SIAM J. NUMER. ANAL.
Vol. 60, No. 3, pp. 970–998

A NEW LAGRANGE MULTIPLIER APPROACH FOR CONSTRUCTING STRUCTURE PRESERVING SCHEMES, II. BOUND PRESERVING*

QING CHENG† AND JIE SHEN†

Abstract. In the second part of this series, we use the Lagrange multiplier approach proposed in the first part [*Comput. Methods Appl. Mech. Engr.*, 391 (2022), 114585] to construct efficient and accurate bound and/or mass preserving schemes for a class of semilinear and quasi-linear parabolic equations. We establish stability results under a general setting and carry out an error analysis for a second-order bound preserving scheme with a hybrid spectral discretization in space. We apply our approach to several typical PDEs which preserve bound and/or mass and also present ample numerical results to validate our approach.

Key words. bound preserving, mass conservation, KKT conditions, Lagrange multiplier, stability, error analysis

AMS subject classifications. 65M70, 65M12, 65N22

DOI. 10.1137/21M144877X

1. Introduction.

Solutions of partial differential equations (PDEs) arising from sciences and engineering applications are often required to be positive or to remain in a bounded interval. It is beneficial, and often necessary, that their numerical approximations preserve the positivity or bound at the discrete level. In recent years, a large effort has been devoted to construct bound preserving schemes for various problems.

In the first part of this series [7], we constructed a class of positivity preserving schemes using a new Lagrange multiplier approach. A main objective of this paper is to extend the approach in [7] to construct bound preserving schemes for a class of nonlinear PDEs in the following form:

$$u_t + \mathcal{L}u + \mathcal{N}(u) = 0 \tag{1.1}$$

with suitable initial and boundary conditions, where $\mathcal{L}$ is a linear or nonlinear nonnegative operator and $\mathcal{N}(u)$ is a semilinear or quasi-linear operator. We assume that the solution of (1.1) is bound preserving, i.e., $a \le u(\boldsymbol{x}, 0) \le b$ for all $\boldsymbol{x} \in \Omega$; then $a \le u(\boldsymbol{x}, t) \le b$ for all $(\boldsymbol{x}, t) \in \Omega \times (0, T)$.

There exists an extensive literature devoted to constructing positivity/bound preserving schemes for (1.1). We refer to the first part of this series [7] (and the references therein) for a summary of existing approaches for constructing positivity/bound preserving schemes. In particular, large efforts have been devoted to construct spatial discretization for (1.1) such that the resulting numerical scheme satisfies a discrete maximum principle (cf., for instance, [14, 8, 9, 13, 24, 32, 23, 22] and the review paper in [15] for a up-to-date summary in this regard). Another popular strategy is to use a

*Received by the editors September 27, 2021; accepted for publication (in revised form) January 4, 2022; published electronically May 5, 2022.
https://doi.org/10.1137/21M144877X
Funding: The work of the first and second authors is partially supported by NSF (National Science Foundation), USA Grant DMS-2012585 and AFOSR (Air Force Office of Scientific Research), USA Grant FA9550-20-1-0309.
†Department of Mathematics, Purdue University, West Lafayette, IN 47907 USA (cheng573@purdue.edu, shen7@purdue.edu).

convex splitting approach to construct positivity/bound preserving schemes (cf., for instance, [6, 12, 11]).

Consider a generic spatial discretization of (1.1):

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

where u_h is in a certain finite dimensional approximation space X_h and $\mathcal{L}_h$ is a certain approximation of $\mathcal{L}$. In general, the solution u_h, if it exists, may not be bound preserving. Oftentimes, (1.2) may not be well posed if the values of u_h go outside of $[a, b]$. For example, a direct finite elements or spectral approximation to the Allen–Cahn or Cahn–Hilliard equation with logarithmic potential may not be well posed. Instead of using special spatial discretizations which satisfy a discrete maximum principle, we aim to develop a bound preserving approach which can be used for a large class of spatial discretizations. To preserve positivity, it suffices to introduce a Lagrange multiplier λ_h. But to preserve bound, we need to introduce an additional quadratic function $g(u) = (b - u)(u - a)$ and consider the following expanded system with a Lagrange multiplier λ_h:

$$\begin{aligned} &\partial_t u_h + \mathcal{L}_h u_h + \mathcal{N}_h(u_h) = \lambda_h g'(u_h), \\ &\lambda_h \geq 0, \quad g(u_h) \geq 0, \quad \lambda_h g(u_h) = 0. \end{aligned} \tag{1.3}$$

The second equation in (1.3) represents the well-known KKT conditions [20, 17, 19, 2] for constrained minimization. The problem (1.3) can be viewed as an approximation to (1.1); it can also be viewed as a discrete problem without a background PDE, e.g., coming from a discrete constrained minimization problem.

Existing approaches for (1.3) usually start with an implicit time discretization scheme so that the nonlinear system at each time step can still be interpreted as a constrained minimization, then apply a suitable iterative procedure (cf. [30]). As in [7], we shall use a different approach which decouples the computation of the Lagrange multiplier λ_h from that of u_h, leading to a much more efficient algorithm.

We recall that for positivity preserving, we simply use $g(u_h) = u_h$ in the above formulation. However, for bound preserving, the nonlinear nature of $g(u_h)$ makes it much harder to prove stability in norms involving derivatives, and mass conservation whenever is necessary. On the other hand, since the numerical solutions remain to be bounded by construction, this allows us to derive more precise stability results, which in turn enable us to obtain optimal error estimates for both semilinear and quasi-linear PDEs. More precisely, the bound preserving schemes that we construct based on the operator splitting approach enjoy all advantages of the positivity preserving schemes in [7], and furthermore, thanks to the bound preserving property, they allow us to prove a more precise stability result (see Theorem 3.2) and to establish rigorous error estimates for a class of semilinear and quasi-linear dissipative equations (see Theorem 4.1).

We would like to point out that the schemes constructed in this paper include the usual cutoff approach [25] as a special case. Therefore, our presentation provides an alternative interpretation of the cutoff approach and allows us to construct new cutoff implicit-explicit (IMEX) schemes with mass conservation.

To validate our schemes, we apply our new schemes to a variety of problems with bound preserving solutions, including the Allen–Cahn [1] and Cahn–Hilliard [3] equations and a class of Fokker–Planck equations [26].

The remainder of the paper is organized as follows. In section 2, we construct bound preserving schemes for general nonlinear systems (1.1) using the Lagrange

multiplier approach. For problems which also conserve mass, we modify our bound preserving schemes so that they also conserve mass. In section 3, we restrict ourselves to second-order parabolic-type equations and establish a stability result for, as an example, the second-order scheme with mass conservation. In section 4, we consider a hybrid spectral method as an example to carry out an error analysis for a fully discretized second-order scheme. In section 5, we describe applications of our schemes to several typical PDEs with bound and/or mass preserving properties. In section 6, we present some numerical simulations to validate the accuracy and robustness of our schemes. And we conclude with some remarks in the final section.

2. Bound preserving schemes. We construct in this section efficient bound preserving schemes for solving (1.3). The key is to adopt an operator splitting approach in which a standard scheme, which is not bound preserving, is used in the first step, while in the second step, the solution is made bound preserving with a simple yet consistent procedure.

We shall first describe a generic spatial discretization with nodal Lagrangian basis functions, followed by time discretization without and with mass conservation.

Let 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 nonessential) boundary condition is prescribed.

We assume that (1.3) is satisfied pointwise as follows:

$$\begin{aligned} &\partial_t u_h(\boldsymbol{z},t) + \mathcal{L}_h u_h + \mathcal{N}_h(u_h) = \lambda_h(\boldsymbol{z},t) g'(u_h) \quad \forall \boldsymbol{z} \in \Sigma_h, \\ &\lambda_h(\boldsymbol{z},t) \geq 0,\ g(u_h(\boldsymbol{z},t)) \geq 0,\ \lambda_h(\boldsymbol{z},t) g(u_h(\boldsymbol{z},t)) = 0 \quad \forall \boldsymbol{z} \in \Sigma_h, \end{aligned} \tag{2.1}$$

with the Dirichlet boundary condition to be satisfied pointwise if the original problem includes a Dirichlet boundary condition at part or all of the boundary. The above scheme includes finite difference schemes, collocation schemes, and Galerkin-type spatial discretization with a Lagrangian basis.

Denote δt the time step and $t^n = n\delta t$ for $n = 0, 1, 2, \ldots, \frac{T}{\delta t}$, where T is the final computational time. Our schemes consist of two steps: in the first step, we use a generic time discretization, which can be implicit, explicit, or IMEX, to find an intermediate solution $\tilde{u}_h^{n+1}$ which is usually not bound preserving; then we introduce a Lagrange multiplier $\lambda_h^{n+1}(\boldsymbol{z})$ to determine a bound preserving u_h^{n+1}, which is a correction to $\tilde{u}_h^{n+1}$. We shall first construct bound preserving schemes which do not necessarily preserve mass; then we introduce a simple modification which allows us to construct bound preserving schemes which can also preserve mass.

For the sake of clarity, we shall restrict ourselves to constructed schemes based on the IMEX-type time discretization since they are most commonly used for parabolic-type systems. It is straightforward to extend the approach below to schemes based on other types of time discretization.

2.1. A class of multistep IMEX schemes. We construct below kth-order bound preserving schemes for (2.1) based on the backward difference formula for the time derivative and Adams–Bashforth extrapolation by using a predictor-corrector approach.

In order to describe the scheme, we define a sequence $\{\alpha_k\}$ with a slight abuse of notation. For any function v, we use $A_k(v^n)$ and $B_{k-1}(v^n)$ to denote two operators depending on $(v^n, \ldots, v^{n-k+1})$ as follows:

$k = 1$:

$$\alpha_1 = 1, \quad A_1(v^n) = v^n, \quad B_0(v^n) = 0; \tag{2.2}$$

$k = 2$:

$$\alpha_2 = \frac{3}{2}, \quad A_2(v^n) = 2v^n - \frac{1}{2}v^{n-1}, \quad B_1(v^n) = v^n; \tag{2.3}$$

$k = 3$:

$$\begin{gathered}\alpha_3 = \frac{11}{6}, \quad A_3(v^n) = 3v^n - \frac{3}{2}v^{n-1} + \frac{1}{3}v^{n-2}, \\ B_2(v^n) = 2v^n - v^{n-1}.\end{gathered} \tag{2.4}$$

The formula for $k = 4, 5, 6$ can be derived similarly with Taylor expansions.

We assume that $u_h^j, j = 0, 1, \ldots, k-1$ are properly initialized. Then we take the following steps.

Step 1 (predictor). Solve $\tilde{u}_h^{n+1}$ from

$$\begin{aligned}&\frac{\alpha_k \tilde{u}_h^{n+1}(\boldsymbol{z}) - A_k(u_h^n(\boldsymbol{z}))}{\delta t} + \mathcal{L}_h \tilde{u}_h^{n+1}(\boldsymbol{z}) + \mathcal{N}_h(B_k(u_h^n(\boldsymbol{z}))) = B_{k-1}(\lambda_h^n g'(u_h^n(\boldsymbol{z}))) \\ &\forall \boldsymbol{z} \in \Sigma_h.\end{aligned} \tag{2.5}$$

Step 2 (corrector). Solve u_h^{n+1} and λ_h^{n+1} from

$$\frac{\alpha_k(u_h^{n+1}(\boldsymbol{z}) - \tilde{u}_h^{n+1}(\boldsymbol{z}))}{\delta t} = \lambda_h^{n+1}(\boldsymbol{z}) g'(u_h^{n+1}(\boldsymbol{z})) - B_{k-1}(\lambda_h^n(\boldsymbol{z}) g'(u_h^n(\boldsymbol{z}))), \tag{2.6a}$$

$$g(u_h^{n+1}(\boldsymbol{z})) \geq 0, \ \lambda_h^{n+1}(\boldsymbol{z}) \geq 0, \ \lambda_h^{n+1}(\boldsymbol{z}) g(u_h^{n+1}(\boldsymbol{z})) = 0 \quad \forall \boldsymbol{z} \in \Sigma_h. \tag{2.6b}$$

The second step can be solved pointwise as follows. We denote

$$\eta_h^{n+1} := -\frac{\delta t}{\alpha_k} B_{k-1}(\lambda_h^n g'(u_h^n)) \tag{2.7}$$

and rewrite (2.6a) as

$$\frac{\alpha_k(u_h^{n+1}(\boldsymbol{z}) - (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z})))}{\delta t} = \lambda_h^{n+1}(\boldsymbol{z}) g'(u_h^{n+1}(\boldsymbol{z})).$$

We find from the above and (2.6b) that

$$(u_h^{n+1}(\boldsymbol{z}), \lambda_h^{n+1}(\boldsymbol{z})) = \begin{cases} (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}), 0) & \text{if} \quad a < \tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}) < b, \\ \left(a, \frac{a - (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}))}{\frac{\delta t}{\alpha_k} g'(a)}\right) & \text{if} \quad \tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}) \leq a, \\ \left(b, \frac{b - (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}))}{\frac{\delta t}{\alpha_k} g'(b)}\right) & \text{if} \quad \tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}) \geq b, \end{cases} \tag{2.8}$$

$\forall \boldsymbol{z} \in \Sigma_h$.

Remark 2.1. It is obvious that the above scheme is a kth-order approximation to (2.1). We would like to point out that it is also a kth-order (in time) approximation plus the spatial discretization error to (1.1).

On the other hand, if we replace $B_{k-1}(\lambda_h^n g'(u_h^n))$ in the above scheme by zero, then it is easy to see that the second step is equivalent to the simple cutoff approach, which is a first-order approximation to (2.1). However, it is easy to see that the error in maximum norm by the cutoff approach is smaller than the error by the corresponding semi-implicit scheme; therefore, the cutoff approach is also a kth-order (in time) approximation plus the spatial discretization error to (1.1).

2.2. Mass conservation. A drawback of the schemes (2.5)–(2.6) is that it does not preserve mass if the exact solution does.

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 second step.

The first step is still exactly the same as (2.5).

Step 1 (predictor). Solve $\tilde{u}_h^{n+1}$ from

$$\begin{aligned}\frac{\alpha_k \tilde{u}_h^{n+1}(\boldsymbol{z}) - A_k(u_h^n(\boldsymbol{z}))}{\delta t} &+ \mathcal{L}_h \tilde{u}_h^{n+1}(\boldsymbol{z}) + \mathcal{N}_h(B_k(u_h^n(\boldsymbol{z}))) \\ &= B_{k-1}(\lambda_h^n(\boldsymbol{z}) g'(u_h^n(\boldsymbol{z}))) + B_{k-1}(\xi_h^n) \ \ \forall \boldsymbol{z} \in \Sigma_h.\end{aligned} \tag{2.9}$$

We introduce another Lagrange multiplier ξ_h^{n+1} in the second step to enforce the mass conservation.

