A NEW SCALAR AUXILIARY VARIABLE APPROACH FOR GENERAL DISSIPATIVE SYSTEMS

by

Fukeng Huang

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Mathematics

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Jie Shen, Chair

Department of Mathematics

Dr. Gregery Buzzard

Department of Mathematics

Dr. Suchuan Dong

Department of Mathematics

Dr. Jingwei Hu

Department of Mathematics

Dr. Peijun Li

Department of Mathematics

Approved by:

Dr. Plamen Stefanov

To my parents

ACKNOWLEDGMENTS

Throughout my graduate study at Purdue University, I have received a great deal of support from lots of people and I would like to express my deepest gratitude to them.

First and foremost, I would like to thank my advisor, Dr. Jie Shen, for his guidance throughout my study at Purdue. Without his lead to this research topic and his extensive experience, it is impossible for me to write down this thesis. His hard working and passion for research has always encouraged me in the academic path.

I would like to thank the professors I met at Purdue. I learnt a lot from them. In particular, I would like to thank my thesis committee members, Dr. Gregery Buzzard, Dr. Suchuan Dong, Dr. Jingwei Hu and Dr. Peijun Li for taking their time on reviewing my thesis and Dr. Zhiqiang Cai for fruitful discussions in my early research topic.

I would like to thank the current and previous group members at Purdue, I gained not only invaluable knowledge but also precious friendship from them. Among them, I want to particularly thank Dr. Zhiguo Yang and Dr. Ke Wu, I learnt lots of useful techniques and interesting research topics from them.

I would like to thank all my current and previous roommates, classmates and friends at Purdue. It is their friendship that provides me the 5-year peaceful and enjoyable Ph.D life at West Lafayette.

I would also like to thank my teachers, Dr. Yuesheng Xu and Dr. Haizhang Zhang, when I was an undergraduate student at SYSU for their encouragement on pursuing a doctorate degree and their recommendation on my application to graduate school.

Most of all, I would like to thank my family members in China for their endless support and concern, especially during this global pandemic period.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this thesis, we first propose a new scalar auxiliary variable (SAV) approach for general dissipative nonlinear systems. This new approach is half computational cost of the original SAV approach [1], can be extended to high order unconditionally energy stable backward differentiation formula (BDF) schemes and not restricted to the gradient flow structure. Rigorous error estimates for this new SAV approach are conducted for the Allen-Cahn and Cahn-Hilliard type equations from the BDF1 to the BDF5 schemes in a unified form. As an application of this new approach, we construct high order unconditionally stable, fully discrete schemes for the incompressible Navier-Stokes equation with periodic boundary condition. The corresponding error estimates for the fully discrete schemes are also reported. Secondly, by combining the new SAV approach with functional transformation, we propose a new method to construct high-order, linear, positivity/bound preserving and unconditionally energy stable schemes for general dissipative systems whose solutions are positivity/bound preserving. We apply this new method to second order equations: the Allen-Cahn equation with logarithm potential, the Poisson-Nernst-Planck equation and the Keller-Segel equations and fourth order equations: the thin film equation and the Cahn-Hilliard equation with logarithm potential. Ample numerical examples are provided to demonstrate the improved efficiency and accuracy of the proposed method.

1. INTRODUCTION

Dissipative physical systems are ubiquitous in the real world, due to the second law of thermodynamics. It is highly desirable for numerical methods targeted on such systems to preserve the discrete energy dissipation law. As such, many efforts to develop energy stable numerical methods have been devoted to this longstanding and active research area. These include, but not limited to, the average vector field (AVF) method [2], [3], the convex splitting method [4]–[7], the stabilization method [8], [9], the Lagrange multiplier method [10] and more recently, the invariant energy quadratization (IEQ) method [11], [12] and the scalar auxiliary variable (SAV) method [1]. Among them, the SAV method [1] is a particular powerful tool for the design of unconditionally energy-stable first- and second-order schemes for a large class of gradient flows.

Very recently, we proposed a new scalar auxiliary variable approach in [13], which provides several essential improvements on the original SAV approach in the sense that

- it only requires solving one linear system with constant coefficients at each time step, which is half computational cost of the original SAV approach;
- it does not require the nonlinear energy functional be bounded from below, and applicable to more general gradient flows, even to general dissipative systems;
- and more importantly, it is extendable to higher-order BDF type schemes with unconditional stability and amenable to higher-order adaptive time stepping.

In this thesis, the new SAV approach has been successfully applied to gradient flow problems: the Allen-Cahn equation and the Cahn-Hilliard equation and the general dissipative systems: the incompressible Navier-Stokes equations with periodic boundary condition. Based on the principal linear operator in the energy, we can prove a uniform bound for the numerical solutions in the new SAV approach, which allow us carrying out rigorous error analysis for the kth-order ($k = 1, 2, 3, 4, 5$) SAV schemes in a unified form by combining a stability results reported in [14].

As an important application of the new SAV approach, we can construct unconditionally energy stable, positivity/bound preserving numerical scheme for complex dissipative

systems. Many problems in sciences and engineering require their solutions to be positive or remain in a prescribed range, such as density, concentration, height, population, etc. Oftentimes, violation of the positivity or bound preserving in their numerical solutions renders the corresponding discrete problems ill posed, although the original problems are well posed. For these type of problems, it is of critical importance for the numerical schemes to be positivity or bound preserving. A particular class of such problems are the Wasserstein gradient flows which are gradient flows over spaces of probability distributions according to the topology defined by the Wasserstein metric [15], [16]. Important examples of Wasserstein gradient flows include the Poisson-Nernst-Planck (PNP) equations [17] and Keller-Segel equations [18], [19]. For these problems, in addition to positivity or bound preserving, it is also important for the numerical schemes to obey a discrete energy law. In this thesis, we construct highly efficient and accurate numerical schemes for the PNP and the Keller-Segel equations, which not only inherits all the advantages of the SAV approach for general dissipative systems, more importantly, it can preserve positivity/bound. Similar techniques are also applied to the fourth-order equation: the thin film equation and the Cahn-Hilliard equation with logarithm potential.

The rest of the thesis is organized as follows. In the second chapter, we describe the construction of the new SAV scheme for the gradient flows systems and introduce the time-adaptive strategy, followed by the extension to general dissipative systems and the rigorous error estimate of the new SAV scheme for the Allen-Cahn and the Cahn-Hilliard type equation in the third chapter. In the fourth chapter, we apply the new SAV approach to construct high order unconditionally stable fully discrete schemes for the incompressible Navier-Stokes equation with periodic boundary condition and the corresponding error estimates are also reported. In the fifth chapter, by combining the functional transformation and the new SAV approach, we construct positivity/bound preserving numerical schemes for the second order dissipative systems with application to the PNP equations and the Keller-Segel equations, followed by the similar techniques applied on the fourth-order dissipative systems in the sixth chapter. Some concluding remarks and future works are given in the last chapter.

2. NEW SAV APPROACH FOR GRADIENT FLOW

In this chapter, we describe the construction of the new SAV approach for gradient flow systems in general form and introduce the time-adaptive strategy. Numerical examples are provided to demonstrate the improved efficiency and accuracy of the proposed method. Most of the results in this chapter are extracted from [13].

2.1 Introduction of the SAV approach

In order to motivate our improvements, we briefly review below the original SAV approach for the general form of gradient flows:

$$\frac{\partial \phi}{\partial t} = -\mathcal{G}\mu, \tag{2.1}$$

where ϕ is the unknown function, $\mathcal{G}$ is a positive operator that gives rise to the dissipative mechanism of the system, e.g. $\mathcal{G} = \mathcal{I}$ in the L^2 gradient flow and $\mathcal{G} = -\Delta$ in the H^{-1} gradient flow, and μ is the so called chemical potential

$$\mu = \frac{\delta E_{tot}}{\delta \phi} = \mathcal{L}\phi + U(\phi), \tag{2.2}$$

with respect to the free energy

$$E_{tot}(\phi) = \frac{1}{2}(\phi, \mathcal{L}\phi) + E_1(\phi), \tag{2.3}$$

where $\mathcal{L}$ is a non-negative linear operator and $E_1(\phi)$ is a nonlinear functional. For the sake of conciseness, homogeneous Neumann or periodic boundary conditions are assumed throughout the thesis such that all boundary terms will vanish when integration by parts are performed.

The key for the SAV approach is to introduce a scalar variable $r(t)$ defined by $r(t) = \sqrt{E_1[\phi] + C_0}$ (C_0 is chosen such that $E_1[\phi] + C_0 > 0$) and its associated dynamical equation

$$\frac{dr}{dt} = \frac{1}{2\sqrt{E_1[\phi] + C_0}} \int_\Omega U[\phi] \frac{\partial \phi}{\partial t} d\Omega, \quad U[\phi] = \frac{\delta E_1}{\delta \phi}. \tag{2.4}$$

Then, a first-order SAV scheme with explicit treatment for all nonlinear terms is as follows:

$$\frac{\phi^{n+1} - \phi^n}{\Delta t} = -\mathcal{G}\mu^{n+1}, \tag{2.5a}$$

$$\mu^{n+1} = \mathcal{L}\phi^{n+1} + \frac{r^{n+1}}{\sqrt{E_1[\phi^n] + C_0}} U(\phi^n), \tag{2.5b}$$

$$\frac{r^{n+1} - r^n}{\Delta t} = \frac{1}{2\sqrt{E_1[\phi^n] + C_0}} \int U(\phi^n) \frac{\phi^{n+1} - \phi^n}{\Delta t} d\Omega. \tag{2.5c}$$

One can eliminate μ^{n+1} and r^{n+1} from the above coupled linear scheme to obtain a linear equation for ϕ only:

$$(I + \Delta t \mathcal{G}\mathcal{L})\phi^{n+1} = \phi^n - r^{n+1} \Delta t \mathcal{G}\Big(\frac{U[\phi^n]}{\sqrt{E_1[\phi^n] + C_0}} \Big). \tag{2.6}$$

Setting $\phi^{n+1} = \phi_1^{n+1} + r^{n+1}\phi_2^{n+1}$, we find that ϕ_1^{n+1} and ϕ_2^{n+1} are solutions of the following two linear equations with constant coefficients

$$(I + \Delta t \mathcal{G}\mathcal{L})\phi_1^{n+1} = \phi^n, \quad (I + \Delta t \mathcal{G}\mathcal{L})\phi_2^{n+1} = -\Delta t \mathcal{G}\Big(\frac{U[\phi^n]}{\sqrt{E_1[\phi^n] + C_0}} \Big). \tag{2.7}$$

Once ϕ_1^{n+1} and ϕ_2^{n+1} are known, we can determine r^{n+1} explicitly from (2.5c) (see more details in [1], [20], [21].

The above SAV approach enjoys the following remarkable properties:

- it requires only the solution of two linear systems with constant coefficients at each time step (efficiency);
- the first- and second-order SAV schemes are unconditionally energy-stable (stability);

- it only requires the nonlinear energy functional $E_1(\phi)$ be bounded from below, so it is applicable to a large class of gradient flows (flexibility).

Our new SAV approach in this thesis is to propose some new essential improvements on the original SAV approach to make it even more efficient and flexible in the sense that

- it only requires solving one linear system with constant coefficients at each time step;
- it does not require the nonlinear energy functional $E_1(\phi)$ be bounded from below, and applicable to more general gradient flows, even to general dissipative systems;
- and more importantly, it is extendable to higher-order BDF type schemes with unconditional stability and amenable to higher-order adaptive time stepping.

2.1.1 First order scheme

The SAV approach requires solving two linear equations at each time step. However, we observe that the two equations for ϕ_1^{n+1} and ϕ_2^{n+1} in (2.7) are different only on the right hand side. This motivates us to employ the auxiliary variable to control not only the nonlinear term $U(\phi^n)$, but also the explicit term ϕ^n, i.e., replace the temporal derivative in (2.5a) by $\frac{\phi^{n+1}-\frac{r^{n+1}}{\sqrt{E_1[\phi^n]+C_0}}\phi^n}{\Delta t}$. Consequently, this gives rise to the counterpart of equation (2.6)

$$(I+\Delta t\mathcal{G}\mathcal{L})\phi^{n+1} = r^{n+1}\Big(\frac{\phi^n}{\sqrt{E_1[\phi^n]+C_0}} - \Delta t\mathcal{G}\Big(\frac{U[\phi^n]}{\sqrt{E_1[\phi^n]+C_0}}\Big)\Big), \tag{2.8}$$

which requires only the solution of the *single* equation

$$(I+\Delta t\mathcal{G}\mathcal{L})\bar{\phi}^{n+1} = \frac{\phi^n}{\sqrt{E_1[\phi^n]+C_0}} - \Delta t\mathcal{G}\Big(\frac{U[\phi^n]}{\sqrt{E_1[\phi^n]+C_0}}\Big), \tag{2.9}$$

and then determine ϕ^{n+1} and r^{n+1} from $\phi^{n+1} = r^{n+1}\bar{\phi}^{n+1}$ and (2.5c).

However, such a naive treatment on the temporal derivative term could not lead to even first-order convergence due to the fact that $\frac{\phi^{n+1}-\frac{r^{n+1}}{\sqrt{E_1[\phi^n]+C_0}}\phi^n}{\Delta t}$ is no longer a first-

order approximation of $\frac{\partial \phi}{\partial t}\Big|^{n+1}$. Specifically, assume that r^{n+1} is approximated such that $\frac{r^{n+1}}{\sqrt{E_1[\phi^n]+C_0}} = 1 + \mathcal{O}(\Delta t)$, we have

$$\frac{\phi^{n+1} - \xi^{n+1}\phi^n}{\Delta t} = \frac{\phi^{n+1} - \phi^n}{\Delta t} + \frac{1 - \xi^{n+1}}{\Delta t}\phi^n = \frac{\partial \phi}{\partial t}\Big|^{n+1} + \mathcal{O}(\phi^n), \tag{2.10}$$

where $\xi^{n+1} := \frac{r^{n+1}}{\sqrt{E_1[\phi^n]+C_0}}$. Actually, in order to achieve first-order approximation of $\frac{\partial \phi}{\partial t}\Big|^{n+1}$ using the novel formula (2.10), ξ^{n+1} need to be approximated such that $\xi^{n+1} = 1+\mathcal{O}(\Delta t^k)$, $k \geq 2$.

This inspires us to replace the controlling factor $\xi^{n+1} = \frac{r^{n+1}}{\sqrt{E_1[\phi^n]+C_0}}$ by

$$\eta^{n+1} = 1 - (1 - \xi^{n+1})^2, \tag{2.11}$$

In this way, it is direct to observe that

$$\frac{\phi^{n+1} - \eta^{n+1}\phi^n}{\Delta t} = \frac{\phi^{n+1} - \phi^n}{\Delta t} + \frac{(1 - \xi^{n+1})^2}{\Delta t}\phi^n = \frac{\partial \phi}{\partial t}\Big|^{n+1} + \mathcal{O}(\Delta t). \tag{2.12}$$

More generally, for any $\eta_k^{n+1} = 1 - (1 - \xi^{n+1})^k$ and $\xi^{n+1} = 1 + \mathcal{O}(\Delta t^n)$, we have

$$\eta_k^{n+1} = 1 + \mathcal{O}(\Delta t^{kn}). \tag{2.13}$$

This observation makes it possible to achieve high-order convergence of ϕ^{n+1} with lower-order approximation for ξ^{n+1}.

The novel approximation (2.12) brings about significant issues in devising energy-stable schemes. In order to overcome this obstacle, we adopt some ideas from the recently proposed gPAV method in [22]. Specifically, in [22] it is suggested to (i) use a shifted total energy $E[\phi] = E_{tot}[\phi] + C_0$ instead of $E_1[\phi]$ in equation (2.3) to define the scalar auxiliary variable; (ii) use the energy balance equation of the gradient flow (2.1) instead of equation (2.4)

to construct the dynamical equation of the auxiliary variable, i.e. we adopt the following equation as the dynamical equation

$$\frac{dE[\phi]}{dt} = \int_\Omega \frac{\delta E}{\delta \phi}\frac{\partial \phi}{\partial t} d\Omega = -\Big(\frac{\delta E}{\delta \phi}, \mathcal{G}\frac{\delta E}{\delta \phi}\Big) = -(\mu, \mathcal{G}\mu) \leq 0; \tag{2.14}$$

and (iii) a delicate treatment of the dynamical equation to preserve the positiveness of the auxiliary variable in the discrete level. With the help of these intuitive thoughts, we are ready to construct new unconditionally energy-stable schemes, which require solving only one linear equation with constant coefficients at each time step. We define a shifted total energy by

$$E[\phi] = E_{tot}[\phi] + C_0 = \frac{1}{2}(\phi, \mathcal{L}\phi) + E_1(\phi) + C_0, \tag{2.15}$$

where C_0 is a chosen scalar such that $E[\phi] > 0$ for all ϕ. Note that for a physically meaningful system, the total energy E_{tot} is bounded from below, thus such a C_0 is always available. To construct our new SAV approach,we introduce a scalar auxiliary variable $r(t) := E[\phi]$, which satisfies the following dynamical equation

$$\frac{dr(t)}{dt} = \frac{dE[\phi]}{dt} = -(\mu, \mathcal{G}\mu). \tag{2.16}$$

Define $\xi(t) = \frac{r(t)}{E(t)}$ and note that $\xi(t) \equiv 1$ at the continuous level, we can reformulate the system (2.1)-(2.2) into the following equivalent form

$$\frac{\partial \phi}{\partial t} = -\mathcal{G}\mu, \tag{2.17a}$$

$$\mu = \mathcal{L}\phi + [\phi], \tag{2.17b}$$

$$\frac{dr}{dt} = -\xi(\mu, \mathcal{G}\mu), \tag{2.17c}$$

$$\eta = 1 - (1-\xi)^2. \tag{2.17d}$$

Our new first-order scheme for (2.17) is as follows:

$$\frac{\phi^{n+1}-\eta^{n+1}\phi^n}{\Delta t}=-\mathcal{G}\mu^{n+1}, \tag{2.18a}$$

$$\mu^{n+1}=\mathcal{L}\phi^{n+1}+\eta^{n+1}U(\phi^n), \tag{2.18b}$$

$$\frac{r^{n+1}-r^n}{\Delta t}=-\xi^{n+1}(\bar{\mu}^{n+1},\mathcal{G}\bar{\mu}^{n+1}), \tag{2.18c}$$

$$\xi^{n+1}=\frac{r^{n+1}}{E[\bar{\phi}^{n+1}]},\quad \eta^{n+1}=1-(1-\xi^{n+1})^2 \tag{2.18d}$$

where $\bar{\phi}^{n+1}$ and $\bar{\mu}^{n+1}$ are to be specified below, together with the initial conditions

$$\phi^0=\phi_0(x,t),\quad r^0=E[\phi_0]. \tag{2.19}$$

Combining equations (2.18a) and (2.18b) leads to the following linear equation

$$\Big(I+\Delta t\mathcal{G}\mathcal{L}\Big)\phi^{n+1}=\eta^{n+1}\Big(\phi^n-\Delta t\mathcal{G}\Big(U[\phi^n]\Big)\Big). \tag{2.20}$$

Setting

$$\phi^{n+1}=\eta^{n+1}\bar{\phi}^{n+1}, \tag{2.21}$$

in the above, we find that $\bar{\phi}^{n+1}$ is determined by

$$\Big(I+\Delta t\mathcal{G}\mathcal{L}\Big)\bar{\phi}^{n+1}=\phi^n-\Delta t\mathcal{G}\Big(U[\phi^n]\Big). \tag{2.22}$$

Once $\bar{\phi}^{n+1}$ is known, we define

$$\bar{\mu}^{n+1}=\mathcal{L}\bar{\phi}^{n+1}+U(\bar{\phi}^{n+1}). \tag{2.23}$$

Note that $\bar{\phi}^{n+1}$ can be viewed as an approximation of $\phi(t^{n+1})$ by a direct semi-implicit method. Thus, $\bar{\phi}^{n+1}$ and $\bar{\mu}^{n+1}$ are first-order approximations of ϕ^{n+1} and μ^{n+1}. Inserting equation (2.18d) into equation (2.18c) leads to

$$\xi^{n+1}=\frac{r^n}{E[\bar{\phi}^{n+1}]+\Delta t(\bar{\mu}^{n+1},\mathcal{G}\bar{\mu}^{n+1})}. \tag{2.24}$$

To summarize, the scheme (2.18a)-(2.18d) can be implemented as follows:

- solve $\bar{\phi}^{n+1}$ from (2.22);
- set $\bar{\mu}^{n+1} = \mathcal{L}\bar{\phi}^{n+1} + U(\bar{\phi}^{n+1})$ and compute ξ^{n+1} from (2.24);
- update $\phi^{n+1} = \eta^{n+1}\bar{\phi}^{n+1}$, and goto the next time step.

We observe that the above procedure only requires solving one linear equation with constant coefficients as in a standard semi-implicit scheme. As for the stability, we have the following result:

Theorem 2.1.1. *Given $r^n \geq 0$, we have $r^{n+1} \geq 0$, $\xi^{n+1} \geq 0$, and the scheme (2.18a)-(2.18d) is unconditionally energy stable in the sense that*

$$r^{n+1} - r^n = -\delta t \xi^{n+1}(\bar{\mu}^{n+1}, \mathcal{G}\bar{\mu}^{n+1}) \leq 0. \tag{2.25}$$

Furthermore, if $E(u) = \frac{1}{2}(\phi, \mathcal{L}\phi) + E_1(\phi)$ with $\mathcal{L}$ positive and $E_1(u)$ bounded from below, there exists $M_1 > 0$ such that

$$(\phi^n, \mathcal{L}\phi^n) \leq M_1^2, \ \forall n. \tag{2.26}$$

Proof. Given $r^n \geq 0$ and since $E[\bar{\phi}^{n+1}] > 0$, it follows from (2.18c) that

$$r^{n+1} = \frac{r^n}{1 + \delta t \frac{(\bar{\mu}^{n+1}, \mathcal{G}\bar{\mu}^{n+1})}{E[\bar{u}^{n+1}]}} \geq 0.$$

Then we derive from (2.18d) that $\xi^{n+1} \geq 0$ and obtain (2.25).

Denote $M := r^0 = E[\phi(\cdot, 0)]$, then (2.25) implies $r^n \leq M$, $\forall n$.

Without loss of generality, we can assume $E_1(\phi) > 1$ for all ϕ. It then follows from (2.18d) that

$$|\xi^{n+1}| = \frac{r^{n+1}}{E(\bar{\phi}^{n+1})} \leq \frac{2M}{(\mathcal{L}\bar{\phi}^{n+1}, \bar{\phi}^{n+1}) + 2}. \tag{2.27}$$

Since $\eta^{n+1} = 1 - (1 - \xi^{n+1})^2$, we have $\eta_k^{n+1} = \xi^{n+1}(2 - \xi^{n+1})$. Then, we derive from (5.74) that there exists $M_1 > 0$ such that

$$|\eta_k^{n+1}| = |\xi^{n+1}(2 - \xi^{n+1})| \leq \frac{M_1}{(\mathcal{L}\bar{\phi}^{n+1}, \bar{\phi}^{n+1}) + 2},$$

which, along with $\phi^{n+1} = \eta_k^{n+1}\bar{\phi}^{n+1}$, implies

$$\begin{aligned}(\mathcal{L}\phi^{n+1}, \phi^{n+1}) &= (\eta^{n+1})^2(\mathcal{L}\bar{\phi}^{n+1}, \bar{\phi}^{n+1}) \\ &\leq \Big(\frac{M_1}{(\mathcal{L}\bar{\phi}^{n+1}, \bar{\phi}^{n+1}) + 2}\Big)^2(\mathcal{L}\bar{\phi}^{n+1}, \bar{\phi}^{n+1}) \leq M_1^2.\end{aligned}$$

The proof is complete. □

2.1.2 High order scheme

The proposed method can be extended to construct high-order unconditionally energy-stable schemes when coupled with k-step backward differentiation formula (BDFk). The essential idea resides in that we can achieve overall kth order accuracy for ϕ by using just a first-order approximation for ξ, if we choose k such that $\eta^n = 1 - (1 - \xi^n)^k$ is a $(k+1)$th order approximation to 1. Actually, for any $\eta^n = 1 - (1 - \xi^n)^k$ and $\xi^{n+1} = 1 + \mathcal{O}(\Delta t^q)$, we have that

$$\eta^n = 1 + \mathcal{O}(\Delta t^{kq}). \tag{2.28}$$

We construct the kth order new SAV schemes based on the implicit-explicit BDF-k formulae in the following unified form:

Given ϕ^n, r^n, we compute $\bar{\phi}^{n+1}$, r^{n+1}, ξ^{n+1} and ϕ^{n+1} consecutively by

$$\frac{\alpha\phi^{n+1} - \eta_k^{n+1}\hat{\phi}^n}{\Delta t} = -\mathcal{G}\mu^{n+1}, \tag{2.29a}$$

$$\mu^{n+1} = \mathcal{L}\phi^{n+1} + \eta_k^{n+1}U(\phi^{*,n+1}), \tag{2.29b}$$

$$\frac{r^{n+1} - r^n}{\Delta t} = -\xi^{n+1}(\bar{\mu}^{n+1}, \mathcal{G}\bar{\mu}^{n+1}), \tag{2.29c}$$

$$\xi^{n+1} = \frac{r^{n+1}}{E[\bar{\phi}^{n+1}]}, \eta_k^{n+1} = 1 - (1 - \xi^{n+1})^{k+1}. \tag{2.29d}$$

Here, α, $\hat{\phi}^n$ and $\phi^{*,n+1}$ in equation (2.29) are defined as follows:

BDF2:

$$\alpha = \frac{3}{2}, \quad \hat{\phi}^n = 2\phi^n - \frac{1}{2}\phi^{n-1}, \quad \phi^{*,n+1} = 2\phi^n - \phi^{n-1}; \tag{2.30}$$

BDF3:

$$\alpha = \frac{11}{6}, \quad \hat{\phi}^n = 3\phi^n - \frac{3}{2}\phi^{n-1} + \frac{1}{3}\phi^{n-2}, \quad \phi^{*,n+1} = 3\phi^n - 3\phi^{n-1} + \phi^{n-2}; \tag{2.31}$$

BDF4:

$$\alpha = \frac{25}{12}, \quad \hat{\phi}^n = 4\phi^n - 3\phi^{n-1} + \frac{4}{3}\phi^{n-2} - \frac{1}{4}\phi^{n-3}, \quad \phi^{*,n+1} = 4\phi^n - 6\phi^{n-1} + 4\phi^{n-2} - \phi^{n-3}. \tag{2.32}$$

We can also use BDF5 and BDF6, but for the sake of brevity, we omit the detailed formula here.

Note that the solution algorithm for the new BDFk scheme is the same as the first-order scheme presented in Section 2.2. In each time step, it requires only the solution of one linear equation with constant coefficients, making the proposed method highly efficient. The new BDFk scheme also enjoys the same stability as the first-order scheme, namely, we can prove the following result using exactly the same procedure as in Section 2.2.

Theorem 2.1.2. *Given $r^n \geq 0$, we have $r^{n+1} \geq 0$, $\xi^{n+1} \geq 0$, and the scheme (2.29) is unconditionally energy stable in the sense that*

$$r^{n+1} - r^n = -\delta t \xi^{n+1}(\bar{\mu}^{n+1}, \mathcal{G}\bar{\mu}^{n+1}) \leq 0. \tag{2.33}$$

Furthermore, if $E(u) = \frac{1}{2}(\phi, \mathcal{L}\phi) + E_1(\phi)$ with $\mathcal{L}$ positive and $E_1(u)$ bounded from below, there exists $M_k > 0$ such that

$$(\phi^n, \mathcal{L}\phi^n) \leq M_k^2, \forall n. \tag{2.34}$$

Note that the stability is built into the scheme in (2.29c), independent of the actual scheme used in (2.29a) and (2.29b). In principle, we can use any linear multistep schemes in place of (2.29a) and (2.29b).

2.2 Time adaptive strategy

To achieve satisfactory numerical results in real simulations efficiently, it is supposed to use small time steps when the energy and solution of gradient flows vary drastically while

using relatively larger time steps when they vary slightly. However, for conditionally stable schemes, the allowable time step is often dictated by the stability constraint, not by accuracy. One salient feature of an unconditionally energy stable scheme is that it allows us to employ an appropriate adaptive time-stepping strategy [23]–[26]. Note that time-adaptivity strategy has been applied to first-order and second-order Crank-Nicolson SAV schemes in [1], [27], [28]. There are essential difficulties to apply adaptive time stepping to other schemes, particularly other second- or higher-order schemes. The main reason is that one does not have robust unconditionally stable second- or higher-order schemes with variable step sizes. In a recent work [23], the authors developed a stabilized second-order BDF scheme with variable step sizes that is stable if $\tau^{n+1} \leq \gamma^* \tau^n$ where $\{\tau^k\}$ are the time step sizes and $\gamma^* \approx 1.5$ for optimal convergence. To the best of our knowledge, there is no unconditionally stable third- or higher-order multistep scheme with variable step sizes.

However, as stated in Remark 2.1.2, we can replace (2.29a) and (2.29b) by any linear multistep schemes without affecting the stability provided by (2.33). In particular, we can replace them by the BDFk schemes (along with kth order extrapolation formula for nonlinear terms) with variable step sizes which we shall derive in the Appendix.

It is crucial to figure out a good indicator for adaptive time-stepping schemes, which suggests us to adjust the time step at reasonable moments. Some observations from abundant numerical experiments are as follows:

(i) to achieve an accurate result, ξ^{n+1} has to be a good approximation to 1;

(ii) ξ^{n+1} starts to deviate from 1 when oscillation or inaccuracy turns to happen, while adopting a smaller time step can avoid such situation.

These observations suggest us that $|1-\xi^{n+1}|$ is a suitable indicator for the time-adaptivity procedure. Roughly speaking, we should decrease the time step whenever $|1-\xi^{n+1}|$ is bigger than a given tolerance while we can maintain a relatively large time step whenever $|1-\xi^{n+1}|$ is small enough.

Based on these observations, we provide an adaptive time-stepping algorithm for the proposed BDFk SAV scheme. Given a default safety coefficient ρ, a reference tolerance tol,

the minimum time steps $\tau_{\min}$ and the maximum time steps $\tau_{\max}$, the adaptivity speed tunable constant r, we can update the time step size by the following formula

$$A_{dp}(\mathrm{e},\tau)=\rho\Big(\frac{tol}{\mathrm{e}}\Big)^{r}\tau. \tag{2.35}$$

The corresponding algorithm is summarized as follows:

Given: the previous time step τ_n.

step 1. compute ξ^n from previous step with time step τ_n;

step 2. calculate $\mathrm{e}_n=|1-\xi^n|$;

step 3. if $\mathrm{e}_n>tol$, **then**

recalculate time step $\tau_n\leftarrow\max\{\tau_{\min},\min\{A_{dp}(\mathrm{e}_n,\tau_n),\tau_{\max}\}\}$;

goto step 1

step 4. else update time step $\tau_{n+1}\leftarrow\max\{\tau_{\min},\min\{A_{dp}(\mathrm{e}_n,\tau_n),\tau_{\max}\}\}$;

step 5. end if

Remark

- It is suggested in [29] that when R^n has an obvious deviation from $E[\phi^n]$, a reset of $R^n=E[\phi^n]$ can be prescribed to improve the long time accuracy of the numerical scheme. This strategy could also be incorporated into the time-adaptivity algorithm. Specifically, R^n is reset to $E[\phi^n]$ when τ_{n+1}/τ_n is less than a threshold value.

- In real implementation, one can also combine the sav indicator with other traditional indicator, like the norm of the difference between two steps: $\|\phi^{n+1}-\phi^n\|$.

The proposed time-stepping strategy will be applied to the simulations of Allen-Cahn and Cahn-Hilliard equations in the next section to show its advantages to achieve high accuracy with low computational cost.

2.3 Numerical examples

In this section, we provide ample numerical examples to demonstrate the improved efficiency and accuracy. of the proposed method.

Let us consider two typical types of gradient flow, i.e. Allen-Cahn equation [30] and Cahn-Hilliard equation [31]. Given the free energy

$$E_{tot}[\phi] = \int_\Omega \frac{\lambda}{2}|\nabla\phi|^2 + E_1[\phi]d\Omega, \quad E_1[\phi] = \frac{\lambda}{4\eta^2}(1-\phi^2)^2, \tag{2.36}$$

the chemical potential in (2.2) takes the form

$$\mu = \frac{\delta E_{tot}}{\delta\phi} = -\lambda\nabla^2 + U[\phi], \quad U[\phi] = \frac{\lambda}{\eta^2}\phi(\phi^2-1). \tag{2.37}$$

Allen-Cahn equation corresponds to the L^2 gradient flow with $\mathcal{G} = m_0 I$, while Cahn-Hilliard equation corresponds the H^{-1} gradient flow with $\mathcal{G} = -m_0\Delta$ in equation (2.1).

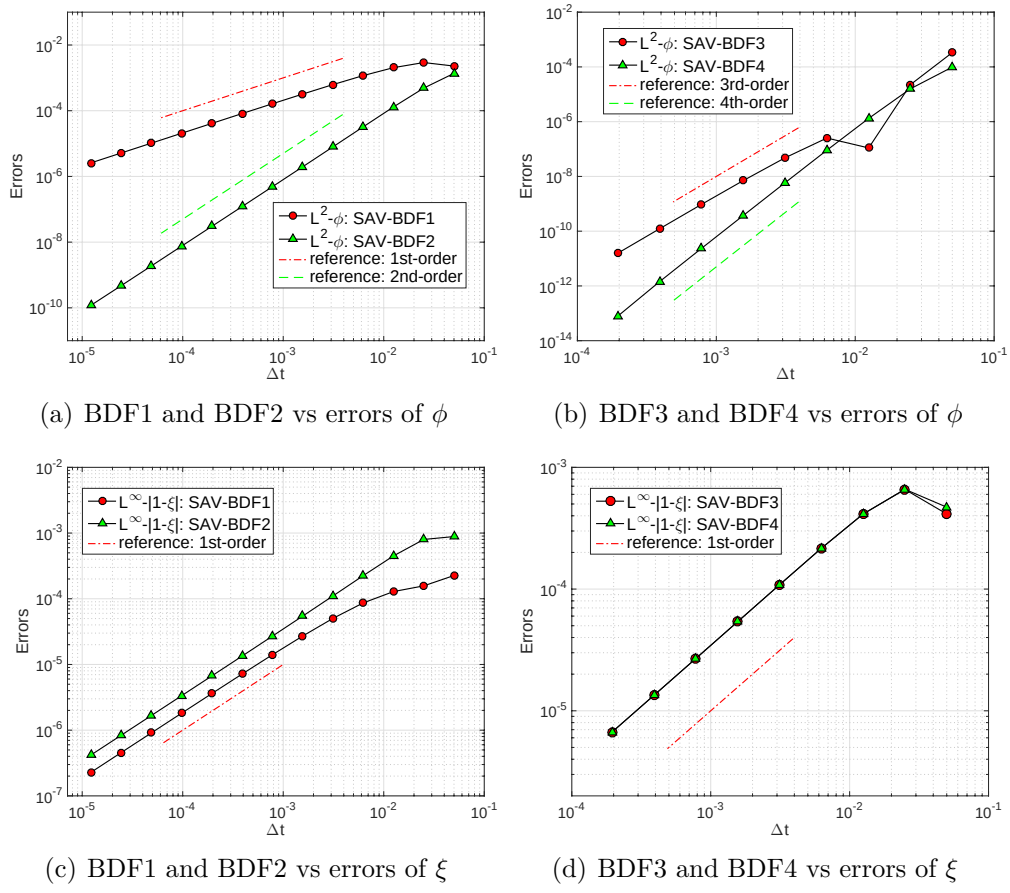
2.3.1 Accuracy test

Example 1. (Convergence rate of the new SAV/BDFk scheme for the Allen-Cahn and Cahn-Hilliard equations) Consider the Allen-Cahn and Cahn-Hilliard equations in the computational domain $\Omega = [0,2]\times[0,2]$ with a contrived exact solution

$$\phi(x,t) = \cos(\pi x)\cos(\pi y)\sin(t)/(1+10t^2), \tag{2.38}$$

and correspondingly, the external source term $f(x,t)$ satisfying $\phi_t = -\mathcal{G}\mu + f$. Fourier spectral method [32] for spatial discretization is employed throughout this section. N_x and N_y denote the number of Fourier collocation points along x and y axis, respectively. In the simulations, we set $(N_x, N_y) = (40, 40)$ with which the spatial discretization error is negligible compared with time discretization error. Other parameters are $\lambda = 0.01$, $m_0 = 0.01$, $\eta = 0.05$ and $C_0 = 0$. The algorithm for the new SAV/BDFk ($k = 1, 2, 3, 4$) is employed to numerically integrate the governing equations in time from $t = 0.1$ to $t = 1.1$. The L^2 errors of ϕ at $t = 1.1$ are plotted respectively in Figure 3.3 (a)-(b) for the Allen-Cahn equation and Figure 3.4 (a)-(b) for the Cahn-Hilliard equation, where we can observe the expected convergence rate of the field variable ϕ for all cases.

Recall that $r(t) = E[\phi]$ in the continuous level and the evolution equation of $R(t)$ is stemmed from this equation. The discretization (2.29c) of $R(t)$ leads to a first order approximation of $E[\phi]$. Consequently, $\xi^{n+1} = \frac{r^{n+1}}{E[\phi^{n+1}]}$ is a first order approximation of 1. We depict the L^∞ error of ξ^{n+1} to 1 for the Allen-Cahn and Cahn-Hillard equations in Figure 3.3 (c)-(d) and Figure 3.4 (c)-(d), respectively. As expected, it can be observed that ξ^{n+1} converges to 1 with a first order convergence rate for all cases. Moreover, this observation implies that ξ^{n+1} can serve as an indicator of the accuracy of the simulations. If the difference of ξ^{n+1} from 1 is small, then the simulation tends to be more accurate. Otherwise, when ξ^{n+1} deviates significantly from 1, the simulation is no longer accurate.



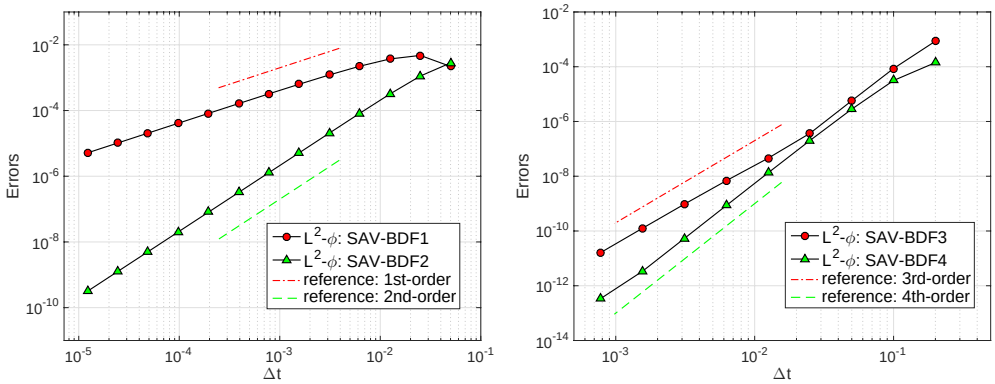
(a) BDF1 and BDF2 vs errors of ϕ

(b) BDF3 and BDF4 vs errors of ϕ

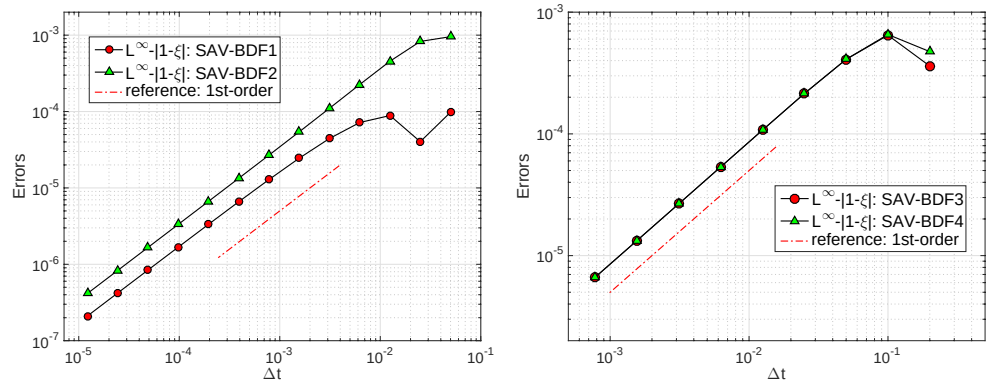
(c) BDF1 and BDF2 vs errors of ξ

(d) BDF3 and BDF4 vs errors of ξ

Figure 2.1. (*Example* 1.) Temporal convergence test for the Allen-Cahn equation using the new SAV/BDFk ($k = 1, 2, 3, 4$). (a)-(b) L^2 errors of ϕ as a function of Δt; (c)-(d) L^∞ erros of ξ as a function of Δt.



(a) BDF1 and BDF2 vs errors of ϕ (b) BDF3 and BDF4 vs errors of ϕ



(c) BDF1 and BDF2 vs errors of ξ (d) BDF3 and BDF4 vs errors of ξ

Figure 2.2. (*Example* 1.) Temporal convergence test for the Cahn-Hilliard equation using the new SAV/BDFk ($k = 1, 2, 3, 4$). (a)-(b) L^2 errors of ϕ as a function of Δt; (c)-(d) L^∞ erros of ξ as a function of Δt.

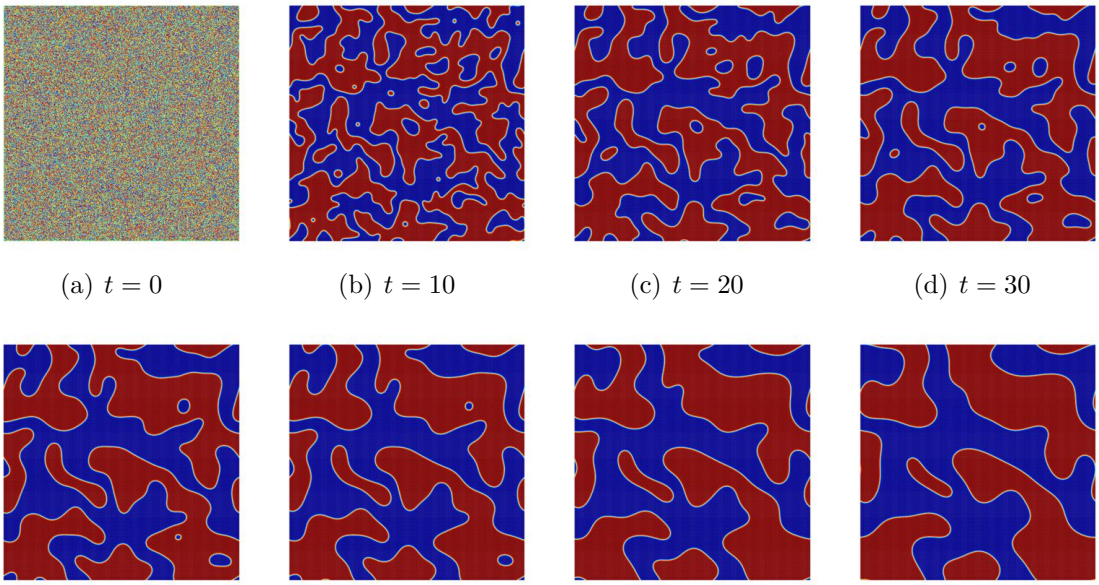
2.3.2 Allen-Cahn equation

Example 2. (Spinodal decomposition for the Allen-Cahn equation) Consider the spinodal decomposition of a homogeneous mixture into two coexisting phases governed by the Allen-Cahn equation as another test of the algorithms developed herein. The computational domain is $[0, 2] \times [0, 2]$ and the initial phase field is given by an uniformly distributed datas between $[-0.5, 0.5]$. We adopt $(N_x, N_y) = (512, 512)$, $\lambda = 1$, $m_0 = 10^{-4}$, $\eta = 0.005$ and $C_0 = 0$ in the forthcoming simulations. The new SAV/BDF2 scheme is employed to numerically integrate this problem.

In Figure 2.3, we depict a temporal sequence of snapshots of the interfaces formed between the two phases. Figure 2.4 shows the time histories of the total energy $E_{tot}[\phi]$ obtained by the current method and the original SAV method using various time steps $\Delta t = 10^{-1}$, 10^{-2}, 10^{-3}. A reference solution obtained with $\Delta t = 10^{-4}$ using the original SAV scheme is also included for comparison. It can be observed in Figure 2.4 (a) that all the energy history curves decrease dramatically at the beginning and level off gradually, indicating the stability of the proposed method. In Figure 2.4 (b), we zoom in the energy history curves at the region $t \in [0, 0.1]$. It shows that the results obtained by the current method and the original SAV method overlap with each other for the same time step size. This implies that the proposed method provides almost the same accuracy with the original SAV method, but with halved computational cost. It also indicates that to achieve acceptable accuracy, the step size should be chosen no bigger than $\Delta t = 10^{-3}$.

Example 3. (Application of adaptive time-stepping strategy to example 2) Next, we use the spinodal decomposition governed by the Allen-Cahn equation to demonstrate the performance of the time adaptivity. The setup is the same as Example 2. We choose $\rho = 0.9$, $tol = 10^{-3}$ and $r = 0.25$ in equation (2.35). The minimum time step is taken as $\tau_{\min} = 10^{-6}$, while the maximum time step is take as $\tau_{\max} = 10^{-3/k}$, $k = 2, 3, 4$, respectively, such that when $\tau_{\max}$ is employed, the errors for BDFk schemes ($k = 2, 3, 4$) are of the same level. The initial time step is taken as $\tau_{\min}$.

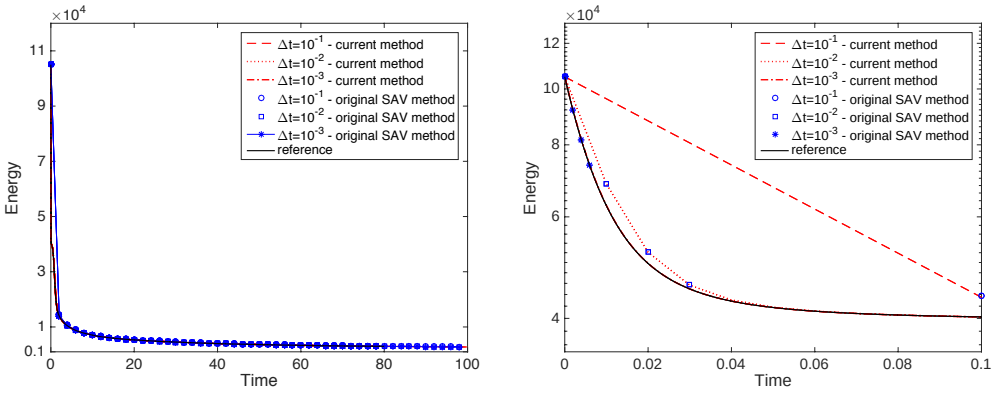
In Figure 2.5 (a), we depict the energy history curves obtained by BDFk ($k = 2, 3, 4$) schemes for long time simulations, and it can be seen that the curves essentially overlap



(a) $t = 0$ (b) $t = 10$ (c) $t = 20$ (d) $t = 30$

(e) $t = 40$ (f) $t = 50$ (g) $t = 70$ (h) $t = 100$

Figure 2.3. (*Example* 2.) Spinodal decomposition governed by the Allen-Cahn equation. The simulation is obtained with $\Delta t = 10^{-3}$.

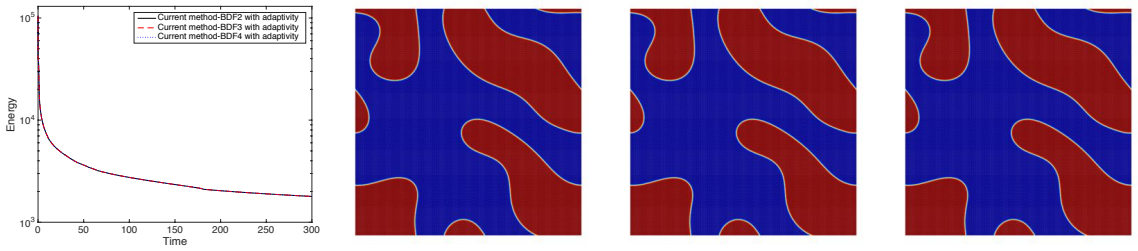


(a) Energy shown in $(0, 100)$ (b) Energy zoomed in $(0, 0.1)$

Figure 2.4. (*Example* 2) Time histories of $E_{tot}[\phi]$ for spinodal decomposition governed by the Allen-Cahn equation obtained using the current method and the original SAV method with $\Delta t = 10^{-1}, 10^{-2}, 10^{-3}$.

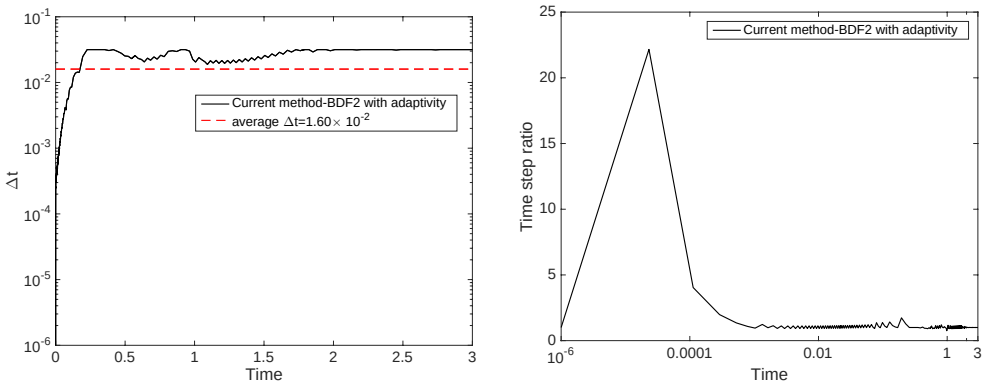
with each other. Correspondingly, the snapshot of the interfaces between these two phases at $t = 300$ are compared in Figure 2.5 (b)-(d) and no noticeable difference can be observed.

In order to demonstrate the efficiency of the time-adaptivity technique, we plot in Figure 2.6-2.8 the time history curves of the time steps and time step ratios between two successive steps in the time window $t \in (0, 3)$, where rapid changes of the phase field variable occur. It can be seen that the time steps gradually increases for all these three solvers and the average time steps in these period are 1.6×10^{-2}, 2.0×10^{-2} and 2.1×10^{-2}. Note that in Example 2, for BDF2 scheme with a fixed time step, it takes at least $\Delta t = 10^{-3}$ to obtain reasonable numerical results. We also record the total wall time of this simulation computed from $t = 0$ to $t = 300$ with BDF2 scheme using a fixed time step $\Delta t = 10^{-3}$, which is 19970.2 seconds. While equipped with the time-adaptivity technique, the total wall time for the BDF2-BDF4 schemes reduces drastically to 929.0 seconds, 318.9 seconds and 429.6 seconds, respectively. The speedups in these simulations are noticeable, compared with the solver without time adaptivity. In [23], it is point out that the variable step BDF2 scheme is stable if $\tau^{n+1} \leq \gamma^* \tau^n$ and $\gamma^* \approx 1.5$ for optimal convergence. While with the current method, we are not restricted by this constraint and the maximum time step ratios are observed to be as large as 22, 41, 41 for $k = 2, 3, 4$.



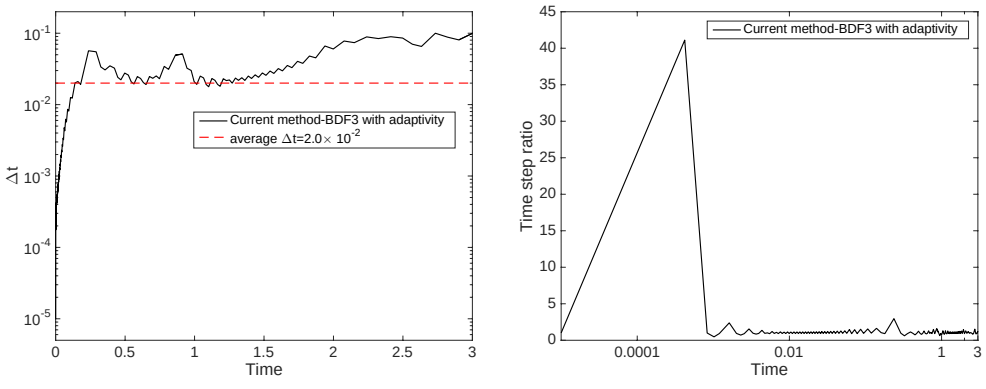
(a) Energy vs time (b) BDF2-adaptivity (c) BDF3-adaptivity (d) BDF4-adaptivity

Figure 2.5. (*Example* 3.) Spinoidal decomposition governed by the Allen-Cahn equation. Simulation results are obtained by the proposed SAV/BDFk ($k = 2, 3, 4$) scheme with adaptive time-stepping technique and the snapshots (b)-(d) of the interfaces between these two phases are depicted at $t = 300$.



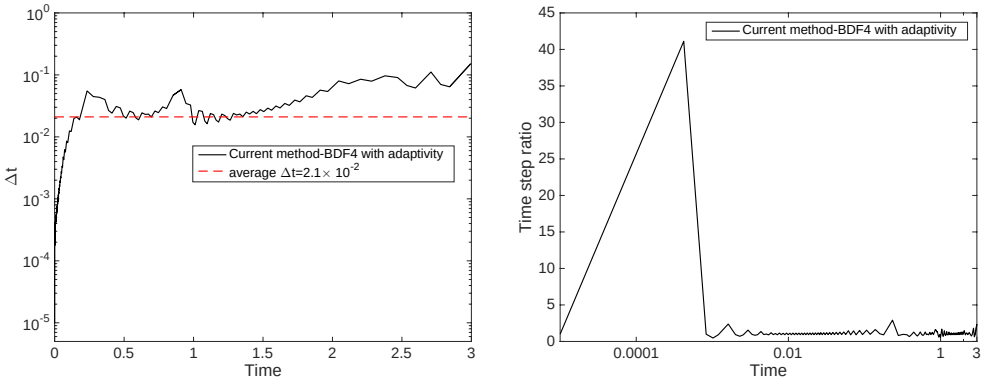
(a) History of Δt at $(0, 3)$ (b) History of τ^{n+1}/τ^n

Figure 2.6. (*Example* 3.) Time histories of time steps and time step ratios in the time window $t \in (0, 3)$ obtained by the BDF2 scheme with adaptive time-stepping technique. The average $\Delta t = 1.60 \times 10^{-2}$.



(a) History of Δt at $(0, 3)$ (b) History of τ^{n+1}/τ^n

Figure 2.7. (*Example* 3.) Time histories of time steps and time step ratios obtained by the BDF3 scheme with adaptive time-stepping technique. The average $\Delta t = 2.0 \times 10^{-2}$.



(a) History of Δt at $(0, 3)$ (b) History of τ^{n+1}/τ^n

Figure 2.8. (*Example* 3.) Time histories of time steps and time step ratios obtained by the BDF4 scheme with adaptive time-stepping technique. The average $\Delta t = 2.1 \times 10^{-2}$.

2.3.3 Cahn-Hilliard equation

Example 4. (Merging of an array of circles for the Cahn-Hilliard equation) We consider, as another test problem, the merging of a rectangular array of 9×9 circles governed by the Cahn-Hilliard equation. The computational domain is $[0, 2] \times [0, 2]$ and the initial phase field is given by

$$\phi_0(x, t) = 80 - \sum_{\mathrm{i}=1}^{9} \sum_{\mathrm{j}=1}^{9} \frac{\tanh\left(\sqrt{(x - x_{\mathrm{i}})^2 + (y - y_{\mathrm{j}})^2} - R_0\right)}{\sqrt{2}\eta}, \tag{2.39}$$

where $R_0 = 0.085$, $x_{\mathrm{i}} = 0.2 \times \mathrm{i}$ and $y_{\mathrm{j}} = 0.2 \times \mathrm{j}$ for $\mathrm{i}, \mathrm{j} = 1, 2, ..., 9$. We adopt $(N_x, N_y) = (512, 512)$, $\lambda = 1$, $m_0 = 10^{-6}$, $\eta = 0.01$ and $C_0 = 0$ in the simulations.

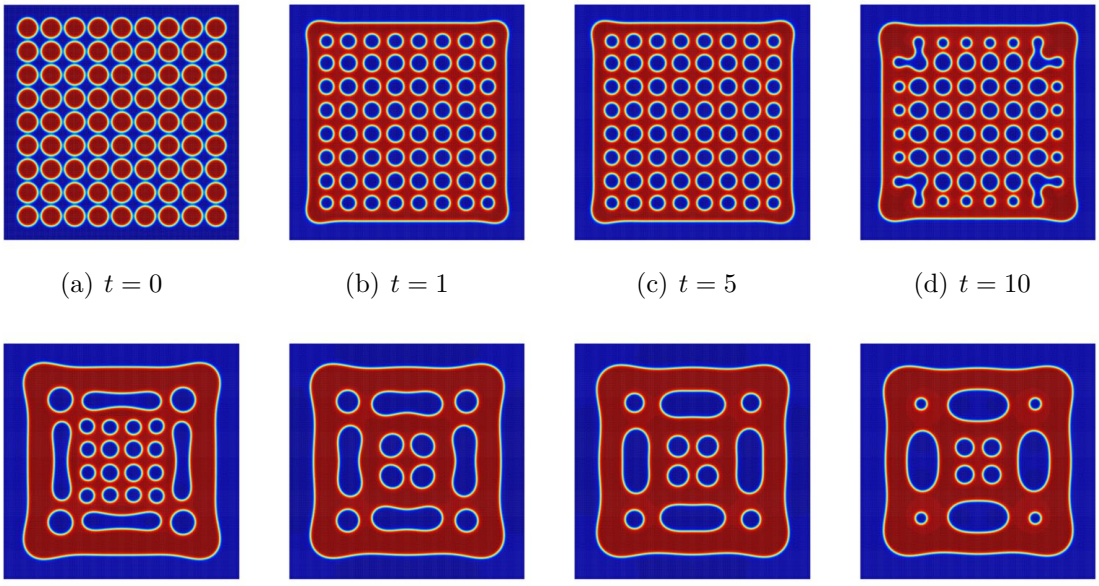
In Figure 2.9, we plot a temporal sequence of snapshots of the interfaces formed between the two phases using the BDF3 scheme with variable time steps using $\rho = 0.95$, $tol = 10^{-3}$, $r = 0.57$, $\tau_{\min} = 10^{-6}$ and $\tau_{\max} = 10^{-1}$ in equation (2.35).

