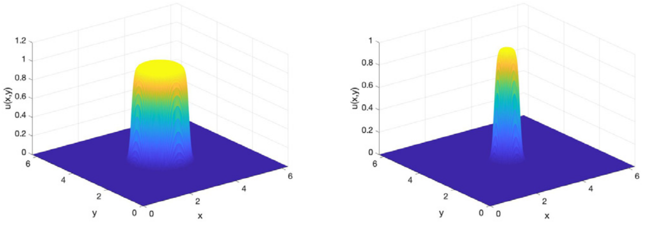
$$u_t - \Delta u + \frac{1}{\epsilon^2} u(u-1)(u-\frac{1}{2}) = 0, \tag{4.1}$$

with periodic boundary condition in $\Omega = [0, 2\pi)^2$. It is well-known that the solution will remain in [0, 1] if the values of the initial condition $u(x, y, 0)$ are in [0, 1] [31]. In particular, it is positivity preserving.

We choose the following initial condition

$$u(x, y, 0) = \frac{1}{2}\big(1 + \tanh(\frac{1 - \sqrt{(x-\pi)^2 + (y-\pi)^2}}{\sqrt{2}\epsilon})\big), \tag{4.2}$$

with $\epsilon^2 = 0.001$ and use 32^2 uniform collocation points in $[0, 2\pi)^2$, i.e., $\Sigma_h = \{x_{jk} = (\frac{j}{2\pi}, \frac{k}{2\pi}) : j, k = 0, 1, \ldots, 31\}$. We note that with this coarse mesh, the usual semi-implicit Fourier-collocation method will produce numerical solutions with negative values, i.e., the spatial discretized problem (2.2) will lead to non zero λ_h. The spatial discretized problem is smooth in time so it can be used to test the convergence rates of the positivity preserving time discretization schemes. On the other hand, the Fourier-spectral method with 32 × 32 uniform collocation points is



(a) u_h at $t = 0.01$

(b) u_h at $t = 0.4$

Fig. 1. Numerical solution of Allen-Cahn equation (4.1) with $\epsilon^2 = 0.001$ at $t = 0.01$ and $t = 0.4$ computed with 32 × 32 Fourier modes but plotted on the 256 × 256 grid.

Table 1
Accuracy test: The L^∞ errors between u_h^n and the reference solution at $t = 0.01$ for the Allen-Cahn equation (4.1) with $\epsilon^2 = 0.001$ using (2.10)–(2.11).

δt	(2.10)–(2.11) $k = 1$	Order	(2.10)–(2.11) $k = 2$	Order
4×10^{-5}	2.71×10^{-4}	–	1.20×10^{-5}	–
2×10^{-5}	1.37×10^{-4}	0.98	2.97×10^{-6}	2.01
1×10^{-5}	6.85×10^{-5}	1.00	7.31×10^{-7}	2.02
5×10^{-6}	3.42×10^{-5}	1.00	1.74×10^{-7}	2.07
2.5×10^{-6}	1.71×10^{-5}	1.00	3.54×10^{-8}	2.30

enough to provide a reasonable approximation to this problem as shown in Fig. 1. As a reference solution, we use the numerical solution computed by the scheme (2.17)–(2.18) with $k = 2$ and $\delta t = 10^{-6}$.

In Table 1, we list the L^∞ errors of numerical solution between the reference solution u_h^{n+1} obtained using the schemes (2.10)–(2.11) with $k = 1, 2$. We observe from Table 1 that the schemes (2.10)–(2.11) are indeed kth order accurate.

4.2. Porous medium equation

In this subsection, we consider the porous medium equation (PME) [17]:

$$\rho_t = \Delta \rho^m = m\nabla \cdot (\rho^{m-1}\nabla \rho), \tag{4.3}$$

with homogeneous Dirichlet boundary condition in $\Omega = (-5, 5)^d$ ($d = 1, 2, 3$) where $m \geq 1$ is a physical parameter. The porous medium equation has wide applications in various areas, including fluid dynamics, heater transfer and image processing. We observe from (4.3) that the PME is degenerate and its solution has to be positive.

We shall use the Legendre–Galerkin method with numerical integration in space. Let P_N be the set of polynomials with degree less than or equal to N in each direction, and let Σ_h be the set of the interior Legendre–Gauss–Lobatto points, i.e., in the one dimensional case, $\Sigma_h = \{x_k : k = 1, 2, \dots, N-1\}$ where $\{x_k\}$ are the roots of $L_N'(x)$ with L_N being the Legendre polynomial of Nth degree, and in the multi-dimensional case, Σ_h is obtained by the tensor product of one-dimensional set. We set $X_h = \{v_h \in P_N : v_h|_{\partial\Omega = 0}\}$, and use the scheme (2.10)–(2.11) with $k = 2$ and $\mathcal{L}_h^n(v_h) = -\nabla \cdot (m(\rho^{n+1,*})^{m-1}\nabla v_h)$. For the reader's convenience, it is explicitly described below:

Find $\rho_h^{n+1} \in X_h$ such that

$$[\frac{3\tilde{\rho}_h^{n+1} - 4\rho_h^n + \rho_h^{n-1}}{2\delta t}, v_h] + m[(\rho_h^{n+1,*})^{m-1}\nabla \rho_h^{n+1}, \nabla v_h] = [\lambda_h^n, v_h], \quad \forall v_h \in X_h, \tag{4.4}$$

$$\frac{3\rho_h^{n+1}(z) - 3\tilde{\rho}_h^{n+1}(z)}{2\delta t} = \lambda_h^{n+1}(z) - \lambda_h^n(z), \quad \forall z \in \Sigma_h, \tag{4.5}$$

$$\lambda_h^{n+1}(z) \geq 0, \ \ \rho_h^{n+1}(z) \geq 0, \ \ \lambda_h^{n+1}(z)\rho_h^{n+1}(z) = 0, \tag{4.6}$$

where

$$\rho_h^{n+1,*} = \begin{cases} 2\rho_h^n - \rho_h^{n-1} & if\ \rho_h^n \geq \rho_h^{n-1}, \\ \frac{1}{2/\rho_h^n - 1/\rho_h^{n-1}} & if\ \rho_h^n < \rho_h^{n-1}. \end{cases} \tag{4.7}$$

At each time step, we need to solve an elliptic equation with variable coefficients in (4.4), which can be efficiently solved by a preconditioned conjugate gradient iteration with a constant coefficient problem as the preconditioner.

4.2.1. *Comparison with a usual semi-implicit scheme*

We now compare the positivity preserving scheme (4.4)–(4.6) with the corresponding usual semi-implicit scheme

$$[\frac{3\rho_h^{n+1} - 4\rho_h^n + \rho_h^{n-1}}{2\delta t}, v_h] + m[(\rho_h^{n+1,*})^{m-1}\nabla\rho_h^{n+1}, \nabla v_h] = 0, \quad \forall v_h \in X_h, \tag{4.8}$$

using the exact solution of the porous medium Eq. (4.3) in the Barenblatt form

$$\rho(x,t) = \frac{1}{t_0^\alpha}\Big(C - \alpha\frac{m-1}{2m}\frac{x^2}{t_0^{2\alpha}}\Big)_+^{\frac{1}{m-1}}, \tag{4.9}$$

where $f_+ = \max\{f, 0\}$, $\alpha = \frac{1}{m+1}$, $C = 1$ and $t_0 = t + 1$. The solution is compactly supported in $(0, 1)$ with the interface moving outward in a finite speed. The initial condition for the numerical simulations is chosen as $\rho(x, 0)$.

In Fig. 2, we plot the L^2 errors by the usual semi-implicit scheme (4.8) and by the positivity preserving (4.4)–(4.6) with $m = 2$ and set $\delta t = 10^{-3}$. We observe that the errors grow rapidly after a short time with $N = 128, 256, 512$ using (4.8) since the numerical solution becomes negative at some places; while the error appears to be under control for at least up to $T = 1$ with $N = 1024$. On the other hand, by using (4.4)–(4.6), the errors remain under control and accurate solutions are obtained for all N. We observe from Fig. 2(c) that, even at $N = 1024$, the Lagrange multiplier λ_h becomes non-zero in order to maintain positivity. Numerical solutions ρ_h at $T = 0.05, 0.1, 1$ are plotted in Fig. 2(d).

In Fig. 3, we consider a more challenging case with $m = 5$ using $\delta t = 10^{-3}$ and $N = 1024$, and plot the numerical solution at $T = 0.1$ using the usual semi-implicit scheme (4.8) and the positivity preserving (4.4)–(4.6) in Fig. 3(a) and (b). We observe that the scheme (4.8) produces negative values near the interface while the scheme (4.4)–(4.6) leads to accurate positive solutions. We also plot the Lagrange multiplier λ_h in Fig. 3(c) which indicates that λ_h becomes larger near the interface to maintain the positivity of ρ_h. In Fig. 3(d), we plot the L^2 errors using (4.4)–(4.6) with λ_h and without λ_h.

