NUMERICAL ANALYSIS OF VARIOUS SINGLE-STEP INTEGRATORS FOR PARABOLIC EQUATIONS

YUAN ZHAOMING

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Applied Mathematics

Numerical Analysis of Various Single-Step Integrators for Parabolic Equations

YUAN Zhaoming

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

October 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

(Signed)

YUAN Zhaoming (Name of student)

Abstract

Parabolic equations are essential in applications like heat conduction, diffusion processes, and financial modeling. They describe how quantities such as temperature, concentration, or option prices evolve over time, making them crucial in engineering, physics, and economics. This thesis aims to develop efficient numerical methods for solving parabolic problems, particularly in phase-field models, ensuring high accuracy while preserving maximum bound and energy decay properties.

In the first part of thesis, we consider the development and analysis of the structure preserving schemes for solving Allen–Cahn equations, that represents an important application of parabolic equations. We apply a k-th order single-step method in time, where the nonlinear term is linearized using multi-step extrapolation. In space, we use a lumped mass finite element method with piecewise r-th order polynomials and Gauss–Lobatto quadrature. At each time level, a cut-off post-processing technique is proposed to eliminate values that violate the maximum bound principle at the finite element nodal points. As a result, the numerical solution satisfies the maximum bound principle at all nodal points, and the optimal error bound $O(\tau^k + h^{r+1})$ is theoretically proven. These time-stepping schemes include algebraically stable collocation-type methods, which can achieve arbitrarily high order in both space and time. By combining the cut-off strategy with the scalar auxiliary variable (SAV) technique, we develop a class of energy-stable and maximum bound preserving schemes that are arbitrarily high-order in time.

In the second part, we present the development and analysis of a class of single-step implicit-explicit schemes for approximately solving linear parabolic equations, which achieves long-time stability and arbitrarily high order in time. This involves splitting the linear operator into symmetric and skew-symmetric components, evaluated implicitly and explicitly, respectively, using the Implicit-Explicit Runge–Kutta Method (IMEX-RK). For the symmetric part, a diagonally implicit method (DIRK) is employed, while

the discretization for skew-symmetric part is designed to satisfy the stage orders. This method is applicable to semilinear problems, such as phase-field models, and our analysis is consistent with existing findings, showing energy stability for certain IMEX-RK schemes. Our results reveal intersections up to at least third order, leading to a scheme that preserves both the original energy decay properties and maximum bound principles.

In the third part of the thesis, we study the parareal algorithm for solving parabolic equations,which enables parallel-in-time computation and significantly accelerates the process. We prove that the parareal method has a robust convergence rate of about 0.3, provided the ratio J of coarse to fine step size exceeds a certain threshold J_*, and the fine propagator meets mild conditions. This convergence is robust even with nonsmooth problem data and boundary condition incompatibilities. Qualified methods include all absolutely stable single-step methods with a stability function satisfying $|r(-\infty)| < 1$, allowing the fine propagator to be of arbitrarily high order. Moreover, we examine popular high-order single-step methods, such as the two-, three-, and four-stage Lobatto IIIC methods, confirming that their corresponding parareal algorithms converge linearly with a factor of 0.31, with a threshold $J_* = 2$.

At the end of each chapter, we present numerical results that support the theoretical findings and inspire future investigations.

Acknowledgements

In this part, I would like to share my appreciate to some people or some organisations that might contribute to the thesis and my study.

I was born in a northern city of China, named Qiqihar, in which the snow covers everything about six months each year. That was tough days for me but thanks to a series of people from my relatives, the communities and the schools, I could enter Beijing Normal University.

I start the first calculus in middle school, and I would call that the start of my mathematical career. I visited every famous bookstore in Qiqihar, and collected the books there, since the e-books were less popular and I cannot get them. Although I have to admit that the books there were bad, they inspired me, and later I started majoring mathematics till now.

People say that Gaokao is important in China, but I did not take it a big deal. It was not hard for me to enter BNU. Not surprisingly, based on my background, I guess I dare to say that I was some kind of outstanding and I honored many awards and scholarships. During that days, I also learnt a lot on many aspects of mathematics.

The professors in BNU are very kind and responsible. Maybe it is because BNU is Teacher's University that almost all of the professors are good and talent in teaching, but that is definitely not the only reason. Some professor will support a discuss session and audit it, and the students would collect the questions, and present them. I learnt LaTeX during the prediscussion, Prof. Liu helped us on the typing and formatting the documents on his spare time. I would also thank other teachers and professors, Prof. Huang, Dr. Liu, Eng. Liu, Dr. Wang, etc. who taught me or helped me. I also learnt a series of softwares or platforms during that days, and some of them I am still using in my research and in my thesis.

I met my current supervisor in BNU. Late Professor Zhang introduced me to Dr. Zhou, and we dis-

cussed about the research plan. Dr. Zhou allowed me to present his Research Assistant and then Dr. Yang and Dr. Zhou offered and funded me their collaborate Ph.D. programme. Dr. Zhou also encouraged me on the preparation and application. They have my deep thanks.

In the days in Hong Kong, from ruins and pandemic we suffered, remotely the studying and researching we did. Everyone, supervisors, professors, instructors, classmates, and friends from chatgroups or internets, supported me and supported each other during that days. I also finished some papers and researches that days. I would also appreciate the help from my supervisor, Dr. Zhou and Dr. Yang on my research, including topics, advising, discussions, answering my questions, and helping me on my papers. I would also thank Dr. Zhang, Dr. Wang, Mr. Li and other teachers and collaborators for the helping and discussion with my research.

I will finish my postgraduate study after the thesis. The new occupation is ready and offered. Many rejection has been received and finally I get the right one. I would like to thank Dr. Zhou, Dr. Yang, Dr. Chen, Dr. Fan on my new work.

At the end of this part, I would like to thank my girlfriend, my love and my wife-to-be, Dr. Fan Ganghua. We met at BNU, and will get married soon. Since then, we have been through everything together. She is the first and the last one who encouraged me on everything, on postgraduate application, on tough days, on research and on my occupation. I can hardly finish all of these on myself without her help.

Contents

Chapter 1

Introduction

1.1 Research Background

Parabolic differential equations are essential in various scientific fields, including physics, engineering, and economics. In physics, they are used to describe heat conduction, diffusion processes, and the behavior of semiconductors. In finance, they are used in the Black-Scholes model for option pricing. In biology, they are used to model population dynamics and the spread of diseases. Additionally, the study of parabolic differential equations has introduced many important mathematical concepts and techniques. The method of separation of variables, commonly used to solve these equations, is a fundamental tool in mathematical analysis. The theory also plays a crucial role in studying stochastic processes and Brownian motion. Parabolic differential equations remain a central topic in mathematical analysis and applied mathematics. They have a rich history and are still an active area of research, offering many challenging problems and applications.

Phase field models are one of the most important applications of parabolic equations, which are mathematical tools widely used in physics, materials science, and other fields to describe the evolution of complex microstructures. They are particularly useful for modeling phase transitions, such as the solidification of a liquid or the formation of crystals in a solution.

Introduced in the late 20th century, the phase field approach addresses challenges associated with traditional methods for modeling phase transitions. Traditional methods often involve sharp interfaces

between different phases, which can be difficult to handle both mathematically and computationally. In contrast, the phase field approach treats interfaces as diffuse regions and describes the microstructure using a continuous field variable, simplifying the mathematical and computational processes. Phase field models have been applied to a wide range of phenomena, including crystal growth, grain boundary motion, and pattern formation in alloys. They are also used in areas such as fluid dynamics, image processing, and tumor growth modeling.

In a phase field model, the microstructure is represented by a phase field variable that varies continuously between phases. The evolution of this phase field is governed by parabolic equations derived from thermodynamic and kinetic principles. These equations are solved numerically using various computational methods.

Structure-preserving schemes are numerical methods used in computational mathematics and physics. They are designed to maintain the inherent geometric or physical properties of the problem being solved. Structure preservation is crucial in many scientific and engineering applications, such as molecular dynamics, fluid dynamics, and electromagnetism, where preserving properties like energy, momentum, or symplectic structure is essential for the physical relevance and accuracy of the solution. These schemes often provide more accurate and physically relevant solutions than traditional numerical methods, especially in long-term simulations and highly nonlinear problems.

For phase field models, researchers focus on developing numerical schemes that preserve **energy dissipation law** and **maximum bound principle**, without strict constraints on time step and space mesh sizes. For example, we focus on the development and analysis of high-order structure-preserving schemes for solving the Allen–Cahn equation:

$$\begin{cases} u_t = \Delta u + f(u) & \text{in } \Omega \times (0,T), \\ u(x,t=0) = u_0(x) & \text{in } \Omega \times \{0\}, \\ \partial_{\mathbf{n}} u = 0 & \text{on } \partial\Omega \times (0,T) \end{cases} \tag{1.1}$$

where Ω is a smooth domain in $\mathbb{R}^d$ with the boundary $\partial\Omega$. Here, $f(u) = -F'(u)$ with a double-well potential F that has two wells at $\pm\alpha$, for some known parameter $\alpha > 0$. It is well-known that the Allen–

Cahn equation (1.1) has the **maximum bound principle** [29]:

$$|u_0(x)| \le \alpha \quad \Rightarrow \quad |u(x,t)| \le \alpha \qquad \text{for all } (x,t) \in \Omega \times (0,T]. \tag{1.2}$$

As a typical L^2 gradient flow associated with the following free energy:

$$E(u) = \int_\Omega \frac{1}{2}|\nabla u|^2 + F(u)\,\mathrm{d}x,$$

the nonlinear **energy dissipation law** holds:

$$\frac{\mathrm{d}}{\mathrm{d}t}E(u) = -\int_\Omega |u_t|^2\,\mathrm{d}x \le 0. \tag{1.3}$$

We aim to develop high-order numerical schemes that preserve both conditions (1.2) and (1.3). Additionally, we will discuss efficient parallel-in-time algorithms for solving the (nonlinear) parabolic equations.

1.2 Literature Review

In this part, we briefly review the existing literature on structure-preserving schemes for solving parabolic equations and phase-filed models.

The backward Euler time-stepping scheme, combined with the central finite difference method in space, effectively preserves the maximum principle for linear parabolic equations [68, Chapter 9]. Additionally, using the backward Euler scheme with the lumped mass linear finite element method (FEM) and simplicial triangulation with acute angles also maintains this principle. In two dimensions, this extends to Delaunay-type triangulations, which is notably sharp [111]. However, without mass lumping, standard Galerkin FEMs generally do not preserve the maximum principle [111, 96]. These methods achieve first-order accuracy in time and second-order accuracy in space.

The development and analysis of maximum bound preserving schemes for Allen–Cahn equations have been intensively studied in existing references. It was proved in [109, 98] that the stabilized semi-implicit Euler time-stepping scheme, with central difference method in space, preserves the maximum principle unconditionally if the stabilizer satisfies certain restrictions. In [30], a stabilized exponential time

differencing scheme was proposed for solving the (nonlocal) Allen–Cahn equation, and the scheme was proved to be unconditionally MBP. See also [29] for the generalization to a class of semilinear parabolic equations.

The development and analysis of maximum bound preserving schemes for Allen–Cahn equations have been intensively studied in existing references. It was proved in [109, 98] that the stabilized semi-implicit Euler time-stepping scheme, with central difference method in space, preserves the maximum principle unconditionally if the stabilizer satisfies certain restrictions. In [30], a stabilized exponential time differencing scheme was proposed for solving the (nonlocal) Allen–Cahn equation, and the scheme was proved to be unconditionally MBP. See also [29] for the generalization to a class of semilinear parabolic equations.

High-order strong stability preserving (SSP) time-stepping methods are widely used in the development of MBP scheme for both parabolic equations and hyperbolic equations (see e.g., [46, 79, 47, 45, 78, 90, 119, 124]). Recently, an SSP integrating factor Runge–Kutta method of up to order four was proposed and analyzed in [58] for semilinear hyperbolic and parabolic equations. For semilinear hyperbolic and parabolic equations with strong stability (possibly in the maximum norm), the method can preserve this property and can avoid the standard parabolic CFL condition $\tau = O(h^2)$, only requiring the stepsize τ to be smaller than some constant depending on the nonlinear source term, also referring to [62]. A nonlinear constraint limiter was introduced in [113] for implicit time-stepping schemes without requiring CFL conditions, which can preserve maximum principle at the discrete level with arbitrarily high-order methods by solving a nonlinearly implicit system.

Very recently, a new class of high-order MBP methods was proposed in [73]. The method consists of a kth-order multistep exponential integrator in time, and a lumped mass finite element method in space with piecewise rth-order polynomials. At every time level, the extra values exceeding the maximum bound are eliminated at the finite element nodal points by a cut-off operation. Then the numerical solution at all nodal points satisfies the MBP, and an error bound of $O(\tau^k + h^r)$ was proved. However, numerical results in [73, Table 4.1] indicates that the error bound is not sharp in space, and how to improve the estimate it is still open. Besides, the aforementioned scheme requires to evaluate some actions of exponential functions of diffusion operators, which might be relatively expensive compared with solving poisson problems, and the generalization to other time stepping schemes is a nontrivial task. Finally, the proposed scheme (with relatively coarse step sizes) might produce a numerical solution with obviously increasing and oscillating

energy. These motivate our current project.

There have been numerous studies focused on developing various numerical schemes that preserve the energy dissipation law at a discrete level. Some notable and widely-used implicit time-stepping methods include convex splitting methods [33, 103] and the Crank-Nicolson type scheme [37, 32]. The main drawback of these methods is the high computational cost associated with solving a nonlinear system of equations at each time step. In contrast, implicit-explicit (IMEX; also known as semi-implicit) methods handle the nonlinear term explicitly and the linear term implicitly, requiring only the solution of a linear system of equations at each time step. These methods can be traced back to the work of Chen and Shen [15] in the context of phase-field models, and since then, many techniques and strategies have been developed to design such schemes, as seen in [1, 20, 36, 48, 50, 74, 100, 102, 107]. Building on the concept of the invariant energy quadratization (IEQ) method [120, 121]. [99, 100] proposed the scalar auxiliary variable (SAV) method, which easily ensures the unconditional energy decay property. Recently, some modified SAV methods have been developed [21, 55, 60, 108]. However, the energy considered in these methods is modified from the original energy. In another direction, exponential time differencing (ETD) methods for the Allen–Cahn equation and other semilinear parabolic equations have garnered significant attention recently. Du et al. [30] demonstrated that ETD and ETDRK2 schemes unconditionally preserve the maximum bound property (MBP) and energy stability (though not the dissipation law). Specifically, [39] establishes the original energy stability for ETDRK2. For the thin film model (or MBE model), interesting results regarding stability analysis and error estimates for the ETD schemes are presented in [23, 28, 61, 71, 118]. Additionally, [51] shows that fully implicit Runge-Kutta (RK) methods can reduce the energy of gradient systems, but the existence and uniqueness of the solution remain unresolved issues. Another class of implicit Runge-Kutta methods for phase-field models is based on the convex splitting approach, which exhibits favorable stability properties [104].

Time-stepping schemes for solving parabolic equations traditionally require sequential computation, which can be time-consuming. However, with modern computing power, parallel-in-time (PinT) methods have become feasible, allowing simultaneous computation of multiple time steps. Originating from Nievergelt's work in 1964 [86], these methods have gained considerable interest. Among them, the parareal method, introduced in 2001 [76], is particularly popular due to its simplicity and adaptability with single-step integrators. It has been effectively applied in various fields such as turbulent plasma [93, 92], structural dynamics [22, 34], molecular dynamics [6], optimal control [80, 82], and fractional models [75, 117].

For more comprehensive insights, readers can consult survey papers [40, 87] and references therein.

1.3 Our Contribution

In this thesis, we discuss the development, analysis, and implementation of structure-preserving single-step integrators for solving parabolic equations, with applications to phase-field models like the Allen-–Cahn equations. Single-step integrators often offer better stability properties than linear multistep methods, particularly for stiff problems, allowing for larger time steps and more efficient simulations. Additionally, they provide superior error control, especially when using adaptive time steps. Compared to linear multistep methods, single-step integrators are also more suitable for applying postprocessing techniques to preserve certain structures, as we will explore in this thesis.

In the first part of thesis, we develop and analyze a class of maximum bound preserving schemes for approximately solving Allen–Cahn equations. We apply a kth-order single-step scheme in time (where the nonlinear term is linearized by multi-step extrapolation), and a lumped mass finite element method in space with piecewise rth-order polynomials and Gauss–Lobatto quadrature. At each time level, a cut-off post-processing is proposed to eliminate extra values violating the maximum bound principle at the finite element nodal points. As a result, the numerical solution satisfies the maximum bound principle (at all nodal points), and the optimal error bound $O(\tau^k + h^{r+1})$ is theoretically proved for a certain class of schemes. The proof is based on energy estimation. Since cut-off itself will not increase the total energy, we just compare each step with its previous step, which is decoupled with postprocessing and complete the proof. These time stepping schemes under consideration includes algebraically stable collocation-type methods, which could be arbitrarily high-order in both space and time. Moreover, combining the cut-off strategy with the scalar auxiliary value (SAV) technique, we also develop a class of energy-stable and maximum bound preserving schemes, which is arbitrarily high-order in time.

In the second part of this thesis, we explore implicit-explicit Runge–Kutta methods for solving parabolic equations. We begin by examining linear parabolic problems where the differential operator may be non-selfadjoint, which is crucial in applications like Stokes–Darcy coupled systems [52, 85, 4]. These systems are often used to model contaminant transport in karst aquifers, where fluid motion in porous media is coupled with conduits. During floods, contaminants can enter the porous media and be released during droughts. Accurate numerical schemes are essential for capturing the long-term retention and release of

contaminants, as fluid motion in porous media is slower than in conduits.

To address this, we develop a class of implicit-explicit Runge–Kutta methods and prove their long-term stability and error estimates. We split the differential operator into symmetric and skew-symmetric parts: the skew-symmetric part and source term are evaluated explicitly, while the symmetric part is evaluated implicitly. Our analysis uses spectral decomposition of the symmetric operator and energy estimation. We establish a novel energy argument to demonstrate long-term stability and optimal error estimates, choosing a special test function inspired by the backward Euler scheme.

This approach can be extended to nonlinear phase-field models. With cutoff postprocessing at each step, we show that the method preserves the maximum bound and can achieve arbitrarily high order. In [38], it was proven that IMEX-RK schemes can maintain energy dissipation laws under certain assumptions. Our assumptions align with theirs, and we have found a third-order scheme that meets all criteria, allowing us to develop time-stepping schemes up to third order that preserve both the maximum bound and original energy dissipation.

In the third part of thesis, we investigate the parareal algorithm for solving parabolic equations, which enables parallel-in-time computation and significantly accelerates the process. For linear problems, we analyzed the robust convergence of a class of parareal algorithms. The coarse propagator is fixed to the backward Euler method and the fine propagator is a high-order single step integrator. Under some conditions on the fine propagator, we show that there exists some critical J_* such that t he parareal solver converges linearly with a convergence rate near 0.3, provided that the ratio between the coarse time step and fine time step named J satisfies $J \geq J_*$. The convergence is robust even if the problem data is nonsmooth and incompatible with boundary conditions. The qualified methods include all absolutely stable single step methods, whose stability function satisfies $|r(-\infty)| < 1$, and hence the fine propagator could be arbitrarily high-order. Moreover, we examine some popular high-order single step methods, e.g., two-, three- and four-stage Lobatto IIIC methods, and verify that the corresponding parareal algorithms converge linearly with a factor 0.31 and the threshold for these cases is $J_* = 2$.

Chapter 2

Arbitrarily High-order Maximum Bound Preserving Schemes with Cut-off Postprocessing

In this chapter, we develop and analyze a class of maximum bound preserving schemes for approximately solving Allen–Cahn equations. We apply a single-step scheme in time with nonlinear term linearized, and a lumped mass finite element method in space. At each time level, a cut-off post-processing is proposed to eliminate extra values violating the maximum bound principle at the finite element nodal points. As a result, the numerical solution satisfies the maximum bound principle (at all nodal points), and the optimal error bound is theoretically proved for a certain class of schemes. Moreover, combining the cut-off strategy with the scalar auxiliary value (SAV) technique, we also develop a class of energy-stable and maximum bound preserving schemes, which is arbitrarily high-order in time.

In Section 2.1 we discuss the time discretization problems and show its convergence. In Section 2.2 and 2.3 we discuss the fully-discrete scheme and prove its convergence. In Section 2.4 we combine our scheme with Scalar Auxiliary Variable method and prove it preserves modified energy decay and arbitrary high order in time.

2.1 Temporal Semi-discrete Cut-off Runge–Kutta Scheme

To begin with, we consider the time discretization for the Allen–Cahn equation (1.1). To this end, we split the interval $(0,T)$ into N subintervals with the uniform mesh size $\tau = T/N$, and set $t^n = n\tau$, $n = 0,1,\ldots,N$. On the time interval $[t^{n-1},t^n]$, we approximate the nonlinear term $f(u(s))$ by the extrapolation polynomial

$$\sum_{j=1}^{k} L_j(s) f(u^{n-j}), \quad \text{with known } u^{n-k},\ldots,u^{n-1}.$$

where $L_j(s)$ is the Lagrange basis polynomials of degree $k-1$ in time, satisfying

$$L_j(t^{n-i}) = \delta_{ij}, \quad i,j = 1,\ldots,k.$$

Thus, on $[t^{n-1},t^n]$, the linearization of (1.1) states as

$$\tilde{u}_t = \Delta\tilde{u} + \sum_{j=1}^{k} L_j(s) f(u^{n-j}).$$

Following Duhamel's principle yields

$$\tilde{u}(t^n) = e^{\tau\Delta}u(t^{n-1}) + \int_0^\tau e^{(\tau-s)\Delta}\sum_{j=1}^{k} L_j(t^{n-1}+s) f(u^{n-j})\mathrm{d}s.$$

Then a framework of a single step scheme of approximating $\tilde{u}(t^n)$ reads:

$$\tilde{u}^n = \sigma(-\tau\Delta)u^{n-1} + \tau\sum_{i=1}^{m} p_i(-\tau\Delta)\Big(\sum_{j=1}^{k} L_j(t^{ni}) f(u^{n-j})\Big), \quad \text{for all } n \geq k, \tag{2.1}$$

with $t^{ni} = t^{n-1} + c_i\tau$. Here, $\sigma(\lambda)$ and $\{p_i(\lambda)\}_{i=1}^m$ are rational functions and c_i are distinct real numbers in $[0,1]$. For simplicity, we assume that the scheme (2.1) satisfies the following assumptions.

(P1) $|\sigma(\lambda)| < 1$ and $|p_i(\lambda)| \leq c$, for all $i = 1,\ldots,m$, uniformly in τ and $\lambda > 0$. Besides, the numerator of $p_i(\lambda)$ is of lower degree than its denominator.

(P2) The time stepping scheme (2.1) is accurate of order k in sense that

$$\sigma(\lambda) = e^{-\lambda} + O(\lambda^{k+1}), \quad \text{as } \lambda \to 0.$$

and, for $0 \le j \le k$

$$\sum_{i=1}^{m} c_i^j p_i(\lambda) - \frac{j!}{(-\lambda)^{j+1}}\Big(e^{-\lambda} - \sum_{\ell=0}^{j} \frac{(-\lambda)^\ell}{\ell!}\Big) = O(\lambda^{k-j}), \quad \text{as } \lambda \to 0.$$

(P3) The time discretization scheme (2.1) is strictly accurate of order q in sense that

$$\sum_{i=1}^{m} c_i^j p_i(\lambda) - \frac{j!}{(-\lambda)^{j+1}}\Big(\sigma(\lambda) - \sum_{\ell=0}^{j} \frac{(-\lambda)^\ell}{\ell!}\Big) = 0, \quad \text{for all } 0 \le j \le q-1.$$

Remark 2.1.1. *In practice, it is convenient to choose $p_i(\lambda)$ that share the same denominator of $\sigma(\lambda)$, for instance:*

$$\sigma(\lambda) = \frac{a_0(\lambda)}{g(\lambda)}, \quad \textit{and} \quad p_i(\lambda) = \frac{a_i(\lambda)}{g(\lambda)}, \quad \textit{for } i = 1, 2, \ldots, m,$$

where $a_i(\lambda)$ and $g(\lambda)$ are polynomials. Then the time stepping scheme (2.1) *could be written as*

$$g(-\tau\Delta)\tilde{u}^n = a_0(-\tau\Delta)u^{n-1} + \tau\sum_{i=1}^{m} a_i(-\tau\Delta)\Big(\sum_{j=1}^{k} L_j(t^{ni}) f(u^{n-j})\Big), \quad \textit{for all } n \ge k.$$

See e.g. [110, pp. 131] for the construction of such rational functions satisfying the Assumptions (P1)-(P3).

Unfortunately, the time stepping scheme (2.1) does not satisfy the maximum bound principle. Therefore, at each time step, we apply the cut-off operation: for $n \ge k$, we find u^n such that

$$\hat{u}^n = \sigma(-\tau\Delta)u^{n-1} + \tau\sum_{i=1}^{m} p_i(-\tau\Delta)\Big(\sum_{j=1}^{k} L_j(t^{ni}) f(u^{n-j})\Big), \tag{2.2}$$

$$u^n = \min(\max(\hat{u}^n, -\alpha), \alpha), \tag{2.3}$$

where α is the maximum bound given in (1.2). The accuracy of this cut-off semi-discrete method is guaranteed by the next theorem.

Theorem 2.1.1. *Suppose that the Assumptions (P1) and (P2) are fulfilled, and (P3) holds for $q = k$. Let $u(t)$ be the solution to the Allen–Cahn equation, and u^n be the solution to the time stepping scheme* (2.2)-(2.3)*. Assume that $|u_0| \leq \alpha$ and the maximum principle* (1.2) *holds, and assume that the starting values u^j, $j = 0, \ldots, k-1$, are given and*

$$|u^j| \leq \alpha, \quad \textit{for all } \ j = 0, \ldots, k-1.$$

Then the semi-discrete solution given by (2.2)-(2.3) *satisfies for all $n \geq k$*

$$|u^n| \leq \alpha,$$

and

$$\|u^n - u(t^n)\| \leq C\tau^k + C\sum_{j=0}^{k-1} \|u^j - u(t^j)\|,$$

provided that f is locally Lipschitz continuous, $\Delta u \in C^k([0,T]; L^2(\Omega))$, $u \in C^{k+1}([0,T]; L^2(\Omega))$ and $f(u) \in C^k([0,T]; L^2(\Omega))$.

Proof. Due to the cut-off operation (2.3), the discrete maximum bound principle follows immediately. Then it suffices to show the error estimate.

Let $e^n = u^n - u(t^n)$ and $\hat{e}^n = \hat{u}^n - u(t^n)$. Since the exact solution satisfies the maximum bound (1.2), we have

$$\|e^n\|_{L^2(\Omega)} \leq \|\hat{e}^n\|_{L^2(\Omega)}.$$

Then it is easy to note that

$$\hat{e}^n = \sigma(-\tau\Delta)e^{n-1} + \varphi^n, \quad n \geq k.$$

where φ^n can be written as

$$\begin{aligned}
\varphi^n &= -u(t^n) + \sigma(-\tau\Delta)u(t^{n-1}) + \tau\sum_{i=1}^{m} p_i(-\tau\Delta)\Big(\sum_{j=1}^{k} L_j(t^{ni}) f(u^{n-j})\Big) \\
&= \tau\sum_{i=1}^{m} p_i(-\tau\Delta)\Big(\sum_{j=1}^{k} L_j(t^{ni}) f(u^{n-j}) - f(t^{ni}))\Big) \\
&\quad + \Big(- u(t^n) + \sigma(-\tau\Delta)u(t^{n-1}) + \tau\sum_{i=1}^{m} p_i(-\tau\Delta)(\partial_t u - \Delta u)(t^{ni})\Big)
\end{aligned}$$

$$=: I + II.$$

Then the bound of I follows from the approximation property of Lagrange interpolation, the maximum bound of u^{n-j} and $u(t^{n-j})$, $j = 1, \dots, k$, the locally Lipschitz continuity of f, and the Assumption (P1):

$$\begin{aligned}\|I\|_{L^2(\Omega)} &\le \tau \sum_{i=1}^{m} \|p_i(-\tau\Delta)\|_{L^2(\Omega)\to L^2(\Omega)} \Big\| \sum_{j=1}^{k} L_j(t^{ni}) f(u(t^{n-j})) - f(u(t^{n-1}+c_i\tau)) \Big\|_{L^2(\Omega)} \\ &\quad + \tau \sum_{i=1}^{m} \|p_i(-\tau\Delta)\|_{L^2(\Omega)\to L^2(\Omega)} \sum_{j=1}^{k} |L_j(t^{ni})| \, \|f(u^{n-j}) - f(u(t^{n-j}))\|_{L^2(\Omega)} \\ &\le C\tau^{k+1} \|f(u)\|_{C^k([t^{n-k},t^n];L^2(\Omega))} + C\tau \sum_{j=1}^{k} \|e^{n-j}\|_{L^2(\Omega)}.\end{aligned}$$

Now we term to the second term II, which can be rewritten by Taylor's expansion at t^{n-1}

$$\begin{aligned}II &= -\sum_{j=0}^{k} \frac{\tau^j}{j!} u^{(j)}(t^{n-1}) + \sigma(-\tau\Delta) u(t^{n-1}) \\ &\quad + \tau \sum_{i=1}^{m} p_i(-\tau\Delta) \sum_{j=0}^{k-1} \frac{(c_i\tau)^j}{j!} (u^{(j+1)} - \Delta u^{(j)})(t^{n-1}) + R_1 + R_2.\end{aligned}$$

where the remainders R_1 and R_2 are

$$\begin{aligned}R_1 &= \int_{t^{n-1}}^{t^n} \frac{(t^n - s)^k}{k!} u^{(k+1)}(s) \, \mathrm{d}s \quad \text{and} \\ R_2 &= \tau \sum_{i=1}^{m} p_i(-\tau\Delta) \int_{t^{n-1}}^{t^{n-1}+c_i\tau} \frac{(t^{n-1} + c_i\tau - s)^{k-1}}{(k-1)!} (u^{(k+1)} - \Delta u^{(k)})(s) \, \mathrm{d}s\end{aligned}$$

respectively. Hereafter, we use $u^{(j)}$ to denote the jth derivative in time. Then Assumption (P1) implies

$$\|R_1 + R_2\|_{L^2(\Omega)} \le C\tau^{k+1} \Big(\|u\|_{C^{k+1}([t^{n-1},t^n];L^2(\Omega))} + \|\Delta u\|_{C^k([t^{n-1},t^n];L^2(\Omega))} \Big).$$

Now we revisit the three leading terms of II. Note that

$$-\sum_{j=0}^{k} \frac{\tau^j}{j!} u^{(j)}(t^{n-1}) + \sigma(-\tau\Delta) u(t^{n-1}) + \tau \sum_{i=1}^{m} p_i(-\tau\Delta) \sum_{j=0}^{k-1} \frac{(c_i\tau)^j}{j!} (u^{(j+1)} - \Delta u^{(j)})(t^{n-1})$$

$$
\begin{aligned}
=&\Big(-I+\sigma(-\tau\Delta)-\tau\sum_{i=1}^{m}p_i(-\tau\Delta)\Delta\Big)u(t^{n-1})\\
&+\sum_{j=1}^{k-1}\frac{\tau^j}{j!}\Big(-I+j\sum_{i=1}^{m}c_i^{j-1}p_i(-\tau\Delta)-\tau\sum_{i=1}^{m}c_i^{j}p_i(-\tau\Delta)\Delta\Big)u^{(j)}(t^{n-1})\\
&+\frac{\tau^k}{k!}\Big(-I+k\sum_{i=1}^{m}c_i^{k-1}p_i(-\tau\Delta)\Big)u^{(k)}(t^{n-1})=\sum_{\ell=1}^{3}II_\ell.
\end{aligned}
$$

Since the time stepping scheme is strictly accurate of order $q=k$ (by Assumption (P3)), we have $II_1=II_2=0$. Meanwhile, we apply Assumption (P3) again to arrive at for $\lambda>0$

$$
-1+k\sum_{i=1}^{m}c_i^{k-1}p_i(\lambda)=\lambda\frac{k!}{(-\lambda)^{k+1}}\Big(\sigma(\lambda)-\sum_{\ell=0}^{k}\frac{(-\lambda)^\ell}{\ell!}\Big)=:\lambda\gamma(\lambda).
$$

Note that $|\gamma(\lambda)|=O(1)$ for $\lambda\to 0$ (by Assumption (P2)) and $|\gamma(\lambda)|\to 0$ for $\lambda\to+\infty$. Hence $|\gamma(\lambda)|$ is bounded uniformly in $[0,\infty)$. Then we arrive at

$$
\|II_3\|_{L^2(\Omega)}\le C\tau^{k+1}\|\Delta u^{(k)}(t^{n-1})\|\le C\tau^{k+1}\|\Delta u\|_{C^k([t^{n-1},t^n];L^2(\Omega))}.
$$

In conclusion, we obtain the following estimate

$$
\|e^n\|_{L^2(\Omega)}\le\|\sigma(-\tau\Delta)e^{n-1}\|_{L^2(\Omega)}+C\tau^{k+1}+C\tau\sum_{j=1}^{k}\|e^{n-j}\|_{L^2(\Omega)}.
$$

Then the assumption (P1) leads to

$$
\|e^n\|_{L^2(\Omega)}\le\|e_h^{n-1}\|_{L^2(\Omega)}+C\tau^{k+1}+C\tau\sum_{j=1}^{k}\|e^{n-j}\|_{L^2(\Omega)}.
$$

Finally, the desired assertion follows immediately by using discrete Gronwall's inequality

$$
\|e^n\|_{L^2(\Omega)}\le Ce^{cT}\tau^k+Ce^{cT}\sum_{j=0}^{k-1}\|e^j\|_{L^2(\Omega)}.
$$

□

Remark 2.1.2. *Theorem 2.1.1 implies that the cut-off operation preserves the maximum bound without*

losing global accuracy. However, the Assumption (P3) is restrictive. It is well-known that a single step method with a given $m \in \mathbb{Z}^+$ could be accurate of order $2m$ (Gauss–Legendre method) [31, Section 2.2], but at most strictly accurate of order $m+1$ [9, Lemma 5]. In general, a collocation-type method is only strictly accurate of order $m+1$.

Without the assumption of strict accuracy, one may still show the error estimate, provided that $f(u)$ satisfies certain compatibility conditions, e.g.,

$$f(u) \in C^\ell([0,T]; Dom(\Delta^{k-\ell})) \quad \text{for all} \quad \ell = 1, 2, \dots, k,$$

that requires $\partial_{\mathbf{n}} \Delta^q f(u) = 0$ for $\ell = 1, 2, \dots, k-1$. Unfortunately, those compatibility conditions cannot be fulfilled in general for semilinear parabolic problems.

Remark 2.1.3. *The same error estimate could be proved by assuming that the scheme satisfies the assumption (P3) with $q = k-1$ and some additional conditions (see e.g. [110, Theorem 8.4] and [88]). However, the proof is not directly applicable when we apply the cut-off operation at each time step. It warrants further investigation to show the sharp convergence rate $O(\tau^k)$ with weaker assumptions.*

2.2 Fully-discrete Cut-off Runge–Kutta Scheme

In this part, we discuss the fully discrete scheme. To illustrate the main idea, we consider the one-dimensional case $\Omega = [a, b]$, and the argument could be straightforwardly extended to multi-dimensional cases, see Remark 2.2.2. We denote by $a = x_0 < x_1 < \cdots < x_{Mr} = b$ a partition of the domain with a uniform mesh size $h = x_{ir} - x_{(i-1)r} = (b-a)/M$, and denote by S_h^r the finite element space of degree $r \geq 1$, i.e.,

$$S_h^r = \{v \in H^1(\Omega) : v|_{I_i} \in P_r, \;\; i = 1, \dots, M\},$$

where $I_i = [x_{(i-1)r}, x_{ir}]$ and P_r denotes the space of polynomials of degree $\leq r$.

Let $x_{(i-1)r+j}$ and $\omega_j, j = 0, \dots, r$, be the quadrature points and weights of the $(r+1)$-point Gauss–Lobatto quadrature on the subinterval I_i, and denote

$$w_{(i-1)r+j} = \begin{cases} \omega_j & \text{for } \; 1 \leq j \leq r-1, \\ 2\omega_j & \text{for } \; j = 0, r. \end{cases}$$

Then we consider the piecewise Gauss–Lobatto quadrature approximation of the inner product, i.e.,

$$(f,g)_h := \sum_{j=0}^{Mr} w_j f(x_j) g(x_j).$$

This discrete inner product induces a norm

$$\|f_h\|_h = \sqrt{(f_h, f_h)_h} \quad \forall\, f_h \in S_h^r.$$

Then we have the following lemma for norm equivalence. The proof follows directly from the positivity of Gauss–Lobatto quadrature weights [91, p. 426].

Lemma 2.2.1. *The discrete norm $\|\cdot\|_h$ is equivalent to usual L^2 norm $\|\cdot\|_{L^2(\Omega)}$ in sense that*

$$C_1\|v_h\|_{L^2(\Omega)} \le \|v_h\|_h \le C_2\|v_h\|_{L^2(\Omega)}, \quad \forall v_h \in S_h^r.$$

where C_1 and C_2 are independent of h.

To develop the fully discrete scheme, we introduce the discrete Laplacian $-\Delta_h : S_h^r \to S_h^r$ such that

$$(-\Delta_h v_h, w_h)_h = (\nabla v_h, \nabla w_h) \qquad \text{for all } \; v_h, w_h \in S_h^r. \tag{2.4}$$

Then at n-th time level, with given $u_h^{n-k}, \dots, u_h^{n-1} \in S_h^r$, we find an intermediate solution $\hat{u}_h^n \in S_h^r$ such that

$$\hat{u}_h^n = \sigma(-\tau\Delta_h)u_h^{n-1} + \tau\sum_{i=1}^{m} p_i(-\tau\Delta_h)\Big(\sum_{j=1}^{k} L_j(t^{ni})\Pi_h f(u_h^{n-j})\Big) \tag{2.5}$$

where $t^{ni} = t^{n-1} + c_i\tau$, and $\Pi_h : C(\overline{\Omega}) \to S_h^r$ is the Lagrange interpolation operator. In order to impose the maximum bound, we apply the cut-off postprocessing: find $u_h^n \in S_h^r$ such that

$$u_h^n(x_j) = \min\big(\max\big(\hat{u}_h^n(x_j), -\alpha\big), \alpha\big), \quad j = 0, \dots, Mr. \tag{2.6}$$

It is equivalent to

$$u_h^n = \Pi_h \min\big(\max\big(\hat{u}_h^n, -\alpha\big), \alpha\big).$$

Essentially, the cut-off operation (2.6) only works on the finite element nodal points.