Step 2 (corrector). Solve $(u_h^{n+1}, \lambda_h^{n+1})$ from

$$\begin{aligned}\frac{\alpha_k (u_h^{n+1}(\boldsymbol{z}) - \tilde{u}_h^{n+1}(\boldsymbol{z}))}{\delta t} &= \lambda_h^{n+1}(\boldsymbol{z}) g'(u_h^{n+1}(\boldsymbol{z})) \\ &\quad - B_{k-1}(\lambda_h^n(\boldsymbol{z}) g'(u_h^n(\boldsymbol{z}))) + \xi_h^{n+1} - B_{k-1}(\xi_h^n) \quad \forall \boldsymbol{z} \in \Sigma_h,\end{aligned} \tag{2.10a}$$

$$\lambda_h^{n+1}(\boldsymbol{z}) \geq 0,\ g(u_h^{n+1}(\boldsymbol{z})) \geq 0,\ \lambda_h^{n+1}(\boldsymbol{z}) g(u_h^{n+1}(\boldsymbol{z})) = 0 \quad \forall \boldsymbol{z} \in \Sigma_h, \tag{2.10b}$$

$$(u_h^{n+1}, 1)_h = (u_h^n, 1)_h, \tag{2.10c}$$

where $(\cdot,\cdot)_h$ is a discrete inner product.

In order to solve the above system, we denote

$$\eta_h^{n+1} := \frac{\delta t}{\alpha_k}(\xi_h^{n+1} - B_{k-1}(\xi_h^n) - B_{k-1}(\lambda_h^n g'(u_h^n))) \tag{2.11}$$

and rewrite (2.10a) as

$$\frac{\alpha_k (u_h^{n+1}(\boldsymbol{z}) - (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z})))}{\delta t} = \lambda_h^{n+1}(\boldsymbol{z}) g'(u_h^{n+1}(\boldsymbol{z})). \tag{2.12}$$

Hence, assuming ξ_h^{n+1} is known, we find from the above and (2.10b) that

$$(u_h^{n+1}(\boldsymbol{z}), \lambda_h^{n+1}(\boldsymbol{z})) = \begin{cases} (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}), 0) & \text{if} \quad a < \tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}) < b, \\ \left(a, \frac{a - (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}))}{\frac{\delta t}{\alpha_k} g'(a)}\right) & \text{if} \quad \tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}) \leq a, \\ \left(b, \frac{b - (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}))}{\frac{\delta t}{\alpha_k} g'(b)}\right) & \text{if} \quad \tilde{u}_h^{n+1}(\boldsymbol{z}) + \eta_h^{n+1}(\boldsymbol{z}) \geq b \end{cases} \tag{2.13}$$

$\forall \boldsymbol{z} \in \Sigma_h$.

It remains to determine ξ_h^{n+1}.

Denote

$$\begin{aligned} {}^a\Sigma_h(\xi) &= \{z \in \Sigma_h : \tilde{u}_h^{n+1}(\boldsymbol{z}) + \delta t \xi \leq a\}, \\ {}^a\Sigma_h^b(\xi) &= \{z \in \Sigma_h : a < \tilde{u}_h^{n+1}(\boldsymbol{z}) + \delta t \xi < b\}, \\ \Sigma_h^b(\xi) &= \{z \in \Sigma_h : \tilde{u}_h^{n+1}(\boldsymbol{z}) + \delta t \xi \geq b\}. \end{aligned} \tag{2.14}$$

Then, thanks to (2.13), the discrete mass conservation (2.10c) can be rewritten as

$$\sum_{z\in {}^a\Sigma_h^b(\eta_h^{n+1})} (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \delta t \eta_h^{n+1})\omega_z + \sum_{z\in\Sigma_h^b(\eta_h^{n+1})} b\,\omega_z + \sum_{z\in {}^a\Sigma_h(\eta_h^{n+1})} a\,\omega_z = (u_h^n, 1)_h. \tag{2.15}$$

Setting

$$\begin{aligned} G_n(\eta) &:= \sum_{z\in {}^a\Sigma_h^b(\eta)} (\tilde{u}_h^{n+1}(\boldsymbol{z}) + \delta t \eta)\omega_z + \sum_{z\in\Sigma_h^b(\eta)} b\,\omega_z + \sum_{z\in {}^a\Sigma_h(\eta)} a\,\omega_z - (u_h^n, 1)_h, \\ F_n(\xi) &:= G_n\left(\frac{\delta t}{\alpha_k}(\xi - B_{k-1}(\xi_h^n) - B_{k-1}(\lambda_h^n g'(u_h^n)))\right), \end{aligned} \tag{2.16}$$

we find from the above and (2.15) that ξ_h^{n+1} is a solution to the nonlinear algebraic equation $F_n(\xi) = 0$. Since $F_n'(\xi)$ may not exist and is difficult to compute if it exists, instead of the Newton iteration, we can use the following secant method:

$$\xi_{k+1} = \xi_k - \frac{F_n(\xi_k)(\xi_k - \xi_{k-1})}{F_n(\xi_k) - F_n(\xi_{k-1})}. \tag{2.17}$$

Since ξ_h^{n+1} is an approximation to zero, we can choose $\xi_0 = 0$ and $\xi_1 = O(\delta t)$. In all our experiments, (2.17) 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 (2.13).

Remark 2.2. It is usually very difficult to construct mass conserved IMEX schemes using the simple cutoff approach. However, replacing $B_{k-1}(\lambda_h^n(\boldsymbol{z})g'(u_h^n(\boldsymbol{z})))$ in (2.9)–(2.10) by zero, we obtain a mass conserved kth-order IMEX cutoff scheme. This is one of the advantages of reformulating the cutoff approach with the operator splitting approach.

3. Stability results. While the schemes constructed in the last section automatically ensure the L^∞ bound for $\{u_h^n\}$, it does not imply any bound on the energy norm $\langle \mathcal{L}\cdot, \cdot\rangle$. In this section, we shall use the energy estimates to derive a bound on the energy norm for $\{\tilde{u}_h^n\}$ as well as a bound on the Lagrange multiplier.

We shall frequently use the following discrete Gronwall lemma [27].

LEMMA 3.1. *Let a_n, b_n, c_n, and d_n be four nonnegative sequences satisfying*

$$a_m + \tau\sum_{n=1}^{m} b_n \le \tau\sum_{n=0}^{m-1} a_n d_n + \tau\sum_{n=0}^{m-1} c_n + C, \qquad m \ge 1,$$

where C and τ are two positive constants. Then

$$a_m + \tau\sum_{n=1}^{m} b_n \le \exp\left(\tau\sum_{n=0}^{m-1} d_n\right)\left(\tau\sum_{n=0}^{m-1} c_n + C\right), \qquad m \ge 1.$$

To fix the idea, we assume that $\mathcal{L}$ is a second-order unbounded positive self-adjoint operator in $L^2(\Omega)$ with domain $D(\mathcal{L})$ and that the nonlinear term can be written as follows:

$$\begin{aligned} &\mathcal{N}(u) = f_1(u) + \nabla\cdot f_2(u) \quad \text{with} \quad f_1(0) = f_2(0) = 0, \\ &\text{and } f_1,\ f_2 \text{ are locally Lipchitz semilinear functions.} \end{aligned} \tag{3.1}$$

Without loss of generality, we assume that $ab \le 0$. Otherwise, we can always find a constant C such that $(a+C)(b+C) \le 0$ and consider the equation for $v = u + C$. Since $ab \le 0$, we have $0 \in (a, b)$. Hence, (3.1) implies in particular

(3.2)
$$|f_1(u)| = |f_1(u) - f_1(0)| \le C_1|u|,\ |f_2(u)| = |f_2(u) - f_2(0)| \le C_2|u| \quad \text{if } a \le u \le b.$$

We observe that the nonlinearities in common nonlinear parabolic equations do satisfy (3.1); see in particular some specific examples given in section 5.

We shall also interpret the first step of the schemes, (2.5) and (2.9), in a Galerkin formulation. More precisely, let $X_h \subset X$ be a subspace with Lagrangian basis functions on Σ_h. We define a discrete inner product on $\Sigma_h = \{\boldsymbol{z}\}$ in $\bar{\Omega}$:

$$(u, v)_h = \sum_{\boldsymbol{z} \in \Sigma_h} \beta_{\boldsymbol{z}} u(\boldsymbol{z}) v(\boldsymbol{z}), \tag{3.3}$$

where we require that the weights $\beta_{\boldsymbol{z}} > 0$. We also denote the induced norm by $\|u\| = (u, u)_h^{\frac{1}{2}}$, and we assume that this norm is equivalent to the L^2 norm for functions in X_h. We denote by $\langle \mathcal{L}_h u_h, v_h \rangle$ the bilinear form on $X_h \times X_h$ based on the discrete inner product after suitable integration by part, and we assume that

$$C_0 \|\nabla u_h\|^2 \le \rangle \mathcal{L}_h u_h, u_h \rangle \quad \forall u_h \in X_h, \tag{3.4}$$

with $C_0 > 0$, which is satisfied by many common spatial discretizations. Hereafter, we shall use C and C_i to denote generic positive constants which are independent of δt and h.

We shall only consider a second-order scheme with mass conservation in this section. It is clear that similar bounds can be derived for second-order schemes without mass conservation and for the first-order schemes, but bounds for higher-order schemes are still elusive. For clarity, we rewrite the second-order version of (2.9)–(2.10) as follows.

Step 1 (predictor). Find $\tilde{u}_h^{n+1} \in X_h$ such that, for all $v_h \in X_h$,

$$\begin{aligned} &\left(\frac{3\tilde{u}_h^{n+1} - 4u_h^n + u_h^{n-1}}{2\delta t}, v_h \right)_h + \langle \mathcal{L}_h \tilde{u}_h^{n+1}, v_h \rangle + (f_1(u_h), v_h)_h - (f_2(u_h), \nabla v_h)_h \\ &\quad = (\lambda_h^n g'(u_h^n) + \xi_h^n, v_h)_h \end{aligned} \tag{3.5}$$

Step 2 (corrector). Find u_h^{n+1}, λ_h^{n+1}, ξ_h^{n+1} from

$$\frac{3(u_h^{n+1}(\boldsymbol{z}) - \tilde{u}_h^{n+1}(\boldsymbol{z}))}{2\delta t}$$

$$= \lambda_h^{n+1}(\boldsymbol{z}) g'(u_h^{n+1}(\boldsymbol{z})) - \lambda_h^n(\boldsymbol{z}) g'(u_h^n(\boldsymbol{z})) + \xi_h^{n+1} - \xi_h^n \quad \forall \boldsymbol{z} \in \Sigma_h, \tag{3.6a}$$

$$\lambda_h^{n+1}(\boldsymbol{z}) \ge 0,\ g(u_h^{n+1}(\boldsymbol{z})) \ge 0,\ \lambda_h^{n+1}(\boldsymbol{z}) g(u_h^{n+1}(\boldsymbol{z})) = 0 \quad \forall \boldsymbol{z} \in \Sigma_h, \tag{3.6b}$$

$$(u_h^{n+1}, 1)_h = (u_h^n, 1)_h \tag{3.6c}$$

and we assume that $\tilde{u}_h^0$ and u_h^0 are computed with the first-order scheme (2.9)–(2.10) with $k = 1$.

THEOREM 3.2. *We assume* (3.1), (3.2), *and* (3.4). *Then, for the scheme* (3.5)–(3.6), *if the generic scheme in* (2.9) *is mass conservative, i.e.,*

$$\langle \mathcal{L}_h \tilde{u}_h^{n+1}, 1 \rangle + (f_1(u_h), 1)_h - (f_2(u_h), \nabla 1)_h = 0, \tag{3.7}$$

then we have

$$4\|u_h^m\|^2 + \|2u_h^m - u_h^{m-1}\|^2 + \frac{4}{3}\delta t^2\|\lambda_h^{n+1}g'(u_h^m) + \xi_h^m\|^2 + 2\delta t\sum_{n=0}^{m-1} C_0\|\nabla\tilde{u}_h^{n+1}\|^2 \le C(T)\|u_h^0\|^2 \ \forall 1 \le m \le T/\delta t.$$

Proof. Choosing $v_h = 4\delta t\tilde{u}_h^{n+1}$ in (3.5), using the assumption (3.4), we obtain

$$\begin{aligned}&(3\tilde{u}_h^{n+1} - 4u_h^n + u_h^{n-1}, 2\tilde{u}_h^{n+1})_h + 4\delta t C_0\|\nabla\tilde{u}_h^{n+1}\|^2\\ &+ 4\delta t(f_1(2u_h^n - u_h^{n-1}), \tilde{u}_h^{n+1})_h - 4\delta t(f_2(2u_h^n - u_h^{n-1}), \nabla\tilde{u}_h^{n+1})_h\\ &\le 4\delta t(\lambda_h^n g'(u_h^n) + \xi_h^n, \tilde{u}_h^{n+1})_h.\end{aligned} \tag{3.8}$$

We start by dealing with the first term in (3.8).

$$\begin{aligned}&(3\tilde{u}_h^{n+1} - 4u_h^n + u_h^{n-1}, 2\tilde{u}_h^{n+1})_h = 2(3u_h^{n+1} - 4u_h^n + u_h^{n-1}, u_h^{n+1})_h\\ &+ 6(\tilde{u}_h^{n+1} - u_h^{n+1}, \tilde{u}_h^{n+1})_h + 2(3u_h^{n+1} - 4u_h^n + u_h^{n-1}, \tilde{u}_h^{n+1} - u_h^{n+1})_h.\end{aligned} \tag{3.9}$$

For the terms on the right-hand side of (3.9), we have

$$\begin{aligned}&2(3u_h^{n+1} - 4u_h^n + u_h^{n-1}, u_h^{n+1})_h = \|u_h^{n+1}\|^2 - \|u_h^n\|^2\\ &+ \|2u_h^{n+1} - u_h^n\|^2 - \|2u_h^n - u_h^{n-1}\|^2 + \|u_h^{n+1} - 2u_h^n + u_h^{n-1}\|^2\end{aligned} \tag{3.10}$$

$$6(\tilde{u}_h^{n+1} - u_h^{n+1}, \tilde{u}_h^{n+1})_h = 3(\|\tilde{u}_h^{n+1}\|^2 - \|u_h^{n+1}\|^2 + \|\tilde{u}_h^{n+1} - u_h^{n+1}\|^2) \tag{3.11}$$

and

$$\begin{aligned}&2(3u_h^{n+1} - 4u_h^n + u_h^{n-1}, \tilde{u}_h^{n+1} - u_h^{n+1})_h\\ &= 2(u_h^{n+1} - 2u_h^n + u_h^{n-1}, \tilde{u}_h^{n+1} - u_h^{n+1})_h + 4(u_h^{n+1} - u_h^n, \tilde{u}_h^{n+1} - u_h^{n+1})_h\\ &\ge -\|u_h^{n+1} - 2u_h^n + u_h^{n-1}\|^2 - \|\tilde{u}_h^{n+1} - u_h^{n+1}\|^2 + 4(u_h^{n+1} - u_h^n, \tilde{u}_h^{n+1} - u_h^{n+1})_h.\end{aligned} \tag{3.12}$$