Next we consider the 2D case with the exact solution in the Barenblatt form

$$\rho(x,y,t)|_{t=0} = \frac{1}{t_0^\alpha}\Big(C - \alpha\frac{m-1}{2m}\frac{x^2+y^2}{t_0^{2\alpha}}\Big)_+^{\frac{1}{m-1}}, \tag{4.10}$$

where $C = 1$, $t_0 = t + 1$ and $\alpha = \frac{1}{m+1}$. We set $N = 200$, $\delta t = 2 \times 10^{-4}$ and consider $m = 2, 5$. We observe from Fig. 4 that correct solutions are obtained by the positivity preserving scheme and that the values of Lagrange multiplier λ_h are quite large near the interface in order to maintain the positivity of ρ_h.

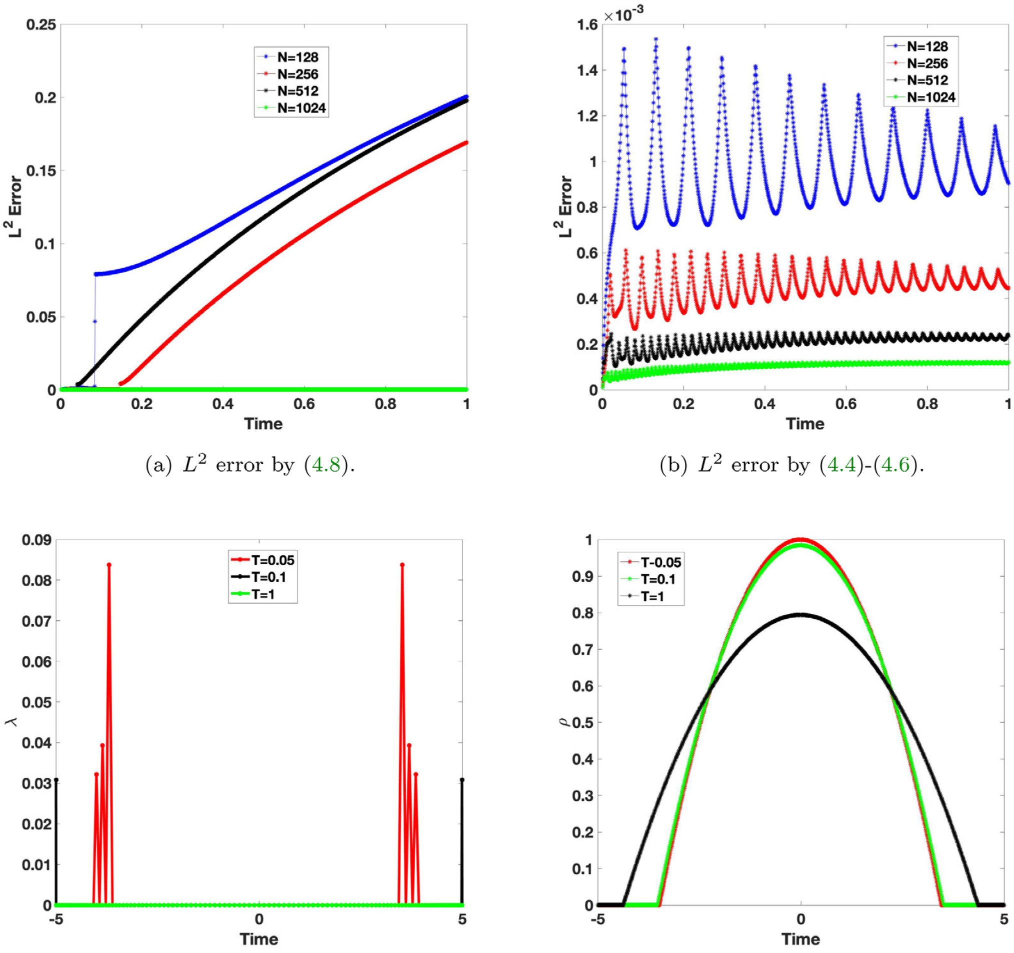
4.2.2. *Effect of mass conservation*

The porous media Eq. (4.3) with homogeneous Dirichlet boundary conditions is mass conserving. So we compare the second-order positivity conserving schemes without mass conservation and with mass conservation for the porous medium equation. The results with $\delta t = 10^{-4}$ and $N = 128$ are plotted in Fig. 5. We observe that the scheme with mass conservation preserves the mass and is slightly more accurate in terms of L^2 error than the scheme without mass conservation whose mass is monotonically increasing. Actually, only a few iterations are needed at each time step to solve ξ using secant method. We can also observe Lagrange multiplier $\xi \leq 0$ in time interval $[0, 2]$.

4.3. *Poisson-Nernst–Planck Equations*

We consider the following Poisson–Nernst–Planck (PNP) system [10,32] which describes the dynamics of ion transport in ion channels:

$$-\epsilon^2\Delta\phi = p - n, \text{in } \Omega_T := (0, T] \times \Omega. \tag{4.11}$$



(a) L^2 error by (4.8).

(b) L^2 error by (4.4)-(4.6).

(c) λ_h by (4.4)-(4.6) with $N = 1024$.

(d) ρ_h by (4.4)-(4.6) with $N = 1024$.

Fig. 2. The L^2 error of numerical solution by (4.8) and by (4.4)–(4.6) with $\delta t = 10^{-3}$ and $m = 2$.

$$p_t = \nabla \cdot (\nabla p + p\nabla\phi), \tag{4.12}$$
$$n_t = \nabla \cdot (\nabla n - n\nabla\phi), \tag{4.13}$$

with initial conditions

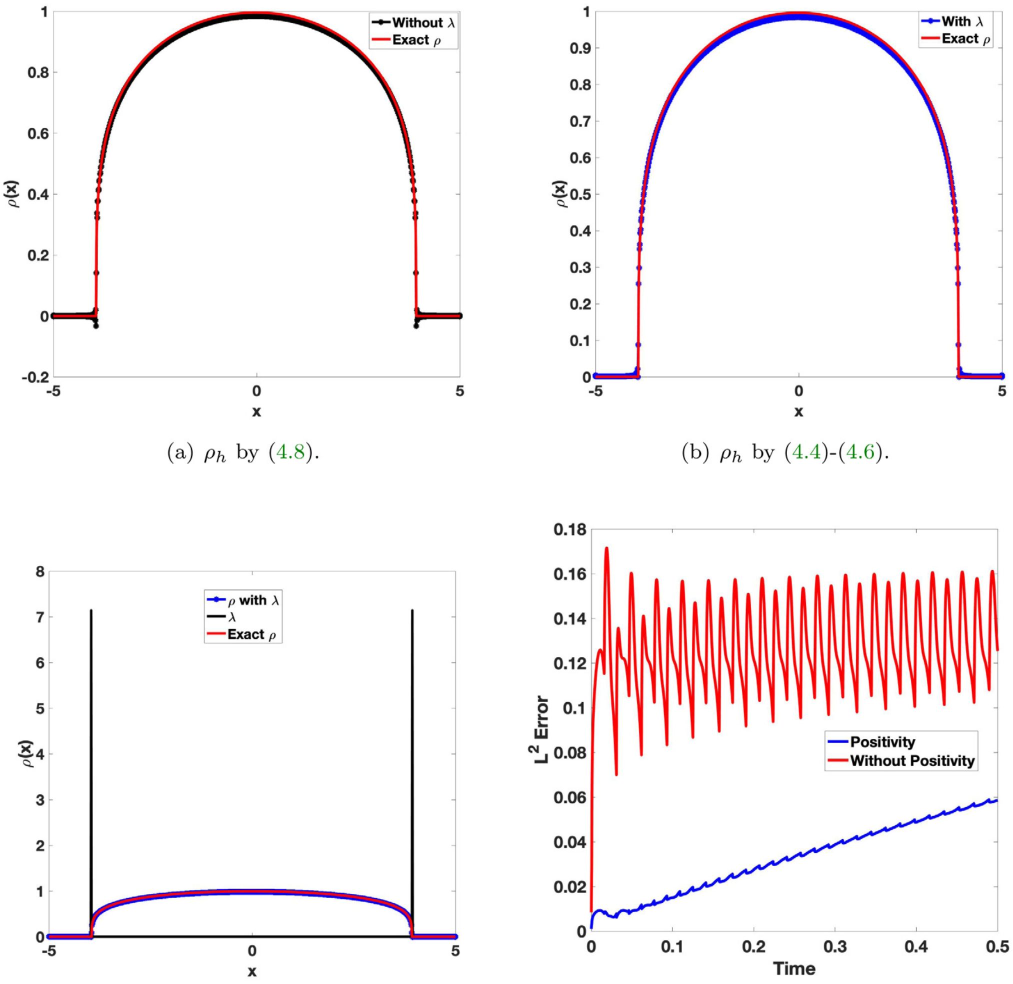
$$p(0, x) = p_0(x) \geq 0, \quad n(0, x) = n_0(x) \geq 0, \quad \text{in } \Omega, \tag{4.14}$$

and homogeneous Neumann boundary conditions

$$\frac{\partial p}{\partial \boldsymbol{n}} = \frac{\partial n}{\partial \boldsymbol{n}} = \frac{\partial \phi}{\partial \boldsymbol{n}} = 0, \quad \text{on} \quad \partial\Omega_T := (0, T] \times \partial\Omega. \tag{4.15}$$

