Efficient and Structure-Preserving Time-Stepping Methods for Nonlinear Evolution Equations and Multiphysics Systems

by

Cao-Kha Doan

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 8, 2026

Keywords: exponential integrators, Lagrange multiplier, transferable neural networks, phase field models, Navier-Stokes equations, hyperbolic conservation laws

Approved by

Thi-Thao-Phuong Hoang, Chair, Associate Professor of Mathematics and Statistics
Yanzhao Cao, Professor of Mathematics and Statistics
Junshan Lin, Professor of Mathematics and Statistics
Hans-Werner van Wyk, Associate Professor of Mathematics and Statistics
Tham C. Hoang, Associate Professor of School of Fisheries, Aquaculture and Aquatic Sciences, University Reader

Abstract

This dissertation is devoted to the design and analysis of efficient and structure-preserving time-stepping methods for nonlinear evolution equations and multiphysics systems. Four novel classes of methods are developed and analyzed: low regularity integrators (LRIs) for phase field models, noniterative localized exponential time differencing (ETD) methods for hyperbolic conservation laws, dynamically regularized Lagrange multiplier (DRLM) methods for the incompressible Navier-Stokes equations and the coupled Cahn-Hilliard-Navier-Stokes system, and generalized transferable neural networks (GTransNet) for the Cahn-Hilliard equation.

We first construct LRI schemes based on Duhamel's principle for the classical and conservative Allen-Cahn equations. The schemes are shown to preserve the maximum bound principle, energy stability, and mass conservation (for the conservative Allen-Cahn equation). Optimal temporal and fully discrete error estimates are rigorously derived under minimal regularity assumptions on the exact solution, requiring only continuity in time for the Allen-Cahn equation and Lipschitz continuity for its conservative counterpart. As a result, LRI schemes outperform classical exponential integrators, particularly when the interfacial parameter approaches zero.

For hyperbolic conservation laws, we combine ETD with the discontinuous Galerkin method and nonoverlapping domain decomposition to develop global and localized ETD schemes that are fully explicit and can be coupled with the total variation bounded (TVB) slope limiter to effectively capture moving shocks. Notably, the localized scheme is noniterative, allowing local problems to be solved in parallel across subdomains while preserving the large time step sizes of the global method. Rigorous temporal error analysis and mass conservation under periodic boundary conditions are established for both schemes.

We then develop the DRLM method for the incompressible Navier-Stokes equations, where a time-dependent Lagrange multiplier and a regularization parameter are introduced to reformulate the equations into an equivalent system. First- and second-order DRLM schemes, with

explicit treatment of the convection term, are shown to be unconditionally energy stable with respect to the original variables. Optimal error estimates are established for the first-order scheme, with error bounds that decay as the regularization parameter increases. The DRLM framework is further extended to the Cahn-Hilliard-Navier-Stokes system, where the corresponding schemes are fully decoupled, mass-conserving, and require no stabilization term.

Finally, we propose a neural network-based framework for the Cahn-Hilliard equation by combining GTransNet for spatial approximation with stabilized backward differentiation formulas (BDF) for time discretization. A key feature of the resulting GTransNet-BDF schemes is the introduction of a mass-conserving projection that corrects output-layer weights produced by the least-squares system to enforce discrete mass conservation at negligible computational cost. In addition, the schemes are shown to be energy stable, and their mesh-free nature and predetermined hidden layers make the method applicable to complex domains and variable mobility.

Artificial Intelligence (AI) Use Disclosure Statement

In the preparation of this dissertation, the following Artificial Intelligence (AI) tools were used: ChatGPT (OpenAI) and Claude (Anthropic). These tools were used primarily to assist with troubleshooting LaTeX errors, bibliography formatting, and proofreading for clarity and consistency. The author acknowledges full responsibility for the intellectual content of this work and has ensured that all AI-assisted sections have been reviewed and revised for accuracy and appropriate academic style. All AI-generated content was reviewed and validated for relevance, appropriateness, and accuracy before incorporation into the final document to maintain scholarly integrity of this research.

Digital Accessibility Disclosure Statement

In the preparation of this dissertation, the following digital accessibility tools were used to ensure this document complies with federal requirements: Adobe Acrobat accessibility tools, LuaLaTeX with PDF/UA-1 tagging, and manual review of headings, figures, tables, captions, and alternative text. The author acknowledges full responsibility for the intellectual content of this work and has made a good faith effort to comply with digital accessibility requirements in publishing, wherein the nature of the content does not significantly change in order to do so. Furthermore, all content has been reviewed and revised to meet these requirements prior to final publication.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Thi-Thao-Phuong Hoang, for her guidance and support throughout my Ph.D. journey. From the very beginning, she provided me with carefully selected materials to initiate my research, always welcomed me to her office with any question I had, and spent countless hours discussing how to approach challenging problems. I am truly grateful for the many opportunities she gave me to attend conferences and present my work through talks and poster presentations, allowing me to expand my professional network and engage with researchers from diverse institutions.

I am also especially thankful to Prof. Lili Ju from the University of South Carolina, whose profound knowledge and passion for research have always been a great source of inspiration throughout my doctoral studies. Despite the geographical distance, we maintained regular meetings, and his insightful ideas, thoughtful suggestions, and timely feedback have had a significant impact on both my research and this dissertation.

I sincerely thank my Ph.D. committee members, Prof. Yanzhao Cao, Prof. Junshan Lin, and Prof. Hans-Werner van Wyk, as well as my university reader, Prof. Tham Hoang, for their valuable time, insightful questions, and constructive comments, which greatly improved this dissertation. I gratefully acknowledge the U.S. National Science Foundation for partially supporting this research under Grant No. DMS-2041884.

I would also like to thank my colleagues and friends at Auburn University, especially Toan Huynh, Nguyen Huynh, Hien Tran, Yi Liu, Pengjun Wang, Li Zhou, Hewan Shemtaga, Chinedu Eleh, Zulaihat Hassan, Ian Ruau, Yuzhuo Tao, Ridvan Ozdemir, Hyelim Jung, Franklin Wang, Risheek Rahul Khanna, and many others. Your friendship and support have made my time at Auburn truly memorable.

Finally, I am deeply grateful to my parents, Mr. Minh Quan Doan and Ms. Thi Lan Le, and to my sisters, Ms. Thi Nguyet Hang Doan, Ms. Thi Nhu Nguyet Doan, and Ms. Thi Hai Yen Doan, for

their unwavering love, encouragement, and support throughout this journey. They have always believed in me, stood by my side through life's greatest challenges, and taught me invaluable lessons that no classroom could ever provide. Without their love and sacrifices, this achievement would not have been possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The numerical solution of nonlinear evolution equations and multiphysics systems is a central topic in computational mathematics, with broad applications ranging from materials science and fluid dynamics to engineering design. Many such problems possess intrinsic physical properties, including energy dissipation laws, maximum bound principles, and mass conservation. A fundamental challenge in the design of numerical methods for these problems is to develop time integration schemes that are accurate, efficient, and structure-preserving at the discrete level. This dissertation focuses on the development and analysis of four novel classes of structure-preserving time-stepping methods: (i) low regularity integrators for phase field models, (ii) noniterative localized exponential time differencing methods for hyperbolic conservation laws, (iii) dynamically regularized Lagrange multiplier methods for incompressible Navier-Stokes equations and the coupled Cahn-Hilliard-Navier-Stokes system, and (iv) generalized transferable neural networks for the Cahn-Hilliard equation.

Phase-field modeling is an important mathematical tool for solving interfacial problems in scientific and engineering fields. Phase-field models were first introduced by Fix [60] and Langer [109], and have gained increasing interest and attention in solidification dynamics [10] and other areas such as fracture mechanics [11, 12], viscous fingering [61, 140], hydrogen embrittlement [136, 92], vesicle dynamics [9, 132], etc. A thorough review on phase-field modeling can be found in [54]. The Allen-Cahn equation, originally introduced by Allen and Cahn in [4], is a well-known phase-field model used to describe the motion of antiphase boundaries in crystalline solids. It belongs to a class of semilinear parabolic equations whose solutions satisfy two important physical properties: the maximum bound principle (MBP) and the energy dissipation law. These properties are highly desired to be preserved when constructing numerical schemes.

Various MBP-preserving schemes for semilinear parabolic equations have been discovered recently. In [174], it was shown that implicit-explicit discretization in time and central finite difference in space for the Allen-Cahn equation preserve the MBP. The generalized Allen-Cahn equation with a nonlinear degenerated mobility was studied in [161], where a first-order semi-implicit scheme was proved to preserve the MBP under some time step constraint. Following these ideas, both linear and nonlinear second-order Crank-Nicolson/Adams-Bashforth schemes were constructed in [90, 89], which guarantee the MBP and energy stability. When solving problems with strong stiff linear terms, exponential-type integrators such as exponential time differencing (ETD) and integrating factor Runge-Kutta (IFRK) methods have been shown to be more effective. In [56], a general framework of ETD schemes for a class of semilinear parabolic equations was established, where linear and nonlinear operators have to satisfy certain conditions to preserve the MBP. ETD methods were also used for the nonlocal Allen-Cahn equation [55] and the conservative Allen-Cahn equation [114]. A combination of ETD with overlapping domain decomposition was studied in [115]. Unlike ETD, the main idea of IFRK methods is to use the integrating factor to eliminate the stiff linear term before applying the conventional Runge-Kutta method. A four-stage, fourth-order IFRK scheme was proposed in [101] which preserves the MBP under some certain condition on the time step size. By adding a stabilizing constant, the stabilized IFRK method was introduced in [117] and shown to unconditionally preserve the MBP up to the third order.

Energy stable numerical methods for Allen-Cahn type equations as well as other gradient flows have also attracted much attention. Examples of traditional approaches include fully implicit methods [58, 191], convex splitting schemes [183, 162, 70, 186], stabilized semi-implicit methods [166, 174, 59, 192], and ETD schemes [102, 207, 103, 56, 55]. Some novel linear schemes have been investigated in the past few years such as the invariant energy quadratization (IEQ) schemes [194, 193, 199] and the scalar auxiliary variable (SAV) schemes [26, 163, 165], which monotonically decrease certain modified energies.

It should be noted that convergence and error estimates play an essential role when analyzing a numerical scheme. In particular, ETD schemes [56, 55] require the exact space-discrete solution to be in $\mathcal{C}^1$ for the first-order scheme and in $\mathcal{C}^2$ for the second-order scheme. Similarly, convergence of the IFRK schemes [101, 117] only holds when the exact solution at least belongs to $\mathcal{C}^p$ in time, where p is the order of the method. Therefore, it is desirable to develop time integration schemes which only require low regularity on the exact solution but hold similar optimal error estimates, while still preserving the MBP and energy stability. The low regularity integrators (LRIs) are recently introduced time discretization techniques for evolution problems, which require weaker regularity assumptions for error analysis than those by classical methods. The so-called resonance based schemes were developed for nonlinear dispersive equations at low regularity, including the Korteweg-de Vries (KdV) [88, 188, 189], Schrödinger [145, 144, 146], Boussinesq [147], and Dirac [156] equations. These schemes are based on Fourier series expansions of the solution (instead of Taylor expansions as in classical methods), thus they are restricted to periodic boundary conditions only. Recently, a new framework of LRIs has been introduced in [154] which does not rely on Fourier expansions, thus can be applied to general boundary conditions. The LRIs can be coupled with various spatial discretization techniques, such as finite element methods as considered in [120], where a first-order semi-implicit LRI scheme was applied to solve the Navier-Stokes equations.

In addition to phase field models, exponential integrators can also be applied to hyperbolic conservation laws, where efficient time integration remains a significant challenge. Exponential integrators, first introduced in 1960 in [20], require no nonlinear solvers and allow large time step sizes without affecting numerical stability. The methods are derived by linearizing the original system at a certain state and applying the variation-of-constants formula to solve exactly the stiff linear part. The main cost lies in the computation of the products of matrix exponentials and vectors. With the improvements in matrix exponential algorithms based on Krylov subspace approximations [155, 84, 173, 142, 64], there has been great interest in developing exponential integrator-based methods for various problems [85, 38, 141, 176, 87, 86, 134, 180, 100]. Among them, Exponential Time Differencing Runge-Kutta methods (ETDRK) have been successfully applied to a large class of semilinear parabolic problems [56]. To accelerate ETDRK methods for large-scale problems, domain decomposition techniques have been applied [205, 83, 115] to reduce the size of the problem and compute matrix exponentials locally, thus much less expensively. Despite considerable progress in ETDRK methods for parabolic equations, their applications to hyperbolic conservation laws are still limited [6, 119, 39].

Hyperbolic conservation laws, a fundamental class of partial differential equations, play an essential role in a wide range of applications, from modeling water flows, traffic flows to gas dynamics and supply chains. As the solution may exhibit shocks or discontinuities even from smooth initial data, various spatial discretization methods have been developed, including ENO/WENO schemes [78, 172, 125, 97, 98] and the discontinuous Galerkin (DG) method [32, 36, 35, 37, 82]. We refer to [34, 139, 170] for a thorough review of different high-order schemes for hyperbolic conservation laws. Regarding time discretization, most existing methods use Strong Stability Preserving Runge-Kutta (SSP-RK) schemes [171, 172, 68], where the time step sizes must satisfy restrictive CFL conditions. Recent studies of ETDRK schemes for oceanic/atmospheric models [150, 17] confirmed their computational efficiency with large time step sizes. ETDRK methods were also studied in [113, 138] with overlapping domain decomposition; however, the overlap size must be sufficiently large to use large time step sizes.

Another main theme of this dissertation concerns the development of energy stable numerical schemes for incompressible fluid flows. The Navier-Stokes (NS) equations are fundamental in computational fluid dynamics, governing the behavior of viscous fluid flows. In the absence of external forces, the NS equations satisfy the energy dissipation law, and it has been highly desirable to develop numerical schemes that preserve this key property. The main difficulty is the treatment of the nonlinear convection term. Fully implicit schemes [135, 178] require solving nonlinear systems, which could be computationally expensive. Semi-implicit schemes [184, 177, 79, 120, 63], on the other hand, solve only linear systems with either constant or variable coefficients, depending on how the convection term is treated. Due to their computational efficiency, we focus on semi-implicit schemes with fully explicit convection. The SAV schemes have recently been developed for gradient flow problems [164, 165, 2, 26] and extended to the NS equations [121, 123, 124]. In [121], an auxiliary variable associated with the kinetic energy was introduced to reformulate the NS equations into an equivalent form with an extra dynamic equation for the auxiliary variable. The resulting schemes are unconditionally energy diminishing, marking the first time that numerical schemes with explicit convection can preserve this property. Besides the SAV approach, the Lagrange multiplier technique also leads to unconditionally energy stable schemes [25, 197], where intrinsic physical properties, such as energy dissipation laws, are incorporated into a reformulated system to preserve energy stability in terms of the original variables. However, both SAV and Lagrange multiplier approaches usually pose three challenges: (i) solving nonlinear algebraic equations for the Lagrange multiplier may produce multiple real or even complex solutions; (ii) the time step size must be sufficiently small to maintain accuracy; (iii) most stability results from the SAV methods concern a modified energy, which is not directly related to the original energy.

The Cahn-Hilliard-Navier-Stokes (CH-NS) system [75, 5, 129, 105, 106] is a well-known phase field model for the mixture of two immiscible and incompressible Newtonian fluids with matched density. When the density difference is small, the model can also be applied with a Boussinesq approximation [158]. Numerical approximations for the nonlinear, coupled CH-NS system pose several challenges. First, to handle the stiffness of the phase equation associated with the interfacial width, numerical schemes are highly desired to be unconditionally stable. Second, the coupling of nonlinear equations introduces significant computational cost, particularly with implicit methods. Third, numerical solutions must preserve key physical principles, such as mass conservation and energy dissipation. During the past two decades, many studies have been carried out to develop and analyze unconditionally energy stable schemes for the CH-NS equations. Popular methods include the convex splitting methods [77, 44, 28, 16, 73], stabilized semi-implicit methods [167, 169, 30, 15], and SAV schemes [123, 124, 200, 131, 118]. Convex splitting schemes guarantee the unique solvability of the numerical solution; however, a nonlinear equation needs to be solved at each time step. Stabilized schemes, on the other hand, introduce artificial numerical dissipation to the system which potentially changes the system dynamics. An issue with SAV-based schemes is that the stability results often apply to a modified energy, which is not directly related or close to the original energy when the time step size is not small enough. We refer to [69, 148] for other energy stable schemes for the CH-NS system, and to [128] for a bound-preserving limiter-based algorithm that employs discontinuous Galerkin discretization in space along with a combination of convex splitting and rotational pressure-correction techniques in time.

In addition to traditional numerical methods, deep learning has recently emerged as a powerful tool for scientific computing, and neural network-based PDE solvers have attracted intense research interest. Physics-informed neural networks (PINNs) [153] embed the governing equations directly into the loss function and train the network via stochastic gradient descent. While PINNs offer the attractive feature of being mesh-free, their application to phase-field equations such as the Cahn-Hilliard equation has proven particularly challenging due to sharp transition layers and the inability of fixed collocation points to adapt to moving features [185, 137, 23]. An alternative paradigm is the class of shallow neural network methods with predetermined hidden-layer parameters, including the Extreme Learning Machine [91, 53], the Random Feature Method [21], and the Transferable Neural Network (TransNet) [201]. In these approaches, the weights and biases of the hidden-layer neurons are fixed in advance, and only the weights of the output layer are optimized, typically by solving a linear least-squares problem. This strategy completely eliminates the need for stochastic gradient descent training, resulting in orders-of-magnitude speedups over deep-network-based solvers. Recently, the Generalized Transferable Neural Network (GTransNet) was proposed in [24] to address the limitation of TransNet in handling problems with highly oscillatory solutions, demonstrating superior performance for a broad class of steady-state PDEs. However, its application to time-dependent fourth-order problems with structure preservation remains unexplored.

The main objective of this dissertation is to overcome the aforementioned difficulties by developing novel, efficient, and structure-preserving time integration methods with rigorous mathematical analysis. Specifically, we address the following four problems:

1. For phase field models of Allen-Cahn type, we construct first- and second-order LRI schemes by repeatedly applying Duhamel's principle to the space-discrete system. The proposed schemes preserve the conditional MBP and energy stability, and their temporal error estimates are derived under the low regularity assumption that the exact solution is only continuous in time, in contrast to the $\mathcal{C}^p$ regularity required by classical ETD and IFRK methods. We also carry out optimal fully discrete error analysis of the first-order LRI schemes and extend the LRI framework to the conservative Allen-Cahn equation with a nonlocal Lagrange multiplier, where the schemes additionally conserve mass while maintaining the MBP and energy stability. These results are presented in Chapter 2.

2. Regarding hyperbolic conservation laws, we combine ETDRK schemes with the DG method and nonoverlapping domain decomposition to develop global and localized ETD methods that significantly relax the CFL conditions imposed by SSP-RK schemes. The localized scheme requires no iteration between subdomains while enabling the same large time step sizes as the global method. Rigorous semi-discrete temporal error analysis is carried out for both schemes. This is the subject of Chapter 3.

3. For the incompressible NS equations and the coupled CH-NS system, we develop the dynamically regularized Lagrange multiplier (DRLM) method. By introducing a dynamic equation involving the kinetic energy, a Lagrange multiplier, and a regularization parameter, we reformulate the original equations into equivalent systems. First- and second-order DRLM schemes are derived and shown to be unconditionally energy stable with respect to the original variables. Optimal error estimates are established for the first-order DRLM scheme based on a uniform bound on the Lagrange multiplier and mathematical induction. The DRLM framework is further extended to the CH-NS system, where the schemes are fully decoupled, mass-conserving, and require no stabilization terms. The details can be found in Chapter 4.

4. For the Cahn-Hilliard equation, we develop a structure-preserving neural network-based

framework that combines GTransNet for spatial approximation with stabilized BDF methods for time discretization. The resulting GTransNet-BDF schemes preserve mass conservation and energy stability at the time-discrete level. Since the collocation-based spatial implementation produces a nonzero least-squares residual that violates mass conservation, a novel post-processing mass-conserving projection is introduced to enforce the mass constraint through a minimization problem with negligible computational cost. The mesh-free nature and predetermined hidden layers make the method applicable to complex domains, variable mobility, and long-time simulations. These results are presented in Chapter 5.

Finally, Chapter 6 summarizes the main results and discusses directions for future research.

Chapter 2

Low Regularity Integrators for Phase Field Models

This chapter is devoted to the construction, analysis, and numerical validation of low regularity integrators (LRIs) for phase field models of Allen-Cahn type. The chapter consists of three parts. In the first part, we construct first- and second-order LRI schemes for the Allen-Cahn equation by iteratively applying Duhamel's formula to the space-discrete system, prove that the proposed schemes preserve the maximum bound principle (MBP) and energy stability, and derive temporal error estimates under the minimal assumption that the exact solution is only continuous in time. In the second part, we carry out fully discrete error analysis of the first-order LRI schemes and establish optimal error estimates in both space and time. In the third part, we extend the LRI framework to the conservative Allen-Cahn equation with a nonlocal Lagrange multiplier, where the schemes additionally conserve mass unconditionally while satisfying the MBP and remaining energy stable. Numerical experiments confirm the theoretical findings and demonstrate the outperformance of the proposed LRI schemes compared to classical exponential time differencing (ETD) schemes, particularly when the interfacial parameter is close to zero. The material in this chapter is based on [45, 46, 47].

2.1 Low regularity integrators for the Allen-Cahn equation

We consider the Allen-Cahn equation [4], a well-known phase-field model that describes the motion of antiphase boundaries in crystalline solids. It belongs to a class of semilinear parabolic equations of the form

$$\partial_t u = \varepsilon^2 \Delta u + f(u), \quad \mathbf{x} \in \Omega,\ t > 0, \tag{2.1}$$

where $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ is an open and bounded Lipschitz domain, $u : \overline{\Omega} \times [0, \infty) \to \mathbb{R}$ is the unknown function, $\varepsilon > 0$ is the interfacial parameter, and $f : \mathbb{R} \to \mathbb{R}$ is a nonlinear function. An initial boundary value problem associated with equation (2.1) is obtained by imposing the initial condition $u(\mathbf{x},0) = u_0(\mathbf{x})$ for any $\mathbf{x} \in \overline{\Omega}$, and suitable boundary conditions (e.g., periodic or homogeneous Neumann or Dirichlet boundary conditions). Throughout this section, we suppose that the nonlinear function f satisfies the following assumptions:

(A1) $f(\beta) \le 0 \le f(-\beta)$ for some $\beta > 0$.

(A2) f is continuously differentiable and nonconstant on $[-\beta, \beta]$.

Then the solution to (2.1) satisfies the maximum bound principle (MBP) [56]

$$\max_{\mathbf{x}\in\overline{\Omega}} |u_0(\mathbf{x})| \le \beta \quad \Longrightarrow \quad \max_{\mathbf{x}\in\overline{\Omega}} |u(\mathbf{x},t)| \le \beta \quad \text{for all } \ t > 0,$$

and energy dissipation law

$$\frac{d}{dt}E(u(t)) = -\int_\Omega |\partial_t u(\mathbf{x},t)|^2 \, d\mathbf{x} \le 0, \ \text{ where } \ E(u) = \int_\Omega \left(\frac{\varepsilon^2}{2}|\nabla u(\mathbf{x})|^2 + F(u(\mathbf{x}))\right) d\mathbf{x}.$$

2.1.1 Space-discrete problem and low regularity integrators

Assume that the spatial domain Ω in $\mathbb{R}^d$ is either a finite interval ($d = 1$), a rectangle ($d = 2$), or a rectangular parallelepiped ($d = 3$). We consider a uniform partition of Ω into rectangular elements; for simplicity, we assume all elements have equal sides and denote by h the length of the side of each element. Let N be the number of grid points of the mesh for unknowns, and denote by

$$\mathbf{u}(t) = (u^1(t), u^2(t), \dots, u^N(t))^T,$$

where $u^j(t)$ is the approximation of $u(\mathbf{x}_j, t)$ for $j = 1, 2, \dots, N$. Using central finite differences for spatial discretization of problem (2.1) with the homogeneous Neumann boundary conditions, we obtain the following space-discrete system:

$$\frac{d\mathbf{u}}{dt} = \mathbf{A}\mathbf{u} + \mathbf{f}(\mathbf{u}), \tag{2.2}$$

where $\mathbf{f}(\mathbf{u}) = (f(u^1), f(u^2), \dots, f(u^N))^T$, $\mathbf{A} = \varepsilon^2 \mathbf{D}_h$, and $\mathbf{D}_h$ is defined as in [104]. Particularly, $\mathbf{D}_h = \lambda_h$ in one dimension, $\mathbf{D}_h = \mathbf{I} \otimes \lambda_h + \lambda_h \otimes \mathbf{I}$ in two dimensions, and $\mathbf{D}_h = \mathbf{I} \otimes \mathbf{I} \otimes \lambda_h +$

$\mathbf{I} \otimes \lambda_h \otimes \mathbf{I} + \lambda_h \otimes \mathbf{I} \otimes I$ in three dimensions. Here $\otimes$ denotes the Kronecker product, $\mathbf{I}$ is the identity matrix, and λ_h is a square matrix of order N given by

$$\lambda_h := \frac{1}{h^2} \begin{bmatrix} -2 & 2 & 0 & \dots & 0 \\ 1 & -2 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & -2 & 1 \\ 0 & \dots & 0 & 2 & -2 \end{bmatrix}.$$

It is well-known that matrix $\mathbf{D}_h$ is diagonally dominant with all diagonal entries negative. We then have the following result [101, Lemma 3.3], which is crucial to the analysis of exponential-type integrators:

Lemma 2.1. For any $\gamma > 0$, we have $\|e^{\gamma \mathbf{D}_h}\|_\infty \leq 1$, and therefore $\|e^{\gamma \mathbf{A}}\|_\infty \leq 1$.

Based on this result, the space-discrete problem (2.2) is well-posed and preserves the MBP as stated below (the proof is similar to Theorem 2 in [115]).

Theorem 2.2. Given any fixed terminal time $T > 0$. Assume that the initial data is bounded by β in absolute value, i.e., $\|u_0\|_{L^\infty} \leq \beta$, then the semi-discrete system (2.2) admits a unique solution $\mathbf{u} \in \mathcal{C}([0,T]; \mathbb{R}^N)$ and

$$\|\mathbf{u}(t)\|_\infty \leq \beta \quad \text{for all} \ \ t \in [0,T].$$

Remark 2.3. The result of Theorem 2.2 still holds when one imposes the periodic boundary condition or the Dirichlet boundary condition (where the Dirichlet data is also assumed to be bounded by β in absolute value); see [56] for more details. In addition, other types of spatial discretizations (e.g., finite volume and lumped-mass finite element) for problem (2.1) can also be considered without affecting the analysis presented in this section, provided that the corresponding matrix $\mathbf{A} = (a_{ij})_{N\times N}$ satisfies the following property [56, Proposition 2.5]:

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{N} |a_{ij}|, \quad a_{ii} < 0, \quad a_{ij} \geq 0 \ (j \neq i). \tag{2.3}$$

A slightly more relaxed condition on $\mathbf{A}$ for Lemma 1 to hold can be further found in [101, Lemma 3.3].

For the time discretization, let us consider a uniform partition of the time interval $[0,T]$: $0 = t_0 < t_1 < \cdots < t_M = T$, with the step size $\Delta t = \frac{T}{M}$. The exact (in time) solution to (2.2) at each time level is given by Duhamel's formula:

$$\mathbf{u}(t_{m+1}) = e^{\Delta t \mathbf{A}}\mathbf{u}(t_m) + \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m + s))\, ds, \quad m = 0, 1, \dots, M-1. \tag{2.4}$$

The main idea of LRIs developed in [154, 120] is to use again Duhamel's formula to compute $\mathbf{u}(t_m + s)$ in (2.4). In particular, we have, for any $s \in [0, \Delta t]$:

$$\mathbf{u}(t_m + s) = e^{s\mathbf{A}}\mathbf{u}(t_m) + \int_0^{s} e^{(s-\sigma)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m + \sigma))\, d\sigma. \tag{2.5}$$

Using (2.5) to approximate $\mathbf{u}(t_m + s)$ in (2.4) leads to the first- and second-order LRIs described below.

For convenience of presentation, we introduce the following notation: for any function $g : \mathbb{R} \to \mathbb{R}$, let $\mathbf{g} : \mathbb{R}^N \to \mathbb{R}^N$ be defined as $\mathbf{g}(\mathbf{v}) := (g(v_1), g(v_2), \dots, g(v_N))^T$, for any $\mathbf{v} = (v_1, v_2, \dots, v_N)^T \in \mathbb{R}^N$; when g is sufficiently smooth, we denote by

$$\frac{\partial^p \mathbf{g}}{\partial \mathbf{v}^p}(\mathbf{v}) := \mathrm{diag}(g^{(p)}(v_1), g^{(p)}(v_2), \dots, g^{(p)}(v_N)), \quad p \in \mathbb{N}. \tag{2.6}$$

Note that when $p = 1$, (2.6) is simply the Jacobian of $\mathbf{g}$. In addition, we make use of the notation $\odot$ for denoting element-wise operations, e.g., $\mathbf{v}^{\odot 2} := (v_1^2, v_2^2, \dots, v_N^2)^T$ and $\mathbf{u} \odot \mathbf{v} := (u_1 v_1, u_2 v_2, \dots, u_N v_N)^T$.

First-order low regularity integrators Let us take $\mathbf{U}_0 = \mathbf{u}(0)$. For first-order schemes, we approximate $\mathbf{u}(t_m + s)$ by $e^{s\mathbf{A}}\mathbf{U}_m$ in (2.4) and obtain: for $m = 0, 1, \dots, M-1$,

$$\mathbf{U}_{m+1} = e^{\Delta t \mathbf{A}}\mathbf{U}_m + \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}}\mathbf{f}(e^{s\mathbf{A}}\mathbf{U}_m)\, ds = e^{\Delta t \mathbf{A}}\mathbf{U}_m + \int_0^{\Delta t} \mathbf{g}(s)\, ds, \tag{2.7}$$

where $\mathbf{g}(s) = e^{(\Delta t - s)\mathbf{A}}\mathbf{f}(e^{s\mathbf{A}}\mathbf{U}_m)$. We then compute the integral on the right-hand side using one-point quadrature rules, which leads to the following schemes:

i) **LRI1a** with $\mathbf{g}(s)$ on $[0, \Delta t]$ approximated by $\mathbf{g}(0)$ (as in [154]):

$$\mathbf{U}_{m+1} = e^{\Delta t \mathbf{A}}(\mathbf{U}_m + \Delta t \mathbf{f}(\mathbf{U}_m)). \tag{2.8}$$

ii) **LRI1b** with $\mathbf{g}(s)$ on $[0, \Delta t]$ approximated by $\mathbf{g}(\Delta t)$:

$$\mathbf{U}_{m+1} = e^{\Delta t \mathbf{A}}\mathbf{U}_m + \Delta t \mathbf{f}(e^{\Delta t \mathbf{A}}\mathbf{U}_m). \tag{2.9}$$

We remark that the LRI1a scheme is identical to the first-order IFRK scheme [94, 101], while the LRI1b scheme (first proposed in [120]) is essentially different from the IFRK or ETD schemes.

Second-order low regularity integrator Let us approximate $\mathbf{u}(t_m + s)$ by $e^{s\mathbf{A}}\mathbf{U}_m + s\mathbf{f}(\mathbf{U}_m)$ from (2.5), and plug that into (2.4) to obtain: for $m = 0, 1, \dots, M-1$,

$$\mathbf{U}_{m+1} = e^{\Delta t \mathbf{A}}\mathbf{U}_m + \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}}\mathbf{f}(e^{s\mathbf{A}}\mathbf{U}_m + s\mathbf{f}(\mathbf{U}_m))\, ds.$$

By replacing $\mathbf{f}(e^{s\mathbf{A}}\mathbf{U}_m + s\mathbf{f}(\mathbf{U}_m))$ by $\mathbf{f}(e^{s\mathbf{A}}\mathbf{U}_m) + s\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m)$, we arrive at

$$\mathbf{U}_{m+1} = e^{\Delta t \mathbf{A}}\mathbf{U}_m + \int_0^{\Delta t} \mathbf{g}(s)\, ds + \int_0^{\Delta t} s e^{(\Delta t - s)\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m)\, ds.$$

Finally, the LRI2 scheme is defined by computing the integral $\int_0^{\Delta t} \mathbf{g}(s)\, ds$ using the trapezoid rule and approximating $e^{(\Delta t - s)\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m)$ by $e^{\Delta t \mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m)$ in the above equation:

$$\mathbf{U}_{m+1} = e^{\Delta t \mathbf{A}}\mathbf{U}_m + \frac{\Delta t}{2}\left[e^{\Delta t \mathbf{A}}\mathbf{f}(\mathbf{U}_m) + \mathbf{f}(e^{\Delta t \mathbf{A}}\mathbf{U}_m)\right] + \frac{\Delta t^2}{2}e^{\Delta t \mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m). \tag{2.10}$$

2.1.2 Properties of the fully discrete numerical solutions

Discrete maximum bound principle

We shall prove that the LRI1 and LRI2 schemes proposed in the previous section satisfy the MBP in the discrete sense. We first state some preliminary result needed for the MBP proof. It is easy to see from assumptions (A1) and (A2) that $\min\limits_{|x| \le \beta} f'(x) < 0$. Define $\omega_0 := -\frac{1}{\min\limits_{|x|\le\beta} f'(x)} > 0$.

Lemma 2.4. It holds that

$$|x + \omega f(x)| \le \beta \quad \text{for all } \; x \in [-\beta, \beta] \; \text{ and } \; \omega \in (0, \omega_0].$$

Proof. Let $f_0(x) = x + \omega f(x)$, we have $f_0'(x) = 1 + \omega f'(x) \ge 0$, for any $\omega \in (0, \omega_0]$ and $x \in [-\beta, \beta]$. Moreover, as $f(\beta) \le 0 \le f(-\beta)$, we deduce that

$$-\beta \le f_0(-\beta) \le f_0(x) \le f_0(\beta) \le \beta \quad \text{for all } \; x \in [-\beta, \beta],$$

which concludes the proof. □

Both LRI1a and LRI1b schemes preserve the discrete MBP under certain time step constraint as stated below.

Theorem 2.5. Suppose that $\|u_0\|_{L^\infty} \le \beta$, then the numerical solution $\{\mathbf{U}_m\}$ generated by the LRI1a scheme (cf. Equation (2.8)) or the LRI1b scheme (cf. Equation (2.9)) is also pointwisely bounded in the absolute value by β, i.e.,

$$\|\mathbf{U}_m\|_\infty \le \beta, \quad m = 0, 1, \dots, M,$$

provided that $0 < \Delta t \le \omega_0$.

Proof. By induction, we only need to prove: if $\|\mathbf{U}_m\|_\infty \le \beta$, then it also holds that $\|\mathbf{U}_{m+1}\|_\infty \le \beta$. For the LRI1a scheme, by applying Lemmas 2.1 and 2.4, we deduce from (2.8) that

$$\|\mathbf{U}_{m+1}\|_\infty \le \|e^{\Delta t \mathbf{A}}\|_\infty \|\mathbf{U}_m + \Delta t \mathbf{f}(\mathbf{U}_m)\|_\infty \le \beta \quad \text{for all } \; \Delta t \in (0, \omega_0].$$

For the LRI1b scheme, as $\|e^{\Delta t \mathbf{A}} \mathbf{U}_m\|_\infty \le \beta$ (Lemma 2.1), together with Lemma 2.4 we obtain

$$\|\mathbf{U}_{m+1}\|_\infty = \|e^{\Delta t \mathbf{A}} \mathbf{U}_m + \Delta t \mathbf{f}(e^{\Delta t \mathbf{A}} \mathbf{U}_m)\|_\infty \le \beta \quad \text{for all } \; \Delta t \in (0, \omega_0].$$

The proof is thus completed. □

Unlike the LRI1 schemes, the LRI2 scheme requires a stronger condition on f to obtain the discrete MBP, namely:

(A3) *f is twice continuously differentiable and nonconstant on* $[-\beta,\beta]$.

The MBP preservation of the LRI2 scheme is guaranteed by the following theorem. Define

$$\omega_1 := -\min_{|x|\le\beta}[f''(x)f(x)].$$

Theorem 2.6. Suppose that f satisfies (A1) and (A3). If $\|u_0\|_{L^\infty}\le\beta$, then the numerical solution $\{\mathbf{U}_m\}$ generated by the LRI2 scheme (cf. Equation (2.10)) is also pointwisely bounded in the absolute value by β, i.e.,

$$\|\mathbf{U}_m\|_\infty\le\beta,\qquad m=0,1,\dots,M,$$

provided that $0<\Delta t\le\delta_0\omega_0$, where δ_0 is given by

$$\delta_0=\begin{cases}\min\{1,\delta\}, & \text{if } \omega_1>0,\\ 1, & \text{if } \omega_1\le 0,\end{cases}$$

with $\delta=\frac{-1+\sqrt{1+7\omega_0^2\omega_1}}{2\omega_0^2\omega_1}$ if $\omega_1>0$.

Proof. By induction, we again only need to show that if $\|\mathbf{U}_m\|_\infty\le\beta$ then $\|\mathbf{U}_{m+1}\|_\infty\le\beta$. Apply Lemma 2.1, for any $\delta_1\in(0,2)$ we have

$$\begin{aligned}
&\|\mathbf{U}_{m+1}\|_\infty\\
&\le\left\|\frac{\delta_1}{2}e^{\Delta t\mathbf{A}}\mathbf{U}_m+\frac{\Delta t}{2}\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{U}_m)\right\|_\infty+\left\|\frac{2-\delta_1}{2}\mathbf{e}^{\Delta t\mathbf{A}}U_m+\frac{\Delta t}{2}e^{\Delta t\mathbf{A}}\mathbf{f}(\mathbf{U}_m)+\frac{\Delta t^2}{2}e^{\Delta t\mathbf{A}}\frac{\partial\mathbf{f}}{\partial\mathbf{u}}(\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m)\right\|_\infty\\
&\le\frac{\delta_1}{2}\left\|e^{\Delta t\mathbf{A}}\mathbf{U}_m+\frac{\Delta t}{\delta_1}\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{U}_m)\right\|_\infty+\frac{1}{2}\left\|(2-\delta_1)\mathbf{U}_m+\Delta t\mathbf{f}(\mathbf{U}_m)+\Delta t^2\frac{\partial\mathbf{f}}{\partial\mathbf{u}}(\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m)\right\|_\infty.
\end{aligned}$$

Notice that $\|e^{\Delta t\mathbf{A}}\mathbf{U}_m\|_\infty\le\beta$. If $\frac{\Delta t}{\delta_1}\in(0,\omega_0]$ (i.e., $\Delta t\in(0,\delta_1\omega_0]$), then we observe by Lemma 2.4 that

$$\left\|e^{\Delta t\mathbf{A}}\mathbf{U}_m+\frac{\Delta t}{\delta_1}\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{U}_m)\right\|_\infty\le\beta.$$

Define $K(x)=(2-\delta_1)x+\Delta tf(x)+\Delta t^2f'(x)f(x)$ for $x\in[-\beta,\beta]$. Rewrite $K(x)$ as follows:

$$K(x)=(2-\delta_1)x+(1+\Delta tf'(x))\Delta tf(x).$$

If $\Delta t \in (0, \omega_0]$, one can apply the property of $f_0'(x)$ in the proof of Lemma 2.4 to obtain

$$-(2-\delta_1)\beta \le K(-\beta), \text{ and } K(\beta) \le (2-\delta_1)\beta. \tag{2.11}$$

On the other hand,

$$\begin{aligned}
K'(x) &= 2 - \delta_1 + \Delta t f'(x) + \Delta t^2 [f'(x)]^2 + \Delta t^2 f''(x) f(x) \\
&= \left(\frac{1}{2} + \Delta t f'(x)\right)^2 + \frac{7}{4} - \delta_1 + \Delta t^2 f''(x) f(x) \\
&\ge \frac{7}{4} - \delta_1 - \Delta t^2 \omega_1 \quad \text{for all } x \in [-\beta, \beta].
\end{aligned}$$

If $\delta_1 < \frac{7}{4}$, and $\Delta t \le \sqrt{\frac{1}{\omega_1}\left(\frac{7}{4} - \delta_1\right)}$ in case of $\omega_1 > 0$, then we have

$$K'(x) \ge 0 \quad \text{for all } x \in [-\beta, \beta].$$

Therefore, K is non-decreasing on $[-\beta, \beta]$. This, combines with (2.11), gives us

$$|K(x)| \le (2-\delta_1)\beta \quad \text{for all } x \in [-\beta, \beta].$$

Thus,

$$\|\mathbf{U}_{m+1}\|_\infty \le \frac{\delta_1}{2}\beta + \frac{2-\delta_1}{2}\beta = \beta, \tag{2.12}$$

provided that the following conditions are all satisfied

$$\begin{aligned}
&\delta_1 \in \left(0, \frac{7}{4}\right), \quad \Delta t \in (0, \delta_1 \omega_0], \quad \Delta t \in (0, \omega_0], \\
&\Delta t \le \sqrt{\frac{1}{\omega_1}\left(\frac{7}{4} - \delta_1\right)} \text{ if } \omega_1 > 0.
\end{aligned}$$

If $\omega_1 \le 0$, we can choose $\delta_1 = 1$ and the condition on the time step size is $\Delta t \in (0, \omega_0]$. If $\omega_1 > 0$ then Δt has to satisfy the constraint below

$$0 < \Delta t \le \min\left\{\omega_0, \delta_1\omega_0, \sqrt{\frac{1}{\omega_1}\left(\frac{7}{4} - \delta_1\right)}\right\}, \quad \delta_1 \in \left(0, \frac{7}{4}\right),$$

or equivalently,

$$0 < \Delta t \le \omega_0 \min\left\{1, \delta_1, \sqrt{\frac{1}{\omega_0^2\omega_1}\left(\frac{7}{4} - \delta_1\right)}\right\}.$$

The best choice of the constant δ_1 in this case is $\delta_1 = \delta$ because of the the following relations:

$$\delta = \sqrt{\frac{1}{\omega_0^2\omega_1}\left(\frac{7}{4} - \delta\right)}, \quad \delta \in \left(0, \frac{7}{4}\right).$$

Hence, we require $0 < \Delta t \le \omega_0 \min\{1, \delta\}$ when $\omega_1 > 0$. Consequently, (2.12) is true for all $\Delta t \in (0, \delta_0\omega_0]$, which completes the proof of Theorem 2.6. □

Remark 2.7. If the nonlinear function f satisfies additionally $f(-\beta) = f(\beta) = 0$ for some $\beta > 0$, then (2.11) is obviously true without the requirement $0 < \Delta t \le \omega_0$. Also in this case, it is not difficult to prove that $\omega_1 > 0$, meaning that δ exists. Thus, the range of Δt can be enlarged, particularly, $0 < \Delta t \le \delta\omega_0$. This is the case for both double-well potential and Flory-Huggins potential functions [56]. The double-well potential function $F(u) = \frac{1}{4}(u^2-1)^2$ gives the nonlinear term $f(u) = u - u^3$, then $f(-1) = f(1) = 0$, i.e., $\beta = 1$. By simple calculations, we obtain

$$\omega_0 = \frac{1}{2}, \quad \omega_1 = \frac{3}{2}, \quad \delta = \frac{-4 + \sqrt{58}}{3} \approx \frac{6}{5}.$$

Therefore, the MBP is preserved when $0 < \Delta t \le \frac{1}{2}$ (for the LRI1 schemes) and $0 < \Delta t \le \frac{3}{5}$ (for the LRI2 scheme). The Flory-Huggins potential function is given by $F(u) = \frac{\theta}{2}[(1+u)\ln(1+u) + (1-u)\ln(1-u)] - \frac{\theta_c}{2}u^2$, where $0 < \theta < \theta_c$. Then $f(u) = -F'(u)$, namely, $f(u) = \frac{\theta}{2}\ln\frac{1-u}{1+u} + \theta_c u$. Consider a particular case [101] where $\theta = 0.8$ and $\theta_c = 1.6$, then $\beta \approx 0.9575$ satisfying $f(-\beta) = f(\beta) = 0$. We can easily verify that

$$\omega_0 = \frac{1-\beta^2}{\theta - \theta_c(1-\beta^2)} \approx 0.1247, \quad \omega_1 \approx -f(0.932)f''(0.932) \approx 13.1739, \quad \delta \approx 1.367.$$

Thus, the conditions on Δt to preserve the MBP are $0 < \Delta t \le 0.1247$ for the LRI1 schemes and $0 < \Delta t \le 0.1705$ for the LRI2 scheme. We shall verify these time step constraints in the numerical experiments (cf. Section 2.1.4).

Discrete energy stability

The discrete energy E_h is defined as

$$E_h(\mathbf{U}) = \sum_{i=1}^{N} F(U^i) - \frac{\varepsilon^2}{2}\mathbf{U}^T\mathbf{D}_h\mathbf{U} \quad \text{for all } \mathbf{U} = (U^1, U^2, \dots, U^N)^T \in \mathbb{R}^N, \tag{2.13}$$

where F is the potential function. In the following, we show that the first- and second-order LRI schemes are conditionally energy stable in the sense that the discrete energy is bounded at all times. We first state some preliminary results needed for the main theorem. Let us define $F_0 = \max_{|x|\le\beta} |f(x)|$ and $F_1 = \max_{|x|\le\beta} |f'(x)|$.

Lemma 2.8. If $\|\mathbf{U}_m\|_\infty \le \beta$ for all $m = 0, 1, \dots, M$, then there exists $\mathbf{W}$ in $\mathbb{R}^N$, $\|\mathbf{W}\|_\infty \le F_0$ such that

$$E_h(\mathbf{U}_{m+1}) - E_h(\mathbf{U}_m) \le (\mathbf{U}_{m+1} - \mathbf{U}_m)^T(\mathbf{W} - \mathbf{A}\mathbf{U}_{m+1}),$$

for $m = 0, 1, \dots, M-1$.

Proof. We first have the following identity from (2.13):

$$E_h(\mathbf{U}_{m+1}) - E_h(\mathbf{U}_m) = \sum_{i=1}^{N}[F(U^i_{m+1}) - F(U^i_m)] - \frac{\varepsilon^2}{2}(\mathbf{U}^T_{m+1}\mathbf{D}_h\mathbf{U}_{m+1} - \mathbf{U}^T_m\mathbf{D}_h\mathbf{U}_m). \tag{2.14}$$

Since $\|\mathbf{U}_m\|_\infty \le \beta$ for $m = 0, 1, \dots, M$, by the mean value theorem, there exist $\gamma_1, \gamma_2, \dots, \gamma_N$ such that $|\gamma_i| \le \beta$ for $i = 1, 2, \dots, N$ and

$$F(U^i_{m+1}) - F(U^i_m) = (U^i_{m+1} - U^i_m)F'(\gamma_i) = -(U^i_{m+1} - U^i_m)f(\gamma_i), \quad i = 1, 2, \dots, N.$$

Let $\mathbf{W} := -[f(\gamma_1), f(\gamma_2), \dots, f(\gamma_N)]^T$, we can see that $\|\mathbf{W}\|_\infty \le F_0$ and

$$\sum_{i=1}^{N}[F(U^i_{m+1}) - F(U^i_m)] = (\mathbf{U}_{m+1} - \mathbf{U}_m)^T\mathbf{W}. \tag{2.15}$$

On the other hand, $\mathbf{D}_h$ is negative semidefinite and symmetric, so

$$\begin{aligned}
0 &\ge (\mathbf{U}_{m+1} - \mathbf{U}_m)^T\mathbf{D}_h(\mathbf{U}_{m+1} - \mathbf{U}_m)\\
&= \mathbf{U}^T_{m+1}\mathbf{D}_h\mathbf{U}_{m+1} + \mathbf{U}^T_m\mathbf{D}_h\mathbf{U}_m - \mathbf{U}^T_{m+1}\mathbf{D}_h\mathbf{U}_m - \mathbf{U}^T_m\mathbf{D}_h\mathbf{U}_{m+1}\\
&= -\mathbf{U}^T_{m+1}\mathbf{D}_h\mathbf{U}_{m+1} + \mathbf{U}^T_m\mathbf{D}_h\mathbf{U}_m + 2(\mathbf{U}_{m+1} - \mathbf{U}_m)^T\mathbf{D}_h\mathbf{U}_{m+1}.
\end{aligned}$$

Therefore,

$$\mathbf{U}_{m+1}^T \mathbf{D}_h \mathbf{U}_{m+1} - \mathbf{U}_m^T \mathbf{D}_h \mathbf{U}_m \geq 2(\mathbf{U}_{m+1} - \mathbf{U}_m)^T \mathbf{D}_h \mathbf{U}_{m+1}. \tag{2.16}$$

From (2.14), (2.15) and (2.16), we get

$$E_h(\mathbf{U}_{m+1}) - E_h(\mathbf{U}_m) \leq (\mathbf{U}_{m+1} - \mathbf{U}_m)^T (\mathbf{W} - \mathbf{A}\mathbf{U}_{m+1}),$$

which leads to the conclusion. □

Lemma 2.9.

(i) If $0 < \Delta t \leq \omega_0$ and f satisfies (A1) and (A2), then the numerical solution $\{\mathbf{U}_m\}$ generated by the LRI1a scheme (2.8) or the LRI1b scheme (2.9) satisfies

$$\|\mathbf{U}_{m+1} - \mathbf{U}_m\|_\infty \leq (\beta \|\mathbf{A}\|_\infty + F_0)\Delta t.$$

(ii) If $0 < \Delta t \leq \delta_0 \omega_0$ and f satisfies (A1) and (A3), then the numerical solution $\{\mathbf{U}_m\}$ generated by the LRI2 scheme (2.10) satisfies

$$\|\mathbf{U}_{m+1} - \mathbf{U}_m\|_\infty \leq (\beta \|\mathbf{A}\|_\infty + F_0 + \delta_0 \omega_0 F_0 F_1)\Delta t.$$

Proof. By Theorem 2.5 or 2.6, we know that the numerical solution $\{\mathbf{U}_m\}$ generated by the schemes (2.8), (2.9) or (2.10) preserves the MBP, meaning that $\|\mathbf{U}_m\|_\infty \leq \beta$ for $m = 0, 1, \dots, M$.

(i) Consider the LRI1a scheme (2.8). Define $\mathbf{q}_1(t) = e^{t\mathbf{A}}\mathbf{U}_m + te^{\Delta t \mathbf{A}}\mathbf{f}(\mathbf{U}_m)$, then $\mathbf{q}_1(0) = \mathbf{U}_m$, $\mathbf{q}_1(\Delta t) = \mathbf{U}_{m+1}$, and

$$\mathbf{q}_1'(t) = \mathbf{A}e^{t\mathbf{A}}\mathbf{U}_m + e^{\Delta t \mathbf{A}}\mathbf{f}(\mathbf{U}_m).$$

By Lemma 2.1, we can easily see that

$$\|\mathbf{q}_1'(t)\|_\infty \leq \beta \|\mathbf{A}\|_\infty + F_0 \quad \text{for all } \ t > 0.$$

Apply the mean value theorem, we get

$$\|\mathbf{U}_{m+1} - \mathbf{U}_m\|_\infty = \|\mathbf{q}_1(\Delta t) - \mathbf{q}_1(0)\|_\infty \leq (\beta \|\mathbf{A}\|_\infty + F_0)\Delta t.$$

Similarly for the LRI1b scheme (2.9), define $\mathbf{q}_2(t) = e^{t\mathbf{A}}\mathbf{U}_m + t\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{U}_m)$, then $\mathbf{q}_2(0) = \mathbf{U}_m$ and $\mathbf{q}_2(\Delta t) = \mathbf{U}_{m+1}$. Moreover,

$$\|\mathbf{q}_2'(t)\|_\infty = \|\mathbf{A}e^{t\mathbf{A}}\mathbf{U}_m + \mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{U}_m)\|_\infty \le \beta\|\mathbf{A}\|_\infty + F_0 \quad \text{for all } \ t > 0.$$

The result of (i) then follows by the mean value theorem.

(ii) For the LRI2 scheme (2.10), define $\mathbf{q}_3(t) = \frac{\mathbf{q}_1(t)+\mathbf{q}_2(t)}{2} + \frac{t^2}{2}e^{\Delta t\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m)$, then $\mathbf{q}_3(0) = \mathbf{U}_m, \mathbf{q}_3(\Delta t) = \mathbf{U}_{m+1}$ and

$$\mathbf{q}_3'(t) = \frac{\mathbf{q}_1'(t) + \mathbf{q}_2'(t)}{2} + te^{\Delta t\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m).$$

Thus we have

$$\|\mathbf{q}_3'(t)\|_\infty \le \frac{\|\mathbf{q}_1'(t)\|_\infty + \|\mathbf{q}_2'(t)\|_\infty}{2} + \delta_0\omega_0F_0F_1 \le \beta\|\mathbf{A}\|_\infty + F_0 + \delta_0\omega_0F_0F_1 \quad \text{for all } \ t \in [0, \Delta t].$$

Then the result of (ii) follows by the mean value theorem. □

Theorem 2.10. Suppose $0 < \Delta t \le \omega_0$ (resp. $0 < \Delta t \le \delta_0\omega_0$) and f satisfies (A1) and (A2) (resp. (A1) and (A3)), then the numerical solution $\{\mathbf{U}_m\}$ generated by the LRI1a scheme (2.8) or the LRI1b scheme (2.9) (resp. the LRI2 scheme (2.10)) satisfies

$$E_h(\mathbf{U}_m) \le E_h(\mathbf{U}_0) + C, \quad m = 0, 1, \dots, M, \tag{2.17}$$

where the constant $C > 0$ is independent of Δt.

Proof. According to Lemma 2.9, there exists $c > 0$ independence of Δt such that

$$\|\mathbf{U}_{m+1} - \mathbf{U}_m\|_\infty \le c\Delta t.$$

This, together with Lemma 2.8, gives us

$$E_h(\mathbf{U}_{m+1}) - E_h(\mathbf{U}_m) \le \|(\mathbf{U}_{m+1} - \mathbf{U}_m)^T\|_\infty(\|\mathbf{W}\|_\infty + \|\mathbf{A}\mathbf{U}_{m+1}\|_\infty) \le cN\Delta t(F_0 + \beta\|\mathbf{A}\|_\infty).$$

Therefore,

$$E_h(\mathbf{U}_m) \le E_h(\mathbf{U}_0) + cNT(F_0 + \beta\|\mathbf{A}\|_\infty) =: E_h(\mathbf{U}_0) + C,$$

where $C = cNT(F_0 + \beta\|\mathbf{A}\|_\infty) > 0$ independence of Δt. $\square$

2.1.3 Temporal error estimates

In this section, we always assume $\|u_0\|_{L^\infty} \le \beta$ so that

$$\mathbf{u} \in \mathcal{C}([0,T];\mathbb{R}^N) \text{ and } \|\mathbf{u}(t)\|_\infty \le \beta \text{ for all } t \in [0,T],$$

by Theorem 2.2 and the discrete MBP of the numerical solution $\{\mathbf{U}_m\}$ also holds by Theorem 2.5 or Theorem 2.6. By using Duhamel's formula as the key idea, we will show that these requirements on $\mathbf{u}$ (the continuity of $\mathbf{u}(t)$ and its MBP) are enough to obtain the temporal error estimates (under a fixed spatial mesh size) for the proposed LRI schemes. Let $\mathbf{e}_m := \mathbf{U}_m - \mathbf{u}(t_m)$ be the error at $t_m = m\Delta t \le T$. Let us define $F_2 = \max_{|x|\le\beta} |f''(x)|$, $F_3 = \max_{|x|\le\beta} |(f'(x)f(x))'|$, $\widetilde{F}_1 = \max_{|x|\le\beta+\delta_0\omega_0F_0} |f'(x)|$, and $\widetilde{F}_2 = \max_{|x|\le\beta+\delta_0\omega_0F_0} |f''(x)|$. The following result will be useful for the error analysis of the LRI schemes.

Lemma 2.11. For each $0 \le s \le \Delta t$, define

$$\boldsymbol{\lambda}(s) = e^{(\Delta t - s)\mathbf{A}}\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m)) \text{ and } \boldsymbol{\psi}(s) = e^{(\Delta t - s)\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m)).$$

Then

(i) $\|\boldsymbol{\lambda}'(s)\|_\infty \le (F_0 + \beta F_1)\|\mathbf{A}\|_\infty$ and $\|\boldsymbol{\lambda}''(s)\|_\infty \le (F_0 + 3\beta F_1 + \beta^2 F_2)\|\mathbf{A}\|_\infty^2$.

(ii) $\|\boldsymbol{\psi}'(s)\|_\infty \le F_0(F_1 + \beta F_2)\|\mathbf{A}\|_\infty$.

Proof. We have

$$\boldsymbol{\lambda}'(s) = e^{(\Delta t - s)\mathbf{A}}\left[-\mathbf{A}\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m)) + \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{u}(t_m))(\mathbf{A}e^{s\mathbf{A}}\mathbf{u}(t_m))\right].$$

Since $\|e^{s\mathbf{A}}\mathbf{u}(t_m)\|_\infty \le \beta$, this implies

$$\|\boldsymbol{\lambda}'(s)\|_\infty \le F_0\|\mathbf{A}\|_\infty + \beta F_1\|\mathbf{A}\|_\infty = (F_0 + \beta F_1)\|\mathbf{A}\|_\infty.$$

On the other hand,

$$\boldsymbol{\lambda}''(s) = e^{(\Delta t - s)\mathbf{A}}\Big[\mathbf{A}^2\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m)) - 2\mathbf{A}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{u}(t_m))(\mathbf{A}e^{s\mathbf{A}}\mathbf{u}(t_m))$$
$$+ \frac{\partial^2 \mathbf{f}}{\partial \mathbf{u}^2}(e^{s\mathbf{A}}\mathbf{u}(t_m))(\mathbf{A}e^{s\mathbf{A}}\mathbf{u}(t_m))^{\odot 2} + \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{u}(t_m))(\mathbf{A}^2e^{s\mathbf{A}}\mathbf{u}(t_m))\Big].$$

Thus we get

$$\|\boldsymbol{\lambda}''(s)\|_\infty \le F_0\|\mathbf{A}\|_\infty^2 + 2\beta F_1\|\mathbf{A}\|_\infty^2 + \beta^2F_2\|\mathbf{A}\|_\infty^2 + \beta F_1\|\mathbf{A}\|_\infty^2 = (F_0 + 3\beta F_1 + \beta^2 F_2)\|\mathbf{A}\|_\infty^2.$$

Similarly for $\boldsymbol{\psi}$, we have

$$\boldsymbol{\psi}'(s) = e^{(\Delta t - s)\mathbf{A}}\Big[-\mathbf{A}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m)) + \frac{\partial^2 \mathbf{f}}{\partial \mathbf{u}^2}(e^{s\mathbf{A}}\mathbf{u}(t_m))(\mathbf{A}e^{s\mathbf{A}}\mathbf{u}(t_m)) \odot \mathbf{f}(\mathbf{u}(t_m))\Big].$$

Therefore,

$$\|\boldsymbol{\psi}'(s)\|_\infty \le F_0(F_1\|\mathbf{A}\|_\infty + \beta F_2\|\mathbf{A}\|_\infty) = F_0(F_1 + \beta F_2)\|\mathbf{A}\|_\infty.$$

The proof is then completed. □

Next, we prove the convergence of the numerical solutions by the LRI1a, LRI1b and LRI2 schemes to the exact solution of the space-discrete problem (2.2) as Δt tends to zero, and their corresponding error estimates.

Theorem 2.12. Given any fixed terminal time $T > 0$ and spatial mesh size $h > 0$. Assume that $\mathbf{u} \in \mathcal{C}([0,T];\mathbb{R}^N)$ is the exact solution of (2.2) and $\{\mathbf{U}_m\}$ is the numerical solution generated by the LRI1a scheme (cf. Equation (2.8)). Then, there exists a constant $C = C(\beta, F_0, F_1, \|\mathbf{A}\|_\infty) > 0$ independent of Δt such that the following estimate is true for all $0 < \Delta t \le \omega_0$:

$$\|\mathbf{U}_m - \mathbf{u}(t_m)\|_\infty \le C(e^{F_1 t_m} - 1)\Delta t, \quad m = 0, 1, \dots, M.$$

Proof. For the LRI1a scheme (2.8), we have

$$\mathbf{u}(t_{m+1}) = e^{\Delta t\mathbf{A}}\left(\mathbf{u}(t_m) + \Delta t\mathbf{f}(\mathbf{u}(t_m))\right) + \mathbf{R}_1(t_m), \tag{2.18}$$

where $\mathbf{R}_1(t_m)$ is the corresponding truncation error. This together with (2.8) gives us

$$\mathbf{e}_{m+1} = e^{\Delta t \mathbf{A}} \left[\mathbf{e}_m + \Delta t \left(\mathbf{f}(\mathbf{U}_m) - \mathbf{f}(\mathbf{u}(t_m))\right)\right] - \mathbf{R}_1(t_m).$$

Since $\|e^{\Delta t \mathbf{A}}\|_\infty \leq 1$, we have

$$\|\mathbf{e}_{m+1}\|_\infty \leq \|\mathbf{e}_m\|_\infty + \Delta t \|\mathbf{f}(\mathbf{U}_m) - \mathbf{f}(\mathbf{u}(t_m))\|_\infty + \|\mathbf{R}_1(t_m)\|_\infty. \tag{2.19}$$

Observe that both the exact and numerical solutions of (2.2) satisfy the MBP when $0 < \Delta t \leq \omega_0$, meaning that $\|\mathbf{U}_m\|_\infty \leq \beta$, $\|\mathbf{u}(t_m)\|_\infty \leq \beta$. By using the Lipschitz continuity of $\mathbf{f}$, we obtain

$$\|\mathbf{f}(\mathbf{U}_m) - \mathbf{f}(\mathbf{u}(t_m))\|_\infty \leq F_1 \|\mathbf{U}_m - \mathbf{u}(t_m)\|_\infty = F_1 \|\mathbf{e}_m\|_\infty. \tag{2.20}$$

On the other hand, by comparing (2.4) and (2.18), we can rewrite $\mathbf{R}_1(t_m)$ as follows:

$$\begin{aligned}
\mathbf{R}_1(t_m) &= \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}} \mathbf{f}(\mathbf{u}(t_m + s))\, ds - \Delta t e^{\Delta t \mathbf{A}} \mathbf{f}(\mathbf{u}(t_m)) \\
&= \int_0^{\Delta t} \left[e^{(\Delta t - s)\mathbf{A}} \mathbf{f}(\mathbf{u}(t_m + s)) - e^{\Delta t \mathbf{A}} \mathbf{f}(\mathbf{u}(t_m)) \right] ds.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\|\mathbf{R}_1(t_m)\|_\infty \leq \int_0^{\Delta t} \|e^{\Delta t \mathbf{A}} \mathbf{f}(\mathbf{u}(t_m)) - e^{(\Delta t - s)\mathbf{A}} \mathbf{f}(\mathbf{u}(t_m + s))\|_\infty \, ds \\
&\leq \int_0^{\Delta t} \left(\|e^{\Delta t \mathbf{A}} \mathbf{f}(\mathbf{u}(t_m)) - e^{(\Delta t - s)\mathbf{A}} \mathbf{f}(e^{s\mathbf{A}} \mathbf{u}(t_m))\|_\infty + \|e^{(\Delta t - s)\mathbf{A}} \mathbf{f}(e^{s\mathbf{A}} \mathbf{u}(t_m)) - e^{(\Delta t - s)\mathbf{A}} \mathbf{f}(\mathbf{u}(t_m + s))\|_\infty \right) ds \\
&\leq \int_0^{\Delta t} \left(\|e^{\Delta t \mathbf{A}} \mathbf{f}(\mathbf{u}(t_m)) - e^{(\Delta t - s)\mathbf{A}} \mathbf{f}(e^{s\mathbf{A}} \mathbf{u}(t_m))\|_\infty + \|\mathbf{f}(e^{s\mathbf{A}} \mathbf{u}(t_m)) - \mathbf{f}(\mathbf{u}(t_m + s))\|_\infty \right) ds \\
&=: \int_0^{\Delta t} (Q_1 + Q_2)\, ds,
\end{aligned}$$

where

$$Q_1 = \|e^{\Delta t \mathbf{A}} \mathbf{f}(\mathbf{u}(t_m)) - e^{(\Delta t - s)\mathbf{A}} \mathbf{f}(e^{s\mathbf{A}} \mathbf{u}(t_m))\|_\infty, \quad Q_2 = \|\mathbf{f}(e^{s\mathbf{A}} \mathbf{u}(t_m)) - \mathbf{f}(\mathbf{u}(t_m + s))\|_\infty.$$

Using Lemma 2.11 and the mean value theorem, we deduce that

$$Q_1 = \|\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}(s)\|_\infty \leq s(F_0 + \beta F_1)\|\mathbf{A}\|_\infty. \tag{2.21}$$

To estimate Q_2, we apply the formula (2.5) to obtain

$$
\begin{aligned}
Q_2 &\le F_1 \|e^{s\mathbf{A}}\mathbf{u}(t_m) - \mathbf{u}(t_m+s)\|_\infty = F_1 \left\| \int_0^s e^{(s-\sigma)\mathbf{A}} \mathbf{f}(\mathbf{u}(t_m+\sigma))\, d\sigma \right\|_\infty \\
&\le F_1 \int_0^s \|e^{(s-\sigma)\mathbf{A}}\|_\infty \|\mathbf{f}(\mathbf{u}(t_m+\sigma))\|_\infty d\sigma \le F_0 F_1 s. \qquad (2.22)
\end{aligned}
$$

Thus we obtain from (2.21) and (2.22) that

$$
\begin{aligned}
\|\mathbf{R}_1(t_m)\|_\infty &\le \int_0^{\Delta t} (Q_1 + Q_2)\, ds \le \int_0^{\Delta t} [(F_0 + \beta F_1)\|\mathbf{A}\|_\infty + F_0 F_1] s\, ds \\
&= \frac{1}{2}[(F_0 + \beta F_1)\|\mathbf{A}\|_\infty + F_0 F_1]\Delta t^2 =: c_0 \Delta t^2. \qquad (2.23)
\end{aligned}
$$

By (2.19), (2.20) and (2.23), the following estimate holds:

$$
\|\mathbf{e}_{m+1}\|_\infty \le (1 + F_1 \Delta t)\|\mathbf{e}_m\|_\infty + c_0 \Delta t^2. \qquad (2.24)
$$

This implies

$$
\|\mathbf{e}_m\|_\infty + \frac{c_0 \Delta t}{F_1} \le (1 + F_1 \Delta t)\left(\|\mathbf{e}_{m-1}\|_\infty + \frac{c_0 \Delta t}{F_1}\right) \le (1 + F_1 \Delta t)^m \left(\|\mathbf{e}_0\|_\infty + \frac{c_0 \Delta t}{F_1}\right).
$$

Note that $\mathbf{e}_0 = 0$, thus

$$
\|\mathbf{e}_m\|_\infty \le \frac{c_0 \Delta t}{F_1}\left[(1 + F_1 \Delta t)^m - 1\right] \le \frac{c_0 \Delta t}{F_1}(e^{F_1 t_m} - 1) =: C(e^{F_1 t_m} - 1)\Delta t,
$$

where $C = \frac{c_0}{F_1} = \frac{1}{2}\left[\left(\frac{F_0}{F_1} + \beta\right)\|\mathbf{A}\|_\infty + F_0\right] > 0.$ □

Theorem 2.13. Given any fixed terminal time $T > 0$ and spatial mesh size $h > 0$. Assume that $\mathbf{u} \in \mathcal{C}([0, T]; \mathbb{R}^N)$ is the exact solution of (2.2) and $\{\mathbf{U}_m\}$ is the numerical solution generated by the LRI1b scheme (cf. Equation (2.9)). Then, there exists a constant $C = C(\beta, F_0, F_1, \|\mathbf{A}\|_\infty) > 0$ independent of Δt such that the following estimate is true for all $0 < \Delta t \le \omega_0$:

$$
\|\mathbf{U}_m - \mathbf{u}(t_m)\|_\infty \le C(e^{F_1 t_m} - 1)\Delta t, \quad m = 0, 1, \ldots, M.
$$

Proof. For the LRI1b scheme (2.9), we have

$$\mathbf{u}(t_{m+1}) = e^{\Delta t\mathbf{A}}\mathbf{u}(t_m) + \Delta t\mathbf{f}\left(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)\right) + \mathbf{R}_2(t_m), \tag{2.25}$$

where $\mathbf{R}_2(t_m)$ is the corresponding truncation error. This together with (2.9) gives us

$$\mathbf{e}_{m+1} = e^{\Delta t\mathbf{A}}\mathbf{e}_m + \Delta t\left[\mathbf{f}\left(e^{\Delta t\mathbf{A}}\mathbf{U}_m\right) - \mathbf{f}\left(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)\right)\right] - \mathbf{R}_2(t_m).$$

Since $\|e^{\Delta t\mathbf{A}}\|_\infty \leq 1$, we have

$$\|\mathbf{e}_{m+1}\|_\infty \leq \|\mathbf{e}_m\|_\infty + \Delta t\left\|\mathbf{f}\left(e^{\Delta t\mathbf{A}}\mathbf{U}_m\right) - \mathbf{f}\left(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)\right)\right\|_\infty + \|\mathbf{R}_2(t_m)\|_\infty. \tag{2.26}$$

Notice that $\|e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)\|_\infty \leq \beta$ and $\|e^{\Delta t\mathbf{A}}\mathbf{U}_m\|_\infty \leq \beta$ under the condition $0 < \Delta t \leq \omega_0$. Using the Lipschitz continuity of $\mathbf{f}$, we obtain

$$\left\|\mathbf{f}\left(e^{\Delta t\mathbf{A}}\mathbf{U}_m\right) - \mathbf{f}\left(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)\right)\right\|_\infty \leq F_1\left\|e^{\Delta t\mathbf{A}}\mathbf{U}_m - e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)\right\|_\infty \leq F_1\|\mathbf{e}_m\|_\infty. \tag{2.27}$$

Compare (2.25) and (2.4), we can rewrite $\mathbf{R}_2(t_m)$ as follows:

$$\begin{aligned}\mathbf{R}_2(t_m) &= \int_0^{\Delta t} e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+s))\,ds - \Delta t\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m))\\ &= \int_0^{\Delta t}\left[e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+s)) - \mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m))\right] ds.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\mathbf{R}_2(t_m)\|_\infty \leq& \int_0^{\Delta t}\Big(\|\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)) - e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m))\|_\infty\\ &\qquad + \|e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m)) - e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+s))\|_\infty\Big)ds\\ \leq& \int_0^{\Delta t}\Big(\|\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)) - e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m))\|_\infty\\ &\qquad + \|\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m)) - \mathbf{f}(\mathbf{u}(t_m+s))\|_\infty\Big)ds =: \int_0^{\Delta t}(Q_0+Q_2)ds,\end{aligned}$$

where

$$Q_0 = \|\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)) - e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m))\|_\infty,$$

and Q_2 is the same term as appeared in the proof of the previous theorem. Using Lemma 2.11 and the mean value theorem, we deduce that

$$Q_0 = \|\boldsymbol{\lambda}(\Delta t) - \boldsymbol{\lambda}(s)\|_\infty \le (\Delta t - s)(F_0 + \beta F_1)\|\mathbf{A}\|_\infty. \tag{2.28}$$

Thus we obtain from (2.28) and (2.22) that

$$\begin{aligned}\|\mathbf{R}_2(t_m)\|_\infty \le \int_0^{\Delta t} (Q_0 + Q_2)\,ds &\le \int_0^{\Delta t} [(\Delta t - s)(F_0 + \beta F_1)\|\mathbf{A}\|_\infty + F_0 F_1 s]\,ds \\ &= \frac{1}{2}[(F_0 + \beta F_1)\|\mathbf{A}\|_\infty + F_0 F_1]\Delta t^2 = c_0 \Delta t^2,\end{aligned} \tag{2.29}$$

where c_0 is the same constant as appeared in the proof of the previous theorem. Combine (2.26), (2.27) and (2.29) we have

$$\|\mathbf{e}_{m+1}\|_\infty \le (1 + F_1\Delta t)\|\mathbf{e}_m\|_\infty + c_0\Delta t^2.$$

By the same arguments as in Theorem 2.12, we obtain

$$\|\mathbf{e}_m\|_\infty \le C(e^{F_1 t_m} - 1)\Delta t,$$

where $C = \dfrac{c_0}{F_1} > 0$ is independent of Δt. □

Theorem 2.14. Given any fixed terminal time $T > 0$ and spatial mesh size $h > 0$. Suppose that f satisfies assumptions (A1) and (A3). Assume that $\mathbf{u} \in \mathcal{C}([0,T];\mathbb{R}^N)$ is the exact solution of (2.2) and $\{\mathbf{U}_m\}$ is the numerical solution generated by the LRI2 scheme (cf. Equation (2.10)). Then, there exists $C = C(\beta, \delta_0, \omega_0, F_0, F_1, \widetilde{F}_1, F_2, \widetilde{F}_2, F_3, \|\mathbf{A}\|_\infty) > 0$ independent of Δt such that the following estimate is true for all $0 < \Delta t \le \delta_0\omega_0$:

$$\|\mathbf{U}_m - \mathbf{u}(t_m)\|_\infty \le C(e^{F_4 t_m} - 1)\Delta t^2, \quad m = 0, 1, \dots, M,$$

where $F_4 := F_1 + \frac{1}{2}\delta_0\omega_0 F_3$.

Proof. Let $\mathbf{R}_3(t_m)$ be the truncation error of the LRI2 scheme (2.10), we have

$$\begin{aligned}\mathbf{u}(t_{m+1}) = e^{\Delta t\mathbf{A}}\mathbf{u}(t_m) + \frac{\Delta t}{2}\left[e^{\Delta t\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m)) + \mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)\right] \\ + \frac{\Delta t^2}{2}e^{\Delta t\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m)) + \mathbf{R}_3(t_m).\end{aligned} \tag{2.30}$$

This together with (2.10) gives us

$$\begin{aligned}\mathbf{e}_{m+1} = e^{\Delta t\mathbf{A}}\mathbf{e}_m + \frac{\Delta t}{2}\left[e^{\Delta t\mathbf{A}}(\mathbf{f}(\mathbf{U}_m) - \mathbf{f}(\mathbf{u}(t_m)) + \mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{U}_m) - \mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m))\right] \\ + \frac{\Delta t^2}{2}e^{\Delta t\mathbf{A}}\left[\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m) - \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m))\right] - \mathbf{R}_3(t_m).\end{aligned}$$

Therefore,

$$\begin{aligned}\|\mathbf{e}_{m+1}\|_\infty \le \|\mathbf{e}_m\|_\infty + \frac{\Delta t}{2}\left[\|\mathbf{f}(\mathbf{U}_m) - \mathbf{f}(\mathbf{u}(t_m)\|_\infty + \|\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{U}_m) - \mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m))\|_\infty\right] \\ + \frac{\Delta t^2}{2}\left\|\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{U}_m)\mathbf{f}(\mathbf{U}_m) - \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m))\right\|_\infty + \|\mathbf{R}_3(t_m)\|_\infty.\end{aligned}$$

Using the similar arguments as in proofs of Theorems 2.12 and 2.13, we can easily obtain

$$\begin{aligned}\|\mathbf{f}(\mathbf{U}_m) - \mathbf{f}(\mathbf{u}(t_m)\|_\infty \le F_1\|\mathbf{U}_m - \mathbf{u}(t_m)\|_\infty = F_1\|\mathbf{e}_m\|_\infty, \\ \|\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{U}_m) - \mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m))\|_\infty \le F_1\|e^{\Delta t\mathbf{A}}\mathbf{U}_m - e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)\|_\infty \le F_1\|\mathbf{e}_m\|_\infty.\end{aligned}$$

Define $\mathbf{l}(\mathbf{v}) = \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{v})\mathbf{f}(\mathbf{v})$ for $\mathbf{v} \in \mathbb{R}^N$. It is easy to verify that $\|\frac{\partial \mathbf{l}}{\partial \mathbf{v}}\|_\infty \le F_3$ when $\|\mathbf{v}\|_\infty \le \beta$, then

$$\|\mathbf{l}(\mathbf{U}_m) - \mathbf{l}(\mathbf{u}(t_m))\|_\infty \le F_3\|\mathbf{U}_m - \mathbf{u}(t_m)\|_\infty = F_3\|\mathbf{e}_m\|_\infty.$$

Consequently, we have

$$\|\mathbf{e}_{m+1}\|_\infty \le \left(1 + F_1\Delta t + \frac{1}{2}F_3\Delta t^2\right)\|\mathbf{e}_m\|_\infty + \|\mathbf{R}_3(t_m)\|_\infty. \tag{2.31}$$

On the other hand, by comparing (2.30) and (2.4), for $\mathbf{R}_3(t_m)$ we arrive at

$$\begin{aligned}\mathbf{R}_3(t_m) &= \int_0^{\Delta t} e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+s))\,ds - \frac{\Delta t}{2}\left[e^{\Delta t\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m)) + \mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)\right] \\ &\qquad - \frac{\Delta t^2}{2}e^{\Delta t\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m)) \\ &= \int_0^{\Delta t}\Big[e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+s)) - se^{\Delta t\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m)) \\ &\qquad - \left(1-\frac{s}{\Delta t}\right)e^{\Delta t\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m)) - \frac{s}{\Delta t}\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m))\Big]\,ds \\ &= \int_0^{\Delta t}(I_1+I_2+I_3+I_4)\,ds,\end{aligned}$$

where $I_i, i = 1,2,3,4$ are defined as follow:

$$\begin{aligned}I_1 &= e^{(\Delta t-s)\mathbf{A}}\left[\mathbf{f}(\mathbf{u}(t_m+s)) - \mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m) + s\mathbf{f}(\mathbf{u}(t_m)))\right], \\ I_2 &= e^{(\Delta t-s)\mathbf{A}}\left[\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m) + s\mathbf{f}(\mathbf{u}(t_m))) - \left(\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m)) + s\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m))\right)\right], \\ I_3 &= se^{(\Delta t-s)\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m)) - se^{\Delta t\mathbf{A}}\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m)), \\ I_4 &= e^{(\Delta t-s)\mathbf{A}}\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m)) - \left(1-\frac{s}{\Delta t}\right)e^{\Delta t\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m)) - \frac{s}{\Delta t}\mathbf{f}(e^{\Delta t\mathbf{A}}\mathbf{u}(t_m)).\end{aligned}$$

Since $\|\mathbf{u}(t_m)\|_\infty \leq \beta$ and $s \leq \Delta t \leq \delta_0\omega_0$, we have

$$\|e^{s\mathbf{A}}\mathbf{u}(t_m) + s\mathbf{f}(\mathbf{u}(t_m))\|_\infty \leq \|e^{s\mathbf{A}}\mathbf{u}(t_m)\|_\infty + s\|\mathbf{f}(\mathbf{u}(t_m))\|_\infty \leq \beta + \delta_0\omega_0 F_0. \tag{2.32}$$

It is easy to see that when $\mathbf{v},\mathbf{w} \in \mathbb{R}^N$ and $\|\mathbf{v}\|_\infty, \|\mathbf{w}\|_\infty \leq \beta + \delta_0\omega_0 F_0$, then

$$\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{w})\|_\infty \leq \widetilde{F}_1\|\mathbf{v}-\mathbf{w}\|_\infty.$$

This, together with the formula (2.5), gives us

$$
\begin{aligned}
\|I_1\|_\infty &\le \left\|\mathbf{f}(\mathbf{u}(t_m+s)) - \mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m) + s\mathbf{f}(\mathbf{u}(t_m)))\right\|_\infty \\
&\le \widetilde{F}_1 \left\|\mathbf{u}(t_m+s) - \left(e^{s\mathbf{A}}\mathbf{u}(t_m) + s\mathbf{f}(\mathbf{u}(t_m))\right)\right\|_\infty \\
&= \widetilde{F}_1 \left\|s\mathbf{f}(\mathbf{u}(t_m)) - \int_0^s e^{(s-\sigma)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+\sigma))\,d\sigma\right\|_\infty \\
&= \widetilde{F}_1 \left\|\int_0^s \left[\mathbf{f}(\mathbf{u}(t_m)) - e^{(s-\sigma)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+\sigma))\right]\,d\sigma\right\|_\infty \\
&\le \widetilde{F}_1 \int_0^s \big[\|\mathbf{f}(\mathbf{u}(t_m)) - e^{s\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m))\|_\infty + \|e^{s\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m)) - e^{(s-\sigma)\mathbf{A}}\mathbf{f}(e^{\sigma\mathbf{A}}\mathbf{u}(t_m))\|_\infty \\
&\qquad\qquad + \|e^{(s-\sigma)\mathbf{A}}\mathbf{f}(e^{\sigma\mathbf{A}}\mathbf{u}(t_m)) - e^{(s-\sigma)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+\sigma))\|_\infty\big]\,d\sigma \\
&=: \widetilde{F}_1 \int_0^s (J_1+J_2+J_3)\,d\sigma,
\end{aligned}
$$

where J_1, J_2 and J_3 are given by

$$
\begin{aligned}
J_1 &= \|\mathbf{f}(\mathbf{u}(t_m)) - e^{s\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m))\|_\infty, \\
J_2 &= \|e^{s\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m)) - e^{(s-\sigma)\mathbf{A}}\mathbf{f}(e^{\sigma\mathbf{A}}\mathbf{u}(t_m))\|_\infty, \\
J_3 &= \|e^{(s-\sigma)\mathbf{A}}\mathbf{f}(e^{\sigma\mathbf{A}}\mathbf{u}(t_m)) - e^{(s-\sigma)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+\sigma))\|_\infty.
\end{aligned}
$$

By the mean value theorem, we have $J_1 \le F_0\|e^{s\mathbf{A}} - I\|_\infty \le sF_0\|\mathbf{A}\|_\infty$. Using Lemma 2.11 and the mean value theorem, we obtain $J_2 \le (F_0 + \beta F_1)\|\mathbf{A}\|_\infty\sigma$. Using the formula (2.5), we have

$$
\begin{aligned}
J_3 &\le F_1\|e^{\sigma\mathbf{A}}\mathbf{u}(t_m) - \mathbf{u}(t_m+\sigma)\|_\infty \\
&= F_1\left\|\int_0^\sigma e^{(\sigma-\xi)\mathbf{A}}\mathbf{f}(\mathbf{u}(t_m+\xi))d\xi\right\|_\infty \le F_1\int_0^\sigma F_0 d\xi = F_0F_1\sigma.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\|I_1\|_\infty &\le \widetilde{F}_1\int_0^s (J_1+J_2+J_3)\,d\sigma \le \widetilde{F}_1\int_0^s [F_0\|\mathbf{A}\|_\infty s + (F_0+\beta F_1)\|\mathbf{A}\|_\infty\sigma + F_0F_1\sigma]d\sigma \\
&= \frac{1}{2}\widetilde{F}_1[(3F_0+\beta F_1)\|\mathbf{A}\|_\infty + F_0F_1]s^2 =: c_1 s^2.
\end{aligned}
\tag{2.33}
$$

By Taylor expansion, there exists $\mathbf{V} \in \mathbb{R}^N$, $\|\mathbf{V}\|_\infty \le \beta + \delta_0\omega_0 F_0$ (from (2.32)) such that

$$\mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m) + s\mathbf{f}(\mathbf{u}(t_m))) = \mathbf{f}(e^{s\mathbf{A}}\mathbf{u}(t_m)) + s\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(e^{s\mathbf{A}}\mathbf{u}(t_m))\mathbf{f}(\mathbf{u}(t_m)) + \frac{\partial^2 \mathbf{f}}{\partial \mathbf{u}^2}(\mathbf{V})\frac{[s\mathbf{f}(\mathbf{u}(t_m))]^{\odot 2}}{2}.$$

Thus one gets

$$\|I_2\|_\infty \le \left\| \frac{\partial^2 \mathbf{f}}{\partial \mathbf{u}^2}(\mathbf{V})\frac{[s\mathbf{f}(\mathbf{u}(t_m))]^{\odot 2}}{2} \right\|_\infty \le \frac{1}{2}F_0^2 \widetilde{F}_2 s^2 =: c_2 s^2. \tag{2.34}$$

To estimate I_3, using Lemma 2.11 and the mean value theorem, we obtain

$$\|I_3\|_\infty = s\|\boldsymbol{\psi}(s) - \boldsymbol{\psi}(0)\|_\infty \le F_0(F_1 + \beta F_2)\|\mathbf{A}\|_\infty s^2 =: c_3 s^2. \tag{2.35}$$

For I_4 notice that

$$\|I_4\|_\infty = \left\| \boldsymbol{\lambda}(s) - \left(1 - \frac{s}{\Delta t}\right)\boldsymbol{\lambda}(0) - \frac{s}{\Delta t}\boldsymbol{\lambda}(\Delta t) \right\|_\infty .$$

Using Taylor expansion, there exist $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{R}^N$ satisfying

$$\|\mathbf{V}_1\|_\infty, \|\mathbf{V}_2\|_\infty \le \max_{s\in[0,\Delta t]} \|\boldsymbol{\lambda}''(s)\|_\infty \overset{\text{Lemma 2.11}}{\le} (F_0 + 3\beta F_1 + \beta^2 F_2)\|\mathbf{A}\|_\infty^2$$

subject to

$$\begin{cases} \boldsymbol{\lambda}(0) = \boldsymbol{\lambda}(s) - s\boldsymbol{\lambda}'(s) + \frac{s^2}{2}\mathbf{V}_1, \\ \boldsymbol{\lambda}(\Delta t) = \boldsymbol{\lambda}(s) + (\Delta t - s)\boldsymbol{\lambda}'(s) + \frac{(\Delta t - s)^2}{2}\mathbf{V}_2. \end{cases}$$

Thus we have

$$\|I_4\|_\infty = \left\| \left(1 - \frac{s}{\Delta t}\right)\frac{s^2}{2}\mathbf{V}_1 + \frac{s}{\Delta t}\frac{(\Delta t - s)^2}{2}\mathbf{V}_2 \right\|_\infty \tag{2.36}$$

$$\le \frac{1}{2}(F_0 + 3\beta F_1 + \beta^2 F_2)\|\mathbf{A}\|_\infty^2 (s\Delta t - s^2) =: c_4(s\Delta t - s^2). \tag{2.37}$$

Combining the above estimates (2.33)-(2.36) for $I_i, i = 1, 2, 3, 4$ yields

$$\|\mathbf{R}_3(t_m)\|_\infty \le \int_0^{\Delta t} (c_1 + c_2 + c_3)s^2 + c_4(s\Delta t - s^2)\, ds = \left(\frac{c_1 + c_2 + c_3}{3} + \frac{c_4}{6}\right)\Delta t^3 =: c\Delta t^3. \tag{2.38}$$

From (2.31) and (2.38), we get

$$\|\mathbf{e}_{m+1}\|_\infty \le \left(1 + F_1\Delta t + \frac{1}{2}F_3\Delta t^2\right)\|\mathbf{e}_m\|_\infty + c\Delta t^3 \le (1 + F_4\Delta t)\|\mathbf{e}_m\|_\infty + c\Delta t^3.$$

Therefore,

$$\|\mathbf{e}_m\|_\infty + \frac{c\Delta t^2}{F_4} \le (1 + F_4\Delta t)^m\left(\|\mathbf{e}_0\|_\infty + \frac{c\Delta t^2}{F_4}\right).$$

Using the fact that $\mathbf{e}_0 = 0$, we deduce that

$$\|\mathbf{e}_m\|_\infty \le \frac{c\Delta t^2}{F_4}[(1 + F_4\Delta t)^m - 1] \le \frac{c\Delta t^2}{F_4}(e^{F_4 t_m} - 1) =: C(e^{F_4 t_m} - 1)\Delta t^2,$$

where $C = \frac{c}{F_4} = \frac{1}{F_4}\left(\frac{c_1+c_2+c_3}{3} + \frac{c_4}{6}\right) > 0$, and c_1, c_2, c_3, c_4, F_4 depend on $\beta, \delta_0, \omega_0, F_0, F_1, \widetilde{F}_1$, $F_2, \widetilde{F}_2, F_3$, and $\|\mathbf{A}\|_\infty$. □

Remark 2.15. For the case of $f(u) = u - u^3$ (i.e., the Allen-Cahn equation with the double-well potential), the estimates derived in Theorems 2.12, 2.13 and 2.14 could be made simpler. In particular, for the LRI1a and LRI1b schemes, there exists $C = C(\|\mathbf{A}\|_\infty) > 0$ such that the following inequality holds for all $0 < \Delta t \le \frac{1}{2}$:

$$\|\mathbf{U}_m - \mathbf{u}(t_m)\|_\infty \le C(e^{2t_m} - 1)\Delta t, \quad m = 0, 1, \dots, M.$$

For the LRI2 scheme, there exists $C = C(\|\mathbf{A}\|_\infty) > 0$ so that for all $0 < \Delta t \le \frac{3}{5}$ we have:

$$\|\mathbf{U}_m - \mathbf{u}(t_m)\|_\infty \le C(e^{\frac{16}{5}t_m} - 1)\Delta t^2, \quad m = 0, 1, \dots, M.$$

2.1.4 Numerical results

We now verify numerically the convergence rates in time, MBP preservation and energy stability of the proposed LRI schemes (LRI1a (2.8), LRI1b (2.9), and LRI2 (2.10)) analyzed in the previous sections. To illustrate the advantages of the LRI schemes, we compare their performance with the ETD schemes [56] of orders 1 and 2 (i.e., ETD1 and ETDRK2) in terms of time discretization errors as the interfacial parameter ε approaches zero. Two test cases in the two dimensional domain $\Omega = (-0.5, 0.5)^2$ are presented: Test case 1 having a traveling wave solution and Test case 2 modeling the coarsen dynamics. For the latter, both double-well potential and Flory-Huggins potential functions will be considered in the numerical simulations. All numerical results are obtained via a fast implementation of the LRI and ETD methods based on matrix decomposition and Discrete Fourier Transform (DFT) as proposed in [104]. The code is implemented in MATLAB on a MacBook Pro with the M1 Pro chip and 32 GB memory.

Test case 1: Traveling wave

To investigate the convergence in time of the proposed schemes, we use the benchmark test [116] where a traveling wave solution of the Allen-Cahn equation (2.1) with $f(u) = u - u^3$ is considered. In particular, if the initial data is given by

$$u_0(x, y) = \frac{1}{2}\left(1 - \tanh\left(\frac{x}{2\sqrt{2}\varepsilon}\right)\right), \ (x, y) \in \overline{\Omega}, \tag{2.39}$$

then the Allen-Cahn equation in the whole space has the traveling wave solution

$$u(x, y, t) = \frac{1}{2}\left(1 - \tanh\left(\frac{x - st}{2\sqrt{2}\varepsilon}\right)\right), \ (x, y) \in \overline{\Omega}, \tag{2.40}$$

where $s = {}^{3\varepsilon}\!/\!\sqrt{2}$ is the speed of the traveling wave. We impose a homogeneous Neumann boundary condition on the whole boundary $\partial\Omega$, so that (2.40) can be used as an *approximate exact solution* (for $\varepsilon \ll 1$) on $\overline{\Omega} \times [0, T]$. We take $\varepsilon \in \{0.02, 0.01, 0.005\}$, and set the ending time $T = {}^{1}\!/\!{}_{4s}$. With a fixed mesh size $h = 1/2048$, we compute the numerical solutions obtained by the first-order methods (ETD1, LRI1a and LRI1b) and the second-order methods (ETDRK2 and LRI2) with decreasing time step sizes. The L^2 and L^∞ errors at time $t = T$ and their convergence rates of the first-order schemes are reported in Tables 2.1 and 2.2, respectively. The corresponding results for the second-order schemes are shown in Table 2.3.

For the first-order schemes, we observe expected first-order temporal convergence rates in both L^2 and L^∞ norms with various values of ε. In addition, the errors by two LRI1 schemes are always smaller than those by ETD1; the performance of LRI1a and LRI1b is quite similar as ε decreases. For the second-order schemes, we first notice that the errors are much smaller

Table 2.1: [Test case 1] L^2 errors and convergence rates in time by the first-order ETD and LRI schemes.

ε	Δt	ETD1		LRI1a		LRI1b	
		L^2 Error	Rate	L^2 Error	Rate	L^2 Error	Rate
0.02	$T/32$	4.24e-2	–	3.35e-2	–	2.86e-2	–
	$T/64$	2.24e-2	0.92	1.75e-2	0.93	1.50e-2	0.93
	$T/128$	1.15e-2	0.96	8.98e-3	0.97	7.66e-3	0.97
	$T/256$	5.83e-3	0.98	4.55e-3	0.98	3.87e-3	0.98
	$T/512$	2.93e-3	0.99	2.29e-3	0.99	1.95e-3	0.99
	$T/1024$	1.47e-3	1.00	1.15e-3	1.00	9.77e-4	1.00
0.01	$T/32$	1.11e-1	–	8.51e-2	–	7.89e-2	–
	$T/64$	6.36e-2	0.81	4.73e-2	0.85	4.37e-2	0.85
	$T/128$	3.40e-2	0.90	2.50e-2	0.92	2.30e-2	0.93
	$T/256$	1.76e-2	0.95	1.28e-2	0.96	1.18e-2	0.96
	$T/512$	8.91e-3	0.98	6.48e-3	0.98	5.96e-3	0.98
	$T/1024$	4.48e-3	0.99	3.25e-3	0.99	2.99e-3	0.99
0.005	$T/32$	2.08e-1	–	1.70e-1	–	1.66e-1	–
	$T/64$	1.45e-1	0.52	1.12e-1	0.61	1.08e-1	0.61
	$T/128$	8.93e-2	0.70	6.51e-2	0.78	6.28e-2	0.79
	$T/256$	4.93e-2	0.86	3.49e-2	0.90	3.36e-2	0.90
	$T/512$	2.57e-2	0.94	1.80e-2	0.96	1.73e-2	0.96
	$T/1024$	1.30e-2	0.98	9.07e-3	0.99	8.69e-3	0.99

Table 2.2: [Test case 1] L^∞ errors and convergence rates in time by the first-order ETD and LRI schemes.

ε	Δt	ETD1		LRI1a		LRI1b	
		L^∞ Error	Rate	L^∞ Error	Rate	L^∞ Error	Rate
0.02	$T/32$	1.61e-1	–	1.28e-1	–	1.12e-1	–
	$T/64$	8.49e-2	0.93	6.70e-2	0.93	5.83e-2	0.93
	$T/128$	4.35e-2	0.96	3.43e-2	0.97	2.97e-2	0.97
	$T/256$	2.20e-2	0.98	1.74e-2	0.98	1.50e-2	0.98
	$T/512$	1.11e-2	0.99	8.73e-3	0.99	7.54e-3	0.99
	$T/1024$	5.55e-3	1.00	4.38e-3	1.00	3.78e-3	1.00
0.01	$T/32$	5.51e-1	–	4.30e-1	–	4.11e-1	–
	$T/64$	3.30e-1	0.74	2.47e-1	0.80	2.32e-1	0.82
	$T/128$	1.78e-1	0.89	1.31e-1	0.91	1.22e-1	0.92
	$T/256$	9.22e-2	0.95	6.76e-2	0.96	6.27e-2	0.97
	$T/512$	4.67e-2	0.98	3.42e-2	0.98	3.16e-2	0.99
	$T/1024$	2.35e-2	0.99	1.72e-2	0.99	1.59e-2	1.00
0.005	$T/32$	9.68e-1	–	9.06e-1	–	9.04e-1	–
	$T/64$	8.49e-1	0.19	7.13e-1	0.34	7.07e-1	0.35
	$T/128$	6.05e-1	0.49	4.57e-1	0.64	4.48e-1	0.66
	$T/256$	3.55e-1	0.77	2.54e-1	0.85	2.47e-1	0.86
	$T/512$	1.88e-1	0.92	1.32e-1	0.94	1.28e-1	0.95
	$T/1024$	9.57e-2	0.97	6.69e-2	0.98	6.43e-2	0.99

Table 2.3: [Test case 1] Errors and convergence rates in time by the second ETD and LRI schemes.

ε	Δt	ETDRK2				LRI2			
		L^2 Error	Rate	L^∞ Error	Rate	L^2 Error	Rate	L^∞ Error	Rate
0.02	$T/16$	8.92e-3	–	3.08e-2	–	6.34e-3	–	2.57e-2	–
	$T/32$	2.50e-3	1.83	8.83e-3	1.80	1.80e-3	1.82	7.24e-3	1.83
	$T/64$	6.65e-4	1.91	2.36e-3	1.90	4.80e-4	1.91	1.92e-3	1.91
	$T/128$	1.71e-4	1.95	6.10e-4	1.95	1.24e-4	1.95	4.94e-4	1.96
	$T/256$	4.57e-5	1.91	1.53e-4	2.00	3.48e-5	1.84	1.43e-4	1.79
	$T/512$	1.95e-5	1.23	1.43e-4	0.10	1.81e-5	0.94	1.43e-4	0.00
0.01	$T/16$	3.75e-2	–	1.80e-1	–	2.97e-2	–	1.62e-1	–
	$T/32$	1.18e-2	1.67	5.92e-2	1.61	9.44e-3	1.65	5.14e-2	1.66
	$T/64$	3.35e-3	1.81	1.70e-2	1.80	2.68e-3	1.82	1.45e-2	1.83
	$T/128$	8.87e-4	1.92	4.52e-3	1.91	7.05e-4	1.92	3.80e-3	1.93
	$T/256$	2.19e-4	2.02	1.12e-3	2.02	1.71e-4	2.04	9.28e-4	2.04
	$T/512$	4.45e-5	2.30	2.30e-4	2.28	3.25e-5	2.40	1.81e-4	2.36
0.005	$T/16$	2.74e-1	–	7.06e-1	–	NaN	–	NaN	–
	$T/32$	4.87e-2	2.49	3.38e-1	1.06	4.23e-2	–	3.13e-1	–
	$T/64$	1.67e-2	1.55	1.17e-1	1.53	1.36e-2	1.64	1.02e-1	1.62
	$T/128$	6.74e-3	1.31	2.77e-2	2.08	3.78e-3	1.84	2.82e-2	1.85
	$T/256$	3.05e-2	-2.18	7.06e-2	-1.35	9.24e-4	2.03	6.90e-3	2.03
	$T/512$	3.61e-2	-0.24	7.65e-2	-0.12	1.54e-4	2.59	1.17e-3	2.57

compared to the first-order schemes with the same time step sizes, especially when ε is close to zero. Importantly, second-order temporal convergence rates of LRI2 are confirmed even for small ε. For $\varepsilon = 0.02$, the errors in space are dominant when Δt is sufficiently small, that is why the convergence rates deteriorate after several refinements in time. Moreover, we remark that the errors are computed using the approximate exact solution (2.40) which is accurate when ε is close enough to zero. Unlike LRI2, ETDRK2 fails to converge when $\varepsilon = 0.005$; a similar behavior was also observed for the Navier-Stokes equations with the classical exponential integrator when the viscosity coefficient is close to zero [120]. Furthermore, as for the first-order schemes, the errors by ETDRK2 are always bigger than those by LRI2 with the same Δt for all considered values of ε; this phenomenon becomes much more significant when ε gets smaller.

We report in Table 2.4 the running times (in seconds) of LRI and ETD methods with fixed $h = 1/2048$ and varying time step sizes. The running times are similar for different values of ε, thus we only show the results when $\varepsilon = 0.01$. We observe that for the first-order schemes, LRI1a and LRI1b have similar running times and are faster than ETD1. For the second-order methods, LRI2 also outperforms ETDRK2 regarding CPU time. Thus, we conclude that LRI1 and LRI2 schemes are more effective than ETD1 and ETDRK2 methods, respectively, in terms of both accuracy and computational cost.

Table 2.4: [Test case 1] Running times (in seconds) of the first and second-order LRI and ETD schemes with $h = 1/2048$.

Δt	ETD1	LRI1a	LRI1b	ETDRK2	LRI2
$T/32$	29.77	24.12	24.13	41.51	35.14
$T/64$	47.08	35.73	35.79	70.57	57.95
$T/128$	82.13	59.42	59.51	129.20	103.76
$T/256$	152.24	107.50	106.96	246.61	194.71
$T/512$	296.41	204.15	203.84	480.35	379.02

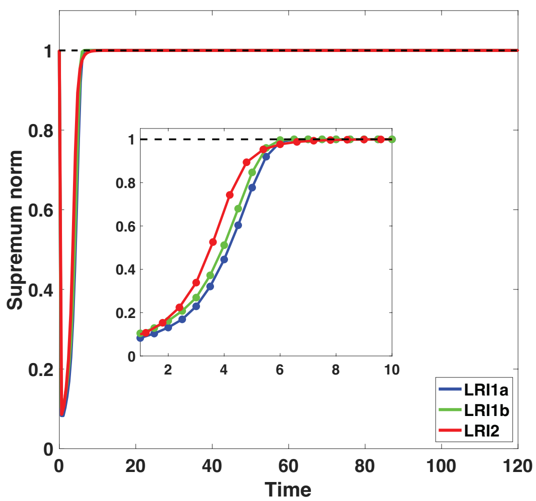
Test case 2: Coarsening dynamics

Next we validate MBP preservation and energy dissipation of the proposed LRIs. Toward that end, we simulate the process of the coarsening dynamics by considering a random initial configuration. The model problem is given by the Allen-Cahn equation (2.1) with periodic boundary conditions. We consider two choices for the nonlinear function $f(u)$:

$$f(u) = f_1(u) = u - u^3, \quad \text{or} \quad f(u) = f_2(u) = \frac{\theta}{2}\ln\frac{1-u}{1+u} + \theta_c u, \tag{2.41}$$

corresponding to the double-well and Flory-Huggins potential cases, respectively. For the latter, the parameters are $\theta = 0.8$ and $\theta_c = 1.6$ as in Remark 2.7. We set $\varepsilon = 0.01$, and fix the spatial mesh $h = 1/1024$. The initial data is generated by random numbers on each mesh point, the range of these numbers will be specified for each nonlinear function considered.

When $f = f_1$ (double-well potential case) in (2.41), we run the simulation with the random initial data ranging from -1 to 1 and a uniform time step of size $\Delta t = 0.5$ for LRI1a and LRI1b, and $\Delta t = 0.6$ for LRI2. Note that these are the upper bounds of the time steps predicted by our theoretical results (cf. Remark 2.7). Evolutions of the supremum norm and the energy of the numerical solutions are depicted in Figure 2.1. It can be seen that the numerical solutions obtained are all bounded by 1 in absolute value, and the associated energies are dissipative over a long time horizon. Figure 2.2 shows the snapshots of the numerical solution at $t = 5, 10, 20, 40, 80,$ and 120, generated by using the LRI2 scheme with a smaller time step $\Delta t = 0.1$; when $\Delta t = 0.05$, we obtain very similar results for the evolution of the numerical solution.



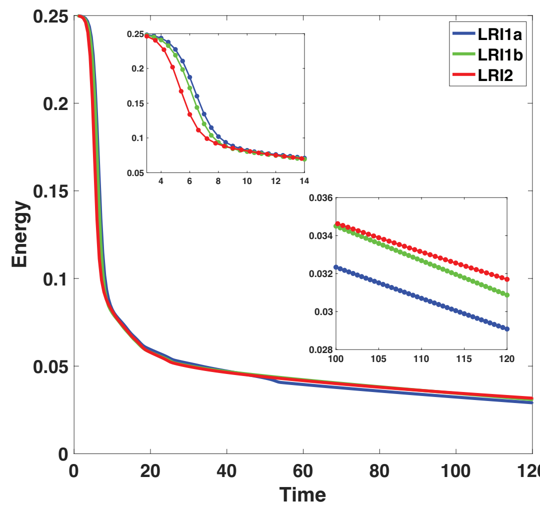


Figure 2.1: [Test case 2 with the double-well potential f_1] Evolutions of the supremum norm (left) and the energy (right) of the numerical solutions by the LRI1a and LRI1b schemes with $\Delta t = 0.5$ and by the LRI2 scheme with $\Delta t = 0.6$.

When $f = f_2$ (Flory-Huggins potential case) in (2.41), we use the random initial data ranging from -0.9 to 0.9. A uniform time step of of size $\Delta t = 0.12$ is used for LRI1a and LRI1b, and $\Delta t = 0.17$ for LRI2, which are close to the upper bounds estimated in Remark 2.7. Evolutions of the supremum norm and the energy of the numerical solutions are plotted in Figures 2.3. We again observe that the LRI schemes preserve MBP very well with the chosen time steps, i.e., the supremum norm of numerical solutions does not exceed $\beta \approx 0.9575$, and the energy monotonically decays along the time.

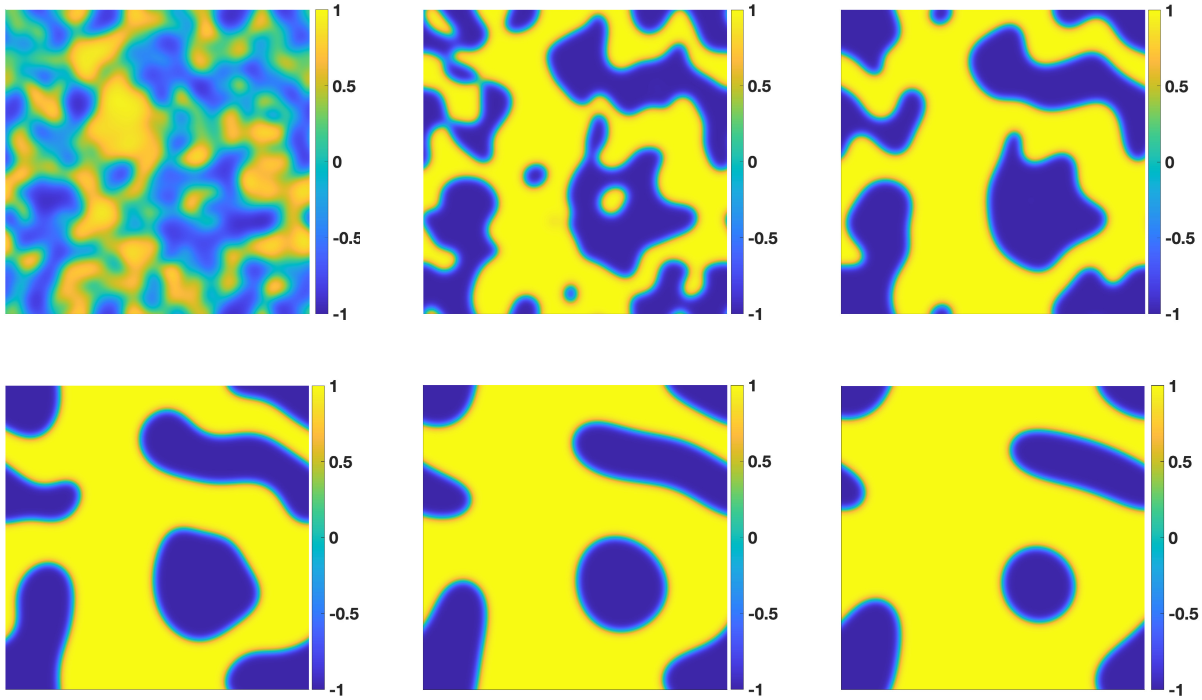


Figure 2.2: [Test case 2 with the the double-well potential f_1] Snapshots of the numerical solution at $t = 5, 10, 20, 40, 80$ and 120 (from left to right and top to bottom) produced by the LRI2 scheme with $\Delta t = 0.1$.

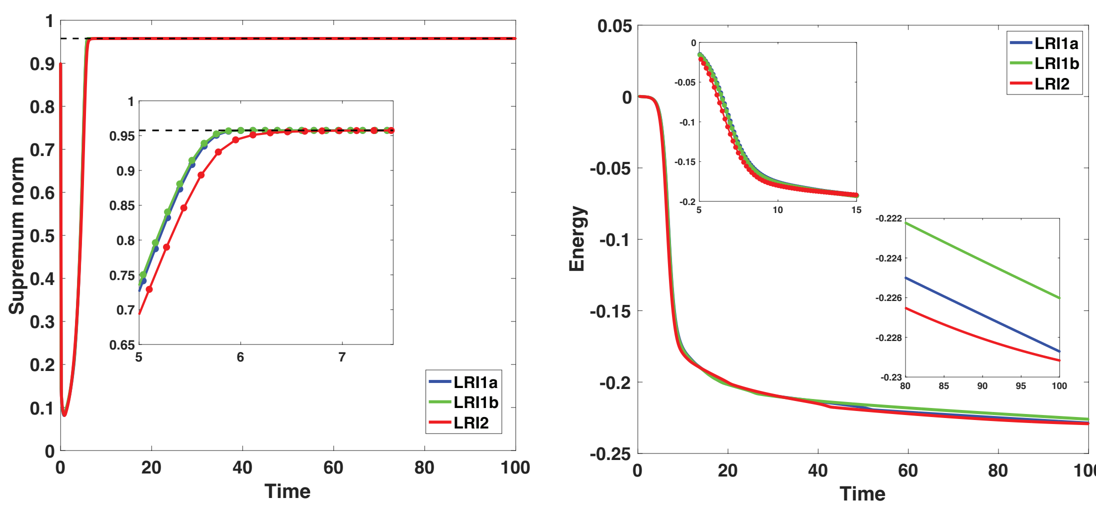


Figure 2.3: [Test case 2 with the Flory-Huggins potential f_2] Evolutions of the supremum norm (left) and the energy (right) of the numerical solutions by the LRI1a and LRI1b schemes with $\Delta t = 0.12$ and by the LRI2 scheme with $\Delta t = 0.17$.

2.2 Fully discrete error analysis of first-order low regularity integrators

In this section, we carry out fully discrete error analysis of the first-order LRI schemes (LRI1a and LRI1b) for the Allen-Cahn equation

$$\partial_t u = \varepsilon^2 \Delta u + f(u), \quad \mathbf{x} \in \Omega,\ t > 0, \tag{2.42}$$

together with the initial condition $u(\mathbf{x}, 0) = u_0(\mathbf{x})$ and the homogeneous Neumann boundary condition $\frac{\partial u}{\partial n}(\mathbf{x}, t) = 0$ for $\mathbf{x} \in \partial\Omega$ and $0 \le t \le T$. Note that periodic and homogeneous Dirichlet boundary conditions also can be imposed instead as discussed later. The nonlinear function f is assumed to satisfy assumptions (A1) and (A2), together with the following additional assumption:

(A4) f' is Lipschitz continuous on $[-\beta, \beta]$.

