Discontinuous Galerkin Methods: Stability of Time Discretizations and Applications to Gradient Flows

by

Zheng Sun

B.Sc., University of Science and Technology of China, Anhui, P.R. China, 2014

M.Sc., Brown University, RI, U.S.A., 2015

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2018

This dissertation by Zheng Sun is accepted in its present form
by the Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Chi-Wang Shu, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Johnny Guzmán, Ph.D., Reader

Date_____ _____

Jérôme Darbon, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Andrew G. Campbell, Dean of the Graduate School

Vita

Education

Brown University, Providence, RI, U.S.A..

Ph.D. in Applied Mathematics (expected in May 2018).
M.Sc. in Applied Mathematics, 2015.
Advisor: Chi-Wang Shu.

University of Science and Technology of China, Hefei, Anhui, P.R. China.

B.Sc. in Mathematics and Applied Mathematics, School of Gifted Young, 2014.
Advisor: Falai Chen.

Professional Experience

Oak Ridge National Laboratory, Oak Ridge, TN, U.S.A..

Intern at Computer Science and Mathematics Division, summer 2017.
Mentor: Cory D. Hauck.

Professional Service

Referee for *Journal of Scientific Computing*.

Publications

1. Z. Sun and C.-W. Shu. Stability analysis and error estimates of Lax–Wendroff discontinuous Galerkin methods for linear conservation laws. *ESAIM: Mathematical Modelling and Numerical Analysis*, 51(3):1063–1087, 2017.

2. Z. Sun and C.-W. Shu. Stability of the fourth order Runge–Kutta method for time-dependent partial differential equations. *Annals of Mathematical Sciences and Applications*, 2(2):255–284, 2017.

3. Z. Sun, J. A. Carrillo, and C.-W. Shu. A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials. *Journal of Computational Physics*, 352:76–104, 2018.

Teaching Experience

Teaching Assistant.

Introduction to scientific computing, Brown University, spring 2016.
Computational probability and statistics, Brown University, fall 2015.
Linear algebra, University of Science and Technology of China, spring 2014.

Awards & Honors

Selectee of the NSF Mathematical Sciences Graduate Internship Program, 2017.
China National Scholarship, 2011, 2012 and 2013.
Mathematical Contest in Modeling, Meritorious Winner, 2012.

Acknowledgments

Firstly and most heartily, I would like to thank my advisor, Professor Chi-Wang Shu, who has taught me, guided me, supported me and encouraged me during my four years at Brown. No matter how busy he is, he always puts his work aside to discuss with me when I stop by, listening to my complaints, providing the most helpful suggestions and sharing my happiness for any incremental progress. He is knowledgable, responsive, modest and open-minded. He acts as my perfect role model as a scholar. It is my great fortune to be his student.

Secondly, I am indebted to Professor Johnny Guzmán and Professor Jérôme Darbon, who have kindly agreed to spend their valuable time reading my thesis and serving in my committee. Both professors have also taught me in class. I admire their insights into the topics and benefit a lot from their instruction.

Thirdly, I would like to express my gratitude to Professor José Carrillo from the Imperial College London and Doctor Cory Hauck from the Oak Ridge National Laboratory (ORNL). It is Prof. Carrillo brought the gradient flow problems into our attention. Our paper and ongoing work (with Prof. Shu) form a considerable part of this thesis. I deeply appreciate his contributions. Dr. Hauck served as my mentor during my internship at ORNL. He assigned me an interesting research project on low memory DG methods, and introduced to me topics on asymptotic-preserving methods and kinetic equations. It is my very valuable experience to work with these excellent scholars.

Then I would like to thank my friends. Among them, my deepest gratitude is given to my former roommate and officemate Tong Qin and his wife Yang Chen. They truly care about me and are always ready to offer me a helping hand. I sincerely appreciate their kindness. I am also grateful to my other roommates Zhizhong Chen and Weifeng Sun. Together we have created so many valuable memories. I want to thank my friends Yiran Lu and Guangyao Zhou as well for their warm invitations to various group activities, my officemates Juntao Huang, Zongyuan Li and Ben Whitney for many helpful discussions and kind reminders, and Zheng Chen and Chongze Hu for their tremendous help during my visit at ORNL. Finally, I would like to thank all my group members, especially Tianheng Chen, Guosheng Fu, Dan Ling, Will Pazner, Cengke Shi, Kunrui Wang and Xinyue Yu, for their inspiration and encouragement over years.

Last but definitely not the least, I keep my special gratitude to my family. I thank my aunt Ning Sun and my uncle Ran Li, as well as my lovely cousins, for providing me with a harbor to anchor oversea. I finish by thanking my parents Juwei Sun and Ritong Tian, who always stand behind me, sorrowing for my sorrow and delighted for my delight, with an exclamation mark: Thank you for your unconditional love!

Abstract of "Discontinuous Galerkin Methods: Stability of Time Discretizations and Applications to Gradient Flows", by Zheng Sun, Ph.D., Brown University, May 2018

This thesis consists of two topics on discontinuous Galerkin (DG) finite element methods for time-dependent problems.

In the first part of the thesis, we analyze the stability of explicit time stepping methods with the DG spatial discretization for linear conservation laws. For the second-order and the third-order schemes, Lax-Wendroff DG methods are analyzed. We firstly note that a class of Lax-Wendroff DG methods is equivalent with the Runge-Kutta DG methods after one full time step. It is then shown that the stability of the fully discretized schemes depends on the choice of numerical fluxes of intermediate variables in time discretizations. Finally, for stable schemes in one dimension, optimal error estimates are given for both the solution and its first-order time derivative. For the fourth-order method, we focus on the Runge-Kutta time discretization. It is shown that the numerical scheme is two-step strongly stable under an appropriate time step constraint. This result applies not only to DG methods, but also to other spatial discretizations that preserve the semi-negativity of the continuum equation.

In the second part, we aim at designing an entropy stable high-order DG method for gradient flow problems. The positivity (non-negativity) of the solution is also expected in various applications. It plays a crucial role for the well-definedness and the dissipative property of the entropy functional. By adopting the Gauss-Lobatto quadrature rule in the local DG scheme and with a suitable choice of numerical fluxes, our semi-discrete scheme is entropy stable for scalar problems without non-smooth interaction kernels. The scheme also features with weak positivity with the Euler forward time discretization. Hence by applying a scaling limiter and with the strong-stability-preserving Runge-Kutta method, the scheme is able to produce non-

negative solutions. The method can also be applied to two-dimensional problems on Cartesian meshes. Numerical examples are given to confirm the high-order accuracy for smooth test cases and to demonstrate the effectiveness for preserving long time asymptotics. Finally, the extension to cross-diffusion gradient flow systems is also discussed.

Contents

List of Tables

List of Figures

INTRODUCTION

The discontinuous Galerkin (DG) methods, a class of finite element methods with discontinuous test and trial function spaces, were pioneered by Reed and Hill in [64] and Lesaint and Raviart in [53] to solve the transport equation in 1970s. Then in 1980s and 1990s, Cockburn et al. developed a general frame work in [29, 28, 27, 26, 31] on using the DG discretization for solving nonlinear hyperbolic conservation laws, by adopting monotone numerical fluxes and combining with the total variation diminishing or total variation bounded Runge-Kutta schemes. Later on, based on successful numerical experiments of Bassi and Rebay in [5], Cockburn and Shu proposed a local DG method [30] for dealing with diffusion terms in convection dominated problems, in which they introduced auxiliary variables to transform the original high-order equation into a first-order system. The development of DG methods for purely elliptic or parabolic problems was independent of the hyperbolic problems and was also initiated in 1970s. Different variants were proposed during the decades and were unified in [4]. Until now, the DG methods have received intensive attention and been widely applied to areas of computational fluid dynamics, semi-conductor device design, computational cosmology and so on. DG methods are still active research areas these days and new variants are kept being proposed. See [23], [58] and [25] for some of the examples.

The success of the DG methods is closely related with various desirable properties that the methods possess. The use of numerical fluxes and limiters, which have been studied intensively in the context of the first-order monotone schemes and the finite volume schemes, ensures the local conservation, suppresses oscillations and effectively captures the shocks. At the same time, the finite element nature allows the methods to handle complex geometry and achieve high formal order of accuracy. The methods also benefit from the compact data structure hence feature with good h-p adaptivity and parallel efficiency. Moreover, the methods also show their flexibility

on preserving structures of the continuum equations, such as the preservation of asymptotics [2, 39], the maximum principle or the invariant bound [88, 89], the well-balanceness at the steady states [77, 78] and the entropy stability [22]. This feature allows the methods to provide accurate results under limited computational resources and be more robust for long time simulations.

When exploiting these benefits in different applications, the key principle for designing usable DG schemes is to ensure the stability. Besides the stability enforced by limiters (total variation and L^1 stability, for example), the semi-discrete DG schemes themselves are usually stable in certain sense. It was shown by Jiang and Shu in [48] that the DG scheme subjects to a square entropy inequality and is L^2 stable for nonlinear hyperbolic conservation laws. This L^2 stability remains true after coupling with implicit time stepping methods, including the Euler backward method and the Crank-Nicholson method. The L^2 stability of local DG methods for time-dependent diffusive and dispersive types of equations were studied extensively around 2000s. We refer to [79] and references therein for further details. For problems governed by special decaying functionals, the L^2 energy may no longer be a good measure to capture the structure of the system. Instead, the dissipation of these special functionals are expected. For systems with high-order spatial derivatives, this can usually be achieved by applying the local DG methods. See [76] for an example for the Cahn-Hilliard equation. For hyperbolic conservation laws, recently, based on the methodology established by Carpenter et al. in [13] and Gassner in [36], Chen and Shu developed a unified framework for designing high-order DG methods which will satisfy entropy inequalities for any given single convex entropy. This thesis is a further exploration of stable DG methods. It is mainly based on our publications [71, 72, 70] and consists of the following two parts.

1. The L^2 stability of DG methods after applying explicit time discretizations.

2. The design of an entropy stable DG scheme for gradient flow problems.

In the first part of the thesis, we focus on linear conservation laws to study the stability of fully discretized DG schemes. The early analyses of fully discretized DG schemes were based on von Neumann analysis, which is limited to uniform meshes and is only necessary but not sufficient. In [84], Zhang and Shu used an energy approach to prove the stability of the DG methods coupling with the second-order and the third-order Runge-Kutta time integrators for scalar conservation laws. The corresponding error analysis and extension to symmetrizable system can be found in [82, 83, 85, 60]. In Chapter 1, we adopt similar energy argument to study the stability of the second-order and the third-order Lax-Wendroff DG methods for linear problems. The Lax-Wendroff DG method we analyze was proposed by Guo et al. in [40], which utilizes the local DG method to discretize all spatial derivatives introduced in the Lax-Wendroff procedure. According to our analysis, the stability and the type of stability of the Lax-Wendroff DG methods depend on the choice of numerical fluxes for auxiliary variables. Furthermore, when all numerical fluxes are upwinding, the Lax-Wendroff DG methods are actually equivalent to the corresponding Runge-Kutta DG methods. The next attempt is to push the envelope to analyze higher-order time discretization. In Chapter 2, we analyze the fourth-order Runge-Kutta time discretization. Instead of focusing on the DG spatial discretization, we consider a general semi-negative ordinary differential equation system, which is typically obtained after applying a stable spatial discretization. We find a counter example that the four stage fourth-order Runge-Kutta method fails to preserve the strong stability in one step. But the eight stage method obtained by doubling the four stage method is proved to be strongly stable.

The second part of the thesis is dedicated to the development of entropy stable DG methods for gradient flow problems. Many classical mathematical physics equa-

tions can be formulated as gradient flows and this structure widely appears in the study of granular flows and aggregation models. Under the simplest scenario, the problem takes form of a continuity equation and the velocity field is determined by the gradient of a potential function. The problem is associated with a decaying entropy functional. The difficulty of the problem is that, in many real life applications, the solution corresponds to density of species or water height, which is expected to be non-negative. Furthermore, the entropy is well-defined and decreases with respect to time only if the solution admits non-negative values. Hence preservation of non-negativity is an essential part of the entropy stability. This forbid us from directly applying the local DG methods. Instead, we apply the quadrature-based DG method as that in [22] to enforce entropy stability. The resulting scheme is also compatible with the high-order positivity-preserving procedure, developed by Zhang and Shu in [87] and then extended for second-order PDEs by Zhang et al. in [86, 69]. The numerical methods for the scalar problems are detailed in Chapter 3. The application to cross-diffusion gradient flow systems is then discussed in Chapter 4.

The rest of the thesis is organized as follows. In Chapter 1, we analyze the stability of the second-order and the third-order Lax-Wendroff DG methods. The optimal error estimates for the unknown and its first-order time derivative are also derived as a corollary of the stability results. In Chapter 2, we prove the two-step strong stability of the fourth-order Runge-Kutta method. In Chapter 3 and Chapter 4, a DG method is introduced for scalar gradient flows and cross-diffusion gradient flow systems respectively. Finally conclusions and future work are given at the very end. To avoid confusion, we remark that all notations are defined locally within each chapter.

Part I

STABILITY OF TIME DISCRETIZATIONS

Chapter One

The second-order and the third-order Lax-Wendroff time discretizations

1.1 Introduction

In this chapter, we analyze the stability of the second-order and the third-order Lax-Wendroff DG (LWDG) methods for solving linear conservation laws. The Runge-Kutta DG (RKDG) methods are covered as special cases. We concentrate our attention in the scalar problem, although the analysis can be easily generalized to one-dimensional linear hyperbolic systems and multi-dimensional symmetric linear systems. To be more specific, we are interested in the following initial value problem

$$\begin{cases} u_t = \nabla \cdot (\boldsymbol{\beta} u), & (\boldsymbol{x}, t) \in \Omega \times (0, T), \\ u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), & \boldsymbol{x} \in \Omega. \end{cases} \tag{1.1}$$

For simplicity, periodic boundary conditions are assumed, but our analysis does not depend on periodicity and can be extended to other types of boundary conditions. Here, $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{x})$ is a vector-valued function with the divergence-free condition $\nabla \cdot \boldsymbol{\beta}(\boldsymbol{x}) = 0$. In this setting, the L^2 energy of the solution to (1.1) is conserved, which facilitates the study of LWDG schemes using energy methods. Note that the divergence-free condition forces the coefficient in (1.1) to be constant in the one-dimensional case, but it allows variable coefficients in multidimensions.

The most widely used time integrators for DG schemes on conservation laws are the strong-stability-preserving Runge-Kutta (SSP-RK) methods [38]. As an alternative, one can also choose the Lax-Wendroff type time discretization, which relies on converting time derivatives in a truncated Taylor expansion into spatial derivatives by using the differential equation repeatedly, resulting in LWDG methods. The first LWDG method for solving hyperbolic conservation laws was proposed by Qiu et al. in [63]. In their schemes, the high-order derivatives are approximated by directly

differentiating the numerical polynomials. In [40], Guo et al. pointed out that this method does not exhibit the superconvergence property, so they developed a new LWDG method by applying the LDG techniques. More specifically, they introduced new auxiliary variables and used DG spatial discretization to approximate the spatial derivatives converted from time derivatives in the Lax-Wendroff procedure. We focus on the LWDG method in [40] and study its properties for solving linear scalar conservation laws.

Firstly, let us remark that there exist close relationships between the LWDG method and the RKDG method. Both of these methods introduce auxiliary or intermediate variables. In the RKDG method, the stage variables correspond to the solution u at different internal time stages between two time levels; while in the LWDG method, the intermediate variables approximate time derivatives of u. In general, these two sets of variables do not contain the same information. But for linear conservation laws not explicitly depending on time (including multi-dimensional systems), if we use the same fluxes throughout the LWDG scheme, then it will be equivalent to the RKDG method after one full time step. More specifically, in this case, the approximations of time derivatives in a LWDG scheme and the stage variables in a RKDG scheme can be expressed as a linear combination of each other, and they lead to the same numerical solution at the next time level. More details are given in Appendix A.1. Therefore, there are strong connections between our analysis and the work by Zhang and Shu in [82, 85, 84], where they have provided the stability analysis and a prior error estimates for the second and third order RKDG schemes. We will comment on these connections later in the chapter.

Next, we move on to the stability analysis. In the LWDG schemes in [40], Guo et al. applied alternating fluxes as that in [30, 80]. However, it can be shown that the LWDG schemes based on these alternating fluxes cannot preserve strong stability

under the standard CFL condition $\tau \leq \lambda h$, where τ and h are the time step and the maximum element length respectively and $\lambda > 0$ is a constant. This reminds us of the well-known fact that the choice of numerical fluxes has an essential influence on the types of stability. We assume the numerical fluxes can be either upwind or downwind for each variable. After a detailed analysis, we find that, if uniform or non-increasing time steps are used, for the second-order schemes (LW2DG) with piecewise linear elements, and the third-order schemes (LW3DG) with arbitrary $\mathcal{P}_k$ elements, as long as we apply the upwind flux for u, then the schemes will be stable in the L^2 norm. Furthermore, if we also use the upwind flux for the second variable p (which approximates u_t), then the schemes will be strongly stable, that is $\|u_h^{n+1}\| \leq \|u_h^n\|$. Notice that, for the linear cases, the RKDG method belongs to this class. However, if we use the downwind flux for p (e.g. according to the choice of alternating fluxes), then the scheme is only stable in a weaker sense, namely $\|u_h^n\| \leq C\|u_h^0\|$, where C is a constant which depends on the CFL number and the inverse estimate constants, but is independent of the total time T. Therefore, with both choices of the fluxes, the energy of the solution is well-controlled after long time integration.

Finally, as a corollary of the stability analysis, we perform error estimates under the same framework of the stability analysis in one dimension. We highlight that optimal error estimates of both u_h and p_h can be obtained, where u_h and p_h are numerical approximations to u and u_t respectively. To be more precise, assuming u is a smooth solution to the one-dimensional problem, under a proper CFL condition $\tau \leq \lambda h$, we have both $\|u - u_h\|_{L^2} \leq C(\tau^2 + h^2)$ and $\|u_t - p_h\| \leq C(\tau^2 + h^2)$ for the LW2DG schemes with piecewise linear elements; and $\|u - u_h\|_{L^2} \leq \mathcal{O}(\tau^3 + h^{k+1})$ and $\|u_t - p_h\| \leq \mathcal{O}(\tau^3 + h^{k+1})$ for the LW3DG schemes with $\mathcal{P}_k$ elements. The error estimates hold for both choices of the numerical fluxes.

The organization of this chapter is as follows. In Section 1.2 , the notations and

preliminaries are introduced for the one-dimensional analysis. In Section 1.3 , we present the stability analysis of the LW2DG schemes and the LW3DG schemes for one-dimensional advection equations, as well as the extension to multidimensions with divergence-free coefficients. In Section 1.4, we state the error estimates for one-dimensional problems and give a detailed proof only for the second-order schemes. In Section 1.5, numerical examples are given to validate our error estimates.

1.2 Notations and preliminaries

1.2.1 Model problem and notations

We study the following model problem in our one-dimensional analysis

$$\begin{cases} u_t = u_x, & (x,t) \in (0,2\pi) \times (0,T), \\ u(x,0) = u_0(x), & x \in (0,2\pi). \end{cases} \tag{1.2}$$

with periodic boundary conditions.

Let $I_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$ and $[0,2\pi] = \cup_{i=1}^{N} I_i$ be a regular partition of the domain. We denote the cell length by $h_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, and define $h = \max_i h_i$. We use the notation $(w,v)_i = \int_{I_i} wv \, dx$ for the L^2 inner product on I_i, and define $(w,v) = \sum_{i=1}^{N} (w,v)_i$. Without specification, the notation $\|\cdot\|$ refers to the L^2 norm.

The associated discontinuous finite element space is defined as

$$V_h = \{v \in L^2(\Omega) : v|_{I_i} \in \mathcal{P}_k(I_i), \forall j\},$$

where $\mathcal{P}_k(I_i)$ is the space of polynomials on I_i of degree at most k. Note that functions in V_h can be double-valued at cell interfaces, so we use v^+ and v^- to represent the right and left limits of v respectively. The jump is denoted by $[v] = v^+ - v^-$. We define the jump semi-norm on V_h as $[\![v]\!] = \sqrt{\sum_{i=1}^{N} [v]^2_{i+\frac{1}{2}}}$, and the associated bilinear form as $[w, v] = \sum_{i=1}^{N} [w]_{i+\frac{1}{2}} [v]_{i+\frac{1}{2}}$.

One has the inverse estimate for polynomials, which states $\forall v \in \mathcal{P}_k(I_i)$,

$$h\|v_x\|_{I_i} + h^{\frac{1}{2}}\|v\|_{\partial I_i} \leq \mu \|v\|_{I_i},$$

where $\|v\|_{\partial I_i} = \sqrt{(v^+_{i-\frac{1}{2}})^2 + (v^-_{i+\frac{1}{2}})^2}$ and μ is the inverse estimate constant, which is independent of h. We denote a constant independent of h but depending on μ by C_μ. Similarly, we may use $C_{\mu,\lambda}$, $C_{\mu,\lambda,\boldsymbol{\beta}}$ and so on. The subscripts indicate the dependency of the constant C on these parameters.

1.2.2 The LWDG method

For simplicity, we assume the time step τ is a constant for different time levels throughout this chapter. The strong stability results however do not need this restriction, while the weaker stability results (the energy is bounded but not necessarily non-increasing) do need the restriction that the time steps are non-increasing, for which the constant time step is a special case.

For the second-order LWDG method, we first approximate the original equation

(1.2) with the following first-order system

$$u^{n+1} = u^n + \tau u_x^n + \frac{\tau^2}{2} p_x^n,$$
$$p^n = u_x^n.$$

Then the fully discretized scheme can be obtained by replacing the spatial derivatives by their DG discretizations. For convenience, we introduce the linear operators $\mathcal{H}^+$ and $\mathcal{H}^-$,

$$\mathcal{H}^{\pm}(w,v) = \sum_{i=1}^{N} \mathcal{H}_i^{\pm}(w,v) = \sum_{i=1}^{N} \big((w, v_x)_i - w_{i+\frac{1}{2}}^{\pm} v_{i+\frac{1}{2}}^{-} + w_{i-\frac{1}{2}}^{\pm} v_{i-\frac{1}{2}}^{+} \big).$$

As a convention, $\mathcal{H}^{\pm 1} = \mathcal{H}^{\pm}$. Then, the schemes will take the form

$$(u_h^{n+1}, \varphi_h) = (u_h^n, \varphi_h) - \tau \mathcal{H}^{\alpha_u}(u_h^n, \varphi_h) - \frac{\tau^2}{2} \mathcal{H}^{\alpha_p}(p_h^n, \varphi_h),$$
$$(p_h^n, \psi_h) = -\mathcal{H}^{\widehat{\alpha_u}}(p_h^n, \psi_h),$$

where $\varphi_h, \psi_h \in V_h$, and α_u, α_p and $\widehat{\alpha_u}$ are equal to $+1$ or -1, corresponding to different numerical fluxes, to be specified later.

In our analysis, we consider different choices of the numerical fluxes, and require $\alpha_u = \widehat{\alpha_u}$. When $\alpha_u = \widehat{\alpha_u} = +1$ and $\alpha_p = -1$, we recover the alternating fluxes used in [40]; when $\alpha_u = \widehat{\alpha_u} = \alpha_p = +1$, we can easily verify that this is actually the classical second-order RKDG scheme. The only superficial difference is that, for the RKDG scheme, one introduces the stage variable $u_h^{(1)} = u_h^n + \tau p_h^n$ instead of p_h^n in our context.

Similarly, one can write down the third-order schemes,

$$
\begin{aligned}
(u_h^{n+1}, \varphi_h) &= (u_h^n, \varphi_h) - \tau \mathcal{H}^{\alpha_u}(u_h^n, \varphi_h) - \frac{\tau^2}{2}\mathcal{H}^{\alpha_p}(p_h^n, \varphi_h) - \frac{\tau^3}{6}\mathcal{H}^{\alpha_q}(q_h^n, \varphi_h), \\
(p_h^n, \psi_h) &= -\mathcal{H}^{\widehat{\alpha_u}}(u_h^n, \psi_h), \\
(q_h^n, \eta_h) &= -\mathcal{H}^{\widehat{\alpha_p}}(p_h^n, \eta_h),
\end{aligned}
$$

where φ_h, ψ_h and η_h are test functions from V_h, and α_u, α_p, $\widehat{\alpha_u}$ and $\widehat{\alpha_p}$ are equal to $+1$ or -1, corresponding to different numerical fluxes. We would again require $\alpha_u = \widehat{\alpha_u}$ and $\alpha_p = \widehat{\alpha_p}$. When $\alpha_u = +1$ and $\alpha_p = \alpha_q = -1$, we restore the alternating fluxes in [40]; while if $\alpha_u = \alpha_p = \alpha_q = +1$, we can easily verify that this is actually the classical third-order RKDG scheme, where the stage variables are $u_h^{(1)} = u_h^n + \tau p_h^n$ and $u_h^{(2)} = u_h^n + \frac{\tau}{2} p_h^n + \frac{\tau^2}{4} q_h^n$.

We are interested in the stability of these schemes. We call a scheme to be stable, if there is a constant C independent of τ and h, but may depend on the total time $T = n\tau$, such that $\|u_h^n\| \leq C\|u_h^0\|$. We say a scheme is strongly stable, if $\|u_h^{n+1}\| \leq \|u_h^n\|$ for any n. As a corollary, we have $C = 1$ for a strongly stable scheme in $\|u_h^n\| \leq C\|u_h^0\|$.

1.2.3 Properties of $\mathcal{H}$

In this subsection, we familiarize the readers with the properties of the operator pair $\mathcal{H}^+$ and $\mathcal{H}^-$, which are fundamental for the analyses later.

The first two lemmas focus on the operators themselves. Lemma 1.2.1 describes the relationships between $\mathcal{H}^+$ and $\mathcal{H}^-$. It states the anti-symmetry, the semi-definiteness and the difference of the operator pair. Lemma 1.2.2 estimates the

L^2 operator norms of $\mathcal{H}^{\pm}(w,\cdot)$ and $\mathcal{H}^{\pm}(\cdot,v)$ on V_h. These lemmas are classical and can be proved straightforwardly, we refer interested readers to, e.g. [81].

Lemma 1.2.1. *Suppose $w, v \in V_h$, then*

$$\begin{aligned}
\mathcal{H}^{-}(w,w) &= -\frac{1}{2}[\![w]\!]^2, \\
\mathcal{H}^{-}(w,v) &= -\mathcal{H}^{+}(v,w), \\
\mathcal{H}^{+}(w,v) &= \mathcal{H}^{-}(w,v) + [w,v].
\end{aligned}$$

Lemma 1.2.2. *Suppose $w, v \in V_h$, then*

$$\begin{aligned}
|\mathcal{H}^{\pm}(w,v)| &\leq (\|w_x\| + C_\mu h^{-\frac{1}{2}}[\![w]\!])\|v\|, \\
|\mathcal{H}^{\pm}(w,v)| &\leq (\|v_x\| + C_\mu h^{-\frac{1}{2}}[\![v]\!])\|w\|.
\end{aligned}$$

The next two lemmas describe the numerical spatial derivatives from *DG* and *LDG* discretizations. Lemma 1.2.3 gives a crude bound for the first-order derivative, and it follows directly from Lemma 1.2.2 and the inverse estimate. Lemma 1.2.4 establishes the connections among different orders of derivatives, which is essentially the discretized version of integration by parts. To cover different choices of fluxes, we introduce undetermined parameters α's in this lemma, which can be either $+1$ or -1. The detailed proof is omitted, as it just amounts to repeatedly using Lemma 1.2.1.

Lemma 1.2.3. *Let $u, p \in V_h$. For any test functions $\psi \in V_h$, $(p,\psi) = -\mathcal{H}^{\alpha_u}(u,\psi)$. Then*

$$\|p\| \leq C_\mu h^{-1}\|u\|.$$

Lemma 1.2.4. *Let $u, p, q, r \in V_h$, such that for any test functions $\psi, \eta, \zeta \in V_h$,*

$$
\begin{aligned}
(p, \psi) &= -\mathcal{H}^{\alpha_u}(u, \psi), \\
(q, \eta) &= -\mathcal{H}^{\alpha_p}(p, \eta), \\
(r, \zeta) &= -\mathcal{H}^{\alpha_q}(q, \zeta),
\end{aligned}
$$

where $\alpha_u, \alpha_p, \alpha_q = -1, +1$. Then we have,

(i)

$$
\begin{aligned}
(p, u) &= -\frac{\alpha_u}{2}[\![u]\!]^2, \\
(q, u) &= -\|p\|^2 - \frac{\alpha_{up}}{2}[u, p], \\
(r, u) &= \frac{\alpha_p}{2}[\![p]\!]^2 - \frac{\alpha_{uq}}{2}[u, q],
\end{aligned}
$$

(ii)

$$
\begin{aligned}
(q, p) &= -\frac{\alpha_p}{2}[\![p]\!]^2, \\
(r, p) &= -\|q\|^2 - \frac{\alpha_{pq}}{2}[p, q],
\end{aligned}
$$

where $\alpha_{wv} = \alpha_w + \alpha_v$, $w, v = u, p, q$.

1.3 Stability analysis

1.3.1 Second-order schemes

The LW2DG schemes can be rewritten as follows. Given u_h^n, find $u_h^{n+1} \in V_h$, such that

$$(u_h^{n+1}, \varphi_h) = (u_h^n, \varphi_h) + \tau(p_h^n, \varphi_h) + \frac{\tau^2}{2}(q_h^n, \varphi_h), \tag{1.3a}$$

$$(p_h^n, \psi_h) = -\mathcal{H}^{\alpha_u}(u_h^n, \psi_h), \tag{1.3b}$$

$$(q_h^n, \eta_h) = -\mathcal{H}^{\alpha_p}(p_h^n, \eta_h), \tag{1.3c}$$

for any $\varphi_h, \psi_h, \eta_h \in V_h$.

Note that (1.3a) implies

$$u_h^{n+1} = u_h^n + \tau p_h^n + \frac{\tau^2}{2} q_h^n. \tag{1.4}$$

In our analysis, we restrict the finite element space to be piecewise linear. This is because when one uses upwind fluxes for both u_h^n and p_h^n, the scheme is equivalent to the second-order RKDG scheme, which, according to the von Neumann analysis in [32], is unstable with high-order ($k > 1$) piecewise polynomial spaces when τ/h is a constant.

The key ingredient for the stability analysis is to use the specialty of piecewise linear elements. In this case, the L^2 norm of $(p_h^n)_x$ can be bounded by the jump of u_h^n, which is not necessarily true in general. Zhang and Shu first used this technique in

[84] to analyze the second-order RKDG scheme, but for a different auxiliary variable.

Here is the outline for the stability analysis of (1.3). We first prove Lemma 1.3.1 to connect $\|(p_h^n)_x\|$ with $[\![u_h^n]\!]$. Then, with this lemma, precise estimates for $\|u_h^{n+1}-u_h^n-\tau p_h^n\|$ and $\|p_h^{n+1}-p_h^n\|$ can be obtained, which are stated in Lemma 1.3.2 and Lemma 1.3.3 respectively. Finally, we carry out the proof of Theorem 1.3.1 to establish the stability of the LW2DG schemes. The first part of this theorem, which essentially rebuilds the result by Zhang and Shu in [84], only uses Lemma 1.3.2. For the second part with alternating fluxes, Lemma 1.3.3 is used as well.

Lemma 1.3.1. *With $\mathcal{P}_1$ elements, $\|(p_h^n)_x\|_i = 6h_i^{-\frac{3}{2}}|[u_h^n]_{i+\frac{\alpha_u}{2}}|$.*

Proof. We only prove the lemma for $\alpha_u = +1$, the case $\alpha_u = -1$ follows along the same line.

For simplicity, we drop all the subscripts h and superscripts n. Let $\{\phi^0, \phi^1\}$ be the normalized Legendre polynomials basis on I_i, and ϕ^i is of degree i. Suppose $p = p_0\phi^0 + p_1\phi^1$ on I_i. Then we have $p_x = 2\sqrt{3}h^{-\frac{3}{2}}p_1$, where p_1 can be obtained through (1.3b),

$$p_1 = (p, \phi^1)_i = -(u, \phi_x^1)_i + u^+_{i+\frac{1}{2}}(\phi^1)^-_{i+\frac{1}{2}} - u^+_{i-\frac{1}{2}}(\phi^1)^+_{i-\frac{1}{2}}.$$

With integration by parts, we get

$$p_1 = (u_x, \phi^1)_i - u^-_{i+\frac{1}{2}}(\phi^1)^-_{i+\frac{1}{2}} + u^+_{i+\frac{1}{2}}(\phi^1)^-_{i+\frac{1}{2}} = [u]_{i+\frac{1}{2}}(\phi^1)^-_{i+\frac{1}{2}}, \tag{1.5}$$

where the second equality holds due to the fact that u_x is a constant. Note that $(\phi^1_{i+\frac{1}{2}})^- = \sqrt{3}h_i^{-\frac{1}{2}}$. Now (1.5) implies $p_1 = \sqrt{3}h_i^{-\frac{1}{2}}[u]_{i+\frac{1}{2}}$. Therefore $p_x = 6h_i^{-2}[u]_{i+\frac{1}{2}}$ and $\|p_x\|_i = 6h_i^{-\frac{3}{2}}|[u]_{i+\frac{1}{2}}|$. □

The following statements hold based on Lemma 1.3.1.

Lemma 1.3.2.

$$\|u_h^{n+1} - u_h^n - \tau p_h^n\|^2 \leq C_\mu(\tau\lambda^3[\![u_h^n]\!]^2 + \tau^3\lambda[\![p_h^n]\!]^2).$$

Proof. Rearranging (1.3) gives

$$(u_h^{n+1} - u_h^n - \tau p_h^n, \varphi_h) = -\frac{\tau^2}{2}\mathcal{H}^{\alpha_p}(p_h^n, \varphi_h).$$

Exploiting Lemma 1.2.2, we obtain

$$(u_h^{n+1} - u_h^n - \tau p_h^n, \varphi_h) \leq \frac{\tau^2}{2}\big(\|(p_h^n)_x\| + C_\mu h^{-\frac{1}{2}}[\![p_h^n]\!]\big)\|\varphi_h\|.$$

Then the proof is completed by applying Lemma 1.3.1 and setting $\varphi_h = u_h^{n+1} - u_h^n - \tau p_h^n$. □

Lemma 1.3.3.

$$\|p_h^{n+1} - p_h^n\|^2 \leq C_\mu(h^{-1}\lambda^2[\![u_h^n]\!]^2 + \tau\lambda[\![p_h^n]\!]^2 + \tau^2\lambda^2\|q_h^n\|^2).$$

Proof. According to (1.4), the following relationship holds

$$(p_h^{n+1} - p_h^n, \psi_h) = -\mathcal{H}^{\alpha_u}(u_h^{n+1} - u_h^n, \psi_h) = -\mathcal{H}^{\alpha_u}(\tau p_h^n + \frac{\tau^2}{2}q_h^n, \psi_h). \tag{1.6}$$

Then we apply Lemma 1.2.2 and Lemma 1.2.3 to obtain

$$(p_h^{n+1} - p_h^n, \psi_h) \leq C_\mu(\tau\|(p_h^n)_x\| + \tau h^{-\frac{1}{2}}[\![p_h^n]\!]^2 + \tau\lambda\|q_h^n\|)\|\psi_h\|.$$

The proof is completed by applying Lemma 1.3.1 and plugging in $\psi_h = p_h^{n+1} - p_h^n$. □

We are now ready to state our main theorem for the stability of the LW2DG schemes.

Theorem 1.3.1 (Stability of the LW2DG schemes)**.** *There exists a constant λ, such that when $\tau \le \lambda h$, the numerical solution of schemes (1.3) satisfies,*

(i) $\|u_h^{n+1}\| \le \|u_h^n\|$, if $\alpha_u = \alpha_p = +1$,

(ii) $\|u_h^n\| \le C_{\mu,\lambda}\|u_h^0\|$, if $\alpha_u = +1$ and $\alpha_p = -1$,

where $C_{\mu,\lambda}$ is a constant depending on λ and the inverse estimate constant μ, but is independent of the total time $T = n\tau$.

Proof. Take $\varphi_h = u_h^n$ in (1.3a). Using Lemma 1.2.4 we get

$$\frac{1}{2}\|u_h^{n+1}\|^2 - \frac{1}{2}\|u_h^n\|^2 + \frac{\tau\alpha_u}{2}[\![u_h^n]\!]^2 + \frac{\tau^2\alpha_{up}}{2}[u_h^n, p_h^n] = \frac{1}{2}\|u_h^{n+1} - u_h^n\|^2 - \frac{\tau^2}{2}\|p_h^n\|^2. \quad (1.7)$$

Note that

$$\|u_h^{n+1} - u_h^n\|^2 - \tau^2\|p_h^n\|^2 \quad = \quad \|u_h^{n+1} - u_h^n - \tau p_h^n\|^2 + 2\tau(u_h^{n+1} - u_h^n - \tau p_h^n, p_h^n),$$

where $(u_h^{n+1} - u_h^n - \tau p_h^n, p_h^n) = \frac{\tau^2}{2}(q_h^n, p_h^n) = -\frac{\tau^2\alpha_p}{4}[\![p_h^n]\!]^2$. Hence

$$\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \tau\alpha_u[\![u_h^n]\!]^2 + \tau^2\alpha_{up}[u_h^n, p_h^n] + \frac{\tau^3\alpha_p}{2}[\![p_h^n]\!]^2 = \|u_h^{n+1} - u_h^n - \tau p_h^n\|^2. \quad (1.8)$$

Exploiting Lemma 1.3.2, we have

$$\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \tau(\alpha_u - C_\mu\lambda^3)[\![u_h^n]\!]^2 + \tau^2\alpha_{up}[u_h^n, p_h^n] + \tau^3(\frac{\alpha_p}{2} - C_\mu\lambda)[\![p_h^n]\!]^2 \le 0. \quad (1.9)$$

(i) When $\alpha_u = \alpha_p = +1$, (1.9) can be rewritten as

$$\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \tau(\frac{1}{4} - C_\mu\lambda^3)[\![u_h^n]\!]^2 + \tau^3(\frac{1}{6} - C_\mu\lambda)[\![p_h^n]\!]^2 + \tau[\![\frac{\sqrt{3}}{2}u_h^n + \tau\frac{\sqrt{3}}{3}p_h^n]\!]^2 \le 0,$$

which implies $\|u_h^{n+1}\| \le \|u_h^n\|$ when λ is sufficiently small.

(ii) When $\alpha_u = +1$ and $\alpha_p = -1$, (1.9) gives

$$\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \tau(1 - C_\mu\lambda^3)[\![u_h^n]\!]^2 - \tau^3(\frac{1}{2} + C_\mu\lambda)[\![p_h^n]\!]^2 \leq 0. \tag{1.10}$$

The coefficient of $[\![p_h^n]\!]^2$ is negative, and one needs to bring in new terms to balance it. To this end, we plug in $\psi_h = p_h^n$ in (1.6) to get

$$\frac{1}{2}\|p_h^{n+1}\|^2 - \frac{1}{2}\|p_h^n\|^2 - \frac{1}{2}\|p_h^{n+1} - p_h^n\|^2 = -\frac{\tau}{2}[\![p_h^n]\!]^2 - \frac{\tau^2}{2}\|q_h^n\|^2.$$

By Lemma 1.3.3, this gives

$$\|p_h^{n+1}\|^2 - \|p_h^n\|^2 - C_\mu\lambda^2 h^{-1}[\![u_h^n]\!]^2 \leq -\tau(1 - C_\mu\lambda)[\![p_h^n]\!]^2 - \tau^2(1 - C_\mu\lambda)\|q_h^n\|^2. \tag{1.11}$$

Multiplying (1.11) with τ^2 and adding it to (1.10), we have

$$\begin{aligned}(\|u_h^{n+1}\|^2 + \tau^2\|p_h^{n+1}\|^2) - (\|u_h^n\|^2 + \tau^2\|p_h^n\|^2) + \tau(1 - C_\mu\lambda^3)[\![u_h^n]\!]^2 \\ +\tau^3(\frac{1}{2} - C_\mu\lambda)[\![p_h^n]\!]^2 + \tau^4(1 - C_\mu\lambda)\|q_h^n\|^2 \leq 0.\end{aligned}$$

When λ is small enough, this gives

$$\|u_h^{n+1}\|^2 + \tau^2\|p_h^{n+1}\|^2 \leq \|u_h^n\|^2 + \tau^2\|p_h^n\|^2.$$

Specially, when uniform time steps are used,

$$\|u_h^n\|^2 + \tau^2\|p_h^n\|^2 \leq \|u_h^0\|^2 + \tau^2\|p_h^0\|^2 \leq (1 + C_\mu\lambda^2)\|u_h^0\|^2.$$

The last inequality holds due to Lemma 1.2.3. This implies the stability result $\|u_h^n\| \leq \sqrt{1 + C_\mu\lambda^2}\|u_h^0\|$. Clearly, the result also holds if the time steps are non-increasing, namely $\tau_{n+1} \leq \tau_n$. □

Remark 1.3.1. *We require that the time steps are non-increasing to ensure the stability in (ii). One should note this is only a sufficient but not necessary condition. Let* $\{\tau_n\}$ *be the sequence of time steps and assume* $0 < \lambda_0 < \frac{\tau_n}{h} \leq \lambda$*. The scheme is still stable in the following cases.*

Case 1: The time steps increase occasionally. Suppose $\tau_{n+1} \leq \tau_n$ *except for* $n = n_0$*, then*

$$\|u_h^{n_0+1}\|^2+\tau_{n_0+1}^2\|p_h^{n_0+1}\|^2 \leq \frac{\tau_{n_0+1}^2}{\tau_{n_0}^2}(\|u_h^{n_0+1}\|^2+\tau_{n_0}^2\|p_h^{n_0+1}\|^2) \leq \frac{\tau_{n_0+1}^2}{\tau_{n_0}^2}(\|u_h^{n_0}\|^2+\tau_{n_0}^2\|p_h^{n_0}\|^2).$$

One can continue the inequality by including the factor $\frac{\tau_{n_0+1}^2}{\tau_{n_0}^2}$*, and finally get*

$$\|u_h^n\|^2 \leq \frac{\tau_{n_0+1}^2}{\tau_{n_0}^2}(1+C_\mu\lambda^2)\|u_h^0\|^2 \leq \frac{\lambda^2}{\lambda_0^2}(1+C_\mu\lambda^2)\|u_h^0\|^2.$$

Similarly, if there is a fixed number m*, such that the time steps increase no more than* m *times, one can show that* $\|u_h^n\|^2 \leq \frac{\lambda^{2m}}{\lambda_0^{2m}}(1+C_\mu\lambda^2)\|u_h^0\|^2$*.*

Case 2: The time steps increase monotonically. Using the argument in case 1, one has

$$\|u_h^n\|^2+\tau_n^2\|p_h^n\|^2 \leq (\prod_{i=1}^{n}\frac{\tau_i^2}{\tau_{i-1}^2})(\|u_h^0\|^2+\tau_0^2\|p_h^0\|^2) = \frac{\tau_n^2}{\tau_0^2}(\|u_h^0\|^2+\tau_0^2\|p_h^0\|^2) \leq \frac{\lambda^2}{\lambda_0^2}(1+C\lambda^2)\|u_h^0\|^2,$$

which is still stable.

However, these arguments can not be used for general time steps $\{\tau_n\}$*. For example,* $\tau_{2n+1} = \lambda h$ *and* $\tau_{2n+2} = \frac{\lambda}{2}h$*. The factor will blow up when* h *goes to* 0*. One would need to make more detailed discussions to analyze the stability.*

In the following sections, we will also assume the non-increasing time steps when

analyzing the stability of the LWDG schemes with alternating fluxes. One should note the argument above will still work, and we will not repeat it.

Remark 1.3.2. *We note that the alternating fluxes in (ii) actually do not preserve the strong stability. To see this, one can plug in a continuous initial condition* u_h^0 *into* (1.7)*. Then, by* (1.8)*,* $\|u_h^1\|^2 - \|u_h^0\|^2 = \frac{\tau^3}{2}[\![p_h^n]\!]^2 + \|u_h^{n+1} - u_h^n - \tau p_h^n\|^2 > 0$.

1.3.2 Third-order schemes

We can rewrite LW3DG schemes as follows. In each step, find $u_h^{n+1} \in V_h$, such that

$$(u_h^{n+1}, \varphi_h) = (u_h^n, \varphi_h) + \tau(p_h^n, \varphi_h) + \frac{\tau^2}{2}(q_h^n, \varphi_h) + \frac{\tau^3}{6}(r_h^n, \varphi_h), \tag{1.12a}$$

$$(p_h^n, \psi_h) = -\mathcal{H}^{\alpha_u}(u_h^n, \psi_h), \tag{1.12b}$$

$$(q_h^n, \eta_h) = -\mathcal{H}^{\alpha_p}(p_h^n, \eta_h), \tag{1.12c}$$

$$(r_h^n, \zeta_h) = -\mathcal{H}^{\alpha_q}(q_h^n, \zeta_h), \tag{1.12d}$$

for any $\varphi_h, \psi_h, \eta_h, \zeta_h \in V_h$. Note that (1.12) implies

$$u_h^{n+1} = u_h^n + \tau p_h^n + \frac{\tau^2}{2} q_h^n + \frac{\tau^3}{6} r_h^n. \tag{1.13}$$

The proof of the third-order schemes are actually easier, since there is inherent numerical viscosity from the time discretization. The stability will hold for piecewise polynomial elements of arbitrary degrees. In this section, $\lambda \leq 1$ is assumed for simplicity.

Before going into the main theorem, we first prove the following lemma.

Lemma 1.3.4.

$$\|p_h^{n+1}-p_h^n\|^2-\tau^2\|q_h^n\|^2\le(\frac{\tau}{4}+C_\mu\lambda)[\![p_h^n]\!]^2+\tau^2C_\mu\lambda\|q_h^n\|^2.$$

Proof. According to (1.12b) and (1.13), the following relationship holds

$$(p_h^{n+1}-p_h^n,\psi_h)=-\mathcal{H}^{\alpha_u}(u_h^{n+1}-u_h^n,\psi_h)=-\mathcal{H}^{\alpha_u}(\tau p_h^n+\frac{\tau^2}{2}q_h^n+\frac{\tau^3}{6}r_h^n,\psi_h). \quad (1.14)$$

Hence

$$\begin{aligned}(p_h^{n+1}-p_h^n-\tau q_h^n,\psi_h)&=\tau\frac{\alpha_p-\alpha_u}{2}[p_h^n,\psi_h^n]-\frac{\tau^2}{2}\mathcal{H}^{\alpha_u}(q_h^n,\psi_h^n)-\frac{\tau^3}{6}\mathcal{H}^{\alpha_u}(r_h^n,\psi_h^n)\\&\le C_\mu(\tau h^{-\frac{1}{2}}[\![p_h^n]\!]+\tau\lambda\|q_h^n\|+\tau^2\lambda\|r_h^n\|)\|\psi_h^n\|.\end{aligned}$$

Using Lemma 1.2.3, we know $\tau^2\lambda\|r_h^n\|\le\tau C_\mu\lambda^2\|q_h^n\|$. Hence

$$(p_h^{n+1}-p_h^n-\tau q_h^n,\psi_h)\le C_\mu(\tau h^{-\frac{1}{2}}[\![p_h^n]\!]+\tau\lambda\|q_h^n\|)\|\psi_h\|. \quad (1.15)$$

Note that

$$\|p_h^{n+1}-p_h^n\|^2-\tau^2\|q_h^n\|^2=\|p_h^{n+1}-p_h^n-\tau q_h^n\|^2+2\tau(p_h^{n+1}-p_h^n-\tau q_h^n,q_h^n). \quad (1.16)$$

By (1.15),

$$\|p_h^{n+1}-p_h^n-\tau q_h^n\|^2\le C_\mu(\tau\lambda[\![p_h^n]\!]^2+\tau^2\lambda\|q_h^n\|^2),$$

$$(p_h^{n+1}-p_h^n-\tau q_h^n,q_h^n)\le C_\mu(\tau h^{-\frac{1}{2}}[\![p_h^n]\!]+\tau\lambda\|q_h^n\|)\|q_h^n\|\le\frac{1}{8}[\![p_h^n]\!]^2+\tau C_\mu\lambda\|q_h^n\|^2.$$

The proof is completed by substituting the estimates above into (1.16). □

Theorem 1.3.2 (Stability of the LW3DG schemes)**.** *There exists a constant λ, such that when $\tau\le\lambda h$, the solution of schemes (1.12) satisfies,*

(i) $\|u_h^{n+1}\|\le\|u_h^n\|$, *if* $\alpha_u=\alpha_p=+1$, $\alpha_q=\pm1$,

(ii) $\|u_h^n\| \le C_{\mu,\lambda}\|u_h^0\|$, *if* $\alpha_u = +1$, $\alpha_p = -1$ *and* $\alpha_q = \pm 1$,

where $C_{\mu,\lambda}$ *is a constant dependent upon* λ *and the inverse estimate constant* μ, *but is independent of the total time* $T = n\tau$.

Proof. Substituting $\varphi_h = u_h^n$ into (1.12) and summing over j, we have

$$\frac{1}{2}\|u_h^{n+1}\|^2 - \frac{1}{2}\|u_h^n\|^2 + \frac{\tau\alpha_u}{2}[\![u_h^n]\!]^2 - \frac{\tau^3\alpha_p}{12}[\![p_h^n]\!]^2 + \frac{\tau^2\alpha_{up}}{4}[u_h^n, p_h^n] + \frac{\tau^3\alpha_{uq}}{12}[u_h^n, q_h^n] = \frac{1}{2}\|u_h^{n+1} - u_h^n\|^2 - \frac{\tau^2}{2}\|p_h^n\|^2. \tag{1.17}$$

Note that

$$\|u_h^{n+1} - u_h^n\|^2 - \tau^2\|p_h^n\|^2 = \|u_h^{n+1} - u_h^n - \tau p_h^n\|^2 + 2\tau(u_h^{n+1} - u_h^n - \tau p_h^n, p_h^n),$$

where

$$(u_h^{n+1} - u_h^n - \tau p_h^n, p_h^n) = \frac{\tau^2}{2}(q_h^n, p_h^n) + \frac{\tau^3}{6}(r_h^n, p_h^n) = -\frac{\tau^2\alpha_p}{4}[\![p_h^n]\!]^2 - \frac{\tau^3}{6}\|q_h^n\|^2 - \frac{\tau^3\alpha_{pq}}{12}[p_h^n, q_h^n].$$

Hence

$$\|u_h^{n+1} - u_h^n\|^2 - \tau^2\|p_h^n\|^2 = \|u_h^{n+1} - u_h^n - \tau p_h^n\|^2 - \frac{\tau^3\alpha_p}{2}[\![p_h^n]\!]^2 - \frac{\tau^4}{3}\|q_h^n\|^2 - \frac{\tau^4\alpha_{pq}}{6}[p_h^n, q_h^n]. \tag{1.18}$$

Using (1.18), the energy estimate (1.17) becomes

$$\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \tau\alpha_u[\![u_h^n]\!]^2 + \frac{\tau^3\alpha_p}{3}[\![p_h^n]\!]^2 + \frac{\tau^2\alpha_{up}}{2}[u_h^n, p_h^n] + \frac{\tau^3\alpha_{uq}}{6}[u_h^n, q_h^n] + \frac{\tau^4\alpha_{pq}}{6}[p_h^n, q_h^n] = \|u_h^{n+1} - u_h^n - \tau p_h^n\|^2 - \frac{\tau^4}{3}\|q_h^n\|^2.$$

Since

$$(u_h^{n+1} - u_h^n - \tau p_h^n, \varphi_h) = \frac{\tau^2}{2}(q_h^n, \varphi_h) + \frac{\tau^3}{6}(r_h^n, \varphi_h),$$

by Lemma 1.2.3, $\|u_h^{n+1} - u_h^n - \tau p_h^n\| \le \tau^2(\frac{1}{2} + C_\mu\lambda)\|q_h^n\|$. Suppose $\lambda \le 1$, then the estimate becomes

$$\begin{aligned}\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \tau\alpha_u[\![u_h^n]\!]^2 + \frac{\tau^3\alpha_p}{3}[\![p_h^n]\!]^2 + \frac{\tau^2\alpha_{up}}{2}[u_h^n, p_h^n] + \frac{\tau^3\alpha_{uq}}{6}[u_h^n, q_h^n] \\ + \frac{\tau^4\alpha_{pq}}{6}[p_h^n, q_h^n] \le \tau^4(-\frac{1}{12} + C_\mu\lambda)\|q_h^n\|^2.\end{aligned} \tag{1.19}$$

Different from (1.7), the right hand side of (1.19) is automatically negative when λ is small. This is due to the numerical dissipation from the third-order time discretization. One can avoid detailed arguments as in Lemma 1.3.1 to absorb terms with jumps.

(i) When $\alpha_u = \alpha_p = +1$, by using the inequality $ab \le \varepsilon a^2 + \frac{b^2}{4\varepsilon}$, we obtain

$$\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \tau C[\![u_h^n]\!]^2 + \tau^3 C[\![p_h^n]\!]^2 \le \tau^4(-\frac{1}{12} + C_\mu\lambda)\|q_h^n\|^2 + \tau^5 C[\![q_h^n]\!]^2,$$

where the C's are positive constants. Using the inverse estimate, $\tau^5 C[\![q_h^n]\!]^2 \le \tau^4 C_\mu\lambda\|q_h^n\|^2$. Hence $\|u_h^{n+1}\|^2 - \|u_h^n\|^2 \le \tau^4(-\frac{1}{12} + C'_\mu\lambda)\|q_h^n\|^2$. The strong stability holds as long as $\lambda \le \frac{1}{12C'_\mu}$.

(ii) When $\alpha_u = +1$ and $\alpha_p = -1$, the estimate becomes

$$\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \tau[\![u_h^n]\!]^2 - \frac{\tau^3}{3}[\![p_h^n]\!]^2 + \frac{\tau^3\alpha_{uq}}{6}[u_h^n, q_h^n] + \frac{\tau^4\alpha_{pq}}{6}[p_h^n, q_h^n] \le \tau^4(-\frac{1}{12} + C_\mu\lambda)\|q_h^n\|^2.$$

Use Lemma 1.2.3 and the inverse estimate,

$$\frac{\tau^3\alpha_{uq}}{6}[u_h^n, q_h^n] \le \frac{\tau}{2}[\![u_h^n]\!]^2 + \tau^4 C_\mu\lambda\|q_h^n\|^2,$$

$$\frac{\tau^3\alpha_{uq}}{6}[p_h^n, q_h^n] \le \frac{\tau}{6}[\![p_h^n]\!]^2 + \tau^4 C_\mu \lambda \|q_h^n\|^2.$$

Hence

$$\|u_h^{n+1}\|^2 - \|u_h^n\|^2 + \frac{\tau}{2}[\![u_h^n]\!]^2 - \frac{\tau^3}{2}[\![p_h^n]\!]^2 \le \tau^4(-\frac{1}{12} + C_\mu\lambda)\|q_h^n\|^2. \tag{1.20}$$

Now the jump term of p_h^n on the left hand side is negative. As we have done in Theorem 1.3.1, we would introduce $\|p_h^n\|$ to balance this term. Substituting $\psi_h = p_h^n$ into (1.14), we have

$$\frac{1}{2}\|p_h^{n+1}\|^2 - \frac{1}{2}\|p_h^n\|^2 - \frac{1}{2}\|p_h^{n+1} - p_h^n\|^2 = -\frac{\tau}{2}[\![p_h^n]\!]^2 - \frac{\tau^2}{2}\|q_h^n\|^2 - \frac{\tau^3}{6}\mathcal{H}^+(r_h^n, p_h^n). \tag{1.21}$$

Note that

$$\frac{\tau^3}{6}\mathcal{H}^+(r_h^n, p_h^n) = \frac{\tau^3}{6}(q_h^n, r_h^n) \le \frac{\tau^3}{6}\|q_h^n\|\|r_h^n\| \le \tau^2 C_\mu \lambda \|q_h^n\|^2, \tag{1.22}$$

where Lemma 1.2.3 is used in the last inequality. Therefore, by using Lemma 1.3.4 and (1.22) in (1.21), one can get

$$\|p_h^{n+1}\|^2 - \|p_h^n\|^2 \le (C_\mu\lambda - \frac{3}{4})\tau[\![p_h^n]\!]^2 + \tau^2 C_\mu\lambda\|q_h^n\|^2. \tag{1.23}$$

Multiplying (1.23) by τ^2 and adding it to (1.20), we get

$$\begin{aligned}&(\|u_h^{n+1}\|^2 + \tau^2\|p_h^{n+1}\|^2) - (\|u_h^n\|^2 + \tau^2\|p_h^n\|^2) + \frac{\tau}{2}[\![u_h^n]\!]^2 + \tau^3(\frac{1}{4} - C_\mu\lambda)[\![p_h^n]\!]^2 \\ &\le \tau^4(-\frac{1}{12} + C_\mu\lambda)\|q_h^n\|^2.\end{aligned}$$

Therefore, when λ is sufficiently small, all jump terms are nonnegative, and the right

hand side is less than zero. So when uniform time steps are used, we get

$$\|u_h^n\|^2 + \tau^2\|p_h^n\|^2 \leq \|u_h^n\|^2 + \tau^2\|p_h^n\|^2 \leq \|u_h^n\|^2 + \tau^2\|p_h^0\|^2.$$

Using Lemma 1.2.3, one has $\tau^2\|p_h^n\|^2 \leq C_\mu\lambda^2\|u_h^n\|^2$. Hence $\|u_h^n\| \leq \sqrt{1+C_\mu\lambda^2}\|u_h^0\|$. Clearly, the result also holds with non-increasing time steps, namely $\tau_{n+1} \leq \tau_n$. □

1.3.3 Multi-dimensional cases

1.3.3.1 Notations and preliminaries

Let us now consider the general linear scalar conservation law (1.1). Recall that $\nabla \cdot \boldsymbol{\beta}(\boldsymbol{x}) = 0$, hence $\nabla \cdot (\boldsymbol{\beta} u) = (\boldsymbol{\beta} \cdot \nabla)u$.

