Structure-preserving Numerical Schemes for Phase Field Models

by
Zhaohui Fu

B.Sc, Zhejiang University, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Mathematics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)
August 2022
©Zhaohui Fu 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Structure-preserving Numerical Schemes for Phase Field Models

Submitted by **Zhaohui Fu** in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Mathematics

Examining Committee:

Brian Wetton, Professor, Mathematics, UBC

Supervisor

Michael J. Ward , Professor, Mathematics, UBC

Committee Member

Colin Macdonald, Professor, Mathematics, UBC

University Examiner

Uri Ascher, Professor, Computer Science, UBC

University Examiner

Zhonghua Qiao, Professor, Mathematics, PolyU

External Examiner

Additional Supervisory Committee Members:

Juncheng Wei, Professor, Mathematics, UBC

Committee Member

Abstract

In this thesis we study how to design accurate, efficient and structure-preserving numerical schemes for phase field models including the Allen–Cahn equation, the Cahn–Hilliard equation and the molecular beam epitaxy equation. These numerical schemes include the explicit Runge–Kutta methods, exponential time differencing (ETD) Runge–Kutta methods and implicit-explicit (IMEX) Runge–Kutta methods. Note that the phase field models under consideration are gradient flows whose energy functionals decrease with time. For the Allen–Cahn equation, it is well known that the solution satisfies the maximum principle; for the Cahn–Hilliard equation, although its solution does not satisfy the maximum principle, the solution is also bounded in time. When designing numerical schemes, we wish to preserve certain stabilities satisfied by the physical solutions. We first make use of strong stability preserving (SSP) Runge-Kutta methods and apply some detailed analysis to derive a class of high-order (up to 4) explicit Runge-Kutta methods which not only decrease the discrete energy but also preserve the maximum principle for the Allen–Cahn equation. Secondly, we prove that the second-order exponential time differencing Runge-Kutta methods decrease the discrete energy for the phase field equations under investigation. Moreover,

it can be shown that the ETDRK methods can also preserve the the maximum bound property for the Allen–Cahn equation. What is more important is that both properties are preserved unconditionally, in the sense that the stability conditions do not depend on the size of time steps. Although the proof is only valid for second-order schemes and still open for higher-order methods, its numerical efficiency has been well observed in computations. The third approach is the implicit-explicit (IMEX) Runge–Kutta (RK) schemes, i.e. taking the linear part in the equation implicitly and the nonlinear part explicitly. A class of high-order IMEX-RK schemes are studied carefully. We demonstrate that some of the IMEX-RK schemes can preserve the energy decreasing property unconditionally for all the phase-field models under investigation.

Lay Summary

Computational simulations can be used to inexpensively understand complex systems such as the stock market prediction and airplane design. The underlying phenomena must be described accurately by mathematical equations in a process called modelling. Then, the mathematical equations are approximated using computational algorithms. The algorithms should be accurate and efficient and also preserve important properties of the solutions of the equations (which are also significant properties inherited from the application). This thesis is dealing with algorithms part with particular attention to preserve properties of the solution. Our special target of complex systems is so-called phase-field models which are useful in modelling interfacial phenomena such as micro-structure evolution and phase transition. As the different equations from the phase-field models are strongly nonlinear, great efforts have to be made to design and analyze the relevant computational methods. This thesis proves rigorously that several proposed methods preserve the important properties of three typical equations in phase-field models.

Preface

This thesis is based on original research projects by the author and the relevant research articles that are published or submitted for publication to research journals. Contributions of collaborators in each research article will be clarified.

Chapter 1 of the thesis is devoted to the purpose of this research together with some background introduction. Chapter 2 provides some notations and useful preliminaries in the rest part of the thesis.

Chapter 3 is based on the paper "Energy plus maximum bound preserving Runge-Kutta methods for the Allen–Cahn equation", which has been published in Journal of Scientific Computing. The development of a new systematic method and the relevant analysis were carried out jointly with the author, J. Yang and T. Tang. The author wrote the majority of the paper and contributes 70% of the research framework. The author played a major role in analyzing the class of Runge–Kutta methods which satisfy the stability conditions obtained in the paper.

Chapter 4 is based on the paper "Energy-decreasing Exponential Time Differencing Runge-Kutta methods for phase-field models", which has been published in Journal of

Computational Physics. The framework and methodology of this project were developed by the author and J. Yang. The author contributes 70% to the work by providing the main ideas, computational details and the majority of the writing up.

Chapter 5 is based on the paper "Unconditionally energy-decreasing high-order Implicit-Explicit Runge-Kutta methods for phase-field models with Lipschitz nonlinearity", which is a preprint and has been put on arxiv. The development of a new systematic analysis tool and the manuscript composition were carried out jointly with the author, J. Yang and T. Tang. The author wrote the majority of the paper and contributes 70% to the research framework and methodology including the rigorous analysis of energy dissipation. The author also played a major role in understanding a class of high-order IMEX Runge-Kutta methods which satisfy the stability conditions obtained in the paper.

Table of Contents

List of Figures

Acknowledgements

Firstly, I would like to thank my supervisors Dr. Brian Wetton, Dr. Jiang Yang and Dr. Tao Tang who helped me to build up qualified mathematics background and skills to find interesting and suitable research problems. Moreover, they spent a lot of their valuable time in providing me with enlightening and significant ideas, which are crucial to the thesis.

Secondly, I would like to express my thanks to my friends, colleagues and collaborators for their friendship which make my research enjoyable and for their discussions which lead to some useful ideas. I give a list of names below and I do apologize if any is missing: Xinyu Chen, Chaoyu Quan, Heyu Wang, Xu Wu and Quanhui Zhu.

Dedication

I dedicate this thesis:

In memory of my grandfather, Mr. Guohao Zhao for his love and support.

To my parents: Mr. Naiyun Fu and Mrs. Jianmei Zhao for always being with me.

To my headteacher and math teachers from my middle schools, Mr. Xiaodong Chen, Mr. Baowei Shen and Mr. Ting Zhao for leading me to learn the beauty of mathematics.

Chapter 1

Introduction

Partial differential equations (PDEs) are widely used to describe mathematical models including physical, biological and financial phenomena. To understand PDEs better helps us to understand the world better.

In the study of PDEs, the main goal is to derive the solution of these equations and properties that they satisfy. The PDEs which describe corresponding physical or mathematical models are the basis of simulations that allow inexpensive virtual experiments and helps understand the behavior of the system. We are often interested in the solutions of PDEs in some specific domain with certain initial conditions and boundary conditions. For example, in the study of lift characteristics of an airplane, the domain is the air around it, the initial condition is the data of the airplane and air (like wind velocity) at the starting time and the boundary condition describes how the air interacts with the body of the airplane. In general, more information about the solutions help the model to be better understood.

For most PDEs used to describe complicated systems, it is difficult to derive explicit solutions, thus there are usually two different approaches to understand them. The first is

to focus on the pure mathematical properties of the solutions like existence, uniqueness, regularity and so on. These results are often investigated with analytic methods and theorems such as fixed point theorems and variation techniques. Some of the basic ideas in this approach are introduced in Chapter 2. The other is to approximate the solutions by using computational methods. As long as the numerical approximation is accurate enough, it could help us understand the analytic solution. Therefore, we are also interested in designing accurate and efficient schemes to approximate solutions.

Our main research object of this thesis is in this spirit of accurate numerical approximation. We are mainly concerned with phase field models such as the Allen–Cahn equation and the Cahn–Hilliard equation, which were first introduced in [1]. They correspond to the research area of material science, which are very popular topics in the study of partial differential equations, see e.g. [35, 23, 25, 6, 18, 24, 33, 38, 48]. The phase field models we are concerned with usually have such a general form,

$$\partial_t u = \mathcal{L}u + \mathcal{G}(u),$$

where $\partial_t u$ means the rate of change of the quantity u with respect to time, $\mathcal{L}$ represents a linear operator such as the Laplacian and biharmonic operator, and $\mathcal{G}$ is a nonlinear one like a scalar function. Phase field methods are popular in modeling nowadays and are often used for describing interfacial phenomena. In phase-field models, each phase takes different constant values, connected by smooth transitions around anti-phase boundaries between them. The phase-field models were originally introduced to describe the micro-

structure evolution [6, 18] and phase transition [24, 33, 38, 48], while these years, they are also applied to many other physical phenomena, including phase separation of block copolymers, solid-solid transitions and infiltration of water into a porous medium, see, e.g. [35, 23, 25].

For example, the Allen–Cahn equation usually reads

$$\partial_t u = \epsilon^2 \Delta u + (u - u^3).$$

Thus, for the Allen–Cahn equation, the linear operator is the Laplacian operator, which indicates a diffusion term, and the nonlinear term is usually taken as $f(u) = u - u^3$, which represents a reaction term and may have different contexts in different equations. In this work we will use different numerical methods to discretize these quantities in the equation, approximate the solutions and study how numerical solutions behave. The Allen–Cahn equation was originally introduced by Allen and Cahn to describe the motion of anti-phase boundaries in crystalline solids. The solution $u(x, t)$ describes the concentration of two crystal orientations of the same material. In this phase model, $u = 1$ represents one orientation and $u = -1$ represents the other. The parameter ϵ here is the width of the interface between two phases, which is positive and small. Since then, the Allen–Cahn equation has been widely applied to many complicated moving interface problems in materials science and fluid dynamics through a phase-field approach.

In this thesis, we focus on structure-preserving numerical schemes for phase field models including the Allen–Cahn equation, Cahn–Hilliard equation and the Molecular Beam

equation, see e.g [1, 6, 18, 16, 17]. Unknowns in PDEs represent certain physically meaningful quantities and satisfy specific structures. For example, if u represents the concentration of oxygen and satisfies some PDEs in the reaction process, then there is no doubt that we do not want u to be negative. Such structures come from the physical background and become mathematical restrictions which our solutions should not violate. The essential features of these phase field models are the maximum bound property for the Allen–Cahn equation and the energy dissipation law for all of three phase field models mentioned above. Therefore, it is important to design numerical schemes satisfying these properties. In Chapter 3-5, we consider three different schemes to numerically solve these phase field models.

The first of them is a particular explicit Runge–Kutta (RK) method which satisfies the discrete maximum bound property and the energy dissipation law for the Allen–Cahn equation. We make use of the Strong Stability Preserving (SSP) RK scheme to preserve the maximum bound property, prove the energy dissipation by detailed analysis and finally derive high-order explicit Runge–Kutta schemes which preserve both structures. Since it is explicit and requires small time steps, it is more of a theoretical result.

The second scheme considered is the second-order exponential time differencing Runge–Kutta (ETDRK2) method which unconditionally satisfies both structures for the AC and the energy dissipation law for all of phase field models. Here "unconditionally" means that the time step can be arbitrarily large without disturbing the discrete structures. Besides this scheme makes use of the Duhamel's principle and is still linear to solve, so it is very stable, efficient and useful in application.

The third scheme is the implicit-explicit (IMEX) Runge–Kutta method, which is also known as the semi-implicit RK scheme. It takes the linear part implicitly and the nonlinear part explicitly to get the numerical solution. In this way, high order solutions which decreases the original energy for a family of phase field models can be derived by a class of IMEX RK methods satisfying certain conditions. This work gives the first scheme which is one-step, high-order, linear to solve and also preserves the original energy dissipation law for a wide class of phase field models.

The structure of the thesis is organized as follows. Chapter 2 introduces notations and preliminaries that will be used in the thesis. The following Chapters 3, 4 and 5 present results related to the explicit Runge–Kutta method, the exponential time differencing Runge–Kutta method and the implicit-explicit Runge–Kutta methods respectively. Conclusions and remarks for future work could be found in Chapter 6.

Chapter 2

Notation and preliminaries

2.1 Notation and definitions

2.1.1 $O(g(x))$ and $o(g(x))$

For functions $f(x)$ and $g(x)$, if there exists a positive constant C such that $|f(x)| \leq C|g(x)|, \forall x$, then $f(x)$ is $O(g(x))$. If $\lim_{x \to x_0} \frac{f(x)}{g(x)} = 0, \forall x$ where x_0 depends on the context, then we say $f(x)$ is $o(g(x))$.

2.1.2 $\lesssim$ and $\ll$

For two quantities A and B, if there exists a positive constant C such that $A \leq CB$, then we denote $A \lesssim B$. Similarly if there exists a positive constant c such that $A \geq cB$, we have $A \gtrsim B$. If $A \lesssim B$ and $A \gtrsim B$, we say $A \sim B$.

We say $A \ll B$ if A/B is very small, which is clear from the context.

2.1.3 L^p Space

On the given domain Ω, for $1 \leq p < \infty$ (in the rest of the thesis we always assume this condition unless otherwise stated), the space $L^p(\Omega)$ consists of all measurable functions which satisfy

$$\int_\Omega |f(x)|^p dx \leq \infty.$$

For $f \in L^p(\Omega)$, we define the L^p norm by

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p dx \right)^{1/p}.$$

2.1.4 Weak Derivatives and Sobolev Space

We use the following notations:

$$x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n,$$

$$\alpha = (\alpha_1, \alpha_2, ..., \alpha_n) \in \mathbb{Z}_+^n,$$

$$\partial^\alpha f = \frac{\partial^{\alpha_1 + ... + \alpha_n}}{\partial_{x_1}^{\alpha_1} ... \partial_{x_n}^{\alpha_n}} f.$$

We define the weak derivative in the following sense: given $u \in L^1_{loc}(\Omega)$, we say a function v is the weak derivative of u if $\forall \phi \in C_0^\infty(\Omega)$,

$$\int_\Omega u(x) \partial^\alpha \phi(x) dx = (-1)^{\alpha_1 + ... + \alpha_n} \int_\Omega v(x) \phi(x) dx,$$

and also we denote $v(x) = \partial^\alpha u(x)$. It is a direct conclusion from the definition that if u is a smooth function, then its classic derivative is also its weak derivative and the above equation is simply repeated integration by parts.

Suppose $u \in L^p(\Omega)$ and all weak derivatives $\partial^\alpha u$ exist for $|\alpha| = \alpha_1 + ... + \alpha_n \leq k$ for some constant k, i.e. $\partial^\alpha \in L^p(\Omega)$ for all $|\alpha| \leq k$, then we say $u \in W^{k,p}(\Omega)$, and such spaces are named Sobolev space. The norm equipped to the Sobolev space is defined as

$$\|u\|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| \leq k} \int_\Omega |\partial^\alpha u|^p dx \right)^{1/p}.$$

When $p = 2$, we use the convention $H^2(\Omega) = W^{k,2}(\Omega)$.

2.1.5 Fourier Transform

Throughout the thesis we use the following convention for the Fourier expansion on $\mathbb{T}^d = (\mathbb{R}/(2\pi))^d$:

$$f(x) = \frac{1}{(2\pi)^{d/2}} \sum_{k \in \mathbb{Z}^d} \hat{f}(k) e^{ikx}, \hat{f}(k) = \int_\Omega f(x) e^{-ikx} dx,$$

where $i = \sqrt{-1}$ is the imaginary unit. Based on the Fourier expansion we define the equivalent H^s norm and $\dot{H}^2$ norm by

$$\|f\|_{H^s} = \frac{1}{(2\pi)^{d/2}} \left(\sum_{k \in \mathbb{Z}^d} (1 + |k|^{2s}) |\hat{f}(k)|^2 \right)^{1/2},$$
$$\|f\|_{\dot{H}^s} = \frac{1}{(2\pi)^{d/2}} \left(\sum_{k \in \mathbb{Z}^d} |k|^{2s} |\hat{f}(k)|^2 \right)^{1/2}.$$

The equivalence of these two norms are well-known.

2.1.6 Convergence of Fourier Series in Periodic Domains

Given a periodic function $f \in L^p(\mathbb{T}^d)$, we denote the Dirichlet partial sum

$$D_N f = \frac{1}{(2\pi)^d} \sum_{|k| \leq N} \hat{f}(k) e^{ikx},$$

and we have the following convergence result

$$\|D_N f - f\|_{L^p(\mathbb{T}^d)} \to 0,$$

and $D_N f \to f$, pointwise almost everywhere.

2.1.7 Duhamul's Formula

Assume a function $u(x,t)$ defined on $\Omega \times (0, +\infty)$ satisfies a linear inhomogeneous evolution equation

$$u_t = Lu + f(x,t),$$

$$u(x,0) = u_0(x),$$

where L is a linear differential operator which involves no time derivatives and the equation is equipped with certain boundary conditions. Then the solution of the system can be written as

$$u(x,t) = e^{Lt} u_0(x) + \int_0^t e^{L(t-s)} f(x,s) ds,$$

where $e^{Lt} u_0$ is equivalent to solving to time t a homogeneous equation $u_t = Lu$ with the initial data u_0.

2.2 Important Inequalities

2.2.1 Hölder's Inequality

Given $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$ such that $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\|fg\|_{L^1(\Omega)} \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}.$$

2.2.2 Young's Inequality

Given a, b, p, q positive real numbers, such that $\frac{1}{p} + \frac{1}{q} = 1$, then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

2.2.3 Morrey's Inequality

Assume Ω is a bounded Lipschitz domain in $\mathbb{R}^d$ where $d \leq 3$ and $f \in H^2(\Omega)$, then

$$\|f\|_{L^\infty(\Omega)} \lesssim \|f\|_{H^2(\Omega)}.$$

A stronger estimate could be proven with the help of Hölder space.

2.2.4 Gagliardo-Nirenberg Interpolation Inequality

Given the function $u : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain, $1 \leq q, r \leq \infty$, and a natural number m. Assume there exist a natural number j and a real number α such that

$$\frac{1}{p} = \frac{j}{d} + (\frac{1}{r} - \frac{m}{d})\alpha + \frac{1-\alpha}{q}$$

and

$$\frac{j}{m} \leq \alpha \leq 1,$$

then we have

$$\|D^j u\|_{L^p} \leq C_1 \|D^m u\|_{L^r}^{\alpha} \|u\|_{L^q}^{1-\alpha} + C_2 \|u\|_{L^s},$$

where $s > 0$ is arbitrary.

Chapter 3

Energy plus maximum bound preserving Runge–Kutta methods for the Allen–Cahn equation

3.1 Introduction

Strong stability preserving Runge–Kutta (SSP-RK) methods have been developed for numerical solution of *hyperbolic partial differential equations*. Starting with Shu [44] it was observed that some Runge–Kutta methods can be decomposed into *convex combinations* of forward Euler steps, and so any convex functional property satisfied by forward Euler will be preserved by these higher-order time discretizations, generally under a different time-step restriction. This approach was used to develop second- and third-order Runge–Kutta methods that preserve the strong stability properties of the spatial discretizations developed in that work. In fact, this approach also guarantees that the intermediate stages in a Runge–Kutta method satisfy the strong stability property as well. More references in

this direction can be found in [29, 30, 34, 39, 37] and a useful survey article of Gottlieb, Shu and Tadmor [28].

The aim of this work is to extend this SSP theory to deal with the nonlinear phase field equation of the Allen-Cahn type. To this end, we consider the numerical approximation of the Allen–Cahn equation

$$u_t = \epsilon \Delta u + \frac{1}{\epsilon} f(u), \qquad x \in \Omega,\ t \in (0, T], \tag{3.1.1}$$

with the initial condition

$$u(x, 0) = u_0(x), \qquad x \in \Omega,$$

and the homogeneous Neumann boundary condition or periodic boundary condition, where Ω is a bounded domain in R^d $(d = 1, 2, 3)$. Notice that the equation (3.1.1) is slightly different from the Allen–Cahn equation in Chapter 1, which has been scaled by the parameter ϵ. This scaling does not affect the intrinsic properties of the system but only means we are concerned with a different time scale. Since the explicit Runge-Kutta methods must have certain restrictions for time steps and we want to point out the relationship between the restrictions and the parameter ϵ, in this chapter we take the Allen–Cahn equation in the form (3.1.1).

In this work, we consider the polynomial double-well potential

$$F(u) = \frac{1}{4}(1 - u^2)^2 \tag{3.1.2}$$

and correspondingly,

$$f(u) = -F'(u) = u - u^3. \tag{3.1.3}$$

The solution $u(x,t)$ describes the concentration of two crystal orientations of the same material. In this phase model, $u = 1$ represents one orientation and $u = -1$ represents the other. The parameter ϵ here is the width of the interface between two phases, which is positive and small.