Next, we shall prove the optimal error estimate of the fully discrete scheme (2.5)-(2.6). To this end, we need the following stability estimate of operators $\sigma(-\tau\Delta_h)$ and $p_i(-\tau\Delta_h)$.

Lemma 2.2.2. *Let Δ_h be the discrete Laplacian defined in* (2.4)*, and $\sigma(\lambda)$ and $p_i(\lambda)$ are rational functions satisfying the Assumption (P1). Then there holds that for all $v_h \in S_h^r$*

$$\|\nabla^q \sigma(-\tau\Delta_h)v_h\|_h \le \|\nabla^q v_h\|_h \quad \textit{and} \quad \|\nabla^q p_i(-\tau\Delta_h)v_h\|_h \le C\|\nabla^q v_h\|_h \tag{2.7}$$

with $i = 1, \dots, m$ and $q = 0, 1$. Meanwhile,

$$\tau\|\nabla^q \Delta_h p_i(-\tau\Delta_h)v_h\|_h \le C\|\nabla^q v_h\|_h \quad i = 1, \dots, m, \;\; q = 0, 1 \tag{2.8}$$

Proof. Let $\{(\lambda_j, \varphi_j^h)\}_{j=1}^{Mr+1}$ be eigenpairs of $-\Delta_h$, where $\{\varphi_j^h\}_{j=1}^{Mr+1}$ forms an orthogonal basis of S_h^r in sense that $(\varphi_i^h, \varphi_j^h)_h = \delta_{i,j}$. Then by the Assumption (P1), we have for any $v_h \in S_h^r$ and $q = 0, 1$

$$\begin{aligned}\|\nabla^q \sigma(-\tau\Delta_h)v_h\|_h^2 &= \sum_{j=1}^{Mr+1} (\lambda_j^h)^q |\sigma(\tau\lambda_j)|^2 |(v_h, \varphi_j^h)_h|^2 \\ &\le \sum_{j=1}^{Mr+1} (\lambda_j^h)^q |(v_h, \varphi_j^h)_h|^2 = \|\nabla^q v_h\|_h^2.\end{aligned}$$

This shows the first estimate. The estimate for p_i follows analogously.

Moreover, the numerator of $p_i(\lambda)$ is of lower degree than its denominator (by Assumption (P1)), and hence there exists constants $C_1, C_2 > 0$ such that

$$|p_i(\lambda)| \le \frac{C_1}{1 + C_2\lambda}, \quad \text{for all } \lambda > 0.$$

Then we derive that for any $v_h \in S_h^r$ and $q = 0, 1$

$$\begin{aligned}\tau^2\|\nabla^q \Delta_h p_i(-\tau\Delta_h)v_h\|_h^2 &= \tau^2 \sum_{j=1}^{Mr+1} (\lambda_j^h)^{q+2} |p_i(\tau\lambda_j)|^2 |(v_h, \varphi_j^h)_h|^2 \\ &\le C\tau^2 \sum_{j=1}^{Mr+1} \frac{(\lambda_j^h)^{q+2}}{(1 + C\tau\lambda_j^h)^2} |(v_h, \varphi_j^h)_h|^2\end{aligned}$$

$$\leq C \sum_{j=1}^{Mr+1} (\lambda_j^h)^q |(v_h, \varphi_j^h)_h|^2 = C\|\nabla^q v_h\|_h^2,$$

where the constant C only depends on C_1 and C_2. This proves the assertion (2.8). □

Lemma 2.2.3. *Let $v \in H^{2r+2}(\Omega)$ with the homogeneous Neumann boundary condition and $\varphi_h \in S_h^r$. Then we have the following estimate*

$$(\Pi_h \Delta v - \Delta_h \Pi_h v, \varphi_h)_h \leq C h^{r+1} \|v\|_{H^{2r+2}} \|\varphi_h\|_{H^1(\Omega)}.$$

Proof. Using the homogeneous Neumann boundary condition and (2.4), we obtain

$$\begin{aligned} &(\Pi_h \Delta v - \Delta_h \Pi_h v, \varphi_h)_h \\ &= (\Pi_h \Delta v, \varphi_h)_h - (\Delta_h \Pi_h v, \varphi_h)_h \\ &= \Big((\Delta v, \varphi_h)_h - (\Delta v, \varphi_h)\Big) + \Big((\Delta v, \varphi_h) - (\Delta_h \Pi_h v, \varphi_h)_h\Big) \\ &= \Big((\Delta v, \varphi_h)_h - (\Delta v, \varphi_h)\Big) + \Big((\partial_x v, \partial_x \varphi_h) - (\partial_x \Pi_h v, \partial_x \varphi_h)\Big) \end{aligned} \tag{2.9}$$

Since the $(r+1)$-point Gauss–Lobatto quadrature on each subinterval I_i is exact for polynomials of degree $2r-1$ [91, pp. 425], employing the Bramble–Hilbert lemma as well as the inverse inequality, we derive that

$$\begin{aligned} |(\Delta v, \varphi_h)_h - (\Delta v, \varphi_h)| &= \Big| \sum_{i=1}^{M} \Big(\sum_{j=0}^{r} \omega_j (\Delta v \varphi_h)(x_{(i-1)r+j}) - \int_{I_i} (\Delta v) \varphi_h \, \mathrm{d}x \Big) \Big| \\ &\leq C h^{2r} \sum_{i=1}^{M} \|\Delta v \varphi_h\|_{W^{2r,1}(I_i)} \leq C h^{2r} \sum_{i=1}^{M} \|v\|_{H^{2r+2}(I_i)} \|\varphi_h\|_{H^r(I_i)} \\ &\leq C h^{r+1} \sum_{i=1}^{M} \|v\|_{H^{2r+2}(I_i)} \|\varphi_h\|_{H^1(I_i)} \leq C h^{r+1} \|v\|_{H^{2r+2}(\Omega)} \|\varphi_h\|_{H^1(\Omega)}. \end{aligned}$$

Similar argument also leads to the estimate for the second term in (2.9) for $r \geq 2$:

$$\begin{aligned} |(\partial_x (v - \Pi_h v), \partial_x \varphi_h)| &= \Big| \sum_{i=1}^{M} \int_{I_i} \partial_x (v - \Pi_h v) \partial_x \varphi_h \, \mathrm{d}x \Big| = \Big| \sum_{i=1}^{M} \int_{I_i} (v - \Pi_h v) \partial_x^2 \varphi_h \, \mathrm{d}x \Big| \\ &= \Big| \sum_{i=1}^{M} \int_{I_i} v \partial_x^2 \varphi_h \, \mathrm{d}x - \sum_{j=0}^{r} \omega_j (v \partial_x^2 \varphi_h)(x_{(i-1)r+j}) \Big| \end{aligned}$$

$$\leq Ch^{2r}\sum_{i=1}^{M}\|v\partial_x^2\varphi_h\|_{W^{2r,1}(I_i)} \leq Ch^{2r}\sum_{i=1}^{M}\|v\|_{H^{2r+2}(I_i)}\|\varphi_h\|_{H^r(I_i)}$$
$$\leq Ch^{r+1}\sum_{i=1}^{M}\|v\|_{H^{2r+2}(I_i)}\|\varphi_h\|_{H^1(I_i)} \leq Ch^{r+1}\|v\|_{H^{2r+2}(\Omega)}\|\varphi_h\|_{H^1(\Omega)}.$$

Finally, in case that $r = 1$, it is easy to observe that

$$(\partial_x(v-\Pi_h v),\partial_x\varphi_h) = \sum_{i=1}^{M}\int_{I_i}\partial_x(v-\Pi_h v)\partial_x\varphi_h\,\mathrm{d}x = -\sum_{i=1}^{M}\int_{I_i}(v-\Pi_h v)\partial_x^2\varphi_h\,\mathrm{d}x = 0.$$

□

To derive an error estimate for the fully discrete scheme (2.5)-(2.6). We need the following extra assumptions on the rational function $\sigma(\lambda)$.

(P4) The rational function $\sigma(\lambda)$ satisfies $|\sigma(\lambda)| \to 0$ as $\lambda \to \infty$.

Note that the Assumption (P4) immediately implies [110, eq. (7.37)]

$$|\sigma(\lambda)| \leq \frac{1}{1+c_0\lambda} \qquad \text{for any } \lambda \geq 0,$$

with a generic constant $c_0 > 0$. This further implies

$$1 - |\sigma(\lambda)|^{-2} \leq -2c_0\lambda \qquad \text{for any } \lambda \geq 0.$$

Therefore, we have for any $v_h \in S_h^r$

$$\|\sigma(-\tau\Delta_h)v_h\|_h^2 = \sum_{j=1}^{Mr+1}|\sigma(\tau\lambda_j)|^2(v_h,\varphi_j^h)_h^2 = \|v_h\|_h^2 + \sum_{j=1}^{Mr+1}(|\sigma(\tau\lambda_j)|^2-1)(v_h,\varphi_j^h)_h^2$$
$$= \|v_h\|_h^2 + \sum_{j=1}^{Mr+1}(1-|\sigma(\tau\lambda_j)|^{-2})|\sigma(\tau\lambda_j)|^2(v_h,\varphi_j^h)_h^2$$
$$\leq \|v_h\|_h^2 - 2c_0\tau\sum_{j=1}^{Mr+1}\lambda_j|\sigma(\tau\lambda_j)|^2(v_h,\varphi_j^h)_h^2 = \|v_h\|_h^2 - 2c_0\tau\|\nabla\sigma(-\tau\Delta_h)v_h\|^2.$$

Then we are ready to state following main theorem.

Theorem 2.2.1. *Suppose that the Assumptions (P1), (P2) and (P4) are fulfilled, and (P3) holds for $q = k$. Assume that $|u_0| \leq \alpha$ and the maximum principle* (1.2) *holds, and assume that the starting values u_h^l, $l = 0, \ldots, k-1$, are given and*

$$|u_h^l(x_j)| \leq \alpha, \quad j = 0, \ldots, Mr, \quad l = 0, \ldots, k-1.$$

Then the fully discrete solution given by (2.5)-(2.6) *satisfies*

$$|u_h^n(x_j)| \leq \alpha, \quad j = 0, \ldots, Mr, \quad n = k, \ldots, N,$$

and for $n = k, \ldots, N$

$$\|u(t^n) - u_h^n\|_{L^2(\Omega)} \leq C(\tau^k + h^{r+1}) + C\sum_{l=0}^{k-1} \|u(t^l) - u_h^l\|_{L^2(\Omega)},$$

provided that $u \in C^{k+1}([0,T]; C(\bar{\Omega})) \cap C^k([0,T]; Dom(\Delta)) \cap C^1([0,T]; H^{2r+2}(\Omega))$, f is locally Lipschitz continuous and $f(u) \in C^k([0,T]; L^2(\Omega)) \cap C([0,T]; H^{2r+2}(\Omega))$.

Proof. In $[t^{n-1}, t^n]$, we note that $\Pi_h u$ satisfies

$$\partial_t \Pi_h u(t) - \Delta_h \Pi_h u(t) = \Pi_h f(u(t)) + g_h(t), \;\; t \in (t^{n-1}, t^n], \quad \text{with } \Pi_h u(t^{n-1}) \text{ given},$$

and $g_h(t) = (\Pi_h \Delta - \Delta_h \Pi_h)u(t)$. Then we define its time stepping approximation w_h^n satisfying

$$w_h^n = \sigma(-\tau\Delta_h)\Pi_h u(t^{n-1}) + \tau \sum_{i=1}^{m} p_i(-\tau\Delta_h)\Big(\Pi_h f(u) + g_h\Big)(t^n + c_i\tau).$$

Then the argument in Theorem 2.1.1 implies that

$$\|\Pi_h u(t^n) - w_h^n\|_h \leq C\tau^{k+1}\Big(\sup_{t^{n-1}\leq t\leq t^n} \|\Pi_h u^{(k+1)}(t)\|_h + \sup_{t^{n-1}\leq t\leq t^n} \|\Delta_h \Pi_h u^{(k)}(t)\|_h\Big).$$

The first term of the right hand side is bounded by $\|u\|_{C^{k+1}([0,T];C(\bar{\Omega}))}$, while the second one is bounded

as

$$\begin{aligned}\|\Delta_h\Pi_h u^{(k)}(t)\|_h &= \sup_{\varphi_h\in S_h^r}\frac{(\Delta_h\Pi_h u^{(k)}(t),\varphi_h)_h}{\|\varphi_h\|_h}\\ &= \sup_{\varphi_h\in S_h^r}\frac{(\nabla(\Pi_h u^{(k)}(t)-u^{(k)}(t)),\nabla\varphi_h)+(\nabla u^{(k)}(t),\nabla\varphi_h)}{\|\varphi_h\|_h}\\ &\le Ch^{-1}\|\nabla(\Pi_h u^{(k)}(t)-u^{(k)}(t))\|_{L^2(\Omega)}+\|\Delta u^{(k)}(t)\|_{L^2(\Omega)}\le C\|u^{(k)}\|_{H^2(\Omega)}.\end{aligned}$$

Therefore, we conclude that

$$\|\Pi_h u(t^n)-w_h^n\|_h\le C\tau^{k+1}\Big(\|u\|_{C^{k+1}([t^{n-1},t^n];C(\bar\Omega))}+\|u\|_{C^k([t^{n-1},t^n];H^2(\Omega))}\Big).$$

Then the simple triangle inequality leads to

$$\begin{aligned}\|\hat u_h^n-\Pi_h u(t^n)\|_h^2&\le\Big(\|\hat u_h^n-w_h^n\|_h+\|w_h^n-\Pi_h u(t^n)\|_h\Big)^2\\ &\le(1+C\tau)\|\hat u_h^n-w_h^n\|_h^2+C\tau^{2k+1}.\end{aligned}\tag{2.10}$$

Let $\rho_h^n=\hat u_h^n-w_h^n$ and $e_h^n=u_h^n-\Pi_h u(t^n)$, then ρ_h^n satisfies

$$\rho_h^n=\sigma(-\tau\Delta_h)e_h^{n-1}+I_1^n+I_2^n\tag{2.11}$$

where

$$I_1^n=\tau\sum_{i=1}^m p_i(-\tau\Delta_h)\Big(\sum_{j=1}^k L_j(t^{n-1}+c_i\tau)\Pi_h f(u_h^{n-j})-\Pi_h f(u(t^{n-1}+c_i\tau))\Big),$$

$$\text{and }\ I_2^n=-\tau\sum_{i=1}^m p_i(-\tau\Delta_h)g_h(t^{n-1}+c_i\tau).$$

Now take the discrete inner product between (2.11) and ρ_h^n

$$\|\rho_h^n\|_h^2=(\sigma(-\tau\Delta_h)e_h^{n-1},\rho_h^n)_h+(I_1^n,\rho_h^n)_h+(I_2^n,\rho_h^n)_h.$$

Then first term, we apply the Assumption (P4) to obtain that

$$\begin{aligned}(\sigma(-\tau\Delta_h)e_h^{n-1},\rho_h^n) &\le \frac{1}{2}\|\sigma(-\tau\Delta_h)e_h^{n-1}\|_h^2+\frac{1}{2}\|\rho_h^n\|_h^2\\ &\le \frac{1}{2}\|e_h^{n-1}\|_h^2-c_0\tau\|\nabla\sigma(-\tau\Delta_h)e_h^{n-1}\|^2+\frac{1}{2}\|\rho_h^n\|_h^2\\ &\le \frac{1}{2}\|e_h^{n-1}\|_h^2-c_0\tau\|\nabla(\rho_h^n-I_1^n-I_2^n)\|^2+||+\frac{1}{2}\|\rho_h^n\|_h^2\\ &\le \frac{1}{2}\|e_h^{n-1}\|_h^2-c_0\tau\|\nabla\rho_h^n\|^2-c_0\tau\|\nabla(I_1^n+I_2^n)\|^2\\ &\quad+2c_0\tau(\nabla\rho_h^n,\nabla(I_1^n+I_2^n))+\frac{1}{2}\|\rho_h^n\|^2\end{aligned}$$

Then applying the definition of Δ_h, we arrive at

$$\begin{aligned}\frac{1}{2}\|\rho_h^n\|_h^2 &\le \frac{1}{2}\|e_h^{n-1}\|_h^2-c_0\tau\|\nabla\rho_h^n\|^2\\ &\quad-2c_0\tau(\rho_h^n,\Delta_h(I_1^n+I_2^n))_h+(I_1^n,\rho_h^n)_h+(I_2^n,\rho_h^n)_h.\end{aligned}\tag{2.12}$$

By using the approximation property of interpolation I_τ^k, Lemma 2.2.2, and the fact that $u_h^{n-k},\dots,u_h^{n-1}$ satisfies the maximum bound, we bound the fourth term in (2.12) as

$$\begin{aligned}|(I_1^n,\rho_h^n)_h| &\le \tau\sum_{i=1}^m\Big|\Big(\sum_{j=1}^k L_j(t^{n-1}+c_i\tau)\Pi_h f(u(t^{n-j}))-\Pi_h f(u(t^{n-1}+c_i\tau),p_i(-\tau\Delta_h)\rho_h^n\Big)_h\Big|\\ &\quad+\tau\sum_{i=1}^m\Big|\Big(\sum_{j=1}^k L_j(t^{n-1}+c_i\tau)(\Pi_h f(u(t^{n-j}))-\Pi_h f(u_h^{n-j})),p_i(-\tau\Delta)\rho_h^n\Big)_h\Big|\\ &\le C\tau\sum_{i=1}^m\|p_i(-\tau\Delta_h)\rho_h^n\|_h\sum_{j=1}^k\|\Pi_h f(u(t^{n-j}))-\Pi_h f(u^{n-j})\|_h\\ &\quad+C\tau^{k+1}\sum_{i=1}^m\|p_i(-\tau\Delta_h)\rho_h^n\|_h\|\Pi_h f(u)\|_{C^k([t^{n-k},t^n];L^2(\Omega))}\\ &\le C\tau^{2k+1}\|\Pi_h f(u)\|^2_{C^k([t^{n-k},t^n];L^2(\Omega))}+C\tau\sum_{j=1}^k\|e_h^{n-j}\|_h^2+C\tau\|\rho_h^n\|_h^2\\ &\le C\tau^{2k+1}\|f(u)\|^2_{C^k([t^{n-k},t^n];C(\bar\Omega))}+C\tau\sum_{j=1}^k\|e_h^{n-j}\|_h^2+C\tau\|\rho_h^n\|_h^2.\end{aligned}$$

The fifth term in (2.12) can be bounded by using lemmas 2.2.2 and 2.2.3, i.e.,

$$\begin{aligned}|(I_2^n,\rho_h^n)_h| &\le C\tau\sum_{i=1}^m |(g_h(t^{n-1}+c_i\tau),p_i(-\tau\Delta_h)\rho_h^n)_h|\\ &\le C\tau\sum_{i=1}^m h^{r+1}\|u(t^{n-1}+c_i\tau)\|_{H^{2r+2}(\Omega)}\|p_i(-\tau\Delta_h)\rho_h^n\|_{H^1(\Omega)}\\ &\le \frac{C\tau h^{2r+2}}{\eta}\|u\|^2_{C([t^{n-1},t^n];H^{2r+2}(\Omega))}+C\tau\eta\|\rho_h^n\|^2_{H^1(\Omega)}.\end{aligned}\tag{2.13}$$

For the third term in the right hand side of (2.12), we shall apply the preceding argument again, together with the stability estimate (2.8), and obtain that

$$\begin{aligned}\tau|(\rho_h^n,\Delta_h(I_1^n+I_2^n))_h| &\le C\tau^2\sum_{i=1}^m\|\Delta_h p_i(-\tau\Delta_h)\rho_h^n\|_h\sum_{j=1}^k\|\Pi_h f(u(t^{n-j}))-\Pi_h f(u^{n-j})\|_h\\ &\quad+C\tau^{k+2}\sum_{i=1}^m\|\Delta_h p_i(-\tau\Delta_h)\rho_h^n\|_h\|\Pi_h f(u)\|_{C^k([t^{n-k},t^n];L^2(\Omega))}\\ &\quad+C\tau^2\sum_{i=1}^m h^{r+1}\|u(t^{n-1}+c_i\tau)\|_{H^{2r+2}(\Omega)}\|\Delta_h p_i(-\tau\Delta_h)\rho_h^n\|_{H^1(\Omega)}\\ &\le C\tau^{2k+1}\|f(u)\|^2_{C^k([t^{n-k},t^n];C(\bar\Omega))}+C\tau\sum_{j=1}^k\|e_h^{n-j}\|_h^2+C\tau\|\rho_h^n\|_h^2\\ &\quad+\frac{C\tau h^{2r+2}}{\eta}\|u\|^2_{C([t^{n-1},t^n];H^{2r+2}(\Omega))}+C\tau\eta\|\rho_h^n\|^2_{H^1(\Omega)}.\end{aligned}\tag{2.14}$$

Then by choosing η small, we arrive at

$$(1-C\tau)\|\rho_h^n\|_h^2\le\|e_h^{n-1}\|_h^2+C\tau\sum_{j=1}^k\|e_h^{n-j}\|_h^2+C\tau(\tau^{2k}+h^{2r+2}).$$

This together with (2.10) and the property of the cut-off operation lead to

$$\begin{aligned}\|e_h^n\|_h^2\le\|\hat u_h^n-\Pi_h u(t^n)\|_h^2&\le(1+C\tau)\|\rho_h^n\|_h^2+c\tau^{2k+1}\\ &\le\|e_h^{n-1}\|_h^2+C\tau\sum_{j=1}^k\|e_h^{n-j}\|_h^2+C\tau(\tau^{2k}+h^{2r+2}),\end{aligned}$$

and hence we rearrange terms and obtain

$$\frac{\|e_h^n\|_h^2 - \|e_h^{n-1}\|_h^2}{\tau} \leq C(\tau^{2k} + h^{2r+2}) + C\sum_{j=1}^{k} \|e_h^{n-j}\|_h^2.$$

Then the discrete Gronwall's inequality implies

$$\|e_h^n\|_h^2 \leq Ce^{cT}(\tau^{2k} + h^{2r+2}) + Ce^{cT}\sum_{j=0}^{k-1} \|e_h^j\|_h^2,$$

and the desired error estimate follows from the equivalence of different norms by Lemma 2.2.1. □

Remark 2.2.1. *In [73], an error estimate $O(\tau^k + h^r)$, which is suboptimal in space, was derived for the multistep exponential integrator method by using energy argument. The loss of the optimal convergence rate is due to the suboptimal estimate of the term $(\partial_x(\Pi_h u - u), \partial_x v_h)$ in [73, eq. (2.6) and (3.22)]. The optimal rate could be also proved by using Lemma 2.2.3.*

The Assumption (P4) , called L-stability, is useful when solving stiff problems. It is also essential in the proof of Theorem 2.2.1 to derive the optimal error estimate of the extrapolated cut-off single step scheme. In particular, Assumption (P4) immediately leads to the estimate

$$\|\sigma(-\tau\Delta_h)v_h\|_h^2 \leq \|v_h\|_h^2 - 2c_0\tau\|\nabla\sigma(-\tau\Delta_h)v_h\|^2,$$

where the second term in the right side is used to handle the term involving $\|\rho_h^n\|_{H^1(\Omega)}$ in (2.13) *and* (2.14)*. Many single step methods, e.g., Lobatto IIIC and Radau IIA methods are L-stable [31, 114]. For both classes, arbitrarily high-order methods can be constructed. Nevertheless, it is not clear how to remove the restriction (P4) in general.*

Remark 2.2.2. *It is straightforward to extend the argument to higher dimensional problems, e.g., Ω is a multi-dimensional rectangular domain $(a, b)^d \subset \mathbb{R}^d$, with $d \geq 2$. Then we can divide Ω in to some small sub-rectangles, called partition $\mathcal{K}$, and apply the tensor-product Lagrange finite elements on the partition $\mathcal{K}$. As a result, Lemma 2.2.3 is still valid, which implies the desired error estimate. See more details about the setting for multi-dimensional problems in [73, Section 2.2].*

2.3 Collocation-type Methods with the Cut-off Postprocessing

Note that the Assumption (P4) excludes some popular methods, e.g., Gauss–Legendre methods. This motivates us to discuss the collocation-type schemes, which belong to implicit Runge–Kutta methods, and derive error estimate without Assumption (P4). This class of time stepping methods is easy to implement, and plays an essential role in the next section to develop an energy-stable scheme. For simplicity, we only present the argument for one-dimensional case, and it can be extended to multi-dimensional cases straightforwardly as mentioned in Remark 2.2.2.

a_{11}	$\dots$	a_{1m}	c_1
$\vdots$		$\vdots$	$\vdots$
a_{m1}	$\dots$	a_{mm}	c_m
b_1	$\dots$	b_m	

Table 2.1: Butcher tableau for Runge–Kutta scheme.

Now we consider an m-stage Runge–Kutta method, described by the Butcher tableau 2.1. Here $\{c_i\}_{i=1}^m$ denotes m distinct quadrature points.

Definition 2.3.1. *We call a Runge–Kutta method is algebraically stable if the method satisfies*

(P5)(a) *The matrix $A = (a_{ij})$, with $i, j = 1, \dots, m$ is invertible;*

(P5)(b) *The coefficients b_i satisfy $b_i > 0$ for $i = 1, 2, \dots, m$;*

(P5)(c) *The symmetric matrix $\mathcal{M} \in \mathbb{R}^{m\times m}$ with entries $m_{ij} := b_i a_{ij} + b_j a_{ji} - b_i b_j$, $i, j = 1, \dots, m$ is positive semidefinite.*

Here we assume that the Runge–Kutta scheme described by Table 2.1 associates with a collocation method, i.e., coefficients a_{ij}, b_i, c_i satisfy

$$\sum_{i=1}^{m} b_i c_i^{l-1} = \frac{1}{l}, \quad l = 1, \cdots, p, \tag{2.15}$$

$$\sum_{j=1}^{m} a_{ij} c_j^{l-1} = \frac{c_i^l}{l}, \quad l = 1, \cdots, m, \tag{2.16}$$

with some integers $p \geq m$. Two popular families of algebraically stable Runge–Kutta methods of collocation type satisfying (2.6) of orders $p = 2m$ and $p = 2m - 1$ are the Gauss–Legendre methods and the

Radau IIA methods respectively. For both classes, arbitrarily high order methods can be constructed, and both of them follow this analysis. Note that the Gauss–Legendre methods are not L-stable [114].

In particular, at level n, with given $u_h^{n-k}, \dots, u_h^{n-1} \in S_h^r$, we find an intermediate solution $\hat{u}_h^n \in S_h^r$ such that

$$\begin{cases} \dot{u}_h^{ni} = \Delta_h u_h^{ni} + \sum_{\ell=1}^{k} L_\ell(t^{n-1} + c_i\tau)\Pi_h f(u_h^{n-\ell}) & \text{for } i = 1, 2, \dots, m, \\ u_h^{ni} = u_h^{n-1} + \tau \sum_{j=1}^{m} a_{ij}\dot{u}_h^{nj} & \text{for } i = 1, 2, \dots, m, \\ \hat{u}_h^n = u_h^{n-1} + \tau \sum_{i=1}^{m} b_i \dot{u}_h^{ni}, \end{cases} \tag{2.17}$$

where $k = \min(p, m+1)$, and $\Pi_h : C(\overline{\Omega}) \to S_h^r$ is the Lagrange interpolation operator. Then we apply the cut-off operation: find $u_h^n \in S_h^r$ such that

$$u_h^n(x_j) = \min\big(\max\big(\hat{u}_h^n(x_j), -\alpha\big), \alpha\big), \quad j = 0, \dots, Mr. \tag{2.18}$$

Remark 2.3.1. *Note that the scheme* (2.17) *is equivalent to* (2.5) *with*

$$(p_1(\lambda), \dots, p_m(\lambda)) = (b_1, \dots, b_m)(I + \lambda A)^{-1}, \qquad \sigma(\lambda) = 1 - \lambda \sum_{j=1}^{m} b_j p_j(\lambda).$$

Then the Assumption (P5), and (2.15)-(2.16) *imply Assumptions (P1), (P2) with order* $k = \min(p, m+1)$ *and (P3) with order* $q = \min(p, m+1)$. *Hence Theorem 2.2.1 indicates the temporal error* $O(\tau^{\min(p,m+1)})$. *This is the reason why we choose* k*-step extrapolation, where* $k = \min(p, m+1)$*, in the time stepping scheme* (2.17).

Next, we shall derive an error estimate for the fully discrete scheme (2.17)-(2.18). To begin with, we shall examine the local truncation error. We define the local truncation error η_{ni} and η_{n+1} as

$$\begin{cases} \dot{u}_*^{ni} = \Delta u(t^{ni}) + \sum_{\ell=1}^{k} L_\ell(t^{ni}) f(u(t^{n-\ell})) & \text{for } i = 1, 2, \dots, m, \\ u(t^{ni}) = u(t^{n-1}) + \tau \sum_{j=1}^{m} a_{ij}\dot{u}_*^{nj} + \eta_{ni} & \text{for } i = 1, 2, \dots, m, \\ u(t^n) = u(t^{n-1}) + \tau \sum_{i=1}^{m} b_i \dot{u}_*^{ni} + \eta_n \end{cases} \tag{2.19}$$

where $t^{ni} = t^{n-1} + c_i\tau$ and $k = \min(p, q+1)$. Then the next lemma give an estimate for the local

truncation error η_{ni} and η_n. We sketch the proof in Appendix for completeness.

Lemma 2.3.1. *Suppose that the Assumption (P5), and relations* (2.15) *and* (2.16) *are valid. Then the local truncation error η_{ni} and η_n, given by* (2.19)*, satisfy the estimate*

$$\|\eta_n\|_{H^1(\Omega)} + \tau \sum_{i=1}^{m} \|\eta_{ni}\|_{H^1(\Omega)} \leq C\tau^{k+1}.$$

with $k = \min(p, q+1)$, provided that $u \in C^{k+1}([0,T]; H^1(\Omega))$ and $f(u) \in C^k([0,T]; H^1(\Omega))$.

Proof. We note that the second relation in equation (2.19) implies

$$u(t^{ni}) - u(t^{n-1}) - \tau \sum_{j=1}^{m} a_{ij} u_t^{nj} = \tau \sum_{j=1}^{m} a_{ij} (\dot{u}_*^{nj} - u_t(t^{nj})) + \eta_{ni} \quad \text{for } i = 1, 2, \ldots, m.$$

Then we substitute the first relation of (2.19) and derive that for $i = 1, 2, \ldots, m$

$$u(t^{ni}) - u(t^{n-1}) - \tau \sum_{j=1}^{m} a_{ij} u_t^{nj} = \tau \sum_{j=1}^{m} a_{ij} \Big(\sum_{\ell=1}^{k} L_\ell(t^{n-1} + c_j\tau) f(u(t^{n-\ell})) - f(t^{nj}) \Big) + \eta_{ni}.$$

Define $\tilde{\eta}_{ni}$ as the left hand side of the above relation. Now we apply Taylor's expansion at t^{n-1} and use (2.16) to derive

$$\begin{aligned}
\tilde{\eta}_{ni} =& \sum_{l=1}^{m} \frac{\tau^l}{(l-1)!} \left(\frac{c_i^l}{l} - \sum_{j=1}^{m} a_{ij} c_j^{l-1} \right) u^{(\ell)}(t^n) + \frac{1}{m!} \int_{t_{n-1}}^{t^{ni}} (t^{ni} - s)^m u^{(m+1)}(s) \mathrm{d}s \\
&+ \frac{\tau}{(m-1)!} \sum_{j=1}^{m} a_{ij} \int_{t^{n-1}}^{t^{nj}} (t^{nj} - s)^{m-1} u^{(m+1)}(s) \mathrm{d}s \\
&= \frac{1}{m!} \int_{t^{n-1}}^{t^{ni}} (t^n - s)^m u^{(m+1)}(s) \mathrm{d}s + \frac{\tau}{(m-1)!} \sum_{j=1}^{m} a_{ij} \int_{t^{n-1}}^{t^{nj}} (t^{nj} - s)^{m-1} u^{(m+1)}(s) \mathrm{d}s
\end{aligned}$$

Then we obtain the estimate for $\tilde{\eta}_{ni}$, with $i = 1, 2, \ldots, m$, that

$$\|\tilde{\eta}_{ni}\|_{H^1(\Omega)} \leq C\tau^{m+1} \|u^{(m+1)}\|_{C([t^{n-1}, t^n]; H^1(\Omega))}.$$

This together with the approximation property of Lagrange interpolation lead to

$$\|\eta_{ni}\|_{H^1(\Omega)} \leq C\Big(\tau^{k+1} \|f(u)\|_{C^k([t^{n-k}, t^n]; H^1(\Omega))} + \tau^{m+1} \|u\|_{C^{(m+1)}([t^{n-1}, t^n]; H^1(\Omega))} \Big).$$

for $i = 1, 2, \ldots, m$. Similarly, we have

$$u(t^n) - u(t^{n-1}) - \tau \sum_{i=1}^{m} b_i u_t^{ni} = \tau \sum_{i=1}^{m} b_i \Big(\sum_{\ell=1}^{k} L_\ell(t^{n-1} + c_i\tau) f(u(t^{n-\ell})) - f(t^{ni}) \Big) + \eta_n.$$

Take the left hand side as $\tilde{\eta_n}$. Then Taylor expansion and (2.15) imply

$$\tilde{\eta}_n = \frac{1}{p!} \int_{t^{n-1}}^{t^n} (t^n - s)^p u^{(p+1)}(s) \mathrm{d}s + \frac{\tau}{(p-1)!} \sum_{i=1}^{m} b_i \int_{t^{n-1}}^{t^{ni}} (t^{ni} - s)^{p-1} u^{(p+1)}(s) \mathrm{d}s.$$

This together with the approximation property of Lagrange interpolation leads to

$$\|\eta_{ni}\|_{H^1(\Omega)} \leq C \Big(\tau^{k+1} \|f(u)\|_{C^k([t^{n-k}, t^n]; H^1(\Omega))} + \tau^{p+1} \|u\|_{C^{p+1}([t^{n-1}, t^n]; H^1(\Omega))} \Big).$$

Using the choice that $k = \min(p, m+1)$, we derive the desired result. □

Then we are ready to present the following theorem, which gives the error estimate for the cut-off Runge–Kutta scheme (2.17)-(2.18).

Theorem 2.3.1. *Suppose that the Runge–Kutta method given by Table 2.1 satisfies Assumption (P5), and relations* (2.15) *and* (2.16) *are valid. Assume that* $|u_0| \leq \alpha$ *and the maximum principle* (1.2) *holds, and assume that the starting values* u_h^n, $l = 0, \ldots, k-1$, *are given and*

$$|u_h^l(x_j)| \leq \alpha, \quad j = 0, \ldots, Mr, \quad l = 0, \ldots, k-1.$$

Then the fully discrete solution given by (2.17)-(2.18) *satisfies*

$$|u_h^n(x_j)| \leq \alpha, \quad j = 0, \ldots, Mr, \quad n = k, \ldots, N,$$

and for $n = k, \ldots, N$

$$\|u(t^n) - u_h^n\|_{L^2(\Omega)} \leq C(\tau^k + h^{r+1}) + C \sum_{l=0}^{k-1} \|u(t^l) - u_h^l\|_{L^2(\Omega)},$$

provided that $u \in C^{k+1}([0,T]; H^1(\Omega)) \cap C^1([0,T]; H^{2r+2}(\Omega))$, f *is locally Lipschitz continuous and* $f(u) \in C^k([0,T]; H^1(\Omega)) \cap C([0,T]; H^{2r+2}(\Omega))$.