In the above, p and n are concentration of positive and negative ions with valence $+1$ and -1, respectively, ϕ is the electrical potential, ϵ is a small positive dimensionless number representing the ratio of the Debye length to the physical characteristic length. The unknown functions p and n have to be positive for the problem to be well posed. Below we use the general approach presented in the last section to construct a positivity preserving scheme for the PNP equations. Since we need to keep both p and n positive, two Lagrange multipliers λ and η are needed . Lagrange multipliers ξ and γ are used to preserve mass.

We set $X_h = P_N \times P_N$, and $\Sigma_h = \{(x_i, x_j),\ 1 \leq i, j \leq N-1\}$, where $\{x_k\}_{k=0}^N$ are the roots of $(1-x^2)L'_N(x)$ with L_N being the Legendre polynomial of Nth degree. And we use the Legendre–Galerkin method with numerical



(a) ρ_h by (4.8).

(b) ρ_h by (4.4)-(4.6).

(c) Lagrange multiplier λ_h.

(d) L^2 errors.

Fig. 3. (a) and (b) Numerical solutions ρ_h at $T = 0.1$ with $m = 5$, $\delta t = 10^{-3}$, $N = 1024$ by (4.8) and by (4.4)–(4.6). (c) Lagrange multiplier λ_h and ρ_h by (4.4)–(4.6). (d) L^2 error with λ_h and without λ_h.

integration in space [29]. Then a second-order positivity preserving scheme based on the scheme (2.25)–(2.26c) with $k = 2$ is as follows: for $\forall q_h, m_h \in X_h$

$$[\frac{3\tilde{p}_h^{n+1} - 4p_h^n + p_h^{n-1}}{2\delta t}, q_h] = [\nabla p_h^{n+1} + p_h^{n+1,\star}\nabla\phi_h^{n+1,\star}, \nabla q_h] + [\lambda_h^n + \xi_h^n, q_h], \tag{4.16}$$

$$\frac{3p_h^{n+1} - \tilde{p}_h^{n+1}}{2\delta t} = \lambda_h^{n+1} - \lambda_h^n + \xi_h^{n+1} - \xi_h^n, \tag{4.17}$$

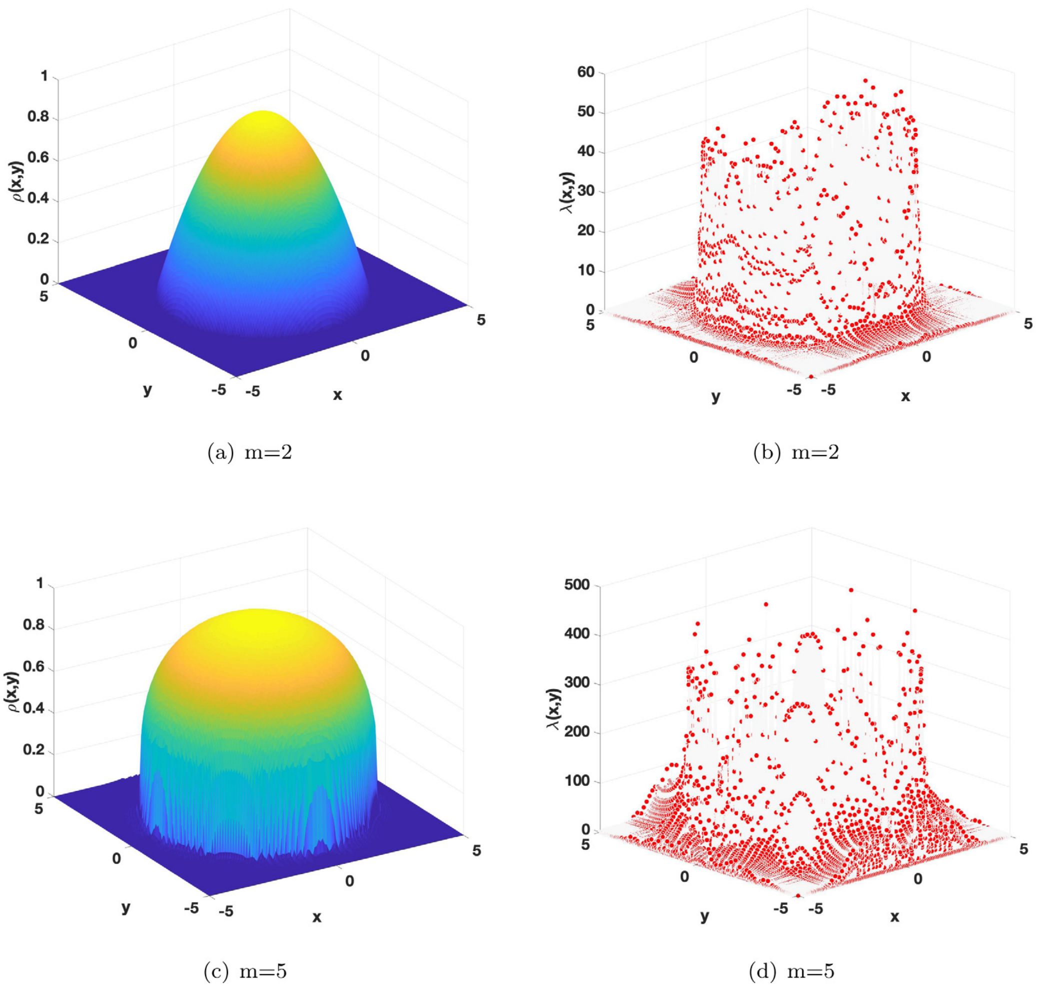
$$\lambda_h^{n+1} \ge 0,\ \ p_h^{n+1} \ge,\ \ \lambda_h^{n+1} p_h^{n+1} = 0,\ \ [p_h^{n+1}, 1] = [p_h^n, 1]; \tag{4.18}$$

$$[\frac{3\tilde{n}_h^{n+1} - 4n_h^n + n_h^{n-1}}{2\delta t}, m_h] = [\nabla n_h^{n+1} - n_h^{n+1,\star}\nabla\phi_h^{n+1,\star}, \nabla m_h] + [\eta_h^n + \gamma_h^n, m_h], \tag{4.19}$$

$$\frac{3n_h^{n+1} - 3\tilde{n}_h^{n+1}}{2\delta t} = \eta_h^{n+1} - \eta_h^n + \gamma_h^{n+1} - \gamma_h^n, \tag{4.20}$$

$$\eta_h^{n+1} \ge 0,\ \ n_h^{n+1} \ge 0,\ \ \eta_h^{n+1} n_h^{n+1} = 0,\ \ [n_h^{n+1}, 1] = [n_h^n, 1]; \tag{4.21}$$

$$\epsilon^2[\nabla\phi_h^{n+1}, \nabla\psi_h] = [p_h^{n+1} - n_h^{n+1}, \psi_h], \quad \forall\psi_h \in X_h; \tag{4.22}$$