The last term in the above needs a special treatment. Using (3.6a) and the fact that $(u_h^{n+1} - u_h^n, 1)_h = 0$, we can write

$$\begin{aligned}&4(u_h^{n+1} - u_h^n, \tilde{u}_h^{n+1} - u_h^{n+1})_h\\ &= -\frac{8\delta t}{3}(u_h^{n+1} - u_h^n, \lambda_h^{n+1}g'(u_h^{n+1}) - \lambda_h^n g'(u_h^n) + \xi_h^{n+1} - \xi_h^n)_h\\ &= -\frac{8\delta t}{3}(u_h^{n+1} - u_h^n, \lambda_h^{n+1}g'(u_h^{n+1}) - \lambda_h^n g'(u_h^n))_h - \frac{8\delta t}{3}(\xi_h^{n+1} - \xi_h^n)(u_h^{n+1} - u_h^n, 1)_h\\ &= -\frac{8\delta t}{3}(u_h^{n+1} - u_h^n, \lambda_h^{n+1}g'(u_h^{n+1}))_h - \frac{8\delta t}{3}(u_h^n - u_h^{n+1}, \lambda_h^n g'(u_h^n))_h := I_1 + I_2.\end{aligned} \tag{3.13}$$

Thanks to $\lambda_h^{n+1}(\boldsymbol{z})g(u_h^{n+1}(\boldsymbol{z})) = 0$, we obtain

$$\begin{aligned}I_1 &= -\frac{8\delta t}{3}(\lambda_h^{n+1}, (u_h^{n+1} - u_h^n)(a + b - 2u_h^{n+1}) - g(u_h^{n+1}))_h\\ &= -\frac{8\delta t}{3}(\lambda_h^{n+1}, -(u_h^{n+1})^2 + ab + 2u_h^n u_h^{n+1} - (a+b)u_h^n)_h\\ &= \frac{8\delta t}{3}(\lambda_h^{n+1}, (u_h^{n+1} - u_h^n)^2)_h - \frac{8\delta t}{3}(\lambda_h^{n+1}, (u_h^n - a)(u_h^n - b))_h \ge 0,\end{aligned}$$

where we used the facts that $a \le u_h^n \le b$ and $\lambda_h^{n+1} \ge 0$. Similarly, we use $\lambda_h^n(\boldsymbol{z})g(u_h^n(\boldsymbol{z})) = 0$ to derive

$$\begin{aligned} I_2 &= -\frac{8\delta t}{3}(\lambda_h^n, (u_h^n - u_h^{n+1})g'(u_h^n) - g(u_h^n))_h \\ &= -\frac{8\delta t}{3}(\lambda_h^n, -(u_h^n)^2 + ab + 2u_h^n u_h^{n+1} - (a+b)u_h^{n+1})_h \\ &= \frac{8\delta t}{3}(\lambda_h^n, (u_h^{n+1} - u_h^n)^2)_h - (\lambda_h^n, (u_h^{n+1} - a)(u_h^{n+1} - b))_h \ge 0, \end{aligned}$$

where we used again the facts that $\lambda_h^n \ge 0$ and $a \le u_h^{n+1} \le b$. We derive from the last two inequalities that

(3.14)
$$4(u_h^{n+1} - u_h^n, \tilde{u}_h^{n+1} - u_h^{n+1})_h = -\frac{8\delta t}{3}(u_h^{n+1} - u_h^n, \lambda_h^{n+1}g'(u_h^{n+1}) - \lambda_h^n g'(u_h^n))_h \ge 0.$$

Combining the above inequalities in (3.9), we find

(3.15)
$$\begin{aligned} (3\tilde{u}_h^{n+1} - 4u_h^n + u_h^{n-1}, 2\tilde{u}_h^{n+1})_h \ge \|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \|2u_h^{n+1} - u_h^n\|^2 - \|2u_h^n - u_h^{n-1}\|^2 \\ + 3(\|\tilde{u}_h^{n+1}\|^2 - \|u_h^{n+1}\|^2) + 2\|\tilde{u}_h^{n+1} - u_h^{n+1}\|^2. \end{aligned}$$

Next, we rewrite (3.6a) as

$$3u_h^{n+1} - 2\delta t(\lambda_h^{n+1}g'(u_h^{n+1}) + \xi_h^{n+1}) = 3\tilde{u}_h^{n+1} - 2\delta t(\lambda_h^n g'(u_h^n) + \xi_h^n). \tag{3.16}$$

Taking the discrete inner product of each side of (3.16) with itself, dividing by 3, we obtain

$$\begin{aligned} &3\|u_h^{n+1}\|^2 - 4\delta t(u_h^{n+1}, \lambda_h^{n+1}g'(u_h^{n+1}) + \xi_h^{n+1})_h + \frac{4}{3}\delta t^2\|\lambda_h^{n+1}g'(u_h^{n+1}) + \xi_h^{n+1}\|^2 \\ &\quad = 3\|\tilde{u}_h^{n+1}\|^2 - 4\delta t(\tilde{u}_h^{n+1}, \lambda_h^n g'(u_h^n) + \xi_h^n)_h + \frac{4}{3}\delta t^2\|\lambda_h^n g'(u_h^n) + \xi_h^n\|^2. \end{aligned} \tag{3.17}$$

Note that we can interpret (3.5) pointwise as

$$\begin{aligned} &\frac{3\tilde{u}_h^{n+1}(\boldsymbol{z}) - 4u_h^n(\boldsymbol{z}) + u_h^{n-1}(\boldsymbol{z})}{2\delta t} + \mathcal{L}_h\tilde{u}_h^{n+1}(\boldsymbol{z}) + \mathcal{N}_h(2u_h^n(\boldsymbol{z}) - u_h^{n-1}(\boldsymbol{z})) \\ &\quad = \lambda_h^n(\boldsymbol{z})g'(u_h^n(\boldsymbol{z})) + \xi_h^n \ \forall \boldsymbol{z} \in \Sigma_h, \end{aligned} \tag{3.18}$$

where $\mathcal{N}_h$ is defined by $(\mathcal{N}_h(u_h), v_h)_h = (f_1(u_h), v_h)_h - (f_2(u_h), \nabla v_h)_h$. Summing up (3.18) and (3.6a), we obtain

$$\begin{aligned}(3.19)\qquad &\frac{3u_h^{n+1}(\boldsymbol{z}) - 4u_h^n(\boldsymbol{z}) + u_h^{n-1}(\boldsymbol{z})}{2\delta t} + \mathcal{L}_h\tilde{u}_h^{n+1}(\boldsymbol{z}) + \mathcal{N}_h(2u_h^n(\boldsymbol{z}) - u_h^{n-1}(\boldsymbol{z}))\\ &\quad = \lambda_h^{n+1}(\boldsymbol{z})g'(u_h^{n+1}(\boldsymbol{z})) + \xi_h^{n+1}\ \forall \boldsymbol{z} \in \Sigma_h.\end{aligned}$$

Taking the discrete inner product of (3.19) with 1 on both sides, using (3.6c) and (3.7), we obtain

$$(3.20)\qquad (\lambda_h^{n+1}g'(u_h^{n+1}) + \xi_h^{n+1}, 1)_h = 0,$$

which implies that

$$(3.21)\qquad \xi_h^{n+1} = -\frac{(\lambda_h^{n+1}g'(u_h^{n+1}), 1)_h}{|\Omega|} = -\frac{(\lambda_h^{n+1}, a + b - 2u_h^{n+1})_h}{|\Omega|},$$

where $|\Omega| := (1,1)_h = \Sigma_{\boldsymbol{z}\in\Sigma_k}\beta_{\boldsymbol{z}} > 0$.

It remains to show that the second term of (3.17) is nonnegative. Using the fact that $\lambda_h^{n+1}(\boldsymbol{z})g(u_h^{n+1}(\boldsymbol{z})) = 0$, we have

$$\begin{aligned}&-4\delta t(u_h^{n+1}, \lambda_h^{n+1}g'(u_h^{n+1}) + \xi_h^{n+1})_h\\ &\quad = -4\delta t(\lambda_h^{n+1}, u_h^{n+1}g'(u_h^{n+1}) - g(u_h^{n+1}))_h - 4\delta t\xi_h^{n+1}(u_h^{n+1}, 1)_h\\ &\quad = -4\delta t(\lambda_h^{n+1}, ab - (u_h^{n+1})^2)_h + \frac{4\delta t}{|\Omega|}(\lambda_h^{n+1}, a + b - 2u_h^{n+1})_h(u_h^{n+1}, 1)_h\\ &\quad = -4\delta t\left(\lambda_h^{n+1}, -\left(u_h^{n+1} - \frac{(u_h^{n+1}, 1)_h}{|\Omega|}\right)^2 + \left(\frac{(u_h^{n+1}, 1)_h}{|\Omega|} - a\right)\left(\frac{(u_h^{n+1}, 1)_h}{|\Omega|} - b\right)\right)_h.\end{aligned}$$

Since $a \le u_h^{n+1} \le b$, we have

$$(3.22)\qquad \left(\frac{(u_h^{n+1}, 1)_h}{|\Omega|} - a\right)\left(\frac{(u_h^{n+1}, 1)_h}{|\Omega|} - b\right) \le 0,$$

which, together with $\lambda_h^{n+1} \ge 0$, implies that

$$(3.23)\qquad -4\delta t(u_h^{n+1}, \lambda_h^{n+1}g'(u_h^{n+1}) + \xi_h^{n+1})_h \ge 0.$$

Then, summing up (3.8) with (3.17) and using (3.14), (3.15), and (3.23), after dropping some unnecessary terms, we obtain

$$\begin{aligned}(3.24)\qquad &4\|u_h^{n+1}\|^2 - 4\|u_h^n\|^2 + \|2u_h^{n+1} - u_h^n\|^2 - \|2u_h^n - u_h^{n-1}\|^2 + 2\|\tilde{u}_h^{n+1} - u_h^{n+1}\|^2\\ &+ \frac{4}{3}\delta t^2(\|\lambda_h^{n+1}g'(u_h^{n+1}) + \xi_h^{n+1}\|^2 - \|\lambda_h^n g'(u_h^n) + \xi_h^n\|^2) + 4\delta t C_0\|\nabla\tilde{u}_h^{n+1}\|^2\\ &\le -4\delta t(f_1(2u_h^n - u_h^{n-1}), \tilde{u}_h^{n+1})_h + 4\delta t(f_2(2u_h^n - u_h^{n-1}), \nabla\tilde{u}_h^{n+1})_h.\end{aligned}$$

Using (3.4), the two terms on the right-hand side above can be bounded as follows:

$$\begin{aligned}(3.25)\qquad 4\delta t(f_1(2u_h^n - u_h^{n-1}), \tilde{u}_h^{n+1})_h &= 4\delta t(f_1(2u_h^n - u_h^{n-1}), \tilde{u}_h^{n+1} - u_h^{n+1})_h\\ &\quad + 4\delta t(f_1(2u_h^n - u_h^{n-1}), u_h^{n+1})_h\\ &\le 2\|\tilde{u}_h^{n+1} - u_h^{n+1}\|^2 + 2C_1^2\delta t^2\|2u_h^n - u_h^{n-1}\|^2\\ &\quad + 2\delta t(C_1^2\|2u_h^n - u_h^{n-1}\|^2 + \|u_h^{n+1}\|^2).\end{aligned}$$

Similarly, we have

$$
\begin{aligned}
4\delta t(f_2(2u_h^n-u_h^{n-1}),\nabla\tilde{u}_h^{n+1})_h &\le 2\delta tC_0\|\nabla\tilde{u}_h^{n+1}\|^2+\frac{2\delta t}{C_0}\|f_2(2u_h^n-u_h^{n-1})\|^2\\
&\le 2\delta tC_0\|\nabla\tilde{u}_h^{n+1}\|^2+\frac{2\delta tC_2^2}{C_0}\|2u_h^n-u_h^{n-1}\|^2.
\end{aligned}
\tag{3.26}
$$

Combining (3.24), (3.25), and (3.26), we obtain

$$
\begin{aligned}
&4\|u_h^{n+1}\|^2-4\|u_h^n\|^2+\|2u_h^{n+1}-u_h^n\|^2-\|2u_h^n-u_h^{n-1}\|^2\\
&\quad+\frac{4}{3}\delta t^2(\|\lambda_h^{n+1}g'(u_h^{n+1})+\xi_h^{n+1}\|^2-\|\lambda_h^ng'(u_h^n)+\xi_h^n\|^2)\\
&\quad+2\delta tC_0\|\nabla\tilde{u}_h^{n+1}\|^2\\
&\quad\le C\delta t\|2u_h^n-u_h^{n-1}\|^2+2\delta t\|u_h^{n+1}\|^2\quad\forall n\ge 1.
\end{aligned}
\tag{3.27}
$$

For $n=0$, we use a first-order scheme, namely, (2.9)–(2.10) with $k=1$, to compute $\tilde{u}_h^1$ and u_h^1. Using a similar (but much simplified) procedure as above, we can obtain

$$
\begin{aligned}
&\|u_h^1\|^2-\|u_h^0\|^2+\delta t^2(\|\lambda_h^1g'(u_h^1)+\xi_h^1\|^2-\|\lambda_h^0g'(u_h^0)+\xi_h^0\|^2)+2\delta tC_0\|\nabla\tilde{u}_h^1\|^2\\
&\quad\le C\delta t\|2u_h^0\|^2+2\delta t\|u_h^1\|^2.
\end{aligned}
\tag{3.28}
$$

Finally summing up (3.28) with (3.27) from $n=1$ to $n=m-1$, we obtain

$$
\begin{aligned}
&4\|u_h^m\|^2+\|2u_h^m-u_h^{m-1}\|^2+\frac{4}{3}\delta t^2\|\lambda_h^mg'(u_h^m)+\xi_h^m\|^2+2\delta t\sum_{n=1}^{m-1}C_0\|\nabla\tilde{u}_h^{n+1}\|^2\\
&\le\|2u_h^1-u_h^0\|^2+4\|u_h^0\|^2+C\delta t\sum_{n=0}^{m-1}\{\|2u_h^n-u_h^{n-1}\|^2+\|u_h^{n+1}\|^2\}.
\end{aligned}
$$

Applying the discrete Gronwall (lemma 3.1) and using (3.28), we arrive at the desired result. ∎

4. Error estimate. The error analysis for the second-order scheme (3.5)–(3.6) with a general spatial discretization is very tedious and may obscure its essential difficulty. Therefore, we shall carry out a complete error analysis for a second-order bound preserving scheme with a hybrid spectral discretization that we shall describe below. To further simplify the presentation, we assume $\mathcal{L}=-\Delta$ with Dirichlet boundary conditions on $\Omega=(-1,1)^d$ $(d=1,2,3)$.