Proof. Due to the cut-off operation (2.3), the discrete maximum bound principle follows immediately. With the notation

$$e_h^{ni} = \Pi_h u(t^{ni}) - u_h^{ni}, \quad \dot{e}_h^{ni} = \Pi_h \dot{u}_*^{ni} - \dot{u}_h^{ni}, \quad e_h^n = \Pi_h u(t^n) - u_h^n, \quad \hat{e}_h^n = \Pi_h u(t^n) - \hat{u}_h^n,$$

we derive the error equations

$$\begin{cases} \dot{e}_h^{ni} = \Delta_h e_h^{ni} + (\Pi_h \Delta - \Delta_h \Pi_h) u(t^{ni}) + \sum_{\ell=1}^k L_\ell(t^{ni}) \Pi_h (f(u(t^{n-\ell})) - f(u_h^{n-\ell})) & \text{for } i = 1, 2, \ldots, m, \\ e_h^{ni} = e_h^{n-1} + \tau \sum_{j=1}^m a_{ij} \dot{e}_h^{nj} + \Pi_h \eta_{ni} & \text{for } i = 1, 2, \ldots, m, \\ \hat{e}_h^n = e_h^{n-1} + \tau \sum_{i=1}^m b_i \dot{e}_h^{ni} + \Pi_h \eta_n. \end{cases} \tag{2.20}$$

Take the square of discrete L^2 norm of both sides of the last relation of (2.20), we obtain

$$\|\hat{e}_h^n\|_h^2 = \|e_h^{n-1} + \tau \sum_{i=1}^m b_i \dot{e}_h^{ni}\|_h^2 + 2(\eta_n, e_h^{n-1} + \tau \sum_{i=1}^m b_i \dot{e}_h^{ni})_h + \|\Pi_h \eta_n\|_h^2. \tag{2.21}$$

For the first term on the right hand side, we expand it and apply the second equation of (2.20) to obtain

$$\begin{aligned} \|e_h^{n-1} + \tau \sum_{i=1}^m b_i \dot{e}_h^{ni}\|_h^2 &= \|e_h^{n-1}\|_h^2 + 2\tau \sum_{i=1}^m b_i (\dot{e}_h^{ni}, e_h^{ni} - \eta_{ni})_h - \tau^2 \sum_{i,j=1}^m m_{ij} (\dot{e}_h^{ni}, \dot{e}_h^{nj})_h \\ &\le \|e_h^{n-1}\|_h^2 + 2\tau \sum_{i=1}^m b_i (\dot{e}_h^{ni}, e_h^{ni} - \eta_{ni})_h, \end{aligned}$$

where in the last inequality we use the positive semi-definiteness of the matrix $\mathcal{M}$ in the Assumption (P5). Next, we note that the first relation of (2.20) implies

$$\begin{aligned} (\dot{e}_h^{ni}, e_h^{ni} - \eta_{ni})_h &= \Big(\Delta_h e_h^{ni} + \sum_{\ell=1}^k L_\ell(t^{ni})(f(u(t^{n-\ell})) - f(u_h^{n-\ell})) + (\Pi_h \Delta - \Delta_h \Pi_h) u(t^{n-1}), e_h^{ni} - \eta_{ni}\Big)_h \\ &= -\|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 + (\nabla e_h^{ni}, \nabla \Pi_h \eta_{ni}) + \Big(\sum_{\ell=1}^k L_\ell(t^{ni})(f(u(t^{n-\ell})) - f(u_h^{n-\ell})), e_h^{ni} - \eta_{ni}\Big)_h \\ &\quad + \Big((\Pi_h \Delta - \Delta_h \Pi_h) u(t^{n-1}), e_h^{ni} - \eta_{ni}\Big)_h \end{aligned}$$

The bound of second term of the right hand side can be derived via Cauchy-Schwarz inequality

$$|(\nabla e_h^{ni}, \nabla \Pi_h \eta_{ni})| \le \frac{1}{4}\|\nabla e_h^{ni}\|^2_{L^2(\Omega)} + C\|\eta_{ni}\|^2_{H^1(\Omega)}.$$

Meanwhile, using the fact that f is locally Lipschitz and the fully disctete solutions satisfy maximum bound principle at the Gauss–Lobatto points, the third term can be bounded as

$$\Big(\sum_{\ell=1}^{k} L_\ell(t^{ni})(f(u(t^{n-\ell})) - f(u_h^{n-\ell})), e_h^{ni} - \eta_{ni}\Big)_h \le C\Big(\|e_h^{ni}\|_h^2 + \|\eta_{ni}\|^2_{H^1(\Omega)} + \sum_{\ell=1}^{k}\|e_h^{n-\ell}\|_h^2\Big)$$

The bound of the last term follows from Lemma 2.2.3

$$\begin{aligned}\Big((\Pi_h\Delta - \Delta_h\Pi_h)u(t^{n-1}), e_h^{ni} - \eta_{ni}\Big)_h &\le Ch^{r+1}\|e_h^{ni} - \Pi_h\eta_{ni}\|_{H^1(\Omega)} \\ &\le \frac{1}{4}\|\nabla e_h^{ni}\|^2_{L^2(\Omega)} + C(\|e_h^{ni}\|_h^2 + \|\eta_{ni}\|^2_{H^1(\Omega)} + h^{2r+2}).\end{aligned}$$

Therefore, we arrive at

$$2(\dot e_h^{ni}, e_h^{ni} - \eta_{ni})_h \le -\|\nabla e_h^{ni}\|^2_{L^2(\Omega)} + C\Big(\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2 + \|e_h^{ni}\|_h^2 + \|\eta_{ni}\|^2_{H^1(\Omega)} + h^{2r+2}\Big),$$

and hence by Lemma 2.3.1, we derive

$$\begin{aligned}\|e_h^{n-1} + \tau\sum_{i=1}^{m} b_i \dot e_h^{ni}\|_h^2 \le \|e_h^{n-1}\|_h^2 &- \tau\sum_{i=1}^{m} b_i\|\nabla e_h^{ni}\|^2_{L^2(\Omega)} + C\tau\sum_{i=1}^{m}\|e_h^{ni}\|_h^2 \\ &+ C\tau\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2 + C\tau(h^{2r+2} + \tau^{2k}).\end{aligned}$$

In view of the first relation of the error equation (2.20), we have the estimate

$$\begin{aligned}(\eta_n, e_h^{n-1} + \tau\sum_{i=1}^{m} b_i\dot e_h^{ni})_h &\le \|\eta_n\|_{H^1(\Omega)}\Big(\|e_h^{n-1}\|_h + C\tau\sum_{i=1}^{m} b_i\Big(\|\nabla e_h^{ni}\|_h + \sum_{j=1}^{k}\|e_h^{n-j}\|_h + h^{2r+2}\Big)\Big) \\ &\le C\tau(h^{2r+2} + \tau^{2k}) + \frac{\tau}{4}\sum_{i=1}^{m} b_i\|\nabla e_h^{ni}\|_h^2 + C\tau\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2\end{aligned}$$

which gives a bound of the second term in (2.21). In conclusion, we obtain that

$$\|\hat{e}_h^n\|_h^2 + \frac{\tau}{2}\sum_{i=1}^m \|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 \le C\tau(h^4+\tau^{2k}) + \|e_h^{n-1}\|_h^2 + C\tau\sum_{i=1}^m \|e_h^{ni}\|_h^2 + C\tau\sum_{j=1}^k \|e_h^{n-j}\|_h^2. \tag{2.22}$$

Next, we shall derive a bound for $\sum_{i=1}^m \|e_h^{ni}\|_h^2$ on the right-hand side. To this end, we test the second relation of (2.20) by e_h^{ni}. This yields

$$\begin{aligned}\sum_{i=1}^m \|e_h^{ni}\|_h^2 &\le C\|e_h^{n-1}\|_h^2 + C\tau\sum_{i,j=1}^m a_{ij}(\dot{e}_h^{nj}, e_h^{ni})_h + C\sum_{i=1}^m \|\Pi_h\eta_{ni}\|_h^2 \\ &\le C\|e_h^{n-1}\|_h^2 + C\tau\sum_{i,j=1}^m a_{ij}(\dot{e}_h^{nj}, e_h^{ni})_h + C\tau^{2k}.\end{aligned}$$

Then, we apply the first relation of (2.20) and Lemma 2.2.3 to derive

$$\begin{aligned}\sum_{i,j=1}^m a_{ij}(\dot{e}_h^{nj}, e_h^{ni})_h &= -\sum_{i,j=1}^m a_{ij}(\nabla e_h^{nj}, \nabla e_h^{ni}) + \sum_{i,j=1}^m a_{ij}\Big(\sum_{\ell=1}^k L_\ell(t^{ni})(f(u(t^{n-\ell})) - f(u_h^{n-\ell})), e_h^{ni}\Big)_h \\ &\quad + \sum_{i,j=1}^m a_{ij}((\Pi_h\Delta - \Delta_h\Pi_h)u(t^{n-1}), e_h^{ni})_h \\ &\le C\sum_{i=1}^m \Big(\|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 + \|e_h^{ni}\|_h^2\Big) + Ch^{2r+2} + C\sum_{\ell=1}^k \|e_h^{n-\ell}\|_h^2.\end{aligned}$$

Therefore, we obtain

$$\sum_{i=1}^m \|e_h^{ni}\|_h^2 \le C(\tau h^{2r+2}+\tau^{2k}) + C\|e_h^{n-1}\|_h^2 + C\tau\sum_{\ell=1}^k \|e_h^{n-\ell}\|_h^2 + C\tau\sum_{i=1}^m \Big(\|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 + \|e_h^{ni}\|_h^2\Big).$$

Then for sufficiently small τ, $C\tau\sum_{i=1}^m \|e_h^{ni}\|_h^2$ on the right-hand side can be absorbed by the left-hand side. Then, we obtain

$$\sum_{i=1}^m \|e_h^{ni}\|_h^2 \le C(\tau h^{2r+2}+\tau^{2k}) + C\|e_h^{n-1}\|_h^2 + C\tau\sum_{\ell=1}^k \|e_h^{n-\ell}\|_h^2 + C\tau\sum_{i=1}^m \|\nabla e_h^{ni}\|_{L^2(\Omega)}^2.$$

Now substituting the above estimate into (2.22), there holds for sufficiently small τ

$$\|\hat{e}_h^n\|_h^2 \leq C\tau(h^{2r+2}+\tau^{2k}) + \|e_h^{n-1}\|_h^2 + C\tau\sum_{\ell=1}^{k}\|e_h^{n-\ell}\|_h^2.$$

Noting that $\|e_h^n\|_h \leq \|\hat{e}_h^n\|_h$ and rearranging terms, we obtain

$$\frac{\|e_h^n\|_h^2 - \|e_h^{n-1}\|_h^2}{\tau} \leq C(h^{2r+2}+\tau^{2k}) + C\sum_{\ell=1}^{k}\|e_h^{n-\ell}\|_h^2.$$

Then the discrete Gronwall's inequality implies

$$\max_{k\leq n\leq N}\|e_h^n\|_h^2 \leq C(h^{2r+2}+\tau^{2k}) + C\sum_{j=0}^{k-1}\|e_h^j\|_h^2.$$

This completes the proof of the theorem. □

Remark 2.3.2. *In Theroem 2.3.1, we discuss the algebraically stable collocation-type method with cut-off technique. We still prove the optiaml error estimate $O(\tau^k+h^{r+1})$, without the L-stability, i.e. Assumption (P4). Note that this class of methods includes Gauss–Legendre and Radau IIA methods [114, Theorem 12.9], while the first one is not L-stable [114, Table 5.13].*

2.4 Fully Discrete Scheme Based on Scalar Auxiliary Variable Method

In the preceding section, we develop and analyze a class of maximum bound preserving schemes. Unfortunately, the proposed scheme (with relatively large time steps) might produce solutions with increasing and oscillating energy, see Figure 2.2. This violates another essential property of the Allen–Cahn model, say energy dissipation. The aim for this section is to develop a high-order time stepping schemes via combining the cut-off strategy and the scalar auxiliary variable (SAV) method.

SAV method is a common-used method for gradient flow models. It was firstly developed in [101, 100] and have motived a sequence of interesting work on the development and analysis of high-order energy-decayed time stepping scheme in recent years [1, 99, 44].

In particular, assuming that $E_1(u(t)) = \int_\Omega F(u(x,t))\mathrm{d}x$ is globally bounded from below, i.e., $E_1(u(t)) >$

$-C_0$. we introduce the following scalar auxiliary variable [101]

$$z(t) = \sqrt{E_1(u(t)) + C_0} \quad \text{and} \quad W(u) = \frac{f(u)}{\sqrt{E_1(u) + C_0}} \tag{2.23}$$

Then the Allen–Cahn equation in (1.1) can be reformulated as

$$\begin{cases} u_t = \Delta u + z(t)W(u) & \text{in } \Omega \times (0,T), \\ u(x, t=0) = u_0(x) & \text{in } \Omega \times \{0\}, \\ \partial_{\mathbf{n}} u = 0 & \text{on } \partial\Omega \times (0,T) \end{cases} \tag{2.24}$$

and the scalar auxiliary variable $z(t)$ satisfies

$$\begin{cases} z'(t) = -\dfrac{1}{2}(W(u(t)), u_t(t)), & \text{in } (0,T), \\ z(0) = \sqrt{E_1(u^0) + C_0}. \end{cases} \tag{2.25}$$

One can easily show that the coupled problem (2.24)-(2.25) is equivalent to the original equation (1.1). Meanwhile, simple calculation leads to the SAV energy dissipation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{1}{2}\|\nabla u\|^2 + |z(t)|^2\Big) = -\|u_t(t)\|^2 \le 0. \tag{2.26}$$

Inspired by [1], we discretize the coupled problem (2.24)-(2.25) by using the m-stage Runge–Kutta method in time (described by Table 2.1) and lumped mass finite element method with $r = 1$ in space discretization. Then the cut-off operation is applied in each time level to remove the value violating the maximum bound principle (at nodal points). For simplicity, we only present the argument for one-dimensional case, and it can be extended to multi-dimensional cases straightforwardly as mentioned in Remark 2.2.2.

Here we assume that the m-stage Runge–Kutta method (described by Table 2.1) satisfies the Assumption (P5) and relations (2.15) and (2.16). Then at n-th time level, with known $u_h^{n-k}, \ldots, u_h^{n-1} \in S_h^r$ and

$z^{n-1} \in \mathbb{R}$, we find $\hat{u}_h^n \in S_h^r$ and $z^n \in \mathbb{R}$ such that

$$\begin{cases} \dot{u}_h^{ni} = \Delta_h u_h^{ni} + z^{ni} W_h^{ni} & \text{for } i = 1, 2, \dots, m, \\ u_h^{ni} = u_h^{n-1} + \tau \sum_{j=1}^m a_{ij} \dot{u}_h^{nj} & \text{for } i = 1, 2, \dots, m, \\ \hat{u}_h^n = u_h^{n-1} + \tau \sum_{i=1}^m b_i \dot{u}_h^{ni}, \end{cases} \tag{2.27}$$

and

$$\begin{cases} \dot{z}^{ni} = -\dfrac{1}{2}(W_h^{ni}, \dot{u}_h^{ni})_h & \text{for } i = 1, 2, \dots, m, \\ z^{ni} = z^{n-1} + \tau \sum_{j=1}^m a_{ij} \dot{z}^{nj} & \text{for } i = 1, 2, \dots, m, \\ z^n = z^{n-1} + \tau \sum_{i=1}^m b_i \dot{z}^{ni}, \end{cases} \tag{2.28}$$

where $\Pi_h : C(\overline{\Omega}) \to S_h^r$ is the Lagrange interpolation operator, and the linearized term W^{ni} is defined by

$$W_h^{ni} = \sum_{\ell=1}^{k} L_\ell(t^{n-1} + c_i\tau)\Pi_h W(u_h^{n-j}), \quad \text{with } k = \min(p, m+1).$$

Then we apply the cut-off operation: find $u_h^n \in S_h^r$ such that

$$u_h^n(x_j) = \min\big(\max\big(\hat{u}_h^n(x_j), -\alpha\big), \alpha\big), \quad j = 0, \dots, Mr. \tag{2.29}$$

Lemma 2.4.1. *For $r = 1$, the cut-off operation* (2.29) *indicates*

$$\|\nabla u_h^n\|_{L^2(\Omega)} \le \|\nabla \hat{u}_h^n\|_{L^2(\Omega)}. \tag{2.30}$$

Proof. Since both $\hat{u}_h^n$ and u_h^n are piecewise linear, it is easy to see that

$$\|\nabla u_h^n\|_{L^2(\Omega)}^2 = \frac{1}{h}\sum_{j=1}^{M} |u_h^n(x_j) - u_h^n(x_{j-1})|^2, \quad \|\hat{u}_h^n\|_{L^2(\Omega)}^2 = \frac{1}{h}\sum_{j=1}^{M} |\hat{u}_h^n(x_j) - \hat{u}_h^n(x_{j-1})|^2.$$

Obviously, the cut-off operation (2.29) derives

$$|u_h^n(x_j) - u_h^n(x_{j-1})| \le |\hat{u}_h^n(x_j) - \hat{u}_h^n(x_{j-1})|, \quad \text{for } j = 1, 2 \cdots, M,$$

which completes the proof. □

The next theorem shows that the cut-off SAV-RK scheme (2.27)-(2.29) satisfies the energy decay property and discrete maximum bound principle.

Theorem 2.4.1. *Suppose that the Runge–Kutta method in Table 2.1 satisfies Assumption (P5), and we apply the lumped mass finite element method with* $r = 1$ *in space discretization. Then, the time stepping scheme* (2.27)-(2.29) *satisfies the energy decay property:*

$$\frac{1}{2}\|\nabla u_h^n\|_{L^2(\Omega)}^2 + |z^n|^2 \le \frac{1}{2}\|\nabla u_h^{n-1}\|_{L^2(\Omega)}^2 + |z^{n-1}|^2, \quad \textit{for all } \ n \ge k. \tag{2.31}$$

Meanwhile, the fully discrete solution (2.27)-(2.29) *satisfies the maximum bound principle*

$$\max_{k\le n\le N} |u_h^n(x)| \le \alpha, \quad \textit{for all } \ x \in \Omega. \tag{2.32}$$

Proof. Due to the cut-off operation in each time level, we know that

$$\max_{k\le n\le N} |u_h^n(x_j)| \le \alpha, \quad \text{for all } \ j = 0, 1, \ldots, M.$$

Since the finite element function is piecewise linear, then for any $x \in (x_{j-1}, x_j)$

$$|u_h^n(x)| \le \max\left(|u_h^n(x_{j-1})|, |u_h^n(x_j)|\right) \le \alpha.$$

Next, we turn to the energy decay property (2.31). According to the third relation of (2.27), we have

$$\nabla \hat{u}_h^n = \nabla u_h^{n-1} + \tau \sum_{i=1}^{m} b_i \nabla \dot{u}_h^{ni}.$$

Squaring the discrete L^2-norms of both sides, yields

$$\|\nabla \hat{u}_h^n\|^2 = \|\nabla u_h^{n-1}\|^2 + 2\tau \sum_{i=1}^{m} b_i (\nabla \dot{u}_h^{ni}, \nabla u_h^{n-1}) + \tau^2 \sum_{i,j=1}^{m} b_i b_j (\nabla \dot{u}_h^{ni}, \nabla \dot{u}_h^{nj}).$$

By the second relation in (2.27), we arrive at

$$\begin{aligned}\|\nabla\hat{u}_h^n\|^2 &= \|\nabla u_h^{n-1}\|^2 + 2\tau\sum_{i=1}^m b_i(\nabla\dot{u}_h^{ni}, \nabla u_h^{ni} - \tau\sum_{j=1}^m a_{ij}\nabla\dot{u}_h^{ni}) + \tau^2\sum_{i,j=1}^m b_i b_j(\nabla\dot{u}_h^{ni}, \nabla\dot{u}_h^{nj})\\ &= \|\nabla u_h^{n-1}\|^2 + 2\tau\sum_{i=1}^m b_i(\nabla\dot{u}_h^{ni}, \nabla u_h^{ni}) - \tau^2\sum_{i,j=1}^m m_{ij}(\nabla\dot{u}_h^{ni}, \nabla\dot{u}_h^{nj})\\ &\le \|\nabla u_h^{n-1}\|^2 + 2\tau\sum_{i=1}^m b_i(\nabla\dot{u}_h^{ni}, \nabla u_h^{ni}),\end{aligned}$$

where we apply the Assumption (P4) in the last inequality. Then we apply the first relation in (2.27) to derive

$$\|\nabla\hat{u}_h^n\|^2 = \|\nabla u_h^{n-1}\|^2 - 2\tau\sum_{i=1}^m b_i\|\dot{u}_h^{ni}\|^2 + 2\tau\sum_{i=1}^m b_i z^{ni}(\dot{u}_h^{ni}, W_h^{ni})_h$$

On the other hand, the similar argument also leads to

$$|z^n|^2 \le |z^{n-1}|^2 - \tau\sum_{i=1}^m b_i z^{ni}(\dot{u}_h^{ni}, W_h^{ni})_h$$

Therefore we conclude that

$$\frac{1}{2}\|\nabla\hat{u}_h^n\|_h^2 + |z^n|^2 \le \frac{1}{2}\|\nabla u_h^{n-1}\|_h^2 + |z^{n-1}|^2 - \tau\sum_{i=1}^m b_i\|\dot{u}_h^{ni}\|_h^2 \le \frac{1}{2}\|\nabla u_h^{n-1}\|_h^2 + |z^{n-1}|^2.$$

which together with (2.30) implies the desired energy decay property immediately. □

Remark 2.4.1. *Note that the energy dissipation law holds valid only if $r = 1$, since in this case the cut-off operation does not enlarge the H^1 semi-norm, which is present as* (2.30) *in Lemma 2.4.1. This property is not clear for finite element method with high degree polynomials. Hence, how to design an spatially high-order (unconditionally) energy dissipative and maximum bound preserving scheme is still unclear and warrants further investigation.*

Next, we shall derive an error estimate for the fully discrete scheme (2.27)-(2.29). To begin with, we

shall examine the local truncation error. We define the local truncation error η_{ni} and η_n as

$$\begin{cases} \dot{u}_*^{ni} = \Delta u(t^{ni}) + z(t^{ni}) W_*^{ni} & \text{for } i = 1, 2, \dots, m, \\ u(t^{ni}) = u(t^{n-1}) + \tau \sum_{j=1}^{m} a_{ij} \dot{u}_*^{nj} + \eta_{ni} & \text{for } i = 1, 2, \dots, m, \\ u(t^n) = u(t^{n-1}) + \tau \sum_{i=1}^{m} b_i \dot{u}_*^{ni} + \eta_n \end{cases} \tag{2.33}$$

where $t^{ni} = t^{n-1} + c_i \tau$ and W_*^{ni} denotes the extrapolation

$$W_*^{ni} = \sum_{\ell=1}^{m} L_\ell(t^{n-1} + c_i \tau) W(u(t^{n-j})).$$

Similarly, we define d_{ni} and d_n as

$$\begin{cases} \dot{z}_*^{ni} = -\dfrac{1}{2}(W_*^{ni}, \dot{u}_*^{ni}) & \text{for } i = 1, 2, \dots, m, \\ z(t^{ni}) = z(t^{n-1}) + \tau \sum_{j=1}^{m} a_{ij} \dot{z}_*^{nj} + d_{ni} & \text{for } i = 1, 2, \dots, m, \\ z(t^n) = z(t^{n-1}) + \tau \sum_{i=1}^{m} b_i \dot{z}_*^{ni} + d_n, \end{cases} \tag{2.34}$$

Provided the assumption (P5) and relations (2.15) and (2.16), the local truncation errors η_{ni}, η_n, d_{ni}, d_n satisfy the estimate

$$\|\eta_n\|_{H^1(\Omega)} + |d_n| + \tau \sum_{i=1}^{m} \Big(\|\eta_{ni}\|_{H^1(\Omega)} + |d_{ni}| \Big) \le C \tau^{k+1}. \tag{2.35}$$

We omit the proof, since it is similar to the one of Lemma 2.3.1, given in Appendix. See also [1, Lemma 3.1].

Theorem 2.4.2. *Suppose that the Runge–Kutta method satisfies Assumption (P4) and the relations* (2.15) *and* (2.16)*. Assume that* $|u_0| \le \alpha$ *and the maximum principle* (1.2) *holds, and assume that the starting values* u_h^l *and* z^l*,* $l = 0, \dots, k-1$*, are given and*

$$|u_h^l(x_j)| \le \alpha, \quad j = 0, \dots, M, \quad l = 0, \dots, k-1.$$

Then the fully discrete solution given by (2.27)-(2.29) *satisfies for* $n = k, \ldots, N$

$$\|u(t^n) - u_h^n\|_{L^2(\Omega)} \leq C(\tau^k + h^2) + C\sum_{l=0}^{k-1} \|u(t^l) - u_h^l\|_{L^2(\Omega)} + C|z(t^{k-1}) - z^{k-1}|, \tag{2.36}$$

provided that u, f *and* $f(u)$ *are sufficiently smooth in both time and space variables.*

Proof. Subtracting (2.27)-(2.28) from (2.33)-(2.34), and with the notation

$$\begin{aligned} &e_h^{ni} = \Pi_h u(t^{ni}) - u_h^{ni}, \qquad \dot{e}_h^{ni} = \Pi_h \dot{u}_*^{ni} - \dot{u}_h^{ni}, \qquad e_h^n = \Pi_h u(t^n) - u_h^n, \quad \hat{e}_h^n = \Pi_h u(t^n) - \hat{u}_h^n, \\ &\xi^{ni} = z(t^{ni}) - z^{ni}, \qquad \dot{\xi}^{ni} = \dot{z}_*^{ni} - \dot{z}^{ni}, \qquad \xi^n = z(t^n) - z^n \quad . \end{aligned}$$

we have the following error equations

$$\begin{cases} \dot{e}_h^{ni} = \Delta_h e_h^{ni} + (z(t^{ni})\Pi_h W_*^{ni} - z^{ni} W_h^{ni}) + (\Pi_h \Delta - \Delta_h \Pi_h) u(t^{n-1}) & \text{for } i = 1, 2, \ldots, m, \\ e_h^{ni} = e_h^{n-1} + \tau \sum_{j=1}^m a_{ij} \dot{e}^{nj} + \Pi_h \eta_{ni} & \text{for } i = 1, 2, \ldots, m, \\ \hat{e}_h^n = e_h^{n-1} + \tau \sum_{i=1}^m b_i \dot{e}_h^{ni} + \Pi_h \eta_n \end{cases} \tag{2.37}$$

and

$$\begin{cases} \dot{\xi}^{ni} = -\dfrac{1}{2}(W_*^{ni}, \dot{u}_*^{ni}) + \dfrac{1}{2}(W_h^{ni}, \dot{u}_h^{ni})_h & \text{for } i = 1, 2, \ldots, m, \\ \xi^{ni} = \xi^{n-1} + \tau \sum_{j=1}^m a_{ij} \dot{\xi}^{nj} + d_{ni} & \text{for } i = 1, 2, \ldots, m, \\ \xi^n = \xi^{n-1} + \tau \sum_{j=1}^m b_i \dot{\xi}^{ni} + d_n, \end{cases} \tag{2.38}$$

Now, take the square of discrete L^2 norm of both sides of the last relation of equation (2.37), we can get

$$\|\hat{e}_h^n\|_h^2 = \|e_h^{n-1} + \tau \sum_{i=1}^m b_i \dot{e}_h^{ni}\|_h^2 + 2(\eta^n, e_h^{n-1} + \tau \sum_{i=1}^m b_i \dot{e}_h^{ni})_h + \|\Pi_h \eta^n\|_h^2. \tag{2.39}$$

For the first term on the right hand side, we expand it and apply the second equation of (2.37) to obtain

$$\|e_h^{n-1} + \tau \sum_{i=1}^m b_i \dot{e}_h^{ni}\|_h^2 = \|e_h^{n-1}\|_h^2 + 2\tau \sum_{i=1}^m b_i (\dot{e}_h^{ni}, e_h^{ni} - \eta_{ni})_h - \tau^2 \sum_{i,j=1}^m m_{ij} (\dot{e}_h^{ni}, \dot{e}_h^{nj})_h$$

$$\leq \|e_h^{n-1}\|_h^2 + 2\tau \sum_{i=1}^{m} b_i(\dot{e}_{ni}, e_h^{ni} - \eta_{ni})_h,$$

where in the last inequality we use the positive semi-definiteness of the matrix $\mathcal{M}$ in Assumption (P4). Next, we note that the relation of (2.37) implies

$$\begin{aligned}(\dot{e}_h^{ni}, e_h^{ni} - \eta_{ni})_h &= \Big(\Delta_h e_h^{ni} + (z(t^{ni})\Pi_h W_*^{ni} - z^{ni} W_h^{ni}) + (\Pi_h\Delta - \Delta_h\Pi_h)u(t^{n-1}), e_h^{ni} - \eta_{ni}\Big)_h \\ &= -\|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 + (\nabla e_h^{ni}, \nabla\Pi_h\eta_{ni}) + \Big(z(t^{ni})\Pi_h W_*^{ni} - z^{ni} W_h^{ni}, e_h^{ni} - \eta_{ni}\Big)_h \\ &\quad + \Big((\Pi_h\Delta - \Delta_h\Pi_h)u(t^{n-1}), e_h^{ni} - \eta_{ni}\Big)_h\end{aligned}$$

The bound of second term of the right hand side can be derived via Cauchy-Schwarz inequality

$$|(\nabla e_h^{ni}, \nabla\Pi_h\eta_{ni})| \leq \frac{1}{4}\|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 + C\|\eta_{ni}\|_{H^1(\Omega)}^2.$$

Then the third term can be bounded as

$$\begin{aligned}\Big(z(t^{ni})\Pi_h W_*^{ni} - z^{ni} W_h^{ni}, e_h^{ni} - \eta_{ni}\Big)_h &\leq z(t^{ni})\Big(\Pi_h W_*^{ni} - W_h^{ni}, e_h^{ni} - \eta_{ni}\Big)_h + \xi^{ni}\Big(W_h^{ni}, e_h^{ni} - \eta_{ni}\Big)_h \\ &\leq C\Big(\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2 + \|e_h^{ni}\|_h^2 + \|\Pi_h\eta_{ni}\|_{L^2(\Omega)}^2 + |\xi^{ni}|^2\Big).\end{aligned}$$

The bound of the last term follows from Lemma 2.2.3

$$\begin{aligned}\Big((\Pi_h\Delta - \Delta_h\Pi_h)u(t^{n-1}), e_h^{ni} - \eta_{ni}\Big)_h &\leq Ch^2\|e_h^{ni} - \eta_{ni}\|_{H^1(\Omega)} \\ &\leq \frac{1}{4}\|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 + C(\|e_h^{ni}\|_h^2 + \|\eta_{ni}\|_{H^1(\Omega)}^2 + h^4)\end{aligned}$$

Therefore, we arrive at

$$2(\dot{e}_h^{ni}, e_h^{ni} - \eta_{ni})_h \leq -\|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 + C\Big(\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2 + \|e_h^{ni}\|_h^2 + |\xi^{ni}|^2 + \|\eta_{ni}\|_{H^1(\Omega)}^2 + h^2\Big),$$

and hence

$$\|e_h^{n-1} + \tau\sum_{i=1}^{m} b_i\dot{e}_h^{ni}\|_h^2 \leq \|e_h^{n-1}\|_h^2 - \tau\sum_{i=1}^{m} b_i\|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 + C\tau\sum_{i=1}^{m}\Big(|\xi^{ni}|^2 + \|e_h^{ni}\|_h^2\Big)$$

$$+ C\tau \sum_{j=1}^{k} \|e_h^{n-j}\|_h^2 + C\tau(h^4 + \tau^{2k}).$$

In view of the first relation of the error equation (2.37), we have the estimate

$$\begin{aligned}(\eta^n, e_h^{n-1} + \tau \sum_{i=1}^{m} b_i \dot{e}_h^{ni})_h &\le \|\eta_n\|_h \|e_h^{n-1}\|_h + C\tau \|\eta_n\|_{H^1(\Omega)} \sum_{i=1}^{m} b_i \Big(\|\nabla e_h^{ni}\|_h + \sum_{j=1}^{k} \|e_h^{n-j}\|_h + |\xi^{ni}| + h^2 \Big) \\ &\le C\tau(h^4 + \tau^{2k}) + \frac{\tau}{4} \sum_{i=1}^{m} b_i \Big(\|\nabla e_h^{ni}\|_h^2 + |\xi^{ni}|^2 \Big) + C\tau \sum_{j=1}^{k} \|e_h^{n-j}\|_h^2\end{aligned}$$

which gives a bound of the second term in (2.39). In conclusion, we obtain that

$$\begin{aligned}\|\hat{e}_h^n\|_h^2 + \frac{\tau}{2} \sum_{i=1}^{m} \|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 &\le C\tau(h^4 + \tau^{2k}) + \|e_h^{n-1}\|_h^2 \\ &\quad + C\tau \sum_{i=1}^{m} \Big(\|e_h^{ni}\|_h^2 + |\xi^{ni}|^2 \Big) + C\tau \sum_{j=1}^{k} \|e_h^{n-j}\|_h^2.\end{aligned} \tag{2.40}$$

Similarly, from (2.38) and (2.35) we can derive

$$\begin{aligned}|\xi^n|^2 &\le C\tau(h^4 + \tau^{2k}) + (1 + c\tau)|\xi^{n-1}|^2 + \frac{\tau}{4} \sum_{i=1}^{m} \|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 \\ &\quad + C\tau \sum_{i=1}^{m} \Big(\|e_h^{ni}\|_h^2 + |\xi^{ni}|^2 \Big) + C\tau \sum_{j=1}^{k} \|e_h^{n-j}\|_h^2,\end{aligned}$$

where we use the estimate that

$$\begin{aligned}(W_*^{ni}, \dot{u}_*^{ni}) - (W_h^{ni}, \dot{u}_h^{ni})_h &= (W_*^{ni}, \dot{u}_*^{ni}) - (W_*^{ni}, \dot{u}_*^{ni})_h + (W_*^{ni} - W_h^{ni}, \dot{u}_*^{ni})_h + (W_h^{ni}, \dot{e}_h^{ni})_h \\ &\le Ch^2 + C \sum_{j=1}^{k} \|e_h^{n-j}\|_h \|\Pi_h \dot{u}_*^{ni}\|_h + (\nabla W_h^{ni}, \nabla e_h^{ni})_h \\ &\quad + (W_h^{ni}, z(t^{ni}) \Pi_h W_*^{ni} - z^{ni} W_h^{ni})_h + (W_h^{ni}, (\Pi_h \Delta - \Delta_h \Pi_h) u(t^{n-1}))_h \\ &\le Ch^2 + C \sum_{j=1}^{k} \|e_h^{n-j}\|_h + C\|\nabla e_h^{ni}\| + C|\xi^{ni}|,\end{aligned}$$

where we use the fact that $\|\nabla u_h^n\| \leq C$ (by Theorem 2.4.1) in the last inequality. To sum up, we arrive at

$$\begin{aligned}\|\hat{e}_h^n\|_h^2 + |\xi^n|^2 + \frac{\tau}{4}\sum_{i=1}^m \|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 \leq & C\tau(h^4+\tau^{2k}) + \|e_h^{n-1}\|_h^2 + (1+c\tau)|\xi^{n-1}|^2 \\ & + C\tau\sum_{i=1}^m \left(\|e_h^{ni}\|_h^2 + |\xi^{ni}|^2\right) + C\tau\sum_{j=1}^k \|e_h^{n-j}\|_h^2.\end{aligned}$$

Note that $|e_h^n(x_j)| \leq |\hat{e}_h^n(x_j)|$ for all $j = 0, 1, \ldots, M$, which implies

$$\begin{aligned}\|e_h^n\|_h^2 + |\xi^n|^2 + \frac{\tau}{4}\sum_{i=1}^m \|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 \leq & C\tau(h^4+\tau^{2k}) + \|e_h^{n-1}\|_h^2 + (1+c\tau)|\xi^{n-1}|^2 \\ & + C\tau\sum_{i=1}^m \left(\|e_h^{ni}\|_h^2 + |\xi^{ni}|^2\right) + C\tau\sum_{j=1}^k \|e_h^{n-j}\|_h^2.\end{aligned} \tag{2.41}$$

Next, we shall derive a bound for $\sum_{i=1}^m \left(\|e_h^{ni}\|_h^2 + |\xi^{ni}|^2\right)$ on the right-hand side. To this end, we test the second relation of (2.37) by e_h^{ni}. This yields

$$\begin{aligned}\sum_{i=1}^m \|e_h^{ni}\|_h^2 &\leq C\|e_h^{n-1}\|_h^2 + C\tau\sum_{i,j=1}^m a_{ij}(\dot{e}_h^{nj}, e_h^{ni}) + C\sum_{i=1}^m \|\Pi_h\eta_{ni}\|_h^2 \\ &\leq C\|e_h^{n-1}\|_h^2 + C\tau\sum_{i,j=1}^m a_{ij}(\dot{e}_h^{nj}, e_h^{ni})_h + C\tau^{2k}.\end{aligned}$$

Then, we apply the first relation of (2.37) and Lemma 2.2.3 to derive

$$\begin{aligned}\sum_{i,j=1}^m a_{ij}(\dot{e}_h^{nj}, e_h^{ni})_h = & -\sum_{i,j=1}^m a_{ij}(\nabla e_h^{nj}, \nabla e_h^{ni}) + \sum_{i,j=1}^m a_{ij}(z(t^{ni})\Pi_h W_*^{ni} - z^{ni}W_h^{ni}, e_h^{ni})_h \\ & + \sum_{i,j=1}^m a_{ij}((\Pi_h\Delta - \Delta_h\Pi_h)u(t^{n-1}), e_h^{ni})_h \\ \leq & C\sum_{i=1}^m \left(\|\nabla e_h^{ni}\|_{L^2(\Omega)}^2 + \|e_h^{ni}\|_h^2 + |\xi^{ni}|^2\right) + Ch^4 + C\sum_{j=1}^k \|e_h^{n-j}\|_h^2.\end{aligned}$$

Therefore, we obtain

$$\sum_{i=1}^m \|e_h^{ni}\|_h^2 \leq C(\tau h^4 + \tau^{2k}) + C\|e_h^{n-1}\|_h^2 + C\tau\sum_{j=1}^k \|e_h^{n-j}\|_h^2 +$$

$$C\tau\sum_{i=1}^{m}\Big(\|\nabla e_h^{ni}\|^2_{L^2(\Omega)}+\|e_h^{ni}\|_h^2+|\xi^{ni}|^2\Big).$$

Similarly, from (2.38) we can derive

$$\begin{aligned}\sum_{i=1}^{m}|\xi^{ni}|^2 &\le C|\xi^{n-1}|^2+C\tau\sum_{i,j=1}^{m}a_{ij}\dot{\xi}^{nj}\xi^{ni}+C\sum_{i=1}^{m}|d_{ni}|^2\\ &\le C(\tau h^4+\tau^{2k})+C|\xi^{n-1}|^2+C\tau\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2+C\tau\sum_{i=1}^{m}\Big(\|\nabla e_h^{ni}\|^2_{L^2(\Omega)}+\|e_h^{ni}\|_h^2+|\xi^{ni}|^2\Big)\end{aligned}$$

Sum up these two estimates and note that, for sufficiently small τ,

$$\sum_{i=1}^{m}\Big(\|e_h^{ni}\|_h^2+|\xi^{ni}|^2\Big)\le C(\tau h^4+\tau^{2k})+C|\xi^{n-1}|^2+C\tau\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2+C\tau\sum_{i=1}^{m}\|\nabla e_h^{ni}\|^2_{L^2(\Omega)}.$$

Now substituting the above estimate into (2.41), we have

$$\begin{aligned}\|e_h^n\|_h^2+|\xi^n|^2+\frac{\tau}{4}\sum_{i=1}^{m}\|\nabla e_h^{ni}\|^2_{L^2(\Omega)}\le&C\tau(h^4+\tau^{2k})+\|e_h^{n-1}\|_h^2+(1+C\tau)|\xi^{n-1}|^2\\&+C\tau^2\sum_{i=1}^{m}\|\nabla e_h^{ni}\|^2_{L^2(\Omega)}+C\tau\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2.\end{aligned}$$

Then for sufficiently small τ, there holds

$$\|e_h^n\|_h^2+|\xi^n|^2\le C\tau(h^4+\tau^{2k})+\|e_h^{n-1}\|_h^2+(1+C\tau)|\xi^{n-1}|^2+C\tau\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2.$$

Rearranging terms, we obtain

$$\frac{(\|e_h^n\|_h^2+|\xi^n|^2)-(\|e_h^{n-1}\|_h^2+|\xi^{n-1}|^2)}{\tau}\le C(h^4+\tau^{2k})+C|\xi^{n-1}|^2+C\sum_{j=1}^{k}\|e_h^{n-j}\|_h^2.$$

Then the discrete Gronwall's inequality implies

$$\max_{k\le n\le N}\Big(\|e_h^n\|_h^2+|\xi^n|^2\Big)\le C(h^4+\tau^{2k})+C|\xi^{k-1}|^2+C\sum_{j=0}^{k-1}\|e_h^j\|_h^2.$$

This completes the proof of the theorem. □

2.5 Numerical Results

In this section, we present numerical results to illustrate the the theoretical results with a one-dimensional example:

$$\begin{cases} \partial_t u = \partial_{xx} u + f(u), & \text{in } \Omega \times (0,T], \\ \partial_x u = 0, & \text{on } \partial\Omega \times (0,T] \\ u(x, t=0) = u_0(x) & \text{in } \Omega, \end{cases} \tag{2.42}$$

where $\Omega = (0,2)$ and $f(u) = \varepsilon^{-2}(u - u^3)$ with $\varepsilon = 0.1$ is the Ginzburg-Landau double-well potential. The initial value satisfies the maximum principle given by

$$u_0(x) = \begin{cases} 1, & \text{if } 0 < x < 1/2, \\ \cos\left(\frac{2}{3}\pi\left(x + \frac{1}{2}\right)\right), & \text{if } 1/2 \le x < 2. \end{cases} \tag{2.43}$$

The smooth initial value is chosen to satisfy the Neumann boundary condition.