Suppose $\mathcal{K} = \{K\}$ is a quasi-uniform partition of the domain Ω with triangular (or rectangular) element K, where $h = \max_K h_K$, with h_K being the diameter of element K. The collection of cell interfaces is denoted by $\mathcal{E}$. We use the notation $(u,v)_K = \int_K uv\ dxdy$ for the L^2 inner product on each element, and $(u,v) = \sum_{K\in\mathcal{K}}(u,v)_K$ for the whole domain. Besides, we denote the integration on element interfaces by $\langle u,v\rangle_e = \int_e uv\ dl$ and jumps by $[u]_e = (u^+ - u^-)_e$. Here, $u^+ = \lim_{\varepsilon\to 0^+} u(\boldsymbol{x}+\varepsilon\boldsymbol{\beta})$, while $u^- = \lim_{\varepsilon\to 0^+} u(\boldsymbol{x}-\varepsilon\boldsymbol{\beta})$. Furthermore, let $[\![u]\!]_{\boldsymbol{\beta},e} = \sqrt{\int_e [u]_e^2|\boldsymbol{\beta}\cdot\boldsymbol{n}|dl}$ and $[\![u]\!]_{\boldsymbol{\beta}} = \sqrt{\sum_{e\in\mathcal{E}}[\![u]\!]^2_{\boldsymbol{\beta},e}}$. The corresponding cross term is denoted by $[u,v]_{\boldsymbol{\beta}} = \sum_{e\in\mathcal{E}}\langle [u],[v]|\boldsymbol{\beta}\cdot\boldsymbol{n}|\rangle$.

Similar to our one-dimensional cases, we introduce the operators $\mathcal{H}^{\pm}_{\boldsymbol{\beta}}$ for conve-

nience,

$$\mathcal{H}^{\pm}_{\boldsymbol{\beta}}(w,v)=\sum_{K\in\mathcal{K}}\mathcal{H}^{\pm}_{\boldsymbol{\beta},K}(w,v)=\sum_{K\in\mathcal{K}}\big((w,\nabla\cdot(\boldsymbol{\beta}v))_K-\langle w^{\pm},v\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial K}\big).$$

When $\boldsymbol{\beta}(\boldsymbol{x})$ is parallel to the cell interfaces, $u^{\pm}$ are not well-defined. But in this case, $\boldsymbol{\beta}(\boldsymbol{x})\cdot\boldsymbol{n}=0$, hence the value of $u^{\pm}$ will not make any difference. When $\boldsymbol{\beta}(\boldsymbol{x})\cdot\boldsymbol{n}$ changes sign on the edge, $\mathcal{H}^{\pm}_{\boldsymbol{\beta}}(w,v)$ can still be defined as above. But one should note $w^{\pm}$ are no longer polynomials and one would need to be careful when applying the inverse estimates. See appendix B.2 for details.

As before, we introduce the auxiliary variables $p=u_t=\nabla\cdot(\boldsymbol{\beta}u)$, $q=u_{tt}=\nabla\cdot(\boldsymbol{\beta}p)$ and $r=u_{ttt}=\nabla\cdot(\boldsymbol{\beta}q)$ (for the third-order schemes). Then the numerical schemes can be defined as in the one-dimensional cases in (1.3) and (1.12), except for replacing $\mathcal{H}^{\pm}$ by $\mathcal{H}^{\pm}_{\boldsymbol{\beta}}$. Moreover, we use the discontinuous finite element space $V_h=\{v\in L^2(\Omega):v|_K\in\mathcal{P}_k(K),\forall K\in\mathcal{K}\}$, where $\mathcal{P}_k(K)$ denotes the space of polynomials on K of degree no more than k.

For $\nabla\cdot\boldsymbol{\beta}(\boldsymbol{x})=0$ and $w,v\in V_h$, the following relationships hold, which extend Lemma 1.2.1 and Lemma 1.2.2 to multidimensions. The proof of Lemma 1.3.5 and 1.3.6 is given in the appendix.

Lemma 1.3.5.

$$\begin{aligned}\mathcal{H}^{-}_{\boldsymbol{\beta}}(w,w)&=-\frac{1}{2}[\![w]\!]^2_{\boldsymbol{\beta}},\\ \mathcal{H}^{-}_{\boldsymbol{\beta}}(w,v)&=-\mathcal{H}^{+}_{\boldsymbol{\beta}}(v,w),\\ \mathcal{H}^{+}_{\boldsymbol{\beta}}(w,v)&=\mathcal{H}^{-}_{\boldsymbol{\beta}}(w,v)+[w,v]_{\boldsymbol{\beta}}.\end{aligned}$$

Lemma 1.3.6.

$$\begin{aligned} |\mathcal{H}^-_{\boldsymbol{\beta}}(w,v)| &\leq (\|(\boldsymbol{\beta}\cdot\nabla)w\| + C_{\mu,\boldsymbol{\beta}} h^{-\frac{1}{2}} [\![w]\!]_{\boldsymbol{\beta}})\|v\|, \\ |\mathcal{H}^-_{\boldsymbol{\beta}}(w,v)| &\leq (\|(\boldsymbol{\beta}\cdot\nabla)v\| + C_{\mu,\boldsymbol{\beta}} h^{-\frac{1}{2}} [\![v]\!]_{\boldsymbol{\beta}})\|w\|. \end{aligned}$$

With the relationships above, we can state a similar lemma to Lemma 1.2.4.

Lemma 1.3.7. *Lemma 1.2.4 holds after replacing $\mathcal{H}$ with $\mathcal{H}_{\boldsymbol{\beta}}$, $[\cdot,\cdot]$ with $[\cdot,\cdot]_{\boldsymbol{\beta}}$ and $[\![\cdot]\!]$ with $[\![\cdot]\!]_{\boldsymbol{\beta}}$.*

Lemma 1.3.1 plays a key role in the stability analysis of the LW2DG schemes. Similarly, we need the following lemma to bound $\|(\boldsymbol{\beta}\cdot\nabla)p_h^n\|_K$ with $[\![u_h^n]\!]_{\boldsymbol{\beta},\partial K}$. The proof of this lemma is given in the appendix.

Lemma 1.3.8. *For the piecewise linear polynomial space in $\mathbb{R}^d$,*
(i) If $\boldsymbol{\beta}$ is a constant vector, then $\|(\boldsymbol{\beta}\cdot\nabla)p_h^n\|_K \leq C_{\mu,\boldsymbol{\beta}} h^{-\frac{3}{2}} [\![u_h^n]\!]_{\boldsymbol{\beta},\partial K}$,
(ii) If $\boldsymbol{\beta}$ is a vector-valued function with $\nabla\cdot\boldsymbol{\beta}(\boldsymbol{x}) = 0$, then

$$\|(\boldsymbol{\beta}\cdot\nabla)p_h^n\|_K \leq C_{\mu,\boldsymbol{\beta}} h^{-\frac{3}{2}} [\![u_h^n]\!]_{\boldsymbol{\beta},\partial K} + C_{\mu,\boldsymbol{\beta}} h^{-1}\|u_h^n\|_K.$$

Remark 1.3.3. *Note that our proof only holds for $\mathcal{P}_1$ elements. It is difficult to extend the proof to $\mathcal{Q}_1$ elements, which are defined by tensor products on rectangular meshes.*

1.3.3.2 Stability results

Theorem 1.3.3 (Stability of the LW2DG schemes in multidimensions)**.** *For multidimensional cases, let u_h^n be the numerical solution of the LW2DG schemes with $\mathcal{P}_1$*

elements. Then under the standard CFL condition $\tau \leq \lambda h$,

(i) When $\boldsymbol{\beta}$ *is constant,* $\|u_h^{n+1}\| \leq \|u_h^n\|$ *if* $\alpha_u = \alpha_p = +1$; $\|u_h^n\| \leq C_{\mu,\lambda,\boldsymbol{\beta}}\|u_h^0\|$ *if* $\alpha_u = +1$ *and* $\alpha_p = -1$, *where* $C_{\mu,\lambda,\boldsymbol{\beta}}$ *depends on* λ, μ *and* $\boldsymbol{\beta}$ *but is independent of* T.

(ii) When $\boldsymbol{\beta}(\boldsymbol{x})$ *is a function with* $\nabla \cdot \boldsymbol{\beta}(\boldsymbol{x}) = 0$, *if* $\alpha_u = +1$ *and* $\alpha_p = \pm 1$, *then* $\|u_h^n\| \leq C_{\mu,\lambda,\boldsymbol{\beta},T}\|u_h^0\|$, *where* $C_{\mu,\lambda,\boldsymbol{\beta},T}$ *depends on* λ, μ, $\boldsymbol{\beta}$ *and* T.

Proof. For constant $\boldsymbol{\beta}$, we proceed in the same way as we have done for Theorem 1.3.1, except for using Lemma 1.3.7 instead of Lemma 1.2.4, and Lemma 1.3.8 instead of Lemma 1.3.1. For non-constant function $\boldsymbol{\beta}$, we will get an extra $\|u\|_K$ term in Lemma 1.3.8. A similar argument will lead to $\|u_h^{n+1}\| \leq (1 + \tau C_{\tau,\lambda,\boldsymbol{\beta}})\|u_h^n\|$ when $\alpha_p = +1$, which will end up with $\|u_h^n\| \leq e^{C_{\mu,\lambda,\boldsymbol{\beta}}T}\|u_h^0\|$. For the same reason, the constant will also depend on T when $\alpha_p = -1$. □

Theorem 1.3.4 (Stability of the LW3DG schemes in multidimensions)**.** *For multi-dimensional cases, assume* $\nabla \cdot \boldsymbol{\beta}(\boldsymbol{x}) = 0$. *Let* u_h^n *be the numerical solution of the LW3DG schemes with* $\mathcal{P}_k$ *elements. Then under the standard CFL condition* $\tau \leq \lambda h$,

(i) $\|u_h^{n+1}\| \leq \|u_h^n\|$, *if* $\alpha_u = \alpha_p = +1$ *and* $\alpha_q = \pm 1$,

(ii) $\|u_h^n\| \leq C_{\mu,\lambda,\boldsymbol{\beta}}\|u_h^0\|$, *if* $\alpha_u = +1$, $\alpha_p = -1$ *and* $\alpha_q = \pm 1$,

where $C_{\mu,\lambda,\boldsymbol{\beta}}$ *depends on* λ, μ *and* $\boldsymbol{\beta}$, *but is independent of* T.

Proof. The proof follows the same lines as in the one-dimensional case and is thus omitted. □

Remark 1.3.4. *By now, we have spent our main effort on some special coefficients* $\boldsymbol{\beta}$. *Actually, the LWDG schemes can be modified to suit general* $\boldsymbol{\beta}(\boldsymbol{x}, t)$, *where* $\boldsymbol{\beta}$ *may depend on time and can have a non-zero divergence. Furthermore, as long as* $\boldsymbol{\beta}$ *is sufficiently smooth, we can obtain the weak stability result* $\|u_h^n\| \leq C\|u_h^0\|$, *where* C *depends on* $\boldsymbol{\beta}$ *and* T, *as well as* μ *and* λ *if alternating fluxes are used.*

We use the LW2DG scheme with upwinding fluxes for a brief illustration. The revised scheme then becomes

$$
\begin{aligned}
(u_h^{n+1}, \varphi_h) &= (u_h^n + \tau p_h^n + \frac{\tau^2}{2} q_h^n, \varphi_h), \\
(p_h^n, \psi_h) &= -\mathcal{H}_{\boldsymbol{\beta}}^+(u_h^n, \psi_h) + (\nabla \cdot \boldsymbol{\beta} u_h^n, \psi_h), \\
(q_h^n, \eta_h) &= -\mathcal{H}_{\boldsymbol{\beta}}^+(p_h^n, \eta_h) + (\nabla \cdot \boldsymbol{\beta} p_h^n, \eta_h) - \mathcal{H}_{\boldsymbol{\beta}_t}^+(u_h^n, \eta_h) + (\nabla \cdot \boldsymbol{\beta}_t u_h^n, \eta_h).
\end{aligned}
$$

Here we have the extra terms $(\nabla \cdot \boldsymbol{\beta} u_h^n, \psi_h)$, $(\nabla \cdot \boldsymbol{\beta} p_h^n, \eta_h)$, $\mathcal{H}_{\boldsymbol{\beta}_t}^+(u_h^n, \eta_h)$ *and* $(\nabla \cdot \boldsymbol{\beta}_t u_h^n, \eta_h)$. *Furthermore, the properties of* $\mathcal{H}$ *has been changed and there will be more terms coming out when one follows the previous approach. But compared with the exponents of* τ *before them, the order of derivatives for all these new terms are relatively low. One can apply the Cauchy-Schwartz inequality, Lemma 1.3.6 and the inverse estimates to show they can be bounded by* $C\tau \|u_h^n\|^2$ *with some proper constant* C. *Hence we will finally obtain* $\|u_h^{n+1}\|^2 \leq (1 + C\tau)\|u_h^n\|^2$, *which indicates the weak stability result.*

1.4 Error estimates

For the one-dimensional problem (1.2), denote $u^n(x) = u(x, t_n)$ and $u_t^n(x) = u_t(x, t_n)$. Assuming the exact solution $u(x, t)$ is smooth, we have the following error estimates.

Theorem 1.4.1 (Error estimates of the LW2DG schemes)**.** *Under the CFL condition* $\tau \leq \lambda h$ *for a proper constant* λ, *using* $\mathcal{P}_1$ *elements, the numerical approximations* u_h^n *and* p_h^n *from the LW2DG schemes* (1.3) *satisfy the following error estimates,*
(i) $\|u^n - u_h^n\| \leq C(\tau^2 + h^2)$,
(ii) $\|u_t^n - p_h^n\| \leq C(\tau^2 + h^2)$,
where C *depends on* μ, λ, T, u, u_t *and their derivatives but is independent of* τ *and* h.

Theorem 1.4.2 (Error estimates of the LW3DG schemes)**.** *Under the CFL condition $\tau \le \lambda h$ for a proper constant λ, using $\mathcal{P}_k$ elements (k is arbitrary), the numerical approximations u_h^n and p_h^n from the LW3DG schemes (1.12) satisfy the following error estimates,*

(i) $\|u^n - u_h^n\| \le C(\tau^3 + h^{k+1})$,

(ii) $\|u_t^n - p_h^n\| \le C(\tau^3 + h^{k+1})$,

where C depends on μ, λ, T, u, u_t and their derivatives but is independent of τ and h.

The error estimate of u follows from the stability results through a standard argument, while the error estimate of u_t relies on the relationship between u_h^n and p_h^n. Here we only give the details of the proof of Theorem 1.4.1 and omit the details of the proof of Theorem 1.4.2, as they can be obtained following similar lines.

We will need the following lemma for the error estimates. The proof of (1.24a) is straightforward, and (1.24b) follows from the standard approximation theorem [24].

Lemma 1.4.1. *Suppose w is a smooth function, then*

$$\mathcal{H}^{\pm}(\pi^{\pm} w - w, v) = 0, \qquad \forall v \in V_h, \tag{1.24a}$$

$$\|\pi^{\pm} w - w\| \le C_w h^{k+1}, \tag{1.24b}$$

where C_w is a constant depending on w; $\pi^{\pm}$ are the Gauss-Radau projections to V_h. More precisely, $\pi^{\pm} w$ is the unique element in V_h such that

$$(\pi^{\pm} w - w, v)_i = 0 \quad \forall v \in \mathcal{P}_{k-1}(I_i), \quad (\pi^{\pm} w)^{\pm}_{i \mp \frac{1}{2}} = w^{\pm}_{i \mp \frac{1}{2}}.$$

Proof of Theorem 1.4.1: Let u be the exact solution of (1.1), $p = u_t$ and $q = u_{tt}$. We

denote the truncation error by

$$\omega^n = u^{n+1} - u^n - \tau p^n - \frac{\tau^2}{2} q^n. \tag{1.25}$$

By the consistency of the schemes, for any $\varphi_h, \psi_h, \eta_h \in V_h$, the following equalities hold

$$(u^{n+1}, \varphi_h) = (u^n + \tau p^n + \frac{\tau^2}{2} q^n, \varphi_h) + (\omega^n, \varphi_h), \tag{1.26a}$$

$$(p^n, \psi_h) = -\mathcal{H}^{\alpha_u}(u^n, \psi_h), \tag{1.26b}$$

$$(q^n, \eta_h) = -\mathcal{H}^{\alpha_p}(p^n, \eta_h). \tag{1.26c}$$

Subtract the system (1.26) from the numerical schemes (1.3), then we get the error equation

$$\begin{aligned}
(u^{n+1} - u_h^{n+1}, \varphi_h) &= (u^n - u_h^n + \tau(p^n - p_h^n) + \frac{\tau^2}{2}(q^n - q_h^n), \varphi_h) + (\omega^n, \varphi_h), \\
(p^n - p_h^n, \psi_h) &= -\mathcal{H}^{\alpha_u}(u^n - u_h^n, \psi_h), \\
(q^n - q_h^n, \eta_h) &= -\mathcal{H}^{\alpha_p}(p^n - p_h^n, \eta_h).
\end{aligned}$$

Denote $e_v^n = \pi^{\alpha_v} v - v_h^n$ and $\varepsilon_v^n = v - \pi^{\alpha_v} v$, with $v = u, p, q$. Applying Lemma 1.4.1, the error equation can be simplified as follows

$$(e_u^{n+1} - e_u^n - \tau e_p^n - \frac{\tau^2}{2} e_q^n, \varphi_h) = -(\varepsilon_u^{n+1} - \varepsilon_u^n - \tau \varepsilon_p^n - \frac{\tau^2}{2} \varepsilon_q^n, \varphi_h) + (\omega^n, \varphi_h), \tag{1.27a}$$

$$(e_p^n, \psi_h) = -\mathcal{H}^{\alpha_u}(e_u^n, \psi_h) - (\varepsilon_p^n, \psi_h), \tag{1.27b}$$

$$(e_q^n, \eta_h) = -\mathcal{H}^{\alpha_p}(e_p^n, \eta_h) - (\varepsilon_q^n, \eta_h). \tag{1.27c}$$

Furthermore, by using (1.25) and (1.27b), one can also obtain the error equation for

u_t

$$(e_p^{n+1} - e_p^n, \psi_h) = -\mathcal{H}^{\alpha_u}(\tau e_p^n + \frac{\tau^2}{2} e_q^n + w, \psi_h) - (\varepsilon_p^{n+1} - \varepsilon_p^n, \psi_h). \tag{1.28}$$

Our error estimates are based on the framework of stability analysis. We need lemmas comparable to Lemma 1.2.3, Lemma 1.2.4, Lemma 1.3.1, Lemma 1.3.2 and Lemma 1.3.3. Among them, Lemma 1.4.2, Lemma 1.4.3 and Lemma 1.4.4 can be proved by using (1.27) and similar techniques as before, and the details are omitted.

Lemma 1.4.2.

$$\|e_p^n\| \le C_\mu h^{-1} \|e_u^n\| + \|\varepsilon_p^n\|, \tag{1.29a}$$

$$\|e_q^n\| \le C_\mu h^{-1} \|e_p^n\| + \|\varepsilon_q^n\|. \tag{1.29b}$$

Lemma 1.4.3.

$$\begin{aligned}
(e_p^n, e_u^n) &= -\frac{\alpha_u}{2} [\![e_u^n]\!]^2 - (\varepsilon_p^n, e_u^n), \\
(e_q^n, e_u^n) &= -\|e_p^n\|^2 - \frac{\alpha_{up}}{2} [e_u^n, e_p^n] - (\varepsilon_q^n, e_u^n) - (\varepsilon_p^n, e_p^n), \\
(e_q^n, e_p^n) &= -\frac{\alpha_p}{2} [\![e_p^n]\!]^2 - (\varepsilon_q^n, e_p^n),
\end{aligned}$$

where $\alpha_{wv} = \alpha_w + \alpha_v$, $w, v = \pm 1$.

Lemma 1.4.4. *For piecewise linear elements,* $\|(e_p^n)_x\| \le C_\mu h^{-\frac{3}{2}} [\![e_u^n]\!] + C_\mu h^{-1} \|\varepsilon_p^n\|$.

Lemma 1.4.5.

$$\|e_u^{n+1} - e_u^n - \tau e_p^n\| \le C_\mu \tau^2 (h^{-\frac{3}{2}} [\![e_u^n]\!] + h^{-\frac{1}{2}} [\![e_p^n]\!]) + C(\tau^3 + \tau h^2).$$

Proof. Substituting (1.27b) into (1.27a), we get

$$\begin{aligned}
&(e_u^{n+1} - e_u^n - \tau e_p^n, \varphi_h) = -(\varepsilon_u^{n+1} - \varepsilon_u^n - \tau\varepsilon_p^n, \varphi_h) - \frac{\tau^2}{2}\mathcal{H}^{\alpha_p}(e_p^n, \varphi_h) + (\omega^n, \varphi_h) \\
&\leq \left(\|\varepsilon_u^{n+1} - \varepsilon_u^n\| + \tau\|\varepsilon_p^n\| + \|\omega^n\| + \tau^2 C_\mu(\|(e_p^n)_x\| + h^{-\frac{1}{2}}[\![e_p^n]\!]) \right) \|\varphi_h\| \\
&\leq \left(\|\varepsilon_u^{n+1} - \varepsilon_u^n\| + \tau(1 + C_\mu\lambda)\|\varepsilon_p^n\| + \|\omega^n\| + \tau^2 C_\mu(h^{-\frac{3}{2}}[\![e_u^n]\!] + h^{-\frac{1}{2}}[\![e_p^n]\!]) \right) \|\varphi_h\|,
\end{aligned} \tag{1.30}$$

where we have used Lemma 1.4.4 in the last inequality. One can see that

$$\|\varepsilon_u^{n+1} - \varepsilon_u^n\| + \tau(1 + C_\mu\lambda)\|\varepsilon_p^n\| + \|\omega^n\| \leq C(\tau^3 + \tau h^2). \tag{1.31}$$

The proof can be completed by using the estimate (1.31) and substituting $\varphi_h = e_u^{n+1} - e_u^n - \tau e_p^n$ into (1.30). □

Lemma 1.4.6.

$$\|e_p^{n+1} - e_p^n\| \leq C_\mu(\tau h^{-\frac{3}{2}}[\![e_u^n]\!] + \tau h^{-\frac{1}{2}}[\![e_p^n]\!] + \tau\lambda\|e_q^n\|) + C(\tau^2 + h^2).$$

Proof. By (1.28), one has

$$\begin{aligned}
&(e_p^{n+1} - e_p^n, \psi_h) = -\mathcal{H}^{\alpha_u}(\tau e_p^n + \frac{\tau^2}{2} e_q^n + w, \psi_h) - (\varepsilon_p^{n+1} - \varepsilon_p^n, \psi_h) \\
&\leq \left(C_\mu(\tau\|(e_p^n)_x\| + \tau h^{-\frac{1}{2}}[\![e_p^n]\!] + \tau\lambda\|e_q^n\| + h^{-1}\|\omega^n\|) + \|\varepsilon_p^{n+1} - \varepsilon_p^n\| \right) \|\psi_h\| \\
&\leq \left(C_\mu(\tau h^{-\frac{3}{2}}[\![e_u^n]\!] + \tau h^{-\frac{1}{2}}[\![e_p^n]\!] + \tau\lambda\|e_q^n\| + \lambda\|\varepsilon_p^n\| + h^{-1}\|\omega^n\|) + \|\varepsilon_p^{n+1} - \varepsilon_p^n\| \right) \|\psi_h\|,
\end{aligned} \tag{1.32}$$

where we have used Lemma 1.4.4 in the last inequality. We see that

$$\lambda\|\varepsilon_p^n\| + h^{-1}\|\omega^n\| + \|\varepsilon_p^{n+1} - \varepsilon_p^n\| \leq C(\tau^2 + h^2). \tag{1.33}$$

The proof is completed by using the estimate (1.33) and substituting $\varphi_h = e_p^{n+1} - e_p^n$ into (1.32). □

Step 1: To prove $\|e_u^n\| \le C(\tau^2 + h^2)$.

Substitute $\varphi_h = e_u^n$ into (1.27) and use Lemma 1.4.3. It follows that

$$\|e_u^{n+1}\|^2 - \|e_u^n\|^2 + \tau\alpha_u[\![e_u^n]\!]^2 + \tau^2\alpha_{up}[e_u^n, e_p^n] + \frac{\tau^3\alpha_p}{2}[\![e_p^n]\!]^2 = \|e_u^{n+1} - e_u^n - \tau e_p^n\|^2 + \gamma_u,$$

where $\gamma_u = -2\tau(\varepsilon_p^n, e_u^n) - 2\tau^2(\varepsilon_q^n, e_u^n) + 2\tau^2(\varepsilon_p^n, e_p^n) - \tau^3(\varepsilon_q^n, e_p^n) + (\varepsilon_u^{n+1} - \varepsilon_u^n, e_u^n) + (\omega^n, e_u^n)$. Using Lemma 1.4.2, one can get an estimate for γ_u

$$\gamma_u \le C_\mu\tau\|e_u^n\|^2 + \tau C(\|\varepsilon_p^n\|^2 + \|\varepsilon_q^n\|^2) + \frac{1}{\tau}\|\varepsilon_u^{n+1} - \varepsilon_u^n\|^2 + \frac{1}{\tau}\|\omega\|^2 \le C_\mu\tau\|e_u^n\|^2 + C(\tau^5 + \tau h^4).$$

Hence, with the estimate above, by using Lemma 1.4.5, we get

$$\begin{aligned} &\|e_u^{n+1}\|^2 + \tau(\alpha_u - C_\mu\lambda)[\![e_u^n]\!]^2 + \tau^2\alpha_{up}[e_u^n, e_p^n] + \tau^3(\frac{\alpha_p}{2} - C_\mu\lambda)[\![e_p^n]\!]^2 \\ &\le (1 + \tau C_\mu)\|e_u^n\|^2 + C(\tau^5 + \tau h^4). \end{aligned} \tag{1.34}$$

Case 1: When λ is small and $\alpha_u = \alpha_p = +1$, all the jump terms can be removed to yield a much simpler inequality

$$\|e_u^{n+1}\|^2 \le (1 + \tau C_\mu)\|e_u^n\|^2 + C(\tau^5 + \tau h^4).$$

Note that $e_u^0 = 0$. (1.34) implies

$$\|e_u^n\| \le C(\tau^2 + h^2).$$

Case 2: When λ is small, $\alpha_u = +1$ and $\alpha_p = -1$, we have

$$\|e_u^{n+1}\|^2+\tau(1-C_\mu\lambda)[\![e_u^n]\!]^2+\tau^3(-\frac{1}{2}-C_\mu\lambda)[\![e_p^n]\!]^2 \le (1+\tau C_\mu)\|e_u^n\|^2+C(\tau^5+\tau h^4). \quad (1.35)$$

Plug in $\psi_h = e_p^n$ in (1.28). It follows that

$$\|e_p^{n+1}\|^2 - \|e_p^n\|^2 = \|e_p^{n+1} - e_p^n\|^2 - \tau^2\|e_q^n\|^2 - \tau\alpha_u[\![e_p^n]\!]^2 + \gamma_p,$$

where $\gamma_p = -\tau^2(\varepsilon_q^n, e_q^n) - \mathcal{H}^+(\omega, e_p^n) - (\varepsilon_p^{n+1} - \varepsilon_p^n, e_p^n)$. Note that

$$\gamma_p \le \tau C_\mu\|e_p^n\|^2 + \tau\|\varepsilon_q^n\|^2 + \frac{1}{\tau}\|\varepsilon_u^{n+1} - \varepsilon_u^n\|^2 + \frac{1}{\tau h^2}\|\omega\|^2 \le C_\mu\tau\|e_p^n\|^2 + C(\tau^5 h^{-2} + \tau h^4).$$

Hence, using Lemma 1.4.6, we get

$$\begin{aligned}\|e_p^{n+1}\|^2 \le&(1 + \tau C_\mu)\|e_p^n\|^2 + C_\mu h^{-1}\lambda^2[\![e_u^n]\!]^2 - \tau^2(1 - C_\mu\lambda^2)\|e_q^n\|^2 \\ &- \tau(1 - C_\mu\lambda)[\![e_p^n]\!]^2 + C(\tau^5 h^{-2} + \tau^4 + h^4).\end{aligned} \quad (1.36)$$

(1.35) and (1.36) lead to the following inequality

$$\|e_u^{n+1}\|^2 + \tau^2\|e_p^{n+1}\|^2 \le (1 + \tau C_\mu)(\|e_u^n\|^2 + \tau^2\|e_p^n\|^2) + C(\tau^5 + \tau h^4).$$

Note that $\|e_u^0\| = 0$, $\|e_p^0\| \le \|\varepsilon_h^0\| \le Ch^2$. Hence we have,

$$\|e_u^n\|^2 + \tau^2\|e_p^n\|^2 \le C(\tau^4 + h^4),$$

which implies $\|e_u^n\| \le C(\tau^2 + h^2)$.

Step 2: We claim that $\|e_u^{n+1} - e_u^n\| \le C(\tau^3 + \tau h^2)$.

Note that $e_u^{n+1} - e_u^n$ satisfies the same error equation as that of e_u^n, except for $\|\omega^n\| \le C\tau^5$. Noting also that $\|e_u^1 - e_u^0\| = \|e_u^1\| \le C\tau h^2$ by estimating (1.27a) directly, we

can prove the statement in the same way as that in Step 1.

Step 3: (1.27) gives that

$$(e_u^{n+1} - e_u^n - \tau e_p^n - \frac{\tau^2}{2} e_q^n, e_p^n) = -(\nu^n, e_p^n),$$

where $\nu^n = \varepsilon_u^{n+1} - \varepsilon_u^n - \tau\varepsilon_p^n - \frac{\tau^2}{2}\varepsilon_q^n + \omega^n$. Hence

$$\|\nu^n\| \le \|\varepsilon_u^{n+1} - \varepsilon_u^n\| + \tau\|\varepsilon_p^n + \frac{\tau}{2}\varepsilon_q^n\| + \|\omega^n\| \le C(\tau^3 + \tau h^2).$$

Rearranging the equality, we have

$$\begin{aligned}
&\tau\|e_p^n\|^2 = (e_u^{n+1} - e_u^n - \frac{\tau^2}{2} e_q^n, e_p^n) - (\nu^n, e_p^n) \\
&\Rightarrow \tau\|e_p^n\| \le \|e_u^{n+1} - e_u^n\| + \frac{\tau^2}{2}\|e_q^n\| + \|\nu^n\| \\
&\Rightarrow \tau(1 - C_\mu\lambda)\|e_p^n\| \le \|e_u^{n+1} - e_u^n\| + \frac{\tau^2}{2}\|\varepsilon_q^n\| + \|\nu^n\|,
\end{aligned}$$

where Lemma 1.4.2 is used in the last step. Using the results in Step 1 and Step 2, one can get

$$\|e_p^n\| \le C(\tau^2 + h^2).$$

Hence by Lemma 1.4.1,

$$\|u^n - u_h^n\| \le \|\varepsilon_u^n\| + \|e_u^n\| \le C(\tau^2 + h^2),$$

$$\|u_t^n - p_h^n\| \le \|\varepsilon_p^n\| + \|e_p^n\| \le C(\tau^2 + h^2).$$

□

Sketch of the proof for Theorem 1.4.2: We adopt similar notations as before. By using the same lines as those in the stability analysis, one can show $\|e_u^n\| \le C(\tau^3 + h^{k+1})$ and $\|e_u^{n+1} - e_u^n\| \le C(\tau^4 + \tau h^{k+1})$. This already implies the error estimate of u.

Like what we have done in Step 3 above, $\tau\|e_p^n\| \leq \|e_u^{n+1} - e_u^n\| + \frac{\tau^2}{2}\|e_q^n\| + \frac{\tau^3}{6}\|e_r^n\| + \|\nu^n\|$, where $\nu^n = \varepsilon_u^{n+1} - \varepsilon_u^n - \tau\varepsilon_p^n - \frac{\tau^2}{2}\varepsilon_q^n - \frac{\tau^3}{6}\varepsilon_r^n + \omega^n$. We can bound $\|e_p^n\|$ by estimating each term here. □

Remark 1.4.1. *Note that our optimal error estimates are only one-dimensional results. Since there are no proper Gauss-Radau projections for $\mathcal{P}_k$ elements in multidimensions to eliminate the cell boundary terms, one may lose the optimal order of accuracy when directly applying trace inverse inequalities. As for rectangular meshes with $\mathcal{Q}_k$ elements, we may obtain optimal error estimates for the third-order schemes using similar arguments as in the one-dimensional case.*

1.5 Numerical tests

The main purpose of this section is to numerically validate the error estimates in Section 1.4. We list the error tables of the second-order schemes with fluxes $(+,+)$ and $(+,-)$, and the third-order schemes with numerical fluxes $(+,+,+)$ and $(+,-,-)$. Here $(+,+)$ means $\alpha_u = +1$ and $\alpha_p = +1$. Others are defined analogously. The piecewise linear polynomial space is used for the second-order schemes and the piecewise quadratic polynomial space is used for the third-order schemes. Furthermore, we use the uniform time steps and uniform spatial meshes for the numerical tests. For two-dimensional problems, the triangular meshes are used and the triangulation is constructed by adding diagonals linking the left-bottom and right-top vertexes in a uniform square mesh.

1.5.1 One-dimensional example with constant coefficients

Consider

$$\begin{cases} u_t = u_x, & (x,t) \in (0,2\pi) \times (0,T), \\ u(x,0) = e^{\sin(x)}, & x \in (0,2\pi), \end{cases}$$

with periodic boundary conditions. The mesh size is chosen as $h = 2\pi/N$, where N is the number of cells and $N = 40, 80, 160, 320, 640$ are used. We compute up to the time $T = \pi/2$, with $\tau = 0.05h$. The numerical results are listed in Table 1.1. The convergence is as we have predicted in Section 1.4.

Table 1.1: One-dimensional accuracy test of LWDG methods for the linear conservation law with constant coefficients. Compute to $T = \frac{\pi}{2}$ with $\lambda = 0.05$.

scheme	N	u L^2 error	order	u_t L^2 error	order
$(+,+)$	40	4.372E-03	-	7.728E-03	-
	80	1.099E-03	1.99	1.890E-03	2.02
	160	2.765E-04	1.99	4.732E-04	2.01
	320	6.941E-05	1.99	1.182E-04	2.00
	640	1.739E-05	2.00	2.957E-05	2.00
$(+,-)$	40	4.381E-03	-	6.688E-03	-
	80	1.100E-03	1.99	1.615E-03	2.05
	160	2.766E-04	1.99	4.001E-04	2.01
	320	6.941E-05	1.99	9.985E-05	2.00
	640	1.739E-05	2.00	2.496E-05	2.00
$(+,+,+)$	40	9.055E-05	-	1.892E-04	-
	80	1.134E-05	3.00	2.402E-05	2.98
	160	1.417E-06	3.00	3.029E-06	2.99
	320	1.772E-07	3.00	3.803E-07	2.99
	640	2.215E-08	3.00	4.764E-08	3.00
$(+,-,-)$	40	9.025E-05	-	1.921E-04	-
	80	1.133E-05	2.99	2.447E-05	2.97
	160	1.417E-06	3.00	3.082E-06	2.99
	320	1.772E-07	3.00	3.865E-07	3.00
	640	2.215E-08	3.00	4.838E-08	3.00

1.5.2 One-dimensional example with variable coefficients

Using the same parameters, we solve the following problem with a variable coefficient

$$\begin{cases} u_t = \sin^2(x)\ u_x, & (x,t) \in (0, 2\pi) \times (0, T), \\ u(x,0) = \sin(x), & x \in (0, 2\pi), \end{cases} \tag{1.37}$$

where periodic boundary conditions are assumed. The exact solution to (1.37) is $u(x,t) = \sin\left(\cot^{-1}(\cot(x) - t)\right)$. Note that our analysis does not cover the problems with variable coefficients in one dimension. This numerical test is given only for showing the generality of our results. The numerical results are listed in Table 1.2. One can observe the designed order of accuracy for all the schemes and for both u and u_t.

1.5.3 Two-dimensional example with constant coefficients

Consider

$$\begin{cases} u_t = \dfrac{\sqrt{2}}{2}\ u_x + \dfrac{\sqrt{2}}{2} u_y, & (x,y,t) \in (0, 2\pi) \times (0, 2\pi) \times (0, T), \\ u(x,0) = \sin(x+y), & (x,y) \in (0, 2\pi) \times (0, 2\pi), \end{cases} \tag{1.38}$$

with periodic boundary conditions. The mesh is constructed by adding diagonals to the uniform square mesh with $N \times N$ elements, where $N = 20, 40, 80, 160, 320$. We compute up to time $T = 1$, with $\tau = 0.05h$. The numerical results are listed in Table 1.3. Once again, the designed order of accuracy can be observed.

Table 1.2: One-dimensional accuracy test of LWDG methods for the linear conservation law with variable coefficients. Compute to $T = \frac{\pi}{2}$ with $\lambda = 0.05$.

scheme	N	u L^2 error	order	u_t L^2 error	order
$(+,+)$	40	4.331E-02	-	1.234E-01	-
	80	1.266E-02	1.77	3.200E-02	1.95
	160	3.413E-03	1.89	8.610E-03	1.89
	320	8.881E-04	1.94	2.243E-03	1.94
	640	2.266E-04	1.97	5.726E-04	1.97
$(+,-)$	40	4.352E-02	-	1.238E-01	-
	80	1.271E-02	1.78	3.133E-02	1.98
	160	3.419E-03	1.89	8.324E-03	1.91
	320	8.888E-04	1.94	2.159E-03	1.95
	640	2.267E-04	1.97	5.503E-04	1.97
$(+,+,+)$	40	6.157E-03	-	1.139E-02	-
	80	8.844E-04	2.80	3.057E-03	1.90
	160	1.180E-04	2.91	4.142E-04	2.88
	320	1.524E-05	2.95	5.355E-05	2.95
	640	1.937E-06	2.98	6.808E-06	2.98
$(+,-,-)$	40	6.143E-03	-	1.110E-02	-
	80	8.804E-04	2.80	2.985E-03	1.90
	160	1.176E-04	2.90	4.104E-04	2.86
	320	1.522E-05	2.95	5.340E-05	2.94
	640	1.936E-06	2.97	6.801E-06	2.97

1.5.4 Two-dimensional example with variable coefficients

We then compute a two-dimensional problem with variable coefficients. Consider

$$\begin{cases} u_t = (-\frac{y}{\pi} + 1)u_x + (\frac{x}{\pi} - 1)u_y, & (x, y, t) \in (0, 2\pi) \times (0, 2\pi) \times (0, T), \\ u(x, 0) = e^{-2(x-\pi)^2 - 4(y-\pi)^2}, & (x, y) \in (0, 2\pi) \times (0, 2\pi), \end{cases}$$

with periodic boundary conditions. The same parameters are used as that of (1.38), except for the time steps $\tau = 0.02h$. When calculating the exact solution, we remove the periodic boundary conditions and compute by tracing the characteristic lines. For general initial inputs, this does cause inconsistency, and the error can be from the mismatch of both u and $\boldsymbol{\beta}$ outside of the domain. However, since our initial condition

Table 1.3: Two-dimensional accuracy test of LWDG methods for linear conservation law with constant coefficients. Compute to $T = 1$ with $\lambda = 0.05$.

scheme	$N \times N$	u L^2 error	order	u_t L^2 error	order
$(+,+)$	20×20	1.288E-01	-	1.816E-01	-
	40×40	3.184E-02	2.02	4.499E-02	2.01
	80×80	7.880E-03	2.01	1.114E-02	2.01
	160×160	1.958E-03	2.01	2.769E-03	2.01
	320×320	4.879e-04	2.00	6.899e-04	2.00
$(+,-)$	20×20	1.289E-01	-	1.789E-01	-
	40×40	3.184E-02	2.02	4.418E-02	2.02
	80×80	7.880E-03	2.01	1.093E-02	2.02
	160×160	1.958E-03	2.01	2.715E-03	2.01
	320×320	4.879E-04	2.00	6.765E-03	2.00
$(+,+,+)$	20×20	2.729E-03	-	3.823E-03	-
	40×40	3.329E-04	3.04	4.741E-04	3.01
	80×80	4.156E-05	3.00	5.890E-05	3.01
	160×160	5.195E-06	3.00	7.362E-06	3.00
	320×320	6.494E-07	3.00	9.203E-07	3.00
$(+,-,-)$	20×20	2.707E-03	-	3.688E-03	-
	40×40	3.327E-04	3.02	4.521E-04	3.03
	80×80	4.155E-05	3.00	5.654E-05	3.00
	160×160	5.194E-06	3.00	7.058E-06	3.00
	320×320	6.494E-07	3.00	8.844E-07	3.00

is intentionally chosen as a Gaussian centered at (π, π), which decays quickly near the boundary, the difference between the exact solution we use and the real exact solution is negligibly small. For simplicity of implementation, we apply upwind and downwind fluxes in terms of the wind direction at the midpoint of each edge, instead of strictly using the definition in our analysis. Since the coefficients are smooth, this should cause negligible difference. The L^2 error of u and u_t is listed in Table 1.4. We can observe the designed convergence of u. As for u_t, the LW2DG schemes exhibit a clear second-order convergence, and the order of accuracy for the LW3DG schemes are also close to the third-order.

Finally, we remark that, even though the choices of the numerical fluxes may have influence on the specific types of stability, they only cause minor differences on

Table 1.4: Two-dimensional accuracy test of LWDG methods for linear conservation law with variable coefficients. Compute to $T = 1$ with $\lambda = 0.02$.

scheme	$N \times N$	u L^2 error	order	u_t L^2 error	order
$(+,+)$	20×20	6.075E-02	-	2.567E-02	-
	40×40	1.607E-02	1.92	6.041E-03	2.09
	80×80	4.095E-03	1.97	1.422E-03	2.09
	160×160	1.030E-03	1.99	3.436E-04	2.05
	320×320	2.582e-04	2.00	8.453E-05	2.02
$(+,-)$	20×20	6.075e-02	-	2.558e-02	-
	40×40	1.607e-02	1.92	6.031e-03	2.08
	80×80	4.096e-03	1.97	1.420e-03	2.09
	160×160	1.030e-03	1.99	3.431e-04	2.05
	320×320	2.582e-04	2.00	8.442e-05	2.02
$(+,+,+)$	20×20	5.609e-03	-	6.554e-03	-
	40×40	6.454e-04	3.12	9.353e-04	2.81
	80×80	7.453e-05	3.11	9.458e-05	3.31
	160×160	9.030e-06	3.05	8.268e-06	3.52
	320×320	1.121e-06	3.01	7.025e-07	3.56
$(+,-,-)$	20×20	5.601E-03	-	6.501E-03	-
	40×40	6.449E-04	3.12	9.258E-04	2.81
	80×80	7.451E-05	3.11	9.360E-05	3.31
	160×160	9.029E-06	3.04	8.213E-06	3.51
	320×320	1.121E-06	3.01	6.999E-07	3.55

the L^2 error according to our numerical tests.

1.6 Concluding remarks

We analyze the stability and estimate the error of the Lax-Wendroff DG method for linear scalar conservation laws. Assume uniform or non-increasing time steps, one can choose between upwind and downwind fluxes for each variable. As we have shown, for both second-order LW2DG schemes with $\mathcal{P}_1$ elements and third-order LW3DG schemes with arbitrary $\mathcal{P}_k$, as long as we use the upwind flux for u, the scheme will be stable. Furthermore, if we also use the upwind flux for p (which approximates u_t), then the scheme will be strongly stable. On the other hand, if the downwind flux is

used for p, then the energy of the numerical solution will be bounded by the initial energy times a constant, which is independent of the total time T. In both cases we have a good control of the L^2 energy after long time integration. The stability results can be extended to high dimensions easily for problems with constant coefficients. For variable coefficient problems with a divergence-free condition $\nabla \cdot \boldsymbol{\beta}(x) = 0$, the previous results still hold for the third-order schemes. But for the second-order schemes, we can only prove the stability in a weaker sense, namely $\|u_h^n\| \leq C\|u_h^0\|$ where C may depend on T.

For the error estimates, we analyze one-dimensional problems with a smooth solution. We obtain optimal error estimates for both u and u_t. Once we choose the upwind flux for u, whatever fluxes we use for the remaining variables, our error estimates will hold. The LW2DG scheme with $\mathcal{P}_1$ elements is of the second-order accuracy, and the LW3DG scheme with P_2 elements is of the third-order accuracy.

Even though we have considered only scalar problems in this chapter, the analysis can be easily generalized to one-dimensional linear hyperbolic systems and multi-dimensional symmetric linear systems. The method also generalizes to nonlinear scalar equations or systems, however stability analysis is expected to be difficult. On the other hand, error estimates for smooth solutions should carry through, along the lines in [83, 60] for RKDG methods.

Chapter Two

The fourth-order Runge-Kutta time discretization

2.1 Introduction

In this chapter, we continue to analyze the fourth-order Runge-Kutta method. The model problem for stability analysis is chosen as a well-posed problem $\partial_t u = L(x,t,\partial_x)u$. But firstly, let us consider a simpler case $\partial_t u = L(x,\partial_x)u$ with $L + L^\top \leq 0$. (Here the superscript $\top$ stands for the adjoint, to be distinguished from T, which stands for the transpose. But for matrices, they are the same under the usual dot product in $\mathbb{R}^m$.) Its semi-discrete scheme corresponds to an autonomous ordinary differential equation (ODE) system

$$\frac{d}{dt}u_N(t) = L_N u_N(t).$$

Here L_N is a constant matrix and the parameter N relates with the degree of freedom in the spatial discretization. Suppose L_N inherits the semi-negativity $L_N + L_N^\top \leq 0$. Then $\frac{d}{dt}\|u_N(t)\|^2 = (L_N u_N, u_N) + (u_N, L_N u_N) \leq 0$, where we temporarily assume $(\cdot,\cdot)$ to be the usual dot product and $\|\cdot\|$ to be the induced norm. Therefore the solution to the semi-discrete scheme is strongly stable. We are interested in whether this strong stability will be preserved after the four-stage fourth-order RK time discretization. In other words, with τ being the time step and

$$u_N^{n+1} = P_4(\tau L_N)u_N^n = \left(I + \tau L_N + \frac{1}{2}(\tau L_N)^2 + \frac{1}{6}(\tau L_N)^3 + \frac{1}{24}(\tau L_N)^4\right)u_N^n,$$

we wonder whether $\|u_N^{n+1}\| \leq \|u_N^n\|$, or equivalently, whether the operator norm of $P_4(\tau L_N)$ is bounded by 1. If this is true, the stability extends to general semi-bounded systems $\partial_t u = L(x,t,\partial_x)u$ by using a standard argument.

For L_N being a scalar and τ being sufficiently small, the proposition above can be justified by analyzing the region of absolute stability, see Chapter IV.2 in [41], for

example. A natural attempt is to apply the eigenvalue analysis to extend the result to systems. However, this technique can only be used for normal matrices L_N, namely $L_N L_N^\top = L_N^\top L_N$. If re-norming is allowed, the method also covers diagonalizable L_N. (But it would still be useless if the diagonalizing matrix is ill-conditioned, see [54].) While for general systems, the naive eigenvalue analysis may end up with misleading results. We refer to Chapter 17.1 in [45] for a specific example.

The analysis for the general systems initiates from the coercive problems, namely $L_N + L_N^\top \leq -\eta L_N^\top L_N$ for some positive constant η. In [54], Levy and Tadmor used the energy method to prove the strong stability of the third-order and the fourth-order RK schemes under the coercivity condition and the time step constraint $\tau \leq c\eta$, where $c = \frac{3}{50}$ for the third-order scheme and $c = \frac{1}{62}$ for the fourth-order scheme. Later in [38] and [73], a simpler proof was provided, with the time step constraint been significantly relaxed to $c = 1$, which extends to general s-stage s-th order linear Runge-Kutta schemes. The authors in [38] pointed out that the coercivity condition implies the strong stability of the forward Euler scheme when $\tau \leq \eta$. The strong stability of the high-order RK schemes follows from the fact that they can be rewritten as convex combinations of the forward Euler steps. These results coincides with the earlier work on contractivity analysis of the numerical solutions to ODE systems, see [68] and [52]. In their theory for contractivity, or strong stability in our context, a circle condition is assumed, which is essentially equivalent to the strong stability assumption for the forward Euler scheme in [38].

In general, the coercivity condition does not hold for method of lines schemes arising from purely hyperbolic problems. Hence it is still imperative to analyze schemes which only satisfy a semi-negativity condition $L_N + L_N^\top \leq 0$. In [73], for the third-order RK scheme, Tadmor successfully removed the coercivity assumption and proved that the third-order RK scheme is strongly stable if $L_N + L_N^\top \leq 0$ and

$\tau\|L_N\| \leq 1$. But the strong stability for the fourth-order RK method remains open, which is the main issue we are concerned with in this chapter.

We find that, although the strong stability of the fourth-order RK method passes the examination of the scalar equation, and can be extended to normal systems, it does NOT hold in general. More specifically, we have the following counter example in Proposition 2.1.1, whose proof is given in Appendix A.3.

Proposition 2.1.1. *Let* $L_N = -\begin{pmatrix} 1 & 2 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$. *Then* $L_N + L_N^T \leq 0$ *and* $\|P_4(\tau L_N)\| > 1$ *for* τ *sufficiently small.*

Interestingly, despite the negative result on the one-step performance, the strong stability property holds in two time steps. In other words, with the constraint $\tau\|L_N\| \leq c_0$ for some positive constant c_0, $\|P_4(\tau L_N)^2\| \leq 1$ and hence $\|u_N^{n+2}\| \leq \|u_N^n\|$. This is still sufficient to justify the stability after long time integration. It can also be used to obtain the stability $\|u_N^n\| \leq K(t^n)\|u_N^0\|$ for semi-bounded and time-dependent L_N. We will apply the results to study the stability of different spatial discretizations coupled with the fourth-order RK approximation.

The remaining of this chapter is organized as follows. In Section 2.2 we exhibit our main results, and prove the stability of the fourth-order RK method with semi-negative and semi-bounded L_N. The case for L_N dependent on time is also discussed. In Section 2.3, we apply our results to several different spatial discretizations. We specifically focus on Galerkin methods as a complementary of [73], including the spectral Galerkin method and the DG method. Finally, we give our concluding remarks in Section 2.4.

2.2 Stability of the fourth-order RK method

In this section, we analyze the stability of the fourth-order RK schemes. For simplicity, we assume everything to be real, but the approach extends to the complex spaces. To facilitate our later discussion on applications to spatial discretizations based on Galerkin methods, we discuss over a general Hilbert space. We avoid clearly characterizing the method of lines scheme in this general setting. But it will not lead to ambiguities, since it only serves as a motivation to derive the fourth-order RK iteration, and will not be used in the stability analysis. To reduce to ODE systems, one simply sets $\mathcal{V} = \mathbb{R}^m$ and $(\cdot, \cdot)$ to be the usual dot product. We also remark that, by using the inner product $(u, v)_H = u^T H v$ with H being a symmetric positive definite matrix, $L_N + L_N^\top \leq 0$ is equivalent to $L_N^T H + H L_N \leq 0$, which is consistent with the assumption in equation (6) of [73].

2.2.1 Notations and the main results

We denote by $\mathcal{V}$ a real Hilbert space equipped with the inner product $(\cdot, \cdot)$. The induced norm is defined as $\|\cdot\| = \sqrt{(\cdot, \cdot)}$. Consider a method of lines scheme defined on $\mathcal{V}$.

$$\partial_t u_N = L_N u_N, \tag{2.1}$$

where $u_N(\cdot, t) \in \mathcal{V}$ and L_N is a bounded linear operator on $\mathcal{V}$. Suppose L_N is independent of t, then the four-stage fourth-order RK approximation can be written

as

$$u_N^{n+1} = P_4(\tau L_N)u_N^n, \qquad P_4(\tau L_N) = I + \tau L_N + \frac{1}{2}(\tau L_N)^2 + \frac{1}{6}(\tau L_N)^3 + \frac{1}{24}(\tau L_N)^4. \tag{2.2}$$

When $L_N = L_N(t)$ depends on time, we denote by

$$u_N^{n+1} = R_\tau(t)u_N^n \tag{2.3}$$

and the specific form of $R_\tau(t)$ will be introduced latter.

For simplicity, we drop all the subscripts N in the remaining parts of the section.

Given an operator L, the operator norm is defined as $\|L\| = \sup_{\|v\|=1} \|Lv\|$. We denote by $\mathcal{B}(\mathcal{V}) = \{L : \|L\| < +\infty\}$ the collection of bounded linear operators on $\mathcal{V}$. For any $L \in \mathcal{B}(\mathcal{V})$, there exists a unique $L^\top \in \mathcal{B}(\mathcal{V})$ such that $(Lw, v) = (w, L^\top v)$. $L^\top$ is referred as the adjoint of L.

If $(v, (L + L^\top)v) \leq 0$, $\forall v \in \mathcal{V}$, denoted by $L + L^\top \leq 0$, then we say L is semi-negative. A special case is when $L + L^\top = 0$, L is referred to be skew-symmetric. If $L + L^\top \leq 2\mu I$ for some positive number μ, namely $L - \mu I$ is semi-negative, then we say L is semi-bounded.

We denote by $[w, v] = -(w, (L + L^\top)v)$ and $[\![w]\!] = \sqrt{[w, w]}$. $[w, v]$ is a bilinear form on $\mathcal{V}$. For $L + L^\top \leq 0$, $[w, v]$ is a semi-inner-product, and one has the Cauchy-Schwarz inequality, $[w, v] \leq [\![w]\!][\![v]\!]$.

To simplify our notation, we define $\mathcal{L} = \tau L$. This notation will be used in the intermediate lemmas and the proofs. For clearness, we restore the notation τL in our main results.

Here we list our stability results on the fourth-order RK approximation. The details will be given in Theorem 2.2.1, Theorem 2.2.2, Theorem 2.2.3, Corollary 2.2.1 and Corollary 2.2.2 respectively. In the following statements, c_0 and K refer to some positive real numbers.

Our main theorem states that

(i) Suppose $L + L^\top \leq 0$. (2.2) *is strongly stable in two steps.* $\|u^{n+2}\| \leq \|u^n\|$, *if* $\tau\|L\| \leq c_0$.

As a corollary of (1), the stability for semi-bounded and time-dependent operator L have also been proved by using a perturbation analysis and a frozen-coefficient argument.

(ii) Suppose $L + L^\top \leq 2\mu I$. (2.2) *is stable,* $\|u^n\| \leq K(t^n)\|u^0\|$, *if* $\tau \leq \frac{c_0}{\|L\|+\mu}$.

(iii) Suppose $L = L(t)$ *satisfies certain regularity assumptions, and* $L+L^\top \leq 2\mu I$. *Then* (2.3) *is stable,* $\|u^n\| \leq K(t^n)\|u^0\|$, *under an appropriate time step constraint.*

As we know, according to the eigenvalue analysis, for normal and semi-negative L, the one-step strong stability will hold under a proper time-step restriction. We also use the energy argument to prove the one-step results, as corollaries of the two-step stability analysis.

(iv) Suppose $L+L^\top = 0$. (2.2) *is strongly stable,* $\|u^{n+1}\| \leq \|u^n\|$, *if* $\tau\|L\| \leq 2\sqrt{2}$.

(v) Suppose $L + L^\top \leq 0$, $LL^\top = L^\top L$. (2.2) *is strongly stable,* $\|u^{n+1}\| \leq \|u^n\|$, *if* $\tau\|L\| \leq c_0$.

For simplicity, we assume the time steps to be uniform in (i), (ii) and (iii). However, these results are essentially built on the two-step performance of the RK time integrator. One only needs the time steps to be uniform for every two steps, namely, $\tau_{2k+1} = \tau_{2k+2}$. As for (iv) and (v), they fit for general time step sizes.

2.2.2 An energy equality

The first thing we would do is to derive an energy equality, which would be useful in understanding the subtle change of the norm of the solution after one time step. To this end, we introduce several identities in Lemma 2.2.1, and then use them to derive the equality in Lemma 2.2.2.

Lemma 2.2.1.

$$
\begin{aligned}
(\mathcal{L}v, v) &= -\frac{\tau}{2}[\![v]\!]^2, && (2.4)\\
(\mathcal{L}^2v, v) &= -\|\mathcal{L}v\|^2 - \tau[\mathcal{L}v, v], && (2.5)\\
(\mathcal{L}^3v, v) &= \frac{\tau}{2}[\![\mathcal{L}v]\!]^2 - \tau[\mathcal{L}^2v, v], && (2.6)\\
(\mathcal{L}^4v, v) &= \|\mathcal{L}^2v\|^2 + \tau[\mathcal{L}^2v, \mathcal{L}v] - \tau[\mathcal{L}^3v, v]. && (2.7)
\end{aligned}
$$

Proof. (1) $(\mathcal{L}v, v) = \frac{1}{2}(\mathcal{L}v, v) + \frac{1}{2}(v, \mathcal{L}v) = \frac{1}{2}(v, (\mathcal{L} + \mathcal{L}^\top)v) = -\frac{\tau}{2}[\![v]\!]^2$.

(2) $(\mathcal{L}^2v, v) = (\mathcal{L}v, \mathcal{L}^\top v) = -(\mathcal{L}v, \mathcal{L}v) + (\mathcal{L}v, (\mathcal{L} + \mathcal{L}^\top)v) = -\|\mathcal{L}v\|^2 - \tau[\mathcal{L}v, v]$.

(3) $(\mathcal{L}^3v, v) = (\mathcal{L}^2v, \mathcal{L}^\top v) = -(\mathcal{L}(\mathcal{L}v), \mathcal{L}v) + (\mathcal{L}^2v, (\mathcal{L} + \mathcal{L}^\top)v) = \frac{\tau}{2}[\![\mathcal{L}v]\!]^2 - \tau[\mathcal{L}^2v, v]$, where we have used (2.4) in the last equality.