In order to investigate the influence of the parameter *tol* to the accuracy and efficiency of the new SAV/BDFk ($k = 2, 3, 4$) schemes with time-adaptivity, we depict in Figure 2.10 the L^2-errors of ϕ as a function of the average of the time step sizes. These errors are obtained through comparing the numerical solutions with a reference solution computed by the new SAV/BDF2 scheme with a fixed small $\Delta t = 10^{-5}$ at $t = 1$. It can be seen that a better accuracy is achieved when we adopt a smaller *tol*.

Figure 2.11 is a demonstration of the stability of the proposed scheme. The time histories of the modified energy $r(t)$ obtained by large time step sizes $\Delta t = 2, 5$ are depicted. At these large time step sizes, we can no longer expect the results to be accurate. But it can be observed that the modified energy decays and remains to be positive for long time simulations.

2.3.4 Application to multiple SAV method

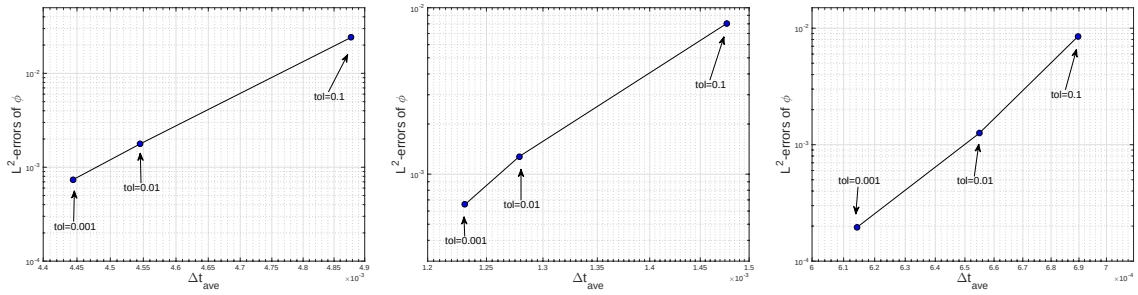
Example 5. (An improved multiple SAV method for one-component Bose-Einstein condensates) We consider the ground state solution of one-component Bose-Einstein condensates in two dimensions [29], [33], [34] as an example to show how our current method can reduce



(a) $t = 0$ (b) $t = 1$ (c) $t = 5$ (d) $t = 10$

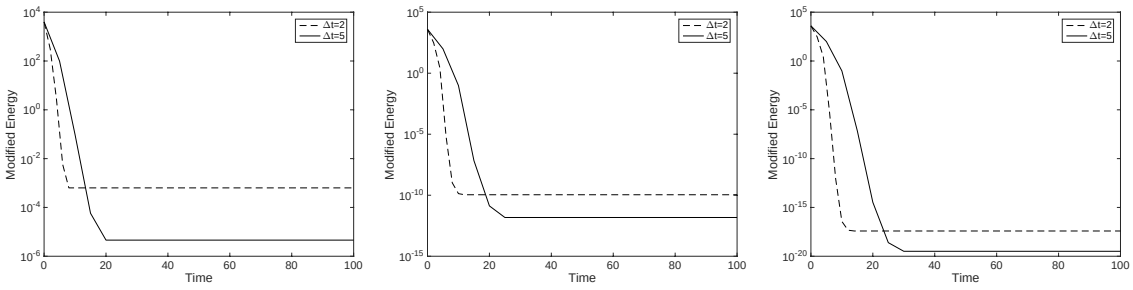
(e) $t = 20$ (f) $t = 30$ (g) $t = 40$ (h) $t = 50$

Figure 2.9. (*Example* 4.) Merging of an array of circles governed by the Cahn-Hilliard equation. Simulations are obtained with the proposed SAV/BDF3 schemes with time-adaptivity technique.



(a) BDF2 scheme with different *tol* (b) BDF3 scheme with different *tol* (c) BDF4 scheme with different *tol*

Figure 2.10. (*Example* 4.) L^2-errors of ϕ as a function of the average of Δt obtained by (a) the BDF2 scheme, (b) the BDF3 scheme, and (c) the BDF4 scheme, respectively. The time-adaptivity technique is employed with $tol = 10^{-1}, 10^{-2}, 10^{-3}$. The parameters (r, ρ) in equation (2.35) are set to be $(0.75, 0.85)$, $(0.57, 0.95)$, $(0.7, 0.85)$ for the BDF2-BDF4 schemes, respectively.



(a) BDF2 (b) BDF3 (c) BDF4

Figure 2.11. (*Example* 4.) Time histories of the modified energy $R(t)$ computed by (a) the BDF2 scheme, (b) the BDF3 scheme, and (c) the BDF4 scheme, using large time step sizes $\Delta t = 2, 5$.

the number of linear equations to be solved in each time step from *three* to *two*, comparing with the classical MSAV [29], [35] approach.

Similar with [29], we consider the penalized energy

$$E(\phi) = \frac{1}{2}(\phi, \mathcal{L}\phi) + \frac{1}{2}\int_\Omega F(|\phi|)^2)d\Omega + \frac{1}{4\epsilon}\Big(\int_\Omega |\phi|^2 d\Omega - 1\Big)^2. \tag{2.40}$$

Here, $F(\phi) = \frac{\beta}{2}\phi^2$, $\mathcal{L} = (-\frac{1}{2}\nabla^2 + V(x,y))\phi$ with $V(x,y) \geq 0$ and $\epsilon \ll 1$. Correspondingly, the governing equation for this gradient flow problem takes the form

$$\frac{\partial \phi}{\partial t} = -\frac{\delta E}{\delta \phi} = -\mathcal{L}\phi - F(|\phi|^2)\phi - \frac{1}{\epsilon}(\|\phi\|^2 - 1)\phi, \tag{2.41}$$

and subject to the constraints

$$\int_\Omega |\phi(x,t)|^2 d\Omega = 1, \tag{2.42a}$$

$$\lim_{|x|\to\infty} \phi(x,t) = 0. \tag{2.42b}$$

In our new MSAV approach, we introduce two scalar auxiliary variables

$$R_1(t) = E_1(\phi), \quad R_2(t) = E_2(\phi). \tag{2.43}$$

where

$$E_1 = \frac{1}{2}(\phi, \mathcal{L}\phi) + \frac{1}{2}\int_\Omega F(|\phi|)^2 d\Omega, \quad E_2 = \frac{1}{4\epsilon}\Big(\int_\Omega |\phi|^2 d\Omega - 1\Big)^2. \tag{2.44}$$

With the same spirit to improve the single SAV approach in Section 2, we employ one of the introduced auxiliary variables to control not only the nonlinear term, but also the

explicit linear term, and consequently, we are only required to solve two linear equations at each time step. The first order new MSAV scheme can be written as

$$\frac{\phi^{n+1}-\eta_1^{n+1}\phi^n}{\Delta t}=-\mu^{n+1}, \tag{2.45a}$$

$$\mu^{n+1}=\mathcal{L}\phi^{n+1}+\eta_1^{n+1}U_1(\phi^n)+\eta_2^{n+1}U_2(\phi^n), \tag{2.45b}$$

$$\frac{R_1^{n+1}-R_1^n}{\Delta t}=-\frac{R_1^{n+1}+R_2^{n+1}}{E_1[\bar{\phi}^{n+1}]+E_2[\bar{\phi}^{n+1}]}\Big(\mathcal{L}\bar{\phi}^{n+1}+U_1(\bar{\phi}^{n+1}),\bar{\mu}^{n+1}\Big), \tag{2.45c}$$

$$\frac{R_2^{n+1}-R_2^n}{\Delta t}=-\frac{R_1^{n+1}+R_2^{n+1}}{E_1[\bar{\phi}^{n+1}]+E_2[\bar{\phi}^{n+1}]}\Big(U_2(\bar{\phi}^{n+1}),\bar{\mu}^{n+1}\Big), \tag{2.45d}$$

$$\xi_1^{n+1}=\frac{R_1^{n+1}}{E_1[\bar{\phi}^{n+1}]},\quad \xi_2^{n+1}=\frac{R_2^{n+1}}{E_2[\bar{\phi}^{n+1}]}, \tag{2.45e}$$

$$\eta_1^{n+1}=1-(1-\xi^{n+1})^2,\quad \eta_2^{n+1}=1-(1-\xi^{n+1})^2, \tag{2.45f}$$

where $\bar{\mu}^{n+1}=\mathcal{L}\bar{\phi}^{n+1}+U_1(\bar{\phi}^{n+1})+U_2(\bar{\phi}^{n+1})$ and $\bar{\phi}^{n+1}$ is defined similar as the single SAV case and

$$U_1(\phi)=F(|\phi|^2)\phi,\quad U_2(\phi)=\frac{1}{\epsilon}(\|\phi\|^2-1)\phi. \tag{2.46}$$

The initial condition is chosen as

$$\phi_0(x,y)=\frac{(\gamma_x\gamma_y)^{1/4}}{\boldsymbol{\pi}^{1/2}}\mathrm{e}^{-(\gamma_x x^2+\gamma_y y^2)/2} \tag{2.47}$$

with two different potential functions:

Case 1. A harmonic oscillator potential

$$V(x,y)=\frac{1}{2}(\gamma_x^2x^2+\gamma_y^2y^2). \tag{2.48}$$

Case 2. A harmonic oscillator potential and a potential of a stirrer corresponding to a far-blue detuned Gaussian laser beam

$$V(x,y)=\frac{1}{2}(\gamma_x^2x^2+\gamma_y^2y^2)+\omega_0\mathrm{e}^{-\delta((x-r_0)^2+y^2)}. \tag{2.49}$$

The parameters are chosen as: $\gamma_x = 1$, $\gamma_y = 4$ and $\beta = 200$ in case 1 and $\gamma_x = 1$, $\gamma_y = 1$, $\omega_0 = 4$, $\delta = r_0 = 1$ and $\beta = 200$ in case 2. We then solve case 1 by spectral-Galerkin method on $\Omega_1 = [-8, 8] \times [-4, 4]$ and case 2 on $\Omega_2 = [-8, 8] \times [-8, 8]$. In both cases, we choose $(N_x, N_y) = (40, 40)$, $\epsilon = 10^{-4}$, time steps $\Delta t = 10^{-4}$ and impose the homogeneous Dirichlet boundary condition.

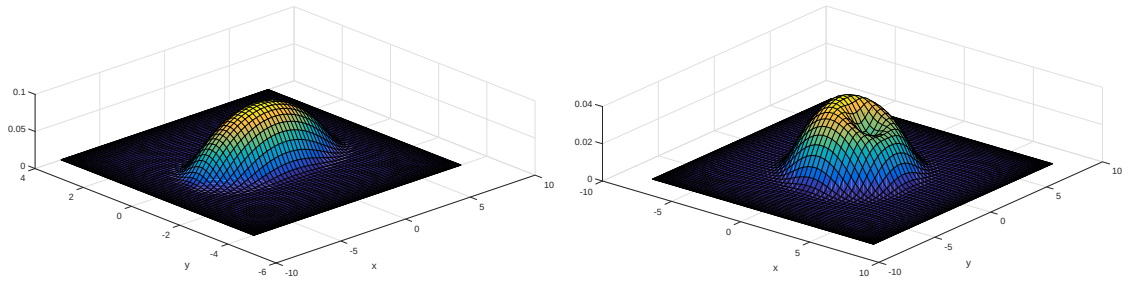
We plot the ground state solutions of both cases in Figure 2.12, and compare the chemical potential and the energy of the ground states with the results obtained by original MSAV method in [29] and by TSSP method in [34] in Tables 1 and 2, where we denote

$$x_{\text{rms}} = \|x\phi\|_{L^2(\Omega)}, \quad y_{\text{rms}} = \|y\phi\|_{L^2(\Omega)}, \tag{2.50}$$

and

$$\begin{aligned} \mu_\beta(\phi) &= \int_\Omega \Big(\frac{1}{2}|\nabla\phi(x)|^2 + V(x)|\phi(x)|^2 + \beta|\phi(x)|^4\Big)d\Omega \\ &= E_\beta(\phi) + \int_\Omega \frac{\beta}{2}|\phi(x)|^4 d\Omega. \end{aligned} \tag{2.51}$$

We observe from Tables 2.1 and 2.2 that the results by our new SAV schemes are consistent with the results obtained by using the original SAV and TSSP methods.



(a) case1 (b) case2

Figure 2.12. Ground state solutions of one-component Bose-Einstein condensates

Some remarks are in order:

Case 1 of BECS				
Scheme	$x_{\rm rms}$	$y_{\rm rms}$	E_β	μ_β
TSSP	2.2734	0.6074	11.1563	16.3377
MSAV	2.2812	0.6096	11.1560	16.3002
New MSAV	2.2710	0.6064	11.1621	16.2514

Table 2.1. Case 1: $\gamma_x = 1$, $\gamma_y = 4$ and $\beta = 200$

Case 2 of BECS				
Scheme	$x_{\rm rms}$	$y_{\rm rms}$	E_β	μ_β
TSSP	1.6951	1.7144	5.8507	8.3269
MSAV	1.6978	1.7169	5.8506	8.3189
New MSAV	1.6933	1.7124	5.8455	8.3202

Table 2.2. Case 2: $\gamma_x = 1$, $\gamma_y = 1$, $\omega_0 = 4$, $\delta = r_0 = 1$ and $\beta = 200$

1. Same as the single SAV approach mentioned in section 2, one can easily prove the scheme (2.45) is unconditionally energy stable in the sense that $R_1^n + R_2^n \geq 0$ for every n, and

$$(R_1^{n+1} + R_2^{n+1}) - (R_1^n + R_2^n) = -\Delta t \frac{R_1^{n+1} + R_2^{n+1}}{E_1[\bar{\phi}^{n+1}] + E_2[\bar{\phi}^{n+1}]} (\bar{\mu}^{n+1}, \bar{\mu}^{n+1}) \leq 0. \tag{2.52}$$

2. The new MSAV scheme (2.45) only requires solving *two* instead of *three* linear equations, compared with in the original MSAV scheme in [29]. In general, if we introduce K auxiliary variables, our new MSAV scheme requires only the solution of K linear equations, while the original SAV scheme requires to solve $K + 1$ linear equations. Hence, the new MSAV schemes is more efficient.

3. We can also construct higher-order unconditionally energy stable MSAV schemes based on the approach presented in Section 3.

2.4 Conclusion of this chapter

We presented in this paper several essential improvements over the original SAV approach, making our new SAV approach even more efficient, flexible and amenable to higher-order. More precisely, our new SAV approach enjoys the following additional advantages:

1. For the case with single SAV, our new method only requires solving one linear equation with constant coefficients, reducing half of the computational cost of the original SAV approach. In other words, the computational cost of the new SAV approach, being unconditionally energy stable, is essentially the same as that of the semi-implicit approach which is only conditionally stable. Furthermore, the new approach does not require the nonlinear part of the free energy to be bounded from below, making it more flexible than the original SAV approach.

2. While the original SAV approach only leads to first- and second-order unconditionally stable BDF type schemes, the new SAV approach allows us to construct higher-order unconditionally energy stable schemes with any multistep schemes. In particular, we

are able to construct, for the first time, unconditionally energy stable higher-order time adaptive schemes based on the BDFk scheme with variable step sizes.

3. For the cases where K SAVs are needed, our new method requires solving K linear equations with constant coefficients, as opposed to $K+1$ linear equations by the original MSAV approach.

The new SAV method is not limited to the realm of gradient flow. Although we only focus on gradient flow problems in the current chapter, the proposed method can be in principle extended to general dissipative systems. We will explore this issue in the following chapter

2.5 Appendix. BDF for variable time step sizes.

Given a successive of variable time steps

$$\tau^{n+1} = t^{n+1} - t^n, \quad \tau^n = t^n - t^{n-1}, \quad \tau^{n-1} = t^{n-1} - t^{n-2}, \quad \tau^{n-2} = t^{n-2} - t^{n-3}, ..., \quad (2.53)$$

the corresponding α, $\hat{\phi}^n$ and $\phi^{*,n+1}$ in equation (2.29) for BDFk scheme with kth order extrapolation can be derived by Taylor expansion. More precisely, if we set $\Delta t = \tau^{n+1}$ in (2.29a), and denote

$$\begin{aligned} x_1 &= -\tau^{n+1}, \\ x_2 &= -\tau^{n+1} - \tau^n, \\ x_3 &= -\tau^{n+1} - \tau^n - \tau^{n-1}, \\ x_4 &= -\tau^{n+1} - \tau^n - \tau^{n-1} - \tau^{n-2}, \\ &\dots . \end{aligned} \quad (2.54)$$

Then, for $k = 2, 3, 4$, the formulae are given below:

BDF2:

$$\alpha = \gamma^{-1}, \quad \hat{\phi}^n = \gamma^{-1}(a\phi^n + b\phi^{n-1}), \quad \phi^{*,n+1} = A\phi^n + B\phi^{n-1}, \quad (2.55)$$

where

$$\gamma = \frac{x_2}{x_1 + x_2}, \quad a = -\frac{x_2^2}{x_1^2 - x_2^2}, \quad b = \frac{x_1^2}{x_1^2 - x_2^2}, \quad A = -\frac{x_2}{x_1 - x_2}, \quad B = \frac{x_1}{x_1 - x_2}. \tag{2.56}$$

BDF3

$$\alpha = \gamma^{-1}, \quad \hat{\phi}^n = \gamma^{-1}(a\phi^n + b\phi^{n-1} + c\phi^{n-2}), \quad \phi^{*,n+1} = A\phi^n + B\phi^{n-1} + C\phi^{n-2}, \tag{2.57}$$

where

$$\begin{aligned} \gamma &= \frac{x_2\, x_3}{x_1x_2 + x_1x_3 + x_2x_3}, \quad a = \frac{x_2^2\, x_3^2}{(x_1 - x_2)(x_1 - x_3)(x_1x_2 + x_1x_3 + x_2x_3)}, \\ b &= -\frac{x_1^2\, x_3^2}{(x_1 - x_2)(x_2 - x_3)(x_1x_2 + x_1x_3 + x_2x_3)}, \\ c &= \frac{x_1^2\, x_2^2}{(x_1 - x_3)(x_2 - x_3)(x_1x_2 + x_1x_3 + x_2x_3)}, \end{aligned} \tag{2.58}$$

and

$$A = \frac{x_2\, x_3}{(x_1 - x_2)(x_1 - x_3)}, \quad B = -\frac{x_1\, x_3}{(x_1 - x_2)(x_2 - x_3)}, \quad C = \frac{x_1\, x_2}{(x_1 - x_3)(x_2 - x_3)}. \tag{2.59}$$

BDF4:

$$\alpha = \gamma^{-1}, \quad \hat{\phi}^n = \gamma^{-1}(a\phi^n + b\phi^{n-1} + c\phi^{n-2} + d\phi^{n-3}), \quad \phi^{*,n+1} = A\phi^n + B\phi^{n-1} + C\phi^{n-2} + D\phi^{n-3}, \tag{2.60}$$

where

$$
\begin{aligned}
\gamma &= \frac{x_2\, x_3\, x_4}{x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4}, \\
a &= -\frac{x_2^2\, x_3^2\, x_4^2}{(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4)}, \\
b &= \frac{x_1^2\, x_3^2\, x_4^2}{(x_1 - x_2)(x_2 - x_3)(x_2 - x_4)(x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4)}, \\
c &= -\frac{x_1^2\, x_2^2\, x_4^2}{(x_1 - x_3)(x_2 - x_3)(x_3 - x_4)(x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4)}, \\
d &= \frac{x_1^2\, x_2^2\, x_3^2}{(x_1 - x_4)(x_2 - x_4)(x_3 - x_4)(x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4)},
\end{aligned}
\tag{2.61}
$$

and

$$
\begin{aligned}
A &= -\frac{x_2\, x_3\, x_4}{(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)}, \quad B = \frac{x_1\, x_3\, x_4}{(x_1 - x_2)(x_2 - x_3)(x_2 - x_4)}, \\
C &= -\frac{x_1\, x_2\, x_4}{(x_1 - x_3)(x_2 - x_3)(x_3 - x_4)}, \quad D = \frac{x_1\, x_2\, x_3}{(x_1 - x_4)(x_2 - x_4)(x_3 - x_4)}.
\end{aligned}
\tag{2.62}
$$

The formulae for higher-order BDFk ($k \geq 5$) with variable step sizes can also be derived similarly. We omit the detail here for brevity.

3. ERROR ANALYSIS FOR THE NEW SAV APPROACH

In this chapter, we construct highly efficient implicit-explicit BDFk scalar auxiliary variable (SAV) schemes for general dissipative systems. We show that the scheme is unconditionally energy stable which leads to a uniform bound for the norm based on the principal linear operator in the energy. Based on this uniform bound, we carry out a rigorous error analysis for the kth-order ($k = 1, 2, 3, 4, 5$) SAV schemes in a unified form for the typical Allen-Cahn type and Cahn-Hilliard type equations. Most of the results in this chapter are extracted from [36].

3.1 Introduction

Analysis of standard semi-implicit schemes for gradient flows usually requires to assume global Lipschitz condition on the nonlinear term (see, for instance, [8], [37], [38]), the convergence of SAV schemes can be established without such assumption thanks to the unconditional energy stability. For examples, rigorous error analysis of the semi-discretized first order original SAV schemes for L^2 and H^{-1} gradient flows with minimum assumptions have been presented in [39], first- and second-order error estimates have been derived for a related semi-discretized gPAV scheme for the Cahn-Hilliard equation in [40], and error analysis of fully discretized SAV schemes with finite differences and finite-elements have also been established in [41] and [42]. On the other hand, error estimates for a Fourier-spectral SAV scheme for the phase-field crystal equation [43] and a MAC-SAV scheme for the Navier-Stokes equation [44] are established. Note that for the original SAV approach, unconditional energy stability can only be established for first- and second-order BDF schemes, although it has been shown in [45] (see also [46]) that the SAV approach coupled with extrapolated and linearized Runge-Kutta methods can achieve arbitrarily high order unconditionally energy stable with a modified energy for the Allen-Cahn and Cahn-Hilliard equations, but require solving coupled linear systems.

In this chapter, we apply the general ideas in [13] to construct a class of explicit-implicit BDFk schemes for general dissipative systems. In particular, we obtain a unconditional and uniform bound on the norm based on principal linear term in the energy functional of the

dissipative system. This bound is essential for the error analysis in this chapter. The main purpose of this chapter is to carry out a rigorous and unified error analysis for the new kth-order ($1 \le k \le 5$) SAV schemes which enjoy several remarkable advantages, including:

- only requires solving, in most common situations, one linear system with constant coefficients at each time step, which is half of the cost for the original SAV approach;
- applicable to general dissipative systems;
- higher-order BDFk SAV schemes are unconditionally stable and amenable to adaptive time stepping without restriction on time step size;
- rigorous error estimates can be established for BDFk ($1 \le k \le 5$) SAV schemes.

While the new schemes are applicable to a large class of dissipative systems, their error analysis is highly non-trivial. As a unified analysis for general dissipative systems will involve complicated assumptions and techniques that may obscure the clarity of presentation, we shall consider the error analysis for two classes of typical dissipative systems: Allen-Cahn type and Cahn-Hilliard type equations. The key ingredients are the uniform H^1 bound derived from the general stability result (see (3.13) in Theorem 4.3.1) and a stability result in [14] (see Lemma 4.4.3 below) for the BDFk ($1 \le k \le 5$) schemes. With a delicate induction argument, we are able to establish optimal error estimates in $L^\infty(0,T;H^2)$ norm for our implicit-explicit BDFk ($1 \le k \le 5$) SAV schemes for both Allen-Cahn type and Cahn-Hilliard type equations.

We use the following notations throughout the chapter. Let $\Omega \in \mathcal{R}^n$ ($n = 1, 2, 3$) be a bounded domain with sufficiently smooth boundary. We denote by $(\cdot,\cdot)$ and $\|\cdot\|$ the inner product and the norm in $L^2(\Omega)$, and by $H^s(\Omega)$ the usual Sobolev spaces with norm $\|\cdot\|_{H^s}$. Let V be a Banach space, we shall also use the standard notations $L^p(0,T;V)$ and $C([0,T];V)$. To simplify the notation, we often omit the spatial dependence in the notation for the exact solution u, namely we denote $u(x,t)$ by $u(t)$. We shall use C to denote a constant which can change from one step to another, but is independent of δt.

3.2 New SAV approach for general dissipative systems

In this section, we describe the new SAV schemes for dissipative systems, show that they are unconditionally energy stable with a modified energy and derive a uniform bound for the norm based on the principal linear term in the energy functional.

Consider the following class of dissipative systems

$$\frac{\partial u}{\partial t} + \mathcal{A}u + g(u) = 0, \tag{3.1}$$

where u is a scalar or vector function, $\mathcal{A}$ is a positive differential operator and $g(u)$ is a nonlinear operator possibly with lower-order derivatives. We assume that the above equation satisfies a dissipative energy law

$$\frac{d\tilde{E}(u)}{dt} = -\mathcal{K}(u), \tag{3.2}$$

where $\tilde{E}(u) > -C_0$ for all u is an energy functional, $\mathcal{K}(u) > 0$ for all $u \neq 0$.

The above class of dissipative systems include in particular gradient flows but also other dissipative systems which do not have the gradient structure, such as viscous Burgers equation, reaction-diffusion equations etc.

3.2.1 The new SAV schemes

The key for the SAV approach is to introduce a scalar auxiliary variable (SAV) to rewrite (3.1) as an expanded system, and to discretize the expanded system instead of the original (3.1). In this paper, we introduce the following new SAV approach inspired by the SAV schemes introduced in [13]

Setting $r(t) = E(u)(t) := \tilde{E}(u)(t) + C_0 > 0$, we rewrite the equation (3.1) with the energy law (3.2) as the following expanded system

$$\frac{\partial u}{\partial t} + \mathcal{A}u + g(u) = 0, \tag{3.3}$$

$$\frac{dE(u)}{dt} = -\frac{r(t)}{E(u)(t)}\mathcal{K}(u). \tag{3.4}$$

We construct the kth order new SAV schemes based on the implicit-explicit BDF-k formulae in the following unified form:

Given u^n, r^n, we compute $\bar{u}^{n+1}$, r^{n+1}, ξ^{n+1} and u^{n+1} consecutively by

$$\frac{\alpha_k \bar{u}^{n+1} - A_k(u^n)}{\delta t} + \mathcal{A}\bar{u}^{n+1} + g[B_k(\bar{u}^n)]) = 0, \tag{3.5a}$$

$$\frac{1}{\delta t}\Big(r^{n+1} - r^n\Big) = -\frac{r^{n+1}}{E(\bar{u}^{n+1})}\mathcal{K}(\bar{u}^{n+1}), \tag{3.5b}$$

$$\xi^{n+1} = \frac{r^{n+1}}{E(\bar{u}^{n+1})}, \tag{3.5c}$$

$$u^{n+1} = \eta_k^{n+1}\bar{u}^{n+1} \ \text{ with } \eta_k^{n+1} = 1 - (1-\xi^{n+1})^{k+1}. \tag{3.5d}$$

where α_k, the operators A_k and B_k ($k = 1, 2, 3, 4, 5$) are given by:

BDF1:

$$\alpha_1 = 1, \quad A_1(u^n) = u^n, \quad B_1(\bar{u}^n) = \bar{u}^n; \tag{3.6}$$

BDF2:

$$\alpha_2 = \frac{3}{2}, \quad A_2(u^n) = 2u^n - \frac{1}{2}u^{n-1}, \quad B_2(\bar{u}^n) = 2\bar{u}^n - \bar{u}^{n-1}; \tag{3.7}$$

BDF3:

$$\alpha_3 = \frac{11}{6}, \quad A_3(u^n) = 3u^n - \frac{3}{2}u^{n-1} + \frac{1}{3}u^{n-2}, \quad B_3(\bar{u}^n) = 3\bar{u}^n - 3\bar{u}^{n-1} + \bar{u}^{n-2}; \tag{3.8}$$

BDF4:

$$\alpha_4 = \frac{25}{12}, \ A_4(u^n) = 4u^n - 3u^{n-1} + \frac{4}{3}u^{n-2} - \frac{1}{4}u^{n-3}, \ B_4(\bar{u}^n) = 4\bar{u}^n - 6\bar{u}^{n-1} + 4\bar{u}^{n-2} - \bar{u}^{n-3}. \tag{3.9}$$

BDF5:

$$\begin{aligned} \alpha_5 = \frac{137}{60}, \quad A_5(u^n) &= 5u^n - 5u^{n-1} + \frac{10}{3}u^{n-2} - \frac{5}{4}u^{n-3} + \frac{1}{5}u^{n-4}, \\ B_5(\bar{u}^n) &= 5\bar{u}^n - 10\bar{u}^{n-1} + 10\bar{u}^{n-2} - 5\bar{u}^{n-3} + \bar{u}^{n-4}. \end{aligned} \tag{3.10}$$

Several remarks are in order:

- Initialization: the second-order scheme can be initialized with a first-order scheme for the first step, the kth-order scheme can be initialized with a $k-1$th-order Runge-Kutta method for the first $k-1$ steps.

- We observe from (3.5b) that r^{n+1} is a first order approximation to $E(u(\cdot, t_{n+1}))$ which implies that ξ^{n+1} is a first order approximation to 1.

- we observe from (3.5a) and (3.5d) that

$$\frac{\alpha_k u_{n+1} - \eta_k^{n+1} A_k(u^n)}{\delta t} + \mathcal{A}u^{n+1} + \eta_k^{n+1} g[B_k(\bar{u}^n)] = 0, \tag{3.11}$$

  which, along with (3.5d), implies that

$$\frac{\alpha_k u^{n+1} - A_k(u^n)}{\delta t} + \mathcal{A}u^{n+1} + g[B_k(u^n)] = O(\delta t^k).$$

  Hence, both u^{n+1} and $\bar{u}^{n+1}$ are formally kth order approximations for $u(\cdot, t^{n+1})$.

- The main difference of the above scheme from the scheme in [13] is the choice of η_k^{n+1}, which can be considered as a special case in [13]. However, as we show below, this choice allows us to obtain a uniform bound on $(\mathcal{L}u^n, u^n)$, which in turn plays a crucial role in the error analysis. Another slight difference is here we use $g[B_k(\bar{u}^n)]$ in (3.5a), which makes the error analysis slightly easier, while $g[B_k(u^n)]$ is used in [13]. Thanks to (3.5d), this does not affect the kth order accuracy nor unconditional energy stability.

- Since the energy stability is achieved through only (3.5b), we can replace (3.5a) by other types of explicit-implicit multistep schemes.

The above scheme can be efficiently implemented as follows:

1. Obtain $\bar{u}^{n+1}$ from (3.5a) by solving an equation of the form

$$(\frac{\alpha_k}{\delta t} I + \mathcal{A})\bar{u}^{n+1} = f^{n+1},$$

where f^{n+1} includes all known terms from previous time steps, and in most cases, this is a linear equation with constant coefficients;

2. With $\bar{u}^{n+1}$ known, determine r^{n+1} explicitly from (3.5b);

3. Compute ξ^{n+1}, η_k^{n+1} and u^{n+1} from (3.5d), goto the next step.

The main computational cost of this scheme is to solve (3.5a) once, while the main computational cost in the original SAV approach is to solve an equation similar to (3.5a) twice. So the cost of this scheme is about half of the original SAV approach while enjoying the same unconditional energy stability as we show below.

3.2.2 A stability result

We have the following results concerning the stability of the above schemes.

Theorem 3.2.1. *Given $r^n \geq 0$, we have $r^{n+1} \geq 0$, $\xi^{n+1} \geq 0$, and the scheme (3.5) for any k is unconditionally energy stable in the sense that*

$$r^{n+1} - r^n = -\delta t \xi^{n+1} \mathcal{K}(\bar{u}^{n+1}) \leq 0. \tag{3.12}$$

Furthermore, if $E(u) = \frac{1}{2}(\mathcal{L}u, u) + E_1(u)$ with $\mathcal{L}$ positive and $E_1(u)$ bounded from below, there exists $M_k > 0$ such that

$$(\mathcal{L}u^n, u^n) \leq M_k^2, \forall n. \tag{3.13}$$

Proof. Given $r^n \geq 0$ and since $E[\bar{u}^{n+1}] > 0$, it follows from (3.5b) that

$$r^{n+1} = \frac{r^n}{1 + \delta t \frac{(\mathcal{K}(\bar{u}^{n+1})}{E[\bar{u}^{n+1}]}} \geq 0.$$

Then we derive from (3.5c) that $\xi^{n+1} \geq 0$ and obtain (3.12).

Denote $M := r^0 = E[u(\cdot, 0)]$, then (3.12) implies $r^n \leq M$, $\forall n$.

Without loss of generality, we can assume $E_1(u) > 1$ for all u. It then follows from (3.5c) that

$$|\xi^{n+1}| = \frac{r^{n+1}}{E(\bar{u}^{n+1})} \leq \frac{2M}{(\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) + 2}. \tag{3.14}$$

Let $\eta_k^{n+1} = 1 - (1-\xi^{n+1})^{k+1}$, we have $\eta_k^{n+1} = \xi^{n+1} P_k(\xi^{n+1})$ with P_k being a polynomial of degree k. Then, we derive from (3.14) that there exists $M_k > 0$ such that

$$|\eta_k^{n+1}| = |\xi^{n+1} P_k(\xi^{n+1})| \le \frac{M_k}{(\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) + 2},$$

which, along with $u^{n+1} = \eta_k^{n+1}\bar{u}^{n+1}$, implies

$$\begin{aligned}(\mathcal{L}u^{n+1}, u^{n+1}) &= (\eta_k^{n+1})^2 (\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) \\ &\le \Big(\frac{M_k}{(\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) + 2}\Big)^2 (\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) \le M_k^2.\end{aligned}$$

The proof is complete. □

From the above proof, we observe that it is essential to introduce $\bar{u}^{n+1}$ and η_k^{n+1} in order to obtain (3.13), and that the bound constant M_k increases as k increases. So while we can replace $k+1$ in η_k^{n+1} by any larger integer without affecting the kth order accuracy, it is best to use the smallest possible integer, which is $k+1$ for kth order accuracy.

Note that the proof of (3.13) does not depend on specific form of (3.5a), so the result is also valid if we replace (3.5a) in the scheme by other implicit-explicit multistep schemes.

3.3 Error analysis for the Allen-Cahn type equation

While the stability results in Theorem 3.2.1 are valid for general dissipative systems, it is cumbersome to carry out error analysis with such generality. So to simplify the presentation, we shall carry out error analysis for two class of typical dissipative equations: Allen-Cahn type equation in this section and Cahn-Hilliard type equation in the next section.

We first recall the following important result. Based on Dahlquist's G-stability theory, Nevanlinna and Odeh [14] proved the following results for BDFk ($1 \le k \le 5$) schemes.

Theorem 3.3.1. *For $1 \le k \le 5$, there exist $0 \le \tau_k < 1$, a positive definite symmetric matrix $G = (g_{ij}) \in \mathcal{R}^{k,k}$ and real numbers $\delta_0, ..., \delta_k$ such that*

$$\Big(\alpha_k u^{n+1} - A_k(u^n), u^{n+1} - \tau_k u^n\Big) = \sum_{i,j=1}^{k} g_{ij}(u^{n+1+i-k}, u^{n+1+j-k}) - \sum_{i,j=1}^{k} g_{ij}(u^{n+i-k}, u^{n+j-k}) + \|\sum_{i=0}^{k} \delta_i u^{n+1+i-k}\|^2,$$

where the smallest possible values of τ_k are

$$\tau_1 = \tau_2 = 0, \quad \tau_3 = 0.0836, \quad \tau_4 = 0.2878, \quad \tau_5 = 0.8160,$$

and α_k, A_k are defined in (3.8)-(3.10).

The above result played a key role in proving the stability of high-order BDF schemes for nonlinear parabolic equations [47], and it plays an important role in our error analysis.

We shall also frequently use the following discrete Gronwall Lemma (see for example, [32], Lemma B.10).

Theorem 3.3.2. ***(Discrete Gronwall Lemma)*** *Let y^k, h^k, g^k, f^k be four nonnegative sequences satisfying*

$$y^n + \delta t \sum_{k=0}^{n} h^k \le B + \delta t \sum_{k=0}^{n} (g^k y^k + f^k) \quad \text{with} \quad \delta t \sum_{k=0}^{T/\delta t} g^k \le M, \, \forall\, 0 \le n \le T/\delta t.$$

We assume $\delta t\, g^k < 1$ and let $\sigma = \max_{0 \le k \le T/\delta t} (1 - \delta t g^k)^{-1}$. Then

$$y^n + \delta t \sum_{k=1}^{n} h^k \le \exp(\sigma M)(B + \delta t \sum_{k=0}^{n} f^k), \ \forall\, n \le T/\delta t.$$

Consider the Allen-Cahn type equation:

$$\frac{\partial u}{\partial t} - \Delta u + \lambda u - g(u) = 0 \quad (x,t) \in \Omega \times (0,T], \tag{3.15}$$

where Ω is an open bounded domain in R^d $(d = 1, 2, 3)$, with the initial condition $u(x, 0) = u^0(x)$, and boundary condition:

$$\text{periodic, } \textit{or } u|_{\partial\Omega} = 0, \textit{ or } \frac{\partial u}{\partial n}|_{\partial\Omega} = 0. \tag{3.16}$$

The above equation is a special case of (3.1) with $\mathcal{A} = -\Delta + \lambda I$, and satisfies the dissipation law (3.2) with $E(u) = \frac{1}{2}(\mathcal{L}u, u)+(G(u), 1)$ where $(\mathcal{L}u, u) = (\nabla u, \nabla u)+\lambda(u, u)$,G(u)=$\int^u g(v)dv$ and $\mathcal{K}(u) = (\frac{\delta E}{\delta u}, \frac{\delta E}{\delta u})$. We assume, without loss of generality,

$$\int_\Omega G(v)dx \geq \underline{C} > 0 \quad \forall v. \tag{3.17}$$

In particular, with $g(u) = (1 - u^2)u$ and $\lambda = 0$, the above equation becomes the celebrated Allen-Cahn equation [30].

We recall the following regularity result for (3.15) (see, for instance, [48]).

Theorem 3.3.3. *Assume $u^0 \in H^2(\Omega)$ and the following holds*

$$|g(x)| < C(|x|^p + 1), \quad p > 0 \textit{ arbitrary} \quad \text{if } d = 1, 2; \quad 0 < p < 4 \quad \text{if } d = 3. \tag{3.18}$$

Then for any $T > 0$, the problem (3.15) has a unique solution in the space

$$C([0, T]; H^2(\Omega)) \cap L^2(0, T; H^3(\Omega)).$$

We also recall a result (see Lemma 2.3 in [39]) which is useful to deal with the nonlinear term in (3.15).

Theorem 3.3.4. *Assume that $\|u\|_{H^1} \leq M$ and (3.18) holds. Then for any $u \in H^3$, there exist $0 \leq \sigma < 1$ and a constant $C(M)$ such that the following inequality holds:*

$$\|\nabla g(u)\|^2 \leq C(M)(1 + \|\nabla\Delta u\|^{2\sigma}).$$

We denote hereafter

$$t^n = n\,\delta t,\ \bar{\mathrm{e}}^n = \bar{u}^n - u(\cdot, t^n),\ \mathrm{e}^n = u^n - u(\cdot, t^n),\ s^n = r^n - r(t^n).$$

In the following, we carry out a unified error analysis for the first- to fifth- order SAV schemes described as in (3.5) with the coefficients defined in (3.8) - (3.10).

For (3.15), the kth-order version of (3.5a) and (3.11) read:

$$\frac{\alpha_k \bar{u}^{n+1} - A_k(u^n)}{\delta t} = \Delta \bar{u}^{n+1} - \lambda \bar{u}^{n+1} + g[B_k(\bar{u}^n)], \tag{3.19}$$

$$\frac{\alpha_k u^{n+1} - \eta_k^{n+1} A_k(u^n)}{\delta t} = \Delta u^{n+1} - \lambda u^{n+1} + \eta_k^{n+1} g[B_k(\bar{u}^n)], \tag{3.20}$$

where α_k, A_k, B_k defined in (3.8) - (3.10).

Theorem 3.3.5. *Given initial condition $\bar{u}^0 = u^0 = u(0)$, $r^0 = E[u^0]$. Let $\bar{u}^{n+1}$ and u^{n+1} be computed with the kth order scheme (3.5a)- (3.5d) $(1 \le k \le 5)$ for (3.15) with*

$$\eta_1^{n+1} = 1 - (1 - \xi^{n+1})^3, \quad \eta_k^{n+1} = 1 - (1 - \xi^{n+1})^{k+1}\ (k = 2, 3, 4, 5).$$

We assume (3.18) holds and

$$u^0 \in H^3,\ \frac{\partial^{\mathrm{j}} u}{\partial t^{\mathrm{j}}} \in L^2(0, T; H^1)\ 1 \le \mathrm{j} \le k+1.$$

Then for $n + 1 \le T/\delta t$ and $\delta t < \min\{\frac{1}{1+2C_0^{k+2}}, \frac{1-\tau_k}{3k}\}$, we have

$$\|\bar{\mathrm{e}}^{n+1}\|_{H^2},\ \|\mathrm{e}^{n+1}\|_{H^2} \le C\delta t^k,$$

where the constants C_0, C are dependent on T, Ω, the $k \times k$ matrix $G = (g_{\mathrm{ij}})$ in Theorem 3.3.1 and the exact solution u but are independent of δt and $0 \le \tau_k < 1$ is the constant in Theorem 3.3.1.

Proof. We assume that $\bar{u}^{\mathrm{i}}$ and u^{i} ($\mathrm{i}=1,\cdots,k-1$) are computed with a proper initialization procedure such that $\|\bar{u}^{\mathrm{i}}-u(t_{\mathrm{i}})\|_{H^2}=O(\delta t^k)$ and $\|u^{\mathrm{i}}-u(t_{\mathrm{i}})\|_{H^2}=O(\delta t^k)$ ($\mathrm{i}=1,\cdots,k-1$). To simplify the presentation, we set $\bar{u}^{\mathrm{i}}=u^{\mathrm{i}}=u(t_{\mathrm{i}})$ and $r^{\mathrm{i}}=E_1[u^{\mathrm{i}}]$ for $\mathrm{i}=1,\cdots,k-1$.

The main task is to prove

$$|1-\xi^q|\le C_0\,\delta t,\ \ \forall q\le T/\delta t. \tag{3.21}$$

where the constant C_0 is dependent on T, Ω and the exact solution u but is independent of δt, and will be defined in the proof process. Below we shall prove (3.21) by induction.

Under the assumption, (3.21) certainly holds for $q=0$. Now suppose we have

$$|1-\xi^q|\le C_0\,\delta t,\ \ \forall q\le m, \tag{3.22}$$

we shall prove below

$$|1-\xi^{m+1}|\le C_0\delta t. \tag{3.23}$$

We shall first consider $k=2,3,4,5$, and point out the necessary modifications for the case $k=1$ later.

Step 1: H^2 bound for u^n and $\bar{u}^n$ for all $n\le m$. For the kth-order schemes, it follows from Theorem 3.2.1 that

$$\|u^q\|_{H^1}\le M_k,\ \forall q\le T/\delta t. \tag{3.24}$$

Under assumption (3.22), if we choose δt small enough such that

$$\delta t\le\frac{1}{2C_0^{k+1}}, \tag{3.25}$$

we have

$$1-\frac{\delta t^k}{2}\le|\eta_k^q|\le 1+\frac{\delta t^k}{2},\ \ |1-\eta_k^q|\le\frac{\delta t^k}{2},\forall q\le m, \tag{3.26}$$

and

$$\|\bar{u}^q\|_{H^1}\le 2M_k,\ \forall q\le m,\ \forall\delta t\le 1. \tag{3.27}$$

Consider (3.20) in step q:

$$\frac{\alpha_k u^q - \eta_k^q A_k(u^{q-1})}{\delta t} = \Delta u^q - \lambda u^q + \eta_k^q g[B_k(\bar{u}^{q-1})]. \tag{3.28}$$

Thanks to Theorem 3.3.4 and (3.26), we have

$$\begin{aligned}
\|\nabla g[B_k(\bar{u}^{q-1})]\|^2 &\le C(M_k)(\|\nabla\Delta B_k(\bar{u}^{q-1})\|^{2\sigma}+1)\\
&\le \gamma_k\|\nabla\Delta B_k(\bar{u}^{q-1})\|^2 + C(M_k,\gamma_k)\\
&\le \gamma_k\|\nabla\Delta B_k(\frac{1}{\eta_k^{q-1}}u^{q-1})\|^2 + C(M_k,\gamma_k)\\
&\le 4\gamma_k\sum_{\mathrm{i}=1}^{k}\|\nabla\Delta u^{q-\mathrm{i}}\|^2 + C(M_k,\gamma_k),
\end{aligned}$$

where γ_k can be any positive constant. Taking the inner product of (3.28) with $\Delta^2 u^q - \tau_k\Delta^2 u^{q-1}$ and using the above inequality, it follows from Theorem 3.3.1 that there exist $0 \le \tau_k < 1$, a positive definite symmetric matrix $G = (g_{\mathrm{ij}}) \in \mathcal{R}^{k,k}$ and $\delta_0, ..., \delta_k$ that

$$\begin{aligned}
&\frac{1}{\delta t}\Big(\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\Delta u^{q+\mathrm{i}-k},\Delta u^{q+\mathrm{j}-k}) - \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\Delta u^{q-1+\mathrm{i}-k},\Delta u^{q-1+\mathrm{j}-k}) + \|\sum_{\mathrm{i}=0}^{k}\delta_{\mathrm{i}}\Delta u^{q+\mathrm{i}-k}\|^2\Big)\\
&\quad+\frac{1}{2}\|\nabla\Delta u^q\|^2 + \frac{\lambda}{2}\|\Delta u^q\|^2\\
&\quad\le \eta_k^q(g[B_k(\bar{u}^{q-1})],\Delta^2 u^q - \tau_k\Delta^2 u^{q-1}) + \frac{\tau_k}{2}\|\nabla\Delta u^{q-1}\|^2 + \frac{\lambda\tau_k}{2}\|\Delta u^{q-1}\|^2\\
&\quad+\frac{(\eta_k^q-1)}{\delta t}(A_k(u^{q-1}),\Delta^2 u^q - \tau_k\Delta^2 u^{q-1})\\
&\quad\le C(\epsilon_k)|\eta_k^q|\|\nabla g[B_k(\bar{u}^{q-1})]\|^2 + \epsilon_k|\eta_k^q|(\|\nabla\Delta u^q\|^2 + \|\nabla\Delta u^{q-1}\|^2) + \frac{\tau_k}{2}\|\nabla\Delta u^{q-1}\|^2\\
&\quad+\frac{\lambda\tau_k}{2}\|\Delta u^{q-1}\|^2 + \frac{|1-\eta_k^q|}{\delta t}\|\nabla A_k(u^{q-1})\|^2 + \frac{|1-\eta_k^q|}{\delta t}(\|\nabla\Delta u^q\|^2 + \|\nabla\Delta u^{q-1}\|^2)\\
&\quad\le C(M_k,\,\epsilon_k,\,\gamma_k) + (4C(\epsilon_k)|\eta_k^q|\gamma_k + \epsilon_k|\eta_k^q| + \frac{|1-\eta_k^q|}{\delta t})\sum_{\mathrm{i}=1}^{k}\|\nabla\Delta u^{q-\mathrm{i}}\|^2\\
&\quad+\frac{\tau_k}{2}\|\nabla\Delta u^{q-1}\|^2 + \frac{\lambda\tau_k}{2}\|\Delta u^{q-1}\|^2 + \frac{|1-\eta_k^q|}{\delta t}\|\nabla A_k(u^{q-1})\|^2,
\end{aligned} \tag{3.29}$$

where ϵ_k can be any positive constant. Note that $\tau_k < 1$, we can choose δt, ϵ_k and γ_k small enough such that

$$\delta t < \frac{1-\tau_k}{3k}, \quad \epsilon_k < \frac{1-\tau_k}{12k}, \quad \gamma_k < \frac{1-\tau_k}{48kC(\epsilon_k)}, \tag{3.30}$$

with the estimate in (3.26), we have

$$\begin{aligned} 4C(\epsilon_k)|\eta_k^q|\gamma_k + \epsilon_k|\eta_k^q| + \frac{1-\eta_k^q}{\delta t} &\le 8C(\epsilon_k)\gamma_k + 2\epsilon_k + \frac{\delta t^{k-1}}{2} \\ &\le \frac{1-\tau_k}{6k} + \frac{1-\tau_k}{6k} + \frac{1-\tau_k}{6k} \\ &\le \frac{1-\tau_k}{2k}. \end{aligned} \tag{3.31}$$

Then, taking the sum (3.29) for q from $k-1$ to n ($\le m$), we obtain

$$\begin{aligned} &\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\Delta u^{n+\mathrm{i}-k}, \Delta u^{n+\mathrm{j}-k}) \\ &\qquad \le C(M_k,\, \tau_k)T + C(u^0, ..., u^{k-1}) + C_{A_k} k\delta t^k \sum_{q=0}^{n-1} \|\nabla u^q\|^2 \\ &\qquad \le C(M_k,\, \tau_k)T + C(u^0, ..., u^{k-1}) + C_{A_k} k\delta t^{k-1} T M_k^2, \end{aligned}$$

where $C(M_k,\, \tau_k)$ is a constant only depends on M_k, τ_k, $C(u^0, ..., u^{k-1})$ only depends on $u^0, ..., u^{k-1}$ and C_{A_k} only depends on the coefficients in A_k. Since $G = (g_{\mathrm{ij}})$ is a positive definite symmetric matrix, we have

$$\begin{aligned} \lambda_G \|\Delta u^n\|^2 &\le \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\Delta u^{n+\mathrm{i}-k}, \Delta u^{n+\mathrm{j}-k}) \\ &\le C(M_k,\, \tau_k)T + C(u^0, ..., u^{k-1}) + C_{A_k} k\delta t^{k-1} T M_k^2. \end{aligned}$$

where $\lambda_G > 0$ is the minimum eigenvalue of $G = (g_{\mathrm{ij}})$. Together with (3.24), the above implies

$$\|u^n\|_{H^2} \le \frac{1}{\lambda_G}\sqrt{C(M_k,\, \tau_k)T + C(u^0, ..., u^{k-1}) + C_{A_k} kTM_k^2} := C_1,\ \forall \delta t < 1,\ n \le m. \tag{3.32}$$

Noting that

$$\|u^n\|_{H^2} = |\eta_k^n| \|\bar{u}^n\|_{H_2},$$

then (3.26) implies

$$\|\bar{u}^n\|_{H^2} \le 2C_1,\ \forall \delta t < 1,\ n \le m. \tag{3.33}$$

Step 2: estimate for $\|\bar{\mathrm{e}}^{n+1}\|_{H^2}$ for all $0 \le n \le m$. By Theorem 3.3.3 and (3.33) we can choose C large enough such that

$$\|u(t)\|_{H^2} \le C,\ \forall t \le T,\ \|\bar{u}^q\|_{H^2} \le C,\ \forall q \le m. \tag{3.34}$$

Since $H^2 \subset L^\infty$, without loss of generality, we can adjust C such that

$$|g^{(\mathrm{i})}[u(t)]|_{L^\infty} \le C,\ \forall t \le T;\ |g^{(\mathrm{i})}(\bar{u}^q)|_{L^\infty} \le C,\ \forall q \le m,\ \mathrm{i} = 0, 1, 2. \tag{3.35}$$

From (3.19), we can write down the error equation as

$$\alpha_k \bar{\mathrm{e}}^{n+1} - A_k(\bar{\mathrm{e}}^n) = A_k(u^n) - A_k(\bar{u}^n) + \delta t \Delta \bar{\mathrm{e}}^{n+1} - \delta t \lambda \bar{\mathrm{e}}^{n+1} + R_k^n + \delta t Q_k^n, \tag{3.36}$$

where $R_k^n,\ Q_k^n$ are given by

$$\begin{aligned} R_k^n &= -\alpha_k u(t^{n+1}) + A_k(u(t^n)) + \delta t u_t(t^{n+1}) \\ &= \sum_{\mathrm{i}=1}^{k} a_{\mathrm{i}} \int_{t^{n+1-\mathrm{i}}}^{t^{n+1}} (t^{n+1-\mathrm{i}} - s)^k \frac{\partial^{k+1} u}{\partial t^{k+1}}(s) ds, \end{aligned} \tag{3.37}$$

with a_{i} being some fixed and bounded constants determined by the truncation errors, and

$$Q_k^n = g[B_k(\bar{u}^n)] - g[u(t^{n+1})]. \tag{3.38}$$

For example, in the case $k = 3$, we have

$$R_3^n = -3 \int_{t^n}^{t^{n+1}} (t^n - s)^3 \frac{\partial^4 u}{\partial t^4}(s) ds + \frac{3}{2} \int_{t^{n-1}}^{t^{n+1}} (t^{n-1} - s)^3 \frac{\partial^4 u}{\partial t^4}(s) ds - \frac{1}{3} \int_{t^{n-2}}^{t^{n+1}} (t^{n-2} - s)^3 \frac{\partial^4 u}{\partial t^4}(s) ds.$$

Taking the inner product of (3.36) with $\bar{\mathrm{e}}^{n+1} - \tau_k \bar{\mathrm{e}}^n$, it follows from Theorem 3.3.1 that

$$\begin{aligned}
&\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\bar{\mathrm{e}}^{n+1+\mathrm{i}-k}, \bar{\mathrm{e}}^{n+1+\mathrm{j}-k}) - \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\bar{\mathrm{e}}^{n+\mathrm{i}-k}, \bar{\mathrm{e}}^{n+\mathrm{j}-k}) \\
&\quad + \|\sum_{\mathrm{i}=0}^{k} \delta_{\mathrm{i}} \bar{\mathrm{e}}^{n+1+\mathrm{i}-k}\|^2 + \delta t \|\nabla \bar{\mathrm{e}}^{n+1}\|^2 + \lambda \delta t \|\bar{\mathrm{e}}^{n+1}\|^2 \\
&\quad = (A_k(u^n) - A_k(\bar{u}^n), \bar{\mathrm{e}}^{n+1} - \tau_k \bar{\mathrm{e}}^n) - \delta t(\Delta \bar{\mathrm{e}}^{n+1}, \tau_k \bar{\mathrm{e}}^n) + \delta t \lambda(\bar{\mathrm{e}}^{n+1}, \tau_k \bar{\mathrm{e}}^n) \\
&\quad + (R_k^n, \bar{\mathrm{e}}^{n+1} - \tau_k \bar{\mathrm{e}}^n) + \delta t(Q_k^n, \bar{\mathrm{e}}^{n+1} - \tau_k \bar{\mathrm{e}}^n).
\end{aligned} \tag{3.39}$$

In the following, we bound the right hand side of (3.39). Note that

$$u^q = \eta_k^q \bar{u}^q, \quad |\eta_k^q - 1| \le C_0^{k+1}\, \delta t^{k+1}, \quad \forall q \le n.$$

hence

$$\begin{aligned}
|(A_k(u^n) - A_k(\bar{u}^n), \bar{\mathrm{e}}^{n+1} - \tau_k \bar{\mathrm{e}}^n)| &\le \frac{\|A_k(u^n) - A_k(\bar{u}^n)\|^2}{2\delta t} + \frac{\delta t}{2} \|\bar{\mathrm{e}}^{n+1} - \tau_k \bar{\mathrm{e}}^n\|^2 \\
&\le C C_0^{2k+2} \delta t^{2k+1} + \delta t \|\bar{\mathrm{e}}^{n+1}\|^2 + \delta t \|\bar{\mathrm{e}}^n\|^2.
\end{aligned} \tag{3.40}$$

It follows from (3.37) that

$$\|R_k^n\|^2 \le C \delta t^{2k+1} \int_{t^{n+1-k}}^{t^{n+1}} \|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\|^2 ds. \tag{3.41}$$

And we can bound Q_k^n based on (3.35) and (3.38) as

$$\begin{aligned}
|Q_k^n| &= \left| g[B_k(\bar{u}^n)] - g[B_k(u(t^n))] + g[B_k(u(t^n))] - g[u(t^{n+1})] \right| \\
&\le C|B_k(\bar{\mathrm{e}}^n)| + C|B_k(u(t^n)) - u(t^{n+1})| \\
&= C|B_k(\bar{\mathrm{e}}^n)| + C\left| \sum_{\mathrm{i}=1}^{k} b_{\mathrm{i}} \int_{t^{n+1-\mathrm{i}}}^{t^{n+1}} (t^{n+1-\mathrm{i}} - s)^{k-1} \frac{\partial^k u}{\partial t^k}(s) ds \right|,
\end{aligned} \tag{3.42}$$

where $b_{\rm i}$ are some fixed and bounded constants determined by the truncation error. For example, in the case $k=3$, we have

$$B_3(u(t^n))-u(t^{n+1})=-\frac{3}{2}\int_{t^n}^{t^{n+1}}(t^n-s)^2\frac{\partial^3 u}{\partial t^3}(s)ds+\frac{3}{2}\int_{t^{n-1}}^{t^{n+1}}(t^{n-1}-s)^2\frac{\partial^3 u}{\partial t^3}ds \\ -\frac{1}{2}\int_{t^{n-2}}^{t^{n+1}}(t^{n-2}-s)^2\frac{\partial^3 u}{\partial t^3}ds.$$

Therefore,

$$\begin{aligned}|\Big(R_k^n,\bar{\rm e}^{n+1}-\tau_k\bar{\rm e}^n\Big)|&\le\frac{1}{2\delta t}\|R_k^n\|^2+\delta t\|\bar{\rm e}^{n+1}\|^2+\delta t\|\bar{\rm e}^n\|^2,\\&\le\delta t\|\bar{\rm e}^{n+1}\|^2+\delta t\|\bar{\rm e}^n\|^2+C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}}\|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|^2ds.\end{aligned}\tag{3.43}$$

$$\delta t|\Big(Q_k^n,\bar{\rm e}^{n+1}-\tau_k\bar{\rm e}^n\Big)|\le C\delta t\Big(\|B_k(\bar{\rm e}^n)\|^2+\|\bar{\rm e}^{n+1}\|^2+\|\bar{\rm e}^n\|^2\Big)+C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}}\|\frac{\partial^k u}{\partial t^k}(s)\|^2ds.\tag{3.44}$$