We solve the problem (2.42) with spatial mesh size $h = 2/N_x$ and temporal mesh size $\tau = T/N_t$, with $T = \varepsilon^2$ and $5\varepsilon^2$. Throughout the section, we shall apply the Gauss–Legendre methods with $m = 1, 2, 3$ and hence $k = 2, 3, 4$. We compute the numerical solution at the first $k-1$ time levels by using the three-stage Gauss–Legendre Runge–Kutta method [114, Table 5.2], that are sufficiently accurate to achieve the optimal convergence rate. Cutting off the numerical solutions at the first $k-1$ time levels does not affect the global accuracy.

Since the closed form of exact solution is unavailable, we compare our numerical solution with a reference solution computed by a high-order method (i.e. cut-off RK method with $r = 3$, $m = 3$) with small mesh sizes. In particular, the temporal error e_τ is computed by fixing the spatial mesh size $h = 2/400$ and comparing the numerical solution with a reference solution (with $\tau = T/1000$). Similarly, the spatial error e_h is computed to by fixing the temporal step size $\tau = T/1000$ and comparing the numerical solutions with a reference solution (with $h = 2/400$).

In Table 2.2, we present the spatial errors of both cut-off RK schemes (2.17)-(2.18) with $r = 1, 2, 3$ and the cut-off SAV-RK scheme (2.27)-(2.29) with $r = 1$. Numerical results show the optimal rate $O(h^{r+1})$, which fully supports our theoretical results in Theorems 2.3.1 and 2.4.2. Temporal errors are presented in 2.3 and 2.4, both of which show the empirical convergence rate $O(\tau^{m+1})$ and hence coincidence to

Theorems 2.3.1 and 2.4.2.

Table 2.2: e_h of cut-off RK (2.17)-(2.18) and cut-off SAV-RK (2.27)-(2.29).

$r\backslash N_x$	T	10	20	40	80	160	rate
RK	0.01	3.03e-2	7.42e-3	1.84e-3	4.60e-4	1.14e-4	$\approx$ 2.00 (2.00)
(r=1)	0.05	1.49e-1	1.03e-2	2.32e-3	5.71e-4	1.43e-4	$\approx$ 2.01 (2.00)
RK	0.01	4.37e-3	4.99e-4	5.90e-5	7.27e-6	9.05e-7	$\approx$ 3.01 (3.00)
(r=2)	0.05	6.15e-2	1.64e-3	1.73e-4	2.09e-5	2.60e-6	$\approx$ 3.03 (3.00)
RK	0.01	5.10e-4	3.19e-5	1.99e-6	1.23e-7	7.74e-9	$\approx$ 4.00 (4.00)
(r=3)	0.05	5.89e-3	1.21e-4	8.12e-6	5.03e-7	3.14e-8	$\approx$ 4.01 (4.00)
SAV-RK	0.01	3.03e-2	7.42e-3	1.84e-2	4.62e-4	1.17e-4	$\approx$ 2.00 (2.00)
(r=1)	0.05	1.49e-1	1.03e-2	2.34e-3	5.85e-4	1.56e-4	$\approx$ 2.01 (2.00)

Table 2.3: e_τ of cut-off RK scheme (2.17)-(2.18), with $\tau = T/N_t$.

$m\backslash N_t$	T	10	20	40	80	160	320	rate
1	0.01	3.76e-4	9.61e-5	2.43e-5	6.10e-5	1.53e-6	3.82e-7	$\approx$ 1.99 (2.00)
	0.05	8.01e-4	5.36e-5	1.16e-5	2.71e-6	6.56e-7	1.61e-7	$\approx$ 2.06(2.00)
2	0.01	4.92e-5	6.20e-6	7.74e-7	9.65e-8	1.21e-8	1.51e-9	$\approx$ 3.00 (3.00)
	0.05	1.73e-2	3.60e-5	1.78e-6	2.08e-7	2.51e-8	3.08e-9	$\approx$ 3.06 (3.00)
3	0.01	1.05e-5	6.83e-7	4.31e-8	2.71e-9	1.69e-10	1.05e-11	$\approx$ 4.00 (4.00)
	0.05	2.88e-2	3.66e-3	3.82e-7	1.56e-8	9.61e-10	6.06e-11	$\approx$ 4.21 (4.00)

Table 2.4: e_τ of cut-off SAV-RK scheme (2.27)-(2.29), with $\tau = T/N_t$.

$m\backslash N_t$	T	10	20	40	80	160	320	rate
1	0.01	8.08e-3	2.23e-3	5.96e-4	1.53e-4	3.79e-5	8.78e-6	$\approx$ 2.03 (2.00)
	0.05	7.94e-4	1.79e-4	4.80e-5	1.24e-5	3.09e-6	7.17e-7	$\approx$ 2.00 (2.00)
2	0.01	5.56e-9	5.95e-4	8.82e-5	1.11e-5	1.37e-6	1.65e-7	$\approx$ 3.02 (3.00)
	0.05	1.47e-2	5.17e-5	7.17e-6	1.00e-6	1.31e-7	1.63e-8	$\approx$ 2.97 (3.00)
3	0.01	6.97e-11	2.56e-4	2.47e-5	1.66e-6	1.06e-7	6.60e-9	$\approx$ 3.95 (4.00)
	0.05	2.45e-2	2.86e-3	7.73e-7	6.16e-8	4.38e-9	2.93e-10	$\approx$ 3.79 (4.00)

In Figure 4.1, we plot the maximal cut-off value at each step

$$\rho^n = \max_{0\le j\le Mr+1} |u_h^n(x_j) - \hat{u}_h^n(x_j)|$$

and the error of the numerical solution $e(x) = u_h^N(x) - u(x,T)$. Our numerical results show that the cut-off operation is active in the computation. Meanwhile, we observe that a coarse step mesh will result in a larger cut-off value, without affecting the convergence rate.

Finally, we test the numerical results in case of relatively large time steps, and compare the numerical solutions of extrapolated RK, cut-off RK (2.17)-(2.18), and cut-off SAV-RK schemes (2.27)-(2.29), with

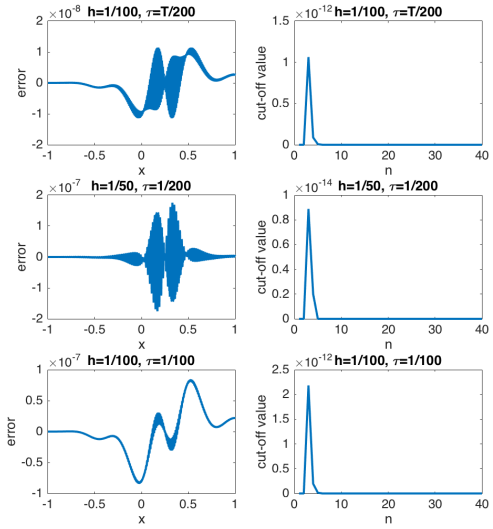
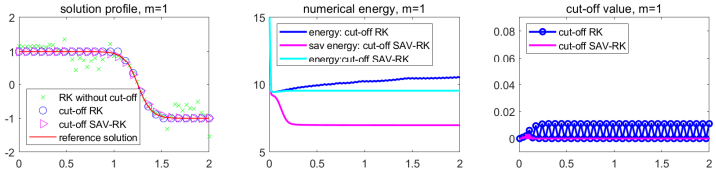


Figure 2.1: Error at $T = 0.01$ and maximal cut-off value at each time level.

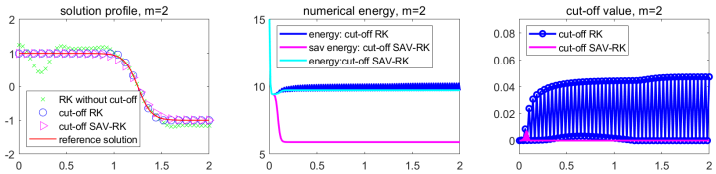
$r = 1$, see Figure 2.2. Without the cut-off postprocessing, the numerical solutions of RK scheme significantly exceed the maximum bound, and present oscillating solution profiles. With the cut-off operation at each time step, the numerical solutions satisfy the maximum bound, and present reasonable solution profiles. However, numerical results show that the cut-off RK scheme might produce a solution with a obviously increasing and oscillating energy curve. This issue could be significantly improved by applying the cut-off SAV-RK method, whose solution satisfy the maximum bound and the numerical energy is more stable. Moreover, the numerical results show that the cut-off SAV-RK scheme will produce a more regular numerical solution and smaller cut-off values, compared with the cut-off RK scheme.

2.6 Conclusion and Comments

In this chapter, we discuss the cut-off postprocessing on a series of single step methods, for Allen–Cahn equation with the nonlinear term linearized. We prove that our scheme can be arbitrarily high order for time discretized problem, and be arbitrarily high order for both space and time for fully discretized problem. A lot of famous schemes are included in our analysis. Combining this strategy with SAV technique, we also develop a class of schemes, which preserve both maximum bound and energy stable. Related numerical are also given to corroborate our analysis.



(a) $m = 1, \varepsilon = 0.1, T = 2, \tau = 1/150$



(b) $m = 2, \varepsilon = 0.1, T = 2, \tau = 1/250$

Figure 2.2: Left: solution profiles of numerical solutions of RK, cut-off RK and cut-off SAV-RK scheme. Middle: solution energy of cut-off RK and cut-off SAV-RK scheme. Right: cut-off values of cut-off RK and cut-off SAV-RK scheme.

Chapter 3

High-order Implicit-Explicit Runge-Kutta Methods for Parabolic Equations

In this chapter, we will develop and study Implicit-Explicit Runge–Kutta method (IMEX-RK) for linear and semilinear parabolic equations. To begin with, we will focus on linear non-selfadjoint equations. It is more than a easier case but itself is also a interesting question and related to some physics problems.

In Section 3.5, we will give a brief introduction to the linear problem and its related background. In Section 3.2 build the IMEX-RK method for linear problem and give its long time error convergence in Section 3.3. In Section 3.4, we extend the analysis to semilinear problem and show it can keep both maximum bound preserving and original energy decay for up to third order. The schemes of this chapter is listed in Section 3.5 which meet all our requirements for the convergence.

3.1 Introduction

In this work, we investigate a numerical approach to solve the following problem. Let $V \subset H = H' \subset V'$ be a Gelfand triple of Hilbert spaces, where the superscript ' denotes the dual. Namely, the embedding

$V \hookrightarrow H$ is continuous and dense, and

$$(u, v)_{V,V'} = (u, v)_H \quad \forall u \in H \hookrightarrow V', v \in V \hookrightarrow H,$$

where $(\cdot, \cdot)_{V,V'}$ is the duality pairing between V' and V, and $(\cdot, \cdot)_H$ is the inner product on H.

We consider an abstract parabolic initial value problem: find

$$u \in L^2((0, T); V) \subset H^1((0, T); V') \hookrightarrow C([0, T]; H)$$

such that

$$\begin{cases} \partial_t u = \mathcal{A}u + f(t) \quad 0 < t < T, \\ u(0) = u_0 \in H \end{cases} \tag{3.1}$$

where $\mathcal{A} : V \to V'$ is a bounded linear operator (possibly non-selfadjoint) with the following property that:

$$\beta^{-1}\|u\|_V^2 \leq -(\mathcal{D}u, u) \leq \beta\|u\|_V^2 \quad \forall u \in V, \tag{3.2}$$

$$|(\mathcal{L}u, v)| \leq C\|u\|_V\|v\|_H, \forall u \in V, v \in H,$$

where $\mathcal{D} = (\mathcal{A} + \mathcal{A}^*)/2, \mathcal{L} = (\mathcal{A} - \mathcal{A}^*)/2$ are the symmetric and skew-symmetric part of the operator $\mathcal{A}$. Furthermore, D is negative definite.

The Stokes-Darcy system

As a multiphysics system, the Stokes-Darcy system is considered in this work which describes a moving fluid governed by the Stokes equations in a free-flow region $\Omega_S \subset \mathbb{R}^d$ and a flow in a neighboring porous media region $\Omega_D \subset \mathbb{R}^d$. Darcy flow and Stokes flow interact through an interface denoted by Γ as shown in Figure 3.1. Applications of such a coupled system are ubiquitous in nature, including groundwater system [54, 56, 69], petroleum extraction [14], and so on.

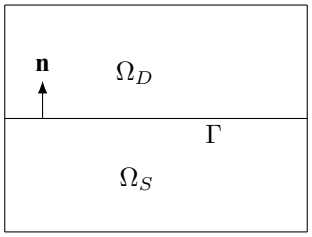


Figure 3.1: Computational domain of the Stokes–Darcy system

This model consists of a parabolic equation

$$\begin{cases} \partial_t \phi - \nabla \cdot (\kappa \nabla \phi) = f_D & \text{in } \Omega_D \times (0,T], \\ \phi = 0 & \text{on } \partial\Omega_D \backslash \overline{\Gamma} \times (0,T], \\ -\kappa \nabla \phi \cdot \mathbf{n} = \mathbf{u} \cdot \mathbf{n} & \text{on } \Gamma \times (0,T], \\ \phi(0) = \phi_0 & \text{in } \Omega_D, \end{cases} \tag{3.3}$$

which describes the Darcy flow in the porous media region Ω_D through the unknown hydraulic head ϕ, and an evolving Stokes equation

$$\begin{cases} \partial_t \mathbf{u} - \nabla \cdot \mathbb{T}(\mathbf{u}, p) = \mathbf{f}_S & \text{in } \Omega_S \times (0,T], \\ \nabla \cdot \mathbf{u} = 0 & \text{in } \Omega_S \times (0,T], \\ \mathbf{u} = 0 & \text{on } \partial\Omega_S \backslash \overline{\Gamma} \times (0,T], \\ -\mathbb{T}(\mathbf{u}, p)\mathbf{n} = g\phi\mathbf{n} + \mu(\mathbf{u} - (\mathbf{u} \cdot \mathbf{n})\mathbf{n}) & \text{on } \Gamma \times (0,T], \\ \mathbf{u}(0) = \mathbf{u}_0 & \text{in } \Omega_S, \end{cases} \tag{3.4}$$

which describes free flow in the region Ω_S through the fluid velocity $\mathbf{u}$, where $\mathbb{T}(\mathbf{u}, p) = 2\nu\mathbb{D}(\mathbf{u}) - p\mathbb{I}$ denotes the stress tensor, in which $\mathbb{D}(\mathbf{u}) = \frac{1}{2}(\nabla\mathbf{u} + (\nabla\mathbf{u})^\top)$ is the deformation tensor and $\mathbb{I}$ is the $d \times d$ identity matrix. The physical parameters κ, g, μ and ν in this model are positive constants, and f_D and $\mathbf{f}_S$ are given source terms.

Homogeneous Dirichlet boundary conditions will be imposed on outer boundaries separately, i.e., on $\partial\Omega_D \backslash \Gamma$ and $\partial\Omega_S \backslash \Gamma$. The interface conditions on Γ in (3.3) and (3.4) represent conservation of mass and

balance of force, respectively, where $\mathbf{n}$ denotes the unit normal vector on $\partial\Omega_S$ as shown in Figure 3.1.

For the ease of error estimate, we will rewrite the Stokes-Darcy system as an equivalent abstract problem.

Let $H = L^2(\Omega_D) \times L^2(\Omega_S)^d$ and $V = H^1_\Gamma(\Omega_D) \times \mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0)$, where

$$H^1_\Gamma(\Omega_D) = \{\varphi \in H^1(\Omega_D) : \varphi = 0 \text{ on } \partial\Omega_D \backslash \Gamma\},$$
$$\mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0) = \{\mathbf{v} \in H^1(\Omega_S)^d : \nabla \cdot \mathbf{v} = 0 \text{ in } \Omega_S \text{ and } \mathbf{v} = 0 \text{ on } \partial\Omega_S \backslash \Gamma\}.$$

The weak formulation of (3.3)-(3.4) reads: find $(\phi, \mathbf{u}) \in L^2((0,T); V) \cap H^1((0,T); V') \hookrightarrow C([0,T]; H)$ satisfying the following equations for all test functions $(\varphi, \mathbf{v}) \in L^2((0,T); V)$:

$$(\partial_t \phi, \varphi)_D + (\kappa \nabla \phi, \nabla \varphi)_D - (\mathbf{u} \cdot \mathbf{n}, \varphi)_\Gamma = (f_D, \varphi) \tag{3.5}$$

$$(\partial_t \mathbf{u}, \mathbf{v})_S + (2\nu \mathbb{D}(\mathbf{u}), \mathbb{D}(\mathbf{v}))_S + (g\phi, \mathbf{v} \cdot \mathbf{n})_\Gamma + \mu(\mathbf{u} - (\mathbf{u} \cdot \mathbf{n})\mathbf{n}, \mathbf{v} - (\mathbf{v} \cdot \mathbf{n})\mathbf{n})_\Gamma \tag{3.6}$$
$$= (\mathbf{f}_S, \mathbf{v}),$$

where $(\cdot,\cdot)_D$ is the pairing between $H^1_\Gamma(\Omega_D)'$ and $H^1_\Gamma(\Omega_D)$, $(\cdot,\cdot)_S$ is the pairing between $\mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0)'$ and $\mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0)$, and $(\cdot,\cdot)_\Gamma$ is the inner product on $L^2(\Gamma)$.

Let the operators $A_1 : H^1_\Gamma(\Omega_D) \to H^1_\Gamma(\Omega_D)'$, $A_2 : \mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0) \to \mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0)'$, $B : \mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0) \to H^1_\Gamma(\Omega_D)'$ and $B^* : H^1_\Gamma(\Omega_D) \to \mathbf{H}^1_\Gamma(\Omega_1; \text{div}_0)'$ be defined via duality by

$$(A_1\phi, \varphi)_D = (\kappa\overline{\nabla\phi}, \nabla\varphi)_D \qquad \forall\, \phi, \varphi \in H^1_\Gamma(\Omega_D),$$
$$(A_2\mathbf{u}, \mathbf{v})_S = (2\nu\overline{\mathbb{D}(\mathbf{u})}, \mathbb{D}(\mathbf{v}))_S + \mu(\overline{\mathbf{u} - (\mathbf{u}\cdot\mathbf{n})\mathbf{n}}, \mathbf{v} - (\mathbf{v}\cdot\mathbf{n})\mathbf{n})_\Gamma \quad \forall\, \mathbf{u}, \mathbf{v} \in \mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0),$$
$$(B\mathbf{u}, \varphi)_D = (\overline{\mathbf{u}\cdot\mathbf{n}}, \varphi)_\Gamma \qquad \forall\, \mathbf{u} \in \mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0) \hookrightarrow L^2(\Gamma),\ \forall\, \varphi \in H^1_\Gamma(\Omega_D) \hookrightarrow L^2(\Gamma),$$
$$(B^*\phi, \mathbf{v})_D = (\phi, \overline{\mathbf{v}\cdot\mathbf{n}})_\Gamma \qquad \forall\, \phi \in H^1_\Gamma(\Omega_D) \hookrightarrow L^2(\Gamma),\ \forall \mathbf{v} \in \mathbf{H}^1_\Gamma(\Omega_S; \text{div}_0) \hookrightarrow L^2(\Gamma).$$

Then the weak formulation (3.5)-(3.6) can be written as

$$\partial_t \phi + A_1\phi - B\mathbf{u} = f_D, \tag{3.7}$$

$$\partial_t \mathbf{u} + A_2\mathbf{u} + gB^*\phi = \mathbf{f}_S. \tag{3.8}$$

By defining notations

$$u = \begin{pmatrix} \phi \\ g^{-\frac{1}{2}}\mathbf{u} \end{pmatrix}, \quad f = \begin{pmatrix} f_D \\ g^{-\frac{1}{2}}\mathbf{f}_S \end{pmatrix} \quad \text{and} \quad \mathcal{A} = -\begin{pmatrix} A_1 & -g^{\frac{1}{2}}B \\ g^{\frac{1}{2}}B^* & A_2 \end{pmatrix}, \tag{3.9}$$

equations (3.7)-(3.8) are equivalently reformulated to the abstract parabolic initial value problem (3.1).

Due to the various applications of Stokes-Darcy system, many different numerical methods are developed and analyzed, including domain decomposition methods [25, 8, 13, 17, 24, 26, 35, 49, 112, 89], Lagrange multiplier methods [70, 5, 42, 57], discontinuous Galerkin methods [63, 77, 94, 95], multigrid methods [4, 83], partitioned time-stepping methods [66, 84, 97, 123], coupled finite element methods [12, 11, 65, 81] and many others [18, 19, 43, 53]. In particular, Kubacki et al [67] presented an overview of non-iterative partitioned methods for such a system. With a time-step restriction for stability, numerical schemes of both first-order and high-order partitioned methods were presented. Gunzburger et al in [49] analyzed a parallel, non iterative, multiphysics domain decomposition method for decoupling the Stokes-Darcy model with multistep backward difference formula (BDF) for the time discretization. Optimal order $O(\tau^k)$ for the k-step BDF scheme were established in a general framework for any $k \leq 5$. Chen et al in [19] proposed two second-order-in-time implicit-explicit methods including 2-step BDF and second-order Adams-Moulton-Bashforth method (AME2) in which coupling term in the interface conditions was treated explicitly and established for both schemes the unconditional and uniform-in-time stability. Error bound was derived with $O(\tau^2)$. An improvement of this work was presented in [18], in which a third-order in time AME algorithm was studied and uniform-in-time error estimate was derived. Recently, authors in [72] presented an implicit-explicit (IMEX) scheme with k-step BDF in time and finite element discretization in space. In that paper, the spatial differential operator $\mathcal{A}$ was split into a symmetric part and an anti-symmetric part on which implicit and explicit schemes were applied respectively. A symmetrized and decoupled temporal k-step BDF scheme was presented and optimal long-time error bound $O(\tau^k + h^2)$ was derived.

We notice that aforementioned high-order-in-time works are conducted with multi-step methods and to our knowledge there is no temporal high-order single-step methods adopted for the coupled Stokes-Darcy system in the literature. This motivates us to apply IMEX Runge-Kutta method on the Stokes-Darcy equations aiming to achieve high-order convergence in time.

c_1	a_{11}	0	$\dots$	0	$\hat{a}_{11}$	0	$\dots$	0
c_2	a_{21}	a_{22}	$\dots$	0	$\hat{a}_{21}$	$\hat{a}_{22}$	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
c_m	a_{m1}	a_{m2}	$\dots$	a_{mm}	$\hat{a}_{m1}$	$\hat{a}_{m2}$	$\dots$	$\hat{a}_{mm}$
	b_1	b_2	$\dots$	b_m	$\hat{b}_1$	$\hat{b}_2$	$\dots$	$\hat{b}_m$

Table 3.1: Butcher tableau for Runge–Kutta scheme.

3.2 Implicit-Explicit Runge-Kutta Methods

In this section, we shall propose the time stepping scheme for solving the Stokes–Darcy system (3.3)-(3.4) by using the IMEX Runge–Kutta method. To this end, we split the interval $[0, T]$ into a sequence of subintervals $[t^i, t^{i+1}]$, for $i = 0, 1, 2, \dots, N-1$. with time levels $0 = t^0 < t^1 < \cdots < t^n = T$. The mesh size is denoted by $\tau := \max_{1\le i\le n} \tau_i$ with $\tau_i = t^i - t^{i-1}$. To simplify the presentation, we will proceed under the assumption that the step size is uniform, i.e., $\tau_i = \tau$ for $i = 1, \dots, n$. Nevertheless, it is important to note that the analysis applies to nonuniform meshes as well, since the proposed schemes are single-step.

For the symmetric part, we consider a m-stage diagonally implicit Runge–Kutta (DIRK) scheme with coefficient $A = \{a_{ij}\}_{m\times m}$, $b = \{b_i\}_{i=1}^m$ and $c = \{c_i\}_{i=1}^m$. For the skew-symmetric part, we make use of a m-stage explicit scheme with coefficients $\hat{A} = \{\hat{a}_{ij}\}_{m\times m}$, $\hat{b} = \{\hat{b}_i\}_{i=1}^m$ and $\hat{c} = c = \{c_i\}_{i=1}^m$. It is important to note that the implicit scheme and explicit scheme share the same internal nodes $t^{n,i} = t^n + c_i\tau_n$. The IMEX Runge–Kutta schemes can be determined by the following Butcher tabular

Throughout, we assume that the scheme is *stiffly accurate*:

Assumption (P1) Assume that b_i and $\hat{b}_i$ are the last columns of A and $\hat{A}$, respectively, which means:

$$b^\top = z^\top A, \quad \hat{b}^\top = z^\top \hat{A} \tag{3.10}$$

where $z = (0, \dots, 0, 1)^\top$.

Remark 3.2.1. *The condition (P1) is a common assumption that improves stability when dealing with stiff problems, ensuring that the method can take larger time steps without losing accuracy. See some useful properties in Proposition 3.2.1 and Corollary 3.2.1.*

Let u^n, $u^{n,i}$ be approximations to $u(t^n)$ and $u(t^{n,i})$, respectively, and $f^{n,i} = f(t^{n,i})$. Then the IMEX

Runge–Kutta scheme for solving the parabolic problem (4.1) can be written as

$$\begin{cases} u^{n,0} = u^n \\ u^{n,i} = u^n + \tau \sum_{j=1}^{i} a_{ij} \mathcal{D} u^{n,j} + \tau \sum_{j=1}^{i} \hat{a}_{ij} (\mathcal{L} u^{n,j-1} + f^{n,j-1}) \quad i = 1, \dots, m \\ u^{n+1} = u^n + \tau \sum_{j=1}^{m} b_j \mathcal{D} u^{n,j} + \tau \sum_{j=1}^{m} \hat{b}_j (\mathcal{L} u^{n,j-1} + f^{n,j-1}). \end{cases} \tag{3.11}$$

Alternatively, we let $U^n = (u^{n,i})_{i=1}^m$, $V^n = (v^{n,i})_{i=1}^m = (\mathcal{L} u^{n,i-1} + f^{n,i-1})_{i=1}^m$. Then the scheme (3.11) could be written in vector form as

$$\begin{cases} U^n = \mathbb{1} u^n + \tau A \mathcal{D} U^n + \tau \hat{A} V^n, \\ u^{n+1} = u^n + \tau b^\top \mathcal{D} U^n + \tau b^\top V^n. \end{cases} \tag{3.12}$$

Note that the first relation in (3.12) leads to

$$U^n = (I - \tau A \mathcal{D})^{-1} (\mathbb{1} u^n + \tau \hat{A} V^n).$$

This together with the second relation in (3.12) and Assumption (P1) yields

$$u^{n+1} = z^\top U^n = z^\top (I - \tau A \mathcal{D})^{-1} \mathbb{1} u^n + \tau z^\top (I - \tau A \mathcal{D})^{-1} \hat{A} V^n \tag{3.13}$$

where $z = (0, \dots, 0, 1)^\top$. Let $\sigma(s) = z^\top (I + sA)^{-1} \mathbb{1}$, and $p_i(s)$ be the i-th entry of $z^\top (I + sA)^{-1} \hat{A}$ for $s > 0$. Then the scheme (3.13) can be written as the following equivalent form

$$u^{n+1} = \sigma(-\tau_n \mathcal{D}) u^n + \tau_n \sum_{i=1}^{m} p_i(-\tau_n \mathcal{D}) (\mathcal{L} u^{n,i} + f^{n,i}). \tag{3.14}$$

Here $\sigma(s)$ and $\{p_i(s)\}_{i=1}^m$ are rational functions.

To guarantee long-time stability, we need the following assumption on the rational function σ.

Assumption (P2): We assume that $0 < \sigma(s) < 1$ for all $s \in (0, \infty)$.

Remark 3.2.2. *The bound $|\sigma(s)| \leq 1$ is typically required for stability. However, we need a stronger assumption, $\sigma(s) > 0$, to ensure that $\sigma(-\tau \mathcal{D})$ is invertible. This is crucial because we use the test*

function $\sigma(-\tau\mathcal{D})^{-1}u^{n+1}$ in the proof of long-time stability (see Theorem 3.3.1). For the backward Euler scheme, $\sigma(-\tau\mathcal{D})^{-1} = (\mathcal{I}-\tau\mathcal{D})$ is well-defined, but the invertibility of $\sigma(-\tau\mathcal{D})$ is not always guaranteed for higher-order Runge–Kutta methods.

Later on, we will show some properties of σ and p_i, which will benefit our further analysis.

Proposition 3.2.1. *Assume that (P1) is valid. Then σ and p_i possess a common denominator with a degree of m, while the degree of their numerators does not exceed $m-1$.*

Proof. Note that $(I+sA)^{-1} = (\det(I+sA))^{-1}(I+sA)^*$, where the star here denotes the cofactor matrix. Since $\sigma(s) = z^\top(I+sA)^{-1}\mathbb{1}$ and $p_i(s)$ are the i-th entry of $z^\top(I+sA)^{-1}\hat{A}$, we observe that σ and p_i share the same denominator $\det(I+sA)$, a polynomial of degree m. Also, we conclude that all entries of the cofactor matrix are polynomials with degree not exceeding $m-1$. It follows that numerators of the polynomials p_i and σ also have a degree of $m-1$. □

As a direct result of Proposition 3.2.1, we have the following estimates for the rational functions σ and p_i.

Corollary 3.2.1. *Let Assumptions (P1) and (P2) be valid. Then there exist positive constants c_0, c_1 and c_2 such that for all $s \geq 0$, the following holds:*

$$\sigma(s) < \frac{1}{1+c_0 s}, \quad s\sigma(s) < c_1 \quad \textit{and} \quad \left|\frac{p_i(s)}{\sigma(s)}\right| < c_2.$$

Proof. In the first estimate, $(1+cs)\sigma(s)$ is a rational function with equal numerator and denominator, finite on $(0,\infty)$. It has finite local maxima on this interval. For $c=0$, all maxima are less than one, so there exists a small c_0 near zero such that $(1+cs)\sigma(s) < 1$.

In the second estimate, $s\sigma(s)$ is a rational function where the numerator equals the denominator, making it finite on $(0,\infty)$. Therefore, $\lim_{s\to+\infty} s\sigma(s)$ is finite, and $s\sigma(s)$ remains finite on $(0,\infty)$.

In the third estimate, p_i/σ is a rational function where the numerator equals the denominator, ensuring it is finite on $(0,\infty)$ and remains finite throughout.

□

To illustrate the order condition of our scheme, we need to make the third assumption:

Assumption (P3) Here we define

$$A_+ = \begin{pmatrix} 0 & 0 \\ 0 & A \end{pmatrix}, \hat{A}_+ = \begin{pmatrix} 0 & 0 \\ \hat{A} & 0 \end{pmatrix}, c_+ = \begin{pmatrix} 0 \\ c \end{pmatrix}.$$

We assume that, for a given positive integer k, the following properties hold valid

$$z^\top (A_*)^{r+1} c_+^l = \frac{l!}{(l+r+1)!}, \quad \forall l \geq 0, r \geq 0, l+r \leq k-1, \tag{3.15}$$

where, each appearance of A_* during the multiplication is either A_+ or $\hat{A}_+$.

We say a scheme is accurate of order k, if

$$\sigma(\lambda) = e^{-\lambda} + O(\lambda^{k+1}) \text{ as } \lambda \to 0, \tag{3.16a}$$

and, for any $0 \leq l \leq k$,

$$\sum_{i=0}^{m} c_i^l p_i(\lambda) = \frac{l!}{(-\lambda)^{l+1}} \Big(e^{-\lambda} - \sum_{l=0}^{j} \frac{(-\lambda)^l}{l!} \Big) + O(\lambda^{k-l}) \text{ as } \lambda \to 0. \tag{3.16b}$$

To achieve k-th order accuracy, the following theorem provides the necessary and sufficient conditions for the Butcher tableau.

Theorem 3.2.1. *The scheme is accurate of order k for linear symmetric problem if Assumption (P3) is valid.*

Proof. Let $\mathcal{A} = \mathcal{D}$ and denote $U^n = \{u^{n,i}\}_{i=0}^m$ and $F^n = \{f(t^{n,i})\}_{i=0}^m$. Then scheme reads

$$U^n = \mathbb{1} u^n + \tau A_+ \mathcal{D} U^n + \tau \hat{A}_+ F^n \quad \text{and} \quad u^{n+1} = u^n + \tau {b_+}^\top \mathcal{D} U + \tau {b_+}^\top F. \tag{3.17}$$

The first relation of (3.17) gives

$$U^n = (I - \tau A_+ \mathcal{D})^{-1} \mathbb{1} u^n + \tau (I - \tau A_+ \mathcal{D})^{-1} \hat{A}_+ F^n.$$

The equation (3.10) guarantees that the last element of U^n happens to be u^{n+1} itself, so finally we can

get

$$u^{n+1} = \sigma(-\tau\mathcal{D})u^n + \tau\sum_{i=0}^{m} p_i(-\tau\mathcal{D})f(t^{n,i}),$$

where $\sigma(\lambda) = z^\top(1+\lambda A_+)^{-1}\mathbb{1}$, $\{p_i(\lambda)\}^\top = z^\top(1+\lambda A_+)^{-1}\hat{A}_+$ for $\lambda \geq 0$. We know the rational real function on the operator $\mathcal{D}$ is well-defined since $\mathcal{D}$ is negative definite such that $-\tau\mathcal{D}$ is positive definite.

To derive (3.16a), since the rational functions and exponential functions are sufficiently smooth and analytic, we can just take the derivative and test whether they are the same at zero.

$$\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\right)^l \sigma = z^\top\left(\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\right)^l (1+\lambda A_+)^{-1}\right)\mathbb{1} = (-1)^l l! z^\top (1+\lambda A_+)^{-(l+1)} {A_+}^l \mathbb{1}.$$

Therefore, (3.16a) is true if and only of

$$\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\right)^l \sigma(0) = (-1)^l l! z^\top {A_+}^l \mathbb{1} = (-1)^l, \quad \forall 0 \leq l \leq k,$$

which means

$$z^\top {A_+}^l \mathbb{1} = \frac{1}{l!}, \quad \forall 0 \leq l \leq k.$$

To evaluate (3.16b), we need some simplification. The equation (3.16b) is equivalent to

$$(-\lambda)^{l+1}\sum_{i=0}^{m} c_i^l p_i(\lambda) = l!\Big(e^{-\lambda} - \sum_{l=0}^{j}\frac{(-\lambda)^l}{l!}\Big) + O(\lambda^{k+1}), \quad as\ \lambda \to 0. \tag{3.18}$$

Similarly, we only need to compare their derivatives at zero. Let

$$LHS = (-\lambda)^{l+1}\sum_{i=0}^{m} c_i^l p_i(\lambda), \quad RHS = l!\Big(e^{-\lambda} - \sum_{l=0}^{j}\frac{(-\lambda)^l}{l!}\Big).$$

It is obviously that LHS has a λ^{l+1} factor so it is zero for no more than l-th order derivative. So does RHS because it is the Taylor's expansion. Thus we only need to test their derivatives with order more than l.

Let $r \geq 0$ and $l+1+r = k$. Taking the $(l+1+r)$-th order derivative on LHS, with the Leibniz

product rule, we can derive that

$$\begin{aligned}\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\right)^{l+1+r} LHS(0) &= C_{l+1+r}^{r}\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\right)^{l+1}(-\lambda)^{l+1}\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\right)^{r}\sum_{i=0}^{m} c_i^l p_i(\lambda)\\ &= \frac{(l+1+r)!}{(l+1)!r!}(-1)^{l+1}(l+1)!\cdot z^\top(-1)^r r!(1+0\cdot A_+)^{-(r+1)}{A_+}^r\hat{A}_+c^l\\ &= (-1)^{l+1+r}(l+1+r)!z^\top{A_+}^r\hat{A}_+c^l.\end{aligned}$$

Therefore, (3.16b) is true if and only if

$$(-1)^{l+1+r}(l+1+r)!z^\top{A_+}^r\hat{A}_+c^l = (-1)^{l+1+r}(l)!,\quad \forall l\geq 0, r\geq 0, l+r\leq k-1,$$

which means

$$z^\top{A_+}^r\hat{A}_+c^l = \frac{l!}{(l+r+1)!},\quad \forall l\geq 0, r\geq 0, l+r\leq k-1.$$

That is guaranteed by the assumption. □

Remark 3.2.3. *If the source term is partially or fully computed implicitly, then an additional requirement*

$$z^\top{A_+}^r A_+c^l = \frac{l!}{(l+r+1)!},\quad \forall l\geq 0, r\geq 0, l+r\leq k-1$$

should be added. The proof is a line-by-line copy of the previous one.

3.3 Implicit-Explicit Runge–Kutta Methods for Linear Problems

Before we illustrate the stability theorem, we will show a stability lemma for the stages, which may be used in the later theorem.

