Higher Order Time-Implicit Bounds Preserving Discontinuous Galerkin Discretizations for Partial Differential Equations

Fengna Yan

Higher Order Time-Implicit Bounds Preserving Discontinuous Galerkin Discretizations for Partial Differential Equations

Fengna Yan

Invitation

To the public defense
of my PhD thesis

Higher Order
Time-Implicit
Bounds Preserving
Discontinuous
Galerkin
Discretizations for
Partial Differential
Equations

On Friday,
June 30th, 2023
at 10:45

A short introduction
will be given at 10:30

Fengna Yan

HIGHER ORDER TIME-IMPLICIT BOUNDS PRESERVING DISCONTINUOUS GALERKIN DISCRETIZATIONS FOR PARTIAL DIFFERENTIAL EQUATIONS

Fengna Yan

HIGHER ORDER TIME-IMPLICIT BOUNDS PRESERVING DISCONTINUOUS GALERKIN DISCRETIZATIONS FOR PARTIAL DIFFERENTIAL EQUATIONS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr.ir. A. Veldkamp,
on account of the decision of the Doctorate Board
to be publicly defended
on Friday the 30$^{\text{th}}$ of June 2023 at 10.45 hours

by

Fengna Yan

born on 05 of January 1991
in Henan, China.

This dissertation has been approved by:

Supervisors:
prof.dr.ir. J.J.W. van der Vegt
prof.dr. Y. Xu.

Cover design: The cover was designed by Fengna Yan.
Printed by: Ipskamp Printing, Enschede, The Netherlands
ISBN: 978-90-365-5676-7
ISBN: 978-90-365-5677-4 (digital)
DOI: 10.3990/1.9789036556774

GRADUATION COMMITTEE:

Chairman/secretary
prof.dr. J.N. Kok (University of Twente)

Supervisors
prof.dr.ir. J.J.W. van der Vegt (University of Twente)
prof.dr. Y. Xu (University of Science and Technology of China)

Members
prof.dr.ir. E.H. van Brummelen (TU Eindhoven)
prof.dr.ir. B. Koren (TU Eindhoven)
prof.dr. I.S. Pop (Hasselt University)
prof.dr. A.A. Stoorvogel (University of Twente)
dr. M. Schlottbom (University of Twente)

This work was funded by a fellowship from the China Scholarship Council (No.201806340058).

It was carried out at the Mathematics of Computational Science (MACS) group, Department of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands, and at the School of Mathematical Sciences, University of Science and Technology of China, No. 96, JinZhai Road, Baohe District, Hefei, Anhui, 230026, P.R. China.

Contents

Chapter 1

Introduction

1.1 Motivation

Many Partial Differential Equations (PDEs) model problems in physics and engineering that must satisfy bounds on some of the variables. For instance, variables must be positive or satisfy a maximum constraint. If one solves these PDEs numerically then it is crucial to satisfy these bounds, otherwise the solution is not physically realizable and frequently the numerical solution process will break down. In order to study bounds preserving numerical discretizations, we will discuss in this PhD thesis two important classes of nonlinear PDEs that have strict bounds on the solution, namely degenerate parabolic PDEs and the hyperbolic compressible reactive Euler equations.

Nonlinear, possibly degenerate, parabolic equations, describe many problems in science and engineering, such as radiative transport in the diffusive limit, flow of electrons and holes in semi-conductor devices, heat and mass transfer, combustion, flow in porous media, displacement of oil by water in oil reservoirs and the evolution of a gas of fermionic and Bose–Einstein particles. These phenomena are modelled for instance by the radiative transport equation in the diffusive limit [65, 119], the drift-diffusion equation for semiconductors [13, 64], the heat equation [19], the porous media equation [2, 112, 131], the Buckley-Leverett equation [13, 73], and the nonlinear Fokker-Plank equation modelling fermion and boson gases [21, 109]. Preserving bounds on the numerical solution of these parabolic and degenerate parabolic equations is non-trivial, but is further complicated by the fact that many bounds preserving numerical discretizations for parabolic equations also have a severe time step constraint.

A second important example of nonlinear PDEs where the solution must satisfy bounds are the chemically reactive Euler equations, which model inviscid, compressible, reacting flows [10, 11, 14, 117, 118]. These equations arise for instance in combustion problems. The study of such problems is of great value in mitigating the risk of accidental fires, preventing the occurrence of gas explosions in industrial production processes, but also for instance in the study of supernova explosions in astrophysics, and many other applications [38, 81, 87]. It is very challenging to numerically simulate these problems since apart from ensuring positivity of density, pressure and internal energy, the mass fractions of the different species must remain in the domain $[0, 1]$. In addition, in high speed chemically reacting flows the reaction speed can be much larger than the gas velocity. This leads to numerical stiffness problems caused by the chemical reactions, which is one of the main numerical challenges when computing reacting flows. Another important issue with the chemically reactive Euler equations is that even a stable numerical discretization can still produce spurious unphysical solutions in the reaction zone [31, 81], unless one is using a sufficiently fine spatial-temporal resolution in the numerical simulations, with time steps close to the very small chemical time scales, or one uses subcell resolution to capture these local phenomena.

So far most bounds preserving numerical discretizations use explicit time integration methods. For many PDEs, especially higher order PDEs, the time step restriction, which is necessary to ensure stability for explicit time integration methods, generally results in excessively small time steps [95, 132, 133, 134], e.g. $\tau \leqslant Ch^p$, with τ the time step, C a positive constant, h the mesh size, and p the highest order of the spatial derivatives in the PDEs. In addition, enforcing positivity or other bounds on the numerical solution frequently imposes further constraints on the time step [83, 84, 94, 131].

An alternative to time-explicit discretizations is to use implicit time integration methods, which generally allow larger time steps, but at the cost of solving each time step a system of algebraic equations. Since bounds preserving numerical discretizations often use limiters, which frequently contain switches or varying stencils, it is nontrivial to combine time-implicit methods with bounds preserving discretizations. The study of implicit bounds preserving numerical discretizations will be an important topic in this thesis.

Several implicit time integration methods, such as some Diagonally Implicit Runge-Kutta (DIRK) methods [5, 18, 99], are stiffly accurate [61], which results in excellent stability properties, especially for singularly per-

turbed problems, but solving the resulting algebraic equations can be difficult and costly. Alternatively, semi-implicit time integration methods, such as semi-implicit Spectral Deferred Correction (SDC) methods [89], can alleviate the complexity of the algebraic equations that must be solved each time step, but SDC methods may not be sufficiently stable for some strongly nonlinear problems or have a relatively severe time step constraint. Choosing a suitable time integration method therefore is nontrivial. For problems with strong nonlinearities such as the nonlinear Fokker-Plank equation with a singular solution or the chemically reactive Euler equations with stiff source terms, fully implicit time integration method is therefore a good choice and will be extensively used in this thesis. For equations that can separate stiff and non-stiff terms, such as the Allen-Cahn equation, semi-implicit methods are more suitable. We will use the Allen-Cahn equation therefore as a model equation for the development and analysis of higher order accurate discretizations using semi-implicit time integration methods, but the Allen-Cahn equation is also interesting in its own respect.

The Allen-Cahn equation was introduced by Allen and Cahn in [6] to describe the motion of anti-phase boundaries in crystalline solids. At present, using the phase field method [78, 97], the Allen-Cahn equation has been widely used to model many complicated moving interface problems, such as the process of phase separation of a binary alloy at a fixed temperature [32, 46], the mixture of two incompressible fluids, phase transitions and interfacial dynamics in materials science [6, 32]. In particular, special phase separations may appear on static and dynamic surfaces, such as phase separation on lipid bilayer membranes [63, 128] and dendritic crystal growth on curved surfaces [93]. We will analyze in this thesis the stability and prove optimal error estimates for higher order accurate semi-implicit numerical discretizations of the Allen-Cahn equation. The main difficulty here consists of the nonlinear term in the Allen-Cahn equation.

1.2 Bounds preserving numerical discretizations

Typically, bounds on the numerical solution of PDEs are enforced using limiters. The main purpose of the limiter is to locally adjust the numerical solution such that it meets the constraints. Since developing accurate and efficient limiters is in general non-trivial, especially for higher order accurate numerical discretizations, there is a vast literature on bounds preserving limiters. In two seminal papers [133, 134] Zhang and Shu proposed limiters and adjustments to the numerical discretization that preserve bounds

for higher order accurate Discontinuus Galerkin (DG) discretizations for conservation laws. The basic idea of these limiters is to first ensure that the element average of the numerical solution obtained with a first order accurate time integration method satisfies the bounds. Next, they limit the higher order accurate polynomial solution at the quadrature points in each element since this are the only data used in the spatial discretization, and finally they use explicit strong stability preserving Runge-Kutta methods to obtain also higher order accuracy in time. The approach of Zhang and Shu provides a clear framework for many types of PDEs, such as the Euler equations of gas dynamics [134], the compressible Navier-Stokes equations [132] and relativistic hydrodynamics [95].

During the past few years many bounds preserving numerical discretizations for nonlinear degenerate parabolic equations, for which preserving positivity of the numerical solution is crucial, have been proposed. In [131], the authors considered time-explicit Local Discontinuous Galerkin (LDG) discretizations for the porous media equation and presented a limiter to ensure the positivity of the solution. The authors in [84] proposed a modified limiter to preserve the maximum principle for time-explicit DG discretizations of the Fokker–Planck equation. This DG method is, however, limited to third order accuracy. A uniformly accurate, entropy satisfying time-explicit DG method for solving the linear Fokker–Planck equation is presented in [85]. An important element in this algorithm is the use of a truncation operator to ensure nonnegative solutions. In [83] the authors developed time-explicit positivity preserving discretizations for the nonlinear Fokker-Plank equation. Positivity of the numerical solution is enforced using a reconstruction algorithm. The main disadvantage of these time-explicit discretizations for degenerate parabolic PEDs is the severe time constraint $\tau \leqslant Ch^2$.

For the chemically reactive Euler equations, which model inviscid compressible flows with chemical reactions, shocks and detonations, the numerical solution must be physically realizable. Many attempts have been made to ensure that the bounds on the solution, such as nonnegative density and pressure, and mass fractions between zero and one, are preserved [36, 37, 113], and to avoid spurious phenomena [10, 11, 108, 117, 118]. For instance, in order to avoid spurious solutions, a second order MinMax scheme [108], a first order random projection method [10, 11], and Harten's essentially non-oscillatory (ENO) subcell resolution technique [117, 118] were used to discretize the reaction part of the chemically reactive Euler equations. Using splitting methods [51], the chemically reactive Euler equations can be divided into homogeneous equations and reaction equations,

which alleviates the stiffness problems, but the authors in [36, 37, 113] all use time-explicit discretizations in their bounds preserving schemes.

So far, nearly all positivity or bounds preserving numerical discretizations only work in combination with explicit time discretizations, which may result in severe time step restrictions to ensure stability of the numerical discretizations. These time step restrictions can be alleviated using time-implicit integration methods.

In [94], Qin and Shu developed an implicit positivity preserving DG discretization for conservation laws. They use an implicit Euler time integration method and the main idea to preserve positivity is to ensure that in each time step the Jacobian matrix is an M-matrix. This approach is, however, not easy to generalize to higher order accuracy in time and more complicated systems such as the chemically reactive Euler equations. The authors in [22] proposed a new Lagrange multiplier approach to construct semi-implicit positivity preserving schemes for parabolic type equations and solved the Lagrange multiplier using a cut-off approach. They further extended this approach in [23] to construct bounds preserving schemes for a class of semilinear and quasi-linear parabolic equations. In [111] an alternative approach to obtain implicit bounds preserving discretizations of PDEs was introduced, called the Karush-Kuhn-Tucker (KKT) limiter. This method works well in combination with time-implicit discretizations. The main idea of the KKT limiter approach is to reformulate time-implicit numerical discretizations with bounds constraints imposed using Lagrange multipliers as a nonlinear mixed complementarity problem. The resulting algebraic equations are then solved using a semi-smooth Newton method. Considering the potential of the KKT limiter approach to be combined with higher order accurate time-implicit DG discretizations and its suitability for large classes of PDEs, we will extensively investigate in this thesis its potential to obtain accurate and efficient bounds preserving numerical discretizations for degenerate parabolic PDEs and the chemically reactive Euler equations.

1.3 Overview of main numerical techniques used in this PhD thesis

In this section, we will give a brief summary of the main numerical techniques used in this PhD thesis, namely the Local Discontinuous Galerkin (LDG) method for spatial discretizations, the Spectral Deferred Correction (SDC) method and the Diagonally Implicit Runge-Kutta (DIRK) method

for time discretizations, and the Karush-Kuhn-Tucker (KKT) equations for the solution of constrained optimization problems.

1.3.1 Local discontinuous Galerkin methods

The local discontinuous Galerkin method is an extension of the discontinuous Galerkin method, which is well suited for PDEs with higher order derivatives. The DG method [34, 66] is a finite element method which uses discontinuous, piecewise polynomials as basis functions. The DG method results in an element-wise conservative numerical discretization, which is particularly important for conservation laws. Due to the use of discontinuous basis functions, DG methods are well suited for *hp*-mesh adaptation, in which the local mesh is refined (*h*-adaptation) or the polynomial order of the basis functions is adjusted (*p*-adaptation), and generally achieve a high degree of parallelization. The DG method was first proposed by Reed and Hill in [96] for the solution of the neutron transport equation. Cockburn et al. [25, 27, 28, 29] subsequently extended the DG method to nonlinear hyperbolic conservation laws, which resulted in many applications, also including bounds preserving discretizations, e.g. [95, 132, 133, 134]. The DG method has many advantages, such as flexibility and efficiency in handling discontinuities and complex geometries, the use of highly nonuniform and unstructured meshes, simple choices of trial and test spaces, and excellent parallelizability.

The LDG method was put forward by Cockburn and Shu in [30] to deal with PDEs that contain second order spatial derivatives. The main idea of the LDG method is to apply the DG method after rewriting the higher order PDEs as a first order set. We refer for general information about the LDG method for linear cases to [35, 114, 125, 135] and for nonlinear cases to [9, 56, 60, 123, 124]. The LDG method not only inherits the advantages of the DG method, but also facilitates efficient handling of some higher order derivative equations.

We take a two-dimensional scalar conservation law as an example to introduce the LDG method

$$u_t + \nabla \cdot \boldsymbol{F}(u) = \nabla \cdot (A \nabla u), \quad \text{in } \Omega \times (0, T], \tag{1.1}$$

with Ω an open bounded domain in $\mathbb{R}^2$, $\boldsymbol{F}(u) : \mathbb{R} \to \mathbb{R}^2$ the flux function and A a nonnegative constant. The subscript t refers to the time derivative and ∇ is the nabla operator. For the LDG discretization, we rewrite (1.1)

as the first order system

$$u_t + \nabla \cdot \boldsymbol{F}(u) = \nabla \cdot (A\boldsymbol{q}), \tag{1.2a}$$

$$\boldsymbol{q} = \nabla u. \tag{1.2b}$$

Let $\mathcal{T}_h$ be a shape-regular tessellation of Ω with convex quadrilateral elements K, and $\mathcal{Q}_k(K)$ denote the space of tensor product polynomials of degree at most k on each element K. The discontinuous finite element spaces for the LDG discretization are defined as

$$V_h^k = \{v \in L^2(\Omega) : \ v|_K \in \mathcal{Q}_k(K), \ \forall K \in \mathcal{T}_h\},$$
$$\boldsymbol{W}_h^k = \{\boldsymbol{w} \in [L^2(\Omega)]^2 : \ \boldsymbol{w}|_K \in [\mathcal{Q}_k(K)]^2, \ \forall K \in \mathcal{T}_h\},$$

which are allowed to have discontinuities across element interfaces. Let e be an interior edge connected to the "left" and "right" elements denoted, respectively, by K_L and K_R. If u is a function on K_L and K_R, we set $u^L := (u|_{K_L})|_e$ and $u^R := (u|_{K_R})|_e$ for the left and right trace of u at e.

The LDG discretization of (1.2) is: Find $u_h \in V_h^k, \boldsymbol{q}_h \in \boldsymbol{W}_h^k$, such that for all $v_h \in V_h^k, \boldsymbol{p}_h \in \boldsymbol{W}_h^k$ and elements $K \in \mathcal{T}_h$,

$$\int_K (u_h)_t v_h dK - \int_K \boldsymbol{F}(u_h) \cdot \nabla v_h dK + \int_{\partial K} \widehat{\boldsymbol{F}}(u_h^L, u_h^R) \cdot \boldsymbol{\nu} v_h ds$$
$$+ \int_K A\boldsymbol{q}_h \cdot \nabla v_h dK - \int_{\partial K} A\widehat{\boldsymbol{q}}_h \cdot \boldsymbol{\nu} v_h ds = 0, \tag{1.3a}$$

$$\int_K \boldsymbol{q}_h \cdot \boldsymbol{p}_h dK + \int_K u_h \nabla \cdot \boldsymbol{p}_h dK - \int_{\partial K} \widehat{u}_h \boldsymbol{p}_h \cdot \boldsymbol{\nu} ds = 0, \tag{1.3b}$$

with $\boldsymbol{\nu}$ the outward normal vector at ∂K. Here $\widehat{\boldsymbol{F}}(u_h^L, u_h^R)$, $\widehat{\boldsymbol{q}}_h$ and $\widehat{u}_h$ are the so-called "numerical fluxes", which should be chosen to ensure stability. The numerical fluxes are single-valued functions defined at the cell edge, and are related, respectively, to the traces of u_h, $\boldsymbol{q}_h$ on both sides of the cell edge. The choice of numerical flux is not unique. For the convection part, we usually choose monotone numerical fluxes satisfying the following conditions [80] :

- Consistency: $\widehat{F}(u, u) = F(u)$.
- Continuity: $\widehat{F}(u^L, u^R)$ is at least Lipschitz continuous in both arguments.
- Monotonicity: $\widehat{F}(u^L, u^R)$ is monotone non-decreasing for the first argument u^L, and monotone non-increasing for the second argument u^R, e.g. $\widehat{F}(\uparrow, \downarrow)$.

The Lax-Friedrichs flux [29] is often be chosen for $\widehat{\boldsymbol{F}}(u_h^L, u_h^R)$, but also other upwind schemes, for instance the Roe flux [80] or the HLLC flux [110], are frequently used. For more details on approximate Riemann solvers, see [80, 107]. Regarding the diffusion part, alternating fluxes are frequently used for $\widehat{\boldsymbol{q}}_h$ and $\widehat{u}_h$, namely $\widehat{\boldsymbol{q}}_h = \boldsymbol{q}_h^L,\ \widehat{u}_h = u_h^R$ or $\widehat{\boldsymbol{q}}_h = \boldsymbol{q}_h^R,\ \widehat{u}_h = u_h^L$, which provide stable and simple numerical fluxes [111, 121, 122, 123, 125].

1.3.2 Time discretizations

In this section, we will introduce semi-implicit Spectral Deferred Correction (SDC) methods and Diagonally Implicit Runge-Kutta (DIRK) methods. SDC methods are well suited for PDEs for which it is easy to separate the stiff and non-stiff terms, such as phase field problems [49, 82] and the phase field crystal equation [59]. For PDEs that need stiffly accurate time discretizations, such as degenerate parabolic equations and the chemically reactive Euler equations, we choose DIRK methods.

1.3.2.1 Spectral deferred correction methods

Spectral deferred correction methods, which were first proposed by Dutt, Greengard and Rokhlin [39], are high order accurate stable time integration methods for stiff and non-stiff problems. Minion extended in [89] SDC methods to semi-implicit SDC methods to solve ODEs containing both stiff and non-stiff terms. The basic idea of the SDC method is to replace the original ODEs by the corresponding Picard integral equation and discretize it using a Legendre–Gauss type quadrature. The resulting system is first solved either by the Euler forward method (for non-stiff problems) or the Euler backward method (for stiff problems). Next the solution is iteratively improved, with each iteration resulting in one more order of accuracy. The SDC method is a one step method and can be easily constructed for any order of accuracy.

We consider the following model ODE system to introduce the semi-implicit SDC method [89]

$$\begin{cases} u_t = F_S(t, u(t)) + F_N(t, u(t)), & t \in (0, T], \\ u(0) = u_0, \end{cases}$$

where F_N is a non-stiff term and F_S a stiff term, t is time and the subscript t refers to the time derivative. In semi-implicit time discretizations, the stiff term F_S will in general be taken implicitly, and the non-stiff term F_N explicitly since for most PDEs the non-stiff term is not the reason for severe

time step constraints. Especially when F_N is a complicated term treating F_N explicitly will make the algebraic equations that must be solved each time step in the SDC method easier to solve.

The SDC method can be summarized as follows. The time interval $[t^n, t^{n+1}]$ is divided into P parts with points $t^{n,m},\ m = 0, 1, \ldots, P$ such that

$$t^n = t^{n,0} < t^{n,1} < \ldots < t^{n,P} = t^{n+1}.$$

Let $\tau^{n,m} = t^{n,m+1} - t^{n,m}$. We denote with $u^k_{n,m},\ k = 1, 2, \ldots, K$ the k-th order approximation to $u(t^{n,m})$, where the points $\{t^{n,m}\}_{m=0}^P$ are chosen as the Legendre-Gauss-Lobatto nodes in the time interval $[t^n, t^{n+1}]$. Suppose u_n is known, we calculate u_{n+1} using Algorithm 1.

Algorithm 1 SDC methods

Compute the initial approximation:

$u^1_{n,0} = u_n$.

For $m = 0, 1, \ldots, P-1$

$u^1_{n,m+1} = u^1_{n,m} + \tau^{n,m}(F_S(t^{n,m+1}, u^1_{n,m+1}) + F_N(t^{n,m}, u^1_{n,m}))$.

Compute successive corrections:

For $k = 1, 2, \ldots, K$

$u^{k+1}_{n,0} = u_n$.

For $m = 0, 1, \ldots, P-1$

$$\begin{aligned} u^{k+1}_{n,m+1} = u^{k+1}_{n,m} &+ \tau^{n,m}(F_S(t^{n,m+1}, u^{k+1}_{n,m+1}) - F_S(t^{n,m+1}, u^k_{n,m+1})) \\ &+ \tau^{n,m}(F_N(t^{n,m}, u^{k+1}_{n,m}) - F_N(t^{n,m}, u^k_{n,m})) \\ &+ I^{m+1}_m(F_S(t, u^k) + F_N(t, u^k)), \end{aligned}$$

where F_S is treated implicitly, F_N is treated explicitly, and $I^{m+1}_m(F_S(t, u^k) + F_N(t, u^k))$ is the integral of the P-th order interpolating polynomial using the $P+1$ points $(t^{n,l}, F_S(t^{n,l}, u^k_{n,l}) + F_N(t^{n,l}, u^k_{n,l}))_{l=0}^P$ over the subinterval $[t^{n,m}, t^{n,m+1}]$.

Finally, $u_{n+1} = u^{K+1}_{n,P}$.

The order of accuracy of the SDC method in Algorithm 1 is $\min(K + 1, P + 1)$. Compared with implicit-explicit (IMEX) methods [16, 100, 114, 115], semi-implicit SDC methods can be constructed easily and systematically for any order of accuracy.

1.3.2.2 Diagonally Implicit Runge-Kutta methods

Diagonally Implicit Runge-Kutta (DIRK) methods, which were introduced by Butcher [18], are very useful for applications that require an implicit time integration method.

We give a description of the DIRK method using the following ODE,

$$u_t = L(u, t).$$

Suppose that the numerical solution u^n at time t^n is known. The numerical solution at time t^{n+1} is obtained with a DIRK method by first solving for each DIRK stage $i, i = 1, \ldots, s$ the following equations.

$$u^{n+1,i} = u^n + \tau^{n+1} \sum_{j=1}^{i} a_{ij} L(u^{n+1,j}, t^n + c_j \tau^{n+1}), \quad i = 1, 2, \cdots, s. \tag{1.4}$$

Next, the solution at t^{n+1} is obtained from $u^{n+1,i}$ using

$$u^{n+1} = u^n + \tau^{n+1} \sum_{i=1}^{s} b_i L(u^{n+1,i}, t^n + c_i \tau^{n+1}), \tag{1.5}$$

with $\tau^{n+1} = t^{n+1} - t^n$. The coefficient matrix $A = (a_{ij})$ and vectors $b = (b_i), c = (c_i)$ describe the Runge-Kutta method and are defined in the Butcher tableau $\begin{array}{c|c} c & A \\ \hline & b \end{array}$. For discussing the stability property of (1.4)-(1.5), we first define its stability function $R(z)$ as

$$R(z) = 1 + z b^T (I - zA)^{-1} (1, \ldots, 1)^T,$$

with A and b given by the Butcher tableau.

Definition 1.3.1 ([61])**.** (A-stable) A time integration method, whose stability domain $S = \{z \in \mathbb{C} : |R(z)| \leqslant 1\}$ satisfies

$$\{z \in \mathbb{C} : \text{Re } z \leqslant 0\} \subset S,$$

is called A-stable.

Definition 1.3.2 ([61])**.** (L-stable) A time integration method is called L-stable if it is A-stable and if in addition

$$\lim_{z \to \infty} R(z) = 0.$$

L-stable methods are well suited for stiff problems.

Runge-Kutta methods satisfying $a_{si} = b_i,\ i = 1, \ldots, s$ are called stiffly accurate [61, 99], which makes A-stable methods L-stable and implies that $u_h^{n+1} = u_h^{n+1,s}$. DIRK methods are easy to implement since the matrix A in

DIRK methods has a lower triangular structure, which permits solving for each stage individually rather than all stages simultaneously. This is computationally more efficient than using fully implicit Runge-Kutta methods such as Gauss-Radau methods that solve all Runge-Kutta stages simultaneously [61, 74]. The disadvantage of DIRK methods compared to fully implicit Runge-Kutta methods is that more stage equations must be solved to obtain the same order of accuracy.

For the bounds preserving implicit discretizations, we choose the stiffly accurate DIRK methods. The Butcher tableaus of the higher order DIRK methods used in this dissertation are:

- Second order DIRK method [5]

$$(a_{ij}) = \begin{pmatrix} \alpha & 0 \\ 1-\alpha & \alpha \end{pmatrix}, (b_j) = \begin{pmatrix} 1-\alpha & \alpha \end{pmatrix}, (c_i) = \begin{pmatrix} \alpha & 1 \end{pmatrix}, \quad (1.6)$$

  where $\alpha = 1 - \dfrac{\sqrt{2}}{2}$.

- Third order DIRK method [99]

$$(a_{ij}) = \begin{pmatrix} \gamma & 0 & 0 \\ 1/2-\gamma/2 & \gamma & 0 \\ 1-\delta-\gamma & \delta & \gamma \end{pmatrix}, (b_j) = \begin{pmatrix} 1-\delta-\gamma & \delta & \gamma \end{pmatrix},$$

$$(c_i) = \begin{pmatrix} \gamma & 1/2+\gamma/2 & 1 \end{pmatrix}, \quad (1.7)$$

  where $\gamma = 0.435866521508,\ \delta = 0.25(5 - 20\gamma + 6\gamma^2)$.

- Fourth order DIRK method [99]

$$(a_{ij}) = \begin{pmatrix} 1/4 & 0 & 0 & 0 & 0 \\ -1/4 & 1/4 & 0 & 0 & 0 \\ 1/8 & 1/8 & 1/4 & 0 & 0 \\ -3/2 & 3/4 & 3/2 & 1/4 & 0 \\ 0 & 1/6 & 2/3 & -1/12 & 1/4 \end{pmatrix},$$

$$(b_j) = \begin{pmatrix} 0 & 1/6 & 2/3 & -1/12 & 1/4 \end{pmatrix},$$

$$(c_i) = \begin{pmatrix} 1/4 & 0 & 1/2 & 1 & 1 \end{pmatrix}.$$

1.3.3 Karush-Kuhn-Tucker system

Bounds on the numerical solutions will be enforced using the Karush-Kuhn-Tucker (KKT) equations, which are frequently used in constrained optimization [40, 41, 42, 77]. Consider the following constrained optimization problem,

$$\begin{aligned} \min_{U} \quad & \theta(U) \\ \text{subject to} \quad & U \in \mathbb{K}, \end{aligned} \tag{1.8}$$

where the objective function $\theta : \mathbb{R}^n \to \mathbb{R}$ is defined and continuously differentiable on the closed set $\mathbb{K}$ with

$$\mathbb{K} := \{U \in \mathbb{R}^n |\ h(U) = 0,\ g(U) \leqslant 0\} \tag{1.9}$$

and $h : \mathbb{R}^n \to \mathbb{R}^l$, $g : \mathbb{R}^n \to \mathbb{R}^m$ are vector-valued continuously differentiable functions.

The general approach for the numerical treatment of (1.8) is based on Lagrange multiplier theory [41, 42, 69]. Lagrange multipliers are of great importance for the analysis of general constrained optimization problems (1.8) and provide efficient and powerful methods for solving such problems. Let $L(U) = \nabla_U \theta(U)$. Assume that Abadie's constraint qualification [41] holds at U, which means that the tangent cone of $\mathbb{K}$ at $U \in \mathbb{K}$ is equal to its linearization cone. If θ is a convex function and $\mathbb{K}$ a convex set, then there exist vectors $\mu \in \mathbb{R}^l$ and $\lambda \in \mathbb{R}^m$ such that [41, 42, 69, 90]

$$\mathcal{L}(U, \lambda) := L(U) + \nabla_U h(U)^T \mu + \nabla_U g(U)^T \lambda = 0, \tag{1.10a}$$

$$-h(U) = 0, \tag{1.10b}$$

$$0 \geqslant g(U) \perp \lambda \geqslant 0, \tag{1.10c}$$

where μ and λ are the Lagrange multipliers used to ensure $h(U) = 0$ and $g(U) \leqslant 0$, respectively. The compatibility condition (1.10c) is equal to

$$g_j(U) \leqslant 0, \quad \lambda_j \geqslant 0, \quad \text{and} \quad g_j(U)\lambda_j = 0, \quad j = 1, 2, \ldots, m.$$

The mixed complementarity problem (1.10) is the so called the KKT system [41].

Note that the KKT system (1.10) is nonlinear and can not be solved using standard Newton methods due to the compatibility condition (1.10c). There are many semi-smooth Newton methods available for constrained optimization problems [41, 42, 69]. In this thesis, we will use the active set semi-smooth Newton algorithm stated in [111], since it provides a robust Newton method with a good mathematical foundation.

1.4 Thesis objectives and outline

We can summarize the main research objectives that will be discussed in this PhD thesis as:

- Developing entropy dissipative higher order accurate time implicit bounds preserving DIRK-LDG discretizations for nonlinear degenerate parabolic equations.
- Theoretically analyze the unique solvability and unconditional stability of the positivity preserving DIRK-LDG discretizations for nonlinear degenerate parabolic equations.
- Developing higher order accurate time implicit bounds preserving DIRK-DG discretizations for the chemically reactive Euler equations.
- Analyzing the stability and obtain optimal error estimates for higher order accurate SDC-LDG discretizations for the Allen-Cahn equation.

This dissertation is organized as follows: in Chapter 2, we will develop entropy dissipative higher order accurate time implicit positivity preserving DIRK-LDG discretizations for nonlinear degenerate parabolic equations with a gradient flow structure. Also, the theoretical analysis of the unique solvability and unconditional entropy dissipation of the numerical discretization will be considered. In Chapter 3, we will construct higher order accurate time implicit bounds preserving DIRK-DG discretizations for the reactive Euler equations modelling multispecies and multireaction chemically reactive flows. Special attention will be given to the elimination of spurious solutions in the chemical reaction zones. In Chapter 4, stability and error estimates of second and third order accurate SDC-LDG discretizations will be analyzed for the Allen-Cahn equation. Conclusions and outlook will be given in Chapter 5.

Chapter 2

Entropy Dissipative Higher Order Accurate Positivity Preserving Time-Implicit Discretizations for Nonlinear Degenerate Parabolic Equations

Abstract

We develop entropy dissipative higher order accurate Local Discontinuous Galerkin (LDG) discretizations coupled with Diagonally Implicit Runge-Kutta (DIRK) methods for nonlinear degenerate parabolic equations with a gradient flow structure. Using the simple alternating numerical flux, we construct DIRK-LDG discretizations that combine the advantages of higher order accuracy, entropy dissipation and proper long-time behavior. The implicit time-discrete methods greatly alleviate the time-step restrictions needed for stability of the numerical discretizations. Also, the larger time step significantly improves computational efficiency. We theoretically prove unconditional entropy dissipation of the implicit Euler-LDG discretization. Next, in order to ensure positivity of the numerical solution, we use the Karush-Kuhn-Tucker (KKT) limiter, which couples the positivity inequality constraint with higher order accurate DIRK-LDG discretizations using Lagrange multipliers. In addition, mass conservation of the positivity limited solution is ensured by imposing a mass conservation equality constraint to the KKT equations. The unique solvability and unconditional entropy dissipation for an implicit first order accurate in time, but higher order accurate in space, KKT-LDG discretizations are proved, which provides a first theoretical analysis of the KKT limiter. Finally, numerical results are shown to demonstrate

the higher order accuracy and entropy dissipation of the KKT-DIRK-LDG discretizations for problems requiring a positivity limiter.

2.1 Introduction

Consider the following degenerate parabolic equation [13]

$$\begin{cases} u_t = \nabla \cdot (f(u)\nabla(\phi(\boldsymbol{x}) + H'(u))), & \text{in } \Omega \times (0,T], \\ u(\boldsymbol{x},0) = u_0(\boldsymbol{x}), & \text{in } \Omega, \end{cases} \tag{2.1}$$

with zero-flux boundary condition

$$\nabla(\phi(\boldsymbol{x}) + H'(u)) \cdot \boldsymbol{\nu} = 0, \quad \text{on } \partial\Omega \times (0,T], \tag{2.2}$$

where Ω is an open bounded domain in $\mathbb{R}^d, d = 1, 2$, with unit outward normal vector $\boldsymbol{\nu}$ at the boundary $\partial\Omega$, $u(\boldsymbol{x},t) \geqslant 0$ is a nonnegative density with time derivative denoted as u_t, $\phi(\boldsymbol{x})$ is a given potential function for $\boldsymbol{x} \in \mathbb{R}^d$, f, H are given functions such that

$$f : \mathbb{R}^+ \to \mathbb{R}^+, \quad H : \mathbb{R}^+ \to \mathbb{R}, \quad f(u)H''(u) \geqslant 0, \tag{2.3}$$

where $\mathbb{R}^+$ is the nonnegative real space. Here $f(u)H''(u)$ can vanish for certain values of u, resulting in degenerate cases. The entropy corresponding to (2.1) is defined by

$$E(u) = \int_\Omega (u\phi(\boldsymbol{x}) + H(u))d\Omega. \tag{2.4}$$

Multiplying (2.1) with $\phi(\boldsymbol{x}) + H'(u)$ and integrating over Ω, with the zero-flux boundary condition (2.2), together with (2.4), we obtain that the time derivative of the entropy satisfies

$$\frac{d}{dt}E(u) = -\int_\Omega f(u)|\nabla(\phi(\boldsymbol{x}) + H'(u))|^2 d\Omega \leqslant 0. \tag{2.5}$$

System (2.1) can represent different physical problems, such as the porous media equation [112, 131], the nonlinear nonlocal equation with a double-well potential [19], the nonlinear Fokker-Plank model for fermion and boson gases [1, 21, 109].

Recently, many numerical discretizations have been proposed for (2.1); e.g. mixed finite element methods [17], finite volume methods [13, 19], Discontinuous Galerkin (DG) methods [83, 84, 85] and LDG methods [131].

Regarding positivity preserving discretizations, Liu and Yu developed in [84, 85], respectively, for the linear Fokker-Plank equation a maximum preserving DG scheme and an entropy satisfying DG scheme, but these discretizations can not be directly applied to the general case given by (2.1). Liu and Wang subsequently developed in [83] an explicit Runge-Kutta (RK) time-discrete method for (2.1) in one dimension together with a positivity preserving high order accurate DG scheme under some Courant-Friedrichs-Lewy (CFL) constraints. For the porous media equation, an LDG discretization coupled with an explicit RK method was considered in [131], which is similar to the DG method in [83], but it uses a special numerical flux to ensure the non-negativity of the numerical solution. Cheng and Shen in [22] propose a Lagrange multiplier approach to construct positivity preserving schemes for a class of parabolic equations, which is different from (4.1), but contains the porous media equation.

For the time-step τ and mesh size h, the condition $\tau = O(h^2)$ is needed for stability in [83] and [131]. These explicit time discretizations therefore suffer from severe time step restrictions, but currently there are no feasible positivity preserving time-implicit LDG discretizations for (2.1). In this chapter, we present therefore higher order accurate Diagonally Implicit Runge-Kutta (DIRK) LDG discretizations, which ensure positivity and mass conservation of the numerical solution without the severe time step restrictions of explicit methods.

The LDG method proposed by Cockburn and Shu in [30] has many advantages, including high parallelizability, high order accuracy, a simple choice of trial and test spaces and easy handling of complicated geometries. We refer to [26, 58, 106, 135] for examples of applications of the LDG method.

For many physical problems, it is crucial that the numerical discretization preserves the positivity properties of the Partial Differential Equations (PDEs). Not only is this necessary to obtain physically meaningful solutions, but also negative values may result in ill-posedness of the problem and divergence of the numerical discretization. Positivity preserving DG methods have been extensively studied by many mathematicians. However, most positivity preserving DG methods are combined with explicit time-discretizations [83, 126, 133, 134], for which numerical stability frequently imposes severe time step restrictions. These severe time-step constraints make explicit methods impractical for parabolic PDEs, such as (2.1).

Recently, Qin and Shu extended in [94] the general framework for establishing positivity-preserving schemes, proposed in [133, 134], from explicit to implicit time discretizations. They developed for one-dimensional con-

servation laws a positivity preserving DG method with high-order spatial accuracy combined with the first-order backward Euler implicit temporal discretization. This approach requires, however, a detailed analysis of the numerical discretization to ensure positivity and it is not straightforward to extend this approach to higher order accurate time-implicit methods. Huang and Shen in [67] constructed higher order linear bound preserving implicit discretizations for the Keller-Segel and Poisson-Nernst-Planck equations. Van der Vegt, Xia and Xu proposed in [111] the KKT limiter concept to construct positivity preserving time-implicit discretizations. The KKT limiter in [111] is obtained by coupling the inequality and equality constraints imposed by the physical problem with higher order accurate DIRK-DG discretizations using Lagrange multipliers. The resulting semi-smooth nonlinear equations are solved by an efficient active set semi-smooth Newton method.

In this chapter, we consider a general class of nonlinear degenerate parabolic equations given by (2.1) and aim at developing higher order accurate entropy dissipative and positivity preserving time-implicit LDG discretizations. For the spatial discretization, we use an LDG method with simple alternating numerical fluxes, which results in entropy dissipation of the semi-discrete LDG discretization. For the temporal discretization, we consider DIRK methods, which significantly enlarge the time step for stability. The unconditional entropy dissipation of the LDG discretization combined with an implicit Euler time integration method is proved theoretically. We construct positivity preserving discretizations using the KKT limiter by imposing the positivity constraint on the numerical discretization using Lagrange multipliers. The unique solvability of the resulting positivity preserving KKT system is proved. We will also prove the unconditional entropy dissipation of the positivity preserving LDG discretization when it is combined with the backward Euler time integration method. Numerical results are given to demonstrate the accuracy and entropy dissipation of the higher order accurate positivity preserving DIRK-LDG discretizations.

This chapter is organized as follows. In Section 2.2, we present the semi-discrete LDG discretization with simple alternating numerical fluxes for the nonlinear degenerate parabolic equation stated in (2.1) and prove that the numerical approximation is entropy dissipative. Higher order accurate DIRK-LDG discretizations, which enlarge the stable time step to a great extent, are discussed in Section 2.3. The unconditional entropy dissipation of the implicit Euler LDG discretizations are proved in Section 2.3.1. In order to ensure positivity of the numerical solution and mass conservation of the positivity limited numerical discretizations, we introduce in Sec-

tion 2.4.1 the KKT system. The higher order DIRK-LDG discretizations with positivity and mass conservation constraints are formulated in Section 2.4.2 as a KKT mixed complementarity problem. The unique solvability and unconditional entropy dissipation of the resulting algebraic system are proved in Section 2.4.3. In Section 2.5, numerical results are provided to demonstrate the higher order accuracy, positivity and entropy dissipation of the positivity preserving KKT-DIRK-LDG discretizations. Concluding remarks are given in Section 2.6.

2.2 Semi-discrete LDG schemes

2.2.1 Definitions, notations

Let $\mathcal{T}_h$ be a shape-regular tessellation of $\Omega \subset \mathbb{R}^d$, $d = 1, 2$, with line or convex quadrilateral elements K. Given the reference element $\widehat{K} = [-1, 1]^d$. Let $\mathcal{Q}_k(\widehat{K})$ denote the space composed of the tensor product of polynomials $\mathcal{P}_k(\widehat{K})$ on $[-1, 1]$ of degree at most $k \geqslant 0$. Here, we choose for $\mathcal{P}_k(\widehat{K})$ Legendre polynomials. The space $\mathcal{Q}_k(K)$ is obtained by using an isoparametric transformation from element K to the reference element $\widehat{K}$. The finite element spaces V_h^k and $\boldsymbol{W}_h^k$ are defined by

$$V_h^k = \{v \in L^2(\Omega) : \ v|_K \in \mathcal{Q}_k(K), \ \forall K \in \mathcal{T}_h\},$$
$$\boldsymbol{W}_h^k = \{\boldsymbol{w} \in [L^2(\Omega)]^d : \ \boldsymbol{w}|_K \in [\mathcal{Q}_k(K)]^d, \ \forall K \in \mathcal{T}_h\},$$

and are allowed to have discontinuities across element interfaces. Let e be an interior edge connected to the "left" and "right" elements denoted, respectively, by K_L and K_R. If u is a function on K_L and K_R, we set $u^L := (u|_{K_L})|_e$ and $u^R := (u|_{K_R})|_e$ for the left and right trace of u at e.

Note that $L^1(\Omega)$, $L^2(\Omega)$ and $L^\infty(\Omega)$ are standard Sobolev spaces, $\|u\|_{L^2(\Omega)}$ is the $L^2(\Omega)$-norm and $(\cdot, \cdot)_\Omega$ is the $L^2(\Omega)$ inner product. For simplicity, we denote the inner product as $(u, v) := (u, v)_\Omega$.

2.2.2 LDG discretization in space

For the LDG discretization of (2.1), we first rewrite this equation as a first order system

$$\begin{aligned} u_t &= \nabla \cdot \boldsymbol{q}, \\ \boldsymbol{q} &= f(u)\boldsymbol{s}, \\ \boldsymbol{s} &= \nabla p, \\ p &= \phi(\boldsymbol{x}) + H'(u). \end{aligned}$$

Then the LDG discretization can be readily obtained by multiplying the above equations with arbitrary test functions, integrating by parts over each element $K \in \mathcal{T}_h$, and finally a summation of element and face contributions. The LDG discretization can be stated as: find $u_h(t), p_h \in V_h^k$, $\boldsymbol{q}_h, \boldsymbol{s}_h \in \boldsymbol{W}_h^k$, such that for all $\rho, \varphi \in V_h^k$ and $\boldsymbol{\theta}, \boldsymbol{\eta} \in \boldsymbol{W}_h^k$, we have

$$(u_{ht}, \rho) + L_h^1(\boldsymbol{q}_h; \rho) = 0, \tag{2.6a}$$

$$(\boldsymbol{q}_h, \boldsymbol{\theta}) + L_h^2(u_h, \boldsymbol{s}_h; \boldsymbol{\theta}) = 0, \tag{2.6b}$$

$$(\boldsymbol{s}_h, \boldsymbol{\eta}) + L_h^3(p_h; \boldsymbol{\eta}) = 0, \tag{2.6c}$$

$$(p_h, \varphi) + L_h^4(u_h; \varphi) = 0, \tag{2.6d}$$

where

$$L_h^1(\boldsymbol{q}_h; \rho) := (\boldsymbol{q}_h, \nabla\rho) - \sum_{K\in\mathcal{T}_h} (\widehat{\boldsymbol{q}}_h \cdot \boldsymbol{\nu}, \rho)_{\partial K}, \tag{2.7a}$$

$$L_h^2(u_h, \boldsymbol{s}_h; \boldsymbol{\theta}) := -(f(u_h)\boldsymbol{s}_h, \boldsymbol{\theta}), \tag{2.7b}$$

$$L_h^3(p_h; \boldsymbol{\eta}) := (p_h, \nabla \cdot \boldsymbol{\eta}) - \sum_{K\in\mathcal{T}_h} (\widehat{p}_h, \boldsymbol{\nu} \cdot \boldsymbol{\eta})_{\partial K}, \tag{2.7c}$$

$$L_h^4(u_h; \varphi) := -\left(\phi(\boldsymbol{x}) + H'(u_h), \varphi\right). \tag{2.7d}$$

Note that $\boldsymbol{\nu}$ is the unit outward normal vector of element K at its boundary ∂K. The "hat" terms in L_h^1 and L_h^3 are the so-called "numerical fluxes", whose choices play an important role in ensuring stability. We remark that the choices for the numerical fluxes are not unique. Here we use the alternating numerical fluxes

$$\widehat{\boldsymbol{q}}_h = \boldsymbol{q}_h^R, \quad \widehat{p}_h = p_h^L, \tag{2.8}$$

or

$$\widehat{\boldsymbol{q}}_h = \boldsymbol{q}_h^L, \quad \widehat{p}_h = p_h^R. \tag{2.9}$$

Considering the zero-flux boundary condition $\nabla(\phi(\boldsymbol{x}) + H'(u)) \cdot \boldsymbol{\nu} = 0$, we take

$$\widehat{\boldsymbol{q}}_h \cdot \boldsymbol{\nu} = 0, \quad p_h = (p_h)^{in} \tag{2.10}$$

at $\partial\Omega$, where "in" refers to the value obtained by taking the boundary trace from the inside of the domain Ω.

2.2.3 Entropy dissipation

Theorem 2.2.1. *For $u_h(t) \in V_h^k$, $\boldsymbol{s}_h \in \boldsymbol{W}_h^k$, the LDG scheme (2.6)-(2.10) with f satisfying (2.3) is entropy dissipative and satisfies*

$$\frac{d}{dt}E(u_h) = -(f(u_h)\boldsymbol{s}_h, \boldsymbol{s}_h) \leqslant 0,$$

which is consistent with the entropy dissipation property (2.5) of the PDE (2.1).