Now, combining (3.39), (3.40), (3.43), (3.44), we arrive at

$$\begin{aligned}&\sum_{\rm i,j=1}^{k}g_{\rm ij}(\bar{\rm e}^{n+1+{\rm i}-k},\bar{\rm e}^{n+1+{\rm j}-k})-\sum_{\rm i,j=1}^{k}g_{\rm ij}(\bar{\rm e}^{n+{\rm i}-k},\bar{\rm e}^{n+{\rm j}-k})\\&\quad+\|\sum_{\rm i=0}^{k}\delta_{\rm i}\bar{\rm e}^{n+1+{\rm i}-k}\|^2+\frac{1}{2}\delta t\|\nabla\bar{\rm e}^{n+1}\|^2+\frac{\lambda}{2}\delta t\|\bar{\rm e}^{n+1}\|^2\\&\quad\le\frac{\tau_k}{2}\delta t\|\nabla\bar{\rm e}^n\|^2+\frac{\lambda\tau_k}{2}\delta t\|\bar{\rm e}^n\|^2+CC_0^{2k+2}\delta t^{2k+1}+C\delta t\sum_{\rm i=0}^{k}\|\bar{\rm e}^{n+1-{\rm i}}\|^2\\&\quad+C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}}(\|\frac{\partial^k u}{\partial t^k}(s)\|^2+\|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|^2)ds.\end{aligned}$$

Taking the sum of the above for n from $k-1$ to m, noting that $G=(g_{\rm ij})$ is a positive definite symmetric matrix with minimum eigenvalue λ_G, we obtain:

$$\begin{aligned}\lambda_G\|\bar{\rm e}^{m+1}\|^2&\le\sum_{\rm i,j=1}^{k}g_{\rm ij}(\bar{\rm e}^{m+1+{\rm i}-k},\bar{\rm e}^{m+1+{\rm j}-k})\\&\le C\delta t\sum_{q=0}^{m+1}\|\bar{\rm e}^q\|^2+C\delta t^{2k}\int_0^T(\|\frac{\partial^k u}{\partial t^k}(s)\|^2+\|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|^2+C_0^{2k+2})ds\end{aligned}\tag{3.45}$$

We can obtain similar inequalities for $\|\nabla\bar{\mathrm{e}}^m\|$ and $\|\Delta\bar{\mathrm{e}}^m\|$ by using essentially the same procedure. Indeed, taking the inner product of (3.36) with $-\Delta\bar{\mathrm{e}}^{n+1}+\tau_k\Delta\bar{\mathrm{e}}^n$, by using Theorem 3.3.1, we obtain

$$\begin{aligned}
&\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla\bar{\mathrm{e}}^{n+1+\mathrm{i}-k},\nabla\bar{\mathrm{e}}^{n+1+\mathrm{j}-k})-\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla\bar{\mathrm{e}}^{n+\mathrm{i}-k},\nabla\bar{\mathrm{e}}^{n+\mathrm{j}-k})+\|\sum_{\mathrm{i}=0}^{k}\delta_{\mathrm{i}}\nabla\bar{\mathrm{e}}^{n+1+\mathrm{i}-k}\|^2\\
&+\delta t\|\Delta\bar{\mathrm{e}}^{n+1}\|^2+\lambda\delta t\|\nabla\bar{\mathrm{e}}^{n+1}\|^2\\
&=(\nabla A_k(u^n)-\nabla A_k(\bar{u}^n),\nabla\bar{\mathrm{e}}^{n+1}-\tau_k\nabla\bar{\mathrm{e}}^n)+\delta t(\Delta\bar{\mathrm{e}}^{n+1},\tau_k\Delta\bar{\mathrm{e}}^n)-\delta t\lambda(\nabla\bar{\mathrm{e}}^{n+1},\tau_k\nabla\bar{\mathrm{e}}^n)\\
&+(R_k^n,-\Delta\bar{\mathrm{e}}^{n+1}+\tau_k\Delta\bar{\mathrm{e}}^n)+\delta t(Q_k^n,-\Delta\bar{\mathrm{e}}^{n+1}+\tau_k\Delta\bar{\mathrm{e}}^n).
\end{aligned}\tag{3.46}$$

Taking the sum of the above for n from $k-1$ to m, using Theorem 3.3.1, (3.41) and (3.42), we can obtain

$$\begin{aligned}
\lambda_G\|\nabla\bar{\mathrm{e}}^{m+1}\|^2&\le\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla\bar{\mathrm{e}}^{m+1+\mathrm{i}-k},\nabla\bar{\mathrm{e}}^{m+1+\mathrm{j}-k})\\
&\le C\delta t\sum_{q=0}^{m+1}\|\nabla\bar{\mathrm{e}}^q\|^2+C\delta t^{2k}\int_0^T(\|\frac{\partial^k u}{\partial t^k}(s)\|^2+\|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|^2+C_0^{2k+2})ds.
\end{aligned}\tag{3.47}$$

On the other hand, taking the inner product of (3.36) with $\Delta^2\bar{\mathrm{e}}^{n+1}-\tau_k\Delta^2\bar{\mathrm{e}}^n$, by using Theorem 3.3.1, we obtain

$$\begin{aligned}
&\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\Delta\bar{\mathrm{e}}^{n+1+\mathrm{i}-k},\Delta\bar{\mathrm{e}}^{n+1+\mathrm{j}-k})-\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\Delta\bar{\mathrm{e}}^{n+\mathrm{i}-k},\Delta\bar{\mathrm{e}}^{n+\mathrm{j}-k})+\|\sum_{\mathrm{i}=0}^{k}\delta_{\mathrm{i}}\Delta\bar{\mathrm{e}}^{n+1+\mathrm{i}-k}\|^2\\
&+\delta t\|\nabla\Delta\bar{\mathrm{e}}^{n+1}\|^2+\lambda\delta t\|\Delta\bar{\mathrm{e}}^{n+1}\|^2\\
&=(\Delta A_k(u^n)-\Delta A_k(\bar{u}^n),\Delta\bar{\mathrm{e}}^{n+1}-\tau_k\Delta\bar{\mathrm{e}}^n)+\delta t(\nabla\Delta\bar{\mathrm{e}}^{n+1},\tau_k\nabla\Delta\bar{\mathrm{e}}^n)-\delta t\lambda(\Delta\bar{\mathrm{e}}^{n+1},\tau_k\Delta\bar{\mathrm{e}}^n)\\
&+(\nabla R_k^n,-\nabla\Delta\bar{\mathrm{e}}^{n+1}+\tau_k\nabla\Delta\bar{\mathrm{e}}^n)+\delta t(\nabla Q_k^n,-\nabla\Delta\bar{\mathrm{e}}^{n+1}+\tau_k\nabla\Delta\bar{\mathrm{e}}^n).
\end{aligned}\tag{3.48}$$

Here, we need to pay attention to the terms with $\nabla\Delta\bar{\mathrm{e}}^{n+1}$ or $\nabla\Delta\bar{\mathrm{e}}^n$. Firstly, we have

$$|\delta t(\nabla\Delta\bar{\mathrm{e}}^{n+1},\tau_k\nabla\Delta\bar{\mathrm{e}}^n)|\le\frac{\delta t}{2}\|\nabla\Delta\bar{\mathrm{e}}^{n+1}\|^2+\frac{\tau_k^2\delta t}{2}\|\nabla\Delta\bar{\mathrm{e}}^n\|^2.$$

It follows from (3.37) and (3.38) that

$$|\nabla R_k^n|^2 =\leq C\delta t^{2k+1}\int_{t^{n+1-k}}^{t^{n+1}}|\nabla\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)|^2ds, \tag{3.49}$$

and

$$\begin{aligned}|\nabla Q_k^n| \leq C(|B_k(\bar{\mathrm{e}}^n)| + |\nabla B_k(\bar{\mathrm{e}}^n)|) + C\Big|\sum_{\mathrm{i}=1}^{k} b_{\mathrm{i}}\int_{t^{n+1-\mathrm{i}}}^{t^{n+1}}(t^{n+1-\mathrm{i}}-s)^{k-1}\frac{\partial^k u}{\partial t^k}(s)ds\Big| \\ + C\Big|\sum_{\mathrm{i}=1}^{k} b_{\mathrm{i}}\int_{t^{n+1-\mathrm{i}}}^{t^{n+1}}(t^{n+1-\mathrm{i}}-s)^{k-1}\nabla\frac{\partial^k u}{\partial t^k}(s)ds\Big|.\end{aligned} \tag{3.50}$$

Therefore,

$$\begin{aligned}|(\nabla R_k^n, -\nabla\Delta\bar{\mathrm{e}}^{n+1}+\tau_k\nabla\Delta\bar{\mathrm{e}}^n)| &\leq \frac{C}{\delta t}\|\nabla R_k^n\|^2 + \frac{\delta t(1-\tau_k^2)}{16}\|-\nabla\Delta\bar{\mathrm{e}}^{n+1}+\tau_k\nabla\Delta\bar{\mathrm{e}}^n\|^2 \\ &\leq C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}}\|\nabla\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|^2ds + \frac{\delta t(1-\tau_k^2)}{8}(\|\nabla\Delta\bar{\mathrm{e}}^{n+1}\|^2+\|\nabla\Delta\bar{\mathrm{e}}^n\|^2),\end{aligned}$$

and

$$\begin{aligned}\delta t|(\nabla Q_k^n, -\nabla\Delta\bar{\mathrm{e}}^{n+1}+\tau_k\nabla\Delta\bar{\mathrm{e}}^n)| &\leq C\delta t\|\nabla Q_k^n\|^2 + \frac{(1-\tau_k^2)\delta t}{16}\|-\nabla\Delta\bar{\mathrm{e}}^{n+1}+\tau_k\nabla\Delta\bar{\mathrm{e}}^n\|^2 \\ &\leq C\delta t\|B_k(\bar{\mathrm{e}}^n)\|_{H^1}^2 + C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}}\|\frac{\partial^k u}{\partial t^k}(s)\|_{H^1}^2ds \\ &+ \frac{(1-\tau_k^2)\delta t}{8}(\|\nabla\Delta\bar{\mathrm{e}}^{n+1}\|^2+\|\nabla\Delta\bar{\mathrm{e}}^n\|^2).\end{aligned}$$

We can bound other terms on the right hand side of (3.48) as before to arrive at

$$\begin{aligned}&\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\Delta\bar{\mathrm{e}}^{n+1+\mathrm{i}-k},\Delta\bar{\mathrm{e}}^{n+1+\mathrm{j}-k}) - \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\Delta\bar{\mathrm{e}}^{n+\mathrm{i}-k},\Delta\bar{\mathrm{e}}^{n+\mathrm{j}-k}) \\ &+ \frac{(1+\tau_k^2)\delta t}{4}\|\nabla\Delta\bar{\mathrm{e}}^{n+1}\|^2 + \frac{\lambda\delta t}{2}\|\Delta\bar{\mathrm{e}}^{n+1}\|^2 \\ &\leq C\delta t(\|B_k(\bar{\mathrm{e}}^n)\|_{H^1}^2 + \|\Delta\bar{\mathrm{e}}^{n+1}\|^2 + \|\Delta\bar{\mathrm{e}}^n\|^2) + \frac{(1+\tau_k^2)\delta t}{4}\|\nabla\Delta\bar{\mathrm{e}}^n\|^2 + \frac{\lambda\tau_k^2\delta t}{2}\|\Delta\bar{\mathrm{e}}^n\|^2 \\ &+ C\delta t^{2k}\int_{t^{n+1-k}}^{t^{n+1}}(\|\frac{\partial^k u}{\partial t^k}(s)\|_{H^1}^2 + \|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|_{H^1}^2 + C_0^{2k+2})ds.\end{aligned}$$

Then, taking the sum of the above for n from $k-1$ to m, we obtain

$$\begin{aligned}\lambda_G\|\Delta\bar{\mathrm{e}}^{m+1}\|^2 &\le \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\Delta\bar{\mathrm{e}}^{m+1+\mathrm{i}-k}, \Delta\bar{\mathrm{e}}^{m+1+\mathrm{j}-k}) \\ &\le C\delta t\sum_{q=0}^{m+1}\|\bar{\mathrm{e}}^q\|_{H^2}^2 + C\delta t^{2k}\int_0^T(\|\frac{\partial^k u}{\partial t^k}(s)\|_{H^1}^2 + \|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|_{H^1}^2 + C_0^{2k+2})ds.\end{aligned} \tag{3.51}$$

Summing up (3.45), (3.47) and (3.51), we obtain

$$\lambda_G\|\bar{\mathrm{e}}^{m+1}\|_{H^2}^2 \le C\delta t\sum_{q=0}^{m+1}\|\bar{\mathrm{e}}^q\|_{H^2}^2 + C\delta t^{2k}\int_0^T(\|\frac{\partial^k u}{\partial t^k}(s)\|_{H^1}^2 + \|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|_{H^1}^2 + C_0^{2k+2})ds \tag{3.52}$$

Finally, we can obtain the following H^2 estimate for $\bar{\mathrm{e}}^{m+1}$ by applying the discrete Gronwall lemma to (3.52) with $\delta t < \frac{1}{2C}$:

$$\begin{aligned}\|\bar{\mathrm{e}}^{m+1}\|_{H^2}^2 &\le C\exp((1-\delta tC)^{-1}))\delta t^{2k}\int_0^T(\|\frac{\partial^k u}{\partial t^k}(s)\|_{H^1}^2 + \|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|_{H^1}^2 + C_0^{2k+2})ds \\ &\le C_2(1+C_0^{2k+2})\delta t^{2k} \quad \forall 0\le n\le m.\end{aligned} \tag{3.53}$$

where C_2 is independent of δt and C_0, can be defined as

$$C_2 := C\exp(2)\max\Big(\int_0^T(\|\frac{\partial^k u}{\partial t^k}(s)\|_{H^1}^2 + \|\frac{\partial^{k+1}u}{\partial t^{k+1}}(s)\|_{H^1}^2)ds, 1\Big). \tag{3.54}$$

then $\delta t < \frac{1}{2C}$ can be guaranteed by

$$\delta t < \frac{1}{C_2}. \tag{3.55}$$

In particular, (3.53) implies

$$\|\bar{\mathrm{e}}^{n+1}\|_{H^2} \le \sqrt{C_2(1+C_0^{2k+2})}\delta t^k, \quad \forall 0\le n\le m. \tag{3.56}$$

Combining (3.34) and (3.56), under the condition (3.25) we obtain

$$\|\bar{u}^{n+1}\|_{H^2} \le \sqrt{C_2(1+C_0^{2k+2})}\delta t^2 + C \le \sqrt{C_2(1+1)} + C := \bar{C} \quad 0\le n\le m. \tag{3.57}$$

Note that $H^2 \subset L^\infty$, without loss of generality, we can adjust $\bar{C}$ so that we have

$$\|g(\bar{u}^{n+1})\|,\ \|g(\bar{u}^{n+1})\| \le \bar{C} \quad \forall 0 \le n \le m. \tag{3.58}$$

As $\bar{C}$ is independent of C_0 and δt, we still use the generic notation C to denote this upper bound.

Step 3: estimate for $|1-\xi^{m+1}|$. By direct calculation,

$$r_{tt} = \int_\Omega \Big(|\nabla u_t|^2 + \nabla u \cdot \nabla u_{tt} + \lambda u_t^2 + \lambda u u_{tt} + g(u)u_t^2 + g(u)u_{tt}\Big)dx. \tag{3.59}$$

It follows from (3.5b) that the equation for the errors can be written as

$$s^{n+1} - s^n = \delta t\Big(\|h[u(t^{n+1})]\|^2 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\|h(\bar{u}^{n+1})\|^2\Big) + T_1^n, \tag{3.60}$$

where $h(u) = \frac{\delta E}{\delta u} = -\Delta u + \lambda u - g(u)$, and

$$T_1^n = r(t^n) - r(t^{n+1}) + \delta t r_t(t^{n+1}) = \int_{t^n}^{t^{n+1}} (s - t^n) r_{tt}(s) ds. \tag{3.61}$$

Taking the sum of (3.60) for n from 0 to m, and noting that $s^0 = 0$, we have

$$s^{m+1} = \delta t \sum_{q=0}^{m} \Big(\|h[u(t^{q+1})]\|^2 - \frac{r^{q+1}}{E(\bar{u}^{q+1})}\|h(\bar{u}^{q+1})\|^2\Big) + \sum_{q=0}^{m} T_1^q. \tag{3.62}$$

We can bound the terms on the right hand side of (3.62) as follow: For T_1^n, noting (3.59) we have

$$|T_1^n| \le C\delta t \int_{t^n}^{t^{n+1}} |r_{tt}| ds \le C\delta t \int_{t^n}^{t^{n+1}} \Big(\|u_t(s)\|_{H^1}^2 + \|u_{tt}(s)\|_{H^1}\Big) ds. \tag{3.63}$$

Next,

$$\begin{aligned}
&\Big|\|h[u(t^{n+1})]\|^2 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\|h(\bar{u}^{n+1})\|^2\Big| \\
&\le \|h[u(t^{n+1})]\|^2\Big|1 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| + \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big|\|h[u(t^{n+1})]\|^2 - \|h(\bar{u}^{n+1})\|^2\Big| \\
&:= P_1^n + P_2^n.
\end{aligned} \tag{3.64}$$

For P_1^n, it follows from (3.34), $E(v) > \underline{C} > 0\,, \forall v$ and Theorem 3.2.1 that

$$\begin{aligned} P_1^n &\le C\Big|1 - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| \\ &\le C\Big|\frac{r(t^{n+1})}{E[u(t^{n+1})]} - \frac{r^{n+1}}{E[u(t^{n+1})]}\Big| + C\Big|\frac{r^{n+1}}{E[u(t^{n+1})]} - \frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| \\ &\le C\Big(|E[u(t^{n+1})] - E(\bar{u}^{n+1})| + |s^{n+1}|\Big). \end{aligned} \tag{3.65}$$

For P_2^n, it follows from (3.34), (3.35), (3.57), (3.58), $E(v) > \underline{C} > 0$ and Theorem 3.2.1 that

$$\begin{aligned} P_2^n &\le C\Big|\|h(\bar{u}^{n+1})\|^2 - \|h[u(t^{n+1})]\|^2\Big| \\ &\le C\|h(\bar{u}^{n+1}) - h[u(t^{n+1})]\|(\|h(\bar{u}^{n+1})\| + \|h[u(t^{n+1})]\|) \\ &\le C\Big(\|\Delta\bar{e}^{n+1}\| + \lambda\|\bar{e}^{n+1}\| + \|g(\bar{u}^{n+1}) - g[u(t^{n+1})]\|\Big) \\ &\le C\Big(\|\Delta\bar{e}^{n+1}\| + \|\bar{e}^{n+1}\|\Big). \end{aligned} \tag{3.66}$$

On the other hand,

$$\begin{aligned} |E[u(t^{n+1})] - E(\bar{u}^{n+1})| \le{}& \frac{1}{2}\Big(\|\nabla u(t^{n+1})\| + \|\nabla\bar{u}^{n+1}\|\Big)\|\nabla u(t^{n+1}) - \nabla\bar{u}^{n+1}\| \\ &+ \frac{\lambda}{2}\Big(\|u(t^{n+1})\| + \|\bar{u}^{n+1}\|\Big)\|u(t^{n+1}) - \bar{u}^{n+1}\| \\ &+ \int F[u(t^{n+1})]dx - \int F(\bar{u}^{n+1})dx \\ &\le C\Big(\|\nabla\bar{e}^{n+1}\| + \|\bar{e}^{n+1}\|\Big). \end{aligned} \tag{3.67}$$

Now, combining (3.56), (3.62)- (3.67), we arrive at

$$\begin{aligned} |s^{m+1}| &\le \delta t\sum_{q=0}^{m}\Big|\|h[u(t^{q+1})]\|^2 - \frac{r^{q+1}}{E(\bar{u}^{q+1})}\|h(\bar{u}^{q+1})\|^2\Big| + \sum_{q=0}^{m}|T_1^q| \\ &\le C\delta t\sum_{q=0}^{m}|s^{q+1}| + C\delta t\sum_{q=0}^{m}\|\bar{e}^{q+1}\|_{H^2} + C\delta t\int_0^T(\|u_t(s)\|_{H^1}^2 + \|u_{tt}(s)\|_{H^1})ds \\ &\le C\delta t\sum_{q=0}^{m}|s^{q+1}| + C\sqrt{C_2(1 + C_0^{2k+2})}\delta t^k + C\delta t. \end{aligned}$$

Applying the discrete Gronwall lemma to the above inequality with $\delta t < \frac{1}{2C}$, we obtain

$$\begin{aligned} |s^{m+1}| &\leq C\exp((1-C\delta t)^{-1})\delta t(\sqrt{C_2(1+C_0^{2k+2})}\delta t^{k-1}+1) \\ &\leq C_3\delta t(\sqrt{C_2(1+C_0^{2k+2})}\delta t^{k-1}+1), \end{aligned} \tag{3.68}$$

where C_3 is independent of C_0 and δt, can be defined as

$$C_3 := C\exp(2), \tag{3.69}$$

then $\delta t < \frac{1}{2C}$ can be guaranteed by

$$\delta t < \frac{1}{C_3}. \tag{3.70}$$

Hence, noting (3.65), (3.67),(3.68) and (3.57), we have

$$\begin{aligned} |1-\xi^{m+1}| &\leq C\Big(|E[u(t^{m+1})]-E(\bar{u}^{m+1})|+|s^{m+1}|\Big) \\ &\leq C(\|\bar{\mathrm{e}}^{m+1}\|_{H^1}+|s^{m+1}|) \\ &\leq C\delta t\Big(\bar{C}\sqrt{C_2(1+C_0^{2k+2})}\delta t^{k-1}+C_3(\sqrt{C_2(1+C_0^{2k+2})}\delta t^{k-1}+1)\Big) \\ &\leq C_4\delta t(\sqrt{1+C_0^{2k+2}}\delta t^{k-1}+1), \end{aligned} \tag{3.71}$$

where the constant C_4 is independent of C_0 and δt. Without loss of generality, we assume $C_4 > \max\{C_2, C_3, 1\}$ to simplify the proof below.

As a result of (3.71), $|1-\xi^{m+1}| \leq C_0\delta t$ if we define C_0 such that

$$C_4(\sqrt{1+C_0^{2k+2}}\delta t^{k-1}+1) \leq C_0. \tag{3.72}$$

For the cases $k \geq 2$, the above can be satisfied if we choose $C_0 = 3C_4$ and $\delta t \leq \frac{1}{1+C_0^{k+1}}$:

$$C_4(\sqrt{1+C_0^{2k+2}}\delta t^{k-1}+1) \leq C_4[(1+C_0^{k+1})\delta t+1] \leq 3C_4 = C_0. \tag{3.73}$$

For the case $k = 1$, we can not define C_0 satisfying (3.72) if $\eta_1^{n+1} = 1-(1-\xi^{n+1})^2$. However, if we choose $\eta_1^{n+1} = 1-(1-\xi^{n+1})^3$, we can repeat the same process above and arrive at a similar version of (3.72) for the first order case:

$$C_4(\sqrt{1+C_0^6}\delta t + 1) \le C_0. \tag{3.74}$$

The above can be satisfied if we choose $C_0 = 3C_4$ and $\delta t < \frac{1}{C_0^3}$ so that

$$C_4(\sqrt{1+C_0^6\delta t^2} + 1) \le C_4[1 + C_0^3\delta t + 1] \le 3C_4 = C_0.$$

To summarize, under the condition

$$\delta t \le \frac{1}{1+C_0^{k+2}}, \quad 1 \le k \le 5, \tag{3.75}$$

we have $|1-\xi^{m+1}| \le C_0\delta t$. Note that with $C_4 > \max\{C_2, C_3, 1\}$, (3.75) also implies (3.55) and (3.70). The induction process for (3.21) is complete.

Finally, thanks to (3.56), it remains to show $\|\mathrm{e}^{m+1}\|_{H^2} \le C\delta t^k$.

We derive from (3.5d) and (3.57) that

$$\|u^{m+1} - \bar{u}^{m+1}\|_{H^2} \le |\eta_k^{m+1} - 1|\|\bar{u}^{m+1}\|_{H^2} \le |\eta_k^{m+1} - 1|\bar{C}. \tag{3.76}$$

On the other hand, we derive from (3.21) that

$$|\eta_k^{m+1} - 1| \le C_0^{k+1}\delta t^{k+1}. \tag{3.77}$$

Then it follows from (3.56), (3.76) and (3.77) and combine the condition (3.25), (3.30) and (3.75) on δt that

$$\begin{aligned}\|\mathrm{e}^{m+1}\|_{H^2}^2 &\le 2\|\bar{\mathrm{e}}^{m+1}\|_{H^2}^2 + 2\|u^{m+1} - \bar{u}^{m+1}\|_{H^2}^2 \\ &\le 2C_2(1 + C_0^{2(k+1)})\delta t^{2k} + 2\bar{C}^2C_0^{2(k+1)}\delta t^{2(k+1)}\end{aligned}$$

holds under the condition $\delta t < \min\{\frac{1}{1+2C_0^{k+2}}, \frac{1-\tau_k}{3k}\}$. The proof is complete. □

Note that we set $\eta_1^{n+1} = 1-(1-\xi^{n+1})^3$ purely for technical reasons in the proof. It is clear that $\eta_1^{n+1} = 1-(1-\xi^{n+1})^2$ leads to first-order accuracy which is confirmed by our numerical tests.

3.4 Error analysis for the Cahn-Hilliard type equation

In this section, we consider the Cahn-Hilliard type equation

$$\frac{\partial u}{\partial t} = -\Delta^2 u + \lambda\Delta u - \Delta g(u) \quad (x,t) \in \Omega \times (0,T], \tag{3.78}$$

where Ω is an open bounded domain in $R^d\,(d=1,2,3)$, with the initial condition $u(x,0) = u^0(x)$ and boundary conditions

$$\text{periodic, or, } \frac{\partial u}{\partial n}|_{\partial\Omega} = \frac{\partial \Delta u}{\partial n}|_{\partial\Omega} = 0. \tag{3.79}$$

The above equation is a special case of (3.1) with $\mathcal{A} = \Delta^2 - \lambda\Delta$ and $g(u)$ replaced by $-\Delta g(u)$. It satisfies the dissipation law (3.2) with $E(u) = \frac{1}{2}(\mathcal{L}u, u) + (G(u), 1)$ where $(\mathcal{L}u, u) = (\nabla u, \nabla u) + \lambda(u,u)$, $G(u) = \int^u g(v)dv$ and $\mathcal{K}(u) = (\nabla\frac{\delta E}{\delta u}, \nabla\frac{\delta E}{\delta u})$.

In particular, with $g(u) = (1-u^2)u$ and $\lambda = 0$, the above equation becomes the celebrated Cahn-Hilliard equation [31].

We first recall the following result (cf. for instance [48]).

Theorem 3.4.1. *Let $u^0 \in H^2$, and (3.18) holds. We assume additionally*

$$|g(x)| < C(|x|^p + 1), \quad p > 0 \text{ arbitrary} \quad \text{if } n = 1,2; \quad 0 < p < 3 \quad \text{if } n = 3. \tag{3.80}$$

Then for any $T > 0$, there exists a unique solution u for (3.78) such that

$$u \in C([0,T]; H^2) \cap L^2(0,T; H^4).$$

We also recall the following result (see Lemma 2.3 in [39]) which we shall use to deal with the nonlinear term.

Theorem 3.4.2. *Assume that* $\|u\|_{H^1} \leq M$, *and that* (3.18) *and* (3.80) *hold. Then for any* $u \in H^4$, *there exist* $0 \leq \sigma < 1$ *and a constant* $C(M)$ *such that the following inequality holds:*

$$\|\Delta g(u)\|^2 \leq C(M)(1 + \|\Delta^2 u\|^{2\sigma}).$$

For (3.78), the kth-order version of (3.5a) and (3.11) read:

$$\frac{\alpha_k \bar{u}^{n+1} - A_k(u^n)}{\delta t} = -\Delta\Big(\Delta \bar{u}^{n+1} - \lambda \bar{u}^{n+1} + g[B_k(\bar{u}^n)]\Big), \tag{3.81}$$

and

$$\frac{\alpha_k u^{n+1} - \eta_k^{n+1} A_k(u^n)}{\delta t} = -\Delta\Big(\Delta u^{n+1} - \lambda u^{n+1} + \eta_k^{n+1} g[B_k(\bar{u}^n)]\Big), \tag{3.82}$$

where α_k, A_k, B_k defined in (3.8) - (3.10).

Theorem 3.4.3. *Given initial condition* $\bar{u}^0 = u^0 = u(0)$, $r^0 = E[u^0]$. *Let* $\bar{u}^{n+1}$ *and* u^{n+1} *be computed with the* k*th order scheme* (3.5a)- (3.5d) $(1 \leq k \leq 5)$ *for* (3.78) *with*

$$\eta_1^{n+1} = 1 - (1 - \xi^{n+1})^3, \quad \eta_k^{n+1} = 1 - (1 - \xi^{n+1})^{k+1} \ (k = 2, 3, 4, 5).$$

We assume (3.18) *and* (3.80) *hold, and*

$$u \in C([0, T]; H^3), \ \frac{\partial^{\mathrm{j}} u}{\partial t^{\mathrm{j}}} \in L^2(0, T; H^2) \ 1 \leq \mathrm{j} \leq k, \ \frac{\partial^{k+1} u}{\partial t^{k+1}} \in L^2(0, T; L^2).$$

Then for $n + 1 \leq T/\delta t$ *and* δt *small enough, we have*

$$\|\bar{\mathrm{e}}^{n+1}\|_{H^2}, \ \|\mathrm{e}^{n+1}\|_{H^2} \leq C\delta t^k,$$

where the constants C_0, C *are dependent on* T, Ω, *the* $k \times k$ *matrix* $G = (g_{\mathrm{ij}})$ *in Theorem 3.3.1 and the exact solution* u *but are independent of* δt.

Proof. For the sake of brevity, we shall only carry out in detail the error analysis for the first-order case. The analysis for the higher-order cases can be carried out by combining the procedures for the first-order case below and for the high-order cases in the proof Theorem 3.3.5. The detail will be left for the interested readers.

As in the proof of Theorem 3.3.5, we will first prove the following by induction:

$$|1-\xi^q| \le C_0\,\delta t,\ \ \forall q \le T/\delta t, \tag{3.83}$$

where the constant C_0 is dependent on T, Ω and the exact solution u but is independent of δt, and will be defined in the proof process.

Under the assumptions, (3.83) certainly holds for $q = 0$. Now suppose we have

$$|1-\xi^q| \le C_0\,\delta t,\ \ \forall q \le m, \tag{3.84}$$

we shall prove below that (3.83) holds for $q = m + 1$, namely,

$$|1-\xi^{m+1}| \le C_0\delta t. \tag{3.85}$$

We will carry out this proof in three steps.

Step 1: H^2 bound for u^n and $\bar{u}^n$ for all $n \le m$. It follows from Theorem 3.2.1 and under condition

$$\delta t \le \min\{\frac{1}{4C_0^3}, 1\}, \tag{3.86}$$

we have

$$\frac{3}{4} \le |\eta_1^q| \le 2,\ \ |1-\eta_1^q| \le \frac{\delta t^2}{4}, \forall q \le m, \tag{3.87}$$

and

$$\|u^q\|_{H^1} \le M_2,\ \ \forall q \le T/\delta t,\ \ \|\bar{u}^q\|_{H^1} \le \frac{4}{3}M_2,\ \forall q \le m. \tag{3.88}$$

Now, consider (3.82) at step q:

$$\frac{u^q - \eta_1^q u^{n-1}}{\delta t} = -\Delta^2 u^q + \lambda\Delta u^q - \eta_1^q \Delta g[\bar{u}^{q-1}] \tag{3.89}$$

Multiply (3.89) with $\Delta^2 u^q$, and by the similar process as step 1 in Theorem 3.3.5, we can obtain

$$\|\Delta u^q\|^2 - \|\Delta u^{q-1}\|^2 + \delta t\|\Delta^2 \bar{u}^q\|^2 - \frac{\delta t}{2}\|\Delta^2 \bar{u}^{q-1}\|^2 \leq C(M_2)\delta t + |1-\eta_1^q|\|u^{q-1}\|^2 \tag{3.90}$$

Taking the sum from 0 to $n\,(\leq m)$ of (3.90), we obtain

$$\begin{aligned}\|\Delta u^n\|^2 + \frac{\delta t}{2}\sum_{q=0}^{n}\|\Delta^2 \bar{u}^q\|^2 &\leq C(M_2)T + C(u^0) + \delta t^2\sum_{q=1}^{n-1}\|u^q\|^2 \\ &\leq C(M_2)T + C(u^0) + \delta t T M_2^2.\end{aligned}$$

with $C(M_2)$ is a constant only depends on M_2 and $C(u^0)$ only depends on u^0. Then together with (3.88) implies

$$\|u^n\|_{H^2} \leq \sqrt{C(M)T + C(u^0) + TM_2^2} + M_2 := C_1, \quad \forall n \leq m. \tag{3.91}$$

As $\|u^n\|_{H^2} = \eta_1^n\|\bar{u}^n\|_{H^2}$, (3.87) implies

$$\|\bar{u}^n\|_{H^2} \leq \frac{4}{3}C_1, \quad \forall n \leq m. \tag{3.92}$$

Step 2: estimates for $\|\bar{e}^{n+1}\|_{H^2}$ and $\|\bar{e}^{n+1}\|_{H^3}$ for all $0 \leq n \leq m$. By given assumption on the exact solution u and (3.92), we can choose C large enough such that

$$\|u(t)\|_{H^3} \leq C,\ \forall t \leq T, \|\bar{u}^q\|_{H^2} \leq C,\ \forall q \leq m, \tag{3.93}$$

and since $H^2 \subset L^\infty$, without loss of generality, we can adjust C such that

$$|g^{(\mathrm{i})}[u(t)]|_{L^\infty} \leq C,\ \forall t \leq T;\ |g^{(\mathrm{i})}(\bar{u}^q)|_{L^\infty} \leq C,\ \forall q \leq m;\ \mathrm{i} = 0,1,2,3. \tag{3.94}$$

From (3.81), we can write down the equation for error as

$$\bar{e}^{n+1} - \bar{e}^n = (\eta_1^n - 1)\bar{u}^n - \delta t\Delta^2 \bar{e}^{n+1} + \lambda\delta t\Delta\bar{e}^{n+1} + R_1^n + \delta t\Delta R_2^n, \tag{3.95}$$

where $R_1^n,\ R_2^n$ are given by

$$R_1^n = u(t^n) - u(t^{n+1}) + \delta t u_t(t^{n+1}) = \int_{t^n}^{t^{n+1}} (s - t^n) u_{tt} ds, \tag{3.96}$$

and

$$R_2^n = -g(\bar{u}^n) + g[u(t^{n+1})]. \tag{3.97}$$

Taking inner product with $\bar{\mathrm{e}}^{n+1} - \Delta\bar{\mathrm{e}}^{n+1} + \Delta^2\bar{\mathrm{e}}^{n+1}$ on both sides of (3.95), we obtain

$$\begin{aligned}
&\frac{1}{2}\Big(\|\bar{\mathrm{e}}^{n+1}\|^2 - \|\bar{\mathrm{e}}^n\|^2\Big) + \frac{1}{2}\|\bar{\mathrm{e}}^{n+1} - \bar{\mathrm{e}}^n\|^2 + \delta t\|\Delta\bar{\mathrm{e}}^{n+1}\|^2 + \lambda\delta t\|\nabla\bar{\mathrm{e}}^{n+1}\|^2 \\
&\quad + \frac{1}{2}\Big(\|\nabla\bar{\mathrm{e}}^{n+1}\|^2 - \|\nabla\bar{\mathrm{e}}^n\|^2\Big) + \frac{1}{2}\|\nabla(\bar{\mathrm{e}}^{n+1} - \bar{\mathrm{e}}^n)\|^2 + \delta t\|\nabla\Delta\bar{\mathrm{e}}^{n+1}\|^2 + \lambda\delta t\|\Delta\bar{\mathrm{e}}^{n+1}\|^2 \\
&\quad + \frac{1}{2}\Big(\|\Delta\bar{\mathrm{e}}^{n+1}\|^2 - \|\Delta\bar{\mathrm{e}}^n\|^2\Big) + \frac{1}{2}\|\Delta(\bar{\mathrm{e}}^{n+1} - \bar{\mathrm{e}}^n)\|^2 + \delta t\|\Delta^2\bar{\mathrm{e}}^{n+1}\|^2 + \lambda\delta t\|\nabla\Delta\bar{\mathrm{e}}^{n+1}\|^2 \\
&\quad = (\eta_1^n - 1)\Big(\bar{u}^n, \bar{\mathrm{e}}^{n+1}\Big) + \Big(R_1^n, \bar{\mathrm{e}}^{n+1}\Big) - \delta t\Big(\nabla R_2^n, \nabla\bar{\mathrm{e}}^{n+1}\Big) \\
&\quad + (\eta_1^n - 1)\Big(\nabla\bar{u}^n, \nabla\bar{\mathrm{e}}^{n+1}\Big) + \Big(R_1^n, -\Delta\bar{\mathrm{e}}^{n+1}\Big) + \delta t\Big(\nabla R_2^n, \nabla\Delta\bar{\mathrm{e}}^{n+1}\Big) \\
&\quad + (\eta_1^n - 1)\Big(\Delta\bar{u}^n, \Delta\bar{\mathrm{e}}^{n+1}\Big) + \Big(R_1^n, \Delta^2\bar{\mathrm{e}}^{n+1}\Big) + \delta t\Big(\Delta R_2^n, \Delta^2\bar{\mathrm{e}}^{n+1}\Big).
\end{aligned} \tag{3.98}$$

In the following, we bound the right hand side of (3.98). Noting that $|\eta_1^n - 1| \le C_0^3\,\delta t^3$, hence

$$|(\eta_1^n - 1)\Big(\bar{u}^n, \bar{\mathrm{e}}^{n+1}\Big)| \le \frac{\|(\eta_1^n - 1)\bar{u}^n\|^2}{\delta t} + \frac{\delta t}{4}\|\bar{\mathrm{e}}^{n+1}\|^2 \le CC_0^6\delta t^5 + \frac{\delta t}{4}\|\bar{\mathrm{e}}^{n+1}\|^2, \tag{3.99}$$

$$|(\eta_1^n - 1)\Big(\nabla\bar{u}^n, \nabla\bar{\mathrm{e}}^{n+1}\Big)| \le CC_0^6\delta t^5 + \frac{\delta t}{4}\|\nabla\bar{\mathrm{e}}^{n+1}\|^2, \tag{3.100}$$

and

$$|(\eta_1^n - 1)\Big(\Delta\bar{u}^n, \Delta\bar{\mathrm{e}}^{n+1}\Big)| \le CC_0^6\delta t^5 + \frac{\delta t}{4}\|\Delta\bar{\mathrm{e}}^{n+1}\|^2. \tag{3.101}$$

It follows from (3.96) that

$$|R_1^n|^2 \le C\delta t^3 \int_{t^n}^{t^{n+1}} |u_{tt}(s)|^2 ds. \tag{3.102}$$

Therefore,

$$|\Big(R_1^n, \bar{\mathrm{e}}^{n+1}\Big)| \leq \frac{1}{2\delta t}\|R_1^n\|^2 + \frac{\delta t}{2}\|\bar{\mathrm{e}}^{n+1}\|^2 \leq \frac{\delta t}{2}\|\bar{\mathrm{e}}^{n+1}\|^2 + C\delta t^2\int_{t^n}^{t^{n+1}}\|u_{tt}(s)\|^2 ds, \tag{3.103}$$

$$|\Big(R_1^n, -\Delta\bar{\mathrm{e}}^{n+1}\Big)| \leq \frac{\delta t}{2}\|\Delta\bar{\mathrm{e}}^{n+1}\|^2 + C\delta t^2\int_{t^n}^{t^{n+1}}\|u_{tt}(s)\|^2 ds, \tag{3.104}$$

and

$$|\Big(R_1^n, \Delta^2\bar{\mathrm{e}}^{n+1}\Big)| \leq \frac{\delta t}{2}\|\Delta^2\bar{\mathrm{e}}^{n+1}\|^2 + C\delta t^2\int_{t^n}^{t^{n+1}}\|u_{tt}(s)\|^2 ds. \tag{3.105}$$

Noting that

$$\begin{aligned}
|\nabla R_2^n| &= |\nabla g(\bar{u}^n) - \nabla g[u(t^n)] + \nabla g[u(t^n)] - \nabla g[u(t^{n+1})]| \\
&\leq |g(\bar{u}^n)\nabla\bar{u}^n - g[u(t^n)]\nabla u(t^n)| + |g[u(t^n)]\nabla u(t^n) - g[u(t^{n+1})]\nabla u(t^{n+1})| \\
&\leq |g(\bar{u}^n)|\Big|\nabla\bar{u}^n - \nabla u(t^n)\Big| + \Big|g(\bar{u}^n) - g[u(t^n)]\Big|\Big|\nabla u(t^n)\Big| \\
&+ \Big|g[u(t^n)] - g[u(t^{n+1})]\Big|\Big|\nabla u(t^n)\Big| + \Big|g[u(t^{n+1})]\Big|\Big|\nabla u(t^n) - \nabla u(t^{n+1})\Big| \\
&\leq C\Big(|\nabla\bar{\mathrm{e}}^n| + |\bar{\mathrm{e}}^n| + \int_{t^n}^{t^{n+1}}\Big(|u_t(s)| + |\nabla u_t(s)|\Big)ds\Big),
\end{aligned} \tag{3.106}$$

then for the terms with ∇R_2^n, it follows from (3.106) that

$$\begin{aligned}
\delta t|(\nabla R_2^n, \nabla\bar{\mathrm{e}}^{n+1})| &\leq \frac{\delta t}{2}\|\nabla R_2^n\|^2 + \frac{\delta t}{2}\|\nabla\bar{\mathrm{e}}^{n+1}\|^2 \\
&\leq C\delta t(\|\nabla\bar{\mathrm{e}}^{n+1}\|^2 + \|\bar{\mathrm{e}}^n\|_{H^1}^2) + C\delta t^2\int_{t^n}^{t^{n+1}}\|u_t(s)\|_{H^1}^2 ds,
\end{aligned} \tag{3.107}$$

and

$$\begin{aligned}
\delta t|(\nabla R_2^n, \nabla\Delta\bar{\mathrm{e}}^{n+1})| &\leq \frac{\delta t}{2}\|\nabla R_2^n\|^2 + \frac{\delta t}{2}\|\nabla\Delta\bar{\mathrm{e}}^{n+1}\|^2 \\
&\leq C\delta t\|\bar{\mathrm{e}}^n\|_{H^1}^2 + C\delta t^2\int_{t^n}^{t^{n+1}}\|u_t(s)\|_{H^1}^2 ds + \frac{\delta t}{2}\|\nabla\Delta\bar{\mathrm{e}}^{n+1}\|^2.
\end{aligned} \tag{3.108}$$

For the term with ΔR_2^n, since

$$|\Delta R_2^n| \leq |-\Delta g(\bar{u}^n) + \Delta g[u(t^n)]| + |-\Delta g[u(t^n)] + \Delta g[u(t^{n+1})]| := Q_1^n + Q_2^n,$$

and note that

$$\Delta g(u) = g(u)|\nabla u|^2 + g(u)\Delta u,$$

by using (3.93) and (3.94), we have

$$\begin{aligned} Q_1^n &\le \Big|g(\bar{u}^n)(|\nabla \bar{u}^n|^2 - |\nabla u(t^n)|^2)\Big| + \Big||\nabla u(t^n)|^2(g(\bar{u}^n) - g[u(t^n)])\Big| \\ &\quad + |g(\bar{u}^n)(\Delta \bar{u}^n - \Delta u(t^n))| + |\Delta u(t^n)(g(\bar{u}^n) - g[u(t^n)])| \\ &\le C\Big(|\nabla \bar{e}^n| + |\bar{e}^n| + |\Delta \bar{e}^n|\Big), \end{aligned}$$

and

$$Q_2^n \le C\Big(\int_{t^n}^{t^{n+1}} |\nabla u_t(s)|^2 ds + \int_{t^n}^{t^{n+1}} |\Delta u_t(s)| ds\Big).$$

Therefore,

$$\begin{aligned} \delta t|\Big(\Delta R_2^n, \Delta^2 \bar{e}^{n+1}\Big)| &\le \delta t|\Big(Q_1^n, \Delta^2 \bar{e}^{n+1}\Big)| + \delta t|\Big(Q_2^n, \Delta^2 \bar{e}^{n+1}\Big)| \\ &\le \delta t\|Q_1^n\|^2 + \frac{\delta t}{4}\|\Delta^2 \bar{e}^{n+1}\|^2 + \delta t\|Q_2^n\|^2 + \frac{\delta t}{4}\|\Delta^2 \bar{e}^{n+1}\|^2 \\ &\le C\delta t(\|\bar{e}^n\|^2 + \|\nabla \bar{e}^n\|^2 + \|\Delta \bar{e}^n\|^2) + \frac{\delta t}{2}\|\Delta^2 \bar{e}^{n+1}\|^2 \\ &\quad + C\delta t^2 \int_{t^n}^{t^{n+1}} \|u_t(s)\|_{H^2}^2 ds, \end{aligned} \tag{3.109}$$

where we used the following inequality

$$\begin{aligned} \int_\Omega \Big(\int_{t^n}^{t^{n+1}} (|\nabla u_t(s)| + |\Delta u_t(s)|) ds\Big)^2 dx &\le \int_\Omega \Big(\int_{t^n}^{t^{n+1}} (|\nabla u_t(s)| + |\Delta u_t(s)|)^2 ds \int_{t^n}^{t^{n+1}} 1 ds\Big) dx \\ &\le C\delta t \int_{t^n}^{t^{n+1}} \|u_t(s)\|_{H^2}^2 ds. \end{aligned}$$

Now, combining (3.98)-(3.108) and (3.109) and dropping some unnecessary terms, we arrive at

$$\begin{aligned} &\|\bar{e}^{n+1}\|^2 - \|\bar{e}^n\|^2 + \|\nabla \bar{e}^{n+1}\|^2 - \|\nabla \bar{e}^n\|^2 + \|\Delta \bar{e}^{n+1}\|^2 - \|\Delta \bar{e}^n\|^2 + \delta t\|\nabla \Delta \bar{e}^{n+1}\|^2 \\ &\qquad \le CC_0^6 \delta t^5 + C\delta t(\|\nabla \bar{e}^{n+1}\|^2 + \|\bar{e}^{n+1}\|^2 + \|\Delta \bar{e}^n\|^2 + \|\nabla \bar{e}^n\|^2 + \|\bar{e}^n\|^2) \\ &\qquad + C\delta t^2 \int_{t^n}^{t^{n+1}} (\|u_t(s)\|_{H^2}^2 + \|u_{tt}(s)\|^2) ds. \end{aligned} \tag{3.110}$$

Taking the sum of the above for n from 0 to m, we obtain

$$\|\bar{\mathrm{e}}^{m+1}\|_{H^2}^2 + \delta t \sum_{q=0}^{m} \|\nabla\Delta\bar{\mathrm{e}}^{q+1}\|^2 \leq C\delta t \sum_{q=0}^{m+1} \|\bar{\mathrm{e}}^{q}\|_{H^2}^2 + C\delta t^2 \int_0^T (\|u_t(s)\|_{H^2}^2 + \|u_{tt}(s)\|^2 + C_0^6\delta t^2)ds. \tag{3.111}$$

Finally, we can obtain the following estimate for $\bar{\mathrm{e}}^{m+1}$ by applying the discrete Gronwall's inequality to (3.111) with $\delta t < \frac{1}{2C}$:

$$\begin{aligned}\|\bar{\mathrm{e}}^{n+1}\|_{H^2}^2 + \delta t \sum_{q=0}^{n} \|\nabla\Delta\bar{\mathrm{e}}^{q+1}\|^2 &\leq C\exp((1-\delta tC)^{-1})\delta t^2 \int_0^T (\|u_t(s)\|_{H^2}^2 + \|u_{tt}(s)\|^2 + C_0^6\delta t^2)ds \\ &\leq C_2(1+C_0^6\delta t^2)\delta t^2, \quad \forall\, 0 \leq n \leq m.\end{aligned} \tag{3.112}$$

where C_2 is independent of δt and C_0, can be defined as

$$C_2 := C\exp(2)\max\Big(\int_0^T (\|u_t(s)\|_{H^2}^2 + \|u_{tt}(s)\|^2)ds, 1\Big), \tag{3.113}$$

and hence $\delta t < \frac{1}{2C}$ can be guaranteed by $\delta t < \frac{1}{C_2}$. In particular, (3.112) implies

$$\|\bar{\mathrm{e}}^{n+1}\|_{H^2},\ \Big(\delta t \sum_{q=0}^{n} \|\nabla\Delta\bar{\mathrm{e}}^{q+1}\|^2\Big)^{1/2} \leq \sqrt{C_2(1+C_0^6\delta t^2)}\delta t, \quad \forall\, 0 \leq n \leq m. \tag{3.114}$$

Combining (3.93) and (3.114), we obtain that for all $\forall\, 0 \leq n \leq m$ and under the condition on δt in (3.86), we have

$$\|\bar{u}^{n+1}\|_{H^2},\ \Big(\delta t \sum_{q=0}^{n} \|\nabla\Delta\bar{u}^{q+1}\|^2\Big)^{1/2} \leq \sqrt{C_2(1+C_0^6\delta t^2)}\delta t + C \leq \sqrt{C_2(1+1)} + C := \bar{C}. \tag{3.115}$$

Note that $H^2 \subset L^\infty$, without loss of generality, we can adjust $\bar{C}$ so that we have

$$\|g(\bar{u}^{n+1})\|,\ \|g(\bar{u}^{n+1})\| \leq \bar{C}, \quad \forall 0 \leq n \leq m. \tag{3.116}$$

As $\bar{C}$ is independent of C_0 and δt, we still use the generic notation C to denote this upper bound.

Step 3: estimate for $|1-\xi^{n+1}|$. It follows from (3.5b) that the equation for the error $\{s^{\mathrm{j}}\}$ can be written as

$$s^{n+1}-s^n=\delta t\Big(\|\nabla h[u(t^{n+1})]\|^2-\frac{r^{n+1}}{E(\bar{u}^{n+1})}\|\nabla h(\bar{u}^{n+1})\|^2\Big)+T_1^n, \tag{3.117}$$

where $h(u)=\frac{\delta E}{\delta u}=-\Delta u+\lambda u-g(u)$ and truncation errors T_1^n is given in (3.61) with a bound given in (3.63).

Taking the sum of (3.117) for n from 0 to m, since $s^0=0$, we have

$$s^{m+1}=\delta t\sum_{q=0}^{m}\Big(\|\nabla h[u(t^{q+1})]\|^2-\frac{r^{q+1}}{E(\bar{u}^{q+1})}\|\nabla h(\bar{u}^{q+1})\|^2\Big)+\sum_{q=0}^{m}T_1^q. \tag{3.118}$$

For $\|\nabla h[u(t^{n+1})]\|^2-\frac{r^{n+1}}{E(\bar{u}^{n+1})}\|\nabla h(\bar{u}^{n+1})\|^2$, we have

$$\begin{aligned}
&\Big|\|\nabla h[u(t^{n+1})]\|^2-\frac{r^{n+1}}{E(\bar{u}^{n+1})}\|\nabla h(\bar{u}^{n+1})\|^2\Big|\\
&\le\|\nabla h[u(t^{n+1})]\|^2\Big|1-\frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big|+\frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big|\|\nabla h[u(t^{n+1})]\|^2-\|\nabla h(\bar{u}^{n+1})\|^2\Big| \qquad (3.119)\\
&:=K_1^n+K_2^n.
\end{aligned}$$

For K_1^n, it follows from (3.93), $E(\bar{u}^{n+1})>\underline{C}>0$ and Theorem 4.3.1 that

$$\begin{aligned}
K_1^n&\le C\Big|1-\frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big|\\
&=C\Big|\frac{r(t^{n+1})}{E[u(t^{n+1})]}-\frac{r^{n+1}}{E[u(t^{n+1})]}\Big|+C\Big|\frac{r^{n+1}}{E[u(t^{n+1})]}-\frac{r^{n+1}}{E(\bar{u}^{n+1})}\Big| \qquad (3.120)\\
&\le C\Big(|E[u(t^{n+1})]-E(\bar{u}^{n+1})|+|s^{n+1}|\Big).
\end{aligned}$$

For K_2^n, it follows from (3.93), (3.94), (3.115), (3.116), $E(\bar{u}^{n+1}) > \underline{C} > 0$ and Theorem 3.2.1 that

$$\begin{aligned}
K_2^n &\le C\Big|\|\nabla h(\bar{u}^{n+1})\|^2 - \|\nabla h[u(t^{n+1})]\|^2\Big| \\
&\le C\|\nabla h(\bar{u}^{n+1}) - \nabla h[u(t^{n+1})]\|(\|\nabla h(\bar{u}^{n+1})\| + \|\nabla h[u(t^{n+1})]\|) \\
&\le C(1+\|\nabla\Delta\bar{u}^{n+1}\|)\Big(\|\nabla\Delta\bar{e}^{n+1}\| + \lambda\|\nabla\bar{e}^{n+1}\| + \|\nabla(g(\bar{u}^{n+1}) - g[u(t^{n+1})])\|\Big) \\
&\le C\Big(\|\nabla\Delta\bar{e}^{n+1}\| + \|\nabla\bar{e}^{n+1}\|\Big) + C\|\nabla\Delta\bar{u}^{n+1}\|\|\nabla\Delta\bar{e}^{n+1}\| + C\|\nabla\Delta\bar{u}^{n+1}\|\|\nabla\bar{e}^{n+1}\|.
\end{aligned} \tag{3.121}$$

It then follows from (3.114), (3.115) and the Cauchy-Schwarz inequality that

$$\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{u}^q\|\|\nabla\bar{e}^q\| \le \Big(\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{u}^q\|^2\delta t\sum_{q=1}^{n+1}\|\nabla\bar{e}^q\|^2\Big)^{1/2} \le C\sqrt{C_2(1+C_0^6\delta t^2)}\delta t,$$

and

$$\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{u}^q\|\|\nabla\Delta\bar{e}^q\| \le \Big(\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{u}^q\|^2\delta t\sum_{q=1}^{n+1}\|\nabla\Delta\bar{e}^q\|^2\Big)^{1/2} \le C\sqrt{C_2(1+C_0^6\delta t^2)}\delta t.$$

For $E[u(t^{n+1})] - E(\bar{u}^{n+1})$, we have estimate (3.67).

Now, we are ready to estimate s^{m+1}. Combine the estimate obtained above, (3.118) leads to

$$\begin{aligned}
|s^{m+1}| &\le \delta t\sum_{q=0}^{m}\Big|\|\nabla h[u(t^{q+1})]\|^2 - \frac{E_0(\bar{u}^{q+1}) + r^{q+1}}{E(\bar{u}^{q+1})}\|\nabla h(\bar{u}^{q+1})\|^2\Big| + \sum_{q=0}^{m}|T_1^q| \\
&\le C\delta t\sum_{q=0}^{m}|s^{q+1}| + C\delta t\sum_{q=0}^{m}\|\bar{e}^{q+1}\|_{H^1} + C\delta t\sum_{q=0}^{m}\|\nabla\Delta e^{q+1}\| \\
&+ C\delta t\sum_{q=1}^{m+1}\|\nabla\Delta\bar{u}^q\|\|\nabla\bar{e}^q\| + C\delta t\sum_{q=1}^{m+1}\|\nabla\Delta\bar{u}^q\|\|\nabla\Delta\bar{e}^q\| \\
&+ C\delta t\int_0^{t^{m+1}}(\|u_t(s)\|_{H^1}^2 + \|u_{tt}(s)\|_{H^1})ds \\
&\le C\delta t\sum_{q=0}^{m}|s^{q+1}| + C\delta t\Big(\sqrt{C_2(1+C_0^6\delta t^2)} + 1\Big)
\end{aligned} \tag{3.122}$$

Finally, applying the discrete Gronwall's inequality on (3.122) with $\delta t < \frac{1}{2C}$, we obtain the following estimate for s^{n+1},

$$\begin{aligned} |s^{n+1}| &\leq C\exp((1-\delta tC)^{-1})\delta t\Big(\sqrt{C_2(1+C_0^6\delta t^2)}+1\Big) \\ &\leq C_3\delta t\Big(\sqrt{C_2(1+C_0^6\delta t^2)}+1\Big), \forall\, 0\leq n\leq m, \end{aligned} \tag{3.123}$$

where C_3 is independent of δt and C_0, can be defined as

$$C_3 := C\exp(2). \tag{3.124}$$

Thanks to (3.123), we can define C_0 and then prove (3.85) by following exactly the same procedure as **Step 3 in Theorem 3.3.5** with the condition

$$\delta t \leq \frac{1}{1+C_0^3} \tag{3.125}$$

The induction process for (3.83) is completed.

Finally, thanks to (3.114), it remains to show $\|\mathrm{e}^{m+1}\|_{H^2} \leq C\delta t^k$.

We derive from (3.115) that

$$\|u^{m+1}-\bar{u}^{m+1}\|_{H^2} \leq |\eta_1^{m+1}-1|\|\bar{u}^{m+1}\|_{H^2} \leq |\eta_1^{m+1}-1|\bar{C}. \tag{3.126}$$

On the other hand, (3.83) implies

$$|\eta_1^{m+1}-1| \leq C_0^3\delta t^3. \tag{3.127}$$

Then it follows from (3.114), (3.126) and (3.127) that

$$\begin{aligned} \|\mathrm{e}^{m+1}\|_{H^2}^2 &\leq 2\|\bar{\mathrm{e}}^{m+1}\|_{H^2}^2 + 2\|u^{m+1}-\bar{u}^{m+1}\|_{H^2}^2 \\ &\leq 2C_2(1+C_0^6\delta t^2)\delta t^2 + 2\bar{C}^2C_0^6\delta t^6. \end{aligned}$$

To summarize, combine the condition (3.86) and (3.125) on δt, we obtain $\|\mathrm{e}^{m+1}\|_{H^2} \leq C\delta t$ with $\delta t < \frac{1}{1+4C_0^3}$. The proof for the case $k=1$ is complete. □

3.5 Numerical examples

In this section, we provide numerical examples to demonstrate the convergence rates for several typical dissipative systems.