(a) m=2

(b) m=2

(c) m=5

(d) m=5

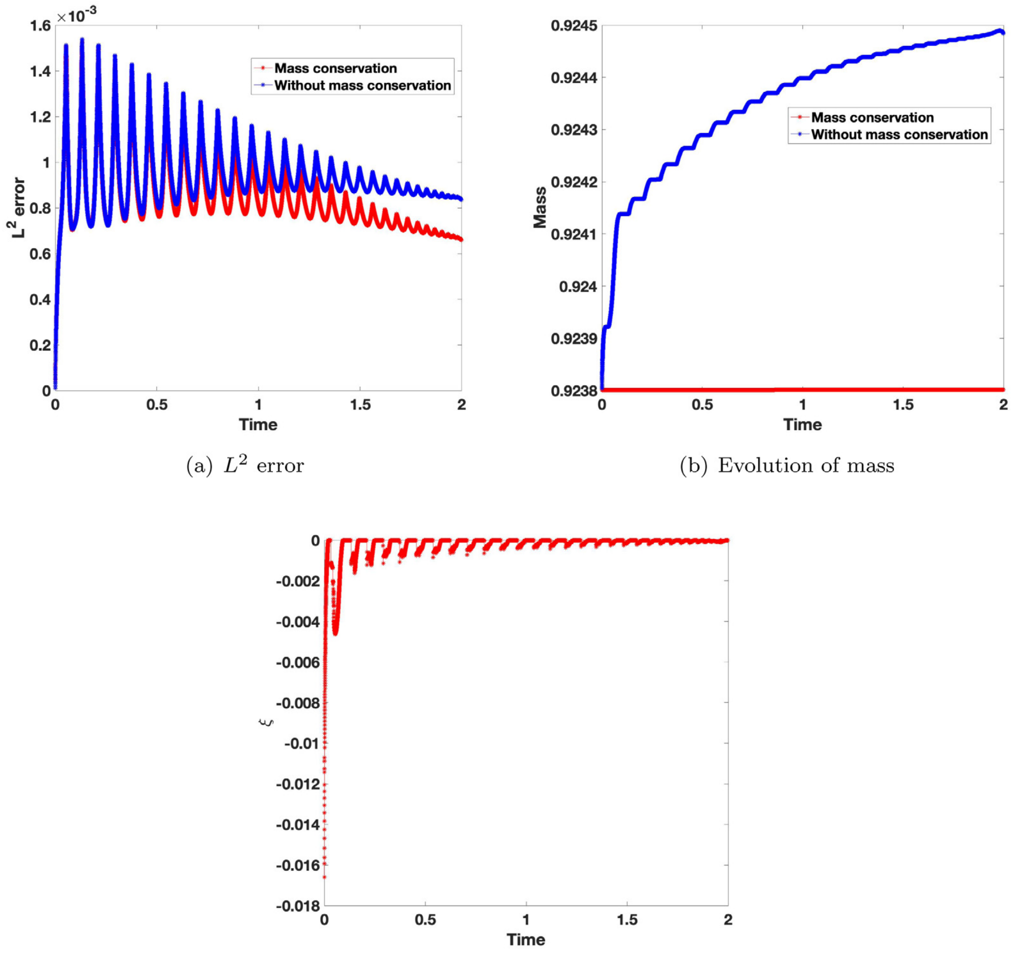
Fig. 4. Numerical solution of 2D porous medium equation at $T = 0.2$ with $\delta t = 2 \times 10^{-4}$ and $N = 200$: (a) ρ_h with $m = 2$. (b) Lagrange multiplier λ_h with $m = 2$. (c) ρ_h with $m = 5$. (d) Lagrange multiplier λ_h with $m = 5$.

where $p_h^{n+1,\star} = 2p_h^n - p_h^{n-1}$ and $n_h^{n+1,\star} = 2n_h^n - n_h^{n-1}$. In the above, p_h^{n+1} and n_h^{n+1} are decoupled and can be determined from (4.16)–(4.18) and (4.19)–(4.21) respectively. Once p_h^{n+1} and n_h^{n+1} are known, ϕ_h^{n+1} can be obtained from (4.22). Hence, the scheme is very efficient.

We set $\Omega = (-1, 1)^2$, $\epsilon = 0.1$, and use $\delta t = 10^{-3}$, $N = 256$ in the above scheme with the initial conditions:

$$\begin{aligned} &p(x, y, 0), n(x, y, 0) = \begin{cases} 1, & x^2 + y^2 \le 0.25, \\ 0, & \text{otherwise}. \end{cases} \\ &\phi(x, y, 0) = \begin{cases} (x-0.5)^2(y-0.5)^2, & x^2 + y^2 \le 0.25, \\ 0, & \text{otherwise}. \end{cases} \end{aligned} \tag{4.23}$$

The numerical solution at different times are plotted in Fig. 6. We observe that p and n are always non-negative. We also plot the Lagrange multipliers λ and η in Fig. 7 at time $t = 3 \times 10^{-3}$. Since the solutions of the PNP system are smooth, the Lagrange multipliers are zero at most places, and are non-zero only at some localized boundary with quite small values.



(a) L^2 error

(b) Evolution of mass

(c) ξ

Fig. 5. (a): L^2 error by second-order positivity schemes with mass conservation and without mass conservation. (b): Evolution of mass with respective to time. (c): Lagrange multiplier ξ for mass conservation.

4.4. Lubrication-type equation

As the last example, we consider the following lubrication equation [23]

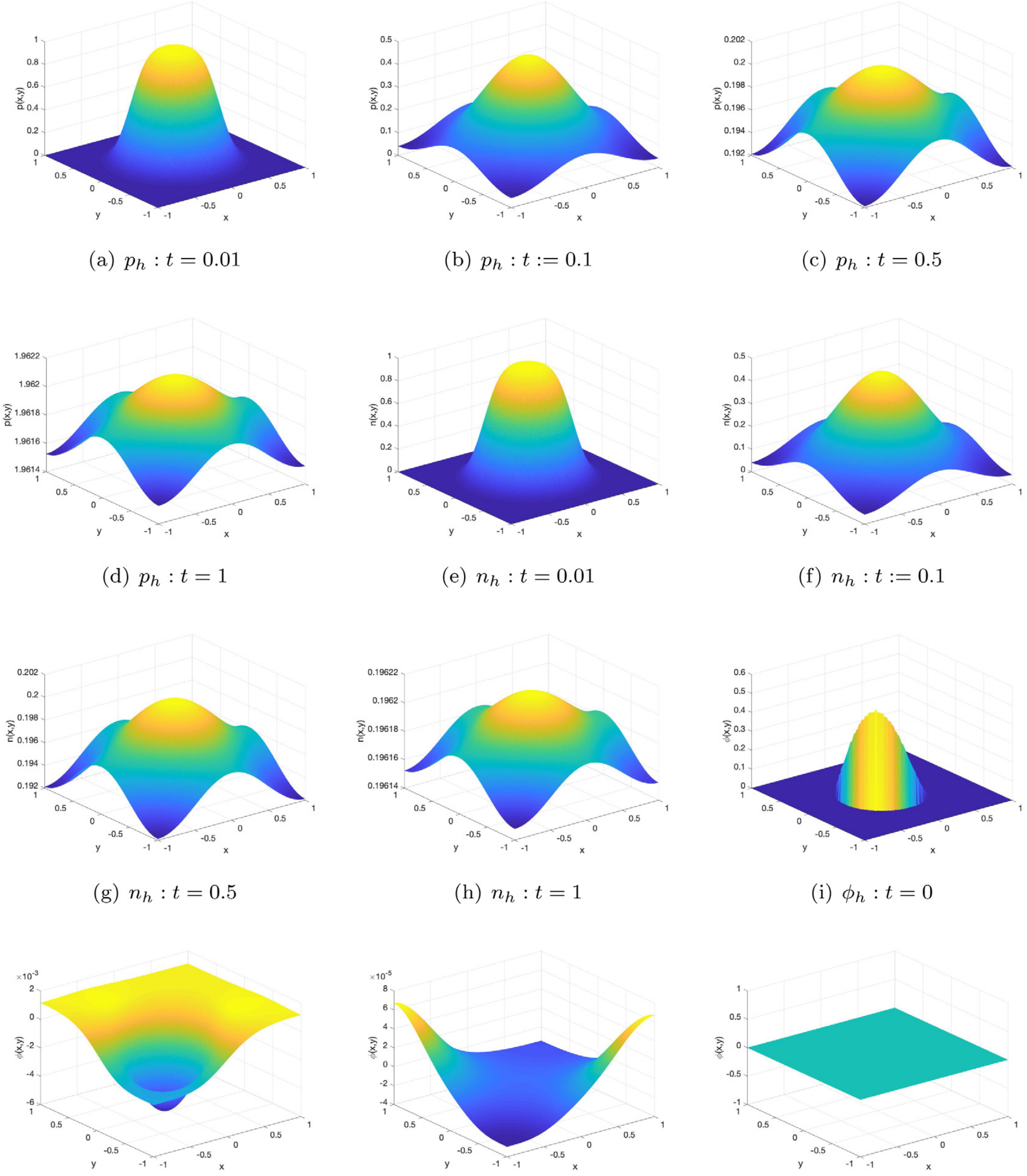
$$\rho_t + \nabla \cdot (f(\rho)\nabla \varDelta \rho) = 0, \tag{4.24}$$

where $f(\rho) \approx \rho^m$ as $\rho \to 0$ with m depending on the boundary condition at the liquid solid interface: $m = 3$ for no-slip boundary condition while $0 < m < 3$ for various other boundary condition. The above equation has been used, e.g., in the study of thin liquid films and fluid interfaces by surface tension [33].