(4) $(\mathcal{L}^4v, v) = (\mathcal{L}^3v, \mathcal{L}^\top v) = -(\mathcal{L}^2(\mathcal{L}v), \mathcal{L}v) + (\mathcal{L}^3v, (\mathcal{L} + \mathcal{L}^\top)v) = \|\mathcal{L}^2v\|^2 + \tau[\mathcal{L}^2v, \mathcal{L}v] - \tau[\mathcal{L}^3v, v]$, where (2.5) is used in the last equality. □

Lemma 2.2.2 (Energy equality)**.**

$$\|u^{n+1}\|^2 - \|u^n\|^2 = Q_1(u^n),$$

where

$$Q_1(u^n) = \frac{1}{576}\|\mathcal{L}^4 u^n\|^2 - \frac{1}{72}\|\mathcal{L}^3 u^n\|^2 + \tau \sum_{i,j=0}^{3} \alpha_{ij}[\mathcal{L}^i u^n, \mathcal{L}^j u^n], \tag{2.8}$$

and

$$A = (\alpha_{ij})_{4\times 4} = -\begin{pmatrix} 1 & 1/2 & 1/6 & 1/24 \\ 1/2 & 1/3 & 1/8 & 1/24 \\ 1/6 & 1/8 & 1/24 & 1/48 \\ 1/24 & 1/24 & 1/48 & 1/144 \end{pmatrix}.$$

Proof. Taking an inner product of (2.2) with u^n, one has

$$\frac{1}{2}\|u^{n+1}\|^2 - \frac{1}{2}\|u^n\|^2 - \frac{1}{2}\|u^{n+1} - u^n\|^2 = ((\mathcal{L} + \frac{1}{2}\mathcal{L}^2 + \frac{1}{6}\mathcal{L}^3 + \frac{1}{24}\mathcal{L}^4)u^n, u^n).$$

Applying (2.4) - (2.7) with $v = u^n$, we obtain

$$\begin{aligned}\|u^{n+1}\|^2 - \|u^n\|^2 = \|u^{n+1} - u^n\|^2 - \|\mathcal{L}u^n\|^2 + \frac{1}{12}\|\mathcal{L}^2 u^n\|^2 - \tau[\![u^n]\!]^2 - \tau[\mathcal{L}u^n, u^n] \\ + \frac{\tau}{6}[\![\mathcal{L}u^n]\!]^2 - \frac{\tau}{3}[\mathcal{L}^2 u^n, u^n] + \frac{\tau}{12}[\mathcal{L}^2 u^n, \mathcal{L}u^n] - \frac{\tau}{12}[\mathcal{L}^3 u^n, u^n].\end{aligned} \tag{2.9}$$

Using (2.4) - (2.6) with $v = \mathcal{L}u^n$, the first two terms on the right can be expanded.

$$\begin{aligned}\|u^{n+1} - u^n\|^2 - \|\mathcal{L}u^n\|^2 &= (u^{n+1} - u^n - \mathcal{L}u^n, u^{n+1} - u^n + \mathcal{L}u^n) \\ &= \|u^{n+1} - u^n - \mathcal{L}u^n\|^2 + ((\frac{1}{2}\mathcal{L}^2 + \frac{1}{6}\mathcal{L}^3 + \frac{1}{24}\mathcal{L}^4)u^n, 2\mathcal{L}u^n) \\ &= \|u^{n+1} - u^n - \mathcal{L}u^n\|^2 - \frac{1}{3}\|\mathcal{L}^2u^n\|^2 - \frac{\tau}{2}[\![\mathcal{L}u^n]\!]^2 \\ &\quad - \frac{\tau}{3}[\mathcal{L}^2u^n, \mathcal{L}u^n] + \frac{\tau}{24}[\![\mathcal{L}^2u^n]\!]^2 - \frac{\tau}{12}[\mathcal{L}^3u^n, \mathcal{L}u^n].\end{aligned} \tag{2.10}$$

Substitute (2.10) into (2.9), one has

$$\begin{aligned}\|u^{n+1}\|^2 - \|u^n\|^2 = \|u^{n+1} - u^n - \mathcal{L}u^n\|^2 &- \frac{1}{4}\|\mathcal{L}^2u^n\|^2 - \tau[\![u^n]\!]^2 \\ &- \tau[\mathcal{L}u^n, u^n] - \frac{\tau}{3}[\![\mathcal{L}u^n]\!]^2 - \frac{\tau}{3}[\mathcal{L}^2u^n, u^n] - \frac{\tau}{4}[\mathcal{L}^2u^n, \mathcal{L}u^n] \\ &- \frac{\tau}{12}[\mathcal{L}^3u^n, u^n] + \frac{\tau}{24}[\![\mathcal{L}^2u^n]\!]^2 - \frac{\tau}{12}[\mathcal{L}^3u^n, \mathcal{L}u^n].\end{aligned} \tag{2.11}$$

Similar as before, one can use (2.4) and (2.5) with $v = \mathcal{L}^2u^n$ to calculate the difference of square terms on the right.

$$\begin{aligned}&\|u^{n+1} - u^n - \mathcal{L}u^n\|^2 - \frac{1}{4}\|\mathcal{L}^2u^n\|^2 \\ =&(u^{n+1} - u^n - \mathcal{L}u^n - \frac{1}{2}\mathcal{L}^2u^n, u^{n+1} - u^n - \mathcal{L}u^n + \frac{1}{2}\mathcal{L}^2u^n) \\ =&\|u^{n+1} - u^n - \mathcal{L}u^n - \frac{1}{2}\mathcal{L}^2u^n\|^2 + ((\frac{1}{6}\mathcal{L}^3 + \frac{1}{24}\mathcal{L}^4)u^n, \mathcal{L}^2u^n) \\ =&\|u^{n+1} - u^n - \mathcal{L}u^n - \frac{1}{2}\mathcal{L}^2u^n\|^2 - \frac{1}{24}\|\mathcal{L}^3u^n\|^2 - \frac{\tau}{12}[\![\mathcal{L}^2u^n]\!]^2 - \frac{\tau}{24}[\mathcal{L}^3u^n, \mathcal{L}^2u^n].\end{aligned} \tag{2.12}$$

Once again, we plug (2.12) into (2.11) to get

$$\begin{aligned}\|u^{n+1}\|^2 - \|u^n\|^2 = &\|u^{n+1} - u^n - \mathcal{L}u^n - \frac{1}{2}\mathcal{L}^2u^n\|^2 - \frac{1}{24}\|\mathcal{L}^3u^n\|^2 \\ &- \tau[\![u^n]\!]^2 - \tau[\mathcal{L}u^n, u^n] - \frac{\tau}{3}[\![\mathcal{L}u^n]\!]^2 - \frac{\tau}{3}[\mathcal{L}^2u^n, u^n] - \frac{\tau}{4}[\mathcal{L}^2u^n, \mathcal{L}u^n] \\ &- \frac{\tau}{12}[\mathcal{L}^3u^n, u^n] - \frac{\tau}{24}[\![\mathcal{L}^2u^n]\!]^2 - \frac{\tau}{12}[\mathcal{L}^3u^n, \mathcal{L}u^n] - \frac{\tau}{24}[\mathcal{L}^3u^n, \mathcal{L}^2u^n].\end{aligned}$$

Finally, note that

$$\begin{aligned}&\|u^{n+1} - u^n - \mathcal{L}u^n - \frac{1}{2}\mathcal{L}^2u^n\|^2 - \frac{1}{36}\|\mathcal{L}^3u^n\|^2 \\ =&\|u^{n+1} - u^n - \mathcal{L}u^n - \frac{1}{2}\mathcal{L}^2u^n - \frac{1}{6}\mathcal{L}^3u^n\|^2 + (\frac{1}{24}\mathcal{L}^4u^n, \frac{1}{3}\mathcal{L}^3u^n) \\ =&\frac{1}{576}\|\mathcal{L}^4u^n\|^2 - \frac{\tau}{144}[\![\mathcal{L}^3u^n]\!]^2.\end{aligned}$$

Therefore

$$\begin{aligned}Q_1(u^n) = &\frac{1}{576}\|\mathcal{L}^4u^n\|^2 - \frac{1}{72}\|\mathcal{L}^3u^n\|^2 - \tau[\![u^n]\!]^2 - \tau[\mathcal{L}u^n, u^n] - \frac{\tau}{3}[\![\mathcal{L}u^n]\!]^2 \\ &- \frac{\tau}{3}[\mathcal{L}^2u^n, u^n] - \frac{\tau}{4}[\mathcal{L}^2u^n, \mathcal{L}u^n] - \frac{\tau}{12}[\mathcal{L}^3u^n, u^n] - \frac{\tau}{24}[\![\mathcal{L}^2u^n]\!]^2 \\ &- \frac{\tau}{12}[\mathcal{L}^3u^n, \mathcal{L}u^n] - \frac{\tau}{24}[\mathcal{L}^3u^n, \mathcal{L}^2u^n] - \frac{\tau}{144}[\![\mathcal{L}^3u^n]\!]^2,\end{aligned}$$

which can be rewritten as (2.8). □

According to Lemma 2.2.2, the energy change $Q_1(u^n)$ consists of two parts, the numerical dissipation $\frac{1}{576}\|\mathcal{L}^4u^n\|^2 - \frac{1}{72}\|\mathcal{L}u^n\|^2$ and the generalized quadratic form $\tau\sum_{i,j=0}^{4}\alpha_{ij}[\mathcal{L}u^i, \mathcal{L}u^j]$. When L is skew-symmetric, the quadratic form is simply 0. Hence one can obtain the one-step strong stability.

Corollary 2.2.1. *Suppose $L+L^\top = 0$, then the fourth-order RK approximation (2.2) to the method of lines scheme (2.1) is strongly stable, $\|u^{n+1}\| \le \|u^n\|$, if $\tau\|L\| \le 2\sqrt{2}$.*

Proof. $L + L^\top = 0$ implies $[w, v] = 0, \forall v, w$. Hence when $\tau\|L\| \le 2\sqrt{2}$,

$$\begin{aligned} \|u^{n+1}\|^2 - \|u^n\|^2 = Q_1(u^n) &= \frac{1}{576}\|\mathcal{L}^4 u^n\|^2 - \frac{1}{72}\|\mathcal{L}^3 u^n\|^2 \\ &\le (\frac{1}{576}\|\mathcal{L}\|^2 - \frac{1}{72})\|\mathcal{L}^3 u^n\|^2 \le 0. \end{aligned}$$

□

2.2.3 Stability for semi-negative L

According to Lemma 2.2.2, for non-skew-symmetric L, we would need to absorb the quadratic form with the help of the numerical dissipation. One can see that the high-order terms $\mathcal{L}^i u^n$ with $i \ge 3$ are easy to control. But there are no other terms to bound u^n, $\mathcal{L}u^n$ and $\mathcal{L}^2 u^n$. Our only hope is that $\tau \sum_{i,j=0}^{2} \alpha_{ij}[\mathcal{L}u^i, \mathcal{L}u^j]$ itself is negative-definite. Lemma 2.2.3 indicates that, to check the negativity of the generalized quadratic form, one only needs to examine the coefficient matrix, as that for the polynomials. In Lemma 2.2.4, we prove a sufficient condition for strong stability, once $\sum_{i,j=0}^{2} \alpha_{ij}[\mathcal{L}u^i, \mathcal{L}u^j]$ is negative-definite, the energy change will be non-positive when $\|\mathcal{L}\|$ is sufficiently small.

Lemma 2.2.3. *Suppose L is semi-negative. Let $M = (m_{ij})$ be a real symmetric semi-negative definite matrix, then $\sum_{i,j} m_{ij}[u_i, u_j] \le 0$.*

Proof. Suppose $M = S^T \Lambda S$, where $S = (s_{ij})$ is an orthogonal matrix and Λ is a diagonal matrix with the diagonal elements $\lambda_i \le 0$. Note that $m_{ij} = \sum_k \lambda_k s_{ki} s_{kj}$.

Therefore

$$\sum_{i,j} m_{ij}[u_i,u_j] = \sum_{i,j}(\sum_k \lambda_k s_{ki}s_{kj})[u_i,u_j] = \sum_k \lambda_k(\sum_{i,j} s_{ki}s_{kj}[u_i,u_j])$$
$$= \sum_k \lambda_k[\sum_i s_{ki}u_i, \sum_j s_{kj}u_j] = \sum_k \lambda_k [\![\sum_i s_{ki}u_i]\!]^2 \le 0.$$

□

Lemma 2.2.4. *Suppose L is semi-negative. Let*

$$Q(u) = \alpha\|\mathcal{L}^3 u\|^2 + \tau \sum_{i,j=0}^{a} \alpha_{ij}[\mathcal{L}^i u, \mathcal{L}^j u],$$

where $\alpha_{ij} = \alpha_{ji}$ and $a \ge 2$. If $\alpha < 0$ and $A_2 = \begin{pmatrix} \alpha_{00} & \alpha_{01} & \alpha_{02} \\ \alpha_{10} & \alpha_{11} & \alpha_{12} \\ \alpha_{20} & \alpha_{21} & \alpha_{22} \end{pmatrix}$ is negative-definite, then there exists $c_0 > 0$, such that $Q(u) \le 0$ as long as $\|\mathcal{L}\| \le c_0$.

Proof. Let $-\varepsilon$ be the largest eigenvalue of the matrix A_2. Then $A_2 + \varepsilon I$ is semi-negative definite. According to Lemma 2.2.3,

$$\begin{aligned} \sum_{i,j=0}^{2} \alpha_{ij}[\mathcal{L}^i u, \mathcal{L}^j u] &= \sum_{i,j=0}^{2} (\alpha_{ij} + \varepsilon\delta_{ij})[\mathcal{L}^i u, \mathcal{L}^j u] - \varepsilon([\![u]\!]^2 + [\![\mathcal{L}u]\!]^2 + [\![\mathcal{L}^2 u]\!]^2) \\ &\le -\varepsilon([\![u]\!]^2 + [\![\mathcal{L}u]\!]^2 + [\![\mathcal{L}^2 u]\!]^2). \end{aligned} \tag{2.13}$$

By Cauchy-Schwartz and the arithmetic-geometric mean inequality,

$$\begin{aligned} \sum_{\max(i,j)>2}^{a} \alpha_{ij}[\mathcal{L}^i u, \mathcal{L}^j u] &\le \varepsilon([\![u]\!]^2 + [\![\mathcal{L}u]\!]^2 + [\![\mathcal{L}^2 u]\!]^2) + \sum_{i=3}^{a} \tilde{\alpha}_i [\![\mathcal{L}^i u]\!]^2 \\ &\le \varepsilon([\![u]\!]^2 + [\![\mathcal{L}u]\!]^2 + [\![\mathcal{L}^2 u]\!]^2) + \sum_{i=3}^{a} 2\tilde{\alpha}_i \|L\|\|\mathcal{L}\|^{2(i-3)}\|\mathcal{L}^3 u\|^2, \end{aligned} \tag{2.14}$$

where we have used the fact $[\![v]\!]^2 \leq 2\|L\|\|v\|^2$ in the last inequality, and $\tilde{\alpha}_i$ are some non-negative constants depending on ε and α_{ij}. Using (2.13) and (2.14), if $\|\mathcal{L}\| = \tau\|L\| \leq c_0$, one has

$$Q(u) \leq (\alpha + \sum_{i=3}^{a} 2\tilde{\alpha}_i c_0^{2(i-3)+1})\|\mathcal{L}^3 u\|^2.$$

Since α is negative, $Q(u)$ is non-positive as long as c_0 is sufficiently small. □

Unfortunately, $Q_1(u^n)$ does not satisfy the assumptions in Lemma 2.2.4. Actually, the three eigenvalues of $A_2 = -\begin{pmatrix} 1 & 1/2 & 1/6 \\ 1/2 & 1/3 & 1/8 \\ 1/6 & 1/8 & 1/24 \end{pmatrix}$ are 0.00560618, −0.0793266 and −1.30128. This motivates us to disprove the strong stability of the fourth-order RK method and we end up with the counter example in Proposition 2.1.1. So we turn to seek the power-boundedness of $P_4(\mathcal{L})$. Surprisingly, though $Q_1(u^n)$ itself fails to pass Lemma 2.2.4, $Q_1(u^{n+1}) + Q_1(u^n)$ succeeds. Hence we obtain the two-step strong stability in Theorem 2.2.1.

Theorem 2.2.1 (Two-step strong stability for semi-negative L)**.** *Suppose $L + L^\top \leq 0$, then the fourth-order RK approximation (2.2) to the method of lines scheme (2.1) is strongly stable in two steps, $\|u^{n+2}\| \leq \|u^n\|$, if $\tau\|L\| \leq c_0$.*

Proof. Let $u^{n+2} = P_4(\mathcal{L})u^{n+1}$. Then

$$\|u^{n+2}\|^2 - \|u^{n+1}\|^2 = Q_1(u^{n+1}).$$

Substitute (2.2) into $Q_1(u^{n+1})$ and rewrite the quadratic form in terms of u^n. By

direct calculation, one can obtain

$$Q_1(u^{n+1}) = \frac{1}{576}\|\mathcal{L}^4 u^{n+1}\|^2 - \frac{1}{72}\|\mathcal{L}^3 u^{n+1}\|^2 + \tau \sum_{i,j=0}^{7} \tilde{\alpha}_{ij}[\mathcal{L}^i u^n, \mathcal{L}^j u^n],$$

where

$$\tilde{A}_2 = (\tilde{\alpha}_{ij})_{3\times 3} = -\begin{pmatrix} 1 & 3/2 & 7/6 \\ 3/2 & 7/3 & 15/8 \\ 7/6 & 15/8 & 37/24 \end{pmatrix}.$$

The complete coefficient matrix $\tilde{A}$ is given in Appendix A.4. While according to Lemma 2.2.2,

$$Q_1(u^n) = \frac{1}{576}\|\mathcal{L}^4 u^n\|^2 - \frac{1}{72}\|\mathcal{L}^3 u^n\|^2 + \tau \sum_{i,j=0}^{3} \hat{\alpha}_{ij}[\mathcal{L}^i u^n, \mathcal{L}^j u^n],$$

where

$$\hat{A}_2 = (\hat{\alpha}_{ij})_{3\times 3} = -\begin{pmatrix} 1 & 1/2 & 1/6 \\ 1/2 & 1/3 & 1/8 \\ 1/6 & 1/8 & 1/24 \end{pmatrix}.$$

When $\|\mathcal{L}\| \le 2$,

$$\frac{1}{576}\|\mathcal{L}^4 u^{n+1}\|^2 - \frac{1}{72}\|\mathcal{L}^3 u^{n+1}\|^2 \le 0,$$

$$\frac{1}{576}\|\mathcal{L}^4 u^n\|^2 - \frac{1}{72}\|\mathcal{L}^3 u^n\|^2 \le -\frac{1}{144}\|\mathcal{L}^3 u^n\|^2.$$

Hence

$$\begin{aligned} \|u^{n+2}\|^2 - \|u^n\|^2 &= Q_1(u^{n+1}) + Q_1(u^n) \\ &\le -\frac{1}{144}\|\mathcal{L}^3 u^n\|^2 + \tau \sum_{i,j=0}^{7} \alpha_{ij}[\mathcal{L}^i u^n, \mathcal{L}^j u^n], \end{aligned}$$

where

$$A_2 = (\alpha_{ij})_{3\times 3} = \tilde{A}_2 + \hat{A}_2 = -\begin{pmatrix} 2 & 2 & 4/3 \\ 2 & 8/3 & 2 \\ 4/3 & 2 & 19/12 \end{pmatrix}.$$

The three eigenvalues of A_2 are -5.73797, -0.499093 and -0.0129329. Hence A_2 is negative definite, and one could complete the proof by applying Lemma 2.2.4. □

Remark 2.2.1. *There is another interpretation of the two-step strong stability. We treat u^{n+1} as a intermediate stage and consider u^{n+2} as the solution of an eight-stage fourth-order RK scheme composed by two four-stage fourth-order RK time stepping. Then Theorem 2.2.1 states this new RK scheme is strongly stable. This coincides with Remark 3 of Section 4 in [73], where the author conjectured the fourth-order RK scheme can preserve the strong stability if allowing additional stages.*

Remark 2.2.2. *Here we only focus on the existence of the constant c_0 instead of giving a specific estimate. The argument above is far from sharp. To obtain a reasonable bound for c_0, one should expand $\frac{1}{576}\|\mathcal{L}^4 u^{n+1}\|^2 - \frac{1}{72}\|\mathcal{L}^3 u^{n+1}\|^2$ to separate the quadratic form involving $[\cdot,\cdot]$, and carefully go through all the constants. To this end, a generalized version of Lemma 2.2.1 is also needed. We leave these to interested readers.*

Realizing the fact that, for a normal operator G, $\|G^2\| \le 1$ implies $\|G\| \le 1$, we prove the one-step strong stability for normal and semi-negative L.

Corollary 2.2.2 (Strong stability for normal and semi-negative L)**.** *Suppose $LL^\top = L^\top L$ and $L + L^\top \le 0$, then the fourth-order RK approximation (2.2) to the method of lines scheme (2.1) is strongly stable, $\|u^{n+1}\| \le \|u^n\|$, if $\tau\|L\| \le c_0$ for some constant c_0.*

Proof. According to Theorem 2.2.1, we have $\|P_4(\mathcal{L})^2\| \le 1$ if $\|\mathcal{L}\| \le c_0$. When $\mathcal{L}$ is

normal, $P_4(\mathcal{L})$ is also normal, $P_4(\mathcal{L})P_4(\mathcal{L})^\top = P_4(\mathcal{L})^\top P_4(\mathcal{L})$. Hence for any $u \in \mathcal{V}$,

$$\begin{aligned}\|P_4(\mathcal{L})u\|^2 &= (P_4(\mathcal{L})u, P_4(\mathcal{L})u) = (u, P_4(\mathcal{L})^\top P_4(\mathcal{L})u)\\ &= (u, P_4(\mathcal{L})P_4(\mathcal{L})^\top u) = \|P_4(\mathcal{L})^\top u\|^2.\end{aligned}$$

Therefore, under the same constraints, $\|\mathcal{L}\| \le c_0$, one has

$$\begin{aligned}\|P_4(\mathcal{L})\|^2 &= \sup_{\|u\|=1} \|P_4(\mathcal{L})u\|^2 = \sup_{\|u\|=1} (u, P_4(\mathcal{L})^\top P_4(\mathcal{L})u)\\ &\le \sup_{\|u\|=1} \|P_4(\mathcal{L})^\top P_4(\mathcal{L})u\| = \sup_{\|u\|=1} \|P_4(\mathcal{L})^2 u\| = \|P_4(\mathcal{L})^2\| \le 1.\end{aligned}$$

In other words, $\|u^{n+1}\| \le \|u^n\|$ if $\|\mathcal{L}\| \le c_0$. □

Remark 2.2.3. *By using scalar eigenvalue analysis, the sharp bound of c_0 in Corollary 2.2.2 is $2\sqrt{2}$, as that in Corollary 2.2.1.*

2.2.4 Stability for semi-bounded L

We then apply the perturbation analysis to obtain the stability for semi-bounded L. In the following proof, one actually needs to assume the time steps satisfies $\tau_{2k+1} = \tau_{2k+2}$ in order to apply Theorem 2.2.1. For simplicity, we use a uniform time stepping, namely $\tau_k = \tau$. The same assumption will be used in Section 2.2.5.

Theorem 2.2.2. *Suppose $L + L^\top \le 2\mu I$, then the fourth-order RK approximation (2.2) to the method of lines scheme (2.1) is stable, $\|u^n\| \le K(t^n)\|u^0\|$, if $\tau \le \frac{c_0}{\|L\|+\mu}$ for some constant c_0. Here $K(t^n)$ is a constant depending on μ and t^n.*

Proof. Note that

$$P_4(\tau L)^2 = P_4(\tau(L-\mu I) + \tau\mu I)^2 = P_4(\tau(L-\mu I))^2 + \tau\mu P(\tau\mu, \tau(L-\mu I)),$$

where $P(s,G)$ is of the form $P(s,G)=\sum_{i,j\geq 0,\ i+j\leq 7}\alpha_{ij}s^iG^j$. Since $L+L^\top\leq 2\mu I$, $(L-\mu I)+(L-\mu I)^\top\leq 0$. According to Theorem 2.2.1, $\|P_4(\tau(L-\mu I))^2\|\leq 1$ if $\tau\|L-\mu I\|\leq c_0$. Under a stricter restriction $\tau\leq\frac{c_0}{\|L\|+\mu}$, one also has

$$\|P(\tau\mu,\tau(L-\mu I))\|\leq\sum_{i,j\geq 0,\ i+j\leq 7}|\alpha_{ij}||\tau\mu|^i\|\tau(L-\mu I)\|^j\leq\sum_{i,j\geq 0,\ i+j\leq 7}|\alpha_{ij}|c_0^{i+j}.$$

Therefore for $\tau\leq\frac{c_0}{\|L\|+\mu}$, $\|P_4(\tau L)^2\|\leq 1+c\mu\tau$, where $c=\sum_{i,j\geq 0,\ i+j\leq 7}|\alpha_{ij}|c_0^{i+j}$. With this one-step estimate, one has

$$\|u^{2k}\|\leq\|P_4(\tau L)^2\|^k\|u^0\|\leq(1+c\mu\tau)^{\frac{t^{2k}}{2\tau}}\|u^0\|\leq K(t^{2k})\|u^0\|,$$

and

$$\|u^{2k+1}\|\leq\|P_4(\tau L)\|\|u^{2k}\|\leq P_4(c_0)K(t^{2k})\|u^0\|,$$

which prove the stability. □

2.2.5 Stability for semi-bounded and time-dependent L

We conclude this section by extending the results to L dependent on time. One should note in this case, the fourth-order RK scheme can no longer be written as the truncated exponential in (2.2). Also different fourth-order RK time integrators are no longer equivalent. One can use the classical four-stage fourth-order RK scheme in (2.15) as an example for our stability analysis. However, our proof does not rely on this specific form and can be used for general cases. The classical four-stage

fourth-order RK scheme is

$$\begin{aligned} k_1 &= L(t^n)u^n, \\ k_2 &= L(t^{n+\frac{1}{2}})(u^n + \frac{\tau}{2}k_1), \\ k_3 &= L(t^{n+\frac{1}{2}})(u^n + \frac{\tau}{2}k_2), \\ k_4 &= L(t^{n+1})(u^n + \tau k_3), \\ u^{n+1} &= u^n + \frac{\tau}{6}(k_1 + 2k_2 + 2k_3 + k_4), \end{aligned} \tag{2.15}$$

where $t^{n+\frac{1}{2}} = t^n + \frac{\tau}{2}$. As a short hand notation, let us denote by

$$u^{n+1} = R_\tau(t^n)u^n,$$

where

$$\begin{aligned} R_\tau(t^n) =& I + \frac{\tau}{6}\left(L(t^n) + 4L(t^{n+\frac{1}{2}}) + L(t^{n+1})\right) \\ &+ \frac{\tau^2}{6}\left(L(t^n)L(t^{n+\frac{1}{2}}) + L(t^{n+\frac{1}{2}})^2 + L(t^{n+\frac{1}{2}})L(t^{n+1})\right) \\ &+ \frac{\tau^3}{12}\left(L(t^n)L(t^{n+\frac{1}{2}})^2 + L(t^{n+\frac{1}{2}})^2L(t^{n+1})\right) \\ &+ \frac{\tau^4}{24}L(t^n)L(t^{n+\frac{1}{2}})^2L(t^{n+1}). \end{aligned} \tag{2.16}$$

The two-step approximation can be written as $u^{n+2} = R_\tau(t^{n+1})R_\tau(t^n)u^n$.

We also assume that L satisfies the Lipschitz continuity condition

$$\|L(t_2) - L(t_1)\| \leq \eta|t_2 - t_1|, \tag{2.17}$$

with $\sup_t \|L(t)\| \leq \eta < +\infty$ for some constant η.

Lemma 2.2.5. *Suppose L satisfies (2.17). Then*

$$\|\prod_{i=1}^{m} L(t+\alpha_i\tau) - (L(t))^m\| \le \sum_{i=1}^{m} |\alpha_i|\eta^m\tau.$$

Proof. We prove by induction. According to assumption (2.17), the lemma holds for $m = 1$. Suppose it holds for m, then

$$\begin{aligned}
&\|\prod_{i=1}^{m+1} L(t+\alpha_i\tau) - (L(t))^{m+1}\| \\
\le&\|\prod_{i=1}^{m+1} L(t+\alpha_i\tau) - L(t)\prod_{i=1}^{m} L(t+\alpha_i\tau)\| + \|L(t)\prod_{i=1}^{m} L(t+\alpha_i\tau) - (L(t))^{m+1}\| \\
\le&\|L(t+\alpha_{m+1}\tau) - L(t)\|\prod_{i=1}^{m}\|L(t+\alpha_i\tau)\| + \|L(t)\|\|\prod_{i=1}^{m} L(t+\alpha_i\tau) - (L(t))^m\| \\
\le&|\alpha_{m+1}|\eta\tau\cdot\eta^m + \eta\cdot(\sum_{i=1}^{m}|\alpha_i|)\eta^m\tau \\
=&\sum_{i=1}^{m+1}|\alpha_i|\eta^{m+1}\tau,
\end{aligned}$$

where we have used the inductive assumption and $\eta > \sup_t \|L(t)\|$ in the second last line. Hence the lemma holds for any positive integer m. □

Theorem 2.2.3. *Suppose $L = L(t)$ and $L + L^\top \le 2\mu I$ for some positive number μ. L satisfies the Lipschitz continuity condition (2.17). Then the four-stage fourth-order RK approximation (2.15) to the method of lines scheme (2.1) is stable, $\|u^n\| \le K(t^n)\|u^0\|$, if $\tau \le \frac{c_0}{\eta+\mu}$. Here $K(t^n)$ is some constant depending on μ and t^n.*

Proof.

$$\begin{aligned}
P_4(\tau L)^2 = I + 2\tau L + 2(\tau L)^2 + \frac{4}{3}(\tau L)^3 + \frac{2}{3}(\tau L)^4 \\
+ \frac{1}{4}(\tau L)^5 + \frac{5}{72}(\tau L)^6 + \frac{1}{72}(\tau L)^7 + \frac{1}{576}(\tau L)^8.
\end{aligned} \tag{2.18}$$

And

$$R_\tau(t^{n+1})R_\tau(t^n) = \sum_{i=0}^{8}\sum_j \alpha_{ij}\prod_{k=1}^{i}(\tau L(t^n+\tilde{\alpha}_{ijk}\tau)), \tag{2.19}$$

with $\{\sum_j \alpha_{ij}\}_{i=0}^8$ take values of 1, 2, 2, $\frac{4}{3}$, $\frac{2}{3}$, $\frac{1}{4}$, $\frac{5}{72}$, $\frac{1}{72}$ and $\frac{1}{576}$ as those in $P_4(\tau L)^2$. (This must be true since when L is independent of time, (2.19) reduces to (2.18).) Applying Lemma 2.2.5, one has

$$\begin{aligned}\|R_\tau(t^{n+1})R_\tau(t^n) - P_4(\tau L)^2\| &= \|\sum_{i=0}^{8}\sum_j \alpha_{ij}\tau^i(\prod_{k=1}^{i} L(t+\tilde{\alpha}_{ijk}\tau) - (L(t))^i)\| \\ &\le \sum_{i=0}^{8}\sum_j |\alpha_{ij}|\tau^i\|\prod_{k=1}^{i} L(t+\tilde{\alpha}_{ijk}\tau) - (L(t))^i\| \\ &\le \tau\sum_{i=0}^{8}\sum_j |\alpha_{ij}|(\sum_k |\tilde{\alpha}_{ijk}|)(\tau\eta)^i = \tilde{c}\tau.\end{aligned}$$

On the other hand, $\|P_4(\tau L)^2\| \le 1 + c\mu\tau$ if $\tau(\eta+\mu) \le c_0$, hence

$$\|R_\tau(t^{n+1})R_\tau(t^n)\| \le \|P_4(\tau L)^2\| + \|R_\tau(t^{n+1})R_\tau(t^n) - P_4(\tau L)^2\| \le 1 + (c\mu + \tilde{c})\tau.$$

As we have done in Theorem 2.2.2, one can prove $\|u^n\| \le K(t^n)\|u^0\|$, if $\tau \le \frac{c_0}{\eta+\mu}$. □

Remark 2.2.4. *For different four-stage fourth-order RK schemes, the values of α_{ij} and $\tilde{\alpha}_{ijk}$ are different. But $\sum_j \alpha_{ij}$ are always the same. Hence the proof above would still work.*

2.3 Applications to hyperbolic problems

In this section, we apply the previous results to several semi-discrete approximations of hyperbolic problems, and justify their stability after coupling with the fourth-order RK time integrator. The time step restriction is also referred as the CFL

condition in this context. In [73], Tadmor has provided detailed examples for spatial discretization based on the nodal-value formulation, including the finite difference method and spectral collocation method. Complementary to [73], we will mainly focus on Galerkin methods, including the global approach, the spectral Galerkin method, and the local approach, the finite element DG method.

The Galerkin methods are spatial discretization techniques based on the weak formulation. Consider the initial value problem,

$$\begin{cases} \partial_t u(x,t) = L(x,t,\partial_x)u(x,t), & (x,t) \in \Omega \times (0,T), \\ u(x,0) = u_0(x), & x \in \Omega. \end{cases}$$

The Galerkin methods seeks a solution of the form $u_N(x,t) = \sum_{k=0}^{N} c_k(t)\phi_k(x)$. Here $\{\phi_k(x)\}_{k=0}^{N}$ forms the basis of a finite dimensional space $\mathcal{V}$ on Ω, equipped with the inner product $(\cdot,\cdot)$. In addition, one requires that

$$\begin{cases} (\partial_t u_N, v_N) = B(u_N, v_N), & \forall v_N \in \mathcal{V}, \\ u_N(x,0) = Pu_0. \end{cases}$$

Here B is a bilinear form on $\mathcal{V}$ with $B(u_N, v_N)$ approximating (Lu_N, v_N), and P is the projection to $\mathcal{V}$.

For spectral Galerkin method, $\mathcal{V}$ is chosen as the span of the trigonometric functions or polynomials over the whole domain. Since the functions in $\mathcal{V}$ are sufficiently smooth, one can directly set $B(u_N, v_N) = (Lu_N, v_N)$. Hence the method can be written as

$$\partial_t u_N = L_N u_N = PLu_N = PLPu_N.$$

It suffices to check whether $L_N = PLP$ satisfies the conditions in Section 2.2 and

examine $\|L_N\|$ to obtain the time step constraint.

For DG method, $\mathcal{V}$ is a piecewise polynomial space based on an appropriate partition of the domain Ω. The associated bilinear form is involved with the so-called numerical flux. We will explain it in the latter part of this section.

2.3.1 Spectral Galerkin method

The following examples are based on Example 3.8 and Example 8.3, 8.4 in [42]. The estimate of $\|L_N\|$ relies on the inverse inequalities on the finite dimensional trigonometric or polynomial spaces, which we refer to [65] for details. For interested readers, we also refer to the same books [42] and [65] for a systematic setup of the spectral methods.

2.3.1.1 Fourier method

Consider the linear problem

$$\partial_t u = a(x,t)\partial_x u + b(x,t)u, \qquad t \in (0,T),\ x \in (0,2\pi), \tag{2.20}$$

with periodic boundary conditions. We assume that the coefficients $a(x,t)$ and $b(x,t)$ are smooth in both x and t, and are periodic with respect to x.

Under these circumstances, $L = a(x,t)\partial_x + b(x,t)I$. Then for any smooth and

periodic functions $v(x)$ and $w(x)$,

$$\begin{aligned}(L^\top v, w) = (v, Lw) = (v, (a\partial_x + b)w) = (-\partial_x(av) + bv, w)\\ = ((-\partial_x a + 2b)\, v, w) - (Lv, w).\end{aligned}$$

Hence $L^\top = (-\partial_x a + 2b)I - L$. Here $(\cdot, \cdot)$ is specified as the L^2 inner product on $[0, 2\pi]$.

In the Fourier Galerkin method, we set $\mathcal{V} = \text{span}\{\sin(kx), \cos(kx)\}_{k=0}^{N}$, and choose $(\cdot, \cdot)$ to be the same L^2 inner product. Noting that P is self-adjoint, one has

$$L_N^\top = PL^\top P = P((-(\partial_x a) + 2b)I - L)P = P(-(\partial_x a) + 2b)P - L_N.$$

Hence

$$L_N + L_N^\top = P(-(\partial_x a) + 2b)P \le (-(\partial_x a) + 2b)I \le (\sup_{x,t} |\partial_x a| + 2 \sup_{x,t} |b|)I,$$

and L_N is semi-bounded.

On the other hand, by using the inverse inequality for trigonometric functions $\|\phi'\| \le N\|\phi\|$, $\forall \phi \in \mathcal{V}$ and the fact $\|P\| \le 1$, one has

$$\|L_N\| \le \|a(\cdot, t)\partial_x + b(\cdot, t)I\| \le \sup_{x,t} |a| N + \sup_{x,t} |b|,$$

and

$$\begin{aligned}\|L_N(t_2) - L_N(t_1)\| \le \|(a(\cdot, t_2) - a(\cdot, t_1))\partial_x + (b(\cdot, t_2) - b(\cdot, t_1))I\|\\ \le (\sup_{x,t} |\partial_t a| N + \sup_{x,t} |\partial_t b|)|t_2 - t_1|.\end{aligned}$$

Hence, the Lipschitz condition holds for $\eta = \max\{\sup_{x,t} |a| N + \sup_{x,t} |b|, \sup_{x,t} |\partial_t a| N +$

$\sup_{x,t}|\partial_t b|\}$. We obtain the following corollary of Theorem 2.2.3.

Corollary 2.3.1. *Consider the Fourier Galerkin approximation of the linear problem (2.20) with the fourth-order RK time discretization,*

$$u_N^{n+1} = R_\tau(t^n)u_N^n, \qquad R_\tau(t^n) \text{ defined in (2.16), with } L_N = P(a(x,t)\partial_x + b(x,t)I)P.$$

This fully discrete scheme is stable,

$$\|u_N^n\| \le K(t^n)\|u_N^0\|,$$

under the CFL condition $\tau(\max\{\sup_{x,t}|a|N + \sup_{x,t}|b|, \sup_{x,t}|\partial_t a|N + \sup_{x,t}|\partial_t b|\} + \frac{1}{2}\sup_{x,t}|\partial_x a| + \sup_{x,t}|b|) \le c_0$, *for some constant* c_0.

Remark 2.3.1. *When a and b are constant, $L_N^\top = 2bI - L_N$. Then $L_N^\top$ commutes with L_N and L_N is normal. Furthermore, if $b \le 0$, $L_N^\top + L_N = 2bI \le 0$, L_N is also semi-negative. One can use Corollary 2.2.2 to prove the one-step strong stability under a sufficiently small time step. In particular, when a is a constant and $b = 0$, L_N is skew-symmetric, we can get a specific estimate of the time step size from Corollary 2.2.1.*

2.3.1.2 Polynomial method

Consider the linear advection equation

$$\partial_t u = \partial_x u, \qquad t \in (0,T),\ x \in (-1,1), \tag{2.21}$$

with the inflow boundary condition $u(1,t) = 0$.

In the polynomial Galerkin method, we set $\mathcal{V} = \{\phi(x) \in \text{span}\{x^k\}_{k=0}^N | \phi(1) = 0\}$

and $(\cdot,\cdot)_w$ is the weighted L^2 inner product, namely $(f,g)_w = \int_{-1}^{1} w(x)f(x)g(x)dx$. The corresponding norm is denoted as $\|\cdot\|_w$. We denote by P_w the projection to $\mathcal{V}$ under $(\cdot,\cdot)_w$.

Different weight functions $w(x)$ correspond to different polynomial methods in the literature. Let us consider the Jacobi method as an example. Here $w(x) = (1+x)^\alpha(1-x)^\beta$ with $\alpha \geq 0$ and $-1 < \beta \leq 0$. When $\alpha = \beta = 0$, $w(x) = 1$, and the method is also referred as the Legendre method.

It can be shown that L_N is semi-negative in this setting. For any $u_N \in \mathcal{V}$, one has

$$\begin{aligned}(u_N, (L_N + L_N^\top)u_N)_w &= (P_w u_N, LP_w u_N)_w + (LP_w u_N, P_w u_N)_w \\ &= 2(u_N, Lu_N)_w = 2\int_{-1}^{1} w u_N \partial_x u_N dx = \int_{-1}^{1} w\partial_x u_N^2 dx \\ &= -\int_{-1}^{1} w' u_N^2 dx - w(-1)u_N(-1,t)^2 \leq -\int_{-1}^{1} w' u_N^2 dx.\end{aligned}$$

For $\alpha > 0$ and $\beta < 0$,

$$\begin{aligned}w'(x) &= \alpha(1+x)^{\alpha-1}(1-x)^\beta - \beta(1+x)^\alpha(1-x)^{\beta-1} \\ &= (1+x)^{\alpha-1}(1-x)^{\beta-1}(\alpha(1-x) - \beta(1+x)) \geq 0.\end{aligned}$$

Similarly, one can prove $w'(x) \geq 0$ for other cases. Hence $L_N + L_N^\top \leq 0$.

Furthermore, we apply the inverse inequality for Jacobi polynomials to obtain $\|L_N\|_w \leq CN^2$ for some constant C. (Refer to Theorem 3.34 in [65].) Therefore, by applying Theorem 2.2.1, one has the following corollary.

Corollary 2.3.2. *Consider the Jacobi Galerkin approximation of the advection equa-*

tion (2.21) with the fourth-order RK time discretization,

$$u_N^{n+1} = P_4(\tau L_N)u_N^n, \qquad L_N = P_w(\partial_x)P_w.$$

This fully discrete scheme is strongly stable in two steps,

$$\|u_N^{n+2}\|_w \leq \|u_N^n\|_w,$$

under the CFL condition $\tau \leq CN^{-2}$, where C depends on the constant in the inverse inequality.

2.3.2 Discontinuous Galerkin method

A detailed discussion of the example in Section 2.3.2.1 can be found in the previous chapter. To be consistent with notations there, we switch the subscript N to h in this section, where h corresponds to the mesh size. Also, we will use the bold fonts for vectors and matrices in this section.

2.3.2.1 Multi-dimensional scalar equation

Consider the linear scalar conservation law

$$\partial_t u(\boldsymbol{x}, t) = \nabla \cdot (\boldsymbol{\beta}(\boldsymbol{x})u(\boldsymbol{x}, t)), \qquad \boldsymbol{x} \in \Omega \subset \mathbb{R}^d, t \in (0, T), \tag{2.22}$$

with the periodic boundary conditions. Here $\boldsymbol{\beta}$ is a smooth function satisfying the divergence-free condition, $\nabla \cdot \boldsymbol{\beta}(\boldsymbol{x}) = 0$.

Suppose $\mathcal{K} = \{K\}$ is a quasi-uniform partition of the domain Ω. We denote by h the largest diameter of the elements. The collection of cell interfaces is denoted by $\mathcal{E}$. Let us define $\mathcal{V} = \{v \in L^2(\Omega) : v|_K \in \mathcal{P}^p(K), \forall K \in \mathcal{K}\}$, where $\mathcal{P}^p(K)$ is the space of polynomials of degree no more than p on K. As for $(\cdot,\cdot)$, we use the L^2 inner product on Ω.

In DG method, one seeks a solution satisfying

$$\int_\Omega \partial_t u_h v_h dx = \sum_K (-\int_K u_h \nabla \cdot (\boldsymbol{\beta} v_h) dx + \int_{\partial K} \widehat{u}_h v_h \boldsymbol{\beta} \cdot \boldsymbol{n} dl), \qquad \forall v_h \in \mathcal{V},$$

where $\widehat{u}_h$ is the numerical flux, approximating the trace of u along the edges. A popular choice is to obey the unwinding principle. In our case, that is $\widehat{u}_h = u_h^+$, where $u_h^+ = \lim_{\varepsilon \to 0^+} u_h(\boldsymbol{x} + \varepsilon \boldsymbol{\beta})$. We also use $u_h^- = \lim_{\varepsilon \to 0^+} u_h(\boldsymbol{x} - \varepsilon \boldsymbol{\beta})$ to represent the trace of u_h from the downwind side. When $\boldsymbol{\beta}(\boldsymbol{x})$ is parallel to the cell interfaces, $u_h^\pm$ are not well-defined. But in this case, $\boldsymbol{\beta}(\boldsymbol{x}) \cdot \boldsymbol{n} = 0$, hence the value of $u_h^\pm$ will not make any difference.

Let us introduce the short hand notation,

$$\mathcal{H}_{\boldsymbol{\beta}}^+(w, v) = \sum_{K \in \mathcal{K}} \big(\int_K w \nabla \cdot (\boldsymbol{\beta} v) dx - \int_{\partial K} w^+ v \boldsymbol{\beta} \cdot \boldsymbol{n} dl \big). \tag{2.23}$$

Then the scheme becomes, find $u_h \in \mathcal{V}$, such that

$$(\partial_t u_h, v_h) = -\mathcal{H}_{\boldsymbol{\beta}}^+(u_h, v_h), \qquad \forall v_h \in \mathcal{V}.$$

$\mathcal{H}_{\boldsymbol{\beta}}^+$ has the following properties. We refer to the previous chapter for details.

Lemma 2.3.1. *For any $w_h, v_h \in \mathcal{V}$, we have*
(1) $\mathcal{H}_{\boldsymbol{\beta}}^+(v_h, v_h) = \frac{1}{2}[\![v_h]\!]_{\boldsymbol{\beta}}^2$, where $[\![v_h]\!]_{\boldsymbol{\beta}} = \sqrt{\sum_{e \in \mathcal{E}} \int_e (v_h^+ - v_h^-)^2 |\boldsymbol{\beta} \cdot \boldsymbol{n}| dl}$.

(2) $|\mathcal{H}^+_{\boldsymbol{\beta}}(w_h, v_h)| \leq Ch^{-1}\|w_h\|\|v_h\|$*, for some constant* C *depending on* $\boldsymbol{\beta}$ *and the constant in the inverse estimate.*

By defining

$$(L_h u_h, v_h) = -\mathcal{H}^+_{\boldsymbol{\beta}}(u_h, v_h), \tag{2.24}$$

we can also rewrite the scheme as

$$\partial_t u_h = L_h u_h.$$

Using Lemma 2.3.1, one has

$$(v_h, (L_h + L_h^\top)v_h) = 2(L_h v_h, v_h) = -2\mathcal{H}^+_{\boldsymbol{\beta}}(v_h, v_h) = -[\![v_h]\!]^2_{\boldsymbol{\beta}} \leq 0, \tag{2.25}$$

$$\|L_h v_h\|^2 = -\mathcal{H}^+_{\boldsymbol{\beta}}(v_h, L_h v_h) \leq Ch^{-1}\|v_h\|\|L_h v_h\|. \tag{2.26}$$

Hence L_h is semi-negative and $\|L_h\| \leq Ch^{-1}$. The stability of the fully discretized scheme now follows as a corollary of Theorem 2.2.1.

Corollary 2.3.3. *Consider the DG approximation of* (2.22) *with the fourth-order RK time discretization,*

$$u_h^{n+1} = P_4(\tau L_h)u_h^n, \qquad L_h \text{ defined in (2.23) and (2.24).}$$

This fully discrete scheme is strongly stable in two steps,

$$\|u_h^{n+2}\| \leq \|u_h^n\|,$$

under the CFL condition $\tau \leq Ch$*, where* C *depends on* $\boldsymbol{\beta}$ *and the constant in the inverse estimate.*

2.3.2.2 Multi-dimensional system

We now consider the symmetric hyperbolic system

$$\begin{aligned} \partial_t \boldsymbol{u}(\boldsymbol{x},t) &= \sum_{i=1}^{d} \boldsymbol{A}_i \partial_{x_i} \boldsymbol{u}(\boldsymbol{x},t), \\ \boldsymbol{x} = (x_1,...,x_d) \in \Omega \subset \mathbb{R}^d, &\qquad \boldsymbol{u} = (u^1,...,u^m)^T \in \mathbb{R}^m, \end{aligned} \tag{2.27}$$

where $\boldsymbol{A}_i$ are $m \times m$ constant real symmetric matrices. For simplicity, we assume Ω to be a hypercube in $\mathbb{R}^d$ and apply the periodic boundary conditions.

Again, we denote by $\mathcal{K} = \{K\}$ a quasi-uniform partition of the domain Ω with the mesh size h. And $\mathcal{E}$ is the collection of the cell interfaces. The space $\mathcal{V}$ is chosen as

$$\mathcal{V} = \{\boldsymbol{v} \in [L^2(\Omega)]^m | \boldsymbol{v} = (v^1,...,v^m)^T, v^k|_K \in \mathcal{P}^p(K), \forall K \in \mathcal{K}\},$$

with the inner product $(\boldsymbol{w},\boldsymbol{v}) = \sum_{K\in\mathcal{K}} (\boldsymbol{w},\boldsymbol{v})_K = \sum_{K\in\mathcal{K}} \int_K \boldsymbol{w}\cdot\boldsymbol{v} dx$ and the induced norm $\|\cdot\| = \sqrt{(\cdot,\cdot)}$. We also use $\langle \boldsymbol{w},\boldsymbol{v}\rangle_e = \int_e \boldsymbol{w}\cdot\boldsymbol{v} dl$ for the integration along the cell interfaces.

To set up the DG approximation, we firstly write down the weak formulation of (2.27)

$$(\partial_t \boldsymbol{u},\boldsymbol{v})_K = -\sum_{i=1}^{d} (\boldsymbol{A}_i\boldsymbol{u}, \partial_{x_i}\boldsymbol{v})_K + \sum_{e\in\partial K} \langle (\sum_{i=1}^{d} n^i_{e,K}\boldsymbol{A}_i)\boldsymbol{u},\boldsymbol{v}\rangle_e,$$

where $\boldsymbol{n}_{e,K} = (n^1_{e,K},...,n^d_{e,K})$ is the outward unit normal vector to the edge e in K. Since $\boldsymbol{A}_i$ are symmetric, for each e there is an orthogonal matrix $\boldsymbol{S}_e$ such that $\boldsymbol{\Lambda}_e = \mathrm{diag}(\lambda^1_e,...,\lambda^m_e) = \boldsymbol{S}_e(\sum_{i=1}^d n^i_{e,K}\boldsymbol{A}_i)\boldsymbol{S}_e^T$ is diagonal. The numerical scheme can

be obtained by applying the upwind flux for each eigen-component, namely

$$(\partial_t \boldsymbol{u}_h, \boldsymbol{v}_h)_K = -\sum_{i=1}^{d} (\boldsymbol{A}_i \boldsymbol{u}_h, \partial_{x_i} \boldsymbol{v}_h)_K + \sum_{e \in \partial K} \langle \boldsymbol{\Lambda}_e \widehat{\boldsymbol{S}_e \boldsymbol{u}_h}, \boldsymbol{S}_e \boldsymbol{v}_h \rangle_e, \tag{2.28}$$

where $\widehat{\boldsymbol{S}_e \boldsymbol{u}_h} = ((\widehat{\boldsymbol{S}_e \boldsymbol{u}_h})^1, ..., (\widehat{\boldsymbol{S}_e \boldsymbol{u}_h})^m)^T$ and

$$(\widehat{\boldsymbol{S}_e \boldsymbol{u}_h})^k = \begin{cases} (\boldsymbol{S}_e \boldsymbol{u}_h^{int})^k, & \lambda_e^k < 0, \\ (\boldsymbol{S}_e \boldsymbol{u}_h^{ext})^k, & \lambda_e^k > 0. \end{cases} \tag{2.29}$$

Here, $\boldsymbol{u}_h^{int}(\boldsymbol{x}, t) = \lim_{\boldsymbol{y} \to \boldsymbol{x}, \boldsymbol{y} \in K} \boldsymbol{u}_h(\boldsymbol{x}, t)$ and $\boldsymbol{u}_h^{ext}(\boldsymbol{x}, t) = \lim_{\boldsymbol{y} \to \boldsymbol{x}, \boldsymbol{y} \in K^c} \boldsymbol{u}_h(\boldsymbol{x}, t)$. When $\lambda_e^k = 0$, $(\widehat{\boldsymbol{S}_e \boldsymbol{u}_h})^k$ will not contribute to the integration, hence one can avoid defining it in this case. Also note that $\sum_{i=1}^d n_{e,K}^i \boldsymbol{A}_i = -\sum_{i=1}^d n_{e,K'}^i \boldsymbol{A}_i$ for $K' \cap K = e$, and the same $\boldsymbol{S}_e$ should be used on both sides. Then $\widehat{\boldsymbol{S}_e \boldsymbol{u}_h}$ defined in K and K' are the same, and the numerical flux is single-valued on the cell interfaces. Furthermore, for the scalar case, (2.29) coincides with the previous definition in Section 2.3.2.1.

As before, we define

$$\begin{aligned} \mathcal{H}(\boldsymbol{w}_h, \boldsymbol{v}_h) &= \sum_{K \in \mathcal{K}} \mathcal{H}_K(\boldsymbol{w}_h, \boldsymbol{v}_h) \\ &= \sum_{K \in \mathcal{K}} \left(\sum_{i=1}^{d} (\boldsymbol{A}_i \boldsymbol{w}_h, \partial_{x_i} \boldsymbol{v}_h)_K - \sum_{e \in \partial K} \langle \boldsymbol{\Lambda}_e \widehat{\boldsymbol{S}_e \boldsymbol{w}_h}, \boldsymbol{S}_e \boldsymbol{v}_h \rangle_e \right). \end{aligned} \tag{2.30}$$

$\mathcal{H}$ has the following properties, and its proof can be found in Appendix A.5.

Lemma 2.3.2. *$\forall \boldsymbol{v}_h, \boldsymbol{w}_h \in \mathcal{V}$, $\mathcal{H}(\boldsymbol{v}_h, \boldsymbol{v}_h) \geq 0$ and $|\mathcal{H}(\boldsymbol{w}_h, \boldsymbol{v}_h)| \leq C h^{-1} \|\boldsymbol{w}_h\| \|\boldsymbol{v}_h\|$, where C depends on $\boldsymbol{A}_i$ and the constant in the inverse estimate.*

With the definition of $\mathcal{H}$, the scheme (2.28) can be written as

$$\partial_t \boldsymbol{u}_h = L_h \boldsymbol{u}_h, \qquad (L_h \boldsymbol{u}_h, \boldsymbol{v}_h) = -\mathcal{H}(\boldsymbol{u}_h, \boldsymbol{v}_h), \qquad \forall \boldsymbol{v}_h \in \mathcal{V}. \tag{2.31}$$

Similar to those in (2.25) and (2.26), Lemma 2.3.2 implies that L_h is semi-negative and $\|L_h\| \leq Ch^{-1}$. Hence, one can use Theorem 2.2.1 to obtain the following corollary.

Corollary 2.3.4. *Consider the DG approximation of (2.27) with the fourth-order RK time discretization,*

$$\boldsymbol{u}_h^{n+1} = P_4(\tau L_h)\boldsymbol{u}_h^n, \qquad L_h \text{ defined in (2.29), (2.30) and (2.31).}$$

The fully discrete scheme is strongly stable in two steps,

$$\|\boldsymbol{u}_h^{n+2}\| \leq \|\boldsymbol{u}_h^n\|,$$

under the CFL condition $\tau \leq Ch$. Here C depends on A_i and the constant in the inverse estimate.

Remark 2.3.2. *The stability extends to symmetrizable hyperbolic systems with constants coefficients. More specifically, if there is a symmetric positive-definite matrix $\boldsymbol{H}$ such that $\boldsymbol{HA}_i$ are symmetric, then under an appropriate CFL condition $\tau \leq Ch^{-1}$, one has $\|\boldsymbol{u}_h^{n+2}\|_{\boldsymbol{H}} \leq \|\boldsymbol{u}_h^n\|_{\boldsymbol{H}}$, where $\|\boldsymbol{v}\|_{\boldsymbol{H}} = \sum_{K\in\mathcal{K}} \int_K \boldsymbol{v}^T\boldsymbol{Hv}dx$.*

Remark 2.3.3. *For symmetric hyperbolic systems with variable coefficients, one can follow similar lines to show L_h is semi-bounded, and prove the stability of the fully discretized scheme. But the extension to symmetrizable hyperbolic systems with variable coefficients is non-trivial.*

2.4 Concluding remarks

We analyze the stability of the fourth-order RK method for integrating method of lines schemes solving the well-posed linear PDE system $\partial_t u = L(x,t,\partial_x)u$. The issue of strong stability is of special interests. We consider the ODE system $\frac{d}{dt}u_N = L_N u_N$, with $L_N + L_N^\top \leq 0$ and L_N independent of time. When L_N is normal, the strong stability of the fourth-order RK approximation has already been justified by the scalar eigenvalue analysis. But for non-normal L_N, we provide a counter example to show that the strong stability can not be preserved whatever small time step we choose. However, the strong stability can actually be obtained in two steps. We prove $\|u_N^{n+2}\| \leq \|u_N^n\|$ under the time step constraint $\tau\|L_N\| \leq c_0$ for some constant c_0. This can also be interpreted as the strong stability of the eight-stage fourth-order RK method composed by two steps of the four-stage method. Then, based on this fact, we extend the stability results to general semi-bounded linear systems after using a perturbation analysis and a frozen-coefficient argument. Finally, we apply the results to justify the stability of the fully discretized schemes combining the fourth-order RK method and different spatial discretizations, including the spectral Galerkin method and the DG method. The corresponding CFL time step restrictions are obtained.

Part II

APPLICATIONS TO GRADIENT FLOWS

Chapter Three

Scalar gradient flow equations

3.1 Introduction

In this chapter, we propose a DG method for solving the initial value problem of scalar gradient flows,

$$\begin{cases} \partial_t \rho = \nabla \cdot \left(f(\rho) \nabla (H'(\rho) + V(\boldsymbol{x}) + W * \rho) \right), \quad \boldsymbol{x} \in \Omega \subset \mathbb{R}^d, \quad t > 0, \\ \rho(\boldsymbol{x}, 0) = \rho_0(\boldsymbol{x}). \end{cases} \tag{3.1}$$

Here $\rho = \rho(\boldsymbol{x}, t) \geq 0$ is a time-dependent density. H' is increasing and $W(\boldsymbol{x}) = W(-\boldsymbol{x})$ is symmetric. We assume $f(\rho) \geq 0$. It either increases with respect to ρ, or it satisfies the property $\frac{f(\rho)}{\rho} \leq C$ for some fixed constant C.

Our concern for (3.1) arises from two typical scenarios. The first case is on nonlinear (possibly degenerate) parabolic equations,

$$\partial_t \rho = \nabla \cdot \left(f(\rho) \nabla \left(H'(\rho) + V(\boldsymbol{x}) \right) \right). \tag{3.2}$$

Many classic problems can be included in this setting, such as the heat equation, the porous media equation, the Fokker-Planck equation and so on.

In the second case, several authors take the interaction term $W * \rho$ into account, while imposing $f(\rho) = \rho$. Hence the equation takes the form of a continuity equation, and the velocity field is determined by the gradient of a potential function.

$$\partial_t \rho = \nabla \cdot (\rho \boldsymbol{u}), \qquad \boldsymbol{u} = \nabla \xi = \nabla \left(H'(\rho) + V(x) + W * \rho \right). \tag{3.3}$$

Here, H is a density of internal energy, V is a confinement potential, and W is an interaction potential. It is used to model, for example, interacting gases [18, 75],

granular flows [7, 6] and aggregation behaviors in biology [61, 74]. This equation is also related with the gradient flow for the Wasserstein metric on the space of probability measures [3].