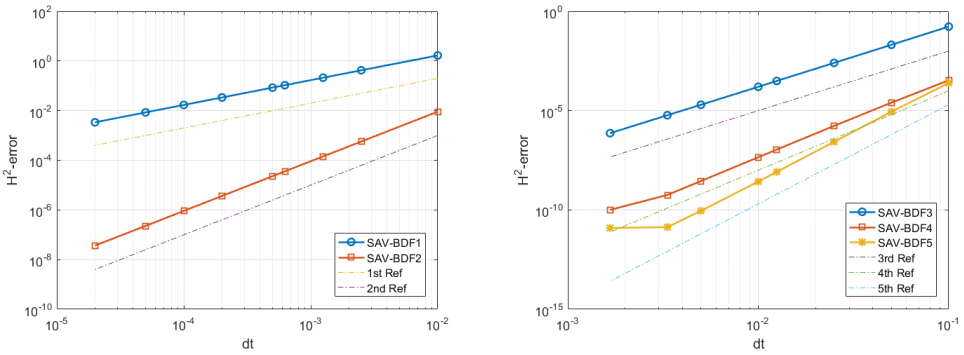
Example 1. Consider the 1-D Burgers equation

$$\frac{\partial u}{\partial t} - \nu u_{xx} + u u_x = f, \tag{3.128}$$

in $\Omega = (-1, 1)$ with Dirichlet boundary condition, and f is chosen such that the exact solution is

$$u(x, t) = \sin(\pi x)\sin(t), \quad u(\pm 1, t) = 0. \tag{3.129}$$

In the test, we choose $\nu = 0.05$ and use the Legendre-Galerkin method [32] with 30 modes for space discretization so that the spatial discretization error is negligible compared with the time discretization error for the range of δt tested. In Figures 3.1, we plot the convergence rate of the H^2 error at $T = 1$ for the Burgers equation. We observe the expected convergence rates for all cases.



(a) BDF1 and BDF2 vs errors of u (b) BDF3, 4, 5 vs errors of u

Figure 3.1. Convergence rate for the Burgers equation using the new SAV/BDFk ($k = 1, 2, 3, 4, 5$). (a)-(b) H^2 errors of u as a function of Δt.

Next, we consider the 1-D Burgers equation

$$\frac{\partial u}{\partial t} - \nu u_{xx} + u u_x = 0, \tag{3.130}$$

in $\Omega = (-1, 1)$ with the initial condition and Dirichlet boundary condition given as

$$u(x, 0) = -\sin(\pi x), \quad u(\pm 1, t) = 0. \tag{3.131}$$

In this test, we use the second order SAV scheme and the corresponding second-order IMEX scheme with $\nu = \frac{1}{314}$, $N = 320$, $\delta t = 8.5 \times 10^{-3}$. The numerical solutions at $T = 1$ are plotted in Fig 3.2 (a) solution obtained by the usual IMEX scheme and (b) solution obtained by the SAV scheme. We observe that the usual IMEX scheme produces oscillatory solutions while the SAV scheme produces the correct solution which is indistinguishable with the reference solution obtained with $\delta t = 10^{-4}$ in Fig 3.2 (c). We also plot in Fig 3.2 (d) the SAV factor $\eta^n = 1 - (1 - \xi^n)^3$. We observe that when the solution exhibits large gradients (for $t \in (0.5, 1)$), the SAV factor η^n deviates slightly from 1 so that the SAV scheme still produces correct result while the corresponding IMEX scheme produces incorrect result.

Example 2. Consider the Allen-Cahn equation

$$\frac{\partial u}{\partial t} = \alpha \Delta u - (1 - u^2)u + f, \tag{3.132}$$

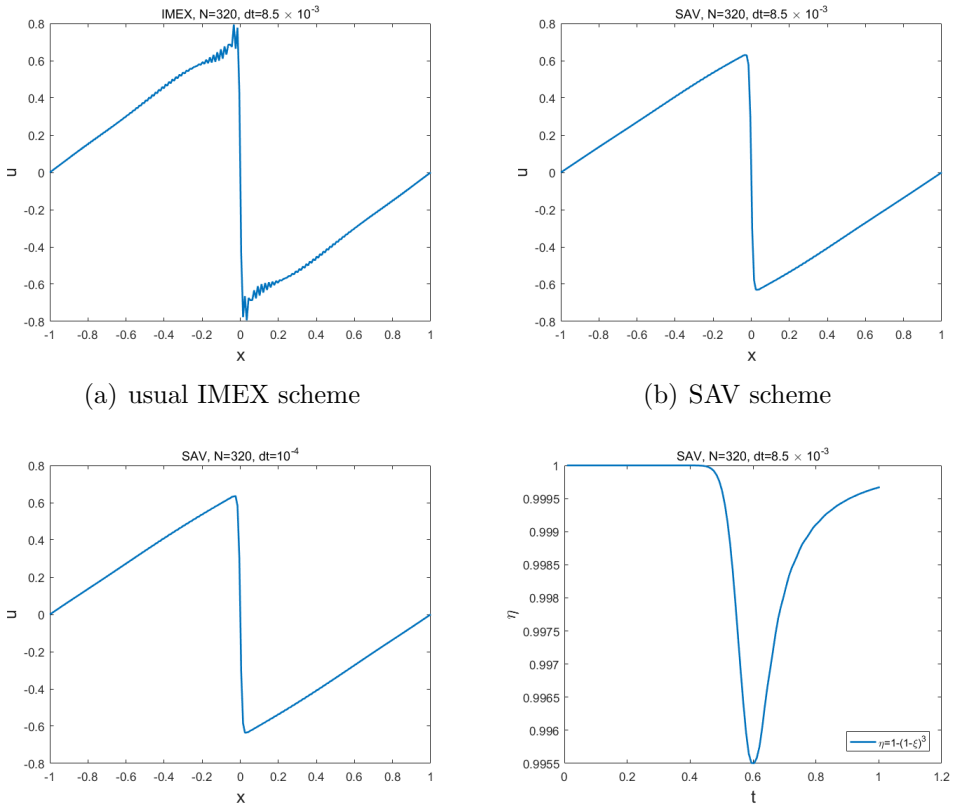
and the Cahn-Hilliard equation

$$\frac{\partial u}{\partial t} = -m_0 \Delta(\alpha \Delta u - (1 - u^2)u) + f, \tag{3.133}$$

in $\Omega = (0, 2) \times (0, 2)$ with periodic boundary condition, and f is chosen such that the exact solution is

$$u(x, y, t) = \exp\Big(\sin(\pi x)\sin(\pi y)\Big)\sin(t). \tag{3.134}$$

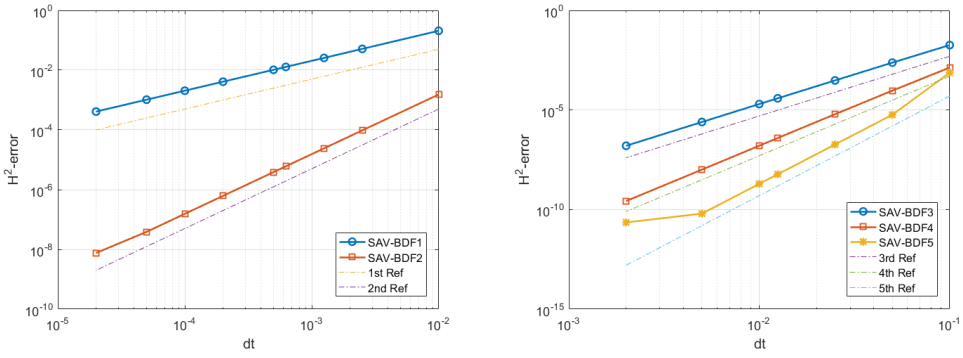
We set $\alpha = 0.01^2$ in (3.132) and $\alpha = 0.04$, $m_0 = 0.005$ in (3.133), and use the Fourier spectral method with 64×64 modes for space discretization. In Figures 3.3 (resp. 3.4), we plot the convergence rate of the H^2 error at $T = 1$ for the Allen-Cahn (resp. Cahn-Hilliard) equation. We also observe the expected convergence rates for all cases.



(a) usual IMEX scheme (b) SAV scheme

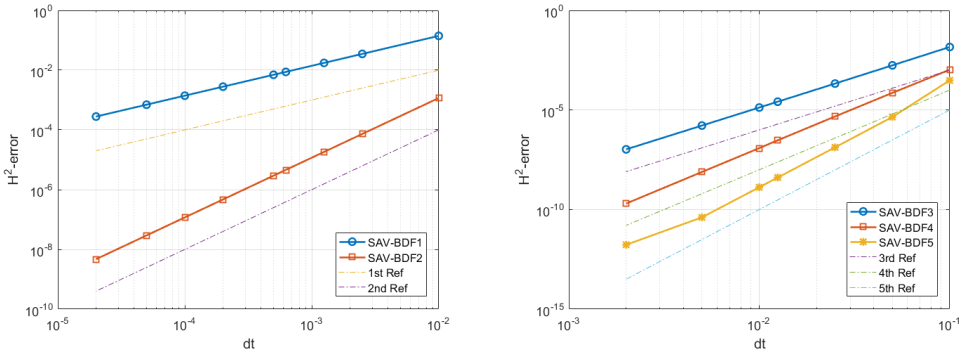
(c) reference solution (d) SAV factor

Figure 3.2. Burgers equation: a comparison of usual IMEX and SAV



(a) BDF1 and BDF2 vs errors of u (b) BDF3, 4, 5 vs errors of u

Figure 3.3. Convergence test for the Allen-Cahn equation using the new SAV/BDFk ($k = 1, 2, 3, 4, 5$). (a)-(b) H^2 errors of u as a function of Δt.



(a) BDF1 and BDF2 vs errors of u

(b) BDF3, 4, 5 vs errors of u

Figure 3.4. Convergence test for the Cahn-Hilliard equation using the new SAV/BDFk ($k = 1, 2, 3, 4, 5$). (a)-(b) H^2 errors of u as a function of Δt.

3.6 Conclusion of this chapter

We constructed a class of implicit-explicit BDFk SAV schemes, based on the schemes in [13], for general linear systems. This class of schemes enjoys the following distinct advantages: (i) it only requires solving, in most common situations, one linear system with constant coefficients at each time step, which is half of the cost for the original SAV approach; (ii) it is not restricted to gradient flows and is applicable to general dissipative systems; and (iii) it can be high-order with unconditional stability and suitable for adaptive time stepping without restriction on time step size; and most importantly, (iv) it leads to unconditional uniform bound, for any order k on the norm based on the principal linear term in the energy functional, which is of critical importance for the convergence and error analysis.

Using the uniform bound on the norm based on the principal linear operator that we derived for the BDFk SAV schemes and to a stability result in [14] for the BDFk ($k = 1, 2, 3, 4, 5$) schemes, we were able to establish, with a delicate inductive argument, rigorous error estimates for the BDFk ($k = 1, 2, 3, 4, 5$) SAV schemes in a unified form for the typical Allen-Cahn and Cahn-Hilliard type equations.

4. NEW SAV APPROACH FOR INCOMPRESSIBLE NAVIER STOKES EQUATION WITH PERIODIC BOUNDARY CONDITION

In this chapter, we construct high-order semi-discrete-in-time and fully discrete (with Fourier-Galerkin in space) schemes for the incompressible Navier-Stokes equations with periodic boundary conditions, and carry out corresponding error analysis. The schemes are of implicit-explicit type based on a scalar auxiliary variable (SAV) approach. It is shown that numerical solutions of these schemes are uniformly bounded without any restriction on time step size. These uniform bounds enable us to carry out a rigorous error analysis for the schemes up to fifth-order in a unified form, and derive global error estimates in $l^\infty(0,T;H^1)\cap l^2(0,T;H^2)$ in the two dimensional case as well as local error estimates in $l^\infty(0,T;H^1)\cap l^2(0,T;H^2)$ in the three dimensional case. We also present numerical results confirming our theoretical convergence rates and demonstrating advantages of higher-order schemes for flows with complex structures in the double shear layer problem. Most of the results in this chapter are extracted from [49].

4.1 Introduction

Numerical approximation of the Navier-Stokes equations has been a subject of intensive study for many decades and continues to attract considerable attention, as it plays a fundamental role in computational fluid dynamics. Most of the work are concerned with the Navier-Stokes equations with non periodic boundary conditions, as is the case with the most applications. An enormous amount of work has been devoted to construct efficient and stable numerical algorithms for solving the incompressible Navier-Stokes equations with non periodic boundary conditions, see [50]–[55] and the references therein. In particular, the papers [56]–[61], among others, are particularly concerned with the error estimates for semi-discrete-in-time or fully discrete schemes.

We consider in this chapter numerical approximation of the incompressible Navier-Stokes equations in primitive formulation:

$$\frac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla) u + \nabla p = 0, \tag{4.1a}$$

$$\nabla \cdot u = 0, \tag{4.1b}$$

with a suitable initial condition $u|_{t=0} = u_0$ in a rectangular domain $\Omega \subset R^d$ $(d = 2, 3)$ with periodic boundary conditions. The unknowns are velocity u and the pressure p which is assumed to have zero mean for uniqueness, $\nu > 0$ is the viscosity. To simplify the presentation, we have set the external force to be zero. But our schemes and analytical results can be naturally extended to the case with a non-zero external force.

The incompressible Navier-Stokes equations with periodic boundary conditions retain the essential mathematical properties/difficulties of the system with non periodic boundary conditions, but are amenable to very efficient numerical algorithms using the Fourier-spectral method, and are particularly useful in the study of homogeneous turbulence [62]–[64].

There exists also a significant number of work devoted to the numerical analysis for Navier-Stokes equations with periodic boundary conditions. For examples, in [65], Hald proved the convergence of semi-discrete Fourier-Galerkin methods in two and three dimensions; in [66], E used semigroup theory to establish convergence and error estimates of the semi-discrete Fourier-Galerkin and Fourier-collocation methods in various energy norms and L^p-norms; in [67], Wang proved uniform bounds and convergence of long time statistics for a semi-discrete second-order implicit-explicit (IMEX) scheme for the 2-D Navier-Stokes equations with periodic boundary conditions in vorticity-stream function formulation, see also related work in [68], [69]; in [70], Cheng and Wang established uniform bounds for semi-discrete higher-order (up to fourth-order) IMEX scheme for the 2-D Navier-Stokes equations with periodic boundary conditions in vorticity-stream function formulation; in [71], Heister et al. proved uniform bounds for a fully discrete finite-element and second-order IMEX scheme for the 2-D Navier-Stokes equations with periodic boundary conditions in vorticity-velocity formulation. Note that the uniform bounds for semi-discrete IMEX schemes obtained in

the above references are for two-dimensional cases only and require that the time step be sufficiently small.

It appears that, except some recently constructed schemes based on the scalar auxiliary variable (SAV) approach [72], [73], all other IMEX type schemes (i.e., the nonlinear term is treated explicitly) for Navier-Stokes equations require the time step to be sufficiently small to have a bounded numerical solution. Furthermore, to the best of our knowledge, there is no error analysis for any IMEX scheme for the three-dimensional Navier-Stokes equations, and no error estimate is available for any higher-order (≥ 3) IMEX scheme.

In this chapter, we construct semi-discrete and fully discrete with Fourier-Galerkin in space SAV IMEX schemes and carry out a unified stability and error analysis. Our main contributions include:

- Our semi-discrete and fully discrete schemes of arbitrary order in time are unconditionally stable without any restriction on time step size;
- Global error estimates in $l^\infty(0,T;H^1)\cap l^2(0,T;H^2)$ up to fifth-order in time are established for the two-dimensional case;
- Local error estimates in $l^\infty(0,T_*;H^1)\cap l^2(0,T_*;H^2)$ (with a $T_*\leq T$) up to fifth-order in time are established for the three-dimensional case.

Our schemes are constructed using the SAV approach proposed in [13] which can be used for general dissipative systems. The main advantages of this approach, compared with other SAV approaches proposed in [72], [73] for Navier-Stokes equations is that our schemes are linear, decoupled and can be high-order. Moreover, in the two dimensional case, we use a stronger energy dissipation law (4.7), which is only true for the 2-D Navier-Stokes equations with periodic boundary conditions, that leads to a uniform bound for the numerical solution in $l^\infty(0,T;H^1)$, as opposed to $l^\infty(0,T;L^2)$ in the three dimensional case.

4.2 Preliminaries

We first introduce some notations. We denote by $(\cdot,\cdot)$ and $\|\cdot\|$ the inner product and the norm in $L^2(\Omega)$, and denote

$$H_p^k(\Omega) = \{u^{\mathrm{j}}\,(\mathrm{j}=0,1,\cdots,k)\ \in L^2(\Omega) : u^{\mathrm{j}}\,(\mathrm{j}=0,1,\cdots,k-1) \quad \text{periodic}\},$$

with norm $\|\cdot\|_k$. For non-integer $s>0$, $H_p^s(\Omega)$ and the corresponding norm $\|\cdot\|_s$ are defined by space interpolation [74]. In particular, we set $H_p^0(\Omega) = L_0^2(\Omega)$.

Let V be a Banach space, we shall also use the standard notations $L^p(0,T;V)$ and $C([0,T];V)$. To simplify the notation, we often omit the spatial dependence for the exact solution u, i.e., $u(x,t)$ is often denoted by $u(t)$. We shall use bold faced letters to denote vectors and vector spaces, and use C to denote a generic positive constant independent of the discretization parameters.

We now define the following spaces which are particularly used for Navier-Stokes equations:

$$\mathbf{H} = \{v \in L_0^2(\Omega) : \nabla\cdot v = 0\}, \quad \mathbf{V} = \{v \in H_p^1(\Omega) : \nabla\cdot v = 0\}.$$

Let $v \in L_0^2(\Omega)$, we define $w := \Delta^{-1}v$ as the solution of

$$\Delta w = v \quad x\in\Omega; \quad w \text{ periodic with zero mean.}$$

Note that in the periodic case, we can define the operators ∇, $\nabla\cdot$ and Δ^{-1} in the Fourier space by expanding functions and their derivatives in Fourier series, and one can easily show that these operators commute with each other.

We define a linear operator $\mathbf{A}$ in $L_0^2(\Omega)$ by

$$\mathbf{A}v := \nabla\times\nabla\times\Delta^{-1}v, \quad \forall v \in L_0^2(\Omega). \tag{4.2}$$

Since

$$\|\Delta w\|^2 = \|\nabla\times\nabla\times w\|^2 + \|\nabla\nabla\cdot w\|^2 \quad \forall w \in H_p^2(\Omega),$$

we derive immediately from the above that

$$\|\mathbf{A}v\| = \|\Delta\Delta^{-1}v\| - \|\nabla\nabla\cdot\Delta^{-1}v\| \le \|v\|,\ \forall v \in L_0^2(\Omega). \tag{4.3}$$

Next, we define the trilinear form $b(\cdot,\cdot,\cdot)$ and $b_{\mathbf{A}}(\cdot,\cdot,\cdot)$ by

$$b(u,v,w) = \int_\Omega (u\cdot\nabla)v\cdot w d,\ b_{\mathbf{A}}(u,v,w) = \int_\Omega \mathbf{A}((u\cdot\nabla)v)\cdot w d.$$

In particular, we have

$$b(u,v,w) = -b(u,w,v),\ \forall u \in \mathbf{H},\ v,w \in H_p^1(\Omega),$$

which implies

$$b(u,v,v) = 0,\ \forall u \in \mathbf{H},\ v \in H_p^1(\Omega). \tag{4.4}$$

In the two-dimensional periodic case, we have also [75]

$$b(u,u,\Delta u) = 0,\ \forall u \in H_p^2(\Omega). \tag{4.5}$$

Taking the inner product of (4.1) with u, thanks to (4.4), we find that solution of the Navier-Stokes equations (4.1) satisfies the energy dissipation law

$$\frac{1}{2}\frac{d}{dt}\|u\|^2 = -\nu\|\nabla u\|^2 \qquad (d=2,3). \tag{4.6}$$

On the other hand, in the two dimensional periodic case, taking the inner product of (4.1) with $-\Delta u$, thanks to (4.5), we derive another energy dissipation law [75]

$$\frac{1}{2}\frac{d}{dt}\|\nabla u\|^2 = -\nu\|\Delta u\|^2 \qquad (d=2). \tag{4.7}$$

Using (4.3), Hölder inequality and Sobolev inequality, we have [75]

$$b(u,v,w),\ b_{\mathbf{A}}(u,v,w) \leq c\|\mathbf{u}\|_1^{1/2}\|\mathbf{u}\|^{1/2}\|\mathbf{v}\|_2^{1/2}\|\mathbf{v}\|_1^{1/2}\|\mathbf{w}\|, \quad d=2; \tag{4.8}$$

$$b(u,v,w),\ b_{\mathbf{A}}(u,v,w) \leq c\|u\|_1\|\nabla v\|_{1/2}\|w\|, \quad d=3. \tag{4.9}$$

We also use frequently the following inequalities [75]:

$$b(u,v,w),\ b_{\mathbf{A}}(u,v,w) \leq \begin{cases} c\|u\|_1\|v\|_1\|w\|_1; \\ c\|u\|_2\|v\|_0\|w\|_1; \\ c\|u\|_2\|v\|_1\|w\|_0; \\ c\|u\|_1\|v\|_2\|w\|_0; \\ c\|u\|_0\|v\|_2\|w\|_1; \end{cases} \quad d\leq 4. \tag{4.10}$$

Note that (4.5), (4.7), and (4.8) enable us to obtain global error estimates in the two-dimensional case.

4.3 The SAV schemes and stability results

In this section, we construct semi-discrete and fully discrete SAV schemes for the incompressible Navier-Stokes equations, and establish stability results for both semi-discrete and fully discrete schemes. More precisely, we shall prove uniform L^2 bound for the SAV scheme based on the dissipation law (4.6) in 3D case, and prove a uniform H^1 bound for the SAV scheme based on the dissipation law (4.7) in 2D case.

4.3.1 The SAV schemes

Following the ideas in [13] for the general dissipative systems, we construct below unconditionally energy stable schemes for (4.1).

For Navier-Stokes equations with periodic boundary conditions, we can explicitly eliminate the pressure from (4.1). Indeed, taking the divergence on both sides of (4.1), we find

$$-\Delta p = \nabla\cdot(u\cdot\nabla u), \tag{4.11}$$

from which we derive

$$\begin{aligned}\nabla p &= \nabla\Delta^{-1}\Delta p = -\nabla\Delta^{-1}\nabla\cdot(u\cdot\nabla u)\\ &= -\nabla\nabla\cdot\Delta^{-1}(u\cdot\nabla u) = -(\Delta+\nabla\times\nabla\times)\Delta^{-1}(u\cdot\nabla u)\\ &= -u\cdot\nabla u - \nabla\times\nabla\times\Delta^{-1}(u\cdot\nabla u) = -u\cdot\nabla u - \mathbf{A}(u\cdot\nabla u),\end{aligned} \tag{4.12}$$

where $\mathbf{A}$ is defined in (4.2). Hence, (4.1) is equivalent to (4.11) and

$$\frac{\partial u}{\partial t} - \nu\Delta u - \mathbf{A}(u\cdot\nabla u) = 0. \tag{4.13}$$

In order to apply the SAV approach, we introduce a SAV, $r(t) = E(u(t))+1$, and expand (4.13) as

$$\frac{\partial u}{\partial t} - \nu\Delta u - \mathbf{A}(u\cdot\nabla u) = 0, \tag{4.14a}$$

$$\frac{dE}{dt} = \begin{cases} -\nu\frac{r(t)}{E(u(t))+1}\|\Delta u\|^2, & d=2,\\ -\nu\frac{r(t)}{E(u(t))+1}\|\nabla u\|^2, & d=3,\end{cases} \tag{4.14b}$$

where

$$E(u) = \begin{cases} \frac{1}{2}\|\nabla u\|^2, & d=2,\\ \frac{1}{2}\|u\|^2, & d=3.\end{cases} \tag{4.15}$$

We construct below semi-discrete and fully discrete schemes for the expanded system (4.14).

Semi-discrete SAV schemes

We consider first the time discretization of (4.14) based on the implicit-explicit BDF-k formulae in the following unified form:

Given r^n, u^{j} ($\mathrm{j} = n, n-1, \cdots, n-k+1$), we compute $\bar{u}^{n+1}$, r^{n+1}, p^{n+1}, ξ^{n+1} and u^{n+1} consecutively by

$$\frac{\alpha_k \bar{u}^{n+1} - A_k(\bar{u}^n)}{\delta t} - \nu \Delta \bar{u}^{n+1} - \mathbf{A}(B_k(u^n) \cdot \nabla B_k(u^n)) = 0, \tag{4.16a}$$

$$\frac{1}{\delta t}\big(r^{n+1} - r^n\big) = \begin{cases} -\nu \frac{r^{n+1}}{E(\bar{u}^{n+1})+1} \|\Delta \bar{u}^{n+1}\|^2, & d = 2, \\ -\nu \frac{r^{n+1}}{E(\bar{u}^{n+1})+1} \|\nabla \bar{u}^{n+1}\|^2, & d = 3; \end{cases} \tag{4.16b}$$

$$\xi^{n+1} = \frac{r^{n+1}}{E(\bar{u}^{n+1})}; \tag{4.16c}$$

$$u^{n+1} = \eta_k^{n+1} \bar{u}^{n+1} \ \text{ with } \eta_k^{n+1} = 1 - (1 - \xi^{n+1})^k. \tag{4.16d}$$

Whenever pressure is needed, it can be computed from

$$\Delta p^{n+1} = -\nabla \cdot (u^{n+1} \cdot \nabla u^{n+1}). \tag{4.17}$$

In the above, α_k, the operators A_k and B_k ($k = 1, 2, 3, 4, 5$) are given by:

first-order:

$$\alpha_1 = 1, \quad A_1(u^n) = u^n, \quad B_1(\bar{u}^n) = \bar{u}^n; \tag{4.18}$$

second-order:

$$\alpha_2 = \frac{3}{2}, \quad A_2(u^n) = 2u^n - \frac{1}{2}u^{n-1}, \quad B_2(\bar{u}^n) = 2\bar{u}^n - \bar{u}^{n-1}; \tag{4.19}$$

third-order:

$$\alpha_3 = \frac{11}{6}, \quad A_3(u^n) = 3u^n - \frac{3}{2}u^{n-1} + \frac{1}{3}u^{n-2}, \quad B_3(\bar{u}^n) = 3\bar{u}^n - 3\bar{u}^{n-1} + \bar{u}^{n-2}; \tag{4.20}$$

fourth-order:

$$\alpha_4 = \frac{25}{12}, \ A_4(u^n) = 4u^n - 3u^{n-1} + \frac{4}{3}u^{n-2} - \frac{1}{4}u^{n-3}, \ B_4(\bar{u}^n) = 4\bar{u}^n - 6\bar{u}^{n-1} + 4\bar{u}^{n-2} - \bar{u}^{n-3}; \tag{4.21}$$

fifth-order:

$$\alpha_5 = \frac{137}{60}, \quad A_5(u^n) = 5u^n - 5u^{n-1} + \frac{10}{3}u^{n-2} - \frac{5}{4}u^{n-3} + \frac{1}{5}u^{n-4}, \\ B_5(\bar{u}^n) = 5\bar{u}^n - 10\bar{u}^{n-1} + 10\bar{u}^{n-2} - 5\bar{u}^{n-3} + \bar{u}^{n-4}. \tag{4.22}$$

Several remarks are in order:

- We observe from (4.16b) that r^{n+1} is a first-order approximation to $E(u(\cdot, t_{n+1})) + 1$ which implies that ξ^{n+1} is a first-order approximation to 1.
- (4.16a) is a kth-order approximation to (4.13) with kth-order BDF for the linear terms and kth-order Adams-Bashforth extrapolation for the nonlinear terms. Hence, $\bar{u}^{n+1}$ is a kth-order approximation to $u(\cdot, t^{n+1})$, which, along with (4.16b) and (4.16a), implies that u^{n+1} and p^{n+1} are kth-order approximations for $u(\cdot, t^{n+1})$ and $p(\cdot, t^{n+1})$.
- The main computational cost is to solve the Poisson type equation (4.16a).

Fully discrete schemes with Fourier spectral method in space

We now consider $\Omega = [0, L_x) \times [0, L_y) \times [0, L_z)$ with periodic boundary conditions. We partition the domain $\Omega = (0, L_x) \times (0, L_y) \times (0, L_z)$ uniformly with size $h_x = L_x/N_x, h_y = L_y/N_y, h_z = L_z/N_z$ and Nx, Ny, Nz are positive even integers. Then the Fourier approximation space can be defined as

$$S_N = \text{span}\{\mathrm{e}^{\mathrm{i}\xi_\mathrm{j} x}\mathrm{e}^{\mathrm{i}\eta_k y}\mathrm{e}^{\mathrm{i}\tau_l z} : -\frac{N_x}{2} \le \mathrm{j} \le \frac{N_x}{2} - 1, -\frac{N_y}{2} \le k \le \frac{N_y}{2} - 1, -\frac{N_z}{2} \le l \le \frac{N_z}{2} - 1\} \backslash R,$$

where $\mathrm{i} = \sqrt{-1}$, $\xi_\mathrm{j} = 2\pi\mathrm{j}/L_x$, $\eta_k = 2\pi k/L_y$ and $\tau_l = 2\pi l/L_z$. Then, any function $u(x, y, z) \in L^2(\Omega)$ can be approximated by:

$$u(x, y, z) \approx u_N(x, y, z) = \sum_{\mathrm{j}=-\frac{Nx}{2}}^{\frac{Nx}{2}-1} \sum_{k=-\frac{Ny}{2}}^{\frac{Ny}{2}-1} \sum_{l=-\frac{Nz}{2}}^{\frac{Nz}{2}-1} \hat{u}_{\mathrm{j},k,l}\mathrm{e}^{\mathrm{i}\xi_\mathrm{j} x}\mathrm{e}^{\mathrm{i}\eta_k y}\mathrm{e}^{\mathrm{i}\tau_l z},$$

with the Fourier coefficients defined as

$$\hat{u}_{\mathrm{j},k,l} = \frac{1}{|\Omega|}\int_\Omega u\mathrm{e}^{-\mathrm{i}(\xi_\mathrm{j}x+\eta_k y+\tau_l z)}dx.$$

In the following, we fix $N_x = N_y = N_z = N$ for simplicity.

Define the L^2-orthogonal projection operator $\Pi_N : L^2(\Omega) \to S_N$ by

$$(\Pi_N u - u, \Psi) = 0, \quad \forall\, \Psi \in S_N, \quad u \in L^2(\Omega),$$

then we have the following approximation results (cf. [76]): For any $0 \le k \le m$, there exists a constant C such that

$$\|\Pi_N u - u\|_k \le C\|u\|_m N^{k-m}, \forall\, u \in H_p^m(\Omega). \tag{4.23}$$

We are now ready to describe our fully discrete schemes.

Given r^n and $u_N^\mathrm{j} \in S_N$ for $\mathrm{j} = n, ..., n-k+1$, we compute $\bar{u}_N^{n+1}$, r^{n+1}, p_N^{n+1}, ξ^{n+1} and u_N^{n+1} consecutively by

$$(\frac{\alpha_k\bar{u}_N^{n+1} - A_k(\bar{u}_N^n)}{\delta t}, v_N) + \nu(\nabla\bar{u}_N^{n+1}, \nabla v_N) - (\mathbf{A}(B_k(u_N^n)\cdot\nabla B_k(u_N^n)), v_N) = 0, \quad \forall v_N \in S_N \tag{4.24a}$$

$$\frac{1}{\delta t}(r^{n+1} - r^n) = \begin{cases} -\nu\frac{r^{n+1}}{E(\bar{u}_N^{n+1})+1}\|\Delta\bar{u}_N^{n+1}\|^2, & d = 2, \\ -\nu\frac{r^{n+1}}{E(\bar{u}_N^{n+1})+1}\|\nabla\bar{u}_N^{n+1}\|^2, & d = 3; \end{cases} \tag{4.24b}$$

$$\xi^{n+1} = \frac{r^{n+1}}{E(\bar{u}_N^{n+1})+1}; \tag{4.24c}$$

$$u_N^{n+1} = \eta_k^{n+1}\bar{u}_N^{n+1} \ \text{ with } \eta_k^{n+1} = 1-(1-\xi^{n+1})^k, \tag{4.24d}$$

where α_k, the operators A_k and B_k $(k = 1, 2, 3, 4, 5)$ are given in (4.18)-(4.22).

Note that Fourier approximation of Poisson type equations leads to diagonal matrix in the frequency space, so the above scheme can be efficiently implemented as follows:

1. Compute $\bar{u}_N^{n+1}$ from (4.24a), which is a Poisson-type equation;

2. With $\bar{u}_N^{n+1}$ known, determine r^{n+1} explicitly from (4.24b);

3. Compute ξ^{n+1}, η_k^{n+1} and u_N^{n+1} from (4.24c) and (4.24d), goto the next step.

Finally, whenever pressure is needed, it can be computed from

$$\Delta p_N^{n+1} = -\Pi_N \nabla \cdot (u_N^{n+1} \cdot \nabla u_N^{n+1}). \tag{4.25}$$

4.3.2 Stability results

We have the following results concerning the stability of the above schemes.

Theorem 4.3.1. *Let $u_0 \in V \cap H_p^2$ if $d = 2$ and $u_0 \in V$ if $d = 3$. Let $\{r^k, \xi^k, \bar{u}_N^k, u_N^k\}$ be the solution of the fully discrete scheme (4.24). Then, given $r^n \geq 0$, we have $r^{n+1} \geq 0$, $\xi^{n+1} \geq 0$, and for any k, the scheme (4.24) is unconditionally energy stable in the sense that*

$$r^{n+1} - r^n = \begin{cases} -\delta t \nu \xi^{n+1} \|\Delta \bar{u}_N^{n+1}\|^2 \leq 0, & d = 2, \\ -\delta t \nu \xi^{n+1} \|\nabla \bar{u}_N^{n+1}\|^2 \leq 0, & d = 3, \end{cases} \quad \forall n. \tag{4.26}$$

Furthermore, there exists $M_k > 0$ such that

$$\begin{aligned} &\|\nabla u_N^{n+1}\|^2 \leq M_k^2, \qquad d = 2, \\ &\|u_N^{n+1}\|^2 \leq M_k^2, \qquad d = 3, \end{aligned} \quad \forall n. \tag{4.27}$$

Same results hold for the semi-discrete schemes (4.16) with $\bar{u}_N^{n+1}$ and u_N^{n+1} in (4.26) and (4.27) be replaced by $\bar{u}^{n+1}$ and u^{n+1}.

Proof. Since the proofs for the fully discrete scheme (4.24) and for the semi-discrete scheme (4.16) are essentially the same, we shall only give the proof for the fully discrete scheme (4.24) below.

Given $r^n \geq 0$. Since $E(\bar{u}_N^{n+1}) > 0$, it follows from (4.24b) that

$$r^{n+1} = \begin{cases} \dfrac{r^n}{1+\delta t \nu \frac{\|\Delta \bar{u}_N^{n+1}\|^2}{E(\bar{u}_N^{n+1})+1}} \geq 0, & d = 2, \\ \dfrac{r^n}{1+\delta t \nu \frac{\|\nabla \bar{u}_N^{n+1}\|^2}{E(\bar{u}_N^{n+1})+1}} \geq 0, & d = 3. \end{cases}$$

Then we derive from (4.24c) that $\xi^{n+1} \geq 0$ and obtain (4.26).

Denote $M := r^0 = E[u(\cdot,0)]$, then (4.26) implies $r^n \leq M, \forall n$. It then follows from (4.24c) that

$$|\xi^{n+1}| = \frac{r^{n+1}}{E(\bar{u}_N^{n+1}) + 1} \leq \begin{cases} \frac{2M}{\|\nabla \bar{u}_N^{n+1}\|^2 + 2}, & d = 2, \\ \frac{2M}{\|\bar{u}_N^{n+1}\|^2 + 2}, & d = 3. \end{cases} \tag{4.28}$$

Since $\eta_k^{n+1} = 1 - (1 - \xi^{n+1})^k$, we have $\eta_k^{n+1} = \xi^{n+1} P_{k-1}(\xi^{n+1})$ with P_{k-1} being a polynomial of degree $k-1$. Then, we derive from (4.28) that there exists $M_k > 0$ such that

$$|\eta_k^{n+1}| = |\xi^{n+1} P_{k-1}(\xi^{n+1})| \leq \begin{cases} \frac{M_k}{\|\nabla \bar{u}_N^{n+1}\|^2 + 2}, & d = 2, \\ \frac{M_k}{\|\bar{u}_N^{n+1}\|^2 + 2}, & d = 3, \end{cases}$$

which, along with $u_N^{n+1} = \eta_k^{n+1} \bar{u}_N^{n+1}$, implies

$$\begin{aligned} &\|\nabla u_N^{n+1}\|^2 = (\eta_k^{n+1})^2 \|\nabla \bar{u}_N^{n+1}\|^2 \leq \left(\tfrac{M_k}{\|\nabla \bar{u}_N^{n+1}\|^2 + 2}\right)^2 \|\nabla \bar{u}_N^{n+1}\|^2 \leq M_k^2, \qquad d = 2, \\ &\|u_N^{n+1}\|^2 = (\eta_k^{n+1})^2 \|\bar{u}_N^{n+1}\|^2 \leq \left(\tfrac{M_k}{\|\bar{u}_N^{n+1}\|^2 + 2}\right)^2 \|\bar{u}_N^{n+1}\|^2 \leq M_k^2, \qquad d = 3. \end{aligned}$$

□

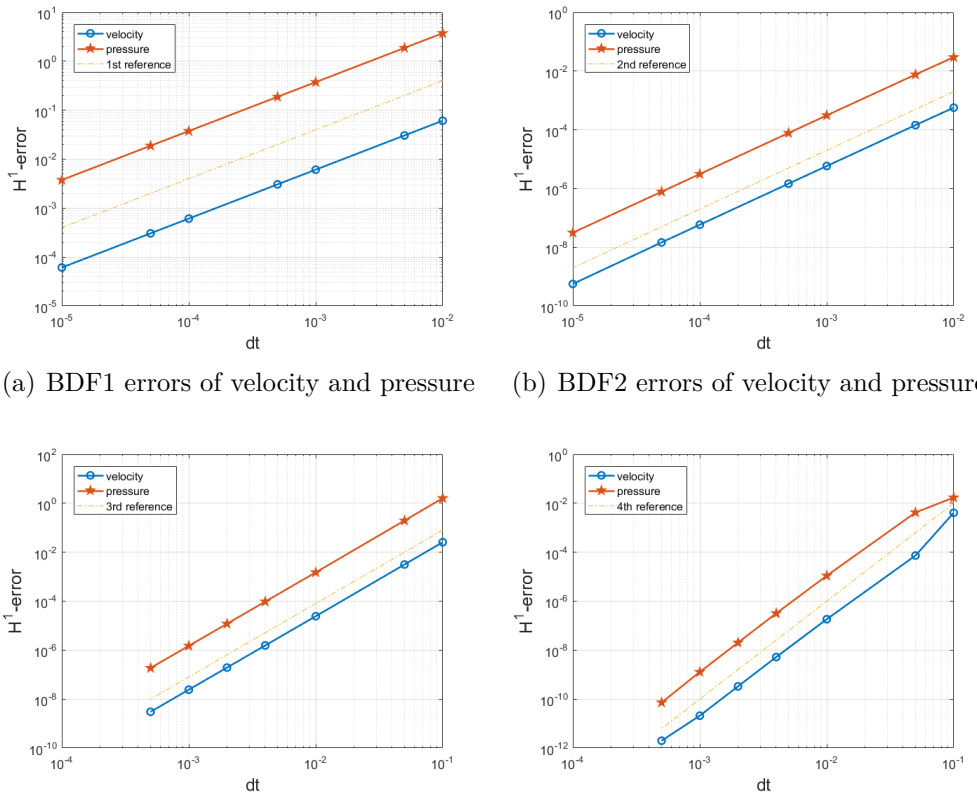
4.3.3 Numerical examples

Before we start the error analysis, we provide numerical examples to demonstrate the convergence rates and compare the performance of the schemes with different orders on a classical benchmark problem.

Example 1: Convergence test. Consider the Navier-Stokes equations (4.1) with an external forcing f in $\Omega = (0,2) \times (0,2)$ with periodic boundary condition such that the exact solution is given by

$$\begin{aligned} u_1(x,y) &= \boldsymbol{\pi} \exp(\sin(\boldsymbol{\pi} x)) \exp(\sin(\boldsymbol{\pi} y)) \cos(\boldsymbol{\pi} y) \sin^2(t); \\ u_2(x,y) &= -\boldsymbol{\pi} \exp(\sin(\boldsymbol{\pi} x)) \exp(\sin(\boldsymbol{\pi} y)) \cos(\boldsymbol{\pi} x) \sin^2(t); \\ p(x,y) &= \exp(\cos(\boldsymbol{\pi} x) \sin(\boldsymbol{\pi} y)) \sin^2(t). \end{aligned}$$

We set $\nu = 1$ in (4.1), and use the Fourier spectral method with 40×40 modes for space discretization so that the spatial discretization error is negligible with respect to the time discretization error. In Figures 4.1, we plot the convergence rate of the H^1 error for the velocity and the pressure at $T = 1$ by using first- to fourth-order schemes. We observe the expected convergence rates for both the velocity and the pressure.



(a) BDF1 errors of velocity and pressure (b) BDF2 errors of velocity and pressure

(c) BDF3 errors of velocity and pressure (d) BDF4 errors of velocity and pressure

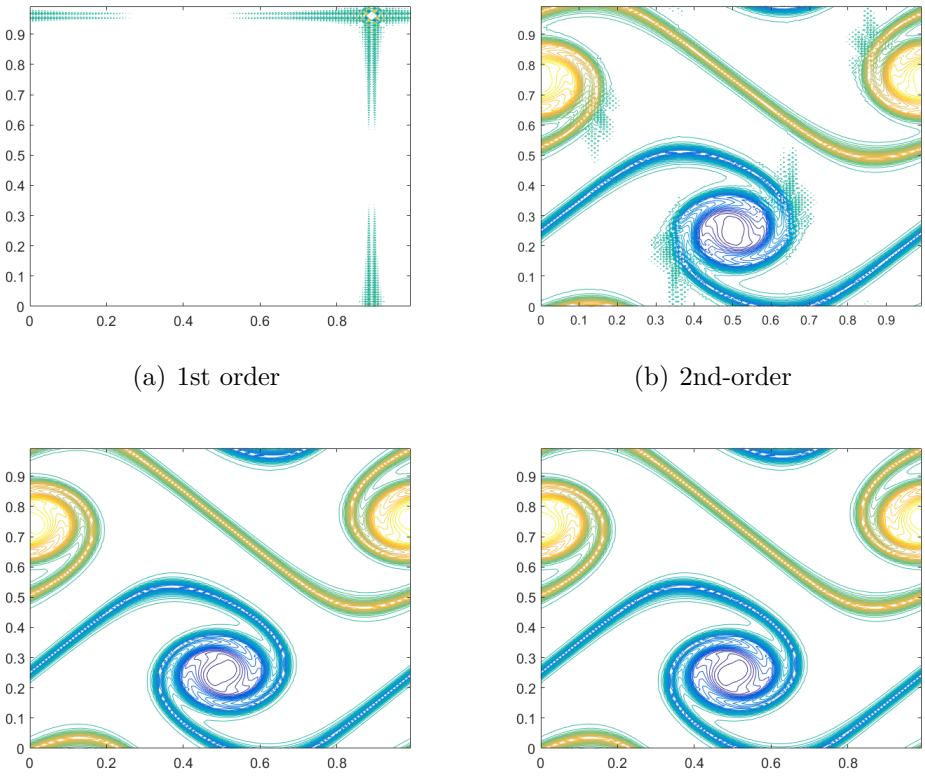
Figure 4.1. Convergence test for the Navier-stokes equations using SAV/BDFk ($k = 1, 2, 3, 4$)

Example 2: Double shear layer problem [77]–[79]. Consider the Navier-Stokes equations (4.1) in $\Omega = (0,1) \times (0,1)$ with periodic boundary conditions and the initial condition given by

$$u_1(x,y,0) = \begin{cases} \tanh(\rho(y-0.25)), & y \leq 0.5 \\ \tanh(\rho(0.75-y)), & y > 0.5 \end{cases},$$

$$u_2(x,y,0) = \delta \sin(2\pi x),$$

where ρ determines the slope of the shear layer and δ represents the size of the perturbation. In our simulations, we fix $\delta = 0.05$. We first test a *thick layer problem* by choosing $\rho = 30$



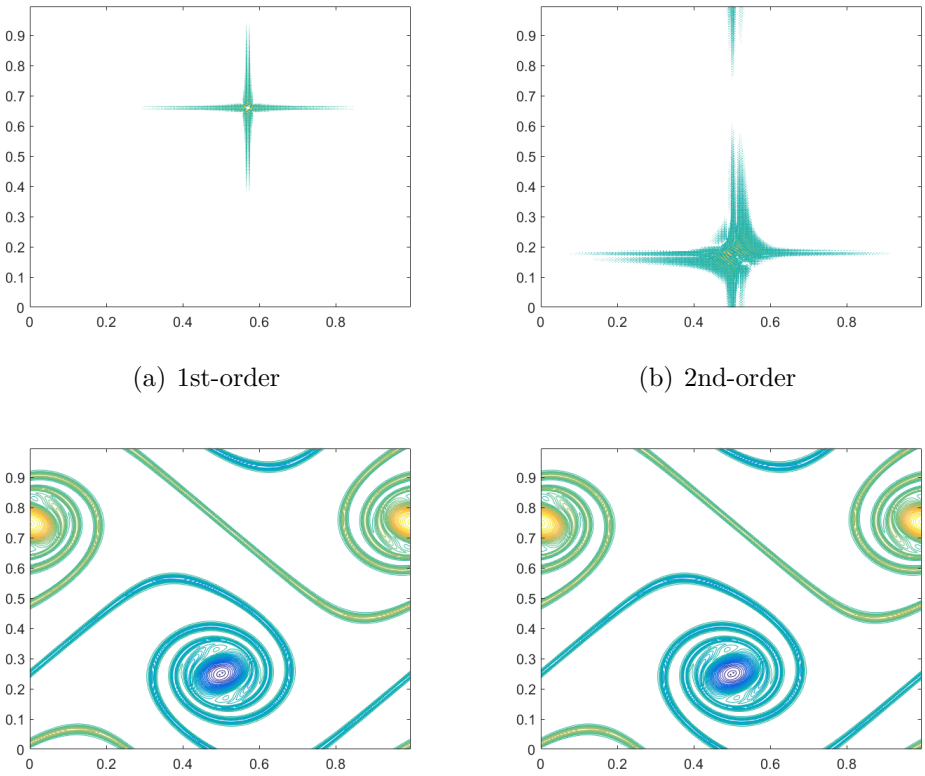
(a) 1st order (b) 2nd-order

(c) 3rd-order (d) 4th-order

Figure 4.2. Thick layer problem: vorticity contours at T=1.2 with $\rho = 30$, $\nu = 0.0001$ and $\delta t = 8 \times 10^{-4}$

and $\nu = 0.0001$. We use the Fourier spectral method with 128×128 modes for the space

discretization, and set $\delta t = 8 \times 10^{-4}$. In Figures 4.2, we show the vorticity contours at $T = 1.2$ obtained with first- to fourth-order schemes. We observe that correct solution is obtained with the third- and fourth-order schemes while the first-order scheme gives totally wrong result and the second-order scheme leads to inaccurate result.



(a) 1st-order (b) 2nd-order

(c) 3rd-order (d) 4th-order

Figure 4.3. Thin layer problem: vorticity contours at T=1.2 with $\rho = 100$, $\nu = 0.00005$ and $\delta t = 3 \times 10^{-4}$

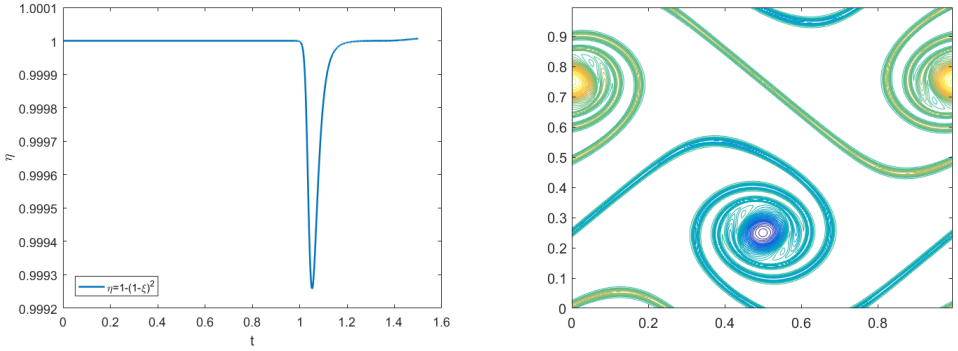
Next, we test a *thin layer problem* by choosing $\rho = 100$ and $\nu = 0.00005$. We use first- to the fourth-order schemes with 256×256 Fourier modes and $\delta t = 3 \times 10^{-4}$. In Figures 4.3, we plot the vorticity contours at $T = 1.2$. We observe that correct solutions are obtained with the third- and fourth-order schemes while first- and second-order schemes lead to wrong results.

In order to examine the effect of SAV approach, we plot in Figure 4.4 evolution of the SAV factor $\eta = 1 - (1 - \xi)^2$ and the vorticity contours at $T = 1.2$, computed with the

second-order scheme with $\delta t = 2.5 \times 10^{-4}$. We observe that at around $t = 1.05$, where the usual semi-implicit second-order scheme blows up, the SAV factor dips slightly to allow the scheme continue to produce correct simulation.

These two tests indicate that for high Reynolds number flows with complex structures, higher-order schemes are preferred over lower-order schemes, as much smaller time steps have to be used to obtain correct solutions with lower-order schemes.

Note that if we use the usual semi-implicit schemes with the same time steps in the above tests, the first- and second-order schemes would blow up. So the SAV approach can effectively prevent the numerical solution from blowing up although sufficient small time steps are needed to capture the correct solution. Thus, one is advised to adopt a suitable adaptive time stepping to take full advantage of the SAV schemes.



(a) Evolution of η (b) Vorticity contours at $t = 1.2$

Figure 4.4. Thin layer problem: second-order scheme with $\rho = 100$, $\nu = 0.00005$ and $\delta t = 2.5 \times 10^{-4}$

4.4 Error analysis

In this section, we carry out a unified error analysis for the fully discrete schemes (4.24) with $1 \leq k \leq 5$, and state, as corollaries, similar results for the semi-discrete schemes (4.16).

We denote

$$
\begin{aligned}
&t^n = n\,\delta t, \quad s^n = r^n - r(t^n),\\
&\bar{\mathrm{e}}_N^n = \bar{u}_N^n - \Pi_N u(\cdot, t^n), \quad \mathrm{e}_N^n = u_N^n - \Pi_N u(\cdot, t^n), \quad \mathrm{e}_\Pi^n = \Pi_N u(\cdot, t^n) - u(\cdot, t^n),\\
&\bar{\mathrm{e}}^n = \bar{u}_N^n - u(\cdot, t^n) = \bar{\mathrm{e}}_N^n + \mathrm{e}_\Pi^n, \quad \mathrm{e}^n = u_N^n - u(\cdot, t^n) = \mathrm{e}_N^n + \mathrm{e}_\Pi^n.
\end{aligned}
$$

To simplify the notations, we dropped the dependence on N for $\bar{\mathrm{e}}^n$ and e^n in the above, and will do so for some other quantities in the sequel.

4.4.1 Several useful lemmas

We will frequently use the following two discrete versions of the Gronwall lemma.

Theorem 4.4.1. ***(Discrete Gronwall Lemma 1 [32])*** *Let y^k, h^k, g^k, f^k be four nonnegative sequences satisfying*

$$
y^n + \delta t \sum_{k=0}^{n} h^k \le B + \delta t \sum_{k=0}^{n} (g^k y^k + f^k) \quad \textit{with} \quad \delta t \sum_{k=0}^{T/\delta t} g^k \le M, \, \forall\, 0 \le n \le T/\delta t.
$$

We assume $\delta t\, g^k < 1$ for all k, and let $\sigma = \max_{0\le k\le T/\delta t}(1-\delta t g^k)^{-1}$. Then

$$
y^n + \delta t \sum_{k=1}^{n} h^k \le \exp(\sigma M)(B + \delta t \sum_{k=0}^{n} f^k), \ \forall\, n \le T/\delta t.
$$

Theorem 4.4.2. ***(Discrete Gronwall Lemma 2 [80])*** *Let , a_n, b_n, c_n, and d_n be four nonnegative sequences satisfying*

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

where C and τ are two positive constants. Then

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

Based on Dahlquist's G-stability theory, Nevanlinna and Odeh [14] proved the following result which plays an essential role in our error analysis.

Theorem 4.4.3. *For $1 \le k \le 5$, there exist $0 \le \tau_k < 1$, a positive definite symmetric matrix $G = (g_{\mathrm{ij}}) \in \mathcal{R}^{k,k}$ and real numbers $\delta_0, ..., \delta_k$ such that*

$$\Big(\alpha_k u^{n+1} - A_k(u^n), u^{n+1} - \tau_k u^n\Big) = \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(u^{n+1+\mathrm{i}-k}, u^{n+1+\mathrm{j}-k}) - \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(u^{n+\mathrm{i}-k}, u^{n+\mathrm{j}-k}) + \|\sum_{\mathrm{i}=0}^{k} \delta_{\mathrm{i}} u^{n+1+\mathrm{i}-k}\|^2,$$

where the smallest possible values of τ_k are

$$\tau_1 = \tau_2 = 0, \quad \tau_3 = 0.0836, \quad \tau_4 = 0.2878, \quad \tau_5 = 0.8160,$$

and α_k, A_k are defined in (6.15)-(4.22).

We also recall the following lemma [81] which will be used to prove local error estimates in the three-dimensional case.

Theorem 4.4.4. *Let $\phi : (0, \infty) \to (0, \infty)$ be continuous and increasing, and let $M > 0$. Given T_* such that $0 < T_* < \int_M^\infty dz/\phi(z)$, there exists $C_* > 0$ independent of $\delta t > 0$ with the following property. Suppose that quantities z_n, $w_n \ge 0$ satisfy*

$$z_n + \sum_{k=0}^{n-1} \delta t w_k \le y_n := M + \sum_{k=0}^{n-1} \delta t \phi(z_k), \ \forall n \le n_*.$$

with $n_* \delta t \le T_*$. Then $y_{n_*} \le C_*$.

4.4.2 Error analysis for the velocity in 2D

Theorem 4.4.5. *Let $d = 2$, $T > 0$, $u_0 \in V \cap H_p^m$ with $m \ge 3$ and u be the solution of (4.1). We assume that $\bar{u}_N^{\mathrm{i}}$ and u_N^{i} $(\mathrm{i} = 1, \cdots, k-1)$ are computed with a proper initialization procedure such that*

$$\begin{aligned} &\|\bar{u}_N^{\mathrm{i}} - u(\cdot, t_{\mathrm{i}})\|_1, \ \|u_N^{\mathrm{i}} - u(t_{\mathrm{i}})\|_1 = O(\delta t^k + N^{1-m}), \\ &\|\bar{u}_N^{\mathrm{i}} - u(\cdot, t_{\mathrm{i}})\|_2, \ \|u_N^{\mathrm{i}} - u(t_{\mathrm{i}})\|_2 = O(\delta t^k + N^{2-m}), \end{aligned} \quad \mathrm{i} = 1, 2, 3, 4, 5. \tag{4.29}$$

Let $\bar{u}_N^{n+1}$ *and* u_N^{n+1} *be computed with the* k*th-order scheme* (4.24) $(1 \le k \le 5)$*, and*

$$\eta_1^{n+1} = 1 - (1-\xi^{n+1})^2, \quad \eta_k^{n+1} = 1 - (1-\xi^{n+1})^k \ (k = 2, 3, 4, 5).$$

Then for $n+1 \le T/\delta t$ *with* $\delta t \le \frac{1}{1+2^{k+2}C_0^{k+1}}$ *and* $N \ge 2^{k+2}C_\Pi^{k+1} + 1$*, we have*

$$\|\bar{u}_N^n - u(\cdot, t^n)\|_1^2,\ \|u_N^n - u(\cdot, t^n)\|_1^2 \le C\delta t^{2k} + CN^{2(1-m)},$$

and

$$\delta t \sum_{q=0}^{n} \|\bar{u}_N^{q+1} - u(\cdot, t^{q+1})\|_2^2,\ \delta t \sum_{q=0}^{n} \|u_N^{q+1} - u(\cdot, t^{q+1})\|_2^2 \le C\delta t^{2k} + CN^{2(2-m)}.$$

where the constants C_0*,* C_Π *and* C *are dependent on* T*,* Ω*, the* $k \times k$ *matrix* $G = (g_{ij})$ *in Theorem 4.4.3 and the exact solution* u*, but are independent of* δt *and* N*.*

Proof. It is shown in [75] that in the periodic case, $u_0 \in H_p^m$ implies that $u(\cdot, t) \in H_p^m$ for all $t \le T$, and furthermore, it is shown in [82] that u has Gevrey class regularity. In particular, we have

$$u \in C([0,T]; H_p^m),\ m \ge 3,\ \frac{\partial^j u}{\partial t^j} \in L^2(0,T; H_p^2)\ 1 \le j \le k,\ \frac{\partial^{k+1} u}{\partial t^{k+1}} \in L^2(0,T; L_0^2). \tag{4.30}$$

To simplify the presentation, we assume $\bar{u}_N^i = u_N^i = \Pi_N u(t_i)$ and $r^i = E_1[u_N^i]$ for $i = 1, \cdots, k-1$ so that (4.29) is obviously satisfied.