We now describe some preliminaries for our hybrid spectral discretization. Letting P_N be the space of polynomials of degree less than or equals to N in each direction, we set

$$
X=H_0^1(\Omega),\quad X_N=\{v\in P_N : v|_{\partial\Omega}=0\}.
\tag{4.1}
$$

We define the projection operator $\Pi_N : X\to X_N$ by

$$
(\nabla(v-\Pi_Nv),\nabla v_N)=0\ \forall v\in X,\ v_N\in X_N,
\tag{4.2}
$$

and recall that for any $r\ge 1$, we have [4]

$$
\|v-\Pi_Nv\|_{H^s}\lesssim N^{s-r}\|v\|_{H^r}\quad\forall v\in H^r(\Omega)\cap X,\ (s=0,\,1),
\tag{4.3}
$$

where $\|\cdot\|_{H^r}$ denotes the usual norm in $H^r(\Omega)$.

Let L_N be the Legendre polynomial of degree N and $\{x_k\}_{0\le k\le N}$ be the roots of $(1-x^2)L'_N(x)$, i.e., the Legendre–Gauss–Lobatto points. We set $\Sigma_N=\{x_k\}_{1\le k\le N-1}$ and $\bar\Sigma_N=\{x_k\}_{0\le k\le N}$ if $d=1$, $\Sigma_N=\{(x_k,x_i)\}_{1\le k,i\le N-1}$ and $\bar\Sigma_N=\{(x_k,x_i)\}_{0\le k,i\le N}$ if $d=2$, and $\Sigma_N=\{(x_k,x_i,x_j)\}_{1\le k,i,j\le N-1}$ and $\bar\Sigma_N=\{(x_k,x_i,x_j)\}_{0\le k,i,j\le N}$ if $d=3$. We define the interpolation operator $I_N: C(\Omega)\to P_N$ by $(I_Nu)(\boldsymbol{z})=u(\boldsymbol{z})$ for all $\boldsymbol{z}\in\bar\Sigma_N$. Then, we also have [4]

$$\|v-I_Nv\|_{H^s}\lesssim N^{s-r}\|v\|_{H^r}\quad\forall v\in H^r(\Omega)\cap X,\ (s=0,\,1). \tag{4.4}$$

Let $(\cdot,\cdot)_N$ be the discrete inner product based on the Gauss–Lobatto quadrature then it is well known that [28]

$$\begin{aligned}&(u_N,v_N)_N=(u_N,v_N)\quad\forall u_N\cdot v_N\in P_{2N-1},\\ &\|v_N\|^2\le(v_N,v_N)_N\le(2+1/N)\|v_N\|^2\quad\forall v_N\in P_N.\end{aligned} \tag{4.5}$$

We observe that the bound preserving is enforced at the second step, so the first step in the bound preserving schemes can be replaced by any other kth-order scheme. We shall consider a second-order modified Crank–Nicholson scheme which is easier to analyze. More precisely, we consider the following modified Crank–Nicholson scheme [18] with a hybrid spectral discretization: find $u_N^{n+1}\in X_N$ such that for all $n\ge1$,

$$\begin{aligned}&\left(\frac{\tilde u_N^{n+1}(\boldsymbol{z})-u_N^n(\boldsymbol{z})}{\delta t},v_N\right)_N+\left(\nabla\frac{3\tilde u_N^{n+1}(\boldsymbol{z})+\tilde u_N^{n-1}(\boldsymbol{z})}{4},\nabla v_N\right)\\ &\qquad+\left(\mathcal{N}\left(\frac{3}{2}u_N^n(\boldsymbol{z})-\frac{1}{2}u_N^{n-1}(\boldsymbol{z})\right),v_N\right)=0\quad\forall v_N\in X_N,\end{aligned} \tag{4.6}$$

and find u_N^{n+1}, λ_N^{n+1} such that

$$\begin{aligned}&\frac{u_N^{n+1}(\boldsymbol{z})-\tilde u_N^{n+1}(\boldsymbol{z})}{\delta t}=\lambda_N^{n+1}(\boldsymbol{z})g'(u_N^{n+1}(\boldsymbol{z}))\quad\forall\boldsymbol{z}\in\Sigma_N,\\ &\qquad\lambda_N^{n+1}(\boldsymbol{z})\ge0,\ g(u_N^{n+1}(\boldsymbol{z}))\ge0,\ \lambda_N^{n+1}(\boldsymbol{z})g(u_N^{n+1}(\boldsymbol{z}))=0\quad\forall\boldsymbol{z}\in\Sigma_N.\end{aligned} \tag{4.7}$$

For $n=0$, we replace $\mathcal{N}(\frac{3}{2}u_N^n(\boldsymbol{z})-\frac{1}{2}u_N^{n-1}(\boldsymbol{z}))$ in (4.6) by $\mathcal{N}(u_N^n(\boldsymbol{z}))$.

To simplify the notation, we shall use $u(t)$ to denote $u(\boldsymbol{x},t)$. We denote

$$\bar e_N^{n+1}=u(t^{n+1})-\Pi_Nu(t^{n+1}),\ \hat e_N^{n+1}=\Pi_Nu(t^{n+1})-u_N^{n+1},\ \tilde e_N^{n+1}=\Pi_Nu(t^{n+1})-\tilde u_N^{n+1}. \tag{4.8}$$

Then, we have

$$u(t^{n+1})-u_N^{n+1}=\bar e_N^{n+1}+\hat e_N^{n+1},\ u(t^{n+1})-\tilde u_N^{n+1}=\bar e_N^{n+1}+\tilde e_N^{n+1}. \tag{4.9}$$

Let $t^k=k\delta t$, $t^{k+\frac12}=\frac12(t^{k+1}+t^k)$ and $u^{n+\frac12}=\frac{u^{n+1}+u^n}{2}$. We denote

$$\begin{aligned}K_N^{n+\frac12}&=\frac{\bar e_N^{n+1}-\bar e_N^n}{\delta t},\\ T_N^{n+\frac12}&=-\Delta\left(u(t^{n+\frac12})-\frac{3u(t^{n+1})+u(t^{n-1})}{4}\right),\\ R_N^{n+\frac12}&=\partial_tu(t^{n+\frac12})-\frac{u(t^{n+1})-u(t^n)}{2},\\ J_N^{n+\frac12}&=u(t^{n+\frac12})-\left(\frac{3}{2}u(t^n)-\frac{1}{2}u(t^{n-1})\right).\end{aligned} \tag{4.10}$$

THEOREM 4.1. *Let* $\tilde{u}_N^{n+1}$, u_N^{n+1}, λ_N^{n+1} *be the solution of* (4.6)–(4.7). *Given* $T \geq 0$, *for some* $l \geq 1$, *assuming* (3.1)–(3.2), *and the exact solution of* (1.1) $u(\boldsymbol{x},t) \in C^2([0,T],H^2(\Omega)) \cap C^1([0,T],H^l(\Omega)) \cap C^3([0,T],L^2(\Omega))$, *then we have the following error estimate:*

$$\|u(t^m) - u_N^m\|^2 + \frac{\delta t}{4}\|\nabla(u(t^m) - \tilde{u}_N^m)\|^2 + \delta t^2 \sum_{n=1}^{m-1} \|\lambda_N^{n+1} g'(u_N^{n+1})\|_N^2$$
$$+ \delta t \sum_{n=1}^{m-1} \|\nabla(u(t^{n+1}) - \tilde{u}_N^{n+1} + u(t^{n-1}) - \tilde{u}_N^{n-1})\|^2 \leq C(\delta t^4 + N^{-2l}) \quad \forall 2 \leq m \leq \frac{T}{\delta t}.$$

Proof. We derive from (1.1) and (4.2) that

$$(\partial_t u, v_N)_N + (\nabla \Pi_N u, \nabla v_N) + (\mathcal{N}(u), v_N) = \epsilon(v_N) \quad \forall v_N \in X_N, \tag{4.11}$$

where

$$\epsilon(v_N) = (\partial_t u, v_N)_N - (\partial_t u, v_N). \tag{4.12}$$

We find from (4.5), the definition of I_N, and (4.3)–(4.4) that

$$\begin{aligned} |\epsilon(v_N)| &= |(u_t, v_N)_N - (u_t, v_N)| = |(u_t - \Pi_{N-1}u_t, v_N)_N + (\Pi_{N-1}u_t - u_t, v_N)| \\ &= |(I_N u_t - \Pi_{N-1}u_t, v_N)_N + (\Pi_{N-1}u_t - u_t, v_N)| \\ &\leq (3\|I_N u_t - \Pi_{N-1}u_t\| + \|\Pi_{N-1}u_t - u_t\|)\|v_N\| \\ &\leq (3\|I_N u_t - u_t\| + 4\|u_t - \Pi_{N-1}u_t\|)\|v_N\| \leq CN^{-l}\|v_N\| \quad \forall v_N \in P_N. \end{aligned} \tag{4.13}$$

Subtracting (4.11) from scheme (4.6), we obtain

$$\begin{aligned} &\left(\frac{\tilde{e}_N^{n+1} - \hat{e}_N^n}{\delta t}, v_N\right)_N + \left(\nabla \frac{3\tilde{e}_N^{n+1} + \tilde{e}_N^{n-1}}{4}, \nabla v_N\right) \\ &\quad + \left(\mathcal{N}(u(t^{n+\frac{1}{2}})) - \mathcal{N}\left(\frac{3}{2}u_N^n - \frac{1}{2}u_N^{n-1}\right), v_N\right) \\ &\quad = (-K_N^{n+\frac{1}{2}}, v_N) - (R_N^{n+\frac{1}{2}}, v_N) - (T_N^{n+\frac{1}{2}}, v_N) + \epsilon(v_N). \end{aligned} \tag{4.14}$$

We also derive from (4.7) that

$$\frac{\hat{e}_N^{n+1}(\boldsymbol{z}) - \tilde{e}_N^{n+1}(\boldsymbol{z})}{\delta t} = s_N^{n+1} \quad \forall \boldsymbol{z} \in \Sigma_N, \tag{4.15}$$

where $s_N^{n+1} = -\lambda_N^{n+1} g'(u_N^{n+1})$. Denoting $Q_N^{n+\frac{1}{2}} = \mathcal{N}(u(t^{n+\frac{1}{2}})) - \mathcal{N}(\frac{3}{2}\Pi_N u(t^n) - \frac{1}{2}\Pi_N u(t^{n-1}))$, we have

$$\begin{aligned} &\left(\mathcal{N}(u(t^{n+\frac{1}{2}})) - \mathcal{N}\left(\frac{3}{2}u_N^n - \frac{1}{2}u_N^{n-1}\right), v_N\right) = (Q_N^{n+\frac{1}{2}}, v_N) \\ &\quad + \left(\mathcal{N}\left(\frac{3}{2}\Pi_N u(t^n) - \frac{1}{2}\Pi_N u(t^{n-1})\right) - \mathcal{N}\left(\frac{3}{2}u_N^n - \frac{1}{2}u_N^{n-1}\right), v_N\right). \end{aligned} \tag{4.16}$$

Then (4.14) can be written as

$$\begin{aligned}
&\left(\frac{\tilde{e}_N^{n+1}-\hat{e}_N^n}{\delta t}, v_N\right)_N + \left(\nabla\frac{3\tilde{e}_N^{n+1}+\tilde{e}_N^{n-1}}{4}, \nabla v_N\right)\\
&+\left(\mathcal{N}\left(\frac{3}{2}\Pi_N u(t^n)-\frac{1}{2}\Pi_N u(t^{n-1})\right)-\mathcal{N}\left(\frac{3}{2}u_N^n-\frac{1}{2}u_N^{n-1}\right), v_N\right)\\
&=-(K_N^{n+\frac{1}{2}}, v_N)-(R_N^{n+\frac{1}{2}}, v_N)-(T_N^{n+\frac{1}{2}}, v_N)-(Q_N^{n+\frac{1}{2}}, v_N)+\epsilon(v_N).
\end{aligned} \tag{4.17}$$

Taking $v_N = 2\delta t\tilde{e}_N^{n+1}$ in (4.17), we obtain

$$\begin{aligned}
&(\tilde{e}_N^{n+1}-\hat{e}_N^n, 2\tilde{e}_N^{n+1})_N + (R_N^{n+\frac{1}{2}}+K_N^{n+\frac{1}{2}}+T_N^{n+\frac{1}{2}}, 2\delta t\tilde{e}_N^{n+1})\\
&+2\delta t\left(\nabla\frac{3\tilde{e}_N^{n+1}+\tilde{e}_N^{n-1}}{4}, \nabla\tilde{e}_N^{n+1}\right)\\
&+\left(\mathcal{N}\left(\frac{3}{2}\Pi_N u(t^n)-\frac{1}{2}\Pi_N u(t^{n-1})\right)-\mathcal{N}\left(\frac{3}{2}u_N^n-\frac{1}{2}u_N^{n-1}\right), 2\delta t\tilde{e}_N^{n+1}\right)\\
&+(Q_N^{n+\frac{1}{2}}, 2\delta t\tilde{e}_N^{n+1}) = 2\delta t\epsilon(\tilde{e}_N^{n+1}).
\end{aligned} \tag{4.18}$$

For the first term in (4.18), we have

$$(\tilde{e}_N^{n+1}-\hat{e}_N^n, 2\tilde{e}_N^{n+1})_N = \|\tilde{e}_N^{n+1}\|_N^2 - \|\hat{e}_N^n\|_N^2 + \|\tilde{e}_N^{n+1}-\hat{e}_N^n\|_N^2. \tag{4.19}$$

We rewrite (4.15) as

$$\hat{e}_N^{n+1}(\boldsymbol{z}) - \delta t s_N^{n+1}(\boldsymbol{z}) = \tilde{e}_N^{n+1}(\boldsymbol{z}) \quad \forall \boldsymbol{z}\in\Sigma_N \tag{4.20}$$

and take the discrete inner product of (4.20) with itself to get

$$\|\hat{e}_N^{n+1}\|_N^2 + \delta t^2\|s_N^{n+1}\|_N^2 - 2\delta t(\hat{e}_N^{n+1}, s_N^{n+1})_N = \|\tilde{e}_N^{n+1}\|_N^2. \tag{4.21}$$

On the other hand,

$$\begin{aligned}
&2\delta t\left(\nabla\frac{3\tilde{e}_N^{n+1}+\tilde{e}_N^{n-1}}{4}, \nabla\tilde{e}_N^{n+1}\right) = \frac{\delta t}{4}\{5(\nabla\tilde{e}_N^{n+1}, \nabla\tilde{e}_N^{n+1}) - (\nabla\tilde{e}_N^{n-1}, \nabla\tilde{e}_N^{n-1})\\
&+(\nabla(\tilde{e}_N^{n+1}+\tilde{e}_N^{n-1}), \nabla(\tilde{e}_N^{n+1}+\tilde{e}_N^{n-1}))\}.
\end{aligned}$$

Combining the above equations, we obtain

$$\begin{aligned}
&\|\hat{e}_N^{n+1}\|_N^2 - \|\hat{e}_N^n\|_N^2 + \|\tilde{e}_N^{n+1}-\hat{e}_N^n\|_N^2 + \delta t^2\|s_N^{n+1}\|_N^2 - 2\delta t(\hat{e}_N^{n+1}, s_h^{n+1})_N\\
&+\frac{\delta t}{4}\{5(\nabla\tilde{e}_N^{n+1}, \tilde{e}_N^{n+1}) - (\nabla\tilde{e}_N^{n-1}, \nabla\tilde{e}_N^{n-1}) + (\nabla(\tilde{e}_N^{n+1}+\tilde{e}_N^{n-1}), \nabla(\tilde{e}_N^{n+1}+\tilde{e}_N^{n-1}))\}\\
&=-(R_N^{n+\frac{1}{2}}+K_N^{n+\frac{1}{2}}+T_N^{n+\frac{1}{2}}+Q_N^{n+\frac{1}{2}}, 2\delta t\tilde{e}_N^{n+1}) - \left(\mathcal{N}\left(\frac{3}{2}\Pi_N u(t^n)-\frac{1}{2}\Pi_N u(t^{n-1})\right)\right.\\
&\left.-\mathcal{N}\left(\frac{3}{2}u_N^n-\frac{1}{2}u_N^{n-1}\right), 2\delta t\tilde{e}_N^{n+1}\right) + 2\delta t\epsilon(\tilde{e}_N^{n+1}).
\end{aligned} \tag{4.22}$$

We now bound the terms on the right-hand side as follows.