Both of the problems (3.2) and (3.3) can be formulated under the framework of (3.1). It has an underlying structure associated with the entropy functional,

$$E = \int_\Omega H(\rho(\boldsymbol{x}))d\boldsymbol{x} + \int_\Omega V(\boldsymbol{x})\rho(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2}\int_\Omega\int_\Omega W(\boldsymbol{x}-\boldsymbol{y})\rho(\boldsymbol{x})\rho(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}.$$

One can show that at least for classical solutions,

$$\frac{d}{dt}E(\rho) = -\int_\Omega f(\rho)|\boldsymbol{u}|^2 d\boldsymbol{x} := -I(\rho) \leq 0. \tag{3.4}$$

Here $\boldsymbol{u}$ is defined as that in (3.3) and I is referred to as the entropy dissipation.

Indeed, (3.4) has provided much insight into the problem and has helped people to study the dynamics of (3.2) and (3.3), see, for example [14, 18, 75]. Hence it is desirable to develop numerical schemes mimicking a similar entropy-entropy dissipation relationship in the discrete sense.

Another challenge for developing the numerical schemes is to ensure the non-negativity of the numerical density without violating the mass conservation. It is not only for the preservation of physical meanings, but also for the well-posedness of the initial value problem. For example, in (3.3), the entropy may not necessarily decay if ρ admits negative values.

Numerical schemes addressing both of these concerns have been studied intensively very recently. In [8], the authors designed a second order finite volume scheme for (3.2). Later on, a direct DG method has been proposed by Liu and Wang in [56].

Their scheme achieves high-order accuracy but the provable positivity-preserving property only holds for certain cases. Recently, this method is generalized for solving the one-dimensional Poisson-Nernst-Planck system [57], which essentially incorporates the interaction through a coupled Poisson equation. As for (3.3), a variety of numerical methods have been developed, including a mixed finite element method [10], a finite volume method [15], a particle method [17], a method of evolving diffeomorphisms [19] and a blob method [33] (for $H = 0$, $V = 0$).

In this chapter, we design a high-order DG method for (3.1), which covers both (3.2) and (3.3). Our idea is to formally treat (3.1) as a classical conservation law and apply the techniques there to overcome the challenges. The main ingredients for our schemes are:

1. Legendre-Gauss-Lobatto quadrature rule for numerical integration,

2. positivity-preserving limiter with SSP-RK time discretization.

The quadrature is used to stabilize the semi-discrete scheme. Such approach has already been studied in different contexts, such as in spectral collocation methods [42] and in nodal DG methods [43]. Recently, based on the methodology established in [13, 36, 37], Chen and Shu proposed a unified framework for designing entropy stable DG scheme for hyperbolic conservation laws using suitable quadrature rules [22]. Their approach is also related with the summation-by-part technique in finite difference methods. The positivity-preserving limiter with provable high-order accuracy is firstly designed by Zhang and Shu in [87] to numerically ensure the maximum principle of scalar hyperbolic conservation laws. Then the methodology has been generalized for developing the bound-preserving schemes for various systems. We refer to [88] and the references therein for more details. Our approach of imple-

menting the positivity-preserving limiter for parabolic problems is mainly inspired by the recent work of Zhang on compressible Navier-Stokes equation [86], in which the author considers the conservative form of the problem and introduces the diffusion flux to handle the second-order derivatives.

Based on these techniques, we propose a DG scheme for (3.1) such that

1. the semi-discrete scheme satisfies an entropy inequality for smooth W,
2. the fully discretized scheme is positivity-preserving.

Our method also has other desired properties. It achieves high-order accuracy, conserves the total mass and preserves numerical steady states. Special care is needed for the case of non-smooth interaction potential, due to the fact that one should adopt exact integration to calculate the convolution, see [15], as the Gauss-Lobatto quadrature is no longer accurate.

The remaining part of the chapter is organized as follows. In Section 3.2, we design the numerical method for one-dimensional problems. We firstly introduce the notations and briefly discuss our motivation on deriving the scheme. Then it follows with the semi-discrete scheme and the discrete entropy inequality. The next part is on the time discretization and the positivity-preserving property of the fully discretized scheme. Finally we outline the matrix formulation and the algorithm flowchart. Section 3.3 is organized similarly for two-dimensional problems on Cartesian meshes, while the implementation details are omitted. Then in Section 3.4 and Section 3.5, we present numerical examples for one-dimensional and two-dimensional problems respectively. Finally conclusions are drawn in Section 3.6.

3.2 Numerical method: one-dimensional case

3.2.1 Notations and motivations

For now, we focus on the one-dimensional case of (3.1)

$$\begin{cases} \partial_t \rho = \partial_x \big(f(\rho) \partial_x (H'(\rho) + V(x) + W * \rho) \big), \quad x \in \Omega \subset \mathbb{R}, \quad t > 0, \\ \rho(x, 0) = \rho_0(x). \end{cases}$$

In general, the problem can be either defined on a connected compact domain with proper boundary conditions, or it can involve the whole real line with solutions vanishing at the infinity. In our numerical scheme, we will always choose Ω to be a connected interval. For simplicity, the periodic or compactly supported boundary conditions are applied. But we remark that our approach can be extended to more general types of boundary conditions, for example, with zero-flux boundaries.

Let $I_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$ and $I = \cup_{i=1}^N I_i$ be a partition of the domain Ω. Denote $h_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ and $h = \max_i h_i$. We will seek a numerical solution in the discontinuous piecewise polynomial space,

$$V_h = \{ v_h : v_h \big|_{I_i} \in P^k(I_i), \text{for } x \in I_i, i = 1, \dots, N \}. \tag{3.5}$$

Here $P^k(I_i)$ is the space of k-th order polynomials on I_i. Note that the functions in V_h can be double-valued at the cell interfaces. Hence the notations v_h^+ and v_h^- are introduced for the right limit and the left limit of v_h. For a function $s = s(\rho)$ or $s = s(\rho, x)$, we denote by $s_h = s(\rho_h)$ or $s_h = s(\rho_h, x)$ respectively. Furthermore, the notation s_h^+ stands for $s(\rho_h^+)$ or $s(\rho_h^+, x)$, and s_h^- stands for $s(\rho_h^-)$ or $s(\rho_h^-, x)$.

To define the DG method, we split the original problem into the following system,

$$\begin{aligned}\partial_t \rho &= \partial_x (f(\rho) u),\\ u &= \partial_x \xi,\\ \xi &= H'(\rho) + V + W * \rho.\end{aligned}$$

By applying the DG approximation, we obtain the following scheme as a *preliminary*. We seek $\rho_h, u_h \in V_h$, such that for any test function $\varphi_h, \psi_h \in V_h$,

$$\int_{I_i} (\partial_t \rho_h)\varphi_h dx = -\int_{I_i} f_h u_h \partial_x \varphi_h dx + \widehat{fu}_{i+\frac{1}{2}} (\varphi_h)^-_{i+\frac{1}{2}} - \widehat{fu}_{i-\frac{1}{2}} (\varphi_h)^+_{i-\frac{1}{2}}, \tag{3.6a}$$

$$\int_{I_i} u_h \psi_h dx = -\int_{I_i} \xi_h \partial_x \psi_h dx + \widehat{\xi}_{i+\frac{1}{2}} (\psi_h)^-_{i+\frac{1}{2}} - \widehat{\xi}_{i-\frac{1}{2}} (\psi_h)^+_{i-\frac{1}{2}}. \tag{3.6b}$$

Here $\xi_h = H'_h + V + \int_\Omega W(x-y)\rho_h(y)dxdy$, while $\widehat{fu}$ and $\widehat{\xi}$ are numerical fluxes.

By setting $\varphi_h(x) = 1_{I_i}(x)$, one will get the following evolution equation for the cell average of $(\rho_h)_i$, which is denoted by $(\bar{\rho}_h)_i$.

$$\frac{d}{dt}(\bar{\rho}_h)_i = \frac{\widehat{fu}_{i+\frac{1}{2}} - \widehat{fu}_{i-\frac{1}{2}}}{h_i}. \tag{3.7}$$

This is very similar to that in hyperbolic conservation laws. In particular, for $k = 0$ and $u \equiv 1$, by using the so called monotone flux, (3.7) will become a monotone scheme, which satisfies many desirable properties.

In order to achieve an entropy-entropy dissipation relationship as that for the exact solution, one may hope to set $\varphi_h = \xi_h$ in (3.6a) and $\psi_h = u_h f_h$ in (3.6b). Unluckily, neither of them falls into the test function space. A natural attempt is to use the usual L^2 projection to enforce this property, as that in [56]. However, the projection will change the values at the cell interfaces, and the desired form

in (3.7) will be violated. This will cause trouble when one seeks to preserve the non-negativity of the solution. Inspired by the recent work of Chen and Shu in [22], we introduce a suitable quadrature to overcome this difficulty. Moreover, the Gauss-Lobatto quadrature is used to preserve the values at the cell ends.

Let us denote by $\{x_i^r\}_{r=1}^{k+1}$ the $k+1$ Gauss-Lobatto quadrature points on I_i and $\{w_r\}_{r=1}^{k+1}$ the $k+1$ Gauss-Lobatto quadrature weights on $[-1,1]$. In particular $x_i^1 = x_{i-\frac{1}{2}}^+$ and $x_i^{k+1} = x_{i+\frac{1}{2}}^-$. On each cell I_i, the operator $\mathcal{I}$ returns the k-th order polynomial interpolating at $\{x_i^r\}_{r=1}^{k+1}$. We will use the notation

$$\widetilde{\int}_{I_i} \eta\zeta dx = \frac{h_i}{2}\sum_{r=1}^{k+1} w_r \eta(x_i^r)\zeta(x_i^r)$$

and

$$\widetilde{\int}_{I_i} \eta\partial_x\zeta dx = \frac{h_i}{2}\sum_{r=1}^{k+1} w_r \eta(x_i^r)(\partial_x\mathcal{I}\zeta)(x_i^r)$$

for the Gauss-Lobatto quadrature. As a convention, $\widetilde{\int}_\Omega$ stands for $\sum_{i=1}^N \widetilde{\int}_{I_i}$.

We will finally come to the fully discretized scheme, for which we denote by τ the time step length and $\lambda_i = \frac{\tau}{h_i}$.

3.2.2 Semi-discrete scheme and entropy inequality

Our scheme is to replace the integrals in (3.6) by the Gauss-Lobatto quadrature. In other words, with V_h defined in (3.5), we seek $\rho_h, u_h \in V_h$, such that for any test functions $\varphi_h, \psi_h \in V_h$,

$$\widetilde{\int}_{I_i} (\partial_t\rho_h)\varphi_h dx = -\widetilde{\int}_{I_i} f_h u_h \partial_x\varphi_h dx + \widehat{fu}_{i+\frac{1}{2}}(\varphi_h)_{i+\frac{1}{2}}^- - \widehat{fu}_{i-\frac{1}{2}}(\varphi_h)_{i-\frac{1}{2}}^+, \tag{3.8a}$$

$$\widetilde{\int_{I_i}} u_h\psi_h dx = -\widetilde{\int_{I_i}} \xi_h\partial_x\psi_h dx + \widehat{\xi}_{i+\frac{1}{2}}(\psi_h)^-_{i+\frac{1}{2}} - \widehat{\xi}_{i-\frac{1}{2}}(\psi_h)^+_{i-\frac{1}{2}}. \tag{3.8b}$$

Here, when the interaction potential W is smooth, we set

$$\xi_h = \xi_h(\rho_h, x) = H_h' + V + \widetilde{\int_\Omega} W(x-y)\rho_h(y)dy.$$

While for non-smooth W, the quadrature may not achieve sufficient accuracy of the convolution. Hence the exact integration is applied

$$\xi_h = \xi_h(\rho_h, x) = H_h' + V + \int_\Omega W(x-y)\rho_h(y)dy.$$

The numerical fluxes are chosen in the following way,

$$\begin{aligned}
\widehat{\xi} &= \frac{1}{2}(\xi_h^+ + \xi_h^-),\\
\widehat{fu} &= \frac{1}{2}(f_h^+u_h^+ + f_h^-u_h^-) + \frac{\alpha}{2}(g_h^+ - g_h^-), \quad \alpha = \max\{|u_h^+|, |u_h^-|\},
\end{aligned}$$

with $g(\rho) = f(\rho)$ if f is increasing and $g(\rho) = C\rho$ if $\frac{f(\rho)}{\rho} \le C$. $\widehat{\xi}$ is the central flux.

Remark 3.2.1. *Although we formally require that φ_h and ψ_h are taken from the finite element space, our scheme (3.8) actually can not distinguish a function from its interpolation polynomial at the Gauss-Lobatto points. Hence, (3.8) will also hold for "test functions" outside V_h. This facilitates our proof of the discrete entropy inequality. This fact can also be justified through the matrix formulation, which is presented in Section 3.2.4. We also remind the readers that the number of quadrature points must be $k+1$. It may result in underdetermined systems if one uses a less accurate quadrature rule, while using more points will mess up with the proof of the entropy inequality.*

Remark 3.2.2. *In general, g is not unique, while it must satisfy* $\mathrm{sign}[g_h] = \mathrm{sign}[\rho_h]$ *or* $\mathrm{sign}[g_h] = 0$, *and* $\alpha g \pm f u \geq 0$. *The correct sign of g_h is needed for the entropy inequality and $\alpha g \pm f u \geq 0$ is used to prove the positivity-preserving property. For most of the applications, $f(\rho)$ is increasing and $g(\rho) = f(\rho)$ should be enough. In particular, in gradient flow type problems where $f(\rho) = \rho$, $g(\rho) = \rho$ gives the local Lax-Friedrichs flux. Similar comments also apply for the two-dimensional problems.*

This semi-discrete scheme satisfies the following entropy inequality.

Theorem 3.2.1. *For smooth interaction kernel W, assume that the semi-discrete scheme (3.8) has a solution, then it satisfies a similar entropy-entropy dissipation relationship, as that in (3.4), given by*

$$\frac{d}{dt}\tilde{E} \leq -\tilde{I},$$

where

$$\tilde{E} = \int_{\Omega}^{\sim} H(\rho_h)dx + \int_{\Omega}^{\sim} V\rho_h dx + \frac{1}{2}\int_{\Omega}^{\sim}\int_{\Omega}^{\sim} W(x-y)\rho_h(x)\rho_h(y)dxdy$$

is the discrete entropy and

$$\tilde{I} = \int_{\Omega}^{\sim} f_h |u_h|^2 dx$$

is the associated discrete entropy dissipation.

Indeed, one can choose larger $\alpha_{i+\frac{1}{2}}$ in the Lax-Friedrichs flux. This will bring in extra numerical dissipation and the entropy inequality will still hold. Moreover, assuming g and H' are strictly increasing, if $\alpha_{i+\frac{1}{2}} > 0$ for all i and the semi-discrete scheme (3.8) achieves a non-negative stationary state ρ_h, then ρ_h is continuous and the piecewise polynomial interpolation of ξ_h is constant in each connected component J of the strict support of ρ_h, which is defined by $J = \cup_{i\in\Lambda} I_i$ for certain set of consecutive indices Λ where $\rho_h > 0$.

Proof. Using the symmetry of W, we have

$$\begin{aligned}
&\frac{d}{dt}\frac{1}{2}\tilde{\int_\Omega}\tilde{\int_\Omega} W(x-y)\rho_h(x)\rho_h(y)dxdy \\
=&\frac{1}{2}\tilde{\int_\Omega}\partial_t\rho_h(x)\left(\tilde{\int_\Omega} W(x-y)\rho_h(y)dy\right)dx + \frac{1}{2}\tilde{\int_\Omega}\partial_t\rho_h(y)\left(\tilde{\int_\Omega} W(x-y)\rho_h(x)dx\right)dy \\
=&\tilde{\int_\Omega}\partial_t\rho_h(x)\left(\tilde{\int_\Omega} W(x-y)\rho_h(y)dy\right)dx.
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{d}{dt}\tilde{E} &= \tilde{\int_\Omega}\partial_t\rho_h(x)\left(H_h' + V + \tilde{\int_\Omega} W(x-y)\rho_h(y)dy\right)dx \\
&= \tilde{\int_\Omega}\partial_t\rho_h\xi_h dx = \sum_{i=1}^N \tilde{\int_{I_i}}\partial_t\rho_h\mathcal{I}(\xi_h)dx \\
&= \sum_{i=1}^N\left(-\tilde{\int_{I_i}}\mathcal{I}(f_hu_h)\partial_x\mathcal{I}(\xi_h)dx + \widehat{fu}_{i+\frac{1}{2}}(\xi_h)^-_{i+\frac{1}{2}} - \widehat{fu}_{i-\frac{1}{2}}(\xi_h)^+_{i-\frac{1}{2}}\right).
\end{aligned}$$

Note $\mathcal{I}(f(\rho_h)u_h)\partial_x\mathcal{I}(\xi_h)$ is a polynomial of degree $2k-1$ and the Gauss-Lobatto quadrature with $k+1$ points is exact. Hence we can replace the quadrature with the exact integral and integrate by parts. Then we obtain

$$\begin{aligned}
\frac{d}{dt}\tilde{E} &= \sum_{i=1}^N\left(-\int_{I_i}\mathcal{I}(f_hu_h)\partial_x\mathcal{I}(\xi_h)dx + \widehat{fu}_{i+\frac{1}{2}}(\xi_h)^-_{i+\frac{1}{2}} - \widehat{fu}_{i-\frac{1}{2}}(\xi_h)^+_{i-\frac{1}{2}}\right) \\
&= \sum_{i=1}^N\Bigg(\int_{I_i}(\partial_x\mathcal{I}(f_hu_h))\mathcal{I}(\xi_h)dx - (f_hu_h\xi_h)^-_{i+\frac{1}{2}} \\
&\qquad + (f_hu_h\xi_h)^+_{i-\frac{1}{2}} + \widehat{fu}_{i+\frac{1}{2}}(\xi_h)^-_{i+\frac{1}{2}} - \widehat{fu}_{i-\frac{1}{2}}(\xi_h)^+_{i-\frac{1}{2}}\Bigg).
\end{aligned}$$

By using the same trick, we change the exact integral back to the Gauss-Lobatto quadrature, and apply the scheme (3.8b) to obtain

$$\frac{d}{dt}\tilde{E} = \sum_{i=1}^N\left(\tilde{\int_{I_i}}\xi_h\partial_x(f_hu_h)dx - (f_hu_h\xi_h)^-_{i+\frac{1}{2}} + (f_hu_h\xi_h)^+_{i-\frac{1}{2}} + \widehat{fu}_{i+\frac{1}{2}}(\xi_h)^-_{i+\frac{1}{2}} - \widehat{fu}_{i-\frac{1}{2}}(\xi_h)^+_{i-\frac{1}{2}}\right)$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \Bigg(- \widetilde{\int_{I_i}} f_h |u_h|^2 dx + \widehat{\xi}_{i+\frac{1}{2}} (f_h u_h)^-_{i+\frac{1}{2}} - \widehat{\xi}_{i-\frac{1}{2}} (f_h u_h)^+_{i-\frac{1}{2}} - (f_h u_h \xi_h)^-_{i+\frac{1}{2}} \\
&\quad + (f_h u_h \xi_h)^+_{i-\frac{1}{2}} + \widehat{fu}_{i+\frac{1}{2}} (\xi_h)^-_{i+\frac{1}{2}} - \widehat{fu}_{i-\frac{1}{2}} (\xi_h)^+_{i-\frac{1}{2}} \Bigg) \\
&= - \widetilde{\int_{\Omega}} f_h |u_h|^2 dx - \sum_{i=1}^{N} \frac{\alpha_{i+\frac{1}{2}}}{2} [g_h]_{i+\frac{1}{2}} [\xi_h]_{i+\frac{1}{2}}, \qquad (3.9)
\end{aligned}
$$

where the bracket represents the jump, $[g_h] = g_h^+ - g_h^-$. According to our choice of g, $\text{sign}[g_h] = \text{sign}[\rho_h]$ or $\text{sign}[g_h] = 0$. Since V and W are single-valued functions, $[\xi_h] = [H'(\rho_h)]$. By using the fact that H' is increasing, we have $\text{sign}[\xi_h] = \text{sign}[H'(\rho_h)] = \text{sign}[\rho_h]$ or $\text{sign}[\xi_h] = 0$. Therefore $\sum_{i=1}^{N} \frac{\alpha_{i+\frac{1}{2}}}{2} [g]_{i+\frac{1}{2}} [\xi_h]_{i+\frac{1}{2}} \geq 0$, which completes the proof of the first claim.

It is easy to see that the entropy inequality will hold as long as the coefficients $\alpha_{i+\frac{1}{2}}$ are non-negative. Let us assume now that $\alpha_{i+\frac{1}{2}} > 0$ for all i and ρ_h is a non-negative stationary state of the semi-discrete scheme (3.8), namely

$$\frac{d}{dt} \tilde{E} = 0.$$

Then from (3.9) we deduce that both terms in the right-hand side must vanish, that is

$$\sum_{i=1}^{N} \widetilde{\int_{I_i}} f_h |u_h|^2 dx = \sum_{i=1}^{N} \frac{\alpha_{i+\frac{1}{2}}}{2} [g_h]_{i+\frac{1}{2}} [\xi_h]_{i+\frac{1}{2}} = 0 \,.$$

Therefore each term in the summations will be zero. On the one hand, if $\alpha_{i+\frac{1}{2}} > 0$ while g and H' are strictly increasing, then $[g_h]_{i+\frac{1}{2}} [\xi_h]_{i+\frac{1}{2}} = 0$ and hence $[\rho_h]_{i+\frac{1}{2}} = 0$. It holds for all i, which means the jumps of ρ_h vanish on all the cell interfaces. This implies the continuity of ρ_h. On the other hand, in the interval J we deduce that $u_h = 0$ due to the positivity of f_h implied by the positivity of ρ_h and its definition.

Hence for all $i \in \Lambda$, by using (3.8b), one can obtain

$$\begin{aligned}\int_{I_i} \partial_x(\mathcal{I}\xi_h)\mathcal{I}\psi_h dx &= -\int_{I_i} \mathcal{I}\xi_h \partial_x(\mathcal{I}\psi_h)dx + (\xi_h\psi_h)^-_{i+\frac{1}{2}} - (\xi_h\psi_h)^+_{i-\frac{1}{2}} \\ &= -\int_{I_i}^{\sim} \xi_h \partial_x\psi_h dx + \widehat{\xi}_{j+\frac{1}{2}}(\psi_h)^-_{i+\frac{1}{2}} - \widehat{\xi}_{i+\frac{1}{2}}(\psi_h)^+_{i-\frac{1}{2}} \\ &= 0.\end{aligned}$$

Here $\xi_h^{\pm} = \widehat{\xi}$ is guaranteed by the continuity of ξ_h (implied by that of ρ_h). Therefore, $\mathcal{I}\xi_h$ is constant on each $i \in \Lambda$. Due to the fact that ξ_h is continuous globally, all these constants must be the same and the piecewise polynomial interpolation of ξ_h is constant on J. □

3.2.3 Time discretization and preservation of positivity

The semi-discrete scheme itself does not guarantee the positivity of the numerical solution. If no special treatment is applied, one may produce nonsense density with negative values and the problem can become illposed. Hence we adopt the methodology developed by Zhang and Shu in [87], which enforces the positivity of the solution without violating the mass conservation. Their idea is to incorporate a positivity-preserving limiter into the SSP-RK time discretization. Under certain time step constraints, each Euler forward step preserves the positivity of the cell average (referred to as the weak positivity in the literature). Then one can scale the solution, without affecting spatial accuracy, to ensure the point-wise non-negativity. SSP-RK time discretization will preserve the non-negativity of the solution in the Euler forward steps.

3.2.3.1 First-order Euler forward in time

Let us firstly consider the Euler forward time stepping. We use the superscript "pre" for the solution obtained by Euler forward method before applying the positivity-preserving limiter. The time discretization of (3.8a) becomes

$$\int_{I_i}^{\sim} \frac{\rho_h^{n+1,\mathrm{pre}} - \rho_h^n}{\tau} \varphi_h dx = -\int_{I_i}^{\sim} f_h^n u_h^n \partial_x \varphi_h dx + (\widehat{fu})_{i+\frac{1}{2}}^n (\varphi_h)_{i+\frac{1}{2}}^- - (\widehat{fu})_{i-\frac{1}{2}}^n (\varphi_h)_{i-\frac{1}{2}}^+. \quad (3.10)$$

Lemma 3.2.1. *Suppose $\rho_h^n(x_i^r) \geq 0$ at the Gauss-Lobatto quadrature points. Then when $\lambda_i = \frac{\tau}{h_i} \leq \min\{(\frac{w_1\rho}{fu+\alpha g})_{i-\frac{1}{2}}^+, (\frac{w_{k+1}\rho}{\alpha g - fu})_{i+\frac{1}{2}}^-\}$, the solution $\rho_h^{n+1,\mathrm{pre}}$ obtained from (3.10) satisfies $(\bar{\rho}_h)_i^{n+1,\mathrm{pre}} \geq 0$. Here, in the constraint of λ_i, we consider $\frac{0}{0} := +\infty$ by convention.*

Proof. We drop all subscripts h in this proof. Take $\varphi = 1$ in (3.10), we have

$$\bar{\rho}_i^{n+1,\mathrm{pre}} = \bar{\rho}_i^n + \lambda_i \left((\widehat{fu})_{i+\frac{1}{2}}^n - (\widehat{fu})_{i-\frac{1}{2}}^n \right).$$

Note that ρ^n is a polynomial of degree k, the Gauss-Lobatto quadrature is exact for evaluating the cell average $\bar{\rho}_i^n$. More specifically, we have

$$\bar{\rho}_i^n = \frac{1}{h_i} \int_{I_i} \rho^n dx = \sum_{r=1}^{k+1} \frac{w_r}{2} \rho^n(x_i^r).$$

The superscripts n will also be omitted for simplicity in the rest. Hence

$$\begin{aligned}
\bar{\rho}_i^{n+1,\text{pre}} =& \sum_{r=2}^{k} \frac{w_r}{2}\rho(x_i^r) + \frac{w_1}{2}\rho^+_{i-\frac{1}{2}} + \frac{w_{k+1}}{2}\rho^-_{i+\frac{1}{2}} \\
&+ \frac{\lambda_i}{2}\left((fu)^+_{i+\frac{1}{2}} + (fu)^-_{i+\frac{1}{2}} + \alpha_{i+\frac{1}{2}}\left(g^+_{i+\frac{1}{2}} - g^-_{i+\frac{1}{2}}\right)\right) \\
&- \frac{\lambda_i}{2}\left((fu)^+_{i-\frac{1}{2}} + (fu)^-_{i-\frac{1}{2}} + \alpha_{i-\frac{1}{2}}\left(g^+_{i-\frac{1}{2}} - g^-_{i-\frac{1}{2}}\right)\right) \\
=& \sum_{r=2}^{k} \frac{w_r}{2}\rho(x_i^r) + \left(\frac{w_{k+1}}{2}\rho^-_{i+\frac{1}{2}} + \frac{\lambda_i}{2}\left((fu)^-_{i+\frac{1}{2}} - \alpha_{i+\frac{1}{2}}g^-_{i+\frac{1}{2}}\right)\right) \\
&+ \left(\frac{\omega_1}{2}\rho^+_{i-\frac{1}{2}} - \frac{\lambda_i}{2}\left((fu)^+_{i-\frac{1}{2}} + \alpha_{i-\frac{1}{2}}g^+_{i-\frac{1}{2}}\right)\right) \\
&+ \frac{\lambda_i}{2}\left((fu)^+_{i+\frac{1}{2}} + \alpha_{i+\frac{1}{2}}g^+_{i+\frac{1}{2}}\right) - \frac{\lambda_i}{2}\left((fu)^-_{i-\frac{1}{2}} - \alpha_{i-\frac{1}{2}}g^-_{i-\frac{1}{2}}\right).
\end{aligned}$$

The first term is automatically non-negative, since the weights $w_r \geq 0$ and the nodal values $\rho(x_i^r) \geq 0$. The positivity of the last two terms is guaranteed by our choice of α and g. One only needs $\lambda_i \leq \min\{(\frac{w_1\rho}{fu+\alpha g})^+_{i-\frac{1}{2}}, (\frac{w_{k+1}\rho}{\alpha g - fu})^-_{i+\frac{1}{2}}\}$ to ensure the second and the third term to be non-negative. (Note that for $\rho^+_{i-\frac{1}{2}}$ or $\rho^-_{i+\frac{1}{2}}$ being 0, the corresponding term is also 0 and there is nothing to impose. Hence we introduce the notation $\frac{0}{0} := +\infty$.) Therefore $\bar{\rho}_i^{n+1,\text{pre}} \geq 0$ under the prescribed time step constraint. □

Remark 3.2.3. *According to the definition of α and g, $(\frac{w_1\rho}{fu+\alpha g})^+_{i-\frac{1}{2}}$ and $(\frac{w_{k+1}\rho}{\alpha g - fu})^-_{i+\frac{1}{2}}$ will always be non-negative.*

Remark 3.2.4. *Although the original equation can be parabolic, we have incorporated the second-order derivative into u, such that one can formally treat it as a hyperbolic problem. This technique is introduced by Zhang for the compressible Navier-Stokes equation [86].*

Here we provide an estimate on the time step for $g(\rho) = f(\rho) = \rho$. One has $\lambda_i = \min\{\left(\frac{w_1}{\alpha+u}\right)^+_{i-\frac{1}{2}}, \left(\frac{w_{k+1}}{\alpha-u}\right)^-_{i+\frac{1}{2}}\}$ in such situations. We assume the mesh is quasi-

uniform, namely, $ch \leq h_i$ for some positive constant c. Then it suffices to satisfy $\tau \leq \frac{c}{\max_i |u^{\pm}_{i\mp\frac{1}{2}}|} h$. Note that

$$\widetilde{\int_{I_i}} u_h \psi_h dx = -\int_{I_i} \mathcal{I}(\xi_h)\partial_x \psi_h dx + \widehat{\mathcal{I}(\xi_h)}_{i+\frac{1}{2}}(\psi_h)^-_{i+\frac{1}{2}} - \widehat{\mathcal{I}(\xi_h)}_{i-\frac{1}{2}}(\psi_h)^+_{i-\frac{1}{2}}. \tag{3.11}$$

If $H' \equiv 0$ in the definition of ξ, then ξ is continuous and (3.11) gives $u_h = \partial_x \mathcal{I}(\xi_h)$ on I_i after integration by parts. Hence $\tau \leq \frac{c}{\max_i \|\partial_x \mathcal{I}(\xi_h)\|_{L^\infty(I_i)}} h$. In particular, for $V = x$, this corresponds to the usual CFL condition for hyperbolic conservation laws.

Otherwise one uses the inverse estimate and the norm equivalence for the time step estimate. Note that $\|v\|_{L^p(I_i)} \leq c h_i^{\frac{1}{p}-\frac{1}{q}} \|v\|_{L^q(I_i)}$, $\forall 1 \leq p, q \leq \infty$, $\forall v \in P^k(I_i)$. Furthermore, $c_1 \|v\|_{L^2_d(I_i)} \leq \|v\|_{L^2(I_i)} \leq c_2 \|v\|_{L^2_d(I_i)}$, with $\|v\|_{L^2_d(I_i)} = \sqrt{\widetilde{\int_{I_i}} v^2 dx}$, $\forall v \in P^k(I_i)$. One has

$$\begin{aligned}
&\|u\|_{L^\infty(I_i)} \leq c h^{-\frac{1}{2}} \|u\|_{L^2(I_i)} \leq c h^{-\frac{1}{2}} \|u\|_{L^2_d(I_i)} \\
\leq & c h^{-\frac{3}{2}} (\|\mathcal{I}(\xi_h)\|_{L^2(I_{i-1})} + \|\mathcal{I}(\xi_h)\|_{L^2(I_i)} + \|\mathcal{I}(\xi_h)\|_{L^2(I_{i+1})}) \\
\leq & c h^{-1} \|\mathcal{I}(\xi_h)\|_{L^\infty(I_{i-1}\cup I_i \cup I_{i+1})}.
\end{aligned}$$

Therefore, $\tau \leq \frac{c}{\max_i \|\mathcal{I}(\xi_h)\|_{L^\infty(I_i)}} h^2$, which is comparable to the time step constraint for parabolic equations.

Lemma 3.2.1 tells us an inherent property of the Euler forward scheme. If the solution is non-negative at the previous time step (at the Gauss-Lobatto quadrature points), as long as the time step is smaller than a threshold, the cell average at next time step will remain non-negative. In order to close the loop, one would need to ensure the nodal values at the quadrature points of the next time step are also non-negative. This indeed can be achieved by applying a scaling limiter, which luckily does not affect the spatial accuracy. We refer to [86] for more details.

Lemma 3.2.2. *Let*

$$\rho_h^{n+1}(x_i^r) = (\bar{\rho}_h)_i^{n+1,\text{pre}} + \theta_i \left(\rho_h^{n+1,\text{pre}}(x_i^r) - (\bar{\rho}_h)_i^{n+1,\text{pre}}\right), \qquad \forall r = 1, \ldots, k+1,$$

with $\theta_i = \min\{\frac{(\bar{\rho}_h)_i^{n+1,\text{pre}}}{(\bar{\rho}_h)_i^{n+1,\text{pre}} - m_i}, 1\}$ *and* $m_i = \min\{\rho_h^{n+1,\text{pre}}(x_i^r)\}_{r=1}^{k+1}$. *Then we have* $\rho_h^{n+1}(x_i^r) \geq 0, \forall r = 1, \ldots, k+1$ *and* $(\bar{\rho}_h)_i^{n+1} = (\bar{\rho}_h)_i^{n+1,\text{pre}}$. *Furthermore, the interpolation polynomial of* $\{\rho_h^{n+1}(x_i^r)\}$ *on* I_i *satisfies*

$$|\rho_h^{n+1}(x) - \rho_h^{n+1,\text{pre}}(x)| \leq C_k \max_{x \in I_i} |\rho(x, t_{n+1}) - \rho_h^{n+1,\text{pre}}(x)|,$$

where $\rho(x, t_{n+1})$ *is the exact solution at time* t_{n+1} *and* C_k *is a constant depending only on the polynomial degree* k.

Remark 3.2.5. *Our scheme only uses the nodal values at the Gauss-Lobatto quadrature points, hence we only need to ensure the non-negativity at these nodes. One can also squash the solution polynomials so that the solution is non-negative everywhere on the domain. The proof will still go through.*

Theorem 3.2.2. *With the scaling limiter in Lemma 3.2.2, the Euler forward time discretization of the semi-discrete scheme is positivity-preserving, provided the time step restriction specified in Lemma 3.2.1 is satisfied.*

3.2.3.2 High-order time discretization

The SSP-RK method will be used for time discretization. We refer readers to [38] for more details. Since the time step usually scales like $\tau = Ch^2$, the Euler forward method will be sufficient for piecewise linear elements to achieve overall second-order

accuracy. For $k = 2, 3$, we will use the second order SSP-RK scheme

$$\rho_h^{(1)} = \rho_h^n + \tau F(\rho_h^n), \tag{3.12a}$$

$$\rho_h^{n+1} = \frac{1}{2}\rho_h^n + \frac{1}{2}\left(\rho_h^{(1)} + \tau F(\rho_h^{(1)})\right). \tag{3.12b}$$

For $k = 4, 5$, the third-order SSP-RK scheme is used

$$\rho_h^{(1)} = \rho_h^n + \tau F(\rho_h^n), \tag{3.13a}$$

$$\rho_h^{(2)} = \frac{3}{4}\rho_h^n + \frac{1}{4}\left(\rho_h^{(1)} + \tau F(\rho_h^{(1)})\right), \tag{3.13b}$$

$$\rho_h^{n+1} = \frac{1}{3}\rho_h^n + \frac{2}{3}\left(\rho_h^{(2)} + \tau F(\rho_h^{(2)})\right). \tag{3.13c}$$

The positivity-preserving limiter should be applied immediately after each Euler forward stage. As one can see, the SSP-RK schemes (3.12) and (3.13) can be rewritten as convex combinations of the Euler forward steps. Since each Euler forward step preserves the positivity, the numerical density at the next time level will remain non-negative.

Theorem 3.2.3. *Consider the SSP-RK time discretization* (3.12) *and* (3.13) *of the semi-discrete scheme* (3.8). *By applying limiters specified in Lemma 3.2.2, the fully discretized scheme preserves non-negativity as long as the time step restriction in Lemma 3.2.1 is satisfied.*

We also mention several other properties of the fully discretized scheme, whose proofs are omitted. Such properties also hold for two-dimensional cases.

1. Mass conservation: $\int_\Omega \rho_h^n(x)dx = \int_\Omega \rho_0(x)dx$.

2. Preservation of numerical steady states: if the numerical potential $\mathcal{I}(\xi_h)$ be-

comes constant on each connected component of the extended support of ρ_h, then we have $\rho_h^{n+1} = \rho_h^n$. Here the extended support is defined as the usual support padded with an extra mesh cell on each side. The preservation of numerical steady state for the fully discretized scheme is related with the semi-discrete version through the profile of ρ_h and ξ_h.

3.2.4 Matrix formulation and implementation

At the end of this section, we would like to introduce the matrix formulation of our numerical scheme and outline the flowchart of the algorithm.

3.2.4.1 Matrix formulation

The derivation of the matrix formulation is similar to that in Section 3.1 of [22]. We refer to that paper for more details.

We omit all the subscripts h. Let $\{\zeta_r\}_{r=1}^{k+1}$ be the Gauss-Lobatto quadrature points on the reference element $[-1, 1]$. We denote by L_r, $r = 1, \ldots, k+1$ the Lagrangian basis polynomials interpolating at these nodes.

$$L_r(\zeta) = \prod_{s=1, s\neq r}^{k+1} \frac{\zeta - \zeta_s}{\zeta_r - \zeta_s}.$$

On each cell, the unknown function can be represented as

$$\rho(x, t) = \sum_{r=1}^{k+1} \rho_i^r(t) L_r(\zeta^i(x)), \qquad x \in I_i.$$

Here ζ^i is the mapping from I_i to $[-1, 1]$. To determine ρ, it suffices to identify the

coefficients $\vec{\rho}_i = [\rho_i^1 \dots \rho_i^{k+1}]^T$. $\vec{u}_i$ and $\vec{\xi}_i$ are defined in a similar fashion.

The matrix formulation can be written as follows.

$$\frac{d}{dt}\vec{\rho}_i = -\frac{2}{h_i}M^{-1}D^T M\vec{fu}_i + \frac{2}{h_i}M^{-1}B\vec{fu}_i^*, \tag{3.14a}$$

$$\vec{u}_i = -\frac{2}{h_i}M^{-1}D^T M\vec{\xi}_i + \frac{2}{h_i}M^{-1}B\vec{\xi}_i^*, \tag{3.14b}$$

$$\xi_i^r = H'(\rho_i^r) + V(x_i^r) + \sum_{j=1}^{N}\rho_j^r\int_{I_j} W(x_i^r - y)L_r(\zeta^j(y))dy. \tag{3.14c}$$

Here $M = \mathrm{diag}\{w_1, \dots, w_{k+1}\}$ and $B = \mathrm{diag}\{-1, 0, \dots, 0, 1\}$. $D = (D_{rs})$ is the difference matrix, and $D_{rs} = L_s'(\zeta_r)$. $\vec{fu}_i$ is the component-wise product of $\vec{f}_i = [f(\rho_i^1) \dots f(\rho_i^{k+1})]^T$ and $\vec{u}_i$. $\vec{\xi}_i^* = [\widehat{\xi}_{i-\frac{1}{2}}\ 0\ \dots\ 0\ \widehat{\xi}_{i+\frac{1}{2}}]^T$ and $\vec{fu}_i^*$ is defined similarly for $\widehat{fu}$. We remind the readers that one should replace $\int_{I_i}$ by $\widetilde{\int}_{I_i}$ in (3.14c) if W is smooth.

3.2.5 Algorithm flowchart

For simplicity, we only consider the Euler forward time stepping. The algorithm with SSP-RK time discretization can be implemented by repeating the following flowchart in each stage.

1. Use (3.14c) to obtain $\{(\vec{\xi}_i)^n\}$.
2. Evaluate the numerical flux $\{(\vec{\xi}_i^*)^n\}$ and use (3.14b) to update $\{(\vec{u}_i)^n\}$.
3. Evaluate $\{(\vec{f}_i)^n\}$, $\{(\vec{fu}_i)^n\}$ and the numerical flux $\{(\vec{fu}_i^*)^n\}$. Use Euler forward time stepping for (3.14a) to calculate $(\vec{\rho}_i)^{n+1,\mathrm{pre}}$.
4. Evaluate $\{\bar{\rho}_i^n\}$ in each cell.

- If $\{\bar{\rho}_i\}$ is a set of non-negative numbers. Apply the positivity preserving limiter to obtain $\{(\vec{\rho}_i)^{n+1}\}$ and enter the next time level.
- Otherwise halve the time step τ and restart from 1.

Remark 3.2.6. *The main advantage for using Gauss-Lobatto interpolation polynomial basis is that all the needed nodal values are automatically acquired. Hence one can save costs on evaluating the numerical fluxes and applying the positivity-preserving limiter.*

Remark 3.2.7. *For $W \neq 0$, the computational bottleneck is to calculate the convolution in step 1. Suppose the $(k+1)$-point quadrature rule is used in each interval to evaluate the integral, the evaluation of the convolution usually takes $\mathcal{O}(N^2k^2)$ operations in each iteration. However, on uniform meshes, the fast Fourier transform (FFT) can be applied to reduce the cost to $\mathcal{O}(Nk(\log N + k))$. The idea is that, for each fixed i, the convolution can be evaluated by,*

$$(\vec{\xi}_i)^n = \sum_{m=1}^{N} K_m(\vec{\rho}_{i+m})^n, \qquad (K_m)_{rs} = \int_{I_1} W\left(x_i^r - ((m-1)h + y)\right) L_s(\zeta^1(y))dy. \tag{3.15}$$

If the convolution kernel is periodic, then (3.15) can be formulated as the multiplication of an $(N \times (k+1)) \times (N \times (k+1))$ block circulant matrix and a vector. Since each block may not be circulant, only one-dimensional FFT can be applied, which results in $\mathcal{O}\left(Nk(\log N + k)\right)$ operations. However, k is usually a much smaller number compared with N.

Although W is not periodic most of the time, ρ is usually a (numerically) compactly supported function. One only needs to evaluate ξ precisely on the same interval. Hence we can simply extend the problem to a larger domain to adopt the previous procedure. For example, if ρ lives on $[-R, R]$. We can consider its zero extension on $[-2R, 2R]$ and assume everything to be $4R$ periodic. When the FFT algorithm

is used to compute the matrix multiplication, it gives exact ξ on $[-R, R]$, because relevant values of W is unchanged on $[-2R, 2R]$. The computational complexity is still $\mathcal{O}(Nk(\log N + k))$.

Remark 3.2.8. *In our numerical tests, both for one-dimensional and two dimensional cases, we will use a sufficiently small time step to avoid the cell average attaining negative values. Also, $g(\rho) = f(\rho)$ will be used to define the numerical flux, unless otherwise stated.*

3.3 Numerical method: two-dimensional case

In this section, we apply our method to solve two-dimensional problems on Cartesian meshes.

3.3.1 Semi-discrete scheme and entropy inequality

Consider the initial value problem,

$$\begin{cases} \partial_t \rho = \nabla \cdot (f(\rho)\nabla(H'(\rho) + V(x,y) + W * \rho)), \quad x, y \in \Omega \subset \mathbb{R}^2, \quad t > 0, \\ \rho(x, y, 0) = \rho_0(x, y). \end{cases}$$

Here $\Omega = I \times J$ is a rectangular domain and the periodic boundary conditions are applied. Let $I_i \times J_j = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ and $\cup_{i=1}^{N^x} \cup_{j=1}^{N^y} I_i \times J_j$ be a partition of the mesh. The mesh size is denoted by $h = \max_{i,j}\{\sqrt{(h_i^x)^2 + (h_j^y)^2}\}$, where $h_i^x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ and $h_j^y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$. The finite element spaces are defined

as

$$V_h = \{v_h : v_h\big|_{I_i\times J_j} \in Q^k(I_i\times J_j), \text{ for all } i = 1,\dots,N^x, j = 1,\dots,N^y\} \text{ and } \boldsymbol{V}_h = V_h\times V_h. \tag{3.16}$$

Here $Q^k(I_i \times J_j)$ is the tensor product space of $P^k(I_i)$ and $P^k(J_j)$.

The semi-discrete DG scheme is formulated as follows. With V_h and $\boldsymbol{V}_h$ defined in (3.16), one needs to find $\rho_h \in V_h$ and $\boldsymbol{u}_h = (u_h^x, u_h^y) \in \boldsymbol{V}_h$, such that for any test functions $\varphi_h \in V_h$ and $(\psi_h^x, \psi_h^y) \in \boldsymbol{V}_h$,

$$\begin{aligned}\widetilde{\int_{J_j}}\widetilde{\int_{I_i}}(\partial_t\rho_h)\varphi_h dxdy = &-\widetilde{\int_{J_j}}\widetilde{\int_{I_i}} f_h\left(u_h^x\partial_x\varphi_h + u_h^y\partial_y\varphi_h\right)dxdy\\ &+\widetilde{\int_{J_j}}\widehat{fu^x}_{i+\frac12}\varphi_h(x^-_{i+\frac12},y) - \widehat{fu^x}_{i-\frac12}\varphi_h(x^+_{i-\frac12},y)dy\\ &+\widetilde{\int_{I_i}}\widehat{fu^y}_{j+\frac12}\varphi_h(x,y^-_{j+\frac12}) - \widehat{fu^y}_{j-\frac12}\varphi_h(x,y^+_{j-\frac12})dx,\end{aligned} \tag{3.17}$$

$$\begin{aligned}\widetilde{\int_{J_j}}\widetilde{\int_{I_i}}u_h^x\psi_h^x + u_h^y\psi_h^y dxdy = &-\widetilde{\int_{J_j}}\widetilde{\int_{I_i}}\xi_h(\partial_x\psi_h^x + \partial_y\psi_h^y)dxdy\\ &+\widetilde{\int_{J_j}}\widehat{\xi}_{i+\frac12}\psi_h^x(x^-_{i+\frac12},y) - \widehat{\xi}_{i-\frac12}\psi_h^y(x^+_{i-\frac12},y)dy\\ &+\widetilde{\int_{I_i}}\widehat{\xi}_{j+\frac12}\psi_h^y(x,y^-_{j+\frac12}) - \widehat{\xi}_{j-\frac12}\psi_h^y(x,y^+_{j-\frac12})dx.\end{aligned} \tag{3.18}$$

Here, when the interaction potential W is smooth, we set

$$\xi_h = \xi_h(\rho_h, x, y) = H_h' + V + \widetilde{\int_J}\widetilde{\int_I} W(x-\tilde{x}, y-\tilde{y})\rho_h(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}.$$

While for non-smooth W, the quadrature may not achieve sufficient accuracy. Hence

the exact integration is applied

$$\xi_h = \xi_h(\rho_h, x, y) = H_h' + V + \int_J \int_I W(x-\tilde{x}, y-\tilde{y})\rho_h(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}.$$

The numerical fluxes are chosen in the following way,

$$\begin{aligned}
\widehat{\xi}_{i+\frac{1}{2}} = \widehat{\xi}_{i+\frac{1}{2}}(y) &= \frac{1}{2}\left(\xi_h(x^+_{i+\frac{1}{2}}, y) + \xi_h(x^-_{i+\frac{1}{2}}, y)\right), \\
\widehat{\xi}_{j+\frac{1}{2}} = \widehat{\xi}_{j+\frac{1}{2}}(x) &= \frac{1}{2}\left(\xi_h(x, y^+_{j+\frac{1}{2}}) + \xi_h(x, y^-_{j+\frac{1}{2}})\right), \\
\widehat{fu^x}_{i+\frac{1}{2}} = \widehat{fu^x}_{i+\frac{1}{2}}(y) &= \frac{1}{2}\left((f_h u_h)(x^+_{i+\frac{1}{2}}, y) + (f_h u_h)(x^-_{i+\frac{1}{2}}, y)\right) \\
&\quad + \frac{\alpha^x_{i+\frac{1}{2}}}{2}\left(g_h(x^+_{i+\frac{1}{2}}, y) - g_h(x^-_{i+\frac{1}{2}}, y)\right), \\
\alpha^x_{i+\frac{1}{2}} = \alpha^x_{i+\frac{1}{2}}(y) &= \max\{|u_h(x^+_{i+\frac{1}{2}}, y)|, |u_h(x^-_{i+\frac{1}{2}}, y)|\}, \\
\widehat{fu^y}_{j+\frac{1}{2}} = \widehat{fu^y}_{j+\frac{1}{2}}(x) &= \frac{1}{2}\left((f_h u_h)(x, y^+_{j+\frac{1}{2}}) + (f_h u_h)(x, y^-_{j+\frac{1}{2}})\right) \\
&\quad + \frac{\alpha^y_{j+\frac{1}{2}}}{2}\left(g_h(x, y^+_{j+\frac{1}{2}}) - g_h(x, y^-_{j+\frac{1}{2}})\right), \\
\alpha^y_{j+\frac{1}{2}} = \alpha^y_{j+\frac{1}{2}}(x) &= \max\{|u_h(x, y^+_{j+\frac{1}{2}})|, |u_h(x, y^-_{j+\frac{1}{2}})|\},
\end{aligned}$$

with $g_h(x,y) = g(\rho_h(x,y))$, where $g(\rho) = f(\rho)$ if f is increasing and $g(\rho) = C\rho$ if $\frac{f(\rho)}{\rho} \le C$.

For smooth W, one can obtain an entropy inequality as we have done for one dimensional problems.

Theorem 3.3.1. *For smooth interaction kernel W, assume that the semi-discrete scheme defined by (3.17) and (3.18) has a solution, then it satisfies the following entropy inequality.*

$$\frac{d}{dt}\tilde{E} \le -\tilde{I}, \tag{3.19}$$

where

$$\tilde{E} = \int_J^{\sim}\int_I^{\sim} H(\rho_h)dxdy + \int_J^{\sim}\int_I^{\sim} V\rho_h dxdy$$
$$+\frac{1}{2}\int_J^{\sim}\int_I^{\sim}\int_J^{\sim}\int_I^{\sim} W(x-\tilde{x}, y-\tilde{y})\rho_h(x,y)\rho_h(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}dxdy$$

is the discrete entropy and

$$\tilde{I} = \int_J^{\sim}\int_I^{\sim} f_h|\boldsymbol{u}_h|^2 dxdy$$

is the associated discrete entropy dissipation. Moreover, assuming g and H' are strictly increasing, if α^x and α^y are all positive and $\rho_h \geq 0$ is a stationary state of the semi-discrete scheme, then ρ_h is continuous and the piecewise polynomial interpolation of ξ_h is constant in each connected component of the support of ρ_h.

Proof. We will focus on the entropy-entropy dissipation relationship and the proof of the second part of the theorem is omitted. Using the symmetry of W, we have

$$\frac{d}{dt}\frac{1}{2}\int_J^{\sim}\int_I^{\sim}\int_J^{\sim}\int_I^{\sim} W(x-\tilde{x}, y-\tilde{y})\rho_h(x,y)\rho_h(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}dxdy$$
$$=\int_J^{\sim}\int_I^{\sim} \partial_t\rho_h(x,y)\left(\int_J^{\sim}\int_I^{\sim} W(x-\tilde{x}, y-\tilde{y})\rho_h(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}\right)dxdy.$$

Hence,

$$\begin{aligned}
\frac{d}{dt}\tilde{E} =& \widetilde{\int_J}\widetilde{\int_I}\partial_t\rho_h(x,y)\left(H_h' + V + \widetilde{\int_J}\widetilde{\int_I}W(x-\tilde{x},y-\tilde{y})\rho_h(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}\right)dxdy \\
=& \widetilde{\int_J}\widetilde{\int_I}\partial_t\rho_h\xi_h dxdy = \sum_{i=1}^{N^x}\sum_{j=1}^{N^y}\widetilde{\int_{J_j}}\widetilde{\int_{I_i}}\partial_t\rho_h\mathcal{I}(\xi_h)dxdy \\
=& \sum_{i=1}^{N^x}\sum_{j=1}^{N^y}\left(-\widetilde{\int_{J_j}}\left(\widetilde{\int_{I_i}}\mathcal{I}(f_h u_h^x)\partial_x\mathcal{I}(\xi_h)dx\right)dy - \widetilde{\int_{I_i}}\left(\widetilde{\int_{J_j}}\mathcal{I}(f_h u_h^y)\partial_y\mathcal{I}(\xi_h)dy\right)dx\right. \\
&+ \widetilde{\int_{J_j}}\widehat{fu^x}_{i+\frac{1}{2}}\xi_h(x^-_{i+\frac{1}{2}},y) - \widehat{fu^x}_{i-\frac{1}{2}}\xi_h(x^+_{i-\frac{1}{2}},y)dy \\
&\left.+\widetilde{\int_{I_i}}\widehat{fu^y}_{j+\frac{1}{2}}\xi_h(x,y^-_{j+\frac{1}{2}}) - \widehat{fu^y}_{j-\frac{1}{2}}\xi_h(x,y^+_{j-\frac{1}{2}})dx\right).
\end{aligned}$$

For fixed y, $\mathcal{I}(f_h u_h^x)\partial_x\mathcal{I}(\xi_h)$ is a polynomial of degree $2k-1$ with respect to x. Hence the Gauss-Lobatto quadrature with $k+1$ nodes is exact. We replace the quadrature with the exact integral, integrate by parts and then change back to the quadrature. The same argument also applies to the second integral. One can then obtain,

$$\begin{aligned}
\frac{d}{dt}\tilde{E} =& \sum_{i=1}^{N^x}\sum_{j=1}^{N^y}\left(\widetilde{\int_{J_j}}\left(\widetilde{\int_{I_i}}\partial_x\mathcal{I}(f_h u_h^x)\mathcal{I}(\xi_h)dx\right)dy + \widetilde{\int_{I_i}}\left(\widetilde{\int_{J_j}}\partial_y\mathcal{I}(f_h u_h^y)\mathcal{I}(\xi_h)dy\right)dx\right. \\
&- \widetilde{\int_{J_j}}(f_h u_h^x\xi_h)(x^-_{i+\frac{1}{2}},y) - (f_h u_h^x\xi_h)(x^+_{i-\frac{1}{2}},y)dy \\
&- \widetilde{\int_{I_i}}(f_h u_h^y\xi_h)(x,y^-_{j+\frac{1}{2}}) - (f_h u_h^y\xi_h)(x,y^+_{j-\frac{1}{2}})dx \\
&+ \widetilde{\int_{J_j}}\widehat{fu^x}_{i+\frac{1}{2}}\xi_h(x^-_{i+\frac{1}{2}},y) - \widehat{fu^x}_{i-\frac{1}{2}}\xi_h(x^+_{i-\frac{1}{2}},y)dy \\
&\left.+\widetilde{\int_{I_i}}\widehat{fu^y}_{j+\frac{1}{2}}\xi_h(x,y^-_{j+\frac{1}{2}}) - \widehat{fu^y}_{j-\frac{1}{2}}\xi_h(x,y^+_{j-\frac{1}{2}})dx\right).
\end{aligned}$$

Use the scheme (3.18) one can get

$$
\begin{aligned}
\frac{d}{dt}\tilde{E} =& -\int_J\!\!\!\!\sim\int_I\!\!\!\!\sim f|\boldsymbol{u}_h|^2 dxdy + \sum_{i=1}^{N^x}\sum_{j=1}^{N^y}\left(\int_{J_j}\!\!\!\!\sim \widehat{\xi}_{i+\frac{1}{2}}(f_h u_h^x)(x^-_{i+\frac{1}{2}},y) - \widehat{\xi}_{i-\frac{1}{2}}(f_h u_h^x)(x^+_{i-\frac{1}{2}},y)dy\right.\\
&+\int_{I_i}\!\!\!\!\sim \widehat{\xi}_{j+\frac{1}{2}}(f_h u_h^y)(x,y^-_{j+\frac{1}{2}}) - \widehat{\xi}_{j-\frac{1}{2}}(f_h u_h^y)(x,y^+_{j-\frac{1}{2}})dx\\
&-\int_{J_j}\!\!\!\!\sim (f_h u_h^x \xi_h)(x^-_{i+\frac{1}{2}},y) - (f_h u_h^x\xi_h)(x^+_{i-\frac{1}{2}},y)dy\\
&-\int_{I_i}\!\!\!\!\sim (f_h u_h^y \xi_h)(x,y^-_{j+\frac{1}{2}}) - (f_h u_h^y\xi_h)(x,y^+_{j-\frac{1}{2}})dx\\
&+\int_{J_j}\!\!\!\!\sim \widehat{fu^x}_{i+\frac{1}{2}}\xi_h(x^-_{i+\frac{1}{2}},y) - \widehat{fu^x}_{i-\frac{1}{2}}\xi_h(x^+_{i-\frac{1}{2}},y)dy\\
&\left.+\int_{I_i}\!\!\!\!\sim \widehat{fu^y}_{j+\frac{1}{2}}\xi_h(x,y^-_{j+\frac{1}{2}}) - \widehat{fu^y}_{j-\frac{1}{2}}\xi_h(x,y^+_{j-\frac{1}{2}})dx\right)\\
=& -\int_J\!\!\!\!\sim\int_I\!\!\!\!\sim f|\boldsymbol{u}_h|^2 dxdy\\
&-\sum_{i=1}^{N^x}\sum_{j=1}^{N^y}\left(\int_{J_j}\!\!\!\!\sim \frac{\alpha^x_{i+\frac{1}{2}}(y)}{2}\left(g_h(x^+_{i+\frac{1}{2}},y)-g_h(x^-_{i+\frac{1}{2}},y)\right)\left(\xi_h(x^+_{i+\frac{1}{2}},y)-\xi_h(x^-_{i+\frac{1}{2}},y)\right)dy\right.\\
&\left.+\int_{I_i}\!\!\!\!\sim \frac{\alpha^y_{j+\frac{1}{2}}(x)}{2}\left(g_h(x,y^+_{j+\frac{1}{2}})-g_h(x,y^-_{j+\frac{1}{2}})\right)\left(\xi_h(x,y^+_{j+\frac{1}{2}})-\xi_h(x,y^-_{j+\frac{1}{2}})\right)dx\right).
\end{aligned}
$$

By our choices of g, the strict monotonicity of H' and the fact that V and W are single-valued, the last term is non-positive, which gives (3.19). □

3.3.2 Time discretization and preservation of positivity

It suffices to ensure the positivity-preserving property of the Euler forward scheme. The high-order case is automatically taken care of by SSP-RK time discretization.

The first step is to show that, provided the solution at the current time level

is non-negative, the cell average at next time level will also be non-negative, if a specific time step restriction is satisfied.