The Allen–Cahn equation can be viewed as the L^2 gradient flow of the Ginzburg-Landau free energy

$$\mathcal{E}(u) = \int_\Omega \left(\frac{\epsilon}{2} |\nabla u|^2 + \frac{1}{\epsilon} F(u) \right) dx. \tag{3.1.4}$$

The L^2 gradient flow structure corresponds to an energy dissipation law. This means that the energy is decreasing as a function of time,

$$\frac{d\mathcal{E}}{dt} = -\int_\Omega \left(\epsilon \Delta u + \frac{1}{\epsilon} f(u) \right)^2 dx \leq 0. \tag{3.1.5}$$

Another significant feature of the Allen–Cahn equation is its maximum bound preserving (MBP) property in the sense

$$\|u(\cdot,t)\|_\infty \leq 1 \tag{3.1.6}$$

provided that the initial and boundary values are bounded by 1.

The present work seems to be the first effort to study high-order time discretizations aiming to preserve both (3.1.5) and (3.1.6). By applying Shu's SSP-RK theory [44], i.e., using the property of the forward Euler method repetitively, we will first obtain a sufficient condition to verify whether a Runge–Kutta method is MBP, and also give a necessary and sufficient condition for s-stage s-th order MBP-RK methods. Both results will be established by using the so-called Butcher Tableau so the results are easy to verify. Moreover,

we will provide a necessary condition to judge whether the MBP-RK solutions preserve the energy dissipation law. Finally, we will provide some RK2, RK3 and RK4 methods which satisfy both both (3.1.5) and (3.1.6). A special RK3 method violating the energy dissipation law will be also reported.

This chapter is organized as follows. Section 3.2 contains some preliminaries and notations. Section 3.3 analyzes high-order MBP-RK methods to the Allen–Cahn equation by using Shu's theory. We build up the relationship between the Butcher Tableau and the so-called Shu-Osher form [45]. Section 3.4 studies how to preserve the energy dissipation law (3.1.5) for the relevant Runge–Kutta methods. Section 3.5 applies the theory of Sections 3.3 and 3.4 to some typical Runge–Kutta schemes for the Allen–Cahn equation.

3.2 Interplay between the Butcher Tableau and Shu-Osher form

The Runge–Kutta methods are a family of implicit and explicit iterative methods used in temporal discretization for the approximate solutions of ordinary differential equations (ODEs). Consider an ODE system in time $u' = G(u)$. An explicit Runge–Kutta method

is commonly written in the form:

$$\begin{aligned} v_0 &= u_n, \\ v_i &= u_n + \tau \sum_{j=0}^{i-1} a_{ij} G(v_j), \qquad 1 \le i \le s-1 \\ u_{n+1} &= u_n + \tau \sum_{j=0}^{s-1} b_j G(v_j). \end{aligned} \tag{3.2.1}$$

In other words, to specify a particular method, one needs to provide the integer s (the number of stages), and the coefficients a_{ij} (for $1 \le j < i \le s$), b_j (for $j = 1, \cdots, s$) and c_j (for $j = 1, \cdots, s-1$). The matrix (a_{ij}) is called the Runge–Kutta matrix, while the b_j and c_j are known as the weights and the nodes [4]. These data are usually arranged in a mnemonic device, known as a *Butcher tableau* (after John C. Butcher):

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ c_1 & a_{1,0} & 0 & & & \\ c_2 & a_{2,0} & a_{2,1} & 0 & & \\ \dots & \dots & \dots & \dots & 0 & \\ c_{s-1} & a_{s-1,0} & a_{s-1,1} & \dots & a_{s-1,s-2} & 0 \\ \hline & b_0 & b_1 & \dots & \dots & b_{s-1} \end{array} \tag{3.2.2}$$

where

$$c_i = \sum_{j=0}^{i-1} a_{ij}, \quad i \ge 1. \tag{3.2.3}$$

If we define $a_{sj} = b_j$ for all $j \geq 0$, then the scheme (3.2.1) becomes

$$\begin{aligned} u_n &= v_0, \\ v_i &= u_n + \tau \sum_{j=0}^{i-1} a_{ij} G(v_j), \qquad 1 \leq i \leq s \\ u_{n+1} &= v_s. \end{aligned} \tag{3.2.4}$$

We further define a strictly lower-triangular matrix A_L as

$$A_L = \begin{bmatrix} 0 & & & & \\ a_{1,0} & 0 & & & \\ a_{2,0} & a_{2,1} & 0 & & \\ \dots & \dots & \dots & \dots & \\ a_{s,0} & a_{s,1} & \dots & a_{s,s-1} & 0 \end{bmatrix}. \tag{3.2.5}$$

On the other hand, the Runge–Kutta method can be written in the Shu–Osher form [45]:

$$\begin{aligned} v_0 &= u_n, \\ v_i &= \sum_{k=0}^{i-1} \Big(\alpha_{ik} v_k + \tau \beta_{ik} G(v_k) \Big), \qquad 1 \leq i \leq s \\ u_{n+1} &= v_s, \end{aligned} \tag{3.2.6}$$

where consistency condition requires

$$\sum_{k=0}^{i-1} \alpha_{ik} = 1, \qquad 1 \leq i \leq s. \tag{3.2.7}$$

It is observed in Shu [44, 45] if all coefficients are positive, i.e., $\alpha_{ik} > 0$ and $\beta_{ik} \geq 0$, then the solution can be viewed as convex combinations of forward Euler solutions. Based on

this theory, the consistency condition (3.2.7) and the positivity conditions $\alpha_{ik} \geq 0$ and $\beta_{ik} \geq 0$ can ensure the Strong Stability Preserving (SSP) properties.

Proposition 2.1. *([44, 45]) If the following so-called RK-SSP condition is satisfied*

$$\sum_{k=0}^{i-1} \alpha_{ik} = 1, \quad 1 \leq i \leq s; \qquad \alpha_{ik} \geq 0, \;\; \beta_{ik} \geq 0, \quad 0 \leq k < i \leq s, \tag{3.2.8}$$

(when $\alpha_{ik} = 0$ then $\beta_{ik} = 0$), then the Runge–Kutta type method of type (3.2.6) satisfies the SSP condition in the sense that

$$\|u_{n+1}\| \leq \|u_n\|,$$

where $\|\cdot\|$ is the maximum norm or in the TV semi-norm.

Below we explore the relationship between the original form (3.2.1) and the Shu–Osher form (3.2.6). We rewrite the Butcher form with the help of the consistency condition (3.2.7):

$$v_i = v_0 + \tau \sum_{j=0}^{i-1} a_{ij} G(v_j) = \alpha_{i0} v_0 + \sum_{j=1}^{i-1} \alpha_{ij} v_0 + \tau \sum_{j=0}^{i-1} a_{ij} G(v_j). \tag{3.2.9}$$

We further use (3.2.1) for the above result to obtain

$$\begin{aligned} v_i &= \alpha_{i0} v_0 + \sum_{j=1}^{i-1} \alpha_{ij} \left(v_j - \tau \sum_{k=0}^{j-1} a_{jk} G(v_k) \right) + \tau \sum_{k=0}^{i-1} a_{ik} G(v_k) \\ &= \sum_{k=0}^{i-1} \left[\alpha_{ik} v_k + \tau \left(a_{ik} - \sum_{j=k+1}^{i-1} \alpha_{ij} a_{jk} \right) G(v_k) \right], \qquad 1 \leq i \leq s. \end{aligned}$$

By defining

$$\beta_{ik} = a_{ik} - \sum_{j=k+1}^{i-1} \alpha_{ij} a_{jk}, \qquad 0 \leq k \leq i-1, \tag{3.2.10}$$

the relationship between the original form (3.2.1) and the Shu–Osher form (3.2.6) is established.

Theorem 2.1. *If all elements in the strictly lower-triangular matrix A_L in (3.2.3) are positive, i.e. $a_{ik} > 0$ for all $0 \le k < i \le s$, then there exist coefficients $\alpha_{ij}, \beta_{ij} \ge 0$ such that the corresponding explicit Runge–Kutta scheme (3.2.1) satisfies the RK-SSP condition.*

Proof. We need to use the given positive elements a_{ij} $(0 \le j < i \le s)$ to construct positive coefficient pairs $(\alpha_{ik}, \beta_{ik})$. Let

$$\delta = \min_{0 \le k < i \le s} \frac{a_{ik}}{\sum_{j=k+1}^{i-1} a_{jk}},$$

and let

$$\begin{aligned} \alpha_{ij} &= \min\left\{\frac{\delta}{2}, \frac{1}{2(i-1)}\right\}, \qquad \forall\, 1 \le i \le s,\ 1 \le j < i, \\ \alpha_{i0} &= 1 - (i-1)\cdot a_{i1}, \qquad 1 \le i \le s. \end{aligned}$$

It is easy to check that $\alpha_{ik} > 0$ for all $0 \le k < i \le s$. Using the relation (3.2.10) and the fact $\alpha_{ij} < \delta$ gives

$$\begin{aligned} \beta_{ik} &= a_{ik} - \sum_{j=k+1}^{i-1} \alpha_{ij} a_{jk} \\ &> a_{ik} - \sum_{j=k+1}^{i-1} \delta a_{jk} \\ &= \sum_{j=k+1}^{i-1} a_{jk} \left(\frac{a_{ik}}{\sum_{j=k+1}^{i-1} a_{jk}} - \delta \right) \ge 0. \end{aligned}$$

Finally, it is easy to observe that

$$\sum_{k=0}^{i-1} \alpha_{ik} = \alpha_{i0} + \sum_{k=1}^{i-1} \alpha_{i1} = 1 - (i-1)\cdot\alpha_{i1} + (i-1)\cdot\alpha_{i1} = 1.$$

This completes the proof of the theorem. □

The above theorem gives a simple sufficient condition which can convert a Runge–Kutta method to be of Shu–Osher type satisfying the RK-SSP condition (3.2.1). Below we derive a sufficient and necessary condition for a wide class of Runge–Kutta method.

Theorem 2.2. *An explicit Runge–Kutta method with non-zero sub-diagonal elements satisfies the RK-SSP condition (3.2.8) if and only if all elements in the strictly lower-triangular part of A_L in (3.2.5) are positive.*

Proof. The sufficient condition is proved in Theorem 2.1. We now prove the necessary condition. In this case, the explicit Runge–Kutta method with non-zero sub-diagonal elements satisfies the RK-SSP condition (3.2.8). Define $order(a_{ik}) = (i-1) * s + k,\ 0 \leq k < i \leq s$. If there exist non-positive elements in A, we take the first of them in the sense of order, $a_{pq} \leq 0$. Since the RK scheme satisfies (3.2.8), we have

$$\alpha_{ik} \geq 0, \quad \beta_{ik} = a_{ik} - \sum_{j=k+1}^{i-1} \alpha_{ij} a_{jk} \geq 0, \quad 0 \leq k \leq i-1.$$

In particular, we have

$$a_{pq} - \sum_{j=q+1}^{p-1} \alpha_{pj} a_{jq} \geq 0. \tag{3.2.11}$$

As a_{pq} is the first non-positive element in the sense of order, all a_{jq} in the summation above are all positive. We then have two cases.

- If $a_{pq} < 0$, then it is easy to see a contradiction to (3.2.11).
- If $a_{pq} = 0$, as the other a_{jq} is positive then it follows from (3.2.11) that all α_{pj} in (3.2.11) are 0, and in particular, $\alpha_{p,p-1} = 0$, which leads to $\beta_{p,p-1} = 0$. Note that by

(3.2.10) we have $\beta_{p,p-1} = a_{p,p-1}$. Consequently, we have $a_{p,p-1} = 0$ which contradicts the non-zero sub-diagonal element assumption.

This completes the proof of the theorem. □

One direct result is the following proposition.

Proposition 2.2. *An s-stage sth-order explicit Runge–Kutta method satisfies the RK-SSP condition (3.2.8) if and only if all elements in the strictly lower-triangular part of A_L are positive.*

Proof. For the s-order RK scheme, in order to match the highest order term in the Taylor expansion, we must have

$$\frac{1}{s!} = a_{1,0}a_{2,1}\cdots a_{s,s-1}$$

which guarantees all sub-diagonal elements $a_{i,i-1}$ are non-zero. □

The following proposition is given in [4, 34], while Theorem 2.2 provides a different perspective.

Proposition 2.3. *There does not exist any 4-stage 4th-order explicit Runge–Kutta method satisfying the RK-SSP condition (3.2.8).*

Proof. The only 4th-order RK whose coefficients are all non-negative is the classic RK4

[4], whose Butcher tableau reads

$$\begin{array}{c|cccc} 0 & 0 & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1/2 & 0 & 0 & 1/2 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} \tag{3.2.12}$$

Note that $a_{21} = a_{30} = a_{31} = 0$, i.e., they are not positive. Consequently, the classical RK4 does not satisfy the RK-SSP condition (3.2.8). □

Proposition 2.4. *Any irreducible RK method whose elements in the strictly lower-triangular part are all positive can not have order greater than 4.*

Proof. It is known that there is no irreducible SSP-RK method which has order greater than 4 [34, 39]. If an explicit irreducible Runge–Kutta method has positive strictly lower-triangular part, then based on Theorem 2.1 it must satisfy the RK-SSP condition, which contradicts the existing theory of [34, 39]. □

Remark 3.2.1. *Theorems in this section could also be derived by the contractivity theory [34, 22], although the approaches and illustrations are different.*

3.3 MBP-RK methods for the Allen-Cahn equation

We will use the central finite difference discretization to the Allen–Cahn equation in space. Without loss of generality, we consider the computational domain $[0, 2\pi]$ with the

periodic boundary condition and let the space mesh size $h = 2\pi/N$. Denote the grid points as $\{x_j = jh, j = 0, 1, ..., N-1\}$ and the forward finite difference matrix of ∂_x by D_1

$$D_1 = \frac{1}{h}\begin{bmatrix} 1 & & & & -1 \\ -1 & 1 & & & \\ & \cdots & \cdots & \cdots & \\ & & & -1 & 1 \end{bmatrix}_{N\times N}. \tag{3.3.1}$$

Thus we have the central difference discretization operator $D = -D_1^T D_1$ for the Laplacian Δ. It is well-known that the discrete operator D is of second-order accuracy to approximate the Laplacian operator.

Lemma 3.1. *Given any vector* $\mathbf{v}$ *and scalar* $\alpha > 2$*, the following inequality holds:*

$$\left\| \left(I + \frac{1}{\alpha} h^2 D \right) \mathbf{v} \right\|_\infty \leq \|\mathbf{v}\|_\infty. \tag{3.3.2}$$

Proof. When $\alpha > 2$, $\alpha I + h^2 D$ is a tri-diagonal matrix whose elements are all positive. Besides, note that the sum of each row of D is zero. Consequently, the sum of every row of $\alpha I + h^2 D$ equals to constant α. Observe that

$$\begin{aligned} \|(\alpha I + h^2 D)\mathbf{v}\|_\infty &= \max_j \|a_j \mathbf{v}_{j-1} + b_j \mathbf{v}_j + c_j \mathbf{v}_{j+1}\|_\infty \\ &\leq (|a_j| + |b_j| + |c_j|) \max_j \|\mathbf{v}_j\|_\infty = \alpha \|\mathbf{v}\|_\infty, \end{aligned} \tag{3.3.3}$$

where a_j, b_j, c_j are corresponding coefficients in the matrix. This completes the proof. □

Lemma 3.2 ([27])**.** *Denote the discrete Fourier transform as* F_N *and the conjugate transpose as* $(\cdot)^H$*, then it holds that*

$$D = F_N^H \Lambda F_N, \quad \Lambda = diag([\lambda_0, \cdots, \lambda_{N-1}]), \tag{3.3.4}$$

where $\lambda_j = -(2-2\cos(jh))/h^2$ *are eigenvalues of* D.

One direct result of this lemma is the following inverse inequality: Given any vector $\mathbf{u}$, it holds that

$$0 \le -\mathbf{u}^T D\mathbf{u} \le \frac{4}{h^2}\mathbf{u}^T\mathbf{u}. \tag{3.3.5}$$

Note that the above property holds for more general boundary conditions and domains, and in these situations the coefficient 4 in (3.3.5) will be replaced by a positive constant C depending only on the boundary conditions and the domain.

For simplicity and for ease of demonstrating the main ideas, in this chapter we only consider the 1D case. For multi-dimension cases, by using the tensor product for the discrete Laplacian operator in 2D and 3D, results similar to the 1D case can be obtained.

The semi-discrete finite difference discretization of the Allen–Cahn equation reads

$$\frac{d}{dt}\mathbf{u} = \epsilon D\mathbf{u} + \frac{1}{\epsilon}f(\mathbf{u}) =: G(\mathbf{u}). \tag{3.3.6}$$

We list following two properties for system (3.3.6) resulting from the so-called method-of-lines approach.

- If the initial value satisfies $\|\mathbf{u}_0\|_\infty \le 1$, then the solution $\mathbf{u}(t)$ given by (3.3.6) satisfies the maximum bound preserving (MBP) property:

$$\|\mathbf{u}(t)\|_\infty \le 1, \quad \forall t \ge 0. \tag{3.3.7}$$

- Let

$$E_h(\mathbf{u}) = \frac{\epsilon}{2}\|D_1\mathbf{u}\|_{l^2}^2 + \frac{1}{4\epsilon}\|1-\mathbf{u}^2\|_{l^2}^2. \tag{3.3.8}$$

Then the solutions of system (3.3.6) satisfy the semi-discrete energy dissipation law

$$\frac{d}{dt}E_h = -\left\|\frac{d\mathbf{u}}{dt}\right\|_{l^2}^2 \leq 0. \tag{3.3.9}$$

Note that the first result can be found in, e.g., [52], and the second result can be obtained by taking the L^2 inner product of (3.3.6) with $\frac{d}{dt}\mathbf{u}$.

In this section, we are concerned about MBP Runge–Kutta method for the Allen–Cahn equation. The main strategy is to extend the Shu-Osher theory for the hyperbolic conservation laws to deal with the Allen–Cahn solutions.

3.3.1 Forward Euler solution

In this section we discretize the semi-discrete system in the time direction by applying forward Euler method.

Before providing a useful theorem, we need the following simple results, which can be obtained by an elementary proof.

Lemma 3.3. *For any positive number a, if $-4a \leq c \leq a/2$, then the function $g(x) = ax + c(x - x^3)$ satisfies*

$$|g(x)| \leq a, \quad \forall x \in [-1, 1]. \tag{3.3.10}$$

The following theorem characterizes the Euler property for the system (3.3.6).

Theorem 3.1. *Consider the ODE system (3.3.6). If $\tau < \tau_0 := \min\{h^2/4\epsilon, \epsilon/4\}$, then for any vector $\mathbf{u}$ satisfying $\|\mathbf{u}\|_\infty \leq 1$, we have*

$$\|\mathbf{u} + \tau G(\mathbf{u})\|_\infty \leq 1. \tag{3.3.11}$$

Proof. Note that

$$\begin{aligned}\|\mathbf{u}+\tau G(\mathbf{u})\|_\infty &= \left\|\mathbf{u}+\tau\left(\epsilon D\mathbf{u}+\frac{1}{\epsilon}f(\mathbf{u})\right)\right\|_\infty \\ &= \left\|\left(\frac{1}{2}\mathbf{u}+\tau\epsilon D\mathbf{u}\right)+\left(\frac{1}{2}\mathbf{u}+\frac{\tau}{\epsilon}f(\mathbf{u})\right)\right\|_\infty \\ &\leq \left\|\frac{1}{2}\mathbf{u}+\tau\epsilon D\mathbf{u}\right\|_\infty+\left\|\frac{1}{2}\mathbf{u}+\frac{\tau}{\epsilon}f(\mathbf{u})\right\|_\infty.\end{aligned}$$

Using Lemma 3.1 and the assumption $\tau < h^2/4\epsilon$ gives

$$\left\|\frac{1}{2}\mathbf{u}+\tau\epsilon D\mathbf{u}\right\|_\infty \leq \frac{1}{2}.$$

Using Lemma 3.3 and the assumption $\tau < \epsilon/4$ yields

$$\left\|\frac{1}{2}\mathbf{u}+\frac{\tau}{\epsilon}f(\mathbf{u})\right\|_\infty \leq \frac{1}{2}.$$

Combining the above three results gives the desired result. □

Remark 3.3.1. *The relationship between the time step and ϵ comes from here and is inevitable if one wants to solve (3.3.6) with explicit methods directly.*

3.3.2 MBP-RK methods

Theorem 3.2. *Consider the Runge–Kutta scheme (3.2.6) with G defined by (3.3.6). If the SSP-RK property (3.2.8)is satisfied, then*

$$\|u^n\|_\infty \leq 1 \quad \Longrightarrow \quad \|u^{n+1}\|_\infty \leq 1 \tag{3.3.12}$$

under the time-step restriction

$$\tau \leq \tau_{SSP} := \min_{0\leq k<i\leq s}\frac{\alpha_{ik}}{\beta_{ik}}\cdot\tau_0, \quad \text{with} \quad \tau_0 = \min\left\{\frac{h^2}{4\epsilon},\frac{\epsilon}{4}\right\}. \tag{3.3.13}$$

Note that the ratio above is understood as infinity whenever $\beta_{ik} = 0$.