Proof. By taking

$$\rho = p_h, \quad \boldsymbol{\theta} = -\boldsymbol{s}_h, \quad \boldsymbol{\eta} = \boldsymbol{q}_h, \quad \varphi = -u_{ht},$$

in (2.6a)-(2.6d), respectively, and after integration by parts, we have

$$\begin{aligned}(\phi(\boldsymbol{x}) + H'(u_h), u_{ht}) =& -(f(u_h)\boldsymbol{s}_h, \boldsymbol{s}_h) - (\boldsymbol{q}_h, \nabla p_h) + \sum_{K\in\mathcal{T}_h} (\widehat{\boldsymbol{q}}_h \cdot \boldsymbol{\nu}, p_h)_{\partial K} \\ & - (p_h, \nabla \cdot \boldsymbol{q}_h) + \sum_{K\in\mathcal{T}_h} (\widehat{p}_h, \boldsymbol{\nu} \cdot \boldsymbol{q}_h)_{\partial K} \\ =& -(f(u_h)\boldsymbol{s}_h, \boldsymbol{s}_h) - \sum_{K\in\mathcal{T}_h} (\boldsymbol{q}_h \cdot \boldsymbol{\nu}, p_h)_{\partial K} \\ & + \sum_{K\in\mathcal{T}_h} (\widehat{\boldsymbol{q}}_h \cdot \boldsymbol{\nu}, p_h)_{\partial K} + \sum_{K\in\mathcal{T}_h} (\widehat{p}_h, \boldsymbol{\nu} \cdot \boldsymbol{q}_h)_{\partial K}. \end{aligned} \tag{2.11}$$

Assume that e is an interior edge shared by elements K_L and K_R, then $\boldsymbol{\nu}^R = -\boldsymbol{\nu}^L$. Replacing $\boldsymbol{\nu}^R$ with $\boldsymbol{\nu}^L$ and using the numerical fluxes (2.8), we obtain

$$\begin{aligned} & - \sum_{K_L \bigcup K_R} (\boldsymbol{q}_h \cdot \boldsymbol{\nu}, p_h)_e + \sum_{K_L \bigcup K_R} (\widehat{\boldsymbol{q}}_h \cdot \boldsymbol{\nu}, p_h)_e + \sum_{K_L \bigcup K_R} (\widehat{p}_h, \boldsymbol{\nu} \cdot \boldsymbol{q}_h)_e \\ =& -(\boldsymbol{q}_h^L \cdot \boldsymbol{\nu}^L, p_h^L)_e + (\boldsymbol{q}_h^R \cdot \boldsymbol{\nu}^L, p_h^R)_e + (\boldsymbol{q}_h^R \cdot \boldsymbol{\nu}^L, p_h^L)_e - (\boldsymbol{q}_h^R \cdot \boldsymbol{\nu}^L, p_h^R)_e \\ & + (\boldsymbol{q}_h^L \cdot \boldsymbol{\nu}^L, p_h^L)_e - (\boldsymbol{q}_h^R \cdot \boldsymbol{\nu}^L, p_h^L)_e = 0. \end{aligned} \tag{2.12}$$

Combining (2.11)-(2.12), using (2.4), boundary conditions (2.10) and the condition on f (2.3), we get

$$\frac{d}{dt}E(u_h) = (\phi(\boldsymbol{x}) + H'(u_h), u_{ht}) = -(f(u_h)\boldsymbol{s}_h, \boldsymbol{s}_h) \leqslant 0.$$

□

Remark 2.2.2. *For brevity, we will only consider in the remaining article the numerical fluxes (2.8) and omit the discussion of the numerical fluxes (2.9), but all results also apply to the numerical fluxes (2.9).*

Remark 2.2.3. *Compared to the spatial discretizations in [83, 131], we choose the simpler alternating numerical fluxes (2.8) and (2.9), which greatly simplifies the theoretical analysis of the entropy dissipation property of the LDG discretization.*

2.3 Time-implicit LDG schemes

The numerical discretization of the nonlinear parabolic equations (2.1) using explicit time discretization methods suffers from the rather severe time-step constraint $\tau = O(h^2)$. In this section, we will discuss therefore implicit time discretizations that will be coupled with positivity constraints in Section 2.4.

We divide the time interval $[0, T]$ into N parts $0 = t_0 < t_1 < \ldots < t_N = T$, with $\tau^n = t_n - t_{n-1}$ $(n = 1, 2, \ldots, N)$. For $n = 0, 1, \ldots, N$, let $u_n = u(\cdot, t_n)$ and u_h^n, respectively, denote the exact and approximate values of u at time t_n.

2.3.1 Backward Euler LDG discretization

Discretizing (2.6) in time with the implicit Euler method gives the following discrete system

$$\left(\frac{u_h^{n+1} - u_h^n}{\tau^{n+1}}, \rho\right) + L_h^1(\boldsymbol{q}_h^{n+1}; \rho) = 0, \tag{2.13a}$$

$$(\boldsymbol{q}_h^{n+1}, \boldsymbol{\theta}) + L_h^2(u_h^{n+1}, \boldsymbol{s}_h^{n+1}; \boldsymbol{\theta}) = 0, \tag{2.13b}$$

$$(\boldsymbol{s}_h^{n+1}, \boldsymbol{\eta}) + L_h^3(p_h^{n+1}; \boldsymbol{\eta}) = 0, \tag{2.13c}$$

$$(p_h^{n+1}, \varphi) + L_h^4(u_h^{n+1}; \varphi) = 0. \tag{2.13d}$$

Define the discrete entropy as

$$E_h(u_h^n) = \int_\Omega (u_h^n \phi(\boldsymbol{x}) + H(u_h^n)) d\Omega. \tag{2.14}$$

We have the following relation for the discrete entropy dissipation.

Theorem 2.3.1. *For all time levels n, the numerical solutions $u_h^n,\ u_h^{n+1} \in V_h^k$ of the LDG discretization (2.13), with boundary condition (2.10) and*

conditions on f, H stated in (2.3), satisfy the following entropy dissipation relation

$$E_h(u_h^{n+1}) \leqslant E_h(u_h^n), \tag{2.15}$$

which implies that the LDG discretization is unconditionally entropy dissipative.

Proof. By choosing, respectively, in (2.13a)-(2.13d) the following test functions

$$\rho = p_h^{n+1}, \quad \boldsymbol{\theta} = -\boldsymbol{s}_h^{n+1}, \quad \boldsymbol{\eta} = \boldsymbol{q}_h^{n+1}, \quad \varphi = -\frac{u_h^{n+1} - u_h^n}{\tau^{n+1}},$$

we get

$$\begin{aligned}
&\left(\phi(\boldsymbol{x}), \frac{u_h^{n+1} - u_h^n}{\tau^{n+1}}\right) + \left(H'(u_h^{n+1}), \frac{u_h^{n+1} - u_h^n}{\tau^{n+1}}\right) \\
=& - \left(f(u_h^{n+1})\boldsymbol{s}_h^{n+1}, \boldsymbol{s}_h^{n+1}\right) - \left(\boldsymbol{q}_h^{n+1}, \nabla p_h^{n+1}\right) + \sum_{K\in\mathcal{T}_h} (\widehat{\boldsymbol{q}}_h^{n+1} \cdot \boldsymbol{\nu}, p_h^{n+1})_{\partial K} \\
&- (p_h^{n+1}, \nabla \cdot \boldsymbol{q}_h^{n+1}) + \sum_{K\in\mathcal{T}_h} (\widehat{p}_h^{n+1}, \boldsymbol{\nu} \cdot \boldsymbol{q}_h^{n+1})_{\partial K} \\
=& - (f(u_h^{n+1})\boldsymbol{s}_h^{n+1}, \boldsymbol{s}_h^{n+1}) - \sum_{K\in\mathcal{T}_h} (\boldsymbol{q}_h^{n+1} \cdot \boldsymbol{\nu}, p_h^{n+1})_{\partial K} \\
&+ \sum_{K\in\mathcal{T}_h} (\widehat{\boldsymbol{q}}_h^{n+1} \cdot \boldsymbol{\nu}, p_h^{n+1})_{\partial K} + \sum_{K\in\mathcal{T}_h} (\widehat{p}_h^{n+1}, \boldsymbol{\nu} \cdot \boldsymbol{q}_h^{n+1})_{\partial K}.
\end{aligned}$$

Together with (2.12), the numerical fluxes (2.8) and the boundary condition (2.10), we obtain then

$$\left(\phi(\boldsymbol{x}), \frac{u_h^{n+1} - u_h^n}{\tau^{n+1}}\right) + \left(H'(u_h^{n+1}), \frac{u_h^{n+1} - u_h^n}{\tau^{n+1}}\right) = - \left(f(u_h^{n+1})\boldsymbol{s}_h^{n+1}, \boldsymbol{s}_h^{n+1}\right).$$

In view of the following Taylor expansion

$$\begin{aligned}
H(u_h^n) =& H(u_h^{n+1}) + H'(u_h^{n+1})(u_h^n - u_h^{n+1}) \\
&+ \frac{1}{2} H''(\xi^{n+1})(u_h^{n+1} - u_h^n)^2, \quad \xi^{n+1} \in (u_h^n, u_h^{n+1}),
\end{aligned}$$

we have, using the conditions on f, H stated in (2.3) and the definition of E_h in (2.14),

$$\begin{aligned}
&E_h(u_h^{n+1}) - E_h(u_h^n) = \left(\phi(\boldsymbol{x}), u_h^{n+1} - u_h^n\right) + \left(H(u_h^{n+1}) - H(u_h^n), 1\right) \\
=& - \tau^{n+1} \left(f(u_h^{n+1})\boldsymbol{s}_h^{n+1}, \boldsymbol{s}_h^{n+1}\right) - \frac{1}{2} \left(H''(\xi^{n+1}), \left(u_h^{n+1} - u_h^n\right)^2\right) \leqslant 0.
\end{aligned}$$

□

2.3.2 Higher order DIRK-LDG discretizations

For higher order accurate implicit in time discretizations of system (2.6), we use a Diagonally Implicit Runge-Kutta (DIRK) method [61]. Assume we know the numerical solution at time level n, we obtain the solution at time level $n+1$ with a DIRK method by solving for each DIRK stage $i, i = 1, 2, \cdots, s$ the following equations.

$$\left(\frac{u_h^{n+1,i} - u_h^n}{\tau^{n+1}}, \rho\right) + \sum_{j=1}^{i} a_{ij} L_h^1(\boldsymbol{q}_h^{n+1,j}; \rho) = 0, \tag{2.16a}$$

$$(\boldsymbol{q}_h^{n+1,i}, \boldsymbol{\theta}) + L_h^2(u_h^{n+1,i}, \boldsymbol{s}_h^{n+1,i}; \boldsymbol{\theta}) = 0, \tag{2.16b}$$

$$(\boldsymbol{s}_h^{n+1,i}, \boldsymbol{\eta}) + L_h^3(p_h^{n+1,i}; \boldsymbol{\eta}) = 0, \tag{2.16c}$$

$$(p_h^{n+1,i}, \varphi) + L_h^4(u_h^{n+1,i}; \varphi) = 0. \tag{2.16d}$$

Then the solution at time t_{n+1} is

$$(u_h^{n+1}, \rho) = (u_h^n, \rho) - \tau \sum_{i=1}^{s} b_i L_h^1(\boldsymbol{q}_h^{n+1,i}; \rho). \tag{2.17}$$

The coefficient matrices (a_{ij}) in (2.16a) and (b_i) in (2.17) are defined in the Butcher tableau. We choose for polynomials of order $k = 1, 2, 3$ the DIRK methods introduced in Section 1.3.2.2, respectively, that satisfy $a_{si} = b_i,\ i = 1, 2, \cdots, s$, which implies $u_h^{n+1} = u_h^{n+1,s}$. The order of these DIRK methods is $k+1$. The above time discretization methods are easy to implement since the matrix (a_{ij}) in the DIRK methods has a lower triangular structure, which means that we can compute the DIRK stages one after another, starting from $i = 1$ up to $i = s$. For detailed information about the DIRK time integration method, we refer to [61].

2.4 Higher order accurate positivity preserving DIRK-LDG discretizations

The positivity constraints on the LDG solution will be enforced by transforming the DIRK-LDG equations with positivity constraints into a mixed complementarity problem using the Karush-Kuhn-Tucker (KKT) equations [41]. In the next sections, we will first define the positivity preserving KKT-DIRK-LDG discretization. Next, we will consider the unique solvability and unconditional entropy dissipation of the discrete KKT system.

2.4.1 KKT-system

For the KKT equations [41], we define the set

$$\mathbb{K} := \{\widetilde{U} \in \mathbb{R}^{dof} |\ h(\widetilde{U}) = 0,\ g(\widetilde{U}) \leqslant 0\}, \tag{2.18}$$

with equality constraints $h : \mathbb{R}^{dof} \to \mathbb{R}^l$ and inequality constraints $g : \mathbb{R}^{dof} \to \mathbb{R}^m$ being vector-valued continuously differentiable functions. The inequality constraints are used to ensure positivity. The equality constraint ensures that the limited DIRK-LDG discretization is mass conservative. Mass conservation is a property of the unlimited DIRK-LDG discretization, but one has to ensure that this property also holds after applying the positivity preserving limiter.

Let L be the LDG discretization (2.16) for each of the DIRK stages $i = 1, 2, \cdots, s$, without a positivity preserving limiter. We assume that L is a continuously differentiable function from $\mathbb{K}$ to $\mathbb{R}^{dof}$. The corresponding KKT-system [41] then is

$$L(\widetilde{U}) + \nabla_{\widetilde{U}} h(\widetilde{U})^T \mu + \nabla_{\widetilde{U}} g(\widetilde{U})^T \lambda = 0, \tag{2.19a}$$

$$-h(\widetilde{U}) = 0, \tag{2.19b}$$

$$0 \geqslant g(\widetilde{U}) \perp \lambda \geqslant 0, \tag{2.19c}$$

where $\mu \in \mathbb{R}^l$ and $\lambda \in \mathbb{R}^m$ are the Lagrange multipliers used to ensure $h(\widetilde{U}) = 0$ and $g(\widetilde{U}) \leqslant 0$, respectively, $\widetilde{U} \in R^{dof}$ are the LDG coefficients in the KKT-DIRK-LDG discretization, and $\nabla_{\widetilde{U}}$ denotes the gradient with respect to $\widetilde{U}$. The compatibility condition (2.19c) is equivalent to

$$g_j(\widetilde{U}) \leqslant 0, \quad \lambda_j \geqslant 0, \quad \text{and} \quad g_j(\widetilde{U})\lambda_j = 0, \quad j = 1, 2, \cdots, m,$$

which can be expressed as

$$\min(-g_j(\widetilde{U}), \lambda_j) = 0, \quad j = 1, 2, \cdots, m.$$

The KKT-system then can be formulated as

$$0 = F(z) = \begin{pmatrix} L(\widetilde{U}) + \nabla_{\widetilde{U}} h(\widetilde{U})^T \mu + \nabla_{\widetilde{U}} g(\widetilde{U})^T \lambda \\ -h(\widetilde{U}) \\ \min(-g(\widetilde{U}), \lambda) \end{pmatrix}. \tag{2.20}$$

Here $z = (\widetilde{U}, \mu, \lambda) \in \mathbb{R}^{dof+l+m}$, and $F : \mathbb{R}^{dof+l+m} \to \mathbb{R}^{dof+l+m}$ represents the DIRK-LDG discretization combined with the positivity and mass conservation constraints. Note, the KKT system (2.20) is nonlinear and $F(z)$ is not continuously differentiable, as is necessary for standard Newton methods, but semi-smooth. We will therefore solve (2.20) with the active set semi-smooth Newton method presented in [111].

2.4.2 Positivity preserving LDG discretizations

In this section, we will provide the details of the higher order accurate positivity preserving DIRK-LDG discretizations (2.16) coupled with the positivity and mass conservation constraints using Lagrange multipliers as stated in (2.19).

Let N_k be the number of basis functions in one element. Let N_e be the number of elements K in the tessellation $\mathcal{T}_h$ of the domain Ω. We introduce the following notation for the element-wise positivity preserving LDG solution

$$U_h|_K := \sum_{j=1}^{N_k} \widetilde{U}_j^K \phi_j^K, \quad \boldsymbol{Q}_h|_K := \sum_{j=1}^{N_k} \widetilde{\boldsymbol{Q}}_j^K \phi_j^K$$

with $K \in \mathcal{T}_h$, ϕ_j^K the tensor product Legendre basis functions in $\mathcal{Q}_k(K)$, and LDG coefficients $\widetilde{U}_j^K \in \mathbb{R}$, $\widetilde{\boldsymbol{Q}}_j^K \in \mathbb{R}^d$. We take in each element $K \in \mathcal{T}_h$ the test function $\rho = \phi_j^K,\ j = 1, 2, \cdots, N_k$ in the operator $L_h^1(\boldsymbol{Q}_h; \rho)$ stated in (2.7a). Since there are $N_k N_e$ choices of ρ, we can define

$$\mathbb{L}_h^1(\widetilde{\boldsymbol{Q}}) := L_h^1(\boldsymbol{Q}_h; \rho) \in \mathbb{R}^{N_k N_e}, \tag{2.21}$$

with similar definitions of $\mathbb{L}_h^k$ for $L_h^k, k = 2, 3, 4$ stated in (2.7b)-(2.7d).

Representing the block-diagonal mass matrices for the scalar and vector variables as $M \in \mathbb{R}^{N_k N_e \times N_k N_e}$ and $\boldsymbol{M} \in \mathbb{R}^{dN_k N_e \times dN_k N_e}$, respectively, the operator L for DIRK stage i ($i = 1, 2, \cdots, s$), as stated in (2.16a), can be expressed as

$$L(\widetilde{U}^{n+1,i}) := M(\widetilde{U}^{n+1,i} - \widetilde{U}^n) + \tau^{n+1} \sum_{j=1}^{i} a_{ij} \mathbb{L}_h^1(\widetilde{\boldsymbol{Q}}^{n+1,j}), \tag{2.22}$$

with LDG coefficients $\widetilde{U}^{n+1,i} \in \mathbb{R}^{N_k N_e}$. Similarly, using (2.16b), (2.16c) and (2.16d), we have

$$\widetilde{\boldsymbol{Q}}^{n+1,i} = -\boldsymbol{M}^{-1} \mathbb{L}_h^2(\widetilde{U}^{n+1,i}, \widetilde{\boldsymbol{S}}^{n+1,i}), \tag{2.23a}$$

$$\widetilde{\boldsymbol{S}}^{n+1,i} = -\boldsymbol{M}^{-1} \mathbb{L}_h^3(\widetilde{P}^{n+1,i}), \tag{2.23b}$$

$$\widetilde{P}^{n+1,i} = -M^{-1} \mathbb{L}_h^4(\widetilde{U}^{n+1,i}), \tag{2.23c}$$

with LDG coefficients $\widetilde{\boldsymbol{Q}}^{n+1,i} \in \mathbb{R}^{dN_k N_e}$, $\widetilde{\boldsymbol{S}}^{n+1,i} \in \mathbb{R}^{dN_k N_e}$, $\widetilde{P}^{n+1,i} \in \mathbb{R}^{N_k N_e}$.

The constraints on the DIRK-LDG discretization can be directly imposed on the DG coefficients for each DIRK stage using the equality and inequality constraints in the KKT-system (2.20). We obtain for each DIRK stage i, with $i = 1, 2, \cdots, s$, the LDG coefficients $\widetilde{U}^{n+1,i}$ by solving the following KKT system for $\widetilde{U}^{n+1,i}$,

$$\begin{pmatrix} L(\widetilde{U}^{n+1,i}) + \nabla_{\widetilde{U}} h(\widetilde{U}^{n+1,i})^T \mu + \nabla_{\widetilde{U}} g(\widetilde{U}^{n+1,i})^T \lambda \\ -h(\widetilde{U}^{n+1,i}) \\ \min(-g(\widetilde{U}^{n+1,i}), \lambda) \end{pmatrix} = 0, \tag{2.24}$$

where the positivity preserving inequality constraint $g(\widetilde{U}^{n+1,i})$ and the mass conservation equality constraint $h(\widetilde{U}^{n+1,i})$ are defined as follows.

1. Positivity preserving inequality constraint

In each element $K \in \mathcal{T}_h$, we define the function g stated in (2.24) as

$$g_p^K(\widetilde{U}^{n+1,i}) = u_{\min} - \sum_{j=1}^{N_k} \widetilde{U}_j^{K,(n+1,i)} \phi_j^K(\boldsymbol{x}_p), \quad p = 1, \cdots, N_p, \tag{2.25}$$

with N_p the number of Gauss-Lobatto quadrature points, and $\boldsymbol{x}_p$ the Gauss-Lobatto quadrature points where the inequality constraints $U_h(\boldsymbol{x}_p) \geqslant u_{\min}$ are imposed. The use of Gauss-Lobatto quadrature rules ensures that the positivity constraint is also imposed in the computation of the numerical fluxes at the element edges where Gauss-Lobatto rules have, next to the element itself, also quadrature points. Note, the Gauss-Lobatto quadrature points $\boldsymbol{x}_p$ are the only points used in the LDG discretization and the positivity constraint $u_{\min}$ therefore only needs to be enforced at these points.

2. Mass conservation equality constraint

In order to ensure mass conservation of the LDG discretization when the positivity constraint is enforced, we impose the following equality constraint, which is obtained by setting $\rho = 1$ in (2.16a) and using the numerical flux (2.8) or (2.9)

$$\begin{aligned} h(\widetilde{U}^{n+1,i}) = \sum_{K \in \mathcal{T}_h} \int_K U_h^n dK + \tau^{n+1} \sum_{j=1}^{i} a_{ij} \sum_{\substack{K \in \mathcal{T}_h \\ \partial K \cap \partial\Omega \neq \emptyset}} (\widehat{\boldsymbol{Q}}_h^{n+1,j} \cdot \boldsymbol{\nu}, 1)_{\partial K} \\ - \sum_{K \in \mathcal{T}_h} \sum_{j=1}^{N_k} \widetilde{U}_j^{K,(n+1,i)} \int_K \phi_j^K(\boldsymbol{x}) dK, \end{aligned} \tag{2.26}$$

with U_h^n the KKT-DIRK-LDG solution at time t_n.

For each DIRK stage i, the KKT-system (2.24) for the higher order accurate positivity preserving LDG discretization is now defined. After solving the KKT equations (2.24) for $i = 1, \cdots, s$, the numerical solution at time t^{n+1} is directly obtained from the last DIRK stage, $U_h^{n+1} = U_h^{n+1,s}$ since we use DIRK methods with $a_{si} = b_i$.

Remark 2.4.1. *In order to ensure the positivity of the discrete initial solution U_h^0, we use the L^2-projection coupled with the positivity constraint (2.25), which is obtained by replacing $\widetilde{U}^{n+1,i}$ with $\widetilde{U}^0$. Mass conservation of the positivity limited initial solution is ensured by the equality constraint*

$$h(\widetilde{U}^0) = \sum_{K\in\mathcal{T}_h} \int_K u_0(\boldsymbol{x})dK - \sum_{K\in\mathcal{T}_h} \sum_{j=1}^{N_k} \widetilde{U}_j^{K,0} \int_K \phi_j^K(\boldsymbol{x})dK.$$

The constraints on the L^2-projection are imposed using KKT equations similar to (2.20). In order to prevent pathological cases, we assume that the limited initial solution satisfies

$$\frac{1}{|\Omega|} \sum_{K\in\mathcal{T}_h} \int_K u_0(\boldsymbol{x})dK \geqslant u_{\min}.$$

Remark 2.4.2. *We emphasize that $u_{\min}$ must be chosen strictly positive to ensure that the positivity of the numerical solution is not violated by errors due to the finite precision of the computer arithmetic.*

2.4.3 Unique solvability and stability of the positivity preserving LDG discretization

In Section 2.4.2, we have presented the positivity preserving LDG discretization for (2.1). In this section, we will consider the unique solvability of the algebraic equations resulting from the backward Euler KKT-LDG discretization. In the theoretical analysis we will also consider the entropy dissipation of the positivity preserving backward Euler LDG discretization and use periodic boundary conditions.

With (2.22)-(2.26), the positivity preserving backward Euler LDG discretization results now in the following KKT system,

$$L(\widetilde{U}^{n+1}) + \nabla_{\widetilde{U}} h(\widetilde{U}^{n+1})^T \mu^{n+1} + \nabla_{\widetilde{U}} g(\widetilde{U}^{n+1})^T \lambda^{n+1} = 0, \tag{2.27a}$$

$$-h(\widetilde{U}^{n+1}) = 0, \tag{2.27b}$$

$$\min(-g(\widetilde{U}^{n+1}), \lambda^{n+1}) = 0. \tag{2.27c}$$

Here $L : \mathbb{R}^{N_k N_e} \to \mathbb{R}^{N_k N_e}$ and

$$L(\widetilde{U}^{n+1}) := M(\widetilde{U}^{n+1} - \widetilde{U}^n) + \tau^{n+1} B\widetilde{\boldsymbol{Q}}^{n+1}, \tag{2.28}$$

$$\boldsymbol{M}\widetilde{\boldsymbol{Q}}^{n+1} = C_d(\widetilde{U}^{n+1})\widetilde{\boldsymbol{S}}^{n+1}, \tag{2.29}$$

$$\boldsymbol{M}\widetilde{\boldsymbol{S}}^{n+1} = A\widetilde{P}^{n+1}, \tag{2.30}$$

$$M\widetilde{P}^{n+1} = D(\widetilde{U}^{n+1}). \tag{2.31}$$

From (2.21)-(2.23), we obtain that

$$B\widetilde{\boldsymbol{Q}}^{n+1} = \mathbb{L}_h^1(\widetilde{\boldsymbol{Q}}^{n+1}) \in \mathbb{R}^{N_k N_e}, \tag{2.32}$$

$$C_d(\widetilde{U}^{n+1})\widetilde{\boldsymbol{S}}^{n+1} = -\mathbb{L}_h^2(\widetilde{U}^{n+1}, \widetilde{\boldsymbol{S}}^{n+1}) \in \mathbb{R}^{dN_k N_e}, \tag{2.33}$$

$$A\widetilde{P}^{n+1} = -\mathbb{L}_h^3(\widetilde{P}^{n+1}) \in \mathbb{R}^{dN_k N_e}, \tag{2.34}$$

$$D(\widetilde{U}^{n+1}) = -\mathbb{L}_h^4(\widetilde{U}^{n+1}) \in \mathbb{R}^{N_k N_e}, \tag{2.35}$$

where

$$C_d(\widetilde{U}^{n+1}) = \begin{pmatrix} C(\widetilde{U}^{n+1}) & & \\ & \ddots & \\ & & C(\widetilde{U}^{n+1}) \end{pmatrix} \in \mathbb{R}^{dN_k N_e \times dN_k N_e}, \tag{2.36}$$

$$C(\widetilde{U}^{n+1}) \in \mathbb{R}^{N_k N_e}. \tag{2.37}$$

The constraints $h : \mathbb{R}^{N_k N_e} \to \mathbb{R},\ g : \mathbb{R}^{N_k N_e} \to \mathbb{R}^{N_p N_e}$ are defined by

$$h(\widetilde{U}^{n+1}) := \sum_{K \in \mathcal{T}_h} \int_K U_h^0 dK - \sum_{K \in \mathcal{T}_h} \sum_{j=1}^{N_k} \widetilde{U}_j^{K,(n+1)} \int_K \phi_j^K(\boldsymbol{x}) dK, \tag{2.38}$$

$$g(\widetilde{U}^{n+1}) := (g_1^{K_1}(\widetilde{U}^{n+1}), \cdots, g_{N_p}^{K_1}(\widetilde{U}^{n+1}), \cdots, g_1^{K_{N_e}}(\widetilde{U}^{n+1}), \cdots, g_{N_p}^{K_{N_e}}(\widetilde{U}^{n+1})), \tag{2.39}$$

with the definition of the constraints $g_p^{K_j},\ 1 \leqslant p \leqslant N_p, 1 \leqslant j \leqslant N_e$ given in (2.25).

2.4.3.1 Auxiliary results used to prove the solvability of the KKT-system

In this section, we will introduce some auxiliary results, which will be used in Section 2.4.3.2 to prove the unique solvability of the KKT-system (2.27).

Definition 2.4.3. [41, Sections 1.1, 3.2] Let $\mathbb{K}$ be given by (2.18), given a map $L : \mathbb{K} \to \mathbb{R}^{dof}$. The Variational Inequality (VI($\mathbb{K}, L$)) is to find $\widetilde{U} \in \mathbb{K}$ such that

$$(y - \widetilde{U})^T L(\widetilde{U}) \geqslant 0, \quad y \in \mathbb{K}. \tag{2.40}$$

The solution of VI($\mathbb{K}, L$) (2.40) is denoted by SOL($\mathbb{K}, L$).

Using the nodal basis function and the definition of g in (2.39) and (2.25), the inequality constraint set in (2.18) can be written as

$$\mathbb{K}_b := \{\widetilde{U} \in \mathbb{R}^{dof} |\ \widetilde{U}_i^{\min} \leqslant \widetilde{U}_i \leqslant \widetilde{U}_i^{\max}, i \in \{1, \cdots, dof\}\}, \tag{2.41}$$

and we write $\mathbb{K}_b$ as

$$\mathbb{K}_b = \prod_{\vartheta=1}^{N} \mathbb{K}_{n_\vartheta}, \tag{2.42}$$

where $\mathbb{K}_{n_\vartheta}$ is a subset of $\mathbb{R}^{n_\vartheta}$ with $\sum_{\vartheta=1}^{N} n_\vartheta = dof$. Thus for a vector $\widetilde{U} \in \mathbb{K}_b$, we write $\widetilde{U} = (\widetilde{U}_\vartheta)$, where each $\widetilde{U}_\vartheta$ belongs to $\mathbb{K}^{n_\vartheta}$.

Definition 2.4.4. [41, Section 3.5.2] Let $\mathbb{K}_b$ be given by (2.41), a map $L : \mathbb{K}_b \to \mathbb{R}^{dof}$ is said to be

a) a P-function on $\mathbb{K}_b$ if for all pairs of distinct vectors $\widetilde{U}$ and $\widetilde{U}'$ in $\mathbb{K}_b$,

$$\max_{1 \leqslant \vartheta \leqslant N} (\widetilde{U}_\vartheta - \widetilde{U}'_\vartheta)^T (L_\vartheta(\widetilde{U}) - L_\vartheta(\widetilde{U}')) > 0,$$

b) a uniformly P-function on $\mathbb{K}_b$ if there exists a constant $\varpi > 0$ such that for all pairs of distinct vectors $\widetilde{U}$ and $\widetilde{U}'$ in $\mathbb{K}_b$,

$$\max_{1 \leqslant \vartheta \leqslant N} (\widetilde{U}_\vartheta - \widetilde{U}'_\vartheta)^T (L_\vartheta(\widetilde{U}) - L_\vartheta(\widetilde{U}')) \geqslant \varpi \|\widetilde{U} - \widetilde{U}'\|^2.$$

Lemma 2.4.5. *[41, Proposition 3.5.10] Let $\mathbb{K}_b$ be given by (2.41).*

a) If L is a P-function on $\mathbb{K}_b$, then VI($\mathbb{K}_b, L$) has at most one solution.

b) If each $\mathbb{K}_{n_\vartheta}$ is closed convex and L is a continuous uniformly P-function on $\mathbb{K}_b$, then the VI($\mathbb{K}_b, L$) has a unique solution.

Lemma 2.4.6. *[41, Proposition 1.3.4] Let $\widetilde{U} \in$ SOL($\mathbb{K}, L$) solve (2.40) with $\mathbb{K}$ given by (2.18). If Abadie's Constraint Qualification holds at $\widetilde{U}$, which means that the tangent cone of $\mathbb{K}$ at $\widetilde{U}$ is equal to its linearization*

cone, then there exist vectors $\mu \in \mathbb{R}^l$ *and* $\lambda \in \mathbb{R}^m$ *satisfying the KKT system (2.27).*

Conversely, if each function h_j $(1 \leqslant j \leqslant l)$ *is affine and each function* g_i $(1 \leqslant i \leqslant m)$ *is convex, and if* $(\widetilde{U}, \mu\, \lambda)$ *satisfies (2.27), then* $\widetilde{U}$ *solves VI*$(\mathbb{K}, L)$ *given by (2.40) with* $\mathbb{K}$ *given by (2.18).*

2.4.3.2 Existence and uniqueness of LDG discretization with positivity and mass conservation constraints

In this section, we will prove existence and uniqueness of the KKT system (2.27)-(2.39) using the unique solvability conditions discussed in Section 2.4.3.1.

Lemma 2.4.7. *For periodic boundary conditions, the matrices* B *in (2.32) and* A *in (2.34) satisfy* $B^T = A$.

Proof. In order to prove the symmetry of B in (2.32) and A in (2.34), we define the bilinear function $a : (V_h^k \times \boldsymbol{W}_h^k) \times (V_h^k \times \boldsymbol{W}_h^k) \to \mathbb{R}$ by

$$\begin{aligned} a(P_h^{n+1}, \boldsymbol{Q}_h^{n+1}; \rho, \boldsymbol{\theta}) =& (\boldsymbol{Q}_h^{n+1}, \nabla\rho) - \sum_{K\in\mathcal{T}_h} (\widehat{\boldsymbol{Q}}_h^{n+1} \cdot \boldsymbol{\nu}, \rho)_{\partial K} \\ & - (P_h^{n+1}, \nabla\cdot\boldsymbol{\theta}) + \sum_{K\in\mathcal{T}_h} (\widehat{P}_h^{n+1}, \boldsymbol{\nu}\cdot\boldsymbol{\theta})_{\partial K}. \end{aligned}$$

Based on the definition of B in (2.32) using (2.7a), A in (2.34) using (2.7c), we rewrite the above bilinear function a as follows:

$$a(P_h^{n+1}, \boldsymbol{Q}_h^{n+1}; \rho, \boldsymbol{\theta}) = (\varrho, \Theta) \begin{pmatrix} 0 & B \\ A & 0 \end{pmatrix} (\widetilde{P}^{n+1}, \widetilde{\boldsymbol{Q}}^{n+1})^T,$$

with ϱ, Θ the LDG coefficients of $\rho, \boldsymbol{\theta}$ and $\widetilde{P}^{n+1}, \widetilde{\boldsymbol{Q}}^{n+1}$ the LDG coefficients of $P_h^{n+1}, \boldsymbol{Q}_h^{n+1}$, respectively.

Interchanging the arguments of a, we get

$$\begin{aligned} a(\rho,\boldsymbol{\theta};P_h^{n+1},\boldsymbol{Q}_h^{n+1}) =&(\boldsymbol{\theta},\nabla P_h^{n+1}) - \sum_{K\in\mathcal{T}_h}(\widehat{\boldsymbol{\theta}}\cdot\boldsymbol{\nu},P_h^{n+1})_{\partial K}\\ &-(\rho,\nabla\cdot\boldsymbol{Q}_h^{n+1}) + \sum_{K\in\mathcal{T}_h}(\widehat{\rho},\boldsymbol{\nu}\cdot\boldsymbol{Q}_h^{n+1})_{\partial K}\\ =&-(P_h^{n+1},\nabla\cdot\boldsymbol{\theta}) + \sum_{K\in\mathcal{T}_h}(\boldsymbol{\theta}\cdot\boldsymbol{\nu},P_h^{n+1})_{\partial K}\\ &-\sum_{K\in\mathcal{T}_h}(\widehat{\boldsymbol{\theta}}\cdot\boldsymbol{\nu},P_h^{n+1})_{\partial K} + (\boldsymbol{Q}_h^{n+1},\nabla\rho)\\ &-\sum_{K\in\mathcal{T}_h}(\rho,\boldsymbol{\nu}\cdot\boldsymbol{Q}_h^{n+1})_{\partial K} + \sum_{K\in\mathcal{T}_h}(\widehat{\rho},\boldsymbol{\nu}\cdot\boldsymbol{Q}_h^{n+1})_{\partial K}, \end{aligned}$$

Using equality (2.12), the alternating numerical fluxes for $\widehat{\boldsymbol{\theta}}$ and $\widehat{\rho}$ in (2.8) or (2.9), and the periodic boundary conditions, we obtain

$$\begin{aligned} &\sum_{K\in\mathcal{T}_h}(\boldsymbol{\theta}\cdot\boldsymbol{\nu},P_h^{n+1})_{\partial K} - \sum_{K\in\mathcal{T}_h}(\widehat{\boldsymbol{\theta}}\cdot\boldsymbol{\nu},P_h^{n+1})_{\partial K} = \sum_{K\in\mathcal{T}_h}(\widehat{P}_h^{n+1},\boldsymbol{\nu}\cdot\boldsymbol{\theta})_{\partial K},\\ &-\sum_{K\in\mathcal{T}_h}(\rho,\boldsymbol{\nu}\cdot\boldsymbol{Q}_h^{n+1})_{\partial K} + \sum_{K\in\mathcal{T}_h}(\widehat{\rho},\boldsymbol{\nu}\cdot\boldsymbol{Q}_h^{n+1})_{\partial K} = -\sum_{K\in\mathcal{T}_h}(\widehat{\boldsymbol{Q}}_h^{n+1}\cdot\boldsymbol{\nu},\rho)_{\partial K}. \end{aligned}$$

Hence,

$$a(P_h^{n+1},\boldsymbol{Q}_h^{n+1};\rho,\boldsymbol{\theta}) = a(\rho,\boldsymbol{\theta};P_h^{n+1},\boldsymbol{Q}_h^{n+1}),$$

which implies

$$\begin{aligned} &(\varrho,\Theta)\begin{pmatrix}0 & B\\ A & 0\end{pmatrix}(\widetilde{P}^{n+1},\widetilde{\boldsymbol{Q}}^{n+1})^T = (\widetilde{P}^{n+1},\widetilde{\boldsymbol{Q}}^{n+1})\begin{pmatrix}0 & B\\ A & 0\end{pmatrix}(\varrho,\Theta)^T\\ =&(\varrho,\Theta)\begin{pmatrix}0 & A^T\\ B^T & 0\end{pmatrix}(\widetilde{P}^{n+1},\widetilde{\boldsymbol{Q}}^{n+1})^T. \end{aligned} \tag{2.43}$$

Since $(P_h^{n+1},\boldsymbol{Q}_h^{n+1})\in V_h^k\times\boldsymbol{W}_h^k$ and $(\rho,\boldsymbol{\theta})\in V_h^k\times\boldsymbol{W}_h^k$ are arbitrary functions, relation (2.43) implies that $A=B^T$. □

Using (2.29)-(2.31) and Lemma 2.4.7, the operator $L(\widetilde{U}^{n+1})$ in (2.28) can be written as

$$\begin{aligned} L(\widetilde{U}^{n+1}) =&M(\widetilde{U}^{n+1}-\widetilde{U}^n)\\ &+\tau^{n+1}B\boldsymbol{M}^{-1}C_d(\widetilde{U}^{n+1})\boldsymbol{M}^{-1}B^TM^{-1}D(\widetilde{U}^{n+1}). \end{aligned} \tag{2.44}$$

Lemma 2.4.8. *Given $\widetilde{U}^n$, the operator L in (2.44) is a uniformly P-function on $\mathbb{K}_b$.*

Proof. Using relation (2.44) for L, for arbitrary $\widetilde{U}_I^{n+1}, \widetilde{U}_{II}^{n+1} \in \mathbb{K}_b$, there holds

$$\begin{aligned} &L(\widetilde{U}_I^{n+1}) - L(\widetilde{U}_{II}^{n+1}) = M(\widetilde{U}_I^{n+1} - \widetilde{U}_{II}^{n+1}) \\ &+ \tau^{n+1} B\boldsymbol{M}^{-1} C_d(\widetilde{U}_I^{n+1})\boldsymbol{M}^{-1} B^T M^{-1} D(\widetilde{U}_I^{n+1}) \\ &- \tau^{n+1} B\boldsymbol{M}^{-1} C_d(\widetilde{U}_{II}^{n+1})\boldsymbol{M}^{-1} B^T M^{-1} D(\widetilde{U}_{II}^{n+1}). \end{aligned} \tag{2.45}$$

After subtracting and adding $\tau^{n+1} B\boldsymbol{M}^{-1} C_d(\widetilde{U}_I^{n+1})\boldsymbol{M}^{-1} B^T M^{-1} D(\widetilde{U}_{II}^{n+1})$ in (2.45), we obtain

$$\begin{aligned} &L(\widetilde{U}_I^{n+1}) - L(\widetilde{U}_{II}^{n+1}) = M(\widetilde{U}_I^{n+1} - \widetilde{U}_{II}^{n+1}) \\ &+ \tau^{n+1} B\boldsymbol{M}^{-1} C_d(\widetilde{U}_I^{n+1})\boldsymbol{M}^{-1} B^T M^{-1} (D(\widetilde{U}_I^{n+1}) - D(\widetilde{U}_{II}^{n+1})) \\ &+ \tau^{n+1} B\boldsymbol{M}^{-1} (C_d(\widetilde{U}_I^{n+1}) - C_d(\widetilde{U}_{II}^{n+1}))\boldsymbol{M}^{-1} B^T M^{-1} D(\widetilde{U}_{II}^{n+1}). \end{aligned} \tag{2.46}$$

For $i \in \{1, \cdots, N_k N_e\}$, with the definition of D in (2.35) using (2.7d), we obtain that

$$\begin{aligned} &(D(\widetilde{U}_I^{n+1}) - D(\widetilde{U}_{II}^{n+1}))_i \\ &= \int_\Omega \left(H'\left(\sum_{j=1}^{N_k N_e} \widetilde{U}_{I,j}^{n+1} \phi_j \right) - H'\left(\sum_{j=1}^{N_k N_e} \widetilde{U}_{II,j}^{n+1} \phi_j \right) \right) \phi_i d\Omega \\ &= \sum_{j=1}^{N_k N_e} (\widetilde{U}_{I,j}^{n+1} - \widetilde{U}_{II,j}^{n+1}) \int_\Omega H''(\xi_1^{n+1}) \phi_j \phi_i d\Omega,\ \xi_1^{n+1} \in (U_{h,I}^{n+1}, U_{h,II}^{n+1}), \end{aligned}$$

and write

$$D(\widetilde{U}_I^{n+1}) - D(\widetilde{U}_{II}^{n+1}) := D_{\widetilde{U}}(\xi_1^{n+1})(\widetilde{U}_I^{n+1}) - \widetilde{U}_{II}^{n+1}). \tag{2.47}$$

Similarly, for $i, j, k \in \{1, \cdots, N_k N_e\}$, from the definition of C_d in (2.33),

(2.36) using (2.7b), we obtain that

$$
\begin{aligned}
&C_d(\widetilde{U}_I^{n+1}) - C_d(\widetilde{U}_{II}^{n+1}) \\
&= \begin{pmatrix} C(\widetilde{U}_I^{n+1}) - C(\widetilde{U}_{II}^{n+1}) & & \\ & \ddots & \\ & & C(\widetilde{U}_I^{n+1}) - C(\widetilde{U}_{II}^{n+1}) \end{pmatrix},
\end{aligned}
$$

$$
\begin{aligned}
&(C(\widetilde{U}_I^{n+1}) - C(\widetilde{U}_{II}^{n+1}))_{ij} \\
&= \int_\Omega \left(f\left(\sum_{k=1}^{N_k N_e} \widetilde{U}_{I,k}^{n+1} \phi_k \right) - f\left(\sum_{k=1}^{N_k N_e} \widetilde{U}_{II,k}^{n+1} \phi_k \right) \right) \phi_j \phi_i d\Omega \\
&= \sum_{k=1}^{N_k N_e} (\widetilde{U}_{I,k}^{n+1} - \widetilde{U}_{II,k}^{n+1}) \int_\Omega f'(\xi_2^{n+1}) \phi_k \phi_j \phi_i d\Omega,\ \xi_2^{n+1} \in (U_{h,I}^{n+1}, U_{h,II}^{n+1}),
\end{aligned}
$$

and write

$$
C(\widetilde{U}_I^{n+1}) - C(\widetilde{U}_{II}^{n+1}) := \sum_{k=1}^{N_k N_e} [C_{d\widetilde{U}}(\xi_2^{n+1})]_k (\widetilde{U}_{I,k}^{n+1}) - \widetilde{U}_{II,k}^{n+1}). \tag{2.48}
$$

In order to estimate (2.46), with (2.47)-(2.48), we assume for arbitrary $\widetilde{U} \in \mathbb{K}_b$ in (2.41), that

$$
\begin{aligned}
|C(\widetilde{U})_{ij}| \leqslant & c, \quad |D(\widetilde{U})_i| \leqslant c, \\
|[C_{\widetilde{U}}(\widetilde{U})_{ij}]_k| \leqslant & c, \quad |D_{\widetilde{U}}(\widetilde{U})_{ij}| \leqslant c,\ i,j,k \in \{1, \cdots, N_k N_e\},
\end{aligned} \tag{2.49}
$$

with c a positive constant, independent of $\widetilde{U}$. In the remainder of this section c is a positive constant, but not necessarily the same.

Using (2.47)-(2.48) and assumption (2.49), we obtain the following two estimates

$$
\begin{aligned}
&(\widetilde{U}_I^{n+1} - \widetilde{U}_{II}^{n+1})^T B \boldsymbol{M}^{-1} C_d(\widetilde{U}_I^{n+1}) \boldsymbol{M}^{-1} B^T M^{-1} (D(\widetilde{U}_I^{n+1}) - D(\widetilde{U}_{II}^{n+1})) \\
&\leqslant \|B\| \|\boldsymbol{M}^{-1}\| \|C_d(\widetilde{U}_I^{n+1})\| \|\boldsymbol{M}^{-1}\| \|B^T\| \|M^{-1}\| \|D_{\widetilde{U}}(\xi_1^{n+1})\| \|\widetilde{U}_I^{n+1} - \widetilde{U}_{II}^{n+1}\|^2 \\
&\leqslant c \|\widetilde{U}_I^{n+1} - \widetilde{U}_{II}^{n+1}\|^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&(\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1})^T B\boldsymbol{M}^{-1}(C_d(\widetilde{U}_I^{n+1})-C_d(\widetilde{U}_{II}^{n+1}))\boldsymbol{M}^{-1}B^T M^{-1}D(\widetilde{U}_{II}^{n+1})\\
\leqslant&\|B\|\|\boldsymbol{M}^{-1}\|\sum_{k=1}^{N_kN_e}\|[C_{d\widetilde{U}}(\xi_2^{n+1})]_k\|\|\boldsymbol{M}^{-1}\|\|B^T\|\|M^{-1}\|\|D(\widetilde{U}_{II}^{n+1})\\
&\|\|\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1}\|^2\\
\leqslant&c\|\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1}\|^2.
\end{aligned}
$$

Then multiplying (2.46) with $(\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1})^T$ gives

$$
\begin{aligned}
&(\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1})^T(L(\widetilde{U}_I^{n+1})-L(\widetilde{U}_{II}^{n+1}))\\
=&(\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1})^T M(\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1})+\tau^{n+1}(\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1})^T B\boldsymbol{M}^{-1}\\
&C_d(\widetilde{U}_I^{n+1})\boldsymbol{M}^{-1}B^T M^{-1}(D(\widetilde{U}_I^{n+1})-D(\widetilde{U}_{II}^{n+1}))+\tau^{n+1}(\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1})^T\\
&B\boldsymbol{M}^{-1}(C_d(\widetilde{U}_I^{n+1})-C_d(\widetilde{U}_{II}^{n+1}))\boldsymbol{M}^{-1}B^T M^{-1}D(\widetilde{U}_{II}^{n+1})\\
\geqslant&\sigma\|\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1}\|^2-2c\tau^{n+1}\|\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1}\|^2,
\end{aligned}
\tag{2.50}
$$

where $\sigma>0$ is the smallest eigenvalue of the symmetric positive mass matrix M.

Choosing $0<\tau^{n+1}\leqslant\dfrac{\sigma}{4c}$, with $\forall\widetilde{U}_I^{n+1},\widetilde{U}_{II}^{n+1}\in\mathbb{K}_b$, we obtain that

$$
(\widetilde{U}_I^{n+1}-\widetilde{U}_I^{n+1})^T(L(\widetilde{U}_I^{n+1})-L(\widetilde{U}_{II}^{n+1}))\geqslant\frac{\sigma}{2}\|\widetilde{U}_I^{n+1}-\widetilde{U}_{II}^{n+1}\|^2, \tag{2.51}
$$

which implies that for τ^{n+1} sufficiently small $L(\widetilde{U}^{n+1})$ is a uniformly function of $\mathbb{K}_b$, □

From Lemmas 2.4.5, 2.4.6 and 2.4.8, we obtain the main result of this section.

Theorem 2.4.9. *Given the DG coefficients $\widetilde{U}^n$ and the positivity preserving backward Euler KKT-LDG discretization (2.27)-(2.39) with equality constraint $h\equiv 0$. If assumption (2.49) is satisfied, then the KKT system (2.27)-(2.39) with periodic boundary conditions has only one solution.*

Corollary 2.4.10. Given the DG coefficients $\widetilde{U}^n$. If assumption (2.49) is satisfied, then for the degenerate parabolic equation (2.1) with periodic boundary conditions there exists only one solution satisfying the higher order accurate in time, positivity preserving KKT-DIRK-LDG discretizations (2.24) with equality constraint $h\equiv 0$.

Proof. Since the DIRK coefficient matrix (a_{ij}) introduced in Section 2.3.2 is a lower triangular matrix, the structure of the DIRK-LDG discretizations is similar to the structure obtained for the backward Euler LDG discretization. The analysis therefore is completely analogous to Theorem 2.4.9. □

2.4.3.3 Stability of the KKT-LDG discretization

Theorem 2.4.11. *Given the numerical solution $U_h^n \in V_h^k$ of the positivity preserving backward Euler KKT-LDG discretization (2.27)-(2.39). If assumption (2.49) is satisfied, then the discrete entropy E_h stated in (2.14) satisfies for $n = 0, 1, \cdots$,*

$$E_h(U_h^{n+1}) \leqslant E_h(U_h^n), \tag{2.52}$$

which implies that the positivity preserving backward Euler KKT-LDG discretization with periodic boundary conditions is unconditionally entropy dissipative.