2.2.1 Fully discrete first-order low regularity integrators

Throughout this section, we assume that the nonlinear function f satisfies assumptions (A1) and (A2) and the initial data is bounded by β in the absolute value, i.e., $\|u_0\|_{L^\infty} \le \beta$. Suppose that the domain Ω is an interval, a rectangle, or a rectangular parallelepiped corresponding to $d = 1, 2$ and 3, respectively. For spatial discretization, we apply a uniform partition of Ω into square elements of length h, and let N be the total number of grid nodes. Denote by I^h the operator restricting a function over the grid points. Using the central finite difference method and the homogeneous Neumann boundary condition, we obtain the corresponding semi-discrete in space problem:

$$\frac{d\mathbf{u}_h}{dt} = \mathbf{A}_h \mathbf{u}_h + \mathbf{F}(\mathbf{u}_h), \quad \mathbf{u}_h(0) = I^h u_0 := (u_0(x_1), u_0(x_2), \dots, u_0(x_N))^T, \tag{2.43}$$

where $\mathbf{u}_h(t) = (u_1(t), u_2(t), \dots, u_N(t))^T$ is the approximation of $u(\cdot, t)$ over the grid points, $\mathbf{F}(\mathbf{u}_h) = (f(u_1), f(u_2), \dots, f(u_N))^T$, $\mathbf{A}_h = \varepsilon^2 \mathbf{D}_h$ with $\mathbf{D}_h = \lambda_h$ in one dimension, $\mathbf{D}_h = \mathbf{I} \otimes \lambda_h + \lambda_h \otimes \mathbf{I}$ in two dimensions, and

$$\mathbf{D}_h = \mathbf{I} \otimes \mathbf{I} \otimes \lambda_h + \mathbf{I} \otimes \lambda_h \otimes \mathbf{I} + \lambda_h \otimes \mathbf{I} \otimes \mathbf{I}$$

in three dimensions. Here $\otimes$ denotes the Kronecker product, $\mathbf{I}$ is the identity matrix of size N, and λ_h is given by

$$\lambda_h = \frac{1}{h^2}\begin{bmatrix} -2 & 2 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \\ 0 & \cdots & 0 & 2 & -2 \end{bmatrix}_{N\times N}.$$

For temporal discretization, let us consider the uniform partition of the time interval $[0,T] : 0 = t_0 < t_1 < \cdots < t_M = T$, with the time step size $\Delta t = T/M$. Denote by $\mathbf{u}_h^m$ the approximation of $\mathbf{u}_h(t_m)$ with $\mathbf{u}_h^0 = \mathbf{u}_h(0)$. Applying the Duhamel's principle to (2.43) gives us

$$\mathbf{u}_h(t_m + s) = e^{s\mathbf{A}_h}\mathbf{u}_h(t_m) + \int_0^s e^{(s-\sigma)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m+\sigma))d\sigma. \tag{2.44}$$

Setting $s = \Delta t$, we deduce that

$$\mathbf{u}_h(t_{m+1}) = e^{\Delta t\mathbf{A}_h}\mathbf{u}_h(t_m) + \int_0^{\Delta t} e^{(\Delta t-s)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m+s))ds. \tag{2.45}$$

By approximating $\mathbf{u}_h(t_m+s)$ in (2.45) with $e^{s\mathbf{A}_h}\mathbf{u}_h(t_m)$ and using one-point quadrature rules to compute the resulting integral, we obtain the following two first-order LRI schemes (the reader is referred to [45] for the details):

i) LRI1a scheme: $$\mathbf{u}_h^{m+1} = e^{\Delta t\mathbf{A}_h}\big(\mathbf{u}_h^m + \Delta t\,\mathbf{F}(\mathbf{u}_h^m)\big), \tag{2.46}$$

ii) LRI1b scheme: $$\mathbf{u}_h^{m+1} = e^{\Delta t\mathbf{A}_h}\mathbf{u}_h^m + \Delta t\,\mathbf{F}(e^{\Delta t\mathbf{A}_h}\mathbf{u}_h^m). \tag{2.47}$$

2.2.2 Optimal error estimates

We assume that *the exact solution u of the model semilinear parabolic equation* (2.42) *belongs to* $\mathcal{C}([0,T],\mathcal{C}^4(\overline{\Omega}))$, and write $u(t)$ in place of $u(\cdot,t)$ for the sake of simplicity. Denote by $\mathbf{u}_h$ the solution of the space-discrete problem (2.43) and by $z_1 \lesssim z_2$ the relation $z_1 \le Cz_2$ for some generic constant $C > 0$ independent of h and Δt.

Preliminaries

Let us first introduce some important lemmas. A key property of $\mathbf{A}_h$ is that its elements a_{ij} satisfy

$$a_{ii} < 0, \quad a_{ii} + \sum_{j=1, j\neq i}^{N} |a_{ij}| \leq 0, \quad i = 1, 2, \dots, N.$$

In other words, $\mathbf{A}_h$ is a diagonally dominant matrix with all diagonal entries negative. As a consequence, we obtain the following result from the conclusion of Lemma 3.3 in [101].

Lemma 2.16. For any $s > 0$, we have $\|e^{s\mathbf{A}_h}\|_\infty \leq 1$.

Using this lemma, the semi-discrete solution of (2.43) is shown to preserve MBP as stated below (see [115, Theorem 2] for the detailed proof).

Lemma 2.17. There exists a unique solution $\mathbf{u}_h \in \mathcal{C}([0, T], \mathbb{R}^N)$ to the semi-discrete problem (2.43) and

$$\|\mathbf{u}_h(t)\|_\infty \leq \beta, \quad \text{for all } t \in [0, T].$$

Moreover, we have the following result on the fully discrete solutions produced by the LRI1 schemes (cf. [45, Theorem 2]).

Lemma 2.18 (Discrete MBP). The fully discrete solutions produced by the LRI1a scheme (2.46) or the LRI1b scheme (2.47) preserve the discrete MBP conditionally:

$$\|\mathbf{u}_h^m\|_\infty \leq \beta, \quad m = 0, 1, \dots, M, \tag{2.48}$$

provided that $0 < \Delta t \leq \omega_0$, where $\omega_0 := -\left(\min_{|x|\leq\beta} f'(x)\right)^{-1} > 0$.

Next, let us state two versions of Gronwall's inequality for use in our error analysis later.

Lemma 2.19.

(i) If a real-value function φ is continuous on $[0, T]$ and there exist two constants $C_1, C_2 > 0$ such that

$$\varphi(t) \leq C_1 + C_2 \int_0^t \varphi(s) ds, \quad \text{for all } t \in [0, T],$$

then

$$\varphi(t) \le C_1 e^{C_2 t}$$

for all $t \in [0, T]$.

(ii) If there exist two constants $C_1, C_2 > 0$ and a nonnegative sequence $\{z_n\}_{n=0}^{\infty}$ satisfying

$$z_{n+1} \le (1 + C_1) z_n + C_2, \quad \text{for all } n \ge 0,$$

then

$$z_n \le (1 + C_1)^n \left(z_0 + \frac{C_2}{C_1} \right) - \frac{C_2}{C_1}$$

for all $n \ge 1$.

It is well known that the central finite difference method is a second-order approximation [55]. Note that periodic boundary condition is considered in [55], but the consistency result still holds true for the homogeneous Neumann boundary condition.

Lemma 2.20. $\|\mathbf{D}_h I^h u(t) - I^h \Delta u(t)\|_\infty \lesssim h^2$ for all $t \in [0, T]$.

This consistency estimate leads to the convergence of the semi-discrete solution $\mathbf{u}_h$ of (2.43) to the exact solution $I^h u$ of (2.42) as demonstrated in the lemma below.

Lemma 2.21. $\|\mathbf{u}_h(t) - I^h u(t)\|_\infty \lesssim h^2$ for all $t \in [0, T]$.

Proof. Let $\mathbf{r}(t) = \mathbf{u}_h(t) - I^h u(t)$, we observe that $\mathbf{r}(0) = 0$. From (2.42) and (2.43), we deduce that

$$\frac{d\mathbf{r}(t)}{dt} = \mathbf{A}_h \mathbf{r}(t) + \delta(t), \quad \forall\, t \in [0, T], \tag{2.49}$$

where $\delta(t) = \varepsilon^2 (\mathbf{D}_h I^h u(t) - I^h \Delta u(t)) + \mathbf{F}(\mathbf{u}_h(t)) - \mathbf{F}(I^h u(t))$. Using Lemmas 2.17, 2.20 and the Lipschitz continuity of f, we obtain

$$\begin{aligned} \|\delta(t)\|_\infty &\le \varepsilon^2 \|\mathbf{D}_h I^h u(t) - I^h \Delta u(t)\|_\infty + \|\mathbf{F}(\mathbf{u}_h(t)) - \mathbf{F}(I^h u(t))\|_\infty \\ &\lesssim h^2 + \|\mathbf{r}(t)\|_\infty, \quad \forall\, t \in [0, T]. \end{aligned} \tag{2.50}$$

Applying the Duhamel's principle for (2.49), we get

$$\mathbf{r}(t) = e^{t\mathbf{A}_h}\mathbf{r}(0) + \int_0^t e^{(t-s)\mathbf{A}_h}\boldsymbol{\delta}(s)ds = \int_0^t e^{(t-s)\mathbf{A}_h}\boldsymbol{\delta}(s)ds, \quad \forall\, t \in [0, T].$$

Taking the infinity norm of both sides, along with using Lemma 2.16 and (2.50), we arrive at

$$\begin{aligned} \|\mathbf{r}(t)\|_\infty &\le \int_0^t \|e^{(t-s)\mathbf{A}_h}\|_\infty \|\boldsymbol{\delta}(s)\|_\infty ds \\ &\lesssim \int_0^t (h^2 + \|\mathbf{r}(s)\|_\infty) ds \lesssim h^2 + \int_0^t \|\mathbf{r}(s)\|_\infty ds, \quad \forall\, t \in [0, T]. \end{aligned} \tag{2.51}$$

Noting that $\|\mathbf{r}(t)\|_\infty$ is continuous, then the proof is completed by using Lemma 2.19-(i) and (2.51). □

Based on Lemma 2.21 and the triangle inequality (cf. Equation (2.53)), there remains to bound $\|\mathbf{u}_h^m - \mathbf{u}_h(t_m)\|_\infty$ to obtain fully discrete error estimates for the LRI1a and LRI1b schemes. The most challenging part is to cancel out the factor h^{-2} under the weak regularity in time of u and $\mathbf{u}_h$ when taking the infinity norm of the matrix $\mathbf{A}_h$ ($\|\mathbf{A}_h\|_\infty \lesssim h^{-2}$, cf. Theorems 5 and 6 in [45]). This will be partly accomplished by the following lemma, which is one of the most crucial results in this section.

Lemma 2.22. The following statements are true for all $s, t \in [0, T]$:

(i) $\|\mathbf{A}_h\mathbf{u}_h(t)\|_\infty \lesssim 1$;

(ii) $\|\mathbf{u}_h(t) - \mathbf{u}_h(s)\|_\infty \lesssim |t - s|$;

(iii) If f additionally satisfies Assumption (A4), then $\|\mathbf{A}_h\mathbf{F}(\mathbf{u}_h(t))\|_\infty \lesssim 1$.

Proof. (i) Note that $\|\mathbf{A}_h I^h u(t)\|_\infty \lesssim 1$ for all $t \in [0, T]$ because of Lemma 2.20 and the boundedness of Δu. By applying Lemma 2.21, we have

$$\begin{aligned}\|\mathbf{A}_h \mathbf{u}_h(t)\|_\infty &\le \|\mathbf{A}_h I^h u(t)\|_\infty + \|\mathbf{A}_h(\mathbf{u}_h(t) - I^h u(t))\|_\infty \\ &\lesssim 1 + \|\mathbf{A}_h\|_\infty \|\mathbf{u}_h(t) - I^h u(t))\|_\infty \\ &\lesssim 1 + h^{-2}h^2 \lesssim 1, \quad \forall\, t \in [0, T].\end{aligned}$$

(ii) Suppose $s < t$ and $s, t \in [0, T]$. Using the Duhamel's principle for (2.43), we obtain

$$\mathbf{u}_h(t) = e^{(t-s)\mathbf{A}_h}\mathbf{u}_h(s) + \int_0^{t-s} e^{(t-s-\sigma)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(s+\sigma))\, d\sigma.$$

This gives us

$$\mathbf{u}_h(t) - \mathbf{u}_h(s) = [e^{(t-s)\mathbf{A}_h} - \mathbf{I}]\mathbf{u}_h(s) + \int_0^{t-s} e^{(t-s-\sigma)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(s+\sigma))\, d\sigma =: K_1 + K_2,$$

where K_1 and K_2 are given by

$$\begin{aligned}K_1 &= [e^{(t-s)\mathbf{A}_h} - \mathbf{I}]\mathbf{u}_h(s), \\ K_2 &= \int_0^{t-s} e^{(t-s-\sigma)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(s+\sigma)).\end{aligned}$$

By Lemmas 2.16 and 2.17 (the boundedness of $\mathbf{u}_h$), we have

$$\|K_2\|_\infty \le \int_0^{t-s} \|e^{(t-s-\sigma)\mathbf{A}_h}\|_\infty \|\mathbf{F}(\mathbf{u}_h(s+\sigma))\|_\infty\, d\sigma \lesssim t - s.$$

To bound $\|K_1\|_\infty$, using the fundamental theorem of calculus and the result of Part (i) yields

$$\|K_1\|_\infty = \left\| \int_0^{t-s} e^{\sigma \mathbf{A}_h}\mathbf{A}_h \mathbf{u}_h(s) d\sigma \right\|_\infty \le \int_0^{t-s} \|e^{\sigma \mathbf{A}_h}\|_\infty \|\mathbf{A}_h \mathbf{u}_h(s)\|_\infty\, d\sigma \lesssim t - s.$$

Thus, we conclude that

$$\|\mathbf{u}_h(t) - \mathbf{u}_h(s)\|_\infty \le \|K_1\|_\infty + \|K_2\|_\infty \lesssim t - s.$$

(iii) We first prove that $\|\mathbf{A}_h I^h[f(u(t))]\|_\infty \lesssim 1$ for all $t \in [0, T]$. It suffices to consider the one dimensional case (the proofs in two and three dimensions can be done in a similar manner).

Denote by $v(x,t) = f(u(x,t))$ then v_x is Lipschitz continuous on Ω with respect to spacial variable because u_x is Lipschitz continuous on the same domain and f' is Lipschitz continuous on $[-\beta,\beta]$ (cf. Assumption (A4)). Moreover, based on the MBP and the regularity of u, one can show that this Lipschitz constant is independent of t. Since u satisfies the homogeneous Neumann boundary condition, so is v, namely, $v_x(x_1,t) = v_x(x_N,t) = 0$. By the definition of $\mathbf{A}_h$ and I^h, we see that

$$\mathbf{A}_h I^h[f(u(t))] = \mathbf{A}_h I^h v(t) = \frac{\varepsilon^2}{h^2}\begin{bmatrix} -2 & 2 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \\ 0 & \cdots & 0 & 2 & -2 \end{bmatrix}\begin{bmatrix} v(x_1,t) \\ v(x_2,t) \\ \vdots \\ v(x_{N-1},t) \\ v(x_N,t) \end{bmatrix} =: \varepsilon^2 \begin{bmatrix} \alpha_1(t) \\ \alpha_2(t) \\ \vdots \\ \alpha_{N-1}(t) \\ \alpha_N(t) \end{bmatrix}. \tag{2.52}$$

For any $j = 2,3,\dots,N-1$ and $t \in [0,T]$, we have

$$\begin{aligned} \alpha_j(t) &= \frac{v(x_{j+1},t) - 2v(x_j,t) + v(x_{j-1},t)}{h^2} \\ &= \frac{[v(x_{j+1},t) - v(x_j,t)] - [v(x_j,t) - v(x_{j-1},t)]}{h^2}. \end{aligned}$$

By the mean value theorem, there exist $\gamma_1 \in (x_j, x_{j+1})$ and $\gamma_2 \in (x_{j-1}, x_j)$ such that

$$v(x_{j+1},t) - v(x_j,t) = hv_x(\gamma_1,t), \quad \text{and} \quad v(x_j,t) - v(x_{j-1},t) = hv_x(\gamma_2,t).$$

This implies that

$$|\alpha_j(t)| = \frac{|v_x(\gamma_1,t) - v_x(\gamma_2,t)|}{h} \lesssim \frac{|\gamma_1 - \gamma_2|}{h} \lesssim 1, \qquad j = 2,3,\dots,N-1.$$

where we have used the Lipschitz continuity of v_x in the second last estimate. Now, when $j = 1$ we have

$$\alpha_1(t) = \frac{2v(x_2,t) - 2v(x_1,t)}{h^2}.$$

By the mean value theorem, there exists $\gamma_3 \in (x_1, x_2)$ such that

$$|v(x_2,t) - v(x_1,t)| = h|v_x(\gamma_3,t)| = h|v_x(\gamma_3,t) - v_x(x_1,t)| \lesssim h|\gamma_3 - x_1| \leq h^2,$$

This shows that $|\alpha_1(t)| \lesssim 1$. By the same argument, we also get $|\alpha_N(t)| \lesssim 1$. The above estimates give us $\|\mathbf{A}_h I^h[f(u(t))]\|_\infty \lesssim 1$ for all $t \in [0,T]$ and therefore

$$\|\mathbf{A}_h \mathbf{F}(I^h u(t))\|_\infty = \|\mathbf{A}_h I^h[f(u(t)]\|_\infty \lesssim 1, \quad t \in [0,T].$$

Finally,

$$\begin{aligned}
\|\mathbf{A}_h \mathbf{F}(\mathbf{u}_h(t))\|_\infty &\leq \|\mathbf{A}_h \mathbf{F}(I^h u(t))\|_\infty + \|\mathbf{A}_h[\mathbf{F}(\mathbf{u}_h(t)) - \mathbf{F}(I^h u(t))]\|_\infty \\
&\lesssim 1 + \|\mathbf{A}_h\|_\infty \|\mathbf{F}(\mathbf{u}_h(t)) - \mathbf{F}(I^h u(t))\|_\infty \\
&\lesssim 1 + \|\mathbf{A}_h\|_\infty \|\mathbf{u}_h(t) - I^h u(t)\|_\infty \\
&\lesssim 1 + h^{-2} h^2 \lesssim 1, \quad t \in [0,T],
\end{aligned}$$

where we have used Lemmas 2.17, 2.21 and the Lipschitz continuity of f. □

Remark 2.23. The analysis presented above can be extended to the periodic and the homogeneous Dirichlet boundary conditions. For the periodic boundary condition, Lemma 2.22 remains valid. For the homogeneous Dirichlet boundary condition, we need to assume an extra condition on the nonlinear function f, namely $f(0) = 0$, to guarantee that $\mathbf{A}_h$ acting on $f(u)$ in (2.52) is bounded in the infinity norm, so that the result (iii) in Lemma 2.22 still holds true. Indeed, in this case, $\mathbf{A}_h$ is given by

$$\mathbf{A}_h = \frac{\varepsilon^2}{h^2} \begin{bmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \\ 0 & \cdots & 0 & 1 & -2 \end{bmatrix}_{N\times N}.$$

Denote by x_0 and x_{N+1} two endpoints of $\overline{\Omega}$ (instead of x_1 and x_N as in the homogeneous

Neumann boundary condition case). By assuming $f(0) = 0$, we have $v(x_0, t) = v(x_{N+1}, t) = 0$. Therefore,

$$\alpha_1(t) = \frac{v(x_2, t) - 2v(x_1, t)}{h^2} = \frac{v(x_2, t) - 2v(x_1, t) + v(x_0, t)}{h^2},$$

and

$$\alpha_N(t) = \frac{v(x_{N-1}, t) - 2v(x_N, t)}{h^2} = \frac{v(x_{N-1}, t) - 2v(x_N, t) + v(x_{N+1}, t)}{h^2}.$$

Thus $|\alpha_1(t)| \lesssim 1$ and $|\alpha_N(t)| \lesssim 1$ by the same arguments as above. Hence $\|\mathbf{A}_h I^h[f(u(t))]\|_\infty \lesssim 1$ for all $t \in [0, T]$, and (iii) in Lemma 2.22 holds. It should be noted that both the double-well and Flory-Huggins potential functions satisfy the extra condition $f(0) = 0$.

Main results

From now on, f is assumed to satisfy additionally Assumption (A4) so that the result (iii) in Lemma 2.22 holds. For any vector $\mathbf{v} = (v_1, v_2, \dots, v_N)^T \in \mathbb{R}^N$, we denote by

$$\frac{\partial \mathbf{F}}{\partial \mathbf{u}}(\mathbf{v}) = \text{diag}(f'(v_1), f'(v_2), \dots, f'(v_N)).$$

Then we have the following optimal error estimates for the fully discrete LRI1a scheme (2.46).

Theorem 2.24 (Error estimate of LRI1a). Let $\{\mathbf{u}_h^m\}$ be generated by the LRI1a scheme (2.46) with $\mathbf{u}_h^0 = \mathbf{u}_h(0)$, then

$$\|\mathbf{u}_h^m - I^h u(t_m)\|_\infty \lesssim h^2 + \Delta t, \quad \forall\, m = 0, 1, \dots, M,$$

provided that $0 < \Delta t \le \omega_0$.

Proof. By the triangle inequality, we can see that

$$\|\mathbf{u}_h^m - I^h u(t_m)\|_\infty \le \|\mathbf{u}_h^m - \mathbf{u}_h(t_m)\|_\infty + \|\mathbf{u}_h(t_m) - I^h u(t_m)\|_\infty. \tag{2.53}$$

The second term $\|\mathbf{u}_h(t_m) - I^h u(t_m)\|_\infty \lesssim h^2$ due to Lemma 2.21, thus it is sufficient to show that

$$\|\mathbf{u}_h^m - \mathbf{u}_h(t_m)\|_\infty \lesssim \Delta t. \tag{2.54}$$

Denote by $\mathbf{e}_m = \mathbf{u}_h^m - \mathbf{u}_h(t_m)$, we first observe that $\mathbf{e}_0 = 0$. Subtracting (2.45) from (2.46), we have

$$\mathbf{e}_{m+1} = e^{\Delta t \mathbf{A}_h}\mathbf{e}_m + \Delta t e^{\Delta t \mathbf{A}_h}(\mathbf{F}(\mathbf{u}_h^m) - \mathbf{F}(\mathbf{u}_h(t_m))) + \mathbf{R}_m^{(1)},$$

where

$$\mathbf{R}_m^{(1)} = \Delta t e^{\Delta t \mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m)) - \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m + s))ds.$$

Using Lemmas 2.16, 2.17 and 2.18 (the boundedness of $\mathbf{u}_h$ and $\mathbf{u}_h^m$) and the Lipschitz continuity of f, we obtain

$$\begin{aligned}
\|\mathbf{e}_{m+1}\|_\infty &\le \|\mathbf{e}_m\|_\infty + \Delta t\|\mathbf{F}(\mathbf{u}_h^m) - \mathbf{F}(\mathbf{u}_h(t_m))\|_\infty + \|\mathbf{R}_m^{(1)}\|_\infty \\
&\le (1 + C_1\Delta t)\|\mathbf{e}_m\|_\infty + \|\mathbf{R}_m^{(1)}\|_\infty,
\end{aligned} \tag{2.55}$$

where $C_1 = \max_{|x|\le\beta}|f'(x)|$. Rewrite $\mathbf{R}_m^{(1)}$ as follows

$$\begin{aligned}
\mathbf{R}_m^{(1)} &= \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}_h}[e^{s\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m)) - \mathbf{F}(\mathbf{u}_h(t_m + s))]ds \\
&= \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}_h}[e^{s\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m)) - \mathbf{F}(e^{s\mathbf{A}_h}\mathbf{u}_h(t_m))]ds \\
&\quad + \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}_h}[\mathbf{F}(e^{s\mathbf{A}_h}\mathbf{u}_h(t_m)) - \mathbf{F}(\mathbf{u}_h(t_m + s))]ds =: I_1 + I_2,
\end{aligned}$$

where I_1 and I_2 are given by

$$\begin{aligned}
I_1 &= \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}_h}[e^{s\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m)) - \mathbf{F}(e^{s\mathbf{A}_h}\mathbf{u}_h(t_m))]ds, \\
I_2 &= \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}_h}[\mathbf{F}(e^{s\mathbf{A}_h}\mathbf{u}_h(t_m)) - \mathbf{F}(\mathbf{u}_h(t_m + s))]ds.
\end{aligned}$$

In order to estimate $\|I_2\|_\infty$, we use Lemmas 2.16 and 2.17, the Lipschitz continuity and the boundedness of f, the Duhamel's formula (2.44) to get

$$\begin{aligned}\|I_2\|_\infty &\lesssim \int_0^{\Delta t} \|e^{(\Delta t-s)\mathbf{A}_h}\|_\infty \|\mathbf{F}(e^{s\mathbf{A}_h}\mathbf{u}_h(t_m)) - \mathbf{F}(\mathbf{u}_h(t_m+s))\|_\infty ds \\ &\lesssim \int_0^{\Delta t} \|e^{s\mathbf{A}_h}\mathbf{u}_h(t_m) - \mathbf{u}_h(t_m+s)\|_\infty ds = \int_0^{\Delta t} \left\| \int_0^s e^{(s-\sigma)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m+\sigma))d\sigma \right\|_\infty ds \\ &\le \int_0^{\Delta t} \left(\int_0^s \|e^{(s-\sigma)\mathbf{A}_h}\|_\infty \|\mathbf{F}(\mathbf{u}_h(t_m+\sigma))\|_\infty d\sigma \right) ds \lesssim \int_0^{\Delta t} s ds \lesssim \Delta t^2. \end{aligned} \tag{2.56}$$

On the other hand, define $\mathbf{G}(s) = e^{s\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m)) - \mathbf{F}(e^{s\mathbf{A}_h}\mathbf{u}_h(t_m))$, clearly $\mathbf{G}(0) = 0$. Therefore,

$$\mathbf{G}(s) = \mathbf{G}(s) - \mathbf{G}(0) = \int_0^s \mathbf{G}'(\sigma)d\sigma,$$

where

$$\mathbf{G}'(\sigma) = e^{\sigma\mathbf{A}_h}\mathbf{A}_h\mathbf{F}(\mathbf{u}_h(t_m)) - \frac{\partial \mathbf{F}}{\partial \mathbf{u}}(e^{s\mathbf{A}_h}\mathbf{u}_h(t_m))e^{\sigma\mathbf{A}_h}\mathbf{A}_h\mathbf{u}_h(t_m).$$

By Lemma 2.22, $\|\mathbf{A}_h\mathbf{F}(\mathbf{u}_h(t_m))\|_\infty \lesssim 1$ and $\|\mathbf{A}_h\mathbf{u}_h(t_m)\|_\infty \lesssim 1$, so $\|\mathbf{G}'(\sigma)\|_\infty \lesssim 1$. Hence,

$$\|\mathbf{G}(s)\|_\infty \le \int_0^s \|\mathbf{G}'(\sigma)\|_\infty d\sigma \lesssim s,$$

which implies

$$\|I_1\|_\infty \le \int_0^{\Delta t} \|e^{(\Delta t-s)\mathbf{A}_h}\|_\infty \|\mathbf{G}(s)\|_\infty ds \lesssim \int_0^{\Delta t} s ds \lesssim \Delta t^2. \tag{2.57}$$

Combining (2.56) and (2.57) gives us $\|\mathbf{R}_m^{(1)}\|_\infty \lesssim \Delta t^2$, or $\|\mathbf{R}_m^{(1)}\|_\infty \le C_2\Delta t^2$ for some $C_2 > 0$. Thus, from (2.55) we further deduce that

$$\|\mathbf{e}_{m+1}\|_\infty \le (1 + C_1\Delta t)\|\mathbf{e}_m\|_\infty + C_2\Delta t^2.$$

The inequality (2.54) is then obtained by using the discrete Gronwall's inequality in Lemma 2.19-(ii), and the proof is completed. □

Following a similar idea as in the proof of the above theorem, we also establish the optimal error estimate for the fully discrete LRI1b scheme.

Theorem 2.25 (Error estimate of LRI1b). Let $\{\mathbf{u}_h^m\}$ be generated by the LRI1b scheme (2.47) with $\mathbf{u}_h^0 = \mathbf{u}_h(0)$, then

$$\|\mathbf{u}_h^m - I^h u(t_m)\|_\infty \lesssim h^2 + \Delta t, \quad \forall\, m = 0, 1, \dots, M,$$

provided that $0 < \Delta t \le \omega_0$.

Proof. As done in Theorem 2.24, the main step is again to prove $\|\mathbf{u}_h^m - \mathbf{u}_h(t_m)\|_\infty \lesssim \Delta t$. Subtracting (2.45) from (2.47) yields

$$\mathbf{e}_{m+1} = e^{\Delta t \mathbf{A}_h}\mathbf{e}_m + \Delta t[\mathbf{F}(e^{\Delta t \mathbf{A}_h}\mathbf{u}_h^m) - \mathbf{F}(e^{\Delta t \mathbf{A}_h}\mathbf{u}_h(t_m))] + \mathbf{R}_m^{(2)},$$

where $\mathbf{R}_m^{(2)} = \Delta t \mathbf{F}(e^{\Delta t \mathbf{A}_h}\mathbf{u}_h(t_m)) - \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m + s))ds$. Thus we deduce that

$$\begin{aligned}
\|\mathbf{e}_{m+1}\|_\infty &\le \|\mathbf{e}_m\|_\infty + \Delta t\|\mathbf{F}(e^{\Delta t \mathbf{A}_h}\mathbf{u}_h^m) - \mathbf{F}(e^{\Delta t \mathbf{A}_h}\mathbf{u}_h(t_m))\|_\infty + \|\mathbf{R}_m^{(2)}\|_\infty \\
&\le \|\mathbf{e}_m\|_\infty + C_1\Delta t\|e^{\Delta t \mathbf{A}_h}(\mathbf{u}_h^m - \mathbf{u}_h(t_m))\|_\infty + \|\mathbf{R}_m^{(2)}\|_\infty \\
&\le (1 + C_1\Delta t)\|\mathbf{e}_m\|_\infty + \|\mathbf{R}_m^{(2)}\|_\infty, \qquad (2.58)
\end{aligned}$$

where C_1 is the same constant as in (2.55). We then rewrite $\mathbf{R}_m^{(2)}$ as follows:

$$\begin{aligned}
\mathbf{R}_m^{(2)} &= \Delta t[\mathbf{F}(e^{\Delta t \mathbf{A}_h}\mathbf{u}_h(t_m)) - \mathbf{F}(\mathbf{u}_h(t_{m+1}))] + \int_0^{\Delta t} \mathbf{F}(\mathbf{u}_h(t_{m+1})) - e^{(\Delta t - s)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m + s))ds \\
&= \Delta t[\mathbf{F}(e^{\Delta t \mathbf{A}_h}\mathbf{u}_h(t_m)) - \mathbf{F}(\mathbf{u}_h(t_{m+1}))] + \int_0^{\Delta t} \mathbf{F}(\mathbf{u}_h(t_{m+1})) - \mathbf{F}(\mathbf{u}_h(t_m + s))ds \\
&\quad + \int_0^{\Delta t} (\mathbf{I} - e^{(\Delta t - s)\mathbf{A}_h})\mathbf{F}(\mathbf{u}_h(t_m + s)))ds \\
&=: J_1 + J_2 + J_3,
\end{aligned}$$

where J_1, J_2 and J_3 are given by

$$\begin{aligned}
J_1 &= \Delta t[\mathbf{F}(e^{\Delta t\mathbf{A}_h}\mathbf{u}_h(t_m)) - \mathbf{F}(\mathbf{u}_h(t_{m+1}))],\\
J_2 &= \int_0^{\Delta t} \mathbf{F}(\mathbf{u}_h(t_{m+1})) - \mathbf{F}(\mathbf{u}_h(t_m+s))ds,\\
J_3 &= \int_0^{\Delta t} (\mathbf{I} - e^{(\Delta t - s)\mathbf{A}_h})\mathbf{F}(\mathbf{u}_h(t_m+s)))ds.
\end{aligned}$$

By applying the Duhamel's principle, we have

$$\begin{aligned}
\|J_1\|_\infty &\lesssim \Delta t\|e^{\Delta t\mathbf{A}_h}\mathbf{u}_h(t_m) - \mathbf{u}_h(t_{m+1})\|_\infty\\
&= \Delta t\left\|\int_0^{\Delta t} e^{(\Delta t-\sigma)\mathbf{A}_h}\mathbf{F}(\mathbf{u}_h(t_m+\sigma))d\sigma\right\|_\infty \lesssim \Delta t^2. \qquad (2.59)
\end{aligned}$$

For $\|J_2\|_\infty$, we use the Lipschitz continuity of f and $\mathbf{u}_h$ mentioned in Lemma 2.22 to obtain

$$\begin{aligned}
\|J_2\|_\infty &\le \int_0^{\Delta t} \|\mathbf{F}(\mathbf{u}_h(t_{m+1})) - \mathbf{F}(\mathbf{u}_h(t_m+s))\|_\infty ds\\
&\lesssim \int_0^{\Delta t} \|\mathbf{u}_h(t_{m+1}) - \mathbf{u}_h(t_m+s)\|_\infty ds\\
&\lesssim \int_0^{\Delta t} [t_{m+1} - (t_m+s)]ds = \int_0^{\Delta t} (\Delta t - s)ds \lesssim \Delta t^2. \qquad (2.60)
\end{aligned}$$

Note that

$$J_3 = -\int_0^{\Delta t}\left(\int_0^{\Delta t - s} e^{\sigma\mathbf{A}_h}\mathbf{A}_h\mathbf{F}(\mathbf{u}_h(t_m+s))d\sigma\right)ds.$$

Now using the fact that $\|\mathbf{A}_h\mathbf{F}(\mathbf{u}_h(t))\|_\infty \lesssim 1$ for all $t \in [0, T]$ (cf. Lemma 2.22), we deduce that

$$\begin{aligned}
\|J_3\|_\infty &\le \int_0^{\Delta t}\left(\int_0^{\Delta t - s} \|e^{\sigma\mathbf{A}_h}\|_\infty\|\mathbf{A}_h\mathbf{F}(\mathbf{u}_h(t_m+s))\|_\infty d\sigma\right)ds\\
&\lesssim \int_0^{\Delta t} (\Delta t - s)ds \lesssim \Delta t^2. \qquad (2.61)
\end{aligned}$$

From (2.59), (2.60) and (2.61), we obtain $\|\mathbf{R}_m^{(2)}\|_\infty \lesssim \Delta t^2$, or $\|\mathbf{R}_m^{(2)}\|_\infty \le C_3\Delta t^2$ for some $C_3 > 0$. Thus it is implied from (2.58) that

$$\|\mathbf{e}_{m+1}\|_\infty \le (1 + C_1\Delta t)\|\mathbf{e}_m\|_\infty + C_3\Delta t^2,$$

which completes the proof after applying the discrete Gronwall's inequality in Lemma 2.19-(ii). $\square$

2.2.3 Numerical results

We verify the error estimates in both space and time of the proposed first-order LRI schemes and compare their performance with the first-order ETD (ETD1) scheme. The benchmark test [116] with a traveling wave solution is considered, where the nonlinear function is $f(u) = u - u^3$ corresponding to the double-well potential function and the spatial domain is $\Omega = (-0.5, 0.5)^2$. Homogeneous Neumann conditions are imposed on the boundary and the initial condition takes the following form

$$u_0(x, y) = \frac{1}{2}\left(1 - \tanh\left(\frac{x}{2\sqrt{2}\varepsilon}\right)\right), \quad (x, y) \in \overline{\Omega}.$$

An approximate exact solution (for $\varepsilon \ll 1$) to the initial-boundary value problem is then given by

$$u(x, y, t) = \frac{1}{2}\left(1 - \tanh\left(\frac{x - st}{2\sqrt{2}\varepsilon}\right)\right), \quad (x, y) \in \overline{\Omega},\ t \geq 0,$$

where $s = 3\varepsilon/\sqrt{2}$ denotes the speed of the traveling wave. We fix $T = 1/4s$ and vary both h and Δt simultaneously so that the ratio between h and Δt is fixed. The L^2 and L^∞ errors of ETD1, LRI1a and LRI1b schemes are computed at the final time T for different values of $\varepsilon \in \{0.01, 0.005\}$. The results are reported in Table 2.5 (for L^2 errors) and Table 2.6 (for L^∞ errors) with corresponding convergence rates. We first observe that all methods, particularly LRI1a and LRI1b, have the overall first-order convergence as proved in the previous section. The errors produced by LRI1a and LRI1b are smaller than those by ETD1, meaning that the proposed LRI schemes are more accurate than ETD1. Moreover, the LRI schemes have better convergence rates than the ETD scheme, especially when ε is sufficiently small. More comparisons regarding the second-order ETD and LRI schemes can by found in [45].

Table 2.5: L^2 errors and convergence rates by the first-order ETD and LRI schemes

ε	h	Δt	ETD1		LRI1a		LRI1b	
			L^2 error	Rate	L^2 error	Rate	L^2 error	Rate
0.01	1/64	$T/32$	1.00e-01		7.36e-02		6.65e-02	
	1/128	$T/64$	6.03e-02	0.73	4.40e-02	0.74	4.01e-02	0.73
	1/256	$T/128$	3.31e-02	0.86	2.41e-02	0.87	2.21e-02	0.86
	1/512	$T/256$	1.73e-02	0.93	1.26e-02	0.94	1.16e-02	0.93
	1/1024	$T/512$	8.87e-03	0.97	6.44e-03	0.97	5.91e-03	0.97
	1/2048	$T/1024$	4.48e-03	0.98	3.25e-03	0.98	2.99e-03	0.98
0.005	1/64	$T/32$	1.63e-01		1.13e-01		1.03e-01	
	1/128	$T/64$	1.27e-01	0.36	8.99e-02	0.33	8.54e-02	0.27
	1/256	$T/128$	8.32e-02	0.61	5.84e-02	0.62	5.60e-02	0.61
	1/512	$T/256$	4.76e-02	0.80	3.32e-02	0.82	3.18e-02	0.81
	1/1024	$T/512$	2.53e-02	0.91	1.76e-02	0.91	1.69e-02	0.91
	1/2048	$T/1024$	1.30e-02	0.96	9.07e-03	0.96	8.69e-03	0.96

Table 2.6: L^∞ errors and convergence rates by the first-order ETD and LRI schemes

ε	h	Δt	ETD1		LRI1a		LRI1b	
			L^∞ error	Rate	L^∞ error	Rate	L^∞ error	Rate
0.01	1/64	$T/32$	5.02e-01		3.73e-01		3.39e-01	
	1/128	$T/64$	3.14e-01	0.68	2.30e-01	0.69	2.13e-01	0.67
	1/256	$T/128$	1.74e-01	0.85	1.27e-01	0.86	1.17e-01	0.86
	1/512	$T/256$	9.10e-02	0.93	6.65e-02	0.93	6.16e-02	0.93
	1/1024	$T/512$	4.65e-02	0.97	3.40e-02	0.97	3.14e-02	0.97
	1/2048	$T/1024$	2.35e-02	0.99	1.72e-02	0.98	1.59e-02	0.99
0.005	1/64	$T/32$	8.60e-01		6.95e-01		6.54e-01	
	1/128	$T/64$	7.79e-01	0.14	5.88e-01	0.24	5.78e-01	0.18
	1/256	$T/128$	5.67e-01	0.46	4.11e-01	0.52	3.97e-01	0.54
	1/512	$T/256$	3.42e-01	0.73	2.42e-01	0.76	2.34e-01	0.77
	1/1024	$T/512$	1.85e-01	0.88	1.30e-01	0.90	1.25e-01	0.90
	1/2048	$T/1024$	9.57e-02	0.95	6.69e-02	0.96	6.43e-02	0.96

2.3 Low regularity integrators for the conservative Allen-Cahn equation with a nonlocal constraint

We consider the conservative Allen-Cahn equation with a nonlocal Lagrange multiplier:

$$\partial_t u = \varepsilon^2 \Delta u + \overline{f}(u), \quad \mathbf{x} \in \Omega,\ t > 0, \tag{2.62}$$

subject to the initial condition $u(\mathbf{x}, 0) = u_0(\mathbf{x})$ and the periodic or homogeneous Neumann boundary conditions. In (2.62), $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ is an open and bounded Lipschitz domain, $\varepsilon > 0$ is the interfacial parameter, $f : \mathbb{R} \to \mathbb{R}$ is a nonlinear function, and $\overline{f}$ is the modified nonlinear function defined by

$$\overline{f}(u) := f(u) - \frac{1}{|\Omega|} \int_\Omega f(u(\mathbf{y}))\, d\mathbf{y}. \tag{2.63}$$

The nonlocal term $\overline{f}$ enforces mass conservation, i.e., $\frac{d}{dt} \int_\Omega u(\mathbf{x}, t)\, d\mathbf{x} = 0$, while the solution still satisfies the maximum bound principle and energy dissipation law.

2.3.1 Low regularity integrator schemes

For a fixed $T > 0$, consider a uniform partition of the time interval $[0, T]$ with $0 = t_0 < t_1 < \dots < t_M = T$ and the time step size $\Delta t := T/M$. Throughout the section, we write $u(t)$ in place of $u(\cdot, t)$ and denote by u^m the approximation of $u(t_m)$. We also write $\int_\Omega v$ instead of $\int_\Omega v(\mathbf{y})\, d\mathbf{y}$ when there is no confusion. By applying Duhamel's principle to (2.62), there holds

$$u(t_{m+1}) = e^{\Delta t \varepsilon^2 \Delta} u(t_m) + \int_0^{\Delta t} e^{(\Delta t - s)\varepsilon^2 \Delta} \overline{f}(u(t_m + s))\, ds. \tag{2.64}$$

We also have from (2.62) and Duhamel's principle that

$$u(t_m + s) = e^{s\varepsilon^2 \Delta} u(t_m) + \int_0^{s} e^{(s - \sigma)\varepsilon^2 \Delta} \overline{f}(u(t_m + \sigma))\, d\sigma. \tag{2.65}$$

By approximating $u(t_m + s)$ in (2.64) by the first term $e^{s\varepsilon^2 \Delta} u(t_m)$ in (2.65) and using one-point quadrature rules as in [45], we obtain the following first-order LRI schemes, namely the LRI1a scheme:

$$u^{m+1} = e^{\Delta t \varepsilon^2 \Delta}(u^m + \Delta t \overline{f}(u^m)), \tag{2.66}$$

and the LRI1b scheme:

$$u^{m+1} = e^{\Delta t \varepsilon^2 \Delta} u^m + \Delta t \overline{f}(e^{\Delta t \varepsilon^2 \Delta} u^m). \tag{2.67}$$

To construct a second-order LRI scheme, we assume that f is a $\mathcal{C}^1$ function. First we approximate $u(t_m + s)$ in (2.64) by $e^{s\varepsilon^2\Delta} u(t_m) + s\overline{f}(u(t_m))$ from (2.65) and obtain

$$u^{m+1} \approx e^{\Delta t \varepsilon^2 \Delta} u^m + \int_0^{\Delta t} e^{(\Delta t - s)\varepsilon^2 \Delta} \overline{f}(e^{s\varepsilon^2\Delta} u^m + s\overline{f}(u^m))\, ds. \tag{2.68}$$

Then by the definition of $\overline{f}$ and Taylor expansion, we observe that

$$\begin{aligned}
&\overline{f}(e^{s\varepsilon^2\Delta} u^m + s\overline{f}(u^m)) = f(e^{s\varepsilon^2\Delta} u^m + s\overline{f}(u^m)) - \frac{1}{|\Omega|}\int_\Omega f(e^{s\varepsilon^2\Delta} u^m + s\overline{f}(u^m)) \\
&\approx f(e^{s\varepsilon^2\Delta} u^m) + s f'(e^{s\varepsilon^2\Delta} u^m)\overline{f}(u^m) - \frac{1}{|\Omega|}\int_\Omega \Big(f(e^{s\varepsilon^2\Delta} u^m) + s f'(e^{s\varepsilon^2\Delta} u^m)\overline{f}(u^m) \Big) \\
&= \overline{f}(e^{s\varepsilon^2\Delta} u^m) + s\left[f'(e^{s\varepsilon^2\Delta} u^m)\overline{f}(u^m) - \frac{1}{|\Omega|}\int_\Omega f'(e^{s\varepsilon^2\Delta} u^m)\overline{f}(u^m) \right].
\end{aligned} \tag{2.69}$$

Combining (2.68) and (2.69) yields

$$\begin{aligned}
u^{m+1} &\approx e^{\Delta t \varepsilon^2 \Delta} u^m + \int_0^{\Delta t} e^{(\Delta t - s)\varepsilon^2 \Delta} \overline{f}(e^{s\varepsilon^2\Delta} u^m)\, ds \\
&\quad + \int_0^{\Delta t} s e^{(\Delta t - s)\varepsilon^2 \Delta} \left[f'(e^{s\varepsilon^2\Delta} u^m)\overline{f}(u^m) - \frac{1}{|\Omega|}\int_\Omega f'(e^{s\varepsilon^2\Delta} u^m)\overline{f}(u^m) \right] ds.
\end{aligned}$$

By using the trapezoidal rule for the second term and setting $s = 0$ in all exponents of the third term on the right hand side of the above approximation, the second-order LRI scheme, namely the LRI2 scheme, is given by

$$u^{m+1} = e^{\Delta t \varepsilon^2 \Delta} u^m + \frac{\Delta t}{2}\left[e^{\Delta t \varepsilon^2 \Delta}\overline{f}(u^m) + \overline{f}(e^{\Delta t \varepsilon^2 \Delta} u^m) \right] + \frac{\Delta t^2}{2} e^{\Delta t \varepsilon^2 \Delta} \overline{g}(u^m). \tag{2.70}$$

Here $\overline{g}$ is defined as $\overline{g}(u^m) = g(u^m) - \dfrac{1}{|\Omega|}\displaystyle\int_\Omega g(u^m)$, where $g := f'\overline{f}$.

We remark that, due to the nonlocal Lagrange term in $\overline{f}$, the formulations of the LRI schemes, especially LRI2, for the conservative Allen-Cahn equation are more complicated than those for the classical Allen-Cahn equation [45]. Consequently, the analysis presented in the following is more challenging and significantly different from [45, 46].

2.3.2 Mass conservation and discrete MBP

We demonstrate in this section the unconditional mass conservation and the conditional MBP of the first- and second-order LRI schemes. The MBP results are of significant importance for our subsequent analysis.

Mass conservation For any two functions $v_1, v_2 \in L^2(\Omega)$, we denote by

$$(v_1, v_2) = \int_\Omega v_1(\mathbf{x}) v_2(\mathbf{x})\, d\mathbf{x}$$

the standard L^2 inner product of v_1 and v_2. The lemma below states a fundamental characteristic of the heat equation under suitable boundary conditions.

Lemma 2.26. For any $v \in L^2(\Omega)$, the exponential of the Laplace operator with the periodic or homogeneous Neumann boundary conditions has the following property

$$(e^{t\Delta} v, 1) = (v, 1), \quad \forall\, t > 0.$$

Proof. Fix $v \in L^2(\Omega)$ and consider the following heat equation:

$$\begin{cases} w_t = \Delta w, & \mathbf{x} \in \Omega, t > 0, \\ w(\mathbf{x}, 0) = v(\mathbf{x}), & \mathbf{x} \in \overline{\Omega}, \end{cases}$$

subject to the periodic or homogeneous Neumann boundary conditions. Taking the L^2 inner product with 1 on both sides of the above equation yields

$$\frac{d}{dt}(w, 1) = (\Delta w, 1) = 0, \quad t > 0.$$

This means that $(w(t), 1)$ is a constant for any $t > 0$. Since $w(t) \to w(0) = v$ almost everywhere as $t \to 0^+$, we deduce that $(w(t), 1) = (v, 1)$ for all $t > 0$, which completes the proof with the notice that $w(t) = e^{t\Delta}w(0) = e^{t\Delta}v$. $\square$

Theorem 2.27. The first-order LRI schemes (2.66) and (2.67) conserve mass unconditionally, i.e., for any time step size $\Delta t > 0$, we have

$$(u^m, 1) = (u^0, 1), \quad m = 1, 2, \dots, M.$$

Proof. We first consider the LRI1a scheme (2.66). By taking the L^2 inner product with 1 on both sides of (2.66), we arrive at

$$(u^{m+1}, 1) = (e^{\Delta t \varepsilon^2 \Delta}(u^m + \Delta t \overline{f}(u^m)), 1). \tag{2.71}$$

From Lemma 2.26 and (2.71), we have

$$(u^{m+1}, 1) = (u^m + \Delta t \overline{f}(u^m), 1) = (u^m, 1),$$

since $(\overline{f}(u^m), 1) = 0$ (cf. (2.63)). Similarly, from the LRI1b scheme (2.67) and Lemma 2.26, we obtain

$$\begin{aligned}(u^{m+1}, 1) &= (e^{\Delta t \varepsilon^2 \Delta}u^m + \Delta t \overline{f}(e^{\Delta t \varepsilon^2 \Delta}u^m), 1) \\ &= (e^{\Delta t \varepsilon^2 \Delta}u^m, 1) + \Delta t(\overline{f}(e^{\Delta t \varepsilon^2 \Delta}u^m), 1) = (u^m, 1),\end{aligned}$$

which concludes the proof. $\square$

Theorem 2.28. The LRI2 scheme (2.70) conserves mass unconditionally, i.e., for any time step size $\Delta t > 0$, we have

$$(u^m, 1) = (u^0, 1), \quad m = 1, 2, \dots, M.$$

Proof. Taking the L^2 inner product with 1 on both sides of (2.70) yields

$$\begin{aligned}(u^{m+1}, 1) = {}& \frac{1}{2}(e^{\Delta t \varepsilon^2 \Delta}(u^m + \Delta t \overline{f}(u^m)), 1) + \frac{1}{2}(e^{\Delta t \varepsilon^2 \Delta}u^m + \Delta t \overline{f}(e^{\Delta t \varepsilon^2 \Delta}u^m), 1) \\ & + \frac{\Delta t^2}{2}(e^{\Delta t \varepsilon^2 \Delta}\overline{g}(u^m), 1).\end{aligned} \tag{2.72}$$

Based on the proof of Theorem 2.27, we observe that

$$(e^{\Delta t\varepsilon^2\Delta}(u^m+\Delta t\overline{f}(u^m)),1)=(u^m,1)=(e^{\Delta t\varepsilon^2\Delta}u^m+\Delta t\overline{f}(e^{\Delta t\varepsilon^2\Delta}u^m),1). \tag{2.73}$$

We now apply Lemma 2.26 with $v=\overline{g}(u^m)$ to obtain

$$(e^{\Delta t\varepsilon^2\Delta}\overline{g}(u^m),1)=(\overline{g}(u^m),1)=0. \tag{2.74}$$

Combining (2.72), (2.73) and (2.74) gives us

$$(u^{m+1},1)=\frac{1}{2}(u^m,1)+\frac{1}{2}(u^m,1)=(u^m,1).$$

The proof is thus completed. □

Discrete MBP We first present some preliminary results that are necessary to prove the discrete MBP as well as the error estimates and energy stability in the next sections. The following property of the Laplace operator from [114, Lemma 2.1] plays a crucial role in our analysis throughout the section.

Lemma 2.29. The Laplace operator with the periodic or homogeneous Neumann boundary conditions generates a contraction semigroup $\{e^{t\Delta}\}_{t\geq 0}$ with respect to the supremum norm on $\mathcal{C}(\overline{\Omega})$. Moreover,

$$\|e^{t\Delta}v\|_{L^\infty}\leq\|v\|_{L^\infty},\quad\forall\, t\geq 0,\ v\in\mathcal{C}(\overline{\Omega}),$$

where $\|v\|_{L^\infty}=\max_{\mathbf{x}\in\overline{\Omega}}|v(\mathbf{x})|$.

Similar to [114, 203], we impose the following assumption on f:

(A1) There exists $\beta>0$ such that $f(\beta)\leq f(w)\leq f(-\beta)$ for all $w\in[-\beta,\beta]$.

Under Assumption (A1) and Lemma 2.29, solution to the original equation (2.62) satisfies the MBP as presented in [114, Theorem 3.1]. Let us denote $\Omega_T=\Omega\times(0,T)$.

Lemma 2.30. If f satisfies Assumption (A1) and $u_0\in\mathcal{C}(\overline{\Omega})$ with $\|u_0\|_{L^\infty}\leq\beta$, then the conservative Allen-Cahn equation (2.62) admits a unique solution $u\in\mathcal{C}(\overline{\Omega_T})$, and moreover,

$\|u(t)\|_{L^\infty} \le \beta$ for all $t \in [0, T]$.

In order to obtain the discrete MBP results, for the first-order LRI schemes we further assume that

(A2) f is continuously differentiable and nonconstant on $[-\beta, \beta]$.

For the second-order LRI scheme, a higher regularity assumption on f is required as follows:

(A3) f is twice continuously differentiable and nonconstant on $[-\beta, \beta]$.

When (A1) and (A2) hold, we define the constants

$$\tau_0 := \max_{|\xi|\le\beta} f'(\xi), \text{ and } \tau_1 := -\left(\min_{|\xi|\le\beta} f'(\xi)\right)^{-1}.$$

Clearly, $\tau_1 > 0.$ In addition, whenever $f \in \mathcal{C}^k([-\beta, \beta])$ for some non-negative integer k, we denote by

$$\alpha_k := \max_{|\xi|\le\beta} |f^{(k)}(\xi)|. \tag{2.75}$$

The technical lemma below is fundamental in establishing the MBP for LRI schemes.

Lemma 2.31. If f satisfies Assumption (A1), then for any $v \in \mathcal{C}(\overline{\Omega})$ with $\|v\|_{L^\infty} \le \beta$, we have

$$\|v + \omega \overline{f}(v)\|_{L^\infty} \le \beta, \quad \forall\, \omega \in (0, \tau_1].$$

Proof. Denote by $f_0(\xi) = \xi + \omega f(\xi)$ for $\xi \in \mathbb{R}$. If $|\xi| \le \beta$ then for all $\omega \in (0, \tau_1]$, we find that

$$f_0'(\xi) = 1 + \omega f'(\xi) \ge 1 + \omega \min_{|\xi|\le\beta} f'(\xi) \ge 1 + \tau_1 \min_{|\xi|\le\beta} f'(\xi) = 0.$$

This shows that f_0 is non-decreasing on $[-\beta, \beta]$, which implies

$$-\beta + \omega f(-\beta) \le f_0(\xi) \le \beta + \omega f(\beta), \quad \forall\, \xi \in [-\beta, \beta].$$

Since $\|v\|_{L^\infty} \le \beta$, that is, $|v(\mathbf{x})| \le \beta$ for all $\mathbf{x} \in \overline{\Omega}$, we obtain

$$-\beta + \omega f(-\beta) \le f_0(v(\mathbf{x})) \le \beta + \omega f(\beta), \quad \forall\, \mathbf{x} \in \overline{\Omega}. \tag{2.76}$$

On the other hand, by the mean value theorem and Assumption (A1), we have

$$f(\beta) \leq \frac{1}{|\Omega|}\int_{\Omega} f(v) \leq f(-\beta). \tag{2.77}$$

Combining (2.76) and (2.77) yields

$$-\beta \leq v(\mathbf{x}) + \omega\left(f(v(\mathbf{x})) - \frac{1}{|\Omega|}\int_{\Omega} f(v)\right) \leq \beta, \quad \forall\, \mathbf{x} \in \overline{\Omega}.$$

This completes the proof of the lemma. □

Theorem 2.32. Suppose that $u_0 \in \mathcal{C}(\overline{\Omega})$ with $\|u_0\|_{L^\infty} \leq \beta$ and f satisfies Assumptions (A1) and (A2), then the numerical solution $\{u^m\}_{0\leq m\leq M}$ given by the LRI1a scheme (2.66) or the LRI1b scheme (2.67) is also bounded by β in the supremum norm:

$$\|u^m\|_{L^\infty} \leq \beta, \quad m = 0, 1, \dots, M,$$

provided that $0 < \Delta t \leq \tau_1$.

Proof. It suffices to show that if $\|u^m\|_{L^\infty} \leq \beta$ then $\|u^{m+1}\|_{L^\infty} \leq \beta$. Indeed, by taking the supremum norm on both sides of the LRI1a scheme, we deduce from Lemmas 2.29 and 2.31 that

$$\|u^{m+1}\|_{L^\infty} = \|e^{\Delta t\varepsilon^2\Delta}(u^m + \Delta t\overline{f}(u^m))\|_{L^\infty} \leq \|u^m + \Delta t\overline{f}(u^m)\|_{L^\infty} \leq \beta.$$

Similarly, regarding the LRI1b scheme, we have

$$\|u^{m+1}\|_{L^\infty} = \|e^{\Delta t\varepsilon^2\Delta}u^m + \Delta t\overline{f}(e^{\Delta t\varepsilon^2\Delta}u^m)\|_{L^\infty} \leq \beta,$$

because of Lemma 2.31 and the fact that $\|e^{\Delta t\varepsilon^2\Delta}u^m\|_{L^\infty} \leq \|u^m\|_{L^\infty} \leq \beta$. □

The LRI2 scheme also preserves the MBP conditionally as stated in the next theorem. Note that under Assumption (A3), the constant α_2 (cf. (2.75)) is well-defined. Moreover, let $\delta_0 \in \left(0, \frac{7}{8}\right)$

and $\tau_2 > 0$ be given by

$$\delta_0 := \frac{1}{2\tau_1}\sqrt{\frac{\frac{7}{4} - 2\delta_0}{\alpha_2(f(-\beta) - f(\beta))}},$$
$$\tau_2 := \min\left\{2\delta_0\tau_1, \left(\frac{1}{\tau_1} + \tau_0 - f'(\beta)\right)^{-1}, \left(\frac{1}{\tau_1} + \tau_0 - f'(-\beta)\right)^{-1}\right\}.$$

Theorem 2.33. Suppose that $u_0 \in \mathcal{C}(\overline{\Omega})$ with $\|u_0\|_{L^\infty} \le \beta$ and f satisfies Assumptions (A1) and (A3), then the numerical solution $\{u^m\}_{0\le m\le M}$ given by the LRI2 scheme (2.70) is also bounded by β in the supremum norm:

$$\|u^m\|_{L^\infty} \le \beta, \quad m = 0, 1, \dots, M,$$

provided that $0 < \Delta t \le \tau_2$.

Proof. Assume that $\|u^m\|_{L^\infty} \le \beta$ for some m, we need to prove that $\|u^{m+1}\|_{L^\infty} \le \beta$. By (2.70) and the triangle inequality, we have

$$\begin{aligned}\|u^{m+1}\|_{L^\infty} &= \left\| e^{\Delta t\varepsilon^2\Delta}u^m + \frac{\Delta t}{2}\left[e^{\Delta t\varepsilon^2\Delta}\overline{f}(u^m) + \overline{f}(e^{\Delta t\varepsilon^2\Delta}u^m)\right] + \frac{\Delta t^2}{2}e^{\Delta t\varepsilon^2\Delta}\overline{g}(u^m)\right\|_{L^\infty}\\ &\le \left\|\delta e^{\Delta t\varepsilon^2\Delta}u^m + \frac{\Delta t}{2}\overline{f}(e^{\Delta t\varepsilon^2\Delta}u^m)\right\|_{L^\infty}\\ &\quad + \left\|(1-\delta)e^{\Delta t\varepsilon^2\Delta}u^m + \frac{\Delta t}{2}e^{\Delta t\varepsilon^2\Delta}\overline{f}(u^m) + \frac{\Delta t^2}{2}e^{\Delta t\varepsilon^2\Delta}\overline{g}(u^m)\right\|_{L^\infty},\end{aligned}$$

for some $\delta \in (0, 1)$ to be determined later. By Lemma 2.29, we arrive at

$$\|u^{m+1}\|_{L^\infty} \le \delta\left\|e^{\Delta t\varepsilon^2\Delta}u^m + \frac{\Delta t}{2\delta}\overline{f}(e^{\Delta t\varepsilon^2\Delta}u^m)\right\|_{L^\infty} + \left\|(1-\delta)u^m + \frac{\Delta t}{2}\overline{f}(u^m) + \frac{\Delta t^2}{2}\overline{g}(u^m)\right\|_{L^\infty}. \tag{2.78}$$

Since $\|u^m\|_{L^\infty} \le \beta$, we have $\|e^{\Delta t\varepsilon^2\Delta}u^m\|_{L^\infty} \le \beta$. So if $\frac{\Delta t}{2\delta} \le \tau_1$, or $\Delta t \le 2\delta\tau_1$, then Lemma 2.31 gives us

$$\left\|e^{\Delta t\varepsilon^2\Delta}u^m + \frac{\Delta t}{2\delta}\overline{f}(e^{\Delta t\varepsilon^2\Delta}u^m)\right\|_{L^\infty} \le \beta. \tag{2.79}$$

To estimate the second term on the right hand side of (2.78), let us introduce some constants:

$$\rho_0 := \frac{1}{|\Omega|}\int_\Omega f(u^m), \quad \rho_1 := \frac{1}{|\Omega|}\int_\Omega f'(u^m), \quad \rho_2 := \frac{1}{|\Omega|}\int_\Omega f(u^m)f'(u^m).$$

Since $\overline{f}(u^m) = f(u^m) - \rho_0$, we can rewrite $\overline{g}(u^m)$ as follows

$$\begin{aligned}
\overline{g}(u^m) &= g(u^m) - \frac{1}{|\Omega|}\int_\Omega g(u^m) = f'(u^m)\overline{f}(u^m) - \frac{1}{|\Omega|}\int_\Omega f'(u^m)\overline{f}(u^m) \\
&= f'(u^m)(f(u^m) - \rho_0) - \frac{1}{|\Omega|}\int_\Omega f'(u^m)(f(u^m) - \rho_0) \\
&= f'(u^m)(f(u^m) - \rho_0) + \rho_0\rho_1 - \rho_2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&(1-\delta)u^m + \frac{\Delta t}{2}\overline{f}(u^m) + \frac{\Delta t^2}{2}\overline{g}(u^m) \\
&= (1-\delta)u^m + \frac{\Delta t}{2}(f(u^m) - \rho_0) + \frac{\Delta t^2}{2}[f'(u^m)(f(u^m) - \rho_0) + \rho_0\rho_1 - \rho_2] \\
&= (1-\delta)u^m + \frac{\Delta t}{2}(f(u^m) - \rho_0)(1 + \Delta t f'(u^m)) + \frac{\Delta t^2}{2}(\rho_0\rho_1 - \rho_2).
\end{aligned} \tag{2.80}$$

The expression (2.80) motivates us to consider the function:

$$K(\xi) = (1-\delta)\xi + \frac{\Delta t}{2}(f(\xi) - \rho_0)(1 + \Delta t f'(\xi)) + \frac{\Delta t^2}{2}(\rho_0\rho_1 - \rho_2),$$

for $\xi \in [-\beta, \beta]$. Notice that

$$K'(\xi) = 1 - \delta + \frac{\Delta t}{2}\left[f'(\xi)(1 + \Delta t f'(\xi)) + \Delta t f''(\xi)(f(\xi) - \rho_0)\right].$$

By the mean value theorem and Assumption (A1), we find that $f(\beta) \le \rho_0 \le f(-\beta)$. This shows that $|f(\xi) - \rho_0| \le f(-\beta) - f(\beta)$ for all $\xi \in [-\beta, \beta]$. Hence, we obtain

$$\begin{aligned}
K'(\xi) &= \frac{7}{8} - \delta + \frac{1}{2}\left(\frac{1}{2} + \Delta t f'(\xi)\right)^2 + \frac{\Delta t^2}{2} f''(\xi)(f(\xi) - \rho_0) \\
&\ge \frac{7}{8} - \delta - \frac{\Delta t^2}{2}\alpha_2(f(-\beta) - f(\beta)), \quad \forall\, \xi \in [-\beta, \beta].
\end{aligned}$$

If $\delta \in \left(0, \frac{7}{8}\right)$ and Δt are chosen such that

$$\Delta t \leq \sqrt{\frac{\frac{7}{4} - 2\delta}{\alpha_2(f(-\beta) - f(\beta))}},$$

then $K'(\xi) \geq 0$ for all $\xi \in [-\beta, \beta]$. Since K is non-decreasing on $[-\beta, \beta]$, it suffices to control the values of K at the endpoints, i.e., we aim to show that $K(\beta) \leq (1-\delta)\beta$ and $K(-\beta) \geq -(1-\delta)\beta$. Observe that

$$\begin{aligned}
K(\beta) &= (1-\delta)\beta + \frac{\Delta t}{2}(f(\beta) - \rho_0)(1 + \Delta t f'(\beta)) + \frac{\Delta t^2}{2}(\rho_0\rho_1 - \rho_2), \\
K(-\beta) &= -(1-\delta)\beta + \frac{\Delta t}{2}(f(-\beta) - \rho_0)(1 + \Delta t f'(-\beta)) + \frac{\Delta t^2}{2}(\rho_0\rho_1 - \rho_2).
\end{aligned}$$

By the definition of ρ_1 and the fact that $-\frac{1}{\tau_1} \leq f'(\xi) \leq \tau_0$ for all $\xi \in [-\beta, \beta]$, we obtain $-\frac{1}{\tau_1} \leq \rho_1 \leq \tau_0$. Hence

$$\begin{aligned}
\rho_0\rho_1 - \rho_2 &= \rho_0\rho_1 - \frac{1}{|\Omega|}\int_\Omega f(u^m)f'(u^m) \\
&= \rho_1(\rho_0 - f(\beta)) - \frac{1}{|\Omega|}\int_\Omega [f(u^m) - f(\beta)]f'(u^m) \\
&\leq \tau_0(\rho_0 - f(\beta)) + \frac{1}{\tau_1}\frac{1}{|\Omega|}\int_\Omega [f(u^m) - f(\beta)] = (\rho_0 - f(\beta))\left(\tau_0 + \frac{1}{\tau_1}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\rho_0\rho_1 - \rho_2 &= \rho_1(\rho_0 - f(-\beta)) + \frac{1}{|\Omega|}\int_\Omega [f(-\beta) - f(u^m)]f'(u^m) \\
&\geq \tau_0(\rho_0 - f(-\beta)) - \frac{1}{\tau_1}\frac{1}{|\Omega|}\int_\Omega [f(-\beta) - f(u^m)] = (\rho_0 - f(-\beta))\left(\tau_0 + \frac{1}{\tau_1}\right).
\end{aligned}$$

Based on the above estimates, we deduce that

$$K(\beta) \leq (1-\delta)\beta + \frac{\Delta t^2}{2}(f(\beta) - \rho_0)\left(\frac{1}{\Delta t} + f'(\beta) - \tau_0 - \frac{1}{\tau_1}\right) \leq (1-\delta)\beta,$$

if Δt satisfies $\frac{1}{\Delta t} + f'(\beta) - \tau_0 - \frac{1}{\tau_1} \geq 0$ or $\Delta t \leq \left(\frac{1}{\tau_1} + \tau_0 - f'(\beta)\right)^{-1}$. By the same argument, we also have

$$K(-\beta) \geq -(1-\delta)\beta + \frac{\Delta t^2}{2}(f(-\beta) - \rho_0)\left(\frac{1}{\Delta t} + f'(-\beta) - \tau_0 - \frac{1}{\tau_1}\right) \geq -(1-\delta)\beta,$$

if $\Delta t \leq \left(\frac{1}{\tau_1} + \tau_0 - f'(-\beta)\right)^{-1}$. Now we can conclude that $|K(\xi)| \leq (1-\delta)\beta$ for all $\xi \in [-\beta, \beta]$ and therefore,

$$\left\|(1-\delta)u^m + \frac{\Delta t}{2}\overline{f}(u^m) + \frac{\Delta t^2}{2}\overline{g}(u^m)\right\|_{L^\infty} \leq (1-\delta)\beta, \tag{2.81}$$

provided that

$$\Delta t \leq \min\left\{\sqrt{\frac{\frac{7}{4} - 2\delta}{\alpha_2(f(-\beta) - f(\beta))}}, \left(\frac{1}{\tau_1} + \tau_0 - f'(\beta)\right)^{-1}, \left(\frac{1}{\tau_1} + \tau_0 - f'(-\beta)\right)^{-1}\right\}.$$

From (2.78), (2.79) and (2.81), we deduce that

$$\|u^{m+1}\|_{L^\infty} \leq \delta\beta + (1-\delta)\beta = \beta,$$

provided that

$$\Delta t \leq \min\left\{2\delta\tau_1, \sqrt{\frac{\frac{7}{4} - 2\delta}{\alpha_2(f(-\beta) - f(\beta))}}, \left(\frac{1}{\tau_1} + \tau_0 - f'(\pm\beta)\right)^{-1}\right\}.$$

To maximize the upper bound for Δt, we pick $\delta = \delta_0$ so that

$$2\delta\tau_1 = \sqrt{\frac{\frac{7}{4} - 2\delta}{\alpha_2(f(-\beta) - f(\beta))}},$$

and the condition on Δt becomes $0 < \Delta t \leq \tau_2$. □

Remark 2.34. The quartic double-well potential function $F(u) = \frac{1}{4}(u^2-1)^2$ has the corresponding nonlinear function $f(u) = -F'(u) = u - u^3$. As shown in [114, 203], for any $\beta \in [\frac{2}{\sqrt{3}}, \infty)$, f satisfies Assumptions (A1), (A2) and (A3), hence we choose $\beta = \frac{2}{\sqrt{3}}$. By simple calculations, we obtain

$$\tau_0 = 1, \quad \tau_1 = \frac{1}{3}, \quad \alpha_2 = 4\sqrt{3}, \quad \delta_0 \approx 0.54, \quad \text{and} \quad \tau_2 = \frac{1}{7}.$$

Thus, by Theorems 2.32 and 2.33, the MBP is preserved if $\Delta t \in (0, \frac{1}{3}]$ for the LRI1 schemes and $\Delta t \in (0, \frac{1}{7}]$ for the LRI2 scheme. Similarly, the Flory-Huggins potential function

$$F(u) = \frac{\theta}{2}\left[(1+u)\ln(1+u) + (1-u)\ln(1-u)\right] - \frac{\theta_c}{2}u^2,$$

corresponds to the nonlinear function $f(u) = -F'(u) = \frac{\theta}{2}\ln\frac{1-u}{1+u} + \theta_c u$, where $0 < \theta < \theta_c$. If $\theta = 0.8$ and $\theta_c = 1.6$, then $\beta \approx 0.987$ (see [114, 203] for a detailed explanation) and so

$$\tau_0 = 0.8, \quad \tau_1 \approx 0.035, \quad \alpha_2 \approx 2290.028, \quad \delta_0 \approx 0.338, \quad \text{and} \quad \tau_2 \approx 0.017.$$

Therefore, the MBP is preserved if $\Delta t \in (0, 0.035]$ for the LRI1 schemes and $\Delta t \in (0, 0.017]$ for the LRI2 scheme.

It should be noted that these upper bounds for Δt (with either double-well or Flory-Huggins potential function) can be improved in the numerical experiments since the supremum norm of numerical solution is usually less than β. In other words, the above upper bounds for Δt are optimal only if the supremum norm is sufficiently close to β.

2.3.3 Convergence analysis

In this section, we assume the nonlinear function f is sufficiently smooth for error analysis purpose; note that the double-well and Flory-Huggins potentials in Remark 2.34 are infinitely differentiable functions in their domains. Throughout the section, Lemma 2.30 is implicitly invoked whenever the exact solution u of (2.62) is present as an argument of the nonlinear functions $f, \overline{f}$ or $g, \overline{g}$. The lemma below shows the Lipschitz continuity of the nonlinear functions $\overline{f}$ and $\overline{g}$.

Lemma 2.35. If $u, v \in \mathcal{C}(\overline{\Omega})$ satisfy $\|u\|_{L^\infty} \le \beta$ and $\|v\|_{L^\infty} \le \beta$ then

(i) $\|\overline{f}(u) - \overline{f}(v)\|_{L^\infty} \le 2\alpha_1 \|u - v\|_{L^\infty}$,

(ii) $\|\overline{g}(u) - \overline{g}(v)\|_{L^\infty} \le 4(\alpha_1^2 + \alpha_0\alpha_2)\|u - v\|_{L^\infty}$.

Proof. By the definition of $\overline{f}$, we have

$$\overline{f}(u) - \overline{f}(v) = f(u) - f(v) - \frac{1}{|\Omega|}\int_\Omega [f(u) - f(v)].$$

Taking the supremum norm on both sides of the above identity gives us

$$\begin{aligned}\|\overline{f}(u) - \overline{f}(v)\|_{L^\infty} &\leq \|f(u) - f(v)\|_{L^\infty} + \frac{1}{|\Omega|}\int_\Omega \|f(u) - f(v)\|_{L^\infty} \\ &= 2\|f(u) - f(v)\|_{L^\infty} \leq 2\alpha_1\|u - v\|_{L^\infty}.\end{aligned}$$

On the other hand, we note that $\|\overline{f}(v)\|_{L^\infty} \leq 2\alpha_0$ and $\|f'(u) - f'(v)\|_{L^\infty} \leq \alpha_2\|u - v\|_{L^\infty}$, so

$$\begin{aligned}\|g(u) - g(v)\|_{L^\infty} &= \|f'(u)\overline{f}(u) - f'(v)\overline{f}(v)\|_{L^\infty} \\ &\leq \|f'(u)(\overline{f}(u) - \overline{f}(v))\|_{L^\infty} + \|(f'(u) - f'(v))\overline{f}(v)\|_{L^\infty} \\ &\leq 2\alpha_1^2\|u - v\|_{L^\infty} + 2\alpha_0\alpha_2\|u - v\|_{L^\infty} = 2(\alpha_1^2 + \alpha_0\alpha_2)\|u - v\|_{L^\infty}.\end{aligned}$$

It follows that

$$\|\overline{g}(u) - \overline{g}(v)\|_{L^\infty} \leq 2\|g(u) - g(v)\|_{L^\infty} \leq 4(\alpha_1^2 + \alpha_0\alpha_2)\|u - v\|_{L^\infty},$$

which completes the proof. □

From this point forward, we assume that the exact solution u of (2.62) is Lipschitz continuous with respect to the time variable so that $\overline{f}(u)$ and $\overline{g}(u)$ also satisfy the same property. More precisely, there exists a constant $L > 0$ such that

$$\|u(s) - u(t)\|_{L^\infty} \leq L|s - t|, \quad \forall\, s, t \in [0, T]. \tag{2.82}$$

By Lemma 2.35, we can assert that

$$\begin{aligned}&\|\overline{f}(u(s)) - \overline{f}(u(t))\|_{L^\infty} \leq 2\alpha_1 L|s - t|, \\ &\|\overline{g}(u(s)) - \overline{g}(u(t))\|_{L^\infty} \leq 4(\alpha_1^2 + \alpha_0\alpha_2)L|s - t|, \quad \forall\, s, t \in [0, T].\end{aligned}$$

The following technical lemma is useful to obtain error estimates for LRI schemes.

Lemma 2.36. Suppose that $u \in \mathcal{C}([0,T], \mathcal{C}^2(\overline{\Omega}))$ is the exact solution of (2.62) and satisfies the Lipschitz condition (2.82). For any $s, t \geq 0$ with $s + t \leq T$, it holds

$$\|e^{s\varepsilon^2\Delta}\overline{f}(u(t)) - \overline{f}(u(t+s))\|_{L^\infty} \leq Cs,$$

where $C = \varepsilon^2 \max_{t\in[0,T]} \|\Delta f(u(t))\|_{L^\infty} + 2\alpha_1 L$ is independent of s and t.

Proof. By the fundamental theorem of calculus and Lemma 2.29, we deduce that

$$\begin{aligned}
\|(e^{s\varepsilon^2\Delta} - 1)\overline{f}(u(t))\|_{L^\infty} &= \left\|\int_0^s e^{\sigma\varepsilon^2\Delta}\varepsilon^2\Delta\overline{f}(u(t))\,d\sigma\right\|_{L^\infty} \\
&= \left\|\varepsilon^2\int_0^s e^{\sigma\varepsilon^2\Delta}\Delta f(u(t))\,d\sigma\right\|_{L^\infty} \\
&\leq \varepsilon^2\int_0^s \|e^{\sigma\varepsilon^2\Delta}\Delta f(u(t))\|_{L^\infty}d\sigma \leq \varepsilon^2 s\|\Delta f(u(t))\|_{L^\infty} \leq \varepsilon^2 c_1 s,
\end{aligned} \tag{2.83}$$

where $c_1 = \max_{t\in[0,T]} \|\Delta f(u(t))\|_{L^\infty}$ depends on the norm of f in $\mathcal{C}^2([-\beta,\beta])$ and the norm of u in $\mathcal{C}([0,T], \mathcal{C}^2(\overline{\Omega}))$.

From Lemma 2.35 and the Lipschitz continuity of u with respect to the time variable (2.82)), there holds

$$\|\overline{f}(u(t)) - \overline{f}(u(t+s))\|_{L^\infty} \leq c_2 s, \tag{2.84}$$

where $c_2 = 2\alpha_1 L$. Combining (2.83) and (2.84) gives us

$$\|e^{s\varepsilon^2\Delta}\overline{f}(u(t)) - \overline{f}(u(t+s))\|_{L^\infty} \leq \|(e^{s\varepsilon^2\Delta} - 1)\overline{f}(u(t))\|_{L^\infty} + \|\overline{f}(u(t)) - \overline{f}(u(t+s))\|_{L^\infty} \leq Cs,$$

where $C = \varepsilon^2 c_1 + c_2$. □

Theorem 2.37. Suppose that the exact solution u to the conservative Allen-Cahn equation (2.62) belongs to $\mathcal{C}([0,T], \mathcal{C}^2(\overline{\Omega}))$ and satisfies the Lipschitz condition (2.82) with the initial data $u_0 \in \mathcal{C}(\overline{\Omega})$ satisfying $\|u_0\|_{L^\infty} \leq \beta$. Denote by $\{u^m\}_{0\leq m\leq M}$ the numerical solution generated by the LRI1a scheme (2.66) or the LRI1b scheme (2.67), then there exists a constant $C > 0$ independent of Δt such that the following estimate holds for all $\Delta t \in (0, \tau_1]$:

$$\|u^m - u(t_m)\|_{L^\infty} \leq C(e^{2\alpha_1 t_m} - 1)\Delta t, \quad m = 0, 1, \ldots, M.$$

Proof. We first consider the LRI1a scheme (2.66). Subtracting (2.64) from (2.66), we obtain

$$\begin{aligned} u^{m+1} - u(t_{m+1}) = e^{\Delta t \varepsilon^2 \Delta}(u^m - u(t_m)) + \Delta t e^{\Delta t \varepsilon^2 \Delta}(\overline{f}(u^m) - \overline{f}(u(t_m))) \\ + \int_0^{\Delta t} e^{(\Delta t - s)\varepsilon^2 \Delta}[e^{s\varepsilon^2\Delta}\overline{f}(u(t_m)) - \overline{f}(u(t_m + s))]ds. \end{aligned}$$

Combing the above equation and Lemmas 2.29, 2.35 and 2.36 gives us

$$\|u^{m+1} - u(t_{m+1})\|_{L^\infty} \leq (1 + 2\alpha_1 \Delta t)\|u^m - u(t_m)\|_{L^\infty} + \frac{c_3 \Delta t^2}{2},$$

where

$$c_3 = \varepsilon^2 \max_{t \in [0,T]} \|\Delta f(u(t))\|_{L^\infty} + 2\alpha_1 L, \tag{2.85}$$

is the constant in Lemma 2.36. This implies that

$$\|u^m - u(t_m)\|_{L^\infty} \leq \frac{c_3 \Delta t}{4\alpha_1}[(1 + 2\alpha_1 \Delta t)^m - 1] \leq \frac{c_3 \Delta t}{4\alpha_1}(e^{2\alpha_1 m \Delta t} - 1) = C(e^{2\alpha_1 t_m} - 1)\Delta t,$$

where $C = \frac{c_3}{4\alpha_1}$.

Next, we show that a similar estimate holds for the LRI1b scheme (2.67). Indeed, by substracting (2.64) from (2.67), we arrive at

$$\begin{aligned} u^{m+1} - u(t_{m+1}) = e^{\Delta t \varepsilon^2 \Delta}(u^m - u(t_m)) + \Delta t[\overline{f}(e^{\Delta t \varepsilon^2 \Delta}u^m) - \overline{f}(e^{\Delta t \varepsilon^2 \Delta}u(t_m))] \\ + \int_0^{\Delta t} [\overline{f}(e^{\Delta t \varepsilon^2 \Delta}u(t_m)) - e^{(\Delta t - s)\varepsilon^2 \Delta}\overline{f}(u(t_m + s))]ds. \end{aligned}$$

Based on Lemmas 2.29 and 2.35, we deduce that

$$\begin{aligned} \|u^{m+1} - u(t_{m+1})\|_{L^\infty} \quad &\leq (1 + 2\alpha_1 \Delta t)\|u^m - u(t_m)\|_{L^\infty} \\ &+ \int_0^{\Delta t} \|\overline{f}(e^{\Delta t \varepsilon^2 \Delta}u(t_m)) - e^{(\Delta t - s)\varepsilon^2 \Delta}\overline{f}(u(t_m + s))\|_{L^\infty} ds. \end{aligned} \tag{2.86}$$

On the other hand, using formula (2.64) and Lemmas 2.35 and 2.36 yields

$$
\begin{aligned}
&\|\overline{f}(e^{\Delta t\varepsilon^2\Delta}u(t_m)) - e^{(\Delta t-s)\varepsilon^2\Delta}\overline{f}(u(t_m+s))\|_{L^\infty} \\
&\leq \|\overline{f}(e^{\Delta t\varepsilon^2\Delta}u(t_m)) - \overline{f}(u(t_{m+1}))\|_{L^\infty} + \|\overline{f}(u(t_{m+1})) - e^{(\Delta t-s)\varepsilon^2\Delta}\overline{f}(u(t_m+s))\|_{L^\infty} \\
&\leq 2\alpha_1\|u(t_{m+1}) - e^{\Delta t\varepsilon^2\Delta}u(t_m)\|_{L^\infty} + c_3(\Delta t - s) \\
&= 2\alpha_1\left\|\int_0^{\Delta t} e^{(\Delta t-s)\varepsilon^2\Delta}\overline{f}(u(t_m+s))\,ds\right\|_{L^\infty} + c_3(\Delta t - s) \\
&\leq 2\alpha_1\int_0^{\Delta t}\|\overline{f}(u(t_m+s))\|_{L^\infty}ds + c_3(\Delta t - s) \leq 4\alpha_0\alpha_1\Delta t + c_3(\Delta t - s),
\end{aligned}
\tag{2.87}
$$

where we have used the fact that $\|\overline{f}(u(t_m+s))\|_{L^\infty} \leq 2\|f(u(t_m+s))\|_{L^\infty} \leq 2\alpha_0$. From (2.86) and (2.87), we obtain

$$
\begin{aligned}
\|u^{m+1} - u(t_{m+1})\|_{L^\infty} &\leq (1+2\alpha_1\Delta t)\|u^m - u(t_m)\|_{L^\infty} + \int_0^{\Delta t}[4\alpha_0\alpha_1\Delta t + c_3(\Delta t - s)]ds \\
&= (1+2\alpha_1\Delta t)\|u^m - u(t_m)\|_{L^\infty} + \left(4\alpha_0\alpha_1 + \frac{c_3}{2}\right)\Delta t^2.
\end{aligned}
$$

By the same argument as in the proof for the LRI1a scheme, the desired inequality holds for $C = 2\alpha_0 + \frac{c_3}{4\alpha_1}$. □

Remark 2.38. The constant C in Theorem 2.37 can be explicitly expressed as:

$$
C = 2\alpha_0 + \frac{\varepsilon^2}{4\alpha_1}\max_{t\in[0,T]}\|\Delta f(u(t))\|_{L^\infty} + \frac{1}{2}L,
$$

where α_0 and α_1 are defined in (2.75), and L denotes the Lipschitz constant in time of u (cf. (2.82)). Therefore, C does not change significantly with different values of ε. However, both $\max_{t\in[0,T]}\|\Delta f(u(t))\|_{L^\infty}$ and L, and consequently C, increase when T becomes larger. We shall verify the dependence of C on ε and T in the numerical experiments (cf. Section 2.3.5).

Theorem 2.39. Suppose that the exact solution u to the conservative Allen-Cahn equation (2.62) belongs to $\mathcal{C}([0,T],\mathcal{C}^4(\overline{\Omega}))$ and satisfies the Lipschitz condition (2.82) with the initial data $u_0 \in \mathcal{C}(\overline{\Omega})$ satisfying $\|u_0\|_{L^\infty} \leq \beta$. Denote by $\{u^m\}_{0\leq m\leq M}$ the numerical solution generated by the

LRI2 scheme (2.70), then there exists a constant $C > 0$ independent of Δt such that the following estimate holds for all $\Delta t \in (0, \tau_2]$:

$$\|u^m - u(t_m)\|_{L^\infty} \le C(e^{2\widetilde{\alpha}_1 t_m} - 1)\Delta t^2, \quad m = 0, 1, \dots, M,$$

where $\widetilde{\alpha}_1 = \alpha_1 + T(\alpha_1^2 + \alpha_0\alpha_2)$.

Proof. Subtracting (2.64) from (2.70), we have

$$\begin{aligned} u^{m+1} - u(t_{m+1}) = {} & e^{\Delta t \varepsilon^2 \Delta}(u^m - u(t_m)) \\ & + \frac{\Delta t}{2}\left[e^{\Delta t\varepsilon^2\Delta}(\overline{f}(u^m) - \overline{f}(u(t_m))) + \overline{f}(e^{\Delta t\varepsilon^2\Delta}u^m) - \overline{f}(e^{\Delta t\varepsilon^2\Delta}u(t_m))\right] \\ & + \frac{\Delta t^2}{2}e^{\Delta t\varepsilon^2\Delta}(\overline{g}(u^m) - \overline{g}(u(t_m))) + R(t_m), \end{aligned} \tag{2.88}$$

where $R(t_m)$ is defined by

$$\begin{aligned} R(t_m) = {} & \frac{\Delta t}{2}[e^{\Delta t\varepsilon^2\Delta}\overline{f}(u(t_m)) + \overline{f}(e^{\Delta t\varepsilon^2\Delta}u(t_m))] + \frac{\Delta t^2}{2}e^{\Delta t\varepsilon^2\Delta}\overline{g}(u(t_m)) \\ & - \int_0^{\Delta t} e^{(\Delta t - s)\varepsilon^2\Delta}\overline{f}(u(t_m + s))\, ds. \end{aligned}$$

Using Lemmas 2.29 and 2.35, we deduce from (2.88) that

$$\|u^{m+1} - u(t_{m+1})\|_{L^\infty} \le \left(1 + 2\alpha_1\Delta t + 2(\alpha_1^2 + \alpha_0\alpha_2)\Delta t^2\right)\|u^m - u(t_m)\|_{L^\infty} + \|R(t_m)\|_{L^\infty}. \tag{2.89}$$

Rewrite $R(t_m)$ as $R(t_m) = R_1(t_m) + R_2(t_m)$, where $R_i(t_m),\ i = 1, 2$, are given by

$$\begin{aligned} R_1(t_m) &= \frac{\Delta t}{2}[e^{\Delta t\varepsilon^2\Delta}\overline{f}(u(t_m)) + \overline{f}(e^{\Delta t\varepsilon^2\Delta}u(t_m))] - \int_0^{\Delta t} e^{(\Delta t - s)\varepsilon^2\Delta}\overline{f}(e^{s\varepsilon^2\Delta}u(t_m))\, ds, \\ R_2(t_m) &= \int_0^{\Delta t} e^{(\Delta t - s)\varepsilon^2\Delta}[\overline{f}(e^{s\varepsilon^2\Delta}u(t_m)) - \overline{f}(u(t_m + s)) + s e^{s\varepsilon^2\Delta}\overline{g}(u(t_m))]ds. \end{aligned}$$

Let $\varphi(s) = e^{(\Delta t - s)\varepsilon^2\Delta}\overline{f}(e^{s\varepsilon^2\Delta}u(t_m))$ with $s \in [0, \Delta t]$, then $R_1(t_m)$ can be written in terms of φ

as follows

$$R_1(t_m) = \frac{\Delta t}{2}[\varphi(0) + \varphi(s)] - \int_0^{\Delta t} \varphi(s)\, ds = \frac{1}{2}\int_0^{\Delta t} s(\Delta t - s)\varphi''(s)\, ds.$$

Thus

$$\|R_1(t_m)\|_{L^\infty} \leq \frac{\Delta t^3}{12} \sup_{s\in[0,\Delta t]} \|\varphi''(s)\|_{L^\infty}. \tag{2.90}$$

By expanding $\varphi''(s)$ using the chain rule, we find that $\sup_{s\in[0,\Delta t]} \|\varphi''(s)\|_{L^\infty} \leq 12c_4$, where c_4 depends on $\mathcal{C}^4([-\beta,\beta])$-norm of f and $\mathcal{C}([0,T],\mathcal{C}^4(\overline{\Omega}))$-norm of u, but is independent of Δt. Therefore, (2.90) gives us

$$\|R_1(t_m)\|_{L^\infty} \leq c_4\Delta t^3. \tag{2.91}$$

To estimate $\|R_2(t_m)\|_{L^\infty}$, we first observe from Lemma 2.29 that

$$\begin{aligned}
\|R_2(t_m)\|_{L^\infty} &\leq \int_0^{\Delta t} \|\overline{f}(e^{s\varepsilon^2\Delta}u(t_m)) - \overline{f}(u(t_m+s)) + se^{s\varepsilon^2\Delta}\overline{g}(u(t_m))\|_{L^\infty} ds \\
&\leq \int_0^{\Delta t} \|\overline{f}(e^{s\varepsilon^2\Delta}u(t_m)) - \overline{f}(u(t_m+s)) + s\overline{g}(u(t_m))\|_{L^\infty} ds \\
&\qquad + \int_0^{\Delta t} s\|(e^{s\varepsilon^2\Delta} - 1)\overline{g}(u(t_m))\|_{L^\infty} ds.
\end{aligned}$$

Similar to (2.83), we have $\|(e^{s\varepsilon^2\Delta}-1)\overline{g}(u(t_m))\|_{L^\infty} \leq \varepsilon^2 c_5 s$, where $c_5 = \max_{t\in[0,T]} \|\Delta g(u(t))\|_{L^\infty}$ depends on $\mathcal{C}^3([-\beta,\beta])$-norm of f and $\mathcal{C}([0,T],\mathcal{C}^2(\overline{\Omega}))$-norm of u. Consequently,

$$\begin{aligned}
\|R_2(t_m)\|_{L^\infty} &\leq 2\int_0^{\Delta t} \|f(e^{s\varepsilon^2\Delta}u(t_m)) - f(u(t_m+s)) + sg(u(t_m))\|_{L^\infty} ds + \frac{\varepsilon^2 c_5}{3}\Delta t^3 \\
&=: 2\int_0^{\Delta t} \|I_1 + I_2 + I_3 + I_4\|_{L^\infty} ds + \frac{\varepsilon^2 c_5}{3}\Delta t^3,
\end{aligned} \tag{2.92}$$

where I_j with $j = 1, 2, 3, 4$ are given by

$$I_1 = f(u(t_m + s)) - f(e^{s\varepsilon^2\Delta}u(t_m)) - f'(e^{s\varepsilon^2\Delta}u(t_m)) \int_0^s e^{(s-\sigma)\varepsilon^2\Delta}\overline{f}(u(t_m + \sigma))\, d\sigma,$$
$$I_2 = [f'(e^{s\varepsilon^2\Delta}u(t_m)) - f'(u(t_m))] \int_0^s e^{(s-\sigma)\varepsilon^2\Delta}\overline{f}(u(t_m + \sigma))\, d\sigma,$$
$$I_3 = f'(u(t_m)) \int_0^s [e^{(s-\sigma)\varepsilon^2\Delta}\overline{f}(u(t_m + \sigma)) - \overline{f}(u(t_m + s))] d\sigma,$$
$$I_4 = s[f'(u(t_m))\overline{f}(u(t_m + s)) - g(u(t_m))].$$

By Taylor's expansion and formula (2.65), there holds

$$\|I_1\|_{L^\infty} \le \frac{\alpha_2}{2} \left\| \int_0^s e^{(s-\sigma)\varepsilon^2\Delta}\overline{f}(u(t_m + \sigma))\, d\sigma \right\|^2 \le \frac{\alpha_2}{2}(2\alpha_0 s)^2 =: c_6 s^2, \tag{2.93}$$

where $c_6 = 2\alpha_0^2\alpha_2$. Based on (2.83), we also have $\|(e^{s\varepsilon^2\Delta} - 1)u(t_m)\|_{L^\infty} \le \varepsilon^2 c_7 s$ where $c_7 = \max_{t\in[0,T]} \|\Delta u(t)\|_{L^\infty}$ depends on $\mathcal{C}([0, T], \mathcal{C}^2(\overline{\Omega}))$-norm of u. We thus obtain

$$\|I_2\|_{L^\infty} \le \alpha_2 \|(e^{s\varepsilon^2\Delta} - 1)u(t_m)\|_{L^\infty} \int_0^s \|\overline{f}(u(t_m + \sigma))\|_{L^\infty} d\sigma \le 2\varepsilon^2\alpha_0\alpha_2 c_7 s^2. \tag{2.94}$$

From Lemma 2.36, it holds that

$$\|I_3\|_{L^\infty} \le \alpha_1 \int_0^s c_3(s - \sigma)\, d\sigma = \frac{\alpha_1 c_3}{2} s^2, \tag{2.95}$$

where c_3 is given by (2.85). Notice that $g = f'\overline{f}$, so we have from Lemma 2.35 that

$$\|I_4\|_{L^\infty} = s\|f'(u(t_m))[\overline{f}(u(t_m + s)) - \overline{f}(u(t_m))]\|_{L^\infty} \le 2\alpha_1^2 L s^2. \tag{2.96}$$

Combining (2.92)-(2.96) and using the triangle inequality lead to

$$\begin{aligned} \|R_2(t_m)\|_{L^\infty} &\le 2 \int_0^{\Delta t} \left(c_6 s^2 + 2\varepsilon^2\alpha_0\alpha_2 c_7 s^2 + \frac{\alpha_1 c_3}{2} s^2 + 2\alpha_1^2 L s^2 \right) ds + \frac{\varepsilon^2 c_5}{3}\Delta t^3 \\ &= \frac{\varepsilon^2 c_5 + 2c_6 + 4\varepsilon^2\alpha_0\alpha_2 c_7 + \alpha_1 c_3 + 4\alpha_1^2 L}{3}\Delta t^3 =: c_8 \Delta t^3. \end{aligned} \tag{2.97}$$

From (2.91) and (2.97), we deduce that $\|R(t_m)\|_{L^\infty} \le (c_4 + c_8)\Delta t^3$. Finally, we get from (2.89) that

$$\begin{aligned}\|u^{m+1} - u(t_{m+1})\|_{L^\infty} &\le \left(1 + 2\alpha_1\Delta t + 2(\alpha_1^2 + \alpha_0\alpha_2)\Delta t^2\right)\|u^m - u(t_m)\|_{L^\infty} + (c_4 + c_8)\Delta t^3 \\ &\le (1 + 2\widetilde{\alpha}_1\Delta t)\|u^m - u(t_m)\|_{L^\infty} + (c_4 + c_8)\Delta t^3,\end{aligned}$$

where $\widetilde{\alpha}_1 = \alpha_1 + T(\alpha_1^2 + \alpha_0\alpha_2)$. Using the discrete Gronwall's inequality, we obtain

$$\|u^m - u(t_m)\|_{L^\infty} \le \frac{c_4 + c_8}{2\widetilde{\alpha}_1}(e^{2\widetilde{\alpha}_1 t_m} - 1)\Delta t^2.$$

The proof is completed by setting $C = \dfrac{c_4 + c_8}{2\widetilde{\alpha}_1}$. □

Remark 2.40. (i) The constant C in Theorem 2.39 can be written as

$$\begin{aligned}C = \frac{1}{2\widetilde{\alpha}_1}c_4 + \frac{1}{6\widetilde{\alpha}_1}\Big(&\varepsilon^2 \max_{t\in[0,T]}\|\Delta g(u(t))\|_{L^\infty} + 4\alpha_0^2\alpha_2 + 4\varepsilon^2\alpha_0\alpha_2 \max_{t\in[0,T]}\|\Delta u(t)\|_{L^\infty} \\ &+\varepsilon^2\alpha_1 \max_{t\in[0,T]}\|\Delta f(u(t))\|_{L^\infty} + 6\alpha_1^2 L\Big).\end{aligned}$$

Thus, as for the LRI1 schemes (cf. Remark 2.38), the error constant C for the LRI2 scheme does not vary considerably with respect to ε; however, it increases when the final time T becomes larger (cf. Section 2.3.5).

(ii) The temporal error analysis of the LRI1 and LRI2 schemes relies on the Duhamel's formula (2.64) and the Lipschitz condition (2.82), so that the exact solution u is only assumed to be Lipschitz continuous in time for both first- and second-order schemes (cf. Theorems 2.37 and 2.39). Therefore, the proposed schemes are called low regularity integrators [154, 45], in contrast to classical methods such as ETD [114] and isIFRK [203] which require higher regularity in time on the exact solution for the convergence analysis (i.e., u needs to be in $\mathcal{C}^p$ in time for the p-order scheme to converge).

2.3.4 Energy stability

Throughout this section, we assume that the spatial domain Ω is either a finite interval ($d = 1$), a rectangle ($d = 2$), or a rectangular parallelepiped ($d = 3$). We first state two preliminary lemmas.

Lemma 2.41. Suppose that Δ is equipped with the periodic boundary conditions on Ω and $v \in \mathcal{C}^1(\overline{\Omega})$ satisfies the same boundary conditions, then $\partial_{x_i}(e^{t\Delta}v) = e^{t\Delta}v_{x_i}$ for all $i = 1, 2, \dots, d$ and $t > 0$, where $v_{x_i} := \partial_{x_i}v$.

Proof. For simplicity, we only consider the one dimensional case, the corresponding proof in higher dimensions can be done in the same way. Without loss of generality, we may assume that $\Omega = (-P, P)$ for some $P > 0$. Let $W(\cdot, t) = e^{t\Delta}v$ and $w(\cdot, t) = e^{t\Delta}v'$, then W and w satisfy the following heat equations:

$$\begin{cases} W_t = W_{xx}, \quad x \in (-P, P),\ t > 0, \\ W(x, 0) = v(x), \quad x \in (-P, P), \end{cases}$$

and

$$\begin{cases} w_t = w_{xx}, \quad x \in (-P, P),\ t > 0, \\ w(x, 0) = v'(x), \quad x \in (-P, P), \end{cases}$$

respectively, subject to the periodic boundary conditions. The general solutions to the two problems above are:

$$W(x,t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{n\pi x}{P}\right) + b_n \sin\left(\frac{n\pi x}{P}\right)\right) e^{-\left(\frac{n\pi}{P}\right)^2 t},$$
$$w(x,t) = \frac{\overline{a_0}}{2} + \sum_{n=1}^{\infty} \left(\overline{a_n} \cos\left(\frac{n\pi x}{P}\right) + \overline{b_n} \sin\left(\frac{n\pi x}{P}\right)\right) e^{-\left(\frac{n\pi}{P}\right)^2 t},$$

where $a_n, \overline{a_n},\ n \geq 0$ and $b_n, \overline{b_n},\ n \geq 1$ are given by

$$a_n = \frac{1}{P}\int_{-P}^{P} v(x) \cos\left(\frac{n\pi x}{P}\right) dx, \quad b_n = \frac{1}{P}\int_{-P}^{P} v(x) \sin\left(\frac{n\pi x}{P}\right) dx,$$
$$\overline{a_n} = \frac{1}{P}\int_{-P}^{P} v'(x) \cos\left(\frac{n\pi x}{P}\right) dx, \quad \overline{b_n} = \frac{1}{P}\int_{-P}^{P} v'(x) \sin\left(\frac{n\pi x}{P}\right) dx.$$

We observe that

$$\partial_x W(x,t) = \sum_{n=1}^{\infty} \frac{n\pi}{P}\left(-a_n \sin\left(\frac{n\pi x}{P}\right) + b_n \cos\left(\frac{n\pi x}{P}\right)\right) e^{-\left(\frac{n\pi}{P}\right)^2 t} = w(x,t),$$

since $\overline{a_0} = \frac{1}{P}[v(P) - v(-P)] = 0$ and $\overline{a_n} = \frac{n\pi}{P} b_n,\ \overline{b_n} = -\frac{n\pi}{P} a_n$ for all $n \geq 1$. The identity in Lemma 2.41 can thus be obtained by taking into account $W(\cdot,t) = e^{t\Delta} v$ and $w(\cdot,t) = e^{t\Delta} v'$. □

Remark 2.42. The two operators ∂_{x_i} and $e^{t\Delta}$ need not be commutative when Δ is equipped with other types of boundary conditions on general domains. Our proof for the energy stability relies on Lemma 2.41, which requires the periodic boundary conditions; it remains an open question to show the same result, i.e., Theorem 2.44, for the homogeneous Neumann boundary conditions.

Lemma 2.43. Let $\{u^m\}_{0 \leq m \leq M}$ be the numerical solution sequence generated by the LRI1a or LRI1b scheme (resp. the LRI2 scheme) and suppose that the exact solution u to (2.62) satisfies the conditions stated in Theorem 2.37 (resp. Theorem 2.39). If $u_0 \in \mathcal{C}^1(\overline{\Omega})$ with $\|u_0\|_{L^\infty} \leq \beta$, then there exist constants c_9 and c_{10} independent of Δt such that

(i) $\|u_{x_i}^m\|_{L^\infty} \leq c_9$ for all $i = 1, 2, \dots, d$ and $m = 0, 1, \dots, M$.

(ii) $\|u_{x_i}^{m+1} - u_{x_i}^m\|_{L^\infty} \leq c_{10}\Delta t$ for all $i = 1, 2, \dots, d$ and $m = 0, 1, \dots, M-1$.

Proof. (i) We first consider the LRI1a scheme (2.66). Taking the derivative on both sides of (2.66) with respect to x_i yields

$$u_{x_i}^{m+1} = e^{\Delta t \varepsilon^2 \Delta}(u_{x_i}^m + \Delta t f'(u^m) u_{x_i}^m), \tag{2.98}$$

where we have used Lemma 2.41 and the fact that $\partial_{x_i} \overline{f}(u^m) = \partial_{x_i} f(u^m)$. From Lemma 2.29 and the boundedness of u^m implied by Theorem 2.32, we obtain

$$\|u_{x_i}^{m+1}\|_{L^\infty} \leq \|u_{x_i}^m + \Delta t f'(u^m) u_{x_i}^m\|_{L^\infty} \leq (1 + \alpha_1 \Delta t)\|u_{x_i}^m\|_{L^\infty}.$$

Similarly, we observe from the LRI1b scheme (2.67) that

$$\|u_{x_i}^{m+1}\|_{L^\infty} = \|e^{\Delta t \varepsilon^2 \Delta} u_{x_i}^m + \Delta t f'(e^{\Delta t \varepsilon^2 \Delta} u^m) e^{\Delta t \varepsilon^2 \Delta} u_{x_i}^m\|_{L^\infty} \leq (1 + \alpha_1 \Delta t)\|u_{x_i}^m\|_{L^\infty}.$$

This gives us $\|u_{x_i}^m\|_{L^\infty} \leq (1+\alpha_1\Delta t)^m \|u_{x_i}^0\|_{L^\infty} \leq e^{\alpha_1 t_m}\|u_{x_i}^0\|_{L^\infty} \leq e^{\alpha_1 T}\|\partial_{x_i}u_0\|_{L^\infty}$ for all $i=1,2,\dots,d$. The result follows by choosing $c_9 = e^{\alpha_1 T}\max_{1\leq i\leq d}\|\partial_{x_i}u_0\|_{L^\infty}$.

For the LRI2 scheme (2.70), based on the above estimates we have that

$$\|u_{x_i}^{m+1}\|_{L^\infty} \leq (1+\alpha_1\Delta t)\|u_{x_i}^m\|_{L^\infty} + \left\|\partial_{x_i}\left(\frac{\Delta t^2}{2}e^{\Delta t\varepsilon^2\Delta}\overline{g}(u^m)\right)\right\|_{L^\infty}. \tag{2.99}$$

Using Lemmas 2.29 and 2.41, we notice that

$$\begin{aligned}\left\|\partial_{x_i}\left(\frac{\Delta t^2}{2}e^{\Delta t\varepsilon^2\Delta}\overline{g}(u^m)\right)\right\|_{L^\infty} &= \left\|\frac{\Delta t^2}{2}e^{\Delta t\varepsilon^2\Delta}(\partial_{x_i}g(u^m))\right\|_{L^\infty} \leq \frac{\Delta t^2}{2}\left\|\partial_{x_i}(f'(u^m)\overline{f}(u^m))\right\|_{L^\infty}\\ &= \frac{\Delta t^2}{2}\left\|[(f'(u^m))^2 + \overline{f}(u^m)f''(u^m)]u_{x_i}^m\right\|_{L^\infty}\\ &\leq \frac{\Delta t^2}{2}(\alpha_1^2 + 2\alpha_0\alpha_2)\|u_{x_i}^m\|_{L^\infty}.\end{aligned} \tag{2.100}$$

From (2.99) and (2.100), we deduce that

$$\|u_{x_i}^{m+1}\|_{L^\infty} \leq \left[1+\alpha_1\Delta t + \left(\frac{\alpha_1^2}{2}+\alpha_0\alpha_2\right)\Delta t^2\right]\|u_{x_i}^m\|_{L^\infty}.$$

The assertion for the LRI2 scheme follows by the same argument as above.

(ii) It should be noted that for any $u_1, u_2, v_1, v_2 \in \mathcal{C}(\overline{\Omega})$ with $\|u_i\|_{L^\infty}, \|v_i\|_{L^\infty} \leq \beta$ for $i=1,2$, we have

$$\begin{aligned}\|f'(u_1)v_1 - f'(u_2)v_2\|_{L^\infty} &\leq \|f'(u_1)(v_1-v_2)\|_{L^\infty} + \|(f'(u_1)-f'(u_2))v_2\|_{L^\infty}\\ &\leq \alpha_1\|v_1-v_2\|_{L^\infty} + \alpha_2\|v_2\|_{L^\infty}\|u_1-u_2\|_{L^\infty}.\end{aligned} \tag{2.101}$$

Theorems 2.37 and 2.39 allow us to have the following estimate

$$\begin{aligned}\|u^{m+1}-u^m\|_{L^\infty} &\leq \|u^{m+1}-u(t_{m+1})\|_{L^\infty} + \|u(t_{m+1})-u(t_m)\|_{L^\infty} + \|u(t_m)-u^m\|_{L^\infty}\\ &= C\Delta t + L\Delta t + C\Delta t = (2C+L)\Delta t.\end{aligned} \tag{2.102}$$

For the LRI1a scheme (2.66), applying (2.98) with m and $m-1$ and taking the difference between the resulting equations, we get

$$u_{x_i}^{m+1} - u_{x_i}^{m} = e^{\Delta t \varepsilon^2 \Delta}\left[u_{x_i}^{m} - u_{x_i}^{m-1} + \Delta t (f'(u^m) u_{x_i}^{m} - f'(u^{m-1}) u_{x_i}^{m-1})\right].$$

The estimates (2.101) and (2.102), together with the result in part (i), lead to

$$\begin{aligned}\|u_{x_i}^{m+1} - u_{x_i}^{m}\|_{L^\infty} &\le \|u_{x_i}^{m} - u_{x_i}^{m-1}\|_{L^\infty} + \Delta t \|f'(u^m) u_{x_i}^{m} - f'(u^{m-1}) u_{x_i}^{m-1}\|_{L^\infty} \\ &\le (1+\alpha_1 \Delta t)\|u_{x_i}^{m} - u_{x_i}^{m-1}\|_{L^\infty} + \alpha_2 c_9 (2C + L)\Delta t^2,\end{aligned}$$

which claims the assertion by the discrete Gronwall's inequality.

Similarly, we also have from the LRI1b scheme (2.67) that

$$\begin{aligned}&\|u_{x_i}^{m+1} - u_{x_i}^{m}\|_{L^\infty} \\ &= \left\|e^{\Delta t \varepsilon^2 \Delta}(u_{x_i}^{m} - u_{x_i}^{m-1}) + \Delta t [f'(e^{\Delta t \varepsilon^2 \Delta} u^m) e^{\Delta t \varepsilon^2 \Delta} u_{x_i}^{m} - f'(e^{\Delta t \varepsilon^2 \Delta} u^{m-1}) e^{\Delta t \varepsilon^2 \Delta} u_{x_i}^{m-1}]\right\|_{L^\infty} \\ &\le \left\|u_{x_i}^{m} - u_{x_i}^{m-1}\right\|_{L^\infty} + \Delta t \left\|f'(e^{\Delta t \varepsilon^2 \Delta} u^m) e^{\Delta t \varepsilon^2 \Delta} u_{x_i}^{m} - f'(e^{\Delta t \varepsilon^2 \Delta} u^{m-1}) e^{\Delta t \varepsilon^2 \Delta} u_{x_i}^{m-1}\right\|_{L^\infty} \\ &\le (1+\alpha_1 \Delta t)\|u_{x_i}^{m} - u_{x_i}^{m-1}\|_{L^\infty} + \alpha_2 c_9 (2C + L)\Delta t^2,\end{aligned}$$

here we have used (2.101) and (2.102) in the last estimate.

Using above estimates and (2.100) for the LRI2 scheme (2.70), we have

$$\begin{aligned}\|u_{x_i}^{m+1} - u_{x_i}^{m}\|_{L^\infty} &\le (1+\alpha_1 \Delta t)\|u_{x_i}^{m} - u_{x_i}^{m-1}\|_{L^\infty} + \alpha_2 c_9 (2C + L)\Delta t^2 \\ &\quad + \left\|\partial_{x_i}\left(\frac{\Delta t^2}{2} e^{\Delta t \varepsilon^2 \Delta} \overline{g}(u^m)\right) - \partial_{x_i}\left(\frac{\Delta t^2}{2} e^{\Delta t \varepsilon^2 \Delta} \overline{g}(u^{m-1})\right)\right\|_{L^\infty} \\ &\le (1+\alpha_1 \Delta t)\|u_{x_i}^{m} - u_{x_i}^{m-1}\|_{L^\infty} + \left[\alpha_2 c_9 (2C + L) + c_9(\alpha_1^2 + 2\alpha_0\alpha_2)\right]\Delta t^2,\end{aligned}$$

which finishes the proof. □

Theorem 2.44. Assume that the initial data $u_0 \in \mathcal{C}^1(\overline{\Omega})$ with $\|u_0\|_{L^\infty} \le \beta$ and the exact solution u of (2.62) satisfies the conditions stated in Theorem 2.37 (resp. Theorem 2.39), then the numerical solution $\{u^m\}_{0 \le m \le M}$ generated by the LRI1a or LRI1b scheme (resp. the LRI2 scheme) satisfies

$$E(u^m) \leq E(u_0) + C,$$

where $C > 0$ is independent of Δt.