Firstly, considering the final term in (4.22), using (4.13), we obtain

$$\begin{aligned}2\delta t\epsilon(\tilde{e}_N^{n+1}) &\le 2C\delta t N^{-l}\|\tilde{e}_N^{n+1}\| \le 2C\delta t N^{-l}\|\tilde{e}_N^{n+1}-\hat{e}_N^n\| + 2C\delta t N^{-l}\|\hat{e}_N^n\| \\ &\le 8C^2\delta t^2 N^{-2l} + \frac{1}{8}\|\tilde{e}_N^{n+1}-\hat{e}_N^n\|^2 + \delta t\|\hat{e}_N^n\|^2 + C^2\delta t N^{-2l}.\end{aligned}$$

Thanks to the KKT condition $\lambda_N^{n+1}\ge 0$ and $a\le \Pi_N u(t^{n+1})\le b$, we find

$$\begin{aligned}&-2\delta t(\hat{e}_N^{n+1}, s_h^{n+1})_N \\ &\quad= -2\delta t(u_N^{n+1}-\Pi_N u(t^{n+1}), \lambda_N^{n+1}g'(u_N^{n+1}))_N + 2\delta t(\lambda_N^{n+1}, g(u_N^{n+1}))_N \\ &\quad= -2\delta t(\lambda_N^{n+1}, -(u_N^{n+1})^2 + ab + 2\Pi_N u(t^{n+1})u_N^{n+1} - (a+b)\Pi_N u(t^{n+1}))_N \\ &\quad= 2\delta t(\lambda_N^{n+1}, (\Pi_N u(t^{n+1}) - u_N^{n+1})^2)_N \\ &\quad- 2\delta t(\lambda_N^{n+1}, (\Pi_N u(t^{n+1}) - a)(\Pi_N u(t^{n+1}) - b))_N \ge 0.\end{aligned}$$

On the other hand, since $\mathcal{N}(u) = f_1(u) + \nabla\cdot f_2(u)$ with (3.1) and (3.2), we have

$$\begin{aligned}&\left(\mathcal{N}\left(\frac{3}{2}\Pi_N u(t^n) - \frac{1}{2}\Pi_N u(t^{n-1})\right) - \mathcal{N}\left(\frac{3}{2}u_N^n - \frac{1}{2}u_N^{n-1}\right), 2\delta t\tilde{e}_N^{n+1}\right) \\ &= \left(f_1\left(\frac{3}{2}\Pi_N u(t^n) - \frac{1}{2}\Pi_N u(t^{n-1})\right) - f_1\left(\frac{3}{2}u_N^n - \frac{1}{2}u_N^{n-1}\right), 2\delta t\tilde{e}_N^{n+1}\right) \\ &- \left(f_2\left(\frac{3}{2}\Pi_N u(t^n) - \frac{1}{2}\Pi_N u(t^{n-1})\right) - f_2\left(\frac{3}{2}u_N^n - \frac{1}{2}u_N^{n-1}\right), 2\delta t\nabla\tilde{e}_N^{n+1}\right).\end{aligned} \tag{4.23}$$

The terms on the right-hand side of (4.23) can be bounded as follows:

$$\begin{aligned}&\left(f_1\left(\frac{3}{2}\Pi_N u(t^n) - \frac{1}{2}\Pi_N u(t^{n-1})\right) - f_1\left(\frac{3}{2}u_N^n - \frac{1}{2}u_N^{n-1}\right), 2\delta t\tilde{e}_N^{n+1}\right) \\ &\le 2C_1\delta t\left(|\frac{3}{2}\hat{e}_N^n - \frac{1}{2}\hat{e}_N^{n-1}|, \tilde{e}_N^{n+1}\right) \\ &= 2C_1\delta t\left(|\frac{3}{2}\hat{e}_N^n - \frac{1}{2}\hat{e}_N^{n-1}|, \tilde{e}_N^{n+1} - \hat{e}_N^n\right) + 2C_1\delta t\left(|\frac{3}{2}\hat{e}_N^n - \frac{1}{2}\hat{e}_N^{n-1}|, \hat{e}_N^n\right) \\ &\le \frac{1}{8}\|\tilde{e}_N^{n+1} - \hat{e}_N^n\|^2 + 8C_1^2\delta t^2\|\frac{3}{2}\hat{e}_N^n - \frac{1}{2}\hat{e}_N^{n-1}\|^2 + C_1\delta t\left(\|\hat{e}_N^n\|^2 + \|\frac{3}{2}\hat{e}_N^n - \frac{1}{2}\hat{e}_N^{n-1}\|^2\right).\end{aligned} \tag{4.24}$$

Similarly,

$$\begin{aligned}&\left(f_2\left(\frac{3}{2}\Pi_N u(t^n) - \frac{1}{2}\Pi_N u(t^{n-1})\right) - f_2\left(\frac{3}{2}u_N^n - \frac{1}{2}u_N^{n-1}\right), 2\delta t\nabla\tilde{e}_N^{n+1}\right) \\ &\le 2\delta t C_2(|\frac{3}{2}\hat{e}_N^n - \frac{1}{2}\hat{e}_N^{n-1}|, \nabla\tilde{e}_N^{n+1}) \le \frac{1}{3}\delta t\|\nabla\tilde{e}_N^{n+1}\|^2 + 3\delta t C_2^2\|\frac{3}{2}\hat{e}_N^n - \frac{1}{2}\hat{e}_N^{n-1}\|^2.\end{aligned} \tag{4.25}$$

It remains to deal with the first term on the right-hand side of (4.22):

$$
\begin{aligned}
&-2\delta t(R_N^{n+\frac{1}{2}}+T_N^{n+\frac{1}{2}},\tilde{e}_N^{n+1})\\
&=-2\delta t(R_N^{n+\frac{1}{2}}+T_N^{n+\frac{1}{2}},\tilde{e}_N^{n+1}-\hat{e}_N^n)-2\delta t(R_N^{n+\frac{1}{2}}+T_N^{n+\frac{1}{2}},\hat{e}_N^n)\\
&\le 4\delta t^2\|R_N^{n+\frac{1}{2}}\|^2+4\delta t^2\|T_N^{n+\frac{1}{2}}\|^2+\frac{1}{4}\|\tilde{e}_N^{n+1}-\hat{e}_N^n\|^2\\
&\qquad+\delta t(\|R_N^{n+\frac{1}{2}}\|^2+\|T_N^{n+\frac{1}{2}}\|^2+\|\hat{e}_N^n\|^2)
\end{aligned}
\tag{4.26}
$$

and

$$
\begin{aligned}
&-2\delta t(K_N^{n+1},\tilde{e}_N^{n+1})=-2\delta t\left(\frac{\bar{e}_N^{n+1}-\bar{e}_N^n}{\delta t},\tilde{e}_N^{n+1}\right)\\
&=-2((I-\Pi_N)(u(t^{n+1})-u(t^n)),\tilde{e}_N^{n+1}-\hat{e}_N^n+\hat{e}_N^n)\\
&\le 2|((I-\Pi_N)(u(t^{n+1})-u(t^n)),\tilde{e}_N^{n+1}-\hat{e}_N^n)|+2|((I-\Pi_N)(u(t^{n+1})-u(t^n)),\hat{e}_N^n)|\\
&\le 8\delta t\int_{t^n}^{t^{n+1}}\|(I-\Pi_N)u_t(t)\|^2dt+\frac{1}{8}\|\tilde{e}_N^{n+1}-\hat{e}_N^n\|^2\\
&\qquad+\int_{t^n}^{t^{n+1}}\|(I-\Pi_N)u_t(t)\|^2dt+\delta t\|\hat{e}_N^n\|^2;
\end{aligned}
\tag{4.27}
$$

and

$$
\begin{aligned}
&\left(Q_N^{n+\frac{1}{2}},2\delta t\tilde{e}_N^{n+1}\right)=\left(\mathcal{N}(u(t^{n+\frac{1}{2}}))-\mathcal{N}\left(\frac{3}{2}\Pi_N u(t^n)-\frac{1}{2}\Pi_N u(t^{n-1})\right),2\delta t\tilde{e}_N^{n+1}\right)\\
&=\left(\mathcal{N}(u(t^{n+\frac{1}{2}}))-\mathcal{N}\left(\frac{3}{2}u(t^n)-\frac{1}{2}u(t^{n-1})\right),2\delta t\tilde{e}_N^{n+1}\right)\\
&+\left(\mathcal{N}\left(\frac{3}{2}u(t^n)-\frac{1}{2}u(t^{n-1})\right)-\mathcal{N}\left(\frac{3}{2}\Pi_N u(t^n)-\frac{1}{2}\Pi_N u(t^{n-1})\right),2\delta t\tilde{e}_N^{n+1}\right).
\end{aligned}
\tag{4.28}
$$

For the first term in right-hand side of (4.28), we have

$$
\begin{aligned}
&\left(\mathcal{N}(u(t^{n+\frac{1}{2}}))-\mathcal{N}\left(\frac{3}{2}u(t^n)-\frac{1}{2}u(t^{n-1})\right),2\delta t\tilde{e}_N^{n+1}\right)\\
&=\left(f_1(u(t^{n+\frac{1}{2}}))-f_1\left(\frac{3}{2}u(t^n)-\frac{1}{2}u(t^{n-1})\right),2\delta t\tilde{e}_N^{n+1}\right)\\
&-\left(f_2(u(t^{n+\frac{1}{2}}))-f_2\left(\frac{3}{2}u(t^n)-\frac{1}{2}u(t^{n-1})\right),2\delta t\nabla\tilde{e}_N^{n+1}\right).
\end{aligned}
\tag{4.29}
$$

Using assumptions (3.1)–(3.2) and Young's inequality, we have

$$
\begin{aligned}
&\left(f_1(u(t^{n+\frac{1}{2}}))-f_1\left(\frac{3}{2}u(t^n)-\frac{1}{2}u(t^{n-1})\right),2\delta t\tilde{e}_N^{n+1}\right)\le C_1(|J_N^{n+\frac{1}{2}}|,2\delta t\tilde{e}_N^{n+1})\\
&=2C_1\delta t(|J_N^{n+\frac{1}{2}}|,\tilde{e}_N^{n+1}-\hat{e}_N^n)+2C_1\delta t(|J_N^{n+\frac{1}{2}}|,\hat{e}_N^n)\\
&\le\frac{1}{4}\|\tilde{e}_N^{n+1}-\hat{e}_N^n\|^2+4C_1^2\delta t^2\|J_N^{n+\frac{1}{2}}\|^2+\delta tC_1(\|\hat{e}_N^n\|^2+\|J_N^{n+\frac{1}{2}}\|^2)
\end{aligned}
\tag{4.30}
$$

and

(4.31)
$$\left(f_2(u(t^{n+\frac{1}{2}})) - f_2\left(\frac{3}{2}u(t^n) - \frac{1}{2}u(t^{n-1})\right), 2\delta t\nabla\tilde{e}_N^{n+1}\right) \le C_2(|J_N^{n+\frac{1}{2}}|, 2\delta t\nabla\tilde{e}_N^{n+1})$$
$$\le 3\delta t C_2^2\|J_N^{n+\frac{1}{2}}\|^2 + \frac{1}{3}\delta t\|\nabla\tilde{e}_N^{n+1}\|^2.$$

For the second term in the right-hand side of (4.28), similar with (4.30) and (4.31), we have

(4.32)
$$\left(\mathcal{N}\left(\frac{3}{2}u(t^n) - \frac{1}{2}u(t^{n-1})\right) - \mathcal{N}\left(\frac{3}{2}\Pi_N u(t^n) - \frac{1}{2}\Pi_N u(t^{n-1})\right), 2\delta t\tilde{e}_N^{n+1}\right)$$
$$\le C_1\left(|\frac{3}{2}\bar{e}_N^n - \frac{1}{2}\bar{e}_N^{n-1}|, 2\delta t\tilde{e}_N^{n+1}\right) + C_2\left(|\frac{3}{2}\bar{e}_N^n - \frac{1}{2}\bar{e}_N^{n-1}|, 2\delta t\nabla\tilde{e}_N^{n+1}\right).$$

For the first term in the right-hand side of (4.32), using assumption (4.3), we have

$$C_1\left(|\frac{3}{2}\bar{e}_N^n - \frac{1}{2}\bar{e}_N^{n-1}|, 2\delta t\tilde{e}_N^{n+1}\right)$$
$$\le 2C_1\delta t\left(|\frac{3}{2}\bar{e}_N^n - \frac{1}{2}\bar{e}_N^{n-1}|, \tilde{e}_N^{n+1} - \hat{e}_N^n\right) + 2C_1\delta t\left(|\frac{3}{2}\bar{e}_N^n - \frac{1}{2}\bar{e}_N^{n-1}|, \hat{e}_N^n\right)$$
$$\le C\delta t N^{-2l} + \frac{1}{8}\|\tilde{e}_N^{n+1} - \hat{e}_N^n\|^2 + C_1\delta t\|\hat{e}_N^n\|^2.$$