Proof. From Lemma 2.4.6, we obtain that the LDG coefficients $\widetilde{U}^{n+1}$ of the positivity preserving solution U_h^{n+1} solve

$$(y - \widetilde{U}^{n+1})^T L(\widetilde{U}^{n+1}) \geqslant 0, \quad \forall y \in \mathbb{K}, \tag{2.53}$$

with L given by (2.44) and $\mathbb{K}$ given by (2.18).

From assumption (2.49), we have that there exists a positive constant $c \geqslant c_0 > 0$ such that

$$\widetilde{U}^{n+1} - cM^{-1}D(\widetilde{U}^{n+1}) \in \mathbb{K}. \tag{2.54}$$

Next, we choose $y = \widetilde{U}^{n+1} - cM^{-1}D(\widetilde{U}^{n+1})$ in (2.53), which implies

$$-c(M^{-1}D(\widetilde{U}^{n+1}))^T L(\widetilde{U}^{n+1}) \geqslant 0. \tag{2.55}$$

Using (2.44) and the fact that $c > 0$, we obtain that (2.55) implies the inequality

$$\begin{aligned} &D(\widetilde{U}^{n+1})^T(\widetilde{U}^{n+1} - \widetilde{U}^n) \\ &+\tau^{n+1}D(\widetilde{U}^{n+1})^T M^{-1}B\boldsymbol{M}^{-1}C_d(\widetilde{U}^{n+1})\boldsymbol{M}^{-1}B^T M^{-1}D(\widetilde{U}^{n+1}) \leqslant 0. \end{aligned} \tag{2.56}$$

From the definition of C_d in (2.33), (2.36) using (2.7b) and the conditions on f stated in (2.3), we obtain that $C_d(\widetilde{U}^{n+1})$ is symmetric positive definite. Hence using $\tau^{n+1} > 0$, we have

$$\tau^{n+1}D(\widetilde{U}^{n+1})^T M^{-1}B\boldsymbol{M}^{-1}C_d(\widetilde{U}^{n+1})\boldsymbol{M}^{-1}B^T M^{-1}D(\widetilde{U}^{n+1}) \geqslant 0,$$

which with (2.56) yields

$$D(\widetilde{U}^{n+1})^T(\widetilde{U}^{n+1} - \widetilde{U}^n) \leqslant 0. \tag{2.57}$$

From the definition of D in (2.35) using (2.7d) and (2.57), we obtain the bound

$$\big(\phi(\boldsymbol{x}), U_h^{n+1} - U_h^n\big) + \big(H'(U_h^{n+1}), U_h^{n+1} - U_h^n\big) \leqslant 0. \tag{2.58}$$

Using the following Taylor expansion

$$\begin{aligned} H(U_h^n) =& H(U_h^{n+1}) + H'(U_h^{n+1})(U_h^n - U_h^{n+1}) \\ &+ \frac{1}{2}H''(\xi_3^{n+1})(U_h^{n+1} - U_h^n)^2,\ \xi_3^{n+1} \in (U_h^n, U_h^{n+1}), \end{aligned}$$

we obtain that (2.58) gives

$$\begin{aligned} &\big(\phi(\boldsymbol{x}), U_h^{n+1} - U_h^n\big) + \big(H(U_h^{n+1}) - H(U_h^n), 1\big) \\ &+ \frac{1}{2}\Big(H''(\xi_3^{n+1}), \big(U_h^{n+1} - U_h^n\big)^2\Big) \leqslant 0, \end{aligned}$$

which implies, using the definition of E_h in (2.14), that

$$E_h(U_h^{n+1}) - E_h(U_h^n) = \big(\phi(\boldsymbol{x}), U_h^{n+1} - U_h^n\big) + \big(H(U_h^{n+1}) - H(U_h^n), 1\big) \leqslant 0,$$

since (2.3) gives $H''(\xi_3^{n+1}) \geqslant 0$. This proves (2.52). □

2.5 Numerical tests

In this section, we will discuss several numerical experiments to demonstrate the performance of the KKT-DIRK-LDG positivity preserving algorithm for the degenerate parabolic equation (2.1). In the computations, we will consider the porous medium equation, the nonlinear diffusion equation with a double-well potential and the nonlinear Fokker-Plank equation for fermion and boson gases. Firstly, we will present in Section 2.5.1 the order of accuracy of the DIRK-LDG discretizations with and without positivity preserving limiter to investigate if the limiter negatively affects the accuracy of the discretizations. Next, we will present in Sections 2.5.3-2.5.5 test cases for which the positivity preserving limiter is essential. Without the positivity constraint it is not possible to obtain a numerical solution or only for extremely small time steps.

In the computations, we take $\tau = \alpha \cdot h$. If the Newton method during strongly nonlinear stages requires a large number of iterations, it is generally more efficient to reduce the time step to $\frac{1}{2}\tau$ and restart the Newton iterations. When the Newton method converges well, then τ is increased each time step to 1.2τ, till the maximum predefined time step is obtained.

In order to avoid round-off effects, a positivity bound $u_{\min} = 10^{-10}$ is used in the numerical simulations, except for Section 2.5.1 where $u_{\min} = 10^{-14}$. If it is not stated otherwise, the numerical results for 1D problems are obtained on a mesh containing 100 elements and Legendre polynomials of order 2. For 2D problems a mesh consisting of 30×30 square elements and tensor product Legendre polynomial basis functions of order 2 are used.

2.5.1 Accuracy tests

For the accuracy test, we use a uniform mesh with M elements and positivity bound $u_{\min} = 10^{-14}$.

Example 2.5.1. We consider (2.1) on the domain $\Omega = (-1, 1)$ with Dirichlet boundary conditions based on the exact solution and select the following parameters

$$f(u) = u, \quad H'(u) = u^2, \quad \phi(x) = 0, \quad x \in \Omega.$$

Then (2.1) with a properly chosen source term has the nonnegative solution

$$u(x, t) = \exp(-t)(1 - x^4)^5, \quad x \in \Omega.$$

We take α in the definition of the time step as $\alpha = 1$. Tables 2.1-2.2 show that the DIRK-LDG discretizations with and without positivity preserving limiter are convergent at the rate $O(h^{k+1})$ for basis functions with polynomial order ranging from 1 to 3. The errors and orders of accuracy presented in Tables 2.1-2.2 indicate that the positivity preserving limiter is necessary and does not negatively affect accuracy.

2.5.2 Porous media equation

For the porous media equation, $f(u)H''(u)$ can locally vanish, resulting in degenerate cases [13]. We test the asymptotic behavior of the numerical solution and will show that the KKT limiter is necessary. The entropy defined in (2.4), which should be non-increasing, is also computed.

Table 2.1: Error in $L^\infty-$ and L^1- norms for Example 2.5.1 at time $T = 1$ without positivity preserving limiter.

$\mathcal{P}_k$	M	$\|u_n - u_h^n\|_{L^\infty(\Omega)}$	Order	$\|u_n - u_h^n\|_{L^1(\Omega)}$	Order	$\min u_h^n$
	40	7.33E-003	–	1.03E-003	–	-8.87e-005
1	80	1.24e-003	2.56	2.27e-004	2.18	-1.08e-005
	160	2.63e-004	2.24	5.44e-005	2.06	-4.41e-007
	320	6.05e-005	2.12	1.35e-005	2.01	-1.57e-008
	40	1.70E-003	–	8.73E-005	–	-1.60e-005
2	80	1.43e-004	3.57	8.07e-006	3.44	-1.79e-007
	160	1.36e-005	3.39	9.40e-007	3.10	-6.24e-009
	320	1.34e-006	3.34	1.16e-007	3.02	-2.07e-010
	40	1.45e-004	–	6.00e-006	–	-2.14e-006
3	80	9.87e-006	3.88	3.11e-007	4.27	-9.56e-008
	160	5.51e-007	4.16	1.76e-008	4.14	-3.51e-009
	320	3.50e-008	3.98	1.11e-009	3.99	-1.19e-010

Table 2.2: Error in $L^\infty-$ and L^1- norms for Example 2.5.1 at time $T = 1$ with positivity preserving limiter.

$\mathcal{P}_k$	M	$\|u_n - U_h^n\|_{L^\infty(\Omega)}$	Order	$\|u_n - U_h^n\|_{L^1(\Omega)}$	Order	$\min U_h^n$
	40	7.33E-003	–	1.05E-003	–	2.05e-005
1	80	1.24e-003	2.56	2.27e-004	2.21	8.15e-007
	160	2.63e-004	2.24	5.44e-005	2.06	2.77e-008
	320	6.05e-005	2.12	1.35e-005	2.01	8.55e-010
	40	1.70E-003	–	8.73E-005	–	6.15e-008
2	80	1.43e-004	3.57	8.08e-006	3.43	3.03e-007
	160	1.36e-005	3.39	9.40e-007	3.10	1.08e-008
	320	1.34e-006	3.34	1.16e-007	3.02	4.55e-010
	40	1.45e-004	–	6.02e-006	–	1.00e-014
3	80	9.87e-006	3.88	3.13e-007	4.27	4.45e-008
	160	5.51e-007	4.16	1.77e-008	4.14	1.21e-009
	320	3.50e-008	3.98	1.11e-009	4.00	2.55e-011

Example 2.5.2. In order to test degenerate cases, we choose the following parameters in (2.1) on the domain $\Omega = (0, 1)$ with zero-flux boundary conditions

$$f(u) = u, \quad H'(u) = \frac{4}{3}\left(u - \frac{1}{2}\right)^3 \max\left(u, \frac{1}{2}\right), \quad \phi(x) = 0, \quad x \in \Omega,$$

and initial data

$$u(x, 0) = \frac{1}{2} - \frac{1}{2}\cos(2\pi x), \quad x \in \Omega.$$

During the computations, the value of α for optimal convergence of the semi-smooth Newton algorithm is most of the time close to 0.1. We present the numerical solution in Figure 2.1 for basis functions with polynomial order ranging from 1 to 3 and with the KKT limiter enforced. Values of the Lagrange multiplier λ larger than 10^{-10} are shown in Figure 2.1, which indicate that the positivity constraint works well since it is only active at locations where the solution is close to the minimum value. The entropy decay using the KKT limiter and polynomial basis functions of order 3 is presented in Figure 2.2, which result is consistent with the stability analysis. In Figure 2.3, the numerical solution without KKT limiter and for polynomial basis functions with order 3 is plotted. This computation breaks down due to unphysical oscillations.

Example 2.5.3. We consider a 2D test case on the domain $\Omega = (-6, 6)^2$ with zero-flux boundary conditions by choosing in (2.1) the following parameters

$$f(u) = u, \quad H'(u) = 2u, \quad \phi(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Omega,$$

and initial data

$$u(\boldsymbol{x}, 0) = \exp\left(-\frac{1}{2}|\boldsymbol{x}|^2\right), \quad \boldsymbol{x} \in \Omega.$$

The value of α in the definition of the time step ranges in this case between 0.1 and 1. Figure 2.4 presents the numerical solution with the KKT limiter active and also the Lagrange multiplier λ. Considering the position of the non-zero Lagrange multipliers, we can see that the limiter also works well in the two dimensional case since it is only active in areas where positivity must be enforced. The entropy decay is plotted in Figure 2.5, which is consistent with the stability result of the numerical solution. Without KKT limiter, there will be unphysical oscillations, and the computation will break down at some point in the computations.

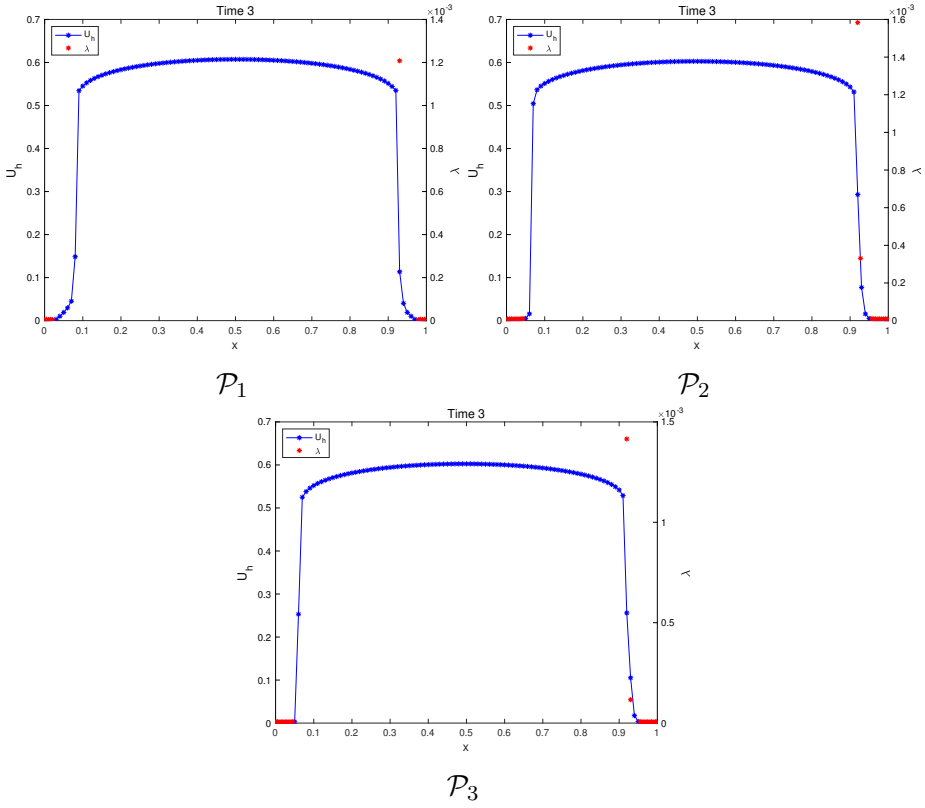


Figure 2.1: (Example 2.5.2) Numerical solution U_h for different orders of polynomial basis functions $\mathcal{P}_1$-$\mathcal{P}_3$ with the KKT limiter enforced and Lagrange multiplier λ (red dots).

2.5.3 Nonlinear diffusion with a double-well potential

Consider the nonlinear diffusion equation with double-well potential [72] on the domain $\Omega = (-1.4, 1.4)$, which is obtained by choosing in (2.1) zero-flux boundary conditions and the following parameters

$$f(u) = u, \quad H'(u) = u, \quad \phi(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2, \quad x \in \Omega. \tag{2.59}$$

This model is taken from [19]. We will test the evolution of the numerical solution with and without KKT limiter, and also the decay of the entropy (2.4). The value of α to compute the time step ranges between 0.01 to 0.1.

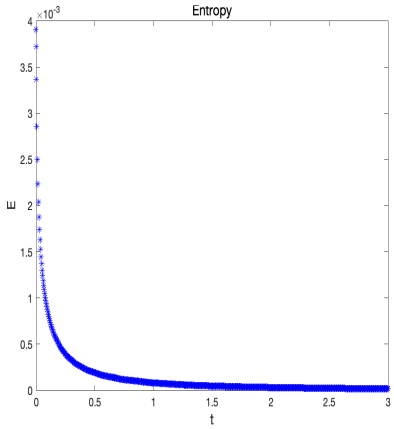


Figure 2.2: (Example 2.5.2) Entropy E_h for $\mathcal{P}_3$ basis functions with the KKT limiter enforced.

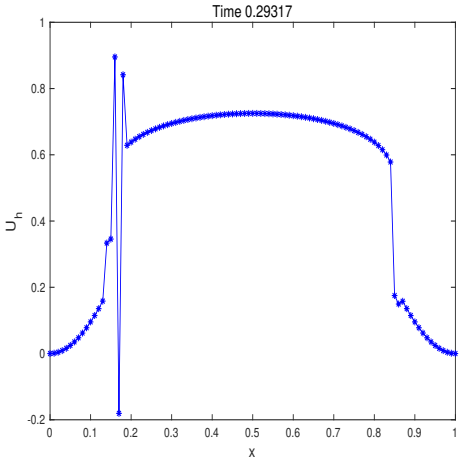


Figure 2.3: (Example 2.5.2) Numerical solution U_h for $\mathcal{P}_3$ basis functions *without* KKT limiter just before blow up.

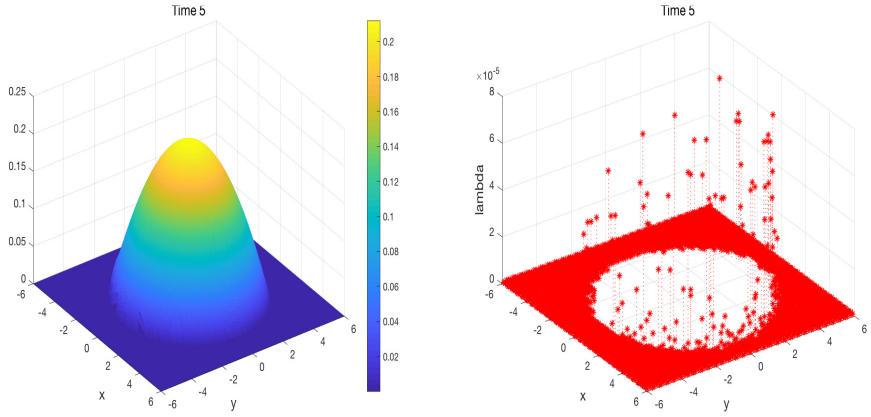


Figure 2.4: (Example 2.5.3) Numerical solution U_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced (Left) and Lagrange multiplier λ (Right).

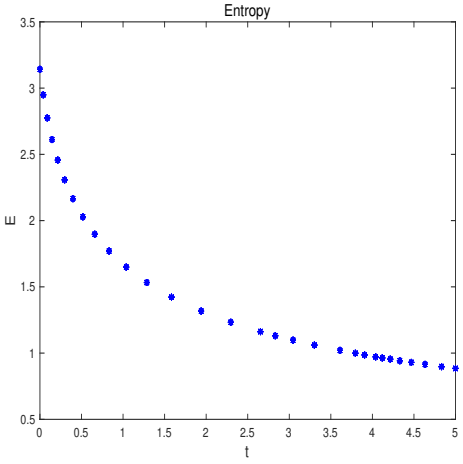


Figure 2.5: (Example 2.5.3) Entropy E_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced.

Example 2.5.4. We consider (4.1) with (2.59) and the initial data

$$u(x,0) = \frac{0.2}{\sqrt{0.4\pi}} \exp\left(-\frac{x^2}{0.4}\right), \quad x \in \Omega.$$

The numerical solution with the KKT limiter enforced and the values of the Lagrange multiplier λ larger than 10^{-10} are shown in Figure 2.6. These results indicate that the numerical solution tends to a steady state and that the KKT limiter is only active at places where the positivity constraint needs to be imposed. The entropy dissipation is presented in Figure 2.7, which uniform decay coincides with our theoretical analysis. For the numerical solution without KKT limiter, we observe that violating the positivity constraint will result in discontinuities in the solution and a breakdown of the computation, even for very small CFL number.

2.5.4 Nonlinear Fokker-Plank equation for fermion gases

We consider the nonlinear Fokker-Plank equation for fermion gases [13] on the domain $\Omega = (-10, 10)^2$, for which we select the following parameters in (2.1)

$$f(u) = u(1-u), \quad H'(u) = \log\frac{u}{1-u}, \quad \phi(\boldsymbol{x}) = \frac{1}{2}|\boldsymbol{x}|^2, \quad \boldsymbol{x} \in \Omega, \tag{2.60}$$

together with zero-flux boundary conditions.

Example 2.5.5. We consider (2.1) with (2.60) and initial data

$$\begin{aligned} u(\boldsymbol{x},0) =& \frac{1}{2\sqrt{2\pi}} \left(\exp\left(-\frac{1}{2}|\boldsymbol{x}-(2,2)|^2\right) + \exp\left(-\frac{1}{2}|\boldsymbol{x}-(2,-2)|^2\right) \right. \\ &\left. + \exp\left(-\frac{1}{2}|\boldsymbol{x}-(-2,2)|^2\right) + \exp\left(-\frac{1}{2}|\boldsymbol{x}-(-2,-2)|^2\right) \right), \ \boldsymbol{x} \in \Omega. \end{aligned}$$

During the computations, the value of α in the definition of the time step ranges between 0.1 and 1, but for most time steps $\alpha = 1$. The numerical solution at several time levels with the KKT limiter enforced and the entropy dissipation are presented in Figures 2.8 and 2.9, respectively, showing the time-asymptotic convergence of the numerical solution towards a steady state. Without the KKT limiter, the computations break down, even for very small CFL numbers.

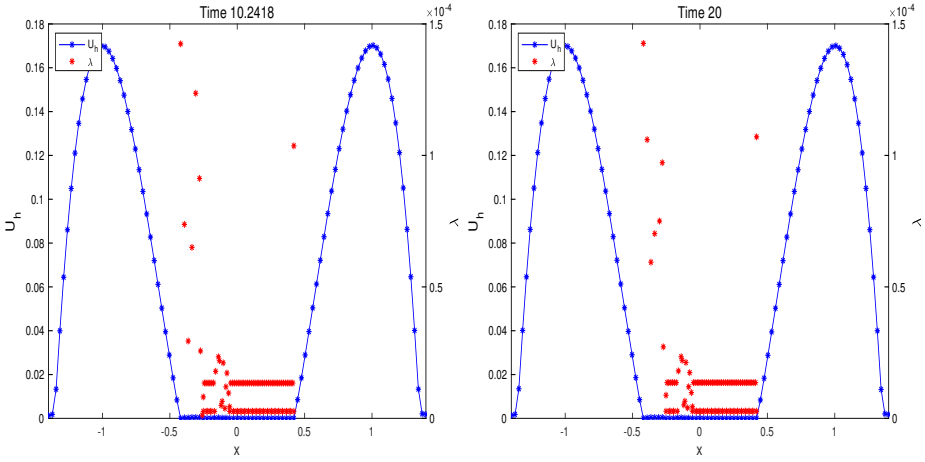


Figure 2.6: (Example 2.5.4) Numerical solution U_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced and Lagrange multiplier λ (red dots).

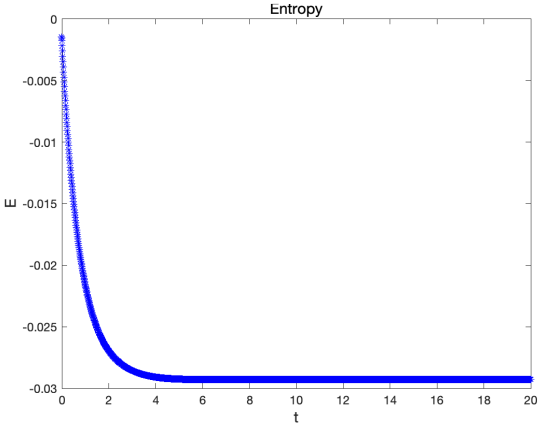


Figure 2.7: (Example 2.5.4) Entropy E_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced.

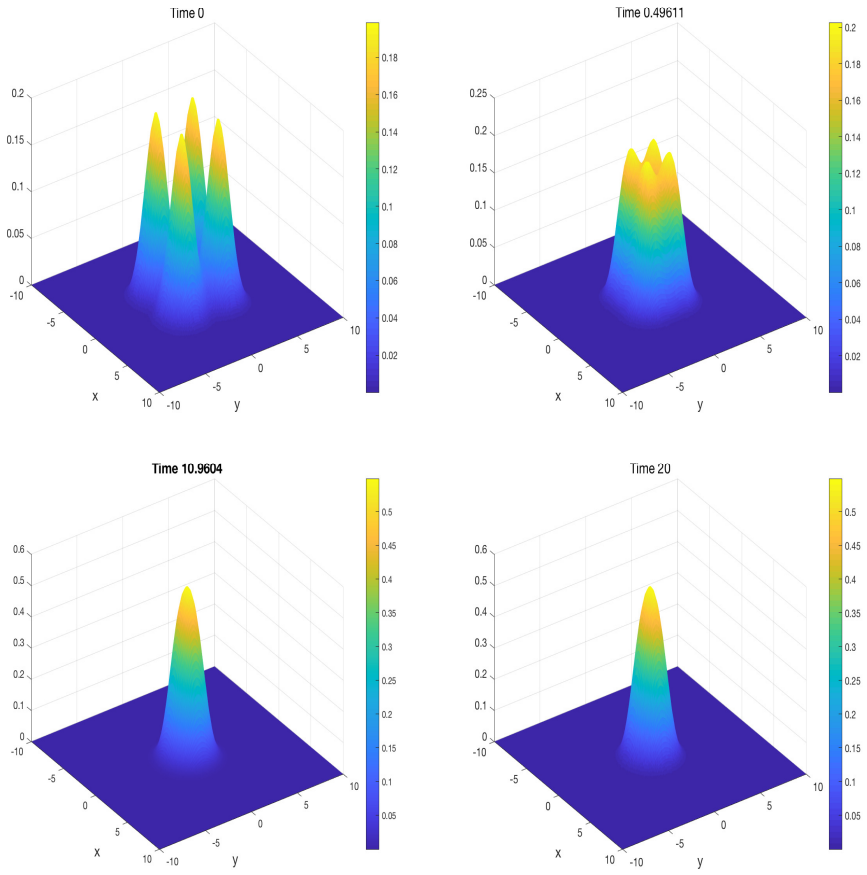


Figure 2.8: (Example 2.5.5) Numerical solution U_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced.

2.5.5 Nonlinear Fokker-Plank equation for boson gases

Example 2.5.6. We consider a nonlinear Fokker-Plank equation for boson gases with zero-flux boundary condition on a domain $\Omega = (-10, 10)$, which requires the following parameters in (2.1)

$$f(u) = u(1 + u^3), \quad H'(u) = \log \frac{u}{(1 + u^3)^{\frac{1}{3}}}, \quad \phi(x) = \frac{x^2}{2}, \quad x \in \Omega.$$

The initial data is [13, 83]

$$u(x, 0) = \frac{M}{2\sqrt{2\pi}} \left(\exp\left(-\frac{(x-2)^2}{2} \right) + \exp\left(-\frac{(x+2)^2}{2} \right) \right), \quad x \in \Omega,$$

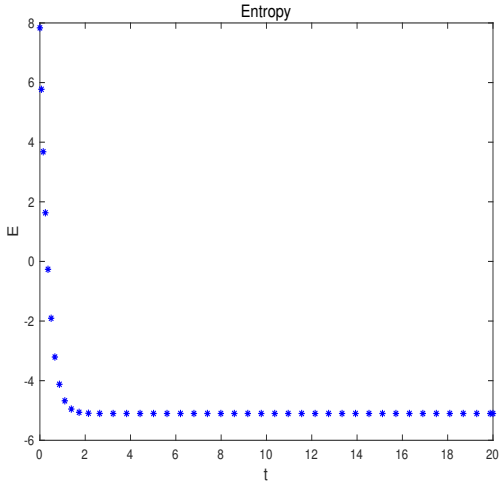


Figure 2.9: (Example 2.5.5) Entropy E_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced.

where $M \geqslant 0$ is the mass of $u(x, 0)$.

For most time steps, the value of α in the definition of the time step is 1. For the case $M = 1$, Figure 2.10 displays the numerical solution at various times. Also, the locations and values of the Lagrange multiplier λ and the entropy with the KKT limiter enforced are shown. The results in Figures 2.10 and 2.11 indicate that the numerical solution tends to a steady state, and that the Lagrange multiplier λ is needed to ensure that the positivity constraint is satisfied. Without KKT limiter, the computations break down, even for very small CFL numbers.

For this model equation, there is a critical mass phenomenon [1], which states that solutions with a large initial mass blows-up in a finite time, while solutions with small mass at initial time will not. The numerical solutions with sub-critical mass $M = 1$ at times $t = 5$ and $t = 10$ and with super-critical mass $M = 10$ at times $t = 0.2$ and $t = 1$ are shown in Figure 2.12 and Figure 2.13, respectively, and are in agreement with the results shown in [1] and the numerical observation in [13, 83].

2.6 Conclusions

The main topic of this chapter is the formulation of higher order accurate positivity preserving DIRK-LDG discretizations for the nonlinear degenerate parabolic equation (2.1). The presented numerical discretizations

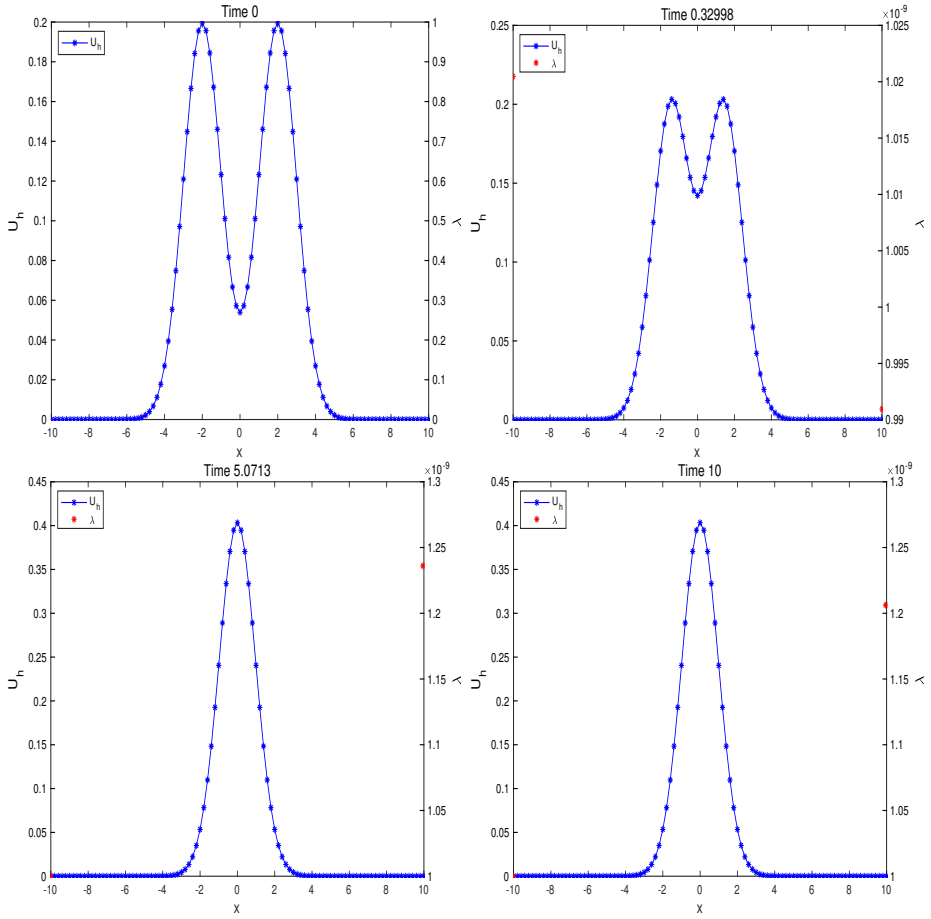


Figure 2.10: (Example 2.5.6): Numerical solution U_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced.

allow the combination of a positivity preserving limiter and time-implicit numerical discretizations for PDEs and alleviate the time step restrictions of currently available positivity preserving DG discretizations, which generally require the use of explicit time integration methods. For the spatial discretization an LDG method combined with a simple alternating numerical flux is used, which simplifies the theoretical analysis for the entropy dissipation. For the temporal discretization, the implicit DIRK methods significantly enlarge the time-step required for stability of the numerical discretization. We prove the existence, uniqueness and unconditional entropy dissipation of the positivity preserving high order accurate KKT-LDG discretization combined with an implicit Euler time discretization. Numerical

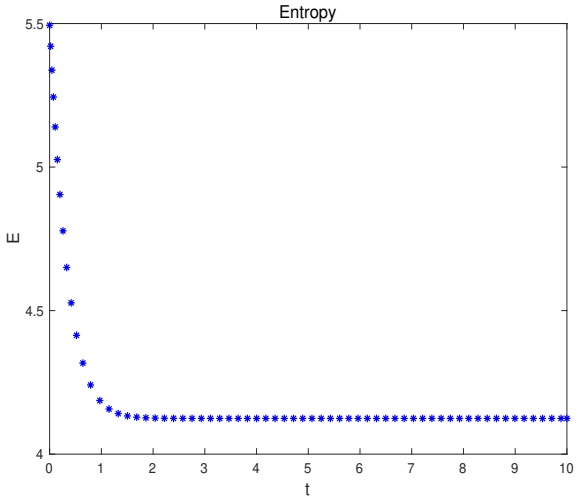


Figure 2.11: (Example 2.5.6): Entropy E_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced.

results are presented to demonstrate the accuracy of the higher order accurate positivity preserving KKT-DIRK-LDG discretizations, which is of optimal order and not affected by the positivity preserving KKT limiter. The numerical solutions satisfy the entropy decay condition.

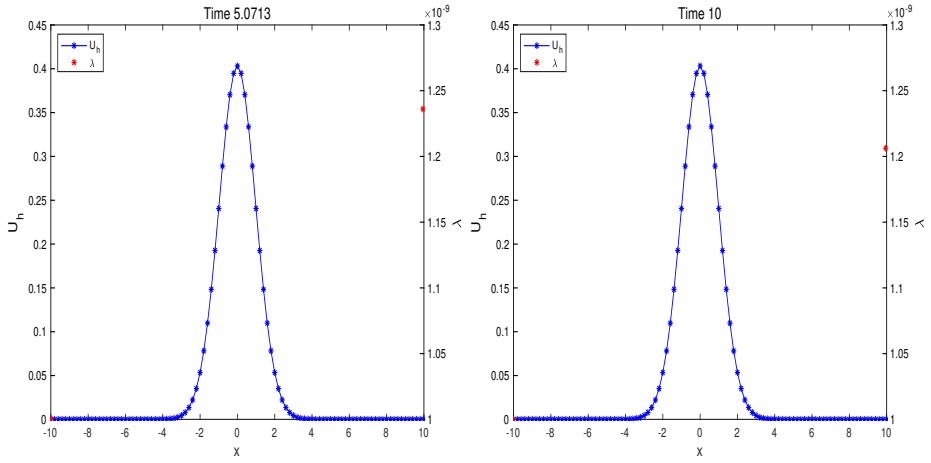


Figure 2.12: (Example 2.5.6: $M = 1$): Numerical solution U_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced.

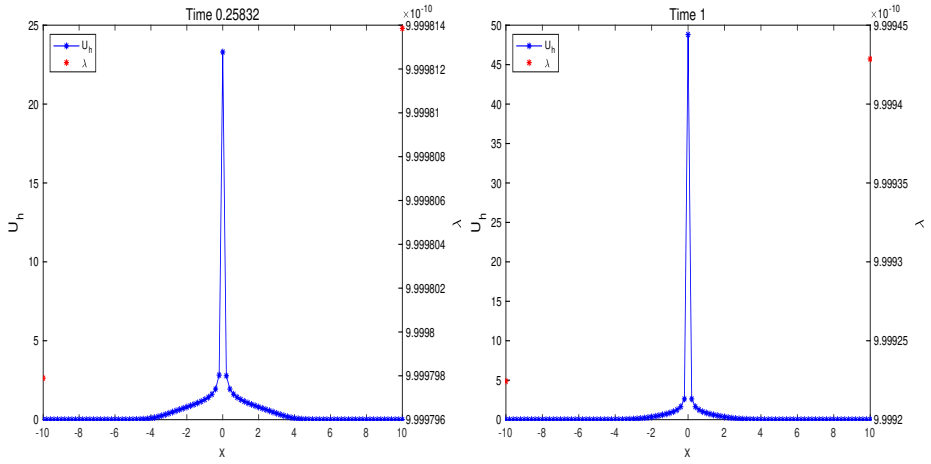


Figure 2.13: (Example 2.5.6: $M = 10$) Numerical solution U_h for $\mathcal{P}_2$ basis functions with KKT limiter enforced.

Chapter 3

Higher Order Accurate Bounds Preserving Time-Implicit Discretizations for the Chemically Reactive Euler Equations

Abstract

We construct higher order accurate bounds preserving time-implicit Discontinuous Galerkin (DG) discretizations for the reactive Euler equations modelling multispecies and multireaction chemically reactive flows. In numerical discretizations of chemically reactive flows, the time step can be significantly limited because of the large difference between the fluid dynamics time scales and the reaction time scales. In addition, the density and pressure should be nonnegative and the mass fractions between zero and one, which imposes constraints on the numerical solution that must be satisfied to obtain physically realizable solutions. We address these issues using the following steps. Firstly, we develop the Karush-Kuhn-Tucker (KKT) limiter for the chemically reactive Euler equations, which imposes bounds on the numerical solution using Lagrange multipliers, and solve the resulting KKT mixed complementarity problem using a semi-smooth Newton method. The disparity in time scales is addressed using a fractional step method, separating the convection and reaction steps, and the use of higher order accurate Diagonally Implicit Runge-Kutta (DIRK) methods. Finally, Harten's subcell resolution technique is used to deal with stiff source terms in chemically reactive flows. Numerical results are shown to demonstrate that the bounds preserving KKT-DIRK-DG discretizations are higher order accurate for smooth solutions and able to capture complicated stiff multispecies and multireaction flows with discontinuities.

3.1 Introduction

Consider the one-dimensional N-species chemically reactive Euler equations [11]

$$\begin{cases} U_t + \mathcal{F}(U)_x = S(U), & (x,t) \in \Omega \times (0, t_T], \\ U(x,0) = U_0(x), & x \in \Omega, \end{cases} \tag{3.1}$$

with Dirichlet boundary conditions, where

$$U = \begin{pmatrix} \rho \\ m \\ E \\ r_1 \\ \dots \\ r_{N-1} \end{pmatrix}, \quad \mathcal{F}(U) = \begin{pmatrix} m \\ \rho u^2 + p \\ (E+p)u \\ \rho u z_1 \\ \dots \\ \rho u z_{N-1} \end{pmatrix}, \quad S(U) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ s_1 \\ \dots \\ s_{N-1} \end{pmatrix}.$$

Here ρ is the density, u the velocity, $m = \rho u$ the momentum, E the total energy, and z_j $(j = 1, \dots, N)$ the mass fraction of the j-th species with $\sum_{j=1}^{N} z_j = 1$, and $r_j = \rho z_j$. In the following, we compute z_N always using

$$z_N = 1 - \sum_{j=1}^{N-1} z_j,$$

which automatically ensures conservation of species. The pressure is obtained from the equation of state

$$p = (E - \frac{1}{2}\rho u^2 - q_1 \rho z_1 - \ldots - q_N \rho z_N)(\gamma - 1), \tag{3.2}$$

where q_j is the enthalpy of formation for the j-th species and γ the ratio of specific heat at constant pressure c_p and constant volume c_v. Physical realizability requires that the density ρ and pressure p are nonnegative and the mass fractions z_j satisfy $z_j \in [0,1]$, $j = 1, \dots, N$. The source terms s_j $(j = 1, \dots, N)$ describe the chemical reactions. For R reactions of the form

$$\nu'_{1,r} X_1 + \nu'_{2,r} X_2 + \ldots + \nu'_{N,r} X_N \to \nu''_{1,r} X_1 + \nu''_{2,r} X_1 + \ldots + \nu''_{N,r} X_N,$$
$$r = 1, \dots, R,$$

for the species X_j and stochiometric coefficients $\nu'_{j,r}$ and $\nu''_{j,r}$, the rate of production of species j for the above chemical reaction can be written as

$$s_j = M_j \sum_{r=1}^{R} (\nu''_{j,r} - \nu'_{j,r}) \left[k_r(T) \prod_{k=1}^{N} \left(\frac{\rho z_k}{M_k} \right)^{\nu'_{k,r}} \right], \quad j = 1, 2, \ldots, N, \quad (3.3)$$

where M_j denotes the molar mass of the j-th species and $k_r(T)$, which is a function of the temperature $T = \dfrac{p}{\rho}$ (see [117, 118]), indicates the reaction rate. In this chapter, we take

$$k_r(T) = \begin{cases} B_r T^{\alpha_r}, & T > T_r, \\ 0, & T \leqslant T_r, \end{cases} \quad (3.4)$$

where for each reaction r, T_r is the ignition temperature, B_r and α_r are the pre-exponential factor and the index of temperature, respectively.

The system of equations (3.1) generally will be stiff [31] since the time scale of the reaction equations $U_t = S(U)$ will be an order of magnitude smaller than the time scale of the homogeneous Euler equations $U_t + \mathcal{F}(U)_x = 0$. This stiffness presents a serious challenge to the design of higher order accurate and efficient numerical discretizations. In particular, in high speed chemically reacting flows, the reaction speed can be much faster than the gas flow, which can easily result in spurious numerical results [31, 81], including wrong propagation speeds and incorrect locations of discontinuities. In addition, the stiff source terms in the chemically reactive Euler equations can severely limit the time step size when explicit time integration methods are used. A second challenge for numerical discretizations of the chemically reactive Euler equations is to ensure that the numerical solution is physically realizable, namely density and pressure must be nonnegative, and mass fractions should be between zero and one.

Many attempts have been made to avoid spurious phenomena when solving the chemically reactive Euler equations [10, 11, 108, 117, 118] and to ensure that the physical bounds are preserved in the numerical discretization [36, 113]. For example, using the splitting methods discussed in [51], the chemically reactive Euler equations can be divided into homogeneous equations and reaction equations. In order to locate the proper position of discontinuities in the reaction part and to avoid spurious solutions, a second order MinMax scheme [108] and a first order random projection method [10, 11] were proposed, but these methods are difficult to generalize to higher order accuracy. Harten's essentially non-oscillatory (ENO) subcell resolution technique [62], which preserves higher order accuracy,

was utilized in [117, 118] for time-explicit finite difference methods in the reaction part.

Positivity preserving DG discretizations of the chemically reactive Euler equations were obtained in [113] by extending the method presented in [134] to preserve the positivity of density, pressure, and mass fractions, except for the mass fraction z_N. The basic idea in [113] is based on the maximum-principle-preserving technique presented in [134], but since individual mass fractions do not satisfy a maximum principle it is not easy to extend this approach to preserve the upper bound for the mass fractions. Using the bounds preserving technique presented in [24, 43, 44, 86], high-order bounds preserving DG methods for multicomponent chemical reactive flows were established in [36, 38].

In general, explicit time discretization methods are chosen to solve the chemically reactive Euler equations with stiff source terms, see e.g. [108, 113, 117, 118, 127], but these methods frequently result in severe time step constraints to ensure stability of the numerical discretization. In [36], the time step was enlarged by modifying the explicit exponential Runge-Kutta (RK)/multistep time-discrete methods as in [68], but this high-order bounds preserving DG method only works on very fine meshes. For implicit time-discrete methods, the backward Euler method was chosen for the reaction equations in [10, 11], but this numerical discretization is only first order accurate. Also, it is difficult to extend the method in [10, 11] to high-order discretizations since the reaction operator in this numerical discretization is highly dependent on the time discretization method. Recently, Qin and Shu in [94] established a positivity-preserving implicit Euler DG time discretization for one-dimensional hyperbolic conservation laws. Since no high-order strong-stability-preserving RK method can be written as a convex combination of backward Euler methods [53], this approach can not be directly extended to high-order methods. Van der Vegt, Xia and Xu in [111] constructed a time-implicit bounds preserving DG discretization for parabolic Partial Differential Equations (PDEs) by coupling the bounds constraints with a higher order accurate Diagonally Implicit Runge-Kutta (DIRK) DG discretization using Lagrange multipliers. The resulting equations are the well-known Karush-Kuhn-Tucker (KKT) equations [41, 42] and the bounds preserving scheme is therefore called KKT-limiter. The KKT equations are solved using a semi-smooth Newton method, which can properly deal with the non-smoothness of the resulting algebraic equations.

In this chapter, we will present novel time-implicit higher order accurate bounds preserving DIRK-DG discretizations by extending the KKT limiter concept proposed in [111] to the chemically reactive Euler equations.

We split these equations into a homogeneous part and a reaction part using fractional step methods. The higher order accurate bounds preserving DIRK-DG discretizations for the homogeneous equations are constructed using the KKT approach in order to ensure that the density and pressure are positive and the mass fractions between zero and one. Also, a constraint projection is used to ensure that the bounds on the mass fractions during the reaction step are obeyed.

Due to the numerical dissipation in shock-capturing schemes, the location of discontinuities in the flow variables can be incorrect, which can activate the source term in an unphysical manner and result in incorrect shock speeds and spurious solutions in the shock region. We address this problem by modifying Harten's higher order ENO subcell resolution method used in [117, 118].

The KKT bounds preserving limiting approach presented in this chapter can be seen as a general framework to enforce bounds on the numerical solution which can be easily adapted to other stiff and non-stiff problems and offers great flexibility to ensure that various constraints on the exact solution are preserved. Since the KKT bounds preserving limiter is connected to the numerical discretization of the Euler equations using Lagrangian multipliers, the limiting procedure is independent of the specific numerical discretization or flow equations. The KKT limiter thus can also be used in combination with discretizations of the chemically reactive Euler or Navier-Stokes equations with a general equation of state and different reaction models. The use of DIRK methods is also interesting for possible extensions to the compressible reactive Navier-Stokes equations, which have an additional time step constraint due to the viscous terms. Also, implicit methods are more efficient when one is interested in steady state solutions. An alternative to the DIRK time discretizations would be the use of Implicit-Explicit (IMEX) time discretizations. Several IMEX schemes can preserve positivity and Total Variation Bounded (TVB) bounds during the time integration step, e.g. [52, 88, 103], but this does not guarantee that these properties are preserved for the fully discrete scheme. Designing IMEX-DG discretizations that inherently preserve bounds using the maximum principle preserving techniques in [132, 133] is non-trivial and still an open question. The KKT limiter concept discussed in this chapter can, however, also be used in combination with IMEX time discretizations, but this is beyond the scope of this chapter.

The organization of this chapter is as follows. Using operator splitting techniques, we split in Section 3.2.1 the chemically reactive Euler equations into a homogeneous part and a reaction part. The higher or-

der DIRK-DG discretizations for both parts are presented in Section 3.2.2. Further, using the KKT limiter, which is discussed in Section 3.3.1, we present in Sections 3.3.2 and 3.3.3, respectively, higher order bounds preserving DIRK-DG discretizations for the homogeneous and reaction parts of the chemically reactive Euler equations. A semi-smooth Newton method to solve the semi-smooth nonlinear algebraic equations resulting from the DIRK-DG discretizations with KKT limiter is presented in Section 3.4. The detailed algorithm, which can be seen as a template to construct time-implicit bounds preserving schemes for chemically reactive flows, is given in Section 3.5. In Section 3.6, numerical results for both the Euler equations and the chemically reactive Euler equations are presented to show that the bounds constraints are necessary and do not negatively affect the higher order accuracy for smooth solutions. For discontinuous solutions, we can see that the algorithm discussed in Section 3.5 accurately captures discontinuities and is more robust and allows a significantly larger time step than the algorithm presented in [36]. Concluding remarks are given in Section 3.7. In this chapter we will focus on the main idea how to develop higher order accurate DIRK-DG discretizations combined with the bounds preserving KKT limiter for the chemically reactive Euler equations, and only discuss the one dimensional case.

3.2 Time-implicit DG discretizations

When computing discontinuous solutions of hyperbolic conservation laws with inhomogeneous stiff source terms, spurious numerical results may be produced due to the different time scales of the homogeneous part and the chemically reactive part. In order to deal with this disparity in time scales, we adopt in Section 3.2.1 fractional step methods to split the chemically reactive Euler equations (3.1) into a homogeneous part and a reaction part. In Section 3.2.2, DIRK-DG discretizations for both parts will be presented.

3.2.1 Fractional step approach

For high speed chemically reacting flows, we use operator splitting algorithms to deal with the stiffness in the equations. With these algorithms, we split the chemically reactive Euler system (3.1) into a homogeneous Euler equation and a reaction equation

$$U_t + \mathcal{F}(U)_x = 0, \tag{3.5a}$$

$$U_t - S(U) = 0. \tag{3.5b}$$

The convection operator $\mathcal{A}$ represents the solution operator of (3.5a), and the reaction operator $\mathcal{R}$ is the solution operator of (3.5b).