Proof. By the definition of E, it follows immediately that

$$E(u^{m+1}) - E(u^m) = \int_\Omega \left[\frac{\varepsilon^2}{2}(|\nabla u^{m+1}|^2 - |\nabla u^m|^2) + F(u^{m+1}) - F(u^m) \right] d\mathbf{x}. \tag{2.103}$$

Note that $F' = -f$ and $\|u^{m+1} - u^m\|_{L^\infty} \leq (2C + L)\Delta t$ by (2.102), hence

$$\int_\Omega [F(u^{m+1}) - F(u^m)]\, d\mathbf{x} \leq \alpha_0 |\Omega| \|u^{m+1} - u^m\|_{L^\infty} \leq \alpha_0 |\Omega| (2C + L)\Delta t. \tag{2.104}$$

On the other hand, it is implied from Lemma 2.43 that

$$\begin{aligned}
\int_\Omega \frac{\varepsilon^2}{2}(|\nabla u^{m+1}|^2 - |\nabla u^m|^2)\, d\mathbf{x} &= \frac{\varepsilon^2}{2} \sum_{i=1}^{d} \int_\Omega (u_{x_i}^{m+1} - u_{x_i}^m)(u_{x_i}^{m+1} + u_{x_i}^m)\, d\mathbf{x} \\
&\leq \frac{\varepsilon^2}{2} \sum_{i=1}^{d} \int_\Omega \|u_{x_i}^{m+1} - u_{x_i}^m\|_{L^\infty} (\|u_{x_i}^{m+1}\|_{L^\infty} + \|u_{x_i}^m\|_{L^\infty})\, d\mathbf{x} \\
&\leq \varepsilon^2 d c_9 c_{10} |\Omega| \Delta t. \qquad (2.105)
\end{aligned}$$

Combining (2.103), (2.104) and (2.105) leads to

$$E(u^{m+1}) - E(u^m) \leq \left[\alpha_0 (2C + L) + \varepsilon^2 d c_9 c_{10} \right] |\Omega| \Delta t,$$

which is the desired estimate. □

2.3.5 Numerical results

We carry out several numerical examples to investigate the convergence of the proposed LRI schemes and three important properties of the discrete solution, namely mass conservation, the MBP and energy stability. While there are various types of spatial discretization that can be coupled with LRIs, we use the central finite difference method to take advantage of FFT-based fast implementation [55, 104, 114, 204]. In all test cases below, the periodic boundary conditions are considered over the spatial domain $\Omega = (-0.5, 0.5)^2$ in two dimensions (2D) and $\Omega = (-0.5, 0.5)^3$ in three dimensions (3D); similar results can be obtained with the homogeneous Neumann boundary conditions. The nonlinear function is either one of the following two types:

$$f_1(u) = u - u^3, \quad \text{and} \quad f_2(u) = \frac{\theta}{2} \ln \frac{1-u}{1+u} + \theta_c u,$$

corresponding to the double-well and Flory-Huggins potential functions, respectively. For the latter, we choose $\theta = 0.8$ and $\theta_c = 1.6$.

Convergence tests We consider the 2D conservative Allen-Cahn equation (2.62) with the initial data given by

$$u_0(x, y) = \frac{1}{2} \cos(2\pi x) \cos(2\pi y), \quad (x, y) \in \overline{\Omega},$$

so that $\|u_0\|_{L^\infty} \leq \beta$, where $\beta \in \{2/\sqrt{3}, 0.987\}$ is a uniform bound mentioned in Remark 2.34. The spatial mesh size $h = h_x = h_y = 1/2048$ is fixed, where h_x and h_y denote the mesh sizes in x and y directions, respectively.

For the polynomial potential function $f = f_1$, the interfacial parameter ε is chosen to be $0.01, 0.005$ or 0.0025 and the final time T is given by $1, 2$ or 4, respectively. We vary the time step size $\Delta t \in \{T/16, T/32, \ldots, T/256\}$ and compute both L^2 and L^∞ errors of the numerical solutions produced by first and second-order LRI schemes at the final time step. The reference solution is obtained using the LRI2 scheme with a refined time step size $\Delta t = T/1024$. To investigate the performance of the LRI schemes, especially LRI2, we compare them with the ETD methods of the same order. The L^2 and L^∞ errors and convergence rates of the first-order ETD and LRI schemes are presented in Tables 2.7 and 2.8, respectively. We clearly observe the first-order convergence of the ETD1, LRI1a and LRI1b schemes, and the errors increase as ε gets closer to zero and T becomes larger simultaneously. Both tables show that the errors produced by the ETD1 scheme are slightly smaller than those generated by the LRI1a scheme, but slightly larger than those produced by the LRI1b scheme. This observation leads to the numerical conclusion that the LRI1b scheme is the best among the three first-order schemes. For the second-order case, the

L^2 and L^∞ errors and the corresponding convergence rates by ETDRK2 and LRI2 are reported in Table 2.9, where the difference between these two schemes becomes clearer. As ε goes to zero, L^2 errors by the LRI2 scheme are nearly half of those by the ETDRK2 scheme, while L^∞ errors by the LRI2 scheme are roughly three quarters of those by its counterpart. Thus, LRI2 performs much better than ETDRK2, especially when ε approaches zero; besides, the second-order accuracy are well observed for both schemes.

Table 2.7: L^2 errors and convergence rates by the first-order ETD and LRI schemes with the double-well potential and $(\varepsilon, T) \in \{(0.01, 1), (0.005, 2), (0.0025, 4)\}$.

ε	Δt	ETD1		LRI1a		LRI1b	
		L^2 error	Rate	L^2 error	Rate	L^2 error	Rate
0.01	$T/16$	4.86e-03		4.98e-03		4.85e-03	
	$T/32$	2.46e-03	0.99	2.52e-03	0.98	2.45e-03	0.99
	$T/64$	1.23e-03	0.99	1.27e-03	0.99	1.23e-03	0.99
	$T/128$	6.19e-04	1.00	6.35e-04	1.00	6.17e-04	1.00
	$T/256$	3.10e-04	1.00	3.18e-04	1.00	3.09e-04	1.00
0.005	$T/16$	1.79e-02		1.81e-02		1.78e-02	
	$T/32$	9.08e-03	0.98	9.21e-03	0.98	9.03e-03	0.98
	$T/64$	4.58e-03	0.99	4.64e-03	0.99	4.55e-03	0.99
	$T/128$	2.30e-03	0.99	2.33e-03	0.99	2.28e-03	0.99
	$T/256$	1.15e-03	1.00	1.17e-03	1.00	1.14e-03	1.00
0.0025	$T/16$	3.83e-02		3.93e-02		3.77e-02	
	$T/32$	1.93e-02	0.99	1.99e-02	0.98	1.88e-02	1.00
	$T/64$	9.64e-03	1.00	1.00e-02	0.99	9.40e-03	1.00
	$T/128$	4.82e-03	1.00	5.01e-03	1.00	4.69e-03	1.00
	$T/256$	2.41e-03	1.00	2.51e-03	1.00	2.34e-03	1.00

For the Flory-Huggins potential function $f = f_2$, we choose $\varepsilon = 0.0025$ or 0.001 and fix $T = 4$ for simplicity. The L^2 and L^∞ errors produced by the first-order schemes, i.e., ETD1, LRI1a and LRI1b, are presented in Tables 2.10 and 2.11, respectively. As expected, we observe the first-order convergence of all three schemes with a quite similar performance in terms of error magnitude. Table 2.12 shows the second-order convergence of both ETDRK2 and LRI2 schemes. However, L^2 errors by LRI2 are approximately half of those by ETDRK2, while L^∞ errors by LRI2 are about two thirds of those by its counterpart. Since T is fixed in this case, the errors produced by our first- and second-order LRI schemes do not change significantly with different values of ε (cf. Remarks 2.38 and 2.40(i)). We refer to [45] for the running time comparison of the LRI and

Table 2.8: L^∞ errors and convergence rates by the first-order ETD and LRI schemes with the double-well potential and $(\varepsilon, T) \in \{(0.01, 1), (0.005, 2), (0.0025, 4)\}$.

ε	Δt	ETD1		LRI1a		LRI1b	
		L^∞ error	Rate	L^∞ error	Rate	L^∞ error	Rate
0.01	$T/16$	7.77e-03		7.92e-03		7.79e-03	
	$T/32$	3.93e-03	0.98	4.01e-03	0.98	3.94e-03	0.98
	$T/64$	1.98e-03	0.99	2.02e-03	0.99	1.98e-03	0.99
	$T/128$	9.93e-04	1.00	1.01e-03	1.00	9.94e-04	1.00
	$T/256$	4.97e-04	1.00	5.07e-04	1.00	4.98e-04	1.00
0.005	$T/16$	3.50e-02		3.53e-02		3.50e-02	
	$T/32$	1.80e-02	0.96	1.81e-02	0.96	1.79e-02	0.96
	$T/64$	9.10e-03	0.98	9.17e-03	0.98	9.09e-03	0.98
	$T/128$	4.58e-03	0.99	4.61e-03	0.99	4.57e-03	0.99
	$T/256$	2.30e-03	1.00	2.31e-03	1.00	2.29e-03	1.00
0.0025	$T/16$	1.47e-01		1.47e-01		1.47e-01	
	$T/32$	7.75e-02	0.92	7.80e-02	0.92	7.74e-02	0.92
	$T/64$	3.98e-02	0.96	4.01e-02	0.96	3.97e-02	0.96
	$T/128$	2.02e-02	0.98	2.03e-02	0.98	2.01e-02	0.98
	$T/256$	1.01e-02	0.99	1.02e-02	0.99	1.01e-02	0.99

Table 2.9: Errors and convergence rates by the second-order ETD and LRI schemes with the double-well potential and $(\varepsilon, T) \in \{(0.01, 1), (0.005, 2), (0.0025, 4)\}$.

ε	Δt	ETDRK2				LRI2			
		L^2 error	Rate	L^∞ error	Rate	L^2 error	Rate	L^∞ error	Rate
0.01	$T/16$	1.41e-04		1.72e-04		1.34e-04		2.33e-04	
	$T/32$	3.56e-05	1.98	4.39e-05	1.97	3.38e-05	1.99	5.85e-05	2.00
	$T/64$	8.93e-06	2.00	1.11e-05	1.99	8.46e-06	2.00	1.46e-05	2.00
	$T/128$	2.21e-06	2.01	2.75e-06	2.01	2.09e-06	2.01	3.61e-06	2.02
	$T/256$	5.28e-07	2.07	6.56e-07	2.07	4.99e-07	2.07	8.60e-07	2.07
0.005	$T/16$	1.07e-03		1.49e-03		5.55e-04		9.85e-04	
	$T/32$	2.69e-04	1.98	3.90e-04	1.94	1.41e-04	1.98	2.55e-04	1.95
	$T/64$	6.76e-05	1.99	9.94e-05	1.97	3.53e-05	1.99	6.47e-05	1.98
	$T/128$	1.68e-05	2.01	2.48e-05	2.00	8.77e-06	2.01	1.61e-05	2.00
	$T/256$	4.00e-06	2.07	5.95e-06	2.06	2.09e-06	2.07	3.86e-06	2.06
0.0025	$T/16$	3.97e-03		1.19e-02		1.97e-03		9.17e-03	
	$T/32$	1.02e-03	1.96	3.25e-03	1.87	5.28e-04	1.90	2.46e-03	1.90
	$T/64$	2.59e-04	1.97	8.46e-04	1.94	1.37e-04	1.95	6.36e-04	1.95
	$T/128$	6.48e-05	2.00	2.14e-04	1.98	3.45e-05	1.99	1.60e-04	1.99
	$T/256$	1.55e-05	2.06	5.15e-05	2.05	8.29e-06	2.06	3.85e-05	2.06

ETD methods, where LRIs require less computational time compared to the ETD schemes of the same order.

Table 2.10: L^2 errors and convergence rates by the first-order ETD and LRI schemes with the Flory-Huggins potential and $T = 4$.

ε	Δt	ETD1		LRI1a		LRI1b	
		L^2 error	Rate	L^2 error	Rate	L^2 error	Rate
0.0025	$T/16$	4.45e-02		4.47e-02		4.42e-02	
	$T/32$	2.24e-02	0.99	2.27e-02	0.98	2.23e-02	0.99
	$T/64$	1.13e-02	0.99	1.15e-02	0.99	1.12e-02	0.99
	$T/128$	5.67e-03	0.99	5.75e-03	0.99	5.62e-03	1.00
	$T/256$	2.84e-03	1.00	2.88e-03	1.00	2.81e-03	1.00
0.001	$T/16$	4.56e-02		4.56e-02		4.56e-02	
	$T/32$	2.30e-02	0.99	2.31e-02	0.98	2.30e-02	0.99
	$T/64$	1.16e-02	0.98	1.17e-02	0.98	1.16e-02	0.99
	$T/128$	5.85e-03	0.99	5.86e-03	0.99	5.84e-03	0.99
	$T/256$	2.93e-03	1.00	2.94e-03	1.00	2.93e-03	1.00

Table 2.11: L^∞ errors and convergence rates by the first-order ETD and LRI schemes with the Flory-Huggins potential and $T = 4$.

ε	Δt	ETD1		LRI1a		LRI1b	
		L^∞ error	Rate	L^∞ error	Rate	L^∞ error	Rate
0.0025	$T/16$	1.37e-01		1.37e-01		1.37e-01	
	$T/32$	7.19e-02	0.93	7.21e-02	0.93	7.19e-02	0.93
	$T/64$	3.68e-02	0.97	3.69e-02	0.97	3.67e-02	0.97
	$T/128$	1.86e-02	0.98	1.87e-02	0.98	1.86e-02	0.98
	$T/256$	9.34e-03	0.99	9.38e-03	0.99	9.33e-03	0.99
0.001	$T/16$	1.38e-01		1.38e-01		1.38e-01	
	$T/32$	7.25e-02	0.93	7.26e-02	0.93	7.25e-02	0.93
	$T/64$	3.71e-02	0.97	3.71e-02	0.97	3.71e-02	0.97
	$T/128$	1.88e-02	0.98	1.88e-02	0.98	1.88e-02	0.98
	$T/256$	9.43e-03	0.99	9.43e-03	0.99	9.43e-03	0.99

Mass conservation, MBP and energy stability tests We now consider both 2D and 3D conservative Allen-Cahn equation (2.62) with a random initial data ranging from -0.9 and 0.9. In the 2D problem with the double-well potential, we test all LRI schemes in terms of mass conservation, MBP and energy stability. For other test cases, the LRI2 scheme is used due to its high accuracy.

We first simulate long time phase separation processes of the model equation in 2D with $\varepsilon = 0.01$ and the double-well potential function. Snapshots of the numerical solution at different

Table 2.12: Errors and convergence rates by the second-order ETD and LRI schemes with the Flory-Huggins potential and $T = 4$.

ε	Δt	ETDRK2				LRI2			
		L^2 error	Rate	L^∞ error	Rate	L^2 error	Rate	L^∞ error	Rate
0.0025	$T/16$	1.50e-02		2.49e-02		3.01e-03		6.97e-03	
	$T/32$	1.21e-03	3.63	3.55e-03	2.81	5.57e-04	2.43	1.84e-03	1.92
	$T/64$	2.77e-04	2.13	7.61e-04	2.22	1.39e-04	2.00	4.72e-04	1.96
	$T/128$	6.69e-05	2.05	1.79e-04	2.09	3.45e-05	2.01	1.18e-04	2.00
	$T/256$	1.58e-05	2.08	4.19e-05	2.10	8.26e-06	2.06	2.84e-05	2.06
0.001	$T/16$	1.53e-02		2.55e-02		3.07e-03		6.99e-03	
	$T/32$	1.25e-03	3.61	3.64e-03	2.81	5.61e-04	2.45	1.84e-03	1.92
	$T/64$	2.84e-04	2.14	7.78e-04	2.23	1.39e-04	2.01	4.72e-04	1.96
	$T/128$	6.84e-05	2.05	1.82e-04	2.09	3.45e-05	2.01	1.18e-04	2.00
	$T/256$	1.61e-05	2.08	4.26e-05	2.10	8.24e-06	2.07	2.84e-05	2.06

times $t = 1, 10, 20, 40, 100$ and 300 are depicted in Figure 2.4; the results are obtained by the LRI2 scheme with the time step size $\Delta t = 0.1$ and the spatial mesh size $h = h_x = h_y = 1/1024$. The evolution of the absolute mass error (the difference between the mass along the time and

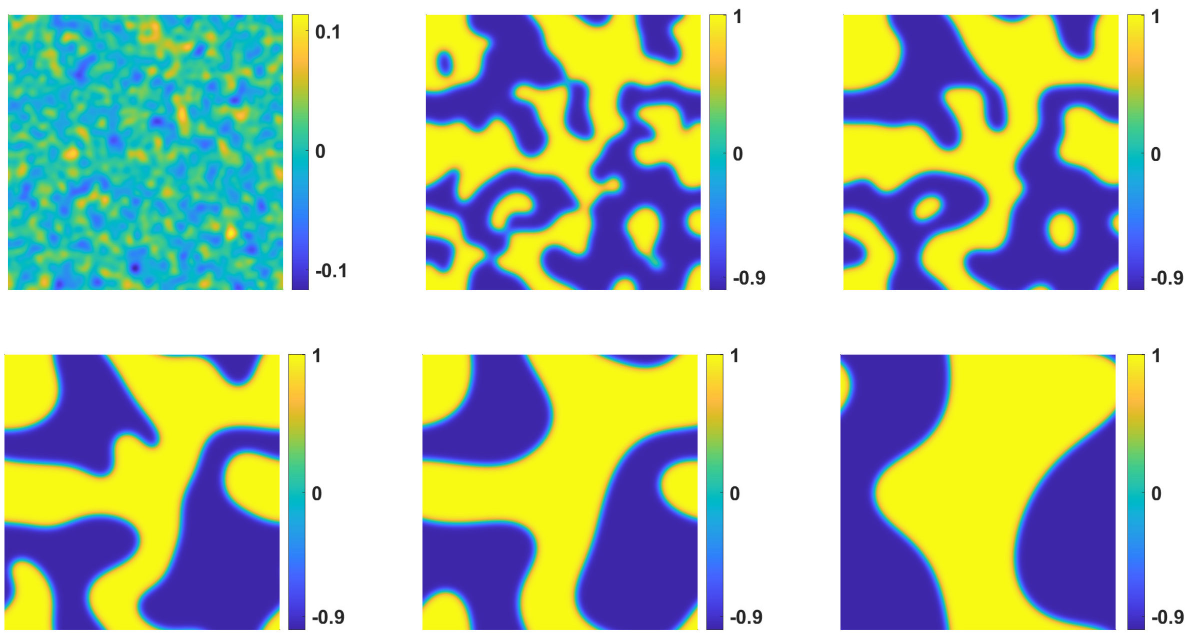


Figure 2.4: Snapshots of the numerical solution at $t = 1, 10, 20, 40, 100$ and 300 (from left to right and top to bottom) produced by the LRI2 scheme with $\Delta t = 0.1$ and the double-well potential.

the initial mass), the supremum norm and the discrete energy of the first- and second-order

LRI schemes with $\Delta t \in \{0.1, 0.2, 0.5\}$ and $T = 100$ is described in Figures 2.5, 2.6 and 2.7. We see that LRIs preserve mass and MBP, and decrease the energy functional over time. Since the supremum norm curves eventually reach values around 1, which is strictly less than $\beta = 2/\sqrt{3}$, we are allowed to take $\Delta t = 0.5$ in this case.

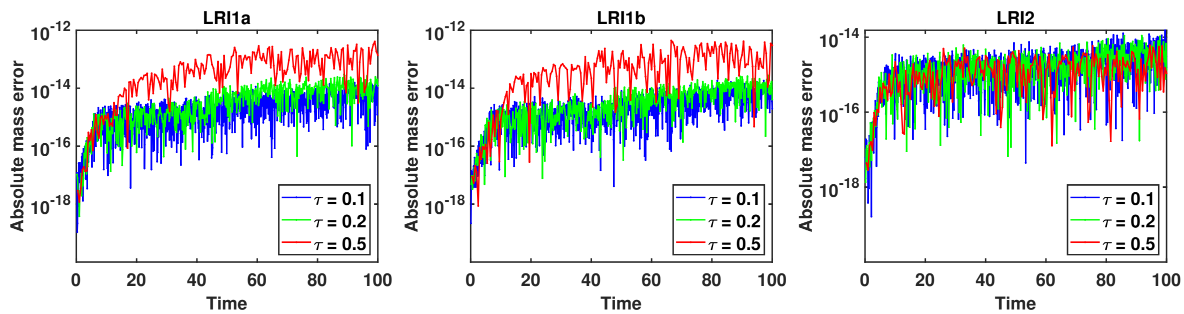


Figure 2.5: Evolution of the absolute mass error of the numerical solutions by the LRI schemes with $\Delta t = 0.1, 0.2, 0.5$ and the double-well potential.

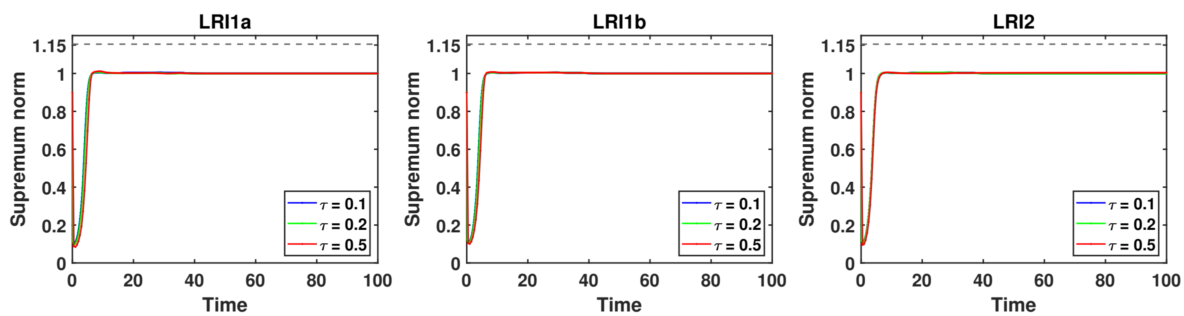


Figure 2.6: Evolution of the supremum norm of the numerical solutions by the LRI schemes with $\Delta t = 0.1, 0.2, 0.5$ and the double-well potential.

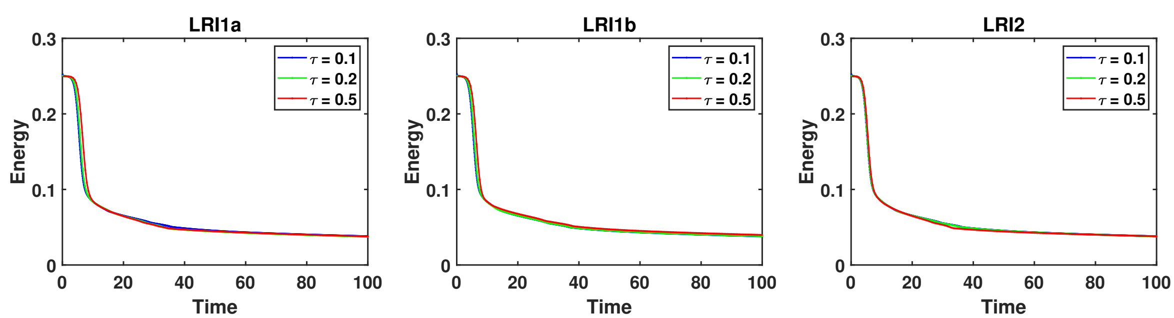


Figure 2.7: Evolution of the energy of the numerical solutions by the LRI schemes with $\Delta t = 0.1, 0.2, 0.5$ and the double-well potential.

We now perform the simulations with the Flory-Huggins potential. Figure 2.8 shows snapshots of the numerical solution by LRI2 with $\Delta t = 0.1$ at time $t = 1, 10, 20, 40, 100$ and 300. The corresponding absolute mass error, supremum norm and energy of the numerical solutions with $\Delta t \in \{0.1, 0.2\}$ are illustrated in Figure 2.9. All expected properties are again satisfied. It should be noted that $\Delta t = 0.5$ does not work in this case because the condition on Δt is more restrictive than the case of double-well potential (cf. Remark 2.34).

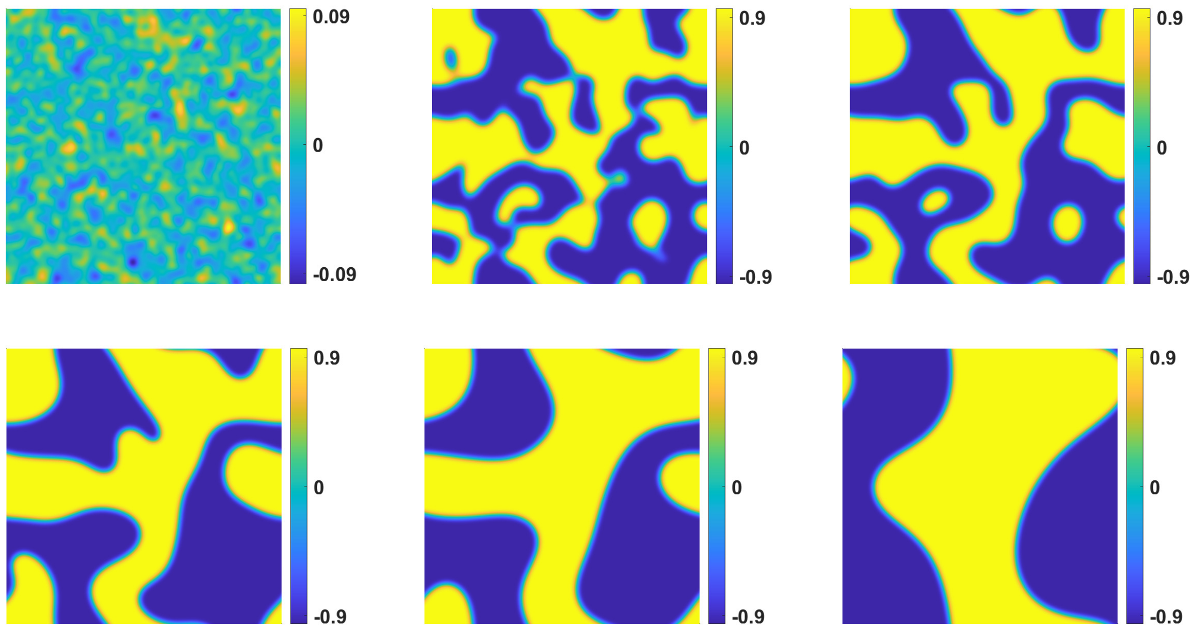


Figure 2.8: Snapshots of the numerical solution at $t = 1$, 10, 20, 40, 100 and 300 (from left to right and top to bottom) produced by the LRI2 scheme with $\Delta t = 0.1$ and the Flory-Huggins potential.

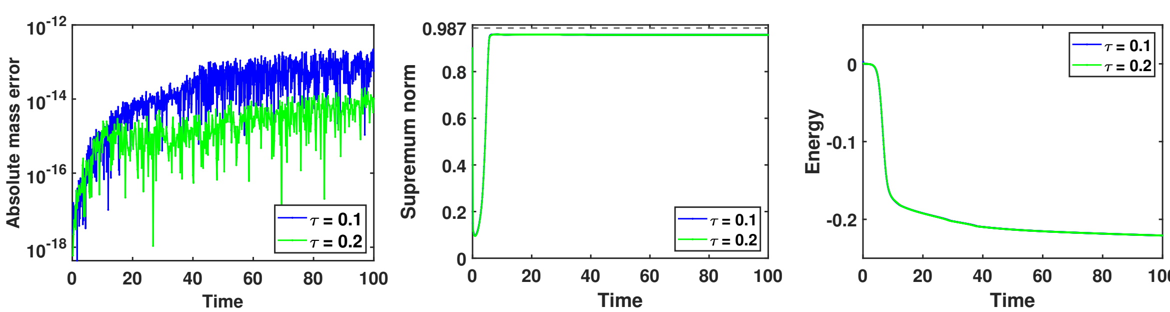


Figure 2.9: Evolution of the absolute mass error (left), the supremum norm (middle) and the energy (right) of the numerical solutions by the LRI2 scheme with $\Delta t = 0.1, 0.2$ and the Flory-Huggins potential.

Next, we proceed with 3D simulations of the conservative Allen-Cahn equation with fixed

$\varepsilon = 0.01$ and spatial mesh size $h = h_x = h_y = h_z = 1/128$. For the double-well potential, the evolution of 3D isosurfaces computed by the LRI2 scheme with $\Delta t = 0.1$ at $t = 1, 40, 100$ and 300 is presented in Figure 2.10. Figure 2.11 demonstrates mass conservation, the MBP property and the energy decay of the LRI2 scheme with different time step sizes $\Delta t \in \{0.1, 0.2, 0.5\}$. For the Flory-Huggins potential, isosurface snapshots by the LRI2 scheme with $\Delta t = 0.1$ at $t = 1, 40, 100$ and 300 are shown in Figure 2.12. All three properties are verified in Figure 2.13 with $\Delta t = 0.1$ and $\Delta t = 0.2$.

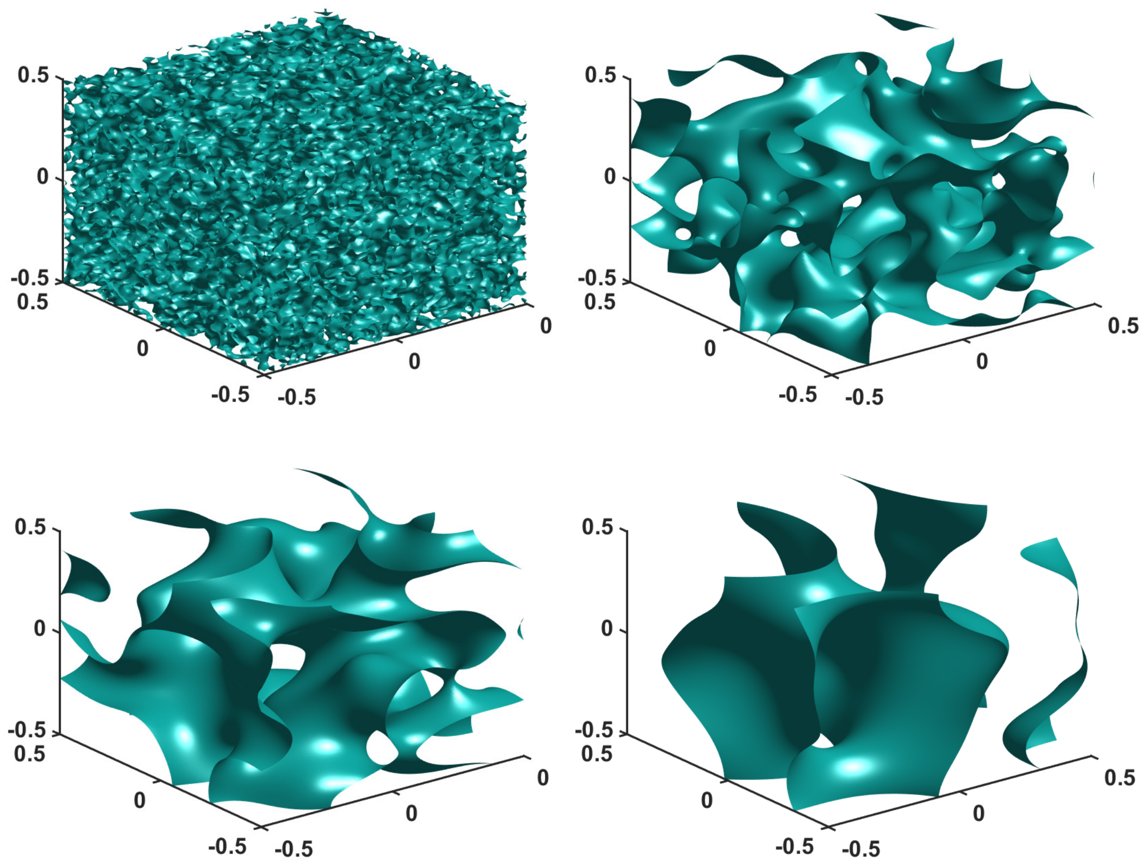


Figure 2.10: Isosurface snapshots of the numerical solution at $t = 1$, 40, 100 and 300 (from left to right and top to bottom) produced by the LRI2 scheme with $\Delta t = 0.1$ and the double-well potential.

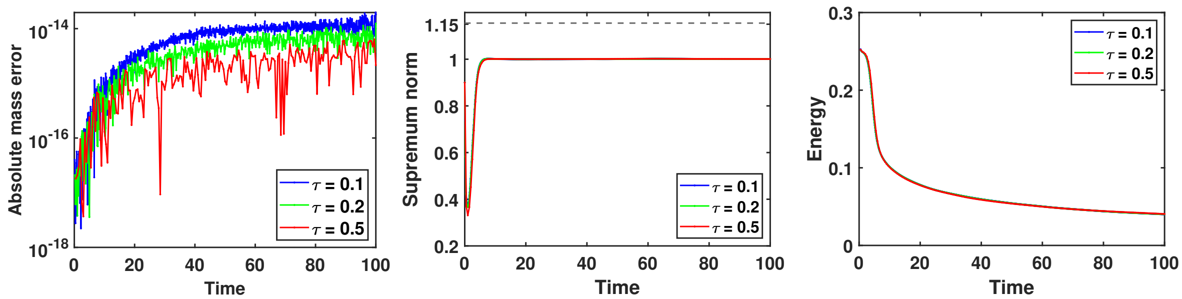


Figure 2.11: Evolution of the absolute mass error (left), the supremum norm (middle) and the energy (right) of the numerical solutions by the LRI2 scheme with $\Delta t = 0.1, 0.2, 0.5$ and the double-well potential in 3D.

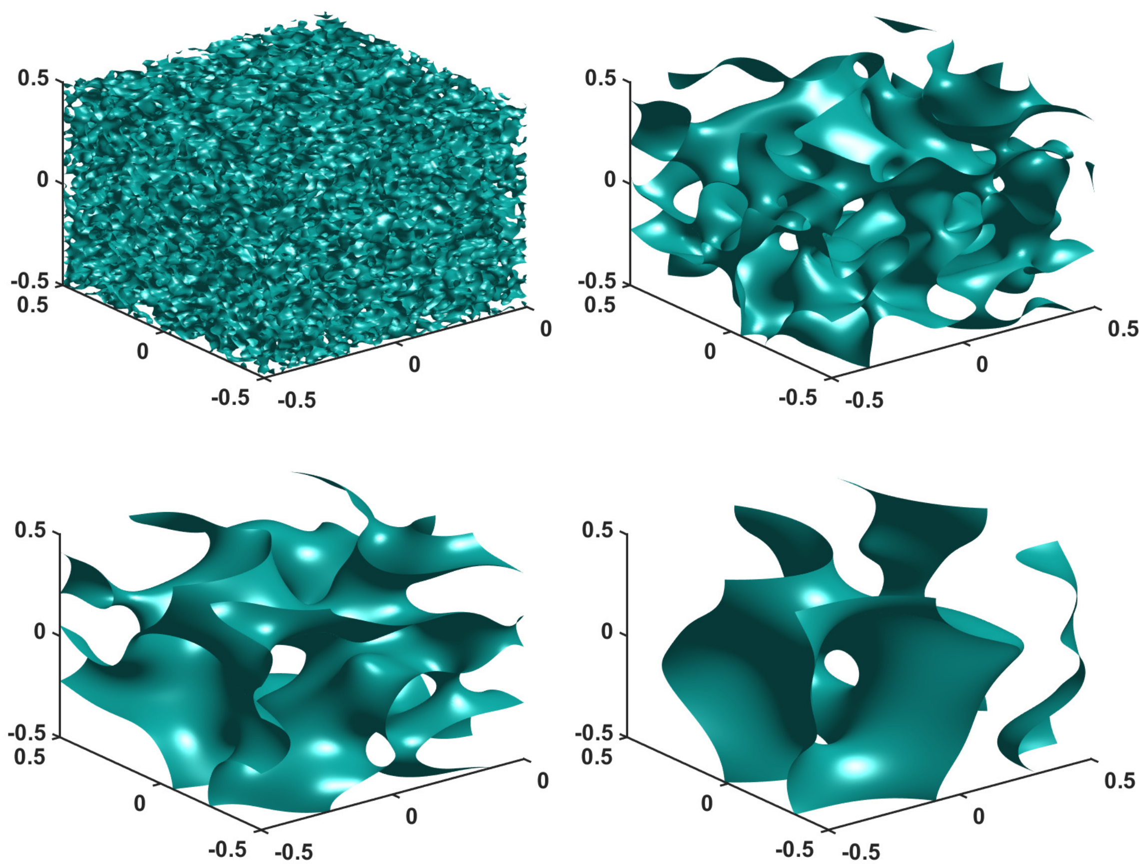


Figure 2.12: Isosurface snapshots of the numerical solution at $t = 1$, 40, 100 and 300 (from left to right and top to bottom) produced by the LRI2 scheme with $\Delta t = 0.1$ and the Flory-Huggins potential.

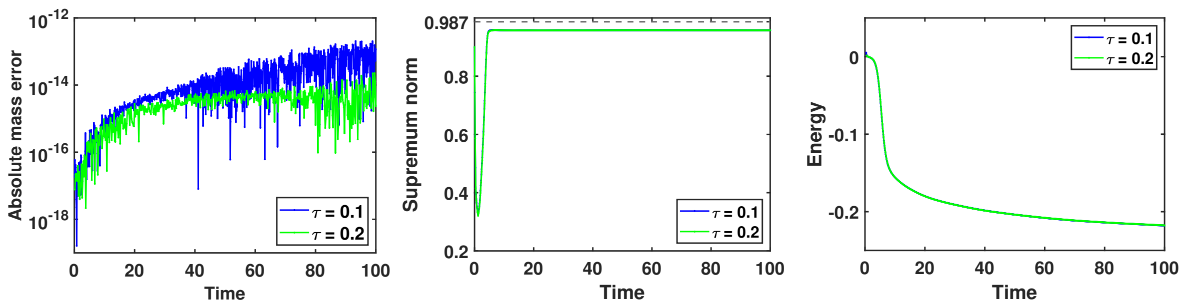


Figure 2.13: Evolution of the absolute mass error (left), the supremum norm (middle) and the energy (right) of the numerical solutions by the LRI2 scheme with $\Delta t = 0.1, 0.2$ and the Flory-Huggins potential in 3D.

Chapter 3

Noniterative Localized Exponential Time Differencing Methods for Hyperbolic Conservation Laws

In this chapter, we develop and analyze noniterative localized exponential time differencing (ETD) methods for scalar hyperbolic conservation laws. The model problem is first discretized in space by the discontinuous Galerkin (DG) method, resulting in a system of nonlinear ordinary differential equations. To solve such a system, the second-order ETDRK scheme is employed with Jacobian linearization at each time step, which can be combined with the standard TVB corrected slope limiter to effectively handle moving shocks. To reduce the computational cost of matrix exponential vector products, we propose a noniterative, nonoverlapping domain decomposition algorithm in which the local Jacobian matrix is taken as the restriction of the global Jacobian to each subdomain. The resulting localized ETDRK2 scheme requires no iteration between subdomains while still enabling the same large time step sizes as the global ETDRK2 scheme. Rigorous semi-discrete temporal error analysis is carried out for both the global and localized schemes, and discrete mass conservation is established under periodic boundary conditions. Numerical experiments for the Burgers' equation in one and two dimensions confirm the theoretical results and illustrate the accuracy and efficiency of the proposed methods. This chapter is based on the work in [52].

3.1 Model problem and spatial discretization

Let Ω be an open, bounded subset of $\mathbb{R}^d$ ($d = 1, 2, 3$) and $T > 0$, consider the following initial-boundary value problem for scalar conservation laws:

$$\partial_t u + \nabla \cdot \mathbf{F}(u) = 0, \quad \text{in } \Omega \times (0, T), \tag{3.1a}$$

$$u = \gamma, \quad \text{on } \partial\Omega \times (0, T), \tag{3.1b}$$

$$u(\cdot, 0) = u_0, \quad \text{in } \Omega, \tag{3.1c}$$

where $\mathbf{F} = (f_1, f_2, \dots, f_d)^T \in (\mathcal{C}^1(\mathbb{R}))^d$, $\gamma \in L^\infty(\partial\Omega \times (0, T))$ and $u_0 \in L^\infty(\Omega)$ are given (nonlinear) functions. We refer to [7] for a correct interpretation of the boundary condition (3.1b); periodic conditions will also be discussed later (cf. Section 3.5).

To construct numerical approximations of the weak solution u to (3.1), we first discretize the equation using the DG method. For simplicity of presentation, we consider the two-dimensional problem (i.e., $d = 2$) with $\Omega = (0, 1)^2$ and $\mathbf{F} = (f_1, f_2)^T$. Let $\mathcal{K}_h$ be a finite element partition of Ω into $N_x N_y$ rectangles $K_{i,j} = (x_{i-1/2}, x_{i+1/2}) \times (y_{j-1/2}, y_{j+1/2})$, for $1 \leq i \leq N_x$, $1 \leq j \leq N_y$, where

$$0 = x_{1/2} < x_{3/2} < \dots < x_{N_x+1/2} = 1, \text{ and } 0 = y_{1/2} < y_{3/2} < \dots < y_{N_y+1/2} = 1.$$

Denote by x_i and y_j the midpoints of the intervals $(x_{i-1/2}, x_{i+1/2})$ and $(y_{j-1/2}, y_{j+1/2})$, respectively, and define $\Delta x_i = x_{i+1/2} - x_{i-1/2}$ and $\Delta y_j = y_{j+1/2} - y_{j-1/2}$. Let $\Delta x = \max_{1 \leq i \leq N_x} \Delta x_i$, $\Delta y = \max_{1 \leq j \leq N_y} \Delta y_j$, and $h = \max\{\Delta x, \Delta y\}$. In this chapter, we consider the finite dimensional space $\mathcal{V}_h$ consisting of discontinuous piecewise linear polynomials:

$$\mathcal{V}_h = \{v_h \in L^1(\Omega) : v_h|_{K_{i,j}} \in \mathcal{P}^1(K_{i,j}),\ 1 \leq i \leq N_x,\ 1 \leq j \leq N_y\},$$

where $\mathcal{P}^1(K_{i,j})$ represents the set of polynomials of degree at most 1 on $K_{i,j}$; for higher order DG methods, we refer to [32]. Let $u_h \in \mathcal{V}_h$ be an approximation of the weak solution u to (3.1), which is uniquely expressed as

$$u_h(x,y,t) = u_{i,j}^{00}(t) + u_{i,j}^{10}(t)\phi_{i,j}^{10}(x,y) + u_{i,j}^{01}(t)\phi_{i,j}^{01}(x,y), \quad \text{for } (x,y) \in K_{i,j},\ t \in (0,T), \tag{3.2}$$

where $\phi_{i,j}^{lk}$, for $(l,k) \in \{(0,0),(1,0),(0,1)\}$, are the local orthogonal basis functions given by

$$\phi_{i,j}^{00}(x,y) = 1,\ \phi_{i,j}^{10}(x,y) = \frac{2(x-x_i)}{\Delta x_i},\ \text{and } \phi_{i,j}^{01}(x,y) = \frac{2(y-y_j)}{\Delta y_j}.$$

The degrees of freedom (DOFs) $u_{i,j}^{lk}$ are defined as

$$\begin{aligned} u_{i,j}^{00}(t) &= \frac{1}{|K_{i,j}|}\int_{K_{i,j}} u(x,y,t)\,dx\,dy, \quad u_{i,j}^{10}(t) = \frac{3}{|K_{i,j}|}\int_{K_{i,j}} u(x,y,t)\phi_{i,j}^{10}(x,y)\,dx\,dy, \\ u_{i,j}^{01}(t) &= \frac{3}{|K_{i,j}|}\int_{K_{i,j}} u(x,y,t)\phi_{i,j}^{01}(x,y)\,dx\,dy, \end{aligned} \tag{3.3}$$

i.e., u_h is the L^2-projection of the exact solution u into $\mathcal{V}_h$.

The DG formulation is obtained by multiplying equation (3.1a) by $v_h \in \mathcal{V}_h$, integrating by parts over each element $K_{i,j}$ and replacing the exact solution u by u_h:

$$\frac{d}{dt}\int_{K_{i,j}} u_h v_h\,dxdy = \int_{K_{i,j}} \mathbf{F}(u_h)\cdot\nabla v_h\,dxdy - \sum_{e\subset\partial K_{i,j}}\int_e \mathbf{F}(u_h)\cdot\mathbf{n}_{e,K_{i,j}} v_h\,ds, \tag{3.4}$$

where $\mathbf{n}_{e,K}$ is the outward unit normal vector to the edge $e \subset \partial K$, for $K \in \mathcal{K}_h$. Since u_h is discontinuous at $\mathbf{x} = (x,y) \in e \subset \partial K$, we replace $\mathbf{F}(u_h)\cdot\mathbf{n}_{e,K}$ in (3.4) by a numerical flux $\mathfrak{h}_{e,K}\left(u_h^{\text{int}(K)}(\mathbf{x},t), u_h^{\text{ext}(K)}(\mathbf{x},t)\right)$ with

$$u_h^{\text{int}(K)}(\mathbf{x},t) = \lim_{\substack{\mathbf{y}\to\mathbf{x}\\ \mathbf{y}\in\mathring{K}}} u_h(\mathbf{y},t),\ \text{and} \quad u_h^{\text{ext}(K)}(\mathbf{x},t) = \begin{cases} \gamma_h(\mathbf{x},t) & \text{if } \mathbf{x}\in\partial\Omega, \\ \lim\limits_{\substack{\mathbf{y}\to\mathbf{x}\\ \mathbf{y}\in(K)^c}} u_h(\mathbf{y},t) & \text{otherwise,} \end{cases} \tag{3.5}$$

where γ_h is the L^2-projection of γ into $\partial\mathcal{V}_h$. Similar to [32], the numerical flux $\mathfrak{h}_{e,K}$ is assumed to satisfy the following conditions:

(i) $\mathfrak{h}_{e,K}(u,u) = \mathbf{F}(u)\cdot\mathbf{n}_{e,K}$.

(ii) $\mathfrak{h}_{e,K}$ is Lipschitz continuous.

(iii) $\mathfrak{h}_{e,K}(u,v)$ is nondecreasing in u and nonincreasing in v.

(iv) For all pairs of neighboring elements (K,K') and $\mathbf{x} \in e = \partial K \cap \partial K'$, we have

$$\mathfrak{h}_{e,K}\left(u_h^{\mathrm{int}(K)}(\mathbf{x},t), u_h^{\mathrm{ext}(K)}(\mathbf{x},t)\right) = -\mathfrak{h}_{e,K'}\left(u_h^{\mathrm{int}(K')}(\mathbf{x},t), u_h^{\mathrm{ext}(K')}(\mathbf{x},t)\right). \tag{3.6}$$

In this chapter, we further assume that the numerical flux is sufficiently smooth in order to construct the approximations in time with exponential integrators (cf. Section 3.2). Examples of such numerical fluxes include the Lax-Friedrichs flux:

$$\mathfrak{h}_{e,K}^{\mathrm{LF}}(u,v) = \frac{1}{2}[\mathbf{F}(u) + \mathbf{F}(v)] \cdot \mathbf{n}_{e,K} - \frac{\alpha}{2}(v-u), \text{ for } \alpha = \max_{w \in [\inf u_0, \sup u_0]} \left|\frac{d}{dw}(\mathbf{F}(w) \cdot \mathbf{n}_{e,K})\right|, \tag{3.7}$$

and the Engquist-Osher flux:

$$\mathfrak{h}_{e,K}^{\mathrm{EO}}(u,v) = \frac{1}{2}[\mathbf{F}(u) + \mathbf{F}(v)] \cdot \mathbf{n}_{e,K} - \frac{1}{2}\int_u^v \left|\frac{d}{dw}(\mathbf{F}(w) \cdot \mathbf{n}_{e,K})\right| dw. \tag{3.8}$$

Next, we let $v_h = \phi_{i,j}^{lk}$ in (3.4) and use (3.2) as well as the orthogonality of the basis functions to obtain

$$\frac{d}{dt}u_{i,j}^{lk}\int_{K_{i,j}} (\phi_{i,j}^{lk})^2\, dxdy = \int_{K_{i,j}} \mathbf{F}(u_h) \cdot \nabla\phi_{i,j}^{lk}\, dxdy - \sum_{e \subset \partial K_{i,j}} \int_e \mathfrak{h}_{e,K_{i,j}}\left(u_h^{\mathrm{int}(K_{i,j})}, u_h^{\mathrm{ext}(K_{i,j})}\right)\phi_{i,j}^{lk} ds,$$

for $1 \le i \le N_x, 1 \le j \le N_y$ and $(l,k) \in \{(0,0),(1,0),(0,1)\}$. As $\int_{K_{i,j}} (\phi_{i,j}^{lk})^2 dxdy = \frac{|K_{i,j}|}{3^{l+k}}$ and $\nabla\phi_{i,j}^{lk} = \left(\frac{2l}{\Delta x_i}, \frac{2k}{\Delta y_j}\right)^T$, we deduce that

$$\begin{aligned}\frac{d}{dt}u_{i,j}^{lk} = \frac{3^{l+k}}{|K_{i,j}|}\bigg[\frac{2l}{\Delta x_i}\int_{K_{i,j}} f_1(u_h) dxdy + \frac{2k}{\Delta y_j}\int_{K_{i,j}} f_2(u_h) dxdy \\ - \sum_{e \subset \partial K_{i,j}} \int_e \mathfrak{h}_{e,K_{i,j}}\left(u_h^{\mathrm{int}(K_{i,j})}, u_h^{\mathrm{ext}(K_{i,j})}\right)\phi_{i,j}^{lk} ds\bigg].\end{aligned} \tag{3.9}$$

The integrals on the right-hand side are computed by quadrature rules:

$$
\begin{aligned}
&\int_K f(u_h(x,y,t))dxdy \approx \mathcal{I}_K(f), \quad \text{with } f = f_1 \text{ or } f = f_2,\\
&\int_e \mathfrak{h}_{e,K}\left(u_h^{\mathrm{int}(K)}(x,y,t), u_h^{\mathrm{ext}(K)}(x,y,t)\right)\phi(x,y)\,ds \approx \mathcal{I}_e(\mathfrak{h}_{e,K}\phi), \quad \text{with } \phi = \phi_{i,j}^{lk},
\end{aligned}
\tag{3.10}
$$

for all elements $K = K_{i,j} \in \mathcal{K}_h$ and edges $e \subset \partial K$, where

$$
\begin{aligned}
\mathcal{I}_K(f) &:= \sum_{q=1}^{N_q} \omega_q f(u_h(x_{Kq}, y_{Kq}, t))|K|,\\
\mathcal{I}_e(\mathfrak{h}_{e,K}\phi) &:= \sum_{q=1}^{\widehat{N}_q} \widehat{\omega}_q \mathfrak{h}_{e,K}\left(u_h^{\mathrm{int}(K)}(x_{eq}, y_{eq}, t), u_h^{\mathrm{ext}(K)}(x_{eq}, y_{eq}, t)\right)\phi(x_{eq}, y_{eq})|e|.
\end{aligned}
\tag{3.11}
$$

Denote by $\mathbf{u}_{i,j}(t) = (u_{i,j}^{00}(t), u_{i,j}^{10}(t), u_{i,j}^{01}(t))^T$ the DOFs of $u_h(t)$ on element $K_{i,j}$, for $1 \le i \le N_x$, $1 \le j \le N_y$. We assemble the DOFs over all elements in the x-direction as follows:

$$
\begin{aligned}
\mathbf{U}_h(t) &= (\mathbf{U}_1(t), \mathbf{U}_2(t), \dots, \mathbf{U}_{N_x}(t))^T,\\
\text{where} \quad \mathbf{U}_i &= (\mathbf{u}_{i,1}, \mathbf{u}_{i,2}, \dots, \mathbf{u}_{i,N_y})^T \in \mathbb{R}^{3N_y}, \text{ for } 1 \le i \le N_x.
\end{aligned}
\tag{3.12}
$$

Using this notation, we obtain from (3.9)-(3.10) the weak formulation:

$$
\frac{d}{dt}u_{i,j}^{lk} = g_{i,j}^{lk}(\mathbf{U}_h, \gamma_h) := \frac{3^{l+k}}{|K_{i,j}|}\left[\frac{2l}{\Delta x_i}\mathcal{I}_{K_{i,j}}(f_1(u_h)) + \frac{2k}{\Delta y_j}\mathcal{I}_{K_{i,j}}(f_2(u_h)) - \sum_{e\subset\partial K_{i,j}} \mathcal{I}_e\left(\mathfrak{h}_{e,K_{i,j}}\phi_{i,j}^{lk}\right)\right],
\tag{3.13}
$$

for $1 \le i \le N_x$, $1 \le j \le N_y$ and $(l,k) \in \{(0,0),(1,0),(0,1)\}$. As for the DOFs of u_h, we define

$$
\begin{aligned}
\mathbf{G}_h &= (\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_{N_x})^T, \quad \mathbf{G}_i = (\mathbf{g}_{i,1}, \mathbf{g}_{i,2}, \dots, \mathbf{g}_{i,N_y})^T,\\
\text{where} \quad &\mathbf{g}_{i,j} = (g_{i,j}^{00}, g_{i,j}^{10}, g_{i,j}^{01})^T, \quad 1 \le i \le N_x,\ 1 \le j \le N_y.
\end{aligned}
\tag{3.14}
$$

The nonlinear equations (3.13) can be recast in the ODE form as follows:

$$
\frac{d}{dt}\mathbf{U}_h = \mathbf{G}_h(\mathbf{U}_h, \gamma_h), \quad \mathbf{U}_h(0) = \mathbf{U}_{h,0},
\tag{3.15}
$$

where $\mathbf{U}_{h,0}$ is the vector of DOFs associated with $u_{h,0}$, the L^2-projection of u_0 into $\mathcal{V}_h$ (cf. (3.3)).

3.2 Exponential time differencing Runge-Kutta method

To solve the nonlinear ODE system (3.15), we consider exponential Rosenbrock-type methods [87] which linearize the equations at each time step via the Jacobian and make use of matrix exponentials and exponential-like φ—functions. These methods are fully explicit and highly accurate while allowing larger time step sizes, compared to standard explicit time-stepping schemes such as Strong Stability Preserving Runge-Kutta (SSP-RK). In the following, we apply the exponential time differencing Runge-Kutta method of order 2 (ETDRK2) to match the second-order DG approximation in space. It should be noted that the exponential Rosenbrock-Euler method, equivalent to the first-order ETD method, is second-order accurate only for autonomous problems [86]. The ODE system (3.15) is nonautonomous in general, especially when a domain decomposition approach is considered (cf. Section 3.3).

3.2.1 Time discretization by ETDRK2

Let $0 = t_0 < t_1 < \ldots < t_M = T$ be a uniform partition of the time interval $[0, T]$ with the time step size $\Delta t = T/M$, and denote by $\mathbf{U}_h^m$ an approximation of $\mathbf{U}_h(t_m)$. For given $\mathbf{U}_h^m$, a linearization of (3.15) is

$$\frac{d}{dt}\mathbf{U}_h = \mathbf{L}_h^m \mathbf{U}_h + \mathbf{R}_h^m(\mathbf{U}_h, \gamma_h), \tag{3.16}$$

where $\mathbf{L}_h^m = \mathbf{L}_h(\mathbf{U}_h^m)$ is some linear operator and $\mathbf{R}_h^m(\mathbf{U}_h, \gamma_h) = \mathbf{G}_h(\mathbf{U}_h, \gamma_h) - \mathbf{L}_h^m \mathbf{U}_h$ is the nonlinear remainder. A common choice for $\mathbf{L}_h^m$ is the Jacobian of $\mathbf{G}_h$ evaluated at t_m:

$$\mathbf{L}_h^m := \mathbf{J}_h^m = \frac{\partial \mathbf{G}_h}{\partial \mathbf{U}_h}(\mathbf{U}_h^m). \tag{3.17}$$

For $\mathbf{U}_h$ given by (3.12) and $\mathbf{G}_h$ by (3.14), $\mathbf{L}_h^m$ is a block tridiagonal matrix of the form:

$$\mathbf{L}_h^m = \frac{\partial \mathbf{G}_h}{\partial \mathbf{U}_h}(\mathbf{U}_h^m) = \begin{pmatrix} \mathbf{A}_{1,1}^m & \mathbf{A}_{1,2}^m & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{A}_{2,1}^m & \mathbf{A}_{2,2}^m & \mathbf{A}_{2,3}^m & \cdots & \mathbf{O} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{O} & \cdots & \mathbf{A}_{N_x-1,N_x-2}^m & \mathbf{A}_{N_x-1,N_x-1}^m & \mathbf{A}_{N_x-1,N_x}^m \\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{A}_{N_x,N_x-1}^m & \mathbf{A}_{N_x,N_x}^m \end{pmatrix}, \tag{3.18}$$

where each block is a $3N_y \times 3N_y$ matrix and

$$\mathbf{A}_{i,i'}^m := \frac{\partial \mathbf{G}_i}{\partial \mathbf{U}_{i'}}(\mathbf{U}_{i'}^m), \quad 1 \le i \le N_x, \quad i' = i-1, i, \text{ or } i+1. \tag{3.19}$$

In general, the matrices $\mathbf{A}_{i,i}^m$ $(1 \le i \le N_x)$ are also block tridiagonal while $\mathbf{A}_{i,i+1}^m$ $(1 \le i \le N_x - 1)$ and $\mathbf{A}_{i-1,i}^m$ $(2 \le i \le N_x)$ are block diagonal.

To find a solution to (3.16) at $t = t_{m+1}$, we apply Duhamel's principle and obtain

$$\mathbf{U}_h(t_{m+1}) = e^{\Delta t \mathbf{L}_h^m}\mathbf{U}_h(t_m) + \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{L}_h^m}\mathbf{R}_h^m(\mathbf{U}_h(t_m + s), \gamma_h(t_m + s))ds. \tag{3.20}$$

The ETDRK2 scheme is derived by approximating $\mathbf{R}_h^m(\mathbf{U}_h, \gamma_h)$ on the interval $[t_m, t_{m+1}]$ by its linear interpolation [38, 86]:

$$\widetilde{\mathbf{U}}_h^{m+1} = \mathbf{U}_h^m + \Delta t \varphi_1(\Delta t \mathbf{L}_h^m)\mathbf{G}_h(\mathbf{U}_h^m, \gamma_h^m), \tag{3.21a}$$

$$\mathbf{U}_h^{m+1} = \widetilde{\mathbf{U}}_h^{m+1} + \Delta t \varphi_2(\Delta t \mathbf{L}_h^m)\left[\mathbf{G}_h\left(\widetilde{\mathbf{U}}_h^{m+1}, \gamma_h^{m+1}\right) - \mathbf{G}_h(\mathbf{U}_h^m, \gamma_h^m) - \mathbf{L}_h^m\left(\widetilde{\mathbf{U}}_h^{m+1} - \mathbf{U}_h^m\right)\right], \tag{3.21b}$$

where $\varphi_1(\mathbf{M}) = \mathbf{M}^{-1}(e^{\mathbf{M}} - \mathbf{I})$ and $\varphi_2(\mathbf{M}) = \mathbf{M}^{-2}(e^{\mathbf{M}} - \mathbf{I} - \mathbf{M})$ for any invertible matrix $\mathbf{M}$.

3.2.2 Slope limiter

Since hyperbolic conservation laws may exhibit discontinuities even when the initial condition is smooth, we use the TVB (Total Variation Bounded) slope limiter as proposed in [171, 33, 31] to capture moving shocks while maintaining the high-order accuracy in smooth regions. To this end, denote by $\widetilde{\mathrm{m}}(a_1, a_2, \dots, a_n)$ the TVB corrected minmod function:

$$\widetilde{\mathrm{m}}(a_1, a_2, \dots, a_n) = \begin{cases} a_1, & \text{if } |a_1| \le C_M h^2, \\ \mathrm{m}(a_1, \dots, a_n), & \text{otherwise,} \end{cases}$$

where C_M is a constant defined in [36] and the minmod function m is given by

$$\mathrm{m}(a_1, \dots, a_n) = s \min_{1 \le i \le n} |a_i|, \text{ with } s = \begin{cases} \mathrm{sign}(a_1), & \text{if } \mathrm{sign}(a_1) = \mathrm{sign}(a_2) = \dots = \mathrm{sign}(a_n), \\ 0, & \text{otherwise.} \end{cases}$$

For $\mathbf{U}_h = (\mathbf{u}_{i,j}) \in \mathbb{R}^{3N_x N_y}$ with $\mathbf{u}_{i,j} = (u_{i,j}^{00}, u_{i,j}^{10}, u_{i,j}^{01})^T$, the TVB corrected slope limiter Λ_h^1 is defined as follows:

$$\mathbf{U}_h^{(\mathrm{mod})} = \Lambda_h^1 \mathbf{U}_h, \tag{3.22}$$

where $\mathbf{U}_h^{(\mathrm{mod})} = \left(\mathbf{u}_{i,j}^{(\mathrm{mod})}\right)_{i,j}$, and $\mathbf{u}_{i,j}^{(\mathrm{mod})} = \left(u_{i,j}^{00}, u_{i,j}^{10(\mathrm{mod})}, u_{i,j}^{01(\mathrm{mod})}\right)^T$ with

$$\begin{aligned} u_{i,j}^{10(\mathrm{mod})} &= \widetilde{\mathrm{m}}\left(u_{i,j}^{10}, u_{i+1,j}^{00} - u_{i,j}^{00}, u_{i,j}^{00} - u_{i-1,j}^{00}\right), \ 2 \le i \le N_x - 1, \ 1 \le j \le N_y, \\ u_{i,j}^{01(\mathrm{mod})} &= \widetilde{\mathrm{m}}\left(u_{i,j}^{01}, u_{i,j+1}^{00} - u_{i,j}^{00}, u_{i,j}^{00} - u_{i,j-1}^{00}\right), \ 1 \le i \le N_x, \ 2 \le j \le N_y - 1. \end{aligned} \tag{3.23}$$

For boundary elements, we use the treatments proposed in [36]. For example, consider elements $K_{1,j}, 1 \le j \le N_y$, near the boundary $x = 0$:

$$u_{1,j}^{10(\mathrm{mod})} = \begin{cases} \widetilde{\mathrm{m}}\left(u_{1,j}^{10}, u_{2,j}^{00} - u_{1,j}^{00}, 2(u_{1,j}^{00} - \gamma(0, y_j, \cdot))\right), & \text{if } x = 0 \text{ is an inflow boundary,} \\ \widetilde{\mathrm{m}}\left(u_{1,j}^{10}, u_{2,j}^{00} - u_{1,j}^{00}\right), & \text{if } x = 0 \text{ is an outflow boundary.} \end{cases}$$

Combining ETDRK2 scheme (3.21) and the limiter (3.22)-(3.23), we obtain the following algorithm:

Algorithm 1 ETDRK2-DG method with TVB slope limiter

1: Compute $\mathbf{U}_h^{0(\text{mod})} = \Lambda_h^1\left(\mathbf{U}_h^0\right)$.

2: **for** $m = 0, 1, \dots, M-1$, **do**

$$\begin{aligned}
\mathbf{L}_h^{m(\text{mod})} &= \mathbf{L}_h\left(\mathbf{U}_h^{m(\text{mod})}\right),\\
\widetilde{\mathbf{U}}_h^{m+1(\text{mod})} &= \Lambda_h^1\left[\mathbf{U}_h^{m(\text{mod})} + \Delta t\varphi_1\left(\Delta t\mathbf{L}_h^{m(\text{mod})}\right)\mathbf{G}_h\left(\mathbf{U}_h^{m(\text{mod})}, \gamma_h^m\right)\right],\\
\mathbf{U}_h^{m+1(\text{mod})} &= \Lambda_h^1\left\{\widetilde{\mathbf{U}}_h^{m+1(\text{mod})} + \Delta t\varphi_2\left(\Delta t\mathbf{L}_h^{m(\text{mod})}\right)\left[\mathbf{G}_h\left(\widetilde{\mathbf{U}}_h^{m+1(\text{mod})}, \gamma_h^{m+1}\right)\right.\right.\\
&\qquad\left.\left. -\mathbf{G}_h\left(\mathbf{U}_h^{m(\text{mod})}, \gamma_h^m\right) - \mathbf{L}_h^{m(\text{mod})}\left(\widetilde{\mathbf{U}}_h^{m+1(\text{mod})} - \mathbf{U}_h^{m(\text{mod})}\right)\right]\right\}.
\end{aligned}$$

3: **end for**

Temporal error estimates for the ETDRK2 scheme as well as its conservation property (under periodic boundary conditions) will be proved in Sections 3.4 and 3.5. Theoretical analysis of the convergence of the approximate solution by Algorithm 1 and its nonlinear stability is still an open question. We shall verify these properties numerically in Section 3.6 where it is shown that the proposed ETDRK-DG method allows much larger time step sizes than the SSP-RK schemes, without affecting the stability of the numerical solution.

3.3 Noniterative localized ETDRK2-DG method

The main cost of the ETDRK2 scheme (3.21) is the computation of actions of matrix exponentials or φ−functions on vectors, especially when the size of $\mathbf{L}_h$ is large. One possible approach to reduce the cost is to use domain decomposition with nonoverlapping subdomains and compute φ−functions of smaller-sized matrices. In the following, we only consider the two-subdomain problem for the sake of presentation; extension to the case of multiple strip subdomains is straightforward.

3.3.1 Space-discrete nonoverlapping domain decomposition formulation

Let $\Omega_1 = (x_{1/2}, x_{N_\Gamma+1/2}) \times (0,1)$ and $\Omega_2 = (x_{N_\Gamma+1/2}, x_{N_x+1/2}) \times (0,1)$ be a partition of Ω with an interface $\Gamma = \{x_{N_\Gamma+1/2}\} \times [0,1]$, where N_Γ is an integer such that $0 < N_\Gamma < N_x$. The meshes in Ω_1 and Ω_2, denoted by $\mathcal{K}_{h,1}$ and $\mathcal{K}_{h,2}$ respectively, are the restriction of the global mesh $\mathcal{K}_h$ to each subdomain:

$$\begin{aligned}
\mathcal{K}_{h,1} &= \left\{K_{i,j} \in \mathcal{K}_h,\ 1 \leq i \leq N_\Gamma,\ 1 \leq j \leq N_y\right\}, \\
\mathcal{K}_{h,2} &= \left\{K_{i,j} \in \mathcal{K}_h,\ N_\Gamma + 1 \leq i \leq N_x,\ 1 \leq j \leq N_y\right\}.
\end{aligned}$$

Let $\mathcal{E}_h^\Gamma$ be the set of edges on the interface:

$$\mathcal{E}_h^\Gamma = \left\{e = \partial K_{N_\Gamma,j} \cap \partial K_{N_\Gamma+1,j},\ 1 \leq j \leq N_y\right\}.$$

To reformulate the space-discrete monodomain problem (3.9) as an equivalent two-subdomain problem, we introduce the interface vector $\boldsymbol{\lambda}_h \in \mathbb{R}^{3N_y}$:

$$\boldsymbol{\lambda}_h = \left(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \dots, \boldsymbol{\lambda}_{N_y}\right)^T, \quad \text{with } \boldsymbol{\lambda}_j = \left(\lambda_j^{00}, \lambda_j^{10}, \lambda_j^{01}\right)^T. \tag{3.24}$$

For $I = 1,2$, denote by $\mathcal{V}_{h,I}$, the restriction of $\mathcal{V}_h$ to Ω_I, and by $w_{h,I} \in \mathcal{V}_{h,I}$, the approximation of the weak solution u to (3.1) in Ω_I. Similarly to u_h (3.12), the vectors of DOFs associated with $w_{h,1}$ and $w_{h,2}$ are:

$$\mathbf{W}_{h,1}(t) = \left(\mathbf{W}_1(t), \mathbf{W}_2(t), \dots, \mathbf{W}_{N_\Gamma}(t)\right)^T, \quad \mathbf{W}_{h,2}(t) = \left(\mathbf{W}_{N_\Gamma+1}(t), \mathbf{W}_{N_\Gamma+2}(t), \dots, \mathbf{W}_{N_x}(t)\right)^T,$$

where

$$\mathbf{W}_i = \left(\mathbf{w}_{i,1}, \mathbf{w}_{i,2}, \dots, \mathbf{w}_{i,N_y}\right)^T,\ \mathbf{w}_{i,j} = \left(w_{i,j}^{00}, w_{i,j}^{10}, w_{i,j}^{01}\right)^T, \quad 1 \leq i \leq N_x,\ 1 \leq j \leq N_y.$$

The unknowns $w_{i,j}^{lk}$ are the solution to the following equations in Ω_1:

$$\frac{d}{dt}w_{i,j}^{lk} = \frac{3^{l+k}}{|K_{i,j}|}\Bigg[\frac{2l}{\Delta x_i}\int_{K_{i,j}} f_1(w_{h,1})dxdy + \frac{2k}{\Delta y_j}\int_{K_{i,j}} f_2(w_{h,1})dxdy$$
$$- \sum_{e\subset\partial K_{i,j}\backslash\mathcal{E}_h^\Gamma}\int_e \mathfrak{h}_{e,K_{i,j}}\left(w_{h,1}^{\mathrm{int}(K_{i,j})}, w_{h,1}^{\mathrm{ext}(K_{i,j})}\right)\phi_{i,j}^{lk}ds - \sum_{e\subset\partial K_{i,j}\cap\mathcal{E}_h^\Gamma}\lambda_j^{lk}\Bigg], \quad 1\leq i\leq N_\Gamma, \tag{3.25}$$

and in Ω_2

$$\frac{d}{dt}w_{i,j}^{lk} = \frac{3^{l+k}}{|K_{i,j}|}\Bigg[\frac{2l}{\Delta x_i}\int_{K_{i,j}} f_1(w_{h,2})dxdy + \frac{2k}{\Delta y_j}\int_{K_{i,j}} f_2(w_{h,2})dxdy$$
$$- \sum_{e\subset\partial K_{i,j}\backslash\mathcal{E}_h^\Gamma}\int_e \mathfrak{h}_{e,K_{i,j}}\left(w_{h,2}^{\mathrm{int}(K_{i,j})}, w_{h,2}^{\mathrm{ext}(K_{i,j})}\right)\phi_{i,j}^{lk}ds + \sum_{e\subset\partial K_{i,j}\cap\mathcal{E}_h^\Gamma}\lambda_j^{lk}\Bigg], \quad N_\Gamma+1\leq i\leq N_x, \tag{3.26}$$

together with the interface condition:

$$\lambda_j^{lk} = \int_e \mathfrak{h}_{e,K_{N_\Gamma,j}}\left(w_{h,1}^{\mathrm{int}(K_{N_\Gamma,j})}, w_{h,2}^{\mathrm{int}(K_{N_\Gamma+1,j})}\right)\phi_{N_\Gamma,j}^{lk}ds, \quad e=\partial K_{N_\Gamma,j}\cap\partial K_{N_\Gamma+1,j}\in\mathcal{E}_h^\Gamma, \tag{3.27}$$

for $1\leq j\leq N_y$, and $(l,k)\in\{(0,0),(1,0),(0,1)\}$. Here the function of the form $w^{\mathrm{int}(K)}(\mathbf{x},t)$ or $w^{\mathrm{ext}(K)}(\mathbf{x},t)$, for $\mathbf{x}\in e\subset\partial K$, is defined as in (3.5).

Lemma 3.1. The two-subdomain problem (3.25)-(3.26) with the interface condition (3.27) is equivalent to the monodomain problem (3.9).

Proof. Suppose u_h is the solution to (3.9). Let $w_{h,1} = u_h|_{\Omega_1}$ and $w_{h,2} = u_h|_{\Omega_2}$, then (3.27) becomes:

$$\begin{aligned}
\lambda_j^{lk} &= \int_e \mathfrak{h}_{e,K_{N_\Gamma,j}}\left(u_h^{\mathrm{int}(K_{N_\Gamma,j})}, u_h^{\mathrm{int}(K_{N_\Gamma+1,j})}\right)\phi_{N_\Gamma,j}^{lk}ds \\
&= \int_e \mathfrak{h}_{e,K_{N_\Gamma,j}}\left(u_h^{\mathrm{int}(K_{N_\Gamma,j})}, u_h^{\mathrm{ext}(K_{N_\Gamma,j})}\right)\phi_{N_\Gamma,j}^{lk}ds \\
&= -\int_e \mathfrak{h}_{e,K_{N_\Gamma+1,j}}\left(u_h^{\mathrm{int}(K_{N_\Gamma+1,j})}, u_h^{\mathrm{ext}(K_{N_\Gamma+1,j})}\right)\phi_{N_\Gamma,j}^{lk}ds,
\end{aligned} \tag{3.28}$$

for $e = \partial K_{N_\Gamma,j} \cap \partial K_{N_\Gamma+1,j} \in \mathcal{E}_h^\Gamma$, $1 \le j \le N_y$, where for the third equality, we have used property (3.6) of the numerical flux on edges shared by two neighboring elements. From (3.28) and (3.9), we deduce that $w_{h,1}$ and $w_{h,2}$ satisfy the two-subdomain problem (3.25)-(3.26).

Next, assume that $w_{h,1}$ and $w_{h,2}$ are the solutions to (3.25)-(3.27). Let $u_h \in \mathcal{V}_h$ such that $u_h|_{\Omega_1} = w_{h,1}$ and $u_h|_{\Omega_2} = w_{h,2}$. By the same argument (i.e. using (3.28)), we conclude that u_h satisfies (3.9). □

Remark 3.2. For the interface unknowns, it should be noted that

$$\lambda_j^{00} = \lambda_j^{10}, \quad 1 \le j \le N_y. \tag{3.29}$$

as $\phi_{i,j}^{00}|_e = \phi_{i,j}^{10}|_e = 1$, for $e = \partial K_{N_\Gamma,j} \cap \partial K_{N_\Gamma+1,j}$, and $i = N_\Gamma$.

The integrals in (3.25)-(3.26) as well as (3.27) (cf. (3.33)) are computed using quadrature rules (3.10)-(3.11) as in the monodomain case. In particular, we have:

$$\begin{aligned}\frac{d}{dt} w_{i,j}^{lk} = \mathrm{g}_{i,j}^{lk}(\mathbf{W}_{h,1}, \boldsymbol{\lambda}_h, \gamma_h) := \frac{3^{l+k}}{|K_{i,j}|}\Bigg[&\frac{2l}{\Delta x_i}\mathcal{I}_{K_{i,j}}(f_1(w_{h,1})) + \frac{2k}{\Delta y_j}\mathcal{I}_{K_{i,j}}(f_2(w_{h,1})) \\ &- \sum_{e\subset\partial K_{i,j}\setminus\mathcal{E}_h^\Gamma} \mathcal{I}_e\left(\mathfrak{h}_{e,K_{i,j}}\phi_{i,j}^{lk}\right) - \sum_{e\subset\partial K_{i,j}\cap\mathcal{E}_h^\Gamma} \lambda_j^{lk}\Bigg], \quad 1 \le i \le N_\Gamma,\end{aligned} \tag{3.30}$$

$$\begin{aligned}\frac{d}{dt} w_{i,j}^{lk} = \mathrm{g}_{i,j}^{lk}(\mathbf{W}_{h,2}, \boldsymbol{\lambda}_h, \gamma_h) := \frac{3^{l+k}}{|K_{i,j}|}\Bigg[&\frac{2l}{\Delta x_i}\mathcal{I}_{K_{i,j}}(f_1(w_{h,2})) + \frac{2k}{\Delta y_j}\mathcal{I}_{K_{i,j}}(f_2(w_{h,2})) \\ &- \sum_{e\subset\partial K_{i,j}\setminus\mathcal{E}_h^\Gamma} \mathcal{I}_e\left(\mathfrak{h}_{e,K_{i,j}}\phi_{i,j}^{lk}\right) + \sum_{e\subset\partial K_{i,j}\cap\mathcal{E}_h^\Gamma} \lambda_j^{lk}\Bigg], \quad N_\Gamma + 1 \le i \le N_x.\end{aligned} \tag{3.31}$$

Let $\mathcal{G}_{h,1} = \left(\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_{N_\Gamma}\right)^T$ and $\mathcal{G}_{h,2} = \left(\mathcal{G}_{N_\Gamma+1}, \mathcal{G}_{N_\Gamma+2}, \dots, \mathcal{G}_{N_x}\right)^T$, with

$$\mathcal{G}_i = \left(\mathbf{g}_{i,1}, \mathbf{g}_{i,2}, \dots, \mathbf{g}_{i,N_y}\right)^T, \ \mathbf{g}_{i,j} = \left(\mathrm{g}_{i,j}^{00}, \mathrm{g}_{i,j}^{10}, \mathrm{g}_{i,j}^{01}\right)^T, \quad 1 \le i \le N_x, \ 1 \le j \le N_y.$$

The two-subdomain formulation can be recast in the ODE form as follows:

$$\frac{d}{dt}\mathbf{W}_{h,1} = \mathcal{G}_{h,1}(\mathbf{W}_{h,1}, \boldsymbol{\lambda}_h, \gamma_h), \quad \frac{d}{dt}\mathbf{W}_{h,2} = \mathcal{G}_{h,2}(\mathbf{W}_{h,2}, \boldsymbol{\lambda}_h, \gamma_h), \tag{3.32}$$

together with the interface condition, for $1 \leq j \leq N_y$, and $(l,k) \in \{(0,0),(1,0),(0,1)\}$:

$$\lambda_j^{lk} = \mathcal{I}_e\left(\mathfrak{h}_{e,K_{N_\Gamma,j}}\left(w_{h,1}^{\mathrm{int}(K_{N_\Gamma,j})}, w_{h,2}^{\mathrm{int}(K_{N_\Gamma+1,j})}\right)\phi_{N_\Gamma,j}^{lk}\right), \quad e = \partial K_{N_\Gamma,j} \cap \partial K_{N_\Gamma+1,j} \in \mathcal{E}_h^\Gamma. \tag{3.33}$$

For convenience of presentation, we shall write $\boldsymbol{\lambda}_h = \mathcal{F}_h^\Gamma(\mathbf{W}_{h,1}, \mathbf{W}_{h,2})$ in place of the interface condition (3.33).

3.3.2 Noniterative localized ETDRK2

We use ETDRK2 for time discretization and decouple the system (3.32)-(3.33) by lagging the interface condition. Let $\mathbf{W}_{h,1}^0$ and $\mathbf{W}_{h,2}^0$ be the restriction of $\mathbf{U}_{h,0}$ (cf. (3.15)) to Ω_1 and Ω_2, respectively. The noniterative localized ETDRK2 method is defined as follows: for $m = 0, 1, \dots, M-1$,

$$\boldsymbol{\lambda}_h^m = \mathcal{F}_h^\Gamma\left(\mathbf{W}_{h,1}^m, \mathbf{W}_{h,2}^m\right), \tag{3.34a}$$

$$\widetilde{\mathbf{W}}_{h,I}^{m+1} = \mathbf{W}_{h,I}^m + \Delta t \varphi_1\left(\Delta t \mathbf{L}_{h,I}^m\right)\mathcal{G}_{h,I}\left(\mathbf{W}_{h,I}^m, \boldsymbol{\lambda}_h^m, \gamma_h^m\right), \quad I = 1,2, \tag{3.34b}$$

$$\widetilde{\boldsymbol{\lambda}}_h^{m+1} = \mathcal{F}_h^\Gamma\left(\widetilde{\mathbf{W}}_{h,1}^{m+1}, \widetilde{\mathbf{W}}_{h,2}^{m+1}\right), \tag{3.34c}$$

$$\begin{aligned}\mathbf{W}_{h,I}^{m+1} = \widetilde{\mathbf{W}}_{h,I}^{m+1} + \Delta t \varphi_2\left(\Delta t \mathbf{L}_{h,I}^m\right)\Big[\mathcal{G}_{h,I}\left(\widetilde{\mathbf{W}}_{h,I}^{m+1}, \widetilde{\boldsymbol{\lambda}}_h^{m+1}, \gamma_h^{m+1}\right) - \mathcal{G}_{h,I}\left(\mathbf{W}_{h,I}^m, \boldsymbol{\lambda}_h^m, \gamma_h^m\right)\\ -\mathbf{L}_{h,I}^m\left(\widetilde{\mathbf{W}}_{h,I}^{m+1} - \mathbf{W}_{h,I}^m\right)\Big], \quad I = 1,2,\end{aligned} \tag{3.34d}$$

where $\mathbf{L}_{h,I}^m$, $I = 1,2$, are some linear operators. A possible choice for $\mathbf{L}_{h,I}^m$ is the Jacobian of $\mathcal{G}_{h,I}$ evaluated at t_m; however, we observe numerically that such a choice leads to some CFL constraint on the time step size, unlike the global ETDRK2 scheme (3.21). We thus construct $\mathbf{L}_{h,1}^m$ and $\mathbf{L}_{h,2}^m$ as the submatrices of the monodomain Jacobian (cf. (3.17)-(3.18)). In particular, we define

$$\mathbf{L}_{h,1}^{m}=\mathbf{L}_{h,1}(\mathbf{W}_{h,1}^{m}):=\begin{pmatrix} \mathbf{A}_{1,1}^{m} & \mathbf{A}_{1,2}^{m} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{A}_{2,1}^{m} & \mathbf{A}_{2,2}^{m} & \mathbf{A}_{2,3}^{m} & \cdots & \mathbf{O} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{O} & \cdots & \mathbf{A}_{N_\Gamma-1,N_\Gamma-2}^{m} & \mathbf{A}_{N_\Gamma-1,N_\Gamma-1}^{m} & \mathbf{A}_{N_\Gamma-1,N_\Gamma}^{m} \\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{A}_{N_\Gamma,N_\Gamma-1}^{m} & \mathbf{A}_{N_\Gamma,N_\Gamma}^{m} \end{pmatrix},$$

and

$$\mathbf{L}_{h,2}^{m}=\mathbf{L}_{h,2}(\mathbf{W}_{h,2}^{m}):=\begin{pmatrix} \mathbf{A}_{N_\Gamma+1,N_\Gamma+1}^{m} & \mathbf{A}_{N_\Gamma+1,N_\Gamma+2}^{m} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{A}_{N_\Gamma+2,N_\Gamma+1}^{m} & \mathbf{A}_{N_\Gamma+2,N_\Gamma+2}^{m} & \mathbf{A}_{N_\Gamma+2,N_\Gamma+3}^{m} & \cdots & \mathbf{O} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{O} & \cdots & \mathbf{A}_{N_x-1,N_x-2}^{m} & \mathbf{A}_{N_x-1,N_x-1}^{m} & \mathbf{A}_{N_x-1,N_x}^{m} \\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{A}_{N_x,N_x-1}^{m} & \mathbf{A}_{N_x,N_x}^{m} \end{pmatrix},$$

where $\mathbf{A}_{i,i'}^{m}:=\dfrac{\partial \mathbf{G}_i}{\partial \mathbf{U}_{i'}}(\mathbf{W}_{i'}^{m})$ as in (3.19). We remark that the computation of the local matrix $\mathbf{L}_{h,I}^{m}$ only involves the local solution $\mathbf{W}_{h,I}^{m}$, for $I=1,2$.

3.3.3 Slope limiter for localized ETDRK2

As for the monodomain case, to handle moving shocks, we incorporate the TVB corrected slope limiter Λ_h^1 (3.22)-(3.23) into the decoupled scheme (3.34). The limiter is applied globally instead of locally on each subdomain to take into account the communication between the local solutions across the interface.

Algorithm 2 Localized ETDRK2 method with TVB slope limiter

1: Compute $\left(\mathbf{W}_{h,1}^{0(\text{mod})},\mathbf{W}_{h,2}^{0(\text{mod})}\right)=\Lambda_h^1\left(\mathbf{W}_{h,1}^0,\mathbf{W}_{h,2}^0\right)$.

2: **for** $m=0,1,\dots,M-1,$ **do**

$$
\begin{aligned}
&\mathbf{L}_{h,I}^{m(\text{mod})}=\mathbf{L}_{h,I}\left(\mathbf{W}_{h,I}^{m(\text{mod})}\right),\ I=1,2,\quad \lambda_h^{m(\text{mod})}=\mathcal{F}_h^\Gamma\left(\mathbf{W}_{h,1}^{m(\text{mod})},\mathbf{W}_{h,2}^{m(\text{mod})}\right),\\
&\widetilde{\mathbf{W}}_{h,I}^{m+1}=\mathbf{W}_{h,I}^{m(\text{mod})}+\Delta t\varphi_1\left(\Delta t\mathbf{L}_{h,I}^{m(\text{mod})}\right)\mathcal{G}_{h,I}\left(\mathbf{W}_{h,I}^{m(\text{mod})},\lambda_h^{m(\text{mod})},\gamma_h^m\right),\quad I=1,2,\\
&\left(\widetilde{\mathbf{W}}_{h,1}^{m+1(\text{mod})},\widetilde{\mathbf{W}}_{h,2}^{m+1(\text{mod})}\right)=\Lambda_h^1\left(\widetilde{\mathbf{W}}_{h,1}^{m+1},\widetilde{\mathbf{W}}_{h,2}^{m+1}\right),\\
&\widetilde{\lambda}_h^{m+1(\text{mod})}=\mathcal{F}_h^\Gamma\left(\widetilde{\mathbf{W}}_{h,1}^{m+1(\text{mod})},\widetilde{\mathbf{W}}_{h,2}^{m+1(\text{mod})}\right),\\
&\mathbf{W}_{h,I}^{m+1}=\widetilde{\mathbf{W}}_{h,I}^{m+1(\text{mod})}+\Delta t\varphi_2\left(\Delta t\mathbf{L}_{h,I}^{m(\text{mod})}\right)\Big[\mathcal{G}_{h,I}\left(\widetilde{\mathbf{W}}_{h,I}^{m+1(\text{mod})},\widetilde{\lambda}_h^{m+1(\text{mod})},\gamma_h^{m+1}\right)\\
&\qquad-\mathcal{G}_{h,I}\left(\mathbf{W}_{h,I}^{m(\text{mod})},\lambda_h^{m(\text{mod})},\gamma_h^m\right)-\mathbf{L}_{h,I}^{m(\text{mod})}\left(\widetilde{\mathbf{W}}_{h,I}^{m+1(\text{mod})}-\mathbf{W}_{h,I}^{m(\text{mod})}\right)\Big],\quad I=1,2,\\
&\left(\mathbf{W}_{h,1}^{m+1(\text{mod})},\mathbf{W}_{h,2}^{m+1(\text{mod})}\right)=\Lambda_h^1\left(\mathbf{W}_{h,1}^{m+1},\mathbf{W}_{h,2}^{m+1}\right).
\end{aligned}
$$

3: **end for**

3.4 Temporal error analysis

We carry out error analysis in time for the ETDRK2 scheme (3.21) and the localized ETDRK2 scheme (3.34), when the solution to (3.1) is sufficiently smooth (and no slope limiter is needed in that case). Throughout this section, we assume that $\mathbf{F}=(f_1,f_2)\in\left(\mathcal{C}^2(\mathbb{R})\right)^2$ and the numerical flux is chosen to be either the Lax-Friedrichs flux (3.7) or the Engquist-Osher flux (3.8). For simplicity of presentation, we consider constant Dirichlet boundary conditions; the same results hold if $\gamma(\mathbf{x},t)$ is bounded or if periodic boundary conditions are imposed (cf. Section 3.5). In addition, we use uniform spatial mesh sizes, namely Δx and Δy in the x- and y-directions respectively. We begin by presenting some preliminary lemmas.

3.4.1 Preliminaries

We follow [202] to modify the flux functions f_1 and f_2 so that their first and second derivatives are bounded. In particular, the model equation (3.1) satisfies the maximum principle in the sense that if the initial and boundary data, u_0 and γ, lie in $[m_0, M_0]$, then the solution to (3.1) is also in this range. Thus we modify f_1 and f_2 outside the interval $[m_0 - 1, M_0 + 1]$ so that the new functions have zero first and second derivatives. For convenience, we still denote these modified functions by f_1 and f_2, and define the constants L_f and L_{ff} as follows:

$$L_f := \max_{u\in\mathbb{R}} \left\{|f_1'(u)|, |f_2'(u)|\right\}, \text{ and } L_{ff} := \max_{u\in\mathbb{R}} \left\{|f_1''(u)|, |f_2''(u)|\right\}. \tag{3.35}$$

Lemma 3.3. For $K \in \mathcal{K}_h$ and $e \subset \partial K$, the numerical flux $\mathfrak{h}_{e,K}$ satisfies

$$|\mathfrak{h}_{e,K}(u,v) - \mathfrak{h}_{e,K}(\hat{u},\hat{v})| \le L_f(|u-\hat{u}| + |v-\hat{v}|), \quad \forall u, v, \hat{u}, \hat{v} \in \mathbb{R}. \tag{3.36}$$

In addition, the partial derivatives $\partial_u \mathfrak{h}_{e,K}$ and $\partial_v \mathfrak{h}_{e,K}$ are Lipschitz continuous.

Proof. We only consider the edge-cell pair (e, K) with $e \subset \partial K$ such that $\mathbf{n}_{e,K} = (1,0)^T$ (similar results can be obtained for other cases, i.e., $\mathbf{n}_{e,K} \in \{(-1,0)^T, (0,1)^T, (0,-1)^T\}$). The numerical flux $\mathfrak{h}_{e,K}$ reduces to

$$\mathfrak{h}^{\mathsf{LF}}_{e,K}(u,v) = \frac{1}{2}[f_1(u) + f_1(v)] - \frac{\alpha}{2}(v-u), \quad \alpha = \max_{[\inf u_0, \sup u_0]} |f_1'|,$$

for the Lax-Friedrichs flux, or

$$\mathfrak{h}^{\mathsf{EO}}_{e,K}(u,v) = \frac{1}{2}[f_1(u) + f_1(v)] - \frac{1}{2}\int_u^v |f_1'(w)| dw,$$

for the Engquist-Osher flux. As a consequence, the partial derivatives are given by

$$\partial_u \mathfrak{h}^{\mathsf{LF}}_{e,K}(u,v) = \frac{f_1'(u)}{2} + \frac{\alpha}{2}, \text{ and } \partial_v \mathfrak{h}^{\mathsf{LF}}_{e,K}(u,v) = \frac{f_1'(v)}{2} - \frac{\alpha}{2}, \tag{3.37}$$

for the Lax-Friedrichs flux, and

$$\partial_u \mathfrak{h}^{\mathsf{EO}}_{e,K}(u,v) = \frac{f_1'(u)}{2} + \frac{|f_1'(u)|}{2} = \max\{f_1'(u), 0\}, \tag{3.38a}$$

$$\partial_v \mathfrak{h}^{\mathsf{EO}}_{e,K}(u,v) = \frac{f_1'(v)}{2} - \frac{|f_1'(v)|}{2} = \min\{f_1'(v), 0\}, \tag{3.38b}$$

for the Engquist-Osher flux. In either case, $|\partial_u \mathfrak{h}_{e,K}(u,v)| \le L_f$ and $|\partial_v \mathfrak{h}_{e,K}(u,v)| \le L_f$ for all

$u, v \in \mathbb{R}$, as $\alpha \leq L_f$. Therefore, for all $u, v, \hat{u}, \hat{v} \in \mathbb{R}$, we deduce that

$$\begin{aligned}|\mathfrak{h}_{e,K}(u,v) - \mathfrak{h}_{e,K}(\hat{u},\hat{v})| &\leq |\mathfrak{h}_{e,K}(u,v) - \mathfrak{h}_{e,K}(u,\hat{v})| + |\mathfrak{h}_{e,K}(u,\hat{v}) - \mathfrak{h}_{e,K}(\hat{u},\hat{v})| \\ &\leq L_f|v-\hat{v}| + L_f|u-\hat{u}|.\end{aligned}$$

Next, the Lipschitz continuity of $\partial_u \mathfrak{h}_{e,K}$ and $\partial_v \mathfrak{h}_{e,K}$ are trivial for the Lax-Friedrichs flux (3.37). For the Engquist-Osher flux, it is implied from (3.38) that

$$\begin{aligned}|\partial_u \mathfrak{h}^{\mathsf{EO}}_{e,K}(u,v) - \partial_u \mathfrak{h}^{\mathsf{EO}}_{e,K}(\hat{u},v)| &\leq \frac{|f_1'(u) - f_1'(\hat{u})|}{2} + \frac{\big||f_1'(u)| - |f_1'(\hat{u})|\big|}{2} \\ &\leq |f_1'(u) - f_1'(\hat{u})| \leq L_{ff}|u-\hat{u}|,\end{aligned}$$

where we have used the inequality $\big||a| - |b|\big| \leq |a-b|$ for all $a, b \in \mathbb{R}$ in the second estimate. Moreover,

$$|\partial_u \mathfrak{h}^{\mathsf{EO}}_{e,K}(u,v) - \partial_u \mathfrak{h}^{\mathsf{EO}}_{e,K}(u,\hat{v})| = 0.$$

Thus $\partial_u \mathfrak{h}^{\mathsf{EO}}_{e,K}$ is Lipschitz continuous, and the same conclusion holds for $\partial_v \mathfrak{h}^{\mathsf{EO}}_{e,K}$. $\square$

Lemma 3.4. For any functions $a, b \in \mathcal{C}^2([0,T])$, there exist positive constants θ_1 and θ_2 such that

$$\left|\frac{\Delta t - s}{\Delta t} f_1(a(t_m)) + \frac{s}{\Delta t} f_1(a(t_{m+1})) - f_1(a(t_m + s))\right| \leq \theta_1 \Delta t^2,$$

and

$$\left|\frac{\Delta t - s}{\Delta t} \mathfrak{h}_{e,K}(a(t_m), b(t_m)) + \frac{s}{\Delta t} \mathfrak{h}_{e,K}(a(t_{m+1}), b(t_{m+1})) - \mathfrak{h}_{e,K}(a(t_m + s), b(t_m + s))\right| \leq \theta_2 \Delta t^2,$$

for all $s \in [0, \Delta t]$ and $m \in \{0, 1, \dots, M-1\}$, where θ_1 and θ_2 depend on L_f, L_{ff} and $\mathcal{C}^2([0,T])$-norms of a, b, and are independent of Δt.

Proof. Since $f_1', \partial_u \mathfrak{h}_{e,K}$ and $\partial_v \mathfrak{h}_{e,K}$ are all Lipschitz continuous, it suffices to show that for any function $p \in \mathcal{C}^1(\mathbb{R})$ with p' being Lipschitz continuous, i.e., $|p'(t) - p'(s)| \leq L_p|t-s|, \forall t, s \in \mathbb{R}$, it holds that

$$\left|\frac{\Delta t - s}{\Delta t} p(t_m) + \frac{s}{\Delta t} p(t_{m+1}) - p(t_m + s)\right| \leq c(L_p)\Delta t^2, \tag{3.39}$$

for some constant $c(L_p) > 0$ depending on L_p. Indeed, by the mean value theorem, there exist $\alpha_m, \beta_m \in [t_m, t_{m+1}]$ satisfying

$$\begin{aligned}
&\left|\frac{\Delta t - s}{\Delta t} p(t_m) + \frac{s}{\Delta t} p(t_{m+1}) - p(t_m + s)\right| \\
&= \left|\frac{\Delta t - s}{\Delta t}(p(t_m) - p(t_m + s)) + \frac{s}{\Delta t}(p(t_{m+1}) - p(t_m + s))\right| \\
&= \left|\frac{\Delta t - s}{\Delta t}(-s)p'(\alpha_m) + \frac{s}{\Delta t}(\Delta t - s)p'(\beta_m)\right| \\
&= \frac{s(\Delta t - s)}{\Delta t}|p'(\beta_m) - p'(\alpha_m)| \le \frac{\Delta t}{4} L_p |\beta_m - \alpha_m| \le \frac{L_p}{4}\Delta t^2,
\end{aligned}$$

where we have used the inequality $s(\Delta t - s) \le \frac{1}{4}\Delta t^2$ for all $s \in [0, \Delta t]$. The proof is completed by applying (3.39) with $p = f_1(a(\cdot))$ or $p = \mathfrak{h}_{e,K}(a(\cdot), b(\cdot))$. $\square$

3.4.2 Temporal error estimates

We present error analysis for the global ETDRK2 scheme (3.21) in Theorems 3.5 and 3.6. Similar results for the localized ETDRK2 scheme (3.34) are proved in Theorem 3.7. Since γ is assumed to be constant, we shall omit the argument γ_h in $\mathbf{G}_h$ or $\mathcal{G}_{h,I}, I = 1, 2$.

Theorem 3.5. Let Δx and Δy be fixed and assume $\mathbf{U}_h \in \mathcal{C}^1([0, T], \mathbb{R}^{3N_xN_y})$ is the exact solution of the space-discrete problem (3.15). The numerical solution $\widetilde{\mathbf{U}}_h^{m+1}$ given by (3.21a) satisfies the following estimate:

$$\|\widetilde{\mathbf{U}}_h^{m+1} - \mathbf{U}_h(t_{m+1})\|_\infty \le (e^{c_0\Delta t} + c_1\Delta t)\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty + c_2\Delta t^2, \tag{3.40}$$

for $m = 0, 1, \dots, M-1$, where c_1 and c_2 are positive constants depending on $\Delta x, \Delta y, L_u, L_f$ and T, and independent of Δt.

Proof. Equation (3.21a) can be rewritten equivalently as

$$\widetilde{\mathbf{U}}_h^{m+1} = e^{\Delta t \mathbf{L}_h^m}\mathbf{U}_h^m + \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{L}_h^m}\mathbf{R}_h^m(\mathbf{U}_h^m)ds. \tag{3.41}$$

Subtracting (3.20) from (3.41), we have

$$\widetilde{\mathbf{U}}_h^{m+1} - \mathbf{U}_h(t_{m+1}) = e^{\Delta t \mathbf{L}_h^m}(\mathbf{U}_h^m - \mathbf{U}_h(t_m)) + \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{L}_h^m}\left[\mathbf{R}_h^m(\mathbf{U}_h^m) - \mathbf{R}_h^m(\mathbf{U}_h(t_m))\right] ds$$
$$+ \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{L}_h^m}\left[\mathbf{R}_h^m(\mathbf{U}_h(t_m)) - \mathbf{R}_h^m(\mathbf{U}_h(t_m + s))\right] ds.$$

Taking the supremum norm on both sides of the above equation yields

$$\begin{aligned}\|\widetilde{\mathbf{U}}_h^{m+1} - \mathbf{U}_h(t_{m+1})\|_\infty \le \|e^{\Delta t \mathbf{L}_h^m}\|_\infty \|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty \\ + \int_0^{\Delta t} \|e^{(\Delta t - s)\mathbf{L}_h^m}\|_\infty \|\mathbf{R}_h^m(\mathbf{U}_h^m) - \mathbf{R}_h^m(\mathbf{U}_h(t_m))\|_\infty ds \\ + \int_0^{\Delta t} \|e^{(\Delta t - s)\mathbf{L}_h^m}\|_\infty \|\mathbf{R}_h^m(\mathbf{U}_h(t_m)) - \mathbf{R}_h^m(\mathbf{U}_h(t_m + s))\|_\infty ds.\end{aligned} \tag{3.42}$$

To bound the terms on the right-hand side, we first deduce from (3.35) and Lemma 3.3 that

$$\|\mathbf{L}_h^m\|_\infty \le c_0, \tag{3.43}$$

for some constant $c_0 = c_0(\Delta x, \Delta y, L_f)$ independent of Δt. Then for any $t > 0$, it follows that

$$\|e^{t\mathbf{L}_h^m}\|_\infty \le e^{t\|\mathbf{L}_h^m\|_\infty} \le e^{c_0 t}. \tag{3.44}$$

For $\mathbf{V}, \widehat{\mathbf{V}} \in \mathbb{R}^{3N_x N_y}$, we have from (3.35) and (3.36) that

$$\|\mathbf{G}_h(\mathbf{V}) - \mathbf{G}_h(\widehat{\mathbf{V}})\|_\infty \le \overline{c_0}\|\mathbf{V} - \widehat{\mathbf{V}}\|_\infty,$$

for some constant $\overline{c_0} = \overline{c_0}(\Delta x, \Delta y, L_f)$ independent of Δt. This, together with (3.43), gives us

$$\begin{aligned}\|\mathbf{R}_h^m(\mathbf{V}) - \mathbf{R}_h^m(\widehat{\mathbf{V}})\|_\infty &\le \|\mathbf{G}_h(\mathbf{V}) - \mathbf{G}_h(\widehat{\mathbf{V}})\|_\infty + \|\mathbf{L}_h^m(\mathbf{V} - \widehat{\mathbf{V}})\|_\infty \\ &\le (c_0 + \overline{c_0})\|\mathbf{V} - \widehat{\mathbf{V}}\|_\infty.\end{aligned} \tag{3.45}$$

Using (3.44) and (3.45), we obtain the following estimate from (3.42):

$$\|\widetilde{\mathbf{U}}_h^{m+1} - \mathbf{U}_h(t_{m+1})\|_\infty \leq e^{c_0\Delta t}\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty + (c_0 + \overline{c_0})\int_0^{\Delta t} e^{c_0(\Delta t - s)} ds \|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty$$
$$+ (c_0 + \overline{c_0})\int_0^{\Delta t} e^{c_0(\Delta t - s)}\|\mathbf{U}_h(t_m) - \mathbf{U}_h(t_m + s)\|_\infty ds.$$

Since $\mathbf{U}_h \in \mathcal{C}^1([0,T], \mathbb{R}^{3N_xN_y})$, there is $L_u > 0$ such that

$$\|\mathbf{U}_h(s) - \mathbf{U}_h(t)\|_\infty \leq L_u|s - t|, \quad \forall s, t \in [0,T]. \tag{3.46}$$

Hence $\|\mathbf{U}_h(t_m) - \mathbf{U}_h(t_m + s)\|_\infty \leq L_u s$. This and the fact that $e^{c_0(\Delta t - s)} \leq e^{c_0 T}$ imply

$$\begin{aligned}\|\widetilde{\mathbf{U}}_h^{m+1} - \mathbf{U}_h(t_{m+1})\|_\infty &\leq (e^{c_0\Delta t} + (c_0 + \overline{c_0})e^{c_0T}\Delta t)\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty + \frac{1}{2}(c_0 + \overline{c_0})L_u e^{c_0T}\Delta t^2 \\ &=: (e^{c_0\Delta t} + c_1\Delta t)\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty + c_2\Delta t^2,\end{aligned}$$

where $c_1 = (c_0 + \overline{c_0})e^{c_0T}$ and $c_2 = \frac{1}{2}(c_0 + \overline{c_0})L_u e^{c_0T}$. $\square$

Theorem 3.6. Let Δx and Δy be fixed and assume $\mathbf{U}_h \in \mathcal{C}^2([0,T], \mathbb{R}^{3N_xN_y})$ is the exact solution of the space-discrete problem (3.15). Let $\left(\mathbf{U}_h^m\right)_{m=1,\ldots,M}$ be the numerical solution given by the global ETDRK2 scheme (3.21), then it holds that

$$\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty \leq C\Delta t^2,$$

for $m = 0, 1, \ldots, M$, where the constant $C > 0$ depends on $\Delta x, \Delta y, L_u, L_f, L_{ff}, T$ and $\mathcal{C}^2([0,T])$-norm of $\mathbf{U}_h$, and is independent of Δt.

Proof. We rewrite equation (3.21b) as

$$\mathbf{U}_h^{m+1} = e^{\Delta t\mathbf{L}_h^m}\mathbf{U}_h^m + \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{L}_h^m}\left[\frac{\Delta t - s}{\Delta t}\mathbf{R}_h^m(\mathbf{U}_h^m) + \frac{s}{\Delta t}\mathbf{R}_h^m(\widetilde{\mathbf{U}}_h^{m+1})\right] ds. \tag{3.47}$$

Subtracting (3.20) from (3.47), we have

$$
\begin{aligned}
&\mathbf{U}_h^{m+1} - \mathbf{U}_h(t_{m+1}) = e^{\Delta t \mathbf{L}_h^m}(\mathbf{U}_h^m - \mathbf{U}_h(t_m)) \\
&+ \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{L}_h^m}\left[\frac{\Delta t - s}{\Delta t}\left(\mathbf{R}_h^m(\mathbf{U}_h^m) - \mathbf{R}_h^m(\mathbf{U}_h(t_m))\right) + \frac{s}{\Delta t}\left(\mathbf{R}_h^m(\widetilde{\mathbf{U}}_h^{m+1}) - \mathbf{R}_h^m(\mathbf{U}_h(t_{m+1}))\right)\right] ds \\
&+ \int_0^{\Delta t} e^{(\Delta t - s)\mathbf{L}_h^m}\left[\frac{\Delta t - s}{\Delta t}\mathbf{R}_h^m(\mathbf{U}_h(t_m)) + \frac{s}{\Delta t}\mathbf{R}_h^m(\mathbf{U}_h(t_{m+1})) - \mathbf{R}_h^m(\mathbf{U}_h(t_m + s))\right] ds.
\end{aligned}
$$

Taking the supremum norm on both sides of the above equation and using (3.44) and (3.45) yields

$$
\begin{aligned}
&\|\mathbf{U}_h^{m+1} - \mathbf{U}_h(t_{m+1})\|_\infty \le e^{c_0 \Delta t}\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty \\
&+ \int_0^{\Delta t} e^{c_0(\Delta t - s)}\left[\frac{\Delta t - s}{\Delta t}(c_0 + \overline{c_0})\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty + \frac{s}{\Delta t}(c_0 + \overline{c_0})\|\widetilde{\mathbf{U}}_h^{m+1} - \mathbf{U}_h(t_{m+1})\|_\infty\right] ds \\
&+ \int_0^{\Delta t} e^{c_0(\Delta t - s)}\left\|\frac{\Delta t - s}{\Delta t}\mathbf{R}_h^m(\mathbf{U}_h(t_m)) + \frac{s}{\Delta t}\mathbf{R}_h^m(\mathbf{U}_h(t_{m+1})) - \mathbf{R}_h^m(\mathbf{U}_h(t_m + s))\right\|_\infty ds.
\end{aligned}
$$

By using $e^{c_0(\Delta t - s)} \le e^{c_0 T}$, we deduce that

$$
\begin{aligned}
&\|\mathbf{U}_h^{m+1} - \mathbf{U}_h(t_{m+1})\|_\infty \le \left(e^{c_0 \Delta t} + \frac{1}{2}(c_0 + \overline{c_0})e^{c_0 T}\Delta t\right)\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty \\
&\quad + \frac{1}{2}(c_0 + \overline{c_0})e^{c_0 T}\Delta t\|\widetilde{\mathbf{U}}_h^{m+1} - \mathbf{U}_h(t_{m+1})\|_\infty \\
&\quad + e^{c_0 T}\int_0^{\Delta t}\left\|\frac{\Delta t - s}{\Delta t}\mathbf{R}_h^m(\mathbf{U}_h(t_m)) + \frac{s}{\Delta t}\mathbf{R}_h^m(\mathbf{U}_h(t_{m+1})) - \mathbf{R}_h^m(\mathbf{U}_h(t_m + s))\right\|_\infty ds. \quad (3.48)
\end{aligned}
$$

It follows from Lemma 3.4 that

$$
\left\|\frac{\Delta t - s}{\Delta t}\mathbf{R}_h^m(\mathbf{U}_h(t_m)) + \frac{s}{\Delta t}\mathbf{R}_h^m(\mathbf{U}_h(t_{m+1})) - \mathbf{R}_h^m(\mathbf{U}_h(t_m + s))\right\|_\infty \le c_3 \Delta t^2, \tag{3.49}
$$

where $c_3 > 0$ depends on L_f, L_{ff} and $\mathcal{C}^2$-norm of $\mathbf{U}_h$. Combining (3.40), (3.48) and (3.49), we obtain

$$\|\mathbf{U}_h^{m+1} - \mathbf{U}_h(t_{m+1})\|_\infty \leq \left(e^{c_0\Delta t} + \frac{1}{2}(c_0 + \overline{c_0})e^{c_0 T}\Delta t(1 + e^{c_0\Delta t} + c_1\Delta t)\right)\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty + e^{c_0 T}\left(\frac{1}{2}(c_0 + \overline{c_0})c_2 + c_3\right)\Delta t^3$$
$$\leq (e^{c_0\Delta t} + c_4\Delta t)\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty + c_5\Delta t^3, \qquad (3.50)$$

where $c_4 = \frac{1}{2}(c_0 + \overline{c_0})e^{c_0 T}(1 + e^{c_0 T} + c_1 T)$ and $c_5 = e^{c_0 T}\left(\frac{1}{2}(c_0 + \overline{c_0})c_2 + c_3\right)$. By repeatedly applying (3.50), noting that $\mathbf{U}_h^0 = \mathbf{U}_h(t_0)$, we arrive at

$$\|\mathbf{U}_h^m - \mathbf{U}_h(t_m)\|_\infty \leq \frac{c_5\Delta t^3}{e^{c_0\Delta t} + c_4\Delta t - 1}\left[(e^{c_0\Delta t} + c_4\Delta t)^m - 1\right]$$
$$\leq \frac{c_5}{c_4}\Delta t^2[e^{c_0 t_m}(1 + c_4\Delta t)^m - 1] \leq \frac{c_5}{c_4}[e^{(c_0+c_4)t_m} - 1]\Delta t^2.$$

The proof is complete by setting $C = \frac{c_5}{c_4}[e^{(c_0+c_4)T} - 1]$. □

Theorem 3.7. Let Δx and Δy be fixed and assume $\mathbf{U}_h \in \mathcal{C}^2([0,T], \mathbb{R}^{3N_xN_y})$ is the exact solution of the space-discrete problem (3.15). Let $\mathbf{W}_h^m = (\mathbf{W}_{h,1}^m, \mathbf{W}_{h,2}^m)^T$ be the numerical solution given by the localized ETDRK2 scheme (3.34), then it holds that

$$\|\mathbf{W}_h^m - \mathbf{U}_h(t_m)\|_\infty \leq C\Delta t^2,$$

for $m = 0, 1, \dots, M$, where the constant $C > 0$ depends on $\Delta x, \Delta y, L_u, L_f, L_{ff}, T$ and $\mathcal{C}^2([0,T])$-norm of $\mathbf{U}_h$, and is independent of Δt.