For the second term in the right-hand side of (4.32), we have

$$C_2\left(|\frac{3}{2}\bar{e}_N^n - \frac{1}{2}\bar{e}_N^{n-1}|, 2\delta t\nabla\tilde{e}_N^{n+1}\right)$$
$$\le \frac{1}{3}\delta t\|\nabla\tilde{e}_N^{n+1}\|^2 + 3\delta t C_2^2\|\frac{3}{2}\bar{e}_N^n - \frac{1}{2}\bar{e}_N^{n-1}\|^2$$
$$\le \frac{1}{3}\delta t\|\nabla\tilde{e}_N^{n+1}\|^2 + C\delta t N^{-2l}.$$

Combining the above relations into (4.22) and using (4.5), we arrive at

(4.33)
$$\|\hat{e}_N^{n+1}\|_N^2 - \|\hat{e}_N^n\|_N^2 + \delta t^2\|s_N^{n+1}\|_N^2 + \frac{\delta t}{4}\{(\nabla\tilde{e}_N^{n+1}, \nabla\tilde{e}_N^{n+1}) - (\nabla\tilde{e}_N^{n-1}, \nabla\tilde{e}_N^{n-1})$$
$$+ (\nabla(\tilde{e}_N^{n+1} + \tilde{e}_N^{n-1}), \nabla(\tilde{e}_N^{n+1} + \tilde{e}_N^{n-1}))\} \le (8C_1^2\delta t^2 + C_1\delta t + 3\delta t C_2^2)\left\|\frac{3}{2}\hat{e}_N^n - \frac{1}{2}\hat{e}_N^{n-1}\right\|^2$$
$$+ 3(C_1+1)\delta t\|\hat{e}_N^n\|_N^2 + (4\delta t^2 + \delta t)\|R_N^{n+\frac{1}{2}}\|^2 + (4\delta t^2 + \delta t)\|T_N^{n+\frac{1}{2}}\|^2$$
$$+ (4C_1^2\delta t^2 + \delta t C_1 + 3\delta t C_2^2)\|J_N^{n+\frac{1}{2}}\|^2 + 8\delta t\int_{t^n}^{t^{n+1}}\|(I-\Pi_N)u_t(t)\|^2 dt$$
$$+ \int_{t^n}^{t^{n+1}}\|(I-\Pi_N)u_t(t)\|^2 dt + (C^2 + 2C + 8C^2\delta t)\delta t N^{-2l} \quad \forall n \ge 1.$$

For $n = 0$, a similar estimate can be easily derived. Summing up (4.33) from $n = 1$ to $n = m - 1$ and its corresponding inequality at $n = 0$, we obtain

(4.34)

$$\begin{aligned}
&\|\hat{e}_N^m\|_N^2 + \delta t^2 \sum_{n=1}^{m-1} \|s_N^{n+1}\|_N^2 + \frac{\delta t}{4}\|\nabla \tilde{e}_N^m\|^2 + \frac{\delta t}{4}\|\nabla \tilde{e}_N^{m-1}\|^2 + \delta t \sum_{n=1}^{m-1} \|\nabla(\tilde{e}_N^{n+1} + \tilde{e}_N^{n-1})\|^2 \\
&\le \|\hat{e}_N^0\|_N^2 + \frac{\delta t}{4}\|\nabla \tilde{e}_N^0\|^2 + \frac{\delta t}{4}\|\nabla \tilde{e}_N^1\|^2 \\
&+ \sum_{n=0}^{m-1} \left\{ (8C_1^2\delta t^2 + C_1\delta t + 3\delta t C_2^2) \left\| \frac{3}{2}\hat{e}_N^n - \frac{1}{2}\hat{e}_N^{n-1} \right\|^2 \right. \\
&+ 3(C_1+1)\delta t \|\hat{e}_N^n\|^2 + (4\delta t^2 + \delta t)\|R_N^{n+\frac{1}{2}}\|^2 + (4\delta t^2 + \delta t)\|T_N^{n+\frac{1}{2}}\|^2 \\
&\left. +(4C_1^2\delta t^2 + \delta t C_1 + 3\delta t C_2^2)\|J_N^{n+\frac{1}{2}}\|^2 \right\} + 8\delta t \int_0^T \|(I - \Pi_N)u_t(t)\|^2 dt \\
&+ \int_0^T \|(I - \Pi_N)u_t(t)\|^2 dt + (C^2 + 2C + 8C^2\delta t) T N^{-2l}.
\end{aligned}$$

For the term in (4.34) with $l \ge 0$, we have

$$\int_0^T \|(I - \Pi_N)u_t(t)\|^2 dt \le C N^{-2l} \|u_t\|^2_{L^2(0,T;H^l)}, \quad \|T_N^{n+\frac{1}{2}}\|^2 \le C\delta t^3 \int_{t^n}^{t^{n+1}} \|u_{tt}\|^2_{H^2} dt,$$

$$\|J_N^{n+\frac{1}{2}}\|^2 \le C\delta t^3 \int_{t^n}^{t^{n+1}} \|u_{tt}\|^2 dt, \quad \|R_N^{n+\frac{1}{2}}\|^2 \le C\delta t^3 \int_{t^n}^{t^{n+1}} \|u_{ttt}\|^2 dt.$$

Finally, applying the discrete Gronwall lemmas (Lemma 3.1) to the above after dropping some unnecessary positive terms, using the norm equivalence (4.5) and the triangular inequality, we obtain the desired result. ∎

Remark 4.2. By following exactly the same procedure, we can also derive a similar error estimate if we use a hybrid Fourier spectral method instead of the hybrid Legendre spectral method.

5. Some typical applications. The bound preserving schemes that we constructed and studied in previous sections can be applied to a large class of PDEs which are bound preserving. We describe applications to several typical examples below.

5.1. Allen–Cahn equation. Consider the Allen–Cahn equation [1]

$$u_t - \Delta u + \frac{1}{\epsilon^2} u(u^2 - 1) = 0 \tag{5.1}$$

with homogeneous Dirichlet, homogeneous Neumann, or periodic boundary condition and where ϵ is a positive constant. It is well known that the above equation satisfies the maximum principle; in particular, if the values of the initial condition u_0 is in $[-1, 1]$, the solution of the Allen–Cahn equation (5.1) will stay within the range $[-1, 1]$. Setting $\mathcal{L} = -\Delta + \frac{1}{\epsilon^2}$ and $\mathcal{N}(u) = f_1(u) = \frac{1}{\epsilon^2}u(u^2 - 1) - \frac{1}{\epsilon^2}$, a second-order scheme based on the modified Crank–Nicholson for (5.1) is

$$\frac{\tilde{u}^{n+1} - u^n}{\delta t} + \mathcal{L}\left(\frac{3}{4}\tilde{u}^{n+1} + \frac{1}{4}\tilde{u}^{n-1}\right) + \mathcal{N}\left(\frac{3}{2}u^n - \frac{1}{2}u^{n-1}\right) = \lambda^n g'(u^n) \tag{5.2}$$

and

$$\begin{aligned} &\frac{u^{n+1}-\tilde{u}^{n+1}}{\delta t}=\frac{1}{2}(\lambda^{n+1}g'(u^{n+1})-\lambda^n g'(u^n)),\\ &\lambda^{n+1}\geq 0,\ g(u^{n+1})\geq 0,\ \lambda^{n+1}g(u^{n+1})=0, \end{aligned} \tag{5.3}$$

where $g(u)=(1+u)(1-u)$.

Similarly, we have its cutoff version:

$$\frac{\tilde{u}^{n+1}-u^n}{\delta t}+\mathcal{L}\left(\frac{3}{4}\tilde{u}^{n+1}+\frac{1}{4}\tilde{u}^{n-1}\right)+\mathcal{N}\left(\frac{3}{2}u^n-\frac{1}{2}u^{n-1}\right)=0 \tag{5.4}$$

and

$$\begin{aligned} &\frac{u^{n+1}-\tilde{u}^{n+1}}{\delta t}=\lambda^{n+1}g'(u^{n+1}),\\ &\lambda^{n+1}\geq 0,\ g(u^{n+1})\geq 0,\ \lambda^{n+1}g(u^{n+1})=0. \end{aligned} \tag{5.5}$$

Since $f_1(u)=0$ and $f_2(u)$ is certainly locally Lipschitz and satisfies (3.1)–(3.4), results which are similar to those in Theorems 3.2 and 4.1 can be derived for the above schemes.

5.2. Cahn–Hilliard equation with variable mobility. Consider the Cahn–Hilliard equation [3] with a logarithmic potential:

$$\begin{aligned} &u_t=\nabla\cdot(M(u)\nabla\mu),\\ &\mu=-\epsilon^2\Delta u+\ln(1+u)-\ln(1-u)-\theta_0 u, \end{aligned} \tag{5.6}$$

where μ is the chemical potential and $M(u)=1-u^2>0$ is the mobility function. θ_0,ϵ are two positive constants. u and μ are prescribed with homogeneous Neumann or periodic boundary condition. The Cahn–Hilliard equation (5.6) is a gradient flow which takes on the form

$$u_t=\nabla\cdot\left(M(u)\nabla\frac{\delta E}{\delta u}\right) \tag{5.7}$$

with the total free energy

$$E(u)=\int_\Omega (1+u)\ln(1+u)+(1-u)\ln(1-u)-\frac{\theta_0}{2}u^2+\frac{\epsilon^2}{2}|\nabla u|^2 d\boldsymbol{x}. \tag{5.8}$$

With a given initial condition $\|u_0\|_{L^\infty}<1-\gamma$ for a constant $\gamma\in(0,1)$, due to the singular logarithmic potential, the solution of Cahn–Hilliard equation (5.6) is expected to remain in the range $(-1+\delta,1-\delta)$ for some $\delta\in(0,1)$ [10, 16]. Note that (5.6) is a fourth-order equation written as a system of two coupled second-order equations, so the approach for constructing bound preserving schemes introduced in section 2 can be directly applied to (5.6). For example, the second-order version of (2.5)–(2.6) for (5.6) is as follows:

$$\begin{aligned} &\frac{3\tilde{u}^{n+1}-4u^n+u^{n-1}}{2\delta t}=\nabla\cdot(M(2u^n-u^{n-1})\nabla\mu^{n+1})+\lambda^n g'(u^n),\\ &\mu^{n+1}=-\epsilon^2\Delta u^{n+1}+\ln(1+2u^n-u^{n-1})-\ln(1-2u^n+u^{n-1})-\theta_0(2u^n-u^{n-1}) \end{aligned} \tag{5.9}$$

and

$$\begin{aligned} &\frac{3u^{n+1}-3\tilde{u}^{n+1}}{2\delta t}=\lambda^{n+1}g'(u^{n+1})-\lambda^n g'(u^n),\\ &\lambda^{n+1}\geq 0,\ g(u^{n+1})\geq 0,\ \lambda^{n+1}g(u^{n+1})=0, \end{aligned} \tag{5.10}$$

where $g(u)=(u+1-\delta)(1-\delta-u)$. Notice that $g(u)=(u+1-\delta)(1-\delta-u)>0$ is equivalent to $-1+\delta\leq u\leq 1-\delta$.

The system (5.6) also preserves mass. Indeed, integrating the first equation in (5.6) over Ω, we obtain $\partial_t\int_\Omega u d\boldsymbol{x}=0$. As described in section 2, we can also easily modify the scheme (5.9)–(5.10) to construct a bound and mass preserving scheme for (5.6).

While the stability results in section 3 was derived only for a second-order equation for the sake of simplicity, since the nonlinear term $\mathcal{N}(u)=f_2(u)=\ln(1+u)-\ln(1-u)-\theta_0 u$ is locally Lipschitz for $u\in(-1,1)$ and satisfies (3.1)–(3.4), a similar procedure can be used to derive a stability result which is similar to Theorem 3.2. However, the error analysis in section 4 can not be easily extended to this case.

5.3. Fokker–Planck equation. Consider the following Fokker–Planck equation:

$$\partial_t u=\partial_x(xu(1-u)+\partial_x u) \tag{5.11}$$

with no flux or periodic boundary conditions, which models the relaxation of fermion and boson gases taking on the form described in [5, 29]. The long time asymptotics of the one dimensional model has been studied in [5].

The Fokker–Planck equation (5.11) can be interpreted as a gradient flow:

$$\partial_t u=\partial_x\left(u(1-u)\partial_x\frac{\delta E}{\delta u}\right), \tag{5.12}$$

with $E(u)$ being the entropy functional

$$E(u)=\int_\Omega\left(\frac{x^2}{2}u+u\log(u)+(1-u)\log(1-u)\right)d\boldsymbol{x}. \tag{5.13}$$

Hence, the solution of (5.11) is expected to take values in $[0,1]$.

The approach for constructing bound preserving schemes introduced in section 2 can be directly applied to (5.11). For example, letting $\mathcal{L}u=-\partial_{xx}u$ and $\mathcal{N}(u)=\partial_x f_2(u)=\partial_x(-xu(1-u))$, a second-order version of (2.5)–(2.6) for (5.11) is as follows:

$$\frac{3u^{n+1}-4u^n+u^{n-1}}{2\delta t}=\partial_x(x(2u^n-u^{n-1})(1-2u^n+u^{n-1})+\partial_x u^{n+1})+\lambda^n g'(u^n) \tag{5.14}$$

and

$$\begin{aligned} &\frac{3\tilde{u}^{n+1}-3\tilde{u}^{n+1}}{2\delta t}=\lambda^{n+1}g'(u^{n+1})-\lambda^n g'(u^n),\\ &\lambda^{n+1}\geq 0,\ g(u^{n+1})\geq 0,\ \lambda^{n+1}g(u^{n+1})=0, \end{aligned} \tag{5.15}$$

where $g(u)=u(1-u)$.

We observe that the Fokker–Plank equation (5.11) with no flux or periodic boundary conditions conserves mass, i.e., $\partial_t \int_\Omega u d\boldsymbol{x} = 0$. The above scheme can be easily modified to be mass conserving as follows:

(5.16)
$$\frac{3u^{n+1} - 4u^n + u^{n-1}}{2\delta t} = \partial_x(x(2u^n - u^{n-1})(1 - 2u^n + u^{n-1}) + \partial_x u^{n+1}) + \lambda^n g'(u^n)$$

and

$$\begin{aligned} &\frac{3u^{n+1} - 3\tilde{u}^{n+1}}{2\delta t} = \lambda^{n+1} g'(u^{n+1}) - \lambda^n g'(u^n) + \xi^{n+1}, \\ &\lambda^{n+1} \geq 0,\ g(u^{n+1}) \geq 0,\ \lambda^{n+1} g(u^{n+1}) = 0,\ (u^{n+1}, 1) = (u^n, 1). \end{aligned} \tag{5.17}$$

It is clear that $f_2(u) = -xu(1-u)$ is locally Lipschitz and satisfies (3.1)–(3.2) with $f_1(u) = 0$. Therefore, a similar result as in Theorem 3.2 can be derived for the scheme (5.14)–(5.15) and (5.16)–(5.17).