The main task is to prove by induction,

$$|1 - \xi^q| \le C_0\, \delta t + C_\Pi N^{2-m},\ \forall q \le T/\delta t, \tag{4.31}$$

where the constant C_0 and C_Π will be defined in the induction process below.

Under the assumption, (4.31) certainly holds for $q = 0$. Now suppose we have

$$|1 - \xi^q| \le C_0\, \delta t + C_\Pi N^{2-m},\ \forall q \le n, \tag{4.32}$$

we shall prove below

$$|1-\xi^{n+1}| \le C_0\delta t + C_\Pi N^{2-m}. \tag{4.33}$$

We shall first consider $k = 2, 3, 4, 5$, and point out the necessary modifications for the case $k = 1$ later.

Step 1: Bounds for $\nabla\bar{u}_N^q$, $\Delta\bar{u}_N^q$ and Δu_N^q, $\forall q \le n$. We first recall the inequality

$$(a+b)^k \le 2^k(a^k+b^k), \quad \forall a, b > 0,\ k \ge 1. \tag{4.34}$$

Under the assumption (4.32), if we choose δt small enough and N large enough such that

$$\delta t \le \min\{\frac{1}{2^{k+2}C_0^k}, 1\}, \quad N \ge \max\{2^{k+2}C_\Pi^k, 1\}, \tag{4.35}$$

we have

$$1-(\frac{1}{2^{k+2}C_0^{k-1}}+\frac{N^{3-m}}{2^{k+2}C_\Pi^{k-1}}) \le |\xi^q| \le 1+(\frac{1}{2^{k+2}C_0^{k-1}}+\frac{N^{3-m}}{2^{k+2}C_\Pi^{k-1}}),\ \forall q \le n, \tag{4.36}$$

and

$$(1-\xi^q)^k \le \frac{\delta t^{k-1}}{4}+\frac{N^{k(2-m)+1}}{4},\ \forall q \le n,$$

and

$$\frac{1}{2} < 1-(\frac{\delta t^{k-1}}{4}+\frac{N^{k(2-m)+1}}{4}) \le |\eta_k^q| \le 1+\frac{\delta t^{k-1}}{4}+\frac{N^{k(2-m)+1}}{4} < 2,\ \forall q \le n.$$

Then it follows from the above and (4.27) that

$$\|\bar{u}_N^q\|_1 \le 2M_k,\ \forall q \le n. \tag{4.37}$$

Moveover, (4.26) and $m \ge 3$ imply that

$$\nu\delta t\sum_{q=1}^{n}\|\Delta\bar{u}_N^q\|^2 \le \frac{2r^0}{|\xi^q|} \le 4r^0,\ C_0 \ge 1,\ C_\Pi \ge 1. \tag{4.38}$$

and

$$\nu\delta t\sum_{q=1}^{n}\|\Delta u_N^q\|^2 \le 16r^0,\ C_0 \ge 1,\ C_\Pi \ge 1. \tag{4.39}$$

Step 2: Estimates for $\nabla\bar{\mathrm{e}}_N^{n+1}$ and $\Delta\bar{\mathrm{e}}_N^{n+1}$. By the assumptions on the exact solution u and (4.37), we can choose C large enough such that

$$\|u(t)\|_{H^2}^2 \le C,\ \forall t \le T,\ \|\bar{u}_N^q\|_1 \le C,\ \forall q \le n. \tag{4.40}$$

From (4.24a), we can write down the error equation as

$$\Big(\alpha_k\bar{\mathrm{e}}^{q+1} - A_k(\bar{\mathrm{e}}^q), v_N\Big) + \delta t\nu\Big(\nabla\bar{\mathrm{e}}^{q+1}, \nabla v_N\Big) = \Big(R_k^q, v_N\Big) + \delta t\Big(Q_k^q, v_N\Big), \quad \forall v_N \in S_N, \tag{4.41}$$

where Q_k^q and R_k^q are given by

$$Q_k^q = -\mathbf{A}\Big((B_k u^q)\cdot\nabla)B_k(u^q)\Big) + \mathbf{A}\Big(u(t^{q+1})\cdot\nabla u(t^{q+1})\Big), \tag{4.42}$$

and

$$\begin{aligned} R_k^q &= -\alpha_k u(t^{q+1}) + A_k(u(t^q)) + \delta t u_t(t^{q+1}) \\ &= \sum_{\mathrm{i}=1}^{k} a_{\mathrm{i}}\int_{t^{q+1-\mathrm{i}}}^{t^{q+1}} (t^{q+1-\mathrm{i}} - s)^k \frac{\partial^{k+1}u}{\partial t^{k+1}}(s)ds, \end{aligned} \tag{4.43}$$

with a_{i} being some fixed and bounded constants determined by the truncation errors, for example, in the case $k = 3$, we have

$$R_3^q = -3\int_{t^q}^{t^{q+1}} (t^q - s)^3 \frac{\partial^4 u}{\partial t^4}(s)ds + \frac{3}{2}\int_{t^{q-1}}^{t^{n+1}} (t^{q-1} - s)^3 \frac{\partial^4 u}{\partial t^4}(s)ds - \frac{1}{3}\int_{t^{q-2}}^{t^{n+1}} (t^{q-2} - s)^3 \frac{\partial^4 u}{\partial t^4}(s)ds.$$

Let $v_N = -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q$ in (4.41), it follows from Theorem 4.4.3 and (4.23) that

$$\begin{aligned}
&\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla \bar{\mathrm{e}}_N^{q+1+\mathrm{i}-k}, \nabla \bar{\mathrm{e}}_N^{q+1+\mathrm{j}-k}) - \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla \bar{\mathrm{e}}_N^{q+\mathrm{i}-k}, \nabla \bar{\mathrm{e}}_N^{q+\mathrm{j}-k}) \\
&\quad + \| \sum_{\mathrm{i}=0}^{k} \delta_{\mathrm{i}} \nabla \bar{\mathrm{e}}_N^{q+1+\mathrm{i}-k} \|^2 + \delta t \nu \| \Delta \bar{\mathrm{e}}_N^{q+1} \|^2 \\
&\quad = \delta t \nu (\Delta \bar{\mathrm{e}}_N^{q+1}, \tau_k \Delta \bar{\mathrm{e}}_N^q) + (R_k^q, -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q) + \delta t (Q_k^n, -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q).
\end{aligned} \tag{4.44}$$

Next, we bound the righthand side of (4.44).

It follows from (4.43) that

$$\| R_k^q \|^2 \leq C \delta t^{2k+1} \int_{t^{q+1-k}}^{t^{q+1}} \| \frac{\partial^{k+1} u}{\partial t^{k+1}}(s) \|^2 ds. \tag{4.45}$$

Therefore,

$$\begin{aligned}
\Big| \Big(R_k^q, -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q \Big) \Big| &\leq \frac{C(\epsilon)}{\delta t} \| R_k^q \|^2 + \delta t \epsilon \| -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q \|^2, \\
&\leq \frac{C(\epsilon)}{\delta t} \| R_k^q \|^2 + 2 \delta t \epsilon \| \Delta \bar{\mathrm{e}}_N^{q+1} \|^2 + 2 \delta t \epsilon \| \Delta \bar{\mathrm{e}}_N^q \|^2, \\
&\leq 2 \delta t \epsilon \| \Delta \bar{\mathrm{e}}_N^{q+1} \|^2 + 2 \delta t \epsilon \| \Delta \bar{\mathrm{e}}_N^q \|^2 + C(\epsilon) \delta t^{2k} \int_{t^{q+1-k}}^{t^{q+1}} \| \frac{\partial^{k+1} u}{\partial t^{k+1}}(s) \|^2 ds.
\end{aligned} \tag{4.46}$$

For the term with Q_k^q, we split it as

$$\begin{aligned}
(Q_k^n, -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \mathrm{e}_N^q) = &\Big(\mathbf{A} \Big([u(t^{q+1}) - B_k(u^q)] \cdot \nabla u(t^{q+1}) \Big), -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \mathrm{e}_N^q \Big) \\
&+ \Big(\mathbf{A} \Big(B_k(u^q) \cdot \nabla [u(t^{q+1}) - B_k(u(t^q))] \Big), -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q \Big) \\
&- \Big(\mathbf{A} \Big(B_k(\mathrm{e}^q) \cdot \nabla B_k(\mathrm{e}^q) \Big), -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q \Big) \\
&- \Big(\mathbf{A} \Big(B_k(u(t^q)) \cdot \nabla B_k(\mathrm{e}^q) \Big), -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q \Big).
\end{aligned} \tag{4.47}$$

We bound the terms on the right hand side of (4.47) with the help of (4.8), (4.10) and (4.40):

$$
\begin{aligned}
&\Big(\mathbf{A}\Big([u(t^{q+1}) - B_k(u^q)] \cdot \nabla u(t^{q+1})\Big), -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q\Big) \\
&\le C\|u(t^{q+1}) - B_k(u^q)\|_1 \|u(t^{q+1})\|_2\| - \Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q\| \\
&\le C(\epsilon)\|u(t^{q+1}) - B_k(u^q)\|_1^2 \|u(t^{q+1})\|_2^2 + \epsilon\| - \Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q\|^2 \\
&\le C(\epsilon)\|u(t^{q+1}) - B_k(u(t^q))\|_1^2 \|u(t^{q+1})\|_2^2 + C(\epsilon)\|B_k(\mathrm{e}^q)\|_1^2 \|u(t^{q+1})\|_2^2 + \epsilon\| - \Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q\|^2 \\
&\le C(\epsilon)\|\sum_{\mathrm{i}=1}^{k} b_{\mathrm{i}} \int_{t^{q+1-\mathrm{i}}}^{t^{q+1}} (t^{q+1-\mathrm{i}} - s)^{k-1} \frac{\partial^k u}{\partial t^k}(s) ds\|_1^2 + C(\epsilon)\|B_k(\mathrm{e}^q)\|_1^2 + 2\epsilon\|\Delta \bar{\mathrm{e}}_N^{q+1}\|^2 + 2\epsilon\|\Delta \bar{\mathrm{e}}_N^q\|^2 \\
&\le C(\epsilon)\delta t^{2k-1} \int_{t^{q+1-k}}^{t^{q+1}} \|\frac{\partial^k u}{\partial t^k}(s)\|_1^2 ds + C(\epsilon)\|B_k(\mathrm{e}^q)\|_1^2 + 2\epsilon\|\Delta \bar{\mathrm{e}}_N^{q+1}\|^2 + 2\epsilon\|\Delta \bar{\mathrm{e}}_N^q\|^2,
\end{aligned}
\tag{4.48}
$$

where b_{i} are some fixed and bounded constants determined by the truncation error. For example, in the case $k = 3$, we have

$$
\begin{aligned}
B_3(u(t^q)) - u(t^{q+1}) = &-\frac{3}{2}\int_{t^q}^{t^{q+1}} (t^q - s)^2 \frac{\partial^3 u}{\partial t^3}(s) ds + \frac{3}{2}\int_{t^{q-1}}^{t^{q+1}} (t^{q-1} - s)^2 \frac{\partial^3 u}{\partial t^3} ds \\
&- \frac{1}{2}\int_{t^{q-2}}^{t^{q+1}} (t^{q-2} - s)^2 \frac{\partial^3 u}{\partial t^3} ds.
\end{aligned}
$$

For the other terms in the righthand side of (4.47), we have

$$
\begin{aligned}
&\Big|\Big(\mathbf{A}\Big(B_k(u^q) \cdot \nabla[u(t^{q+1}) - B_k(u(t^q))]\Big), -\Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q\Big)\Big| \\
&\qquad \le C\|B_k(u^q)\|_1 \|u(t^{q+1}) - B_k(u(t^q))\|_2\| - \Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q\| \\
&\qquad \le C(\epsilon)\|B_k(u^q)\|_1^2 \|u(t^{q+1}) - B_k(u(t^q))\|_2^2 + \epsilon\| - \Delta \bar{\mathrm{e}}_N^{q+1} + \tau_k \Delta \bar{\mathrm{e}}_N^q\|^2 \\
&\qquad \le C(\epsilon)\delta t^{2k-1} \int_{t^{q+1-k}}^{t^{q+1}} \|\frac{\partial^k u}{\partial t^k}(s)\|_2^2 ds + 2\epsilon\|\Delta \bar{\mathrm{e}}_N^{q+1}\|^2 + 2\epsilon\|\Delta \bar{\mathrm{e}}_N^q\|^2;
\end{aligned}
\tag{4.49}
$$

Since $d = 2$, we can use (4.8) to obtain

$$\begin{aligned}
&\left|\left(\mathbf{A}\Big(B_k(\mathrm{e}^q)\cdot\nabla B_k(\mathrm{e}^q)\Big), -\Delta\bar{\mathrm{e}}_N^{q+1}+\tau_k\Delta\bar{\mathrm{e}}_N^q\right)\right| \\
&\quad\le C\|B_k(\bar{\mathrm{e}}^q)\|_1^{1/2}\|B_k(\bar{\mathrm{e}}^q)\|^{1/2}\|B_k(\bar{\mathrm{e}}^q)\|_2^{1/2}\|B_k(\bar{\mathrm{e}}^q)\|_1^{1/2}\| - \Delta\bar{\mathrm{e}}_N^{q+1}+\tau_k\Delta\bar{\mathrm{e}}_N^q\| \\
&\quad\le C\|B_k(\mathrm{e}^q)\|_1\|B_k(\mathrm{e}^q)\|_2\| - \Delta\bar{\mathrm{e}}_N^{q+1}+\tau_k\Delta\bar{\mathrm{e}}_N^q\| \quad \text{(true in 2d and 3d)} \\
&\quad\le C(\epsilon)\|B_k(\mathrm{e}^q)\|_1^2\|B_k(\mathrm{e}^q)\|_2^2+\epsilon\| - \Delta\bar{\mathrm{e}}_N^{q+1}+\tau_k\Delta\bar{\mathrm{e}}_N^q\|^2 \\
&\quad\le C(\epsilon)\|B_k(\mathrm{e}^q)\|_1^2\|B_k(\mathrm{e}^q)\|_2^2+2\epsilon\|\Delta\bar{\mathrm{e}}_N^{q+1}\|^2+2\epsilon\|\Delta\bar{\mathrm{e}}_N^q\|^2;
\end{aligned} \tag{4.50}$$

Thanks to (4.10), we have

$$\begin{aligned}
&\left|\left(\mathbf{A}\Big(B_k(u(t^q))\cdot\nabla B_k(\bar{\mathrm{e}}^q)\Big), -\Delta\bar{\mathrm{e}}_N^{q+1}+\tau_k\Delta\bar{\mathrm{e}}_N^q\right)\right| \\
&\quad\le C\|B_k(u(t^q))\|_2\|B_k(\mathrm{e}^q)\|_1\| - \Delta\bar{\mathrm{e}}_N^{q+1}+\tau_k\Delta\bar{\mathrm{e}}_N^q\| \\
&\quad\le C(\epsilon)\|B_k(u(t^q))\|_2^2\|B_k(\mathrm{e}^q)\|_1^2+\epsilon\| - \Delta\bar{\mathrm{e}}_N^{q+1}+\tau_k\Delta\bar{\mathrm{e}}_N^q\|^2 \\
&\quad\le C(\epsilon)\|B_k(\mathrm{e}^q)\|_1^2+2\epsilon\|\Delta\bar{\mathrm{e}}_N^{q+1}\|^2+2\epsilon\|\Delta\bar{\mathrm{e}}_N^q\|^2.
\end{aligned} \tag{4.51}$$

On the other hand, we derive from (4.34) and (4.32) that

$$|\eta_k^q - 1| \le 2^k C_0^k\,\delta t^k + 2^k C_\Pi^k N^{k(2-m)}, \quad \forall q \le n.$$

Note that $u_N^q = \eta_k^q \bar{u}_N^q$, we can estimate $\|B_k(\mathrm{e}^q)\|_1^2$ by

$$\begin{aligned}
\|B_k(\mathrm{e}^q)\|_1^2 &= \|B_k(u_N^q - \bar{u}_N^q) + B_k(\bar{\mathrm{e}}_N^q) + B_k(\mathrm{e}_\Pi^q)\|_1^2 \\
&\le CC_0^{2k}\delta t^{2k} + CC_\Pi^{2k}N^{2k(2-m)} + C\|B_k(\bar{\mathrm{e}}_N^q)\|_1^2 + C\|u(t^q)\|_m^2 N^{2-2m}.
\end{aligned} \tag{4.52}$$

Combining (4.44)-(4.52) and dropping some unnecessary terms, we arrive at

$$\begin{aligned}
&\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla \bar{\mathrm{e}}_N^{q+1+\mathrm{i}-k}, \nabla \bar{\mathrm{e}}_N^{q+1+\mathrm{j}-k}) - \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla \bar{\mathrm{e}}_N^{q+\mathrm{i}-k}, \nabla \bar{\mathrm{e}}_N^{q+\mathrm{j}-k}) + \delta t(\frac{\nu}{2} - 10\epsilon)\|\Delta \bar{\mathrm{e}}_N^{q+1}\|^2 \\
&\quad \le \delta t(\frac{\nu\tau_k^2}{2} + 10\epsilon)\|\Delta \bar{\mathrm{e}}_N^{q}\|^2 + C(\epsilon)\delta t\|B_k(\bar{\mathrm{e}}_N^{q})\|_1^2 + C(\epsilon)\delta t\|B_k(\bar{\mathrm{e}}_N^{q})\|_1^2\|B_k(\mathrm{e}_N^{q})\|_2^2 \\
&\quad + C(\epsilon)\delta t^{2k}\int_{t^{q+1-k}}^{t^{q+1}} (\|\frac{\partial^k u}{\partial t^k}(s)\|_2^2 + \|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\|^2)ds \\
&\quad + C(\epsilon)C_0^{2k}\delta t^{2k+1}(1 + \|B_k(\mathrm{e}^q)\|_2^2) + \delta t C(\epsilon)C_\Pi^{2k}N^{2k(2-m)}(1 + \|B_k(\mathrm{e}^q)\|_2^2) \\
&\quad + \delta t C(\epsilon)\|u(t^q)\|_m^2 N^{2-2m}(1 + \|B_k(\mathrm{e}^q)\|_2^2).
\end{aligned} \tag{4.53}$$

Since $\tau_k < 1$, we can choose ϵ small enough such that

$$\frac{\nu}{2} - 10\epsilon > \frac{\nu\tau_k^2}{2} + 10\epsilon + \frac{\nu(1-\tau_k^2)}{4}, \tag{4.54}$$

and then taking the sum of (4.53) on q from $k-1$ to n, noting that $G = (g_{\mathrm{ij}})$ is a symmetric positive definite matrix with minimum eigenvalue λ_G, we obtain:

$$\begin{aligned}
&\lambda_G\|\nabla \bar{\mathrm{e}}_N^{n+1}\|^2 + \frac{\delta t\nu(1-\tau_k^2)}{4}\sum_{q=0}^{n+1}\|\Delta \bar{\mathrm{e}}_N^{q}\|^2 \\
&\qquad \le \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla \bar{\mathrm{e}}_N^{n+1+\mathrm{i}-k}, \nabla \bar{\mathrm{e}}_N^{n+1+\mathrm{j}-k}) + \frac{\delta t\nu(1-\tau_k^2)}{4}\sum_{q=0}^{n+1}\|\Delta \bar{\mathrm{e}}_N^{q}\|^2 \\
&\qquad \le C\delta t\sum_{q=0}^{n}\|\bar{\mathrm{e}}_N^{q}\|_1^2(\|B_k(\mathrm{e}^q)\|_2^2 + 1) \\
&\qquad + C\delta t^{2k}\Big(\int_0^T (\|\frac{\partial^k u}{\partial t^k}(s)\|_2^2 + \|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\|^2)ds + C_0^{2k}(T + \delta t\sum_{q=0}^{n}\|B_k(\mathrm{e}^q)\|_2^2)\Big) \\
&\qquad + C\Big(C_\Pi^{2k}N^{2k(2-m)} + N^{2-2m}\Big)\Big(T + \delta t\sum_{q=0}^{n}\|B_k(\mathrm{e}^q)\|_2^2\Big).
\end{aligned} \tag{4.55}$$

Noting that (4.39) and (4.40) imply $\delta t \sum_{q=0}^{n} \|B_k(\mathrm{e}^q)\|_2^2 < C_{H^2}$ for some constant C_{H^2} depends only on the exact solution u. Applying the discrete Gronwall Lemma 4.4.2 to (4.55), we obtain

$$\begin{aligned}
&\|\bar{\mathrm{e}}_N^{n+1}\|_1^2 + \delta t \sum_{q=0}^{n+1} \|\bar{\mathrm{e}}_N^q\|_2^2 \\
&\leq C \exp\Big(C_{H^2} + 1)\Big)\delta t^{2k} \int_0^T (\|\frac{\partial^k u}{\partial t^k}(s)\|_2^2 + \|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\|^2) ds \\
&+ C \exp\Big(C_{H^2} + 1)\Big)(\delta t^{2k} C_0^{2k} + C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m})(T + C_{H^2}) \\
&\leq C_1(1 + C_0^{2k})\delta t^{2k} + C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}),
\end{aligned} \tag{4.56}$$

where C_1 is independent of δt, C_0, C_Π, and can be defined as

$$C_1 := C \exp(C_{H^2} + 1) \max\Big(\int_0^T (\|\frac{\partial^k u}{\partial t^k}(s)\|_2^2 + \|\frac{\partial^{k+1} u}{\partial t^{k+1}}(s)\|^2) ds,\ 1,\ T + C_{H^2} \Big). \tag{4.57}$$

Therefore, (4.56) implies

$$\|\bar{\mathrm{e}}_N^{n+1}\|_1^2,\ \delta t \sum_{q=0}^{n+1} \|\bar{\mathrm{e}}_N^q\|_2^2 \leq C_1(1 + C_0^{2k})\delta t^{2k} + C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}). \tag{4.58}$$

Since $\bar{\mathrm{e}}^q = \bar{\mathrm{e}}_N^q + \bar{\mathrm{e}}_\Pi^q$, it follows from the triangle inequality that

$$\|\bar{\mathrm{e}}^{n+1}\|_1^2 \leq C_1(1 + C_0^{2k})\delta t^{2k} + C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}) + C N^{2(1-m)}, \tag{4.59}$$

and

$$\delta t \sum_{q=0}^{n+1} \|\bar{\mathrm{e}}^q\|_2^2 \leq C_1(1 + C_0^{2k})\delta t^{2k} + C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}) + C N^{2(2-m)}. \tag{4.60}$$

Combining (4.40), (4.59) and (4.60), we find that, under the condition (4.35) and $m \geq 3$, we have

$$\begin{aligned}
\|\bar{u}_N^{n+1}\|_1^2,\ \delta t \sum_{q=0}^{n+1} \|\bar{u}_N^q\|_2^2 &\leq C_1(1 + C_0^{2k} \frac{1}{2^{2k(k+2)} C_0^{2k^2}}) + C_1(C_\Pi^{2k} 2^{-4k(k+1)} C_\Pi^{-4k^2} + 1) + C \\
&\leq 4C_1 + C := \bar{C}.
\end{aligned} \tag{4.61}$$

Step 3: Estimate for $|1-\xi^{n+1}|$. It follows from (4.24b) that the equation for $\{s^{\mathrm{j}}\}$ can be written as

$$s^{q+1}-s^q=\delta t\nu\Big(\|\Delta u(t^{q+1})\|^2-\frac{r^{q+1}}{E(\bar{u}_N^{q+1})+1}\|\Delta\bar{u}_N^{q+1}\|^2\Big)+T_q,\ \forall q\le n, \tag{4.62}$$

where T_q is the truncation error

$$T_q=r(t^q)-r(t^{q+1})+\delta t r_t(t^{q+1})=\int_{t^q}^{t^{q+1}}(s-t^q)r_{tt}(s)ds. \tag{4.63}$$

Taking the sum of (4.62) for q from 0 to n, and noting that $s^0=0$, we have

$$s^{n+1}=\delta t\nu\sum_{q=0}^{n}\Big(\|\Delta u(t^{q+1})\|^2-\frac{r^{q+1}}{E(\bar{u}_N^{q+1})+1}\|\Delta\bar{u}_N^{q+1}\|^2\Big)+\sum_{q=0}^{n}T_q, \tag{4.64}$$

We bound the righthand side of (4.64) as follows. By direct calculation, we have

$$r_{tt}=\int_\Omega((\nabla u)_t^2+\nabla u(\nabla u)_{tt})dx, \tag{4.65}$$

then from (4.63), we have

$$|T_q|\le C\delta t\int_{t^q}^{t^{q+1}}|r_{tt}|ds\le C\delta t\int_{t^q}^{t^{q+1}}(\|u_t\|_1^2+\|u_{tt}\|_1^2)ds,\ \forall q\le n.$$

By triangular inequality,

$$\begin{aligned}&\Big|\|\Delta u(t^{q+1})\|^2-\frac{r^{q+1}}{E(\bar{u}_N^{q+1})+1}\|\Delta\bar{u}_N^{q+1}\|^2\Big|\\&\le\|\Delta u(t^{q+1})\|^2\Big|1-\frac{r^{q+1}}{E(\bar{u}_N^{q+1})+1}\Big|+\frac{r^{q+1}}{E(\bar{u}_N^{q+1})+1}\Big|\|\Delta u(t^{q+1})\|^2-\|\Delta\bar{u}_N^{q+1}\|^2\Big|\\&:=K_1^q+K_2^q.\end{aligned} \tag{4.66}$$

It follows from (4.40) and Theorem 4.3.1 that

$$\begin{aligned} K_1^q &\le C\Big|1-\frac{r^{q+1}}{E(\bar{u}_N^{q+1})+1}\Big| \\ &= C\Big|\frac{r(t^{q+1})}{E[u(t^{q+1})]+1}-\frac{r^{q+1}}{E[u(t^{q+1})]+1}\Big|+C\Big|\frac{r^{q+1}}{E[u(t^{q+1})]+1}-\frac{r^{q+1}}{E(\bar{u}_N^{q+1})+1}\Big| \\ &\le C\Big(|E[u(t^{q+1})]-E(\bar{u}_N^{q+1})|+|s^{q+1}|\Big),\ \forall q\le n, \end{aligned} \tag{4.67}$$

and it follows from (4.40) and Theorem 4.3.1 that

$$\begin{aligned} K_2^q &\le C\Big|\|\Delta\bar{u}_N^{q+1}\|^2-\|\Delta u(t^{q+1})\|^2\Big| \\ &\le C\|\Delta\bar{u}_N^{q+1}-\Delta u(t^{q+1})\|(\|\Delta\bar{u}^{q+1}\|+\|\Delta u(t^{q+1})\|) \\ &\le C\|\Delta\bar{u}_N^{q+1}\|\|\Delta\bar{\mathrm{e}}^{q+1}\|+C\|\Delta\bar{\mathrm{e}}^{q+1}\|,\ \forall q\le n. \end{aligned} \tag{4.68}$$

We derive from the definition of $E(u)$ that

$$|E(u(t^{q+1}))-E(\bar{u}_N^{q+1})|\le\frac{1}{2}(\|\nabla u(t^{q+1})\|+\|\nabla\bar{u}_N^{q+1}\|)\|\nabla u(t^{q+1})-\nabla\bar{u}_N^{q+1}\|\le C\|\nabla\bar{\mathrm{e}}^{q+1}\|. \tag{4.69}$$

It follows from (4.60), (4.61) and the Cauchy-Schwarz inequality that

$$\begin{aligned} \delta t\sum_{q=0}^{n}\|\Delta\bar{u}_N^{q+1}\|\|\Delta\bar{\mathrm{e}}^{q+1}\| &\le\Big(\delta t\sum_{q=0}^{n}\|\Delta\bar{u}^{q+1}\|^2\delta t\sum_{q=0}^{n}\|\Delta\bar{\mathrm{e}}^{q+1}\|^2\Big)^{1/2} \\ &\le C\sqrt{C_1(1+C_0^{2k})\delta t^{2k}+C_1(C_\Pi^{2k}N^{2k(2-m)}+N^{2-2m})+N^{2(2-m)}}. \end{aligned} \tag{4.70}$$

Now, we are ready to estimate s^{n+1}. Combining the estimates obtained above, (4.64) leads to

$$\begin{aligned}
|s^{n+1}| &\le \delta t\nu \sum_{q=0}^{n} \Big| \|\nabla u(t^{q+1})\|^2 - \frac{r^{q+1}}{E(\bar{u}^{q+1})+1} \|\nabla \bar{u}_N^{q+1}\|^2 \Big| + \sum_{q=0}^{n} |T^q| \\
&\le C\delta t \sum_{q=0}^{n} |s^{q+1}| + C\delta t \sum_{q=0}^{n} \|\bar{e}^{q+1}\|_2 + C\delta t \sum_{q=0}^{n} \|\Delta \bar{u}_N^{q+1}\| \|\Delta \bar{e}^{q+1}\| \\
&+ C\delta t \int_0^{t^{n+1}} (\|u_t\|_1^2 + \|u_{tt}\|_1^2) ds \\
&\le C\sqrt{C_1(1+C_0^{2k})\delta t^{2k} + C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}) + N^{2(2-m)}} \\
&+ C\delta t \sum_{q=0}^{n} |s^{q+1}| + C\delta t.
\end{aligned} \tag{4.71}$$

Finally, applying Theorem 4.4.1 on (4.71) with $\delta t < \frac{1}{2C}$, we obtain the following estimate for s^{n+1}:

$$\begin{aligned}
|s^{n+1}| &\le C\exp((1-\delta t C)^{-1})\Big(\sqrt{C_1(1+C_0^{2k})\delta t^{2k} + C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}) + N^{2(2-m)}} + \delta t\Big) \\
&\le C_2\Big(\sqrt{C_1(1+C_0^{2k})\delta t^{2k} + C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}) + N^{2(2-m)}} + \delta t\Big) \\
&\le C_2\delta t^k \sqrt{C_1(1+C_0^{2k})} + C_2\sqrt{C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}) + N^{2(2-m)}} + C_2\delta t,
\end{aligned} \tag{4.72}$$

where $C_2 := C\exp(2)$ is independent of δt and C_0. then $\delta t < \frac{1}{2C}$ can be guaranteed by

$$\delta t < \frac{1}{C_2}. \tag{4.73}$$

Thanks to (4.58), (4.67), (4.69), (4.72) and $m \ge 3$, we have

$$\begin{aligned}
|1-\xi^{n+1}| &\le C\Big(|E[u(t^{n+1})] - E(\bar{u}^{n+1})| + |s^{n+1}|\Big) \\
&\le C(\|\nabla \bar{e}^{n+1}\| + |s^{n+1}|) \\
&\le C\sqrt{C_1(1+C_0^{2k})\delta t^{2k} + C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}) + CN^{2(1-m)}} \\
&+ C_2\delta t^k \sqrt{C_1(1+C_0^{2k})} + C_2\sqrt{C_1(C_\Pi^{2k} N^{2k(2-m)} + N^{2-2m}) + N^{2(2-m)}} + C_2\delta t \\
&\le C_3\delta t(\sqrt{1+C_0^{2k}}\delta t^{k-1} + 1) + C_3 N^{2-m}\Big(\sqrt{C_\Pi^{2k} N^{(4-2m)(k-1)} + N^{-2} + 1}\Big),
\end{aligned} \tag{4.74}$$

where the constant C_3 is independent of C_0, C_Π , δt and N. Without loss of generality, we assume $C_3 > \max\{C_1, C_2, 1\}$ to simplify the proof below.

For the cases $k = 2, 3, 4, 5$, we choose $C_0 = 2C_3$ and $\delta t \le \frac{1}{1+C_0^k}$ to obtain

$$C_3(\sqrt{1 + C_0^{2k}}\delta t^{k-1} + 1) \le C_3[(1 + C_0^k)\delta t + 1] \le 2C_3 = C_0, \tag{4.75}$$

and since $m \ge 3$, we can choose $C_\Pi = 3C_3$ and $N \ge C_\Pi^k + 1$ to obtain

$$C_3\Big(\sqrt{C_\Pi^{2k} N^{(4-2m)(k-1)} + N^{-2} + 1}\Big) \le C_3[C_\Pi^k N^{2-m} + 2] \le 3C_3 = C_\Pi. \tag{4.76}$$

For the case $k = 1$, since $\eta_1^{n+1} = 1 - (1 - \xi^{n+1})^2$, we choose $C_0 = 2C_3$ and $\delta t \le \frac{1}{1+C_0^2}$ so that

$$C_3(\sqrt{1 + C_0^4}\delta t + 1) \le C_3[(1 + C_0^2)\delta t + 1] \le 2C_3 = C_0,$$

and since $m \ge 3$, we choose $C_\Pi = 3C_3$ and $N \ge C_\Pi^2 + 1$ to obtain

$$C_3\Big(\sqrt{C_\Pi^4 N^{(4-2m)} + N^{-2} + 1}\Big) \le C_3[C_\Pi^2 N^{2-m} + 2] \le 3C_3 = C_\Pi. \tag{4.77}$$

To summarize, combining the above with (4.74), we derive from (4.74) that

$$|1 - \xi^{n+1}| \le C_0\delta t + C_\Pi N^{2-m}$$

under the conditions

$$\delta t \le \frac{1}{1 + 2^{k+2}C_0^{k+1}}, \quad , N \ge 2^{k+2}C_\Pi^{k+1} + 1 \quad 1 \le k \le 5. \tag{4.78}$$

Note that the above implies (4.35), and with $C_3 > \max\{C_1, C_2, 1\}$, it also implies (4.73). The induction process for (4.31) is complete.

We derive from (4.24d) and (4.61) that

$$\|u_N^{n+1} - \bar{u}_N^{n+1}\|_1^2 \le |\eta_k^{n+1} - 1|^2 \|\bar{u}_N^{n+1}\|_1^2 \le |\eta_k^{n+1} - 1|^2 C, \tag{4.79}$$

and

$$\begin{aligned}\delta t\sum_{q=0}^{n}\|u_N^{q+1}-\bar{u}_N^{q+1}\|_2^2 &\le \delta t\sum_{q=0}^{n}|\eta_k^{q+1}-1|^2\|\bar{u}_N^{q+1}\|_2^2\\ &\le \max_q|\eta_k^{q+1}-1|^2\delta t\sum_{q=0}^{n}\|\bar{u}_N^{q+1}\|_2^2\\ &\le \max_q|\eta_k^{q+1}-1|^2 C.\end{aligned}\tag{4.80}$$

On the other hand, we derive from (4.31) that

$$|\eta_1^{q+1}-1|\le 2^2C_0^2\delta t^2+2^2C_\Pi^2N^{2(2-m)},\qquad \forall q\le n\quad k=1,\tag{4.81a}$$

$$|\eta_k^{q+1}-1|\le 2^kC_0^k\delta t^k+2^kC_\Pi^kN^{k(2-m)},\qquad \forall q\le n\quad k=2,3,4,5.\tag{4.81b}$$

Therefore, we derive from (4.59), (4.60), (4.79), (4.80), (4.81) and the triangle inequality that

$$\|\mathrm{e}^{n+1}\|_1^2\le\|\bar{\mathrm{e}}^{n+1}\|_1^2+\|u_N^{n+1}-\bar{u}_N^{n+1}\|_1^2,$$

and

$$\|\mathrm{e}^{q+1}\|_2^2\le\|\bar{\mathrm{e}}^{q+1}\|_2^2+\|u_N^{q+1}-\bar{u}_N^{q+1}\|_2^2,\quad \forall q\le n,$$

under the condition (4.78) on δt and N. The proof is now complete since we already proved (4.59) and (4.60). □

Using exactly the same procedure above without the spatial discretization, we can prove the following result for the semi-discrete schemes (4.16). Let $d=2$, $T>0$, $u_0\in V\cap H_p^2$ and u be the solution of (4.1). We assume that $\bar{u}^{\mathrm{i}}$ and u^{i} $(\mathrm{i}=1,\cdots,k-1)$ are computed with a proper initialization procedure such that for $(\mathrm{i}=1,\cdots,k-1)$,

$$\|\bar{u}^{\mathrm{i}}-u(t_{\mathrm{i}})\|_1,\ \|u^{\mathrm{i}}-u(t_{\mathrm{i}})\|_1=O(\delta t^k);\quad \|\bar{u}^{\mathrm{i}}-u(t_{\mathrm{i}})\|_2,\ \|u^{\mathrm{i}}-u(t_{\mathrm{i}})\|_2=O(\delta t^k),\quad \mathrm{i}=1,2,3,4,5.$$

Let $\bar{u}^{n+1}$ and u^{n+1} be computed with the $k-$th order scheme (4.16) $(1\le k\le 5)$, and

$$\eta_1^{n+1}=1-(1-\xi^{n+1})^2,\quad \eta_k^{n+1}=1-(1-\xi^{n+1})^k\ (k=2,3,4,5).$$

Then for $n+1 \le T/\delta t$ and $\delta t \le \frac{1}{1+2^{k+2}C_0^{k+1}}$, we have

$$\|\bar{u}^n - u(\cdot,t^n)\|_1^2,\ \|u^n - u(\cdot,t^n)\|_1^2 \le C\delta t^{2k},$$

and

$$\delta t\sum_{q=0}^{n}\|\bar{u}^{q+1} - u(\cdot,t^{q+1})\|_2^2,\ \delta t\sum_{q=0}^{n}\|u^{q+1} - u(\cdot,t^{q+1})\|_2^2 \le C\delta t^{2k}.$$

where the constants C_0 and C are dependent on T, Ω, the $k\times k$ matrix $G=(g_{\mathrm{ij}})$ in Lemma 4.4.3 and the exact solution u, but are independent of δt.

4.4.3 Error analysis for the velocity in 3D

In the three-dimensional case, it is no longer possible to obtain the global estimates (4.37), (4.38) and (4.39) as in the two-dimensional case. Instead, we shall derive local estimates in analogy to the local existence of strong solution for the 3-D Navier-Stokes equations.

Theorem 4.4.6. *Let $d=3$, $T>0$, $u_0 \in V\cap H_p^m$ with $m\ge 3$. We assume that (4.1) admits a unique strong solution u in $C([0,T];H_p^1)\cap L^2(0,T;H_p^2)$. We assume (4.29) as in* **Theorem 2**, *and let $\bar{u}_N^{n+1}$ and u_N^{n+1} be computed using the kth-order scheme (4.24) $(1\le k\le 5)$, and*

$$\eta_1^{n+1} = 1-(1-\xi^{n+1})^2,\quad \eta_k^{n+1} = 1-(1-\xi^{n+1})^k\ (k=2,3,4,5).$$

Then, there exits $T_*>0$ such that for $0<T<T_*$, $n+1\le T/\delta t$ and $\delta t \le \frac{1}{1+2^{k+2}C_0^{k+1}}$, $N\ge 2^{k+2}C_\Pi^{k+1}+1$, we have

$$\|\bar{u}_N^n - u(\cdot,t^n)\|_1^2,\ \|u_N^n - u(\cdot,t^n)\|_1^2 \le C\delta t^{2k} + CN^{2(1-m)}, \tag{4.82}$$

and

$$\delta t\sum_{q=0}^{n}\|\bar{u}_N^{q+1} - u(\cdot,t^{q+1})\|_2^2,\ \delta t\sum_{q=0}^{n}\|u_N^{q+1} - u(\cdot,t^{q+1})\|_2^2 \le C\delta t^{2k} + CN^{2(2-m)}, \tag{4.83}$$

where the constants C_0, C_Π, C are dependent on T, Ω, the $k\times k$ matrix $G=(g_{\mathrm{ij}})$ in Lemma 4.4.3 and the exact solution u, but are independent of δt and N.

Proof. The proof follows essentially the same procedure as the proof for **Theorem 4.4.5**. However, since we only has the weak version of the stability in **Theorem 1** and (4.8) is not valid when $d = 3$, we can only get a local version of (4.37) and (4.38). To simplify the presentation, we shall only point out below the main differences with the proof for **Theorem 4.4.5**.

With $u_0 \in H_p^m$ and the existence of a unique strong solution u in $C([0,T]; H_p^1) \cap L^2(0,T; H_p^2)$, regularity results in [75], [82] imply that (4.30) is also valid in the three-dimensional case.

In **Step 1**, we still assume (4.32) holds and choose δt and N satisfies (4.35). Let $v_N = -\Delta \bar{u}^{n+1} + \tau_k \Delta \bar{u}^n$ in (4.24a), it follows from Lemma 4.4.3 that

$$\begin{aligned}
&\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla \bar{u}_N^{q+1+\mathrm{i}-k}, \nabla \bar{u}_N^{q+1+\mathrm{j}-k}) - \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla \bar{u}_N^{q+\mathrm{i}-k}, \nabla \bar{u}_N^{q+\mathrm{j}-k}) \\
&\qquad + \|\sum_{\mathrm{i}=0}^{k} \delta_{\mathrm{i}} \nabla \bar{u}_N^{q+1+\mathrm{i}-k}\|^2 + \delta t \nu \|\Delta \bar{u}_N^{q+1}\|^2 \\
&\qquad = \delta t \nu (\Delta \bar{u}_N^{q+1}, \tau_k \Delta \bar{u}_N^q) + \delta t (\mathbf{A}\Big((B_k(u_N^q) \cdot \nabla) B_k(u_N^q)\Big), -\Delta \bar{u}_N^{q+1} + \tau_k \Delta \bar{u}_N^q).
\end{aligned} \tag{4.84}$$

We now bound the right hand side of (4.84). Note that (4.35) implies

$$\frac{1}{2} < 1 - (\frac{\delta t^{k-1}}{4} + \frac{N^{k(2-m)+1}}{4}) \le |\eta_k^q| \le 1 + \frac{\delta t^{k-1}}{4} + \frac{N^{k(2-m)+1}}{4} < 2, \ \forall q \le n.$$

First, we have

$$|\delta t \nu (\Delta \bar{u}_N^{q+1}, \tau_k \Delta \bar{u}_N^q)| \le \delta t \frac{\nu}{2} \|\Delta \bar{u}_N^{q+1}\|^2 + \delta t \frac{\nu \tau_k}{2} \|\Delta \bar{u}_N^q\|^2. \tag{4.85}$$

Next, it follows from (4.9) that

$$
\begin{aligned}
&|(\mathbf{A}\Big((B_k(u_N^q)\cdot\nabla)B_k(u_N^q)\Big), -\Delta\bar{u}_N^{q+1}+\tau_k\Delta\bar{u}_N^q)| \\
&\quad\le C\|B_k(u_N^q)\|_1\|B_k(\nabla u_N^q)\|_{1/2}\| - \Delta\bar{u}_N^{q+1}+\tau_k\Delta\bar{u}_N^q\| \\
&\quad\le C\|B_k(u_N^q)\|_1\|B_k(u_N^q)\|_1^{1/2}\|B_k(u_N^q)\|_2^{1/2}\| - \Delta\bar{u}_N^{q+1}+\tau_k\Delta\bar{u}_N^q\| \\
&\quad\le C(\epsilon)\|B_k(u_N^q)\|_1^3\|B_k(u_N^q)\|_2+\epsilon\| - \Delta\bar{u}_N^{q+1}+\tau_k\Delta\bar{u}_N^q\|^2 \\
&\quad\le C(\epsilon)\|B_k(u_N^q)\|_1^6+\epsilon\|B_k(u_N^q)\|_2^2+2\epsilon\|\Delta\bar{u}_N^{q+1}\|^2+2\epsilon\|\Delta\bar{u}_N^q\|^2.
\end{aligned}
\tag{4.86}
$$

Now, combining (4.84)-(4.86) and noting that $u_N^q = \eta_k^q\bar{u}_N^q$, we find after dropping some unnecessary terms that

$$
\begin{aligned}
&\sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla\bar{u}_N^{q+1+\mathrm{i}-k},\nabla\bar{u}_N^{q+1+\mathrm{j}-k}) - \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla\bar{u}_N^{q+\mathrm{i}-k},\nabla\bar{u}_N^{q+\mathrm{j}-k}) + \delta t(\frac{\nu}{2}-2\epsilon)\|\Delta\bar{u}_N^{q+1}\|^2 \\
&\quad\le \delta t(\frac{\nu\tau_k}{2}+2\epsilon)\|\Delta\bar{u}_N^q\|^2+\epsilon\delta t\|B_k(u_N^q)\|_2^2+C(\epsilon)\delta t\|B_k(u_N^q)\|_1^6 \\
&\quad\le \delta t(\frac{\nu\tau_k}{2}+2\epsilon)\|\Delta\bar{u}_N^q\|^2+2^2\epsilon\delta t\|B_k(\bar{u}_N^q)\|_2^2+2^6C(\epsilon)\delta t\|B_k(\bar{u}_N^q)\|_1^6
\end{aligned}
\tag{4.87}
$$

Taking the sum of (4.87) for q from $k-1$ to $n-1$, noting that $G=(g_{\mathrm{ij}})$ is a symmetric positive definite matrix with the minimum eigenvalue λ_G and $\tau_k<1$, we can choose ϵ small enough such that:

$$
\begin{aligned}
&\lambda_G\|\bar{u}_N^n\|_1^2+\frac{\delta t\nu(1-\tau_k)}{4}\sum_{q=0}^{n}\|\Delta\bar{u}_N^q\|^2 \\
&\qquad\le \sum_{\mathrm{i,j}=1}^{k} g_{\mathrm{ij}}(\nabla\bar{u}^{n+\mathrm{i}-k},\nabla\bar{u}^{n+\mathrm{j}-k})+\frac{\delta t\nu(1-\tau_k)}{4}\sum_{q=0}^{n}\|\Delta\bar{u}_N^q\|^2 \\
&\qquad\le C\delta t\sum_{q=0}^{n-1}\|\bar{u}_N^q\|_1^6+M_0,
\end{aligned}
$$

where $M_0>0$ is a constant only depends on $\bar{u}_N^0,...,\bar{u}_N^k$, g_{ij}. If we define ϕ as $\phi(x)=x^6$ and let

$$
0<T_*<\int_{M_0}^{\infty} dz/\phi(z), \tag{4.88}
$$

then Theorem 4.4.4 implies that there exist $C_* > 0$ independent of δt such that

$$\|\bar{u}_N^n\|_1^2 + \delta t \sum_{q=0}^{n} \|\Delta \bar{u}_N^q\|^2 \le C_*, \quad \forall n < T_*/\delta t. \tag{4.89}$$

With (4.89) holds true, we can then prove (4.82) and (4.83) by following the same procedures in **Step 2** and **Step 3** in the proof of **Theorem 4.4.5**. □

Similarly, we can prove the following result for the semi-discrete scheme (4.16). Let $d = 3$, $T > 0$, $u_0 \in V \cap H_p^m$ with $m \ge 3$. We assume that (4.1) admits a unique strong solution u in $C([0,T]; H_p^1) \cap L^2(0,T; H_p^2)$. We assume (4.29) as in **Theorem 2**, and let $\bar{u}^{n+1}$ and u^{n+1} be computed using the kth-order schemes (4.16), and

$$\eta_1^{n+1} = 1 - (1-\xi^{n+1})^2, \quad \eta_k^{n+1} = 1 - (1-\xi^{n+1})^k \ (k = 2,3,4,5).$$

Then, there exits $T_* > 0$ such that for $0 < T < T_*$, $n+1 \le T/\delta t$ and $\delta t \le \frac{1}{1+2^{k+2}C_0^{k+1}}$, we have

$$\|\bar{u}^n - u(\cdot, t^n)\|_1^2, \ \|u^n - u(\cdot, t^n)\|_1^2 \le C\delta t^{2k},$$

and

$$\delta t \sum_{q=0}^{n} \|\bar{u}^{q+1} - u(\cdot, t^{q+1})\|_2^2, \ \delta t \sum_{q=0}^{n} \|u^{q+1} - u(\cdot, t^{q+1})\|_2^2 \le C\delta t^{2k},$$

where T_* is defined in (4.88), the constants C_0, C_Π, C are dependent on T_*, Ω, the $k \times k$ matrix $G = (g_{ij})$ in Lemma 4.4.3 and the exact solution u, but are independent of δt.

4.4.4 Error analysis for the pressure

With the established error estimates for the velocity u, the error estimate for the pressure p can be derived directly from (4.17) or (4.25).

We denote

$$\mathrm{e}_{pN}^n := p_N^n - \Pi_N p(\cdot, t^n), \ \mathrm{e}_{p\Pi}^n := \Pi_N p(\cdot, t^n) - p(\cdot, t^n), \ \text{ and } \mathrm{e}_p^n = \mathrm{e}_{pN}^n + \mathrm{e}_{p\Pi}^n$$

.

Theorem 4.4.7. *Under the same assumptions as in* **Theorem 4.4.5 and Theorem 4.4.6**, *we have*

$$\|p_N^{n+1} - p(\cdot, t^{n+1})\|^2 \leq \begin{cases} C\delta t^{2k} + CN^{2(1-m)}, \ \forall n \leq T/\delta t, & d = 2, \\ C\delta t^{2k} + CN^{2(1-m)}, \ \forall n \leq T_*/\delta t, & d = 3. \end{cases} \tag{4.90}$$

and

$$\delta t \sum_{q=0}^{n} \|\nabla(p_N^{n+1} - p(\cdot, t^{n+1}))\|^2 \leq \begin{cases} C\delta t^{2k} + CN^{2(2-m)}, \ \forall n \leq T/\delta t, & d = 2, \\ C\delta t^{2k} + CN^{2(2-m)}, \ \forall n \leq T_*/\delta t, & d = 3. \end{cases} \tag{4.91}$$

where p_N^{n+1} is computed from (4.25), T_* is defined in (4.88) and C is a constant independent of δt and N.

Proof. From (4.25), we can write down the error equation for p_N^{n+1} as

$$\left(\nabla \mathrm{e}_p^{q+1}, \nabla v_N\right) = \left(u_N^{q+1} \cdot \nabla u_N^{q+1} - u(t^{q+1}) \cdot \nabla u(t^{q+1}), \nabla v_N\right), \forall v_N \in S_N, \forall q+1 \leq n. \tag{4.92}$$

To prove (4.90), we set $v_N = \Delta^{-1}\mathrm{e}_{pN}^{q+1}$ in (4.92) to obtain

$$\begin{aligned} \|\mathrm{e}_{pN}^{q+1}\|^2 = &\left(u_N^{q+1} \cdot \nabla[u_N^{q+1} - u(t^{q+1})], \Delta^{-\frac{1}{2}}\mathrm{e}_{pN}^{q+1}\right) \\ &- \left([u(t^{q+1}) - u_N^{q+1}] \cdot \nabla u(t^{q+1}), \Delta^{-\frac{1}{2}}\mathrm{e}_{pN}^{q+1}\right) \end{aligned} \tag{4.93}$$

We can bound the righthand side of (4.93) by using (4.10), the stability result Theorem 4.3.1 and error analysis for the velocity, namely, we can obtain

$$\begin{aligned} \left|\left(u_N^{q+1} \cdot \nabla[u_N^{q+1} - u(t^{q+1})], \Delta^{-\frac{1}{2}}\mathrm{e}_{pN}^{q+1}\right)\right| &\leq C(\epsilon)\|u_N^{q+1}\|_1^2\|\|\mathrm{e}^{q+1}\|_1^2 + \epsilon\|\nabla \mathrm{e}_{pN}^{q+1}\|^2 \\ &\leq C(\epsilon)(\delta t^{2k} + N^{2(1-m)}) + \epsilon\|\mathrm{e}_{pN}^{q+1}\|^2; \end{aligned} \tag{4.94}$$

and

$$\begin{aligned} \left| - \left([u(t^{q+1}) - u_N^{q+1}] \cdot \nabla u(t^{q+1}), \Delta^{-\frac{1}{2}}\mathrm{e}_{pN}^{q+1}\right)\right| &\leq C(\epsilon)\|u(t^{q+1})\|_1^2\|\|\mathrm{e}^{q+1}\|_1^2 + \epsilon\|\nabla \mathrm{e}_{pN}^{q+1}\|^2 \\ &\leq C(\epsilon)(\delta t^{2k} + N^{2(1-m)}) + \epsilon\|\mathrm{e}_{pN}^{q+1}\|^2; \end{aligned} \tag{4.95}$$

Combining (4.93)-(4.95) with $\epsilon = \frac{1}{4}$ we obtain

$$\|\mathrm{e}_{pN}^{q+1}\|^2 \le C\delta t^{2k} + CN^{2(1-m)}, \quad \forall q \le n. \tag{4.96}$$

To prove (4.91), we set $v_N = \mathrm{e}_{pN}^{q+1}$ in (4.92) to obtain

$$\begin{aligned} \|\nabla \mathrm{e}_{pN}^{q+1}\|^2 = &\left(u_N^{q+1} \cdot \nabla[u_N^{q+1} - u(t^{q+1})], \nabla \mathrm{e}_{pN}^{q+1}\right) \\ &- \left([u(t^{q+1}) - u_N^{q+1}] \cdot \nabla u(t^{q+1}), \nabla \mathrm{e}_{pN}^{q+1}\right) \end{aligned} \tag{4.97}$$

Again, we can bound the righthand side of (4.97) in a similar fashion as in (4.94)-(4.95), namely, we can obtain

$$\begin{aligned} \left|\left(u_N^{q+1} \cdot \nabla[u_N^{q+1} - u(t^{q+1})], \nabla \mathrm{e}_{pN}^{q+1}\right)\right| &\le C(\epsilon)\|u_N^{q+1}\|_1^2 \|\|\mathrm{e}^{q+1}\|_2^2 + \epsilon\|\nabla \mathrm{e}_{pN}^{q+1}\|^2 \\ &\le C(\epsilon)\|\mathrm{e}^{q+1}\|_2^2 + \epsilon\|\nabla \mathrm{e}_{pN}^{q+1}\|^2; \end{aligned} \tag{4.98}$$

and

$$\begin{aligned} \left| - \left([u(t^{q+1}) - u_N^{q+1}] \cdot \nabla u(t^{q+1}), \nabla \mathrm{e}_{pN}^{q+1}\right)\right| &\le C(\epsilon)\|u(t^{q+1})\|_2^2 \|\|\mathrm{e}^{q+1}\|_1^2 + \epsilon\|\nabla \mathrm{e}_{pN}^{q+1}\|^2 \\ &\le C(\epsilon)(\delta t^{2k} + N^{2(1-m)}) + \epsilon\|\nabla \mathrm{e}_{pN}^{q+1}\|^2; \end{aligned} \tag{4.99}$$

Combining (4.97)-(4.99) with $\epsilon = \frac{1}{4}$, we obtain

$$\|\nabla \mathrm{e}_{pN}^{q+1}\|^2 \le C\|\mathrm{e}^{q+1}\|_2^2 + C\delta t^{2k} + CN^{2(1-m)}, \quad \forall q \le n. \tag{4.100}$$

Taking the sum of (4.53) for q from 0 to n and multiplying δt on both sides, we arrive at

$$\delta t \sum_{q=0}^{n} \|\nabla \mathrm{e}_{pN}^{q+1}\|^2 \le C\delta t \sum_{q=0}^{n} \|\mathrm{e}^{q+1}\|_2^2 + C\delta t^{2k} + CN^{2(1-m)}. \tag{4.101}$$

Now, with the estimates on $\|\mathrm{e}^n\|_2^2$ in Theorem 4.4.5 or Theorem 4.4.6, (4.101) leads to

$$\delta t \sum_{q=0}^{n} \|\nabla \mathrm{e}_{pN}^{q+1}\|^2 \le C\delta t^{2k} + CN^{2(2-m)}. \tag{4.102}$$

Finally, we can obtain (4.90) and (4.91) from (4.96), (4.102) and

$$\|\nabla \mathrm{e}_{p\Pi}^q\|^2 \leq CN^{2(1-m)}.$$

□

Similarly, we can derive the following results for the semi-discrete scheme (4.16).

Under the same assumptions as in **Corollary 1 and Corollary 2**, we have

$$\|p^{n+1} - p(\cdot, t^{n+1})\|^2 \leq \begin{cases} C\delta t^{2k}, \ \forall n \leq T/\delta t, & d = 2, \\ C\delta t^{2k}, \ \forall n \leq T_*/\delta t, & d = 3. \end{cases}$$

and

$$\delta t \sum_{q=0}^{n} \|\nabla(p^{q+1} - p(\cdot, t^{n+1}))\|^2 \leq \begin{cases} C\delta t^{2k}, \ \forall n \leq T/\delta t, & d = 2, \\ C\delta t^{2k}, \ \forall n \leq T_*/\delta t, & d = 3. \end{cases}$$

where p^{n+1} is computed from (4.17), T_* is defined in (4.88) and C is a constant independent of δt.

4.5 Conclusion of this chapter

We considered numerical approximation of the incompressible Navier-Stokes equations with periodic boundary conditions for which the pressure can be explicitly eliminated, allowing us to construct very efficient IMEX type schemes using Fourier-Galerkin approximation in space. Our high-order semi-discrete-in-time and fully discrete IMEX schemes are based on a scalar auxiliary variable (SAV) approach which enables us to derive uniform bounds for the numerical solution without any restriction on time step size. We also take advantage of an additional energy dissipation law (4.7), which is only valid for the two-dimensional Navier-Stokes equations with periodic boundary conditions, leading to a uniform bound in H^1-norm, instead of the usual L^2-norm. By using these uniform bounds and a delicate induction process, we derived global error estimates in $l^\infty(0,T;H^1) \cap l^2(0,T;H^2)$ in the two dimensional case as well as local error estimates in $l^\infty(0,T;H^1) \cap l^2(0,T;H^2)$ in the three dimensional case for our semi-discrete-in-time and fully discrete IMEX schemes up

to fifth-order. We also validated our schemes with manufactured exact solutions and with the double shear layer problem. Our numerical results for the double shear layer problem indicate that the SAV approach can effectively prevent numerical solution from blowing up, and that higher-order schemes are preferable for flows with complex structures such as the double shear layer problem with thin layers.

To the best of our knowledge, our numerical schemes are the first unconditionally stable high-order IMEX type schemes for Navier-Stokes equations without any restriction on time step size, and our error estimates are the first for any IMEX type scheme for the Navier-Stokes equations in the three-dimensional case.

While the stability results can be extended to similar schemes for the Navier-Stokes equations with non-periodic boundary conditions, it is non trivial to carry out the corresponding error analysis which will be left as a subject of future endeavor.