Proof. For $m = 0, 1, \dots, M-1$, denote by $\widetilde{\mathbf{W}}_h^{m+1} = (\widetilde{\mathbf{W}}_{h,1}^{m+1}, \widetilde{\mathbf{W}}_{h,2}^{m+1})^T$, and let

$$\boldsymbol{\mathcal{L}}_h^m := \begin{pmatrix} \mathbf{L}_{h,1}^m & \mathbf{O} \\ \mathbf{O} & \mathbf{L}_{h,2}^m \end{pmatrix}. \quad \text{Then} \quad \varphi_l(\Delta t\boldsymbol{\mathcal{L}}_h^m) = \begin{pmatrix} \varphi_l(\Delta t\mathbf{L}_{h,1}^m) & \mathbf{O} \\ \mathbf{O} & \varphi_l(\Delta t\mathbf{L}_{h,2}^m) \end{pmatrix}, \quad l = 1, 2. \qquad (3.51)$$

In addition, given the relation (3.33) between $\boldsymbol{\lambda}_h$ and the subdomain DOF vectors $\mathbf{W}_{h,1}^m$ and $\mathbf{W}_{h,2}^m$, we deduce from the right-hand sides of (3.30)-(3.31) and (3.13) that

$$\begin{pmatrix} \mathcal{G}_{h,1}\left(\mathbf{W}_{h,1}^m, \lambda^m\right) \\ \mathcal{G}_{h,2}\left(\mathbf{W}_{h,2}^m, \lambda^m\right) \end{pmatrix} = \begin{pmatrix} \mathcal{G}_{h,1}\left(\mathbf{W}_{h}^m\right) \\ \mathcal{G}_{h,2}\left(\mathbf{W}_{h}^m\right) \end{pmatrix} = \mathbf{G}_h\left(\mathbf{W}_h^m\right). \tag{3.52}$$

With (3.51) and (3.52), we can rewrite the localized ETDRK2 scheme (3.34) equivalently as follows:

$$\begin{aligned} \widetilde{\mathbf{W}}_h^{m+1} &= \mathbf{W}_h^m + \Delta t \varphi_1(\Delta t \boldsymbol{\mathcal{L}}_h^m)\mathbf{G}_h(\mathbf{W}_h^m), \\ \mathbf{W}_h^{m+1} &= \widetilde{\mathbf{W}}_h^{m+1} + \Delta t \varphi_2(\Delta t \boldsymbol{\mathcal{L}}_h^m)\left[\mathbf{G}_h(\widetilde{\mathbf{W}}_h^{m+1}) - \mathbf{G}_h(\mathbf{W}_h^m) - \boldsymbol{\mathcal{L}}_h^m(\widetilde{\mathbf{W}}_h^{m+1} - \mathbf{W}_h^m)\right]. \end{aligned} \tag{3.53}$$

Hence, the proof for Theorem 3.7 can be done in a similar manner as Theorems 3.5 and 3.6, noting that

$$\|\boldsymbol{\mathcal{L}}_h^m\|_\infty \le c_0^*, \text{ and } \|\boldsymbol{\mathcal{R}}_h^m(\mathbf{V}) - \boldsymbol{\mathcal{R}}_h^m(\widehat{\mathbf{V}})\|_\infty \le (c_0^* + \overline{c_0})\|\mathbf{V} - \widehat{\mathbf{V}}\|_\infty,$$

for all $\mathbf{V}, \widehat{\mathbf{V}} \in \mathbb{R}^{3N_xN_y}$, where c_0^* is some positive constant independent of Δt, and $\boldsymbol{\mathcal{R}}_h^m(\mathbf{V}) = \mathbf{G}_h(\mathbf{V}) - \boldsymbol{\mathcal{L}}_h^m\mathbf{V}$. □

Remark 3.8. The estimates in Theorems 3.6 and 3.7 still hold in the L^1-norm (instead of the L^∞-norm). In particular, for the former we have

$$\begin{aligned} u_h(t_m)|_{K_{i,j}} &= u_{i,j}^{00}(t_m) + u_{i,j}^{10}(t_m)\phi_{i,j}^{10} + u_{i,j}^{01}(t_m)\phi_{i,j}^{01}, \\ u_h^m|_{K_{i,j}} &= u_{i,j}^{00,m} + u_{i,j}^{10,m}\phi_{i,j}^{10} + u_{i,j}^{01,m}\phi_{i,j}^{01}, \end{aligned}$$

for all elements $K_{i,j} \in \mathcal{K}_h$, where $\mathbf{U}_h(t_m) = (u_{i,j}^{lk}(t_m))_{i,j,lk}$ is the solution of the space-discrete problem (3.15) and $\mathbf{U}_h^m = (u_{i,j}^{lk,m})_{i,j,lk}$ is the numerical solution by the ETDRK2 scheme (3.21).

Let $\mathbf{U}_h^m - \mathbf{U}_h(t_m) = (r_{i,j}^{lk,m})_{i,j,lk}$, then

$$\begin{aligned}
\|u_h^m - u_h(t_m)\|_{L^1(\Omega)} &= \sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\int_{K_{i,j}} \left| r_{i,j}^{00,m} + r_{i,j}^{10,m}\phi_{i,j}^{10} + r_{i,j}^{01,m}\phi_{i,j}^{01}\right| dxdy \\
&\le \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} |K_{i,j}| \left(|r_{i,j}^{00,m}| + |r_{i,j}^{10,m}| + |r_{i,j}^{01,m}|\right) \\
&\le \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} |K_{i,j}| 3C\Delta t^2 = 3|\Omega|C\Delta t^2,
\end{aligned}$$

where we have used Theorem 3.6 and the facts that $|\phi_{i,j}^{10}| \le 1$ and $|\phi_{i,j}^{01}| \le 1$ on $K_{i,j} \in \mathcal{K}_h$.

3.5 Periodic boundary conditions and discrete conservation

In this section, we extend the global and localized ETDRK2 methods presented in previous sections to the case of periodic boundary conditions. For spatial discretization, such boundary conditions are incorporated in the numerical flux $\mathfrak{h}_{e,K}\left(u_h^{\text{int}(K)}(\mathbf{x},t), u_h^{\text{ext}(K)}(\mathbf{x},t)\right)$, where

$$u_h^{\text{int}(K)}(\mathbf{x},t) = \lim_{\substack{\mathbf{y}\to\mathbf{x}\\ \mathbf{y}\in\mathring{K}}} u_h(\mathbf{y},t), \text{ and } \quad u_h^{\text{ext}(K)}(\mathbf{x},t) = \lim_{\substack{\mathbf{y}\to\mathbf{x}\\ \mathbf{y}\in(K)^c}} u_h(\mathbf{y},t). \tag{3.54}$$

Note that to compute $u_h^{\text{ext}(K_{i,j})}(\mathbf{x},t)$ when $\mathbf{x} \in \partial\Omega$, the standard periodic extension for u_h is applied to define $u_h(\mathbf{y},\cdot)$ with $\mathbf{y} \notin \Omega$. As a consequence, the ODE system (3.15) becomes autonomous:

$$\frac{d}{dt}\mathbf{U}_h = \mathbf{G}_h^{\text{per}}(\mathbf{U}_h), \quad \mathbf{U}_h(0) = \mathbf{U}_{h,0}, \tag{3.55}$$

where $\mathbf{G}_h^{\text{per}}$ is defined similarly to (3.14) with the numerical flux computed based on (3.54) (i.e., no Dirichlet data). In the following, we present the formulations of the global ETDRK2 and localized ETDRK2 schemes corresponding to (3.55), and prove the discrete conservation property for each scheme. Note that compared to Dirichlet boundary conditions, the main difference lies in the formulation of the linear operator $\mathbf{L}_h$ or $\mathbf{L}_{h,i}, I = 1, 2$. The application of the TVB-corrected slope limiter is done in the same manner as (3.22)-(3.23), and the results for temporal error estimates in Section 3.4 can be extended straightforwardly to the case of periodic boundary conditions. For simplicity of presentation, we again consider uniform mesh sizes Δx and Δy.

3.5.1 Global ETDRK2 method

For $m = 0, 1, \dots, M-1$,

$$\widetilde{\mathbf{U}}_h^{m+1} = \mathbf{U}_h^m + \Delta t \varphi_1 \left(\Delta t \mathbf{L}_h^{\text{per},m} \right) \mathbf{G}_h^{\text{per}}(\mathbf{U}_h^m),$$
$$\mathbf{U}_h^{m+1} = \widetilde{\mathbf{U}}_h^{m+1} + \Delta t \varphi_2 \left(\Delta t \mathbf{L}_h^{\text{per},m} \right) \left[\mathbf{G}_h^{\text{per}} \left(\widetilde{\mathbf{U}}_h^{m+1} \right) - \mathbf{G}_h^{\text{per}}(\mathbf{U}_h^m) - \mathbf{L}_h^{\text{per},m} \left(\widetilde{\mathbf{U}}_h^{m+1} - \mathbf{U}_h^m \right) \right], \tag{3.56}$$

where the linear operator $\mathbf{L}_h^{\text{per},m}$ is the Jacobian of $\mathbf{G}_h^{\text{per}}$ at $t = t_m$:

$$\mathbf{L}_h^{\text{per},m} = \frac{\partial \mathbf{G}_h^{\text{per}}}{\partial \mathbf{U}_h}(\mathbf{U}_h^m) = \begin{pmatrix} \mathbf{A}_{1,1}^m & \mathbf{A}_{1,2}^m & \mathbf{O} & \dots & \mathbf{A}_{1,N_x}^m \\ \mathbf{A}_{2,1}^m & \mathbf{A}_{2,2}^m & \mathbf{A}_{2,3}^m & \dots & \mathbf{O} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{O} & \dots & \mathbf{A}_{N_x-1,N_x-2}^m & \mathbf{A}_{N_x-1,N_x-1}^m & \mathbf{A}_{N_x-1,N_x}^m \\ \mathbf{A}_{N_x,1}^m & \dots & \mathbf{O} & \mathbf{A}_{N_x,N_x-1}^m & \mathbf{A}_{N_x,N_x}^m \end{pmatrix}. \tag{3.57}$$

Here the block matrix $\mathbf{A}_{i,i'}^m := \dfrac{\partial \mathbf{G}_i^{\text{per}}}{\partial \mathbf{U}_{i'}}(\mathbf{U}_{i'}^m)$.

Lemma 3.9. The ETDRK2 scheme (3.56) is conservative for any time step size Δt:

$$\int_\Omega u_h^m \, dx \, dy = \int_\Omega u_h^0 \, dx \, dy, \quad \forall m = 1, \dots, M. \tag{3.58}$$

Proof. We have

$$\int_\Omega u_h^m \, dx \, dy = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} |K_{i,j}| u_{i,j}^{00,m} = \mathbf{Z}^T \mathbf{U}_h^m,$$

where

$$\mathbf{Z} = \left(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{N_x} \right)^T, \ \mathbf{z}_i = \left(|K_{i,1}|, 0, 0, |K_{i,2}|, 0, 0, \dots, |K_{i,N_y}|, 0, 0 \right)^T \in \mathbb{R}^{3N_y}, \ 1 \le i \le N_x.$$

From (3.13), it is implied that

$$|K_{i,j}|\mathrm{g}_{i,j}^{00} = -\sum_{e\subset\partial K_{i,j}} \mathcal{J}_e\left(\mathfrak{h}_{e,K_{i,j}}\left(u_h^{\mathrm{int}(K_{i,j})}, u_h^{\mathrm{ext}(K_{i,j})}\right)\right), \tag{3.59}$$

for every element $K_{i,j}$. Taking the sum of (3.59) over $i \in \{1,\dots,N_x\}$ and $j \in \{1,\dots,N_y\}$ yields

$$\mathbf{Z}^T\mathbf{G}_h^{\mathrm{per}}(\mathbf{U}_h) = \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} |K_{i,j}|\mathrm{g}_{i,j}^{00} = 0, \tag{3.60}$$

where we have used the definition of $\mathbf{Z}$ and property (3.6) of the numerical flux $\mathfrak{h}_{e,K}$. Note that (3.60) holds for any vector $\mathbf{U}_h \in \mathbb{R}^{3N_xN_y}$, thus $\mathbf{Z}^T\mathbf{G}_h^{\mathrm{per}} \equiv 0$. As a consequence,

$$\mathbf{Z}^T\mathbf{L}_h^{\mathrm{per},m} = \frac{\partial(\mathbf{Z}^T\mathbf{G}_h^{\mathrm{per}})}{\partial\mathbf{U}_h}(\mathbf{U}_h^m) = \mathbf{0}. \tag{3.61}$$

Let $\mathbf{Z}^T\mathbf{L}_h^{\mathrm{per},m} = \begin{pmatrix}\zeta_{1,m}^T & \zeta_{2,m}^T & \cdots & \zeta_{N_x,m}^T\end{pmatrix}$ where $\zeta_{i,m} \in \mathbb{R}^{3N_y}$ satisfies

$$\begin{aligned}
\zeta_{1,m}^T &:= \mathbf{z}_1^T\mathbf{A}_{1,1}^m + \mathbf{z}_2^T\mathbf{A}_{2,1}^m + \mathbf{z}_{N_x}^T\mathbf{A}_{N_x,1}^m = \mathbf{0},\\
\zeta_{i,m}^T &:= \mathbf{z}_{i-1}^T\mathbf{A}_{i-1,i}^m + \mathbf{z}_i^T\mathbf{A}_{i,i}^m + \mathbf{z}_{i+1}^T\mathbf{A}_{i+1,i}^m = \mathbf{0}, \text{ for } 2 \le i \le N_x - 1,\\
\zeta_{N_x,m}^T &:= \mathbf{z}_1^T\mathbf{A}_{1,N_x}^m + \mathbf{z}_{N_x-1}^T\mathbf{A}_{N_x-1,N_x}^m + \mathbf{z}_{N_x}^T\mathbf{A}_{N_x,N_x}^m = \mathbf{0}.
\end{aligned} \tag{3.62}$$

By Taylor's expansion, there exist square matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ such that

$$\begin{aligned}
\varphi_1(\Delta t\mathbf{L}_h^{\mathrm{per},m}) &= (\Delta t\mathbf{L}_h^{\mathrm{per},m})^{-1}\left(e^{\Delta t\mathbf{L}_h^{\mathrm{per},m}} - \mathbf{I}\right) = \mathbf{I} + \mathbf{L}_h^{\mathrm{per},m}\mathbf{B}_1,\\
\varphi_2(\Delta t\mathbf{L}_h^{\mathrm{per},m}) &= (\Delta t\mathbf{L}_h^{\mathrm{per},m})^{-2}\left(e^{\Delta t\mathbf{L}_h^{\mathrm{per},m}} - \mathbf{I} - \Delta t\mathbf{L}_h^{\mathrm{per},m}\right) = \frac{1}{2}\mathbf{I} + \mathbf{L}_h^{\mathrm{per},m}\mathbf{B}_2.
\end{aligned}$$

Therefore, the following identities hold:

$$\mathbf{Z}^T\varphi_1(\Delta t\mathbf{L}_h^{\mathrm{per},m}) = \mathbf{Z}^T, \text{ and } \mathbf{Z}^T\varphi_2(\Delta t\mathbf{L}_h^{\mathrm{per},m}) = \frac{1}{2}\mathbf{Z}^T. \tag{3.63}$$

Multiply $\mathbf{Z}^T$ on both sides of equations (3.56) respectively, we then obtain

$$\begin{aligned}
\mathbf{Z}^T\widetilde{\mathbf{U}}_h^{m+1} &= \mathbf{Z}^T\mathbf{U}_h^m + \Delta t\mathbf{Z}^T\mathbf{G}_h^{\mathrm{per}}(\mathbf{U}_h^m),\\
\mathbf{Z}^T\mathbf{U}_h^{m+1} &= \mathbf{Z}^T\widetilde{\mathbf{U}}_h^{m+1} + \frac{\Delta t}{2}\left[\mathbf{Z}^T\mathbf{G}_h^{\mathrm{per}}(\widetilde{\mathbf{U}}_h^{m+1}) - \mathbf{Z}^T\mathbf{G}_h^{\mathrm{per}}(\mathbf{U}_h^m)\right].
\end{aligned}$$

Note that $\mathbf{Z}^T\mathbf{G}_h^{\mathrm{per}}(\mathbf{U}_h^m) = \mathbf{Z}^T\mathbf{G}_h^{\mathrm{per}}(\widetilde{\mathbf{U}}_h^{m+1}) = \mathbf{0}$ due to (3.60), we deduce that

$$\mathbf{Z}^T\mathbf{U}_h^{m+1} = \mathbf{Z}^T\widetilde{\mathbf{U}}_h^{m+1} = \mathbf{Z}^T\mathbf{U}_h^m,$$

for all $m = 0, 1, \dots, M-1$. □

3.5.2 Localized ETDRK2 method

Due to the periodic boundary conditions, we have another interface, denoted by $\underline{\Gamma} = \{x_{1/2}\} \times [0,1] \equiv \{x_{N_x+1/2}\} \times [0,1]$. Let $\underline{\boldsymbol{\lambda}}_h$ be the interface vector associated with $\underline{\Gamma}$. Similarly to (3.32), we obtain the two-subdomain formulation which is equivalent to (3.55) and consists of solving the local ODE systems:

$$\frac{d}{dt}\mathbf{W}_{h,1} = \mathcal{G}_{h,1}^{\mathrm{per}}(\mathbf{W}_{h,1}, \boldsymbol{\lambda}_h, \underline{\boldsymbol{\lambda}_h}), \quad \frac{d}{dt}\mathbf{W}_{h,2} = \mathcal{G}_{h,2}^{\mathrm{per}}(\mathbf{W}_{h,2}, \boldsymbol{\lambda}_h, \underline{\boldsymbol{\lambda}_h}), \tag{3.64}$$

together with the interface conditions, for $1 \le j \le N_y$, and $(l,k) \in \{(0,0),(1,0),(0,1)\}$:

$$\lambda_j^{lk} = \mathcal{I}_e\left(\mathfrak{h}_{e,K_{N_\Gamma,j}}\left(w_{h,1}^{\mathrm{int}(K_{N_\Gamma,j})}, w_{h,2}^{\mathrm{int}(K_{N_\Gamma+1,j})}\right)\phi_{N_\Gamma,j}^{lk}\right), \quad e = \partial K_{N_\Gamma,j} \cap \partial K_{N_\Gamma+1,j} \in \mathcal{E}_h^\Gamma, \tag{3.65a}$$

$$\underline{\lambda}_j^{lk} = \mathcal{I}_e\left(\mathfrak{h}_{e,K_{N_x,j}}\left(w_{h,2}^{\mathrm{int}(K_{N_x,j})}, w_{h,1}^{\mathrm{int}(K_{1,j})}\right)\phi_{N_x,j}^{lk}\right), \quad e = \partial K_{N_x,j} \cap \underline{\Gamma}. \tag{3.65b}$$

Again, we write $\boldsymbol{\lambda}_h = \mathcal{F}_h^\Gamma(\mathbf{W}_{h,1}, \mathbf{W}_{h,2})$ and $\underline{\boldsymbol{\lambda}}_h = \underline{\mathcal{F}_h^\Gamma}(\mathbf{W}_{h,1}, \mathbf{W}_{h,2})$ in place of the interface conditions (3.65).

As in the Dirichlet case (cf. Subsection 3.3.2), we construct the local linear operators $\mathbf{L}_{h,1}^{\mathrm{per},m}$ and $\mathbf{L}_{h,2}^{\mathrm{per},m}$ from the monodomain Jacobian $\mathbf{L}_h^{\mathrm{per},m}$. In addition, to ensure that the numerical solution given by the localized ETDRK2 is conservative, we further modify $\mathbf{L}_{h,1}^{\mathrm{per},m}$ and $\mathbf{L}_{h,2}^{\mathrm{per},m}$ so that they preserve (3.62). In particular, the local matrices need to satisfy

$$\begin{aligned} \mathbf{Z}_1^T\mathbf{L}_{h,1}^{\mathrm{per},m} &= \begin{pmatrix} \zeta_{1,m}^T & \zeta_{2,m}^T & \dots & \zeta_{N_\Gamma,m}^T \end{pmatrix} = \mathbf{0}, \\ \mathbf{Z}_2^T\mathbf{L}_{h,2}^{\mathrm{per},m} &= \begin{pmatrix} \zeta_{N_\Gamma+1,m}^T & \zeta_{N_\Gamma+2,m}^T & \dots & \zeta_{N_x,m}^T \end{pmatrix} = \mathbf{0}, \end{aligned} \tag{3.66}$$

where $\mathbf{Z}_1 = \left(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{N_\Gamma}\right)^T$, and $\mathbf{Z}_2 = \left(\mathbf{z}_{N_\Gamma+1}, \mathbf{z}_{N_\Gamma+2}, \dots, \mathbf{z}_{N_x}\right)^T$. This can be obtained by defining $\mathbf{L}_{h,1}^{\text{per},m}$ and $\mathbf{L}_{h,2}^{\text{per},m}$ as follows:

$$\mathbf{L}_{h,1}^{\text{per},m} = \begin{pmatrix} \mathbf{A}_{1,1}^m & \mathbf{A}_{1,2}^m & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{A}_{2,1}^m & \mathbf{A}_{2,2}^m & \mathbf{A}_{2,3}^m & \cdots & \mathbf{O} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{O} & \cdots & \mathbf{A}_{N_\Gamma-1,N_\Gamma-2}^m & \mathbf{A}_{N_\Gamma-1,N_\Gamma-1}^m & \mathbf{A}_{N_\Gamma-1,N_\Gamma}^m \\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{A}_{N_\Gamma,N_\Gamma-1}^m & \mathbf{A}_{N_\Gamma,N_\Gamma}^m \end{pmatrix} + \mathbf{M}_{h,1}^m,$$

and

$$\mathbf{L}_{h,2}^{\text{per},m} = \begin{pmatrix} \mathbf{A}_{N_\Gamma+1,N_\Gamma+1}^m & \mathbf{A}_{N_\Gamma+1,N_\Gamma+2}^m & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{A}_{N_\Gamma+2,N_\Gamma+1}^m & \mathbf{A}_{N_\Gamma+2,N_\Gamma+2}^m & \mathbf{A}_{N_\Gamma+2,N_\Gamma+3}^m & \cdots & \mathbf{O} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{O} & \cdots & \mathbf{A}_{N_x-1,N_x-2}^m & \mathbf{A}_{N_x-1,N_x-1}^m & \mathbf{A}_{N_x-1,N_x}^m \\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{A}_{N_x,N_x-1}^m & \mathbf{A}_{N_x,N_x}^m \end{pmatrix} + \mathbf{M}_{h,2}^m,$$

where $\mathbf{M}_{h,1}^m$ and $\mathbf{M}_{h,2}^m$ are sparse matrices given by

$$\mathbf{M}_{h,1}^m = \begin{pmatrix} \frac{1}{n_1}\mathbf{A}_{N_x,1}^m & \mathbf{O} & \cdots & \mathbf{O} \\ \frac{1}{n_1}\mathbf{A}_{N_x,1}^m & \mathbf{O} & \cdots & \mathbf{O} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \frac{1}{n_1}\mathbf{A}_{N_\Gamma+1,N_\Gamma}^m \\ \mathbf{O} & \mathbf{O} & \cdots & \frac{1}{n_1}\mathbf{A}_{N_\Gamma+1,N_\Gamma}^m \end{pmatrix}, \quad \mathbf{M}_{h,2}^m = \begin{pmatrix} \frac{1}{n_2}\mathbf{A}_{N_\Gamma,N_\Gamma+1}^m & \mathbf{O} & \cdots & \mathbf{O} \\ \frac{1}{n_2}\mathbf{A}_{N_\Gamma,N_\Gamma+1}^m & \mathbf{O} & \cdots & \mathbf{O} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \frac{1}{n_2}\mathbf{A}_{1,N_x}^m \\ \mathbf{O} & \mathbf{O} & \cdots & \frac{1}{n_2}\mathbf{A}_{1,N_x}^m \end{pmatrix}, \tag{3.67}$$

with

$$n_1 := \text{floor}(N_\Gamma/2), \quad \text{and} \quad n_2 := \text{floor}((N_x - N_\Gamma)/2). \tag{3.68}$$

In (3.67), only the first n_I block matrices in the first column of $\mathbf{M}_{h,I}^m$ and the last n_I block matrices in the last column of $\mathbf{M}_{h,I}^m$ are nonzero, for $I = 1, 2$.

The noniterative, localized ETDRK2 scheme with periodic boundary conditions is given by: for $m = 0, 1, \ldots, M-1$,

$$\lambda_h^m = \mathcal{F}_h^{\Gamma}\left(\mathbf{W}_{h,1}^m, \mathbf{W}_{h,2}^m\right),\ \underline{\lambda}_h^m = \underline{\mathcal{F}}_h^{\Gamma}\left(\mathbf{W}_{h,1}^m, \mathbf{W}_{h,2}^m\right), \tag{3.69a}$$

$$\widetilde{\mathbf{W}}_{h,I}^{m+1} = \mathbf{W}_{h,I}^m + \Delta t \varphi_1\left(\Delta t \mathbf{L}_{h,I}^{\text{per},m}\right) \mathcal{G}_{h,I}^{\text{per}}\left(\mathbf{W}_{h,I}^m, \lambda_h^m, \underline{\lambda}_h^m\right), \quad I = 1, 2, \tag{3.69b}$$

$$\widetilde{\lambda}_h^{m+1} = \mathcal{F}_h^{\Gamma}\left(\widetilde{\mathbf{W}}_{h,1}^{m+1}, \widetilde{\mathbf{W}}_{h,2}^{m+1}\right),\ \widetilde{\underline{\lambda}}_h^{m+1} = \underline{\mathcal{F}}_h^{\Gamma}\left(\widetilde{\mathbf{W}}_{h,1}^{m+1}, \widetilde{\mathbf{W}}_{h,2}^{m+1}\right), \tag{3.69c}$$

$$\begin{aligned}\mathbf{W}_{h,I}^{m+1} &= \widetilde{\mathbf{W}}_{h,I}^{m+1} + \Delta t \varphi_2\left(\Delta t \mathbf{L}_{h,I}^{\text{per},m}\right)\Big[\mathcal{G}_{h,I}^{\text{per}}\left(\widetilde{\mathbf{W}}_{h,I}^{m+1}, \widetilde{\lambda}_h^{m+1}, \widetilde{\underline{\lambda}}_h^{m+1}\right) - \mathcal{G}_{h,I}^{\text{per}}\left(\mathbf{W}_{h,I}^m, \lambda_h^m, \underline{\lambda}_h^m\right)\\ &\qquad - \mathbf{L}_{h,I}^{\text{per},m}\left(\widetilde{\mathbf{W}}_{h,I}^{m+1} - \mathbf{W}_{h,I}^m\right)\Big], \quad I = 1, 2.\end{aligned} \tag{3.69d}$$

Lemma 3.10. The localized ETDRK2 scheme (3.69) is conservative for any time step size Δt:

$$\int_{\Omega_1} w_{h,1}^m \, dx\, dy + \int_{\Omega_2} w_{h,2}^m \, dx\, dy = \int_{\Omega_1} w_{h,1}^0 \, dx\, dy + \int_{\Omega_2} w_{h,2}^0 \, dx\, dy, \quad \forall m = 1, \ldots, M. \tag{3.70}$$

Proof. As in the monodomain case (cf. Lemma 3.9), we use (3.13) and (3.6) as well as the interface conditions (3.65) to deduce that

$$\mathbf{Z}_1^T \mathcal{G}_{h,1}^{\text{per}}\left(\mathbf{W}_{h,1}^m, \lambda_h^m, \underline{\lambda}_h^m\right) = \sum_{i=1}^{N_\Gamma} \sum_{j=1}^{N_y} |K_{i,j}| g_{i,j}^{00,m} = -\sum_{j=1}^{N_y} \lambda_j^{00,m} + \sum_{j=1}^{N_y} \underline{\lambda}_j^{00,m} \tag{3.71}$$

and

$$\mathbf{Z}_2^T \mathcal{G}_{h,2}^{\text{per}}\left(\mathbf{W}_{h,2}^m, \lambda_h^m, \underline{\lambda}_h^m\right) = \sum_{i=N_\Gamma+1}^{N_x} \sum_{j=1}^{N_y} |K_{i,j}| g_{i,j}^{00,m} = \sum_{j=1}^{N_y} \lambda_j^{00,m} - \sum_{j=1}^{N_y} \underline{\lambda}_j^{00,m}. \tag{3.72}$$

The combination of (3.71) and (3.72) results in

$$\sum_{I=1}^{2} \mathbf{Z}_I^T \mathcal{G}_{h,I}^{\text{per}}\left(\mathbf{W}_{h,I}^m, \lambda_h^m, \underline{\lambda}_h^m\right) = 0. \tag{3.73}$$

Using the same argument, we also have

$$\sum_{I=1}^{2} \mathbf{Z}_I^T \mathcal{G}_{h,I}^{\text{per}}\left(\widetilde{\mathbf{W}}_{h,I}^{m+1}, \widetilde{\lambda}_h^{m+1}, \widetilde{\underline{\lambda}}_h^{m+1}\right) = 0. \tag{3.74}$$

We remark that, by the construction of $\mathbf{L}_{h,I}^{\text{per},m}$, $\mathbf{Z}_1^T \mathbf{L}_{h,1}^{\text{per},m} = \mathbf{0}$ and $\mathbf{Z}_2^T \mathbf{L}_{h,2}^{\text{per},m} = \mathbf{0}$ (cf. (3.66)-(3.67).

Therefore, as in (3.63), there holds

$$\mathbf{Z}_I^T \varphi_1(\Delta t \mathbf{L}_{h,I}^{\text{per},m}) = \mathbf{Z}_I^T, \text{ and } \mathbf{Z}_I^T \varphi_2(\Delta t \mathbf{L}_{h,I}^{\text{per},m}) = \frac{1}{2}\mathbf{Z}_I^T, \quad I = 1, 2. \tag{3.75}$$

Multiply $\mathbf{Z}_I^T$ on both sides of (3.69b) and (3.69d) respectively, we obtain

$$\begin{aligned}
&\mathbf{Z}_I^T \widetilde{\mathbf{W}}_{h,I}^{m+1} = \mathbf{Z}_I^T \mathbf{W}_{h,I}^m + \Delta t \mathbf{Z}_I^T \mathcal{G}_{h,I}^{\text{per}}(\mathbf{W}_{h,I}^m, \lambda_h^m, \underline{\lambda}_h^m), \\
&\mathbf{Z}_I^T \mathbf{W}_{h,I}^{m+1} = \mathbf{Z}_I^T \widetilde{\mathbf{W}}_{h,I}^{m+1} + \frac{\Delta t}{2}\left[\mathbf{Z}_I^T \mathcal{G}_{h,I}^{\text{per}}\left(\widetilde{\mathbf{W}}_{h,I}^{m+1}, \widetilde{\lambda}_h^{m+1}, \widetilde{\underline{\lambda}}_h^{m+1}\right) - \mathbf{Z}_I^T \mathcal{G}_{h,I}^{\text{per}}(\mathbf{W}_{h,I}^m, \lambda_h^m, \underline{\lambda}_h^m)\right].
\end{aligned} \tag{3.76}$$

Taking the sum over $I \in \{1, 2\}$ in (3.76) and using (3.73) and (3.74), we deduce that

$$\sum_{I=1}^{2} \mathbf{Z}_I^T \mathbf{W}_{h,I}^{m+1} = \sum_{I=1}^{2} \mathbf{Z}_I^T \widetilde{\mathbf{W}}_{h,I}^{m+1} = \sum_{I=1}^{2} \mathbf{Z}_I^T \mathbf{W}_{h,I}^m,$$

which completes the proof. □

Remark 3.11. Matrices $\mathbf{M}_{h,I}^m$ can be constructed in a different way than (3.67)-(3.68) to preserve discrete conservation. However, numerically we observe that the presented formulation for $\mathbf{M}_{h,I}^m$ leads to satisfactory results in terms of accuracy and stability of the noniterative localized ETDRK2 scheme, especially when the solution is discontinuous.

3.6 Numerical results

We consider the Burgers' equation in both one and two dimensions to illustrate the performance of the ETDRK2 scheme and the noniterative, localized ETDRK2 scheme with multiple subdomains. In the numerical experiments, we use uniform spatial meshes with $h = \Delta x = \Delta y$ and choose the Engquist-Osher numerical flux to handle the solution with discontinuities (similar results can also be obtained with the Lax-Friedrichs flux). The integrals in the weak formulation (cf. (3.10)-(3.11)) are computed by the two-point Gaussian rule in the $x-$ and $y-$directions. To efficiently compute actions of matrix exponentials or $\varphi-$functions on vectors, we use the Krylov subspace projection method [155] with the incomplete orthogonalization procedure (phipm/IOM2) proposed in [173] (see also [130, 138]). In the following test cases, we also compute the Courant number:

$$\mathcal{C} := \frac{\Delta t}{h} \max_u |f_1'(u)| \text{ in 1D}, \qquad \mathcal{C} := \frac{\Delta t}{h} \max_u \left(|f_1'(u)|, |f_2'(u)|\right) \text{ in 2D},$$

to confirm that the proposed global and localized ETDRK2 schemes allow much larger time step sizes than standard explicit time-stepping methods.

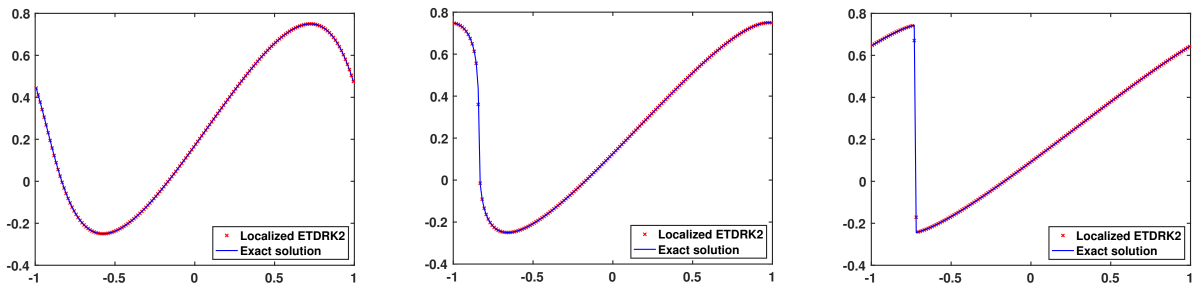
3.6.1 One-dimensional Burgers' equation

We consider the initial-boundary value problem:

$$\begin{aligned} & u_t + \left(\frac{u^2}{2}\right)_x = 0, && (x,t) \in (-1,1) \times (0,T), \\ & u(x,0) = \frac{1}{4} + \frac{1}{2}\sin(\pi x), && x \in (-1,1), \end{aligned} \tag{3.77}$$

subject to the periodic boundary conditions. The case of Dirichlet boundary conditions will also be discussed later (cf. Table 3.5). The exact solution to (3.77) is given in [78, 82], which is smooth up to $T = 2/\pi$. When $T = 1.1 > 2/\pi$, the solution becomes discontinuous near $x_{\text{shock}} \approx -0.72$ [82]. To capture the moving shock, we apply the slope limiter with fixed $C_M = 50$ (cf. Subsection 3.2.2). Figure 3.1 shows the snapshots at different times $T = 0.3, 2/\pi, 1.1$ of the exact solution and the numerical solution computed by the localized ETDRK2 scheme with 4 equal subdomains on a spatial mesh of size $h = 1/80$ and time step size $\Delta t = 1/80$. Note that we do not plot the numerical solution by the global ETDRK2 scheme since it is indistinguishable from the one by the localized ETDRK2. In Figure 3.2, we plot the evolution of mass by the global and localized ETDRK2 schemes, which confirms that mass is conserved up to machine precision.

To verify the convergence rates in time when the solution is smooth, we fix $T = 0.3$ and the spatial mesh size $h = 1/2048$ while varying the time step size Δt. For the noniterative, localized ETDRK2 scheme, we consider a decomposition of $(-1,1)$ into N_S equal subdomains/intervals, where $N_S = 2, 4, 8$. The L^1-errors of the numerical solutions given by the global and localized ETDRK2 schemes at $T = 0.3$ are reported in Table 3.1, while the corresponding running times are shown in Table 3.2 - note that these results are obtained with sequential computing. We observe second-order convergence in time for all schemes, with the time step sizes significantly



At $T = 0.3$. At $T = 2/\pi$. At $T = 1.1$.

Figure 3.1: [1D Burgers' equation] Snapshots of the exact solution and the numerical solution by the noniterative, localized ETDRK2 scheme with 4 subdomains and $h = \Delta t = 1/80$.

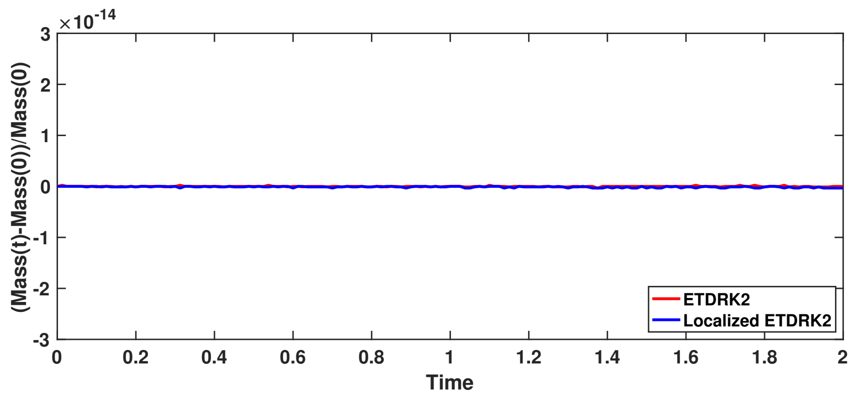


Figure 3.2: [1D Burgers' equation] Mass evolution of the global and localized ETDRK2 solutions with 4 subdomains and $h = \Delta t = 1/80$.

larger than the spatial mesh size, allowing the Courant number up to 115.2. Regarding accuracy, the errors become slightly larger when the number of subdomains increases since there is no iteration in the localized ETDRK2 scheme and large time step sizes were used. However, localized ETDRK2 is computationally efficient with similar running times as the global ETDRK2 scheme, even when the subdomain problems are solved sequentially.

Next, we compare the accuracy and efficiency of the proposed ETDRK2 schemes with the SSP-RK2 algorithm [36] for $T = 0.3$. We use the same spatial mesh size and different time step sizes for SSP-RK2 and ETDRK2 schemes since SSP-RK2 is restricted by a CFL condition for stability [36, 37]; the corresponding L^1-errors are reported in Table 3.3. We first observe that second-order convergence is confirmed when the spatial mesh size and time step size are refined simultaneously. Importantly, the errors by all ETDRK2 schemes with much larger time step sizes

Table 3.1: [1D Burgers' equation] L^1-errors at $T = 0.3$ and corresponding convergence rates in time by the global and localized ETDRK2 schemes with $h = 1/2048$.

Δt	$\mathcal{C}$	ETDRK2	Localized ETDRK2		
			2 subdomains	4 subdomains	8 subdomains
T/4	115.2	4.73e-04 –	9.00e-04 –	1.88e-03 –	3.97e-03 –
T/8	57.6	1.17e-04 [2.02]	2.25e-04 [2.00]	4.49e-04 [2.07]	8.50e-04 [2.22]
T/16	28.8	2.90e-05 [2.01]	5.50e-05 [2.03]	1.08e-04 [2.06]	1.90e-04 [2.16]
T/32	14.4	7.24e-06 [2.00]	1.32e-05 [2.06]	2.54e-05 [2.09]	4.15e-05 [2.19]

Table 3.2: [1D Burgers' equation] Running times (in seconds) of the global and localized ETDRK2 schemes with $h = 1/2048$ and $T = 0.3$.

Δt	ETDRK2	Localized ETDRK2		
		2 subdomains	4 subdomains	8 subdomains
T/4	2.28	2.07	2.21	2.41
T/8	2.22	2.70	2.55	2.85
T/16	2.75	2.37	2.74	2.84
T/32	2.27	2.40	2.54	2.84

are comparable to those by the SSP-RK2 algorithm. Moreover, the noniterative, localized ETDRK2 solution with multiple subdomains is as accurate as the global ETDRK2 solution.

Table 3.3: [1D Burgers' equation] L^1-errors and convergence rates in both space and time by the SSP-RK2, global and localized ETDRK2 schemes at $T = 0.3$. The Courant numbers for the SSP-RK2 and global/localized ETDRK2 schemes are $\mathcal{C}_{\text{SSP-RK2}} = 0.331$ [36, 37] and $\mathcal{C} = 0.75$, respectively.

h	$\Delta t_{\text{SSP-RK2}}$	SSP-RK2	Δt	ETDRK2	Localized ETDRK2		
					2 subdomains	4 subdomains	8 subdomains
1/64	1/145	1.57e-04 –	1/60	1.62e-04 –	1.66e-04 –	1.70e-04 –	1.82e-04 –
1/128	1/290	3.91e-05 [2.01]	1/120	4.02e-05 [2.01]	4.14e-05 [2.00]	4.22e-05 [2.01]	4.47e-05 [2.03]
1/256	1/580	9.75e-06 [2.00]	1/240	1.00e-05 [2.01]	1.03e-05 [2.01]	1.05e-05 [2.01]	1.11e-05 [2.01]
1/512	1/1160	2.44e-06 [2.00]	1/480	2.50e-06 [2.00]	2.58e-06 [2.00]	2.63e-06 [2.00]	2.76e-06 [2.01]

We now investigate the performance of ETDRK2 schemes when a moving shock is de-

Table 3.4: [1D Burgers' equation] L^1-global errors and errors in smooth regions (where $|x - x_{\text{shock}}| \geq 0.1$) by the global and localized ETDRK2 schemes at $T = 1.1$. The corresponding convergence rates are shown in square brackets. The Courant number is $\mathcal{C} = 0.75$.

h	Δt	ETDRK2		Localized ETDRK2			
				2 subdomains		4 subdomains	
		Global	Smooth	Global	Smooth	Global	Smooth
1/40	1/40	1.03e-02	5.04e-05	1.03e-02	1.01e-04	1.04e-02	1.24e-04
		–	–	–	–	–	–
1/80	1/80	5.12e-03	1.23e-05	5.12e-03	2.50e-05	5.13e-03	2.98e-05
		[1.01]	[2.03]	[1.01]	[2.01]	[1.02]	[2.06]
1/160	1/160	2.55e-03	3.03e-06	2.55e-03	6.21e-06	2.55e-03	7.26e-06
		[1.01]	[2.02]	[1.01]	[2.01]	[1.01]	[2.04]
1/320	1/320	1.27e-03	7.50e-07	1.27e-03	1.55e-06	1.27e-03	1.80e-06
		[1.01]	[2.01]	[1.01]	[2.00]	[1.01]	[2.01]

veloped. We let $T = 1.1$ and compute the L^1-errors of the solutions by global and localized ETDRK2 over the whole spatial domain and in smooth regions (where $|x - x_{\text{shock}}| \geq 0.1$). The results are reported in Table 3.4, where Δt is chosen to be smaller than Δx to capture the moving shock. Since low order approximation is applied near the shock, we observe the global first-order accuracy of the global and localized ETDRK2 with two and four subdomains. In smooth regions, both methods still preserve the expected second-order convergence.

Finally, we consider the case of Dirichlet boundary conditions (instead of periodic conditions) for problem (3.77):

$$u(-1,t) = \gamma(t) = v(-1,t), \quad t \in (0,T),$$

where $v(x,t)$ is the exact solution to (3.77) (see [36, Example 2]). Note that $x = -1$ is an inflow boundary while $x = 1$ is an outflow boundary. For $T = 0.3$ and fixed $h = 1/2048$, we verify the expected convergence rates in time of the global and localized ETDRK2 as shown in Table 3.5. We notice that the errors are slightly larger than the case with periodic conditions.

Table 3.5: [1D Burgers' equation with Dirichlet conditions] L^1-errors at $T = 0.3$ and corresponding convergence rates in time by the global and localized ETDRK2 schemes with $h = 1/2048$.

Δt	$\mathcal{C}$	ETDRK2	Localized ETDRK2		
			2 subdomains	4 subdomains	8 subdomains
T/4	115.2	6.37e-04 –	6.92e-04 –	8.67e-04 –	1.29e-03 –
T/8	57.6	1.57e-04 [2.02]	1.67e-04 [2.05]	2.08e-04 [2.06]	3.11e-04 [2.05]
T/16	28.8	3.87e-05 [2.02]	4.02e-05 [2.05]	4.96e-05 [2.07]	7.47e-05 [2.06]
T/32	14.4	9.43e-06 [2.04]	9.61e-06 [2.06]	1.18e-05 [2.07]	1.75e-05 [2.09]

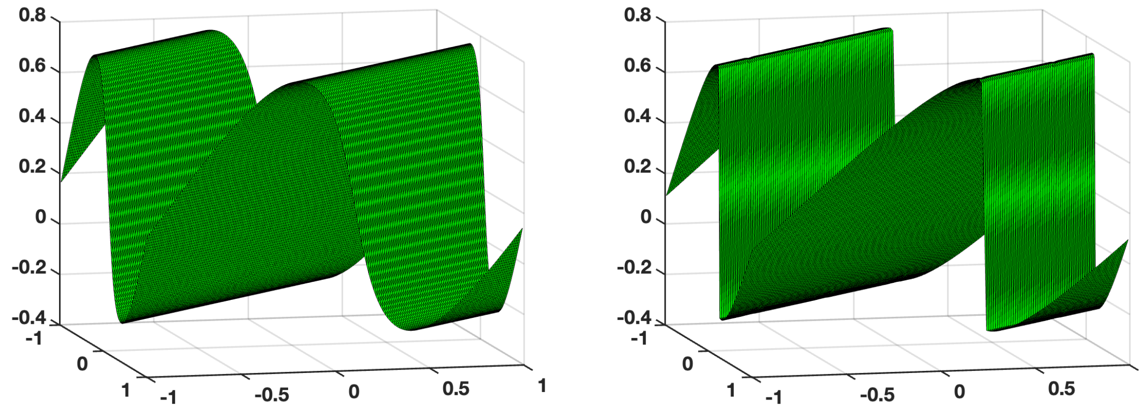
3.6.2 Two-dimensional Burgers' equation

We now test the performance of the global and localized ETDRK2 schemes on a two-dimensional counterpart of the previous example:

$$\begin{aligned} & u_t + \left(\frac{u^2}{2}\right)_x + \left(\frac{u^2}{2}\right)_y = 0, && (x, y, t) \in (-1, 1)^2 \times (0, T), \\ & u(x, y, 0) = \frac{1}{4} + \frac{1}{2} \sin\left(\pi(x + y)\right), && (x, y) \in (-1, 1)^2, \end{aligned} \tag{3.78}$$

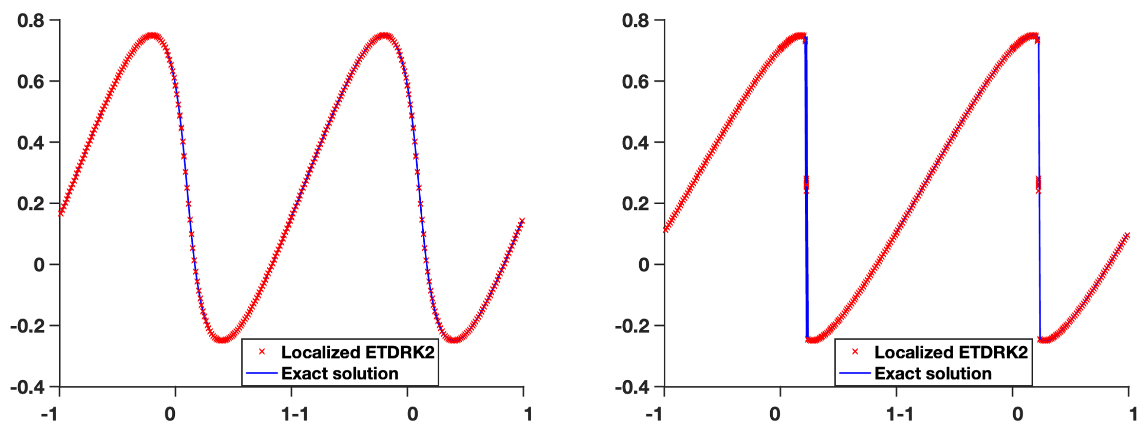
subject to the periodic boundary conditions. For localized ETDRK2, we decompose the spatial domain into N_S equal strips of the form $(x_I, x_{I+1}) \times (-1, 1)$ for $I = 1, \dots, N_S$, with $N_S = 2, 4, 8$. The numerical solution by the localized ETDRK2 scheme with 4 strips/subdomains is displayed in Figure 3.3 at $T = 0.2$ and $T = 0.45$. Their corresponding diagonal cuts along with those of the exact solution are shown in Figure 3.4. We can observe that the numerical solution approximates well the entropy solution; it is free off any unphysical undershoot or overshoot and captures well the discontinuities. In addition, the evolution of mass by the global and localized ETDRK2 solutions is depicted in Figure 3.5, confirming that discrete mass is conserved over time.

We first consider $T = 0.2$ when the solution is smooth and verify the order of convergence in time of the ETDRK2 schemes using a fixed spatial mesh size $h = 1/400$ and varying large time step sizes Δt, with the Courant number up to 30. The L^1-errors between the exact solution and numerical solution are reported in Table 3.6, where second-order convergence is observed for the



At $T = 0.2$. At $T = 0.45$.

Figure 3.3: [2D Burgers' equation] Snapshots of the numerical solution by the localized ETDRK2 with 4 subdomains and $h = \Delta t = 1/80$.



At $T = 0.2$. At $T = 0.45$.

Figure 3.4: [2D Burgers' equation] Snapshots of the diagonal cuts of the exact solution and the numerical solution by localized ETDRK2 with 4 subdomains and $h = \Delta t = 1/80$.

global and localized ETDRK2 schemes. Moreover, we remark that the accuracy of the localized ETDRK2 scheme (without subdomain iterations) is almost the same as global ETDRK2, even with many subdomains. The computational times needed to run the convergence in time test are reported in Table 3.7. We notice that for the localized ETDRK2 scheme, the subdomain problems are solved sequentially, thus the overall cost is higher than the global scheme. However, the running times of localized ETDRK2 only increase slightly as the number of subdomain increases, which shows the great potential of the method for parallel simulations. Next, we focus on the

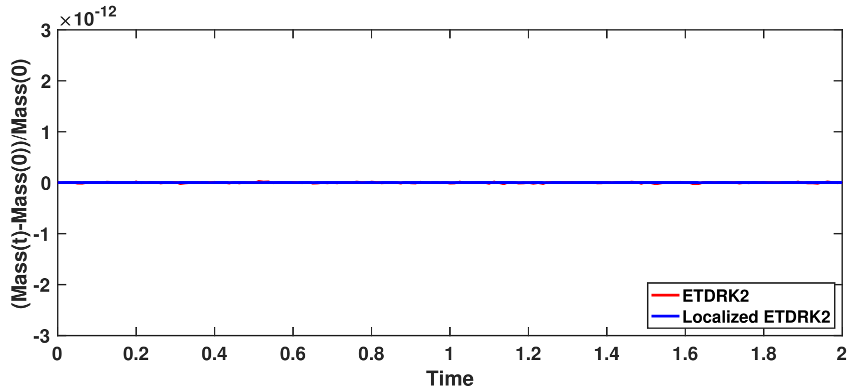


Figure 3.5: [2D Burgers' equation] Mass evolution of the global and localized ETDRK2 solutions with 4 subdomains and $h = \Delta t = 1/80$.

case $T = 0.45$ where discontinuity curves appear in the solution. Table 3.8 shows the L^1-errors in both space and time from the global and localized ETDRK2 schemes. Similarly to the 1D problem, both schemes achieve first-order convergence over the whole region due to lower order approximation near the discontinuities, and maintain second-order convergence over the smooth regions.

Table 3.6: [2D Burgers' equation] L^1-errors at $T = 0.2$ and corresponding convergence rates in time by the global and localized ETDRK2 schemes with $h = 1/400$.

Δt	$\mathcal{C}$	ETDRK2	Localized ETDRK2		
			2 subdomains	4 subdomains	8 subdomains
$T/2$	30	7.51e-03 –	9.24e-03 –	1.27e-02 –	2.37e-02 –
$T/4$	15	1.91e-03 [1.98]	2.25e-03 [2.04]	2.93e-03 [2.12]	4.82e-03 [2.30]
$T/8$	7.5	4.74e-04 [2.01]	5.46e-04 [2.04]	6.83e-04 [2.10]	1.02e-03 [2.24]
$T/16$	3.75	1.19e-04 [1.99]	1.36e-04 [2.01]	1.65e-04 [2.05]	2.18e-04 [2.23]

Finally, we replace the periodic boundary conditions by the Dirichlet conditions on the inflow boundaries at $x = -1$ and $y = -1$. We fix the spatial mesh $h = 1/400$ and compute the L^1−errors in time at $T = 0.2$ between the numerical solutions by the global and localized ETDRK2 schemes and a reference solution with a fine time step size $\Delta t_{\text{ref}} = T/512$. The results are reported in Table 3.9, where the second-order convergence in time of both ETDRK schemes is observed as expected (cf. Theorems 3.6 and 3.7).

Table 3.7: [2D Burgers' equation] Running times (in seconds) of the global and localized ETDRK2 schemes with $h = 1/400$ and $T = 0.2$.

Δt	ETDRK2	Localized ETDRK2		
		2 subdomains	4 subdomains	8 subdomains
$T/2$	17.13	29.41	27.95	29.12
$T/4$	19.61	29.17	29.85	30.72
$T/8$	20.59	29.43	29.18	30.35
$T/16$	24.61	40.24	39.61	42.79

Table 3.8: [2D Burgers' equation] L^1-global errors and errors in the smooth region $[-0.2, 0.4]^2$ by the global and localized ETDRK2 schemes at $T = 0.45$. The corresponding convergence rates are shown in square brackets. The Courant number is $\mathcal{C} = 0.5625$.

h	Δt	ETDRK2		Localized ETDRK2			
				2 subdomains		4 subdomains	
		Global	Smooth	Global	Smooth	Global	Smooth
1/10	$T/6$	7.67e-02 –	1.10e-04 –	7.66e-02 –	2.05e-04 –	7.77e-02 –	4.33e-04 –
1/20	$T/12$	4.18e-02 [0.88]	3.45e-05 [1.67]	4.21e-02 [0.86]	4.59e-05 [2.16]	4.26e-02 [0.87]	8.08e-05 [2.42]
1/40	$T/24$	1.81e-02 [1.21]	9.63e-06 [1.84]	1.82e-02 [1.21]	1.11e-05 [2.05]	1.83e-02 [1.22]	1.85e-05 [2.13]
1/80	$T/48$	8.90e-03 [1.02]	2.51e-06 [1.94]	8.96e-03 [1.02]	2.74e-06 [2.02]	9.00e-03 [1.02]	4.42e-06 [2.07]

Table 3.9: [2D Burgers' equation with Dirichlet conditions] L^1-errors at $T = 0.2$ and corresponding convergence rates in time by the global and localized ETDRK2 schemes with $h = 1/400$.

Δt	$\mathcal{C}$	ETDRK2	Localized ETDRK2		
			2 subdomains	4 subdomains	8 subdomains
$T/2$	30	8.68e-03 –	9.31e-03 –	1.06e-02 –	1.31e-02 –
$T/4$	15	2.17e-03 [2.00]	2.32e-03 [2.00]	2.60e-03 [2.03]	3.15e-03 [2.06]
$T/8$	7.5	5.27e-04 [2.04]	5.63e-04 [2.04]	6.29e-04 [2.05]	7.57e-04 [2.06]
$T/16$	3.75	1.29e-04 [2.03]	1.38e-04 [2.03]	1.54e-04 [2.03]	1.85e-04 [2.03]

Chapter 4

Dynamically Regularized Lagrange Multiplier Methods for Incompressible Fluid Flows

This chapter presents dynamically regularized Lagrange multiplier (DRLM) methods for the incompressible Navier-Stokes (NS) equations and the coupled Cahn-Hilliard-Navier-Stokes (CH-NS) system. The chapter is organized into three parts. In the first part, we propose the DRLM framework for the incompressible NS equations: a time-dependent Lagrange multiplier and a regularization parameter are introduced to reformulate the NS equations into an equivalent system that incorporates the kinetic energy evolution. First- and second-order DRLM schemes based on backward differentiation formulas are derived and shown to be unconditionally energy stable with respect to the original variables. Fully discrete energy stability is also proved with the Marker-and-Cell (MAC) discretization in space. In the second part, we carry out rigorous temporal error analysis for the first-order DRLM scheme. Based on a uniform bound on the Lagrange multiplier and mathematical induction, we establish optimal error estimates for the velocity, Lagrange multiplier, and pressure. These estimates are robust with respect to the regularization parameter θ, meaning that the error bounds decay as θ increases. In the third part, we extend the DRLM framework to the coupled CH-NS system by introducing a Lagrange multiplier for each of the Cahn-Hilliard and Navier-Stokes equations. The resulting DRLM schemes are fully decoupled, mass-conserving, unconditionally energy stable, and require no stabilization terms. Various numerical experiments in two and three dimensions verify the theoretical findings and illustrate the accuracy and robustness of the proposed DRLM schemes. The results presented in this chapter can be found in [48, 50, 49].

4.1 DRLM method for the incompressible Navier-Stokes equations

We consider the incompressible NS equations, which govern the behavior of viscous fluid flows and take the following form:

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f}, \quad \text{in } \Omega\times(0,T], \tag{4.1a}$$

$$\nabla\cdot\mathbf{u} = 0, \quad \text{in } \Omega\times(0,T], \tag{4.1b}$$

subject to the initial condition $\mathbf{u}(\mathbf{x},0) = \mathbf{u}_0(\mathbf{x})$ and the homogeneous Dirichlet or periodic boundary conditions. In (4.1), Ω is an open bounded domain in $\mathbb{R}^d$ ($d = 2,3$), $\mathbf{u} = (u_1,\dots,u_d)^T$ and p are the unknown velocity field and pressure respectively, $(\mathbf{u}\cdot\nabla)\mathbf{u}$ denotes the nonlinear convection, $\mathbf{f}$ is an external force, and ν represents the kinematic viscosity which is inversely proportional to the Reynolds number Re. It is well known that, in the absence of external forces, the incompressible NS equations (4.1) under the prescribed boundary conditions satisfy the energy dissipation law. Therefore, it has been highly desirable to develop numerical schemes for (4.1) that preserve this key property.

4.1.1 The DRLM formulation and time-discrete schemes

For any two functions $\mathbf{u},\mathbf{v}\in(L^2(\Omega))^d$, denote by $(\mathbf{u},\mathbf{v})$ the standard L^2 inner product of $\mathbf{u}$ and $\mathbf{v}$ on Ω, and $\|\mathbf{u}\|_0 = \sqrt{(\mathbf{u},\mathbf{u})}$ is the corresponding L^2 norm of $\mathbf{u}$. Let $\mathcal{K}(\mathbf{u}) = \frac{1}{2}\int_\Omega |\mathbf{u}|^2 d\mathbf{x}$ be the kinetic energy associated with (4.1), it is clear that

$$\frac{d\mathcal{K}(\mathbf{u})}{dt} = (\mathbf{u}_t,\mathbf{u}) \quad \text{and} \quad ((\mathbf{u}\cdot\nabla)\mathbf{u},\mathbf{u}) = 0. \tag{4.2}$$

In [197], a Lagrange multiplier $q(t)$ was introduced to reformulate the NS equations (4.1) as follows:

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + q(\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f}, \tag{4.3a}$$

$$\nabla\cdot\mathbf{u} = 0, \tag{4.3b}$$

$$\frac{d\mathcal{K}(\mathbf{u})}{dt} = (\mathbf{u}_t + q(\mathbf{u}\cdot\nabla)\mathbf{u},\mathbf{u}). \tag{4.3c}$$

However, the system (4.3) is not equivalent to the original system (4.1) because any constant function $q(t)$ satisfies (4.3c). To ensure the uniqueness of q, we introduce a dynamic term for q in (4.3c) and consider instead the following equations:

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + q(\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f}, \tag{4.4a}$$

$$\nabla\cdot\mathbf{u} = 0, \tag{4.4b}$$

$$\frac{d\mathcal{K}(\mathbf{u})}{dt} + \theta\frac{dq^2}{dt} = (\mathbf{u}_t + q(\mathbf{u}\cdot\nabla)\mathbf{u}, \mathbf{u}), \tag{4.4c}$$

where $q(0) = 1$ and $\theta > 0$ is a regularization parameter to be thoroughly studied in the numerical experiments (cf. Section 4.1.3). From (4.2) and (4.4c), we can verify that at the continuous level, $q(t) = 1$ for all $t > 0$. Consequently, the new system (4.4) is equivalent to the original NS equations (4.1). It should be noted that, unlike the SAV approach, the function q serves as a Lagrange multiplier to enforce dissipation of the original energy. Additionally, the parameter θ is crucial in our formulation to guarantee the uniqueness of the Lagrange multiplier at the discrete level, and thus the numerical solutions. Let us next assume a uniform partition of the time interval $[0, T]$: $0 = t_0 < t_1 < \ldots < t_{N_t} = T$ with the time step size $\Delta t = {}^{T}/_{N_t}$ and consider the time discretization of the DRLM system (4.4).

Remark 4.1. In the case of inhomogeneous Dirichlet boundary conditions where $\mathbf{u}|_{\partial\Omega} = \mathbf{u}_b$, (4.4c) is replaced by

$$\frac{d\mathcal{K}(\mathbf{u})}{dt} + \theta\frac{dq^2}{dt} = (\mathbf{u}_t, \mathbf{u}) + q\left[((\mathbf{u}\cdot\nabla)\mathbf{u}, \mathbf{u}) - \frac{1}{2}(\mathbf{u}_b\cdot\mathbf{n}, |\mathbf{u}_b|^2)_{\partial\Omega}\right],$$

where $(\cdot,\cdot)_{\partial\Omega}$ denotes the L^2 inner product on $\partial\Omega$ and $\mathbf{n}$ represents the outward unit normal vector to $\partial\Omega$.

The first-order in time DRLM scheme By applying the first-order BDF (i.e., backward Euler) method to equations (4.4a)-(4.4c) with an explicit treatment of the nonlinear convection term, we obtain the following first-order in time DRLM scheme: for $0 \le n \le N_t - 1$,

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu\Delta\mathbf{u}^{n+1} + q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n + \nabla p^{n+1} = \mathbf{f}^{n+1}, \tag{4.5a}$$

$$\nabla \cdot \mathbf{u}^{n+1} = 0, \tag{4.5b}$$

$$\frac{\mathcal{K}(\mathbf{u}^{n+1}) - \mathcal{K}(\mathbf{u}^n)}{\Delta t} + \theta\frac{(q^{n+1})^2 - (q^n)^2}{\Delta t} = \left(\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n, \mathbf{u}^{n+1}\right), \tag{4.5c}$$

where $\mathbf{u}^{n+1}$ satisfies either the periodic or homogeneous Dirichlet boundary conditions, $\mathbf{f}^{n+1} = \mathbf{f}(t_{n+1})$ is given, $\mathbf{u}^n$ and p^n are approximations of $\mathbf{u}(t_n)$ and $p(t_n)$, respectively.

Following [123, 124, 121, 197], an efficient implementation of the first-order DRLM scheme (4.5) can be achieved by decomposing each of the unknown quantities, the velocity $\mathbf{u}^{n+1}$ and the pressure p^{n+1}, into two components as follows:

$$\mathbf{u}^{n+1} = \mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1}, \tag{4.6}$$

$$p^{n+1} = p_1^{n+1} + q^{n+1}p_2^{n+1}. \tag{4.7}$$

Plugging the above equations into the system (4.5), we arrive at

$$\frac{\mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1} - \mathbf{u}^n}{\Delta t} - \nu\Delta(\mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1}) + q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n + \nabla(p_1^{n+1} + q^{n+1}p_2^{n+1}) = \mathbf{f}^{n+1}, \tag{4.8a}$$

$$\nabla \cdot (\mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1}) = 0, \tag{4.8b}$$

$$\begin{aligned}&\frac{\mathcal{K}(\mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1}) - \mathcal{K}(\mathbf{u}^n)}{\Delta t} + \theta\frac{(q^{n+1})^2 - (q^n)^2}{\Delta t} \\ &\qquad = \left(\frac{\mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1} - \mathbf{u}^n}{\Delta t} + q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n, \mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1}\right).\end{aligned} \tag{4.8c}$$

We then split (4.8a) and (4.8b) into two systems of generalized Stokes equations:

$$\frac{\mathbf{u}_1^{n+1} - \mathbf{u}^n}{\Delta t} - \nu\Delta\mathbf{u}_1^{n+1} + \nabla p_1^{n+1} = \mathbf{f}^{n+1}, \tag{4.9a}$$

$$\nabla \cdot \mathbf{u}_1^{n+1} = 0, \tag{4.9b}$$

and

$$\frac{\mathbf{u}_2^{n+1}}{\Delta t} - \nu\Delta\mathbf{u}_2^{n+1} + (\mathbf{u}^n \cdot \nabla)\mathbf{u}^n + \nabla p_2^{n+1} = 0, \tag{4.10a}$$

$$\nabla \cdot \mathbf{u}_2^{n+1} = 0, \tag{4.10b}$$

where $\mathbf{u}_1^{n+1}$ and $\mathbf{u}_2^{n+1}$ satisfy the same boundary conditions as $\mathbf{u}^{n+1}$, i.e., either homogeneous Dirichlet or periodic boundary conditions. For the NS equations with inhomogeneous Dirichlet data $\mathbf{u}|_{\partial\Omega} = \mathbf{u}_b$, we impose $\mathbf{u}_1^{n+1} = \mathbf{u}_b(t_{n+1})$ and $\mathbf{u}_2^{n+1} = 0$ on $\partial\Omega$.

Note that the two systems (4.9) and (4.10) are linear and can be solved independently of q^{n+1} (cf. Subsection 4.1.2 and [123, 149]). Once $\mathbf{u}_1^{n+1}, \mathbf{u}_2^{n+1}$, p_1^{n+1}, and p_2^{n+1} are determined, we compute q^{n+1} from (4.8c) by solving the quadratic equation:

$$A_{1,n+1}(q^{n+1})^2 + B_{1,n+1}q^{n+1} + C_{1,n+1} = 0, \tag{4.11}$$

where

$$\begin{aligned}
A_{1,n+1} &= \frac{1}{2}\|\mathbf{u}_2^{n+1}\|_0^2 + \theta + \Delta t\nu\|\nabla\mathbf{u}_2^{n+1}\|_0^2,\\
B_{1,n+1} &= -(\mathbf{u}_1^{n+1} - \mathbf{u}^n, \mathbf{u}_2^{n+1}) - \Delta t((\mathbf{u}^n\cdot\nabla)\mathbf{u}^n, \mathbf{u}_1^{n+1}),\\
C_{1,n+1} &= -\frac{1}{2}\|\mathbf{u}_1^{n+1} - \mathbf{u}^n\|_0^2 - \theta(q^n)^2.
\end{aligned}$$

To derive $A_{1,n+1}$, we have used the property $(\frac{1}{\Delta t}\mathbf{u}_2^{n+1} + (\mathbf{u}^n\cdot\nabla)\mathbf{u}^n, \mathbf{u}_2^{n+1}) = -\nu\|\nabla\mathbf{u}_2^{n+1}\|_0^2$ obtained from (4.10) and integration by parts.

We remark that $A_{1,n+1} > 0$ and $C_{1,n+1} < 0$ for any $\theta > 0$. Therefore, the quadratic equation (4.11) has a *unique positive solution* q^{n+1} for any positive regularization parameter θ. Once q^{n+1} is known, we update $\mathbf{u}^{n+1}$ and p^{n+1} using equations (4.6) and (4.7), respectively.

The second-order in time DRLM scheme To construct the second-order DRLM scheme, we assume $\mathbf{u}^{n-1}$ and $\mathbf{u}^n$ are given and denote by $\tilde{\mathbf{u}}^{n+1} = 2\mathbf{u}^n - \mathbf{u}^{n-1}$ the second-order extrapolation to be used for the approximation of the nonlinear convection term. By applying the BDF2 method to equations (4.4a)-(4.4c), we obtain the following second-order DRLM scheme:

for $1 \leq n \leq N_t - 1$,

$$\frac{3\mathbf{u}^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} - \nu\Delta\mathbf{u}^{n+1} + q^{n+1}(\tilde{\mathbf{u}}^{n+1} \cdot \nabla)\tilde{\mathbf{u}}^{n+1} + \nabla p^{n+1} = \mathbf{f}^{n+1}, \tag{4.12a}$$

$$\nabla \cdot \mathbf{u}^{n+1} = 0, \tag{4.12b}$$

$$\begin{aligned}\frac{3\mathcal{K}(\mathbf{u}^{n+1}) - 4\mathcal{K}(\mathbf{u}^n) + \mathcal{K}(\mathbf{u}^{n-1})}{2\Delta t} &+ \theta\frac{3(q^{n+1})^2 - 4(q^n)^2 + (q^{n-1})^2}{2\Delta t} \\ &= \left(\frac{3\mathbf{u}^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} + q^{n+1}(\tilde{\mathbf{u}}^{n+1} \cdot \nabla)\tilde{\mathbf{u}}^{n+1}, \mathbf{u}^{n+1}\right),\end{aligned} \tag{4.12c}$$

where $\mathbf{u}^{n+1}$ satisfies either the periodic or homogeneous Dirichlet boundary conditions, $\mathbf{u}^1$ and q^1 are computed using the first-order DRLM scheme (4.5).

The second-order DRLM scheme (4.12) can also be implemented efficiently by again decomposing each of the two unknowns $\mathbf{u}^{n+1}$ and p^{n+1} into two components, as in (4.6) and (4.7). From (4.12a) and (4.12b), we obtain the generalized Stokes systems for $\mathbf{u}_i^{n+1}$ and p_i^{n+1} $(i = 1, 2)$ as follows:

$$\frac{3\mathbf{u}_1^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} - \nu\Delta\mathbf{u}_1^{n+1} + \nabla p_1^{n+1} = \mathbf{f}^{n+1}, \tag{4.13a}$$

$$\nabla \cdot \mathbf{u}_1^{n+1} = 0, \tag{4.13b}$$

and

$$\frac{3\mathbf{u}_2^{n+1}}{2\Delta t} - \nu\Delta\mathbf{u}_2^{n+1} + (\tilde{\mathbf{u}}^{n+1} \cdot \nabla)\tilde{\mathbf{u}}^{n+1} + \nabla p_2^{n+1} = 0, \tag{4.14a}$$

$$\nabla \cdot \mathbf{u}_2^{n+1} = 0. \tag{4.14b}$$

After determining $\mathbf{u}_i^{n+1}$ and p_i^{n+1} $(i = 1, 2)$ from the linear systems (4.13) and (4.14), the discrete Lagrange multiplier q^{n+1} is obtained from (4.12c) by solving the quadratic equation:

$$A_{2,n+1}(q^{n+1})^2 + B_{2,n+1}q^{n+1} + C_{2,n+1} = 0, \tag{4.15}$$

where

$$
\begin{aligned}
A_{2,n+1} &= \frac{3}{2}\|\mathbf{u}_2^{n+1}\|_0^2 + 3\theta + 2\Delta t\nu\|\nabla \mathbf{u}_2^{n+1}\|_0^2, \\
B_{2,n+1} &= -(3\mathbf{u}_1^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}, \mathbf{u}_2^{n+1}) - 2\Delta t((\tilde{\mathbf{u}}^{n+1}\cdot\nabla)\tilde{\mathbf{u}}^{n+1}, \mathbf{u}_1^{n+1}), \\
C_{2,n+1} &= -2\|\mathbf{u}_1^{n+1} - \mathbf{u}^n\|_0^2 + \frac{1}{2}\|\mathbf{u}_1^{n+1} - \mathbf{u}^{n-1}\|_0^2 - \theta\left[4(q^n)^2 - (q^{n-1})^2\right].
\end{aligned}
$$

To derive $A_{2,n+1}$, we have used the fact that $(\frac{3}{2\Delta t}\mathbf{u}_2^{n+1} + (\tilde{\mathbf{u}}^{n+1}\cdot\nabla)\tilde{\mathbf{u}}^{n+1}, \mathbf{u}_2^{n+1}) = -\nu\|\nabla\mathbf{u}_2^{n+1}\|_0^2$ due to (4.14) and integration by parts. Since $A_{2,n+1} > 0$ for all $\theta > 0$, the algebraic equation (4.15) attains two solutions:

$$
q_{\pm}^{n+1} = \frac{-B_{2,n+1} \pm \sqrt{B_{2,n+1}^2 - 4A_{2,n+1}C_{2,n+1}}}{2A_{2,n+1}}.
$$

Because the exact value of q is 1, we choose the root that is closer to 1. We observe numerically that $C_{2,n+1} < 0$ if θ is sufficiently large, making $q^{n+1} = q_+^{n+1}$ the unique positive solution. Finally, $\mathbf{u}^{n+1}$ and p^{n+1} are computed from (4.6) and (4.7), respectively.

Remark 4.2. Another possible way to construct the DRLM scheme of order 2 is based on the Crank-Nicolson method, where the time-discrete version of (4.4) is given by

$$
\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu\Delta\mathbf{u}^{n+1/2} + q^{n+1/2}(\tilde{\mathbf{u}}^{n+1/2}\cdot\nabla)\tilde{\mathbf{u}}^{n+1/2} + \nabla p^{n+1/2} = \mathbf{f}^{n+1/2}, \tag{4.16a}
$$

$$
\nabla\cdot\mathbf{u}^{n+1/2} = 0, \tag{4.16b}
$$

$$
\frac{\mathcal{K}(\mathbf{u}^{n+1}) - \mathcal{K}(\mathbf{u}^n)}{\Delta t} + \theta\frac{(q^{n+1})^2 - (q^n)^2}{\Delta t} = \left(\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + q^{n+1/2}(\tilde{\mathbf{u}}^{n+1/2}\cdot\nabla)\tilde{\mathbf{u}}^{n+1/2}, \mathbf{u}^{n+1/2}\right), \tag{4.16c}
$$

where $\mathbf{f}^{n+1/2} = \mathbf{f}(t_{n+1/2})$, $\tilde{\mathbf{u}}^{n+1/2} = \frac{1}{2}(3\mathbf{u}^n - \mathbf{u}^{n-1})$, and $\psi^{n+1/2} = \frac{1}{2}(\psi^n + \psi^{n+1})$ for $\psi \in \{\mathbf{u}, p, q\}$. Note that $\tilde{\mathbf{u}}^{1/2}$ can be computed by the following first-order scheme [123]:

$$
\frac{\tilde{\mathbf{u}}^{1/2} - \mathbf{u}^0}{\Delta t/2} - \nu\Delta\tilde{\mathbf{u}}^{1/2} + (\mathbf{u}^0\cdot\nabla)\mathbf{u}^0 + \nabla\tilde{p}^{1/2} = \mathbf{f}^{1/2}, \quad \nabla\cdot\tilde{\mathbf{u}}^{1/2} = 0,
$$

which has a local truncation error of $\mathcal{O}(\Delta t^2)$. We remark that all the terms involving the kinetic energy in (4.16c) disappear after simplification. Indeed, by the definition of $\mathcal{K}(\mathbf{u})$ we have:

$$\frac{\mathcal{K}(\mathbf{u}^{n+1})-\mathcal{K}(\mathbf{u}^{n})}{\Delta t}=\frac{1}{2\Delta t}\left[(\mathbf{u}^{n+1},\mathbf{u}^{n+1})-(\mathbf{u}^{n},\mathbf{u}^{n})\right]=\left(\frac{\mathbf{u}^{n+1}-\mathbf{u}^{n}}{\Delta t},\mathbf{u}^{n+1/2}\right).$$

Therefore, (4.16c) is reduced to $\theta\frac{(q^{n+1})^2-(q^n)^2}{\Delta t} = \left(q^{n+1/2}(\tilde{\mathbf{u}}^{n+1/2}\cdot\nabla)\tilde{\mathbf{u}}^{n+1/2},\mathbf{u}^{n+1/2}\right)$, or equivalently (note that $q^{n+1}+q^n=2q^{n+1/2}\neq 0$),

$$\theta\frac{q^{n+1}-q^{n}}{\Delta t}=\frac{1}{2}\left((\tilde{\mathbf{u}}^{n+1/2}\cdot\nabla)\tilde{\mathbf{u}}^{n+1/2},\mathbf{u}^{n+1/2}\right). \tag{4.17}$$

From this equation q^{n+1} is uniquely determined. Note that (4.17) is equivalent to the Crank-Nicolson discretization of equation $\theta\frac{dq}{dt}=\frac{1}{2}((\mathbf{u}\cdot\nabla)\mathbf{u},\mathbf{u})$.

Remark 4.3. The proposed DRLM schemes involve solving two systems of generalized Stokes equations at each time step. Alternatively, one can replace the Stokes solver with two Poisson solvers via the pressure-correction approach; we refer to [71, 72, 124, 126, 197] for a detailed discussion of this approach in the context of the NS equations, and [27, 127, 182] for other incompressible fluid models. To that end, a pressure-correction variant of the first-order DRLM scheme (4.5) is given by

$$\frac{\overline{\mathbf{u}}^{n+1}-\mathbf{u}^{n}}{\Delta t}-\nu\Delta\overline{\mathbf{u}}^{n+1}+q^{n+1}(\mathbf{u}^{n}\cdot\nabla)\mathbf{u}^{n}+\nabla p^{n}=\mathbf{f}^{n+1}, \tag{4.18a}$$

$$\frac{\mathbf{u}^{n+1}-\overline{\mathbf{u}}^{n+1}}{\Delta t}+\nabla(p^{n+1}-p^{n})=\mathbf{0}, \tag{4.18b}$$

$$\nabla\cdot\mathbf{u}^{n+1}=0, \tag{4.18c}$$

$$\frac{\mathcal{K}(\mathbf{u}^{n+1})-\mathcal{K}(\mathbf{u}^{n})}{\Delta t}+\theta\frac{(q^{n+1})^2-(q^{n})^2}{\Delta t}=\left(\frac{\mathbf{u}^{n+1}-\mathbf{u}^{n}}{\Delta t}+q^{n+1}(\mathbf{u}^{n}\cdot\nabla)\mathbf{u}^{n},\overline{\mathbf{u}}^{n+1}\right), \tag{4.18d}$$

and a rotational pressure-correction variant of the second-order DRLM scheme (4.12) by

$$\frac{3\overline{\mathbf{u}}^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} - \nu\Delta\overline{\mathbf{u}}^{n+1} + q^{n+1}(\tilde{\mathbf{u}}^{n+1}\cdot\nabla)\tilde{\mathbf{u}}^{n+1} + \nabla p^n = \mathbf{f}^{n+1}, \tag{4.19a}$$

$$\frac{3\mathbf{u}^{n+1} - 3\overline{\mathbf{u}}^{n+1}}{2\Delta t} + \nabla(p^{n+1} - p^n + \nu\nabla\cdot\overline{\mathbf{u}}^{n+1}) = \mathbf{0}, \tag{4.19b}$$

$$\nabla\cdot\mathbf{u}^{n+1} = 0, \tag{4.19c}$$

$$\begin{aligned}\frac{3\mathcal{K}(\mathbf{u}^{n+1}) - 4\mathcal{K}(\mathbf{u}^n) + \mathcal{K}(\mathbf{u}^{n-1})}{2\Delta t} &+ \theta\frac{3(q^{n+1})^2 - 4(q^n)^2 + (q^{n-1})^2}{2\Delta t} \\ &= \left(\frac{3\mathbf{u}^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} + q^{n+1}(\tilde{\mathbf{u}}^{n+1}\cdot\nabla)\tilde{\mathbf{u}}^{n+1}, \overline{\mathbf{u}}^{n+1}\right),\end{aligned} \tag{4.19d}$$

where $\overline{\mathbf{u}}^{n+1}$ and $\mathbf{u}^{n+1}$ in (4.18) and (4.19) satisfy the periodic boundary conditions or $\overline{\mathbf{u}}^{n+1}|_{\partial\Omega} = 0$ and $\mathbf{u}^{n+1}\cdot\mathbf{n}|_{\partial\Omega} = 0$. Note that we have used the rotational form in (4.19) since the standard second-order pressure-correction approach may limit the accuracy of the scheme due to certain artificial boundary conditions, as discussed in [71, 72]. The stability of the pressure-correction DRLM schemes (4.18) and (4.19) will be discussed in Remark 4.8.

Time-discrete energy stability In this subsection, we aim to establish energy stability of the proposed first- and second-order DRLM schemes.

Theorem 4.4. In the absence of the external force $\mathbf{f}$, the first-order DRLM scheme (4.5) is unconditionally stable in the sense that

$$\mathcal{K}(\mathbf{u}^{n+1}) + \theta[(q^{n+1})^2 - 1] \le \mathcal{K}(\mathbf{u}^n) + \theta[(q^n)^2 - 1], \quad n = 0, 1, \cdots, N_t - 1.$$

Proof. Taking the L^2 inner product on both sides of (4.5a) with $\mathbf{u}^{n+1}$, we have

$$\left(\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + q^{n+1}(\mathbf{u}^n\cdot\nabla)\mathbf{u}^n, \mathbf{u}^{n+1}\right) = (\nu\Delta\mathbf{u}^{n+1} - \nabla p^{n+1}, \mathbf{u}^{n+1}) = -\nu\|\nabla\mathbf{u}^{n+1}\|_0^2, \tag{4.20}$$

where we used the fact that $(\nabla p^{n+1}, \mathbf{u}^{n+1}) = 0$ due to the divergence-free condition (4.5b), boundary conditions of $\mathbf{u}^{n+1}$, and integration by parts. The combination of (4.5c) and (4.20) yields

$$\frac{\mathcal{K}(\mathbf{u}^{n+1})-\mathcal{K}(\mathbf{u}^n)}{\Delta t}+\theta\frac{(q^{n+1})^2-(q^n)^2}{\Delta t}=-\nu\|\nabla\mathbf{u}^{n+1}\|_0^2\le 0.$$

This implies that

$$\mathcal{K}(\mathbf{u}^{n+1})+\theta(q^{n+1})^2\le\mathcal{K}(\mathbf{u}^n)+\theta(q^n)^2,$$

which completes the proof of Theorem 4.4. □

Remark 4.5. We have from Theorem 4.4 that

$$0\le\mathcal{K}(\mathbf{u}^n)+\theta(q^n)^2\le\mathcal{K}(\mathbf{u}^{n-1})+\theta(q^{n-1})^2\le\ldots\le\mathcal{K}(\mathbf{u}^0)+\theta(q^0)^2=\mathcal{K}(\mathbf{u}^0)+\theta.$$

Thus, the discrete energy $\{\mathcal{K}(\mathbf{u}^n)\}_{n=0}^{N_t}$ is uniformly bounded for the first-order scheme (4.5).

Theorem 4.6. In the absence of the external force $\mathbf{f}$, the second-order DRLM scheme (4.12) is unconditionally stable in the sense that

$$\begin{aligned}\frac{3}{2}\mathcal{K}(\mathbf{u}^{n+1})-\frac{1}{2}\mathcal{K}(\mathbf{u}^n)+\theta\left[\frac{3}{2}(q^{n+1})^2-\frac{1}{2}(q^n)^2-1\right]&\\ \le\frac{3}{2}\mathcal{K}(\mathbf{u}^n)-\frac{1}{2}\mathcal{K}(\mathbf{u}^{n-1})+\theta\left[\frac{3}{2}(q^n)^2-\frac{1}{2}(q^{n-1})^2-1\right],&\quad n=1,2,\cdots,N_t-1.\end{aligned}\tag{4.21}$$

Proof. Multiplying both sides of equation (4.12a) with $\mathbf{u}^{n+1}$ and using (4.12b), we find that

$$\begin{aligned}\left(\frac{3\mathbf{u}^{n+1}-4\mathbf{u}^n+\mathbf{u}^{n-1}}{2\Delta t}+q^{n+1}(\tilde{\mathbf{u}}^{n+1}\cdot\nabla)\tilde{\mathbf{u}}^{n+1},\mathbf{u}^{n+1}\right)&=(\nu\Delta\mathbf{u}^{n+1}-\nabla p^{n+1},\mathbf{u}^{n+1})\\&=-\nu\|\nabla\mathbf{u}^{n+1}\|_0^2\le 0.\end{aligned}$$

This, together with (4.12c), gives us

$$\frac{3\mathcal{K}(\mathbf{u}^{n+1})-4\mathcal{K}(\mathbf{u}^n)+\mathcal{K}(\mathbf{u}^{n-1})}{2\Delta t}+\theta\frac{3(q^{n+1})^2-4(q^n)^2+(q^{n-1})^2}{2\Delta t}\le 0,$$

or equivalently,

$$3\mathcal{K}(\mathbf{u}^{n+1})-\mathcal{K}(\mathbf{u}^n)+\theta\left[3(q^{n+1})^2-(q^n)^2\right]\le 3\mathcal{K}(\mathbf{u}^n)-\mathcal{K}(\mathbf{u}^{n-1})+\theta\left[3(q^n)^2-(q^{n-1})^2\right].$$

Thus, we obtain the desired estimate. $\square$

Remark 4.7. (i) We note that (4.21) is equivalent to the following estimate

$$\mathcal{K}(\tilde{\mathbf{u}}^{n+3/2}) + \theta[(\tilde{q}^{n+3/2})^2 - 1] \leq \mathcal{K}(\tilde{\mathbf{u}}^{n+1/2}) + \theta[(\tilde{q}^{n+1/2})^2 - 1] + \varepsilon^{n+1},$$

where $\tilde{\mathbf{u}}^{n+1/2} = \frac{1}{2}(3\mathbf{u}^n - \mathbf{u}^{n-1})$ and $\tilde{q}^{n+1/2} = \frac{1}{2}(3q^n - q^{n-1})$ for $n \geq 1$ are second-order approximations of $\mathbf{u}(t_{n+1/2})$ and $q(t_{n+1/2})$, respectively, and

$$\varepsilon^{n+1} = \frac{3}{4}\left[\mathcal{K}(\mathbf{u}^{n+1} - \mathbf{u}^n) - \mathcal{K}(\mathbf{u}^n - \mathbf{u}^{n-1}) + \theta(q^{n+1} - q^n)^2 - \theta(q^n - q^{n-1})^2\right].$$

Therefore, if $\|\mathbf{u}^{n+1} - \mathbf{u}^n\|_0 \leq c\Delta t$ and $|q^{n+1} - q^n| \leq c\Delta t$ for all $n \geq 0$, then $\varepsilon^{n+1} = \mathcal{O}(\Delta t^2)$.

(ii) It follows from Theorem 4.6 that

$$\frac{3}{2}\mathcal{K}(\mathbf{u}^{n+1}) + \frac{3}{2}\theta(q^{n+1})^2 \leq \frac{1}{2}\mathcal{K}(\mathbf{u}^n) + \frac{1}{2}\theta(q^n)^2 + C, \tag{4.22}$$

where $C = \frac{3}{2}\mathcal{K}(\mathbf{u}^1) - \frac{1}{2}\mathcal{K}(\mathbf{u}^0) + \theta\left[\frac{3}{2}(q^1)^2 - \frac{1}{2}(q^0)^2\right]$. By repeatedly applying (4.22), we deduce that

$$\mathcal{K}(\mathbf{u}^n) + \theta(q^n)^2 \leq \frac{1}{3^n}\left(\mathcal{K}(\mathbf{u}^0) + \theta(q^0)^2 - C\right) + C,$$

which indicates the boundedness of the energy functional for the second-order scheme (4.12).

(iii) By using similar arguments as in the proofs of Theorems 4.4 and 4.6, we can show that the second-order DRLM scheme based on the Crank-Nicolson method (4.16) is unconditionally stable in the sense that

$$\mathcal{K}(\mathbf{u}^{n+1}) + \theta[(q^{n+1})^2 - 1] \leq \mathcal{K}(\mathbf{u}^n) + \theta[(q^n)^2 - 1], \quad n = 0, 1, \cdots, N_t - 1.$$

Note that the energy stability result is improved with the Crank-Nicolson scheme, compared to the BDF2 method (cf. (4.21)). However, for dissipative systems, the second-order BDF method usually has better numerical performance than the Crank-Nicolson counterpart [25]. Therefore, in the following sections, we focus on the DRLM schemes based on the first- and second-order BDF methods.

Remark 4.8. For the pressure-correction versions of the proposed DRLM schemes, the following energy stability results hold (assuming $\mathbf{f} = \mathbf{0}$):

$$\mathcal{K}(\mathbf{u}^{n+1}) + \frac{\Delta t^2}{2}\|\nabla p^{n+1}\|_0^2 + \theta[(q^{n+1})^2 - 1] \leq \mathcal{K}(\mathbf{u}^n) + \frac{\Delta t^2}{2}\|\nabla p^n\|_0^2 + \theta[(q^n)^2 - 1], \quad (4.23)$$

for the first-order pressure-correction DRLM scheme (4.18), and

$$\begin{aligned}
&\frac{3}{2}\mathcal{K}(\mathbf{u}^{n+1}) - \frac{1}{2}\mathcal{K}(\mathbf{u}^n) + \frac{\Delta t^2}{3}\|\nabla(p^{n+1} + \nu s^{n+1})\|_0^2 + \frac{\Delta t\nu}{2}\|s^{n+1}\|_0^2 + \theta\left[\frac{3}{2}(q^{n+1})^2 - \frac{1}{2}(q^n)^2 - 1\right] \\
&\quad \leq \frac{3}{2}\mathcal{K}(\mathbf{u}^n) - \frac{1}{2}\mathcal{K}(\mathbf{u}^{n-1}) + \frac{\Delta t^2}{3}\|\nabla(p^n + \nu s^n)\|_0^2 + \frac{\Delta t\nu}{2}\|s^n\|_0^2 + \theta\left[\frac{3}{2}(q^n)^2 - \frac{1}{2}(q^{n-1})^2 - 1\right],
\end{aligned} \quad (4.24)$$

for the second-order rotational pressure-correction DRLM scheme (4.19), where $s^n := \nabla \cdot \sum_{i=1}^{n} \overline{\mathbf{u}}^i$ for $1 \leq n \leq N_t$. Detailed proofs of these estimates are provided in Appendix A and Appendix B, respectively.

4.1.2 Fully discrete DRLM schemes

In this section, we apply spatial discretization for the time-discrete problems (4.5) and (4.12) to obtain fully discrete DRLM schemes. While the proposed DRLM approach can be coupled with finite difference, finite element, or finite volume methods, we adopt the finite difference approximation on a MAC staggered grid for simplicity. The corresponding energy stability of the numerical solutions is then rigorously established. The proposed schemes require solving two generalized Stokes systems and a quadratic equation at each time step, with the former being the main computational cost, especially for large Reynolds numbers. Therefore, we discuss preconditioned iterative solvers for Stokes systems at the end of the section.

The MAC scheme for spatial discretization To simplify the presentation, we consider the 2D NS equations (4.1) with the spatial domain $\Omega = (0,1)^2$, the external force $\mathbf{f} = [f_1, f_2]^T$, and the velocity vector field $\mathbf{u} = [u, v]^T$. The extension to the 3D case can be done in a similar manner. Suppose that Ω is partitioned into $N_x N_y$ rectangles of the form $(x_{i-1}, x_i) \times (y_{j-1}, y_j)$

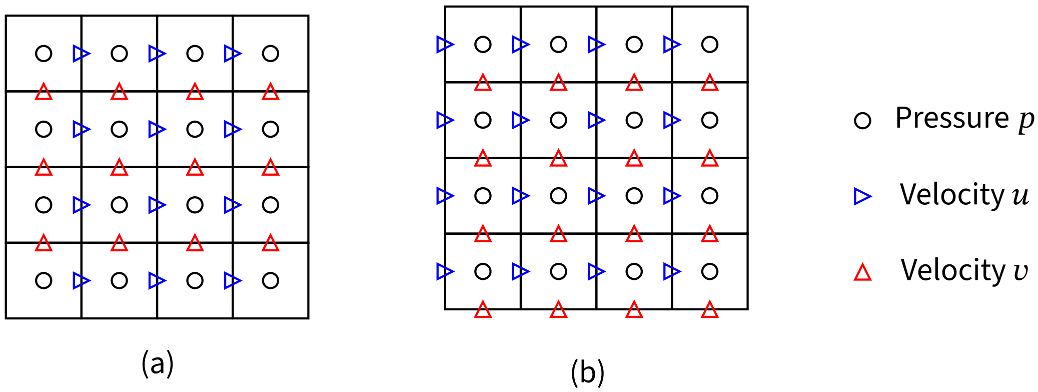


Figure 4.1: Staggered grid ($N_x = N_y = 4$) for spatial discretization with positions of unknowns u, v, and p. (a) Homogeneous Dirichlet boundary conditions. (b) Periodic boundary conditions.

for $1 \leq i \leq N_x$ and $1 \leq j \leq N_y$, where

$$0 = x_0 < x_1 < \ldots < x_{N_x} = 1, \quad \text{and} \quad 0 = y_0 < y_1 < \ldots < y_{N_y} = 1.$$

Let $h_x = \frac{1}{N_x}$ and $h_y = \frac{1}{N_y}$ be the uniform mesh sizes in the x- and y-directions, respectively. For $\psi \in \{u, v, p\}$, let $\Omega_{h,\psi}$ represent the collection of discrete points in Ω associated with ψ (see Figure 4.1); in particular, $\Omega_{h,p}$ corresponds to black circle points, $\Omega_{h,u}$ to blue right-pointing triangles, and $\Omega_{h,v}$ to red upward-pointing triangles. Denote by $\mathbf{U}^n, \mathbf{V}^n$, and $\mathbf{P}^n$ $(0 \leq n \leq N_t)$ the approximations of u^n, v^n, and p^n over $\Omega_{h,u}, \Omega_{h,v}$, and $\Omega_{h,p}$, respectively. For any two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size, we define

$$(\mathbf{A}, \mathbf{B})_{l^2} := h_x h_y \mathrm{Tr}(\mathbf{A}^T \mathbf{B}), \qquad \|\mathbf{A}\|_{l^2} := \sqrt{(\mathbf{A}, \mathbf{A})_{l^2}},$$

as the discrete l^2 inner product and corresponding l^2 norm.

Next, we introduce differentiation matrices for Δu, Δv, and ∇p, utilizing the following matrices for any positive integer N and real numbers h and a:

$$\mathbf{K}^{\mathsf{Dir}}_{N,h,a} = \frac{1}{h^2}\begin{bmatrix} -a & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & -2 & 1 \\ 0 & \dots & 0 & 1 & -a \end{bmatrix}_{N\times N} \qquad \mathbf{M}^{\mathsf{Dir}}_{N,h} = \frac{1}{h}\begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix}_{(N-1)\times N},$$

$$\mathbf{K}^{\mathsf{per}}_{N,h} = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & \dots & 1 \\ 1 & -2 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & -2 & 1 \\ 1 & \dots & 0 & 1 & -2 \end{bmatrix}_{N\times N} \qquad \mathbf{M}^{\mathsf{per}}_{N,h} = \frac{1}{h}\begin{bmatrix} 1 & 0 & 0 & \dots & -1 \\ -1 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & -1 & 1 & 0 \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix}_{N\times N}.$$

For homogeneous Dirichlet boundary conditions, we define the matrices $\mathbf{A}_u, \mathbf{B}_u, \mathbf{A}_v, \mathbf{B}_v, \mathbf{A}_p$, and $\mathbf{B}_p$ as follows:

$$\mathbf{A}_u = \mathbf{K}^{\mathsf{Dir}}_{N_x-1,h_x,2}, \qquad \mathbf{A}_v = \mathbf{K}^{\mathsf{Dir}}_{N_x,h_x,3}, \qquad \mathbf{A}_p = \mathbf{M}^{\mathsf{Dir}}_{N_x,h_x}, \tag{4.25}$$

$$\mathbf{B}_u = \mathbf{K}^{\mathsf{Dir}}_{N_y,h_y,3}, \qquad \mathbf{B}_v = \mathbf{K}^{\mathsf{Dir}}_{N_y-1,h_y,2}, \qquad \mathbf{B}_p = \mathbf{M}^{\mathsf{Dir}}_{N_y,h_y}. \tag{4.26}$$

For periodic boundary conditions, we set

$$\mathbf{A}_u = \mathbf{A}_v = \mathbf{K}^{\mathsf{per}}_{N_x,h_x}, \qquad \mathbf{A}_p = \mathbf{M}^{\mathsf{per}}_{N_x,h_x}, \tag{4.27}$$

$$\mathbf{B}_u = \mathbf{B}_v = \mathbf{K}^{\mathsf{per}}_{N_y,h_y}, \qquad \mathbf{B}_p = \mathbf{M}^{\mathsf{per}}_{N_y,h_y}. \tag{4.28}$$

Applying the central finite difference method on the staggered grid (cf. Figure 4.1) to system (4.5) results in the fully discrete first-order DRLM scheme:

$$\frac{\mathbf{U}^{n+1}-\mathbf{U}^n}{\Delta t} - \nu\left(\mathbf{A}_u\mathbf{U}^{n+1} + \mathbf{U}^{n+1}\mathbf{B}_u^T\right) + q^{n+1}\mathbf{F}_1(\mathbf{U}^n,\mathbf{V}^n) + \mathbf{A}_p\mathbf{P}^{n+1} = \mathcal{F}_1^{n+1}, \tag{4.29a}$$

$$\frac{\mathbf{V}^{n+1}-\mathbf{V}^n}{\Delta t} - \nu\left(\mathbf{A}_v\mathbf{V}^{n+1} + \mathbf{V}^{n+1}\mathbf{B}_v^T\right) + q^{n+1}\mathbf{F}_2(\mathbf{U}^n,\mathbf{V}^n) + \mathbf{P}^{n+1}\mathbf{B}_p^T = \mathcal{F}_2^{n+1}, \tag{4.29b}$$

$$(-\mathbf{A}_p^T)\mathbf{U}^{n+1} + \mathbf{V}^{n+1}(-\mathbf{B}_p) = \mathbf{O}, \tag{4.29c}$$

$$\begin{aligned}&\frac{1}{2\Delta t}\left(\|\mathbf{U}^{n+1}\|_{l^2}^2 + \|\mathbf{V}^{n+1}\|_{l^2}^2 - \|\mathbf{U}^n\|_{l^2}^2 - \|\mathbf{V}^n\|_{l^2}^2\right) + \theta\frac{(q^{n+1})^2-(q^n)^2}{\Delta t}\\ &= \left(\frac{\mathbf{U}^{n+1}-\mathbf{U}^n}{\Delta t} + q^{n+1}\mathbf{F}_1(\mathbf{U}^n,\mathbf{V}^n), \mathbf{U}^{n+1}\right)_{l^2} + \left(\frac{\mathbf{V}^{n+1}-\mathbf{V}^n}{\Delta t} + q^{n+1}\mathbf{F}_2(\mathbf{U}^n,\mathbf{V}^n), \mathbf{V}^{n+1}\right)_{l^2}.\end{aligned} \tag{4.29d}$$

Here, $\mathcal{F}_1$ and $\mathcal{F}_2$ represent the discrete evaluations of f_1 and f_2 over $\Omega_{h,u}$ and $\Omega_{h,v}$, respectively. Similarly, $\mathbf{F}_1(\mathbf{U},\mathbf{V})$ and $\mathbf{F}_2(\mathbf{U},\mathbf{V})$ approximate the nonlinear convection terms $uu_x + vu_y$ and $uv_x + vv_y$ on the grids $\Omega_{h,u}$ and $\Omega_{h,v}$, respectively.

For the second-order case (4.12), we define $\tilde{\mathbf{U}}^{n+1} = 2\mathbf{U}^n - \mathbf{U}^{n-1}$ and $\tilde{\mathbf{V}}^{n+1} = 2\mathbf{V}^n - \mathbf{V}^{n-1}$ for $n \geq 1$. The fully discrete second-order DRLM scheme is given by:

$$\frac{3\mathbf{U}^{n+1}-4\mathbf{U}^n+\mathbf{U}^{n-1}}{2\Delta t} - \nu\left(\mathbf{A}_u\mathbf{U}^{n+1} + \mathbf{U}^{n+1}\mathbf{B}_u^T\right) + q^{n+1}\mathbf{F}_1(\tilde{\mathbf{U}}^{n+1},\tilde{\mathbf{V}}^{n+1}) + \mathbf{A}_p\mathbf{P}^{n+1} = \mathcal{F}_1^{n+1}, \tag{4.30a}$$

$$\frac{3\mathbf{V}^{n+1}-4\mathbf{V}^n+\mathbf{V}^{n-1}}{2\Delta t} - \nu\left(\mathbf{A}_v\mathbf{V}^{n+1} + \mathbf{V}^{n+1}\mathbf{B}_v^T\right) + q^{n+1}\mathbf{F}_2(\tilde{\mathbf{U}}^{n+1},\tilde{\mathbf{V}}^{n+1}) + \mathbf{P}^{n+1}\mathbf{B}_p^T = \mathcal{F}_2^{n+1}, \tag{4.30b}$$

$$(-\mathbf{A}_p^T)\mathbf{U}^{n+1} + \mathbf{V}^{n+1}(-\mathbf{B}_p) = \mathbf{O}, \tag{4.30c}$$

$$\begin{aligned}&\frac{1}{4\Delta t}\left(3\|\mathbf{U}^{n+1}\|_{l^2}^2 + 3\|\mathbf{V}^{n+1}\|_{l^2}^2 - 4\|\mathbf{U}^n\|_{l^2}^2 - 4\|\mathbf{V}^n\|_{l^2}^2 + \|\mathbf{U}^{n-1}\|_{l^2}^2 + \|\mathbf{V}^{n-1}\|_{l^2}^2\right)\\ &+\theta\frac{3(q^{n+1})^2-4(q^n)^2+(q^{n-1})^2}{2\Delta t} = \left(\frac{3\mathbf{U}^{n+1}-4\mathbf{U}^n+\mathbf{U}^{n-1}}{2\Delta t} + q^{n+1}\mathbf{F}_1(\tilde{\mathbf{U}}^{n+1},\tilde{\mathbf{V}}^{n+1}), \mathbf{U}^{n+1}\right)_{l^2}\\ &\qquad + \left(\frac{3\mathbf{V}^{n+1}-4\mathbf{V}^n+\mathbf{V}^{n-1}}{2\Delta t} + q^{n+1}\mathbf{F}_2(\tilde{\mathbf{U}}^{n+1},\tilde{\mathbf{V}}^{n+1}), \mathbf{V}^{n+1}\right)_{l^2}.\end{aligned} \tag{4.30d}$$

The implementation of the fully discrete DRLM schemes (4.29) and (4.30) can be carried out in the same manner as in the semi-discrete case and is thus omitted.

Fully discrete energy stability In this subsection, we show that the fully discrete DRLM schemes (4.29) and (4.30) are also unconditionally energy stable.

Theorem 4.9. In the absence of the external force $\mathbf{f}$, the fully discrete first-order DRLM scheme

(4.29) is unconditionally stable in the sense that, for $n = 0, 1, \cdots, N_t - 1$,

$$\frac{1}{2}\left(\|\mathbf{U}^{n+1}\|_{l^2}^2 + \|\mathbf{V}^{n+1}\|_{l^2}^2\right) + \theta\left[(q^{n+1})^2 - 1\right] \le \frac{1}{2}\left(\|\mathbf{U}^{n}\|_{l^2}^2 + \|\mathbf{V}^{n}\|_{l^2}^2\right) + \theta\left[(q^{n})^2 - 1\right].$$

Proof. Since $\mathbf{f} = [0, 0]^T$, both $\mathcal{F}_1^{n+1}$ and $\mathcal{F}_2^{n+1}$ ($n \ge 0$) are zero matrices. Taking the discrete l^2 inner product of (4.29a) and (4.29b) with $\mathbf{U}^{n+1}$ and $\mathbf{V}^{n+1}$, respectively, we obtain

$$\begin{aligned}&\left(\frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} + q^{n+1}\mathbf{F}_1(\mathbf{U}^n, \mathbf{V}^n), \mathbf{U}^{n+1}\right)_{l^2} \\ &\qquad = \nu\left(\mathbf{A}_u\mathbf{U}^{n+1} + \mathbf{U}^{n+1}\mathbf{B}_u^T, \mathbf{U}^{n+1}\right)_{l^2} - (\mathbf{A}_p\mathbf{P}^{n+1}, \mathbf{U}^{n+1})_{l^2}, \qquad (4.31)\\ &\left(\frac{\mathbf{V}^{n+1} - \mathbf{V}^n}{\Delta t} + q^{n+1}\mathbf{F}_2(\mathbf{U}^n, \mathbf{V}^n), \mathbf{V}^{n+1}\right)_{l^2} \\ &\qquad = \nu\left(\mathbf{A}_v\mathbf{V}^{n+1} + \mathbf{V}^{n+1}\mathbf{B}_v^T, \mathbf{V}^{n+1}\right)_{l^2} - (\mathbf{P}^{n+1}\mathbf{B}_p^T, \mathbf{V}^{n+1})_{l^2}. \qquad (4.32)\end{aligned}$$

For any matrices $\mathbf{A}, \mathbf{B}, \mathbf{X}$, and $\mathbf{Y}$, we have

$$(\mathbf{AX}, \mathbf{Y})_{l^2} = h_x h_y \mathrm{Tr}\left((\mathbf{AX})^T\mathbf{Y}\right) = h_x h_y \mathrm{Tr}\left(\mathbf{X}^T\mathbf{A}^T\mathbf{Y}\right) = (\mathbf{X}, \mathbf{A}^T\mathbf{Y})_{l^2}, \qquad (4.33)$$

$$(\mathbf{XB}^T, \mathbf{Y})_{l^2} = h_x h_y \mathrm{Tr}\left((\mathbf{XB}^T)^T\mathbf{Y}\right) = h_x h_y \mathrm{Tr}\left(\mathbf{BX}^T\mathbf{Y}\right) = h_x h_y \mathrm{Tr}\left(\mathbf{X}^T\mathbf{YB}\right) = (\mathbf{X}, \mathbf{YB})_{l^2}, \qquad (4.34)$$

provided that the above matrix products are well-defined. Using (4.33)-(4.34) and the fact that $\mathbf{A}_p^T\mathbf{U}^{n+1} + \mathbf{V}^{n+1}\mathbf{B}_p = \mathbf{O}$ obtained from (4.29c), yields

$$\begin{aligned}(\mathbf{A}_p\mathbf{P}^{n+1}, \mathbf{U}^{n+1})_{l^2} + (\mathbf{P}^{n+1}\mathbf{B}_p^T, \mathbf{V}^{n+1})_{l^2} &= (\mathbf{P}^{n+1}, \mathbf{A}_p^T\mathbf{U}^{n+1})_{l^2} + (\mathbf{P}^{n+1}, \mathbf{V}^{n+1}\mathbf{B}_p)_{l^2} \\ &= (\mathbf{P}^{n+1}, \mathbf{A}_p^T\mathbf{U}^{n+1} + \mathbf{V}^{n+1}\mathbf{B}_p)_{l^2} = 0. \qquad (4.35)\end{aligned}$$

It follows from the definitions of $\mathbf{A}_u$, $\mathbf{B}_v$, $\mathbf{A}_p$, and $\mathbf{B}_p$ (cf. (4.25)-(4.28)) that $\mathbf{A}_u = -\mathbf{A}_p\mathbf{A}_p^T$ and

$\mathbf{B}_v = -\mathbf{B}_p\mathbf{B}_p^T$. Consequently,

$$\begin{aligned}\left(\mathbf{A}_u\mathbf{U}^{n+1}, \mathbf{U}^{n+1}\right)_{l2} &= -\left(\mathbf{A}_p\mathbf{A}_p^T\mathbf{U}^{n+1}, \mathbf{U}^{n+1}\right)_{l2} \\ &= -\left(\mathbf{A}_p^T\mathbf{U}^{n+1}, \mathbf{A}_p^T\mathbf{U}^{n+1}\right)_{l2} = -\left\|\mathbf{A}_p^T\mathbf{U}^{n+1}\right\|_{l2}^2 \le 0, \end{aligned} \tag{4.36}$$

$$\begin{aligned}\left(\mathbf{V}^{n+1}\mathbf{B}_v^T, \mathbf{V}^{n+1}\right)_{l2} &= -\left(\mathbf{V}^{n+1}\mathbf{B}_p\mathbf{B}_p^T, \mathbf{V}^{n+1}\right)_{l2} \\ &= -\left(\mathbf{V}^{n+1}\mathbf{B}_p, \mathbf{V}^{n+1}\mathbf{B}_p\right)_{l2} = -\left\|\mathbf{V}^{n+1}\mathbf{B}_p\right\|_{l2}^2 \le 0, \end{aligned} \tag{4.37}$$

where we have applied identities (4.33) and (4.34) to derive the above equalities. On the other hand, both $-\mathbf{B}_u$ and $-\mathbf{A}_v$ are symmetric positive definite matrices, they admit Cholesky decompositions $-\mathbf{B}_u = \mathbf{R}_u\mathbf{R}_u^T$ and $-\mathbf{A}_v = \mathbf{R}_v\mathbf{R}_v^T$. By using the same arguments as in (4.36) and (4.37), we obtain

$$\left(\mathbf{U}^{n+1}\mathbf{B}_u^T, \mathbf{U}^{n+1}\right)_{l2} = -\left\|\mathbf{U}^{n+1}\mathbf{R}_u\right\|_{l2}^2 \le 0, \tag{4.38}$$

$$\left(\mathbf{A}_v\mathbf{V}^{n+1}, \mathbf{V}^{n+1}\right)_{l2} = -\left\|\mathbf{R}_v^T\mathbf{V}^{n+1}\right\|_{l2}^2 \le 0. \tag{4.39}$$

Combining (4.31)-(4.32) and (4.35)-(4.39) gives us

$$\left(\frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} + q^{n+1}\mathbf{F}_1(\mathbf{U}^n, \mathbf{V}^n), \mathbf{U}^{n+1}\right)_{l2} + \left(\frac{\mathbf{V}^{n+1} - \mathbf{V}^n}{\Delta t} + q^{n+1}\mathbf{F}_2(\mathbf{U}^n, \mathbf{V}^n), \mathbf{V}^{n+1}\right)_{l2} \le 0.$$

This, together with (4.29d), leads to

$$\frac{1}{2\Delta t}\left(\|\mathbf{U}^{n+1}\|_{l2}^2 + \|\mathbf{V}^{n+1}\|_{l2}^2 - \|\mathbf{U}^n\|_{l2}^2 - \|\mathbf{V}^n\|_{l2}^2\right) + \theta\frac{(q^{n+1})^2 - (q^n)^2}{\Delta t} \le 0,$$

or equivalently,

$$\frac{1}{2}\left(\|\mathbf{U}^{n+1}\|_{l2}^2 + \|\mathbf{V}^{n+1}\|_{l2}^2\right) + \theta(q^{n+1})^2 \le \frac{1}{2}\left(\|\mathbf{U}^n\|_{l2}^2 + \|\mathbf{V}^n\|_{l2}^2\right) + \theta(q^n)^2,$$

which completes the proof of Theorem 4.9. □

Theorem 4.10. In the absence of the external force $\mathbf{f}$, the fully discrete second-order DRLM scheme (4.30) is unconditionally stable in the sense that, for $n = 1, 2, \cdots, N_t - 1$,

$$\begin{aligned}
&\frac{3}{4}\left(\|\mathbf{U}^{n+1}\|_{l2}^2 + \|\mathbf{V}^{n+1}\|_{l2}^2\right) - \frac{1}{4}\left(\|\mathbf{U}^{n}\|_{l2}^2 + \|\mathbf{V}^{n}\|_{l2}^2\right) + \theta\left[\frac{3}{2}(q^{n+1})^2 - \frac{1}{2}(q^{n})^2 - 1\right] \\
&\quad \leq \frac{3}{4}\left(\|\mathbf{U}^{n}\|_{l2}^2 + \|\mathbf{V}^{n}\|_{l2}^2\right) - \frac{1}{4}\left(\|\mathbf{U}^{n-1}\|_{l2}^2 + \|\mathbf{V}^{n-1}\|_{l2}^2\right) + \theta\left[\frac{3}{2}(q^{n})^2 - \frac{1}{2}(q^{n-1})^2 - 1\right].
\end{aligned}$$

Proof. By taking the discrete l^2 inner product of (4.30a) and (4.30b) with $\mathbf{U}^{n+1}$ and $\mathbf{V}^{n+1}$, respectively, and noting that $\mathcal{F}_1^{n+1}$ and $\mathcal{F}_2^{n+1}$ ($n \geq 0$) are zero matrices, we arrive at

$$\begin{aligned}
&\left(\frac{3\mathbf{U}^{n+1} - 4\mathbf{U}^{n} + \mathbf{U}^{n-1}}{2\Delta t} + q^{n+1}\mathbf{F}_1(\tilde{\mathbf{U}}^{n+1}, \tilde{\mathbf{V}}^{n+1}), \mathbf{U}^{n+1}\right)_{l2} \\
&\qquad = \nu\left(\mathbf{A}_u\mathbf{U}^{n+1} + \mathbf{U}^{n+1}\mathbf{B}_u^T, \mathbf{U}^{n+1}\right)_{l2} - (\mathbf{A}_p\mathbf{P}^{n+1}, \mathbf{U}^{n+1})_{l2}, \\
&\left(\frac{3\mathbf{V}^{n+1} - 4\mathbf{V}^{n} + \mathbf{V}^{n-1}}{2\Delta t} + q^{n+1}\mathbf{F}_2(\tilde{\mathbf{U}}^{n+1}, \tilde{\mathbf{V}}^{n+1}), \mathbf{V}^{n+1}\right)_{l2} \\
&\qquad = \nu\left(\mathbf{A}_v\mathbf{V}^{n+1} + \mathbf{V}^{n+1}\mathbf{B}_v^T, \mathbf{V}^{n+1}\right)_{l2} - (\mathbf{P}^{n+1}\mathbf{B}_p^T, \mathbf{V}^{n+1})_{l2}.
\end{aligned}$$

Based on the proof of Theorem 4.9, we find that

$$\begin{aligned}
&\nu\left(\mathbf{A}_u\mathbf{U}^{n+1} + \mathbf{U}^{n+1}\mathbf{B}_u^T, \mathbf{U}^{n+1}\right)_{l2} - (\mathbf{A}_p\mathbf{P}^{n+1}, \mathbf{U}^{n+1})_{l2} \\
&\quad + \nu\left(\mathbf{A}_v\mathbf{V}^{n+1} + \mathbf{V}^{n+1}\mathbf{B}_v^T, \mathbf{V}^{n+1}\right)_{l2} - (\mathbf{P}^{n+1}\mathbf{B}_p^T, \mathbf{V}^{n+1})_{l2} \leq 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left(\frac{3\mathbf{U}^{n+1} - 4\mathbf{U}^{n} + \mathbf{U}^{n-1}}{2\Delta t} + q^{n+1}\mathbf{F}_1(\tilde{\mathbf{U}}^{n+1}, \tilde{\mathbf{V}}^{n+1}), \mathbf{U}^{n+1}\right)_{l2} \\
&\quad + \left(\frac{3\mathbf{V}^{n+1} - 4\mathbf{V}^{n} + \mathbf{V}^{n-1}}{2\Delta t} + q^{n+1}\mathbf{F}_2(\tilde{\mathbf{U}}^{n+1}, \tilde{\mathbf{V}}^{n+1}), \mathbf{V}^{n+1}\right)_{l2} \leq 0.
\end{aligned} \tag{4.40}$$

The combination of (4.30d) and (4.40) results in

$$\begin{aligned}
&\frac{1}{4\Delta t}\left(3\|\mathbf{U}^{n+1}\|_{l2}^2 + 3\|\mathbf{V}^{n+1}\|_{l2}^2 - 4\|\mathbf{U}^{n}\|_{l2}^2 - 4\|\mathbf{V}^{n}\|_{l2}^2 + \|\mathbf{U}^{n-1}\|_{l2}^2 + \|\mathbf{V}^{n-1}\|_{l2}^2\right) \\
&\qquad + \theta\frac{3(q^{n+1})^2 - 4(q^{n})^2 + (q^{n-1})^2}{2\Delta t} \leq 0.
\end{aligned}$$

This implies that

$$
\begin{aligned}
&\frac{3}{4}\left(\|\mathbf{U}^{n+1}\|_{l^2}^2+\|\mathbf{V}^{n+1}\|_{l^2}^2\right)-\frac{1}{4}\left(\|\mathbf{U}^{n}\|_{l^2}^2+\|\mathbf{V}^{n}\|_{l^2}^2\right)+\theta\left[\frac{3}{2}(q^{n+1})^2-\frac{1}{2}(q^{n})^2\right]\\
&\qquad\le\frac{3}{4}\left(\|\mathbf{U}^{n}\|_{l^2}^2+\|\mathbf{V}^{n}\|_{l^2}^2\right)-\frac{1}{4}\left(\|\mathbf{U}^{n-1}\|_{l^2}^2+\|\mathbf{V}^{n-1}\|_{l^2}^2\right)+\theta\left[\frac{3}{2}(q^{n})^2-\frac{1}{2}(q^{n-1})^2\right],
\end{aligned}
$$

which concludes the proof of Theorem 4.10. □

Remark 4.11. (i) Let $\mathcal{E}_1^n$ and $\mathcal{E}_2^n$ $(n \geq 1)$ be the discrete error functions for q associated with the first- and second-order DRLM schemes:

$$
\mathcal{E}_1^n(q^n,\theta) := \theta\left[(q^n)^2-1\right],\quad \mathcal{E}_2^n(q^n,q^{n-1},\theta) := \theta\left[\frac{3}{2}(q^n)^2-\frac{1}{2}(q^{n-1})^2-1\right].
$$

Under suitable choices of θ, it will be demonstrated in the numerical experiments (cf. Section 4.1.3) that both $\mathcal{E}_1^n$ and $\mathcal{E}_2^n$ are significantly smaller than the main energy functional. Therefore, they can be considered negligible in Theorems 4.4 and 4.6 (for the semi-discrete case) and Theorems 4.9 and 4.10 (for the fully discrete case).