The time interval $[0, t_T]$ is divided using the time steps τ^n such that $\sum_n \tau^n = t_T$ and $t^n = \sum_{j=1}^{n} \tau^j$. Given the solution $\mathcal{U}^n$ at time t^n, the solution at time t^{n+1} can be computed with the second order Strang-splitting algorithm in [102] as

$$\mathcal{U}^{n+1} = \mathcal{A}\left(\frac{\tau^{n+1}}{2}\right) \mathcal{R}^{N_r}(\tau^{n+1}) \mathcal{A}\left(\frac{\tau^{n+1}}{2}\right) \mathcal{U}^n + O(\tau^2), \tag{3.6}$$

where

$$\mathcal{R}^{N_r}(\tau) = \underbrace{\mathcal{R}\left(\frac{\tau}{N_r}\right) \cdots \mathcal{R}\left(\frac{\tau}{N_r}\right)}_{N_r}.$$

Here $N_r = 1$ is the original Strang-splitting algorithm. For test cases with fast reaction rates, N_r should be larger than one, see [118] and also [76] for a study of N_r–values, otherwise the reaction zone will be under resolved in time, which results in inaccurate solutions and possibly also in spurious numerical solutions.

For third order accurate discretizations, we will use the third order operator splitting algorithm presented in [70], but also other third order operator splitting algorithms could be used [15, 33, 51, 79].

3.2.2 DIRK-DG discretizations

In this section, we will summarize the DG discretization of the chemically reactive Euler equations combined with higher order DIRK time integration methods [61]. The numerical discretizations of (3.5a) will be combined in Section 3.3.2 with the KKT limiter, resulting in bounds preserving time-implicit DG discretizations. In order to deal with stiff source terms, we will discuss in Section 3.3.3 the use of Harten's subcell resolution technique in the numerical discretization of the reaction part.

For the DG discretization, we consider a partition of the domain $\Omega = [a, b]$ using $N_e + 1$ points

$$a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N_e+\frac{1}{2}} = b,$$

and denote element $K_j,\ 1 \leqslant j \leqslant N_e$ as $K_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$. The reference mesh size h is denoted as

$$h = \max_{1 \leqslant j \leqslant N_e} |x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}|.$$

The discontinuous finite element spaces are defined as

$$V_h^k = \{v \in L^2(\Omega) :\ v|_{K_j} \in P_k(K_j),\ j = 1, \cdots, N_e\},$$
$$\boldsymbol{V}_h^k = (V_h^k)^{N+2},$$

where $P_k(K_j)$ denote the polynomials of degree $k \geq 0$ on each element K_j.

In the numerical discretizations, we will first consider the homogeneous Euler equations (3.5a) with N species. The DG discretization is: Find $\mathcal{U}_h(t) \in \boldsymbol{V}_h^k$, such that for all $V_h \in \boldsymbol{V}_h^k$

$$\int_\Omega (\mathcal{U}_h)_t V_h d\Omega + H(\mathcal{U}_h; V_h) = 0, \tag{3.7}$$

is satisfied, where

$$H(\mathcal{U}_h; V_h) = -\int_\Omega \mathcal{F}(\mathcal{U}_h)(V_h)_x d\Omega + \sum_{j=1}^{N_e} \left(\widehat{\mathcal{F}}_{j+\frac{1}{2}} V_h(x^-_{j+\frac{1}{2}}) - \widehat{\mathcal{F}}_{j-\frac{1}{2}} V_h(x^+_{j-\frac{1}{2}}) \right),$$

here we use the local Lax-Friedrichs flux [98] defined as

$$\begin{aligned}\widehat{\mathcal{F}}_{j+\frac{1}{2}} =& \widehat{\mathcal{F}}\left(\mathcal{U}_h(x^-_{j+\frac{1}{2}}), \mathcal{U}_h(x^+_{j+\frac{1}{2}})\right) \\ =& \frac{1}{2}\left(\mathcal{F}(\mathcal{U}_h(x^-_{j+\frac{1}{2}})) + \mathcal{F}(\mathcal{U}_h(x^+_{j+\frac{1}{2}})) - \alpha_j \left(\mathcal{U}_h(x^+_{j+\frac{1}{2}}) - \mathcal{U}_h(x^-_{j+\frac{1}{2}})\right)\right),\end{aligned}$$

with $\alpha_j = \| \sqrt{\gamma p/\rho} + |u| \|_{L^\infty(x^-_{j+\frac{1}{2}}, x^+_{j+\frac{1}{2}})}$ the maximum wave speed in the homogeneous Euler equations and $x^\pm_{j+\frac{1}{2}} = \lim_{\varepsilon \downarrow 0}(x_{j+\frac{1}{2}} \pm \varepsilon)$ at $x_{j+\frac{1}{2}},\ j = 0, \cdots, N_e$.

Next, we will discretize the semi-discrete formulation (3.7) in time using DIRK methods. The approximate solution at time t^n is denoted as $\mathcal{U}_h^n = \mathcal{U}_h(\cdot, t^n)$. Assume that we know the numerical solution at time level n, then we can compute the solution at time level $n+1$ using the following steps. Given the Butcher tableau with matrix (a_{ij}) and vector (b_i), with

$a_{si} = b_i,\ i = 1, \cdots, s$. We compute the s intermediate DIRK stages i ($i = 1, \cdots, s$) by solving the nonlinear equations

$$\int_\Omega (\mathcal{U}_h^{n+1,i} - \mathcal{U}_h^n) V_h d\Omega + \tau^{n+1} \sum_{j=1}^{i} a_{ij} H(\mathcal{U}_h^{n+1,j}; V_h) = 0, \tag{3.8}$$

for the intermediate solutions $\mathcal{U}_h^{n+1,i}$. The intermediate solutions $\mathcal{U}_h^{n+1,i}$ are obtained from (3.8) using a semi-smooth Newton method after they are coupled with the bounds preserving KKT limiter, which will be discussed in Sections 3.3 and 3.4. Then the solution at time $t = t^{n+1}$ is equal to

$$\mathcal{U}_h^{n+1} = \mathcal{U}_h^{n+1,s}. \tag{3.9}$$

We use the DIRK methods presented in Section 1.3.2.2. For polynomials of order $k = 1$ and 2, we use, respectively, the Butcher tableaux (1.6) and (1.7). Note these DIRK schemes satisfy $a_{si} = b_i,\ i = 1, \cdots, s$ and are stiffly accurate. The order of accuracy of these DIRK methods is $k + 1$. Since the matrix (a_{ij}) in DIRK methods has a lower triangular structure, which means $a_{ij} = 0$ if $j > i$, DIRK methods can be easily implemented by successively solving the DIRK stages for $i = 1, \cdots, s$. For detailed information about DIRK time-discretization method, we refer to [61].

For the reaction equations (3.5b), the DIRK-DG discretization can be straightforwardly obtained by taking the L^2-inner product in space with test functions $W_h \in \boldsymbol{V}_h^k$ combined with the DIRK time integration methods,

$$\int_\Omega (\mathcal{U}_h^{n+1,i} - \mathcal{U}_h^n) W_h d\Omega - \tau^{n+1} \sum_{j=1}^{i} a_{ij} \int_\Omega S(\mathcal{U}_h^{n+1,j}) W_h d\Omega = 0,$$
$$i = 1, \cdots, s, \tag{3.10}$$

with the solution at time t^{n+1} given by (3.9).

3.3 Bounds preserving DG discretization

Physical realizability requires that the solution of (3.1) has a nonnegative density ρ and pressure p and that the mass fractions z_j ($1 \leqslant j \leqslant N$) are in the interval $[0, 1]$, with $\sum_{j=1}^{N} z_j = 1$. In general, it is very difficult to develop time-implicit DG discretizations of the chemically reactive Euler equations that intrinsically satisfy these constraints. In this chapter we will therefore

follow a different approach. Following the procedure we outlined in [111] for the construction of bounds preserving DG discretizations of parabolic PDEs using the KKT limiter concept, we impose the constraints on ρ, p and z_j $(1 \leqslant j \leqslant N-1)$ explicitly using Lagrange multipliers. The resulting KKT equations are, however, only semi-smooth and can not be solved efficiently with standard Newton methods. We use therefore the active set semi-smooth Newton method presented in [111] to solve the nonlinear algebraic equations of the KKT-DIRK-DG discretization of the chemically reactive Euler equations with explicitly imposed bounds.

In Section 3.3.1, we will summarize the KKT equations that result from the combination of the DIRK-DG discretization with the bounds constraints. Next, we will discuss in Section 3.3.2 the constraints on ρ, p and z_j $(1 \leqslant j \leqslant N-1)$ and how to impose these constraints on the stages in the DIRK-DG discretization of the homogeneous Euler equations (3.8). Since discontinuities that appear in the convection step will be smeared due to numerical dissipation, this can result in incorrect shock positions and reaction rates. We will discuss in Section 3.3.3 how to deal with this issue and also present a constrained L^2-projection to preserve the bounds on the mass fractions z_j $(1 \leqslant j \leqslant N)$ in the reaction equations (3.10).

3.3.1 Imposing bounds on the DG discretization

The numerical solution of the DG discretization in each element K is expressed as

$$U_h|_K := \sum_{j=1}^{N_k} \widehat{U}_j^K \boldsymbol{\phi}_j^K, \tag{3.11}$$

with basis functions $\boldsymbol{\phi}_j^K \in \boldsymbol{V}_h^k$ and DG coefficients $\widehat{U}_j^K$. We collect all DG coefficients in the vector $\widehat{U} \in \mathbb{R}^{dof}$, with dof the number of DG coefficients. In order to ensure that the DG discretization satisfies the bounds on ρ, p and z_j $(1 \leqslant j \leqslant N)$, we need to impose additional constraints on $\widehat{U}$. Define the set

$$\mathbb{K} := \{\widehat{U} \in \mathbb{R}^{dof} |\ g(\widehat{U}) \leqslant 0\},$$

with $g : \mathbb{R}^{dof} \to \mathbb{R}^{dof'}$ a differentiable function with the inequality constraints, see Section 3.3.2, that must be imposed on the DG coefficients $\widehat{U}$. Denote $\nabla_{\widehat{U}}$ the gradient with respect to $\widehat{U}$. Let $L : \mathbb{R}^{dof} \to \mathbb{R}^{dof}$ be the unconstrained DIRK-DG discretization (3.8)-(3.9), which is continuously

differentiable. The corresponding Karush-Kuhn-Tucker (KKT) equations [41] are then

$$\mathcal{L}(\widehat{U},\lambda) := L(\widehat{U}) + \nabla_{\widehat{U}} g(\widehat{U})^T \lambda = 0, \tag{3.12a}$$

$$0 \geqslant g(\widehat{U}) \perp \lambda \geqslant 0, \tag{3.12b}$$

where $\lambda \in \mathbb{R}^{dof'}$ are the Lagrange multipliers used to ensure that the constraint $g(\widehat{U}) \leqslant 0$ is satisfied. Here $\perp$ denotes the perpendicularity between two vectors. For the compatibility condition (3.12b), we have that

$$g_j(\widehat{U}) \leqslant 0, \quad \lambda_j \geqslant 0, \quad \text{and} \quad g_j(\widehat{U})\lambda_j = 0, \quad j = 1, \cdots, dof',$$

which is equivalent to

$$\min(-g_j(\widehat{U}), \lambda_j) = 0, \quad j = 1, \cdots, dof'.$$

Then with $F : \mathbb{R}^{dof+dof'} \to \mathbb{R}^{dof+dof'}$, the KKT-system (3.12) can be formulated as

$$0 = F(z) = \begin{pmatrix} \mathcal{L}(\widehat{U},\lambda) \\ \min(-g(\widehat{U}),\lambda) \end{pmatrix}, \tag{3.13}$$

with $z = (\widehat{U},\lambda)$. Note the KKT system (3.13) is nonlinear and $F(z)$ is only semi-smooth [69, 42] due to the compatibility condition (3.12b). This implies that standard Newton methods, which require F to be continuously differentiable, will not be efficient to solve (3.13). In Section 3.4 we will therefore discuss the global active set semi-smooth Newton method presented in [111] and adapt this method to solve the bounds preserving DIRK-DG discretization for the chemically reactive Euler equations.

Remark 3.3.1. *The bounds constraints in the KKT limiter are only active at grid points where the numerical solution does not satisfy the physical bounds. Imposing the exact physical constraints at these nodes does not influence the numerical discretization elsewhere and preserves the conservation properties of the DG discretization. This was investigated in detail in [111], where it was shown that enforcing element wise conservation in the KKT limiter, next to the physical bounds, has no effect on the numerical solution. Hence, the conservation properties of the DG discretization, and in particular the shock speeds and shock positions, are not affected by the KKT limiter.*

3.3.2 Constraints on the homogeneous Euler equations

Since the density ρ and pressure p must be nonnegative, and mass fractions z_j $(1 \leqslant j \leqslant N)$ must be in the interval $[0, 1]$ and also satisfy the equality constraint $\sum_{j=1}^{N} z_j = 1$, we need to impose these constraints explicitly in the KKT algorithm discussed in Section 3.3.1.

Imposing these constraints at all points in an element would be prohibitively expensive and also very difficult, especially for higher order accurate discretizations. This is, however, not necessary since in the DG discretization only data at the element and face quadrature points are used. The physical constraints therefore only need to be imposed at the N_q quadrature points used in the DG discretization. These constraints then are expressed in terms of the DG coefficients $\widehat{U}$ in (3.11) since these are the unknown variables in the DG discretization. We use Gauss-Lobatto quadrature points, which are both inside the element and at element faces. This ensures that the bounds are also preserved in the flux calculations.

After setting the test functions $V_h = \boldsymbol{\phi}_h^k \in \boldsymbol{V}_h^k$, the unconstrained DIRK-DG discretization (3.8) for stage i $(i = 1, \cdots, s)$ can be expressed as

$$L(\widehat{U}^{n+1,i}) = M(\widehat{U}^{n+1,i} - \widehat{U}^n) + \tau^{n+1} \sum_{j=1}^{i} a_{ij} \widetilde{H}(\widehat{U}^{n+1,j}), \tag{3.14}$$

with

$$\widetilde{H}(\widehat{U}) := H(U_h; \boldsymbol{\phi}_h^k),$$

and mass matrix M.

Using the expression for U_h in terms of the basis functions (3.11), we can explicitly express each component of U_h in terms of the basis functions, e.g. for the i-th DIRK stage with $\phi_j^K \in V_h^k$, we have for the density in element K the expression

$$\rho_h^{K,(n+1,i)}(x) = \sum_{j=1}^{N_k} \widehat{\rho}_j^{K,(n+1,i)} \phi_j^K(x),$$

with similar expressions for $m_h, E_h, r_{h,1}, \cdots, r_{h,N-1}$. The inequality constraints on ρ, p, z_j $(1 \leqslant j \leqslant N)$ are imposed in each element K at the quadrature points x_l, $1 \leqslant l \leqslant N_q$, e.g.

$$g_\rho^K = (g_{\rho_1}^K, \ldots, g_{\rho_{N_q}}^K).$$

The bounds on the DG coefficients $\widehat{U}^{n+1,i} = (\widehat{\rho}^{n+1,i}, \widehat{m}^{n+1,i}, \widehat{E}^{n+1,i}, \widehat{r}_1^{n+1,i}, \ldots, \widehat{r}_{N-1}^{n+1,i})$ in the DIRK-DG discretization (3.8) can now be stated as

i. Positivity Constraints

$$g_{\rho_l}^K(\widehat{\rho}^{n+1,i}) = \rho_{\min} - \rho_h^{K,(n+1,i)}(x_l), \quad l = 1, \ldots, N_q, \tag{3.15}$$

$$g_{p_l}^K(\widehat{U}^{n+1,i}) = p_{\min} - p_h^{K,(n+1,i)}(x_l), \quad l = 1, \ldots, N_q, \tag{3.16}$$

$$g_{(z_j)_l}^K(\widehat{\rho}^{n+1,i}, \widehat{r}_j^{n+1,i}) = -\frac{r_{h,j}^{K,(n+1,i)}(x_l)}{\rho_h^{K,(n+1,i)}(x_l)}, \quad l = 1, \ldots, N_q, \quad j = 1, \ldots, N-1, \tag{3.17}$$

where we use the equation of state (3.2) and $z_N = 1 - \sum_{j=1}^{N-1} z_j$ in (3.16) to express the dependence of the pressure on the conservative variables.

ii. Maximum Constraints

$$g_{(z_N)_l}^K(\widehat{\rho}^{n+1,i}, \widehat{r}_1^{n+1,i}, \ldots, \widehat{r}_{N-1}^{n+1,i}) = \sum_{j=1}^{N-1} \frac{r_{h,j}^{K,(n+1,i)}(x_l)}{\rho_h^{K,(n+1,i)}(x_l)} - 1, \quad l = 1, \ldots, N_q. \tag{3.18}$$

Note, for the constraints on the mass fractions we impose $z_j \geqslant 0$ ($j = 1, \ldots, N-1$) and $\sum_{j=1}^{N-1} z_j \leqslant 1$ at all quadrature points. The condition $z_N = 1 - \sum_{j=1}^{N-1} z_j$ will ensure then that all mass fractions are in the interval $[0, 1]$ and total mass is conserved.

The constants $\rho_{\min}$, $p_{\min}$ are the bounds imposed on the density or pressure. In order to prevent that the density or pressure become negative due to small numerical truncation errors, if not stated otherwise, we set $\rho_{\min} = p_{\min} = 10^{-10}$.

In order to simplify notation, we combine all inequality constraints into a single vector

$$g = (g_\rho, g_p, g_{z_1}, \ldots, g_{z_N})^T \tag{3.19}$$

with $g_\rho = (g_\rho^{K_1}, \ldots, g_\rho^{K_{N_e}})$ and similar expressions for the other terms $g_p, g_{z_1}, \ldots, g_{z_N}$. For the semi-smooth Newton method discussed in Section 3.4, it is crucial to note that g depends on the DG coefficient $\widehat{U}$.

For each DIRK i-stage, the KKT equations for the bounds preserving DIRK-DG discretizations can now be expressed as

$$\mathcal{L}(\widehat{U}^{n+1,i}, \lambda) := L(\widehat{U}^{n+1,i}) + \nabla_{\widehat{U}} g(\widehat{U}^{n+1,i})^T \lambda = 0, \tag{3.20a}$$

$$0 \geqslant g(\widehat{U}^{n+1,i}) \perp \lambda \geqslant 0, \tag{3.20b}$$

with

$$\nabla_{\widehat{U}} g\left(\widehat{U}^{n+1,i}\right) = \frac{\partial g}{\partial \widehat{U}^{n+1,i}} \in \mathbb{R}^{(N+2)N_q N_e \times (N+2)N_k N_e}.$$

Here L: $\mathbb{R}^{(N+2)N_k N_e} \to \mathbb{R}^{(N+2)N_k N_e}$ is the DIRK-DG discretization (3.14), g: $\mathbb{R}^{(N+2)N_k N_e} \to \mathbb{R}^{(N+2)N_q N_e}$ the inequality constraints (3.19) and $\lambda \in \mathbb{R}^{(N+2)N_q N_e}$ the Lagrange multipliers.

The KKT equations for DIRK stage i can be concisely expressed as the following system of non-smooth algebraic equations

$$0 = F(\widehat{U}^{n+1,i}, \lambda) = \begin{pmatrix} \mathcal{L}(\widehat{U}^{n+1,i}, \lambda) \\ \min(-g(\widehat{U}^{n+1,i}), \lambda) \end{pmatrix}. \tag{3.21}$$

After solving (3.21) with the semi-smooth Newton algorithm, discussed is Section 3.4, we obtain the DG coefficients $\widehat{U}^{n+1,i}$, which gives using (3.11) the DG solution $U_h^{n+1,i*}$. Next, monotonicity of the numerical solution is enforced by applying the TVB limiter [27, 29] to $U_h^{n+1,i*}$, which results in the updated DG coefficients $\widehat{U}^{n+1,i}$, and after (3.21) is solved for all Runge-Kutta stages gives the numerical solution U_h^{n+1}.

The physical constraints are thus imposed implicitly, coupled with the DIRK-DG discretization, whereas the monotonicity constraint is enforced after the solution of the KKT-DIRK-DG equations for each Runge-Kutta stage is obtained.

Solving the DIRK-DG equations without imposing the physical constraints coupled with the DIRK-DG discretization can easily result in unphysical solutions and thus in the breakdown of the Newton algorithm. Including the TVB limiter as a constraint to the DIRK-DG equations would, however, make the algorithm unnecessarily complex and is not necessary for stability.

3.3.3 Constraints on the reaction equations

When solving the homogeneous Euler equations (3.5a), all standard shock-capturing schemes will produce smeared discontinuities in the shock regions. For stiff reaction equations, the smeared discontinuities will then result in inaccurate shock locations. This activates the source terms in an unphysical way, which can result in incorrect shock speeds and spurious solutions. Since it is impractical to resolve the very small chemical reaction scales, a stable and accurate algorithm must be obtained based on the data from the shock capturing scheme. We will follow here the process outlined in [117, 118, 127] using the Harten's subcell resolution technique [62]. First, we will discuss an algorithm that can identify elements that contain discontinuities. Next, after computing the correct shock locations in the reaction zone, we will reconstruct the temperature, mass fractions and density on both sides of the discontinuity using data in the neighboring elements, which are less polluted by the numerical dissipation in the discontinuity.

In order to better explain the reaction operator, we rewrite the source term $S(U)$ in (3.5b) as $S(T, \rho, z_1, \ldots, z_N)$ and denote its non-zero components s_k $(k = 1, 2, \ldots, N)$.

We will use the following algorithm:

1. **Identification of correct position of discontinuities**

a.) We use one mass fraction to identify elements that contain true discontinuities inside the region with the smeared discontinuities. For this identification, we choose a mass fraction z_k that has a zero value in the left-hand side state. If there is more than one such mass fraction or none, we choose the mass fraction with the biggest jump in the smeared discontinuities.

b.) Next, we use the minmod-based discontinuity indicator [117] on the selected mass fraction z_k to identify elements with discontinuities. Element $K_l = [x_{l-\frac{1}{2}}, x_{l+\frac{1}{2}}]$, $l = 1, \ldots, N_e$, is identified as being in the shock domain if $|C_l| \geqslant |C_{l-1}|$ and $|C_l| \geqslant |C_{l+1}|$ (with at least one strict inequality), where

$$C_l = \text{minmod}\{(\bar{z}_k)_{l+1} - (\bar{z}_k)_l, (\bar{z}_k)_l - (\bar{z}_k)_{l-1}\},$$

with $(\bar{z}_k)_l = \dfrac{1}{|K_l|}\displaystyle\int_{K_l} z_k dK_l$ the element average of the mass fraction z_k in element K_l that is selected under 1.a. We also check whether neighboring elements K_{l-1} and K_{l+1} also contain a discontinuity, but for simplicity of

exposition, we assume in the remainder that this is not the case.

2. **Improve accuracy of temperature, mass fractions and density in discontinuities**

If element K_l is identified to contain a discontinuity, we obtain modified values of the temperature T, mass fractions z_k $(k = 1, 2, \ldots, N)$ and density ρ in element K_l using

$$\begin{cases} \widetilde{T}(x) = p_{l-1}(x;T), & \widetilde{z}_k(x) = p_{l-1}(x;z_k), & \widetilde{\rho}(x) = p_{l-1}(x;\rho), & \text{if } x \leqslant x_t \\ \widetilde{T}(x) = p_{l+1}(x;T), & \widetilde{z}_k(x) = p_{l+1}(x;z_k), & \widetilde{\rho}(x) = p_{l+1}(x;\rho), & \text{if } x > x_t, \end{cases} \tag{3.22}$$

with $p_{l\pm1}(x;y)$ the DG solution for variable y in elements $K_{l\pm1}$, with y either the temperature T, mass fractions z_k $(k = 1, \ldots, N)$ or density ρ. The position x_t of the shock location is the solution of the following conservation relation

$$\int_{x_{l-\frac{1}{2}}}^{x_t} p_{l-1}(x;E)dx + \int_{x_t}^{x_{l+\frac{1}{2}}} p_{l+1}(x;E)dx - \bar{E}_l \Delta x = 0, \tag{3.23}$$

where $p_l(x;E)$ is the DG solution of the energy E in element K_l in the convection step and $\bar{E}_l$ the average energy in element K_l. The energy E is chosen because it is a conserved variable.

It is not necessary to compute the exact shock location x_t in element K_l by solving (3.23) accurately. One only needs to know if a quadrature point in element K_l is on the left or right side of the shock location x_t to determine if the temperature, mass fractions and density need to be computed from either the left or right element connected to element K_l. In order to decide this, we use the criterion $x > x_t$ if $\chi(x_{l-\frac{1}{2}})\chi(x) < 0$, and $x \leqslant x_t$ otherwise, where

$$\chi(x) = \int_{x_{l-\frac{1}{2}}}^{x} p_{l-1}(x;E)dx + \int_{x}^{x_{l+\frac{1}{2}}} p_{l+1}(x;E)dx - \bar{E}_l \Delta x.$$

Note, when Δx is small enough then (3.23) will have a unique solution.

3. **Modify DIRK-DG discretization for reaction step**

Next, we use $\widetilde{T}$, $\widetilde{z}_k$ $(k = 1, \ldots, N)$, $\widetilde{\rho}$, obtained from (3.22), instead of T, z_k $(k = 1, \ldots, N)$, ρ in the explicit parts of the DIRK-DG discretization of the reaction step (3.10). At every DIRK stage i $(i = 1, \ldots, s)$, we modify

(3.10) therefore as

$$\int_\Omega U_h^{n+1,i} V_h d\Omega = \int_\Omega U_h^n V_h d\Omega + \tau^{n+1} \sum_{j=1}^{i-1} a_{ij} \int_\Omega S(\widetilde{T}_h^{n+1,j}, \widetilde{\rho}_h^{n+1,j}, (\widetilde{z}_1)_h^{n+1,j}, \ldots, (\widetilde{z}_N)_h^{n+1,j}) V_h d\Omega + \tau^{n+1} a_{ii} \int_\Omega S(T_h^{n+1,i}, \rho_h^{n+1,i}, (z_1)_h^{n+1,i}, \ldots, (z_N)_h^{n+1,i}) V_h d\Omega. \quad (3.24)$$

Then after modifying $U_h^{n+1,s}$ obtained from (3.24) using Steps 1 and 2, the solution of the reaction equations at time t^{n+1} is

$$\int_\Omega U_h^{n+1} V_h d\Omega = \int_\Omega U_h^n V_h d\Omega + \tau^{n+1} \sum_{i=1}^{s} a_{si} \int_\Omega S(\widetilde{T}_h^{n+1,i}, \widetilde{\rho}_h^{n+1,i}, (\widetilde{z}_1)_h^{n+1,i}, \ldots, (\widetilde{z}_N)_h^{n+1,i}) V_h d\Omega. \quad (3.25)$$

4. **Impose constraints on the mass fractions**

Finally, for the reaction equations, we apply a constrained L^2-projection of U_h^{n+1} computed in (3.25) to obtain the DG coefficients $\widehat{U}^{n+1}$ that also ensure U_h^{n+1} satisfies the bounds on the mass fractions stated in Section 3.3.2. The constrained L^2-projection is obtained by replacing $L(\widehat{U})$ in (3.12) with the L^2-projection and using the same constraints on the mass fractions as discussed in Section 3.3.2 for the KKT-DIRK-DG discretizations.

3.4 Semi-smooth Newton method

The nonlinear algebraic equations $F(\widehat{U}^{n+1,i}, \lambda)$ (3.21), resulting from the KKT equations (3.20), are only semi-smooth. A function $F(x)$ is semi-smooth at $x \in \mathbb{R}^d$ if F is locally Lipschitz continuous and directional differentiable at x and

$$\lim_{x+h \in D_F, |h| \to 0} \frac{F'(x+h;h) - F'(x;h)}{|h|} = 0, \quad (3.26)$$

with D_F the set of points at which F is differentiable, $h \in \mathbb{R}^d$ and F' the directional derivative of F, see [69, Theorem 8.2]. For the rather technical definition of semi-smoothness, we refer to [42]. The semi-smoothness of $F(\widehat{U}^{n+1,i}, \lambda)$ prevents the usage of a standard Newton method, which requires F to be continuously differentiable in order to converge. We use

therefore the active set semi-smooth Newton method presented in [111] to solve (3.21). This semi-smooth Newton method uses the concept of a quasi-directional derivative [69].

We assume that $D \subset \mathbb{R}^{dof+dof'}$, with $dof = (N+2)N_kN_e$, $dof' = (N+2)N_qN_e$, is an open set and $F : D \to \mathbb{R}^{dof+dof'}$ is directionally differentiable and locally Lipschitz continuous. Assume that there exists a $z^0 \in D$ such that

$$S := \{z = (x, \lambda) \in D |\ |F(z)| \leqslant |F(z^0)|\}$$

is bounded. The quasi-directional derivative [69] satisfies the following conditions.

Definition 3.4.1. $G : S \times \mathbb{R}^{dof+dof'} \to \mathbb{R}^{dof+dof'}$ is a quasi-directional derivative of $F : D \to \mathbb{R}^{dof+dof'}$ on S, when for all $z, z^* \in S$, the following three conditions hold

$$\begin{aligned}
&(F(z), F'(z;d)) \leqslant (F(z), G(z;d)),\\
&G(z;td) = tG(z;d), \text{ for all } d \in \mathbb{R}^{dof+dof'}, z \in S \text{ and } t \geqslant 0,\\
&(F(z^*), F^0(z^*;d^*)) \leqslant \limsup_{z\to z^*, d\to d^*}(F(z), G(z;d)), \text{ for all } z \to z^*,\\
&\qquad\qquad\qquad\qquad d \to d^*,
\end{aligned}$$

where $F'(z;d)$ is the directional derivative of F at z in the direction d, $F^0(z^*;d^*)$ is the Clarke generalized directional derivative of F at z^* in the direction d^*, which is defined as

$$F^0(z^*;d^*) = \lim_{y\to z^*} \sup_{t\downarrow 0^+} \frac{F(y+td^*) - F(y)}{t}.$$

The search direction d in the semi-smooth Newton method is then the solution of the mixed linear complementarity problem

$$F(z) + G(z;d) = 0, \quad z \in S,\ d \in \mathbb{R}^{dof+dof'}. \tag{3.27}$$

Given the global merit function $\theta(z) = \frac{1}{2}|F(z)|^2$, it is proven in [111] that the use of a quasi-directional derivative ensures a bound on the Clarke generalized directional derivative of $\theta(z)$,

$$\theta^0(z^*;d^*) \leqslant -2\theta(z^*).$$

Hence the search direction d obtained from (3.27) always provides a descent direction for the global merit function $\theta(z^*)$. If $\theta(z) = 0$, then this also implies $F(z) = 0$.

The crucial element in the semi-smooth Newton method (3.27) is the quasi-directional derivative G. Based on the definition of L stated in (3.20a), we take $\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_{(N+2)N_kN_e})^T$. Set $z = (\widehat{U}^{n+1,i}, \lambda) \in \mathbb{R}^{(N+2)N_kN_e+(N+2)N_qN_e}$ and the search direction

$$d = (d_U, d_\lambda) \in \mathbb{R}^{(N+2)N_kN_e+(N+2)N_qN_e},$$

where

$$d_U = (d_\rho, d_m, d_E, d_{r_1}, \ldots, d_{r_{N-1}}) \in \mathbb{R}^{(N+2)N_kN_e}$$

is the search direction with respect to $\widehat{U}^{n+1,i} = (\widehat{\rho}^{n+1,i}, \widehat{m}^{n+1,i}, \widehat{E}^{n+1,i}, \widehat{r}_1^{n+1,i}, \ldots, \widehat{r}_{N-1}^{n+1,i})$ and $d_\lambda \in \mathbb{R}^{(N+2)N_qN_e}$ is the search direction with respect to λ.

Based on the analysis in [111] and using $D_{\widehat{U}^{n+1,i}}\mathcal{L}(z)\cdot d_U = D_{\widehat{\rho}^{n+1,i}}\mathcal{L}(z)\cdot d_\rho + D_{\widehat{m}^{n+1,i}}\mathcal{L}(z)\cdot d_m + D_{\widehat{E}^{n+1,i}}\mathcal{L}(z)\cdot d_E + D_{\widehat{z}_1^{n+1,i}}\mathcal{L}(z)\cdot d_{z_1} + \ldots + D_{\widehat{z}_{N-1}^{n+1,i}}\mathcal{L}(z)\cdot d_{z_{N-1}}$, we obtain the following expression for the quasi-directional derivative

$$\begin{aligned}
G_l(z;d) =& D_{\widehat{U}^{n+1,i}}\mathcal{L}_l(z)\cdot d_U + D_\lambda\mathcal{L}_l(z)\cdot d_\lambda, \\
& l \in \mathcal{N}_{(N+2)N_kN_e}, && \text{(3.28a)}\\
G_{l+(N+2)N_kN_e}(z;d) =& - D_{\widehat{U}^{n+1,i}} g_l(\widehat{U}^{n+1,i})\cdot d_U,\ l \in \alpha_\delta(z), && \text{(3.28b)}\\
G_{l+(N+2)N_kN_e}(z;d) =& \max(-D_{\widehat{U}^{n+1,i}} g_l(\widehat{U}^{n+1,i})\cdot d_U, (d_\lambda)_l), \\
& l \in \beta_{1\delta}(z), && \text{(3.28c)}\\
G_{l+(N+2)N_kN_e}(z;d) =& \min(-D_{\widehat{U}^{n+1,i}} g_l(\widehat{U}^{n+1,i})\cdot d_U, (d_\lambda)_l), \\
& l \in \beta_{2\delta}(z), && \text{(3.28d)}\\
G_{l+(N+2)N_kN_e}(z;d) =& (d_\lambda)_l,\ l \in \gamma_\delta(z), && \text{(3.28e)}
\end{aligned}$$

where the following sets are used to define $G(z;d)$

$$\begin{aligned}
\mathcal{N}_n =& \left\{ j \in \mathbb{N} \mid 1 \leqslant j \leqslant n \right\}, \\
\alpha_\delta(z) =& \left\{ j \in \mathcal{N}_{(N+2)N_qN_e} \mid \lambda_j > -g_j(x) + \delta \right\}, \\
\beta_{1\delta}(z) =& \left\{ j \in \mathcal{N}_{(N+2)N_qN_e} \mid -g_j(x) - \delta \leqslant \lambda_j \leqslant -g_j(x) + \delta, \right. \\
& \left. F_{j+(N+2)N_kN_e}(z) > 0 \right\}, \\
\beta_{2\delta}(z) =& \left\{ j \in \mathcal{N}_{(N+2)N_qN_e} \mid -g_j(x) - \delta \leqslant \lambda_j \leqslant -g_j(x) + \delta, \right. \\
& \left. F_{j+(N+2)N_kN_e}(z) \leqslant 0 \right\}, \\
\gamma_\delta(z) =& \left\{ j \in \mathcal{N}_{(N+2)N_qN_e} \mid \lambda_j < -g_j(x) - \delta \right\}.
\end{aligned}$$

Using the procedure outlined in [111] Appendix, it can be shown that G satisfies, for any $\delta > 0$, the conditions stated in Definition 3.4.1. Therefore, G provides a suitable search direction for the global semi-smooth Newton method. In order to use the quasi-directional derivative G given by (3.28) in the semi-smooth Newton method, we introduce the following sets

$$\begin{aligned}
\Lambda^{11}_{\beta_\delta}(z,d) &:= \{j \in \beta_{1\delta}(z) \mid -D_{\widehat{U}^{n+1,i}} g_j(\widehat{U}^{n+1,i}) \cdot d_U > (d_\lambda)_j\},\\
\Lambda^{12}_{\beta_\delta}(z,d) &:= \{j \in \beta_{1\delta}(z) \mid -D_{\widehat{U}^{n+1,i}} g_j(\widehat{U}^{n+1,i}) \cdot d_U \leqslant (d_\lambda)_j\},\\
\Lambda^{21}_{\beta_\delta}(z,d) &:= \{j \in \beta_{2\delta}(z) \mid -D_{\widehat{U}^{n+1,i}} g_j(\widehat{U}^{n+1,i}) \cdot d_U > (d_\lambda)_j\},\\
\Lambda^{22}_{\beta_\delta}(z,d) &:= \{j \in \beta_{2\delta}(z) \mid -D_{\widehat{U}^{n+1,i}} g_j(\widehat{U}^{n+1,i}) \cdot d_U \leqslant (d_\lambda)_j\},
\end{aligned}$$

and combine these sets into

$$\Lambda^1_\delta(z,d) := \alpha_\delta(z) \cup \Lambda^{11}_{\beta_\delta}(z,d) \cup \Lambda^{22}_{\beta_\delta}(z,d), \tag{3.29a}$$

$$\Lambda^2_\delta(z,d) := \gamma_\delta(z) \cup \Lambda^{12}_{\beta_\delta}(z,d) \cup \Lambda^{21}_{\beta_\delta}(z,d). \tag{3.29b}$$

The quasi-directional derivative G in (3.28) can then be written as the Jacobi matrix

$$G(z;d) = \widehat{G}(z)d,$$

with

$$\widehat{G}(z) = \begin{pmatrix} D_{\widehat{U}^{n+1,i}} \mathcal{L}_l(z)|_{l \in \mathcal{N}_{(N+2)N_k N_e}} & D_\lambda \mathcal{L}_l(z)|_{l \in \mathcal{N}_{(N+2)N_k N_e}} \\ -D_{\widehat{U}^{n+1,i}} g_l(\widehat{U}^{n+1,i})|_{l \in \Lambda^1_\delta(z,d)} & \delta_{lj}|_{l,j \in \Lambda^2_\delta(z,d)} \end{pmatrix}, \tag{3.30}$$

where δ_{lj} is the Kronecker delta function. Hence, the equations for the search direction d are equal to

$$F(z) + \widehat{G}(z)d = 0. \tag{3.31}$$

The details of the semi-smooth Newton algorithm are given in [111].

During the Newton iterations both the Jacobian matrix (3.30) and the sets (3.29) are continuously updated. In general, after a few iterations, the proper sets will be obtained and the Jacobian matrix elements will only change smoothly. The semi-smooth Newton method then closely resembles a standard Newton method. It is possible that during the Newton iterations the matrix $\widehat{G}(z)$ is poorly conditioned. The linear system (3.31) is therefore solved using a minimum norm least squares method, which basically makes the algorithm a Gauss-Newton method [69]. To further improve the condition number, we apply simultaneously iterative row and column scaling in the L^∞-norm, which is described in detail in [7, 111].

3.5 Algorithm for stiff multispecies detonation problems

We split the reactive Euler equations as discussed in Section 3.2.1 into a homogeneous part and a reaction part. The DIRK-DG discretizations for these two parts are given in Section 3.2.2. In order to obtain physically relizable numerical solutions, bounds constraints using the KKT limiter are imposed on the DIRK-DG discretization of the convection part in Section 3.3.2 and on the reaction part in Section 3.3.3.

The chemically reactive Euler equations are now solved using the following steps. Assume we know the numerical solution at time t^n, then the numerical solution at time t^{n+1} is obtained by successively computing the convection and reaction operators with the fractional step approach. Here we take the fractional step method (3.6) as an example to describe the algorithm.

Algorithm 2 Bounds preserving KKT-DIRK-DG method based on splitting method (3.6)

Given the DG coefficients $\widehat{U}^n$ at time t^n.

Homogeneous part:

For all DIRK stages $i = 1, 2, \ldots, s$, do

(i) Solve (3.21) for a time step $\tau^{n+1}/2$ with the semi-smooth Newton method discussed in Section 3.4. This will provide the DG coefficients $\widehat{U}^{n+\frac{1}{2},i*}$.

(ii) In order to ensure monotonicity of the numerical solution, apply a TVB limiter [27, 29] to the DG solution obtained from the DG coefficients $\widehat{U}^{n+\frac{1}{2},i*}$, which results in the updated DG coefficients $\widehat{U}^{n+\frac{1}{2},i}$.

end

$\widehat{U}^{n+\frac{1}{2}} = \widehat{U}^{n+\frac{1}{2},s}$.

Reaction part:

For N_r steps, do

(iii) Solve the DIRK-DG discretizations (3.24) and (3.25) of the reaction equations with a time step τ^{n+1}/N_r.

(iv) Apply the L^2-projection with the mass fraction constraints enforced.

end

Homogeneous part:

(v) Repeat steps (i)-(ii).

The numerical solution $\widehat{U}^{n+1}$ for the chemically reactive Euler equations at time t^{n+1} using the fractional step method (3.6) is obtained.

Remark 3.5.1. *Algorithm 2 results in a higher order bounds preserving DG discretization for stiff and non-stiff problems. For non-stiff problems,*

we only need to do steps (i)-(ii) of Algorithm 2 for the homogeneous part.

3.6 Numerical tests

In this section, we will present tests of the bounds preserving KKT-DIRK-DG discretizations of the chemically reactive Euler equations, both for problems with smooth and discontinuous solutions. We will also show test cases for the homogeneous Euler equations. In order to demonstrate the necessity of preserving the bounds on the solution and the accuracy-preserving property of the KKT-DIRK-DG discretizations, we will compare for test cases with a smooth solution the minimum and maximum values of the solution and the errors in the numerical solution of the KKT-DIRK-DG discretization with and without bounds constraints.

The time step τ is based on the CFL number, $\tau = \text{CFL} \cdot h/v_{ref}$, with reference velocity $v_{ref} = \left\| \sqrt{\gamma p_h/\rho_h} + |u_h| \right\|_\infty$. For numerical efficiency, it is important to have a good balance between the number of Newton iterations in each DIRK stage and the time step. If the Newton method does not converge within a predefined number of iterations, we restart the computation with time step $\tau/2$, while if the Newton algorithm converges well then the time step will be increased to 1.2τ until the maximum CFL number is obtained. In practice, this provides a good balance between the number of Newton iterations and the time step size, which is constantly adjusted to account for the dynamics in the reactive flow.

3.6.1 Euler equations

We first consider test cases of the homogeneous Euler equations.

Example 3.6.1. (Accuracy test) In order to demonstrate the necessity of imposing bounds on the density and pressure and the higher order accuracy of the KKT-DIRK-DG discretizations, we consider the homogeneous Euler equations for $\gamma = 1.4$ and the exact solution

$$\rho(x,t) = (1 + 0.9999\sin(x-t))/10,\ u(x,t) = 1,\ p(x,t) = 1,\ x \in [0, 2\pi].$$

We compare numerical results of the higher order DIRK-DG discretizations with and without imposing the positivity constraints. The positivity constraint $\rho_{\min}$ is taken here as 10^{-14} in order to account for small truncation errors. The CFL numbers are chosen as 1 for P_1 polynomials, 0.5

for P_2 polynomials. In Tables 3.1 and 3.2, we present the order of accuracy and the minimum value of ρ_h, with and without the KKT limiter. These tables show that the positivity-preserving KKT limiter preserves the nonnegativity of the density and does not negatively affect the order of accuracy. Without the KKT limiter, on relatively coarse meshes for P_1 basis functions there are negative values of ρ_h, which cases are marked with a cross. When the mesh resolution increases or for P_2 basis functions, which are more accurate, the negative density values disappear.

Table 3.1: (Example 3.6.1) Accuracy test for the homogeneous Euler equations without KKT limiter at time $t_T = 1$, a $\times$ indicates a result with a negative density.

	N_e	$L^\infty(\Omega)$ norm	Order	$L^1(\Omega)$ norm	Order	Minimum ρ_h
P_1	10	$\times$	–	$\times$	–	$\times$
	20	$\times$	$\times$	$\times$	$\times$	$\times$
	40	$\times$	$\times$	$\times$	$\times$	$\times$
	80	$\times$	$\times$	$\times$	$\times$	$\times$
	160	2.64E-005	2.01	4.54E-005	1.99	9.537068E-07
P_2	10	3.71E-004	–	5.88E-004	–	1.292250E-04
	20	4.98E-005	2.90	7.32E-005	3.01	6.385523E-05
	40	6.40E-006	2.96	9.14E-006	3.00	2.118774E-05
	80	8.11E-007	2.98	1.14E-006	3.00	1.018187E-05
	160	1.02E-007	2.99	1.43E-007	2.99	1.275602E-05

Table 3.2: (Example 3.6.1) Accuracy test for the homogeneous Euler equations with KKT limiter at time $t_T = 1$.

k	N_e	$L^\infty(\Omega)$ norm	Order	$L^1(\Omega)$ norm	Order	Minimum ρ_h
P_1	10	6.53E-003	–	1.15E-002	–	1.000068E-14
	20	1.78E-003	1.88	3.13E-003	1.88	1.000025E-14
	40	4.70E-004	1.92	7.09E-004	2.14	2.064498E-05
	80	1.25E-004	1.91	1.89E-004	1.91	2.998577E-05
	160	3.01E-005	2.05	4.63E-005	2.03	1.629531E-05
P_2	10	3.71E-004	–	6.04E-004	–	1.734736E-04
	20	4.98E-005	2.90	7.32E-005	3.04	6.385523E-05
	40	6.40E-006	2.96	9.14E-006	3.00	2.118774E-05
	80	8.11E-007	2.98	1.14E-006	3.00	1.018187E-05
	160	1.02E-007	2.99	1.43E-007	2.99	1.275602E-05

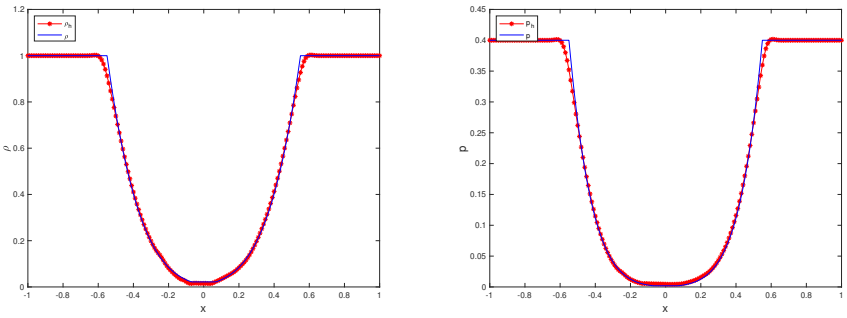
In Examples 3.6.2 and 3.6.3, the test cases are computed using P_2 poly-

nomials and the third order DIRK time-integration method [99].

Example 3.6.2. (Double rarefaction) In this test case, we consider the homogeneous Euler equations with initial solution

$$(\rho_L, u_L, p_L) = (1, -2, 0.4) \text{ in } [-1, 0], \quad (\rho_R, u_R, p_R) = (1, 2, 0.4) \text{ in } [0, 1].$$

We use 200 elements, $\gamma = 1.4$ and CFL $= 1.2$ for most time steps. The density and pressure are shown in Figure 3.1 together with the exact solution. The positivity preserving KKT limiter works well and ensures positivity of density and pressure. Without the KKT limiter, there will be some unphysical negative values of the density and pressure.



(a) (b)

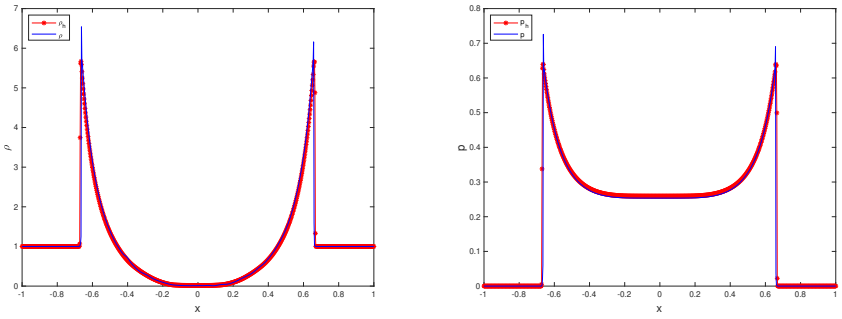
Figure 3.1: (Example 3.6.2) Solution of homogeneous Euler equations at time $t_T = 0.2$. (a) KKT-limited numerical solution ρ_h and exact solution ρ, (b) KKT-limited numerical solution p_h and exact solution p.

Example 3.6.3. (Sedov blast wave problem [75]) Next, we consider the homogeneous Euler equations for the initial solution

$$(\rho, u, p) = (1, 0, 10^{-9}) \text{ in } [-1, 1],$$

except in the central cell, where we set $p = 100$. These initial conditions result in the Sedov blast wave problem.