Below, we consider (4.24) in $\varOmega = (-1, 1)$ and $\varOmega = (-\pi, \pi)^2$ with periodic boundary conditions, and use a Fourier collocation method in space. Since Eq. (4.24) may develop a singularity in finite time, it is a common practice to regularize it [23,33]. In [33], the equation is regularized by replacing $f(\rho)$ with $f_\eta(\rho) = \frac{\rho^4 f(\rho)}{\eta f(\rho)+\rho^4}$ and it is shown in [33] that the regularized problem is well posed for all time. On the other hand, one can also regularize the equation by requiring the solution to be bounded away from zero, namely, $\rho(z) \geq \epsilon$ for a prescribed ϵ.

Hence, a second-order scheme based on (2.25)–(2.26c) with $k = 2$ is as follows:

$$\frac{3\tilde{\rho}_h^{n+1}(z) - 4\rho_h^n(z) + \rho_h^{n-1}(z)}{2\delta t} + \mathcal{L}_h^n \tilde{\rho}_h^{n+1}(z) = \lambda_h^n(z) + \xi_h^n, \quad \forall z \in \varSigma_h, \tag{4.25}$$



(a) $p_h : t = 0.01$ (b) $p_h : t := 0.1$ (c) $p_h : t = 0.5$

(d) $p_h : t = 1$ (e) $n_h : t = 0.01$ (f) $n_h : t := 0.1$

(g) $n_h : t = 0.5$ (h) $n_h : t = 1$ (i) $\phi_h : t = 0$

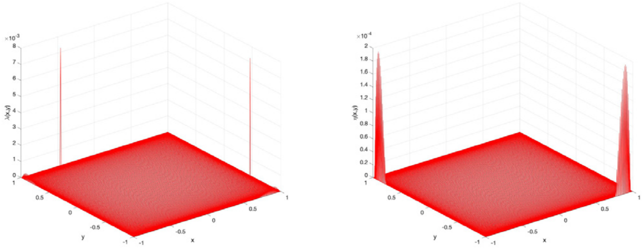
(j) $\phi_h : t := 0.01$ (k) $\phi_h : t = 0.1$ (l) $\phi_h : t = 1$

Fig. 6. The numerical solution of PNP Eqs. (4.11)–(4.13) with time step 1×10^{-3} and $N = 256$ in 2D by using scheme (4.16)–(4.22). Parameter $\epsilon = 0.1$.

and

$$\begin{aligned}
&\frac{3\rho_h^{n+1}(z) - 3\tilde{\rho}_h^{n+1}(z)}{2\delta t} = \lambda_h^{n+1}(z) - \lambda_h^n(z) + \xi_h^{n+1} - \xi_h^n, \quad \forall z \in \Sigma_h, \\
&\rho_h^{n+1}(z) \geq \epsilon, \ \lambda_h^{n+1}(z) \geq 0, \ \lambda_h^{n+1}(\rho^{n+1}(z) - \epsilon) = 0, \quad \forall z \in \Sigma_h, \\
&[\rho_h^{n+1}, 1] = [\rho_h^n, 1],
\end{aligned} \tag{4.26}$$

where $\mathcal{L}_h^n$ is defined as follows depending on the type of regularization:



(a) λ_h (b) η_h

Fig. 7. The Lagrange multipliers of PNP Eqs. (4.11)–(4.13) λ and η at time $t = 3 \times 10^{-3}$ for numerical simulations at Fig. 6.

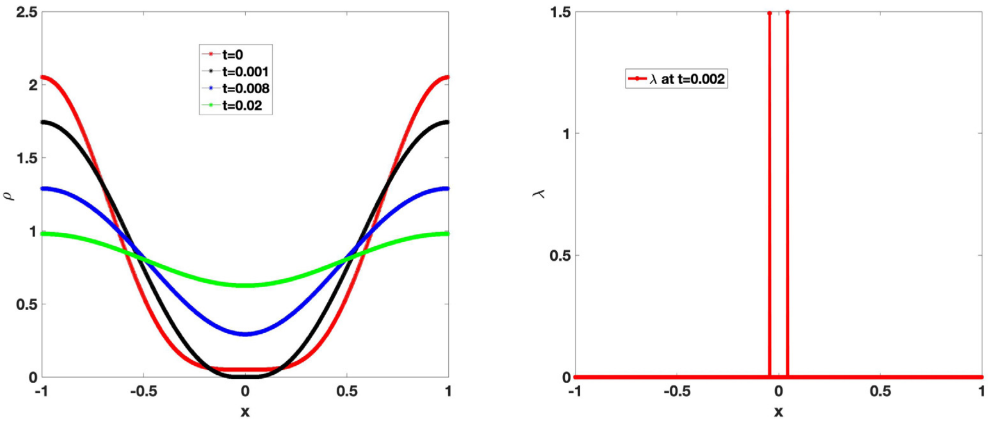


Fig. 8. Numerical solutions ρ and Lagrange multiplier λ of positivity preserving scheme computed with $\eta = 10^{-12}$, $\epsilon = 0$ and $\delta t = 2 \times 10^{-8}$.

- $\epsilon = 0$ and $\mathcal{L}_h^n \tilde{\rho}_h^{n+1} = \nabla \cdot (f_\eta(\rho_h^{n+1,*}) \nabla \Delta \tilde{\rho}_h^{n+1})$ with $\rho_h^{n+1,*}$ defined in (4.7).
- $\epsilon > 0$ and $\mathcal{L}_h^n \tilde{\rho}_h^{n+1} = \nabla \cdot (f(\rho_h^{n+1,*}) \nabla \Delta \tilde{\rho}_h^{n+1})$ with $\rho_h^{n+1,*}$ defined in (4.7).

The second step (4.26) can be implemented as

$$(\rho_h^{n+1}, \lambda_h^{n+1}) = \begin{cases} (\tilde{\rho}_h^{n+1} - \frac{2}{3}\delta t(\lambda_h^n - \xi_h^{n+1} + \xi_h^n), 0) & \text{if} \quad \epsilon \le \tilde{\rho}_h^{n+1} - \frac{2}{3}\delta t(\lambda_h^n - \xi_h^{n+1} + \xi_h^n), \\ (\epsilon, \lambda_h^n + \xi_h^n - \xi_h^{n+1} + \frac{3}{2\delta t}(\epsilon - \tilde{\rho}_h^{n+1})) & \text{otherwise.} \end{cases} \tag{4.27}$$

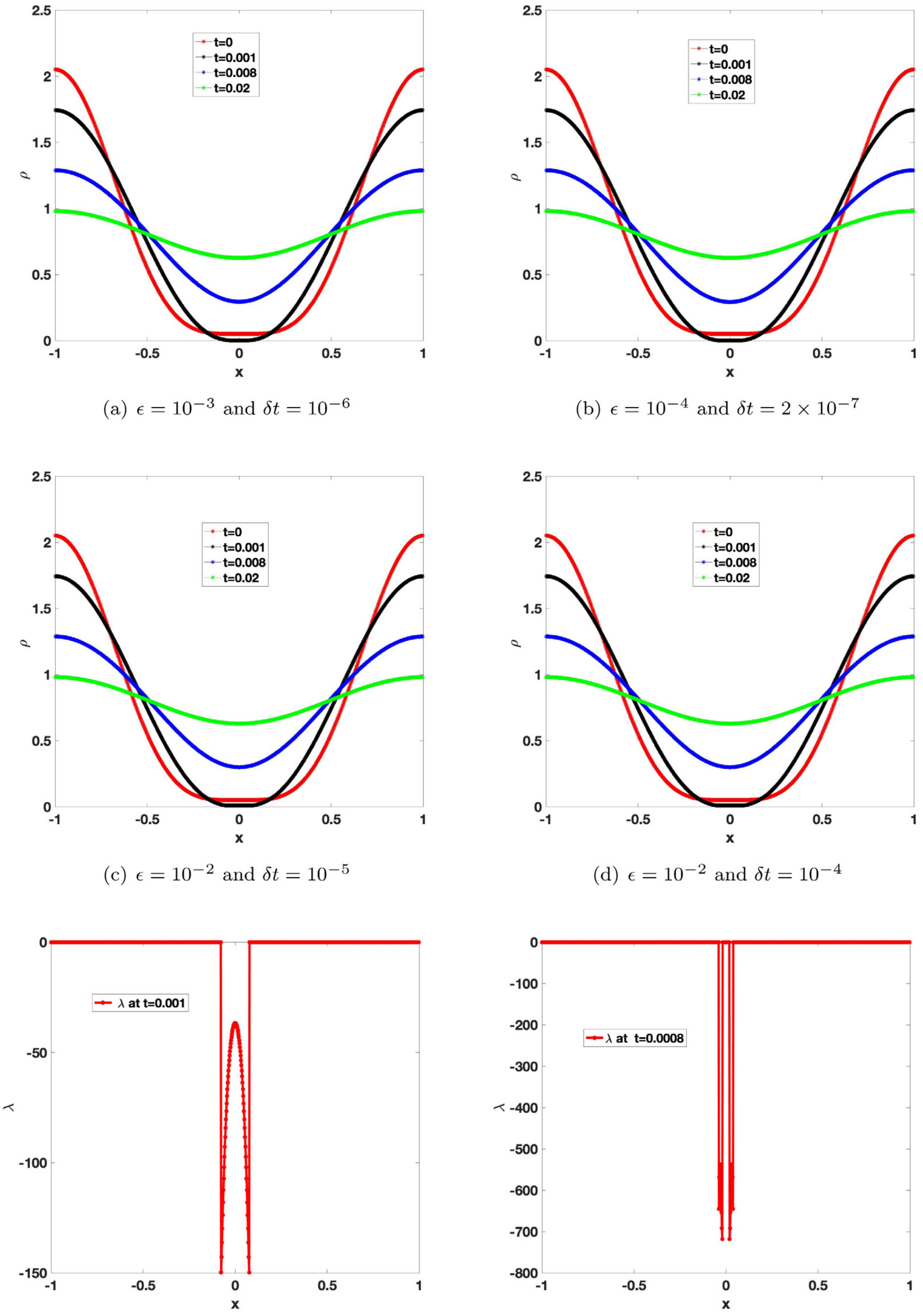
We consider first the one-dimensional case with $f(\rho) = \rho^{\frac{1}{2}}$ in $(-1, 1)$ with periodic boundary conditions and the initial condition

