High Order Numerical Methods for Hyperbolic Equations: Bound-preserving and Riemann Invariant Based System Solvers

by

Ziyao Xu

B.Eng., China University of Petroleum (East China), Shandong, P.R. China, 2016

M.Sc., Michigan Technological University, MI, U.S.A., 2019

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 2023

This dissertation by Ziyao Xu 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_____ _____

Mark Ainsworth, Ph.D., Reader

Date_____ _____

Johnny Guzmán, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Thomas A. Lewis, Dean of the Graduate School

Vita

Education

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

Ph.D. in Applied Mathematics (expected in May 2023).
Advisor: Chi-Wang Shu.

Michigan Technological University, Houghton, MI, U.S.A..

M.Sc. in Mathematical Sciences, 2019.
Advisor: Yang Yang.

China University of Petroleum (East China), Qingdao, Shandong, P.R. China.

B.Eng. in Petroleum Engineering, 2016.
Advisor: Yajun Li.

Publications and Preprints

1. Z. Xu and Y. Yang. The hybrid-dimensional Darcy's law: A non-conforming reinterpreted discrete fracture model (RDFM) for the compressible miscible displacement and multicomponent gas flow in fractured media. submitted.
2. X. Wu, H. Guo, Z. Xu and Y. Yang. A reinterpreted discrete fracture model for Darcy-Forchheimer flow in fractured porous media. submitted.

3. Z. Xu and C.-W. Shu. Local characteristic decomposition free high order finite difference WENO schemes for hyperbolic systems endowed with a coordinate system of Riemann invariants. submitted.

4. Z. Xu and C.-W. Shu. On the conservation property of positivity-preserving discontinuous Galerkin methods for stationary hyperbolic equations. submitted.

5. Z. Xu, Z. Huang and Y. Yang. The hybrid-dimensional Darcy's law: A non-conforming reinterpreted discrete fracture model (RDFM) for single-phase flow in fractured media. *Journal of Computational Physics*, v473 (2023), 111749.

6. Z. Xu and C.-W. Shu. Third order maximum-principle-satisfying and positivity-preserving Lax-Wendroff discontinuous Galerkin methods for hyperbolic conservation laws. *Journal of Computational Physics*, v470 (2022), 111591.

7. Z. Xu and C.-W. Shu. High order conservative positivity-preserving discontinuous Galerkin method for stationary hyperbolic equations. *Journal of Computational Physics*, v466 (2022), 111410.

8. H. Guo, W. Feng, Z. Xu and Y. Yang. Conservative numerical methods for the reinterpreted discrete fracture model on non-conforming meshes and their applications in contaminant transportation in fractured porous media. *Advances in Water Resources*, v153 (2021), 103951.

9. Z. Xu and Y. Yang. The hybrid dimensional representation of permeability tensor: A reinterpretation of the discrete fracture model and its extension on nonconforming meshes. *Journal of Computational Physics*, v415 (2020), 109523.

10. Z. Xu, Y. Yang and H. Guo. High-order bound-preserving discontinuous Galerkin methods for wormhole propagation on triangular meshes. *Journal of Computational Physics*, v390 (2019), pp.323-341.

11. N. Chuenjarern, Z. Xu and Y. Yang. High-order bound-preserving discontinuous Galerkin methods for compressible miscible displacements in porous media on triangular meshes. *Journal of Computational Physics*, v378 (2019), pp.110-128.

12. H. Guo, L. Tian, Z. Xu, Y. Yang and N. Qi. High-order local discontinuous Galerkin method for simulating wormhole propagation. *Journal of Computational and Applied Mathematics*, v350 (2019), pp.247-261.

Teaching Experience

Brown University

Applied Ordinary Differential Equations, Spring 2021.
Introduction to Computational Linear Algebra, Fall 2020.

Michigan Technological University

Calculus 2, Spring 2019.
Calculus for Life Sciences, Fall 2018.
Calculus 2, Spring 2018.

Awards & Honors

Outstanding Research Award, Department of Mathematical Sciences, Michigan Technological University, 2019.
First Prize Scholarship, China University of Petroleum (East China), 2015.
China National Scholarship, 2013 and 2014.

Acknowledgments

First and foremost, I would like to thank my advisor, Prof. Chi-Wang Shu, for his exceptional mentorship throughout my Ph.D. journey. His constant encouragement and invaluable guidance in the face of challenges have been truly remarkable. He is incredibly supportive of my academic growth. I am truly fortunate to have the opportunity to study with him.

I am also grateful to my dissertation readers, Prof. Mark Ainsworth and Prof. Johnny Guzmán, for their invaluable feedback and insightful suggestions. Their expertise has significantly contributed to the quality of this dissertation.

I would like to extend my appreciation to the staff of the division, including Candida Hall, Jean Radican, Stephanie Han, Rosanna Wertheimer, and others, for their consistent assistance and support.

Throughout my four years at Brown, I have been fortunate to have many incredible friends. I would like to thank my friends, including Zongren Zou, Xiaoyu Xie, Enrui Zhang, Tianmin Yu, Moyi Tian, Zhen Zhang, Mengjie Liu, Sining Gong, Sicheng Liu, Hanye Zhu, Qian Zhang, and many others, for the memorable joyous moments and inspiring discussions on mathematics we have had. I also wish to thank the graduates of the division, Zheng Sun, Xinyue Yu, Tingwei Meng, and Kun Meng, who generously offered help in various aspects.

Lastly, I would like to convey my heartfelt appreciation to my family. To my father, Changqi Xu, whose memory will forever live in my heart; my mother, Haiping Li, for her unconditional love and unwavering support; my young sister, Zixi Xu, for the happiness she brings to our family; and my loving wife, Wenjing Liu, for her constant encouragement, companionship, and devotion.

Abstract of "High Order Numerical Methods for Hyperbolic Equations: Bound-preserving and Riemann Invariant Based System Solvers", by Ziyao Xu, Ph.D., Brown University, May 2023

This dissertation consists of three topics on bound-preserving discontinuous Galerkin (DG) methods for time-dependent and stationary hyperbolic equations, and efficient finite difference weighted essentially non-oscillatory (WENO) schemes for hyperbolic systems.

In Chapter 2, we propose third order bound-preserving DG schemes for scalar conservation laws and the Euler equations based on the Lax-Wendroff time discretization. We first establish the maximum-principle-satisfying DG scheme for scalar conservation laws in one dimension. The scheme develops the idea from the direct discontinuous Galerkin (DDG) method for heat equations to discretize high order spatial derivatives resulting from the Lax-Wendroff procedure. When it extends to multi-dimensions, we avoid the appearance of mixed derivatives in the numerical schemes based on carefully designed expansions of high order derivatives. The positivity-preserving schemes for the Euler equations are constructed in a similar manner.

In Chapter 3, we follow the work of Yuan et al. (2016) [90] and Ling et al. (2018) [46] to investigate the positivity-preserving DG methods for stationary hyperbolic equations. High order conservative positivity-preserving DG methods for variable coefficient and nonlinear stationary hyperbolic equations in one dimension, and constant coefficient stationary hyperbolic equations in two and three dimensions are constructed, via suitable quadratures. In Chapter 4, we continue the study in Chapter 3 and clarify a more appropriate definition of mass conservation, rather than preserving cell averages, for stationary hyperbolic equations. The genuinely conservative high-order positivity-preserving DG methods based on the new definition are

constructed, which are able to preserve the positivity of more general types of equations with much simpler implementations and easier proofs for the Lax-Wendroff theorem. Novel conservative positivity-preserving limiters are designed to accommodate for the new definition of conservation.

In Chapter 5, we investigate local characteristic decomposition free WENO schemes for a special class of hyperbolic systems endowed with a coordinate system of Riemann invariants. We apply the WENO procedure to the coordinate system of Riemann invariants instead of the local characteristic variables to save the expensive computational cost on local characteristic decomposition but meanwhile, maintain the essentially non-oscillatory performance. The efficiency and good performance of our method are demonstrated by extensive numerical tests, which indicate the coordinate system of Riemann invariants is a good alternative to local characteristic variables for the WENO procedure with higher efficiency.

Contents

List of Tables

List of Figures

Chapter One

Introduction

Hyperbolic conservation/balance laws are fundamental tools to investigate the phenomena of flow and transport. When numerically solving hyperbolic equations, high order accurate numerical methods are preferred due to their vast advantages in efficiency and fidelity. However, many high order methods suffer from strong oscillations near discontinuities of solutions, which is the so-called Gibbs phenomenon. Some important physical quantities may be out of their physical range due to spurious oscillations, e.g. density or pressure of fluid becomes negative if the density or pressure is low near shocks. Once these happen, not only physically the quantities are no longer meaningful, but numerically the hyperbolicity of differential equations is also changed, which often leads to the NaN outputs and simulation failure. To reduce the spurious oscillations, many remarkable robust algorithms and stabilizing techniques were developed, e.g. artificial viscosity methods [79], total variation diminishing (TVD) [31, 32] and total variation bounded (TVB) [69] schemes, essentially non-oscillatory (ENO) [33] and weighted essentially non-oscillatory (WENO) [48, 35] schemes, and discontinuous Galerkin (DG) [62, 40] methods with various of limiters, etc.

The discontinuous Galerkin method, first proposed by Reed and Hill in [62] for the neutron transport equation on triangular meshes, is a class of finite element methods with discontinuous function space of test and trial functions. It was developed into the Runge-Kutta discontinuous Galerkin (RKDG) methods by Cockburn et al. in a series of work [17, 16, 14, 13, 18] to solve nonlinear hyperbolic problems. Since then, the DG methods have been widely used in numerical simulation for hyperbolic equations, due to their advantages in high order accuracy, flexibility in complex geometry, and easiness to be parallelized. Limiters such as the total variation bounded (TVB) limiter [18] are usually applied after each Runge-Kutta stage to stabilize the solution.

Due to the lack of provable bound-preserving property, common DG methods designed for general purpose are not enough in some extreme cases, e.g. very high Mach jets in astrophysics, point blast problems, and vacuum near shocks in compressible gas dynamics. The simulation may blow up if no ad hoc positivity correction is made to avoid possible negative density or pressure. Bound-preserving methods, also referred to as the maximum-principle preserving methods in some contexts, are numerical methods that are carefully designed to ensure the physical quantities strictly fall into their physical bounds during the entire simulation. If the upper bound of a physical quantity is infinity, the methods are called positivity-preserving. In 2010, the genuinely high order bound-preserving DG methods were constructed by Zhang and Shu in [93, 94]. The basic idea of the Zhang-Shu framework is to take the test function in the DG scheme to be 1 in each cell to yield an equation satisfied by the cell average of the target variable, and prove the desired boundedness of the cell average under the certain choice of numerical fluxes and suitable CFL conditions. Then a scaling limiter that does not affect the accuracy and cell average can be used to modify the variable to obtain physically relevant bounds for the entire solution.

In order to gain high order accuracy, the bound-preserving schemes also need to combine with temporal discretization whose order is consistent with the order of spatial discretization. Almost all temporal discretizations in the existing bound-preserving methods is based on the method of lines, which treats the spatially discretized equations as ODE systems and uses appropriate time marching approaches to evolve in time. In particular, the strong stability preserving (SSP) Runge-Kutta or multi-step methods, which are convex combinations of Euler-forward time discretization, are primary time marching approaches used in bound-preserving DG methods for time-dependent hyperbolic problems. In this dissertation, we study the bound-preserving DG methods based on the Lax-Wendroff time discretization, which

haven't been studied adequately before. The Lax-Wendroff discontinuous Galerkin (LWDG) methods [58, 30] first take the Taylor expansion of the solution in time to gain high order temporal accuracy. Then, the methods utilize PDE itself to replace the temporal derivatives in the expansion with spacial derivatives, followed by spatial discretization by the DG methods. Therefore, the LWDG is a high order single-stage method, with more compact stencils in each time step. The main challenge of the bound-preserving technique for LWDG is the appearance of high order derivatives and mixed derivatives in high dimensions due to the Lax-Wendroff procedure, which is not commonly encountered when dealing with first order hyperbolic equations or convection-diffusion equations with the method of lines. To resolve the problem, the bound-preserving DDG [11] flux will play a key role in the construction of the numerical schemes.

Besides the LWDG for time-dependent problems, we also study the positivity-preserving DG for stationary hyperbolic equations in this dissertation. The stationary hyperbolic equations could appear in flow and transport problems at steady-state, the stationary radiative transfer equations (RTE) discretized by the discrete-ordinate method (DOM), or the implicit time discretization/space-time DG methods for time-dependent hyperbolic equations. A distinct difference between the DG methods for the stationary and time-dependent hyperbolic equations is the marching direction of the computation. For time-dependent problems, the DG methods march in time, so the cell averages of solution at the current time stage depend on the cell averages in previous stages, thereby the bounds of cell averages are easier to be ensured, provided the bounds are satisfied by the solutions at previous times. However, in stationary problems, the sweeping direction of the computation follows the wind direction of the flux in space. Therefore, the cell average of the solution on a target cell depends only on the inflow fluxes and the source term, which gives rise to new challenges to

the bound-preserving technique. Such a feature makes the proof of the positivity of cell averages very complicated even for the simple equation $u_x + \gamma u = s(x)$, see [46]. As shown in the studies in Chapter 3 and [46], for more complex equations or higher dimensions, the positivity of cell averages obtained by unmodulated DG methods generally fails! Some complicated non-conservative positivity-preserving limiters are proposed in [90, 91] to address such an issue. In this dissertation, we shall propose new techniques and methodologies to preserve the positivity of the solution without hurting the conservation property of DG schemes.

Besides DG, the ENO and WENO methods are also prevalent numerical methods for hyperbolic conservation laws. The ENO methods, first developed by Harten et al. [33], use an adaptive strategy to choose the smoothest stencil among several candidates to reconstruct the solution from its cell averages, hence the methods yield essentially non-oscillatory approximation near shocks. The original ENO scheme was based on the framework of finite volume methods with Lax-Wendroff time discretization. Later on, Shu and Osher proposed the finite difference framework in [73] by ENO interpolation for nodal values of solutions and high order approximation for spatial derivatives of fluxes, which saves considerable computational cost in multi-dimensions. Their subsequent work in [74] developed a simpler finite difference framework based on the Shu-Osher lemma to approximate the fluxes at cell interfaces by standard reconstruction procedure for fluxes at cell centers. The WENO methods [48, 35] were developed upon ENO, with the idea of using a convex combination of all candidate stencils rather than only one stencil in the original ENO scheme to gain higher order of accuracy. The WENO procedure can be used in the framework of finite volume or finite difference methods like the ENO.

The high order ENO/WENO finite volume/difference methods have to be used in cooperation with the local characteristic decomposition, when solving hyperbolic

systems of conservation laws, as the reconstruction/interpolation on conservative variables typically produces much worse numerical fidelity near shocks compared with the characteristic variables, especially when shocks of different characteristic fields interact. However, the local characteristic decomposition is computationally costly since the eigendecomposition is needed on each interface of cells. In this dissertation, we look for alternative variables for the WENO procedure, such that the expansive computation of local characteristic decomposition is exempted, meanwhile the non-oscillatory fashion of the WENO schemes is not affected. An ideal candidate is the coordinate system of Riemann invariants, which only admits one major discontinuity in each component when shocks of different characteristic fields appear in Riemann problems. Due to the nonlinear algebraic relation between the Riemann invariants and conservative variables, we establish our local characteristic free scheme within the finite difference framework of [73].

The rest of the dissertation is organized as follows. In Chapter 2, we study the bound-preserving LWDG methods for scalar conservation laws and the Euler equations. In Chapter 3, we construct positivity-preserving DG methods for stationary hyperbolic balance laws with nonnegative source terms and initial conditions, via a suitable quadrature rule. We further study this topic in Chapter 4, where a more appropriate notion of mass conservation for stationary hyperbolic equations is clarified and the corresponding conservative limiters are given. In Chapter 5, we propose an efficient local characteristic decomposition free finite difference WENO method for a particular class of hyperbolic conservation systems. Finally, we summarize the dissertation with some conclusions in Chapter 6.

To avoid confusion, we remark that we define and use notations locally in each chapter. Some notations may have different meanings in different chapters.

Chapter Two

Bound-preserving Lax-Wendroff discontinuous Galerkin methods for time-dependent hyperbolic equations

2.1 Introduction

Hyperbolic conservation laws are basic tools to characterize the phenomena of flow and transport, e.g. the Burgers' equation for traffic flow and the Buckley-Leverett equation for two phase flow as the scalar cases, and the Euler equations for compressible gas dynamics and shallow water equations for water with shallow depth as the system cases.

The scalar conservation laws are known to satisfy the maximum-principle, e.g. for the one dimensional scalar equation

$$u_t + f(u)_x = 0, \quad x \in \mathbb{R},\ t > 0, \tag{2.1}$$

with initial condition

$$u(x, 0) = u_0(x), \quad x \in \mathbb{R},$$

the entropy solution satisfies $m \leq u(x,t) \leq M, \forall x \in \mathbb{R}, t > 0$, where $m = \min_{x\in\mathbb{R}} u_0(x)$ and $M = \max_{x\in\mathbb{R}} u_0(x)$. Same results also hold for periodic boundary conditions, bounded domain with compactly supported solution, and higher dimensions.

Similarly, the positivity of certain important physical quantities are satisfied by some hyperbolic systems, e.g. for the Euler equations

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{0}, \quad x \in \mathbb{R}, t > 0 \tag{2.2}$$

where

$$\mathbf{u} = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix}, \qquad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} m \\ \rho u^2 + p \\ (E+p)u \end{pmatrix},$$

with

$$m = \rho u, \quad E = \frac{1}{2}\rho u^2 + \rho e, \quad p = (\gamma - 1)\rho e,$$

in which ρ is the density of fluid, m is the momentum, u is the velocity, E is the total energy, p is the pressure, e is the specific internal energy, and $\gamma > 1$ is the ratio of specific heats, it is well-known that the physical solution $\mathbf{u} \in G$ for all $t > 0$ if it holds at $t = 0$, where G is the admissible set of solutions defined as

$$G = \{\mathbf{u} : \rho \geq 0, p(\mathbf{u}) \geq 0\}. \tag{2.3}$$

Rigorously preserving these physical bounds of solutions is of great importance for the robustness of numerical algorithms, in that once the quantities were out of their physical range, the hyperbolicity of equations is lost, which often leads to the simulation failure. There have been intensive studies on the maximum-principle-satisfying and positivity-preserving numerical methods for hyperbolic conservation laws. In 2010, the genuinely maximum-principle satisfying high-order discontinuous Galerkin (DG) and finite volume methods for scalar conservation laws were proposed by Zhang and Shu in [93]. The algorithm is composed of two steps under the DG framework. The first step is to prove desired physical bounds for the cell averages of numerical solutions are automatically satisfied by the unmodulated high order DG scheme with appropriate CFL conditions and numerical fluxes. Then a scaling limiter, which does not destroy accuracy and mass conservation, are adopted to modify the solution such that the physical bounds satisfied by cell averages are extended

to the entire solution. Based on this simple and general framework, the high order maximum-principle-satisfying and positivity-preserving numerical schemes have been rapidly developed for different problems ever since, for instance, the Euler equations [94, 95, 80], the Navier-Stokes equations [92], the shallow water equations [82, 81, 45], convection diffusion equations [96, 42, 11], and hyperbolic equations involving $\delta-$singularities [97, 89], etc. For convenience, we call both maximum-principle-satisfying and positivity-preserving techniques the bound-preserving methods.

It should be noted that, in order to gain high order accuracy, the bound-preserving schemes also need to combine with temporal discretization whose order is consistent with the order of spatial discretization. Almost all time discretizations in the aforementioned bound-preserving methods are based on the method of lines, which treats the spatially discretized equation as ODE systems and use appropriate time marching approaches to evolve in time. In particular, the strong stability preserving Runge-Kutta (SSP-RK) methods or the SSP multi-step methods [25, 27, 70] are preferable because they are convex combinations of forward Euler time discretization, which greatly simplifies the proof of the bound-preserving since all analysis only need to be carried out on a single forward Euler time step. Besides the explicit methods, there are also studies on backward Euler time discretization [55, 46].

As an alternative to method of lines, the Lax-Wendroff methods are also widely used in the computation of time-dependent partial differential equations, for instance, the combination of Lax-Wendroff type time discretization with DG (LWDG) methods [58, 57, 30] or with the WENO schemes [60, 56], the two-stage fourth-order methods [53, 43], the arbitrary high order derivative Riemann problem (ADER) approach [77, 23, 22], and its variant based on the Galerkin space-time predictor [20, 7, 21], etc. The Lax-Wendroff methods utilizes the information of the partial differential equations

to replace temporal derivatives by spatial derivatives in the Taylor expansion of the solution in time. Therefore, the Lax-Wendroff methods are one-stage, explicit, high order methods, and only need the stabilizing scaling limiters once per time step.

Regarding to the situation that there are very limited researches on bound-preserving techniques for Lax-Wendroff schemes, we study the LWDG to construct third order maximum-principle-satisfying and positivity-preserving LWDG schemes for scalar conservation laws and the Euler equations in one and two space dimensions. Different to the previous works [50, 66] on positivity-preserving Lax-Wendroff type methods, our algorithm does not rely on the flux limiter that needs to combine low order positivity-preserving flux and high order flux together, hence the high order accuracy of our approach is easier to guarantee.

The construction of our numerical schemes is based on the third order Taylor expansion of solution in time

$$u(x,t^{n+1}) = u(x,t^n) + \Delta t u_t(x,t^n) + \frac{\Delta t^2}{2} u_{tt}(x,t^n) + \frac{\Delta t^3}{6} u_{ttt}(x,t^n) + O(\Delta t^4), \quad (2.4)$$

where $\Delta t = t^{n+1} - t^n$. Due to the Lax-wendroff procedure, there will be many spatial derivatives to replace the original time derivatives in (2.4), especially for the system case in high dimensions. In this chapter, we adopt the discontinuous Galerkin methods for the spatial discretization of the derivatives. In 1973, Reed and Hill [62] proposed the first discontinuous Galerkin method to solve the steady linear transport problem. It was developed into Runge-Kutta discontinuous galerkin methods (RKDG) by Cockburn et al. in a series papers [17, 16, 14, 13, 18] to solve nonlinear hyperbolic conservation laws. Limiters such as the total variation bounded (TVB) limiter [18] are usually applied to stabilize the solution near shocks after each Runge-Kutta stage. Discontinuous Galerkin methods have been widely

used in computational fluid dynamics due to their advantages in high order accuracy, flexibility in complex geometry and easiness to be parallelized, and is one of the most common choices in developing bound-preserving schemes.

In our work, we develop the idea of bound-preserving direct discontinuous Galerkin (DDG) method from [11] to resolve the difficulty caused by high order spatial derivatives produced by the Lax-Wendroff procedure. When it extends to multi-dimensions, we avoid the appearance of mixed derivatives in our numerical schemes based on carefully designed expansions of high order temporal derivatives in the Lax-Wendroff procedure, which is the key for the success of bound-preserving in high dimensions. We only demonstrate the treatments in two dimensions but the technique can be generalized into three dimensions directly.

It is worth mentioning that, the tedious CFL conditions to be derived for bound-preserving in the chapter is not explicitly used in the implementation. But rather, they are used as a theoretical guarantee. In practice, one can use standard CFL conditions in computation, and rewind the computation back to the beginning of the step with halved time step-size when the cell averages exceeds their desired bounds at that step. The theoretical results in the chapter guarantee that one only needs to halve the step-size finite number of times. Moreover, since the LWDG is an explicit single stage method, the temporal derivatives of the solution only need to be computed once per time step, which makes the cost of rewinding computation very cheap.

The rest of the chapter is organized as follows. In Section 2.2, we first introduce the notations to be used throughout the chapter, and then construct the maximum-principle-satisfying LWDG methods for scalar conservation laws in one and two space dimensions. In Section 2.3, we establish the positivity-preserving LWDG schemes

for the Euler equations in one and two dimensional spaces. The scaling limiters are introduced in Section 2.4 to ensure the boundedness and stability of the numerical solution. In Section 2.5, we give extensive numerical examples to demonstrate the effectiveness of our algorithm. We end up with some concluding remarks in Section 2.6. The discussion in the above sections are based on uniform meshes. In the Appendix A.3, we give illustrations on how to extend the algorithms to nonuniform meshes and take the one dimensional scalar conservation law as an example.

2.2 Maximum-principle-preserving for scalar conservation laws

In this section, we study the maximum-principle-satisfying LWDG methods for scalar conservation laws. Based on the framework of [93], we only need to put our effort on attaining the maximum-principle for cell averages of the solution, i.e. $m \leq \bar{u}^{n+1} \leq M$, provided $m \leq u^n \leq M$, where the superscripts n and $n+1$ denote the time level t^n and t^{n+1}, respectively. The slope limiters introduced in Section 2.4 will make up the gap between the maximum-principles of $\bar{u}^{n+1}$ and u^{n+1}.

For simplicity, we only discuss the one and two dimensional problems with periodic boundary conditions on uniform meshes, but the results can be directly extended to three space dimensions and non-periodic cases. However, the extension from uniform meshes to nonuniform meshes is not trivial, which will be demonstrated in the Appendix A.3 with one dimensional space as an example.

We first introduce the notations to be used throughout the chapter, then construct and prove the maximum-principle-satisfying LWDG schemes.

2.2.1 Notations

In the one dimensional space, we assume the domain $\Omega = [a, b]$ is discretized by $a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots x_{N+\frac{1}{2}} = b$, and denote by $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ the cells on Ω for $j = 1, 2, \ldots N$. Moreover, we denote the length and center of the cell I_j by $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ and $x_j = \frac{1}{2}\left(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}}\right)$, respectively, and let $u_j = u(x_j)$

Similarly, in the two dimensional space, we assume $\Omega = [a, b] \times [c, d]$ is discretized by $a = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N_x+\frac{1}{2}} = b$ and $c = y_{\frac{1}{2}} < y_{\frac{3}{2}} < \cdots < y_{N_y+\frac{1}{2}} = d$ in the x and y directions, respectively. We denote by $K_{i,j} = I_i \times J_j = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ the cells in Ω for $i = 1, \ldots, N_x, j = 1, \ldots, N_y$, and by $\Delta x_i \Delta y_j = (x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}})(y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}})$, $(x_i, y_j) = (\frac{1}{2}(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}}), \frac{1}{2}(y_{j-\frac{1}{2}} + y_{j+\frac{1}{2}}))$ the area and center of the cell $K_{i,j}$, respectively, and let $u_{i,j} = u(x_i, y_j)$.

We only consider the uniform meshes in this section and the next section to simplify the discussion, i.e. $\Delta x_i \equiv \Delta x$ and $\Delta y_j \equiv \Delta y$, for $i = 1, \ldots, N_x, j = 1, \ldots, N_y$. The case of nonuniform meshes will be discussed in the appendices.

The finite element spaces in the DG schemes are taken as $V = \{v \in L^2 : v|_{I_j} \in P^2(I_j),\ j = 1, 2, \ldots, N\}$ and $W = \{v \in L^2 : v|_{K_{i,j}} \in Q^2(K_{i,j}),\ i = 1, \ldots, N_x, j = 1, \ldots, N_y\}$ in one and two dimensional spaces, respectively, where $P^2(I)$ is the space of quadratic polynomials on interval I and $Q^2(K)$ is the tensor product space of quadratic polynomials on rectangle K.

Due to discontinuities, functions in the schemes may have double values on cell interfaces. In one space dimension, we denote by $v^-_{j+\frac{1}{2}}$ and $v^+_{j+\frac{1}{2}}$ the left and right limits of v at $x_{j+\frac{1}{2}}$, respectively, i.e. $v^\pm_{j+\frac{1}{2}} = v(x_{j+\frac{1}{2}} \pm 0)$. Moreover, we denote the average and jump of v at $x_{j+\frac{1}{2}}$ by $\{v\}_{j+\frac{1}{2}} = \frac{1}{2}\left(v^-_{j+\frac{1}{2}} + v^+_{j+\frac{1}{2}}\right)$ and $[v]_{j+\frac{1}{2}} = v^+_{j+\frac{1}{2}} - v^-_{j+\frac{1}{2}}$, respec-

tively. Similarly, in two space dimensions, we denote the left/right and lower/upper limits of v on vertical and horizontal cell interfaces by $v(x^{\pm}_{i+\frac{1}{2}}, y) = v(x_{i+\frac{1}{2}} \pm 0, y)$ and $v(x, y^{\pm}_{j+\frac{1}{2}}) = v(x, y_{j+\frac{1}{2}} \pm 0)$, respectively. The averages and jumps of v on vertical and horizontal cell interfaces are defined as $\{v\}(x_{i+\frac{1}{2}}, y) = \frac{1}{2}\left(v(x^-_{i+\frac{1}{2}}, y) + v(x^+_{i+\frac{1}{2}}, y)\right)$, $[v](x_{i+\frac{1}{2}}, y) = v(x^+_{i+\frac{1}{2}}, y) - v(x^-_{i+\frac{1}{2}}, y)$ and $\{v\}(x, y_{j+\frac{1}{2}}) = \frac{1}{2}\left(v(x, y^-_{j+\frac{1}{2}}) + v(x, y^+_{j+\frac{1}{2}})\right)$, $[v](x, y_{j+\frac{1}{2}}) = v(x, y^+_{j+\frac{1}{2}}) - v(x, y^-_{j+\frac{1}{2}})$, respectively. For simplicity, these notations will be abbreviated as $v^{\pm}$, $\{v\}$ and $[v]$ when the cell interface is clear from the context.

We denote the L^2 inner product on cell I_j in one space dimension as

$$(u, v)_{I_j} = \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x)v(x)dx,$$

and on $K_{i,j}$ in two space dimensions as

$$(u, v)_{K_{i,j}} = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u(x, y)v(x, y)dxdy,$$

for $u, v \in L^2(\Omega)$.

We use the Gauss-Lobatto quadrature of $2N_q - 1$ points to evaluate integrals in one dimensional cells, where N_q is taken such that the third order accuracy is attained in the scheme, e.g. $N_q = 3$. We denote the quadrature points on I_j as $\{\hat{x}_\gamma, \gamma = 1, \ldots, 2N_q - 1\}$, and let $\{\hat{\omega}_\gamma, \gamma = 1, \ldots, 2N_q - 1\}$ be the corresponding quadrature weights satisfying $\sum_{\gamma=1}^{2N_q-1} \hat{\omega}_\gamma = 1$. In particular, $\hat{x}_1 = x_{j-\frac{1}{2}}, \hat{x}_{N_q} = x_j$ and $\hat{x}_{2N_q-1} = x_{j+\frac{1}{2}}$. We denote $\hat{u}^\gamma = u(\hat{x}_\gamma)$, for $\gamma = 1, \ldots, 2N_q - 1$. The quadrature rule adopted in two dimensional cells follows from tensor product and we denote $\hat{u}^{\beta,\gamma} = u(\hat{x}_\beta, \hat{y}_\gamma)$, for $\beta, \gamma = 1, \ldots, 2N_q - 1$, on the cell $K_{i,j}$.

2.2.2 Scalar conservation laws in one dimension

Consider the scalar conservation law (2.1). Direct computation gives the expressions of u_t, u_{tt} and u_{ttt} as follows:

$$u_t = -f(u)_x, \tag{2.5}$$

$$u_{tt} = ((f')^2 u_x)_x \tag{2.6}$$

$$u_{ttt} = -\left(3f''(f')^2 u_x^2 + (f')^3 u_{xx}\right)_x \tag{2.7}$$

Based on the expansions (2.5), (2.6) and (2.7), the third order maximum-principle-satisfying LWDG scheme of (2.1) at time level t^n is to find $u^{n+1} \in V$, s.t. $\forall \xi \in V$, the equation

$$\begin{aligned}(u^{n+1}, \xi)_{I_j} =& (u, \xi)_{I_j} + \Delta t (f(u), \xi_x)_{I_j} - \frac{\Delta t^2}{2}((f')^2 u_x, \xi_x)_{I_j} \\ &+ \frac{\Delta t^3}{6}(3f''(f')^2 u_x^2 + (f')^3 u_{xx}, \xi_x)_{I_j} - \Delta t \hat{F}_{j+\frac{1}{2}} \xi^-_{j+\frac{1}{2}} + \Delta t \hat{F}_{j-\frac{1}{2}} \xi^+_{j-\frac{1}{2}},\end{aligned} \tag{2.8}$$

holds for $j = 1, 2, \ldots, N$, where the superscript n denoting time level t^n on the right hand side is omitted.

In the scheme (2.8), $\hat{F}_{j+\frac{1}{2}}$ is the numerical flux at $x_{j+\frac{1}{2}}$ defined as

$$\hat{F}_{j+\frac{1}{2}} = \hat{f}^{\mathrm{LF}}_{j+\frac{1}{2}} - \frac{\Delta t}{2}\{f'^2\}_{j+\frac{1}{2}} \widehat{u_x}^{\mathrm{DDG}}_{j+\frac{1}{2}} + \frac{\Delta t^2}{6}\{3f'^2 f'' u_x^2 + f'^3 u_{xx}\}_{j+\frac{1}{2}}, \tag{2.9}$$

where

$$\hat{f}^{\mathrm{LF}}_{j+\frac{1}{2}} = \{f\}_{j+\frac{1}{2}} - \frac{\alpha}{2}[u]_{j+\frac{1}{2}}, \quad \alpha = \max_u |f'(u)| \tag{2.10}$$

is the Lax-Friedriches flux as used in [93], and

$$\widehat{u_x}^{\mathrm{DDG}}_{j+\frac{1}{2}} = \beta_0 \frac{[u]_{j+\frac{1}{2}}}{\Delta x} + \{u_x\}_{j+\frac{1}{2}} + \beta_1 \Delta x [u_{xx}]_{j+\frac{1}{2}} \tag{2.11}$$

is the bound-preserving direct discontinuous Galerkin (DDG) flux [47, 11], with β_0, β_1 satisfying

$$\frac{1}{8} < \beta_1 < \frac{1}{4}, \quad \beta_0 > \frac{3}{2} - 4\beta_1 \tag{2.12}$$

The following lemmas are useful in the proofs of maximum-principle-satisfying and positivity-preserving in this section and the next section.

Lemma 2.2.1. *For $u \in V$, the DDG flux $\widehat{u_x}^{DDG}_{j+\frac{1}{2}}$ defined in* (2.11) *can be expanded on uniform meshes as*

$$\begin{aligned} \widehat{u_x}^{DDG}_{j+\frac{1}{2}} =& \frac{1}{\Delta x}\Bigg((\frac{1}{2} - 4\beta_1) u^+_{j-\frac{1}{2}} + (-2 + 8\beta_1) u_j + (-\beta_0 + \frac{3}{2} - 4\beta_1) u^-_{j+\frac{1}{2}} \\ &+ (\beta_0 - \frac{3}{2} + 4\beta_1) u^+_{j+\frac{1}{2}} + (2 - 8\beta_1) u_{j+1} + (-\frac{1}{2} + 4\beta_1) u^-_{j+\frac{3}{2}} \Bigg) \end{aligned} \tag{2.13}$$

Proof. Since the mesh is uniform and u is piecewise quadratic, it follows from direct calculations. □

Lemma 2.2.2. *If $u \in V$ and $m \le u \le M$, then*

$$|\frac{du}{dx}| \le \frac{5(M-m)}{\Delta x_j}, \quad \forall x \in I_j. \tag{2.14}$$

Proof. We first consider $v \in P^2([-1,1])$ with $-\frac{R}{2} \le v \le \frac{R}{2}$. The Lagrange interpolation gives

$$v(r) = v(-1)L_{-1}(r) + v(0)L_0(r) + v(1)L_1(r), \quad r \in [-1,1], \tag{2.15}$$

where $L_{-1}(r) = \frac{1}{2}r(r-1), L_0(r) = -(r+1)(r-1), L_1(r) = \frac{1}{2}r(r+1)$.

Therefore, $|v'(r)| \le |v(-1)| \cdot |L'_{-1}(r)| + |v(0)| \cdot |L'_0(r)| + |v(1)| \cdot |L'_1(r)| \le \frac{R}{2} \times \frac{3}{2} + \frac{R}{2} \times 2 + \frac{R}{2} \times \frac{3}{2} = \frac{5R}{2}, \forall r \in [-1, 1]$. Then (2.14) follows from changing of variables and the chain rule. □

We now state our main result for the LWDG scheme (2.8).

Theorem 2.2.3. *Given $m \le u^n \le M$, the cell averages $\bar{u}_j^{n+1}, j = 1, \ldots, N$ of the solution of scheme* (2.8) *are bounded between m and M under the CFL condition* (2.16).

$$\lambda \le \min\left\{q_1, q_2, \ldots, q_6\right\}, \tag{2.16}$$

where $\lambda = \frac{\Delta t}{\Delta x}$, $q_1 = \frac{\hat{\omega}_1}{2M_1}$, $q_2 = \frac{4\beta_1 - \frac{1}{2}}{5(M-m)M_2 + \frac{4}{3}M_1}$, $q_3 = \frac{2 - 8\beta_1}{20(M-m)M_2 + \frac{8}{3}M_1}$, $q_4 = \frac{\beta_0 - \frac{3}{2} + 4\beta_1}{15(M-m)M_2 + \frac{4}{3}M_1}$, $q_5 = \frac{\hat{\omega}_1^{1/2}}{M_1(\beta_0 - 1 + 4\beta_1)^{1/2}}$, $q_6 = \frac{\hat{\omega}_{N_q}^{1/2}}{M_1(6 - 24\beta_1)^{1/2}}$, *and* $M_1 = \max_{m \le u \le M} |f'(u)|$, $M_2 = \max_{m \le u \le M} |f''(u)|$.

Proof. Take the test function $\xi = 1$ on I_j and zero anywhere else in the scheme (2.8) and denote $\lambda = \frac{\Delta t}{\Delta x}$, we obtain the equation satisfied by cell average of u^{n+1} on cell I_j,

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \lambda \hat{F}_{j+\frac{1}{2}} + \lambda \hat{F}_{j-\frac{1}{2}} = \mathrm{I} + \mathrm{II}, \tag{2.17}$$

where

$$\mathrm{I} = \frac{1}{2}\left(\bar{u}_j^n - 2\lambda \hat{f}^{\mathrm{LF}}_{j+\frac{1}{2}} + 2\lambda \hat{f}^{\mathrm{LF}}_{j-\frac{1}{2}}\right), \tag{2.18}$$

and

$$
\begin{aligned}
\mathrm{II} =& \frac{1}{2}\sum_{\gamma=1}^{2N_q-1}\hat{\omega}_\gamma\hat{u}^\gamma \\
& -\lambda\left(-\frac{\Delta t}{4}(f'^{2-}_{j+\frac{1}{2}}+f'^{2+}_{j+\frac{1}{2}})\widehat{u}^{\,\mathrm{DDG}}_{x\,j+\frac{1}{2}}+\frac{\Delta t^2}{12}\times\right. \\
& \left.(3f'^{2-}_{j+\frac{1}{2}}f''^{-}_{j+\frac{1}{2}}u^{2-}_{xj+\frac{1}{2}}+3f'^{2+}_{j+\frac{1}{2}}f''^{+}_{j+\frac{1}{2}}u^{2+}_{xj+\frac{1}{2}}+f'^{3-}_{j+\frac{1}{2}}u_{xx}{}^{-}_{j+\frac{1}{2}}+f'^{3+}_{j+\frac{1}{2}}u_{xx}{}^{+}_{j+\frac{1}{2}})\right) \\
& +\lambda\left(-\frac{\Delta t}{4}(f'^{2-}_{j-\frac{1}{2}}+f'^{2+}_{j-\frac{1}{2}})\widehat{u}^{\,\mathrm{DDG}}_{x\,j-\frac{1}{2}}+\frac{\Delta t^2}{12}\times\right. \\
& \left.(3f'^{2-}_{j-\frac{1}{2}}f''^{-}_{j-\frac{1}{2}}u^{2-}_{xj-\frac{1}{2}}+3f'^{2+}_{j-\frac{1}{2}}f''^{+}_{j-\frac{1}{2}}u^{2+}_{xj-\frac{1}{2}}+f'^{3-}_{j-\frac{1}{2}}u_{xx}{}^{-}_{j-\frac{1}{2}}+f'^{3+}_{j-\frac{1}{2}}u_{xx}{}^{+}_{j-\frac{1}{2}})\right)
\end{aligned}
$$

Note that the cell average $\bar{u}_j^n$ is split equally in I and II just for the ease of written, rather than to obtain an optimal CFL condition, which is the same case for all other proofs in this chapter.

Since I has exactly the same form as in [93], we have $\frac{1}{2}m \le \mathrm{I} \le \frac{1}{2}M$, under the condition $\lambda \le q_1$ based on the conclusion therein. One can refer to [93] for more details.

As for II, it can be expanded as follows:

$$
\begin{aligned}
\mathrm{II} =& \frac{1}{2}\sum_{\gamma=2}^{N_q-1}\hat{\omega}_\gamma\hat{u}^\gamma+\frac{1}{2}\sum_{\gamma=N_q+1}^{2N_q-2}\hat{\omega}_\gamma\hat{u}^\gamma+z_1u^+_{j-\frac{3}{2}}+z_2u_{j-1}+z_3u^-_{j-\frac{1}{2}}+z_4u^+_{j-\frac{1}{2}} \\
& +z_5u_j+z_6u^-_{j+\frac{1}{2}}+z_7u^+_{j+\frac{1}{2}}+z_8u_{j+1}+z_9u^-_{j+\frac{3}{2}},
\end{aligned}
\tag{2.19}
$$

where

$$
z_1=\frac{\lambda^2}{4}f'^{2-}_{j-\frac{1}{2}}\left((4\beta_1-\frac{1}{2})+\Delta t f''^{-}_{j-\frac{1}{2}}u_{x}{}^{-}_{j-\frac{1}{2}}+\frac{4\lambda}{3}f'^{-}_{j-\frac{1}{2}}\right)+\frac{\lambda^2}{4}f'^{2+}_{j-\frac{1}{2}}(4\beta_1-\frac{1}{2}),
$$

$$
z_2=\frac{\lambda^2}{4}f'^{2-}_{j-\frac{1}{2}}\left((2-8\beta_1)-4\Delta t f''^{-}_{j-\frac{1}{2}}u_{x}{}^{-}_{j-\frac{1}{2}}-\frac{8\lambda}{3}f'^{-}_{j-\frac{1}{2}}\right)+\frac{\lambda^2}{4}f'^{2+}_{j-\frac{1}{2}}(2-8\beta_1)
$$

$$
\begin{aligned}
z_3 =& \frac{\lambda^2}{4} {f'^2}^-_{j-\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 3\Delta t {f''}^-_{j-\frac{1}{2}} {u_x}^-_{j-\frac{1}{2}} + \frac{4\lambda}{3} {f'}^-_{j-\frac{1}{2}} \right) \\
& + \frac{\lambda^2}{4} {f'^2}^+_{j-\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \\
z_4 =& \frac{1}{2}\hat{\omega}_1 - \frac{\lambda^2}{4} {f'^2}^-_{j-\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \\
& - \frac{\lambda^2}{4} {f'^2}^+_{j-\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 3\Delta t {f''}^+_{j-\frac{1}{2}} {u_x}^+_{j-\frac{1}{2}} - \frac{4\lambda}{3} {f'}^+_{j-\frac{1}{2}} \right) \\
& - \frac{\lambda^2}{4} {f'^2}^-_{j+\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) + \Delta t {f''}^-_{j+\frac{1}{2}} {u_x}^-_{j+\frac{1}{2}} + \frac{4\lambda}{3} {f'}^-_{j+\frac{1}{2}} \right) - \frac{\lambda^2}{4} {f'^2}^+_{j+\frac{1}{2}} (4\beta_1 - \frac{1}{2}) \\
z_5 =& \frac{1}{2}\hat{\omega}_{N_q} - \frac{\lambda^2}{4} {f'^2}^-_{j-\frac{1}{2}} (2 - 8\beta_1) - \frac{\lambda^2}{4} {f'^2}^+_{j-\frac{1}{2}} \left((2 - 8\beta_1) - 4\Delta t {f''}^+_{j-\frac{1}{2}} {u_x}^+_{j-\frac{1}{2}} + \frac{8\lambda}{3} {f'}^+_{j-\frac{1}{2}} \right) \\
& - \frac{\lambda^2}{4} {f'^2}^-_{j+\frac{1}{2}} \left((2 - 8\beta_1) - 4\Delta t {f''}^-_{j+\frac{1}{2}} {u_x}^-_{j+\frac{1}{2}} - \frac{8\lambda}{3} {f'}^-_{j+\frac{1}{2}} \right) - \frac{\lambda^2}{4} {f'^2}^+_{j+\frac{1}{2}} (2 - 8\beta_1) \\
z_6 =& \frac{1}{2}\hat{\omega}_{2N_q-1} - \frac{\lambda^2}{4} {f'^2}^-_{j-\frac{1}{2}} (4\beta_1 - \frac{1}{2}) \\
& - \frac{\lambda^2}{4} {f'^2}^+_{j-\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) + \Delta t {f''}^+_{j-\frac{1}{2}} {u_x}^+_{j-\frac{1}{2}} - \frac{4\lambda}{3} {f'}^+_{j-\frac{1}{2}} \right) \\
& - \frac{\lambda^2}{4} {f'^2}^-_{j+\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 3\Delta t {f''}^-_{j+\frac{1}{2}} {u_x}^-_{j+\frac{1}{2}} + \frac{4\lambda}{3} {f'}^-_{j+\frac{1}{2}} \right) \\
& - \frac{\lambda^2}{4} {f'^2}^+_{j+\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \\
z_7 =& \frac{\lambda^2}{4} {f'^2}^-_{j+\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) + \frac{\lambda^2}{4} {f'^2}^+_{j+\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 3\Delta t {f''}^+_{j+\frac{1}{2}} {u_x}^+_{j+\frac{1}{2}} - \frac{4\lambda}{3} {f'}^+_{j+\frac{1}{2}} \right) \\
z_8 =& \frac{\lambda^2}{4} {f'^2}^-_{j+\frac{1}{2}} (2 - 8\beta_1) + \frac{\lambda^2}{4} {f'^2}^+_{j+\frac{1}{2}} \left((2 - 8\beta_1) - 4\Delta t {f''}^+_{j+\frac{1}{2}} {u_x}^+_{j+\frac{1}{2}} + \frac{8\lambda}{3} {f'}^+_{j+\frac{1}{2}} \right) \\
z_9 =& \frac{\lambda^2}{4} {f'^2}^-_{j+\frac{1}{2}} (4\beta_1 - \frac{1}{2}) + \frac{\lambda^2}{4} {f'^2}^+_{j+\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) + \Delta t {f''}^+_{j+\frac{1}{2}} {u_x}^+_{j+\frac{1}{2}} - \frac{4\lambda}{3} {f'}^+_{j+\frac{1}{2}} \right)
\end{aligned}
$$

It is not difficult to verify that

$$
\frac{1}{2} \sum_{\gamma=2}^{N_q-1} \hat{\omega}_\gamma + \frac{1}{2} \sum_{\gamma=N_q+1}^{2N_q-2} \hat{\omega}_\gamma + z_1 + z_2 + \cdots + z_9 = \frac{1}{2},
$$

Moreover, we claim that $z_1, z_2, \ldots, z_9 \geq 0$ under the CFL conditions (2.16). In fact,

the following estimates can be made under the CFL conditions,

$$
\begin{aligned}
z_1 \geq & \frac{\lambda^2}{4} f'^{2-}_{j-\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) - 5\lambda(M-m)M_2 - \frac{4\lambda}{3} M_1 \right) + \frac{\lambda^2}{4} f'^{2+}_{j-\frac{1}{2}} (4\beta_1 - \frac{1}{2}) \geq 0, \\
z_2 \geq & \frac{\lambda^2}{4} f'^{2-}_{j-\frac{1}{2}} \left((2 - 8\beta_1) - 20\lambda(M-m)M_2 - \frac{8\lambda}{3} M_1 \right) + \frac{\lambda^2}{4} f'^{2+}_{j-\frac{1}{2}} (2 - 8\beta_1) \geq 0, \\
z_3 \geq & \frac{\lambda^2}{4} f'^{2-}_{j-\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) - 15\lambda(M-m)M_2 - \frac{4\lambda}{3} M_1 \right) \\
& + \frac{\lambda^2}{4} f'^{2+}_{j-\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \geq 0, \\
z_4 \geq & \frac{1}{2}\hat{\omega}_1 - \frac{\lambda^2}{4} M_1^2 (\beta_0 - \frac{3}{2} + 4\beta_1) \\
& - \frac{\lambda^2}{4} M_1^2 \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 15\lambda(M-m)M_2 + \frac{4\lambda}{3} M_1 \right) \\
& - \frac{\lambda^2}{4} M_1^2 \left((4\beta_1 - \frac{1}{2}) + 5\lambda(M-m)M_2 + \frac{4\lambda}{3} M_1 \right) - \frac{\lambda^2}{4} M_1^2 (4\beta_1 - \frac{1}{2}) \geq 0, \\
z_5 \geq & \frac{1}{2}\hat{\omega}_N - \frac{\lambda^2}{4} M_1^2 (2 - 8\beta_1) - \frac{\lambda^2}{4} M_1^2 \left((2 - 8\beta_1) + 20\lambda(M-m)M_2 + \frac{8\lambda}{3} M_1 \right) \\
& - \frac{\lambda^2}{4} M_1^2 \left((2 - 8\beta_1) + 20\lambda(M-m)M_2 + \frac{8\lambda}{3} M_1 \right) - \frac{\lambda^2}{4} M_1^2 (2 - 8\beta_1) \geq 0, \\
z_6 \geq & \frac{1}{2}\hat{\omega}_{2N_q-1} - \frac{\lambda^2}{4} M_1^2 (4\beta_1 - \frac{1}{2}) - \frac{\lambda^2}{4} M_1^2 \left((4\beta_1 - \frac{1}{2}) + 5\lambda(M-m)M_2 + \frac{4\lambda}{3} M_1 \right) \\
& - \frac{\lambda^2}{4} M_1^2 \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 15\lambda(M-m)M_2 + \frac{4\lambda}{3} M_1 \right) - \frac{\lambda^2}{4} M_1^2 (\beta_0 - \frac{3}{2} + 4\beta_1) \\
\geq & 0, \\
z_7 \geq & \frac{\lambda^2}{4} f'^{2-}_{j+\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \\
& + \frac{\lambda^2}{4} f'^{2+}_{j+\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) - 15\lambda(M-m)M_2 - \frac{4\lambda}{3} M_1 \right) \geq 0, \\
z_8 \geq & \frac{\lambda^2}{4} f'^{2-}_{j+\frac{1}{2}} (2 - 8\beta_1) + \frac{\lambda^2}{4} f'^{2+}_{j+\frac{1}{2}} \left((2 - 8\beta_1) - 20\lambda(M-m)M_2 - \frac{8\lambda}{3} M_1 \right) \geq 0, \\
z_9 \geq & \frac{\lambda^2}{4} f'^{2-}_{j+\frac{1}{2}} (4\beta_1 - \frac{1}{2}) + \frac{\lambda^2}{4} f'^{2+}_{j+\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) - 5\lambda(M-m)M_2 - \frac{4\lambda}{3} M_1 \right) \geq 0.
\end{aligned}
$$

Therefore, II is one half of a convex combination of values of u^n at different quadrature points, which implies $\frac{1}{2}m \leq \mathrm{II} \leq \frac{1}{2}M$ since we assume $m \leq u^n \leq M$.

Since $\bar{u}_j^{n+1} = \mathrm{I} + \mathrm{II}$, we finish the proof by summing up the inequalities of I and

II . □

Remark 2.2.1. *The CFL condition in Theorem 2.2.3 is not sharp, because we split the cell average $\bar{u}_j^n$ equally into* I *and* II *for the convenience of the proof. The same case applies to all later theorems. In order to get a sharp CFL condition, one has to analyze/estimate $\bar{u}^{n+1}$ or $\bar{\mathbf{u}}^{n+1}$ as a whole, which makes the proof extremely tedious. But even if we did so, the task of finding optimal β_0 and β_1 to obtain an exact upper bound of the CFL number would still be very difficult if not impossible, since the CFL condition also depends on the lower and upper bounds of the solution, the maximum norms of the first and second derivatives of the flux function, and the quadrature rule, etc.*

However, we can get an intuition about the CFL constraints in the LWDG by analyzing the equation $u_t + u_x = 0$. In this case, the Lax-Friedrichs flux becomes the upwind flux, and the upper bound on the time step constraints can be computed exactly. Calculation shows that, using the 5-point Gauss-Lobatto quadrature, the CFL number of the LWDG is $\frac{\Delta t}{\Delta x} = 0.049917$, with the optimal parameters $\beta_0 = 0.999978, \beta_1 = 0.133326$. In comparison, under the same quadrature rule, the CFL numbers of the maximum-principle-satisfying DG schemes [93] are $\frac{\Delta t}{\Delta x} = 0.05$ and $\frac{\Delta t}{\Delta x} = 0.016666$, for the SSP-RK3 method (three stages) and SSP3 multi-step method (single-stage), respectively.

2.2.3 Scalar conservation laws in two dimensions

Consider the scalar conservation law in two space dimensions

$$u_t + f(u)_x + g(u)_y = 0. \tag{2.20}$$

Direct computation gives the expressions of u_t, u_{tt}, u_{ttt} as follows:

$$u_t = -f(u)_x - g(u)_y, \tag{2.21}$$

$$u_{tt} = \left(f'^2 u_x\right)_x + \left(f'g'u_y\right)_x + \left(f'g'u_x\right)_y + \left(g'^2 u_y\right)_y, \tag{2.22}$$

$$\begin{aligned} u_{ttt} = &- \left(3f'^2 f'' u_x^2 + 6f'g'g'' u_y^2 + 3g'^2 f'' u_y^2 + f'^3 u_{xx} + 3f'g'^2 u_{yy}\right)_x \\ &- \left(6f'g'f'' u_x^2 + 3f'^2 g'' u_x^2 + 3g'^2 g'' u_y^2 + 3f'^2 g' u_{xx} + g'^3 u_{yy}\right)_y \end{aligned} \tag{2.23}$$

Note that there are different ways to expand u_{ttt}, among which we choose the one that avoids the appearance of mixed derivatives in the numerical scheme.

Based on the expansions (2.21), (2.22) and (2.23), the third order maximum-principle-preserving LWDG scheme of (2.20) at time level t^n is to find $u^{n+1} \in W$, s.t. $\forall \xi \in W$, the equation

$$\begin{aligned} (u^{n+1}, \xi)_{K_{i,j}} =& (u, \xi)_{K_{i,j}} + \Delta t (f(u), \xi_x)_{K_{i,j}} + \Delta t (g(u), \xi_y)_{K_{i,j}} \\ &- \frac{\Delta t^2}{2} (f'^2 u_x + f'g'u_y, \xi_x)_{K_{i,j}} - \frac{\Delta t^2}{2} (f'g'u_x + g'^2 u_y, \xi_y)_{K_{i,j}} \\ &+ \frac{\Delta t^3}{6} \left(3f'^2 f'' u_x^2 + 6f'g'g'' u_y^2 + 3g'^2 f'' u_y^2 + f'^3 u_{xx} + 3f'g'^2 u_{yy}, \xi_x\right)_{K_{i,j}} \\ &+ \frac{\Delta t^3}{6} \left(6f'g'f'' u_x^2 + 3f'^2 g'' u_x^2 + 3g'^2 g'' u_y^2 + 3f'^2 g' u_{xx} + g'^3 u_{yy}, \xi_y\right)_{K_{i,j}} \\ &- \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}_{i+\frac{1}{2},j}\ \xi(x^-_{i+\frac{1}{2}}, y) dy + \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}_{i-\frac{1}{2},j}\ \xi(x^+_{i-\frac{1}{2}}, y) dy \\ &- \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}_{i,j+\frac{1}{2}}\ \xi(x, y^-_{j+\frac{1}{2}}) dx + \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}_{i,j-\frac{1}{2}}\ \xi(x, y^+_{j-\frac{1}{2}}) dx \end{aligned} \tag{2.24}$$

holds for $i = 1, \ldots, N_x, j = 1, \ldots, N_y$. In the scheme, $\hat{F}_{i+\frac{1}{2},j}$ and $\hat{G}_{i,j+\frac{1}{2}}$ are numerical

fluxes defined as

$$\hat{F}_{i+\frac{1}{2},j} = \hat{F}^0_{i+\frac{1}{2},j} + \hat{F}^1_{i+\frac{1}{2},j}, \quad \hat{G}_{i,j+\frac{1}{2}} = \hat{G}^0_{i,j+\frac{1}{2}} + \hat{G}^1_{i,j+\frac{1}{2}},$$

where

$$\hat{F}^0_{i+\frac{1}{2},j} = \hat{f}^{\mathrm{LF}}_{i+\frac{1}{2},j} - \frac{\Delta t}{2}\{f'^2\}_{i+\frac{1}{2},j}\widehat{u_x}^{\mathrm{DDG}}_{i+\frac{1}{2},j} + \frac{\Delta t^2}{6}\{3f'^2f''u_x^2 + f'^3u_{xx}\}_{i+\frac{1}{2},j}, \tag{2.25}$$

$$\begin{aligned}\hat{F}^1_{i+\frac{1}{2},j} =& -\frac{1}{2}\alpha^1_x[u]_{i+\frac{1}{2},j} - \frac{\Delta t}{2}\{f'g'u_y\}_{i+\frac{1}{2},j} \\ &+ \frac{\Delta t^2}{6}\{6f'g'g''u_y^2 + 3g'^2f''u_y^2 + 3f'g'^2u_{yy}\}_{i+\frac{1}{2},j},\end{aligned} \tag{2.26}$$

$$\hat{G}^0_{i,j+\frac{1}{2}} = \hat{g}^{\mathrm{LF}}_{i,j+\frac{1}{2}} - \frac{\Delta t}{2}\{g'^2\}_{i,j+\frac{1}{2}}\widehat{u_y}^{\mathrm{DDG}}_{i,j+\frac{1}{2}} + \frac{\Delta t^2}{6}\{3g'^2g''u_y^2 + g'^3u_{yy}\}_{i,j+\frac{1}{2}}, \tag{2.27}$$

$$\begin{aligned}\hat{G}^1_{i,j+\frac{1}{2}} =& -\frac{1}{2}\alpha^1_y[u]_{i,j+\frac{1}{2}} - \frac{\Delta t}{2}\{f'g'u_x\}_{i,j+\frac{1}{2}} \\ &+ \frac{\Delta t^2}{6}\{6f'g'f''u_x^2 + 3f'^2g''u_x^2 + 3f'^2g'u_{xx}\}_{i,j+\frac{1}{2}},\end{aligned} \tag{2.28}$$

in which the Lax-Friedrichs fluxes and DDG fluxes are defined the same way as before, and α^1_x, α^1_y are positive viscosity constants that can be taken as $0.05\max_u|f'(u)|$ and $0.05\max_u|g'(u)|$ for instance. In fact, any constants strictly positive should be enough for positivity-preserving, which just makes difference on the CFL numbers and the dissipation effect.

We now state the main result for the LWDG scheme (2.24).

Theorem 2.2.4. *Given $m \le u^n \le M$, the cell averages $\bar{u}^{n+1}_{i,j}$, $i = 1,\dots,N_x, j = 1,\dots,N_y$ of the solution of scheme* (2.24) *are bounded between m and M under the CFL condition* (2.29)*:*

$$\lambda_x \le \min\{Q_1, Q_3\}, \quad \lambda_y \le \min\{Q_2, Q_4\}, \tag{2.29}$$

where $\lambda_x = \frac{\Delta t}{\Delta x}$, $\lambda_y = \frac{\Delta t}{\Delta y}$, and the definitions of Q_1, Q_2, Q_3, Q_4 are given in Appendix A.1.1.

The proof is very similar to that of the one dimensional case, except that the expansions are much more tedious, which results in much more complicated CFL conditions.

Proof. Take the test function $\xi = 1$ on $K_{i,j}$ and zero anywhere else in the scheme (2.24) and denote by $\lambda_x = \frac{\Delta t}{\Delta x}$, $\lambda_y = \frac{\Delta t}{\Delta y}$, we obtain

$$\bar{u}_{i,j}^{n+1} = \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV},$$

where

$$\mathrm{I} = \frac{1}{4}\bar{u}_{i,j}^{n} - \lambda_x \frac{1}{\Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}^0_{i+\frac{1}{2},j}dy + \lambda_x \frac{1}{\Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}^0_{i-\frac{1}{2},j}dy,$$

$$\mathrm{II} = \frac{1}{4}\bar{u}_{i,j}^{n} - \lambda_x \frac{1}{\Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}^1_{i+\frac{1}{2},j}dy + \lambda_x \frac{1}{\Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}^1_{i-\frac{1}{2},j}dy,$$

$$\mathrm{III} = \frac{1}{4}\bar{u}_{i,j}^{n} - \lambda_y \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}^0_{i,j+\frac{1}{2}}dx + \lambda_y \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}^0_{i,j-\frac{1}{2}}dx$$

$$\mathrm{IV} = \frac{1}{4}\bar{u}_{i,j}^{n} - \lambda_y \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}^1_{i,j+\frac{1}{2}}dx + \lambda_y \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}^1_{i,j-\frac{1}{2}}dx$$

It suffices to shown $\frac{1}{4}m \leq \mathrm{I}, \mathrm{II} \leq \frac{1}{4}M$ under the CFL condition (2.29), due to the symmetry in x and y directions.

It is clear that I can be decomposed in the form of convex combination

$$\mathrm{I} = \frac{1}{4}\sum_{\gamma=1}^{2N_q-1}\hat{\omega}_\gamma H_\gamma$$

where

$$H_\gamma = \sum_{\beta=1}^{2N_q-1}\hat{\omega}_\beta\hat{u}^{\beta,\gamma} - 4\lambda_x\hat{F}^0_{i+\frac{1}{2},j}(x_{i+\frac{1}{2}},\hat{y}_\gamma) + 4\lambda_x\hat{F}^0_{i-\frac{1}{2},j}(x_{i-\frac{1}{2}},\hat{y}_\gamma),$$

Notice that H_γ has exactly the same structure as (2.17). Therefore, $\mathrm{I}\in[\frac{1}{4}m,\frac{1}{4}M]$, under the CFL condition (2.16) for one dimensional scalar case with λ replaced by $4\lambda_x$, i.e. $\lambda_x \le Q_1$.

As for the term II, it can be expanded as follows,

$$\begin{aligned}
\mathrm{II} =& \frac{1}{4}\sum_{\alpha=2}^{2N_q-2}\sum_{\beta=1}^{2N_q-1}\hat{\omega}_\alpha\hat{\omega}_\beta\hat{u}^{\alpha,\beta} + \sum_{\beta=2}^{N_q-1}\frac{\lambda_x}{2}\hat{\omega}_\beta\alpha^1_x u(x^-_{i-\frac{1}{2}},\hat{y}_\beta) + \sum_{\beta=N_q+1}^{2N_q-2}\frac{\lambda_x}{2}\hat{\omega}_\beta\alpha^1_x u(x^-_{i-\frac{1}{2}},\hat{y}_\beta)\\
&+ \sum_{\beta=2}^{N_q-1}\frac{\lambda_x}{2}\hat{\omega}_\beta\alpha^1_x u(x^+_{i+\frac{1}{2}},\hat{y}_\beta) + \sum_{\beta=N_q+1}^{2N_q-2}\frac{\lambda_x}{2}\hat{\omega}_\beta\alpha^1_x u(x^+_{i+\frac{1}{2}},\hat{y}_\beta)\\
&+ z_1u(x^-_{i-\frac{1}{2}},y^+_{j-\frac{1}{2}}) + z_2u(x^-_{i-\frac{1}{2}},y_j) + z_3u(x^-_{i-\frac{1}{2}},y^-_{j+\frac{1}{2}}) + z_4u(x^+_{i-\frac{1}{2}},y^+_{j-\frac{1}{2}})\\
&+ z_5u(x^+_{i-\frac{1}{2}},y_j) + z_6u(x^+_{i-\frac{1}{2}},y^-_{j+\frac{1}{2}}) + z_7u(x^-_{i+\frac{1}{2}},y^+_{j-\frac{1}{2}}) + z_8u(x^-_{i+\frac{1}{2}},y_j)\\
&+ z_9u(x^-_{i+\frac{1}{2}},y^-_{j+\frac{1}{2}}) + z_{10}u(x^+_{i+\frac{1}{2}},y^+_{j-\frac{1}{2}}) + z_{11}u(x^+_{i+\frac{1}{2}},y_j) + z_{12}u(x^+_{i+\frac{1}{2}},y^-_{j+\frac{1}{2}})\\
&+ \sum_{\beta=2}^{N_q-1}\hat{\omega}_\beta z_{13,\beta}u(x^-_{i+\frac{1}{2}},\hat{y}_\beta) + \sum_{\beta=N_q+1}^{2N_q-2}\hat{\omega}_\beta z_{13,\beta}u(x^-_{i+\frac{1}{2}},\hat{y}_\beta)\\
&+ \sum_{\beta=2}^{N_q-1}\hat{\omega}_\beta z_{14,\beta}u(x^+_{i-\frac{1}{2}},\hat{y}_\beta) + \sum_{\beta=N_q+1}^{2N_q-2}\hat{\omega}_\beta z_{14,\beta}u(x^+_{i-\frac{1}{2}},\hat{y}_\beta),
\end{aligned} \tag{2.30}$$

where the expressions of $z_1,\ldots,z_{14,\beta}$ are given in Appendix A.1.2.

It can be verified that the following equality holds,

$$\frac{1}{4}\sum_{\alpha=2}^{2N_q-2}\sum_{\beta=1}^{2N_q-1}\hat{\omega}_\alpha\hat{\omega}_\beta+\sum_{\beta=2}^{N_q-1}\frac{\lambda_x}{2}\hat{\omega}_\beta\alpha_x^1+\sum_{\beta=N_q+1}^{2N_q-2}\frac{\lambda_x}{2}\hat{\omega}_\beta\alpha_x^1+\sum_{\beta=2}^{N_q-1}\frac{\lambda_x}{2}\hat{\omega}_\beta\alpha_x^1+\sum_{\beta=N_q+1}^{2N_q-2}\frac{\lambda_x}{2}\hat{\omega}_\beta\alpha_x^1$$

$$+z_1+z_2+z_3+z_4+z_5+z_6+z_7+z_8+z_9+z_{10}+z_{11}+z_{12}$$

$$+\sum_{\beta=2}^{N_q-1}\hat{\omega}_\beta z_{13,\beta}+\sum_{\beta=N_q+1}^{2N_q-2}\hat{\omega}_\beta z_{13,\beta}+\sum_{\beta=2}^{N_q-1}\hat{\omega}_\beta z_{14,\beta}+\sum_{\beta=N_q+1}^{2N_q-2}\hat{\omega}_\beta z_{14,\beta}=\frac{1}{4},$$

Moreover, all z's are nonnegative under the CFL condition (2.29). The detailed estimates can be found in Appendix A.1.2

To sum up, II can be written as one fourth of a convex combination of point values of u^n under the CFL condition (2.29), which implies $\frac{1}{4}m \le \mathrm{II} \le \frac{1}{4}M$ since $m \le u^n \le M$. The similar arguments apply to III and IV

Since $\bar{u}_{i,j}^{n+1} = \mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}$, we finish the proof by summing up the inequalities of I, II, III and IV. □

2.3 Positivity-preserving for the Euler equations

2.3.1 The Euler equations in one dimension

Consider the Euler equations (2.2). Direct computation gives the expressions of ρ_t, ρ_{tt} and ρ_{ttt} as follows:

$$\rho_t = -(\rho u)_x, \tag{2.31}$$

$$\rho_{tt} = \left((\rho u^2)_x + \hat{\gamma}(\rho e)_x\right)_x, \tag{2.32}$$

$$\begin{aligned}\rho_{ttt} = &- \big(u_{xx}(\rho u^2) + 2u_x(\rho u^2)_x + u(\rho u^2)_{xx} \\ &+\hat{\gamma}\gamma u_{xx}(\rho e) + \hat{\gamma}(3+\gamma)u_x(\rho e)_x + 3\hat{\gamma}u(\rho e)_{xx}\big)_x\end{aligned} \tag{2.33}$$

where $\hat{\gamma} = \gamma - 1$. Moreover,

$$m_t = A^1_x, \quad m_{tt} = A^2_x, \quad m_{ttt} = A^3_x,$$

and

$$E_t = B^1_x, \quad E_{tt} = B^2_x, \quad E_{ttt} = B^3_x,$$

where $A^1, A^2, A^3, B^1, B^2, B^3$ are shorthand notations introduced for convenience of later discussion. For the full expressions of m_t, m_{tt}, m_{ttt}, and E_t, E_{tt}, E_{ttt}, see Appendix A.2.1.

The positivity-preserving LWDG scheme of (2.2) for ρ at time level t^n is to find $\rho^{n+1} \in V$, s.t. $\forall \xi \in V$, the equation

$$\begin{aligned}(\rho^{n+1}, \xi)_{I_j} =&(\rho, \xi)_{I_j} + \Delta t(\rho u, \xi_x)_{I_j} - \frac{\Delta t^2}{2}((\rho u^2)_x + \hat{\gamma}(\rho e)_x, \xi_x)_{I_j} \\ &+ \frac{\Delta t^3}{6}\big(u_{xx}(\rho u^2) + 2u_x(\rho u^2)_x + u(\rho u^2)_{xx} \\ &+\hat{\gamma}\gamma u_{xx}(\rho e) + \hat{\gamma}(3+\gamma)u_x(\rho e)_x + 3\hat{\gamma}u(\rho e)_{xx}, \xi_x\big)_{I_j} \\ &- \Delta t\hat{F}_{j+\frac{1}{2}}\xi^-_{j+\frac{1}{2}} + \Delta t\hat{F}_{j-\frac{1}{2}}\xi^+_{j-\frac{1}{2}},\end{aligned} \tag{2.34}$$

holds for $j = 1, 2, \ldots, N$. In the scheme, $\hat{F}_{j+\frac{1}{2}}$ is the numerical flux of ρ at $x_{j+\frac{1}{2}}$ defined as

$$\begin{aligned}\hat{F}_{j+\frac{1}{2}} =&\hat{f}^{\mathrm{LF}}_{j+\frac{1}{2}} - \frac{\Delta t}{2}\widehat{(\mathcal{I}(\rho u^2))}^{\mathrm{DDG}}_{xj+\frac{1}{2}} - \frac{\Delta t}{2}\hat{\gamma}\widehat{(\mathcal{I}(\rho e))}^{\mathrm{DDG}}_{xj+\frac{1}{2}} \\ &+ \frac{\Delta t^2}{6}\{u_{xx}(\rho u^2) + 2u_x\left(\mathcal{I}(\rho u^2)\right)_x + u\left(\mathcal{I}(\rho u^2)\right)_{xx}\}_{j+\frac{1}{2}} \\ &+ \frac{\Delta t^2}{6}\{\hat{\gamma}\gamma u_{xx}(\rho e) + \hat{\gamma}(3+\gamma)u_x\left(\mathcal{I}(\rho e)\right)_x + 3\hat{\gamma}u\left(\mathcal{I}(\rho e)\right)_{xx}\}_{j+\frac{1}{2}}\end{aligned} \quad , \tag{2.35}$$

where

$$\hat{f}^{\mathrm{LF}}_{j+\frac{1}{2}} = \{\rho u\}_{j+\frac{1}{2}} - \frac{1}{2}\alpha[\rho]_{j+\frac{1}{2}}, \quad \alpha = ||(|u| + c)||_{\infty}, \tag{2.36}$$

is the Lax-Friedriches flux used in the positivity-preserving for the Euler equations in [94], $c = \sqrt{\frac{\gamma p}{\rho}}$ is the sound speed, $(\widehat{\mathcal{I}(\rho u^2)})^{\mathrm{DDG}}_{xj+\frac{1}{2}}$ and $(\widehat{\mathcal{I}(\rho e)})^{\mathrm{DDG}}_{xj+\frac{1}{2}}$ are the DDG fluxes defined in (2.11), with u replaced by $\mathcal{I}(\rho u^2)$ and $\mathcal{I}(\rho e)$, respectively, where $\mathcal{I}$ is the quadratic interpolation operator with interpolation points at $x^+_{j-\frac{1}{2}}, x_j$, and $x^-_{j+\frac{1}{2}}$ on I_j, in order to get the similar expansions of the DDG flux as in (2.13).