Proof. The proof is based on the original SSP machinery, see, e.g., [28, 30]. In particular, note that for each i, we have

$$\|v_i\|_\infty = \left\| \sum_{k=0}^{i-1} (\alpha_{ik} v_k + \tau \beta_{ik} G(v_k)) \right\|_\infty \leq \left\| \sum_{k=0}^{i-1} \alpha_{ik} \left(v_k + \tau \frac{\beta_{ik}}{\alpha_{ik}} G(v_k) \right) \right\|_\infty . \tag{3.3.14}$$

Under the assumption (3.3.13), we have $\tau \beta_{ik}/\alpha_{ik} \leq \tau_0$. Then using Theorem 3.1 gives

$$\|v_i\|_\infty \leq \sum_{k=0}^{i-1} \alpha_{ik} \left\| v_k + \tau \frac{\beta_{ik}}{\alpha_{ik}} G(v_k) \right\|_\infty \leq \sum_{k=0}^{i-1} \alpha_{ik} \cdot 1 = 1. \tag{3.3.15}$$

This yields the desired result (3.3.12). □

3.4 The discrete energy dissipation law

The discrete energy is defined as follows

$$E(\mathbf{u}) = -\frac{\epsilon}{2} \mathbf{u}^T D \mathbf{u} + \frac{1}{\epsilon} \sum_{j=1}^{N} F(\mathbf{u}_j). \tag{3.4.1}$$

For ease of our derivation, we will consider a special class of Runge–Kutta schemes.

Lemma 4.1. *Given a Butcher tableau (3.2.2) and the corresponding RK scheme (3.2.4). By suitably arranging coefficients* $\{c_{ik}\}$*, we can obtain a class of RK scheme of the following form:*

$$v_i = \sum_{k=0}^{i-1} p_{ik} v_k + d_i \tau G(v_{i-1}), \qquad 1 \leq i \leq s. \tag{3.4.2}$$

Proof. We wish to convert the RK formula (3.2.4) into the Shu-Osher format. It follows from (3.2.10) that

$$\begin{aligned} v_i &= \sum_{k=0}^{i-1} \left[\alpha_{ik} v_k + \left(a_{ik} - \sum_{l=k+1}^{i-1} a_{lk}\alpha_{il} \right) \tau G(v_k) \right] \\ &= \sum_{k=0}^{i-1} \alpha_{ik} v_k + \sum_{k=0}^{i-2} \left(a_{ik} - \sum_{l=k+1}^{i-1} a_{lk}\alpha_{il} \right) \tau G(v_k) + a_{i,i-1}\tau G(v_{i-1}). \end{aligned} \tag{3.4.3}$$

By forcing the second last term in (3.4.3) to 0, a set of values of $\{\alpha_{ik}\}$ can be determined by $\{a_{ik}\}$. This will leave only the last G-term in (3.4.3). Therefore, we derive $p_{ik} = \alpha_{ik}$ and $d_i = a_{i,i-1}$ and thus the scheme has the unique form (3.4.2). □

Note that the consistency condition requires $\sum_{k=0}^{i-1} p_{ik} = 1$, but now the coefficients in (3.4.2) may be negative.

Before we present the main result of this section, we state the following result whose proof is quite straightforward:

$$\frac{1}{4}[(a^2-1)^2 - (b^2-1)^2] \le (b^3-b)(a-b) + (a-b)^2, \quad \forall a, b \in [-1,1]. \tag{3.4.4}$$

Theorem 4.1. *For a given SSP-RK solution which has the form (3.4.2), we define an upper triangular matrix* Φ *given by*

$$\Phi_{ij} = \sum_{k=0}^{i-1} \frac{p_{jk}}{d_j}, \qquad i \le j \tag{3.4.5}$$

and the energy discriminant

$$\Delta_E = \frac{1}{2}(\Phi + \Phi^T). \tag{3.4.6}$$

If Δ_E is positive-definite, then the energy is non-increasing under the time step restriction

$$\tau \leq \min\left\{\frac{\lambda}{\frac{1}{\epsilon}+\frac{2\epsilon}{h^2}},\ \tau_{SSP}\right\}, \tag{3.4.7}$$

where τ_{SSP} is the SSP-RK time-restriction given by (3.3.13), and λ is the smallest eigenvalue of Δ_E.

Proof. Rewrite (3.4.2) by using the form of G (for simplicity we drop the ϵ scale in this proof and notice that here $\mathbf{v}_i$ are vectors) and the consistency condition:

$$\begin{aligned} f(\mathbf{v}_i) &= \frac{1}{d_{i+1}\tau}\left(\mathbf{v}_{i+1}-\sum_{k=0}^{i} p_{i+1,k}\mathbf{v}_k\right) - D\mathbf{v}_i \\ &= \frac{\mathbf{v}_{i+1}-\mathbf{v}_i}{d_{i+1}\tau} + \frac{1}{d_{i+1}\tau}\sum_{k=0}^{i-1} p_{i+1,k}(\mathbf{v}_i-\mathbf{v}_k) - D\mathbf{v}_i. \end{aligned}$$

By using the definition of the potential F and by using (3.4.4), we obtain

$$\begin{aligned} &\sum_{j=1}^{N} F((\mathbf{u}_{n+1})_j) - F((\mathbf{u}_n)_j) = \sum_{i=0}^{s-1}\sum_{j=1}^{N} F((\mathbf{v}_{i+1})_j) - F((\mathbf{v}_i)_j) \\ &\leq \sum_{i=0}^{s-1} -(\mathbf{v}_{i+1}-\mathbf{v}_i)^T f(\mathbf{v}_i) + (\mathbf{v}_{i+1}-\mathbf{v}_i)^2 =: J_1+J_2+J_3, \end{aligned} \tag{3.4.8}$$

where

$$\begin{aligned} J_1 &= \sum_{i=0}^{s-1}\left(1-\frac{1}{d_{i+1}\tau}\right)(\mathbf{v}_{i+1}-\mathbf{v}_i)^2, \qquad J_2 = \sum_{i=0}^{s-1}(\mathbf{v}_{i+1}-\mathbf{v}_i)^T D\mathbf{v}_i, \\ J_3 &= -\sum_{i=0}^{s-1}\frac{\sum_{k=0}^{i-1} p_{i+1,k}(\mathbf{v}_i-\mathbf{v}_k)^T}{d_{i+1}\tau}(\mathbf{v}_{i+1}-\mathbf{v}_i). \end{aligned}$$

It can be easily seen that J_1 is simply quadratic, which will be negative for sufficiently

small τ. By denoting $\mathbf{w}_i = \mathbf{v}_i - \mathbf{v}_{i-1}$, we have

$$\begin{aligned} J_2 &= \sum_{i=0}^{s-1} (\mathbf{v}_{i+1} - \mathbf{v}_i)^T D \left(\frac{\mathbf{v}_{i+1} + \mathbf{v}_i}{2} - \frac{\mathbf{v}_{i+1} - \mathbf{v}_i}{2} \right) \\ &= \sum_{i=0}^{s-1} \frac{1}{2} \left(\mathbf{v}_{i+1}^T D \mathbf{v}_{i+1} - \mathbf{v}_i^T D \mathbf{v}_i - \mathbf{w}_{i+1}^T D \mathbf{w}_{i+1} \right) \\ &= \frac{1}{2} (\mathbf{u}_{n+1}^T D \mathbf{u}_{n+1} - \mathbf{u}_n^T D \mathbf{u}_n) - \frac{1}{2} \sum_{i=1}^{s} \mathbf{w}_i^T D \mathbf{w}_i; \end{aligned}$$

and also

$$\begin{aligned} J_3 &= \sum_{i=0}^{s-1} \frac{1}{d_{i+1}\tau} \sum_{k=0}^{i-1} p_{i+1,k} \left(\sum_{m=k+1}^{i} \mathbf{w}_m^T \right) \mathbf{w}_{i+1} \\ &= \sum_{i=1}^{s} \frac{1}{d_i \tau} \sum_{k=0}^{i-2} \sum_{m=k+1}^{i-1} p_{ik} \mathbf{w}_m^T \mathbf{w}_i = \sum_{i=1}^{s} \frac{1}{d_i \tau} \sum_{m=1}^{i-1} \sum_{k=0}^{m-1} p_{ik} \mathbf{w}_m^T \mathbf{w}_i. \end{aligned}$$

Combining all these results together, we obtain

$$\begin{aligned} E_{n+1} - E_n &= \sum_{j=1}^{N} F((\mathbf{u}_{n+1})_j) - F((\mathbf{u}_n)_j) - \frac{1}{2} (\mathbf{u}_{n+1}^T D \mathbf{u}_{n+1} - \mathbf{u}_n^T D \mathbf{u}_n) \\ &\le \sum_{i=1}^{s} (1 - \frac{1}{d_i \tau}) \mathbf{w}_i^2 - \sum_{i=1}^{s} \frac{1}{d_i \tau} \sum_{m=1}^{i-1} \sum_{k=0}^{m-1} p_{ik} \mathbf{w}_m^T \mathbf{w}_i - \frac{1}{2} \sum_{i=1}^{s} \mathbf{w}_i^T D \mathbf{w}_i \qquad (3.4.9) \\ &= \sum_{i=1}^{s} \mathbf{w}_i^2 - \frac{1}{\tau} \sum_{m,i=1}^{s} \mathbf{w}_m^T \Phi_{mi} \mathbf{w}_i - \frac{1}{2} \sum_{i=1}^{s} \mathbf{w}_i^T D \mathbf{w}_i, \end{aligned}$$

where we have defined an upper triangle matrix Φ by (notice that $\sum_{k=0}^{i-1} p_{ik} = 1$)

$$\Phi_{ij} = \sum_{k=0}^{i-1} p_{jk}/d_j, \qquad i \le j. \qquad (3.4.10)$$

Consider the energy discriminant Δ_E defined by (3.4.6). Recall that we dropped the ϵ scale in the very beginning. If we keep ϵ in the derivation, the change of the energy becomes

$$E_{n+1} - E_n \le \sum_{i=1}^{s} \mathbf{w}_i^2 - \frac{\epsilon}{\tau} \sum_{i,j=1}^{s} \mathbf{w}_i^T \Phi_{ij} \mathbf{w}_j - \frac{\epsilon^2}{2} \sum_{i=1}^{s} \mathbf{w}_i^T D \mathbf{w}_i. \qquad (3.4.11)$$

If Δ_E is positive-definite and λ is the smallest eigenvalue of Δ_E, then we have

$$\sum_{i,j=1}^{s} \mathbf{w}_i^T \Phi_{ij} \mathbf{w}_j = \sum_{i,j=1}^{s} \mathbf{w}_i^T \Delta_{Eij} \mathbf{w}_j \geq \lambda \sum_{i=1}^{s} \mathbf{w}_i^2. \tag{3.4.12}$$

It follows from (3.3.5) that

$$-\sum_{i=1}^{s} \mathbf{w}_i^T D\mathbf{w}_i \leq 4h^{-2} \sum_{i=1}^{s} \mathbf{w}_i^2. \tag{3.4.13}$$

Thus, by (3.4.11), to make sure the energy dissipation we only need

$$1 - \frac{\epsilon\lambda}{\tau} + \frac{2\epsilon^2}{h^2} \leq 0, \tag{3.4.14}$$

which is true under the assumption (3.4.7). □

3.5 Some energy plus MBP RK methods

In this section we present some RK2, RK3 and RK4 methods which are maximum bound preserving and energy dissipation. We will also give an RK3 method which is MBP but does not satisfy our condition for the energy dissipation law. All examples in this section are existing schemes, and the special 5-stage 4th-order example comes from [30].

3.5.1 An RK2 satisfying energy-dissipation and MBP

The Butcher Tableau is as follows:

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

The corresponding (3.4.2) form is given by

$$\mathbf{v}_0 = \mathbf{u}_n,$$

$$\mathbf{v}_1 = \mathbf{v}_0 + \tau G(\mathbf{v}_0),$$

$$\mathbf{v}_2 = \mathbf{v}_0 + \frac{\tau}{2}G(\mathbf{v}_0) + \frac{\tau}{2}G(\mathbf{v}_1) = \frac{1}{2}\mathbf{v}_0 + \frac{1}{2}\mathbf{v}_1 + \frac{\tau}{2}G(\mathbf{v}_1).$$

It follows from the theory in Section 3.3 the scheme is MBP.

The energy form coincides with (3.4.5) and the energy discriminant is

$$\Phi = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}, \quad \Delta_E = \frac{1}{2}\left(\Phi + \Phi^T\right) = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{pmatrix}.$$

Note that Δ_E is positive-definite and the smallest eigenvalue of Δ_E is $\frac{1}{2}(3-\sqrt{2})$. Hence with suitably small time-step, this MBP-RK2 scheme preserves both maximum bound and the energy dissipation law.

3.5.2 An RK3 satisfying energy-dissipation and MBP

Consider the Butcher tableau:

0			
1	1		
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	
	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{3}$

The corresponding (3.4.2) form is given by

$$\begin{aligned}
\mathbf{v}_0 &= \mathbf{u}_n, \\
\mathbf{v}_1 &= \mathbf{v}_0 + \tau G(\mathbf{v}_0), \\
\mathbf{v}_2 &= \frac{3}{4}\mathbf{v}_0 + \frac{1}{4}\mathbf{v}_1 + \frac{\tau}{4}G(\mathbf{v}_1), \\
\mathbf{v}_3 &= \frac{1}{3}\mathbf{v}_0 + \frac{2}{3}\mathbf{v}_2 + \frac{2\tau}{3}G(\mathbf{v}_2).
\end{aligned}$$

The energy discriminant reads

$$\Phi = \begin{pmatrix} 1 & 3 & \frac{1}{2} \\ 0 & 4 & \frac{1}{2} \\ 0 & 0 & \frac{3}{2} \end{pmatrix}, \qquad \Delta_E = \frac{1}{2}\left(\Phi + \Phi^T\right) = \begin{pmatrix} 1 & \frac{3}{2} & \frac{1}{4} \\ \frac{3}{2} & 4 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{3}{2} \end{pmatrix}.$$

Note that Δ_E is positive-definite and the smallest eigenvalue of Δ_E is about 0.362228. Again based on the theory in Section 3.4, with sufficiently small time-step, this MBP-RK3 scheme preserves both maximum bound and the energy dissipation law.

3.5.3 An RK3 satisfying MBP but not sure energy-dissipation

Consider the Butcher tableau:

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ 2 & 1 & 1 & \\ \hline & \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \end{array}$$

The corresponding (3.4.2) form is given by

$$\begin{aligned}
\mathbf{v}_0 &= \mathbf{u}_n, \\
\mathbf{v}_1 &= \mathbf{v}_0 + \tau G(\mathbf{v}_0), \\
\mathbf{v}_2 &= \mathbf{v}_1 + \tau G(\mathbf{v}_1), \\
\mathbf{v}_3 &= \frac{1}{3}\mathbf{v}_0 + \frac{1}{2}\mathbf{v}_1 + \frac{1}{6}\mathbf{v}_2 + \frac{\tau}{6}G(\mathbf{v}_2).
\end{aligned}$$

However, It can be shown that the energy discriminant is not positive-definite in this case. Note

$$\Phi = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 5 \\ 0 & 0 & 6 \end{pmatrix}, \quad \Delta_E = \frac{1}{2}\left(\Phi + \Phi^T\right) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & \frac{5}{2} \\ 1 & \frac{5}{2} & 6 \end{pmatrix}.$$

The smallest eigenvalue of Δ_E is $\frac{1}{2}(7-3\sqrt{6}) \approx -0.174$. Thus, this scheme is not guaranteed to decrease the energy by our approach.

3.5.4 An 5-stage RK4 satisfying MBP and energy-dissipation

It is well known that there is no 4-stage 4th-order SSP-RK4. We then just consider the following 5-stage RK scheme:

$$\begin{aligned}
\mathbf{v}_0 &= \mathbf{u}_n, \\
\mathbf{v}_1 &= \mathbf{v}_0 + d_1\tau G(\mathbf{v}_0), \\
\mathbf{v}_2 &= p_{20}\mathbf{v}_0 + p_{21}\mathbf{v}_1 + d_2\tau G(\mathbf{v}_1), \\
\mathbf{v}_3 &= p_{30}\mathbf{v}_0 + p_{32}\mathbf{v}_2 + d_3\tau G(\mathbf{v}_2),
\end{aligned}$$

$$\mathbf{v}_4 = p_{40}\mathbf{v}_0 + p_{43}\mathbf{v}_3 + d_4\tau G(\mathbf{v}_3),$$

$$\mathbf{v}_5 = p_{52}\mathbf{v}_2 + p_{53}\mathbf{v}_3 + d_{53}\tau G(\mathbf{v}_3) + p_{54}\mathbf{v}_4 + d_{54}\tau G(\mathbf{v}_4),$$

where

$$d_1 = 0.391752226571890, \quad p_{20} = 0.444370493651235,$$

$$p_{21} = 0.555629506348765, \quad d_2 = 0.368410593050371,$$

$$p_{30} = 0.620101851488403, \quad p_{32} = 0.379898148511597,$$

$$d_3 = 0.251891774271694, \quad p_{40} = 0.178079954393132,$$

$$p_{43} = 0.821920045606868, \quad d_4 = 0.544974750228521,$$

$$p_{52} = 0.517231671970585, \quad p_{53} = 0.096059710526147,$$

$$d_{53} = 0.063692468666290, \quad p_{54} = 0.386708617503269,$$

$$d_{54} = 0.226007483236906.$$

Using the theory of Section 3.3, it is known that this scheme satisfies MBP. In order to obtain the energy form, we rewrite the last line as

$$\begin{aligned}\mathbf{v}_5 &= p_{52}\mathbf{v}_2 + p_{53}\mathbf{v}_3 + d_{53}\frac{\mathbf{v}_4 - p_{40}\mathbf{v}_0 - p_{43}\mathbf{v}_3}{d_4} + p_{54}\mathbf{v}_4 + d_{54}\tau G(\mathbf{v}_4)\\ &= -\frac{d_{53}p_{40}}{d_4}\mathbf{v}_0 + p_{52}\mathbf{v}_2 + \left(p_{53} - \frac{d_{53}p_{43}}{d_4}\right)\mathbf{v}_3 + \left(p_{54} + \frac{d_{53}}{d_4}\right)\mathbf{v}_4 + d_{54}\tau G(\mathbf{v}_4).\end{aligned}$$

Thus the energy discriminant is

$$\Phi = \begin{pmatrix} \frac{1}{d_1} & \frac{p_{20}}{d_2} & \frac{p_{30}}{d_3} & \frac{p_{40}}{d_4} & -\frac{d_{53}p_{40}}{d_4} \\ 0 & \frac{1}{d_2} & \frac{p_{30}}{d_3} & \frac{p_{40}}{d_4} & -\frac{d_{53}p_{40}}{d_4} \\ 0 & 0 & \frac{1}{d_3} & \frac{p_{40}}{d_4} & p_{52} - \frac{d_{53}p_{40}}{d_4} \\ 0 & 0 & 0 & \frac{1}{d_4} & p_{52} + p_{53} - \frac{d_{53}}{d_4} \\ 0 & 0 & 0 & 0 & \frac{1}{d_5} \end{pmatrix}, \quad \Delta_E = \frac{1}{2}\left(\Phi + \Phi^T\right).$$

The smallest eigenvalue of Δ_E is about 1.706. Hence, based on the theory of Section 3.4, then with sufficiently small time-step this 5-stage RK4 preserves both maximum bound and energy dissipation law.