Lemma 3.3.1. *Let* $\lambda_i^x = \frac{\tau}{h_i^x}$ *and* $\lambda_j^y = \frac{\tau}{h_j^y}$. *Suppose* $\rho_h^n(x_i^r, y_j^s) \geq 0$, $r, s = 1, \ldots, k+1$ *and*

$$\lambda_i^x \leq \min_s \left\{ \left(\frac{w_1 \rho}{2(fu^x + \alpha^x g)} \right) (x_{i-\frac{1}{2}}^+, y_j^s), \left(\frac{w_{k+1} \rho}{2(\alpha^x g - fu^x)} \right) (x_{i+\frac{1}{2}}^-, y_j^s) \right\}, \tag{3.20a}$$

$$\lambda_j^y \leq \min_r \left\{ \left(\frac{w_1 \rho}{2(fu^y + \alpha^y g)} \right) (x_i^r, y_{j-\frac{1}{2}}^+), \left(\frac{w_{k+1} \rho}{2(\alpha^y g - fu^y)} \right) (x_i^r, y_{j+\frac{1}{2}}^-) \right\}, \tag{3.20b}$$

then the solution $(\rho_h)_{i,j}^{n+1,\mathrm{pre}}$ *obtained from* (3.17) *satisfies* $(\bar{\rho}_h)_{i,j}^{n+1,\mathrm{pre}} \geq 0$. *Here, in the constraint of* λ_i^x *and* λ_j^x, *we formally denote by* $\frac{0}{0} := +\infty$.

Proof. As before, we drop all the subscripts h in this proof. The superscript n will also be omitted for simplicity. Take $\varphi = 1$ in (3.17), we have

$$\bar{\rho}_{i,j}^{n+1,\mathrm{pre}} = \bar{\rho}_{i,j} + \frac{\tau}{h_i^x h_j^y} \int_{J_j}^{\sim} \widehat{fu^x}_{i+\frac{1}{2}} - \widehat{fu^x}_{i-\frac{1}{2}} dy + \frac{\tau}{h_i^x h_j^y} \int_{I_i}^{\sim} \widehat{fu^y}_{j+\frac{1}{2}} - \widehat{fu^y}_{j-\frac{1}{2}} dx.$$

Note that

$$\bar{\rho}_{i,j} = \frac{1}{h_i^x h_j^y} \int_{J_j}^{\sim} \int_{I_i}^{\sim} \rho dx dy = \frac{1}{4} \sum_{r=1}^{k+1} \sum_{s=1}^{k+1} w_r w_s \rho(x_i^r, y_j^s).$$

Then we have

$$
\begin{aligned}
\bar{\rho}_{i,j}^{n+1,\mathrm{pre}} &= \frac{1}{4}\sum_{r=1}^{k+1}\sum_{s=1}^{k+1} w_r w_s \rho(x_i^r, y_j^s) \\
&+ \frac{\lambda_i^x}{4}\sum_{s=1}^{k+1} w_s \left((fu^x)(x_{i+\frac{1}{2}}^+, y_j^s) + (fu^x)(x_{i+\frac{1}{2}}^-, y_j^s) + \alpha_{i+\frac{1}{2}}^x \left(g_h(x_{i+\frac{1}{2}}^+, y_j^s) - g_h(x_{i+\frac{1}{2}}^-, y_j^s) \right) \right) \\
&- \frac{\lambda_i^x}{4}\sum_{s=1}^{k+1} w_s \left((fu^x)(x_{i-\frac{1}{2}}^+, y_j^s) + (fu^x)(x_{i-\frac{1}{2}}^-, y_j^s) + \alpha_{i-\frac{1}{2}}^x \left(g_h(x_{i-\frac{1}{2}}^+, y_j^s) - g_h(x_{i-\frac{1}{2}}^-, y_j^s) \right) \right) \\
&+ \frac{\lambda_j^y}{4}\sum_{r=1}^{k+1} w_r \left((fu^y)(x_i^r, y_{j+\frac{1}{2}}^+) + (fu^y)(x_i^r, y_{j+\frac{1}{2}}^-) + \alpha_{j+\frac{1}{2}}^y \left(g_h(x_i^r, y_{j+\frac{1}{2}}^+) - g_h(x_i^r, y_{j+\frac{1}{2}}^-) \right) \right) \\
&- \frac{\lambda_j^y}{4}\sum_{r=1}^{k+1} w_r \left((fu^y)(x_i^r, y_{j-\frac{1}{2}}^+) + (fu^y)(x_i^r, y_{j-\frac{1}{2}}^-) + \alpha_{j-\frac{1}{2}}^y \left(g_h(x_i^r, y_{j-\frac{1}{2}}^+) - g_h(x_i^r, y_{j-\frac{1}{2}}^-) \right) \right) \\
&\geq \frac{1}{4}\sum_{r=2}^{k}\sum_{s=2}^{k} w_r w_s \rho(x_i^r, y_j^s) \\
&+ \sum_{s=1}^{k+1} w_s \left(\frac{w_{k+1}}{8}\rho(x_{i+\frac{1}{2}}^-, y_j^s) + \frac{\lambda_i^x}{4}\left((fu^x)(x_{i+\frac{1}{2}}^-, y_j^s) - \alpha_{i+\frac{1}{2}}^x g_h(x_{i+\frac{1}{2}}^-, y_j^s) \right) \right) \\
&+ \sum_{s=1}^{k+1} w_s \left(\frac{w_1}{8}\rho(x_{i-\frac{1}{2}}^+, y_j^s) - \frac{\lambda_i^x}{4}\left((fu^x)(x_{i-\frac{1}{2}}^+, y_j^s) + \alpha_{i-\frac{1}{2}}^x g_h(x_{i-\frac{1}{2}}^+, y_j^s) \right) \right) \\
&+ \sum_{r=1}^{k+1} w_r \left(\frac{w_{k+1}}{8}\rho(x_i^r, y_{j+\frac{1}{2}}^-) + \frac{\lambda_j^y}{4}\left((fu^y)(x_i^r, y_{j+\frac{1}{2}}^-) - \alpha_{j+\frac{1}{2}}^y g_h(x_i^r, y_{j+\frac{1}{2}}^-) \right) \right) \\
&+ \sum_{r=1}^{k+1} w_r \left(\frac{w_1}{8}\rho(x_i^r, y_{j-\frac{1}{2}}^+) - \frac{\lambda_j^y}{4}\left((fu^y)(x_i^r, y_{j-\frac{1}{2}}^+) + \alpha_{j-\frac{1}{2}}^y g_h(x_i^r, y_{j-\frac{1}{2}}^+) \right) \right) \\
&+ \frac{\lambda_i^x}{4}\sum_{s=1}^{k+1} w_s \left((fu^x)(x_{i+\frac{1}{2}}^+, y_j^s) + \alpha_{i+\frac{1}{2}}^x g_h(x_{i+\frac{1}{2}}^+, y_j^s) \right) \\
&- \frac{\lambda_i^x}{4}\sum_{s=1}^{k+1} w_s \left((fu^x)(x_{i-\frac{1}{2}}^-, y_j^s) - \alpha_{i-\frac{1}{2}}^x g_h(x_{i-\frac{1}{2}}^-, y_j^s) \right) \\
&+ \frac{\lambda_j^y}{4}\sum_{r=1}^{k+1} w_r \left((fu^y)(x_i^r, y_{j+\frac{1}{2}}^+) + \alpha_{j+\frac{1}{2}}^y g_h(x_i^r, y_{j+\frac{1}{2}}^+) \right) \\
&- \frac{\lambda_j^y}{4}\sum_{r=1}^{k+1} w_r \left((fu^y)(x_i^r, y_{j-\frac{1}{2}}^-) - \alpha_{j-\frac{1}{2}}^y g_h(x_i^r, y_{j-\frac{1}{2}}^-) \right).
\end{aligned}
$$

The first term is automatically non-negative, since the weights $w_r, w_s \geq 0$ and

the nodal values $\rho(x_i^r, y_j^s) \geq 0$. The positivity of the last four terms is guaranteed by our choice of α and g. One only needs (3.20) to ensure the second and the third term to be non-negative. And as before, one can check the convention $\frac{0}{0} = +\infty$ does make sense. Hence $\bar{\rho}_{i,j}^{n+1} \geq 0$ under the prescribed time step constraint. □

Then, as we have done in the one-dimensional case, a scaling limiter is applied to sure the numerical polynomial solution takes non-negative values at the quadrature points. Hence the assumption in Lemma 3.3.1 is met and the fully discretized scheme is positivity-preserving.

Theorem 3.3.2. *Let*

$$\rho_h^{n+1}(x_i^r, y_j^s) = (\bar{\rho}_h)_{i,j}^{n+1,\text{pre}} + \theta_{i,j}\left(\rho_h^{n+1,\text{pre}}(x_i^r, y_j^s) - (\bar{\rho}_h)_{i,j}^{n+1,\text{pre}}\right), \qquad \forall r, s = 1, \ldots, k+1,$$

with $\theta_{i,j} = \min\{\frac{(\bar{\rho}_h)_{i,j}^{n+1,\text{pre}}}{(\bar{\rho}_h)_{i,j}^{n+1,\text{pre}} - m_{i,j}}, 1\}$ *and* $m_{i,j} = \min\{\rho_h^{n+1,\text{pre}}(x_i^r, y_j^s)\}_{r,s=1}^{k+1}$. *Then we have* $\rho_h^{n+1}(x_i^r, y_j^s) \geq 0, \forall r, s = 1, \ldots, k+1$, $\bar{\rho}_h^{n+1} = \bar{\rho}_h^{n+1,\text{pre}}$. *Hence the resulting fully discretized scheme using Euler forward or SSP-RK time discretization preserves the non-negativity of the solution, if*

$$\tau \leq \min_{i,j}\{h_i^x, h_j^y\} \cdot \min_{i,j,s,r}\left\{\left(\frac{w_1\rho}{2(fu^x + \alpha^x g)}\right)(x_{i-\frac{1}{2}}^+, y_j^s), \left(\frac{w_{k+1}\rho}{2(\alpha^x g - fu^x)}\right)(x_{i+\frac{1}{2}}^-, y_j^s)\right.$$
$$\left.\left(\frac{w_1\rho}{2(fu^y + \alpha^y g)}\right)(x_i^r, y_{j-\frac{1}{2}}^+), \left(\frac{w_{k+1}\rho}{2(\alpha^y g - fu^y)}\right)(x_i^r, y_{j+\frac{1}{2}}^-)\right\}.$$

Remark 3.3.1. *As that in the one-dimensional case, we expect the time step constraint to be* $\tau \leq ch$ *if* $H' \equiv 0$ *and* $\tau \leq ch^2$ *otherwise.*

3.4 One-dimensional numerical tests

3.4.1 Accuracy tests

In this part, we examine the accuracy of the numerical schemes with P^1, P^2, P^3 and P^4 elements. The error is measured in the discrete norms.

$$
\begin{aligned}
e_{L^1} &= \sum_{i=1}^{N} \int_{I_i}^{\sim} |u_h(x,t) - u(x,t)|\, dx, \\
e_{L^2} &= \sqrt{\sum_{i=1}^{N} \int_{I_i}^{\sim} |u_h(x,t) - u(x,t)|^2\, dx}, \\
e_{L^\infty} &= \max_{x \in \{x_i^r\}_{i,r}} |u_h(x,t) - u(x,t)|\,.
\end{aligned}
$$

Example 3.4.1.1 (Advection equation)**.** The first numerical test is done for the linear advection equation

$$
\begin{cases}
\partial_t \rho = \partial_x \rho, \quad x \in [-\pi, \pi], \\
\rho(x,0) = 1 + \sin(x).
\end{cases}
$$

The problem has an exact solution $u(x,t) = 1 + \sin(x+t)$. In this test, $f(\rho) = \rho$, $H'(\rho) = 0$, $V(x) = x$ and $W(x) = 0$. To be consistent at the boundaries, one needs to manually impose $\widehat{\xi}(\pi) = \pi$ and $\widehat{\xi}(-\pi) = -\pi$. (This will gives $u \equiv 1$ and the scheme is equivalent to the usual upwinding DG method with a *mass lumping* treatment.) We compute up to $t = 2$ and the time step is $\tau = 0.02h^2$.

Due to our choice of the initial condition, the solution has point vacuum and the numerical solution may become negative in its neighborhood. We perform numerical tests without and with the positivity-preserving limiter and the results are listed

in Table 3.1 and Table 3.2 respectively. As one can see, without the limiter, the convergence rate is optimal. The rate degenerates a little bit for P^1 and P^4 schemes when one applies the limiter.

Table 3.1: Accuracy test of the linear advection equation in Example 3.4.1.1: without limiters.

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	20	1.555E-01	-	6.893E-02	-	4.169E-02	-
	40	4.039E-02	1.94	1.793E-02	1.94	1.062E-02	1.97
	80	1.023E-02	1.98	4.536E-03	1.98	2.657E-03	2.00
	160	2.567E-03	1.99	1.138E-03	1.99	6.633E-04	2.00
2	20	1.837E-03	-	1.042E-03	-	1.241E-03	-
	40	2.225E-04	3.05	1.306E-04	3.00	1.588E-04	2.97
	80	2.738E-05	3.02	1.633E-05	3.00	2.002E-05	2.99
	160	3.394E-06	3.01	2.041E-06	3.00	2.513E-06	2.99
3	20	2.995E-05	-	1.763E-05	-	2.705E-05	-
	40	1.877E-06	4.00	1.104E-06	4.00	1.700E-06	3.99
	80	1.177E-07	4.00	6.899E-08	4.00	1.061E-07	4.00
	160	7.363E-09	4.00	4.312E-09	4.00	6.622E-09	4.00
4	20	4.510E-08	-	2.528E-07	-	4.294E-07	-
	40	1.330E-09	5.08	7.985E-09	4.98	1.432E-08	4.91
	80	4.153E-11	5.00	2.470E-10	5.01	4.443E-10	5.01
	160	1.297E-12	5.00	7.719E-12	5.00	1.390E-11	5.00

Example 3.4.1.2 (Heat equation)**.** We then examine the heat equation,

$$\begin{cases} \partial_t \rho = \partial_{xx} \rho, \quad x \in [-\pi, \pi], \\ \rho(x, 0) = 2 + \sin(x), \end{cases}$$

with periodic boundary conditions. The decomposition of the equation into the desired form is not unique. Let us consider two test cases,

(i) $f(\rho) = \rho$, $H'(\rho) = \log(\rho)$ and $V(x) = W(x) = 0$,

(ii) $f(\rho) = \sqrt{\rho}$, $H'(\rho) = 2\sqrt{\rho}$ and $V(x) = W(x) = 0$.

Note that for both of the cases, the schemes are nonlinear, although the original problem is linear. The exact solution to the problem is $\rho(x, t) = 2 + e^{-t}\sin(x)$. It attains values away from 0. Hence the positivity-preserving limiter will not be

Table 3.2: Accuracy test of the linear advection equation in Example 3.4.1.1: with limiters.

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	20	1.494E-01	-	6.679E-02	-	4.333E-02	-
	40	4.009E-02	1.90	1.814E-02	1.88	1.368E-02	1.66
	80	1.047E-02	1.94	4.734E-03	1.94	5.148E-03	1.41
	160	2.686E-03	1.96	1.227E-03	1.95	1.761E-03	1.55
2	20	1.835E-03	-	1.048E-03	-	1.241E-03	-
	40	2.231E-04	3.04	1.307E-04	3.00	1.588E-04	2.97
	80	2.743E-05	3.02	1.633E-05	3.00	2.002E-05	2.99
	160	3.395E-06	3.01	2.041E-06	3.00	2.513E-06	2.99
3	20	3.136E-05	-	1.885E-05	-	3.593E-05	-
	40	1.990E-06	3.98	1.173E-06	4.01	1.827E-06	4.30
	80	1.217E-07	4.03	7.196E-08	4.02	1.629E-07	3.49
	160	7.591E-09	4.00	4.463E-09	4.01	9.453E-09	4.11
4	20	1.661E-06	-	1.605E-06	-	2.855E-06	-
	40	5.831E-08	4.83	7.580E-08	4.40	2.044E-07	3.80
	80	1.996E-09	4.87	3.591E-09	4.40	1.385E-08	3.88
	160	7.072E-11	4.82	1.702E-10	4.40	9.038E-10	3.94

activated as long as the numerical approximations are reasonably accurate. We compute to $t = 2$ and use a time step $\tau = 0.01h^2$ for accuracy tests. Error tables are given in Table 3.3 and Table 3.4 respectively. According to our numerical results, we see that different choices of decomposition lead to negligible difference. For both of the tests, P^2 and P^4 schemes are of the optimal rate of convergence, but the order for P^1 and P^3 schemes seems to be reduced. For this specific problem, the order degeneracy should be due to the fact that the constant in the Lax-Friedrichs flux is not large enough. We refer to our numerical tests and comments in Section 4.4.1 for further details.

Example 3.4.1.3 (Evolution equation with interaction potentials)**.** Our final tests are designed for problems with interaction potentials.

$$\begin{cases} \partial_t \rho = \partial_x \left(\rho \partial_x (W * \rho) \right), \quad x \in [-1, 1], \\ \rho(x, 0) = \left(\dfrac{e^{-\frac{x^2}{0.1}}}{\sqrt{0.1\pi}} \right)^4. \end{cases}$$

Table 3.3: Accuracy test of the heat equation in Example 3.4.1.2: (i) $\partial_t \rho = \partial_x \left(\rho \partial_x \log(\rho)\right)$.

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	20	7.95669E-03	-	3.69447E-03	-	2.28808E-03	-
	40	2.00183E-03	1.99	9.88037E-04	1.90	6.64459E-04	1.78
	80	5.52074E-04	1.86	2.83256E-04	1.80	2.02063E-04	1.72
	160	1.72193E-04	1.68	8.55650E-05	1.73	6.22412E-05	1.70
	320	5.38010E-05	1.68	2.59228E-05	1.72	1.87767E-05	1.73
2	20	1.534E-04	-	9.350E-05	-	9.010E-05	-
	40	1.679E-05	3.19	1.131E-05	3.05	1.108E-05	3.02
	80	1.956E-06	3.10	1.403E-06	3.01	1.382E-06	3.00
	160	2.358E-07	3.05	1.751E-07	3.00	1.727E-07	3.00
	320	2.895E-08	3.02	2.188E-08	3.00	2.158E-08	3.00
3	20	1.622E-05	-	7.803E-06	-	7.896E-06	-
	40	1.805E-06	3.17	8.674E-07	3.17	8.771E-07	3.17
	80	1.851E-07	3.29	8.922E-08	3.28	9.100E-08	3.27
	160	1.713E-08	3.43	8.331E-09	3.42	8.658E-09	3.39
	320	1.425E-09	3.59	7.054E-10	3.56	7.564E-10	3.52
4	20	3.576E-08	-	2.373E-08	-	4.104E-08	-
	40	1.040E-09	5.10	7.208E-10	5.04	1.255E-09	5.03
	80	3.152E-11	5.04	2.237E-11	5.01	3.900E-11	5.01
	160	9.711E-13	5.02	6.981E-13	5.00	1.218E-12	5.00
	320	3.014E-14	5.01	2.181E-14	5.00	3.804E-14	5.00

Periodic boundary conditions are applied for the problem. We consider both the smooth case $W(x) = 0.2\frac{e^{-\frac{x^2}{0.1}}}{\sqrt{0.1\pi}}$ and the nonsmooth case $W(x) = \max\{0.2 - |x|, 0\}$. The convolution integrals are evaluated by quadrature and exact integration respectively. We compute to $t = 0.2$ with $\tau = 0.4h^2$ and use the numerical solution with P^4 elements on $N = 1280$ mesh as the reference solution to evaluate the accuracy. We do not impose the limiter in the tests. The order of accuracy seems to be optimal for odd k, while the order degenerates for even k. See Table 3.5 and Table 3.6. We speculate that the insufficient accuracy of the Gauss-Lobatto quadrature. The $(k+1)$-point Gauss-Lobatto quadrature is only exact for polynomials of degree $2k-1$, while in general, a quadrature with algebraic degree of accuracy $2k$ is expected (see [26] and [44]). In [22], reduced accuracy is reported when the authors solve conservation laws with a DG method using suboptimal quadrature rules. The

Table 3.4: Accuracy test of the heat equation in Example 3.4.1.2: (ii) $\partial_t \rho = \partial_x \left(\sqrt{\rho}\partial_x(2\sqrt{\rho})\right)$.

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	20	8.427E-03	-	3.783E-03	-	2.239E-03	-
	40	2.107E-03	2.01	9.678E-04	1.97	6.069E-04	1.88
	80	5.314E-04	2.00	2.583E-04	1.92	1.743E-04	1.80
	160	1.430E-04	1.89	7.340E-05	1.83	5.221E-05	1.74
	320	4.390E-05	1.70	2.195E-05	1.74	1.592E-05	1.71
2	20	1.572E-04	-	9.394E-05	-	8.929E-05	-
	40	1.699E-05	3.21	1.132E-05	3.05	1.101E-05	3.02
	80	1.971E-06	3.11	1.402E-06	3.01	1.369E-06	3.01
	160	2.369E-07	3.06	1.749E-07	3.00	1.711E-07	3.00
	320	2.904E-08	3.03	2.185E-08	3.00	2.138E-08	3.00
3	20	1.740E-05	-	8.302E-06	-	8.097E-06	-
	40	2.042E-06	3.09	9.706E-07	3.10	9.538E-07	3.09
	80	2.265E-07	3.17	1.075E-07	3.17	1.054E-07	3.18
	160	2.319E-08	3.29	1.101E-08	3.29	1.078E-08	3.29
	320	2.147E-09	3.43	1.018E-08	3.43	9.981E-10	3.43
4	20	3.528E-08	-	2.189E-08	-	3.378E-08	-
	40	1.033E-09	5.09	6.681E-10	5.03	1.050E-09	5.01
	80	3.138E-11	5.04	2.076E-11	5.01	3.269E-11	5.01
	160	9.679E-13	5.02	6.478E-13	5.00	1.022E-12	5.00
	320	3.006E-14	5.01	2.024E-14	5.00	3.192E-14	5.00

degeneracy may be due to the same reason here.

3.4.2 Fokker-Planck type equations

Example 3.4.2.1 (Porous media equation)**.** Let us consider the porous media equation

$$\partial_t \rho = \partial_x \left(\rho \partial_x \left(\frac{m}{m-1} \rho^{m-1} + \frac{x^2}{2} \right) \right),$$

which is used to model the flow of a gas through a porous interface. The equation fits our model with $H(\rho) = \frac{1}{m-1}\rho^m$ and $V(x) = \frac{x^2}{2}$. The property of the equation is studied by Carrillo and Toscani in [21] using an entropy approach. They have proved that the equation converges to a unique steady state given by a Barenblatt-Pattle

Table 3.5: Accuracy test for Example 3.4.1.3 with a smooth kernel $W(x) = 0.2\frac{e^{-\frac{x^2}{0.1}}}{\sqrt{0.1\pi}}$.

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	40	8.598E-02	-	9.874E-02	-	1.944E-01	-
	80	2.866E-02	1.59	3.552E-02	1.48	7.714E-02	1.33
	160	8.129E-03	1.82	1.064E-02	1.74	2.432E-02	1.67
	320	2.130E-03	1.93	2.842E-03	1.90	6.868E-03	1.82
	640	5.394E-04	1.98	7.253E-04	1.97	1.790E-03	1.94
2	40	2.615E-02	-	6.900E-02	-	4.582E-01	-
	80	4.037E-03	2.70	1.309E-02	2.40	1.224E-01	1.91
	160	6.680E-04	2.60	2.369E-03	2.47	3.104E-02	1.98
	320	1.271E-04	2.39	4.266E-04	2.47	7.807E-03	1.99
3	40	7.679E-04	-	1.255E-03	-	1.004E-02	-
	80	4.486E-05	4.10	6.716E-05	4.22	6.499E-04	3.95
	160	2.450E-06	4.19	3.570E-06	4.23	4.094E-05	3.99
	320	1.312E-07	4.22	1.934E-07	4.21	2.509E-06	4.03
	640	7.257E-09	4.18	1.069E-08	4.18	1.019E-07	4.62
4	40	8.266E-04	-	2.369E-04	-	2.832E-03	-
	80	3.017E-05	4.78	1.146E-05	4.37	1.955E-04	3.86
	160	1.110E-06	4.76	5.189E-07	4.46	1.258E-05	3.96
	320	4.332E-08	4.68	2.407E-08	4.43	8.466E-07	3.89

type formula,

$$\rho_\infty(x) = \left(C - \frac{m-1}{2m}|x|^2\right)^{\frac{1}{m-1}}.$$

Here the constant C is determined by ensuring the mass conservation. Furthermore, the relative entropy $E(t|\infty) = E(\rho(t)) - E(\rho_\infty)$ decays exponentially, $E(t|\infty) \leq E(0|\infty)e^{-2t}$ and the rate -2 is sharp.

We particularly choose $m = 2$ in our numerical test.

$$\begin{cases} \partial_t\rho = \partial_x(\rho\partial_x(2\rho + \frac{x^2}{2})), \quad x \in [-2, 2], \\ \rho(x, 0) = \max\{1 - |x|, 0\}, \end{cases}$$

Table 3.6: Accuracy test for Example 3.4.1.3 with a nonsmooth kernel $W(x) = \max\{0.2 - |x|, 0\}$.

k	N	L^1error	order	L^2 error	order	L^∞ error	order
1	40	1.072E-01	-	1.292E-01	-	2.764E-01	-
	80	3.794E-02	1.50	5.095E-02	1.34	1.462E-01	0.92
	160	1.133E-02	1.74	1.671E-02	1.61	5.300E-02	1.46
	320	3.064E-03	1.89	4.805E-03	1.80	1.613E-02	1.72
	640	7.809E-04	1.97	1.267E-03	1.92	4.463E-03	1.85
2	40	2.779E-02	-	3.634E-02	-	1.352E-01	-
	80	5.066E-03	2.46	7.037E-03	2.37	3.271E-02	2.05
	160	1.135E-03	2.16	1.619E-03	2.12	7.916E-03	2.05
	320	2.774E-04	2.03	4.085E-04	1.99	2.206E-03	1.84
	640	7.017E-05	1.98	1.049E-04	1.96	5.825E-04	1.92
3	40	1.838E-03	-	3.473E-03	-	2.408E-02	-
	80	1.811E-04	3.34	3.735E-04	3.22	3.650E-03	2.72
	160	1.088E-05	4.06	2.334E-05	4.00	2.541E-04	3.84
	320	6.814E-07	4.00	1.527E-06	3.93	1.859E-05	3.77
4	40	3.851E-04	-	8.153E-04	-	7.931E-03	-
	80	1.717E-05	4.49	2.443E-05	5.06	1.304E-04	5.93
	160	1.008E-06	4.09	1.690E-06	3.85	1.366E-05	3.26
	320	5.854E-08	4.11	1.331E-07	3.67	1.520E-06	3.17

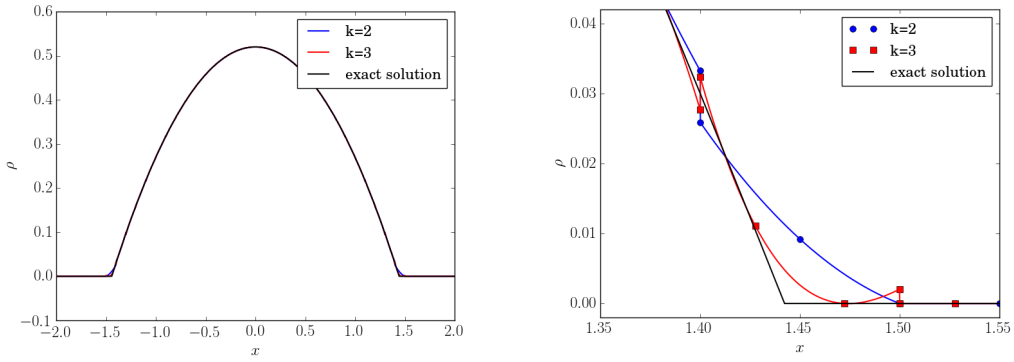
with periodic boundary conditions. The stationary solution is

$$\rho_\infty = \max\left\{\left(\frac{3}{8}\right)^{\frac{2}{3}} - \frac{x^2}{4}, 0\right\}.$$

We compute up to $t = 5$ with the number of cells $N = 40$ and the time step $\tau = 0.005h^2$. The positivity preserving limiter keeps being invoked in the test. (If we manually turn off the limiter, the solution blows up.) The profiles of the solution polynomials with $k = 2$ and $k = 3$ are given in Figure 3.1a. As one can see, the numerical solutions converge well to the exact steady state in the smooth region. We also provide a zoomed-in snapshot to exhibit the capture of singularity near $x = 3^{\frac{1}{3}}$ in Figure 3.1b.

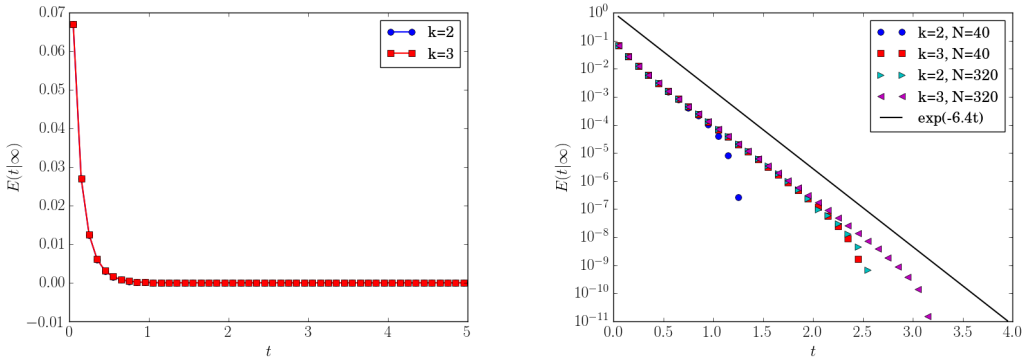
In Figure 3.2a and Figure 3.2b, we plot the entropy and the relative entropy respectively. The entropy profiles for $k = 2$ and $k = 3$ are almost identical, and

they both approach to zero. We then evaluate the relative entropy using that of the numerical steady state as a reference. But we can only plot up to a certain time, before the relative entropy becomes slightly negative, since the entropy decay of the semi-discrete scheme may not be preserved after applying the time discretization and the limiter. We stop the plotting after negative relative entropy appears, not only because it can not be depicted in the logarithm scale, but also because the unresolved tail should be regarded as a discretization error. If we choose $N = 320$ while keeping $\tau = 0.005h^2$, the decay will continue further.



(a) Steady state. (b) Zoomed in figure near $x = 3\frac{1}{3}$.

Figure 3.1: Profiles of the numerical steady states for (3.4.2.1) in Example 3.4.2.1, with $\rho(x, 0) = \max\{1 - |x|, 0\}$.



(a) $\rho(x, 0) = \max\{1 - |x|, 0\}$. (b) $\rho(x, 0) = \max\{1 - |x|, 0\}$.

Figure 3.2: The entropy and the relative entropy for (3.4.2.1) in Example 3.4.2.1, with $\rho(x, 0) = \max\{1 - |x|, 0\}$.

For the symmetric initial condition, our numerical tests indicate the decay rate

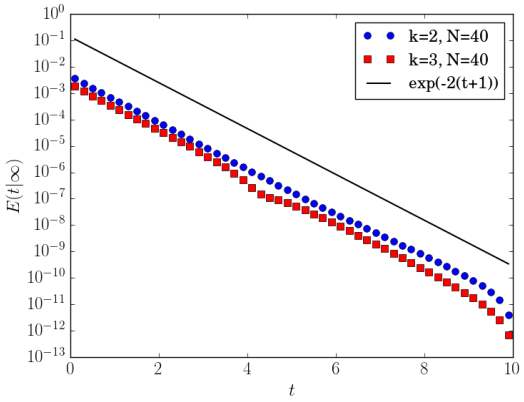


Figure 3.3: The relative entropy for porous equation in Example 3.4.2.1, with $\rho(x,0) = \max\{1 - |x - \frac{1}{2}|, 0\}$.

is around $e^{-6.4t}$. Indeed, symmetric initial data converge faster to equilibrium than the sharp rate since they preserve the invariance of the center mass, see [16] for more details. We then test the problem under the same settings, except for the initial condition shifted to the right $\rho(x,0) = \max\{1 - |x - \frac{1}{2}|, 0\}$ and the final time set to $t = 10$. The corresponding plot of $E(t|\infty)$ is given in Figure 3.3 with the exponential decay rate -2, which coincides with the result in [21]. Similar numerical test can be found in [8].

Example 3.4.2.2 (Fokker-Planck equation)**.** In this numerical test, we consider the Fokker-Planck equation for modeling the relaxation of fermion and boson gases. The equation takes the form

$$\partial_t \rho = \partial_x \left(x\rho \left(1 + \kappa\rho\right) + \partial_x \rho \right).$$

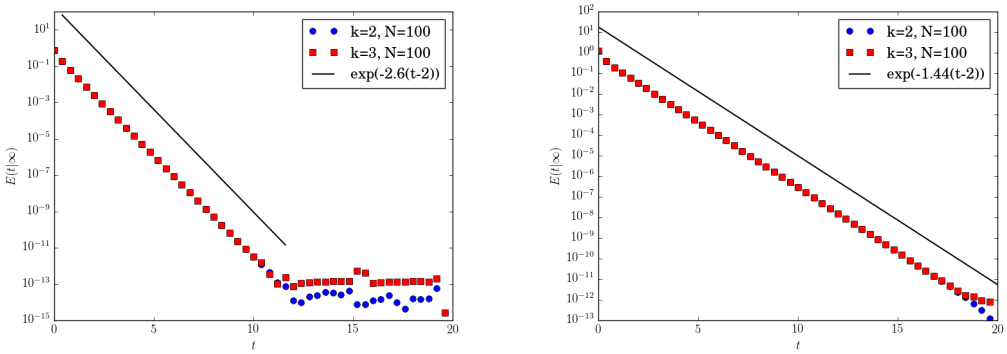
Here $\kappa = 1$ corresponds to boson gases and $\kappa = -1$ relates to fermion gases. The long time asymptotics of the one-dimensional model has been studied in [20] for $0 \le \rho \le 1$. The authors point out the equation evolves to a steady state $\rho_\infty(x) = \frac{1}{\beta e^{\frac{x^2}{2}} - \kappa}$. β is chosen such that $\int_{-\infty}^{\infty} \rho_\infty(x)dx = \int_{-\infty}^{\infty} \rho(x,0)dx$. The stationary solution minimizes

the entropy functional

$$E = \int \left(\frac{|x|^2}{2}\rho + \rho \log(\rho) - \kappa(1+\kappa\rho)\log(1+\kappa\rho) \right) dx.$$

The relative entropy decays at an exponential rate $E(t|\infty) \leq E(0|\infty)e^{-2Ct}$, with $C = 1$ for the boson case and $0 < C < 1$ for the fermion case. In our numerical test, we study the same entropy functional and set $f(\rho) = \rho(1+\kappa\rho)$, $H'(\rho) = \log\left(\frac{\rho}{1+\kappa\rho}\right)$, $V = \frac{x^2}{2}$ in our numerical scheme. The limiter is turned on in the computation. The initial condition is chosen as $\rho(x,0) = \frac{1}{0.4\pi}e^{-\frac{(x-1)^2}{0.4}}$. We compute on the domain $[-10, 10]$ with 100 mesh cells, and march towards the steady state with $\tau = 0.0002h^2$. For both boson and fermin cases, $g(\rho) = 2\rho$ is used in the numerical flux.

For both of the test cases, we use numerical steady states as references to calculate the relative entropy. The result for the boson case is given in Figure 3.4a. As one can see, the decay rate is around -2.6. While for the fermion case, which is exhibited in Figure 3.4b, the relative entropy decays at a slower rate of -1.44.



(a) Bose gas. (b) Fermi gas.

Figure 3.4: Decay of the relative entropy for Fokker-Planck equation in Example 3.4.2.2.

Example 3.4.2.3 (Generalized Fokker-Planck equation for the boson gas)**.** Let us now consider the generalized Fokker-Planck equation with linear diffusion and su-

perlinear drift

$$\partial_t \rho = \partial_x \left(x\rho(1+\rho^N) + \partial_x \rho \right),$$

with N being a positive constant. For $N > 2$, it is reported in [1] that a critical mass phenomenon exists for one-dimensional problems. An initial distribution with supercritical mass will evolve a singularity at the origin, which has been confirmed numerically in [8] and [56]. In this test, we repeat the numerical experiment in [8] and [56], setting

$$f(\rho) = \rho\left(1+\rho^3\right), H'(\rho) = \log \frac{\rho}{\sqrt[3]{1+\rho^3}} \text{ and } V(x) = \frac{x^2}{2}.$$

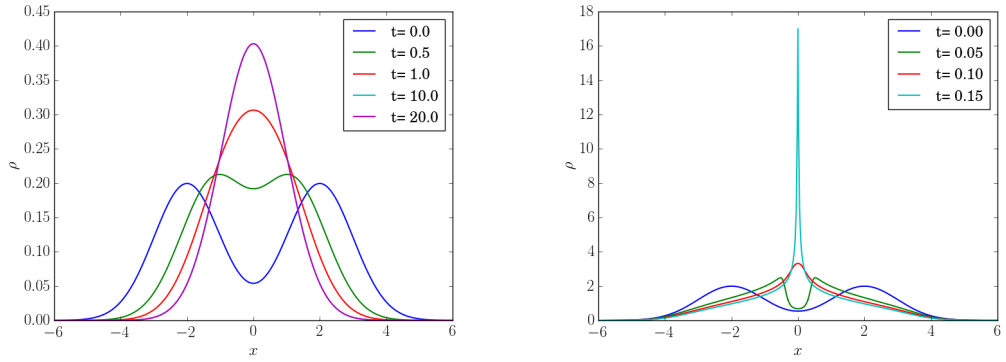
The initial datum is chosen as

$$\rho(x,0) = \frac{M}{2\sqrt{2\pi}} \left(e^{-\frac{(x-2)^2}{2}} + e^{-\frac{(x+2)^2}{2}} \right).$$

We test with both the subcritical case with $M = 1$ and the supercritical case with $M = 10$. The P^4 elements are used in our numerical scheme and we compute on the domain $[-6, 6]$ with $N = 120$. For $M = 1$, the time step is chosen as $\tau = 0.003h^2$ and for $M = 10$, it is $\tau = 0.0005h^2$. And we use $g(\rho) = f(\rho)$ when defining the numerical flux. According to the numerical results in Figure 3.5, our scheme does capture the asymptotics of the equation.

3.4.3 Aggregation models

Example 3.4.3.1 (Nonlinear diffusion with smooth attraction kernel)**.** This numerical test is to study the dynamics of the equation with competing nonlinear diffusion



(a) $M = 1$. (b) $M = 10$.

Figure 3.5: Evolution of ρ of subcritical mass $M = 1$ and supercritical mass $M = 10$ for generalized Fokker-Planck equation for the boson gas in Example 3.4.2.3.

and smooth nonlocal attraction,

$$\partial_t \rho = \partial_x \left(\rho \partial_x \left(\nu \rho^{m-1} + W * \rho \right) \right).$$

Here $0 \leq \rho \leq 1$. $\nu > 0$ and $m > 1$ are parameters to be specified. The convolution kernel W in this example is nonlocal and smooth. Under this setting, the attraction effect is weak and the solution would either end up with a steady state or spread out in the whole domain with bounded initial data. The compactly supported steady state is of special interest, due to its application for modeling the biological aggregation, such as flocks and swarms. Indeed, such stationary solution can be reached for $m > 2$ with arbitrary ν. While for $1 < m \leq 2$, the long time behavior of the solution can be sophisticated. We refer to [11] and [12] for details. For our numerical test, we focus on the specific setting,

$$\begin{cases} \partial_t \rho = \partial_x(\rho \partial_x(\nu \rho^{m-1} + W * \rho)), W(x) = -\dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in [-6, 6], \\ \rho(x, 0) = \dfrac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(x-\frac{5}{2})^2}{2}} + e^{-\frac{(x+\frac{5}{2})^2}{2}} \right). \end{cases} \tag{3.21}$$

We apply periodic boundary conditions and use a mesh with $N = 120$ for computation. The P^2 scheme is used for the numerical test, and the time step is chosen as $\tau = 0.05h^2$. In Figure 3.6, we depict the numerical solution at $t = 1800$ with $m = 1.5$, $\nu = 0.33$, $m = 2$, $\nu = 0.48$ and $m = 3$, $\nu = 1.48$, which are used as the reference steady states when evaluating the relative entropy.

According to the plot, one can see that a larger m corresponds to a steady state with a sharper transition along the boundary of the support. Indeed, one should expect the Hölder continuity with the exponent $\alpha = \min\{1, 1/(m-1)\}$.

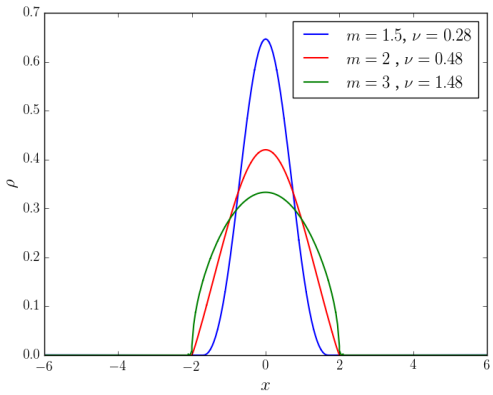
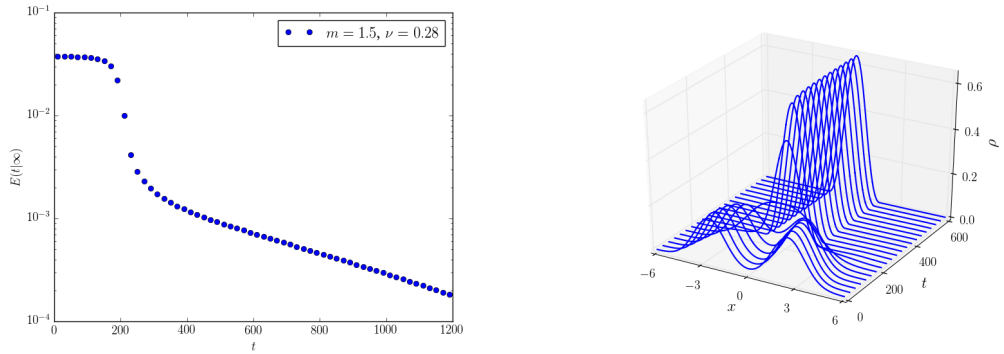


Figure 3.6: Solution profile of the equation (3.21) with nonlinear diffusion with smooth attraction kernel in Example 3.4.3.1 at $t = 1800$.



(a) Relative entropy. (b) Evolution.

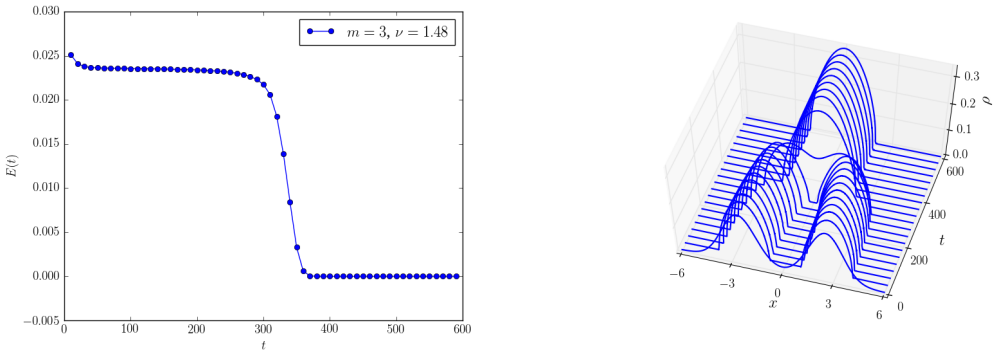
Figure 3.7: The relative entropy and the evolution of the solution with $m = 1.5$, $\nu = 0.28$. in Example 3.4.3.1.

We track the solution profile and the relative entropy. For $m = 1.5$, as one can see from Figure 3.7a, the dynamics of the problem are distinguished from the test cases for Fokker-Planck type equations. The relative entropy decays slowly at first, then it follows with a steep drop at a certain time. After that, the relative entropy decays exponentially. The behavior can be explained with Figure 3.7b. At the beginning, the two bumps of the initial condition stay away from each other, their interaction is weak hence the equation evolves at a slow rate. When they get closer, the attraction becomes strong. A sudden decay of the relative entropy occurs when the two bumps merge. After that, the contribution of the interaction potential to the total energy becomes small. The equation is again dominated by the diffusion term, and the relative entropy decays exponentially as we have seen before.

We omit the plots for $m = 2$. And for $m = 3$, the diffusion is relatively weak and the exponential decay after the steep drop is hard to observe. Hence we only plot the entropy in the normal scale. But still we can see the sharp drop when the bumps merge in Figure 3.8.

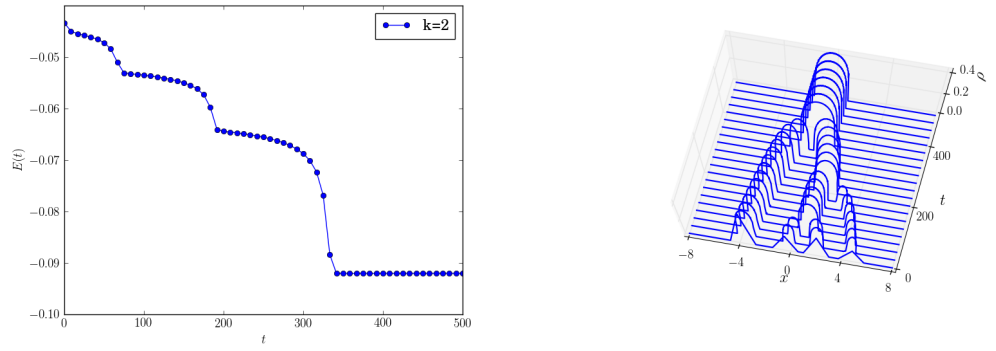
The initial stage featured with the weak long-range-interaction is referred as metastability. If multiple bumps exist, the relative entropy can decay in a staircase fashion. For example, we test the problem with $m = 6$, $\nu = 6$ and $\rho(x, 0) = \frac{3}{20}\max\{1 - |x - \frac{19}{4}|, 0\} + \frac{1}{4}\max\{1 - |x - 2|, 0\} + \frac{1}{5}\max\{1 - |x + \frac{17}{40}|, 0\} + \frac{2}{5}\max\{1 - |x + \frac{15}{4}|, 0\}$. The relative entropy and dynamics are given in Figure 3.9.

Example 3.4.3.2 (Nonlinear diffusion with compactly supported attraction kernel)**.** In the previous test, the attraction effect is global and the steady state will be connected for one-dimensional problems. But when W is local, the connectivity of the equilibrium can be affected by the initial mass distribution. Let us consider the



(a) Entropy. (b) Evolution.

Figure 3.8: The entropy and evolution of the solution with $m = 3$, $\nu = 1.48$ in Example 3.4.3.1.



(a) Entropy. (b) Evolution.

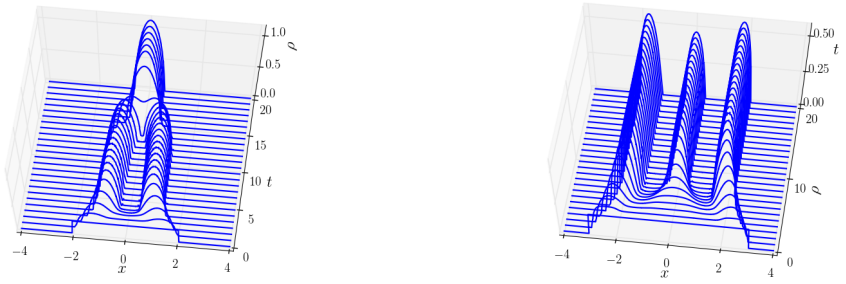
Figure 3.9: Stepwise decay of the entropy and the evolution of the solution with $m = 6$, $\nu = 6$ in Example 3.4.3.1.

following problem

$$\begin{cases} \partial_t \rho = \partial_x \left(\rho \partial_x \left(\frac{1}{4}\rho^2 + W * \rho \right) \right), \quad W = -\max\{1 - |x|, 0\}, \quad x \in [-4, 4], \\ \rho(x, 0) = \chi_{[-a,a]}(x), \end{cases} \tag{3.22}$$

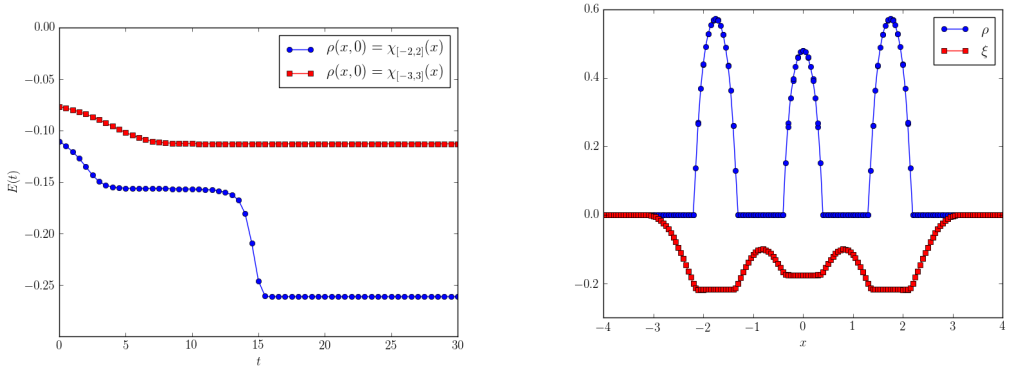
with periodic boundary conditions. We compute with $k = 2$ with the number of cells $N = 80$.

To convince the readers that the disconnected profile in Figure 3.10b is indeed the



(a) $\rho(x,0) = \chi_{[-2,2]}(x)$ (b) $\rho(x,0) = \chi_{[-3,3]}(x)$

Figure 3.10: Evolution of (3.22) with different initial conditions.



(a) Entropy

(b) ξ and ρ at $t = 30$ for $\rho(x,0) = \chi_{[-3,3]}(x)$

Figure 3.11: Entropy and the steady state of (3.22).

stationary solution, we plot ρ and ξ in Figure 3.11b. As one can observe, $\rho\partial_x\xi \approx 0$. Hence $\partial_t\rho \approx 0$ and ρ will be trapped in this steady state. Therefore, the observation in Figure 3.10 confirms our previous claim, that different initial density distributions may end up with steady states with distinct connectivity. Let us remark that such phenomenon has also been explored numerically in [15].

3.5 Two-dimensional numerical tests

Example 3.5.1 (Accuracy test)**.** We consider the initial value problem with a source term,

$$\begin{cases} \partial_t \rho = \nabla \cdot (\rho \nabla (\log(\rho) + \sin(x+y) + W * \rho)) + F, \\ \qquad (x,y) \in (-\pi,\pi) \times (-\pi,\pi), t > 0 \\ W(x,y) = \cos(x+y), \\ F(x,y) = 4\sin(x+y) + \cos(x+y+t) + \left(2 + 8\pi^2\right)\sin(x+y+t) \\ \qquad - 2\cos\left(2(x+y)+t\right) - 4\pi^2 \cos\left(2\left(x+y+t\right)\right) \\ \rho(x,y,0) = \sin(x+y) + 2. \end{cases} \tag{3.23}$$

Here periodic boundary conditions are applied and $W * \rho = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} W(x-\tilde{x}, y-\tilde{y})\rho(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}$. One can check that the exact solution to (3.23) is $\rho(x,y,t) = 2 + \sin(x+y+t)$. We use the time step $\tau = 0.0005(h^x)^2$ for the calculation. The error table is given in Table 3.7.

Table 3.7: Two-dimensional accuracy test, with smooth data and periodic boundary conditions.

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	10×10	5.117	-	1.029	-	4.008E-01	-
	20×20	1.152	2.15	2.319E-01	2.15	9.442E-02	2.09
	40×40	3.011E-01	1.94	6.216E-02	1.90	2.420E-01	1.96
2	10×10	1.453	-	3.070E-01	-	1.675E-01	-
	20×20	2.223E-01	2.71	4.315E-02	2.83	2.972E-02	2.49
	40×40	3.563E-02	2.64	7.203E-03	2.58	5.135E-03	2.53
3	10×10	3.778E-02	-	7.929E-03	-	4.396E-03	-
	20×20	2.292E-03	4.04	5.255E-04	3.92	2.941E-04	3.90
	40×40	1.373E-04	4.06	3.333E-05	3.98	2.054E-05	3.83
4	10×10	2.240E-03	-	5.113E-04	-	2.941E-04	-
	20×20	6.765E-05	5.05	1.439E-05	5.15	1.152E-05	4.67
	40×40	2.439E-06	4.80	5.242E-07	4.78	4.503E-07	4.68

Example 3.5.2 (Dumbbell model for polymers)**.** The dumbbell model is widely used to describe the rheological behavior of dilute polymer solutions. In this model, the polymer molecular is treated as a dumbbell made of two beads jointed by a spring. We will consider the simplest case, in which the flow is homogeneous and the scaling constant is set to 1. Then the configuration probability density is governed by the Fokker-Plank equation,

$$\partial_t \rho(\boldsymbol{x}, t) = \nabla \cdot (\rho \nabla (U - \frac{1}{2} \boldsymbol{x} K \boldsymbol{x})) + \Delta \rho. \tag{3.24}$$

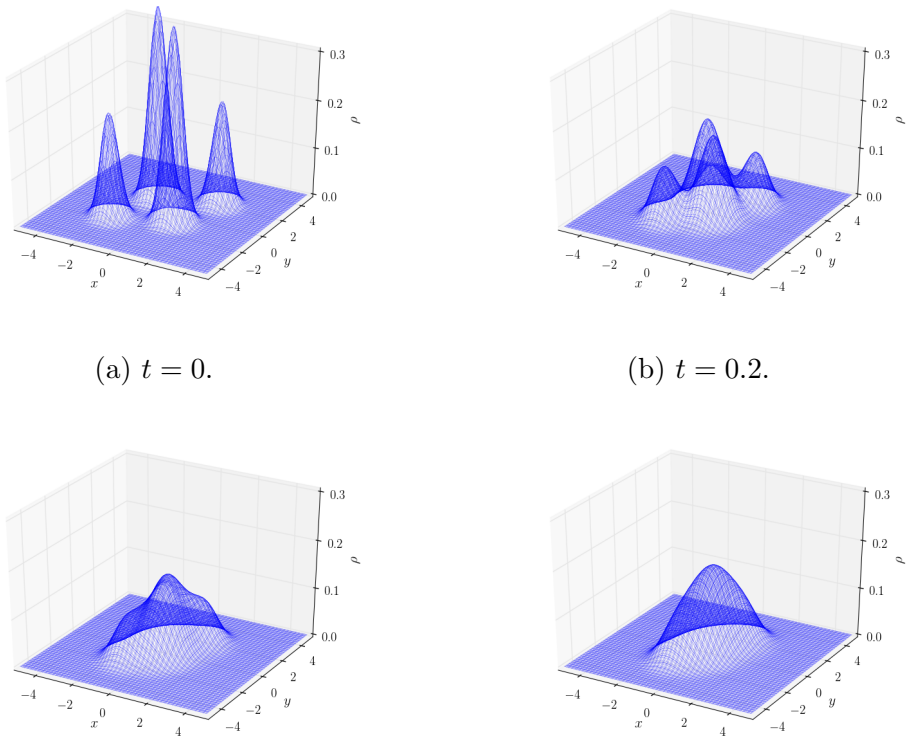
Here $\boldsymbol{x} = (x, y)$ corresponds to the direction vector of the molecule, while U is the spring potential and the 2×2 matrix K is the velocity gradient of the background flow. For the incompressible flow, $\mathrm{Tr}(K) = 0$. In our numerical test, we consider the finitely extensible nonlinear elastic (FENE) model. The potential U is given by

$$U(\boldsymbol{x}) = -\frac{r^2}{2} \log \left(1 - \frac{|\boldsymbol{x}|^2}{r^2}\right). \tag{3.25}$$

It is close to the Hookean potential when $|\boldsymbol{x}| \ll r$, while the distance between the two beads are restricted within r. Rigorously, one should consider the equation on the ball $\{\boldsymbol{x} : |\boldsymbol{x}| \leq r\}$, and the singularity near the boundary will cause challenges both analytically and numerically, see [34, 55, 66, 59] and the references therein. While in our numerical test, we only consider a simpler case, that the solution seems to be supported within the ball and it hardly reaches the boundaries. More specifically, we solve (3.24)-(3.25) with $r = 5$ and $K = \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & -0.3 \end{pmatrix}$. The initial condition is set as

$$\begin{aligned} \rho(x, y, 0) =& c \max\{24 - (x^2 + y^2), 0\} \\ & \left(e^{-\frac{(x-2)^2+(y-2)^2}{2\sigma^2}} + e^{-\frac{(x+2)^2+(y+2)^2}{2\sigma^2}} + e^{-\frac{(x-1)^2+(y-1)^2}{2\sigma^2}} + e^{-\frac{(x+1)^2+(y+1)^2}{2\sigma^2}} \right), \end{aligned}$$

where $\sigma^2 = 0.2$ and c the normalization constant. We use the P^3 DG scheme on a $[-5,5] \times [-5,5]$ domain with 50×50 mesh cells. The time step is chosen as $\tau = 2 \times 10^{-6}$. In Figure 3.12, we plot the evolution from $t = 0$ to $t = 0.6$. It seems the numerical solution merges to a single peak.



(a) $t = 0$. (b) $t = 0.2$.

(c) $t = 0.4$. (d) $t = 0.6$.

Figure 3.12: Evolution of the dumbbell model (3.24) in Example 3.5.2 with the FENE potential (3.25).

Example 3.5.3 (Patlak-Keller-Segel system for chemotaxis)**.** Chemotaxis is defined as a move of an organism along a chemical concentration gradient. Bacteria can produce this chemo-attractant themselves, creating thus a long range nonlocal interaction between them. The Patlak-Keller-Segel system is a mathematical model to

describe the motion of the organism. Its simplified version is given by

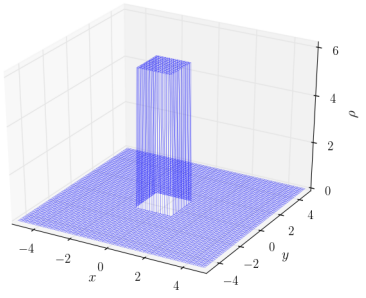
$$\begin{cases} \partial_t \rho = \Delta \rho - \nabla \cdot (\rho \nabla c), & (x,y) \in \mathbb{R}^2, t > 0, \\ -\Delta c = \rho, & (x,y) \in \mathbb{R}^2, t > 0, \\ \rho(x,y,0) = \rho_0(x,y). \end{cases}$$

The equation can be rewritten in a compact way

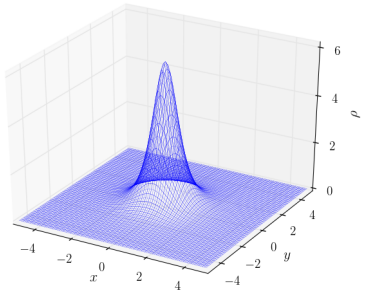
$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \left(\log(\rho) + W * \rho\right)\right), \quad W(x,y) = \frac{1}{2\pi} \log(\sqrt{x^2+y^2}) \quad (x,y) \in \mathbb{R}^2, t > 0. \tag{3.26}$$

Such system has been studied intensively in the past decades. It has been shown that the behavior of the equation (3.26) is determined by its initial mass (see [9], for example). If the initial value M is smaller than a critical value $M_c = 8\pi$, then the solution will exist globally. Otherwise, if M lies beyond M_c, the solution will blow up in a finite time, which is referred as chemotactic collapse.