5. POSITIVITY/BOUND PRESERVING SAV SCHEMES: WITH APPLICATION TO SECOND ORDER EQUATION

In this chapter, we apply new SAV approach to construct high-order, linear, positivity/bound preserving and unconditionally energy stable schemes for general dissipative systems whose solutions are positivity/bound preserving. The method is based on applying a new SAV approach to the transformed system with a suitable functional transformation. In particular, we applied this method to the Poisson-Nernst-Planck equation and the Keller-Segel equation. Most of the results in this chapter are extracted from [83].

5.1 Introduction

For the PNP equations, a quite complicated entropy-based scheme with regularized free energy is constructed in [84] along with rigorous numerical analyses for a set of finite-element approximations; a mass-conservative finite difference scheme is constructed in [85]; an arbitrary-order energy dissipative schemes are constructed using a discontinuous Galerkin (DG) method for 1-D PNP systems [86]; and most recently a fully discrete positivity-preserving and energy-dissipative finite difference scheme is developed in [87]. On the other hand, there exists a large number of numerical work for the PNP equations in the electric and medical engineering literature, see, for examples, [88]–[90] and the references therein.

For the Keller-Segel equations and related models, a finite volume scheme is developed with convergence proof in [91]; a second-order positivity preserving central-upwind scheme is constructed in [92] (see also [93], [94]); finite volume methods for a Keller-Segel system are considered with discrete energy dissipation and error estimates in [95]; and a positivity-preserving and asymptotic preserving method is constructed for a reformulated Keller-Segel system in [96] [95], [97]. We refer to the aforementioned papers and the references therein for more details on existing numerical schemes for Keller-Segel equations.

Some of these numerical schemes preserve positivity and/or some form of energy dissipation under certain conditions and specific spatial discretization. Oftentimes one needs to solve nonlinear systems at each time step. Very recently, an interesting approach is proposed

to construct unconditionally energy stable and positivity/bound preserving for Keller-Segel equations in [98] and for PNP equations in [99]. However, these schemes require solving, at each time step, a nonlinear system which is a unique minimizer of a strictly convex functional. The question we would like to address in this chapter is: for PDEs which preserve positivity or bound and satisfy an energy dissipation law, how to construct numerical schemes which are linear, positivity/bound preserving and unconditionally energy stable for any consistent spatial discretization?

The recently proposed scalar auxiliary variable (SAV) approach [1], [20] is a powerful tool to design unconditionally energy stable, linear schemes to a large class of gradient flows, and has been applied successfully to many challenging problems. However, it does not have mechanism to preserve bounds or positivity. On the other hand, a common strategy to enforce solutions to preserve bounds or positivity is to use a suitable function transform. A drawback of this approach is that the transformed equation becomes very complicated that it is very difficult to construct efficient and energy stable schemes for the transformed equation.

In this chapter, we propose a new class of bound/positivity preserving and energy stable schemes by combining the SAV approach and the function transform approach:

- make a suitable function transform to ensure positivity or bound preserving;
- use a recently proposed SAV approach [13] to design linear and unconditionally energy stable schemes for the transformed equation.

Our new schemes will enjoy the following remarkable properties:

- it can be used with high-order semi-implicit (i.e., IMEX) schemes;
- it is positivity or bound preserving;
- it is unconditionally energy dissipative;
- it only requires solving one set (instead of two in the original SAV approach) decoupled linear equations with constant coefficients at each time step, so the coding and computational complexity is similar to that of semi-implicit schemes;

- for problems with mass conservation as in PNP and KS equations, it also conserves mass.

5.2 Positivity/bound preserving SAV schemes for second order nonlinear systems

In order to clearly describe our idea, we consider a semi-linear or quasilinear parabolic system in the form

$$\frac{\partial u}{\partial t} - \Delta u + g(u) = 0, \tag{5.1}$$

with either periodic or homogeneous Neumann boundary condition, where $g(u)$ is a nonlinear function. The following discussions are still valid if we replace $-\Delta$ in (5.1) with more general or higher-order linear elliptic operators.

We assume that the above system satisfies a dissipation law in the form

$$\frac{dE(u)}{dt} = -\Big(\mathcal{G}u, u\Big), \tag{5.2}$$

where $E(u)$ is a typical energy functional given by

$$E[u] = \int_\Omega \Big(\frac{1}{2}\mathcal{L}u \cdot u + F(u)\Big)dx := E_0(u) + E_1(u), \tag{5.3}$$

$\mathcal{G}$ is a non-negative operator and $\mathcal{L}$ is a self-adjoint, linear, non-negative operator.

Note that the above framework includes, as special cases, the L^2 gradient flows for which $g(u) = F'(u)$ where $F(u)$ is a given nonlinear function, $L = -\Delta$ and $(\mathcal{G}u, u) = (-\Delta u + g(u), -\Delta u + g(u))$.

Solutions of (5.1) is often bound/positivity preserving. It is desirable, and sometimes necessary such as in the case of PNP and Keller-Segel equations, for the numerical solutions to be also bound/positivity preserving. While it is possible to construct some fully discrete numerical methods which preserve the bounds/positivity using finite-differences or piecewise linear finite-elements for a class of (5.1) satisfying a maximum principle, it is in general

very difficult to construct higher-order finite-elements or spectral methods which preserve bounds/positivity as well as energy dissipation.

While the SAV approach [1] provided a powerful approach to design numerical schemes which preserve energy dissipation, it does not have mechanism to preserve bounds or positivity. A common strategy to enforce solutions to preserve bounds or positivity is to use a suitable function transform. More precisely, given a prescribed range interval I which could be open, closed or half open, we can construct an invertible mapping $T : R \to I$, and make the function transform $u = T(v)$ in (5.1), leading to

$$\frac{\partial v}{\partial t} - \Delta v - \frac{T''(v)}{T'(v)}|\nabla v|^2 + \frac{1}{T'(v)} g(T(v)) = 0, \tag{5.4}$$

with either periodic or homogeneous Neumann boundary condition, since $\frac{\partial u}{\partial n} = T'(v)\frac{\partial v}{\partial n}$. After we solve v from the above, we get $u = T(v)$ whose range is included in I. Two typical cases are:

- $I = (a, b)$: a suitable choice is $T(v) = \frac{b-a}{2}\tanh(v) + \frac{b+a}{2}$ so that the range of $u = T(v)$ is still in I.
- $I = (0, \infty)$: a suitable choice is $T(v) = \exp(v/M)$, where M is a tunable parameter to prevent $T(v)$ increases too fast, so that $u = T(v)$ is always positive.

The main difficulty with this transformed approach is that the transformed equation (5.4) is much more complicated than (5.1), and it is difficult to design efficient and energy dissipative schemes. Fortunately, the recently proposed SAV approach [13] can provide a satisfactory solution as we show below.

As in the usual SAV approach, we introduce a SAV to enforce energy dissipation (5.2). More precisely, we set $r(t) = \int_\Omega F(u)dx + C_0$ with $C_0 > |E[u^0]|$ so that $E(u(\cdot,t)) + C_0 > 0$, and expand (5.4) with (5.2) as

$$\frac{\partial v}{\partial t} - \Delta v - \frac{T''(v)}{T'(v)}|\nabla v|^2 + \frac{1}{T'(v)}g(T(v)) = 0, \tag{5.5a}$$

$$u = T(v), \tag{5.5b}$$

$$\frac{dE_0(u)}{dt} + \frac{dr}{dt} = -\frac{E_0(u) + r(t)}{E(u)+C_0}\Big(\mathcal{G}u, u\Big), \tag{5.5c}$$

with $r(0) = \int_\Omega F(u(x,0))dx + C_0$, it is clear that the above system is equivalent to (5.4) with (5.2). However, discretizing the above will allow us to easily construct schemes which is energy dissipative, in addition to bound/positivity prerserving which is built into the system. We construct below k-th order BDF-Adams-Bashforth SAV schemes for (5.52b) in a uniform setting: treat the linear term Δv implicitly and use Adams-Bashforth extrapolation to deal with all nonlinear terms.

More precisely, given r^n and $(w^{\mathrm{j}}, v^{\mathrm{j}})$ for $\mathrm{j} = n, \cdots, n-k+1$, we find $(v^{n+1}, u^{n+1}, r^{n+1}, \xi^{n+1})$ such that

$$\frac{\alpha_k v^{n+1} - A_k(v^n)}{\delta t} - \Delta v^{n+1} = \frac{T''(B_k(v^n))}{T'(B_k(v^n))}|\nabla B_k(v^n)|^2 - \frac{1}{T'(B_k(v^n))}g(B_k(u^n)), \tag{5.6}$$

$$\bar{u}^{n+1} = T(v^{n+1}), \tag{5.7}$$

$$\begin{aligned}\frac{1}{\delta t}\Big(\frac{1}{2}\int_\Omega (\mathcal{L}\bar{u}^{n+1}\cdot\bar{u}^{n+1} - \mathcal{L}\bar{u}^n\cdot\bar{u}^n)dx + r^{n+1} - r^n\Big)& \\ = -\frac{\frac{1}{2}\int_\Omega \mathcal{L}\bar{u}^{n+1}\cdot\bar{u}^{n+1}dx + r^{n+1}}{E[\bar{u}^{n+1}]+C_0}\Big(\mathcal{G}\bar{u}^{n+1}, \bar{u}^{n+1}\Big),&\end{aligned} \tag{5.8}$$

$$\xi^{n+1} = \frac{\int_\Omega \frac{1}{2}\mathcal{L}\bar{u}^{n+1}\cdot\bar{u}^{n+1}dx + r^{n+1}}{E[\bar{u}^{n+1}]+C_0}, \tag{5.9}$$

$$u^{n+1} = \eta_k^{n+1}\bar{u}^{n+1} \text{ with } \eta_k^{n+1} = 1-(1-\xi^{n+1})^{I_k},\ I_k = \begin{cases} k+1,\ k \text{ odd} \\ k,\ k \text{ even}\end{cases}, \tag{5.10}$$

where the constant α_k, operators A_k, B_k are defined by

BDF1:

$$\alpha_1 = 1,\ A_1(v^n) = v^n,\ B_1(h^n) = h^n; \tag{5.11}$$

BDF2:

$$\alpha_2 = \frac{3}{2},\quad A_2(v^n) = 2v^n - \frac{1}{2}v^{n-1},\quad B_2(h^n) = 2h^n - h^{n-1}; \tag{5.12}$$

BDF3:

$$\alpha_3 = \frac{11}{6},\quad A_3(v^n) = 3v^n - \frac{3}{2}v^{n-1} + \frac{1}{3}v^{n-2},\quad B_3(h^n) = 3h^n - 3h^{n-1} + h^{n-2}; \tag{5.13}$$

BDF4:

$$\alpha_4 = \frac{25}{12},\quad A_4(v^n) = 4v^n - 3v^{n-1} + \frac{4}{3}v^{n-2} - \frac{1}{4}v^{n-3},\quad B_4(h^n) = 4h^n - 6h^{n-1} + 4h^{n-2} - h^{n-3}. \tag{5.14}$$

The formulae for $k = 5$ and $k = 6$ can be derived similarly.

Several remarks are in oder:

- Since we assume T is invertible, $T'(v) \neq 0$ so the above scheme is well defined. The range of the approximate solution $\bar{u}^{n+1} = T(v^{n+1})$ is obviously included in I.
- (5.6) is a k-th order approximation to (5.5a) with k-th order BDF for the linear terms and k-th order Adams-Bashforth extrapolation for the nonlinear terms. Hence, v^{n+1} is a k-th order approximation to $v(t_{n+1})$.
- (5.8) is a first-order approximation to (5.5c). Hence, r^{n+1} is a first order approximation to $E_1(u(\cdot, t_{n+1}))$ which implies that ξ^{n+1} is a first order approximation to 1. Hence, $\eta_k^{n+1} = 1 + O(\delta t)^{I_k}$ which implies that both $\bar{u}^{n+1}$ and u^{n+1} are k-th order approximation of $u(t_{n+1})$.
- The above scheme can be efficiently implemented as follows:
  - determine v^{n+1} from (5.6);
  - set $\bar{u}^{n+1} = T(v^{n+1})$;

– with $\bar{u}^{n+1}$ known, determine r^{n+1} explicitly from (5.8), and compute ξ^{n+1} from (5.9);

– update u^{n+1} using (5.10), goto the next step.

The main cost is to solve v^{n+1} from (5.6) which is a linear equation with constant coefficients.

Theorem 5.2.1. *Without loss of generality, we assume $ab \leq 0$ if $I = (a, b)$. Given u^{i} with range in I, $v^{\mathrm{i}} = T^{-1}(u^{\mathrm{i}})$ and r^{i} for $\mathrm{i} = 0, 1, \ldots, k-1$. The scheme (5.6)-(5.10) admits a unique solution satisfying the following properties unconditionally:*

1. *Positivity or bound preserving: i.e., the range of $\bar{u}^{n+1}$ and u^{n+1} is in I.*

2. *Unconditionally energy dissipation with a modified energy defined by $\bar{E}^n = \int_\Omega \frac{1}{2}\mathcal{L}\bar{u}^n \cdot \bar{u}^n dx + r^n$: More precisely, if $\bar{E}^n \geq 0$, we have $\bar{E}^{n+1} \geq 0$ and*

$$\bar{E}^{n+1} - \bar{E}^n \leq -\delta t \frac{\bar{E}^{n+1}}{E[\bar{u}^{n+1}]+C_0}\left(\mathcal{G}\bar{u}^{n+1}, \bar{u}^{n+1}\right) \leq 0. \tag{5.15}$$

3. *Furthermore, if $E_1(u) = \int_\Omega F(u)dx$ is bounded from below, then for the k-th order schemes, there exists constant M_k, such that*

$$(\mathcal{L}u^n, u^n)^{1/2} \leq M_k, \forall n. \tag{5.16}$$

Proof. By construction, the scheme is obviously positivity or bound preserving for $\bar{u}^{n+1}$.

We derive from (5.8) that

$$\bar{E}^{n+1} = \bar{E}^n / \Big(1 + \frac{\delta t}{E[\bar{u}^{n+1}]+C_0}(\mathcal{G}\bar{u}^{n+1}, \bar{u}^{n+1})\Big).$$

Hence, if $\bar{E}^n \geq 0$, we have $\bar{E}^{n+1} \geq 0$, and (5.15) follows directly from (5.8). It follows from (5.9), (5.15) and $C_0 = E[u^0]$, $E[\bar{u}^{n+1}] > 0$ that

$$0 < \xi^{n+1} \leq \frac{E[u^0] + C_0}{E[\bar{u}^{n+1}] + C_0} \leq 2, \tag{5.17}$$

which together with (5.10) imply

$$0 < (1 - \xi^{n+1})^{I_k} < 1, \ 0 < \eta_k^{n+1} < 1. \tag{5.18}$$

Hence, the range of u^{n+1} is also in I as $u^{n+1} = \eta_k^{n+1}\bar{u}^{n+1}$ for $I = (0, \infty)$ or $I = (a, b)$ with $ab \leq 0$.

If $E_1(u) = \int_\Omega F(u)dx$ is bounded from below, without loss of generality, we assume $E_1(u) > 1$. Denote $M := \bar{E}[u(\cdot, 0)]$, then (5.15) implies $\bar{E}^n \leq M$, $\forall n$. Now, it follows from (5.9) and the assumption of $E_1(u) > 1$ that

$$|\xi^{n+1}| = \frac{\bar{E}^{n+1}}{E[\bar{u}^{n+1}] + C_0} \leq \frac{2M}{(\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) + 2}. \tag{5.19}$$

Since $\eta_k^{n+1} = 1 - (1 - \xi^{n+1})^{I_k}$, there exists a polynomial P_k of degree $I_k - 1$ and a constant $M_k > 0$ such that

$$|\eta_k^{n+1}| = |\xi^{n+1} P_k(\xi^{n+1})| \leq \frac{M_k}{(\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1}) + 2}. \tag{5.20}$$

Therefore, by the fact $\sqrt{A} \leq A + 2$ for all $A \geq 0$, we have

$$(\mathcal{L}u^{n+1}, u^{n+1})^{1/2} = \eta_k^{n+1}(\mathcal{L}\bar{u}^{n+1}, \bar{u}^{n+1})^{1/2} \leq M_k. \tag{5.21}$$

□

The above scheme can be directly applied to bound/positivity preserving L^2 gradient flows, including in particular the Allen-Cahn equation. In the following two sections, we shall extend the approach presented in this section to construct positivity preserving and energy stable schemes for Poisson-Nernst-Planck and Keller-Segel equations for which it is essential to preserve positivity.

We emphasize that both $\bar{u}^{n+1}$ and u^{n+1} are k-th order approximation to $u(\cdot, t_{n+1})$.

We only considered the time discretization in this section. However, it is clear from the proof of the above theorem that, as long as the spatial approximations of $\mathcal{G}$ and $\mathcal{L}$ are still positive definite, the results of Theorem 5.2.1 also holds for the fully discrete schemes.

5.3 Positivity preserving schemes for the Poisson-Nernst-Planck equation

We consider in this section the Poisson-Nernst-Planck (PNP) equation which describes the dynamics of N species of charged particles driven by Brownian motion and electric field (cf. [100]–[102] and the references therein). To simplify the presentation, we will focus on the two-component system ($N = 2$). The schemes can be easily extended to more general PNP system with N components.

5.3.1 Poisson-Nernst-Planck equation

We consider a two-component PNP system in the following form:

$$\frac{\partial c_1}{\partial t} = D_1 \nabla \cdot (\nabla c_1 + \chi_1 z_1 c_1 \nabla \phi), \tag{5.22a}$$

$$\frac{\partial c_2}{\partial t} = D_2 \nabla \cdot (\nabla c_2 + \chi_1 z_2 c_2 \nabla \phi), \tag{5.22b}$$

$$-\Delta \phi = \chi_2 (z_1 c_1 + z_2 c_2), \tag{5.22c}$$

in an open bounded domain $\Omega \subset R^d$ ($d = 1, 2, 3$) and supplemented with either periodic boundary condition, or no flux boundary conditions

$$\frac{\partial c_{\mathrm{i}}}{\partial \vec{n}}|_{\partial\Omega} = 0, \ \mathrm{i} = 1, 2; \frac{\partial \phi}{\partial \vec{n}}|_{\partial\Omega} = 0. \tag{5.23}$$

It is also possible to use the Dirichlet boundary condition $\phi|_{\partial\Omega} = 0$ or a Robin type boundary condition $(\alpha\phi + \beta\frac{\partial\phi}{\partial\vec{n}})|_{\partial\Omega} = 0$.

In the above, the unknown are c_{i}, the density of the i-th species, and ϕ, the internal electric potential, $D_{\mathrm{i}} > 0$ is the diffusion constant of the i-th specie ($\mathrm{i} = 1, 2$), z_{i} are the valence constant and χ_1, χ_2 are dimensionless parameters. To make the formulas below more concise, in the following we fix $z_1 = 1$, $z_2 = -1$ and $\chi_1 = \chi_2 = 1$.

Using the identity $\nabla\psi = \psi\nabla\log\psi$, we can rewrite (5.22) as a Wasserstein gradient flow

$$\frac{\partial c_1}{\partial t} = D_1\nabla\cdot(c_1\nabla\log c_1 + c_1\nabla\phi), \tag{5.24a}$$

$$\frac{\partial c_2}{\partial t} = D_2\nabla\cdot(c_2\nabla\log c_2 - c_2\nabla\phi), \tag{5.24b}$$

$$-\Delta\phi = c_1 - c_2, \tag{5.24c}$$

with the free energy

$$E(c_1, c_2, \phi) = \int_\Omega c_1(\log c_1 - 1) + c_2(\log c_2 - 1) + \frac{1}{2}|\nabla\phi|^2 dx. \tag{5.25}$$

Indeed, taking the inner product of (5.24a) with $\log c_1 + \phi$ and of (5.24b) with $\log c_2 - \phi$, summing them up along with $(-\Delta\partial_t\phi = \partial_t(c_1 - c_2), \phi)$, we obtain the following energy law:

$$\frac{dE(c_1, c_2, \phi)}{dt} = -\int_\Omega \Big(D_1\, c_1|\nabla(\log c_1 + \phi)|^2 + D_2\, c_2|\nabla(\log c_2 - \phi)|^2\Big)dx. \tag{5.26}$$

Note that the form of the free energy, as well as the well-posedness of (5.24), requires $c_1, c_2 > 0$. Therefore, it is of critical importance that numerical schemes for the PNP system preserve positivity.

On the other hand, we also derive from (5.24) and (5.23) that

$$\frac{d}{dt}\int_\Omega c_{\mathrm{i}} dx = 0, \quad \mathrm{i} = 1, 2, \tag{5.27}$$

i.e., the mass for each component is conserved.

5.3.2 Positivity preserving SAV scheme

As explained in Section 2, we can preserve the positivity using suitable function transforms. Since only c_1, c_2 are positivity preserving, we only make function transform for c_1, c_2. More precisely, we introduce two new functions p_1 and p_2 through

$$c_{\mathrm{i}} = T(p_{\mathrm{i}}) := \exp(p_{\mathrm{i}}), \ \mathrm{i} = 1, 2, \tag{5.28}$$

which implies in particular $c_{\rm i} > 0$, ${\rm i} = 1, 2$.

Substituting (5.28) into (5.24a)-(5.24b), we obtain

$$\frac{\partial p_1}{\partial t} = D_1(\Delta p_1 + |\nabla p_1|^2 + \nabla p_1 \cdot \nabla \phi + \Delta \phi), \tag{5.29a}$$

$$\frac{\partial p_2}{\partial t} = D_2(\Delta p_2 + |\nabla p_2|^2 - \nabla p_1 \cdot \nabla \phi - \Delta \phi). \tag{5.29b}$$

Note that for this transform, we have $T'(p_{\rm i}) = T''(p_{\rm i}) = T(p_{\rm i})$, so the transformed equations are not too complicated.

Next we split the free energy $E(c_1, c_2, \phi)$ into the sum of $E_0(\phi) := \frac{1}{2}(\nabla\phi, \nabla\phi)$ and $E_1(c_1, c_2) := \int_\Omega c_1(\log c_1 - 1) + c_2(\log c_2 - 1)dx$. It is clear that $E_1(c_1, c_2)$ is convex and bounded from below in the admissible set $\mathcal{D} := \{(c_1, c_2) : c_1, c_2 > 0\}$, so we assume that for some $C_0 > 0$,

$$E_1(c_1, c_2) \geq -C_0 + 1, \tag{5.30}$$

and define a SAV $r(t) = E_1(c_1, c_2) + C_0 > 1$. Then, the total free energy E and its time derivative can be rewritten as

$$E(c_1, c_2, \phi) = \frac{1}{2}(\nabla\phi, \nabla\phi) + r(t) = E_0(\phi) + r(t), \tag{5.31a}$$

$$\frac{dE}{dt} = \frac{dE_0}{dt} + r_t. \tag{5.31b}$$

Denote $\mu_1 = \log c_1 + \phi$, $\mu_2 = \log c_2 - \phi$, we can reformulate (5.24) and (5.26) as

$$\frac{\partial p_1}{\partial t} = D_1(\Delta p_1 + |\nabla p_1|^2 + \nabla p_1 \cdot \nabla \phi + \Delta \phi), \tag{5.32a}$$

$$\frac{\partial p_2}{\partial t} = D_2(\Delta p_2 + |\nabla p_2|^2 - \nabla p_2 \cdot \nabla \phi - \Delta \phi), \tag{5.32b}$$

$$c_1 = \exp(p_1),\ \ c_2 = \exp(p_2), \tag{5.32c}$$

$$-\Delta\phi = c_1 - c_2, \tag{5.32d}$$

$$\frac{dE_0}{dt} + r_t = -\frac{E_0(\phi) + r(t)}{E(c_1, c_2, \phi) + C_0} \int_\Omega \Big(D_1\, c_1 |\nabla\mu_1|^2 + D_2\, c_2 |\nabla\mu_2|^2\Big) dx. \tag{5.32e}$$

We remark that since the above system is equivalent to the original system (5.24), the masses of $c_{\rm i}$ are still conserved, but that of $p_{\rm i}$ are not.

We now construct k-th order SAV schemes ($1 \le k \le 6$) for the above system in a uniform setting.

Given $(c_{\rm i}^{\rm j}, p_{\rm i}^{\rm j}, \phi^{\rm j}, r^{\rm j}, \xi^{\rm j})$, ${\rm i} = 1, 2,\ {\rm j} = n, n-1, \cdots, n-k+1$ such that

$$\int_\Omega c_{\rm i}^{\rm j} dx = \int_\Omega c_{\rm i}^0 dx,\ {\rm i} = 1, 2,\ {\rm j} = n, n-1, \cdots, n-k+1, \tag{5.33}$$

we determine $(c_{\rm i}^{n+1}, p_{\rm i}^{n+1}, \lambda_{\rm i}^{n+1})$, ${\rm i} = 1, 2$ and $(\phi^{n+1}, r^{n+1}, \xi^{n+1})$ as follows:

$$\frac{\alpha_k p_{\rm i}^{n+1} - A_k(p_{\rm i}^n)}{\delta t} - D_{\rm i}\Delta p_{\rm i}^{n+1} = g_{\rm i}(B_k(p_{\rm i}^n), B_k(\phi^n)), \quad {\rm i} = 1, 2, \tag{5.34}$$

$$\bar{c}_{\rm i}^{n+1} = \exp(p_{\rm i}^{n+1}),\ {\rm i} = 1, 2, \tag{5.35}$$

$$\lambda_{\rm i}^{n+1} \int_\Omega \alpha_k \bar{c}_{\rm i}^{n+1} dx - \int_\Omega A_k(c_{\rm i}^n) dx = 0,\ {\rm i} = 1, 2, \tag{5.36}$$

$$c_{\rm i}^{n+1} = \lambda_{\rm i}^{n+1} \bar{c}_{\rm i}^{n+1},\ {\rm i} = 1, 2, \tag{5.37}$$

$$-\Delta\bar{\phi}^{n+1} = c_1^{n+1} - c_2^{n+1}, \tag{5.38}$$

$$\begin{aligned} &\frac{1}{\delta t}\Big(E_0(\bar{\phi}^{n+1}) - E_0(\bar{\phi}^n) + r^{n+1} - r^n\Big) \\ &= -\frac{E_0(\bar{\phi}^{n+1}) + r^{n+1}}{E(c_1^{n+1}, c_2^{n+1}, \bar{\phi}^{n+1}) + C_0} \int_\Omega \Big(D_1\, c_1^{n+1} |\nabla\mu_1^{n+1}|^2 + D_2\, c_2^{n+1} |\nabla\mu_2^{n+1}|^2\Big) dx, \end{aligned} \tag{5.39}$$

$$\xi^{n+1} = \frac{E_0(\bar{\phi}^{n+1}) + r^{n+1}}{E(c_1^{n+1}, c_2^{n+1}, \bar{\phi}^{n+1}) + C_0}, \tag{5.40}$$

$$\phi^{n+1} = \eta_k^{n+1}\bar{\phi}^{n+1} \text{ with } \eta_k^{n+1} = 1 - (1 - \xi^{n+1})^k, \tag{5.41}$$

together with homogeneous Neumann boundary conditions

$$\frac{\partial p_{\rm i}^{n+1}}{\partial \vec{n}}|_{\partial\Omega} = 0,\ {\rm i} = 1, 2; \frac{\partial \phi^{n+1}}{\partial \vec{n}}|_{\partial\Omega} = 0, \tag{5.42}$$

where $\mu_1^{n+1} = \log c_1^{n+1} + \bar{\phi}^{n+1}$, $\mu_2^{n+1} = \log c_2^{n+1} - \bar{\phi}^{n+1}$, α_k, A_k and B_k are the same as in the last section, and

$$g_1(p_1, \phi) = D_1(|\nabla p_1|^2 + \nabla p_1 \cdot \nabla\phi + \Delta\phi),$$

$$g_2(p_2, \phi) = D_1(|\nabla p_2|^2 - \nabla p_2 \cdot \nabla\phi - \Delta\phi).$$

Similar to the last section, we have the following remarks:

- Clearly, (5.34) is a k-th order semi-implicit scheme for (5.32a)-(5.32b). We then derive from (5.35)-(5.38) that $\lambda_{\mathrm{i}}^{n+1}$ is k-th order approximations to 1, c_{i}^{n+1} and $\bar{\phi}^{n+1}$ are k-th order approximations to $c_{\mathrm{i}}(t_{n+1})$ and $\phi(t_{n+1})$.

- (5.39) is a first-order approximation to (5.32e), so r^{n+1} is a first-order approximation to $E_1(c_1^{n+1}, c_2^{n+1})$ and $\xi^{n+1} = 1 + O(\delta t)$ which implies that $\eta_k^{n+1} = 1 + O(\delta t^k)$. Therefore, ϕ^{n+1} is also a k-th order approximation of $\phi(t_{n+1})$.

- The scheme (5.34)- (5.41) can be efficiently implemented by the following steps:

  1. solve p_{i}^{n+1} from (5.34);
  2. compute $\bar{c}_1^{n+1}, \bar{c}_2^{n+1}$ from (5.35) and compute $\lambda_{\mathrm{i}}^{n+1}$ explicitly from (5.36);
  3. update c_1^{n+1}, c_2^{n+1} from (5.37) and solve $\bar{\phi}$ from (5.38);
  4. compute r^{n+1} explicitly from (5.39) and then obtain ξ^{n+1} from (5.40);
  5. update ϕ^{n+1} from (5.41), goto next step.

  The main computational cost is to solve the linear equations with constant coefficients in (5.34) and (5.38).

We have the following results:

Theorem 5.3.1. *Given $c_{\mathrm{i}}^{\mathrm{j}} > 0$, $p_{\mathrm{i}}^{\mathrm{j}} = \log c_{\mathrm{i}}^{\mathrm{j}}$, ϕ^{j}, and r^{j} such that $\int_\Omega c_{\mathrm{i}}^{\mathrm{j}} dx = \int_\Omega c_{\mathrm{i}}^0 dx$ for $\mathrm{i} = 1, 2$ and $\mathrm{j} = n, n-1, \ldots, n-k+1$. The scheme (5.34)-(5.41) admits a unique solution satisfying the following properties unconditionally:*

1. *Positivity preserving: c_1^{n+1}, $c_2^{n+1} > 0$.*
2. *Mass conserving: $\int_\Omega c_{\mathrm{i}}^{n+1} dx = \int_\Omega c_{\mathrm{i}}^0 dx$ for $\mathrm{i} = 1, 2$.*
3. *Unconditionally energy dissipation with a modified energy defined by $\bar{E}^n = E_0(\bar{\phi}^n) + r^n$: More precisely, if $\bar{E}^n \geq 0$, we have $\bar{E}^{n+1} \geq 0$, $\xi^{n+1} \geq 0$ and*

$$\bar{E}^{n+1} - \bar{E}^n = -\xi^{n+1} \int_\Omega \Big(D_1\, c_1^{n+1} |\nabla \mu_1^{n+1}|^2 + D_2\, c_2^{n+1} |\nabla \mu_2^{n+1}|^2 \Big) dx \leq 0. \tag{5.43}$$

4. There exists constant M_k, such that

$$\sqrt{E_0[\phi^n]} \leq M_k, \ \forall n. \tag{5.44}$$

Proof. From (5.35), we obviously have $\bar{c}_1^{n+1}$, $\bar{c}_2^{n+1} > 0$.

We derive from the assumption that $\int_\Omega c_{\mathrm{i}}^{\mathrm{j}} dx = \int_\Omega c_{\mathrm{i}}^0 dx$ for $\mathrm{i} = 1, 2$ and $\mathrm{j} = n, n-1, \ldots, n-k+1$, and the definition of coefficients α_k and A_k in section 2 that

$$\int_\Omega A_k(c_{\mathrm{i}}^n) dx = \alpha_k \int_\Omega c_{\mathrm{i}}^0 dx.$$

It then follows from (5.33) and (5.36) that

$$\alpha_k \lambda_{\mathrm{i}}^{n+1} \int_\Omega \bar{c}_{\mathrm{i}}^{n+1} dx = \alpha_k \int_\Omega c_{\mathrm{i}}^0 dx, \tag{5.45}$$

which, along with $\bar{c}_{\mathrm{i}}^{n+1} > 0$, implies that $\lambda_{\mathrm{i}}^{n+1} > 0$. Hence, we have c_1^{n+1}, $c_2^{n+1} > 0$, and we derive from the above and (5.37) that $\int_\Omega c_{\mathrm{i}}^{n+1} dx = \int_\Omega c_{\mathrm{i}}^0 dx$ for $\mathrm{i} = 1, 2$.

It follows from (5.39) that

$$E_0(\bar{\phi}^{n+1}) + r^{n+1} = \frac{E_0(\bar{\phi}^n) + r^n}{1 + \delta t \frac{\int_\Omega \left(D_1 c_1^{n+1} |\nabla \mu_1^{n+1}|^2 + D_2 c_2^{n+1} |\nabla \mu_2^{n+1}|^2 \right) dx}{E(c_1^{n+1}, c_2^{n+1}, \bar{\phi}^{n+1}) + C_0}} \geq 0. \tag{5.46}$$

Therefore, we derive from (5.40) that $\xi^{n+1} \geq 0$, which, together with (5.39), implies (5.43).

Denote $M := \bar{E}^0$, then (5.43) implies $\bar{E}^n \leq M$, $\forall n$. It follows from (5.39) and (5.30) that

$$|\xi^{n+1}| = \frac{\bar{E}^{n+1}}{E(c_1{}^{n+1}, c_2{}^{n+1}, \bar{\phi}^{n+1}) + C_0} \leq \frac{M}{E_0(\bar{\phi}^{n+1}) + 1}. \tag{5.47}$$

Since $\eta_k^{n+1} = 1 - (1 - \xi^{n+1})^k$, there exists a polynomial P_{k-1} of degree $k-1$ and a constant $M_k > 0$ such that

$$|\eta_k^{n+1}| = |\xi^{n+1} P_{k-1}(\xi^{n+1})| \leq \frac{M_k}{E_0(\bar{\phi}^{n+1}) + 1}. \tag{5.48}$$

Therefore, by the fact that $\sqrt{A} \le A+1$ for all $A \ge 0$, we derive

$$\sqrt{E_0[\phi^{n+1}]} = |\eta_k^{n+1}| \sqrt{E_0[\bar{\phi}^{n+1}]} \le M_k.$$

□

We emphasize that both $\bar{c}_{\rm i}^{n+1}$ (resp. $\bar{\phi}_{\rm i}^{n+1}$) and $c_{\rm i}^{n+1}$ (resp. ϕ^{n+1}) are k-th order approximation to $c_{\rm i}(\cdot, t_{n+1})$ (resp. $\phi(\cdot, t_{n+1})$), ${\rm i} = 1, 2$.

Obviously, the positivity of $c_{\rm i}$ will be preserved with any spatial approximation of the schemes (5.34)-(5.41).

It is clear from the proof of the above theorem that the mass conservation and the energy dissipation (5.43) still hold for any fully discrete schemes.

5.4 Bound preserving schemes for the Keller-Segel equation

We first introduce the Keller-Segel equations, followed by the construction of bound preserving schemes for one particular case of the Keller-Segel equations whose solution is bound preserving.

5.4.1 Keller-Segel equations

To fix the idea, we consider the following Keller-Segel system with only one organism and one chemoattractant in a bounded domain Ω:

$$\frac{\partial u}{\partial t} = D\Big(\gamma \Delta u - \chi \nabla \cdot (\eta(u) \nabla \phi)\Big), \tag{5.49a}$$

$$\tau \frac{\partial \phi}{\partial t} = \mu \Delta \phi - \alpha \phi + \chi u, \tag{5.49b}$$

with either periodic boundary conditions, or no-flux boundary conditions on u and the Neumann boundary conditions on ϕ,

$$\gamma \frac{\partial u}{\partial \vec{n}} - \chi \eta(u) \frac{\partial \phi}{\partial \vec{n}} = 0, \ \frac{\partial \phi}{\partial \vec{n}} = 0 \ on \ \partial\Omega. \tag{5.50}$$

Here, the unknown are u, the concentration of the organism, and ϕ, the concentration of the chemoattractant. The parameters $D, \gamma, \chi, \tau, \mu, \alpha$ are all positive. The function $\eta(u) \geq 0$ describes the concentration-dependent mobility. It is a smooth function with $\eta(0) = 0$.

The model is a parabolic-parabolic system when $\tau > 0$, and a parabolic-elliptic system when $\tau = 0$.

The system (5.49) with (5.50) can be interpreted as a gradient flow about (u, ϕ). To this end, we choose $f(u)$ such that $f(u) = 1/\eta(u)$, and define the free energy

$$E[u, \phi] = \int_\Omega (\gamma f(u) - \chi u\phi + \frac{\mu}{2}|\nabla\phi|^2 + \frac{\alpha}{2}\phi^2)dx. \tag{5.51}$$

Then writing $\Delta u = \nabla \cdot \big(\frac{1}{f(u)}\nabla f(u)\big)$, we can rewrite (5.49) as

$$\frac{\partial u}{\partial t} = D\nabla \cdot \Big(\frac{1}{f(u)}\nabla(\gamma f(u) - \chi\phi)\Big) = D\nabla \cdot \Big(\frac{1}{f(u)}\nabla\frac{\delta E}{\delta u}\Big), \tag{5.52a}$$

$$\tau\frac{\partial \phi}{\partial t} = \mu\Delta\phi - \alpha\phi + \chi u = -\frac{\delta E}{\delta \phi}. \tag{5.52b}$$

Taking the inner products of (5.52a) with $\frac{\delta E}{\delta u}$, and of (5.52b) with $\frac{\partial \phi}{\partial t}$, and summing up the results, we obtain the energy dissipation law:

$$\frac{dE[u(t), \phi(t)]}{dt} = -\int_\Omega [D\frac{1}{f(u)}\Big(\nabla\frac{\delta E}{\delta u}\Big)^2 + \tau\Big(\frac{\partial\phi}{\partial t}\Big)^2]dx. \tag{5.53}$$

We now consider several typical choices of $\eta(u)$ and the corresponding function $f(u)$.

1. The classical Keller-Segel system: $\eta(u) = u$. We can choose $f(u) = u\log u - u$ with the domain of definition $(0, +\infty)$. In this case, it is known that its solution can blow up in finite time if the initial mass is large enough [103]–[105].

2. Keller-Segel system with a bounded mobility: a typical choice [106], [107] is $\eta(u) = \frac{u}{1+\kappa u}$ ($\kappa > 0$). In this case, we can choose $f(u) = u\log u - u + \kappa u^2/2$ with the domain of definition $(0, +\infty)$.

3. Keller-Segel system with a saturation concentration: $\eta(u) = u(1 - u/M)$, where$M > 0$ is the saturation concentration, and the mobility tends to zero when it is near saturation

[108], [109]. In this case, we can choose $f(u) = u\log u + (M-u)\log(1-u/M)$ with the domain of definition $(0, M)$.

Hence, the solution of the Keller-Segel system is positivity preserving in cases (i) and (ii), and bound preserving in case (iii). Furthermore, we observe from (5.49) that

$$\frac{d}{dt}\int_\Omega u dx = 0. \tag{5.54}$$

To simplify the presentation, we shall only consider the third case where the solution is bound preserving. For the first and second order cases, the solution is positivity preserving, so one can construct positivity preserving schemes for these two cases similarly by replacing the mapping below with $Y(v) = \exp(v)$ as in the last section.

5.4.2 Bound preserving SAV schemes

We set $\eta(u) = u(1-u/M)$ and $f(u) = u\log u + (M-u)\log(1-u/M)$, and split $E[u,\phi]$ into two parts as follows

$$E[u,\phi] = \int_\Omega(\gamma f(u) - \chi u\phi + \frac{\alpha}{4}\phi^2)dx + \int_\Omega(\frac{\mu}{2}|\nabla\phi|^2 + \frac{\alpha}{4}\phi^2)dx = E_1[u,\phi] + E_0[\phi]. \tag{5.55}$$

Note that $f(u) = u\log u + (M-u)\log(1-u/M)$ implies that $u \in (0, M)$. Along with $\alpha > 0$ and f is strictly convex, it is easy to see that E_1 is bounded from below. Hence, there exists $C_0 > 0$ such that,

$$E_1[u,\phi] \geq -C_0 + 1. \tag{5.56}$$

Due to the form of $f(u)$, it is necessary that the range of numerical solution is also in $(0, M)$. To this end, we consider the transform

$$u = T(v) := \frac{M}{2}\tanh(v) + \frac{M}{2}. \tag{5.57}$$

As $\tanh(x) \in (-1, 1)$, $\forall x \in (-\infty, +\infty)$, then for $v \in (-\infty, +\infty)$, we have $u \in (0, M)$. Since ϕ is not bound preserving, we do not need to transform ϕ.

Substituting (5.57) into (5.49a), we obtain the equation for v

$$\frac{\partial v}{\partial t} = D\gamma\Delta v + D\gamma\frac{\tanh''(v)}{\tanh'(v)}|\nabla v|^2 - \frac{2D\chi}{M\tanh(v)}\nabla\cdot\Big(\eta(u)\nabla\phi\Big). \tag{5.58}$$

Note that $\tanh'(x) = 1-\tanh^2(x)$, we know $\tanh'(v)\neq 0$ and (5.58) is well-defined.

We introduce $r(t) = E_1(u,\phi) + C_0 \geq 1$. Then, we have

$$E[u,\phi] = \frac{\mu}{2}\Big(\phi, -\Delta\phi\Big)_\Omega + \frac{\alpha}{4}\Big(\phi,\phi\Big)_\Omega + r = E_0(\phi) + r, \tag{5.59a}$$

$$\frac{d}{dt}E[u,\phi] = \mu\Big(\phi_t, -\Delta\phi\Big)_\Omega + \frac{\alpha}{2}\Big(\phi_t,\phi\Big)_\Omega + r_t = \frac{dE_0(\phi)}{dt} + r_t. \tag{5.59b}$$

We can reformulate (5.49) and (5.53) as

$$\frac{\partial v}{\partial t} = D\gamma\Delta v + \Big(D\gamma\frac{\tanh''(v)}{\tanh'(v)}|\nabla v|^2 - \frac{2D\chi}{M\tanh'(v)}\nabla\cdot(\eta(u)\nabla\phi)\Big), \tag{5.60a}$$

$$u = \frac{M}{2}\tanh(v) + \frac{M}{2}, \tag{5.60b}$$

$$\tau\frac{\partial\phi}{\partial t} = \mu\Delta\phi - \alpha\phi + \chi u, \tag{5.60c}$$

$$\frac{dE_0(\phi)}{dt} + r_t = -\frac{E_0(\phi)+r(t)}{E(u,\phi)+C_0}\int_\Omega [D\frac{1}{f''(u)}\Big(\nabla\frac{\delta E}{\delta u}\Big)^2 + \tau\Big(\frac{\partial\phi}{\partial t}\Big)^2]dx. \tag{5.60d}$$

We now construct k-th order schemes for (5.60) in a uniform setting.

Given $(v^{\mathrm{i}}, u^{\mathrm{i}}, \phi^{\mathrm{i}}, r^{\mathrm{i}})$, $\mathrm{i} = n, n-1, \cdots, n-k+1$, we find $(v^{n+1}, u^{n+1}, \phi^{n+1}, r^{n+1})$ as follows:

$$\frac{\alpha_k v^{n+1} - A_k(v^n)}{\delta t} - D\gamma\Delta v^{n+1} = g(B_k(v^n), B_k(u^n), B_k(\phi^n)), \tag{5.61}$$

$$\bar{u}^{n+1} = \frac{M}{2}\tanh(v^{n+1}) + \frac{M}{2}, \tag{5.62}$$

$$\lambda^{n+1}\int_\Omega \alpha_k \bar{u}^{n+1}dx - \int_\Omega A_k(u^n)dx = 0, \tag{5.63}$$

$$u^{n+1} = \lambda^{n+1}\bar{u}^{n+1}, \tag{5.64}$$

$$\tau\frac{\alpha_k\bar{\phi}^{n+1} - A_k(\bar{\phi}^n)}{\delta t} = \mu\Delta\bar{\phi}^{n+1} - \alpha\bar{\phi}^{n+1} + \chi u^{n+1}, \tag{5.65}$$

$$\begin{aligned}&\frac{1}{\delta t}\Big(E_0(\bar{\phi}^{n+1}) - E_0(\bar{\phi}^n) + r^{n+1} - r^n\Big)\\ &= -\frac{E_0(\bar{\phi}^{n+1}) + r^{n+1}}{E[\bar{u}^{n+1}, \bar{\phi}^{n+1}] + C_0}\int_\Omega [\frac{D}{f''(\bar{u}^{n+1})}\Big(\nabla\frac{\delta E}{\delta u}(\bar{u}^{n+1})\Big)^2 + \tau\Big(\frac{\bar{\phi}^{n+1} - \bar{\phi}^n}{\delta t}\Big)^2]dx,\end{aligned} \tag{5.66}$$

$$\xi^{n+1} = \frac{E_0(\bar{\phi}^{n+1}) + r^{n+1}}{E[\bar{u}^{n+1}, \bar{\phi}^{n+1}] + C_0}, \tag{5.67}$$

$$\phi^{n+1} = \eta_k^{n+1}\bar{\phi}^{n+1} \text{ with } \eta_k^{n+1} = 1 - (1-\xi^{n+1})^k, \tag{5.68}$$

where the constant α_k, operators A_k, B_k are defined in Section 2, and

$$g(u, v, \phi) = D\gamma\frac{\tanh''(v)}{\tanh'(v)}|\nabla v|^2 - \frac{2D\chi}{M\tanh'(v)}\nabla\cdot(\eta(u)\nabla\phi). \tag{5.69}$$

Essential properties of the above schemes are as follows:

- (5.61) and (5.65) are k-th order semi-implicit schemes for (5.60a) and (5.60c), (5.63) is k-th order approximation to (5.54), which imply that $v^{n+1}, \lambda^{n+1}, u^{n+1}, \bar{\phi}^{n+1}$ are k-th order approximations to $v(t_{n+1}), 1, u(t_{n+1}), \phi(t_{n+1})$.

- (5.66) is a first-order approximation to (5.60d), which implies that r^{n+1} is a first-order approximation to $r(t_{n+1})$. Then, (5.67) implies that $\xi^{n+1} = 1 + O(\delta t)$, which in turn implies $\eta_k^{n+1} = 1 + O(\delta t)^k$ and ϕ^{n+1} is a k-th order approximations to $\phi(t_{n+1})$.

- The above scheme can be efficiently implemented as follows:

  1. solve v^{n+1} from (5.61);

  2. compute $\bar{u}^{n+1}$ from (5.62) and compute λ^{n+1} explicitly from (5.63);

3. update u^{n+1} from (5.64);

4. with u^{n+1} known, solve $\bar{\phi}^{n+1}$ from (5.65);

5. with $\bar{u}^{n+1}$, $\bar{\phi}^{n+1}$ known, determine r^{n+1} explicitly from (5.66);

6. compute ξ^{n+1} from (5.67) and update ϕ^{n+1} from (5.68), goto the next step.

We have the following results:

Theorem 5.4.1. *Given u^{i}, ϕ^{i}, v^{i} and r^{i} such that*

$$\int_\Omega u^{\mathrm{i}} dx = \int_\Omega u^0 dx, \quad \mathrm{i} = n, n-1, \ldots, n-k+1. \tag{5.70}$$

Then, the scheme (5.61)-(5.68) admits a unique solution satisfying the following properties unconditionally:

1. *Bound preserving for $\bar{u}^{n+1}$: i.e., the range of $\bar{u}^{n+1}$ is in $(0, M)$.*

2. *Mass conservation: i.e., $\int_\Omega u^{n+1} dx = \int_\Omega u^0 dx$.*

3. *Unconditionally energy dissipation with a modified energy defined by $\bar{E}^n = E_0(\bar{\phi}^{n+1}) + r^n$: More precisely, if $\bar{E}^n \geq 0$, we have $\bar{E}^{n+1} \geq 0$, $\xi^{n+1} \geq 0$ and*

$$\bar{E}^{n+1} - \bar{E}^n = -\xi^{n+1} \int_\Omega [\frac{1}{f''(\bar{u}^{n+1})} \Big(\nabla \frac{\delta E}{\delta u}(\bar{u}^{n+1})\Big)^2 + \tau \Big(\frac{\bar{\phi}^{n+1} - \bar{\phi}^n}{\delta t}\Big)^2] dx \leq 0. \tag{5.71}$$

4. *There exists constant M_k, such that*

$$\sqrt{E_0[\phi^n]} = \sqrt{\int_\Omega (\frac{\mu}{2}|\nabla \phi^n|^2 + \frac{\alpha}{4}(\phi^n)^2) dx} \leq M_k, \ \forall n. \tag{5.72}$$

Proof. The proof is essentially the same as that of Theorem 5.3.1. For the readers' convenience, we still carry it out below.

We derive from (5.62) that the range of $\bar{u}^{n+1}$ is in $(0, M)$.

Noting the definition of coefficients α_k and A_k in Section 2, it follows from (5.70) and (5.63) that

$$\alpha_k \lambda^{n+1} \int_\Omega \bar{u}^{n+1} dx = \alpha_k \int_\Omega u^0 dx, \tag{5.73}$$

which implies $\lambda^{n+1} > 0$, and consequently $u^{n+1} > 0$. Furthermore, along with (5.64), it also implies $\int_\Omega u^{n+1} dx = \int_\Omega u^0 dx$.

It follows from (5.66) that

$$E_0(\bar{\phi}^{n+1}) + r^{n+1} = \frac{E_0(\bar{\phi}^n) + r^n}{1 + \frac{\delta t \int_\Omega [\frac{D}{f''(\bar{u}^{n+1})} \left(\nabla \frac{\delta E}{\delta u}(\bar{u}^{n+1})\right)^2 + \tau \left(\frac{\bar{\phi}^{n+1} - \bar{\phi}^n}{\delta t}\right)^2] dx}{E(\bar{u}^{n+1}, \bar{\phi}^{n+1}) + C_0}} \geq 0.$$

Therefore, we derive from (5.67) that $\xi^{n+1} \geq 0$, which, together with (5.66), implies the energy dissipation.

Denote $M := \bar{E}^0$, then (5.71) implies $\bar{E}^n \leq M$, $\forall n$. Now, it follows from (5.67) and (5.56) that

$$|\xi^{n+1}| = \frac{\bar{E}^{n+1}}{E(\bar{u}^{n+1}, \bar{\phi}^{n+1}) + C_0} \leq \frac{M}{E_0(\bar{\phi}^{n+1}) + 1}. \tag{5.74}$$

Since $\eta_k^{n+1} = 1 - (1 - \xi^{n+1})^k$, there exists a polynomial P_{k-1} of degree $k-1$ and a constant $M_k > 0$ such that

$$|\eta_k^{n+1}| = |\xi^{n+1} P_{k-1}(\xi^{n+1})| \leq \frac{M_k}{E_0(\bar{\phi}^{n+1}) + 1}. \tag{5.75}$$

Therefore, by the fact that $\sqrt{A} \leq A + 1$ for all $A \geq 0$, we obtain

$$\sqrt{E_0[\phi^{n+1}]} = |\eta_k^{n+1}| \sqrt{E_0[\bar{\phi}^{n+1}]} \leq M_k.$$

□

We only consider the semi-discretization in time in this paper. As for fully discretizations, we have the following remarks: *We emphasize that both $\bar{u}^{n+1}$ (resp. $\bar{\phi}_{\mathrm{i}}^{n+1}$) and u^{n+1} (resp. ϕ^{n+1}) are k-th order approximation to $u(\cdot, t_{n+1})$ (resp. $\phi(\cdot, t_{n+1})$). While only the range of $\bar{u}^{n+1}$ is guaranteed in $(0, M)$, the range of u^{n+1} is in $(0, M + O(\delta t^k))$.*

The positivity of $\bar{u}^{n+1}$ and u^{n+1} will be preserved with any spatial approximation of the schemes (5.61)-(5.68).

It is also clear from the proof of the above theorem that the mass conservation and the energy dissipation (5.71) still hold for any fully discrete schemes.

One can easily extend these schemes to deal with Keller-Segel equations with multiple organisms. We leave the detail to the interested readers.

5.5 Numerical examples

In this section, we provide some numerical examples to validate our numerical schemes.

5.5.1 Allen-Cahn equation with a singular potential

We first use the schemes presented in Section 2 to solve the Allen-Cahn equation with a singular potential. In all examples for the Allen-Cahn equation, we consider problems with periodic boundary conditions and use a Fourier-spectral method to discretize in space.

Example 1. We consider the Allen-Cahn equation [30]

$$\partial_t u = -\frac{\delta E}{\delta u} = \epsilon^2 \Delta u + \lambda u - \ln(1+u) + \ln(1-u), \tag{5.76}$$

where $\epsilon > 0$, $\lambda > 0$ and

$$E(u) = \int_\Omega \Big(\frac{\epsilon^2}{2} |\nabla \phi|^2 - \frac{\lambda}{2} u^2 + (1+\phi)\ln(1+u) + (1-u)\ln(1-u) \Big) dx, \tag{5.77}$$

is the free energy with a singular potential. The well posedness of the above equation requires that $u \in (-1, 1)$.

We use the transformation $u = \tanh(v)$, in the scheme (5.6)-(5.10).

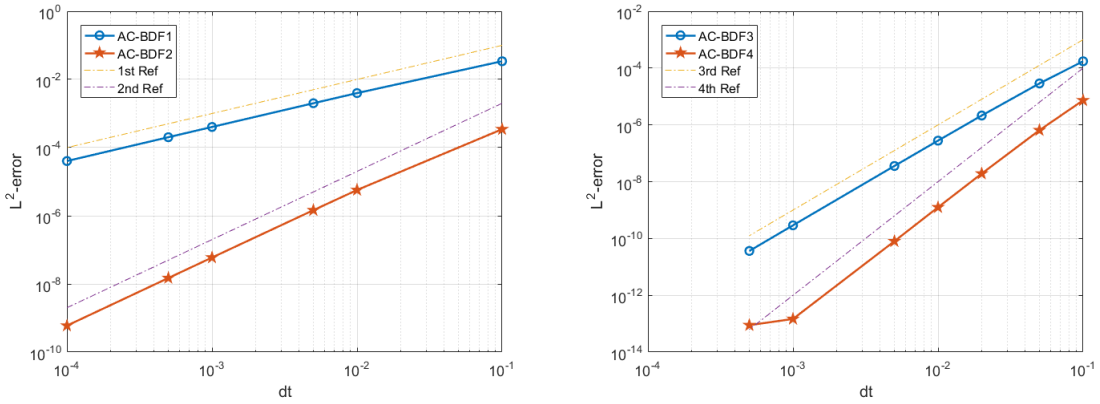
We first test the accuracy with the following exact solution and the corresponding external forcing f

$$u(x, y, t) = \Big(\exp(-\sin^2(\pi x)) - \exp(-\sin^2(\pi y)) \Big) \sin(t),$$

$$f = \partial_t u + \frac{\delta E}{\delta u}.$$

The parameters are chosen as $\epsilon = 0.1$, $\lambda = 3$ and the computational domain is $(0, 2) \times (0, 2)$. Fourier spectral method with 96×96 modes is used for special discretization. we plot in Figure 5.1 (a) the errors of the first- and second-order schemes at $t_n = 1$, and in Figure 5.1

(b), the errors of the third- and fourth-order schemes at $t_n = 1$. Expected convergence rates are observed for all cases.



(a) BDF1 and BDF2 for Allen-Cahn

(b) BDF3 and BDF4 for Allen-Cahn

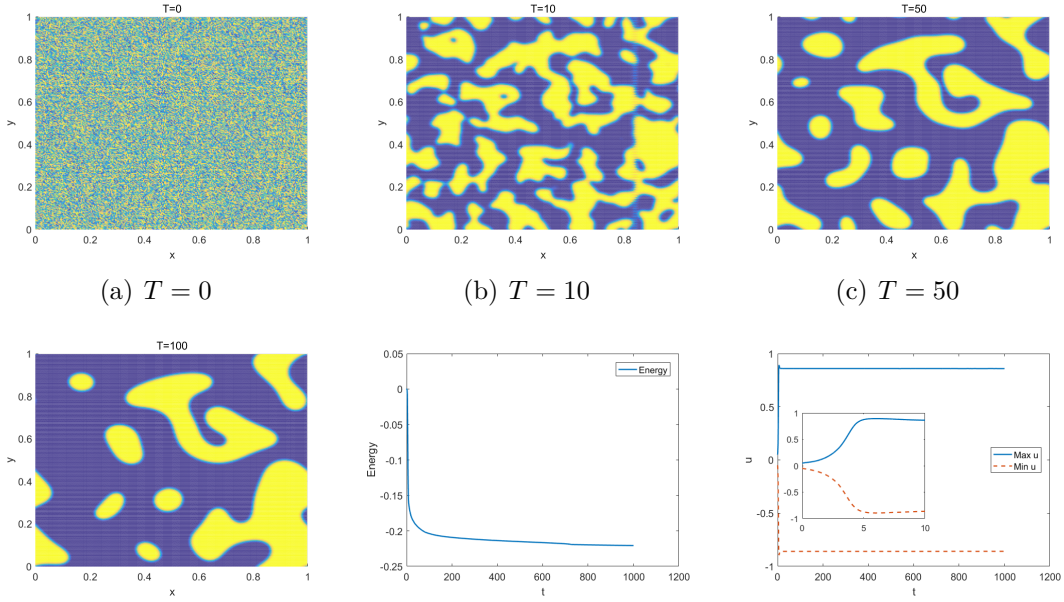
Figure 5.1. (*Example* 1.) Accuracy test for the Allen-Cahn equation using the new SAV/BDFk schemes ($k = 1, 2, 3, 4$).

Next, we consider the spinodal decomposition of a homogeneous mixture into two coexisting phases governed by the Allen-Cahn equation. The parameters are chosen as $\epsilon = 0.005$, $\lambda = 3$ and the computational domain is $(0, 1) \times (0, 1)$. The time step is set to $\delta t = 0.001$. Fourier spectral method with 256×256 modes is used for space discretization. The initial condition is chosen as a random variable with uniform distribution in $[-0.05, 0.05]$. We plot the evolution of energy, the evolution of $\max u$, $\min u$ and four snapshots in Figure 5.2.

5.5.2 Two-component PNP system

We present here numerical results of using the scheme (5.34)-(5.41) to solve the two-component PNP system (5.22).