Chapter 4

Energy-decreasing Exponential Time Differencing Runge–Kutta methods for phase-field models

4.1 Introduction

In this work, we will restrict our concentration on three of the most popular phase-field models, namely the Allen–Cahn equation,

$$\frac{\partial u}{\partial t} = \epsilon^2 \Delta u - f(u), \quad x \in \Omega,\ t \in (0, T], \tag{4.1.1}$$

the Cahn–Hilliard equation,

$$\frac{\partial u}{\partial t} = \Delta(-\epsilon^2 \Delta u + f(u)), \quad x \in \Omega,\ t \in (0, T], \tag{4.1.2}$$

and the thin film model, which is also named the molecular beam epitaxy (MBE) model,

$$\frac{\partial u}{\partial t} = -\epsilon^2 \Delta^2 u + \nabla \cdot f(\nabla u), \quad x \in \Omega,\ t \in (0, T], \tag{4.1.3}$$

with the initial value

$$u(x,0) = u_0(x), \quad x \in \bar{\Omega}.$$

Here Ω is a bounded domain in R^d $(d = 1, 2, 3)$ and T is a finite time. For simplicity, we impose periodic boundary conditions or homogeneous Neumann boundary conditions for all these equations. In the Allen–Cahn model (4.1.1) and the Cahn–Hilliard model (4.1.2), the solution $u(x,t)$ describes the concentration of two crystal orientations of the same material. In the phase model, $u = 1$ represents one orientation and $u = -1$ represents the other and the parameter ϵ measures the interfacial width, which is small compared to the characteristic length of the laboratory scale. In the thin film model (4.1.3), the function u is the scaled height of epitaxial growth thin films in a co-moving frame. It is well-known that these models satisfy the energy dissipation law, since all of them can be viewed as the gradient flows with the following energy functionals respectively:

$$E(u) = \int_\Omega \left(\frac{\epsilon^2}{2} |\nabla u|^2 + F(u) \right) dx \tag{4.1.4}$$

in L^2 for the Allen–Cahn equation,

$$E(u) = \int_\Omega \left(\frac{\epsilon^2}{2} |\nabla u|^2 + F(u) \right) dx \tag{4.1.5}$$

in H^{-1} for the Cahn–Hilliard equation, and

$$E(u) = \int_\Omega \left(\frac{\epsilon^2}{2} |\Delta u|^2 + F(\nabla u) \right) dx \tag{4.1.6}$$

for the thin film model both with and without slope selection, where $F(u)$ is a given energy potential satisfying $F'(u) = f(u)$. The nonlinear term can be taken as

$$F(u) = \frac{1}{4}(u^2 - 1)^2, \quad f(u) = u^3 - u, \tag{4.1.7}$$

for the Allen–Cahn equation and the Cahn–Hilliard equation as in most of the literature. For the MBE model with slope selection,

$$F(\nabla u) = \frac{1}{4}(|\nabla u|^2 - 1)^2, \quad f(\nabla u) = (|\nabla u|^2 - 1)\nabla u, \tag{4.1.8}$$

and for the MBE model without slope selection,

$$F(\nabla u) = -\frac{1}{2}\ln(|\nabla u|^2 + 1), \quad f(\nabla u) = -\frac{\nabla u}{1 + |\nabla u|^2}. \tag{4.1.9}$$

These gradient flows share the following general form:

$$u_t = G(-\epsilon^2 Du + f(u)), \quad (x, t) \in \Omega \times [0, T] \tag{4.1.10}$$

where G and D are negative and dissipative operators, and the eigenvalues of D are also eigenvalues of G. (Notice that in the MBE model without slope selection, the nonlinear term $f(u)$ is different from f above). It is easy to see that

- for the Allen–Cahn equation, $G = -1, D = \Delta, f(u) = u^3 - u$;
- for the Cahn–Hilliard equation, $G = \Delta, D = \Delta, f(u) = u^3 - u$;
- for the MBE model without slope selection, $G = -1, D = -\Delta^2, f(u) = -\nabla\cdot(\frac{\nabla u}{1+|\nabla u|^2})$.

As these equations involve the perturbed (i.e., the scaling coefficient $\epsilon^2 \ll 1$) Laplacian or biharmonic operators and strong nonlinearities, it is difficult to design an efficient time discretization scheme which is able to resolve dynamics and steady states of the corresponding phase-field models. In addition, it is also a challenging issue to guarantee the energy dissipation which is intrinsic to all these models for numerical approximations. Numerical evidence has shown that non-physical oscillations may happen when the nonlinear energy stability is violated. Consequently, a satisfactory numerical strategy needs to balance accuracy, efficiency and nonlinear stability of the solution.

In our work, we prove the energy dissipation law unconditionally guaranteed by the ETDRK2 scheme for a class of gradient flows. Thus, the ETDRK2 method becomes the first second-order linear scheme which decreases the original energy of the gradient flows. For the Allen–Cahn equation specially, it is already known that the ETDRK2 preserves the maximum bound principle (MBP), thus the ETDRK2 becomes the first second-order scheme proven to unconditionally preserve both the MBP and energy dissipation properties.

This chapter is organized as follows. In Section 4.2, we provide the detailed proof of the main theorem, as well as some discussions on assumptions. Section 4.3 presents some numerical experiments to illustrate the behavior of ETDRK2 solutions for different gradient flows and its properties.

4.2 Exponential time differencing Runge–Kutta methods

In this section we first introduce the second-order exponential time differencing Runge–Kutta methods (ETDRK2) and then prove that the discrete energy decreases. We consider the general form of gradient flows

$$u_t = G(-\epsilon^2 Du + f(u)), \quad (x,t) \in \Omega \times [0,T] \tag{4.2.1}$$

where G and D are negative and dissipative operators, the eigenvalues of D are also eigenvalues of G, and the function f satisfies the Lipschitz condition with the Lipschitz constant β, i.e. $\forall u, v$, we have $|f(u) - f(v)| \leq \beta |u - v|$.

The energy of the gradient flow is

$$E(u) = \int_\Omega \left(\frac{\epsilon^2}{2} |D_{1/2} u|^2 + F(u) \right) \, dx, \tag{4.2.2}$$

where $D_{1/2} D_{1/2}^* = D_{1/2}^* D_{1/2} = -D$ and $F' = f$. Consider the natural splitting of the energy $E(u) = E_l - E_n$ with

$$E_l(u) = \int_\Omega \left(\frac{\epsilon^2}{2} |D_{1/2} u|^2 + \frac{\beta}{2} |u|^2 \right) \, dx,$$
$$E_n(u) = \int_\Omega \left(-\frac{\epsilon^2}{2} F(u) + \frac{\beta}{2} |u|^2 \right) \, dx.$$

In the ETDRK, E_l is treated implicitly and E_n is treated explicitly, which leads to a linearly implicit scheme. From this perspective, the gradient flow can also be written as

$$u_t = G(Lu - g(u)), \tag{4.2.3}$$

where $L = \beta I - \epsilon^2 D$ and $g = \beta I - f$.

The key idea of ETDRK is to consider Duhamel's principle for the gradient flow

$$u(x,t) = e^{GL(t-t_0)}u(x,t_0) - e^{GL(t-t_0)}\int_{t_0}^{t} e^{-GL(s-t_0)}Gg(u(x,s))ds \tag{4.2.4}$$

and approximate the function $g(u)$ in the integral to get enough accuracy. For example, assuming that we have discretized the time and want to solve u_{n+1} from u_n, the easiest way is to approximate the function by a constant $g(u_n)$ which leads to the first-order ETD (ETD1) scheme:

$$u_{n+1} = e^{\tau GL}u_n + \left(I - e^{\tau GL}\right)L^{-1}g(u_n). \tag{4.2.5}$$

The energy dissipation law and MBP for the Allen–Cahn equation specially have been proved, see, e.g. [13]. Hence, by linearly approximating $g(u)$ based on u_n and the solution of ETD, for the gradient flow (4.1.10), the second order ETDRK (ETDRK2) reads

$$\begin{aligned} v &= e^{\tau GL}u_n + \left(I - e^{\tau GL}\right)L^{-1}g(u_n), \\ u_{n+1} &= v - \frac{1}{\tau}\left(e^{\tau GL} - I - \tau GL\right)(GL)^{-2}(Gg(v) - Gg(u_n)), \end{aligned} \tag{4.2.6}$$

where τ is the time step, $L = \beta I - \epsilon^2 D$ and $g = \beta I - f$.

Remark 4.2.1. *Here, β serves as a stabilization to enhance the dissipation of linear part, so as to bound the Lipschitz growing nonlinear term and derive a monotonic function in the analysis. As in [12], this stabilization is necessary to guarantee the MBP for Allen–Cahn equations. [41] provides a numerical evidence to illustrate that this stabilization can improve the numerical performance significantly.*

4.2.1 Main theorem

Before we introduce the theorem and its proof, we start with a useful lemma.

Lemma 2.1. *Consider the positive-definite operator $L = \beta I - \epsilon^2 \Delta$ and let f be an analytic function defined on the spectrum of L, i.e. the values $\{f(\lambda_i)\}_{i\in\mathcal{N}}$ exist, where $\{\lambda_i\}_{i\in\mathcal{N}}$ are the eigenvalues of L. Then, the eigenvalues of $f(L)$ are $\{f(\lambda_i)\}_{i\in\mathcal{N}}$.*

Furthermore, if f is a positive function, then $f(L)$ is also a positive-definite operator.

Theorem 2.1. *For the gradient flow*

$$u_t = G(-\epsilon^2 Du + f(u)), \quad (x,t) \in \Omega \times [0,T] \tag{4.2.7}$$

where G and D are negative and dissipative operators, the eigenvalues of D are also eigenvalues of G, and the function f satisfies the Lipschitz condition, the second-order ETDRK (ETDRK2) unconditionally decreases the energy.

Proof. We only need to calculate the difference of the energy and prove that it is non-positive. For simplicity we denote $(u,v) = \int_\Omega uv\ dx$.

Since the function f is Lipschitz continuous, for the nonlinear part in the energy we have

$$\begin{aligned}(F(v) - F(u_n), 1) &\le (v - u_n, f(u_n)) + \frac{\beta}{2}(v - u_n, v - u_n) \\ &= -(v - u_n, g(u_n)) + \beta(v - u_n, u_n) + \frac{\beta}{2}(v - u_n, v - u_n).\end{aligned} \tag{4.2.8}$$

For the first part in the energy we have

$$\begin{aligned}\frac{\epsilon^2}{2}\left(\int_\Omega |D_{1/2}v|^2 - |D_{1/2}u_n|^2 dx\right) &= -\frac{\epsilon^2}{2}\left((v, Dv) - (u_n, Du_n)\right) \\ &= -\epsilon^2(v - u_n, Dv) + \frac{\epsilon^2}{2}(v - u_n, D(v - u_n)) \\ &= (v - u_n, Lv) - \beta(v - u_n, v) + \frac{\epsilon^2}{2}(v - u_n, D(v - u_n)),\end{aligned} \tag{4.2.9}$$

where we use the identity

$$[a, a] - [b, b] = 2[a - b, a] - [a - b, a - b]$$

for all a, b and inner products $[\cdot, \cdot]$. Therefore, combining these two parts we derive

$$\begin{aligned}E(v) - E(u_n) &\le (v - u_n, Lv - g(u_n)) - \frac{\beta}{2}(v - u_n, v - u_n) + \frac{\epsilon^2}{2}(v - u_n, D(v - u_n)) \\ &\le (v - u_n, Lv - g(u_n)).\end{aligned} \tag{4.2.10}$$

According to the scheme we obtain

$$\begin{aligned}E(v) - E(u_n) &\le (v - u_n, Lv - g(u_n)) \\ &= \left(v - u_n, Lv - L\left(I - e^{\tau GL}\right)^{-1}\left(v - e^{\tau GL}u_n\right)\right) \\ &= \left(v - u_n, \left[L - L\left(I - e^{\tau GL}\right)^{-1}\right](v - u_n)\right) \\ &= \left(v - u_n, \Delta_1(v - u_n)\right),\end{aligned} \tag{4.2.11}$$

where $\Delta_1 := L - L\left(I - e^{\tau GL}\right)^{-1}$.

Since the function $y_1 = x - \frac{x}{1-e^x}$ is non-negative for all x and $\tau G\Delta_1 = y_1(\tau GL)$, we derive that the operator Δ_1 is non-positive, i.e. the energy is decreasing from u_n to v.

Similarly, we compute the energy difference between other two stages:

$$\begin{aligned}
E(u_{n+1}) - E(v) &\le (u_{n+1} - v, Lu_{n+1} - g(v)) \\
&= \left(u_{n+1} - v, Lu_{n+1} - g(u_n) + \tau GL^2 \left(e^{\tau GL} - I - \tau GL\right)^{-1} (u_{n+1} - v)\right) \\
&= \left(u_{n+1} - v, \Delta_1(v - u_n) + \left[L + \tau GL^2 \left(e^{\tau GL} - I - \tau GL\right)^{-1}\right] (u_{n+1} - v)\right) \\
&:= (u_{n+1} - v, \Delta_1(v - u_n)) + (u_{n+1} - v, \Delta_2(u_{n+1} - v)),
\end{aligned} \tag{4.2.12}$$

where $\Delta_2 = L + \tau GL^2 \left(e^{\tau GL} - I - \tau GL\right)^{-1}$.

In conclusion,

$$\begin{aligned}
&E(u_{n+1}) - E(u_n) \\
\le& (v - u_n, \Delta_1(v - u_n)) + (u_{n+1} - v, \Delta_1(v - u_n)) + (u_{n+1} - v, \Delta_2(u_{n+1} - v)) \\
\le& \frac{1}{2}(v - u_n, \Delta_1(v - u_n)) + (u_{n+1} - v, (\Delta_2 - \frac{1}{2}\Delta_1)(u_{n+1} - v)) \\
&+ \frac{1}{2}((v - u_n, \Delta_1(v - u_n)) + 2(u_{n+1} - v, \Delta_1(v - u_n)) + (u_{n+1} - v, \Delta_1(u_{n+1} - v))) \\
=& \frac{1}{2}(v - u_n, \Delta_1(v - u_n)) + (u_{n+1} - v, (\Delta_2 - \frac{1}{2}\Delta_1)(u_{n+1} - v)) + \frac{1}{2}(u_{n+1} - u_n, \Delta_1(u_{n+1} - u_n)).
\end{aligned} \tag{4.2.13}$$

Among these three terms above, the first and third terms are non-positive since Δ_1 is negative-definite. For the second one we only need to consider the function

$$\begin{aligned}
y_2 &= x + \frac{x^2}{e^x - 1 - x} - \frac{1}{2}y_1 \\
&= \frac{1}{2}x + \frac{x^2}{e^x - 1 - x} + \frac{x}{2(1 - e^x)} \\
&= \frac{x(e^x(e^x + x - 3) + 2)}{(e^x - 1)(e^x - 1 - x)}.
\end{aligned} \tag{4.2.14}$$

Since the functions $\frac{x}{e^x - 1}$ and $(e^x - 1 - x)$ are both always non-negative and $\frac{d}{dx}(e^x(e^x +$

$x-3+2e^{-x})) = e^x(2e^x+x-2) = 0$ only happens at $x = 0$, where $(e^x(e^x+x-3)+2)$ reaches its minimum 0, i.e. it is always non-negative. Notice that $\tau G(\Delta_2 - \frac{1}{2}\Delta_1) = y_2(\tau GL)$, so we can derive that $(\Delta_2 - \frac{1}{2}\Delta_1)$ is negative-definite, which indicates that the second term is also negative so that the energy is decreasing from u_n to u_{n+1}.

Thus, we prove that

$$E(u_{n+1}) - E(u_n) \leq 0.$$

□

Remark 4.2.2. *The convergence of ETD1 and ETDRK2 for Allen–Cahn equations has been well studied, see, e.g. [12].*

Remark 4.2.3. *Notice that we have not considered the spatial discretization. If we consider the finite difference method, then the proof also holds as long as we re-define the inner product as* $(u, v) = u^T v$.

Remark 4.2.4. *For the MBE model without slope selection we may notice that the nonlinear term does not directly satisfy the Lipschitz condition as a function of* u. *However, it is Lipschitz continuous as a function of* ∇u, *and meanwhile the convex splitting of the energy is also different (see [32] for more details), but all analysis in the proof can be carried out in the similar way. To keep the presentation short, we omit the proof.*

4.2.2 Discussions on assumptions

As we have put our main theorem in a very general framework, including extensive phase field gradient flows, it is necessary to supplement extra discussions when we apply it to a specific phase field model. In this section we discuss the restrictions of the theorem and illustrate that such assumptions are reasonable.

For many gradient flows, it is not difficult to satisfy the restriction that the eigenvalues of D are also eigenvalues of G. Both of them are usually some power of the Laplacian operator and G is a lower one. Besides, as long as these two operators are dissipative and satisfy the assumption, the conclusion still holds. In fact, we did not make use of more information about the Laplacian operator in the proof. Hence, our framework can also works for nonlocal models [14].

As for the Lipschitz condition, it is satisfied by many physically relevant potentials by restricting them to be quadratic for $|u| > M$, given some constant M big enough. This assumption is automatically satisfied for the Allen-Cahn equation because of the maximum principle. For the numerical approach, the finite difference method (second-order accuracy in space) applied to the spatial direction guarantees the discrete maximal bound principle (see [13]). For the Cahn-Hilliard equation, it has not been proven that the numerical solution has the maximal bound so that the nonlinear term does not satisfy the Lipschitz condition directly. One common practice is to modify the potential by cutting off it in $|x| > M$ and replacing it with a quadratic function smoothly connected to the inner part

for some M big enough

$$\hat{F}(u) = \begin{cases} \frac{3M^2-1}{2}u^2 - 2M^3u + \frac{1}{4}(3M^4+1), & u > M \\ \frac{1}{4}(u^2-1)^2, & |u| \le M \\ \frac{3M^2-1}{2}u^2 + 2M^3u + \frac{1}{4}(3M^4+1), & u < -M \end{cases}, \tag{4.2.15}$$

and the corresponding $f(u)$ is replaced by

$$\hat{f}(u) = \hat{F}'(u) = \begin{cases} (3M^2-1)u - 2M^3, & u > M \\ (u^2-1)u, & |u| \le M \\ (3M^2-1)u + 2M^3, & u < -M \end{cases} \tag{4.2.16}$$

Other methods can also be found in [10, 11, 50].

For the MBE model without slope selection, the Lipschitz condition is automatically satisfied when the nonlinear term is viewed as a function of ∇u. For the original functions f and F in 4.1.9, we have

$$\|\partial^2_{\nabla u} F(\nabla u)\|_2 = \frac{1}{(|\nabla u|^2+1)} \le 1, \tag{2.1}$$

(i.e. the eigenvalues of the matrix are smaller than 1), so the energy of the nonlinear term still satisfies the following inequality

$$(F(\nabla v) - F(\nabla u_n), 1) \le (\nabla(v-u_n), f(\nabla u_n)) + \frac{\beta}{2}(\nabla(v-u_n), \nabla(v-u_n)). \tag{4.2.17}$$

All other steps can be carried out in the same way.

4.3 Numerical Experiments

In this section we carry out some numerical experiments to illustrate the convergence and energy decay property of ETDRK2 for different phase-field models. We first verify the temporal convergence rates of spectral collocation methods with a smooth initial data for the Allen–Cahn and Cahn–Hilliard equations. Next we check the energy dissipation law for all these examples. Finally we present an adaptive time stepping example which makes full use of the unconditional energy stability and serves as a good application. The 2D domain $\Omega = (0, 2\pi) \times (0, 2\pi)$ will be used in all following examples.