6. Numerical results. In this section, we will present various numerical experiments to validate the proposed bound preserving schemes. For the examples presented below, if not specified, by default we assume periodic boundary conditions in $\Omega = [0, 2\pi)^d$ and use a Fourier spectral method for spatial approximation.

6.1. Allen–Cahn equation. The first example is the Allen–Cahn equation (5.1).

6.1.1. Accuracy test. We first verify the convergence rate for the scheme (5.2)–(5.3) and its first-order version for (2.1) in the domain $\Omega = [0, 2\pi]^2$ with the initial condition

$$u(x, y, 0) = \tanh\left(\frac{1 - \sqrt{(x-\pi)^2 + (y-\pi)^2}}{\sqrt{2}\epsilon}\right). \tag{6.1}$$

We use 128^2 uniform collocation points in $[0, 2\pi]^2$, i.e., $\Sigma_N = \{x_{jk} = (\frac{j}{2\pi}, \frac{k}{2\pi}); j, k = 0, 1, \ldots, 128\}$ so that the spatial discretization error is negligible compared with the time discretization error. We shall test their accuracy as approximations of (2.1) and (1.1), respectively.

First, we consider these schemes as approximations of (2.1) and use the reference solution computed by (5.2)–(5.3) with a very small time step $\delta t = 10^{-6}$. We observe from Table 1 that the scheme (5.2)–(5.3) (resp., its first-order version) achieves a second-order (resp., first-order) convergence rate in time. The scheme (5.4)–(5.5) only achieves first-order convergence in time. We plot in Figure 1 the profile of numerical solution u and the Lagrange multiplier λ at $T = 0.001$.

TABLE 1
Accuracy test for approximations to (2.1)*: The* L^∞ *errors at* $t = 0.01$ *with* $\epsilon^2 = 0.001$.

δt	First-order version of (5.2)–(5.3)	Order	(5.2)–(5.3)	Order	(5.4)–(5.5)	Order
4×10^{-5}	$4.89E(-3)$	–	$3.56E(-4)$	–	$1.36E(-3)$	–
2×10^{-5}	$2.47E(-3)$	0.98	$9.50E(-5)$	1.90	$6.75E(-4)$	1.01
1×10^{-5}	$1.24E(-3)$	0.99	$2.31E(-5)$	2.04	$3.24E(-4)$	1.06
5×10^{-6}	$6.22E(-4)$	0.99	$5.84E(-6)$	1.98	$1.44E(-4)$	1.17
2.5×10^{-6}	$3.11E(-4)$	1.00	$1.25E(-6)$	2.22	$5.43E(-5)$	1.40

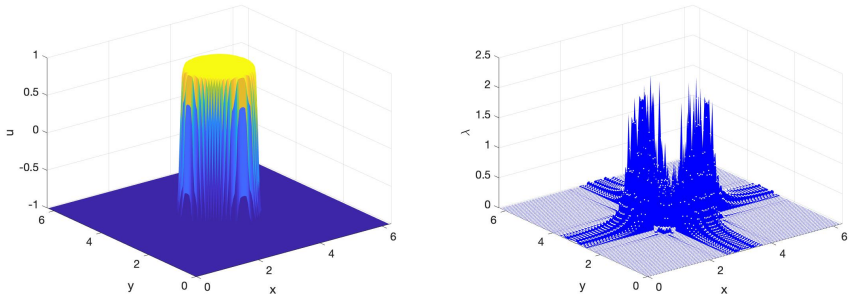


FIG. 1. *Numerical solution u and Lagrange multiplier λ at $T = 0.001$ computed by scheme (2.5)–(2.6) with $k = 2$ and $\delta t = 10^{-6}$.*

TABLE 2
Accuracy test for approximations to (1.1): The L^∞ errors at $t = 0.01$ with $\epsilon^2 = 0.001$.

δt	(5.2)–(5.3)	Order	(5.4)–(5.5)	Order
4×10^{-5}	$1.05E(-4)$	–	$1.05E(-4)$	–
2×10^{-5}	$4.25E(-5)$	1.30	$4.25E(-5)$	1.30
1×10^{-5}	$1.00E(-5)$	2.08	$1.00E(-5)$	2.08
5×10^{-6}	$2.76E(-6)$	1.86	$2.76E(-6)$	1.86
2.5×10^{-6}	$6.29E(-7)$	2.13	$6.29E(-7)$	2.13

We then consider these schemes as approximations of (1.1) and use the reference solution as a highly accurate approximation to the original PDE (1.1) which is computed by a standard semi-implicit scheme with $\delta t = 10^{-8}$. We compare the accuracy between the scheme (5.2)–(5.3) and its cutoff version (5.4)–(5.5). The results are reported in Table 2. We observe that both schemes have essentially the same accuracy and are second-order in time, which is consistent with the error estimates in Theorem 4.1.

The results reported in Tables 1 and 2 are consistent with Remark 2.1.

6.1.2. Comparison with a usual semi-implicit scheme. We consider the Allen–Cahn equation with $\epsilon^2 = 0.001$ and the initial condition

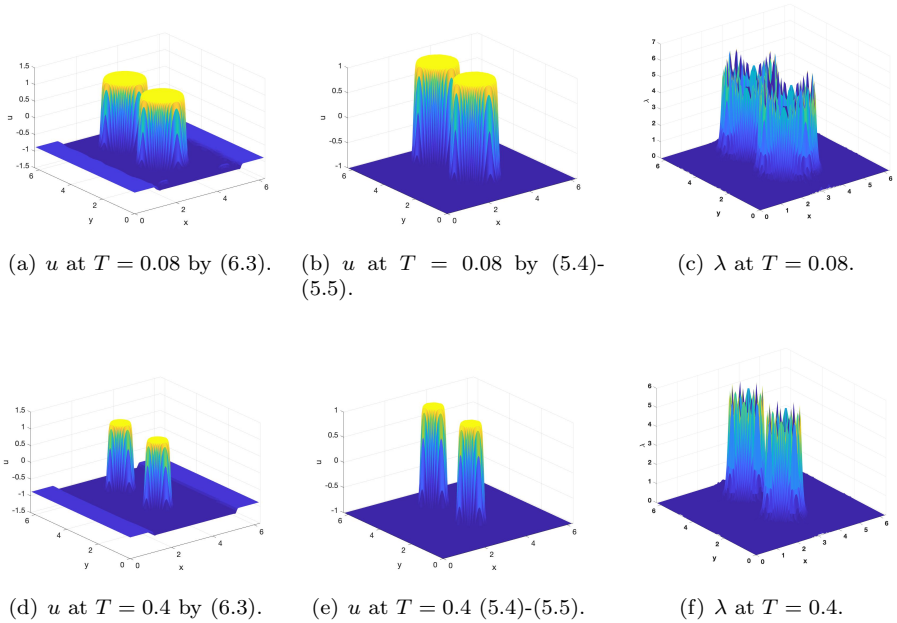
$$u(x, y, 0) = \tanh\left(\frac{1 - \sqrt{(x-\pi)^2 + (y - 3\pi/2)^2}}{\sqrt{2}\epsilon}\right) + \tanh\left(\frac{1 - \sqrt{(x-\pi)^2 + (y - 3\pi/4)^2}}{\sqrt{2}\epsilon}\right) + 1. \tag{6.2}$$

We use the scheme (5.4)–(5.5) and its usual semi-implicit version:

$$\frac{u^{n+1} - u^n}{\delta t} + \mathcal{L}\left(\frac{3}{4}u^{n+1} + \frac{1}{4}u^{n-1}\right) + \mathcal{N}\left(\frac{3}{2}u^n - \frac{1}{2}u^{n-1}\right) = 0 \tag{6.3}$$

with time step $\delta t = 8 \times 10^{-4}$ and 128^2 Fourier modes.

In Figure 2, we plot the numerical solution u at $T = 0.08$ and $T = 0.4$ using the semi-implicit scheme (6.3) and the bound preserving scheme (5.4)–(5.5). It is observed that the numerical solution by the bound preserving scheme stays within $[-1, 1]$, while that by the semi-implicit scheme (6.3) violates this property. The Lagrange multiplier



(a) u at $T = 0.08$ by (6.3). (b) u at $T = 0.08$ by (5.4)-(5.5). (c) λ at $T = 0.08$.

(d) u at $T = 0.4$ by (6.3). (e) u at $T = 0.4$ (5.4)-(5.5). (f) λ at $T = 0.4$.

FIG. 2. (a) and (d): *Numerical solutions at* $T = 0.08, 0.4$ *computed by* (6.3). (b)–(c) *and* (e)–(f): *numerical solutions and Lagrange multiplier* λ *at* $T = 0.08, 0.4$ *computed by* (5.4)–(5.5).

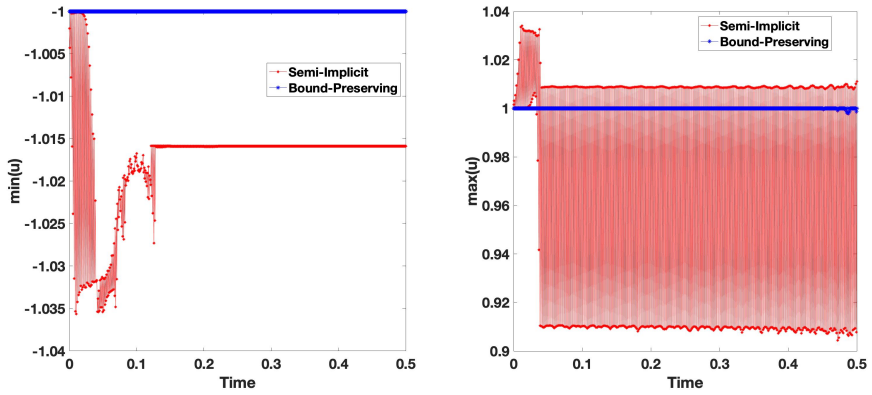


FIG. 3. *Evolution of* $\max\{u\}$ *and* $\min\{u\}$ *with respect to time for the semi-implicit scheme* (6.3) *and the bound preserving scheme* (5.4)–(5.5).

λ by the bound preserving scheme (5.4)–(5.5) is also shown in Figure 2. In Figure 3, we plot the evolution of $\max\{u\}$ and $\min\{u\}$ by both schemes.

6.2. Porous medium equation. For our bound preserving schemes, the L^∞ stability is guaranteed without any time step restriction; we shall test the performance of our bound preserving schemes in accuracy for a porous medium equation [31] with larger time steps. We consider the porous medium equation which takes on the form

$$u_t = \Delta u^m = m\nabla \cdot (u^{m-1}\nabla u) \tag{6.4}$$

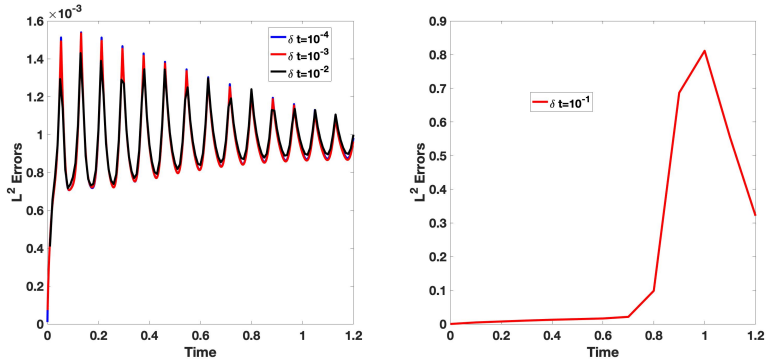


FIG. 4. *L^2 errors with different time steps for porous medium equation* (6.4) *using a second-order bound preserving scheme.*

with Dirichlet or Neumann boundary conditions. $m \geq 1$ is a physical parameter. For the predictor step, we construct the following scheme:

$$\frac{3\tilde{u}^{n+1} - 4u^n + u^{n-1}}{2\delta t} = m\nabla \cdot ((u^{n+1,*})^{m-1}\nabla \tilde{u}^{n+1}) + \lambda^n g'(u^n), \tag{6.5}$$

where

$$u^{n+1,*} = \begin{cases} 2u^n - u^{n-1} & \text{if } u^n \geq u^{n-1}, \\ \frac{1}{2/u^n - 1/u^{n-1}} & \text{if } u^n < u^{n-1}. \end{cases} \tag{6.6}$$

The correction step can be constructed as (5.15).

The exact solution of the porous medium equation (6.4) in the Barenblatt form [31] is

$$u(x,t) = \frac{1}{t_0^\alpha}\Big(C - \alpha\frac{m-1}{2m}\frac{x^2}{t_0^{2\alpha}}\Big)_+^{\frac{1}{m-1}}, \tag{6.7}$$

where $f_+ = \max\{f, 0\}$, $\alpha = \frac{1}{m+1}$, $C = 1$, and $t_0 = t + 1$.

We consider the Dirichlet boundary condition and use the Legendre spectral method [28] with $N = 128$ spectral collocations in domain $[-5, 5]$. The initial condition is set to be $u^0(x) = u(x,t)|_{t=0}$ in (6.7). In Figure 4, we depict L^2 errors for porous medium equation (6.4) with $m = 2$ using various time steps. It is observed that time steps $\delta t = 10^{-2}, 10^{-3}, 10^{-4}$ can be allowed to obtain good accuracy. If time step $\delta t = 10^{-1}$ is chosen, the error will be extremely large, but the bound preserving scheme is still stable.

6.3. Cahn–Hilliard equation. We now consider the Cahn–Hilliard equation (5.6) with the initial condition

$$u_0(x,y) = 0.2 + 0.05\,\mathrm{rand}(x,y), \tag{6.8}$$

where function $\mathrm{rand}(x,y)$ is a uniformed distributed random function with values in $(-1,1)$. We set $\theta_0 = 5$ and $\epsilon = 0.1$ and use $\delta t = 10^{-5}$ with 128^2 Fourier modes in

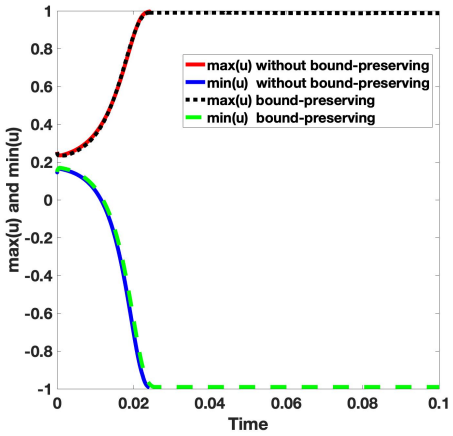
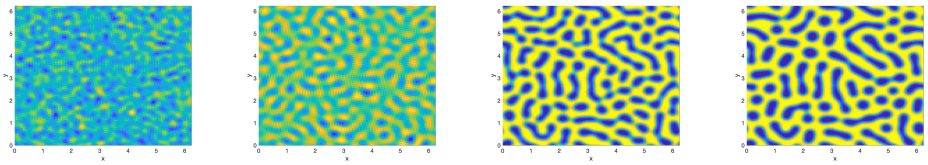


FIG. 5. *The evolution of* $\max_{i,j} u^n_{i,j}$ *and* $\min_{i,j} u^n_{i,j}$ *with respect to time computed by the scheme* (6.9) *up to* $t \approx .025$ *and by the scheme* (5.9)–(5.10) *up to* $t = 0.1$.