(ii) The energy stability of the proposed DRLM schemes remains valid under certain mixed boundary conditions. One typical example is the 2D Kelvin-Helmholtz instability (cf. Subsection 4.1.3), where periodic conditions are applied along the vertical boundaries while free-slip conditions are imposed at the horizontal boundaries.

(iii) Due to the coupling between the Lagrange multiplier, velocity, and pressure, the fully discrete convergence analysis of the proposed DRLM schemes (4.29) and (4.30) is challenging and remains an open problem. We refer to [123] and [29] for error analysis of the MAC method applied to the NS equations and other coupled physical systems involving fluid motion, such as the Cahn-Hilliard-Hele-Shaw equation.

Preconditioned iterative solvers for generalized Stokes systems The proposed first- and second-order DRLM schemes involve solving the following generalized Stokes system at each time step (cf. (4.9), (4.10), (4.13), and (4.14)):

$$\alpha \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{g}, \tag{4.41a}$$

$$\nabla \cdot \mathbf{u} = 0, \tag{4.41b}$$

where $\mathbf{u} = [u, v]^T$ and p denote the velocity field and pressure, ν is the viscosity coefficient, $\mathbf{g} = [g_1, g_2]^T$ is a given vector function, $\alpha = \frac{1}{\Delta t}$ for the first-order DRLM scheme, and $\alpha = \frac{3}{2\Delta t}$ for the second-order DRLM scheme. After spatial discretization on the MAC staggered grid, system (4.41) becomes

$$\alpha \mathbf{U} - \nu(\mathbf{A}_u \mathbf{U} + \mathbf{U}\mathbf{B}_u^T) + \mathbf{A}_p \mathbf{P} = \mathcal{G}_1, \tag{4.42a}$$

$$\alpha \mathbf{V} - \nu(\mathbf{A}_v \mathbf{V} + \mathbf{V}\mathbf{B}_v^T) + \mathbf{P}\mathbf{B}_p^T = \mathcal{G}_2, \tag{4.42b}$$

$$(-\mathbf{A}_p^T)\mathbf{U} + \mathbf{V}(-\mathbf{B}_p) = \mathbf{O}, \tag{4.42c}$$

where $\mathbf{U}, \mathbf{V}$, and $\mathbf{P}$ are the approximations of u, v, and p over $\Omega_{h,u}, \Omega_{h,v}$, and $\Omega_{h,p}$, respectively; $\mathcal{G}_1$ and $\mathcal{G}_2$ are the values of g_1 and g_2 over $\Omega_{h,u}$ and $\Omega_{h,v}$, respectively. The matrices $\mathbf{A}_u, \mathbf{B}_u, \mathbf{A}_v, \mathbf{B}_v, \mathbf{A}_p$, and $\mathbf{B}_p$ are defined in (4.25)-(4.28).

System (4.42) can be transformed into the following saddle point problem:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathcal{G} \\ \mathbf{O} \end{bmatrix}, \tag{4.43}$$

where $\mathbf{x} = \begin{bmatrix} \text{vec}(\mathbf{U}) \\ \text{vec}(\mathbf{V}) \end{bmatrix}$, $\mathbf{y} = \text{vec}(\mathbf{P})$, and $\mathcal{G} = \begin{bmatrix} \text{vec}(\mathcal{G}_1) \\ \text{vec}(\mathcal{G}_2) \end{bmatrix}$, with vec denoting the vectorization operator. The coefficient matrix $\mathbf{A}$ is defined as

$$\mathbf{A} = \begin{bmatrix} \alpha \mathbf{I}_{(N_x-1)N_y} - \nu(\mathbf{I}_{N_y} \otimes \mathbf{A}_u + \mathbf{B}_u \otimes \mathbf{I}_{N_x-1}) & \mathbf{O} \\ \mathbf{O} & \alpha \mathbf{I}_{N_x(N_y-1)} - \nu(\mathbf{I}_{N_y-1} \otimes \mathbf{A}_v + \mathbf{B}_v \otimes \mathbf{I}_{N_x}) \end{bmatrix},$$

for the homogeneous Dirichlet boundary condition, and

$$\mathbf{A} = \begin{bmatrix} \alpha \mathbf{I}_{N_x N_y} - \nu(\mathbf{I}_{N_y} \otimes \mathbf{A}_u + \mathbf{B}_u \otimes \mathbf{I}_{N_x}) & \mathbf{O} \\ \mathbf{O} & \alpha \mathbf{I}_{N_x N_y} - \nu(\mathbf{I}_{N_y} \otimes \mathbf{A}_v + \mathbf{B}_v \otimes \mathbf{I}_{N_x}) \end{bmatrix},$$

for the periodic boundary condition, where $\otimes$ denotes the Kronecker product. The matrix $\mathbf{B}$ is given by:

$$\mathbf{B} = \begin{bmatrix} \mathbf{I}_{N_y} \otimes \mathbf{A}_p^T & \mathbf{B}_p^T \otimes \mathbf{I}_{N_x} \end{bmatrix}.$$

Since the pressure p is defined up to a constant, $\mathbf{B}^T$ has a one-dimensional kernel. By fixing $\mathbf{P}(1,1) = 0$, we may assume without loss of generality that $\mathbf{B}^T$ has full rank.

Suppose that $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{m \times n}$ (with $m < n$), and that $\mathbf{A}$ is decomposed as $\mathbf{A} = \mathbf{D} - \mathbf{E}$, where $\mathbf{D}$ is the diagonal matrix obtained from $\mathbf{A}$. Let $\hat{\mathbf{S}} = \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^T$ be an approximation of the Schur complement matrix $\mathbf{S} = \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T$. As discussed in [8, 43], we consider the following block preconditioner:

$$\begin{bmatrix} \mathbf{D} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{O} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{B}^T\hat{\mathbf{S}}^{-1}\mathbf{B})\mathbf{D}^{-1} & \mathbf{D}^{-1}\mathbf{B}^T\hat{\mathbf{S}}^{-1} \\ \hat{\mathbf{S}}^{-1}\mathbf{B}\mathbf{D}^{-1} & -\hat{\mathbf{S}}^{-1} \end{bmatrix}. \tag{4.44}$$

Preconditioning (4.43) from the left by (4.44) yields

$$\begin{bmatrix} \mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{B}^T\hat{\mathbf{S}}^{-1}\mathbf{B})\mathbf{D}^{-1}\mathbf{E} & \mathbf{O} \\ -\hat{\mathbf{S}}^{-1}\mathbf{B}\mathbf{D}^{-1}\mathbf{E} & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} (\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{B}^T\hat{\mathbf{S}}^{-1}\mathbf{B})\mathbf{D}^{-1}\mathcal{G} \\ \hat{\mathbf{S}}^{-1}\mathbf{B}\mathbf{D}^{-1}\mathcal{G} \end{bmatrix}. \tag{4.45}$$

Therefore, solving (4.43) or (4.45) is equivalent to performing the following two steps:

(i) Solve for $\mathbf{x}$ the linear system

$$\left[\mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{B}^T\hat{\mathbf{S}}^{-1}\mathbf{B})\mathbf{D}^{-1}\mathbf{E}\right]\mathbf{x} = (\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{B}^T\hat{\mathbf{S}}^{-1}\mathbf{B})\mathbf{D}^{-1}\mathcal{G}. \tag{4.46}$$

(ii) Compute $\mathbf{y}$ from $\mathbf{x}$ by $\mathbf{y} = \hat{\mathbf{S}}^{-1}\mathbf{B}\mathbf{D}^{-1}(\mathbf{E}\mathbf{x} + \mathcal{G})$.

The linear system (4.46) can be solved using various iterative solvers; in the numerical experiments, we employ the GMRES method. It is worth noting that $\hat{\mathbf{S}} \in \mathbb{R}^{m \times m}$ is a sparse symmetric positive definite matrix, allowing efficient computation of $\hat{\mathbf{S}}^{-1}\mathbf{z}$ for any $\mathbf{z} \in \mathbb{R}^m$, such as via the incomplete Cholesky decomposition.

4.1.3 Numerical results

In this section, we carry out several numerical experiments to verify the accuracy and energy stability of the proposed first- and second-order DRLM schemes with difference choices of the regularization parameter and Reynolds number. Various benchmark test cases in 2D and 3D, including lid-driven cavity flow and Kelvin-Helmholtz instability, are also considered to validate the performance of the proposed schemes, particularly the second-order one. For simplicity, we choose the uniform MAC spatial mesh size $h = h_x = h_y$ in 2D and $h = h_x = h_y = h_z$ in 3D. Unless otherwise stated, the viscosity coefficient is defined as $\nu = \frac{1}{Re}$ for a given Reynolds number Re.

Convergence test We set the spatial domain $\Omega = (0,1)^2$ and the final time $T = 1$. The external force $\mathbf{f}$ is computed according to the following analytic solution [100], whose velocity satisfies homogeneous Dirichlet boundary conditions:

$$\begin{cases} p(x,y,t) = \cos(\pi x)\sin(\pi y)\sin(t), \\ u(x,y,t) = \pi \sin^2(\pi x)\sin(2\pi y)\sin(t), \\ v(x,y,t) = -\pi \sin(2\pi x)\sin^2(\pi y)\sin(t). \end{cases}$$

We first consider the case $Re = 10$ (i.e., $\nu = 0.1$). To study the convergence of the DRLM schemes, we vary both the spatial mesh size and time step size with $h = \frac{\Delta t}{8}$ and $\Delta t \in \left\{\frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}\right\}$. The L^2 errors of the velocity and pressure, as well as the absolute errors of the Lagrange multiplier q with different values of $\theta \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ are shown in Figure 4.2 for both first- and second-order DRLM schemes. We observe that the errors in velocity, pressure, and q are large for small values of θ, such as $\theta \in \{10^{-2}, 10^{-1}\}$. We remark that when $\theta = 10^{-3}$ and $\Delta t \in \{\frac{1}{8}, \frac{1}{16}\}$, the second-order scheme produces complex values for q after a certain number of iterations; therefore, we do not consider $\theta = 10^{-3}$ in this case. Conversely, when $\theta \in \{1, 10, 10^2\}$, q stabilizes closer to 1, resulting in the expected convergence rates for the proposed schemes. This highlights the critical role of θ in our DRLM formulations: if θ is too small or absent, the schemes produce incorrect solutions, but they perform significantly better when θ is large enough, with a minimum value of 1.

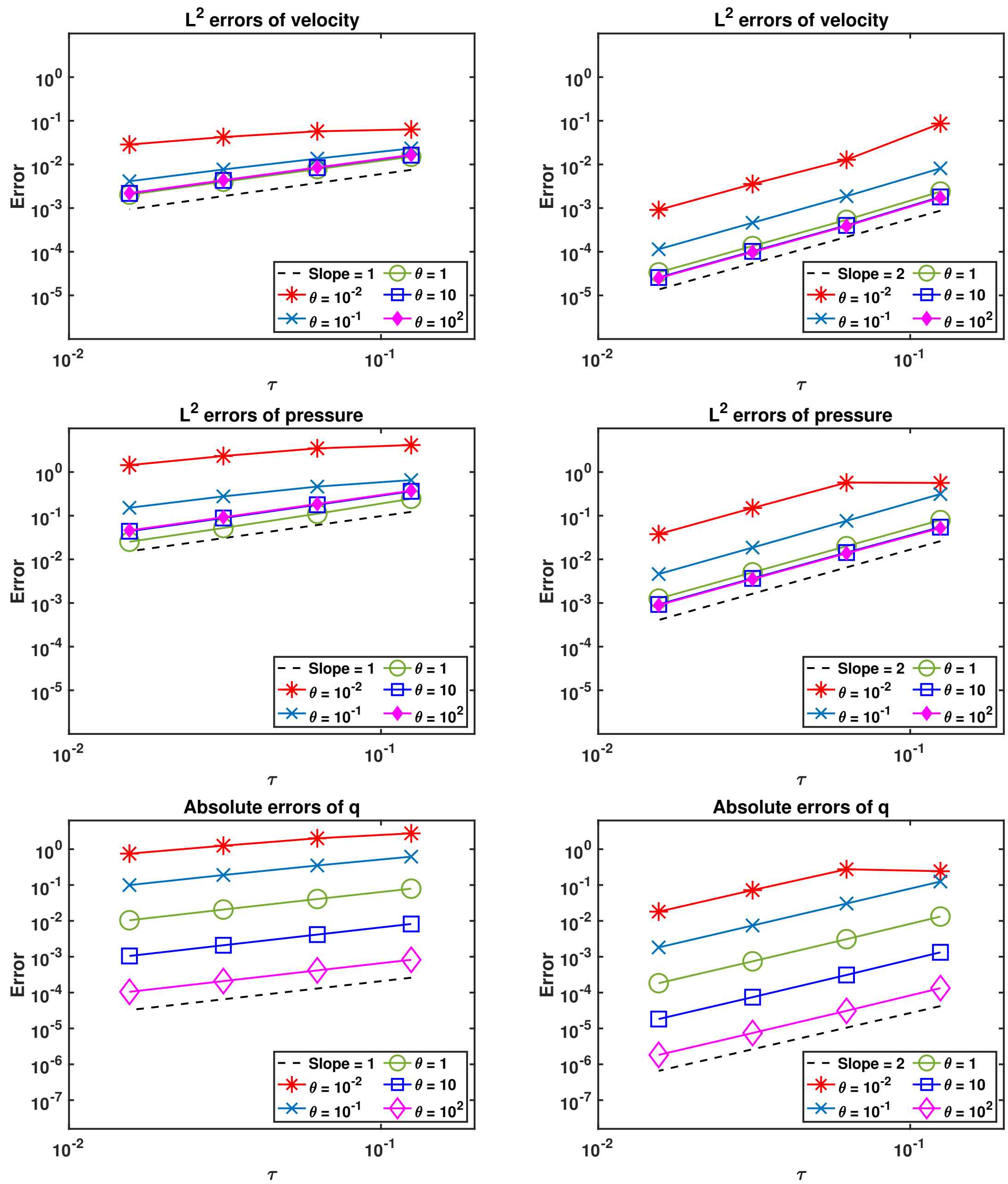


Figure 4.2: [Convergence test, $Re = 10$] Errors of the velocity, pressure, and Lagrange multiplier by the first-order (left) and second-order (right) DRLM schemes with $h = \frac{\Delta t}{8}$ and difference values of θ.

Next, we increase the Reynolds number to $Re = 1000$ (i.e., $\nu = 0.001$). Due to the explicit treatment of the convection term which requires some CFL condition, we consider $h = 4\Delta t$ and $\Delta t \in \left\{\frac{1}{128}, \frac{1}{256}, \frac{1}{512}, \frac{1}{1024}\right\}$. The corresponding errors of the velocity, pressure, and Lagrange multiplier for different values of $\theta \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ are presented in Figure 4.3. As θ increases, errors of the concerned quantities become smaller, and both first- and second-order DRLM schemes exhibit the correct order of convergence when $\theta \geq 1$. Note that the results with $\theta = 10^2$ are almost identical to those with $\theta = 10$, thus omitted here. Moreover, for this case with a large Reynolds number, $\theta = 10^{-3}$ could be used without leading to complex values for q; nevertheless, the expected accuracy of the numerical solution is only achieved when θ is sufficiently large.

Energy dissipation test We now examine the Taylor-Green vortex problem in $\Omega = (0,1)^2$ with zero external force, $\mathbf{f} = [0,0]^T$, and the periodic boundary conditions. The exact solution to this problem is given by

$$\begin{cases} u(x,y,t) = \sin(2\pi x)\cos(2\pi y)e^{-8\pi^2 \nu t}, \\ v(x,y,t) = -\cos(2\pi x)\sin(2\pi y)e^{-8\pi^2 \nu t}, \\ p(x,y,t) = \dfrac{1}{4}\left[\cos(4\pi x) + \cos(4\pi y)\right] e^{-16\pi^2 \nu t}. \end{cases}$$

The original kinetic energy can be computed explicitly as follows:

$$\mathcal{K}(\mathbf{u}(t)) = \frac{1}{2}\int_\Omega \left[u^2(x,y,t) + v^2(x,y,t)\right] dxdy = \frac{1}{4}e^{-16\pi^2 \nu t}.$$

We set $T = 20$, $Re = 1000$, and $h = 1.5\Delta t = 0.01$. The evolutions of $\mathcal{K}(\mathbf{u}^n)$, $\mathcal{K}(\mathbf{u})$, and $\mathcal{E}_i^n$ ($i = 1,2$) generated by the first- and second-order DRLM schemes with $\theta \in \{1, 100\}$ are shown in Figure 4.4, where we clearly observe the energy decay property of the numerical solutions. Figure 4.4 also confirms that the additional terms, $\mathcal{E}_1^n$ and $\mathcal{E}_2^n$, in our energy stability analysis (cf. Theorems 4.4-4.6 and 4.9-4.10) are negligible due to their small magnitudes. Furthermore, the results for $\theta = 1$ and $\theta = 100$ are quite similar, with the second-order DRLM scheme exhibiting better performance.

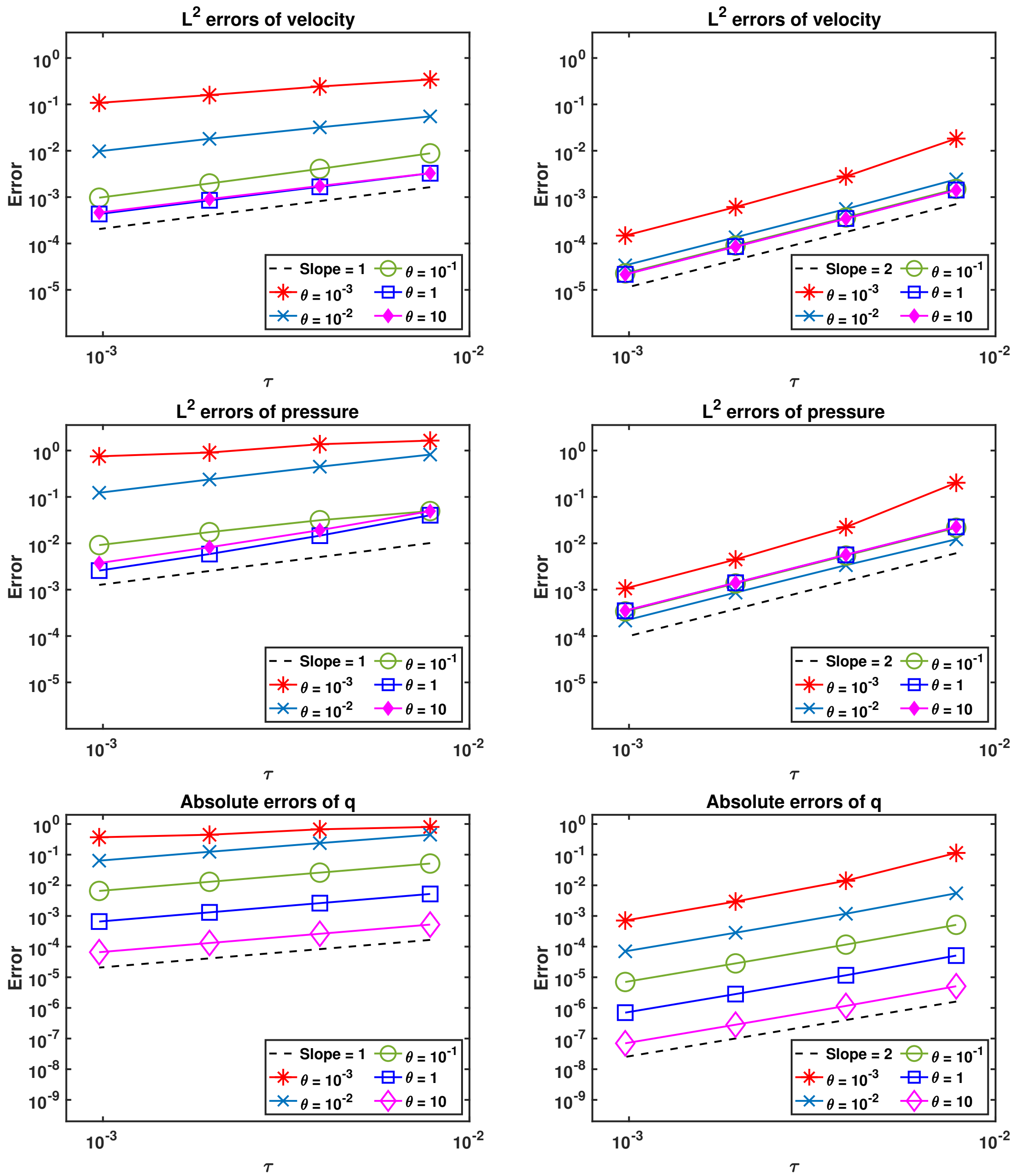


Figure 4.3: [Convergence test, $Re = 1000$] Errors of the velocity, pressure, and Lagrange multiplier by the first-order (left) and second-order (right) DRLM schemes with $h = 4\Delta t$ and difference values of θ.

Lid-driven cavity flow Next, we demonstrate the accuracy of our new schemes through realistic physical simulations, the well-known lid-driven cavity flow [65, 100, 122, 124] in 2D and [3, 108, 187, 74] in 3D. For the 2D problem, the computational domain $\Omega = (0,1)^2$ consists of three rigid walls (at $x = 0$, $x = 1$, and $y = 0$) with no-slip boundary conditions and a lid (at $y = 1$)

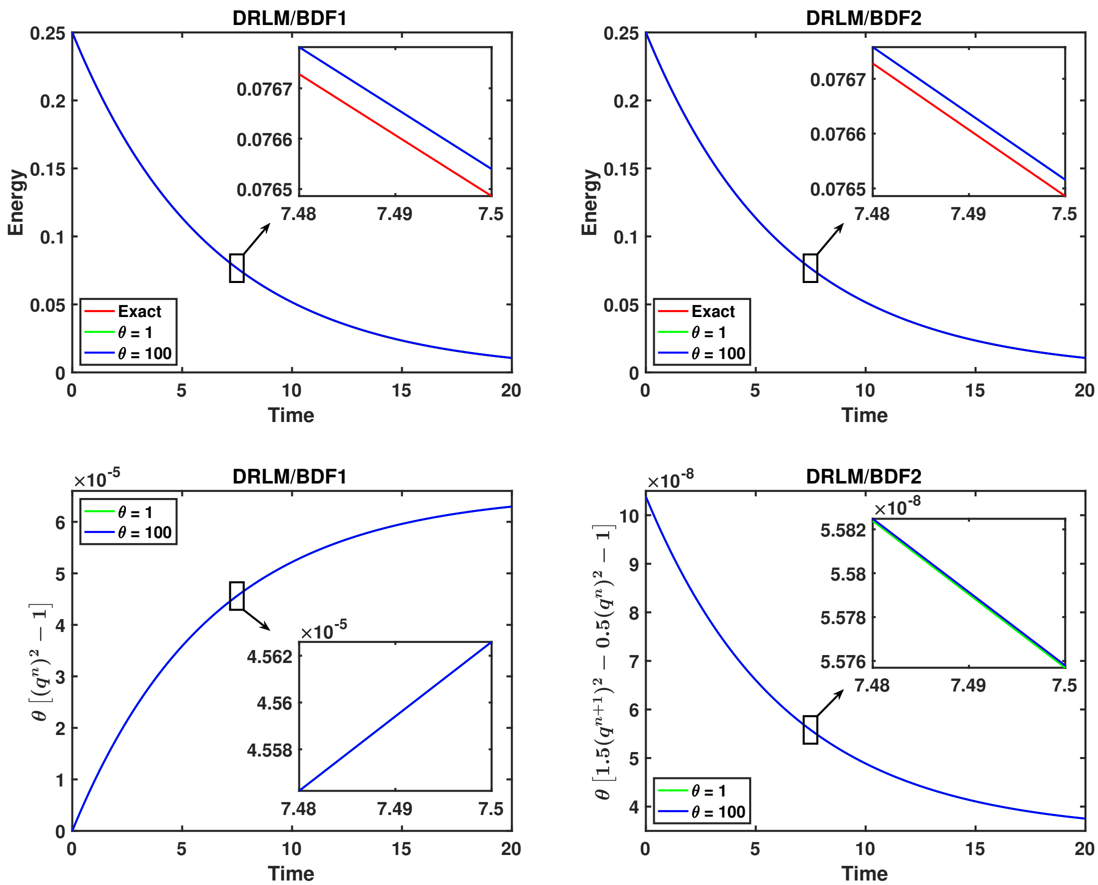


Figure 4.4: [Taylor-Green vortex, $Re = 1000$] Evolution of the energy functional and error function $\mathcal{E}_i^n$ ($i = 1, 2$) by the first-order (left) and second-order (right) DRLM schemes with $h = 1.5\Delta t = 0.01$.

moving with a tangential unit velocity. Since the exact solution to this problem is not available, we compare our numerical results with the benchmark method in [65]. Using the second-order DRLM scheme with $Re = 5000$, $\theta = 100$, and $h = 2\Delta t = \frac{1}{256}$, we perform the simulation until the steady state is reached, characterized by $\|\mathbf{u}^n - \mathbf{u}^{n-1}\|_\infty \leq 10^{-6}$. Contour plots of the velocity magnitude at times $t = 4$, $t = 6$, and $t = 10$ are presented in Figure 4.5. Figure 4.6 illustrates the velocity magnitude, vorticity, and stream function at the final steady state, along with velocity components at the cavity centerlines, which closely match the benchmark results [65]. As reported in [65] and [100], in addition to the primary and secondary vortices at the bottom corners, a third vortex emerges in the upper left corner at $Re = 5000$. Thus, our numerical simulation based on the DRLM approach accurately captures the dynamic evolution of the velocity field, achieving the steady state that compares well with existing results.

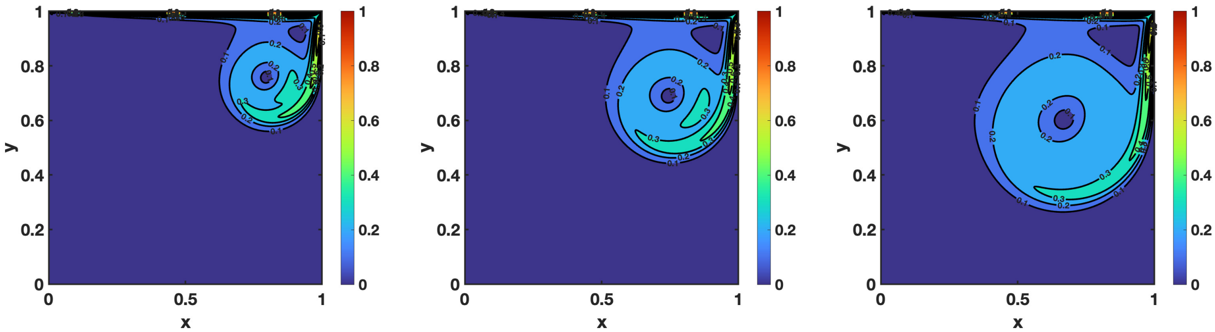


Figure 4.5: [2D lid-driven cavity, $Re = 5000$] Contour plots of the velocity magnitude at times $t = 4, 6$, and 10 by the second-order DRLM scheme with $h = 2\Delta t = \frac{1}{256}$.

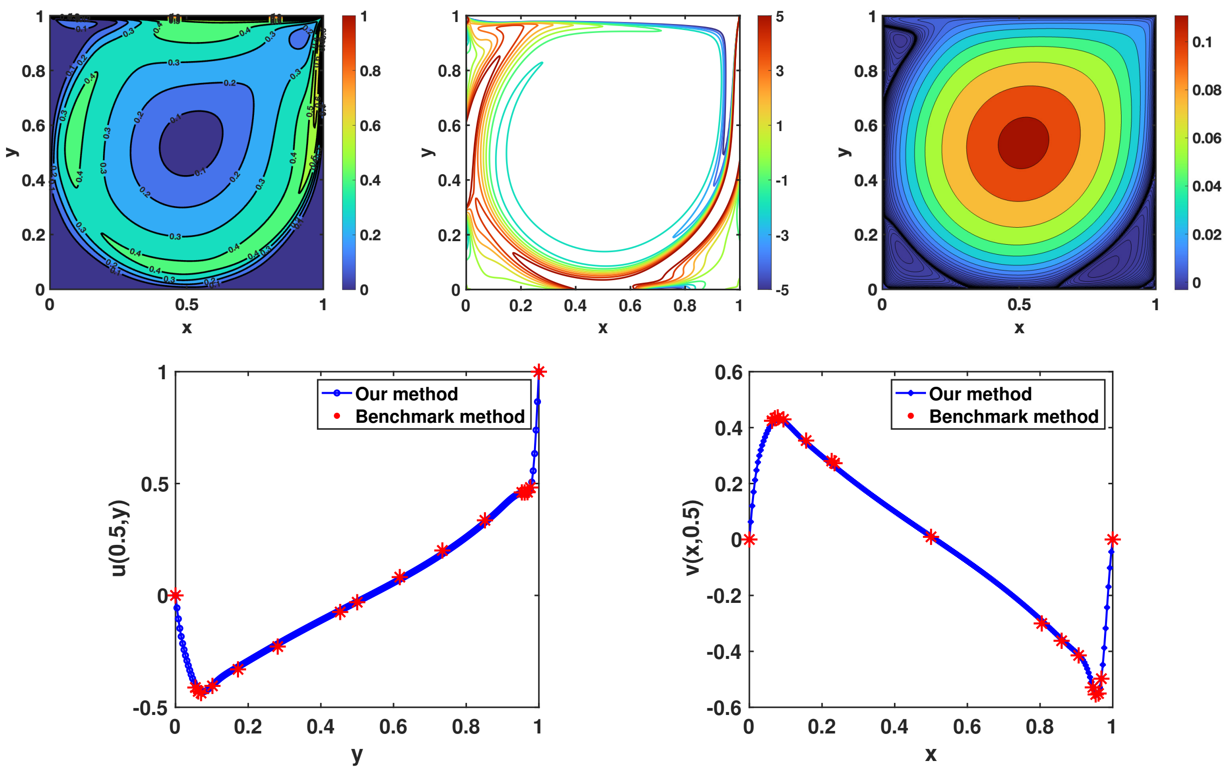


Figure 4.6: [2D lid-driven cavity, $Re = 5000$] Contour plots of the velocity magnitude, vorticity, and stream function (top) and the velocity components along the centerlines (bottom) at the steady state by the second-order DRLM scheme with $h = 2\Delta t = \frac{1}{256}$.

We proceed to consider the 3D lid-driven flow in a cubic cavity $\Omega = (0,1)^3$, a natural extension of the 2D driven cavity test case. The fluid is initially at rest and starts moving as the top lid (at $y = 1$) is dragged at a constant unit speed in the positive x-direction. On all other sides, the velocity satisfies homogeneous Dirichlet boundary conditions. With $Re = 1000$, $\theta = 100$, and $h = 1.5\Delta t = 0.01$ fixed, we simulate the flow evolution using the second-order DRLM scheme until the steady state is achieved, indicated by $\|\mathbf{u}^n - \mathbf{u}^{n-1}\|_0 \leq 10^{-6}$. Figure 4.7 displays contour plots of the vorticity components $[\omega_1, \omega_2, \omega_3]^T = \nabla \times \mathbf{u}$ on the midplanes $x = 0.5$, $y = 0.5$, and $z = 0.5$, respectively. The flow is symmetric about the plane $z = 0.5$, and the presence of Taylor-Görtler-like vortices in the bottom region of the cavity is observed besides corner vortices, which are consistent with those reported in [96, 143]. The distributions of the u- and v-velocity components along the vertical ($x = z = 0.5$) and horizontal ($y = z = 0.5$) plane centerlines are provided in Figure 4.8, demonstrating strong agreement with the benchmark results [3].

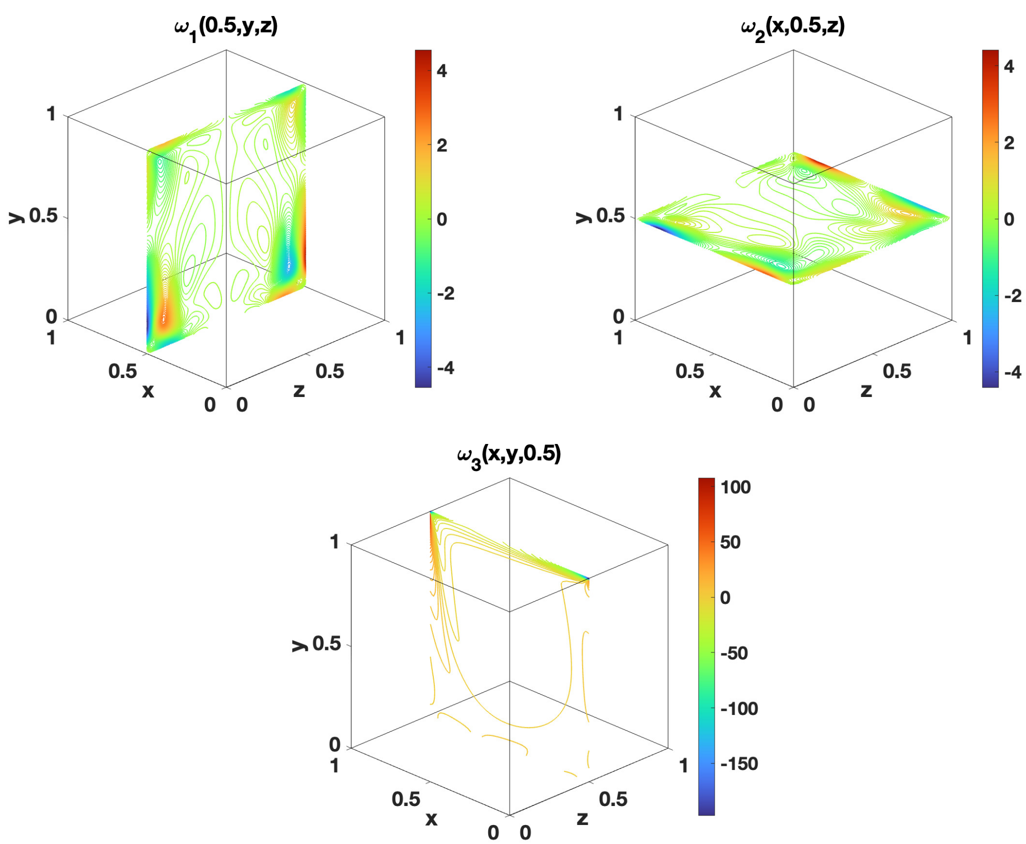


Figure 4.7: [3D lid-driven cavity, $Re = 1000$] Contour plots of the vorticity components on the midplanes at the steady state by the second-order DRLM scheme with $h = 1.5\Delta t = 0.01$.

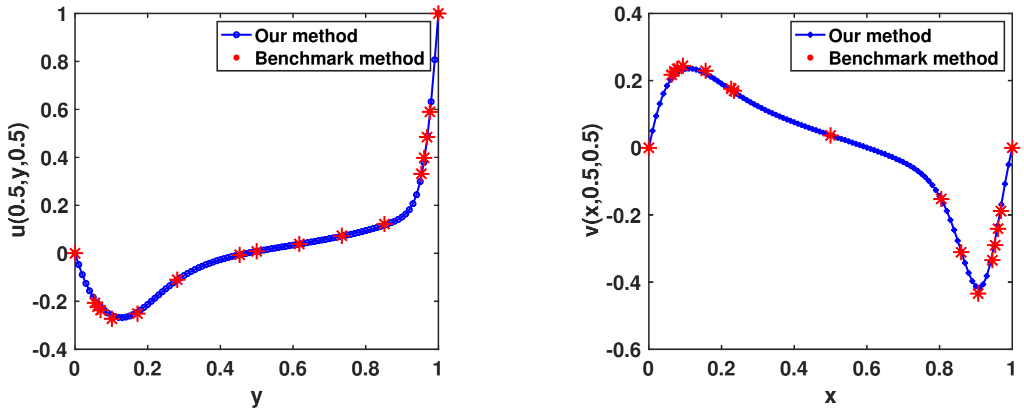


Figure 4.8: [3D lid-driven cavity, $Re = 1000$] Velocity components along the centerlines at the steady state by the second-order DRLM scheme with $h = 1.5\Delta t = 0.01$.

Kelvin–Helmholtz instability When there is an initial velocity difference across a shear layer, small disturbances can grow over time, leading to the formation of vortices. This phenomenon is known as the Kelvin-Helmholtz instability. We consider the 2D NS equations in $\Omega = (0,1)^2$ with $\mathbf{f} = [0,0]^T$. At the top and bottom boundaries of the domain Ω, we apply the conditions $\frac{\partial u}{\partial y} = 0$ and $v = 0$, while at the left and right sides of Ω, we impose periodic boundary conditions. Since no body forces are present, the entire motion is driven by the initial condition, which is given by [157]:

$$\mathbf{u}_0(x,y) = \begin{bmatrix} u_\infty \tanh\left(\frac{2y-1}{\delta_0}\right) \\ 0 \end{bmatrix} + c_n \begin{bmatrix} \partial_y \psi(x,y) \\ -\partial_x \psi(x,y) \end{bmatrix},$$

with the corresponding stream function

$$\psi(x,y) = u_\infty \exp\left(-\frac{(y-0.5)^2}{\delta_0^2}\right)[\cos(8\pi x) + \cos(20\pi x)].$$

Here, $\delta_0 = \frac{1}{28}$ denotes the initial vorticity thickness, $u_\infty = 1$ is a reference velocity, and $c_n = 10^{-3}$ is a scaling factor. With fixed Reynolds number $Re = 100$, the kinematic viscosity is given by $\nu = \frac{\delta_0 u_\infty}{Re} = \frac{1}{2800}$. We take $h = \frac{1}{256}$, $\Delta t = \frac{1}{560}$, $\theta = 100$, and $T = 200\bar{t}$ with $\bar{t} = \frac{\delta_0}{u_\infty} = \frac{1}{28}$ being the time unit. Evolution of the vorticity produced by the second-order DRLM scheme at different times is illustrated in Figure 4.9. We observe that four vortices gradually emerge from the initial condition, exhibiting instability and a tendency to merge into two larger vortices. These two ellipsoidal vortices remain separated for a long time, with their magnitudes decreasing until the final time $T = 200\bar{t}$. Figure 4.10 confirms the energy dissipation of the system, in agreement with the benchmark results [157], and shows the evolution of the discrete error function $\mathcal{E}_2^n$, which remains insignificant compared to the total energy.

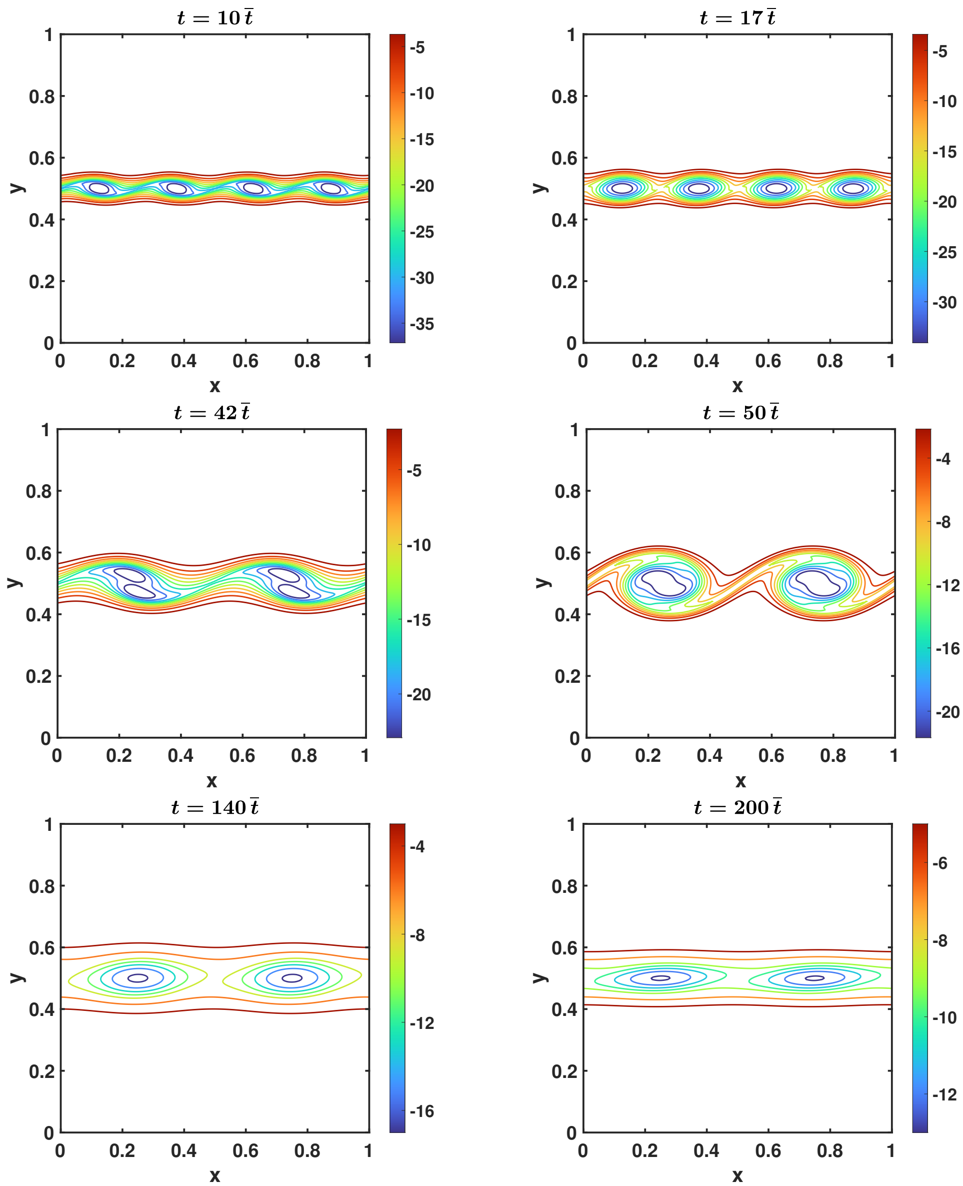


Figure 4.9: [Kelvin-Helmholtz instability, $Re = 100$] Evolution of the vorticity at different times $t = 10\bar{t}$, $17\bar{t}$, $42\bar{t}$, $50\bar{t}$, $140\bar{t}$, and $200\bar{t}$ by the second-order DRLM scheme with $h = \dfrac{1}{256}$ and $\Delta t = \dfrac{1}{560}$.

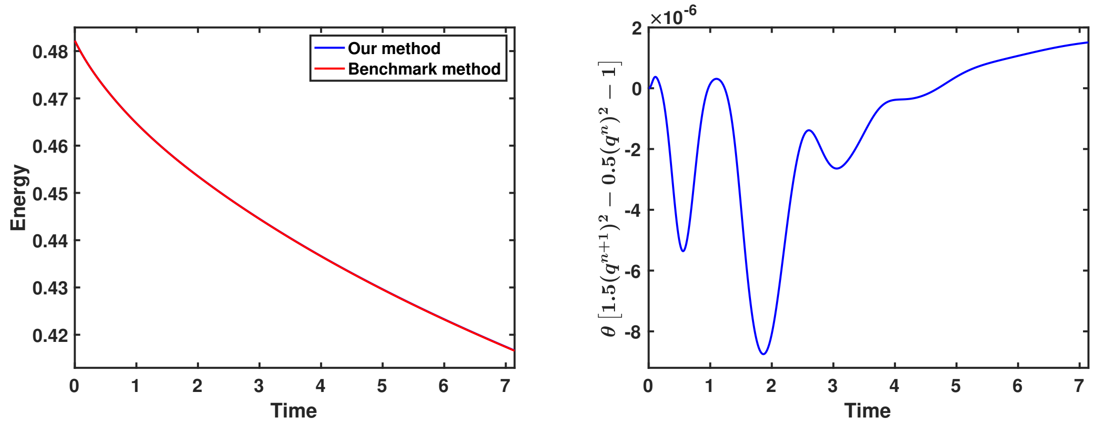


Figure 4.10: [Kelvin-Helmholtz instability, $Re = 100$] Evolution of the energy functional and error function $\mathcal{E}_2^n$ by the second-order DRLM scheme with $h = \dfrac{1}{256}$ and $\Delta t = \dfrac{1}{560}$.

4.2 Convergence analysis of the DRLM method for the Navier-Stokes equations

4.2.1 Model problem

For an open, bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) and a terminal time $T > 0$, we consider the following incompressible NS equations:

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f}, \quad \text{in } \Omega \times (0, T], \tag{4.47a}$$

$$\nabla \cdot \mathbf{u} = 0, \quad \text{in } \Omega \times (0, T], \tag{4.47b}$$

subject to the initial condition $\mathbf{u}(\cdot, 0) = \mathbf{u}_0$ and no-slip boundary condition, i.e. $\mathbf{u}|_{\partial\Omega} = 0$. In (4.47), $\mathbf{u}$ and p are the unknown velocity vector and pressure, respectively, $\mathbf{f}$ is an external force, and ν represents the kinematic viscosity.

Let $L^2(\Omega)$, $H^k(\Omega)$, and $H_0^k(\Omega)$ denote the standard Sobolev spaces over Ω, with $\mathbf{L}^2(\Omega)$, $\mathbf{H}^k(\Omega)$, and $\mathbf{H}_0^k(\Omega)$ as their d-dimensional extensions, e.g., $\mathbf{H}^k(\Omega) = (H^k(\Omega))^d$. The norms on $H^k(\Omega)$ and $\mathbf{H}^k(\Omega)$ are indicated by $\|\cdot\|_k$. For $L^2(\Omega)$ and $\mathbf{L}^2(\Omega)$, we denote by $(\cdot, \cdot)$ and $\|\cdot\|_0$ the inner product and the associated norm, respectively. Since the pressure is defined up to an additive constant in the NS equations, we define $H^k(\Omega)/\mathbb{R}$ as the quotient space consisting of equivalence classes of elements of $H^k(\Omega)$ differing by constants. When $k = 0$, the quotient space is denoted by $L^2(\Omega)/\mathbb{R}$. For $\varphi \in H^k(\Omega)/\mathbb{R}$, its norm is given by $\|\varphi\|_{H^k/\mathbb{R}} = \inf_{c\in\mathbb{R}} \|\varphi + c\|_k$.

Denote by $\mathbf{H}$ and $\mathbf{V}$ the following Hilbert spaces:

$$\mathbf{H} = \{\mathbf{u} \in \mathbf{L}^2(\Omega) : \nabla \cdot \mathbf{u} = 0,\ \mathbf{u} \cdot \mathbf{n}|_{\partial\Omega} = 0\}, \quad \mathbf{V} = \{\mathbf{v} \in \mathbf{H}_0^1(\Omega) : \nabla \cdot \mathbf{v} = 0\}.$$

Let us state a result on the existence and uniqueness of a strong solution to the NS equations (4.47), which can be found, for example, in [80, 81, 175, 13].

Theorem 4.12. Assume $\mathbf{u}_0 \in \mathbf{V}$ and $\mathbf{f} \in L^\infty(0, T; \mathbf{L}^2(\Omega))$. There exists a positive time T^*, with $T^* = T$ if $d = 2$ and $T^* = T^*(\mathbf{u}_0) \leq T$ if $d = 3$, such that the NS equations (4.47) admit a unique strong solution $(\mathbf{u}, p)$ satisfying

$$\begin{aligned} &\mathbf{u} \in L^2(0, T^*; \mathbf{H}^2(\Omega)) \cap C([0, T^*]; \mathbf{V}), \quad \mathbf{u}_t \in L^2(0, T^*; \mathbf{L}^2(\Omega)), \\ &p \in L^2(0, T^*; H^1(\Omega)/\mathbb{R}). \end{aligned}$$

Moreover, if $d = 2$ or, in the case $d = 3$, if $\|\mathbf{u}_0\|_1$ and $\|\mathbf{f}\|_{L^\infty(0,T;\mathbf{L}^2(\Omega))}$ are sufficiently small, then the solution $(\mathbf{u}, p)$ exists for all $t \in [0, T]$, i.e. $T^* = T$ for $d \in \{2, 3\}$, and

$$\sup_{t\in[0,T]} \|\mathbf{u}(t)\|_1 < \infty. \tag{4.48}$$

The goal of this section is to establish temporal error estimates for the DRLM method applied to the NS equations. Toward that end, throughout this section we assume $\mathbf{u}_0$ and $\mathbf{f}$ satisfy the following regularity conditions:

$$\mathbf{u}_0 \in \mathbf{V} \cap \mathbf{H}^2(\Omega), \text{ and } \mathbf{f} \in L^\infty(0, T; \mathbf{L}^2(\Omega)). \tag{4.49}$$

4.2.2 DRLM method and properties

In this section, we recall the DRLM formulation introduced in [48], which is equivalent to the original NS equations (4.47). Based on this reformulation, we present the first-order DRLM scheme and analyze its fundamental properties. In particular, we establish the unique solvability of the scheme and derive a uniform bound on the Lagrange multiplier, which plays a crucial role in the subsequent error analysis. Although the DRLM framework naturally extends to higher-order schemes, their rigorous analysis is highly challenging and remains an open problem.

DRLM formulation and discretization schemes The idea of the DRLM approach [48] is to introduce a time-dependent Lagrange multiplier $q(t)$ with $q(0) = 1$ and a constant regularization parameter $\theta > 0$ to reformulate the NS equations (4.47) equivalently as follows:

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + q(\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f}, \qquad \text{in } \Omega\times(0,T], \tag{4.50a}$$

$$\nabla\cdot\mathbf{u} = 0, \qquad \text{in } \Omega\times(0,T], \tag{4.50b}$$

$$\frac{d\mathcal{K}(\mathbf{u})}{dt} + \theta\frac{dq^2}{dt} = (\mathbf{u}_t + q(\mathbf{u}\cdot\nabla)\mathbf{u}, \mathbf{u}), \quad \text{in } (0,T], \tag{4.50c}$$

under the same initial and boundary conditions. In (4.50c), $\mathcal{K}(\mathbf{u}) = \frac{1}{2}\|\mathbf{u}\|_0^2$ denotes the kinetic energy associated with (4.47) and the notation $(\cdot,\cdot)$ represents the standard $\mathbf{L}^2$ inner product.

For temporal discretization, let us consider a uniform partition of the time interval $[0,T]$: $0 = t_0 < t_1 < \cdots < t_N = T$ with the time step size $\Delta t = T/N$. From the reformulated system (4.50), DRLM schemes can be derived using BDF methods with explicit treatment of the nonlinear convection term [48]. In particular, the first-order DRLM scheme, based on the first-order BDF (i.e., backward Euler) method, is given by

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t} - \nu\Delta\mathbf{u}^{n+1} + q^{n+1}(\mathbf{u}^n\cdot\nabla)\mathbf{u}^n + \nabla p^{n+1} = \mathbf{f}(t_{n+1}), \tag{4.51a}$$

$$\nabla\cdot\mathbf{u}^{n+1} = 0, \quad \mathbf{u}^{n+1}|_{\partial\Omega} = 0, \tag{4.51b}$$

$$\frac{\mathcal{K}(\mathbf{u}^{n+1})-\mathcal{K}(\mathbf{u}^n)}{\Delta t} + \theta\frac{(q^{n+1})^2-(q^n)^2}{\Delta t} = \left(\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t} + q^{n+1}(\mathbf{u}^n\cdot\nabla)\mathbf{u}^n, \mathbf{u}^{n+1}\right). \tag{4.51c}$$

We refer to [48] for the formulation of the second-order DRLM scheme, as well as the unconditional energy stability results for both first- and second-order schemes. Note that the initial conditions for the velocity and Lagrange multiplier are $\mathbf{u}^0 = \mathbf{u}_0$ and $q^0 = 1$. An efficient implementation of the DRLM scheme (4.51), which involves solving two generalized Stokes systems and a quadratic equation for the Lagrange multiplier at each time step, is presented in [48] and thus omitted.

Hereafter, for analysis purposes, we assume that θ is bounded from below by a positive constant, i.e., $\theta \geq \theta_{\text{tol}}$ for some $\theta_{\text{tol}} > 0$. For simplicity, we take $\theta_{\text{tol}} = 1$ so that $\theta \geq 1$. It should be noted that when θ is too small or zero, DRLM schemes produce inaccurate values for the Lagrange multiplier unless the time step size is sufficiently small, which affects the accuracy of the velocity and pressure. More details about numerical performance of the first- and second-order DRLM schemes with different values of θ can be found in [48].

Properties We begin with the unique solvability and regularity results of the DRLM scheme (4.51).

Lemma 4.13. The DRLM scheme (4.51) admits a unique solution $(\mathbf{u}^n, p^n, q^n)$ with

$$\mathbf{u}^n \in \mathbf{V} \cap \mathbf{H}^2(\Omega), \quad p^n \in H^1(\Omega)/\mathbb{R}, \quad q^n > 0,$$

for every $n \in \{1, 2, \ldots, N\}$.

Proof. It is clear from the initial conditions and assumption (4.49) that $\mathbf{u}^0 = \mathbf{u}_0 \in \mathbf{V} \cap \mathbf{H}^2(\Omega)$ and $q^0 = 1 > 0$. Suppose $(\mathbf{u}^n, q^n)$ with $\mathbf{u}^n \in \mathbf{V} \cap \mathbf{H}^2(\Omega)$ and $q^n > 0$ is uniquely determined from (4.51) for some nonnegative integer n. We consider the following generalized Stokes systems:

$$\frac{\mathbf{u}_1^{n+1} - \mathbf{u}^n}{\Delta t} - \nu \Delta \mathbf{u}_1^{n+1} + \nabla p_1^{n+1} = \mathbf{f}(t_{n+1}), \qquad \nabla \cdot \mathbf{u}_1^{n+1} = 0, \quad \mathbf{u}_1^{n+1}|_{\partial\Omega} = 0, \tag{4.52}$$

$$\frac{\mathbf{u}_2^{n+1}}{\Delta t} - \nu \Delta \mathbf{u}_2^{n+1} + (\mathbf{u}^n \cdot \nabla)\mathbf{u}^n + \nabla p_2^{n+1} = 0, \quad \nabla \cdot \mathbf{u}_2^{n+1} = 0, \quad \mathbf{u}_2^{n+1}|_{\partial\Omega} = 0. \tag{4.53}$$

We know that $\frac{1}{\Delta t}\mathbf{u}^n + \mathbf{f}(t_{n+1}) \in \mathbf{L}^2(\Omega)$ and $-(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n \in \mathbf{L}^2(\Omega)$ by the inductive hypothesis and assumption (4.49). Based on the well-posedness and regularity results for Stokes systems (see, for instance, [19, 80, 81]), we conclude that (4.52) and (4.53) admit unique solutions, $(\mathbf{u}_1^{n+1}, p_1^{n+1})$ and $(\mathbf{u}_2^{n+1}, p_2^{n+1})$, respectively, satisfying

$$\mathbf{u}_i^{n+1} \in \mathbf{V} \cap \mathbf{H}^2(\Omega) \text{ and } p_i^{n+1} \in H^1(\Omega)/\mathbb{R} \text{ for } i = 1, 2. \tag{4.54}$$

Setting $\mathbf{u}^{n+1} = \mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1}$ and plugging it in (4.51c) yields

$$\begin{aligned} &\frac{\|\mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1}\|_0^2 - \|\mathbf{u}^n\|_0^2}{2\Delta t} + \theta \frac{(q^{n+1})^2 - (q^n)^2}{\Delta t} \\ &= \left(\frac{\mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1} - \mathbf{u}^n}{\Delta t} + q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n, \mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1} \right), \end{aligned}$$

which can be simplified as follows:

$$A_{n+1}(q^{n+1})^2 + B_{n+1}q^{n+1} + C_{n+1} = 0, \tag{4.55}$$

where the quadratic coefficients are given by

$$\begin{aligned}
A_{n+1} &= \theta + \frac{1}{2}\|\mathbf{u}_2^{n+1}\|_0^2 + \Delta t\nu\|\nabla\mathbf{u}_2^{n+1}\|_0^2, \\
B_{n+1} &= -\left(\mathbf{u}_1^{n+1} - \mathbf{u}^n, \mathbf{u}_2^{n+1}\right) - \Delta t\left((\mathbf{u}^n \cdot \nabla)\mathbf{u}^n, \mathbf{u}_1^{n+1}\right), \\
C_{n+1} &= -\theta(q^n)^2 - \frac{1}{2}\|\mathbf{u}_1^{n+1} - \mathbf{u}^n\|_0^2.
\end{aligned}$$

We observe that $A_{n+1} > 0$ and $C_{n+1} < 0$ for all $\theta > 0$. As a consequence, (4.55) has a unique positive solution, namely $q^{n+1} > 0$. Let $p^{n+1} := p_1^{n+1} + q^{n+1}p_2^{n+1}$. We find that $\mathbf{u}^{n+1} \in \mathbf{V} \cap \mathbf{H}^2(\Omega)$ and $p^{n+1} \in H^1(\Omega)/\mathbb{R}$ due to (4.54). Furthermore, $(\mathbf{u}^{n+1}, p^{n+1}, q^{n+1})$ satisfies the DRLM scheme (4.51), thereby verifying the existence of a solution to (4.51) at t_{n+1}.

To prove the uniqueness, we claim that if (4.51) admits a solution $(\mathbf{u}^{n+1}, p^{n+1}, q^{n+1})$, then

$$\mathbf{u}^{n+1} = \mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1}, \tag{4.56a}$$

$$p^{n+1} = p_1^{n+1} + q^{n+1}p_2^{n+1} + c^{n+1}, \tag{4.56b}$$

for any real number c^{n+1} (i.e., the pressure is unique up to a constant). Indeed, let $\widehat{\mathbf{u}}^{n+1} = \mathbf{u}^{n+1} - \mathbf{u}_1^{n+1} - q^{n+1}\mathbf{u}_2^{n+1}$ and $\widehat{p}^{n+1} = p^{n+1} - p_1^{n+1} - q^{n+1}p_2^{n+1}$. The combination of (4.51a), (4.51b), (4.52), and (4.53) results in

$$\frac{\widehat{\mathbf{u}}^{n+1}}{\Delta t} - \nu\Delta\widehat{\mathbf{u}}^{n+1} + \nabla\widehat{p}^{n+1} = \mathbf{0}, \tag{4.57a}$$

$$\nabla \cdot \widehat{\mathbf{u}}^{n+1} = 0, \quad \widehat{\mathbf{u}}^{n+1}|_{\partial\Omega} = 0. \tag{4.57b}$$

Taking the $\mathbf{L}^2$ inner product of (4.57a) with $\widehat{\mathbf{u}}^{n+1}$ and utilizing (4.57b) gives us

$$\frac{\|\widehat{\mathbf{u}}^{n+1}\|_0^2}{\Delta t} + \nu\|\nabla\widehat{\mathbf{u}}^{n+1}\|_0^2 = 0,$$

which leads to $\widehat{\mathbf{u}}^{n+1} = \mathbf{0}$, and therefore, $\nabla\widehat{p}^{n+1} = \mathbf{0}$ due to (4.57a). This means that $\mathbf{u}^{n+1}$ and p^{n+1} satisfy the decompositions (4.56). Hence, if both $(\overline{\mathbf{u}}^{n+1}, \overline{p}^{n+1}, \overline{q}^{n+1})$ and

$(\tilde{\mathbf{u}}^{n+1}, \tilde{p}^{n+1}, \tilde{q}^{n+1})$ are the solutions of (4.51), where $\overline{q}^{n+1} > 0$ and $\tilde{q}^{n+1} > 0$, then we have

$$\overline{\mathbf{u}}^{n+1} = \mathbf{u}_1^{n+1} + \overline{q}^{n+1}\mathbf{u}_2^{n+1}, \tag{4.58a}$$

$$\overline{p}^{n+1} = p_1^{n+1} + \overline{q}^{n+1}p_2^{n+1} + \overline{c}^{n+1}, \tag{4.58b}$$

and

$$\tilde{\mathbf{u}}^{n+1} = \mathbf{u}_1^{n+1} + \tilde{q}^{n+1}\mathbf{u}_2^{n+1}, \tag{4.59a}$$

$$\tilde{p}^{n+1} = p_1^{n+1} + \tilde{q}^{n+1}p_2^{n+1} + \tilde{c}^{n+1}. \tag{4.59b}$$

By substituting (4.58a) and (4.59a) into (4.51c) (with $(\mathbf{u}^{n+1}, q^{n+1})$ replaced by $(\overline{\mathbf{u}}^{n+1}, \overline{q}^{n+1})$ or $(\tilde{\mathbf{u}}^{n+1}, \tilde{q}^{n+1})$), we find that $\overline{q}^{n+1}$ and $\tilde{q}^{n+1}$ are positive solutions of (4.55). Thus, we must have $\overline{q}^{n+1} = \tilde{q}^{n+1}$, and consequently, $\overline{\mathbf{u}}^{n+1} = \tilde{\mathbf{u}}^{n+1}$ and $\overline{p}^{n+1} = \tilde{p}^{n+1} - \tilde{c}^{n+1} + \overline{c}^{n+1}$. $\square$

Next, we show that the Lagrange multiplier in the DRLM scheme (4.51) is uniformly bounded for any time step size $\Delta t > 0$.

Lemma 4.14. Let $\{q^n\}_{n=0}^N$ be the positive sequence generated by the DRLM scheme (4.51). There exists a constant $c_0 \geq 1$, depending only on T, $\|\mathbf{u}_0\|_0$, and $\|\mathbf{f}\|_{L^\infty(0,T;\mathbf{L}^2(\Omega))}$, such that

$$0 < q^n \leq c_0, \quad n = 0, 1, \dots, N.$$

Proof. Taking the $\mathbf{L}^2$ inner product of (4.51a) with $\mathbf{u}^{n+1}$ yields

$$\left(\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n, \mathbf{u}^{n+1}\right) = -\nu\|\nabla\mathbf{u}^{n+1}\|_0^2 + \left(\mathbf{f}(t_{n+1}), \mathbf{u}^{n+1}\right). \tag{4.60}$$

Let $c_{\mathbf{f}} = \|\mathbf{f}\|_{L^\infty(0,T;\mathbf{L}^2(\Omega))}$. From (4.51c) and (4.60), we obtain

$$\begin{aligned}\frac{\|\mathbf{u}^{n+1}\|_0^2 - \|\mathbf{u}^n\|_0^2}{2\Delta t} + \theta\frac{(q^{n+1})^2 - (q^n)^2}{\Delta t} &= -\nu\|\nabla\mathbf{u}^{n+1}\|_0^2 + \left(\mathbf{f}(t_{n+1}), \mathbf{u}^{n+1}\right) \\ &\leq \|\mathbf{f}(t_{n+1})\|_0\|\mathbf{u}^{n+1}\|_0 \leq c_{\mathbf{f}}\|\mathbf{u}^{n+1}\|_0.\end{aligned} \tag{4.61}$$

Summing (4.61) over $i = 0, 1, \dots, n$, noting that $q^0 = 1$, gives us

$$\frac{\|\mathbf{u}^{n+1}\|_0^2 - \|\mathbf{u}^0\|_0^2}{2\Delta t} + \theta\frac{(q^{n+1})^2 - 1}{\Delta t} \leq c_{\mathbf{f}}\sum_{i=0}^{n}\|\mathbf{u}^{i+1}\|_0,$$

or equivalently,

$$\|\mathbf{u}^{n+1}\|_0^2 + 2\theta(q^{n+1})^2 \le \|\mathbf{u}^0\|_0^2 + 2\theta + 2c_{\mathbf{f}}\Delta t \sum_{i=0}^{n} \|\mathbf{u}^{i+1}\|_0. \tag{4.62}$$

Let $\eta_n := \max_{1\le i\le n} \|\mathbf{u}^i\|_0$ for $n \in \{1, 2, \dots, N\}$. It follows from (4.62) that

$$\|\mathbf{u}^{n+1}\|_0^2 + 2\theta(q^{n+1})^2 \le \|\mathbf{u}^0\|_0^2 + 2\theta + 2c_{\mathbf{f}}T\eta_{n+1}. \tag{4.63}$$

Since $\{\eta_n\}$ is a non-decreasing sequence, it is implied from (4.63) for $0 \le i \le n$ that

$$\|\mathbf{u}^{i+1}\|_0^2 \le \|\mathbf{u}^0\|_0^2 + 2\theta + 2c_{\mathbf{f}}T\eta_{i+1} \le \|\mathbf{u}^0\|_0^2 + 2\theta + 2c_{\mathbf{f}}T\eta_{n+1},$$

which leads to

$$\eta_{n+1}^2 \le \|\mathbf{u}^0\|_0^2 + 2\theta + 2c_{\mathbf{f}}T\eta_{n+1}. \tag{4.64}$$

Note that the quadratic equation $\eta_{n+1}^2 = \|\mathbf{u}^0\|_0^2 + 2\theta + 2c_{\mathbf{f}}T\eta_{n+1}$ has two distinct solutions of opposite signs. Therefore, η_{n+1} satisfying (4.64) is bounded by the positive solution, i.e.,

$$\eta_{n+1} \le c_{\mathbf{f}}T + \sqrt{(c_{\mathbf{f}}T)^2 + \|\mathbf{u}^0\|_0^2 + 2\theta} \le 2c_{\mathbf{f}}T + \|\mathbf{u}^0\|_0 + \sqrt{2\theta}. \tag{4.65}$$

The combination of (4.63) and (4.65) results in

$$\|\mathbf{u}^{n+1}\|_0^2 + 2\theta(q^{n+1})^2 \le \|\mathbf{u}^0\|_0^2 + 2\theta + 2c_{\mathbf{f}}T(2c_{\mathbf{f}}T + \|\mathbf{u}^0\|_0 + \sqrt{2\theta}). \tag{4.66}$$

In particular, noting that $\theta \ge 1$, we have

$$\begin{aligned}(q^{n+1})^2 &\le 1 + \frac{\|\mathbf{u}^0\|_0^2 + 2c_{\mathbf{f}}T(2c_{\mathbf{f}}T + \|\mathbf{u}^0\|_0 + \sqrt{2\theta})}{2\theta} \\ &\le 1 + \frac{\|\mathbf{u}^0\|_0^2 + 2c_{\mathbf{f}}T(2c_{\mathbf{f}}T + \|\mathbf{u}^0\|_0 + \sqrt{2})}{2}.\end{aligned}$$

The proof is thus completed. □

4.2.3 Error analysis

Let $\{\mathbf{u}^n, p^n, q^n\}$ be the numerical solution at time t_n obtained by the DRLM scheme (4.51), we define the errors of the velocity, pressure, and Lagrange multiplier as follows:

$$e_{\mathbf{u}}^n = \mathbf{u}^n - \mathbf{u}(t_n), \;\; e_p^n = p^n - p(t_n), \text{ and } \; e_q^n = q^n - 1.$$

In particular, we have $e_{\mathbf{u}}^0 = \mathbf{0}$ and $e_q^0 = 0$. Before stating the main results of this section, let us first recall some useful inequalities.

Preliminary inequalities Since the homogeneous Dirichlet boundary condition is imposed for the velocity field, we invoke the Poincaré inequality and obtain, for some constant c_1 depending only on the domain Ω:

$$\|\mathbf{v}\|_1 \le (1 + c_1)\|\nabla \mathbf{v}\|_0, \quad \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega). \tag{4.67}$$

For $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{H}_0^1(\Omega)$, we define the trilinear form $b(\cdot,\cdot,\cdot)$ by

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

It can be shown that b is skew-symmetric with respect to its last two arguments, i.e.,

$$b(\mathbf{u}, \mathbf{v}, \mathbf{w}) = -b(\mathbf{u}, \mathbf{w}, \mathbf{v}), \quad \forall\, \mathbf{u} \in \mathbf{V}, \; \forall\, \mathbf{v}, \mathbf{w} \in \mathbf{H}_0^1(\Omega). \tag{4.68}$$

By setting $\mathbf{w} = \mathbf{v}$ in (4.68), we have $b(\mathbf{u}, \mathbf{v}, \mathbf{v}) = 0, \; \forall\, \mathbf{u} \in \mathbf{V}, \; \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega)$. Furthermore, for $d \le 3$, there exists a constant c_2 depending on Ω such that (cf. [159, 175, 177, 13, 124])

$$|b(\mathbf{u}, \mathbf{v}, \mathbf{w})| \le c_2 \begin{cases} \|\mathbf{u}\|_1 \|\mathbf{v}\|_1 \|\mathbf{w}\|_1, \\ \|\mathbf{u}\|_0 \|\mathbf{v}\|_2 \|\mathbf{w}\|_1, \\ \|\mathbf{u}\|_1 \|\mathbf{v}\|_2 \|\mathbf{w}\|_0, \\ \|\mathbf{u}\|_0 \|\mathbf{v}\|_1 \|\mathbf{w}\|_2, \\ \|\mathbf{u}\|_2 \|\mathbf{v}\|_1 \|\mathbf{w}\|_0. \end{cases} \tag{4.69}$$

A discrete version of Gronwall's inequality (see, for example, [44, 79, 152]) is stated below:

Lemma 4.15. Let $\{\tilde{a}_n\}_{n=0}^K$, $\{\tilde{b}_n\}_{n=1}^K$, $\{\tilde{c}_n\}_{n=1}^K$, and $\{\omega_n\}_{n=0}^K$ be nonnegative sequences with $\omega_0\tilde{a}_0 + \tilde{c}_1 \le \tilde{c}_1 e^{\omega_0}$. If $\{\tilde{c}_n\}$ is non-decreasing and

$$\tilde{a}_n + \tilde{b}_n \le \tilde{c}_n + \sum_{i=0}^{n-1} \omega_i \tilde{a}_i, \quad \forall\, 1 \le n \le K,$$

then we have

$$\tilde{a}_n + \tilde{b}_n \le \tilde{c}_n \exp\left(\sum_{i=0}^{n-1} \omega_i\right), \quad \forall\, 1 \le n \le K. \tag{4.70}$$

We note that the condition $\omega_0\tilde{a}_0 + \tilde{c}_1 \le \tilde{c}_1 e^{\omega_0}$ in Lemma 4.15 is needed to ensure (4.70) holds for $n = 1$. This condition is satisfied if $\tilde{a}_0 \le \tilde{c}_1$.

Optimal error estimates for the velocity and Lagrange multiplier In what follows, we assume that the solution $(\mathbf{u}, p)$ to the NS equations (4.47) satisfies the following regularity conditions:

$$\begin{aligned} &\mathbf{u} \in L^\infty(0,T;\mathbf{H}^2(\Omega)),\ \mathbf{u}_t \in L^2(0,T;\mathbf{H}^1(\Omega)), \\ &\mathbf{u}_{tt} \in L^2(0,T;\mathbf{H}^{-1}(\Omega)),\ p \in L^2(0,T;H^1(\Omega)/\mathbb{R}). \end{aligned} \tag{4.71}$$

Let M be a positive constant such that

$$\max\left\{ \sup_{t\in[0,T]} \|\mathbf{u}\|_2,\ \int_0^T \|\mathbf{u}_t\|_1^2\, dt,\ \int_0^T \|\mathbf{u}_{tt}\|_{-1}^2\, dt \right\} \le M. \tag{4.72}$$

Theorem 4.16. Let $\{\mathbf{u}^n\}, \{p^n\}$, and $\{q^n\}$ be generated by the first-order DRLM scheme (4.51). There exist positive constants τ_0, C_1, and C_2 depending on Ω, T, ν, M, $\mathbf{u}_0$, and $\mathbf{f}$ but independent of Δt, θ, and n such that the following error estimates hold for all $\Delta t \le \tau_0$:

$$\|e_{\mathbf{u}}^{n+1}\|_0^2 + \sum_{i=0}^n \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^i\|_0^2 + \Delta t \nu \sum_{i=0}^n \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 \le C_1\left(1 + \frac{1}{\theta}\right)\Delta t^2, \quad 0 \le n \le N-1,$$

and

$$|e_q^n| \le \frac{C_2}{\theta}\Delta t, \quad 0 \le n \le N.$$

Proof. To derive optimal error estimates for both the velocity and the Lagrange multiplier, we perform the following steps. A velocity estimate is first obtained based on its error equation (integrated against $e_{\mathbf{u}}^{n+1}$) and the use of the viscosity term to control the truncation error and nonlinear convection terms. Since the resulting estimate involves the Lagrange multiplier error, we bound it in terms of the velocity error. This is a challenging task due to the presence of both q^{n+1} and $(q^{n+1})^2$ in the DRLM scheme and requires combining the dynamic equation for q with the momentum equation and the velocity error equation, tested with $\mathbf{u}^{n+1}-\mathbf{u}^n$ and $e_{\mathbf{u}}^{n+1}-e_{\mathbf{u}}^n$, respectively. Once these estimates are established, the discrete Gronwall's inequality is applied to both the velocity and the Lagrange multiplier, followed by an inductive argument to achieve optimal convergence rates of the concerned quantities under an $\mathcal{O}(1)$ time step size requirement.

Step 1: Estimates for the velocity. Subtracting the momentum equation (4.47a) at t_{n+1} from (4.51a), we derive the error equation for the velocity and pressure:

$$\frac{e_{\mathbf{u}}^{n+1}-e_{\mathbf{u}}^n}{\Delta t}-\nu\Delta e_{\mathbf{u}}^{n+1}+\nabla e_p^{n+1}=\mathbf{R}_{\mathbf{u}}^{n+1}+(\mathbf{u}(t_{n+1})\cdot\nabla)\mathbf{u}(t_{n+1})-q^{n+1}(\mathbf{u}^n\cdot\nabla)\mathbf{u}^n, \quad (4.73)$$

where the truncation error $\mathbf{R}_{\mathbf{u}}^{n+1}$ is given by

$$\mathbf{R}_{\mathbf{u}}^{n+1}=\mathbf{u}_t(t_{n+1})-\frac{\mathbf{u}(t_{n+1})-\mathbf{u}(t_n)}{\Delta t}=\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}(t-t_n)\mathbf{u}_{tt}(t)\,dt.$$

Integrating (4.73) against $e_{\mathbf{u}}^{n+1}$ and noting that $(\nabla e_p^{n+1},e_{\mathbf{u}}^{n+1})=0$ due to $\nabla\cdot e_{\mathbf{u}}^{n+1}=0$ (cf. the incompressibility conditions (4.50b) and (4.51b)), we find

$$\begin{aligned}&\frac{\|e_{\mathbf{u}}^{n+1}\|_0^2-\|e_{\mathbf{u}}^n\|_0^2}{2\Delta t}+\frac{\|e_{\mathbf{u}}^{n+1}-e_{\mathbf{u}}^n\|_0^2}{2\Delta t}+\nu\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2\\&\quad=(\mathbf{R}_{\mathbf{u}}^{n+1},e_{\mathbf{u}}^{n+1})+b(\mathbf{u}(t_{n+1}),\mathbf{u}(t_{n+1}),e_{\mathbf{u}}^{n+1})-q^{n+1}b(\mathbf{u}^n,\mathbf{u}^n,e_{\mathbf{u}}^{n+1}).\end{aligned}\quad (4.74)$$

Since $\mathbf{u}(t_{n+1})\in\mathbf{H}_0^1(\Omega)$ and $\mathbf{u}^{n+1}\in\mathbf{H}_0^1(\Omega)$ by Lemma 4.13, we have $e_{\mathbf{u}}^{n+1}\in\mathbf{H}_0^1(\Omega)$. Hence, the first term on the right-hand side of (4.74) can be bounded by

$$(\mathbf{R}_{\mathbf{u}}^{n+1},e_{\mathbf{u}}^{n+1})\le\left\|\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}(t-t_n)\mathbf{u}_{tt}\,dt\right\|_{-1}\|e_{\mathbf{u}}^{n+1}\|_1\le\left(\int_{t_n}^{t_{n+1}}\|\mathbf{u}_{tt}\|_{-1}dt\right)\|e_{\mathbf{u}}^{n+1}\|_1. \quad (4.75)$$

The last two terms on the right-hand side of (4.74) can be split as follows

$$b(\mathbf{u}(t_{n+1}),\mathbf{u}(t_{n+1}),e_{\mathbf{u}}^{n+1})-q^{n+1}b(\mathbf{u}^n,\mathbf{u}^n,e_{\mathbf{u}}^{n+1})=B_1+B_2+B_3+B_4, \quad (4.76)$$

where

$$
\begin{aligned}
B_1 &= b(\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n), \mathbf{u}(t_{n+1}), e_{\mathbf{u}}^{n+1}) + b(\mathbf{u}(t_n), \mathbf{u}(t_{n+1}) - \mathbf{u}(t_n), e_{\mathbf{u}}^{n+1}),\\
B_2 &= -q^{n+1} b(e_{\mathbf{u}}^{n}, \mathbf{u}(t_n), e_{\mathbf{u}}^{n+1}) - q^{n+1} b(\mathbf{u}(t_n), e_{\mathbf{u}}^{n}, e_{\mathbf{u}}^{n+1}),\\
B_3 &= -q^{n+1} b(e_{\mathbf{u}}^{n}, e_{\mathbf{u}}^{n}, e_{\mathbf{u}}^{n+1}),\\
B_4 &= -e_q^{n+1} b(\mathbf{u}(t_n), \mathbf{u}(t_n), e_{\mathbf{u}}^{n+1}).
\end{aligned}
$$

It follows from properties of the operator b in (4.68)-(4.69) and the regularity assumption (4.72) that

$$
\begin{aligned}
B_1 &= b(\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n), \mathbf{u}(t_{n+1}), e_{\mathbf{u}}^{n+1}) - b(\mathbf{u}(t_n), e_{\mathbf{u}}^{n+1}, \mathbf{u}(t_{n+1}) - \mathbf{u}(t_n))\\
&\le c_2 \|\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n)\|_0 \|\mathbf{u}(t_{n+1})\|_2 \|e_{\mathbf{u}}^{n+1}\|_1 + c_2 \|\mathbf{u}(t_n)\|_2 \|e_{\mathbf{u}}^{n+1}\|_1 \|\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n)\|_0\\
&\le 2c_2 M \|\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n)\|_0 \|e_{\mathbf{u}}^{n+1}\|_1\\
&\le 2c_2 M \left(\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0 \, dt \right) \|e_{\mathbf{u}}^{n+1}\|_1.
\end{aligned}
\tag{4.77}
$$

Similarly, noting that $0 < q^{n+1} \le c_0$ from Lemma 4.14, we obtain

$$
\begin{aligned}
B_2 &= -q^{n+1} b(e_{\mathbf{u}}^{n}, \mathbf{u}(t_n), e_{\mathbf{u}}^{n+1}) + q^{n+1} b(\mathbf{u}(t_n), e_{\mathbf{u}}^{n+1}, e_{\mathbf{u}}^{n})\\
&\le c_0 c_2 \|e_{\mathbf{u}}^{n}\|_0 \|\mathbf{u}(t_n)\|_2 \|e_{\mathbf{u}}^{n+1}\|_1 + c_0 c_2 \|\mathbf{u}(t_n)\|_2 \|e_{\mathbf{u}}^{n+1}\|_1 \|e_{\mathbf{u}}^{n}\|_0\\
&\le 2c_0 c_2 M \|e_{\mathbf{u}}^{n}\|_0 \|e_{\mathbf{u}}^{n+1}\|_1,
\end{aligned}
\tag{4.78}
$$

$$
B_3 = -q^{n+1} b(e_{\mathbf{u}}^{n}, e_{\mathbf{u}}^{n}, e_{\mathbf{u}}^{n+1}) \le c_0 c_2 \|e_{\mathbf{u}}^{n}\|_1 \|e_{\mathbf{u}}^{n}\|_1 \|e_{\mathbf{u}}^{n+1}\|_1 = c_0 c_2 \|e_{\mathbf{u}}^{n}\|_1^2 \|e_{\mathbf{u}}^{n+1}\|_1. \tag{4.79}
$$

We remark that the estimation of B_3 involves $\|e_{\mathbf{u}}^{n}\|_1^2$, which eventually becomes $\|e_{\mathbf{u}}^{n}\|_1^4$ after the use of Young's inequality. Alternatively, one may estimate B_3 using a stronger inequality $B_3 \le c_0 c_2 \|e_{\mathbf{u}}^{n}\|_0^{1/2} \|e_{\mathbf{u}}^{n}\|_1^{1/2} \|e_{\mathbf{u}}^{n}\|_0^{1/2} \|e_{\mathbf{u}}^{n}\|_1^{1/2} \|e_{\mathbf{u}}^{n+1}\|_1$, which simplifies to $B_3 \le c_0 c_2 \|e_{\mathbf{u}}^{n}\|_0 \|e_{\mathbf{u}}^{n}\|_1 \|e_{\mathbf{u}}^{n+1}\|_1$ and (4.79) follows. Although this estimate avoids the difficulty associated with the quartic term $\|e_{\mathbf{u}}^{n}\|_1^4$, it is only valid in two dimensions ($d = 2$), as shown in [124] (where the study is restricted to the 2D setting). To ensure the analysis in the current paper applies to both two and three dimensions, we still use (4.79) for estimating B_3. Now for B_4, it can be bounded by

$$B_4 = -e_q^{n+1} b(\mathbf{u}(t_n), \mathbf{u}(t_n), e_{\mathbf{u}}^{n+1}) \leq c_2 |e_q^{n+1}| \|\mathbf{u}(t_n)\|_2 \|\mathbf{u}(t_n)\|_1 \|e_{\mathbf{u}}^{n+1}\|_0 \leq c_2 M^2 |e_q^{n+1}| \|e_{\mathbf{u}}^{n+1}\|_0. \tag{4.80}$$

In addition to B_3, the estimate for B_4 introduces another challenge to the analysis due to the product of implicit error terms, i.e., $|e_q^{n+1}|\|e_{\mathbf{u}}^{n+1}\|_0$. We remark that for the first-order SAV scheme in [124], a term similar to B_4 (involving both $e_{\mathbf{u}}^{n+1}$ and e_{SAV}^{n+1}) can be easily handled by invoking the dynamic equation for the auxiliary variable. However, applying the same approach to B_4 in our case leads to suboptimal error estimates for both the velocity and the Lagrange multiplier.

The combination of (4.74)-(4.80) results in

$$\begin{aligned}
&\frac{\|e_{\mathbf{u}}^{n+1}\|_0^2 - \|e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \frac{\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \nu \|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 \\
&\quad \leq \left(\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1} \, dt + 2c_2 M \int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0 \, dt + 2c_0 c_2 M \|e_{\mathbf{u}}^{n}\|_0 + c_0 c_2 \|e_{\mathbf{u}}^{n}\|_1^2 \right) \|e_{\mathbf{u}}^{n+1}\|_1 \\
&\qquad + c_2 M^2 |e_q^{n+1}| \|e_{\mathbf{u}}^{n+1}\|_0.
\end{aligned} \tag{4.81}$$

By the Young's and Cauchy-Schwarz inequalities, noting that $\|e_{\mathbf{u}}^{n+1}\|_1 \leq (1 + c_1)\|\nabla e_{\mathbf{u}}^{n+1}\|_0$ due to (4.67), the first term on the right-hand side of (4.81) is bounded by

$$
\begin{aligned}
&\left(\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}\,dt + 2c_2M\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0\,dt + 2c_0c_2M\|e_{\mathbf{u}}^n\|_0 + c_0c_2\|e_{\mathbf{u}}^n\|_1^2\right)\|e_{\mathbf{u}}^{n+1}\|_1 \\
&\quad\le \frac{(1+c_1)^2}{2\nu}\left(\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}\,dt + 2c_2M\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0\,dt + 2c_0c_2M\|e_{\mathbf{u}}^n\|_0 + c_0c_2\|e_{\mathbf{u}}^n\|_1^2\right)^2 \\
&\qquad + \frac{\nu}{2(1+c_1)^2}\|e_{\mathbf{u}}^{n+1}\|_1^2 \\
&\quad\le \frac{(1+c_1)^2}{2\nu}\left(1+4c_2^2M^2+4c_0^2c_2^2M^2+c_0^2c_2^2\right)\Bigg[\left(\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}\,dt\right)^2 + \left(\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0\,dt\right)^2 \\
&\qquad + \|e_{\mathbf{u}}^n\|_0^2 + \|e_{\mathbf{u}}^n\|_1^4\Bigg] + \frac{\nu}{2}\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 \\
&\quad\le c_3\left(\Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}^2\,dt + \Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0^2\,dt + \|e_{\mathbf{u}}^n\|_0^2 + \|e_{\mathbf{u}}^n\|_1^4\right) + \frac{\nu}{2}\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2,
\end{aligned}
\tag{4.82}
$$

where $c_3 = \frac{(1+c_1)^2}{2\nu}\left(1+4c_2^2M^2+4c_0^2c_2^2M^2+c_0^2c_2^2\right)$. It follows from (4.81) and (4.82) that

$$
\begin{aligned}
&\frac{\|e_{\mathbf{u}}^{n+1}\|_0^2 - \|e_{\mathbf{u}}^n\|_0^2}{2\Delta t} + \frac{\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^n\|_0^2}{2\Delta t} + \frac{\nu}{2}\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 \\
&\quad\le c_3\left(\Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}^2\,dt + \Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0^2\,dt + \|e_{\mathbf{u}}^n\|_0^2 + \|e_{\mathbf{u}}^n\|_1^4\right) + c_2M^2|e_q^{n+1}|\|e_{\mathbf{u}}^{n+1}\|_0.
\end{aligned}
\tag{4.83}
$$

Summing (4.83) over $i = 0, 1, \dots, n$ and multiplying the resulting estimate by $2\Delta t$, we deduce that

$$
\begin{aligned}
&\|e_{\mathbf{u}}^{n+1}\|_0^2 + \sum_{i=0}^n \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^i\|_0^2 + \Delta t\nu\sum_{i=0}^n \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 \\
&\quad\le 2c_3\Delta t\left(\Delta t\int_0^T \|\mathbf{u}_{tt}\|_{-1}^2\,dt + \Delta t\int_0^T \|\mathbf{u}_t\|_0^2\,dt + \sum_{i=0}^n \|e_{\mathbf{u}}^i\|_0^2 + \sum_{i=0}^n \|e_{\mathbf{u}}^i\|_1^4\right) \\
&\qquad + 2c_2M^2\Delta t\sum_{i=0}^n |e_q^{i+1}|\|e_{\mathbf{u}}^{i+1}\|_0 \\
&\quad\le 2c_3\Delta t\left(2M\Delta t + \sum_{i=0}^n \|e_{\mathbf{u}}^i\|_0^2 + \sum_{i=0}^n \|e_{\mathbf{u}}^i\|_1^4\right) + 2c_2M^2\Delta t\sum_{i=0}^n |e_q^{i+1}|\|e_{\mathbf{u}}^{i+1}\|_0,
\end{aligned}
\tag{4.84}
$$

where we have used the regularity assumption (4.72) in the last estimate. By Young's inequality, the last summation on the right-hand side of (4.84) is bounded by

$$
\begin{aligned}
2c_2M^2\Delta t\sum_{i=0}^{n}|e_q^{i+1}|\|e_{\mathbf{u}}^{i+1}\|_0 &= 2c_2M^2\Delta t|e_q^{n+1}|\|e_{\mathbf{u}}^{n+1}\|_0 + 2c_2M^2\Delta t\sum_{i=0}^{n}|e_q^{i}|\|e_{\mathbf{u}}^{i}\|_0 \\
&\le \frac{1}{2}\|e_{\mathbf{u}}^{n+1}\|_0^2 + 2c_2^2M^4\Delta t^2|e_q^{n+1}|^2 + c_2M^2\Delta t\sum_{i=0}^{n}|e_q^{i}|^2 + c_2M^2\Delta t\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_0^2.
\end{aligned}
\tag{4.85}
$$

From (4.84) and (4.85), we arrive at

$$
\begin{aligned}
&\frac{1}{2}\|e_{\mathbf{u}}^{n+1}\|_0^2 + \sum_{i=0}^{n}\|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_0^2 + \Delta t\nu\sum_{i=0}^{n}\|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 \\
&\le 4c_3M\Delta t^2 + (2c_3 + c_2M^2)\Delta t\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_0^2 + 2c_3\Delta t\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_1^4 \\
&\quad + c_2M^2\Delta t\sum_{i=0}^{n}|e_q^{i}|^2 + 2c_2^2M^4\Delta t^2|e_q^{n+1}|^2.
\end{aligned}
\tag{4.86}
$$

Step 2: Estimates for the Lagrange multiplier. Since the right-hand side of (4.86) involves the error term $|e_q^{n+1}|^2$ of the Lagrange multiplier, we proceed by estimating this term based on the dynamic equation (4.51c). This can be achieved by bounding $\left|(q^{n+1})^2 - 1\right|$, noting that the exact value of q is 1 and $q^0 = 1$. Observe that (4.51c) can be written as follows

$$
\theta\frac{(q^{n+1})^2 - (q^n)^2}{\Delta t} = \frac{\|\mathbf{u}^{n+1} - \mathbf{u}^n\|_0^2}{2\Delta t} + q^{n+1}b(\mathbf{u}^n, \mathbf{u}^n, \mathbf{u}^{n+1}).
\tag{4.87}
$$

Taking the $\mathbf{L}^2$ inner product of (4.51a) with $\mathbf{u}^{n+1} - \mathbf{u}^n$ and using the properties $b(\mathbf{u}^n, \mathbf{u}^n, \mathbf{u}^n) = 0$ and $(\nabla p^{n+1}, \mathbf{u}^{n+1} - \mathbf{u}^n) = 0$, we find

$$
\begin{aligned}
&\frac{\|\mathbf{u}^{n+1} - \mathbf{u}^n\|_0^2}{\Delta t} + \frac{\nu}{2}\left(\|\nabla\mathbf{u}^{n+1}\|_0^2 - \|\nabla\mathbf{u}^n\|_0^2 + \|\nabla(\mathbf{u}^{n+1} - \mathbf{u}^n)\|_0^2\right) + q^{n+1}b(\mathbf{u}^n, \mathbf{u}^n, \mathbf{u}^{n+1}) \\
&\qquad = \left(\mathbf{f}(t_{n+1}), \mathbf{u}^{n+1} - \mathbf{u}^n\right).
\end{aligned}
\tag{4.88}
$$

It follows from (4.87) and (4.88) that

$$
\begin{aligned}
\theta\frac{(q^{n+1})^2 - (q^n)^2}{\Delta t} &= \left(\mathbf{f}(t_{n+1}), \mathbf{u}^{n+1} - \mathbf{u}^n\right) - \frac{\|\mathbf{u}^{n+1} - \mathbf{u}^n\|_0^2}{2\Delta t} \\
&\quad - \frac{\nu}{2}\left(\|\nabla\mathbf{u}^{n+1}\|_0^2 - \|\nabla\mathbf{u}^n\|_0^2 + \|\nabla(\mathbf{u}^{n+1} - \mathbf{u}^n)\|_0^2\right).
\end{aligned}
\tag{4.89}
$$

Notice that $q^0 = 1$, the sum of (4.89) over $i = 0, 1, \dots, n$ leads to

$$\theta \frac{(q^{n+1})^2 - 1}{\Delta t} = \sum_{i=0}^{n} \left(\mathbf{f}(t_{i+1}), \mathbf{u}^{i+1} - \mathbf{u}^i\right) - \frac{1}{2\Delta t} \sum_{i=0}^{n} \|\mathbf{u}^{i+1} - \mathbf{u}^i\|_0^2 \\ - \frac{\nu}{2}(\|\nabla \mathbf{u}^{n+1}\|_0^2 - \|\nabla \mathbf{u}^0\|_0^2) - \frac{\nu}{2} \sum_{i=0}^{n} \|\nabla(\mathbf{u}^{i+1} - \mathbf{u}^i)\|_0^2. \tag{4.90}$$

Recalling $c_{\mathbf{f}} = \|\mathbf{f}\|_{L^\infty(0,T;\mathbf{L}^2(\Omega))}$, the first term on the right-hand side of (4.90) can be bounded by

$$\begin{aligned} \left|\sum_{i=0}^{n} \left(\mathbf{f}(t_{i+1}), \mathbf{u}^{i+1} - \mathbf{u}^i\right)\right| &\le \sum_{i=0}^{n} \|\mathbf{f}(t_{i+1})\|_0 \|\mathbf{u}^{i+1} - \mathbf{u}^i\|_0 \\ &\le \sum_{i=0}^{n} c_{\mathbf{f}} \sqrt{\Delta t} \cdot \frac{1}{\sqrt{\Delta t}} \|\mathbf{u}^{i+1} - \mathbf{u}^i\|_0 \\ &\le \frac{1}{2} \sum_{i=0}^{n} c_{\mathbf{f}}^2 \Delta t + \frac{1}{2\Delta t} \sum_{i=0}^{n} \|\mathbf{u}^{i+1} - \mathbf{u}^i\|_0^2 \\ &\le \frac{1}{2} c_{\mathbf{f}}^2 T + \frac{1}{2\Delta t} \sum_{i=0}^{n} \|\mathbf{u}^{i+1} - \mathbf{u}^i\|_0^2. \end{aligned} \tag{4.91}$$

Observing that $\mathbf{u}^{i+1} - \mathbf{u}^i = (\mathbf{u}(t_{i+1}) - \mathbf{u}(t_i)) + (e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^i)$ and using the inequality $\|\mathbf{u} + \mathbf{v}\|_0^2 \le 2(\|\mathbf{u}\|_0^2 + \|\mathbf{v}\|_0^2)$, we can estimate the second term on the right-hand side of (4.90) (as well as (4.91)) as follows

$$\begin{aligned} \frac{1}{2\Delta t} \sum_{i=0}^{n} \|\mathbf{u}^{i+1} - \mathbf{u}^i\|_0^2 &\le \frac{1}{\Delta t} \sum_{i=0}^{n} \|\mathbf{u}(t_{i+1}) - \mathbf{u}(t_i)\|_0^2 + \frac{1}{\Delta t} \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^i\|_0^2 \\ &\le \frac{1}{\Delta t} \sum_{i=0}^{n} \left(\int_{t_i}^{t_{i+1}} \|\mathbf{u}_t\|_0 \, dt\right)^2 + \frac{1}{\Delta t} \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^i\|_0^2 \\ &\le \int_0^T \|\mathbf{u}_t\|_0^2 \, dt + \frac{1}{\Delta t} \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^i\|_0^2 \\ &\le M + \frac{1}{\Delta t} \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^i\|_0^2. \end{aligned} \tag{4.92}$$

Similarly, for the term $\frac{\nu}{2}\|\nabla \mathbf{u}^{n+1}\|_0^2$ and the last summation on the right-hand side of (4.90), we have

$$\frac{\nu}{2}\|\nabla\mathbf{u}^{n+1}\|_0^2 \le \nu\|\nabla\mathbf{u}(t_{n+1})\|_0^2 + \nu\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 \\ \le \nu M^2 + \nu\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2, \tag{4.93}$$

and

$$\begin{aligned}\frac{\nu}{2}\sum_{i=0}^{n}\|\nabla(\mathbf{u}^{i+1}-\mathbf{u}^i)\|_0^2 &\le \nu\sum_{i=0}^{n}\|\nabla(\mathbf{u}(t_{i+1})-\mathbf{u}(t_i))\|_0^2 + \nu\sum_{i=0}^{n}\|\nabla(e_{\mathbf{u}}^{i+1}-e_{\mathbf{u}}^i)\|_0^2\\ &\le \nu\Delta t\int_0^T\|\nabla\mathbf{u}_t\|_0^2\,dt + \nu\sum_{i=0}^{n}\|\nabla(e_{\mathbf{u}}^{i+1}-e_{\mathbf{u}}^i)\|_0^2\\ &\le \nu M\Delta t + \nu\sum_{i=0}^{n}\|\nabla(e_{\mathbf{u}}^{i+1}-e_{\mathbf{u}}^i)\|_0^2.\end{aligned} \tag{4.94}$$

Combining (4.90)-(4.94), we obtain

$$\begin{aligned}\theta\frac{|(q^{n+1})^2-1|}{\Delta t} \le{}& \left|\sum_{i=0}^{n}\left(\mathbf{f}(t_{i+1}),\mathbf{u}^{i+1}-\mathbf{u}^i\right)\right| + \frac{1}{2\Delta t}\sum_{i=0}^{n}\|\mathbf{u}^{i+1}-\mathbf{u}^i\|_0^2\\ &+\frac{\nu}{2}\left(\|\nabla\mathbf{u}^{n+1}\|_0^2+\|\nabla\mathbf{u}^0\|_0^2\right)+\frac{\nu}{2}\sum_{i=0}^{n}\|\nabla(\mathbf{u}^{i+1}-\mathbf{u}^i)\|_0^2\\ \le{}& \frac{1}{2}c_{\mathbf{f}}^2T + 2M + \nu M^2 + \frac{\nu}{2}\|\nabla\mathbf{u}^0\|_0^2 + \nu M\Delta t\\ &+\frac{2}{\Delta t}\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i+1}-e_{\mathbf{u}}^i\|_0^2 + \nu\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 + \nu\sum_{i=0}^{n}\|\nabla(e_{\mathbf{u}}^{i+1}-e_{\mathbf{u}}^i)\|_0^2.\end{aligned} \tag{4.95}$$

Notice that $e_q^{n+1} = q^{n+1} - 1 \ge -1$, which implies

$$\begin{aligned}\theta\frac{|(q^{n+1})^2-1|}{\Delta t} &= \theta\frac{|(e_q^{n+1}+1)^2-1|}{\Delta t}\\ &= \theta\frac{|(e_q^{n+1})^2+2e_q^{n+1}|}{\Delta t}\\ &\ge \theta\frac{\max\{|e_q^{n+1}|^2,|e_q^{n+1}|\}}{\Delta t}.\end{aligned} \tag{4.96}$$

In the last estimate of (4.96), we have used the fact that $|x+2| \ge \max\{|x|,1\}$ for all $x \ge -1$. From (4.95) and (4.96), we deduce

$$\theta\frac{\max\{|e_q^{n+1}|^2,|e_q^{n+1}|\}}{\Delta t} \le c_4 + \nu M\Delta t + S^{n+1}, \tag{4.97}$$

where $c_4 = \frac{1}{2}c_{\mathbf{f}}^2 T + 2M + \nu M^2 + \frac{\nu}{2}\|\nabla \mathbf{u}^0\|_0^2$ and

$$S^{n+1} = \frac{2}{\Delta t}\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_0^2 + \nu\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 + \nu \sum_{i=0}^{n} \|\nabla(e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i})\|_0^2.$$

Step 3: Estimates for S^{n+1}. By taking the $\mathbf{L}^2$ inner product of (4.73) with $e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}$ and using similar arguments as in the estimates (4.75)-(4.82), we obtain

$$\begin{aligned}
&\frac{\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0^2}{\Delta t} + \frac{\nu}{2}\left(\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 - \|\nabla e_{\mathbf{u}}^{n}\|_0^2 + \|\nabla(e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n})\|_0^2\right) \\
&\quad = (\mathbf{R}_{\mathbf{u}}^{n+1}, e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}) + b(\mathbf{u}(t_{n+1}), \mathbf{u}(t_{n+1}), e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}) - q^{n+1} b(\mathbf{u}^n, \mathbf{u}^n, e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}) \\
&\quad \leq \left(\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}\, dt + 2c_2 M \int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0\, dt + 2c_0c_2M\|e_{\mathbf{u}}^{n}\|_0 + c_0c_2\|e_{\mathbf{u}}^{n}\|_1^2\right)\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_1 \\
&\qquad + c_2M^2|e_q^{n+1}|\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0 \\
&\quad \leq 2c_3\left(\Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}^2\, dt + \Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0^2\, dt + \|e_{\mathbf{u}}^{n}\|_0^2 + \|e_{\mathbf{u}}^{n}\|_1^4\right) + \frac{\nu}{4}\|\nabla(e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n})\|_0^2 \\
&\qquad + \frac{\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \frac{c_2^2M^4\Delta t}{2}|e_q^{n+1}|^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
&\frac{\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \frac{\nu}{2}\left(\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 - \|\nabla e_{\mathbf{u}}^{n}\|_0^2\right) + \frac{\nu}{4}\|\nabla(e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n})\|_0^2 \\
&\quad \leq 2c_3\left(\Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}^2\, dt + \Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0^2\, dt + \|e_{\mathbf{u}}^{n}\|_0^2 + \|e_{\mathbf{u}}^{n}\|_1^4\right) + \frac{c_2^2M^4\Delta t}{2}|e_q^{n+1}|^2.
\end{aligned} \tag{4.98}$$

The sum of (4.98) over $i = 0, 1, \ldots, n$ gives us

$$\begin{aligned}
&\frac{1}{2\Delta t}\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_0^2 + \frac{\nu}{2}\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 + \frac{\nu}{4}\sum_{i=0}^{n} \|\nabla(e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i})\|_0^2 \\
&\quad \leq 2c_3\left(2M\Delta t + \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_0^2 + \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^4\right) + \frac{c_2^2M^4\Delta t}{2}\sum_{i=0}^{n} |e_q^{i+1}|^2.
\end{aligned} \tag{4.99}$$

By multiplying both sides of (4.99) by 4, we can bound S^{n+1} as follows

$$\begin{aligned} S^{n+1} &\leq \frac{2}{\Delta t}\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i+1}-e_{\mathbf{u}}^{i}\|_0^2 + 2\nu\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 + \nu\sum_{i=0}^{n}\|\nabla(e_{\mathbf{u}}^{i+1}-e_{\mathbf{u}}^{i})\|_0^2 \\ &\leq 8c_3\left(2M\Delta t + \sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_0^2 + \sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_1^4\right) + 2c_2^2M^4\Delta t\sum_{i=0}^{n}|e_q^{i+1}|^2. \end{aligned} \tag{4.100}$$

Step 4: Simultaneous estimates for the velocity and Lagrange multiplier. In light of Step 2 and Step 3, let us first simplify the estimate for $\theta|e_q^{n+1}|^2$. From (4.97) and (4.100), we find

$$\begin{aligned} &\theta\frac{\max\{|e_q^{n+1}|^2, |e_q^{n+1}|\}}{\Delta t} \leq c_4 + \nu M\Delta t + S^{n+1} \\ &\leq c_4 + (16c_3+\nu)M\Delta t + 8c_3\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_0^2 + 8c_3\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_1^4 + 2c_2^2M^4\Delta t\sum_{i=0}^{n}|e_q^{i+1}|^2. \end{aligned} \tag{4.101}$$

Since $\frac{x+y}{2} \leq \max\{x,y\}$ for all $x, y \in \mathbb{R}$, multiplying (4.101) by $2\Delta t$ yields

$$\begin{aligned} \theta|e_q^{n+1}|^2 + \theta|e_q^{n+1}| &\leq 2c_4\Delta t + (32c_3+2\nu)M\Delta t^2 + 16c_3\Delta t\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_0^2 \\ &\quad + 16c_3\Delta t\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_1^4 + 4c_2^2M^4\Delta t^2\sum_{i=0}^{n}|e_q^{i+1}|^2. \end{aligned} \tag{4.102}$$

If $\theta|e_q^{n+1}| > 2c_4\Delta t$ then (4.102) reduces to

$$\theta|e_q^{n+1}|^2 \leq (32c_3+2\nu)M\Delta t^2 + 16c_3\Delta t\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_0^2 + 16c_3\Delta t\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_1^4 + 4c_2^2M^4\Delta t^2\sum_{i=0}^{n}|e_q^{i+1}|^2.$$

If $\theta|e_q^{n+1}| \leq 2c_4\Delta t$ then $\theta|e_q^{n+1}|^2 \leq \frac{4c_4^2}{\theta}\Delta t^2$. In both cases, we have

$$\begin{aligned} \theta|e_q^{n+1}|^2 &\leq \left(32c_3M + 2\nu M + \frac{4c_4^2}{\theta}\right)\Delta t^2 + 16c_3\Delta t\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_0^2 \\ &\quad + 16c_3\Delta t\sum_{i=0}^{n}\|e_{\mathbf{u}}^{i}\|_1^4 + 4c_2^2M^4\Delta t^2\sum_{i=0}^{n}|e_q^{i+1}|^2. \end{aligned} \tag{4.103}$$

Taking the sum of (4.86) and (4.103), we obtain

$$
\begin{aligned}
&\frac{1}{2}\|e_{\mathbf{u}}^{n+1}\|_0^2 + \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_0^2 + \Delta t\nu \sum_{i=0}^{n} \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 + \theta|e_q^{n+1}|^2 \\
&\quad \le \left(36c_3M + 2\nu M + \frac{4c_4^2}{\theta}\right)\Delta t^2 + (18c_3 + c_2M^2)\Delta t \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_0^2 + 18c_3\Delta t \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^4 \qquad (4.104) \\
&\qquad + (c_2M^2 + 4c_2^2M^4\Delta t)\Delta t \sum_{i=0}^{n} |e_q^{i}|^2 + 6c_2^2M^4\Delta t^2|e_q^{n+1}|^2.
\end{aligned}
$$

If Δt satisfies $6c_2^2M^4\Delta t^2 \le \frac{\theta}{2}$, or

$$
\Delta t \le \frac{\sqrt{\theta}}{2\sqrt{3}c_2M^2}, \tag{4.105}
$$

then multiplying (4.104) by 2 yields

$$
\begin{aligned}
&\|e_{\mathbf{u}}^{n+1}\|_0^2 + 2\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_0^2 + 2\Delta t\nu \sum_{i=0}^{n} \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 + \theta|e_q^{n+1}|^2 \\
&\quad \le \left(72c_3M + 4\nu M + \frac{8c_4^2}{\theta}\right)\Delta t^2 + (36c_3 + 2c_2M^2)\Delta t \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_0^2 + 36c_3\Delta t \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^4 \\
&\qquad + (2c_2M^2 + 8c_2^2M^4\Delta t)\Delta t \sum_{i=0}^{n} |e_q^{i}|^2 \qquad (4.106) \\
&\quad \le \max\left\{36c_3 + 2c_2M^2, \frac{2c_2M^2 + 8c_2^2M^4\Delta t}{\theta}\right\}\Delta t \sum_{i=0}^{n} \left(\|e_{\mathbf{u}}^{i}\|_0^2 + \theta|e_q^{i}|^2\right) \\
&\qquad + \left(72c_3M + 4\nu M + \frac{8c_4^2}{\theta}\right)\Delta t^2 + 36c_3\Delta t \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^4.
\end{aligned}
$$

From (4.105) and the fact $\theta \ge 1$, we know that

$$
\frac{2c_2M^2 + 8c_2^2M^4\Delta t}{\theta} = 2c_2M^2\frac{1 + 4c_2M^2\Delta t}{\theta} \le 2c_2M^2\frac{1 + \frac{2\sqrt{\theta}}{\sqrt{3}}}{\theta} \le 2c_2M^2\left(1 + \frac{2}{\sqrt{3}}\right) \le 5c_2M^2.
$$

Therefore, (4.106) can be simplified as

$$
\begin{aligned}
&\|e_{\mathbf{u}}^{n+1}\|_0^2 + 2\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_0^2 + 2\Delta t\nu \sum_{i=0}^{n} \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 + \theta|e_q^{n+1}|^2 \\
&\le (36c_3 + 5c_2M^2)\Delta t \sum_{i=0}^{n} \left(\|e_{\mathbf{u}}^{i}\|_0^2 + \theta|e_q^{i}|^2\right) + \left(72c_3M + 4\nu M + \frac{8c_4^2}{\theta}\right)\Delta t^2 + 36c_3\Delta t \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^4.
\end{aligned} \tag{4.107}
$$

Applying the discrete Gronwall's inequality in Lemma 4.15 to (4.107), with $\tilde{a}_n = \|e_{\mathbf{u}}^{n}\|_0^2 + \theta|e_q^{n}|^2$

for $0 \leq n \leq N$, we obtain

$$
\begin{aligned}
&\|e_{\mathbf{u}}^{n+1}\|_0^2 + 2\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_0^2 + 2\Delta t\nu \sum_{i=0}^{n} \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 + \theta |e_q^{n+1}|^2 \\
&\leq \left[\left(72c_3M + 4\nu M + \frac{8c_4^2}{\theta}\right)\Delta t^2 + 36c_3\Delta t \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^4\right] \exp\left(T(36c_3 + 5c_2M^2)\right) \qquad (4.108)\\
&\leq \left(2c_5M + \frac{c_6}{\theta}\right)\Delta t^2 + c_5\Delta t \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^4,
\end{aligned}
$$

where $c_5 = (36c_3 + 2\nu)\exp\left(T(36c_3 + 5c_2M^2)\right)$ and $c_6 = 8c_4^2 \exp\left(T(36c_3 + 5c_2M^2)\right)$.