For this case, we use 800 elements, $\gamma = 1.4$ and CFL $= 1$ for most time steps. Since the initial solution for this test case is related to the number of elements, we compute the reference solutions also for 800 elements and CFL $= 0.05$ using the bounds preserving DG method in [36]. Figure 3.2 shows that the bounds preserving DIRK-DG discretization ensures positivity of the density and pressure. Without the positivity constraints, there will be some unphysical negative values of density and pressure.



(a) (b)

Figure 3.2: (Example 3.6.3) Solution of the homogeneous Euler equations for the Sedov blast wave problem at time $t_T = 0.5$. (a) KKT-limited numerical solution ρ_h and reference solution ρ, (b) KKT-limited numerical solution p_h and reference solution p.

3.6.2 Chemically reactive Euler equations

In this section, we will investigate the performance of the KKT-DIRK-DG algorithm on the chemically reactive Euler equations. As reference solutions, we use the algorithm in [36] with piecewise P_1 polynomials on a mesh with 5000 elements for Examples 3.6.5-3.6.7 and 2000 elements for Examples 3.6.8-3.6.9. For the time integration of the reference solution, we use a second-order accurate multistep time integration method with CFL $= 0.05$ for Examples 3.6.5-3.6.7 and CFL $= 0.01$ for Examples 3.6.8-3.6.9. In the following, we compute the test cases in Examples 3.6.5-3.6.7 with the KKT-DIRK-DG method using Algorithm 2 with P_2 polynomials, the third order fractional step method [70] and the third order DIRK time-discrete method [99]. The test cases in Examples 3.6.8-3.6.9 are computed with P_1 polynomials, the second order fractional step method (3.6) and the second order DIRK time-discrete method [5].

Example 3.6.4. (Accuracy test) By taking $N = 2$ and $u = 1,\ p = 0,\ s_1 = -cr^7$ in (3.1), we obtain the following convection-reaction system [36]

$$\begin{cases} \rho_t + \rho_x = 0, \\ r_t + r_x = -cr^7. \end{cases} \tag{3.32}$$

The parameter c is a constant and can be used to adjust the stiffness of the equations. The equations become more stiff as c increases. In this test

case, we take $c = 10000$, which makes (3.32) a very difficult test case to compute. The initial solutions are given as

$$\rho(x,0) = (2+\sin(x)+\cos(x))/10, \quad r(x,0) = (1+\sin(x))/10, \quad x \in [0, 2\pi].$$

We consider second and third order accurate DIRK-DG discretizations with and without the bounds preserving KKT limiter. We take CFL = 1 for P_1 polynomials and CFL = 0.5 for P_2 polynomials. The number of intermediate reaction steps $N_r = 2$. The mass fraction r_h/ρ_h should be between zero and one. The minimum and maximum values of r_h/ρ_h, with and without KKT limiter, are shown in Tables 4.3 and 4.4. These tables show that Algorithm 2 maintains an order of accuracy $O(h^{k+1})$ for a polynomial order k and preserves the bounds. This indicates that the KKT limiter works properly and does not harm the accuracy. Without the KKT limiter, the bounds are violated on relatively coarse meshes.

Table 4.3: (Example 3.6.4) Accuracy test of convection-reaction system (3.32) without KKT limiter at time $t_T = 1$.

	N_e	$L^\infty(\Omega)$ norm	Order	$L^1(\Omega)$ norm	Order	Minimum r_h/ρ_h	Maximum r_h/ρ_h
P_1	10	5.82E-003	–	9.48E-003	–	-6.210755E-02	1.031343E+00
	20	1.60E-003	1.86	2.33E-003	2.02	-1.521221E-02	1.004196E+00
	40	4.38E-004	1.87	5.74E-004	2.02	-2.557660E-03	9.943711E-01
	80	1.15E-004	1.93	1.43E-004	2.01	-7.806174E-04	9.925336E-01
	160	2.91E-005	1.98	3.58E-005	2.00	-9.150088E-05	9.916861E-01
P_2	10	9.06E-004	–	7.70E-004	–	3.176584E-04	9.923696E-01
	20	1.19E-004	2.93	8.59E-005	3.16	4.304911E-04	9.917059E-01
	40	1.63E-005	2.87	1.07E-005	3.01	1.010852E-04	9.918443E-01
	80	2.08E-006	2.97	1.33E-006	3.01	2.563580E-07	9.916889E-01
	160	2.62E-007	2.99	1.68E-007	2.98	2.756807E-05	9.916604E-01

Table 4.4: (Example 3.6.4) Accuracy test of convection-reaction system (3.32) with KKT limiter at time $t_T = 1$.

	N_e	$L^\infty(\Omega)$ norm	Order	$L^1(\Omega)$ norm	Order	Minimum r_h/ρ_h	Maximum r_h/ρ_h
P_1	10	6.13E-003	–	1.14E-002	–	9.993497E-15	9.861364E-01
	20	1.67E-003	1.88	2.72E-003	2.07	1.000095E-14	9.959969E-01
	40	4.38E-004	1.93	6.34E-004	2.10	4.972616E-04	9.933924E-01
	80	1.15E-004	1.93	1.52E-004	2.06	6.788640E-05	9.923583E-01
	160	2.91E-005	1.98	3.71E-005	2.03	1.083892E-04	9.916715E-01
P_2	10	9.08E-004	–	7.87E-004	–	1.352766E-03	9.921466E-01
	20	1.19E-004	2.93	8.70E-005	3.18	4.833784E-04	9.916945E-01
	40	1.63E-005	2.87	1.16E-005	2.91	8.211523E-05	9.914553E-01
	80	2.08E-006	2.97	1.36E-006	3.09	1.964713E-06	9.916889E-01
	160	2.62E-007	2.99	1.71E-007	2.99	2.794314E-05	9.916604E-01

Example 3.6.5. (Detonation in two species gas [10]) In this test case, we consider two species in the chemically reactive Euler equations (3.1) with source term

$$s_1 = -K(T)\rho z,$$

where $\gamma = 1.2,\ q_1 = 50,\ q_2 = 0$ and

$$K(T) = \begin{cases} 230.75, & T > 3, \\ 0, & T \leqslant 3. \end{cases}$$

The computational domain is $[0, 100]$ and the initial solution is defined as

$$(\rho, u, p, z_1, z_2) = \begin{cases} (2.0, 4.0, 40.0, 0.0, 1.0), & x \leqslant 10, \\ (3.64282, 6.2489, 54.8244, 0.0, 1.0), & 10 < x \leqslant 20, \\ (1.0, 0.0, 1.0, 1.0, 0.0), & x > 20. \end{cases}$$

The exact solution consists of a right moving detonation wave, a right moving rarefaction wave, a right moving contact discontinuity, and a left moving rarefaction wave before the right moving rarefaction catches the detonation wave.

For this test case, we use 400 elements. In order to have a good balance between the number of Newton iterations and the time step, we take the maximum CFL number as CFL $= 0.1$. In this test case, $N_r = 1$ intermediate reaction steps already ensures the correct propagation speed of discontinuities, but in order to obtain a more accurate numerical solution, we take $N_r = 10$. Figure 3.3 shows that the bounds for density, pressure and mass fraction are preserved and all waves are captured correctly, which implies that Algorithm 2 is able to compute the correct position and speed of the detonation wave.

Example 3.6.6. (Two detonations in a two species gas [10]) The parameters $\gamma,\ q_1,\ q_2,\ K(T)$ used in this test case are similar to those in Example 3.6.5. The computational domain is $[0, 100]$ and the initial solution is defined as

$$(\rho, u, p, z_1, z_2) = \begin{cases} (1.79463, 3.0151, 30.0, 0.0, 1.0), & x \leqslant 10, \\ (1.0, 0.0, 1.0, 1.0, 0.0), & 10 < x \leqslant 90, \\ (1.79463, -8.0, 21.53134, 0.0, 1.0), & x > 90. \end{cases}$$

The exact solution contains a right moving detonation and a left moving strong detonation. After some time, there is a collision between the two detonations.

In this example, we use 400 elements and the number of intermediate reaction steps is set to $N_r = 24$. In order to have a good balance between the number of Newton iterations and the time step, we take the maximum

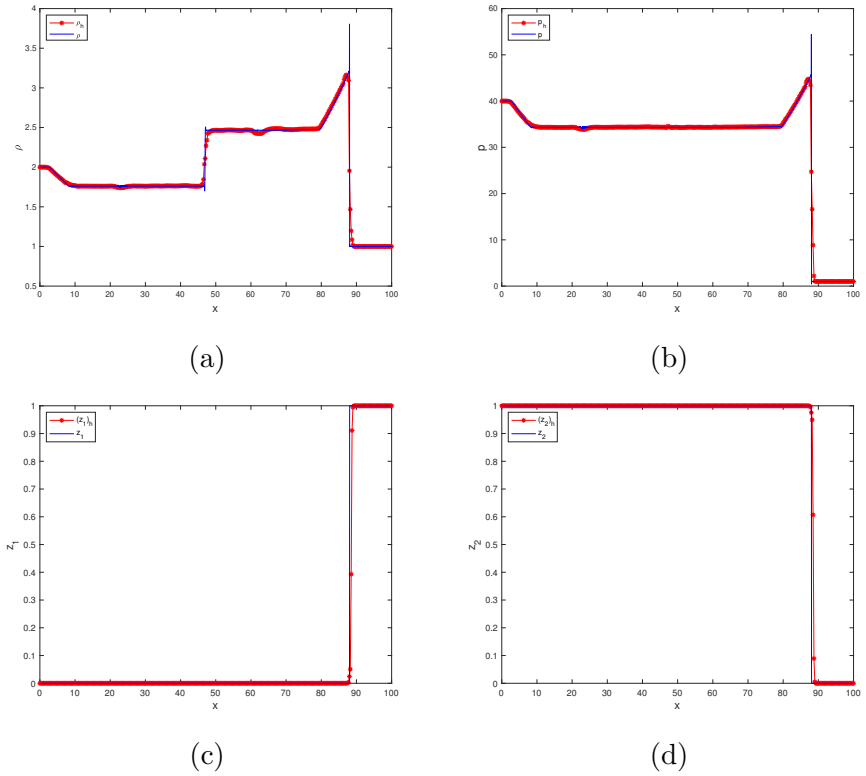


Figure 3.3: (Example 3.6.5) KKT-limited numerical solution of chemically reactive Euler equations at time $t_T = 8$, mesh 400 elements, CFL $= 0.1$, $N_r = 10$. Reference solution at time $t_T = 8$, mesh 5000 elements, CFL $= 0.05$ obtained using the algorithm in [36]. (a) KKT-limited numerical solution ρ_h and reference solution ρ, (b) KKT-limited numerical solution p_h and reference solution p, (c-d) KKT-limited numerical solutions $(z_1)_{h2}$, $(z_2)_{h2}$ and reference solutions z_1, z_2.

CFL number as CFL $= 0.1$. The profiles of density, pressure, and mass fraction are shown in Figure 3.4. Clearly, the shock speed and position are captured well. Also, the density and pressure are positive, and all mass fractions are between zero and one.

Example 3.6.7. (Detonation wave with three species and one reaction [11, 117]) We consider the one-step chemical model (3.1) for a hydrogen-oxygen mixture

$$2H_2 + O_2 \rightarrow 2H_2O.$$

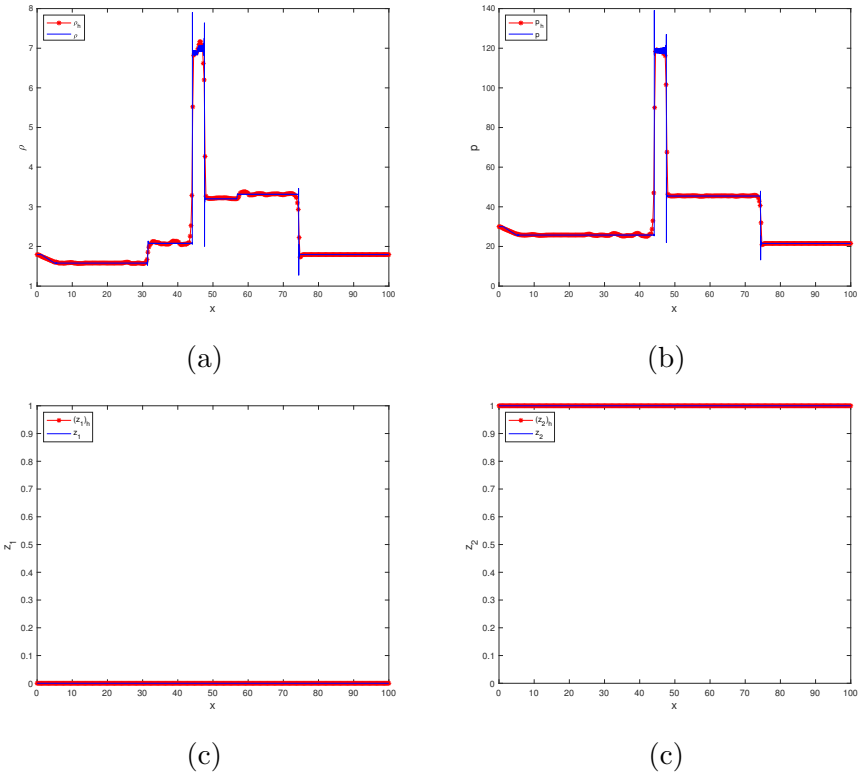


Figure 3.4: (Example 3.6.6) KKT-limited numerical solution of chemically reactive Euler equations at time $t_T = 6$, mesh 400 elements, CFL $= 0.1$, $N_r = 24$. Reference solution at time $t_T = 6$, mesh 5000 elements, CFL $= 0.05$ obtained using the algorithm in [36]. (a) KKT-limited numerical solution ρ_h and reference solution ρ, (b) KKT-limited numerical solution p_h and reference solution p, (c) KKT-limited numerical solutions $(z_1)_{h2}$, $(z_2)_{h2}$ and reference solutions z_1, z_2.

The parameters in (3.2)-(3.4) are chosen as $\gamma = 1.4$, $T_1 = 2$, $B_1 = 10^6$, $\alpha_1 = 0$, $q_1 = 100$, $q_2 = q_3 = 0$, $M_1 = 2$, $M_2 = 32$, $M_3 = 18$. The computational domain is $[0, 50]$ and the initial solution is defined as

$$(\rho, u, p, z_1, z_2, z_3) = \begin{cases} (2.0, 8.0, 20.0, 0.0, 0.0, 1.0), & x \leqslant 2.5, \\ (1.0, 0.0, 1.0, 1/9, 8/9, 0.0), & x > 2.5, \end{cases}$$

with z_1 the mass fraction of H_2, z_2 the mass fraction of O_2, and z_3 the mass fraction of H_2O.

The exact solution consists of a detonation wave, followed by a contact discontinuity and a shock, all moving to the right. In this example, we use

400 elements and the number of intermediate reaction steps $N_r = 20$. In order to have a good balance between the number of Newton iterations and the time step, we take the maximum CFL number as CFL $= 0.1$. Also, in this case $N_r = 1$ is already sufficient to obtain the correct propagation speed of discontinuities. We compare the results with the algorithm in [36] for the same number of elements and CFL number as used for the KKT-DIRK-DG algorithm. The reference solution is obtained using the algorithm in [36] on a mesh with 5000 elements and CFL $= 0.05$.

We observe in Figure 3.5 spurious numerical solutions when the bounds preserving DG method in [36] is used on the same mesh and CFL number as the KKT-DIRK-DG method. On a much finer mesh with 5000 elements, CFL $= 0.05$ the results of [36] are the same as for the KKT-DIRK-DG discretization. All discontinuities for density, pressure and mass fractions are captured correctly by the KKT-DIRK-DG discretization on the 400 element mesh, which indicates that our algorithm is already accurate on a considerably coarser mesh than the method presented in [36].

Example 3.6.8. (Detonation wave with four species and one reaction [11, 117]) We consider the chemically reactive Euler equations (3.1) with four species and the reaction

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O.$$

The parameters in (3.2)-(3.4) are chosen as $\gamma = 1.4$, $T_1 = 2$, $B_1 = 10^6$, $\alpha_1 = 0$, $q_1 = 500$, $q_2 = q_3 = q_4 = 0$, $M_1 = 16$, $M_2 = 32$, $M_3 = 44$, $M_4 = 18$. The computational domain is $[0, 10]$ and the initial solution is given as

$$(\rho, u, p, z_1, z_2, z_3, z_4) = \begin{cases} (2.0, 10.0, 40.0, 0.0, 0.2, 0.475, 0.325), & x \leqslant 2.5, \\ (1.0, 0.0, 1.0, 0.1, 0.6, 0.2, 0.1), & x > 2.5, \end{cases}$$

with z_1 the mass fraction of CH_4, z_2 the mass fraction of O_2, z_3 the mass fraction of CO_2, and z_4 the mass fraction of H_2O.

The exact solution consists of a detonation wave followed by a contact discontinuity and a shock, all moving to the right. In this example, we use 300 elements and take $N_r = 1$, CFL $= 0.2$. The reference solution is obtained using the algorithm in [36] on a mesh with 2000 elements and CFL $= 0.01$. Figure 3.6 shows that all shock waves and wave speeds for the density, pressure and mass fractions are captured correctly and the bounds preserving KKT-DIRK-DG discretization works well in this multispecies example.

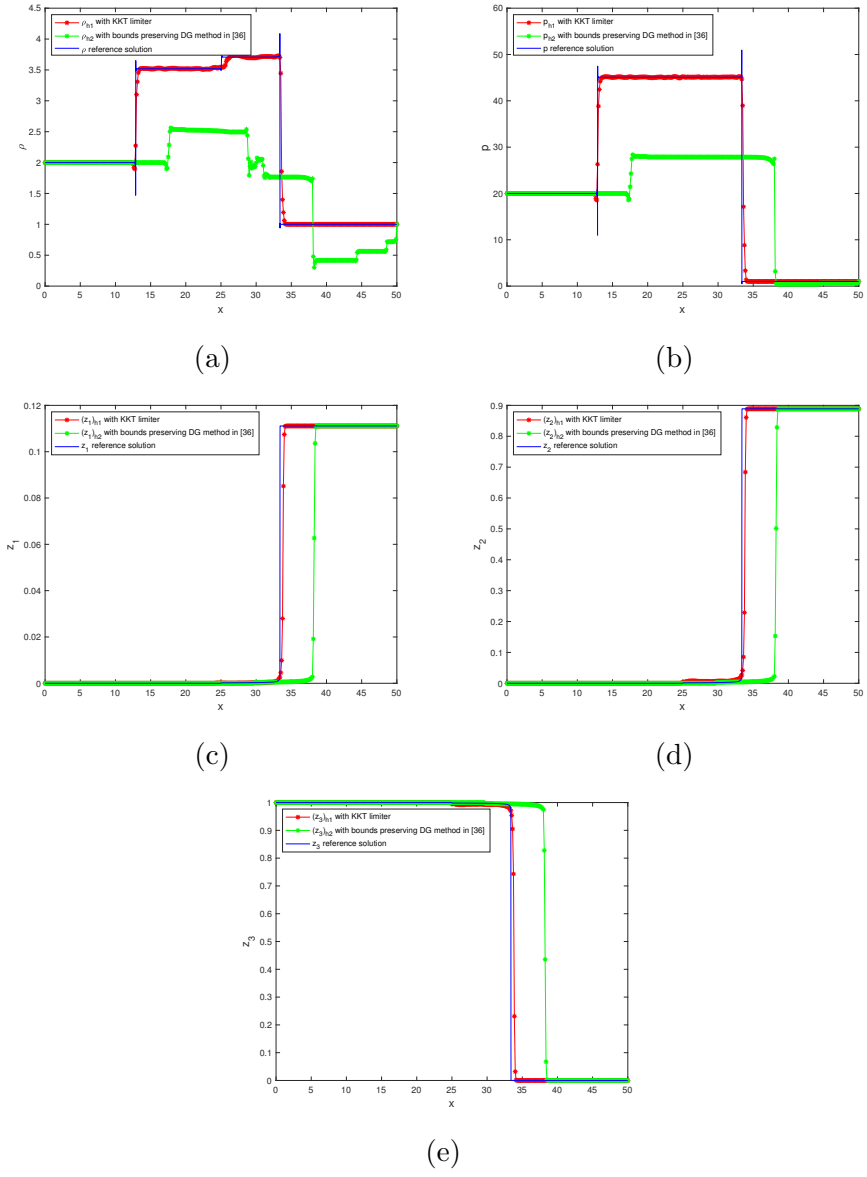


Figure 3.5: (Example 3.6.7) Numerical solutions of chemically reactive Euler equations at time $t_T = 4$, mesh 400 elements, CFL $= 0.1$, $N_r = 20$. Also shown, solutions at time $t_T = 4$ using the algorithm [36] on a mesh with 400 elements, CFL $= 0.1$, and for 5000 elements, CFL $= 0.05$ (reference solution). (a) numerical solution ρ_{h1} obtained with KKT-DIRK-DG discretization, numerical solution ρ_{h2} obtained with bounds preserving DG method in [36] and reference solution ρ, (b) numerical solution p_{h1} obtained with KKT-DIRK-DG discretization, numerical solution p_{h2} obtained with bounds preserving DG method in [36] and reference solution p, (c-e) numerical solutions $(z_1)_{h1},\ (z_2)_{h1},\ (z_3)_{h1}$ obtained with KKT-DIRK-DG discretization, numerical solutions $(z_1)_{h2},\ (z_2)_{h2},\ (z_3)_{h2}$ obtained with bounds preserving DG method in [36], and reference solutions $z_1,\ z_2,\ z_3$, with $z_1,\ z_2,\ z_3$, respectively, the H_2, O_2 and H_2O mass fraction.

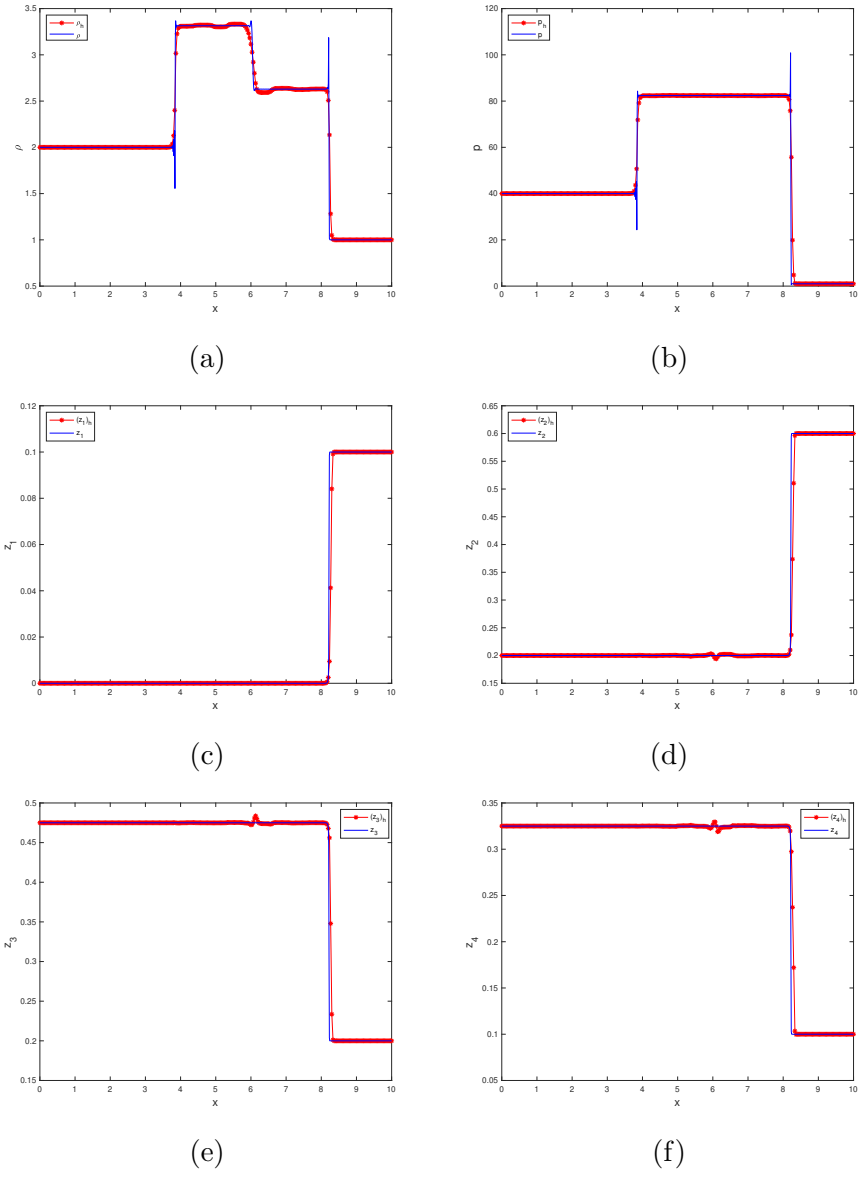


Figure 3.6: (Example 3.6.8) KKT-limited numerical solution of chemically reactive Euler equations at time $t_T = 0.5$, mesh 300 elements, CFL $= 0.2$, $N_r = 1$. Reference solution at time $t_T = 0.5$, mesh 2000 elements, CFL $= 0.01$ using the algorithm in [36]. (a) KKT-limited numerical solution ρ_h and reference solution ρ, (b) KKT-limited numerical solution p_h and reference solution p, (c-f) KKT-limited numerical solutions $(z_1)_h$, $(z_2)_h$, $(z_3)_h$, $(z_4)_h$ and the corresponding reference solutions. Here $z_1, \ldots, z_4$ denote, respectively, the mass fractions of CH_4, O_2, CO_2 and H_2O.

Example 3.6.9. (Detonation wave with five species and two reactions [11, 36, 117]) Consider the chemically reactive Euler equations (3.1) with a two-step chemical model with 5 species for a hydrogen-oxygen-nitrogen mixture

$$H_2 + O_2 \to 2OH, \quad 2OH + H_2 \to 2H_2O,$$

with nitrogen appearing as a catalyst. The parameters in (3.2)-(3.4) are chosen as $\gamma = 1.4,\ T_1 = 2,\ T_2 = 10,\ B_1 = B_2 = 10^6,\ \alpha_1 = \alpha_2 = 0,\ q_1 = q_2 = q_5 = 0,\ q_3 = -20,\ q_4 = -100,\ M_1 = 2,\ M_2 = 32,\ M_3 = 17,\ M_4 = 18,\ M_5 = 28$. The computational domain is $[0, 10]$ and the initial solution is given as

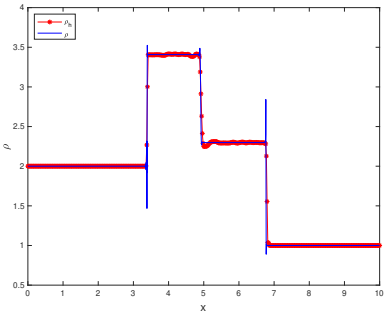
$$(\rho, u, p, z_1, z_2, z_3, z_4, z_5) = \begin{cases} (2.0, 10.0, 40.0, 0.0, 0.0, 0.17, 0.63, 0.2), & x \leqslant 2.5, \\ (1.0, 0.0, 1.0, 0.08, 0.72, 0.0, 0.0, 0.2), & x > 2.5, \end{cases}$$

with z_1 the mass fraction of H_2, z_2 the mass fraction of O_2, z_3 the mass fraction of OH, z_4 the mass fraction of H_2O, and z_5 the mass fraction of N_2.

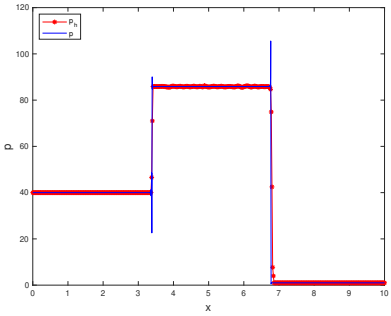
The exact solution consists of a detonation wave followed by a rarefaction wave and a shock, all moving to the right. In this example, we use 500 elements and take $N_r = 20$. For most time steps CFL $= 0.1$. The reference solution is obtained using the algorithm in [36] on a mesh with 2000 elements and CFL $= 0.01$. Figure 3.7 shows that the density and pressure are positive, and all mass fractions are between zero and one. Algorithm 2 is able to capture the correct propagation speed and position of the detonation wave and works well in this multispecies and multireaction example.

3.7 Conclusions

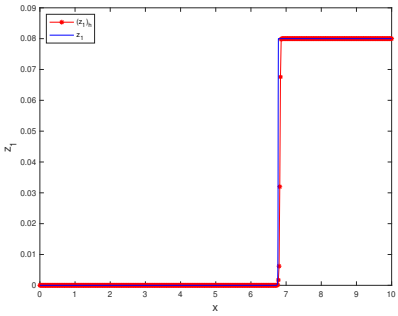
In this chapter, we propose a higher order bounds preserving time-implicit KKT-DIRK-DG algorithm for the chemically reactive Euler equations modelling multispecies and multireaction chemically reactive flows. This algorithm combines several important features when solving stiff chemically reacting gas flows, namely, higher order accuracy, preservation of the physical bounds on the density, pressure and mass fractions, good accuracy on coarse meshes and large time steps compared to existing time explicit methods, e.g. [36]. In addition, we can consider Algorithm 2 as a template



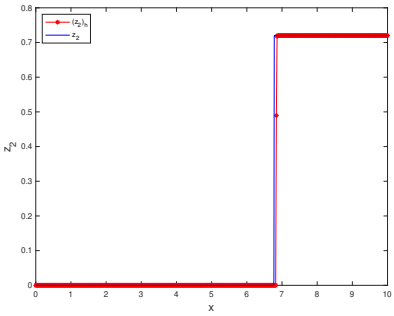
(a)



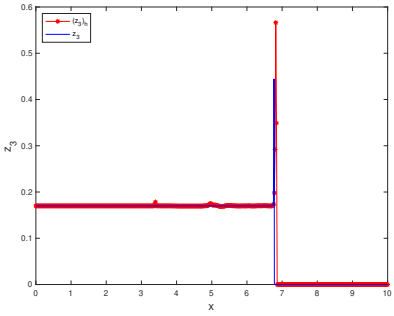
(b)



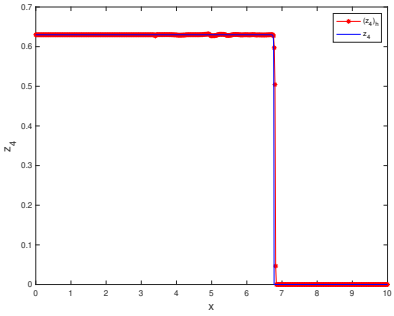
(c)



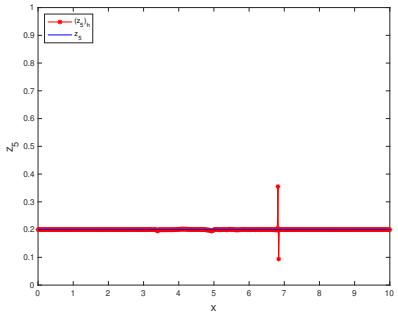
(d)



(e)



(f)



(g)

Figure 3.7: (Example 3.6.9) KKT-limited numerical solution of chemically reactive Euler equations at time $t_T = 0.35$, mesh 500 elements, CFL $= 0.1$, $N_r = 20$. Reference solution at time $t_T = 0.35$, mesh 2000 elements, CFL $= 0.01$ using the algorithm in [36]. (a) KKT-limited numerical solution ρ_h and reference solution ρ, (b) KKT-limited numerical solution p_h and reference solution p, (c-g) KKT-limited numerical solutions $(z_1)_h$, $(z_2)_h$, $(z_3)_h$, $(z_4)_h$, $(z_5)_h$ and the corresponding reference solutions. Here $z_1, \ldots, z_5$ denote, respectively, the mass fractions of H_2, O_2, OH, H_2O and N_2.

to solve both stiff and non-stiff chemically reacting gases with strict preservation in the numerical solution of the physical bounds. The KKT-DIRK-DG algorithm is already used in [111] for two dimensional parabolic PDEs and extension of the algorithm to the two dimensional chemically reactive Euler equations will be considered in future work. The extension of the constraints imposed on the reaction equations from one to two dimensions is discussed in [117, 127]. A disadvantage of the presented KKT-DIRK-DG algorithm is its dependence on operator splitting methods. Higher order accurate splitting methods become rather involved, requiring many intermediate steps. Numerical results demonstrate the optimal order of accuracy for smooth problems and excellent preservation of the bounds when using the KKT-DIRK-DG discretizations for the chemically reactive Euler equations.

Chapter 4

Stability Analysis and Error Estimates of Local Discontinuous Galerkin Methods with Semi-Implicit Spectral Deferred Correction Time-Marching for the Allen-Cahn Equation[1]

Abstract

This chapter is concerned with stability and error estimates of Local Discontinuous Galerkin (LDG) discretizations coupled with semi-implicit Spectral Deferred Correction (SDC) time integration methods up to third order accuracy for the Allen-Cahn equation. Since the SDC method is based on a first order convex splitting scheme, the implicit treatment of the nonlinear terms results each time step in a nonlinear system of equations, which increases the difficulty of the theoretical analysis. For the LDG discretizations coupled with second and third order accurate SDC methods, we prove the unique solvability of the numerical solutions through a standard fixed point argument in finite dimensional spaces. By carefully choosing the test functions, we prove energy stability with an upper bound for the time step that is independent of the mesh size. In addition, we derive optimal error estimates for the fully discrete LDG-SDC discretization. Numerical examples are presented to illustrate our theoretical results.

[1]Based on: F. Yan, Y. Xu. Stability Analysis and Error Estimates of Local Discontinuous Galerkin Methods with Semi-Implicit Spectral Deferred Correction Time-Marching for the Allen-Cahn Equation. *Journal of Computational and Applied Mathematics*, 376(2020), 112857.

4.1 Introduction

Let Ω be a bounded domain with dimension $d \leq 3$ and $0 < T < \infty$. We analyze Local Discontinuous Galerkin (LDG) discretizations coupled with semi-implicit Spectral Deferred Correction (SDC) time integration methods for the Allen-Cahn equation

$$\begin{cases} u_t - \Delta u + \dfrac{1}{\varepsilon^2} f(u) = 0, & \text{in } \Omega \times (0, T], \\ u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), & \text{in } \Omega, \end{cases} \tag{4.1}$$

with the Neumann boundary condition

$$\frac{\partial u}{\partial \nu} = 0, \quad \text{at } \partial\Omega \times (0, T], \tag{4.2}$$

where $f(u) = \Psi'(u)$ and $\Psi(u) = \dfrac{1}{4}(1 - u^2)^2$. Since Allen-Cahn equation satisfies a maximum principle [82], the solution u in (4.1) will take values in a bounded interval. In the bounded interval, we take $f(u) = \Psi'(u)$, while outside the interval, f is chosen such that f is globally derivable Lipschitz continuous. Then we can assume,

$$\max_{u \in \mathbb{R}} |f'(u)| \leq C_L, \tag{4.3}$$

where C_L is a positive constant, and we might as well require $C_L > 1$.

In order to describe the motion of anti-phase boundaries in crystalline solids, Allen and Cahn [6] originally proposed the well-known Allen-Cahn equation. Subsequently, numerous numerical studies have been devoted to the Allen-Cahn equation, for instance, using finite difference methods [3, 12], finite element methods [47, 48, 50, 130], Discontinuous Galerkin (DG) methods [46, 129] and LDG methods [55]. For the time integration first order accurate time integration methods [3, 46, 47, 130], first and second-order accurate implicit-explicit (IMEX) methods [48], implicit Additive Runge-Kutta (ARK) methods and Diagonally Implicit Runge-Kutta (DIRK) methods [129] have been used. Recently, Guo *et al.* [55] presented LDG schemes for the Allen-Cahn equation that are coupled with semi-implicit SDC time integration methods. The authors in [55] did not theoretically analyze the stability and error results of the second and third order accurate SDC-LDG discretizations.

For some Partial Differential Equations (PDEs), especially those with nonlinear terms, we often need to use higher order accurate numerical dis-

cretizations in space and time to get sufficiently accurate numerical solutions. Recently, the stability was analyzed and error estimates were obtained for LDG discretizations combined with IMEX Runge-Kutta (RK) time discretizations up to third order accuracy for the one-dimensional linear advection-diffusion equation [114], the one-dimensional nonlinear convection-diffusion equation [115] and the multi-dimensional nonlinear convection-diffusion equation [116]. Stability and error estimates were obtained in the sense that the time step Δt is only required to be upper-bounded by a positive constant independent of the mesh size h. In [48], the authors consider first and second-order accurate IMEX finite element discretizations in one and multiple dimensions for the Allen-Cahn equation and prove energy stability in a similar sense as in [114, 115, 116]. The purpose of this chapter is to study the stability and obtain error estimates for LDG discretizations combined with second and third order accurate SDC time integration methods for the Allen-Cahn equation.

The SDC method, as well as the integral deferred correction (InDC) method [16], is based on low order time integration methods, followed by iterative accuracy improvements. In comparison with RK methods, the SDC method is easy to construct for any order of accuracy. More general information about semi-implicit SDC methods coupled with an LDG discretization can be found in [57, 120]. Applications of the SDC method are presented in [49, 55, 59, 82].

The LDG method belongs to the class of DG methods. DG methods are finite element methods with discontinuous, piecewise polynomials as basis functions, which were first proposed by Reed and Hill in [96]. DG methods have many advantages over other finite element methods, such as suitability for highly nonuniform and unstructured meshes, mesh adaptation and parallel computing. For more information, we refer to [25, 27, 28, 29]. By extending the DG method, Cockburn and Shu in [30] introduced the LDG method to deal with PDEs that contain second order spatial derivatives. The idea of the LDG method is to apply the DG method after rewriting higher order equations as a system of first order equations. We refer for general information about the LDG method for linear cases to [26, 35, 114, 125] and for nonlinear cases to [9, 56, 60, 121, 122, 123] .

The main contribution of this chapter is to prove stability and error estimates of LDG discretizations coupled with second and third order accurate SDC time discretizations for the Allen-Cahn equation (4.1)-(4.2). Since the implicit treatment of the nonlinear term u^3 results in a nonlinear system, we prove the unique solvability of the fully-discrete numerical discretizations by using a standard fixed point argument in finite dimensional

spaces. Compared with Runge-Kutta type semi-implicit time integration methods [114], the SDC time discretization is, however, more difficult to analyze. For the stability analysis of the third order accurate SDC-LDG scheme and the error estimates of the second and third order accurate SDC-LDG schemes, we will extensively use property (4.3) to deal with the nonlinear term in the Allen-Cahn equation. By a careful selection of the test functions, energy stability and error estimates for the second and third order accurate time-discrete LDG schemes are obtained in the sense that the time step Δt requires only a positive upper bound, which is independent of the mesh size h.

The rest of this chapter is organized as follows. In Section 4.2, we will introduce some notations, projection operators, and the SDC scheme that will be used in the following analysis. In Section 4.3, we will present the LDG discretization combined with a second order semi-implicit SDC method for the Allen-Cahn equation (4.1), and prove unique solvability, stability and error estimates. A similar analysis for the third order SDC-LDG discretization will be presented in Section 4.4. In Section 4.5, numerical results are provided to verify the theoretical analysis. Concluding remarks are given in Section 4.6.

4.2 Preliminaries

In this section, we will introduce the finite element spaces, some notations, the definition of norms, and the SDC scheme to be used later in this chapter. We will also present some projection operators and related interpolation properties for the finite element spaces that will be used in the error analysis.

4.2.1 Finite element spaces

Let $\mathcal{T}_h$ be a regular subdivision of Ω with line, rectangular or cubic elements K in, respectively, 1D, 2D or 3D, Γ denotes the union of the boundary of elements $K \in \mathcal{T}_h$, i.e. $\Gamma = \cup_{K\in\mathcal{T}_h} \partial K$, and $\mathcal{Q}_{kk}(K)$ denotes the space of tensor product polynomials of degree at most $k \geq 0$ on each element K. In particular, we have $\mathcal{Q}_{kk}(K) = \mathcal{P}_k(K)$ in one dimension.

The finite element spaces V_h^k and $\boldsymbol{W}_h^k$ are defined as

$$V_h^k = \{v \in L^2(\Omega) : \quad v|_K \in \mathcal{Q}_{kk}(K), \quad \forall K \in \mathcal{T}_h\},$$
$$\boldsymbol{W}_h^k = \{\boldsymbol{w} \in [L^2(\Omega)]^d : \quad \boldsymbol{w}|_K \in [\mathcal{Q}_{kk}(K)]^d, \quad \forall K \in \mathcal{T}_h\},$$

which spaces are allowed to have discontinuities across element faces. Let e be an interior edge shared by the "left" and "right" elements, denoted K_L and K_R. If u is a function on K_L and K_R, we set $u^L \doteq (u|_{K_L})|_e$ and $u^R \doteq (u|_{K_R})|_e$.

4.2.2 Notations

For a positive integer N, let $0 = t_0 < t_1 < \ldots < t_N = T$ be a given partition of $[0, T]$ with time step $\Delta t = \dfrac{T}{N}$, and $t_n = n\Delta t,\ n = 0, 1, \ldots, N$. Note that $u_n = u(\cdot, t_n),\ \boldsymbol{q}_n = \boldsymbol{q}(\cdot, t_n)$. We denote U_n and $\boldsymbol{Q}_n$ as the approximate values of u and $\boldsymbol{q}$ at t_n $(n = 0, 1, \ldots, N)$, respectively.

For convenience in the analysis, we denote throughout this chapter by C a positive constant independent of h , which may depend on the solutions of our problems.

4.2.3 Inner products and norms

The inner products are denoted by

$$(u, v)_K = \int_K uv dK, \quad (u, v)_{\partial K} = \int_{\partial K} uv ds,$$
$$(\boldsymbol{p}, \boldsymbol{q})_K = \int_K \boldsymbol{p} \cdot \boldsymbol{q} dK, \quad (\boldsymbol{p}, \boldsymbol{q})_{\partial K} = \int_{\partial K} \boldsymbol{p} \cdot \boldsymbol{q} ds,$$

for the scalar variables u, v and the vector variables $\boldsymbol{p}$, $\boldsymbol{q}$, respectively. For any positive integer i, we define some norms over the domain Ω as

$$\|\eta\|_{L^2(\Omega)} = \left(\sum_{K \in \mathcal{T}_h} \|\eta\|^2_{L^2(K)} \right)^{\frac{1}{2}},$$
$$\|\eta\|_{H^i(\Omega)} = \left(\sum_{K \in \mathcal{T}_h} \|\eta\|^2_{H^i(K)} \right)^{\frac{1}{2}},$$
$$\|\eta\|_{L^\infty(\Omega)} = \max_{K \in \mathcal{T}_h} \left(\operatorname{ess\,sup}_{x \in K} |\eta(x)| \right),$$

where

$$\|\eta\|_{L^2(K)} = (\eta, \eta)^{\frac{1}{2}}_K, \quad \|\eta\|_{H^i(K)} = \left(\sum_{|\alpha| \le i} \|D^\alpha \eta\|^2_{L^2(K)} \right)^{\frac{1}{2}}.$$

For simplicity, we denote $\|\eta\| := \|\eta\|_{L^2(\Omega)},\ (u, v) := (u, v)_\Omega$.

4.2.4 Projections and properties

In what follows, we will introduce projections for one-dimensional and multi-dimensional problems.

- **One-dimension**

For all $u \in H^1(\Omega)$, we define the interpolation operators $P^{\pm}$ as

$$P^{\pm} : H^1(\Omega) \to V_h^k,$$

equipped with

$$\left(P^+u, v\right)_{K_j} = (u, v)_{K_j}, \quad \forall v \in \mathcal{P}_{k-1}(K_j), \quad P^+u(x_{j-1}) = u(x_{j-1}), \tag{4.4}$$

$$\left(P^-u, v\right)_{K_j} = (u, v)_{K_j}, \quad \forall v \in \mathcal{P}_{k-1}(K_j), \quad P^-u(x_j) = u(x_j), \tag{4.5}$$

where $K_j = (x_{j-1}, x_j)$. If $u \in H^{k+1}(\Omega)$, there holds (see [35])

$$\|u - P^{\pm}u\| \le Ch^{k+1}\|u\|_{H^{k+1}(\Omega)}.$$

- **Multi-dimensions**

For the two-dimensional case, we describe the projection operator P^- for scalar functions as

$$P^- = P_x^- \otimes P_y^-,$$

where the subscripts x and y denote the one-dimensional projections defined in (4.5) on a rectangular element $\mathcal{X} \otimes \mathcal{Y} = [x_{j-1}, x_j] \times [y_{j-1}, y_j]$.

Given that π_x and π_y are the standard L^2 projections in the x and y direction, respectively, the projection Π^+ for vector-valued functions $\boldsymbol{\phi} = (\phi_1(x, y), \phi_2(x, y)) \in [H^1(\Omega)]^2$ is defined by

$$\Pi^+\boldsymbol{\phi} = \left(P_x^+ \otimes \pi_y\phi_1, \pi_x \otimes P_y^+\phi_2\right) : [H^1(\Omega)]^2 \to [\mathcal{Q}_{kk}(\mathcal{X} \otimes \mathcal{Y})]^2,$$

which satisfies

$$\left(\Pi^+\boldsymbol{\phi} - \boldsymbol{\phi}, \nabla w\right)_{\mathcal{X}\otimes\mathcal{Y}} = 0, \quad \forall w \in \mathcal{Q}_{kk}(\mathcal{X} \otimes \mathcal{Y}),$$

and

$$\left(\left(\Pi^+\boldsymbol{\phi}(x_{i-1}, \cdot) - \boldsymbol{\phi}(x_{i-1}, \cdot)\right) \cdot \boldsymbol{\nu}, w(x_{i-1}^+, \cdot)\right)_{\mathcal{Y}} = 0, \quad \forall w \in \mathcal{Q}_{kk}(\mathcal{X} \otimes \mathcal{Y}),$$

$$\left(\left(\Pi^+\boldsymbol{\phi}(\cdot, y_{j-1}) - \boldsymbol{\phi}(\cdot, y_{j-1})\right) \cdot \boldsymbol{\nu}, w(\cdot, y_{j-1}^+)\right)_{\mathcal{X}} = 0, \quad \forall w \in \mathcal{Q}_{kk}(\mathcal{X} \otimes \mathcal{Y}).$$

For the three-dimensional case, we refer to [26].

The projections defined above have the following approximation properties. If $u \in H^{k+1}(\Omega)$, $\boldsymbol{\phi} \in [H^{k+1}(\Omega)]^2$, we have (see [35])

$$\|P^- u - u\| \leq Ch^{k+1}\|u\|_{H^{k+1}(\Omega)}, \tag{4.6}$$

$$\|\Pi^+ \boldsymbol{\phi} - \boldsymbol{\phi}\| \leq Ch^{k+1}\|\boldsymbol{\phi}\|_{H^{k+1}(\Omega)}. \tag{4.7}$$

The projection P^- on Cartesian meshes has the following superconvergence property (see Lemma 3.7 in [35]).

Lemma 4.2.1. *Assume $\eta \in H^{k+2}(\Omega)$, $\boldsymbol{\rho} \in \boldsymbol{W}_h^k$, then the projection P^- satisfies*

$$|(\eta - P^-\eta, \nabla \cdot \boldsymbol{\rho})_\Omega - (\eta - \widehat{P^-\eta}, \boldsymbol{\rho} \cdot \boldsymbol{\nu})_\Gamma| \leq Ch^{k+1}\|\eta\|_{H^{k+2}(\Omega)}\|\boldsymbol{\rho}\|_\Omega,$$

with $\widehat{P^-\eta} = (P^-\eta)^L$.

4.2.5 Spectral deferred correction scheme

Dutt, Greengard and Rokhlin in [39] constructed the SDC method to obtain high order accurate stable time integration methods. Next, Minion in [89] presented the semi-implicit SDC time integration method. Here we will only discuss the second and third order semi-implicit SDC methods proposed by Minion in [89].