(a) $t = 0.001$ (b) $t = 0.02$ (c) $t = 0.05$ (d) $t = 0.1$

FIG. 6. *Numerical solutions of the Cahn–Hilliard equation at* $t = 0.001, 0.02, 0.05, 0.1$ *computed by the scheme* (5.9)–(5.10).

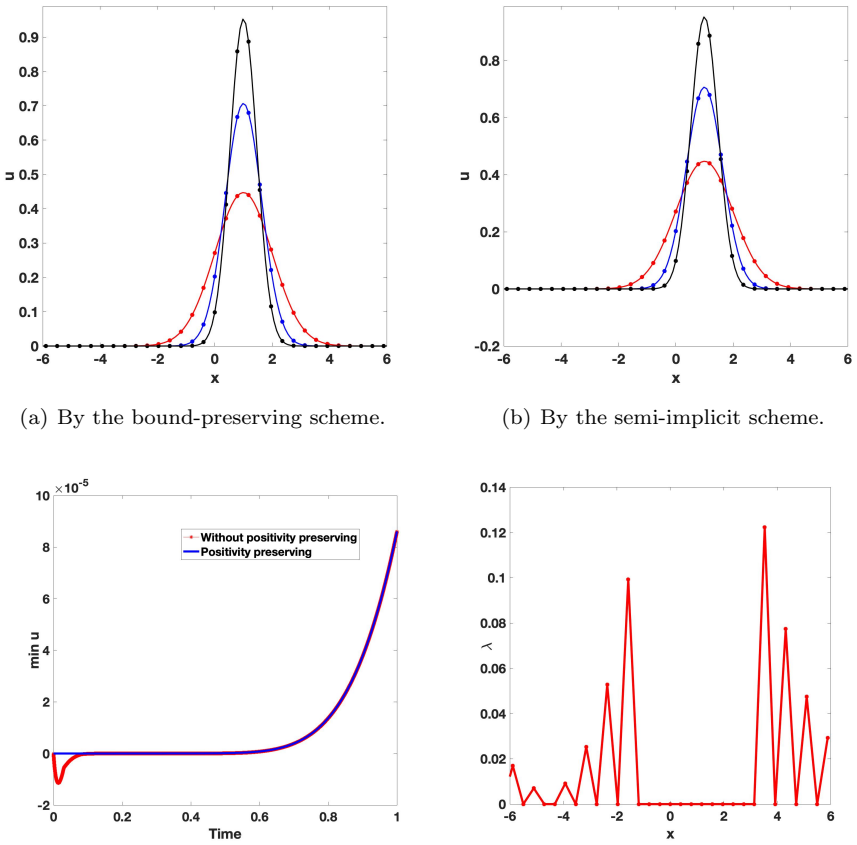
$(0, 2\pi)^2$. We first use the following semi-implicit scheme:

(6.9)
$$\frac{3u^{n+1} - 4u^n + u^{n-1}}{2\delta t} = \nabla \cdot (M(2u^n - u^{n-1})\nabla \mu^{n+1}),$$
$$\mu^{n+1} = -\epsilon^2 \Delta u^{n+1} + \ln(1 + 2u^n - u^{n-1}) - \ln(1 - 2u^n + u^{n-1}) - \theta_0(2u^n - u^{n-1})$$

and found that it blows up at $t \approx .025$ when $\|u^n\|_{l^\infty} > 1$ due to the singular potential. We then use the bound-preserving scheme (5.9)–(5.10) with $\delta = 0.01$ to compute up to $t = 0.1$ and plot in Figure 5 the evolution of $\max_{\boldsymbol{z} \in \Sigma_N} u^n(\boldsymbol{z})$ and $\min_{\boldsymbol{z} \in \Sigma_N} u^{n+1}(\boldsymbol{z})$ by the scheme (6.9) up to $t \approx .025$, and by the scheme (5.9)–(5.10) up to $t = 0.1$. In Figure 6, we plot the numerical solutions at various times which depict the coarsening process.

6.4. Fokker–Planck equation. As the final example, we consider the Fokker-Planck equation (5.11) with periodic boundary condition whose solution remains in $[0, 1]$ and is mass preserving. We present below simulations of (5.11) on the domain $(-2\pi, 2\pi)$ with the initial condition $u(x, 0) = -e^{-\frac{(x-1)^2}{0.4}}$ using three second-order schemes: a usual semi-implicit scheme

$$\frac{3u^{n+1} - 4u^n + u^{n-1}}{2\delta t} = \partial_x(x(2u^n - u^{n-1})(1 - 2u^n + u^{n-1}) + \partial_x u^{n+1}); \tag{6.10}$$



(a) By the bound-preserving scheme.

(b) By the semi-implicit scheme.

(c) Evolutions of minimum values

(d) λ at $t = 0.01$

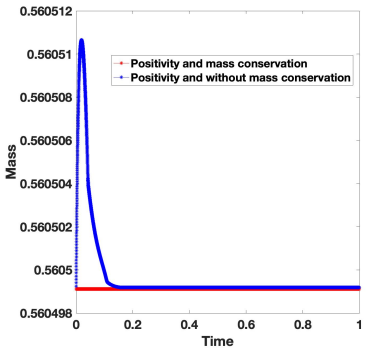
FIG. 7. (a)–(b): *Numerical solutions computed with* 32 *Fourier modes plotted on the* 256 *uniform grids using* (5.14)–(5.15) *and* (6.10). (c): *Evolutions of minimal values using* (5.14)–(5.15) *and* (6.10). (d): *Lagrange multiplier lambda at* $t = 0.01$ *using* (5.14)–(5.15).

the bound preserving scheme (5.14)–(5.15); and the mass conservative, bound preserving scheme (5.16)–(5.17).

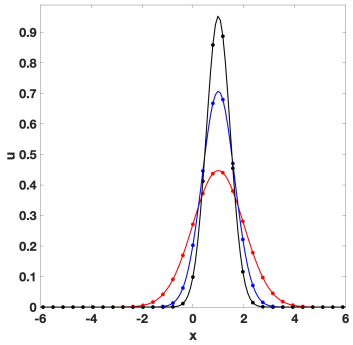
In Figure 7, we plot the numerical results using the semi-implicit scheme (6.10) and the bound preserving scheme (5.14)–(5.15) with 32 Fourier modes and $\delta t = 10^{-4}$. We observe that while the two numerical solutions look very similar, the minimum value by the semi-implicit scheme (6.10) does become negative in a short period at the beginning, while the numerical solutions by (5.14)–(5.15) remain in $[0, 1]$.

In Figure 8, we plot the numerical results using the bound preserving scheme (5.14)–(5.15) and the mass conservative, bound preserving scheme (5.16)–(5.17) with 32 Fourier modes and $\delta t = 10^{-4}$. We observe that (5.14)–(5.15) cannot preserve mass, while (5.16)–(5.17) preserves mass exactly. Only a few iterations are needed to compute the Lagrange multiplier ξ at each time step by using the mass conservative, bound-preserving scheme (5.16)–(5.17).

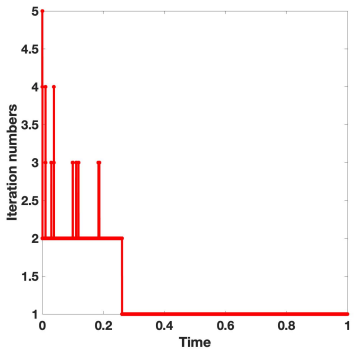
7. Concluding remarks. We constructed efficient and accurate bound and/or mass preserving schemes for a class of semilinear and quasi-linear parabolic equations using the Lagrange multiplier approach.



(a) Evolution of mass



(b) Solution profiles: u at $t = 0.01, 0.1, 0.4$.



(c) Iteration number

FIG. 8. (a): *Evolution of mass by* (5.14)–(5.15) *and* (5.16)–(5.17). (b): *Solution profiles by* (5.16)–(5.17). (c): *Iteration numbers for solving* ξ^{n+1} *at each time step of* (5.16)–(5.17).

First, we constructed a class of multistep IMEX schemes (2.5)–(2.6) for the semi-discrete problem (2.1) with a Lagrange multiplier to enforce bound preserving, which is an approximation to the original PDE (1.1). Hence, the scheme (2.5)–(2.6) is a kth-order approximation in time for both (2.1) and (1.1). In particular, (2.5)–(2.6) can be very useful if one is interested in the discrete problem (2.1) without a background PDE.

Then, we pointed out in Remark 2.1 that by dropping out the term $B_{k-1}(\lambda_h^n g'(u_h^n))$ in (2.5) and (2.6), we recover the usual cutoff scheme which is a kth-order approximation in time for (1.1) but only a first-order approximation in time for (2.1). Thus, our presentation provided an alternative interpretation of the cutoff approach and, moreover, allowed us to construct new cutoff IMEX schemes with mass conservation.

We also established some stability results involving norms with derivatives under a general setting, and derived optimal error estimates for a second-order bound preserving scheme with a hybrid spectral discretization in space.

Finally, we applied our approach to several typical PDEs which preserve bound and/or mass, and presented ample numerical results to validate our approach. The

approach presented in this paper is quite general and can be used to develop bound preserving schemes for other bound preserving PDEs such as the Keller-Segel equations [21].

REFERENCES

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

[2] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[3] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system.* I. *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[4] C. Canuto, M. Y. Hussaini, A. Quarteroni, and A. Thomas, Jr., *Spectral Methods in Fluid Dynamics*, Springer Science & Business Media, New York, 2012.

[5] J. A. Carrillo, J. Rosado, and F. Salvarani, 1*D nonlinear Fokker–Planck equations for fermions and bosons*, Appl. Math. Lett., 21 (2008), pp. 148–154.

[6] W. Chen, C. Wang, X. Wang, and S. M. Wise, *Positivity-preserving, energy stable numerical schemes for the Cahn-Hilliard equation with logarithmic potential*, J. Comput. Phys. X, 3 (2019), 100031, https://doi.org/10.1016/j.jcpx.2019.100031.

[7] Q. Cheng and J. Shen, *A new Lagrange multiplier approach for constructing structure preserving schemes,* I. *Positivity preserving*, Comput. Methods Appl. Mech. Engrg., 391 (2022), 114585.

[8] P. G. Ciarlet, *Discrete maximum principle for finite-difference operators*, Aequationes Math., 4 (1970), pp. 266–268.

[9] P. G. Ciarlet and P.-A. Raviart, *Maximum principle and uniform convergence for the finite element method*, Comput. Methods Appl. Mech. Engrg., 2 (1973), pp. 17–31.

[10] A. Debussche and L. Dettori, *On the Cahn-Hilliard equation with a logarithmic free energy*, Nonlinear Anal., 24 (1995), pp. 1491–1514.

[11] L. Dong, C. Wang, S. M. Wise, and Z. Zhang, *A positivity-preserving, energy stable scheme for a ternary Cahn-Hilliard system with the singular interfacial parameters*, J. Comput. Phys., 442 (2021), 110451, https://doi.org/10.1016/j.jcp.2021.110451.

[12] L. Dong, C. Wang, H. Zhang, and Z. Zhang, *A positivity-preserving, energy stable and convergent numerical scheme for the Cahn-Hilliard equation with a Flory-Huggins-deGennes energy*, Commun. Math. Sci., 17 (2019), pp. 921–939, https://doi.org/10.4310/CMS.2019.v17.n4.a3.

[13] J. Droniou and C. L. Potier, *Construction and convergence study of schemes preserving the elliptic local maximum principle*, SIAM J. Numer. Anal., 49 (2011), pp. 459–490.

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

[15] Q. Du, L. Ju, X. Li, and Z. Qiao, *Maximum bound principles for a class of semilinear parabolic equations and exponential time-differencing schemes*, SIAM Rev., 63 (2021), pp. 317–359, https://doi.org/10.1137/19M1243750.

[16] C. M. Elliott and H. Garcke, *On the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Math. Anal., 27 (1996), pp. 404–423.

[17] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer Science & Business Media, New York, 2007.

[18] S. Gottlieb and C. Wang, *Stability and convergence analysis of fully discrete Fourier collocation spectral method for* 3*-D viscous Burgers' equation*, J. Sci. Comput., 53 (2012), pp. 102–128.

[19] P. T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications*, Math. Program., 48 (1990), pp. 161–220.

[20] K. Ito and K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, SIAM, Philadelphia, 2008.

[21] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415, https://doi.org/10.1016/0022-5193(70)90092-5.

[22] M. Li, Y. Cheng, J. Shen, and X. Zhang, *A bound-preserving high order scheme for variable density incompressible Navier-Stokes equations*, J. Comput. Phys., 425 (2021), 109906, https://doi.org/10.1016/j.jcp.2020.109906.

[23] H.-L. Liao, T. Tang, and T. Zhou, *A second-order and nonuniform time-stepping maximum-principle preserving scheme for time-fractional Allen-Cahn equations*, J. Comput. Phys., 414 (2020), 109473.

[24] H. Liu and H. Yu, *Maximum-principle-satisfying third order discontinuous Galerkin schemes for Fokker–Planck equations*, SIAM J. Sci. Comput., 36 (2014), pp. A2296–A2325.

[25] C. Lu, W. Huang, and E. 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. Computat. Phys., 242 (2013), pp. 24–36, https://doi.org/10.1016/j.jcp.2013.01.052.

[26] H. Risken, *Fokker-Planck equation*, in The Fokker-Planck Equation, Springer, Cham, 1996, pp. 63–95.

[27] J. Shen, *Long time stability and convergence for fully discrete nonlinear Galerkin methods*, Appl. Anal., 38 (1990), pp. 201–229, https://doi.org/10.1080/00036819008839963.

[28] J. Shen, T. Tang, and L.-L. Wang, *Spectral Methods: Algorithms, Analysis and Applications*, Springer Science & Business Media, New York, 2011.

[29] Z. Sun, J. A. Carrillo, and C.-W. Shu, *A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials*, J. Comput. Phys., 352 (2018), pp. 76–104.

[30] J. J. van der Vegt, Y. Xia, and Y. Xu, *Positivity preserving limiters for time-implicit higher order accurate discontinuous Galerkin discretizations*, SIAM J. Sci. Comput., 41 (2019), pp. A2037–A2063.

[31] J. L. Vázquez, *The Porous Medium Equation: Mathematical Theory*, Oxford University Press, Oxford, UK, 2006.

[32] H. Zhou, Z. Sheng, and G. Yuan, *Physical-bound-preserving finite volume methods for the Nagumo equation on distorted meshes*, Comput. Math. Appl., 77 (2019), pp. 1055–1070, https://doi.org/10.1016/j.camwa.2018.10.038.