Step 5: Estimates for $\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^4$ by induction. It follows from (4.108) and Poincaré inequality (4.67) that

$$
\frac{2\Delta t\nu}{(1+c_1)^2} \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1}\|_1^2 \leq 2\Delta t\nu \sum_{i=0}^{n} \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 \leq \left(2c_5M + \frac{c_6}{\theta}\right)\Delta t^2 + c_5\Delta t \left(\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^2\right)^2,
$$

which leads to

$$
\sum_{i=0}^{n+1} \|e_{\mathbf{u}}^{i}\|_1^2 \leq \frac{(1+c_1)^2}{2\nu}\left[2c_5M + \frac{c_6}{\theta} + \frac{c_5}{\Delta t}\left(\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^2\right)^2\right]\Delta t. \qquad (4.109)
$$

Using (4.109), we will prove by induction that, for sufficiently small Δt, the following holds for $0 \leq n \leq N$:

$$
\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^2 \leq \frac{(1+c_1)^2}{2\nu}\left(2c_5M + \frac{c_6}{\theta} + c_5\right)\Delta t =: \alpha\Delta t. \qquad (4.110)
$$

Clearly, (4.110) holds for $n = 0$ since $e_{\mathbf{u}}^0 = \mathbf{0}$. Suppose it is true for some nonnegative integer $n < N$. In light of (4.109), for Δt satisfying

$$
\Delta t \leq \frac{1}{\alpha^2} = \frac{4\nu^2}{(1+c_1)^4\left(c_5 + 2c_5M + \frac{c_6}{\theta}\right)^2}, \qquad (4.111)
$$

we find that

$$
\sum_{i=0}^{n+1} \|e_{\mathbf{u}}^{i}\|_1^2 \leq \frac{(1+c_1)^2}{2\nu}\left(2c_5M + \frac{c_6}{\theta} + c_5\alpha^2\Delta t\right)\Delta t \leq \frac{(1+c_1)^2}{2\nu}\left(2c_5M + \frac{c_6}{\theta} + c_5\right)\Delta t = \alpha\Delta t,
$$

which verifies the statement (4.110), provided that $\Delta t \leq \frac{1}{\alpha^2}$. Therefore, we have the following estimate

$$\sum_{i=0}^{N} \|e_{\mathbf{u}}^i\|_1^4 \leq \left(\sum_{i=0}^{N} \|e_{\mathbf{u}}^i\|_1^2\right)^2 \leq \alpha^2 \Delta t^2 \leq \Delta t. \tag{4.112}$$

Step 6: Optimal convergence rates for the velocity and Lagrange multiplier. Combining (4.108) and (4.112), we end up with

$$\|e_{\mathbf{u}}^{n+1}\|_0^2 + 2\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^i\|_0^2 + 2\Delta t \nu \sum_{i=0}^{n} \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 + \theta |e_q^{n+1}|^2 \leq \left(c_5 + 2c_5 M + \frac{c_6}{\theta}\right)\Delta t^2, \tag{4.113}$$

provided that Δt satisfies the following condition (cf. (4.105) and (4.111) and the fact $\theta \geq 1$)

$$\Delta t \leq \min\left\{\frac{1}{2\sqrt{3}c_2 M^2}, \frac{4\nu^2}{(1+c_1)^4 (c_5 + 2c_5 M + c_6)^2}, T\right\} =: \tau_0.$$

Setting $c_7 = \max\{c_5 + 2c_5 M, c_6\}$, we obtain from (4.113) that for $0 \leq n < N$,

$$\|e_{\mathbf{u}}^{n+1}\|_0^2 + 2\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^i\|_0^2 + 2\Delta t \nu \sum_{i=0}^{n} \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 \leq c_7 \left(1 + \frac{1}{\theta}\right)\Delta t^2, \tag{4.114}$$

and

$$|e_q^{n+1}|^2 \leq c_7 \left(\frac{1}{\theta} + \frac{1}{\theta^2}\right)\Delta t^2. \tag{4.115}$$

In light of (4.112), (4.114), and (4.115), the bound for $|e_q^{n+1}|$ can be further improved (especially when $\theta \gg 1$) using (4.101) as follows

$$\begin{aligned}
\theta \frac{|e_q^{n+1}|}{\Delta t} &\leq c_4 + (16c_3 + \nu) M \Delta t + 8c_3 \sum_{i=0}^{n} \|e_{\mathbf{u}}^i\|_0^2 + 8c_3 \sum_{i=0}^{n} \|e_{\mathbf{u}}^i\|_1^4 + 2c_2^2 M^4 \Delta t \sum_{i=0}^{n} |e_q^{i+1}|^2 \\
&\leq c_4 + (16c_3 + \nu) M \Delta t + 8c_3 c_7 \left(1 + \frac{1}{\theta}\right) T \Delta t + 8c_3 \Delta t + 2c_2^2 M^4 c_7 \left(\frac{1}{\theta} + \frac{1}{\theta^2}\right) T \Delta t^2 \\
&\leq c_4 + c_8 \Delta t + c_9 \Delta t^2,
\end{aligned}$$

where $c_8 = (16c_3 + \nu)M + 16c_3 c_7 T + 8c_3$ and $c_9 = 4c_2^2 M^4 c_7 T$. This means that

$$|e_q^{n+1}| \leq \frac{c_4 + c_8 \Delta t + c_9 \Delta t^2}{\theta} \Delta t,$$

and the proof of Theorem 4.16 is completed with $C_1 = c_7$ and $C_2 = c_4 + c_8 \tau_0 + c_9 \tau_0^2$. □

Remark 4.17. It is necessary to follow Steps 2-6 in the proof above to bound the Lagrange multiplier error e_q^{n+1} from both above and below, especially the latter. Note that an upper bound for e_q^{n+1} can be easily obtained by combining (4.90) and (4.91) which yields

$$\begin{aligned} \theta\frac{(q^{n+1})^2-1}{\Delta t} &\leq \frac{1}{2}c_{\mathbf{f}}^2 T + \frac{\nu}{2}\|\nabla \mathbf{u}^0\|_0^2 - \frac{\nu}{2}\|\nabla \mathbf{u}^{n+1}\|_0^2 - \frac{\nu}{2}\sum_{i=0}^{n}\|\nabla(\mathbf{u}^{i+1}-\mathbf{u}^i)\|_0^2 \\ &\leq \frac{1}{2}c_{\mathbf{f}}^2 T + \frac{\nu}{2}\|\nabla \mathbf{u}^0\|_0^2 =: \hat{c}. \end{aligned} \tag{4.116}$$

Recall that $q^{n+1}-1 = e_q^{n+1}$ and $q^{n+1}+1>1$, (4.116) implies

$$\theta\frac{e_q^{n+1}}{\Delta t} \leq \hat{c}, \text{ or equivalently, } e_q^{n+1} \leq \frac{\hat{c}}{\theta}\Delta t.$$

Optimal error estimates for the pressure We begin with a key lemma on the discrete derivative of the Lagrange multiplier error, which is crucial for the pressure error estimate. For any sequence $\{g^n\}_{n\geq 0}$, let us define

$$d_t g^{n+1} := \frac{g^{n+1}-g^n}{\Delta t}, \quad \text{for any } n \geq 0.$$

Lemma 4.18. Under the regularity assumptions (4.71) and $\mathbf{u}_t \in L^4(0,T;\mathbf{L}^2(\Omega))$, there exists a positive constant C_3 independent of Δt and θ such that for all $\Delta t \leq \min\{\tau_0, \frac{\theta}{2C_2}\}$, we have

$$\sum_{n=0}^{N-1}|d_t e_q^{n+1}|^2 \leq \frac{C_3}{\theta^2}\Delta t,$$

where the constants τ_0 and C_2 are defined in Theorem 4.16.

Proof. Let $\widetilde{M}$ be a constant such that $\int_0^T \|\mathbf{u}_t\|_0^4\,dt \leq \widetilde{M}$. Since $q^{n+1}-q^n = e_q^{n+1}-e_q^n$, it follows from (4.89) that

$$\theta(q^{n+1}+q^n)d_t e_q^{n+1} = \big(\mathbf{f}(t_{n+1}), \mathbf{u}^{n+1}-\mathbf{u}^n\big) - \frac{\|\mathbf{u}^{n+1}-\mathbf{u}^n\|_0^2}{2\Delta t} - \nu(\nabla\mathbf{u}^{n+1}, \nabla(\mathbf{u}^{n+1}-\mathbf{u}^n)). \tag{4.117}$$

Recalling that $c_{\mathbf{f}} = \|\mathbf{f}\|_{L^\infty(0,T;\mathbf{L}^2(\Omega))}$, the first term on the right-hand side of (4.117) is bounded by

$$
\begin{aligned}
\left|\left(\mathbf{f}(t_{n+1}), \mathbf{u}^{n+1}-\mathbf{u}^n\right)\right| &\leq \|\mathbf{f}(t_{n+1})\|_0 \|\mathbf{u}^{n+1}-\mathbf{u}^n\|_0 \\
&\leq c_{\mathbf{f}} \|\mathbf{u}^{n+1}-\mathbf{u}^n\|_0 \\
&\leq \frac{1}{2} c_{\mathbf{f}}^2 \Delta t + \frac{1}{2\Delta t} \|\mathbf{u}^{n+1}-\mathbf{u}^n\|_0^2 .
\end{aligned} \tag{4.118}
$$

It is implied from (4.117) and (4.118) that

$$
\theta(q^{n+1}+q^n)|d_t e_q^{n+1}| \leq \frac{1}{2} c_{\mathbf{f}}^2 \Delta t + \frac{1}{\Delta t} \|\mathbf{u}^{n+1}-\mathbf{u}^n\|_0^2 + \nu \|\nabla \mathbf{u}^{n+1}\|_0 \|\nabla(\mathbf{u}^{n+1}-\mathbf{u}^n)\|_0 . \tag{4.119}
$$

To estimate the right-hand side of (4.119), we observe from Theorem 4.16 that

$$
\begin{aligned}
\|\mathbf{u}^{n+1}-\mathbf{u}^n\|_0^2 &\leq \left(\|e_{\mathbf{u}}^{n+1}-e_{\mathbf{u}}^n\|_0 + \|\mathbf{u}(t_{n+1})-\mathbf{u}(t_n)\|_0\right)^2 \\
&\leq \left(\sqrt{2C_1}\,\Delta t + \int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0 \, dt\right)^2 \\
&\leq (2C_1+1)\left(\Delta t^2 + \Delta t \int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0^2 \, dt\right), \\
\|\nabla \mathbf{u}^{n+1}\|_0 &\leq \|\nabla e_{\mathbf{u}}^{n+1}\|_0 + \|\nabla \mathbf{u}(t_{n+1})\|_0 \\
&\leq \sqrt{\frac{2C_1\Delta t}{\nu}} + M \leq \sqrt{\frac{2C_1 T}{\nu}} + M, \\
\|\nabla(\mathbf{u}^{n+1}-\mathbf{u}^n)\|_0 &\leq \|\nabla(e_{\mathbf{u}}^{n+1}-e_{\mathbf{u}}^n)\|_0 + \|\nabla(\mathbf{u}(t_{n+1})-\mathbf{u}(t_n))\|_0 \\
&\leq \|\nabla e_{\mathbf{u}}^{n+1}\|_0 + \|\nabla e_{\mathbf{u}}^n\|_0 + \int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_1 \, dt .
\end{aligned} \tag{4.120}
$$

When $\Delta t \leq \frac{\theta}{2C_2}$, we have $|e_q^n| \leq \frac{1}{2}$ for $0 \leq n \leq N$. Therefore, it holds $\frac{1}{2} \leq q^n \leq \frac{3}{2}$ for $0 \leq n \leq N$, and thus $q^{n+1}+q^n \geq 1$. Collecting (4.119) and (4.120) yields

$$
\begin{aligned}
\theta|d_t e_q^{n+1}| &\leq \theta(q^{n+1}+q^n)|d_t e_q^{n+1}| \\
&\leq \left(\frac{1}{2} c_{\mathbf{f}}^2 + 2C_1 + 1\right)\Delta t + (2C_1+1)\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0^2 \, dt \\
&\quad + \left(\sqrt{2C_1 T \nu} + M\nu\right)\left(\|\nabla e_{\mathbf{u}}^{n+1}\|_0 + \|\nabla e_{\mathbf{u}}^n\|_0 + \int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_1 \, dt\right).
\end{aligned} \tag{4.121}
$$

Squaring both sides of (4.121) and applying the Cauchy-Schwarz inequality, we find

$$
\begin{aligned}
\theta^2 |d_t e_q^{n+1}|^2 \leq & \left[\left(\frac{1}{2} c_{\mathbf{f}}^2 + 2C_1 + 1 \right)^2 + (2C_1 + 1)^2 + 3 \left(\sqrt{2C_1 T \nu} + M\nu \right)^2 \right] \\
& \cdot \left[\Delta t^2 + \Delta t \int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0^4 \, dt + \|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 + \|\nabla e_{\mathbf{u}}^n\|_0^2 + \Delta t \int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_1^2 \, dt \right].
\end{aligned}
\tag{4.122}
$$

By the regularity of $\mathbf{u}$ and Theorem 4.16, we have

$$
\int_0^T \|\mathbf{u}_t\|_0^4 \, dt \leq \widetilde{M}, \quad \int_0^T \|\mathbf{u}_t\|_1^2 \, dt \leq M, \quad \sum_{n=0}^{N} \|\nabla e_{\mathbf{u}}^n\|_0^2 \leq \frac{2C_1}{\nu} \Delta t. \tag{4.123}
$$

The proof of Lemma 4.18 is completed by summing (4.122) over $n = 0, 1, \ldots, N-1$ and using (4.123), where the resulting constant C_3, given by

$$
C_3 = \left[\left(\frac{1}{2} c_{\mathbf{f}}^2 + 2C_1 + 1 \right)^2 + (2C_1 + 1)^2 + 3 \left(\sqrt{2C_1 T \nu} + M\nu \right)^2 \right] \left(T + \widetilde{M} + M + \frac{4C_1}{\nu} \right),
$$

is independent of Δt and θ. □

Next, we state the main result in this subsection and present a detailed proof of the pressure error estimate. In addition to (4.71), the following extra regularity assumptions are imposed (cf. Remark 4.20):

$$
\begin{aligned}
& \mathbf{u}_t \in L^2(0, T; \mathbf{H}^2(\Omega)) \cap L^4(0, T; \mathbf{H}^1(\Omega)) \cap L^\infty(0, \tau_0; \mathbf{H}^1(\Omega)), \\
& \mathbf{u}_{tt} \in L^2(0, T; \mathbf{L}^2(\Omega)) \cap L^\infty(0, \tau_0; \mathbf{L}^2(\Omega)), \quad \mathbf{u}_{ttt} \in L^2(0, T; \mathbf{H}^{-1}(\Omega)).
\end{aligned}
\tag{4.124}
$$

Let $\widehat{M}$ be a constant such that

$$
\begin{aligned}
& \max \left\{ \sup_{t \in [0, \tau_0]} \|\mathbf{u}_t\|_1, \int_0^T \|\mathbf{u}_t\|_2^2 \, dt, \int_0^T \|\mathbf{u}_t\|_1^4 \, dt \right\} \leq \widehat{M}, \\
& \max \left\{ \sup_{t \in [0, \tau_0]} \|\mathbf{u}_{tt}\|_0, \int_0^T \|\mathbf{u}_{tt}\|_0^2 \, dt, \int_0^T \|\mathbf{u}_{ttt}\|_{-1}^2 \, dt \right\} \leq \widehat{M}.
\end{aligned}
\tag{4.125}
$$

Theorem 4.19. Let $\{\mathbf{u}^n\}$, $\{p^n\}$, and $\{q^n\}$ be generated by the first-order DRLM scheme (4.51). There exist positive constants $\hat{\tau}_0$ and C_4 depending on Ω, T, ν, M, $\widehat{M}$, $\mathbf{u}_0$, and $\mathbf{f}$ but independent of Δt,

θ, and n such that the following error estimate holds for all $\Delta t \leq \min\{\tau_0, \hat{\tau}_0, \frac{\theta}{2C_2}\}$:

$$\Delta t \sum_{i=0}^{n} \|e_p^{i+1}\|_{L^2/\mathbb{R}}^2 \leq C_4 \left(1 + \frac{1}{\theta}\right) \Delta t^2, \quad 0 \leq n \leq N-1,$$

where the constants τ_0 and C_2 are given in Theorem 4.16.

Proof. To obtain optimal error estimates for the pressure, we first apply the inf-sup condition to derive an upper bound for the pressure error that involves only the velocity and Lagrange multiplier. This leads to estimating the discrete time derivative of the velocity error – the main step in the proof, which can be carried out using Lemma 4.18 and the regularity assumptions on $\mathbf{u}$. Finally, the desired estimate is obtained using Gronwall's inequality in Lemma 4.15.

Step 1: Estimates for the pressure. A standard approach to estimating the pressure error is to apply the inf-sup (or LBB) condition after establishing the velocity error bound (cf. [123, 124]). Equivalently, we note from [13, Lemma IV.1.9] that there exists a constant $\beta > 0$ such that for all $\varphi \in L^2(\Omega)/\mathbb{R}$, we have

$$\frac{1}{\beta}\|\nabla\varphi\|_{-1} \leq \|\varphi\|_{L^2/\mathbb{R}} \leq \beta\|\nabla\varphi\|_{-1}. \tag{4.126}$$

In order to bound $\|e_p^n\|_{L^2/\mathbb{R}}$ for $1 \leq n \leq N$, it suffices to estimate $\|\nabla e_p^n\|_{-1}$. From (4.73), we have

$$\nabla e_p^{n+1} = -\frac{e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^n}{\Delta t} + \nu\Delta e_{\mathbf{u}}^{n+1} + \mathbf{R}_{\mathbf{u}}^{n+1} + (\mathbf{u}(t_{n+1})\cdot\nabla)\mathbf{u}(t_{n+1}) - q^{n+1}(\mathbf{u}^n\cdot\nabla)\mathbf{u}^n. \tag{4.127}$$

For any $\mathbf{v} \in \mathbf{H}_0^1(\Omega)$, the $\mathbf{L}^2$ inner product of (4.127) with $\mathbf{v}$ leads to

$$\begin{aligned}
\left(\nabla e_p^{n+1}, \mathbf{v}\right) &= -\frac{1}{\Delta t}\left(e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^n, \mathbf{v}\right) - \nu\left(\nabla e_{\mathbf{u}}^{n+1}, \nabla\mathbf{v}\right) + \left(\mathbf{R}_{\mathbf{u}}^{n+1}, \mathbf{v}\right) \\
&\quad + b\left(\mathbf{u}(t_{n+1}), \mathbf{u}(t_{n+1}), \mathbf{v}\right) - q^{n+1} b\left(\mathbf{u}^n, \mathbf{u}^n, \mathbf{v}\right) \\
&\leq \frac{1}{\Delta t}\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^n\|_{-1}\|\mathbf{v}\|_1 + \nu\|\nabla e_{\mathbf{u}}^{n+1}\|_0\|\mathbf{v}\|_1 + \Bigg(\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}\,dt \\
&\quad + 2c_2 M\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0\,dt + 2c_0c_2M\|e_{\mathbf{u}}^n\|_0 + c_0c_2\|e_{\mathbf{u}}^n\|_1^2 + c_2M^2|e_q^{n+1}|\Bigg)\|\mathbf{v}\|_1,
\end{aligned} \tag{4.128}$$

where in the last inequality, we have used estimates (4.75)-(4.80) with $e_{\mathbf{u}}^{n+1}$ replaced by $\mathbf{v}$. By the definition of $\mathbf{H}^{-1}$-norm, we know that

$$\|\nabla e_p^{n+1}\|_{-1} = \inf\left\{c \geq 0 \,:\, \left(\nabla e_p^{n+1}, \mathbf{v}\right) \leq c\|\mathbf{v}\|_1,\ \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega)\right\}.$$

Therefore, (4.128) gives us

$$\begin{aligned}\|\nabla e_p^{n+1}\|_{-1} \leq{}& \frac{1}{\Delta t}\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_{-1} + \nu\|\nabla e_{\mathbf{u}}^{n+1}\|_0 + \int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}\,dt \\ &+ 2c_2M\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0\,dt + 2c_0c_2M\|e_{\mathbf{u}}^{n}\|_0 + c_0c_2\|e_{\mathbf{u}}^{n}\|_1^2 + c_2M^2|e_q^{n+1}|.\end{aligned} \tag{4.129}$$

Squaring both sides of (4.129) and applying the Cauchy-Schwarz inequality, we deduce that

$$\begin{aligned}\|\nabla e_p^{n+1}\|_{-1}^2 \leq{}& \left(1 + \nu + 1 + 4c_2^2M^2 + 4c_0^2c_2^2M^2 + c_0^2c_2^2 + c_2^2M^4\right) \\ &\cdot\left(\frac{1}{\Delta t^2}\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_{-1}^2 + \nu\|\nabla e_{\mathbf{u}}^{n+1}\|_0^2 + \Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_{tt}\|_{-1}^2\,dt\right. \\ &\left.+ \Delta t\int_{t_n}^{t_{n+1}} \|\mathbf{u}_t\|_0^2\,dt + \|e_{\mathbf{u}}^{n}\|_0^2 + \|e_{\mathbf{u}}^{n}\|_1^4 + |e_q^{n+1}|^2\right).\end{aligned} \tag{4.130}$$

Let $c_{10} = 2 + \nu + 4c_2^2M^2 + 4c_0^2c_2^2M^2 + c_0^2c_2^2 + c_2^2M^4$. The sum of (4.130) over $i = 0, 1, \ldots, n$ yields

$$\begin{aligned}\sum_{i=0}^{n} \|\nabla e_p^{i+1}\|_{-1}^2 \leq{}& c_{10}\left(\frac{1}{\Delta t^2}\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_{-1}^2 + \nu\sum_{i=0}^{n} \|\nabla e_{\mathbf{u}}^{i+1}\|_0^2 + \Delta t\int_0^T \|\mathbf{u}_{tt}\|_{-1}^2\,dt\right. \\ &\left.+ \Delta t\int_0^T \|\mathbf{u}_t\|_0^2\,dt + \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_0^2 + \sum_{i=0}^{n} \|e_{\mathbf{u}}^{i}\|_1^4 + \sum_{i=0}^{n} |e_q^{i+1}|^2\right).\end{aligned} \tag{4.131}$$

By using Theorem 4.16, the regularity assumption (4.72), and the estimate (4.112), we obtain from (4.131) that

$$\begin{aligned}\sum_{i=0}^{n} \|\nabla e_p^{i+1}\|_{-1}^2 \leq{}& c_{10}\left[\frac{1}{\Delta t^2}\sum_{i=0}^{n} \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_{-1}^2 + C_1\left(1 + \frac{1}{\theta}\right)\Delta t + 2M\Delta t\right. \\ &\left.+ C_1\left(1 + \frac{1}{\theta}\right)T\Delta t + \Delta t + \frac{C_2^2}{\theta^2}T\Delta t\right].\end{aligned} \tag{4.132}$$

Step 2: Estimates for $\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0$. It remains to estimate $\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_{-1}$ for $0 \leq n \leq N-1$. Since $\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_{-1} \leq \|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0$, we will bound $\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0$ instead. Taking the difference of (4.73) between two consecutive time steps yields

$$\frac{d_t e_{\mathbf{u}}^{n+1} - d_t e_{\mathbf{u}}^{n}}{\Delta t} - \nu \Delta d_t e_{\mathbf{u}}^{n+1} + \nabla d_t e_p^{n+1} = d_t \mathbf{R}_{\mathbf{u}}^{n+1} + \frac{(\mathbf{u}(t_{n+1}) \cdot \nabla)\mathbf{u}(t_{n+1}) - (\mathbf{u}(t_n) \cdot \nabla)\mathbf{u}(t_n)}{\Delta t} - \frac{q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n - q^n(\mathbf{u}^{n-1} \cdot \nabla)\mathbf{u}^{n-1}}{\Delta t}. \tag{4.133}$$

Taking the $\mathbf{L}^2$ inner product of (4.133) with $d_t e_{\mathbf{u}}^{n+1}$ gives us

$$\begin{aligned}
&\frac{\|d_t e_{\mathbf{u}}^{n+1}\|_0^2 - \|d_t e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \frac{\|d_t e_{\mathbf{u}}^{n+1} - d_t e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \nu \|\nabla d_t e_{\mathbf{u}}^{n+1}\|_0^2 \\
&\quad = (d_t \mathbf{R}_{\mathbf{u}}^{n+1}, d_t e_{\mathbf{u}}^{n+1}) + \frac{b(\mathbf{u}(t_{n+1}), \mathbf{u}(t_{n+1}), d_t e_{\mathbf{u}}^{n+1}) - b(\mathbf{u}(t_n), \mathbf{u}(t_n), d_t e_{\mathbf{u}}^{n+1})}{\Delta t} \\
&\qquad - \frac{q^{n+1} b(\mathbf{u}^n, \mathbf{u}^n, d_t e_{\mathbf{u}}^{n+1}) - q^n b(\mathbf{u}^{n-1}, \mathbf{u}^{n-1}, d_t e_{\mathbf{u}}^{n+1})}{\Delta t} \\
&\quad = (d_t \mathbf{R}_{\mathbf{u}}^{n+1}, d_t e_{\mathbf{u}}^{n+1}) + S_1 + S_2,
\end{aligned} \tag{4.134}$$

where S_1 and S_2 are defined as

$$\begin{aligned}
S_1 &= \frac{b(\mathbf{u}(t_{n+1}), \mathbf{u}(t_{n+1}), d_t e_{\mathbf{u}}^{n+1}) - 2b(\mathbf{u}(t_n), \mathbf{u}(t_n), d_t e_{\mathbf{u}}^{n+1}) + b(\mathbf{u}(t_{n-1}), \mathbf{u}(t_{n-1}), d_t e_{\mathbf{u}}^{n+1})}{\Delta t} \\
S_2 &= -\frac{q^{n+1} b(\mathbf{u}^n, \mathbf{u}^n, d_t e_{\mathbf{u}}^{n+1}) - b(\mathbf{u}(t_n), \mathbf{u}(t_n), d_t e_{\mathbf{u}}^{n+1})}{\Delta t} \\
&\qquad + \frac{q^n b(\mathbf{u}^{n-1}, \mathbf{u}^{n-1}, d_t e_{\mathbf{u}}^{n+1}) - b(\mathbf{u}(t_{n-1}), \mathbf{u}(t_{n-1}), d_t e_{\mathbf{u}}^{n+1})}{\Delta t}.
\end{aligned}$$

By noting that $d_t e_{\mathbf{u}}^{n+1} \in \mathbf{H}_0^1(\Omega)$ and

$$\begin{aligned}
d_t \mathbf{R}_{\mathbf{u}}^{n+1} &= \frac{1}{\Delta t^2} \left[\int_{t_n}^{t_{n+1}} (t - t_n) \mathbf{u}_{tt}(t) dt - \int_{t_{n-1}}^{t_n} (t - t_{n-1}) \mathbf{u}_{tt}(t) dt \right] \\
&= \frac{1}{\Delta t^2} \int_{t_{n-1}}^{t_n} \int_r^{r+\Delta t} \int_s^{r+\Delta t} \mathbf{u}_{ttt}(\xi)\, d\xi\, ds\, dr,
\end{aligned}$$

we obtain

$$(d_t \mathbf{R}_{\mathbf{u}}^{n+1}, d_t e_{\mathbf{u}}^{n+1}) \le \|d_t \mathbf{R}_{\mathbf{u}}^{n+1}\|_{-1} \|d_t e_{\mathbf{u}}^{n+1}\|_1 \le \left(\int_{t_{n-1}}^{t_{n+1}} \|\mathbf{u}_{ttt}\|_{-1} dt \right) \|d_t e_{\mathbf{u}}^{n+1}\|_1. \tag{4.135}$$

Due to the linearity of the operator b, S_1 can be written as follows

$$S_1 = \frac{b(\mathbf{u}(t_{n+1}) - 2\mathbf{u}(t_n) + \mathbf{u}(t_{n-1}), \mathbf{u}(t_{n-1}), d_t e_{\mathbf{u}}^{n+1})}{\Delta t} + \frac{b(\mathbf{u}(t_n), \mathbf{u}(t_{n+1}) - 2\mathbf{u}(t_n) + \mathbf{u}(t_{n-1}), d_t e_{\mathbf{u}}^{n+1})}{\Delta t} + \frac{b(\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n), \mathbf{u}(t_{n+1}) - \mathbf{u}(t_{n-1}), d_t e_{\mathbf{u}}^{n+1})}{\Delta t}.$$

Using the properties of b in (4.68)-(4.69) and the regularity assumption (4.72), we find

$$\begin{aligned} S_1 \leq\ & \frac{c_2}{\Delta t}\|\mathbf{u}(t_{n+1}) - 2\mathbf{u}(t_n) + \mathbf{u}(t_{n-1})\|_0\|\mathbf{u}(t_{n-1})\|_2\|d_t e_{\mathbf{u}}^{n+1}\|_1 \\ & + \frac{c_2}{\Delta t}\|\mathbf{u}(t_n)\|_2\|\mathbf{u}(t_{n+1}) - 2\mathbf{u}(t_n) + \mathbf{u}(t_{n-1})\|_0\|d_t e_{\mathbf{u}}^{n+1}\|_1 \\ & + \frac{c_2}{\Delta t}\|\mathbf{u}(t_{n+1}) - \mathbf{u}(t_n)\|_1\|\mathbf{u}(t_{n+1}) - \mathbf{u}(t_{n-1})\|_1\|d_t e_{\mathbf{u}}^{n+1}\|_1 \\ \leq\ & \left(2c_2 M\int_{t_{n-1}}^{t_{n+1}} \|\mathbf{u}_{tt}\|_0\,dt + 2c_2\int_{t_{n-1}}^{t_{n+1}} \|\mathbf{u}_t\|_1^2\,dt\right)\|d_t e_{\mathbf{u}}^{n+1}\|_1. \end{aligned} \tag{4.136}$$

On the other hand, S_2 can be expressed as

$$\begin{aligned} S_2 =\ & -d_t e_q^{n+1} b(\mathbf{u}^n, \mathbf{u}^n, d_t e_{\mathbf{u}}^{n+1}) \\ & - e_q^n \frac{b(\mathbf{u}(t_n), \mathbf{u}(t_n) - \mathbf{u}(t_{n-1}), d_t e_{\mathbf{u}}^{n+1}) + b(\mathbf{u}(t_n) - \mathbf{u}(t_{n-1}), \mathbf{u}(t_{n-1}), d_t e_{\mathbf{u}}^{n+1})}{\Delta t} \\ & - q^n \frac{b(e_{\mathbf{u}}^{n-1}, \mathbf{u}(t_n) - \mathbf{u}(t_{n-1}), d_t e_{\mathbf{u}}^{n+1}) + b(\mathbf{u}(t_n) - \mathbf{u}(t_{n-1}), e_{\mathbf{u}}^n, d_t e_{\mathbf{u}}^{n+1})}{\Delta t} \\ & - q^n \left[b(d_t e_{\mathbf{u}}^n, \mathbf{u}(t_n), d_t e_{\mathbf{u}}^{n+1}) + b(\mathbf{u}(t_{n-1}), d_t e_{\mathbf{u}}^n, d_t e_{\mathbf{u}}^{n+1})\right] \\ & - q^n \left[b(d_t e_{\mathbf{u}}^n, e_{\mathbf{u}}^{n-1}, d_t e_{\mathbf{u}}^{n+1}) + b(e_{\mathbf{u}}^n, d_t e_{\mathbf{u}}^n, d_t e_{\mathbf{u}}^{n+1})\right]. \end{aligned}$$

By using (4.68)-(4.69) and (4.72) again, along with Lemma 4.14, one arrives at

$$\begin{aligned} S_2 \leq\ & c_2|d_t e_q^{n+1}|\|\mathbf{u}^n\|_1^2\|d_t e_{\mathbf{u}}^{n+1}\|_1 + \frac{2c_2 M}{\Delta t}|e_q^n|\left(\int_{t_{n-1}}^{t_n} \|\mathbf{u}_t\|_0\,dt\right)\|d_t e_{\mathbf{u}}^{n+1}\|_1 \\ & + \frac{c_0 c_2}{\Delta t}(\|e_{\mathbf{u}}^{n-1}\|_0 + \|e_{\mathbf{u}}^n\|_0)\left(\int_{t_{n-1}}^{t_n} \|\mathbf{u}_t\|_2\,dt\right)\|d_t e_{\mathbf{u}}^{n+1}\|_1 \\ & + 2c_0 c_2 M\|d_t e_{\mathbf{u}}^n\|_0\|d_t e_{\mathbf{u}}^{n+1}\|_1 + c_0 c_2(\|e_{\mathbf{u}}^{n-1}\|_1 + \|e_{\mathbf{u}}^n\|_1)\|d_t e_{\mathbf{u}}^n\|_1\|d_t e_{\mathbf{u}}^{n+1}\|_1. \end{aligned} \tag{4.137}$$

Collecting (4.134)-(4.137), we find

$$
\begin{aligned}
&\frac{\|d_t e_{\mathbf{u}}^{n+1}\|_0^2 - \|d_t e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \frac{\|d_t e_{\mathbf{u}}^{n+1} - d_t e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \nu\|\nabla d_t e_{\mathbf{u}}^{n+1}\|_0^2 \\
&\quad \le \Big[\int_{t_{n-1}}^{t_{n+1}} \|\mathbf{u}_{ttt}\|_{-1} dt + 2c_2 M \int_{t_{n-1}}^{t_{n+1}} \|\mathbf{u}_{tt}\|_0 \, dt + 2c_2 \int_{t_{n-1}}^{t_{n+1}} \|\mathbf{u}_t\|_1^2 \, dt \\
&\qquad + c_2 |d_t e_q^{n+1}| \|\mathbf{u}^n\|_1^2 + \frac{2c_2 M}{\Delta t}|e_q^n| \int_{t_{n-1}}^{t_n} \|\mathbf{u}_t\|_0 \, dt + \frac{c_0 c_2}{\Delta t}(\|e_{\mathbf{u}}^{n-1}\|_0 + \|e_{\mathbf{u}}^{n}\|_0) \int_{t_{n-1}}^{t_n} \|\mathbf{u}_t\|_2 \, dt \\
&\qquad + 2c_0 c_2 M \|d_t e_{\mathbf{u}}^n\|_0 + c_0 c_2 (\|e_{\mathbf{u}}^{n-1}\|_1 + \|e_{\mathbf{u}}^{n}\|_1)\|d_t e_{\mathbf{u}}^n\|_1 \Big] \|d_t e_{\mathbf{u}}^{n+1}\|_1.
\end{aligned} \tag{4.138}
$$

Using similar arguments as in (4.82) and the facts (cf. Theorem 4.16) that for $0 \le n \le N$,

$$
\begin{aligned}
&|e_q^n| \le \frac{C_2}{\theta}\Delta t, \quad \|e_{\mathbf{u}}^n\|_0^2 \le C_1\left(1 + \frac{1}{\theta}\right)\Delta t^2, \quad \|e_{\mathbf{u}}^n\|_1^2 \le \frac{2(1+c_1)^2 C_1}{\nu}\Delta t, \\
&\|\mathbf{u}^n\|_1^2 \le (\|\mathbf{u}(t_n)\|_1 + \|e_{\mathbf{u}}^n\|_1)^2 \le 2\left(M^2 + \frac{2(1+c_1)^2 C_1 T}{\nu}\right),
\end{aligned}
$$

we obtain from (4.138) that

$$
\begin{aligned}
&\frac{\|d_t e_{\mathbf{u}}^{n+1}\|_0^2 - \|d_t e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \frac{\|d_t e_{\mathbf{u}}^{n+1} - d_t e_{\mathbf{u}}^{n}\|_0^2}{2\Delta t} + \frac{\nu}{2}\|\nabla d_t e_{\mathbf{u}}^{n+1}\|_0^2 \\
&\quad \le c_{11}\Big[\Delta t \int_{t_{n-1}}^{t_{n+1}} \|\mathbf{u}_{ttt}\|_{-1}^2 dt + \Delta t \int_{t_{n-1}}^{t_{n+1}} \|\mathbf{u}_{tt}\|_0^2 \, dt + \Delta t \int_{t_{n-1}}^{t_{n+1}} \|\mathbf{u}_t\|_1^4 \, dt \\
&\qquad + |d_t e_q^{n+1}|^2 + \frac{\Delta t}{\theta^2} \int_{t_{n-1}}^{t_n} \|\mathbf{u}_t\|_0^2 \, dt + \left(1 + \frac{1}{\theta}\right)\Delta t \int_{t_{n-1}}^{t_n} \|\mathbf{u}_t\|_2^2 \, dt \\
&\qquad + \|d_t e_{\mathbf{u}}^n\|_0^2 + \Delta t \|d_t e_{\mathbf{u}}^n\|_1^2 \Big],
\end{aligned} \tag{4.139}
$$

with

$$
\begin{aligned}
c_{11} =& \frac{(1+c_1)^2}{2\nu}\Big[2 + 8c_2^2 M^2 + 8c_2^2 + 4c_2^2\left(M^2 + \frac{2(1+c_1)^2 C_1 T}{\nu}\right)^2 + 4C_2^2 c_2^2 M^2 + 4c_0^2 c_2^2 C_1 \\
&+ 4c_0^2 c_2^2 M^2 + \frac{8c_0^2 c_2^2 (1+c_1)^2 C_1}{\nu}\Big].
\end{aligned}
$$

It should be noted that the term $\Delta t\|d_t e_{\mathbf{u}}^n\|_1^2$ on the right-hand side of (4.139) introduces additional complexity to the analysis, which does not occur in the 2D case [124]. Taking the sum of (4.139) over $i = 1, 2, \ldots, n$ and multiplying the resulting inequality by $2\Delta t$ yields

$$
\begin{aligned}
&\|d_t e_{\mathbf{u}}^{n+1}\|_0^2 + \sum_{i=1}^{n} \|d_t e_{\mathbf{u}}^{i+1} - d_t e_{\mathbf{u}}^{i}\|_0^2 + \Delta t \nu \sum_{i=1}^{n} \|\nabla d_t e_{\mathbf{u}}^{i+1}\|_0^2 \\
&\quad \le \|d_t e_{\mathbf{u}}^{1}\|_0^2 + 2c_{11}\Delta t \bigg[2\Delta t \int_0^T \|\mathbf{u}_{ttt}\|_{-1}^2 dt + 2\Delta t \int_0^T \|\mathbf{u}_{tt}\|_0^2 \, dt + 2\Delta t \int_0^T \|\mathbf{u}_t\|_1^4 \, dt \\
&\qquad\qquad + \sum_{i=1}^{n} |d_t e_q^{i+1}|^2 + \frac{\Delta t}{\theta^2} \int_0^T \|\mathbf{u}_t\|_0^2 \, dt + \left(1 + \frac{1}{\theta}\right) \Delta t \int_0^T \|\mathbf{u}_t\|_2^2 \, dt \\
&\qquad\qquad + \sum_{i=1}^{n} \|d_t e_{\mathbf{u}}^{i}\|_0^2 + \Delta t \sum_{i=1}^{n} \|d_t e_{\mathbf{u}}^{i}\|_1^2 \bigg] \\
&\quad \le \|d_t e_{\mathbf{u}}^{1}\|_0^2 + 2c_{11}\Delta t^2 \|d_t e_{\mathbf{u}}^{1}\|_1^2 + 2c_{11}\Delta t \bigg[6\widehat{M}\Delta t + \frac{C_3 \Delta t}{\theta^2} + \frac{M\Delta t}{\theta^2} + \left(1 + \frac{1}{\theta}\right) \widehat{M}\Delta t \\
&\qquad\qquad + \sum_{i=1}^{n} \|d_t e_{\mathbf{u}}^{i}\|_0^2 + \Delta t \sum_{i=2}^{n} \|d_t e_{\mathbf{u}}^{i}\|_1^2 \bigg],
\end{aligned}
\tag{4.140}
$$

where we have used Lemma 4.18 and regularity assumptions (4.72) and (4.125) in the last estimate. Note that, by convention, the last summation on the right-hand side of (4.140) is set to zero when $n = 1$. To bound $\|d_t e_{\mathbf{u}}^1\|_0^2 + 2c_{11}\Delta t^2 \|d_t e_{\mathbf{u}}^1\|_1^2$ in (4.140), we let $n = 0$ in (4.74) and obtain

$$
\frac{\|e_{\mathbf{u}}^1\|_0^2}{\Delta t} + \nu \|\nabla e_{\mathbf{u}}^1\|_0^2 = (\mathbf{R}_{\mathbf{u}}^1, e_{\mathbf{u}}^1) + b(\mathbf{u}(t_1), \mathbf{u}(t_1), e_{\mathbf{u}}^1) - q^1 b(\mathbf{u}(t_0), \mathbf{u}(t_0), e_{\mathbf{u}}^1). \tag{4.141}
$$

Using similar arguments as in (4.75)-(4.80) and the regularity of $\mathbf{u}$ in (4.125), we deduce

$$
\begin{aligned}
\frac{\|e_{\mathbf{u}}^1\|_0^2}{\Delta t} + \nu \|\nabla e_{\mathbf{u}}^1\|_0^2 &\le \left(\int_0^{t_1} \|\mathbf{u}_{tt}\|_0 \, dt + 2c_2 M \int_0^{t_1} \|\mathbf{u}_t\|_1 \, dt + c_2 M^2 |e_q^1| \right) \|e_{\mathbf{u}}^1\|_0 \\
&\le \Delta t \left(\widehat{M} + 2c_2 M \widehat{M} + \frac{C_2 c_2 M^2}{\theta} \right) \|e_{\mathbf{u}}^1\|_0 \le c_{12} \left(1 + \frac{1}{\theta}\right) \Delta t \|e_{\mathbf{u}}^1\|_0,
\end{aligned}
\tag{4.142}
$$

where $c_{12} = \max\{\widehat{M} + 2c_2 M\widehat{M}, C_2 c_2 M^2\}$. By dropping the non-negative term $\nu\|\nabla e_{\mathbf{u}}^1\|_0^2$ on the left-hand side of (4.142), we find

$$
\|e_{\mathbf{u}}^1\|_0 \le c_{12} \left(1 + \frac{1}{\theta}\right) \Delta t^2 \quad \text{and} \quad \|d_t e_{\mathbf{u}}^1\|_0 = \frac{1}{\Delta t} \|e_{\mathbf{u}}^1\|_0 \le c_{12} \left(1 + \frac{1}{\theta}\right) \Delta t. \tag{4.143}
$$

It follows from (4.142) and (4.143) that

$$
\Delta t^2 \|d_t e_{\mathbf{u}}^1\|_1^2 = \|e_{\mathbf{u}}^1\|_1^2 \le (1 + c_1)^2 \|\nabla e_{\mathbf{u}}^1\|_0^2 \le \frac{(1 + c_1)^2}{\nu} c_{12}^2 \left(1 + \frac{1}{\theta}\right)^2 \Delta t^3. \tag{4.144}
$$

From (4.140), (4.143), and (4.144), one obtains

$$\begin{aligned}
&\|d_t e_{\mathbf{u}}^{n+1}\|_0^2 + \sum_{i=1}^{n} \|d_t e_{\mathbf{u}}^{i+1} - d_t e_{\mathbf{u}}^{i}\|_0^2 + \Delta t \nu \sum_{i=1}^{n} \|\nabla d_t e_{\mathbf{u}}^{i+1}\|_0^2 \\
&\quad \le c_{13}\left(1 + \frac{1}{\theta} + \frac{1}{\theta^2}\right)\Delta t^2 + 2c_{11}\Delta t \sum_{i=1}^{n} \|d_t e_{\mathbf{u}}^{i}\|_0^2 + 2c_{11}\Delta t^2 \sum_{i=2}^{n} \|d_t e_{\mathbf{u}}^{i}\|_1^2,
\end{aligned} \tag{4.145}$$

where $c_{13} = 2c_{12}^2 + \frac{4(1+c_1)^2 c_{11} c_{12}^2 T}{\nu} + 2c_{11}(C_3 + M + 7\widehat{M})$. If

$$\Delta t \le \frac{\nu}{2c_{11}(1+c_1)^2} =: \hat{\tau}_0,$$

the last summation in (4.145) can be bounded using Poincaré inequality (4.67) by

$$2c_{11}\Delta t^2 \sum_{i=2}^{n} \|d_t e_{\mathbf{u}}^{i}\|_1^2 \le 2c_{11}(1+c_1)^2 \Delta t^2 \sum_{i=2}^{n} \|\nabla d_t e_{\mathbf{u}}^{i}\|_0^2 \le \Delta t \nu \sum_{i=1}^{n} \|\nabla d_t e_{\mathbf{u}}^{i+1}\|_0^2. \tag{4.146}$$

In light of (4.146), we derive from (4.145) that

$$\|d_t e_{\mathbf{u}}^{n+1}\|_0^2 \le c_{13}\left(1 + \frac{1}{\theta} + \frac{1}{\theta^2}\right)\Delta t^2 + 2c_{11}\Delta t \sum_{i=1}^{n} \|d_t e_{\mathbf{u}}^{i}\|_0^2, \quad n \ge 1. \tag{4.147}$$

By applying Gronwall's inequality in Lemma 4.15 to (4.147), noting from (4.143) and the definition of c_{13} that $\|d_t e_{\mathbf{u}}^{1}\|_0^2 \le c_{13}\left(1 + \frac{1}{\theta} + \frac{1}{\theta^2}\right)\Delta t^2$, we end up with

$$\frac{1}{\Delta t^2}\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0^2 = \|d_t e_{\mathbf{u}}^{n+1}\|_0^2 \le c_{13}\left(1 + \frac{1}{\theta} + \frac{1}{\theta^2}\right) e^{2c_{11}T}\Delta t^2, \quad n \ge 0. \tag{4.148}$$

Step 3: Optimal convergence rate for the pressure. Finally, the combination of (4.132), (4.148), and the fact $\|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_{-1} \le \|e_{\mathbf{u}}^{i+1} - e_{\mathbf{u}}^{i}\|_0$ results in

$$\begin{aligned}
\sum_{i=0}^{n} \|\nabla e_p^{i+1}\|_{-1}^2 &\le c_{10}\Bigg[c_{13}\left(1 + \frac{1}{\theta} + \frac{1}{\theta^2}\right) T e^{2c_{11}T} + C_1\left(1 + \frac{1}{\theta}\right) + 2M \\
&\qquad + C_1\left(1 + \frac{1}{\theta}\right) T + 1 + \frac{C_2^2}{\theta^2} T\Bigg]\Delta t \\
&\le c_{14}\left(1 + \frac{1}{\theta}\right)\Delta t,
\end{aligned} \tag{4.149}$$

provided that $\Delta t \le \min\{\tau_0, \hat{\tau}_0, \frac{\theta}{2C_2}\}$, where $c_{14} = c_{10}\left(2c_{13}Te^{2c_{11}T} + C_1 + 2M + C_1 T + 1 + C_2^2 T\right)$ (note that $\frac{1}{\theta^2} \le \frac{1}{\theta}$). The proof of Theorem 4.19 is thus completed by combining (4.126) and (4.149) with $C_4 = c_{14}\beta^2$. $\square$

Remark 4.20. (i) The pressure error estimate in Theorem 4.19 requires higher regularity on the exact solution than the velocity error estimate in Theorem 4.16. The main reason is that we estimate $\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_0$ in Step 2 of the above proof, rather than directly bounding $\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_{-1}$. As discussed in [159, 160], the desired estimate $\|e_{\mathbf{u}}^{n+1} - e_{\mathbf{u}}^{n}\|_{-1} \leq C\Delta t^2$ cannot be obtained using standard techniques involving Stokes operator A: $D(A) = \mathbf{V} \cap \mathbf{H}^2(\Omega) \to \mathbf{H}$ due to the following incorrect inequality [160]:

$$\gamma\|\mathbf{v}\|_{-1}^2 \leq \left(A^{-1}\mathbf{v}, \mathbf{v}\right), \quad \forall\, \mathbf{v} \in \mathbf{H}, \tag{4.150}$$

for any constant $\gamma > 0$.

(ii) Extension of the presented semi-discrete analysis to fully discrete schemes could be carried out with inf-sup stable finite elements (e.g., Taylor-Hood elements) for spatial discretization. The key difference is the use of mathematical induction to first derive suboptimal convergence results, which serve as a basis for bounding the numerical solutions via inverse inequalities. Then optimal error estimates for velocity and pressure can be obtained, provided with the uniformly bounded numerical solutions. We refer to [110] for some related technical details.

(iii) Error analysis for high-order DRLM schemes, in both semi-discrete and fully discrete settings, is highly challenging. Unlike the first-order DRLM scheme (4.51) where the Lagrange multiplier is uniquely determined from the quadratic equation (4.55), the second-order DRLM scheme [48] leads to a more complicated quadratic equation for the Lagrange multiplier whose unique positive solution is not always guaranteed. This is beyond the scope of this work and will be part of our future investigation.

4.2.4 Numerical results

We first verify the convergence in time of the DRLM scheme (4.51) and demonstrate the role of the regularization parameter θ via a manufactured solution. Then we apply the scheme to simulate the well-known lid-driven cavity flow; additional tests on energy stability and benchmark problems in two and three dimensions, such as the Kelvin-Helmholtz instability, can be found in [48]. Note that various spatial discretization techniques, such as finite difference, finite element, or finite volume methods, can be used to construct fully discrete DRLM schemes. For simplicity, we adopt the finite difference approximation on a Marker-and-Cell (MAC) staggered grid with a uniform mesh size in both x- and y-directions, i.e., $h = h_x = h_y$.

Convergence test The external force $\mathbf{f}$ is computed from the following analytic solution, in which the velocity field satisfies homogeneous Dirichlet boundary conditions:

$$\mathbf{u} = \begin{pmatrix} 5\sin^2(\pi x)\sin(2\pi y)e^{-t} \\ -5\sin(2\pi x)\sin^2(\pi y)e^{-t} \end{pmatrix}, \quad p = \cos(\pi x)\sin(\pi y)e^{-t}.$$

We set the computational domain $\Omega = (0,1)^2$, the terminal time $T = 1$, and the viscosity coefficient $\nu = 0.1$. We vary both the time step size Δt and the spatial mesh size h with $\Delta t = 2h$ so that spatial errors are negligible compared to temporal errors. The $\mathbf{L}^2$ and $\mathbf{H}^1$ errors of the velocity, L^2 errors of the pressure, and absolute errors of the Lagrange multiplier at the final time with different values of θ are reported in Table 4.1. We clearly observe first-order convergence for all variables. As θ increases, errors of the Lagrange multiplier (see the last column in Table 4.1) decrease roughly proportionally, with a tenfold increase in θ resulting in nearly a tenfold decrease in the Lagrange multiplier error. Consequently, the velocity and pressure errors are improved when θ becomes larger, especially from $\theta = 0.1$ to $\theta = 1$. These observations agree well with the theoretical results in Theorems 4.16 and 4.19.

Lid-driven cavity flow We consider the lid-driven cavity flow [65, 100, 122, 124] to validate the long-time performance of the proposed DRLM scheme. The computational domain is the unit square $\Omega = (0,1)^2$ where the top lid moves with a tangential unit velocity and the remaining three walls satisfy no-slip boundary conditions. We run the simulation with parameters $\nu = 1/5000$ (i.e., the Reynolds number $Re = 5000$), $\theta = 100$, and $h = 2\Delta t = \frac{1}{256}$ until the steady state is

Table 4.1: Errors and convergence rates of the velocity, pressure, and Lagrange multiplier with $\Delta t = 2h \in \{\frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}\}$ and $\theta \in \{0.1, 1, 10, 100\}$.

θ	Δt	h	$\|e_{\mathbf{u}}\|_0$	$\|e_{\mathbf{u}}\|_1$	$\|e_p\|_0$	$\|e_q\|$
	1/8	1/16	6.38e-02	9.78e-01	9.78e-01	5.51e-01
	1/16	1/32	1.66e-02 [1.94]	2.42e-01 [2.01]	6.47e-01 [0.60]	4.51e-01 [0.29]
0.1	1/32	1/64	9.64e-03 [0.78]	1.50e-01 [0.69]	3.50e-01 [0.89]	2.66e-01 [0.76]
	1/64	1/128	4.99e-03 [0.95]	7.94e-02 [0.92]	1.81e-01 [0.95]	1.44e-01 [0.89]
	1/128	1/256	2.56e-03 [0.96]	4.11e-02 [0.95]	9.27e-02 [0.97]	7.53e-02 [0.94]
	1/8	1/16	3.48e-02	3.56e-01	4.21e-01	9.99e-02
	1/16	1/32	1.29e-02 [1.43]	1.29e-01 [1.46]	2.08e-01 [1.02]	5.77e-02 [0.79]
1	1/32	1/64	5.51e-03 [1.23]	5.82e-02 [1.15]	1.02e-01 [1.03]	3.04e-02 [0.92]
	1/64	1/128	2.51e-03 [1.13]	2.75e-02 [1.08]	5.05e-02 [1.01]	1.55e-02 [0.97]
	1/128	1/256	1.20e-03 [1.06]	1.34e-02 [1.04]	2.51e-02 [1.01]	7.84e-03 [0.98]
	1/8	1/16	3.39e-02	3.12e-01	3.08e-01	1.06e-02
	1/16	1/32	1.26e-02 [1.43]	1.13e-01 [1.47]	1.50e-01 [1.04]	5.95e-03 [0.83]
10	1/32	1/64	5.28e-03 [1.25]	4.94e-02 [1.19]	7.34e-02 [1.03]	3.09e-03 [0.95]
	1/64	1/128	2.38e-03 [1.15]	2.29e-02 [1.11]	3.63e-02 [1.02]	1.56e-03 [0.99]
	1/128	1/256	1.12e-03 [1.09]	1.10e-02 [1.06]	1.80e-02 [1.01]	7.87e-04 [0.99]
	1/8	1/16	3.38e-02	3.08e-01	2.96e-01	1.07e-03
	1/16	1/32	1.26e-02 [1.42]	1.12e-01 [1.46]	1.44e-01 [1.04]	5.97e-04 [0.84]
100	1/32	1/64	5.26e-03 [1.26]	4.86e-02 [1.20]	7.04e-02 [1.03]	3.09e-04 [0.95]
	1/64	1/128	2.37e-03 [1.15]	2.25e-02 [1.11]	3.48e-02 [1.02]	1.57e-04 [0.98]
	1/128	1/256	1.12e-03 [1.08]	1.08e-02 [1.06]	1.73e-02 [1.01]	7.87e-05 [1.00]

reached, defined by $\|\mathbf{u}^n - \mathbf{u}^{n-1}\|_\infty \leq 10^{-6}$.

Figure 4.11 illustrates the evolution of the velocity magnitude at times $t = 4, 6,$ and 10. The steady state contours for velocity magnitude, vorticity, and stream function are shown in Figure 4.12. We observe that the DRLM scheme accurately captures the expected flow structures (cf. the benchmark results in [65]), including the primary vortex, the secondary vortices near the bottom corners, and the tertiary vortex in the upper left region. Moreover, these results demonstrate strong agreement with those obtained by the second-order DRLM scheme in [48].

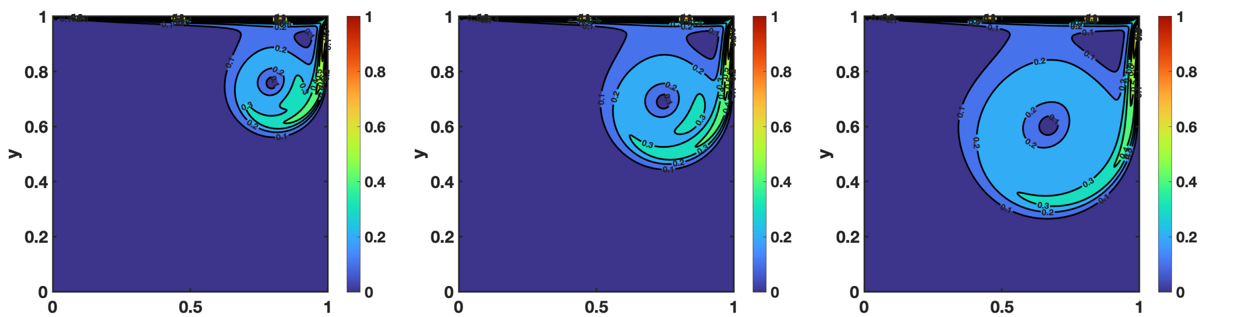


Figure 4.11: Contour plots of the velocity magnitude at times $t = 4, 6,$ and 10 with $h = 2\Delta t = \dfrac{1}{256}$.

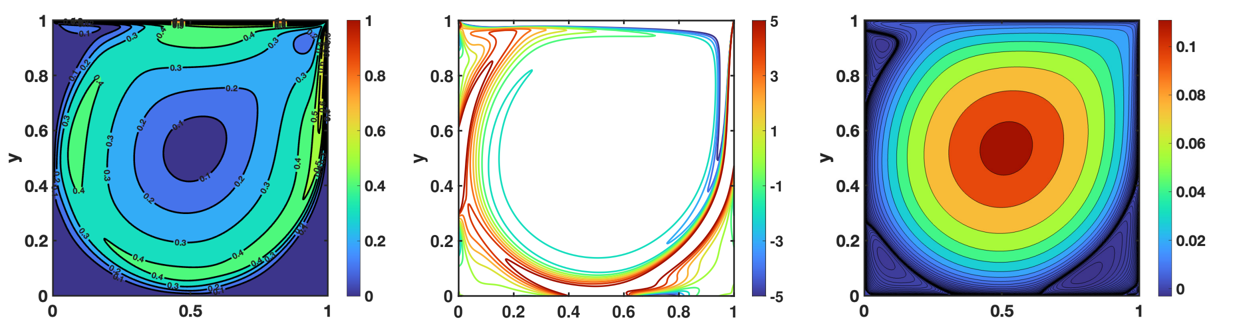


Figure 4.12: Contour plots of the velocity magnitude, vorticity, and stream function at the steady state with $h = 2\Delta t = \dfrac{1}{256}$.

4.3 DRLM method for the Cahn-Hilliard-Navier-Stokes system

4.3.1 The Cahn-Hilliard-Navier-Stokes model and equivalent formulation

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be an open and bounded domain, and $T > 0$ be the terminal time, we consider the following coupled Cahn-Hilliard-Navier-Stokes system:

$$\frac{\partial \phi}{\partial t} + \mathbf{u} \cdot \nabla \phi = M \Delta \mu, \qquad \text{in } \Omega \times (0, T], \tag{4.151a}$$

$$\mu = \lambda \left(-\Delta \phi + F'(\phi) \right), \qquad \text{in } \Omega \times (0, T], \tag{4.151b}$$

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mu \nabla \phi, \qquad \text{in } \Omega \times (0, T], \tag{4.151c}$$

$$\nabla \cdot \mathbf{u} = 0, \qquad \text{in } \Omega \times (0, T], \tag{4.151d}$$

subject to the initial conditions

$$\phi|_{t=0} = \phi_0, \quad \mathbf{u}|_{t=0} = \mathbf{u}_0, \quad \text{in } \Omega, \tag{4.152}$$

and suitable boundary conditions. Throughout this section, we consider either the periodic boundary conditions for all variables or the following no-flux/no-slip boundary conditions:

$$\frac{\partial \phi}{\partial \mathbf{n}} = \frac{\partial \mu}{\partial \mathbf{n}} = 0, \ \mathbf{u} = 0, \quad \text{on } \partial\Omega \times (0, T], \tag{4.153}$$

where $\mathbf{n}$ denotes the outward unit normal vector to $\partial\Omega$. In (4.151), the unknown quantities ϕ, μ, $\mathbf{u}$, and p represent the phase variable, the chemical potential, the velocity vector field, and the pressure, respectively. The given positive parameters are M (the mobility constant), λ (the mixing coefficient), and ν (the viscosity). The nonlinear free energy density is given by

$$F(\phi) = \frac{1}{4\varepsilon^2} (\phi^2 - 1)^2,$$

with ε denoting the thickness of the diffuse interface. The system (4.151) describes the motion of two incompressible, macroscopically immiscible Newtonian fluids with matched density (which is taken to be 1) and viscosity ν. We refer to [75, 5, 129] for the derivation of the model (and its extensions to other cases, e.g., with variable density and variable viscosity) and to [57, 1, 66] for the existence and uniqueness of the solution to the model.

It is well-known that the system (4.151) admits an energy dissipation law and is mass conservative. In particular, the total energy associated with (4.151) is defined as

$$\begin{aligned} E(\phi, \mathbf{u}) &= E_{CH}(\phi) + E_{NS}(\mathbf{u}) \\ &= \lambda \int_\Omega \Big(\frac{1}{2}|\nabla\phi|^2 + F(\phi)\Big)\, d\mathbf{x} + \frac{1}{2}\int_\Omega |\mathbf{u}|^2 \, d\mathbf{x}. \end{aligned} \tag{4.154}$$

By taking the L^2 inner product of (4.151a), (4.151b), and (4.151c) with μ, $\frac{\partial\phi}{\partial t}$, and $\mathbf{u}$, respectively, and using the incompressibility condition (4.151d), we obtain the energy dissipation law:

$$\frac{d}{dt}E(\phi, \mathbf{u}) = -M\|\nabla\mu\|_0^2 - \nu\|\nabla\mathbf{u}\|_0^2 \leq 0, \tag{4.155}$$

where $\|\cdot\|_0$ denotes the standard L^2 norm on Ω. The mass conservation property holds by taking the L^2 inner product of (4.151a) with 1 which yields

$$\frac{d}{dt}\int_\Omega \phi \, d\mathbf{x} = M\int_{\partial\Omega} \nabla\mu\cdot\mathbf{n}\, ds - \int_{\partial\Omega} \phi\mathbf{u}\cdot\mathbf{n}\, ds = 0, \tag{4.156}$$

where $\mathbf{u}\cdot\nabla\phi = \nabla\cdot(\phi\mathbf{u})$ due to (4.151d).

The goal of this section is to construct efficient, linear, decoupled numerical methods that preserve the two intrinsic properties (4.155) and (4.156) at the discrete level and accurately capture the dynamics of the model, especially when ε is small. To that end, by introducing two time-dependent Lagrange multipliers $\eta(t)$ and $q(t)$ with $\eta(0) = q(0) = 1$, we reformulate the CH-NS model (4.151) into the following system, subject to the same initial and boundary conditions:

$$
\begin{aligned}
&\frac{\partial \phi}{\partial t} + \eta\, \mathbf{u} \cdot \nabla \phi = M \Delta \mu, && \text{in } \Omega \times (0, T], && (4.157\text{a})\\
&\mu = \lambda \left(-\Delta \phi + \eta F'(\phi) \right), && \text{in } \Omega \times (0, T], && (4.157\text{b})\\
&\frac{\partial \mathbf{u}}{\partial t} + q\, \mathbf{u} \cdot \nabla \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \eta\, \mu \nabla \phi, && \text{in } \Omega \times (0, T], && (4.157\text{c})\\
&\nabla \cdot \mathbf{u} = 0, && \text{in } \Omega \times (0, T], && (4.157\text{d})\\
&\frac{d}{dt} \int_\Omega F(\phi)\, d\mathbf{x} + \theta_\eta \frac{d\eta^2}{dt} = \eta \int_\Omega F'(\phi) \frac{\partial \phi}{\partial t}\, d\mathbf{x}, && \text{in } (0, T], && (4.157\text{e})\\
&\frac{d}{dt} \int_\Omega \frac{1}{2} |\mathbf{u}|^2\, d\mathbf{x} + \theta_q \frac{dq^2}{dt} \\
&\quad = \int_\Omega [\mathbf{u}_t + q(\mathbf{u} \cdot \nabla \mathbf{u})] \cdot \mathbf{u}\, d\mathbf{x} + \eta \int_\Omega [(\mathbf{u} \cdot \nabla \phi)\mu - (\mu \nabla \phi) \cdot \mathbf{u}]\, d\mathbf{x}, && \text{in } (0, T]. && (4.157\text{f})
\end{aligned}
$$

where θ_η and θ_q are two positive regularization parameters associated with the CH and NS parts, respectively. We remark that the role of η and q is to enforce dissipation of the total energy while the regularization parameters θ_η and θ_q are crucial for relaxing time step size constraints caused by explicit treatment of nonlinear functions in the proposed DRLM schemes derived later in the next section. Moreover, θ_q guarantees the uniqueness of the Lagrange multiplier q, so that the following lemma holds.

Lemma 4.21. The system (4.157) is equivalent to the original CH-NS model (4.151).

Proof. If $\phi, \mu, \mathbf{u}$, and p solve the original model (4.151), then they, together with $\eta \equiv 1$ and $q \equiv 1$, satisfy the reformulated system (4.157). Conversely, if (4.157) admits a solution $\phi, \mu, \mathbf{u}, p, \eta$, and q, we observe that

$$
\int_\Omega (\mathbf{u} \cdot \nabla \mathbf{u}) \cdot \mathbf{u}\, d\mathbf{x} = 0, \tag{4.158}
$$

$$
\int_\Omega [(\mathbf{u} \cdot \nabla \phi)\mu - (\mu \nabla \phi) \cdot \mathbf{u}]\, d\mathbf{x} = 0. \tag{4.159}
$$

From (4.157f), (4.158) and (4.159), along with $q(0) = 1$, we find that $q(t) \equiv 1$ for all $t > 0$. On the other hand, (4.157e) can be simplified as

$$\eta\frac{d\eta}{dt} = \alpha(\eta - 1),$$

where $\alpha = \frac{1}{2\theta_\eta}\frac{d}{dt}\int_\Omega F(\phi)\,d\mathbf{x}$ is a time-dependent function. It follows that

$$\begin{aligned}
&\frac{d}{dt}\left[(\eta(t)-1)\exp\left(\eta(t)-\int_0^t \alpha(s)ds\right)\right]\\
&= \exp\left(\eta(t)-\int_0^t \alpha(s)ds\right)\left[\frac{d\eta(t)}{dt}+(\eta(t)-1)\left(\frac{d\eta(t)}{dt}-\alpha(t)\right)\right]\\
&= \exp\left(\eta(t)-\int_0^t \alpha(s)ds\right)\left[\eta(t)\frac{d\eta(t)}{dt}-\alpha(t)(\eta(t)-1)\right]=0.
\end{aligned} \tag{4.160}$$

Since $\eta(0) = 1$, we obtain from (4.160) that $\eta(t) \equiv 1$ for all $t > 0$. Consequently, $\phi, \mu, \mathbf{u}$, and p satisfy the system (4.151), which completes the proof of the lemma. □

4.3.2 Dynamically regularized Lagrange multiplier schemes

We consider a uniform partition of the time interval $[0, T]$: $0 = t_0 < t_1 < \cdots < t_N = T$ with the time step size $\Delta t = T/N$ for the time discretization of the system (4.157) with the two Lagrange multipliers. In the following, we derive the first- and second-order dynamically regularized Lagrange multiplier (DRLM) schemes, along with their efficient numerical implementation.

The first-order DRLM scheme For $\psi \in \{\phi, \mu, \mathbf{u}, p, \eta, q\}$ and $0 \leq n \leq N$, let ψ^n be an approximation of $\psi(t_n)$. By applying the first-order BDF (i.e., backward Euler) scheme to the system (4.157) with explicit treatment of the coupling term $\mathbf{u}\cdot\nabla\phi$ and nonlinear terms $F'(\phi)$ and $\mathbf{u}\cdot\nabla\mathbf{u}$, we obtain the following first-order (in time) DRLM scheme: for $0 \leq n \leq N-1$,

$$\frac{\phi^{n+1}-\phi^n}{\Delta t} + \eta^{n+1}\mathbf{u}^n \cdot \nabla\phi^n = M\Delta\mu^{n+1}, \tag{4.161a}$$

$$\mu^{n+1} = \lambda\left(-\Delta\phi^{n+1} + \eta^{n+1}F'(\phi^n)\right), \tag{4.161b}$$

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t} + q^{n+1}\mathbf{u}^n \cdot \nabla\mathbf{u}^n - \nu\Delta\mathbf{u}^{n+1} + \nabla p^{n+1} = \eta^{n+1}\mu^{n+1}\nabla\phi^{n+1}, \tag{4.161c}$$

$$\nabla \cdot \mathbf{u}^{n+1} = 0, \tag{4.161d}$$

$$\int_\Omega \left[F(\phi^{n+1}) - F(\phi^n)\right] d\mathbf{x} + \theta_\eta\left[(\eta^{n+1})^2 - (\eta^n)^2\right] = \eta^{n+1}\int_\Omega F'(\phi^n)\left(\phi^{n+1}-\phi^n\right) d\mathbf{x}, \tag{4.161e}$$

$$\begin{aligned}\int_\Omega \frac{|\mathbf{u}^{n+1}|^2 - |\mathbf{u}^n|^2}{2}\, d\mathbf{x} + \theta_q\left[(q^{n+1})^2 - (q^n)^2\right] &= \int_\Omega \left(\mathbf{u}^{n+1} - \mathbf{u}^n + \Delta t q^{n+1}\mathbf{u}^n \cdot \nabla\mathbf{u}^n\right) \cdot \mathbf{u}^{n+1} d\mathbf{x} \\ &+ \Delta t\eta^{n+1}\int_\Omega \left[(\mathbf{u}^n \cdot \nabla\phi^n)\mu^{n+1} - (\mu^{n+1}\nabla\phi^{n+1}) \cdot \mathbf{u}^{n+1}\right] d\mathbf{x}.\end{aligned} \tag{4.161f}$$

We note that the right-hand side of (4.157c) is treated implicitly in (4.161c) because the CH variables at t_{n+1}, namely ϕ^{n+1}, μ^{n+1}, and η^{n+1}, can be computed from (4.161a), (4.161b), and (4.161e) before solving the NS equations (4.161c), (4.161d), and (4.161f) for $\mathbf{u}^{n+1}$, p^{n+1}, and q^{n+1}. In particular, let us decompose each of the unknown quantities ϕ^{n+1}, μ^{n+1}, $\mathbf{u}^{n+1}$, and p^{n+1} into two components as follows:

$$\phi^{n+1} = \phi_1^{n+1} + \eta^{n+1}\phi_2^{n+1}, \tag{4.162a}$$

$$\mu^{n+1} = \mu_1^{n+1} + \eta^{n+1}\mu_2^{n+1}, \tag{4.162b}$$

$$\mathbf{u}^{n+1} = \mathbf{u}_1^{n+1} + q^{n+1}\mathbf{u}_2^{n+1}, \tag{4.162c}$$

$$p^{n+1} = p_1^{n+1} + q^{n+1}p_2^{n+1}. \tag{4.162d}$$

Plugging (4.162a)-(4.162b) into (4.161a)-(4.161b), we obtain the following two linear equations in Ω for $(\phi_i^{n+1}, \mu_i^{n+1})$, $i = 1, 2$:

$$\frac{\phi_1^{n+1} - \phi^n}{\Delta t} = M\Delta\mu_1^{n+1}, \qquad \mu_1^{n+1} = -\lambda\Delta\phi_1^{n+1}; \tag{4.163}$$

$$\frac{\phi_2^{n+1}}{\Delta t} + \mathbf{u}^n \cdot \nabla\phi^n = M\Delta\mu_2^{n+1}, \qquad \mu_2^{n+1} = \lambda\left(-\Delta\phi_2^{n+1} + F'(\phi^n)\right), \tag{4.164}$$

where ϕ_i^{n+1} and μ_i^{n+1} ($i = 1, 2$) satisfy the same boundary conditions as ϕ^{n+1} and μ^{n+1}, respectively, i.e., either periodic or no-flux boundary conditions.

With $(\phi_i^{n+1}, \mu_i^{n+1})$, $i = 1, 2$, determined from the above equations (4.163) and (4.164), we compute the Lagrange multiplier η^{n+1} from (4.161e) by solving the following nonlinear algebraic equation:

$$\int_\Omega F(\phi_1^{n+1} + \eta^{n+1}\phi_2^{n+1})\, d\mathbf{x} + A_{1,n+1}^{\eta}(\eta^{n+1})^2 + B_{1,n+1}^{\eta}\eta^{n+1} + C_{1,n+1}^{\eta} = 0, \tag{4.165}$$

where the quadratic coefficients are given by

$$\begin{aligned}
A_{1,n+1}^{\eta} &= \theta_\eta - \int_\Omega F'(\phi^n)\phi_2^{n+1}\, d\mathbf{x},\\
B_{1,n+1}^{\eta} &= -\int_\Omega F'(\phi^n)\left(\phi_1^{n+1} - \phi^n\right) d\mathbf{x},\\
C_{1,n+1}^{\eta} &= -\theta_\eta(\eta^n)^2 - \int_\Omega F(\phi^n)\, d\mathbf{x}.
\end{aligned}$$

The complexity of the nonlinear equation (4.165) depends on the potential function $F(\phi)$. In this section, we focus on the commonly used double-well potential; hence, (4.165) is a fourth-order algebraic equation. Given that the exact value of η is 1, we solve (4.165) using Newton's iterative method with an initial guess of 1. Once η^{n+1} is known, we update ϕ^{n+1} and μ^{n+1} from (4.162a) and (4.162b), respectively.

Regarding the NS equations, substituting (4.162c)-(4.162d) into (4.161c)-(4.161d) yields the following two linear systems of generalized Stokes equations in Ω for $(\mathbf{u}_i^{n+1}, p_i^{n+1})$, $i = 1, 2$:

$$\frac{\mathbf{u}_1^{n+1} - \mathbf{u}^n}{\Delta t} - \nu\Delta\mathbf{u}_1^{n+1} + \nabla p_1^{n+1} = \eta^{n+1}\mu^{n+1}\nabla\phi^{n+1}, \qquad \nabla\cdot\mathbf{u}_1^{n+1} = 0; \tag{4.166}$$

$$\frac{\mathbf{u}_2^{n+1}}{\Delta t} + \mathbf{u}^n\cdot\nabla\mathbf{u}^n - \nu\Delta\mathbf{u}_2^{n+1} + \nabla p_2^{n+1} = 0, \qquad \nabla\cdot\mathbf{u}_2^{n+1} = 0, \tag{4.167}$$

where $\mathbf{u}_1^{n+1}$ and $\mathbf{u}_2^{n+1}$ satisfy the same boundary conditions as $\mathbf{u}^{n+1}$, i.e., either periodic or no-slip boundary conditions.

Once $\mathbf{u}_i^{n+1}$ and p_i^{n+1} ($i = 1, 2$) are determined from the above systems (4.166) and (4.167), the Lagrange multiplier q^{n+1} is computed from (4.161f) by solving the quadratic equation:

$$A_{1,n+1}^{q}(q^{n+1})^2 + B_{1,n+1}^{q}q^{n+1} + C_{1,n+1}^{q} = 0, \tag{4.168}$$

where the quadratic coefficients are defined as:

$$
\begin{aligned}
A^{q}_{1,n+1} &= \theta_q + \frac{1}{2}\|\mathbf{u}_2^{n+1}\|_0^2 + \Delta t\nu\|\nabla\mathbf{u}_2^{n+1}\|_0^2,\\
B^{q}_{1,n+1} &= -\int_\Omega (\mathbf{u}_1^{n+1} - \mathbf{u}^n)\cdot\mathbf{u}_2^{n+1}d\mathbf{x} + \Delta t\int_\Omega \left[\eta^{n+1}\left(\mu^{n+1}\nabla\phi^{n+1}\right)\cdot\mathbf{u}_2^{n+1} - (\mathbf{u}^n\cdot\nabla\mathbf{u}^n)\cdot\mathbf{u}_1^{n+1}\right]d\mathbf{x},\\
C^{q}_{1,n+1} &= -\theta_q(q^n)^2 - \frac{1}{2}\|\mathbf{u}_1^{n+1} - \mathbf{u}^n\|_0^2 - \Delta t\eta^{n+1}\int_\Omega \left[(\mathbf{u}^n\cdot\nabla\phi^n)\mu^{n+1} - (\mu^{n+1}\nabla\phi^{n+1})\cdot\mathbf{u}_1^{n+1}\right]d\mathbf{x}.
\end{aligned}
$$

Note that in the derivation of $A^{q}_{1,n+1}$, we have used the fact that

$$
\Delta t\int_\Omega \left(\mathbf{u}^n\cdot\nabla\mathbf{u}^n\right)\cdot\mathbf{u}_2^{n+1}\,d\mathbf{x} = -\|\mathbf{u}_2^{n+1}\|_0^2 - \Delta t\nu\|\nabla\mathbf{u}_2^{n+1}\|_0^2
$$

obtained from (4.167) and integration by parts. Since $A^{q}_{1,n+1} > 0$ for all $\theta_q > 0$, (4.168) has two solutions, and q^{n+1} is chosen as the one closer to 1. Finally, we update $\mathbf{u}^{n+1}$ and p^{n+1} via (4.162c) and (4.162d), respectively.

The second-order DRLM scheme Let us denote by $\tilde{\phi}^{n+1} = 2\phi^n - \phi^{n-1}$ and $\tilde{\mathbf{u}}^{n+1} = 2\mathbf{u}^n - \mathbf{u}^{n-1}$ the second-order extrapolations to approximate $\mathbf{u}\cdot\nabla\phi$, $F'(\phi)$, and $\mathbf{u}\cdot\nabla\mathbf{u}$ at t_{n+1}. Applying the BDF2 scheme to the system (4.157) yields the following second-order (in time) DRLM scheme:

for $1 \leq n \leq N-1$,

$$\frac{3\phi^{n+1}-4\phi^{n}+\phi^{n-1}}{2\Delta t}+\eta^{n+1}\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\phi}^{n+1}=M\Delta\mu^{n+1}, \tag{4.169a}$$

$$\mu^{n+1}=\lambda\left(-\Delta\phi^{n+1}+\eta^{n+1}F'(\tilde{\phi}^{n+1})\right), \tag{4.169b}$$

$$\frac{3\mathbf{u}^{n+1}-4\mathbf{u}^{n}+\mathbf{u}^{n-1}}{2\Delta t}+q^{n+1}\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\mathbf{u}}^{n+1}-\nu\Delta\mathbf{u}^{n+1}+\nabla p^{n+1}=\eta^{n+1}\mu^{n+1}\nabla\phi^{n+1}, \tag{4.169c}$$

$$\nabla\cdot\mathbf{u}^{n+1}=0, \tag{4.169d}$$

$$\begin{aligned}\int_{\Omega}\left[3F(\phi^{n+1})-4F(\phi^{n})+F(\phi^{n-1})\right]d\mathbf{x}+\theta_{\eta}\left[3(\eta^{n+1})^2-4(\eta^{n})^2+(\eta^{n-1})^2\right]&\\ =\eta^{n+1}\int_{\Omega}F'(\tilde{\phi}^{n+1})\left(3\phi^{n+1}-4\phi^{n}+\phi^{n-1}\right)d\mathbf{x},&\end{aligned} \tag{4.169e}$$

$$\begin{aligned}\int_{\Omega}\frac{3|\mathbf{u}^{n+1}|^2-4|\mathbf{u}^{n}|^2+|\mathbf{u}^{n-1}|^2}{2}\,d\mathbf{x}+\theta_{q}\left[3(q^{n+1})^2-4(q^{n})^2+(q^{n-1})^2\right]&\\ =\int_{\Omega}\left(3\mathbf{u}^{n+1}-4\mathbf{u}^{n}+\mathbf{u}^{n-1}+2\Delta t q^{n+1}\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\mathbf{u}}^{n+1}\right)\cdot\mathbf{u}^{n+1}d\mathbf{x}&\\ +2\Delta t\eta^{n+1}\int_{\Omega}\left[\left(\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\phi}^{n+1}\right)\mu^{n+1}-\left(\mu^{n+1}\nabla\phi^{n+1}\right)\cdot\mathbf{u}^{n+1}\right]d\mathbf{x}.&\end{aligned} \tag{4.169f}$$

In order to initialize the second-order DRLM scheme (4.169), we use $\{\phi^0, \mathbf{u}^0, \eta^0 = 1, q^0 = 1\}$ from the initial conditions (4.152) and $\{\phi^1, \mathbf{u}^1, \eta^1, q^1\}$ computed by the first-order DRLM scheme (4.161). It should be noted that the pressure p and chemical potential μ at t_{n-1} and t_n are not required to compute the solution $\{\phi^{n+1}, \mu^{n+1}, \eta^{n+1}, \mathbf{u}^{n+1}, p^{n+1}, q^{n+1}\}$ at time t_{n+1}. We now describe how to efficiently implement the second-order DRLM scheme (4.169). Similarly to the first-order case, we again consider the decomposition (4.162) for the unknown quantities ϕ^{n+1}, μ^{n+1}, $\mathbf{u}^{n+1}$, and p^{n+1}. By substituting (4.162a)-(4.162b) into (4.169a)-(4.169b), we get the following two linear equations for solving $(\phi_i^{n+1}, \mu_i^{n+1})$, $i = 1, 2$:

$$\frac{3\phi_1^{n+1}-4\phi^{n}+\phi^{n-1}}{2\Delta t}=M\Delta\mu_1^{n+1}, \qquad \mu_1^{n+1}=-\lambda\Delta\phi_1^{n+1}; \tag{4.170}$$

$$\frac{3\phi_2^{n+1}}{2\Delta t}+\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\phi}^{n+1}=M\Delta\mu_2^{n+1}, \qquad \mu_2^{n+1}=\lambda\left(-\Delta\phi_2^{n+1}+F'(\tilde{\phi}^{n+1})\right). \tag{4.171}$$

Next, we update the Lagrange multiplier η^{n+1} from (4.169e), which can be written as:

$$3\int_{\Omega}F\left(\phi_1^{n+1}+\eta^{n+1}\phi_2^{n+1}\right)d\mathbf{x}+A^{\eta}_{2,n+1}(\eta^{n+1})^2+B^{\eta}_{2,n+1}\eta^{n+1}+C^{\eta}_{2,n+1}=0, \tag{4.172}$$

where

$$
\begin{aligned}
A^{\eta}_{2,n+1} &= 3\theta_\eta - 3\int_\Omega F'(\tilde{\phi}^{n+1})\phi_2^{n+1} d\mathbf{x},\\
B^{\eta}_{2,n+1} &= -\int_\Omega F'(\tilde{\phi}^{n+1})\left(3\phi_1^{n+1} - 4\phi^n + \phi^{n-1}\right) d\mathbf{x},\\
C^{\eta}_{2,n+1} &= \theta_\eta\left[-4(\eta^n)^2 + (\eta^{n-1})^2\right] + \int_\Omega \left[-4F(\phi^n) + F(\phi^{n-1})\right] d\mathbf{x}.
\end{aligned}
$$

Note that (4.172) can be solved iteratively using Newton's method with an initial guess of 1. After determining η^{n+1} from (4.172), we update ϕ^{n+1} and μ^{n+1} according to (4.162a) and (4.162b), respectively. The combination of (4.162c)-(4.162d) and (4.169c)-(4.169d) results in the following two linear Stokes systems for $(\mathbf{u}_i^{n+1}, p_i^{n+1})$, $i = 1, 2$:

$$
\frac{3\mathbf{u}_1^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} - \nu\Delta\mathbf{u}_1^{n+1} + \nabla p_1^{n+1} = \eta^{n+1}\mu^{n+1}\nabla\phi^{n+1}, \qquad \nabla\cdot\mathbf{u}_1^{n+1} = 0; \tag{4.173}
$$

$$
\frac{3\mathbf{u}_2^{n+1}}{2\Delta t} + \tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\mathbf{u}}^{n+1} - \nu\Delta\mathbf{u}_2^{n+1} + \nabla p_2^{n+1} = 0, \qquad \nabla\cdot\mathbf{u}_2^{n+1} = 0. \tag{4.174}
$$

Given $(\mathbf{u}_i^{n+1}, p_i^{n+1})$ for $i = 1, 2$ from (4.173)-(4.174) and simplifying (4.169f), we find that q^{n+1} satisfies the following quadratic equation:

$$
A^{q}_{2,n+1}(q^{n+1})^2 + B^{q}_{2,n+1}q^{n+1} + C^{q}_{2,n+1} = 0, \tag{4.175}
$$

where

$$
\begin{aligned}
A^{q}_{2,n+1} &= 3\theta_q + \frac{3}{2}\|\mathbf{u}_2^{n+1}\|_0^2 + 2\Delta t\nu\|\nabla\mathbf{u}_2^{n+1}\|_0^2,\\
B^{q}_{2,n+1} &= -\int_\Omega \left(3\mathbf{u}_1^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}\right)\cdot\mathbf{u}_2^{n+1} d\mathbf{x}\\
&\qquad + 2\Delta t\int_\Omega \left[\eta^{n+1}\left(\mu^{n+1}\nabla\phi^{n+1}\right)\cdot\mathbf{u}_2^{n+1} - \left(\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\mathbf{u}}^{n+1}\right)\cdot\mathbf{u}_1^{n+1}\right] d\mathbf{x},\\
C^{q}_{2,n+1} &= \theta_q\left[-4(q^n)^2 + (q^{n-1})^2\right] - 2\|\mathbf{u}_1^{n+1} - \mathbf{u}^n\|_0^2 + \frac{1}{2}\|\mathbf{u}_1^{n+1} - \mathbf{u}^{n-1}\|_0^2\\
&\qquad - 2\Delta t\eta^{n+1}\int_\Omega \left[\left(\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\phi}^{n+1}\right)\mu^{n+1} - \left(\mu^{n+1}\nabla\phi^{n+1}\right)\cdot\mathbf{u}_1^{n+1}\right] d\mathbf{x}.
\end{aligned}
$$

To derive $A^{q}_{2,n+1}$, we have used (4.174) and integration by parts. Since $A^{q}_{2,n+1} > 0$ for all $\theta_q > 0$, equation (4.175) has two solutions, with q^{n+1} selected as the one closer to 1. Finally, we determine $\mathbf{u}^{n+1}$ and p^{n+1} from (4.162c) and (4.162d), respectively.

4.3.3 Properties of the time-discrete solutions

In the following, the first- and second-order DRLM schemes (cf. (4.161) and (4.169), respectively) are shown to be mass conserving and unconditionally energy stable. We remark that, differently from existing studies on SAV-based schemes, our energy stability results are established with respect to the original variables rather than the modified ones.

Mass conservation We verify that, under the prescribed boundary conditions, the proposed time-discrete DRLM schemes conserve mass over time.

Theorem 4.22. The proposed first-order (4.161) and second-order (4.169) DRLM schemes satisfy mass conservation:

$$\int_\Omega \phi^{n+1} d\mathbf{x} = \int_\Omega \phi^n d\mathbf{x}, \quad n = 0, 1, \dots, N-1. \tag{4.176}$$

Proof. For the first-order scheme (4.161), taking the integral of (4.161a) over Ω gives us

$$\frac{1}{\Delta t}\int_\Omega (\phi^{n+1} - \phi^n)\, d\mathbf{x} + \eta^{n+1}\int_\Omega \mathbf{u}^n \cdot \nabla\phi^n \, d\mathbf{x} = M\int_\Omega \Delta\mu^{n+1} d\mathbf{x}. \tag{4.177}$$

Since μ^{n+1} satisfies either periodic or homogeneous Neumann boundary conditions, integration by parts yields

$$\int_\Omega \Delta\mu^{n+1} d\mathbf{x} = \int_{\partial\Omega} \nabla\mu^{n+1} \cdot \mathbf{n}\, ds = 0. \tag{4.178}$$

In light of the incompressibility condition (4.161d) and the boundary conditions of $\mathbf{u}^n$ and ϕ^n, we observe that

$$\int_\Omega \mathbf{u}^n \cdot \nabla\phi^n \, d\mathbf{x} = \int_\Omega \nabla \cdot (\phi^n \mathbf{u}^n)\, d\mathbf{x} = \int_{\partial\Omega} (\phi^n \mathbf{u}^n) \cdot \mathbf{n}\, ds = 0. \tag{4.179}$$

The combination of (4.177)-(4.179) results in

$$\int_{\Omega} \left(\phi^{n+1} - \phi^{n}\right) d\mathbf{x} = 0,$$

which indicates the mass conservation of the first-order scheme (4.161).

For the second-order counterpart (4.169), it is clear that (4.176) holds for $n = 0$. The proof is completed by induction, noting from (4.169a) that

$$\frac{1}{2\Delta t}\int_{\Omega} \left(3\phi^{n+1} - 4\phi^{n} + \phi^{n-1}\right) d\mathbf{x} = M\int_{\Omega} \Delta\mu^{n+1} d\mathbf{x} - \eta^{n+1}\int_{\Omega} \nabla\cdot\left(\tilde{\phi}^{n+1}\tilde{\mathbf{u}}^{n+1}\right) d\mathbf{x} = 0,$$

where we have used the fact that $\tilde{\mathbf{u}}^{n+1} = 2\mathbf{u}^{n} - \mathbf{u}^{n-1}$ is divergence-free. □

Energy stability Our goal in this subsection is to show that the two proposed DRLM schemes satisfy energy stability with respect to the original energy functional (4.154).

Theorem 4.23. The first-order DRLM scheme (4.161) is unconditionally energy stable in the sense that for $0 \le n \le N-1$, we have

$$E(\phi^{n+1}, \mathbf{u}^{n+1}) + \mathcal{E}_1(\eta^{n+1}, q^{n+1}) \le E(\phi^{n}, \mathbf{u}^{n}) + \mathcal{E}_1(\eta^{n}, q^{n}), \tag{4.180}$$

where

$$\mathcal{E}_1(\eta^{n}, q^{n}) := \lambda\theta_\eta\left[(\eta^{n})^2 - 1\right] + \theta_q\left[(q^{n})^2 - 1\right].$$

Proof. Taking the L^2 inner product of (4.161b) with $\phi^{n+1} - \phi^{n}$ and (4.161a) with $\Delta t\mu^{n+1}$ yields

$$\begin{aligned}
\lambda\int_{\Omega} &\left(-\Delta\phi^{n+1} + \eta^{n+1}F'(\phi^{n})\right)\left(\phi^{n+1} - \phi^{n}\right) d\mathbf{x}\\
&= \int_{\Omega} \mu^{n+1}\left(\phi^{n+1} - \phi^{n}\right) d\mathbf{x}\\
&= -\Delta t\eta^{n+1}\int_{\Omega} (\mathbf{u}^{n}\cdot\nabla\phi^{n})\mu^{n+1} d\mathbf{x} - \Delta t M\|\nabla\mu^{n+1}\|_0^2\\
&\le -\Delta t\eta^{n+1}\int_{\Omega} (\mathbf{u}^{n}\cdot\nabla\phi^{n})\mu^{n+1} d\mathbf{x}.
\end{aligned} \tag{4.181}$$

It follows from integration by parts that

$$
\begin{aligned}
\int_\Omega (-\Delta\phi^{n+1})(\phi^{n+1}-\phi^n)\,d\mathbf{x} &= \int_\Omega \nabla\phi^{n+1}\cdot\nabla(\phi^{n+1}-\phi^n)\,d\mathbf{x} \\
&= \frac{1}{2}\left(\|\nabla\phi^{n+1}\|_0^2-\|\nabla\phi^n\|_0^2\right)+\frac{1}{2}\|\nabla(\phi^{n+1}-\phi^n)\|_0^2 \\
&\geq \frac{1}{2}\left(\|\nabla\phi^{n+1}\|_0^2-\|\nabla\phi^n\|_0^2\right).
\end{aligned}
\tag{4.182}
$$

The combination of (4.181) and (4.182) results in

$$
\begin{aligned}
&\lambda\left[\frac{1}{2}\left(\|\nabla\phi^{n+1}\|_0^2-\|\nabla\phi^n\|_0^2\right)+\eta^{n+1}\int_\Omega F'(\phi^n)(\phi^{n+1}-\phi^n)\,d\mathbf{x}\right] \\
&\qquad\leq -\Delta t\eta^{n+1}\int_\Omega (\mathbf{u}^n\cdot\nabla\phi^n)\mu^{n+1}d\mathbf{x}.
\end{aligned}
$$

This, together with (4.161e), implies that

$$
E_{CH}(\phi^{n+1})-E_{CH}(\phi^n)+\lambda\theta_\eta\left[(\eta^{n+1})^2-(\eta^n)^2\right]\leq -\Delta t\eta^{n+1}\int_\Omega (\mathbf{u}^n\cdot\nabla\phi^n)\mu^{n+1}d\mathbf{x}. \tag{4.183}
$$

On the other hand, taking the L^2 inner product of (4.161c) with $\Delta t\mathbf{u}^{n+1}$ gives us

$$
\begin{aligned}
&\int_\Omega \left(\mathbf{u}^{n+1}-\mathbf{u}^n+\Delta t q^{n+1}\mathbf{u}^n\cdot\nabla\mathbf{u}^n\right)\cdot\mathbf{u}^{n+1}d\mathbf{x}-\Delta t\eta^{n+1}\int_\Omega (\mu^{n+1}\nabla\phi^{n+1})\cdot\mathbf{u}^{n+1}d\mathbf{x} \\
&\qquad = \Delta t\nu\int_\Omega \Delta\mathbf{u}^{n+1}\cdot\mathbf{u}^{n+1}d\mathbf{x} = -\Delta t\nu\|\nabla\mathbf{u}^{n+1}\|_0^2\leq 0,
\end{aligned}
\tag{4.184}
$$

where we have used the fact that $(\nabla p^{n+1},\mathbf{u}^{n+1}) = 0$ due to the incompressibility condition (4.161d) and integration by parts. From (4.161f) and (4.184), we obtain

$$
E_{NS}(\mathbf{u}^{n+1})-E_{NS}(\mathbf{u}^n)+\theta_q\left[(q^{n+1})^2-(q^n)^2\right]\leq \Delta t\eta^{n+1}\int_\Omega (\mathbf{u}^n\cdot\nabla\phi^n)\mu^{n+1}d\mathbf{x}. \tag{4.185}
$$

The sum of (4.183) and (4.185) leads to

$$
E(\phi^{n+1},\mathbf{u}^{n+1})+\lambda\theta_\eta(\eta^{n+1})^2+\theta_q(q^{n+1})^2\leq E(\phi^n,\mathbf{u}^n)+\lambda\theta_\eta(\eta^n)^2+\theta_q(q^n)^2,
$$

which completes the proof of Theorem 4.23. □

Remark 4.24. From Theorem 4.23, we deduce that

$$\begin{aligned} E(\phi^n, \mathbf{u}^n) + \lambda\theta_\eta(\eta^n)^2 + \theta_q(q^n)^2 &\leq E(\phi^0, \mathbf{u}^0) + \lambda\theta_\eta(\eta^0)^2 + \theta_q(q^0)^2 \\ &= E(\phi^0, \mathbf{u}^0) + \lambda\theta_\eta + \theta_q, \end{aligned} \tag{4.186}$$

for $0 \leq n \leq N$. Therefore, the discrete energy sequence $\{E(\phi^n, \mathbf{u}^n)\}_{n=0}^N$ is uniformly bounded.

Remark 4.25. Energy stability of the first-order DRLM scheme can also be proven for the CH-NS system with a forcing term $\mathbf{f}$ (cf. [129] and (4.217)), i.e., the momentum equation (4.151c) is replaced by

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = \mu\nabla\phi + \mathbf{f}. \tag{4.187}$$

In particular, the counterpart of (4.186) takes the following form

$$E(\phi^n, \mathbf{u}^n) + \lambda\theta_\eta(\eta^n)^2 + \theta_q(q^n)^2 \leq e^{2T}\left(E(\phi^0, \mathbf{u}^0) + \lambda\theta_\eta + \theta_q + \frac{1}{2}\|\mathbf{f}\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))}\right), \tag{4.188}$$

provided that $\Delta t \leq \frac{1}{2}$ and $\mathbf{f} \in L^\infty(0, T; \mathbf{L}^2(\Omega))$. Furthermore, under the no-flux/no-slip boundary conditions (4.153), the scheme is unconditionally energy stable if $\mathbf{f} \in L^\infty(0, T; \mathbf{H}^{-1}(\Omega))$; that is, (4.186) becomes

$$E(\phi^n, \mathbf{u}^n) + \lambda\theta_\eta(\eta^n)^2 + \theta_q(q^n)^2 \leq E(\phi^0, \mathbf{u}^0) + \lambda\theta_\eta + \theta_q + \frac{C_\Omega^2\, T}{4\nu}\|\mathbf{f}\|^2_{L^\infty(0,T;\mathbf{H}^{-1}(\Omega))}, \tag{4.189}$$

for some constant $C_\Omega > 0$ only depending on Ω.

To prove (4.188), we first use similar arguments as in the proof of Theorem 4.23 to obtain

$$\begin{aligned} &E(\phi^{n+1}, \mathbf{u}^{n+1}) + \lambda\theta_\eta(\eta^{n+1})^2 + \theta_q(q^{n+1})^2 \\ &\leq E(\phi^n, \mathbf{u}^n) + \lambda\theta_\eta(\eta^n)^2 + \theta_q(q^n)^2 - \Delta t\nu\|\nabla\mathbf{u}^{n+1}\|_0^2 + \Delta t\int_\Omega \mathbf{f}(t_{n+1}) \cdot \mathbf{u}^{n+1} d\mathbf{x}. \end{aligned} \tag{4.190}$$

It is implied by Young's inequality and the definition of E in (4.154) that

$$\Delta t \int_\Omega \mathbf{f}(t_{n+1}) \cdot \mathbf{u}^{n+1} d\mathbf{x} \leq \frac{\Delta t}{2}\|\mathbf{f}(t_{n+1})\|_0^2 + \frac{\Delta t}{2}\|\mathbf{u}^{n+1}\|_0^2 \leq \frac{\Delta t}{2}\|\mathbf{f}(t_{n+1})\|_0^2 + \Delta t E(\phi^{n+1}, \mathbf{u}^{n+1}). \tag{4.191}$$

Combining (4.190) and (4.191), we end up with

$$\begin{aligned}(1-\Delta t)E(\phi^{n+1}, \mathbf{u}^{n+1}) + \lambda\theta_\eta(\eta^{n+1})^2 + \theta_q(q^{n+1})^2 \\ \leq E(\phi^n, \mathbf{u}^n) + \lambda\theta_\eta(\eta^n)^2 + \theta_q(q^n)^2 + \frac{\Delta t}{2}\|\mathbf{f}(t_{n+1})\|_0^2,\end{aligned}$$

or

$$\begin{aligned}&E(\phi^{n+1}, \mathbf{u}^{n+1}) + \lambda\theta_\eta(\eta^{n+1})^2 + \theta_q(q^{n+1})^2 \\ &\quad \leq \frac{1}{1-\Delta t}\left[E(\phi^n, \mathbf{u}^n) + \lambda\theta_\eta(\eta^n)^2 + \theta_q(q^n)^2\right] + \frac{\Delta t}{2(1-\Delta t)}\|\mathbf{f}(t_{n+1})\|_0^2.\end{aligned} \tag{4.192}$$

By repeatedly applying (4.192), one arrives at

$$\begin{aligned}&E(\phi^n, \mathbf{u}^n) + \lambda\theta_\eta(\eta^n)^2 + \theta_q(q^n)^2 \\ &\quad \leq \frac{1}{(1-\Delta t)^n}\left[E(\phi^0, \mathbf{u}^0) + \lambda\theta_\eta + \theta_q + \frac{\Delta t}{2}\sum_{i=0}^{n-1}(1-\Delta t)^i\|\mathbf{f}(t_{i+1})\|_0^2\right].\end{aligned}$$

This leads to (4.188), noting that $\frac{1}{(1-\Delta t)^n} \leq (1+2\Delta t)^n \leq e^{2T}$ (since $\Delta t \leq \frac{1}{2}$) and $\Delta t \sum_{i=0}^{n-1}(1-\Delta t)^i \leq 1$.

For the proof of (4.189), we use the Poincaré inequality for $\mathbf{u}^{n+1} \in \mathbf{H}_0^1(\Omega)$ (due to the no-flux/no-slip boundary conditions (4.153)):

$$\|\mathbf{u}^{n+1}\|_1 \leq C_\Omega\|\nabla\mathbf{u}^{n+1}\|_0. \tag{4.193}$$

Therefore, the integral term on the right-hand side of (4.190) can be estimated as follows

$$\begin{aligned}&\int_\Omega \mathbf{f}(t_{n+1}) \cdot \mathbf{u}^{n+1} d\mathbf{x} \leq \|\mathbf{f}(t_{n+1})\|_{-1}\|\mathbf{u}^{n+1}\|_1 \\ &\qquad \leq C_\Omega\|\mathbf{f}(t_{n+1})\|_{-1}\|\nabla\mathbf{u}^{n+1}\|_0 \leq \frac{C_\Omega^2}{4\nu}\|\mathbf{f}(t_{n+1})\|_{-1}^2 + \nu\|\nabla\mathbf{u}^{n+1}\|_0^2.\end{aligned} \tag{4.194}$$

From (4.190) and (4.194), we obtain

$$
\begin{aligned}
&E(\phi^{n+1}, \mathbf{u}^{n+1}) + \lambda\theta_\eta(\eta^{n+1})^2 + \theta_q(q^{n+1})^2 \\
&\quad \le E(\phi^n, \mathbf{u}^n) + \lambda\theta_\eta(\eta^n)^2 + \theta_q(q^n)^2 + \frac{C_\Omega^2 \Delta t}{4\nu}\|\mathbf{f}(t_{n+1})\|_{-1}^2.
\end{aligned}
$$

Theorem 4.26. The second-order DRLM scheme (4.169) is unconditionally energy stable in the sense that for $1 \le n \le N-1$, we have

$$
\begin{aligned}
&\frac{3}{2}E(\phi^{n+1}, \mathbf{u}^{n+1}) - \frac{1}{2}E(\phi^n, \mathbf{u}^n) + \frac{\lambda}{2}\|\nabla(\phi^{n+1} - \phi^n)\|_0^2 + \mathcal{E}_2(\eta^{n+1}, \eta^n, q^{n+1}, q^n) \\
&\quad \le \frac{3}{2}E(\phi^n, \mathbf{u}^n) - \frac{1}{2}E(\phi^{n-1}, \mathbf{u}^{n-1}) + \frac{\lambda}{2}\|\nabla(\phi^n - \phi^{n-1})\|_0^2 + \mathcal{E}_2(\eta^n, \eta^{n-1}, q^n, q^{n-1}),
\end{aligned}
\tag{4.195}
$$

where

$$
\mathcal{E}_2(\eta^{n+1}, \eta^n, q^{n+1}, q^n) := \lambda\theta_\eta\left[\frac{3}{2}(\eta^{n+1})^2 - \frac{1}{2}(\eta^n)^2 - 1\right] + \theta_q\left[\frac{3}{2}(q^{n+1})^2 - \frac{1}{2}(q^n)^2 - 1\right].
$$

Proof. By taking the L^2 inner product of (4.169b) with $3\phi^{n+1} - 4\phi^n + \phi^{n-1}$ and (4.169a) with $2\Delta t\mu^{n+1}$, we obtain

$$
\begin{aligned}
&\lambda\int_\Omega \left(-\Delta\phi^{n+1} + \eta^{n+1}F'(\tilde{\phi}^{n+1})\right)\left(3\phi^{n+1} - 4\phi^n + \phi^{n-1}\right)d\mathbf{x} \\
&\qquad = \int_\Omega \mu^{n+1}\left(3\phi^{n+1} - 4\phi^n + \phi^{n-1}\right)d\mathbf{x} \\
&\qquad = -2\Delta t\eta^{n+1}\int_\Omega (\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\phi}^{n+1})\mu^{n+1}d\mathbf{x} - 2\Delta tM\|\nabla\mu^{n+1}\|_0^2 \\
&\qquad \le -2\Delta t\eta^{n+1}\int_\Omega (\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\phi}^{n+1})\mu^{n+1}d\mathbf{x}.
\end{aligned}
\tag{4.196}
$$

We observe from integration by parts that

$$
\begin{aligned}
&\int_\Omega \left(-\Delta\phi^{n+1}\right)\left(3\phi^{n+1} - 4\phi^n + \phi^{n-1}\right)d\mathbf{x} = \int_\Omega \nabla\phi^{n+1}\cdot\nabla\left(3\phi^{n+1} - 4\phi^n + \phi^{n-1}\right)d\mathbf{x} \\
&\quad \ge \frac{1}{2}\left(3\|\nabla\phi^{n+1}\|_0^2 - 4\|\nabla\phi^n\|_0^2 + \|\nabla\phi^{n-1}\|_0^2\right) + \|\nabla(\phi^{n+1} - \phi^n)\|_0^2 - \|\nabla(\phi^n - \phi^{n-1})\|_0^2,
\end{aligned}
\tag{4.197}
$$

where in the last estimate, we have applied the following inequality

$$a(3a-4b+c) \geq \frac{1}{2}(3a^2-4b^2+c^2)+(a-b)^2-(b-c)^2, \quad \forall\, a,b,c \in \mathbb{R}.$$

Combining (4.196) and (4.197), we arrive at

$$\begin{aligned}
&\lambda\Big[\frac{1}{2}\big(3\|\nabla\phi^{n+1}\|_0^2-4\|\nabla\phi^{n}\|_0^2+\|\nabla\phi^{n-1}\|_0^2\big)+\|\nabla(\phi^{n+1}-\phi^{n})\|_0^2-\|\nabla(\phi^{n}-\phi^{n-1})\|_0^2\\
&+\eta^{n+1}\int_\Omega F'(\tilde{\phi}^{n+1})\big(3\phi^{n+1}-4\phi^{n}+\phi^{n-1}\big)\,d\mathbf{x}\Big] \leq -2\Delta t\eta^{n+1}\int_\Omega(\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\phi}^{n+1})\mu^{n+1}d\mathbf{x}.
\end{aligned}$$

This, together with (4.169e), leads to

$$\begin{aligned}
3E_{CH}(\phi^{n+1})-4E_{CH}(\phi^{n})+E_{CH}(\phi^{n-1})+\lambda\|\nabla(\phi^{n+1}-\phi^{n})\|_0^2-\lambda\|\nabla(\phi^{n}-\phi^{n-1})\|_0^2&\\
+\lambda\theta_\eta\left[3(\eta^{n+1})^2-4(\eta^{n})^2+(\eta^{n-1})^2\right] \leq -2\Delta t\eta^{n+1}\int_\Omega(\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\phi}^{n+1})\mu^{n+1}d\mathbf{x}.&
\end{aligned} \tag{4.198}$$

Taking the L^2 inner product of (4.169c) with $2\Delta t\mathbf{u}^{n+1}$ yields

$$\begin{aligned}
&\int_\Omega\big(3\mathbf{u}^{n+1}-4\mathbf{u}^{n}+\mathbf{u}^{n-1}+2\Delta tq^{n+1}\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\mathbf{u}}^{n+1}\big)\cdot\mathbf{u}^{n+1}\,d\mathbf{x}\\
&\qquad-2\Delta t\eta^{n+1}\int_\Omega(\mu^{n+1}\nabla\phi^{n+1})\cdot\mathbf{u}^{n+1}d\mathbf{x}=-2\Delta t\nu\|\nabla\mathbf{u}^{n+1}\|_0^2\leq 0.
\end{aligned} \tag{4.199}$$

From (4.169f) and (4.199), we deduce that

$$\begin{aligned}
3E_{NS}(\mathbf{u}^{n+1})-4E_{NS}(\mathbf{u}^{n})+E_{NS}(\mathbf{u}^{n-1})+\theta_q\left[3(q^{n+1})^2-4(q^{n})^2+(q^{n-1})^2\right]&\\
\leq 2\Delta t\eta^{n+1}\int_\Omega(\tilde{\mathbf{u}}^{n+1}\cdot\nabla\tilde{\phi}^{n+1})\mu^{n+1}d\mathbf{x}.&
\end{aligned} \tag{4.200}$$

Adding (4.198) and (4.200) results in

$$\begin{aligned}
3E(\phi^{n+1},\mathbf{u}^{n+1})-4E(\phi^{n},\mathbf{u}^{n})+E(\phi^{n-1},\mathbf{u}^{n-1})+\lambda\|\nabla(\phi^{n+1}-\phi^{n})\|_0^2-\lambda\|\nabla(\phi^{n}-\phi^{n-1})\|_0^2&\\
+\lambda\theta_\eta\left[3(\eta^{n+1})^2-4(\eta^{n})^2+(\eta^{n-1})^2\right]+\theta_q\left[3(q^{n+1})^2-4(q^{n})^2+(q^{n-1})^2\right]\leq 0,&
\end{aligned}$$

which concludes the proof of Theorem 4.26. □

Remark 4.27. It follows from Theorem 4.26 that

$$\begin{aligned}\frac{3}{2}E(\phi^n,\mathbf{u}^n)-\frac{1}{2}E(\phi^{n-1},\mathbf{u}^{n-1})+\frac{\lambda}{2}\|\nabla(\phi^n-\phi^{n-1})\|_0^2\\+\lambda\theta_\eta\left[\frac{3}{2}(\eta^n)^2-\frac{1}{2}(\eta^{n-1})^2\right]+\theta_q\left[\frac{3}{2}(q^n)^2-\frac{1}{2}(q^{n-1})^2\right]\le C^*,\end{aligned}\tag{4.201}$$

for $1\le n\le N$, where C^* is given by

$$C^* := \frac{3}{2}E(\phi^1,\mathbf{u}^1)-\frac{1}{2}E(\phi^0,\mathbf{u}^0)+\frac{\lambda}{2}\|\nabla(\phi^1-\phi^0)\|_0^2+\lambda\theta_\eta\left[\frac{3}{2}(\eta^1)^2-\frac{1}{2}\right]+\theta_q\left[\frac{3}{2}(q^1)^2-\frac{1}{2}\right].\tag{4.202}$$

For $0\le n\le N$, we define

$$z^n := E(\phi^n,\mathbf{u}^n)+\lambda\theta_\eta(\eta^n)^2+\theta_q(q^n)^2.\tag{4.203}$$

We obtain from (4.201) that

$$z^n\le\frac{1}{3}z^{n-1}+\frac{2}{3}C^*,\quad 1\le n\le N,$$

which implies that

$$z^n\le C^*+\frac{1}{3^n}(z^0-C^*),\quad 0\le n\le N.\tag{4.204}$$

Thus, $\{z^n\}_{n=0}^N$ and, consequently, $\{E(\phi^n,\mathbf{u}^n)\}_{n=0}^N$ are uniformly bounded.