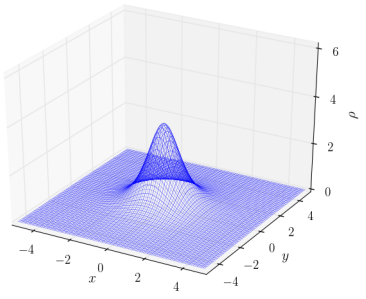
In our numerical test, we consider both the subcritical case $\rho_0(x) = 2(\pi - 0.2)1_{[-1,1]\times[-1,1]}(x,y)$ and the super-critical case $\rho_0(x) = 2(\pi + 0.2)1_{[-1,1]\times[-1,1]}(x,y)$. The computational domain is set as $[-5,5]\times[-5,5]$ and $[-\frac{3}{2},\frac{3}{2}]\times[-\frac{3}{2},\frac{3}{2}]$ respectively. We use the P^2 scheme for computation and $N^x = N^y = 50$. The time step is set as $\tau = 0.0005(h^x)^2$. The plots are given in Figure 3.13 and Figure 3.14. As one can see, the numerical solution dissipates for $\rho_0(x) = 2(\pi - 0.2)1_{[-1,1]\times[-1,1]}(x,y)$ and it evolves to a spike centered at the origin for $\rho_0(x) = 2(\pi + 0.2)1_{[-1,1]\times[-1,1]}(x,y)$.



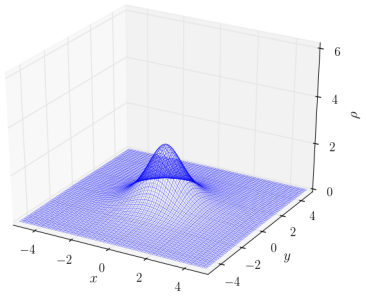
(a) $t = 0$.



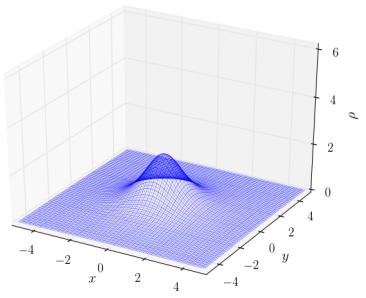
(b) $t = 4$.



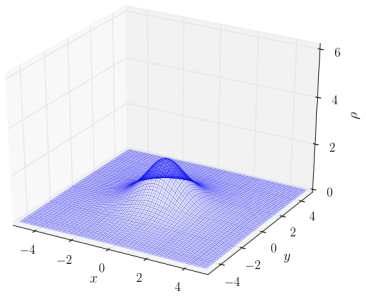
(c) $t = 8$.



(d) $t = 12$.

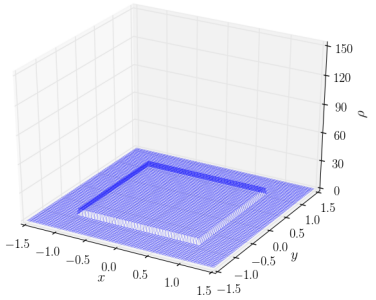


(e) $t = 16$.

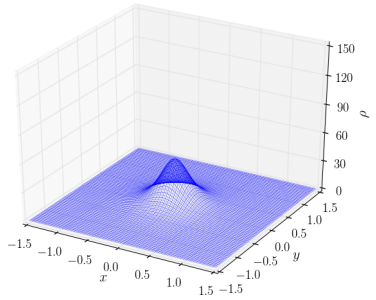


(f) $t = 20$.

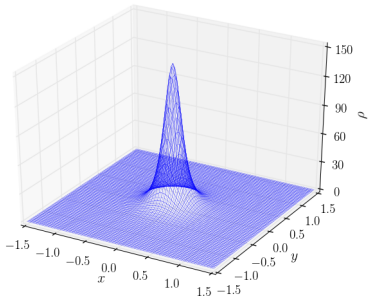
Figure 3.13: Evolution of Patlak-Keller-Segel equation (3.26) in Example 3.5.3 with subcritical mass.



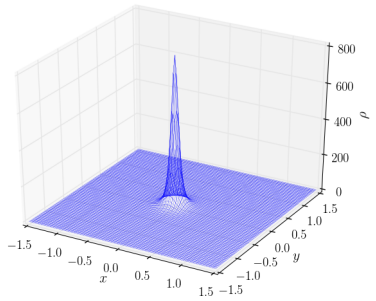
(a) $t = 0$.



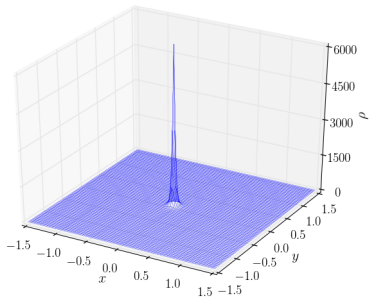
(b) $t = 0.5$.



(c) $t = 1$.



(d) $t = 1.5$.



(e) $t = 2$.

Figure 3.14: Evolution of Patlak-Keller-Segel equation (3.26) in Example 3.4.2.2 with supercritical mass.

3.6 Concluding remarks

In this chapter, we develop a high-order DG method for solving a class of parabolic equations and gradient flow problems with interaction potentials. Such equations are

governed by an entropy-entropy dissipation relationship and are featured with non-negative solutions. By applying the Gauss-Lobatto quadrature rule, our numerical scheme admits an entropy inequality for problems with smooth interaction kernels. Furthermore, with the SSP-RK time discretization and the positivity-preserving limiter, the fully discretized scheme preserves the non-negativity of the numerical density. It also conserves mass, and preserves numerical steady states for certain problems. We apply the method to two-dimensional problems on Cartesian meshes as well. Numerical examples are given to demonstrate the performance of the scheme.

Chapter Four

Cross-diffusion gradient flow systems

4.1 Introduction

Cross-diffusion systems are widely used to model multi-species interactions in various applications, such as population dynamics in biological systems [67], chemotactic cell migration [51] and spread of surfactant on the membrane [47]. In many situations, the systems are associated with neither symmetric nor positive-definite diffusion matrices, which not only complicate the mathematical analysis but also hinder the development of numerical methods. Recently, progress has been made in analyzing a class of cross-diffusion systems as a gradient flow associated with a dissipative entropy (or free energy) functional. See [49] and references therein. We exploit this entropy structure to develop a high-order stable DG method for these cross-diffusion systems.

Let Ω be a bounded domain in $\mathbb{R}^d$. We are interested in solving the following initial value problem of the cross-diffusion system.

$$\begin{cases} \partial_t \boldsymbol{\rho} = \nabla \cdot (F(\boldsymbol{\rho}) \nabla \boldsymbol{\xi}(\boldsymbol{\rho})) := \sum_{l=1}^{m} \partial_{x_l} \left(F(\boldsymbol{\rho}) \partial_{x_l} \boldsymbol{\xi}(\boldsymbol{\rho}) \right), & (\boldsymbol{x}, t) \in \Omega \times [0, \infty), \\ \boldsymbol{\rho}(\boldsymbol{x}, 0) = \boldsymbol{\rho}_0(\boldsymbol{x}). \end{cases} \tag{4.1}$$

Here $\boldsymbol{\rho} = \boldsymbol{\rho}(\boldsymbol{x}, t)$ and $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_m)^T$ is a vector-valued function. $\boldsymbol{\xi} = (\partial_{\rho_1} e, \cdots, \partial_{\rho_m} e)^T$ and $e = e(\boldsymbol{\rho})$ is a scalar-valued convex function that is twice differentiable. We also assume F to be an $m \times m$ positive-semidefinite matrix, in the sense that $\boldsymbol{z} \cdot F\boldsymbol{z} \geq 0$. Note F can be non-symmetric. In the divergence form $\partial_t \boldsymbol{\rho} = \nabla \cdot (A(\boldsymbol{\rho}) \nabla \boldsymbol{\rho})$, $A(\boldsymbol{\rho}) = F(\boldsymbol{\rho}) D\boldsymbol{\xi}(\boldsymbol{\rho})$ can be neither symmetric nor positive-definite.

(4.1) possesses a formal gradient flow structure governed by the entropy func-

tional

$$E = \int_\Omega e(\boldsymbol{\rho}) dx. \tag{4.2}$$

The system can be rewritten as $\partial_t \boldsymbol{\rho} = \nabla \cdot \left(F(\boldsymbol{\rho}) \nabla \frac{\delta E}{\delta \boldsymbol{\rho}} \right)$. One can see at least for classical solutions

$$\frac{d}{dt} E = \int_\Omega \partial_t \boldsymbol{\rho} \cdot \boldsymbol{\xi} dx = -\sum_{l=1}^{m} \int_\Omega \partial_{x_l} \boldsymbol{\xi} \cdot F \partial_{x_l} \boldsymbol{\xi} dx \leq 0. \tag{4.3}$$

(4.2) and (4.3) indicate certain well-posedness of the initial value problem (4.1). However, as that in the scalar case, in applications with $\boldsymbol{\rho}$ representing non-negative physics quantities, such as species densities, mass fractions and water heights, the well-posedness usually relies heavily on the positivity of the solution. For example, with $F(\boldsymbol{\rho}) = \mathrm{diag}(\boldsymbol{\rho})$, $F(\boldsymbol{\rho})$ is semi-positive definite only when $\boldsymbol{\rho}$ is non-negative. Violating the non-negativity may result in a non-decreasing entropy in (4.3). Furthermore, in problems that a logarithm entropy $E = \int_\Omega \sum_{l=1}^{m} \rho_l (\log \rho_l - 1) dx$ is considered, the entropy may not even be well-defined as $\boldsymbol{\rho}$ admits negative values.

In this chapter, we extend the DG method in Chapter 3 for scalar gradient flows to cross-diffusion systems. Our idea for is to formally rewrite the problem into decoupled equations and apply the previous numerical methods to each component of the unknown. For example, in one-dimensional case, (4.1) can be rewritten as

$$\partial_t \boldsymbol{\rho} = \partial_x (\mathrm{diag}(\boldsymbol{\rho}) \boldsymbol{v}), \qquad \boldsymbol{v} = \mathrm{diag}(\boldsymbol{\rho})^{-1} F(\boldsymbol{\rho}) \boldsymbol{u}, \qquad \boldsymbol{u} = \partial_x \boldsymbol{\xi}.$$

We apply the previous scheme to

$$\partial_t \rho_l = \partial_x (\rho_l v_l).$$

The numerical fluxes are properly chosen to ensure the entropy decay and the weak positivity. In this approach, the boundedness of $\mathrm{diag}(\boldsymbol{\rho})^{-1}F(\boldsymbol{\rho})$ is required for $\boldsymbol{v}$ to be well-defined. This assumption does hold in various applications. See examples in Section 4.4 and Section 4.5.

The rest of the chapter is organized as follows. In Section 4.2 and Section 4.3, we propose our DG schemes for one-dimensional and two-dimensional gradient flow systems respectively. Each section is composed of three parts: notations, semi-discrete scheme and entropy inequality, and fully discrete scheme as well as positivity-preserving techniques for producing non-negative solutions (with central and Lax-Friedrichs fluxes). In Section 4.4 and Section 4.5, we show several numerical examples to examine the performance of the numerical schemes. Finally, concluding remarks are given in Section 4.6. All proofs are documented in Appendix B.

4.2 Numerical method: one-dimensional case

4.2.1 Notations

In this section, we consider the one-dimensional problem

$$\begin{cases} \partial_t \boldsymbol{\rho} = \partial_x \left(F \partial_x \boldsymbol{\xi} \right), x \in \Omega \subset \mathbb{R}, t > 0. \\ \boldsymbol{\rho}(x, 0) = \boldsymbol{\rho}^0(x). \end{cases} \tag{4.4}$$

For simplicity, the compact support or periodic boundary condition is assumed. While it can also be extended to the zero-flux boundary condition.

Let $I_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$ and $I = \cup_{i=1}^{N} I_i$ be a regular partition of the domain Ω. The

length of the mesh cell is denoted by $h_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$. We assume $h = \max_{i=1,\ldots,N} h_i$ and $h_i \geq c_{mesh} h$ for some fixed constant c_{mesh}. The numerical solutions are defined in the tensor product space of discontinuous piecewise polynomials

$$\boldsymbol{V}_h = \prod_{l=1}^{m} V_h, \qquad V_h = \{v_h(x) : v_h|_{I_i} \in P^k(I_i)\}.$$

$P^k(I_i)$ is the space of polynomials of degree k on I_i. Functions in $\boldsymbol{V}_h$ (V_h) can be double-valued at cell interfaces. We use $\boldsymbol{v}_h^-$ ($v_{h,l}^-$) and $\boldsymbol{v}_h^+$ ($v_{h,l}^+$) to represent the left and right limit of $\boldsymbol{v}_h \in \boldsymbol{V}_h$ ($v_{h,l} \in V_h$) respectively. We also introduce notations $\{\boldsymbol{v}_h\} = \frac{1}{2}(\boldsymbol{v}_h^+ + \boldsymbol{v}_h^-)$ ($\{v_{h,l}\} = \frac{1}{2}(v_{h,l}^+ + v_{h,l}^-)$) for averages and $[\boldsymbol{v}] = \boldsymbol{v}_h^+ - \boldsymbol{v}_h^-$ ($[v_h] = v_h^+ - v_h^-$) for jumps.

Let $\{x_i^r\}_{r=1}^{k+1}$ be the $k+1$ Gauss-Lobatto quadrature points on I_i and $\{w_i^r\}_{r=1}^{k+1}$ are the corresponding weights associated with the normalized interval $[-1, 1]$. We introduce the following notations for the quadrature rule.

$$\widetilde{\int_{I_i}} \boldsymbol{\varphi} \cdot \boldsymbol{\psi} dx := \frac{h_i}{2} \sum_{r=1}^{k+1} w_r \boldsymbol{\varphi}(x_i^r) \cdot \boldsymbol{\psi}(x_i^r).$$

$$\widetilde{\int_{I_i}} \boldsymbol{\varphi} \cdot \partial_x \boldsymbol{\psi} dx := \frac{h_i}{2} \sum_{r=1}^{k+1} w_r \boldsymbol{\varphi}(x_i^r) \cdot \partial_x (\mathcal{I}\boldsymbol{\psi})(x_i^r).$$

Here $\mathcal{I}$ is a component-wise interpolation operator. Namely, $\mathcal{I}\boldsymbol{\psi} : \mathbb{R} \to \mathbb{R}^m$, and the l-th component of $\mathcal{I}\boldsymbol{\psi}$ is the k-th order interpolation polynomial of ψ_l at Gauss-Lobatto points. We also define $\widetilde{\int_\Omega} \cdot dx = \sum_{i=1}^N \widetilde{\int_{I_i}} \cdot dx$.

4.2.2 Semi-discrete scheme and entropy stability

We firstly rewrite (4.4) into a first-order system.

$$\partial_t \boldsymbol{\rho} = \partial_x (F\boldsymbol{u}),$$
$$\boldsymbol{u} = \partial_x \boldsymbol{\xi}.$$

On each mesh cell I_i, we multiply with a test function and integrate by parts to get

$$\int_{I_i} \partial_t \boldsymbol{\rho} \cdot \boldsymbol{\varphi} dx = -\int_{I_i} F\boldsymbol{u} \cdot \partial_x \boldsymbol{\varphi} dx + (F\boldsymbol{u} \cdot \boldsymbol{\varphi})^-_{i+\frac{1}{2}} - (F\boldsymbol{u} \cdot \boldsymbol{\varphi})^+_{i-\frac{1}{2}}, \tag{4.5a}$$
$$\int_{I_i} \boldsymbol{u} \cdot \boldsymbol{\psi} dx = -\int_{I_i} \boldsymbol{\xi} \cdot \partial_x \boldsymbol{\psi} dx + (\boldsymbol{\xi} \cdot \boldsymbol{\psi})^-_{i+\frac{1}{2}} - (\boldsymbol{\xi} \cdot \boldsymbol{\psi})^+_{i-\frac{1}{2}}. \tag{4.5b}$$

The numerical scheme is obtain by taking trial and test functions from the finite element space, replacing cell interface values with numerical fluxes and applying the quadrature rule. More precisely, we seek $\boldsymbol{\rho}_h, \boldsymbol{u}_h \in \boldsymbol{V}_h$ such that for all $\boldsymbol{\varphi}_h, \boldsymbol{\psi}_h \in \boldsymbol{V}_h$,

$$\int_{I_i}^{\sim} \partial_t \boldsymbol{\rho}_h \cdot \boldsymbol{\varphi}_h dx = -\int_{I_i}^{\sim} F_h \boldsymbol{u}_h \cdot \partial_x \boldsymbol{\varphi}_h dx + (\widehat{F\boldsymbol{u}} \cdot \boldsymbol{\varphi}_h^-)_{i+\frac{1}{2}} - (\widehat{F\boldsymbol{u}} \cdot \boldsymbol{\varphi}_h^+)_{i-\frac{1}{2}}, \tag{4.6a}$$
$$\int_{I_i}^{\sim} \boldsymbol{u}_h \cdot \boldsymbol{\psi}_h dx = -\int_{I_i}^{\sim} \boldsymbol{\xi}_h \cdot \partial_x \boldsymbol{\psi}_h dx + (\widehat{\boldsymbol{\xi}} \cdot \boldsymbol{\psi}_h^-)_{i+\frac{1}{2}} - (\widehat{\boldsymbol{\xi}} \cdot \boldsymbol{\psi}_h^+)_{i-\frac{1}{2}}. \tag{4.6b}$$

Here $F_h = F(\boldsymbol{\rho}_h)$, $\boldsymbol{\xi}_h = \boldsymbol{\xi}(\boldsymbol{\rho}_h)$, $\boldsymbol{\rho}_h = (\rho_{h,1}, \cdots, \rho_{h,m})^T$ and $\boldsymbol{u}_h$, $\boldsymbol{\varphi}_h$, $\boldsymbol{\psi}_h$ are defined similarly. We have the following choices for $\widehat{F\boldsymbol{u}}$ and $\widehat{\boldsymbol{\xi}}$.

(1) Central and Lax-Friedrichs fluxes:

$$\widehat{\boldsymbol{\xi}} = \{\boldsymbol{\xi}_h\}, \qquad \widehat{F\boldsymbol{u}} = \{F_h \boldsymbol{u}_h\} + \frac{\alpha}{2}[\boldsymbol{\rho}_h], \tag{4.7}$$

where

$$\alpha = \max(|\boldsymbol{v}_h^+|_\infty, |\boldsymbol{v}_h^-|_\infty), \qquad \boldsymbol{v}_h = \mathrm{diag}(\boldsymbol{\rho}_h)^{-1} F_h \boldsymbol{u}_h,$$

and $|\cdot|_\infty$ is the l^∞ norm on $\mathbb{R}^m$.

(2) Alternating fluxes:

$$\widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}_h^{\mp}, \qquad \widehat{F\boldsymbol{u}} = (F_h \boldsymbol{u}_h)^{\pm}. \tag{4.8}$$

Remark 4.2.1. *Note that* $F_h \boldsymbol{u}_h = \mathrm{diag}(\boldsymbol{\rho}_h)\boldsymbol{v}_h$. *Hence* (4.6a) *is equivalent to*

$$\widetilde{\int_{I_i}} \partial_t \rho_{h,l} \varphi_{h,l} dx = -\widetilde{\int_{I_i}} \rho_{h,l} v_{h,l} \partial_x \varphi_{h,l} dx + (\widehat{\rho v}_l \varphi_{h,l}^-)_{i+\frac{1}{2}} - (\widehat{\rho v}_l \varphi_{h,l}^+)_{i-\frac{1}{2}}$$

with

$$\widehat{\rho v}_l = \frac{1}{2}(\rho_{h,l}^+ v_{h,l}^+ + \rho_{h,l}^- v_{h,l}^-) + \frac{\alpha}{2}(\rho_{h,l}^+ - \rho_{h,l}^-),$$

which is formally reduced to the scalar case.

We define the discrete entropy

$$E_h = \widetilde{\int_\Omega} e(\boldsymbol{\rho}_h) dx. \tag{4.9}$$

As is stated in Theorem 4.2.1, the numerical scheme has a decaying entropy as that for the continuum system. The proof is given in Appendix B.1.1.

Theorem 4.2.1 (Entropy inequality)**.** *Let* $\boldsymbol{\rho}_h$ *and* $\boldsymbol{u}_h$ *be obtained from* (4.6), (4.7) *and* (4.8). E_h *is the associated discrete entropy defined in* (4.9). *Then*

(1) for central and Lax-Friedrichs fluxes,

$$\frac{d}{dt}E_h = -\widetilde{\int_\Omega} \boldsymbol{u}_h \cdot F_h \boldsymbol{u}_h dx - \frac{1}{2}\sum_{i=1}^{N} \alpha_{i+\frac{1}{2}} [\boldsymbol{\xi}_h]_{i+\frac{1}{2}} \cdot [\boldsymbol{\rho}_h]_{i+\frac{1}{2}} \leq 0,$$

where $[\boldsymbol{v}_h] = \boldsymbol{v}_h^+ - \boldsymbol{v}_h^-$.

(2) for alternating fluxes,

$$\frac{d}{dt}E_h = -\widetilde{\int_\Omega} \boldsymbol{u}_h \cdot F_h \boldsymbol{u}_h dx.$$

Remark 4.2.2. *Both choices of numerical fluxes are entropy stable, while they have different advantages and disadvantages. For central and Lax-Friedrichs fluxes defined in (4.7), they preserve positive cell averages as time steps are small. Hence the positivity-preserving limiter can be applied for producing non-negative solutions. Details are given in the next section. However, they are limited to problems satisfying* $|\operatorname{diag}(\boldsymbol{\rho})^{-1} F(\boldsymbol{\rho})| \leq C$ *for some matrix norm* $|\cdot|$*, so that* $\boldsymbol{v}_h$ *can be well-defined. Furthermore, for odd-order polynomials, one would observe a reduced order of accuracy with this particular choice of* α*. See accuracy tests in Section 4.4 and Section 4.5. Alternating fluxes do not have such order reduction and restriction on* F*, while it fails to preserve non-negative cell averages after one Euler forward step.*

Remark 4.2.3. *Due to the possible degeneracy of the problem, in general it is not easy to extend the entropy decay property to fully discrete explicit schemes. While when* F *is uniformly positive-definite and* e *is strongly convex, a fully discrete entropy inequality can be derived. See Appendix B.1.3, where we consider the Euler forward time discretization with central and Lax-Friedrichs fluxes.*

4.2.3 Fully discrete scheme and preservation of positivity

4.2.3.1 Euler forward time stepping

When central and Lax-Friedrichs fluxes are used, one can adopt the methodology introduced by Zhang et.al in [87, 86, 69] to enforce positivity of the solution.

It can be shown, when the solution achieves non-negative values at the current time level, with fluxes defined in (4.7) and a sufficiently small time step, the Euler forward time discretization will produce a solution with non-negative cell averages at the next time level. This is referred to as the weak positivity. Then we can apply a scaling limiter, squashing the solution polynomials towards cell averages to avoid inadmissible negative values. It is also shown in [86] that the limiter preserves the high-order spatial accuracy.

Theorem 4.2.2. *1. Suppose* $\rho_{h,l}^n(x_i^r) \geq 0$ *for all* i *and* r*. Take*

$$\tau \leq \min_{i=1,\ldots,N} \left(\left(\frac{w_1 h_i}{\alpha + v_{h,l}^+} \right)_{i-\frac{1}{2}}, \left(\frac{w_{k+1} h_i}{\alpha - v_{h,l}^-} \right)_{i+\frac{1}{2}} \right),$$

Then the preliminary solution $\rho_{h,l}^{n+1,\mathrm{pre}}$ *obtained through the Euler forward time stepping*

$$\int_{I_i}^{\sim} \frac{\boldsymbol{\rho}_h^{n+1,\mathrm{pre}} - \boldsymbol{\rho}_h^n}{\tau} \cdot \boldsymbol{\varphi} dx = -\int_{I_i}^{\sim} F\boldsymbol{u} \cdot \partial_x \boldsymbol{\varphi} dx + (\widehat{Fu} \cdot \boldsymbol{\varphi}_h^-)_{i+\frac{1}{2}} - (\widehat{Fu} \cdot \boldsymbol{\varphi}_h^+)_{i-\frac{1}{2}}.$$

has non-negative cell averages.

2. Let

$$\theta_{l,i} = \min\left(\frac{(\bar{\rho}_{h,l})_i^{n+1,\text{pre}}}{(\bar{\rho}_{h,l})_i^{n+1,\text{pre}} - \rho_{l,i}^{\min}}, 1\right), \qquad \rho_{l,i}^{\min} = \min_r\left(\rho_{h,l}^{n+1,\text{pre}}(x_i^r)\right).$$

Define $\rho_{h,l}^{n+1}$ *so that*

$$\rho_{h,l}^{n+1}(x_i^r) = (\bar{\rho}_{h,l})_i^{n+1,\text{pre}} + \theta_{l,i}(\rho_{h,l}^{n+1,\text{pre}}(x_i^r) - (\bar{\rho}_{h,l})_i^{n+1,\text{pre}}).$$

Then $\rho_{h,l}^{n+1}(x_i^r)$ *is non-negative. Furthermore, the interpolation polynomial* $\rho_{h,l}^{n+1}(x)$ *maintains spatial accuracy in the sense that*

$$|(\rho_{h,l})_i^{n+1}(x) - (\rho_{h,l})_i^{n+1,\text{pre}}(x)| \le C_k \max_{x\in I_i} |\rho_l(x, t_{n+1}) - (\rho_{h,l})_i^{n+1,\text{pre}}(x)|,$$

where C_k *is a constant depending only on the polynomial degree* k*.*

The proof of the first part of Theorem 4.2.2 is given in Appendix B.2.1. The non-negativity of the solution is ensured through the definition of $\theta_{l,i}$. For the accuracy result we refer to Theorem 4 in [86].

Remark 4.2.4 (Pointwise non-negativity)**.** *The procedure stated in Theorem 4.2.2 only guarantees the non-negativity of the solution at all Gauss-Lobatto points. This would be enough for most scenarios, since the scheme is defined only through these points. One can also ensure pointwise non-negativity of solution polynomials on the whole interval by taking* $\rho_{i,l}^{min} = \inf_{x\in I_i} \rho_{h,l}^{n+1,\text{pre}}(x)$*.*

Remark 4.2.5 (Time steps)**.** *As has been analyzed for the scalar case, we expect the time step to be* $\tau \le \mu h^2$ *for some constant* μ*. In practice, we assume a diffusion number* μ*. If a negative cell average emerges, we halve the time step and redo the computation. Theorem 4.2.2 guarantees that the halving procedure will end after finite loops.*

Remark 4.2.6 (The scaling parameter)**.** *For robustness of the algorithm, especially for dealing with logarithm type entropy functionals, we introduce a small parameter* $\varepsilon_{l,i} = \min\left(10^{-13}, (\bar{\rho}_{h,l})_i^{n+1,\mathrm{pre}}\right)$ *and use* $\theta_{l,i}^{\varepsilon} = \min\left(\frac{(\bar{\rho}_{h,l})_i^{n+1,\mathrm{pre}} - \varepsilon_{l,i}}{(\bar{\rho}_{h,l})_i^{n+1,\mathrm{pre}} - \rho_{l,i}^{\min}}, 1\right)$ *instead of* θ *in our computation. In other words, if* $(\bar{\rho}_{h,l})_i^{n+1,\mathrm{pre}} > 10^{-13}$*, we scale the polynomial to require it takes values not smaller than* 10^{-13} *at Gauss-Lobatto points; otherwise, we set the solution to be the constant* $(\bar{\rho}_{h,l})_i^{n+1,\mathrm{pre}}$ *in the interval. Note as long as* $\bar{\rho}_{h,l}(x) \geq 10^{-13}$*, the accuracy could still be maintained.*

Remark 4.2.7 (Other bounds)**.** *In general, it would be difficult to preserve other bounds besides positivity though this procedure. Similar difficulty has also been encountered in [69].*

4.2.3.2 High-order time discretization

We adopt SSP-RK methods for high-order time discretizations. Since the time step will be chosen as $\tau = \mu h^2$. The first-order Euler forward time stepping would be sufficiently accurate for $k = 1$ to achieve the second-order accuracy. For $k = 2$ and $k = 3$, we apply the second-order SSP-RK method

$$\begin{aligned}
\boldsymbol{\rho}_h^{(1)} &= \boldsymbol{\rho}_h^n + \tau \boldsymbol{F}(\boldsymbol{\rho}_h^n), \\
\boldsymbol{\rho}_h^{n+1} &= \frac{1}{2}\boldsymbol{\rho}_h^n + \frac{1}{2}\left(\boldsymbol{\rho}_h^{(1)} + \tau \boldsymbol{F}(\boldsymbol{\rho}_h^{(1)})\right).
\end{aligned}$$

For $k = 4$ and $k = 5$, a third-order time discretization method should be used

$$\begin{aligned}
\boldsymbol{\rho}_h^{(1)} &= \boldsymbol{\rho}_h^n + \tau \boldsymbol{F}(\boldsymbol{\rho}_h^n), \\
\boldsymbol{\rho}_h^{(2)} &= \frac{3}{4}\boldsymbol{\rho}_h^n + \frac{1}{4}\left(\boldsymbol{\rho}_h^{(1)} + \tau \boldsymbol{F}(\boldsymbol{\rho}_h^{(1)})\right), \\
\boldsymbol{\rho}_h^{n+1} &= \frac{1}{3}\boldsymbol{\rho}_h^n + \frac{2}{3}\left(\boldsymbol{\rho}_h^{(2)} + \tau \boldsymbol{F}(\boldsymbol{\rho}_h^{(2)})\right).
\end{aligned}$$

As one can see, SSP-RK methods can be rewritten as convex combinations of Euler forward steps. Hence if the time step is chosen to be sufficiently small, and the scaling limiter is applied at each Euler forward stage, then the solution will remain non-negative after one full time step.

4.3 Numerical method: two-dimensional case

4.3.1 Notations

In this section, we focus on two-dimensional systems.

$$\partial_t \boldsymbol{\rho} = \partial_x(F\partial_x \boldsymbol{\xi}) + \partial_y(F\partial_y \boldsymbol{\xi}),$$

The problem is set on a rectangular domain $\Omega = I \times J$ with the compact support or periodic boundary condition. We consider a regular Cartesian mesh on Ω, with $I = \cup_{i=1}^{N^x} I_i$, $I_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, $J = \cup_{j=1}^{N^y} J_j$ and $J_j = (y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}})$. Let $h_i^x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ and $h_j^y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$. Then the solution is sought in the following finite element space.

$$\boldsymbol{V}_h = \prod_{l=1}^{N} V_h, \qquad V_h = \{v_h : v_h|_{I_i \times J_j} \in Q^k(I_i \times J_j)\}.$$

$Q^k(I_i \times J_j)$ is the tensor product space of $P^k(I_i)$ and $P^k(J_j)$. For $\boldsymbol{v}_h(x, y) \in \boldsymbol{V}_h$,

$$\{\boldsymbol{v}_h\}_{i+\frac{1}{2}}(y) = \frac{1}{2}\left(\boldsymbol{v}_h(x^+_{i+\frac{1}{2}}, y) + \boldsymbol{v}_h(x^-_{i+\frac{1}{2}}, y)\right),$$

$$\{\boldsymbol{v}_h\}_{j+\frac{1}{2}}(x) = \frac{1}{2}\left(\boldsymbol{v}_h(x, y^+_{j+\frac{1}{2}}) + \boldsymbol{v}_h(x, y^-_{j+\frac{1}{2}})\right),$$

$$[\boldsymbol{v}_h]_{i+\frac{1}{2}}(y) = \boldsymbol{v}_h(x^+_{i+\frac{1}{2}}, y) - \boldsymbol{v}_h(x^-_{i+\frac{1}{2}}, y),$$

$$[\boldsymbol{v}_h]_{j+\frac{1}{2}}(x) = \boldsymbol{v}_h(x, y^+_{j+\frac{1}{2}}) - \boldsymbol{v}_h(x, y^-_{j+\frac{1}{2}}).$$

Same notation will be applied to the scalar case $v_{h,l} \in V_h$. Finally, for the quadrature rule, we denote by $\widetilde{\int}_J \cdot dy = \sum_{j=1}^{N^y} \widetilde{\int}_{J_j} \cdot dy$ and $\widetilde{\int}_I \cdot dx = \sum_{i=1}^{N^x} \widetilde{\int}_{I_i} \cdot dx$.

4.3.2 Semi-discrete scheme and entropy inequality

The semi-discrete scheme is given as follows. Find $\boldsymbol{\rho}_h$, $\boldsymbol{u}^x_h$ and $\boldsymbol{u}^y_h$ in $\boldsymbol{V}_h$, such that for all $\boldsymbol{\varphi}_h$, $\boldsymbol{\psi}^x_h$ and $\boldsymbol{\psi}^y_h$ in $\boldsymbol{V}_h$, we have

$$\begin{aligned}
\widetilde{\int_{J_j}}\widetilde{\int_{I_i}} \partial_t \boldsymbol{\rho}_h \cdot \boldsymbol{\varphi}_h dxdy &= -\widetilde{\int_{J_j}}\widetilde{\int_{I_i}} F_h \boldsymbol{u}^x_h \cdot \partial_x \boldsymbol{\varphi}_h + F_h \boldsymbol{u}^y_h \cdot \partial_y \boldsymbol{\varphi}_h dxdy \\
&\quad + \widetilde{\int_{J_j}} \widehat{F\boldsymbol{u}^x}_{i+\frac{1}{2}}(y) \cdot \boldsymbol{\varphi}_h(x^-_{i+\frac{1}{2}}, y) - \widehat{F\boldsymbol{u}^x}_{i-\frac{1}{2}}(y) \cdot \boldsymbol{\varphi}_h(x^+_{i-\frac{1}{2}}, y) dy \\
&\quad + \widetilde{\int_{I_i}} \widehat{F\boldsymbol{u}^y}_{j+\frac{1}{2}}(x) \cdot \boldsymbol{\varphi}_h(x, y^-_{j+\frac{1}{2}}) - \widehat{F\boldsymbol{u}^y}_{j-\frac{1}{2}}(x) \cdot \boldsymbol{\varphi}_h(x, y^+_{j-\frac{1}{2}}) dx, \\
\widetilde{\int_{J_j}}\widetilde{\int_{I_i}} \boldsymbol{u}^x_h \cdot \boldsymbol{\psi}^x_h + \boldsymbol{u}^y_h \cdot \boldsymbol{\psi}^y_h dxdy &= -\widetilde{\int_{J_j}}\widetilde{\int_{I_i}} \boldsymbol{\xi}_h \cdot \left(\partial_x \boldsymbol{\psi}^x_h + \partial_y \boldsymbol{\psi}^y_h\right) dxdy \\
&\quad + \widetilde{\int_{J_j}} \widehat{\boldsymbol{\xi}}_{i+\frac{1}{2}}(y) \cdot \boldsymbol{\psi}^x_h(x^-_{i+\frac{1}{2}}, y) - \widehat{\boldsymbol{\xi}}_{i-\frac{1}{2}}(y) \cdot \boldsymbol{\psi}^x_h(x^+_{i-\frac{1}{2}}, y) dy \\
&\quad + \widetilde{\int_{I_i}} \widehat{\boldsymbol{\xi}}_{j+\frac{1}{2}}(x) \cdot \boldsymbol{\psi}^y_h(x, y^-_{j+\frac{1}{2}}) - \widehat{\boldsymbol{\xi}}_{j-\frac{1}{2}}(x) \cdot \boldsymbol{\psi}^y_h(x, y^+_{j-\frac{1}{2}}) dx.
\end{aligned}$$

As that in the one-dimensional case, two choices of numerical fluxes can be used.

(1) Central and Lax-Friedrichs flux

$$\begin{aligned}
\widehat{\boldsymbol{\xi}}_{i+\frac{1}{2}}(y) &= \{\boldsymbol{\xi}_h\}_{i+\frac{1}{2}}(y), \qquad \widehat{\boldsymbol{\xi}}_{j+\frac{1}{2}}(x) = \{\boldsymbol{\xi}_h\}_{j+\frac{1}{2}}(x), \\
\widehat{F\boldsymbol{u}^x}_{i+\frac{1}{2}}(y) &= \left(\{F_h \boldsymbol{u}^x_h\}_{i+\frac{1}{2}} + \frac{\alpha^x_{i+\frac{1}{2}}}{2} [\boldsymbol{\rho}_h]_{i+\frac{1}{2}} \right)(y),
\end{aligned}$$

$$\alpha^x_{i+\frac{1}{2}}(y) = \max(|\boldsymbol{v}^x_h(x^+_{i+\frac{1}{2}}, y)|_\infty, |\boldsymbol{v}^x_h(x^-_{i+\frac{1}{2}}, y)|_\infty), \qquad \boldsymbol{v}^x_h = \mathrm{diag}(\boldsymbol{\rho}_h)^{-1} F_h \boldsymbol{u}^x_h,$$

$$\widehat{F\boldsymbol{u}^y}_{j+\frac{1}{2}}(x) = \left(\{F_h \boldsymbol{u}^y_h\}_{j+\frac{1}{2}} + \frac{\alpha^y_{j+\frac{1}{2}}}{2} [\boldsymbol{\rho}_h]_{j+\frac{1}{2}} \right)(x),$$

$$\alpha^y_{j+\frac{1}{2}}(x) = \max(|\boldsymbol{v}^y_h(x, y^+_{j+\frac{1}{2}})|_\infty, |\boldsymbol{v}^y_h(x, y^-_{j+\frac{1}{2}})|_\infty), \qquad \boldsymbol{v}^y_h = \mathrm{diag}(\boldsymbol{\rho}_h)^{-1} F_h \boldsymbol{u}^y_h.$$

(2) Alternating fluxes

$$\widehat{\boldsymbol{\xi}}_{i+\frac{1}{2}}(y) = \boldsymbol{\xi}_h(x^\mp_{i+\frac{1}{2}}, y), \qquad \widehat{F\boldsymbol{u}^x}_{i+\frac{1}{2}}(y) = (F_h \boldsymbol{u}^x_h)(x^\pm_{i+\frac{1}{2}}, y),$$

$$\widehat{\boldsymbol{\xi}}_{j+\frac{1}{2}}(x) = \boldsymbol{\xi}_h(x, y^\mp_{j+\frac{1}{2}}), \qquad \widehat{F\boldsymbol{u}^y}_{j+\frac{1}{2}}(x) = (F_h \boldsymbol{u}^y_h)(x, y^\pm_{j+\frac{1}{2}}).$$

The numerical scheme mimics a similar entropy decay behavior as the continuum equation. We define the discrete entropy

$$E_h = \int_J \int_I e(\boldsymbol{\rho}_h) dx dy. \tag{4.10}$$

We state the following entropy decay property for the semi-discrete scheme.

Theorem 4.3.1. *Let $\boldsymbol{\rho}_h$, $\boldsymbol{u}^x_h$ and $\boldsymbol{u}^y_h$ be obtained from the semi-discrete scheme for two-dimensional problems. E_h defined in (4.10) is the discrete entropy.*

(1) Suppose central and Lax-Friedrichs fluxes are used, then

$$\begin{aligned}
\frac{d}{dt} E_h = &- \int_J \int_I \left(\boldsymbol{u}^x_h \cdot F_h \boldsymbol{u}^x_h + \boldsymbol{u}^y_h \cdot F_h \boldsymbol{u}^y_h \right) dx dy \\
&- \frac{1}{2} \sum_{i=1}^{N^x} \int_J \alpha^x_{i+\frac{1}{2}} [\boldsymbol{\xi}_h]_{i+\frac{1}{2}}(y) \cdot [\boldsymbol{\rho}_h]_{i+\frac{1}{2}}(y) dy \\
&- \frac{1}{2} \sum_{j=1}^{N^y} \int_I \alpha^y_{j+\frac{1}{2}} [\boldsymbol{\xi}_h]_{j+\frac{1}{2}}(x) \cdot [\boldsymbol{\rho}_h]_{j+\frac{1}{2}}(x) dx \le 0.
\end{aligned}$$

(2) Suppose alternating fluxes are used, then

$$\frac{d}{dt}E_h = -\int_J^{\sim}\int_I^{\sim} (\boldsymbol{u}_h^x \cdot F_h \boldsymbol{u}_h^x + \boldsymbol{u}_h^y \cdot F_h \boldsymbol{u}_h^y)\, dxdy \leq 0.$$

4.3.3 Fully discrete scheme and preservation of positivity

One can adopt similar positivity-preserving techniques as in the one dimensional case, when central and Lax-Friedrichs fluxes are used. Again, the key step is to ensure the non-negativity of the first-order Euler forward time discretization. The positivity will be automatically preserved by the SSP-RK time discretization. For the first-order scenario, we could prove the following theorem.

Theorem 4.3.2. *1. Suppose $\rho_{h,l}(x_i^r, y_j^s) \geq 0$ for all i, j, r and s. Take*

$$\tau \leq \frac{1}{2}\min_{r,s,i,j}\left(\left(\frac{w_1 h_i^x}{\alpha^x + v_{h,l}^x}\right)(x_{i-\frac{1}{2}}^+, y_j^s), \left(\frac{w_{k+1} h_i^x}{\alpha^x - v_{h,l}^x}\right)(x_{i+\frac{1}{2}}^-, y_j^s), \right.$$
$$\left.\left(\frac{w_1 h_j^y}{\alpha^y + v_{h,l}^y}\right)(x_i^r, y_{j-\frac{1}{2}}^+), \left(\frac{w_{k+1} h_j^y}{\alpha^y - v_{h,l}^y}\right)(x_i^r, y_{j+\frac{1}{2}}^-)\right).$$

Then the preliminary solution $(\rho_{h,l})_{i,j}^{n+1,\mathrm{pre}}$ obtained through the Euler forward time discretization has non-negative cell averages.

2. Let

$$\theta_{l,i,j} = \min\left(\frac{(\bar{\rho}_{h,l})_{i,j}^{n+1,\mathrm{pre}}}{(\bar{\rho}_{h,l})_{i,j}^{n+1,\mathrm{pre}} - m_{l,i,j}}, 1\right), \qquad \rho_{l,i,j}^{\min} = \min_{r,s}\left(\rho_{h,l}^{n+1,\mathrm{pre}}(x_i^r, y_j^s)\right).$$

Define $\rho_{h,l}^{n+1}$ so that

$$\rho_{h,l}^{n+1}(x_i^r, y_j^s) = (\bar{\rho}_{h,l})_{i,j}^{n+1,\mathrm{pre}} + \theta_{l,i,j}(\rho_{h,l}^{n+1,\mathrm{pre}}(x_i^r, y_j^s) - (\bar{\rho}_{h,l})_{i,j}^{n+1,\mathrm{pre}}).$$

Then $\rho_{h,l}^{n+1}(x_i^r, y_j^s)$ *is non-negative. Furthermore,* $\rho_{h,l}^{n+1}$ *would maintain the spatial accuracy achieved by* $\rho_{h,l}^{n+1,\mathrm{pre}}$.

Remark 4.3.1. *As before, pointwise non-negative solutions can be obtained by taking* $\rho_{l,i,j}^{\min} = \inf_{(x,y)\in I_i\times J_j} \rho_{h,l}^{n+1,\mathrm{pre}}(x,y)$. *The time step* $\tau = \mu h^2$ *will be used for time marching. As negative cell averages emerge, we halve the time step. We will also use* $\theta_{l,i,j}^{\varepsilon} = \min\left(\frac{(\bar{\rho}_{h,l})_{i,j}^{n+1,\mathrm{pre}} - \varepsilon_{l,i,j}}{(\bar{\rho}_{h,l})_{i,j}^{n+1,\mathrm{pre}} - \rho_{l,i,j}^{\min}}, 1\right)$ *instead of* $\theta_{l,i,j}$ *when applying the scaling limiter. Here* $\varepsilon_{l,i,j} = \min\left(10^{-13}, (\bar{\rho}_{h,l})_{i,j}^{n+1,\mathrm{pre}}\right)$.

4.4 One-dimensional numerical tests

In this section, we apply the entropy stable DG method for solving one-dimensional problems. In Example 4.4.1 and Example 4.4.2, the accuracy of the scheme is examined. The error is measured with the discrete norm associated with the Gauss-Lobatto quadrature rule. We do not invoke the positivity-preserving limiter in either tests, to exclude order degeneracy due to the temporal order reduction. See [87, 69] for relevant numerical experiments. Then we consider systems from tumor encapsulation (Example 4.4.3) and surfactant spreading (Example 4.4.4) . In these tests, leading fronts are formed and the issue of positivity arises.

Example 4.4.1 (Heat equations)**.** Let us first consider the initial value problem with decoupled heat equations,

$$\partial_t \rho_l = \partial_{xx}\rho_l, \qquad l = 1, 2,$$

on $[-1, 1]$ with the periodic boundary condition. $\rho_1(x, 0) = \sin(\pi x)+2$ and $\rho_2(x, 0) = \cos(\pi x) + 2$ are taken as our initial condition. The system is associated with the entropy $E = \int_\Omega \rho_1(\log\rho_1 - 1) + \rho_2(\log\rho_2 - 1)dx$. Hence $\boldsymbol{\xi} = (\log\rho_1, \log\rho_2)^T$ and

$F = \mathrm{diag}(\boldsymbol{\rho})$. We compute up to $t = 0.002$, with time step set as $\tau = 0.001h^2$. Positivity-preserving limiter is not activated. We use the Lax-Friedrichs flux and alternating fluxes respectively for computation. As one can see from Table 4.1 and Table 4.2. Optimal order of convergence is achieved with the alternating fluxes. The order of accuracy seems to be degenerated when we use central and Lax-Friedrichs fluxes with odd-order polynomials. This may be related with the fact that the problem is smooth hence the jump term $\frac{\alpha}{2}(\boldsymbol{\rho}^+ - \boldsymbol{\rho}^-)$ is too small, hence the Lax-Friedrichs flux $\widehat{F\boldsymbol{u}}$ is close to the central flux when the mesh is not well refined, which would cause order reduction. In Table 4.3, we document the error with different Lax-Friedrichs constant, with α replaced by $\tilde{\alpha} = 0, 2\alpha, 10\alpha$. As one can see, the optimal order is retrieved as $\tilde{\alpha}$ becomes large. We also remark, choosing a larger Lax-Friedrichs constant does not affect the compatibility with positivity-preserving procedure, while it does make the time step more restrictive.

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	80	8.802E-04	-	4.958E-04	-	4.226E-04	-
	160	3.948E-04	1.16	2.224E-04	1.16	1.899E-04	1.15
	320	1.609E-04	1.30	9.077E-05	1.29	7.821E-05	1.28
	640	5.647E-05	1.51	3.201E-05	1.50	2.806E-05	1.48
2	80	7.760E-06	-	8.881E-06	-	1.456E-05	-
	160	9.600E-07	3.02	1.113E-06	3.00	1.825E-06	3.00
	320	1.194E-07	3.01	1.395E-07	3.00	2.285E-07	3.00
	640	1.489E-08	3.00	1.745E-08	3.00	2.859E-08	3.00
3	80	6.027E-07	-	4.185E-07	-	8.166E-07	-
	160	5.914E-08	3.35	4.133E-08	3.34	8.365E-08	3.29
	320	5.003E-09	3.56	3.551E-09	3.54	7.567E-09	3.47
	640	3.702E-10	3.76	2.685E-10	3.73	6.054E-10	3.64
4	80	6.210E-10	-	6.542E-10	-	2.194E-09	-
	160	1.887E-11	5.04	2.030E-11	5.01	6.809E-11	5.01
	320	5.823E-13	5.02	6.338E-13	5.00	2.128E-12	5.00
	640	1.808E-14	5.01	1.981E-14	5.00	6.653E-14	5.00

Table 4.1: Accuracy test of heat equations in Example 4.4.1, with central flux for $\widehat{\boldsymbol{\xi}}$ and Lax-Friedrichs flux for $\widehat{F\boldsymbol{u}}$.

Example 4.4.2 (SKT population model)**.** We use the population model of Shigesada, Kawashima and Teramoto [67] for the second accuracy test. All the cross-

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	80	4.027E-03	-	2.214E-03	-	1.591E-03	-
	160	1.006E-03	2.00	5.530E-04	2.00	4.004E-04	1.99
	320	2.515E-04	2.00	1.382E-04	2.00	1.003E-04	2.00
	640	6.288E-05	2.00	3.455E-05	2.00	2.508E-05	2.00
2	80	1.541E-05	-	1.418E-05	-	3.049E-05	-
	160	1.912E-06	3.01	1.769E-06	3.00	3.736E-06	3.03
	320	2.383E-07	3.00	2.211E-07	3.00	4.624E-07	3.01
	640	2.975E-08	3.00	2.763E-08	3.00	5.752E-08	3.01
3	80	1.219E-07	-	1.116E-07	-	4.006E-07	-
	160	7.609E-09	4.00	6.966E-09	4.00	2.513E-08	4.00
	320	4.754E-10	4.00	4.353E-10	4.00	1.572E-09	4.00
	640	2.971E-11	4.00	2.720E-11	4.00	9.828E-11	4.00
4	80	1.09E-09	-	1.08E-09	-	4.41E-09	-
	160	3.401E-11	5.00	3.365E-11	5.00	1.374E-10	5.00
	320	1.062E-12	5.00	1.052E-12	5.00	4.278E-12	5.01
	640	3.319E-14	5.00	3.286E-14	5.00	1.334E-13	5.00

Table 4.2: Accuracy test of heat equations in Example 4.4.1, with alternating fluxes $\widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}_h^-$ and $\widehat{F\boldsymbol{u}} = (F_h \boldsymbol{u}_h)^+$.

$\tilde{\alpha}$	N	L^1 error	order	L^2 error	order	L^∞ error	order
0	80	8.029E-07	-	5.549E-07	-	1.031E-06	-
	160	1.007E-07	3.00	6.949E-08	3.00	1.294E-07	2.99
	320	1.259E-08	3.00	8.689E-09	3.00	1.618E-08	3.00
	640	1.575E-09	3.00	1.086E-09	3.00	2.022E-09	3.00
2α	80	4.721E-07	-	3.300E-07	-	6.683E-07	-
	160	4.003E-08	3.56	2.841E-08	3.54	6.043E-08	3.47
	320	2.952E-09	3.76	2.148E-09	3.73	4.843E-09	3.64
	640	2.028E-10	3.86	1.502E-10	3.84	3.560E-10	3.77
10α	80	1.572E-07	-	1.150E-07	-	2.605E-07	-
	160	1.053E-08	3.90	7.906E-09	3.86	1.892E-08	3.78
	320	6.830E-10	3.95	5.221E-10	3.92	1.300E-09	3.86
	640	4.356E-11	3.97	3.366E-11	3.96	8.619E-11	3.92

Table 4.3: Accuracy test of heat equations in Example 4.4.1, with central flux for $\widehat{\boldsymbol{\xi}}$ and Lax-Friedrichs flux for $\widehat{F\boldsymbol{u}} = \frac{1}{2}(F_h \boldsymbol{u}_h^+ + F_h \boldsymbol{u}_h^-) + \frac{\tilde{\alpha}}{2}(\boldsymbol{\rho}_h^+ - \boldsymbol{\rho}_h^-)$. Here $\tilde{\alpha} = 0, 2\alpha, 10\alpha$.

diffusion and self-diffusion coefficents are taken as 1. The system is written as follows.

$$\begin{cases} \partial_t \rho_1 = \partial_x \left((2\rho_1 + \rho_2)\partial_x \rho_1 + \rho_1 \partial_x \rho_2 \right), & \text{(4.11a)} \\ \partial_t \rho_2 = \partial_x \left(\rho_2 \partial_x \rho_1 + (\rho_1 + 2\rho_2)\partial_x \rho_2 \right). & \text{(4.11b)} \end{cases}$$

(4.11) is governed by the entropy $E = \int_\Omega \rho_1(\log\rho_1 - 1) + \rho_2(\log\rho_2 - 1)dx$. Again $\boldsymbol{\xi} = (\log\rho_1, \log\rho_2)^T$. $F = \mathrm{diag}(\boldsymbol{\rho}) \begin{pmatrix} 2\rho_1 + \rho_2 & \rho_2 \\ \rho_1 & 2\rho_2 + \rho_1 \end{pmatrix}$ and $\boldsymbol{z} \cdot F\boldsymbol{z} = 2\rho_1^2 z_1^2 + 2\rho_2^2 z_2^2 + \rho_1\rho_2(z_1 + z_2)^2 \geq 0$. The computational domain is taken as $[-\pi, \pi]$. Let $\rho_1(x, 0) = e^{\frac{1}{2}\sin x}$ and $\rho_2(x, 0) = e^{\frac{1}{2}\cos(2x)}$. We assume periodic boundary condition and compute to $t = 0.2$ with $\tau = 0.0002h^2$. The numerical solution at the next mesh level is set as a reference to evaluate the error. The error with central and Lax-Friedrichs fluxes is documented in Table 4.4, and that with alternating fluxes is in Table 4.5. The exhibited order of accuracy is similar to that in Example 4.4.1. We again compute the error for $k = 3$ with different constants in the Lax-Friedrichs flux. As one can see, the order of accuracy gets close to 4 as $\tilde{\alpha}$ increases.

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	20	2.852E-02	-	1.009E-02	-	8.675E-03	-
	40	9.370E-03	1.61	3.461E-03	1.54	2.918E-03	1.57
	80	3.031E-03	1.63	1.149E-03	1.59	9.292E-04	1.65
	160	9.527E-04	1.67	3.609E-04	1.67	2.765E-04	1.75
2	20	1.093E-03	-	4.283E-04	-	4.440E-04	-
	40	1.022E-04	3.42	4.329E-05	3.31	4.291E-05	3.37
	80	1.164E-05	3.13	5.223E-06	3.05	5.183E-06	3.05
	160	1.414E-06	3.04	6.480E-07	3.01	6.428E-07	3.01
3	20	7.543E-05	-	2.840E-05	-	3.058E-05	-
	40	8.282E-06	3.19	3.208E-06	3.15	3.901E-06	2.97
	80	8.588E-07	3.27	3.501E-07	3.20	4.431E-07	3.14
	160	8.748E-08	3.30	3.642E-08	3.27	4.812E-08	3.20
4	20	2.170E-06	-	9.649E-07	-	1.752E-06	-
	40	3.209E-08	6.08	1.746E-08	5.79	3.606E-08	5.60
	80	8.787E-10	5.19	5.031E-10	5.12	1.066E-09	5.08
	160	2.620E-11	5.07	1.542E-11	5.03	3.288E-11	5.02

Table 4.4: Accuracy test of the SKT population model in Example 4.4.2, with central flux for $\widehat{\boldsymbol{\xi}}$ and Lax-Friedrichs flux for $\widehat{F\boldsymbol{u}}$.

k	N	L^1 error	order	L^2 error	order	L^∞ error	order
1	20	8.476E-02	-	3.128E-02	-	2.359E-02	-
	40	2.055E-02	2.04	7.346E-03	2.090	5.189E-03	2.18
	80	5.088E-03	2.01	1.803E-03	2.027	1.235E-03	2.07
	160	1.268E-03	2.00	4.486E-04	2.007	3.049E-04	2.02
2	20	1.313E-03	-	5.584E-04	-	6.613E-04	-
	40	1.490E-04	3.14	6.531E-05	3.096	7.530E-05	3.14
	80	1.815E-05	3.04	8.039E-06	3.022	9.240E-06	3.03
	160	2.250E-06	3.01	1.001E-06	3.005	1.146E-06	3.01
3	20	4.008E-05	-	1.908E-05	-	3.948E-05	-
	40	2.420E-06	4.05	1.160E-06	4.040	2.645E-06	3.90
	80	1.503E-07	4.01	7.200E-08	4.010	1.682E-07	3.98
	160	9.374E-09	4.00	4.493E-09	4.002	1.056E-08	3.99
4	20	1.603E-06	-	8.465E-07	-	1.774E-06	-
	40	4.754E-08	5.08	2.615E-08	5.017	6.252E-08	4.83
	80	1.469E-09	5.02	8.149E-10	5.004	2.028E-09	4.95
	160	4.577E-11	5.00	2.544E-11	5.001	6.375E-11	4.99

Table 4.5: Accuracy test of the SKT population model in Example 4.4.2, with alternating fluxes $\widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}_h^-$ and $\widehat{F\boldsymbol{u}} = (F_h \boldsymbol{u}_h)^+$.

$\tilde{\alpha}$	N	L^1 error	order	L^2 error	order	L^∞ error	order
0	20	9.502E-05	-	3.419E-05	-	3.141E-05	-
	40	1.196E-05	2.99	4.260E-06	3.00	4.207E-06	2.90
	80	1.501E-06	2.99	5.331E-07	3.00	5.222E-07	3.01
	160	1.878E-07	3.00	6.667E-08	3.00	6.516E-08	3.00
10α	20	3.941E-05	-	1.692E-05	-	2.287E-05	-
	40	3.833E-06	3.36	1.644E-06	3.36	2.245E-06	3.35
	80	3.351E-07	3.52	1.477E-07	3.48	2.136E-07	3.39
	160	2.720E-08	3.62	1.238E-08	3.58	1.865E-08	3.52
900α	20	2.873E-06	-	1.096E-06	-	1.981E-06	-
	40	1.558E-07	4.21	6.967E-08	3.98	1.213E-07	4.03
	80	1.075E-08	3.86	4.892E-09	3.83	8.305E-09	3.87
	160	7.043E-10	3.93	3.261E-10	3.91	5.546E-10	3.90

Table 4.6: Accuracy test of the SKT population model in Example 4.4.2 with $k = 3$, with central flux for $\widehat{\boldsymbol{\xi}}$ and the Lax-Friedrichs flux for $\widehat{F\boldsymbol{u}} = \frac{1}{2}((F_h \boldsymbol{u}_h)^+ + (F_h \boldsymbol{u}_h)^-) + \frac{\tilde{\alpha}}{2}(\boldsymbol{\rho}_h^+ - \boldsymbol{\rho}_h^-)$. Here $\tilde{\alpha} = 0, 10\alpha, 900\alpha$ respectively.