Consider the ODE system

$$\begin{cases} u_t = F_S(t, u(t)) + F_N(t, u(t)), \quad t \in [0, T], \\ u(0) = u_0, \end{cases} \tag{4.8}$$

where F_N is a non-stiff term and F_S is a stiff term. For the Allen-Cahn equation (4.1), we have

$$F_N = \frac{1}{\varepsilon^2}u, \quad F_S = \Delta u - \frac{1}{\varepsilon^2}u^3.$$

We subdivide the time interval $[t_n, t_{n+1}]$ using the points $t_{n,m}$ for $m = 0, 1, \ldots, P$ such that

$$t_n = t_{n,0} < t_{n,1} < \ldots < t_{n,P} = t_{n+1}.$$

Let $\Delta t_{n,m} = t_{n,m+1} - t_{n,m}$ and $u_{n,m}^k$ denote the k-th order approximation to $u(t_{n,m})$. We choose the points $\{t_{n,m}\}_{m=0}^P$ as the Gauss-Lobatto nodes in

$[t_n, t_{n+1}]$. Starting from u_n, the second and third order time accurate SDC algorithms to calculate u_{n+1} are

• Second order accurate SDC scheme

$$\begin{aligned}
u^1_{n,0} =& u_n,\\
u^1_{n,1} =& u^1_{n,0} + \Delta t_{n,0}(F_S(t_{n,1}, u^1_{n,1}) + F_N(t_{n,0}, u^1_{n,0})),\\
u^2_{n,0} =& u_n,\\
u^2_{n,1} =& u^2_{n,0} + \Delta t_{n,0}(F_S(t_{n,1}, u^2_{n,1}) - F_S(t_{n,1}, u^1_{n,1}))\\
&+ I^1_0(F_S(t, u^1) + F_N(t, u^1)),
\end{aligned}$$

where $I^1_0(F_S(t, u^1) + F_N(t, u^1))$ is the integral of the linear interpolating polynomial using the two points $(t_{n,l}, F_S(t_{n,l}, u^1_{n,l}) + F_N(t_{n,l}, u^1_{n,l}))$, $(l = 0, 1)$ over the subinterval $[t_{n,0}, t_{n,1}]$.

Finally, we have $u_{n+1} = u^2_{n,1}$.

• Third order accurate SDC scheme

Compute initial approximation:

$$u^1_{n,0} = u_n.$$

For $m = 0, 1$

$$u^1_{n,m+1} = u^1_{n,m} + \Delta t_{n,m}(F_S(t_{n,m+1}, u^1_{n,m+1}) + F_N(t_{n,m}, u^1_{n,m})).$$

Compute successive corrections:

For $k = 1, 2$

$$u^{k+1}_{n,0} = u_n.$$

For $m = 0, 1$

$$\begin{aligned}
u^{k+1}_{n,m+1} =& u^{k+1}_{n,m} + \Delta t_{n,m}(F_S(t_{n,m+1}, u^{k+1}_{n,m+1}) - F_S(t_{n,m+1}, u^k_{n,m+1}))\\
&+ \Delta t_{n,m}(F_N(t_{n,m}, u^{k+1}_{n,m}) - F_N(t_{n,m}, u^k_{n,m}))\\
&+ I^{m+1}_m(F_S(t, u^k) + F_N(t, u^k)),
\end{aligned}$$

where $I^{m+1}_m(F_S(t, u^k) + F_N(t, u^k))$ is the integral of the quadratic interpolating polynomial using the three points $(t_{n,l}, F_S(t_{n,l}, u^k_{n,l}) + F_N(t_{n,l}, u^k_{n,m}))$ $(l = 0, 1, 2)$ over the subinterval $[t_{n,m}, t_{n,m+1}]$.

Finally, we have $u_{n+1} = u^3_{n,2}$.

4.3 LDG discretization combined with second order accurate SDC time integration method

In this section, we will present the second order time accurate SDC-LDG scheme for the Allen-Cahn equation (4.1)-(4.2) in $\Omega \in R^d$ with $d \leq 3$.

4.3.1 Fully-discrete SDC-LDG scheme

We use the second order semi-implicit SDC method introduced in Section 4.2.5 for the time discretization. Then the fully-discrete SDC-LDG discretization for (4.1) reads as: find $U_{n,1}, U_{n+1} \in V_h^k$, $\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n+1} \in \boldsymbol{W}_h^k$, such that for all $v \in V_h^k$ and $\boldsymbol{\phi} \in \boldsymbol{W}_h^k$, we have

$$\begin{aligned}
(U_{n,1} - U_n, v)_K =& -\Delta t[(\boldsymbol{Q}_{n,1}, \nabla v)_K - (\widehat{\boldsymbol{Q}}_{n,1} \cdot \boldsymbol{\nu}, v)_{\partial K}] \\
&- \frac{\Delta t}{\varepsilon^2}(U_{n,1}^3 - U_n, v)_K, \qquad (4.9)\\
(U_{n+1} - U_n, v)_K =& -\Delta t[(\boldsymbol{Q}_{n+1}, \nabla v)_K - (\widehat{\boldsymbol{Q}}_{n+1} \cdot \boldsymbol{\nu}, v)_{\partial K}] + \frac{\Delta t}{2}[(\boldsymbol{Q}_{n,1}, \nabla v)_K \\
&- (\widehat{\boldsymbol{Q}}_{n,1} \cdot \boldsymbol{\nu}, v)_{\partial K}] - \frac{\Delta t}{2}[(\boldsymbol{Q}_n, \nabla v)_K - (\widehat{\boldsymbol{Q}}_n \cdot \boldsymbol{\nu}, v)_{\partial K}] \\
&- \frac{\Delta t}{\varepsilon^2}(U_{n+1}^3 - U_n, v)_K + \frac{\Delta t}{\varepsilon^2}(U_{n,1}^3 - U_n, v)_K \\
&- \frac{\Delta t}{2\varepsilon^2}(U_{n,1}^3 - U_{n,1}, v)_K - \frac{\Delta t}{2\varepsilon^2}(U_n^3 - U_n, v)_K, \qquad (4.10)\\
(\boldsymbol{Q}_{n,1}, \boldsymbol{\phi})_K =& -(U_{n,1}, \nabla \cdot \boldsymbol{\phi})_K + (\widehat{U}_{n,1}, \boldsymbol{\nu} \cdot \boldsymbol{\phi})_{\partial K}, \qquad (4.11)\\
(\boldsymbol{Q}_{n+1}, \boldsymbol{\phi})_K =& -(U_{n+1}, \nabla \cdot \boldsymbol{\phi})_K + (\widehat{U}_{n+1}, \boldsymbol{\nu} \cdot \boldsymbol{\phi})_{\partial K}. \qquad (4.12)
\end{aligned}$$

Here $\boldsymbol{\nu}$ is the outward unit vector of element K at ∂K. The "hat" terms at ∂K in (4.9)-(4.12) are the so-called "numerical fluxes", which are functions that should be chosen to ensure stability. We remark that the selection of the numerical fluxes is not unique. Here we make the following simple choices:

$$\widehat{\boldsymbol{Q}}_{n,1} = \boldsymbol{Q}_{n,1}^R, \quad \widehat{\boldsymbol{Q}}_{n+1} = \boldsymbol{Q}_{n+1}^R, \quad \widehat{U}_{n,1} = U_{n,1}^L, \quad \widehat{U}_{n+1} = U_{n+1}^L. \qquad (4.13)$$

In view of the boundary condition (4.2), we take at $\partial\Omega$,

$$\widehat{\boldsymbol{Q}}_{n,1} \cdot \boldsymbol{\nu} = 0, \quad \widehat{\boldsymbol{Q}}_{n+1} \cdot \boldsymbol{\nu} = 0, \quad \widehat{U}_{n,1} = (U_{n,1})^{in}, \quad \widehat{U}_{n+1} = (U_{n+1})^{in}, \quad (4.14)$$

where $(U_{n,1})^{in}$ and $(U_{n+1})^{in}$ refer to values obtained from the interior of the boundary elements.

For convenience in the analysis, we set

$$\begin{aligned}\Gamma_K^+(\boldsymbol{\phi}, v) &:= -(\boldsymbol{\phi}, \nabla v)_K + (\boldsymbol{\phi}^R \cdot \boldsymbol{\nu}, v)_{\partial K},\\ \Gamma_K^-(v, \boldsymbol{\phi}) &:= -(v, \nabla \cdot \boldsymbol{\phi})_K + (v^L, \boldsymbol{\nu} \cdot \boldsymbol{\phi})_{\partial K}.\end{aligned} \tag{4.15}$$

Then equations (4.9)-(4.12) can be written as

$$(U_{n,1} - U_n, v)_K = \Delta t \Gamma_K^+(\boldsymbol{Q}_{n,1}, v) - \frac{\Delta t}{\varepsilon^2}(U_{n,1}^3 - U_n, v)_K, \tag{4.16}$$

$$\begin{aligned}(U_{n+1} - U_n, v)_K =& \Delta t \Gamma_K^+(\boldsymbol{Q}_{n+1}, v) - \frac{\Delta t}{2}\Gamma_K^+(\boldsymbol{Q}_{n,1}, v) + \frac{\Delta t}{2}\Gamma_K^+(\boldsymbol{Q}_n, v)\\ &- \frac{\Delta t}{\varepsilon^2}(U_{n+1}^3 - U_n, v)_K + \frac{\Delta t}{\varepsilon^2}(U_{n,1}^3 - U_n, v)_K\\ &- \frac{\Delta t}{2\varepsilon^2}(U_{n,1}^3 - U_{n,1}, v)_K - \frac{\Delta t}{2\varepsilon^2}(U_n^3 - U_n, v)_K,\end{aligned} \tag{4.17}$$

$$(\boldsymbol{Q}_{n,1}, \boldsymbol{\phi})_K = \Gamma_K^-(U_{n,1}, \boldsymbol{\phi}), \tag{4.18}$$

$$(\boldsymbol{Q}_{n+1}, \boldsymbol{\phi})_K = \Gamma_K^-(U_{n+1}, \boldsymbol{\phi}). \tag{4.19}$$

We define $F_{\partial K}$ as

$$F_{\partial K}(\boldsymbol{Q}, U) \doteq (\boldsymbol{Q}^R \cdot \boldsymbol{\nu}, U)_{\partial K} + (\boldsymbol{Q} \cdot \boldsymbol{\nu}, U^L)_{\partial K} - (\boldsymbol{Q} \cdot \boldsymbol{\nu}, U)_{\partial K}. \tag{4.20}$$

Using $\boldsymbol{\nu}^R = -\boldsymbol{\nu}^L$, the following property for $F_{\partial K}(\boldsymbol{Q}, U)$ is easy to show.

Lemma 4.3.1. *Assume e is an internal face shared by the elements K_L and K_R, then we have*

$$F_{\partial K_L \cap e}(\boldsymbol{Q}, U) + F_{\partial K_R \cap e}(\boldsymbol{Q}, U) = 0, \quad \forall \boldsymbol{Q} \in \boldsymbol{W}_h^k,\ U \in V_h^k.$$

4.3.2 Existence and uniqueness

In the following, we assume that U_n and $\boldsymbol{Q}_n$ are known and we will prove existence and uniqueness of the numerical solutions at time t^{n+1} for system (4.16)-(4.19).

Theorem 4.3.2. *The second order semi-implicit SDC-LDG scheme (4.16)-(4.19) is uniquely solvable if the time step satisfies the condition*

$$\Delta t < \frac{\varepsilon^2}{3C_L - 1},$$

with C_L the Lipschitz constant in (4.3) and ε the coefficient in the Allen-Cahn equation (4.1).

Proof. With U_n and $\boldsymbol{Q}_n$ known, we first prove that $(U_{n,1}, \boldsymbol{Q}_{n,1})$ is well-defined.

• **Existence of** $(U_{n,1}, \boldsymbol{Q}_{n,1})$

Let $\boldsymbol{X^n} \doteq (U_{n,1}, \Delta t^{\frac{1}{2}} \boldsymbol{Q}_{n,1})$ and $S_h^k \doteq V_h^k \times \Delta t^{\frac{1}{2}} \boldsymbol{W}_h^k$. After multiplying (4.18) by Δt, we sum equations of (4.16) and (4.18) over the elements $K \in \mathcal{T}_h$, and write this expression for $G_h : S_h^k \to S_h^k$ as

$$(G_h(\boldsymbol{X^n}), \boldsymbol{\chi}) = 0, \quad \forall \boldsymbol{\chi} \in S_h^k.$$

It is obvious that G_h is continuous.

Using Lemma 1.4 in Chapter 2 of [104], by Schauder's fixed point theorem, $G_h(\boldsymbol{\varpi}) = 0$ has a solution $\boldsymbol{\varpi} \in B_q = \{\boldsymbol{\chi} = \left(\chi_1, \Delta t^{\frac{1}{2}} \boldsymbol{\chi}_2\right) \in S_h^k : \|\boldsymbol{\chi}\|^2 = \|\chi_1\|^2 + \Delta t \|\boldsymbol{\chi}_2\|^2 \leq q^2\}$ if $(G_h(\boldsymbol{\chi}), \boldsymbol{\chi}) > 0$ for $\|\boldsymbol{\chi}\| = q$. For more detailed information, we refer to [45] and Chapter 13 in [105].

To prove $(G_h(\boldsymbol{\chi}), \boldsymbol{\chi}) > 0$, recalling the boundary conditions in (4.14) and Lemma 4.3.1, there holds

$$\begin{aligned}
&\sum_K \Gamma_K^-(\chi_1, \boldsymbol{\chi}_2) = \sum_K (-(\chi_1, \nu \cdot \boldsymbol{\chi}_2)_{\partial K} + (\boldsymbol{\chi}_2, \nabla \chi_1)_K + (\widehat{\chi}_1, \boldsymbol{\nu} \cdot \boldsymbol{\chi}_2)_{\partial K}) \\
=&\sum_K (-(\widehat{\boldsymbol{\chi}}_2 \cdot \boldsymbol{\nu}, \chi_1)_{\partial K} + (\boldsymbol{\chi}_2, \nabla \chi_1)_K) = -\sum_K \Gamma_K^+(\boldsymbol{\chi}_2, \chi_1). \qquad (4.21)
\end{aligned}$$

Then we have

$$\begin{aligned}
&(G_h(\boldsymbol{\chi}), \boldsymbol{\chi}) = (\chi_1 - U_n, \chi_1) + \Delta t \|\boldsymbol{\chi}_2\|^2 + \frac{\Delta t}{\varepsilon^2}(f(\chi_1) - f(0), \chi_1) \\
&+ \frac{\Delta t}{\varepsilon^2}(\chi_1 - U_n, \chi_1) \\
\geqslant &\left(\frac{1}{2} + \frac{\Delta t}{2\varepsilon^2}\right) \|\chi_1\|^2 - \left(\frac{1}{2} + \frac{\Delta t}{2\varepsilon^2}\right) \|U_n\|^2 + \Delta t \|\boldsymbol{\chi}_2\|^2 - \frac{C_L \Delta t}{\varepsilon^2} \|\chi_1\|^2 \\
\geqslant &\left(\frac{1}{2} + \frac{(1 - 2C_L)\Delta t}{2\varepsilon^2}\right) \|\chi_1\|^2 - \left(\frac{1}{2} + \frac{\Delta t}{2\varepsilon^2}\right) \|U_n\|^2 + \Delta t \|\boldsymbol{\chi}_2\|^2, \qquad (4.22)
\end{aligned}$$

which is positive if $\|\boldsymbol{\chi}\|$ is large enough, provided

$$\Delta t < \frac{\varepsilon^2}{2C_L - 1}. \qquad (4.23)$$

• **Uniqueness of** $(U_{n,1}, \boldsymbol{Q}_{n,1})$

Assuming that $U_{n,1}, \boldsymbol{Q}_{n,1}$ and $\tilde{U}_{n,1}, \tilde{\boldsymbol{Q}}_{n,1}$ both satisfy (4.16) and (4.18), we have

$$\begin{aligned}(U_{n,1} - \tilde{U}_{n,1}, v)_K =& \Delta t \Gamma_K^+(\boldsymbol{Q}_{n,1} - \tilde{\boldsymbol{Q}}_{n,1}, v) \\ &- \frac{\Delta t}{\varepsilon^2}(f(U_{n,1}) - f(\tilde{U}_{n,1}) + U_{n,1} - \tilde{U}_{n,1}, v)_K, \\ (\boldsymbol{Q}_{n,1} - \tilde{\boldsymbol{Q}}_{n,1}, \boldsymbol{\phi})_K =& \Gamma_K^-(U_{n,1} - \tilde{U}_{n,1}, \boldsymbol{\phi}).\end{aligned}$$

Let

$$v = U_{n,1} - \tilde{U}_{n,1}, \quad \boldsymbol{\phi} = \Delta t(\boldsymbol{Q}_{n,1} - \tilde{\boldsymbol{Q}}_{n,1}).$$

Using (4.21) and summation over all elements $K \in \mathcal{T}_h$ yields

$$\begin{aligned}&\|U_{n,1} - \tilde{U}_{n,1}\|^2 + \Delta t \|\boldsymbol{Q}_{n,1} - \tilde{\boldsymbol{Q}}_{n,1}\|^2 \\ &+ \frac{\Delta t}{\varepsilon^2}(f(U_{n,1}) - f(\tilde{U}_{n,1}) + U_{n,1} - \tilde{U}_{n,1}, U_{n,1} - \tilde{U}_{n,1}) = 0.\end{aligned}$$

Due to the Lipschitz condition (4.3) on f, we have

$$\left(1 + \frac{\Delta t}{\varepsilon^2} - \frac{C_L \Delta t}{\varepsilon^2}\right) \|U_{n,1} - \tilde{U}_{n,1}\|^2 + \Delta t \|\boldsymbol{Q}_{n,1} - \tilde{\boldsymbol{Q}}_{n,1}\|^2 \leqslant 0,$$

which implies uniqueness of $(U_{n,1}, \boldsymbol{Q}_{n,1})$ if the time step satisfies the condition

$$\Delta t < \frac{\varepsilon^2}{C_L - 1}. \tag{4.24}$$

Next, we will give a similar proof for the well posedness of $(U_{n+1}, \boldsymbol{Q}_{n+1})$.

• **Existence and uniqueness of** $(U_{n+1}, \boldsymbol{Q}_{n+1})$

For the existence, we need to prove a condition similar to (4.22). From the above analysis, we know that there exist unique numerical solutions of (4.16) and (4.18), then by taking $v = U_{n,1},\ \boldsymbol{\phi} = \boldsymbol{Q}_{n,1}$, we obtain

$$\begin{aligned}&\frac{\|U_{n,1}\|^2 - \|U_n\|^2 + \|U_{n,1} - U_n\|^2}{2\Delta t} + \|\boldsymbol{Q}_{n,1}\|^2 \\ &+ \frac{1}{\varepsilon^2}(f(U_{n,1}) - f(0) + U_{n,1} - U_n, U_{n,1}) = 0.\end{aligned}$$

By a simple use of the Cauchy and Young inequalities, there holds

$$\left(\frac{1}{2} + \frac{(1 - 2C_L)\Delta t}{2\varepsilon^2}\right) \|U_{n,1}\|^2 + \Delta t \|\boldsymbol{Q}_{n,1}\|^2 \leqslant \left(\frac{1}{2} + \frac{\Delta t}{2\varepsilon^2}\right) \|U_n\|^2. \tag{4.25}$$

Hence, for equations (4.17) and (4.19), similar to (4.22), we have

$$\begin{aligned}(G_h(\boldsymbol{\chi}),\boldsymbol{\chi}) =&(\chi_1 - U_n,\chi_1) + \Delta t\|\boldsymbol{\chi}_2\|^2 - \frac{\Delta t}{2}(\boldsymbol{Q}_{n,1},\boldsymbol{\chi}_2) + \frac{\Delta t}{2}(\boldsymbol{Q}_n,\boldsymbol{\chi}_2)\\ &+ \frac{\Delta t}{\varepsilon^2}(f(\chi_1),\chi_1) - \frac{\Delta t}{2\varepsilon^2}(f(U_{n,1}),\chi_1) + \frac{\Delta t}{2\varepsilon^2}(f(U_n),\chi_1)\\ &- \frac{\Delta t}{\varepsilon^2}(U_{n,1},\chi_1) + \frac{\Delta t}{\varepsilon^2}(\chi_1,\chi_1)\\ \geqslant&\left(\frac{1}{2} + \frac{\Delta t}{2\varepsilon^2} - \frac{3C_L\Delta t}{2\varepsilon^2}\right)\|\chi_1\|^2 - \left(\frac{\Delta t}{2\varepsilon^2} + \frac{C_L\Delta t}{4\varepsilon^2}\right)\|U_{n,1}\|^2\\ &- \left(\frac{1}{2} + \frac{C_L\Delta t}{4\varepsilon^2}\right)\|U_n\|^2 + \frac{\Delta t}{2}\|\boldsymbol{\chi}_2\|^2 - \frac{\Delta t}{4}\|\boldsymbol{Q}_{n,1}\|^2 - \frac{\Delta t}{4}\|\boldsymbol{Q}_n\|^2\\ \geqslant&\left(\frac{1}{2} + \frac{\Delta t}{2\varepsilon^2} - \frac{3C_L\Delta t}{2\varepsilon^2}\right)\|\chi_1\|^2 - \left(C_0 + \frac{C_L\Delta t}{4\varepsilon^2} + \frac{C_0\Delta t}{\varepsilon^2}\right)\|U_n\|^2\\ &+ \frac{\Delta t}{2}\|\boldsymbol{\chi}_2\|^2 - \frac{\Delta t}{4}\|\boldsymbol{Q}_n\|^2,\end{aligned}$$

where C_0 is a fixed positive constant generated by (4.25) and the last inequality is based on (4.23) and (4.25).
Let

$$\Delta t < \frac{\varepsilon^2}{3C_L - 1}, \tag{4.26}$$

then (4.23) and (4.24) are satisfied.
So if $\|\boldsymbol{\chi}\|$ is large enough, together with condition (4.26), we have

$$(G_h(\boldsymbol{\chi}),\boldsymbol{\chi}) > 0,$$

which completes the proof of the existence of the numerical solutions.

The proof of the uniqueness is similar to the proof for $(U_{n,1},\boldsymbol{Q}_{n,1})$ and we omit the details. □

4.3.3 Stability

Theorem 4.3.3. *If $\Delta t < \frac{3}{5}\varepsilon^2$, numerical solutions of the second order accurate semi-implicit SDC-LDG discretization (4.16)-(4.19) of the Allen-Cahn equation (4.1) satisfy the stability estimate*

$$\begin{aligned}&\frac{1}{10}\|U_{n+1}\|^2 + \frac{\Delta t}{4}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{7\Delta t}{8\varepsilon^2}\|U_{n+1}^2\|^2\\ \leqslant&\exp\left(\frac{40T}{\varepsilon^2}\right)\left(\|U_0\|^2 + \frac{\Delta t}{4}\|\boldsymbol{Q}_0\|^2 + \frac{7\Delta t}{8\varepsilon^2}\|U_0^2\|^2\right).\end{aligned}$$

Proof. We rewrite (4.16)-(4.19) as the following system

$$(U_{n,1}-U_n,v)_K = \Delta t\Gamma_K^+(\boldsymbol{Q}_{n,1},v) - \frac{\Delta t}{\varepsilon^2}(U_{n,1}^3-U_n,v)_K, \tag{4.27}$$

$$\begin{aligned}(U_{n+1}-U_{n,1},v)_K =& \Delta t\Gamma_K^+(\boldsymbol{Q}_{n+1},v) - \frac{3\Delta t}{2}\Gamma_K^+(\boldsymbol{Q}_{n,1},v) + \frac{\Delta t}{2}\Gamma_K^+(\boldsymbol{Q}_n,v)\\ &- \frac{\Delta t}{\varepsilon^2}(U_{n+1}^3-U_n,v)_K + \frac{2\Delta t}{\varepsilon^2}(U_{n,1}^3-U_n,v)_K\\ &- \frac{\Delta t}{2\varepsilon^2}(U_{n,1}^3-U_{n,1},v)_K - \frac{\Delta t}{2\varepsilon^2}(U_n^3-U_n,v)_K,\end{aligned} \tag{4.28}$$

$$(\boldsymbol{Q}_{n,l},\boldsymbol{\phi})_K = \Gamma_K^-(U_{n,l},\boldsymbol{\phi}), \quad l=1,2, \tag{4.29}$$

where $\boldsymbol{Q}_{n,2}=\boldsymbol{Q}_{n+1},\ U_{n,2}=U_{n+1}$.

Choose $v=-\frac{1}{4}U_{n+1}+\frac{3}{2}U_{n,1}-\frac{1}{4}U_n$ in (4.27) and $v=U_{n+1}$ in (4.28), respectively. After summation of the above equations over all elements $K\in\mathcal{T}_h$ and splitting $v=-\frac{1}{4}U_{n+1}+\frac{3}{2}U_{n,1}-\frac{1}{4}U_n$ into three parts: $v=-\frac{1}{4}(U_{n+1}-U_{n,1})$, $v=U_{n,1}$, $v=\frac{1}{4}(U_{n,1}-U_n)$, we have

$$\begin{aligned}&\frac{\|U_{n+1}\|^2-\|U_n\|^2+\|U_{n+1}-U_{n,1}\|^2+\|U_{n,1}-U_n\|^2}{2\Delta t}\\ &-\frac{(U_{n,1}-U_n,U_{n+1}-U_{n,1})}{4\Delta t}+\frac{\|U_{n,1}-U_n\|^2}{4\Delta t}=\sum_{i=1}^{2}\mathcal{E}_i,\end{aligned} \tag{4.30}$$

where

$$\begin{aligned}\mathcal{E}_1 =& \frac{1}{4\varepsilon^2}(U_{n,1}^3-U_n,U_{n+1}-U_{n,1}) - \frac{1}{\varepsilon^2}\|U_{n,1}^2\|^2 + \frac{1}{\varepsilon^2}(U_n,U_{n,1})\\ &- \frac{1}{4\varepsilon^2}(U_{n,1}^3-U_n,U_{n,1}-U_n) - \frac{1}{\varepsilon^2}\|U_{n+1}^2\|^2 + \frac{1}{\varepsilon^2}(U_n,U_{n+1})\\ &+ \frac{2}{\varepsilon^2}(U_{n,1}^3-U_n,U_{n+1}) - \frac{1}{2\varepsilon^2}(U_{n,1}^3-U_{n,1},U_{n+1})\\ &- \frac{1}{2\varepsilon^2}(U_n^3-U_n,U_{n+1}),\\ \mathcal{E}_2 =& -\frac{1}{4}\sum_K\Gamma_K^+(\boldsymbol{Q}_{n,1},U_{n+1}-U_{n,1}) + \sum_K\Gamma_K^+(\boldsymbol{Q}_{n,1},U_{n,1})\\ &+ \frac{1}{4}\sum_K\Gamma_K^+(\boldsymbol{Q}_{n,1},U_{n,1}-U_n) + \sum_K\Gamma_K^+(\boldsymbol{Q}_{n+1},U_{n+1})\\ &- \frac{3}{2}\sum_K\Gamma_K^+(\boldsymbol{Q}_{n,1},U_{n+1}) + \frac{1}{2}\sum_K\Gamma_K^+(\boldsymbol{Q}_n,U_{n+1}).\end{aligned}$$

• **Estimates for $\mathcal{E}_1$**

$$
\begin{aligned}
\mathcal{E}_1 =& -\frac{3}{2\varepsilon^2}\|U_{n,1}^2\|^2 - \frac{1}{\varepsilon^2}\|U_{n+1}^2\|^2 + \frac{7}{4\varepsilon^2}(U_{n,1}^3, U_{n+1}) + \frac{1}{4\varepsilon^2}(U_{n,1}^3, U_n)\\
&- \frac{1}{2\varepsilon^2}(U_n^3, U_{n+1}) - \frac{3}{4\varepsilon^2}(U_n, U_{n+1}) + \frac{3}{2\varepsilon^2}(U_n, U_{n,1})\\
&+ \frac{1}{2\varepsilon^2}(U_{n,1}, U_{n+1}) - \frac{1}{4\varepsilon^2}(U_n, U_n)\\
\leqslant& -\frac{3}{2\varepsilon^2}\|U_{n,1}^2\|^2 - \frac{1}{\varepsilon^2}\|U_{n+1}^2\|^2 + \frac{7}{4\varepsilon^2}\left(\frac{3}{4}\|U_{n,1}^2\|^2 + \frac{1}{4}\|U_{n+1}^2\|^2\right)\\
&+ \frac{1}{4\varepsilon^2}\left(\frac{3}{4}\|U_{n,1}^2\|^2 + \frac{1}{4}\|U_n^2\|^2\right) + \frac{1}{2\varepsilon^2}\left(\frac{3}{4}\|U_n^2\|^2 + \frac{1}{4}\|U_{n+1}^2\|^2\right)\\
&- \frac{3}{4\varepsilon^2}(U_n, U_{n+1}) + \frac{3}{2\varepsilon^2}(U_n, U_{n,1}) + \frac{1}{2\varepsilon^2}(U_{n,1}, U_{n+1}) - \frac{1}{4\varepsilon^2}(U_n, U_n)\\
=& -\frac{7}{16\varepsilon^2}(\|U_{n+1}^2\|^2 - \|U_n^2\|^2) + \frac{1}{2\varepsilon^2}(U_{n,1} - U_n, U_{n+1})\\
&+ \frac{1}{4\varepsilon^2}(U_{n,1} - U_n, U_n) - \frac{1}{4\varepsilon^2}(U_{n+1} - U_{n,1}, U_n) + \frac{1}{\varepsilon^2}(U_n, U_{n,1} - U_{n+1})\\
&+ \frac{1}{\varepsilon^2}(U_n, U_{n+1})\\
\leqslant& -\frac{7}{16\varepsilon^2}(\|U_{n+1}^2\|^2 - \|U_n^2\|^2) + \frac{3}{4\varepsilon^2}\|U_{n+1}\|^2 + \frac{5}{4\varepsilon^2}\|U_n\|^2\\
&+ \frac{5}{8\varepsilon^2}\|U_{n+1} - U_{n,1}\|^2 + \frac{3}{8\varepsilon^2}\|U_{n,1} - U_n\|^2,
\end{aligned}
\tag{4.31}
$$

where we have used the Cauchy and Young inequalities in the first estimate.

• **Estimates for $\mathcal{E}_2$**

Recalling the boundary conditions in (4.14), Lemma 4.3.1 and (4.29), with $0 \leqslant i, j \leqslant 2$, there holds

$$
\begin{aligned}
&\sum_K \Gamma_K^+(\boldsymbol{Q}_{n,i}, U_{n,j}) = \sum_K \left(-(\boldsymbol{Q}_{n,i}, \nabla U_{n,j})_K + (\widehat{\boldsymbol{Q}}_{n,i}\cdot\boldsymbol{\nu}, U_{n,j})_{\partial K}\right)\\
=&\sum_K \left(-(U_{n,j}, \nu\cdot\boldsymbol{Q}_{n,i})_{\partial K} + (U_{n,j}, \nabla\cdot\boldsymbol{Q}_{n,i})_K + (\widehat{\boldsymbol{Q}}_{n,i}\cdot\boldsymbol{\nu}, U_{n,j})_{\partial K}\right)\\
=&-\sum_K \Gamma_K^-(U_{n,j}, \boldsymbol{Q}_{n,i}) = -(\boldsymbol{Q}_{n,j}, \boldsymbol{Q}_{n,i}) = -(\boldsymbol{Q}_{n,i}, \boldsymbol{Q}_{n,j}),
\end{aligned}
\tag{4.32}
$$

where $\boldsymbol{Q}_{n,0} = \boldsymbol{Q}_n,\ U_{n,0} = U_n$. Using the Cauchy and Young inequalities, we obtain that

$$\begin{aligned}
\mathcal{E}_2 =& \frac{1}{4}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n+1} - \boldsymbol{Q}_{n,1}) - \|\boldsymbol{Q}_{n,1}\|^2 - \frac{1}{4}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n) - \|\boldsymbol{Q}_{n+1}\|^2 \\
&+ \frac{3}{2}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n+1}) - \frac{1}{2}(\boldsymbol{Q}_n, \boldsymbol{Q}_{n+1}) \\
=& -\frac{5}{4}\|\boldsymbol{Q}_{n,1}\|^2 - \|\boldsymbol{Q}_{n+1}\|^2 + \frac{1}{2}(\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n, \boldsymbol{Q}_{n+1}) + \frac{5}{4}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n+1}) \\
&- \frac{1}{4}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n) \\
\leqslant& -\frac{5}{4}\|\boldsymbol{Q}_{n,1}\|^2 - \frac{1}{2}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{1}{8}\|\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n\|^2 + \frac{5}{4}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n+1}) \\
&- \frac{1}{8}\left(\|\boldsymbol{Q}_{n,1}\|^2 - \|\boldsymbol{Q}_n\|^2 + \|\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n\|^2\right) \\
=& -\frac{1}{8}\left(\|\boldsymbol{Q}_{n+1}\|^2 - \|\boldsymbol{Q}_n\|^2\right) - S,
\end{aligned}$$

where

$$S = \frac{11}{8}\|\boldsymbol{Q}_{n,1}\|^2 + \frac{3}{8}\|\boldsymbol{Q}_{n+1}\|^2 - \frac{5}{4}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n+1}).$$

We set $\boldsymbol{X} = (\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n+1})$, and $S = \int_\Omega \boldsymbol{X} M \boldsymbol{X}^T dx$ with

$$M = \begin{pmatrix} 11/8 & -5/8 \\ -5/8 & 3/8 \end{pmatrix}. \tag{4.33}$$

It is easy to prove that M is positive definite, which shows that $S \geqslant 0$.

Inserting the estimates of $\mathcal{E}_1$, $\mathcal{E}_2$ into (4.30), we obtain

$$\begin{aligned}
&\left(1 - \frac{3\Delta t}{2\varepsilon^2}\right)\|U_{n+1}\|^2 - \left(1 + \frac{5\Delta t}{2\varepsilon^2}\right)\|U_n\|^2 + \left(\frac{3}{4} - \frac{5\Delta t}{4\varepsilon^2}\right)\|U_{n+1} - U_{n,1}\|^2 \\
&+ \left(\frac{5}{4} - \frac{3\Delta t}{4\varepsilon^2}\right)\|U_{n,1} - U_n\|^2 + \frac{\Delta t}{4}\left(\|\boldsymbol{Q}_{n+1}\|^2 - \|\boldsymbol{Q}_n\|^2\right) \\
&+ \frac{7\Delta t}{8\varepsilon^2}\left(\|U_{n+1}^2\|^2 - \|U_n^2\|^2\right) \leqslant 0.
\end{aligned}$$

Then

$$\begin{aligned}
&\left(1 - \frac{3\Delta t}{2\varepsilon^2}\right)\|U_{n+1}\|^2 + \left(\frac{3}{4} - \frac{5\Delta t}{4\varepsilon^2}\right)\|U_{n+1} - U_{n,1}\|^2 \\
&+ \left(\frac{5}{4} - \frac{3\Delta t}{4\varepsilon^2}\right)\|U_{n,1} - U_n\|^2 + \frac{\Delta t}{4}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{7\Delta t}{8\varepsilon^2}\|U_{n+1}^2\|^2 \\
\leqslant& \left(1 + \frac{5\Delta t}{2\varepsilon^2}\right)\|U_n\|^2 + \frac{\Delta t}{4}\|\boldsymbol{Q}_n\|^2 + \frac{7\Delta t}{8\varepsilon^2}\|U_n^2\|^2.
\end{aligned} \tag{4.34}$$

If $\Delta t < \dfrac{3\varepsilon^2}{5}$, the coefficients of the left hand side terms in (4.34) are all positive, hence we conclude that

$$\begin{aligned}&\|U_{n+1}\|^2 + \frac{\Delta t}{4\alpha}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{7\Delta t}{8\alpha\varepsilon^2}\|U_{n+1}^2\|^2\\ \leqslant &\left(1 + \frac{4\Delta t}{\alpha\varepsilon^2}\right)\|U_n\|^2 + \frac{\Delta t}{4\alpha}\|\boldsymbol{Q}_n\|^2 + \frac{7\Delta t}{8\alpha\varepsilon^2}\|U_n^2\|^2,\end{aligned}$$

where $\alpha = 1 - \dfrac{3\Delta t}{2\varepsilon^2}$. Note that $\Delta t < \dfrac{3\varepsilon^2}{5}$ is equivalent to $\alpha > \dfrac{1}{10}$, then

$$\begin{aligned}&\|U_{n+1}\|^2 + \frac{\Delta t}{4\alpha}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{7\Delta t}{8\alpha\varepsilon^2}\|U_{n+1}^2\|^2\\ \leqslant &\left(1 + \frac{40\Delta t}{\varepsilon^2}\right)\left(\|U_n\|^2 + \frac{\Delta t}{4\alpha}\|\boldsymbol{Q}_n\|^2 + \frac{7\Delta t}{8\alpha\varepsilon^2}\|U_n^2\|^2\right).\end{aligned}$$

Summing the above equation from 0 to $n < N$ yields

$$\begin{aligned}&\|U_{n+1}\|^2 + \frac{\Delta t}{4\alpha}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{7\Delta t}{8\alpha\varepsilon^2}\|U_{n+1}^2\|^2\\ \leqslant &\exp\left(\frac{40T}{\varepsilon^2}\right)\left(\|U_0\|^2 + \frac{\Delta t}{4\alpha}\|\boldsymbol{Q}_0\|^2 + \frac{7\Delta t}{8\alpha\varepsilon^2}\|U_0^2\|^2\right).\end{aligned}$$

That is

$$\begin{aligned}&\frac{1}{10}\|U_{n+1}\|^2 + \frac{\Delta t}{4}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{7\Delta t}{8\varepsilon^2}\|U_{n+1}^2\|^2\\ \leqslant &\exp\left(\frac{40T}{\varepsilon^2}\right)\left(\|U_0\|^2 + \frac{\Delta t}{4}\|\boldsymbol{Q}_0\|^2 + \frac{7\Delta t}{8\varepsilon^2}\|U_0^2\|^2\right),\end{aligned}$$

which completes the proof of Theorem 4.3.3. □

Remark 4.3.4. *To prove stability for the second order time accurate SDC-LDG discretization, the choice of the test functions is non-trivial, especially for the equations that contain nonlinear terms.*

The above proof mainly contains two parts. Firstly, without loss of generality, we choose the test function $v = U_{n,1},\ v = U_{n+1}$ *in (4.27), (4.28) respectively. Secondly, by analyzing the energy equation obtained in the first step, we take* $v = -\dfrac{1}{4}(U_{n+1} - U_{n,1}),\ v = \dfrac{1}{4}(U_{n,1} - U_n)$ *in (4.27) to eliminate several terms, which simplifies the stability analysis.*

Remark 4.3.5. *For the nonlinear terms containing* U*, we use the following inequality*

$$(u^3, v) \leqslant \frac{1}{2}\|u^2\|^2 + \frac{1}{2}\|uv\|^2 \leqslant \frac{3}{4}\|u^2\|^2 + \frac{1}{4}\|v^2\|^2. \tag{4.35}$$

4.3.4 Error estimates

Assume that the solution u of equation (4.1) is sufficiently smooth and satisfies

$$\begin{aligned} &u \in L^\infty((0,T);H^{k+2}(\Omega)),\ u_t \in L^\infty((0,T);H^{k+1}(\Omega)),\\ &u_{ttt} \in L^\infty((0,T);L^2(\Omega)). \end{aligned} \tag{4.36}$$

For simplicity, we use the following notations in the error analysis

$$\begin{aligned} e_u^n &:= u_n - U_n = u_n - Pu_n + Pu_n - U_n := u_n - Pu_n + Pe_{u_n},\\ e_{\boldsymbol{q}}^n &:= \boldsymbol{q}_n - \boldsymbol{Q}_n = \boldsymbol{q}_n - \Pi\boldsymbol{q}_n + \Pi\boldsymbol{q}_n - \boldsymbol{Q}_n := \boldsymbol{q}_n - \Pi\boldsymbol{q}_n + \Pi e_{\boldsymbol{q}_n}, \end{aligned}$$

with similar relations for other variables. Here we choose

$$\begin{aligned} &(P,\Pi) = (P^-,P^+) \ \text{ in one dimension},\\ &(P,\Pi) = (P^-,\Pi^+) \ \text{ in multi-dimensions}, \end{aligned}$$

which are defined in Section 4.2.4.

We rewrite the second order accurate SDC-LDG discretization for the Allen-Cahn equation (4.1) into a similar form as (4.27)-(4.29), which gives

$$(\partial_t u_{n,1}, v)_K = \Gamma_K^+(\boldsymbol{q}_{n,1}, v) - \frac{1}{\varepsilon^2}(u_{n,1}^3 - u_n, v)_K, \tag{4.37}$$

$$\begin{aligned} (\partial_t u_{n+1}, v)_K =& \Gamma_K^+(\boldsymbol{q}_{n+1}, v) - \frac{3}{2}\Gamma_K^+(\boldsymbol{q}_{n,1}, v) + \frac{1}{2}\Gamma_K^+(\boldsymbol{q}_n, v) + (\iota^n, v)_K\\ &- \frac{1}{\varepsilon^2}(u_{n+1}^3 - u_n, v)_K + \frac{2}{\varepsilon^2}(u_{n,1}^3 - u_n, v)_K\\ &- \frac{1}{2\varepsilon^2}(u_{n,1}^3 - u_{n,1}, v)_K - \frac{1}{2\varepsilon^2}(u_n^3 - u_n, v)_K, \end{aligned} \tag{4.38}$$

$$(\boldsymbol{q}_{n,l}, \boldsymbol{\phi})_K = \Gamma_K^-(u_{n,l}, \boldsymbol{\phi}), \quad l = 1, 2, \tag{4.39}$$

where

$$\partial_t u_{n,1} = \frac{u_{n,1} - u_n}{\Delta t},\ \partial_t u_{n+1} = \frac{u_{n+1} - u_{n,1}}{\Delta t}$$

and $\|\iota^n\| \le C\Delta t^2$ is the local truncation error for the second order accurate SDC time discretization [120].

Subtracting (4.27)-(4.29) from (4.37)-(4.39), we obtain the error equations

$$\begin{aligned}(\partial_t(u_{n,1}-U_{n,1}),v)_K =&\Gamma_K^+(\boldsymbol{q}_{n,1}-\boldsymbol{Q}_{n,1},v)\\ &-\frac{1}{\varepsilon^2}(u_{n,1}^3-u_n-(U_{n,1}^3-U_n),v)_K, \qquad (4.40)\\ (\partial_t(u_{n+1}-U_{n+1}),v)_K =&\Gamma_K^+(\boldsymbol{q}_{n+1}-\boldsymbol{Q}_{n+1},v)-\frac{3}{2}\Gamma_K^+(\boldsymbol{q}_{n,1}-\boldsymbol{Q}_{n,1},v)\\ &+\frac{1}{2}\Gamma_K^+(\boldsymbol{q}_n-\boldsymbol{Q}_n,v)-\frac{1}{\varepsilon^2}(u_{n+1}^3-u_n-U_{n+1}^3\\ &+U_n,v)_K+\frac{2}{\varepsilon^2}(u_{n,1}^3-u_n-(U_{n,1}^3-U_n),v)_K\\ &-\frac{1}{2\varepsilon^2}(u_{n,1}^3-u_{n,1}-(U_{n,1}^3-U_{n,1}),v)_K+(\iota^n,v)_K\\ &-\frac{1}{2\varepsilon^2}(u_n^3-u_n-(U_n^3-U_n),v)_K, \qquad (4.41)\\ (\boldsymbol{q}_{n,l}-\boldsymbol{Q}_{n,l},\boldsymbol{\phi})_K =&\Gamma_K^-(u_{n,l}-U_{n,l},\boldsymbol{\phi}),\quad l=1,2. \qquad (4.42)\end{aligned}$$

By choosing

$$U_0=Pu_0,\quad (\boldsymbol{Q}_0,\boldsymbol{\phi})_K=\Gamma_K^-(U_0,\boldsymbol{\phi}), \qquad (4.43)$$

it is easy to show that $\boldsymbol{Q}_0$ is well-defined. In addition, from the interpolation properties of P and Π, we have that

$$\|\Pi e_{\boldsymbol{q}_0}\|_\Omega \le Ch^{k+1}. \qquad (4.44)$$

Next, we present error estimates for the system (4.40)-(4.42). The proof of the error estimates follows the same line as for the stability analysis.

Theorem 4.3.6. *Let u be the exact solution of the Allen-Cahn equation (4.1)-(4.2), which satisfies the smoothness assumptions (4.36), and U_n be the numerical solution of the second order accurate semi-implicit SDC-LDG scheme (4.16)-(4.19). Then for $n=1,2,\ldots,N$, there exist positive constants h and Δt_0, where Δt_0 depends on ε, but is independent of h, such that for $\Delta t\le\Delta t_0$ the error in the second order accurate SDC-LDG discretization of the Allen-Cahn equation is bounded as*

$$\max_{n\Delta t\le T}\|e_u^n\|\le Ch^{k+1}+C\Delta t^2, \qquad (4.45)$$

where C depends on $\|u\|_{L^\infty((0,T);H^{k+2}(\Omega))}$, $\|u_t\|_{L^\infty((0,T);H^{k+1}(\Omega))}$, $\|u_{ttt}\|_{L^\infty((0,T);L^2(\Omega))}$, ε and T.