The variables m and E are discretized by the standard discontinuous Galerkin method with the first order flux terms adopting the Lax-Friedriches flux and high-order flux terms adopting the average flux, i.e.

$$\begin{aligned}
(m^{n+1}, \xi)_{I_j} =& (m, \xi)_{I_j} - \Delta t(A^1, \xi_x)_{I_j} \\
&+ \Delta t\{A^1\}_{j+\frac{1}{2}} + \Delta t\alpha[m]_{j+\frac{1}{2}} \\
&- \Delta t\{A^1\}_{j-\frac{1}{2}} - \Delta t\alpha[m]_{j-\frac{1}{2}} \\
&- \frac{\Delta t^2}{2}(A^2, \xi_x)_{I_j} + \frac{\Delta t^2}{2}\{A^2\}_{j+\frac{1}{2}}\xi^-_{j+\frac{1}{2}} - \frac{\Delta t^2}{2}\{A^2\}_{j-\frac{1}{2}}\xi^+_{j-\frac{1}{2}} \\
&- \frac{\Delta t^3}{6}(A^3, \xi_x)_{I_j} + \frac{\Delta t^3}{6}\{A^3\}_{j+\frac{1}{2}}\xi^-_{j+\frac{1}{2}} - \frac{\Delta t^3}{6}\{A^3\}_{j-\frac{1}{2}}\xi^+_{j-\frac{1}{2}}
\end{aligned} \tag{2.37}$$

$$\begin{aligned}
(E^{n+1}, \xi)_{I_j} =& (E, \xi)_{I_j} - \Delta t(B^1, \xi_x)_{I_j} \\
&+ \Delta t\{B^1\}_{j+\frac{1}{2}} + \Delta t\alpha[E]_{j+\frac{1}{2}} \\
&- \Delta t\{B^1\}_{j-\frac{1}{2}} - \Delta t\alpha[E]_{j-\frac{1}{2}} \\
&- \frac{\Delta t^2}{2}(B^2, \xi_x)_{I_j} + \frac{\Delta t^2}{2}\{B^2\}_{j+\frac{1}{2}}\xi^-_{j+\frac{1}{2}} - \frac{\Delta t^2}{2}\{B^2\}_{j-\frac{1}{2}}\xi^+_{j-\frac{1}{2}} \\
&- \frac{\Delta t^3}{6}(B^3, \xi_x)_{I_j} + \frac{\Delta t^3}{6}\{B^3\}_{j+\frac{1}{2}}\xi^-_{j+\frac{1}{2}} - \frac{\Delta t^3}{6}\{B^3\}_{j-\frac{1}{2}}\xi^+_{j-\frac{1}{2}}
\end{aligned} \tag{2.38}$$

We now state the result for the positivity-preserving of $\bar{\rho}_j^{n+1}$.

Theorem 2.3.1. *Given $\mathbf{u}^n \in G$, the cell averages $\bar{\rho}_j^{n+1}, j = 1, \ldots, N$ of the solution*

of scheme (2.34) *are nonnegative under the CFL condition* (2.39)*:*

$$\lambda \leq \min\{q_1, q_2, \ldots, q_{11}\}, \tag{2.39}$$

where $q_1 = \frac{\hat{\omega}_1}{2||(|u|+c)||_\infty}$, $q_2 = \frac{6(\beta_0-\frac{3}{2}+4\beta_1)}{\Delta x^2||u_{xx}||_\infty+6\Delta x||u_x||_\infty+4||u||_\infty}$, $q_3 = \frac{3(2-8\beta_1)}{4(\Delta x||u_x||_\infty+||u||_\infty)}$, $q_4 = \frac{3(4\beta_1-\frac{1}{2})}{\Delta x||u_x||_\infty+2||u||_\infty}$, $q_5 = \frac{1}{2||u||_\infty}\left(\frac{\omega_1}{\beta_0-2+8\beta_1}\right)^{\frac{1}{2}}$, $q_6 = \frac{1}{2||u||_\infty}\left(\frac{\omega_{N_q}}{2(2-8\beta_1)}\right)^{\frac{1}{2}}$, $q_7 = \frac{6(4\beta_1-\frac{1}{2})}{(3+\gamma)\Delta x||u_x||_\infty+12||u||_\infty}$, $q_8 = \frac{3(2-8\beta_1)}{2(3+\gamma)\Delta x||u_x||_\infty+12||u||_\infty}$, $q_9 = \frac{6(\beta_0-\frac{3}{2}+4\beta_1)}{\gamma\Delta x^2||u_{xx}||_\infty+3(3+\gamma)\Delta x||u_x||_\infty+12||u||_\infty}$, $q_{10} = \left(\frac{\hat{\omega}_1}{4\hat{\gamma}(\beta_0-2+8\beta_1)||e||_\infty}\right)^{\frac{1}{2}}$, $q_{11} = \left(\frac{\hat{\omega}_{N_q}}{8\hat{\gamma}(2-8\beta_1)||e||_\infty}\right)^{\frac{1}{2}}$.

Proof. Take $\xi = 1$ on I_j and zero on other cells in the scheme (2.34), we obtain

$$\bar{\rho}_j^{n+1} = \mathrm{I} + \mathrm{II} + \mathrm{III}, \tag{2.40}$$

where

$$\mathrm{I} = \frac{1}{2}\left(\bar{\rho}_j^n - 2\lambda\hat{f}^{\mathrm{LF}}_{j+\frac{1}{2}} + 2\lambda\hat{f}^{\mathrm{LF}}_{j-\frac{1}{2}}\right)$$

$$\begin{aligned}\mathrm{II} = \frac{1}{4}\bar{\rho}^n - \lambda\Bigg(&-\frac{\Delta t}{2}\widehat{(\mathcal{I}(\rho u^2))}^{\mathrm{DDG}}_{xj+\frac{1}{2}} \\ &+\frac{\Delta t^2}{6}\left\{u_{xx}(\rho u^2) + 2u_x\left(\mathcal{I}(\rho u^2)\right)_x + u\left(\mathcal{I}(\rho u^2)\right)_{xx}\right\}_{j+\frac{1}{2}}\Bigg) \\ &+\lambda\Bigg(-\frac{\Delta t}{2}\widehat{(\mathcal{I}(\rho u^2))}^{\mathrm{DDG}}_{xj-\frac{1}{2}} \\ &+\frac{\Delta t^2}{6}\left\{u_{xx}(\rho u^2) + 2u_x\left(\mathcal{I}(\rho u^2)\right)_x + u\left(\mathcal{I}(\rho u^2)\right)_{xx}\right\}_{j-\frac{1}{2}}\Bigg)\end{aligned}$$

$$\begin{aligned}\mathrm{III} = \frac{1}{4}\bar{\rho}^n - \lambda \Bigg(& -\frac{\Delta t}{2}\hat{\gamma}\widehat{(\mathcal{I}(\rho e))}^{\mathrm{DDG}}_{x j+\frac{1}{2}} \\ & + \frac{\Delta t^2}{6}\left\{\hat{\gamma}\gamma u_{xx}(\rho e) + \hat{\gamma}(3+\gamma)u_x\left(\mathcal{I}(\rho e)\right)_x + 3\hat{\gamma}u\left(\mathcal{I}(\rho e)\right)_{xx}\right\}_{j+\frac{1}{2}}\Bigg) \\ & + \lambda\Bigg(-\frac{\Delta t}{2}\hat{\gamma}\widehat{(\mathcal{I}(\rho e))}^{\mathrm{DDG}}_{x j-\frac{1}{2}} \\ & + \frac{\Delta t^2}{6}\left\{\hat{\gamma}\gamma u_{xx}(\rho e) + \hat{\gamma}(3+\gamma)u_x\left(\mathcal{I}(\rho e)\right)_x + 3\hat{\gamma}u\left(\mathcal{I}(\rho e)\right)_{xx}\right\}_{j-\frac{1}{2}}\Bigg)\end{aligned}$$

Since I has exactly the same form as in [94], $I \geq 0$ is guaranteed under the condition $\lambda \leq q_1$ from the conclusion therein. Now we expand II as follows,

$$\begin{aligned}\mathrm{II} = & \frac{1}{4}\sum_{\gamma=2}^{N_q-1}\hat{\omega}_\gamma\hat{\rho}^\gamma + \frac{1}{4}\sum_{\gamma=N_q+1}^{2N_q-2}\hat{\omega}_\gamma\hat{\rho}^\gamma + z_1\rho^+_{j-\frac{3}{2}} + z_2\rho_{j-1} + z_3\rho^-_{j-\frac{1}{2}} + z_4\rho^+_{j-\frac{1}{2}} \\ & + z_5\rho_j + z_6\rho^-_{j+\frac{1}{2}} + z_7\rho^+_{j+\frac{1}{2}} + z_8\rho_{j+1} + z_9\rho^-_{j+\frac{3}{2}},\end{aligned}$$

where

$$\begin{aligned}z_1 = & \lambda^2\left(\frac{1}{2}(4\beta_1 - \frac{1}{2}) + \frac{\Delta t}{6}\left(u_x\right)^-_{j-\frac{1}{2}} + \frac{\lambda}{3}u^-_{j-\frac{1}{2}}\right)\left(u^+_{j-\frac{3}{2}}\right)^2 \\ z_2 = & \lambda^2\left(\frac{1}{2}(2 - 8\beta_1) - \frac{2\Delta t}{3}\left(u_x\right)^-_{j-\frac{1}{2}} - \frac{2\lambda}{3}u^-_{j-\frac{1}{2}}\right)\left(u_{j-1}\right)^2 \\ z_3 = & \lambda^2\left(\frac{1}{2}(\beta_0 - \frac{3}{2} + 4\beta_1) + \frac{\Delta t^2}{12\lambda}\left(u_{xx}\right)^-_{j-\frac{1}{2}} + \frac{\Delta t}{2}\left(u_x\right)^-_{j-\frac{1}{2}} + \frac{\lambda}{3}\left(u^-_{j-\frac{1}{2}}\right)\right)\left(u^-_{j-\frac{1}{2}}\right)^2 \\ z_4 = & \frac{1}{4}\hat{\omega}_1 - \lambda^2\left(\frac{1}{2}(4\beta_1 - \frac{1}{2}) + \frac{\Delta t}{6}\left(u_x\right)^-_{j+\frac{1}{2}} + \frac{\lambda}{3}u^-_{j+\frac{1}{2}}\right. \\ & \left. + \frac{1}{2}(\beta_0 - \frac{3}{2} + 4\beta_1) - \frac{\Delta t^2}{12\lambda}\left(u_{xx}\right)^+_{j-\frac{1}{2}} + \frac{\Delta t}{2}\left(u_x\right)^+_{j-\frac{1}{2}} - \frac{\lambda}{3}\left(u^+_{j-\frac{1}{2}}\right)\right)\left(u^+_{j-\frac{1}{2}}\right)^2 \\ z_5 = & \frac{1}{4}\hat{\omega}_{N_q} - \lambda^2\left(\frac{1}{2}(2 - 8\beta_1) - \frac{2\Delta t}{3}\left(u_x\right)^-_{j+\frac{1}{2}} - \frac{2}{3}\lambda u^-_{j+\frac{1}{2}}\right. \\ & \left. + \frac{1}{2}(2 - 8\beta_1) - \frac{2\Delta t}{3}\left(u_x\right)^+_{j-\frac{1}{2}} + \frac{2}{3}\lambda u^+_{j-\frac{1}{2}}\right)\left(u_j\right)^2\end{aligned}$$

$$
\begin{aligned}
z_6 =& \frac{1}{4}\hat{\omega}_{2N_q-1} - \lambda^2\left(\frac{1}{2}(\beta_0 - \frac{3}{2} + 4\beta_1) + \frac{1}{2}(4\beta_1 - \frac{1}{2}) + \frac{\Delta t^2}{12\lambda}\left(u_{xx}\right)^-_{j+\frac{1}{2}}\right.\\
&\left. + \frac{\Delta t}{2}\left(u_x\right)^-_{j+\frac{1}{2}} + \frac{\Delta t}{6}\left(u_x\right)^+_{j-\frac{1}{2}} + \frac{\lambda}{3}\left(u^-_{j+\frac{1}{2}}\right) - \frac{\lambda}{3}u^+_{j-\frac{1}{2}}\right)\left(u^-_{j+\frac{1}{2}}\right)^2\\
z_7 =& \lambda^2\left(\frac{1}{2}(\beta_0 - \frac{3}{2} + 4\beta_1) - \frac{\Delta t^2}{12\lambda}\left(u_{xx}\right)^+_{j+\frac{1}{2}} + \frac{\Delta t}{2}\left(u_x\right)^+_{j+\frac{1}{2}} - \frac{\lambda}{3}\left(u^+_{j+\frac{1}{2}}\right)\right)\left(u^+_{j+\frac{1}{2}}\right)^2\\
z_8 =& \lambda^2\left(\frac{1}{2}(2 - 8\beta_1) - \frac{2\Delta t}{3}\left(u_x\right)^+_{j+\frac{1}{2}} + \frac{2\lambda}{3}u^+_{j+\frac{1}{2}}\right)\left(u_{j+1}\right)^2\\
z_9 =& \lambda^2\left(\frac{1}{2}(4\beta_1 - \frac{1}{2}) + \frac{\Delta t}{6}\left(u_x\right)^+_{j+\frac{1}{2}} - \frac{\lambda}{3}u^+_{j+\frac{1}{2}}\right)\left(u^-_{j+\frac{3}{2}}\right)^2
\end{aligned}
$$

We claim that $z_1, z_2, \ldots, z_9 \geq 0$ under the CFL condition $\lambda \leq \min\{q_2, q_3, \ldots, q_6\}$. In fact, we have the following estimates

$$
\begin{aligned}
z_1 \geq& \lambda^2\left(\frac{1}{2}(4\beta_1 - \frac{1}{2}) - \frac{\Delta t}{6}||u_x||_\infty - \frac{\lambda}{3}||u||_\infty\right)\left(u^+_{j-\frac{3}{2}}\right)^2 \geq 0,\\
z_2 \geq& \lambda^2\left(\frac{1}{2}(2 - 8\beta_1) - \frac{2\Delta t}{3}||u_x||_\infty - \frac{2\lambda}{3}||u||_\infty\right)\left(u_{j-1}\right)^2 \geq 0,\\
z_3 \geq& \lambda^2\left(\frac{1}{2}(\beta_0 - \frac{3}{2} + 4\beta_1) - \frac{\Delta t^2}{12\lambda}||u_{xx}||_\infty - \frac{\Delta t}{2}||u_x||_\infty - \frac{\lambda}{3}||u||_\infty\right)\left(u^-_{j-\frac{1}{2}}\right)^2 \geq 0,\\
z_4 \geq& \frac{1}{4}\hat{\omega}_1 - \lambda^2\left(\frac{1}{2}(\beta_0 - 2 + 8\beta_1) + \frac{\Delta t^2}{12\lambda}||u_{xx}||_\infty + \frac{2\Delta t}{3}||u_x||_\infty + \frac{2\lambda}{3}||u||_\infty\right)||u||^2_\infty \geq 0,\\
z_5 \geq& \frac{1}{4}\hat{\omega}_{N_q} - \lambda^2\left((2 - 8\beta_1) + \frac{4\Delta t}{3}||u_x||_\infty + \frac{4}{3}\lambda||u||_\infty\right)||u||^2_\infty \geq 0,\\
z_6 \geq& \frac{1}{4}\hat{\omega}_{2N_q-1} - \lambda^2\left(\frac{1}{2}(\beta_0 - 2 + 8\beta_1) + \frac{\Delta t^2}{12\lambda}||u_{xx}||_\infty + \frac{2\Delta t}{3}||u_x||_\infty + \frac{2\lambda}{3}||u||_\infty\right)||u||^2_\infty \geq 0,\\
z_7 \geq& \lambda^2\left(\frac{1}{2}(\beta_0 - \frac{3}{2} + 4\beta_1) - \frac{\Delta t^2}{12\lambda}||u_{xx}||_\infty - \frac{\Delta t}{2}||u_x||_\infty - \frac{\lambda}{3}||u||_\infty\right)\left(u^+_{j+\frac{1}{2}}\right)^2 \geq 0,\\
z_8 \geq& \lambda^2\left(\frac{1}{2}(2 - 8\beta_1) - \frac{2\Delta t}{3}||u_x||_\infty - \frac{2\lambda}{3}||u||_\infty\right)\left(u_{j+1}\right)^2 \geq 0,\\
z_9 \geq& \lambda^2\left(\frac{1}{2}(4\beta_1 - \frac{1}{2}) - \frac{\Delta t}{6}||u_x||_\infty - \frac{\lambda}{3}||u||_\infty\right)\left(u^-_{j+\frac{3}{2}}\right)^2 \geq 0,
\end{aligned}
$$

Similarly, we can expand III as

$$\begin{aligned}\text{III} =&\frac{1}{4}\sum_{\gamma=2}^{N_q-1}\hat{\omega}_\gamma\hat{\rho}^\gamma+\frac{1}{4}\sum_{\gamma=N_q+1}^{2N_q-2}\hat{\omega}_\gamma\hat{\rho}^\gamma+z_{10}\rho^+_{j-\frac{3}{2}}+z_{11}\rho_{j-1}+z_{12}\rho^-_{j-\frac{1}{2}}+z_{13}\rho^+_{j-\frac{1}{2}}\\&+z_{14}\rho_j+z_{15}\rho^-_{j+\frac{1}{2}}+z_{16}\rho^+_{j+\frac{1}{2}}+z_{17}\rho_{j+1}+z_{18}\rho^-_{j+\frac{3}{2}},\end{aligned}\tag{2.41}$$

and $z_{10},\ldots,z_{18}\geq 0$ under the condition $\lambda\leq\min\{q_7,q_8,q_9,q_{10},q_{11}\}$. The expressions and estimates of $z_{10},\ldots,z_{18}$ are similar to those of $z_1,\ldots,z_9$, thus are given in Appendix A.1.3.

By the same arguments as in the scalar cases, we have $\text{II},\text{III}\geq 0$, provided the positivity of ρ^n. Since $\bar{\rho}_j^{n+1}=\text{I}+\text{II}+\text{III}$, we finish the proof by collecting the results for I, II and III. □

The remaining task is to preserve the positivity of internal energy of cell averages of the solution, i.e. $e(\bar{\mathbf{u}}_j^{n+1})\geq 0$. We have the results as follows.

Theorem 2.3.2. *Given* $\mathbf{u}^n\in G$*, the specific internal energy of the cell averages* $e(\bar{\mathbf{u}}_j^{n+1}), j=1,\ldots,N$ *of scheme* (2.34)*,* (2.37) *and* (2.38) *are nonnegative under the CFL condition* (2.42)*:*

$$\lambda\leq\frac{\gamma+1}{2\alpha^2(\gamma-1)}\min_j\left\{\frac{(p^-_{j+\frac{1}{2}})^2}{C^-_{j+\frac{1}{2}}},\frac{(p^+_{j+\frac{1}{2}})^2}{C^+_{j+\frac{1}{2}}}\right\},\tag{2.42}$$

where

$$
\begin{aligned}
C^-_{j+\frac{1}{2}} =& \frac{\Delta x}{\alpha}\left((2E^-_{j+\frac{1}{2}}+p^-_{j+\frac{1}{2}})\left(|\tilde{f}^1_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^1_{j+\frac{1}{2}}|\right)+2\rho^-_{j+\frac{1}{2}}\left(|\tilde{f}^3_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^3_{j+\frac{1}{2}}|\right)\right.\\
&+Q_1\frac{\Delta x}{\alpha}\left(|\tilde{f}^1_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^1_{j+\frac{1}{2}}|\right)\left(|\tilde{f}^3_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^3_{j+\frac{1}{2}}|\right)\\
&+\frac{1}{2}Q_1\frac{\Delta x}{\alpha}\left(|\tilde{f}^2_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^2_{j+\frac{1}{2}}|\right)^2\\
&\left.+(2|m^-_{j+\frac{1}{2}}|+\frac{p^-_{j+\frac{1}{2}}}{\alpha})\left(|\tilde{f}^2_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^2_{j+\frac{1}{2}}|\right)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
C^+_{j+\frac{1}{2}} =& \frac{\Delta x}{\alpha}\left((2E^+_{j+\frac{1}{2}}+p^+_{j+\frac{1}{2}})\left(|\tilde{f}^1_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^1_{j+\frac{1}{2}}|\right)+2\rho^+_{j+\frac{1}{2}}\left(|\tilde{f}^3_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^3_{j+\frac{1}{2}}|\right)\right.\\
&+Q_1\frac{\Delta x}{\alpha}\left(|\tilde{f}^1_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^1_{j+\frac{1}{2}}|\right)\left(|\tilde{f}^3_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^3_{j+\frac{1}{2}}|\right)\\
&+\frac{1}{2}Q_1\frac{\Delta x}{\alpha}\left(|\tilde{f}^2_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^2_{j+\frac{1}{2}}|\right)^2\\
&\left.+(2|m^+_{j+\frac{1}{2}}|+\frac{p^+_{j+\frac{1}{2}}}{\alpha})\left(|\tilde{f}^2_{j+\frac{1}{2}}|+Q_1\Delta x|\check{f}^2_{j+\frac{1}{2}}|\right)\right)
\end{aligned}
$$

Proof. Take $\xi = 1$ on I_j and zero anywhere else in the scheme (2.34),(2.37) and (2.38), we can obtain the following vector equation satisfied by the cell average of $\mathbf{u}^{n+1}$ on I_j,

$$
\bar{\mathbf{u}}^{n+1}_j = \bar{\mathbf{u}}^n_j - \lambda\left(\hat{\mathbf{f}}^{\mathrm{LF}}_{j+\frac{1}{2}}+\Delta t\tilde{\mathbf{f}}_{j+\frac{1}{2}}+\Delta t^2\check{\mathbf{f}}_{j+\frac{1}{2}}\right)+\lambda\left(\hat{\mathbf{f}}^{\mathrm{LF}}_{j-\frac{1}{2}}+\Delta t\tilde{\mathbf{f}}_{j-\frac{1}{2}}+\Delta t^2\check{\mathbf{f}}_{j-\frac{1}{2}}\right),
$$

where $\hat{\mathbf{f}}^{\mathrm{LF}}_{j+\frac{1}{2}} = \frac{1}{2}\left(\mathbf{f}(\mathbf{u}^-_{j+\frac{1}{2}})+\mathbf{f}(\mathbf{u}^+_{j+\frac{1}{2}})-\alpha\left(\mathbf{u}^+_{j+\frac{1}{2}}-\mathbf{u}^-_{j+\frac{1}{2}}\right)\right), \alpha = ||(|u|+c)||_\infty$, is the standard Lax-Friedriches flux, which is the leading term in the total flux constructed in the LWDG scheme (2.34)-(2.38), $\tilde{\mathbf{f}}_{j+\frac{1}{2}} = (\tilde{f}^1_{j+\frac{1}{2}}, \tilde{f}^2_{j+\frac{1}{2}}, \tilde{f}^3_{j+\frac{1}{2}})$ and $\check{\mathbf{f}}_{j+\frac{1}{2}} = (\check{f}^1_{j+\frac{1}{2}}, \check{f}^2_{j+\frac{1}{2}}, \check{f}^3_{j+\frac{1}{2}})$ are the remaining second and third order terms contained in the flux of (2.35), (2.37) and (2.38), in which the abbreviated terms can be found in Appendix A.2.1, respectively.

Similar to [94], we have the decomposition

$$\begin{aligned}\bar{\mathbf{u}}_j^{n+1} = &\sum_{\gamma=2}^{2N_q-2} \hat{\omega}_\gamma \mathbf{u}^\gamma + \hat{\omega}_1 \left(1 - \frac{\alpha\lambda}{\hat{\omega}_1}\right) \mathbf{u}^+_{j-\frac{1}{2}} + \hat{\omega}_{2N_q-1} \left(1 - \frac{\alpha\lambda}{\hat{\omega}_{2N_q-1}}\right) \mathbf{u}^-_{j+\frac{1}{2}} \\ &+ \frac{\alpha\lambda}{2}\left(\mathbf{u}^-_{j+\frac{1}{2}} - \frac{1}{\alpha}\mathbf{f}(\mathbf{u}^-_{j+\frac{1}{2}}) - \frac{\Delta t}{\alpha}\left(\tilde{\mathbf{f}}_{j+\frac{1}{2}} + \Delta t \check{\mathbf{f}}_{j+\frac{1}{2}}\right)\right) \\ &+ \frac{\alpha\lambda}{2}\left(\mathbf{u}^+_{j+\frac{1}{2}} - \frac{1}{\alpha}\mathbf{f}(\mathbf{u}^+_{j+\frac{1}{2}}) - \frac{\Delta t}{\alpha}\left(\tilde{\mathbf{f}}_{j+\frac{1}{2}} + \Delta t \check{\mathbf{f}}_{j+\frac{1}{2}}\right)\right) \\ &+ \frac{\alpha\lambda}{2}\left(\mathbf{u}^-_{j-\frac{1}{2}} + \frac{1}{\alpha}\mathbf{f}(\mathbf{u}^-_{j-\frac{1}{2}}) + \frac{\Delta t}{\alpha}\left(\tilde{\mathbf{f}}_{j-\frac{1}{2}} + \Delta t \check{\mathbf{f}}_{j-\frac{1}{2}}\right)\right) \\ &+ \frac{\alpha\lambda}{2}\left(\mathbf{u}^+_{j-\frac{1}{2}} + \frac{1}{\alpha}\mathbf{f}(\mathbf{u}^+_{j-\frac{1}{2}}) + \frac{\Delta t}{\alpha}\left(\tilde{\mathbf{f}}_{j-\frac{1}{2}} + \Delta t \check{\mathbf{f}}_{j-\frac{1}{2}}\right)\right)\end{aligned}$$

Since $\hat{\omega}_\gamma \geq 0, \gamma = 1, \ldots, 2N_q - 1$ and $(1 - \frac{\alpha\lambda}{\hat{\omega}_1}), (1 - \frac{\alpha\lambda}{\hat{\omega}_{2N_q-1}}) \geq 0$ from the CFL condition (2.39), by convexity of G, it suffices to show

$$\mathbf{u}^+_{j+\frac{1}{2}} \pm \frac{1}{\alpha}\mathbf{f}(\mathbf{u}^+_{j+\frac{1}{2}}) \pm \frac{\Delta t}{\alpha}\left(\tilde{\mathbf{f}}_{j+\frac{1}{2}} + \Delta t \check{\mathbf{f}}_{j+\frac{1}{2}}\right) \in G,$$

provided $\mathbf{u}^+_{j+\frac{1}{2}} \in G$. For simplicity, we omit the superscripts and subscripts in the following proof.

Using the equality $\rho^2 e = \rho E - \frac{1}{2}m^2$, one can calculate that

$$\begin{aligned}&\rho^2 e\left(\mathbf{u} \pm \frac{1}{\alpha}\mathbf{f}(\mathbf{u}) \pm \frac{\Delta t}{\alpha}\left(\tilde{\mathbf{f}} + \Delta t \check{\mathbf{f}}\right)\right) \\ =&\frac{p\rho}{\alpha^2(\gamma-1)}\left((\alpha \pm u)^2 - \frac{\gamma-1}{2\gamma}c^2\right) \pm \frac{\Delta t}{\alpha}(\tilde{f}^1 + \Delta t \check{f}^1)\left((1 \pm \frac{u}{\alpha})E \pm \frac{u}{\alpha}p\right) \\ &\pm \frac{\Delta t}{\alpha}(\tilde{f}^3 + \Delta t \check{f}^3)\left((1 \pm \frac{u}{\alpha})\rho\right) + \frac{\Delta t^2}{\alpha^2}(\tilde{f}^1 + \Delta t \check{f}^1)(\tilde{f}^3 + \Delta t \check{f}^3) - \frac{1}{2}\frac{\Delta t^2}{\alpha^2}(\tilde{f}^2 + \Delta t \check{f}^2)^2 \\ &\mp \frac{\Delta t}{\alpha}(\tilde{f}^2 + \Delta t \check{f}^2)\left((1 \pm \frac{u}{\alpha})m \pm \frac{1}{\alpha}p\right) \\ \geq&\frac{\gamma+1}{2\alpha^2(\gamma-1)}p^2 - C\lambda,\end{aligned}$$

where

$$\begin{aligned} C =& \frac{\Delta x}{\alpha}\left((2E+p)\left(|\tilde{f}^1|+Q_1\Delta x|\check{f}^1|\right)+2\rho\left(|\tilde{f}^3|+Q_1\Delta x|\check{f}^3|\right)\right.\\ &+Q_1\frac{\Delta x}{\alpha}\left(|\tilde{f}^1|+Q_1\Delta x|\check{f}^1|\right)\left(|\tilde{f}^3|+Q_1\Delta x|\check{f}^3|\right)\\ &+\frac{1}{2}Q_1\frac{\Delta x}{\alpha}\left(|\tilde{f}^2|+Q_1\Delta x|\check{f}^2|\right)^2\\ &\left.+(2|m|+\frac{p}{\alpha})\left(|\tilde{f}^2|+Q_1\Delta x|\check{f}^2|\right)\right). \end{aligned}$$

Under the CFL condition (2.42), we can get the positivity of $\rho^2 e$, which finishes the proof. □

Collecting the above two theorems, we reach our final result.

Theorem 2.3.3. *Given* $\mathbf{u}^n \in G$, *we have* $\bar{\mathbf{u}}_j^{n+1} \in G, j = 1, \ldots, N$ *for scheme* (2.34), (2.37) *and* (2.38), *under the CFL conditions* (2.39) *and* (2.42).

2.3.2 The Euler equations in two dimensions

Consider the Euler equations in two space dimensions

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x + \mathbf{g}(\mathbf{u})_y = \mathbf{0}, \tag{2.43}$$

where

$$\mathbf{u} = \begin{pmatrix} \rho \\ m \\ n \\ E \end{pmatrix}, \quad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (E+p)u \end{pmatrix}, \quad \mathbf{g}(\mathbf{u}) = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ (E+p)v \end{pmatrix},$$

with

$$m = \rho u, \quad n = \rho v, \quad E = \frac{1}{2}\rho u^2 + \frac{1}{2}\rho v^2 + \rho e, \quad p = (\gamma - 1)\rho e,$$

in which u and v are velocities in x and y directions, respectively, and m and n are momentums in x and y directions, respectively.

Direct computation gives the expressions of ρ_t, ρ_{tt} and ρ_{ttt} as follows:

$$\rho_t = -(\rho u)_x - (\rho v)_y, \tag{2.44}$$

$$\rho_{tt} = \left((\rho u^2)_x + \hat{\gamma}(\rho e)_x\right)_x + 2\left(\rho u v\right)_{xy} + \left((\rho v^2)_y + \hat{\gamma}(\rho e)_y\right)_y, \tag{2.45}$$

$$\begin{aligned}
\rho_{ttt} = &- \big(u_{xx}(\rho u^2) + 2u_x(\rho u^2)_x + u(\rho u^2)_{xx} \\
&+\hat{\gamma}\gamma u_{xx}(\rho e) + (\hat{\gamma}(3+\gamma)u_x + \hat{\gamma}^2 v_y)(\rho e)_x + 3\hat{\gamma}u(\rho e)_{xx}\big)_x \\
&- \big(v_{yy}(\rho v^2) + 2v_y(\rho v^2)_y + v(\rho v^2)_{yy} \\
&+\hat{\gamma}\gamma v_{yy}(\rho e) + (\hat{\gamma}(3+\gamma)v_y + \hat{\gamma}^2 u_x)(\rho e)_y + 3\hat{\gamma}v(\rho e)_{yy}\big)_y \\
&- \big(\big(\gamma\hat{\gamma}e v_{xx} + \hat{\gamma}(\gamma+3)e_x v_x + 6v u_x^2 + 12 u u_x v_x \\
&+3\hat{\gamma}v e_{xx} + 3u^2 v_{xx} + 6uv u_{xx} - \hat{\gamma}^2 u_y e_x\big)\,\rho\big)_y \\
&- \big(\big(6\hat{\gamma}v e_x + \hat{\gamma}(\gamma+3)e v_x + 6u(u v_x + 2v u_x) - \hat{\gamma}^2 u_y e\big)\,\rho_x\big)_y \\
&- \big((3(\hat{\gamma}e + u^2)v)\rho_{xx}\big)_y \\
&- \big(\big(\gamma\hat{\gamma}e u_{yy} + \hat{\gamma}(\gamma+3)e_y u_y + 6u v_y^2 + 12 v u_y v_y \\
&+3\hat{\gamma}u e_{yy} + 3v^2 u_{yy} + 6uv v_{yy} - \hat{\gamma}^2 v_x e_y\big)\,\rho\big)_x \\
&- \big(\big(6\hat{\gamma}u e_y + \hat{\gamma}(\gamma+3)e u_y + 6v(v u_y + 2u v_y) - \hat{\gamma}^2 v_x e\big)\,\rho_y\big)_x \\
&- \big(\big(3(\hat{\gamma}e + v^2)u\big)\,\rho_{yy}\big)_x
\end{aligned} \tag{2.46}$$

where $\hat{\gamma} = \gamma - 1$. Note that there are a lot of ways to expand ρ_{ttt}, among which we choose the one that avoids the appearance of mixed derivatives in the LWDG scheme.

Moreover,

$$m_t = B_x^1 + B_y^2, \quad m_{tt} = B_x^3 + B_y^4, \quad m_{ttt} = B_x^5 + B_y^6,$$

$$n_t = C_x^1 + C_y^2, \quad n_{tt} = C_x^3 + C_y^4, \quad n_{ttt} = C_x^5 + C_y^6,$$

and

$$E_t = D_x^1 + D_y^2, \quad E_{tt} = D_x^3 + D_y^4, \quad E_{ttt} = D_x^5 + D_y^6,$$

where $B^1, B^2, B^3, B^4, B^5, B^6, C^1, C^2, C^3, C^4, C^5, C^6, D^1, D^2, D^3, D^4, D^5, D^6$, are shorthand notations introduced for convenience of later discussion. For the full expressions of m_t, m_{tt}, m_{ttt}, n_t, n_{tt}, n_{ttt},and E_t, E_{tt}, E_{ttt}, see Appendix A.2.2.

The positivity-preserving LWDG of ρ at time level t^n is to find $\rho^{n+1} \in W$, s.t. $\forall \xi \in W$, the equation

$$\begin{aligned}(\rho^{n+1}, \xi)_{K_{i,j}} =& (\rho, \xi)_{K_{i,j}} + \Delta t(\rho u, \xi_x)_{K_{i,j}} + \Delta t(\rho v, \xi_y)_{K_{i,j}} \\ &- \frac{\Delta t^2}{2}((\rho u^2)_x + \hat{\gamma}(\rho e)_x + (\rho uv)_y, \xi_x)_{K_{i,j}} \\ &- \frac{\Delta t^2}{2}((\rho v^2)_y + \hat{\gamma}(\rho e)_y + (\rho uv)_x, \xi_y)_{K_{i,j}} \\ &- \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}_{i+\frac{1}{2},j}\ \xi(x_{i+\frac{1}{2}}^-, y)dy + \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}_{i-\frac{1}{2},j}\ \xi(x_{i-\frac{1}{2}}^+, y)dy \\ &- \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}_{i,j+\frac{1}{2}}\ \xi(x, y_{j+\frac{1}{2}}^-)dx + \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}_{i,j-\frac{1}{2}}\ \xi(x, y_{j-\frac{1}{2}}^+)dx\end{aligned} \tag{2.47}$$

holds for $i = 1, 2, \ldots, N_x, j = 1, 2, \ldots, N_y$. $\hat{F}_{i+\frac{1}{2},j}$ and $\hat{G}_{i,j+\frac{1}{2}}$ are numerical fluxes defined as

$$\hat{F}_{i+\frac{1}{2},j} = \hat{F}^0_{i+\frac{1}{2},j} + \hat{F}^1_{i+\frac{1}{2},j}, \tag{2.48}$$

and

$$\hat{G}_{i,j+\frac{1}{2}} = \hat{G}^0_{i,j+\frac{1}{2}} + \hat{G}^1_{i,j+\frac{1}{2}}, \tag{2.49}$$

where

$$\begin{aligned}\hat{F}^0_{i+\frac{1}{2},j} =& \{\rho u\}_{i+\frac{1}{2},j} - \frac{1}{2}\alpha^0_x[\rho]_{i+\frac{1}{2},j} - \frac{\Delta t}{2}\widehat{(\mathcal{I}(\rho u^2))}^{\text{DDG}}_{xi+\frac{1}{2},j} - \frac{\Delta t}{2}\hat{\gamma}\widehat{(\mathcal{I}(\rho e))}^{\text{DDG}}_{xi+\frac{1}{2},j} \\ &+ \frac{\Delta t^2}{6}\{u_{xx}(\rho u^2) + 2u_x\left(\mathcal{I}(\rho u^2)\right)_x + u\left(\mathcal{I}(\rho u^2)\right)_{xx}\}_{i+\frac{1}{2},j} \\ &+ \frac{\Delta t^2}{6}\{\hat{\gamma}\gamma u_{xx}(\rho e) + \left(\hat{\gamma}(3+\gamma)u_x + \hat{\gamma}^2 v_y\right)\left(\mathcal{I}(\rho e)\right)_x + 3\hat{\gamma}u\left(\mathcal{I}(\rho e)\right)_{xx}\}_{i+\frac{1}{2},j}\end{aligned} \tag{2.50}$$

$$\begin{aligned}\hat{F}^1_{i+\frac{1}{2},j} =& -\frac{1}{2}\alpha^1_x[\rho]_{i+\frac{1}{2},j} - \frac{\Delta t}{2}\{\rho_y uv + \rho(u_y v + u v_y)\} + \frac{\Delta t^2}{6}\left\{A^1\rho + A^2\rho_y + A^3\rho_{yy}\right\} \\ \hat{G}^0_{i,j+\frac{1}{2}} =& \{\rho v\}_{i,j+\frac{1}{2}} - \frac{1}{2}\alpha^0_y[\rho]_{i,j+\frac{1}{2}} - \frac{\Delta t}{2}\widehat{(\mathcal{I}(\rho v^2))}^{\text{DDG}}_{y i,j+\frac{1}{2}} - \frac{\Delta t}{2}\hat{\gamma}\widehat{(\mathcal{I}(\rho e))}^{\text{DDG}}_{y i,j+\frac{1}{2}} \\ &+ \frac{\Delta t^2}{6}\{v_{yy}(\rho v^2) + 2v_y\left(\mathcal{I}(\rho v^2)\right)_y + v\left(\mathcal{I}(\rho v^2)\right)_{yy}\}_{i,j+\frac{1}{2}} \\ &+ \frac{\Delta t^2}{6}\{\hat{\gamma}\gamma v_{yy}(\rho e) + \left(\hat{\gamma}(3+\gamma)v_y + \hat{\gamma}^2 u_x\right)\left(\mathcal{I}(\rho e)\right)_y + 3\hat{\gamma}v\left(\mathcal{I}(\rho e)\right)_{yy}\}_{i,j+\frac{1}{2}}\end{aligned} \tag{2.51}$$

$$\hat{G}^1_{i,j+\frac{1}{2}} = -\frac{1}{2}\alpha^1_y[\rho]_{i,j+\frac{1}{2}} - \frac{\Delta t}{2}\{\rho_x uv + \rho(uv_x + u_x v)\} + \frac{\Delta t^2}{6}\left\{A^4\rho + A^5\rho_x + A^6\rho_{xx}\right\}$$

in which $\alpha^0_x = ||(|u| + c)||_\infty$, $\alpha^0_y = ||(|v| + c)||_\infty$, $\alpha^1_x, \alpha^1_y > 0$, and

$$A^1 = \left(\gamma\hat{\gamma}eu_{yy} + \hat{\gamma}(\gamma+3)e_y u_y + 6uv_y^2 + 12vu_y v_y + 3\hat{\gamma}ue_{yy} + 3v^2 u_{yy} + 6uvv_{yy} - \hat{\gamma}^2 v_x e_y\right)$$

$$A^2 = \left(6\hat{\gamma}ue_y + \hat{\gamma}(\gamma+3)eu_y + 6v(vu_y + 2uv_y) - \hat{\gamma}^2 v_x e\right)$$

$$A^3 = \left(3(\hat{\gamma}e + v^2)u\right)$$

$$A^4 = \left(\gamma\hat{\gamma}ev_{xx} + \hat{\gamma}(\gamma+3)e_x v_x + 6vu_x^2 + 12uu_x v_x + 3\hat{\gamma}ve_{xx} + 3u^2 v_{xx} + 6uvu_{xx} - \hat{\gamma}^2 u_y e_x\right)$$

$$A^5 = \left(6\hat{\gamma}ve_x + \hat{\gamma}(\gamma+3)ev_x + 6u(uv_x + 2vu_x) - \hat{\gamma}^2 u_y e\right)$$

$$A^6 = (3(\hat{\gamma}e + u^2)v)$$

The variables m, n and E are discretized by the standard discontinuous Galerkin method with the first order flux terms adopting the Lax-Friedriches flux, in which the viscosity constant $\alpha_x = \alpha_x^0 + \alpha_x^1$ for the vertical cell interfaces and $\alpha_y = \alpha_y^0 + \alpha_y^1$ for the horizontal cell interfaces, and high-order flux terms adopting the average flux, i.e.

$$
\begin{aligned}
(m^{n+1},\xi)_{K_{i,j}} =& (m,\xi)_{K_{i,j}} - \Delta t(B^1,\xi_x)_{K_{i,j}} - \Delta t(B^2,\xi_y)_{K_{i,j}} \\
&- \frac{\Delta t^2}{2}(B^3,\xi_x)_{K_{i,j}} - \frac{\Delta t^2}{2}(B^4,\xi_y)_{K_{i,j}} \\
&- \frac{\Delta t^3}{6}(B^5,\xi_x)_{K_{i,j}} - \frac{\Delta t^3}{6}(B^6,\xi_y)_{K_{i,j}} \\
&+ \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\{B^1\} + \alpha_x[m] + \frac{\Delta t}{2}\{B^3\} + \frac{\Delta t^2}{6}\{B^5\} \right) \xi(x^-_{i+\frac{1}{2}}, y) dy \\
&- \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\{B^1\} + \alpha_x[m] + \frac{\Delta t}{2}\{B^3\} + \frac{\Delta t^2}{6}\{B^5\} \right) \xi(x^+_{i-\frac{1}{2}}, y) dy \\
&+ \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\{B^2\} + \alpha_y[m] + \frac{\Delta t}{2}\{B^4\} + \frac{\Delta t^2}{6}\{B^6\} \right) \xi(x, y^-_{j+\frac{1}{2}}) dx \\
&- \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\{B^2\} + \alpha_y[m] + \frac{\Delta t}{2}\{B^4\} + \frac{\Delta t^2}{6}\{B^6\} \right) \xi(x, y^+_{j-\frac{1}{2}}) dx
\end{aligned}
\tag{2.52}
$$

$$
\begin{aligned}
(n^{n+1},\xi)_{K_{i,j}} =& (n,\xi)_{K_{i,j}} - \Delta t(C^1,\xi_x)_{K_{i,j}} - \Delta t(C^2,\xi_y)_{K_{i,j}} \\
&- \frac{\Delta t^2}{2}(C^3,\xi_x)_{K_{i,j}} - \frac{\Delta t^2}{2}(C^4,\xi_y)_{K_{i,j}} \\
&- \frac{\Delta t^3}{6}(C^5,\xi_x)_{K_{i,j}} - \frac{\Delta t^3}{6}(C^6,\xi_y)_{K_{i,j}} \\
&+ \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\{C^1\} + \alpha_x[n] + \frac{\Delta t}{2}\{C^3\} + \frac{\Delta t^2}{6}\{C^5\} \right) \xi(x^-_{i+\frac{1}{2}}, y)dy \\
&- \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\{C^1\} + \alpha_x[n] + \frac{\Delta t}{2}\{C^3\} + \frac{\Delta t^2}{6}\{C^5\} \right) \xi(x^+_{i-\frac{1}{2}}, y)dy \\
&+ \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\{C^2\} + \alpha_y[n] + \frac{\Delta t}{2}\{C^4\} + \frac{\Delta t^2}{6}\{C^6\} \right) \xi(x, y^-_{j+\frac{1}{2}})dx \\
&- \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\{C^2\} + \alpha_y[n] + \frac{\Delta t}{2}\{C^4\} + \frac{\Delta t^2}{6}\{C^6\} \right) \xi(x, y^+_{j-\frac{1}{2}})dx
\end{aligned}
\tag{2.53}
$$

and

$$
\begin{aligned}
(E^{n+1},\xi)_{K_{i,j}} =& (E,\xi)_{K_{i,j}} - \Delta t(D^1,\xi_x)_{K_{i,j}} - \Delta t(D^2,\xi_y)_{K_{i,j}} \\
&- \frac{\Delta t^2}{2}(D^3,\xi_x)_{K_{i,j}} - \frac{\Delta t^2}{2}(D^4,\xi_y)_{K_{i,j}} \\
&- \frac{\Delta t^3}{6}(D^5,\xi_x)_{K_{i,j}} - \frac{\Delta t^3}{6}(D^6,\xi_y)_{K_{i,j}} \\
&+ \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\{D^1\} + \alpha_x[E] + \frac{\Delta t}{2}\{D^3\} + \frac{\Delta t^2}{6}\{D^5\} \right) \xi(x^-_{i+\frac{1}{2}}, y)dy \\
&- \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\{D^1\} + \alpha_x[E] + \frac{\Delta t}{2}\{D^3\} + \frac{\Delta t^2}{6}\{D^5\} \right) \xi(x^+_{i-\frac{1}{2}}, y)dy \\
&+ \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\{D^2\} + \alpha_y[E] + \frac{\Delta t}{2}\{D^4\} + \frac{\Delta t^2}{6}\{D^6\} \right) \xi(x, y^-_{j+\frac{1}{2}})dx \\
&- \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\{D^2\} + \alpha_y[E] + \frac{\Delta t}{2}\{D^4\} + \frac{\Delta t^2}{6}\{D^6\} \right) \xi(x, y^+_{j-\frac{1}{2}})dx
\end{aligned}
\tag{2.54}
$$

Similar to the one dimensional Euler equations, we have the results for positivity of $\bar{\rho}^{n+1}$ as follows.

Theorem 2.3.4. *Given $\mathbf{u}^n \in G$, the cell averages $\bar{\rho}_{i,j}^{n+1}, i = 1, \ldots, N_x, j = 1, \ldots, N_y$ of the solution of scheme* (2.47) *are nonnegative under the CFL condition* (2.55)*:*

$$\lambda_x \leq \min\{Q_1, Q_3\}, \quad \lambda_y \leq \min\{Q_2, Q_4\} \tag{2.55}$$

where the definitions of $Q_1, \ldots, Q_4$ are given in Appendix A.1.4.

Proof. Take $\xi = 1$ in $K_{i,j}$ and zero on other cells in (2.47), we obtain

$$\bar{\rho}_{i,j}^{n+1} = \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}, \tag{2.56}$$

where

$$\mathrm{I} = \frac{1}{4}\bar{\rho}_{i,j}^{n} - \lambda_x \frac{1}{\Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}^0_{i+\frac{1}{2},j}dy + \lambda_x \frac{1}{\Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}^0_{i-\frac{1}{2},j}dy,$$

$$\mathrm{II} = \frac{1}{4}\bar{\rho}_{i,j}^{n} - \lambda_x \frac{1}{\Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}^1_{i+\frac{1}{2},j}dy + \lambda_x \frac{1}{\Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{F}^1_{i-\frac{1}{2},j}dy,$$

$$\mathrm{III} = \frac{1}{4}\bar{\rho}_{i,j}^{n} - \lambda_y \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}^0_{i,j+\frac{1}{2}}dx + \lambda_y \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}^0_{i,j-\frac{1}{2}}dx$$

$$\mathrm{IV} = \frac{1}{4}\bar{\rho}_{i,j}^{n} - \lambda_y \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}^1_{i,j+\frac{1}{2}}dx + \lambda_y \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{G}^1_{i,j-\frac{1}{2}}dx$$

It suffices to show $\mathrm{I}, \mathrm{II} \geq 0$ under the CFL condition (2.55), due to the symmetry in the x and y directions.

One can observe that I can be decomposed in the form of convex combination

$$\mathrm{I} = \frac{1}{4}\sum_{\gamma=1}^{2N_q-1} \hat{\omega}_\gamma H_\gamma,$$

where

$$H_\gamma = \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \hat{\rho}^{\beta,\gamma} - 4\lambda_x \hat{F}^0_{i+\frac{1}{2},j}(x_{i+\frac{1}{2}}, \hat{y}_\gamma) + 4\lambda_x \hat{F}^0_{i-\frac{1}{2},j}(x_{i-\frac{1}{2}}, \hat{y}_\gamma),$$

Notice that H_γ has the same structure as (2.40). Thus $I \geq 0$ provided $\lambda_x \leq Q_1$. We omit the proof since it is almost the same with that of the one dimensional Euler equations.

As for II, we have the expansion as follows.

$$\begin{aligned}
\mathrm{II} =& \frac{1}{4} \sum_{\alpha=2}^{2N_q-2} \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\alpha \hat{\omega}_\beta \hat{\rho}^{\alpha,\beta} \\
&+ z_1 \rho(x^-_{i-\frac{1}{2}}, y^+_{j-\frac{1}{2}}) + z_2 \rho(x^-_{i-\frac{1}{2}}, y_j) + z_3 \rho(x^-_{i-\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
&+ z_4 \rho(x^+_{i-\frac{1}{2}}, y^+_{j-\frac{1}{2}}) + z_5 \rho(x^+_{i-\frac{1}{2}}, y_j) + z_6 \rho(x^+_{i-\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
&+ z_7 \rho(x^-_{i+\frac{1}{2}}, y^+_{j-\frac{1}{2}}) + z_8 \rho(x^-_{i+\frac{1}{2}}, y_j) + z_9 \rho(x^-_{i+\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
&+ z_{10} \rho(x^+_{i+\frac{1}{2}}, y^+_{j-\frac{1}{2}}) + z_{11} \rho(x^+_{i+\frac{1}{2}}, y_j) + z_{12} \rho(x^+_{i+\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
&+ \sum_{\beta=2}^{N_q-1} \hat{\omega}_\beta z_{13,\beta} \rho(x^-_{i-\frac{1}{2}}, \hat{y}_\beta) + \sum_{\beta=N_q+1}^{2N_q-2} \hat{\omega}_\beta z_{13,\beta} \rho(x^-_{i-\frac{1}{2}}, \hat{y}_\beta) \\
&+ \sum_{\beta=2}^{N_q-1} \hat{\omega}_\beta z_{14,\beta} \rho(x^+_{i+\frac{1}{2}}, \hat{y}_\beta) + \sum_{\beta=N_q+1}^{2N_q-2} \hat{\omega}_\beta z_{14,\beta} \rho(x^+_{i+\frac{1}{2}}, \hat{y}_\beta) \\
&+ \sum_{\beta=2}^{N_q-1} \hat{\omega}_\beta z_{15,\beta} \rho(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \sum_{\beta=N_q+1}^{2N_q-2} \hat{\omega}_\beta z_{15,\beta} \rho(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) \\
&+ \sum_{\beta=2}^{N_q-1} \hat{\omega}_\beta z_{16,\beta} \rho(x^+_{i-\frac{1}{2}}, \hat{y}_\beta) + \sum_{\beta=N_q+1}^{2N_q-2} \hat{\omega}_\beta z_{16,\beta} \rho(x^+_{i-\frac{1}{2}}, \hat{y}_\beta),
\end{aligned} \tag{2.57}$$

The expressions of $z_1, \ldots, z_{16,\beta}$ and their estimates can be found in Appendix A.1.5. The conclusion is that all coefficients of point values of ρ^n appearing in (2.57) are nonnegative under the CFL condition (2.55), which implies the nonnegativity of II. Similar arguments also apply to III and IV.

Since $\bar{\rho}_{i,j}^{n+1} = \mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}$, we finish the proof of positivity of $\bar{\rho}_{i,j}^{n+1}$ by summing up the inequalities of $\mathrm{I}, \mathrm{II}, \mathrm{III}$ and IV. □

It remains to show the positivity of specific internal energy of cell averages. Similar to Theorem 2.3.2, we have the result as follows.

Theorem 2.3.5. *Given $\mathbf{u}^n \in G$, the specific internal energy of the cell averages $e(\bar{\mathbf{u}}_{i,j}^{n+1}), i = 1, 2, \dots N_x, j = 1, 2, \dots, N_y$ of scheme* (2.47),(2.52) ,(2.53) *and* (2.54) *are nonnegative under the CFL condition* (2.58).

$$\begin{aligned}
\lambda_x \leq & \frac{\gamma+1}{4\alpha_x^2(\gamma-1)} \min_{i,\beta} \left\{ \frac{p(x^-_{i+\frac{1}{2}}, \hat{y}_\beta)^2}{C(x^-_{i+\frac{1}{2}}, \hat{y}_\beta)}, \frac{p(x^+_{i+\frac{1}{2}}, \hat{y}_\beta)^2}{C(x^+_{i+\frac{1}{2}}, \hat{y}_\beta)} \right\}, \\
\lambda_y \leq & \frac{\gamma+1}{4\alpha_y^2(\gamma-1)} \min_{\alpha,j} \left\{ \frac{p(\hat{x}_\alpha, y^-_{j+\frac{1}{2}})^2}{D(\hat{x}_\alpha, y^-_{j+\frac{1}{2}})}, \frac{p(\hat{x}_\alpha, y^+_{j+\frac{1}{2}})^2}{D(\hat{x}_\alpha, y^+_{j+\frac{1}{2}})} \right\},
\end{aligned} \tag{2.58}$$

where the definitions of the constants are given in Appendix A.1.6.

Proof. By taking $\xi = 1$ on $K_{i,j}$ and zero anywhere else in (2.47), (2.52), (2.53) and (2.54), we have the decomposition of $\bar{\mathbf{u}}_{i,j}^{n+1}$ in x and y directions:

$$\bar{\mathbf{u}}_{i,j}^{n+1} = \mathrm{I} + \mathrm{II},$$

where

$$\mathrm{I} = \frac{1}{2}\bar{\mathbf{u}}_{i,j}^n - \lambda_x \frac{1}{\Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{\mathbf{F}}_{i+\frac{1}{2},j} dy + \lambda_x \frac{1}{\Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{\mathbf{F}}_{i-\frac{1}{2},j} dy,$$

$$\mathrm{II} = \frac{1}{2}\bar{\mathbf{u}}_{i,j}^n - \lambda_y \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{\mathbf{G}}_{i,j+\frac{1}{2}} dx + \lambda_y \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{\mathbf{G}}_{i,j-\frac{1}{2}} dx,$$

where $\hat{\mathbf{F}}_{i+\frac{1}{2},j} = \hat{\mathbf{f}}^{\mathrm{LF}}_{i+\frac{1}{2},j} + \Delta t \tilde{\mathbf{f}}_{i+\frac{1}{2},j} + \Delta t^2 \check{\mathbf{f}}_{i+\frac{1}{2},j}$, $\hat{\mathbf{G}}_{i,j+\frac{1}{2}} = \hat{\mathbf{g}}^{\mathrm{LF}}_{i,j+\frac{1}{2}} + \Delta t \tilde{\mathbf{g}}_{i,j+\frac{1}{2}} + \Delta t^2 \check{\mathbf{g}}_{i,j+\frac{1}{2}}$ are the total fluxes of LWDG defined before, $\hat{\mathbf{f}}^{\mathrm{LF}}_{i+\frac{1}{2},j}, \hat{\mathbf{g}}^{\mathrm{LF}}_{i,j+\frac{1}{2}}$ are Lax-Friedriches fluxes,

$\tilde{\mathbf{f}}_{i+\frac{1}{2},j}, \tilde{\mathbf{g}}_{i,j+\frac{1}{2}}$ and $\check{\mathbf{f}}_{i+\frac{1}{2},j}, \check{\mathbf{g}}_{i,j+\frac{1}{2}}$ are the second and third order terms in the total flux.

By symmetry and concaveness of the internal energy ρe, it suffices to show $\rho e(\mathrm{I}) \geq 0$. We can decompose the term I as

$$\begin{aligned}\mathrm{I} =& \frac{1}{2} \sum_{\alpha=1}^{2N_q-1} \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\alpha \hat{\omega}_\beta \mathbf{u}^{\alpha,\beta} - \lambda_x \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \hat{\mathbf{F}}(x_{i+\frac{1}{2}}, \hat{y}_\beta) + \lambda_x \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \hat{\mathbf{F}}(x_{i-\frac{1}{2}}, \hat{y}_\beta) \\ =& \frac{1}{2} \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \mathbf{H}_\beta,\end{aligned}$$

where $\mathbf{H}_\beta = \sum_{\alpha=1}^{2N_q-1} \hat{\omega}_\alpha \mathbf{u}^{\alpha,\beta} - 2\lambda_x \left(\hat{\mathbf{f}}^{\mathrm{LF}}_{j+\frac{1}{2}} + \Delta t \tilde{\mathbf{f}}_{j+\frac{1}{2}} + \Delta t^2 \check{\mathbf{f}}_{j+\frac{1}{2}} \right) + 2\lambda_x \left(\hat{\mathbf{f}}^{\mathrm{LF}}_{j-\frac{1}{2}} + \Delta t \tilde{\mathbf{f}}_{j-\frac{1}{2}} + \Delta t^2 \check{\mathbf{f}}_{j-\frac{1}{2}} \right)$

Following the same lines as the proof of Theorem (2.3.2), we can show $\rho e(\mathbf{H}_\beta) \geq 0$, which implies $\rho e(\mathrm{I}) \geq 0$ □

Collecting the above two theorems, we reach our final result.

Theorem 2.3.6. *Given $\mathbf{u}^n \in G$, we have $\bar{\mathbf{u}}^{n+1}_{i,j} \in G, i = 1, \ldots, N_x, j = 1, \ldots, N_y$ for the schemes* (2.47), (2.52), (2.53) *and* (2.54), *under the CFL conditions* (2.55) *and* (2.58).

Remark 2.3.1. *To this end, we would like to comment on the CFL conditions obtained in this chapter. These conditions are not optimal for bound-preserving since the splitting of cell averages in the proofs are just for the ease of writing and the bounds may not be sharp in some of the estimates. Moreover, the expressions of the CFL conditions are too tedious to be coded up in practice. Therefore, we actually take the CFL conditions of the bound-preserving Euler forward DG schemes derived in [93, 94] as the initial guess in practice, since the Euler forward methods are the first order approximation of the LWDG in our work. Once the initial step size is not small enough to obtain boundedness of the cell averages, we rewind the computation*

back to the beginning of the time step with a halved step-size of time. The value of the theoretical proofs in this chapter is that we can be guaranteed to obtain bound-preserving cell averages with finitely many halvings of the time step-size.

We also want to note that, for simplicity, we take the viscosity parameter in the Lax-Friedrichs flux to be global in all proofs. However, the local Lax-Friedrichs flux can be used in the bound-preserving technique as well. In practice, the global Lax-Friedrichs flux is more dissipative, thus it may preserve the bounds of target variables more easily, but may result in a more smeared solution.

2.4 Scaling limiters

In the Sections 2.2 and 2.3, we have constructed the maximum-principle-satisfying and positivity-preserving LWDG schemes for hyperbolic equations of scalar and system cases. The cell averages of the target variables fall into their physical bounds under appropriate CFL conditions, provided these bounds are satisfied by the entire solution at the previous time level. To close the cycle of the algorithm, it remains to use appropriate scaling limiters to achieve the bound-preserving for the entire solution.

We adopt the following maximum-principle-satisfying limiter for scalar conservation laws. Given $u \in V$ with $m \leq \bar{u}_j \leq M, j = 1, 2, \ldots, N$, define the modified solution $\tilde{u} \in V$ as follows:

$$\tilde{u}_j(x) = \theta_j \left(u_j(x) - \bar{u}_j\right) + \bar{u}_j, \quad \theta_j = \min\left\{1, \frac{M - \bar{u}_j}{M_j - \bar{u}_j}, \frac{\bar{u}_j - m}{\bar{u}_j - m_j}\right\},$$

$$M_j = \max_{x \in I_j} u_j(x), \quad m_j = \min_{x \in I_j} u_j(x), \quad j = 1, 2, \ldots, N.$$

It is clear that the modified solution $\tilde{u}_j(x) \in [m, M], j = 1, \ldots, N$ and it preserves the cell average. Moreover, it was proved in [92] that such a limiter does not destroy the order of convergence, i.e. $||u - \tilde{u}||_\infty = O(\Delta x^{k+1})$, where k is the order of polynomial space V, which is 2 in this chapter. In practice, one usually take the max and min in the definition of M_j and m_j only over the quadrature points, i.e. $M_j = \max_{1\le\gamma\le 2N_q-1} u_j(\hat{x}_\gamma), m_j = \min_{1\le\gamma\le 2N_q-1} u_j(\hat{x}_\gamma)$, as we only need to control the values at quadrature points. Such a treatment does not affect the accuracy and cell average of the modified solution, as indicated in [93], and we shall use this definition in the numerical section.

For the solution $\mathbf{u} = (\rho, m, E)^T \in V \times V \times V$ of the Euler equations with $\bar{\mathbf{u}}_j \in G, j = 1, 2, \ldots, N$, we adopt the following limiting process which is introduced in [94] and modified in [80].

First, enforce the positivity of the density function ρ by,

$$\hat{\rho}_j(x) = \theta_j^\rho \left(\rho_j(x) - \bar{\rho}_j\right) + \bar{\rho}_j, \quad \theta_j^\rho = \min\left\{1, \frac{\bar{\rho}_j}{\bar{\rho}_j - \min_{1\le\gamma\le 2N_q-1} \rho(\hat{x}_\gamma)}\right\},$$
$$j = 1, 2, \ldots, N.$$

Then let $\hat{\mathbf{u}}_j = (\hat{\rho}_j, m_j, E_j)^T$ and define

$$\tilde{\mathbf{u}}_j(x) = \theta_j^e \left(\hat{\mathbf{u}}_j(x) - \bar{\mathbf{u}}_j\right) + \bar{\mathbf{u}}_j, \quad \theta_j^e = \min\left\{1, \frac{\rho e(\bar{\mathbf{u}}_j)}{\rho e(\bar{\mathbf{u}}_j) - \min_{1\le\gamma\le 2N_q-1} \rho e(\hat{\mathbf{u}}_j(\hat{x}_\gamma))}\right\},$$
$$j = 1, 2, \ldots, N.$$

It follows from the concaveness of the function $\rho e(\mathbf{u})$ that $\tilde{\mathbf{u}}_j(\hat{x}_\gamma) \in G, \gamma = 1, 2, \ldots, 2N_q - 1$, and also it does not destroy accuracy of the solution, see the detailed proof in [94] and [80].

The above limiters are demonstrated based on one space dimension but can be directly extended to multi-dimensions. In implementation, to enhance the stability of algorithms, we can set a threshold $\epsilon = 10^{-10}$ and let $\tilde{u}_j = \bar{u}_j$ if $M - \bar{u}_j < \epsilon$ or $\bar{u}_j - m < \epsilon$ for scalar conservation law, and $\tilde{\mathbf{u}}_j = \bar{\mathbf{u}}_j$ if $\bar{\rho}_j < \epsilon$ or $\rho e(\bar{\mathbf{u}}_j) < \epsilon$ for the Euler equations.

2.5 Numerical tests

In this section, we demonstrate the accuracy and effectiveness of the third order maximum-principle-satisfying and positivity-preserving LWDG schemes by ample numerical tests. The tests are presented from scalar to systems and from one space dimension to two space dimensions with an order of increasing complexity. Most of them can be found in [93, 94, 92, 80].

We have tried both global Lax-Friedrichs and local Lax-Friedrichs fluxes in simulations. The plots of their solutions are very close. However, the accuracy and order of convergence of the global one may be not as good as the local one for some nonlinear problems when the order of DG polynomial space is even, see [15], which is our case. We demonstrate this phenomenon in the tests for Burgers' equation. For all other tests, we only present the results computed using the local Lax-Friedrichs flux to save space. In all the tests, we take the parameters $\beta_0 = 1, \beta_1 = \frac{1}{6}$ in the DDG fluxes.

As mentioned in Remark 3.1, we take the initial guess of CFL numbers in our tests the same as the bound-preserving Euler forward DG schemes [93, 94], and rewind the computation back to the beginning of the time step with a halved step-size of

time if the cell average of solutions exceed their bounds. We report the number of times that the rewinding happens, together with the total number of time steps in each test. As we will see, the actual CFL conditions of LWDG are almost the same with that of the bound-preserving Euler forward DG schemes in most cases.

2.5.1 Scalar conservation laws

Example 2.5.1. *We solve the linear equation* $u_t + u_x = 0$ *in the domain* $\Omega = [-1, 1]$ *with periodic boundary conditions.*

To test the accuracy, we take the smooth initial condition $u_0(x) = \sin(\pi x)$ *and the terminal time* $T = 1$*.*

To show the effect of maximum-principle-preserving, we adopt the discontinuous initial condition

$$u_0(x) = \begin{cases} 1, & -1 \le x \le 0, \\ -1, & 0 \le x \le 1, \end{cases}$$

and take the terminal time $T = 100$*.*

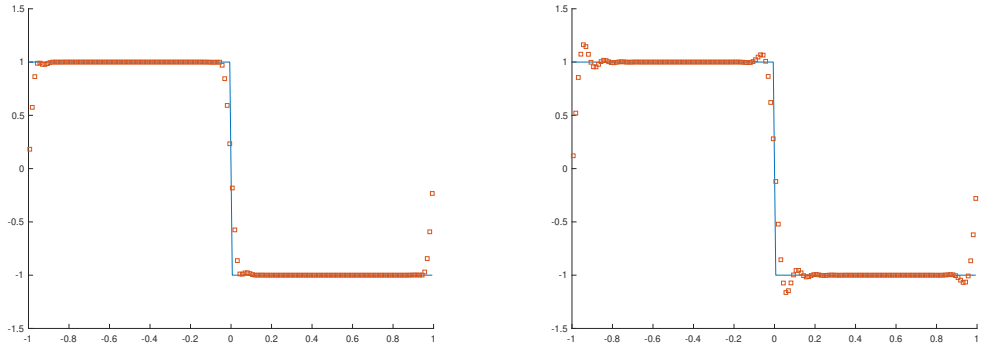
The errors and order of convergence of the problem with the smooth initial condition are given in Table 2.1, from which the third order accuracy can be clearly observed.

The results of the problem with the discontinuous initial condition is shown in Figure 2.1, where a comparison with the exact solution and the result of the unlimited LWDG solution are given. The effect of maximum-principle-preserving is obvious by comparison.

No rewinding of computation happens in this test.

N	L^1 error	order	L^∞ error	order
20	2.06E-04	–	5.09E-04	–
40	2.48E-05	3.05	6.38E-05	3.00
80	3.08E-06	3.01	7.97E-06	3.00
160	3.85E-07	3.00	9.97E-07	3.00
320	4.81E-08	3.00	1.25E-07	3.00
640	6.01E-09	3.00	1.56E-08	3.00

Table 2.1: Results of Example 2.5.1 with smooth initial condition



(a) with limiter (b) without limiter

Figure 2.1: Results of Example 2.5.1 for discontinuous initial condition. $N = 160$. Solid line: exact solution; Squares: numerical solution (cell averages).

Example 2.5.2. *We solve the Burgers' equation* $u_t + \left(\frac{u^2}{2}\right)_x = 0$ *in the domain* $\Omega = [0, 2\pi]$ *with initial condition* $u_0(x) = \frac{1}{2} + \sin(x)$ *and periodic boundary conditions.*

The solution is smooth up to $t = 1$*, when shock appears. We list the errors and order of convergence at* $T = 0.3$ *in Table 2.2 for both the local Lax-Friedrichs flux and global Lax-Friedrichs flux, which shows third order accuracy, and plot the comparison of the numerical solution based on the local Lax-Friedrichs flux with the exact solution at* $T = 2.0$ *in Figure 2.2.*

No rewinding of computation happens in this test.

	Local Lax–Friedrichs Flux				Global Lax–Friedrichs Flux			
N	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
20	9.05E-04	–	1.40E-03	–	1.05E-03	–	1.46E-03	–
40	1.13E-04	3.00	2.35E-04	2.58	1.53E-04	2.77	2.81E-04	2.38
80	1.37E-05	3.05	3.23E-05	2.87	2.24E-05	2.78	4.60E-05	2.61
160	1.66E-06	3.04	4.23E-06	2.93	3.23E-06	2.79	7.20E-06	2.68
320	2.04E-07	3.03	5.38E-07	2.98	4.59E-07	2.82	1.09E-06	2.72
640	2.52E-08	3.02	6.78E-08	2.99	6.43E-08	2.84	1.66E-07	2.72

Table 2.2: Results of Example 2.5.2 at $T = 0.3$

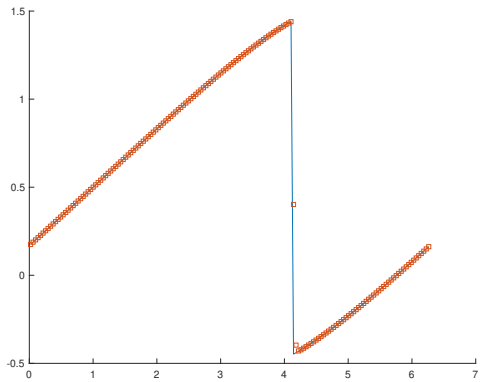


Figure 2.2: Results of Example 2.5.2 at $T = 2.0$. $N = 160$. Solid line: exact solution; Squares: numerical solution (cell averages).

Example 2.5.3. *We solve the two dimensional linear equation* $u_t + u_x + u_y = 0$ *in the domain* $\Omega = [-1, 1] \times [-1, 1]$ *with periodic boundary conditions.*

To show the accuracy, we take the smooth initial condition $u_0(x, y) = \sin(\pi(x+y))$ *and the terminal time* $T = 1$.

To test the effect of maximum-principle-preserving, we adopt a discontinuous initial condition

$$u_0(x) = \begin{cases} 1, & (x, y) \in [-\frac{1}{2}, \frac{1}{2}]^2 \\ -1, & \textit{elsewhere}, \end{cases}$$

and take the terminal time $T = 100$.

The errors and order of convergence for the smooth initial condition are given in Table 2.3, from which the third order accuracy can be observed.

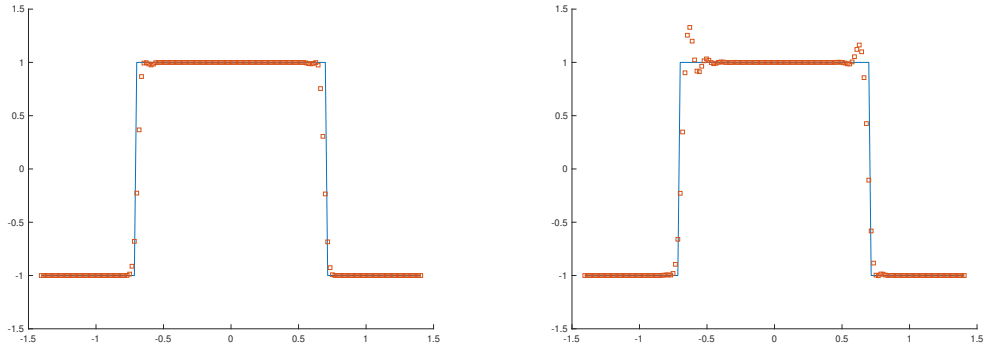
The results of the problem with the discontinuous initial condition is shown in Figure 2.3, where a comparison with the exact solution and the result of the unlimited LWDG solution are given, from which we can see the maximum-principle-preserving limiter works effectively.

No rewinding of computation happens in this test.