$$\rho_0(x) = 0.8 - \cos(\pi x) + 0.25 \cos(2\pi x). \tag{4.28}$$

This example has been well studied in [23], and the original equation $f(\rho) = \rho^{\frac{1}{2}}$ will develop a singularity at $t \approx 0.00074$. However, with a regularization, the solution can be continued beyond the singularity.

In Fig. 8, numerical solutions ρ and Lagrange multiplier λ are shown at different times computed by regularized positivity preserving scheme with $N = 1000$, $\eta = 10^{-12}$, $\epsilon = 0$ and time step $\delta t = 2 \times 10^{-8}$. Numerical solutions in Fig. 8 are indistinguishable from results computed with (4.25).

In Fig. 9(a–d), we plot the numerical solutions at different times computed with (4.25) using 1000 Fourier modes with various ϵ and δt. We observe that the numerical solutions are indistinguishable with ϵ ranging from 10^{-2} to 10^{-4}. However, as we decrease ϵ, smaller time steps has to be used. The Lagrange multiplier λ_h at time $t = 0.001$ and $t = 0.0008$ are plotted in Fig. 9(e–f). We observe that λ_h becomes large near the places where the solution approaches zero.



(a) $\epsilon = 10^{-3}$ and $\delta t = 10^{-6}$

(b) $\epsilon = 10^{-4}$ and $\delta t = 2 \times 10^{-7}$

(c) $\epsilon = 10^{-2}$ and $\delta t = 10^{-5}$

(d) $\epsilon = 10^{-2}$ and $\delta t = 10^{-4}$

(e) $\epsilon = 10^{-2}$ and $\delta t = 10^{-4}$

(f) $\epsilon = 10^{-4}$ and $\delta t = 2 \times 10^{-7}$

Fig. 9. (a)–(d): numerical solution ρ_h of positivity preserving scheme (4.25) for lubrication-type Eq. (4.24) in 1D at various time with different time steps and ϵ. (e)–(f): Lagrange multiplier λ at time $t = 0.001$ and $t = 0.0008$ using different time steps and ϵ.

Next, we consider a 2D example with the initial condition

$$\rho(x,y)=\begin{cases}(x-0.5)^2(y-0.5)^2, & x^2+y^2\le 0.25,\\ 0, & \text{otherwise},\end{cases} \tag{4.29}$$

in the domain $[-\pi,\pi)^2$.

We first take $f(\rho)=\rho$ and use the following usual semi-implicit scheme:

$$\frac{3\rho_h^{n+1}(z)-4\rho_h^n(z)+\rho_h^{n-1}(z)}{2\delta t}+\mathcal{L}_h^n\rho_h^{n+1}(z)=0,\quad \forall z\in\Sigma_h. \tag{4.30}$$

The scheme failed to converge with $\delta t=10^{-5}$. However, by using the 2D version of the scheme (4.25) with $\epsilon=0$ and 128×128 Fourier modes, correct results can be obtained with $\delta t=10^{-5}$. In Fig. 10(a–d), we plot the initial condition and numerical solutions at $t=0.001, 0.01, 0.1$, while we plot in Fig. 10(e–f) the Lagrange multipliers λ_h at $t=0.001, 0.1$. We observe that the Lagrange multiplier takes nonzero values at a significant part of the domain which explains why the usual semi-implicit scheme failed to converge.

4.5. Cahn–Hilliard equation in 3D

For three-dimensional case, we consider the Cahn–Hilliard equation [34] with a logarithmic potential in the domain $[0,1]^3$:

$$\begin{aligned}&u_t=\nabla\cdot(M(u)\nabla(\mu-\frac{1}{\alpha}\Delta u)),\\ &\mu=\log(\frac{u}{1-u})+2\theta(1-2u),\end{aligned} \tag{4.31}$$

where $\theta=\frac{T_c}{T}$ is a dimensionless number which represents the ratio between the critical temperature T_c and the absolute temperature T. The mobility function is $M(u)=\alpha u(1-u)$. With a given initial condition $u_0>0$, due to the singular logarithmic potential, the solution of Cahn–Hilliard equation (4.31) is expected to be $u(t)>0$. We choose initial condition to be

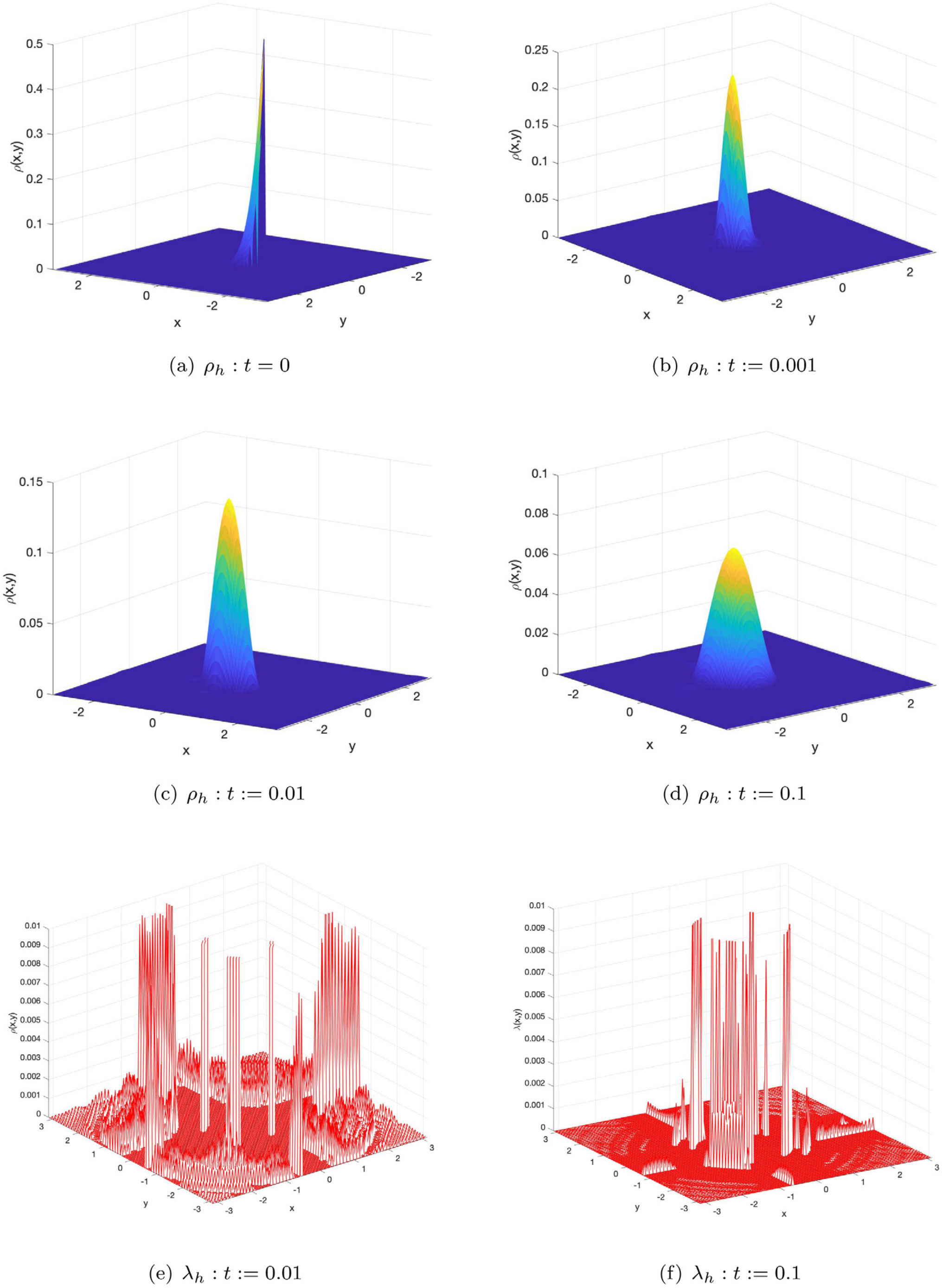
$$u(t=0)=\widehat{u}_0+0.05\,\mathrm{rand}(x,y), \tag{4.32}$$

where $\mathrm{rand}(x,y)$ represents random values with uniform distribution in $[-1,1]$.