Remark 4.28. For the CH-NS system with an extra forcing term $\mathbf{f}$ (cf. (4.187) in Remark 4.25), the energy stability result (4.204) of the second-order DRLM scheme becomes

$$z^n\le\widehat{C}+\frac{1}{3^n}(z^0-\widehat{C}),\quad 0\le n\le N,\tag{4.205}$$

provided that $\Delta t\le\frac{4}{5}$ and $\mathbf{f}\in L^\infty(0,T;\mathbf{L}^2(\Omega))$, where

$$\widehat{C} := \frac{15}{14}e^{2T}\left[(3-2\Delta t)z^1+\lambda\|\nabla(\phi^1-\phi^0)\|_0^2+\frac{1}{2}\|\mathbf{f}\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))}\right].$$

Moreover, the second-order DRLM scheme is unconditionally energy stable with no-flux/no-slip boundary conditions if $\mathbf{f} \in L^\infty(0, T; \mathbf{H}^{-1}(\Omega))$:

$$z^n \le \widetilde{C} + \frac{1}{3^n}(z^0 - \widetilde{C}), \quad 0 \le n \le N, \tag{4.206}$$

where $\widetilde{C} := C^* + \frac{C_\Omega^2 T}{4\nu}\|\mathbf{f}\|^2_{L^\infty(0,T;\mathbf{H}^{-1}(\Omega))}$, C^* is given in (4.202), and C_Ω denotes the Poincaré constant defined in (4.193).

The proof of (4.206) is done in the same manner as in Remarks 4.25 and 4.27 and thus omitted. To derive (4.205), we follow the proof of Theorem 4.26 and obtain

$$\begin{aligned}
&3E(\phi^{n+1}, \mathbf{u}^{n+1}) - 4E(\phi^n, \mathbf{u}^n) + E(\phi^{n-1}, \mathbf{u}^{n-1}) + \lambda\|\nabla(\phi^{n+1} - \phi^n)\|_0^2 - \lambda\|\nabla(\phi^n - \phi^{n-1})\|_0^2 \\
&\qquad + \lambda\theta_\eta\left[3(\eta^{n+1})^2 - 4(\eta^n)^2 + (\eta^{n-1})^2\right] + \theta_q\left[3(q^{n+1})^2 - 4(q^n)^2 + (q^{n-1})^2\right] \\
&\le -2\Delta t\nu\|\nabla\mathbf{u}^{n+1}\|_0^2 + 2\Delta t\int_\Omega \mathbf{f}(t_{n+1})\cdot\mathbf{u}^{n+1}d\mathbf{x}.
\end{aligned} \tag{4.207}$$

Recalling the definition of $\{z^n\}$ in (4.203) and Young's inequality (4.191), we deduce from (4.207) that

$$(3 - 2\Delta t)z^{n+1} - 4z^n + z^{n-1} + \lambda\|\nabla(\phi^{n+1} - \phi^n)\|_0^2 - \lambda\|\nabla(\phi^n - \phi^{n-1})\|_0^2 \le \Delta t\|\mathbf{f}(t_{n+1})\|_0^2,$$

which implies

$$\begin{aligned}
&(3 - 2\Delta t)z^{n+1} - (2 - \sqrt{1 + 2\Delta t})z^n + \lambda\|\nabla(\phi^{n+1} - \phi^n)\|_0^2 \\
&\le \frac{2 + \sqrt{1 + 2\Delta t}}{3 - 2\Delta t}\left[(3 - 2\Delta t)z^n - (2 - \sqrt{1 + 2\Delta t})z^{n-1} + \lambda\|\nabla(\phi^n - \phi^{n-1})\|_0^2\right] + \Delta t\|\mathbf{f}(t_{n+1})\|_0^2 \\
&\le (1 + 2\Delta t)\left[(3 - 2\Delta t)z^n - (2 - \sqrt{1 + 2\Delta t})z^{n-1} + \lambda\|\nabla(\phi^n - \phi^{n-1})\|_0^2\right] + \Delta t\|\mathbf{f}(t_{n+1})\|_0^2.
\end{aligned} \tag{4.208}$$

Note that in the last inequality, we have used the fact that $\Delta t \le \frac{4}{5}$. Repeatedly applying the estimate (4.208) yields

$$
\begin{aligned}
&(3-2\Delta t)z^{n+1}-(2-\sqrt{1+2\Delta t})z^n+\lambda\|\nabla(\phi^{n+1}-\phi^n)\|_0^2\\
&\le (1+2\Delta t)^n\Bigg[(3-2\Delta t)z^1-(2-\sqrt{1+2\Delta t})z^0+\lambda\|\nabla(\phi^1-\phi^0)\|_0^2\\
&\qquad\qquad +\Delta t\sum_{i=1}^{n}(1+2\Delta t)^{-i}\|\mathbf{f}(t_{i+1})\|_0^2\Bigg]\\
&\le e^{2T}\left[(3-2\Delta t)z^1+\lambda\|\nabla(\phi^1-\phi^0)\|_0^2+\frac{1}{2}\|\mathbf{f}\|^2_{L^\infty(0,T;\mathbf{L}^2(\Omega))}\right]=:\widehat{C}_0.
\end{aligned}
\tag{4.209}
$$

By dropping the non-negative term $\lambda\|\nabla(\phi^{n+1}-\phi^n)\|_0^2$ on the left-hand side of (4.209), we obtain

$$
\begin{aligned}
z^{n+1}&\le\frac{2-\sqrt{1+2\Delta t}}{3-2\Delta t}z^n+\frac{\widehat{C}_0}{3-2\Delta t}\\
&=\frac{1}{2+\sqrt{1+2\Delta t}}z^n+\frac{\widehat{C}_0}{3-2\Delta t}\le\frac{1}{3}z^n+\frac{\widehat{C}_0}{3-2\Delta t},\quad n\ge 0.
\end{aligned}
\tag{4.210}
$$

Since $\Delta t\le\frac{4}{5}$, we have $3-2\Delta t\ge\frac{7}{5}$. Hence, we deduce from (4.210) that

$$
z^{n+1}\le\frac{1}{3}z^n+\frac{5}{7}\widehat{C}_0,\quad n\ge 0.
$$

With $\widehat{C}=\frac{15}{14}\widehat{C}_0$, the estimate (4.205) is thus proved.

Remark 4.29. While the convergence analysis of the first-order DRLM scheme for the incompressible NS equations is established in Section 4.2.3, the corresponding analysis of the proposed DRLM schemes (cf. (4.161) and (4.169)) for the coupled CH-NS system remains an open question and is left for future research. The main challenges arise from (i) the explicit treatment of nonlinear terms $F'(\phi)$ and $\mathbf{u}\cdot\nabla\mathbf{u}$, (ii) the complicated coupling between primary variables ($\phi,\mu,\mathbf{u}$, p) and scalar Lagrange multipliers (η, q), and (iii) the presence of quadratic terms involving the Lagrange multipliers, such as $(\eta^{n+1})^2$ and $(q^{n+1})^2$. We note that for numerical schemes where the nonlinear terms are treated fully implicitly or semi-implicitly, error analysis can be carried out as in [44, 22]. For SAV-based schemes with explicit treatment of nonlinear terms [206, 93, 118], although energy stability is obtained with respect to a modified energy, certain error estimates can be derived due to the linear structure of auxiliary variables. We refer to [198] for the analysis of a fully discrete SAV scheme that incorporates both a semi-implicit treatment of the nonlinear convection term $\mathbf{u} \cdot \nabla \mathbf{u}$ and an explicit treatment of $F'(\phi)$ via a suitable auxiliary variable.

4.3.4 Numerical results

We carry out various numerical experiments to illustrate the performance of the two proposed DRLM schemes (4.161) and (4.169). In the first test case, a manufactured solution is considered to verify the temporal convergence of both first- and second-order DRLM schemes and to compare their performance with the SAV-based schemes. We also investigate the role of the regularization parameters in DRLM schemes to ensure the accuracy of the Lagrange multipliers while allowing for large time step sizes (cf. Tables 4.6 and 4.7). The mass conservation and energy stability of the DRLM schemes are confirmed in the second test case. We then apply the second-order DRLM scheme for more realistic test cases, including shape relaxation, spinodal decomposition, and dripping droplet in two and three dimensions, to demonstrate its robustness and ability to capture the interface dynamics of two-phase systems.

For spatial discretization, different methods such as finite difference, finite element, or finite volume methods, can be applied. For simplicity, we consider the Marker-And-Cell (MAC) discretization on a staggered grid, where scalar variables such as the phase variable, chemical potential, and pressure are located at the cell centers, while the velocity components are positioned at the cell faces. We refer to [123, 48] for the details on MAC discretization. In the following, we consider no-flux boundary conditions for the phase variable and chemical potential, and no-slip conditions for the velocity (cf. (4.153)); similar results can be obtained with periodic boundary conditions. The spatial mesh sizes are uniform, $h = h_x = h_y$ in 2D and $h = h_x = h_y = h_z$ in 3D.

Note that the two DRLM schemes involve solving the following type of fourth-order linear equation (cf. (4.163)-(4.164) and (4.170)-(4.171)):

$$\left(\alpha I + \lambda M \Delta^2\right) \phi = f, \tag{4.211}$$

and the generalized Stokes system (cf. (4.166)-(4.167) and (4.173)-(4.174))

$$\alpha \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{g}, \quad \nabla \cdot \mathbf{u} = 0, \tag{4.212}$$

where f and $\mathbf{g}$ are given functions, $\alpha = \frac{1}{\Delta t}$ for the first-order DRLM scheme (4.161) and $\alpha = \frac{3}{2\Delta t}$ for the second-order counterpart (4.169). With the MAC spatial discretization, (4.211) can be solved via the Discrete Fourier Transform (for periodic boundary conditions) and the Discrete Cosine Transform (for homogeneous Neumann boundary conditions (4.153)); we refer to [104] and [103] for an overview of Fast Fourier Transform-based algorithms for the Allen-Cahn and Cahn-Hilliard equations. On the other hand, the Stokes system (4.212) can be transformed into a saddle point problem, where various effective preconditioners [8, 105, 48] are available to enhance the performance of iterative solvers such as GMRES.

For comparison purposes, we consider the multiple scalar auxiliary variables (MSAV) approach introduced in [124], which is briefly summarized as follows. Let $\kappa > 0$ be a stabilization constant, and define

$$G(\phi) := F(\phi) - \frac{\kappa}{2}\phi^2, \quad \text{and} \quad E_1(\phi) = \int_\Omega G(\phi) d\mathbf{x}.$$

The two extra space-independent variables are defined by

$$r(t) = \sqrt{E_1(\phi) + \delta}, \quad s(t) = e^{-t/T}, \tag{4.213}$$

for some constant $\delta > 0$ to ensure that $E_1(\phi) + \delta > 0$. In light of (4.213), the original CH-NS system (4.151) is reformulated into the following equivalent form

$$\frac{\partial \phi}{\partial t} + \frac{r}{\sqrt{E_1(\phi)+\delta}} \mathbf{u} \cdot \nabla \phi = M\Delta\mu, \qquad \text{in } \Omega \times (0, T], \tag{4.214a}$$

$$\mu = \lambda\left(-\Delta\phi + \kappa\phi + \frac{r}{\sqrt{E_1(\phi)+\delta}} G'(\phi)\right), \qquad \text{in } \Omega \times (0, T], \tag{4.214b}$$

$$\frac{\partial \mathbf{u}}{\partial t} + e^{t/T} s(t)\, \mathbf{u} \cdot \nabla \mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = \frac{r}{\sqrt{E_1(\phi)+\delta}} \mu\nabla\phi, \qquad \text{in } \Omega \times (0, T], \tag{4.214c}$$

$$\nabla \cdot \mathbf{u} = 0, \qquad \text{in } \Omega \times (0, T], \tag{4.214d}$$

$$\frac{dr}{dt} = \frac{1}{2\sqrt{E_1(\phi)+\delta}}\bigg[\int_\Omega G'(\phi)\frac{\partial\phi}{\partial t}\, d\mathbf{x} + \frac{1}{\lambda}\int_\Omega \big((\mathbf{u}\cdot\nabla\phi)\mu - (\mu\nabla\phi)\cdot\mathbf{u}\big)d\mathbf{x}\bigg], \qquad \text{in } (0, T], \tag{4.214e}$$

$$\frac{ds}{dt} = -\frac{1}{T}s + e^{t/T}\int_\Omega (\mathbf{u}\cdot\nabla\mathbf{u})\cdot\mathbf{u}\, d\mathbf{x}, \qquad \text{in } (0, T], \tag{4.214f}$$

with the initial conditions for the scalar variables: $r(0) = \sqrt{E_1(\phi|_{t=0}) + \delta}$ and $s(0) = 1$. From (4.214), the first- and second-order MSAV schemes can be derived based on the associated BDF methods, as detailed in [124].

Convergence test By adding suitable forcing terms to the right-hand sides of (4.151a) and (4.151c), we suppose that the exact solution of the CH-NS equations (4.151) is given by:

$$\begin{aligned} \phi(x, y, t) &= \cos(\pi x)\cos(\pi y)\sin(t), \\ p(x, y, t) &= \cos(\pi x)\sin(\pi y)\sin(t), \\ \mathbf{u}(x, y, t) &= \left[\pi\sin^2(\pi x)\sin(2\pi y)\sin(t), -\pi\sin(2\pi x)\sin^2(\pi y)\sin(t)\right]^T. \end{aligned} \tag{4.215}$$

We take the domain $\Omega = (0, 1)^2$, $\varepsilon = 0.1$, $M = 1$, $\lambda = 0.001$, $\nu = 0.1$, $\theta_\eta = 100/\varepsilon^2$, $\theta_q = 100$ and vary the time step size and spatial mesh size such that $h = \frac{\sqrt{\Delta t}}{4}$ for the first-order DRLM scheme and $h = \frac{\Delta t}{2}$ for the second-order one, where $\Delta t \in \left\{\frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}, \frac{1}{256}\right\}$. The L^2 and L^∞ errors of all variables, along with the absolute errors of the Lagrange multipliers at the final time $T = 1$, are reported in Table 4.2 for the first-order DRLM scheme and in Table 4.3 for the second-order counterpart. The results confirm the expected order of convergence for all quantities. We note that, in contrast to other approaches that employ projection methods [77, 124, 200], which often result in a lower convergence order for pressure due to artificial boundary conditions, our DRLM schemes maintain consistent convergence for pressure and other variables.

Table 4.2: Errors and convergence rates of the CH-NS variables and Lagrange multipliers by the first-order DRLM scheme (4.161) with $h = \frac{\sqrt{\Delta t}}{4}$.

Δt	$\|e_\phi\|_{L^2}$	$\|e_\phi\|_{L^\infty}$	$\|e_\mu\|_{L^2}$	$\|e_\mu\|_{L^\infty}$	$\vert\eta^N - 1\vert$
1/16	1.45e-01	2.30e-01	7.05e-03	1.71e-02	3.56e-05
1/32	5.27e-02 [1.46]	8.31e-02 [1.47]	2.31e-03 [1.61]	5.54e-03 [1.63]	1.16e-05 [1.62]
1/64	2.25e-02 [1.23]	3.84e-02 [1.11]	9.17e-04 [1.33]	2.02e-03 [1.46]	4.95e-06 [1.23]
1/128	1.06e-02 [1.09]	1.87e-02 [1.04]	4.12e-04 [1.15]	8.62e-04 [1.23]	2.37e-06 [1.06]
1/256	5.21e-03 [1.02]	9.21e-03 [1.02]	1.96e-04 [1.07]	4.08e-04 [1.08]	1.17e-06 [1.02]
Δt	$\|e_{\mathbf{u}}\|_{L^2}$	$\|e_{\mathbf{u}}\|_{L^\infty}$	$\|e_p\|_{L^2}$	$\|e_p\|_{L^\infty}$	$\vert q^N - 1\vert$
1/16	8.95e-03	1.51e-02	2.95e-01	9.50e-01	4.28e-04
1/32	4.41e-03 [1.02]	7.53e-03 [1.00]	1.47e-01 [1.00]	4.89e-01 [0.96]	2.12e-04 [1.01]
1/64	2.38e-03 [0.89]	4.08e-03 [0.88]	7.41e-02 [0.99]	2.44e-01 [1.00]	1.06e-04 [1.00]
1/128	1.22e-03 [0.96]	2.10e-03 [0.96]	3.71e-02 [1.00]	1.23e-01 [0.99]	5.27e-05 [1.01]
1/256	6.00e-04 [1.02]	1.03e-03 [1.03]	1.85e-02 [1.00]	6.13e-02 [1.00]	2.63e-05 [1.00]

Table 4.3: Errors and convergence rates of the CH-NS variables and Lagrange multipliers by the second-order DRLM scheme (4.169) with $h = \frac{\Delta t}{2}$.

Δt	$\|e_\phi\|_{L^2}$	$\|e_\phi\|_{L^\infty}$	$\|e_\mu\|_{L^2}$	$\|e_\mu\|_{L^\infty}$	$\vert\eta^N - 1\vert$
1/16	3.93e-03	9.23e-03	3.84e-04	7.79e-04	4.27e-06
1/32	9.54e-04 [2.04]	2.09e-03 [2.14]	6.47e-05 [2.57]	1.36e-04 [2.52]	1.14e-06 [1.91]
1/64	2.67e-04 [1.84]	5.98e-04 [1.81]	1.69e-05 [1.94]	3.81e-05 [1.84]	2.95e-07 [1.95]
1/128	6.98e-05 [1.94]	1.58e-04 [1.92]	4.31e-06 [1.97]	9.95e-06 [1.94]	7.53e-08 [1.97]
1/256	1.78e-05 [1.97]	4.06e-05 [1.96]	1.09e-06 [1.98]	2.54e-06 [1.97]	1.90e-08 [1.99]
Δt	$\|e_{\mathbf{u}}\|_{L^2}$	$\|e_{\mathbf{u}}\|_{L^\infty}$	$\|e_p\|_{L^2}$	$\|e_p\|_{L^\infty}$	$\vert q^N - 1\vert$
1/16	4.05e-03	6.35e-03	4.13e-03	9.07e-03	3.11e-05
1/32	1.02e-03 [1.99]	1.59e-03 [2.00]	1.19e-03 [1.80]	2.56e-03 [1.82]	7.46e-06 [2.06]
1/64	2.55e-04 [2.00]	3.99e-04 [1.99]	3.14e-04 [1.92]	6.70e-04 [1.93]	1.83e-06 [2.03]
1/128	6.38e-05 [2.00]	9.99e-05 [2.00]	8.04e-05 [1.97]	1.71e-04 [1.97]	4.52e-07 [2.02]
1/256	1.59e-05 [2.00]	2.50e-05 [2.00]	2.03e-05 [1.99]	4.32e-05 [1.98]	1.12e-07 [2.01]

Next, we compare DRLM schemes with the MSAV schemes [124] derived from the reformulated system (4.214); for consistency, the velocity and pressure are coupled via a Stokes system rather than the original pressure-correction approach [124]. The parameters $\varepsilon, M, \lambda, \nu, T$, as well as the mesh size h and time step size Δt, are chosen as before, while the stabilization constant $\varkappa$ and the positive constant δ are given by $\varkappa = 5/\varepsilon^2$ and $\delta = 35/(4\varepsilon^2)$ as in [124]. Errors by the first- and second-order MSAV schemes are reported in Tables 4.4 and 4.5, respectively; expected order of convergence is confirmed for all quantities. We observe that DRLM schemes provide significantly improved accuracy for the CH-related variables compared to MSAV counterparts, as DRLM schemes require no stabilization. Both DRLM and MSAV schemes provide comparable results for the NS-related variables.

Table 4.4: Errors and convergence rates of the CH-NS variables and scalar auxiliary variables by the first-order MSAV scheme with $h = \frac{\sqrt{\Delta t}}{4}$.

Δt	$\|e_\phi\|_{L^2}$	$\|e_\phi\|_{L^\infty}$	$\|e_\mu\|_{L^2}$	$\|e_\mu\|_{L^\infty}$	$\|e_r\|$
1/16	2.40e-01	4.59e-01	2.91e-02	4.65e-02	5.94e-02
1/32	1.51e-01 [0.67]	3.14e-01 [0.55]	1.50e-02 [0.96]	2.76e-02 [0.75]	1.64e-01 [-1.47]
1/64	7.77e-02 [0.96]	1.69e-01 [0.89]	6.08e-03 [1.30]	1.28e-02 [1.11]	1.31e-01 [0.32]
1/128	3.57e-02 [1.12]	8.16e-02 [1.05]	2.29e-03 [1.41]	5.13e-03 [1.32]	6.64e-02 [0.98]
1/256	1.66e-02 [1.10]	3.91e-02 [1.06]	9.74e-04 [1.23]	2.16e-03 [1.25]	3.22e-02 [1.04]
Δt	$\|e_{\mathbf{u}}\|_{L^2}$	$\|e_{\mathbf{u}}\|_{L^\infty}$	$\|e_p\|_{L^2}$	$\|e_p\|_{L^\infty}$	$\|e_s\|$
1/16	1.02e-02	1.70e-02	2.17e-01	7.20e-01	1.00e-02
1/32	4.17e-03 [1.29]	6.90e-03 [1.30]	1.04e-01 [1.06]	3.58e-01 [1.01]	5.42e-03 [0.88]
1/64	2.05e-03 [1.02]	3.25e-03 [1.09]	5.29e-02 [0.98]	1.80e-01 [0.99]	2.80e-03 [0.95]
1/128	1.07e-03 [0.94]	1.68e-03 [0.95]	2.66e-02 [0.99]	9.08e-02 [0.99]	1.42e-03 [0.98]
1/256	5.45e-04 [0.97]	8.51e-04 [0.98]	1.32e-02 [1.01]	4.49e-02 [1.02]	7.14e-04 [0.99]

Table 4.5: Errors and convergence rates of the CH-NS variables and scalar auxiliary variables by the second-order MSAV scheme with $h = \frac{\Delta t}{2}$.

Δt	$\|e_\phi\|_{L^2}$	$\|e_\phi\|_{L^\infty}$	$\|e_\mu\|_{L^2}$	$\|e_\mu\|_{L^\infty}$	$\|e_r\|$
1/16	7.54e-02	1.35e-01	7.93e-03	1.69e-02	2.14e-01
1/32	1.05e-02 [2.84]	2.16e-02 [2.64]	7.16e-04 [3.47]	1.43e-03 [3.56]	2.59e-02 [3.05]
1/64	2.25e-03 [2.22]	4.73e-03 [2.19]	1.20e-04 [2.58]	2.62e-04 [2.45]	6.91e-03 [1.91]
1/128	5.78e-04 [1.96]	1.19e-03 [1.99]	2.99e-05 [2.00]	6.76e-05 [1.95]	1.78e-03 [1.96]
1/256	1.47e-04 [1.98]	3.01e-04 [1.98]	7.53e-06 [1.99]	1.72e-05 [1.97]	4.50e-04 [1.98]
Δt	$\|e_{\mathbf{u}}\|_{L^2}$	$\|e_{\mathbf{u}}\|_{L^\infty}$	$\|e_p\|_{L^2}$	$\|e_p\|_{L^\infty}$	$\|e_s\|$
1/16	2.92e-03	4.39e-03	4.92e-03	9.45e-03	6.91e-04
1/32	9.32e-04 [1.65]	1.41e-03 [1.64]	1.67e-03 [1.56]	3.86e-03 [1.29]	1.56e-04 [2.15]
1/64	2.43e-04 [1.94]	3.72e-04 [1.92]	4.96e-04 [1.75]	1.18e-03 [1.71]	3.80e-05 [2.04]
1/128	6.11e-05 [1.99]	9.35e-05 [1.99]	1.27e-04 [1.97]	3.05e-04 [1.95]	9.41e-06 [2.01]
1/256	1.53e-05 [2.00]	2.34e-05 [2.00]	3.20e-05 [1.99]	7.69e-05 [1.99]	2.35e-06 [2.00]

To investigate the effect of the regularization parameters on the proposed DRLM schemes, we vary θ_η and θ_q and compute absolute errors of the Lagrange multipliers at the final time $T = 1$. These errors are presented in Tables 4.6 and 4.7 for the first- and second-order DRLM schemes, respectively. We observe that as θ_η and θ_q increase, the errors in both Lagrange multipliers become much smaller, resulting in better accuracy of the numerical solutions, especially with large time step sizes. Hence, we will choose sufficiently large values for θ_η and θ_q in the subsequent numerical experiments.

Table 4.6: Errors of the Lagrange multipliers by the first-order DRLM scheme (4.161) with $h = \frac{\sqrt{\Delta t}}{4}$ and different regularization parameters.

Δt	$\theta_\eta = 0.1/\varepsilon^2,\ \theta_q = 0.1$		$\theta_\eta = 1/\varepsilon^2,\ \theta_q = 1$		$\theta_\eta = 10/\varepsilon^2,\ \theta_q = 10$	
	$\lvert\eta^N - 1\rvert$	$\lvert q^N - 1\rvert$	$\lvert\eta^N - 1\rvert$	$\lvert q^N - 1\rvert$	$\lvert\eta^N - 1\rvert$	$\lvert q^N - 1\rvert$
1/16	2.77e-02	3.65e-01	3.46e-03	4.20e-02	3.55e-04	4.27e-03
1/32	8.54e-03	1.94e-01	1.13e-03	2.10e-02	1.16e-04	2.12e-03
1/64	3.63e-03	1.01e-01	4.80e-04	1.05e-02	4.93e-05	1.05e-03
1/128	1.75e-03	5.14e-02	2.30e-04	5.25e-03	2.37e-05	5.27e-04
1/256	8.71e-04	2.60e-02	1.14e-04	2.63e-03	1.17e-05	2.63e-04

Table 4.7: Errors of the Lagrange multipliers by the second-order DRLM scheme (4.169) with $h = \frac{\Delta t}{2}$ and different regularization parameters.

Δt	$\theta_\eta = 0.1/\varepsilon^2,\ \theta_q = 0.1$		$\theta_\eta = 1/\varepsilon^2,\ \theta_q = 1$		$\theta_\eta = 10/\varepsilon^2,\ \theta_q = 10$	
	$\lvert\eta^N - 1\rvert$	$\lvert q^N - 1\rvert$	$\lvert\eta^N - 1\rvert$	$\lvert q^N - 1\rvert$	$\lvert\eta^N - 1\rvert$	$\lvert q^N - 1\rvert$
1/16	3.22e-03	3.09e-02	4.15e-04	3.11e-03	4.26e-05	3.11e-04
1/32	8.41e-04	7.45e-03	1.10e-04	7.46e-04	1.13e-05	7.46e-05
1/64	2.18e-04	1.83e-03	2.86e-05	1.83e-04	2.94e-06	1.83e-05
1/128	5.54e-05	4.52e-04	7.30e-06	4.52e-05	7.51e-07	4.52e-06
1/256	1.40e-05	1.12e-04	1.84e-06	1.12e-05	1.90e-07	1.12e-06

Mass conservation and energy dissipation test Next, we test the properties of mass conservation and energy dissipation by the proposed DRLM schemes. Let us consider the CH-NS system

(4.151) in the domain $\Omega = (0,1)^2$ with the following initial conditions [77]:

$$\begin{aligned} \phi_0(x,y) &= 0.24\cos(2\pi x)\cos(2\pi y) + 0.4\cos(\pi x)\cos(3\pi y), \\ \mathbf{u}_0(x,y) &= \left[-\sin^2(\pi x)\sin(2\pi y), \sin(2\pi x)\sin^2(\pi y)\right]^T. \end{aligned} \tag{4.216}$$

The parameters are set as follows:

$$\varepsilon = 0.1, \quad M = 1, \quad \lambda = 0.01, \quad \nu = 0.01, \quad \theta_\eta = 1000/\varepsilon^2, \quad \theta_q = 100.$$

The evolution of mass, original energy, and error functions $\mathcal{E}_1$ and $\mathcal{E}_2$ associated with the Lagrange multipliers (cf. Theorems 4.23 and 4.26) up to $T = 10$ is shown in Figure 4.13 for both DRLM schemes with different spatial mesh sizes and time step sizes, where $h = \Delta t \in \{4 \times 10^{-3}, 2 \times 10^{-3}\}$. We clearly observe the conservation of mass and dissipation of energy by the numerical solutions along the time. Additionally, the error functions $\mathcal{E}_1$ and $\mathcal{E}_2$ in the energy stability results (4.180) and (4.195) are negligible compared to the energy functional, with their magnitudes getting closer to zero as h and Δt are refined.

Shape relaxation We now simulate the evolution of a square-shaped fluid bubble in the domain $\Omega = (0,1)^2$. The phase variable is initialized with a value of 1 inside the square bubble $[0.25, 0.75]^2$ and -1 outside. The initial velocity is set to be zero and other parameters are given by

$$\varepsilon \in \{0.01, 0.02\}, \quad M = 0.002, \quad \lambda = 0.01, \quad \nu = 1, \quad \theta_\eta = 1000/\varepsilon^2, \quad \theta_q = 100.$$

We perform the simulation until $T = 1$ using the second-order DRLM scheme with $h = 4 \times 10^{-3}$, $\Delta t = 2 \times 10^{-3}$ for $\varepsilon = 0.02$, and $\Delta t = 2 \times 10^{-4}$ for $\varepsilon = 0.01$. The evolution of the phase variable at different times for $\varepsilon \in \{0.01, 0.02\}$ is shown in Figure 4.14, where the dashed circles, centered at $(0.5, 0.5)$ with radius $\frac{1}{2\sqrt{\pi}}$, indicate the interface at steady state. Under the effect of surface tension, the initial square shape eventually transitions into a circle that closely matches the dashed one, with the circle forming at an earlier stage for $\varepsilon = 0.01$ than for $\varepsilon = 0.02$. Our second-order DRLM scheme accurately captures this dynamic within a short time interval. Figure 4.15 confirms the energy decay property of the numerical solutions, with the total energy stabilizing

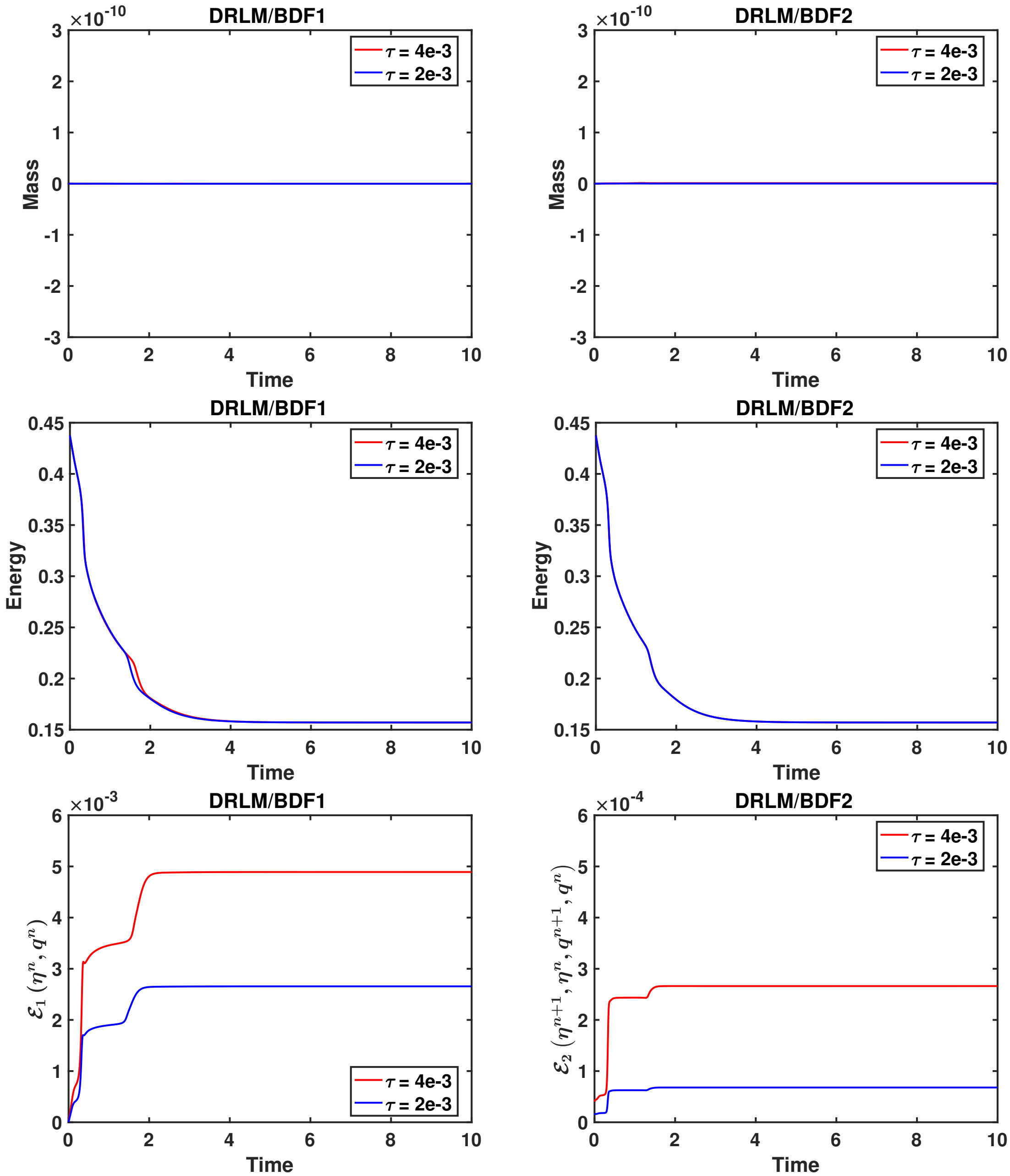


Figure 4.13: Evolution of mass, original energy, and error functions $\mathcal{E}_i$ $(i = 1, 2)$ produced by the DRLM schemes with $h = \Delta t \in \{4 \times 10^{-3}, 2 \times 10^{-3}\}$. Left column: the first-order scheme (4.161); Right column: the second-order scheme (4.169).

at approximately $t = 1$ for $\varepsilon = 0.02$ and near $t = 0.6$ for $\varepsilon = 0.01$. In contrast, existing stabilized schemes (e.g., [169, 123]) allow for relatively larger time step sizes but require significantly longer times for the shape relaxation process due to artificial dissipation introduced by stabilization constants.

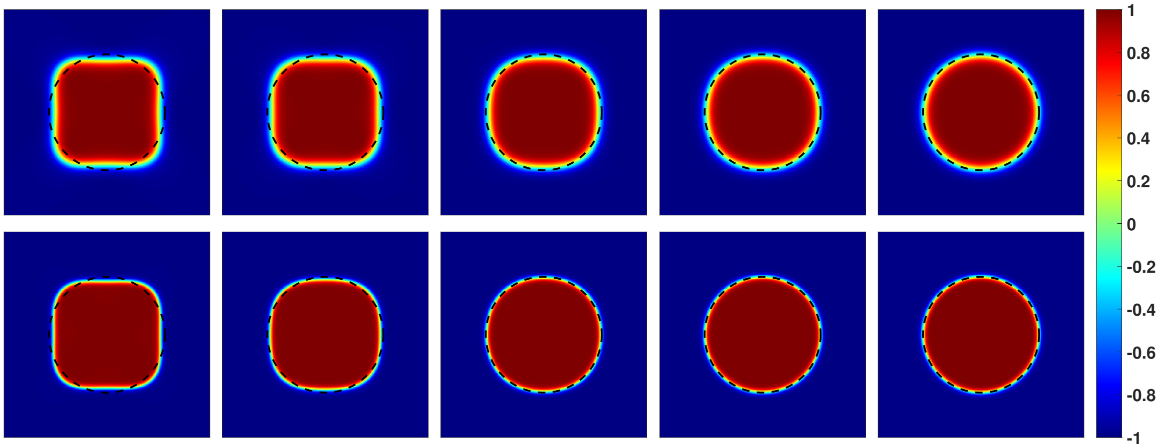


Figure 4.14: Shape relaxation of surface tension-driven flow at $t = 0.1, 0.2, 0.4, 0.6, 1$ (from left to right) by the second-order DRLM scheme (4.169) with $\varepsilon = 0.02$ (top) and $\varepsilon = 0.01$ (bottom).

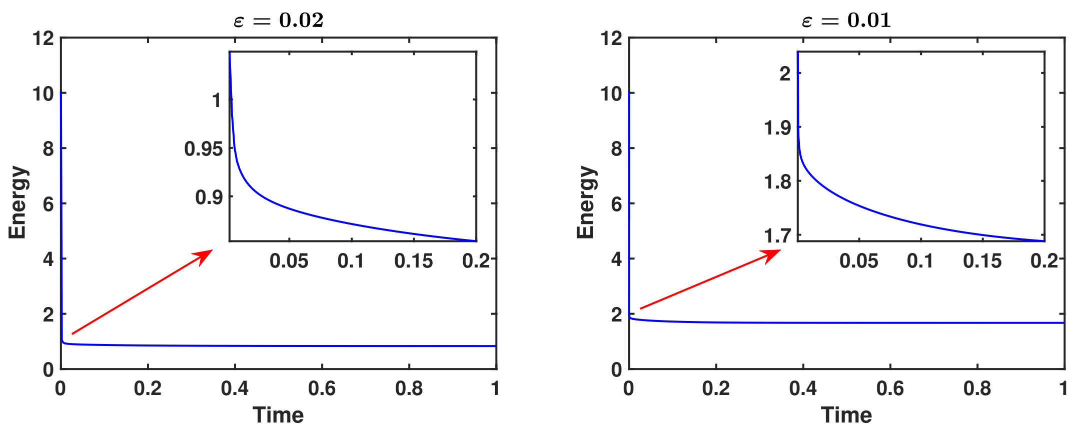


Figure 4.15: Evolution of the original energy during the shape relaxation by the second-order DRLM scheme (4.169) with $\varepsilon = 0.02$ (left) and $\varepsilon = 0.01$ (right).

Spinodal decomposition The CH-NS system (4.151) can serve as a model for spinodal decomposition in a binary fluid mixture [14, 107, 77, 200]. This decomposition begins with a small perturbation to the initial concentration field, set as $\phi_0(x, y) = \overline{\phi}_0 + 0.001\,\mathrm{rand}(x, y)$, where $\mathrm{rand}(x, y)$ denotes a random number in $[-1, 1]$ following the normal distribution, and the volume fraction $\overline{\phi}_0 \in \{0, 0.5\}$. The initial velocity is set to zero, i.e., $\mathbf{u}_0 = \mathbf{0}$. We take the domain $\Omega = (0, 1)^2$ with the following parameters:

$$\varepsilon = 0.01, \quad M = 0.001, \quad \lambda = 0.01, \quad \nu = 1, \quad \theta_\eta = 1000/\varepsilon^2, \quad \theta_q = 1000.$$

We perform the simulation until $T = 25$ for $\overline{\phi}_0 = 0$ and $T = 50$ for $\overline{\phi}_0 = 0.5$ using the second-order DRLM scheme with $h = 4 \times 10^{-3}$ and $\Delta t = 4 \times 10^{-4}$. Snapshots of the phase variable at different times are shown in Figure 4.16 for $\overline{\phi}_0 = 0$ and in Figure 4.17 for $\overline{\phi}_0 = 0.5$. Both simulations exhibit the phase separation process, with the later (i.e., $\overline{\phi}_0 = 0.5$) requiring a longer time to reach the equilibrium state. At equilibrium, a pipe-like phase structure is observed in the case of $\overline{\phi}_0 = 0$ while a circular phase structure is observed in the case of $\overline{\phi}_0 = 0.5$. Figure 4.18 displays the original energy evolution for each case, where the energy functional decreases significantly at the early state before gradually stabilizing as it approaches the final time.

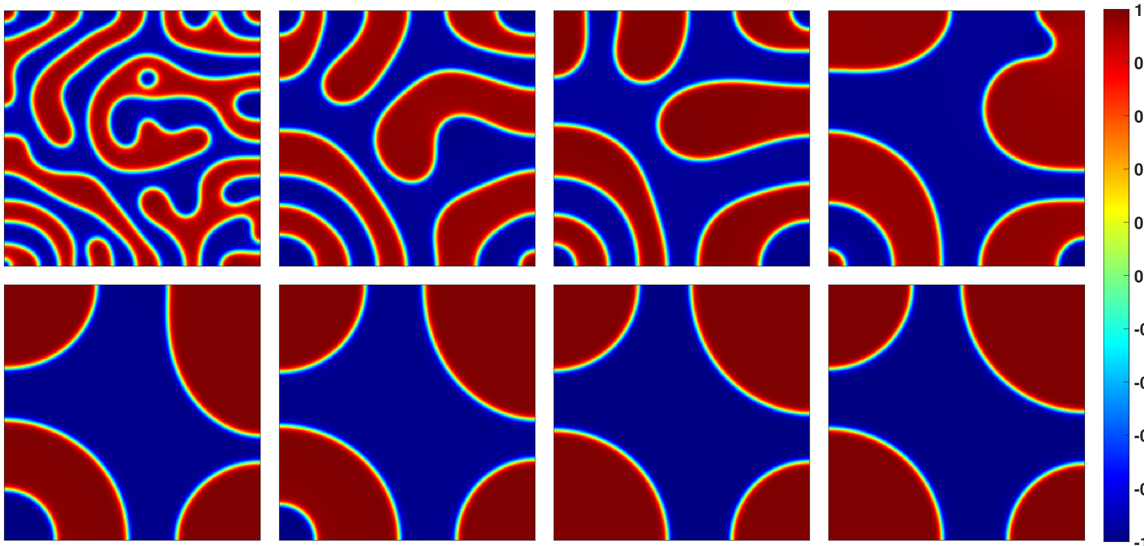


Figure 4.16: Snapshots of the phase variable during the spinodal decomposition at $t = 0.25, 1,$ $2, 6, 10, 15, 20,$ and 25 (from left to right and top to bottom) by the second-order DRLM scheme (4.169) with $\overline{\phi}_0 = 0$.

Dripping droplet We finally perform two-dimensional and three-dimensional simulations of the dripping dynamics of a droplet falling from the top wall under the gravitational force [200, 131]. By assuming the density difference of the liquid drop and the ambient fluid is small, we reformulate the momentum equation (4.151c) using Boussinesq approximations [71, 111, 168] as follows:

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mu \nabla \phi + \chi \rho(\phi) \mathbf{g}, \tag{4.217}$$

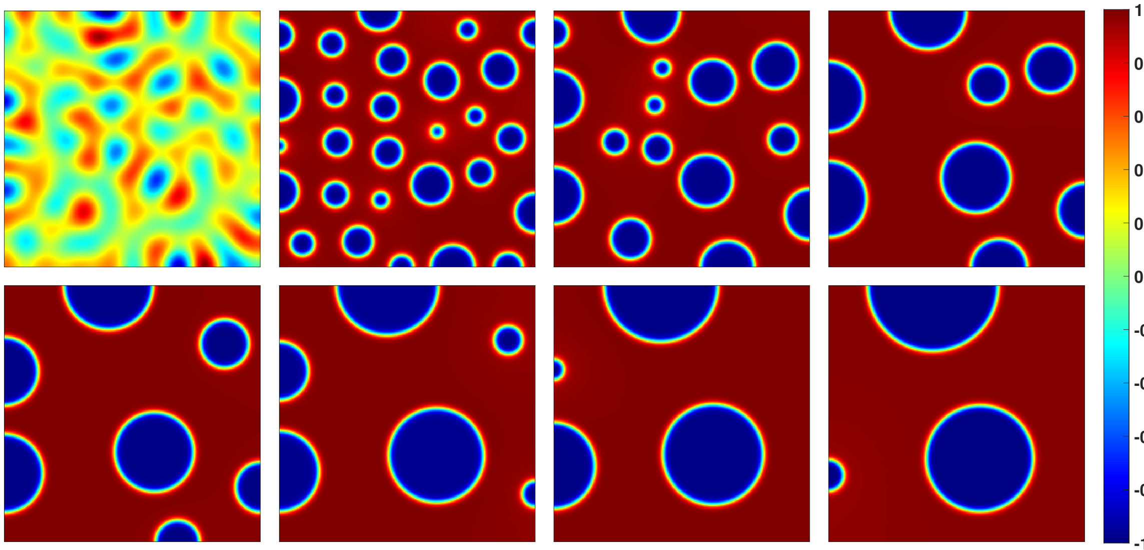


Figure 4.17: Snapshots of the phase variable during the spinodal decomposition at $t = 0.25, 1$, $2, 6, 10, 15, 20$, and 50 (from left to right and top to bottom) by the second-order DRLM scheme (4.169) with $\overline{\phi}_0 = 0.5$.

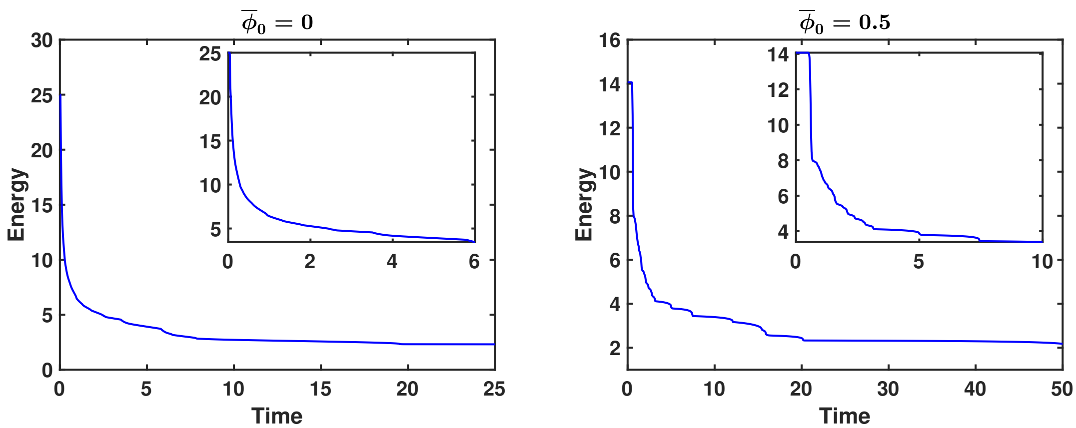


Figure 4.18: Evolution of the total energy during the spinodal decomposition by the second-order scheme (4.169) with $\overline{\phi}_0 = 0$ (left) and $\overline{\phi}_0 = 0.5$ (right).

where $\chi\rho(\phi)\mathbf{g}$ is a buoyancy term with $\chi\rho(\phi) = \chi(\phi - \overline{\phi})$, $\mathbf{g} = [0, -1]^T$ in 2D and $\mathbf{g} = [0, 0, -1]^T$ in 3D.

For the two-dimensional case, we set the computational domain $\Omega = (0, 1) \times (0, 2)$, the initial conditions:

$$\phi_0(x, y) = \tanh\left(\frac{0.32 - \sqrt{(x - 0.5)^2 + (y - 2.1)^2}}{\varepsilon}\right), \quad \mathbf{u}_0(x, y) = \mathbf{0}, \tag{4.218}$$

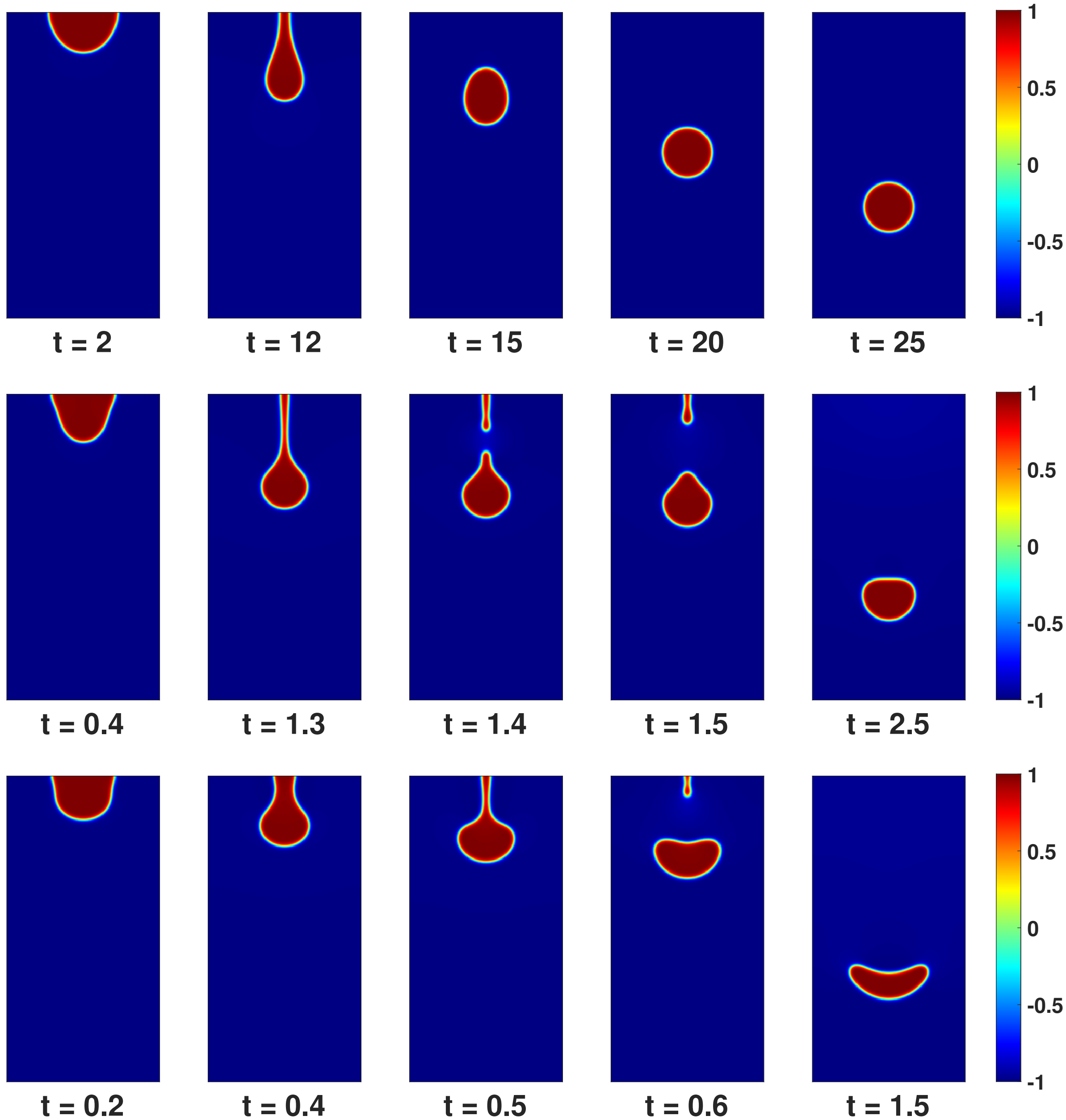


Figure 4.19: Evolution of the 2D droplet falling at different times with $Re = 1, 10$, and 50 (from top row to bottom row) by the second-order DRLM scheme (4.169).

and the parameters:

$$\varepsilon = 0.01, \quad M = 0.01, \quad \lambda = 0.001, \quad \theta_\eta = 1000/\varepsilon^2, \quad \theta_q = 1000, \quad \chi = 10, \quad \overline{\phi} = 0.$$

The spatial mesh size and time step size are given by $h = 4 \times 10^{-3}$ and $\Delta t = 4 \times 10^{-4}$, respectively. Snapshots of the phase evolution with different Reynolds numbers $Re \in \{1, 10, 50\}$, corresponding to $\nu \in \{1, 0.1, 0.02\}$, by the second-order DRLM scheme are presented in Figure 4.19. In all cases, the droplet first elongates under the effect of gravity, then pinches off and moves downward to the bottom of the domain. As the Reynolds number increases, the pinch-off behavior occurs earlier, and the droplet descends more quickly. Moreover, a circular shape for the droplet is observed at $Re = 1$, while a bean-like shape appears at $Re = 50$, as illustrated in the last subfigure of each row in Figure 4.19.

We further extend the simulation in three dimensions with the domain $\Omega = (0,1) \times (0,1) \times (0,2)$ and the initial conditions:

$$\phi_0(x,y,z) = \tanh\left(\frac{0.32 - \sqrt{(x-0.5)^2 + (y-0.5)^2 + (z-2.1)^2}}{\varepsilon}\right), \quad \mathbf{u}_0(x,y,z) = \mathbf{0}. \tag{4.219}$$

We fix $\varepsilon = 0.025$ and $Re = 50$ (i.e., $\nu = 0.02$), with other parameters chosen as in the two-dimensional case. Isosurface snapshots of the droplet (i.e., $\{\phi = 0\}$) at different times by the second-order DRLM scheme with $h = 1.25 \times 10^{-2}$ and $\Delta t = 2 \times 10^{-3}$ are shown in Figure 4.20. An elongated liquid filament gradually forms and pinches off at around $t = 0.7$, resulting in a bowl-shaped droplet. This behavior is consistent with the two-dimensional simulation, where a similar process produces a droplet with a comparable concave shape.

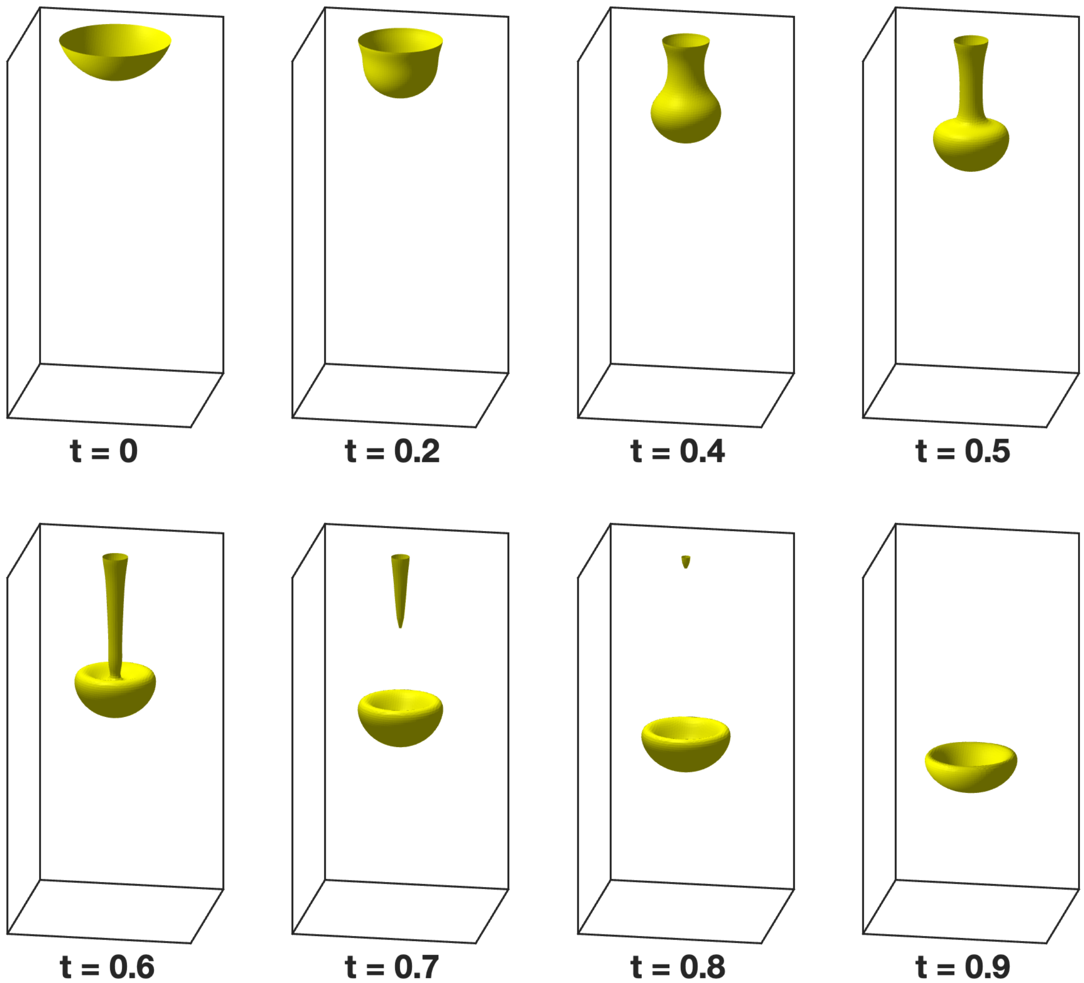


Figure 4.20: Evolution of the 3D droplet falling at different times with $Re = 50$ by the second-order DRLM scheme (4.169).

Chapter 5

Generalized Transferable Neural Networks for the Cahn-Hilliard Equation

This chapter is devoted to a structure-preserving neural network-based framework with a mass-conserving projection for the Cahn-Hilliard equation. We employ generalized transferable neural networks (GTransNet) for spatial approximation and stabilized backward differentiation formulas (BDF) for temporal discretization of the equation written in the mixed formulation. The resulting first- and second-order GTransNet-BDF schemes are shown to conserve mass and satisfy energy stability at the time-discrete level. The schemes are implemented by a collocation-based method, in which a least-squares system with constant coefficient matrix needs to be solved at each time step. Since the solution of this system, which determines the output-layer weights of the network, violates mass conservation due to the expected nonzero least-squares residual, we introduce a novel post-processing mass-conserving projection that enforces the mass constraint through a minimization problem, whose solution can be computed at negligible computational cost. A key advantage of the proposed method lies in its predetermined hidden layers and mesh-free nature, making it applicable to complex domains, variable mobility, and long-time simulations. Numerical experiments in two and three dimensions verify convergence, mass conservation, and energy dissipation as well as demonstrate the accuracy and robustness of the proposed methods. The material in this chapter is based on the work in [51].

5.1 Generalized transferable neural networks

5.1.1 Original TransNet

For simplicity of presentation, we describe TransNet on the unit ball $B_1(\mathbf{0})$; the general case follows by an affine shift (see Remark 5.2 below). The idea of transferable neural networks [201] is to fix the hidden layer parameters in advance, without relying on any PDE-specific information, and to optimize only the output weights, with the tanh activation function. The resulting approximate solution u_{NN} to a time-dependent PDE takes the following form:

$$u_{\text{NN}}(\mathbf{x}, t) = \sum_{m=1}^{N} \alpha_m(t) \tanh(\mathbf{w}_m^T \mathbf{x} + b_m) = \sum_{m=1}^{N} \alpha_m(t) \tanh(\gamma_m(\mathbf{a}_m^T \mathbf{x} + r_m)), \tag{5.1}$$

where N denotes the number of hidden neurons, $\mathbf{w}_m$ and b_m are the weight and bias of the m-th hidden neuron, respectively, and $\alpha_1, \alpha_2, \dots, \alpha_N$ are the time-dependent weights of the output layer. Note that the decomposition of $(\mathbf{w}_m, b_m)$ into the location parameter $(\mathbf{a}_m, r_m)$ with $\|\mathbf{a}_m\|_{l2} = 1$ and the shape parameter $\gamma_m > 0$ in (5.1) is based on the following relations:

$$\gamma_m = \|\mathbf{w}_m\|_{l2}, \quad \mathbf{a}_m = \frac{\mathbf{w}_m}{\gamma_m}, \quad r_m = \frac{b_m}{\gamma_m}, \quad 1 \le m \le N.$$

Geometrically, $\mathbf{a}_m$ represents the normal direction of the *partition hyperplane*, defined by $\mathbf{a}_m^T \mathbf{x} + r_m = 0$, while r_m denotes its distance from the origin. The shape parameter γ_m determines the steepness of the pre-activation value $\mathbf{w}_m^T \mathbf{x} + b_m$ along the normal direction $\mathbf{a}_m$. We refer to [201] for a geometric visualization of these parameters.

To measure the density of hidden neurons within a specific region, we recall the partition hyperplane density function $D_N^\tau(\mathbf{x})$, defined as

$$D_N^\tau(\mathbf{x}) = \frac{1}{N} \sum_{m=1}^{N} \chi_{\{d_m(\mathbf{x}) < \tau\}}(\mathbf{x}), \quad \text{where } d_m(\mathbf{x}) = |\mathbf{a}_m^T \mathbf{x} + r_m|.$$

Here $d_m(\mathbf{x})$ measures the distance from $\mathbf{x}$ to the m-th partition hyperplane, and χ denotes the standard indicator function. The following theorem (cf. [201, 133]) shows that uniform sampling of the location parameters distributes the partition hyperplanes evenly throughout the domain, ensuring that the basis functions provide balanced resolution everywhere.

Theorem 5.1. If $\{\mathbf{a}_m\}_{m=1}^N$ are i.i.d. and uniformly distributed on the d-dimensional unit sphere, and $\{r_m\}_{m=1}^N$ are i.i.d. and uniformly distributed in $[0, 1]$, then for any $\tau \in (0, 1)$,

$$\mathbb{E}\left[D_N^\tau(\mathbf{x})\right] = \tau,$$

for all $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x}\|_{l2} \le 1 - \tau$.

Once the location parameters $\{(\mathbf{a}_m, r_m)\}_{m=1}^N$ are determined as in Theorem 5.1, we use Gaussian random fields [201] to tune the shape parameters $\{\gamma_m\}_{m=1}^N$, which are assumed to be uniform (i.e., $\gamma = \gamma_1 = \dots = \gamma_N$). The *neural feature space* is then defined by

$$\begin{aligned}\mathcal{P}_{\text{NN}} &= \text{span}\left\{\tanh(\gamma(\mathbf{a}_1^T\mathbf{x} + r_1)), \dots, \tanh(\gamma(\mathbf{a}_N^T\mathbf{x} + r_N))\right\} \\ &=: \text{span}\left\{\psi_1(\mathbf{x}), \dots, \psi_N(\mathbf{x})\right\},\end{aligned}$$

where $\psi_m(\mathbf{x}) = \tanh(\gamma(\mathbf{a}_m^T\mathbf{x} + r_m))$ for $1 \le m \le N$.

Remark 5.2. For a general domain $\Omega \subset B_R(\mathbf{x}_c)$, the m-th hidden neuron is defined as

$$\psi_m(\mathbf{x}) = \tanh(\gamma(\mathbf{a}_m^T(\mathbf{x} - \mathbf{x}_c) + Rr_m)), \quad 1 \le m \le N,$$

where $\{\mathbf{a}_m\}_{m=1}^N$ and $\{r_m\}_{m=1}^N$ are sampled as in Theorem 5.1. Equivalently, the TransNet approximation can be written compactly as

$$\begin{cases}\boldsymbol{\psi}(\mathbf{x}) = \tanh(\gamma(\mathcal{A}(\mathbf{x} - \mathbf{x}_c) + R\mathbf{r})), \\ u_{\text{NN}}(\mathbf{x}, t) = \boldsymbol{\alpha}(t)^T\boldsymbol{\psi}(\mathbf{x}),\end{cases}$$

where $\mathcal{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N]^T \in \mathbb{R}^{N\times d}$, $\mathbf{r} = (r_1, r_2, \dots, r_N)^T \in \mathbb{R}^N$, $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots, \psi_N)^T \in \mathbb{R}^N$, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_N)^T \in \mathbb{R}^N$.

5.1.2 Generalized TransNet

Although the original TransNet performs well for problems with relatively smooth solutions, its accuracy may degrade when dealing with problems exhibiting sharp interfaces, steep gradients, or high-frequency features. As discussed in [24], when the shape parameter γ increases, the activation values of the single hidden layer tend to cluster in the saturation regions of the tanh function, thereby limiting the network's expressive capacity. To address this limitation, we adopt the generalized TransNet (i.e., GTransNet) framework [24] to augment the original TransNet with additional hidden layers while preserving the interpretable feature-generation mechanism.

Network architecture Let N_l denote the number of neurons in the l-th hidden layer. The GTransNet with $L \geq 2$ hidden layers takes the following form

$$\begin{cases} \boldsymbol{\psi}_1(\mathbf{x}) = \tanh(\gamma(\mathcal{A}(\mathbf{x} - \mathbf{x}_c) + R\mathbf{r})), \\ \boldsymbol{\psi}_l(\mathbf{x}) = \tanh(\mathbf{W}_l \boldsymbol{\psi}_{l-1}(\mathbf{x})), \quad l = 2, \dots, L, \\ u_{\mathsf{NN}}(\mathbf{x}, t) = \boldsymbol{\alpha}(t)^T \boldsymbol{\psi}_L(\mathbf{x}), \end{cases} \tag{5.2}$$

where $\mathcal{A}$ and $\mathbf{r}$ are defined as in Remark 5.2 with N replaced by N_1, $\boldsymbol{\psi}_l = (\psi_1^{(l)}, \dots, \psi_{N_l}^{(l)})^T \in \mathbb{R}^{N_l}$ for $1 \leq l \leq L$, $\mathbf{W}_l = (W_{ij}^{(l)}) \in \mathbb{R}^{N_l \times N_{l-1}}$ for $2 \leq l \leq L$, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{N_L})^T \in \mathbb{R}^{N_L}$ collects the time-dependent output-layer weights. Note that no bias terms are used in the second and subsequent hidden layers.

The first hidden layer of GTransNet follows the same sampling strategy as in TransNet for the direction vectors $\{\mathbf{a}_m\}_{m=1}^{N_1}$ (cf. Theorem 5.1), but the neuron biases are drawn from a centrally symmetric distribution, i.e.,

$$r_m \sim \mathcal{U}[-1, 1], \quad m = 1, \dots, N_1, \tag{5.3}$$

instead of $r_m \sim \mathcal{U}[0, 1]$ in the original TransNet. This symmetric choice makes the first-layer activations zero-mean (cf. Theorem 5.3), which the one-sided distribution of TransNet does not, and this property propagates through the deeper hidden layers (cf. Theorem 5.4). For the subsequent hidden layers, the weight matrices $\{\mathbf{W}_l\}_{l=2}^{L}$ are predetermined through a variance-controlled sampling strategy inspired by the Xavier initialization [67]:

$$W_{ij}^{(l)} \sim \mathcal{N}\left(0, \sigma_l^2\right) \quad \text{with} \quad \sigma_l = \sqrt{\frac{\delta}{N_{l-1}}},$$

where $0 < \delta \leq 1$ is a variance control parameter.

Theoretical properties

Theorem 5.3 (Zero mean of the first hidden layer). For any fixed $\mathbf{x} \in \mathbb{R}^d$, the random variable

$\psi_m^{(1)}(\mathbf{x})$ defined by (5.2) with the sampling (5.3) satisfies

$$\psi_m^{(1)}(\mathbf{x}) \overset{d}{=} -\psi_m^{(1)}(\mathbf{x}), \quad m = 1, \dots, N_1,$$

where $\overset{d}{=}$ denotes equality in distribution. Consequently, $\mathbb{E}[\boldsymbol{\psi}_1(\mathbf{x})] = \mathbf{0}$.

Theorem 5.4 (Zero mean propagation). The means of the neuron pre-activations and activations in the subsequent hidden layers of GTransNet (5.2) are all zeros, i.e.,

$$\mathbb{E}[\mathbf{W}_l \boldsymbol{\psi}_{l-1}] = \mathbf{0} \quad \text{and} \quad \mathbb{E}[\boldsymbol{\psi}_l] = \mathbf{0}, \quad l = 2, \dots, L.$$

Theorem 5.5 (Controlled variance propagation). Assume that each component of $\boldsymbol{\psi}_1$ has a variance at most σ_0^2. Then the variances of the neuron activations in the subsequent hidden layers of GTransNet (5.2) satisfy

$$\mathrm{Var}[\psi_i^{(l)}] \leq \delta^{l-1}\sigma_0^2, \quad i = 1, \dots, N_l, \quad l = 2, \dots, L.$$

Theorem 5.5 shows that by properly choosing the parameter $0 < \delta \leq 1$, the variance of neuron values in the deeper hidden layers can be effectively reduced, preventing them from clustering in the saturation regions. This mechanism is particularly beneficial for capturing sharp interfaces and steep gradients that arise in the solutions of the CH equation.

5.2 GTransNet-BDF schemes for the Cahn-Hilliard equation

5.2.1 Model problem

For a bounded domain $\Omega \subset \mathbb{R}^d$ $(d \leq 3)$ and a terminal time $T > 0$, we consider the following CH equation:

$$\frac{\partial u}{\partial t} = D\Delta(-\varepsilon^2 \Delta u + f(u)), \quad (\mathbf{x}, t) \in \Omega \times (0, T], \tag{5.4}$$

subject to the initial condition $u(\cdot, 0) = u_0$ and periodic or homogeneous Neumann boundary conditions. In (5.4), $u(\mathbf{x}, t)$ denotes the order parameter defined as the concentration difference between the two phases, ε represents the interfacial thickness, D is the diffusion coefficient, and $f(u) = F'(u)$, where $F(u)$ is a double-well potential. In the mixed formulation, the fourth-order CH equation (5.4) is split into a system of two second-order equations:

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\Delta\mu, & \text{in } \Omega \times (0, T], \\ \mu = -\varepsilon^2 \Delta u + f(u), & \text{in } \Omega \times (0, T], \end{cases} \tag{5.5}$$

where μ denotes the chemical potential. Note that (5.4), or equivalently (5.5), can be viewed as an H^{-1} gradient flow with respect to the Ginzburg-Landau free energy functional:

$$E(u) = \int_\Omega \left(\frac{\varepsilon^2}{2} |\nabla u|^2 + F(u) \right) d\mathbf{x}. \tag{5.6}$$

Under the prescribed boundary conditions, the CH equation possesses two intrinsic properties, namely mass conservation and energy dissipation:

$$\frac{d}{dt} \int_\Omega u(\mathbf{x}, t)\, d\mathbf{x} = 0, \qquad \frac{d}{dt} E(u) \leq 0, \quad \forall\, t \in (0, T]. \tag{5.7}$$

Next, we construct the GTransNet-BDF schemes for the mixed system (5.5) and establish discrete analogues of both properties (5.7) in Section 5.3.

5.2.2 GTransNet-BDF schemes

We now develop time-stepping schemes that combine the GTransNet basis from Section 5.1 for spatial approximation with first- and second-order BDF for time integration. Consider a uniform partition of the time interval $0 = t_0 < t_1 < \cdots < t_K = T$ with the time step size $\Delta t = T/K$. Let $\kappa \geq 0$ be a stabilization constant and $f_\kappa(u) = f(u) - \kappa u$, then (5.5) can be written equivalently as

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\Delta\mu, & \text{in } \Omega \times (0, T], \\ \mu = -\varepsilon^2 \Delta u + \kappa u + f_\kappa(u), & \text{in } \Omega \times (0, T]. \end{cases} \tag{5.8}$$

In what follows, let $\{\phi_j(\mathbf{x})\}_{j=1}^{N_L}$ denote the components of the last hidden layer $\boldsymbol{\psi}_L$ of the GTransNet (5.2); that is, $\phi_j := \psi_j^{(L)}$ for $j = 1, \ldots, N_L$. The approximate solutions u_{NN} and μ_{NN} for the order parameter u and the chemical potential μ are then expressed as

$$u(\mathbf{x},t) \approx u_{\mathsf{NN}}(\mathbf{x},t) = \sum_{j=1}^{N_L} \alpha_j(t)\,\phi_j(\mathbf{x}), \quad \mu(\mathbf{x},t) \approx \mu_{\mathsf{NN}}(\mathbf{x},t) = \sum_{j=1}^{N_L} \beta_j(t)\,\phi_j(\mathbf{x}). \tag{5.9}$$

Substituting the GTransNet approximations (5.9) into the stabilized mixed system (5.8) and discretizing in time by the BDF method with explicit treatment of the nonlinear term, we obtain the first-order GTransNet-BDF1 scheme

$$\sum_{j=1}^{N_L} \frac{\alpha_j^{n+1} - \alpha_j^n}{\Delta t} \phi_j(\mathbf{x}) = D \sum_{j=1}^{N_L} \beta_j^{n+1} \Delta \phi_j(\mathbf{x}), \tag{5.10a}$$

$$\sum_{j=1}^{N_L} \beta_j^{n+1} \phi_j(\mathbf{x}) = \sum_{j=1}^{N_L} \alpha_j^{n+1} (-\varepsilon^2 \Delta + \kappa) \phi_j(\mathbf{x}) + f_\kappa(u_{\mathsf{NN}}^n(\mathbf{x})), \tag{5.10b}$$

and the second-order GTransNet-BDF2 scheme

$$\sum_{j=1}^{N_L} \frac{3\alpha_j^{n+1} - 4\alpha_j^n + \alpha_j^{n-1}}{2\Delta t} \phi_j(\mathbf{x}) = D \sum_{j=1}^{N_L} \beta_j^{n+1} \Delta \phi_j(\mathbf{x}), \tag{5.11a}$$

$$\sum_{j=1}^{N_L} \beta_j^{n+1} \phi_j(\mathbf{x}) = \sum_{j=1}^{N_L} \alpha_j^{n+1} (-\varepsilon^2 \Delta + \kappa) \phi_j(\mathbf{x}) + f_\kappa(2u_{\mathsf{NN}}^n(\mathbf{x}) - u_{\mathsf{NN}}^{n-1}(\mathbf{x})), \tag{5.11b}$$

where $u_{\mathsf{NN}}^k(\mathbf{x}) = \sum_{j=1}^{N_L} \alpha_j^k\, \phi_j(\mathbf{x})$ for $k \in \{n-1, n\}$. The schemes are initialized by fitting the initial condition u_0 in the least-squares sense to obtain $\{\alpha_j^0\}_{j=1}^{N_L}$, and for the GTransNet-BDF2 scheme, the solution $\{\alpha_j^1\}_{j=1}^{N_L}$ at $t = t_1$ is computed using the GTransNet-BDF1 scheme (5.10). We note that the nonlinear term in (5.11b) is treated as $f_\kappa(2u_{\mathsf{NN}}^n - u_{\mathsf{NN}}^{n-1})$, rather than by the standard extrapolation $2f_\kappa(u_{\mathsf{NN}}^n) - f_\kappa(u_{\mathsf{NN}}^{n-1})$, to relax the time step size restriction in the energy stability analysis (cf. Remark 5.9).

For homogeneous Neumann boundary conditions, we impose

$$\sum_{j=1}^{N_L} \alpha_j^n \nabla \phi_j(\mathbf{x}) \cdot \mathbf{n} = \sum_{j=1}^{N_L} \beta_j^n \nabla \phi_j(\mathbf{x}) \cdot \mathbf{n} = 0, \quad \forall \mathbf{x} \in \partial\Omega,\ 1 \le n \le K, \tag{5.12}$$

where $\mathbf{n}$ denotes the outward unit normal vector to $\partial\Omega$. For periodic boundary conditions, we take $\Omega = (0,1)^d$ for the purpose of presentation and impose

$$\begin{aligned} &\sum_{j=1}^{N_L} \alpha_j^n \left(\partial_{x_i}^k \phi_j|_{x_i=0} - \partial_{x_i}^k \phi_j|_{x_i=1} \right) = 0, \quad 1 \le n \le K, \\ &\sum_{j=1}^{N_L} \beta_j^n \left(\partial_{x_i}^k \phi_j|_{x_i=0} - \partial_{x_i}^k \phi_j|_{x_i=1} \right) = 0, \quad 1 \le n \le K, \end{aligned} \tag{5.13}$$

for each coordinate direction $i \in \{1, 2, \dots, d\}$ and $k \in \{0, 1\}$, here $\mathbf{x} = (x_1, x_2, \dots, x_d)^T \in \partial\Omega$.

5.2.3 Implementation

Next, we describe how the GTransNet-BDF schemes (5.10)–(5.11) are implemented in practice. Let $\{\mathbf{x}_i\}_{i=1}^{K_{\text{int}}}$ and $\{\mathbf{x}_i^{\text{bd}}\}_{i=1}^{K_{\text{bd}}}$ denote the sets of interior and boundary collocation points, respectively. We then define the interior basis matrices and the boundary basis matrix as

$$\begin{gathered} \mathbf{\Phi} = \left(\phi_j(\mathbf{x}_i)\right) \in \mathbb{R}^{K_{\text{int}} \times N_L}, \quad \mathbf{\Phi}_\Delta = \left(\Delta\phi_j(\mathbf{x}_i)\right) \in \mathbb{R}^{K_{\text{int}} \times N_L}, \\ \mathbf{\Phi}_{\text{bd}} = \left(\partial_{\mathbf{n}}\phi_j(\mathbf{x}_i^{\text{bd}})\right) \in \mathbb{R}^{K_{\text{bd}} \times N_L}, \text{ where } \partial_{\mathbf{n}}\phi_j := \nabla\phi_j \cdot \mathbf{n}, \end{gathered}$$

in the homogeneous Neumann case (and analogously, $\mathbf{\Phi}_{\text{bd}}$ collects the value and first-derivative differences across paired faces in the periodic case). With $\boldsymbol{\alpha}^n = (\alpha_1^n, \dots, \alpha_{N_L}^n)^T$ and $\boldsymbol{\beta}^n = (\beta_1^n, \dots, \beta_{N_L}^n)^T$ the coefficient vectors of u_{NN}^n and μ_{NN}^n, respectively, the unknown vector $\mathbf{c}^{n+1} = ((\boldsymbol{\alpha}^{n+1})^T, (\boldsymbol{\beta}^{n+1})^T)^T \in \mathbb{R}^{2N_L}$ is obtained for both schemes by solving the linear least-squares system

$$\mathbf{A}\,\mathbf{c}^{n+1} = \mathbf{b}^{n+1}. \tag{5.14}$$

For the GTransNet-BDF1 scheme (5.10), the system matrix and right-hand side are

$$\mathbf{A} = \mathbf{A}_{\text{BDF1}} := \begin{pmatrix} \mathbf{\Phi} & -D\Delta t\,\mathbf{\Phi}_\Delta \\ \varepsilon^2\mathbf{\Phi}_\Delta - \kappa\mathbf{\Phi} & \mathbf{\Phi} \\ \mathbf{\Phi}_{\text{bd}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_{\text{bd}} \end{pmatrix}, \qquad \mathbf{b}^{n+1} = \mathbf{b}_{\text{BDF1}}^{n+1} := \begin{pmatrix} \mathbf{\Phi}\boldsymbol{\alpha}^n \\ f_\kappa(\mathbf{\Phi}\boldsymbol{\alpha}^n) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

For the GTransNet-BDF2 scheme (5.11),

$$\mathbf{A} = \mathbf{A}_{\mathsf{BDF2}} := \begin{pmatrix} \frac{3}{2}\mathbf{\Phi} & -D\Delta t\,\mathbf{\Phi}_{\Delta} \\ \varepsilon^2\mathbf{\Phi}_{\Delta} - \kappa\mathbf{\Phi} & \mathbf{\Phi} \\ \mathbf{\Phi}_{\mathsf{bd}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_{\mathsf{bd}} \end{pmatrix}, \qquad \mathbf{b}^{n+1} = \mathbf{b}^{n+1}_{\mathsf{BDF2}} := \begin{pmatrix} 2\mathbf{\Phi}\boldsymbol{\alpha}^n - \frac{1}{2}\mathbf{\Phi}\boldsymbol{\alpha}^{n-1} \\ f_\kappa(2\mathbf{\Phi}\boldsymbol{\alpha}^n - \mathbf{\Phi}\boldsymbol{\alpha}^{n-1}) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

In both cases, the third and fourth block rows of $\mathbf{A}$ impose the boundary conditions on $\boldsymbol{\alpha}^{n+1}$ and $\boldsymbol{\beta}^{n+1}$, respectively. Since the coefficient matrix $\mathbf{A}$ in (5.14) (that is, $\mathbf{A}_{\mathsf{BDF1}}$ or $\mathbf{A}_{\mathsf{BDF2}}$) is independent of n, we precompute its reduced QR factorization $\mathbf{A} = \mathbf{QR}$ once and then solve

$$\mathbf{R}\mathbf{c}^{n+1} = \mathbf{Q}^T\mathbf{b}^{n+1} \tag{5.15}$$

via back substitution at each time step. Consequently, the per-step cost is dominated by the matrix-vector product $\mathbf{Q}^T\mathbf{b}^{n+1}$, which is more efficient than directly solving the least-squares system (5.14) at each time step.

At $t = 0$, the initial condition $u_0(\mathbf{x})$ is projected onto the GTransNet basis by solving the least-squares problem

$$\boldsymbol{\alpha}^0 = \arg\min_{\alpha\in\mathbb{R}^{N_L}} \|\mathbf{\Phi}_{\mathsf{init}}\boldsymbol{\alpha} - \mathbf{u}_0\|^2_{l^2},$$

where $\mathbf{\Phi}_{\mathsf{init}} = (\phi_j(\mathbf{x}_i^{\mathsf{init}}))_{i,j}$ is the basis matrix evaluated at a set of initialization points $\{\mathbf{x}_i^{\mathsf{init}}\}$, which includes both interior and boundary points, and $\mathbf{u}_0 = (u_0(\mathbf{x}_i^{\mathsf{init}}))_i$.

5.3 Properties of the GTransNet-BDF schemes

In this section, we establish the mass conservation and energy stability of the first- and second-order GTransNet-BDF schemes (5.10)–(5.11). To that end, we define the time-discrete mass and free energy (cf. (5.6)) as

$$\mathcal{M}(u^n_{\mathsf{NN}}) = \int_\Omega u^n_{\mathsf{NN}}(\mathbf{x})\,d\mathbf{x}, \quad E(u^n_{\mathsf{NN}}) = \int_\Omega \left(\frac{\varepsilon^2}{2}|\nabla u^n_{\mathsf{NN}}|^2 + F(u^n_{\mathsf{NN}})\right)d\mathbf{x}.$$

Throughout this section, $(\cdot,\cdot)$ and $\|\cdot\|_0$ denote the standard $L^2(\Omega)$ inner product and associated norm, respectively. We shall restrict our attention to potential functions $F(u)$ whose derivative $f(u) = F'(u)$ satisfies the following condition

$$L_f := \max_{u\in\mathbb{R}} |f'(u)| < \infty, \tag{5.16}$$

which holds under a truncation of F (and hence f) outside a bounded interval; we refer to [166, 62] for the explicit form of a truncated double-well potential with quadratic growth at infinity. For the analysis of the GTransNet-BDF2 scheme, we further recall the inverse Laplacian operator $(-\Delta)^{-1}$ defined on the mean-zero subspace

$$L_0^2(\Omega) := \Big\{ v \in L^2(\Omega) : \int_\Omega v \, d\mathbf{x} = 0 \Big\}.$$

For $v \in L_0^2(\Omega)$, we define $(-\Delta)^{-1} v = \varphi$ with $\varphi \in H^1(\Omega) \cap L_0^2(\Omega)$ solving the Poisson's equation $-\Delta\varphi = v$ under either periodic or homogeneous Neumann boundary conditions. The associated H^{-1}-norm is given by $\|v\|_{-1} := \sqrt{(v, (-\Delta)^{-1} v)}$. One easily verifies the interpolation inequality

$$\|v\|_0^2 \le \|\nabla v\|_0 \cdot \|v\|_{-1}, \quad \forall\, v \in H^1(\Omega) \cap L_0^2(\Omega).$$

As a consequence, applying Young's inequality yields

$$a\|v\|_{-1}^2 + b\|\nabla v\|_0^2 \ge 2\sqrt{ab}\,\|v\|_0^2, \quad \forall\, a, b \ge 0,\ \forall\, v \in H^1(\Omega) \cap L_0^2(\Omega). \tag{5.17}$$

Finally, we denote $\delta_t v^{n+1} := v^{n+1} - v^n$ and $\delta_{tt} v^{n+1} := v^{n+1} - 2v^n + v^{n-1}$ for any sequence $\{v^n\}$.

By integrating the first equations (5.10a) and (5.11a) of the GTransNet-BDF1 and GTransNet-BDF2 schemes over Ω, applying the divergence theorem, and invoking the boundary conditions (5.12) or (5.13) on μ_{NN}^{n+1}, we obtain the following discrete mass conservation.

Theorem 5.6 (Mass conservation). The first- and second-order GTransNet-BDF schemes (cf. (5.10) and (5.11)) conserve mass unconditionally, i.e., for any time step size $\Delta t > 0$, we have

$$\mathcal{M}(u_{\mathsf{NN}}^{n+1}) = \mathcal{M}(u_{\mathsf{NN}}^{n}), \quad 0 \le n \le K - 1.$$

Following the standard energy-estimate technique for the stabilized BDF1 discretization of the CH equation developed in [166], we have the following result.

Theorem 5.7 (Unconditional energy stability of GTransNet-BDF1). If the stabilization constant satisfies $\kappa \geq L_f/2$, then the GTransNet-BDF1 scheme (5.10) is unconditionally energy stable, that is,

$$E(u_{\mathrm{NN}}^{n+1}) \leq E(u_{\mathrm{NN}}^{n}), \quad 0 \leq n \leq K-1.$$

For the GTransNet-BDF2 scheme, the analysis is more involved and yields only conditional energy stability under a time step size restriction, as stated in the following theorem.

Theorem 5.8 (Conditional energy stability of GTransNet-BDF2). If the stabilization constant and the time step size satisfy

$$\kappa \geq L_f, \quad \Delta t \leq \frac{8\varepsilon^2}{DL_f^2}, \tag{5.18}$$

then the GTransNet-BDF2 scheme (5.11) is energy stable in the sense that

$$\widetilde{E}(u_{\mathrm{NN}}^{n+1}) \leq \widetilde{E}(u_{\mathrm{NN}}^{n}), \quad 1 \leq n \leq K-1, \tag{5.19}$$

where the modified energy is defined by

$$\widetilde{E}(u_{\mathrm{NN}}^{n}) := E(u_{\mathrm{NN}}^{n}) + \frac{1}{4D\Delta t}\|\delta_t u_{\mathrm{NN}}^{n}\|_{-1}^2 + \frac{\kappa + L_f}{2}\|\delta_t u_{\mathrm{NN}}^{n}\|_0^2.$$

Proof. By the mass conservation of GTransNet-BDF schemes in Theorem 5.6, $\delta_t u_{\mathrm{NN}}^{n+1} \in L_0^2(\Omega)$, so $(-\Delta)^{-1}\delta_t u_{\mathrm{NN}}^{n+1}$ is well-defined. Taking the L^2 inner product of (5.11a) with $(-\Delta)^{-1}\delta_t u_{\mathrm{NN}}^{n+1}$, applying the divergence theorem together with the boundary conditions, and dividing both sides by D, we obtain

$$\frac{1}{2D\Delta t}\big(3u_{\mathrm{NN}}^{n+1} - 4u_{\mathrm{NN}}^{n} + u_{\mathrm{NN}}^{n-1}, (-\Delta)^{-1}\delta_t u_{\mathrm{NN}}^{n+1}\big) = -(\mu_{\mathrm{NN}}^{n+1}, \delta_t u_{\mathrm{NN}}^{n+1}). \tag{5.20}$$

Note that (5.11b) can be rewritten, using $f_\kappa(u) = f(u) - \kappa u$, as

$$\mu_{\mathrm{NN}}^{n+1} = -\varepsilon^2 \Delta u_{\mathrm{NN}}^{n+1} + f(2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1}) + \kappa \delta_{tt} u_{\mathrm{NN}}^{n+1}. \tag{5.21}$$

Taking the L^2 inner product of (5.21) with $\delta_t u_{\mathrm{NN}}^{n+1}$ and applying integration by parts, we obtain

$$\big(\mu_{\mathrm{NN}}^{n+1}, \delta_t u_{\mathrm{NN}}^{n+1}\big) = \varepsilon^2\big(\nabla u_{\mathrm{NN}}^{n+1}, \nabla \delta_t u_{\mathrm{NN}}^{n+1}\big) + \big(f(2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1}), \delta_t u_{\mathrm{NN}}^{n+1}\big) + \kappa\big(\delta_{tt} u_{\mathrm{NN}}^{n+1}, \delta_t u_{\mathrm{NN}}^{n+1}\big). \tag{5.22}$$

Combining (5.20) and (5.22) yields

$$\begin{aligned} &\frac{1}{2D\Delta t}\big(3u_{\mathrm{NN}}^{n+1} - 4u_{\mathrm{NN}}^n + u_{\mathrm{NN}}^{n-1}, (-\Delta)^{-1}\delta_t u_{\mathrm{NN}}^{n+1}\big) + \varepsilon^2\big(\nabla u_{\mathrm{NN}}^{n+1}, \nabla \delta_t u_{\mathrm{NN}}^{n+1}\big) \\ &\quad + \kappa\big(\delta_{tt} u_{\mathrm{NN}}^{n+1}, \delta_t u_{\mathrm{NN}}^{n+1}\big) = -\big(f(2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1}), \delta_t u_{\mathrm{NN}}^{n+1}\big). \end{aligned} \tag{5.23}$$

For the first term in (5.23), by decomposing $3u_{\mathrm{NN}}^{n+1} - 4u_{\mathrm{NN}}^n + u_{\mathrm{NN}}^{n-1} = 2\delta_t u_{\mathrm{NN}}^{n+1} + \delta_{tt} u_{\mathrm{NN}}^{n+1}$ and applying the identity $a(a-b) = \frac{1}{2}[a^2 - b^2 + (a-b)^2]$ in the H^{-1} inner product, we obtain

$$\begin{aligned} &\frac{1}{2D\Delta t}\big(3u_{\mathrm{NN}}^{n+1} - 4u_{\mathrm{NN}}^n + u_{\mathrm{NN}}^{n-1}, (-\Delta)^{-1}\delta_t u_{\mathrm{NN}}^{n+1}\big) \\ &\quad = \frac{1}{D\Delta t}\|\delta_t u_{\mathrm{NN}}^{n+1}\|_{-1}^2 + \frac{1}{4D\Delta t}\big(\|\delta_t u_{\mathrm{NN}}^{n+1}\|_{-1}^2 - \|\delta_t u_{\mathrm{NN}}^n\|_{-1}^2 + \|\delta_{tt} u_{\mathrm{NN}}^{n+1}\|_{-1}^2\big). \end{aligned} \tag{5.24}$$

For the remaining two terms on the left-hand side of (5.23), we again apply the identity $a(a-b) = \frac{1}{2}[a^2 - b^2 + (a-b)^2]$, with $\delta_{tt} u_{\mathrm{NN}}^{n+1} = \delta_t u_{\mathrm{NN}}^{n+1} - \delta_t u_{\mathrm{NN}}^n$, to deduce that

$$\begin{aligned} \varepsilon^2\big(\nabla u_{\mathrm{NN}}^{n+1}, \nabla \delta_t u_{\mathrm{NN}}^{n+1}\big) &= \frac{\varepsilon^2}{2}\big(\|\nabla u_{\mathrm{NN}}^{n+1}\|_0^2 - \|\nabla u_{\mathrm{NN}}^n\|_0^2 + \|\nabla \delta_t u_{\mathrm{NN}}^{n+1}\|_0^2\big), \\ \kappa\big(\delta_{tt} u_{\mathrm{NN}}^{n+1}, \delta_t u_{\mathrm{NN}}^{n+1}\big) &= \frac{\kappa}{2}\big(\|\delta_t u_{\mathrm{NN}}^{n+1}\|_0^2 - \|\delta_t u_{\mathrm{NN}}^n\|_0^2 + \|\delta_{tt} u_{\mathrm{NN}}^{n+1}\|_0^2\big). \end{aligned} \tag{5.25}$$

For the right-hand side of (5.23), we apply the Taylor expansion of F around $2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1}$. Noting that $u_{\mathrm{NN}}^{n+1} - (2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1}) = \delta_{tt} u_{\mathrm{NN}}^{n+1}$ and $u_{\mathrm{NN}}^n - (2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1}) = -\delta_t u_{\mathrm{NN}}^n$, the bound (5.16) on $F'' = f'$ yields

$$\begin{aligned} F(u_{\mathrm{NN}}^{n+1}) &\le F(2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1}) + f(2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1})\,\delta_{tt} u_{\mathrm{NN}}^{n+1} + \frac{L_f}{2}|\delta_{tt} u_{\mathrm{NN}}^{n+1}|^2, \\ F(u_{\mathrm{NN}}^n) &\ge F(2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1}) - f(2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1})\,\delta_t u_{\mathrm{NN}}^n - \frac{L_f}{2}|\delta_t u_{\mathrm{NN}}^n|^2. \end{aligned} \tag{5.26}$$

Subtracting the second inequality from the first in (5.26) and using $\delta_{tt}u_{\mathrm{NN}}^{n+1}+\delta_t u_{\mathrm{NN}}^n=\delta_t u_{\mathrm{NN}}^{n+1}$ yields

$$F(u_{\mathrm{NN}}^{n+1})-F(u_{\mathrm{NN}}^n)\le f(2u_{\mathrm{NN}}^n-u_{\mathrm{NN}}^{n-1})\,\delta_t u_{\mathrm{NN}}^{n+1}+\frac{L_f}{2}\big(|\delta_{tt}u_{\mathrm{NN}}^{n+1}|^2+|\delta_t u_{\mathrm{NN}}^n|^2\big). \tag{5.27}$$

Integrating (5.27) over Ω and rearranging gives

$$-\big(f(2u_{\mathrm{NN}}^n-u_{\mathrm{NN}}^{n-1}),\delta_t u_{\mathrm{NN}}^{n+1}\big)\le-\big(F(u_{\mathrm{NN}}^{n+1})-F(u_{\mathrm{NN}}^n),1\big)+\frac{L_f}{2}\big(\|\delta_{tt}u_{\mathrm{NN}}^{n+1}\|_0^2+\|\delta_t u_{\mathrm{NN}}^n\|_0^2\big). \tag{5.28}$$

Substituting (5.24), (5.25), and (5.28) into (5.23) yields

$$\begin{aligned}
&E(u_{\mathrm{NN}}^{n+1})-E(u_{\mathrm{NN}}^n)+\frac{\varepsilon^2}{2}\|\nabla\delta_t u_{\mathrm{NN}}^{n+1}\|_0^2+\frac{1}{D\Delta t}\|\delta_t u_{\mathrm{NN}}^{n+1}\|_{-1}^2\\
&\quad+\frac{1}{4D\Delta t}\big(\|\delta_t u_{\mathrm{NN}}^{n+1}\|_{-1}^2-\|\delta_t u_{\mathrm{NN}}^n\|_{-1}^2+\|\delta_{tt}u_{\mathrm{NN}}^{n+1}\|_{-1}^2\big)\\
&\quad+\frac{\kappa}{2}\big(\|\delta_t u_{\mathrm{NN}}^{n+1}\|_0^2-\|\delta_t u_{\mathrm{NN}}^n\|_0^2+\|\delta_{tt}u_{\mathrm{NN}}^{n+1}\|_0^2\big)\le\frac{L_f}{2}\big(\|\delta_{tt}u_{\mathrm{NN}}^{n+1}\|_0^2+\|\delta_t u_{\mathrm{NN}}^n\|_0^2\big).
\end{aligned}\tag{5.29}$$

Finally, applying the inequality (5.17) with $a=1/(D\Delta t)$ and $b=\varepsilon^2/2$ yields, under the time step constraint (5.18),

$$\frac{\varepsilon^2}{2}\|\nabla\delta_t u_{\mathrm{NN}}^{n+1}\|_0^2+\frac{1}{D\Delta t}\|\delta_t u_{\mathrm{NN}}^{n+1}\|_{-1}^2\ge\varepsilon\sqrt{\frac{2}{D\Delta t}}\,\|\delta_t u_{\mathrm{NN}}^{n+1}\|_0^2\ge\frac{L_f}{2}\|\delta_t u_{\mathrm{NN}}^{n+1}\|_0^2. \tag{5.30}$$

Combining (5.29) and (5.30) with $\kappa\ge L_f$, we arrive at (5.19). $\square$

Remark 5.9.

(i) The proof of Theorem 5.8 partially adapts the energy-estimate framework developed in [179], which analyzes a more general family of stabilized second-order semi-implicit BDF schemes for the CH equation. Particularly, in addition to the stabilization term $\kappa(u_{\mathrm{NN}}^{n+1}-2u_{\mathrm{NN}}^n+u_{\mathrm{NN}}^{n-1})$, the schemes proposed in [179] incorporate an additional Laplacian-type stabilization of the form $-\hat{\kappa}\Delta t\Delta(u_{\mathrm{NN}}^{n+1}-u_{\mathrm{NN}}^n)$ with $\hat{\kappa}>0$ in the chemical potential equation. This extra term contributes to the left-hand side of (5.29) an amount of $\hat{\kappa}\Delta t\|\nabla\delta_t u_{\mathrm{NN}}^{n+1}\|_0^2$, yielding unconditional energy stability with an appropriate choice of $\hat{\kappa}$.

(ii) If the second-order extrapolation $f(2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1})$ in the GTransNet-BDF2 scheme (5.11) is replaced by the standard linearization $2f(u_{\mathrm{NN}}^n) - f(u_{\mathrm{NN}}^{n-1})$, an analogous energy stability result holds only under the more restrictive time step constraint $\Delta t \leq 8\varepsilon^2/(9DL_f^2)$ (cf. [166, 179]). Moreover, this constraint cannot be relaxed by increasing the stabilization constant κ. The form $f(2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1})$ adopted here produces the term $\|\delta_{tt}u_{\mathrm{NN}}^{n+1}\|_0^2$ in the nonlinear estimate (5.28), which can be absorbed by the corresponding stabilization term in (5.29) provided that $\kappa \geq L_f$, thereby relaxing the time step constraint by a factor of 9.

5.4 Discrete mass-conserving projection

While Theorem 5.6 guarantees mass conservation in the space-continuous sense, the fully discrete mass computed from $\mathbf{c}_{\mathrm{unc}}^{n+1} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{b}^{n+1}$ (cf. (5.15)) may not be conserved exactly, since the over-determined least-squares system (5.14) generally has nonzero residuals at the collocation points. In this section, we describe a post-processing projection that enforces exact discrete mass conservation with minimal perturbation of the least-squares approximation, and establish the key properties of this projection that justify its use in long-time simulations.

For all numerical computations and for the projection introduced below, we denote by $\mathcal{M}^n$ an approximation of $\mathcal{M}(u_{\mathrm{NN}}^n)$, i.e.,

$$\mathcal{M}(u_{\mathrm{NN}}^n) \approx \mathcal{M}^n := \sum_q \omega_q\, u_{\mathrm{NN}}^n(\mathbf{x}_q),$$

where $\{\mathbf{x}_q, \omega_q\}$ denotes a set of quadrature points and weights; in the numerical experiments, we use the tensor-product two-point Gauss-Legendre quadrature rule. Substituting the basis expansion $u_{\mathrm{NN}}^n(\mathbf{x}) = \sum_{j=1}^{N_L} \alpha_j^n \phi_j(\mathbf{x})$ into the definition of $\mathcal{M}^n$ gives the following identity

$$\mathcal{M}^n = \mathbf{g}^T\boldsymbol{\alpha}^n, \text{ where } \mathbf{g} = (g_1, \ldots, g_{N_L})^T \in \mathbb{R}^{N_L} \text{ and } g_j = \sum_q \omega_q \phi_j(\mathbf{x}_q),\ 1 \leq j \leq N_L.$$

With $\hat{\mathbf{g}} := (\mathbf{g}^T, \mathbf{0}^T)^T \in \mathbb{R}^{2N_L}$, the discrete mass-conservation requirement $\mathcal{M}^{n+1} = \mathcal{M}^0$ becomes the single linear constraint

$$\hat{\mathbf{g}}^T\mathbf{c}^{n+1} = \mathcal{M}^0, \quad \text{where } \mathbf{c}^{n+1} = ((\boldsymbol{\alpha}^{n+1})^T, (\boldsymbol{\beta}^{n+1})^T)^T. \tag{5.31}$$

5.4.1 Constrained least-squares formulation

The unconstrained solution $\mathbf{c}_{\text{unc}}^{n+1}$ approximates the discretized PDE in the least-squares sense and satisfies (5.31) up to a small residual at the level of the least-squares error. We therefore correct $\mathbf{c}_{\text{unc}}^{n+1}$ by the smallest perturbation, in the weighted norm induced by $\mathbf{R}^T\mathbf{R}$, that exactly restores (5.31):

$$\mathbf{c}^{n+1} = \arg\min_{\mathbf{c}\in\mathbb{R}^{2N_L}} \|\mathbf{c} - \mathbf{c}_{\text{unc}}^{n+1}\|_{\mathbf{R}^T\mathbf{R}}^2 \quad \text{subject to} \quad \hat{\mathbf{g}}^T\mathbf{c} = \mathcal{M}^0, \tag{5.32}$$

where $\|\mathbf{v}\|_{\mathbf{R}^T\mathbf{R}}^2 = \mathbf{v}^T\mathbf{R}^T\mathbf{R}\mathbf{v} = \|\mathbf{R}\mathbf{v}\|_{l^2}^2$ for all $\mathbf{v} \in \mathbb{R}^{2N_L}$. The choice of the $\mathbf{R}^T\mathbf{R}$-norm is natural for least-squares problems: $\|\mathbf{R}(\mathbf{c} - \mathbf{c}_{\text{unc}}^{n+1})\|_{l^2}$ measures exactly the increase in the least-squares residual norm caused by the correction, so among all feasible coefficient vectors the minimizer $\mathbf{c}^{n+1}$ is the one that least perturbs the discretized-PDE fit. The associated Lagrangian

$$\mathcal{L}(\mathbf{c}, \lambda) = \frac{1}{2}\|\mathbf{c} - \mathbf{c}_{\text{unc}}^{n+1}\|_{\mathbf{R}^T\mathbf{R}}^2 + \lambda\big(\hat{\mathbf{g}}^T\mathbf{c} - \mathcal{M}^0\big)$$

yields the Karush-Kuhn-Tucker (KKT) system $\mathbf{R}^T\mathbf{R}(\mathbf{c} - \mathbf{c}_{\text{unc}}^{n+1}) + \lambda\hat{\mathbf{g}} = \mathbf{0}$ together with the mass constraint $\hat{\mathbf{g}}^T\mathbf{c} = \mathcal{M}^0$, from which the closed form

$$\mathbf{c}^{n+1} = \mathbf{c}_{\text{unc}}^{n+1} - \frac{\hat{\mathbf{g}}^T\mathbf{c}_{\text{unc}}^{n+1} - \mathcal{M}^0}{\hat{\mathbf{g}}^T\mathbf{z}}\,\mathbf{z}, \qquad \mathbf{z} := (\mathbf{R}^T\mathbf{R})^{-1}\hat{\mathbf{g}}, \tag{5.33}$$

follows. In (5.33), the vector $\mathbf{z} \in \mathbb{R}^{2N_L}$ is computed once for all time steps by two triangular solves (one with $\mathbf{R}^T$ and one with $\mathbf{R}$); and the projection then requires only one inner product and a scalar correction along $\mathbf{z}$, adding $\mathcal{O}(N_L)$ operations per time step. Since this cost is dominated by the matrix-vector product $\mathbf{Q}^T\mathbf{b}^{n+1}$ and back substitution when computing $\mathbf{c}_{\text{unc}}^{n+1}$, the projection adds negligible computational cost.

5.4.2 Properties of the projection

Let $H_{\mathcal{M}^0} := \{\mathbf{c} \in \mathbb{R}^{2N_L} \,:\, \hat{\mathbf{g}}^T\mathbf{c} = \mathcal{M}^0\}$ denote the affine hyperplane of coefficient vectors that satisfy the discrete mass constraint (5.31), and define the projection operator $\mathcal{P}$:

$\mathbb{R}^{2N_L} \to H_{\mathcal{M}^0}$ by

$$\mathcal{P}(\mathbf{c}) := \mathbf{c} - \frac{\hat{\mathbf{g}}^T\mathbf{c} - \mathcal{M}^0}{\hat{\mathbf{g}}^T\mathbf{z}}\,\mathbf{z}, \quad \forall\, \mathbf{c} \in \mathbb{R}^{2N_L}, \tag{5.34}$$

so that $\mathbf{c}^{n+1} = \mathcal{P}(\mathbf{c}_{\text{unc}}^{n+1})$ (cf. (5.33)). Note that $\hat{\mathbf{g}}^T\mathbf{z} = \hat{\mathbf{g}}^T(\mathbf{R}^T\mathbf{R})^{-1}\hat{\mathbf{g}} > 0$ since $\mathbf{R}^T\mathbf{R}$ is symmetric positive definite and $\hat{\mathbf{g}} \neq \mathbf{0}$, so $\mathcal{P}$ is well defined. The map $\mathcal{P}$ possesses the following properties:

(P1) Exact discrete mass conservation. A direct computation gives

$$\hat{\mathbf{g}}^T\mathcal{P}(\mathbf{c}) = \hat{\mathbf{g}}^T\mathbf{c} - (\hat{\mathbf{g}}^T\mathbf{c} - \mathcal{M}^0) = \mathcal{M}^0, \quad \forall\, \mathbf{c} \in \mathbb{R}^{2N_L}.$$

Therefore, the projected coefficient vector $\mathbf{c}^{n+1}$ satisfies mass conservation $\mathcal{M}^{n+1} = \hat{\mathbf{g}}^T\mathbf{c}^{n+1} = \hat{\mathbf{g}}^T\mathcal{P}(\mathbf{c}_{\text{unc}}^{n+1}) = \mathcal{M}^0$ at every time step.

(P2) Idempotency. If $\mathbf{c} \in H_{\mathcal{M}^0}$, then $\hat{\mathbf{g}}^T\mathbf{c} = \mathcal{M}^0$, so $\mathcal{P}(\mathbf{c}) = \mathbf{c}$ due to (5.34). Since $\mathcal{P}(\mathbf{c}) \in H_{\mathcal{M}^0}$ for all $\mathbf{c} \in \mathbb{R}^{2N_L}$ by (P1), we have $\mathcal{P} \circ \mathcal{P} = \mathcal{P}$.