$N_x \times N_y$	L^1 error	order	L^∞ error	order
20×20	7.49E-04	–	1.11E-03	–
40×40	7.99E-05	3.23	1.29E-04	3.11
80×80	9.71E-06	3.04	1.61E-05	3.00
160×160	1.21E-06	3.01	2.01E-06	3.00
320×320	1.51E-07	3.00	2.51E-07	3.00
640×640	1.89E-08	3.00	3.14E-08	3.00

Table 2.3: Results of Example 2.5.3 with smooth initial condition

Example 2.5.4. *We solve the two dimensional Burgers' equation* $u_t + \left(\frac{u^2}{2}\right)_x +$



(a) with limiter (b) without limiter

Figure 2.3: Results of Example 2.5.3 with discontinuous initial condition cut along the diagonal ($x = y$) of Ω. $N_x = 160, N_y = 160$. Solid line: exact solution; Squares: numerical solution (cell averages).

$\left(\frac{u^2}{2}\right)_y = 0$ *in the domain* $\Omega = [0, 2\pi] \times [0, 2\pi]$ *with the initial condition* $u_0(x, y) = \frac{1}{2} + \sin(x + y)$ *and periodic boundary conditions.*

The solution is smooth up to $t = 0.5$*, when shock appears. We list the errors and order of convergence for both the local Lax-Friedrichs flux and global Lax-Friedrichs flux, at* $T = 0.2$ *under the* L^1 *and* L^∞ *norms in Table 2.4, and plot the comparison of the numerical solution based on the local Lax-Friedrichs flux with the exact solution at* $T = 1.0$ *along the diagonal of* Ω *in Figure 2.4.*

No rewinding of computation happens in this test.

	Local Lax–Friedrichs Flux				Global Lax–Friedrichs Flux			
$N_x \times N_y$	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
20×20	1.06E-02	–	5.33E-03	–	1.15E-02	–	5.38E-03	–
40×40	1.33E-03	2.99	7.67E-04	2.80	1.63E-03	2.82	8.54E-04	2.66
80×80	1.67E-04	3.00	1.12E-04	2.77	2.43E-04	2.75	1.42E-04	2.59
160×160	2.09E-05	3.00	1.52E-05	2.89	3.61E-05	2.75	2.30E-05	2.62
320×320	2.59E-06	3.01	1.95E-06	2.97	5.26E-06	2.78	3.47E-06	2.73
640×640	3.20E-07	3.01	2.45E-07	2.99	7.46E-07	2.82	4.99E-07	2.80

Table 2.4: Results of Example 2.5.4 at $T = 0.2$

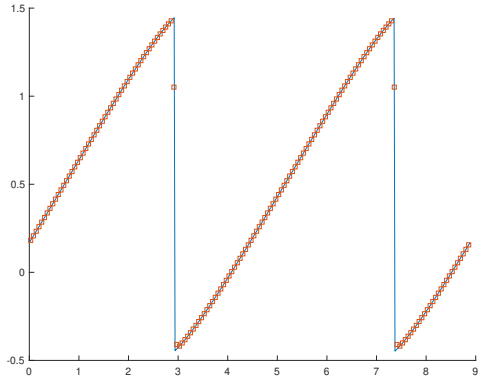


Figure 2.4: Results of Example 2.5.4 cut along the diagonal ($x = y$) of Ω at $T = 1.0$. $N_x = 160, N_y = 160$. Solid line: exact solution; Squares: numerical solution (cell averages).

2.5.2 The Euler equations

Example 2.5.5. *We solve the one dimensional problem in the domain* $\Omega = [0, 2\pi]$ *with the initial condition*

$$\rho_0(x) = 1 + 0.999\sin(x), \quad u_0(x) = 1, \quad p_0(x) = 1$$

and periodic boundary conditions. The ratio of specific heat is $\gamma = 1.4$.

The exact solution of the problem is

$$\rho(x,t) = 1 + 0.999\sin(x-t), \quad u(x,t) = 1, \quad p(x,t) = 1. \tag{2.59}$$

This is a low density problem with the minimum density 0.001. *The positivity of density is preserved during simulation and the third order convergence of density at time* $T = 1$ *is shown in Table 2.5.*

No rewinding of computation happens in this test.

N	L^1 error	order	L^∞ error	order
20	1.13E-03	–	8.60E-04	–
40	1.40E-04	3.01	1.07E-04	3.01
80	1.72E-05	3.02	1.34E-05	3.00
160	2.14E-06	3.01	1.65E-06	3.02
320	2.67E-07	3.00	2.04E-07	3.01
640	3.33E-08	3.00	2.55E-08	3.01

Table 2.5: Results of Example 2.5.5 at $T = 1$

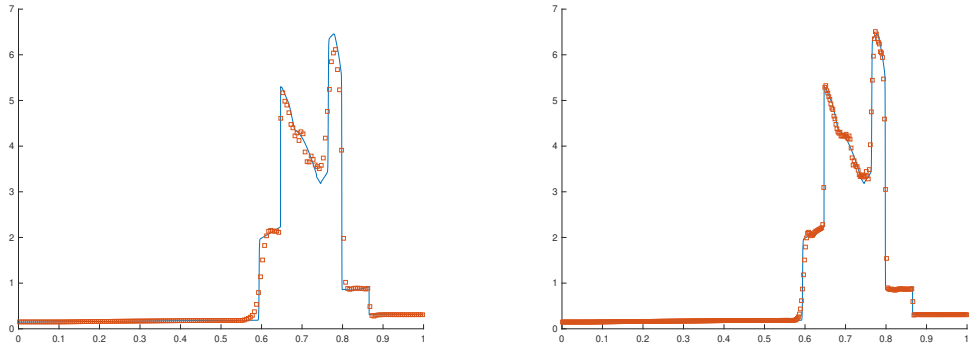
Example 2.5.6. *We solve the one dimensional problem of blast waves in the domain* $\Omega = [0, 1]$ *with initial condition*

$$(\rho_0, u_0, p_0) = \begin{cases} (1, 0, 10^3) & 0 \le x < 0.1, \\ (1, 0, 10^{-2}) & 0.1 \le x < 0.9 \\ (1, 0, 10^2), & 0.9 \le x < 1 \end{cases}$$

and reflective boundary condition. The ratio of specific heat is $\gamma = 1.4$.

We plot the density of numerical solutions at $T = 0.38$ *for* $N = 200, N = 400$, *and compare them with the reference solution, which is computed by the WENO-5 scheme on a very fine mesh with* $16,000$ *cells, in Figure 2.5. Since the positivity-preserving limiter only works when the density or pressure is close to zero and no other limiters are used to stabilize shocks in this test, we can observe some oscillations in the figures.*

In the test for $N = 200$, *there are* 8 *times of rewinding of computation, among a total number of* $6,535$ *time steps. In the test for* $N = 400$, *there are* 15 *times of rewinding of computation, among a total number of* $13,061$ *time steps.*



(a) $N = 200$ (b) $N = 400$

Figure 2.5: Results of Example 2.5.6 at $T = 0.038$. Solid line: reference solution; Squares: numerical solution (cell averages).

Example 2.5.7. *We solve two extreme Riemann problems in one space dimension. The first one is a double rarefaction problem in the domain* $\Omega = [-1, 1]$ *with initial condition*

$$(\rho_0, u_0, p_0) = \begin{cases} (7, -1, 0.2), & x < 0 \\ (7, 1, 0.2), & x > 0. \end{cases}$$

The second one is the Leblanc shock tube problem in the domain $\Omega = [-10, 10]$ *with initial condition*

$$(\rho_0, u_0, p_0) = \begin{cases} (2, 0, 10^9), & x < 0 \\ (10^{-3}, 0, 1), & x > 0. \end{cases}$$

We take the ratio of specific heat $\gamma = 1.4$ *for both cases. In the first test example, vacuum (zero density) will be generated around the origin in the exact solution. For both problems, simulation will blow up without the positivity-preserving limiter in the tests.*

We plot the density of numerical solution of the double rarefaction problem at $T = 0.6$ *on* $N = 200$ *and* $N = 400$ *meshes, and compare them with the reference solution, which is obtained from the exact Riemann solver [78], in Figure 2.6. The*

density of the numerical solution of the Leblanc shock tube problem at $T = 0.0001$ *on* $N = 800$ *and* $N = 1,600$ *meshes, together with the exact solution from the exact Riemann solver, are shown in Figure 2.7, where the y-axis uses* log *scales. From the figures, we can see that the positivity of density and pressure in both cases are preserved, and the numerical solutions agree with the exact solution well.*

No rewinding of computation happens in this test.

Example 2.5.8. *We solve the one dimensional Sedov point-blast wave problem [67] in the domain* $\Omega = [-2, 2]$ *with the initial condition*

$$\rho_0 = 1, \quad u_0 = 0, \quad E_0 = \begin{cases} \frac{3200000}{\Delta x}, & |x| \leq \frac{\Delta x}{2} \\ 10^{-12}, & otherwise. \end{cases}$$

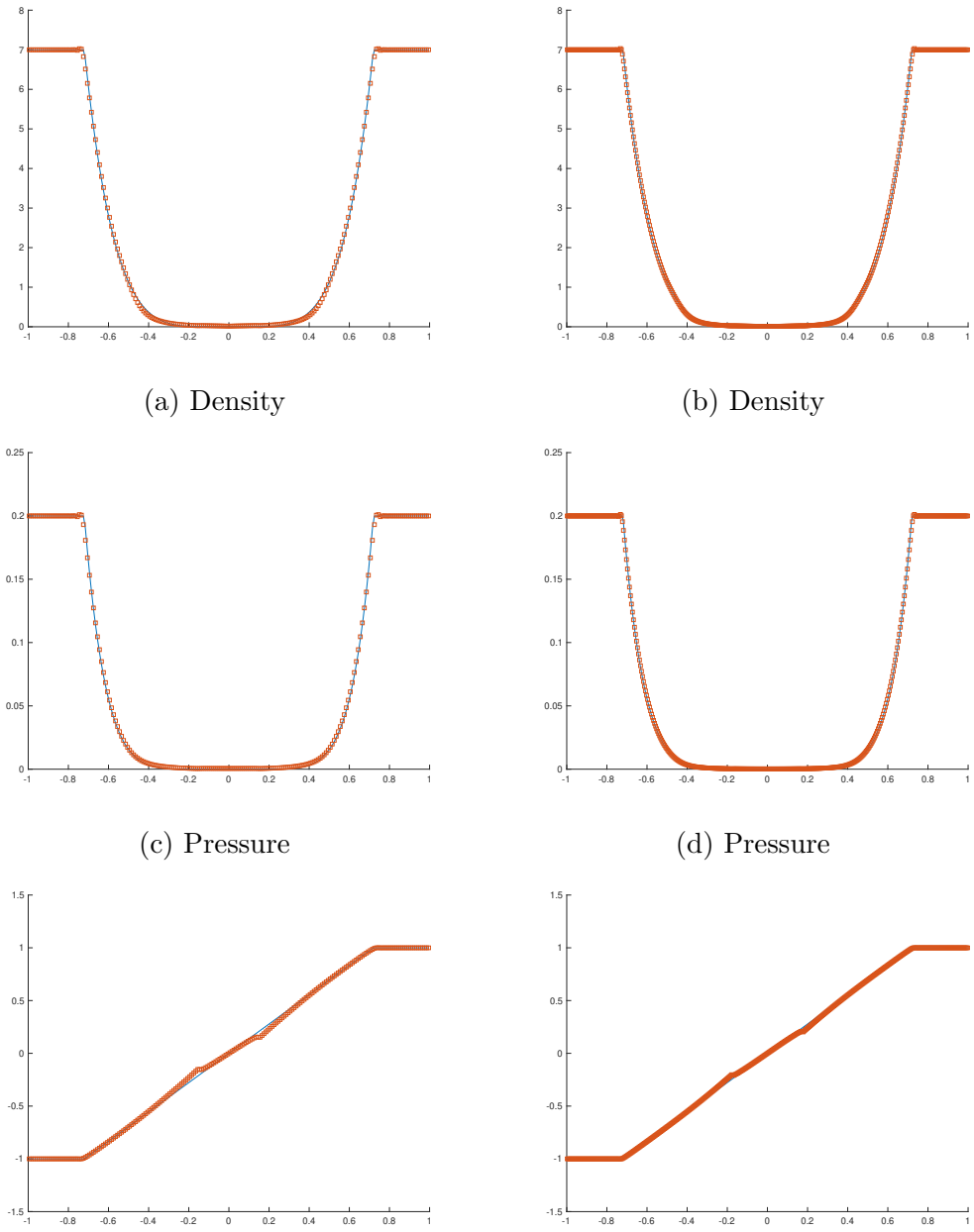
The ratio of specific heat is $\gamma = 1.4$.

This example simulates the point-blast in air, which produces very low density after shock. The simulation will blow up without the positivity-preserving limiter due to the very low density in the exact solution. We plot the simulation results of density, pressure and velocity on $N = 201$ *and* $N = 401$ *meshes at* $T = 0.001$ *in Figure 2.8.*

In the test for $N = 201$*, there is only once of rewinding of computation, among a total number of* $7,377$ *time steps. In the test for* $N = 401$*, there is only once of rewinding of computation, among a total number of* $18,661$ *time steps.*

Example 2.5.9. *We solve the two dimensional problem in the domain* $[0, 2\pi]^2$ *with the initial condition*

$$\rho_0(x, y) = 1 + 0.999 \sin(x + y), \quad u_0 = v_0 = p_0 = 1.$$



(a) Density

(b) Density

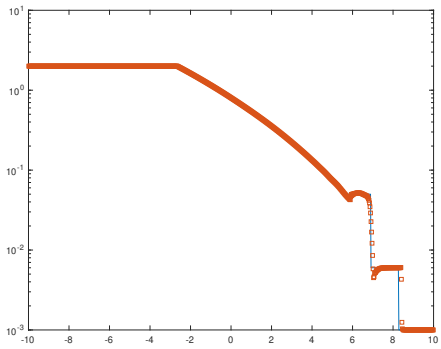
(c) Pressure

(d) Pressure

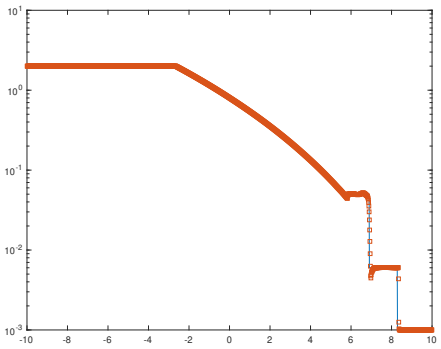
(e) Velocity

(f) Velocity

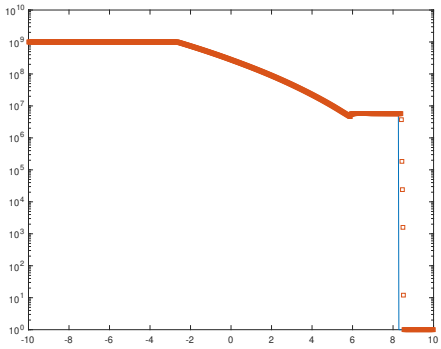
Figure 2.6: Results of Example 2.5.7, the double rarefaction problem, at $T = 0.6$. Solid line: reference solution; Squares: numerical solution (cell averages). Left: $N = 200$; Right:$N = 400$.



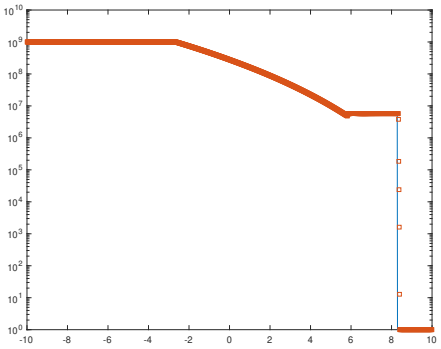
(a) Density, log scale



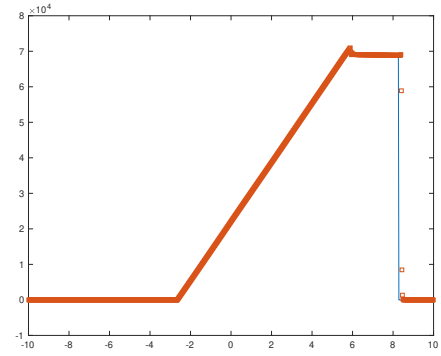
(b) Density, log scale



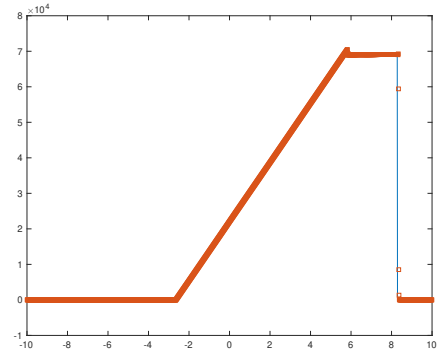
(c) Pressure, log scale



(d) Pressure, log scale

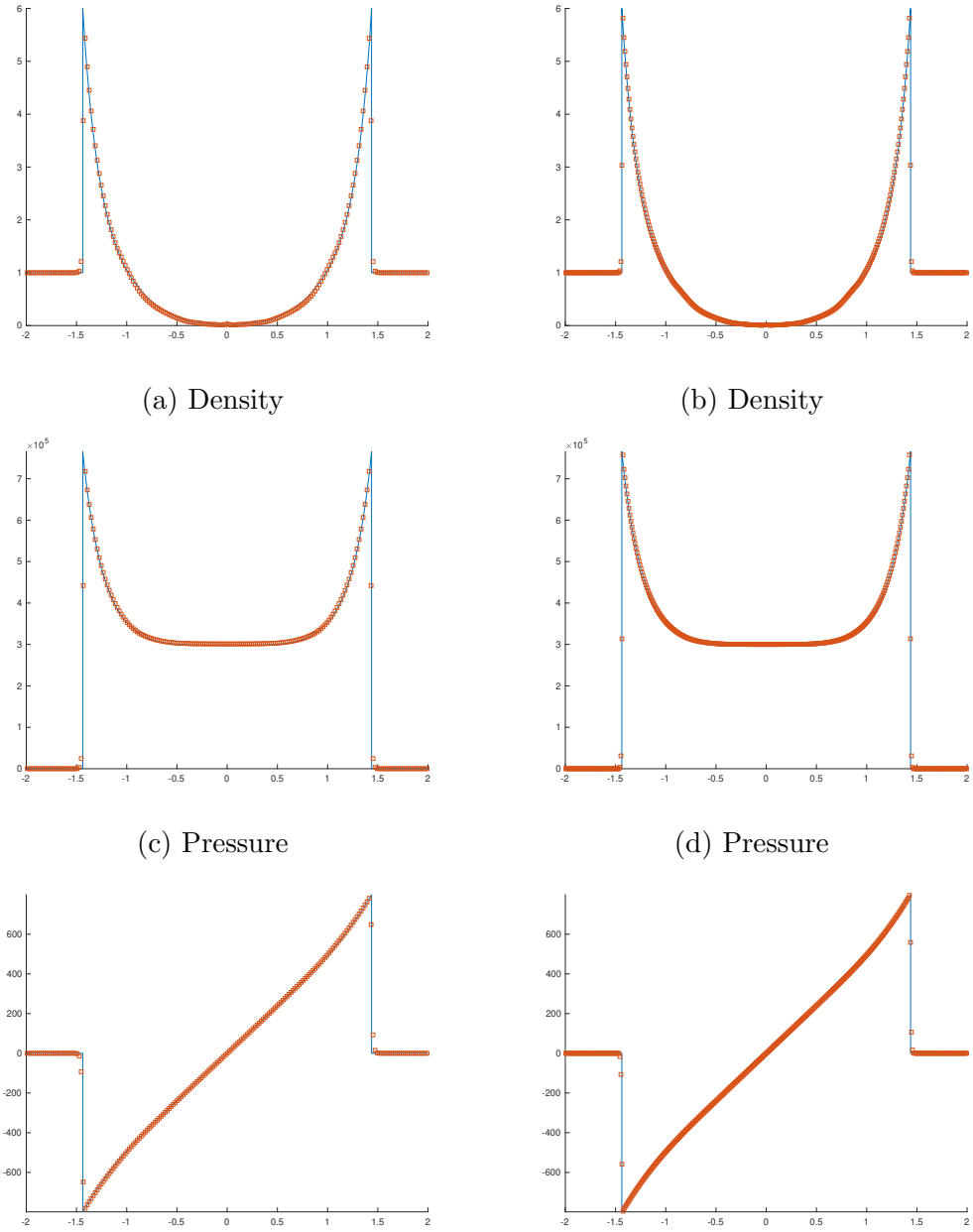


(e) Velocity



(f) Velocity

Figure 2.7: Results of Example 2.5.7, Leblanc shock tube problem, at $T = 0.0001$. Solid line: reference solution; Squares: numerical solution (cell averages). Left: $N = 800$; Right: $N = 1,600$.



(a) Density

(b) Density

(c) Pressure

(d) Pressure

(e) Velocity

(f) Velocity

Figure 2.8: Results of Example 2.5.8 at $T = 0.001$. Solid line: reference solution; Squares: numerical solution (cell averages). Left: $N = 201$; Right: $N = 401$.

and periodic boundary conditions. The ratio of specific heat is $\gamma = 1.4$.

The exact solution of the problem is

$$\rho(x, y, t) = 1 + 0.999 \sin(x + y - 2t), \quad u(x, y, t) = v(x, y, t) = p(x, y, t) = 1.$$

This is a low density problem with the minimum density 0.001. The positivity of density is preserved during simulation and the third order convergence of density at time $T = 0.1$ *is shown in Table 2.6.*

No rewinding of computation happens in this test.

$N_x \times N_y$	L^1 error	order	L^∞ error	order
20×20	8.64E-03	–	1.23E-03	–
40×40	1.37E-03	2.65	2.12E-04	2.53
80×80	1.79E-04	2.94	2.71E-05	2.97
160×160	2.23E-05	3.00	3.33E-06	3.03
320×320	2.75E-06	3.02	4.12E-07	3.02

Table 2.6: Results of Example 2.5.9 at $T = 0.1$

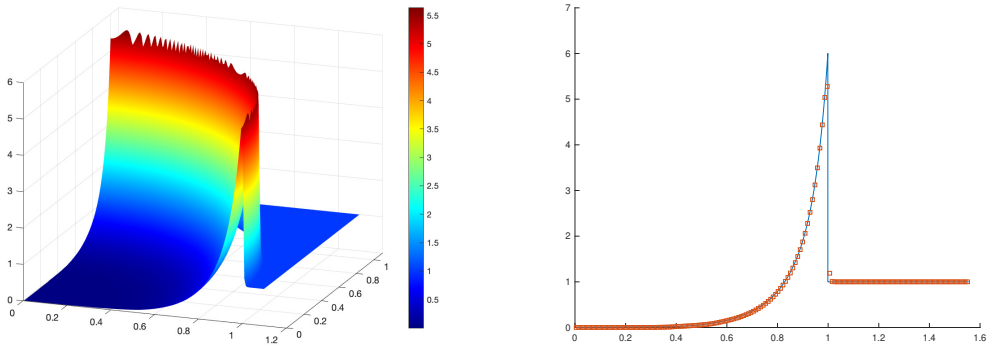
Example 2.5.10. *We solve the two dimensional Sedov point-blast wave problem [67] in the domain* $\Omega = [0, 1.1] \times [0, 1.1]$ *with the initial condition*

$$\rho_0 = 1, \quad u_0 = v_0 = 0, \quad E_0 = \begin{cases} \frac{0.244816}{\Delta x \Delta y}, & (x, y) \in [0, \Delta x] \times [0, \Delta y] \\ 10^{-12}, & otherwise, \end{cases}$$

and the left and bottom boundary the reflective boundary, and other boundaries the outflow boundary. The ratio of specific heat is $\gamma = 1.4$.

We plot the density on Ω *and its profile cut along the diagonal of* Ω *at* $T = 1$ *on the* $N_x = 160, N_y = 160$ *mesh, see Figure 2.9. The simulation blows up if the positivity-preserving limiter is not used in the test.*

In this test, there are 605 *times of rewinding of computation, among a total number of* 344,226 *time steps.*



(a) Density in Ω

(b) Cut of density along diagonal ($x = y$) of Ω

Figure 2.9: Results of Example 2.5.10 at $T = 1$. Solid line: reference solution; Squares: numerical solution (cell averages).

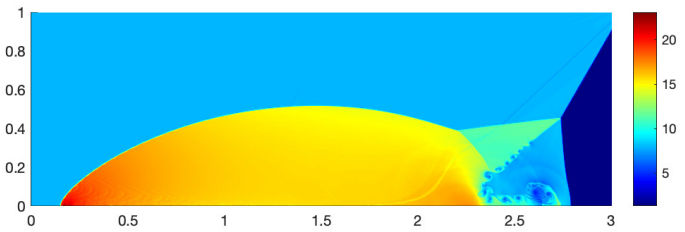
Example 2.5.11. *Consider the two-dimensional double Mach reflection problem with a Mach 10 shock in the domain* $\Omega = [0,4] \times [0,1]$*, with the initial condition*

$$(\rho_0, u_0, v_0, p_0) = \begin{cases} (8, \frac{33\sqrt{3}}{8}, -\frac{33}{8}, 116.5), & y > \sqrt{3}(x - \frac{1}{6}) \quad \textit{(post-shock)} \\ (1.4, 0, 0, 1), & y < \sqrt{3}(x - \frac{1}{6}) \quad \textit{(pre-shock).} \end{cases}$$

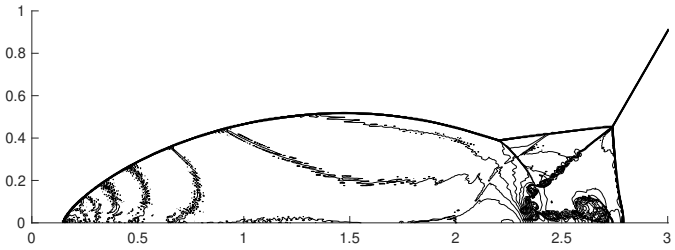
The left boundary is the inflow boundary, the right boundary is the outflow boundary, $\{0 \le x < \frac{1}{6}, y = 0\}$ *on the bottom is the boundary with post-shock condition,* $\{\frac{1}{6} < x \le 4, y = 0\}$ *on the bottom is the reflective boundary, and the condition on top boundary follows the motion of the shock. We show the results at* $T = 0.2$ *on the* $N_x = 960, N_y = 240$ *mesh in Figure 2.10. The results are comparable with the results in [92].*

No rewinding of computation happens in this test.

Example 2.5.12. *We solve the two dimensional problem of shock passing a backward facing corner in the domain* $\Omega = [1,13] \times [0,11] \cup [0,1] \times [6,11]$*, with the initial*



(a) Density on $[0, 3] \times [0, 1]$



(b) 30 equally spaced contour lines from 1.394 to 23.083 for density

Figure 2.10: Results of Example 2.5.11 at $T = 0.2$ on $N_x = 960, N_y = 240$ mesh.

condition

$$(\rho_0, u_0, v_0, p_0) = \begin{cases} (\rho_*, u_*, v_*, p_*), & x < 0.5 \quad \textit{(post-shock)} \\ (1.4, 0, 0, 1), & x > 0.5 \quad \textit{(pre-shock)} \end{cases},$$

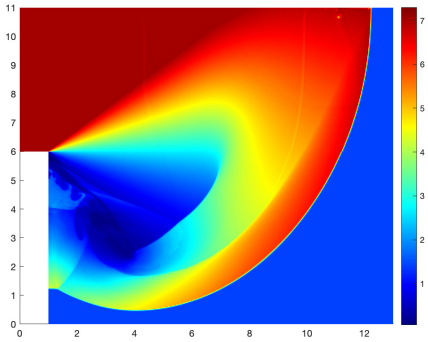
where $(\rho_*, u_*, v_*, p_*) = (7.041132906907898, 4.07794695481336, 0, 30.05945)$ *are taken such that the shock is right-moving with Mach number 5.09. The boundary* $\{x = 0, 6 \leq y \leq 11\}$ *is the inflow boundary,* $\{0 \leq x \leq 1, y = 6\}$ *and* $\{x = 1, 0 \leq y \leq 6\}$ *are reflexive boundaries,* $\{x = 13, 0 \leq y \leq 11\}$ *and* $\{1 \leq x \leq 13, y = 0\}$ *are outflow boundaries, and the boundary condition on* $\{0 \leq x \leq 13, y = 11\}$ *follows the motion of the shock.*

The density and pressure at $T = 2.3$ *with* $\Delta x = \Delta y = \frac{1}{32}$ *are presented in Figure 2.11. The results are comparable with the results in [92, 94]*

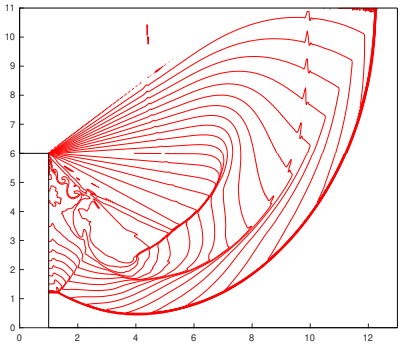
No rewinding of computation happens in this test.

Example 2.5.13. *Consider the two-dimensional astrophysical jets problems with very high Mach number. We set the domain* $\Omega = [0, 0.5] \times [0, 0.25]$ *with initial condition* $\rho_0(x, y) = 0.5, u_0(x, y) = v_0(x, y) = 0, p_0(x, y) = 0.4127$*. The boundary conditions of the right and top are outflow; the bottom boundary is reflexive; the left boundary is inflow with* $(\rho, u, v, p) = (5, 800, 0, 0.4127)$ *if* $0 \leq y \leq 0.05$*, which corresponds to a jet flow of Mach number* 2000*, while* $(\rho, u, v, p) = (0.5, 0, 0, 0.4127)$ *otherwise. The ratio of specific heat is* $\gamma = 5/3$*.*

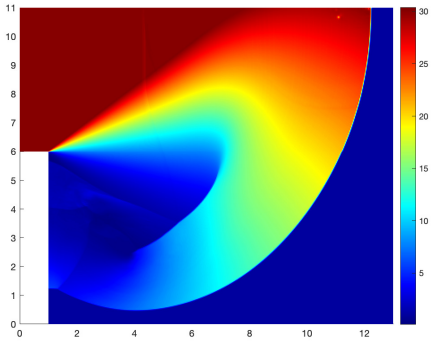
A combination of the total variation bounded limiter [18] and the flux limiter [98] are used before applying the positivity-preserving limiter in each time stage to reduce the spurious oscillations where the density and pressure are far above zero. We would like to note that, the positivity of density and pressure are preserved during



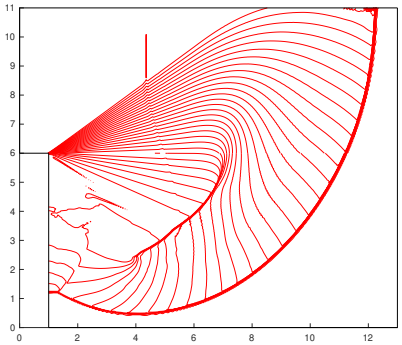
(a) Density in Ω



(b) 20 equally spaced contour lines from 0.066227 to 7.0668 for density



(c) Pressure in Ω

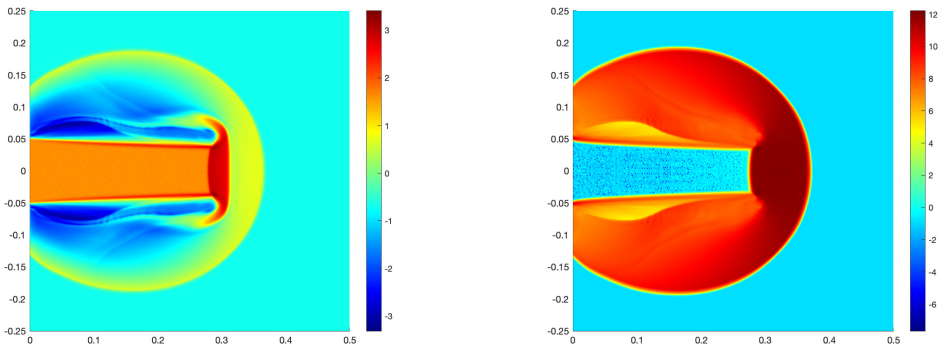


(d) 40 equally spaced contour lines from 0.091 to 37 for pressure

Figure 2.11: Results of Example 2.5.12 at $T = 2.3$.

simulation if only the positivity-preserving limiter is used, however, the simulation blows up very soon without the positivity-preserving limiter. We compute the solution on $N_x \times N_y = 320 \times 160$ *grid, and show the density and pressure at* $T = 5 \times 10^{-4}$ *in Figure 2.12.*

In this test, there are $1,968,558$ *times of rewinding of computation, resulting in a total number of* $356,643$ *time steps. The unusually small CFL number is caused by the TVB limiter adopted, without which there is no rewinding of computation and the CFL number is almost* 10 *times larger, but the result is oscillatory, though the positivity is preserved. Since the scope of this chapter is on positivity-preserving algorithms, we do not further study more compatible slope limiters for this example here.*



(a) Density with log scale, lower part flipped from the upper part

(b) Pressure with log scale, lower part flipped from the upper part

Figure 2.12: Results of Example 2.5.13 at $T = 5 \times 10^{-4}$.

2.6 Concluding remarks

In this chapter, we have proposed the third order maximum-principle-satisfying and positivity-preserving discontinuous Galerkin methods for scalar conservation laws

and the Euler equations, respectively, based on the Lax-Wendroff time discretization. The approach here is specified for DG methods with the use of DDG discretization for the second temporal derivative terms. The main contribution of the work is to prove rigorously that, under suitable CFL conditions, the cell average of the unmodulated LWDG scheme at the next time step is bounded, provided the solution stay in the desired bounds at the current time step. The scaling limiters, which were proved not to affect the high order accuracy and mass conservation, can then be used to enforce the bounds for the whole solution at the next time step, hence closing the loop of the bound-preserving LWDG algorithm.

Several possible extensions could be made in future works. For instance, it is of great importance to extend the algorithm to schemes with accuracy higher than third order. It is also meaningful to extend the algorithm from structured grids to unstructured meshes for geometry flexibility. The 3D case of the algorithm will also be studied in the future.

Chapter Three

Positivity-preserving discontinuous Galerkin methods for stationary hyperbolic equations

3.1 Introduction

In this chapter, we are interested in numerical methods for stationary hyperbolic equations. In the one dimensional space, we consider the variable coefficient and nonlinear stationary hyperbolic equations

$$(a(x)u)_x + \lambda u = f(x), \quad x \in \Omega = [0,1], \tag{3.1}$$

where $a(x)$ does not change sign and, without loss of generality $a(x) > 0$, and

$$(a(u)u)_x + \lambda u = f(x), \quad x \in \Omega = [0,1], \tag{3.2}$$

where $a(u)$ does not change sign and, without loss of generality $a(u) > 0$. Here $\lambda \geq 0$ is a constant. In two and three dimensional spaces, we consider the constant coefficient stationary hyperbolic equations

$$au_x + bu_y + \lambda u = f(x,y), \quad (x,y) \in \Omega = [0,1]^2, \tag{3.3}$$

and

$$au_x + bu_y + cu_z + \lambda u = f(x,y,z), \quad (x,y,z) \in \Omega = [0,1]^3, \tag{3.4}$$

respectively, where $\lambda \geq 0$ is a constant and, without loss of generality, we assume $a, b, c > 0$.

The stationary hyperbolic equations (3.1)-(3.4) have wide applications in steady-state transport problems. Moreover, the equations form the building block of the linear radiative transfer equation (RTE), which is an integro-differential equation that describes the distribution of radiative intensity in a medium, based on the

discrete-ordinate method (DOM) [24, 37] and iterative procedure on the source terms, see [90, 46] for more details.

The discontinuous Galerkin (DG) method is one of the most popular numerical methods to solve hyperbolic equations, for its advantages in obtaining high order accuracy, flexibility for complex geometry and easiness to be parallelized. In 1973, Reed and Hill [62] proposed the first DG scheme to solve the linear steady-state RTE for neutron transport problems. It was later developed into Runge-Kutta discontinuous Galerkin (RKDG) methods by Cockburn et al. in a series of papers [17, 16, 14, 13, 18] to solve time-dependent hyperbolic equations such as the Burgers equation, Euler equations, and shallow water equations, etc. In this chapter, we will adopt the classic DG method to solve the stationary hyperbolic equations.

For stationary hyperbolic equations, it is well-known that their physical solutions satisfy the positivity-preserving property, i.e. the solutions are nonnegative, provided the corresponding boundary conditions and source terms are nonnegative. When designing numerical methods, one naturally wants to maintain the positivity-preserving property on the numerical solution, since negative values are not only physically unacceptable, but also may cause severe robustness issues in the simulations, especially when coupled with other physical systems.

There have been intensive studies on positivity-preserving DG methods. In 2010, the genuinely maximum-principle-satisfying DG method was proposed by Zhang et al. in [93] for time-dependent scalar hyperbolic equations. The method is called positivity-preserving when the lower bound in the maximum-principle is zero, which is the case in our problems. The general framework of the positivity-preserving method is composed of two parts. The first part is to obtain the solution at the next time step with nonnegative cell averages from the original, unlimited DG scheme,

probably under certain step-size conditions. Once the cell averages of solution are guaranteed nonnegative, the scaling limiter in [93], which maintains the high order accuracy and mass conservation, is applied to modify the solution such that the entire solution becomes nonnegative. Based on this simple but powerful framework, positivity-preserving and maximum-principle-satisfying DG methods for time-dependent problems have been rapidly developed later, e.g. for the Euler equations [94, 95], Navier-Stokes equations [92, 44], shallow water equations[82, 81], convection-diffusion equations [96, 83], and compressible miscible displacements [29], among others.

In 2016, Yuan et al. [90] proposed a high order positivity-preserving DG method for constant coefficient stationary hyperbolic equations. Taking the one dimensional case as an example, their algorithm is as follows: Firstly, they proved a fundamental result that the numerical solution $u(x)$ solved from the unmodulated DG method satisfies $\max\{\bar{u}_K, u_K(x_c)\} \geq 0$ on every cell K of the mesh, where $\bar{u}_K$ is the cell average on K, and x_c is the right end point (the downwind point) of K. They then modify the solution $u_K(x)$ on cell K based on the principle that, if $\bar{u}_K \geq 0$, the conservative scaling limiter [93]

$$\tilde{u}_K(x) = \theta\left(u_K(x) - \bar{u}_K\right) + \bar{u}_K, \text{where } \theta = \min\{\frac{\bar{u}_K}{\bar{u}_K - \min_K u_K(x)}, 1\} \tag{3.5}$$

is applied, otherwise a non-conservative rotational limiter [90] centered at x_c is used. Their algorithm can maintain positivity without affecting high order accuracy, however, since the cell average $\bar{u}_K$ can be changed by the rotation, the algorithm is not conservative in general, which is also true when the algorithm is extended to two-dimensional rectangular [90] or triangular [91] meshes. In 2018, Ling et al. [46] improved the result by rigorously proving that the solution of the unmodulated DG method in one dimension actually satisfies $\bar{u}_K \geq 0$ for all K. Therefore the scaling

limiter (3.5) can always be used, which yields a high order conservative positivity-preserving DG method. In their work, a special test function ξ that recovers cell averages $\bar{u}_K$ from the left hand side of the DG scheme was proved to be nonnegative, which implies $\bar{u}_K \geq 0$ since the source term and boundary terms on the right hand side of the DG scheme are both nonnegative, see more details in [46]. Unfortunately, direct extension to two dimensions fails due to the fact that such test function ξ is no longer nonnegative over the cell in rectangular meshes, even for second order DG method with P^1 or Q^1 spaces. Instead, the authors obtained a second order positivity-preserving conservative scheme on rectangular meshes by augmenting the P^1 finite element space, but the extension of this approach to higher space dimensions or to higher order schemes was not carried out in [46] and is highly nontrivial.

In this chapter, we further investigate high order conservative positivity-preserving DG method for stationary hyperbolic equations. We put our effort on proving the positivity of cell averages of the scheme so that the conservative scaling limiter (3.5) can be applied directly to maintain high order accuracy and positivity. The main difficulty is that the unmodulated DG method fails in positivity-preserving for cell averages in all the equations we consider in this chapter, which will be illustrated by concrete examples in later sections. To resolve this difficulty, we modify the original DG method by adopting appropriate quadrature rules to replace the exact integrals in the schemes, which is a common practice in the implementation of DG schemes, not only because the exact integral is often difficult to obtain, but also for the purpose of achieving specific properties, e.g. maximum-principle-satisfying [93] or entropy stability [10]. The quadrature rules adopted in the schemes are easy to implement and can be directly extended to high dimensions. More importantly, we will show that the cell averages of the DG schemes with such quadrature rules are positive, by proving the positivity of the test function that recovers the cell average

from the left hand side of the schemes.

The rest of the chapter is organized as follows. In Section 3.2, we propose the conservative positivity-preserving method in the one dimensional space by introducing the desired quadrature rules in the DG formulation, which do not evaluate the integrals in the DG scheme exactly. We give an example to explain why such quadratures are necessary, and rigorously prove the positivity-preserving property of our method. In Section 3.3, we propose the positivity-preserving DG methods for two and three space dimensions, based on direct extensions from the 1D algorithm. We detail the implementation of the positivity-preserving scaling limiter (3.5) and summarize the complete positivity-preserving algorithm in Section 3.4. The good performance of the schemes are demonstrated by ample numerical experiments in Section 3.5. Due to the inaccurate quadrature, the order of convergence is suboptimal in two and three space dimensions, but we observe optimal convergence in all one dimensional tests. Finally, we end in Section 3.6 with concluding remarks.

3.2 Numerical algorithm in one space dimension

In this section, we construct high order conservative positivity-preserving DG methods for stationary hyperbolic equations (3.1) and (3.2) in the one dimensional space. The schemes can be arbitrarily high order for the case of (3.1) with $\lambda = 0$, but for the other cases we are only able to prove the positivity-preserving property for P^1 and P^2 (second and third order) DG schemes.

3.2.1 Notations

We take the partition $0 = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N+\frac{1}{2}} = 1$ on $\Omega = [0,1]$, and denote the j-th cell by $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, with the cell size $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ and the cell center $x_j = \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})$ for $j = 1, 2, \ldots, N$.

The finite element space of P^k-DG scheme is defined as

$$V_h^k = \{v \in L^2([0,1]) : v|_{I_j} \in P^k(I_j), j = 1, 2, \ldots, N\}, \tag{3.6}$$

where $P^k(I)$ is the polynomial space of order no greater than k on I. For $v \in V_h^k$, we define the cell average $\bar{v}_j = \frac{1}{\Delta x_j} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x)dx$ on I_j. Moreover, we denote by $v_{j+\frac{1}{2}}^-$ and $v_{j+\frac{1}{2}}^+$ the left and right limits of v at $x_{j+\frac{1}{2}}$, respectively, i.e. $v_{j+\frac{1}{2}}^{\pm} = v(x_{j+\frac{1}{2}} \pm 0)$.

For the purpose of positivity-preserving, we adopt the Gauss-Legendre quadrature rule of k points to evaluate volume integrals in the P^k-DG scheme, and denote this quadrature by $⨍_{I_j} v(x)dx = \Delta x_j \sum_{\alpha=1}^k \hat{\omega}_\alpha v(\hat{x}_\alpha)$, where $\{\hat{x}_\alpha, \alpha = 1, \ldots, k\}$ are the quadrature points on I_j and $\{\hat{\omega}_\alpha, \alpha = 1, \ldots, k\}$ are the quadrature weights satisfying $\sum_{\alpha=1}^k \hat{\omega}_\alpha = 1$.

3.2.2 Variable coefficient stationary hyperbolic equation in one space dimension

Consider the variable coefficient stationary hyperbolic equation (3.1) with $f(x) \geq 0$ in Ω. As mentioned before, without loss of generality we assume $a(x) > 0$ and the corresponding boundary condition $u(0) = u_0 \geq 0$. The case $a(x) < 0$ with the boundary condition $u(1) = u_0 \geq 0$ can be obtained by the change of variable

$x' = 1 - x$.

Firstly, we give an example to show that the original DG scheme with exact integrals may produce negative cell averages, even when the upwind boundary condition and the source term are both positive. The original P^k-DG scheme of the equation (3.1) is to seek $u \in V_h^k$, s.t. $\forall w \in V_h^k$,

$$-\int_{I_j} (a(x)uw_x - \lambda uw)dx + a(x_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}}w^-_{j+\frac{1}{2}} = a(x_{j-\frac{1}{2}})u^-_{j-\frac{1}{2}}w^+_{j-\frac{1}{2}} + \int_{I_j} fwdx, \quad (3.7)$$

for $j = 1, 2, \ldots, N$, where we let $u^-_{\frac{1}{2}} = u_0$. We adopt the P^1-DG scheme and take $a(x) = 1 + x, \lambda = 0$ and $u_0 > 0$. It is easy to check that $\xi(x) = \frac{6+5\Delta x_1}{6+8\Delta x_1+2\Delta x_1^2} - \frac{3x}{\Delta x_1(3+\Delta x_1)}$ is the unique function in $P^1(I_1)$ such that $-\int_{I_1} a(x)v\xi_x dx + a(x_{\frac{3}{2}})v^-_{\frac{3}{2}}\xi^-_{\frac{3}{2}} = \bar{v}_1$ for all $v \in V_h^1$, and $\xi(x_{\frac{3}{2}}) = -\frac{\Delta x_1}{2(3+4\Delta x_1+\Delta x_1^2)} < 0$. By taking the test function $w = \xi$ (where we extend $w = 0$ outside I_1) in the scheme, we can construct $f(x) \geq 0$ that takes large values around $x_{\frac{3}{2}}$ such that $\bar{u}_1 = a(0)u_0\xi(x_{\frac{1}{2}}) + \int_{I_1} f\xi dx < 0$. One can check that if we adopt P^2, P^3, P^4, P^5-DG schemes and take $a(x) = 1+x^2, a(x) = 1 + x^3, a(x) = 1 + x^4, a(x) = 1 + x^5$, respectively, negative cell averages may also appear following the same lines, see the details in Appendix B.2.

However, we are going to show that the positivity-preserving property can be achieved simply by replacing the exact integrals in the scheme by the Gauss-Legendre quadratures of k points. The positivity-preserving P^k-DG scheme of (3.1) is to seek $u \in V_h^k$, s.t. $\forall w \in V_h^k$,

$$-\fint_{I_j} (a(x)uw_x - \lambda uw)\, dx + a(x_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}}w^-_{j+\frac{1}{2}} = a(x_{j-\frac{1}{2}})u^-_{j-\frac{1}{2}}w^+_{j-\frac{1}{2}} + \fint_{I_j} fwdx, \quad (3.8)$$

for $j = 1, 2, \ldots, N$.

Cockburn et al. have proved in [13] that a sufficient condition for the quadra-

ture in P^k-DG scheme to attain optimal convergence is to have algebraic degree of accuracy $2k$. Though this condition is not satisfied by the quadrature in (3.8), we observe optimal order of convergence in all one dimensional tests.

Based on the framework of [93], we only need to put our effort on proving the positivity of cell averages of the scheme (3.8), then the scaling limiter (3.5) can be used to achieve positivity of the entire solution without losing mass conservation and accuracy. Same as in [46], it suffices to prove the positivity of the test function $\xi \in V_h^k$ that recovers the cell average of the solution from the left hand side of the scheme (3.8).

We assume that $a(x) \in C^k(I_j), j = 1, 2, \ldots, N$, in the P^k-DG scheme to make sense of some norms to be used. We first consider the case $\lambda = 0$ and give the main result as follows.

Lemma 3.2.1. *Define* $\xi(x) = \frac{1}{\Delta x_j} \int_x^{x_{j+\frac{1}{2}}} \mathcal{L}[\frac{1}{a(t)}]dt$ *for* $x \in I_j$, *where* $\mathcal{L}[\cdot]$ *is the Lagrange interpolation operator at the Gauss-Legendre points* $\{\hat{x}_\alpha\}_{\alpha=1}^k$, *then* ξ *is the unique function in* $P^k(I_j)$ *that satisfies*

$$-\int_{I_j} a(x)v\xi_x dx + a(x_{j+\frac{1}{2}})v^-_{j+\frac{1}{2}}\xi^-_{j+\frac{1}{2}} = \bar{v}_j, \quad \forall v \in P^k(I_j). \tag{3.9}$$

Moreover, for $k = 1$, $\xi \geq 0$ *on* I_j; *for* $k \geq 2$, $\xi \geq 0$ *on* I_j *if the mesh size satisfies*

$$\Delta x_j \leq \left(\frac{(2k)!}{k!||a(x)||_{L^\infty(I_j)}||\frac{d^k}{dx^k}\left(\frac{1}{a(x)}\right)||_{L^\infty(I_j)}} \right)^{\frac{1}{k}}. \tag{3.10}$$

Proof. By definition, $\xi \in P^k(I_j)$, $\xi_x(x) = -\frac{1}{\Delta x_j}\mathcal{L}[\frac{1}{a(t)}](x)$, and $\xi^-_{j+\frac{1}{2}} = 0$. Therefore,

it follows from direct computation that, $\forall v \in P^k(I_j)$,

$$\begin{aligned}
&-\fint_{I_j} a(x)v\xi_x dx + a(x_{j+\frac{1}{2}})v^-_{j+\frac{1}{2}}\xi^-_{j+\frac{1}{2}} \\
=&\sum_{\alpha=1}^{k} \hat{\omega}_\alpha a(\hat{x}_\alpha)v(\hat{x}_\alpha)\mathcal{L}[\frac{1}{a(t)}](\hat{x}_\alpha) + 0 \\
=&\sum_{\alpha=1}^{k} \hat{\omega}_\alpha v(\hat{x}_\alpha) = \bar{v}_j,
\end{aligned}$$

where the last equality holds because the k-point Gauss-Legendre quadrature is exact for integrals of polynomials of order at most k.

As for the uniqueness, we consider the corresponding homogeneous linear problem: Find $\eta \in P^k(I_j)$, s.t.

$$-\fint_{I_j} a(x)v\eta_x dx + a(x_{j+\frac{1}{2}})v^-_{j+\frac{1}{2}}\eta^-_{j+\frac{1}{2}} = 0, \quad \forall v \in P^k(I_j).$$

If we take v as the $k+1$ Lagrange basis at $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k, x_{j+\frac{1}{2}}$, the above linear problem is converted to the system of linear equations

$$\begin{cases}
\eta_x(\hat{x}_\alpha) = 0, \quad \alpha = 1, 2, \ldots, k \\
\eta(x_{j+\frac{1}{2}}) = 0.
\end{cases}$$

Since $\eta_x \in P^{k-1}(I_j)$, we have $\eta_x \equiv 0$ from the uniqueness of Lagrange interpolation, which implies $\eta \equiv 0$ since $\eta(x_{j+\frac{1}{2}}) = 0$. Therefore, the function satisfying (3.9) is unique in $P^k(I_j)$.

To show the positivity of ξ, it suffices to prove its integrand $\mathcal{L}[\frac{1}{a(x)}] \geq 0$ on I_j. When $k = 1$, this is clear because the Lagrange interpolant $\mathcal{L}[\frac{1}{a(x)}] = \frac{1}{a(\hat{x}_1)}$ is a constant. When $k \geq 2$, we need the error formula[6] of the Lagrange polynomial for

$g(x) \in C^k(I_j)$ interpolating at $\hat{x}_1, \ldots, \hat{x}_k$,

$$g(x) - \mathcal{L}[g](x) = \frac{g^{(k)}(\zeta(x))}{k!}(x - \hat{x}_1)(x - \hat{x}_2)\cdots(x - \hat{x}_k),$$

where $\zeta(x) \in I_j$ is generally unknown. Moreover, let us recall that the standard k-th order Legendre polynomial satisfies $|P_k(r)| \leq 1$ for $r \in [-1, 1]$, and has the explicit formula

$$P_k(r) = \frac{(2k)!}{2^k(k!)^2}(r - \hat{r}_1)(r - \hat{r}_2)\cdots(r - \hat{r}_k),$$

where $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_k$ are the roots of the k-th order Legendre polynomial. The properties of the Legendre polynomials imply $|\frac{1}{k!}(x - \hat{x}_1)(x - \hat{x}_2)\cdots(x - \hat{x}_k)| = \frac{(\Delta x_j)^k k!}{(2k)!}$ $\left|P_k\left(\frac{x - \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})}{\Delta x_j/2}\right)\right| \leq \frac{(\Delta x_j)^k k!}{(2k)!}$. Therefore, we have the lower bound estimates for $\mathcal{L}[\frac{1}{a(x)}]$ on I_j as follows,

$$\begin{aligned}
\mathcal{L}[\frac{1}{a(t)}](x) &= \frac{1}{a(x)} - \frac{d^k}{dx^k}\left(\frac{1}{a(x)}\right)\Big|_{x=\zeta} \cdot \frac{1}{k!}(x - \hat{x}_1)(x - \hat{x}_2)\cdots(x - \hat{x}_k) \\
&\geq \frac{1}{||a(x)||_{L^\infty(I_j)}} - \frac{(\Delta x_j)^k k!}{(2k)!}||\frac{d^k}{dx^k}\left(\frac{1}{a(x)}\right)||_{L^\infty(I_j)} \\
&\geq 0, \quad \forall x \in I_j,
\end{aligned}$$

under the condition $\Delta x_j \leq \left(\frac{(2k)!}{k!||a(x)||_{L^\infty(I_j)}||\frac{d^k}{dx^k}\left(\frac{1}{a(x)}\right)||_{L^\infty(I_j)}}\right)^{\frac{1}{k}}$ on the mesh size. □

Remark 3.2.1. *The condition* (3.10) *is drawn from the requirement that the Lagrange interpolation* $\mathcal{L}[\frac{1}{a(x)}]$ *being nonnegative on* I_j. *Since we have assumed the smoothness of* $a(x)$, *which implies* $\frac{1}{a(x)}$ *is smooth and lower bounded away from zero, the mesh size condition should not be severe. Indeed, since we merely need the integration* $\int_x^{x_{j+\frac{1}{2}}} \mathcal{L}[\frac{1}{a(t)}]dt \geq 0, x \in I_j$, *to guarantee the positivity of* ξ, *the actual condition needed on the mesh size may be even more relaxed.*

Based on the lemma above, if we assume the inflow condition $u^-_{j-\frac{1}{2}} \geq 0$, we can

immediately obtain the positivity of $\bar{u}_j$ by taking the test function $w = \xi$ (extend $\xi = 0$ outside I_j) in the scheme (3.8) and using the fact that the source term f and coefficient $a(x)$ are positive. We can therefore obtain the result for the positivity-preserving property of the scheme (3.8) with $\lambda = 0$ as follows.

Theorem 3.2.2. *For the variable coefficient stationary hyperbolic equation* (3.1) *with* $\lambda = 0$*, if the source term and inflow conditions from upstream cells (including the inflow condition on the first cell) are positive, then the cell averages of the scheme* (3.8) *are positive, under the mesh size condition in Lemma 3.2.1.*

We then consider the case $\lambda > 0$ and give the main result as follows.

Lemma 3.2.3. *Define the functions*

$$\xi_1(x) = \frac{2(x_{j+\frac{1}{2}} - x)}{\Delta x_j(2a(\hat{x}_1) + \lambda \Delta x_j)}, \quad x \in I_j,$$

and

$$\xi_2(x) = \frac{6(x_{j+\frac{1}{2}} - x)\left(\tilde{\lambda}(x - x_{j-\frac{1}{2}}) + a(\hat{x}_1) + a(\hat{x}_2)\right)}{\Delta x_j\left(12a(\hat{x}_1)a(\hat{x}_2) + 3\Delta x_j\lambda(a(\hat{x}_1) + a(\hat{x}_2)) + \Delta x_j^2\lambda^2\right)}, \quad x \in I_j,$$

where $\tilde{\lambda} = \lambda + \frac{\sqrt{3}(a(\hat{x}_1) - a(\hat{x}_2))}{\Delta x_j}$*, for* P^1*-DG and* P^2*-DG schemes, respectively, then* ξ_1 *and* ξ_2 *are the unique functions in* $P^k(I_j)$ *that satisfies*

$$-\fint_{I_j} (a(x)v\xi_x - \lambda v\xi)\, dx + a(x_{j+\frac{1}{2}})v^-_{j+\frac{1}{2}}\xi^-_{j+\frac{1}{2}} = \bar{v}_j, \quad \forall v \in P^k(I_j), \tag{3.11}$$

for $k = 1$ *and* $k = 2$*, respectively.*

Moreover, $\xi_1 \geq 0$ *on* I_j*;* $\xi_2 \geq 0$ *on* I_j *if* $\lambda \geq p_j(a)$*, or otherwise* $\Delta x_j \leq$

$\frac{2\min_{x\in I_j} a(x)}{p_j(a)-\lambda}$, *where* $p_j(\cdot)$ *is the one-sided Lipschitz seminorm [5] defined as*

$$p_j(v) = \sup_{x,y\in I_j, x\neq y} \left(\frac{v(x)-v(y)}{x-y}\right)_+, \text{ where } z_+ = \max(0,z).$$

Proof. It is easy to check by solving the linear equation/system that $\xi_1(x)$ and $\xi_2(x)$ are the unique solutions of the linear problem (3.11) for $k=1$ and $k=2$, respectively.

It is also clear that $\xi_1(x) \geq 0$ on I_j, since $a(x), \lambda > 0$ by assumption.

As for $k=2$, the positivity of $\xi_2(x)$ is always the same to its factor $\tilde{\lambda}(x-x_{j-\frac{1}{2}})+a(\hat{x}_1)+a(\hat{x}_2)$. Note that $\tilde{\lambda} = \lambda - \frac{a(\hat{x}_1)-a(\hat{x}_2)}{\hat{x}_1-\hat{x}_2} \geq \lambda - p_j(a)$, thereby $\tilde{\lambda}(x-x_{j-\frac{1}{2}})+a(\hat{x}_1)+a(\hat{x}_2) \geq a(\hat{x}_1)+a(\hat{x}_2) \geq 0$ if $\lambda \geq p_j(a)$, or $\tilde{\lambda}(x-x_{j-\frac{1}{2}})+a(\hat{x}_1)+a(\hat{x}_2) \geq a(\hat{x}_1)+a(\hat{x}_2)-(p_j(a)-\lambda)\Delta x_j \geq 0$ if $\lambda < p_j(a)$. Both cases indicate that $\xi_2(x) \geq 0$ on I_j. □

Following the same arguments as before, we can immediately get the positivity of $\bar{u}_j$ if we assume the positivity of the inflow condition and the source term. We can therefore obtain the result for the positivity-preserving property of the scheme (3.8) with $\lambda > 0$ (in fact it also applies to the case of $\lambda = 0$) as follows.

Theorem 3.2.4. *For the variable coefficient stationary hyperbolic equation* (3.1) *with* $\lambda > 0$*, if the source term and the inflow conditions from upstream cells (including the inflow condition on the first cell) are positive, then the cell averages of the scheme* (3.8) *are positive, under the conditions in Lemma 3.2.3.*

Remark 3.2.2. *We are only able to prove the positivity-preserving property for* P^k*-DG methods with* $k=1$ *and* $k=2$ *here. For the cases* $k \geq 3$*, the positivity of test function* ξ *satisfying* (3.11) *is too complicated to be analyzed generally. However,*

we have investigated these cases for some special $a(x)$ and the results are promising, which are shown in Appendix B.2.

3.2.3 Nonlinear stationary hyperbolic equation in one space dimension

Consider the nonlinear stationary hyperbolic equation (3.2) with $f \geq 0$ in Ω. We assume $a(u) \geq c > 0$, $\frac{d(a(u)u)}{du} > 0$ for all u, and the boundary condition $u(0) \geq 0$.

Formally, we still have the same positivity-preserving results as in the variable coefficient case if we adopt the scheme: seek $u \in V_h^k$, s.t. $\forall w \in V_h^k$,

$$-\fint_{I_j} (a(u)uw_x - \lambda uw)\,dx + a(u^-_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}}w^-_{j+\frac{1}{2}} = a(u^-_{j-\frac{1}{2}})u^-_{j-\frac{1}{2}}w^+_{j-\frac{1}{2}} + \fint_{I_j} fwdx, \quad (3.12)$$

for $j = 1, 2, \ldots, N$, since $a(u)$ in the scheme can be regarded as $a(u(x))$ in the variable coefficient case. However, because $u(x)$ is unknown, the mesh size conditions established before for positivity-preserving is unavailable for $k \geq 2$. To resolve this difficulty, we give a P^2-DG scheme which is positivity-preserving on arbitrary meshes: seek $u \in V_h^2$, s.t. $\forall w \in V_h^2$,

$$-\fint_{I_j} (a(u)uw_x - \lambda uw)\,dx + a(u^-_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}}w^-_{j+\frac{1}{2}} = a(u^-_{j-\frac{1}{2}})u^-_{j-\frac{1}{2}}w^+_{j-\frac{1}{2}} + \fint_{I_j} fwdx, \quad (3.13)$$

for $j = 1, 2, \ldots, N$, where $\fint_{I_j}$ denotes the Simpson's quadrature rule.

We give the main result for the P^2-DG scheme (3.13) as follows.

Lemma 3.2.5. *Let $u(x)$ be the solution of the scheme* (3.13) *and define the function*

$$\xi(x) = \frac{6(x_{j+\frac{1}{2}} - x)\left(\tilde{\lambda}(x - x_{j-\frac{1}{2}}) + a(u(\hat{x}_1)) + a(u(\hat{x}_2))\right)}{\Delta x_j\left(12a(u(\hat{x}_1))a(u(\hat{x}_2)) + 3\Delta x_j\lambda(a(u(\hat{x}_1)) + a(u(\hat{x}_2))) + \Delta x_j^2\lambda^2\right)}, \quad x \in I_j, \tag{3.14}$$

where $\tilde{\lambda} = \lambda + \frac{\sqrt{3}(a(u(\hat{x}_1)) - a(u(\hat{x}_2)))}{\Delta x_j}$, *then* $\xi \in P^2(I_j)$ *satisfies*

$$-\fint_{I_j}(a(u)v\xi_x - \lambda v\xi)\,dx + a(u^-_{j+\frac{1}{2}})v^-_{j+\frac{1}{2}}\xi^-_{j+\frac{1}{2}} = \bar{v}_j, \quad \forall v \in P^2(I_j). \tag{3.15}$$

Moreover, $\xi \geq 0$ *at the points* $\{x_{j-\frac{1}{2}}, x_j, x_{j+\frac{1}{2}}\}$.

Proof. It can be verified by direct computations similar to the proofs before. □

Following the same arguments as in the variable coefficient case, we immediately get the positivity of $\bar{u}_j$, if we assume the positivity of inflow condition and source term. Though the expression of ξ in (3.14) contains the unknown solution u, it is not a problem since we actually do not use ξ in the implementation of the positivity-preserving algorithm. We can therefore obtain the result for the positivity-preserving property of the schemes (3.12) for $k = 1$ and (3.13) for $k = 2$ as follows.

Theorem 3.2.6. *For the nonlinear stationary hyperbolic equation* (3.2)*, if the source term and inflow conditions from upstream cells (including the inflow condition on the first cell) are positive, then the cell averages of the schemes* (3.12) *for* $k = 1$ *and* (3.13) *for* $k = 2$ *are positive on arbitrary meshes.*

3.3 Numerical algorithm in two and three space dimensions

In this section, we construct high order conservative positivity-preserving DG schemes for constant coefficient stationary hyperbolic equations (3.3) and (3.4) in two and three dimensions, respectively. The schemes are direct extensions from the algorithm in one space dimension. We are only able to give rigorous proofs of positivity-preserving for limited cases but numerical computation shows strong evidence that the schemes are positivity-preserving for Q^k-DG for arbitrary k in two dimensions, and for odd $k = 1, 3, 5, 7, \ldots$ in three dimensions.

3.3.1 Notations

We take the partition $0 = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N_x+\frac{1}{2}} = 1$, $0 = y_{\frac{1}{2}} < y_{\frac{3}{2}} < \cdots < y_{N_y+\frac{1}{2}} = 1$, and $0 = z_{\frac{1}{2}} < z_{\frac{3}{2}} < \cdots < z_{N_z+\frac{1}{2}} = 1$ in the x, y and z directions, respectively, and define the mesh sizes $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, i = 1, \ldots, N_x$, $\Delta y_j = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}, j = 1, \ldots, N_y$, and $\Delta z_l = z_{l+\frac{1}{2}} - z_{l-\frac{1}{2}}, l = 1, \ldots, N_z$, with cell centers $x_i = \frac{1}{2}(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}}), i = 1, \ldots, N_x$, $y_j = \frac{1}{2}(y_{j-\frac{1}{2}} + y_{j+\frac{1}{2}}), j = 1, \ldots, N_y$, and $z_l = \frac{1}{2}(z_{l-\frac{1}{2}} + z_{l+\frac{1}{2}}), l = 1, \ldots, N_z$. Moreover, we denote by $K_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}], i = 1, \ldots, N_x, j = 1, \ldots, N_y$ the cells in the two dimensional domain $\Omega = [0, 1]^2$, and $K_{i,j,l} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}] \times [z_{l-\frac{1}{2}}, z_{l+\frac{1}{2}}], i = 1, \ldots, N_x, j = 1, \ldots, N_y, l = 1, \ldots, N_z$ the cells in the three dimensional domain $\Omega = [0, 1]^3$.

The finite element spaces of the Q^k-DG scheme are defined as

$$V_h^k = \{v \in L^2([0,1]^2) : v|_{K_{i,j}} \in Q^k(K_{i,j}), i = 1, \ldots, N_x, j = 1, \ldots, N_y\}, \tag{3.16}$$

and

$$V_h^k = \{v \in L^2([0,1]^3) : v|_{K_{i,j,l}} \in Q^k(K_{i,j,l}), i = 1, \ldots, N_x, j = 1, \ldots, N_y, l = 1, \ldots, N_z\}, \tag{3.17}$$

in two and three dimensional domains, respectively, where $Q^k(K)$ is the tensor product polynomial space of order no greater than k on the cell K. For $v \in V_h^k$, we denote the cell average by $\bar{v}_{i,j}$ on $K_{i,j}$, and $\bar{v}_{i,j,l}$ on $K_{i,j,l}$. In two space dimensions, we define the left/right and lower/upper limits of v on the vertical and horizontal cell interfaces by $v(x^{\pm}_{i+\frac{1}{2}}, y) = v(x_{i+\frac{1}{2}} \pm 0, y)$ and $v(x, y^{\pm}_{j+\frac{1}{2}}) = v(x, y_{j+\frac{1}{2}} \pm 0)$, respectively. In three space dimensions, the limits on cell interfaces are defined similarly.

We let $\{\hat{r}_\alpha, \hat{\omega}_\alpha\}_{\alpha=1}^k$ and $\{\tilde{r}_\alpha, \tilde{\omega}_\alpha\}_{\alpha=1}^{k+1}$ be the Gauss-Legendre quadrature rules with k and $k+1$ quadrature points on $[-1, 1]$, respectively. As in the previous section, we use the notation $\fint$ to denote the approximate integration via the k-point Gauss-Legendre quadrature. If not otherwise stated, the usual integral notation $\int$ stands for the exact integral, which can be evaluated by the $k+1$ point Gauss-Legendre quadrature in the Q^k-DG scheme for the constant coefficient problems. Finally, we denote by $\{\ell_i(x), i = 1, \ldots, k\}$ the Lagrange interpolation basis at $\{\hat{r}_\alpha\}_{\alpha=1}^k$ with $\ell_i(\hat{r}_\alpha) = \delta_{i,\alpha}$, and by $\ell_i'(x)$ the derivative of $\ell_i(x)$.

3.3.2 Constant coefficient stationary hyperbolic equation in two space dimensions

Consider the constant coefficients stationary hyperbolic equation (3.3) with $f(x,y) \geq 0$ in Ω. As mentioned before, without loss of generality, we may assume $a, b > 0$, because the other cases can be obtained by the change of variables $x' = 1 - x$ and/or $y' = 1 - y$. The corresponding boundary conditions are given by $u(0,y) = g_1(y), u(x,0) = g_2(x)$, where $g_1, g_2 \geq 0$.

Firstly, we would like to remark that the original DG methods are not positivity-preserving for the cell averages in general, even for the P^1-DG or Q^1-DG schemes. One can refer to the counterexamples constructed in [46].

The positivity-preserving Q^k-DG scheme of (3.3) is to seek $u \in V_h^k$ s.t. $\forall w \in V_h^k$,

$$\begin{aligned}
&-\oint_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \oint_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (auw_x + buw_y - \lambda uw)\, dxdy \\
&+ \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} au(x^-_{i+\frac{1}{2}}, y) w(x^-_{i+\frac{1}{2}}, y) dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} bu(x, y^-_{j+\frac{1}{2}}) w(x, y^-_{j+\frac{1}{2}}) dx \\
&= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} au(x^-_{i-\frac{1}{2}}, y) w(x^+_{i-\frac{1}{2}}, y) dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} bu(x, y^-_{j-\frac{1}{2}}) w(x, y^+_{j-\frac{1}{2}}) dx \\
&+ \oint_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \oint_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} fw dxdy,
\end{aligned} \tag{3.18}$$

for $i = 1, \ldots, N_x, j = 1, \ldots, N_y$. If $x_{i-\frac{1}{2}} = 0$, we let $u(x^-_{i-\frac{1}{2}}, y) = g_1(y)$, similarly if $y_{j-\frac{1}{2}} = 0$, we let $u(x, y^-_{j-\frac{1}{2}}) = g_2(x)$. The quadrature adopted in (3.18) does not satisfy the condition for optimal convergence established in [13], which results in sub-optimal convergence as we will show in the numerical tests.

Without loss of generality, we only consider scheme (3.18) on the reference cell

$K=[-1,1]\times[-1,1]$, as any cell $K_{i,j}=[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]\times[y_{j-\frac{1}{2}},y_{j+\frac{1}{2}}]$ can be transferred to K by changing of coordinates which only rescales a,b,λ,f without altering their signs. We give the main result as follows.

Lemma 3.3.1. *Define* $\xi(x,y;a,b,\lambda)=(1-x)(1-y)\eta(x,y;a,b,\lambda)$ *for* $(x,y)\in[-1,1]^2$, *where* $\eta(x,y;a,b,\lambda)=\sum_{i,j=1}^{k}\eta_{ij}(a,b,\lambda)\ell_i(x)\ell_j(y)$, *and* $\{\eta_{ij}(a,b,\lambda)\}_{i,j=1}^{k}$ *is the solution of the linear system*

$$\begin{aligned}\sum_{i,j=1}^{k}\big(a\left((1-\hat{r}_\alpha)(1-\hat{r}_\beta)\ell_i'(\hat{r}_\alpha)\delta_{\beta,j}-(1-\hat{r}_\beta)\delta_{\alpha,i}\delta_{\beta,j}\right)\\ +b\left((1-\hat{r}_\alpha)(1-\hat{r}_\beta)\ell_j'(\hat{r}_\beta)\delta_{\alpha,i}-(1-\hat{r}_\alpha)\delta_{\alpha,i}\delta_{\beta,j}\right)\\ -\lambda(1-\hat{r}_\alpha)(1-\hat{r}_\beta)\delta_{\alpha,i}\delta_{\beta,j}\big)\,\eta_{ij}=-\frac{1}{4},\quad \alpha,\beta=1,2,\ldots,k,\end{aligned}\tag{3.19}$$

then $\xi(x,y;a,b,\lambda)\in Q^k([-1,1]^2)$ *satisfies*

$$\begin{aligned}&-\fint_{-1}^{1}\fint_{-1}^{1}(av\xi_x+bv\xi_y-\lambda v\xi)\,dxdy+\int_{-1}^{1}av(1,y)\xi(1,y)dy+\int_{-1}^{1}bv(x,1)\xi(x,1)dx\\&=\frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}vdxdy,\end{aligned}\tag{3.20}$$

for any $v\in Q^k([-1,1]^2)$.

Moreover, for $k=1,2$, *we have* $\xi(x,y;a,b,\lambda)\geq 0$ *on* $[-1,1]^2$; *for* $k=3$, *we can show* $\xi(\hat{r}_\alpha,\hat{r}_\beta;a,b,0)\geq 0$, $\alpha,\beta=1,2,3$ *and* $\xi(-1,\tilde{r}_\alpha;a,b,0),\xi(\tilde{r}_\alpha,-1;a,b,0)\geq 0$, $\alpha=1,2,3,4$.