4.3.1 Convergence tests

We consider the Allen–Cahn (4.1.1) and Cahn–Hilliard (4.1.2) equations with the smooth initial data $u_0 = 0.5 \sin x \sin y$ and the periodic boundary condition. To compute the errors and the convergence rate, we set the number of grid points $N = 256$, the interfacial parameter $\epsilon = 0.1$ and the terminal time $T = 0.5$. With these settings we calculate the numerical solutions with various time steps $dt = 0.01/2^{-k}$ with $k = 0, 1, ..., 6$ and calculate the relative errors to get the convergence rate. Since we are applying the spectral method, fast algorithms (FFT) can be used.

First we solve the Allen–Cahn equation. The maximum and L^2 norms are considered to calculate the convergence rates.

dt=0.01	L^∞ **err**	**rate**	L^2 **err**	**rate**
dt	2.060e-05	-	1.156e-06	-
dt/2	5.206e-06	1.9846	2.915e-07	1.9846
dt/4	1.309e-06	1.9923	7.326e-08	1.9923
dt/8	3.280e-07	1.9961	1.836e-08	1.9961
dt/16	8.212e-08	1.9981	4.597e-09	1.9981
dt/32	2.054e-08	1.9991	1.150e-09	1.9991

Table 4.1: ETDRK2 solution errors and convergence rates for the Allen–Cahn equation

Next we solve the Cahn–Hilliard equation with the same initial data and settings. The maximum and L^2 norms are also considered to calculate the convergence rates.

dt=0.01	L^∞ **err**	**rate**	L^2 **err**	**rate**
dt	4.363e-01	-	7.736e-03	-
dt/2	1.211e-01	1.8498	2.206e-03	1.8104
dt/4	3.441e-02	1.8147	6.333e-04	1.8002
dt/8	9.450e-03	1.8644	1.745e-04	1.8595
dt/16	2.507e-03	1.9145	4.635e-05	1.9127
dt/32	6.488e-04	1.9499	1.201e-05	1.9490

Table 4.2: ETDRK2 solution errors and convergence rates for the Cahn–Hilliard equation

It can be observed in both cases that the convergence rate approaches the theoretical value 2 as the time step becomes smaller.

4.3.2 Dynamics and energy evolution of gradient flows

In this section we present some numerical examples to show the dynamics and energy evolution of gradient flows starting with specific and random initial data. For the Allen–Cahn equation we set $\beta = 1$ and for the Cahn–Hilliard equation $\beta = 2$.

We first consider the Allen–Cahn and Cahn–Hilliard equation with $f(u) = u^3 - u$ and $\epsilon = 0.1$ with the smooth initial condition

$$u_0 = 0.05 \sin(x) \sin(y)$$

and the periodic boundary condition. We take the numerical solutions on a 512×512 mesh, the uniform step $\tau = 0.001$ and $T = 8$. The energy curves of these two solutions are also given in Fig 2, which indicate the decreasing energy.

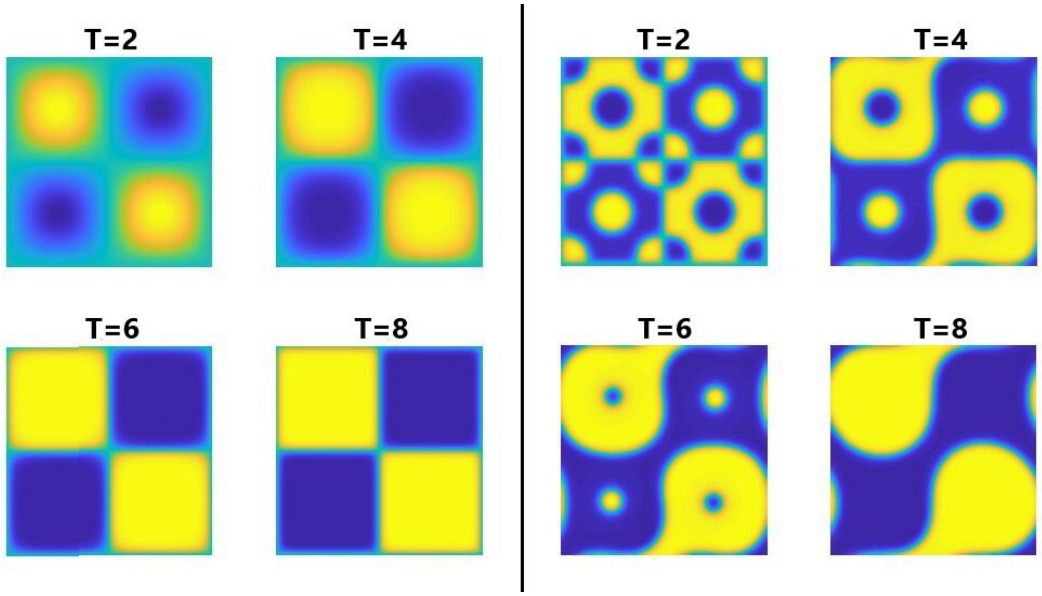

Figure 1: Numerical solutions for the Allen–Cahn (left) and Cahn–Hilliard (right) equations at $T = 2, 4, 6, 8$

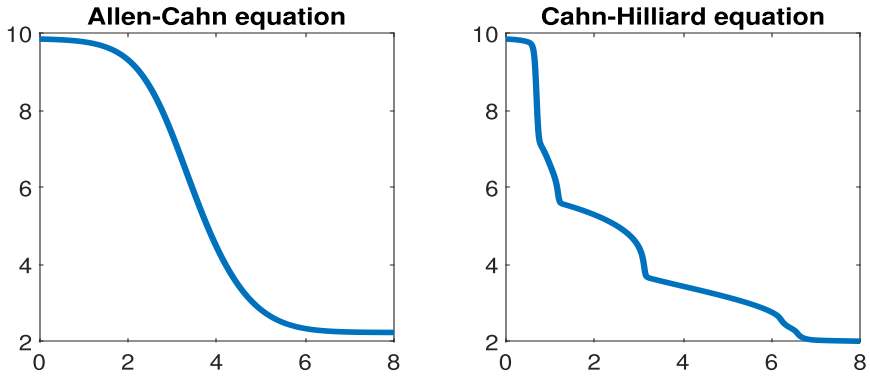

Figure 2: Energy of solutions for the Allen–Cahn and Cahn–Hilliard equations

Next we simulate the thin film models without slope selection. We take the same smooth initial condition, boundary condition and the same physical parameters as the previous example. We also use the same computational grids in space and take the uniform time step $\tau = 0.001$, but the parameter β in the numerical scheme is set to be $1/8$. The snapshots of the numerical solution and the corresponding energy curve are presented in Fig 3.

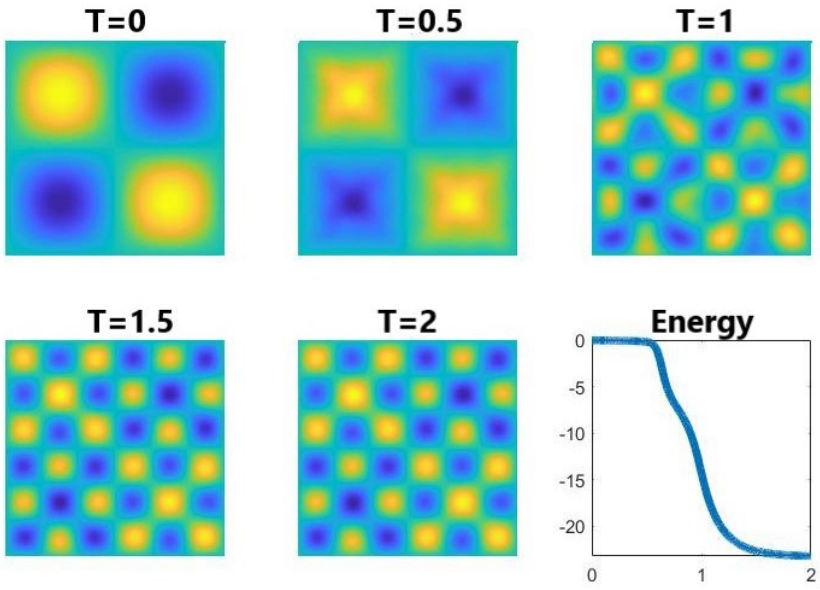

Figure 3: Numerical solutions and energy curve for the MBE model without slope selection

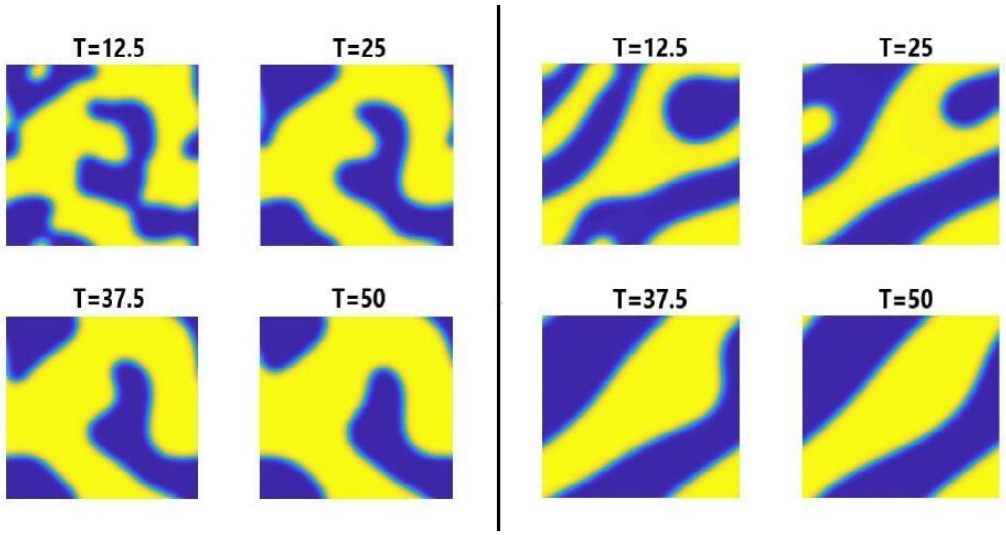

Figure 4: Numerical solutions for the Allen–Cahn (left) and Cahn–Hilliard (right) equations with random initial data

Then we simulate the Allen–Cahn and Cahn–Hilliard equations with a random initial data ranging from -1 to 1 with the same basic setting $N = 512$ and $\epsilon = 0.1$. To see the long-time energy evolution, we take $T = 50$ the time step $dt = 0.01$. The dynamics of the numerical solutions and the energy curves are shown in Fig 4 and 5.

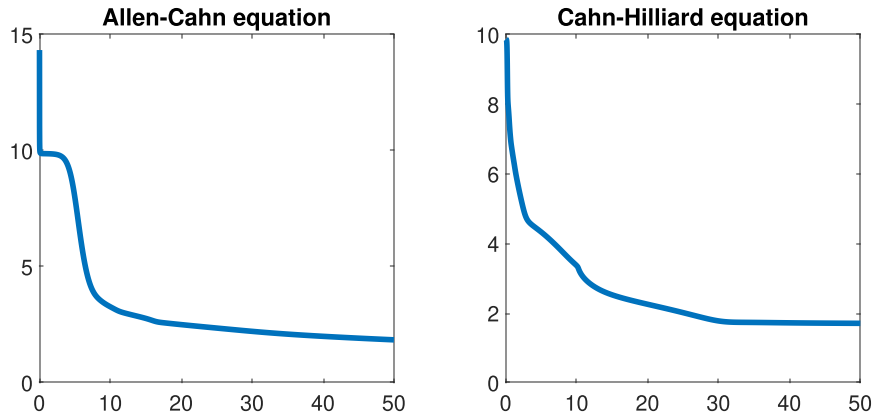

Figure 5: Energy of the Allen–Cahn and Cahn–Hilliard solutions with random initial data

Finally we test the thin film model without slope selection with a random initial data. For this case, we set the parameter $\epsilon = 0.1$, the scheme constant $\beta = 0.5$ and also use $N = 512$. The initial data now is ranging from -0.001 to 0.001 and the boundary condition is also periodic. We take the time step $dt = 0.01$ and $T = 50$ to see the development of the energy. It can be observed from the snapshots of the solution that the number of hills and valleys is decreasing, which is the long time feature of the thin film model.

4.3.3 Adaptive time stepping

As we have presented above, the solution of gradient flows can vary sharply during a very short time, but only slightly elsewhere. One major advantage of unconditional energy stable schemes is that they can be easily used to construct an adaptive time stepping algorithm, in which the time step is only dictated by accuracy rather than by stability as with conditionally stable schemes.

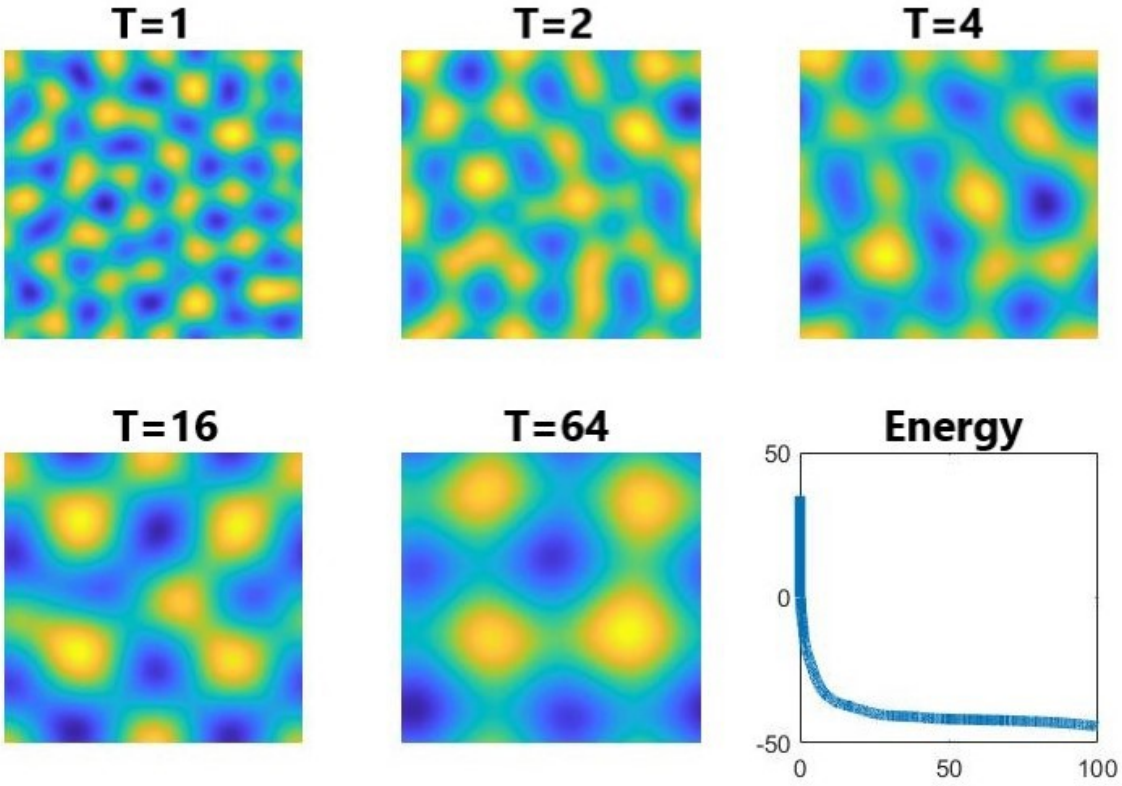

Figure 6: Numerical solutions and energy curve for the thin film model without slope selection with random initial data

For gradient flows, there are several adaptive time stepping strategies. Here we follow the strategy in [40] summarized in the following Algorithm 1, which has been shown to be effective for Allen-Cahn equations. In Step 4 and 6, the time step size which needs to be updated is given by the formula

$$A_{dp}(e,\tau) = \rho\left(\frac{tol}{e}\right)^{1/2}\tau,$$

along with the restriction of the minimum and maximum time steps. In the above formula,

ρ is a default safety coefficient, tol is a reference tolerance to be set, and e is the relative error computed at each time level in Step 3. In our numerical example, we set $\rho = 0.9$ and $tol = 10^{-3}$, and the minimum and maximum time steps are chosen to be 10^{-5} and 10^{-2}, respectively. The initial time step is taken as the minimum time step.

Algorithm 1 Adaptive time stepping procedure

Given: U^n, τ_n

Step 1. Compute U_1^{n+1} by the first-order ETD scheme with τ_n.

Step 2. Compute U_2^{n+1} by the second-order ETDRK scheme with τ_n.

Step 3. Calculate $e_{n+1} = \frac{\|U_1^{n+1} - U_2^{n+1}\|}{\|U_2^{n+1}\|}$.

Step 4. If $e_{n+1} > tol$, recalculate the time step $\tau_n \leftarrow \max\{\tau_{min}, \min\{A_{dp}(e_{n+1}, \tau_n), \tau_{max}\}\}$,

Step 5. goto Step 1.

Step 6. else, update the time step $\tau_{n+1} \leftarrow \max\{\tau_{min}, \min\{A_{dp}(e_{n+1}, \tau_n), \tau_{max}\}\}$.

Step 7. endif

We take the two-dimensional Cahn–Hilliard equation as our example to demonstrate the performance of the adaptive time stepping algorithm. We consider the Cahn-Hilliard equation with the periodic boundary condition and random initial data and take $\epsilon = 0.1$, $N = 512$ and $\beta = 2$. As comparison, we compute two ETDRK2 solutions with a small uniform time step $\tau = 10^{-5}$ and a large uniform time step $\tau = 10^{-2}$ as reference.