Example 2. We test accuracy by considering the two-component PNP system (5.24), i.e. we fix $z_1 = 1$, $z_2 = -1$ and $\chi_1 = \chi_2 = 1$ in (5.22). We first consider the following



(a) $T = 0$ (b) $T = 10$ (c) $T = 50$

(d) $T = 100$ (e) evolution of origina energy (f) evolution of u_{max} and u_{min}

Figure 5.2. *Example* 1. Spinodal decomposition by the Allen-Cahn equation. The simulation is obtained with $\delta t = 0.001$ using the scheme (5.6)-(5.10).

manufactured exact solutions in $\Omega = (-0.5, 0.5) \times (-0.5, 0.5)$ with suitable external forcing:

$$c_1(x, y, t) = 1.1 + \sin(\pi x)\sin(\pi y)\sin(t), \tag{5.78a}$$

$$c_2(x, y, t) = 1.1 - \sin(\pi x)\sin(\pi y)\sin(t), \tag{5.78b}$$

$$\phi(x, y, t) = \frac{1}{\pi^2}\sin(\pi x)\sin(\pi y)\sin(t). \tag{5.78c}$$

In this example, we use Legendre spectral-Galerkin method and $(N_x, N_y) = (40, 40)$. Other parameters are $D_1 = D_2 = 1$. Define the L^2-error at t_n as $\sqrt{\|c_1^n - c_1(t_n)\|^2 + \|c_2^n - c_2(t_n)\|^2}$, we plot in Figure 5.3 (a) the errors of the first- and second-order schemes at $t_n = 1$, and in Figure 5.3 (b), the errors of the third- and fourth-order schemes at $t_n = 10$. Expected convergence rates are observed for all cases.

Next, we test the accuracy in the computational domain $\Omega = (0, 2\pi) \times (0, 2\pi)$ with periodic boundary condition and the initial conditions are given by

$$c_1(x, y, 0) = 1.1 + \sin(x)\cos(y), \tag{5.79a}$$

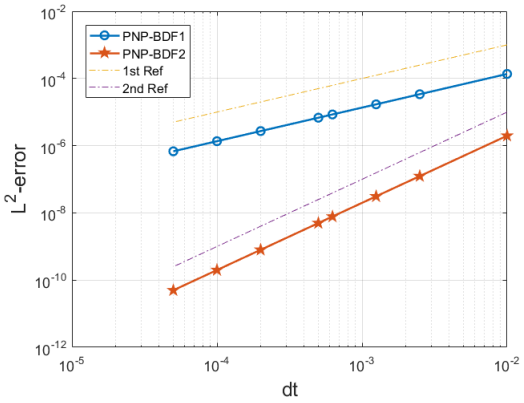
$$c_2(x, y, 0) = 1.1 - \sin(x)\cos(y). \tag{5.79b}$$

In this example, we use Fourier-spectral method to discretize in space and $(N_x, N_y) = (128, 128)$. Other parameters are $D_1 = D_2 = 1$. We generate the reference solution by the fourth-order scheme with $\delta t = 0.0001$. Define the L^2-error at t_n as above, we plot in Figure 5.3 (c) the errors of the first- and second-order schemes at $t_n = 0.1$, and in Figure 5.3 (d), the errors of the third- and fourth-order schemes at $t_n = 0.1$. Expected convergence rates are observed for all cases.

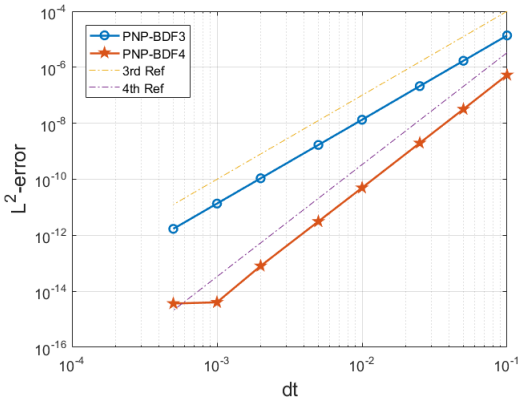
Example 3. In this example, we test the so-called Gouy-Chapman model [85] which is used to describe the evolution of the distributions of the ions.

We consider the PNP system (5.22) in $(-1, 1)$ with the following parameters: $D_1 = D_2 = 1$, $z_1 = 1$, $z_2 = -1$, and $\chi_1 = 3.1$, $\chi_2 = 125.4$. The boundary conditions for c_i and ϕ are given as

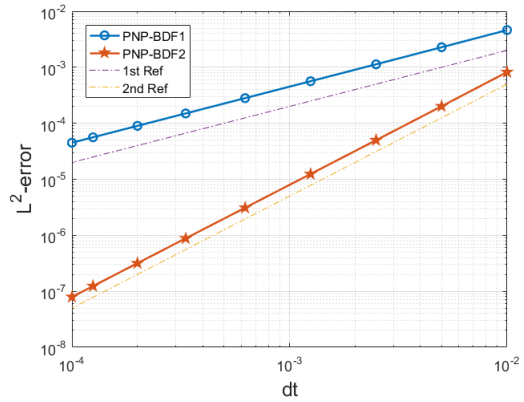
$$\partial_x c_\mathrm{i} + z_\mathrm{i}\chi_1 c_\mathrm{i}\partial_x\phi = 0, \ \mathrm{i} = 1, 2, \tag{5.80}$$



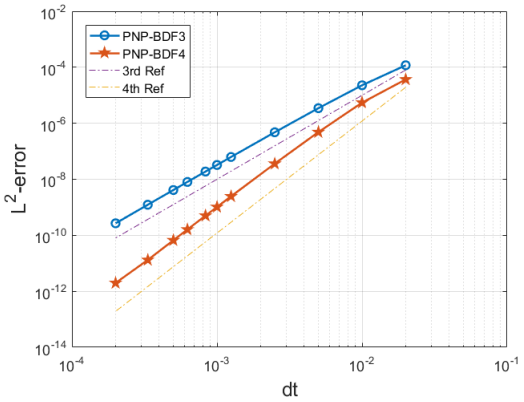
(a) BDF1 and BDF2 for PNP with known exact solution



(b) BDF3 and BDF4 for PNP with known exact solution



(c) BDF1 and BDF2 for PNP with unknown exact solution

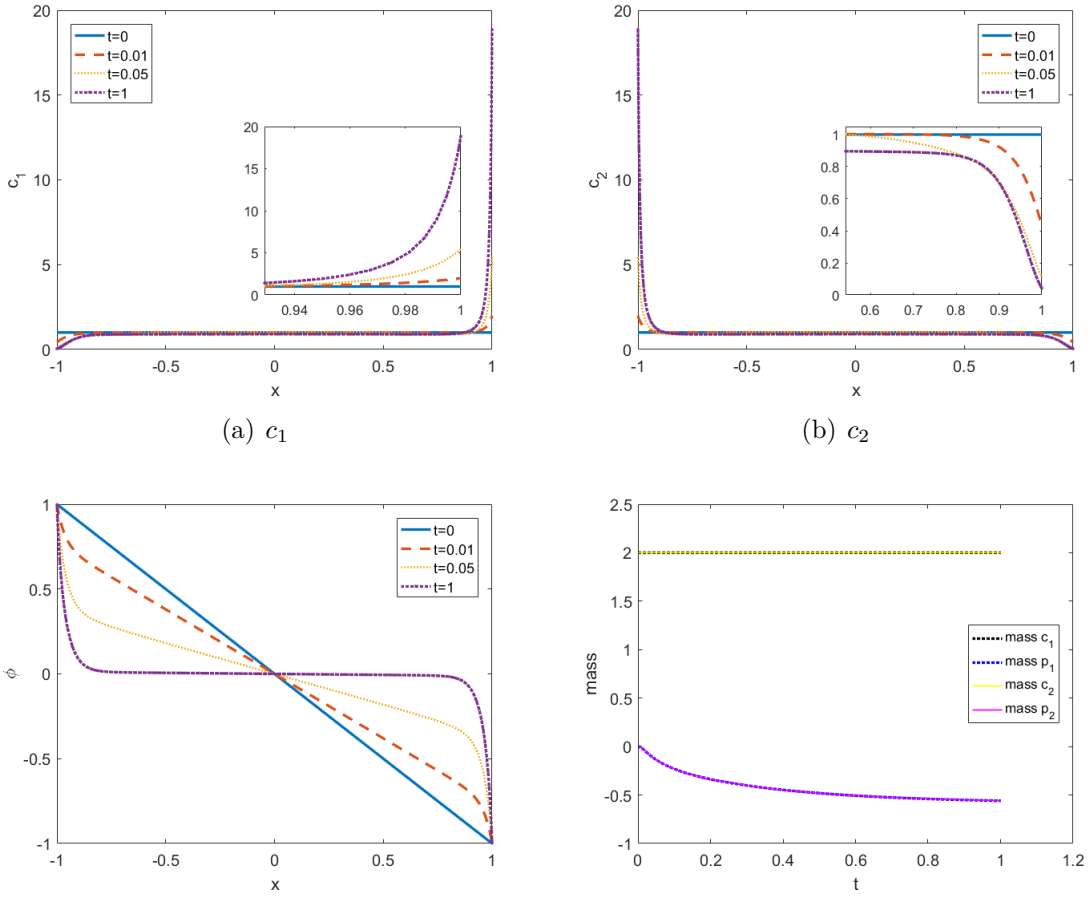


(d) BDF3 and BDF4 for PNP with unknown exact solution

Figure 5.3. *Example* 2. Accuracy test for PNP equation using the SAV/BDFk schemes ($k = 1, 2, 3, 4$).

$$\alpha\phi(t,-1) - \beta\phi_x(t,-1) = f_{-1},\ \alpha\phi(t,1) + \beta\phi_x(t,1) = f_1,\ t \geq 0, \tag{5.81}$$

with $\alpha = 1$, $\beta = 4.63 \times 10^{-5}$, $f_{-1} = 1$ and $f_1 = -1$. For space discretization, we use Legendre spectral-Galerkin method. We set $\delta t = 0.001$ and used 80 nodes in space. The initial condition on c_{i} are $c_{\mathrm{i}}(x,0) = 1, \mathrm{i} = 1,2$ for all $-1 \leq x \leq 1$. The profiles of c_1, c_2 and ϕ at different times are plotted in Fig 5.4, which are consistent with the results in [85]. In Fig 5.4(d), we also plot the mass evolution of c_{i} and p_{i} with $c_{\mathrm{i}} = \exp(p_{\mathrm{i}}), \mathrm{i} = 1,2$. We can see the masses of c_{i} are well conserved, but that of p_{i} are not.



(a) c_1 (b) c_2

(c) ϕ (d) mass evolution

Figure 5.4. *Example* 3. Gouy-Chapman model: Profiles of c_1, c_2 and ϕ

5.5.3 Keller-Segel equations

In this subsection, we present numerical results of using scheme (5.61)-(5.68) to solve the Keller-Segel equations (5.49).

Example 4. We test the accuracy of the scheme. First consider the one-specie parabolic-elliptic ($\tau = 0$) Keller-Segel equations (5.49) in $\Omega = (-0.5, 0.5) \times (-0.5, 0.5)$ with external forcing such that the exact solutions are given by

$$u(x, y, t) = \sin(\pi x) \sin(\pi y) \sin(t) + 1.1, \tag{5.82a}$$

$$\phi(x, y, t) = \frac{1}{2\pi^2} \sin(\pi x) \sin(\pi y) \sin(t) + 1.1. \tag{5.82b}$$

Other parameters are $D = \gamma = \mu = \alpha = \chi = 1$, $M = 5$. We use Legendre spectral-Galerkin method and $(N_x, N_y) = (40, 40)$ in space. Define the L^2-error as $\sqrt{\|u^n - u(t_n)\|^2 + \|\phi^n - \phi(t_n)\|^2}$, we plot in Figure 5.5 (a) the errors at $t_n = 1$ for the first- and second-order schemes, and in Figure 5.5 (b) the errors at $t_n = 10$ for the third- and fourth-order schemes.

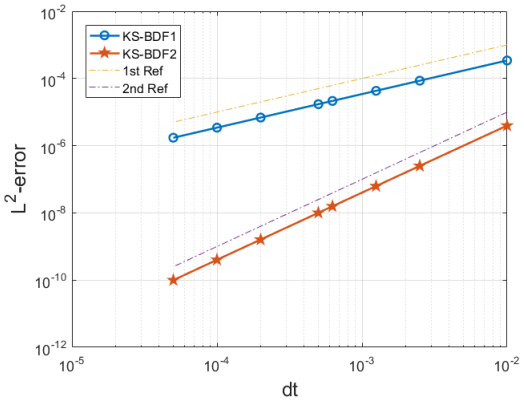
Next, we test the accuracy in $\Omega = (0, 2\pi) \times (0, 2\pi)$ with periodic boundary condition and the initial conditions are given by

$$u(x, y, 0) = \sin(x) \sin(y) + 1.1. \tag{5.83}$$

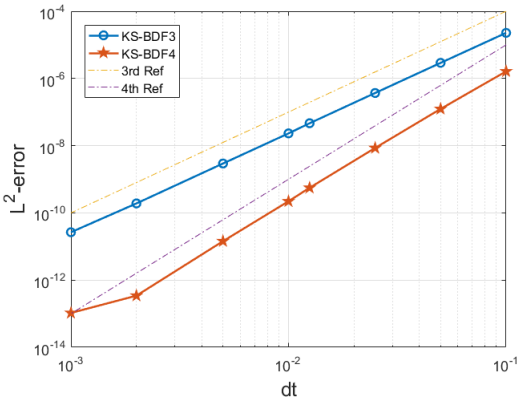
In this example, we use Fourier-spectral method to discretize in space and $(N_x, N_y) = (128, 128)$. Other parameters are $D = \gamma = \mu = \alpha = \chi = 1$, $M = 3$. We generate the reference solution by the fourth-order scheme with $\delta t = 0.0001$. Define the L^2-error as above, we plot in Figure 5.5 (c) the errors at $t_n = 0.1$ for the first- and second-order schemes, and in Figure 5.5 (d) the errors at $t_n = 0.1$ for the third- and fourth-order schemes. As in the Allen-Cahn case and the PNP case, the expected convergence rates are observed for all cases.

Example 5. In this example, we consider the one-specie parabolic-elliptic ($\tau = 0$) Keller-Segel equations with the following initial condition

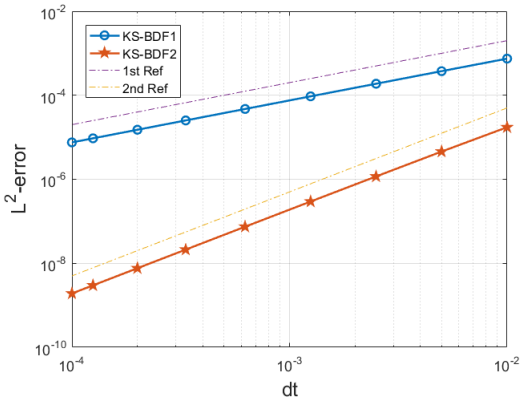
$$u(x, y, 0) = 4 \exp\Big(- \frac{(x - L/2)^2 + (y - L/2)^2}{4} \Big) \tag{5.84}$$



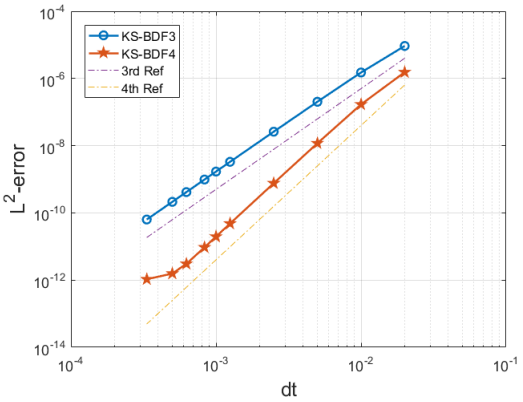
(a) BDF1 and BDF2 for Keller-Segel with known exact solution



(b) BDF3 and BDF4 for Keller-Segel with known exact solution



(c) BDF1 and BDF2 for Keller-Segel with unknown exact solution



(d) BDF3 and BDF4 for Keller-Segel with unknown exact solution

Figure 5.5. *Example* 4. Accuracy test for Keller-Segel equations using the SAV/BDFk (5.61)-(5.68) ($k = 1, 2, 3, 4$).

such that the total mass is large enough that chemotaxis happens, in $(0, 2\pi) \times (0, 2\pi)$ with the homogeneous Neumann boundary conditions. We use Legendre spectral-Galerkin method method with $(N_x, N_y) = (64, 64)$ nodes to discretize in space, and the second-order scheme with time step $\delta t = 0.001$. The parameters are chosen as $D = \gamma = \mu = 1$, $\chi = 1$, $M = 100$, $\alpha = 0.1$ and $L = 2\pi$.

We carry out simulation until the system reaches steady state at $t = 8$. Several snapshots of concentration at different times are shown in Figure 5.6, where we plot the snapshots by using smaller time steps and more nodes in the right hand side , and evolutions of max u, mass of u, mass of v and energy are shown in Figure 5.7. These results agree well with those in [98] computed with a nonlinear scheme. In particular, the energy is dissipative at all time, and the mass of u is conserved up to machine accuracy.

Example 6. We consider the one specie parabolic-elliptic system with an initial condition with two bulges, given by

$$u(x, y, 0) = 2\exp(-\frac{(x - 3L/8)^2 + (y - 3L/8)^2}{4}) + 2\exp(-\frac{(x - 5L/8)^2 + (y - 5L/8)^2}{4}) \tag{5.85}$$

with $L = 4\pi$. We take $M = 50$ while all other settings are the same as Example 5. We use the third-order scheme, and plot the evolution of energy, maximum concentration and four snapshots of u in Figure 5.8. We observe that the energy is dissipative at all times, the maximum of u increases while the support of u shrinks to maintain the mass conservation.

Example 7. In this example, we consider the parabolic-elliptic Keller-Segel system with two species:

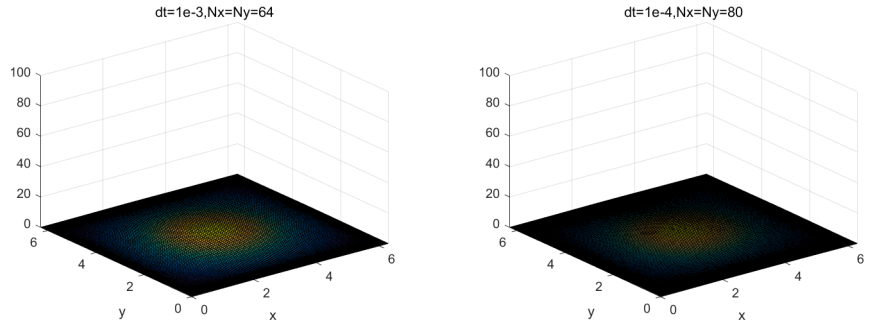
$$\frac{\partial u_1}{\partial t} = D_1\Big(\gamma_1 \Delta u_1 - \chi_1 \nabla \cdot (\eta_1(u_1)\nabla\phi)\Big), \tag{5.86a}$$

$$\frac{\partial u_2}{\partial t} = D_2\Big(\gamma_2 \Delta u_2 - \chi_2 \nabla \cdot (\eta_2(u_2)\nabla\phi)\Big), \tag{5.86b}$$

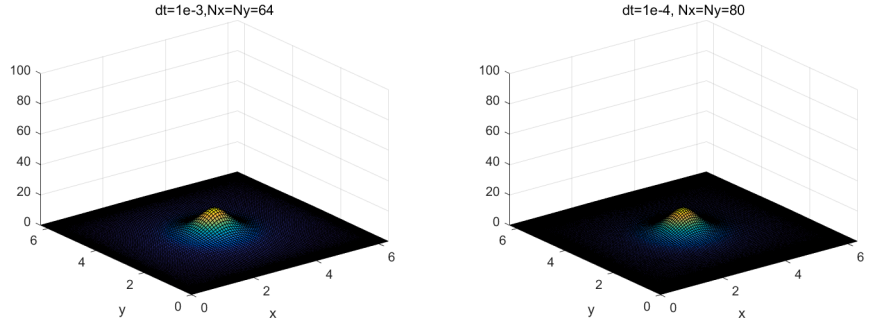
$$0 = \mu\Delta\phi - \alpha\phi + \chi_1 u_1 + \chi_2 u_2, \tag{5.86c}$$

with the initial conditions

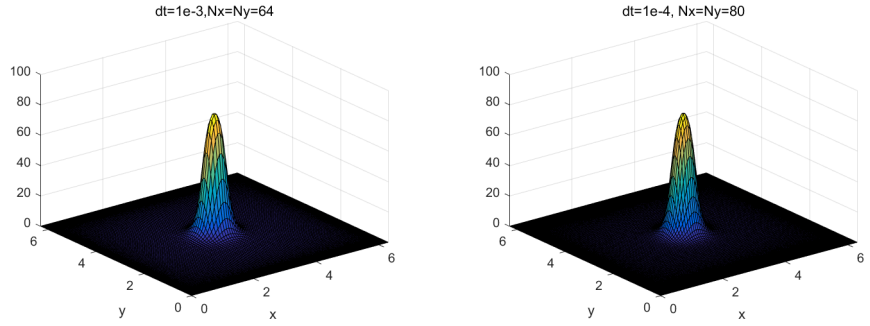
$$u_1(x, y, 0) = u_2(x, y, 0) = \phi(x, y, 0) = 4\exp\Big(-\frac{(x - L/2)^2 + (y - L/2)^2}{4}\Big). \tag{5.87}$$



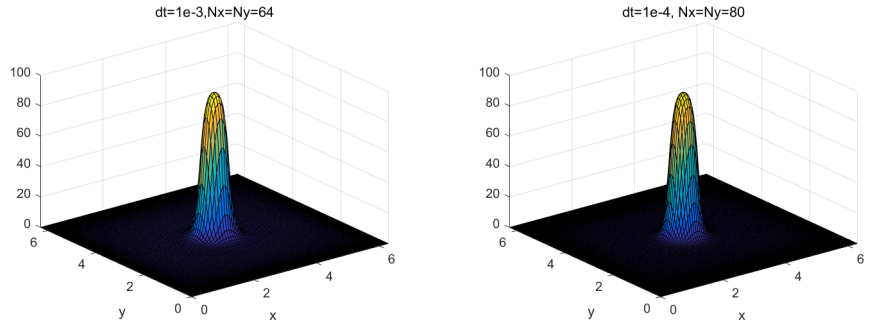
(a) T=0 (b) T=0



(c) T=1 (d) T=1

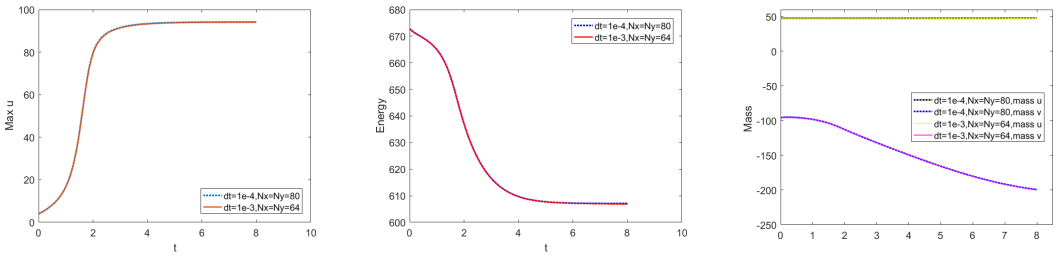


(e) T=2 (f) T=2



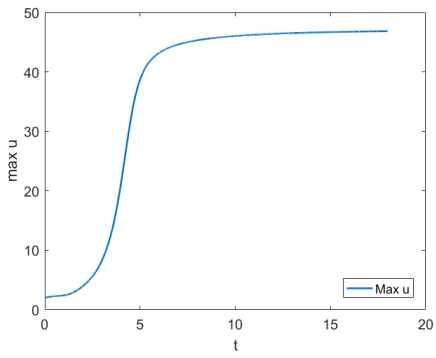
(g) T=8 (h) T=8

Figure 5.6. *Example* 5. Simulation of Keller-Segel equations with chemotaxis.

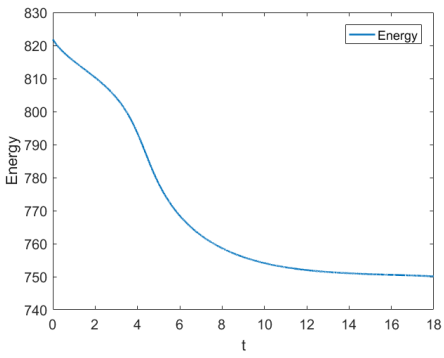


(a) max u evolution (b) energy evolution (c) mass evolution

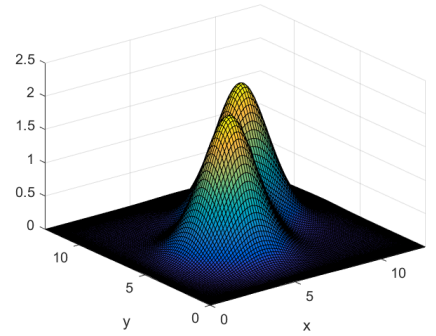
Figure 5.7. *Example* 5. Simulation of Keller-Segel equations with chemotaxis.



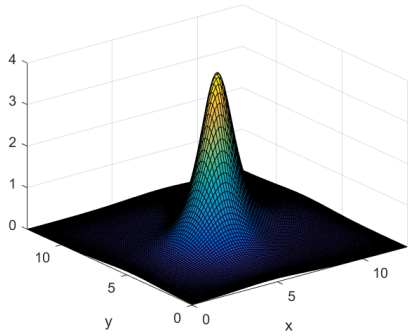
(a) Max u evolution



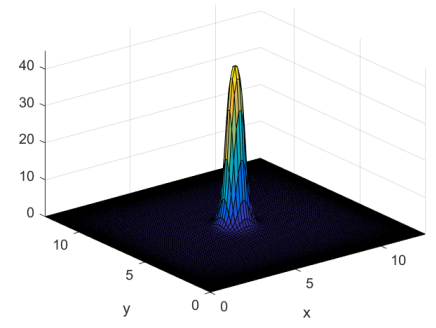
(b) Energy evolution



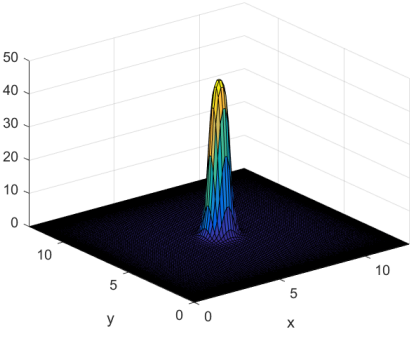
(c) t=0.1



(d) t=2



(e) t=6



(f) t=18

Figure 5.8. *Example* 6. Simulation of Keller-Segel equations with initial condition (5.85)

The parameters are chosen as $D_1 = D_2 = \gamma_1 = \gamma_2 = \mu = \chi_1 = 1$, $\alpha = 0.1$ with all other settings are the same as Example 5. We use the first order scheme for this example. The results with two different chemotactic sensitivities with $\chi_2 = 0.1$ and $\chi_2 = 0.01$ are plotted in Figure 5.9 and 5.10 respectively.

In both cases, we observe accumulation for u_1, while for u_2, it diffuses first and then accumulates in the case $\chi_2 = 0.1$, and it keeps diffusing in the case $\chi_2 = 0.01$. These results are consistent with the results in [98].

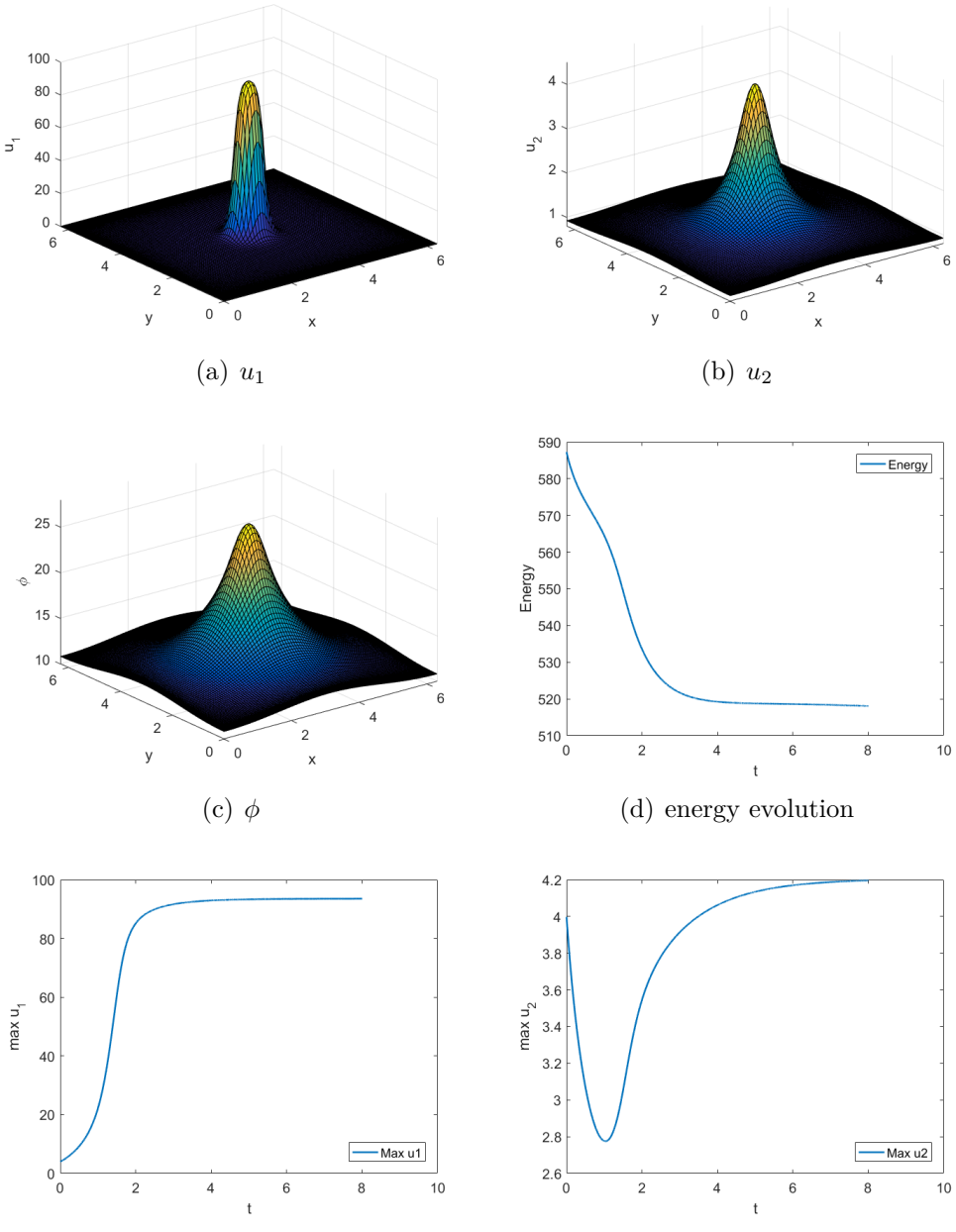
5.6 Conclusion of this chapter

For PDEs whose solutions are required to be positive or in a prescribed range, it is of critical importance to construct numerical schemes which are positivity or bound preserving. If the PDEs are also energy dissipative and/or mass conservative, it is important that the numerical schemes would be energy dissipative and/or mass conservative at the discrete level.

In this chapter, we proposed a new approach to construct linear, positivity/bound preserving and unconditionally energy stable schemes for general dissipative systems whose solutions are positivity/bound preserving. The essential ideas of this new approach are (i) to first make a function transform so that the solution will always be positivity/bound preserving, and (ii) apply a new SAV approach presented in [13] to the transformed system and the original energy dissipation law to construct efficient and accurate time discretization schemes.

The resulting schemes enjoy remarkable properties such as positivity/bound preserving, unconditionally energy stable, can achieve high-order and with computational complexity similar to a semi-implicit scheme. We applied this approach to Allen-Cahn equation with a singular potential, and to Keller-Segel and Poisson-Nernst-Planck (PNP) equations which can be classified as Wasserstein gradient flows with an additional property of mass conservation.

While we only discussed semi-discretization in time in this chapter, we pointed out that the energy dissipation, positivity or bound preserving and mass conservation can all be naturally carried over to consistent fully discretizations.

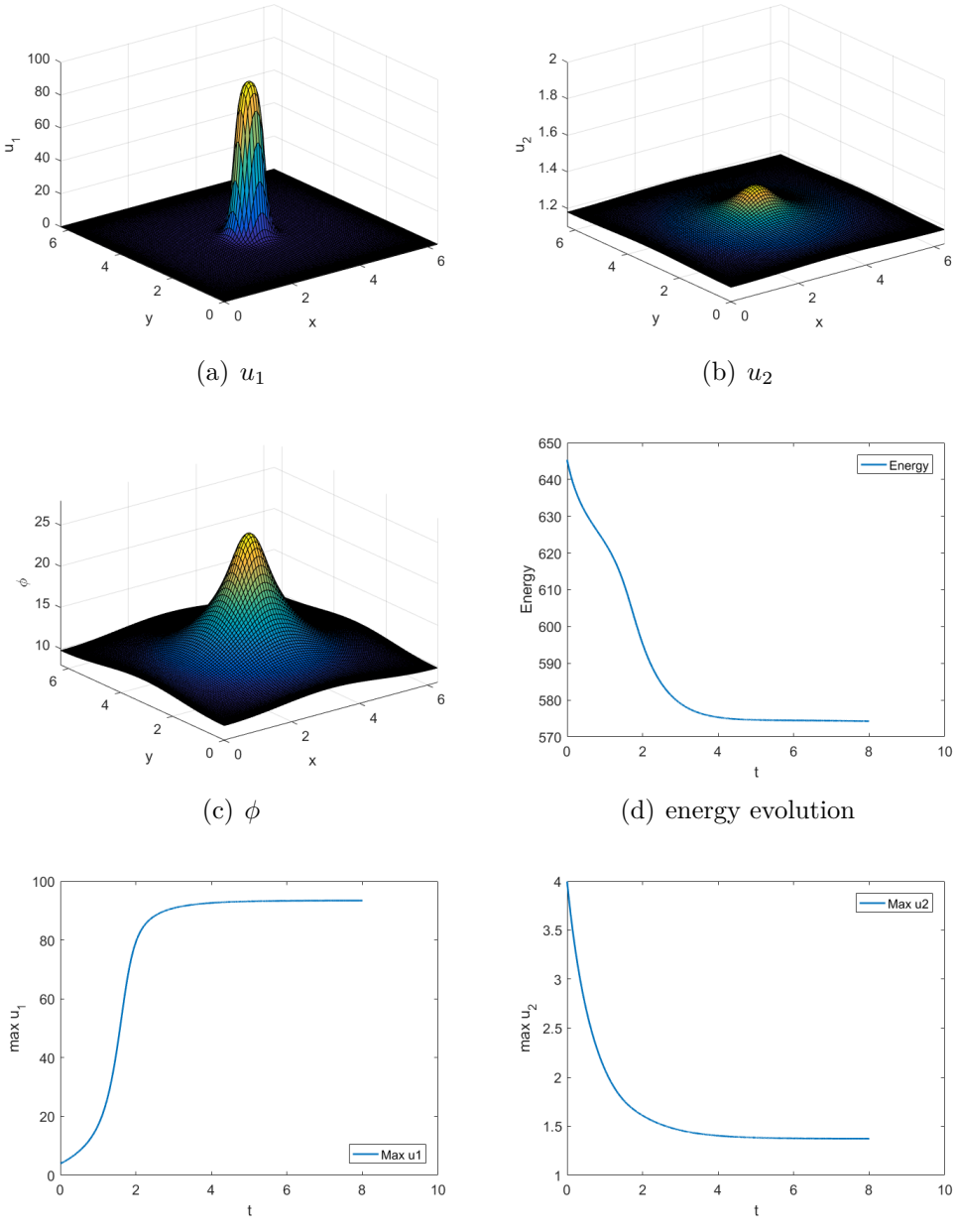


(a) u_1 (b) u_2

(c) ϕ (d) energy evolution

(e) max u_1 evolution (f) max u_2 evolution

Figure 5.9. *Example* 7. Simulation with $\chi_2 = 0.1$



(a) u_1 (b) u_2

(c) ϕ (d) energy evolution

(e) max u_1 evolution (f) max u_2 evolution

Figure 5.10. *Example* 7. Simulation with $\chi_2 = 0.01$

6. POSITIVITY/BOUND PRESERVING SAV SCHEMES: WITH APPLICATION TO FOURTH ORDER EQUATION

In this chapter, we continue to apply the techniques presented in the last chapter to the fourth order equation: the thin film type equation and Cahn-Hilliard equation with singular potential. In particular, we show numerical examples to demonstrate the effect of functional transformation and SAV. Most of the results in this chapter are extracted from a working paper with Dr. Jie Shen and Dr. Ke Wu.

6.1 Introduction

In order to clearly describe our ideas, we consider the following fourth order nonlinear equations in a bounded domain $\Omega \subset \mathcal{R}^d (d = 1, 2, 3)$ arise in the thin films and phase field models [110], whose general form can be written as

$$\phi_t = m\nabla \cdot (f(\phi)\nabla\mu), \tag{6.1a}$$

$$\mu := \frac{\delta E}{\delta \phi} = -\frac{1}{\alpha}\Delta\phi + h(\phi), \tag{6.1b}$$

with either periodic or homogeneous Neumann boundary condition, where m is a positive constant, $\phi > 0$ and $f(\phi)$ is a positive mobility function and E is the dimensionless free energy given as

$$E(\phi) = \int_\Omega \frac{1}{2\alpha}|\nabla\phi|^2 d\Omega + \int_\Omega H(\phi) d\Omega = E_0(\phi) + E_1(\phi). \tag{6.2}$$

Throughout this chapter, we focus on the following two types equations:

- Case 1. The thin film equations [111]: Set $f(\phi) = \phi^n$, $n > 0$ and $h(\phi) = 0$ in (6.1).

- Case 2. The Cahn-Hilliard equation with singular potential [112]: Set $f(\phi) = \phi(1-\phi)$ in (6.1) and $H(\phi) = \phi\log\phi + (1-\phi)\log(1-\phi) + 3\phi(1-\phi)$ in (6.2).

In both cases, with $f(\phi) > 0$, we have the following energy dissipation law

$$\frac{dE}{dt} = -m \int_\Omega f(\phi)|\nabla \mu|^2 d\Omega \leq 0. \tag{6.3}$$

Numerical solutions of (6.1) is necessary positivity preserving ($\phi > 0$) in Case 1 and bound preserving ($0 < \phi < 1$) in Case 2 to ensure the equations are well-defined and the energy law (6.3) holds true.

6.2 Positivity/bound preserving SAV scheme

In this section, we present two methods to construct our positivity/bound preserving SAV for fourth-order equation: in the first method, we need to solve one fourth-order equation, which is highly efficient by using Fourier spectral method; in the second method, we need to solve two coupled second-order equations, which can be implemented easily by using finite element method.

6.2.1 Method1: solve one fourth-order equation

Following the ideas in last chapter, which deals with the second order equations, we can first make a suitable invertible mapping $\phi = T(u)$, then the equation (6.1) on ϕ is transformed to an equation on u and we finally get ϕ in a desired interval.

In particularly, we have

$$\Delta^2 \phi = T'(u)\Delta^2 u + T''(u)C_2(u) + T'''(u)C_3(u) + T''''(u)C_4(u), \tag{6.4}$$

with C_2, C_3 and C_4 defined in the appendix.

After the transformation, the equation on u becomes

$$\frac{\partial u}{\partial t} = -\frac{m}{\alpha} f(\phi)\Delta^2 u + \frac{m}{T'(u)} g(u, \phi), \tag{6.5}$$

where $f(\phi)$ and $g(u,\phi)$ will be treated explicitly in our numerical schemes and defined as

$$g(u,\phi) = f(\phi)\Delta h(\phi) + \nabla f(\phi)\cdot\nabla\mu - \frac{f(\phi)}{\alpha}T''(u)C_2(u) - \frac{f(\phi)}{\alpha}T'''(u)C_3(u) - \frac{f(\phi)}{\alpha}T''''(u)C_4(u). \tag{6.6}$$

As shown below, for the constant mobility case, $f(\phi) = 1$, we can construct linear schemes with constant coefficients for (6.5) and for the variable mobility function f, we can construct linear schemes with variable coefficients. On the other hand, for the variable mobility case, similar to the method presented in [113], we can first rewrite (6.1a) as

$$\partial_t\phi = m\nabla\cdot((S + f(\phi) - S)\nabla\mu), \tag{6.7}$$

where S is a positive constant served as a stabilizer. Then equation on u becomes

$$\frac{\partial u}{\partial t} = -\frac{mS}{\alpha}\Delta^2 u + \frac{m}{T'(u)}G_s(u,\phi), \tag{6.8}$$

where $G_s(u,\phi)$ will be treated explicitly in our numerical schemes and defined as

$$G_s(u,\phi) = \nabla\cdot((f(\phi) - S)\nabla\mu) + S\Delta h(\phi) + \frac{S}{\alpha}T''(u)C_2(u) - \frac{S}{\alpha}T'''(u)C_3(u) - \frac{S}{\alpha}T''''(u)C_4(u). \tag{6.9}$$

As a result, we can also construct linear schemes with constant coefficients for the variable mobility function. Note that T is an invertible mapping, we know $T(u) \neq 0$ and hence both (6.5) and (6.8) are well-defined.

Now, we construct our new SAV schemes for (6.5), one can deal with (6.8) by the same way. Note that in two cases we consider above, $H(\phi)$ is either 0 or a strictly convex function, hence there exists constant $\underline{C} > 0$ such that

$$E_1(\phi) \geq -\underline{C} + 1. \tag{6.10}$$

To construct our numerical schemes, we introduce $r(t) = E(\phi) + C_0$ with $C_0 = 2\underline{C} + |E(\phi^0)|$ and rewrite (6.1) and (6.3) as

$$\frac{\partial u}{\partial t} = -\frac{m}{\alpha} f(\phi)\Delta^2 u + \frac{m}{T'(u)} g(u,\phi), \tag{6.11a}$$

$$\phi = T(u), \tag{6.11b}$$

$$\frac{dr}{dt} = -m\frac{r}{E(\phi)+C_0}\Big(f(\phi)\nabla\mu, \nabla\mu\Big). \tag{6.11c}$$

With $r(t) = E(\phi) + C_0$, it is clear that the above system is equivalent to (6.1) with (6.3) while discretizing the above systems can allow us to construct energy dissipative and bound preserving numerical schemes. We construct below k-th order BDF-Adams-Bashforth SAV schemes with stabilizer terms for (6.5) in a uniform setting: treat the linear term implicitly and use Adams-Bashforth extrapolation to deal with all nonlinear terms.

More precisely, given r^n and $(u^{\mathrm{j}}, \phi^{\mathrm{j}})$ for $\mathrm{j} = n, \cdots, n-k+1$, we find $(\phi^{n+1}, u^{n+1}, r^{n+1}, \xi^{n+1})$ such that

$$\frac{\alpha_k u^{n+1} - A_k(u^n)}{\delta t} + \frac{m}{\alpha} f(B_k(\phi^n))\Delta^2 u^{n+1} = \frac{m}{T'(B_k(u^n))} g\Big(B_k(u^n), B_k(\phi^n)\Big), \tag{6.12a}$$

$$\bar{\phi}^{n+1} = T(u^{n+1}), \tag{6.12b}$$

$$\frac{r^{n+1} - r^n}{\delta t} = -m\frac{r^{n+1}}{E(\bar{\phi}^{n+1}) + C_0}\Big(f(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1}, \nabla\bar{\mu}^{n+1}\Big), \tag{6.12c}$$

$$\xi^{n+1} = \frac{r^{n+1}}{E(\bar{\phi}^{n+1}) + C_0},\ \eta_k^{n+1} = 1 - (1-\xi^{n+1})^{I_k}, I_k = \begin{cases} k+1,\ k \text{ odd} \\ k,\ k \text{ even} \end{cases}, \tag{6.12d}$$

$$\phi^{n+1} = \eta_k^{n+1}\bar{\phi}^{n+1}, \tag{6.12e}$$

where the constant α_k, I_k, operators A_k, B_k are defined by

BDF1:

$$\alpha_1 = 1,\ A_1(v^n) = v^n,\ B_1(h^n) = h^n; \tag{6.13}$$

BDF2:

$$\alpha_2 = \frac{3}{2}, \quad A_2(v^n) = 2v^n - \frac{1}{2}v^{n-1}, \quad B_2(h^n) = 2h^n - h^{n-1}; \tag{6.14}$$

BDF3:

$$\alpha_3 = \frac{11}{6}, \quad A_3(v^n) = 3v^n - \frac{3}{2}v^{n-1} + \frac{1}{3}v^{n-2}, \quad B_3(h^n) = 3h^n - 3h^{n-1} + h^{n-2}; \quad (6.15)$$

BDF4:

$$\alpha_4 = \frac{25}{12}, \quad A_4(v^n) = 4v^n - 3v^{n-1} + \frac{4}{3}v^{n-2} - \frac{1}{4}v^{n-3}, \quad B_4(h^n) = 4h^n - 6h^{n-1} + 4h^{n-2} - h^{n-3}. \quad (6.16)$$

The formulae for $k = 5$ and $k = 6$ can be derived similarly.

Several remarks are in oder:

- We choose I_k to be even and satisfies $I_k \geq k$ to guarantee the whole scheme is k-th order accuracy and $0 < \eta_k^{n+1} < 1$. As a result, both $\bar{\phi}^{n+1}$ and ϕ^{n+1} are included in I as shown below.

- (6.12a) is a k-th order approximation to (6.11a) with k-th order BDF for the linear terms and k-th order Adams-Bashforth extrapolation for the nonlinear terms. Hence, u^{n+1} is a k-th order approximation to $u(t_{n+1})$.

- (6.12c) is a first-order approximation to (6.11c). Hence, r^{n+1} is a first order approximation to $E(\phi(\cdot, t^{n+1}))$ which implies that ξ^{n+1} is a first order approximation to 1. Hence, $\eta_k^{n+1} = 1 + O(\delta t)^{k+1}$, which implies that both $\bar{\phi}^{n+1}$ and ϕ^{n+1} are k-th order approximation of $u(t_{n+1})$.

- The above scheme can be efficiently implemented as follows:

  - determine u^{n+1} from (6.12a);
  - set $\bar{\phi}^{n+1} = T(u^{n+1})$;
  - with $\bar{\phi}^{n+1}$ known, determine r^{n+1} explicitly from (6.12c), and compute ξ^{n+1} from (6.12d);
  - update u^{n+1} and ϕ^{n+1} using (6.12e), goto the next step.

The main cost is to solve u^{n+1} from (6.12a) which is a linear equation with constant coefficients for the constant mobility case and is a linear equation with variable coefficients for the variable mobility case.

It is well-known that adding suitable stabilization terms can improve the performance in real simulations. For the schemes 6.12, one effective way to add the stabilization terms in our numerical simulation is rewriting (6.12a) as

$$\frac{\alpha_k u^{n+1} - A_k(u^n)}{\delta t} + \frac{m}{\alpha} f(B_k(\phi^n))\Delta^2 u^{n+1} + S\Delta^2 u^{n+1} = \frac{m}{T(B_k(u^n))} g(B_k(u^n), B_k(\phi^n)) + S\Delta^2 u^n. \tag{6.17}$$

6.2.2 Method2: solve two coupled second-order equations

In the second method, instead of obtaining u^{n+1} by solving a fourth order equation as in (6.12a), we solve two coupled second order equation to obtain u^{n+1}. To this end, we first write (6.1) as weak form: We denote the solution space as V and we would like to find $\phi \in V$ such that

$$\Big(\phi_t, v\Big) = -m\Big(f(\phi)\nabla\mu, \nabla v\Big), \quad \forall v \in V, \tag{6.18a}$$

$$\Big(\mu, v\Big) = \frac{1}{\alpha}\Big(\nabla\phi, \nabla v\Big) + \Big(h(\phi), v\Big), \quad \forall v \in V. \tag{6.18b}$$

Let $\phi = T(u)$ with T is an invertible mapping as before, then the equation on ϕ in (6.18) is transformed to the equation on u as follows:

$$\Big(T'(u)u_t, v\Big) = -m\Big(f(\phi)\nabla\mu, \nabla v\Big), \quad \forall v \in V, \tag{6.19a}$$

$$\Big(\mu, v\Big) = \frac{1}{\alpha}\Big(T'(u)\nabla u, \nabla v\Big) + \Big(h(\phi), v\Big), \quad \forall v \in V, \tag{6.19b}$$

Now, suppose our discrete numerical solution space is V_N, we would like to find $u^{n+1}, \mu^{n+1} \in V_N$ and ϕ^{n+1} such that

$$\Big(T'(B_k(u^n))\frac{\alpha_k u^{n+1} - A_k(u^n)}{\delta t}, v^{n+1}\Big) = -m\Big(f(B_k(\phi^n))\nabla\bar{\mu}^{n+1}, \nabla v^{n+1}\Big), \quad \forall v^{n+1} \in V_N, \tag{6.20a}$$

$$\Big(\bar{\mu}^{n+1}, v^{n+1}\Big) = \frac{1}{\alpha}\Big(T'(B_k(u^n)))\nabla u^{n+1}, \nabla v^{n+1}\Big) + \Big(h(B_k(\phi^n)), v^{n+1}\Big), \quad \forall v^{n+1} \in V_N, \tag{6.20b}$$

$$\bar{\phi}^{n+1} = T(u^{n+1}), \tag{6.20c}$$

$$\frac{r^{n+1} - r^n}{\delta t} = -m\frac{r^{n+1}}{E(\bar{\phi}^{n+1}) + C_0}\Big(f(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1}, \nabla\bar{\mu}^{n+1}\Big), \tag{6.20d}$$

$$\xi^{n+1} = \frac{r^{n+1}}{E(\bar{\phi}^{n+1}) + C_0}, \ \eta_k^{n+1} = 1 - (1 - \xi^{n+1})^{I_k}, I_k = \begin{cases} k+1, \ k \text{ odd} \\ k, \ k \text{ even} \end{cases}, \tag{6.20e}$$

$$\phi^{n+1} = \eta_k^{n+1}\bar{\phi}^{n+1}, \tag{6.20f}$$

The above scheme can be efficiently implemented by the same process as (6.12).

6.2.3 Stability results

For the numerical scheme (6.12) and (6.20), we have the following bound and stability results.

Theorem 6.2.1. *Given ϕ^{i} with range in $I = (0,1)$ or $I = (0,\infty)$, $u^{\mathrm{i}} = T^{-1}(\phi^{\mathrm{i}})$ and r^{i} for $\mathrm{i} = 0, 1, \ldots, k-1$. The scheme (6.12a)-(6.12e) admits a unique solution satisfying the following properties unconditionally:*

1. *Bound preserving: i.e., the range of $\bar{\phi}^{n+1}$ and ϕ^{n+1} are in I.*

2. *Energy dissipation: Given $r^n \geq 0$ we have $r^{n+1} \geq 0$, $\xi^{n+1} \geq 0$ and*

$$r^{n+1} - r^n = -\delta t m \xi^{n+1}\Big(f(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1}, \nabla\bar{\mu}^{n+1}\Big) \leq 0. \tag{6.21}$$

3. Furthermore, for the k-th order schemes, there exists constant M_k, such that

$$\|\nabla\phi^n\| \le M_k,\ \forall n. \tag{6.22}$$

Proof. It is obviously that the range of $\bar{\phi}^{n+1}$ is in I from (6.12b).

We derive from (6.12c) that

$$r^{n+1} = r^n \Big/ \Big(1 + \frac{m\delta t}{E(\bar{\phi}^{n+1}) + C_0}\big(f(\bar{\phi}^{n+1})\nabla\bar{\mu}^{n+1}, \nabla\bar{\mu}^{n+1}\big)\Big).$$

Hence, if $r^n \ge 0$, we have $r^{n+1} \ge 0$, and (6.21) follows directly from (6.12c).

It follows from (6.12d), (6.21) and the definition of C_0 that

$$0 < \xi^{n+1} \le \frac{r^0}{E(\bar{\phi}^{n+1}) + C_0} < \frac{2|E(\phi^0)| + 2\underline{C}}{1 + |E(\phi^0)| + \underline{C}} < 2 \tag{6.23}$$

As a result, (6.12d) and (6.23) together imply

$$0 < (1-\xi^{n+1})^{I_k} < 1,\ 0 < \eta_k^{n+1} < 1. \tag{6.24}$$

Hence, the range of ϕ^{n+1} is also in I as $\phi^{n+1} = \eta_k^{n+1}\bar{\phi}^{n+1}$.

Now, it follows from (6.12d) and the assumption on $E_1(u)$ that

$$|\xi^{n+1}| = \frac{r^{n+1}}{E(\bar{\phi}^{n+1}) + C_0} \le \frac{2r^0}{\|\nabla\bar{\phi}^{n+1}\|^2 + 2}. \tag{6.25}$$

Since $\eta_k^{n+1} = 1 - (1-\xi^{n+1})^{I_k}$, there exists a polynomial P_k of degree $I_k - 1$ and a constant $M_k > 0$ such that

$$|\eta_k^{n+1}| = |\xi^{n+1}P_k(\xi^{n+1})| \le \frac{M_k}{\|\nabla\bar{\phi}^{n+1}\|^2 + 2}. \tag{6.26}$$

Therefore, by the fact $\sqrt{A} \le A + 2$ for all $A \ge 0$, we have

$$\|\nabla\phi^{n+1}\| = \eta_k^{n+1}\|\nabla\bar{\phi}^{n+1}\| \le M_k. \tag{6.27}$$

□

It is exactly the same procedure to construct k-th order BDF-Adams-Bashforth SAV schemes for (6.8). The only difference in 6.12 is (6.12a) becomes

$$\frac{\alpha_k \bar{u}^{n+1} - A_k(u^n)}{\delta t} + \frac{mS}{\alpha} f(B_k(\phi^n))\Delta^2 \bar{u}^{n+1} = \frac{m}{T(B_k(u^n))} G_s\Big(B_k(u^n), B_k(\phi^n)\Big), \tag{6.28}$$

and all the properties in Thm6.2.1 still hold true.

6.3 Numerical examples

In this section, we present ample numerical examples to verify our numerical schemes. For all the examples in this section, periodic boundary conditions are imposed in all directions and for space discretization, we employ the Fourier-Spectral method [32] and we use Nx, Ny, Nz to denote the number of nodes in x, y, z direction respectively.

6.3.1 Accuracy test

Example1, accuracy test by solving linear equation with constant coefficients. We consider the two-dimension Cahn-Hilliard equation with singular potential and constant mobility, i.e. Case 2 in section 2 with $f(\phi) = 1$. We test the accuracy with the following exact solution and the corresponding external forcing

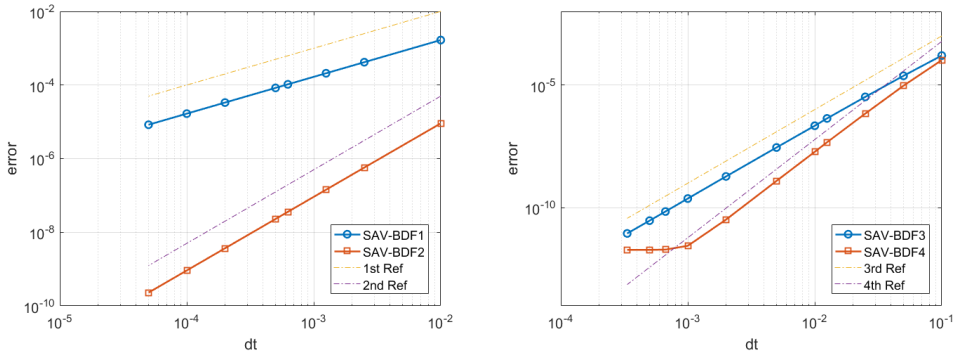
$$\phi(x, y, t) = 0.9 \exp(-\sin^2(\pi x) - \cos^2(\pi y)) \cos(t), \tag{6.29}$$

and the transformation function we used in this example is

$$\phi = T(u) := \frac{1}{1 + \exp(-u)}. \tag{6.30}$$

The parameters are chosen as $\alpha = 1000$, $m = 0.001$ and the computational domain is $(0, 2) \times (0, 2)$. We choose $(Nx, Ny) = (128, 128)$ for the first- and second- order schemes and $(Nx, Ny) = (150, 150)$ for the third- and fourth- order schemes. We plot in Figure 6.1 (a) the errors of the first- and second-order schemes at $t_n = 1$, and in Figure 6.1 (b), the errors

of the third- and fourth-order schemes at $t_n = 1$. Expected convergence rates are observed for all cases.



(a) BDF1 and BDF2 for Cahn-Hilliard (b) BDF3 and BDF4 for Cahn-Hilliard

Figure 6.1. (*Example* 1.) Accuracy test for the 2-D Cahn-Hilliard equation using the new SAV/BDFk schemes ($k = 1, 2, 3, 4$).

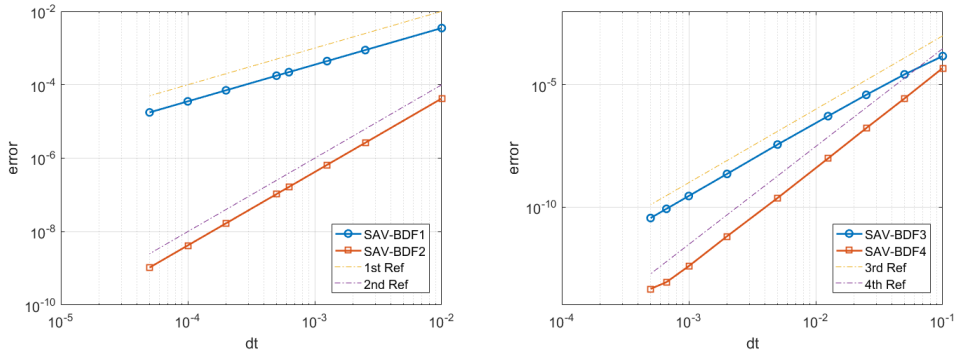
Example2, accuracy test by solving linear equation with variable coefficients. We consider the one-dimension Lubrication-type equation with $f(\phi) = \phi$, i.e. Case 1 in section 2 with $f(\phi) = \phi$. We test the accuracy with the following exact solution and the corresponding external forcing

$$\phi(x, t) = \exp(\sin(\pi x)) \cos(t), \tag{6.31}$$

and the transformation function we used in this example is

$$\phi = T(u) := \exp(u). \tag{6.32}$$

The parameters are chosen as $\alpha = 1$, $m = 0.0001$ and the computational domain is $(0, 2)$. We choose $Nx = 32$ for all the schemes. As the mobility function f is not constant, scheme 6.12 leads to a linear equation with variable coefficients. We use **bicgstab** in **Matlab** to solve those linear equations with variable coefficients by setting $tol = 1\text{e} - 16$ and the maximum iteration $it = 5$. We plot in Figure 6.2 (a) the errors of the first- and second-order schemes at $t_n = 1$, and in Figure 6.2 (b), the errors of the third- and fourth-order schemes at $t_n = 1$. Expected convergence rates are observed for all cases.