Lemma 3.3.1. *If a series of solutions satisfy*

$$\psi = \phi + c_0\tau\mathcal{D}\psi + \tau\sum_{i=1}^{m} a_i Dv^i + \tau\sum_{i=1}^{m}\hat{a}_i\mathcal{L}v^i,$$

for some constant $c_0 > 0$, then there exists a constant C such that

$$\|\psi\|_H \le C\Big(\|\phi\|_H + \sum_{i=1}^{m}\|v_i\|_H\Big) \quad \textit{and} \quad \|\psi\|_V \le C\Big(\|\phi\|_V + \sum_{i=1}^{m}\|v_i\|_V\Big),$$

Proof. Take the inverse of $(I - c_0\tau\mathcal{D})$, we can get

$$\psi = (I - c_0\tau\mathcal{D})^{-1}\Big(\phi + \tau\sum_{i=1}^{m} a_i\mathcal{D}v^i\Big) + \tau\sum_{i=1}^{m}\hat{a}_i(I - c_0\tau\mathcal{D})^{-1}Lv^i.$$

Test with ψ and we can get

$$\|\psi\|_H^2 \le C\|\psi\|_H\Big(\|\phi\|_H + \sum_{i=1}^{m}\|v_i\|_H\Big) + \tau\sum_{i=1}^{m}\hat{a}_i\Big(\mathcal{L}v^i, (I - c_0\tau\mathcal{D})^{-1}\psi\Big)$$

and

$$\tau\Big(\mathcal{L}v^i, (I - c_0\tau\mathcal{D})^{-1}\psi\Big) \le C\tau\|v^i\|_H\|(I - c_0\tau\mathcal{D})^{-1}\psi\|_V \le C\|v^i\|_H\|\psi\|_H.$$

which gives us the first relation.

To get the second relation, instead of test ψ, we will now test $-D\psi$ to the above equation. It turns to be

$$\|\psi\|_V^2 \le C\|\psi\|_V\Big(\|\phi\|_V + \sum_{i=1}^{m}\|v_i\|_V\Big) - \tau\sum_{i=1}^{m}\hat{a}_i\Big(\mathcal{L}v^i, \mathcal{D}(I - c_0\tau\mathcal{D})^{-1}\psi\Big)$$

and

$$\tau\Big(\mathcal{L}v^i, \mathcal{D}(I - c_0\tau\mathcal{D})^{-1}\psi\Big) \le C\tau\|v^i\|_V\|\mathcal{D}(I - c_0\tau\mathcal{D})^{-1}\psi\|_H \le C\|v^i\|_V\|\psi\|_V.$$

which gives us the second relation. □

Theorem 3.3.1. *If u^i and $u^{n,i}$ are the solutions generated by 3.11, with $f = 0$, and the Assumption (P1) (P2) hold, then*

$$\|u^n\|_H^2 \le \|u^{n-1}\|_H^2, \forall n > 1$$

when $\tau \le \tau^*$, where the constant τ^* is only related to the scheme and β in equation (3.2)*, and not related to u or u^n.*

Proof. Test equation 3.14 with $\sigma^{-1}(-\tau\mathcal{D})u^n$, we can get

$$\left(u^n, \sigma^{-1}u^n\right) = \left(\sigma u^{n-1}, \sigma^{-1}u^n\right) + \tau \sum_{i=1}^{m} \left(p_i \mathcal{L} u^{n,i-1}, \sigma^{-1}u^n\right)$$

Due to the symmetry of D, all the rational combinations are symmetric and commutable, so

$$\left(u^n, \sigma^{-1}u^n\right) = \left(u^{n-1}, u^n\right) + \tau \sum_{i=1}^{m} \left(\mathcal{L} u^{n,i-1}, p_i\sigma^{-1}u^n\right)$$

Since $\mathcal{L}$ is skew-symmetric, $(\alpha\mathcal{L}w, w) = \alpha(\mathcal{L}w, w) = -\alpha(w, \mathcal{L}w) = 0$ for all α, w, so

$$\left(u^n, \sigma^{-1}u^n\right) = \left(u^{n-1}, u^n\right) + \tau \sum_{i=1}^{m} \left(\mathcal{L}\big(u^{n,i-1} - p_i(0)^{-1}p_i\sigma^{-1}u^n\big), p_i\sigma^{-1}u^n\right).$$

For our assumption, there exist an $c_0' = c_0/\beta$ s.t. $0 < \sigma(s) < 1/(1 + c_0'\beta s)$, so the LHS of the above equation can be bounded by

$$\left(u^n, \sigma^{-1}u^n\right) > \left\|u^n\right\|_H^2 + c_0'\tau\left\|u^n\right\|_V^2$$

and

$$\left(u^{n-1}, u^n\right) = \frac{1}{2}\left\|u^n\right\|_H^2 + \frac{1}{2}\left\|u^{n-1}\right\|_H^2 - \frac{1}{2}\left\|u^n - u^{n-1}\right\|_H^2.$$

For the second term on the RHS, use the property of L, we can get

$$\begin{aligned}
&\sum_{i=1}^{m} \left(\mathcal{L}\big(u^{n,i-1} - p_i(0)^{-1}p_i\sigma^{-1}u^n\big), p_i\sigma^{-1}u^n\right) \\
&\quad\le \varepsilon \sum_{i=1}^{m} \left\|p_i\sigma^{-1}u^n\right\|_V^2 + C_\varepsilon \sum_{i=1}^{m} \left\|u^{n,i-1} - p_i(0)^{-1}p_i\sigma^{-1}u^n\right\|_H^2 \\
&\quad\le \varepsilon \sum_{i=1}^{m} \left\|p_i\sigma^{-1}u^n\right\|_V^2 + 2C_\varepsilon \sum_{i=1}^{m} \left\|u^{n,i-1} - u^n\right\|_H^2 + 2C_\varepsilon \sum_{i=1}^{m} \left\|u^n - p_i(0)^{-1}p_i\sigma^{-1}u^n\right\|_H^2 \\
&\quad= I_1 + I_2 + I_3
\end{aligned}$$

By corollary 3.2.1, we can get

$$I_1 = \varepsilon \sum_{i=1}^{m} \left\|p_i\sigma^{-1}u^n\right\|_V^2 < \varepsilon_1 C\left\|u^n\right\|_V^2.$$

Then we are going to estimate I_2. The second relation in 3.11 gives us

$$u^{n,i} - u^n = (u^{n-1} - u^n) + \tau \sum_{j=1}^{i} a_{ij}\mathcal{D}(u^{n,j} - u^n) + \tau \sum_{j=1}^{i} \hat{a}_{ij}\mathcal{L}(u^{n,j-1} - u^n) + \tau c_i(\mathcal{D} + \mathcal{L})u^n.$$

Define $w^{n,i} = u^{n,i} - u^n$, so

$$w^{n,i} = w^{n,0} + \tau \sum_{j=1}^{i} a_{ij}\mathcal{D}w^{n,j} + \tau \sum_{j=1}^{i} \hat{a}_{ij}\mathcal{L}w^{n,j-1} + \tau c_i(\mathcal{D} + \mathcal{L})u^n.$$

Similar with Lemma 3.3.1, taking the inverse of $\big(I - a_{ii}\tau\mathcal{D}\big)$, we can derive that

$$\begin{aligned} w^{n,i} = {} & \tau \sum_{j=1}^{i-1} a_{ij}\big(I - a_{ii}\tau\mathcal{D}\big)^{-1}\mathcal{D}w^{n,j} + \tau \sum_{j=1}^{i} \hat{a}_{ij}\big(I - a_{ii}\tau\mathcal{D}\big)^{-1}\mathcal{L}w^{n,j-1} \\ & + \big(I - a_{ii}\tau\mathcal{D}\big)^{-1}w^{n,0} + \tau c_i\big(I - a_{ii}\tau\mathcal{D}\big)^{-1}\mathcal{D}u^n + \tau c_i\big(I - a_{ii}\tau\mathcal{D}\big)^{-1}\mathcal{L}u^n. \end{aligned}$$

Test with $w^{n,i}$, we can derive that

$$\begin{aligned} \|w^{n,i}\|_H^2 \le {} & C\sum_{j=1}^{i-1} a_{ij}\|w^{n,j}\|_H^2 + \tau \sum_{j=1}^{i} \hat{a}_{ij}\big(\mathcal{L}w^{n,j-1}, \big(I - a_{ii}\tau\mathcal{D}\big)^{-1}w^{n,i}\big) \\ & + \|w^{n,0}\|_H\|w^{n,i}\|_H + C\tau^{1/2}\|u^n\|_V\|w^{n,i}\|_H + \tau c_i\big(\mathcal{L}u^n, \big(I - a_{ii}\tau\mathcal{D}\big)^{-1}w^{n,i}\big), \end{aligned}$$

and

$$\big(\mathcal{L}\phi, \big(I - a_{ii}\tau\mathcal{D}\big)^{-1}\chi\big) \le C\|\phi\|_H\|\big(I - a_{ii}\tau\mathcal{D}\big)^{-1}\chi\|_V \le C\|\phi\|_H\|\chi\|_H,$$

for any ϕ, χ. Therefore

$$\|w^{n,i}\|_H \le C\|w^{n,0}\|_H + C\sum_{j=1}^{i-1} a_{ij}\|w^{n,j}\|_H + C\tau^{1/2}\|u^n\|_V.$$

Accumulate from 1 to i, we can derive that

$$\|w^{n,i}\|_H \le C\|w^{n,0}\|_H + C\tau^{1/2}\|u^n\|_V,$$

which means that

$$\|u^{n,i} - u^n\|_H^2 \leq C\|u^{n-1} - u^n\|_H^2 + C\tau\|u^n\|_V^2.$$

For I_3, since $p_i\sigma^{-1}$ is bounded, we know that $\left(1 - \frac{p_i(s)}{p_i(0)\sigma(s)}\right)^2$ is also bounded for $s \geq 0$. Further, it is a rational function with value 0 when $s = 0$, so

$$\left(1 - \frac{p_i(s)}{p_i(0)\sigma(s)}\right)^2 \leq Cs, \forall s \geq 0.$$

which means that

$$I_3 \leq C_\varepsilon\tau\left\|u^n\right\|_V^2.$$

Now, we can get that

$$\begin{aligned}\left\|u^n\right\|_H^2 =& - c_0\tau\left\|u^n\right\|_V^2 + \frac{1}{2}\left\|u^n\right\|_H + \frac{1}{2}\left\|u^{n-1}\right\|_H^2 - \frac{1}{2}\left\|u^n - u^{n-1}\right\|_H^2 \\ &+ \tau\varepsilon C\sum_{i=1}^{m}\left\|u^n\right\|_V^2 + C_\varepsilon\tau\|u^{n-1} - u^n\|_H^2 + C_\varepsilon\tau^2\|u^n\|_V^2 \\ &+ C_\varepsilon\tau^2\left\|u^n\right\|_V^2\end{aligned}$$

For a given IMEX–RK scheme, ε can be fixed so that C_ε is also a fixed value, which is non-related to the time step τ and solution u. Then a small τ can guarantee that $C_\varepsilon\tau\|u^{n-1} - u^n\|_H^2$ and $C_\varepsilon\tau^2\left\|u^n\right\|_V^2$ are bounded by the negative terms. So finally we find

$$\left\|u^n\right\|_H^2 \leq \left\|u^{n-1}\right\|_H^2,$$

which agrees with our claimant. □

Next, we shall derive an error estimate for the scheme.

Theorem 3.3.2. *Suppose that Assumptions (P1)-(P3) are valid, u is the solution of 3.1, and u^n is the solution of 3.11. We can then derive the following error estimate:*

$$\|u(t^n) - u^n\|_H \leq C\tau^k\left(\int_0^{t^n}\|\mathcal{A}^{k+1}u(s)\|_H + \|u^{(k+1)}(s)\|_H + \|\mathcal{A}^{k+1}u'(s)\|_H\,\mathrm{d}s + \sum_{l=0}^{k-1}\int_0^{t^n}\|\mathcal{A}^l f^{(k-l)}(s)\|_H\,\mathrm{d}s\right)$$

when $\tau \leq \tau^*$, where the constant τ^* is related only to the scheme and β in equation (3.2)*, and not to u or u^n.*

Proof. To begin with, we shall examine the truncation error for our scheme. Let $U_*^n = \{u(t^{n,i}\}_{i=0}^m$ be the exact solution. We define the local truncation error R^n as

$$U_*^n = \mathbb{1}u(t^n) + \tau A_+ \mathcal{D} U_*^n + \tau \hat{A}_+ (\mathcal{L} U_*^n + F^n) + R^n. \tag{3.19}$$

together with

$$U^n = \mathbb{1}u^n + \tau A_+ \mathcal{D} U^n + \tau \hat{A}_+ (\mathcal{L} U^n + F^n). \tag{3.20}$$

Define $e^{n,i} = u(t^{n,i}) - u^{n,i}$, and vector E^n be consisted with $e^{n,i}$. we can get

$$E^n = \mathbb{1}e^n + \tau A_+ \mathcal{D} E^n + \tau \hat{A}_+ \mathcal{L} E^n + R^n.$$

Divide E^n into two vectors, such that

$$E_1^n = \mathbb{1}e^n + \tau A_+ \mathcal{D} E_1^n + \tau \hat{A}_+ \mathcal{L} E_1^n,$$

$$E_2^n = \tau A_+ \mathcal{D} E_2^n + \tau \hat{A}_+ \mathcal{L} E_2^n + R^n.$$

Obviously $E^n = E_1^n + E_2^n$, and $\|z^\top E_1^n\|_H \leq \|e^n\|_H$ when $\tau < \tau^*$ by Theorem 3.3.1.

From equation (3.19), we can get

$$R^n = (I - \tau A_+ \mathcal{D} - \tau \hat{A}_+ \mathcal{L}) U_*^n - \mathbb{1}u(t^n) - \tau \hat{A}_+ F^n.$$

Substitute this into the relation of E^2, we can get

$$E_2^n = U_*^n - (I - \tau A_+ \mathcal{D} - \tau \hat{A}_+ \mathcal{L})^{-1} (\mathbb{1}u(t^n) + \tau \hat{A}_+ F^n).$$

The Taylor expansion gives us that

$$U_*^n = \mathbb{1}u(t^n) + \sum_{l=1}^{k} \frac{1}{l!} \tau^l c^l u^{(l)}(t^n) + O(\tau^{k+1}).$$

and

$$\begin{aligned} u^{(l)}(t^n) &= (\mathcal{D}+\mathcal{L})u^{(l-1)}(t^n) + f^{(l-1)}(t^n) = \dots \\ &= (\mathcal{D}+\mathcal{L})^l u(t^n) + \sum_{p=0}^{l-1} (\mathcal{D}+\mathcal{L})^{l-1-p} f^{(p)}(t^n). \end{aligned}$$

Although $A_+D + \hat{A}_+L$ can be non-self-adjoint and unbounded, the laws of finite products hold, so that

$$(I-X)^{-1} = (I + X + \cdots + X^k) + (I-X)^{-1}X^{k+1}$$

where X can be $(\tau A_+\mathcal{D} + \tau\hat{A}_+\mathcal{L})$.

Finally, the Taylor expansion of F is given by:

$$F^n = \mathbb{1} f(t^n) + \sum_{l=1}^{k-1} \frac{1}{l!}\tau^l c^l f^{(l)}(t^n) + O(\tau^k).$$

Note that we only need the last element of E_2, and the Assumptions (P3) guarantee that:

$$z^\top (A_+\mathcal{D} + \hat{A}_+\mathcal{L})^l \mathbb{1} = \frac{1}{l!}(\mathcal{D}+\mathcal{L})^l = \frac{1}{l!}\mathcal{A}^l$$

and

$$z^\top (A_+\mathcal{D} + \hat{A}_+\mathcal{L})^r \hat{A}_+ c^l = \frac{l!}{(l+1+r)!}(\mathcal{D}+\mathcal{L})^r = \frac{l!}{(l+1+r)!}\mathcal{A}^r.$$

Combining all the above equations, we observe that all the lower-order terms are canceled, so

$$z^\top E_2 = C\tau^{k+1}$$

and

$$\begin{aligned} \|e^{n+1}\| \le& \|e^n\| + C\tau^{k+1}\|\mathcal{A}^{k+1}u(t^n)\|_H + C\tau^k \int_{t^n}^{t^{n+1}} \|u^{(k+1)}\|_H \,\mathrm{d}s \\ &+ C\tau^k \sum_{l=0}^{k-1} \int_{t^n}^{t^{n+1}} \|\mathcal{A}^l f^{(k-l)}\|_H \,\mathrm{d}s, \end{aligned}$$

where the constant C is related only to the scheme itself.

Furthermore, we have

$$\begin{aligned}\tau\|\mathcal{A}^{k+1}u(t^n)\|_H &= \int_{t^n}^{t^{n+1}} \|\mathcal{A}^{k+1}u(s)\|_H \mathrm{d}s \\ &\quad + \int_{t^n}^{t^{n+1}} \|\mathcal{A}^{k+1}u(t^n)\|_H - \|\mathcal{A}^{k+1}u(s)\|_H \mathrm{d}s \\ &\leq \int_{t^n}^{t^{n+1}} \|\mathcal{A}^{k+1}u(s)\|_H \mathrm{d}s + \tau \int_{t^n}^{t^{n+1}} \|\mathcal{A}^{k+1}u'(s)\|_H \mathrm{d}s.\end{aligned}$$

As a result, if $\tau \leq \tau^*$ which is introduced in Theorem 3.3.1, we can get

$$\begin{aligned}\|e^n\| \leq &\|e^0\| + C\tau^k \int_0^{t^n} \|\mathcal{A}^{k+1}u(s)\|_H \mathrm{d}s + C\tau^{k+1} \int_0^{t^n} \|\mathcal{A}^{k+1}u'(s)\|_H \mathrm{d}s \\ &+ C\tau^k \int_0^{t^n} \|u^{(k+1)}\|_H \mathrm{d}s + C\tau^k \sum_{l=0}^{k-1} \int_0^{t^n} \|\mathcal{A}^l f^{(k-l)}\|_H \mathrm{d}s.\end{aligned}$$

If the exact solution and source term is bounded in $(0, \infty)$, then we can get the long time error estimate. Here the constant C is only related to the scheme, and τ^* is only related to the scheme and β in equation (3.2). Neither of them are related to the source term f, the exact solution u, the numerical solution u^n, or the mesh size τ. □

Remark 3.3.1. *Note that each possible combination of Assumption (P3) has appeared in the above proof during the Taylor's expansion, so the stage order Assumption (P3) is also necessary conditions.*

3.4 Implicit-Explicit Runge–Kutta Methods for Semilinear Problems

In this section, we will extend IMEX-RK method to equation (1.1). Unlike the simple cut-off post-processing used before, we first perform a modification on the potential term to ensure their solvability on the stages, and then apply cut-off post-processing at the final stage of each step. Combined with [38], we can show that IMEX-RK can preserve both the maximum bound and the original energy dissipation.

To solve the Allen-Cahn equation (1.1), we begin with defining a modification nonlinear term $\hat{f}$ as

$$\hat{f}(v) = \begin{cases} f(v), & \text{if } |v| \le \alpha, \\ f'(\alpha)(v-\alpha), & \text{if } v > \alpha, \\ f'(-\alpha)(v+\alpha), & \text{if } v < -\alpha, \end{cases} \tag{3.21}$$

and consider the modified model

$$u_t = \Delta u + \hat{f}(u). \tag{3.22}$$

Note that $f(u) = \hat{f}(u)$ for $|u| \le \alpha$, thus the equation (1.1) and (3.22) share the same exact solution. Moreover, since the modification is tangent cutoff of the original function, we can know that $f \in H^2(\mathbb{R})$.

The IMEX-RK method for solving equation (3.22) is

$$\begin{cases} u^{ni} = u^{n-1} + \tau \sum_{j=0}^{m} a_{ij} \Delta u^{nj} + \tau \sum_{j=0}^{m} \hat{a}_{ij} \hat{f}(u^{nj}) & \text{for } i = 1, 2, \dots, m, \\ E(\tau) u^{n-1} = u^{n-1} + \tau \sum_{i=0}^{m} b_i \Delta u^{ni} + \tau \sum_{i=0}^{m} \hat{b}_i \hat{f}(u^{ni}), \\ u^n = E(\tau) u^{n-1}. \end{cases} \tag{3.23}$$

This scheme naturally defines a solution map $E(\tau) : u^{n-1} \mapsto u^n$. The map $E(\tau)$ satisfies the following Lipschitz condition.

Theorem 3.4.1. *The operator $E(\tau)$ defined in equation* (3.23) *satisfies that*

$$\|E(\tau)v - E(\tau)w\| \le (1 + C\tau)\|v - w\|$$

for all $v, w \in L^2(\Omega)$. Here $\|\cdot\|$ refers to L^2 norm.

Proof. Define $\mathbf{e}_i$ as the vector with the $(i+1)$-th entry as 1 and others as 0. Let $U^n = [u^{n,0}, u^{n,1}, \dots]^\top$, where $u^{n,i} = \mathbf{e}_i^\top U^n$, and $\hat{f}(U^n) = [\hat{f}(u^{n,0}), \hat{f}(u^{n,1}), \dots]^\top$. In vector form, we have:

$$U^n = \mathbb{1} u^{n-1} + \tau A \Delta U^n + \tau \hat{A} f(U^n),$$

and hence,

$$U^n = (I - \tau A\Delta)^{-1} \mathbb{1} u^{n-1} + \tau (I - \tau A\Delta)^{-1} \hat{A} \hat{f}(U^n).$$

Test the above realtion with $\mathbf{e}_i$, similar to the linear case we can get

$$\begin{aligned} u^{ni} &= \mathbf{e}_i^\top (I - \tau A\Delta)^{-1} \mathbb{1} u^{n-1} + \tau \mathbf{e}_i^\top (I - \tau A\Delta)^{-1} \hat{A} \hat{f}(U^n), \\ &=: \sigma_i(-\tau\Delta) u^{n-1} + \tau \sum_{j=0}^{i-1} p_{ij}(-\tau\Delta) \hat{f}(u^{nj}). \end{aligned}$$

Define v^i and w^i as the stages corresponding to v and w. Similar to the linear case, σ_i and p_{ij} are also bounded operators. We have the inequality

$$|v^i - w^i| \le |\sigma_i(v-w)| + \tau \sum_{j=0}^{i-1} \left| p_{ij} \left(\hat{f}(v^j) - \hat{f}(w^j) \right) \right| \le |v-w| + CL\tau \sum_{j=0}^{i-1} \left| v^j - w^j \right|$$

Combining these inequalities for all i and tracing back to the first stage, we conclude:

$$\|E(\tau)(v-w)\| \le (1 + C\tau)\|v-w\|,$$

where the constant C is related to the scheme and Lipschitz constant of the source term.

□

Later we will show the consistency error of semilinear IMEX-RK method.

Theorem 3.4.2. *Suppose that the Assumption (P1)-(P3) are valid for $k \ge 3$, u is sufficiently smooth on both space and time, then the operator $E(\tau)$ defined in equation* (3.23) *satisfies that*

$$\|E(\tau)u(t^{n-1}) - u(t^n)\| \le C\tau^{k+1}, \qquad k = 1, 2, 3 \tag{3.24}$$

for all $n \ge 1$. Here $\|\cdot\|$ refers to L^2 norm. Furthermore, in condition that

$$(b_* \cdot c)^\top A_* c = \frac{1}{8}$$

works for all $A_* \in \{A_+, \hat{A}_+\}$, $b_* \in \{b_+, \hat{b}_+\}$, equation (3.24) *hold for $k = 4$. Here $v \cdot w$ means the multiplication elementwisely.*

Proof. Equation (3.4) gives us that

$$U^n - \mathbb{1}u(t^{n-1}) = \tau A\Delta U^n + \tau\hat{A}\hat{f}(U^n) = O(\tau). \tag{3.25}$$

and

$$\begin{aligned} U^n - \mathbb{1}u(t^{n-1}) = &\ \tau\left(A\mathbb{1}\Delta u(t^{n-1}) + \hat{A}\mathbb{1}\hat{f}(u(t^{n-1}))\right) \\ &+ \tau A\Delta\left(U^n - \mathbb{1}u(t^{n-1})\right) + \tau\hat{A}\left(\hat{f}(U^n) - \mathbb{1}\hat{f}(u(t^{n-1}))\right) \\ = &\ \tau u_t^0 + O(\tau^2), \end{aligned} \tag{3.26}$$

where $u_t^0 = A\Delta\mathbb{1}u(t^{n-1}) + \hat{A}\mathbb{1}f(u(t^{n-1}))$.

Substitute equation (3.26) into itself again, we can derive that

$$\begin{aligned} U^n - \mathbb{1}u(t^{n-1}) = &\ \tau\left(A\mathbb{1}\Delta u(t^{n-1}) + \hat{A}\mathbb{1}\hat{f}(u(t^{n-1}))\right) \\ &+ \tau A\Delta\left(U^n - \mathbb{1}u(t^{n-1})\right) \\ &+ \tau\hat{A}\left(\hat{f}_u\left(U^n - \mathbb{1}u(t^{n-1})\right) + O(\tau^2)\right) \\ = &\ \tau u_t^0 + \tau A\Delta\cdot\tau u_t^0 + \tau\hat{A}\hat{f}_u\cdot\tau u_t^0 + O(\tau^3) \\ = &\ \tau u_t^0 + \tau^2 u_{tt}^0 + O(\tau^3), \end{aligned} \tag{3.27}$$

where $u_{tt}^0 = A\Delta u_t^0 + \hat{A}\hat{f}_u u_t^0$.

Denote $v^{\cdot i}$ as the i-th power elementwisely and pointwisely, then

$$\begin{aligned} U^n - \mathbb{1}u(t^{n-1}) = &\ \tau\left(A\mathbb{1}\Delta u(t^{n-1}) + \hat{A}\mathbb{1}\hat{f}(u(t^{n-1}))\right) \\ &+ \tau A\Delta\left(U^n - \mathbb{1}u(t^{n-1})\right) \\ &+ \tau\hat{A}\left(\hat{f}_u\left(U^n - \mathbb{1}u(t^{n-1})\right) + \frac{1}{2}\hat{f}_{uu}\left(U^n - \mathbb{1}u(t^{n-1})\right)^{\cdot 2} + O(\tau^3)\right) \\ = &\ \tau u_t^0 + \tau A\Delta\cdot\left(\tau u_t^0 + \tau^2 u_{tt}^0\right) \\ &+ \tau\hat{A}\hat{f}_u\cdot\left(\tau u_t^0 + \tau^2 u_{tt}^0\right) + \tau\hat{A}\frac{1}{2}\hat{f}_{uu}(\tau u_t^0)^{\cdot 2} + O(\tau^4) \\ = &\ \tau u_t^0 + \tau^2 u_{tt}^0 + \tau^3 u_{ttt}^0 + O(\tau^4), \end{aligned} \tag{3.28}$$

where $u^0_{ttt} = A\Delta u^0_{tt} + \hat{A}\hat{f}_u u^0_{tt} + \frac{1}{2}\hat{A}\hat{f}_{uu}(u^0_t)^{\cdot 2}$. Similarily, we reached that

$$U^n - \mathbb{1}u(t^{n-1}) = \tau u^0_t + \tau^2 u^0_{tt} + \tau^3 u^0_{ttt} + \tau^4 u^0_{tttt} + O(\tau^5), \tag{3.29}$$

where $u^0_{tttt} = A\Delta u^0_{ttt} + \hat{A}\hat{f}_u u^0_{ttt} + \hat{A}\hat{f}_{uu} u^0_t \cdot u^0_{tt} + \frac{1}{6}\hat{A}(u^0_t)^{\cdot 3}$. Test the above equations with $\mathbf{e}_m$, and we can get

$$E(\tau)u(t^{n-1}) = u(t^{n-1}) + \mathbf{e}_m^\top \left(\tau u^0_t + \tau^2 u^0_{tt} + \tau^3 u^0_{ttt} + \tau^4 u^0_{tttt}\right) + O(\tau^5),$$

where the RHS is only depend on $u(t^{n-1})$. Assume that the exact solution u is smooth enough, and compare it with the Taylor's expansion, we can find that all the low-order terms disappeared, thus the theorem is proved. □

After proving consistency, the final convergence result follows naturally.

Corollary 3.4.1. *Suppose that the Assumption (P1), (P2) and (P3) are valid for $k \geq 4$, and $(b_* \cdot c)^\top A_* c = \frac{1}{8}$ works for all $A_* \in \{A_\sigma, \hat{A}_\sigma\}$, $b_* \in \{b_\sigma, \hat{b}_\sigma\}$ if $k = 4$. The exace solution u of equation* (1.1) *is sufficiently smooth on both space and time u^n is the solution of* (3.23)*, then*

$$\|u^n - u(t^n)\| \leq C\tau^k. \tag{3.30}$$

Proof. Combining the estimate of consistency error in Theorem 3.4.2 and the stability estimate in Theorem 3.4.1, we derive

$$\begin{aligned} \|u^n - u(t^n)\| &= \|E(\tau)u^{n-1} - u(t^n)\| \\ &\leq \|E(\tau)u^{n-1} - E(\tau)u(t^{n-1})\| + \|E(\tau)u(t^{n-1}) - u(t^n)\| \\ &\leq (1 + C\tau)\|u^{n-1} - u(t^{n-1})\| + C\tau^{k+1}. \end{aligned} \tag{3.31}$$

This estimate, along with Grönwall's inequality, completes the proof.

□

Remark 3.4.1. *Since this is a single-step method, as what we have done in equation* (2.3)*, we can build a scheme that*

$$u^n = \min(\max(E(\tau)u^{n-1}, -\alpha), \alpha), \tag{3.32}$$

The exact solution will always meet the maximum bound condition, so

$$\|u^n - u(t^n)\| \leq \|E(\tau)u^{n-1} - u(t^n)\|.$$

The rest part of the proof is the same to (3.31).

Remark 3.4.2. *This analysis is consistent with the work by Fu and Yang [38], where they prove that certain IMEX-RK schemes can maintain the energy dissipation law using a stabilizer. We have identified schemes of first, second, and third order, demonstrating that a third-order IMEX-RK scheme can preserve both the maximum bound principle and the energy dissipation law.*

3.5 Construction and List of Implicit-Explicit Runge–Kutta Schemes

In this part we make a short introduction to how we find the table of IMEX-RK and list some qualified IMEX-RK schemes.

In this section, we first provide a brief introduction to searching the IMEX Runge–Kutta table, followed by some typical examples of IMEX Runge–Kutta schemes that satisfy Assumptions (P1)-(P3).

Our searching algorithm is based on undetermined coefficient method. The algorithm is also listed in Algorithm 1. We aim to find a k-th order scheme in this subsection.

We note that for a m-stage IMEX Runge–Kutta table, the total degree of freedoms is about m^2, but the number of equations is about 2^k. Because the all possibilities of combination from $b, \hat{b}$ and $A, \hat{A}$ should keep the relations. The number of stages will be far more than the scheme order.

However, when we focus on the implicit table A itself, the scheme is reduced to a DIRK scheme, and the total number of relations is algebraic to k. In this case we choose a small m fixed and find a candidate A to the DIRK scheme, which is not difficult to achieve.

In fact, plenty of degree of freedoms are available in the DIRK table searching. Since c_i is high order in the stage order requirements, In Line 1 we fix it at the beginning. Then we solve A and b in Line 2 and 6. If there are other requirements on A, like $\sigma(\lambda) > 0$, we will test it here in Line 10.

What follows is to solve $\hat{A}$ once A is generated in a small size. The current number of stages may not be enough fulfill all requirements of the stages orders, so additional stages are necessary. Note that since the number of stages is always far more than the scheme order for high order schemes, it is reasonable to

add more stages.

To add more stages, we only need to repeat the last stage in the implicit table A. The scheme in Table 3.9 is an example for what we should do to add a stage. If we only focus on the implicit part, literally we did nothing and all the stages order conditions and requirements on σ will not change, so we do not need to retest any former conditions for they will be kept naturally, which is what we are doing in Line 19 and 24. However extra stages give us extra degree of freedoms so we can solve a larger system, the unsolvable equations may turn to be solvable.

Last step is to is to search $\hat{a}$ which can be accessed by undetermined coefficient method, in Line 22. Since we can always add extra stages, which will provide more degree of freedoms for solving, but will not raise the requirements, it is reasonable to make a conjecture that there exist arbitrary high order IMEX Runge–Kutta schemes.

The followings are some IMEX-RK schemes that satisfy our Assumption (P1)-(P3).

(i) **First-order scheme** The following Butcher tableau Tab.3.2 gives us a first order IMEX RK scheme.

This scheme agrees with Remark 3.4.2. In this example,

$$\sigma(\lambda) = \frac{1}{1+\lambda}, \forall \lambda > 0.$$

(ii) **Second-order scheme** The following Butcher tableau Tab.3.3 gives us a second order IMEX RK scheme.

in which $\gamma = 1 + \frac{\sqrt{2}}{2}$, $\delta = 1 - \frac{1}{2\gamma}$. This scheme agrees with Remark 3.4.2. In this example,

$$\sigma(\lambda) = \frac{1 + (1+\sqrt{2})\lambda}{(1 + (1+\sqrt{2}/2)\lambda)^2}, \forall \lambda > 0$$

(iii) **Third-order scheme** The following Butcher tableau Tab.3.4 gives us a third order IMEX Runge–Kutta scheme.

This scheme does not agree with Remark 3.4.2. In this scheme,

$$\sigma(\lambda) = \frac{16(48 - 6\lambda^2 + \lambda^3)}{3(4+\lambda)^4}, \forall \lambda > 0.$$

Algorithm 1: Searching Algorithm for Qualified IMEX-RK Schemes

Input : Given order k
Output: A, $\hat{A}$, b, $\hat{b}$, c

1 Choose one possible c with $m = len(c)$, which is the number of the stages;
2 Try to solve b from $b^\top c^l = \frac{1}{l+1}$;
3 **if** b *is not solvable* **then**
4 | $m = m + 1$; Goto Line 1 for a new c;
5 **end**
6 Solve A from $z^\top A^{r+1} c^l = \frac{l!}{(l+r+1)!}$ with $z^\top A = b^\top$;
7 **if** A *is not solvable* **then**
8 | $m = m + 1$; Goto Line 1 for a new c;
9 **end**
10 Evaluate $\sigma(\lambda) = z^\top (1 + \lambda A_+)^{-1} \mathbb{1}$;
11 **if** $0 < \sigma < 1$ *fails* **then**
12 | **if** $0 < \sigma < 1$ *fails too many times* **then** `// Ususlly, we do not need it`
13 | | $m = m + 1$;
14 | **end**
15 | Goto Line 1 for a new c;
16 **end**
17 Solve $\hat{b}$ from the order relations with $\hat{b}^\top = z^\top \hat{A}$. ;
18 **if** $\hat{b}$ *is not solvable* **then**
19 | $m = m + 1$; Duplicate the final stage of A and c; Insert a 0 before the last element of b;
20 | Goto Line 17.
21 **end**
22 Solve $\hat{A}$ from the order relations with $\hat{b}^\top = z^\top \hat{A}$. ;
23 **if** $\hat{A}$ *is not solvable* **then**
24 | $m = m + 1$; Duplicate the final stage of A and c; Insert a 0 before the last element of b;
25 | Goto Line 17.
26 **end**
27 Print A, $\hat{A}$, b, $\hat{b}$, c

1	1	1
	1	1

Table 3.2: Butcher tableau for first order IMEX-RK

γ	γ		γ	
1	$1-\gamma$	γ	δ	$1-\delta$
	$1-\gamma$	γ	δ	$1-\delta$

Table 3.3: Butcher tableau for second order IMEX-RK

Although there are some negative coefficients in the numerator, it is still positive for all $\lambda \geq 0$.

(iv) **Third-order energy diminishing scheme**

The following Butcher tableau Tab.3.5 gives us a third order IMEXRK scheme.

This scheme agrees with Remark 3.4.2. In this example,

$$\sigma(\lambda) = \frac{1228800 + 33778560\lambda + 55256268\lambda^2 + 5250325\lambda^3}{1228800 + 35007360\lambda + 89649228\lambda^2 + 77600673\lambda^3 + 21689019\lambda^4}, \forall \lambda > 0$$

(v) **Forth-order scheme for linear problem**

The following Butcher tableau Tab.3.6 gives us an IMEXRK scheme which is fourth order for linear problem and third order for semilinear problem.

This scheme does not agree with Remark 3.4.2. In this example,

$$\sigma(\lambda) = \frac{75000 - 7500\lambda^2 + 1000\lambda^3 + 225\lambda^4}{24(5+\lambda)^5}, \forall \lambda > 0.$$

Although there are some negative coefficients in the numerator, it is still positive for all $\lambda \geq 0$.

(vi) **Forth-order IMEX scheme for linear and semilinear problem**

The following Butcher tableau Tab.3.7 gives us an IMEXRK scheme which is fourth order for linear and semilinear problem.

Related coefficients are listed in Table 27. This scheme does not agree with Remark 3.4.2. In this example,

$$\sigma(\lambda) = -\Big(- 18533185137819\lambda^5 + 245682733504208\lambda^4 - 1917903570331840\lambda^3$$

$1/4$	$1/4$				$1/4$	$1/4$			
$1/2$	$1/4$	$1/4$			$1/2$	$1/4$	$1/4$		
$3/4$	$-4/5$	$7/4$	$1/4$		$3/4$	1	$1/4$	$-1/2$	
1	$5/12$	$5/12$	$-1/12$	$1/4$	1	$5/6$	$-11/6$	$13/6$	$-1/6$
	$5/12$	$5/12$	$-1/12$	$1/4$		$5/6$	$-11/6$	$13/6$	$-1/6$

Table 3.4: Butcher tableau for four-stage third-order IMEX Runge–Kutta scheme

$3/5$	$3/5$			
$3/2$	$15/32$	$33/32$		
$19/20$	$2/5$	$-357/640$	$709/640$	
1	$2825/756$	$-232/297$	$-6400/231$	$103/4$
	$2825/756$	$-232/297$	$-6400/231$	$103/4$
$3/5$	$3/5$			
$3/2$	$51/64$	$45/64$		
$19/20$	$2/5$	$4841/11520$	$299/2304$	
1	$103/342$	$125/378$	$-26/297$	$2000/4389$
	$103/342$	$125/378$	$-26/297$	$2000/4389$

Table 3.5: Butcher tableau for third order IMEX-RK

$$+ 4891758175886400\lambda^2 + 9437315024592000\lambda - 61167782566800000\Big)$$
$$/\Big(14794928512(9\lambda+25)(\lambda+7)^2(\lambda+15)^3\Big), \forall\lambda > 0.$$

Although there are some negative coefficients in the numerator, it is still positive for all $\lambda \geq 0$.