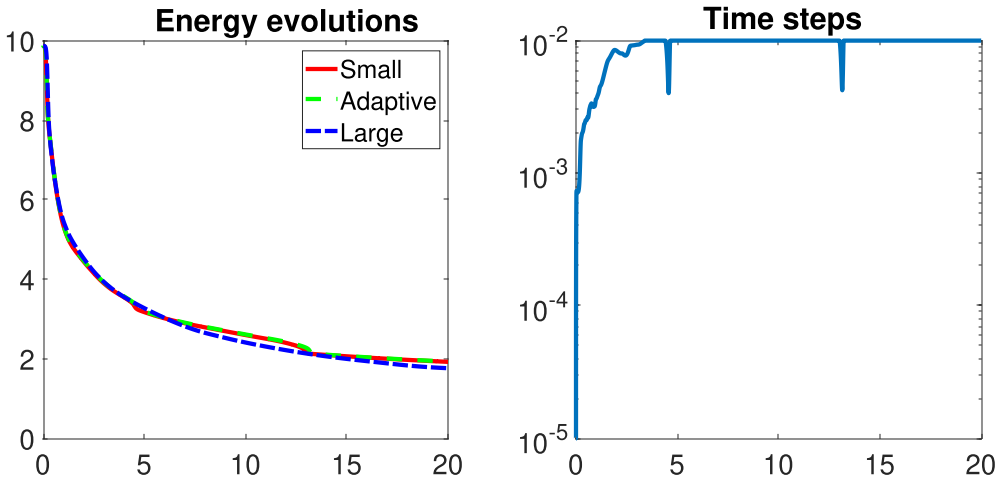

Figure 7: Energy curves among small time steps, adaptive time steps and large time steps and the size of time steps in the adaptive procedure

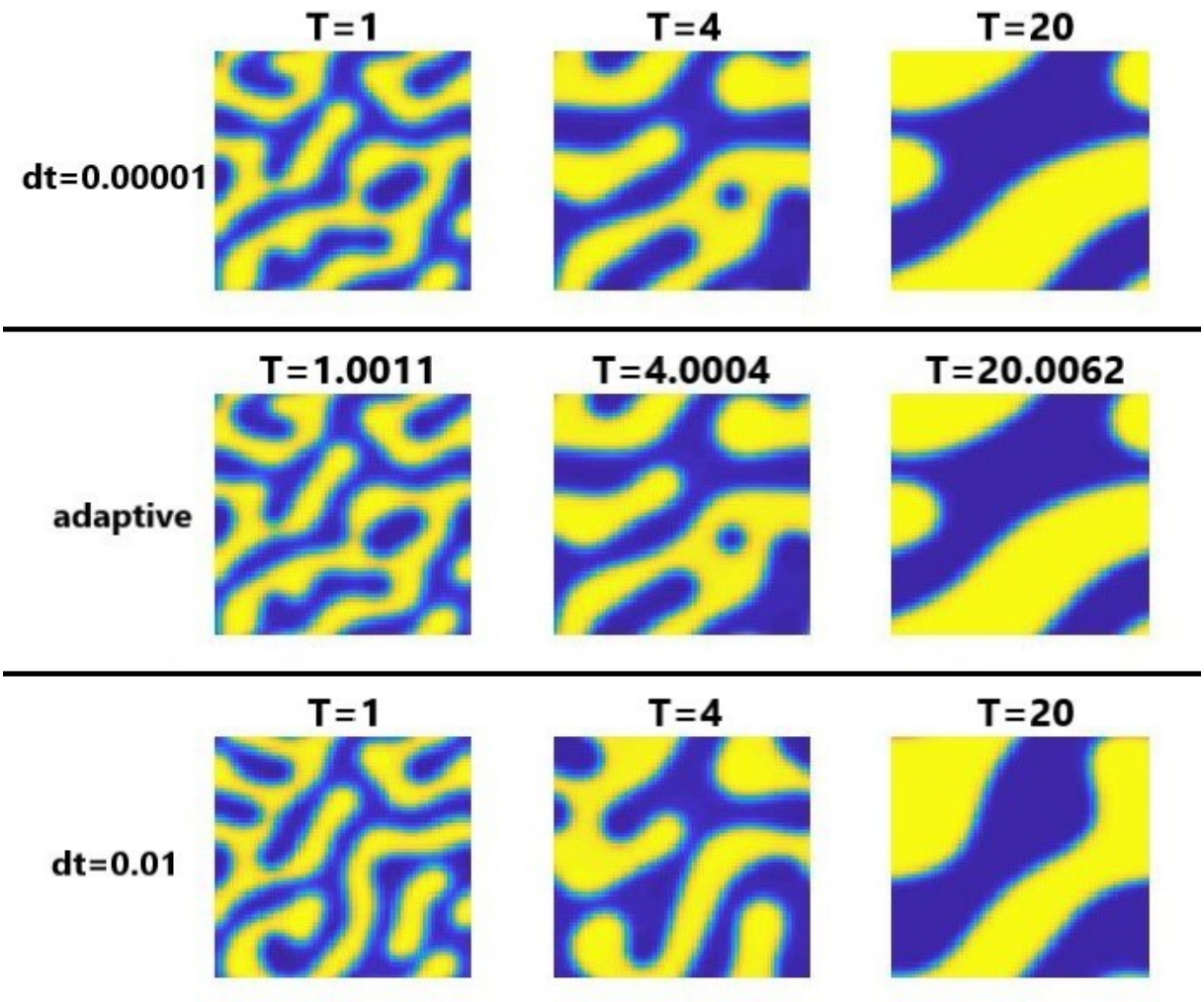

Figure 8: Numerical solutions for the Cahn–Hilliard equation among small time steps, adaptive time steps and large time steps

In the Fig.7 and Fig.8 we show the energy development, the size of time steps and snapshots of phase evolution in the adaptive experiment. It can be observed that the topology of the large time step solution is very different from the topology of the small

time step one, while the adaptive time solution is in good agreement with the reference solution. This is also indicated by the energy evolution. Note also that adaptive time steps change accordingly with the energy evolution. In addition it can be noticed that the adaptive time steps basically lie in the interval $[10^{-3}, 10^{-2}]$, but the result of the adaptive time stepping strategy is rather close to the small step reference solution, which indicates a fast and efficient computation.

Chapter 5

Unconditionally energy-decreasing high-order Implicit-Explicit Runge-Kutta methods for phase-field models with Lipschitz nonlinearity

5.1 Introduction

In this chapter, we also fix our concentration on the following general form for phase-field models:

$$u_t = G(-Du + f(u)), \quad (x,t) \in \Omega \times [0,T] \tag{5.1.1}$$

where Ω is a bounded domain in R^d $(d = 1, 2, 3)$, T is a finite time, both G and D are negative definite and dissipative operators, and the eigenvectors of D are also eigenvectors of G (both of them are usually related to the Laplacian operator). As we have mentioned, for some well-known phase-field equations, we have

- for the Allen–Cahn equation, $G = -1, D = \epsilon^2\Delta, f(u) = u^3 - u$;

$$\frac{\partial u}{\partial t} = \epsilon^2\Delta u - f(u);$$

- for the Cahn–Hilliard equation, $G = \Delta, D = \epsilon^2\Delta, f(u) = u^3 - u$;

$$\frac{\partial u}{\partial t} = \Delta(-\epsilon^2\Delta u + f(u));$$

- for the MBE model without slope selection, $G = -1, D = -\epsilon^2\Delta^2, f(u) = -\nabla \cdot (\frac{\nabla u}{1+|\nabla u|^2})$

$$\frac{\partial u}{\partial t} = -\epsilon^2\Delta^2 u + \nabla \cdot (\frac{\nabla u}{1 + |\nabla u|^2}).$$

We consider the equation with certain initial value

$$u(t_0, x) = u_0(x), \quad x \in \Omega,$$

and impose periodic boundary conditions or homogeneous Neumann boundary conditions for simplicity. It is well-known that these models satisfy the energy dissipation law, since all of them can be viewed as the gradient flows with the following energy functionals respectively:

$$E(u) = \int_\Omega \left(\frac{\epsilon^2}{2}|\nabla u|^2 + F(u)\right) dx \tag{5.1.2}$$

in L^2 for the Allen–Cahn equation, in H^{-1} for the Cahn–Hilliard equation, and

$$E(u) = \int_\Omega \left(\frac{\epsilon^2}{2} |\Delta u|^2 + F(\nabla u) \right) dx \tag{5.1.3}$$

for the thin film model, where $F(u)$ is a given energy potential satisfying $F'(u) = f(u)$.

Due to the perturbed (i.e., the scaling coefficient $\epsilon^2 \ll 1$) Laplacian or biharmonic operators and strong nonlinearities involved in the equations, it is difficult to design an efficient and accurate time discretization scheme which resolve dynamics and steady states of the corresponding phase-field models. Moreover, another challenging issue for numerical approximations is to preserve the energy dissipation law which intrinsically holds for all these models. Numerical evidence has shown that non-physical oscillations may happen when the energy stability is violated. Therefore, a satisfactory numerical strategy needs to balance accuracy, efficiency and nonlinear stability of the solution.

There have been a wide range of studies for the construction of various numerical schemes preserving the energy dissipation law at a discrete level. Some popular and significant implicit time stepping work includes convex splitting methods [20] and Crank-Nicolson type schemes [16, 17]. The deficiency of these methods is the expense of solving a nonlinear system of equations at each time step. While in contrast to fully implicit schemes, implicit-explicit (also named semi-implicit) methods treats the nonlinear term explicitly and the linear term implicitly, and to solve these linearly implicit problems, it only requires solving a linear system of equations at every time step. Such methods may date back to the work of Chen and Shen [7] in the phase-field context, and so far there have been developed many

techniques and skills to design such schemes, see, e.g. [19, 21, 26, 31, 42, 46, 47, 43]. Based on the idea of the invariant energy quadratization (IEQ) method (see, e.g. [53, 54]), Shen et al. [8, 41] proposed the scalar auxiliary variable (SAV) method, which can easily render the unconditional energy decay property. However, the energy considered in these methods is modified and different from the exact original energy. In another direction, exponential time differencing (ETD) methods for the Allen–Cahn equation and other semilinear parabolic equations have attracted much attention. Du et al. [12] has shown that ETD and ETDRK2 schemes unconditionally preserve maximum bound property (MBP) and energy stability (but not the dissipation law). For the thin film model (or MBE model), authors of [32, 51, 36] offer nice results concerning stability analysis and error estimates of the ETD schemes. For more information, stability analysis and more applications of ETD schemes can be found in [9, 15].

In this chapter, we prove the energy dissipation law unconditionally guaranteed by a class of high-order IMEX-RK schemes for gradient flows with Lipschitz nonlinearity. We make use of convex splitting and as long as the conditions for the RK method is satisfied, the scheme unconditionally decreases the original energy of the phase-field equation. In addition, we also give detailed analysis and derive inequalities which constrains the coefficients of splitting and the time step for the Allen–Cahn and Cahn–Hilliard equation respectively. Thus, the IMEX-RK method becomes the first high-order linear one-step scheme which unconditionally decreases the **original energy** of the gradient flows. Furthermore, we estimate the error and prove the error analysis theoretically which indicates

the accuracy.

This chapter is organized as follows. In Section 5.2, we introduce some preliminaries concerning convex splitting and IMEX-RK schemes. Section 5.3 presents our main theorem and several theorems related to Allen–Cahn and Cahn–Hilliard equation and their proof. In Section 5.4 we give the error analysis, to show the accuracy of the method. Section 5.5 offers some IMEX-RK examples to help illustrate the requirements of the theorems.

5.2 Preliminaries: Convex splitting and IMEX-RK

In this section we introduce some preliminaries concerning operator splitting and implicit-explicit (IMEX) Runge-Kutta (RK) schemes. We consider the general form of gradient flows

$$u_t = G(-Du + f(u)), \quad (x,t) \in \Omega \times [0,T] \tag{5.2.1}$$

where G and D are negative and dissipative operators, the eigenvalues of D are also eigenvalues of G, and the function f satisfies the Lipschitz condition with the Lipschitz constant L, i.e. given any v, u, we want

$$(F(v) - F(u), 1) \leq (v - u, f(u)) + \frac{L}{2}(v - u, v - u). \tag{5.2.2}$$

for the Allen–Cahn and Cahn–Hilliard equations and

$$(F(\nabla v) - F(\nabla u), 1) \leq (\nabla(v - u), f(\nabla u)) + \frac{L}{2}(\nabla(v - u), \nabla(v - u)). \tag{5.2.3}$$

for the MBE model without slope selection, where $(\cdot, \cdot)$ represents the spatial inner product.

The energy of the gradient flow is

$$E(u) = \int_\Omega \left(\frac{1}{2}|D_{1/2}u|^2 + F(u) \right) \, dx, \tag{5.2.4}$$

where $D_{1/2}D_{1/2}^* = D_{1/2}^* D_{1/2} = D$ and $F' = f$. Consider the splitting of the energy $E(u) = E_l - E_n$ with

$$E_l(u) = \int_\Omega \left(\frac{1}{2}|D_{1/2}u|^2 + \frac{\alpha}{2}|D_{1/2}u|^2 + \frac{\beta}{2}|u|^2 \right) \, dx,$$
$$E_n(u) = \int_\Omega \left(-F(u) + \frac{\alpha}{2}|D_{1/2}u|^2 + \frac{\beta}{2}|u|^2 \right) \, dx,$$

where both E_l and E_n are convex funcationals. In the computation, E_l is treated implicitly and E_n is treated explicitly, which leads to an implicit scheme but still linear to solve. From this perspective, the gradient flow can be equivalently written as

$$u_t = G(D_s u - f_s(u)), \tag{5.2.5}$$

where $D_s = -(1+\alpha)D + \beta I$ and $f_s = -f - \alpha D + \beta I$.

Here, β serves as the enhancement of the dissipation of the linear part, so as to bound the Lipschitz growing nonlinear term in the analysis. Meanwhile, α is not necessary but in our framework, for certain Runge-Kutta schemes, this term helps preserve the stability. [41] provides a numerical evidence to illustrate that the stabilization can improve the numerical performance significantly.

5.2.1 Implicit-Explicit Runge-Kutta schemes

For the linear term $GD_s u$ in the gradient flow (5.2.5), we consider an s-stage diagonally implicit Runge-Kutta (DIRK) scheme with coefficients $A = (a_{ij})_{s\times s} \in \mathcal{R}^{s\times s}, c, b \in \mathcal{R}^s$, in

the usual Butcher notation. For the nonlinear term $Gf_s(u)$ in (5.2.5), we make use of an s-stage explicit scheme with the same abscissae $\hat{c} = c$ and coefficient $\hat{A} = (\hat{a}_{ij})_{s\times s} \in \mathcal{R}^{s\times s}, \hat{b} \in \mathcal{R}^s$. Thus, the IMEX-RK scheme can be determined by the following Butcher notation

$$\begin{array}{c|ccccc} 0 & 0 & 0 & ... & ... & 0 \\ c_1 & 0 & a_{11} & 0 & ... & 0 \\ c_2 & 0 & a_{21} & a_{22} & ... & 0 \\ ... & 0 & ... & ... & ... & ... \\ c_s & 0 & a_{s1} & a_{s2} & ... & a_{ss} \\ \hline & 0 & b_1 & b_2 & ... & b_s \end{array} \qquad \begin{array}{c|ccccc} 0 & 0 & 0 & 0 & ... & 0 \\ \hat{c}_1 & \hat{a}_{11} & 0 & 0 & ... & 0 \\ \hat{c}_2 & \hat{a}_{21} & \hat{a}_{22} & 0 & ... & 0 \\ ... & ... & ... & ... & ... & 0 \\ \hat{c}_s & \hat{a}_{s1} & \hat{a}_{s2} & ... & \hat{a}_{ss} & 0 \\ \hline & \hat{b}_1 & \hat{b}_2 & ... & \hat{b}_s & 0 \end{array} . \tag{5.2.6}$$

where $c_i = \hat{c}_i = \sum_j a_{ij} = \sum_j \hat{a}_{ij}$. Here we only consider such special IMEX-RK schemes: $b_j = a_{sj}, \hat{b}_j = \hat{a}_{sj}, j = 1, ..., s$, indicating the implicit scheme is stiffly accurate. Besides, we require that the coefficient matrix of the implicit scheme begins with a zero column, i.e. we are considering type ARS (see [2]). In addition, we also require that the matrix $\hat{A}$ is invertible.

Consider a model equation:

$$u_t = \mathcal{L}u + N(u), \tag{5.2.7}$$

where $\mathcal{L}$ represents a linear operator and N indicates a nonlinear one. Applying the IMEX-RK (5.2.6) to the model equation, we derive the following system of equations: (to solve

u_{n+1} starting from u_n)

$$\begin{aligned} v_0 &= u_n, \\ v_i &= v_0 + \tau \left(\sum_{j=1}^{i} a_{ij} \mathcal{L} v_j + \sum_{j=1}^{i} \hat{a}_{ij} N(v_{j-1}) \right), \quad 1 \le i \le s, \\ v_s &= u_{n+1}. \end{aligned} \tag{5.2.8}$$

There have been some existing work related to such IMEX-RK schemes, such as asymtotic behavior, error analysis and studies of different kinds of stability, etc. See, e.g. [2, 3, 5, 49].

5.3 Energy decreasing property

In this section we present our main theorem and its proof.

5.3.1 Main Theorem

Theorem 3.1. *The IMEX-RK scheme (5.2.8) unconditionally decreases the energy of the phase field model (5.2.5) if the following three matrices are positive-definite:*

$$\begin{aligned} H_0 &= (\hat{A})^{-1} E_L, \\ H_1(\beta) &= \beta Q - \frac{L}{2} I, \\ H_2(\alpha) &= \alpha Q + (\hat{A})^{-1} A E_L - \frac{1}{2} E_1, \end{aligned} \tag{5.3.1}$$

where $E_1 = \mathbf{1}_{s\times s}$, $E_L = (\mathbf{1}_{i\ge j})_{s\times s}$ *represents the lower triangle matrix full of the element* 1, $I = (\mathbf{1}_{i=j})_{s\times s}$ *represents the identity matrix, and the determinant matrix* Q *is defined by*

$$Q = \left((\hat{A})^{-1} A - I \right) E_L + I. \tag{5.3.2}$$

In other words, if both Q and H_0 are positive-definite, then we only need to take α and β big enough:

$$\begin{aligned} \beta \geq \beta_0 &= \frac{L}{2\lambda_{min}(Q)}, \\ \alpha \geq \alpha_0 &= -\frac{\lambda_{min}((\hat{A})^{-1}AE_L - \frac{1}{2}E_1)}{\lambda_{min}(Q)}, \end{aligned} \tag{5.3.3}$$

and thus the scheme unconditionally decreases the energy.

Remark 5.3.1. *Here we say a matrix, M is positive-definite when $\frac{1}{2}(M+M^T)$ is positive-definite.*

Remark 5.3.2. *In fact the splitting $\frac{\alpha}{2}|D_{1/2}u|^2$ is evitable for certain Runge–Kutta methods. We may notice that $H_2(0) > 0$ guarantees the positive-definiteness of $H_2(\alpha)$ when Q is positive-definite. However, in order to derive unconditional energy dissipation, the splitting $\frac{\beta}{2}|u|^2$ is, in a sense, necessary. This is because the term has to counter the effect of the Lipschitz nonlinear term. Besides, [41] provides a numerical evidence to illustrate that this stabilization can improve the numerical performance significantly.*

Proof. The system of the IMEX-RK schemes (5.2.8) can formally be written as

$$\begin{pmatrix} v_1 \\ v_2 \\ \cdots \\ v_s \end{pmatrix} = \begin{pmatrix} v_0 \\ v_0 \\ \cdots \\ v_0 \end{pmatrix} + \tau \left(A \begin{pmatrix} \mathcal{L}v_1 \\ \mathcal{L}v_2 \\ \cdots \\ \mathcal{L}v_s \end{pmatrix} + \hat{A} \begin{pmatrix} N(v_0) \\ N(v_1) \\ \cdots \\ N(v_{s-1}) \end{pmatrix} \right). \tag{5.3.4}$$

Here we have not done the spatial discretization, if we want to do this with fully discretization we could use Kronecker products to get a bigger equation system rigorously.

Thus, we have

$$\frac{1}{\tau}\begin{pmatrix} v_1 - v_0 \\ v_2 - v_0 \\ \cdots \\ v_s - v_0 \end{pmatrix} = A \begin{pmatrix} \mathcal{L}v_1 \\ \mathcal{L}v_2 \\ \cdots \\ \mathcal{L}v_s \end{pmatrix} + \hat{A} \begin{pmatrix} N(v_0) \\ N(v_1) \\ \cdots \\ N(v_s) \end{pmatrix}. \tag{5.3.5}$$

For the phase field model, we have $\mathcal{L}v = GD_s v, N(v) = -Gf_s(v)$, where $D_s = -(1+\alpha)D + \beta I$ and $f_s = -f - \alpha D + \beta I$, so we can derive the following equation

$$\begin{aligned} \frac{1}{\tau}\begin{pmatrix} v_1 - v_0 \\ v_2 - v_0 \\ \cdots \\ v_s - v_0 \end{pmatrix} =& - A \begin{pmatrix} GDv_1 \\ GDv_2 \\ \cdots \\ GDv_s \end{pmatrix} + \hat{A} \begin{pmatrix} Gf(v_0) \\ Gf(v_1) \\ \cdots \\ Gf(v_{s-1}) \end{pmatrix} - \alpha \left[A \begin{pmatrix} GDv_1 \\ GDv_2 \\ \cdots \\ GDv_s \end{pmatrix} - \hat{A} \begin{pmatrix} GDv_0 \\ GDv_1 \\ \cdots \\ GDv_{s-1} \end{pmatrix} \right] \\ &+ \beta \left[A \begin{pmatrix} Gv_1 \\ Gv_2 \\ \cdots \\ Gv_s \end{pmatrix} - \hat{A} \begin{pmatrix} Gv_0 \\ Gv_1 \\ \cdots \\ Gv_{s-1} \end{pmatrix} \right]. \end{aligned} \tag{5.3.6}$$

Then we are going to reformulate the system. For simplicity, we denote the inverse matrix of $\hat{A}$ as $\hat{B} \in \mathcal{R}^{s\times s}$,

$$\hat{A}\hat{B} = \hat{B}\hat{A} = I. \tag{5.3.7}$$

Besides, we list some simple lemmas to help the simplification.