(a) BDF1 and BDF2 for Lubrication-type equation

(b) BDF3 and BDF4 for Lubrication-type equation

Figure 6.2. (*Example* 2.) Accuracy test for the 1-D Lubrication-type equation using the new SAV/BDFk schemes ($k = 1, 2, 3, 4$).

6.3.2 Thin film equation

Example3. In this example, we illustrate the advantages of our positivity preserving SAV scheme over the usual semi-implicit scheme without functional transformation and the functional transformation schemes without SAV. We consider the following 1D lubrication-type equation

$$\phi_t + \partial_x(f(\phi)\partial_x^3\phi) = 0, \;\; f(\phi) = \phi^{1/2} \tag{6.33}$$

with positive initial condition

$$\phi_0(x) = 0.8 - cos(\pi x) + 0.25cos(2\pi x). \tag{6.34}$$

It was shown in [114] that the solution of this problem develops singularity in finite time and one common way to compute the solution after the singularity time is to introduce the regularization

$$\partial_t\phi_\eta + \partial_x(f(\phi_\eta)\partial_x^3\phi_\eta) = 0, \;\; f(\phi_\eta) = \frac{\phi_\eta^4 f(\phi_\eta)}{\eta f(\phi_\eta) + \phi_\eta^4}, \tag{6.35}$$

and take η close to zero and the analytical solution of the regularized problem is positive. However, in our numerical test, we can take $\eta = 0$ and keep the numerical solution positive by using the positivity preserving SAV scheme. In the following, we fix the computational domain as $(-1, 1)$, $\eta = 0$ and use **bicgstab** in **Matlab** to solve the linear equation with variable coefficients and then test three different first order schemes for this problem. The parameters in **bicgstab** are fixed as $tol = 1e-7$ and maximum iteration is 10 for all the test.

- (scheme I) The usual semi-implicit scheme without functional transformation, i.e. given ϕ_η^n, solve ϕ_η^{n+1} by

$$\frac{\phi_\eta^{n+1} - \phi_\eta^n}{\delta t} + f(\phi_\eta^n)\partial_x^4\phi_\eta^{n+1} + \partial_x f(\phi_\eta^n)\partial_x^3\phi_\eta^n = 0. \tag{6.36}$$

We plot the maximum value of the imaginary part of the solution in Figure 6.3 (a), we can see the imaginary part come out at finite time, which means the numerical solution is not positivity preserving, even we use smaller time steps and more nodes.

- (scheme II) Functional transformation scheme without SAV, i.e. fix $\xi^{n+1} = 1$ in scheme 6.12. We set $Nx = 128$, $\delta t = 5\mathrm{e}-8$, $A = 1$ and the transformation function we used in this example is

$$\phi = T(u) := \exp(u). \tag{6.37}$$

We plot the energy evolution in Figure 6.3 (b) and the minimum value of the numerical solution in Figure 6.3 (c). We can see although the numerical solution is positivity preserving, the scheme is unstable when time arrive at the singularity and the solution is wrong after that.

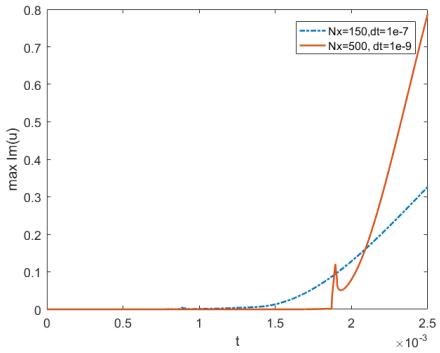
- (scheme III) Positivity preserving SAV scheme, i.e. scheme 6.12 in section 2. All the settings are the same as scheme II. We plot the minimum value of the numerical solution in Figure 6.4 (a) and the sav factor ξ in Figure 6.4 (b). We can see the effect of SAV as time close to the singularity time and it helps the whole scheme passing the singularity time successfully. Finally, we plot the solution at different stages up to a steady state in Figure 6.5 (a) and the energy evolution in Figure 6.5 (b).

6.3.3 Cahn-Hilliard equation

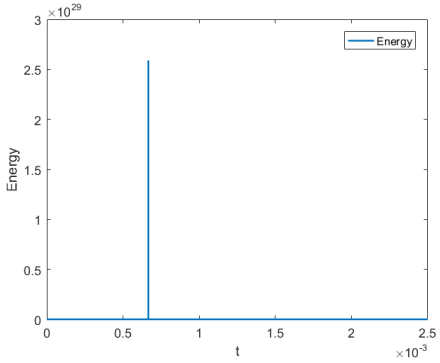
In this subsection, we consider the Cahn-Hilliard equation with singular potential. i.e. Case 2 mentioned in the introduction, the main settings are the same as those in [112]: Set $f(\phi) = \phi(1-\phi)$ in (6.1), $H(\phi) = \phi\log\phi + (1-\phi)\log(1-\phi) + 3\phi(1-\phi)$ in (6.2), the initial condition is given as

$$\phi_0 = \overline{c} + r, \tag{6.38}$$

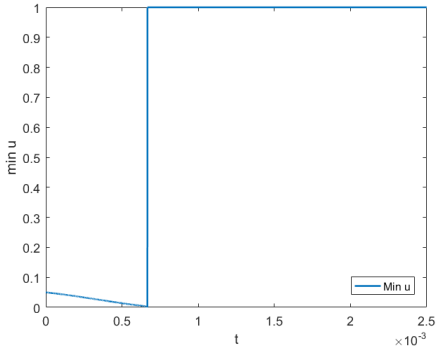
with r is a random variable with uniform distribution in $[-0.05, 0.05]$. All the examples in this subsection are computed by the second-order scheme with stabilization terms by solving



(a) Maximum value of imaginary part from scheme I

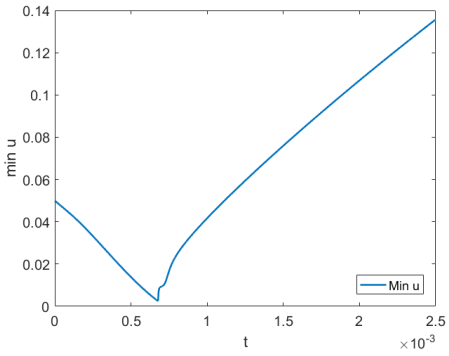


(b) Energy evolution from scheme II

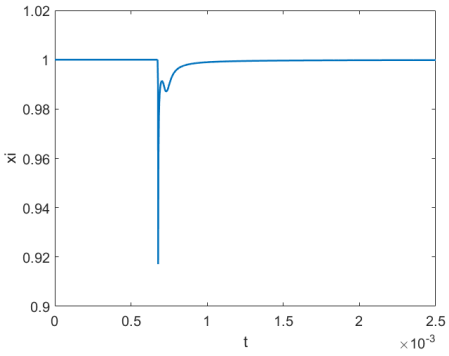


(c) Minimum value from scheme II

Figure 6.3. (*Example* 3.) Failure to compute the solution of equation (6.35) by using scheme I and scheme II

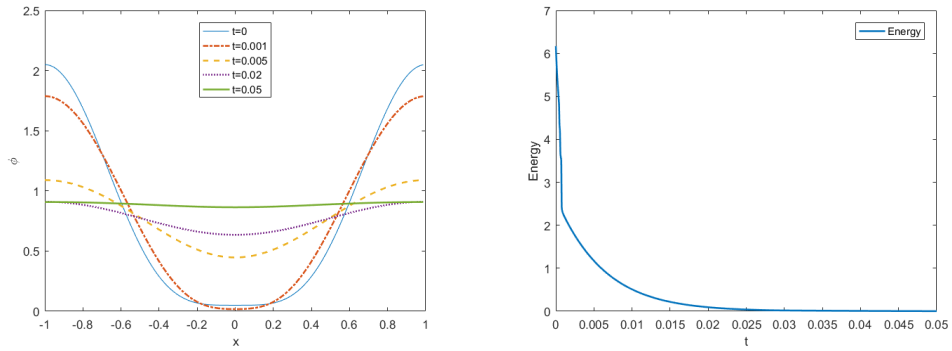


(a) Minimum value from scheme III



(b) Sav factor ξ from scheme III

Figure 6.4. (*Example* 3.) Successful computational of equation (6.35) by using scheme III



(a) Solutions from scheme III (b) Energy evolution from scheme III

Figure 6.5. (*Example* 3.) Successful computational of equation (6.35) by using scheme III

a fourth-order linear equation with constant coefficients and the transformation function we used in this example is

$$\phi = T(u) := \frac{1}{1+\exp(-u)}. \tag{6.39}$$

Example4. In this example, we consider the 2-dimension case with constant mobility $f(\phi) = 1$ and variable mobility $f(\phi) = \phi(1-\phi)$. We fix $\delta t = 5\mathrm{e}-9$ and $Nx = Ny = 96$. For the constant mobility $f(\phi) = 1$ case, we adopt scheme 6.12 with stabilization terms,i.e. replacing 6.12a by 6.17 and take $S = 1$. For the variable mobility $f(\phi) = \phi(1-\phi)$ case, we adopt scheme 6.12 and replacing 6.12a by 6.28. As a result, we only need to solve linear equations with constant coefficients at each time steps for both the constant and variable mobility cases, which is highly efficient. In Fig 6.6, we plot the energy evolution for two cases, which show energy dissipative for all the time. We plot the snapshots for the constant mobility case at different times in Fig 6.7 and the snapshots for the variable mobility case in Fig 6.8, which show two cases have the similar dynamic process while it takes much longer time for the variable mobility case to converge.

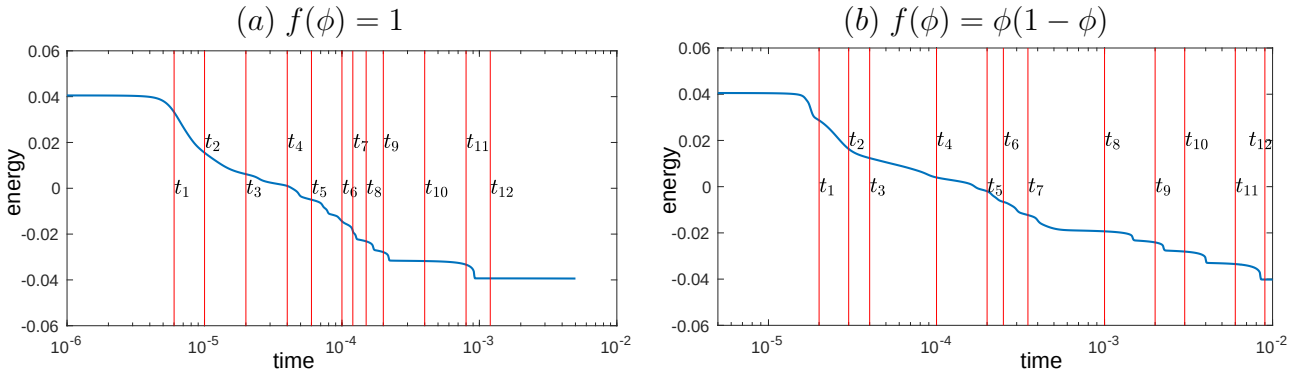


Figure 6.6. Time series of total energy plot at $\alpha = 1000$ and mean initial condition $\bar{c} = 0.63$ with different mobilities as indicated.

Example5. In this example, we consider the 3-dimension case with constant mobility $f(\phi) = 1$ and different mean initial value $\bar{c}$. We fix $\delta t = 1\mathrm{e}-9$ and $Nx = Ny = 96$. For the constant mobility $f(\phi) = 1$ case, we adopt scheme 6.12 with stabilization terms and take $S = 1$. In Fig 6.9, we plot the energy evolution for two cases, which shows energy dissipative for all the time. We plot the snapshots for the case $\bar{c} = 0.63$ at different times in Fig 6.10

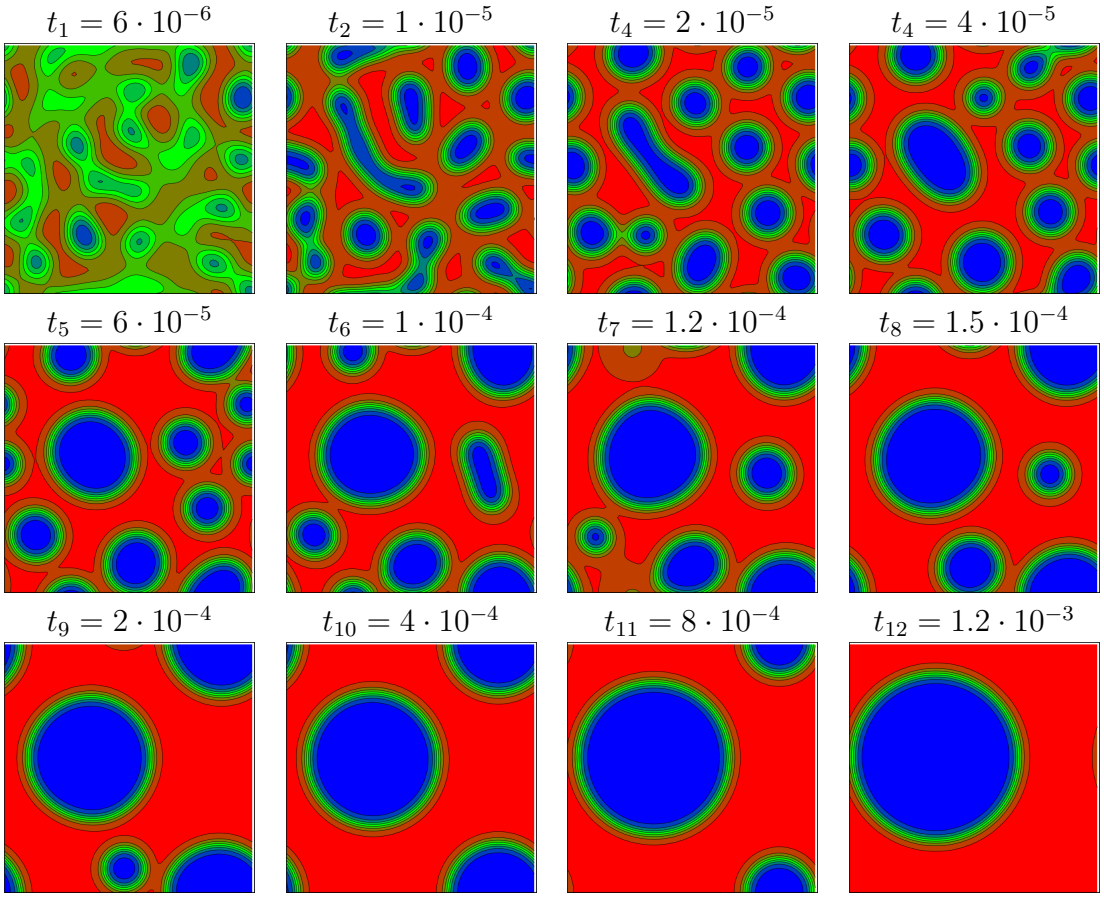


Figure 6.7. Snapshots of ϕ at different time instants (indicated in Figure 6.6(a)) at $\alpha = 1000$, mobility $f(\phi) = 1$ and mean initial condition $\bar{c} = 0.63$. The contour levels are equally spaced in $[0, 1]$ and the colormap is consistent with the contour levels 0(blue), 0.5(green), 1(red).

and the snapshots for the case $\bar{c} = 0.35$ at different times in Fig 6.11, which shows two cases have the different dynamic processes and different steady states.

6.4 Appendix. Coefficients in the bilaplace operator after transformation

The coefficients C_2, C_3 and C_4 in (6.4) are defined as:

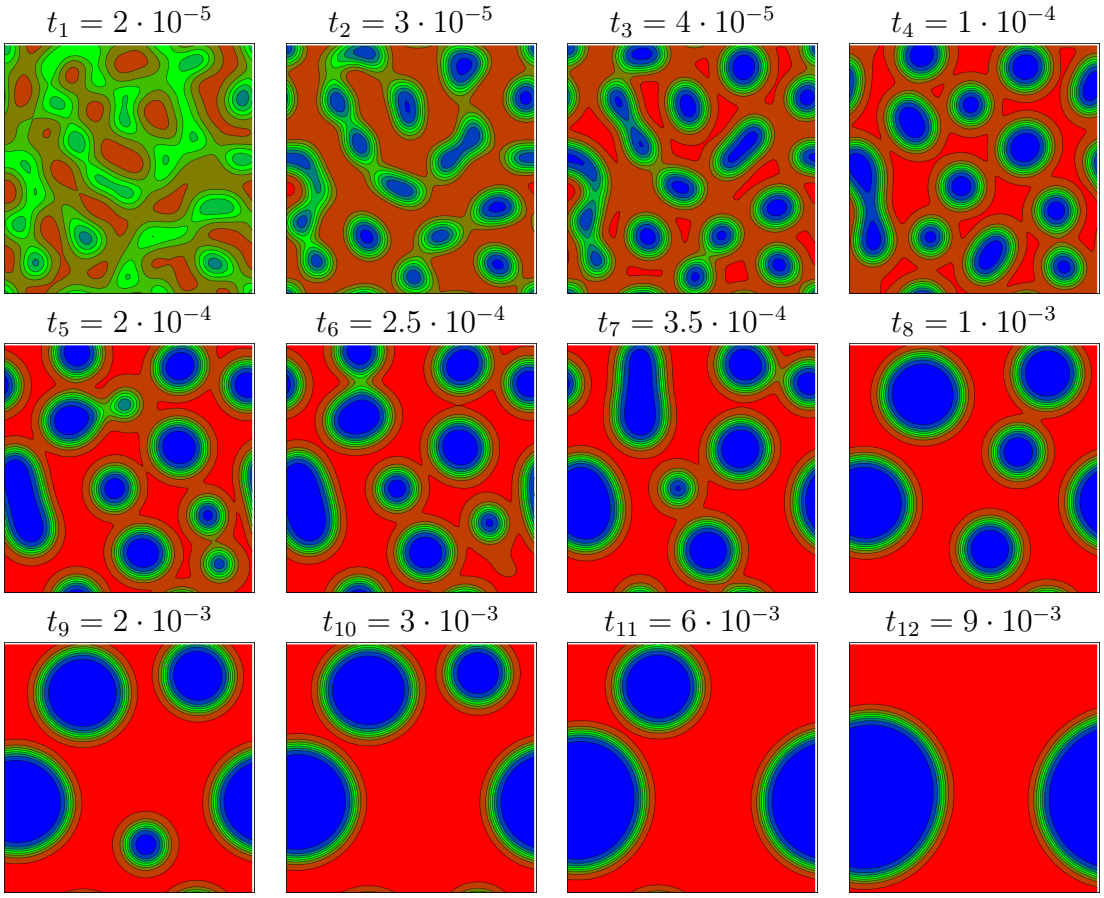


Figure 6.8. Snapshots of ϕ at different time instants (indicated in Figure 6.6(b)) at $\alpha = 1000$, mobility $f(\phi) = \phi(1-\phi)$ and mean initial condition $\bar{c} = 0.63$. The contour levels are equally spaced in $[0, 1]$ and the colormap is consistent with the contour levels 0(blue), 0.5(green), 1(red).

For 2-dimension case:

$$C_2(u) = 4u_x(\Delta u)_x + 4u_y(\Delta u)_y + (\Delta u)^2 + 2u_{xx}^2 + 2u_{yy}^2 + 4u_{xy}^2, \tag{6.40a}$$

$$C_3(u) = 4(u_x)^2 u_{xx} + 4(u_y)^2 u_{yy} + 2|\nabla u|^2 \Delta u + 8u_x u_y u_{xy}, \tag{6.40b}$$

$$C_4(u) = ((u_x)^2 + (u_y)^2)^2. \tag{6.40c}$$

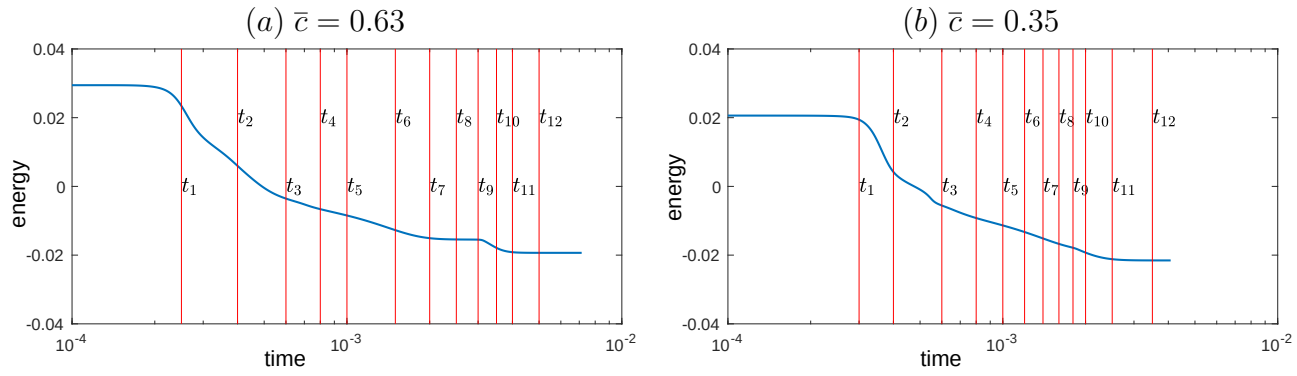


Figure 6.9. Time series of total energy plot at $\alpha = 200$ and mobility $f(\phi) = 1$ with different mean initial conditions as indicated.

For 3-dimension case:

$$C_2(u) = 4\nabla u \cdot \nabla(\Delta u) + (\Delta u)^2 + 2u_{xx}^2 + 2u_{yy}^2 + 2u_{zz}^2 + 4u_{xy}^2 + 4u_{xz}^2 + 4u_{yz}^2, \tag{6.41a}$$

$$C_3(u) = 4(u_x)^2 u_{xx} + 4(u_y)^2 u_{yy} + 4(u_z)^2 u_{zz} + 2|\nabla u|^2 \Delta u + 8u_x u_y u_{xy} + 8u_x u_z u_{xz} + 8u_y u_z u_{yz}, \tag{6.41b}$$

$$C_4(u) = ((u_x)^2 + (u_y)^2 + (u_z)^2)^2. \tag{6.41c}$$

6.5 Conclusion of this chapter

In this section, we continue to apply the positivity/bound preserving SAV scheme for the four-order equation. Depending on specific problems, one can construct linear, decoupled schemes with constant coefficients or variable coefficients and linear, coupled schemes with variable coefficients. Those schemes are unconditionally stable and preserve positivity or bound at the same time. Our numerical examples show that those schemes can be applied to tough problems and both the functional transformation and SAV play an important role.

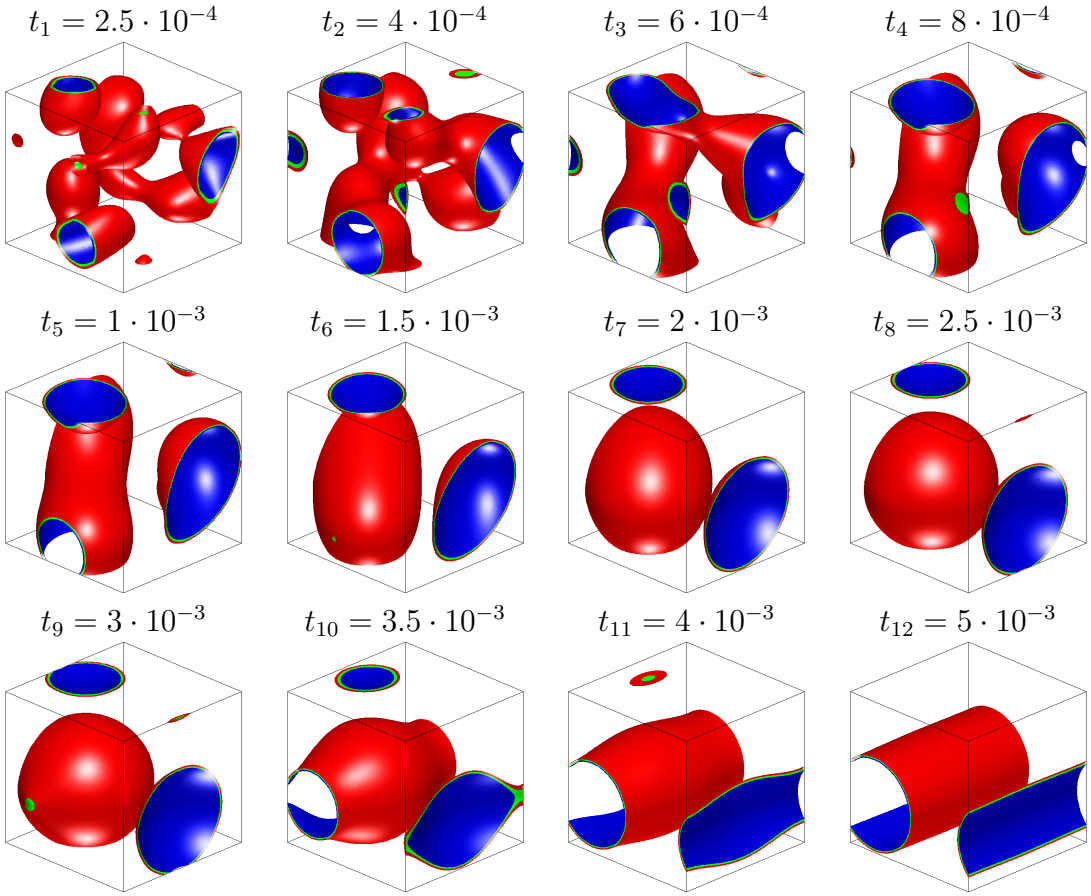


Figure 6.10. Snapshots of ϕ at different time instants (indicated in Figure 6.9(a)) at $\alpha = 200$, mobility $f(\phi) = 1$ and mean initial condition $\bar{c} = 0.63$. The iso-surfaces are plotted at $\phi = -0.1$(blue), $\phi = 0$(green) and $\phi = 0.1$(red).

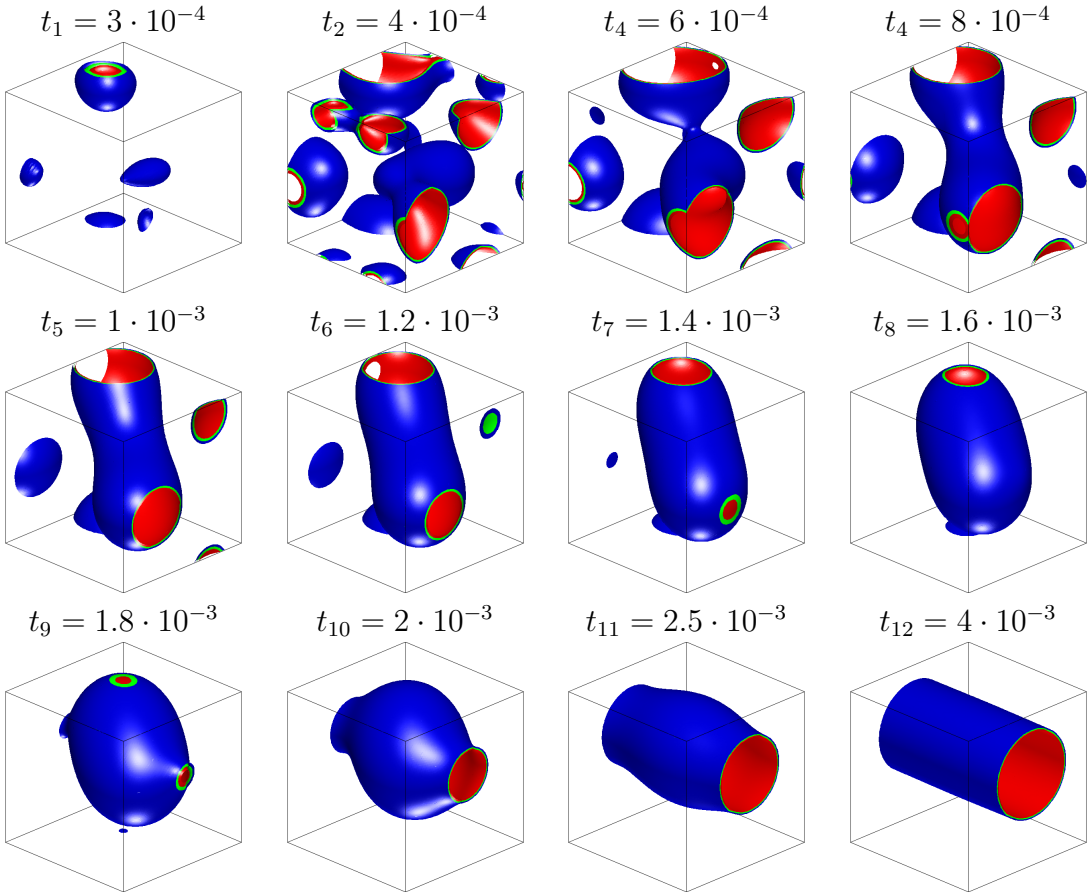


Figure 6.11. Snapshots of ϕ at different time instants (indicated in Figure 6.9(b)) at $\alpha = 200$, mobility $f(\phi) = 1$ and mean initial condition $\overline{c} = 0.35$. The iso-surfaces are plotted at $\phi = -0.1$(blue), $\phi = 0$(green) and $\phi = 0.1$(red).

7. CONCLUDING REMARKS AND FUTURE WORKS

In this thesis, we construct a novel scalar auxiliary variable approach for general dissipative systems. The new SAV approach enjoys lots of remarkable properties:

- it only requires solving one linear system with constant coefficients at each time step, which is half computational cost of the original SAV approach [1];
- it is not limited to the gradient flows systems and it is applicable to general dissipative systems;
- it is extendable to higher-order BDF type schemes with unconditional stability and amenable to higher-order adaptive time stepping;
- it can be used to construct positivity/bound preserving schemes while keeps all the advantages of new SAV approach.

The new SAV approach have the same computational cost as the usual IMEX schemes while we can prove the unconditionally stability for not only the first- and second- order scheme, but also the high order schemes. With the unconditionally stability, we prove the rigorous error analysis for the Allen-Cahn, Cahn-Hilliard type equations and the Navier-Stokes equation with periodic boundary condition. Undoubtedly, the similar ideas can be applied to prove the error analysis for other general dissipative systems. Ample numerical examples show that the new SAV approach have the same good performance as the usual IMEX schemes for simple problems and the new SAV approach has better performance than the usual IMEX problems in some extreme cases, see example 1 in chapter2, example 2 in chapter3 and example 3 in chapter 5.

Following the works presented in this thesis, lots of relevant topics are worth investigating in the future, including but not limited to:

- SAV approach on conserved systems;
- time adaptivity strategies on positivity/bound preserving SAV schemes;
- error analysis for the positivity/bound preserving SAV schemes;

- SAV approach on non-dissipative/non-conserved systems.

REFERENCES

[1] J. Shen, J. Xu, and J. Yang, "A new class of efficient and robust energy stable schemes for gradient flows," *SIAM Review*, vol. 61, no. 3, pp. 474–506, 2019.

[2] E. Celledoni, V. Grimm, R. I. McLachlan, D. McLaren, D. O'Neale, B. Owren, and G. Quispel, "Preserving energy resp. dissipation in numerical pdes using the "average vector field" method," *Journal of Computational Physics*, vol. 231, no. 20, pp. 6770–6789, 2012.

[3] G. Quispel and D. I. McLaren, "A new class of energy-preserving numerical integration methods," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 4, p. 045 206, 2008.

[4] J. W. Barrett and J. F. Blowey, "Finite element approximation of a model for phase separation of a multi-component alloy with non-smooth free energy," *Numerische Mathematik*, vol. 77, no. 1, pp. 1–34, 1997.

[5] J. W. Barrett and J. F. Blowey, "Finite element approximation of a model for phase separation of a multi-component alloy with a concentration-dependent mobility matrix," *IMA journal of numerical analysis*, vol. 18, no. 2, pp. 287–328, 1998.

[6] C. M. Elliott and A. Stuart, "The global dynamics of discrete semilinear parabolic equations," *SIAM journal on numerical analysis*, vol. 30, no. 6, pp. 1622–1663, 1993.

[7] D. J. Eyre, "Unconditionally gradient stable time marching the cahn-hilliard equation," in *Materials Research Society Symposium Proceedings*, Materials Research Society, vol. 529, 1998, pp. 39–46.

[8] J. Shen and X. Yang, "Numerical approximations of allen-cahn and cahn-hilliard equations," *Discrete & Continuous Dynamical Systems-A*, vol. 28, no. 4, p. 1669, 2010.

[9] J. Zhu, L.-Q. Chen, J. Shen, and V. Tikare, "Coarsening kinetics from a variable-mobility cahn-hilliard equation: Application of a semi-implicit fourier spectral method," *Physical Review E*, vol. 60, no. 4, p. 3564, 1999.

[10] F. Guillén-González and G. Tierra, "On linear schemes for a cahn–hilliard diffuse interface model," *Journal of Computational Physics*, vol. 234, pp. 140–171, 2013.

[11] X. Yang, "Linear, first and second-order, unconditionally energy stable numerical schemes for the phase field model of homopolymer blends," *Journal of Computational Physics*, vol. 327, pp. 294–316, 2016.

[12] J. Zhao, Q. Wang, and X. Yang, "Numerical approximations for a phase field dendritic crystal growth model based on the invariant energy quadratization approach," *International Journal for Numerical Methods in Engineering*, vol. 110, no. 3, pp. 279–300, 2017.

[13] F. Huang, J. Shen, and Z. Yang, "A highly efficient and accurate new scalar auxiliary variable approach for gradient flows," *SIAM Journal on Scientific Computing*, vol. 42, no. 4, A2514–A2536, 2020.

[14] O. Nevanlinna and F. Odeh, "Multiplier techniques for linear multistep methods," *Numerical Functional Analysis and Optimization*, vol. 3, no. 4, pp. 377–423, 1981.

[15] R. Jordan, D. Kinderlehrer, and F. Otto, "The variational formulation of the Fokker-Planck equation," *SIAM J. Math. Anal.*, vol. 29, no. 1, pp. 1–17, 1998, ISSN: 0036-1410. DOI: 10.1137/S0036141096303359. [Online]. Available: https://doi.org/10.1137/S0036141096303359.

[16] F. Santambrogio, "{euclidean, metric, and wasserstein} gradient flows: An overview," *Bulletin of Mathematical Sciences*, vol. 7, no. 1, pp. 87–154, 2017.

[17] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor equations*. Springer-Verlag, Vienna, 1990, pp. x+248, ISBN: 3-211-82157-0. DOI: 10.1007/978-3-7091-6961-2. [Online]. Available: https://doi.org/10.1007/978-3-7091-6961-2.

[18] E. F. Keller and L. A. Segel, "Initiation of slime mold aggregation viewed as an instability," *Journal of Theoretical Biology*, vol. 26, no. 3, pp. 399–415, 1970, ISSN: 0022-5193. DOI: https://doi.org/10.1016/0022-5193(70)90092-5. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022519370900925.

[19] T. Hillen and K. J. Painter, "A user's guide to PDE models for chemotaxis," *Journal of mathematical biology*, vol. 58, no. 1-2, p. 183, 2009.

[20] J. Shen, J. Xu, and J. Yang, "The scalar auxiliary variable (SAV) approach for gradient flows," *J. Comput. Phys.*, vol. 353, pp. 407–416, 2018, ISSN: 0021-9991. DOI: 10.1016/j.jcp.2017.10.021. [Online]. Available: https://doi.org/10.1016/j.jcp.2017.10.021.

[21] Z. Yang, L. Lin, and S. Dong, "A family of second-order energy-stable schemes for cahn–hilliard type equations," *Journal of Computational Physics*, vol. 383, pp. 24–54, 2019.

[22] Z. Yang and S. Dong, "A roadmap for discretely energy-stable schemes for dissipative systems based on a generalized auxiliary variable with guaranteed positivity," *Journal of Computational Physics*, vol. 404, p. 109 121, 2020.

[23] W. Chen, X. Wang, Y. Yan, and Z. Zhang, "A second order bdf numerical scheme with variable steps for the cahn–hilliard equation," *SIAM Journal on Numerical Analysis*, vol. 57, no. 1, pp. 495–525, 2019.

[24] Y. He, Y. Liu, and T. Tang, "On large time-stepping methods for the cahn–hilliard equation," *Applied Numerical Mathematics*, vol. 57, no. 5-7, pp. 616–628, 2007.

[25] F. Luo, T. Tang, and H. Xie, "Parameter-free time adaptivity based on energy evolution for the cahn-hilliard equation," *Communications in Computational Physics*, vol. 19, no. 5, pp. 1542–1563, 2016.

[26] Z. Zhang and Z. Qiao, "An adaptive time-stepping strategy for the cahn-hilliard equation," *Communications in Computational Physics*, vol. 11, no. 4, pp. 1261–1278, 2012.

[27] Q. Cheng, J. Shen, and X. Yang, "Highly efficient and accurate numerical schemes for the epitaxial thin film growth models by using the sav approach," *Journal of Scientific Computing*, vol. 78, no. 3, pp. 1467–1487, 2019.

[28] X. Li, J. Shen, and H. Rui, "Stability and error analysis of a second-order sav scheme with block-centered finite differences for gradient flows," *Math. Comp*, vol. 88, p. 2047, 2019.

[29] Q. Zhuang and J. Shen, "Efficient sav approach for imaginary time gradient flows with applications to one-and multi-component bose-einstein condensates," *Journal of Computational Physics*, vol. 396, pp. 72–88, 2019.

[30] S. M. Allen and J. W. Cahn, "A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening," *Acta Metall. Mater.*, vol. 27, pp. 1085–1095, 1979.

[31] J. W. Cahn and J. E. Hilliard, "Free energy of a nonuniform system. i. interfacial free energy," *The Journal of chemical physics*, vol. 28, no. 2, pp. 258–267, 1958.

[32] J. Shen, T. Tang, and L.-L. Wang, *Spectral methods: algorithms, analysis and applications*. Springer Science & Business Media, 2011, vol. 41.

[33] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell, "Observation of bose-einstein condensation in a dilute atomic vapor," *science*, vol. 269, no. 5221, pp. 198–201, 1995.

[34] W. Bao and Q. Du, "Computing the ground state solution of bose–einstein condensates by a normalized gradient flow," *SIAM Journal on Scientific Computing*, vol. 25, no. 5, pp. 1674–1697, 2004.

[35] Q. Cheng and J. Shen, "Multiple scalar auxiliary variable (msav) approach and its application to the phase-field vesicle membrane model," *SIAM Journal on Scientific Computing*, vol. 40, no. 6, A3982–A4006, 2018.

[36] F. Huang and J. Shen, "Implicit-explicit bdf$k$ sav schemes for general dissipative systems and their error analysisl," submitted, 2020.

[37] D. Kessler, R. H. Nochetto, and A. Schmidt, "A posteriori error control for the allen–cahn problem: Circumventing gronwall's inequality," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 38, no. 1, pp. 129–142, 2004.

[38] N. Condette, C. Melcher, and E. Süli, "Spectral approximation of pattern-forming nonlinear evolution equations with double-well potentials of quadratic growth," *Mathematics of computation*, vol. 80, no. 273, pp. 205–223, 2011.

[39] J. Shen and J. Xu, "Convergence and error analysis for the scalar auxiliary variable (sav) schemes to gradient flows," *SIAM Journal on Numerical Analysis*, vol. 56, no. 5, pp. 2895–2912, 2018.

[40] Y. Qian, Z. Yang, F. Wang, and S. Dong, "Gpav-based unconditionally energy-stable schemes for the cahn–hilliard equation: Stability and error analysis," *Computer Methods in Applied Mechanics and Engineering*, vol. 372, p. 113 444, 2020.

[41] X. Li, J. Shen, and H. Rui, "Energy stability and convergence of SAV block-centered finite difference method for gradient flows," *Math. Comp.*, vol. 88, no. 319, pp. 2047–2068, 2019, ISSN: 0025-5718. DOI: 10.1090/mcom/3428. [Online]. Available: https://doi.org/10.1090/mcom/3428.

[42] H. Chen, J. Mao, and J. Shen, "Optimal error estimates for the scalar auxiliary variable finite-element schemes for gradient flows," *Numer. Math.*, vol. 145, no. 1, pp. 167–196, 2020, ISSN: 0029-599X. DOI: 10.1007/s00211-020-01112-4. [Online]. Available: https://doi.org/10.1007/s00211-020-01112-4.

[43] X. Li and J. Shen, "Stability and error estimates of the SAV Fourier-spectral method for the phase field crystal equation," *Adv. Comput. Math.*, vol. 46, no. 3, Paper No. 48, 20, 2020, ISSN: 1019-7168. DOI: 10.1007/s10444-020-09789-9. [Online]. Available: https://doi.org/10.1007/s10444-020-09789-9.

[44] X. Li and J. Shen, "Error Analysis of the SAV-MAC Scheme for the Navier–Stokes Equations," *SIAM J. Numer. Anal.*, vol. 58, no. 5, pp. 2465–2491, 2020, ISSN: 0036-1429. DOI: 10.1137/19M1288267. [Online]. Available: https://doi.org/10.1137/19M1288267.

[45] G. Akrivis, B. Li, and D. Li, "Energy-decaying extrapolated RK-SAV methods for the Allen-Cahn and Cahn-Hilliard equations," *SIAM J. Sci. Comput.*, vol. 41, no. 6, A3703–A3727, 2019, ISSN: 1064-8275. DOI: 10.1137/19M1264412. [Online]. Available: https://doi.org/10.1137/19M1264412.

[46] D. Li and W. Sun, "Linearly implicit and high-order energy-conserving schemes for nonlinear wave equations," *J. Sci. Comput.*, vol. 83, no. 3, Paper No. 65, 17, 2020, ISSN: 0885-7474. DOI: 10.1007/s10915-020-01245-6. [Online]. Available: https://doi.org/10.1007/s10915-020-01245-6.

[47] G. Akrivis, "Stability of implicit-explicit backward difference formulas for nonlinear parabolic equations," *SIAM Journal on Numerical Analysis*, vol. 53, no. 1, pp. 464–484, 2015.

[48] R. Temam, *Infinite-dimensional dynamical systems in mechanics and physics*. Springer Science & Business Media, 2012, vol. 68.

[49] F. Huang and J. Shen, "Stability and error analysis of a class of high-order imex schemes for navier-stokes equations with periodic boundary conditions," submitted, 2021.

[50] V. Girault and P.-A. Raviart, "Finite element approximation of the navier-stokes equations," *Lecture Notes in Mathematics, Berlin Springer Verlag*, vol. 749, 1979.

[51] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis*. North-Holland, Amsterdam, 1984.

[52] M. O. Deville, P. F. Fischer, P. F. Fischer, E. Mund, *et al.*, *High-order methods for incompressible fluid flow*, 9. Cambridge university press, 2002.

[53] R. Glowinski, "Finite element methods for incompressible viscous flow," *Handbook of numerical analysis*, vol. 9, pp. 3–1176, 2003.

[54] M. D. Gunzburger, *Finite element methods for viscous incompressible flows: a guide to theory, practice, and algorithms*. Elsevier, 2012.

[55] R. Peyret, *Spectral methods for incompressible viscous flow*. Springer Science & Business Media, 2013, vol. 148.

[56] G. A. Baker, V. A. Dougalis, and O. A. Karakashian, "On a higher order accurate fully discrete galerkin approximation to the navier-stokes equations," *Mathematics of Computation*, vol. 39, no. 160, pp. 339–375, 1982.

[57] J. G. Heywood and R. Rannacher, "Finite-element approximation of the nonstationary navier–stokes problem. part iv: Error analysis for second-order time discretization," *SIAM Journal on Numerical Analysis*, vol. 27, no. 2, pp. 353–384, 1990.

[58] W. E and J.-G. Liu, "Projection method i: Convergence and numerical boundary layers," *SIAM journal on numerical analysis*, pp. 1017–1057, 1995.

[59] J.-L. Guermond, P. Minev, and J. Shen, "An overview of projection methods for incompressible flows," *Computer methods in applied mechanics and engineering*, vol. 195, no. 44-47, pp. 6011–6045, 2006.

[60] Y. He and W. Sun, "Stability and convergence of the Crank-Nicolson/Adams-Bashforth scheme for the time-dependent Navier-Stokes equations," *SIAM J. Numer. Anal.*, vol. 45, no. 2, pp. 837–869, 2007, ISSN: 0036-1429. DOI: 10.1137/050639910. [Online]. Available: https://doi.org/10.1137/050639910.

[61] J. de Frutos, B. Garcıa-Archilla, and J. Novo, "Postprocessing finite-element methods for the Navier-Stokes equations: The fully discrete case," *SIAM J. Numer. Anal.*, vol. 47, no. 1, pp. 596–621, 2008/09, ISSN: 0036-1429. DOI: 10.1137/070707580. [Online]. Available: https://doi.org/10.1137/070707580.

[62] S. A. Orszag and G. Patterson Jr, "Numerical simulation of three-dimensional homogeneous isotropic turbulence," *Physical Review Letters*, vol. 28, no. 2, p. 76, 1972.

[63] Z.-S. She, E. Jackson, and S. A. Orszag, "Structure and dynamics of homogeneous turbulence: Models and simulations," *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, vol. 434, no. 1890, pp. 101–124, 1991.

[64] P. Moin and K. Mahesh, "Direct numerical simulation: A tool in turbulence research," *Annual review of fluid mechanics*, vol. 30, no. 1, pp. 539–578, 1998.

[65] O. H. Hald, "Convergence of fourier methods for navier-stokes equations," *Journal of Computational Physics*, vol. 40, no. 2, pp. 305–317, 1981.

[66] W. E, "Convergence of Fourier methods for the Navier-Stokes equations," *SIAM J. Numer. Anal.*, vol. 30, no. 3, pp. 650–674, 1993, ISSN: 0036-1429. DOI: 10.1137/0730032. [Online]. Available: https://doi.org/10.1137/0730032.

[67] X. Wang, "An efficient second order in time scheme for approximating long time statistical properties of the two dimensional navier–stokes equations," *Numerische Mathematik*, vol. 121, no. 4, pp. 753–779, 2012.

[68] S. Gottlieb, F. Tone, C. Wang, X. Wang, and D. Wirosoetisno, "Long time stability of a classical efficient scheme for two-dimensional navier–stokes equations," *SIAM Journal on Numerical Analysis*, vol. 50, no. 1, pp. 126–150, 2012.

[69] F. Tone, X. Wang, and D. Wirosoetisno, "Long-time dynamics of 2d double-diffusive convection: Analysis and/of numerics," *Numerische Mathematik*, vol. 130, no. 3, pp. 541–566, 2015.

[70] K. Cheng and C. Wang, "Long time stability of high order multistep numerical schemes for two-dimensional incompressible Navier-Stokes equations," *SIAM J. Numer. Anal.*, vol. 54, no. 5, pp. 3123–3144, 2016, ISSN: 0036-1429. DOI: 10.1137/16M1061588. [Online]. Available: https://doi.org/10.1137/16M1061588.

[71] T. Heister, M. A. Olshanskii, and L. G. Rebholz, "Unconditional long-time stability of a velocity–vorticity method for the 2d navier–stokes equations," *Numerische Mathematik*, vol. 135, no. 1, pp. 143–167, 2017.

[72] L. Lin, Z. Yang, and S. Dong, "Numerical approximation of incompressible navier-stokes equations based on an auxiliary energy variable," *Journal of Computational Physics*, vol. 388, pp. 1–22, 2019.

[73] X. Li and J. Shen, "Error analysis of the sav-mac scheme for the navier–stokes equations," *SIAM Journal on Numerical Analysis*, vol. 58, no. 5, pp. 2465–2491, 2020.

[74] R. A. Adams and J. J. Fournier, *Sobolev spaces*. Elsevier, 2003.

[75] R. Temam, *Navier-Stokes Equations and Nonlinear Functional Analysis*. SIAM, Philadelphia, 1983.

[76] H.-O. Kreiss and J. Oliger, "Stability of the fourier method," *SIAM Journal on Numerical Analysis*, vol. 16, no. 3, pp. 421–433, 1979.

[77] J. B. Bell, P. Colella, and H. M. Glaz, "A second-order projection method for the incompressible navier-stokes equations," *Journal of Computational Physics*, vol. 85, no. 2, pp. 257–283, 1989.

[78] D. L. Brown, "Performance of under-resolved two-dimensional incompressible flow simulations," *Journal of Computational Physics*, vol. 122, no. 1, pp. 165–183, 1995.

[79] Y. Di, R. Li, T. Tang, and P. Zhang, "Moving mesh finite element methods for the incompressible navier–stokes equations," *SIAM Journal on Scientific Computing*, vol. 26, no. 3, pp. 1036–1056, 2005.

[80] J. Shen, "Long time stability and convergence for fully discrete nonlinear galerkin methods," *Applicable Analysis*, vol. 38, no. 4, pp. 201–229, 1990.

[81] J.-G. Liu, R. Pego, *et al.*, "Stable discretization of magnetohydrodynamics in bounded domains," *Communications in Mathematical Sciences*, vol. 8, no. 1, pp. 235–251, 2010.

[82] C. Foias and R. Temam, "Gevrey class regularity for the solutions of the navier-stokes equations," *Journal of Functional Analysis*, vol. 87, no. 2, pp. 359–369, 1989.

[83] F. Huang and J. Shen, "Bound/positivity preserving and energy stable sav schemes for dissipative systems: Applications to keller-segel and poisson-nernst-planck equations," *To appear in SIAM Journal on Scientific Computing*, 2021.

[84] A. Prohl and M. Schmuck, "Convergent discretizations for the Nernst–Planck–Poisson system," *Numerische Mathematik*, vol. 111, no. 4, pp. 591–630, 2009.

[85] A. Flavell, M. Machen, B. Eisenberg, J. Kabre, C. Liu, and X. Li, "A conservative finite difference scheme for Poisson-Nernst-Planck equations," *Journal of Computational Electronics*, vol. 13, no. 1, pp. 235–249, 2014.

[86] H. Liu and Z. Wang, "A free energy satisfying discontinuous Galerkin method for one-dimensional Poisson-Nernst-Planck systems," *J. Comput. Phys.*, vol. 328, pp. 413–437, 2017, ISSN: 0021-9991. DOI: 10.1016/j.jcp.2016.10.008. [Online]. Available: https://doi.org/10.1016/j.jcp.2016.10.008.

[87] J. Hu and X. Huang, "A fully discrete positivity-preserving and energy-dissipative finite difference scheme for Poisson-Nernst-Planck equations," *Numer. Math.*, vol. 145, no. 1, pp. 77–115, 2020, ISSN: 0029-599X. DOI: 10.1007/s00211-020-01109-z. [Online]. Available: https://doi.org/10.1007/s00211-020-01109-z.

[88] C. L. Gardner, W. Nonner, and R. S. Eisenberg, "Electrodiffusion model simulation of ionic channels: 1D simulations," *Journal of Computational Electronics*, vol. 3, no. 1, pp. 25–31, 2004.

[89] C. L. Lopreore, T. M. Bartol, J. S. Coggan, D. X. Keller, G. E. Sosinsky, M. H. Ellisman, and T. J. Sejnowski, "Computational modeling of three-dimensional electrodiffusion in biological systems: Application to the node of ranvier," *Biophysical journal*, vol. 95, no. 6, pp. 2624–2635, 2008.

[90] T.-L. Horng, T.-C. Lin, C. Liu, and B. Eisenberg, "PNP equations with steric effects: A model of ion flow through channels," *The Journal of Physical Chemistry B*, vol. 116, no. 37, pp. 11 422–11 441, 2012.

[91] F. Filbet, "A finite volume scheme for the Patlak–Keller–Segel chemotaxis model," *Numerische Mathematik*, vol. 104, no. 4, pp. 457–488, 2006.

[92] A. Chertock and A. Kurganov, "A second-order positivity preserving central-upwind scheme for chemotaxis and haptotaxis models," *Numerische Mathematik*, vol. 111, no. 2, p. 169, 2008.

[93] Y. Epshteyn and A. Kurganov, "New interior penalty discontinuous Galerkin methods for the Keller–Segel chemotaxis model," *SIAM Journal on Numerical Analysis*, vol. 47, no. 1, pp. 386–408, 2009.

[94] Y. Epshteyn, "Upwind-difference potentials method for Patlak-Keller-Segel chemotaxis model," *Journal of Scientific Computing*, vol. 53, no. 3, pp. 689–713, 2012.

[95] G. Zhou and N. Saito, "Finite volume methods for a Keller–Segel system: Discrete energy, error estimates and numerical blow-up analysis," *Numerische Mathematik*, vol. 135, no. 1, pp. 265–311, 2017.

[96] J.-G. Liu, L. Wang, and Z. Zhou, "Positivity-preserving and asymptotic preserving method for 2D Keller-Segal equations," *Math. Comp.*, vol. 87, no. 311, pp. 1165–1189, 2018, ISSN: 0025-5718. DOI: 10.1090/mcom/3250. [Online]. Available: https://doi.org/10.1090/mcom/3250.

[97] L. N. De Almeida, F. Bubba, B. Perthame, and C. Pouchol, "Energy and implicit discretization of the Fokker-Planck and Keller-Segel type equations," *arXiv preprint arXiv:1803.10629*, 2018.

[98] J. Shen and J. Xu, "Unconditionally bound preserving and energy dissipative schemes for a class of Keller–Segel Equations," *SIAM Journal on Numerical Analysis*, vol. 58, no. 3, pp. 1674–1695, 2020.

[99] J. Shen and J. Xu, "Unconditionally positivity preserving and energy dissipative schemes for Poisson–Nernst–Planck equations," *arXiv preprint arXiv:2007.06132*, 2020.

[100] M. Z. Bazant, K. Thornton, and A. Ajdari, "Diffuse-charge dynamics in electrochemical systems," *Physical review E*, vol. 70, no. 2, p. 021 506, 2004.

[101] H. Gajewski and K. Gröger, "On the basic equations for carrier transport in semiconductors," *Journal of mathematical analysis and applications*, vol. 113, no. 1, pp. 12–35, 1986.

[102] B. Eisenberg, "Ionic channels in biological membranes-electrostatic analysis of a natural nanotube," *Contemporary Physics*, vol. 39, no. 6, pp. 447–466, 1998.

[103] P. Biler and T. Nadzieja, "Existence and nonexistence of solutions for a model of gravitational interaction of particles, i," in *Colloquium Mathematicae*, vol. 66, 1993, pp. 319–334.

[104] A. Blanchet, J. Dolbeault, and B. Perthame, "Two-dimensional Keller-Segel model: Optimal critical mass and qualitative properties of the solutions.," *Electronic Journal of Differential Equations (EJDE)[electronic only]*, vol. 2006, Paper–No, 2006.

[105] V. Calvez, L. Corrias, *et al.*, "The parabolic-parabolic Keller-Segel model in R2," *Communications in Mathematical Sciences*, vol. 6, no. 2, pp. 417–447, 2008.

[106] J. J. Velázquez, "Point dynamics in a singular limit of the Keller–Segel model 1: Motion of the concentration regions," *SIAM Journal on Applied Mathematics*, vol. 64, no. 4, pp. 1198–1223, 2004.

[107] J. J. Velázquez, "Point dynamics in a singular limit of the Keller–Segel model 2: Formation of the concentration regions," *SIAM Journal on Applied Mathematics*, vol. 64, no. 4, pp. 1224–1248, 2004.

[108] Y. Dolak and C. Schmeiser, "The Keller–Segel model with logistic sensitivity function and small diffusivity," *SIAM Journal on Applied Mathematics*, vol. 66, no. 1, pp. 286–308, 2005.

[109] T. Hillen and K. Painter, "Global existence for a parabolic chemotaxis model with prevention of overcrowding," *Advances in Applied Mathematics*, vol. 26, no. 4, pp. 280–301, 2001.

[110] C. M. Elliott and H. Garcke, "On the cahn–hilliard equation with degenerate mobility," *Siam journal on mathematical analysis*, vol. 27, no. 2, pp. 404–423, 1996.

[111] P. Neogi and C. Miller, "Spreading kinetics of a drop on a smooth solid surface," *Journal of Colloid and Interface Science*, vol. 86, no. 2, pp. 525–538, 1982.

[112] H. Gómez, V. M. Calo, Y. Bazilevs, and T. J. Hughes, "Isogeometric analysis of the cahn–hilliard phase-field model," *Computer methods in applied mechanics and engineering*, vol. 197, no. 49-50, pp. 4333–4352, 2008.

[113] A. L. Bertozzi, N. Ju, and H.-W. Lu, "A biharmonic-modified forward time stepping method for fourth order nonlinear diffusion equations," *Discrete Contin. Dyn. Syst*, vol. 29, no. 4, pp. 1367–1391, 2011.

[114] L. Zhornitskaya and A. L. Bertozzi, "Positivity-preserving numerical schemes for lubrication-type equations," *SIAM Journal on Numerical Analysis*, vol. 37, no. 2, pp. 523–555, 1999.

VITA

Fukeng Huang (Chinese: 黄富铿, born in October, 1993) is a Ph.D candidate in the Department of Mathematics, Purdue University. He was born and grew up in Foshan, China. He obtained his bachelor degree in Mathematics and Applied Mathematics from Sun Yat-sen University in June, 2016 and has become a graduate student at Purdue University since August, 2016.