Example 4.4.3 (Tumor encapsulation)**.**

$$\begin{cases} \partial_t \rho_1 = \partial_x \left((2\rho_1(1-\rho_1) - \beta\gamma\rho_1\rho_2^2)\partial_x\rho_1 - 2\beta\rho_1\rho_2(1+\gamma\rho_1)\partial_x\rho_2 \right), & \text{(4.12a)} \\ \partial_t \rho_2 = \partial_x \left((-2\rho_1\rho_2 + \beta\gamma(1-\rho_2)\rho_2^2)\partial_x\rho_1 + 2\beta\rho_2(1-\rho_2)(1+\gamma\rho_1)\partial_x\rho_2 \right). & \text{(4.12b)} \end{cases}$$

System (4.12) comes from the tumor encapsulation model proposed by Jackson and Byrne in [46], which describes the formation of a dense, fibrous connective tissue surrounding the benign neoplastic mass. In the system, ρ_1 corresponds to the concentration of tumors, and ρ_2 is the concentration of the surrounding tissue. In [50], the authors pointed out the system can be formulated as a gradient flow with the entropy

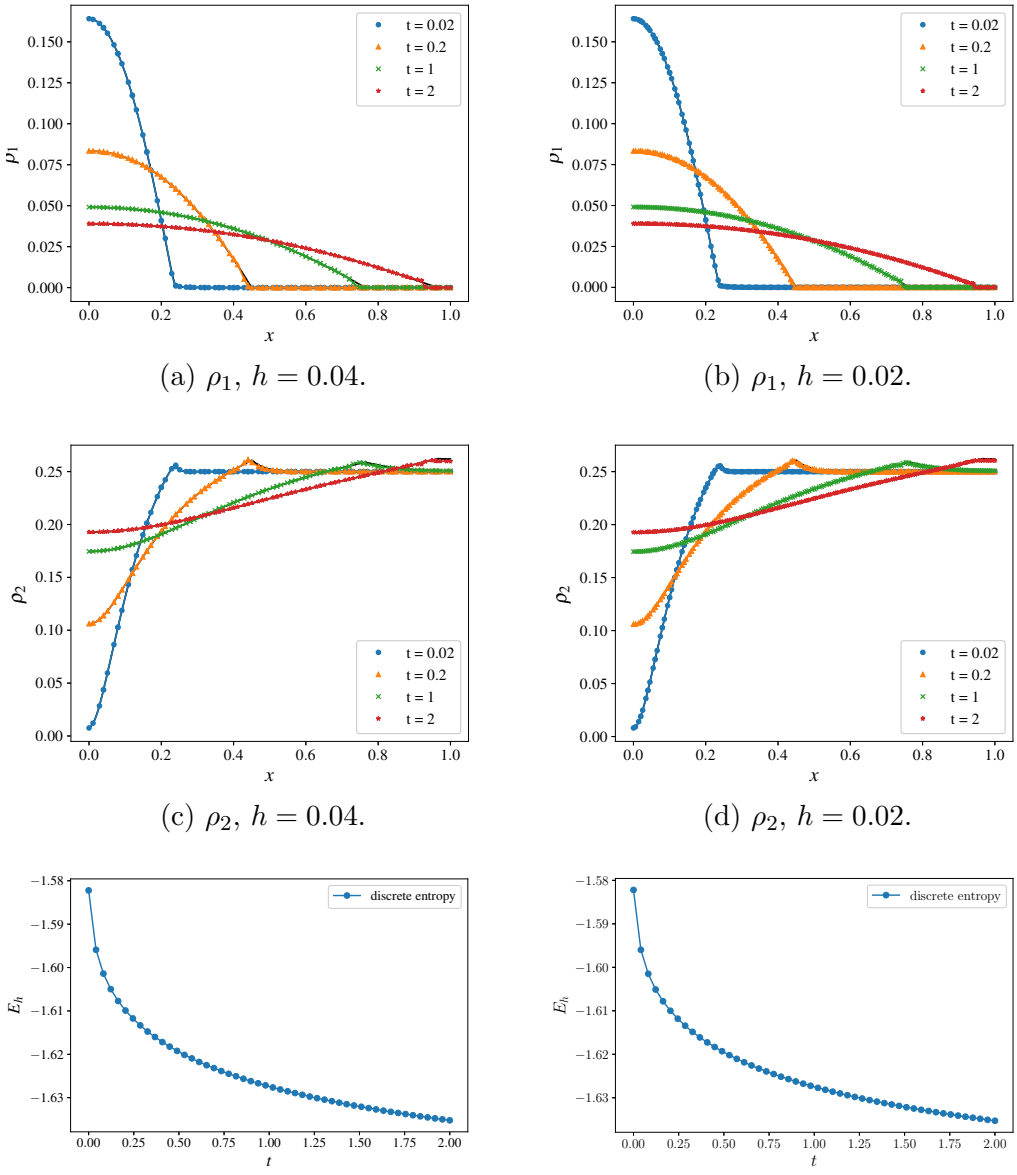
$$E = \int_\Omega \rho_1(\log \rho_1 - 1) + \rho_2(\log \rho_2 - 1) + (1-\rho_1-\rho_2)(\log(1-\rho_1-\rho_2) - 1)dx,$$

if $0 \le \gamma < 4/\sqrt{\beta}$. Then $\boldsymbol{\xi} = (\log \frac{\rho_1}{1-\rho_1-\rho_2}, \log \frac{\rho_2}{1-\rho_1-\rho_2})^T$ and the corresponding F is semi-positive definite with the prescribed parameters. In the numerical test, we firstly consider the case $\beta = 0.0075$ and $\gamma = 10$. The initial condition is set as

$$\rho_1(x,0) = \frac{1}{8}\left(1 + \tanh\frac{0.1-x}{0.05}\right), \qquad \rho_2(x,0) = \frac{1}{8}\left(1 - \tanh\frac{0.1-x}{0.05}\right).$$

The zero-flux boundary condition is applied in the simulation. The computational domain is set as $\Omega = [0,1]$ with $h = 0.02$ and $h = 0.04$ respectively. We choose the time step to be $\tau = 0.02h^2$. The numerical results are given in Figure 4.1. We then consider problems with a strong cell-induced pressure, with $\beta = 0.0075$ and $\gamma = 1000$. These parameters are rescaled from [46] and we drop the source term in their simulation. In the numerical test, $0.95\theta^\varepsilon_{l,i}$ instead of $\theta^\varepsilon_{l,i}$ is used in the scaling limiter for robustness and all other settings are the same. The numerical results are given in Figure 4.2. Although the system may no long possess a decaying entropy since $\gamma > 4/\sqrt{\beta}$, it seems that our numerical method still produces satisfying results and captures the sharp leading front of ρ_2.

Example 4.4.4 (Surfactant spreading)**.** We perform numerical simulations of a system modelling the surfactant spreading on a thin viscous film, which can be used for analyzing the delivery of aerosol for curing the respiratory distress syndrome. The



(a) ρ_1, $h = 0.04$.

(b) ρ_1, $h = 0.02$.

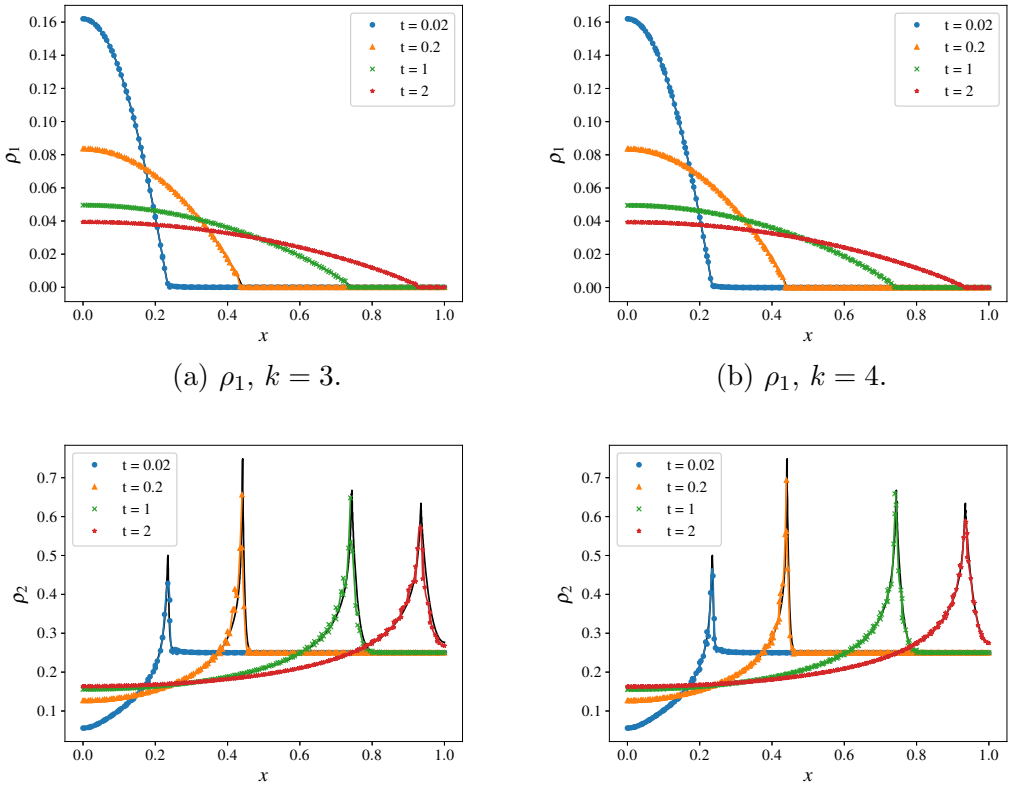
(c) ρ_2, $h = 0.04$.

(d) ρ_2, $h = 0.02$.

(e) E_h, $h = 0.04$.

(f) E_h, $h = 0.02$.

Figure 4.1: Numerical solutions to the tumor encapsulation problem in Example 4.4.3 with $\beta = 0.0075$ and $\gamma = 10$ at $t = 0.02$, $t = 0.2$, $t = 1$ and $t = 2$. We use piecewise cubic polynomials in the scheme. The mesh size is $h = 0.04$ in Figure 4.1a and Figure 4.1c, and is $h = 0.02$ in Figure 4.1b and Figure 4.1d. The corresponding entropy profiles are given in Figure 4.1e and Figure4.1f. The reference solutions are given in black lines obtained by using the P^1 scheme on a mesh with $h = 0.002$.



(a) ρ_1, $k = 3$.

(b) ρ_1, $k = 4$.

(c) ρ_2, $k = 3$.

(d) ρ_2, $k = 4$.

Figure 4.2: Numerical solutions to the tumor encapsulation problem in Example 4.4.3 with $\beta = 0.0075$ and $\gamma = 1000$ at $t = 0.02$, $t = 0.2$, $t = 1$ and $t = 2$. The mesh size is $h = 0.02$ and the time step is $\tau = 0.02h^2$. The scaling paramter is set as $0.95\theta^{\varepsilon}_{l,i}$ in the positivity-preserving procedure. Piecewise cubic polynomials are used in Figure 4.2a and Figure 4.2c, and piecewise quartic polynomials are used in Figure 4.2b and Figure 4.2d. The reference solution is given in black lines obtained by using the P^1 scheme on a mesh with $h = 0.001$.

system was derived by Jensen and Grotberg in [47] and was then analyzed by Escher et al. in [35]. In this system, ρ_1 represents the film thickness and ρ_2 is the concentration of the surfactant. The parameter g corresponds to a gravitational force.

$$\begin{cases} \partial_t \rho_1 = \partial_x \left(\frac{g}{3} \rho_1^3 \partial_x \rho_1 + \frac{1}{2} \rho_1^2 \partial_x \rho_2 \right), & \text{(4.13a)} \\ \partial_t \rho_2 = \partial_x \left(\frac{g}{2} \rho_1^2 \rho_2 \partial_x \rho_1 + \rho_1 \rho_2 \partial_x \rho_2 \right). & \text{(4.13b)} \end{cases}$$

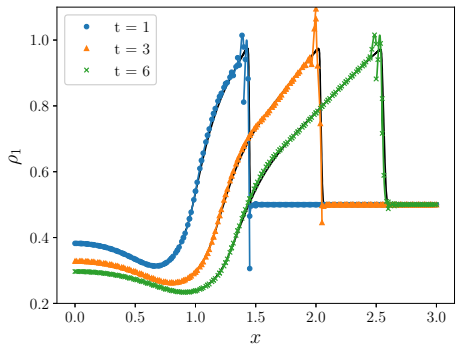
Following the derivation in [35], the system is associated with the following entropy functional

$$E = \int_\Omega \frac{g}{2} \rho_1^2 + \rho_2 (\log \rho_2 - 1) dx.$$

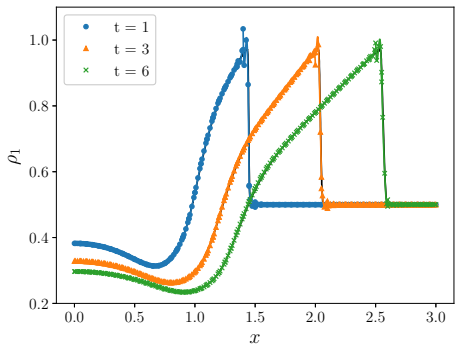
Accordingly, $\boldsymbol{\xi} = (g\rho_1, \log \rho_2)^T$, $F = \mathrm{diag}(\boldsymbol{\rho}) \begin{pmatrix} \frac{1}{3}\rho_1^2 & \frac{1}{2}\rho_1\rho_2 \\ \frac{1}{2}\rho_1^2 & \rho_1\rho_2 \end{pmatrix}$ and $\boldsymbol{z} \cdot F\boldsymbol{z} = \frac{1}{12}\rho_1^3 z_1^2 + \rho_1(\frac{1}{2}\rho_1 z_1 + \rho_2 z_2)^2 \geq 0$. It is shown in [47] that the similarity solutions to (4.13) would develop shocks as $g = 0$. In this numerical test, we consider a weak gravitational effect with $g = 0.02$. A uniform mesh with $h = 0.05$ is used on the spatial domain $[0, 3]$. The time step is set as $\tau = 0.02h^2$. We apply the zero-flux boundary condition and assume the initial condition to be $\rho_1 = 0.5$ and $\rho_2 = 0.5(1 - \tanh \frac{x-0.5}{0.1})$. Numerical solutions obtain with $k = 3$ and $k = 4$ are given in Figure 4.3. Under the same setting without applying positivity-preserving limiter, the solutions blow up shortly after $t = 0.1718$ for $k = 3$ and $t = 0.1691$ for $k = 4$.

4.5 Two-dimensional numerical tests

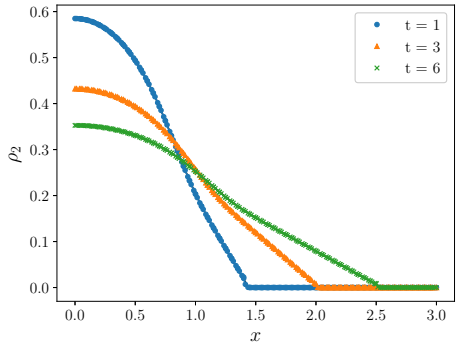
In this section, we provide several numerical examples of two-dimensional problems on Cartesian meshes. The accuracy test is given in Example 4.5.1. Then we apply the numerical scheme to solve the two-dimensional surfactant spreading problem in



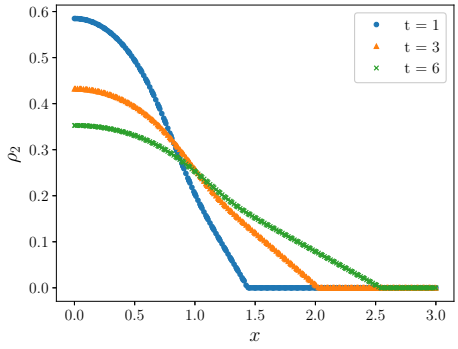
(a) ρ_1, $k = 3$.



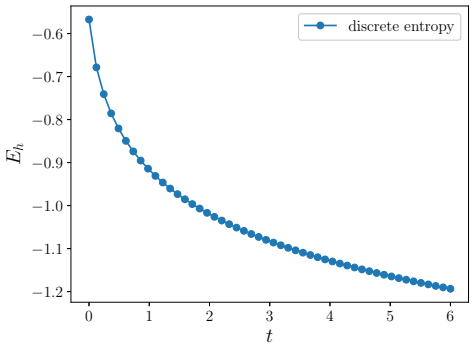
(b) ρ_1, $k = 4$.



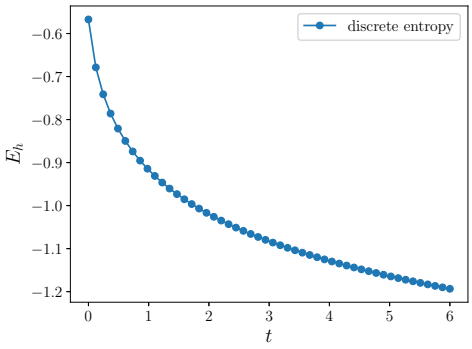
(c) ρ_2, $k = 3$.



(d) ρ_2, $k = 4$.



(e) E_h, $k = 3$.



(f) E_h, $k = 4$.

Figure 4.3: Numerical solutions to the surfactant spreading problem in Example 4.4.4 at $t = 1$, $t = 3$ and $t = 6$. The mesh size is $h = 0.05$ and the time step is $\tau = 0.02h^2$. We apply piecewise cubic polynomials for producing Figure 4.3a and Figure 4.3c. Piecewise quartic polynomials are used in Figure 4.3b and Figure 4.3d. Positivity-preserving limiter is activated mainly near the leading front of ρ_2. The reference solutions are given in black lines obtained with P^1 scheme on a mesh with $h = 0.0025$. The profiles of discrete entropy are depicted in Figure 4.3e and Figure 4.3f.

Example 4.5.2 and the seawater intrusion model in Example 4.5.3.

Example 4.5.1. We start with the accuracy test and consider the two-dimensional version of Example 4.4.2 with a source term $\boldsymbol{s} = (s_1, s_2)^T$.

$$\begin{cases} \partial_t \rho_1 = \nabla \cdot ((2\rho_1 + \rho_2)\nabla\rho_1 + \rho_1 \nabla \rho_2) + s_1, \\ \partial_t \rho_2 = \nabla \cdot (\rho_2 \nabla \rho_1 + (\rho_1 + 2\rho_2)\nabla \rho_2) + s_2. \end{cases}$$

The problem is again associated with the logarithm entropy.

$$E = \iint_\Omega \rho_1(\log \rho_1 - 1) + \rho_2(\log \rho_2 - 1) dxdy.$$

Then $\boldsymbol{\xi} = (\log \rho_1, \log \rho_2)^T$ and $F = \mathrm{diag}(\boldsymbol{\rho}) \begin{pmatrix} 2\rho_1 + \rho_2 & \rho_2 \\ \rho_1 & 2\rho_2 + \rho_1 \end{pmatrix}$. We assume the exact solution $\boldsymbol{\rho} = (\rho_1, \rho_2)^T = (0.5\sin(\pi(x+y+t)) + 1, 0.5\cos(\pi(x-y-0.5t)) + 1)^T$. The source terms $\boldsymbol{s} = (s_1, s_2)^T$ is computed accordingly. We compute to $t = 0.03$. The time step is set as $\tau = 0.0003h^2$ for $k = 1, 2, 3$ and $\tau = 0.0001h^2$ for $k = 4$.

k	N^x	L^1 error	order	L^2 error	order	L^∞ error	order
1	10	1.185E-01	-	5.349E-02	-	5.732E-02	-
	20	4.656E-02	1.35	2.147E-02	1.32	2.265E-02	1.34
	40	1.773E-02	1.39	8.266E-03	1.38	8.692E-03	1.38
	80	6.218E-03	1.51	2.921E-03	1.50	3.084E-03	1.50
2	10	8.206E-03	-	4.362E-03	-	8.402E-03	-
	20	9.124E-04	3.17	5.659E-04	2.95	9.942E-04	3.08
	40	1.077E-04	3.08	7.186E-05	2.98	1.213E-04	3.04
	80	1.315E-05	3.03	9.068E-06	2.99	1.515E-05	3.00
3	10	9.481E-04	-	5.256E-04	-	9.990E-04	-
	20	1.061E-04	3.16	5.847E-05	3.17	1.112E-04	3.17
	40	1.128E-05	3.23	6.169E-06	3.25	1.216E-05	3.19
	80	1.119E-06	3.33	6.042E-07	3.35	1.232E-06	3.30
4	10	4.685E-05	-	2.649E-05	-	7.459E-05	-
	20	1.212E-06	5.27	8.159E-07	5.02	2.837E-06	4.72
	40	3.328E-08	5.19	2.273E-08	5.17	7.754E-08	5.19

Table 4.7: Accuracy test of the cross-diffusion system in Example 4.5.1, with central flux for $\widehat{\boldsymbol{\xi}}$ and Lax-Friedrichs flux for $\widehat{F\boldsymbol{u}}$.

k	N^x	L^1 error	order	L^2 error	order	L^∞ error	order
1	10	3.848E-01	-	1.853E-01	-	2.415E-01	-
	20	9.608E-02	2.00	4.397E-02	2.08	4.372E-02	2.47
	40	2.376E-02	2.02	1.075E-02	2.03	1.014E-02	2.11
	80	5.916E-03	2.01	2.667E-03	2.01	2.525E-03	2.01
2	10	1.510E-02	-	8.132E-03	-	2.121E-02	-
	20	1.673E-03	3.17	9.563E-04	3.09	2.328E-03	3.19
	40	1.923E-04	3.12	1.169E-04	3.03	2.662E-04	3.13
	80	2.294E-05	3.07	1.453E-05	3.01	3.192E-05	3.06
3	10	8.253E-04	-	4.525E-04	-	1.449E-03	-
	20	4.998E-05	4.05	2.922E-05	3.95	1.044E-04	3.80
	40	3.075E-06	4.02	1.850E-06	3.98	6.675E-06	3.97
	80	1.912E-07	4.01	1.161E-07	4.00	4.216E-07	3.99
4	10	5.174E-05	-	3.287E-05	-	1.448E-04	-
	20	1.683E-06	4.94	1.127E-06	4.87	4.972E-06	4.86
	40	5.276E-08	5.00	3.620E-08	4.96	1.485E-07	5.07

Table 4.8: Accuracy test of the cross-diffusion system in Example 4.5.1 with alternating fluxes $\widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}_h^-$ and $\widehat{F\boldsymbol{u}} = (F_h \boldsymbol{u}_h)^+$.

$\tilde{\alpha}$	N^x	L^1 error	order	L^2 error	order	L^∞ error	order
0	10	1.082E-03	-	6.170E-04	-	1.139E-03	-
	20	1.328E-04	3.07	7.682E-05	3.01	1.373E-04	3.05
	40	1.651E-05	3.01	9.595E-06	3.00	1.722E-05	3.00
	80	2.063E-06	3.00	1.199E-06	3.00	2.157E-06	3.00
100α	10	2.505E-04	-	1.266E-04	-	1.528E-04	-
	20	1.234E-05	4.34	6.425E-06	4.30	1.234E-05	3.63
	40	6.709E-07	4.20	3.609E-07	4.15	9.094E-07	3.76
	80	3.923E-08	4.10	2.178E-08	4.05	6.480E-08	3.81

Table 4.9: Accuracy test for Example 4.5.1, with central flux for $\widehat{\boldsymbol{\xi}}$ and Lax-Friedrichs flux for $\widehat{F\boldsymbol{u}} = \frac{1}{2}((F_h \boldsymbol{u}_h)^+ + (F_h \boldsymbol{u}_h)^-) + \frac{\tilde{\alpha}}{2}(\boldsymbol{\rho}_h^+ - \boldsymbol{\rho}_h^-)$. Here $\tilde{\alpha} = 0, 100\alpha$.

Example 4.5.2. In this numerical example, we simulate the two-dimensional surfactant spreading.

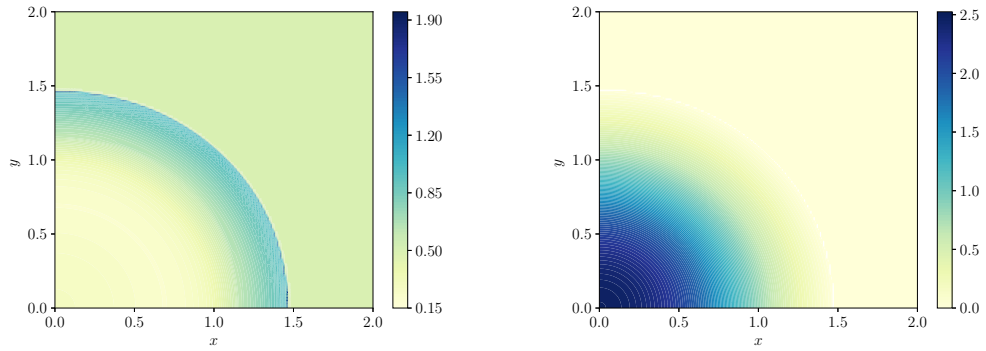
$$\begin{cases} \partial_t \rho_1 = \nabla \cdot \left(\dfrac{g}{3} \rho_1^3 \nabla \rho_1 + \dfrac{1}{2} \rho_1^2 \nabla \rho_2 \right), & \text{(4.14a)} \\ \partial_t \rho_2 = \nabla \cdot \left(\dfrac{g}{2} \rho_1^2 \rho_2 \nabla \rho_1 + \rho_1 \rho_2 \nabla \rho_2 \right). & \text{(4.14b)} \end{cases}$$

Again ρ_1 and ρ_2 correspond to the film thickness and surfactant concentration re-

spectively. The associated entropy is

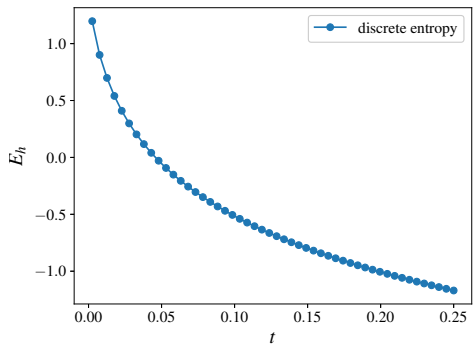
$$E = \iint_\Omega \frac{g}{2}\rho_1^2 + \rho_2(\log \rho_2 - 1)dxdy.$$

As before, $\boldsymbol{\xi} = (g\rho_1, \log \rho_2)^T$ and $F(\boldsymbol{\rho}) = \text{diag}(\boldsymbol{\rho}) \begin{pmatrix} \frac{1}{3}\rho_1^2 & \frac{1}{2}\rho_1\rho_2 \\ \frac{1}{2}\rho_1^2 & \rho_1\rho_2 \end{pmatrix}$. We assume zero-flux boundary condition on $\Omega = [0, 2] \times [0, 2]$ and compute to $t = 0.25$ with $h^x = h^y = 0.02$ and $\tau = 0.003(h^x)^2$. The gravitational coefficient is set as $g = 0.001$. Numerical results are given in Figure 4.4. The numerical method does capture the sharp transition of the bowl-shaped leading front of ρ_1, though with oscillations.



(a) ρ_1, $k = 3$.

(b) ρ_2, $k = 3$.



(c) E_h, $k = 3$.

Figure 4.4: Numerical solutions to the surfactant spreading problem (4.14) at $t = 0.25$. The solution is computed with piecewise cubic polynomials $k = 3$, mesh size $h = 0.02$ and time step $\tau = 0.003h^2$.

Example 4.5.3 (Seawater intrusion)**.** This test case is taken from [62] on solving a cross-diffusion system modelling the seawater intrusion in an unconfined aquifer.

$$\begin{cases} \partial_t \rho_1 = \nabla \cdot (\mu \rho_1 \nabla(\rho_1 + \rho_2 + b)), & (4.15a) \\ \partial_t \rho_2 = \nabla \cdot (\rho_2 \nabla(\mu \rho_1 + \rho_2 + b)). & (4.15b) \end{cases}$$

Here $z = b(x, y)$ gives the impermeable interface between the seawater and the bedrock. The saltwater sits on the bedrock between $z = b(x, y)$ and $z = b(x, y) + \rho_2(x, y, t)$. From $z = b(x, y) + \rho_2(x, y, t)$ to $z = b(x, y) + \rho_2(x, y, t) + \rho_1(x, y, t)$ is the freshwater, which is immiscible with the saltwater. The parameter $\mu \in (0, 1)$ is the mass density ratio between the freshwater and the saltwater. (4.15) is associated with the energy functional

$$E = \iint_\Omega \frac{\mu}{2}(\rho_1 + \rho_2 + b)^2 + \frac{1-\mu}{2}(\rho_2 + b)^2 dxdy.$$

It can be show that there exists a non-negative solution to (4.15) satisfies the energy decay property

$$\frac{d}{dt}E = -\iint_\Omega \mu^2 \rho_1 |\nabla(\rho_1 + \rho_2 + b)|^2 + \rho_2 |\nabla(\mu\rho_1 + \rho_2 + b)|^2 dxdy.$$

We assume the zero-flux boundary condition and use a 20×20 square mesh on $[0, 1] \times [0, 1]$. We compute to $t = 12$ with $\tau = 0.002(h^x)^2$. The parameter is set as $\mu = 0.9$. The initial condition $\boldsymbol{\rho}_0 = (\rho_1^0, \rho_2^0)^T$, with

$$\rho_1^0(x, y) = \begin{cases} 0.5 & \text{if } x \le 0.25 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\rho_2^0(x,y) = \begin{cases} b(0.5,0) - b(x,y) - (x-0.5) & \text{if } x \le 0.5 \\ 0 & \text{otherwise} \end{cases},$$

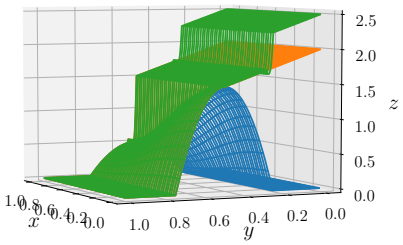
is used in our numerical simulation. Here the seabed function is

$$b(x,y) = \max(0, 0.5(1-16(x-0.5)^2)(\cos(\pi y)+2)).$$

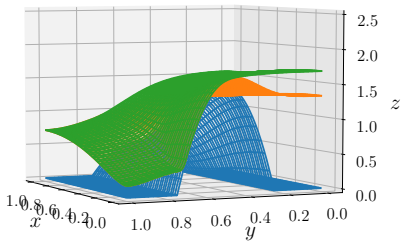
Numerical solutions at different time are depicted in Figure 4.5. One can see that solutions converge to a steady state as t becomes large.

4.6 Concluding remarks

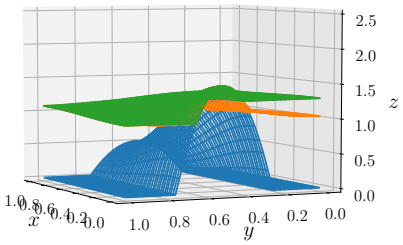
In this chapter, we develop a DG method for solving cross-diffusion systems with a gradient flow structure. The difficulty is that the non-negativity of solutions is usually an essential part for the entropy stability. One needs to design numerical schemes to preserve the entropy structure, as well as ensures the positivity of solutions. We adopt the Gauss-Lobatto quadrature rule in the DG method, so that the resulting semi-discrete schemes are subject to an entropy inequality consistent with that of the continuum system. Furthermore, for a class of problems, with a suitable choice of numerical fluxes, the numerical method is compatible with the positivity-preserving procedure established in [87] and [86]. The extension to two-dimensional problems on Cartesian meshes is also discussed. In our numerical tests, we observe that the methods achieve high-order accuracy. Though the convergence rate is reduced with odd-order piecewise polynomials, the optimal rate can be retrieved by imposing a larger Lax-Friedrichs constant in the numerical fluxes. Numerical simulations to problems with positivity-preserving issues have also been performed.



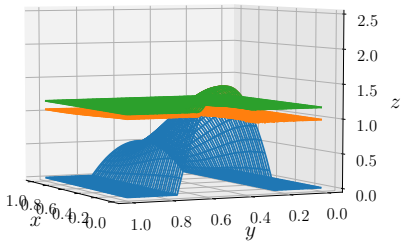
(a) $t = 0$.



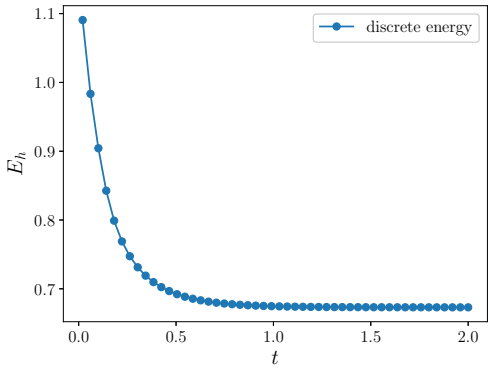
(b) $t = 0.2$.



(c) $t = 0.79$.



(d) $t = 12$.



(e) E_h.

Figure 4.5: Numerical solutions to the seawater intrusion problem in Example 4.5.3 at $t = 0, 0.2, 0.79, 12$. Solutions are obtained with piecewise cubic polynomials on a uniform square mesh on $[0, 1] \times [0, 1]$, with $h^x = h^y = 0.05$. The time step is set as $\tau = 0.002(h^x)^2$ in the simulation. b, $b + \rho_2$ and $b + \rho_1 + \rho_2$ are depicted in blue, orange and green respectively.

CONCLUSIONS AND FUTURE WORK

The major contributions of the thesis are to analyze the L^2 stability of DG methods coupling with several explicit time discretization methods and to design a entropy stable DG methods for gradient flow problems that is compatible with a positivity-preserving procedure. The thesis is mainly based on our publications [71, 72, 70].

On stability of time discretizations, we firstly consider the second-order and the third-order Lax-Wendroff DG methods, which are originally proposed by Guo et al. in [40]. The linear advection equation is used as a model problem for our analysis. Under these contexts, Lax-Wendroff DG methods with upwind fluxes for all intermediate variables are shown to be equivalent with the Runge-Kutta DG methods, whose stability has already been studied in [84]. In general, different choices of numerical fluxes can be used. Among all these choices, we prove that, with uniform or non-increasing time steps satisfying a usual CFL condition $\tau \leq Ch$, the second-order Lax-Wendroff DG schemes with piecewise linear polynomials and the third-order Lax-Wendroff DG schemes with arbitrary polynomial elements are L^2 stable, as long as the numerical flux for the unknown u is upwinding. Furthermore, if the next auxiliary variable, which approximates u_t, is also upwinding, then the scheme would be strongly stable, in the sense that L^2 norm of the numerical solution will not increase at the next time level. While if downwind flux is used, the scheme is L^2 stable in a weaker sense. More specifically, the L^2 norm of the solution is bounded by that of the initial condition multiplied by a constant. These results are also generalized to multi-dimensions for advection equations with a divergence-free velocity field. Finally, as a corollary of the stability results, optimal error estimates are provided for stable Lax-Wendroff DG schemes in one dimension, and the convergence results are verified numerically both in one dimension and two dimensions.

The next step is to analyze the fourth-order Runge-Kutta method. We mainly consider the semi-negative linear autonomous ODE system, which can be resulted from different strongly stable spatial discretizations, including the DG method. A counter example is given for showing that the four stage fourth-order DG method does not preserve the strong stability of the ODE system, whatever small time step is used. However, the strong stability will be achieved for every two time steps under a time step restriction. We then extend the result to more general well-posed and non-autonomous problems and state the stability of fully discrete schemes with Galerkin type spatial discretizations.

On the development of DG methods for gradient flows, we start with scalar problems, including both the nonlinear parabolic equations, such as porous media equations and Fokker-Planck equations, and Wasserstein gradient flows that may involving nonlocal interaction effects. Noting that a direct application of the local DG method may break the weak positivity of the scheme, we apply the technique in [22] and introduce the Gauss-Lobatto quadrature rule. The resulting semi-discrete preserve the entropy decay property of the continuum equation, if the problem does not involve a non-smooth interaction potential. Furthermore, it can be shown that under sufficiently small time step, the method preserves the weak positivity with the Euler forward time discretization. Hence one can combine the scaling limiter and the SSP-RK method, as have been done in [87] and [86], to ensure that the fully discrete scheme produces non-negative solutions. The method is also extended to two dimensions with Cartesian meshes. Numerical examples are shown to validate the high-order accuracy and the ability for preserving long time asymptotics.

We then apply the method to cross-diffusion systems, which feature with a similar entropy decay property and the positivity is usually expected as well. The idea is to introduce a new velocity function to formally decouple the system. Then one

can apply the method in Chapter 3 for each equation. The resulting semi-discrete scheme is also associated with a decaying entropy and it is also compatible with the positivity-preserving procedure stated before. We validate the performance of the method with both one-dimensional and two-dimensional numerical tests.

This thesis can serve as a starting point for a few future projects. For time discretizations, our ongoing work is to analyze the stability of implicit-explicit time stepping methods with the DG spatial discretization for KdV type equations. We would also like to study the L^2 stability of other multi-stage and multi-derivative time discretizations. For gradient flows, one future direction is to extent the method to unstructured meshes, to fit more complicated geometry and to resolve the blow up behaviors. Another direction is to develop a well-balanced numerical scheme based on the current method, to preserve the steady state for problems with source terms. We will also explore the possibility of applying the method to other entropy or energy dissipative systems with non-negative solutions.

APPENDIX

Appendix A

Appendix for Part I

A.1 Relationships between RKDG and LWDG methods

We only consider the third order temporal discretization. The relationships of the second order schemes follow from the same lines.

Consider the time-dependent problem $\boldsymbol{u}_t = L\boldsymbol{u}$. One can apply Lax-Wendroff procedure to write down a semi-discretized scheme with respect to time.

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \tau \boldsymbol{p}^n + \frac{\tau^2}{2}\boldsymbol{q}^n + \frac{\tau^3}{6} L\boldsymbol{q}^n,$$

$$\boldsymbol{p}^n = L\boldsymbol{u}^n, \qquad \boldsymbol{q}^n = L\boldsymbol{p}^n,$$

where τ is the time step.

In general, to obtain a fully discretized scheme, one can replace L with different consistent spatial discretization operators. For example, for the DG method, one can apply different numerical fluxes, and the final scheme will still approximate the original equation.

Now we assume the same spatial discretization L_N is used for each L. Then the scheme becomes

$$\begin{aligned} \boldsymbol{u}^{n+1} &= \boldsymbol{u}^n + \tau \boldsymbol{p}^n + \frac{\tau^2}{2}\boldsymbol{q}^n + \frac{\tau^3}{6} L_N\boldsymbol{q}^n, \\ \boldsymbol{p}^n &= L_N\boldsymbol{u}^n, \qquad \boldsymbol{q}^n = L_N\boldsymbol{p}^n, \end{aligned} \tag{A.1}$$

which corresponds to the truncated Taylor expansion for the semi-discretized scheme $\boldsymbol{u}_t = L_N\boldsymbol{u}$.

Assume L_N to be linear and independent of t. If the third order SSP Runge-Kutta method is applied for time stepping, one will get

$$\begin{aligned}
\boldsymbol{u}^{(1)} &= \boldsymbol{u}^n + \tau L_N \boldsymbol{u}^n = \boldsymbol{u}^n + \tau \boldsymbol{p}^n, \\
\boldsymbol{u}^{(2)} &= \frac{3}{4}\boldsymbol{u}^n + \frac{1}{4}(\boldsymbol{u}^{(1)} + \tau L_N \boldsymbol{u}^{(1)}) \\
&= \frac{3}{4}\boldsymbol{u}^n + \frac{1}{4}(\boldsymbol{u}^n + \tau \boldsymbol{p}^n + \tau L_N \boldsymbol{u}^n + \tau^2 L_N \boldsymbol{p}^n) \\
&= \boldsymbol{u}^n + \frac{\tau}{2}\boldsymbol{p}^n + \frac{\tau^2}{4}\boldsymbol{q}^n, \\
\boldsymbol{u}^{n+1} &= \frac{1}{3}\boldsymbol{u}^n + \frac{2}{3}(\boldsymbol{u}^{(2)} + \tau L \boldsymbol{u}^{(2)}) \\
&= \frac{1}{3}\boldsymbol{u}^n + \frac{2}{3}(\boldsymbol{u}^n + \frac{\tau}{2}\boldsymbol{p}^n + \frac{\tau^2}{4}\boldsymbol{q}^n + \tau L_N \boldsymbol{u}^n + \frac{\tau^2}{2} L_N \boldsymbol{p}^n + \frac{\tau^3}{4} L_N \boldsymbol{q}^n) \\
&= \boldsymbol{u}^n + \tau \boldsymbol{p}^n + \frac{\tau^2}{2}\boldsymbol{q}^n + \frac{\tau^3}{6} L_N \boldsymbol{q}^n.
\end{aligned} \tag{A.2}$$

After comparing (A.1) and (A.2), we will see, if the same spatial discretization L_N is used, with L_N being linear and independent of time t, then the fully discretized schemes obtain from the Lax-Wendroff procedure and the Runge-Kutta method are equivalent after one full time step.

In particular, let L_N corresponds to the DG discretization for multidimensional system $\boldsymbol{u}_t = \sum_{i=1}^{d} (A_i(\boldsymbol{x})\boldsymbol{u})_{x_i}$. One can check this specific L_N also satisfies the assumptions above. If this L_N is used through out the Lax-Wendroff procedure, then the scheme will be equivalent to the corresponding RKDG scheme.

A.2 Proofs of several lemmas in Chapter 1

A.2.1 Proof of Lemma 1.3.5

Proof. We restore the integral notation in this proof.

(1) In $\mathcal{H}_{\boldsymbol{\beta}}^{-}$, $w = w^{+}$ if $\boldsymbol{\beta}\cdot\boldsymbol{n} < 0$, $w = w^{-}$ if $\boldsymbol{\beta}\cdot\boldsymbol{n} > 0$. Hence

$$\begin{aligned}
\mathcal{H}_{\boldsymbol{\beta}}^{-}(w,w) &= \sum_{K\in\mathcal{K}}\frac{1}{2}\int_{K}(\boldsymbol{\beta}\cdot\nabla)w^{2}dx - \sum_{K\in\mathcal{K}}\int_{\partial K} w^{-}w\boldsymbol{\beta}\cdot\boldsymbol{n}dl \\
&= \sum_{K\in\mathcal{K}}\int_{\partial K}(\frac{1}{2}w - w^{-})w\boldsymbol{\beta}\cdot\boldsymbol{n}dl \\
&= \sum_{e\in\mathcal{E}}\int_{e} -\frac{1}{2}(w^{-})^{2}|\boldsymbol{\beta}\cdot\boldsymbol{n}| - (\frac{1}{2}w^{+} - w^{-})w^{+}|\boldsymbol{\beta}\cdot\boldsymbol{n}|dl \\
&= \sum_{e\in\mathcal{E}}\int_{e} -\frac{1}{2}(w^{-} - w^{+})^{2}|\boldsymbol{\beta}\cdot\boldsymbol{n}|dl \\
&= -\frac{1}{2}[\![w]\!]_{\boldsymbol{\beta}}^{2}.
\end{aligned}$$

(2) Since $\nabla\cdot\boldsymbol{\beta} = 0$, $(\boldsymbol{\beta}\cdot\nabla)v = \nabla\cdot(\boldsymbol{\beta}v)$

$$\begin{aligned}
\mathcal{H}_{\boldsymbol{\beta}}^{-}(w,v) &= \sum_{K\in\mathcal{K}}\int_{K} w\nabla\cdot(\boldsymbol{\beta}v)dx - \sum_{K\in\mathcal{K}}\int_{\partial K} w^{-}v\boldsymbol{\beta}\cdot\boldsymbol{n}dl \\
&= -\sum_{K\in\mathcal{K}}\int_{K} v(\boldsymbol{\beta}\cdot\nabla)wdx - \sum_{K\in\mathcal{K}}\int_{\partial K}(w^{-} - w)v\boldsymbol{\beta}\cdot\boldsymbol{n}dl \\
&= -\sum_{K\in\mathcal{K}}\int_{K} v(\boldsymbol{\beta}\cdot\nabla)wdx + \sum_{e\in\mathcal{E}}\int_{e}(w^{-} - w^{+})v^{+}|\boldsymbol{\beta}\cdot\boldsymbol{n}|dl \qquad \text{(A.3)} \\
&= -\sum_{K\in\mathcal{K}}\int_{K} v(\boldsymbol{\beta}\cdot\nabla)wdx + \sum_{K\in\mathcal{K}}\int_{\partial K} wv^{+}\boldsymbol{\beta}\cdot\boldsymbol{n}dl \\
&= -\mathcal{H}_{\boldsymbol{\beta}}^{+}(v,w).
\end{aligned}$$

(3)

$$\begin{aligned}\mathcal{H}_{\boldsymbol{\beta}}^{-}(w,v) &= \sum_{K\in\mathcal{K}}\int_K w(\boldsymbol{\beta}\cdot\nabla)v dx - \sum_{K\in\mathcal{K}}\int_{\partial K} w^{-}v\boldsymbol{\beta}\cdot\boldsymbol{n}dl \\ &= \sum_{K\in\mathcal{K}}\int_K w(\boldsymbol{\beta}\cdot\nabla)v dx - \sum_{K\in\mathcal{K}}\int_{\partial K} w^{+}v\boldsymbol{\beta}\cdot\boldsymbol{n}dl + \sum_{K\in\mathcal{K}}\int_{\partial K}[w]v\boldsymbol{\beta}\cdot\boldsymbol{n}dl \\ &= \mathcal{H}_{\boldsymbol{\beta}}^{+}(w,v) + \sum_{e\in\mathcal{E}}\int_e [w](v^{-}-v^{+})|\boldsymbol{\beta}\cdot\boldsymbol{n}|dl \\ &= \mathcal{H}_{\boldsymbol{\beta}}^{+}(w,v) - [w,v]_{\boldsymbol{\beta}}.\end{aligned}$$

□

A.2.2 Proof of Lemma 1.3.6

Proof. (1) According to (A.3),

$$\mathcal{H}_{\boldsymbol{\beta}}^{-}(w,v) = -\sum_{K\in\mathcal{K}}\int_K v(\boldsymbol{\beta}\cdot\nabla)w dx - \sum_{e\in\mathcal{E}}\int_e [w]v^{+}|\boldsymbol{\beta}\cdot\boldsymbol{n}|dl.$$

Therefore, by the Cauchy-Schwartz inequality,

$$\begin{aligned}|\mathcal{H}_{\boldsymbol{\beta}}^{-}(w,v)| &\le \sum_{K\in\mathcal{K}}\|v\|_K\|(\boldsymbol{\beta}\cdot\nabla)w\|_K + \sqrt{\sum_{e\in\mathcal{E}}\int_e [w]^2|\boldsymbol{\beta}\cdot\boldsymbol{n}|dl}\sqrt{\sum_{e\in\mathcal{E}}\int_e (v^{+})^2|\boldsymbol{\beta}\cdot\boldsymbol{n}|dl} \\ &\le \|(\boldsymbol{\beta}\cdot\nabla)w\|\|v\| + C_{\boldsymbol{\beta}}[\![w]\!]_{\boldsymbol{\beta}}\sqrt{\sum_{e\in\mathcal{E}}\int_e (v^{+})^2 dl}.\end{aligned}$$

We would like to bound the term $\sqrt{\sum_{e\in\mathcal{E}}\int_e (v^{+})^2 dl}$ with $\|v\|$. However, v^{+} will no longer be a polynomial if $\boldsymbol{\beta}(\boldsymbol{x})\cdot\boldsymbol{n}$ changes sign along the edge, and in this case one can not directly apply the trace inverse inequality. Therefore, we introduce v^{l} and

v^r to denote the left and right limit of v along the edge. Then it gives

$$|\mathcal{H}^-_{\boldsymbol{\beta}}(w,v)| \le \|(\boldsymbol{\beta}\cdot\nabla)w\|\|v\| + C_{\boldsymbol{\beta}}[\![w]\!]_{\boldsymbol{\beta}}\sqrt{\sum_{e\in\mathcal{E}}\int_e 2(v^l)^2+2(v^r)^2 dl},$$
$$\le (\|(\boldsymbol{\beta}\cdot\nabla)w\| + C_{\mu,\boldsymbol{\beta}}h^{-\frac{1}{2}}[\![w]\!]_{\boldsymbol{\beta}})\|v\|.$$

(2) Use $\mathcal{H}^-_{\boldsymbol{\beta}}(w,v) = -\mathcal{H}^+_{\boldsymbol{\beta}}(v,w)$. The remaining steps are similar. □

A.2.3 Proof of Lemma 1.3.8

Proof. For simplicity, we drop all the subscripts h and superscripts n. Let $\boldsymbol{x}_0$ be the centroid of the element K in d-dimensions. Denote $\boldsymbol{\beta}_0 = \boldsymbol{\beta}(\boldsymbol{x}_0)$. Let $\mathcal{P}_1(K) = V_0 + V_1 + V_2$, where

$$V_0 = \{y \in \mathcal{P}_1(K) | y \text{ is a constant}\},$$

$$V_1 = \{y \in \mathcal{P}_1(K) | (y,z)_K = 0, \forall z \in V_0 \text{ and } (\boldsymbol{\beta}_0\cdot\nabla)y = 0\},$$

$$V_2 = \{y \in \mathcal{P}_1(K) | (y,z)_K = 0, \forall z \in V_0 \cup V_1\}.$$

Then we can assume $p = p_0\phi^0 + p_1\phi^1 + p_2\phi^2$, where $\phi^i \in V_i$ and $\|\phi^i\|_K = 1$. Note that $\nabla\phi^0 = \vec{0}$ and $(\boldsymbol{\beta}_0\cdot\nabla)\phi^1 = 0$. Using the inverse inequality, we get

$$\|(\boldsymbol{\beta}\cdot\nabla)p\|_K \le \|p_2(\boldsymbol{\beta}\cdot\nabla)\phi^2\|_K + \|p_1((\boldsymbol{\beta}-\boldsymbol{\beta}_0)\cdot\nabla)\phi^1\|_K$$
$$\le C_{\mu,\boldsymbol{\beta}}h^{-1}|p_2|\|\phi^2\|_K + C_\mu h^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_{L^\infty}|p_1|\|\phi^1\|_K$$
$$\le C_{\mu,\boldsymbol{\beta}}h^{-1}|p_2| + C_\mu h^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_{L^\infty}|p_1|.$$

Denote $e^+ = \{x \in \partial K | \boldsymbol{\beta}(\boldsymbol{x})\cdot\boldsymbol{n} > 0\}$, and $e^- = \partial K \backslash e^+$. We integrate by parts to get p_2.

$$p_2 = (p,\phi^2)_K = -(u,(\boldsymbol{\beta}\cdot\nabla)\phi^2)_K + \langle u^+, \phi^2\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial K}$$

$$
\begin{aligned}
&= ((\boldsymbol{\beta}\cdot\nabla)u, \phi^2)_K - \langle u, \phi^2\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial K} + \langle u^+, \phi^2\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial K} \\
&= ((\boldsymbol{\beta}\cdot\nabla)u, \phi^2)_K - \langle u^-, \phi^2|\boldsymbol{\beta}\cdot\boldsymbol{n}|\rangle_{e^+} + \langle u^+, \phi^2|\boldsymbol{\beta}\cdot\boldsymbol{n}|\rangle_{e^-} + \langle u^+, \phi^2\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial K} \\
&= (((\boldsymbol{\beta}-\boldsymbol{\beta}_0)\cdot\nabla)u, \phi^2)_K + \langle [u], \phi^2|\boldsymbol{\beta}\cdot\boldsymbol{n}|\rangle_{e^+},
\end{aligned}
$$

where we have used the fact $(\boldsymbol{\beta}_0\cdot\nabla)u \in V_0$. By the trace inverse inequality,

$$
\begin{aligned}
|p_2| &\le C_{\mu,\boldsymbol{\beta}}\|[\![u]\!]_{\boldsymbol{\beta},\partial K}\|\phi^2\|_{\partial K} + C_\mu h^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_{L^\infty}\|u\|_K \\
&\le C_{\mu,\boldsymbol{\beta}}h^{-\frac{1}{2}}[\![u]\!]_{\boldsymbol{\beta},\partial K} + C_\mu h^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_{L^\infty}\|u\|_K.
\end{aligned}
$$

Similarly, one can show

$$
|p_1| \le C_{\mu,\boldsymbol{\beta}}h^{-\frac{1}{2}}[\![u]\!]_{\boldsymbol{\beta},\partial K} + C_\mu h^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_{L^\infty}\|u\|_K.
$$

(i) When $\boldsymbol{\beta}$ is a constant, $\boldsymbol{\beta}-\boldsymbol{\beta}_0 = 0$. Therefore, $\|(\boldsymbol{\beta}\cdot\nabla)p\|_K \le C_{\mu,\boldsymbol{\beta}}h^{-1}|p_2| \le C_{\mu,\boldsymbol{\beta}}h^{-\frac{3}{2}}[\![u]\!]_{\boldsymbol{\beta},\partial K}$. (ii) For non-constant $\boldsymbol{\beta}$, by the regularity of $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_{L^\infty} \le C_{\boldsymbol{\beta}}h$. Therefore

$$
\begin{aligned}
\|(\boldsymbol{\beta}\cdot\nabla)p\|_K &\le C_{\mu,\boldsymbol{\beta}}h^{-1}|p_2| + C_{\mu,\boldsymbol{\beta}}|p_1| \\
|p_2| &\le C_{\mu,\boldsymbol{\beta}}h^{-\frac{1}{2}}[\![u]\!]_{\boldsymbol{\beta},\partial K} + C_{\mu,\boldsymbol{\beta}}\|u\|_K \\
|p_1| &\le C_{\mu,\boldsymbol{\beta}}h^{-\frac{1}{2}}[\![u]\!]_{\boldsymbol{\beta},\partial K} + C_{\mu,\boldsymbol{\beta}}\|u\|_K,
\end{aligned}
$$

which implies $\|(\boldsymbol{\beta}\cdot\nabla)p\|_K \le C_{\mu,\boldsymbol{\beta}}h^{-\frac{3}{2}}[\![u]\!]_{\boldsymbol{\beta},\partial K} + C_{\mu,\boldsymbol{\beta}}h^{-1}\|u\|_K$. □

A.3 Proof of Proposition 2.1.1

Proof. (1) To prove $L_N + L_N^T \leq 0$.

$$L_N + L_N^T = -\begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix} = -2\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \end{pmatrix}.$$

Hence for any $u = (u_1, u_2, u_3)^T \in \mathbb{R}^3$,

$$u^T(L_N + L_N^T)u = -2(u_1 + u_2 + u_3)^2 \leq 0.$$

(2) To prove $\|P_4(\tau L_N)\| > 1$.

By definition,

$$P_4(\tau L_N) = \begin{pmatrix} 1 - \tau + \frac{\tau^2}{2} - \frac{\tau^3}{6} + \frac{\tau^4}{24} & -2\tau + 2\tau^2 - \tau^3 + \frac{\tau^4}{3} & -2\tau + 4\tau^2 - 3\tau^3 + \frac{4\tau^4}{3} \\ 0 & 1 - \tau + \frac{\tau^2}{2} - \frac{\tau^3}{6} + \frac{\tau^4}{24} & -2\tau + 2\tau^2 - \tau^3 + \frac{\tau^4}{3} \\ 0 & 0 & 1 - \tau + \frac{\tau^2}{2} - \frac{\tau^3}{6} + \frac{\tau^4}{24} \end{pmatrix}.$$

Since $\|P_4(\tau L_N)\| = \sqrt{\lambda_{\max}(P_4(\tau L_N)^T P_4(\tau L_N))}$ is the square root of the largest eigenvalue of $P_4(\tau L_N)^T P_4(\tau L_N)$, we define $f(\lambda, \tau) = \det(\lambda I - P_4(\tau L_N)^T P_4(\tau L_N))$. It suffices to show, for any sufficiently small τ, $f(\cdot, \tau)$ has a root larger than 1.

After direct calculation, one can obtain

$$f(\lambda, \tau) = \lambda^3 + c_2\lambda^2 + c_1\lambda + c_0,$$

where

$$c_0 = -\left(1 - \tau + \frac{1}{2}\tau^2 - \frac{1}{6}\tau^3 + \frac{1}{24}\tau^4\right)^6,$$

$$\begin{aligned} c_1 = 3 - 12\tau + 36\tau^2 - 72\tau^3 + 100\tau^4 - \frac{211\tau^5}{2} + \frac{1069\tau^6}{12} \\ - \frac{257\tau^7}{4} + \frac{4081\tau^8}{96} - \frac{3827\tau^9}{144} + \frac{2225\tau^{10}}{144} - \frac{1133\tau^{11}}{144} \\ + \frac{1895\tau^{12}}{576} - \frac{3709\tau^{13}}{3456} + \frac{197\tau^{14}}{768} - \frac{851\tau^{15}}{20736} + \frac{385\tau^{16}}{110592}, \end{aligned}$$

$$c_2 = -3 + 6\tau - 18\tau^2 + 36\tau^3 - 46\tau^4 + \frac{163\tau^5}{4} - \frac{589\tau^6}{24} + \frac{75\tau^7}{8} - \frac{385\tau^8}{192}.$$

One can see

$$\begin{aligned} f(1,\tau) = -\frac{8\tau^9}{27} + \frac{\tau^{10}}{3} - \frac{\tau^{11}}{9} - \frac{59\tau^{12}}{216} + \frac{341\tau^{13}}{864} - \frac{247\tau^{14}}{864} + \frac{1435\tau^{15}}{10368} \\ - \frac{19\tau^{16}}{384} + \frac{2299\tau^{17}}{165888} - \frac{4757\tau^{18}}{1492992} + \frac{79\tau^{19}}{124416} - \frac{107\tau^{20}}{995328} \\ + \frac{179\tau^{21}}{11943936} - \frac{13\tau^{22}}{7962624} + \frac{\tau^{23}}{7962624} - \frac{\tau^{24}}{191102976}, \end{aligned}$$

$$\begin{aligned} f(2,\tau) =& 1 + 6\tau - 18\tau^2 + 36\tau^3 - 38\tau^4 + \frac{67\tau^5}{4} + \frac{371\tau^6}{24} - \frac{289\tau^7}{8} + \frac{7007\tau^8}{192} \\ & - \frac{11609\tau^9}{432} + \frac{2273\tau^{10}}{144} - \frac{383\tau^{11}}{48} + \frac{5213\tau^{12}}{1728} - \frac{2345\tau^{13}}{3456} - \frac{203\tau^{14}}{6912} \\ & + \frac{673\tau^{15}}{6912} - \frac{5087\tau^{16}}{110592} + \frac{2299\tau^{17}}{165888} - \frac{4757\tau^{18}}{1492992} + \frac{79\tau^{19}}{124416} - \frac{107\tau^{20}}{995328} \\ & + \frac{179\tau^{21}}{11943936} - \frac{13\tau^{22}}{7962624} + \frac{\tau^{23}}{7962624} - \frac{\tau^{24}}{191102976}. \end{aligned}$$

Hence, for any sufficiently small τ,

$$f(1,\tau) = -\frac{8\tau^8}{27} + \mathcal{O}(\tau^9) < 0, \qquad f(2,\tau) = 1 + 6\tau + \mathcal{O}(\tau^2) > 1.$$

Note $f(\lambda,\tau)$ is a polynomial with respect to λ and τ, hence it is continuous. By the intermediate value theorem, there exists $\lambda(\tau) \in (1,2)$ such that $f(\lambda(\tau),\tau) = 0$. Therefore

$$\|P_4(\tau L_N)\| \geq \sqrt{\lambda(\tau)} > 1.$$

□

Remark A.3.1. *We also provide the numerical examination of Proposition 2.1.1. The values of* $\|P_4(\tau L_N)\| - 1$ *and* $\|P_4(\tau L_N)^2\| - 1$ *with a decreasing sequence of* τ *are listed in Table A.1. As we can see, for this specific case, the one-step strong stability does fail, but the two-step strong stability holds.*

Table A.1: The numerical examination of Proposition 2.1.1 and Theorem 2.2.1.