Lemma 3.1.

$$\begin{pmatrix} v_1 - v_0 \\ v_2 - v_0 \\ \cdots \\ v_s - v_0 \end{pmatrix} = E_L \begin{pmatrix} v_1 - v_0 \\ v_2 - v_1 \\ \cdots \\ v_s - v_{s-1} \end{pmatrix}. \tag{5.3.8}$$

Lemma 3.2.

$$(A - \hat{A})\mathbf{1}_s = c - \hat{c} = 0, \tag{5.3.9}$$

where $\mathbf{1}_s = (1, 1, ..., 1)^T \in \mathcal{R}^{s\times 1}$. *This lemma shows these two matrices share the same eigenvector* $\mathbf{1}_s$ *and helps much in simplification. Taking the second term on the right hand side of (5.3.6) as an example,*

$$
\begin{aligned}
A\begin{pmatrix} GDv_1 \\ GDv_2 \\ \cdots \\ GDv_s \end{pmatrix} =& A\begin{pmatrix} GDv_1 \\ GDv_2 \\ \cdots \\ GDv_s \end{pmatrix} - (A-\hat{A})\begin{pmatrix} GDv_0 \\ GDv_0 \\ \cdots \\ GDv_0 \end{pmatrix} \\
=& A\begin{pmatrix} GDv_1 - GDv_0 \\ GDv_2 - GDv_0 \\ \cdots \\ GDv_s - GDv_0 \end{pmatrix} + \hat{A}\begin{pmatrix} GDv_0 \\ GDv_0 \\ \cdots \\ GDv_0 \end{pmatrix} \\
=& AE_l\begin{pmatrix} GD(v_1 - v_0) \\ GD(v_2 - v_1) \\ \cdots \\ GD(v_s - v_{s-1}) \end{pmatrix} + \hat{A}\begin{pmatrix} GDv_0 \\ GDv_0 \\ \cdots \\ GDv_0 \end{pmatrix}.
\end{aligned} \tag{5.3.10}
$$

Lemma 3.3.

$$A\begin{pmatrix} Gv_1 \\ Gv_2 \\ \cdots \\ Gv_s \end{pmatrix} - \hat{A}\begin{pmatrix} Gv_0 \\ Gv_1 \\ \cdots \\ Gv_{s-1} \end{pmatrix} = \left[A\begin{pmatrix} Gv_1 \\ Gv_2 \\ \cdots \\ Gv_s \end{pmatrix} - \hat{A}\begin{pmatrix} Gv_0 \\ Gv_1 \\ \cdots \\ Gv_{s-1} \end{pmatrix}\right] - \hat{A}\begin{pmatrix} Gv_1 \\ Gv_2 \\ \cdots \\ Gv_s \end{pmatrix} + \hat{A}\begin{pmatrix} Gv_1 \\ Gv_2 \\ \cdots \\ Gv_s \end{pmatrix}$$

$$=(A-\hat{A})\begin{pmatrix} Gv_1 \\ Gv_2 \\ \cdots \\ Gv_s \end{pmatrix} + \hat{A}\begin{pmatrix} Gv_1 - Gv_0 \\ Gv_2 - Gv_1 \\ \cdots \\ Gv_s - Gv_{s-1} \end{pmatrix}$$

$$=(A-\hat{A})\left[\begin{pmatrix} Gv_1 \\ Gv_2 \\ \cdots \\ Gv_s \end{pmatrix} - \begin{pmatrix} Gv_0 \\ Gv_0 \\ \cdots \\ Gv_0 \end{pmatrix}\right] + \hat{A}\begin{pmatrix} Gv_1 - Gv_0 \\ Gv_2 - Gv_1 \\ \cdots \\ Gv_s - Gv_{s-1} \end{pmatrix}$$

$$= (A-\hat{A})E_L(Gw) + \hat{A}(Gw) = \hat{A}Q(Gw), \tag{5.3.11}$$

where for simplity we denote $w = (v_1 - v_0, ..., v_s - v_{s-1})^T$ *and*$(Pw) = (P(v_1 - v_0), ..., P(v_s - v_{s-1}))^T$ *for all operators* P, *and* $Q = \left((\hat{A})^{-1}A - I\right)E_L + I$.

Therefore, using all these lemmas, (5.3.6) can be reformulated in the following way

$$\frac{1}{\tau}E_L(G^{-1}w) = -AE_L(Dw) - \hat{A}\begin{pmatrix} Dv_0 \\ Dv_0 \\ \cdots \\ Dv_0 \end{pmatrix} + \hat{A}\begin{pmatrix} f(v_0) \\ f(v_1) \\ \cdots \\ f(v_{s-1}) \end{pmatrix} - \alpha\hat{A}Q(Dw) + \beta\hat{A}Q(w) \quad (5.3.12)$$

$$\begin{pmatrix} f(v_0) \\ f(v_1) \\ \cdots \\ f(v_{s-1}) \end{pmatrix} = \frac{1}{\tau}\hat{B}E_L(G^{-1}w) + \hat{B}AE_L(Dw) + \begin{pmatrix} Dv_0 \\ Dv_0 \\ \cdots \\ Dv_0 \end{pmatrix} \quad (5.3.13)$$

$$+ \alpha Q(Dw) - \beta Q(w).$$

Next we focus on the difference of the energy, which reads

$$E_{n+1} - E_n = -\frac{1}{2}\left((u_{n+1}, Du_{n+1}) - (u_n, Du_n)\right) + (F_{n+1} - F_n, \mathbf{1}), \quad (5.3.14)$$

where (u, v) represents the inner product in space.

Notice that given any $a, b \in [-1, 1]$, we have

$$F(a) - F(b) \leq f(b)(a-b) + \frac{L}{2}(a-b)^2,$$

where L is the Lipschitz constant of the function f. Thus,

$$\begin{aligned}(F_{n+1}-F_n,\mathbf{1}) &= \sum_{i=0}^{s-1}(F(v_{i+1})-F(v_i),\mathbf{1}) \\ &\leq \begin{pmatrix} v_1-v_0, v_2-v_1, ..., v_s-v_{s-1}\end{pmatrix}\begin{pmatrix} f(v_0)\\ f(v_1)\\ ...\\ f(v_{s-1})\end{pmatrix} + \frac{L}{2}\begin{pmatrix} v_1-v_0, v_2-v_1, ..., v_s-v_{s-1}\end{pmatrix}^2 \\ &= w^T\begin{pmatrix} f(v_0)\\ f(v_1)\\ ...\\ f(v_{s-1})\end{pmatrix} + \frac{L}{2}w^2,\end{aligned} \tag{5.3.15}$$

where $u^2 = u^T u$ for simplicity. Now we substitute $(f(v_0), f(v_1), ..., f(v_s))^T$ by using (5.3.13),

$$\begin{aligned}(F_{n+1}-F_n,\mathbf{1}) \leq & \frac{L}{2}w^2 + w^T\begin{pmatrix} \\ \frac{1}{\tau}\hat{B}E_L(G^{-1}w) + \hat{B}AE_L(Dw) + \begin{pmatrix} Dv_0\\ Dv_0\\ ...\\ Dv_0\end{pmatrix}\\ \\ \end{pmatrix} \\ & + \alpha w^T Q(Dw) - \beta w^T Q(w)\end{aligned} \tag{5.3.16}$$

Notice

$$w^T \begin{pmatrix} Dv_0 \\ Dv_0 \\ \cdots \\ Dv_0 \end{pmatrix} = (v_s - v_0, Dv_0), \tag{5.3.17}$$

and combining this term with the other term in the energy difference gives

$$\begin{aligned} & -\frac{1}{2}\left((v_s, Dv_s) - (v_0, Dv_0)\right) + (v_s - v_0, Dv_0) \\ =& -\frac{1}{2}(v_s, Dv_s) - \frac{1}{2}(v_0, Dv_0) + (v_s, Dv_0) \\ =& -\frac{1}{2}(v_s - v_0, D(v_s - v_0)) \\ =& -\frac{1}{2} w^T E_1(Dw). \end{aligned} \tag{5.3.18}$$

To conclude, the energy difference becomes

$$\begin{aligned} E_{n+1} - E_n \leq & -\frac{1}{2} w^T E_1(Dw) + \frac{L}{2} w^2 + w^T \left(\frac{1}{\tau}\hat{B}E_L(G^{-1}w) + \hat{B}AE_L(Dw)\right) \\ & + \alpha w^T Q(Dw) - \beta w^T Q(w) \\ =& \frac{1}{\tau} w^T H_0(G^{-1}w) - w^T H_1 w + w^T H_2(Dw) \end{aligned} \tag{5.3.19}$$

where $H_0 = \hat{B}E_L, H_1 = \beta Q - \frac{L}{2}I, H_2 = \alpha Q + \hat{B}AE_L - \frac{1}{2}E_1$.

If here H_0, H_1 and H_2 are positive-definite, then the energy difference is negative. In order to satisfy the condition, the only requirement which is also a must, is that Q and H_0 are positive-definite, and after that we only need to set α and β big enough. The energy

unconditionally decreases as long as

$$\begin{aligned} \beta &\geq \beta_0 = \frac{L}{2\lambda_{min}(Q)}, \\ \alpha &\geq \alpha_0 = -\frac{\lambda_{min}(\hat{B}AE_L - \frac{1}{2}E_1)}{\lambda_{min}(Q)}, \end{aligned} \tag{5.3.20}$$

where λ_{min} represents the smallest eigenvalue, α_0 and β_0 are constants only dependent on the Runge–Kutta scheme. □

Remark 5.3.3. *For the MBE model without slope selection, the proof is slightly different but shares the same key idea so we omit it here.*

Remark 5.3.4. *In (5.3.19), the matrices H_0, H_1 and H_2 can be viewed as coefficients or some combination of H^{-k}, L^2 and H^1 norms of the difference of the solution at different time steps, where k depends on the operator G. Therefore, we do not have to require all three terms to be positive-definite, but could use H_0^{-k}, H_0^1 terms to cover the L^2 term.*

Theorem 3.2. *(For the Allen–Cahn equation) For the IMEX-RK scheme (5.2.8), if $H_0 = (\hat{A})^{-1}E_L$ and $Q = \left((\hat{A})^{-1}A - I\right)E_L + I$ are both positive-definite, then it decreases the energy of the Allen–Cahn equation when following inequalities hold:*

$$\begin{aligned} &\frac{1}{\tau}\lambda_{min}(H_0) + \beta\lambda_{min}(Q) \geq \frac{L}{2} \\ &\alpha\lambda_{min}(Q) \geq -\lambda_{min}((\hat{A})^{-1}AE_L - \frac{1}{2}E_1). \end{aligned}$$

Theorem 3.3. *(For the Cahn–Hilliard equation) For the IMEX-RK scheme (5.2.8), if $H_0 = (\hat{A})^{-1}E_L$ and $Q = \left((\hat{A})^{-1}A - I\right)E_L + I$ are both positive-definite, then it decreases*

the energy of the Cahn–Hilliard equation when the following inequality holds:

$$\frac{4\epsilon^2}{\tau}\lambda_{min}(H_0)\lambda_{min}(H_2) + \beta\lambda_{min}(Q) \geq \frac{L}{2}$$

where $H_2 = \alpha Q + (\hat{A})^{-1}AE_L - \frac{1}{2}E_1$.

5.4 Error Analysis

In this section, we analyze a class of s-stage pth-order IMEX-RK schemes (5.2.6) for the phase field model (5.2.5) with requirements in the Theorem 3.1. The error analysis is straightforward with the energy stability result, although it is still technical to deal with some details to derive the optimal estimate for smooth solutions. In this section we use C to represent a generic positive constant which is independent of τ and ϵ but may depend on the Runge-Kutta scheme. It may also have a different value in each occurrence.

We denote the exact solution as $u(t)$ and at each stage time step $t_{ni} = t_n + c_i\tau, 1 \leq i \leq s$, we define reference solutions $\bar{v}_i$ by

$$\begin{aligned} &\bar{v}_0 = u(t_n), \\ &\begin{pmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \cdots \\ \bar{v}_s \end{pmatrix} = \begin{pmatrix} \bar{v}_0 \\ \bar{v}_0 \\ \cdots \\ \bar{v}_0 \end{pmatrix} + \tau \left(A \begin{pmatrix} \mathcal{L}\bar{v}_1 \\ \mathcal{L}\bar{v}_2 \\ \cdots \\ \mathcal{L}\bar{v}_s \end{pmatrix} + \hat{A} \begin{pmatrix} N(\bar{v}_0) \\ N(\bar{v}_1) \\ \cdots \\ N(\bar{v}_{s-1}) \end{pmatrix} \right), \end{aligned} \tag{5.4.1}$$

and the exact solution satisfies the following equation

$$u(t_{n+1}) = u(t_n) + \tau \left(b^T \begin{pmatrix} \mathcal{L}\bar{v}_1 \\ \mathcal{L}\bar{v}_2 \\ \cdots \\ \mathcal{L}\bar{v}_s \end{pmatrix} + \hat{b}^T \begin{pmatrix} N(\bar{v}_0) \\ N(\bar{v}_1) \\ \cdots \\ N(\bar{v}_{s-1}) \end{pmatrix} \right) + r_{n+1}. \tag{5.4.2}$$

Consider the Taylor expansion of (5.4.1) at $t = t_n$ and use asymptotic analysis, applying the order conditions of the IMEX-RK scheme

$$b_\sigma^T c_\sigma^{j-1} = \hat{b_\sigma}^T c_\sigma^{j-1} = \tfrac{1}{j},$$

$$b_\sigma^T A_\sigma^{j-1} e_\sigma = \hat{b_\sigma}^T A_\sigma^{j-1} e_\sigma = b_\sigma^T \hat{A_\sigma}^{j-1} e_\sigma = \hat{b_\sigma}^T \hat{A_\sigma}^{j-1} e_\sigma = \tfrac{1}{j!}, 1 \le j \le p,$$

where $\sigma = s+1, c_\sigma^j = (0, c_1^j, c_2^j, ..., c_s^j)^T$ and $b_\sigma, \hat{b_\sigma}, A_\sigma, \hat{A}_\sigma$ represent $\sigma \times 1$ vectors and $\sigma \times \sigma$ matrices in (5.2.6) respectively. Coefficients of $u^{(k)}(t_n)\tau^k$ on both side are equal up to order p and thus we derive $r_{n+1} = C\tau^{p+1}$.

In order to obtain error estimates, we define $e_j = \bar{v}_j - v_j, 0 \le j \le s-1$ and $e_s = u(t_{n+1}) - u_{n+1} = \bar{v}_s - v_s - r_{n+1}$. The difference of (5.4.1) and (5.3.4) shows

$$\begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_s \end{pmatrix} = \begin{pmatrix} e_0 \\ e_0 \\ \cdots \\ e_0 \end{pmatrix} + \tau \left(A \begin{pmatrix} \mathcal{L}e_1 \\ \mathcal{L}e_2 \\ \cdots \\ \mathcal{L}e_s \end{pmatrix} + \hat{A} \begin{pmatrix} N(\bar{v}_0) - N(v_0) \\ N(\bar{v}_1) - N(v_1) \\ \cdots \\ N(\bar{v}_{s-1}) - N(v_{s-1}) \end{pmatrix} \right) + \begin{pmatrix} 0 \\ 0 \\ \cdots \\ r_{n+1} \end{pmatrix}. \tag{5.4.3}$$

Using $\mathcal{L}v = GD_s v, N(v) = -Gf_s(v)$, where $D_s = -(1+\alpha)D + \beta I$ and $f_s = -f - \alpha D + \beta I$

and following all steps in the proof of Theorem 3.1, we derive

$$\begin{aligned}\frac{1}{2}\|\nabla e_s\|_2^2 - \frac{1}{2}\|\nabla e_0\|_2^2 =& \frac{1}{\tau}p^T H_0(G^{-1}p) - \beta p^T Q p + p^T H_2(Dp) \\ & - p^T \begin{pmatrix} f(\bar{v}_0) - f(v_0) \\ f(\bar{v}_1) - f(v_1) \\ \cdots \\ f(\bar{v}_{s-1}) - f(v_{s-1}) \end{pmatrix} - \frac{1}{\tau}p^T(\hat{A})^{-1} \begin{pmatrix} 0 \\ 0 \\ \cdots \\ r_{n+1} \end{pmatrix},\end{aligned} \tag{5.4.4}$$

where $p = (e_1 - e_0, e_2 - e_1, ..., e_s - e_{s-1})^T$. According to Theorem 3.1, here H_0, Q and H_2 are all positive-definite, so the first three terms on the right-hand side are all negative. For the other two terms,

$$-\frac{1}{\tau}p^T(\hat{A})^{-1} \begin{pmatrix} 0 \\ 0 \\ \cdots \\ r_{n+1} \end{pmatrix} \leq \frac{C_1}{\tau}p^T p + \frac{C_2}{\tau}\|r_{n+1}\|_2^2, \tag{5.4.5}$$

$$\begin{aligned} -p^T \begin{pmatrix} f(\bar{v}_0) - f(v_0) \\ f(\bar{v}_1) - f(v_1) \\ \cdots \\ f(\bar{v}_{s-1}) - f(v_{s-1}) \end{pmatrix} &\leq \frac{C_3}{\tau}p^T p + C_4\tau\sum_{i=0}^{s-1} e_i^2 \\ &\leq \frac{C_3}{\tau}p^T p + C_4\tau\sum_{i=0}^{s-1}((e_i - e_0) + e_0)^2 \\ &\leq \frac{C_3}{\tau}p^T p + C_5\tau p^T p + C_6\tau\|e_0\|_2^2. \end{aligned} \tag{5.4.6}$$

Besides, notice that

$$\begin{aligned}\frac{1}{2}\|e_s\|_2^2 &\leq (\frac{1}{2}+C\tau)\|e_0\|_2^2+(\frac{1}{2}+\frac{1}{2C\tau})\|e_s-e_0\|_2^2\\ &=(\frac{1}{2}+C\tau)\|e_0\|_2^2+(\frac{1}{2}+\frac{1}{2C\tau})p^TE_1p.\end{aligned} \tag{5.4.7}$$

For the Allen–Cahn equation, adding up four inequalities above leads to the following result

$$\|e_s\|_2^2+\|\nabla e_s\|_2^2\leq(1+C\tau)(\|e_0\|_2^2+\|\nabla e_0\|_2^2)+\frac{C'}{\tau}\|r_{n+1}\|_2^2, \tag{5.4.8}$$

where $e_s=u(t_{n+1})-u_{n+1}$ and $e_0=u(t_n)-u_n$. For the Cahn–Hilliard equation, we need one more inequality

$$\|(\nabla)^{-1}p_i\|_2\|\nabla p_i\|_2\geq\|p_i\|_2^2,\quad 1\leq i\leq s \tag{5.4.9}$$

which leads to

$$\left(p^T((-G)^{-1}p)\right)\left(p^T((-D)p)\right)\geq\left(p^Tp\right)^2, \tag{5.4.10}$$

to obtain (5.4.8).

Therefore, by Gronwall's inequality, we derive the error estimate

$$\begin{aligned}\|u(t_n)-u_n\|_{H^1}^2&\leq Ce^{CT}\tau^{2p},\\ \|u(t_n)-u_n\|_{H^1}&\leq Ce^{CT}\tau^{p}.\end{aligned} \tag{5.4.11}$$

5.5 Runge–Kutta schemes

In this section we present some Runge–Kutta schemes.

5.5.1 Example 1: first-order IMEX

This simple one-step case corresponds to the following IMEX scheme whose tableau reads:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0 & 1 \\ \hline & 0 & 1 \end{array}, \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1 & 0 \end{array}. \tag{5.5.1}$$

Here $A = \hat{A} = 1$, and thus $H_0 = Q = 1$ and $H_2(0) = 1/2$. Therefore, for the Allen-Cahn equation, we only need

$$\frac{1}{dt} + \beta \geq 1$$

to guarantee the energy dissipation law; while for the Cahn–Hilliard equation, we only need

$$\frac{2\epsilon^2}{dt} + \beta \geq \frac{L}{2}.$$

5.5.2 Example 2: a second-order IMEX

Consider this second-order IMEX-RK scheme with coefficients (see, eg.[2])

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & 0 & \gamma & 0 \\ 1 & 0 & 1-\gamma & \gamma \\ \hline & 0 & 1-\gamma & \gamma \end{array}, \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & \gamma & 0 & 0 \\ 1 & \delta & 1-\delta & 0 \\ \hline & \delta & 1-\delta & 0 \end{array}, \tag{5.5.2}$$

where $\gamma = 1 - \frac{\sqrt{2}}{2}, \delta = 1 - \frac{1}{2\gamma}$.