Proof. Based on relation (4.32), with (4.42), we obtain

$$\sum_K \Gamma_K^+(\Pi e_{\boldsymbol{q}_{n,i}}, Pe_{u_{n,j}}) = -\sum_K \Gamma_K^-(Pe_{u_{n,j}}, \Pi e_{\boldsymbol{q}_{n,i}}) = -(\boldsymbol{q}_{n,j} - \boldsymbol{Q}_{n,j}, \Pi e_{\boldsymbol{q}_{n,i}}) + \sum_K \Gamma_K^-(u_{n,j} - Pu_{n,j}, \Pi e_{\boldsymbol{q}_{n,i}}), \quad 0 \leqslant i, j \leqslant 2. \tag{4.46}$$

Let $v = -\frac{1}{4}Pe_{u_{n+1}} + \frac{3}{2}Pe_{u_{n,1}} - \frac{1}{4}Pe_{u_n}$ in (4.40) and $v = Pe_{u_{n+1}}$ in (4.41), respectively, dividing $v = -\frac{1}{4}Pe_{u_{n+1}} + \frac{3}{2}Pe_{u_{n,1}} - \frac{1}{4}Pe_{u_n}$ into three parts: $v = -\frac{1}{4}(Pe_{u_{n+1}} - Pe_{u_{n,1}})$, $v = Pe_{u_{n,1}}$, $v = \frac{1}{4}(Pe_{u_{n,1}} - Pe_{u_n})$, together with the definition of Π and (4.46), we have

$$\begin{aligned} &\frac{\|Pe_{u_{n+1}}\|^2 - \|Pe_{u_n}\|^2 + \|Pe_{u_{n+1}} - Pe_{u_{n,1}}\|^2 + \|Pe_{u_{n,1}} - Pe_{u_n}\|^2}{2\Delta t} \\ &- \frac{(Pe_{u_{n,1}} - Pe_{u_n}, Pe_{u_{n+1}} - Pe_{u_{n,1}})}{4\Delta t} + \frac{\|Pe_{u_{n,1}} - Pe_{u_n}\|^2}{4\Delta t} \\ =&RHS := \sum_{i=1}^{6} \mathcal{F}_i + (\iota^n, Pe_{u_{n+1}}), \end{aligned} \tag{4.47}$$

where

$$\begin{aligned} \mathcal{F}_1 =& -\|\Pi e_{\boldsymbol{q}_{n,1}}\|^2 + \frac{1}{4}(\Pi e_{\boldsymbol{q}_{n,1}}, \Pi e_{\boldsymbol{q}_{n+1}} - \Pi e_{\boldsymbol{q}_{n,1}}) - \frac{1}{8}(\|\Pi e_{\boldsymbol{q}_{n,1}}\|^2 - \|\Pi e_{\boldsymbol{q}_n}\|^2 \\ &+ \|\Pi e_{\boldsymbol{q}_{n,1}} - \Pi e_{\boldsymbol{q}_n}\|^2) - \|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 + \frac{3}{2}(\Pi e_{\boldsymbol{q}_{n,1}}, \Pi e_{\boldsymbol{q}_{n+1}}) \\ &- \frac{1}{2}(\Pi e_{\boldsymbol{q}_n}, \Pi e_{\boldsymbol{q}_{n+1}}), \\ \mathcal{F}_2 =& -(\boldsymbol{q}_{n,1} - \Pi \boldsymbol{q}_{n,1}, \Pi e_{\boldsymbol{q}_{n,1}}) + \frac{1}{4}(\boldsymbol{q}_{n+1} - \Pi \boldsymbol{q}_{n+1} - (\boldsymbol{q}_{n,1} - \Pi \boldsymbol{q}_{n,1}), \Pi e_{\boldsymbol{q}_{n,1}}) \\ &- \frac{1}{4}(\boldsymbol{q}_{n,1} - \Pi \boldsymbol{q}_{n,1} - (\boldsymbol{q}_n - \Pi \boldsymbol{q}_n), \Pi e_{\boldsymbol{q}_{n,1}}) - (\boldsymbol{q}_{n+1} - \Pi \boldsymbol{q}_{n+1}, \Pi e_{\boldsymbol{q}_{n+1}}) \\ &+ \frac{3}{2}(\boldsymbol{q}_{n+1} - \Pi \boldsymbol{q}_{n+1}, \Pi e_{\boldsymbol{q}_{n,1}}) - \frac{1}{2}(\boldsymbol{q}_{n+1} - \Pi \boldsymbol{q}_{n+1}, \Pi e_{\boldsymbol{q}_n}), \end{aligned}$$

$$
\begin{aligned}
\mathcal{F}_3 =& -\frac{1}{\varepsilon^2}(f(u_{n,1}) - f(U_{n,1}), Pe_{u_{n,1}}) \\
&+ \frac{1}{4\varepsilon^2}(f(u_{n,1}) - f(U_{n,1}), Pe_{u_{n+1}} - Pe_{u_{n,1}}) \\
&- \frac{1}{4\varepsilon^2}(f(u_{n,1}) - f(U_{n,1}), Pe_{u_{n,1}} - Pe_{u_n}) \\
&- \frac{1}{\varepsilon^2}(f(u_{n+1}) - f(U_{n+1}), Pe_{u_{n+1}}) \\
&+ \frac{3}{2\varepsilon^2}(f(u_{n,1}) - f(U_{n,1}), Pe_{u_{n+1}}) - \frac{1}{2\varepsilon^2}(f(u_n) - f(U_n), Pe_{u_{n+1}}), \\
\mathcal{F}_4 =& -\frac{\Delta t}{\varepsilon^2}(\partial_t Pe_{u_{n,1}}, Pe_{u_{n,1}}) + \frac{\Delta t}{4\varepsilon^2}(\partial_t Pe_{u_{n,1}}, Pe_{u_{n+1}} - Pe_{u_{n,1}}) \\
&- \frac{\Delta t}{4\varepsilon^2}(\partial_t Pe_{u_{n,1}}, Pe_{u_{n,1}} - Pe_{u_n}) - \frac{\Delta t}{\varepsilon^2}(\partial_t Pe_{u_{n+1}}, Pe_{u_{n+1}}) \\
&+ \frac{2\Delta t}{\varepsilon^2}(\partial_t Pe_{u_{n,1}}, Pe_{u_{n+1}}), \\
\mathcal{F}_5 =& -(\partial_t(u_{n,1} - Pu_{n,1}), Pe_{u_{n,1}}) - \frac{\Delta t}{\varepsilon^2}(\partial_t(u_{n,1} - Pu_{n,1}), Pe_{u_{n,1}}) \\
&+ \frac{1}{4}(\partial_t(u_{n,1} - Pu_{n,1}), Pe_{u_{n+1}} - Pe_{u_{n,1}}) \\
&+ \frac{\Delta t}{4\varepsilon^2}(\partial_t(u_{n,1} - Pu_{n,1}), Pe_{u_{n+1}} - Pe_{u_{n,1}}) \\
&- \frac{1}{4}(\partial_t(u_{n,1} - Pu_{n,1}), Pe_{u_{n,1}} - Pe_{u_n}) \\
&- \frac{\Delta t}{4\varepsilon^2}(\partial_t(u_{n,1} - Pu_{n,1}), Pe_{u_{n,1}} - Pe_{u_n}) \\
&- (\partial_t(u_{n+1} - Pu_{n+1}), Pe_{u_{n+1}}) - \frac{\Delta t}{\varepsilon^2}(\partial_t(u_{n+1} - Pu_{n+1}), Pe_{u_{n+1}}) \\
&+ \frac{2\Delta t}{\varepsilon^2}(\partial_t(u_{n,1} - Pu_{n,1}), Pe_{u_{n+1}}), \\
\mathcal{F}_6 =& \sum_K (\Gamma_K^-(u_{n,1} - Pu_{n,1}, \Pi e_{\boldsymbol{q}_{n,1}}) \\
&- \frac{1}{4}\Gamma_K^-(u_{n+1} - Pu_{n+1} - (u_{n,1} - Pu_{n,1}), \Pi e_{\boldsymbol{q}_{n,1}})) \\
&+ \frac{1}{4}\sum_K (\Gamma_K^-(u_{n,1} - Pu_{n,1} - (u_n - Pu_n), \Pi e_{\boldsymbol{q}_{n,1}}) \\
&+ \Gamma_K^-(u_{n+1} - Pu_{n+1}, \Pi e_{\boldsymbol{q}_{n+1}})) \\
&- \frac{3}{2}\sum_K (\Gamma_K^-(u_{n+1} - Pu_{n+1}, \Pi e_{\boldsymbol{q}_{n,1}}) + \frac{1}{2}\Gamma_K^-(u_{n+1} - Pu_{n+1}, \Pi e_{\boldsymbol{q}_n})).
\end{aligned}
$$

Next, we estimate each term $\mathcal{F}_i,\ 1 \leqslant i \leqslant 6$. It is easy to see that

$$\begin{aligned}\mathcal{F}_1 =& -\frac{11}{8}\|\Pi e_{\boldsymbol{q}_{n,1}}\|^2 + \frac{1}{4}(\Pi e_{\boldsymbol{q}_{n,1}}, \Pi e_{\boldsymbol{q}_{n+1}}) + \frac{1}{8}\|\Pi e_{\boldsymbol{q}_n}\|^2 - \frac{1}{8}\|\Pi e_{\boldsymbol{q}_{n,1}} - \Pi e_{\boldsymbol{q}_n}\|^2 \\ & - \|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 + (\Pi e_{\boldsymbol{q}_{n,1}}, \Pi e_{\boldsymbol{q}_{n+1}}) + \frac{1}{2}(\Pi e_{\boldsymbol{q}_{n,1}} - \Pi e_{\boldsymbol{q}_n}, \Pi e_{\boldsymbol{q}_{n+1}}) \\ \leqslant& -\frac{11}{8}\|\Pi e_{\boldsymbol{q}_{n,1}}\|^2 + \frac{5}{4}(\Pi e_{\boldsymbol{q}_{n,1}}, \Pi e_{\boldsymbol{q}_{n+1}}) + \frac{1}{8}\|\Pi e_{\boldsymbol{q}_n}\|^2 - \frac{1}{2}\|\Pi e_{\boldsymbol{q}_{n+1}}\|^2.\end{aligned}$$

Based on the interpolation properties of the projections P and Π, as well as the Lipschitz continuity property (4.3) of f, there holds

$$\begin{aligned}|\sum_{i=2}^{5}\mathcal{F}_i| \leqslant& Ch^{2k+2} + C(\|Pe_{u_{n+1}}\|^2 + \|Pe_{u_{n,1}}\|^2 + \|Pe_{u_n}\|^2) \\ & + \frac{\epsilon_1}{2}(\|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 + \|\Pi e_{\boldsymbol{q}_{n,1}}\|^2 + \|\Pi e_{\boldsymbol{q}_n}\|^2),\end{aligned}$$

where C depends on ε and $\epsilon_1 > 0$ is a small enough constant generated by Young's inequality.

In one-dimension, $\mathcal{F}_6 = 0$. In multi-dimensions, using Lemma 4.2.1 we have

$$|\mathcal{F}_6| \leqslant Ch^{2k+2} + \frac{\epsilon_1}{2}(\|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 + \|\Pi e_{\boldsymbol{q}_{n,1}}\|^2 + \|\Pi e_{\boldsymbol{q}_n}\|^2).$$

Inserting the estimates of $\mathcal{F}_i,\ i = 1, \ldots, 6$ into (4.47), together with the error estimate for ι^n, we have

$$\begin{aligned}RHS \leqslant& Ch^{2k+2} + C\Delta t^4 + C(\|Pe_{u_{n+1}}\|^2 + \|Pe_{u_{n,1}}\|^2 + \|Pe_{u_n}\|^2) \\ & + \epsilon_1(\|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 + \|\Pi e_{\boldsymbol{q}_{n,1}}\|^2 + \|\Pi e_{\boldsymbol{q}_n}\|^2) - \frac{11}{8}\|\Pi e_{\boldsymbol{q}_{n,1}}\|^2 \\ & + \frac{5}{4}(\Pi e_{\boldsymbol{q}_{n,1}}, \Pi e_{\boldsymbol{q}_{n+1}}) + \frac{1}{8}\|\Pi e_{\boldsymbol{q}_n}\|^2 - \frac{1}{2}\|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 \\ \leqslant& Ch^{2k+2} + C\Delta t^4 + C\|Pe_{u_{n+1}}\|^2 + C\|Pe_{u_{n,1}} - Pe_{u_n}\|^2 \\ & + C\|Pe_{u_n}\|^2 - \left(\frac{1}{8} + \epsilon_1\right)(\|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 - \|\Pi e_{\boldsymbol{q}_n}\|^2) - SS, \quad (4.48)\end{aligned}$$

where

$$SS = \left(\frac{3}{8} - 2\epsilon_1\right)\|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 - \frac{5}{4}(\Pi e_{\boldsymbol{q}_{n,1}}, \Pi e_{\boldsymbol{q}_{n+1}}) + \left(\frac{11}{8} - \epsilon_1\right)\|\Pi e_{\boldsymbol{q}_{n,1}}\|^2.$$

Similar to (4.33), we have $SS \geqslant 0$ provided by $\left(\frac{3}{8} - 2\epsilon_1\right)\left(\frac{11}{8} - \epsilon_1\right) \geqslant \frac{25}{64}$, i.e. $0 < \epsilon_1 \leqslant \frac{25 - \sqrt{561}}{32}$. Combining (4.47) with (4.48), gives

$$(1 - C\Delta t)\|Pe_{u_{n+1}}\|^2 - (1 + C\Delta t)\|Pe_{u_n}\|^2 + \Delta t\left(\frac{1}{4} + 2\epsilon_1\right)(\|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 - \|\Pi e_{\boldsymbol{q}_n}\|^2) + (\frac{5}{4} - C\Delta t)\|Pe_{u_{n,1}} - Pe_{u_n}\|^2 \leqslant C\Delta t(h^{2k+2} + \Delta t^4).$$

Let $\Delta t \leqslant \Delta t_0 < \frac{1}{C}$, where Δt_0 depends on ε, but is independent of h. Summing the above equation from 0 to $n < N$ yields

$$(1 - C\Delta t)\|Pe_{u_{n+1}}\|^2 + \Delta t\left(\frac{1}{4} + 2\epsilon_1\right)\|\Pi e_{\boldsymbol{q}_{n+1}}\|^2 \leqslant C\Delta t\sum_{i=0}^{n}(h^{2k+2} + \Delta t^4) + C\Delta t\sum_{i=0}^{n}\|Pe_{u_i}\|^2 + \|Pe_{u_0}\|^2 + \Delta t\left(\frac{1}{4} + 2\epsilon_1\right)\|\Pi e_{\boldsymbol{q}_0}\|^2.$$

Using (4.43), (4.44) and Gronwall's inequality, we have

$$\|Pe_{u_{n+1}}\|^2 \leqslant C(h^{2k+2} + \Delta t^4),$$

which completes the proof. □

4.4 LDG discretization combined with third order accurate SDC time integration method

In this section, we will discuss stability and error estimates of the third order time accurate SDC-LDG scheme for the Allen-Cahn equation (4.1)-(4.2) in $\Omega \subset \mathbb{R}^d$ with $d \leqslant 3$.

4.4.1 Fully-discrete numerical scheme

We use the following numerical fluxes in the LDG discretization

$$\widehat{\boldsymbol{Q}}_{n,l} = \boldsymbol{Q}^R_{n,l}, \quad \widehat{U}_{n,l} = U^L_{n,l}, \quad l = 1, 2, \ldots, 6. \tag{4.49}$$

In view of the boundary condition (4.2), we take

$$\widehat{\boldsymbol{Q}}_{n,l} \cdot \boldsymbol{\nu} = 0, \quad \widehat{U}_{n,l} = (U_{n,l})^{in}, \quad l = 1, 2, \ldots, 6 \tag{4.50}$$

at $\partial\Omega$, where $(U_{n,l})^{in},\ l = 1, \ldots, 6$, refer to values obtained from the interior of the boundary elements. The numerical solutions $U_{n,l},\ l = 1, \ldots, 6$ correspond to $u^k_{n,m}$ in the third order SDC discretization, with $U_{n,1}$ corresponding to $u^1_{n,1}$, $U_{n,2}$ to $u^1_{n,2}$, $U_{n,3}$ to $u^2_{n,1}$, $U_{n,4}$ to $u^2_{n,2}$, $U_{n,5}$ to $u^3_{n,1}$ and $U_{n,6}$ to $u^3_{n,2}$.

We use the third order semi-implicit SDC method introduced in Section 4.2.5 for the time discretization. Using (4.15), the fully-discrete third order accurate SDC-LDG approximation scheme to the Allen-Cahn equation (4.1) is given as: find $U_{n,l} \in V^k_h$, $\boldsymbol{Q}_{n,l} \in \boldsymbol{W}^k_h,\ l = 1, 2, \ldots, 6$, such that, for all $v \in V^k_h$ and $\boldsymbol{\phi} \in \boldsymbol{W}^k_h$, we have

$$(U_{n,1} - U_n, v)_K = \frac{\Delta t}{2}\Gamma^+_K(\boldsymbol{Q}_{n,1}, v) - \frac{\Delta t}{2\varepsilon^2}(U^3_{n,1} - U_n, v)_K, \tag{4.51}$$

$$(U_{n,2} - U_{n,1}, v)_K = \frac{\Delta t}{2}\Gamma^+_K(\boldsymbol{Q}_{n,2}, v) - \frac{\Delta t}{2\varepsilon^2}(U^3_{n,2} - U_{n,1}, v)_K, \tag{4.52}$$

$$\begin{aligned}(U_{n,3} - U_{n,2}, v)_K =& \frac{\Delta t}{2}\Gamma^+_K(\boldsymbol{Q}_{n,3}, v) - \frac{13\Delta t}{24}\Gamma^+_K(\boldsymbol{Q}_{n,2}, v) - \frac{2\Delta t}{3}\Gamma^+_K(\boldsymbol{Q}_{n,1}, v)\\ &+ \frac{5\Delta t}{24}\Gamma^+_K(\boldsymbol{Q}_n, v) - \frac{\Delta t}{2\varepsilon^2}(U^3_{n,3}, v)_K + \frac{13\Delta t}{24\varepsilon^2}(U^3_{n,2}, v)_K\\ &+ \frac{2\Delta t}{3\varepsilon^2}(U^3_{n,1}, v)_K - \frac{5\Delta t}{24\varepsilon^2}(U^3_n, v)_K\\ &- \frac{\Delta t}{\varepsilon^2}(\frac{1}{24}U_{n,2} + \frac{1}{6}U_{n,1} + \frac{7}{24}U_n, v)_K,\end{aligned} \tag{4.53}$$

$$\begin{aligned}(U_{n,4} - U_{n,3}, v)_K =& \frac{\Delta t}{2}\Gamma^+_K(\boldsymbol{Q}_{n,4}, v) - \frac{7\Delta t}{24}\Gamma^+_K(\boldsymbol{Q}_{n,2}, v) + \frac{\Delta t}{3}\Gamma^+_K(\boldsymbol{Q}_{n,1}, v)\\ &- \frac{\Delta t}{24}\Gamma^+_K(\boldsymbol{Q}_n, v) - \frac{\Delta t}{2\varepsilon^2}(U^3_{n,4}, v)_K + \frac{7\Delta t}{24\varepsilon^2}(U^3_{n,2}, v)_K\\ &- \frac{\Delta t}{3\varepsilon^2}(U^3_{n,1}, v)_K + \frac{\Delta t}{24\varepsilon^2}(U^3_n, v)_K\\ &+ \frac{\Delta t}{\varepsilon^2}(\frac{1}{2}U_{n,3} + \frac{5}{24}U_{n,2} - \frac{1}{6}U_{n,1} - \frac{1}{24}U_n, v)_K,\end{aligned} \tag{4.54}$$

$$\begin{aligned}(U_{n,5} - U_{n,4}, v)_K =& \frac{\Delta t}{2}\Gamma^+_K(\boldsymbol{Q}_{n,5}, v) - \frac{13\Delta t}{24}\Gamma^+_K(\boldsymbol{Q}_{n,4}, v) - \frac{2\Delta t}{3}\Gamma^+_K(\boldsymbol{Q}_{n,3}, v)\\ &+ \frac{\Delta t}{3}\Gamma^+_K(\boldsymbol{Q}_{n,2}, v) - \frac{\Delta t}{6}\Gamma^+_K(\boldsymbol{Q}_{n,1}, v) + \frac{\Delta t}{24}\Gamma^+_K(\boldsymbol{Q}_n, v)\\ &- \frac{\Delta t}{2\varepsilon^2}(U^3_{n,5}, v)_K + \frac{13\Delta t}{24\varepsilon^2}(U^3_{n,4}, v)_K + \frac{2\Delta t}{3\varepsilon^2}(U^3_{n,3}, v)_K\\ &- \frac{\Delta t}{3\varepsilon^2}(U^3_{n,2}, v)_K + \frac{\Delta t}{6\varepsilon^2}(U^3_{n,1}, v)_K - \frac{\Delta t}{24\varepsilon^2}(U^3_n, v)_K\end{aligned}$$

$$- \frac{\Delta t}{\varepsilon^2}(\frac{1}{24}U_{n,4} + \frac{1}{6}U_{n,3} + \frac{1}{6}U_{n,2} + \frac{1}{6}U_{n,1} - \frac{1}{24}U_n, v)_K, \tag{4.55}$$

$$\begin{aligned}(U_{n,6} - U_{n,5}, v)_K =& \frac{\Delta t}{2}\Gamma_K^+(\boldsymbol{Q}_{n,6}, v) - \frac{7\Delta t}{24}\Gamma_K^+(\boldsymbol{Q}_{n,4}, v) + \frac{\Delta t}{3}\Gamma_K^+(\boldsymbol{Q}_{n,3}, v) \\ & - \frac{\Delta t}{24}\Gamma_K^+(\boldsymbol{Q}_n, v) - \frac{\Delta t}{2\varepsilon^2}(U_{n,6}^3, v)_K + \frac{7\Delta t}{24\varepsilon^2}(U_{n,4}^3, v)_K \\ & - \frac{\Delta t}{3\varepsilon^2}(U_{n,3}^3, v)_K + \frac{\Delta t}{24\varepsilon^2}(U_n^3, v)_K \\ & + \frac{\Delta t}{\varepsilon^2}(\frac{1}{2}U_{n,5} + \frac{5}{24}U_{n,4} - \frac{1}{6}U_{n,3} - \frac{1}{24}U_n, v)_K, \end{aligned} \tag{4.56}$$

$$(\boldsymbol{Q}_{n,l}, \boldsymbol{\phi})_K = \Gamma_K^-(U_{n,l}, \boldsymbol{\phi}), \quad l = 1, 2, \ldots, 6, \tag{4.57}$$

where $U_{n+1} = U_{n,6}$, $\boldsymbol{Q}_{n+1} = \boldsymbol{Q}_{n,6}$.

4.4.2 Existence and Uniqueness

The proof of the well-posedness for the LDG method combined with the third order SDC time integration method is similar to the proof in Section 4.3.2 for the second order discretization. We only give the main result and skip the proof details.

Theorem 4.4.1. *There exists a positive constant C_1 independent of ε and h, such that if $\Delta t < \dfrac{\varepsilon^2}{C_1}$, the third order accurate semi-implicit SDC-LDG discretization (4.51)-(4.57) for the Allen-Cahn equation (4.1) is well-defined.*

4.4.3 Stability

Theorem 4.4.2. *If $\Delta t < \dfrac{\varepsilon^2}{C_3}$, numerical solutions of the third order accurate semi-implicit SDC-LDG discretization (4.51)-(4.57) for the Allen-Cahn equation (4.1) satisfy the stability estimate*

$$\begin{aligned} & C_4\|U_{n+1}\|^2 + \frac{\Delta t}{2}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{\Delta t}{2\varepsilon^2}\|U_{n+1}^2\|^2 \\ \leqslant & \exp\left(\frac{C_5 T}{\varepsilon^2}\right)(\|U_0\|^2 + \frac{\Delta t}{2}\|\boldsymbol{Q}_0\|^2 + \frac{\Delta t}{2\varepsilon^2}\|U_0^2\|^2), \end{aligned} \tag{4.58}$$

where C_3, C_4, C_5 are positive constants and independent of ε and h.

Proof. The proof of Theorem 4.4.2 is given in Appendix 4.A. □

4.4.4 Error estimates

For the error estimates, we assume that the exact solution u has the following smoothness:

$$u \in L^\infty((0,T);H^{k+2}(\Omega)),\ u_t \in L^\infty((0,T);H^{k+1}(\Omega)),$$
$$u_{tttt} \in L^\infty((0,T);L^2(\Omega)). \tag{4.59}$$

In Section 4.3.4, we obtained error estimates for the second order accurate SDC scheme. The same ideas can be used to obtain error estimates for the third order accurate SDC-LDG discretization. We omit the proof details and only give the error estimate.

Theorem 4.4.3. *Let u be the exact solution of the Allen-Cahn equation (4.1)-(4.2) satisfying the smoothness condition (4.59), and U_n be the numerical solution of the third order accurate semi-implicit SDC-LDG scheme (4.51)-(4.57). Then for $n = 1,\dots,N$, there exist positive constants h and Δt_0, where Δt_0 depends on ε, but is independent of h, such that for $\Delta t \leqslant \Delta t_0$ the error in the third order accurate SDC-LDG discretization of the Allen-Cahn equation is bounded as*

$$\max_{n\Delta t \leqslant T} \|e_u^n\| \leqslant C(h^{k+1} + \Delta t^3),$$

where C depends on $\|u\|_{L^\infty((0,T);H^{k+2}(\Omega))}$, $\|u_t\|_{L^\infty((0,T);H^{k+1}(\Omega))}$, $\|u_{tttt}\|_{L^\infty((0,T);L^2(\Omega))}$, ε, T.

4.5 Numerical tests

In this section, we will provide some numerical results to confirm the theoretical analysis. For more numerical simulations of the Allen-Cahn equation using the SDC-LDG discretization, we refer to [55].

4.5.1 Accuracy test

We consider the Allen-Cahn equation

$$u_t - \Delta u + \frac{1}{\varepsilon^2} f(u) = g(t,x,y), \quad \text{in } \ \Omega \times (0,T] \tag{4.60}$$

with periodic boundary condition on the domain $\Omega = [0,1] \times [0,1]$. We take $\varepsilon = 0.1$, and the exact solution as

$$u(t,x,y) = (1 + 0.1\sin(2\pi(x+y)) + 0.3\sin(4\pi(x+y)))\cos(t),$$

Table 4.1: Error and order of accuracy for the second order accurate SDC-LDG discretization at $T = 0.5$.

	$\mathcal{Q}_{11}$			
$m \times m$	$\|u_n - U_n\|$	Order	$\|u_n - U_n\|_{L^\infty(\Omega)}$	Order
8×8	2.70E-002	–	5.04E-002	–
16×16	7.58E-003	1.83	1.57E-002	1.68
32×32	1.94E-003	1.97	4.11E-003	1.93
64×64	4.89E-004	1.99	1.05E-003	1.97

Table 4.2: Error and order of accuracy for the third order accurate SDC-LDG discretization at $T = 0.5$.

	$\mathcal{Q}_{22}$			
$m \times m$	$\|u_n - U_n\|$	Order	$\|u_n - U_n\|_{L^\infty(\Omega)}$	Order
8×8	3.26E-003	–	7.39E-003	–
16×16	4.57E-004	2.83	1.06E-003	2.80
32×32	5.89E-005	2.96	1.37E-004	2.95
64×64	7.46E-006	2.98	1.75E-005	2.97

for which we can easily calculate the function $g(t, x, y)$.

In the computations, a uniform rectangular mesh with $m + 1$ nodes in each direction is used. The time step Δt is chosen as $\Delta t = 0.0005$ for the LDG discretization of the Allen-Cahn equation (4.60) combined with the second and third order accurate semi-implicit SDC time integration methods. The error and order of accuracy at time $T = 0.5$ are shown in Table 4.1 and Table 4.2 for, respectively, the linear and quadratic tensor product basis functions $\mathcal{Q}_{11}$ and $\mathcal{Q}_{22}$.

Tables 4.1-4.2 show that $\|u_n - U_n\|$ and $\|u_n - U_n\|_{L^\infty(\Omega)}$ converge at the rate $O(h^2)$ for the second order accurate SDC-LDG scheme using $\mathcal{Q}_{11}$ basis functions, and $\|u_n - U_n\|$ and $\|u_n - U_n\|_{L^\infty(\Omega)}$ convergence at the rate $O(h^3)$ for the third order accurate SDC-LDG scheme using $\mathcal{Q}_{22}$ basis functions, which is consistent with our theoretical analysis.

4.5.2 Dependence of stability on the ε parameter in the Allen-Cahn equation

Next, we study the dependence of the stability of the SDC-LDG discretization on the parameter ε in the Allen-Cahn equation (4.1). We consider a

Table 4.3: Maximum stable time step Δt_0 as a function of the parameter ε in the Allen-Cahn equation.

	$\mathcal{Q}_{11}$			$\mathcal{Q}_{22}$		
ε	1.00E-006	3.00E-006	4.00E-006	4.00E-006	6.00E-006	8.00E-006
Δt_0	1.482	14.149	27.295	2.321	10.208	32.547

one-dimensional problem with periodic boundary conditions and take

$$u(t,x) = \sin(2\pi x)\exp(-2t), \quad x \in [0,1]$$

as the exact solution by choosing the appropriate source term in (4.1).

In the computations, we use 200 elements in the domain $[0,1]$. The final simulation time $T = 5000$. Table 4.3 shows the maximum stable time step Δt_0 that can be chosen for the second order accurate semi-implicit SDC-LDG scheme with $\mathcal{Q}_{11}$ basis functions and the third order accurate semi-implicit SDC-LDG scheme with $\mathcal{Q}_{22}$ basis functions. From the values of Δt_0, we can observe that the time step depends on ε, which confirms the theoretical results stated in Theorems 4.3.3 and 4.4.2.

4.6 Conclusion

The semi-implicit SDC-LDG discretization provides an accurate and robust numerical method when solving the Allen-Cahn equation. The Allen-Cahn equation has a clear separation between stiff and non-stiff terms, which makes the semi-implicit SDC time integration method a good choice to solve this equation in combination with an LDG discretization. In addition, it is easy to construct SDC time discretizations for any order of accuracy, with the order increasing with one after each iteration. The LDG method is easy to use in domains with a complicated geometry.

The following results were obtained for the second and third order accurate SDC time integration methods combined with a LDG spatial discretization. For the nonlinear fully-discrete SDC-LDG discretization, we proved existence and uniqueness of the numerical solutions by making use of a standard fixed point argument in finite dimensional spaces. Stability of the SDC-LDG discretization was proven on Cartesian meshes, in the sense that stability is guaranteed if $\Delta t \leqslant \Delta t_0$, where $\Delta t_0 > 0$ depends on ε, but is independent of h. Finally, with the above time step condition, we obtained error estimates that show the optimal order of accuracy $k+1$.

4.A Proof of Theorem 4.4.2

Proof. For the following analysis, we set

$$U_{n,0} = U_n, \quad \boldsymbol{Q}_{n,0} = \boldsymbol{Q}_n.$$

• **Energy inequality**

Step 1.

Choosing v in (4.51)-(4.56, respectively, as $v = 2U_{n,1},\ 2U_{n,2},\ 2U_{n,3},\ 2U_{n,4},\ 2U_{n,5},\ 2U_{n+1}$, together with (4.32), we have

$$\begin{aligned} LHS :=& \frac{\|U_{n+1}\|^2 - \|U_n\|^2}{\Delta t} + \sum_{l=1}^{6}\left(\frac{\|U_{n,l} - U_{n,l-1}\|^2}{\Delta t} + \|\boldsymbol{Q}_{n,l}\|^2 + \frac{1}{\varepsilon^2}\|U_{n,l}^2\|^2\right) \\ =& \sum_{i=1}^{4}(\mathcal{Q}_i(\boldsymbol{Q}_{n,i+2}) + \mathcal{U}_i(U_{n,i+2})) + \mathcal{A}_1, \end{aligned} \tag{4.61}$$

where

$$\mathcal{Q}_1(\boldsymbol{\phi}) = \frac{13}{12}(\boldsymbol{Q}_{n,2}, \boldsymbol{\phi}) + \frac{4}{3}(\boldsymbol{Q}_{n,1}, \boldsymbol{\phi}) - \frac{5}{12}(\boldsymbol{Q}_n, \boldsymbol{\phi}), \tag{4.62}$$

$$\mathcal{Q}_2(\boldsymbol{\phi}) = \frac{7}{12}(\boldsymbol{Q}_{n,2}, \boldsymbol{\phi}) - \frac{2}{3}(\boldsymbol{Q}_{n,1}, \boldsymbol{\phi}) + \frac{1}{12}(\boldsymbol{Q}_n, \boldsymbol{\phi}), \tag{4.63}$$

$$\begin{aligned} \mathcal{Q}_3(\boldsymbol{\phi}) =& \frac{13}{12}(\boldsymbol{Q}_{n,4}, \boldsymbol{\phi}) + \frac{4}{3}(\boldsymbol{Q}_{n,3}, \boldsymbol{\phi}) - \frac{2}{3}(\boldsymbol{Q}_{n,2}, \boldsymbol{\phi}) + \frac{1}{3}(\boldsymbol{Q}_{n,1}, \boldsymbol{\phi}) \\ & - \frac{1}{12}(\boldsymbol{Q}_n, \boldsymbol{\phi}), \end{aligned} \tag{4.64}$$

$$\mathcal{Q}_4(\boldsymbol{\phi}) = \frac{7}{12}(\boldsymbol{Q}_{n,4}, \boldsymbol{\phi}) - \frac{2}{3}(\boldsymbol{Q}_{n,3}, \boldsymbol{\phi}) + \frac{1}{12}(\boldsymbol{Q}_n, \boldsymbol{\phi}), \tag{4.65}$$

$$\tag{4.66}$$

$$\mathcal{U}_1(v) = \frac{13}{12\varepsilon^2}(U_{n,2}^3, v) + \frac{4}{3\varepsilon^2}(U_{n,1}^3, v) - \frac{5}{12\varepsilon^2}(U_n^3, v), \tag{4.67}$$

$$\mathcal{U}_2(v) = \frac{7}{12\varepsilon^2}(U_{n,2}^3, v) - \frac{2}{3\varepsilon^2}(U_{n,1}^3, v) + \frac{1}{12\varepsilon^2}(U_n^3, v), \tag{4.68}$$

$$\begin{aligned} \mathcal{U}_3(v) =& \frac{13}{12\varepsilon^2}(U_{n,4}^3, v) + \frac{4}{3\varepsilon^2}(U_{n,3}^3, v) - \frac{2}{3\varepsilon^2}(U_{n,2}^3, v) + \frac{1}{3\varepsilon^2}(U_{n,1}^3, v) \\ & - \frac{1}{12\varepsilon^2}(U_n^3, v), \end{aligned} \tag{4.69}$$

$$\tag{4.70}$$

$$\mathcal{U}_4(v) = \frac{7}{12\varepsilon^2}(U_{n,4}^3, v) - \frac{2}{3\varepsilon^2}(U_{n,3}^3, v) + \frac{1}{12\varepsilon^2}(U_n^3, v), \tag{4.71}$$

and

$$\begin{aligned}\mathcal{A}_1 =&\frac{1}{\varepsilon^2}(U_n, U_{n,1}) + \frac{1}{\varepsilon^2}(U_{n,1}, U_{n,2}) - \frac{1}{12\varepsilon^2}(U_{n,2}, U_{n,3}) - \frac{1}{3\varepsilon^2}(U_{n,1}, U_{n,3})\\ &- \frac{7}{12\varepsilon^2}(U_n, U_{n,3}) + \frac{1}{\varepsilon^2}(U_{n,3}, U_{n,4}) + \frac{5}{12\varepsilon^2}(U_{n,2}, U_{n,4})\\ &- \frac{1}{3\varepsilon^2}(U_{n,1}, U_{n,4}) - \frac{1}{12\varepsilon^2}(U_n, U_{n,4}) - \frac{1}{12\varepsilon^2}(U_{n,4}, U_{n,5})\\ &- \frac{1}{3\varepsilon^2}(U_{n,3}, U_{n,5}) - \frac{1}{3\varepsilon^2}(U_{n,2}, U_{n,5}) - \frac{1}{3\varepsilon^2}(U_{n,1}, U_{n,5})\\ &+ \frac{1}{12\varepsilon^2}(U_n, U_{n,5}) + \frac{1}{\varepsilon^2}(U_{n,5}, U_{n+1}) + \frac{5}{12\varepsilon^2}(U_{n,4}, U_{n+1})\\ &- \frac{1}{3\varepsilon^2}(U_{n,3}, U_{n+1}) - \frac{1}{12\varepsilon^2}(U_n, U_{n+1}).\end{aligned}$$

The terms $\frac{1}{\varepsilon^2}(U_{n,i}, U_{n,j}),\ i,j = 0,1,2,\ldots,6$ can be controlled by the left hand side of (4.61) under the condition $\Delta t \leqslant \frac{\varepsilon^2}{\widetilde{C}}$, where $\widetilde{C}$ is a positive constant generated by Young's inequality. For example

$$\begin{aligned}(U_{n,3}, U_{n,5}) =&(U_{n,3} - U_{n,2}, U_{n,5} - U_{n+1}) + (U_{n,2} - U_{n,1}, U_{n,5} - U_{n+1})\\ &+ (U_{n,1} - U_n, U_{n,5} - U_{n+1}) + (U_n, U_{n,5} - U_{n+1})\\ &+ (U_{n,3} - U_{n,2}, U_{n+1}) + (U_{n,2} - U_{n,1}, U_{n+1})\\ &+ (U_{n,1} - U_n, U_{n+1}) + (U_n, U_{n+1})\\ \leqslant&\|U_n\|^2 + 2\|U_{n+1}\|^2 + 2\|U_{n,5} - U_{n+1}\|^2 + \|U_{n,3} - U_{n,2}\|^2\\ &+ \|U_{n,2} - U_{n,1}\|^2 + \|U_{n,1} - U_n\|^2. \qquad (4.72)\end{aligned}$$

In the following, we denote linear combinations of $\frac{1}{\varepsilon^2}(U_{n,i}, U_{n,j}),\ i,j = 0,1,2,\ldots,6$ as $\mathcal{A}_2,\ \mathcal{A}_3,\ \mathcal{A}_4,\ldots$.

Step 2.

Next, choose $v = -\frac{1}{2}(U_{n,5} - U_{n+1})$ in (4.55). Using (4.32), we can eliminate then $\frac{1}{3}(\boldsymbol{Q}_{n,3}, \boldsymbol{Q}_{n,5} - \boldsymbol{Q}_{n+1})$ with the following equation

$$\begin{aligned}\mathcal{B}_1 :=& -\frac{\left(U_{n,5} - U_{n,4}, \frac{1}{2}(U_{n,5} - U_{n+1})\right)}{\Delta t} = \frac{1}{4}(\boldsymbol{Q}_{n,5}, \boldsymbol{Q}_{n,5} - \boldsymbol{Q}_{n+1}) + \mathcal{A}_2\\ &- \frac{1}{4}\mathcal{Q}_3(\boldsymbol{Q}_{n,5} - \boldsymbol{Q}_{n+1}) + \frac{1}{4\varepsilon^2}(U_{n,5}^3, U_{n,5} - U_{n+1}) - \frac{1}{4}\mathcal{U}_3(U_{n,5} - U_{n+1}).\end{aligned}$$

After summation of $\mathcal{B}_1$ and (4.61), we obtain

$$LHS + \mathcal{B}_1 = \sum_{i=1}^{4} (\mathcal{Q}_i(\boldsymbol{Q}_{n,i+2}) + \mathcal{U}_i(U_{n,i+2})) + \mathcal{B}_1 + \mathcal{A}_1. \tag{4.73}$$

Step 3.

For the terms $\frac{4}{9}(\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n, \boldsymbol{Q}_{n,3})$, $\frac{4}{9}(\boldsymbol{Q}_{n,2} - \boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n,4})$ and $\frac{1}{3}(\boldsymbol{Q}_{n,4} - \boldsymbol{Q}_{n,3}, \boldsymbol{Q}_{n+1})$, we have the estimate

$$\begin{aligned}
\frac{4}{9}(\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n, \boldsymbol{Q}_{n,3}) &\leqslant \frac{2}{9}\|\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n\|^2 + \frac{2}{9}\|\boldsymbol{Q}_{n,3}\|^2, \\
\frac{4}{9}(\boldsymbol{Q}_{n,2} - \boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n,4}) &\leqslant \frac{2}{9}\|\boldsymbol{Q}_{n,2} - \boldsymbol{Q}_{n,1}\|^2 + \frac{2}{9}\|\boldsymbol{Q}_{n,4}\|^2, \\
\frac{1}{3}(\boldsymbol{Q}_{n,4} - \boldsymbol{Q}_{n,3}, \boldsymbol{Q}_{n+1}) &\leqslant \frac{1}{6}\|\boldsymbol{Q}_{n,4} - \boldsymbol{Q}_{n,3}\|^2 + \frac{1}{6}\|\boldsymbol{Q}_{n+1}\|^2.
\end{aligned} \tag{4.74}$$

In order to eliminate the left three terms in (4.74), we choose, respectively,

$$v = \frac{8}{9}(U_{n,1} - U_n), v = \frac{8}{9}(U_{n,2} - U_{n,1}), v = \frac{2}{3}(U_{n,4} - U_{n,3})$$

in (4.51), (4.52), (4.54) to obtain

$$\begin{aligned}
\mathcal{B}_2 :=& \frac{\left(U_{n,1} - U_n, \frac{8}{9}(U_{n,1} - U_n)\right)}{\Delta t} \\
=& -\frac{2}{9}(\|\boldsymbol{Q}_{n,1}\|^2 - \|\boldsymbol{Q}_n\|^2 + \|\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n\|^2) \\
& - \frac{4}{9\varepsilon^2}(U_{n,1}^3 - U_n, U_{n,1} - U_n), \\
\mathcal{B}_3 :=& \frac{\left(U_{n,2} - U_{n,1}, \frac{8}{9}(U_{n,2} - U_{n,1})\right)}{\Delta t} \\
=& -\frac{2}{9}(\|\boldsymbol{Q}_{n,2}\|^2 - \|\boldsymbol{Q}_{n,1}\|^2 + \|\boldsymbol{Q}_{n,2} - \boldsymbol{Q}_{n,1}\|^2) \\
& - \frac{4}{9\varepsilon^2}(U_{n,2}^3 - U_{n,1}, U_{n,2} - U_{n,1}), \\
\mathcal{B}_4 :=& \frac{\left(U_{n,4} - U_{n,3}, \frac{2}{3}(U_{n,4} - U_{n,3})\right)}{\Delta t} = -\frac{1}{6}(\|\boldsymbol{Q}_{n,4}\|^2 - \|\boldsymbol{Q}_{n,3}\|^2 \\
& + \|\boldsymbol{Q}_{n,4} - \boldsymbol{Q}_{n,3}\|^2) + \frac{1}{3}\mathcal{Q}_2(\boldsymbol{Q}_{n,4} - \boldsymbol{Q}_{n,3}) \\
& - \frac{1}{3\varepsilon^2}(U_{n,4}^3, U_{n,4} - U_{n,3}) + \frac{1}{3}\mathcal{U}_2(U_{n,4} - U_{n,3}) + \mathcal{A}_3.
\end{aligned}$$

Then from (4.73), we get

$$LHS + \sum_{j=1}^{4} \mathcal{B}_j = \sum_{i=1}^{4} (\mathcal{Q}_i(\boldsymbol{Q}_{n,i+2}) + \mathcal{U}_i(U_{n,i+2})) + \sum_{j=1}^{4} \mathcal{B}_j + \mathcal{A}_1. \tag{4.75}$$

Adding and subtracting $\frac{4}{9}(\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n, \boldsymbol{Q}_{n,3})$, $\frac{4}{9}(\boldsymbol{Q}_{n,2} - \boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n,4})$, $\frac{1}{3}(\boldsymbol{Q}_{n,4} - \boldsymbol{Q}_{n,3}, \boldsymbol{Q}_{n+1})$ to (4.75), together with (4.74), gives

$$\begin{aligned} LHS + \sum_{j=1}^{4} \mathcal{B}_j \leqslant & \left(\sum_{i=1}^{4} (\mathcal{Q}_i(\boldsymbol{Q}_{n,i+2}) + \mathcal{U}_i(U_{n,i+2})) - \frac{4}{9}(\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n, \boldsymbol{Q}_{n,3}) \right. \\ & \left. - \frac{4}{9}(\boldsymbol{Q}_{n,2} - \boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n,4}) - \frac{1}{3}(\boldsymbol{Q}_{n,4} - \boldsymbol{Q}_{n,3}, \boldsymbol{Q}_{n+1}) \right) \\ & + \frac{2}{9}\|\boldsymbol{Q}_{n,1} - \boldsymbol{Q}_n\|^2 + \frac{2}{9}\|\boldsymbol{Q}_{n,2} - \boldsymbol{Q}_{n,1}\|^2 + \frac{1}{6}\|\boldsymbol{Q}_{n,4} - \boldsymbol{Q}_{n,3}\|^2 \\ & + \frac{2}{9}\|\boldsymbol{Q}_{n,3}\|^2 + \frac{2}{9}\|\boldsymbol{Q}_{n,4}\|^2 + \frac{1}{6}\|\boldsymbol{Q}_{n+1}\|^2 + \sum_{j=1}^{4} \mathcal{B}_j + \mathcal{A}_1. \end{aligned} \tag{4.76}$$

Step 4.