τ	$\|P_4(\tau L_N)\| - 1$	$\|P_4(\tau L_N)^2\| - 1$
0.5	1.2794e-3	-1.0428e-3
0.2	8.3857e-6	-1.8788e-5
0.1	2.2173e-7	-6.6719e-7
0.05	6.3437e-9	-2.2029e-8
0.02	6.1504e-11	-2.3254e-10
0.01	1.8868e-12	-7.3375e-12

A.4 Coefficient matrix $\tilde{A}$

$$\tilde{A} = -\begin{pmatrix} 1 & 3/2 & 7/6 & 5/8 & 1/4 & 5/72 & 1/72 & 1/576 \\ 3/2 & 7/3 & 15/8 & 25/24 & 31/72 & 1/8 & 5/192 & 1/288 \\ 7/6 & 15/8 & 37/24 & 127/144 & 3/8 & 65/576 & 7/288 & 1/288 \\ 5/8 & 25/24 & 127/144 & 25/48 & 131/576 & 61/864 & 1/64 & 1/432 \\ 1/4 & 31/72 & 3/8 & 131/576 & 175/1728 & 37/1152 & 25/3456 & 5/4608 \\ 5/72 & 1/8 & 65/576 & 61/864 & 37/1152 & 1/96 & 11/4608 & 5/13824 \\ 1/72 & 5/192 & 7/288 & 1/64 & 25/3456 & 11/4608 & 23/41472 & 7/82944 \\ 1/576 & 1/288 & 1/288 & 1/432 & 5/4608 & 5/13824 & 7/82944 & 1/82944 \end{pmatrix}.$$

A.5 Proof of Lemma 2.3.2

Proof. (1) To prove $\mathcal{H}(\boldsymbol{v}_h, \boldsymbol{v}_h) \geq 0$.

We denote by $\boldsymbol{v}_{e,h} = \boldsymbol{S}_e \boldsymbol{v}_h$ and $\widehat{\boldsymbol{v}}_{e,h} = \widehat{\boldsymbol{S}_e \boldsymbol{v}_h}$. Since $\boldsymbol{A}_i$ are symmetric. Integrating by parts, one has

$$\sum_{i=1}^{d} (\boldsymbol{A}_i \boldsymbol{v}_h, \partial_{x_i} \boldsymbol{v}_h)_K = \sum_{e \in \partial K} \langle \frac{1}{2} \sum_{i=1}^{d} n_{e,K}^i \boldsymbol{A}_i \boldsymbol{v}_h, \boldsymbol{v}_h \rangle_e = \sum_{e \in \partial K} \langle \frac{1}{2} \boldsymbol{\Lambda}_e \boldsymbol{v}_{e,h}, \boldsymbol{v}_{e,h} \rangle_e. \tag{A.4}$$

Substitute (A.4) into (2.30), then we obtain

$$\mathcal{H}(\boldsymbol{v}_h, \boldsymbol{v}_h) \quad = \quad \sum_{K \in \mathcal{K}} \sum_{e \in \partial K} \langle \boldsymbol{\Lambda}_e (\frac{1}{2} \boldsymbol{v}_{e,h} - \widehat{\boldsymbol{v}}_{e,h}), \boldsymbol{v}_{e,h} \rangle_e.$$

Each cell interface e is shared by two elements. We sit in either side and call $\boldsymbol{v}_{e,h}$ defined in this element to be $\boldsymbol{v}_{e,h}^{int}$. $\boldsymbol{v}_{e,h}$ defined from the other side is referred as $\boldsymbol{v}_{e,h}^{ext}$. Then the summation can be rewritten as

$$\begin{aligned} \mathcal{H}(\boldsymbol{v}_h, \boldsymbol{v}_h) &= \sum_{e \in \mathcal{E}} \left(\langle \boldsymbol{\Lambda}_e (\frac{1}{2} \boldsymbol{v}_{e,h}^{int} - \widehat{\boldsymbol{v}}_{e,h}), \boldsymbol{v}_{e,h}^{int} \rangle_e + \langle -\boldsymbol{\Lambda}_e (\frac{1}{2} \boldsymbol{v}_{e,h}^{ext} - \widehat{\boldsymbol{v}}_{e,h}), \boldsymbol{v}_{e,h}^{ext} \rangle_e \right) \\ &= \sum_{e \in \mathcal{E}} (\frac{1}{2} \langle \boldsymbol{\Lambda}_e \boldsymbol{v}_{e,h}^{int}, \boldsymbol{v}_{e,h}^{int} \rangle_e - \frac{1}{2} \langle \boldsymbol{\Lambda}_e \boldsymbol{v}_{e,h}^{ext}, \boldsymbol{v}_{e,h}^{ext} \rangle_e + \langle \boldsymbol{\Lambda}_e \widehat{\boldsymbol{v}}_{e,h}, \boldsymbol{v}_{e,h}^{ext} - \boldsymbol{v}_{e,h}^{int} \rangle_e). \end{aligned}$$

By checking the definition of $\widehat{\boldsymbol{v}}_{e,h}$, one can prove that

$$\mathcal{H}(\boldsymbol{v}_h, \boldsymbol{v}_h) = \frac{1}{2} \sum_{e \in \mathcal{E}} \langle \text{abs}(\boldsymbol{\Lambda}_e)(\boldsymbol{v}_{e,h}^{ext} - \boldsymbol{v}_{e,h}^{int}), (\boldsymbol{v}_{e,h}^{ext} - \boldsymbol{v}_{e,h}^{int}) \rangle_e,$$

where $\text{abs}(\boldsymbol{\Lambda}_e) = \text{diag}(|\lambda_e^1|, ..., |\lambda_e^m|)$. Hence $\mathcal{H}(\boldsymbol{v}_h, \boldsymbol{v}_h)$ is non-negative.

(2) To prove $|\mathcal{H}(\boldsymbol{w}_h, \boldsymbol{v}_h)| \leq C h^{-1} \|\boldsymbol{w}_h\| \|\boldsymbol{v}_h\|$.

We will use the notation $|\cdot|$ for different meanings. For scalars, $|a|$ is the absolute value of a; for vectors, $|\boldsymbol{a}|$ stands for the Euclidean norm of $\boldsymbol{a}$; and for matrices, $|\boldsymbol{A}|$ is the operator norm of $\boldsymbol{A}$. We also use the notation $\|\cdot\|_K = \sqrt{(\cdot,\cdot)_K}$.

By using the Cauchy- Schwarz inequality and the inverse estimate, one has

$$\begin{aligned} |\sum_{K\in\mathcal{K}}\sum_{i=1}^{d}(\boldsymbol{A}_i\boldsymbol{w}_h, \partial_{x_i}\boldsymbol{v}_h)_K| &\le \sum_{K\in\mathcal{K}}\sum_{i=1}^{d}|\boldsymbol{A}_i|\|\boldsymbol{w}_h\|_K\|\partial_{x_i}\boldsymbol{v}_h\|_K \\ &\le \left(Ch^{-1}\sum_{i=1}^{d}|\boldsymbol{A}_i|\right)\sum_{K\in\mathcal{K}}\|\boldsymbol{w}_h\|_K\|\boldsymbol{v}_h\|_K \\ &\le \left(Ch^{-1}\sum_{i=1}^{d}|\boldsymbol{A}_i|\right)\|\boldsymbol{w}_h\|\|\boldsymbol{v}_h\|. \end{aligned} \tag{A.5}$$

And

$$|\sum_{K\in\mathcal{K}}\sum_{e\in\partial K}\langle\boldsymbol{\Lambda}_e\widehat{\boldsymbol{S}_e\boldsymbol{w}_h}, \boldsymbol{S}_e\boldsymbol{v}_h\rangle_e| \le \sup_{k,e}|\lambda_e^k|\sum_{K\in\mathcal{K}}\sum_{e\in\partial K}\sqrt{\int_e|\widehat{\boldsymbol{S}_e\boldsymbol{w}_h}|^2dl}\sqrt{\int_e|\boldsymbol{S}_e\boldsymbol{v}_h|^2dl}.$$

Note that $\boldsymbol{S}_e$ are orthogonal, $|\boldsymbol{S}_e\boldsymbol{v}_h| = |\boldsymbol{v}_h|$. By using the inverse inequality, we obtain

$$\sqrt{\int_e|\boldsymbol{S}_e\boldsymbol{v}_h|^2dl} = \sqrt{\int_e|\boldsymbol{v}_h|^2dl} \le Ch^{-\frac{1}{2}}\|\boldsymbol{v}_h\|_K,$$

and

$$\sqrt{\int_e|\widehat{\boldsymbol{S}_e\boldsymbol{w}_h}|^2dl} \le \sqrt{\int_e|\boldsymbol{w}_h^{int}|^2+|\boldsymbol{w}_h^{ext}|^2dl} \le Ch^{-\frac{1}{2}}\|\boldsymbol{w}_h\|_{\mathcal{N}(K)},$$

where $\mathcal{N}(K)$ is the union of K and its neighboring elements. Hence

$$\begin{aligned} |\sum_{K\in\mathcal{K}}\sum_{e\in\partial K}\langle\boldsymbol{\Lambda}_e\widehat{\boldsymbol{S}_e\boldsymbol{w}_h}, \boldsymbol{S}_e\boldsymbol{v}_h\rangle_e| \le& Ch^{-1}\sup_{k,e}|\lambda_e^k|\sum_{K\in\mathcal{K}}\|\boldsymbol{w}_h\|_{\mathcal{N}(K)}\|\boldsymbol{v}_h\|_K \\ \le& Ch^{-1}\sup_{k,e}|\lambda_e^k|\sqrt{\sum_{K\in\mathcal{K}}\|\boldsymbol{w}_h\|^2_{\mathcal{N}(K)}}\sqrt{\sum_{K\in\mathcal{K}}\|\boldsymbol{v}_h\|^2_K} \\ \le& Ch^{-1}\sup_{k,e}|\lambda_e^k|\|\boldsymbol{w}_h\|\|\boldsymbol{v}_h\|. \end{aligned} \tag{A.6}$$

Combining (A.5) and (A.6), we get

$$|\mathcal{H}(\boldsymbol{w}_h, \boldsymbol{v}_h)| \leq Ch^{-1}\|\boldsymbol{w}_h\|\|\boldsymbol{v}_h\|.$$

□

Appendix B

Appendix for Part II

B.1 Proofs of theorems on entropy inequalities

B.1.1 Proof of Theorem 4.2.1

Proof. $\frac{d}{dt}E_h = \sum_{i=1}^{N}\int_{I_i}\partial_t\boldsymbol{\rho}_h\cdot\boldsymbol{\xi}_h dx$. Then, with $\boldsymbol{\varphi}_h = \boldsymbol{\xi}_h$ in (4.5a), we have

$$\begin{aligned}\int_{I_i}^{\sim}\partial_t\boldsymbol{\rho}_h\cdot\boldsymbol{\xi}_h dx &= -\int_{I_i}^{\sim} F_h\boldsymbol{u}_h\cdot\partial_x\boldsymbol{\xi}_h dx + (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^-)_{i+\frac{1}{2}} - (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^+)_{i-\frac{1}{2}}\\ &= -\int_{I_i}\mathcal{I}(F_h\boldsymbol{u}_h)\cdot\partial_x\mathcal{I}(\boldsymbol{\xi}_h)dx + (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^-)_{i+\frac{1}{2}} - (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^+)_{i-\frac{1}{2}}.\end{aligned}$$

Here we have used the fact that $\mathcal{I}(F_h\boldsymbol{u}_h)\cdot\partial_x\mathcal{I}(\boldsymbol{\xi}_h)$ is a polynomial of degree $2k-1$ and Gauss-Lobatto quadrature rule is exact. Then by integrating by parts and again use the exactness of the quadrature, one can get

$$\begin{aligned}\int_{I_i}^{\sim}\partial_t\boldsymbol{\rho}_h\cdot\boldsymbol{\xi}_h dx &= \int_{I_i}\partial_x\mathcal{I}(F_h\boldsymbol{u}_h)\cdot\mathcal{I}(\boldsymbol{\xi}_h)dx - (F_h\boldsymbol{u}_h\cdot\boldsymbol{\xi}_h)^-_{i+\frac{1}{2}} + (F_h\boldsymbol{u}_h\cdot\boldsymbol{\xi}_h)^+_{i-\frac{1}{2}}\\ &\quad + (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^-)_{i+\frac{1}{2}} - (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^+)_{i-\frac{1}{2}}\\ &= \int_{I_i}^{\sim}\boldsymbol{\xi}_h\cdot\partial_x(F_h\boldsymbol{u}_h)dx - (F_h\boldsymbol{u}_h\cdot\boldsymbol{\xi}_h)^-_{i+\frac{1}{2}} + (F_h\boldsymbol{u}_h\cdot\boldsymbol{\xi}_h)^+_{i-\frac{1}{2}}\\ &\quad + (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^-)_{i+\frac{1}{2}} - (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^+)_{i-\frac{1}{2}}\\ &= -\int_{I_i}^{\sim}\boldsymbol{u}_h\cdot F_h\boldsymbol{u}_h dx + (\widehat{\boldsymbol{\xi}}\cdot(F_h\boldsymbol{u}_h)^-)_{i+\frac{1}{2}} - (\widehat{\boldsymbol{\xi}}\cdot(F_h\boldsymbol{u}_h)^+)_{i-\frac{1}{2}}\\ &\quad - (F_h\boldsymbol{u}_h\cdot\boldsymbol{\xi}_h)^-_{i+\frac{1}{2}} + (F_h\boldsymbol{u}_h\cdot\boldsymbol{\xi}_h)^+_{i-\frac{1}{2}} + (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^-)_{i+\frac{1}{2}} - (\widehat{F\boldsymbol{u}}\cdot\boldsymbol{\xi}_h^+)_{i-\frac{1}{2}}.\end{aligned}$$

After summing over the index i and using the periodicity, we obtain

$$\begin{aligned}\frac{d}{dt}E_h &= \sum_{i=1}^{N}\int_{I_i}^{\sim}\partial_t\boldsymbol{\rho}_h\cdot\boldsymbol{\xi}_h dx\\ &= -\int_{\Omega}^{\sim}\boldsymbol{u}_h\cdot F_h\boldsymbol{u}_h dx+\sum_{i=1}^{N}\left(((F_h\boldsymbol{u}_h)^- - \widehat{F\boldsymbol{u}})_{i+\frac{1}{2}}\cdot[\boldsymbol{\xi}_h]_{i+\frac{1}{2}}+[F_h\boldsymbol{u}_h]_{i+\frac{1}{2}}\cdot(\boldsymbol{\xi}_h^+-\widehat{\boldsymbol{\xi}})_{i+\frac{1}{2}}\right).\end{aligned}$$

The proof is completed by substituting numerical fluxes in (4.7) and (4.8). Note $\sum_{i=1}^{N}\alpha_{i+\frac{1}{2}}[\boldsymbol{\xi}_h]_{i+\frac{1}{2}}\cdot[\boldsymbol{\rho}_h]_{i+\frac{1}{2}} = \sum_{i=1}^{N}\alpha_{i+\frac{1}{2}}[\boldsymbol{\rho}_h]_{i+\frac{1}{2}}\cdot D^2e(\boldsymbol{\zeta}_{i+\frac{1}{2}})[\boldsymbol{\rho}_h]_{i+\frac{1}{2}}\geq 0$ due to the convexity of e. Here $\boldsymbol{\zeta}_{i+\frac{1}{2}}$ lies between $(\boldsymbol{\rho}_h)^-_{i+\frac{1}{2}}$ and $(\boldsymbol{\rho}_h)^+_{i+\frac{1}{2}}$. □

B.1.2 Proof of Theorem 4.3.1

Proof. $\frac{d}{dt}E_h=\sum_{j=1}^{N^y}\sum_{i=1}^{N^x}\int_{I_i}\partial_t\boldsymbol{\rho}_h\cdot\boldsymbol{\xi}_h dx$. Then

$$\begin{aligned}\int_{J_j}^{\sim}\int_{I_i}^{\sim}\partial_t\boldsymbol{\rho}_h\cdot\boldsymbol{\xi}_h dx =& -\int_{J_j}^{\sim}\int_{I_i}^{\sim}F_h\boldsymbol{u}_h^x\cdot\partial_x\boldsymbol{\xi}_h+F_h\boldsymbol{u}_h^y\cdot\partial_y\boldsymbol{\xi}_h dxdy\\ &+\int_{J_j}^{\sim}\widehat{F\boldsymbol{u}^x}_{i+\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x^-_{i+\frac{1}{2}},y)-\widehat{F\boldsymbol{u}^x}_{i-\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x^+_{i-\frac{1}{2}},y)dy\\ &+\int_{I_i}^{\sim}\widehat{F\boldsymbol{u}^y}_{j+\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x,y^-_{j+\frac{1}{2}})-\widehat{F\boldsymbol{u}^y}_{j-\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x,y^+_{j-\frac{1}{2}})dx,\end{aligned}$$

For each fixed y, each component of $\mathcal{I}(F_h\boldsymbol{u}^x)\partial_x\mathcal{I}(\boldsymbol{\xi}_h)$ is a polynomial of degree $2k-1$ with respect to x. $\mathcal{I}(F_h\boldsymbol{u}^y)\partial_y\mathcal{I}(\boldsymbol{\xi}_h)$ is a $(2k-1)$-th order polynomial with respect to y for each fixed x. Hence the Gauss-Lobatto quadrature with $k+1$ node is exact. We then replace the quadrature rule with the exact integral, and integrate by parts

to get

$$
\begin{aligned}
\frac{d}{dt}E =& \sum_{i=1}^{N^x}\sum_{j=1}^{N^y}\left(-\int_{J_j}^{\sim}\left(\int_{I_i}^{\sim}\mathcal{I}(F_h\boldsymbol{u}_h^x)\cdot\partial_x\mathcal{I}(\boldsymbol{\xi}_h)dx\right)dy\right.\\
&-\int_{I_i}^{\sim}\left(\int_{J_j}^{\sim}\mathcal{I}(F_h\boldsymbol{u}_h^y)\cdot\partial_y\mathcal{I}(\boldsymbol{\xi}_h)dy\right)dx\\
&+\int_{J_j}^{\sim}\widehat{F\boldsymbol{u}^x}_{i+\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x^-_{i+\frac{1}{2}},y)-\widehat{F\boldsymbol{u}^x}_{i-\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x^+_{i-\frac{1}{2}},y)dy\\
&\left.+\int_{I_i}^{\sim}\widehat{F\boldsymbol{u}^y}_{j+\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x,y^-_{j+\frac{1}{2}})-\widehat{F\boldsymbol{u}^y}_{j-\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x,y^+_{j-\frac{1}{2}})dx\right)\\
=&\sum_{i=1}^{N^x}\sum_{j=1}^{N^y}\left(\int_{J_j}^{\sim}\left(\int_{I_i}^{\sim}\partial_x\mathcal{I}(F_h\boldsymbol{u}_h^x)\cdot\mathcal{I}(\boldsymbol{\xi}_h)dx\right)dy\right.\\
&+\int_{I_i}^{\sim}\left(\int_{J_j}^{\sim}\partial_y\mathcal{I}(F_h\boldsymbol{u}_h^y)\cdot\mathcal{I}(\boldsymbol{\xi}_h)dy\right)dx\\
&-\int_{J_j}^{\sim}(F_h\boldsymbol{u}_h^x\cdot\boldsymbol{\xi}_h)(x^-_{i+\frac{1}{2}},y)-(F_h\boldsymbol{u}_h^x\cdot\boldsymbol{\xi}_h)(x^+_{i-\frac{1}{2}},y)dy\\
&-\int_{I_i}^{\sim}(F_h\boldsymbol{u}_h^y\cdot\boldsymbol{\xi}_h)(x,y^-_{j+\frac{1}{2}})-(F_h\boldsymbol{u}_h^y\cdot\boldsymbol{\xi}_h)(x,y^+_{j-\frac{1}{2}})dx\\
&+\int_{J_j}^{\sim}\widehat{F\boldsymbol{u}^x}_{i+\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x^-_{i+\frac{1}{2}},y)-\widehat{F\boldsymbol{u}^x}_{i-\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x^+_{i-\frac{1}{2}},y)dy\\
&\left.+\int_{I_i}^{\sim}\widehat{F\boldsymbol{u}^y}_{j+\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x,y^-_{j+\frac{1}{2}})-\widehat{F\boldsymbol{u}^y}_{j-\frac{1}{2}}\cdot\boldsymbol{\xi}_h(x,y^+_{j-\frac{1}{2}})dx\right).
\end{aligned}
$$

Again by changing back to the quadrature rule and applying the scheme, we obtain

$$\begin{aligned}
\frac{d}{dt}E =& \sum_{i=1}^{N^x}\sum_{j=1}^{N^y}\left(-\int_{J_j}\int_{I_i} \boldsymbol{u}_h^x \cdot F_h \boldsymbol{u}_h^x + \boldsymbol{u}_h^y \cdot F_h \boldsymbol{u}_h^y dxdy \right. \\
&+ \int_{J_j} (F_h \boldsymbol{u}_h^x)(x^-_{i+\frac{1}{2}}, y) \cdot \widehat{\boldsymbol{\xi}}_{i+\frac{1}{2}} - (F_h \boldsymbol{u}_h^x)(x^+_{i-\frac{1}{2}}, y) \cdot \widehat{\boldsymbol{\xi}}_{i-\frac{1}{2}} dy \\
&+ \int_{I_i} (F_h \boldsymbol{u}_h^y)(x, y^-_{j+\frac{1}{2}}) \cdot \widehat{\boldsymbol{\xi}}_{j+\frac{1}{2}} - (F_h \boldsymbol{u}_h^y)(x, y^+_{j-\frac{1}{2}}) \cdot \widehat{\boldsymbol{\xi}}_{j-\frac{1}{2}} dx \\
&- \int_{J_j} (F_h \boldsymbol{u}_h^x \cdot \boldsymbol{\xi}_h)(x^-_{i+\frac{1}{2}}, y) - (F_h \boldsymbol{u}_h^x \cdot \boldsymbol{\xi}_h)(x^+_{i-\frac{1}{2}}, y) dy \\
&- \int_{I_i} (F_h \boldsymbol{u}_h^y \cdot \boldsymbol{\xi}_h)(x, y^-_{j+\frac{1}{2}}) - (F_h \boldsymbol{u}_h^y \cdot \boldsymbol{\xi}_h)(x, y^+_{j-\frac{1}{2}}) dx \\
&+ \int_{J_j} \widehat{F\boldsymbol{u}^x}_{i+\frac{1}{2}} \cdot \boldsymbol{\xi}_h(x^-_{i+\frac{1}{2}}, y) - \widehat{F\boldsymbol{u}^x}_{i-\frac{1}{2}} \cdot \boldsymbol{\xi}_h(x^+_{i-\frac{1}{2}}, y) dy \\
&+ \left. \int_{I_i} \widehat{F\boldsymbol{u}^y}_{j+\frac{1}{2}} \cdot \boldsymbol{\xi}_h(x, y^-_{j+\frac{1}{2}}) - \widehat{F\boldsymbol{u}^y}_{j-\frac{1}{2}} \cdot \boldsymbol{\xi}_h(x, y^+_{j-\frac{1}{2}}) dx \right).
\end{aligned}$$

Since we have assumed periodicity, all cell interface terms are canceled out with alternating fluxes.

$$\frac{d}{dt}E_h = -\int_J\int_I \left(\boldsymbol{u}_h^x \cdot F_h \boldsymbol{u}_h^x + \boldsymbol{u}_h^y \cdot F_h \boldsymbol{u}_h^y\right) dxdy.$$

For central and Lax-Friedrichs fluxes, there remain additional penalty terms.

$$\begin{aligned}
\frac{d}{dt}E_h =& -\int_J\int_I \left(\boldsymbol{u}_h^x \cdot F_h \boldsymbol{u}_h^x + \boldsymbol{u}_h^y \cdot F_h \boldsymbol{u}_h^y\right) dxdy \\
&- \sum_{i=1}^{N^x} \int_J \frac{\alpha^x_{i+\frac{1}{2}}(y)}{2} \left(\boldsymbol{\xi}_h(x^+_{i+\frac{1}{2}}, y) - \boldsymbol{\xi}_h(x^-_{i+\frac{1}{2}}, y)\right) \cdot \left(\boldsymbol{\rho}_h(x^+_{i+\frac{1}{2}}, y) - \boldsymbol{\rho}_h(x^-_{i+\frac{1}{2}}, y)\right) dy \\
&- \sum_{j=1}^{N^y} \int_I \frac{\alpha^y_{j+\frac{1}{2}}(x)}{2} \left(\boldsymbol{\xi}_h(x, y^+_{j+\frac{1}{2}}) - \boldsymbol{\xi}_h(x, y^-_{i+\frac{1}{2}})\right) \cdot \left(\boldsymbol{\rho}_h(x, y^+_{j+\frac{1}{2}}) - \boldsymbol{\rho}_h(x, y^-_{j+\frac{1}{2}})\right) dx.
\end{aligned}$$

$-\widetilde{\int_J}\widetilde{\int_I}(\boldsymbol{u}_h^x \cdot F_h \boldsymbol{u}_h^x + \boldsymbol{u}_h^y \cdot F_h \boldsymbol{u}_h^y)\,dxdy \le 0$ since F_h is positive-semidefinite. Note the rest terms are also non-positive, due to the convexity of e. □

B.1.3 Fully discrete entropy inequality

Theorem B.1.1. *Consider the Euler forward time discretization. Let us assume $h_i = h$. Suppose $\boldsymbol{z} \cdot F\boldsymbol{z} \ge \beta_F|\boldsymbol{z}|^2$, $\boldsymbol{z} \cdot D\boldsymbol{\xi}\boldsymbol{z} \ge \beta_{D\boldsymbol{\xi}}|\boldsymbol{z}|^2$, $|F| \le \beta^F$ and $|D\boldsymbol{\xi}| \le \beta^{D\boldsymbol{\xi}}$ uniformly in $\boldsymbol{\rho}$ for some fixed constants β_F, $\beta_{D\boldsymbol{\xi}}$, β^F and $\beta^{D\boldsymbol{\xi}}$. Then when $\tau \le \min_i(\frac{\beta_F h^2}{8c_0^2 c_{inv}^2 (\beta^F)^2 \beta^{D\boldsymbol{\xi}}}, \frac{\beta_{D\boldsymbol{\xi}} h}{4c_0 c_{inv} \beta^{D\boldsymbol{\xi}} \alpha_{i+\frac{1}{2}}})$,*

$$\frac{E^{n+1} - E^n}{\tau} \le -\frac{\beta_F}{2}\widetilde{\int_\Omega}|\boldsymbol{u}|^2 dx - \frac{\beta_{D\boldsymbol{\xi}}}{4}\sum_{i=1}^N \alpha_{i+\frac{1}{2}}|[\boldsymbol{\rho}]_{i+\frac{1}{2}}|^2.$$

Here $|F|$ and $|D\boldsymbol{\xi}|$ are l^2 matrix norms of F and $D\boldsymbol{\xi}$ respectively.

Proof. We omit all subscripts h and superscripts n in this proof.

$$\begin{aligned}
E^{n+1} - E &= \sum_{i=1}^N \widetilde{\int_{I_i}} e(\boldsymbol{\rho}^{n+1}) - e(\boldsymbol{\rho}) dx \\
&= \sum_{i=1}^N \widetilde{\int_{I_i}} e(\boldsymbol{\rho}^{n+1}) - e(\boldsymbol{\rho}) - (\boldsymbol{\rho}^{n+1} - \boldsymbol{\rho}) \cdot \boldsymbol{\xi} dx + \sum_{i=1}^N \widetilde{\int_{I_i}} (\boldsymbol{\rho}^{n+1} - \boldsymbol{\rho}) \cdot \boldsymbol{\xi} dx \\
&= \sum_{i=1}^N \widetilde{\int_{I_i}} (\boldsymbol{\rho}^{n+1} - \boldsymbol{\rho}) \cdot D\boldsymbol{\xi}(\boldsymbol{\zeta})(\boldsymbol{\rho}^{n+1} - \boldsymbol{\rho}) dx + \sum_{i=1}^N \widetilde{\int_{I_i}} (\boldsymbol{\rho}^{n+1} - \boldsymbol{\rho}) \cdot \boldsymbol{\xi} dx \\
&:= A + B.
\end{aligned}$$

Here we have applied the mean value theorem and $\boldsymbol{\zeta}$ is on the line segment between

$\boldsymbol{\rho}^{n+1}$ and $\boldsymbol{\rho}$. Let $\boldsymbol{\eta} = \frac{\boldsymbol{\rho}^{n+1}-\boldsymbol{\rho}}{\tau}$. Then

$$A \le \max_{\boldsymbol{\zeta}} |D\boldsymbol{\xi}(\boldsymbol{\zeta})| \int_\Omega |\boldsymbol{\rho}^{n+1} - \boldsymbol{\rho}|^2 dx \le \tau^2 \beta^{D\boldsymbol{\xi}} \int_\Omega |\boldsymbol{\eta}|^2 dx,$$

and

$$\begin{aligned} B &= \tau \left(-\sum_{i=1}^N \widetilde{\int_{I_i}} \boldsymbol{u} \cdot F\boldsymbol{u} dx - \frac{1}{2} \sum_{i=1}^N \alpha_{i+\frac{1}{2}} [\boldsymbol{\xi}]_{i+\frac{1}{2}} \cdot [\boldsymbol{\rho}]_{i+\frac{1}{2}} \right) \\ &= \tau \left(-\widetilde{\int_\Omega} \boldsymbol{u} \cdot F\boldsymbol{u} dx - \frac{1}{2} \sum_{i=1}^N \alpha_{i+\frac{1}{2}} [\boldsymbol{\rho}]_{i+\frac{1}{2}} \cdot D\boldsymbol{\xi}(\boldsymbol{\zeta}_{i+\frac{1}{2}}) [\boldsymbol{\rho}]_{i+\frac{1}{2}} \right) \\ &\le -\tau \beta_F \widetilde{\int_\Omega} |\boldsymbol{u}|^2 dx - \frac{\tau}{2} \beta_{D\boldsymbol{\xi}} \sum_{i=1}^N \alpha_{i+\frac{1}{2}} |[\boldsymbol{\rho}]|^2_{i+\frac{1}{2}}. \end{aligned}$$

Here $\boldsymbol{\zeta}_{i+\frac{1}{2}}$ lies between $\boldsymbol{\rho}^-_{i+\frac{1}{2}}$ and $\boldsymbol{\rho}^+_{i+\frac{1}{2}}$.

Our main task is to estimate $\widetilde{\int_\Omega} |\boldsymbol{\eta}|^2 dx$.

$$\begin{aligned} \widetilde{\int_\Omega} |\boldsymbol{\eta}|^2 dx =& \sum_{i=1}^N \widetilde{\int_{I_i}} \frac{\boldsymbol{\rho}^{n+1} - \boldsymbol{\rho}^n}{\tau} \cdot \boldsymbol{\eta} dx \\ =& \sum_{i=1}^N \left(-\widetilde{\int_{I_i}} F\boldsymbol{u} \cdot \partial_x \boldsymbol{\eta} dx + (\widehat{F\boldsymbol{u}} \cdot \boldsymbol{\eta})^-_{i+\frac{1}{2}} - (\widehat{F\boldsymbol{u}} \cdot \boldsymbol{\eta})^+_{i-\frac{1}{2}} \right) \\ =& -\sum_{i=1}^N \widetilde{\int_{I_i}} F\boldsymbol{u} \cdot \partial_x \boldsymbol{\eta} dx - \sum_{i=1}^N \widehat{F\boldsymbol{u}}_{i+\frac{1}{2}} \cdot [\boldsymbol{\eta}]_{i+\frac{1}{2}} \\ \le& c_1 \sum_{i=1}^N \widetilde{\int_{I_i}} |F\boldsymbol{u}|^2 dx + \frac{1}{4c_1} \sum_{i=1}^N \widetilde{\int_{I_i}} |\partial_x \boldsymbol{\eta}|^2 dx \\ &+ 2c_2 \left(\sum_{i=1}^N |\{F\boldsymbol{u}\}|^2_{i+\frac{1}{2}} + \sum_{i=1}^N \frac{\alpha_i^2}{4} |[\boldsymbol{\rho}]|^2_{i+\frac{1}{2}} \right) + \frac{1}{4c_2} \sum_{i=1}^N |[\boldsymbol{\eta}]|^2_{i+\frac{1}{2}}. \end{aligned}$$

Here c_1 and c_2 are constants to be determined. Note that $\widetilde{\int_{I_i}} |\cdot|^2 dx$ defines a norm

on $P^k(I_i)$. Using norm equivalence and inverse estimates, we have

$$c_0^{-1}\int_{I_i}|p|^2dx \le \widetilde{\int_{I_i}}|p|^2dx \le c_0\int_{I_i}|p|^2dx, \qquad \forall p\in P^k(I_i).$$

$$\sum_{i=1}^N\widetilde{\int_{I_i}}|\partial_x\boldsymbol{\eta}|^2dx \le c_0\sum_{i=1}^N\int_{I_i}|\partial_x\boldsymbol{\eta}|^2dx \le \frac{c_0c_{inv}}{h^2}\sum_{i=1}^N\int_{I_i}|\boldsymbol{\eta}|^2dx \le \frac{c_0^2c_{inv}}{h^2}\sum_{i=1}^N\widetilde{\int_{I_i}}|\boldsymbol{\eta}|^2dx.$$

$$\sum_{i=1}^N[\boldsymbol{\eta}]^2_{i+\frac{1}{2}} \le \frac{c_{inv}}{h}\sum_{i=1}^N\int_{I_i}|\boldsymbol{\eta}|^2dx \le \frac{c_0c_{inv}}{h}\sum_{i=1}^N\widetilde{\int_{I_i}}|\boldsymbol{\eta}|^2dx.$$

c_{inv} is a constant no less than 1 in the inverse estimates. Choose $c_1 = \frac{c_0^2c_{inv}}{h^2}$ and $c_2 = \frac{c_0c_{inv}}{h}$. Then

$$\widetilde{\int_\Omega}|\boldsymbol{\eta}|^2dx \le \frac{2c_0^2c_{inv}}{h^2}\widetilde{\int_\Omega}|F\boldsymbol{u}|^2dx + \frac{4c_0c_{inv}}{h}\left(\sum_{i=1}^N|\{F\boldsymbol{u}\}|^2_{i+\frac{1}{2}} + \sum_{i=1}^N\frac{\alpha^2_{i+\frac{1}{2}}}{4}|[\boldsymbol{\rho}]|^2_{i+\frac{1}{2}}\right).$$

Since

$$\begin{aligned}\sum_{i=1}^N|\{F\boldsymbol{u}\}_{i+\frac{1}{2}}|^2 \le \frac{1}{2}\sum_{i=1}^N\left(|(F\boldsymbol{u})^+_{i-\frac{1}{2}}|^2 + |(F\boldsymbol{u})^-_{i+\frac{1}{2}}|^2\right) &\le \frac{c_{inv}}{2h}\sum_{i=1}^N\int_{I_i}|\mathcal{I}(F\boldsymbol{u})|^2dx\\ &\le \frac{c_0c_{inv}}{2h}\sum_{i=1}^N\widetilde{\int_{I_i}}|F\boldsymbol{u}|^2dx,\end{aligned}$$

we have

$$\begin{aligned}\widetilde{\int_\Omega}|\boldsymbol{\eta}|^2dx &\le \frac{4c_0^2c_{inv}^2}{h^2}\widetilde{\int_\Omega}|F\boldsymbol{u}|^2dx + \frac{c_0c_{inv}}{h}\sum_{i=1}^N\alpha^2_{i+\frac{1}{2}}|[\boldsymbol{\rho}]|^2_{i+\frac{1}{2}}\\ &\le \frac{4c_0^2c_{inv}^2(\beta^F)^2}{h^2}\widetilde{\int_\Omega}|\boldsymbol{u}|^2dx + \frac{c_0c_{inv}}{h}\sum_{i=1}^N\alpha^2_{i+\frac{1}{2}}|[\boldsymbol{\rho}]|^2_{i+\frac{1}{2}}.\end{aligned}$$

Therefore,

$$\frac{E^{n+1}-E^n}{\tau} \leq -\left(\beta_F - 4c_0^2 c_{inv}^2 (\beta^F)^2 \beta^{D\boldsymbol{\xi}} \frac{\tau}{h^2}\right) \int_\Omega |\boldsymbol{u}|^2 dx \\ -\sum_{i=1}^{N} \left(\frac{\beta_{D\boldsymbol{\xi}}}{2} - c_0 c_{inv} \beta^{D\boldsymbol{\xi}} \alpha_{i+\frac{1}{2}} \frac{\tau}{h}\right) \alpha_{i+\frac{1}{2}} |[\boldsymbol{\rho}]_{i+\frac{1}{2}}|^2.$$

The proof is completed by substituting the time step restriction into the inequality. □

B.2 Proofs of theorems on the preservation of positivity

B.2.1 Proof of weak positivity in Theorem 4.2.2

Proof. Note $\widehat{F\boldsymbol{u}} = \frac{1}{2}\left(\mathrm{diag}(\boldsymbol{\rho}_h^+)\boldsymbol{v}_h^+ + \mathrm{diag}(\boldsymbol{\rho}_h^-)\boldsymbol{v}_h^-\right) + \frac{\alpha}{2}(\boldsymbol{\rho}_h^+ - \boldsymbol{\rho}_h^-)$. Hence

$$(\widehat{\rho v})_l := (\widehat{F\boldsymbol{u}})_l = \frac{1}{2}\left(\rho_{h,l}^+ v_{h,l}^+ + \rho_{h,l}^- v_{h,l}^-\right) + \frac{\alpha}{2}(\rho_{h,l}^+ - \rho_{h,l}^-).$$

Then

$$\frac{(\bar{\rho}_{h,l})_i^{n+1} - (\bar{\rho}_{h,l})_i^n}{\tau} = \frac{(\widehat{F\boldsymbol{u}})_{l,i+\frac{1}{2}}^n - (\widehat{F\boldsymbol{u}})_{l,i-\frac{1}{2}}^n}{h} = \frac{(\widehat{\rho v})_{l,i+\frac{1}{2}}^n - (\widehat{\rho v})_{l,i-\frac{1}{2}}^n}{h},$$

According to the result we have proved for the scalar case in Lemma 3.2.1, we assert $(\bar{\rho}_{h,l})_i^{n+1} \geq 0$. □

B.2.2 Proof of weak positivity in Theorem 4.3.2

Proof.

$$\begin{aligned}\frac{(\bar{\rho}_{h,l})_{i,j}^{n+1,\mathrm{pre}}-(\bar{\rho}_{h,l})_{i,j}^{n}}{\tau} =&\frac{1}{h_i^x h_j^y}\widetilde{\int_{J_j}}(\widehat{F\boldsymbol{u}^x})_{l,i+\frac{1}{2}}-(\widehat{F\boldsymbol{u}^x})_{l,i-\frac{1}{2}}dy\\ &+\frac{1}{h_i^x h_j^y}\widetilde{\int_{I_i}}(\widehat{F\boldsymbol{u}^y})_{l,j+\frac{1}{2}}-(\widehat{F\boldsymbol{u}^y})_{l,j-\frac{1}{2}}dx.\end{aligned}$$

Note that the numerical fluxes can be rewritten as

$$\widehat{F\boldsymbol{u}^x}=\frac{1}{2}\left((\mathrm{diag}(\boldsymbol{\rho}_h)\boldsymbol{v}_h^x)^+ +(\mathrm{diag}(\boldsymbol{\rho}_h^-)\boldsymbol{v}_h^x)^-\right)+\frac{\alpha^x}{2}\mathrm{diag}(\boldsymbol{\rho}_h^+-\boldsymbol{\rho}_h^-),$$

$$\widehat{F\boldsymbol{u}^y}=\frac{1}{2}\left((\mathrm{diag}(\boldsymbol{\rho}_h^+)\boldsymbol{v}_h^y)^+ +(\mathrm{diag}(\boldsymbol{\rho}_h^-)\boldsymbol{v}_h^y)^-\right)+\frac{\alpha^y}{2}\mathrm{diag}(\boldsymbol{\rho}_h^+-\boldsymbol{\rho}_h^-).$$

Let

$$(\widehat{\rho v^x})_l:=(\widehat{F\boldsymbol{u}^x})_l=\frac{1}{2}\left((\rho_{h,l}v_{h,l}^x)^+ +(\rho_{h,l}v_{h,l}^x)^-\right)+\frac{1}{2}\alpha^x(\rho_{h,l}^+-\rho_{h,l}^-),$$

and

$$(\widehat{\rho v^y})_l:=(\widehat{F\boldsymbol{u}^y})_l=\frac{1}{2}\left((\rho_{h,l}v_{h,l}^y)^+ +(\rho_{h,l}v_{h,l}^y)^-\right)+\frac{1}{2}\alpha^y(\rho_{h,l}^+-\rho_{h,l}^-).$$

Then we have

$$\begin{aligned}(\bar{\rho}_{h,l})_{i,j}^{n+1,\mathrm{pre}} =&(\bar{\rho}_{h,l})_{i,j}^{n}+\frac{\tau}{h_i^x h_j^y}\widetilde{\int_{J_j}}(\widehat{\rho v^x})_{l,i+\frac{1}{2}}-(\widehat{\rho v^x})_{l,i-\frac{1}{2}}dy\\ &+\frac{\tau}{h_i^x h_j^y}\widetilde{\int_{I_i}}(\widehat{\rho v^y})_{l,j+\frac{1}{2}}-(\widehat{\rho v^y})_{l,j-\frac{1}{2}}dx,\end{aligned}$$

which formally reduces to the scalar case. One can follow the proof of Theorem 3.20 to show $\bar{\rho}_{h,l}^{n+1,\mathrm{pre}}\geq 0$, if the prescribed time step restriction is satisfied. □

Bibliography

[1] N. B. Abdallah, I. M. Gamba, and G. Toscani. On the minimization problem of sub-linear convex functionals. *Kinetic & Relatated Models*, 4(4):857–871, 2011.

[2] M. L. Adams. Discontinuous finite element transport solutions in thick diffusive problems. *Nuclear Science and Engineering*, 137(3):298–333, 2001.

[3] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[4] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM Journal on Numerical Analysis*, 39(5):1749–1779, 2002.

[5] F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations. *Journal of Computational Physics*, 131(2):267–279, 1997.

[6] D. Benedetto, E. Caglioti, J. A. Carrillo, and M. Pulvirenti. A non-Maxwellian steady distribution for one-dimensional granular media. *Journal of Statistical Physics*, 91(5-6):979–990, 1998.

[7] D. Benedetto, E. Caglioti, and M. Pulvirenti. A kinetic equation for granular media. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(2):439–441, 1999.

[8] M. Bessemoulin-Chatard and F. Filbet. A finite volume scheme for nonlinear degenerate parabolic equations. *SIAM Journal on Scientific Computing*, 34(5):B559–B583, 2012.

[9] A. Blanchet, J. Dolbeault, and B. Perthame. Two-dimensional Keller–Segel model: Optimal critical mass and qualitative properties of the solutions. *Electronic Journal of Differential Equations (EJDE)[electronic only]*, 2006:Paper–No, 2006.

[10] M. Burger, J. A. Carrillo, and M.-T. Wolfram. A mixed finite element method for nonlinear diffusion equations. *Kinetic & Related Models*, 3(1):59–83, 2010.

[11] M. Burger, R. Fetecau, and Y. Huang. Stationary states and asymptotic behavior of aggregation models with nonlinear local repulsion. *SIAM Journal on Applied Dynamical Systems*, 13(1):397–424, 2014.

[12] M. Burger, M. d. Francesco, and M. Franek. Stationary states of quadratic diffusion equations with long-range attraction. *Communications in Mathematical Sciences*, 11(3):709–738, 2013.

[13] M. H. Carpenter, T. C. Fisher, E. J. Nielsen, and S. H. Frankel. Entropy stable spectral collocation schemes for the Navier–Stokes equations: discontinuous interfaces. *SIAM Journal on Scientific Computing*, 36(5):B835–B867, 2014.

[14] J. Carrillo, A. Jüngel, P. Markowich, G. Toscani, and A. Unterreiter. Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities. *Monatshefte für Mathematik*, 133(1):1–82, 2001.

[15] J. A. Carrillo, A. Chertock, and Y. Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17(1):233–258, 2015.

[16] J. A. Carrillo, M. Di Francesco, and G. Toscani. Strict contractivity of the 2-Wasserstein distance for the porous medium equation by mass-centering. *Proceedings of the American Mathematical Society*, 135(2):353–363, 2007.

[17] J. A. Carrillo, Y. Huang, F. Saverio Patacchini, and G. Wolansky. Numerical study of a particle method for gradient flows. *Kinetic & Related Models*, 10(3), 2017.

[18] J. A. Carrillo, R. J. McCann, C. Villani, et al. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Revista Matematica Iberoamericana*, 19(3):971–1018, 2003.

[19] J. A. Carrillo, H. Ranetbauer, and M.-T. Wolfram. Numerical simulation of nonlinear continuity equations by evolving diffeomorphisms. *Journal of Computational Physics*, 327:186–202, 2016.

[20] J. A. Carrillo, J. Rosado, and F. Salvarani. 1d nonlinear Fokker–Planck equations for fermions and bosons. *Applied Mathematics Letters*, 21(2):148–154, 2008.

[21] J. A. Carrillo and G. Toscani. Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity. *Indiana University Mathematics Journal*, pages 113–142, 2000.

[22] T. Chen and C.-W. Shu. Entropy stable high order discontinuous Galerkin methods with suitable quadrature rules for hyperbolic conservation laws. *Journal of Computational Physics*, 345:427–461, 2017.

[23] Y. Cheng and C.-W. Shu. A discontinuous Galerkin finite element method for time dependent partial differential equations with higher order derivatives. *Mathematics of Computation*, 77(262):699–730, 2008.

[24] P. G. Ciarlet. The finite element method for elliptic problems. *Classics in Applied Mathematics*, 40:1–511, 2002.

[25] B. Cockburn. Static condensation, hybridization, and the devising of the HDG methods. In *Building bridges: connections and challenges in modern approaches to numerical partial differential equations*, pages 129–177. Springer, 2016.

[26] 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. *Mathematics of Computation*, 54(190):545–581, 1990.

[27] B. Cockburn, S.-Y. Lin, and C.-W. Shu. TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one-dimensional systems. *Journal of Computational Physics*, 84(1):90–113, 1989.

[28] B. Cockburn and C.-W. Shu. TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework. *Mathematics of Computation*, 52(186):411–435, 1989.

[29] B. Cockburn and C.-W. Shu. The Runge–Kutta local projection P^1-discontinuous-Galerkin finite element method for scalar conservation laws. *ESAIM: Mathematical Modelling and Numerical Analysis*, 25(3):337–361, 1991.

[30] B. Cockburn and C.-W. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM Journal on Numerical Analysis*, 35(6):2440–2463, 1998.

[31] B. Cockburn and C.-W. Shu. The Runge–Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems. *Journal of Computational Physics*, 141(2):199–224, 1998.

[32] B. Cockburn and C.-W. Shu. Runge–Kutta discontinuous Galerkin methods for convection-dominated problems. *Journal of Scientific Computing*, 16(3):173–261, 2001.

[33] K. Craig and A. Bertozzi. A blob method for the aggregation equation. *Mathematics of computation*, 85(300):1681–1717, 2016.

[34] Q. Du, C. Liu, and P. Yu. FENE dumbbell model and its several linear and nonlinear closure approximations. *Multiscale Modeling & Simulation*, 4(3):709–731, 2005.

[35] J. Escher, M. Hillairet, P. Laurencot, and C. Walker. Global weak solutions for a degenerate parabolic system modeling the spreading of insoluble surfactant. *Indiana University Mathematics Journal*, pages 1975–2019, 2011.

[36] G. J. Gassner. A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP–SAT finite difference methods. *SIAM Journal on Scientific Computing*, 35(3):A1233–A1253, 2013.

[37] G. J. Gassner, A. R. Winters, and D. A. Kopriva. A well balanced and entropy conservative discontinuous Galerkin spectral element method for the shallow water equations. *Applied Mathematics and Computation*, 272:291–308, 2016.

[38] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Review*, 43(1):89–112, 2001.

[39] J.-L. Guermond and G. Kanschat. Asymptotic analysis of upwind discontinuous Galerkin approximation of the radiative transport equation in the diffusive limit. *SIAM Journal on Numerical Analysis*, 48(1):53–78, 2010.

[40] W. Guo, J.-M. Qiu, and J. Qiu. A new Lax–Wendroff discontinuous Galerkin method with superconvergence. *Journal of Scientific Computing*, 65(1):299–326, 2015.

[41] E. Hairer and G. Wanner. Solving ordinary differential equations II. *New York*, 1991.

[42] J. S. Hesthaven, S. Gottlieb, and D. Gottlieb. *Spectral methods for time-dependent problems*, volume 21. Cambridge University Press, 2007.

[43] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods: algorithms, analysis, and applications*. Springer Science & Business Media, 2007.

[44] J. Huang and C.-W. Shu. Error estimates to smooth solutions of semi-discrete discontinuous Galerkin methods with quadrature rules for scalar conservation laws. *Numerical Methods for Partial Differential Equations*, 33(2):467–488, 2017.

[45] A. Iserles. *A first course in the numerical analysis of differential equations*. Number 44. Cambridge University Press, 2009.

[46] T. L. Jackson and H. M. Byrne. A mechanical model of tumor encapsulation and transcapsular spread. *Mathematical Biosciences*, 180(1-2):307–328, 2002.

[47] O. Jensen and J. Grotberg. Insoluble surfactant spreading on a thin viscous film: shock evolution and film rupture. *Journal of Fluid Mechanics*, 240:259–288, 1992.

[48] G. S. Jiang and C.-W. Shu. On a cell entropy inequality for discontinuous Galerkin methods. *Mathematics of Computation*, 62(206):531–538, 1994.

[49] A. Jüngel. The boundedness-by-entropy method for cross-diffusion systems. *Nonlinearity*, 28(6):1963, 2015.

[50] A. Jüngel and I. V. Stelzer. Entropy structure of a cross-diffusion tumor-growth model. *Mathematical Models and Methods in Applied Sciences*, 22(07):1250009, 2012.

[51] E. F. Keller and L. A. Segel. Initiation of slime mold aggregation viewed as an instability. *Journal of Theoretical Biology*, 26(3):399–415, 1970.

[52] J. F. B. M. Kraaijevanger. Contractivity of Runge–Kutta methods. *BIT Numerical Mathematics*, 31(3):482–528, 1991.

[53] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical aspects of finite elements in partial differential equations*, pages 89–123. Elsevier, 1974.

[54] D. Levy and E. Tadmor. From semidiscrete to fully discrete: stability of Runge–Kutta schemes by the energy method. *SIAM Review*, 40(1):40–73, 1998.

[55] C. Liu and H. Liu. Boundary conditions for the microscopic FENE models. *SIAM Journal on Applied Mathematics*, 68(5):1304–1315, 2008.

[56] H. Liu and Z. Wang. An entropy satisfying discontinuous Galerkin method for nonlinear Fokker–Planck equations. *Journal of Scientific Computing*, 68(3):1217–1240, 2016.

[57] H. Liu and Z. Wang. A free energy satisfying discontinuous Galerkin method for one-dimensional Poisson–Nernst–Planck systems. *Journal of Computational Physics*, 328:413–437, 2017.

[58] H. Liu and J. Yan. The direct discontinuous Galerkin (DDG) methods for diffusion problems. *SIAM Journal on Numerical Analysis*, 47(1):675–698, 2009.

[59] H. Liu and H. Yu. Maximum–principle-satisfying third order discontinuous Galerkin schemes for Fokker–Planck equations. *SIAM Journal on Scientific Computing*, 36(5):A2296–A2325, 2014.

[60] J. Luo, C.-W. Shu, and Q. Zhang. A priori error estimates to smooth solutions of the third order Runge–Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(4):991–1018, 2015.

[61] A. Mogilner and L. Edelstein-Keshet. A non-local model for a swarm. *Journal of Mathematical Biology*, 38(6):534–570, 1999.

[62] A. A. H. Oulhaj. A finite volume scheme for a seawater intrusion model with cross-diffusion. In C. Cancès and P. Omnes, editors, *Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects*, pages 421–429, Cham, 2017. Springer International Publishing.

[63] J. Qiu, M. Dumbser, and C.-W. Shu. The discontinuous Galerkin method with lax–wendroff type time discretizations. *Computer Methods in Applied Mechanics and Engineering*, 194(42-44):4528–4543, 2005.

[64] W. H. Reed and T. Hill. Triangular mesh methods for the neutron transport equation. Technical report, Los Alamos Scientific Lab., N. Mex.(USA), 1973.

[65] J. Shen, T. Tang, and L.-L. Wang. *Spectral methods: algorithms, analysis and applications*, volume 41. Springer Science & Business Media, 2011.

[66] J. Shen and H. Yu. On the approximation of the Fokker–Planck equation of the finitely extensible nonlinear elastic dumbbell model I: a new weighted formulation and an optimal spectral–Galerkin algorithm in two dimensions. *SIAM Journal on Numerical Analysis*, 50(3):1136–1161, 2012.

[67] N. Shigesada, K. Kawasaki, and E. Teramoto. Spatial segregation of interacting species. *Journal of Theoretical Biology*, 79(1):83–99, 1979.

[68] M. Spijker. Contractivity in the numerical solution of initial value problems. *Numerische Mathematik*, 42(3):271–290, 1983.

[69] S. Srinivasana, J. Poggiea, and X. Zhang. A positivity-preserving high order discontinuous Galerkin scheme for convection-diffusion equations. *Journal of Computational Physics*, to appear.

[70] Z. Sun, J. A. Carrillo, and C.-W. Shu. A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials. *Journal of Computational Physics*, 352:76–104, 2018.

[71] Z. Sun and C.-W. Shu. Stability analysis and error estimates of Lax–Wendroff discontinuous Galerkin methods for linear conservation laws. *ESAIM: Mathematical Modelling and Numerical Analysis*, 51(3):1063–1087, 2017.

[72] Z. Sun and C.-W. Shu. Stability of the fourth order Runge–Kutta method for time-dependent partial differential equations. *Annals of Mathematical Sciences and Applications*, 2(2):255–284, 2017.

[73] E. Tadmor. From semidiscrete to fully discrete: Stability of Runge–Kutta schemes by the energy method. II. *Collected Lectures on the Preservation of Stability under Discretization*, 109:25, 2002.

[74] C. M. Topaz, A. L. Bertozzi, and M. A. Lewis. A nonlocal continuum model for biological aggregation. *Bulletin of Mathematical Biology*, 68(7):1601, 2006.

[75] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[76] Y. Xia, Y. Xu, and C.-W. Shu. Local discontinuous Galerkin methods for the Cahn–Hilliard type equations. *Journal of Computational Physics*, 227(1):472–491, 2007.

[77] Y. Xing and C.-W. Shu. High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms. *Journal of Computational Physics*, 214(2):567–598, 2006.

[78] Y. Xing and C.-W. Shu. A new approach of high order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms. *Communications of Computational Physics*, 1(1):100–134, 2006.

[79] Y. Xu and C.-W. Shu. Local discontinuous Galerkin methods for high-order time-dependent partial differential equations. *Communications in Computational Physics*, 7(1):1, 2010.

[80] J. Yan and C.-W. Shu. A local discontinuous Galerkin method for KdV type equations. *SIAM Journal on Numerical Analysis*, 40(2):769–791, 2002.

[81] Q. Zhang and F. Gao. A fully-discrete local discontinuous Galerkin method for convection-dominated Sobolev equation. *Journal of Scientific Computing*, 51(1):107–134, 2012.

[82] Q. Zhang and C.-W. Shu. Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws. *SIAM Journal on Numerical Analysis*, 42(2):641–666, 2004.

[83] Q. Zhang and C.-W. Shu. Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws. *SIAM Journal on Numerical Analysis*, 44(4):1703–1720, 2006.

[84] Q. Zhang and C.-W. Shu. Stability analysis and a priori error estimates of the third order explicit Runge–Kutta discontinuous Galerkin method for scalar conservation laws. *Brown University Scientific Computing Report*, 2009(28), 2009.

[85] Q. Zhang and C.-W. Shu. Stability analysis and a priori error estimates of the third order explicit Runge–Kutta discontinuous Galerkin method for scalar conservation laws. *SIAM Journal on Numerical Analysis*, 48(3):1038–1063, 2010.

[86] X. Zhang. On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier–Stokes equations. *Journal of Computational Physics*, 328:301–343, 2017.

[87] X. Zhang and C.-W. Shu. On maximum-principle-satisfying high order schemes for scalar conservation laws. *Journal of Computational Physics*, 229(9):3091–3120, 2010.

[88] X. Zhang and C.-W. Shu. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 467, pages 2752–2776. The Royal Society, 2011.

[89] Y. Zhang, X. Zhang, and C.-W. Shu. Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection–diffusion equations on triangular meshes. *Journal of Computational Physics*, 234:295–316, 2013.