3.6 Numerical Result

We consider the Stokes-Darcy coupled system described by Eqs. (3.3)-(3.4) in this example. Numerical scheme (3.11) applied on this system is formulated as follows:

$$\begin{cases} \phi_{n,0} = \phi^{n-1} \\ \phi_{n,i} = \phi_{n,0} + \tau \sum_{j=1}^{i} a_{ij}(-A_1)\phi_{n,j} + \tau \sum_{j=1}^{i} \hat{a}_{ij}(B\mathbf{u}_{n,j-1} + f_D(t^{n,j-1})), \quad i = 1, 2, \cdots, s \\ \phi^n = \phi_{n,s} \end{cases} \tag{3.33}$$

$$
\begin{array}{c|ccccc}
\frac{1}{5} & \frac{1}{5} & & & & \\
\frac{2}{5} & \frac{1}{5} & \frac{1}{5} & & & \\
\frac{3}{5} & \frac{7}{5} & -1 & \frac{1}{5} & & \\
\frac{4}{5} & \frac{3989}{1025} & -\frac{3343}{1025} & -\frac{31}{1025} & \frac{1}{5} & \\
1 & \frac{79}{120} & -\frac{91}{120} & \frac{149}{120} & -\frac{41}{120} & \frac{1}{5} \\
1 & \frac{79}{120} & -\frac{91}{120} & \frac{149}{120} & -\frac{41}{120} & 0 & \frac{1}{5} \\
\hline
 & \frac{79}{120} & -\frac{91}{120} & \frac{149}{120} & -\frac{41}{120} & 0 & \frac{1}{5}
\end{array}
\qquad
\begin{array}{c|ccccc}
\frac{1}{5} & \frac{1}{5} & & & & \\
\frac{2}{5} & \frac{5299}{30120} & \frac{1357}{6042} & & & \\
\frac{3}{5} & \frac{8681383}{58516770} & \frac{5413353}{9752795} & -\frac{6051439}{58516770} & & \\
\frac{4}{5} & 0 & \frac{25905769}{16101930} & -\frac{9808291}{16101930} & -\frac{10113}{50635} & \\
1 & 0 & \frac{79}{120} & -\frac{55873}{32520} & \frac{96299}{32520} & -\frac{5863}{6504} \\
1 & 0 & \frac{79}{120} & -\frac{91}{120} & \frac{149}{120} & -\frac{41}{120} & \frac{1}{5} \\
\hline
 & 0 & \frac{79}{120} & -\frac{91}{120} & \frac{149}{120} & -\frac{41}{120} & \frac{1}{5}
\end{array}
$$

Table 3.6: Butcher tableau for fourth order IMEX-RK

$$
\begin{cases}
\mathbf{u}_{n,0} = \mathbf{u}^{n-1} \\
\mathbf{u}_{n,i} = \mathbf{u}_{n,0} + \tau \sum_{j=1}^{i} a_{ij}(-A_2)\mathbf{u}_{n,j} + \tau \sum_{j=1}^{i} \hat{a}_{ij}(-gB^*\phi_{n,j-1} + \mathbf{f}_S(t^{n,j-1})), \quad i = 1, 2, \cdots, s \\
\mathbf{u}^n = \mathbf{u}_{n,s}
\end{cases}
\tag{3.34}
$$

which is a decoupled and linear scheme. We test temporal convergence of numerical schemes aforementioned.

Parameters are set $\nu = \mu = \kappa = g = 1$ for simplicity. Problem domain is set to be a unit square centered at the origin. Darcy flow and Stokes flow occupy the upper half and lower half domain as shown in Fig.3.1. To verify the temporal convergence order, we fix the spatial step size $h = 1/100$ and calculate numerical solutions $(\phi_n^\tau, \mathbf{u}_n^\tau)$ with various τ and the convergence order is obtained as

$$
\text{order } = \log\left(\frac{\|w_n^\tau - w_n^{\tau/2}\|}{\|w_n^{\tau/2} - w_n^{\tau/4}\|}\right) \Big/ \log(2) \tag{3.35}
$$

where w denotes either ϕ or $\mathbf{u}$. In addition to second-order scheme described in Tab.3.3, we numerically test convergence order for four-stage third-order scheme and six-stage fourth-order scheme as well described in Tab.3.9.

$\frac{1}{7}$	$\frac{1}{7}$						
$\frac{2}{7}$	$\frac{1}{7}$	$\frac{1}{7}$					
$\frac{11}{25}$	$\frac{1354}{4375}$	$-\frac{104}{4375}$	$\frac{27}{175}$				
$\frac{4}{5}$	a_{41}	a_{42}	a_{43}	$\frac{9}{25}$			
$\frac{13}{15}$	a_{51}	a_{52}	a_{53}	1	$\frac{1}{15}$		
$\frac{14}{15}$	$-\frac{14605}{25536}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$-\frac{50399}{127680}$	$\frac{1}{15}$	
1	a_{71}	a_{72}	a_{73}	a_{74}	0	$\frac{1}{15}$	$\frac{1}{15}$
	a_{71}	a_{72}	a_{73}	a_{74}	0	$\frac{1}{15}$	$\frac{1}{15}$

$\frac{1}{7}$	$\frac{1}{7}$						
$\frac{2}{7}$	$\frac{1}{7}$	$\frac{1}{7}$					
$\frac{11}{25}$	$-\frac{3229}{8750}$	1	$-\frac{1671}{8750}$				
$\frac{4}{5}$	-1	$\hat{a}_{41}$	$\hat{a}_{42}$	$\hat{a}_{43}$			
$\frac{13}{15}$	1	$\hat{a}_{51}$	$\hat{a}_{52}$	$\hat{a}_{53}$	$\hat{a}_{54}$		
$\frac{14}{15}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\hat{a}_{63}$	$\hat{a}_{64}$	$\hat{a}_{65}$	
1	$\hat{a}_{70}$	$\hat{a}_{71}$	$\hat{a}_{72}$	$\hat{a}_{73}$	$\hat{a}_{74}$	$\hat{a}_{75}$	$\hat{a}_{76}$
	$\hat{a}_{70}$	$\hat{a}_{71}$	$\hat{a}_{72}$	$\hat{a}_{73}$	$\hat{a}_{74}$	$\hat{a}_{75}$	$\hat{a}_{76}$

Table 3.7: Butcher tableau for fourth order IMEX-RK

where the coefficients

$\hat{a}_{21} = 0.112.897320084$ $\hat{a}_{22} = 0.2212436013249$

$\hat{a}_{31} = 0.0875607487880$ $\hat{a}_{32} = 0.3795156599067$ $\hat{a}_{33} = 0.0329235913053$

$\hat{a}_{41} = -0.0166666666667$ $\hat{a}_{42} = 4.5354817628720$

$\hat{a}_{43} = -7.3386840918385$ $\hat{a}_{44} = 3.4865356622999$

$\hat{a}_{51} = 15.0076923076923$ $\hat{a}_{52} = -151.6742055202485$

$\hat{a}_{53} = 266.9663261813719$ $\hat{a}_{54} = -137.6176704347547$ $\hat{a}_{55} = 8.1178574659391$

$\hat{a}_{61} = -1.0372960372960$ $\hat{a}_{62} = 4.4428904428904$ $\hat{a}_{63} = -5.9195804195804$

$\hat{a}_{64} = 4.0093240093240$ $\hat{a}_{65} = -0.9324009324009$ $\hat{a}_{66} = 4.370629370629$

Results are shown in Tab.3.10 for the two-stage second-order scheme described in Tab.3.3 at $t^n = 0.08$, and in Tables 3.12 and 3.13 for the four-stage third-order scheme Tab.3.4 at $t^n = 0.08$ and six-stage fourth-order scheme Tab.3.9 at $t^n - = 0.008$, respectively. To verify the long-time convergence, we test with the second-order schemes at $t^n = 10$ and the results are shown in Table 3.11. Numerical experiments confirm the proposed theorem.

a_{41}	$24058589213/79084733000$
a_{42}	$299597908/760430125$
a_{43}	$-23336559/90382552$
a_{51}	$-419174425120649204186119/79218920548212919110000$
a_{52}	$477902525326525491346619/41132901053879784922500$
a_{53}	$-558440160387285872007719/85556434192069952638800$
a_{71}	$4922197/9687600$
a_{72}	$-2268847/3936600$
a_{73}	$3238175/4094064$
a_{73}	$869063/6036120$
$\hat{a}_{41}$	$55740853326839983334621/27520768939834456977600$
$\hat{a}_{42}$	$-11060173572755944147541/1428963002645250650760 0$
$\hat{a}_{43}$	$232934489391190945570 7/4246061493574459076544$
$\hat{a}_{51}$	$-71186087675929044223656193931944384573278473/34385108850831845244624041532001641729750 00$
$\hat{a}_{52}$	$14268306999047363197901818351205461479990553/39921554948648535090509517879963683422500 0$
$\hat{a}_{53}$	$-16419684395723702322742309390143828209405 3/11625156801046453418356371606645424612632$
$\hat{a}_{54}$	$-27778778074096247550359633 89/26523372117018237983981550 00$
$\hat{a}_{63}$	$-14789107485345372789 59/506708836650241118208$
$\hat{a}_{64}$	$29121578321788444001/11310465103800024960$
$\hat{a}_{65}$	$-10436070861733408099/18766993950008930304$
$\hat{a}_{70}$	$-124322508365216540633 95/85826994234468729923442$
$\hat{a}_{71}$	$819449656406591531728406 3497/92384176593982140889592968 80$
$\hat{a}_{72}$	$-29969058445578885999181 37227/37540727278156622468513530 80$
$\hat{a}_{73}$	$139201741583094829137674 68375/19521178184641443683627036016$
$\hat{a}_{74}$	$138984614055560715629066 1143/57562448493173487785054080 56$
$\hat{a}_{75}$	$84635598876295853750/1430449903907812165390 7$
$\hat{a}_{76}$	$136909318359415531976 1/14304499039078121653907$

Table 3.8: related coefficients in Table 3.7

3.7 Conclusion and Comments

In this chapter, we discuss the IMEX-RK schemes on linear and semilinear equations. We divide the operator into implicit part and explicit part, which can benifit us for both linear or semilianer cases. For linear problem, like Stokes-Darcy equations, the division can decouple the system. By spectral decomposition, we prove that the schemes have long time stability and give its convergence. In semilinear case, IMEX-RK schemes avoid solving nonlinear systems. Combining existence analysis, we can build a framework to build such schemes that hold both maximum bound preserving and original energy decay up to third order.

$1/6$	$1/6$					
$1/3$	$1/6$	$1/6$				
$1/2$	$3/10$	$1/30$	$1/6$			
$2/3$	$47/30$	$-12/5$	$4/3$	$1/6$		
$4/5$	$1/2$	-1	$4/3$	$-1/6$	$2/15$	
1	$4/5$	$-3/2$	2	$-1/2$	0	$1/5$
	$4/5$	$-3/2$	2	$-1/2$	0	$1/5$

$1/6$	$1/6$					
$1/3$	$\hat{a}_{21}$	$\hat{a}_{22}$				
$1/2$	$\hat{a}_{31}$	$\hat{a}_{32}$	$\hat{a}_{33}$			
$2/3$	$\hat{a}_{41}$	$\hat{a}_{42}$	$\hat{a}_{43}$	$\hat{a}_{44}$		
$4/5$	$\hat{a}_{51}$	$\hat{a}_{52}$	$\hat{a}_{53}$	$\hat{a}_{54}$	$\hat{a}_{55}$	
1	$\hat{a}_{61}$	$\hat{a}_{62}$	$\hat{a}_{63}$	$\hat{a}_{64}$	$\hat{a}_{65}$	$\hat{a}_{66}$
	$\hat{a}_{61}$	$\hat{a}_{62}$	$\hat{a}_{63}$	$\hat{a}_{64}$	$\hat{a}_{65}$	$\hat{a}_{66}$

Table 3.9: Butcher tableau for six-stage fourth-order IMEX-RK

Table 3.10: Errors and convergence rates at $t^n = 0.08$ with $h = 1/100$ for two-stage second-order scheme described in Tab.3.3.

τ	$\|\phi^\tau - \phi^{\tau/2}\|$	order	$\|\mathbf{u}^\tau - \mathbf{u}^{\tau/2}\|$	order
$1/800$	7.51e-07	-	6.35e-05	-
$1/1600$	2.05e-07	1.87	1.74e-05	1.87
$1/3200$	5.38e-08	1.93	4.57e-06	1.93
$1/6400$	1.38e-08	1.96	1.17e-06	1.96

Table 3.11: Errors and convergence rates at $t^n = 10.0$ with $h = 1/100$ for two-stage second-order scheme described in Tab.3.3.

τ	$\|\phi^\tau - \phi^{\tau/2}\|$	order	$\|\mathbf{u}^\tau - \mathbf{u}^{\tau/2}\|$	order
$1/400$	5.23e-08	-	1.99e-08	-
$1/800$	1.42e-08	1.89	5.85e-09	1.77
$1/1600$	3.70e-09	1.94	1.60e-09	1.87
$1/3200$	9.45e-10	1.97	4.20e-10	1.93

Table 3.12: Errors and convergence rates at $t^n = 0.08$ with $h = 1/100$ for four-stage third-order scheme described in Tab.3.4.

τ	$\|\phi^\tau - \phi^{\tau/2}\|$	order	$\|\mathbf{u}^\tau - \mathbf{u}^{\tau/2}\|$	order
$1/100$	7.19e-08	-	6.27e-06	-
$1/200$	1.01e-08	2.84	8.78e-07	2.84
$1/400$	1.33e-09	2.92	1.16e-07	2.91
$1/800$	1.72e-10	2.96	1.50e-08	2.96

Table 3.13: Errors and convergence rates at $t^n = 0.008$ with $h = 1/100$ for six-stage fourth-order scheme described in Tab.3.9

τ	$\|\phi^\tau - \phi^{\tau/2}\|$	order	$\|\mathbf{u}^\tau - \mathbf{u}^{\tau/2}\|$	order
$1/250$	1.33e-07	-	1.21e-04	-
$1/500$	5.58e-09	4.58	1.96e-06	5.95
$1/1000$	3.40e-10	4.04	1.15e-07	4.09
$1/2000$	2.12e-11	4.00	7.17e-09	4.01

Chapter 4

Robust Convergence of Parareal Algorithms with Arbitrarily High-order Fine Propagators

The main focus of this part is to study the convergence of a class of parareal solver for the parabolic problems. Specifically, we let $T > 0, u^0 \in H$, and consider the initial value problem of seeking $u \in C((0,T]; D(A)) \cap C([0,T]; H)$ satisfying

$$\begin{cases} u'(t) + Au(t) = f(t), \quad 0 < t < T, \\ u(0) = u^0, \end{cases} \tag{4.1}$$

where A is a positive definite, selfadjoint, linear operator with a compact inverse, defined in Hilbert space $(H, (\cdot,\cdot))$ with domain $D(A)$ dense in H. Here $u^0 \in H$ is a given initial condition and $f : [0,T] \to H$ is a given forcing term. Throughout this part, $\|\cdot\|$ denotes the norm of the space H.

The parareal algorithm is defined by using two time propagators, $\mathcal{G}$ and $\mathcal{F}$, associated with the large step size ΔT and the small step size Δt respectively, where we assume that the ratio $J = \Delta T/\Delta t$ is an integer greater than 1. The fine time propagator $\mathcal{F}$ is operated with small step size Δt in each coarse sub-interval parallelly, after which the coarse time propagator $\mathcal{G}$ is operated with large step size ΔT se-

quentially for corrections. In general, the coarse propagator $\mathcal{G}$ is assumed to be much cheaper than the fine propagator $\mathcal{F}$. Therefore, throughout this part, we fix $\mathcal{G}$ to the backward-Euler method and study the choices of $\mathcal{F}$. Then a natural question arises related to convergence of the parareal algorithm. For parabolic type problems, in the pioneer work [7], Bal proved a fast convergence of the parareal method with a strongly stable coarse propagator and the exact fine propagator, provided some regularity assumptions on the problem data. The analysis works for both linear and nonlinear problems. This convergence behavior is clearly observed in numerical experiments, see e.g. Figure 4.2. However, without those regularity assumptions, the convergence observed from the empirical experiments will be much slower than expected, cf. Figure 4.3. See also some rigorous analysis in [27, 105, 37].

This interesting phenomenon motivates the current work, where we aim to study the convergence of parareal algorithm which is expected to be robust in the case of nonsmooth / incompatible problem data, that is related to various applications, e.g., optimal control, inverse problems, and stochastic models. There have existed some case studies. In [82], Mathew, Sarkis and Schaerer considered the backward Euler method as the fine propagator and proved the robust convergence of the parareal algorithm with a convergence factor 0.298 (for all $J \geq 2$); see also [41, 106] for some related discussion. In [115], Wu showed that the convergence factors for the second-order diagonal implicit Runge–Kutta method and a single step TR/BDF2 method (i.e., the ode23tb solver for ODEs in MATLAB) are 0.316 (with $J \geq 2$) and 0.333 (with $J \geq 2$), respectively. These error bounds might be slightly improved by increasing J_*. See also [116] for the analysis for a third-order diagonal implicit Runge–Kutta method with a convergence factor 0.333 ($J_* = 4$). For fourth-order Gauss–Runge–Kutta integrator, in [116] Wu and Zhou showed that the threshold depends on both the largest eigenvalue of operator A and the step size Δt. Note that the eigenvalues of A may approach infinity, e.g. $A = -\Delta$ with homogeneous boundary conditions. Therefore, this kind of integrators might not be suitable for the parareal algorithm.

Then a natural question arises: *in what case there exists a threshold $J_* > 0$ (independent of step sizes ΔT, Δt, terminal time T, problem data u^0 and f, as well as the distribution of spectrum of the elliptic operator A), such that for any $J \geq J_*$, the parareal algorithm for solving the parabolic equation* (4.1) *converges robustly*? Our study provides a positive answer to this question: if the fine propagator is strongly stable, in sense that the stability function satisfies $|r(-\infty)| \in [0, 1)$, then there must exist such a positive threshold J_* so that for all $J \geq J_*$ the parareal algorithm converges linearly with convergence factor close to 0.3. The convergence is robust even if the initial data is nonsmooth or incompatible with

boundary conditions. Noting that all L-stable Runge–Kutta schemes satisfy that condition, so the fine propagator can be arbitrarily high-order. As examples, we analyzed three popular L-stable schemes, i.e., two-, three-, four-stage Lobatto IIIC schemes. We show that for all these cases the parareal algorithm converges linearly with factor less than 0.31 and $J_* = 2$. Our theoretical results are fully supported by numerical experiments.

The rest of the chapter is organized as follows. In Section 4.1, we introduce singe step integrators and parareal algorithms for solving the parabolic problem. Then we show the convergence of the algorithm in Section 4.2 by using the spectrum decomposition. Moreover, in Section 4.3, we present case studies on three popular L-stable Runge–Kutta schemes, and show a sharper estimate for the threshold J_*. Finally, in Section 4.4, we present some numerical results to illustrate and complement the theoretical analysis.

4.1 Single-Step Methods and Parareal Algorithm

In this section, we present the basic setting of the single step time stepping methods for solving the parabolic equation (4.1) and the parareal algorithm. See more detailed discussion in the monograph [110, Chapter 7-9] and the comprehensive survey paper [40].

4.1.1 Single-Step Integrators for Solving Parabolic Equations

To begin with, we consider the time discretization for the parabolic equation (4.1). We split the interval $(0, T)$ into N subintervals with the uniform mesh size $\Delta t = T/N$, and set $t^n = n\Delta t$, $n = 0, 1, \ldots, N$. Then a framework of a single step scheme approximating $u(t^n)$ reads:

$$u^{n+1} = r(-\Delta t A)u^n + \Delta t \sum_{i=1}^{m} p_i(-\Delta t A) f(t^n + c_i \Delta t), \quad \text{for all } 0 \le n \le N-1, \tag{4.2}$$

Here, $r(\lambda)$ and $\{p_i(\lambda)\}_{i=1}^m$ are rational functions and c_i are distinct real numbers in $[0, 1]$. Throughout this thesis, we assume that the scheme (4.2) satisfies the following assumptions.

(P1) $|r(-\lambda)| < 1$ and $|p_i(-\lambda)| \le c$, for all $i = 1, \ldots, m$, uniformly in Δt and $\lambda > 0$. Besides, the numerator of $p_i(\lambda)$ is of lower degree than its denominator.

(P2) The time stepping scheme (4.2) is *accurate* of order q in sense that

$$r(-\lambda) = e^{-\lambda} + O(\lambda^{q+1}), \quad \text{as } \lambda \to 0.$$

and for $0 \le j \le q$

$$\sum_{i=1}^{m} c_i^j p_i(-\lambda) - \frac{j!}{(-\lambda)^{j+1}}\Big(e^{-\lambda} - \sum_{\ell=0}^{j} \frac{(-\lambda)^\ell}{\ell!}\Big) = O(\lambda^{q-j}), \quad \text{as } \lambda \to 0.$$

(P3) The rational function $r(\lambda)$ is *strongly stable* in sense that $|r(-\infty)| < 1$.

Remark 4.1.1. *Condition (P3) is essential for the convergence of parareal iteration. If $|r(\infty)| = 1$, e.g., Crank-Nicolson method and implicit Runge-Kutta methods of Gauss type, the parareal method converges only if the eigenvalues of A is bounded from above (which is not true for parabolic equations) and the ratio between the coarse step size and the fine step size is sufficiently large (depending on the upper bound of eigenvalues of A). Besides, this condition is also important in case that problem data is nonsmooth, e.g., $u^0 \in H$. Time stepping schemes violating this condition may lose the optimal convergence rate in the nonsmooth data case [110, Chapter 8].*

Practically, it is convenient to choose $p_i(\lambda)$ that share the same denominator of $r(\lambda)$:

$$r(\lambda) = \frac{a_0(\lambda)}{g(\lambda)}, \quad \text{and} \quad p_i(\lambda) = \frac{a_i(\lambda)}{g(\lambda)}, \quad \text{for } i = 1, 2, \ldots, m,$$

where $a_i(\lambda)$ and $g(\lambda)$ are polynomials. Then the integrator (4.2) could be written as

$$g(-\Delta t A)u^{n+1} = a_0(-\Delta t A)u^n + \Delta t \sum_{i=1}^{m} a_i(-\Delta t A) f(t^n + c_i \Delta t), \quad \text{for all } 1 \le n \le N.$$

See e.g. [110, pp. 131] for the construction of such rational functions satisfying (P1)-(P3).

Under those conditions, there holds the following error estimate for the time stepping scheme (4.2). The proof is given in [110, Theorems 7.2 and 8.3].

Lemma 4.1.1. *Suppose that the Conditions (P1)-(P3) are fulfilled. Let $u(t)$ be the solution to parabolic*

equation (4.1)*, and* u^n *be the solution to the time stepping scheme* (4.2)*. Then there holds*

$$\|u^n - u(t^n)\| \le c\,(\Delta t)^q \Big((t^n)^{-q}\|u^0\| + t^n \sum_{\ell=0}^{q-1} \sup_{s\le t^n} \|A^{q-\ell} f^{(\ell)}(s)\| + \int_0^{t^n} \|f^{(q)}(s)\|\,\mathrm{d}s \Big),$$

provided that $v \in H$*,* $f^{(\ell)} \in C([0,T]; Dom(A^{q-\ell})$ *with* $0 \le \ell \le q-1$ *and* $f^{(q)} \in L^1(0,T;H)$ *for all* $\ell < q$*.*

Remark 4.1.2. *Lemma 4.1.1 indicates that, under Conditions (P1)-(P3), the solution of the time stepping scheme* (4.2) *converges to the exact solution with order* q *provided that the source term* f *and initial condition* u_0 *satisfy certain compatibility conditions. For example, if we consider the parabolic equation where* $A = -\Delta$ *with homogeneous Dirichlet boundary condition, it requires* $(-\Delta)^\ell f^{(q-\ell)} = 0$ *on the boundary* $\partial\Omega$ *for* $0 \le \ell \le q$*. In order to avoid the restrictive compatibility conditions, we shall assume that the time discretization scheme* (4.2) *is strictly accurate of order* q *in sense that*

$$\sum_{i=1}^m c_i^j p_i(-\lambda) - \frac{j!}{(-\lambda)^{j+1}} \Big(r(-\lambda) - \sum_{\ell=0}^{j} \frac{(-\lambda)^\ell}{\ell!} \Big) = 0, \quad \textit{for all } 0 \le j \le q-1.$$

It is well-known that a single step method with a given $m \in \mathbb{Z}^+$ *could be accurate of order* $2m$ *(Gauss–Legendre method) [31, Section 2.2], but at most strictly accurate of order* $m+1$ *[9, Lemma 5].*

Remark 4.1.3. *The error estimate in Lemma 4.1.1 could be slightly improved if the time integrator is L-stable, i.e.* $r(-\infty) = 0$*; see e.g., [110, Theorem 7.2].*

4.1.2 Parareal Algorithm

Next, we state the parareal solver for the single step scheme (4.2). Let $\Delta T = J\Delta t$, with a positive integer $J \ge 2$, be the coarse step size. Without loss of generality, we assume that $N_c = T/\Delta T$ is an integer, and let $T^n = n\Delta T$ Then, two numerical propagators $\mathcal{G}$ and $\mathcal{F}$ are assigned to the coarse and fine time grids, where $\mathcal{G}$ is usually a low-order and inexpensive numerical method (such as backward Euler scheme), and $\mathcal{F}$ is given by the single step integrator (4.2). Specifically, for $v \in H$ and $f \in C([0,T];H)$, letting I denote the identity operator, we define the coarse and finer propagator as

$$\mathcal{G}(T^n, \Delta T, v, f) = (I + \Delta T A)^{-1}(v + \Delta T f(T^n)).$$

and

$$\mathcal{F}(t^n, \Delta t, v, f) = r(-\Delta t A)v + \Delta t \sum_{i=1}^{m} p_i(-\Delta t A) f(t^n + c_i \Delta t).$$

respectively. Then, the parareal solver is described in Algorithm 2.

Algorithm 2: Parareal solver for the single step scheme (4.2)

Input : u^0, $\mathcal{F}$, $\mathcal{G}$, K, N_c, J
Output: U_K^n

$U_0^0 = u^0$;
for $n = 0, 1, ..., N_c - 1$ **do**
  $U_0^{n+1} = \mathcal{G}(T^n, \Delta T, U_0^n, f)$;
end
for $k = 0, 1, \ldots, K - 1$ **do**
  for $n = 0, 1, ..., N_c - 1$ **do** parallel
    for $j = 0, 1, 2, \ldots, J - 1$ **do**
      $\widetilde{U}^{n,j+1} = \mathcal{F}(T^n + j\Delta t, \Delta t, \widetilde{U}^{n,j}, f)$ with initial value $\widetilde{U}^{n,0} = U_k^n$;
    end
    $\widetilde{U}^{n+1} = \widetilde{U}^{n,J}$;
  end
  for $n = 0, 1, ..., N_c - 1$ **do**
    $U_{k+1}^{n+1} = \mathcal{G}(T^n, \Delta T, U_{k+1}^n, f) + \widetilde{U}^{n+1} - \mathcal{G}(T^n, \Delta T, U_k^n, f)$
  end
end

The aim of this chapter is to show that the iterative solution U_k^n, generated by the parareal algorithm, linearly converges to the exact time stepping solution u^{nJ} of the single step integrator (4.2) with fine time step Δt, i.e.,

$$\max_{1 \le n \le N} \|U_k^n - u^{nJ}\| \le c\, \gamma^k, \tag{4.3}$$

with some convergence factor γ strictly smaller than 1. We shall prove that there exists a positive threshold J_*, independent of ΔT, Δt and the upper bound of spectrum of A, such that if $J \ge J_*$, then (4.3) is true with $\gamma \approx 0.3$, under conditions (P1)-(P3).

4.2 Convergence Analysis

Next, we briefly test the convergence factor of parareal iteration. Taking comparision with the exact time stepping solution in (4.2), we arrive at

$$\begin{aligned}
U_{k+1}^{n+1} - u^{(n+1)J} &= (I + \Delta T A)^{-1}\Big[(U_{k+1}^{n} - u^{nJ}) - (U_{k}^{n} - u^{nJ})\Big] \\
&\quad + F(T^n + (J-1)\Delta t, \Delta t, \widetilde{U}^{n,J-1}, f) - F(T^n + (J-1)\Delta t, \Delta t, u^{nJ-1}, f)) \\
&= (I + \Delta T A)^{-1}\Big[(U_{k+1}^{n} - u^{nJ}) - (U_{k}^{n} - u^{nJ})\Big] + r(-\Delta t A)(\widetilde{U}^{n,J-1} - u^{(n+1)J-1}) \\
&= \cdots \\
&= (I + \Delta T A)^{-1}\Big[(U_{k+1}^{n} - u^{nJ}) - (U_{k}^{n} - u^{nJ})\Big] + r(-\Delta t A)^J(\widetilde{U}^{n,0} - u^{nJ}) \\
&= (I + \Delta T A)^{-1}\Big[(U_{k+1}^{n} - u^{nJ}) - (U_{k}^{n} - u^{nJ})\Big] + r(-\Delta t A)^J(U_{k}^{n} - u^{nJ}),
\end{aligned}$$

For the sake of simplicity, we define $E_k^n = U_k^n - u^{nJ}$ and rewrite the above equation as

$$E_{k+1}^{n+1} = (I + \Delta T A)^{-1}(E_{k+1}^{n} - E_{k}^{n}) + r(-\Delta t A)^J E_k^n$$

Recall that the operator A is a positive definite, selfadjoint, linear operator with a compact inverse, defined in Hilbert space $(H, (\cdot, \cdot))$. Then by the spectral theory, A has positive eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$, where $0 < \lambda_1 \leq \lambda_2 \leq \ldots$ and $\lambda_j \to \infty$, and the corresponding eigenfunctions $\{\phi_j\}_{j=1}^{\infty}$ form an orthonormal basis of the Hilbert space H. Then, letting $e_{k,j}^n = (E_k^n, \phi_j)$, by means of spectrum decomposition, we derive

$$e_{k+1,j}^{n+1} = \frac{e_{k+1,j}^{n} - e_{k,j}^{n}}{1 + \Delta T \lambda_j} + r(-\Delta t \lambda_j)^J e_{k,j}^n.$$

By letting $d_j = \Delta T \lambda_j$, we have

$$e_{k+1,j}^{n+1} = (1 + d_j)^{-1} e_{k+1,j}^{n} + (r(-d_j/J)^J - (1 + d_j)^{-1}) e_{k,j}^n.$$

We apply the recursion and use the fact that $e^0_{k+1,j} = 0$, and hence obtain

$$\begin{aligned} e^{n+1}_{k+1,j} &= (r(-d_j/J)^J - (1+d_j)^{-1})e^n_{k,j} + (1+d_j)^{-1}e^n_{k+1,j} \\ &= (r(-d_j/J)^J - (1+d_j)^{-1})\Big(e^n_{k,j} + (1+d_j)^{-1}e^{n-1}_{k,j}\Big) + (1+d_j)^{-2}e^{n-1}_{k+1,j} \\ &= \dots \\ &= (r(-d_j/J)^J - (1+d_j)^{-1})\Big(e^n_{k,j} + (1+d_j)^{-1}e^{n-1}_{k,j} + \cdots + (1+d_j)^{-(n-1)}e^1_{k,j}\Big). \end{aligned}$$

Now taking the absolute value on the both sides yields

$$\begin{aligned} |e^{n+1}_{k+1,j}| \le& |r(-d_j/J)^J - (1+d_j)^{-1}| \cdot (1 + (1+d_j)^{-1} + \cdots + (1+d_j)^{-(n-1)}) \max_{1\le n\le N} |e^n_{k,j}| \\ \le& \frac{|r(-d_j/J)^J - (1+d_j)^{-1}|}{1-(1+d_j)^{-1}} \max_{1\le n\le N} |e^n_{k,j}| \\ =& \left|\frac{(1+d_j)r(-d_j/J)^J - 1}{d_j}\right| \max_{1\le n\le N} |e^n_{k,j}| \\ \le& \sup_{s\in(0,\infty)} \left|\frac{(1+s)r(-s/J)^J - 1}{s}\right| \max_{1\le n\le N} |e^n_{k,j}|. \end{aligned} \tag{4.4}$$

If the the leading factor is strictly smaller than one, i.e.,

$$\sup_{s\in(0,\infty)} \left|\frac{(1+s)r(-s/J)^J - 1}{s}\right| \le \gamma < 1, \tag{4.5}$$

then e^n_k converges to zero linearly with a factor (smaller than) γ, and hence the parareal iteration converges linearly to the time stepping solution (4.2) in sense of (4.3).

Our convergence analysis in this and next sections heavily depend on the constant κ_α defined by

$$\kappa_\alpha := \sup_{s\in(0,\infty)} \left|\frac{(1+s)e^{-\alpha s} - 1}{s}\right|, \qquad \text{for } \alpha \in [0,2]. \tag{4.6}$$

To begin with, we establish a simple upper bound for the constant κ_α.

Lemma 4.2.1. *Let $\alpha \in [0,2]$, and κ_α be the constant defined in* (4.6)*. Then there holds*

$$\kappa_\alpha \le \begin{cases} e^{\alpha-2}, & \alpha \in [1,2]; \\ \max(e^{\alpha-2}, 1-\alpha), & \alpha \in [0,1). \end{cases}$$

Proof. First of all, we show the claim that

$$\frac{(1+s)e^{-\alpha s}-1}{s} \geq -e^{\alpha-2}. \tag{4.7}$$

To this end, we define the auxiliary function

$$g(s) = (1+s)e^{-\alpha s} + e^{\alpha-2}s.$$

Then a simple computation yields

$$g'(s) = (1-\alpha-\alpha s)e^{-\alpha s} + e^{\alpha-2} \quad \text{and} \quad g''(s) = (\alpha^2+\alpha^2 s-2\alpha)e^{-\alpha s}.$$

It is easy to observe that $g''(s)$ admits a single root at $s=(2-\alpha)/\alpha$, and

$$g'(x) \geq g'((2-\alpha)/\alpha) = 0.$$

Therefore $g(s)$ is increasing in $[0,\infty)$. As a result, $g(s) \geq g(0) = 1$, and hence

$$(1+s)e^{-\alpha s} - 1 \geq -e^{\alpha-2}s \qquad \forall\ s \geq 0,$$

which implies (4.7). Moreover, for $\alpha \geq 1$, we observe that

$$1+s \leq e^s \leq e^{\alpha s} \qquad \forall s \geq 0,$$

which immediately leads to $(1+s)e^{-\alpha s} - 1 \leq 0$ for all $s \geq 0$. This completes the proof for the desired results in case that $\alpha \in [1,2]$.

Now we turn to the case that $\alpha \in [0,1)$. Let $\kappa_\alpha^* = \max\{e^{\alpha-2}, 1-\alpha\}$ and define

$$g(x) = (1+s)e^{-\alpha s} - \kappa_\alpha^* s.$$

Then the simple computation yields

$$g'(s) = (1-\alpha-\alpha s)e^{-\alpha s} - \kappa_\alpha^* \quad \text{and} \quad g''(s) = (\alpha^2 + \alpha^2 s - 2\alpha)e^{-\alpha s}.$$

Noting that $g'(0) = 1-\alpha-\kappa_\alpha^* \leq 0$, $g'(\infty) = -\kappa_\alpha^* < 0$ and $g'((2-\alpha)/\alpha) \leq 0$. These imply $g'(s) \leq 0$ and hence g is decreasing function in $[0,\infty)$. Therefore $g(s) \leq g(0) = 1$, which further implies

$$(1+s)e^{-\alpha s} - \kappa_\alpha^* s \leq 1.$$

This leads to the desired assertion for the case that $\alpha \in [0,1)$. □

Lemma 4.2.1 only provides a rough upper bound for κ_α. In fact, for a fixed α we can further improve the upper bound via a more careful computation. In Figure 4.1, we numerically compute the constant κ_α for $\alpha \in [0,2]$ and plot those values.

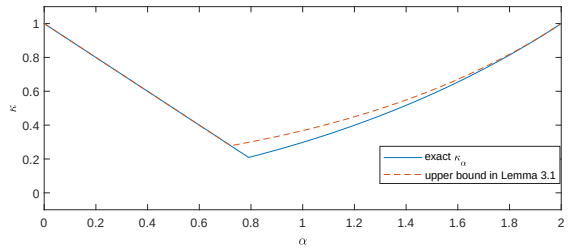


Figure 4.1: Plot of κ_α defined in lemma 4.2.1.

The next lemma provides a sharper estimate for κ_1.

Lemma 4.2.2. *Let κ_α be the constant defined in* (4.6)*. Then $\kappa_1 \approx 0.2984$.*

Proof. To show the sharp estimate, we note that for $g(s) = 1-(1+s)e^{-s}$, there holds $g'(s) = se^{-s} \geq 0$ and $g(0) = 1$. Therefore $g(s) \geq 0$ for all $s \in (0,\infty)$. This further implies $\psi(s) = (1-(1+s)e^{-s})/s \geq 0$ for $s \in (0,\infty)$.

Meanwhile, we note that

$$\varphi(s) := s^2 e^s \psi'(s) = s^2 + s + 1 - e^s.$$

By checking the monotonicity, it is easy to verify that $\varphi(s)$ has a unique root, denoted by s_*, in $(0,\infty)$.

It further implies the function $\psi(s)$ achieves its maximum at s^*, i.e., $\kappa_1 = \psi(s_*)$. Finally, the fixed-point iteration $s_{k+1} = \ln(s_k^2 + s_k + 1)$ and the contraction mapping theorem for $s \in [1.5, 2]$ provide $s^* \approx 1.793$, and hence $\kappa_1 \approx 0.2984$. □

Now we state our main theorem which verifies the desired result (4.5) with $\gamma \approx 0.3$.