(P3) Non-expansiveness in the $\mathbf{R}^T\mathbf{R}$-norm. For all $\mathbf{c}, \mathbf{c}' \in \mathbb{R}^{2N_L}$,

$$\|\mathcal{P}(\mathbf{c}) - \mathcal{P}(\mathbf{c}')\|_{\mathbf{R}^T\mathbf{R}} \le \|\mathbf{c} - \mathbf{c}'\|_{\mathbf{R}^T\mathbf{R}}. \tag{5.35}$$

Indeed, $\mathcal{P}(\mathbf{c}) - \mathcal{P}(\mathbf{c}')$ is the $\mathbf{R}^T\mathbf{R}$-orthogonal projection of $\mathbf{c} - \mathbf{c}'$ onto the linear subspace $\hat{\mathbf{g}}^\perp = \{\mathbf{v} \in \mathbb{R}^{2N_L} : \hat{\mathbf{g}}^T\mathbf{v} = 0\}$, and the standard Pythagorean identity for orthogonal projections gives

$$\|\mathbf{c} - \mathbf{c}'\|^2_{\mathbf{R}^T\mathbf{R}} = \|\mathcal{P}(\mathbf{c}) - \mathcal{P}(\mathbf{c}')\|^2_{\mathbf{R}^T\mathbf{R}} + \frac{\left(\hat{\mathbf{g}}^T(\mathbf{c} - \mathbf{c}')\right)^2}{\hat{\mathbf{g}}^T\mathbf{z}}.$$

In particular, taking $\mathbf{c} = \mathbf{c}_{\text{unc}}^{n+1}$ and $\mathbf{c}' = \mathbf{c}^n$ in (5.35), noting that $\mathbf{c}^n = \mathcal{P}(\mathbf{c}_{\text{unc}}^n) \in H_{\mathcal{M}^0}$ by (P1) and $\mathcal{P}(\mathbf{c}^n) = \mathbf{c}^n$ by (P2), we obtain

$$\|\mathbf{c}^{n+1} - \mathbf{c}^n\|_{\mathbf{R}^T\mathbf{R}} \le \|\mathbf{c}_{\text{unc}}^{n+1} - \mathbf{c}^n\|_{\mathbf{R}^T\mathbf{R}}.$$

(P4) Bounded perturbation. Substituting (5.33) into the objective function of (5.32) and using the identity $\mathbf{z}^T\mathbf{R}^T\mathbf{R}\mathbf{z} = \hat{\mathbf{g}}^T\mathbf{z}$ (which follows from $\mathbf{z} = (\mathbf{R}^T\mathbf{R})^{-1}\hat{\mathbf{g}}$), the optimal value is

$$\|\mathbf{c}^{n+1} - \mathbf{c}_{\text{unc}}^{n+1}\|_{\mathbf{R}^T\mathbf{R}}^2 = \frac{\left(\hat{\mathbf{g}}^T\mathbf{c}_{\text{unc}}^{n+1} - \mathcal{M}^0\right)^2}{\hat{\mathbf{g}}^T\mathbf{z}},$$

i.e., the perturbation $\|\mathbf{c}^{n+1} - \mathbf{c}_{\text{unc}}^{n+1}\|_{\mathbf{R}^T\mathbf{R}}$ is proportional to the absolute mass error $|\hat{\mathbf{g}}^T\mathbf{c}_{\text{unc}}^{n+1} - \mathcal{M}^0|$ of the unconstrained solution, and vanishes when $\mathbf{c}_{\text{unc}}^{n+1}$ already satisfies the mass constraint.

Together, properties (P1)–(P4) show that the projection (5.33) enforces exact discrete mass conservation while introducing the smallest possible correction in the $\mathbf{R}^T\mathbf{R}$-norm. In contrast to a global a posteriori approach that uniformly shifts the solution to match the target mass and ignores the PDE residual, the proposed projection (5.33) corrects $\mathbf{c}_{\text{unc}}^{n+1}$ along the optimal direction $\mathbf{z}$ that minimizes the disturbance to this residual. The projection is applied at every time step in the numerical experiments of Section 5.5. We refer to [112] for a maximum-bound-principle-preserving and mass-conservative projection method for the conservative Allen–Cahn equation and to [151] for a mass-projection strategy for the CH equation on surfaces.

5.5 Numerical experiments

In this section, we demonstrate the performance of the proposed GTransNet-BDF2 scheme (5.11) coupled with the mass-conserving projection (cf. Section 5.4) on a range of test cases. For the two-dimensional CH equation, we verify the temporal convergence of the scheme via a manufactured solution and illustrate its ability to capture the correct dynamics while preserving mass conservation and energy dissipation through the shape relaxation and coarsening dynamics tests. The method is further applied to solve the CH equation on irregular domains, including circular and amoeba-shaped ones. In three dimensions, the temporal convergence of the GTransNet-BDF2 scheme is confirmed, followed by a coarsening dynamics test that demonstrates the robustness of the scheme in preserving the two intrinsic properties for long-time simulations. Finally, we extend the GTransNet-BDF framework to the CH equation with variable mobility (see Appendix C for the detailed formulation and implementation) and show its numerical performance for the degenerate mobility $M(u) = |1 - u^2|$.

Unless otherwise stated, we consider the standard double-well potential $F(u) = \frac{1}{4}(u^2 - 1)^2$, with $f(u) = F'(u) = u^3 - u$ and wells at $u = \pm 1$. The interior collocation points $\{\mathbf{x}_i\}_{i=1}^{K_{\text{int}}}$ are chosen according to the boundary condition: a uniform vertex grid on $\overline{\Omega}$ in the periodic case, and the cell centers of a uniform partition of Ω in the homogeneous Neumann counterpart; for irregular domains, we apply the same uniform partition to a bounding box of Ω and retain only the cell centers inside Ω. The boundary collocation points $\{\mathbf{x}_i^{\text{bd}}\}_{i=1}^{K_{\text{bd}}}$ are equally spaced for rectangular domains, and are placed at equally spaced parameter values along the parametrized boundaries of irregular domains. In addition, the GTransNet hidden-layer neurons are generated on a covering ball $B_R(\mathbf{x}_c) \supset \Omega$ (cf. Remark 5.2), where $\mathbf{x}_c$ is the center of Ω (or of its bounding box), and the radius R, specified for each example, is chosen so that $B_R(\mathbf{x}_c)$ is slightly larger than Ω.

We employ the GTransNet basis with $L = 2$ hidden layers in most test cases. As discussed in [24], this choice achieves a favorable balance between accuracy and computational efficiency. Adding more hidden layers (e.g., $L = 3$) may offer some accuracy improvements, but at the expense of increased computational cost in assembling the least-squares system due to the evaluation of higher-order derivatives via the chain rule. The case $L = 3$ is included only in the selected convergence tests for comparison.

5.5.1 Two-dimensional Cahn-Hilliard equation

Convergence test We set the domain $\Omega = (-0.5, 0.5)^2$ and the terminal time $T = 1$. By adding an external forcing term to the right-hand side of the first equation in (5.8), we take the exact solution to be

$$u(x, y, t) = \sin(2\pi x)\cos(2\pi y)\cos(t),$$

subject to the periodic boundary conditions. We fix $D = 0.5$, $\kappa = 0$, $\delta = 0.5$, $R = 0.75$, and vary the interfacial width $\varepsilon \in \{0.2, 0.1\}$. The collocation points are sampled from a uniform $n_x \times n_y$ mesh, and the test data consist of $4n_x n_y$ uniformly distributed random points in Ω. The parameters used for each value of ε are listed in Table 5.1.

Table 5.1: [2D convergence test] Parameter settings for $\varepsilon \in \{0.2, 0.1\}$.

(a) $L = 2$ hidden layers

ε	N_1	N_2	γ	$n_x = n_y$
0.2	600	500	3	50
0.1	1500	1000	4	80

(b) $L = 3$ hidden layers

ε	N_1	N_2	N_3	γ	$n_x = n_y$
0.2	600	500	500	2.8	50
0.1	1500	1000	1000	4	80

Table 5.2 presents L^∞ errors of the phase variable at the final time by the GTransNet-BDF2 scheme with $L = 2$ and $L = 3$ hidden layers for $\varepsilon \in \{0.2, 0.1\}$. In all cases, the scheme achieves the expected second-order convergence in time. Moreover, the errors for $L = 2$ and $L = 3$ are nearly identical, indicating that the spatial approximation error with two hidden layers is already negligible relative to the temporal error. We note that similar convergence results hold for the homogeneous Neumann boundary conditions; a corresponding three-dimensional convergence test under the same boundary conditions is presented in Section 5.5.2.

Table 5.2: [2D convergence test] L^∞ errors of the phase variable at the final time for $\varepsilon \in \{0.2, 0.1\}$.

(a) $\varepsilon = 0.2$

Δt	GTransNet-BDF2 $L = 2$ hidden layers	GTransNet-BDF2 $L = 3$ hidden layers
1/10	1.01e-03	1.01e-03
1/20	2.40e-04 [2.07]	2.40e-04 [2.07]
1/40	5.82e-05 [2.04]	5.83e-05 [2.04]
1/80	1.44e-05 [2.02]	1.44e-05 [2.02]
1/160	3.60e-06 [2.00]	3.64e-06 [1.98]
1/320	9.09e-07 [1.99]	9.32e-07 [1.97]

(b) $\varepsilon = 0.1$

Δt	GTransNet-BDF2 $L = 2$ hidden layers	GTransNet-BDF2 $L = 3$ hidden layers
1/10	5.87e-03	5.87e-03
1/20	1.54e-03 [1.93]	1.54e-03 [1.96]
1/40	3.84e-04 [2.00]	3.84e-04 [2.00]
1/80	9.54e-05 [2.01]	9.54e-05 [2.01]
1/160	2.37e-05 [2.01]	2.38e-05 [2.00]
1/320	6.08e-06 [1.96]	6.18e-06 [1.95]

Shape relaxation We next apply the GTransNet-BDF2 scheme with 2 hidden layers to a shape relaxation test case [76, 49] under homogeneous Neumann boundary conditions. We set $\Omega = (0, 1)^2$, $D = 0.05$, $\kappa = 2$, $\delta = 0.5$, $R = 0.75$, and $\varepsilon \in \{0.02, 0.01\}$. The initial condition has a value of 1 inside the square $[0.25, 0.75]^2$ and -1 outside.

For $\varepsilon = 0.02$, we run the simulation until $T = 1.5$ using $N_1 = 2000$, $N_2 = 1000$, $\gamma = 12$, $\Delta t = 1 \times 10^{-2}$, $K_{\text{int}} = 150^2$ interior and $K_{\text{bd}} = 600$ boundary collocation points. For $\varepsilon = 0.01$, we run until $T = 3$ with $N_1 = 3000$, $N_2 = 1500$, $\gamma = 16$, $\Delta t = 5 \times 10^{-3}$, $K_{\text{int}} = 200^2$, and $K_{\text{bd}} = 800$.

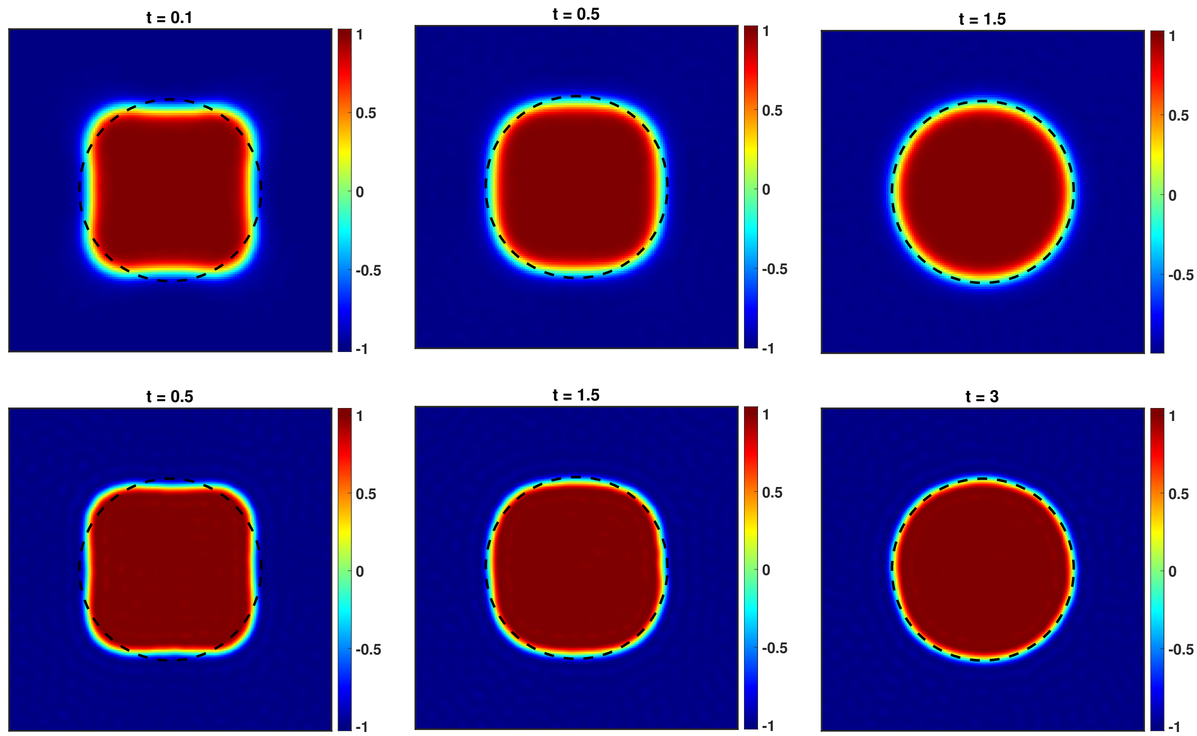


Figure 5.1: [Shape relaxation] Snapshots of the phase variable by the GTransNet-BDF2 scheme. Top: $\varepsilon = 0.02$ at $t = 0.1, 0.5$, and 1.5. Bottom: $\varepsilon = 0.01$ at $t = 0.5, 1.5$, and 3.

Snapshots of the phase variable at different times for $\varepsilon \in \{0.02, 0.01\}$ are shown in Figure 5.1, where the dashed circles, centered at $(0.5, 0.5)$ with radius $\frac{1}{2\sqrt{\pi}}$, indicate the interface at steady state. We observe that the initial square gradually transitions into a circle that closely matches the dashed reference, with the circular shape forming at an earlier stage for $\varepsilon = 0.02$ than for $\varepsilon = 0.01$. Figure 5.2 confirms mass conservation and energy dissipation of the numerical solutions, with the former ensured by the projection introduced in Section 5.4.

Coarsening dynamics We simulate the long-time coarsening dynamics of the phase separation process with different volume fractions in $\Omega = (0, 1)^2$ under homogeneous Neumann boundary conditions. The initial condition is a small random perturbation of the uniform state

$$u_0(x, y) = \overline{u} + 0.05\,\text{rand}(x, y),$$

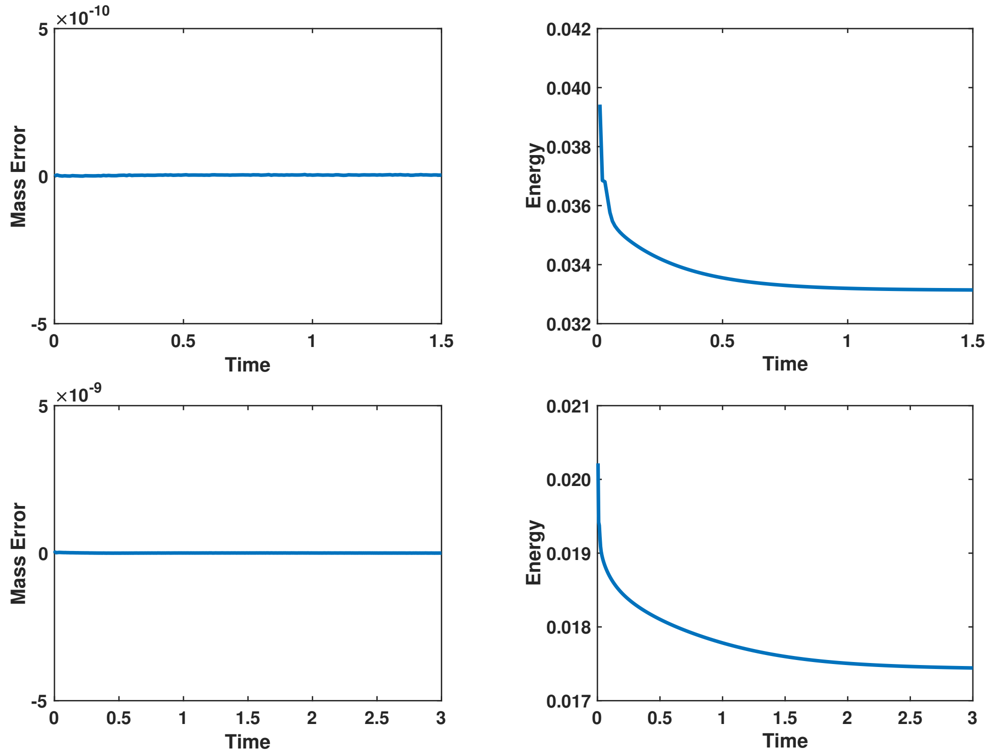


Figure 5.2: [Shape relaxation] Evolution of the absolute mass error and free energy by the GTransNet-BDF2 scheme with $\varepsilon = 0.02$ (top) and $\varepsilon = 0.01$ (bottom).

where rand(x, y) is uniformly distributed in $[-1, 1]$. We fix $\varepsilon = 0.02$, $D = 0.05$, $\varkappa = 2$, $\delta = 0.5$, $R = 0.75$, $K_{\text{int}} = 250^2$, $K_{\text{bd}} = 1000$, and consider $\overline{u} \in \{0, 0.5\}$. The simulation is run until $T = 40$ using the GTransNet-BDF2 scheme with $N_1 = 2000$, $N_2 = 1000$, $\gamma = 12$, $\Delta t = 1 \times 10^{-3}$ for $\overline{u} = 0$, and $N_1 = 3000$, $N_2 = 1500$, $\gamma = 14$, $\Delta t = 5 \times 10^{-3}$ for $\overline{u} = 0.5$.

Evolution of the phase variable at $t = 5, 20$, and 40 for $\overline{u} \in \{0, 0.5\}$ is reported in Figure 5.3. As expected, for $\overline{u} = 0$, the two phases separate into large interconnected regions, whereas for $\overline{u} = 0.5$, the minority phase forms isolated circular droplets within the majority phase. Additionally, the discrete mass is conserved over time while the free energy decays monotonically, as shown in Figure 5.4.

Cahn-Hilliard equation on irregular domains A key advantage of the proposed GTransNet-BDF method is its mesh-free nature, as the hidden-layer neurons are generated from a single domain-covering ball, allowing the method to be applied directly to domains of arbitrary shape.

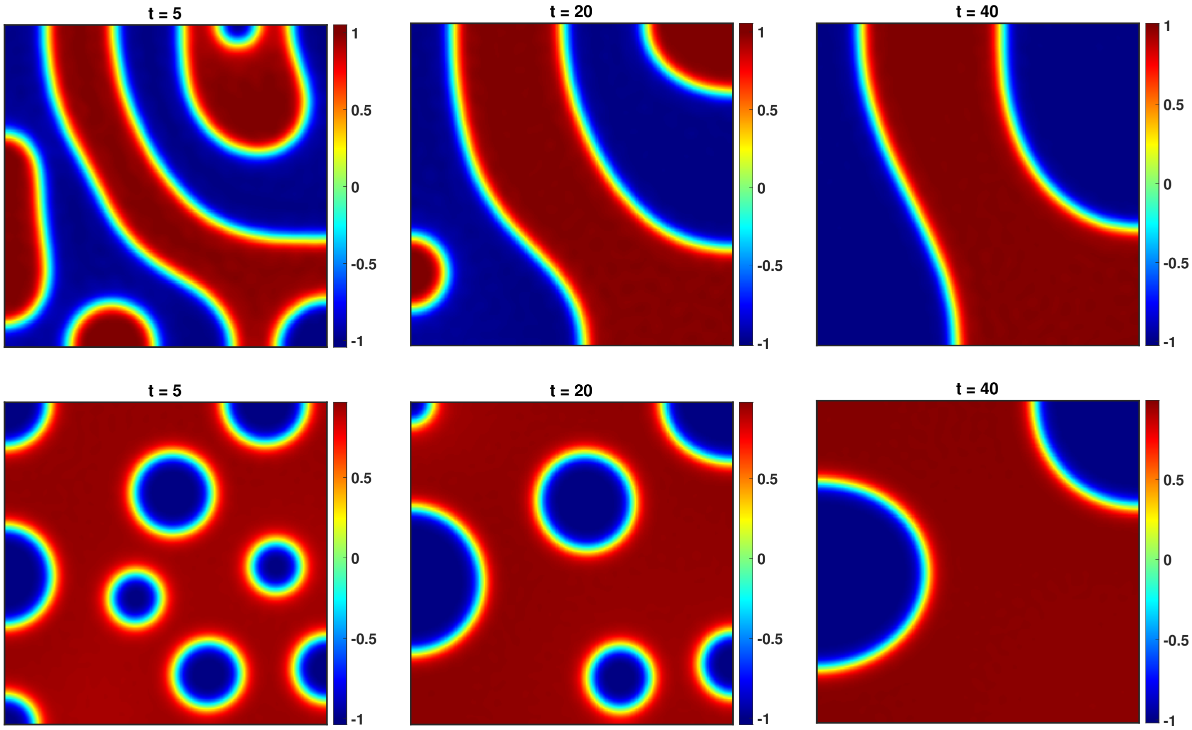


Figure 5.3: [2D coarsening dynamics] Snapshots of the phase variable by the GTransNet-BDF2 scheme with $\overline{u} = 0$ (top) and $\overline{u} = 0.5$ (bottom) at $t = 5, 20$, and 40.

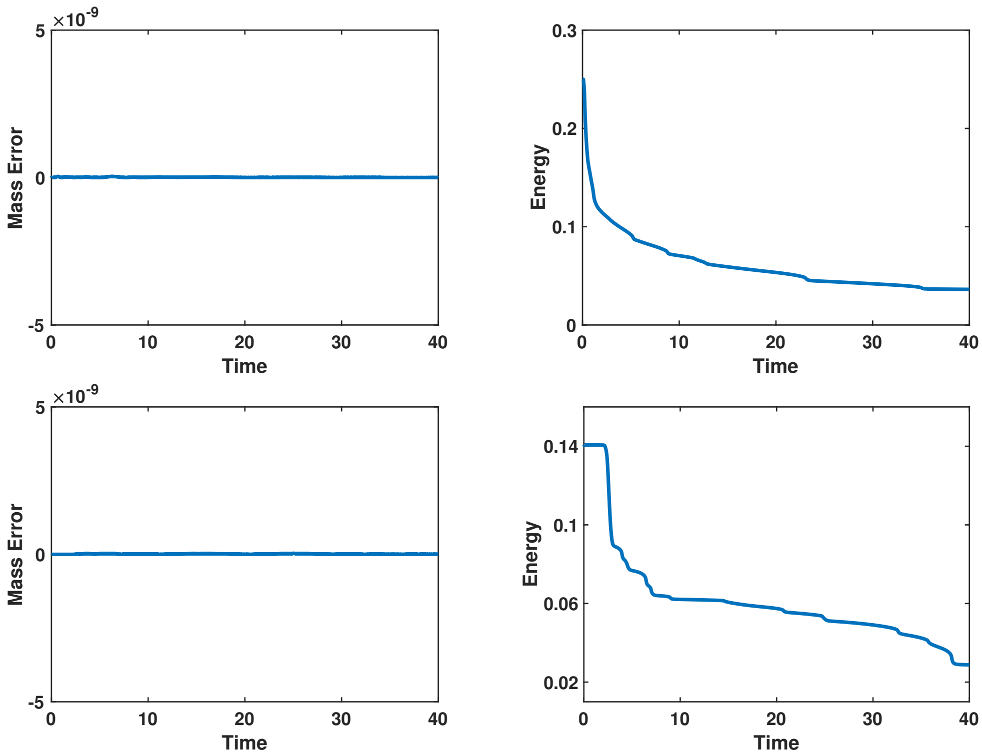


Figure 5.4: [2D coarsening dynamics] Evolution of the absolute mass error and free energy by the GTransNet-BDF2 scheme with $\overline{u} = 0$ (top) and $\overline{u} = 0.5$ (bottom).

To highlight this flexibility, we simulate the coarsening dynamics governed by the CH equation on three computational domains – a square, a disk, and an amoeba-shaped region with non-convex boundary – under homogeneous Neumann boundary conditions. We use the double-well potential $F(u) = \frac{1}{4}u^2(u-1)^2$ with wells at $u = 0$ and $u = 1$ and consider the following initial condition for all three domains [41, 42, 18]:

$$u_0(x, y) = 0.5 + 0.17\cos(\pi x)\cos(2\pi y) + 0.2\cos(3\pi x)\cos(\pi y).$$

We fix the common parameters $\varepsilon = 0.01$, $D = 0.25$, $\kappa = 2$, $\delta = 0.5$, $N_1 = 2000$, and $N_2 = 1000$. Other domain-specific parameters are listed below.

- **Square domain:** $\Omega = (0,1)^2$, $R = 0.75$, $\gamma = 12$, $K_{\text{int}} = 200^2$, $K_{\text{bd}} = 800$, $\Delta t = 2 \times 10^{-3}$.
- **Circular domain:** $\Omega = \{(x,y) : (x-0.5)^2 + (y-0.5)^2 < 0.25\}$, $R = 0.6$, $\gamma = 12$, $K_{\text{int}} = 31428$, $K_{\text{bd}} = 800$, $\Delta t = 2 \times 10^{-3}$.
- **Amoeba-shaped domain** with boundary

$$\partial\Omega = \left\{(x,y) : x = \frac{1}{5}\rho(\theta)\cos(\theta) + \frac{2}{5},\ y = \frac{1}{5}\rho(\theta)\sin(\theta) + \frac{2}{5},\ 0 \le \theta < 2\pi\right\},$$
$$\rho(\theta) = e^{\sin(\theta)}\sin^2(2\theta) + e^{\cos(\theta)}\cos^2(2\theta),$$

  and $\mathbf{x}_c = (0.53, 0.48)$, $R = 0.65$, $\gamma = 13.76$, $K_{\text{int}} = 31259$, $K_{\text{bd}} = 800$, $\Delta t = 5 \times 10^{-4}$.

Figure 5.5 illustrates the phase variable at $t = 0.3$, 0.7, and 1.5. In all three cases, the GTransNet-BDF2 scheme captures the early-stage spinodal decomposition and subsequent coarsening, producing smooth and well-resolved interfaces that conform to the domain geometry. The results for the square and circular domains are in strong agreement with those of the radial basis function (RBF) methods in [41, 42]. For the amoeba-shaped domain, the GTransNet-BDF2 solution remains much closer to the physical range $[0,1]$ than the one reported in [18]. Figure 5.6 further confirms mass conservation and energy dissipation of the numerical solutions across all three domains.

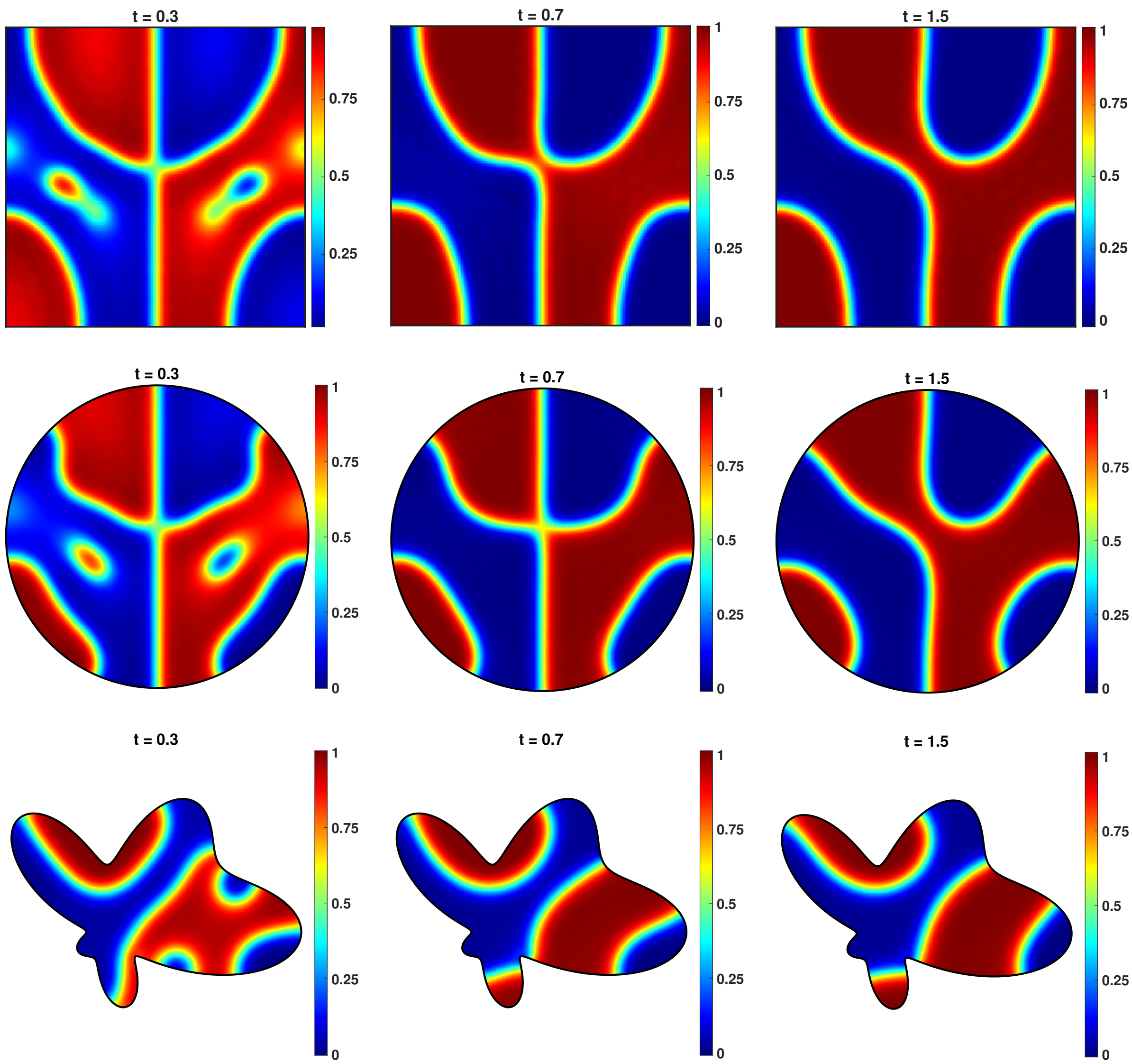


Figure 5.5: [Regular and irregular domains] Snapshots of the phase variable by the GTransNet-BDF2 scheme at $t = 0.3, 0.7$, and 1.5 for the square, circular, and amoeba-shaped domains (top to bottom).

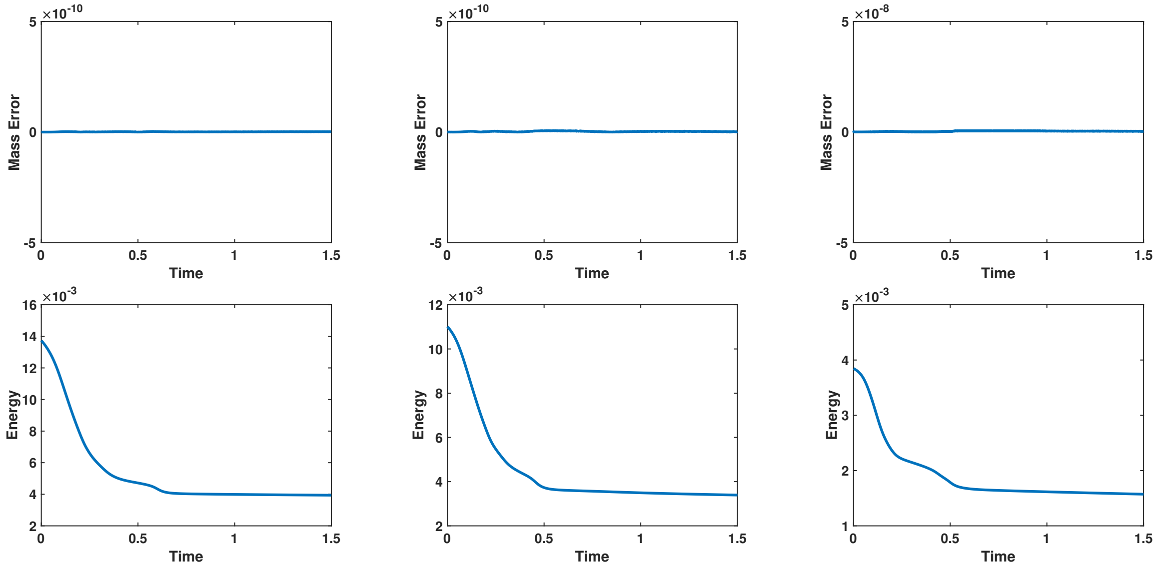


Figure 5.6: [Regular and irregular domains] Evolution of the absolute mass error (top) and free energy (bottom) by the GTransNet-BDF2 scheme for the square, circular, and amoeba-shaped domains (from left to right).

5.5.2 Three-dimensional Cahn-Hilliard equation

Convergence test We next examine the temporal convergence of the GTransNet-BDF2 scheme in $\Omega = (0,1)^3$ under homogeneous Neumann boundary conditions, using a manufactured solution

$$u(x, y, z, t) = \cos(\pi x)\cos(\pi y)\cos(\pi z)\sin(t),$$

We set $T = 1, D = 0.1, \kappa = 0, \delta = 0.5, R = 0.9, K_{\text{int}} = 30^3, K_{\text{bd}} = 5402$, and $\varepsilon \in \{0.2, 0.1\}$, with the number of neurons in each hidden layer and the shape parameter listed in Table 5.3 for each ε. The test data consist of $8K_{\text{int}}$ uniformly distributed random points in Ω.

Table 5.3: [3D convergence test] Parameter settings for $\varepsilon \in \{0.2, 0.1\}$.

(a) $L = 2$ hidden layers

ε	N_1	N_2	γ
0.2	2500	1500	1
0.1	3000	1500	3

(b) $L = 3$ hidden layers

ε	N_1	N_2	N_3	γ
0.2	1500	1500	1500	1
0.1	1500	1500	1500	3

Table 5.4 reports the L^∞ errors of the phase variable at the final time by the GTransNet-

BDF2 scheme with $L = 2$ and $L = 3$ hidden layers for $\varepsilon \in \{0.2, 0.1\}$. As in the two-dimensional case, we observe second-order convergence of the scheme, with comparable errors for $L = 2$ and $L = 3$.

Table 5.4: [3D convergence test] L^∞ errors of the phase variable at the final time for $\varepsilon \in \{0.2, 0.1\}$.

(a) $\varepsilon = 0.2$

Δt	GTransNet-BDF2 $L = 2$ hidden layers	GTransNet-BDF2 $L = 3$ hidden layers
$1/10$	8.58e-03	8.58e-03
$1/20$	1.99e-03 [2.11]	1.99e-03 [2.11]
$1/40$	4.90e-04 [2.02]	4.89e-04 [2.02]
$1/80$	1.23e-04 [1.99]	1.22e-04 [2.00]
$1/160$	3.09e-05 [1.99]	3.11e-05 [1.97]
$1/320$	7.48e-06 [2.05]	8.47e-06 [1.88]

(b) $\varepsilon = 0.1$

Δt	GTransNet-BDF2 $L = 2$ hidden layers	GTransNet-BDF2 $L = 3$ hidden layers
$1/10$	8.19e-02	8.19e-02
$1/20$	2.33e-02 [1.81]	2.33e-02 [1.81]
$1/40$	6.18e-03 [1.91]	6.18e-03 [1.91]
$1/80$	1.58e-03 [1.97]	1.58e-03 [1.97]
$1/160$	3.98e-04 [1.99]	3.97e-04 [1.99]
$1/320$	1.05e-04 [1.92]	9.66e-05 [2.04]

Coarsening dynamics We proceed to consider the three-dimensional coarsening dynamics in $\Omega = (0, 1)^3$ under homogeneous Neumann boundary conditions, a natural extension of the two-dimensional case. The initial condition is again a small random perturbation of a uniform state

$$u_0(x, y, z) = \overline{u} + 0.05\,\mathrm{rand}(x, y, z),$$

with $\mathrm{rand}(x, y, z)$ uniformly distributed in $[-1, 1]$. We take $\varepsilon = 0.04$, $\overline{u} = 0$, $D = 0.05$, $\varkappa = 2$, $\delta = 0.5$, $R = 0.9$, $K_{\mathrm{int}} = 50^3$, and $K_{\mathrm{bd}} = 15002$. The simulation is run until $T = 20$ using the GTransNet-BDF2 scheme with $N_1 = 4000$, $N_2 = 2000$, $\gamma = 7$, and $\Delta t = 5 \times 10^{-3}$.

Isosurface snapshots of the phase variable (the zero level set $\{u = 0\}$) at different times by the GTransNet-BDF2 scheme are shown in Figure 5.7. Starting from the random initial configuration, the interface coarsens over time, with the fine, highly connected structure at $t = 2$ progressively simplifying into a smoother surface with fewer features by $t = 20$. Figure 5.8 reports the corresponding evolution of the absolute mass error and free energy, confirming the preservation of the two intrinsic properties.

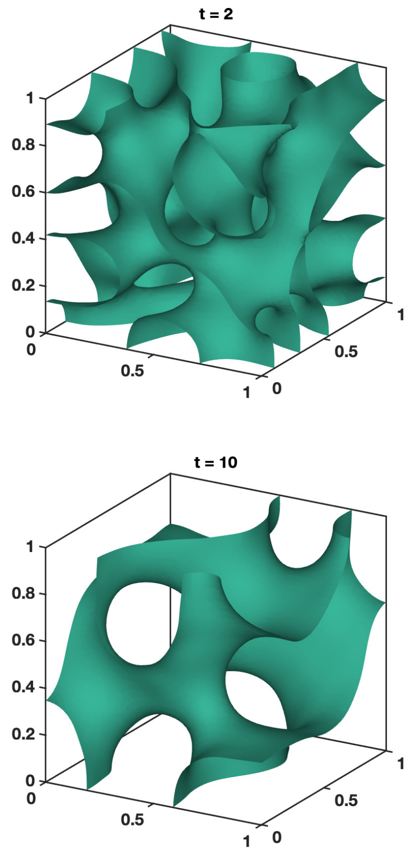


Figure 5.7: [3D coarsening dynamics] Zero-isosurfaces of the phase variable by the GTransNet-BDF2 scheme at $t = 2, 5, 10$, and 20.

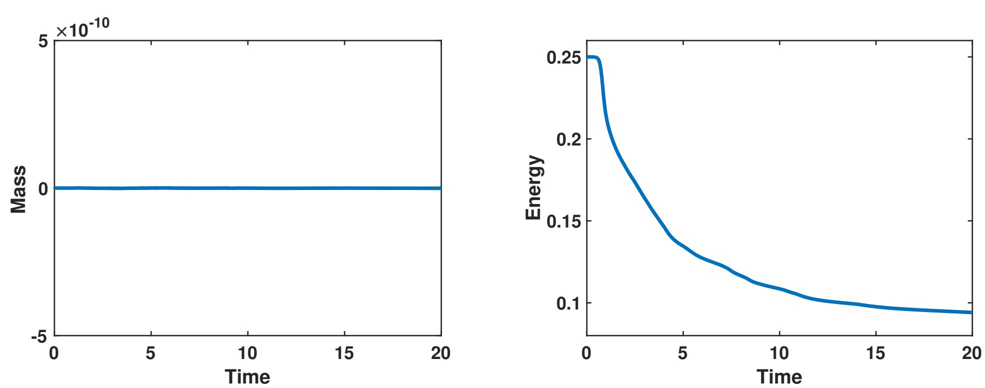


Figure 5.8: [3D coarsening dynamics] Evolution of the absolute mass error and free energy by the GTransNet-BDF2 scheme.

5.5.3 Cahn-Hilliard equation with degenerate mobility

Finally, we extend the proposed GTransNet-BDF framework to the CH equation with variable mobility $M(u)$, for which the corresponding first- and second-order schemes are derived in Appendix C. We consider the symmetric degenerate mobility $M(u) = |1 - u^2|$, which vanishes at the pure phases $u = \pm 1$ and attains its maximum $\Gamma = \max_{u \in [-1,1]} M(u) = 1$ at $u = 0$. The degeneracy suppresses diffusion in the bulk regions and confines mass transport to the thin interfacial layers, so that coarsening proceeds mainly through surface diffusion and the late-stage dynamics are considerably slower than in the constant-mobility case (i.e., $M(u) \equiv 1$).

We solve the CH equation in $\Omega = (0,1)^2$ under homogeneous Neumann boundary conditions, with the following initial data [103]:

$$u_0(x, y) = \overline{u} + 0.2\,\mathrm{rand}(x, y),$$

where $\mathrm{rand}(x, y)$ is uniformly distributed in $[-1, 1]$. We take $\overline{u} = 0.5$, $\varepsilon = 0.02$, $\kappa = 2$, $\delta = 0.5$, $R = 0.75$, $K_{\mathrm{int}} = 250^2$, and $K_{\mathrm{bd}} = 1000$, and run the simulation until $T = 5$ using the GTransNet-BDF2 scheme with $N_1 = 4000$, $N_2 = 2000$, $\gamma = 16$, and $\Delta t = 5 \times 10^{-4}$.

The phase variable at $t = 0.2, 0.5, 1, 2.5$, and 5 is plotted in Figure 5.9. The off-critical mixture ($\overline{u} = 0.5$) rapidly separates into circular droplets of the minority phase, which subsequently coarsen over time. Figure 5.10 verifies mass conservation and energy dissipation of the numerical solution, demonstrating that the framework extends robustly to the degenerate-mobility regime.

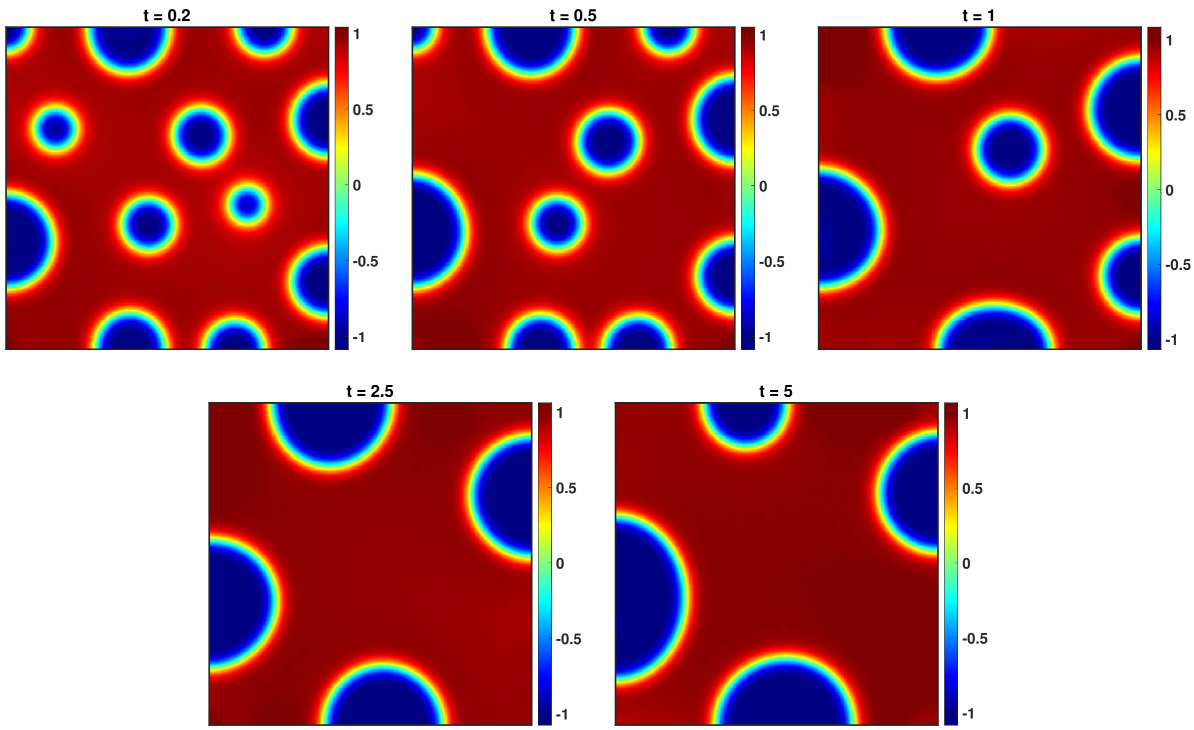


Figure 5.9: [Degenerate mobility] Snapshots of the phase variable by the GTransNet-BDF2 scheme with $\overline{u} = 0.5$ and $M(u) = |1 - u^2|$ at $t = 0.2, 0.5, 1, 2.5$, and 5.

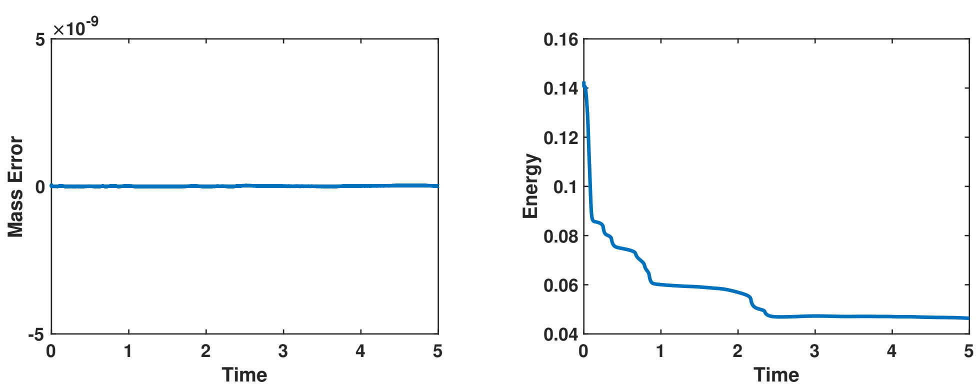


Figure 5.10: [Degenerate mobility] Evolution of the absolute mass error and free energy by the GTransNet-BDF2 scheme with $\overline{u} = 0.5$ and $M(u) = |1 - u^2|$.

Chapter 6

Conclusions and Future Work

In this dissertation, we have developed and analyzed four novel classes of efficient and structure-preserving time-stepping methods for nonlinear evolution equations and multiphysics systems: low regularity integrators for phase field models, noniterative localized exponential time differencing methods for hyperbolic conservation laws, dynamically regularized Lagrange multiplier methods for incompressible fluid flows, and generalized transferable neural networks for the Cahn-Hilliard equation.

In Chapter 2, we constructed first- and second-order LRI schemes for the Allen-Cahn and conservative Allen-Cahn equations by iteratively applying Duhamel's formula. The proposed schemes were shown to preserve the maximum bound principle and energy stability under mild time step size constraints. Optimal temporal error estimates were derived under the minimal assumption that the exact solution is only continuous in time for the classical Allen-Cahn equation and Lipschitz continuous for the conservative variant. We also carried out the fully discrete error analysis of the first-order LRI schemes by utilizing fundamental properties of the discrete system and Duhamel's principle, establishing optimal error estimates in both time and space under the same low regularity assumption. For the conservative Allen-Cahn equation, the LRI schemes were further shown to unconditionally preserve mass. Numerical experiments confirmed the theoretical findings and demonstrated the outperformance of LRI schemes compared to classical ETD schemes, particularly when the interfacial parameter approaches zero.

In Chapter 3, we developed global and localized ETDRK2 schemes for scalar hyperbolic conservation laws with DG spatial discretization. The proposed schemes are fully explicit and require no nonlinear or linear solver while allowing much larger time step sizes than standard SSP-RK methods. To reduce the cost of computing products of matrix exponentials and vectors, a noniterative localized ETDRK2 scheme was proposed via nonoverlapping domain decomposition with interface conditions defined by the numerical flux. The localized scheme requires no overlap and no iterations between subdomains while preserving the accuracy and stability of the global method. Rigorous semi-discrete temporal error analysis was carried out for both schemes, and mass conservation was proved under periodic boundary conditions.

In Chapter 4, we proposed the DRLM method for the incompressible Navier-Stokes equations. By introducing a dynamic equation involving the kinetic energy, a Lagrange multiplier, and a regularization parameter, the NS equations were reformulated into an equivalent system. First- and second-order DRLM schemes based on the BDF method with explicit treatment of the nonlinear convection term were derived and shown to unconditionally satisfy the energy dissipation law with respect to the original variables. Fully discrete energy stability was also established with the Marker-and-Cell spatial discretization. Unlike existing Lagrange multiplier methods, the regularization parameter enables the use of large time step sizes while ensuring the unique solvability of the Lagrange multiplier. We then established rigorous temporal error estimates for the first-order DRLM scheme, deriving optimal convergence rates for the velocity, Lagrange multiplier, and pressure. All error constants exhibit the desired dependence on the regularization parameter, and the analysis is valid in both two and three dimensions. Finally, we extended the DRLM framework to the coupled Cahn-Hilliard-Navier-Stokes system, where the proposed schemes are fully decoupled, mass-conserving, unconditionally energy stable, and require no stabilization term. Numerical experiments in 2D and 3D verified the theoretical findings and demonstrated the accuracy and robustness of the proposed methods.

In Chapter 5, we proposed the first- and second-order GTransNet-BDF schemes coupled with a mass-conserving projection for solving the Cahn-Hilliard equation. The method combines the GTransNet architecture for spatial approximation with stabilized BDF for time integration; the first hidden layer is predetermined as in the original TransNet with a symmetric bias distribution, followed by a variance-controlled weight sampling strategy for the subsequent hidden layers. The schemes preserve two important properties of the Cahn-Hilliard equation at the time-discrete level, namely mass conservation and energy stability. By employing a collocation-based method, a least-squares system for the unknown output-layer weights is solved at each time step, which can be implemented efficiently using a precomputed QR factorization. To mitigate the effect of least-squares error, we introduced a mass-conserving projection that corrects the least-squares solution along a single direction with minimal perturbation so that the discrete mass is conserved exactly. Numerical experiments in two and three dimensions confirmed the theoretical findings and illustrated the performance of the GTransNet-BDF methods for the Cahn-Hilliard equation with both constant and degenerate mobility.

Several directions for future research naturally arise from this dissertation:

1. Regarding the low regularity integrators, we plan to develop new LRI schemes for the Cahn-Hilliard equation. The current LRI formulations for the Allen-Cahn equation are not directly applicable to the Cahn-Hilliard equation due to the presence of fourth-order spatial derivatives. Constructing LRI schemes that preserve the energy dissipation law and mass conservation for the Cahn-Hilliard equation, together with rigorous error analysis under low regularity, is an interesting and challenging problem. In addition, extension of the fully discrete error analysis to the second-order LRI scheme for the Allen-Cahn equation is still an open question. Further investigations include the energy stability of LRIs for the conservative Allen-Cahn equation on general domains with the homogeneous Neumann boundary conditions and the development of high-order LRIs for the Navier-Stokes equations [120] and the coupled Navier-Stokes/Allen-Cahn system [95].

2. For the localized ETD methods, we are currently extending the global and localized ETDRK schemes to the shallow water equations. Compared to the scalar conservation laws studied in Chapter 3, solving the shallow water equations is more involved as one has to employ the well-balanced numerical flux [190] to preserve the still water steady-state solution, leading to more complicated linear operators at each time step. Preliminary results in one dimension indicate that the ETDRK2 methods allow for a much larger CFL number than SSP-RK2 while still achieving comparable accuracy. Other future directions include

nonlinear stability analysis of ETDRK schemes, their convergence to the entropy solution, and extension to systems of hyperbolic conservation laws in higher spatial dimensions.

3. For the DRLM methods, error analysis of the second-order DRLM scheme for both the incompressible NS equations and the coupled CH-NS system, in both semi-discrete and fully discrete settings, remains highly challenging and will be a subject of our future studies. Unlike the first-order scheme where the Lagrange multiplier is uniquely determined from a quadratic equation, the second-order scheme leads to a more complicated algebraic equation for the Lagrange multiplier whose unique positive solution is not always guaranteed. Extension of the current approach to higher-order DRLM schemes using the BDF method will also be investigated; we refer to [40] for a discussion of high-order BDF schemes with time filters and to [99] for the development of energy stable SAV-BDF schemes up to fifth-order accuracy in time. Other research directions include error analysis of the DRLM methods for the incompressible magnetohydrodynamic (MHD) equations [181] and the extension of the DRLM approach to more complex interfacial flow problems, such as the phase field elastic bending energy model [195] and the binary fluid-surfactant phase field model [196]. Although it has been observed numerically that large regularization parameters guarantee the uniqueness and accuracy of the Lagrange multipliers, providing theoretical lower bounds for these parameters is still challenging and requires further investigation.

4. For the GTransNet-BDF methods, two limitations of the current framework deserve further investigation. First, the collocation-based implementation requires a sufficiently large number of collocation points, making the QR factorization and matrix-vector products in the least-squares solver computationally expensive. The development of GTransNet-BDF variants that retain comparable accuracy with significantly fewer collocation points would therefore be highly desirable. Second, an explicit formula for the optimal shape parameter γ in the GTransNet activation is not yet available; numerical experiments indicate

that it depends on the number of neurons in the first hidden layer N_1, the interfacial thickness ε, and is sensitive to the time step size Δt, especially when $\Delta t \ll 1$. Establishing a precise relationship between γ and these parameters is therefore an important direction for future work. We also plan to extend the proposed framework to other complicated time-dependent PDEs, such as the incompressible Navier-Stokes equations [48] and the coupled Cahn-Hilliard-Navier-Stokes system [49]. Other future directions include the combination with domain decomposition techniques [133] for improved computational efficiency on large-scale problems, and rigorous convergence analysis of the GTransNet-BDF schemes.

References

[1] H. Abels, On a diffuse interface model for two-phase flows of viscous, incompressible fluids with matched densities. Arch. Rational Mech. Anal. **194**, 463–506 (2009).

[2] G. Akrivis, B. Li, D. Li, Energy-decaying extrapolated RK-SAV methods for the Allen-Cahn and Cahn-Hilliard equations. SIAM J. Sci. Comput. **41**, A3703–A3727 (2019).

[3] S. Albensoeder and H.C. Kuhlmann, Accurate three-dimensional lid-driven cavity flow. J. Comput. Phys. **206**(2), 536–558 (2005).

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

[5] D.M. Anderson, G.B. McFadden, A.A. Wheeler, Diffuse-interface methods in fluid mechanics. Annu. Rev. Fluid Mech. **30**, 139–165 (1998).

[6] P. Aursand, S. Evje, T. Flåtten, K.E.T. Giljarhus, S.T. Munkejord, An exponential time-differencing method for monotonic relaxation systems. Appl. Numer. Math. **80**, 1–21 (2014).

[7] C. Bardos, A.Y. LeRoux, J.C. Nedelec, First order quasilinear equations with boundary conditions. Comm. Partial Differential Equations **4**, 1017–1034 (1979).

[8] M. Benzi, G.H. Golub, J. Liesen, Numerical solution of saddle point problems. Acta Numer. **14**, 1–137 (2005).

[9] T. Biben, K. Kassner, C. Misbah, Phase-field approach to three-dimensional vesicle dynamics. Phys. Rev. E **72**(4), 041921 (2005).

[10] W.J. Boettinger, J.A. Warren, C. Beckermann, A. Karma, Phase-field simulation of solidification. Annu. Rev. Mater. Res. **32**, 163–194 (2002).

[11] B. Bourdin, Numerical implementation of the variational formulation for quasi-static brittle fracture. Interfaces Free Boundaries, 411–430 (2007).

[12] B. Bourdin, G.A. Francfort, J.J. Marigo, The variational approach to fracture. J. Elast. **91**, 5–148 (2008).

[13] F. Boyer and P. Fabrie, Mathematical Tools for the Study of the Incompressible Navier-Stokes Equations and Related Models. Applied Mathematical Sciences, vol. 183, Springer, 2013.

[14] J.W. Cahn, On spinodal decomposition. Acta Metall. **9**, 795–801 (1961).

[15] Y. Cai and J. Shen, Error estimates for a fully discretized scheme to a Cahn-Hilliard phase-field model for two-phase incompressible flows. Math. Comp. **87**, 2057–2090 (2018).

[16] W. Cai, W. Sun, J. Wang, Z. Yang, Optimal error estimates of unconditionally stable finite element schemes for the Cahn–Hilliard–Navier–Stokes system. SIAM J. Numer. Anal. **61**, 1218–1245 (2023).

[17] S. Calandrini, K. Pieper, M. Gunzburger, Numerical analyses of exponential time-differencing schemes for the solution of atmospheric models. Q. J. R. Meteorol. Soc. **147**(736), 1477–1496 (2021).

[18] D. Cao, X. Li, and H. Zhu, A polynomial-augmented RBF collocation method using fictitious centres for solving the Cahn-Hilliard equation. Eng. Anal. Bound. Elem. **137**, 41–55 (2022).

[19] L. Cattabriga, Su un problema al contorno relativo al sistema di equazioni di Stokes. Rend. Sem. Mat. Univ. Padova **31**, 308–340 (1961).

[20] J. Certaine, The solution of ordinary differential equations with large time constants. Mathematical Methods for Digital Computers, Wiley, 128–132 (1960).

[21] J. Chen, X. Chi, W. E, and Z. Yang, Bridging traditional and machine learning-based algorithms for solving PDEs: The random feature method. J. Mach. Learn. **1**, 268–298 (2022).

[22] W. Chen, J. Jing, Q. Liu, C. Wang, X. Wang, Convergence analysis of a second order numerical scheme for the Flory-Huggins-Cahn-Hilliard-Navier-Stokes system. J. Comput. Appl. Math. **450**, 115981 (2024).

[23] N. Chen, S. Lucarini, R. Ma, A. Chen, and C. Cui, PF-PINNs: Physics-informed neural networks for solving coupled Allen-Cahn and Cahn-Hilliard phase field equations. J. Comput. Phys. **529**, 113843 (2025).

[24] T. Cheng, L. Ju, Z. Qiao, and X. Zhang, Generalized transferable neural networks for steady-state partial differential equations. Preprint, arXiv:2604.03020 (2026).

[25] Q. Cheng, C. Liu, J. Shen, A new Lagrange multiplier approach for gradient flows. Comput. Methods Appl. Mech. Engrg. **367**, 113070 (2020).

[26] Q. Cheng, C. Wang, Error estimate of a second order accurate scalar auxiliary variable (SAV) numerical method for the epitaxial thin film equation. Adv. Appl. Math. Mech. **13**, 1318–1354 (2021).

[27] W. Chen, J. Jing, Q. Liu, C. Wang, X. Wang, A second order numerical scheme of the Cahn-Hilliard-Navier-Stokes system with Flory-Huggins potential. Commun. Comput. Phys. **35**(3), 633–661 (2024).

[28] W. Chen, J. Jing, C. Wang, X. Wang, A positivity preserving, energy stable finite difference scheme for the Flory-Huggins-Cahn-Hilliard-Navier-Stokes system. J. Sci. Comput. **92**, 31 (2022).

[29] W. Chen, Y. Liu, C. Wang, S. Wise, Convergence analysis of a fully discrete finite difference scheme for the Cahn-Hilliard-Hele-Shaw equation. Math. Comp. **85**(301), 2231–2257 (2016).

[30] Y. Chen and J. Shen, Efficient, adaptive energy stable schemes for the incompressible Cahn–Hilliard Navier–Stokes phase-field models. J. Comput. Phys. **308**, 40–56 (2016).

[31] B. Cockburn, Devising discontinuous Galerkin methods for non-linear hyperbolic conservation laws. J. Comput. Appl. Math. **128**(1-2), 187–204 (2001).

[32] B. Cockburn, S. Hou, C.-W. Shu, The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case. Math. Comp. **54**(190), 545–581 (1990).

[33] B. Cockburn, S. Hou, C.-W. Shu, The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws V: Multidimensional systems. J. Comput. Phys. **141**, 199–224 (1998).

[34] B. Cockburn, G. Karniadakis, C.-W. Shu, The development of discontinuous Galerkin methods, in Discontinuous Galerkin Methods: Theory, Computation and Applications, Lecture Notes in Computational Science and Engineering, vol. 11, Springer, 2000, pp. 3–50.

[35] B. Cockburn, S.Y. Lin, C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. III: One-dimensional systems. J. Comput. Phys. **84**(1), 90–113 (1989).

[36] B. Cockburn, C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. Math. Comp. **52**(186), 411–435 (1989).

[37] B. Cockburn, C.-W. Shu, The Runge-Kutta local projection-discontinuous-Galerkin finite element method for scalar conservation laws. ESAIM: Math. Model. Numer. Anal. **25**(3), 337–361 (1991).

[38] S.M. Cox, P.C. Matthews, Exponential time differencing for stiff systems. J. Comput. Phys. **176**(2), 430–455 (2002).

[39] S. Dai, S. Wu, Explicit large time stepping with third-order exponential time differencing scheme for hypersonic chemical non-equilibrium flow. J. Appl. Math. Phys. **6**(1), 174–182 (2018).

[40] V. DeCaria, A. Guzel, W. Layton, Y. Li, A variable stepsize, variable order family of low complexity. SIAM J. Sci. Comput. **43**(3), A2130–A2160 (2021).

[41] M. Dehghan and V. Mohammadi, The numerical solution of Cahn-Hilliard (CH) equation in one, two and three-dimensions via globally radial basis functions (GRBFs) and RBFs-differential quadrature (RBFs-DQ) methods. Eng. Anal. Bound. Elem. **51**, 74–100 (2015).

[42] M. Dehghan and M. Abbaszadeh, The meshless local collocation method for solving multi-dimensional Cahn-Hilliard, Swift-Hohenberg and phase field crystal equations. Eng. Anal. Bound. Elem. **78**, 49–64 (2017).

[43] E. de Sturler, J. Liesen, Block-diagonal and constraint preconditioners for nonsymmetric indefinite linear systems. Part I: Theory. SIAM J. Sci. Comput. **26**(5), 1598–1619 (2005).

[44] A.E. Diegel, C. Wang, X. Wang, S.M. Wise, Convergence analysis and error estimates for a second order accurate finite element method for the Cahn–Hilliard–Navier–Stokes system. Numer. Math. **137**(3), 495–534 (2017).

[45] C.-K. Doan, T.-T.-P. Hoang, L. Ju, K. Schratz, Low regularity integrators for semilinear parabolic equations with maximum bound principles. BIT Numer. Math. **63**(1), 2 (2023).

[46] C.-K. Doan, T.-T.-P. Hoang, L. Ju, Fully discrete error analysis of first-order low regularity integrators for the Allen-Cahn equation. Numer. Methods Partial Differ. Equ. **39**(5), 3594–3608 (2023).

[47] C.-K. Doan, T.-T.-P. Hoang, L. Ju, Low regularity integrators for the conservative Allen-Cahn equation with a nonlocal constraint. J. Sci. Comput. **101**, 57 (2024).

[48] C.-K. Doan, T.-T.-P. Hoang, L. Ju, R. Lan, Dynamically regularized Lagrange multiplier schemes with energy dissipation for the incompressible Navier-Stokes equations. J. Comput. Phys. **521**, 113550 (2025).

[49] C.-K. Doan, T.-T.-P. Hoang, L. Ju, R. Lan, Dynamically regularized Lagrange multiplier method for the Cahn-Hilliard-Navier-Stokes system. Int. J. Numer. Methods Engrg. **126**(13), e70074 (2025).

[50] C.-K. Doan, T.-T.-P. Hoang, L. Ju, R. Lan, Convergence analysis of the dynamically regularized Lagrange multiplier method for the incompressible Navier-Stokes equations. IMA J. Numer. Anal., drag016 (2026).

[51] C.-K. Doan, T.-T.-P. Hoang, L. Ju, S. Wang, Structure-preserving generalized transferable neural networks for the Cahn-Hilliard equation. Submitted (2026).

[52] C.-K. Doan, P.-T. Huynh, T.-T.-P. Hoang, Noniterative localized exponential time differencing methods for hyperbolic conservation laws. Adv. Comput. Math. **51**(3), 26 (2025).

[53] S. Dong and Z. Li, Local extreme learning machines and domain decomposition for solving linear and nonlinear partial differential equations. Comput. Methods Appl. Mech. Engrg. **387**, 114129 (2021).

[54] Q. Du, X. Feng, The phase field method for geometric moving interfaces and their numerical approximations. Handb. Numer. Anal. **21**, 425–508 (2020).

[55] Q. Du, L. Ju, X. Li, Z. Qiao, Maximum principle preserving exponential time differencing schemes for the nonlocal Allen-Cahn equation. SIAM J. Numer. Anal. **57**(2), 875–898 (2019).

[56] Q. Du, L. Ju, X. Li, Z. Qiao, Maximum bound principles for a class of semilinear parabolic equations and exponential time-differencing schemes. SIAM Rev. **63**(2), 317–359 (2021).

[57] X. Feng, Fully discrete finite element approximations of the Navier-Stokes-Cahn-Hilliard diffuse interface model for two-phase fluid flows. SIAM J. Numer. Anal. **44**, 1049–1072 (2006).

[58] X. Feng, A. Prohl, Numerical analysis of the Allen-Cahn equation and approximation for mean curvature flows. Numer. Math. **94**(1), 33–65 (2003).

[59] X. Feng, T. Tang, J. Yang, Stabilized Crank-Nicolson/Adams-Bashforth schemes for phase field models. East Asian J. Appl. Math. **3**, 59–80 (2013).

[60] G.J. Fix, Phase field methods for free boundary problems. Free Boundary Problems: Theory and Applications, Pitman, 580–589 (1982).

[61] R. Folch, J. Casademunt, A. Hernandez-Machado, L. Ramirez-Piscina, Phase-field model for Hele-Shaw flows with arbitrary viscosity contrast. II. Numerical study. Phys. Rev. E **60**(2), 1734–1740 (1999).

[62] Z. Fu and J. Yang, Energy-decreasing exponential time differencing Runge–Kutta methods for phase-field models. J. Comput. Phys. **454**, 110943 (2022).

[63] B. García-Archilla and J. Novo, Robust error bounds for the Navier–Stokes equations using implicit-explicit second-order BDF method with variable steps. IMA J. Numer. Anal. **43**(5), 2892–2933 (2023).

[64] S. Gaudreault, G. Rainwater, M. Tokman, KIOPS: A fast adaptive Krylov subspace solver for exponential integrators. J. Comput. Phys. **372**, 236–255 (2018).

[65] U. Ghia, K.N. Ghia, C.T. Shin, High-Re solutions for incompressible flow using the Navier-Stokes equations and a multigrid method. J. Comput. Phys. **48**(3), 387–411 (1982).

[66] A. Giorgini, A. Miranville, R. Temam, Uniqueness and regularity for the Navier-Stokes-Cahn-Hilliard system. SIAM J. Math. Anal. **51**, 2535–2574 (2019).

[67] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks. JMLR Proc. Track **9**, 249–256 (2010).

[68] S. Gottlieb, C.-W. Shu, Total variation diminishing Runge-Kutta schemes. Math. Comp. **67**(221), 73–85 (1998).

[69] J. Goulding, M. Ayazi, T. Shinar, C. Schroeder, Simplified conservative discretization of the Cahn-Hilliard-Navier-Stokes equations. J. Comput. Phys. **519**, 113382 (2024).

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

[71] J.L. Guermond, P. Minev, J. Shen, An overview of projection methods for incompressible flows. Comput. Methods Appl. Mech. Engrg. **195**(44-47), 6011–6045 (2006).

[72] J.L. Guermond, J. Shen, On the error estimates for the rotational pressure-correction projection methods. Math. Comp. **73**(248), 1719–1737 (2004).

[73] F. Guillén-González and G. Tierra, Structure preserving finite element schemes for the Navier-Stokes-Cahn-Hilliard system with degenerate mobility. Comput. Math. Appl. **172**, 181–201 (2024).

[74] G. Guj and F. Stella, A vorticity-velocity method for the numerical solution of 3D incompressible flows. J. Comput. Phys. **106**(2), 286–298 (1993).

[75] M.E. Gurtin, D. Polignone, J. Viñals, Two-phase binary fluids and immiscible fluids described by an order parameter. Math. Models Meth. Appl. Sci. **6**, 815–831 (1996).

[76] S. Ham, Y. Li, D. Jeong, C. Lee, S. Kwak, Y. Hwang, and J. Kim, An explicit adaptive finite difference method for the Cahn-Hilliard equation. J. Nonlinear Sci. **32**, 80 (2022).

[77] D. Han and X. Wang, A second order in time, uniquely solvable, unconditionally stable numerical scheme for Cahn–Hilliard–Navier–Stokes equation. J. Comput. Phys. **290**, 139–156 (2015).

[78] A. Harten, B. Engquist, S. Osher, S.R. Chakravarthy, Uniformly high order accurate essentially non-oscillatory schemes III. J. Comput. Phys. **71**, 231–303 (1987).

[79] Y. He, W. Sun, Stability and convergence of the Crank–Nicolson/Adams–Bashforth scheme for the time-dependent Navier–Stokes equations. SIAM J. Numer. Anal. **45**(2), 837–869 (2007).

[80] J.G. Heywood, R. Rannacher, Finite element approximation of the nonstationary Navier–Stokes problem. I. Regularity of solutions and second-order error estimates for spatial discretization. SIAM J. Numer. Anal. **19**(2), 275–311 (1982).

[81] J.G. Heywood, R. Rannacher, Finite-element approximation of the nonstationary Navier–Stokes problem. Part IV: Error analysis for second-order time discretization. SIAM J. Numer. Anal. **27**(2), 353–384 (1990).

[82] T.T.P. Hoang, L. Ju, W. Leng, Z. Wang, High order explicit local time stepping methods for hyperbolic conservation laws. Math. Comput. **89**(324), 1807–1842 (2020).

[83] T.T.P. Hoang, L. Ju, Z. Wang, Nonoverlapping localized exponential time differencing methods for diffusion problems. J. Sci. Comput. **82**, 1–27 (2020).

[84] M. Hochbruck, C. Lubich, On Krylov subspace approximations to the matrix exponential operator. SIAM J. Numer. Anal. **34**, 1911–1925 (1997).

[85] M. Hochbruck, C. Lubich, H. Selhofer, Exponential integrators for large systems of differential equations. SIAM J. Sci. Comput. **19**, 1552–1574 (1998).

[86] M. Hochbruck, A. Ostermann, Exponential integrators. Acta Numer. **19**, 209–286 (2010).

[87] M. Hochbruck, A. Ostermann, J. Schweitzer, Exponential Rosenbrock-type methods. SIAM J. Numer. Anal. **47**(1), 786–803 (2009).

[88] M. Hofmanova, K. Schratz, An exponential-type integrator for the KdV equation. Numer. Math. **136**, 1117–1137 (2017).

[89] T. Hou, H. Leng, Numerical analysis of a stabilized Crank-Nicolson/Adams-Bashforth finite difference scheme for Allen-Cahn equations. Appl. Math. Lett. **102**, 106150 (2020).

[90] T. Hou, T. Tang, J. Yang, Numerical analysis of fully discretized Crank-Nicolson scheme for fractional-in-space Allen-Cahn equations. J. Sci. Comput. **72**(3), 1214–1231 (2017).

[91] G. Huang, Q. Zhu, and C. Siew, Extreme learning machine: Theory and applications. Neurocomputing **70**, 489–501 (2006).

[92] C. Huang, X. Gao, Phase field modeling of hydrogen embrittlement. Int. J. Hydrog. Energy (2020).

[93] F. Huang and J. Shen, A class of IMEX schemes and their error analysis for the Navier-Stokes Cahn-Hilliard system. Ann. Math. Sci. Appl. **9**, 185–235 (2024).

[94] L. Isherwood, Z.J. Grant, S. Gottlieb, Strong stability preserving integrating factor Runge-Kutta methods. SIAM J. Numer. Anal. **56**(6), 3276–3307 (2018).

[95] K. Jiang, L. Ju, J. Li, X. Li, Unconditionally stable exponential time differencing schemes for the mass-conserving Allen–Cahn equation with nonlocal and local effects. Numer. Methods Partial Differ. Equ. **38**(6), 1636–1657 (2022).

[96] B.N. Jiang, T.L. Lin, L.A. Povinelli, Large-scale computation of incompressible viscous flow by least-squares finite element method. Comput. Methods Appl. Mech. Engrg. **114**(3-4), 213–231 (1994).

[97] G.S. Jiang, C.-W. Shu, Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**(1), 202–228 (1996).

[98] Y. Jiang, C.-W. Shu, M. Zhang, An alternative formulation of finite difference weighted ENO schemes with Lax–Wendroff time discretization for conservation laws. SIAM J. Sci. Comput. **35**(2), A1137–A1160 (2013).

[99] B. Ji, H.-L. Liao, A unified L^2 norm error analysis of SAV-BDF schemes for the incompressible Navier–Stokes equations. J. Sci. Comput. **100**(1), 5 (2024).

[100] L. Ju, Z. Wang, Exponential time differencing Gauge method for incompressible viscous flows. Commun. Comput. Phys. **22**(2), 517–541 (2017).

[101] L. Ju, X. Li, Z. Qiao, J. Yang, Maximum bound principle preserving integrating factor Runge-Kutta methods for semilinear parabolic equations. J. Comp. Phys. **439**, 110405 (2021).

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

[103] L. Ju, J. Zhang, Q. Du, Fast and accurate algorithms for simulating coarsening dynamics of Cahn-Hilliard equations. Comput. Mater. Sci. **108**, 272–282 (2015).

[104] L. Ju, J. Zhang, L. Zhu, Q. Du, Fast explicit integration factor methods for semilinear parabolic equations. J. Sci. Comput. **62**, 431–455 (2015).

[105] D. Kay and R. Welford, Efficient numerical solution of Cahn–Hilliard–Navier–Stokes fluids in 2D. SIAM J. Sci. Comput. **29**, 2241–2257 (2007).

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

[107] J. Kim, K. Kang, J. Lowengrub, Conservative multigrid methods for Cahn–Hilliard fluids. J. Comput. Phys. **193**, 511–543 (2004).

[108] H.C. Ku, R.S. Hirsh, T.D. Taylor, A pseudospectral method for solution of the three-dimensional incompressible Navier-Stokes equations. J. Comput. Phys. **70**(2), 439–462 (1987).

[109] J.S. Langer, Models of pattern formation in first-order phase transitions. Directions in Condensed Matter Physics, 165–186 (1986).

[110] R. Lan, L. Ju, M. Liu, Robust error analysis of stabilized EMAC-ESAV finite element schemes for the incompressible Navier-Stokes equations. Math. Comp. (2025).

[111] H.G. Lee, J. Kim, A comparison study of the Boussinesq and the variable density models on buoyancy-driven flows. J. Eng. Math. **75**, 15–27 (2012).

[112] J. Li, J. Li, and F. Tong, Maximum bound principle preserving and mass conservative projection method for the conservative Allen–Cahn equation. Appl. Math. Lett. **156**, 109151 (2024).

[113] S. Liang, J. Zhang, X.Z. Liu, X.D. Hu, W. Yuan, Domain decomposition based exponential time differencing method for fluid dynamics problems with smooth solutions. Comput. Fluids **194**, 104307 (2019).

[114] J. Li, L. Ju, Y. Cai, X. Feng, Unconditionally maximum bound principle preserving linear schemes for the conservative Allen–Cahn equation with nonlocal constraint. J. Sci. Comput. **87**(3), 1–32 (2021).

[115] X. Li, L. Ju, T.T.P. Hoang, Overlapping domain decomposition based exponential time differencing methods for semilinear parabolic equations. BIT Numer. Math. **61**, 1–36 (2021).

[116] Y. Li, H.G. Lee, D. Jeong, J. Kim, An unconditionally stable hybrid numerical method for solving the Allen-Cahn equation. Comput. Math. Appl. **60**, 1591–1606 (2010).

[117] J. Li, X. Li, L. Ju, X. Feng, Stabilized integrating factor Runge-Kutta method and unconditional preservation of maximum bound principle. SIAM J. Sci. Comput. **43**(3), A1780–A1802 (2021).

[118] X. Li, Z. Liu, J. Shen, N. Zheng, On a class of higher-order fully decoupled schemes for the Cahn–Hilliard–Navier–Stokes system. J. Sci. Comput. **103**, 27 (2025).

[119] S.J. Li, L.S. Luo, Z.J. Wang, L. Ju, An exponential time-integrator scheme for steady and unsteady inviscid flows. J. Comput. Phys. **365**, 206–225 (2018).

[120] B. Li, S. Ma, K. Schratz, A semi-implicit exponential low-regularity integrator for the Navier–Stokes equations. SIAM J. Numer. Anal. **60**(4), 2273–2292 (2022).

[121] L. Lin, Z. Yang, S. Dong, Numerical approximation of incompressible Navier-Stokes equations based on an auxiliary energy variable. J. Comput. Phys. **388**, 1–22 (2019).

[122] X. Li, H. Rui, Superconvergence of characteristics marker and cell scheme for the Navier–Stokes equations on nonuniform grids. SIAM J. Numer. Anal. **56**(3), 1313–1337 (2018).

[123] X. Li, J. Shen, Error analysis of the SAV-MAC scheme for the Navier-Stokes equations. SIAM J. Numer. Anal. **58**(5), 2465–2491 (2020).

[124] X. Li, J. Shen, Z. Liu, New SAV-pressure correction methods for the Navier-Stokes equations: stability and error analysis. Math. Comp. **91**(333), 141–167 (2022).

[125] X.D. Liu, S. Osher, T. Chan, Weighted essentially non-oscillatory schemes. J. Comput. Phys. **115**(1), 200–212 (1994).

[126] C. Liu, F. Frank, F.O. Alpak, B. Rivière, An interior penalty discontinuous Galerkin approach for 3D incompressible Navier–Stokes equation for permeability estimation of porous media. J. Comput. Phys. **396**, 669–686 (2019).

[127] C. Liu, D. Ray, C. Thiele, L. Lin, B. Riviere, A pressure-correction and bound-preserving discretization of the phase-field method for variable density two-phase flows. J. Comput. Phys. **449**, 110769 (2022).

[128] C. Liu, B. Riviere, J. Shen, X. Zhang, A simple and efficient convex optimization based bound-preserving high order accurate limiter for Cahn–Hilliard–Navier–Stokes system. SIAM J. Sci. Comput. **46**, A1923–A1948 (2024).

[129] C. Liu and J. Shen, A phase field model for the mixture of two incompressible fluids and its approximation by a Fourier-spectral method. Phys. D **179**, 211–228 (2003).

[130] S.J. Li, Z.J. Wang, L. Ju, L.S. Luo, Explicit large time stepping with a second-order exponential time integrator scheme for unsteady and steady flows. 55th AIAA Aerosp. Sci. Meet., 0753 (2017).

[131] Y. Li and J. Yang, Consistency-enhanced SAV BDF2 time-marching method with relaxation for the incompressible Cahn–Hilliard–Navier–Stokes binary fluid model. Commun. Nonlinear Sci. Numer. Simul. **118**, 107055 (2023).

[132] J.S. Lowengrub, A. Ratz, A. Voigt, Phase-field modeling of the dynamics of multicomponent vesicles: Spinodal decomposition, coarsening, budding, and fission. Phys. Rev. E **79**(3) (2009).

[133] T. Lu, L. Ju, and L. Zhu, A multiple transferable neural network method with domain decomposition for elliptic interface problems. J. Comput. Phys. **530**, 113902 (2025).

[134] V.T. Luan, A. Ostermann, Exponential Runge-Kutta methods of high-order for parabolic problems. J. Comput. Appl. Math. **256**, 168–179 (2014).

[135] M.L. Mansour, A. Hamed, Implicit solution of the incompressible Navier-Stokes equations on a non-staggered grid. J. Comput. Phys. **86**(1), 147–167 (1990).

[136] E. Martinez-Paneda, A. Golahmar, C. Niordson, A phase field formulation for hydrogen assisted cracking. Comput. Methods Appl. Mech. Engrg. **342**, 742–761 (2018).

[137] R. Mattey and S. Ghosh, A novel sequential method to train physics informed neural networks for Allen Cahn and Cahn Hilliard equations. Comput. Methods Appl. Mech. Engrg. **390**, 114474 (2022).

[138] X. Meng, T.T.P. Hoang, Z. Wang, L. Ju, Localized exponential time differencing method for shallow water equations: Algorithms and numerical study. Commun. Comput. Phys. **29**(1) (2020).

[139] K.W. Morton, T. Sonar, Finite volume methods for hyperbolic conservation laws. Acta Numer. **16**, 155–238 (2007).

[140] S. Nguyen, R. Folch, V.K. Verma, H. Henry, M. Plapp, Phase-field simulations of viscous fingering in shear-thinning fluids. Phys. Fluids **22**(10), 103102 (2010).

[141] Q. Nie, F.Y.M. Wan, Y.-T. Zhang, X.-F. Liu, Compact integration factor methods in high spatial dimensions. J. Comput. Phys. **227**, 5238–5255 (2008).

[142] J. Niesen, W.M. Wright, Algorithm 919: A Krylov subspace algorithm for evaluating the φ-functions appearing in exponential integrators. ACM Trans. Math. Softw. **38**, 22:1–22:19 (2012).

[143] G. Orlando, A. Della Rocca, P.F. Barbante, L. Bonaventura, N. Parolini, An efficient and accurate implicit DG solver for the incompressible Navier–Stokes equations. Int. J. Numer. Methods Fluids **94**(9), 1484–1516 (2022).

[144] A. Ostermann, F. Rousset, K. Schratz, Fourier integrator for periodic NLS: low regularity estimates via discrete Bourgain spaces. J. Eur. Math. Soc. **25**(10), 3913–3952 (2023).

[145] A. Ostermann, F. Rousset, K. Schratz, Error estimates of a Fourier integrator for the cubic Schrödinger equation at low regularity. Found. Comput. Math. **21**(3), 725–765 (2021).

[146] A. Ostermann, K. Schratz, Low regularity exponential-type integrators for semilinear Schrödinger equations. Found. Comput. Math. **18**, 731–755 (2018).

[147] A. Ostermann, C. Su, Two exponential-type integrators for the “good” Boussinesq equation. Numer. Math. **143**, 683–712 (2019).

[148] Q. Pan, Y. Huang, T. Rabczuk, X. Yang, The subdivision-based IGA-EIEQ numerical scheme for the Navier-Stokes equations coupled with Cahn-Hilliard phase-field model of two-phase incompressible flow on complex curved surfaces. Comput. Methods Appl. Mech. Eng. **424**, 116901 (2024).

[149] R. Peyret, T.D. Taylor, Computational methods for fluid flow. Springer Science & Business Media, 2012.

[150] K. Pieper, K.C. Sockwell, M. Gunzburger, Exponential time differencing for mimetic multi-layer ocean models. J. Comput. Phys. **398**, 108900 (2019).

[151] J. Qiao, Y. Qiao, and Y. He, Stabilized radial basis function finite difference schemes with mass conservation for the Cahn–Hilliard equation on surfaces. Entropy **27**, 1206 (2025).

[152] A. Quarteroni, A. Valli, Numerical Approximation of Partial Differential Equations. Springer, Berlin, Heidelberg, 2008.

[153] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. **378**, 686–707 (2019).

[154] F. Rousset, K. Schratz, A general framework of low regularity integrators. SIAM J. Numer. Anal. **59**(3), 1735–1768 (2021).

[155] Y. Saad, Analysis of some Krylov subspace approximations to the matrix exponential operator. SIAM J. Numer. Anal. **29**(1), 209–228 (1992).

[156] K. Schratz, Y. Wang, X. Zhao, Low-regularity integrators for nonlinear Dirac equations. Math. Comp. **90**(327), 189–214 (2021).

[157] P.W. Schroeder, V. John, P.L. Lederer, C. Lehrenfeld, G. Lube, J. Schöberl, On reference solutions and the sensitivity of the 2D Kelvin–Helmholtz instability problem. Comput. Math. Appl. **77**(4), 1010–1028 (2019).

[158] J. Shen, Modeling and numerical approximation of two-phase incompressible flows by a phase-field approach. Multiscale Model. Anal. Mater. Simul., 147–195 (2012).

[159] J. Shen, On error estimates of projection methods for Navier–Stokes equations: First-order schemes. SIAM J. Numer. Anal. **29**(1), 57–77 (1992).

[160] J. Shen, Remarks on the pressure error estimates for the projection methods. Numer. Math. **67**(4), 513–520 (1994).

[161] J. Shen, T. Tang, J. Yang, On the maximum principle preserving schemes for the generalized Allen-Cahn equation. Commun. Math. Sci. **14**(6), 1517–1534 (2016).

[162] J. Shen, C. Wang, X. Wang, S.M. Wise, Second-order convex splitting schemes for gradient flows with Ehrlich-Schwoebel type energy: application to thin film epitaxy. SIAM J. Numer. Anal. **50**(1), 105–125 (2012).

[163] J. Shen, J. Xu, Convergence and error analysis for the scalar auxiliary variable (SAV) schemes to gradient flows. SIAM J. Numer. Anal. **56**, 2895–2912 (2018).

[164] J. Shen, J. Xu, J. Yang, The scalar auxiliary variable (SAV) approach for gradient flows. J. Comput. Phys. **353**, 407–416 (2018).

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

[166] J. Shen, X. Yang, Numerical approximations of Allen-Cahn and Cahn-Hilliard equations. Discrete Contin. Dyn. Syst. **28**(4), 1669–1691 (2010).

[167] J. Shen and X. Yang, A phase-field model and its numerical approximation for two-phase incompressible flows with different densities and viscosities. SIAM J. Sci. Comput. **32**, 1159–1179 (2010).

[168] J. Shen and X. Yang, Decoupled energy stable schemes for phase-field models of two-phase complex fluids. SIAM J. Sci. Comput. **36**, B122–B145 (2014).

[169] J. Shen and X. Yang, Decoupled, energy stable schemes for phase-field models of two-phase incompressible flows. SIAM J. Numer. Anal. **53**, 279–296 (2015).

[170] C.-W. Shu, High order weighted essentially nonoscillatory schemes for convection dominated problems. SIAM Rev. **51**, 82–126 (2009).

[171] C.-W. Shu, TVB uniformly high-order schemes for conservation laws. Math. Comp. **49**(179), 105–121 (1987).

[172] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**(2), 439–471 (1988).

[173] R.B. Sidje, EXPOKIT: A software package for computing matrix exponentials. ACM Trans. Math. Softw. **24**, 130–156 (1998).

[174] T. Tang, J. Yang, Implicit-explicit scheme for the Allen-Cahn equation preserves the maximum principle. J. Comp. Math. **34**(5), 451–461 (2016).

[175] R. Temam, Navier–Stokes Equations and Nonlinear Functional Analysis. SIAM, 1995.

[176] M. Tokman, Efficient integration of large stiff systems of ODEs with exponential propagation iterative (EPI) methods. J. Comput. Phys. **213**, 748–776 (2006).

[177] F. Tone, Error analysis for a second order scheme for the Navier–Stokes equations. Appl. Numer. Math. **50**(1), 93–119 (2004).

[178] S. Turek, A comparative study of time-stepping techniques for the incompressible Navier-Stokes equations: from fully implicit non-linear schemes to semi-implicit projection methods. Int. J. Numer. Methods Fluids **22**(10), 987–1011 (1996).

[179] L. Wang and H. Yu, On efficient second order stabilized semi-implicit schemes for the Cahn-Hilliard phase-field equation. J. Sci. Comput. **77**, 1185–1209 (2018).

[180] X. Wang, L. Ju, Q. Du, Efficient and stable exponential time differencing Runge-Kutta methods for phase field elastic bending energy models. J. Comput. Phys. **316**, 21–38 (2016).

[181] S. Wang, W. Wang, J. Li, An efficient dynamically regularized Lagrange multiplier method for the incompressible MHD equations. Commun. Nonlinear Sci. Numer. Simul. **152**, 109235 (2026).

[182] C. Wang, J. Wang, Z. Xia, L. Xu, Optimal error estimates of a second-order projection finite element method for magnetohydrodynamic equations. Math. Model. Numer. Anal. **56**(3), 767–789 (2022).

[183] C. Wang, S.M. Wise, An energy stable and convergent finite-difference scheme for the modified phase field crystal equation. SIAM J. Numer. Anal. **49**(3), 945–969 (2011).

[184] Weinan E, J.G. Liu, Projection method I: convergence and numerical boundary layers. SIAM J. Numer. Anal. **32**(4), 1017–1057 (1995).

[185] C. L. Wight and J. Zhao, Solving Allen-Cahn and Cahn-Hilliard equations using the adaptive physics informed neural networks. Commun. Comput. Phys. **29**, 930–954 (2021).

[186] S.M. Wise, C. Wang, J.S. Lowengrub, An energy stable and convergent finite difference scheme for the phase field crystal equation. SIAM J. Numer. Anal. **47**, 2269–2288 (2009).

[187] K.L. Wong, A.J. Baker, A 3D incompressible Navier–Stokes velocity–vorticity weak form finite element algorithm. Int. J. Numer. Methods Fluids **38**(2), 99–123 (2002).

[188] Y. Wu, X. Zhao, Optimal convergence of a second-order low-regularity integrator for the KdV equation. IMA J. Numer. Anal. **42**(4), 3499–3528 (2022).

[189] Y. Wu, X. Zhao, Embedded exponential-type low-regularity integrators for KdV equation under rough data. BIT Numer. Math. **62**(3), 1049–1090 (2022).

[190] Y. Xing, X. Zhang, C.-W. Shu, Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations. Adv. Water Resour. **33**(12), 1476–1493 (2010).

[191] J. Xu, Y. Li, S. Wu, A. Bousquet, On the stability and accuracy of partially and fully implicit schemes for phase field modeling. Comput. Methods Appl. Mech. Engrg. **345**, 826–853 (2019).

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

[193] Z. Xu, X. Yang, H. Zhang, Z. Xie, Efficient and linear schemes for anisotropic Cahn-Hilliard model using the stabilized-invariant energy quadratization (S-IEQ) approach. Comput. Phys. Commun. **238**, 36–49 (2019).

[194] X. Yang, Linear, first and second-order, unconditionally energy stable numerical schemes for the phase field model of homopolymer blends. J. Comput. Phys. **327**, 294–316 (2016).

[195] X. Yang, L. Ju, Linear and unconditionally energy stable schemes for the binary fluid–surfactant phase field model. Comput. Methods Appl. Mech. Engrg. **318**, 1005–1029 (2017).

[196] X. Yang, L. Ju, Efficient linear schemes with unconditional energy stability for the phase field elastic bending energy model. Comput. Methods Appl. Mech. Engrg. **315**, 691–712 (2017).

[197] J. Yang, Z. Tan, J. Kim, Original variables based energy-stable time-dependent auxiliary variable method for the incompressible Navier–Stokes equation. Comput. & Fluids **240**, 105432 (2022).

[198] J. Yang, N. Yi, Y. Chen, Optimal error estimates of a SAV–FEM for the Cahn-Hilliard-Navier-Stokes model. J. Comput. Appl. Math. **438**, 115577 (2024).

[199] X. Yang, G. Zhang, Convergence analysis for the invariant energy quadratization (IEQ) schemes for solving the Cahn-Hilliard and Allen-Cahn equations with general nonlinear potential. J. Sci. Comput. **82**(3), 55 (2020).

[200] Q. Ye, Z. Ouyang, C. Chen, X. Yang, Efficient decoupled second-order numerical scheme for the flow-coupled Cahn–Hilliard phase-field model of two-phase flows. J. Comput. Appl. Math. **405**, 113875 (2022).

[201] Z. Zhang, F. Bao, L. Ju, G. Zhang, Transferable neural networks for partial differential equations. J. Sci. Comput. **99**, 2 (2024).

[202] Q. Zhang, C.-W. Shu, Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin methods for scalar conservation laws. SIAM J. Numer. Anal. **42**(2), 641–666 (2004).

[203] H. Zhang, J. Yan, X. Qian, X. Chen, S. Song, Explicit third-order unconditionally structure-preserving schemes for conservative Allen–Cahn equations. J. Sci. Comput. **90**, 1–29 (2022).

[204] H. Zhang, J. Yan, X. Qian, S. Song, Numerical analysis and applications of explicit high order maximum principle preserving integrating factor Runge–Kutta schemes for Allen–Cahn equation. Appl. Numer. Math. **161**, 372–390 (2021).

[205] J. Zhang, C. Zhou, Y. Wang, L. Ju, Q. Du, X. Chi, D. Xu, D. Chen, Y. Liu, Z. Liu, Extreme-scale phase field simulations of coarsening dynamics on the Sunway TaihuLight supercomputer, in Proceedings of SC'16, Article #4, 12 pages, 2016.

[206] D. Zhu, X. Feng, L. Qian, Error analysis of Crank-Nicolson-Leapfrog scheme for the two-phase Cahn-Hilliard-Navier-Stokes incompressible flows. Comput. Math. Appl. **172**, 78–93 (2024).

[207] L. Zhu, L. Ju, W. Zhao, Fast high-order compact exponential time differencing Runge-Kutta methods for second-order semilinear parabolic equations. J. Sci. Comput. **67**, 1043–1065 (2016).

Appendix A

Proof of energy stability for the first-order pressure-correction DRLM scheme

Proof. Adding equations (4.18a) and (4.18b) together, noting that $\mathbf{f} = \mathbf{0}$, yields

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu\Delta\overline{\mathbf{u}}^{n+1} + q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n + \nabla p^{n+1} = \mathbf{0}.$$

This implies that

$$\begin{aligned}\left(\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n, \overline{\mathbf{u}}^{n+1}\right) &= \left(\nu\Delta\overline{\mathbf{u}}^{n+1} - \nabla p^{n+1}, \overline{\mathbf{u}}^{n+1}\right) \\ &= -\nu\left\|\nabla\overline{\mathbf{u}}^{n+1}\right\|_0^2 - \left(\nabla p^{n+1}, \overline{\mathbf{u}}^{n+1} - \mathbf{u}^{n+1}\right), \qquad \text{(A.1)}\end{aligned}$$

where in the last identity, we have used integration by parts and the fact that $\left(\nabla p^{n+1}, \mathbf{u}^{n+1}\right) = 0$ due to the incompressibility condition (4.18c). From (4.18b), we have $\overline{\mathbf{u}}^{n+1} - \mathbf{u}^{n+1} = \Delta t \nabla(p^{n+1} - p^n)$. Therefore, (A.1) can be written as

$$\begin{aligned}\left(\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + q^{n+1}(\mathbf{u}^n \cdot \nabla)\mathbf{u}^n, \overline{\mathbf{u}}^{n+1}\right) &= -\nu\left\|\nabla\overline{\mathbf{u}}^{n+1}\right\|_0^2 - \Delta t\left(\nabla p^{n+1}, \nabla(p^{n+1} - p^n)\right) \\ &= -\nu\left\|\nabla\overline{\mathbf{u}}^{n+1}\right\|_0^2 - \frac{\Delta t}{2}\left(\|\nabla p^{n+1}\|_0^2 - \|\nabla p^n\|_0^2 + \|\nabla(p^{n+1} - p^n)\|_0^2\right) \\ &\leq -\frac{\Delta t}{2}\left(\|\nabla p^{n+1}\|_0^2 - \|\nabla p^n\|_0^2\right). \qquad \text{(A.2)}\end{aligned}$$

Combining (4.18d) and (A.2) gives us

$$\frac{\mathcal{K}(\mathbf{u}^{n+1}) - \mathcal{K}(\mathbf{u}^n)}{\Delta t} + \theta\frac{(q^{n+1})^2 - (q^n)^2}{\Delta t} \leq -\frac{\Delta t}{2}\left(\|\nabla p^{n+1}\|_0^2 - \|\nabla p^n\|_0^2\right),$$

which leads to

$$\mathcal{K}(\mathbf{u}^{n+1}) + \frac{\Delta t^2}{2}\|\nabla p^{n+1}\|_0^2 + \theta(q^{n+1})^2 \leq \mathcal{K}(\mathbf{u}^n) + \frac{\Delta t^2}{2}\|\nabla p^n\|_0^2 + \theta(q^n)^2.$$

Thus, (4.23) holds true for $0 \leq n \leq N_t - 1$. □

Appendix B

Proof of energy stability for the second-order rotational pressure-correction DRLM scheme

Proof. Let $s^n := \nabla \cdot \sum_{i=1}^{n} \overline{\mathbf{u}}^i$ for $n \geq 1$, we have $\nabla \cdot \overline{\mathbf{u}}^{n+1} = s^{n+1} - s^n$. By taking the sum of equations (4.19a) and (4.19b), we obtain

$$\frac{3\mathbf{u}^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} - \nu\Delta\overline{\mathbf{u}}^{n+1} + q^{n+1}(\tilde{\mathbf{u}}^{n+1} \cdot \nabla)\tilde{\mathbf{u}}^{n+1} + \nabla(p^{n+1} + \nu s^{n+1} - \nu s^n) = \mathbf{0}.$$

It follows that

$$\begin{aligned}
&\left(\frac{3\mathbf{u}^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} + q^{n+1}(\tilde{\mathbf{u}}^{n+1} \cdot \nabla)\tilde{\mathbf{u}}^{n+1}, \overline{\mathbf{u}}^{n+1}\right) \\
&= \left(\nu\Delta\overline{\mathbf{u}}^{n+1} - \nabla(p^{n+1} + \nu s^{n+1} - \nu s^n), \overline{\mathbf{u}}^{n+1}\right) \\
&= -\nu\|\nabla\overline{\mathbf{u}}^{n+1}\|_0^2 + \nu\left(\nabla s^n, \overline{\mathbf{u}}^{n+1}\right) - \left(\nabla(p^{n+1} + \nu s^{n+1}), \overline{\mathbf{u}}^{n+1} - \mathbf{u}^{n+1}\right),
\end{aligned} \tag{B.1}$$

where in the last equality, we have used integration by parts and the fact that

$$\left(\nabla(p^{n+1} + \nu s^{n+1}), \mathbf{u}^{n+1}\right) = 0$$

due to (4.19c). It is implied from (4.19b) that

$$\overline{\mathbf{u}}^{n+1} - \mathbf{u}^{n+1} = \frac{2\Delta t}{3}\nabla(p^{n+1} + \nu s^{n+1} - p^n - \nu s^n). \tag{B.2}$$

The combination of (B.1) and (B.2) results in

$$\begin{aligned}
&\left(\frac{3\mathbf{u}^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} + q^{n+1}(\tilde{\mathbf{u}}^{n+1} \cdot \nabla)\tilde{\mathbf{u}}^{n+1}, \overline{\mathbf{u}}^{n+1}\right) \\
&= -\nu\|\nabla\overline{\mathbf{u}}^{n+1}\|_0^2 + \nu\left(\nabla s^n, \overline{\mathbf{u}}^{n+1}\right) - \frac{2\Delta t}{3}\left(\nabla(p^{n+1} + \nu s^{n+1}), \nabla(p^{n+1} + \nu s^{n+1} - p^n - \nu s^n)\right).
\end{aligned} \tag{B.3}$$

Since $\nabla \cdot \overline{\mathbf{u}}^{n+1} = s^{n+1} - s^n$ and $\|\nabla \cdot \overline{\mathbf{u}}^{n+1}\|_0 \le \|\nabla \overline{\mathbf{u}}^{n+1}\|_0$, the second term on the right hand side of (B.3) can be estimated as

$$
\begin{aligned}
\nu\left(\nabla s^n, \overline{\mathbf{u}}^{n+1}\right) &= -\nu(s^n, \nabla \cdot \overline{\mathbf{u}}^{n+1}) = -\nu(s^n, s^{n+1} - s^n) \\
&= -\frac{\nu}{2}\left(\|s^{n+1}\|_0^2 - \|s^n\|_0^2 - \|s^{n+1} - s^n\|_0^2\right) \\
&= -\frac{\nu}{2}\left(\|s^{n+1}\|_0^2 - \|s^n\|_0^2\right) + \frac{\nu}{2}\|\nabla \cdot \overline{\mathbf{u}}^{n+1}\|_0^2 \\
&\le -\frac{\nu}{2}\left(\|s^{n+1}\|_0^2 - \|s^n\|_0^2\right) + \frac{\nu}{2}\|\nabla \overline{\mathbf{u}}^{n+1}\|_0^2. \qquad \text{(B.4)}
\end{aligned}
$$

Using the inequality $a(a-b) \ge \frac{1}{2}(a^2 - b^2)$, the last term on the right hand side of (B.3) can be bounded by

$$
\begin{aligned}
&-\frac{2\Delta t}{3}\left(\nabla(p^{n+1} + \nu s^{n+1}), \nabla(p^{n+1} + \nu s^{n+1} - p^n - \nu s^n)\right) \\
&\le -\frac{\Delta t}{3}\left(\left\|\nabla(p^{n+1} + \nu s^{n+1})\right\|_0^2 - \left\|\nabla(p^n + \nu s^n)\right\|_0^2\right). \qquad \text{(B.5)}
\end{aligned}
$$

From (B.3), (B.4), and (B.5), we derive

$$
\begin{aligned}
&\left(\frac{3\mathbf{u}^{n+1} - 4\mathbf{u}^n + \mathbf{u}^{n-1}}{2\Delta t} + q^{n+1}(\tilde{\mathbf{u}}^{n+1} \cdot \nabla)\tilde{\mathbf{u}}^{n+1}, \overline{\mathbf{u}}^{n+1}\right) \\
&\le -\frac{\nu}{2}\left(\|s^{n+1}\|_0^2 - \|s^n\|_0^2\right) - \frac{\Delta t}{3}\left(\left\|\nabla(p^{n+1} + \nu s^{n+1})\right\|_0^2 - \left\|\nabla(p^n + \nu s^n)\right\|_0^2\right). \qquad \text{(B.6)}
\end{aligned}
$$

Combining (4.19d) and (B.6) yields

$$
\begin{aligned}
&\frac{3\mathcal{K}(\mathbf{u}^{n+1}) - 4\mathcal{K}(\mathbf{u}^n) + \mathcal{K}(\mathbf{u}^{n-1})}{2\Delta t} + \theta\frac{3(q^{n+1})^2 - 4(q^n)^2 + (q^{n-1})^2}{2\Delta t} \\
&\le -\frac{\nu}{2}\left(\|s^{n+1}\|_0^2 - \|s^n\|_0^2\right) - \frac{\Delta t}{3}\left(\left\|\nabla(p^{n+1} + \nu s^{n+1})\right\|_0^2 - \left\|\nabla(p^n + \nu s^n)\right\|_0^2\right),
\end{aligned}
$$

or equivalently,

$$
\begin{aligned}
&\frac{3}{2}\mathcal{K}(\mathbf{u}^{n+1}) - \frac{1}{2}\mathcal{K}(\mathbf{u}^n) + \frac{\Delta t^2}{3}\left\|\nabla(p^{n+1} + \nu s^{n+1})\right\|_0^2 + \frac{\Delta t\nu}{2}\|s^{n+1}\|_0^2 + \theta\left[\frac{3}{2}(q^{n+1})^2 - \frac{1}{2}(q^n)^2\right] \\
&\le \frac{3}{2}\mathcal{K}(\mathbf{u}^n) - \frac{1}{2}\mathcal{K}(\mathbf{u}^{n-1}) + \frac{\Delta t^2}{3}\left\|\nabla(p^n + \nu s^n)\right\|_0^2 + \frac{\Delta t\nu}{2}\|s^n\|_0^2 + \theta\left[\frac{3}{2}(q^n)^2 - \frac{1}{2}(q^{n-1})^2\right].
\end{aligned}
$$

Thus, we obtain (4.24) for $1 \le n \le N_t - 1$. □

Appendix C

GTransNet-BDF schemes for the Cahn-Hilliard equation with variable mobility

We extend the GTransNet-BDF framework to the CH equation with variable mobility, which takes the following form

$$\begin{cases} \dfrac{\partial u}{\partial t} = \nabla \cdot \big(M(u)\, \nabla \mu\big), & \text{in } \Omega \times (0, T], \\ \mu = -\varepsilon^2 \Delta u + f(u), & \text{in } \Omega \times (0, T], \end{cases} \tag{C.1}$$

subject to the same initial and boundary conditions as in Section 5.2. We remark that a direct application of the GTransNet-BDF method to (C.1) leads to solution-dependent coefficients in the least-squares matrix (5.14). To recover the one-time factorization of the coefficient matrix for efficient implementation, we adopt the mobility splitting strategy introduced in [103]. Toward that end, let $\Gamma := \max_u M(u)$ and consider the decomposition

$$M(u)\, \nabla \mu = \Gamma\, \nabla \mu + \big(M(u) - \Gamma\big) \nabla \mu,$$

so that the first equation in (C.1) can be written equivalently as

$$\frac{\partial u}{\partial t} = \Gamma \Delta \mu + \nabla \cdot \Big[\big(M(u) - \Gamma\big) \nabla \mu\Big]. \tag{C.2}$$

The first term on the right-hand side of (C.2) carries a constant coefficient and is treated implicitly; the second term, in which $M(u) - \Gamma \leq 0$, is treated explicitly via extrapolation. As in Section 5.2, we let $\kappa \geq 0$ be a stabilization constant, set $f_\kappa(u) = f(u) - \kappa u$, and approximate both u and μ by their GTransNet expansions (5.9) in the basis (5.2).

C.1 GTransNet-BDF schemes

Applying the backward Euler method to (C.1)-(C.2) with explicit treatment of the nonlinear term and the variable mobility correction term yields the first-order GTransNet-BDF1 scheme

$$\sum_{j=1}^{N_L} \frac{\alpha_j^{n+1} - \alpha_j^n}{\Delta t} \phi_j(\mathbf{x}) = \Gamma \sum_{j=1}^{N_L} \beta_j^{n+1} \Delta \phi_j(\mathbf{x}) + \nabla \cdot \Big[\big(M(u_{\mathrm{NN}}^n(\mathbf{x})) - \Gamma \big) \nabla \mu_{\mathrm{NN}}^{\star,n}(\mathbf{x}) \Big], \tag{C.3a}$$

$$\sum_{j=1}^{N_L} \beta_j^{n+1} \phi_j(\mathbf{x}) = \sum_{j=1}^{N_L} \alpha_j^{n+1} (-\varepsilon^2 \Delta + \kappa) \phi_j(\mathbf{x}) + f_\kappa(u_{\mathrm{NN}}^n(\mathbf{x})), \tag{C.3b}$$

where $\mu_{\mathrm{NN}}^{\star,n} = -\varepsilon^2 \Delta u_{\mathrm{NN}}^n + f(u_{\mathrm{NN}}^n)$. Similarly, the corresponding second-order GTransNet-BDF2 scheme is given by

$$\sum_{j=1}^{N_L} \frac{3\alpha_j^{n+1} - 4\alpha_j^n + \alpha_j^{n-1}}{2\Delta t} \phi_j(\mathbf{x}) = \Gamma \sum_{j=1}^{N_L} \beta_j^{n+1} \Delta \phi_j(\mathbf{x}) + \nabla \cdot \Big[\big(M(\tilde{u}_{\mathrm{NN}}^{n+1}(\mathbf{x})) - \Gamma \big) \nabla \tilde{\mu}_{\mathrm{NN}}^{n+1}(\mathbf{x}) \Big], \tag{C.4a}$$

$$\sum_{j=1}^{N_L} \beta_j^{n+1} \phi_j(\mathbf{x}) = \sum_{j=1}^{N_L} \alpha_j^{n+1} (-\varepsilon^2 \Delta + \kappa) \phi_j(\mathbf{x}) + f_\kappa(\tilde{u}_{\mathrm{NN}}^{n+1}(\mathbf{x})), \tag{C.4b}$$

where $\tilde{u}_{\mathrm{NN}}^{n+1} = 2u_{\mathrm{NN}}^n - u_{\mathrm{NN}}^{n-1}$ and $\tilde{\mu}_{\mathrm{NN}}^{n+1} = -\varepsilon^2 \Delta \tilde{u}_{\mathrm{NN}}^{n+1} + f(\tilde{u}_{\mathrm{NN}}^{n+1})$. For the initialization step, $\{\alpha_j^1\}_{j=1}^{N_L}$ is computed by the GTransNet-BDF1 scheme (C.3).

C.2 Implementation

The least-squares systems for the schemes (C.3) and (C.4) retain the structure of (5.14) with two modifications: the diffusion coefficient D in the implicit term is replaced by Γ, and the first block of the right-hand side vector includes an explicit variable-mobility contribution. For the GTransNet-BDF1 scheme (C.3), we solve for $\mathbf{c}^{n+1} \in \mathbb{R}^{2N_L}$ the following least-squares system

$$\begin{pmatrix} \mathbf{\Phi} & -\Gamma \Delta t \, \mathbf{\Phi}_\Delta \\ \varepsilon^2 \mathbf{\Phi}_\Delta - \kappa \mathbf{\Phi} & \mathbf{\Phi} \\ \mathbf{\Phi}_{\mathrm{bd}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_{\mathrm{bd}} \end{pmatrix} \mathbf{c}^{n+1} = \begin{pmatrix} \mathbf{\Phi} \boldsymbol{\alpha}^n + \Delta t \, \mathbf{g}_{\mathrm{var}}^n \\ f_\kappa(\mathbf{\Phi} \boldsymbol{\alpha}^n) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \tag{C.5}$$

where $\mathbf{g}_{\mathsf{var}}^{n} \in \mathbb{R}^{K_{\mathsf{int}}}$ collects the values of $\nabla\cdot\left[(M(u_{\mathsf{NN}}^{n}) - \Gamma)\,\nabla\mu_{\mathsf{NN}}^{\star,n}\right]$ at the interior collocation points. For the GTransNet-BDF2 scheme (C.4), the least-squares system reads

$$\begin{pmatrix} \frac{3}{2}\mathbf{\Phi} & -\Gamma\Delta t\,\mathbf{\Phi}_{\Delta} \\ \varepsilon^2\mathbf{\Phi}_{\Delta} - \kappa\mathbf{\Phi} & \mathbf{\Phi} \\ \mathbf{\Phi}_{\mathsf{bd}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_{\mathsf{bd}} \end{pmatrix} \mathbf{c}^{n+1} = \begin{pmatrix} 2\mathbf{\Phi}\boldsymbol{\alpha}^{n} - \frac{1}{2}\mathbf{\Phi}\boldsymbol{\alpha}^{n-1} + \Delta t\,\tilde{\mathbf{g}}_{\mathsf{var}}^{n+1} \\ f_{\kappa}(2\mathbf{\Phi}\boldsymbol{\alpha}^{n} - \mathbf{\Phi}\boldsymbol{\alpha}^{n-1}) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \tag{C.6}$$

where $\tilde{\mathbf{g}}_{\mathsf{var}}^{n+1} \in \mathbb{R}^{K_{\mathsf{int}}}$ collects the values of $\nabla\cdot\left[(M(\tilde{u}_{\mathsf{NN}}^{n+1}) - \Gamma)\,\nabla\tilde{\mu}_{\mathsf{NN}}^{n+1}\right]$ at the interior collocation points. The systems (C.5)-(C.6) are solved exactly as in Section 5.2.3, where the QR factorizations are precomputed once and each time step requires only a matrix-vector product followed by back substitution.