Proof. By definition of $\xi(x,y)$, we can compute that $\xi_x(x,y)=(1-x)(1-y)\sum_{i,j=1}^{k}\eta_{ij}\ell_i'(x)\ell_j(y)-(1-y)\sum_{i,j=1}^{k}\eta_{ij}\ell_i(x)\ell_j(y)$ and $\xi_y(x,y)=(1-x)(1-y)\sum_{i,j=1}^{k}\eta_{ij}\ell_i(x)\ell_j'(y)-(1-x)\sum_{i,j=1}^{k}\eta_{ij}\ell_i(x)\ell_j(y)$, thereby it can be checked that $\{\eta_{ij}\}_{i,j=1}^{k}$ is the solution

of the linear system (3.19) if and only if ξ satisfies

$$a\xi_x(\hat{r}_\alpha, \hat{r}_\beta) + b\xi_y(\hat{r}_\alpha, \hat{r}_\beta) - \lambda\xi(\hat{r}_\alpha, \hat{r}_\beta) = -\frac{1}{4}, \quad \alpha, \beta = 1, 2, \ldots, k.$$

Moreover, we have $\xi(1, y) = \xi(x, 1) = 0$ from the definition. Therefore, it follows from direct computation that

$$\begin{aligned}
&-\int_{-1}^{1}\int_{-1}^{1} (av\xi_x + bv\xi_y - \lambda v\xi)\, dxdy + \int_{-1}^{1} av(1, y)\xi(1, y)dy + \int_{-1}^{1} bv(x, 1)\xi(x, 1)dx \\
=& -4 \sum_{\alpha,\beta=1}^{k} \hat{\omega}_\alpha\hat{\omega}_\beta v(\hat{r}_\alpha, \hat{r}_\beta) \left(a\xi_x(\hat{r}_\alpha, \hat{r}_\beta) + b\xi_y(\hat{r}_\alpha, \hat{r}_\beta) - \lambda\xi(\hat{r}_\alpha, \hat{r}_\beta)\right) + 0 + 0 \\
=& \sum_{\alpha,\beta=1}^{k} \hat{\omega}_\alpha\hat{\omega}_\beta v(\hat{r}_\alpha, \hat{r}_\beta) = \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} v dxdy, \qquad \forall v \in Q^k([-1, 1]^2),
\end{aligned}$$

where the last equality follows from the fact that the tensor product of k-point Gauss-Legendre quadrature is accurate for $v \in Q^k([-1, 1]^2)$.

It remains to show the positivity of ξ, or equivalently η.

When $k = 1$, by solving the linear equation (3.19), we have $\eta(x, y; a, b, \lambda) = \frac{1}{4(a+b+\lambda)} > 0$.

When $k = 2$, by solving the linear system (3.19), we have $\eta(x, y; a, b, \lambda) = C^{-1}(6a^3 + 15a^2b + 15ab^2 + 6b^3 + 9a^2\lambda + 17ab\lambda + 9b^2\lambda + 5a\lambda^2 + 5b\lambda^2 + \lambda^3 + 3a^2bx + 9ab^2x + 6b^3x + 3a^2\lambda x + 9ab\lambda x + 9b^2\lambda x + 3a\lambda^2x + 5b\lambda^2x + \lambda^3x + 6a^3y + 9a^2by + 3ab^2y + 9a^2\lambda y + 9ab\lambda y + 3b^2\lambda y + 5a\lambda^2y + 3b\lambda^2y + \lambda^3y + 9a^2bxy + 9ab^2xy + 3a^2\lambda xy + 9ab\lambda xy + 3b^2\lambda xy + 3a\lambda^2xy + 3b\lambda^2xy + \lambda^3xy)$, where $C = \frac{16}{9}(3a^2 + 3ab + 3b^2 + 3a\lambda + 3b\lambda + \lambda^2)(3a^2 + 6ab + 3b^2 + 3a\lambda + 3b\lambda + \lambda^2) > 0$. Since $\eta \in Q^1([-1, 1]^2)$ and $\eta(-1, -1) = C^{-1}(12a^2b + 12ab^2 + 8ab\lambda) > 0$, $\eta(-1, 1) = C^{-1}(12a^3 + 12a^2b + 12a^2\lambda + 8ab\lambda + 4a\lambda^2) > 0$, $\eta(1, -1) = C^{-1}(12ab^2 + 12b^3 + 8ab\lambda + 12b^2\lambda + 4b\lambda^2) > 0$, $\eta(1, 1) =$

$C^{-1}(12a^3+36a^2b+36ab^2+12b^3+24a^2\lambda+44ab\lambda+24b^2\lambda+16a\lambda^2+16b\lambda^2+4\lambda^3)>0$, we have $\eta(x,y;a,b,\lambda)>0$ for $(x,y)\in[-1,1]^2$.

Now we consider the case $k=3$ with $\lambda=0$. Firstly, we note that from the definition, $\xi(x,y;a,b,\lambda)=C\xi(x,y;Ca,Cb,C\lambda)$ and $\eta(x,y;a,b,\lambda)=C\eta(x,y;Ca,Cb,C\lambda)$, $\forall C>0$. Therefore it suffices to investigate the case $a=1,b>0$ since $\xi(x,y;a,b,0)=\frac{1}{a}\xi(x,y;1,\frac{b}{a},0)$. By solving the linear system (3.19), we get $\eta_{ij}(1,b,0)=\frac{P_{ij}(b)}{Q(b)},i,j=1,2,3$, where $P_{ij}(b)$ and $Q(b)$ are polynomials defined as:

$$
\begin{aligned}
P_{11}(b) =&2(5(5-\sqrt{15})+5(17-4\sqrt{15})b+(195-31\sqrt{15})b^2+(240-38\sqrt{15})b^3\\
&+(195-31\sqrt{15})b^4+5(17-4\sqrt{15})b^5+5(-5+\sqrt{15})b^6)\\
P_{12}(b) =&20+(95+3\sqrt{15})b+180b^2+14(15-\sqrt{15})b^3+(195-29\sqrt{15})b^4\\
&+25(5-\sqrt{15})b^5+10(5-\sqrt{15})b^6\\
P_{13}(b) =&2(5(5+\sqrt{15})+5(8+\sqrt{15})b+(45+\sqrt{15})b^2+30b^3+(45-\sqrt{15})b^4\\
&+5(8-\sqrt{15})b^5+5(5-\sqrt{15})b^6)\\
P_{21}(b) =&10(5-\sqrt{15})+25(5-\sqrt{15})b+(195-29\sqrt{15})b^2+14(15-\sqrt{15})b^3\\
&+180b^4+(95+3\sqrt{15})b^5+20b^6\\
P_{22}(b) =&20+95b+198b^2+249b^3+198b^4+95b^5+20b^6\\
P_{23}(b) =&10(5+\sqrt{15})+25(5+\sqrt{15})b+(195+29\sqrt{15})b^2+14(15+\sqrt{15})b^3\\
&+180b^4+(95-3\sqrt{15})b^5+20b^6\\
P_{31}(b) =&2(5(5-\sqrt{15})+5(8-\sqrt{15})b+(45-\sqrt{15})b^2+30b^3+(45+\sqrt{15})b^4\\
&+5(8+\sqrt{15})b^5+5(5+\sqrt{15})b^6)\\
P_{32}(b) =&20+(95-3\sqrt{15})b+180b^2+14(15+\sqrt{15})b^3+(195+29\sqrt{15})b^4\\
&+25(5+\sqrt{15})b^5+10(5+\sqrt{15})b^6\\
P_{33}(b) =&2(5(5+\sqrt{15})+5(17+4\sqrt{15})b+(195+31\sqrt{15})b^2+(240+38\sqrt{15})b^3\\
&+(195+31\sqrt{15})b^4+5(17+4\sqrt{15})b^5+5(5+\sqrt{15})b^6)
\end{aligned}
$$

$$Q(b) = 16(1+b)(5+15b+27b^2+31b^3+27b^4+15b^5+5b^6)$$

One can observe that all coefficients in the above polynomials are positive. Therefore, we have $\eta(\hat{r}_\alpha, \hat{r}_\beta; a, b, 0) = \frac{1}{a}\eta(\hat{r}_\alpha, \hat{r}_\beta; 1, \frac{b}{a}, 0) = \frac{1}{a}\frac{P_{\alpha,\beta}(b/a)}{Q(b/a)} > 0$, for $\alpha, \beta = 1, 2, 3$. Further more, since $\eta(x, y; 1, b, 0) = \sum_{i,j=1}^{3} \eta_{ij}(1, b, 0)\ell_i(x)\ell_j(y) = \frac{\sum_{i,j=1}^{3} P_{ij}(b)\ell_i(x)\ell_j(y)}{Q(b)}$, the values of η at the quadrature points $\{(-1, \tilde{r}_\alpha), \alpha = 1, 2, 3, 4\}$ and $\{(\tilde{r}_\alpha, -1), \alpha = 1, 2, 3, 4\}$ are also rational functions of b. By direct computation, one can check that the coefficients of these rational functions are all positive, which implies the positivity of $\eta(x, y; a, b, 0)$ at these points. We omit the details of computation since it is straightforward but lengthy. □

Remark 3.3.1. *By the Cramer's rule, we always have* $\eta_{ij}(1, b, 0) = \frac{P_{ij}(b)}{Q(b)}$, *where* $P_{ij}(b)$ *and* $Q(b)$ *are polynomials,* $i, j = 1, 2, \ldots, k$, *for general* k. *However, Mathematica is unable to afford the symbolic calculation for* $k > 3$. *We sample some values of* b *and solve the corresponding values of* $P_{i,j}(b)$ *and* $Q(b)$ *numerically. By interpolation, we recover the expressions of* $P_{i,j}(b)$ *and* $Q(b)$, *and find that all coefficients of them are nonnegative for* $k = 4$. *Unfortunately, even numerical computation are difficult for the case* $k \geq 5$.

Based on the lemma above, if we assume the positivity of the inflow conditions $u(x^-_{i-\frac{1}{2}}, y)$ and $u(x, y^-_{j-\frac{1}{2}})$, we can prove the positivity of $\bar{u}_{i,j}$ by taking the test function $w = \xi$ (extend $\xi = 0$ outside $K_{i,j}$) in the scheme (3.18) and using the fact that the source term f and coefficients a, b are positive. We can therefore obtain the result for the positivity-preserving property of the scheme (3.18) as follows.

Theorem 3.3.2. *For the constant coefficient stationary hyperbolic equation* (3.3), *if the source term and inflow conditions from upstream cells (including the inflow conditions on inflow boundary cells) are positive, then the cell averages of the scheme* (3.18) *are positive for the* Q^1, Q^2*-DG schemes with* $\lambda \geq 0$, *and* Q^3*-DG scheme with*

$\lambda = 0$.

Though we are not able to give rigorous proofs for the positivity-preserving property of the scheme (3.18) with $k = 3, \lambda > 0$ or $k > 3, \lambda \geq 0$ due to the difficulty of symbolically solving the large linear system (3.19), we can still investigate these cases numerically.

For any given values of a, b, λ, we can always solve for $\{\eta_{ij}\}_{i,j=1}^{k}$ numerically from the linear system (3.19) to obtain the values of η at the quadrature points $\{(\hat{r}_\alpha, \hat{r}_\beta)\}_{\alpha,\beta=1}^{k}$, $\{(-1, \tilde{r}_\alpha)\}_{\alpha=1}^{k+1}$ and $\{(\tilde{r}_\alpha, -1)\}_{\alpha=1}^{k+1}$ used on the right hand side of (3.18). The scheme is positivity-preserving if η is positive at all these quadrature points. Moreover, we can take advantage of the relationship $\eta(x, y; a, b, \lambda) = C\eta(x, y; Ca, Cb, C\lambda)$, $\forall C > 0$, to reduce the computation. If $\lambda \geq \max\{a, b\}$, we use $\eta(x, y; a, b, \lambda) = \frac{1}{\lambda}\eta(x, y; \frac{a}{\lambda}, \frac{b}{\lambda}, 1)$; otherwise we assume $a \geq \max\{b, \lambda\}$ without loss of generality and use $\eta(x, y; a, b, \lambda) = \frac{1}{a}\eta(x, y; 1, \frac{b}{a}, \frac{\lambda}{a})$. Therefore, we only need to numerically investigate the positivity of η in the two cases $0 \leq a, b \leq 1, \lambda = 1$ and $a = 1, 0 \leq b, \lambda \leq 1$.

We define

$$\eta_1(k) = \min_{0\leq a,b\leq 1} \min_{1\leq\alpha\leq k+1} \{\eta(-1, \tilde{r}_\alpha; a, b, 1), \eta(\tilde{r}_\alpha, -1; a, b, 1)\},$$

$$\eta_2(k) = \min_{0\leq b,\lambda\leq 1} \min_{1\leq\alpha\leq k+1} \{\eta(-1, \tilde{r}_\alpha; 1, b, \lambda), \eta(\tilde{r}_\alpha, -1; 1, b, \lambda)\},$$

$$\eta_3(k) = \min_{0\leq a,b\leq 1} \min_{1\leq\alpha,\beta\leq k} \eta(\hat{r}_\alpha, \hat{r}_\beta; a, b, 1),$$

$$\eta_4(k) = \min_{0\leq b,\lambda\leq 1} \min_{1\leq\alpha,\beta\leq k} \eta(\hat{r}_\alpha, \hat{r}_\beta; 1, b, \lambda),$$

and equally space 1000×1000 points of (a, b) or (b, λ) on $[0, 1] \times [0, 1]$ to approximate $\min_{0\leq a,b\leq 1}$ and $\min_{0\leq b,\lambda\leq 1}$, and give the approximate values $\tilde{\eta}_i(k), i = 1, 2, 3, 4$ in Table 3.1 and Table 3.2 for odd and even k, respectively. From the tables, we can observe that the minimum value of η at the quadrature points is

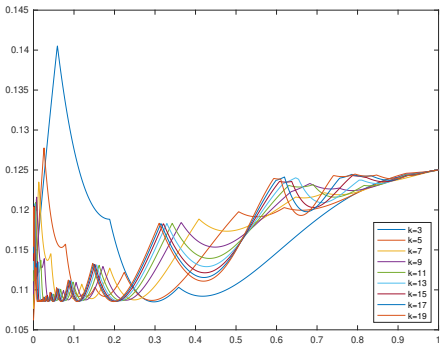
zero (machine epsilon) on boundaries when k is even, and strictly positive in all other cases. Moreover, we visualize a particular case $\lambda = 0$, and plot $h_1^k(b) = \min_{1\le\alpha\le k+1}\{\eta(-1,\tilde{r}_\alpha;1,b,0),\eta(\tilde{r}_\alpha,-1;1,b,0)\}$, $h_2^k(b) = \min_{1\le\alpha,\beta\le k}\eta(\hat{r}_\alpha,\hat{r}_\beta;1,b,0)$ for $b\in[0,1]$ in the Figure 3.1, from which we can observe the same pattern as shown in the tables.

k	3	5	7	9	11	13
$\tilde{\eta}_1$	4.75E-02	4.59E-02	4.65E-02	4.73E-02	4.80E-02	4.86E-02
$\tilde{\eta}_2$	4.75E-02	4.59E-02	4.65E-02	4.73E-02	4.80E-02	4.86E-02
$\tilde{\eta}_3$	5.67E-02	5.17E-02	5.01E-02	4.93E-02	4.90E-02	4.88E-02
$\tilde{\eta}_4$	5.67E-02	5.17E-02	5.01E-02	4.93E-02	4.90E-02	4.88E-02
k	15	17	19	-	-	-
$\tilde{\eta}_1$	4.91E-02	4.93E-02	4.92E-02	-	-	-
$\tilde{\eta}_2$	4.91E-02	4.93E-02	4.92E-02	-	-	-
$\tilde{\eta}_3$	4.86E-02	4.85E-02	4.85E-02	-	-	-
$\tilde{\eta}_4$	4.86E-02	4.85E-02	4.85E-02	-	-	-

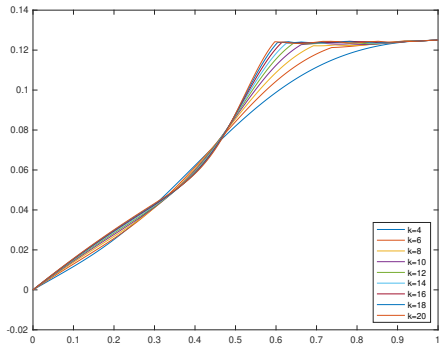
Table 3.1: $\tilde{\eta}_i(k), i = 1,2,3,4$ with odd k

k	4	6	8	10	12	14
$\tilde{\eta}_1$	-1.11E-15	-1.78E-15	-2.66E-15	-4.44E-15	-5.33E-15	-3.02E-14
$\tilde{\eta}_2$	-2.22E-16	-2.78E-16	-3.89E-16	-2.36E-16	-4.72E-16	-1.05E-15
$\tilde{\eta}_3$	5.98E-02	5.64E-02	5.51E-02	5.44E-02	5.40E-02	5.37E-02
$\tilde{\eta}_4$	5.98E-02	5.64E-02	5.51E-02	5.44E-02	5.40E-02	5.37E-02
k	16	18	20	-	-	-
$\tilde{\eta}_1$	-2.84E-14	-5.68E-14	-3.20E-14	-	-	-
$\tilde{\eta}_2$	-7.77E-16	-1.16E-15	-7.22E-16	-	-	-
$\tilde{\eta}_3$	5.33E-02	5.29E-02	5.27E-02	-	-	-
$\tilde{\eta}_4$	5.33E-02	5.29E-02	5.27E-02	-	-	-

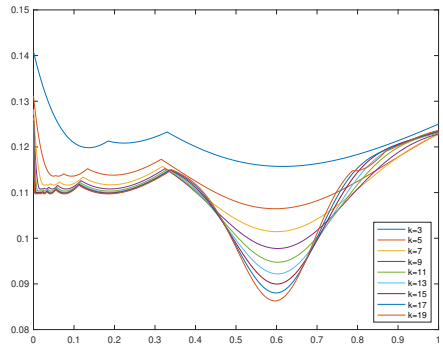
Table 3.2: $\tilde{\eta}_i(k), i = 1,2,3,4$ with even k



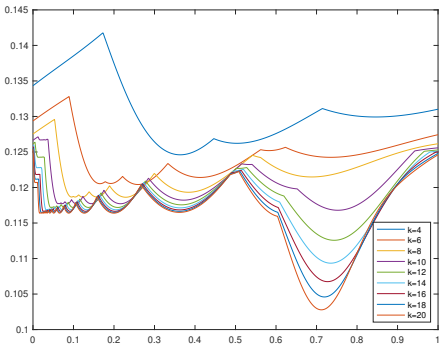
(a) $h_1^k(b)$, k is odd



(b) $h_1^k(b)$, k is even



(c) $h_2^k(b)$, k is odd



(d) $h_2^k(b)$, k is even

Figure 3.1: $h_1^k(b)$ and $h_2^k(b)$ for different k, 1000 points equally spaced on $[0, 1]$

3.3.3 Constant coefficient stationary hyperbolic equation in three space dimensions

Consider the constant coefficient stationary hyperbolic equation (3.4) with $f(x,y,z) \geq 0$ in Ω. Without loss of generality, we assume $a,b,c > 0$. The corresponding boundary conditions are given by $u(0,y,z) = g_1(y,z), u(x,0,z) = g_2(x,z)$ and $u(x,y,0) = g_3(x,y)$, where $g_1, g_2, g_3 \geq 0$.

The positivity-preserving Q^k-DG scheme of (3.4) is to seek $u \in V_h^k$, where k is odd, s.t. $\forall w \in V_h^k$

$$
\begin{aligned}
&-\oint_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \oint_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \oint_{z_{l-\frac{1}{2}}}^{z_{l+\frac{1}{2}}} (auw_x + buw_y + cuw_z - \lambda uw)\, dxdydz \\
&+ \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{z_{l-\frac{1}{2}}}^{z_{l+\frac{1}{2}}} au(x^-_{i+\frac{1}{2}}, y, z) w(x^-_{i+\frac{1}{2}}, y, z) dydz \\
&+ \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{z_{l-\frac{1}{2}}}^{z_{l+\frac{1}{2}}} bu(x, y^-_{j+\frac{1}{2}}, z) w(x, y^-_{j+\frac{1}{2}}, z) dxdz \\
&+ \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} cu(x, y, z^-_{l+\frac{1}{2}}) w(x, y, z^-_{l+\frac{1}{2}}) dxdy \\
&= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{z_{l-\frac{1}{2}}}^{z_{l+\frac{1}{2}}} au(x^-_{i-\frac{1}{2}}, y, z) w(x^+_{i-\frac{1}{2}}, y, z) dydz \\
&+ \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{z_{l-\frac{1}{2}}}^{z_{l+\frac{1}{2}}} bu(x, y^-_{j-\frac{1}{2}}, z) w(x, y^+_{j-\frac{1}{2}}, z) dxdz \\
&+ \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} cu(x, y, z^-_{l-\frac{1}{2}}) w(x, y, z^+_{l-\frac{1}{2}}) dxdy \\
&+ \oint_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \oint_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \oint_{z_{l-\frac{1}{2}}}^{z_{l+\frac{1}{2}}} fw dxdydz,
\end{aligned}
\tag{3.21}
$$

for $i = 1, \ldots, N_x, j = 1, \ldots, N_y, l = 1, \ldots, N_z$. If $x_{i-\frac{1}{2}} = 0$, we let $u(x^-_{i-\frac{1}{2}}, y, z) = g_1(y,z)$, similarly, if $y_{j-\frac{1}{2}} = 0$ or $z_{l-\frac{1}{2}} = 0$, we let $u(x, y^-_{j-\frac{1}{2}}, z) = g_2(x,z)$ or $u(x,y,z^-_{l-\frac{1}{2}}) = g_3(x,y)$, respectively. The sub-optimal convergence is observed in

numerical experiments due to the inaccurate quadrature rule adopted in the scheme.

Without loss of generality, we only consider the scheme (3.21) on the reference cell $K = [-1, 1]^3$, as any cell $K_{i,j,l}$ can be transferred to K by changing of coordinates with only rescales a, b, c, λ, f without altering their signs. We give the main results as follows.

Lemma 3.3.3. *Define* $\xi(x, y, z; a, b, c, \lambda) = (1 - x)(1 - y)(1 - z)\eta(x, y, z; a, b, c, \lambda)$, *where* $\eta(x, y, z; a, b, c, \lambda) = \sum_{i,j,l=1}^{k} \eta_{ijl}(a, b, c, \lambda)\ell_i(x)\ell_j(y)\ell_l(z)$, *and* $\{\eta_{ijl}(a, b, c, \lambda)\}_{i,j,l=1}^{k}$ *is the solution of the linear system*

$$\begin{aligned}
\sum_{i,j,l=1}^{k} \big(& a\left((1-\hat{r}_\alpha)(1-\hat{r}_\beta)(1-\hat{r}_\gamma)\ell_i'(x_\alpha)\delta_{\beta,j}\delta_{\gamma,l} - (1-\hat{r}_\beta)(1-\hat{r}_\gamma)\delta_{\alpha,i}\delta_{\beta,j}\delta_{\gamma,l}\right) \\
&+ b\left((1-\hat{r}_\alpha)(1-\hat{r}_\beta)(1-\hat{r}_\gamma)\ell_j'(x_\beta)\delta_{\alpha,i}\delta_{\gamma,l} - (1-\hat{r}_\alpha)(1-\hat{r}_\gamma)\delta_{\alpha,i}\delta_{\beta,j}\delta_{\gamma,l}\right) \\
&+ c\left((1-\hat{r}_\alpha)(1-\hat{r}_\beta)(1-\hat{r}_\gamma)\ell_l'(x_\gamma)\delta_{\alpha,i}\delta_{\beta,j} - (1-\hat{r}_\alpha)(1-\hat{r}_\beta)\delta_{\alpha,i}\delta_{\beta,j}\delta_{\gamma,l}\right) \\
&- \lambda(1-\hat{r}_\alpha)(1-\hat{r}_\beta)(1-\hat{r}_\gamma)\delta_{\alpha,i}\delta_{\beta,j}\delta_{\gamma,l}\big)\,\eta_{ijl} \\
&= -\frac{1}{8}, \quad \alpha, \beta, \gamma = 1, 2, \ldots, k,
\end{aligned} \tag{3.22}$$

then $\xi(x, y, z; a, b, c, \lambda) \in Q^k([-1, 1]^3)$ *satisfies*

$$\begin{aligned}
&-\fint_{-1}^{1}\fint_{-1}^{1}\fint_{-1}^{1} (av\xi_x + bv\xi_y + cv\xi_z - \lambda v\xi)\, dxdydz \\
&+ \int_{-1}^{1}\int_{-1}^{1} av(1, y, z)\xi(1, y, z)dydz + \int_{-1}^{1}\int_{-1}^{1} bv(x, 1, z)\xi(x, 1, z)dxdz \\
&+ \int_{-1}^{1}\int_{-1}^{1} cv(x, y, 1)\xi(x, y, 1)dxdy = \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} vdxdy,
\end{aligned} \tag{3.23}$$

for any $v \in Q^k([-1, 1]^3)$.

Moreover, for $k = 1$, *we have* $\xi(x, y, z; a, b, c, \lambda) \geq 0$ *for* $(x, y) \in [-1, 1]^3$.

Proof. By definition of $\xi(x, y, z)$, we can compute that $\xi_x(x, y, z) = (1-x)(1-y)(1-$

$z)\sum_{i,j,l=1}^{k}\eta_{ijl}\ell_i'(x)\ell_j(y)\ell_l(z)-(1-y)(1-z)\sum_{i,j,l=1}^{k}\eta_{ijl}\ell_i(x)\ell_j(y)\ell_l(z)$, $\xi_y(x,y,z)=(1-x)(1-y)(1-z)\sum_{i,j,l=1}^{k}\eta_{ijl}\ell_i(x)\ell_j'(y)\ell_l(z)-(1-x)(1-z)\sum_{i,j,l=1}^{k}\eta_{ijl}\ell_i(x)\ell_j(y)\ell_l(z)$, and $\xi_z(x,y,z)=(1-x)(1-y)(1-z)\sum_{i,j,l=1}^{k}\eta_{ijl}\ \ell_i(x)\ell_j(y)\ell_l'(z)-(1-x)(1-y)\sum_{i,j,l=1}^{k}\eta_{ijl}\ell_i(x)\ell_j(y)\ell_l(z)$, thereby it is easy to check that $\{\eta_{ijl}\}_{i,j,l=1}^{k}$ is the solution of the linear system (3.22) if and only if ξ satisfies

$$a\xi_x(\hat{r}_\alpha,\hat{r}_\beta,\hat{r}_\gamma)+b\xi_y(\hat{r}_\alpha,\hat{r}_\beta,\hat{r}_\gamma)+c\xi_z(\hat{r}_\alpha,\hat{r}_\beta,\hat{r}_\gamma)-\lambda\xi(\hat{r}_\alpha,\hat{r}_\beta,\hat{r}_\gamma)=-\frac{1}{8},\quad \alpha,\beta,\gamma=1,2,\ldots,k.$$

Moreover, $\xi(1,y,z)=\xi(x,1,z)=\xi(x,y,1)=0$ from the definition. Therefore, it follows from direct computation that (3.23) holds. When $k=1$, we can solve ξ from (3.22) to obtain $\xi(x,y,z)=\frac{1}{8(a+b+c+\lambda)}(1-x)(1-y)(1-z)\geq 0$ in $[-1,1]^3$. □

Based on the above lemma, we can obtain the result for the positivity-preserving property of the scheme (3.21) as follows.

Theorem 3.3.4. *For the constant coefficient stationary hyperbolic equation* (3.4)*, if the source term and inflow conditions from upstream cells (including the inflow conditions on inflow boundary cells) are positive, then the cell averages of the scheme* (3.21) *are positive for the Q^1-DG scheme.*

We are of course not satisfied with only Q^1-DG positivity-preserving scheme, which has first order convergence rate by numerical experiments. Similar to the two dimensional case, we numerically investigate the positivity of $\eta(x,y,z)$ at the quadrature points used on the right hand side of (3.21) for larger k. It suffices to consider two cases: $0\leq a,b,c\leq 1,\lambda=1$ and $a=1,0\leq b,c,\lambda\leq 1$ because of the property $\xi(x,y,z;a,b,c,\lambda)=C\xi(x,y,z;Ca,Cb,Cc,C\lambda),\forall C>0$ and the symmetry in x,y,z directions.

We define

$$\eta_1(k) = \min_{0\le a,b,c\le 1}\ \min_{1\le \alpha,\beta\le k+1}\{\eta(-1,\hat{r}_\alpha,\hat{r}_\beta;a,b,c,1),\eta(\hat{r}_\alpha,-1,\hat{r}_\beta;a,b,c,1),$$
$$\eta(\hat{r}_\alpha,\hat{r}_\beta,-1;a,b,c,1)\},$$
$$\eta_2(k) = \min_{0\le b,c,\lambda\le 1}\ \min_{1\le \alpha,\beta\le k+1}\{\eta(-1,\hat{r}_\alpha,\hat{r}_\beta;1,b,c,\lambda),\eta(\hat{r}_\alpha,-1,\hat{r}_\beta;1,b,c,\lambda),$$
$$\eta(\hat{r}_\alpha,\hat{r}_\beta,-1;1,b,c,\lambda)\},$$
$$\eta_3(k) = \min_{0\le a,b,c\le 1}\ \min_{1\le \alpha,\beta,\gamma\le k}\eta(\hat{r}_\alpha,\hat{r}_\beta,\hat{r}_\gamma;a,b,c,1),$$
$$\eta_4(k) = \min_{0\le b,c,\lambda\le 1}\ \min_{1\le \alpha,\beta,\gamma\le k}\eta(\hat{r}_\alpha,\hat{r}_\beta,\hat{r}_\gamma;1,b,c,\lambda)$$

and equally space $100\times100\times100$ points for $k=2,3,4$, $30\times30\times30$ points for $k=5,6,\ldots,10$, of (a,b,c) or (b,c,λ) on $[0,1]^3$ to approximate $\min_{0\le a,b,c\le1}$ and $\min_{0\le b,c,\lambda\le1}$. We give the approximate values $\tilde{\eta}_i(k), i=1,2,3,4$ in Table 3.3. From the table, we can observe that the minimum value of η at quadrature points is negative on boundaries when k is even, and strictly positive in all other cases, which suggest that we should use odd k for the purpose of positivity-preserving.

k	2	3	4	5	6	7
$\tilde{\eta}_1$	-4.44E-16	1.04E-02	-4.00E-15	9.75E-03	-1.60E-14	1.00E-02
$\tilde{\eta}_2$	-3.97E-06	1.04E-02	-1.63E-03	9.75E-03	-6.15E-03	1.00E-02
$\tilde{\eta}_3$	2.61E-02	1.39E-02	1.61E-02	1.20E-02	1.43E-02	1.14E-02
$\tilde{\eta}_4$	2.61E-02	1.39E-02	1.61E-02	1.20E-02	1.43E-02	1.14E-02
k	8	9	10	-	-	-
$\tilde{\eta}_1$	-1.70E-03	1.04E-02	-6.05E-03	-	-	-
$\tilde{\eta}_2$	-1.25E-02	1.04E-02	-2.01E-02	-	-	-
$\tilde{\eta}_3$	1.38E-02	1.18E-02	1.36E-02	-	-	-
$\tilde{\eta}_4$	1.38E-02	1.18E-02	1.36E-02	-	-	-

Table 3.3: $\tilde{\eta}_i(k), i=1,2,3,4$

3.4 Implementation of the algorithms

In this section, we summarize the results obtained in the previous sections and illustrate the implementation of the positivity-preserving algorithms.

Firstly, we introduce a robust version of the positivity-preserving limiter (3.5) used in practice. We set a small threshold $\epsilon > 0$, e.g. $\epsilon = 10^{-14}$, and denote by S the set of points where we want to preserve the positivity of function values. The set S must include the quadrature points used on the inflow boundaries in the schemes for the purpose of positivity-preserving. To be more precise, S must include the point $x_{i+\frac{1}{2}}$ on I_i in one space dimension, the points $\{(x_{i+\frac{1}{2}}, \tilde{y}_\alpha)\}_{\alpha=1}^{k+1}$, $\{(\tilde{x}_\alpha, y_{j+\frac{1}{2}})\}_{\alpha=1}^{k+1}$ on $K_{i,j}$ in two space dimensions, and the points $\{(x_{i+\frac{1}{2}}, \tilde{y}_\alpha, \tilde{z}_\beta)\}_{\alpha,\beta=1}^{k+1}$, $\{(\tilde{x}_\alpha, y_{j+\frac{1}{2}}, \tilde{z}_\beta)\}_{\alpha,\beta=1}^{k+1}$, $\{(\tilde{x}_\alpha, \tilde{y}_\beta, z_{l+\frac{1}{2}})\}_{\alpha,\beta=1}^{k+1}$ on $K_{i,j,l}$ in three space dimensions, where $\tilde{x}_\alpha = x_i + \frac{1}{2}\Delta x_i \tilde{r}_\alpha$, $\tilde{y}_\alpha = y_j + \frac{1}{2}\Delta y_j \tilde{r}_\alpha$, $\tilde{z}_\alpha = z_l + \frac{1}{2}\Delta z_l \tilde{r}_\alpha$, $\alpha = 1, 2, \ldots, k+1$ are the $(k+1)$-point Gauss-Legendre quadrature points in different directions. On a cell K with the cell average $\bar{u}_K \geq 0$, if $\bar{u}_K \leq \epsilon$, we take the modified solution $\tilde{u}_K(\mathbf{x}) \equiv \bar{u}_K$, otherwise, we take the modified solution as

$$\tilde{u}_K(\mathbf{x}) = \theta\left(u_K(\mathbf{x}) - \bar{u}_K\right) + \bar{u}_K, \text{where } \theta = \min\{\frac{\bar{u}_K - \epsilon}{\bar{u}_K - \min_{\mathbf{x}\in S} u_K(\mathbf{x})}, 1\}, \qquad (3.24)$$

where $\mathbf{x}$ denotes the coordinates in one, two or three space dimensions.

In one dimensional space, we compute the solution u_i on cell I_i based on the solution $\tilde{u}_{i-1}$ with $\tilde{u}_{i-1}(x) \geq 0, x \in S$. Once u_i is obtained from the scheme with $\bar{u}_i \geq 0$, we apply the above limiter to obtain a modified solution $\tilde{u}_i$, which will be used in the computation on the next cell.

Similarly, in two dimensional space, we compute the solution $u_{i,j}$ on cell $K_{i,j}$ based on the solution $\tilde{u}_{i-1,j}, \tilde{u}_{i,j-1}$ with $\tilde{u}_{i-1,j}(x,y), \tilde{u}_{i,j-1}(x,y) \geq 0, (x,y) \in S$. Once $u_{i,j}$ is obtained, we apply the above limiter to obtain the modified solution $\tilde{u}_{i,j}$, which will be used in later computations. In three dimensional space, we compute the solution $u_{i,j,l}$ on cell $K_{i,j,l}$ based on the solution $\tilde{u}_{i-1,j,l}, \tilde{u}_{i,j-1,l}, \tilde{u}_{i,j,l-1}$ with $\tilde{u}_{i-1,j,l}(x,y,z), \tilde{u}_{i,j-1,l}(x,y,z), \tilde{u}_{i,j,l-1}(x,y,z) \geq 0, (x,y,z) \in S$. Once $u_{i,j,l}$ is obtained, we apply the above limiter to obtain the modified solution $\tilde{u}_{i,j,l}$ and use it in the later computations.

We would like to remark that, under certain mesh size conditions, the positivity of the solution at the interfaces $x^-_{j+\frac{1}{2}}, j = 1, 2, \ldots, N$ in one dimensional space is automatically maintained even without the positivity-preserving limiter, i.e. $u^-_{j+\frac{1}{2}} \geq 0, j = 1, 2, \ldots, N$ provided $f, u_0 \geq 0$. This fact allows us to apply the positivity-preserving limiter simultaneously for all cells after the DG solution has been obtained for all cells. The detailed theorem and its proof are given in Appendix B.1.

3.5 Numerical tests

In this section, we perform numerical experiments to show the good performance of the positivity-preserving methods established in the previous sections. Many of the examples are taken from [90, 46, 91]. We take the set S in the positivity-preserving limiter of the Section 3.4 as the union of the necessary points introduced therein and 100 equally spaced points on 1D cells, or 50×50 equally spaced points on 2D cells, or $20 \times 20 \times 20$ equally spaced points on 3D cells. If not otherwise stated, we use uniform meshes with mesh sizes satisfying the conditions of positivity-preserving established in the previous sections.

Example 3.5.1. *We solve the equation* (3.1) *with* $a(x) = \frac{1}{2+\sin(4\pi x)}, \lambda = 0$ *and* $f(x) = x^2$ *on the domain* $\Omega = [0, 1]$. *The boundary condition is given by* $u(0) = 0$ *and the exact solution is* $u(x) = \frac{2}{3}x^3 + \frac{1}{3}\sin(4\pi x)x^3$. *We compute the solution based on the positivity-preserving scheme* (3.8) *and give the errors, order of convergence, and data about positivity in Tables 3.4 and 3.5 for the cases without and with the limiter, respectively. From the tables, we can see that the orders of convergence are optimal, and the negative values of the solution of the scheme without limiter are eliminated by the positivity-preserving limiter.*

k	N	L^1 error	order	L^∞ error	order	$\min u_h$
1	20	1.78E-03	-	2.89E-02	-	-8.71E-06
	40	4.41E-04	2.01	7.27E-03	1.99	-5.96E-07
	80	1.10E-04	2.00	1.83E-03	1.99	-3.81E-08
	160	2.75E-05	2.00	4.57E-04	2.00	-2.39E-09
	320	6.88E-06	2.00	1.14E-04	2.00	-1.50E-10
2	20	8.34E-05	-	1.53E-03	-	-3.46E-06
	40	1.06E-05	2.98	2.10E-04	2.86	-4.17E-07
	80	1.32E-06	3.01	2.91E-05	2.85	-5.24E-08
	160	1.64E-07	3.00	3.81E-06	2.93	-6.63E-09
	320	2.05E-08	3.00	4.86E-07	2.97	-8.38E-10
3	20	4.42E-06	-	1.15E-04	-	-9.64E-07
	40	2.76E-07	4.00	7.31E-06	3.98	-7.63E-08
	80	1.72E-08	4.01	4.57E-07	4.00	-5.03E-09
	160	1.07E-09	4.00	2.86E-08	4.00	-3.18E-10
	320	6.70E-11	4.00	1.79E-09	4.00	-2.00E-11
4	20	1.36E-07	-	3.41E-06	-	-6.96E-08
	40	4.23E-09	5.01	1.09E-07	4.97	-1.14E-09
	80	1.34E-10	4.98	3.39E-09	5.00	-1.80E-11
	160	4.17E-12	5.00	1.06E-10	5.00	-2.81E-13
	320	1.30E-13	5.00	3.32E-12	5.00	-4.40E-15

Table 3.4: Results of Example 3.5.1 without limiter

Example 3.5.2. *We solve the equation* (3.1) *with* $a(x) = 1, \lambda = 6000$ *and* $f(x) = \lambda\left(\frac{1}{9}\cos^4(x) + \epsilon\right) - \frac{4}{9}\cos^3(x)\sin(x)$ *on the domain* $\Omega = [0, \pi]$. *We take* $\epsilon = 10^{-14}$ *such that the source term is nonnegative. The boundary condition is given by* $u(0) = \frac{1}{9} + \epsilon$ *and the exact solution is* $u(x) = \frac{1}{9}\cos^4(x) + \epsilon$. *This example has been tested in*

k	N	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20	1.78E-03	-	2.89E-02	-	5.00
	40	4.41E-04	2.01	7.27E-03	1.99	2.50
	80	1.10E-04	2.00	1.83E-03	1.99	1.25
	160	2.75E-05	2.00	4.57E-04	2.00	0.63
	320	6.88E-06	2.00	1.14E-04	2.00	0.31
2	20	8.41E-05	-	1.52E-03	-	5.00
	40	1.07E-05	2.97	2.10E-04	2.85	2.50
	80	1.34E-06	3.00	2.92E-05	2.85	1.25
	160	1.67E-07	3.00	3.82E-06	2.93	0.63
	320	2.09E-08	3.00	4.88E-07	2.97	0.31
3	20	5.45E-06	-	1.13E-04	-	5.00
	40	3.84E-07	3.83	7.17E-06	3.98	2.50
	80	2.52E-08	3.93	4.46E-07	4.01	1.25
	160	1.61E-09	3.97	2.79E-08	4.00	0.63
	320	1.02E-10	3.99	1.74E-09	4.00	0.31
4	20	2.35E-07	-	3.50E-06	-	5.00
	40	5.81E-09	5.33	1.10E-07	4.99	2.50
	80	1.54E-10	5.23	3.42E-09	5.01	1.25
	160	4.44E-12	5.12	1.07E-10	5.00	0.63
	320	1.47E-13	4.92	3.34E-12	5.00	0.31

Table 3.5: Results of Example 3.5.1 with limiter

[46] with a rigorously proved high order conservative positivity-preserving method. However, since the inaccurate integral is adopted in our scheme, the results of our algorithm will be different. We collect the numerical errors, orders of convergence, and data about positivity in Tables 3.6 and 3.7 for the schemes (3.8) *without and with the limiter, respectively, from which we can observe the optimal convergence, and the negative values of the solution being eliminated by the positivity-preserving limiter.*

k	N	L^1 error	order	L^∞ error	order	$\min u_h$
1	20	2.14E-03	-	2.64E-03	-	-1.33E-03
	40	4.66E-04	2.20	6.76E-04	1.96	-2.90E-04
	80	7.94E-05	2.55	1.70E-04	2.00	-4.34E-05
	160	1.15E-05	2.78	4.21E-05	2.01	-1.41E-06
	320	2.41E-06	2.26	1.04E-05	2.02	-1.20E-10
	640	5.94E-07	2.02	2.51E-06	2.05	-1.66E-11
2	20	3.40E-04	-	4.59E-04	-	-2.68E-04
	40	6.79E-05	2.32	1.04E-04	2.15	-4.43E-05
	80	1.00E-05	2.76	1.86E-05	2.48	-8.12E-08
	160	1.25E-06	3.00	2.12E-06	3.13	-3.03E-07
	320	9.42E-08	3.72	1.55E-07	3.78	-8.65E-09
	640	6.07E-09	3.96	9.80E-09	3.98	-1.52E-10

Table 3.6: Results of Example 3.5.2 without limiter

k	N	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20	1.20E-03	-	2.64E-03	-	10.00
	40	2.97E-04	2.02	6.76E-04	1.96	5.00
	80	6.59E-05	2.17	1.70E-04	2.00	2.50
	160	1.14E-05	2.53	4.21E-05	2.01	1.25
	320	2.41E-06	2.24	1.04E-05	2.02	0.31
	640	5.94E-07	2.02	2.51E-06	2.05	0.16
2	20	2.41E-04	-	4.59E-04	-	15.00
	40	5.56E-05	2.11	1.04E-04	2.15	7.50
	80	9.98E-06	2.48	1.86E-05	2.48	2.50
	160	1.24E-06	3.01	2.12E-06	3.13	1.88
	320	9.41E-08	3.72	1.55E-07	3.78	0.94
	640	6.07E-09	3.95	9.80E-09	3.98	0.31

Table 3.7: Results of Example 3.5.2 with limiter

Example 3.5.3. *We solve the equation* (3.1) *with* $a(x) = 1 + x, \lambda = 10000$ *and* $f(x) = (\lambda+1)\left(\frac{1}{9}\cos^4(x) + \epsilon\right) - (1+x)\left(\frac{4}{9}\cos^3(x)\sin(x)\right)$ *on the domain* $\Omega = [0, 2\pi]$. *We take* $\epsilon = 2 \times 10^{-14}$ *such that the source term is nonnegative. The boundary condition is given by* $u(0) = \frac{1}{9} + \epsilon$ *and the exact solution is* $u(x) = \frac{1}{9}\cos^4(x) + \epsilon$. *We compute the solution using the scheme* (3.8) *and give the numerical errors, orders of convergence, and data about positivity in Tables 3.8 and 3.9 for the case without and with the limiter, respectively. From the tables, we can see that the orders of convergence are optimal, and that negative values appear without limiter and the positivity is maintained under the modification of the limiter.*

k	N	L^1 error	order	L^∞ error	order	$\min u_h$
1	20	8.35E-03	-	1.07E-02	-	-3.03E-03
	40	1.76E-03	2.25	2.83E-03	1.93	-5.01E-04
	80	4.10E-04	2.10	7.02E-04	2.01	-1.10E-04
	160	9.38E-05	2.13	1.73E-04	2.02	-1.60E-05
	320	2.15E-05	2.13	4.28E-05	2.02	-4.49E-07
	640	5.19E-06	2.05	1.05E-05	2.02	-3.21E-10
2	20	1.62E-03	-	1.26E-03	-	-4.51E-04
	40	3.73E-04	2.11	2.89E-04	2.12	-1.03E-04
	80	7.96E-05	2.23	7.02E-05	2.04	-1.62E-05
	160	1.29E-05	2.62	1.32E-05	2.41	-1.19E-06
	320	1.35E-06	3.26	1.85E-06	2.83	-1.45E-07
	640	1.03E-07	3.71	1.66E-07	3.48	-3.78E-09

Table 3.8: Results of Example 3.5.3 without limiter

Example 3.5.4. *We solve the equation* (3.2) *with* $a(u) = u^2 + 0.01, \lambda = 5$ *and* $f(x) = -8\sin(x)\cos^7(x)\left(3(\cos^8(x) + \epsilon)^2 + 0.01\right) + \lambda\left(\cos^8(x) + \epsilon\right)$ *on the domain* $\Omega = [0, \pi]$. *We take* $\epsilon = 10^{-14}$ *such that the source term is nonnegative. The boundary condition is given by* $u(0) = 1 + \epsilon$ *and the exact solution is* $u(x) = \cos^8(x) + \epsilon$. *We give the errors, orders of convergence, and data about positivity in Tables 3.10 and 3.11 for the scheme* (3.12) *with* $k = 1$ *and scheme* (3.13) *with* $k = 2$ *in the case without and with the limiter, respectively, with the same conclusion about accuracy and positivity-preserving as before.*

k	N	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20	6.44E-03	-	1.07E-02	-	10.00
	40	1.53E-03	2.08	2.83E-03	1.93	5.00
	80	3.75E-04	2.03	7.02E-04	2.01	3.75
	160	9.12E-05	2.04	1.73E-04	2.02	1.88
	320	2.15E-05	2.09	4.28E-05	2.02	0.94
	640	5.19E-06	2.05	1.05E-05	2.02	0.31
2	20	1.44E-03	-	1.37E-03	-	20.00
	40	3.36E-04	2.10	3.25E-04	2.07	10.00
	80	7.62E-05	2.14	7.37E-05	2.14	6.25
	160	1.29E-05	2.56	1.32E-05	2.49	3.13
	320	1.35E-06	3.26	1.85E-06	2.83	1.56
	640	1.03E-07	3.71	1.66E-07	3.48	0.78

Table 3.9: Results of Example 3.5.3 with limiter

k	N	L^1 error	order	L^∞ error	order	$\min u_h$
1	20	1.20E-01	-	1.28E-01	-	-1.17E-01
	40	8.11E-03	3.89	1.86E-02	2.78	-2.97E-03
	80	1.55E-03	2.39	4.39E-03	2.08	-2.03E-07
	160	3.87E-04	2.01	1.09E-03	2.02	-1.04E-13
	320	9.66E-05	2.00	2.71E-04	2.00	8.96E-15
	640	2.41E-05	2.00	6.76E-05	2.00	9.86E-15
2	20	8.56E-02	-	1.76E-01	-	-7.20E-02
	40	1.16E-02	2.88	4.47E-02	1.98	-1.44E-02
	80	9.28E-04	3.65	4.78E-03	3.22	-9.60E-05
	160	8.20E-05	3.50	4.31E-04	3.47	-4.44E-09
	320	8.27E-06	3.31	4.41E-05	3.29	-7.82E-14
	640	9.25E-07	3.16	4.89E-06	3.17	9.98E-15

Table 3.10: Results of Example 3.5.4 without limiter

k	N	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20	3.85E-02	-	1.28E-01	-	15.00
	40	7.02E-03	2.46	1.86E-02	2.78	12.50
	80	1.55E-03	2.18	4.39E-03	2.08	3.75
	160	3.87E-04	2.01	1.09E-03	2.02	1.25
	320	9.66E-05	2.00	2.71E-04	2.00	0.00
	640	2.41E-05	2.00	6.76E-05	2.00	0.00
2	20	1.89E-02	-	7.65E-02	-	50.00
	40	6.35E-03	1.58	4.26E-02	0.85	37.50
	80	9.15E-04	2.80	4.78E-03	3.15	18.75
	160	8.20E-05	3.48	4.31E-04	3.47	6.25
	320	8.27E-06	3.31	4.41E-05	3.29	1.25
	640	9.25E-07	3.16	4.89E-06	3.17	0.00

Table 3.11: Results of Example 3.5.4 with limiter

Example 3.5.5. *We solve the equation* (3.3) *with* $a = 0.7, b = 0.3, \lambda = 1.0$ *and* $f = 0$ *on the domain* $\Omega = [0,1] \times [0,1]$. *The boundary conditions are given by* $u(x,0) = 0$ *for* $0 \le x \le 1$ *and* $u(0,y) = \sin^6(\pi y)$ *for* $0 \le y \le 1$. *It is easy to check that the exact solution of the problem is*

$$u(x,y) = \begin{cases} 0, & y < \frac{b}{a}x \\ \sin^6(\pi(y - \frac{b}{a}x))e^{-\frac{\lambda}{a}x} & y \ge \frac{b}{a}x \end{cases}$$

We compute the solution based on the scheme (3.18) *with* $k = 1,2,3,4,5$. *The errors, orders of convergence and data about positivity are given in Tables 3.12 and 3.13 for the cases without and with positivity-preserving limiter, respectively, from which the sub-optimal convergence can be observed. Moreover, we plot the results of the scheme with the limiter for* $k = 1,2,3,4$ *on the* 40×40 *mesh in Figure 3.2, in which we put white dots on those cells where negative values appear before the limiting process.*

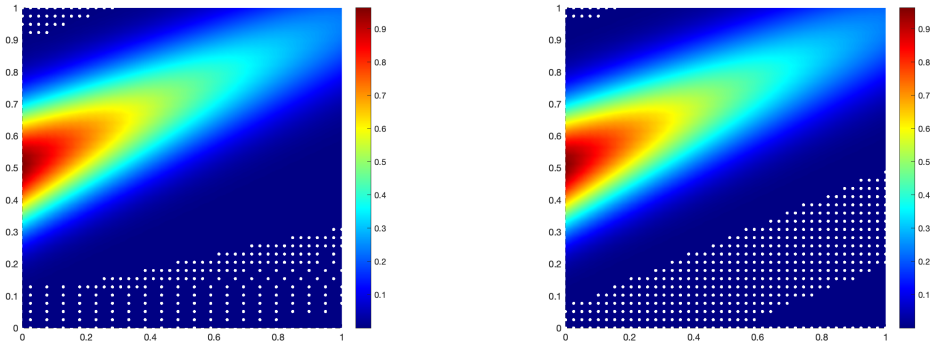
Example 3.5.6. *We solve the equation* (3.3) *with* $a = 0.6, b = 0.4, \lambda = 0$ *and* $f = 0$ *on the domain* $\Omega = [0,1]^2$. *The boundary condition is given by* $u(x,0) = 1$ *for*

k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	$\min u_h$
1	10×10	1.87E-02	-	2.96E-01	-	-1.01E-01
	20×20	6.04E-03	1.63	1.08E-01	1.46	-6.43E-03
	40×40	2.45E-03	1.30	4.80E-02	1.16	-2.42E-04
	80×80	1.14E-03	1.11	2.36E-02	1.02	-4.50E-06
	160×160	5.56E-04	1.03	1.18E-02	1.00	-7.34E-08
	320×320	2.77E-04	1.01	5.89E-03	0.99	-1.16E-09
2	10×10	1.70E-03	-	3.97E-02	-	-8.78E-03
	20×20	3.77E-04	2.17	1.27E-02	1.65	-2.37E-03
	40×40	9.11E-05	2.05	3.48E-03	1.86	-1.70E-04
	80×80	2.27E-05	2.01	9.17E-04	1.93	-2.53E-06
	160×160	5.68E-06	2.00	2.35E-04	1.97	-1.14E-07
	320×320	1.42E-06	2.00	5.95E-05	1.98	-2.89E-09
3	10×10	1.45E-04	-	4.56E-03	-	-6.59E-04
	20×20	1.49E-05	3.29	4.76E-04	3.26	-4.35E-05
	40×40	1.70E-06	3.13	6.13E-05	2.96	-7.75E-07
	80×80	2.06E-07	3.04	7.78E-06	2.98	-1.22E-08
	160×160	2.56E-08	3.01	9.80E-07	2.99	-1.92E-10
	320×320	3.19E-09	3.00	1.23E-07	2.99	-3.01E-12
4	10×10	1.23E-05	-	4.28E-04	-	-7.15E-05
	20×20	6.95E-07	4.14	3.63E-05	3.56	-2.49E-06
	40×40	4.13E-08	4.07	2.55E-06	3.83	-4.55E-08
	80×80	2.53E-09	4.03	1.68E-07	3.93	-7.49E-10
	160×160	1.57E-10	4.01	1.07E-08	3.97	-1.19E-11
	320×320	9.77E-12	4.00	6.77E-10	3.99	-1.87E-13

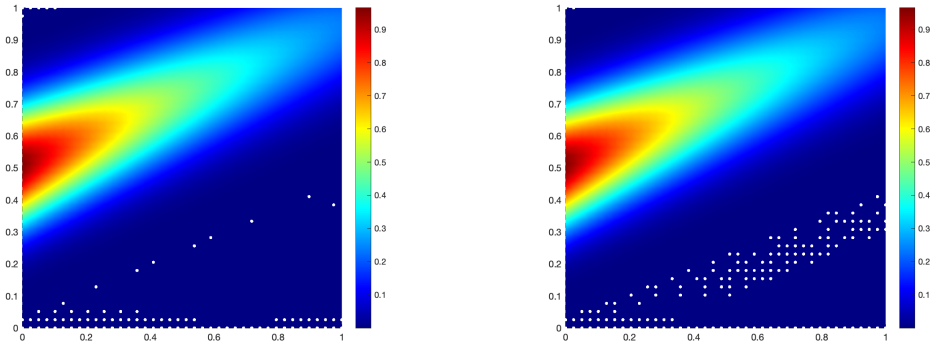
Table 3.12: Results of Example 3.5.5 without limiter

k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	Limited cells (%)
1	10×10	1.98E-02	-	3.17E-01	-	36.00
	20×20	6.13E-03	1.69	1.08E-01	1.56	21.25
	40×40	2.45E-03	1.32	4.80E-02	1.16	14.69
	80×80	1.14E-03	1.11	2.36E-02	1.02	5.27
	160×160	5.56E-04	1.03	1.18E-02	1.00	1.21
	320×320	2.77E-04	1.01	5.89E-03	0.99	0.23
2	10×10	2.34E-03	-	3.92E-02	-	49.00
	20×20	3.91E-04	2.58	1.27E-02	1.63	37.25
	40×40	9.12E-05	2.10	3.48E-03	1.86	25.50
	80×80	2.27E-05	2.01	9.17E-04	1.93	13.34
	160×160	5.68E-06	2.00	2.35E-04	1.97	6.25
	320×320	1.42E-06	2.00	5.95E-05	1.98	3.43
3	10×10	2.69E-04	-	5.18E-03	-	27.00
	20×20	1.80E-05	3.90	5.01E-04	3.37	13.25
	40×40	1.72E-06	3.39	6.13E-05	3.03	5.81
	80×80	2.06E-07	3.06	7.78E-06	2.98	3.97
	160×160	2.56E-08	3.01	9.80E-07	2.99	2.95
	320×320	3.19E-09	3.00	1.23E-07	2.99	2.45
4	10×10	3.29E-05	-	8.94E-04	-	29.00
	20×20	1.01E-06	5.03	3.81E-05	4.55	14.00
	40×40	4.37E-08	4.53	2.55E-06	3.90	9.31
	80×80	2.55E-09	4.10	1.68E-07	3.93	4.20
	160×160	1.57E-10	4.02	1.07E-08	3.97	2.28
	320×320	9.77E-12	4.01	6.77E-10	3.99	1.67

Table 3.13: Results of Example 3.5.5 with limiter



(a) $k = 1$ (b) $k = 2$



(c) $k = 3$ (d) $k = 4$

Figure 3.2: Solutions of Example 3.5.5 with limiter

$0 < x \leq 1$ *and* $u(0, y) = 0$ *for* $0 \leq y \leq 1$. *The exact solution of the problem is*

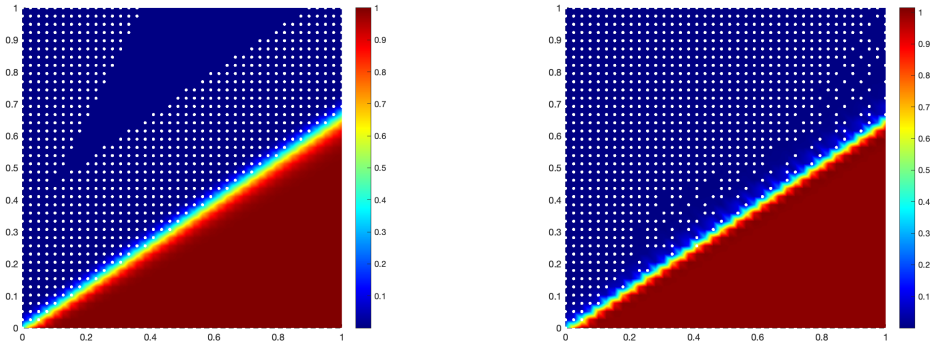
$$u(x, y) = \begin{cases} 1, & y < \frac{b}{a}x \\ 0, & y \geq \frac{b}{a}x \end{cases}$$

This problem can be interpreted as a two-dimensional radiative transfer model in transparent medium, see [46]. We plot the contours of the numerical solution solved from the scheme (3.18) *with positivity-preserving limiter for* $k = 1, 2, 3, 4$ *on* 40×40 *rectangular mesh in Figure 3.3, where white dots are drawn on the cells with negative values appearing before the limiting process. Moreover, we cut the profile of the solutions along the line* $x = 0.5$, *and compare them with the exact solution and the numerical solution solved without limiter in Figure 3.4, from which we can clearly see that the scheme without limiter produces negative values while the positivity of the solution is maintained with the limiter.*

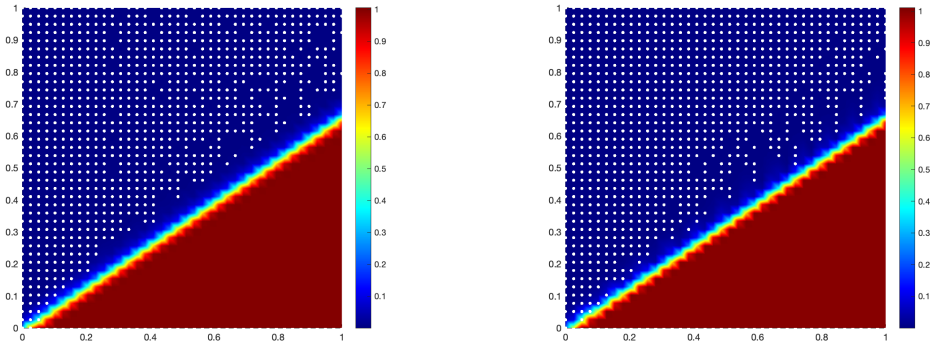
Example 3.5.7. *We solve the equation* (3.3) *with* $a = 0.6, b = 0.4, \lambda = 1$ *and* $f = 0$ *on the domain* $\Omega = [0, 1]^2$. *The boundary condition is given by* $u(x, 0) = 1$ *for* $0 < x \leq 1$ *and* $u(0, y) = 0$ *for* $0 \leq y \leq 1$. *The exact solution of the problem is*

$$u(x, y) = \begin{cases} e^{-\frac{\lambda}{b}y}, & y < \frac{b}{a}x \\ 0, & y \geq \frac{b}{a}x \end{cases}$$

The problem can be viewed as a two-dimensional radiative transfer model in purely absorbing medium, see [46]. We plot the contour of the numerical solution solved from the scheme (3.18) *with positivity-preserving limiter for* $k = 1, 2, 3, 4$ *on* 40×40 *rectangular mesh in Figure 3.5, where white dots are drawn on the cells with negative values appearing before the limiting process. Moreover, we cut the profile of the solution along the line* $x = 0.5$, *and compare them with the exact solutions and the numerical solutions solved without limiter in Figure 3.6, from which we can see the*

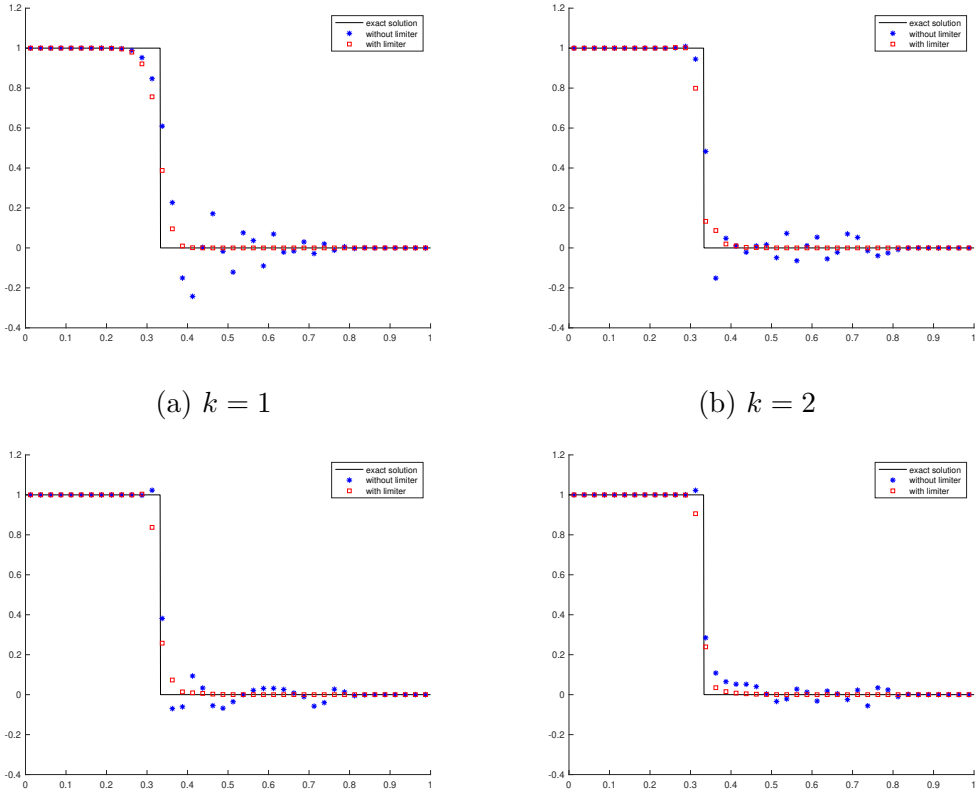


(a) $k = 1$ (b) $k = 2$



(c) $k = 3$ (d) $k = 4$

Figure 3.3: Solutions of Example 3.5.6 with limiter

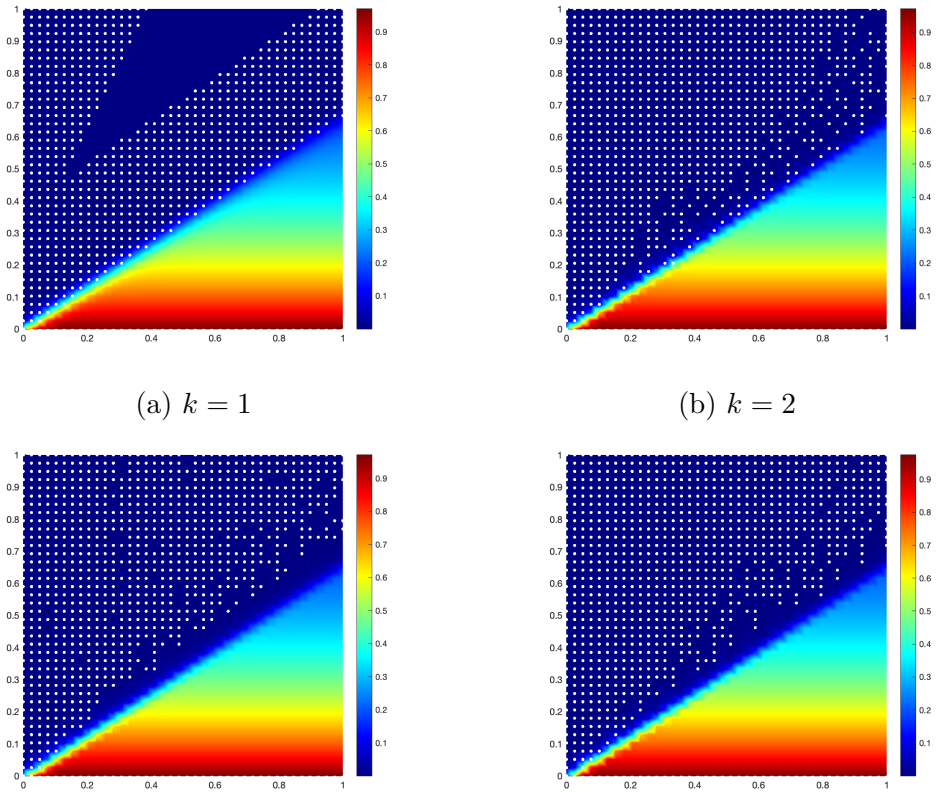


(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

Figure 3.4: Solutions of Example 3.5.6 cut along $x = 0.5$

positivity of solution is attained under the positivity-preserving limiter.



(a) $k = 1$ (b) $k = 2$

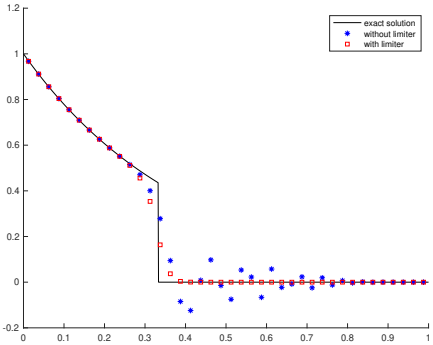
(c) $k = 3$ (d) $k = 4$

Figure 3.5: Solutions of Example 3.5.7 with limiter

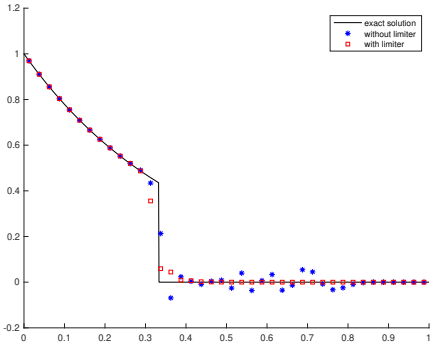
Example 3.5.8. *We consider the time-dependent linear problem* $u_t + u_x = 0$ *on the domain* $\Omega = [0, 2]$ *with boundary condition* $u(0) = 0$ *and discontinuous initial condition*

$$u_0(x) = \begin{cases} 1, & x \in [\frac{1}{4}, \frac{3}{4}] \\ 0, & \textit{otherwise.} \end{cases}$$

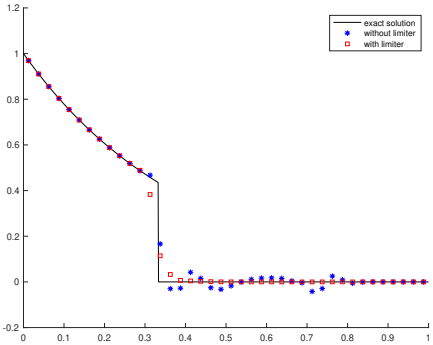
The solution of the problem is $u(x, t) = u_0(x - t)$. *We use the space-time DG approach that treats the time as an extra dimension, and solve the problem based on the scheme* (3.3). *The mesh is* 80×40 *on the space-time domain* $\Omega \times [0, T]$. *We plot the numerical solutions at* $t = 1$ *and compare it with the exact solution and the solution solved without limiter in Figure 3.7. From the figures, we can see*



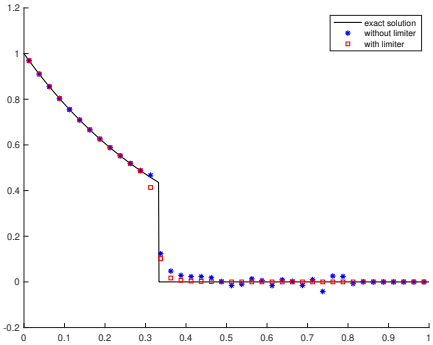
(a) $k = 1$



(b) $k = 2$



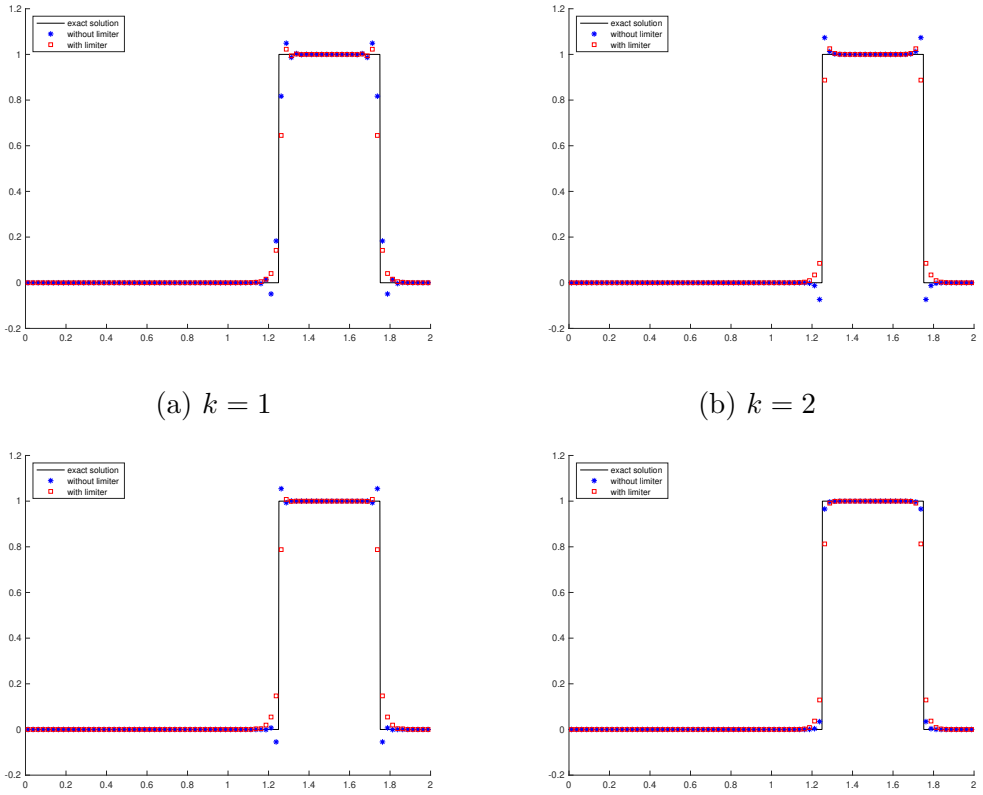
(c) $k = 3$



(d) $k = 4$

Figure 3.6: Solutions of Example 3.5.7 cut along $x = 0.5$

the solutions have negative values without the positivity-preserving limiter, while the positivity is maintained after the limiting process.