Theorem 4.2.1. *Let conditions (P1)-(P3) hold valid. Then there exists a threshold $J_* > 0$ such that for all $J \geq J_*$*

$$\sup_{s\in(0,\infty)} \left| \frac{(1+s)r(-s/J)^J - 1}{s} \right| \leq 0.3,$$

Proof. First of all, we aim to show that

$$\lim_{J\to\infty} \sup_{s\geq 0} \frac{(1+s)}{s} \left| r(-s/J)^J - e^{-s} \right| = 0. \tag{4.8}$$

For a given $\delta > 0$, for any $s > \delta$, it is obvious that $(1+s)/s$ is bounded by a constant $C(\delta)$. Meanwhile, note that conditions (P1)-(P3) are fullfilled. Then by means of the nonsmooth data error estimate [110, Theorem 7.2], there holds

$$\left| r(-s/J)^J - e^{-s} \right| \leq cJ^{-q},$$

where c is independent of s. Then we derive

$$\lim_{J\to\infty} \sup_{s>\delta} \frac{(1+s)}{s} \left| r(-s/J)^J - e^{-s} \right| = 0.$$

For $0 < s \leq \delta$, conditions (P1) and (P2) imply

$$\begin{aligned}
&\frac{(1+s)}{s} \left| r(-s/J)^J - e^{-s} \right| \\
&= \frac{(1+s)}{s} |r(-s/J) - e^{-s/J}| \left| \sum_{i=0}^{J-1} r(-s/J)^i e^{-(J-1-i)s/J} \right| \\
&\leq \frac{(1+s)}{s} \cdot C(\alpha,\delta) \Big(\frac{s}{J}\Big)^{q+1} \cdot J \\
&\leq C(\alpha,\delta) J^{-q}.
\end{aligned}$$

Therefore we arrive at

$$\lim_{J\to\infty} \sup_{0<s\leq\delta} \frac{(1+s)}{s} \left| r(-s/J)^J - e^{-s} \right| = 0,$$

which completes the proof of the (4.8). This together with Lemma 4.2.2 implies that

$$\lim_{J\to\infty}\sup_{s\in(0,\infty)}\left|\frac{(1+s)r(-s/J)^J-1}{s}\right|=\kappa_1<0.3,$$

which completes the proof of the lemma. □

Next, using Theorem 4.2.1, we are able to show the linear convergence of the parareal iteration 2.

Theorem 4.2.2. *Let conditions (P1)-(P3) be fullfilled and the data regularity in Lemma 4.1.1 hold valid. Let u^n be the solution to the time stepping scheme* (4.2)*, and U_k^n be the solution obtained from the parareal algorithm 2. Then there exists a threshold $J_*>0$ such that for all $J\geq J_*$, we have*

$$\max_{1\leq n\leq N_c}\|U_k^n-u^{nJ}\|\leq c\gamma^k \quad \textit{with} \quad \gamma=0.3.$$

Proof. In Algorithm 2, the initial guess U_0^n is obtained by the coarse propagator, i.e., the backward Euler scheme. Then Lemma 4.1.1 (with $q=1$) implies the estimate

$$\begin{aligned}\|U_0^n-u^{nJ}\|&\leq\|U_0^n-u(T^n)\|+\|u(T^n)-u^{nJ}\|\\&\leq c\big((\Delta T)(T^n)^{-1}+(\Delta t)(t^{nJ})^{-1}\big)\leq cn^{-1}.\end{aligned}\tag{4.9}$$

Let $E_k^n=U_k^n-u^{nJ}$ and $e_{k,j}^n=(E_k^n,\phi_j)$. The the relation (4.4) and Theorem 4.2.1 imply

$$\begin{aligned}\sup_{1\leq n\leq N_c}\|E_k^n\|^2&\leq\sum_{j=1}^{\infty}\sup_{1\leq n\leq N_c}|e_{k,j}^n|^2\leq\gamma^2\sum_{j=1}^{\infty}\sup_{1\leq n\leq N_c}|e_{k-1,j}^n|^2\\&\leq\cdots\leq\gamma^{2k}\sum_{j=1}^{\infty}\sup_{1\leq n\leq N_c}|e_{0,j}^n|^2\end{aligned}$$

with $\gamma=0.3$. This together with the estimate $\sup_{1\leq n\leq N_c}|e_{0,j}^n|^2\leq\sum_{n=1}^{N_c}|e_{0,j}^n|^2$ leads to

$$\sup_{1\leq n\leq N_c}\|E_k^n\|^2\leq c\gamma^{2k}\sum_{j=1}^{\infty}\sum_{n=1}^{N_c}|e_{0,j}^n|^2\leq c\gamma^{2k}\sum_{n=1}^{N_c}\|E_0^n\|^2\leq c\gamma^{2k}\sum_{n=1}^{N_c}n^{-2}\leq c\gamma^{2k},$$

where in the second last inequality we apply the estimate (4.9). This completes the proof of the theorem.

□

Remark 4.2.1. *Theorem 4.2.2 provides an useful upper bound of the convergence factor for all single step integrators (satisfying (P1)-(P3)), which might not be sharp for specific one. For example, in [82, Lemma 4.3], Mathew, Sarkis and Schaerer considered the backward Euler method and proved that the convergence factor of the Parareal algorithm is around* 0.298 *(with* $J_* = 2$*). In [115], Wu showed that convergence factors are* 0.316 *(with* $J_* = 2$*) and* 0.333 *(with* $J_* = 2$*) for the second-order diagonal implicit Runge–Kutta method and a single step TR/BDF2 method (i.e., the ode23tb solver for ODEs in MATLAB), respectively. These error bounds might be slightly improved by increasing* J_**. See also [116] for the analysis for a third-order diagonal implicit Runge–Kutta method with a convergence factor* 0.333 *and* $J_* = 4$*.*

Remark 4.2.2. *Theorem 4.2.2 only provides the existence of the threshold* J_* *without any upper bound estimate. It is obvious that a huge* J_* *may destroy the parallelism of the algorithm. Then a question arise naturally: is it possible to find* J_* *for a given scheme satisfying conditions (P1)-(P3)? This is the focus of Section 4.3.*

4.3 Case Studies for Several High-Order Single-Step Integrators

In this section, we shall study some popular single step methods. As we mentioned in Remark 4.2.2, Theorem 4.2.2 did not provide a sharp estimate for the threshold J_*. In fact, there is no universal estimate for all single step methods. Fortunately, for any given single step integrator satisfying conditions (P1)-(P3) and fixed convergence rate $\gamma > 0.2984$, we have a regular routine to find a sharper estimate for J_*.

We consider three time-stepping methods, namely the the two-, three-, four-stage Lobatto IIIC methods, which are respectively second-, fourth- and sixth-order accurate, to the initial and boundary value problem (4.1). For the reader's convenience, we present the Butcher tableaus of the two-, three-, four-stage Lobatto IIIC methods, respectively,

$$\begin{array}{cc|c} \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 \\ \hline \frac{1}{2} & \frac{1}{2} & \end{array} =: \begin{array}{c|c} \mathcal{A} & c \\ \hline b^\top & \end{array} \tag{4.10}$$

,

$$\begin{array}{cccc|c} \frac{1}{6} & -\frac{3}{1} & \frac{1}{6} & & 0 \\ \frac{1}{6} & \frac{5}{12} & -\frac{1}{12} & & \frac{1}{2} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & & 1 \\ \hline \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & & \end{array} =: \begin{array}{c|c} \mathcal{A} & c \\ \hline b^\top & \end{array} \tag{4.11}$$

and

$$\begin{array}{cccc|c} \frac{1}{12} & -\frac{\sqrt{5}}{12} & \frac{\sqrt{5}}{12} & -\frac{1}{12} & 0 \\ \frac{1}{12} & \frac{1}{4} & \frac{10-7\sqrt{5}}{60} & \frac{\sqrt{5}}{60} & \frac{1}{2}-\frac{\sqrt{5}}{10} \\ \frac{1}{12} & \frac{10+7\sqrt{5}}{60} & \frac{1}{4} & -\frac{\sqrt{5}}{60} & \frac{1}{2}+\frac{\sqrt{5}}{10} \\ \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12} & 1 \\ \hline \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12} & \end{array} =: \begin{array}{c|c} \mathcal{A} & c \\ \hline b^\top & \end{array}. \tag{4.12}$$

Let us also briefly recall some well-known facts about Lobatto IIIC; for details we refer to [114].

These methods can be viewed as discontinuous collocation methods. The *order* of the m-stage Lobatto IIIC methods is $q = 2m - 2$. In particular, the methods are *algebraically stable* and *L-stable*, that makes them suitable for stiff problems. The stability functions r,

$$r(z) := 1 + zb^\top (I - z\mathcal{A})^{-1}\mathbb{1} \quad \text{with} \quad \mathbb{1} := (1, \dots, 1)^\top \in \mathbb{R}^q,$$

is given by the $(m-2, m)$-Padé approximation to e^z and vanishes at infinity, i.e., $r(\infty) = 1 - b^T \mathcal{A}^{-1}\mathbb{1} = 0$. Note that the computational cost of implicit Runge-Kutta methods increases fast with the stage number, and we refer to [10, 64, 59] and the reference therein for some efficient implementations.

The following argument highly depends on the upper bound for the constant κ_α defined in (4.6). From Figure 4.1, we observe that Lemma 4.2.1 gives an sharp estimate for κ_α for $\alpha < 0.7$, while the estimate for $\alpha > 1$ could be further improved. The next lemma provides an estimate for $\alpha = 1.02$, which is useful in the analysis of convergence rate.

Lemma 4.3.1. *Let κ_α be the constant defined in* (4.6)*. Then $\kappa_{1.02} \approx 0.3078 < 0.31$.*

Proof. With $\beta = 1.02 \geq 1$ and $\psi(s) = \frac{(1+s)e^{-\beta s}-1}{s}$, we observe that $\psi(0+) = 1 - \beta$ and $\psi(\infty) = 0$. Meanwhile, since $e^{-\beta s} \leq e^{-s} \leq (1+s)^{-1}$ for $s \geq 0$, we derive that $\psi(s) \leq 0$. Now we intend to show

that $\psi'(s)$ admits a unique root in $(0,\infty)$, denoted as x_*. Then $\kappa_\beta = \psi(x_*)$. Noting that

$$\psi'(s) = \frac{1-(1+\beta s+\beta s^2)e^{-\beta s}}{s^2},$$

it suffices to show that $g(s) = 1-(1+\beta s+\beta s^2)e^{-\beta s}$ has a unique root in $(0,\infty)$. It is straightforward to see that the function

$$g'(s) = (-2+\beta+\beta s)\beta s e^{-\beta s}$$

admits a unique root in $(0,\infty)$. Then by the fact that $g(0)=0$ and Rolle's theorem, we conclude that g has at most one root in $(0,\infty)$. Meanwhile, we observe $\psi'(1) = 1-(1+2\beta)e^{-\beta} \approx -0.0962 < 0$ and $\psi'(2) = \frac{1-(1+6\beta)e^{-2\beta}}{4} \approx 0.01855 > 0$. Therefore, there exists a unique root of ψ' in $(0,\infty)$, named as x_*, which lies in $(1,2)$. Then the fixed-point iteration $s_{k+1} = \ln(\beta s_k^2+\beta s_k+1)/\beta$ and the contraction mapping theorem for $s\in[1.5,2]$ provide $s_* \approx 1.715$, and hence $\kappa_\beta = f(x_*) \approx 0.3078 \le 0.31$. □

Proposition 4.3.1. *Let u^n be the solution to the time stepping scheme* (4.2) *using the two-stage Lobatto IIIC method* (4.10)*, and U_k^n be the solution obtained from the parareal algorithm 2. Then for all $J\ge 2$, there holds*

$$\max_{1\le n\le N_c} \|U_k^n - u^n\| \le c\gamma^k \quad \textit{with} \quad \gamma = 0.31.$$

Proof. It suffices to show that for any $J\ge 2$, there holds

$$\sup_{s\in(0,\infty)} \left|\frac{(1+s)r(-s/J)^J - 1}{s}\right| \le 0.31, \quad \text{where } r(-s) = \frac{2}{s^2+2s+2}. \tag{4.13}$$

To this end, we define $\alpha = 0.69$ and $\beta = 1.02$. Then Lemma 4.2.1 implies that $\kappa_\alpha \le 1-0.69 = 0.31$, and meanwhile Lemma 4.3.1 indicates $\kappa_\beta \le 0.31$.

Next, we aim to show that $e^{-\beta s} \le r(-s) \le e^{-\alpha s}$ for all $s\in(0,s_*)$ with $s_* = 3.2$. First of all, using the fact that $\frac{2}{s^2+2s+2} \ge e^{-s} \ge e^{-\beta s}$ for all $s>0$, we derive the first inequality $e^{-\beta s} \le r(-s)$. Then we turn to the second inequality $r(-s) \le e^{-\alpha s}$, equivalent to $g(s) := 2e^{\alpha s} - (s^2+2s+2) \le 0$ in $(0,s_*)$. Noting that $g''(s) = 2\alpha^2 e^{\alpha s} - 2$, which admits a unique root at $-\frac{2\ln\alpha}{\alpha}$. Meanwhile, we observe that $g''(s) < 0$ in $(0,-\frac{2\ln\alpha}{\alpha})$, and $g''(s) > 0$ in $(-\frac{2\ln\alpha}{\alpha},\infty)$. Besides, since $g'\left(-\frac{2\ln\alpha}{\alpha}\right) = \frac{4\ln\alpha+2-2\alpha}{\alpha} < 0$ and $g'(0) = 2\alpha - 2 < 0$, we conclude that $g'(s) < 0$ in $(0,-\frac{2\ln\alpha}{\alpha})$, and $g'(s)$ has a unique root in $(-\frac{2\ln\alpha}{\alpha},\infty)$. Moreover, the facts $g'(2) \approx -0.5 < 0$ and $g'(3) \approx 2.9 > 0$ implies that there exists a

constant $s_1 \in (2,3)$, s.t. $g'(s_1) = 0$, and $g'(s) < 0$ in $(0, s_1)$, $g'(s) > 0$ in (s_1, ∞). Then we note that $g(s_*) = -0.445 < 0$ and conclude that $g(s) := 2e^{\alpha s} - (s^2 + 2s + 2) \leq 0$ in $(0, s_*)$, which implies $r(s) \leq e^{-\alpha s}$ in $(0, s_*)$. As a result, we arrive at $e^{-\beta s} \leq r(-s/J)^J \leq e^{-\alpha s}$ for all $s \in (0, Js_*)$ which implies

$$\sup_{s\in(0,Js_*)} \left| \frac{(1+s)r(-s/J)^J - 1}{s} \right| \leq 0.31.$$

Besides, we observe the fact that

$$\sup_{(s_*,\infty)} \left| \frac{2}{s^2+2s+2} \right| = \left| \frac{2}{s_*^2+2s_*+2} \right| \approx 0.1073 < 0.11.$$

Then we derive for $s \in (Js_*, \infty)$ and $J \geq 2$

$$\begin{aligned} \left| \frac{(1+s)r(-s/J)^J - 1}{s} \right| &\leq \frac{1+s}{s} |r(-s/J)^J| + s^{-1} \leq \frac{1+Js_*}{Js_*} (0.11)^J + (Js_*)^{-1} \\ &\leq \frac{1+6.4}{6.4} (0.11)^2 + (2 \times 3.2)^{-1} \approx 0.1702 \leq 0.31. \end{aligned}$$

This completes the proof of (4.13). □

The argument could be further extended to the high-order time stepping scheme, e.g., 3- or 4-stage Lobatto IIIC method. This result is given in the following proposition.

Proposition 4.3.2. *Let u^n be the solution to the time stepping scheme* (4.2) *using the three-stage Lobatto IIIC method* (4.11) *or the four-stage Lobatto IIIC method* (4.12)*, and U_k^n be the solution obtained from the parareal algorithm 2. Then for all $J \geq 2$, there holds*

$$\max_{1\leq n\leq N_c} \|U_k^n - u^n\| \leq c\gamma^k \quad \textit{with} \quad \gamma = 0.31.$$

Proof for the 3-stage Lobatto IIIC scheme (4.11):

Similar to the proof of proposition 4.3.1, we aim to show that for any $J \geq 2$

$$\sup_{s\in(0,\infty)} \left| \frac{(1+s)r(-s/J)^J - 1}{s} \right| \leq 0.31, \quad \text{where } r(-s) = \frac{24-6s}{s^3+6s^2+18s+24}. \tag{4.14}$$

Letting $\alpha = 0.69$ and $\beta = 1.02$, Lemmas 4.2.1 and 4.3.1 implies that $\kappa_\alpha \leq 1 - 0.69 = 0.31$ and $\kappa_\beta \leq 0.31$, respectively.

Next, we show the claim that $e^{-\beta s} \leq r(-s) \leq e^{-\alpha s}$ for all $s \in (0, s_*)$ with $s_* = 2$. To begin with, we shall prove that $r(-s) > e^{-\beta s}$ for all $s \in (0, \infty)$, which is equivalent to

$$\psi(s) = e^{\beta s}(-6s + 24) - (s^3 + 6s^2 + 18s + 24) > 0 \quad \forall\, s \in (0, s_*).$$

We note that

$$\psi^{(4)}(s) = 6\beta^3 e^{\beta s}(-\beta s + 4\beta - 4).$$

has a unique root in $(0, \infty)$, namely $s_0 = \frac{4\beta-4}{\beta} \approx 0.0784$, and hence $\psi^{(4)}(s) > 0$ for all $s \in (0, s_0)$. Besides, we observe that $\psi^{(3)}(0) = 0.742 > 0$, $\psi^{(3)}(s_0) \approx 0.762 > 0$. Therefore $\psi^{(3)}(s)$ has a unique root in $(s_0, +\infty)$, denoted as s_1. By means of the fixed point iteration, we know that $s_1 \approx 0.4832$. Similarly, since $\psi''(0) = 0.730 > 0$ and $\psi''(s_1) = 1.00 > 0$, we conclude that $\psi''(s)$ is always positive in $[0, s_1]$ and it has a unique root $s_2 \in (s_1, \infty)$, and we find $s_2 \approx 0.9980$. Repeating the argument, we are able to show that $\psi'(s)$ keeps positive in $[0, s_2]$ and the unique root in (s_2, ∞) locates at $s_3 \approx 1.5344$. Finally, we observe that $\psi(0) = 0$ and $\psi(s_3) \approx 1.401 > 0$, so ψ is positive in $(0, s_3]$ and it admits a unique root at $s_4 \in (s_3, \infty)$. Noting that $\psi(s_*) \approx 0.2873 > 0$, we conclude that $r(-s) > e^{-\beta s}$ for all $s \in (0, s_0)$.

Next we will show that $r(-s) < e^{-\alpha s}$ in $(0, s_*)$, which is equivalent to show

$$\varphi(s) = e^{\alpha s}(-6s + 24) - (s^3 + 6s^2 + 18s + 24) < 0 \quad \text{for all } s \in (0, s_*).$$

We note the fact that

$$\varphi^{(4)}(s) = 6\alpha^3 e^{\alpha s}(-\alpha s + 4\alpha - 4) < 0 \quad \text{for all } s \in (0, \infty).$$

Meanwhile, we have $\varphi^{(4)}(0) < 0$, $\varphi^{(3)}(0) < 0$, $\varphi''(0) < 0$, $\varphi'(0) < 0$, and $\varphi(0) < 0$. Those together imply $\varphi(s) < 0$ for any $s \in (0, \infty)$. Therefore $r(-s) < e^{-\alpha s}$.

As a result, for any $J \geq 2$, we arrive at $e^{-\beta s} \leq r(-s/J)^J \leq e^{-\alpha s}$ for all $s \in (0, Js_*)$, that further implies the estimate

$$\sup_{s \in (0, Js_*)} \left| \frac{(1+s)r(-s/J)^J - 1}{s} \right| \leq 0.31.$$

Next, we aim to prove the claim that

$$\sup_{(s_*,\infty)} \big|r(-s)\big| = \sup_{(s_*,\infty)} \Big|\frac{24-6s}{s^3+6s^2+18s+24}\Big| \le 0.15. \tag{4.15}$$

To begin with, we show that $\frac{\mathrm{d}}{\mathrm{d}s} r(-s)$ admits a unique root in $(2,\infty)$. We note that

$$\frac{\mathrm{d}}{\mathrm{d}s} r(-s) = \frac{12(s^3-3s^2-24s-48)}{(s^3+6s^2+18s+24)^2}.$$

and hence it is sufficient to show that $\eta(s) = s^3-3s^2-24s-48$ has a unique root in $(2,\infty)$. Since $\eta'(s)$ has two roots, -2 and 4, and $\eta(-2) = -20 < 0$, $\eta(4) = -128 < 0$, we conclude that $\eta(s) < 0$ for all $s \in [-2,4]$, and $\eta(s)$ admits a unique root in $(4,\infty)$, namely $s_5 \approx 7.235$. Therefore $r(-s)$ is decreasing in (s_*, s_5) and increasing in $[s_5,\infty)$. Noting that fact that $r(-s_*) \approx 0.130$, $r(-s_5) \approx -0.0229$ and $r(-\infty) = 0$, we obtian $\sup_{(s_*,\infty)} \big|r(-s)\big| = \big|r(-s_*)\big| \le 0.15$.

Therefore, we derive for $s \in (Js_*,\infty)$ and $J \ge 2$

$$\begin{aligned}\left|\frac{(1+s)r(-s/J)^J - 1}{s}\right| &\le \frac{1+s}{s}|r(-s/J)^J| + s^{-1} \le \frac{1+Js_*}{Js_*}(0.15)^J + (Js_*)^{-1} \\ &\le \frac{5}{4}(0.02)^2 + 4^{-1} \approx 0.251 \le 0.31.\end{aligned}$$

This completes the proof of (4.14) as well as the proposition. □

Proof for the 4-stage Lobatto IIIC scheme (4.12): Similar to the proof of Proposition 4.3.1, we aim to show that for any $J \ge 2$

$$\sup_{s\in(0,\infty)} \left|\frac{(1+s)r(-s/J)^J - 1}{s}\right| \le 0.31, \ \text{ where } \ r(-s) = \frac{12s^2-120s+360}{s^4+12s^3+72s^2+240s+360}. \tag{4.16}$$

Letting $\alpha = 0.69$ and $\beta = 1.02$, Lemmas 4.2.1 and 4.3.1 implies that $\kappa_\alpha \le 1-0.69 = 0.31$ and $\kappa_\beta \le 0.31$, respectively. Next, we show the claim that

$$e^{-\beta s} \le r(-s) \le e^{-\alpha s} \ \text{ for all } \ s \in (0, s_*) \ \text{ with } s_* = 6.8. \tag{4.17}$$

To begin with, we show that, for $s \in (0, \infty)$, $r(-s) \geq e^{-\beta s}$, which is equivalent to

$$\psi(s) = (12s^2 - 120s + 360)e^{\beta s} - (s^4 + 12s^3 + 72s^2 + 240s + 360) \geq 0.$$

Define $h(s) = \beta^2 s^2 + (10\beta - 10\beta^2)s + (30\beta^2 - 50\beta + 20)$, then we have

$$\psi^{(5)}(s) = 12\beta^3 e^{\beta s}\left[\beta^2 s^2 + (10\beta - 10\beta^2)s + (30\beta^2 - 50\beta + 20)\right] = 12\beta^3 e^{\beta s} h(s).$$

Here $h(s)$ is a quadratic polynomial, whose minimum locates at $\frac{5\beta-5}{\beta}$. Therefore $h(s) \geq h(\frac{5\beta-5}{\beta}) > 0$. Then $\psi^{(5)}(s) = 12\beta^2 e^{\beta x} h(s) - 24 > 12 \times 2 - 24 > 0$. Meanwhile, simple computation yields

$$\psi(0) = 0 \qquad \text{and} \qquad \psi^{(k)}(0) > 0 \qquad \text{with} \qquad 1 \leq k \leq 5.$$

Then we conclude that $\psi(s) > 0$ for all $s \in (0, \infty)$, and hence $r(-s) \geq e^{-\beta s}$ in $(0, \infty)$.

Next we show the bound that $r(s) \leq e^{-\alpha s}$ for $s \in (0, s_*)$, which is equivalent to show

$$\varphi(s) = (12s^2 - 120s + 360)e^{\alpha s} - (s^4 + 12s^3 + 72s^2 + 240s + 360) \leq 0 \qquad \forall\, s \in (0, s_*).$$

Similar to the preceding argument, let $g(s) = (20 - 50\alpha + 30\alpha^2 + 10\alpha s - 10\alpha^2 s + \alpha^2 s^2)$. Then

$$\varphi^{(5)}(s) = 12\alpha^3 e^{\alpha s}(20 - 50\alpha + 30\alpha^2 + 10\alpha s - 10\alpha^2 s + \alpha^2 s^2) = 12\alpha^3 e^{\alpha s} g(s).$$

Here g is a quadratic polynomial with minimum at $\frac{5\alpha-5}{\alpha}$. Therefore, $g(s) \geq g(\frac{5\alpha-5}{\alpha}) \approx -2.62$. Meanwhile, we observe that $g(0) = -0.217 < 0$, so there is a unique root of g in $(0, \infty)$. It is easy to find that, by means of the fixed point iteration, that root locates at $s_0 \approx 0.0993$. Then $\varphi^{(5)}(s) \leq 0$ for all $s \in [0, s_0]$ and $\varphi^{(5)}(s) \geq 0$ for $s \in (s_0, \infty)$. Noting that $\varphi^{(4)}(s_0) < 0$ and $\varphi^{(4)}(0) < 0$, so $\varphi^{(4)}(s) < 0$ in $[0, s_0]$ and $\varphi^{(4)}$ admits a unique root in (s_0, ∞), named as s_1. Then the fixed point iteration implies $s_1 \approx 1.6849$. Repeating this argument, we are able to show that $\varphi^{(3)}(s) < 0$ in $[0, s_1]$ and $\varphi^{(3)}$ has a unique root (s_1, ∞), namely $s_2 \approx 3.0558$. Then we derive that $\varphi''(s) < 0$ in $[0, s_2]$ and $\varphi''(s)$ has a unique root in (s_2, ∞), denoted as $s_3 \approx 4.3640$. Similarly, $\varphi'(s) < 0$ in $[0, s_3]$ and $\varphi'(s)$ has a unique root in (s_3, ∞), named as $s_4 \approx 5.6285$. Finally, since $\psi(0) = 0$ and $\varphi(s_4) < 0$, we conclude that

$\varphi(s) < 0$ in $(0, s_4]$, and $\varphi(s)$ has a unique root in (s_4, ∞). Then the fact that $\varphi(s_*) \approx -447.65 < 0$ implies $\varphi(s) < 0$ in $(0, s_*)$. This completes the proof of the claim (4.17). As a result, for any $J \geq 2$, we arrive at $e^{-\beta s} \leq r(-s/J)^J \leq e^{-\alpha s}$ for all $s \in (0, Js_*)$ which implies

$$\sup_{s\in(0,Js_*)} \left| \frac{(1+s)r(-s/J)^J - 1}{s} \right| \leq 0.31.$$

Next, we intend to show the claim that

$$\sup_{(s_*,\infty)} \big|r(-s)\big| = \sup_{(s_*,\infty)} \Big|\frac{12s^2 - 120s + 360}{s^4 + 12s^3 + 72s^2 + 240s + 360}\Big| \leq 0.02. \tag{4.18}$$

In order to establish a bound for the supremum, we note

$$\frac{d}{ds} r(-s) = \frac{-24s^5 + 216s^4 + 1440s^3 - 1440s^2 - 43200s - 129600}{(s^4 + 12s^3 + 72s^2 + 240s + 360)^2},$$

and we will show that it admits a unique root in (s_*, ∞), denoted by s_5. Noting that, with $\mu(s) = -s^5 + 9s^4 + 60s^3 - 60s^2 - 1800s - 5400$, we have

$$\frac{d}{ds} r(-s) = \frac{24\mu(s)}{(s^4 + 12s^3 + 72s^2 + 240s + 360)^2}.$$

So it suffices to show that $\mu(s)$ admits a unique root in (s_*, ∞). Since $\mu^{(3)}(s) = -60s^2 + 216s + 360$, being a quadratic polynomial, it gains the maximum at 1.8 where $\mu^{(3)}(1.8) = 554.4 > 0$. Noting that $s_* > 1.8$ and $\mu^{(3)}(s_*) = -945.6 < 0$, we conclude that $\mu^{(3)}(s) < 0$ for all $s \in (s_*, \infty)$. Moreover, since $\mu''(s_*) \approx 1033.3 > 0$ and $\mu''(\infty) = -\infty$, $\mu''(s)$ admits a unique root $s_6 \approx 7.6503 \in (s_*, \infty)$. Then we know that $\mu''(s) > 0$ in (s_*, s_6) and $\mu''(s) < 0$ in (s_6, ∞). Similarly, we have $\mu'(s_*) > 0$ and $\mu'(s_6) > 0$, so $\mu'(s) > 0$ in $[s_*, s_6]$. This together with the fact $\mu'(\infty) < 0$ implies that $\mu'(s)$ has a unique root $s_7 \approx 10.166 \in (s_6, \infty)$. Finally, using the facts that $\mu(s_*) > 0$ and $\mu(s_7) > 0$, we know $\mu(s) > 0$ in (s_*, s_7) and $\mu(s)$ has a unique root $s_8 \approx 12.28 \in (s_7, \infty)$. Therefore $r(-s)$ is increasing in (s_*, s_8), and decreasing in (s_8, ∞). Noting that $r(-s) \approx 0.0088$, $r(-s_8) \approx 0.0118$, and $r(-\infty) = 0$, we arrive at $\sup_{(s_*,\infty)} \big|r(-s)\big| \leq 0.02$.

As a result, the estimate (4.18) implies that for $s \in (Js_*, \infty)$ and $J \geq 2$

$$\left|\frac{(1+s)r(-s/J)^J - 1}{s}\right| \leq \frac{1+s}{s}|r(-s/J)^J| + s^{-1} \leq \frac{1+Js_*}{Js_*}(0.02)^J + (Js_*)^{-1}$$
$$\leq \frac{14.6}{13.6}(0.02)^2 + 13.6^{-1} \approx 0.074 \leq 0.31.$$

This completes the proof of (4.16) as well as the proposition. □

Remark 4.3.1. *Propositions 4.3.1 and 4.3.2 show that, for two-, three-, four-stage Lobbatto IIIC schemes, the convergence factor is (at worst)* 0.31*, and there is no restriction on the ratio between the coarse time step and fine time step. It is still possible to improve those estimations, by means of Theorem 4.2.1. For example, one may obtain a smaller convergence factor* γ *by choosing a bigger* α *and a smaller* β*, which might not affect the threshold* $J_* = 2$*.*

Remark 4.3.2. *In the proof of Propositions 4.3.1 and 4.3.2, we employ the L-stability (*$r(-\infty) = 0$*) of the two-, three-, four-stage Lobbatto IIIC schemes. If the* $r(-\infty) \in (0, 1)$*, the analysis might be more technical, and the convergence might be slow for small step ratio* J*; see e.g. Figure 4.3 for the Calahan scheme* (4.20)–(4.21)*. However, Theorem 4.2.1 guarantees the existence of the threshold* J_* *such that for any* $J \geq J_*$ *the convergence factor is close to* 0.3*.*

Remark 4.3.3. *The previous analysis shows that different* J *lead to almost the same convergence rate. For a smaller* J*, or more corase intervals, the algorithm will be guaranteed a higher parallel ability and takes less CPU time if we have plenty CPU cores.*

4.4 Numerical Results

In this section, we shall present some numerical examples to illustrate and complement our theoretical results. To begin with, we use the one-dimensional diffusion models to show the sharpness of our convergence analysis in Sections 4.2 and 4.3.

Example 1. Linear Diffusion Models We consider the following initial-boundary value problem of parabolic equations

$$\begin{cases} \partial_t u(x,t) - \partial_{xx} u(x,t) = f(x,t), & 0 < t < T, \\ u(x,t) = 0, & x \in \partial\Omega,\ 0 < t < T, \\ u(x,0) = u^0(x), & x \in \Omega. \end{cases} \tag{4.19}$$

where $\Omega = (0,\pi)$ and $T = 1$. We consider the following two sets of problem data

(a) $u^0(x) = x^5(1-x)^5/(\pi/2)^{10}$ and $f \equiv 0$;

(b) $u^0(x) = \chi_{(0,\frac{\pi}{2})}(x)$ and $f = \cos(t)\sin(x)$, where $\chi_{(0,\frac{\pi}{2})}(x)$ denotes the step function:

$$\chi_{(0,\pi/2)}(x) = \begin{cases} 1, & x \in (0,\pi/2), \\ 0, & \text{elsewise.} \end{cases}$$

In the computation, we divided the domain Ω into with M equal subintervals of length $h = \pi/M$ and apply the Galerkin finite element with piesewise linear polynomials to discretize in space. We examine the error between the parareal iterative solution U_k^n and the exact time stepping solution U^n, i.e.,

$$\text{error} = \max_{1\le n\le N_c} \|U_k^n - u^{nJ}\|_{L^2(\Omega)}.$$

In our computation, we fixed spatial mesh size $h = \pi/1000$, and choose the initial guess $U_0^n = u^0$ for all $n = 0,\dots,N$.

In example (a), the data is sufficiently smooth and compatible to the homogeneous Dirichlet boundary condition. In fact, it is easy to show show that $u^0 \in \text{Dom}((-\Delta)^{3+\varepsilon})$ with $\varepsilon \in (0,1/4)$ (see e.g. Lemma [110, Lemma 3.1]). For this case of regular data, Bal showed that the first several parareal iterations converge linearly with the rate $O(\Delta T)$; see cf. [7]. This is fully supported by the numerical results presented in Figure 4.2, where we show the convergence of parareal algorithm for 2- and 3-stage Lobatto IIIC methods with fixed $J = 10$ and $\Delta T = 1/100,\ 1/300,\ 1/600$ (and correspondingly $\Delta t = 1/1000,\ 1/3000,\ 1/6000$). We observe that the convergence of the first several iterations is faster for smaller coarse step size, but the convergence then deteriorates for the later iterations.

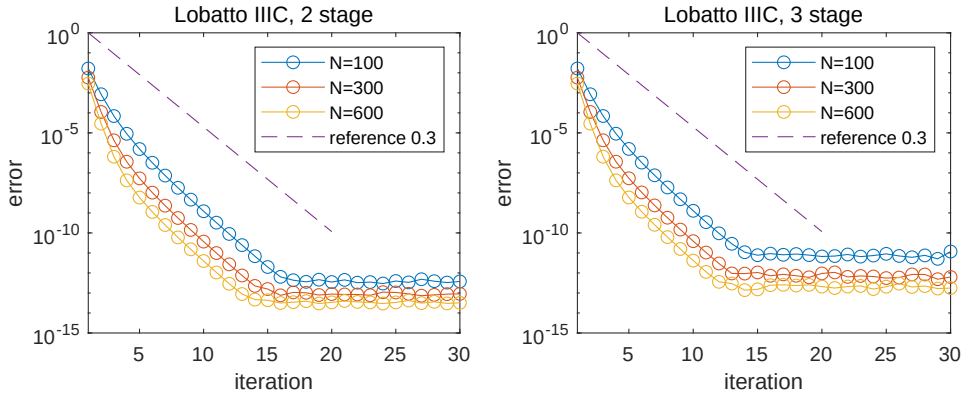


Figure 4.2: Example 1 (a): smooth data. Convergence of the parareal algorithm for 2- and 3-stage Lobatto IIIC methods with fixed mesh ratio $J = 10$ and various coarse step sizes $1/N$, $N = 100, 300, 600$.

In Figure 4.3, we show the convergence of parareal algorithm for 2-, 3-, 4-stage Lobatto IIIC methods solving parabolic equation with nonsmooth initial data, i.e. Example 1 (b). We fixed the fine step size $\Delta t = 1/3000$ and use different step ratios $J = 2$, 3 and 10. The numerical experiments clearly show that the parareal iterations converge linearly with convergence factor near 0.3 for all $J \geq 2$. Meanwhile, we observe that the convergence factor is independent of the ratio between coarse and find step sizes. These phenomenon fully support our theoretical findings in Propositions 4.3.1 and 4.3.2. Moreover, we test another time integrator, called Calahan scheme [125, eq. (1.9)], defined by

$$r(-s) = 1 - \frac{s}{1+bs} - \frac{\sqrt{3}}{6}\Big(\frac{s}{1+bs}\Big)^2, \qquad \text{with } b = \frac{1}{2}\Big(1 + \frac{\sqrt{3}}{3}\Big) \tag{4.20}$$

and

$$\begin{aligned} c_1 &= \frac{1}{3}, \qquad p_1(-s) = \frac{(1/2+\sqrt{3}) + (\sqrt{3}/2)s}{(1+bs)^2}; \\ c_2 &= \frac{2}{3}, \qquad p_2(-s) = \frac{(1/2-\sqrt{3}) + (1/2-\sqrt{3}/2)s}{(1+bs)^2}. \end{aligned} \tag{4.21}$$

The Butcher tableau is given by

$$\begin{array}{cc|c} \frac{1}{6}\sqrt{3}+\frac{2}{3} & -\frac{1}{6}\sqrt{3}-\frac{1}{3} & \frac{1}{3} \\ -\frac{1}{6}\sqrt{3}+\frac{1}{3} & \frac{1}{6}\sqrt{3}+\frac{1}{3} & \frac{2}{3} \\ \hline \frac{1}{2} & \frac{1}{2} & \end{array} =: \begin{array}{c|c} \mathcal{A} & c \\ \hline b^\top & \end{array} \tag{4.22}$$

It is easy to see that $r(-s)$ is a decreasing function on $(0,\infty)$ and $r(-\infty) = 1-\sqrt{3} \in (-1,0)$, so it is *A-stable*, but not *L-stable*. Besides, the scheme is accurate of order $k = 3$. Therefore, the Calahan scheme satisfies Conditions (P1)–(P3). Numerical results show that the converegnce of the corresponding parareal iterations is much slower than 0.3 for small J. This might be due to the fact that $|r(-\infty)| > 0$; see Remark 4.3.2. However, for large J, the numerical results indicate a convergence rate close to 0.3, as predicted by Theorem 4.2.2.