We develop the following positivity-preserving scheme

$$\begin{aligned}&\frac{3\tilde{u}_h^{n+1}-4u_h^n+u_h^{n-1}}{2\delta t}+\Delta^2\tilde{u}_h^{n+1}-\Delta\tilde{u}_h^{n+1}=\Delta^2u_h^{n+1,\star}-\Delta u_h^{n+1,\star}\\ &+\nabla\cdot(M_h^{n+1,\star}\nabla(\mu_h^{n+1,\star}-\frac{1}{\alpha}\Delta\tilde{u}_h^{n+1,\star}))+\lambda_h^n,\\ &\mu_h^{n+1,\star}=\log(\frac{u_h^{n+1,\star}}{1-u_h^{n+1,\star}})+2\theta(1-2u_h^{n+1,\star}),\\ &\frac{3u_h^{n+1}(z)-3\tilde{u}_h^{n+1}(z)}{2\delta t}=\lambda_h^{n+1}(z)-\lambda_h^n(z),\\ &\lambda_h^{n+1}(z)\ge 0,\ u_h^{n+1}(z)\ge\epsilon,\ \lambda_h^{n+1}(z)(u_h^{n+1}(z)-\epsilon)=0,\end{aligned} \tag{4.33}$$

where $u^{n+1,\star}=2u^n-u^{n-1}$ and $M_h^{n+1,\star}=\alpha u_h^{n+1,\star}(1-u_h^{n+1,\star})$. We develop the linear no-iteration scheme (4.33) by adding two stabilized terms [21]. For computations, we implement Fourier spectral method in space $[0,1]^3$ with resolution 128^3. We set parameters to be $\widehat{u}_0=0.63$, $\theta=\frac{3}{2}$, $\epsilon=10^{-6}$ and $\alpha=200$ which are same with [34]. InFig. 11, we depict iso-surface of $u_h=0.635$ and $u_h=0.63$ at time $t=2.5\times 10^{-4}, 1.183\times 10^{-3}, 1.1514\times 10^{-3}$, 2.368×10^{-3}, 5.240×10^{-3}, 4×10^{-2} using positive-preserving scheme (4.33). We obtain similar results with Fig. 27 in [34]. We also show the evolution of minimum value of u_h in Fig. 12 which indicates that u_h will be positive from initial to steady state. Computations are implemented from initial time $t=0$ to steady state $t=0.04$ using time step $\delta t=10^{-6}$. CPU time is 16.774 h for 40 000 time steps on the Mac: Intel(R) Core(TM) i5-8279U CPU @ 2.40 GHz.



(a) $\rho_h : t = 0$

(b) $\rho_h : t := 0.001$

(c) $\rho_h : t := 0.01$

(d) $\rho_h : t := 0.1$

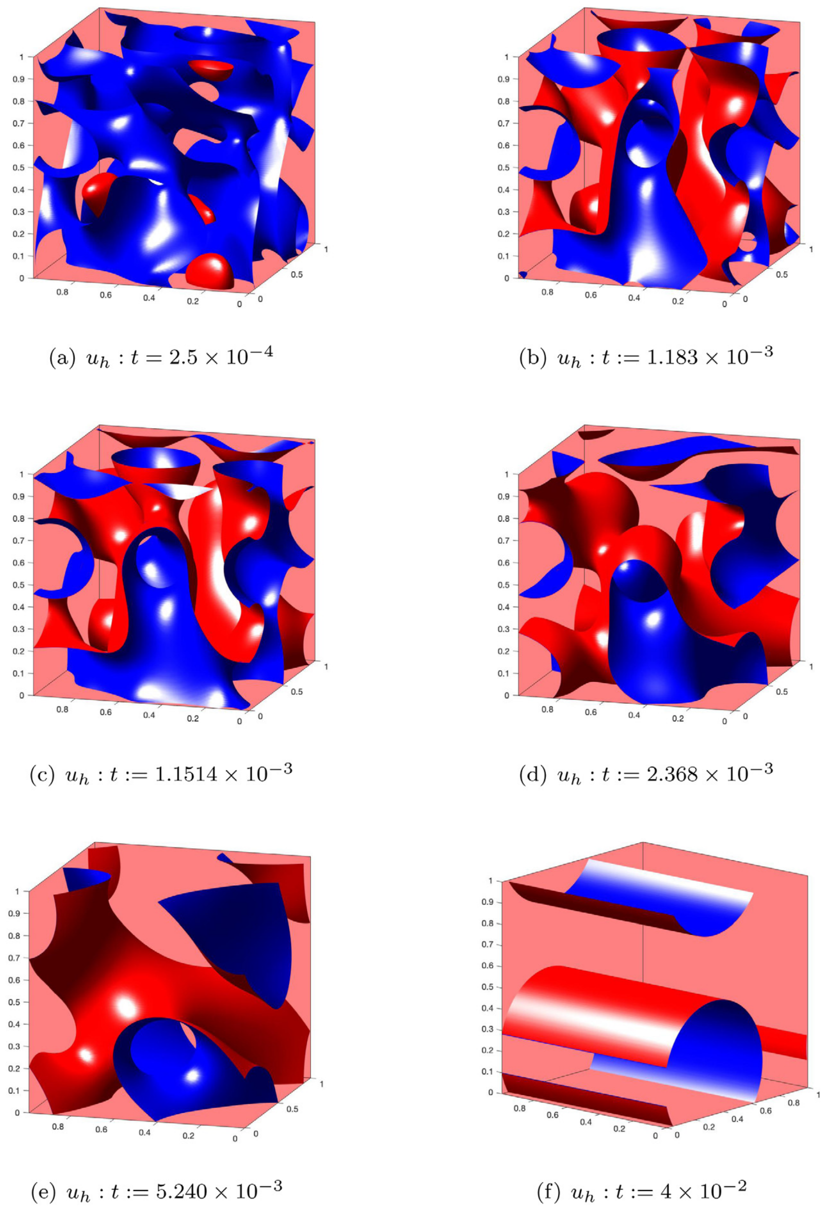
(e) $\lambda_h : t := 0.01$

(f) $\lambda_h : t := 0.1$

Fig. 10. Numerical solutions ρ_h with positivity preserving scheme at $t = 0, 0.001, 0.01, 0.1$, and Lagrange multiplier λ_h at $t = 0.01, 0.1$.

5. Concluding remarks

If a PDE requires its solution to be positive, a generic numerical scheme for the PDE usually cannot preserve the positivity. We presented in this paper a new approach to construct positivity preserving schemes for parabolic type equations by a simple modification to generic numerical schemes. More precisely, we introduce a space–time



(a) $u_h : t = 2.5 \times 10^{-4}$

(b) $u_h : t := 1.183 \times 10^{-3}$

(c) $u_h : t := 1.1514 \times 10^{-3}$

(d) $u_h : t := 2.368 \times 10^{-3}$

(e) $u_h : t := 5.240 \times 10^{-3}$

(f) $u_h : t := 4 \times 10^{-2}$

Fig. 11. The iso-surfaces of numerical solutions $u_h = 0.63$ in blue and $u_h = 0.635$ in red for Cahn–Hilliard equation with logarithmic potential computed by positivity preserving scheme with time step $\delta t = 10^{-6}$ at $t = 2.5 \times 10^{-4}$, 1.183×10^{-3}, 1.1514×10^{-3}, 2.368×10^{-3}, 5.240×10^{-3}, 4×10^{-2}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Lagrange multiplier function to enforce the positivity, and expand the underlying PDE using the KKT conditions. The key question is how to solve the expanded system efficiently with essentially the same cost as the generic numerical scheme.