(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

Figure 3.7: Solutions of Example 3.5.8 at $T = 1$

Example 3.5.9. *We consider the time-dependent linear problem* $u_t + au_x + bu_y + \lambda u = f$ *with* $a = 0.7, b = 0.3, \lambda = 0.5$ *on the domain* $\Omega = [0, 1]^2$. *The initial condition is*

$$u_0(x, y) = \begin{cases} 0, & y < \frac{b}{a}x, \\ \sin^6(\pi(y - \frac{b}{a}x))e^{-2\lambda x}, & y \geq \frac{b}{a}x. \end{cases}$$

The exact solution of the problem is $u(x, y, t) = u_0(x - at, y - bt)e^{-\lambda t}$. *The boundary conditions are given according to the exact solution on the inflow boundaries. We use the space-time DG approach that treats the time as an extra dimension, and solve the problem based on the scheme* (3.4) *on the space-time domain* $\Omega \times T$ *with*

$T = 0.5$. *The errors, orders of convergence and data about positivity on the whole space-time domain* $\Omega \times [0, T]$ *are given in Tables 3.14 and 3.15 for the cases without and with limiter, respectively, from which we can observe sub-optimal convergence and the positivity of solution being maintained by the positivity-preserving limiter.*

k	$N_x \times N_y \times N_t$	L^1 error	order	L^∞ error	order	$\min u_h$
1	$10 \times 10 \times 5$	1.17E-02	-	8.74E-01	-	-4.52E-01
	$20 \times 20 \times 10$	4.80E-03	1.29	3.61E-01	1.27	-8.95E-02
	$40 \times 40 \times 20$	2.27E-03	1.08	1.87E-01	0.95	-7.47E-03
	$80 \times 80 \times 40$	1.12E-03	1.02	9.51E-02	0.98	-2.51E-04
	$160 \times 160 \times 80$	5.55E-04	1.01	4.80E-02	0.99	-2.10E-05
	$320 \times 320 \times 160$	2.77E-04	1.00	2.41E-02	0.99	-3.49E-06
3	$10 \times 10 \times 5$	1.19E-04	-	1.55E-02	-	-2.63E-03
	$20 \times 20 \times 10$	1.29E-05	3.20	1.82E-03	3.09	-1.40E-04
	$40 \times 40 \times 20$	1.52E-06	3.09	2.27E-04	3.00	-2.63E-06
	$80 \times 80 \times 40$	1.86E-07	3.03	2.89E-05	2.98	-4.65E-08
	$160 \times 160 \times 80$	2.31E-08	3.01	3.66E-06	2.98	-7.58E-10
	$320 \times 320 \times 160$	2.89E-09	3.00	4.61E-07	2.99	-3.68E-11
5	$10 \times 10 \times 5$	8.73E-07	-	1.43E-04	-	-5.46E-05
	$20 \times 20 \times 10$	2.47E-08	5.14	3.71E-06	5.27	-5.46E-07
	$40 \times 40 \times 20$	7.35E-10	5.07	1.14E-07	5.02	-1.11E-08
	$80 \times 80 \times 40$	2.25E-11	5.03	3.63E-09	4.98	-2.02E-10
	$160 \times 160 \times 80$	7.01E-13	5.01	1.15E-10	4.98	-3.37E-12
	$320 \times 320 \times 160$	2.20E-14	4.99	3.62E-12	4.99	-5.44E-14

Table 3.14: Results of Example 3.5.9 without limiter

3.6 Concluding remarks

In this chapter, we have constructed the high order conservative positivity-preserving DG method for stationary hyperbolic equations, via suitable quadrature rules in the DG framework.

In one space dimension, we propose the conservative positivity-preserving scheme with arbitrary high order for the variable coefficient equation (3.1) with $\lambda = 0$, and

k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	Limited cells (%)
1	$10 \times 10 \times 5$	1.28E-02	-	7.43E-01	-	78.20
	$20 \times 20 \times 10$	5.11E-03	1.33	4.81E-01	0.63	58.20
	$40 \times 40 \times 20$	2.29E-03	1.16	1.87E-01	1.36	40.61
	$80 \times 80 \times 40$	1.12E-03	1.03	9.51E-02	0.98	29.66
	$160 \times 160 \times 80$	5.55E-04	1.01	4.80E-02	0.99	21.89
	$320 \times 320 \times 160$	2.77E-04	1.00	2.41E-02	0.99	17.11
3	$10 \times 10 \times 5$	8.84E-04	-	1.02E-01	-	48.40
	$20 \times 20 \times 10$	5.36E-05	4.04	1.85E-02	2.46	34.55
	$40 \times 40 \times 20$	2.18E-06	4.62	8.90E-04	4.38	25.93
	$80 \times 80 \times 40$	1.92E-07	3.51	2.89E-05	4.94	22.40
	$160 \times 160 \times 80$	2.32E-08	3.05	3.66E-06	2.98	19.09
	$320 \times 320 \times 160$	2.89E-09	3.01	4.61E-07	2.99	16.19
5	$10 \times 10 \times 5$	3.35E-04	-	9.43E-02	-	42.40
	$20 \times 20 \times 10$	3.02E-05	3.47	3.03E-02	1.64	33.68
	$40 \times 40 \times 20$	1.01E-06	4.90	1.55E-03	4.29	27.34
	$80 \times 80 \times 40$	1.28E-08	6.30	3.83E-05	5.34	22.33
	$160 \times 160 \times 80$	1.24E-10	6.69	6.96E-07	5.78	18.93
	$320 \times 320 \times 160$	1.40E-12	6.47	1.15E-08	5.92	16.36

Table 3.15: Results of Example 3.5.9 with limiter

second and third orders for the variable coefficient equation (3.1) with $\lambda > 0$ and nonlinear equation (3.2) with $\lambda \geq 0$, which is a vast extension of the previous works in [46, 90] since only constant coefficient equations were addressed therein.

We also propose the conservative positivity-preserving scheme for constant coefficient equations with arbitrary high order in two space dimensions, and arbitrary odd order in three space dimensions, which improves the existing results in [46, 90] that are either non-conservative with high order accuracy or conservative with second order accuracy. We only give rigorous proofs for limited cases but the results of numerical experiments in Section 3.3 for general cases are very promising.

Finally, we would like to mention that, even though we have not discussed it in this chapter, one important application of the positivity-preserving schemes for stationary hyperbolic equations is to radiative transfer equations. One can refer to

[46, 90, 91] for details.

Chapter Four

On the conservation property of positivity-preserving discontinuous Galerkin methods for stationary hyperbolic equations

4.1 Introduction

The hyperbolic balance laws are important tools to investigate the phenomenon of flow and transport. In one space dimension, the scalar hyperbolic balance law is typically written in the form of

$$u_t + f(u)_x = s, \tag{4.1}$$

where u is the balanced quantity, f is the flux function, and s is the source term. In particular, if $s = 0$, the equation is called a hyperbolic conservation law and u is the conserved quantity.

Integrated over the spatial interval $[x_1, x_2]$, the hyperbolic equation (4.1) is transformed to the conservative formulation satisfied by the average of u on $[x_1, x_2]$

$$\frac{d\bar{u}}{dt} + \frac{1}{\Delta x}\left(f(x_2) - f(x_1)\right) = \bar{s}, \tag{4.2}$$

where $\Delta x = x_2 - x_1$, $\bar{u}(t) = \frac{1}{\Delta x}\int_{x_1}^{x_2} u(x,t)dx$, $f(x_i) = f(u(x_i,t)), i = 1, 2$, and $\bar{s} = \frac{1}{\Delta x}\int_{x_1}^{x_2} s(x,t)dx$.

Drawn from the formulation (4.2), numerous numerical schemes have been designed for the hyperbolic equation (4.1) in the conservative form

$$\frac{d\bar{u}_j}{dt} + \frac{1}{\Delta x_j}\left(\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}\right) = \bar{s}_j, \tag{4.3}$$

under the partition $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}], j = 0, \pm 1, \pm 2, \ldots$, for space, where $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$, $\hat{f}_{j\pm\frac{1}{2}}$ are numerical fluxes at $x_{j\pm\frac{1}{2}}$, $\bar{u}_j$ and $\bar{s}_j$ are cell averages of the numerical solution and the source term on I_j, respectively.

Conservation is of great importance for numerical methods for hyperbolic equations, as it is not only a numerical analogy of the theoretical property of hyperbolic balance laws, but more importantly also the Lax-Wendroff theorem [39], which can be briefly stated as follows,

Theorem 4.1.1. *Consider a sequence of grids with grid sizes $\Delta x_l, \Delta t_l$ converging to zero as $l \to \infty$, and a sequence of numerical solutions $U_l(x,t), l = 1, 2, \ldots$ computed from a consistent and conservative scheme for a hyperbolic equation on these grids. If U_l converges boundedly a.e. to a function u as $l \to \infty$, then u is a weak solution of the hyperbolic equation.*

Roughly speaking, conservative schemes guarantee correct shock speed determined by the Rankine-Hugoniot jump condition thanks to the mass conservation. To make it clear, we sum the equation (4.3) over the cells $I_j, I_{j+1}, \ldots, I_{j+r}$ to obtain the equation

$$\frac{d}{dt}\int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} u dx + \left(\hat{f}_{j+r+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}\right) = \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} s dx, \tag{4.4}$$

which enforces the correct speed of the shock (if there is a shock in the interval $[x_{j-\frac{1}{2}}, x_{j+r+\frac{1}{2}}]$) since the total mass $\int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} u dx$ depends on the shock location. On the other hand, non-conservative schemes could produce shocks with totally wrong speed and converge to a spurious solution. A well-known example [41] is the Burgers' equation in the non-conservative form $u_t + uu_x = 0$ discretized by a natural upwinding finite difference scheme $u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} u_j^n \left(u_j^n - u_{j-1}^n\right)$ with the initial condition $u_j^0 = \mathbf{1}(j < 0)$, where $\mathbf{1}(\cdot)$ is the indicator function. It's easy to check that $u_j^n \equiv u_j^0, \forall n$ for the scheme, which is wrong as the physical solution $u(x,t)$ with the initial condition $u_0(x) = \mathbf{1}(x < 0)$ is $u(x,t) = u_0(x - \frac{1}{2}t)$. For deeper discussions about conservative schemes and their significance for time-dependent hyperbolic equations, one can refer

to Chapter 12 in the monograph [41].

The discontinuous Galerkin (DG) method is one of the most popular numerical methods solving hyperbolic equations for its advantages in geometric flexibility, local mass conservation, easiness of parallelization and high order accuracy. The DG method was first proposed in 1973 by Reed et al. [62] to compute the stationary linear transport equation, and first analyzed by Lesaint et al. [40] in 1974. It was later developed into the Runge-Kutta discontinuous Galerkin (RKDG) method in a series of papers by Cockburn et al. [17, 16, 14, 13, 18] for time-dependent nonlinear hyperbolic problems. The classic DG scheme for the hyperbolic equation (4.1) is to find $u \in V$, such that

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_t v dx - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} f(u) v_x dx + \hat{f}_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} = \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} s v dx, \quad \forall v \in V, \tag{4.5}$$

for all j, where V is a piecewise polynomial space and $v^{\pm}_{j+\frac{1}{2}} = \lim_{\epsilon \to 0^+} v(x_{j+\frac{1}{2}} \pm \epsilon)$ denote the right and left limits of v at $x_{j+\frac{1}{2}}$. Taking $v = 1$ on I_j and zero anywhere else in (4.5), we recover the conservative formulation (4.3) satisfied by cell averages. Therefore, the unmodulated DG scheme is conservative for hyperbolic equations.

However, conservation is not the only issue we need to consider for numerical schemes. It is well-known that the scalar hyperbolic conservation laws satisfy the maximum-principle, e.g. its physical solution satisfies $m \leq u(x,t) \leq M, \forall x \in \mathbb{R}, t > 0$, where $m = \min_{x \in \mathbb{R}} u(x,0)$ and $M = \max_{x \in \mathbb{R}} u(x,0)$. These results hold also for periodic boundary condition and for compactly supported solutions, as well as in higher dimensions. If $m = 0$, the property is also called positivity-preserving. For the hyperbolic balance law (4.1) with $s \geq 0$, the solution is positivity-preserving, provided the initial condition and inflow boundary conditions are nonnegative. It is important to keep the positivity/maximum-principle, besides mass conservation, in

numerical schemes, otherwise the numerical solution is not only physically unacceptable, but also may cause severe robustness issues due to the change of hyperbolicity, or when coupled with other physical systems.

There have been intensive studies on positivity-preserving and maximum-principle-satisfying methods. The genuinely high order maximum-principle-satisfying DG method was proposed in 2010 by Zhang et al. [93] for scalar hyperbolic equations, and is rapidly developed for different problems ever since, e.g. for the Euler equations [94, 95], Navier-Stokes equations [92], shallow water equations [82], convection-diffusion equations [96, 11], and fluid flow in porous media [29, 12, 88, 28], among others.

The framework of the positivity-preserving DG methods is composed of two parts. The first part is problem-dependent, which is to obtain the solution with provable nonnegative cell averages, probably under certain mesh-size conditions, from the unmodulated DG scheme. Once the cell averages are guaranteed nonnegative, a scaling limiter, which preserves cell averages and does not destroy the original accuracy of the solution [93], [92], is employed such that the entire solution is modified into nonnegative. It is of great importance for the scaling limiter to preserve cell averages for time-dependent problems. We explain the significance of this principle by an example of the positivity-preserving DG method for (4.1) based on the forward Euler or backward Euler time discretization. The equation satisfied by the cell average of the solution on I_j is given as follows

$$\frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} + \frac{1}{\Delta x_j}\left(\hat{f}^m_{j+\frac{1}{2}} - \hat{f}^m_{j-\frac{1}{2}}\right) = \bar{s}_j^m, \tag{4.6}$$

where n denotes the time level t^n and m is taken as n or $n+1$ in the forward-Euler or backward-Euler time discretization, respectively. We denote the modified solution by

$\tilde{u}$ to distinguish it from the unmodulated solution u. Since $\bar{\tilde{u}}_j^n = \bar{u}_j^n$ and $\bar{\tilde{u}}_j^{n+1} = \bar{u}_j^{n+1}$ from the property of the limiter, we have the same equation satisfied by the modified solution:

$$\frac{\bar{\tilde{u}}_j^{n+1} - \bar{\tilde{u}}_j^n}{\Delta t} + \frac{1}{\Delta x_j}\left(\hat{f}^m_{j+\frac{1}{2}} - \hat{f}^m_{j-\frac{1}{2}}\right) = \bar{s}_j^m. \tag{4.7}$$

Thus the Lax-Wendroff theorem and a discrete analogy of (4.4) are satisfied by the modified solution $\tilde{u}$ as well, which guarantees the numerical solution (if it converges) converging to a weak solution with the correct shock speed. This is why preserving cell averages is desired in positivity-preserving/maximum-principle satisfying limiters for time-dependent hyperbolic equations.

Besides time-dependent problems, the stationary hyperbolic equations have also attracted the attention of many researchers. The stationary hyperbolic equations have wide applications in steady-state flow and transport problems. Moreover, they are building blocks of the discrete-ordinate method (DOM) for radiative transfer equations (RTE), see [24, 37]. They are also encountered in implicit time-discretization for time-dependent hyperbolic problems. Similar to the time-dependent problems, the physical solutions of stationary hyperbolic equations are also positivity-preserving, provided the inflow boundary conditions and source terms are nonnegative. There is a series of works on the positivity-preserving DG methods for stationary hyperbolic equations to enhance the stability of numerical algorithms. In 2016, Yuan et al. [90] proposed a rotational limiter based non-conservative positivity-preserving algorithm for constant coefficients stationary hyperbolic equations in one and two space dimensions on structured meshes. Later on, the algorithm is extended to triangular meshes in two space dimensions by Zhang et al. [91] based on a rotational limiter defined on triangles, which is still non-conservative. In 2018, Ling et al. [46] improved the results in [90] in one dimensional space by proving the positivity of cell averages of the unmodulated DG scheme, which results in a high order

conservative positivity-preserving DG method by adopting the scaling limiter [93] from time-dependent problems. However, the unmodulated scheme fails to preserve the positivity of cell averages in two space dimensions [46], thus only a second order conservative positivity-preserving scheme was proposed therein by an augmentation of the DG function space. The above works only focus on equations with constant coefficients, and higher than second order conservative methods are unavailable in two and three space dimensions. In Chapter 3, we developed high order conservative positivity-preserving algorithms for linear variable coefficient and nonlinear stationary hyperbolic equations in one dimension, and constant coefficients equations in two and three dimensions in [86].

Here, we would like to note that, the notion of conservation in the aforementioned works for stationary hyperbolic equations are different from the notion to be clarified in this chapter. The previous notion of conservation in positivity-preserving limiters, coming directly from time-dependent problems to preserve the cell average, is not very suitable for stationary problems.

To show this, we consider the stationary equation

$$f(u)_x + \lambda u = s(x), \tag{4.8}$$

where $f(u)$ is a smooth flux function with unchanged wind direction: $f'(u) > 0, \forall u$, and $\lambda, s(x) \geq 0$ are nonnegative coefficient and source, respectively. The equation (4.8) could come from the backward Euler discretization of the time-dependent problem (4.1), with the correspondence $u = u^n, \lambda = \frac{1}{\Delta t}$ and $s(x) = \frac{1}{\Delta t} u^{n-1}(x) + s(x, t^n)$. The linear stationary hyperbolic equations, with the main applications in RTE [90, 46, 91], will also be discussed in later sections. Throughout the chapter, we always assume the wind direction of flux in hyperbolic equations does not change,

for both nonlinear equations and linear ones, and always use the upwind flux in the DG schemes.

The unmodulated DG scheme with the upwind flux for the equation (4.8) is to find $u \in V$, such that

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \lambda uv dx - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} f(u) v_x dx + f(u^-_{j+\frac{1}{2}}) v^-_{j+\frac{1}{2}} - f(u^-_{j-\frac{1}{2}}) v^+_{j-\frac{1}{2}} = \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} svdx, \quad \forall v \in V. \tag{4.9}$$

In the implementation, because of the upwind mechanism of the equation and scheme, we sweep the computation from the left to the right cells, i.e. we obtain the solution u_{j-1} on I_{j-1} before computing u_j on I_j, and then solve u_{j+1} on I_{j+1}, and so forth. Same as the time-dependent cases, by taking $v = 1$ on I_j and zeros on other cells, we obtain the conservation equation satisfied by the cell averages as follows

$$\lambda \Delta x_j \bar{u}_j + f(u^-_{j+\frac{1}{2}}) = f(u^-_{j-\frac{1}{2}}) + \Delta x_j \bar{s}_j,$$

where the right hand side is known when solving u_j. In the positivity-preserving algorithms, the limiter has been used for u_{j-1} on the upstream cell I_{j-1} when computing u_j, to provide a physically relevant inflow flux, thus the actual conservation equation satisfied by the cell average on I_j is

$$\lambda \Delta x_j \bar{u}_j + f(u^-_{j+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \Delta x_j \bar{s}_j, \tag{4.10}$$

where $\tilde{u}^-_{j-\frac{1}{2}} = \tilde{u}_{j-1}(x_{j-\frac{1}{2}})$ denotes the value of the modified solution on I_{j-1} evaluated at $x_{j-\frac{1}{2}}$.

If the limiter is "conservative" in the sense of preserving cell averages, i.e. $\bar{\tilde{u}}_j = \bar{u}_j$,

then from (4.10) we have the following equation satisfied by the cell average

$$\lambda\Delta x_j\bar{\tilde{u}}_j + f(\tilde{u}^-_{j+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \Delta x_j\bar{s}_j + \left(f(\tilde{u}^-_{j+\frac{1}{2}}) - f(u^-_{j+\frac{1}{2}})\right), \tag{4.11}$$

Summing the above equations over cells $I_j, I_{j+1}, \ldots, I_{j+r}$ yields

$$\begin{aligned}\lambda\int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} \tilde{u}dx + f(\tilde{u}^-_{j+r+\frac{1}{2}}) &= f(\tilde{u}^-_{j-\frac{1}{2}}) + \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} s dx + \sum_{i=j}^{j+r}\left(f(\tilde{u}^-_{i+\frac{1}{2}}) - f(u^-_{i+\frac{1}{2}})\right)\\ &\neq f(\tilde{u}^-_{j-\frac{1}{2}}) + \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} s dx.\end{aligned} \tag{4.12}$$

We shall give concrete examples in the numerical section to show that the limiter preserving cell averages for stationary hyperbolic equations could produce solutions with wrong total mass/ shock location, even for the simplest hyperbolic equation $u_t + u_x = 0$ discretized implicitly in time.

On the other hand, if we define the local mass in stationary hyperbolic equations as the sum of the cell average and the outflow flux, and develop limiters such that the modified solution $\tilde{u}$ preserves the local mass on I_j in the sense that

$$\lambda\Delta x_j\bar{\tilde{u}}_j + f(\tilde{u}^-_{j+\frac{1}{2}}) = \lambda\Delta x_j\bar{u}_j + f(u^-_{j+\frac{1}{2}}), \tag{4.13}$$

then we have the local conservation formulation

$$\lambda\Delta x_j\bar{\tilde{u}}_j + f(\tilde{u}^-_{j+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \Delta x_j\bar{s}_j, \tag{4.14}$$

and the global conservation formulation

$$\lambda\int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} \tilde{u}dx + f(\tilde{u}^-_{j+r+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} s dx, \tag{4.15}$$

satisfied by the modified solution $\tilde{u}$.

Moreover, we can easily prove the Lax-Wendroff theorem for the modified solution $\tilde{u}$. Indeed, $\forall \phi \in C_c^\infty(\mathbb{R})$, we can multiply $\phi(x_j)$ on both sides of the equation (4.14), sum the resulting equations over all cells, and use the summation by parts, to obtain

$$\lambda \sum_j \Delta x_j \bar{\tilde{u}}_j \phi(x_j) + \sum_j f(\tilde{u}^-_{j-\frac{1}{2}}) \left(\phi(x_{j-1}) - \phi(x_j)\right) = \sum_j \Delta x_j \bar{s}_j \phi(x_j). \tag{4.16}$$

We can rewrite the equation in the integration form as

$$\begin{aligned} &\lambda \int_{-\infty}^{\infty} \sum_j \bar{\tilde{u}}_j \phi(x_j) \mathbf{1}(x_{j-\frac{1}{2}} \le x \le x_{j+\frac{1}{2}}) dx - \int_{-\infty}^{\infty} \sum_j f(\tilde{u}^-_{j-\frac{1}{2}}) \phi_x(x_{j-\frac{1}{2}}) \mathbf{1}(x_{j-1} \le x \le x_j) dx \\ &- \Delta x \int_{-\infty}^{\infty} r(x) dx = \int_{-\infty}^{\infty} \sum_j \bar{s}_j \phi(x_j) \mathbf{1}(x_{j-\frac{1}{2}} \le x \le x_{j+\frac{1}{2}}) dx, \end{aligned} \tag{4.17}$$

by the Taylor expansion of $\phi(x)$, where $x_j = \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})$, $\Delta x = \max_j \Delta x_j$, and $r(x)$ is the remainder of the Taylor expansion, which is uniformly bounded and compactly supported for any partition with $\Delta x \le 1$. If the modified solution $\tilde{u}$ converges to a function u almost everywhere with uniformly bounded total variation as the mesh size refines to zero, then applying the dominated convergence theorem (DCT) for (4.17), we yield

$$-\int_{-\infty}^{\infty} f(u) \phi_x dx + \lambda \int_{-\infty}^{\infty} u \phi dx = \int_{-\infty}^{\infty} s \phi dx, \tag{4.18}$$

which is the definition of the weak solution of (4.8).

To this end, we would like to give a remark on the definition (4.13) for conservative limiters. Indeed, it is quite reasonable to preserve the sum of the cell average and outflow fluxes in limiters, as any decrease in cell average caused by limiters should

be remedied to the mass on the downstream cells via increasing the outflow fluxes in the current cell, and vise versa.

As we will see in later sections, based on this novel definition of conservation, the positivity-preserving DG methods for stationary hyperbolic equations are straightforward and their implementations are simple. We only discuss the linear stationary hyperbolic equations in one and two space dimensions, and nonlinear stationary equations in one dimension to save space, but the method can be directly extended to higher dimensions with various meshes and a class of nonlinear hyperbolic systems with eigenvalues being of the same sign. As important applications, the algorithms developed in this chapter can be used in the positivity-preserving algorithm for radiative transfer equations and implicit time discretization for time-dependent hyperbolic problems, see the numerical section and refer to [46] for more details.

The rest of the chapter is organized as follows. In Section 4.2, we establish the positivity-preserving discontinuous Galerkin method for stationary linear hyperbolic equations in one space dimension and construct the conservative limiters with rigorous proofs for the accuracy. We extend the method and limiters to rectangular meshes and triangular meshes in two dimensions in Section 4.3 and Section 4.4, respectively. The positivity-preserving technique for stationary nonlinear hyperbolic equations is studied in Section 4.5, which is focused on one dimension to save space but the method can be extended to higher dimensions directly as in the linear case. In Section 4.6, we give ample numerical tests to demonstrate the accuracy and effectiveness of our positivity-preserving methods for stationary equations as well as the applications in implicit time discretization for time-dependent problems. Finally, we end up with some concluding remarks in Section 4.7.

4.2 Linear stationary hyperbolic equations in one dimension

In this section, we study the high order conservative positivity-preserving discontinuous Galerkin method for the linear stationary hyperbolic equation

$$(a(x)u)_x + \lambda u = s(x), \quad x \in \Omega = (0,1), \tag{4.19}$$

with $0 < a_* \le a(x) \le a^*$ for some positive constants a_*, a^*, and $\lambda, s(x) \ge 0$. We assign the inflow boundary condition $u(0) = u_0 \ge 0$ for the equation. The other case $a(x) < 0$ with boundary condition $u(1) = u_0 \ge 0$ can be transformed to this case by the change of variable $x' = 1 - x$ in (4.19), thus we omit the discussion. We assume λ is constant for simplicity, as we are mainly concerned with the applications of the model in the discrete-ordinate method (DOM) for radiative transfer equations (RTE) and implicit time-discretization for time-dependent hyperbolic problems, where λ is constant for both cases. However, there is not essential difficulty to extend the positivity-preserving technique to the variable case $\lambda(x) \ge 0$.

We adopt the partition $0 = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N+\frac{1}{2}} = 1$ for Ω and denote the j-th cell by $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ with the length $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$, for $j = 1, 2, \ldots, N$. The function space V of the P^k-DG scheme is defined as

$$V = \left\{ v \in L^2(\Omega) : v|_{I_j} \in P^k(I_j), j = 1, 2, \ldots, N \right\},$$

where $P^k(I_j)$ denotes the space of polynomials of order no greater than k on the cell I_j. We define the cell average of $v \in V$ on I_j as $\bar{v}_j = \frac{1}{\Delta x_j} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x)dx$, and its left and right limits at the interface $x_{j+\frac{1}{2}}$ as $v^{\pm}_{j+\frac{1}{2}} = v(x_{j+\frac{1}{2}} \pm 0)$. Moreover, we denote

by $v_j = v|_{I_j}$ for $v \in V$, $j = 1, 2, \ldots, N$, for convenience.

The positivity-preserving P^k-DG scheme of the equation (4.19) is to find $u \in V$, such that

$$\begin{aligned} &- \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} a(x) u v_x dx + a(x_{j+\frac{1}{2}}) u^-_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \lambda u v dx \\ &= a(x_{j-\frac{1}{2}}) \tilde{u}^-_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} s v dx, \quad \forall v \in P^k(I_j) \end{aligned} \tag{4.20}$$

for $j = 1, 2, \ldots, N$, where we define $u^-_{\frac{1}{2}} = u_0$. We would like to emphasize that, the calculation of u_j is based on the modified solution on the upstream cells, thus we use $\tilde{u}_{j-1}$ on the right hand side of the scheme (4.20). Once u_j is solved from the scheme, we employ the positivity-preserving limiter to be introduced later to obtain the modified solution $\tilde{u}_j$, and use it in the calculation of u_{j+1}, and so forth.

Assume the quadrature rules adopted in the scheme (4.20) is accurate for integrals of k-th order polynomials. Taking the test function $v = 1$ on I_j in the scheme (4.20), we obtain the following equation satisfied by the local mass

$$\lambda \Delta x_j \bar{u}_j + a(x_{j+\frac{1}{2}}) u^-_{j+\frac{1}{2}} = a(x_{j-\frac{1}{2}}) \tilde{u}^-_{j-\frac{1}{2}} + \Delta x_j \bar{s}_j, \tag{4.21}$$

For convenience, we define $LHS(w_j) = \lambda \Delta x_j \bar{w}_j + a(x_{j+\frac{1}{2}}) w^-_{j+\frac{1}{2}}$, for $w_j \in P^k(I_j)$, to be the amount of local mass of w_j on I_j. Since $\tilde{u}^-_{j-\frac{1}{2}} \geq 0$ on the right hand side of (4.21), we have $LHS(u_j) \geq 0$. The conservative limiter should satisfy $LHS(\tilde{u}_j) = LHS(u_j)$, where u_j and $\tilde{u}_j$ are the unmodulated and modified solutions on I_j, respectively.

There are two types of limiters to be developed throughout the chapter, where

the type-1 limiter requires the DG scheme to use the Gauss-Radau quadrature rule of $k+1$ points for numerical integration and only guarantees the positivity of modified solution at the Gauss-Radau points (in other parts, it can be negative), while the type-2 limiter does not. We denote the Gauss-Radau points on I_j by $\hat{x}_\alpha, \alpha = 1, 2, \ldots, k+1$ with $\hat{x}_{k+1} = x_{j+\frac{1}{2}}$, and the corresponding weights by $\hat{\omega}_\alpha, \alpha = 1, 2, \ldots, k+1$ with $\sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha = 1$.

The type-1 limiter for u_j is defined as follows:

$$\tilde{u}_j(x) = \theta_j \hat{u}_j(x), \quad \hat{u}_j(x) = \sum_{\alpha=1}^{k+1} u_j^+(\hat{x}_\alpha)\ell_\alpha(x), \tag{4.22}$$

where $z^+ = \max\{z, 0\}$ is the positive part of a real number z, $\ell_\alpha(x)$ is the Lagrange basis at the Gauss-Radau points $\{\hat{x}_\beta\}_{\beta=1}^{k+1}$ with $\ell_\alpha(\hat{x}_\beta) = \delta_{\alpha,\beta}$, and $\theta_j = \frac{LHS(u_j)}{LHS(\hat{u}_j)} \in [0, 1]$. Note that the integral in $LHS(\cdot)$ is evaluated by the Gauss-Radau quadrature, thus $0 \leq LHS(u_j) \leq LHS(\hat{u}_j)$. In the case $LHS(u_j) = LHS(\hat{u}_j) = 0$, we take $\theta_j = 1$. In practice, this case can be avoided by taking $\theta_j = \frac{LHS(u_j)+\varepsilon}{LHS(\hat{u}_j)+\varepsilon}$, where ε is a very small positive number, e.g. $\varepsilon = 10^{-16}$.

It is clear that the limiter (4.22) is conservative in the sense that $LHS(\tilde{u}_j) = LHS(u_j)$, and $\tilde{u}_j \geq 0$ at the Gauss-Radau points $\{\hat{x}_\alpha\}_{\alpha=1}^{k+1}$. More importantly, we have the result of accuracy for the limiter as follows:

Lemma 4.2.1. *Consider the solution u_j of the scheme* (4.20) *with accuracy $O(\Delta x_j^{k+1})$. If $\lambda = 0$, the error introduced by the limiter* (4.22) *is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^{k+1})$. If $\lambda > 0$, the error introduced by the limiter* (4.22) *is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^k)$, but the error is optimal at the downstream point, i.e. $|\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}| = O(\Delta x_j^{k+1})$.*

Proof. We can decompose the error as

$$e = u_j - \tilde{u}_j = (u_j - \hat{u}_j) + (\hat{u}_j - \tilde{u}_j) = e_1 + e_2. \tag{4.23}$$

For $e_1 = u_j - \hat{u}_j$, we have the estimate

$$\begin{aligned} |e_1(x)| &= |\hat{u}_j(x) - u_j(x)| \\ &= \left| \sum_{\alpha=1}^{k+1} u_j^+(\hat{x}_\alpha)\ell_\alpha(x) - \sum_{\alpha=1}^{k+1} u_j(\hat{x}_\alpha)\ell_\alpha(x) \right| \\ &= \left| \sum_{\alpha=1}^{k+1} u_j^-(\hat{x}_\alpha)\ell_\alpha(x) \right| \\ &\le \sum_{\alpha=1}^{k+1} |\ell_\alpha(x)| \cdot \max_{1\le\alpha\le k+1} u_j^-(\hat{x}_\alpha) \\ &\le \Lambda_k \cdot O(\Delta x_j^{k+1}) = O(\Delta x_j^{k+1}), \quad \forall x \in I_j, \end{aligned} \tag{4.24}$$

where $z^- = -\min\{z, 0\}$ denotes the negative part of a real number z and $\Lambda_k = \max_{x\in I_j} \sum_{\alpha=1}^{k+1} |\ell_\alpha(x)|$ is the Lebesgue constant. Note that $u_j^-(\hat{x}_\alpha) = O(\Delta x_j^{k+1}), \alpha = 1, \ldots, k+1$, since the exact solution is nonnegative. Therefore $||e_1||_{L^\infty(I_j)} = O(\Delta x_j^{k+1})$.

For e_2, we have $e_2 = \hat{u}_j - \tilde{u}_j = (1-\theta_j)\hat{u}_j$. If $\lambda = 0$, we have $e_2 \equiv 0$ since $\theta_j = 1$, which follows from the observation that $u^-_{j+\frac{1}{2}} = \hat{u}^-_{j+\frac{1}{2}} \ge 0$ due to (4.20), (4.21). If $\lambda > 0$, we have the estimate for $e_2(x)$ as follows,

$$\begin{aligned} |e_2(x)| = (1-\theta_j)\,|\hat{u}_j(x)| &= \left(1 - \frac{LHS(u_j)}{LHS(\hat{u}_j)}\right)|\hat{u}_j(x)| = \frac{LHS(\hat{u}_j - u_j)}{LHS(\hat{u}_j)}\,|\hat{u}_j(x)| \\ &= \frac{\lambda\Delta x_j\left(\bar{\hat{u}}_j - \bar{u}_j\right) + a(x_{j+\frac{1}{2}})\left(\hat{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}\right)}{\lambda\Delta x_j\bar{\hat{u}}_j + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\,|\hat{u}_j(x)| \\ &\le \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\bar{\hat{u}}_j + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\,|\hat{u}_j(x)| \end{aligned} \tag{4.25}$$

In particular, at the Gauss-Radau points, we have the following estimates from (4.25),

$$\begin{aligned}
|e_2(\hat{x}_\beta)| &\le \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\bar{\hat{u}}_j}\hat{u}_j(\hat{x}_\beta) \\
&= \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha\hat{u}_j(\hat{x}_\alpha)}\hat{u}_j(\hat{x}_\beta) \\
&\le \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\hat{\omega}_\beta\hat{u}_j(\hat{x}_\beta)}\hat{u}_j(\hat{x}_\beta) \\
&\le \left(\hat{\omega}_\beta^{-1} + \frac{a^*}{\lambda\hat{\omega}_\beta\Delta x_j}\right)||e_1||_{L^\infty(I_j)} \\
&= O(\Delta x_j^k), \quad \text{for } \beta = 1, 2, \ldots, k+1,
\end{aligned} \tag{4.26}$$

thus,

$$\begin{aligned}
|e_2(x)| =& |\sum_{\alpha=1}^{k+1} e_2(\hat{x}_\alpha)\ell_\alpha(x)| \le \sum_{\alpha=1}^{k+1}|\ell_\alpha(x)| \cdot \max_{1\le\alpha\le k+1}|e_2(\hat{x}_\alpha)| \\
&\le \Lambda_k \cdot O(\Delta x_j^k) = O(\Delta x_j^k), \quad \forall x \in I_j,
\end{aligned} \tag{4.27}$$

i.e. $||e_2||_{L^\infty(I_j)} = O(\Delta x_j^k)$.

In particular, at the downstream point $x_{j+\frac{1}{2}}$, it follows from (4.25) that

$$\begin{aligned}
|e_2(x_{j+\frac{1}{2}})| \le& \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha\hat{u}_j(\hat{x}_\alpha) + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} \\
=& \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha\hat{u}_j(\hat{x}_\alpha) + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} \\
&+ \frac{a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha\hat{u}_j(\hat{x}_\alpha) + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} \\
\le& \frac{\Delta x_j||e_1||_{L^\infty(I_j)}}{\Delta x_j\hat{\omega}_{k+1}\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} + \frac{a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} \\
\le& \left(1 + \hat{\omega}_{k+1}^{-1}\right)||e_1||_{L^\infty(I_j)} = O(\Delta x_j^{k+1}).
\end{aligned} \tag{4.28}$$

Gathering all results above and using the triangle inequalities, we finish the proof of Lemma 4.2.1. □

We would like to note that, the error estimates in Lemma 4.2.1 is sharp and the result cannot be improved by any conservative limiters, which can be illustrated by a concrete example given as follows.

Example 4.2.1. *We assume $\lambda > 0$. Consider the numerical approximation $u_j(x) = x_{j+\frac{1}{2}} - x - \Delta x_j^{k+1}$ of the exact solution $u_j^{exact}(x) = x_{j+\frac{1}{2}} - x$ on the cell $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. The modified solution $\tilde{u}_j$ of any conservative limiters should satisfy $\lambda\Delta x_j\bar{\tilde{u}}_j + a(x_{j+\frac{1}{2}})\tilde{u}^-_{j+\frac{1}{2}} = \lambda\Delta x_j\bar{u}_j + a(x_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}}$. Since $u^-_{j+\frac{1}{2}} = -\Delta x_j^{k+1}$ and $\tilde{u}^-_{j+\frac{1}{2}} \geq 0$, we have $\bar{u}_j - \bar{\tilde{u}}_j = \frac{a(x_{j+\frac{1}{2}})}{\lambda\Delta x_j}\left(\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}\right) \geq \frac{a(x_{j+\frac{1}{2}})}{\lambda\Delta x_j}\left(0 - (-\Delta x_j^{k+1})\right) = a(x_{j+\frac{1}{2}})\lambda^{-1}\Delta x_j^k$, which implies that $\tilde{u}_j$ is at most k-th order accurate.*

The type-1 limiter only preserves the positivity of modified solutions at the Gauss-Radau points and we must use the Gauss-Radau quadrature to evaluate integrals in the scheme (4.20), which may not be satisfactory in some applications. We now introduce the type-2 limiter, which is positivity-preserving on the whole cell or at any desired points, and exempts the requirement on quadrature rules.

The type-2 limiter is defined as follows,

$$\tilde{u}_j(x) = \theta_j\hat{u}_j(x), \quad \hat{u}_j(x) = u_j(x) + \epsilon_j, \tag{4.29}$$

where $\epsilon_j = -\min\{\min_{x\in S} u_j(x), 0\}$, $S \subset I_j$ is the set of points where we want to preserve the positivity of the solution, and $\theta_j = \frac{LHS(u_j)}{LHS(\hat{u}_j)} \in [0, 1]$.

It is clear that the limiter (4.29) is conservative in the sense that $LHS(\tilde{u}_j) = LHS(u_j)$ and $\tilde{u}_j \geq 0$ on S. More importantly, we have the accuracy result for the

limiter as follows:

Lemma 4.2.2. *Consider the solution u_j of the scheme* (4.20) *with accuracy* $O(\Delta x_j^{k+1})$. *If $\lambda = 0$, the error introduced by the limiter* (4.29) *is* $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j)$. *If $\lambda > 0$, the error introduced by the limiter* (4.29) *is* $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^k)$. *Nevertheless, at the downstream point, the errors in both cases are optimal, i.e.* $|\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}| = O(\Delta x_j^{k+1})$.

Proof. For simplicity, we assume $S = I_j$. We have the same decomposition $e = u_j - \tilde{u}_j = (u_j - \hat{u}_j) + (\hat{u}_j - \tilde{u}_j) = e_1 + e_2$ for the error as (4.23).

It is clear that $||e_1||_{L^\infty(I_j)} = ||u_j - \hat{u}_j||_{L^\infty(I_j)} = \epsilon_j = O(\Delta x_j^{k+1})$ by the definitions. For e_2, we have $e_2 = \hat{u}_j - \tilde{u}_j = (1 - \theta_j)\hat{u}_j$. If $\lambda = 0$, we have $||e_2||_{L^\infty(I_j)} = (1 - \theta_j)||\hat{u}_j||_{L^\infty(I_j)} = O(\Delta x_j)$, since $||\hat{u}_j||_{L^\infty(I_j)} = O(\Delta x_j)$ if $\theta_j < 1$. If $\lambda > 0$, we have the estimates for e_2 exactly the same as (4.25), (4.26), (4.27), and end up with the result $||e_2||_{L^\infty(I_j)} = O(\Delta x_j^k)$.

At the downstream point $x_{j+\frac{1}{2}}$, the estimate for e_2 is exactly the same as (4.28), thereby $|e_2(x_{j+\frac{1}{2}})| = O(\Delta x_j^{k+1})$.

Gathering all results above and using the triangle inequalities, we finish the proof of Lemma 4.2.2. □

The estimates in Lemma (4.2.2) is sharp, i.e. it could happen that $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j)$ if $\lambda = 0$, which can be illustrated by the following example.

Example 4.2.2. *We assume $\lambda = 0$. Consider the exact solution u_j^{exact} on I_j with $u_j^{exact}(\hat{x}_\alpha) = \Delta x_j$ for $1 \le \alpha \le k-1$ and $u_j^{exact}(\hat{x}_\alpha) = 0$ for $\alpha = k, k+1$, and its numerical approximation $u_j = \sum_{\alpha=1}^{k+1} u_j^{exact}(\hat{x}_\alpha)\ell_\alpha(x) - \Delta x_j^{k+1}\ell_k(x)$. It is clear that*

u_j is flattened to $\tilde{u}_j \equiv 0$ by the limiter (4.29)*, which is only of the accuracy $O(\Delta x)$.*

Remark 4.2.1. *The above discussions are based on the assumption that λ is a constant. However, in the backward Euler discretization for time-dependent problems, λ is of the order $\frac{1}{\Delta t}$, as demonstrated in the introduction. If we take the common CFL condition $\Delta t \propto \Delta x$ in this case, the accuracy of both the type-1 and type-2 limiters is optimal, which is clear from the estimates in the proofs. The same conclusion applies to later sections.*

Since the accuracy of both type-1 and type-2 limiters is optimal at the downstream points of cells, the possible non-optimal errors introduced by the limiters do not propagate to downstream cells, which makes the limited positivity-preserving DG solution having the optimal order of accuracy in the sense of downstream points of cells.

Collecting the Lemma 4.2.1 and 4.2.2, we attain the following theorem for the positivity-preserving DG method of the equation (4.19).

Theorem 4.2.3. *For the linear stationary hyperbolic equation* (4.19)*, if the source term and inflow boundary condition are nonnegative, then the solution of the scheme* (4.20) *modified by the limiter* (4.22) *or* (4.29) *is nonnegative, with the local accuracy established in Lemma 4.2.1 and 4.2.2, respectively.*

4.3 Linear stationary hyperbolic equations in two dimensions on rectangular meshes

In this section, we study the high order conservative positivity-preserving discontinuous Galerkin method in two space dimensions on rectangular meshes for the linear stationary hyperbolic equation

$$(a(x,y)u)_x + (b(x,y)u)_y + \lambda u = s(x,y), \quad (x,y) \in \Omega = (0,1)^2, \tag{4.30}$$

with $0 < a_* \le a(x,y) \le a^*$ and $0 < b_* \le b(x,y) \le b^*$ for some positive constants a_*, a^*, b_*, b^*, and $\lambda, s(x,y) \ge 0$. We assign the inflow boundary conditions $u(x,0) = g_1(x) \ge 0$ and $u(0,y) = g_2(y) \ge 0$ for the equation. The cases $a(x,y) < 0$ and/or $b(x,y) < 0$ can be transformed to this case by the change of variables $x' = 1 - x$ and/or $y' = 1 - y$, thus we omit the discussion.

We partition the domain Ω by $0 \le x_{\frac{1}{2}} < x_{\frac{3}{2}} < \ldots < x_{N_x+\frac{1}{2}} = 1$ and $0 \le y_{\frac{1}{2}} < y_{\frac{3}{2}} < \ldots < y_{N_y+\frac{1}{2}} = 1$ in x and y directions, respectively, and denote by $K_{i,j} = I_i \times J_j = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ the cells in Ω with the area $|K_{i,j}| = \Delta x_i \Delta y_j$, where $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \Delta y_j = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$, $i = 1, 2, \ldots N_x, j = 1, 2, \ldots, N_y$. Moreover, we assume the meshes are regular in the refinement, i.e. $\max_{i,j}\{\Delta x_i, \Delta y_j\} \le \rho \min_{i,j}\{\Delta x_i, \Delta y_j\}$ for some constant ρ that is independent of mesh sizes, and denote by $h = \min_{i,j}\{\Delta x_i, \Delta y_j\}$. The function space V of the Q^k-DG scheme is defined as

$$V = \left\{ v \in L^2(\Omega) : v|_{K_{i,j}} \in Q^k(K_{i,j}), i = 1, 2, \ldots, N_x, j = 1, 2, \ldots, N_y \right\},$$

where $Q^k(K)$ denotes the space of tensor products of polynomials of order no greater than k on the cell K.

Similar to the one space dimension, we define the cell average of $v \in V$ on $K_{i,j}$ as $\bar{v}_{i,j} = \frac{1}{\Delta x_i \Delta y_j} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} v(x,y)dxdy$, and its left/right and lower/upper limits on the vertical and horizontal cell interfaces by $v(x^{\pm}_{i+\frac{1}{2}}, y) = \lim_{\epsilon \to 0^+} v(x_{i+\frac{1}{2}} \pm \epsilon, y)$ and $v(x, y^{\pm}_{j+\frac{1}{2}}) = \lim_{\epsilon \to 0^+} v(x, y_{j+\frac{1}{2}} \pm \epsilon)$, respectively. Moreover, we denote by $v_{i,j} = v|_{K_{i,j}}$ for $v \in V, i = 1, 2, \ldots, N_x, j = 1, 2, \ldots, N_y$.

The positivity-preserving Q^k-DG scheme of the equation (4.30) on rectangular meshes is to find $u \in V$, such that

$$
\begin{aligned}
&- \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(auv_x + buv_y - \lambda uv\right) dxdy \\
&+ \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} au(x^-_{i+\frac{1}{2}}, y)v(x^-_{i+\frac{1}{2}}, y)dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} bu(x, y^-_{j+\frac{1}{2}})v(x, y^-_{j+\frac{1}{2}})dx \\
&= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} a\tilde{u}(x^-_{i-\frac{1}{2}}, y)v(x^+_{i-\frac{1}{2}}, y)dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} b\tilde{u}(x, y^-_{j-\frac{1}{2}})v(x, y^+_{j-\frac{1}{2}})dx \\
&+ \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} svdxdy, \quad \forall v \in Q^k(K_{i,j}),
\end{aligned}
\tag{4.31}
$$

for $i = 1, 2, \ldots, N_x, j = 1, 2, \ldots, N_y$, where we define $u(x, y^-_{\frac{1}{2}}) = \mathcal{I}\left(g_1\right)(x)$ and $u(x^-_{\frac{1}{2}}, y) = \mathcal{I}\left(g_2\right)(y)$ on the inflow boundaries, with $\mathcal{I}$ denoting the polynomial interpolation at the quadrature points on cell interfaces. In the computation, we solve $u_{i,j}$ on cell $K_{i,j}$ based on the modified solutions $\tilde{u}_{i-1,j}$ and $\tilde{u}_{i,j-1}$ on upstream cells. Once $u_{i,j}$ is obtained, we employ the positivity-preserving limiters to get the modified solution $\tilde{u}_{i,j}$ and use it in the computations on the downstream cells.

Taking the test function $v = 1$ on $K_{i,j}$ in the scheme (4.31), we obtain the

following equation satisfied by the local mass

$$\begin{aligned}&\lambda\Delta x_i\Delta y_j\bar{u}_{i,j}+\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}au(x^-_{i+\frac{1}{2}},y)dy+\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}bu(x,y^-_{j+\frac{1}{2}})dx\\&=\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}a\tilde{u}(x^-_{i-\frac{1}{2}},y)dy+\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}b\tilde{u}(x,y^-_{j-\frac{1}{2}})dx+\Delta x_i\Delta y_j\bar{s}_{i,j}.\end{aligned}\tag{4.32}$$

We define $LHS(w_{i,j})=\lambda\Delta x_i\Delta y_j\bar{w}_{i,j}+\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}aw(x^-_{i+\frac{1}{2}},y)dy+\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}bw(x,y^-_{j+\frac{1}{2}})dx$, for $w_{i,j}\in Q^k(K_{i,j})$, to be the amount of local mass of $w_{i,j}$ on $K_{i,j}$. Moreover, we define $LHS^b(w_{i,j})=\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}aw(x^-_{i+\frac{1}{2}},y)dy+\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}bw(x,y^-_{j+\frac{1}{2}})dx$ for the total outflow flux. Since $\tilde{u}_{i-1,j},\tilde{u}_{i,j-1}\geq 0$ on the right hand side of (4.32), we have $LHS(u_{i,j})\geq 0$. In particular, if $\lambda=0$, then $LHS^b(u_{i,j})=LHS(u_{i,j})\geq 0$.

Similar to the one dimensional case, there are two types of limiters, in which the type-1 limiter depends on the Gauss-Radau quadrature while the type-2 limiter does not.

The type-1 limiter for $u_{i,j}$ is defined as follows:

$$\begin{aligned}&\tilde{u}_{i,j}(x,y)=\theta^2_{i,j}\hat{u}_{i,j}(x,y),\quad \hat{u}_{i,j}(x,y)=\hat{u}^o_{i,j}(x,y)+\theta^1_{i,j}\hat{u}^b_{i,j}(x,y),\\&\hat{u}^o_{i,j}(x,y)=\sum_{\alpha=1}^{k}\sum_{\beta=1}^{k}u^+_{i,j}(\hat{x}_\alpha,\hat{y}_\beta)\ell_\alpha(x)\ell_\beta(y),\\&\hat{u}^b_{i,j}(x,y)=\sum_{\alpha=1}^{k}u^+_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})\ell_\alpha(x)\ell_{k+1}(y)+\sum_{\beta=1}^{k+1}u^+_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)\ell_{k+1}(x)\ell_\beta(y),\end{aligned}\tag{4.33}$$

where $\{\hat{x}_\alpha\}_{\alpha=1}^{k+1}$ and $\{\hat{y}_\beta\}_{\beta=1}^{k+1}$ are the Gauss-Radau points on the intervals $[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]$ and $[y_{j-\frac{1}{2}},y_{j+\frac{1}{2}}]$, respectively, with $\hat{x}_{k+1}=x_{i+\frac{1}{2}}$ and $\hat{y}_{k+1}=y_{j+\frac{1}{2}}$, we abuse notations to denote by $\ell_\alpha(x)$ and $\ell_\beta(y)$ the Lagrange basis at $\{\hat{x}_\alpha\}_{\alpha=1}^{k+1}$ and $\{\hat{y}_\beta\}_{\beta=1}^{k+1}$, respectively, $\theta^1_{i,j}=\max\{\frac{LHS^b(u^b_{i,j})}{LHS^b(\hat{u}^b_{i,j})},0\}\in[0,1]$, $\theta^2_{i,j}=\frac{LHS(u_{i,j})}{LHS(\hat{u}_{i,j})}\in[0,1]$, $u^b_{i,j}(x,y)=\sum_{\alpha=1}^{k}u_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})\ell_\alpha(x)\ell_{k+1}(y)+\sum_{\beta=1}^{k+1}u_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)\ \ell_{k+1}(x)\ell_\beta(y)$. In particular, if

$\lambda = 0$, we have $\theta^1_{i,j} = \frac{LHS^b(u^b_{i,j})}{LHS^b(\hat{u}^b_{i,j})}$ and $\theta^2_{i,j} = 1$. We denote $u^o_{i,j}(x,y) = \sum_{\alpha=1}^k \sum_{\beta=1}^k u_{i,j}(\hat{x}_\alpha, \hat{y}_\beta)$ $\ell_\alpha(x)\ell_\beta(y)$ for the convenience of later discussion.

We have the accuracy results for the conservative positivity-preserving limiter (4.33) as follows:

Lemma 4.3.1. *Consider the solution $u_{i,j}$ of the scheme (4.31) with accuracy $O(h^{k+1})$. If $\lambda = 0$, the error introduced by the limiter (4.33) is $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(K_{i,j})} = O(h^{k+1})$. If $\lambda > 0$, the error introduced by the limiter (4.33) is $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(K_{i,j})} = O(h^k)$, but the error is optimal on the downstream edges, i.e. $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(I_{i+\frac{1}{2}} \cup J_{j+\frac{1}{2}})} = O(h^{k+1})$, where $I_{i+\frac{1}{2}}$ and $J_{j+\frac{1}{2}}$ denote the right and upper edges of $K_{i,j}$, respectively.*

Proof. We decompose the error as

$$e = u_{i,j} - \tilde{u}_{i,j} = (u_{i,j} - \hat{u}_{i,j}) + (\hat{u}_{i,j} - \tilde{u}_{i,j}) = e_1 + e_2, \tag{4.34}$$

and

$$e_1 = u_{i,j} - \hat{u}_{i,j} = \left(u^o_{i,j} - \hat{u}^o_{i,j}\right) + \left(u^b_{i,j} - \hat{u}^b_{i,j}\right) + \left(\hat{u}^b_{i,j} - \theta^1_{i,j}\hat{u}^b_{i,j}\right) = e_{1,1} + e_{1,2} + e_{1,3}. \tag{4.35}$$

Using similar arguments as in (4.24), it is easy to prove that $||e_{1,1}||_{L^\infty(K_{i,j})} = O(h^{k+1})$ and $||e_{1,2}||_{L^\infty(K_{i,j})} = O(h^{k+1})$. As for $e_{1,3} = \left(1 - \theta^1_{i,j}\right)\hat{u}^b_{i,j}$, we consider two cases.

Case I: $\theta^1_{i,j} = 0$. We have $LHS^b(u^b_{i,j}) \leq 0$, i.e.

$$\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) u_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) u_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \leq 0,$$

thus

$$\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) u^+_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) u^+_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \leq$$
$$\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) u^-_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) u^-_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}),$$

which implies

$$\begin{aligned}
&\sum_{\beta=1}^{k+1} \hat{\omega}_\beta u^+_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha u^+_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \\
\leq &\frac{\rho \max\{a^*, b^*\}}{\min\{a_*, b_*\}} \left(\sum_{\beta=1}^{k+1} \hat{\omega}_\beta u^-_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha u^-_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \right) = O(h^{k+1}).
\end{aligned}$$

By the definition of $\hat{u}^b_{i,j}$, we have $||\hat{u}^b_{i,j}||_{L^\infty(K_{i,j})} = O(h^{k+1})$, therefore $||e_{1,3}||_{L^\infty(K_{i,j})} = O(h^{k+1})$.

Case II: $\theta^1_{i,j} > 0$. We have $LHS^b(u^b_{i,j}) > 0$ and $\theta^1_{i,j} = \frac{LHS^b(u^b_{i,j})}{LHS^b(\hat{u}^b_{i,j})}$. Therefore,

$$
\begin{aligned}
&|e_{1,3}| \\
=&(1-\theta^1_{i,j})|\hat{u}^b_{i,j}| \\
=&\frac{LHS^b(\hat{u}^b_{i,j}-u^b_{i,j})}{LHS^b(\hat{u}^b_{i,j})}|\hat{u}^b_{i,j}| \\
=&\left(\Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta a(x_{i+\frac{1}{2}},\hat{y}_\beta)\hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)+\Delta x_i\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha b(\hat{x}_\alpha,y_{j+\frac{1}{2}})\hat{u}^b_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})\right)^{-1}\times \\
&\left(\Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta a(x_{i+\frac{1}{2}},\hat{y}_\beta)\left(\hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)-u^b_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)\right)\right. \\
&\left.+\Delta x_i\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha b(\hat{x}_\alpha,y_{j+\frac{1}{2}})\left(\hat{u}^b_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})-u^b_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})\right)\right)\cdot|\hat{u}^b_{i,j}| \\
\le&\frac{\left(\Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta a(x_{i+\frac{1}{2}},\hat{y}_\beta)+\Delta x_i\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha b(\hat{x}_\alpha,y_{j+\frac{1}{2}})\right)||e_{1,2}||_{L^\infty(K_{i,j})}}{\Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta a(x_{i+\frac{1}{2}},\hat{y}_\beta)\hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)+\Delta x_i\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha b(\hat{x}_\alpha,y_{j+\frac{1}{2}})\hat{u}^b_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})}\cdot|\hat{u}^b_{i,j}|
\end{aligned}
\tag{4.36}
$$

In particular, $e_{1,3}(\hat{x}_{\gamma_1},\hat{y}_{\gamma_2})=0$ for $\gamma_1,\gamma_2=1,2,\ldots,k$, since $\hat{u}^b_{i,j}=0$ at these points by definition. Moreover, for $\gamma=1,2,\ldots,k+1$, we have the following estimates from (4.36),

$$
\begin{aligned}
&|e_{1,3}(x^-_{i+\frac{1}{2}},\hat{y}_\gamma)| \\
\le&\frac{\left(\Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta a(x_{i+\frac{1}{2}},\hat{y}_\beta)+\Delta x_i\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha b(\hat{x}_\alpha,y_{j+\frac{1}{2}})\right)||e_{1,2}||_{L^\infty(K_{i,j})}}{\Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta a(x_{i+\frac{1}{2}},\hat{y}_\beta)\hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)+\Delta x_i\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha b(\hat{x}_\alpha,y_{j+\frac{1}{2}})\hat{u}^b_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})}\hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\gamma) \\
\le&\frac{(\Delta y_j a^*+\Delta x_i b^*)\,||e_{1,2}||_{L^\infty(K_{i,j})}}{\Delta y_j a_*\sum_{\beta=1}^{k+1}\hat{\omega}_\beta\hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)}\hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\gamma) \\
\le&\frac{\rho\,(a^*+b^*)}{a_*\hat{\omega}_\gamma}||e_{1,2}||_{L^\infty(K_{i,j})}=O(h^{k+1}),
\end{aligned}
\tag{4.37}
$$

and similarly, $|e_{1,3}(\hat{x}_\gamma,y^-_{j+\frac{1}{2}})|=O(h^{k+1}),\gamma=1,2,\ldots,k+1$. Therefore, following the

similar argument as (4.27), we have $||e_{1,3}||_{L^\infty(K_{i,j})} = \Lambda_k^2 \cdot O(h^{k+1}) = O(h^{k+1})$.

To sum up, we have $||e_1||_{L^\infty(K_{i,j})} \leq ||e_{1,1}||_{L^\infty(K_{i,j})} + ||e_{1,2}||_{L^\infty(K_{i,j})} + ||e_{1,3}||_{L^\infty(K_{i,j})} = O(h^{k+1})$.

We now estimate e_2 as follows. If $\lambda = 0$, then $\theta_{i,j}^2 = 1$, thus $e_2 = (1-\theta_{i,j}^2)\hat{u}_{i,j} = 0$. If $\lambda > 0$, we have

$$
\begin{aligned}
&|e_2(x,y)| \\
&= \left(1 - \frac{LHS(u_{i,j})}{LHS(\hat{u}_{i,j})}\right) |\hat{u}_{i,j}(x,y)| = \frac{LHS(\hat{u}_{i,j} - u_{i,j})}{LHS(\hat{u}_{i,j})} |\hat{u}_{i,j}(x,y)| \\
&= \Bigg(\lambda \Delta x_i \Delta y_j \sum_{\alpha=1}^{k+1} \sum_{\beta=1}^{k+1} \hat{\omega}_\alpha \hat{\omega}_\beta \left(\hat{u}_{i,j}(\hat{x}_\alpha, \hat{y}_\beta) - u_{i,j}(\hat{x}_\alpha, \hat{y}_\beta) \right) \\
&\quad + \Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) \left(\hat{u}_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) - u_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) \right) \\
&\quad + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) \left(\hat{u}_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) - u_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \right) \Bigg) \times \\
&\quad \Bigg(\lambda \Delta x_i \Delta y_j \sum_{\alpha=1}^{k+1} \sum_{\beta=1}^{k+1} \hat{\omega}_\alpha \hat{\omega}_\beta \hat{u}_{i,j}(\hat{x}_\alpha, \hat{y}_\beta) + \Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) \hat{u}_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) \\
&\quad + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) \hat{u}_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \Bigg)^{-1} \cdot |\hat{u}_{i,j}(x,y)| \\
&\leq (\lambda \Delta x_i \Delta y_j + a^* \Delta y_j + b^* \Delta x_i) \cdot ||e_1||_{L^\infty(K_{i,j})} \cdot \Bigg(\lambda \Delta x_i \Delta y_j \sum_{\alpha=1}^{k+1} \sum_{\beta=1}^{k+1} \hat{\omega}_\alpha \hat{\omega}_\beta \hat{u}_{i,j}(\hat{x}_\alpha, \hat{y}_\beta) \\
&\quad + a_* \Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta \hat{u}_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + b_* \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha \hat{u}_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \Bigg)^{-1} \cdot |\hat{u}_{i,j}(x,y)|
\end{aligned}
\tag{4.38}
$$

In particular, at the Gauss-Radau points, we have the following estimates from

(4.38),

$$
\begin{aligned}
&|e_2(\hat{x}_{\gamma_1}, \hat{y}_{\gamma_2})| \\
\leq &\left(\lambda \Delta x_i \Delta y_j + a^* \Delta y_j + b^* \Delta x_i\right) ||e_1||_{L^\infty(K_{i,j})} \times \\
&\left(\lambda \Delta x_i \Delta y_j \sum_{\alpha=1}^{k+1} \sum_{\beta=1}^{k+1} \hat{\omega}_\alpha \hat{\omega}_\beta \hat{u}_{i,j}(\hat{x}_\alpha, \hat{y}_\beta)\right)^{-1} \hat{u}_{i,j}(\hat{x}_{\gamma_1}, \hat{y}_{\gamma_2}) \\
\leq &\frac{\lambda \rho^2 h^2 + a^* \rho h + b^* \rho h}{\lambda h^2 \hat{\omega}_{\gamma_1} \hat{\omega}_{\gamma_2}} ||e_1||_{L^\infty(K_{i,j})} \\
= &\left(\rho^2 \hat{\omega}_{\gamma_1}^{-1} \hat{\omega}_{\gamma_2}^{-1} + \frac{a^* \rho + b^* \rho}{\lambda \hat{\omega}_{\gamma_1} \hat{\omega}_{\gamma_2}} \frac{1}{h}\right) ||e_1||_{L^\infty(K_{i,j})} \\
= &O(h^k), \quad \text{for } \gamma_1, \gamma_2 = 1, 2, \ldots, k+1,
\end{aligned}
\tag{4.39}
$$

therefore, following the similar argument as (4.27), we have $||e_2||_{L^\infty(K_{i,j})} = \Lambda_k^2 \cdot O(h^k) = O(h^k)$.

In particular, on the downstream edge $I_{i+\frac{1}{2}}$, it follows from (4.38) that

$$
\begin{aligned}
&|e_2(x_{i+\frac{1}{2}}, \hat{y}_\gamma)| \\
\leq &\left(\lambda \Delta x_i \Delta y_j + a^* \Delta y_j + b^* \Delta x_i\right) ||e_1||_{L^\infty(K_{i,j})} \times \\
&\left(a_* \Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta \hat{u}_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta)\right)^{-1} \cdot \hat{u}_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\gamma) \\
\leq &\left(\lambda \Delta x_i \Delta y_j + a^* \Delta y_j + b^* \Delta x_i\right) \cdot ||e_1||_{L^\infty(K_{i,j})} \times \\
&\left(a_* \Delta y_j \hat{\omega}_\gamma \hat{u}_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\gamma)\right)^{-1} \cdot \hat{u}_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\gamma) \\
\leq &\frac{\lambda \rho^2 h + a^* \rho + b^* \rho}{a_* \hat{\omega}_\gamma} ||e_1||_{L^\infty(K_{i,j})} = O(h^{k+1}),
\end{aligned}
\tag{4.40}
$$

for $\gamma = 1, 2, \ldots, k+1$. Similarly, on the downstream edge $J_{j+\frac{1}{2}}$, we have $|e_2(\hat{x}_\gamma, y_{j+\frac{1}{2}})| = O(h^{k+1})$, for $\gamma = 1, 2, \ldots, k+1$. Following the same lines as in (4.27), we have the estimate $||e_2||_{L^\infty(I_{i+\frac{1}{2}} \cup J_{j+\frac{1}{2}})} = \Lambda_k \cdot O(h^{k+1}) = O(h^{k+1})$. Thus $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(I_{i+\frac{1}{2}} \cup J_{j+\frac{1}{2}})} = O(h^{k+1})$ by the triangle inequality.

Gathering all results above, we finish the proof of Lemma 4.3.1. □

The type-2 limiter is defined as follows,

$$\tilde{u}_{i,j}(x,y) = \theta_{i,j}\hat{u}_{i,j}(x,y), \quad \hat{u}_{i,j}(x,y) = u_{i,j}(x,y) + \epsilon_{i,j}, \tag{4.41}$$

where $\epsilon_{i,j} = -\min\{\min_{(x,y)\in S} u_{i,j}(x,y), 0\}$, $S \subset K_{i,j}$ is the set of points where we want to preserve the positivity of solutions, and $\theta_{i,j} = \frac{LHS(u_{i,j})}{LHS(\hat{u}_{i,j})} \in [0,1]$.

We have the accuracy result for the conservative positivity-preserving limiter as follows:

Lemma 4.3.2. *Consider the solution $u_{i,j}$ of the scheme* (4.31) *with accuracy $O(h^{k+1})$. If $\lambda = 0$, the error introduced by the limiter* (4.41) *is $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(K_{i,j})} = O(h)$. If $\lambda > 0$, the error introduced by the limiter* (4.41) *is $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(K_{i,j})} = O(h^k)$. Nevertheless, on the downstream edges, the errors in both cases are optimal, i.e. $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(I_{i+\frac{1}{2}} \cup J_{j+\frac{1}{2}})} = O(h^{k+1})$.*

Proof. For simplicity, we assume $S = K_{i,j}$. Same as (4.34), we decompose the error as $e = u_{i,j} - \tilde{u}_{i,j} = (u_{i,j} - \hat{u}_{i,j}) + (\hat{u}_{i,j} - \tilde{u}_{i,j}) = e_1 + e_2$. It is clear that $|e_1| = \epsilon_{i,j} = O(h^{k+1})$. Consider $e_2 = (1-\theta_{i,j})\,\hat{u}_{i,j}$. If $\lambda = 0$, we have $||e_2||_{L^\infty(K_{i,j})} = (1-\theta_{i,j})\,||\hat{u}_{i,j}||_{L^\infty(K_{i,j})} = O(h)$, since $||\hat{u}_{i,j}||_{L^\infty(K_{i,j})} = O(h)$ if $\theta_{i,j} < 1$. If $\lambda > 0$, we have the same estimates for e_2 as (4.38) and (4.39). The estimates for e_2 on the downstream edges are exactly the same as (4.40) for both the cases $\lambda = 0$ and $\lambda > 0$.

Collecting all results above, we finish the proof of Lemma 4.3.2. □

Since the accuracy of both type-1 and type-2 limiters is optimal on the downstream edges, we do not need to worry about the pollution of the non-optimal errors

introduced by the limiters to the downstream cells. Thus we have the following theorem for the positivity-preserving DG method of the equation (4.30).

Theorem 4.3.3. *For the linear stationary hyperbolic equation* (4.30)*, if the source term and inflow boundary conditions are nonnegative, then the solution of the scheme* (4.31) *modified by the limiter* (4.33) *or* (4.41) *is nonnegative, with the local accuracy established in Lemmas 4.3.1 and 4.3.2, respectively.*

Remark 4.3.1. *In particular, in the space-time DG discretization for the equation of the form* $u_t + (a(x)u)_x = s(x,t)$*, the accuracy of the solution at the terminal time is optimal, as the terminal time is indeed an outflow boundary.*

4.4 Linear stationary hyperbolic equations in two dimensions on triangular meshes

In this section, we study the high order conservative positivity-preserving discontinuous Galerkin method in two space dimensions on triangular meshes for the linear stationary hyperbolic equation (4.30) with nonnegative source term and the inflow boundary condition $u|_{\Gamma^{\text{in}}}(x,y) = g(x,y) \geq 0$, where $\Gamma^{\text{in}} \subset \partial\Omega$ is the inflow boundary. We still assume $\lambda \geq 0$ in (4.30) but $a(x,y)$ and $b(x,y)$ are not necessarily positive (or negative).

Consider a regular triangulation Ω_h of Ω which satisfies $\text{diam}(K) \leq \rho h, \forall K \in \Omega_h$ for some $\rho \geq 1$ independent of the refinement, where $\text{diam}(K)$ is the diameter of an element K, $h = \min_{K\in\Omega_h} h_K$ and h_K is the radius of the largest ball inscribed in K. For any triangle element $K \in \Omega_h$, we denote by $|K|$ the area of K, and $e_K^i, i = 1,2,3$ the three edges of K, with length ℓ_K^i, unit outer normal $n_K^i = (n_{x,K}^i, n_{y,K}^i)^T$

and neighboring cells K_i, $i = 1, 2, 3$. We assume that the coefficients $a(x, y)$ and $b(x, y)$ in (4.30) satisfy $c_* \leq \left|a(x, y)n^i_{x,K} + b(x, y)n^i_{y,K}\right| \leq c^*, \forall K \in \Omega_h, (x, y) \in \Omega, i = 1, 2, 3$, for some positive constants c_*, c^*. This assumption was adopted in the optimal order error estimate for the DG method in [64], as the optimal accuracy is unavailable for general meshes [54]. The assumption can be satisfied, for instance, by the conditions on the coefficients $a(x, y), b(x, y)$ in Section 4.3, together with the triangulation obtained by splitting each cell therein from the skew diagonal of cells, see Figure 4.1 for an illustration. The function space V of the P^k-DG scheme is defined as

$$V = \left\{v \in L^2(\Omega) : v|_K \in P^k(K), \forall K \in \Omega_h\right\},$$

where $P^k(K)$ denotes the space of polynomials of order no greater than k on the element K. We define the cell average of $v \in V$ on K as $\bar{v}_K = \frac{1}{|K|}\iint_K v(x, y)dxdy$, and denote by $v_K = v|_K$ for $v \in V$.

To save space, we only discuss the case that e^1_K is the upstream edge and e^2_K, e^3_K are the downstream edges, as the discussion of the case of two upstream edges and one downstream edge is almost the same with the first case.

The positivity-preserving P^k-DG scheme of the equation (4.30) on triangular meshes is to find $u \in V$, such that

$$\begin{aligned}
&-\iint_K \left(auv_x + buv_y - \lambda uv\right) dxdy \\
&+\int_{e^2_K} \left(an^2_{x,K} + bn^2_{y,K}\right) u_K v ds + \int_{e^3_K} \left(an^3_{x,K} + bn^3_{y,K}\right) u_K v ds \\
&= -\int_{e^1_K} \left(an^1_{x,K} + bn^1_{y,K}\right) \tilde{u}_{K_1} v ds + \iint_K svdxdy, \quad \forall v \in P^k(K),
\end{aligned} \tag{4.42}$$

for $K \in \Omega_h$, where we define $\tilde{u}_{K_1}|_{e^1_K} = \mathcal{I}(g)$ if $e^1_K \subset \Gamma^{\text{in}}$, with $\mathcal{I}$ denoting the polynomial interpolation at the quadrature points on cell interfaces. In the computation,

we solve u_K on cell K based on the modified solution on upstream cells. Once u_K is obtained, we employ the positivity-preserving limiter to obtain the modified solution $\tilde{u}_K$, and use it in the computation on the downstream cells.