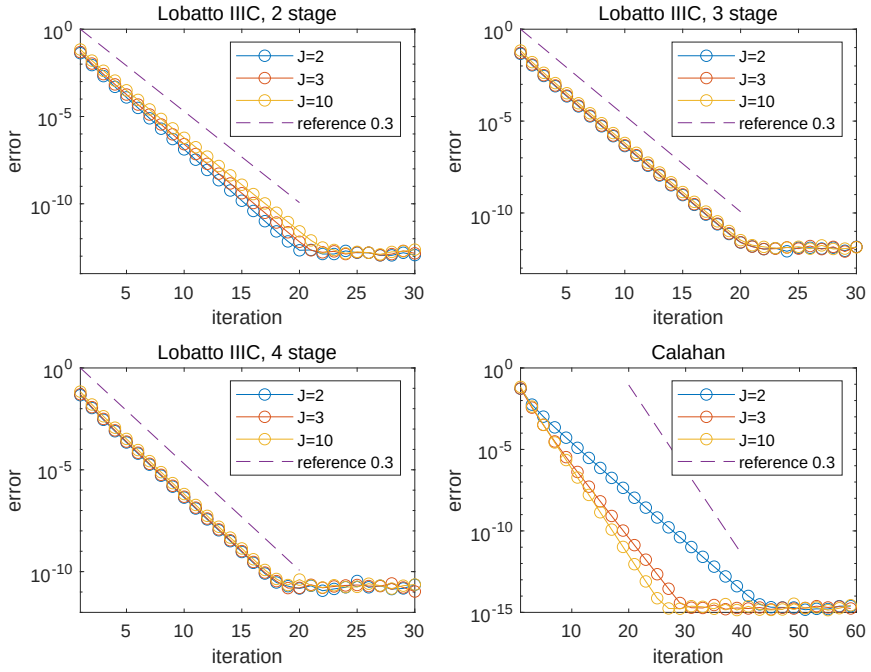


Figure 4.3: Example 1 (b): nonsmooth data. Convergence of the parareal algorithm for 2-, 3-, 4-stage Lobatto IIIC methods and Calahan method with fixed fine step size $\Delta t = 1/3000$ and various ratios of coarse step size and fine step size.

Example 2. Semilinear Parabolic Equations In this part, we shall examine the convergence of parareal algorithm for solving the initial-boundary value problem of the semilinear parabolic equations

$$\begin{cases} \partial_t u - \partial_{xx} u = \dfrac{1}{\varepsilon^2}(u - u^3) =: f(u), & \text{for all } x \in \Omega, t \in (0, T], \\ \partial_x u(x, t) = 0, & \text{for all } x \in \{0, \pi\}, t \in (0, T], \\ u(x, 0) = u^0(x), & \text{for all } x \in \Omega. \end{cases} \tag{4.23}$$

The model (4.23), called Allen–Cahn equation, was originally introduced by Allen and Cahn in [2] to describe the motion of anti-phase boundaries in crystalline solids. In the context, u represents the concentration of one of the two metallic components of the alloy and the parameter ε involved in the nonlinear term represents the width of interface. Recent decades, the Allen–Cahn equation has become one of basic phase-field equations, which has been widely applied to many complicated moving interface problems in materials science and fluid dynamics [3, 16, 122].

In our numerical scheme, the coarse propagator is the semli-implicit backward Euler scheme: for given u^n, look for u^{n+1} such that for all $\phi \in H^1(0, \pi)$

$$(u^{n+1}, \phi) + \Delta T(\partial_x u^{n+1}, \partial_x \phi) = (u^n, \phi) + \Delta T(f(u^n), \phi),$$

which is uniquely solvable and first-order accurate, see e.g. [110, Theorem 14.7]. Meanwhile, the fine propagator is an arbitrary fully implicit high-order single step integrator (such as the Lobatto IIIC schemes or the fully implicit Calahan scheme): for given u^n, look for u^{n+1} such that for all $\phi \in H^1(0, \pi)$

$$\begin{cases} (u^{ni}, \phi) = (u^{n-1}, \phi) + \Delta t \displaystyle\sum_{j=1}^{m} a_{ij} \Big(- (\partial_x u^{nj}, \partial_x \phi) + (f(u^{nj}), \phi) \Big) \ \text{ for } 1 \le i \le q, \\ (u^n, \phi) = (u^{n-1}, \phi) + \Delta t \displaystyle\sum_{i=1}^{m} b_i \Big(- (\partial_x u^{ni}, \partial_x \phi) + (f(u^{ni}), \phi) \Big), \end{cases} \tag{4.24}$$

where the nonlinear system is uniquely solvable for sufficiently small step size, and we solve it by use Newton's algorithm. Note that the fine propagator is fully nonlinear and hence time consuming where the coarse propagator is a linear scheme, so the application of parareal algorithm is able to significantly improve the efficiency.

In Figure 4.4, we show the convergence of parareal algorithm for 2-, 3-, 4-stage Lobatto IIIC methods

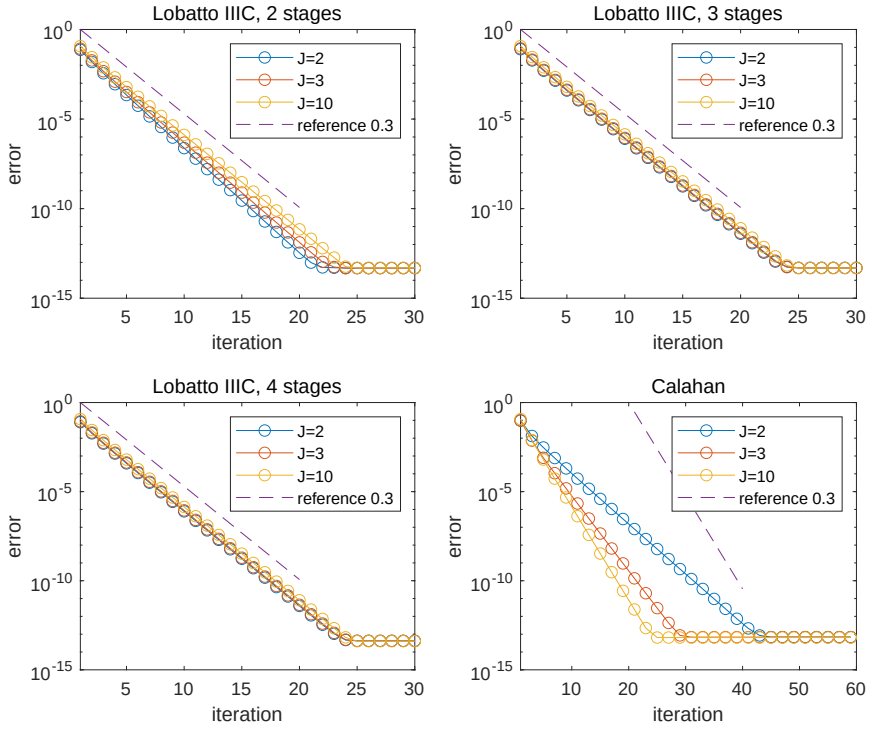


Figure 4.4: Example 2. Convergence of the parareal algorithm for 2-, 3-, 4-stage Lobatto IIIC methods and Calahan method with fine step size $\Delta t = 1/600$ and various step ratios $J = 2, 3, 10$.

and the Calahan method solving the semilinear parabolic equation (1.1) with $\varepsilon = 1$ and $T = 0.1$. The fine step size is fixed and we examine the convergence for different step ratios. Similar to the linear problem, for Lobatto IIIC methods, the numerical experiments clearly show that the parareal iterations converge linearly with convergence factor near 0.3 for all $J \geq 2$, while for the Calahan method the parareal iterations converge slowly for a small J. The convergence analysis for the nonlinear problem warrants further investigation in our future studies.

4.5 Conclusion and Comments

In this chapter, we study a parallel-in-time algorithm, named parareal algorithm, to speed up our simulation. We will first start on linear equations. The prove is based on spectrum decomposition. We prove that, for a fixed coarse propagator (Implicit Euler Method), for any single step method fine propagator, as long as the scheme satisfy the given assumptions, we can always find a threshold J_*, such that if the mesh ratio $J \geq J_*$, the convergence rate for the iteration is bounded by a given constant about 0.3. A lot of famous schemes agree with our assumptions. We also tested Allen-Cahn equation and it also works on it.

Chapter 5

Conclusions and Future Work

This thesis aims to develop efficient single-step methods for solving parabolic problems, particularly in phase-field models, and ensure high accuracy while preserving maximum bound and energy dissipation.

In the first part of the thesis, we focus on the development and analysis of structure-preserving schemes for solving Allen–Cahn equations, a significant application of parabolic equations. We employ a k-th order single-step method in time, linearizing the nonlinear term using multi-step extrapolation. In space, we use a lumped mass finite element method with piecewise r-th order polynomials and Gauss–Lobatto quadrature. A cut-off post-processing technique is proposed at each time level to eliminate values violating the maximum bound principle at finite element nodal points. Consequently, the numerical solution adheres to the maximum bound principle at all nodal points, and the optimal error bound $O(\tau^k + h^{r+1})$ is theoretically proven. These time-stepping schemes include algebraically stable collocation-type methods, achieving high order in both space and time. By integrating the cut-off strategy with the scalar auxiliary variable (SAV) technique, we develop energy-stable and maximum bound preserving schemes of arbitrarily high order in time.

In the second part, we start to develop and analyze a class of single-step implicit-explicit schemes for approximately solving linear parabolic equations, achieving long-time stability and arbitrarily high order in time. This approach involves splitting the linear operator into symmetric and skew-symmetric components, which are evaluated implicitly and explicitly respectively using IMEX-RK. For the symmetric part, a diagonally implicit method is employed, while the discretization for the skew-symmetric part is

designed to satisfy the stage orders. This method is applicable to semilinear problems, such as phase-field models, and our analysis aligns with existing findings, demonstrating energy stability for certain IMEX-RK schemes. Our results reveal intersections up to at least third order, leading to a scheme that preserves both the original energy decay properties and maximum bound principles.

In the third part of the thesis, we study the parareal algorithm for solving parabolic equations, enabling parallel-in-time computation and significantly accelerating the process. We prove that the parareal method has a robust convergence rate of about 0.3, provided the ratio J of coarse to fine step size exceeds a certain threshold J_*, and the fine propagator meets mild conditions. This convergence holds even with nonsmooth problem data and boundary condition incompatibilities. Qualified methods include all absolutely stable single-step methods with a stability function satisfying $|r(-\infty)| < 1$, allowing the fine propagator to be of arbitrarily high order. We also examine popular high-order single-step methods, such as the two-, three-, and four-stage Lobatto IIIC methods, confirming that their corresponding parareal algorithms converge linearly with a factor of 0.31 and a threshold $J_* = 2$. At the end of each chapter, we present numerical results that support the theoretical findings and inspire future investigations.

Next we list several perspectives for the future research:

1. Cut-off postprocessing can naturally ensure maximum bound preservation and maintain energy decay when the function is continuous in space or uses piecewise linear FEM. However, for higher-order finite elements, like $P^2(I)$ for $I \subset \mathbb{R}^1$, a simple cut-off on u may increase $|\nabla u|$, causing our analysis to fail. Thus, for both cut-off RK-SAV, which preserves maximum bounds and modified energy decay, and cut-off IMEX-RK, which preserves maximum bounds and original energy decay, we achieve at most second-order spatial accuracy in fully discretized problems. We need to explore alternative methods, such as enhanced cut-off or other postprocessing techniques, to achieve higher-order spatial convergence, requiring further study.

2. In the proof of long-time error estimate for linear implicit-explicit Runge-Kutta methods, we assume $0 < \sigma(\lambda) < 1$, which is equivalent to $|\sigma| < 1$ and $\sigma > 0$. We require $\sigma > 0$ to ensure the inverse operator and norms are well-posed. Although this condition seems redundant, the proof fails without it. We hope to address this issue in the future.

3. We use the undetermined coefficients method to explore implicit-explicit Runge-Kutta schemes. Although there are more degrees of freedom than equations, we have not shown that these schemes

can achieve arbitrarily high order. Unlike the usual Runge-Kutta method, where high-order schemes can be systematically built, this approach does not work for implicit-explicit Runge-Kutta. Developing an algorithm to construct high-order implicit-explicit Runge-Kutta schemes, rather than searching for them, could be beneficial.

4. When applying IMEX-RK schemes to solve gradient flow models, preserving the energy dissipation property is crucial. Unfortunately, this requires strict conditions, including the positive-definiteness of coefficient matrices, which are more complex than algebraic stage-order conditions. Consequently, our strategy fails to ensure the energy dissipation property for high-order schemes. While we have not ruled out the possibility of IMEX-RK achieving higher order for original energy decay, we have only found first, second, and third-order results. Discovering fourth or higher-order schemes would be both important and interesting.

5. We established the robust convergence rate for the parareal method using a single-step approach. Extending this work to multi-step methods, such as BDF, is intriguing. Developing the coarse propagator carefully is crucial to prevent algorithm instability. Understanding the conditions for convergence and creating a suitable coarse propagator is essential for applying the parareal method to a wider range of scenarios.

6. Although we have only proven robust convergence for the parareal method on linear equations, numerical experiments indicate it also performs well on nonlinear models, such as the Allen-Cahn equations, with similar convergence behavior. Various coarse propagators can be used, each behaving differently, and some achieving excellent convergence. Developing the theory and analysis for the parareal method on semilinear equations with nonsmooth data is needed.

References

[1] Georgios Akrivis, Buyang Li, and Dongfang Li. “Energy-decaying extrapolated RK–SAV methods for the Allen–Cahn and Cahn–Hilliard equations”. In: *SIAM Journal on Scientific Computing* 41.6 (2019), A3703–A3727.

[2] Samuel M Allen and John W Cahn. “A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening”. In: *Acta metallurgica* 27.6 (1979), pp. 1085–1095.

[3] Daniel M Anderson, Geoffrey B McFadden, and Adam A Wheeler. “Diffuse-interface methods in fluid mechanics”. In: *Annual review of fluid mechanics* 30.1 (1998), pp. 139–165.

[4] Todd Arbogast and Mario San Martin Gomez. “A discretization and multigrid solver for a Darcy–Stokes system of three dimensional vuggy porous media”. In: *Computational Geosciences* 13.3 (2009), pp. 331–348.

[5] Ivo Babuška and Gabriel N Gatica. “A residual-based a posteriori error estimator for the Stokes–Darcy coupled problem”. In: *SIAM Journal on Numerical Analysis* 48.2 (2010), pp. 498–523.

[6] Leonardo Baffico et al. “Parallel-in-time molecular-dynamics simulations”. In: *Physical Review E* 66.5 (2002), p. 057701.

[7] Guillaume Bal. “On the convergence and the stability of the parareal algorithm to solve partial differential equations”. In: *Domain decomposition methods in science and engineering*. Springer, 2005, pp. 425–432.

[8] Yassine Boubendir and Svetlana Tlupova. “Domain decomposition methods for solving Stokes–Darcy problems with boundary integrals”. In: *SIAM Journal on Scientific Computing* 35.1 (2013), B82–B106.

[9] Philip Brenner, Michel Crouzeix, and Vidar Thomée. “Single step methods for inhomogeneous linear differential equations in Banach space”. In: *RAIRO. Analyse numérique* 16.1 (1982), pp. 5–26.

[10] John C Butcher. “On the implementation of implicit Runge-Kutta methods”. In: *BIT Numerical Mathematics* 16.3 (1976), pp. 237–240.

[11] Jessika Camano et al. “New fully-mixed finite element methods for the Stokes–Darcy coupling”. In: *Computer Methods in Applied Mechanics and Engineering* 295 (2015), pp. 362–395.

[12] Yanzhao Cao et al. “Finite element approximations for Stokes–Darcy flow with Beavers–Joseph interface conditions”. In: *SIAM Journal on Numerical Analysis* 47.6 (2010), pp. 4239–4256.

[13] Yanzhao Cao et al. “Robin–Robin domain decomposition methods for the steady-state Stokes–Darcy system with the Beavers–Joseph interface condition”. In: *Numerische Mathematik* 117.4 (2011), pp. 601–629.

[14] Jie Chen, Shuyu Sun, and Xiao-Ping Wang. “A numerical method for a model of two-phase flow in a coupled free flow and porous media system”. In: *Journal of Computational Physics* 268 (2014), pp. 1–16.

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

[16] Long-Qing Chen. “Phase-field models for microstructure evolution”. In: *Annual review of materials research* 32.1 (2002), pp. 113–140.

[17] Wenbin Chen et al. “A parallel Robin–Robin domain decomposition method for the Stokes–Darcy system”. In: *SIAM Journal on Numerical Analysis* 49.3 (2011), pp. 1064–1084.

[18] Wenbin Chen et al. “An efficient and long-time accurate third-order algorithm for the Stokes–Darcy system”. In: *Numerische Mathematik* 134 (2016), pp. 857–879.

[19] Wenbin Chen et al. “Efficient and long-time accurate second-order methods for the Stokes–Darcy system”. In: *SIAM Journal on Numerical Analysis* 51.5 (2013), pp. 2563–2584.

[20] Wenbin Chen et al. “Energy stable higher-order linear ETD multi-step methods for gradient flows: application to thin film epitaxy”. In: *Research in the Mathematical Sciences* 7 (2020), pp. 1–27.

[21] Qing Cheng and Jie Shen. “Multiple scalar auxiliary variable (MSAV) approach and its application to the phase-field vesicle membrane model”. In: *SIAM Journal on Scientific Computing* 40.6 (2018), A3982–A4006.

[22] Julien Cortial and Charbel Farhat. “A time-parallel implicit method for accelerating the solution of non-linear structural dynamics problems”. In: *International Journal for Numerical Methods in Engineering* 77.4 (2009), pp. 451–470.

[23] Steven M Cox and Paul C Matthews. “Exponential time differencing for stiff systems”. In: *Journal of Computational Physics* 176.2 (2002), pp. 430–455.

[24] Marco Discacciati, Edie Miglio, and Alfio Quarteroni. “Mathematical and numerical models for coupling surface and groundwater flows”. In: *Applied Numerical Mathematics* 43.1-2 (2002), pp. 57–74.

[25] Marco Discacciati and Alfio Quarteroni. “Convergence analysis of a subdomain iterative method for the finite element approximation of the coupling of Stokes and Darcy equations”. In: *Computing and Visualization in Science* 6 (2004), pp. 93–103.

[26] Marco Discacciati, Alfio Quarteroni, and Alberto Valli. “Robin–Robin domain decomposition methods for the Stokes–Darcy coupling”. In: *SIAM Journal on Numerical Analysis* 45.3 (2007), pp. 1246–1268.

[27] Veselin A Dobrev et al. “Two-level convergence theory for multigrid reduction in time (MGRIT)”. In: *SIAM Journal on Scientific Computing* 39.5 (2017), S501–S527.

[28] Qiang Du and Wen-xiang Zhu. “Stability analysis and application of the exponential time differencing schemes”. In: *Journal of Computational Mathematics* (2004), pp. 200–209.

[29] Qiang Du et al. “Maximum bound principles for a class of semilinear parabolic equations and exponential time-differencing schemes”. In: *SIAM review* 63.2 (2021), pp. 317–359.

[30] Qiang Du et al. “Maximum principle preserving exponential time differencing schemes for the nonlocal Allen–Cahn equation”. In: *SIAM Journal on numerical analysis* 57.2 (2019), pp. 875–898.

[31] Byron L Ehle. *On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems*. 1969.

[32] Charles M Elliott. “The Cahn-Hilliard model for the kinetics of phase separation”. In: (1989), pp. 35–73.

[33] David J Eyre. “An unconditionally stable one-step scheme for gradient systems”. In: *Unpublished article* 6 (1998).

[34] Charbel Farhat and Marion Chandesris. “Time-decomposed parallel time-integrators: theory and feasibility studies for fluid, structure, and fluid–structure applications”. In: *International Journal for Numerical Methods in Engineering* 58.9 (2003), pp. 1397–1434.

[35] Wenqiang Feng et al. “Non-iterative domain decomposition methods for a non-stationary Stokes–Darcy model with Beavers–Joseph interface condition”. In: *Applied Mathematics and Computation* 219.2 (2012), pp. 453–463.

[36] Xinlong Feng, Tao Tang, and Jiang Yang. “Long time numerical simulations for phase-field problems using p-adaptive spectral deferred correction methods”. In: *SIAM Journal on Scientific Computing* 37.1 (2015), A271–A294.

[37] Stephanie Friedhoff and Ben S Southworth. “On “Optimal” h-independent convergence of Parareal and multigrid-reduction-in-time using Runge-Kutta time integration”. In: *Numerical Linear Algebra with Applications* 28.3 (2021), e2301.

[38] Zhaohui Fu, Tao Tang, and Jiang Yang. “Energy diminishing implicit-explicit Runge–Kutta methods for gradient flows”. In: *Mathematics of Computation* (2024).

[39] Zhaohui Fu and Jiang Yang. “Energy-decreasing exponential time differencing Runge–Kutta methods for phase-field models”. In: *Journal of Computational Physics* 454 (2022), p. 110943.

[40] Martin J Gander. “50 years of time parallel time integration”. In: *Multiple Shooting and Time Domain Decomposition Methods: MuS-TDD, Heidelberg, May 6-8, 2013*. Springer, 2015, pp. 69–113.

[41] Martin J Gander and Stefan Vandewalle. “Analysis of the parareal time-parallel time-integration method”. In: *SIAM Journal on Scientific Computing* 29.2 (2007), pp. 556–578.

[42] Gabriel N Gatica, Salim Meddahi, and Ricardo Oyarzúa. “A conforming mixed finite-element method for the coupling of fluid flow with porous media flow”. In: *IMA Journal of Numerical Analysis* 29.1 (2009), pp. 86–108.

[43] Vivette Girault, Danail Vassilev, and Ivan Yotov. “Mortar multiscale finite element methods for Stokes–Darcy flows”. In: *Numerische Mathematik* 127.1 (2014), pp. 93–165.

[44] Yuezheng Gong, Jia Zhao, and Qi Wang. “Arbitrarily high-order unconditionally energy stable schemes for thermodynamically consistent gradient flow models”. In: *SIAM Journal on Scientific Computing* 42.1 (2020), B135–B156.

[45] Sigal Gottlieb, David Ketcheson, and Chi-Wang Shu. *Strong stability preserving Runge-Kutta and multistep time discretizations*. World Scientific, 2011.

[46] Sigal Gottlieb and Chi-Wang Shu. “Total variation diminishing Runge-Kutta schemes”. In: *Mathematics of computation* 67.221 (1998), pp. 73–85.

[47] Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor. “Strong stability-preserving high-order time discretization methods”. In: *SIAM review* 43.1 (2001), pp. 89–112.

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

[49] Max Gunzburger, Xiaoming He, and Buyang Li. “On Stokes–Ritz projection and multistep backward differentiation schemes in decoupling the Stokes–Darcy model”. In: *SIAM Journal on Numerical Analysis* 56.1 (2018), pp. 397–427.

[50] Ruihan Guo and Yan Xu. “Local discontinuous Galerkin method and high order semi-implicit scheme for the phase field crystal equation”. In: *SIAM Journal on Scientific Computing* 38.1 (2016), A105–A127.

[51] Ernst Hairer and Christian Lubich. “Energy-diminishing integration of gradient systems”. In: *IMA J. Numer. Anal.* 34 (2014), pp. 452–461.

[52] NS Hanspal et al. “Numerical analysis of coupled Stokes/Darcy flows in industrial filtrations”. In: *Transport in porous media* 64 (2006), pp. 73–101.

[53] Xiaoming He, Nan Jiang, and Changxin Qiu. “An artificial compressibility ensemble algorithm for a stochastic Stokes-Darcy model with random hydraulic conductivity and interface conditions”. In: *International Journal for Numerical Methods in Engineering* 121.4 (2020), pp. 712–739.

[54] Ronald HW Hoppe, Paulo Porta, and Yuri Vassilevski. “Computational issues related to iterative coupling of subsurface and channel flows”. In: *Calcolo* 44 (2007), pp. 1–20.

[55] Dianming Hou, Hongyi Zhu, and Chuanju Xu. “Highly efficient schemes for time-fractional Allen-Cahn equation using extended SAV approach”. In: *Numerical Algorithms* (2021), pp. 1–32.

[56] Jiangyong Hou et al. “A dual-porosity-Stokes model and finite element method for coupling dual-porosity flow and free flow”. In: *SIAM Journal on Scientific Computing* 38.5 (2016), B710–B739.

[57] Peiqi Huang, Jinru Chen, and Mingchao Cai. “A Mixed and Nonconforming FEM with Non-matching Meshes for a Coupled Stokes-Darcy Model”. In: *Journal of scientific computing* 53.2 (2012), pp. 377–394. ISSN: 0885-7474.

[58] Leah Isherwood, Zachary J Grant, and Sigal Gottlieb. “Strong stability preserving integrating factor Runge–Kutta methods”. In: *SIAM Journal on Numerical Analysis* 56.6 (2018), pp. 3276–3307.

[59] Kenneth R Jackson and Syvert Paul Nørsett. “The potential for parallelism in Runge–Kutta methods. Part 1: RK formulas in standard form”. In: *SIAM journal on numerical analysis* 32.1 (1995), pp. 49–82.

[60] Lili Ju, Xiao Li, and Zhonghua Qiao. “Generalized SAV-exponential integrator schemes for Allen–Cahn type gradient flows”. In: *SIAM journal on numerical analysis* 60.4 (2022), pp. 1905–1931.

[61] Lili Ju et al. “Energy stability and error estimates of exponential time differencing schemes for the epitaxial growth model without slope selection”. In: *Mathematics of Computation* 87.312 (2018), pp. 1859–1885.

[62] Lili Ju et al. “Maximum bound principle preserving integrating factor Runge–Kutta methods for semilinear parabolic equations”. In: *Journal of Computational Physics* 439 (2021), p. 110405.

[63] Guido Kanschat and Béatrice Riviere. “A strongly conservative finite element method for the coupling of Stokes and Darcy flow”. In: *Journal of Computational Physics* 229.17 (2010), pp. 5933–5943.

[64] Ohannes A Karakashian and William Rust. “On the parallel implementation of implicit Runge–Kutta methods”. In: *SIAM journal on scientific and statistical computing* 9.6 (1988), pp. 1085–1090.

[65] Trygve Karper, Kent-Andre Mardal, and Ragnar Winther. “Unified finite element discretizations of coupled Darcy–Stokes flow”. In: *Numerical Methods for Partial Differential Equations: An International Journal* 25.2 (2009), pp. 311–326.

[66] Michaela Kubacki, Marina Moraiti, et al. “Analysis of a second-order, unconditionally stable, partitioned method for the evolutionary Stokes–Darcy model”. In: *Int. J. Numer. Anal. Model* 12.4 (2015), pp. 704–730.

[67] Michaela Kubacki and Hoang Tran. “Non-Iterative Partitioned Methods for Uncoupling Evolutionary Groundwater–Surface Water Flows”. In: *Fluids* 2.3 (2017), p. 47.

[68] Stig Larsson and Vidar Thomée. *Partial differential equations with numerical methods*. Vol. 45. Springer, 2003.

[69] William Layton, Hoang Tran, and Catalin Trenchea. “Analysis of long time stability and errors of two partitioned methods for uncoupling evolutionary groundwater–surface water flows”. In: *SIAM Journal on Numerical Analysis* 51.1 (2013), pp. 248–272.

[70] William J Layton, Friedhelm Schieweck, and Ivan Yotov. “Coupling fluid flow with porous media flow”. In: *SIAM Journal on Numerical Analysis* 40.6 (2002), pp. 2195–2218.

[71] Bo Li and Jian-Guo Liu. “Thin film epitaxy with or without slope selection”. In: *European Journal of Applied Mathematics* 14.6 (2003), pp. 713–743.

[72] Buyang Li, Kai Wang, and Zhi Zhou. “Long-time accurate symmetrized implicit-explicit BDF methods for a class of parabolic equations with non-self-adjoint operators”. In: *SIAM Journal on Numerical Analysis* 58.1 (2020), pp. 189–210.

[73] Buyang Li, Jiang Yang, and Zhi Zhou. “Arbitrarily high-order exponential cut-off methods for preserving maximum principle of parabolic equations”. In: *SIAM Journal on Scientific Computing* 42.6 (2020), A3957–A3978.

[74] Dong Li, Chaoyu Quan, and Jiao Xu. “Stability and convergence of Strang splitting. Part I: scalar Allen-Cahn equation”. In: *Journal of Computational Physics* 458 (2022), p. 111087.

[75] Xianjuan Li, Tao Tang, and Chuanju Xu. “Parallel in time algorithm with spectral-subdomain enhancement for Volterra integral equations”. In: *SIAM Journal on Numerical Analysis* 51.3 (2013), pp. 1735–1756.

[76] Jacques-Louis Lions, Yvon Maday, and Gabriel Turinici. “Résolution d’EDP par un schéma en temps «pararéel»”. In: *Comptes Rendus de l’Académie des Sciences-Series I-Mathematics* 332.7 (2001), pp. 661–668.

[77] Konstantin Lipnikov, Danail Vassilev, and Ivan Yotov. “Discontinuous Galerkin and mimetic finite difference methods for coupled Stokes–Darcy flows on polygonal and polyhedral grids”. In: *Numerische Mathematik* 126.2 (2014), pp. 321–360.

[78] Xu-Dong Liu and Stanley Osher. “Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes I”. In: *SIAM Journal on Numerical Analysis* 33.2 (1996), pp. 760–779.

[79] Hailiang Liu and Hui Yu. “Maximum-principle-satisfying third order discontinuous Galerkin schemes for Fokker–Planck equations”. In: *SIAM Journal on Scientific Computing* 36.5 (2014), A2296–A2325.

[80] Yvon Maday, Julien Salomon, and Gabriel Turinici. “Monotonic parareal control for quantum systems”. In: *SIAM Journal on Numerical Analysis* 45.6 (2007), pp. 2468–2482.

[81] Antonio Márquez, Salim Meddahi, and Francisco-Javier Sayas. “Strong coupling of finite element methods for the Stokes–Darcy problem”. In: *IMA Journal of Numerical Analysis* 35.2 (2015), pp. 969–988.

[82] Tarek P Mathew, Marcus Sarkis, and Christian E Schaerer. “Analysis of block parareal preconditioners for parabolic optimal control problems”. In: *SIAM Journal on Scientific Computing* 32.3 (2010), pp. 1180–1200.

[83] Mo Mu and Jinchao Xu. “A two-grid method of a mixed Stokes–Darcy model for coupling fluid flow with porous media flow”. In: *SIAM journal on numerical analysis* 45.5 (2007), pp. 1801–1813.

[84] Mo Mu and Xiaohong Zhu. “Decoupled schemes for a non-stationary mixed Stokes-Darcy model”. In: *Mathematics of Computation* 79.270 (2010), pp. 707–731.

[85] V Nassehi. “Modelling of combined Navier–Stokes and Darcy flows in crossflow membrane filtration”. In: *Chemical Engineering Science* 53.6 (1998), pp. 1253–1265.

[86] Jürg Nievergelt. “Parallel methods for integrating ordinary differential equations”. In: *Communications of the ACM* 7.12 (1964), pp. 731–733.

[87] Benjamin W Ong and Jacob B Schroder. “Applications of time parallelization”. In: *Computing and Visualization in Science* 23 (2020), pp. 1–15.

[88] Alexander Ostermann and Michel Roche. “Runge-Kutta methods for partial differential equations and fractional orders of convergence”. In: *Mathematics of computation* 59.200 (1992), pp. 403–420.

[89] Changxin Qiu et al. “A domain decomposition method for the time-dependent Navier-Stokes-Darcy model with Beavers-Joseph interface condition and defective boundary condition”. In: *Journal of Computational Physics* 411 (2020), p. 109400.

[90] Jianxian Qiu and Chi-Wang Shu. “Runge–Kutta discontinuous Galerkin method using WENO limiters”. In: *SIAM Journal on Scientific Computing* 26.3 (2005), pp. 907–929.

[91] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*. Vol. 37. Springer Science & Business Media, 2010.

[92] José Miguel Reynolds-Barredo, David E Newman, and Raúl Sanchez. “An analytic model for the convergence of turbulent simulations time-parallelized via the parareal algorithm”. In: *Journal of Computational Physics* 255 (2013), pp. 293–315.

[93] José Miguel Reynolds-Barredo et al. “Mechanisms for the convergence of time-parallelized, parareal turbulent plasma simulations”. In: *Journal of Computational Physics* 231.23 (2012), pp. 7851–7867.

[94] Béatrice Rivière. “Analysis of a discontinuous finite element method for the coupled Stokes and Darcy problems”. In: *Journal of Scientific Computing* 22.1-3 (2005), pp. 479–500.

[95] Béatrice Rivière and Ivan Yotov. “Locally conservative coupling of Stokes and Darcy flows”. In: *SIAM Journal on Numerical Analysis* 42.5 (2005), pp. 1959–1977.

[96] Alfred H Schatz, Vidar Thomée, and Lars B Wahlbin. “On positivity and maximum-norm contractivity in time stepping methods for parabolic equations”. In: *Computational Methods in Applied Mathematics* 10.4 (2010), pp. 421–443.

[97] Li Shan and Haibiao Zheng. “Partitioned time stepping method for fully evolutionary Stokes–Darcy flow with Beavers–Joseph interface conditions”. In: *SIAM Journal on Numerical Analysis* 51.2 (2013), pp. 813–839.

[98] Jie Shen, Tao Tang, and Jiang Yang. “On the maximum principle preserving schemes for the generalized Allen-Cahn equation”. In: *Commun. Math. Sci.* 14.6 (2016), pp. 1517–1534. ISSN: 1539-6746. DOI: 10.4310/CMS.2016.v14.n6.a3. URL: https://doi-org.ezproxy.lb.polyu.edu.hk/10.4310/CMS.2016.v14.n6.a3.

[99] Jie Shen and Jie Xu. “Convergence and error analysis for the scalar auxiliary variable (SAV) schemes to gradient flows”. In: *SIAM Journal on Numerical Analysis* 56.5 (2018), pp. 2895–2912.

[100] Jie Shen, Jie Xu, and Jiang Yang. “A new class of efficient and robust energy stable schemes for gradient flows”. In: *SIAM Review* 61.3 (2019), pp. 474–506.

[101] Jie Shen, Jie Xu, and Jiang Yang. “The scalar auxiliary variable (SAV) approach for gradient flows”. In: *Journal of Computational Physics* 353 (2018), pp. 407–416.

[102] Jie Shen and Xiaofeng Yang. “Numerical approximations of Allen-Cahn and Cahn-Hilliard equations”. In: *Discrete Contin. Dyn. Syst* 28.4 (2010), pp. 1669–1691.

[103] Jie Shen et al. “Second-order convex splitting schemes for gradient flows with Ehrlich–Schwoebel type energy: application to thin film epitaxy”. In: *SIAM Journal on Numerical Analysis* 50.1 (2012), pp. 105–125.

[104] Jaemin Shin, Hyun Geun Lee, and June-Yub Lee. “Convex splitting Runge–Kutta methods for phase-field models”. In: *Computers & Mathematics with Applications* 73.11 (2017), pp. 2388–2403.

[105] Ben S Southworth. “Necessary conditions and tight two-level convergence bounds for parareal and multigrid reduction in time”. In: *SIAM Journal on Matrix Analysis and Applications* 40.2 (2019), pp. 564–608.

[106] Gunnar Andreas Staff and Einar M Rønquist. “Stability of the parareal algorithm”. In: *Domain decomposition methods in science and engineering*. Springer, 2005, pp. 449–456.

[107] Tao Tang. “Revisit of semi-implicit schemes for phase-field equations”. In: *arXiv preprint arXiv: 2006.06990* (2020).

[108] Tao Tang, Xu Wu, and Jiang Yang. “Arbitrarily High Order and Fully Discrete Extrapolated RK–SAV/DG Schemes for Phase-field Gradient Flows”. In: *Journal of Scientific Computing* 93.2 (2022), p. 38.

[109] Tao Tang and Jiang Yang. “Implicit-explicit scheme for the Allen-Cahn equation preserves the maximum principle”. In: *Journal of Computational Mathematics* (2016), pp. 451–461.

[110] Vidar Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Second. Springer-Verlag, Berlin, 2006, pp. xii+370. ISBN: 978-3-540-33121-6; 3-540-33121-2.

[111] Vidar Thomée and Lars Wahlbin. “On the existence of maximum principles in parabolic finite element equations”. In: *Mathematics of computation* 77.261 (2008), pp. 11–19.

[112] Danail Vassilev, ChangQing Wang, and Ivan Yotov. “Domain decomposition for coupled Stokes and Darcy flows”. In: *Computer Methods in Applied Mechanics and Engineering* 268 (2014), pp. 264–283.

[113] Jaap JW van der Vegt, Yinhua Xia, and Yan Xu. “Positivity preserving limiters for time-implicit higher order accurate discontinuous Galerkin discretizations”. In: *SIAM journal on scientific computing* 41.3 (2019), A2037–A2063.

[114] Gerhard Wanner and Ernst Hairer. *Solving ordinary differential equations II*. Vol. 375. Springer Berlin Heidelberg New York, 1996.

[115] Shu-Lin Wu. “Convergence analysis of some second-order parareal algorithms”. In: *IMA Journal of Numerical Analysis* 35.3 (2015), pp. 1315–1341.

[116] Shu-Lin Wu and Tao Zhou. “Convergence analysis for three parareal solvers”. In: *SIAM Journal on Scientific Computing* 37.2 (2015), A970–A992.

[117] Shu-Lin Wu and Tao Zhou. “Fast parareal iterations for fractional diffusion equations”. In: *Journal of Computational Physics* 329 (2017), pp. 210–226.

[118] Chuanju Xu and Tao Tang. “Stability analysis of large time-stepping methods for epitaxial growth models”. In: *SIAM Journal on Numerical Analysis* 44.4 (2006), pp. 1759–1779.

[119] Zhengfu Xu. “Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: one-dimensional scalar problem”. In: *Mathematics of Computation* 83.289 (2014), pp. 2213–2238.

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

[121] Xiaofeng Yang et al. “Numerical approximations for a three-component Cahn–Hilliard phase-field model based on the invariant energy quadratization method”. In: *Mathematical Models and Methods in Applied Sciences* 27.11 (2017), pp. 1993–2030.

[122] Pengtao Yue et al. “A diffuse-interface method for simulating two-phase flows of complex fluids”. In: *Journal of Fluid Mechanics* 515 (2004), p. 293.

[123] Jingyuan Zhang, Hongxing Rui, and Yanzhao Cao. “A partitioned method with different time steps for coupled Stokes and Darcy flows with transport”. In: *Int. J. Numer. Anal. Model* 15 (2019), pp. 463–498.

[124] Xiangxiong Zhang and Chi-Wang Shu. “On maximum-principle-satisfying high order schemes for scalar conservation laws”. In: *Journal of Computational Physics* 229.9 (2010), pp. 3091–3120.

[125] Miloš Zlámal. “Finite element methods for parabolic equations”. In: *mathematics of computation* 28.126 (1974), pp. 393–404.