We constructed a new class of positivity preserving schemes by using the predictor–corrector approach to the expanded system: the prediction step can be a generic semi-implicit or implicit scheme, while the correction step is used to enforce the positivity and can be implemented as a simple pointwise update with negligible cost. This new approach is not restricted to any particular spatial discretization and can be combined with various time discretization schemes. It can be applied to a large class of parabolic PDEs which require solutions to be positive. It is also non

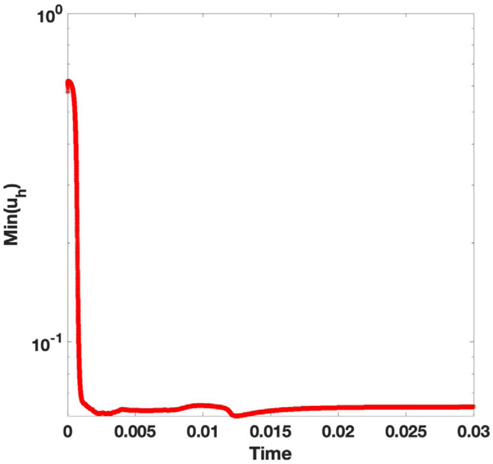


Fig. 12. The minimum value of u_h for Cahn–Hilliard equation in 3D using positive-preserving scheme (4.33).

intrusive as you can easily modify your non-positivity preserving schemes for them to become positivity preserving. In addition, we also presented a modification to the above approach so that the schemes can also preserve mass if the underlying PDE is mass conserving.

An interesting and useful observation is that the ad-hoc cut-off approach can be interpreted as a special case of our predictor–corrector approach. Hence, it provides a different justification for the cut-off approach, moreover allows us to modify the cut-off approach so that it becomes mass conserving, and opens new avenue for further exploration.

We established stability results for the first- and second-order schemes based on the new approach under a general setting, and presented ample numerical experiments to validate the new approach. Our numerical results indicate that the new approach is very effective for the variety of problems that we tested.

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] Changna Lu, Weizhang Huang, Erik S. Van Vleck, The cutoff method for the numerical computation of nonnegative solutions of parabolic PDEs with application to anisotropic diffusion and lubrication-type equations, J. Comput. Phys. 242 (2013) 24–36.

[2] Buyang Li, Jiang Yang, Zhi Zhou, Arbitrarily high-order exponential cut-off methods for preserving maximum principle of parabolic equations, SIAM J. Sci. Comput. 42 (6) (2020) A3957–A3978.

[3] Qiang Du, Lili Ju, Xiao Li, Zhonghua Qiao, Maximum bound principles for a class of semilinear parabolic equations and exponential time differencing schemes, 2020, arXiv preprint arXiv:2005.11465.

[4] Hao Li, Shusen Xie, Xiangxiong Zhang, A high order accurate bound-preserving compact finite difference scheme for scalar convection diffusion equations, SIAM J. Numer. Anal. 56 (6) (2018) 3308–3345.

[5] Hao Li, Xiangxiong Zhang, On the monotonicity and discrete maximum principle of the finite difference implementation of C^0-Q^2 finite element method, Numer. Math. 145 (2) (2020) 437–472.

[6] Xiangxiong Zhang, Chi-Wang Shu, Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments, Proc. R. Soc. A 467 (2134) (2011) 2752–2776.

[7] Xiangxiong Zhang, Chi-Wang Shu, On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes, J. Comput. Phys. 229 (23) (2010) 8918–8934.

[8] Wenbin Chen, Cheng Wang, Xiaoming Wang, Steven M. Wise, Positivity-preserving, energy stable numerical schemes for the Cahn-Hilliard equation with logarithmic potential, J. Comput. Phys.: X 3 (2019) 100031.

[9] Jian-Guo Liu, Li Wang, Zhennan Zhou, Positivity-preserving and asymptotic preserving method for 2D Keller-Segal equations, Math. Comp. 87 (311) (2018) 1165–1189.

[10] Jingwei Hu, Xiaodong Huang, A fully discrete positivity-preserving and energy-dissipative finite difference scheme for Poisson–Nernst–Planck equations, Numer. Math. (2020) 1–39.

[11] Fukeng Huang, Jie Shen, Bound/positivity preserving and energy stable SAV schemes for dissipative systems: applications to keller-segel and Poisson-Nernst-Planck equations, SIAM J. Sci. Comput. 43 (3) (2021) A1832–A1857.

[12] Jaap J.W. van der Vegt, Yinhua Xia, Yan Xu, Positivity preserving limiters for time-implicit higher order accurate discontinuous Galerkin discretizations, SIAM J. Sci. Comput. 41 (3) (2019) A2037–A2063.

[13] Kazufumi Ito, Karl Kunisch, On a semi-smooth Newton method and its globalization, Math. Program. 118 (2) (2009) 347–370.

[14] Kazufumi Ito, Karl Kunisch, Lagrange Multiplier Approach to Variational Problems and Applications, SIAM, 2008.

[15] Patrick T. Harker, Jong-Shi Pang, Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications, Math. Program. 48 (1) (1990) 161–220.

[16] Francisco Facchinei, Jong-Shi Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer Science & Business Media, 2007.

[17] Juan Luis Vázquez, The Porous Medium Equation: Mathematical Theory, Oxford University Press on Demand, 2007.

[18] Maïtine Bergounioux, Kazufumi Ito, Karl Kunisch, Primal-dual strategy for constrained optimal control problems, SIAM J. Control Optim. 37 (4) (1999) 1176–1194.

[19] Charles Francis Curtiss, Joseph O. Hirschfelder, Integration of stiff equations, Proc. Natl. Acad. Sci. USA 38 (3) (1952) 235.

[20] Qing Cheng, Jie Shen, Global constraints preserving scalar auxiliary variable schemes for gradient flows, SIAM J. Sci. Comput. 42 (4) (2020) A2489–A2513.

[21] Qing Cheng, Jie Shen, Multiple scalar auxiliary variable (MSAV) approach and its application to the phase-field vesicle membrane model, SIAM J. Sci. Comput. 40 (6) (2018) A3982–A4006.

[22] Yuanyuan Liu, Chi-Wang Shu, Mengping Zhang, High order finite difference WENO schemes for nonlinear degenerate parabolic equations, SIAM J. Sci. Comput. 33 (2) (2011) 939–965.

[23] Liya Zhornitskaya, Andrea L. Bertozzi, Positivity-preserving numerical schemes for lubrication-type equations, SIAM J. Numer. Anal. 37 (2) (1999) 523–555.

[24] C.M. Elliott, A.M. Stuart, The global dynamics of discrete semilinear parabolic equations, SIAM J. Numer. Anal. 30 (6) (1993) 1622–1663.

[25] Jie Shen, Jie Xu, Jiang Yang, A new class of efficient and robust energy stable schemes for gradient flows, SIAM Rev. 61 (3) (2019) 474–506.

[26] Qing Cheng, Chun Liu, Jie Shen, Generalized SAV approaches for gradient systems, J. Comput. Appl. Math. 394 (2021) 113532.

[27] John Charles Butcher, Numerical Methods for Ordinary Differential Equations, John Wiley & Sons, 2016.

[28] Jean-Luc Guermond, Peter Minev, Jie Shen, An overview of projection methods for incompressible flows, Comput. Methods Appl. Mech. Engrg. 195 (44–47) (2006) 6011–6045.

[29] Jie Shen, Tao Tang, Li-Lian Wang, Spectral Methods: Algorithms, Analysis and Applications, Vol. 41, Springer Science & Business Media, 2011.

[30] Samuel M. Allen, John W. Cahn, A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening, Acta Metall. 27 (6) (1979) 1085–1095.

[31] Roger Temam, Infinite-Dimensional Dynamical Systems in Mechanics and Physics, Vol. 68, Springer Science & Business Media, 2012.

[32] Dongdong He, Kejia Pan, Xiaoqiang Yue, A positivity preserving and free energy dissipative difference scheme for the Poisson–Nernst–Planck system, J. Sci. Comput. 81 (1) (2019) 436–458.

[33] Andrea L. Bertozzi, Mary Pugh, The lubrication approximation for thin viscous films: Regularity and long-time behavior of weak solutions, Comm. Pure Appl. Math. 49 (2) (1996) 85–123.

[34] Héctor Gómez, Victor M. Calo, Yuri Bazilevs, Thomas J.R. Hughes, Isogeometric analysis of the Cahn–Hilliard phase-field model, Comput. Methods Appl. Mech. Engrg. 197 (49–50) (2008) 4333–4352.