If we take $v = 1$ on K in the scheme (4.42), the following equation satisfied by the local mass can be obtained

$$\begin{aligned} &\lambda|K|\bar{u}_K + \int_{e_K^2} \left(an_{x,K}^2 + bn_{y,K}^2\right) u_K ds + \int_{e_K^3} \left(an_{x,K}^3 + bn_{y,K}^3\right) u_K ds \\ &= -\int_{e_K^1} \left(an_{x,K}^1 + bn_{y,K}^1\right) \tilde{u}_{K_1} ds + |K|\bar{s}_K. \end{aligned} \tag{4.43}$$

We define $LHS(w_K) = \lambda|K|\bar{w}_K + \int_{e_K^2} \left(an_{x,K}^2 + bn_{y,K}^2\right) w_K ds + \int_{e_K^3} \left(an_{x,K}^3 + bn_{y,K}^3\right) w_K ds$, for $w_K \in P^k(K)$, to be the amount of local mass of w_K on K. Since $\tilde{u}_{K_1} \geq 0$ and $an_{x,K}^1 + bn_{y,K}^1 < 0$ on the upstream edge in (4.43), we have the $LHS(u_K) \geq 0$.

Due to the lack of suitable quadrature rules, we do not have the type-1 limiter available. The type-2 limiter is defined as follows,

$$\tilde{u}_K(x,y) = \theta_K \hat{u}_K(x,y), \quad \hat{u}_K(x,y) = u_K(x,y) + \epsilon_K, \tag{4.44}$$

where $\epsilon_K = -\min\{\min_{(x,y)\in S} u_K(x,y), 0\}$, $S \subset K$ is the set of points where we want to preserve the positivity of solutions, and $\theta_K = \frac{LHS(u_K)}{LHS(\hat{u}_K)} \in [0,1]$.

We have the accuracy result for the conservative positivity-preserving limiter as follows:

Lemma 4.4.1. *Consider the solution u_K of the scheme (4.42) with accuracy $O(h^{k+1})$. If $\lambda = 0$, the error introduced by the limiter (4.44) is $||\tilde{u}_K - u_K||_{L^\infty(K)} = O(h)$. If $\lambda > 0$, the error introduced by the limiter (4.44) is $||\tilde{u}_K - u_K||_{L^\infty(K)} = O(h^k)$. Nevertheless, on the downstream edges, the errors in both cases are optimal, i.e.*

$||\tilde{u}_K - u_K||_{L^\infty(e_K^2 \cup e_K^3)} = O(h^{k+1})$.

Proof. For simplicity, we assume $S = K$. We decompose the error as

$$e = u_K - \tilde{u}_K = (u_K - \hat{u}_K) + (\hat{u}_K - \tilde{u}_K) = e_1 + e_2 \tag{4.45}$$

It is clear that $||e_1||_{L^\infty(K)} = \epsilon_K = O(h^{k+1})$.

For e_2, we have $e_2 = \hat{u}_K - \tilde{u}_K = (1 - \theta_K)\hat{u}_K$. If $\lambda = 0$, we have $||e_2||_{L^\infty(K)} = (1 - \theta_K)||\hat{u}_K||_{L^\infty(K)} = O(h)$, since $||\hat{u}_K||_{L^\infty(K)} = O(h)$ if $\theta_K < 1$. If $\lambda > 0$, we have the estimate for e_2 as follows,

$$\begin{aligned}
&|e_2| \\
=&(1 - \theta_K)\hat{u}_K = \left(1 - \frac{LHS(u_K)}{LHS(\hat{u}_K)}\right)\hat{u}_K = \frac{LHS(\hat{u}_K - u_K)}{LHS(\hat{u}_K)}\hat{u}_K \\
=&\left(\lambda|K|\bar{\hat{u}}_K + \int_{e_K^2}\left(an_{x,K}^2 + bn_{y,K}^2\right)\hat{u}_K ds + \int_{e_K^3}\left(an_{x,K}^3 + bn_{y,K}^3\right)\hat{u}_K ds\right)^{-1} \times \\
&\left(\lambda|K|\left(\bar{\hat{u}}_K - \bar{u}_K\right) + \int_{e_K^2}\left(an_{x,K}^2 + bn_{y,K}^2\right)(\hat{u}_K - u_K)\, ds\right. \\
&\left. + \int_{e_K^3}\left(an_{x,K}^3 + bn_{y,K}^3\right)(\hat{u}_K - u_K)\, ds\right)\hat{u}_K \\
&\leq \frac{\left(\lambda|K| + c^*\ell_K^2 + c^*\ell_K^3\right)||e_1||_{L^\infty(K)}}{\lambda|K|\bar{\hat{u}}_K + c_*\ell_K^2\bar{\hat{u}}_{e_K^2} + c_*\ell_K^3\bar{\hat{u}}_{e_K^3}}\hat{u}_K,
\end{aligned} \tag{4.46}$$

where $\bar{v}_{e_K^i} = \frac{1}{\ell_K^i}\int_{e_K^i} v_K ds$, for $v \in V$, $i = 1, 2, 3$.

By the equivalence of norms in the finite-dimensional space $P^k(K)$ and the rescaling argument, we have $||v||_{L^\infty(K)} \leq \frac{C_k}{|K|}||v||_{L^1(K)}$ and $||v||_{L^\infty(e_K^i)} \leq \frac{C_k'}{\ell_K^i}||v||_{L^1(e_K^i)}$, $\forall v \in P^k(K), i = 1, 2, 3$, for some positive constants C_k and C_k' depending only on k.

Therefore,

$$\begin{aligned} ||e_2||_{L^\infty(K)} &\leq \frac{(\lambda|K| + c^*\ell_K^2 + c^*\ell_K^3)\,||e_1||_{L^\infty(K)}}{\lambda|K|}\frac{||\hat{u}_K||_{L^\infty(K)}}{\bar{\hat{u}}_K} \\ &\leq \frac{(\lambda|K| + 2c^*\rho h)\,||e_1||_{L^\infty(K)}}{\lambda|K|}C_k \\ &\leq C_k\left(1 + \frac{2c^*\rho}{\pi\lambda}\frac{1}{h}\right)||e_1||_{L^\infty(K)} \\ &= O(h^k), \end{aligned} \tag{4.47}$$

where we have used the fact that $\hat{u}_K \geq 0$. Moreover, we have

$$\begin{aligned} ||e_2||_{L^\infty(e_K^2)} &\leq \frac{(\lambda|K| + c^*\ell_K^2 + c^*\ell_K^3)\,||e_1||_{L^\infty(K)}}{c_*\ell_K^2}\frac{||\hat{u}_K||_{L^\infty(e_K^2)}}{\bar{\hat{u}}_{e_K^2}} \\ &\leq \frac{(\lambda\rho^2h^2 + 2c^*\rho h)\,||e_1||_{L^\infty(K)}}{c_*h}C_k' \\ &= C_k'\frac{(\lambda\rho^2 h + 2c^*\rho)}{c_*}||e_1||_{L^\infty(K)} \\ &= O(h^{k+1}), \end{aligned} \tag{4.48}$$

and, similarly, $||e_2||_{L^\infty(e_K^3)} = O(h^{k+1})$.

Gathering all results above and using triangle inequalities, we finish the proof of Lemma 4.4.1. □

Since the accuracy of the limiter (4.44) is optimal on the downstream edges, we have the following theorem for the positivity-preserving DG method of the equation (4.30).

Theorem 4.4.2. *For the linear stationary hyperbolic equation* (4.30)*, if the source term and inflow boundary condition are nonnegative, then the solution of the scheme* (4.42) *modified by the limiter* (4.44) *is nonnegative, with the local accuracy established in Lemma 4.4.1.*

4.5 Nonlinear stationary hyperbolic equations in one dimension

In this section, we study the high order conservative positivity-preserving discontinuous Galerkin method for the nonlinear stationary hyperbolic equation

$$f(u)_x + \lambda u = s(x), \quad x \in \Omega = (0,1), \tag{4.49}$$

where $0 \le f'(u) \le a^*, \forall u$, and $\lambda, s(x) \ge 0$. We assign the inflow boundary condition $u(0) = u_0 \ge 0$ for the equation. We would like to note that, the assumption on invariant sign of $f'(u)$ for all u is essential, otherwise the stationary hyperbolic equation may need boundary conditions from both sides for the problem to be well-posed, see [38, 63] for instance. This condition is also necessary for the limiters to be well-defined.

We adopt the partition for Ω and the function space V exactly the same as in Section 4.2, as well as the notations if not otherwise stated.

The positivity-preserving P^k-DG scheme of the equation (4.49) is to find $u \in V$, such that

$$\begin{aligned} &-\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} f(u)v_x dx + f(u^-_{j+\frac{1}{2}})v^-_{j+\frac{1}{2}} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \lambda uv dx \\ &= f(\tilde{u}^-_{j-\frac{1}{2}})v^+_{j-\frac{1}{2}} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} sv dx, \quad \forall v \in P^k(I_j), \end{aligned} \tag{4.50}$$

for $j = 1, 2, \ldots, N$, where we define $u^-_{\frac{1}{2}} = u_0$. Note that the upstream cells adopt the modified solution in the scheme.

If we take the test function $v = 1$ on I_j in the scheme (4.50), the following equation satisfied by the local mass is obtained,

$$\lambda \Delta x_j \bar{u}_j + f(u^-_{j+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \Delta x_j \bar{s}_j. \tag{4.51}$$

Same as the linear case, we define $LHS(w_j) = \lambda \Delta x_j \bar{w}_j + f(w^-_{j+\frac{1}{2}})$, for $w_j \in P^k(I_j)$, to be the amount of local mass of w_j on I_j. A notable difference is that, we no longer have $LHS(u_j) \geq 0$.

The type-1 limiter for u_j is defined as follows,

$$\tilde{u}_j(x) = \theta_j \hat{u}_j(x), \quad \hat{u}_j(x) = \sum_{\alpha=1}^{k+1} u_j^+(\hat{x}_\alpha)\ell_\alpha(x), \tag{4.52}$$

where $\theta_j \in [0,1]$ is taken such that the local mass is conservative, i.e. $LHS(\tilde{u}_j) = LHS(u_j)$. Same as before, the type-1 limiter must be used in cooperation with the Gauss-Radau quadrature.

If $\lambda > 0$, $\theta_j \in [0,1]$ is uniquely determined. To see this, we define $h(\theta) = LHS(\theta \hat{u}_j) - LHS(u_j)$. It is clear that $h(0) = f(0) - f(\tilde{u}^-_{j-\frac{1}{2}}) - \Delta x_j \bar{s}_j \leq 0$, $h(1) = \lambda \Delta x_j \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha u_j^-(\hat{x}_\alpha) + f((u^-_{j+\frac{1}{2}})^+) - f(u^-_{j+\frac{1}{2}}) \geq 0$, and $h'(\theta) > 0$ for $\theta \in [0,1]$. Therefore, the existence and uniqueness of θ_j is guaranteed by the mean value theorem and monotonicity of $h(\theta)$. If $\lambda = 0$, we always take $\theta_j = 1$, since $u^-_{j+\frac{1}{2}} \geq 0$, which implies $LHS(\hat{u}_j) = f(u^-_{j+\frac{1}{2}}) = LHS(u_j)$. Moreover, we have the accuracy result of the limiter as follows:

Lemma 4.5.1. *Consider the solution u_j of the scheme* (4.50) *with accuracy $O(\Delta x_j^{k+1})$. If $\lambda = 0$, the error introduced by the limiter* (4.52) *is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^{k+1})$. If $\lambda > 0$, the error introduced by the limiter* (4.52) *is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^k)$, but the error is optimal at the downstream point, i.e. $|\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}| = O(\Delta x_j^{k+1})$.*

Proof. From $\theta_j\lambda\Delta x_j\bar{\hat{u}}_j + f(\theta_j\hat{u}^-_{j+\frac{1}{2}}) = \lambda\Delta x_j\bar{u}_j + f(u^-_{j+\frac{1}{2}})$, we have the expression of θ_j as follows,

$$\theta_j = \frac{\lambda\Delta x_j\bar{u}_j + f'(c_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}}}{\lambda\Delta x_j\bar{\hat{u}}_j + f'(c_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}, \tag{4.53}$$

where $c_{j+\frac{1}{2}} \in [u^-_{j+\frac{1}{2}}, \hat{u}^-_{j+\frac{1}{2}}]$ satisfies the Lagrange mean value theorem $f(\theta_j\hat{u}^-_{j+\frac{1}{2}}) - f(u^-_{j+\frac{1}{2}}) = f'(c_{j+\frac{1}{2}})(\theta_j\hat{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}})$. Then the estimates are almost the same to those in the proof of Lemma 4.2.1, except that $a(x_{j+\frac{1}{2}})$ is replaced by $f'(c_{j+\frac{1}{2}})$. □

The type-2 limiter for u_j is defined as follows,

$$\tilde{u}_j(x) = \theta_j\hat{u}_j(x), \quad \hat{u}_j(x) = u_j(x) + \epsilon_j, \tag{4.54}$$

where $\epsilon_j = -\min\{\min_{x\in S} u_j(x), 0\}$, $S \subset I_j$ is the set of points where we want to preserve the positivity of solutions, and $\theta_j \in [0,1]$ is uniquely determined by $LHS(\tilde{u}_j) = LHS(u_j)$.

We have the accuracy result for the limiter as follows,

Lemma 4.5.2. *Consider the solution u_j of the scheme* (4.50) *with accuracy $O(\Delta x_j^{k+1})$. If $\lambda = 0$, the error introduced by the limiter* (4.54) *is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j)$. If $\lambda > 0$, the error introduced by the limiter* (4.54) *is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^k)$. Nevertheless, at the downstream point, the errors in both cases are optimal, i.e. $|\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}| = O(\Delta x_j^{k+1})$.*

Proof. We have the same expression of θ_j as in (4.53). Therefore the estimates are almost the same with those in the proof of Lemma 4.2.2, except that $a(x_{j+\frac{1}{2}})$ is replaced by $f'(c_{j+\frac{1}{2}})$. □

Since the accuracy of both type-1 and type-2 limiters is optimal at the down-

stream point of cells, the assumption on the optimal accuracy of the unmodulated DG solution is appropriate. Collecting the Lemma 4.5.1 and 4.5.2, we attain the following theorem for the positivity-preserving DG method for the equation (4.49).

Theorem 4.5.3. *For the nonlinear stationary hyperbolic equation* (4.49)*, if the source term and inflow boundary condition are nonnegative, then the solution of the scheme* (4.50) *modified by the limiters* (4.52) *or* (4.54) *is nonnegative, with the local accuracy established in Lemma 4.5.1 and 4.5.2, respectively.*

4.6 Numerical tests

In this section, we show the accuracy and effectiveness of the conservative positivity-preserving DG methods established in previous sections for stationary hyperbolic equations and time-dependent problems with implicit time discretization by ample numerical tests. Most of the tests are taken from [46, 86, 91]. For simplicity, the triangular meshes adopted in the two dimensional tests are obtained by splitting the rectangular grids by the skew diagonals of every cells, see Figure 4.1 for an illustration of a 6×6 mesh. To save space, we only present the results of the type-2 limiters, as those of the type-1 limiters are almost the same (even thought the type-1 limiter is formally more accurate than the type-2).

We would like to note that, though the sub-optimal error estimates of the limiters are sharp by artificial examples in 4.2.1 and 4.2.2, in numerical tests we have not observed any degeneracy of orders of accuracy.

Example 4.6.1. ***(Comparison of the conservation property for different limiters)***

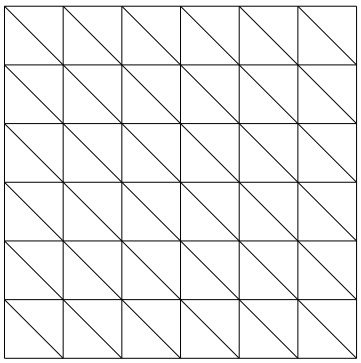


Figure 4.1: A typical triangular mesh in the tests

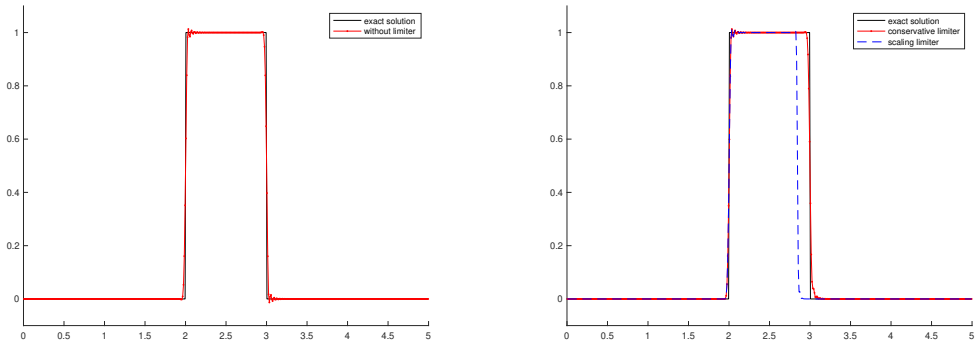
We solve the simple hyperbolic equation $u_t + u_x = 0$ *with implicit time discretization by a variety of positivity-preserving schemes, and compare the results of different positivity-preserving limiters. We first compute the solutions using the scaling limiter [93] that preserves cell averages. Then, we replace the scaling limiter in these algorithms by our conservative limiter that preserves the sum of cell average and outflow fluxes. Since the only difference is in the use of limiters, it would be convincing that our notion of conservation is more appropriate if the results of the conservative limiters are better than those of the scaling limiter.*

The initial and boundary condition are given below

$$u(x,0) = \begin{cases} 1, & 0 < x \leq 1 \\ 0, & otherwise \end{cases}, \quad x \in \Omega; \quad u(0,t) = 0, \quad t \in [0,T].$$

We first compute the equation on the domain $\Omega = [0,5]$*, with backward Euler time discretization, CFL number* $\frac{\Delta t}{\Delta x} = 0.01$ *and spatial partition* $N = 500$*, to the terminal time* $T = 2$*, based on the positivity-preserving* P^2*-DG scheme proposed in [86] for one dimensional linear equations. We plot the cell averages of the numerical*

solutions at the terminal time for the cases with no limiter, with the scaling limiter and with the conservative limiter, and compare them with the exact solution. The results are shown in Figure 4.2, from which we can clearly observe a wrong shock location with the use of the scaling limiter. The total mass of the exact solution at the terminal time is $\int_\Omega u(x,T)dx = 1$. *In the numerical solutions, the total mass has changed* 3.10×10^{-12}, 2.94×10^{-12} *and* 1.54×10^{-1} *for the cases with no limiter, with the conservative limiter and with the scaling limiter, respectively.*

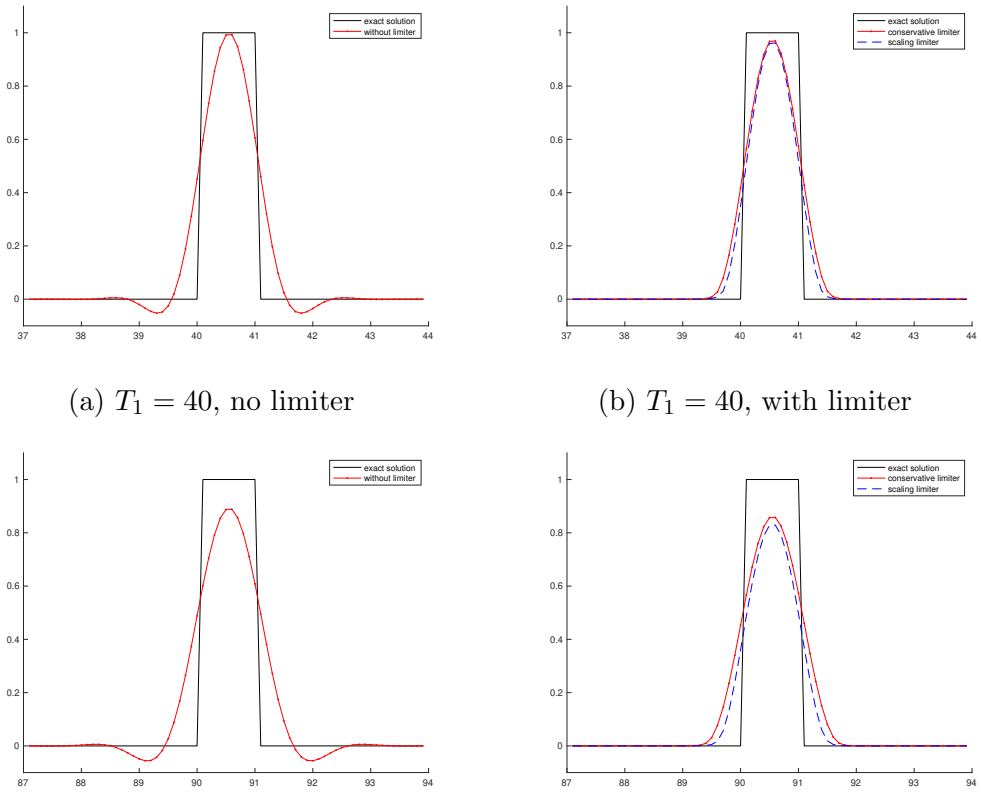


(a) No limiter (b) With limiter

Figure 4.2: Comparison of results for different limiters in the scheme of [86]

We then compute the equation on the space-time box $\Omega \times [0,T]$ *by the space-time DG discretization, based on the positivity-preserving* R^1*-DG scheme proposed in [46] for two dimensional linear equations. We take two space-time boxes* $\Omega_1 = [0,50], T_1 = 40$ *and* $\Omega_2 = [0,100], T_2 = 90$, *on the uniform meshes* $N_x^1 \times N_t^1 = 500 \times 400$ *and* $N_x^2 \times N_t^2 = 1000 \times 900$, *respectively. We plot the cell averages of the numerical solutions at the terminal times for the cases with no limiter, with the conservative limiter and with the scaling limiter, and compare them with the exact solution. The results are shown in the Figure 4.3, from which we can clearly observe the loss of mass with the use of the scaling limiter. In the numerical solutions, the total mass in the domain has changed* 4.97×10^{-14}, 4.77×10^{-14} *and* 8.59×10^{-2} *at* T_1 *for the cases with no limiter, with the conservative limiter and with the scaling limiter,*

respectively, and 1.01×10^{-13}, 1.11×10^{-13} *and* 1.47×10^{-1} *at* T_2 *for the cases with no limiter, with the conservative limiter and with the scaling limiter, respectively.*



(a) $T_1 = 40$, no limiter

(b) $T_1 = 40$, with limiter

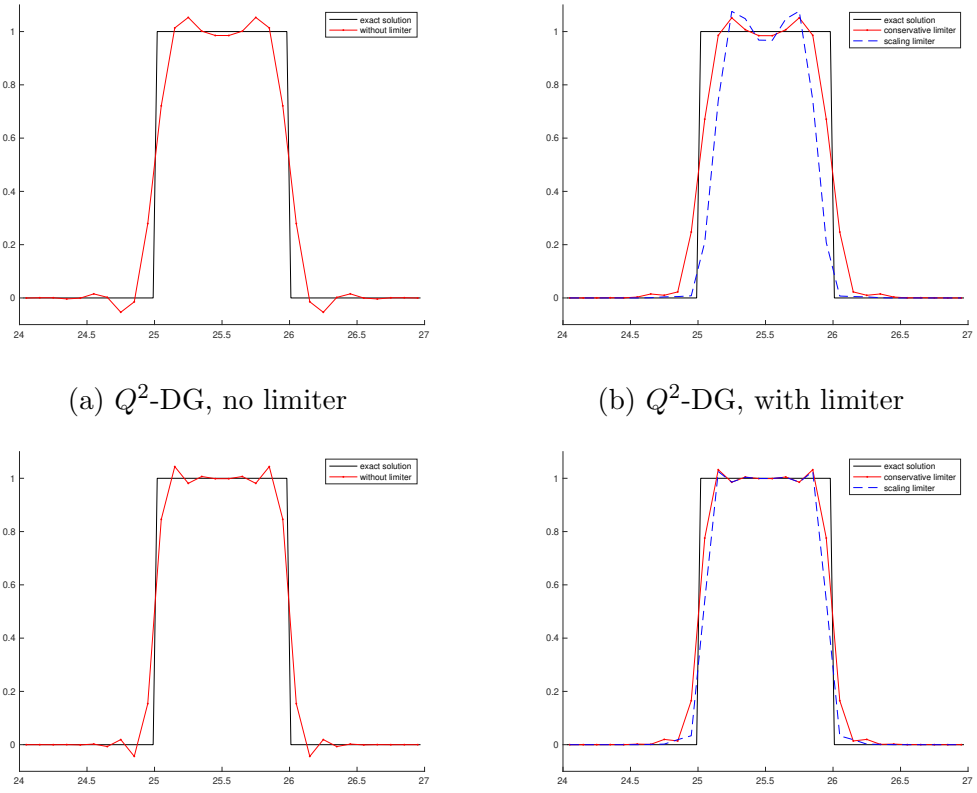
(c) $T_2 = 90$, no limiter

(d) $T_2 = 90$, with limiter

Figure 4.3: Comparison of results for different limiters in the scheme of [46]

Finally, we compute the equation by the space-time DG based on the Q^2 *and* Q^3 *schemes* (4.31). *Note that these two schemes are not positivity-preserving for cell averages in general, namely, there is no theoretical guarantee that the cell averages always remain nonnegative with the use of the scaling limiter. However, by trial and error, we find a setting that keeps the cell averages nonnegative during simulation, with the use of positivity-preserving scaling limiter. We take the space-time box* $\Omega = [0, 30], T = 25$ *on the uniform mesh* $N_x \times N_t = 300 \times 250$. *We plot the cell averages of the numerical solutions at the terminal time for the cases with no limiter, with the conservative limiter and with the scaling limiter, and compare them with the*

exact solution. The results are shown in Figure 4.4, from which we can clearly observe the loss of mass with the use of the scaling limiter. The total mass in the domain have changed 8.48×10^{-14}, 8.53×10^{-14} *and* 1.88×10^{-1} *in the* Q^2*-DG scheme for the cases with no limiter, with the conservative limiter and with the scaling limiter, respectively, and* 1.87×10^{-13}, 1.88×10^{-13} *and* 7.78×10^{-2} *in the* Q^3*-DG scheme for the cases with no limiter, with the conservative limiter and with the scaling limiter, respectively.*



(a) Q^2-DG, no limiter

(b) Q^2-DG, with limiter

(c) Q^3-DG, no limiter

(d) Q^3-DG, with limiter

Figure 4.4: Comparison of results for different limiters in the scheme (4.31)

Example 4.6.2. ***A linear stationary hyperbolic equation in one dimension***

We solve the equation (4.19) *with* $a(x) = 1$, $\lambda = 6000$ *and* $s(x) = \lambda\left(\frac{1}{9}\cos^4(x) + \epsilon\right) - \frac{4}{9}\cos^3(x)\sin(x)$ *on the domain* $\Omega = [0, \pi]$, *where* $\epsilon = 10^{-14}$ *is taken such that the source term is nonnegative. The boundary condition of the problem is* $u(0) = \frac{1}{9} + \epsilon$

and the exact solution is $u(x) = \frac{1}{9}\cos^4(x) + \epsilon$.

We compute the equation using the P^k*-DG scheme* (4.20) *with* $k = 1, 2, 3, 4$. *The errors, orders of convergence and data about positivity are given in Table 4.1, in which the category name LC (%) stands for the percentage of limited cells in the mesh. We can observe from the table that the negative values of the original scheme are eliminated by the limiter and the order of accuracy remains optimal.*

		no Limiter					with Limiter				
k	N	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	LC (%)
1	20	4.62E-04	-	9.00E-04	-	-4.67E-05	4.63E-04	-	9.00E-04	-	20.00
	40	1.16E-04	1.99	2.28E-04	1.98	-3.34E-06	1.16E-04	2.00	2.28E-04	1.98	10.00
	80	2.90E-05	2.00	5.83E-05	1.97	-2.19E-07	2.90E-05	2.00	5.83E-05	1.97	5.00
	160	7.26E-06	2.00	1.50E-05	1.96	-1.43E-08	7.26E-06	2.00	1.50E-05	1.96	2.50
	320	1.82E-06	2.00	3.90E-06	1.94	-9.54E-10	1.82E-06	2.00	3.90E-06	1.94	1.25
2	20	2.04E-05	-	3.88E-05	-	-2.30E-06	2.05E-05	-	3.88E-05	-	10.00
	40	2.54E-06	3.01	4.84E-06	3.00	-1.49E-07	2.54E-06	3.01	4.84E-06	3.00	5.00
	80	3.19E-07	2.99	5.98E-07	3.02	-9.58E-09	3.19E-07	2.99	5.98E-07	3.02	2.50
	160	4.01E-08	2.99	7.33E-08	3.03	-6.28E-10	4.01E-08	2.99	7.33E-08	3.03	1.25
	320	5.09E-09	2.98	8.86E-09	3.05	-4.25E-11	5.09E-09	2.98	8.86E-09	3.05	0.63
3	20	7.72E-07	-	1.57E-06	-	-9.43E-07	9.58E-07	-	4.24E-06	-	10.00
	40	4.79E-08	4.01	1.03E-07	3.93	-6.21E-08	5.38E-08	4.16	2.76E-07	3.94	5.00
	80	3.01E-09	3.99	6.74E-09	3.93	-4.05E-09	3.19E-09	4.07	1.73E-08	4.00	2.50
	160	1.89E-10	4.00	4.48E-10	3.91	-2.69E-10	1.94E-10	4.04	1.06E-09	4.02	1.25
	320	1.19E-11	3.98	3.06E-11	3.87	-1.83E-11	1.21E-11	4.00	6.46E-11	4.04	0.63
4	20	2.44E-08	-	4.80E-08	-	-1.12E-08	2.71E-08	-	4.80E-08	-	10.00
	40	7.60E-10	5.01	1.47E-09	5.03	-1.83E-10	7.81E-10	5.12	1.47E-09	5.03	5.00
	80	2.39E-11	4.99	4.45E-11	5.04	-3.02E-12	2.41E-11	5.02	4.45E-11	5.04	2.50
	160	7.57E-13	4.98	1.32E-12	5.07	-4.17E-14	7.58E-13	4.99	1.32E-12	5.07	1.25
	320	2.43E-14	4.96	3.97E-14	5.06	9.10E-15	2.43E-14	4.96	3.97E-14	5.06	0.00

Table 4.1: Results of Example 4.6.2

Example 4.6.3. ***A nonlinear stationary hyperbolic equation in one dimension***

We solve the equation (4.49) *with* $f(u) = u^3 + 0.01u$, $\lambda = 5$ *and* $s(x) = -8\sin(x)\cos^7(x)\left(3(\cos^8(x)+\epsilon)^2 + 0.01\right) + \lambda\left(\cos^8(x)+\epsilon\right)$ *on the domain* $\Omega = [0, \pi]$, *where* $\epsilon = 10^{-14}$ *is taken such that the source term is nonnegative. The boundary condition of the problem is* $u(0) = 1 + \epsilon$ *and the exact solution is* $u(x) = \cos^8(x) + \epsilon$.

We compute the equation using the P^k*-DG scheme* (4.50) *with* $k = 1, 2, 3, 4$. *The errors, orders of convergence and data about positivity are given in the Table*

4.2. Same to the linear case, we can observe that the negative values of the original scheme are eliminated by the limiter and the order of accuracy remains optimal.

		no Limiter					with Limiter				
k	N	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	LC (%)
1	20	7.45E-03	-	3.02E-02	-	-1.49E-04	7.45E-03	-	3.02E-02	-	30.00
	40	1.91E-03	1.96	7.97E-03	1.92	-8.48E-07	1.91E-03	1.96	7.97E-03	1.92	17.50
	80	4.91E-04	1.96	2.03E-03	1.97	-6.90E-09	4.91E-04	1.96	2.03E-03	1.97	8.75
	160	1.26E-04	1.96	5.11E-04	1.99	-5.30E-11	1.26E-04	1.96	5.11E-04	1.99	3.13
	320	3.22E-05	1.97	1.28E-04	2.00	-3.36E-13	3.22E-05	1.97	1.28E-04	2.00	0.63
2	20	4.57E-04	-	2.12E-03	-	-1.27E-06	4.57E-04	-	2.12E-03	-	20.00
	40	5.72E-05	3.00	2.62E-04	3.01	-1.18E-08	5.72E-05	3.00	2.62E-04	3.01	10.00
	80	7.22E-06	2.98	3.20E-05	3.03	-7.14E-11	7.22E-06	2.98	3.20E-05	3.03	5.00
	160	9.17E-07	2.98	3.96E-06	3.01	-2.48E-13	9.17E-07	2.98	3.96E-06	3.01	1.88
	320	1.16E-07	2.98	4.92E-07	3.01	9.44E-15	1.16E-07	2.98	4.92E-07	3.01	0.00
3	20	2.30E-05	-	1.19E-04	-	-9.63E-07	2.30E-05	-	1.19E-04	-	20.00
	40	1.44E-06	3.99	7.83E-06	3.93	-5.10E-09	1.44E-06	3.99	7.83E-06	3.93	10.00
	80	9.24E-08	3.97	4.95E-07	3.98	-2.42E-11	9.24E-08	3.97	4.95E-07	3.98	5.00
	160	5.86E-09	3.98	3.11E-08	3.99	-1.07E-13	5.86E-09	3.98	3.11E-08	3.99	0.63
	320	3.71E-10	3.98	1.94E-09	4.00	9.43E-15	3.71E-10	3.98	1.94E-09	4.00	0.00
4	20	1.05E-06	-	6.46E-06	-	-6.21E-09	1.05E-06	-	6.46E-06	-	10.00
	40	3.31E-08	4.98	1.88E-07	5.10	-3.05E-11	3.31E-08	4.98	1.88E-07	5.10	5.00
	80	1.05E-09	4.97	5.80E-09	5.02	-8.17E-14	1.05E-09	4.97	5.80E-09	5.02	2.50
	160	3.35E-11	4.98	1.79E-10	5.02	9.65E-15	3.35E-11	4.98	1.79E-10	5.02	0.00
	320	1.06E-12	4.98	5.54E-12	5.01	1.00E-14	1.06E-12	4.98	5.54E-12	5.01	0.00

Table 4.2: Results of Example 4.6.3

Example 4.6.4. ***A nonlinear time-dependent hyperbolic equation in one dimension with backward Euler time discretization***

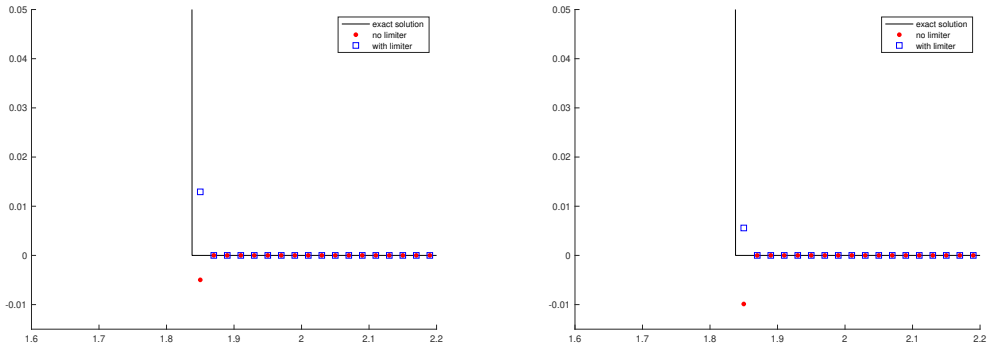
We solve the equation (4.1) *with backward Euler time discretization, and take* $f(u) = \frac{u^3}{3}$, $s(x) = 0$. *The initial and boundary condition of the equation are given below*

$$u(x,0) = \begin{cases} 1, & x \leq 1 \\ 0, & otherwise \end{cases}, \quad x \in \Omega; \quad u(0,t) = 1, \quad t \in [0,T],$$

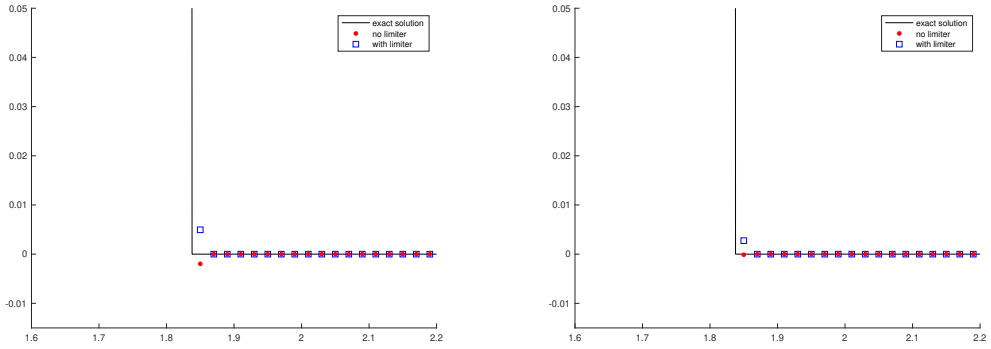
where $\Omega = [0,3]$ *and* $T = 2.5$.

We compute the equation using the P^k*-DG scheme* (4.50) *with* $k = 1,2,3,4$, *CFL number* $\frac{\Delta t}{\Delta x} = 0.5$ *and spatial partition* $N = 150$. *We zoom in the pre-shock zone and draw the cell averages of the numerical solutions in this area in Figure 4.5, with a comparison with the exact solution and the results without limiter. From the figures,*

we can observe that the negative cell averages of the original numerical scheme are eliminated by the limiter.



(a) P^1-DG (b) P^2-DG



(c) P^3-DG (d) P^4-DG

Figure 4.5: Results of Example 4.6.4

Example 4.6.5. ***Linear stationary hyperbolic equations in two dimensions with smooth solutions***

We solve the equation (4.30) *with constant coefficients* $a(x,y) \equiv a = 0.7, b(x,y) \equiv b = 0.3$ *and* $s(x,y) \equiv 0$ *in the domain* $\Omega = [0,1] \times [0,1]$. *The inflow boundary conditions are* $u(x,0) = 0, 0 < x \leq 1$ *and* $u(0,y) = \sin^6(\pi y), 0 \leq y \leq 1$. *The exact solution of the problem is*

$$u(x,y) = \begin{cases} 0, & y < \frac{b}{a}x \\ \sin^6(\pi(y - \frac{b}{a}x))e^{-\frac{\lambda}{a}x} & y \geq \frac{b}{a}x. \end{cases}$$

We take $\lambda = 1$, which corresponds to the purely absorbing medium in RTE, and $\lambda = 0$, which corresponds to the transparent medium in RTE, in the tests.

We compute the equations using the Q^k-DG scheme (4.31) *on rectangular meshes, and the P^k-DG scheme* (4.42) *on triangular meshes. The errors, orders of convergence and data about positivity are given in the Table 4.3 - Table 4.6, from which we can observe that the positivity and optimal accuracy are both attained by the algorithms.*

		no Limiter					with Limiter				
k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	LC (%)
1	20×20	1.43E-03	-	2.66E-02	-	-3.07E-04	1.42E-03	-	2.66E-02	-	35.00
	40×40	3.38E-04	2.08	6.94E-03	1.94	-2.06E-05	3.38E-04	2.07	6.94E-03	1.94	29.25
	80×80	8.21E-05	2.04	1.76E-03	1.98	-1.03E-06	8.21E-05	2.04	1.76E-03	1.98	25.98
	160×160	2.03E-05	2.02	4.42E-04	1.99	-4.61E-08	2.03E-05	2.02	4.42E-04	1.99	24.05
	320×320	5.04E-06	2.01	1.11E-04	2.00	-1.99E-09	5.04E-06	2.01	1.11E-04	2.00	22.95
2	20×20	6.32E-05	-	1.26E-03	-	-1.59E-06	6.32E-05	-	1.26E-03	-	14.50
	40×40	7.79E-06	3.02	1.63E-04	2.96	-2.73E-08	7.79E-06	3.02	1.63E-04	2.96	13.13
	80×80	9.71E-07	3.01	2.07E-05	2.98	-4.36E-10	9.71E-07	3.01	2.07E-05	2.98	12.58
	160×160	1.21E-07	3.00	2.60E-06	2.99	-6.86E-12	1.21E-07	3.00	2.60E-06	2.99	12.23
	320×320	1.51E-08	3.00	3.26E-07	3.00	-1.15E-13	1.51E-08	3.00	3.26E-07	3.00	12.06
3	20×20	2.77E-06	-	6.61E-05	-	-3.64E-07	2.78E-06	-	6.61E-05	-	18.50
	40×40	1.72E-07	4.01	4.35E-06	3.92	-7.14E-09	1.72E-07	4.01	4.35E-06	3.92	16.56
	80×80	1.07E-08	4.00	2.76E-07	3.98	-1.21E-10	1.07E-08	4.00	2.76E-07	3.98	15.55
	160×160	6.71E-10	4.00	1.73E-08	3.99	-1.94E-12	6.71E-10	4.00	1.73E-08	3.99	15.20
	320×320	4.19E-11	4.00	1.08E-09	4.00	-3.06E-14	4.19E-11	4.00	1.08E-09	4.00	14.98
4	20×20	1.08E-07	-	2.56E-06	-	-2.03E-08	1.09E-07	-	2.56E-06	-	19.50
	40×40	3.37E-09	5.01	8.33E-08	4.94	-3.45E-10	3.38E-09	5.01	8.33E-08	4.94	17.94
	80×80	1.05E-10	5.00	2.62E-09	4.99	-5.55E-12	1.05E-10	5.00	2.62E-09	4.99	17.31
	160×160	3.29E-12	5.00	8.19E-11	5.00	-9.54E-14	3.29E-12	5.00	8.19E-11	5.00	16.75
	320×320	1.07E-13	4.94	2.57E-12	5.00	-2.15E-15	1.07E-13	4.94	2.57E-12	5.00	16.53

Table 4.3: Results of Example 4.6.5 on rectangular meshes with $\lambda = 1$

Example 4.6.6. ***Linear stationary hyperbolic equations in two dimensions with discontinuous solutions***

We solve the equation (4.30) *with constant coefficients $a(x, y) \equiv a = 0.7, b(x, y) \equiv b = 0.3, s(x, y) \equiv 0$ in the domain $\Omega = [0, 1] \times [0, 1]$. The inflow boundary conditions are $u(x, 0) = 0, 0 < x \leq 1$ and $u(0, y) = 1, 0 \leq y \leq 1$. The exact solution of the problem is*

$$u(x, y) = \begin{cases} 0, & y < \frac{b}{a}x \\ e^{-\frac{\lambda}{a}x} & y \geq \frac{b}{a}x. \end{cases}$$

		no Limiter					with Limiter				
k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	LC (%)
1	20×20	1.92E-03	-	2.37E-02	-	-2.89E-03	1.89E-03	-	2.37E-02	-	22.13
	40×40	4.40E-04	2.13	6.85E-03	1.79	-1.32E-04	4.40E-04	2.10	6.85E-03	1.79	17.22
	80×80	1.05E-04	2.07	1.81E-03	1.92	-5.02E-06	1.05E-04	2.07	1.81E-03	1.92	14.23
	160×160	2.57E-05	2.03	4.67E-04	1.96	-1.88E-07	2.57E-05	2.03	4.67E-04	1.96	12.62
	320×320	6.38E-06	2.01	1.18E-04	1.98	-7.32E-09	6.38E-06	2.01	1.18E-04	1.98	11.77
2	20×20	9.43E-05	-	1.92E-03	-	-2.44E-05	9.47E-05	-	1.92E-03	-	10.38
	40×40	1.14E-05	3.05	2.62E-04	2.87	-4.59E-07	1.14E-05	3.05	2.62E-04	2.87	8.28
	80×80	1.42E-06	3.01	3.45E-05	2.92	-7.72E-09	1.42E-06	3.01	3.45E-05	2.92	7.74
	160×160	1.77E-07	3.00	4.39E-06	2.97	-1.23E-10	1.77E-07	3.00	4.39E-06	2.97	7.42
	320×320	2.21E-08	3.00	5.54E-07	2.99	-1.95E-12	2.21E-08	3.00	5.54E-07	2.99	7.34
3	20×20	5.87E-06	-	1.04E-04	-	-1.03E-05	6.77E-06	-	1.04E-04	-	14.75
	40×40	3.70E-07	3.99	7.30E-06	3.83	-2.32E-07	3.78E-07	4.16	7.30E-06	3.83	11.94
	80×80	2.31E-08	4.00	4.85E-07	3.91	-4.05E-09	2.32E-08	4.03	4.85E-07	3.91	10.50
	160×160	1.44E-09	4.00	3.13E-08	3.96	-6.52E-11	1.44E-09	4.00	3.13E-08	3.96	9.86
	320×320	9.02E-11	4.00	1.98E-09	3.98	-1.03E-12	9.02E-11	4.00	1.98E-09	3.98	9.43
4	20×20	3.16E-07	-	5.85E-06	-	-5.81E-07	3.32E-07	-	5.85E-06	-	9.75
	40×40	9.83E-09	5.01	2.05E-07	4.83	-1.08E-08	9.97E-09	5.06	2.05E-07	4.83	8.75
	80×80	3.03E-10	5.02	6.72E-09	4.93	-1.89E-10	3.04E-10	5.03	6.72E-09	4.93	8.59
	160×160	9.41E-12	5.01	2.21E-10	4.93	-3.08E-12	9.41E-12	5.01	2.21E-10	4.93	8.36
	320×320	3.06E-13	4.94	7.25E-12	4.93	-4.90E-14	3.06E-13	4.95	7.25E-12	4.93	8.31

Table 4.4: Results of Example 4.6.5 on triangular meshes with $\lambda = 1$

		no Limiter					with Limiter				
k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	LC (%)
1	20×20	2.66E-03	-	2.81E-02	-	-1.12E-03	2.64E-03	-	2.81E-02	-	35.00
	40×40	6.10E-04	2.12	7.23E-03	1.96	-8.44E-05	6.10E-04	2.11	7.23E-03	1.96	29.25
	80×80	1.46E-04	2.06	1.82E-03	1.99	-4.12E-06	1.46E-04	2.06	1.82E-03	1.99	25.97
	160×160	3.59E-05	2.03	4.56E-04	2.00	-1.87E-07	3.59E-05	2.03	4.56E-04	2.00	24.05
	320×320	8.89E-06	2.01	1.14E-04	2.00	-8.21E-09	8.89E-06	2.01	1.14E-04	2.00	22.95
2	20×20	1.14E-04	-	1.32E-03	-	-1.65E-06	1.14E-04	-	1.32E-03	-	14.50
	40×40	1.40E-05	3.03	1.67E-04	2.98	-2.78E-08	1.40E-05	3.03	1.67E-04	2.98	13.19
	80×80	1.74E-06	3.01	2.09E-05	3.00	-4.39E-10	1.74E-06	3.01	2.09E-05	3.00	12.56
	160×160	2.17E-07	3.00	2.61E-06	3.00	-1.32E-11	2.17E-07	3.00	2.61E-06	3.00	12.21
	320×320	2.71E-08	3.00	3.27E-07	3.00	-4.61E-13	2.71E-08	3.00	3.27E-07	3.00	12.07
3	20×20	4.96E-06	-	7.10E-05	-	-4.14E-07	4.97E-06	-	7.10E-05	-	19.25
	40×40	3.08E-07	4.01	4.48E-06	3.98	-7.61E-09	3.08E-07	4.01	4.48E-06	3.98	16.69
	80×80	1.92E-08	4.00	2.81E-07	4.00	-1.24E-10	1.92E-08	4.00	2.81E-07	4.00	15.53
	160×160	1.20E-09	4.00	1.76E-08	4.00	-3.10E-12	1.20E-09	4.00	1.76E-08	4.00	15.21
	320×320	7.50E-11	4.00	1.10E-09	4.00	-8.31E-14	7.50E-11	4.00	1.10E-09	4.00	15.01
4	20×20	1.94E-07	-	2.68E-06	-	-2.70E-08	1.95E-07	-	2.68E-06	-	19.25
	40×40	6.04E-09	5.01	8.42E-08	4.99	-5.42E-10	6.05E-09	5.01	8.42E-08	4.99	17.88
	80×80	1.88E-10	5.00	2.63E-09	5.00	-1.18E-11	1.88E-10	5.00	2.63E-09	5.00	17.30
	160×160	5.88E-12	5.00	8.24E-11	5.00	-2.73E-13	5.88E-12	5.00	8.24E-11	5.00	16.74
	320×320	1.92E-13	4.93	2.78E-12	4.89	-6.56E-15	1.92E-13	4.93	2.80E-12	4.88	16.48

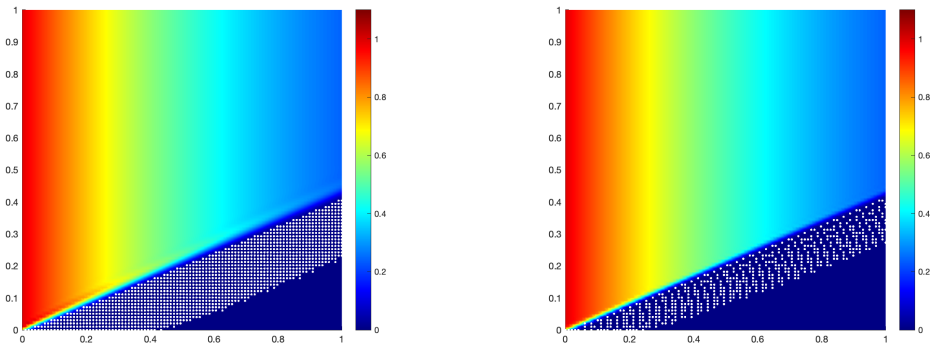
Table 4.5: Results of Example 4.6.5 on rectangular meshes with $\lambda = 0$

		no Limiter					with Limiter				
k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	LC (%)
1	20×20	3.67E-03	-	2.70E-02	-	-5.59E-03	3.58E-03	-	2.70E-02	-	22.38
	40×40	8.02E-04	2.19	7.41E-03	1.87	-4.09E-04	8.02E-04	2.16	7.41E-03	1.87	17.22
	80×80	1.86E-04	2.11	1.91E-03	1.96	-1.85E-05	1.87E-04	2.10	1.91E-03	1.96	14.21
	160×160	4.52E-05	2.04	4.80E-04	1.99	-7.38E-07	4.52E-05	2.04	4.80E-04	1.99	12.61
	320×320	1.12E-05	2.02	1.20E-04	2.00	-2.94E-08	1.12E-05	2.02	1.20E-04	2.00	11.77
2	20×20	1.68E-04	-	2.05E-03	-	-2.81E-05	1.69E-04	-	2.05E-03	-	11.63
	40×40	2.02E-05	3.05	2.64E-04	2.96	-4.80E-07	2.02E-05	3.06	2.64E-04	2.96	9.25
	80×80	2.50E-06	3.02	3.32E-05	2.99	-7.89E-09	2.50E-06	3.02	3.32E-05	2.99	8.54
	160×160	3.12E-07	3.00	4.16E-06	3.00	-1.25E-10	3.12E-07	3.00	4.16E-06	3.00	8.27
	320×320	3.89E-08	3.00	5.21E-07	3.00	-1.96E-12	3.89E-08	3.00	5.21E-07	3.00	8.18
3	20×20	1.02E-05	-	1.14E-04	-	-1.22E-05	1.19E-05	-	1.14E-04	-	15.38
	40×40	6.43E-07	3.99	7.90E-06	3.85	-2.52E-07	6.58E-07	4.17	7.90E-06	3.85	12.50
	80×80	4.02E-08	4.00	5.10E-07	3.95	-4.20E-09	4.03E-08	4.03	5.10E-07	3.95	11.00
	160×160	2.51E-09	4.00	3.21E-08	3.99	-6.67E-11	2.51E-09	4.00	3.21E-08	3.99	10.34
	320×320	1.57E-10	4.00	2.01E-09	4.00	-1.05E-12	1.57E-10	4.00	2.01E-09	4.00	9.87
4	20×20	5.55E-07	-	6.22E-06	-	-6.07E-07	5.85E-07	-	6.22E-06	-	10.13
	40×40	1.71E-08	5.02	2.16E-07	4.85	-1.19E-08	1.73E-08	5.08	2.16E-07	4.85	9.34
	80×80	5.26E-10	5.02	6.96E-09	4.96	-1.99E-10	5.28E-10	5.04	6.96E-09	4.96	8.84
	160×160	1.63E-11	5.01	2.20E-10	4.98	-3.16E-12	1.63E-11	5.01	2.20E-10	4.98	8.48
	320×320	5.36E-13	4.93	7.06E-12	4.96	-4.96E-14	5.36E-13	4.93	7.06E-12	4.96	8.32

Table 4.6: Results of Example 4.6.5 on triangular meshes with $\lambda = 0$

We test the cases of $\lambda = 1$ and $\lambda = 0$, which correspond to the purely absorbing medium and transparent medium in RTE, respectively.

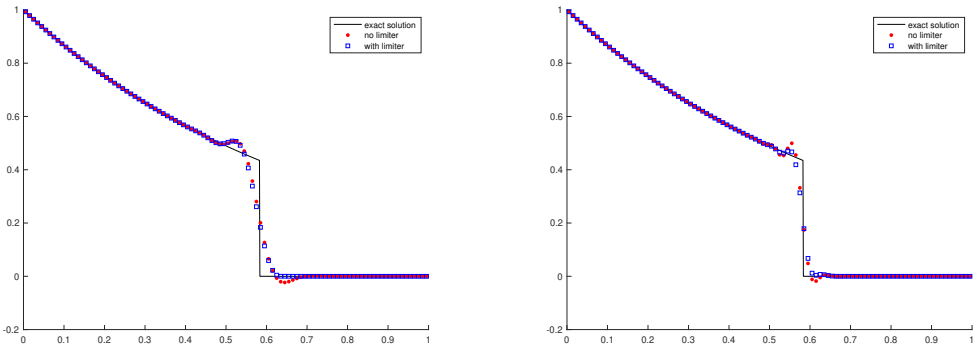
The solutions are computed by the Q^k-DG scheme (4.31) *on rectangular meshes, and by the P^k-DG scheme* (4.42) *on triangular meshes, with the spatial partition $N_x \times N_y = 100 \times 100$. We draw the contours of the solutions on rectangular meshes in Figures 4.6 and 4.8 for the cases $\lambda = 1$ and $\lambda = 0$, respectively. The contours of the solutions on triangular meshes are given in Figures 4.10 and 4.12 for the cases $\lambda = 1$ and $\lambda = 0$, respectively. Moreover, we slice the solutions along $y = 0.25$ and plot the averages of the solution along the line in Figures 4.7, 4.9, 4.11 and 4.13. From the figures, we can observe that the negative averages of the solution in the original scheme are eliminated by the positivity-preserving technique.*



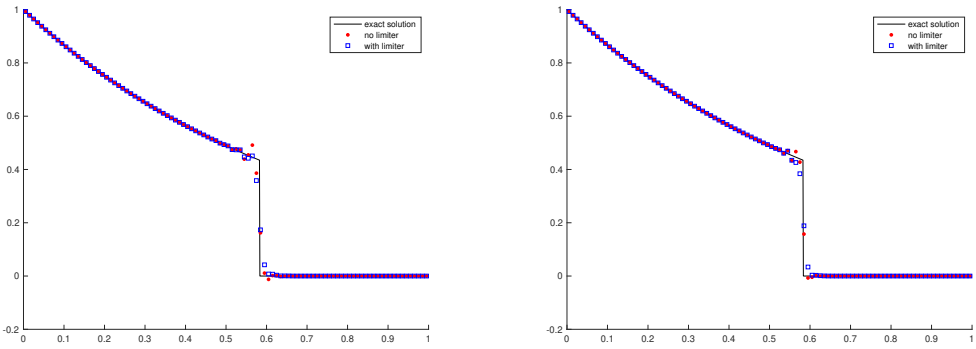
(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

Figure 4.6: Solutions of Example 4.6.6 on rectangular meshes with $\lambda = 1$

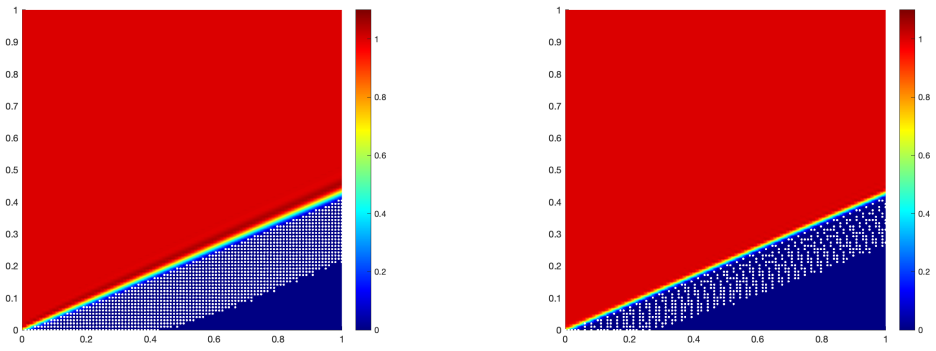


(a) $k = 1$ (b) $k = 2$

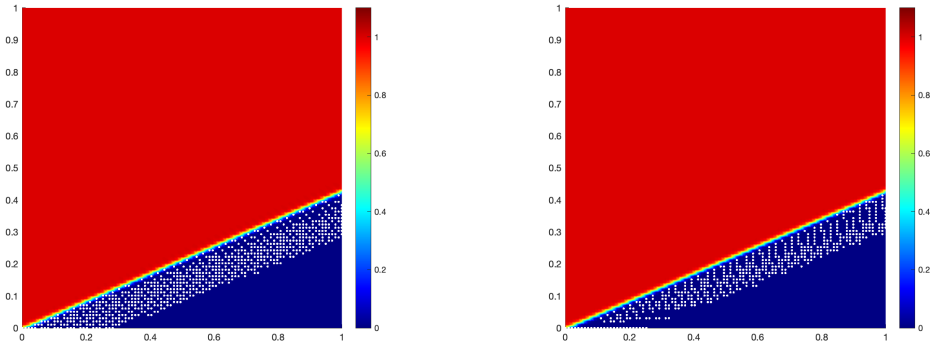


(c) $k = 3$ (d) $k = 4$

Figure 4.7: Solutions of Example 4.6.6 on rectangular meshes with $\lambda = 1$, cut along $y = 0.25$

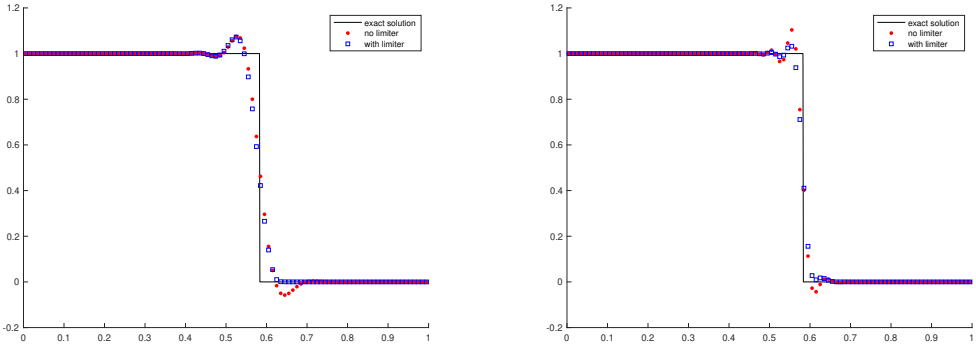


(a) $k = 1$ (b) $k = 2$

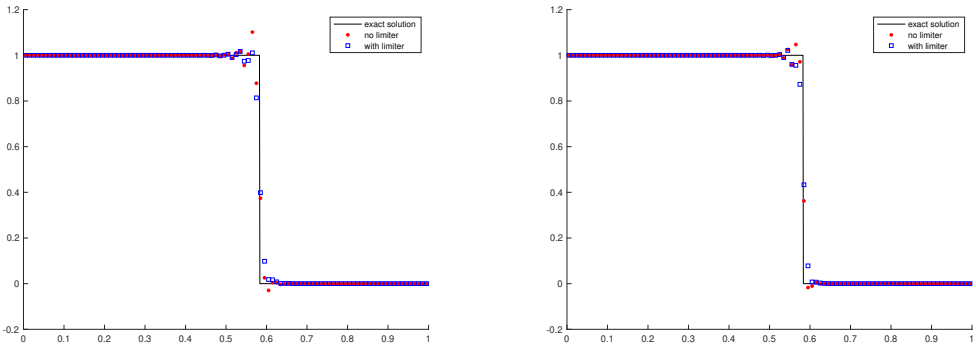


(c) $k = 3$ (d) $k = 4$

Figure 4.8: Solutions of Example 4.6.6 on rectangular meshes with $\lambda = 0$

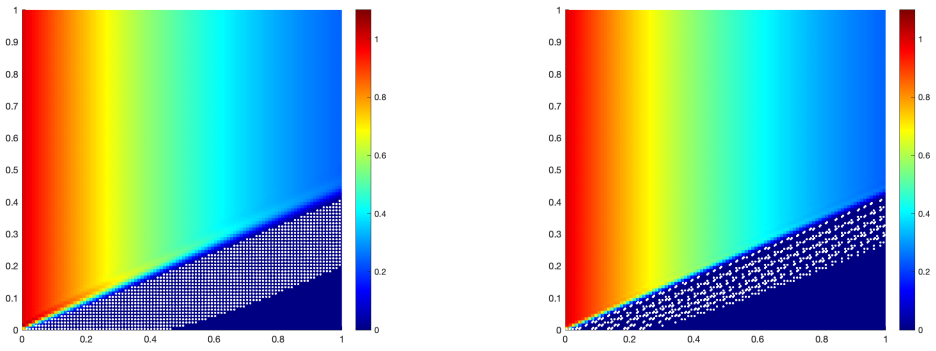


(a) $k = 1$ (b) $k = 2$

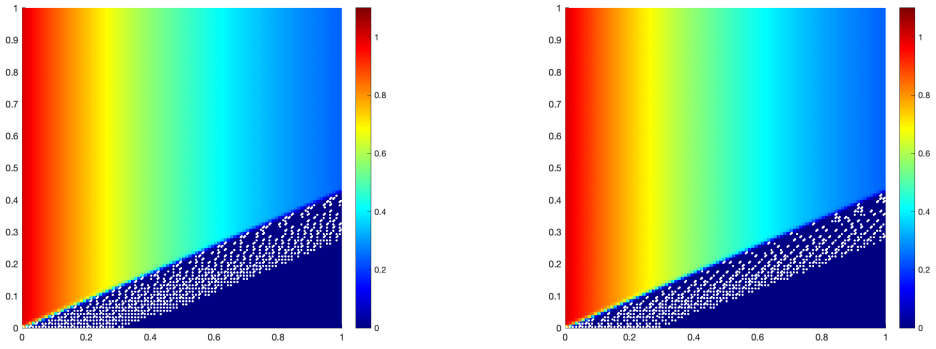


(c) $k = 3$ (d) $k = 4$

Figure 4.9: Solutions of Example 4.6.6 on rectangular meshes with $\lambda = 0$, cut along $y = 0.25$

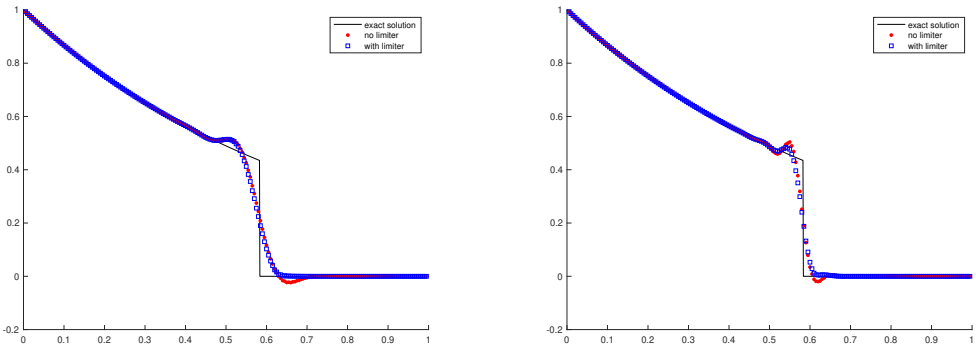


(a) $k = 1$ (b) $k = 2$

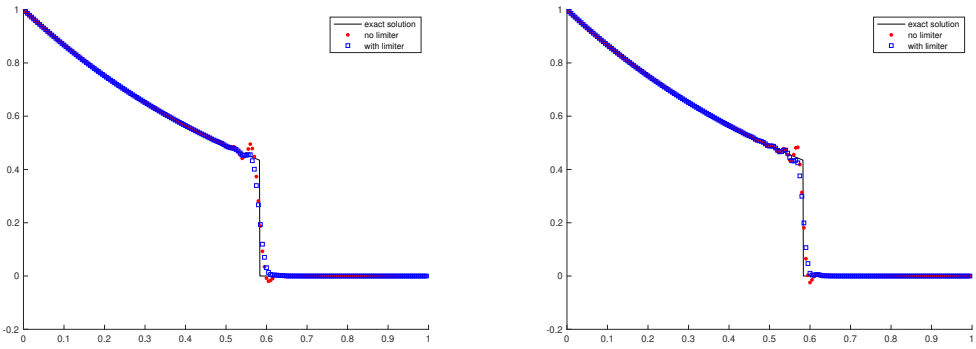


(c) $k = 3$ (d) $k = 4$

Figure 4.10: Solutions of Example 4.6.6 on triangular meshes with $\lambda = 1$

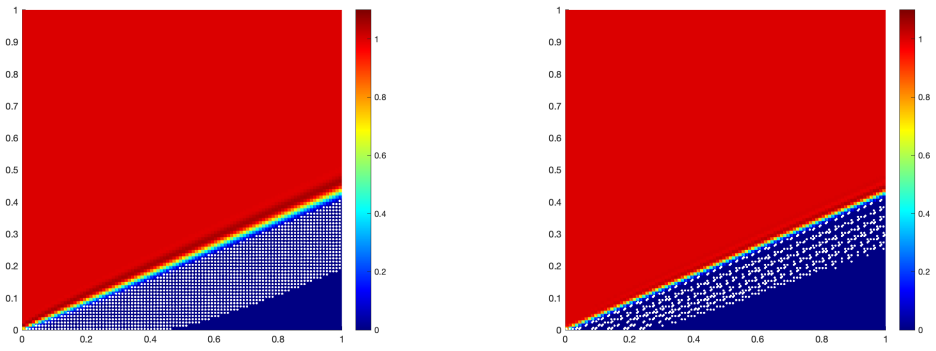


(a) $k = 1$ (b) $k = 2$

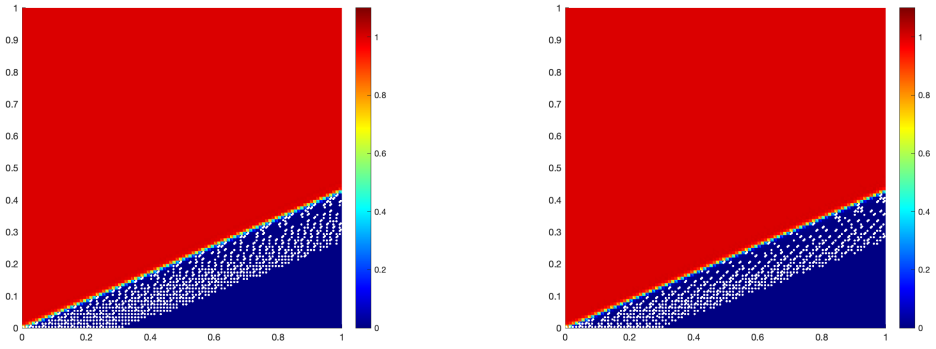


(c) $k = 3$ (d) $k = 4$

Figure 4.11: Solutions of Example 4.6.6 on triangular meshes with $\lambda = 1$, cut along $y = 0.25$

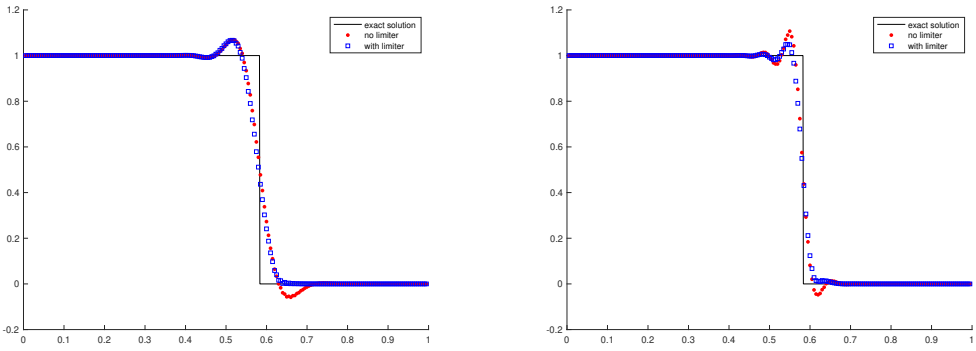


(a) $k = 1$ (b) $k = 2$

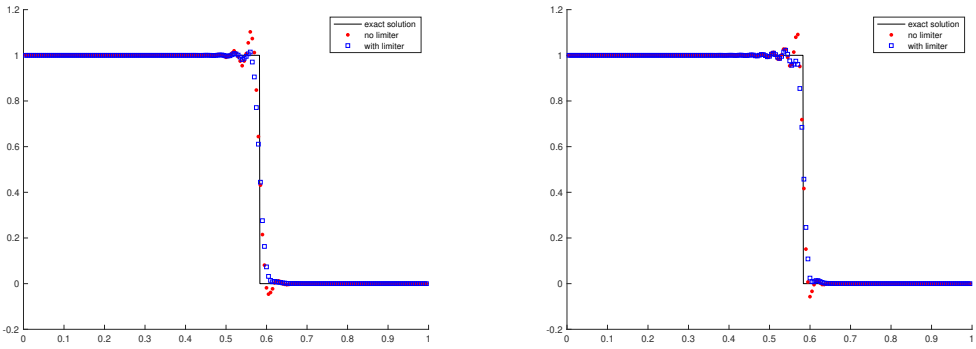


(c) $k = 3$ (d) $k = 4$

Figure 4.12: Solutions of Example 4.6.6 on triangular meshes with $\lambda = 0$



(a) $k = 1$ (b) $k = 2$



(c) $k = 3$ (d) $k = 4$

Figure 4.13: Solutions of Example 4.6.6 on triangular meshes with $\lambda = 0$, cut along $y = 0.25$

4.7 Concluding remarks

In this chapter, we have constructed high order conservative positivity-preserving discontinuous Galerkin methods for various stationary hyperbolic equations in one and two space dimensions, based on a novel definition of conservation for stationary equations. Two types of conservative positivity-preserving limiters are introduced, where the type-1 limiter relies on particular Gauss-Radau quadratures for the schemes while the type-2 limiter does not. The errors introduced by the limiters are of optimal order on downstream edges, thus the limiter does not pollute the original high order accuracy on downstream cells. Moreover, for time-dependent hyperbolic problems with implicit time discretization, the errors introduced by limiters are always optimal.

The positivity-preserving technique proposed in this chapter is easy to implement, simple to prove for the positivity, and applicable for general types of stationary hyperbolic equations, compared with the previous work.

Chapter Five

Local characteristic decomposition free finite difference WENO schemes

5.1 Introduction

It has long been recognized that, the solutions of nonlinear hyperbolic equations can develop discontinuities (shocks) in finite time, even if the initial condition is smooth. Such a phenomenon greatly challenges the robustness of high order numerical methods, as spurious oscillations typically appear near shocks in numerical approximations (the Gibbs phenomenon), and may blow/mess up the simulation in later times. There have been numerous high order numerical methods developed to address this issue, among which the essentially non-oscillatory (ENO)/weighted essentially non-oscillatory (WENO) schemes have gained great success and have been widely used in applications.