Here

$$A = \begin{pmatrix} \gamma & 0 \\ 1-\gamma & \gamma \end{pmatrix}, \quad \hat{A} = \begin{pmatrix} \gamma & 0 \\ \delta & 1-\delta \end{pmatrix}, \tag{5.5.3}$$

and thus

$$H_0 = \begin{pmatrix} 2+\sqrt{2} & 0 \\ 2 & 2-\sqrt{2} \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 \\ 0 & 3-2\sqrt{2} \end{pmatrix}, \quad H_2 = \begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 5/2-2\sqrt{2} \end{pmatrix}. \tag{5.5.4}$$

The corresponding smallest eigenvalues of their symmetrizers are

$$\lambda_{min}(H_0) = 2-\sqrt{3}, \quad \lambda_{min}(Q) = 3-2\sqrt{2}, \quad \lambda_{min}(H_2(0)) = 5/2-2\sqrt{2}. \tag{5.5.5}$$

Therefore, we need the $\alpha|D_{1/2}u|^2$ term in the splitting. For the Allen-Cahn equation, we need

$$\begin{aligned} &\frac{2-\sqrt{3}}{dt} + (3-2\sqrt{2})\beta \geq 1, \\ &\alpha \geq \alpha_0 = \frac{2\sqrt{2}-5/2}{3-2\sqrt{2}} \approx 1.9142. \end{aligned} \tag{5.5.6}$$

For the Cahn-Hilliar equation, we need

$$4(2-\sqrt{3})(\alpha+5/2-2\sqrt{2})\frac{\epsilon^2}{dt} + (3-2\sqrt{2})\beta \geq \frac{L}{2}. \tag{5.5.7}$$

However, since $\lambda_{min}(Q)$ for this scheme is too small, if we want to get unconditional energy dissipation we have to set $\beta > \frac{1}{\lambda_{min}(Q)} = 3+2\sqrt{2}$, which is too large and may cause unwanted error.

5.5.3 Third-order schemes

There is no pair of a three-stage, L-stable DIRK (diagonally implicit Runge–Kutta) and a four-stage ERK (explicit Runge–Kutta) with a combined third-order accuracy (see [2]), so we have to consider 4-stage schemes. However, as far as we search, no existing 4-stage third-order ARS Runge–Kutta scheme satisfies the conditions in Theorem 3.1 and we construct a new one. Here we list two common 4-stage third-order Runge–Kutta schemes as examples to illustrate why they fail, and then present our new 4-stage 3rd-order scheme and introduce our strategy to construct.

Example 3

$$A = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 1/24 & 11/24 & 1/4 & 0 \\ 11/24 & 1/6 & 1/8 & 1/4 \end{pmatrix}, \quad \hat{A} = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 13/4 & -3 & 0 & 0 \\ 1/4 & 0 & 1/2 & 0 \\ 0 & 1/3 & 1/6 & 1/2 \end{pmatrix}. \tag{5.5.8}$$

The corresponding smallest eigenvalues of symmetrizers of the determinants are approximately

$$\lambda_{min}(H_0) = -1.496826, \quad \lambda_{min}(Q) = -0.165679, \quad \lambda_{min}(H_2(0)) = -0.665679. \tag{5.5.9}$$

Since here both H_0 and Q are negative-definite, this scheme does not satisfy the conditions of our theorem.

Example 4

$$A = \begin{pmatrix} 1/2 & 0 & 0 & 0 \\ 1/6 & 1/2 & 0 & 0 \\ -1/2 & 1/2 & 1/2 & 0 \\ 3/2 & -3/2 & 1/2 & 1/2 \end{pmatrix}, \quad \hat{A} = \begin{pmatrix} 1/2 & 0 & 0 & 0 \\ 11/18 & 1/18 & 0 & 0 \\ 5/6 & -5/6 & 1/2 & 0 \\ 1/4 & 7/4 & 3/4 & -7/4 \end{pmatrix}. \tag{5.5.10}$$

The corresponding smallest eigenvalues of symmetrizers of the determinants are approximately

$$\lambda_{min}(H_0) = -15.242727, \quad \lambda_{min}(Q) = -7.706226, \quad \lambda_{min}(H_2(0)) = -8.206226. \tag{5.5.11}$$

Since here both H_0 and Q are also negative-definite, this scheme also does not satisfy the conditions of our theorem.

Example 5: Energy decreasing 4-stage third-order IMEX-RK scheme

Here we present a 4-stage third-order ARS IMEX-RK scheme which has rational coefficients in c which are rational and not too large.

0	0	0	0	0	0
3/5	0	0.6	0	0	0
3/2	0	0.46875	1.03125	0	0
19/20	0	0.4	−0.5578125	1.1078125	0
1	0	a_{41}	a_{42}	a_{43}	25.75
	0	a_{41}	a_{42}	a_{43}	25.75

(5.5.12)

0	0	0	0	0	0
3/5	0.6	0	0	0	0
3/2	0.796875	0.703125	0	0	0
19/20	0.4	$\hat{a}_{42}$	$\hat{a}_{43}$	0	0
1	$\hat{a}_{41}$	$\hat{a}_{42}$	$\hat{a}_{43}$	$\hat{a}_{44}$	0
	$\hat{a}_{41}$	$\hat{a}_{42}$	$\hat{a}_{43}$	$\hat{a}_{44}$	0

(5.5.13)

where

$$\begin{aligned}
a_{41} &= 3.736772486772523;\\
a_{42} &= -0.781144781144795;\\
a_{43} &= -27.705627705628103;\\
\hat{a}_{42} &= 0.420225694444444;\\
\hat{a}_{43} &= 0.129774305555556;\\
\hat{a}_{41} &= 0.301169590643275;\\
\hat{a}_{42} &= 0.330687830687831;\\
\hat{a}_{43} &= -0.087542087542087;\\
\hat{a}_{44} &= 0.455684666210982;
\end{aligned} \tag{5.5.14}$$

The corresponding smallest eigenvalues of symmetrizers of the determinants are approximately

$$\lambda_{min}(H_0) = 0.087230, \quad \lambda_{min}(Q) = 1, \quad \lambda_{min}(H_2(0)) = 0.5, \tag{5.5.15}$$

which are all positive. Therefore, we only need to set

$$\beta \geq \frac{1}{\lambda_{min}(Q)} = 1,$$

so that this scheme unconditionally decreases the energy of phase fields with Lipschitz nonlinear terms.

Here we also simply illustrate our strategy to search coefficients in the tableau. In order to construct 4-stage third-order IMEX-RK schemes, we only need to solve a linear

problem. The order condition of IMEX-RK3 is

$$\begin{aligned} &A_\sigma e_\sigma = \hat{A}_\sigma e_\sigma = c_\sigma, \\ &b_\sigma^T e_\sigma = \hat{b_\sigma}^T e_\sigma = 1, b_\sigma^T c_\sigma = \hat{b_\sigma}^T c_\sigma = \frac{1}{2}, b_\sigma^T c_\sigma^2 = \hat{b_\sigma}^T c_\sigma^2 = \frac{1}{3}, \\ &b_\sigma^T A_\sigma c_\sigma = \hat{b_\sigma}^T A_\sigma c_\sigma = b_\sigma^T \hat{A}_\sigma c_\sigma = \hat{b_\sigma}^T \hat{A}_\sigma c_\sigma = \frac{1}{6}. \end{aligned} \tag{5.5.16}$$

There are 16 different equations for 23 variables. Although the whole system seems to be nonlinear, we could set free parameters to make it a linear problem. First we set $c_\sigma = (0, c_2, c_3, c_4, 1)$, and then let

$$b_\sigma^T c_\sigma^3 = \zeta, \quad \hat{b_\sigma}^T c_\sigma^3 = \hat{\zeta}. \tag{5.5.17}$$

Combining this condition with (5.5.16), we could solve b_σ and $\hat{b_\sigma}$ in terms of c_σ, where the linear equation involves the Vandermonde determinant. The last step is to solve A_σ and $\hat{A}_\sigma$, where the nonlinear system becomes a linear one.

However, such trick does not work for fourth and higher order situations, since the corresponding order condition of (5.5.16) will have such forms

$$b_\sigma^T A_\sigma^{p-2} c_\sigma = \hat{b_\sigma}^T A_\sigma^{p-2} c_\sigma = b_\sigma^T \hat{A_\sigma}^{p-2} c_\sigma = \hat{b_\sigma}^T \hat{A_\sigma}^{p-2} c_\sigma = \frac{1}{p!}, \tag{5.5.18}$$

where even if we already know all variables in b and c unknowns, a nonlinear problem still needs to be solved. Therefore, we do not present a fourth-order scheme here.

Chapter 6

Conclusions and Future work

Throughout the thesis, we studied three different schemes to solve phase field models with gradient flow structures.

6.1 Explicit Runge–Kutta methods

In Chapter 3, we offer detailed analysis and prove a family of explicit Runge–Kutta methods which preserve both the maximum bound property and the energy dissipation law for the Allen–Cahn equation. However, due to the order limit determined by the strong stability preserving (SSP) Runge–Kutta methods, there do not exist schemes of higher order than three which could preserve both structures. Explicit methods also suffer from time step restrictions and can not be efficient for applications.

However, we believe when small time steps have to be taken, these explicit Runge–Kutta methods are the simplest and also most useful methods to implement. But when should such situations happen? The answer is the popular adaptive time-stepping methods. As is shown in Chapter 4, the curve of these phase field models is usually flat and rarely

turns out to have dramatic variations. When the curve is flat, we only need to use some efficient and stable numerical methods with large time steps and when it is going to vary dramatically, we could use explicit Runge–Kutta methods with small time steps. Whether explicit RK methods can be better than other methods is still an open problem and needs to be verified.

6.2 Exponential time differencing Runge–Kutta methods

In Chapter 4, we prove that the second-order exponential time differencing Runge–Kutta (ETDRK) methods unconditionally preserve both MBP and the energy dissipation law and thus becomes almost the best second-order scheme. Therefore, the main problem is whether we can derive higher-order ETDRK schemes which still preserve the structures. This problem, somehow, has intrinsic difficulties. The key idea of ETDRK approaches is applying discrete integral in the Duhamel's Formula, and constant and linear interpolations lead to first and second-order schemes correspondingly. It is also not hard to realize the constant and linear interpolations do not break the maximum bound property. However, when we step to third-order schemes and use second-order polynomials (parabolas), the MBP can be violated. Furthermore, the loss of the MBP could also affect the proof of the energy dissipation. In conclusion, we need some more techniques and approaches to get higher-order structure-preserving ETDRK schemes.

6.3 Implicit-explicit Runge–Kutta methods

In Chaper 5, we prove and find a class of implicit-explicit Runge–Kutta methods which unconditionally decrease the original energy of phase field models with gradient flow structures. The techniques used here could also apply to other PDEs with similar structure. I believe a general result can be shown for a wide class of problems.

Bibliography

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

[2] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Applied Numerical Mathematics, 25 (1997), pp. 151–167.

[3] K. Burrage, W. H. Hundsdorfer, and J. G. Verwer, *A study of B-convergence of Runge-Kutta methods*, Computing, 36 (1986), p. 1734.

[4] J. C. Butcher, *Numerical methods for ordinary differential equations*, Wiley, (2003).

[5] J. H. Chaudhry, J. Collins, and J. N. Shadid, *A posteriori error estimation for multi-stage Runge-Kutta IMEX schemes*, Applied Numerical Mathematics, 117 (2017), pp. 36–49.

[6] L. Chen, *Phase-field models for microstructural evolution*, Ann. Rev. Mater. Res., 32 (2002), pp. 113–140.

[7] L. Chen and J. Shen, *Applications of semi-implicit Fourier-spectral method to phase field equations*, Comput. Phys. Commun., 108 (1998), pp. 147–158.

[8] 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, 40 (2018), pp. A3982–A4006.

[9] S. M. Cox and P. C. Matthews, *Exponential time differencing for stiff systems*, J. Comput. Phys., 176 (2002), pp. 430–455.

[10] L. Dong and Q. Zhonghua, *On second order semi-implicit Fourier spectral methods for 2D Cahn-Hilliard equations*, J. Sci. Comp., 70 (2017), pp. 301–341.

[11] L. Dong, Q. Zhonghua, and T. Tao, *Characterizing the stabilization size for semi-implicit Fourier-spectral method to phase field equations*, SIAM Journal on Numerical Analysis, 54 (2016), pp. 1653–1681.

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

[13] Q. Du, L. Ju, X. Li, and Z. Qiao, *Maximum bound principles for a class of semi-linear parabolic equations and exponential time differencing schemes*, SIAM Review, 63 (2021), pp. 317–359.

[14] Q. Du and J. Yang, *Asymptotically compatible Fourier spectral approximations of nonlocal Allen–Cahn equations*, SIAM J. Numer. Anal., 54 (2016), pp. 1899–1919.

[15] Q. Du and W. Zhu, *Stability analysis and application of the exponential time differencing schemes*, J. Comput. Math., 22 (2004), pp. 200–209.

[16] C. M. Elliott, *The Cahn-Hilliard model for the kinetics of phase separation*, Mathematical models for phase change problems (Obidos, 1988), Internat. Ser. Numer. Math., vol. 88, Birkhauser, Basel, 88 (1989), pp. 35–73.

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

[18] H. Emmerich, *The diffuse interface approach in materials science*, Springer, New York, (2003).

[19] D. J. Eyre, *An unconditionally stable one-step scheme for gradient systems*, unpublished,http://www.math.utah.edu/eyre/research/methods/stable.ps., (1997).

[20] D. J. Eyre, *Unconditionally gradient stable time marching the Cahn-Hilliard equation*, Mater. Res. Soc. Symp. Proc., 529 (1998), pp. 39–46.

[21] X. Feng, T. Tang, and J. Yang, *Long time numerical simulations for phase-field problems using p-adaptive spectral deferred correction methods*, SIAM Journal on Scientific Computing, 37 (2015), pp. A271–A294.

[22] L. Ferracina and M. N. Spijker, *Stepsize restrictions for the total-variation-diminishing property in general Runge–Kutta methods*, SIAM Journal on Numerical Analysis, 42 (2004), pp. 1073–1093.

[23] E. Fried and M. E. Gurtin, *Dynamic solid-solid transitions with phase characterized by an order parameter*, Physics D: Nonlinear Phenomena, 72 (1994), pp. 287–308.

[24] L. Golubovic, A. Levandovsky, and D. Moldovan, *Interface dynamics and far-from-equilibrium phase transitions in multilayer epitaxial growth and erosion on crystal surfaces: Continuum theory insights*, East Asian J. Appl. Math., 1 (2011), pp. 297–371.

[25] H. Gomez, L. Cueto-Felgueroso, and R. Juanes, *Three-dimensional simulation of unstable gravity-driven infiltration of water into a porous medium*, Journal of Computation Physics, 238 (2013), pp. 217–239.

[26] H. Gomez and T. Hughes, *Provably unconditionally stable, second-order time-accurate, mixed variational methods for phase-field models*, J. Comput. Phys., 230 (2011), pp. 5310–5327.

[27] Y. Gong, Q. Wang, Y. Wang, and J. Cai, *A conservative Fourier pseudo-spectral method for the nonlinear Schrodinger equation*, Journal of Computational Physics, 328 (2017), pp. 354–370.

[28] C. Gottlieb, S amd Shu and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[29] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu, *Strong stability preserving Runge–Kutta and multistep time discretizations*, World Scientific Press, (2011).

[30] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge–Kutta schemes*, Mathematics of Computation, 67 (1998), pp. 73–85.

[31] Z. Guan, C. Wang, and S. Wise, *A convergent convex splitting scheme for the periodic nonlocal Cahn–Hilliard equation*, Numer. Math., 128 (2014), pp. 377–406.

[32] L. Ju, X. Li, Z. Qiao, and H. Zhang, *Energy stability and error estimates of exponential time differencing schemes for the epitaxial growth model without slope selection*, Math. Comp., 87 (2018), pp. 1859–1885.

[33] J. Kim, *Phase-field models for multi-component fluid flows*, Commun. Comput. Phys., 12 (2012), pp. 613–661.

[34] J. F. B. M. Kraaijevanger, *Contractivity of Runge–Kutta methods*, BIT, 31 (1991), p. 482528.

[35] L. Leibler, *Theory of microphase separation in block copolymers*, Macromolecules, 13 (6) (1980), pp. 1602–1617.

[36] B. Li and J. Liu, *Thin film epitaxy with or without slope selection*, European J. Appl. Math., 14 (2003), pp. 713–743.

[37] J. Lili, L. Xiao, Q. Zhonghua, and J. Yang, *Maximum bound principle preserving integrating factor RungeKutta methods for semilinear parabolic equations*, J. Comp. Phys., 439 (2021).

[38] O. Penrose and P. C. Fife, *Thermodynamically consistent models of phase-field type for the kinetics of phase transition*, Phys. D, 43 (1990), pp. 44–62.

[39] S. J. Ruuth and R. J. Spiteri, *Two barriers on strong-stability-preserving time discretization methods*, J. Sci. Comput., 17 (2002), pp. 211–220.

[40] J. Shen, T. Tang, and J. Yang, *On the maximum principle preserving schemes for the generalized Allen–Cahn equation*, Comm. Math. Sci., 14 (2016), pp. 1517–1534.

[41] J. Shen, J. Xu, and J. Yang, *The scalar auxiliary variable (SAV) approach for gradient flows*, Journal of Computational Physics, 353 (2018), pp. 407–416.

[42] J. Shen, J. Xu, and J. Yang, *A new class of efficient and robust energy stable schemes for gradient flows*, SIAM Rev., 61 (2019), pp. 474–506.

[43] J. Shen and X. Yang, *Numerical approximations of Allen–Cahn and Cahn–Hilliard equations*, Discret. Contin. Dyn. Syst., 28 (2010), pp. 1669–1691.

[44] C. Shu, *Total-variation-diminishing time discretizations*, SIAM Journal on Scientific and Statistical Computing, 9 (1988), pp. 1073–1084.

[45] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, Journal of Computational Physics, 77 (1988), pp. 439–471.

[46] T. Tang, *Revisit of semi-implicit schemes for phase-field equations*, Anal. Theory Appl., 36(3) (2020), pp. 235–242.

[47] T. Tang and J. Yang, *Implicit-explicit scheme for the Allen–Cahn equation preserves the maximum principle*, J. Comput. Math., 34 (2016), pp. 451–461.

[48] J. D. van der Waals, *The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density*, J. Stat. Phys., 20 (1979), pp. 197–244.

[49] J. G. Verwer, *Convergence and order reduction of diagonally implicit Runge-Kutta schemes in the method of lines*, Proc. Dundee Numerical Analysis Conference, 140 (1985), pp. 220–237.

[50] L. Xiao, Q. Zhonghua, and W. Cheng, *Convergence analysis for a stabilized linear semi-implicit numerical scheme for the nonlocal CahnHilliard equation*, Math. Comp., 90 (2021), pp. 171–188.

[51] C. Xu and T. Tang, *Stability analysis of large time-stepping methods for epitaxial growth models*, SIAM J. Numer. Anal., 44 (2006), pp. 1759–1779.

[52] J. Yang, Q. Du, and W. Zhang, *Uniform l^p -bound of the Allen–Cahn equation and its numerical discretization*, International Journal of Numerical Analysis and Modeling, 15 (1-2) (2018), pp. 213–227.

[53] X. Yang, *Linear, first and second-order, unconditionally energy stable numerical schemes for the phase field model of homopolymer blends*, Journal of Computational Physics, 327 (2016), pp. 294–316.

[54] X. Yang, J. Zhao, Q. Wang, and J. Shen, *Numerical approximations for a three-component Cahn–Hilliard phase-field model based on the invariant energy quadratization method*, Mathematical Models and Methods in Applied Sciences, 27 (2017), pp. 1993–2030.