Next, in order to deal with terms containing $\boldsymbol{Q}_{n,1}$ and $\boldsymbol{Q}_{n,2}$, we choose v in (4.51) and (4.52), respectively, as

$$\begin{aligned} v =& \frac{8}{9}(U_{n,3} - U_{n,4}) + \frac{2}{3}(U_{n,3} - U_{n,1}), \\ v =& \frac{13}{9}(U_{n,3} - U_{n,2}) + \frac{1}{3}(U_{n,3} - U_{n,5}) + \frac{2}{3}(U_{n,4} - U_{n,5}), \end{aligned}$$

which gives

$$\begin{aligned}
\mathcal{B}_5 :=& \frac{\left(U_{n,1}-U_n, \frac{8}{9}(U_{n,3}-U_{n,4})+\frac{2}{3}(U_{n,3}-U_{n,1})\right)}{\Delta t} \\
&+ \frac{\left(U_{n,2}-U_{n,1}, \frac{13}{9}(U_{n,3}-U_{n,2})+\frac{1}{3}(U_{n,3}-U_{n,5})+\frac{2}{3}(U_{n,4}-U_{n,5})\right)}{\Delta t} \\
=& -\left(\boldsymbol{Q}_{n,1}, \frac{4}{9}(\boldsymbol{Q}_{n,3}-\boldsymbol{Q}_{n,4})+\frac{1}{3}(\boldsymbol{Q}_{n,3}-\boldsymbol{Q}_{n,1})\right) \\
&- \frac{1}{\varepsilon^2}\left(U_{n,1}^3-U_n, \frac{4}{9}(U_{n,3}-U_{n,4})+\frac{1}{3}(U_{n,3}-U_{n,1})\right) \\
&- \left(\boldsymbol{Q}_{n,2}, \frac{13}{18}(\boldsymbol{Q}_{n,3}-\boldsymbol{Q}_{n,2})+\frac{1}{6}(\boldsymbol{Q}_{n,3}-\boldsymbol{Q}_{n,5})+\frac{1}{3}(\boldsymbol{Q}_{n,4}-\boldsymbol{Q}_{n,5})\right) - \frac{1}{\varepsilon^2} \\
&\left(U_{n,2}^3-U_{n,1}, \frac{13}{18}(U_{n,3}-U_{n,2})+\frac{1}{6}(U_{n,3}-U_{n,5})+\frac{1}{3}(U_{n,4}-U_{n,5})\right).
\end{aligned}$$

Using the definitions of $\mathcal{Q}_i$, $\mathcal{U}_i$ $(1 \leqslant i \leqslant 4)$ given in (4.62)-(4.71), the right hand side of $\mathcal{B}_j$ $(1 \leqslant j \leqslant 5)$ and (4.76), we obtain the energy inequality

$$\begin{aligned}
LHS + \sum_{j=1}^{5} \mathcal{B}_j =& \sum_{i=1}^{4} (\mathcal{Q}_i(\boldsymbol{Q}_{n,i+2}) + \mathcal{U}_i(U_{n,i+2})) + \sum_{j=1}^{5} \mathcal{B}_j + \mathcal{A}_1 \\
\leqslant& \sum_{k=1}^{3} \mathcal{C}_k + \sum_{l=1}^{5} \mathcal{D}_l, \qquad (4.77)
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{C}_1 =& \frac{1}{6}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{1}{4}\|\boldsymbol{Q}_{n,5}\|^2 + \frac{1}{18}\|\boldsymbol{Q}_{n,4}\|^2 + \frac{7}{18}\|\boldsymbol{Q}_{n,3}\|^2 + \frac{1}{2}\|\boldsymbol{Q}_{n,2}\|^2 \\
&+ \frac{1}{3}\|\boldsymbol{Q}_{n,1}\|^2 + \frac{2}{9}\|\boldsymbol{Q}_n\|^2, \\
\mathcal{C}_2 =& \frac{1}{3}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n,3}) + \frac{1}{9}(\boldsymbol{Q}_n, \boldsymbol{Q}_{n,4}) + \frac{1}{4}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n,5}) - \frac{1}{6}(\boldsymbol{Q}_{n,2}, \boldsymbol{Q}_{n+1}) \\
&+ \frac{1}{12}(\boldsymbol{Q}_{n,1}, \boldsymbol{Q}_{n+1}), \\
\mathcal{C}_3 =& \frac{13}{16}(\boldsymbol{Q}_{n,4}, \boldsymbol{Q}_{n,5}) + (\boldsymbol{Q}_{n,3}, \boldsymbol{Q}_{n,5}) - \frac{1}{16}(\boldsymbol{Q}_n, \boldsymbol{Q}_{n,5}) - \frac{1}{4}(\boldsymbol{Q}_{n,5}, \boldsymbol{Q}_{n+1}) \\
&+ \frac{25}{48}(\boldsymbol{Q}_{n,4}, \boldsymbol{Q}_{n+1}) + \frac{1}{16}(\boldsymbol{Q}_n, \boldsymbol{Q}_{n+1}),
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{D}_1 =&\frac{4}{9\varepsilon^2}(f(U_{n,1}),U_n)+\frac{4}{9\varepsilon^2}(f(U_{n,2}),U_{n,1})-\frac{1}{9\varepsilon^2}(f(U_{n,1}),U_{n,1})\\
&+\frac{5}{18\varepsilon^2}(f(U_{n,2}),U_{n,2})+\mathcal{A}_4,\\
\mathcal{D}_2 =&\frac{1}{3\varepsilon^2}(f(U_{n,4}),U_{n,3})+\frac{7}{9\varepsilon^2}(f(U_{n,1}),U_{n,3})-\frac{4}{9\varepsilon^2}(f(U_n),U_{n,3}),\\
\mathcal{D}_3 =&-\frac{1}{3\varepsilon^2}(f(U_{n,4}),U_{n,4})+\frac{4}{9\varepsilon^2}(f(U_{n,2}),U_{n,4})-\frac{4}{9\varepsilon^2}(f(U_{n,1}),U_{n,4})\\
&+\frac{1}{9\varepsilon^2}(f(U_n),U_{n,4}),\\
\mathcal{D}_4 =&\frac{1}{4\varepsilon^2}(f(U_{n,5}),U_{n,5})+\frac{13}{16\varepsilon^2}(f(U_{n,4}),U_{n,5})+\frac{1}{\varepsilon^2}(f(U_{n,3}),U_{n,5})\\
&+\frac{1}{4\varepsilon^2}(f(U_{n,1}),U_{n,5})-\frac{1}{16\varepsilon^2}(f(U_n),U_{n,5}),\\
\mathcal{D}_5 =&-\frac{1}{4\varepsilon^2}(f(U_{n,5}),U_{n+1})+\frac{41}{48\varepsilon^2}(f(U_{n,4}),U_{n+1})-\frac{1}{3\varepsilon^2}(f(U_{n,3}),U_{n+1})\\
&-\frac{1}{6\varepsilon^2}(f(U_{n,2}),U_{n+1})+\frac{1}{12\varepsilon^2}(f(U_{n,1}),U_{n+1})+\frac{1}{16\varepsilon^2}(f(U_n),U_{n+1}).
\end{aligned}$$

• **Estimates for the energy inequality (4.77)**

a. Estimates for $\mathcal{B}_j$ $(1\leqslant j\leqslant 5)$ and $\mathcal{C}_k$ $(1\leqslant k\leqslant 3)$

For $\mathcal{C}_2$, using the Cauchy and Young inequalities we obtain the estimate

$$\begin{aligned}
\mathcal{C}_2\leqslant&\left(\frac{1}{24}+\frac{1}{48}\right)\|\boldsymbol{Q}_{n+1}\|^2+\frac{1}{16}\|\boldsymbol{Q}_{n,5}\|^2+\frac{1}{18}\|\boldsymbol{Q}_{n,4}\|^2+\frac{1}{12}\|\boldsymbol{Q}_{n,3}\|^2\\
&+\frac{1}{6}\|\boldsymbol{Q}_{n,2}\|^2+\left(\frac{1}{3}+\frac{1}{4}+\frac{1}{12}\right)\|\boldsymbol{Q}_{n,1}\|^2+\frac{1}{18}\|\boldsymbol{Q}_n\|^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathcal{C}_1+\mathcal{C}_2\leqslant&\frac{11}{48}\|\boldsymbol{Q}_{n+1}\|^2+\frac{5}{16}\|\boldsymbol{Q}_{n,5}\|^2+\frac{1}{9}\|\boldsymbol{Q}_{n,4}\|^2+\frac{17}{36}\|\boldsymbol{Q}_{n,3}\|^2\\
&+\frac{2}{3}\|\boldsymbol{Q}_{n,2}\|^2+\|\boldsymbol{Q}_{n,1}\|^2+\frac{5}{18}\|\boldsymbol{Q}_n\|^2. \qquad (4.78)
\end{aligned}$$

In addition, the following lower bounds are obtained for $\mathcal{B}_j$ $(1 \leqslant j \leqslant 5)$

$$\begin{aligned}
\sum_{j=1}^{5} \mathcal{B}_j \geqslant & \frac{\left(\frac{8}{9}-\frac{2}{9}-\frac{1}{2}-\frac{1}{3}\right)\|U_{n,1}-U_n\|^2}{\Delta t} \\
& + \frac{\left(\frac{8}{9}-\frac{13}{18}-\frac{1}{12}-\frac{1}{3}-\frac{1}{3}\right)\|U_{n,2}-U_{n,1}\|^2 - \left(\frac{13}{18}+\frac{2}{9}\right)\|U_{n,3}-U_{n,2}\|^2}{\Delta t} \\
& + \frac{\left(\frac{2}{3}-\frac{1}{3}-\frac{8}{9}\right)\|U_{n,4}-U_{n,3}\|^2 - \left(\frac{3}{4}+\frac{1}{8}\right)\|U_{n,5}-U_{n,4}\|^2}{\Delta t} \\
& - \frac{\frac{1}{2}\|U_{n+1}-U_{n,5}\|^2}{\Delta t}. \qquad (4.79)
\end{aligned}$$

Inserting LHS, $\mathcal{C}_3$, (4.78) and (4.79) into (4.77), we obtain then the estimate

$$\begin{aligned}
& \frac{\|U_{n+1}\|^2 - \|U_n\|^2 + \frac{1}{2}\|U_{n+1}-U_{n,5}\|^2 + \frac{1}{8}\|U_{n,5}-U_{n,4}\|^2 + \frac{4}{9}\|U_{n,4}-U_{n,3}\|^2}{\Delta t} \\
& + \frac{\frac{1}{18}\|U_{n,3}-U_{n,2}\|^2 + \frac{5}{12}\|U_{n,2}-U_{n,1}\|^2 + \frac{5}{6}\|U_{n,1}-U_n\|^2}{\Delta t} \\
& + \frac{1}{2}(\|\boldsymbol{Q}_{n+1}\|^2 - \|\boldsymbol{Q}_n\|^2) + \frac{13}{48}\|\boldsymbol{Q}_{n+1}\|^2 + \frac{11}{16}\|\boldsymbol{Q}_{n,5}\|^2 + \frac{8}{9}\|\boldsymbol{Q}_{n,4}\|^2 \\
& + \frac{19}{36}\|\boldsymbol{Q}_{n,3}\|^2 + \frac{2}{9}\|\boldsymbol{Q}_n\|^2 + \frac{1}{\varepsilon^2}\sum_{i=1^6}\|U_{n,i}^2\|^2 \\
\leqslant & \frac{13}{16}(\boldsymbol{Q}_{n,4},\boldsymbol{Q}_{n,5}) + (\boldsymbol{Q}_{n,3},\boldsymbol{Q}_{n,5}) - \frac{1}{16}(\boldsymbol{Q}_n,\boldsymbol{Q}_{n,5}) - \frac{1}{4}(\boldsymbol{Q}_{n,5},\boldsymbol{Q}_{n+1}) \\
& + \frac{25}{48}(\boldsymbol{Q}_{n,4},\boldsymbol{Q}_{n+1}) + \frac{1}{16}(\boldsymbol{Q}_n,\boldsymbol{Q}_{n+1}) + \sum_{l=1}^{5}\mathcal{D}_l. \qquad (4.80)
\end{aligned}$$

b. Estimates for $\mathcal{D}_l$ $(1 \leqslant l \leqslant 5)$

For $\mathcal{D}_1$, using the Lipschitz condition (4.3) of f and (4.35), we have the

estimate

$$\begin{aligned}
\mathcal{D}_1 \leqslant & \frac{C_L}{18\varepsilon^2}\left(\|U_{n,2}-U_{n,1}\|^2+\|U_{n,1}-U_n\|^2+\|U_n\|^2\right)+\frac{4}{9\varepsilon^2}(f(U_{n,1}),U_n)\\
&+\frac{1}{3\varepsilon^2}(f(U_{n,2}),U_{n,1})+\frac{5}{18\varepsilon^2}(f(U_{n,2}),U_{n,2})+\mathcal{A}_4\\
=&\frac{C_L}{18\varepsilon^2}\left(\|U_{n,2}-U_{n,1}\|^2+\|U_{n,1}-U_n\|^2+\|U_n\|^2\right)+\frac{4}{9\varepsilon^2}(U_{n,1}^3,U_n)\\
&+\frac{1}{3\varepsilon^2}(U_{n,2}^3,U_{n,1})+\frac{5}{18\varepsilon^2}(U_{n,2}^3,U_{n,2})+\mathcal{A}_5\\
\leqslant&\frac{C_L}{18\varepsilon^2}\left(\|U_{n,2}-U_{n,1}\|^2+\|U_{n,1}-U_n\|^2+\|U_n\|^2\right)+\left(\frac{1}{3}+\frac{1}{12}\right)\|U_{n,1}^2\|^2\\
&+\left(\frac{1}{4}+\frac{5}{18}\right)\|U_{n,2}^2\|^2+\frac{1}{9}\|U_n^2\|^2+\mathcal{A}_5.
\end{aligned}$$

Analogously to the estimate for $\mathcal{D}_1$, we obtain the estimates

$$\begin{aligned}
\sum_{l=2}^{5}\mathcal{D}_l \leqslant&\frac{C_2}{\varepsilon^2}\left(\sum_{l=1}^{6}\|U_{n,l}-U_{n,l-1}\|^2+\|U_n\|^2+\|U_{n+1}\|^2\right)+\frac{1}{3\varepsilon^2}(U_{n,4}^3,U_{n,3})\\
&+\frac{1}{3\varepsilon^2}(U_{n,1}^3,U_{n,3})-\frac{2}{9\varepsilon^2}(U_{n,4}^3,U_{n,4})+\frac{1}{4\varepsilon^2}(U_{n,5}^3,U_{n,5})\\
&+\frac{3}{4\varepsilon^2}(U_{n,4}^3,U_{n,5})+\frac{1}{\varepsilon^2}(U_{n,3}^3,U_{n,5})+\frac{1}{4\varepsilon^2}(U_{n,1}^3,U_{n,5})\\
&+\frac{5}{48\varepsilon^2}(U_{n,4}^3,U_{n+1})+\frac{1}{12\varepsilon^2}(U_{n,1}^3,U_{n+1})+\frac{1}{16\varepsilon^2}(U_n^3,U_{n+1})+\mathcal{A}_6\\
\leqslant&\frac{C_2}{\varepsilon^2}\left(\sum_{l=1}^{6}\|U_{n,l}-U_{n,l-1}\|^2+\|U_n\|^2+\|U_{n+1}\|^2\right)+\frac{1}{\varepsilon^2}\\
&\Big[\left(\frac{5}{192}+\frac{1}{48}+\frac{1}{64}\right)\|U_{n+1}^2\|^2+\left(\frac{1}{4}+\frac{3}{16}+\frac{1}{4}+\frac{1}{16}\right)\|U_{n,5}^2\|^2\\
&+\left(\frac{1}{4}-\frac{2}{9}+\frac{9}{16}+\frac{5}{64}\right)\|U_{n,4}^2\|^2+\left(\frac{1}{12}+\frac{1}{12}+\frac{3}{4}\right)\|U_{n,3}^2\|^2\\
&+\left(\frac{1}{4}+\frac{3}{16}+\frac{1}{16}\right)\|U_{n,1}^2\|^2+\frac{3}{64}\|U_n^2\|^2\Big]+\mathcal{A}_6,
\end{aligned}$$

where C_2 is a positive constant generated by Young's inequality.
Inserting the estimates for $\mathcal{D}_l$ $(1 \leqslant l \leqslant 5)$ and (4.72) into (4.80), with

$\Delta t < \dfrac{\varepsilon^2}{C_3}$, we obtain

$$\begin{aligned}
&\frac{\left(1-\frac{C_3\Delta t}{\varepsilon^2}\right)\|U_{n+1}\|^2-\left(1+\frac{C_3\Delta t}{\varepsilon^2}\right)\|U_n\|^2}{\Delta t}+\frac{1}{2}(\|\boldsymbol{Q}_{n+1}\|^2-\|\boldsymbol{Q}_n\|^2)\\
&+W+\frac{1}{\varepsilon^2}\sum_{i=1^6}\|U_{n,i}^2\|^2\\
\leqslant&\frac{1}{\varepsilon^2}\Big[\frac{1}{16}\|U_{n+1}^2\|^2+\frac{3}{4}\|U_{n,5}^2\|^2+\frac{385}{576}\|U_{n,4}^2\|^2+\frac{11}{12}\|U_{n,3}^2\|^2+\frac{19}{36}\|U_{n,2}^2\|^2\\
&+\frac{11}{12}\|U_{n,1}^2\|^2+\frac{91}{576}\|U_n^2\|^2\Big],
\end{aligned}$$

where

$$\begin{aligned}
W=&\frac{13}{48}\|\boldsymbol{Q}_{n+1}\|^2+\frac{11}{16}\|\boldsymbol{Q}_{n,5}\|^2+\frac{8}{9}\|\boldsymbol{Q}_{n,4}\|^2+\frac{19}{36}\|\boldsymbol{Q}_{n,3}\|^2+\frac{2}{9}\|\boldsymbol{Q}_n\|^2\\
&-\frac{13}{16}(\boldsymbol{Q}_{n,4},\boldsymbol{Q}_{n,5})-(\boldsymbol{Q}_{n,3},\boldsymbol{Q}_{n,5})+\frac{1}{16}(\boldsymbol{Q}_n,\boldsymbol{Q}_{n,5})+\frac{1}{4}(\boldsymbol{Q}_{n,5},\boldsymbol{Q}_{n+1})\\
&-\frac{25}{48}(\boldsymbol{Q}_{n,4},\boldsymbol{Q}_{n+1})-\frac{1}{16}(\boldsymbol{Q}_n,\boldsymbol{Q}_{n+1}).
\end{aligned}$$

We denote $\boldsymbol{Y}=(\boldsymbol{Q}_{n+1},\boldsymbol{Q}_{n,5},\boldsymbol{Q}_{n,4},\boldsymbol{Q}_{n,3},\boldsymbol{Q}_n)$, and $W=\int_\Omega \boldsymbol{Y}D\boldsymbol{Y}^T dx$ with

$$D=\begin{pmatrix}
13/48 & 1/8 & -25/96 & 0 & -1/32\\
1/8 & 11/16 & -13/32 & -1/2 & 1/32\\
-25/96 & -13/32 & 8/9 & 0 & 0\\
0 & -1/2 & 0 & 19/36 & 0\\
-1/32 & 1/32 & 0 & 0 & 2/9
\end{pmatrix}.$$

It is easy to see that W is positive definite, which shows that if $\Delta t < \dfrac{\varepsilon^2}{C_3}$, we have the estimate

$$\begin{aligned}
&\left(1-\frac{C_3\Delta t}{\varepsilon^2}\right)\|U_{n+1}\|^2-\left(1+\frac{C_3\Delta t}{\varepsilon^2}\right)\|U_n\|^2+\frac{\Delta t}{2}(\|\boldsymbol{Q}_{n+1}\|^2-\|\boldsymbol{Q}_n\|^2)\\
&+\frac{\Delta t}{2\varepsilon^2}(\|U_{n+1}^2\|^2-\|U_n^2\|^2)\leqslant 0,
\end{aligned}$$

Finally, similar to the analysis of (4.34), we obtain

$$\begin{aligned}
&C_4\|U_{n+1}\|^2+\frac{\Delta t}{2}\|\boldsymbol{Q}_{n+1}\|^2+\frac{\Delta t}{2\varepsilon^2}\|U_{n+1}^2\|^2\\
\leqslant&\exp\left(\frac{C_5T}{\varepsilon^2}\right)(\|U_0\|^2+\frac{\Delta t}{2}\|\boldsymbol{Q}_0\|^2+\frac{\Delta t}{2\varepsilon^2}\|U_0^2\|^2).
\end{aligned}$$

□

Chapter 5

Conclusions and Outlook

In this dissertation, we study higher order accurate time-implicit Discontinuous Galerkin (DG) discretizations for several classes of nonlinear partial differential equations (PDEs). The main conclusions are as follows:

- Based on the Karush-Kuhn-Tucker (KKT) limiter, which imposes bounds on the numerical solution using Lagrange multipliers, higher order accurate bounds preserving time-implicit Local Discontinuous Galerkin (LDG) and DG discretizations were constructed, respectively, for nonlinear degenerate parabolic equations and the chemically reactive Euler equations. Numerical results demonstrate that the bounds preserving time-implicit discretizations are of optimal order of accuracy and accurate in preserving the bounds on the numerical solution, even on coarse meshes and for relatively large time steps.

- The unique solvability and unconditional entropy dissipation of the positivity preserving KKT-LDG discretizations were proven for nonlinear degenerate parabolic equations. This analysis gives the theoretical support for the use of the KKT limiter, that is, the KKT-LDG discretizations preserve the positivity of the numerical solutions and are numerically stable.

- Stability and error estimates for second and third order accurate time-implicit Spectral Deferred Correction (SDC) LDG discretizations were proven for the Allen-Cahn equation. The theoretical analysis addresses the fully discrete analysis, both in space and time, of the higher order accurate time-implicit discretizations for the Allen-

Cahn equation. Numerical examples are presented to illustrate the theoretical results.

Regarding bounds preserving limiters and error estimates for higher order accurate time-implicit DG discretizations, there are several topics that are interesting for further research:

- Proving error estimates for the higher order bounds preserving KKT-DIRK-LDG discretizations for nonlinear degenerate parabolic PDEs.
- Developing higher order time-implicit bounds preserving discretizations for the compressible Navier-Stokes equations modelling multi-species chemically reactive flows.
- Proving unconditional energy stability and optimal error estimates for higher order accurate time-implicit LDG discretizations for the Cahn-Hilliard equation. This will provide an important extension of the results obtained in Chapter 4 for the Allen-Cahn equation.

Summary

This dissertation discusses higher order accurate time-implicit Discontinuous Galerkin (DG) discretizations for several classes of nonlinear Partial Differential Equations (PDEs). The two main topics considered are bounds preserving limiters combined with Diagonally Implicit Runge-Kutta (DIRK) methods, and novel efficient higher order accurate semi-implicit Spectral Deferred Correction (SDC) DG discretizations, including error estimates.

In Chapter 2, positivity constraints are imposed on time-implicit Local Discontinuous Galerkin (LDG) discretizations of degenerate parabolic equations using Lagrange multipliers. In addition, mass conservation of the positivity limited solution is ensured by imposing a mass conservation equality constraint. This results in a mixed complementarity problem, which is expressed by the Karush-Kuhn-Tucker (KKT) equations. This approach results in a direct coupling of the bounds constraints and the DG discretization and is well suited for time-implicit discretizations. We call this approach to enforce bounds constraints "KKT-limiter", which differs significantly from frequently used limiters in combination with explicit time integration methods that generally suffer from serious time step constraints for parabolic or stiff hyperbolic PDEs. The KKT-DIRK-LDG discretizations preserve higher order accuracy, and allow for a significantly larger time step than the time-explicit bounds preserving DG discretizations. We prove entropy stability, both for the unlimited DIRK-LDG discretization and the KKT-DIRK-LDG discretization, and unique solvability of the KKT-DIRK-LDG discretizations. Finally, numerical results are shown which illustrate the higher order accuracy and entropy dissipation of the positivity preserving KKT-DIRK-LDG discretizations.

In Chapter 3, we develop higher order accurate bounds preserving time-implicit DG discretizations for the chemically reactive Euler equations. These equations are used to describe inviscid, compressible chemically reacting flows, including detonations. Since in chemically reactive flows, the

time step can be significantly limited by the large difference between the fluid dynamics time scales and the reaction time scales, we use a fractional step method, which separates the convection and reaction steps, and combine this with higher order accurate DIRK methods for the time discretization. In order to ensure that the density and pressure are nonnegative, and mass fractions are in the range between zero and one, the KKT limiter is adopted to construct bounds preserving DIRK-DG discretizations for the reactive Euler equations. In order to deal with the stiff source terms in chemically reactive flows, we use Harten's subcell resolution technique in the reaction step. This ensures proper wave speeds in the reaction zone and improves stability. Numerical examples demonstrate that the KKT-DIRK-DG discretization results in the correct wave speed, preserves the physical bounds on the solution, and compares well with exact solutions and accurate reference solutions obtained with explicit bounds preserving discretizations for the chemically reactive Euler equations. Without the KKT limiter, most computations break down due to unphysical solutions.

In Chapter 4, we prove stability and error estimates for second and third order accurate semi-implicit SDC-LDG discretizations of the Allen-Cahn equation. For the numerical discretization of this parabolic equation, implicit time integration methods, which alleviate the time-step restrictions, result in a nonlinear system of equations that must be solved each time step. We first prove the unique solvability of the implicit SDC-LDG discretizations through a standard fixed point argument in finite dimensional space. Next, by a careful selection of the test functions, stability and error estimates for second and third order accurate time-implicit SDC-LDG discretizations are obtained in the sense that the time step only requires a positive upper bound and is independent of the mesh size. Also, numerical examples are presented that illustrate the theoretical results.

Samenvatting

Dit proefschrift bespreekt tijdsimpliciete discontinue Galerkin (DG) methoden voor verschillende klassen van hogere orde niet-lineaire partiële differentiaalvergelijkingen (PDVs). De twee belangrijkste onderwerpen zijn discontinue Galerkin discretizaties met limiters die de numerieke oplossing binnen de fysische grenzen moeten houden in combinatie met diagonaal impliciete Runge-Kutta (DIRK) tijdsintegratie methoden, en nieuwe efficiënte hogere orde nauwkeurige semi-impliciete spectral deferred correction (SDC) tijdsintegratiemethoden in combinatie met local discontinue Galerkin (LDG) discretizaties, inclusief foutafschattingen.

In hoofdstuk 2 worden via limiters positiviteitsvoorwaarden opgelegd aan tijdsimpliciete local discontinuous Galerkin discretizaties die met behulp van Lagrange multipliers zorgen dat de numerieke oplossing van ontaarde parabolische vergelijkingen positief is. Bovendien wordt massabehoud van de numerieke discretizatie met de positiviteits limiter gegarandeerd door massabehoud als extra voorwaarde in de numerieke discretizatie op te leggen. Dit resulteert in een gemengd complementariteitsprobleem dat wordt beschreven door de Karush-Kuhn-Tucker (KKT) vergelijkingen. Deze aanpak resulteert in een directe koppeling van de positiviteitseisen waaraan de numerieke oplossing moet voldoen met de DG discretizatie en is zeer geschikt voor tijdsimpliciete numerieke discretizaties. Wij noemen deze aanpak om positiviteitseisen aan de numerieke oplossing op te leggen "KKT-limiter". Deze aanpak verschilt sterk van de vaak gebruikte technieken om positiviteitseisen aan de numerieke oplossing op te leggen bij tijdsexpliciete numerieke discretizaties. De KKT-DIRK-LDG numerieke discretizaties hebben een hogere orde nauwkeurigheid en staan een aanzienlijk grotere tijdstap toe dan wanneer tijdsexpliciete positiviteitbehoudende DG discretizaties worden gebruikt. We bewijzen entropie-stabiliteit, zowel voor de DIRK-LDG discretizaties als voor de KKT-DIRK-LDG discretizaties, en ook de uniciteit van de oplossing van de KKT-DIRK-LDG discretizaties. Tenslotte worden numerieke resultaten getoond die de ho-

gere orde nauwkeurigheid, entropiedissipatie en de positiviteitbehoudende eigenschappen van de KKT-DIRK-LDG discretizaties illustreren.

In hoofdstuk 3 ontwikkelen we hogere orde nauwkeurige tijdsimpliciete positiviteitsbehoudende DG discretizaties voor de chemisch reactieve Euler vergelijkingen. Deze vergelijkingen worden gebruikt om niet-visceuze, samendrukbare chemisch reagerende stromingen, inclusief detonaties, te beschrijven. Aangezien in stromingen met chemische reacties de tijdstap in de numerieke tijdsdiscretizatie aanzienlijk kan worden beperkt door het grote verschil tussen de vloeistofdynamische tijdschalen en de reactie tijdschalen gebruiken we een fractionele stapmethode, die de convectie en reactiestappen scheidt, en combineren deze methode met hogere orde nauwkeurige DIRK tijdsintegratiemethoden. Om ervoor te zorgen dat de dichtheid en de druk niet negatief worden, en de massafracties tussen nul en één liggen, wordt de KKT-limiter toegepast om te zorgen dat numerieke oplossingen van de DIRK-DG numerieke discretizaties van de reactieve Euler vergelijkingen aan deze fysische eisen voldoet. Om stijve brontermen in chemisch reactieve stromingen nauwkeurig te kunnen discretizeren gebruiken we in de reactiestap de subcelresolutie-techniek van Harten. Dit zorgt voor de juiste golfsnelheden en reactiesnelheden in de reactiezone en verbetert de stabiliteit van de numerieke methode. Numerieke simulaties tonen aan dat de KKT-DIRK-DG discretizatie resulteert in de juiste golfsnelheid, dat de numerieke oplossing voldoet aan de fysische grenzen en goed overeenkomt met exacte oplossingen en nauwkeurige referentieoplossingen die verkregen zijn met positiviteitsbehoudende tijdsexpliciete numerieke discretizaties van de chemisch reactieve Euler vergelijkingen. Zonder de KKT-limiter lopen de meeste berekeningen stuk op niet-fysische oplossingen.

In hoofdstuk 4 bewijzen we de numerieke stabiliteit en geven foutschattingen voor tweede en derde orde nauwkeurige semi-impliciete SDC-LDG discretizaties van de Allen-Cahn vergelijking. Impliciete tijdsdiscretizaties van deze parabolische vergelijking resulteren in een niet-lineair stelsel van vergelijkingen dat elke tijdstap moet worden opgelost. Eerst bewijzen we de uniciteit van de oplossing van de semi-impliciete SDC-LDG discretizaties in een eindig dimensionale ruimte door middel van een standaard dekpunt argument. Vervolgens worden door een zorgvuldige selectie van de testfuncties voorwaarden voor de numerieke stabiliteit en foutschattingen voor de tweede en derde orde nauwkeurige tijdsimpliciete SDC-LDG discretizaties verkregen. Hierbij worden condities voor de grootte van de tijdstap afgeleid die onafhankelijk zijn van de maaswijdte van het rekenrooster. Ook worden numerieke simulaties gepresenteerd die de theoretische resultaten illustreren.

Bibliography

[1] N. Abdallah, I. Gamba and G. Toscani, *On the minimization problem of sub-linear convex functionals*, Kinet. Relat. Mod., 4 (2011), 857–871.

[2] R. Abedian and M. Dehghan, *A RBF-WENO finite difference scheme for non-linear degenerate parabolic equations*, J. Sci. Comput., 93 (2022), 60.

[3] A. Aderogba and M. Chapwanya, *An explicit nonstandard finite difference scheme for the Allen-Cahn equation*, J. Difference Equ. Appl., 21 (2015), 875–886.

[4] R. Agarwal and D. O'Regan, *Ordinary and partial differential equations*, Springer Science & Business Media, 2009.

[5] R. Alexander, *Diagonally implicit Runge-Kutta methods for stiff ODE's*, SIAM J. Numer. Anal., 14 (1977), 1006–1021.

[6] S. Allen and J. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metall., 27 (1979), 1084–1095.

[7] P. Amestoy, I. Duff, D. Ruiz and B. Uçar, *A parallel matrix scaling algorithm*, in International Conference on High Performance Computing for Computational Science, Springer, 2008, 301–313.

[8] D. Arnold, F. Brezzi, B. Cockburn and L. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), 1749–1779.

[9] M. Baccouch, *Optimal energy-conserving local discontinuous Galerkin method for the one-dimensional sine-Gordon equation*, Int. J. comput. Math., 94 (2017), 316–344.

[10] W. Bao and S. Jin, *The random projection method for stiff detonation capturing*, SIAM J. Sci. Comput., 23 (2001), 1000–1026.

[11] W. Bao and S. Jin, *The random projection method for stiff multispecies detonation capturing*, J. Comput. Phys., 178 (2002), 37–57.

[12] M. Beneš, V. Chalupecký and K. Mikula, *Geometrical image segmentation by the Allen-Cahn equation*, Appl. Numer. Math., 51 (2004), 187–205.

[13] M. Bessemoulin-Chatard and F. Filbet, *A finite volume scheme for nonlinear degenerate parabolic equations*, SIAM J. Sci. Comput., 34 (2012), B559–B583.

[14] B. Bihari and D. Schwendeman, *Multiresolution schemes for the reactive Euler equations*, J. Comput. Phys., 154 (1999), 197–230.

[15] S. Blanes, F. Casas, P. Chartier and A. Murua, *Optimized high-order splitting methods for some classes of parabolic equations*, Math. Comput., 82 (2013), 1559–1576.

[16] S. Boscarino, J. Qiu and G. Russo, *Implicit-Explicit integral deferred correction methods for stiff problems*, SIAM J. Sci. Comput., 40 (2018), A787–A816.

[17] M. Burger, J. Carrillo and M.-T. Wolfram, *A mixed finite element method for nonlinear diffusion equations*, Kinet. Relat. Mod., 3 (2010), 59–83.

[18] J. Butcher, *Coefficients for the study of Runge-Kutta integration processes*, J. Australian Math. Soc., 3 (1963), 185–201.

[19] J. Carrillo, A. Chertock and Y. Huang, *A finite-volume method for nonlinear nonlocal equations with a gradient flow structure*, Commun. Comput. Phys., 17 (2015), 233–258.

[20] J. Carrillo, A. Jüngel, P. Markowich, G. Toscani and A. Unterreiter, *Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities*, Monatsh. Math., 133 (2001), 1–82.

[21] J. Carrillo, P. Laurençot and J. Rosado, *Fermi-Dirac-Fokker-Planck equation: Well-posedness and long-time asymptotics*, J. Differ. Equ., 247 (2009), 2209–2234.

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

[23] Q. Cheng and J. Shen, *A new Lagrange multiplier approach for constructing structure preserving schemes, I. Bound preserving*, SIAM J. Numer. Anal., 60 (2022), 970-998.

[24] N. Chuenjarern, Z. Xu and Y. Yang, *High-order bound-preserving discontinuous Galerkin methods for compressible miscible displacements in porous media on triangular meshes*, J. Comput. Phys., 378 (2019), 110–128.

[25] B. Cockburn, S. Hou and C.-W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case*, Math. Comp., 54 (1990), 545–581.

[26] B. Cockburn, G. Kanschat, I. Perugia and D. Schötzau, *Superconvergence of the local discontinuous Galerkin method for elliptic problems on Cartesian grids.*, SIAM J. Numer. Anal., 39 (2001), 264–285.

[27] B. Cockburn, S. Lin and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one dimensional systems*, J. Comput. Phys., 84 (1989), 90–113.

[28] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems*, J. Comput. Phys., 141 (1989), 199–224.

[29] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework*, Math. Comp., 52 (1989), 411–435.

[30] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), 2440–2463.

[31] P. Colella, A. Majda and V. Roytburd, *Theoretical and numerical structure for reacting shock waves*, SIAM J. Sci. and Stat. Comput., 7 (1986), 1059–1080.

[32] D. Deng and Z. Zhao, *Efficiently energy-dissipation-preserving ADI methods for solving two-dimensional nonlinear Allen-Cahn equation*, Comput. Math. Appl., 128 (2022), 249–272.

[33] S. Descombes, *Convergence of a splitting method of high order for reaction-diffusion system*, Math. Comput., 70 (2000), 1481–1501.

[34] D. Di Pietro and A. Ern, *Mathematical aspects of discontinuous Galerkin methods*, Spring, 2010.

[35] B. Dong and C.-W. Shu, *Analysis of a local discontinuous Galerkin method for linear time-dependent fourth-order problems*, SIAM J. Numer. Anal., 47 (2009), 3240–3268.

[36] J. Du, C. Wang, C. Qian and Y. Yang, *High-order bound-preserving discontinuous Galerkin methods for stiff multispecies detonation*, SIAM J. Sci. Comput., 41 (2019), B250–B273.

[37] J. Du and Y. Yang, *Third-order conservative sign-preserving and steady-state-preserving time integrations and applications in stiff multispecies and multireaction detonations*, J. Comput. Phys., 395 (2019), 489–510.

[38] J. Du and Y. Yang, *High-order bound-preserving discontinuous Galerkin methods for multicomponent chemically reacting flows*, J. Comput. Phys., 469 (2022), 111548.

[39] A. Dutt, L. Greengard and V. Rokhlin, *Spectral deferred correction methods for ordinary differential equations*, BIT Numer. Math., 40 (2000), 241–266.

[40] J. Evans, T. Hughes and G. Sangalli, *Enforcement of constraints and maximum principles in the variational multiscale method*, Comput. Methods Appl. Mech. Engrg., 199 (2009), 61–76.

[41] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems, Volume I*, Springer, 2003.

[42] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems, Volume II*, Springer, 2003.

[43] W. Feng, H. Guo, Z. Xu and Y. Yang, *Conservative numerical methods for the reinterpreted discrete fracture model on non-conforming meshes and their applications in contaminant transportation in fractured porous media*, Adv. Water Resour., 153 (2021), 103951.

[44] W. Feng, H. Guo, Y. Kang and Y. Yang, *Bound-preserving discontinuous Galerkin methods with second-order implicit pressure explicit concentration time marching for compressible miscible displacements in porous media*, J. Comput. Phys., 463 (2022), 111240.

[45] X. Feng and O. Karakashian, *Fully discrete dynamic mesh discontinuous Galerkin methods for the Cahn-Hilliard equation of phase transition*, Math. Comp., 76 (2007), 1093–1117.

[46] X. Feng and Y. Li, *Analysis of symmetric interior penalty discontinuous Galerkin methods for the Allen-Cahn equation and the mean curvature flow*, IMA J. Numer. Anal., 35 (2015), 1622–1651.

[47] X. Feng, Y. Li and Y. Zhang, *Finite element methods for the stochastic Allen-Cahn equation with gradient-type multiplicative noise*, IMA J. Numer. Anal., 55 (2017), 194–216.

[48] X. Feng, H. Song, T. Tang and J. Yang, *Nonlinear stability of the implicit-explicit methods for the Allen-Cahn equation*, Inverse Probl. Imaging, 7 (2013), 679–695.

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

[50] X. Feng and H. Wu, *A posteriori error estimates and an adaptive finite element method for the Allen-Cahn equation and the mean curvature flow*, J. Sci. Comput., 24 (2005), 121–146.

[51] R. Glowinski, S. J. Osher and W. Yin, *Splitting methods in communication, imaging, science, and engineering*, Springer, 2017.

[52] S. Gottlieb, Z. Grant, J. Hu and R. Shu, *High order strong stability preserving multiderivative implicit and IMEX Runge–Kutta methods with asymptotic preserving properties*, SIAM J. Numer. Anal., 60 (2022), 423–449.

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

[54] R. Guo, F. Filbet and Y. Xu, *Efficient high order semi-implicit time discretization and local discontinuous Galerkin methods for highly nonlinear PDEs*, J. Sci. Comput., 68 (2016), 1029–1054.

[55] R. Guo, L. Ji and Y. Xu, *High order local discontinuous Galerkin methods for the Allen-Cahn equation: analysis and simulation*, J. Comput. Math., 34 (2016), 135–158.

[56] R. Guo, Y. Xia and Y. Xu, *An efficient fully-discrete local discontinuous Galerkin method for the Cahn-Hilliard-Hele-Shaw system*, J. Comput. Phys., 264 (2014), 23–40.

[57] R. Guo, Y. Xia and Y. Xu, *Semi-implicit spectral deferred correction methods for highly nonlinear partial differential equations*, J. Comput. Phys., 338 (2017), 269–284.

[58] R. Guo and Y. Xu, *A high order adaptive time-stepping strategy and local discontinuous Galerkin method for the modified phase field crystal equation*, Commun. Comput. Phys., 24 (2018), 123–151.

[59] R. Guo and Y. Xu, *Local discontinuous Galerkin method and high order semi-implicit scheme for the phase field crystal equation*, SIAM J. Sci. Comput., 38 (2016), A105–A127.

[60] R. Guo, Y. Xu and Z. F. Xu, *Local discontinuous Galerkin methods for the functionalized Cahn-Hilliard equation*, J. Sci. Comput., 63 (2015), 913–937.

[61] E. Hairer and G. Wanner, *Solving ordinary differential equations II. Stiff and differential-algebraic problem*, Springer Science & Business Media, 2010.

[62] A. Harten, *ENO schemes with subcell resolution*, J. Comput. Phys., 83 (1989), 148–184.

[63] F. Heberle and G. Feigenson, *Phase separation in lipid membranes*, Cold Spring Harb. Perspect. Biol., 3 (2011), a004630.

[64] A. Heibig and A. Petrov, *Local existence result in time for a drift-diffusion system with Robin boundary conditions*, Z. Angew. Math. Phys., 70 (2019), 162.

[65] V. Heningburg and C. Hauck, *A hybrid finite-volume, discontinuous Galerkin discretization for the radiative transport equation*, Multiscale Model. Simul., 19 (2021), 1–24.

[66] J. Hesthaven and T. Warburton, *Nodal discontinuous Galerkin methods*, Spring, 2008.

[67] F. Huang and J. Shen, *Bound/Positivity preserving and energy stable scalar auxiliary variable schemes for dissipative systems: applications to Keller-Segel and Poisson-Nernst-Planck equations*, SIAM J. Sci. Comput., 43 (2021), A1832–A1857.

[68] J. Huang and C.-W. Shu, *Bound-preserving modified exponential Runge-Kutta discontinuous Galerkin methods for scalar hyperbolic equations with stiff source terms*, J. Comput. Phys., 361 (2018), 111–135.

[69] K. Ito and K. Kunisch, *Lagrange multiplier approach to variational problems and applications*, SIAM, 2008.

[70] H. Jia and K. Li, *A third accurate operator splitting method*, Math. Comput. Model., 53 (2011), 387–396.

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

[72] Y. Kalmykov, W. Coffey and S. Titov, *On the Brownian motion in a double-well potential in the overdamped limit*, Physica A, 377 (2007), 412–420.

[73] C. Kao, A. Kurganov, Z. Qu and Y. Wang, *A fast explicit operator splitting method for modified Buckley-Leverett equations*, J. Sci. Comput, 64 (2015), 837–857.

[74] C. Kennedy and M. Carpenter, *Diagonally implicit Runge-Kutta methods for ordinary differential equations. A review*, NASA Langley Research Center Hampton, 2016.

[75] V. Korobeinikov, *Problems of point blast theory*, AIP Press, New York, 1991.

[76] D. Kotov, H. Yee, W. Wang and C.W. Shu, *On spurious numerics in solving reactive equations*, Proceedings of the ASTRONUM-2012, The Big Island, Hawaii.

[77] P. Kuberry, P. Bochev and K. Peterson, *An optimization-based approach for elliptic problems with interfaces*, SIAM J. Sci. Comput, 39 (2017), S757–S781.

[78] R. Lan, J. Li, Y. Cai and L. Ju, *Operator splitting based structure-preserving numerical schemes for the mass-conserving convective Allen-Cahn equation*, J. Comput. Phys., 472 (2023), 111695.

[79] J. Lee and B. Fornberg, *A split step approach for the 3-D Maxwell's equations*, J. Comput. Appl. Math., 158 (2003), 485–505.

[80] R. LeVeque, *Numerical methods for conservation laws*, Spring, 1992.

[81] R. LeVeque and H. Yee, *A study of numerical methods for hyperbolic conservation laws with stiff source terms*, J. Comput. Phys., 86 (1990), 187–210.

[82] F. Liu and J. Shen, *Stabilized semi-implicit spectral deferred correction methods for Allen-Cahn and Cahn-Hilliard equations*, Math. Methods Appl. Sci., 38 (2015), 4564–4575.

[83] H. Liu and Z. Wang, *An entropy satisfying discontinuous Galerkin method for nonlinear Fokker-Planck equations*, J. Sci. Comput., 68 (2016), 1217–1240.

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

[85] H. Liu and H. Yu, *The entropy satisfying discontinuous Galerkin method for Fokker-Planck equations*, J. Sci. Comput., 62 (2015), 803–830.

[86] X. Liu, Y. Yang and H. Guo, *High-order bound-preserving finite difference methods for incompressible wormhole propagation*, J. Sci. Comput., 89 (2021), 7.

[87] Y. Lv and M. Ihme, *Discontinuous Galerkin method for multicomponent chemically reacting flows and combustion*, J. Comput. Phys., 270 (2014), 105–137.

[88] V. Michel-Dansac and A. Thomann, *TVD-MOOD schemes based on implicit-explicit time integration*, Appl. Math. Comput., 433 (2022), 127397.

[89] M. Minion, *Semi-implicit spectral deferred correction methods for ordinary differential equations*, Commun. Math. Sci., 1 (2003), 471–500.

[90] J. Nachbar, *Finite Dimensional Optimization Part I: The KKT Theorem*, Washington University, 2018.

[91] J. Nachbar, *Finite Dimensional Optimization Part II: Sufficiency*, Washington University, 2020.

[92] J. Nocedal and S. Wright, *Numerical Optimization*, Springer-Verlag, New York, Berlin, 2006.

[93] L. Ortellado and L. Góme, *Phase-field modeling of dendritic growth on spherical surfaces*, Front. Mater., 7 (2020), 00163.

[94] T. Qin and C.-W. Shu, *Implicit positivity-preserving high-order discontinuous Galerkin methods for conservation laws*, SIAM J. Sci. Comput., 40 (2018), A81–A107.

[95] T. Qin, C.-W. Shu and Y. Yang, *Bound-preserving discontinuous Galerkin methods for relativistic hydrodynamics*, J. Comput. Phys., 315 (2016), 323–347.

[96] W. Reed and T. Hill, *Triangular mesh method for the neutron transport equation*, Technical report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[97] 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 (2010), 1159–1179.

[98] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), 439–471.

[99] L. Skvortsov, *Diagonally implicit Runge-Kutta methods for stiff problems*, Comput. Math. Math. Phys., 46 (2006), 2110–2123.

[100] H. Song and C.-W. Shu, *Unconditional energy stability analysis of a second order implicit-explicit local discontinuous Galerkin method for the Cahn-Hilliard equation*, J. Sci. Comput., 73 (2017), 1178–1203.

[101] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, Springer-Verlag, New York, Berlin, 1991.

[102] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), 506–517.

[103] M. Svärd and S. Mishra, *Implicit-explicit schemes for flow equations with stiff source terms*, J. Comput. Appl. Math., 235 (2011), 1564–1577.

[104] R. Temam, *Navier-Stokes equations, theory and numerical analysis.*, AMS Chelsea Publishing, Providence, RI, 2001.

[105] V. Thomée, *Galerkin finite element methods for parabolic problems*, Springer, 2006.

[106] L. Tian, Y. Xu, J. G. Kuerten and J. J. W. van der Vegt, *An h-adaptive local discontinuous Galerkin method for the Navier-Stokes-Korteweg equations*, J. Comput. Phys., 319 (2016), 242–265.

[107] E. Toro, *Riemann Solvers and Numerical Methods For Dynamic*, Spring, 2009

[108] L. Tosatto and L. Vigevano, *Numerical solution of under-resolved detonations*, J. Comput. Phys., 227 (2008), 2317–2343.

[109] G. Toscani, *Finite time blow up in Kaniadakis-Quarati model of Bose-Einstein particles*, Commun. Part. Diff. Eq., 37 (2012), 77–87.

[110] J. J. W. van der Vegt and H. van der Ven, *Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows*, J. Comput. Phys., 182 (2002), 546–585.

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

[112] J. Vázquez, *The porous medium equation: mathematical theory*, Oxford University Press, 2007.

[113] C. Wang, X. Zhang, C.-W. Shu and J. Ning, *Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations*, J. Comput. Phys., 231 (2012), 653–665.

[114] H. Wang, C.-W. Shu and Q. Zhang, *Stability and error estimate of local discontinuous Galerkin methods with implicit-explicit time-marching for advection-diffusion problems*, SIAM J. Numer. Anal., 53 (2015), 206–227.

[115] H. Wang, C.-W. Shu and Q. Zhang, *Stability and error estimate of local discontinuous Galerkin methods with implicit-explicit time-marching for nonlinear convection-diffusion problems*, Appl. Math. Comput., 272 (2016), 237–258.

[116] H. Wang, S. Wang, Q. Zhang and C.-W. Shu, *Local discontinuous Galerkin methods with implicit-explicit time-marching for multi-dimensional convection-diffusion problems*, ESAIM: Math. Model. Numer. Anal., 50 (2016), 1083–1105.

[117] W. Wang, C.-W. Shu, H. C. Yee, D. V. Kotov and B. Sjögreen, *High order finite difference methods with subcell resolution for stiff multi-species discontinuity capturing*, Commun. Comput. Phys., 17 (2015), 317–336.

[118] W. Wang, C.-W. Shu, H. C. Yee and B. Sjögreen, *High order finite difference methods with subcell resolution for advection equations with stiff source terms*, J. Comput. Phys., 231 (2012), 190–214.

[119] Y. Wang, M. Tang and J. Fu, *Uniform convergent scheme for discrete-ordinate radiative transport equation with discontinuous coefficients on unstructured quadrilateral meshes*, Partial Differ. Equ. Appl., 3 (2022), 61.

[120] Y. Xia, Y. Xu and C.-W. Shu, *Efficient time discretization for local discontinuous Galerkin methods*, Discrete Contin. Dyn. Syst. Ser. B, 8 (2007), 677–693.

[121] Y. Xia, Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for the Cahn-Hilliard type equations*, J. Comput. Phys., 227 (2007), 472–491.

[122] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for nonlinear Schrödinger equations*, J. Comput. Phys., 205 (2005), 72–97.

[123] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin method for surface diffusion and Willmore flow of graphs*, J. Sci. Comput., 40 (2009), 375–390.

[124] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for the degasperis-procesi equation*, Commun. Comput. Phys., 7 (2010), 1–46.

[125] Y. Xu and C.-W. Shu, *Optimal error estimate of the semi-discrete local discontinuous Galerkin methods for high order wave equations*, SIAM J. Numer. Anal., 50 (2012), 79–104.

[126] Y. Yang, D. Wei and C.-W. Shu, *Discontinuous Galerkin method for Krause's consensus models and pressureless Euler equations*, J. Comput. Phys., 252 (2013), 109–127.

[127] H. Yee, D. Kotov, W. Wang and C.-W. Shu, *Spurious behavior of shock-capturing methods by the fractional step approach: problems containing stiff source terms and discontinuities*, J. Comput. Phys., 241 (2013), 266–291.

[128] K. Yokota and T. Ogino, *Phase separation in lipid bilayer membranes induced by intermixing at a boundary of two phases with different components*, Chem. Phys. Lipids, 191 (2015), 147–152.

[129] F. Zhang, Y. Xu, F. Chen and R. Guo, *Interior penalty discontinuous Galerkin based isogeometric analysis for Allen-Cahn equations on surfaces*, Commun. Comput. Phys., 18 (2015), 1380–1416.

[130] J. Zhang and Q. Du, *Numerical studies of discrete approximations to the Allen-Cahn equation in the sharp interface limit*, SIAM J. Sci. Comput., 31 (2009), 3042–3063.

[131] Q. Zhang and Z.-L. Wu, *Numerical simulation for porous medium equation by local discontinuous Galerkin finite element method*, J. Sci. Comput., 38 (2009), 127–148.

[132] X. Zhang, *On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier-Stokes equations*, J. Comput. Phys., 328 (2017), 301–343.

[133] X. Zhang and C.-W. Shu, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, J. Comput. Phys., 229 (2010), 3091–3120.

[134] X. Zhang and C.-W. Shu, *Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms*, J. Comput. Phys., 230 (2011), 1238–1248.

[135] L. Zhou and Y. Xu, *Stability analysis and error estimates of semi-implicit spectral deferred correction coupled with local discontinuous*

Galerkin method for linear convection–diffusion equations, J. Sci. Comput., 77 (2018), 1001–1029.

Acknowledgements

The two years at the University of Twente have been a wonderful experience, which were full of challenges and nice experiences. I am very grateful to my supervisors Professors Jaap van der Vegt and Yan Xu for giving me the chance to be a joint PhD student between the University of Twente and the University of Science and Technology of China.

First, I want to express my sincere gratitude to Jaap for his continuous encouragement and guidance during this project. He always shows constant patience to introduce me to the topic of bounds preserving limiters, and to teach me how to write and present papers. I can hardly describe in words how his consistent support and criticism helped me to improve my research. Besides research, I also want to thank Jaap and his wife Phing for helping me to adapt to the life in the Netherlands. They always care about my daily life and introduced me to some nice food and supermarkets.

At the same time, my sincere thanks go to Prof. Yan Xu for her insightful guidance and suggestions in my research. She not only gives me a lot of academic freedom, but always provides me with many useful references and nice ideas when my research is stagnant.

I want to thank the secretaries Marielle and Linda who helped me with all the administrative issues at the University of Twente. I want to thank my group members Kaifang, Xiangyi and Xiaoyu for all the assistance in both research and life, and I also want to thank Lars, Marek, Nishant, Olena, Poorvi and Sjoerd who helped to create a warm working atmosphere.

Finally, I want to give my special gratitude to my parents and my husband for their unconditional love, understanding and support. For better or worse, they always be there for me.