The ENO methods, first developed by Harten et al. [33], use adaptive strategy to choose the smoothest stencil among several candidates to reconstruct the solution from its cell averages, hence the methods yield essentially non-oscillatory approximation near shocks. The original ENO scheme was based on the framework of finite volume methods, where the numerical fluxes at cell interfaces are obtained through reconstructed solution. Later, Shu and Osher proposed the finite difference ENO scheme in [73] based on ENO interpolation for nodal values and high order finite difference approximation for spatial derivatives of fluxes, which saves considerable computational cost in multi-dimensions, as the derivatives can be approximated dimension by dimension in finite difference schemes. Their subsequent work in [74] developed a simpler finite difference ENO scheme based on the Shu-Osher lemma to approximate the fluxes at cell interfaces by standard reconstruction for fluxes at grid points. The WENO methods were developed upon ENO, with the idea of using a convex combination of all candidate stencils rather than only one stencil in the original ENO scheme. In the pioneer work of WENO schemes, Liu et al. [48] used

linear weights to combine the candidate stencils in r-th order ENO schemes to yield $(r+1)$-th order of accuracy. It was later improved by Jiang and Shu [35] to achieve $(2r-1)$-th order of accuracy on the same stencils, by adopting nonlinear weights based on smoothness indicators designed for optimal accuracy in smooth regions and essentially non-oscillatory fashion near discontinuities. Thereafter, intensive modifications and improvements of the WENO procedure have been developed, e.g. the mapped WENO [34], WENO-Z [4, 9], modified WENO to handle negative weights [68], multi-resolution WENO [99], Hermite WENO [61], among other variants. Both finite volume [33] and finite difference [73, 74] frameworks for ENO can be used with the above WENO procedures. In our work, we use the classic WENO-JS procedure [35], as it is most widely used and relatively simple to code. For more details about the history and development of ENO and WENO methods, one can refer to the surveys [71, 72].

The ENO/WENO methods perform very well for scalar conservation laws as they achieve uniformly high order accuracy in smooth regions and resolve shocks sharply with essentially non-oscillatory quality. However, when dealing with hyperbolic systems, the component-wise ENO/WENO procedure often produces oscillatory results near shocks, especially when waves corresponding to different characteristic fields interact, such as in Riemann problems. The primary approach to resolve this problem is to apply the ENO/WENO procedure to the local characteristic fields of the system obtained by local characteristic decomposition for the conserved variables/fluxes, and transform the results back to the conserved variables/fluxes afterwards. Below, we briefly review how the WENO methods for hyperbolic systems are used in cooperation with the local characteristic decomposition. For the ease of comparison with the algorithm to be developed in this chapter, we demonstrate it as per example of the alternative formulation of finite difference WENO scheme developed in [36] from

[73], which will be introduced with more details in Section 5.3.

We consider the hyperbolic system of m ($m > 1$) components

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{0}, \tag{5.1}$$

in one space dimension, where $\mathbf{u} = (u_1, \ldots, u_m) \in \mathbb{R}^m$ are the conserved variables and $\mathbf{f}(\mathbf{u}) = (f_1(\mathbf{u}), \ldots, f_m(\mathbf{u})) \in \mathbb{R}^m$ are the fluxes. Now and henceforth, we use bold face font to denote vectors or matrices.

Consider uniform grids with the grid point $x_j = j\Delta x$ centering in the cell $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] = [(j-\frac{1}{2})\Delta x, (j+\frac{1}{2})\Delta x], \forall j \in \mathbb{Z}$. The semi-discrete $(2r-1)$-th order alternative formulation of finite difference WENO scheme for (5.1) is formulated as

$$\frac{d\mathbf{u}_j}{dt} + \frac{1}{\Delta x}\left(\hat{\mathbf{f}}_{j+\frac{1}{2}} - \hat{\mathbf{f}}_{j-\frac{1}{2}}\right) = \mathbf{0}, \tag{5.2}$$

where $\mathbf{u}_j$ is the approximation to $\mathbf{u}(x_j, t)$, $\hat{\mathbf{f}}_{j+\frac{1}{2}} = \hat{\mathbf{f}}(\mathbf{u}^-_{j+\frac{1}{2}}, \mathbf{u}^+_{j+\frac{1}{2}}, \cdots)$ is the numerical flux, whose definition and arguments omitted for brevity will be detailed in later sections, and $\mathbf{u}^{\pm}_{j+\frac{1}{2}}$ are approximations to $\mathbf{u}(x_{j+\frac{1}{2}}, t)$ from interpolants on I_j and I_{j+1}. We denote the WENO interpolation for a scalar-valued grid function v at $x_{j+\frac{1}{2}}$ on I_j by $v^-_{j+\frac{1}{2}} = \text{weno}(v_{j-r+1}, \ldots, v_{j+r-1})$, whose implementation will be detailed in Section 5.3. The WENO interpolation for $v^+_{j-\frac{1}{2}}$ follows from mirror symmetry, i.e. $v^+_{j-\frac{1}{2}} = \text{weno}(v_{j+r-1}, \ldots, v_{j-r+1})$. We shall abuse the notation to also let it denote the component-wise WENO interpolation for vectors, e.g. $\mathbf{v}^-_{j+\frac{1}{2}} = \text{weno}(\mathbf{v}_{j-r+1}, \ldots, \mathbf{v}_{j+r-1})$.

The flowchart of the alternative formulation of finite difference WENO algorithm (5.2) with local characteristic decomposition, based on the nodal values $\{\mathbf{u}^n_j\}_{j\in\mathbb{Z}}$ at time level t^n, is given as follows, where the superscript n is omitted for brevity and

the computation is carried out for all $j \in \mathbb{Z}$:

1. Approximate the solution at $x_{j+\frac{1}{2}}$ by the arithmetic mean $\mathbf{u}_{j+\frac{1}{2}} = \frac{1}{2}(\mathbf{u}_j + \mathbf{u}_{j+1})$, or the Roe's average [65] satisfying $\mathbf{f}(\mathbf{u}_{j+1}) - \mathbf{f}(\mathbf{u}_j) = \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u}_{j+\frac{1}{2}})(\mathbf{u}_{j+1} - \mathbf{u}_j)$, if it is available.

2. Perform the eigendecomposition on the Jacobian matrix: $\frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u}_{j+\frac{1}{2}}) = \mathbf{R}_{j+\frac{1}{2}} \mathbf{\Lambda}_{j+\frac{1}{2}} \mathbf{R}^{-1}_{j+\frac{1}{2}}$, where $\mathbf{\Lambda}_{j+\frac{1}{2}}$ and $\mathbf{R}_{j+\frac{1}{2}}$ are the diagonal matrix containing all eigenvalues and the eigenmatrix consist of a complete set of eigenvectors as its columns, respectively, of the Jacobian matrix.

3. Calculate the local characteristic variables: $\mathbf{v}_i = \mathbf{R}^{-1}_{j+\frac{1}{2}} \mathbf{u}_i$, on the stencils $i = j - r + 1, \ldots, j + r$.

4. Perform the WENO interpolation for the local characteristic variables to obtain $\mathbf{v}^-_{j+\frac{1}{2}} = \text{weno}(\mathbf{v}_{j-r+1}, \ldots, \mathbf{v}_{j+r-1})$ and $\mathbf{v}^+_{j+\frac{1}{2}} = \text{weno}(\mathbf{v}_{j+r}, \ldots, \mathbf{v}_{j-r+2})$.

5. Transform the local characteristic variables back to the conserved variables: $\mathbf{u}^\pm_{j+\frac{1}{2}} = \mathbf{R}_{j+\frac{1}{2}} \mathbf{v}^\pm_{j+\frac{1}{2}}$.

6. Calculate the numerical fluxes $\hat{\mathbf{f}}_{j+\frac{1}{2}}$ to evolve the scheme (5.2) in time.

As we can see, the steps 1, 2, 3 and 5 are extra costs due to the local characteristic decomposition. In particular, there are $2r$ matrix-vector multiplications at every cell interface $x_{j+\frac{1}{2}}$ at the step 3, which is responsible for most of the floating point operations.

There have been some attempts on avoiding or reducing the costs on local characteristic decomposition in numerical schemes, meanwhile maintaining the essentially non-oscillatory performance, but only limited successes were achieved. In [35], Jiang

and Shu computed the weights in WENO from entropy and pressure instead of the characteristic variables for Euler systems, to reduce part of the operations in local characteristic decomposition. In [98], Zheng et al. argued that at the contact discontinuity on interface of two-medium flow, direct WENO interpolation for primary variables is better than component-wise interpolation for conserved variables, but local characteristic decomposition was still applied therein to the primitive variables to get more satisfactory results. Low order central schemes [52, 49] can be used without local characteristic decomposition. However, the local characteristic decomposition is still necessary to control spurious oscillations when orders of the schemes are high [59].

In this chapter, we propose an efficient implementation of finite difference WENO schemes that is local characteristic decomposition free, for a special class of hyperbolic systems endowed with a coordinate system of Riemann invariants. Examples of such systems include all two-component hyperbolic systems and some multi-component systems to be introduced in Section 5.2. The key idea of the method is to apply the WENO procedure to the nodal values of the coordinate system of Riemann invariants, which are (one-to-one) nonlinear algebraic functions of the conserved variables, and transform the interpolated values back to the conserved variables in the calculation of fluxes. The improvement in efficiency is due to the fact that, the characteristic decomposition for the WENO procedure is calculated locally, namely the conserved variables/fluxes at every node need to be projected onto local characteristic fields by different inverse eigenmatrices at different cell interfaces, while the Riemann invariants have definite algebraic relation with the conserved variables thus only need to be calculated once per node. A comparison of floating point operations in these two methods are shown in Appendix C.1. The good non-oscillatory performance of such treatment is justified by both theoretical properties of hyperbolic

systems and numerical tests.

Due to the nonlinearity of the algebraic relation between Riemann invariants and conserved variables/fluxes, we cannot use any reconstruction based numerical schemes like the finite volume WENO or the traditional Shu-Osher lemma based finite difference WENO, as we cannot directly transfer the cell averages between Riemann invariants and conserved variables/fluxes. On the other hand, the transform between nodal values is straightforward, thus we adopt the alternative formulation of finite difference WENO scheme [36], which is based on WENO interpolation for nodal values. Its implementation will be demonstrated in Section 5.3. For detailed introduction and comparison with the traditional finite difference WENO for the alternative formulation, one can refer to [36].

The rest of the chapter is organized as follows. In Section 5.2, we review the definition of Riemann invariants and their important properties, and give examples of hyperbolic systems endowed with a coordinate system of Riemann invariants. In Section 5.3, we give a detailed description for our algorithm. We use numerical tests in Section 5.4 to demonstrate the efficiency and good performance of our methods. Finally, we end up with some concluding remarks in Section 5.5.

5.2 Riemann invariants

In this section, we review the definition and important properties of Riemann invariants of hyperbolic system of conservation laws.

We consider the hyperbolic system (5.1), with $\mathbf{u} = (u_1, \ldots, u_m)^T$ the conserved variables taking values in an open set $\mathscr{O} \subset \mathbb{R}^m$, and $\mathbf{f}(\mathbf{u}) = (f_1(\mathbf{u}), \ldots, f_m(\mathbf{u}))^T$ a

smooth flux function on $\mathscr{O}$. From hyperbolicity, the Jacobian matrix $\frac{\partial \mathbf{f}}{\partial \mathbf{u}}$ has a complete set of eigenvectors $\mathbf{r}_1(\mathbf{u}), \mathbf{r}_2(\mathbf{u}), \ldots, \mathbf{r}_m(\mathbf{u})$ corresponding to the real eigenvalues $\lambda_1(\mathbf{u}) \leq \lambda_2(\mathbf{u}) \leq \ldots \leq \lambda_m(\mathbf{u})$, for all $\mathbf{u} \in \mathscr{O}$.

The Riemann invariants of the hyperbolic system (5.1) is defined as follows [75]:

Definition 5.2.1. *An i-Riemann invariant* $(1 \leq i \leq m)$ *of the hyperbolic system* (5.1) *is a scalar-valued function* $w(\mathbf{u})$ *on* $\mathscr{O}$*, such that* $\nabla w(\mathbf{u}) \cdot \mathbf{r}_i(\mathbf{u}) = 0$*,* $\forall u \in \mathscr{O}$*, where* $\mathbf{r}_i(\mathbf{u})$ *is an eigenvector of the Jacobian matrix* $\frac{\partial \mathbf{f}}{\partial \mathbf{u}}$ *corresponding to the eigenvalue* $\lambda_i(\mathbf{u})$*.*

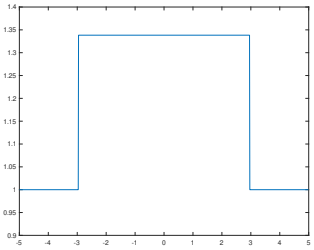
Riemann invariants are closely related to the Riemann problem, which is a Cauchy problem of the hyperbolic system (5.1) with the initial condition

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}_l, & x < 0 \\ \\ \mathbf{u}_r, & x > 0 \end{cases}, \tag{5.3}$$

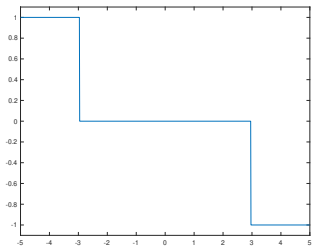
where $\mathbf{u}_l$ and $\mathbf{u}_r$ are constant states. It is well-known that the solution $\mathbf{u}(x,t)$ of the Riemann problem typically develops from the initial discontinuity at the origin into $m+1$ constant states in sector regions separated by the i-shock, contact or rarefaction wave, for $i = 1, 2, \ldots, m$, which is a characterization of the fundamental behavior of solutions of hyperbolic systems involving discontinuities. An important property of Riemann invariants across waves is stated as follows [75]:

Theorem 5.2.2. *The change of an i-Riemann invariant* w *of the hyperbolic system* (5.1) *across an i-shock wave is of third order in* ϵ*, i.e.* $|w(\mathbf{u}_l) - w(\mathbf{u}_r)| = O(\epsilon^3)$*, where* $\mathbf{u}_l$ *and* $\mathbf{u}_r$ *are the states on the left and right sides of the i-shock, respectively, and* $\epsilon = |\lambda_i(\mathbf{u}_l) - \lambda_i(\mathbf{u}_r)|$ *is a measure of the strength of the i-shock. In addition, the i-Riemann invariant is unchanged across an i-rarefaction or contact wave.*

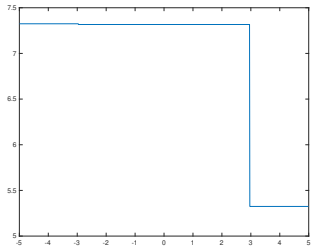
Roughly speaking, the i-Riemann invariant is unchanged or almost unchanged across an i-wave, consult Figure 5.1, where h, hu are the conserved variables, and w_1, w_2 are the 1 and 2-Riemann invariants, respectively, in a Riemann problem of the shallow water equations (5.5).



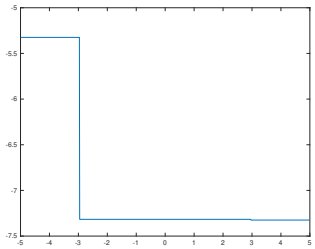
(a) Water depth h



(b) Discharge hu



(c) 1-Riemann invariant $w_1 = u - 2\sqrt{gh}$



(d) 2-Riemann invariant $w_2 = u + 2\sqrt{gh}$

Figure 5.1: Conserved variables and Riemann invariants in a Riemann problem of the shallow water equations

The WENO interpolation/reconstruction procedure performs very well if there is only one discontinuity in the stencil. However, the results turn out to be less satisfactory when there are multiple shocks in the stencil. The property of Riemann invariants in Theorem 5.2.2 gives us a hint to perform the WENO procedure on the $1, 2$-Riemann invariants of hyperbolic systems when $m = 2$, as there is only one major discontinuity in each Riemann invariant in Riemann problems. We shall show in numerical section that such a treatment yields very satisfactory non-oscillatory results.

A direct extension of the above approach to hyperbolic systems with $m \geq 3$ is to perform the WENO procedure on m variables, each of which only admits one major jump in stencils. An ideal choice is the coordinate system of Riemann invariants, which is defined as follows [19]:

Definition 5.2.3. *The system* (5.1) *is endowed with a coordinate system of Riemann invariants if there exist* m *scalar-valued functions* $w_1(\mathbf{u}), w_2(\mathbf{u}), \ldots, w_m(\mathbf{u})$ *on* $\mathscr{O}$ *such that,*

$$\nabla w_i(\mathbf{u}) \cdot \mathbf{r}_j(\mathbf{u}) = \delta_{i,j}, \quad i, j = 1, 2, \ldots, m,$$

where δ *is the Kronecker delta,* $\mathbf{r}_j(\mathbf{u})$ *is an eigenvector of the Jacobian matrix* $\frac{\partial \mathbf{f}}{\partial \mathbf{u}}$ *corresponding to the eigenvalue* $\lambda_j(\mathbf{u}), 1 \leq j \leq m$*. The variables* $(w_1(\mathbf{u}), w_2(\mathbf{u}), \ldots, w_m(\mathbf{u}))$ *are called a coordinate system of Riemann invariants of* (5.1)*.*

To this end, we give some examples of hyperbolic systems of conservation laws endowed with a coordinate system of Riemann invariants.

Example 5.2.1. *The linear hyperbolic system*

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = \mathbf{0}, \tag{5.4}$$

where $\mathbf{A} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^{-1}$ *for some diagonal matrix* $\mathbf{\Lambda}$ *and eigenmatrix* $\mathbf{R}$*, has a coordinate system of Riemann invariants* $(w_1, w_2, \ldots, w_m)$ *with* $w_i(\mathbf{u}) = \mathbf{l}_i\mathbf{u}, 1 \leq i \leq m$*, where* $\mathbf{l}_i$ *is the* i*-th row of* $\mathbf{R}^{-1}$*.*

Example 5.2.2. *The shallow water equations in one dimension*

$$\begin{pmatrix} h \\ hu \end{pmatrix}_t + \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \end{pmatrix}_x = \mathbf{0} \tag{5.5}$$

where h *is the water height,* u *is the velocity of the fluid, and* g *is the gravitational*

constant, is endowed with a coordinate system of Riemann invariants $(w_1, w_2) = (u + 2\sqrt{gh}, u - 2\sqrt{gh})$.

The shallow water equations in two dimensions

$$\begin{pmatrix} h \\ hu \\ hv \end{pmatrix}_t + \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{pmatrix}_x + \begin{pmatrix} hv \\ huv \\ hv^2 + \frac{1}{2}gh^2 \end{pmatrix}_y = \mathbf{0}, \tag{5.6}$$

where u *and* v *are velocities of the fluid in* x *and* y *directions, respectively, has coordinate systems of Riemann invariants* $(w_1, w_2, w_3) = (u - 2\sqrt{gh}, v, u + 2\sqrt{gh})$ *and* $(w_1, w_2, w_3) = (v - 2\sqrt{gh}, u, v + 2\sqrt{gh})$ *in* x *and* y *directions, respectively, in the sense that the states of fluid are constant in the other direction (in this case, the system is of the form of one dimensional equations, which is known as the split multi-dimensional problem).*

Example 5.2.3. *The hyperbolic system of electrophoresis of* m *components*

$$\partial_t u_i + \partial_x \left(\frac{c_i u_i}{\sum_{j=1}^{n} u_j} \right) = 0, \quad i = 1, 2, \ldots, m, \tag{5.7}$$

where $c_1 < c_2 < \cdots < c_m$ *are positive constants, is endowed with a coordinate system of Riemann invariants* $(w_1, w_2, \ldots, w_m)$*, where* $w_i \in (c_i, c_{i+1})$ *is the solution of the equation* $\sum_{j=1}^{m} \frac{u_j}{c_j - w} = 0$*, for* $i = 1, 2, \ldots, m-1$*, and* $w_m = \sum_{j=1}^{m} \frac{u_j}{c_j}$.

This system models the separation of ionized chemical compounds in solution driven by an electric filed, where c_i *and* u_i *denote the electrophoretic mobility and concentration of the* i*-th component, respectively, see [2] for more details about its physical backgrounds.*

Example 5.2.4. *The hyperbolic system of planar electromagnetic waves in nonlinear*

isotropic dielectrics

$$\begin{pmatrix} B_1 \\ B_2 \\ D_1 \\ D_2 \end{pmatrix}_t + \begin{pmatrix} -\frac{\Psi'(r)}{r} D_2 \\ \frac{\Psi'(r)}{r} D_1 \\ \frac{\Psi'(r)}{r} B_2 \\ -\frac{\Psi'(r)}{r} B_1 \end{pmatrix}_x = \mathbf{0}, \tag{5.8}$$

where $B = (B_1, B_2)^T$ *is the magnetic induction,* $D = (D_1, D_2)$ *is the electric displacement,* $\Psi(r)$ *is the electromagnetic energy, and* $r = \sqrt{B_1^2 + B_2^2 + D_1^2 + D_2^2}$*, is endowed with a coordinate system of Riemann invariants* (w_1, w_2, w_3, w_4)*.*

If we define a, b, p, q *by* $pe^{ia} = \frac{1}{\sqrt{2}}(B_2 + D_1 - i(B_1 - D_2))$ *and* $qe^{ib} = \frac{1}{\sqrt{2}}(-B_2 + D_1 + i(B_1 + D_2))$*, then* $w_1 = a, w_2 = b$*, and* w_3, w_4 *are the* $1,2$*-Riemann invariants of the smaller hyperbolic system*

$$\begin{pmatrix} p \\ q \end{pmatrix}_t + \begin{pmatrix} \frac{\Psi'(r)}{r} p \\ -\frac{\Psi'(r)}{r} q \end{pmatrix}_x = \mathbf{0}, \quad r = \sqrt{p^2 + q^2}. \tag{5.9}$$

5.3 The algorithms

In this section, we overview the WENO-JS interpolation, and establish our algorithms in the framework of alternative formulation of finite difference WENO scheme in one and two space dimensions. We shall assume the grids are uniform and, for simplicity, only consider periodic boundaries.

5.3.1 Overview of the WENO-JS interpolation

The $(2r-1)$-th order WENO-JS interpolation for a scalar-valued grid function v is described as follows.

First, we define the small stencils $S_k = \{x_{j-r+k}, \ldots, x_{j-1+k}\}$ to calculate the $(r-1)$-th order polynomial interpolant $p^{(k)}(x)$ of v on I_j, for $k = 1, 2, \ldots, r$, and the big stencil $S_0 = \cup_{k=1}^{r} S_k = \{x_{j-r+1}, \ldots, x_{j+r-1}\}$ to calculate the $(2r-2)$-th order polynomial interpolant $p^{(0)}(x)$ of v on I_j, such that

$$p^{(k)}(x_{j-r+k+m-1}) = v_{j-r+k+m-1}, \quad m = 1, 2, \ldots, r,$$

for $k = 1, 2, \ldots, r$, and

$$p^{(0)}(x_{j-r+m}) = v_{j-r+m}, \quad m = 1, 2, \ldots, 2r-1,$$

so that we yield

$$v_{j+\frac{1}{2}}^{-(k)} = p^{(k)}(x_{j+\frac{1}{2}}) = \sum_{m=1}^{r} a_m^{(k)} v_{j-r+k+m-1} = v(x_{j+\frac{1}{2}}) + O(\Delta x^r), \quad k = 1, 2, \ldots, r, \tag{5.10}$$

and

$$v_{j+\frac{1}{2}}^{-(0)} = p^{(0)}(x_{j+\frac{1}{2}}) = \sum_{k=1}^{r} \gamma_k v_{j+\frac{1}{2}}^{-(k)} = v(x_{j+\frac{1}{2}}) + O(\Delta x^{2r-1}), \tag{5.11}$$

where $\{\gamma_k\}_{k=1}^{r}$ are the so-called optimal linear weights with $\gamma_k \geq 0$, for $k = 1, 2, \ldots, r$ [8] and $\sum_{k=1}^{r} \gamma_k = 1$, and $\{a_m^{(k)}\}_{m,k=1}^{r}$ are constant coefficients.

Then, we introduce the nonlinear weights $\{\omega_k\}_{k=1}^{r}$, which is designed in the principle that, in smooth regions w_k is close to γ_k to achieve optimal accuracy while,

if containing discontinuities, w_k is close to zero to minimize the contribution of the stencil containing discontinuities in WENO interpolation:

$$\omega_k = \frac{\tilde{\omega}_k}{\sum_{m=1}^{r} \tilde{\omega}_m}, \quad \tilde{\omega}_k = \frac{\gamma_k}{(\beta_k + \epsilon)^2}, \quad k = 1, 2, \ldots, r, \tag{5.12}$$

where ϵ is a small positive number, e.g. $\epsilon = 10^{-6}$, to avoid the case of linear weights being divided by zero, and $\{\beta_k\}_{k=1}^{r}$ are the smoothness indicators of the polynomial interpolant $p^{(k)}(x)$ on I_j:

$$\beta_k = \sum_{\ell=1}^{r} \Delta x^{2\ell-1} \int_{I_j} \left(\frac{d^\ell}{dx^\ell} p^{(k)}(x) \right)^2 dx. \tag{5.13}$$

Finally, the WENO-JS interpolation $v^{-}_{j+\frac{1}{2}}$ is calculated by

$$v^{-}_{j+\frac{1}{2}} = \sum_{k=1}^{r} \omega_k v^{-(k)}_{j+\frac{1}{2}}. \tag{5.14}$$

For instance, in the fifth order ($r = 3$) WENO-JS interpolation, we have

$$\begin{aligned}
v^{-(1)}_{j+\frac{1}{2}} &= \frac{3}{8} v_{j-2} - \frac{5}{4} v_{j-1} + \frac{15}{8} v_j, \\
v^{-(2)}_{j+\frac{1}{2}} &= -\frac{1}{8} v_{j-1} + \frac{3}{4} v_j + \frac{3}{8} v_{j+1}, \\
v^{-(3)}_{j+\frac{1}{2}} &= \frac{3}{8} v_j + \frac{3}{4} v_{j+1} - \frac{1}{8} v_{j+2},
\end{aligned}$$

and

$$\gamma_1 = \frac{1}{16}, \quad \gamma_2 = \frac{5}{8}, \quad \gamma_3 = \frac{5}{16},$$

and

$$\beta_1 = \frac{13}{12}(v_{j-2} - 2v_{j-1} + v_j)^2 + \frac{1}{4}(v_{j-2} - 4v_{j-1} + 3v_j)^2,$$
$$\beta_2 = \frac{13}{12}(v_{j-1} - 2v_j + v_{j+1})^2 + \frac{1}{4}(v_{j-1} - v_{j+1})^2,$$
$$\beta_3 = \frac{13}{12}(v_j - 2v_{j+1} + v_{j+2})^2 + \frac{1}{4}(3v_j - 4v_{j+1} + v_{j+2})^2.$$

For expressions of smoothness indicators in higher order WENO-JS interpolations, one can refer to [3].

5.3.2 The algorithm in one dimension

For the domain $[x_a, x_b]$, we take the uniform partition $x_a = x_0 < x_1 < \ldots < x_N = x_b$, and denote $\Delta x \equiv x_j - x_{j-1}$, $x_{j-\frac{1}{2}} = \frac{1}{2}(x_{j-1} + x_j)$, for $j = 1, 2, \ldots, N$. In the finite difference WENO scheme, we seek $\mathbf{u}_j$ to approximate $\mathbf{u}(x_j, t)$, and $\mathbf{u}^{\pm}_{j+\frac{1}{2}}$ to approximate the solution at $x_{j+\frac{1}{2}}$ from I_j and I_{j+1}, respectively. For the ease of writing, we shall use subscript indices exceeding the domain in the cyclic sense.

The semi-discrete $(2r-1)$-th order alternative formulation of finite difference WENO scheme for the hyperbolic system (5.1) in one dimensions is given by (5.2), in which we define

$$\hat{\mathbf{f}}_{j+\frac{1}{2}} = \mathbf{h}(\mathbf{u}^-_{j+\frac{1}{2}}, \mathbf{u}^+_{j+\frac{1}{2}}) + \sum_{m=1}^{r-1} a_{2m} \Delta x^{2m} \left(\frac{\partial^{2m}}{\partial x^{2m}} \mathbf{f} \right)_{j+\frac{1}{2}}, \tag{5.15}$$

where $\mathbf{h}(\cdot, \cdot)$ is the numerical flux based on exact or approximate Riemann solvers, e.g. the Godunov flux, the Lax-Friedrichs flux, or the HLLC-type fluxes, among others, and the coefficients $a_2 = -\frac{1}{24}, a_4 = \frac{7}{5760}, a_6 = -\frac{31}{967680}, a_8 = \frac{127}{154828800}, a_{10} = -\frac{73}{3503554560}, \ldots$, are obtained through Taylor expansion to approximate the spacial

derivative of flux with high accuracy, see [73].

Following the practice in [60, 36], we calculate $\mathbf{u}^{\pm}_{j+\frac{1}{2}}$ in $\mathbf{h}(\mathbf{u}^{-}_{j+\frac{1}{2}}, \mathbf{u}^{+}_{j+\frac{1}{2}})$ by WENO interpolation, while use simple central difference to approximate the spatial derivatives of $\mathbf{f}$ in the remaining terms to save computational costs, as these terms contain at least Δx^2 in the coefficients, which is expected to contribute much less oscillations. To attain enough accuracy, we use the stencil $\{x_{j-r+1}, \ldots, x_j, \ldots, x_{j+r}\}$ in the central difference approximation for $\left(\frac{\partial^{2m}}{\partial x^{2m}}\mathbf{f}\right)_{j+\frac{1}{2}}$.

For instance, in the fifth order finite difference WENO, we use

$$\left(\frac{\partial^2}{\partial x^2}\mathbf{f}\right)_{j+\frac{1}{2}} = \frac{1}{\Delta x^2}\left(-\frac{5}{48}\mathbf{f}_{j-2} + \frac{13}{16}\mathbf{f}_{j-1} - \frac{17}{24}\mathbf{f}_j - \frac{17}{24}\mathbf{f}_{j+1} + \frac{13}{16}\mathbf{f}_{j+2} - \frac{5}{48}\mathbf{f}_{j+3}\right),$$
$$\left(\frac{\partial^4}{\partial x^4}\mathbf{f}\right)_{j+\frac{1}{2}} = \frac{1}{\Delta x^4}\left(\frac{1}{2}\mathbf{f}_{j-2} - \frac{3}{2}\mathbf{f}_{j-1} + \mathbf{f}_j + \mathbf{f}_{j+1} - \frac{3}{2}\mathbf{f}_{j+2} + \frac{1}{2}\mathbf{f}_{j+3}\right).$$

If the hyperbolic system (5.1) is endowed with a coordinate system of Riemann invariants $\mathbf{w}$ with the one-to-one algebraic relation $\mathbf{w} = \mathbf{w}(\mathbf{u})$ and $\mathbf{u} = \mathbf{u}(\mathbf{w})$ to the conserved variables $\mathbf{u}$, the $(2r-1)$-th order alternative formulation of finite difference WENO scheme based on the nodal values $\{\mathbf{u}^n_j\}_{j=1}^N$ at time level t^n is given as follows, where the superscript n is omitted for simplicity and computation is carried out for all $j = 1, 2, \ldots, N$:

1. Calculate the coordinate system of Riemann invariants $\mathbf{w}_j = \mathbf{w}(\mathbf{u}_j)$.
2. Perform the WENO interpolation introduced in Section 5.3.1 on $\{\mathbf{w}_j\}_{j=1}^N$ to obtain $\mathbf{w}^{-}_{j+\frac{1}{2}} = \text{weno}(\mathbf{w}_{j-r+1}, \ldots, \mathbf{w}_{j+r-1})$ and $\mathbf{w}^{+}_{j+\frac{1}{2}} = \text{weno}(\mathbf{w}_{j+r}, \ldots, \mathbf{w}_{j-r+2})$.
3. Transform the results back to the conserved variables by $\mathbf{u}^{\pm}_{j+\frac{1}{2}} = \mathbf{u}\left(\mathbf{w}^{\pm}_{j+\frac{1}{2}}\right)$.
4. Calculate the numerical fluxes $\hat{\mathbf{f}}_{j+\frac{1}{2}}$ to evolve the scheme (5.2) in time.

To this end, we would like to introduce the time-marching approach used the algorithm. For the ODE system,

$$\mathbf{u}_t = \mathbf{L}(\mathbf{u}), \tag{5.16}$$

which is obtained from the semi-discrete finite difference scheme, we adopt the 4-th order 5 stage strong stability preserving Runge-Kutta (SSPRK$(4,5)$) method [76],

$$\begin{aligned}
\mathbf{u}^{(1)} &= \mathbf{u}^n + 0.39175222700392\Delta t\mathbf{L}(\mathbf{u}^n),\\
\mathbf{u}^{(2)} &= 0.44437049406734\mathbf{u}^n + 0.55562950593266\mathbf{u}^{(1)} + 0.36841059262959\Delta t\mathbf{L}(\mathbf{u}^{(1)}),\\
\mathbf{u}^{(3)} &= 0.62010185138540\mathbf{u}^n + 0.37989814861460\mathbf{u}^{(2)} + 0.25189177424738\Delta t\mathbf{L}(\mathbf{u}^{(2)}),\\
\mathbf{u}^{(4)} &= 0.17807995410773\mathbf{u}^n + 0.82192004589227\mathbf{u}^{(3)} + 0.54497475021237\Delta t\mathbf{L}(\mathbf{u}^{(3)}),\\
\mathbf{u}^{n+1} &= 0.00683325884039\mathbf{u}^n + 0.51723167208978\mathbf{u}^{(2)} + 0.12759831133288\mathbf{u}^{(3)}\\
&\quad + 0.34833675773694\mathbf{u}^{(4)} + 0.08460416338212\Delta t\mathbf{L}(\mathbf{u}^{(3)})\\
&\quad + 0.22600748319395\Delta t\mathbf{L}(\mathbf{u}^{(4)}),
\end{aligned}$$

where $\mathbf{u}^n$ and $\mathbf{u}^{n+1}$ are solutions at the time level t^n and t^{n+1}, respectively, and $\Delta t = t^{n+1} - t^n$. We refer to [25] and [26] for more details about the strong stability preserving (SSP), also called the total variation diminishing (TVD), Runge-Kutta or multi-step time discretization approaches.

In the numerical section, we shall use WENO schemes with spatial accuracy higher than fourth order (the temporal accuracy), as in applications it is usually the spatial accuracy that restricts the resolution of simulations.

5.3.3 The algorithm in two dimensions

For the two dimensional domain $[x_a, x_b] \times [y_a, y_b]$, we take the uniform partition $x_a = x_0 < x_1 < \cdots < x_N = x_b$ and $y_a = y_0 < y_1 < \cdots < y_M = y_b$ in x and y directions, respectively, and denote by $\Delta x \equiv x_i - x_{i-1}$, $x_{i-\frac{1}{2}} = \frac{1}{2}(x_{i-1} + x_i)$ for $i = 1, 2, \ldots, N$, and $\Delta y \equiv y_j - y_{j-1}$, $y_{j-\frac{1}{2}} = \frac{1}{2}(y_{j-1} + y_j)$ for $j = 1, 2, \ldots, M$. We seek $\mathbf{u}_{i,j}$ to approximate $\mathbf{u}(x_i, y_j, t)$, and $\mathbf{u}^{\pm}_{i+\frac{1}{2},j}$ and $\mathbf{u}^{\pm}_{i,j+\frac{1}{2}}$ to approximate $\mathbf{u}(x_{i+\frac{1}{2}}, y_j, t)$ and $\mathbf{u}(x_i, y_{j+\frac{1}{2}}, t)$, respectively, from different sides, in the finite difference WENO schemes.

The semi-discrete $(2r-1)$-th order alternative formulation of finite difference WENO scheme for the hyperbolic system

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x + \mathbf{g}(\mathbf{u})_y = \mathbf{0}, \tag{5.17}$$

in two dimensions is formulated as

$$\frac{d\mathbf{u}_{i,j}}{dt} + \frac{1}{\Delta x}\left(\hat{\mathbf{f}}_{i+\frac{1}{2},j} - \hat{\mathbf{f}}_{i-\frac{1}{2},j}\right) + \frac{1}{\Delta y}\left(\hat{\mathbf{g}}_{i,j+\frac{1}{2}} - \hat{\mathbf{g}}_{i,j-\frac{1}{2}}\right) = \mathbf{0}, \tag{5.18}$$

for $i = 1, 2, \ldots, N, j = 1, 2, \ldots, M$, where the fluxes are defined the same way as in one dimensional case, thanks to the advantage of finite difference schemes.

If the x-split problem of (5.17) is endowed with a coordinate system of Riemann invariants $\mathbf{w}$ and the y-split problem of (5.17) is endowed with a coordinate system of Riemann invariants $\mathbf{v}$, the algorithm based on the nodal values $\{\mathbf{u}^n_{i,j}\}^{N,M}_{i=1,j=1}$ at time level t^n is given as follows, where the superscript n is omitted for brevity and computation is carried out for all $i = 1, 2, \ldots, N, j = 1, 2, \ldots, M$:

1. Calculate the coordinate systems of Riemann invariants $\mathbf{w}_{i,j} = \mathbf{w}\left(\mathbf{u}_{i,j}\right)$ and $\mathbf{v}_{i,j} = \mathbf{v}\left(\mathbf{u}_{i,j}\right)$.

2. Perform the WENO interpolation introduced in Section 5.3.1 on $\{\mathbf{w}_{i,j}\}_{i=1,j=1}^{N,M}$ and $\{\mathbf{v}_{i,j}\}_{i=1,j=1}^{N,M}$ to obtain $\mathbf{w}^{-}_{i+\frac{1}{2},j} = \text{weno}(\mathbf{w}_{i-r+1,j}, \ldots, \mathbf{w}_{i+r-1,j})$, $\mathbf{w}^{+}_{i+\frac{1}{2},j} = \text{weno}(\mathbf{w}_{i+r,j}, \ldots, \mathbf{w}_{i-r+2,j})$, and $\mathbf{v}^{-}_{i,j+\frac{1}{2}} = \text{weno}(\mathbf{v}_{i,j-r+1}, \ldots, \mathbf{v}_{i,j+r-1})$, $\mathbf{v}^{+}_{i,j+\frac{1}{2}} = \text{weno}(\mathbf{v}_{i,j+r}, \ldots, \mathbf{v}_{i,j-r+2})$.

3. Calculate $\mathbf{u}^{\pm}_{i+\frac{1}{2},j} = \mathbf{u}\left(\mathbf{w}^{\pm}_{i+\frac{1}{2},j}\right)$ and $\mathbf{u}^{\pm}_{i,j+\frac{1}{2}} = \mathbf{u}\left(\mathbf{v}^{\pm}_{i,j+\frac{1}{2}}\right)$.

4. Calculate the numerical fluxes $\hat{\mathbf{f}}_{i+\frac{1}{2},j}$ and $\hat{\mathbf{g}}_{i,j+\frac{1}{2}}$ to evolve the scheme (5.18) in time.

We adopt the same time marching approach in the algorithm as in the one space dimension.

5.4 Numerical tests

In this section, we study the accuracy, efficiency and essentially non-oscillatory performance of the algorithm established in the previous sections, and compare them with those of the component-wise and local characteristic decomposition based WENO methods. For convenience, the component-wise WENO, local characteristic decomposition based WENO and Riemann invariants based WENO methods shall be abbreviated to CW-WENO, LCD-WENO and RI-WENO, respectively. We adopt the Lax-Friedrichs flux as the lowest order term in the flux (5.15). The numerical tests are carried out for examples given in Section 5.2, except for the first one, as the RI-WENO and LCD-WENO are exactly the same for linear hyperbolic systems.

Example 5.4.1. *(Accuracy and efficiency)*

In this example, we compare the accuracy and efficiency of RI-WENO with those of the CW-WENO and LCD-WENO for the one dimensional shallow water equations (5.5).

It is easy to verify that, if $v(x,t)$ *is a classic solution of the inviscid Burgers' equation* $v_t + \left(\frac{v^2}{2}\right)_x = 0$, *then* $h(x,t) = \frac{4}{9}v^2(x,t)$ *and* $u(x,t) = \frac{2}{3}v(x,t)$ *are solutions of the shallow water equations with the gravitational constant* $g = \frac{1}{4}$, *thus we let* $v(x,0) = \frac{1}{2}\sin(x) + 1$ *to determine the corresponding initial conditions of* h *and* u. *We set the the domain* $\Omega = [0, 2\pi]$ *and enforce the periodic boundary condition in the tests. The CFL conditions are taken as* $\Delta t = \frac{1}{10\lambda_{\max}}\Delta x^{\frac{2r-1}{4}}$ *in accuracy tests and* $\Delta t = \frac{1}{10\lambda_{\max}}\Delta x$ *in efficiency tests, where* $\lambda_{\max} = ||\left(|u| + \sqrt{gh}\right)||_\infty$, *and the terminal time is* $T = 0.1$.

The errors and orders of convergence of CW-WENO, RI-WENO and LCD-WENO for h *are given in Table 5.1, from which we can clearly observe that RI-WENO has the same orders of convergence as those of CW-WENO.*

Moreover, we compare the CPU times of CW-WENO, RI-WENO and LCD-WENO on different grids for different orders. The code is run on Oscar[1] with 1 *core and* $8GB$ *memory, and we count the CPU times by taking the average of* 1000 *trials of the complete computation. The results are given in Table 5.2, from which we can see that RI-WENO has roughly the same efficiency as CW-WENO while reduces considerable computational costs from LCD-WENO.*

Example 5.4.2. *(Shallow water equations in one dimension)*

In this test, we compare the essentially non-oscillatory performance of RI-WENO

	method	CW-WENO		RI-WENO		LCD-WENO	
r	N	L^1 error	order	L^1 error	order	L^1 error	order
3	20	5.08E-04	-	1.07E-04	-	1.29E-03	-
	40	1.53E-05	5.05	3.12E-06	5.10	9.23E-05	3.80
	80	4.12E-07	5.22	9.18E-08	5.08	6.03E-06	3.94
	160	1.16E-08	5.16	2.79E-09	5.04	3.64E-07	4.05
	200	3.71E-09	5.09	9.28E-10	4.93	1.23E-07	4.87
4	10	4.09E-03	-	8.43E-04	-	4.53E-03	-
	20	7.82E-05	5.71	6.90E-06	6.93	2.88E-04	3.98
	40	7.66E-07	6.67	6.42E-08	6.75	1.78E-05	4.01
	60	5.81E-08	6.36	5.61E-09	6.01	3.63E-06	3.92
5	10	1.75E-03	-	3.42E-04	-	2.12E-03	-
	20	8.11E-06	7.76	8.91E-07	8.58	3.89E-05	5.77
	30	1.87E-07	9.30	2.04E-08	9.32	3.26E-06	6.12
	40	1.41E-08	9.00	1.28E-09	9.62	3.65E-07	7.61
6	12	2.11E-04	-	3.86E-05	-	2.84E-04	-
	20	1.93E-06	9.20	2.68E-07	9.73	9.25E-06	6.70
	30	1.89E-08	11.40	2.92E-09	11.15	4.76E-07	7.32
	40	9.34E-10	10.46	1.09E-10	11.41	3.81E-08	8.78

Table 5.1: Accuracy of h of different WENO methods in Example 5.4.1

	method	CW-WENO	RI-WENO	LCD-WENO
r	N	CPU time (s)	CPU time (s)	CPU time (s)
3	50	1.58E-03	1.74E-03	3.55E-03
	100	6.22E-03	6.52E-03	1.44E-02
	150	9.94E-03	1.07E-02	2.77E-02
	200	1.72E-02	1.84E-02	4.87E-02
4	50	2.78E-03	2.97E-03	5.27E-03
	100	1.08E-02	1.13E-02	2.11E-02
	150	2.03E-02	2.10E-02	4.35E-02
	200	3.35E-02	3.69E-02	6.76E-02
5	50	3.78E-03	3.95E-03	6.50E-03
	100	1.47E-02	1.52E-02	2.60E-02
	150	2.88E-02	2.96E-02	5.42E-02
	200	4.65E-02	5.18E-02	8.38E-02
6	50	4.84E-03	5.01E-03	7.94E-03
	100	1.90E-02	1.95E-02	3.18E-02
	150	3.81E-02	3.88E-02	6.71E-02
	200	6.57E-02	6.76E-02	1.03E-01

Table 5.2: CPU times of different WENO methods in Example 5.4.1

with that of CW-WENO and LCD-WENO for the shallow water equations (5.5) *in one dimension.*

We first solve a Riemann problem with $g = 10$ *and the initial condition*

$$h(x,0) = \begin{cases} 0.125, & x < 0 \\ 1.000, & x > 0 \end{cases}, \quad u(x,0) = 0,$$

on the domain $\Omega = [-5,5]$ *with the partition* $N = 200$*. The plots of* h *of different methods at* $T = 1$ *are compared in Figure 5.2, where the reference solution are given by the exact Riemann solver.*

We then solve a periodic boundary problem with $g = 1$ *and the initial condition*

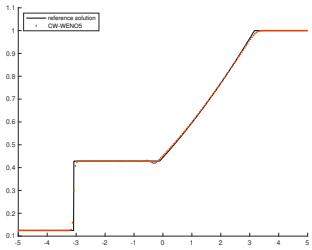
$$h(x,0) = \begin{cases} 2.0, & 0 < x < 10 \\ 1.5, & 10 < x < 20 \end{cases}, \quad u(x,0) = 0,$$

on the domain $\Omega = [0,20]$ *with the partition* $N = 200$*. The plots of* h *of different methods at* $T = 20$ *are compared in Figure 5.3, where the reference solution is obtained from the fifth order LCD-WENO on a grid containing* 10000 *cells.*

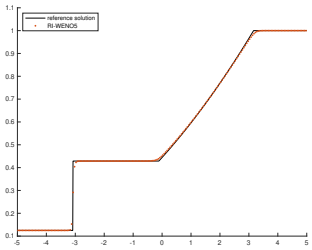
By comparison, we observe the essentially non-oscillatory effect of RI-WENO is much better than CW-WENO, and similar to LCD-WENO.

Example 5.4.3. *(Shallow water equations in two dimensions)*

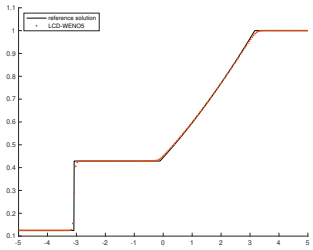
In this test, we compare the essentially non-oscillatory performance of RI-WENO with that of CW-WENO and LCD-WENO for the shallow water equations (5.6) *in two dimensions.*



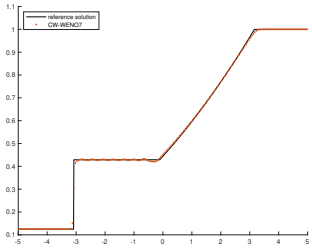
(a) CW-WENO, $r = 3$



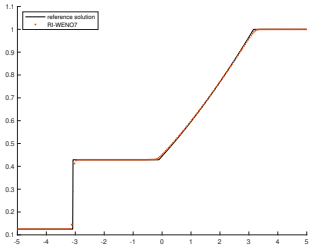
(b) RI-WENO, $r = 3$



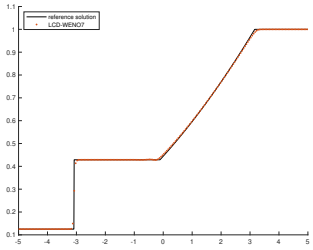
(c) LCD-WENO, $r = 3$



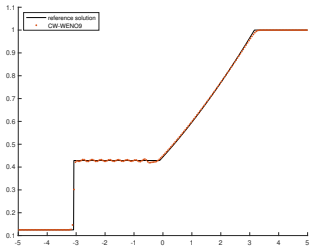
(d) CW-WENO, $r = 4$



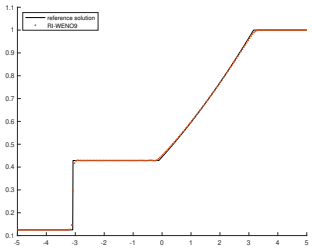
(e) RI-WENO, $r = 4$



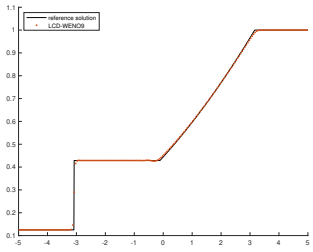
(f) LCD-WENO, $r = 4$



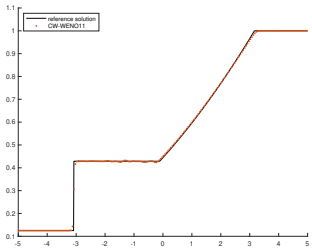
(g) CW-WENO, $r = 5$



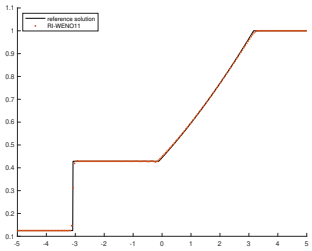
(h) RI-WENO, $r = 5$



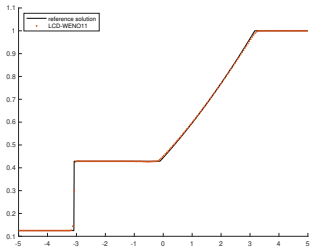
(i) LCD-WENO, $r = 5$



(j) CW-WENO, $r = 6$

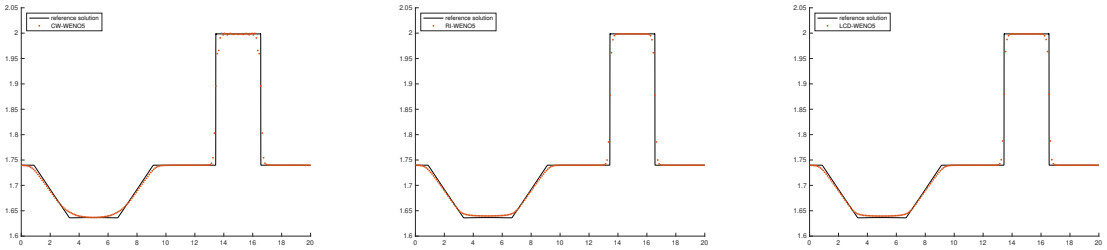


(k) RI-WENO, $r = 6$

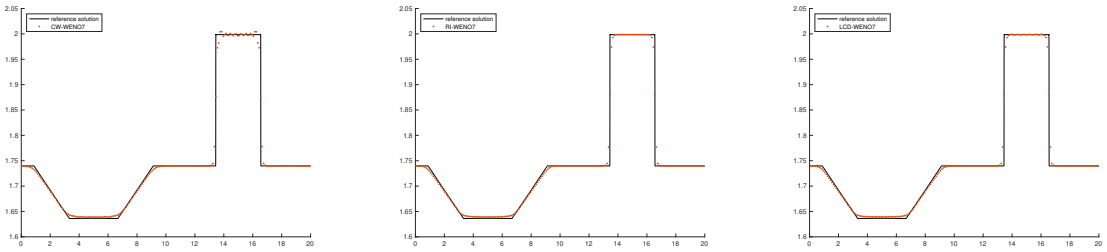


(l) LCD-WENO, $r = 6$

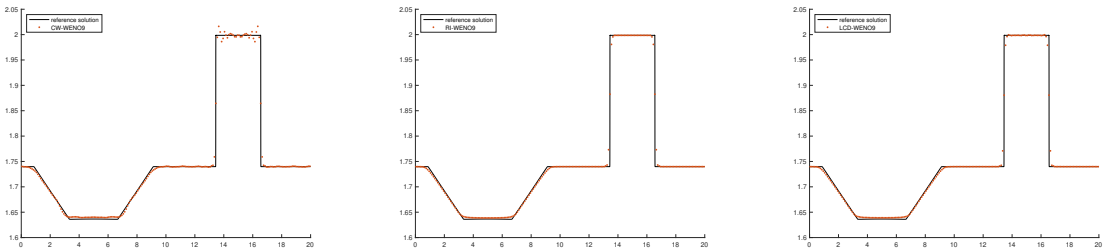
Figure 5.2: Solution h of different WENO methods for the Riemann problem in Example 5.4.2.



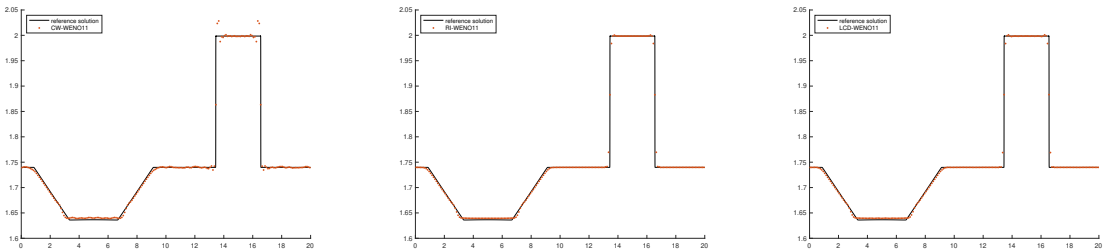
(a) CW-WENO, $r = 3$ (b) RI-WENO, $r = 3$ (c) LCD-WENO, $r = 3$



(d) CW-WENO, $r = 4$ (e) RI-WENO, $r = 4$ (f) LCD-WENO, $r = 4$



(g) CW-WENO, $r = 5$ (h) RI-WENO, $r = 5$ (i) LCD-WENO, $r = 5$



(j) CW-WENO, $r = 6$ (k) RI-WENO, $r = 6$ (l) LCD-WENO, $r = 6$

Figure 5.3: Solution h of different WENO methods for the periodic boundary problem in Example 5.4.2.

We solve a periodic boundary problem with $g = 1$ *and the initial condition*

$$h(x,y,0) = \begin{cases} 2.5, & 0 < x < 10, 0 < y < 10 \\ 2.0, & 0 < x < 10, 10 < y < 20 \\ 0.5, & 10 < x < 20, 0 < y < 10 \\ 1.5, & 10 < x < 20, 10 < y < 20 \end{cases}, \quad u(x,y,0) = v(x,y,0) = 0,$$

on the domain $\Omega = [0,20]^2$ *with* $N = M = 200$.

The contours of h *of different methods at* $T = 5$ *are shown in Figure 5.4, from which we can observe oscillations in the fourth quadrant in CW-WENO are eliminated by RI-WENO and LCD-WENO. Moreover, we plot the cut of* h *along* $y = 10$ *for different methods, and compare them with the reference solution obtained from the fifth order LCD-WENO on a* 1000×1000 *grid in Figure 5.5, from which we can see the non-oscillatory fashion of RI-WENO.*

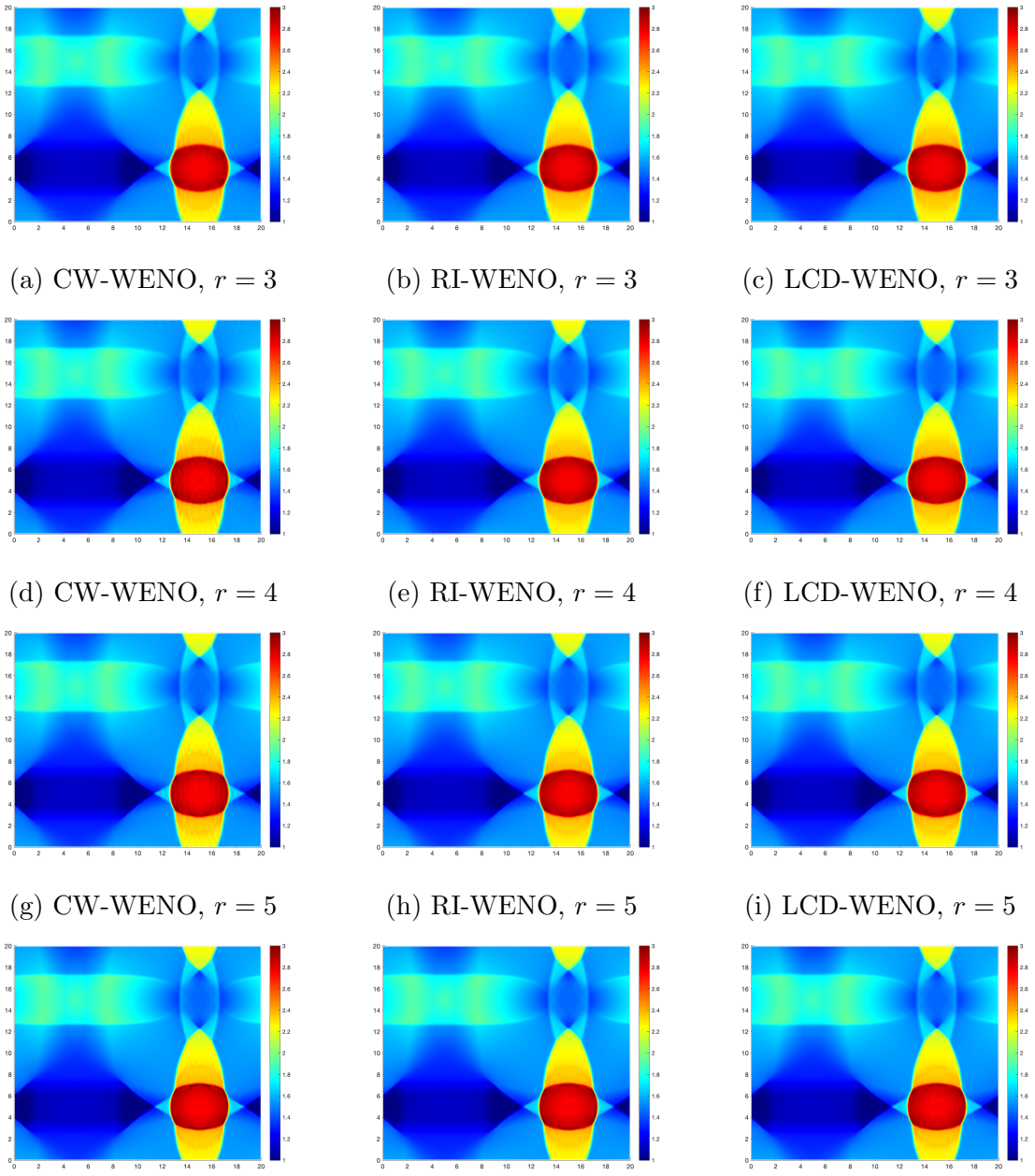
Example 5.4.4. *(Equations of electrophoresis)*

In this test, we compare the essentially non-oscillatory performance of RI-WENO with that of CW-WENO and LCD-WENO for the electrophoresis equations (5.7).

We solve the three-component periodic boundary problem with the electrophoretic mobilities $c_1 = 2, c_2 = 4, c_3 = 5$, *and the initial condition*

$$u_1(x,0) = \begin{cases} 1, & 0 < x < \frac{\pi}{2} \\ 0.01, & \frac{\pi}{2} < x < 2\pi \end{cases}, u_2(x,0) = \begin{cases} 0.01, & 0 < x < \frac{3\pi}{2} \\ 1, & \frac{3\pi}{2} < x < 2\pi \end{cases}, u_3(x,0) = 1,$$

on the domain $\Omega = [0, 2\pi]$ *with* $N = 200$.



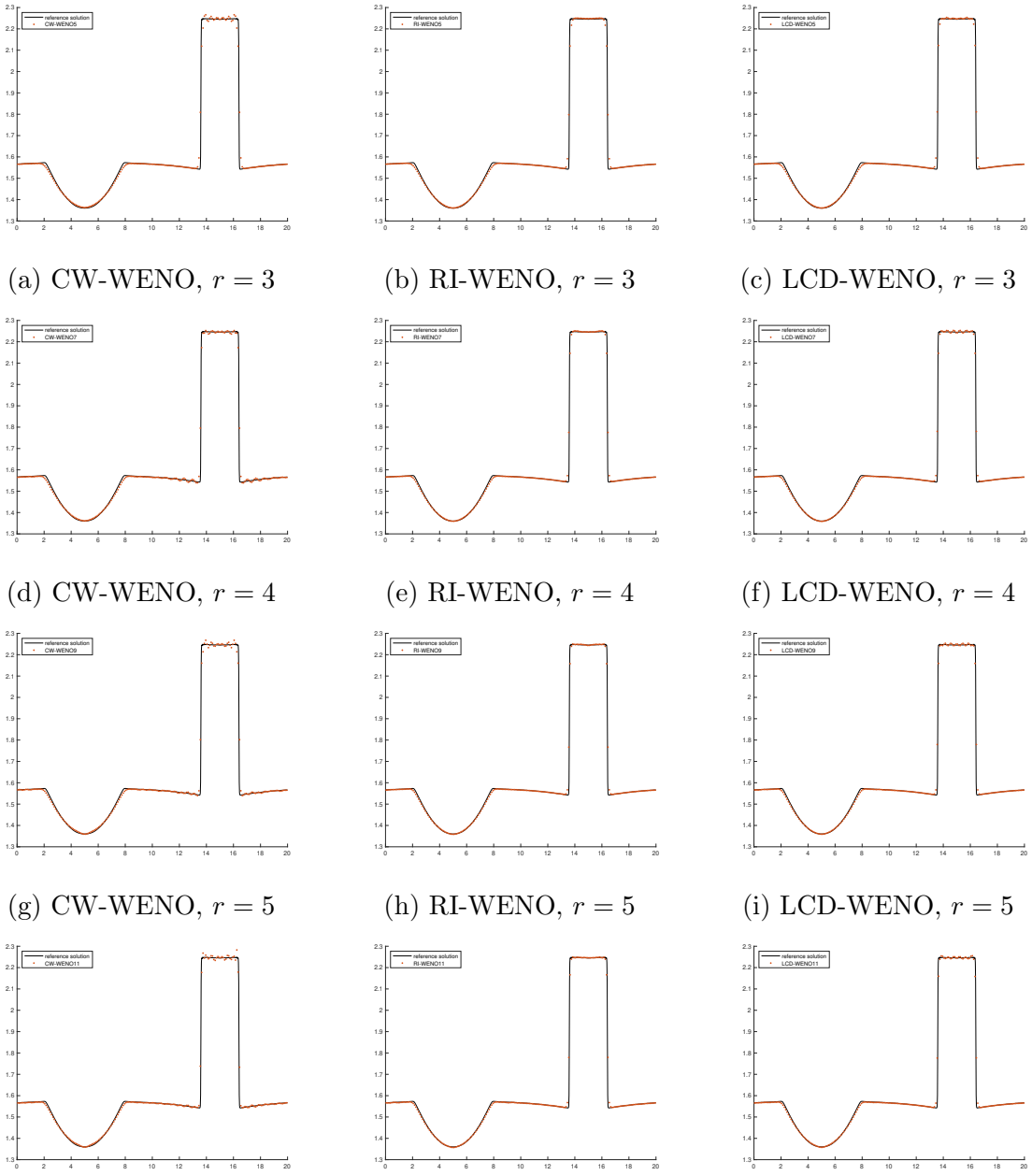
(a) CW-WENO, $r = 3$ (b) RI-WENO, $r = 3$ (c) LCD-WENO, $r = 3$

(d) CW-WENO, $r = 4$ (e) RI-WENO, $r = 4$ (f) LCD-WENO, $r = 4$

(g) CW-WENO, $r = 5$ (h) RI-WENO, $r = 5$ (i) LCD-WENO, $r = 5$

(j) CW-WENO, $r = 6$ (k) RI-WENO, $r = 6$ (l) LCD-WENO, $r = 6$

Figure 5.4: Contours of h of difference WENO methods in Example 5.4.3.



(a) CW-WENO, $r = 3$ (b) RI-WENO, $r = 3$ (c) LCD-WENO, $r = 3$

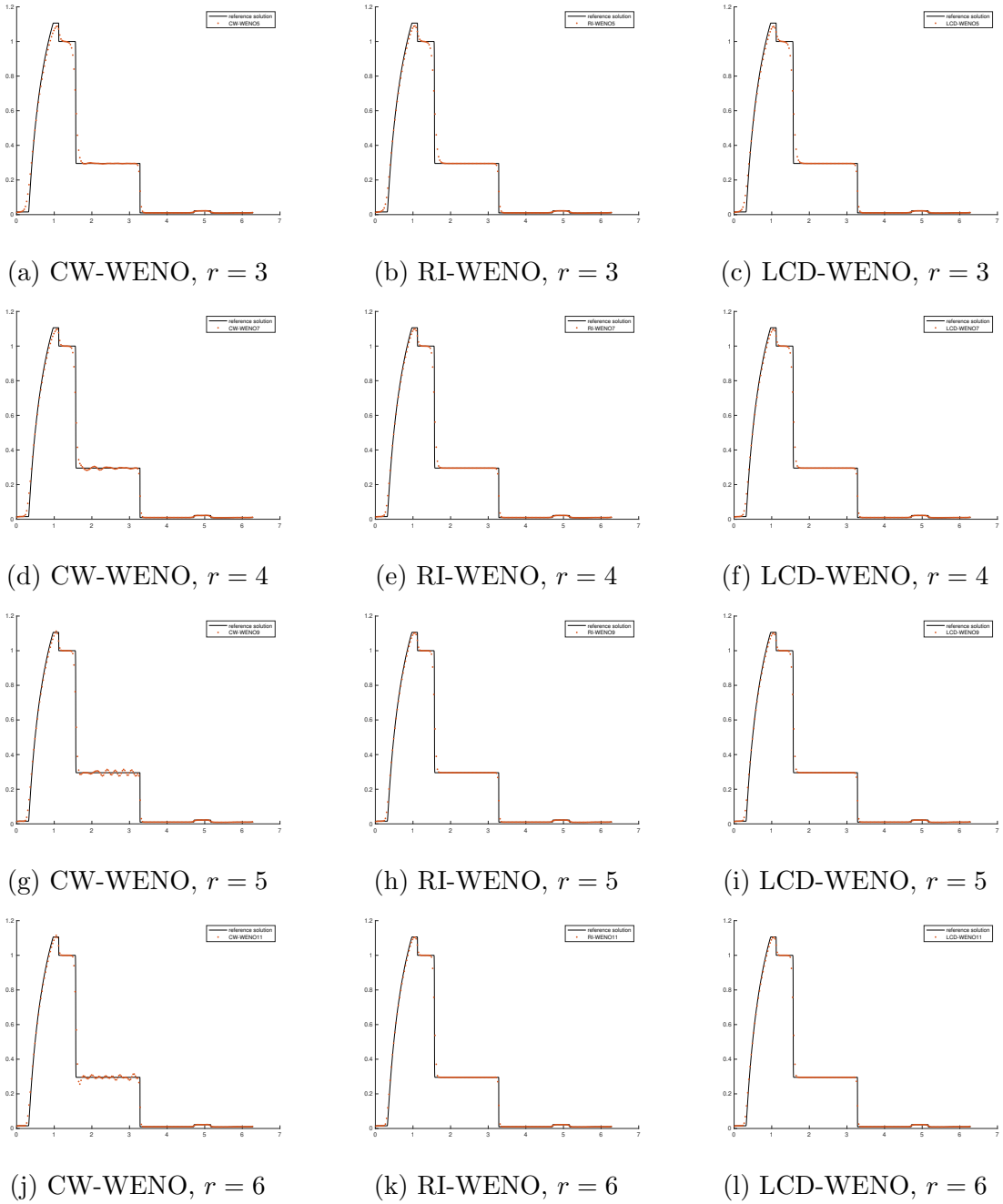
(d) CW-WENO, $r = 4$ (e) RI-WENO, $r = 4$ (f) LCD-WENO, $r = 4$

(g) CW-WENO, $r = 5$ (h) RI-WENO, $r = 5$ (i) LCD-WENO, $r = 5$

(j) CW-WENO, $r = 6$ (k) RI-WENO, $r = 6$ (l) LCD-WENO, $r = 6$

Figure 5.5: Cut of h along $y = 10$ of difference WENO methods in Example 5.4.3.

The plots of u_1 of different methods at $T = 0.5$ are compared in Figure 5.6, where the reference solution is obtained from the fifth order LCD-WENO on a grid containing 10000 *cells. The results of RI-WENO apparently have much less oscillation compared with those of CW-WENO and similar fashion with LCD-WENO.*



(a) CW-WENO, $r = 3$ (b) RI-WENO, $r = 3$ (c) LCD-WENO, $r = 3$

(d) CW-WENO, $r = 4$ (e) RI-WENO, $r = 4$ (f) LCD-WENO, $r = 4$

(g) CW-WENO, $r = 5$ (h) RI-WENO, $r = 5$ (i) LCD-WENO, $r = 5$

(j) CW-WENO, $r = 6$ (k) RI-WENO, $r = 6$ (l) LCD-WENO, $r = 6$

Figure 5.6: Solution u_1 of different WENO methods in Example 5.4.4.

Example 5.4.5. *(Equations of planar electromagnetic wave)*

In this test, we compare the essentially non-oscillatory performance of RI-WENO with that of CW-WENO and LCD-WENO for the planar electromagnetic wave equations (5.8)*. One can check that, if the electromagnetic energy satisfies* $\frac{\Psi'(r)}{r} = r^\alpha$ *for some* $\alpha > 0$*, the* $1, 2$*-Riemann invariants of the smaller hyperbolic system in Example 5.2.4 have the expressions* $w_3(p,q) = p - qG^{-1}(\log \frac{1}{q})$ *and* $w_4(p,q) = p + qG^{-1}(\log \frac{1}{q})$*, where* $G(\cdot)$ *is defined in the Appendix C.2.*

We solve the periodic boundary problem with $\alpha = 2$ *and the initial condition*

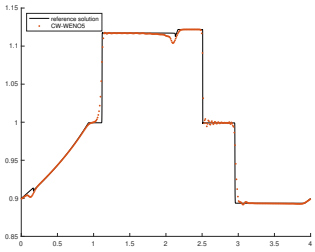
$$B_1(x,0) = \begin{cases} 1, & 0 < x < 2 \\ 0, & 2 < x < 4 \end{cases}, \quad B_2(x,0) = D_1(x,0) = D_2(x,0) = 1,$$

on the domain $\Omega = [0,4]$ *with* $N = 400$.

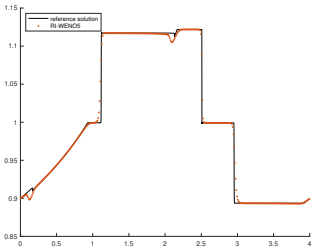
The plots of D_1 *of different methods at* $T = 0.3$ *are compared in Figure 5.7, where the reference solution is obtained from the fifth order LCD-WENO on a grid containing* 10000 *cells. From the comparison, we can see that RI-WENO has excellent essentially non-oscillatory performance.*

5.5 Concluding remarks

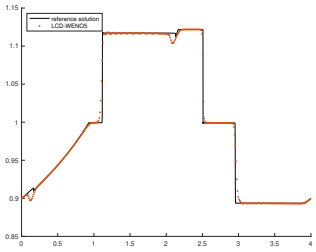
In this chapter, we establish a local characteristic decomposition free WENO method for hyperbolic system of conservation laws endowed with a coordinate system of Riemann invariants. We apply the WENO procedure to the coordinate system of Riemann invariants instead of the local characteristic fields of the hyperbolic system, thereby the efficiency is improved significantly. Due to the nonlinear algebraic relation of Riemann invariants and conserved variables/fluxes, we have to adopt the



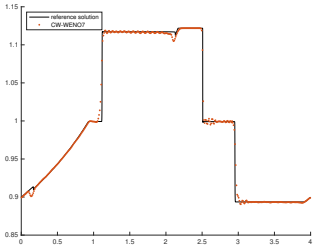
(a) CW-WENO, $r = 3$



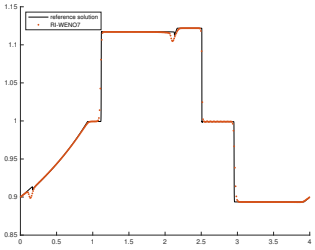
(b) RI-WENO, $r = 3$



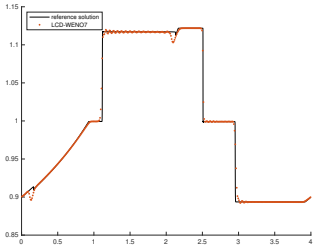
(c) LCD-WENO, $r = 3$



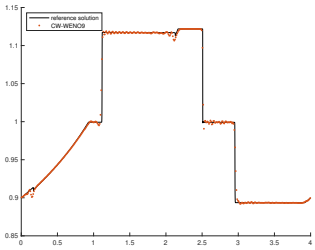
(d) CW-WENO, $r = 4$



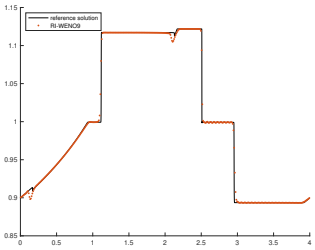
(e) RI-WENO, $r = 4$



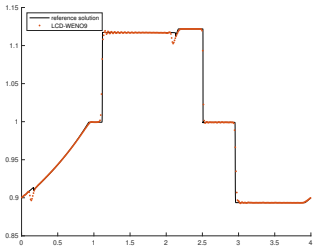
(f) LCD-WENO, $r = 4$



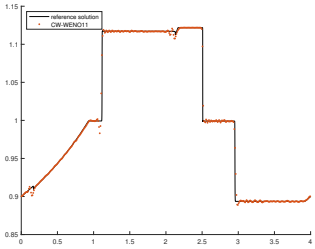
(g) CW-WENO, $r = 5$



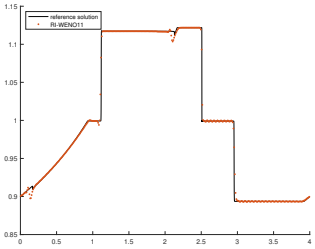
(h) RI-WENO, $r = 5$



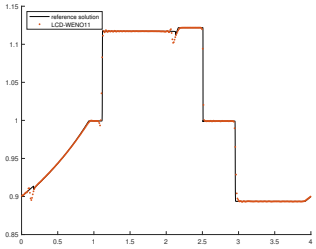
(i) LCD-WENO, $r = 5$



(j) CW-WENO, $r = 6$



(k) RI-WENO, $r = 6$



(l) LCD-WENO, $r = 6$

Figure 5.7: Solution D_1 of different WENO methods in Example 5.4.5.

interpolation based alternative formulation of finite difference WENO method. Numerical tests show that the Riemann invariants based WENO method has optimal order of convergence and roughly the same efficiency as that of the components-wise WENO, but its essentially non-oscillatory fashion is similar to that of local characteristic decomposition based WENO.

Chapter Six

Conclusion

This dissertation is based on the work in [87, 86, 85, 84], with the main contributions on designing third order bound-preserving Lax-Wendroff discontinuous Galerkin methods for scalar conservation laws and the Euler equations, establishing positivity-preserving discontinuous Galerkin methods for stationary hyperbolic balance laws, and proposing an efficient (local characteristic decomposition free) finite difference weighted essentially non-oscillatory scheme for the systems of hyperbolic conservation laws endowed with a coordinate system of Riemann invariants.

In the first topic of this dissertation, we have designed third order maximum-principle-satisfying DG methods for scalar conservation laws and positivity-preserving DG methods for the Euler equations based on the Lax-Wendroff temporal discretization, within the Zhang–Shu bound-preserving framework [93, 94]. For the first order spatial derivatives in the equations, we adopt the classic Lax-Friedrichs flux used in the bound-preserving work[93, 94], so that the proofs of the bound-preserving of this part can be omitted. The main difficulty of the bound-preserving Lax-Wendroff DG methods is the appearance of high order derivatives and mixed derivatives (in high dimensions) resulting from the Lax-Wendroff procedure. We adopt the bound-preserving DDG flux in [11] and the average fluxes to discretize the second and third order derivatives, respectively, such that the high order part is also bound-preserving. As for the mixed derivatives, by carefully designed expansions of high order temporal derivatives, we avoid their appearance in our numerical schemes, which is the key to the success of bound-preserving in high dimensions. Finally, we prove that, under suitable CFL conditions, the cell average of the LWDG scheme at the next time step is bounded, provided the solution stays in the desired bounds at the current time step. The scaling limiters, which were proved not to affect the high order accuracy and mass conservation of solutions, can then be used to enforce the bounds for the whole solution at the next time step, hence closing the loop of the bound-preserving

LWDG algorithm.

For stationary hyperbolic equations, it is known that their physical solutions are nonnegative, provided the corresponding boundary conditions and source terms are nonnegative. In the second topic, we establish the positivity-preserving DG methods for this kind of equation. We first follow the studies of Yuan et al. (2016) [90] and Ling et al. (2018) [46] in the Zhang–Shu bound-preserving framework [93], to rigorously preserve the positivity of cell averages of the designed DG scheme so that the conservative scaling limiter can be used to attain positivity without affecting accuracy. Via suitable quadrature rules in the DG formulation, we have successfully constructed the positivity-preserving scheme with the accuracy of arbitrarily high order for the variable coefficient equation (3.1) with $\lambda = 0$, and second and third orders for the variable coefficient equation (3.1) with $\lambda > 0$ and the nonlinear equation (3.2) with $\lambda \geq 0$, in one space dimension. Moreover, we have proposed the positivity-preserving scheme for constant coefficient equations with arbitrarily high order accuracy in two space dimensions and arbitrary odd order accuracy in three space dimensions. In a further study of this topic, we clarify a more appropriate definition of mass conservation for stationary hyperbolic equations. Instead of preserving the cell averages as we did in the previous work, we preserve the sum of the cell average and outflow fluxes in each cell. Novel conservative positivity-preserving limiters are proposed to accommodate the new definition of conservation, and their accuracy is investigated. The genuinely conservative high order positivity-preserving DG methods are established based on this definition. The new methods are able to preserve the positivity of more general types of equations with much simpler implementations and easier proofs for accuracy and the Lax-Wendroff theorem, compared with the previous methods.

In the last topic, we propose a local characteristic decomposition free finite dif-

ference WENO scheme for a particular class of systems of hyperbolic conservation laws, including the shallow water equations, equations of electrophoresis, and the planar electromagnetic waves in isotropic dielectrics, etc. As observed from numerical practices, e.g. [51, 59], the reconstruction performed on conservative variables is worse than the reconstruction on characteristic variables for the shock capturing fidelity of numerical methods for hyperbolic systems. Due to this reason, the local characteristic decomposition technique is widely used in the computation of compressible fluid dynamics, though it is computationally expensive. The main goal of our work is to find an alternative to the characteristic variables in the WENO reconstruction/interpolation procedure to avoid the expensive computational cost spent on local characteristic decomposition. As analyzed per the example of Riemann problems, the coordinate system of Riemann invariants admits only one major discontinuity in each component, thus it is expected to provide good shock capturing fidelity when WENO procedure is applied to it. We have verified this conjecture through extensive numerical experiments. By comparison, the Riemann invariants based WENO method has roughly the same efficiency as that of the components-wise WENO, which saves roughly half of the simulation time from the local characteristic decomposition based WENO scheme, but its essentially non-oscillatory fashion is similar to that of the local characteristic decomposition based WENO.

APPENDIX

Appendix A

Appendix for Chapter 2

A.1 Skipped details of CFL conditions and proofs of bound-preserving for the scalar conservation law and Euler equations

A.1.1 Constants in the CFL condition (2.29)

Denote

$M_1^f = \max_{m\le u\le M} |f'(u)|$,

$M_2^f = \max_{m\le u\le M} |f''(u)|$,

$M_1^g = \max_{m\le u\le M} |g'(u)|$,

$M_2^g = \max_{m\le u\le M} |g''(u)|$,

then the constants Q_1 and Q_2 in the CFL condition (2.29) are defined as:

$Q_1 = \min\{q_1^1, q_2^1, \ldots, q_6^1\}$, where

$q_1^1 = \frac{1}{8M_1^f} \min_\gamma \hat{\omega}_\gamma$,

$q_2^1 = \frac{1}{4} \frac{4\beta_1 - \frac{1}{2}}{5(M-m)M_2^f + \frac{4}{3}M_1^f}$,

$q_3^1 = \frac{1}{4} \frac{2 - 8\beta_1}{20(M-m)M_2^f + \frac{8}{3}M_1^f}$,

$q_4^1 = \frac{1}{4} \frac{\beta_0 - \frac{3}{2} + 4\beta_1}{15(M-m)M_2^f + \frac{4}{3}M_1^f}$,

$q_5^1 = \frac{1}{4} \frac{\hat{\omega}_1^{1/2}}{M_1^f(\beta_0 - 1 + 4\beta_1)^{1/2}}$,

$q_6^1 = \frac{1}{4} \frac{\hat{\omega}_{N_q}^{1/2}}{M_1^f(6 - 24\beta_1)^{1/2}}$,

$Q_2 = \min\{q_1^2, q_2^2, \ldots, q_6^2\}$, where

$q_1^2 = \frac{1}{8M_1^g} \min_\gamma \hat{\omega}_\gamma$,

$q_2^2 = \frac{1}{4} \frac{4\beta_1 - \frac{1}{2}}{5(M-m)M_2^g + \frac{4}{3}M_1^g}$,

$q_3^2 = \frac{1}{4} \frac{2 - 8\beta_1}{20(M-m)M_2^g + \frac{8}{3}M_1^g}$,

$q_4^2 = \frac{1}{4} \frac{\beta_0 - \frac{3}{2} + 4\beta_1}{15(M-m)M_2^g + \frac{4}{3}M_1^g}$,

$q_5^2 = \frac{1}{4} \frac{\hat{\omega}_1^{1/2}}{M_1^g(\beta_0 - 1 + 4\beta_1)^{1/2}}$,

$q_6^2 = \frac{1}{4}\frac{\hat{\omega}_{N_q}^{1/2}}{M_1^g(6-24\beta_1)^{1/2}}$,

Define

$c_1 = M_1^f M_1^g + Q_2(10(M-m)M_1^f M_1^g M_2^g + 5(M-m)M_1^{g^2} M_2^f + 2M_1^f M_1^{g^2})$,

$c_2 = M_1^f M_1^g + Q_1(10(M-m)M_1^f M_1^g M_2^f + 5(M-m)M_1^{f^2} M_2^g + 2M_1^g M_1^{f^2})$,

then Q_3 and Q_4 in (2.29) are defined as:

$Q_3 = \min\{q_1^3, q_2^3, q_3^3, q_4^3\}$, where

$q_1^3 = \frac{\hat{\omega}_1^2}{2\hat{\omega}_1\alpha_x^1 + 4Q_2c_1}$,

$q_2^3 = \frac{\hat{\omega}_1\hat{\omega}_{N_q}}{2\hat{\omega}_{N_q}\alpha_x^1 + 4Q_2c_1}$,

$q_3^3 = \frac{\hat{\omega}_1\alpha_y^1}{2c_2}$,

$q_4^3 = \frac{\hat{\omega}_{N_q}\alpha_y^1}{2c_2}$,

$Q_4 = \min\{q_1^4, q_2^4, q_3^4, q_4^4\}$, where

$q_1^4 = \frac{\hat{\omega}_1^2}{2\hat{\omega}_1\alpha_y^1 + 4Q_1c_2}$,

$q_2^4 = \frac{\hat{\omega}_1\hat{\omega}_{N_q}}{2\hat{\omega}_{N_q}\alpha_y^1 + 4Q_1c_2}$,

$q_3^4 = \frac{\hat{\omega}_1\alpha_x^1}{2c_1}$,

$q_4^4 = \frac{\hat{\omega}_{N_q}\alpha_x^1}{2c_1}$.

A.1.2 Coefficients in the expansion (2.30)

For convenience, we introduce the constants

$$d_1^\gamma = 2L'_{-1}(\hat{r}_\gamma), \quad d_2^\gamma = 2L'_0(\hat{r}_\gamma), \quad d_3^\gamma = 2L'_1(\hat{r}_\gamma), \quad \gamma = 1, 2, \ldots, 2N_q - 1,$$

where L_{-1}, L_0, L_1 are the Lagrange basis in (2.15) and $\{\hat{r}_\gamma, \gamma = 1, \ldots, 2N_q - 1\}$ are the Gauss-Lobatto points on $[-1, 1]$. It is clear that $|d_i^\gamma| \leq 4$, for $i = 1, 2, 3, \gamma = 1, 2, \ldots, 2N_q - 1$.

The coefficients $z_1, \ldots, z_{14,\beta}$ in the expansion (2.30) are defined as follows.

$$
\begin{aligned}
z_1 =&\lambda_x \left(\frac{1}{2}\hat{\omega}_1 \alpha_x^1 + \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \right. \\
&\left. \left(-\frac{1}{4} f' g' d_1^\beta + \frac{\Delta t}{12}(6 f' g' g'' u_y + 3 g'^2 f'' u_y) d_1^\beta + \lambda_y f' g'^2 \right) (x^-_{i-\frac{1}{2}}, \hat{y}_\beta) \right) \\
z_2 =&\lambda_x \left(\frac{1}{2}\hat{\omega}_{N_q} \alpha_x^1 + \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \right. \\
&\left. \left(-\frac{1}{4} f' g' d_2^\beta + \frac{\Delta t}{12}(6 f' g' g'' u_y + 3 g'^2 f'' u_y) d_2^\beta - 2\lambda_y f' g'^2 \right) (x^-_{i-\frac{1}{2}}, \hat{y}_\beta) \right) \\
z_3 =&\lambda_x \left(\frac{1}{2}\hat{\omega}_{2N_q-1} \alpha_x^1 + \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \right. \\
&\left. \left(-\frac{1}{4} f' g' d_3^\beta + \frac{\Delta t}{12}(6 f' g' g'' u_y + 3 g'^2 f'' u_y) d_3^\beta + \lambda_y f' g'^2 \right) (x^-_{i-\frac{1}{2}}, \hat{y}_\beta) \right) \\
z_4 =&\frac{1}{4}\hat{\omega}_1^2 - \frac{1}{2}\lambda_x \hat{\omega}_1 \alpha_x^1 + \lambda_x \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \\
&\left(-\frac{1}{4} f' g' d_1^\beta + \frac{\Delta t}{12}(6 f' g' g'' u_y + 3 g'^2 f'' u_y) d_1^\beta + \lambda_y f' g'^2 \right) (x^+_{i-\frac{1}{2}}, \hat{y}_\beta) \\
z_5 =&\frac{1}{4}\hat{\omega}_1 \hat{\omega}_{N_q} - \frac{1}{2}\lambda_x \hat{\omega}_{N_q} \alpha_x^1 + \lambda_x \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \\
&\left(-\frac{1}{4} f' g' d_2^\beta + \frac{\Delta t}{12}(6 f' g' g'' u_y + 3 g'^2 f'' u_y) d_2^\beta - 2\lambda_y f' g'^2 \right) (x^+_{i-\frac{1}{2}}, \hat{y}_\beta) \\
z_6 =&\frac{1}{4}\hat{\omega}_1 \hat{\omega}_{2N_q-1} - \frac{1}{2}\lambda_x \hat{\omega}_{2N_q-1} \alpha_x^1 + \lambda_x \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \\
&\left(-\frac{1}{4} f' g' d_3^\beta + \frac{\Delta t}{12}(6 f' g' g'' u_y + 3 g'^2 f'' u_y) d_3^\beta + \lambda_y f' g'^2 \right) (x^+_{i-\frac{1}{2}}, \hat{y}_\beta) \\
z_7 =&\frac{1}{4}\hat{\omega}_1 \hat{\omega}_{2N_q-1} - \frac{1}{2}\lambda_x \hat{\omega}_1 \alpha_x^1 + \lambda_x \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \\
&\left(\frac{1}{4} f' g' d_1^\beta - \frac{\Delta t}{12}(6 f' g' g'' u_y + 3 g'^2 f'' u_y) d_1^\beta - \lambda_y f' g'^2 \right) (x^-_{i+\frac{1}{2}}, \hat{y}_\beta)
\end{aligned}
$$

$$
\begin{aligned}
z_8 =& \frac{1}{4}\hat{\omega}_{N_q}\hat{\omega}_{2N_q-1} - \frac{1}{2}\lambda_x\hat{\omega}_{N_q}\alpha_x^1 + \lambda_x\lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \\
& \left(\frac{1}{4}f'g'd_2^\beta - \frac{\Delta t}{12}(6f'g'g''u_y + 3g'^2f''u_y)d_2^\beta + 2\lambda_y f'g'^2\right)(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) \\
z_9 =& \frac{1}{4}\hat{\omega}_{2N_q-1}\hat{\omega}_{2N_q-1} - \frac{1}{2}\lambda_x\hat{\omega}_{2N_q-1}\alpha_x^1 + \lambda_x\lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \\
& \left(\frac{1}{4}f'g'd_3^\beta - \frac{\Delta t}{12}(6f'g'g''u_y + 3g'^2f''u_y)d_3^\beta - \lambda_y f'g'^2\right)(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) \\
z_{10} =& \lambda_x\left(\frac{1}{2}\hat{\omega}_1\alpha_x^1 + \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \right. \\
& \left.\left(\frac{1}{4}f'g'd_1^\beta - \frac{\Delta t}{12}(6f'g'g''u_y + 3g'^2f''u_y)d_1^\beta - \lambda_y f'g'^2\right)(x^+_{i+\frac{1}{2}}, \hat{y}_\beta)\right) \\
z_{11} =& \lambda_x\left(\frac{1}{2}\hat{\omega}_{N_q}\alpha_x^1 + \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \right. \\
& \left.\left(\frac{1}{4}f'g'd_2^\beta - \frac{\Delta t}{12}(6f'g'g''u_y + 3g'^2f''u_y)d_2^\beta + 2\lambda_y f'g'^2\right)(x^+_{i+\frac{1}{2}}, \hat{y}_\beta)\right) \\
z_{12} =& \lambda_x\left(\frac{1}{2}\hat{\omega}_{2N_q-1}\alpha_x^1 + \lambda_y \sum_{\beta=1}^{2N_q-1} \hat{\omega}_\beta \times \right. \\
& \left.\left(\frac{1}{4}f'g'd_3^\beta - \frac{\Delta t}{12}(6f'g'g''u_y + 3g'^2f''u_y)d_3^\beta - \lambda_y f'g'^2\right)(x^+_{i+\frac{1}{2}}, \hat{y}_\beta)\right), \\
z_{13,\beta} =& \frac{1}{4}\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\alpha_x^1 \\
z_{14,\beta} =& \frac{1}{4}\hat{\omega}_1 - \frac{\lambda_x}{2}\alpha_x^1
\end{aligned}
$$

Moreover, we have the following lower bound estimates for $z_1, \ldots, z_{14,\beta}$ under the

CFL condition (2.29).

$$
\begin{aligned}
z_1 \geq & \lambda_x \left(\frac{1}{2}\hat{\omega}_1 \alpha_x^1 - \lambda_y \times \right. \\
& \left. \left(M_1^f M_1^g + Q_2(10(M-m) M_1^f M_1^g M_2^g + 5(M-m) M_1^{g2} M_2^f) + Q_2 M_1^f M_1^{g2} \right) \right) \geq 0 \\
z_2 \geq & \lambda_x \left(\frac{1}{2}\hat{\omega}_{N_q} \alpha_x^1 - \lambda_y \times \right. \\
& \left. \left(M_1^f M_1^g + Q_2(10(M-m) M_1^f M_1^g M_2^g + 5(M-m) M_1^{g2} M_2^f) + 2Q_2 M_1^f M_1^{g2} \right) \right) \geq 0 \\
z_3 \geq & \lambda_x \left(\frac{1}{2}\hat{\omega}_{2N_q-1} \alpha_x^1 - \lambda_y \times \right. \\
& \left. \left(M_1^f M_1^g + Q_2(10(M-m) M_1^f M_1^g M_2^g + 5(M-m) M_1^{g2} M_2^f) + Q_2 M_1^f M_1^{g2} \right) \right) \geq 0 \\
z_4 \geq & \frac{1}{4}\hat{\omega}_1^2 - \frac{1}{2}\lambda_x \hat{\omega}_1 \alpha_x^1 - \lambda_x Q_2 \times \\
& \left(M_1^f M_1^g + Q_2(10(M-m) M_1^f M_1^g M_2^g + 5(M-m) M_1^{g2} M_2^f) + Q_2 M_1^f M_1^{g2} \right) \geq 0 \\
z_5 \geq & \frac{1}{4}\hat{\omega}_1 \hat{\omega}_{N_q} - \frac{1}{2}\lambda_x \hat{\omega}_{N_q} \alpha_x^1 - \lambda_x Q_2 \times \\
& \left(M_1^f M_1^g + Q_2(10(M-m) M_1^f M_1^g M_2^g + 5(M-m) M_1^{g2} M_2^f) + 2Q_2 M_1^f M_1^{g2} \right) \geq 0 \\
z_6 \geq & \frac{1}{4}\hat{\omega}_1 \hat{\omega}_{2N_q-1} - \frac{1}{2}\lambda_x \hat{\omega}_{2N_q-1} \alpha_x^1 - \lambda_x Q_2 \times \\
& \left(M_1^f M_1^g + Q_2(10(M-m) M_1^f M_1^g M_2^g + 5(M-m) M_1^{g2} M_2^f) + Q_2 M_1^f M_1^{g2} \right) \geq 0
\end{aligned}
$$

$$
\begin{aligned}
z_7 \geq & \frac{1}{4}\hat{\omega}_1\hat{\omega}_{2N_q-1} - \frac{1}{2}\lambda_x\hat{\omega}_1\alpha_x^1 - \lambda_x Q_2 \times \\
& \left(M_1^f M_1^g + Q_2(10(M-m)M_1^f M_1^g M_2^g + 5(M-m)M_1^{g2}M_2^f) + Q_2 M_1^f M_1^{g2}\right) \geq 0 \\
z_8 \geq & \frac{1}{4}\hat{\omega}_{N_q}\hat{\omega}_{2N_q-1} - \frac{1}{2}\lambda_x\hat{\omega}_{N_q}\alpha_x^1 - \lambda_x Q_2 \times \\
& \left(M_1^f M_1^g + Q_2(10(M-m)M_1^f M_1^g M_2^g + 5(M-m)M_1^{g2}M_2^f) + 2Q_2 M_1^f M_1^{g2}\right) \geq 0 \\
z_9 \geq & \frac{1}{4}\hat{\omega}_{2N_q-1}^2 - \frac{1}{2}\lambda_x\hat{\omega}_{2N_q-1}\alpha_x^1 - \lambda_x Q_2 \times \\
& \left(M_1^f M_1^g + Q_2(10(M-m)M_1^f M_1^g M_2^g + 5(M-m)M_1^{g2}M_2^f) + Q_2 M_1^f M_1^{g2}\right) \geq 0 \\
z_{10} \geq & \lambda_x \left(\frac{1}{2}\hat{\omega}_1\alpha_x^1 - \lambda_y \times \right. \\
& \left.\left(M_1^f M_1^g + Q_2(10(M-m)M_1^f M_1^g M_2^g + 5(M-m)M_1^{g2}M_2^f) + Q_2 M_1^f M_1^{g2}\right)\right) \geq 0 \\
z_{11} \geq & \lambda_x \left(\frac{1}{2}\hat{\omega}_{N_q}\alpha_x^1 - \lambda_y \times \right. \\
& \left.\left(M_1^f M_1^g + Q_2(10(M-m)M_1^f M_1^g M_2^g + 5(M-m)M_1^{g2}M_2^f) + 2Q_2 M_1^f M_1^{g2}\right)\right) \geq 0 \\
z_{12} \geq & \lambda_x \left(\frac{1}{2}\hat{\omega}_{2N_q-1}\alpha_x^1 - \lambda_y \times \right. \\
& \left.\left(M_1^f M_1^g + Q_2(10(M-m)M_1^f M_1^g M_2^g + 5(M-m)M_1^{g2}M_2^f) + Q_2 M_1^f M_1^{g2}\right)\right) \geq 0
\end{aligned}
$$

and $z_{13,\beta}, z_{14,\beta} \geq 0, \forall\beta$.

A.1.3 Coefficients in the expansion (2.41)

The coefficients of the expansion (2.41) are

$$
\begin{aligned}
z_{10} =&\lambda^2\left(\frac{\hat{\gamma}}{2}(4\beta_1-\frac{1}{2})+\frac{\Delta t}{12}\hat{\gamma}(3+\gamma)\left(u_x\right)^-_{j-\frac{1}{2}}+\lambda\hat{\gamma}u^-_{j-\frac{1}{2}}\right)e^+_{j-\frac{3}{2}}\\
z_{11} =&\lambda^2\left(\frac{\hat{\gamma}}{2}(2-8\beta_1)-\frac{\Delta t}{3}\hat{\gamma}(3+\gamma)\left(u_x\right)^-_{j-\frac{1}{2}}-2\lambda\hat{\gamma}u^-_{j-\frac{1}{2}}\right)e_{j-1}\\
z_{12} =&\lambda^2\left(\frac{\hat{\gamma}}{2}(\beta_0-\frac{3}{2}+4\beta_1)+\frac{\Delta t^2}{12\lambda}\hat{\gamma}\gamma\left(u_{xx}\right)^-_{j-\frac{1}{2}}+\frac{\Delta t}{4}\hat{\gamma}(3+\gamma)\left(u_x\right)^-_{j-\frac{1}{2}}+\lambda\hat{\gamma}u^-_{j-\frac{1}{2}}\right)e^-_{j-\frac{1}{2}}\\
z_{13} =&\frac{1}{4}\hat{\omega}_1-\lambda^2\left(\frac{\hat{\gamma}}{2}(4\beta_1-\frac{1}{2})+\frac{\hat{\gamma}}{2}(\beta_0-\frac{3}{2}+4\beta_1)+\frac{\Delta t}{12}\hat{\gamma}(3+\gamma)\left(u_x\right)^-_{j+\frac{1}{2}}\right.\\
&\left.-\frac{\Delta t^2}{12\lambda}\hat{\gamma}\gamma\left(u_{xx}\right)^+_{j-\frac{1}{2}}+\frac{\Delta t}{4}\hat{\gamma}(3+\gamma)\left(u_x\right)^+_{j-\frac{1}{2}}+\lambda\hat{\gamma}u^-_{j+\frac{1}{2}}-\lambda\hat{\gamma}u^+_{j-\frac{1}{2}}\right)e^+_{j-\frac{1}{2}}\\
z_{14} =&\frac{1}{4}\hat{\omega}_{N_q}-\lambda^2\left(\frac{\hat{\gamma}}{2}(2-8\beta_1)+\frac{\hat{\gamma}}{2}(2-8\beta_1)-\frac{\Delta t}{3}\hat{\gamma}(3+\gamma)\left(u_x\right)^+_{j-\frac{1}{2}}\right.\\
&\left.-\frac{\Delta t}{3}\hat{\gamma}(3+\gamma)\left(u_x\right)^-_{j+\frac{1}{2}}-2\lambda\hat{\gamma}u^-_{j+\frac{1}{2}}+2\lambda\hat{\gamma}u^+_{j-\frac{1}{2}}\right)e_j\\
z_{15} =&\frac{1}{4}\hat{\omega}_{2N_q-1}-\lambda^2\left(\frac{\hat{\gamma}}{2}(\beta_0-\frac{3}{2}+4\beta_1)+\frac{\hat{\gamma}}{2}(4\beta_1-\frac{1}{2})+\frac{\Delta t^2}{12\lambda}\hat{\gamma}\gamma\left(u_{xx}\right)^-_{j+\frac{1}{2}}\right.\\
&\left.+\frac{\Delta t}{12}\hat{\gamma}(3+\gamma)\left(u_x\right)^+_{j-\frac{1}{2}}+\frac{\Delta t}{4}\hat{\gamma}(3+\gamma)\left(u_x\right)^-_{j+\frac{1}{2}}-\lambda\hat{\gamma}u^+_{j-\frac{1}{2}}+\lambda\hat{\gamma}u^-_{j+\frac{1}{2}}\right)e^-_{j+\frac{1}{2}}\\
z_{16} =&\lambda^2\left(\frac{\hat{\gamma}}{2}(\beta_0-\frac{3}{2}+4\beta_1)-\frac{\Delta t^2}{12\lambda}\hat{\gamma}\gamma\left(u_{xx}\right)^+_{j+\frac{1}{2}}+\frac{\Delta t}{4}\hat{\gamma}(3+\gamma)\left(u_x\right)^+_{j+\frac{1}{2}}-\lambda\hat{\gamma}u^+_{j+\frac{1}{2}}\right)e^+_{j+\frac{1}{2}}\\
z_{17} =&\lambda^2\left(\frac{\hat{\gamma}}{2}(2-8\beta_1)-\frac{\Delta t}{3}\hat{\gamma}(3+\gamma)\left(u_x\right)^+_{j+\frac{1}{2}}+2\lambda\hat{\gamma}u^+_{j+\frac{1}{2}}\right)e_{j+1}\\
z_{18} =&\lambda^2\left(\frac{\hat{\gamma}}{2}(4\beta_1-\frac{1}{2})+\frac{\Delta t}{12}\hat{\gamma}(3+\gamma)\left(u_x\right)^+_{j+\frac{1}{2}}-\lambda\hat{\gamma}u^+_{j+\frac{1}{2}}\right)e^-_{j+\frac{3}{2}}
\end{aligned}
$$

Under the condition $\lambda \leq \min\{q_7, q_8, q_9, q_{10}, q_{11}\}$, we have the estimates as follows

$$
\begin{aligned}
z_{10} \geq & \lambda^2 \left(\frac{\hat{\gamma}}{2}(4\beta_1 - \frac{1}{2}) - \frac{\Delta t}{12}\hat{\gamma}(3+\gamma)||u_x||_\infty - \lambda\hat{\gamma}||u||_\infty \right) e^+_{j-\frac{3}{2}} \geq 0, \\
z_{11} \geq & \lambda^2 \left(\frac{\hat{\gamma}}{2}(2 - 8\beta_1) - \frac{\Delta t}{3}\hat{\gamma}(3+\gamma)||u_x||_\infty - 2\lambda\hat{\gamma}||u||_\infty \right) e_{j-1} \geq 0, \\
z_{12} \geq & \lambda^2 \left(\frac{\hat{\gamma}}{2}(\beta_0 - \frac{3}{2} + 4\beta_1) - \frac{\Delta t^2}{12\lambda}\hat{\gamma}\gamma||u_{xx}||_\infty - \frac{\Delta t}{4}\hat{\gamma}(3+\gamma)||u_x||_\infty - \lambda\hat{\gamma}||u||_\infty \right) e^-_{j-\frac{1}{2}} \geq 0, \\
z_{13} \geq & \frac{1}{4}\hat{\omega}_1 - \lambda^2 \times \\
& \left(\frac{\hat{\gamma}}{2}(\beta_0 - 2 + 8\beta_1) + \frac{\Delta t^2}{12\lambda}\hat{\gamma}\gamma||u_{xx}||_\infty + \frac{\Delta t}{3}\hat{\gamma}(3+\gamma)||u_x||_\infty + 2\lambda\hat{\gamma}||u||_\infty \right) ||e||_\infty \geq 0, \\
z_{14} \geq & \frac{1}{4}\hat{\omega}_{N_q} - \lambda^2 \left(\hat{\gamma}(2 - 8\beta_1) + \frac{2\Delta t}{3}\hat{\gamma}(3+\gamma)||u_x||_\infty + 4\lambda\hat{\gamma}||u||_\infty \right) ||e||_\infty \geq 0, \\
z_{15} \geq & \frac{1}{4}\hat{\omega}_{2N_q-1} - \lambda^2 \times \\
& \left(\frac{\hat{\gamma}}{2}(\beta_0 - 2 + 8\beta_1) + \frac{\Delta t^2}{12\lambda}\hat{\gamma}\gamma||u_{xx}||_\infty + \frac{\Delta t}{3}\hat{\gamma}(3+\gamma)||u_x||_\infty + 2\lambda\hat{\gamma}||u||_\infty \right) ||e||_\infty \geq 0, \\
z_{16} \geq & \lambda^2 \left(\frac{\hat{\gamma}}{2}(\beta_0 - \frac{3}{2} + 4\beta_1) - \frac{\Delta t^2}{12\lambda}\hat{\gamma}\gamma||u_{xx}||_\infty - \frac{\Delta t}{4}\hat{\gamma}(3+\gamma)||u_x||_\infty - \lambda\hat{\gamma}||u||_\infty \right) e^+_{j+\frac{1}{2}} \geq 0, \\
z_{17} \geq & \lambda^2 \left(\frac{\hat{\gamma}}{2}(2 - 8\beta_1) - \frac{\Delta t}{3}\hat{\gamma}(3+\gamma)||u_x||_\infty - 2\lambda\hat{\gamma}||u||_\infty \right) e_{j+1} \geq 0, \\
z_{18} \geq & \lambda^2 \left(\frac{\hat{\gamma}}{2}(4\beta_1 - \frac{1}{2}) - \frac{\Delta t}{12}\hat{\gamma}(3+\gamma)||u_x||_\infty - \lambda\hat{\gamma}||u||_\infty \right) e^-_{j+\frac{3}{2}} \geq 0,
\end{aligned}
$$

A.1.4 Constants in the CFL condition (2.55)

$Q_1 = \min\{q_1^1, q_2^1, \ldots, q_{11}^1\}$, where

$q_1^1 = \frac{\hat{\omega}_1}{8||(|u|+c)||_\infty}$,

$q_2^1 = \frac{1}{4}\frac{6(\beta_0 - \frac{3}{2} + 4\beta_1)}{\Delta x^2||u_{xx}||_\infty + 6\Delta x||u_x||_\infty + 4||u||_\infty}$,

$q_3^1 = \frac{1}{4}\frac{3(2-8\beta_1)}{4(\Delta x||u_x||_\infty + ||u||_\infty)}$,

$q_4^1 = \frac{1}{4}\frac{3(4\beta_1 - \frac{1}{2})}{\Delta x||u_x||_\infty + 2||u||_\infty}$,

$q_5^1 = \frac{1}{8||u||_\infty}\left(\frac{\hat{\omega}_1}{\beta_0 - 2 + 8\beta_1}\right)^{\frac{1}{2}}$,

$q_6^1 = \frac{1}{8||u||_\infty}\left(\frac{\hat{\omega}_{N_q}}{2(2-8\beta_1)}\right)^{\frac{1}{2}}$,

$q_7^1 = \frac{1}{4}\frac{6(4\beta_1-\frac{1}{2})}{(3+\gamma)\Delta x||u_x||_\infty+\hat{\gamma}\Delta x||v_y||_\infty+12||u||_\infty}$,

$q_8^1 = \frac{1}{4}\frac{3(2-8\beta_1)}{2(3+\gamma)\Delta x||u_x||_\infty+2\hat{\gamma}\Delta x||v_y||_\infty+12||u||_\infty}$,

$q_9^1 = \frac{1}{4}\frac{6(\beta_0-\frac{3}{2}+4\beta_1)}{\gamma\Delta x^2||u_{xx}||_\infty+3(3+\gamma)\Delta x||u_x||_\infty+3\hat{\gamma}\Delta x||v_y||_\infty+12||u||_\infty}$,

$q_{10}^1 = \frac{1}{4}\left(\frac{\hat{\omega}_1}{4\hat{\gamma}(\beta_0-2+8\beta_1)||e||_\infty}\right)^{\frac{1}{2}}$,

$q_{11}^1 = \frac{1}{4}\left(\frac{\hat{\omega}_{N_q}}{8\hat{\gamma}(2-8\beta_1)||e||_\infty}\right)^{\frac{1}{2}}$,

$Q_2 = \min\{q_1^2, q_2^2, \ldots, q_{11}^2\}$, where

$q_1^2 = \frac{\hat{\omega}_1}{8||(|v|+c)||_\infty}$,

$q_2^2 = \frac{1}{4}\frac{6(\beta_0-\frac{3}{2}+4\beta_1)}{\Delta y^2||v_{yy}||_\infty+6\Delta y||v_y||_\infty+4||v||_\infty}$,

$q_3^2 = \frac{1}{4}\frac{3(2-8\beta_1)}{4(\Delta y||v_y||_\infty+||v||_\infty)}$,

$q_4^2 = \frac{1}{4}\frac{3(4\beta_1-\frac{1}{2})}{\Delta y||v_y||_\infty+2||v||_\infty}$,

$q_5^2 = \frac{1}{8||v||_\infty}\left(\frac{\hat{\omega}_1}{\beta_0-2+8\beta_1}\right)^{\frac{1}{2}}$,

$q_6^2 = \frac{1}{8||v||_\infty}\left(\frac{\hat{\omega}_{N_q}}{2(2-8\beta_1)}\right)^{\frac{1}{2}}$,

$q_7^2 = \frac{1}{4}\frac{6(4\beta_1-\frac{1}{2})}{(3+\gamma)\Delta y||v_y||_\infty+\hat{\gamma}\Delta y||u_x||_\infty+12||v||_\infty}$,

$q_8^2 = \frac{1}{4}\frac{3(2-8\beta_1)}{2(3+\gamma)\Delta y||v_y||_\infty+2\hat{\gamma}\Delta y||u_x||_\infty+12||v||_\infty}$,

$q_9^2 = \frac{1}{4}\frac{6(\beta_0-\frac{3}{2}+4\beta_1)}{\gamma\Delta y^2||v_{yy}||_\infty+3(3+\gamma)\Delta y||v_y||_\infty+3\hat{\gamma}\Delta y||u_x||_\infty+12||v||_\infty}$,

$q_{10}^2 = \frac{1}{4}\left(\frac{\hat{\omega}_1}{4\hat{\gamma}(\beta_0-2+8\beta_1)||e||_\infty}\right)^{\frac{1}{2}}$,

$q_{11}^2 = \frac{1}{4}\left(\frac{\hat{\omega}_{N_q}}{8\hat{\gamma}(2-8\beta_1)||e||_\infty}\right)^{\frac{1}{2}}$,

Let

$c_1 = 3\hat{\omega}_1\Delta x||(v_xu+vu_x)||_\infty+\hat{\omega}_1Q_1\Delta x^2||A_4||_\infty+12(||uv||_\infty+\frac{Q_1}{3}\Delta x||A_5||_\infty+\frac{Q_1}{3}||A_6||_\infty)$

$c_1' = 3\hat{\omega}_1\Delta y||(u_yv+uv_y)||_\infty+\hat{\omega}_1Q_2\Delta y^2||A_1||_\infty+12(||uv||_\infty+\frac{Q_2}{3}\Delta y||A_2||_\infty+\frac{Q_2}{3}||A_3||_\infty)$

$c_2 = 3\hat{\omega}_{N_q}\Delta x||(v_xu+vu_x)||_\infty+\hat{\omega}_{N_q}Q_1\Delta x^2||A_4||_\infty+12(||uv||_\infty+\frac{Q_1}{3}\Delta x||A_5||_\infty+\frac{2Q_1}{3}||A_6||_\infty)$

$c_2' = 3\hat{\omega}_{N_q}\Delta y||(u_yv+uv_y)||_\infty+\hat{\omega}_{N_q}Q_2\Delta y^2||A_1||_\infty+12(||uv||_\infty+\frac{Q_2}{3}\Delta y||A_2||_\infty+\frac{2Q_2}{3}||A_3||_\infty)$

$c_3 = 6\hat{\omega}_1\alpha_x^1+3\hat{\omega}_1Q_2\Delta y||(u_yv+uv_y)||_\infty+\hat{\omega}_1Q_2^2\Delta y^2||A_1||_\infty+12Q_2(||uv||_\infty+\frac{Q_2}{3}\Delta y||A_2||_\infty+\frac{Q_2}{3}||A_3||_\infty)$

$c_3' = 6\hat{\omega}_1\alpha_y^1+3\hat{\omega}_1Q_1\Delta x||(v_xu+vu_x)||_\infty+\hat{\omega}_1Q_1^2\Delta x^2||A_4||_\infty+12Q_1(||uv||_\infty+\frac{Q_1}{3}\Delta y||A_5||_\infty+$

$\frac{Q_1}{3}||A_6||_\infty)$

$c_4 = 6\hat{\omega}_{N_q}\alpha_x^1 + 3\hat{\omega}_{N_q}Q_2\Delta y||(u_y v + uv_y)||_\infty + \hat{\omega}_{N_q}Q_2^2\Delta y^2||A_1||_\infty + 12Q_2(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{2Q_2}{3}||A_3||_\infty)$

$c_4' = 6\hat{\omega}_{N_q}\alpha_y^1 + 3\hat{\omega}_{N_q}Q_1\Delta x||(v_x u + vu_x)||_\infty + \hat{\omega}_{N_q}Q_1^2\Delta x^2||A_4||_\infty + 12Q_1(||uv||_\infty + \frac{Q_1}{3}\Delta x||A_5||_\infty + \frac{2Q_1}{3}||A_6||_\infty)$,

then

$Q_3 = \min\{q_1^3, q_2^3, q_3^3, q_4^3\}$, where

$q_1^3 = \frac{6\hat{\omega}_1\alpha_y^1}{c_1}$,

$q_2^3 = \frac{6\hat{\omega}_{N_q}\alpha_y^1}{c_2}$,

$q_3^3 = \frac{3\hat{\omega}_1^2}{c_3}$,

$q_4^3 = \frac{3\hat{\omega}_1\hat{\omega}_{N_q}}{c_4}$,

$Q_4 = \min\{q_1^4, q_2^4, q_3^4, q_4^4\}$, where

$q_1^4 = \frac{6\hat{\omega}_1\alpha_x^1}{c_1'}$,

$q_2^4 = \frac{6\hat{\omega}_{N_q}\alpha_x^1}{c_2'}$,

$q_3^4 = \frac{3\hat{\omega}_1^2}{c_3'}$,

$q_4^4 = \frac{3\hat{\omega}_1\hat{\omega}_{N_q}}{c_4'}$.

A.1.5 Coefficients in the expansion (2.57)

The coefficients $z_1, \ldots, z_{16,\beta}$ in the expansion (2.57) are defined as follows.

$$
\begin{aligned}
z_1 =&\lambda_x \left(\frac{1}{2}\hat{\omega}_1 \alpha_x^1 - \hat{\omega}_1 \frac{\lambda_y}{4}\Delta y (u_y v + u v_y)(x^-_{i-\frac{1}{2}}, y^+_{j-\frac{1}{2}}) + \hat{\omega}_1 \frac{\lambda_y^2}{12}\Delta y^2 A_1 (x^-_{i-\frac{1}{2}}, y^+_{j-\frac{1}{2}}) \right. \\
&\left. + \lambda_y \sum_{\gamma=1}^{2N_q-1} \hat{\omega}_\gamma \left(-\frac{1}{4} d_1^\gamma uv + \frac{\lambda_y}{12} d_1^\gamma \Delta y A_2 + \frac{\lambda_y}{3} A_3 \right) (x^-_{i-\frac{1}{2}}, \hat{y}_\gamma) \right) \\
z_2 =&\lambda_x \left(\frac{1}{2}\hat{\omega}_{N_q} \alpha_x^1 - \hat{\omega}_{N_q} \frac{\lambda_y}{4}\Delta y (u_y v + u v_y)(x^-_{i-\frac{1}{2}}, y_j) + \hat{\omega}_{N_q} \frac{\lambda_y^2}{12}\Delta y^2 A_1 (x^-_{i-\frac{1}{2}}, y_j) \right. \\
&\left. + \lambda_y \sum_{\gamma=1}^{2N_q-1} \hat{\omega}_\gamma \left(-\frac{1}{4} d_2^\gamma uv + \frac{\lambda_y}{12} d_2^\gamma \Delta y A_2 - \frac{2\lambda_y}{3} A_3 \right) (x^-_{i-\frac{1}{2}}, \hat{y}_\gamma) \right) \\
z_3 =&\lambda_x \left(\frac{1}{2}\hat{\omega}_{2N_q-1} \alpha_x^1 - \hat{\omega}_{2N_q-1} \frac{\lambda_y}{4}\Delta y (u_y v + u v_y)(x^-_{i-\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \right. \\
&+ \hat{\omega}_{2N_q-1} \frac{\lambda_y^2}{12}\Delta y^2 A_1 (x^-_{i-\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
&\left. + \lambda_y \sum_{\gamma=1}^{2N_q-1} \hat{\omega}_\gamma \left(-\frac{1}{4} d_3^\gamma uv + \frac{\lambda_y}{12} d_3^\gamma \Delta y A_2 + \frac{\lambda_y}{3} A_3 \right) (x^-_{i-\frac{1}{2}}, \hat{y}_\gamma) \right) \\
z_4 =&\frac{1}{4}\hat{\omega}_1^2 - \frac{\lambda_x}{2}\hat{\omega}_1 \alpha_x^1 - \hat{\omega}_1 \frac{\lambda_x \lambda_y}{4}\Delta y (u_y v + u v_y)(x^+_{i-\frac{1}{2}}, y^+_{j-\frac{1}{2}}) + \hat{\omega}_1 \frac{\lambda_x \lambda_y^2}{12}\Delta y^2 A_1 (x^+_{i-\frac{1}{2}}, y^+_{j-\frac{1}{2}}) \\
&+ \lambda_x \lambda_y \sum_{\gamma=1}^{2N_q-1} \hat{\omega}_\gamma \left(-\frac{1}{4} d_1^\gamma uv + \frac{\lambda_y}{12} d_1^\gamma \Delta y A_2 + \frac{\lambda_y}{3} A_3 \right) (x^+_{i-\frac{1}{2}}, \hat{y}_\gamma) \\
z_5 =&\frac{1}{4}\hat{\omega}_1 \hat{\omega}_{N_q} - \frac{\lambda_x}{2}\hat{\omega}_{N_q} \alpha_x^1 - \hat{\omega}_{N_q} \frac{\lambda_x \lambda_y}{4}\Delta y (u_y v + u v_y)(x^+_{i-\frac{1}{2}}, y_j) + \hat{\omega}_{N_q} \frac{\lambda_x \lambda_y^2}{12}\Delta y^2 A_1 (x^+_{i-\frac{1}{2}}, y_j) \\
&+ \lambda_x \lambda_y \sum_{\gamma=1}^{2N_q-1} \hat{\omega}_\gamma \left(-\frac{1}{4} d_2^\gamma uv + \frac{\lambda_y}{12} d_2^\gamma \Delta y A_2 - \frac{2\lambda_y}{3} A_3 \right) (x^+_{i-\frac{1}{2}}, \hat{y}_\gamma) \\
z_6 =&\frac{1}{4}\hat{\omega}_1 \hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\hat{\omega}_{2N_q-1} \alpha_x^1 - \hat{\omega}_{2N_q-1} \frac{\lambda_x \lambda_y}{4}\Delta y (u_y v + u v_y)(x^+_{i-\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
&+ \hat{\omega}_{2N_q-1} \frac{\lambda_x \lambda_y^2}{12}\Delta y^2 A_1 (x^+_{i-\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
&+ \lambda_x \lambda_y \sum_{\gamma=1}^{2N_q-1} \hat{\omega}_\gamma \left(-\frac{1}{4} d_3^\gamma uv + \frac{\lambda_y}{12} d_3^\gamma \Delta y A_2 + \frac{\lambda_y}{3} A_3 \right) (x^+_{i-\frac{1}{2}}, \hat{y}_\gamma)
\end{aligned}
$$

$$
\begin{aligned}
z_7 =& \frac{1}{4}\hat{\omega}_1\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\hat{\omega}_1\alpha_x^1 + \hat{\omega}_1\frac{\lambda_x\lambda_y}{4}\Delta y(u_y v + u v_y)(x^-_{i+\frac{1}{2}}, y^+_{j-\frac{1}{2}}) \\
& - \hat{\omega}_1\frac{\lambda_x\lambda_y^2}{12}\Delta y^2 A_1(x^-_{i+\frac{1}{2}}, y^+_{j-\frac{1}{2}}) \\
& + \lambda_x\lambda_y\sum_{\gamma=1}^{2N_q-1}\hat{\omega}_\gamma\left(\frac{1}{4}d_1^\gamma uv - \frac{\lambda_y}{12}d_1^\gamma\Delta y A_2 - \frac{\lambda_y}{3}A_3\right)(x^-_{i+\frac{1}{2}}, \hat{y}_\gamma) \\
z_8 =& \frac{1}{4}\hat{\omega}_{N_q}\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\hat{\omega}_{N_q}\alpha_x^1 + \hat{\omega}_{N_q}\frac{\lambda_x\lambda_y}{4}\Delta y(u_y v + u v_y)(x^-_{i+\frac{1}{2}}, y_j) \\
& - \hat{\omega}_{N_q}\frac{\lambda_x\lambda_y^2}{12}\Delta y^2 A_1(x^-_{i+\frac{1}{2}}, y_j) \\
& + \lambda_x\lambda_y\sum_{\gamma=1}^{2N_q-1}\hat{\omega}_\gamma\left(\frac{1}{4}d_2^\gamma uv - \frac{\lambda_y}{12}d_2^\gamma\Delta y A_2 + \frac{2\lambda_y}{3}A_3\right)(x^-_{i+\frac{1}{2}}, \hat{y}_\gamma) \\
z_9 =& \frac{1}{4}\hat{\omega}_{2N_q-1}\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\hat{\omega}_{2N_q-1}\alpha_x^1 + \hat{\omega}_{2N_q-1}\frac{\lambda_x\lambda_y}{4}\Delta y(u_y v + u v_y)(x^-_{i+\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
& - \hat{\omega}_{2N_q-1}\frac{\lambda_x\lambda_y^2}{12}\Delta y^2 A_1(x^-_{i+\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
& + \lambda_x\lambda_y\sum_{\gamma=1}^{2N_q-1}\hat{\omega}_\gamma\left(\frac{1}{4}d_3^\gamma uv - \frac{\lambda_y}{12}d_3^\gamma\Delta y A_2 - \frac{\lambda_y}{3}A_3\right)(x^-_{i+\frac{1}{2}}, \hat{y}_\gamma) \\
z_{10} =& \lambda_x\left(\frac{1}{2}\hat{\omega}_1\alpha_x^1 + \hat{\omega}_1\frac{\lambda_y}{4}\Delta y(u_y v + u v_y)(x^+_{i+\frac{1}{2}}, y^+_{j-\frac{1}{2}}) - \hat{\omega}_1\frac{\lambda_y^2}{12}\Delta y^2 A_1(x^+_{i+\frac{1}{2}}, y^+_{j-\frac{1}{2}})\right. \\
& \left. + \lambda_y\sum_{\gamma=1}^{2N_q-1}\hat{\omega}_\gamma\left(\frac{1}{4}d_1^\gamma uv - \frac{\lambda_y}{12}d_1^\gamma\Delta y A_2 - \frac{\lambda_y}{3}A_3\right)(x^+_{i+\frac{1}{2}}, \hat{y}_\gamma)\right) \\
z_{11} =& \lambda_x\left(\frac{1}{2}\hat{\omega}_{N_q}\alpha_x^1 + \hat{\omega}_{N_q}\frac{\lambda_y}{4}\Delta y(u_y v + u v_y)(x^+_{i+\frac{1}{2}}, y_j) - \hat{\omega}_{N_q}\frac{\lambda_y^2}{12}\Delta y^2 A_1(x^+_{i+\frac{1}{2}}, y_j)\right. \\
& \left. + \lambda_y\sum_{\gamma=1}^{2N_q-1}\hat{\omega}_\gamma\left(\frac{1}{4}d_2^\gamma uv - \frac{\lambda_y}{12}d_2^\gamma\Delta y A_2 + \frac{2\lambda_y}{3}A_3\right)(x^+_{i+\frac{1}{2}}, \hat{y}_\gamma)\right) \\
z_{12} =& \lambda_x\left(\frac{1}{2}\hat{\omega}_{2N_q-1}\alpha_x^1 + \hat{\omega}_{2N_q-1}\frac{\lambda_y}{4}\Delta y(u_y v + u v_y)(x^+_{i+\frac{1}{2}}, y^-_{j+\frac{1}{2}})\right. \\
& - \hat{\omega}_{2N_q-1}\frac{\lambda_y^2}{12}\Delta y^2 A_1(x^+_{i+\frac{1}{2}}, y^-_{j+\frac{1}{2}}) \\
& \left. + \lambda_y\sum_{\gamma=1}^{2N_q-1}\hat{\omega}_\gamma\left(\frac{1}{4}d_3^\gamma uv - \frac{\lambda_y}{12}d_3^\gamma\Delta y A_2 - \frac{\lambda_y}{3}A_3\right)(x^+_{i+\frac{1}{2}}, \hat{y}_\gamma)\right)
\end{aligned}
$$

$$
\begin{aligned}
z_{13,\beta} =&\lambda_x \left(\frac{1}{2}\alpha_x^1 - \frac{\lambda_y}{4}\Delta y(u_y v + u v_y)(x^-_{i-\frac{1}{2}}, \hat{y}_\beta) + \frac{\lambda_y^2}{12}\Delta y^2 A_1(x^-_{i-\frac{1}{2}}, \hat{y}_\beta) \right) \\
z_{14,\beta} =&\lambda_x \left(\frac{1}{2}\alpha_x^1 + \frac{\lambda_y}{4}\Delta y(u_y v + u v_y)(x^+_{i+\frac{1}{2}}, \hat{y}_\beta) - \frac{\lambda_y^2}{12}\Delta y^2 A_1(x^+_{i+\frac{1}{2}}, \hat{y}_\beta) \right) \\
z_{15,\beta} =&\frac{1}{4}\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\alpha_x^1 + \frac{\lambda_x\lambda_y}{4}\Delta y(u_y v + u v_y)(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) - \frac{\lambda_x\lambda_y^2}{12}\Delta y^2 A_1(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) \\
z_{16,\beta} =&\frac{1}{4}\hat{\omega}_1 - \frac{\lambda_x}{2}\alpha_x^1 - \frac{\lambda_x\lambda_y}{4}\Delta y(u_y v + u v_y)(x^+_{i-\frac{1}{2}}, \hat{y}_\beta) + \frac{\lambda_x\lambda_y^2}{12}\Delta y^2 A_1(x^+_{i-\frac{1}{2}}, \hat{y}_\beta)
\end{aligned}
$$

Under the CFL condition (2.55), we have the following estimates.

$$
\begin{aligned}
z_1 \geq&\lambda_x \left(\frac{1}{2}\hat{\omega}_1\alpha_x^1 - \hat{\omega}_1\frac{\lambda_y}{4}\Delta y||(u_y v + u v_y)||_\infty - \hat{\omega}_1\frac{\lambda_y}{12}Q_2\Delta y^2||A_1||_\infty \right. \\
&\left. -\lambda_y \left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{Q_2}{3}||A_3||_\infty \right) \right) \geq 0 \\
z_2 \geq&\lambda_x \left(\frac{1}{2}\hat{\omega}_{N_q}\alpha_x^1 - \hat{\omega}_{N_q}\frac{\lambda_y}{4}\Delta y||(u_y v + u v_y)||_\infty - \hat{\omega}_{N_q}\frac{\lambda_y}{12}Q_2\Delta y^2||A_1||_\infty \right. \\
&\left. -\lambda_y \left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{2Q_2}{3}||A_3||_\infty \right) \right) \geq 0 \\
z_3 \geq&\lambda_x \left(\frac{1}{2}\hat{\omega}_{2N_q-1}\alpha_x^1 - \hat{\omega}_{2N_q-1}\frac{\lambda_y}{4}\Delta y||(u_y v + u v_y)||_\infty - \hat{\omega}_{2N_q-1}\frac{\lambda_y}{12}Q_2\Delta y^2||A_1||_\infty \right. \\
&\left. -\lambda_y \left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{Q_2}{3}||A_3||_\infty \right) \right) \geq 0 \\
z_4 \geq&\frac{1}{4}\hat{\omega}_1^2 - \frac{\lambda_x}{2}\hat{\omega}_1\alpha_x^1 - \hat{\omega}_1\frac{\lambda_x}{4}Q_2\Delta y||(u_y v + u v_y)||_\infty - \hat{\omega}_1\frac{\lambda_x}{12}Q_2^2\Delta y^2||A_1||_\infty \\
&- \lambda_x Q_2 \left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{Q_2}{3}||A_3||_\infty \right) \geq 0 \\
z_5 \geq&\frac{1}{4}\hat{\omega}_1\hat{\omega}_{N_q} - \frac{\lambda_x}{2}\hat{\omega}_{N_q}\alpha_x^1 - \hat{\omega}_{N_q}\frac{\lambda_x}{4}Q_2\Delta y||(u_y v + u v_y)||_\infty - \hat{\omega}_{N_q}\frac{\lambda_x}{12}Q_2^2\Delta y^2||A_1||_\infty \\
&- \lambda_x Q_2 \left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{2Q_2}{3}||A_3||_\infty \right) \geq 0 \\
z_6 \geq&\frac{1}{4}\hat{\omega}_1\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\hat{\omega}_{2N_q-1}\alpha_x^1 - \hat{\omega}_{2N_q-1}\frac{\lambda_x}{4}Q_2\Delta y||(u_y v + u v_y)||_\infty \\
&- \hat{\omega}_{2N_q-1}\frac{\lambda_x}{12}Q_2^2\Delta y^2||A_1||_\infty \\
&- \lambda_x Q_2 \left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{Q_2}{3}||A_3||_\infty \right) \geq 0
\end{aligned}
$$

$$
\begin{aligned}
z_7 \geq & \frac{1}{4}\hat{\omega}_1\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\hat{\omega}_1\alpha_x^1 - \hat{\omega}_1\frac{\lambda_x}{4}Q_2\Delta y||(u_y v + uv_y)||_\infty - \hat{\omega}_1\frac{\lambda_x}{12}Q_2^2\Delta y^2||A_1||_\infty \\
& - \lambda_x Q_2\left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{Q_2}{3}||A_3||_\infty\right) \geq 0 \\
z_8 \geq & \frac{1}{4}\hat{\omega}_{N_q}\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\hat{\omega}_{N_q}\alpha_x^1 - \hat{\omega}_{N_q}\frac{\lambda_x}{4}Q_2\Delta y||(u_y v + uv_y)||_\infty - \hat{\omega}_{N_q}\frac{\lambda_x}{12}Q_2^2\Delta y^2||A_1||_\infty \\
& - \lambda_x Q_2\left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{2Q_2}{3}||A_3||_\infty\right) \geq 0 \\
z_9 \geq & \frac{1}{4}\hat{\omega}_{2N_q-1}\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\hat{\omega}_{2N_q-1}\alpha_x^1 - \hat{\omega}_{2N_q-1}\frac{\lambda_x}{4}Q_2\Delta y||(u_y v + uv_y)||_\infty \\
& - \hat{\omega}_{2N_q-1}\frac{\lambda_x}{12}Q_2^2\Delta y^2||A_1||_\infty \\
& - \lambda_x Q_2\left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{Q_2}{3}||A_3||_\infty\right) \geq 0 \\
z_{10} \geq & \lambda_x\left(\frac{1}{2}\hat{\omega}_1\alpha_x^1 - \hat{\omega}_1\frac{\lambda_y}{4}\Delta y||(u_y v + uv_y)||_\infty - \hat{\omega}_1\frac{\lambda_y}{12}Q_2\Delta y^2||A_1||_\infty\right. \\
& \left. -\lambda_y\left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{Q_2}{3}||A_3||_\infty\right)\right) \geq 0 \\
z_{11} \geq & \lambda_x\left(\frac{1}{2}\hat{\omega}_{N_q}\alpha_x^1 - \hat{\omega}_{N_q}\frac{\lambda_y}{4}\Delta y||(u_y v + uv_y)||_\infty - \hat{\omega}_{N_q}\frac{\lambda_y}{12}Q_2\Delta y^2||A_1||_\infty\right. \\
& \left. -\lambda_y\left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{2Q_2}{3}||A_3||_\infty\right)\right) \geq 0 \\
z_{12} \geq & \lambda_x\left(\frac{1}{2}\hat{\omega}_{2N_q-1}\alpha_x^1 - \hat{\omega}_{2N_q-1}\frac{\lambda_y}{4}\Delta y||(u_y v + uv_y)||_\infty - \hat{\omega}_{2N_q-1}\frac{\lambda_y}{12}Q_2\Delta y^2||A_1||_\infty\right. \\
& \left. -\lambda_y\left(||uv||_\infty + \frac{Q_2}{3}\Delta y||A_2||_\infty + \frac{Q_2}{3}||A_3||_\infty\right)\right) \geq 0
\end{aligned}
$$

$$
\begin{aligned}
z_{13,\beta} \geq & \lambda_x\left(\frac{1}{2}\alpha_x^1 - \frac{\lambda_y}{4}\Delta y||(u_y v + uv_y)||_\infty - \frac{\lambda_y}{12}Q_2\Delta y^2||A_1||_\infty\right) \geq 0, \quad \forall\beta \\
z_{14,\beta} \geq & \lambda_x\left(\frac{1}{2}\alpha_x^1 - \frac{\lambda_y}{4}\Delta y||(u_y v + uv_y)||_\infty - \frac{\lambda_y}{12}Q_2\Delta y^2||A_1||_\infty\right) \geq 0, \quad \forall\beta \\
z_{15,\beta} \geq & \frac{1}{4}\hat{\omega}_{2N_q-1} - \frac{\lambda_x}{2}\alpha_x^1 - \frac{\lambda_x}{4}Q_2\Delta y||(u_y v + uv_y)||_\infty - \frac{\lambda_x}{12}Q_2^2\Delta y^2||A_1||_\infty \geq 0, \quad \forall\beta \\
z_{16,\beta} \geq & \frac{1}{4}\hat{\omega}_1 - \frac{\lambda_x}{2}\alpha_x^1 - \frac{\lambda_x}{4}Q_2\Delta y||(u_y v + uv_y)||_\infty - \frac{\lambda_x}{12}Q_2^2\Delta y^2||A_1||_\infty \geq 0, \quad \forall\beta
\end{aligned}
$$

A.1.6 Constants in the CFL condition (2.58)

The constants appearing in the CFL condition (2.58) are defined as follows.

$$
\begin{aligned}
&C(x^-_{i+\frac{1}{2}},\hat{y}_\beta) = \\
&\frac{\Delta x}{\alpha_x}\Big((2E(x^-_{i+\frac{1}{2}},\hat{y}_\beta)+p(x^-_{i+\frac{1}{2}},\hat{y}_\beta))\Big(|\tilde{f}^1(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^1(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\\
&+2\rho(x^-_{i+\frac{1}{2}},\hat{y}_\beta)\Big(|\tilde{f}^4(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^4(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\\
&+Q_1\frac{\Delta x}{\alpha_x}\Big(|\tilde{f}^1(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^1(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\Big(|\tilde{f}^4(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^4(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\\
&+\frac{1}{2}Q_1\frac{\Delta x}{\alpha_x}\Big(|\tilde{f}^2(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^2(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)^2\\
&+\frac{1}{2}Q_1\frac{\Delta x}{\alpha_x}\Big(|\tilde{f}^3(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^3(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)^2\\
&+(2|m(x^-_{i+\frac{1}{2}},\hat{y}_\beta)|+\frac{p(x^-_{i+\frac{1}{2}},\hat{y}_\beta)}{\alpha_x})\Big(|\tilde{f}^2(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^2(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\\
&+2|n(x^-_{i+\frac{1}{2}},\hat{y}_\beta)|\Big(|\tilde{f}^3(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^3(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&C(x^+_{i+\frac{1}{2}},\hat{y}_\beta) = \\
&\frac{\Delta x}{\alpha_x}\Big((2E(x^+_{i+\frac{1}{2}},\hat{y}_\beta)+p(x^+_{i+\frac{1}{2}},\hat{y}_\beta))\Big(|\tilde{f}^1(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^1(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\\
&+2\rho(x^+_{i+\frac{1}{2}},\hat{y}_\beta)\Big(|\tilde{f}^4(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^4(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\\
&+Q_1\frac{\Delta x}{\alpha_x}\Big(|\tilde{f}^1(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^1(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\Big(|\tilde{f}^4(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^4(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\\
&+\frac{1}{2}Q_1\frac{\Delta x}{\alpha_x}\Big(|\tilde{f}^2(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^2(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)^2\\
&+\frac{1}{2}Q_1\frac{\Delta x}{\alpha_x}\Big(|\tilde{f}^3(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^3(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)^2\\
&+(2|m(x^+_{i+\frac{1}{2}},\hat{y}_\beta)|+\frac{p(x^+_{i+\frac{1}{2}},\hat{y}_\beta)}{\alpha_x})\Big(|\tilde{f}^2(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^2(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\\
&+2|n(x^+_{i+\frac{1}{2}},\hat{y}_\beta)|\Big(|\tilde{f}^3(x_{i+\frac{1}{2}},\hat{y}_\beta)|+Q_1\Delta x|\check{f}^3(x_{i+\frac{1}{2}},\hat{y}_\beta)|\Big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&D(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) = \\
&\frac{\Delta x}{\alpha_y}\Big((2E(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) + p(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}))\Big(|\tilde{f}^1(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^1(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big) \\
&+2\rho(\hat{x}_\alpha, y^-_{j+\frac{1}{2}})\Big(|\tilde{f}^4(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^4(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big) \\
&+Q_2\frac{\Delta y}{\alpha_y}\Big(|\tilde{f}^1(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^1(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big)\Big(|\tilde{f}^4(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^4(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big) \\
&+\frac{1}{2}Q_2\frac{\Delta y}{\alpha_y}\Big(|\tilde{f}^2(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^2(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big)^2 \\
&+\frac{1}{2}Q_2\frac{\Delta y}{\alpha_y}\Big(|\tilde{f}^3(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^3(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big)^2 \\
&+2|m(\hat{x}_\alpha, y^-_{j+\frac{1}{2}})|\Big(|\tilde{f}^2(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^2(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big) \\
&+(2|n(\hat{x}_\alpha, y^-_{j+\frac{1}{2}})| + \frac{p(\hat{x}_\alpha, y^-_{j+\frac{1}{2}})}{\alpha_y})\Big(|\tilde{f}^3(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^3(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&D(\hat{x}_\alpha, y^+_{j+\frac{1}{2}}) = \\
&\frac{\Delta x}{\alpha_y}\Big((2E(\hat{x}_\alpha, y^+_{j+\frac{1}{2}}) + p(\hat{x}_\alpha, y^+_{j+\frac{1}{2}}))\Big(|\tilde{f}^1(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^1(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big) \\
&+2\rho(\hat{x}_\alpha, y^+_{j+\frac{1}{2}})\Big(|\tilde{f}^4(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^4(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big) \\
&+Q_2\frac{\Delta y}{\alpha_y}\Big(|\tilde{f}^1(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^1(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big)\Big(|\tilde{f}^4(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^4(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big) \\
&+\frac{1}{2}Q_2\frac{\Delta y}{\alpha_y}\Big(|\tilde{f}^2(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^2(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big)^2 \\
&+\frac{1}{2}Q_2\frac{\Delta y}{\alpha_y}\Big(|\tilde{f}^3(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^3(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big)^2 \\
&+2|m(\hat{x}_\alpha, y^+_{j+\frac{1}{2}})|\Big(|\tilde{f}^2(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^2(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big) \\
&+2(|n(\hat{x}_\alpha, y^+_{j+\frac{1}{2}})| + \frac{p(\hat{x}_\alpha, y^+_{j+\frac{1}{2}})}{\alpha_y})\Big(|\tilde{f}^3(\hat{x}_\alpha, y_{j+\frac{1}{2}})| + Q_2\Delta y|\check{f}^3(\hat{x}_\alpha, y_{j+\frac{1}{2}})|\Big)\Big)
\end{aligned}
$$

A.2 Derivatives in the Euler equations

To simplify the derivation and coding, we need to compute a lot of intermediate variables before finally obtaining m_t, m_{tt}, m_{ttt}, (and n_t, n_{tt}, n_{ttt} in $2D$), and E_t, E_{tt}, E_{ttt} to be used in the Lax-Wendroff procedure. The expressions of the intermediate and target variables are given as follows.

A.2.1 One dimensional space

$u = \frac{m}{\rho}$,

$u_x = \frac{m_x}{\rho} - \frac{u\rho_x}{\rho}$,

$u_{xx} = -\frac{2m_x\rho_x}{\rho^2} + \frac{m_{xx}}{\rho} + m(\frac{2\rho_x^2}{\rho^3} - \frac{\rho_{xx}}{\rho^2})$,

$\rho_t = -m_x$,

$m_t = -\big(\hat{\gamma}E_x + \frac{3-\gamma}{2}m_x u + \frac{3-\gamma}{2}mu_x\big)$,

$E_t = -\big(\gamma E_x u + \gamma E u_x - \frac{\hat{\gamma}}{2}m_x u^2 - \hat{\gamma}muu_x\big)$,

$u_t = \frac{m_t}{\rho} - \frac{u\rho_t}{\rho}$,

$\rho_{tx} = -m_{xx}$,

$m_{tx} = -\big(\hat{\gamma}E_{xx} + \frac{3-\gamma}{2}m_{xx}u + (3-\gamma)m_x u_x + \frac{3-\gamma}{2}mu_{xx}\big)$,

$E_{tx} = -\big(\gamma E_{xx}u + 2\gamma E_x u_x + \gamma E u_{xx} - \frac{\hat{\gamma}}{2}m_{xx}u^2 - 2\hat{\gamma}m_x uu_x - \hat{\gamma}mu_x^2 - \hat{\gamma}muu_{xx}\big)$,

$u_{tx} = \frac{m_{tx}}{\rho} - \frac{m_x\rho_t}{\rho^2} - \frac{m_t\rho_x + m\rho_{tx}}{\rho^2} + \frac{2u\rho_x\rho_t}{\rho^2}$,

$\rho_{tt} = -m_{tx}$,

$m_{tt} = -\big(\hat{\gamma}E_{tx} + \frac{3-\gamma}{2}m_{tx}u + \frac{3-\gamma}{2}m_x u_t + \frac{3-\gamma}{2}m_t u_x + \frac{3-\gamma}{2}mu_{tx}\big)$,

$E_{tt} = -\big(\gamma E_{tx}u + \gamma E_x u_t + \gamma E_t u_x + \gamma E u_{tx} - \frac{\hat{\gamma}}{2}m_{tx}u^2 - \hat{\gamma}m_x uu_t - \hat{\gamma}m_t uu_x - \hat{\gamma}mu_t u_x$
$-\hat{\gamma}muu_{tx}\big)$,

$u_{tt} = -\frac{2m_t\rho_t}{\rho^2} + \frac{m_{tt}}{\rho} + u(\frac{2\rho_t^2}{\rho^2} - \frac{\rho_{tt}}{\rho})$,

$\rho_{ttt} = -(m_{tt})_x$,

$$m_{ttt} = -\left(\hat{\gamma}E_{tt} + \tfrac{3-\gamma}{2}m_{tt}u + \tfrac{3-\gamma}{2}mu_{tt} + (3-\gamma)m_t u_t\right)_x,$$
$$E_{ttt} = -\left(\gamma E_{tt}u + \gamma E u_{tt} + 2\gamma E_t u_t - \tfrac{\hat{\gamma}}{2}m_{tt}u^2 - \hat{\gamma}m(u_t^2 + uu_{tt}) - 2\hat{\gamma}m_t u u_t\right)_x.$$

A.2.2 Two dimensional space

$$u = \tfrac{m}{\rho},$$
$$v = \tfrac{n}{\rho},$$
$$u_x = \tfrac{m_x}{\rho} - \tfrac{u\rho_x}{\rho},$$
$$u_y = \tfrac{m_y}{\rho} - \tfrac{u\rho_y}{\rho},$$
$$v_x = \tfrac{n_x}{\rho} - \tfrac{v\rho_x}{\rho},$$
$$v_y = \tfrac{n_y}{\rho} - \tfrac{v\rho_y}{\rho},$$
$$u_{xx} = -\tfrac{2m_x\rho_x}{\rho^2} + \tfrac{m_{xx}}{\rho} + m(\tfrac{2\rho_x^2}{\rho^3} - \tfrac{\rho_{xx}}{\rho^2}),$$
$$u_{yy} = -\tfrac{2m_y\rho_y}{\rho^2} + \tfrac{m_{yy}}{\rho} + m(\tfrac{2\rho_y^2}{\rho^3} - \tfrac{\rho_{yy}}{\rho^2}),$$
$$u_{xy} = -\tfrac{\rho_y m_x}{\rho^2} - \tfrac{m_y\rho_x}{\rho^2} + \tfrac{2m\rho_y\rho_x}{\rho^3} + \tfrac{m_{xy}}{\rho} - \tfrac{m\rho_{xy}}{\rho^2},$$
$$v_{xx} = -\tfrac{2n_x\rho_x}{\rho^2} + \tfrac{n_{xx}}{\rho} + n(\tfrac{2\rho_x^2}{\rho^3} - \tfrac{\rho_{xx}}{\rho^2}),$$
$$v_{yy} = -\tfrac{2n_y\rho_y}{\rho^2} + \tfrac{n_{yy}}{\rho} + n(\tfrac{2\rho_y^2}{\rho^3} - \tfrac{\rho_{yy}}{\rho^2}),$$
$$v_{xy} = -\tfrac{\rho_y n_x}{\rho^2} - \tfrac{n_y\rho_x}{\rho^2} + \tfrac{2n\rho_y\rho_x}{\rho^3} + \tfrac{n_{xy}}{\rho} - \tfrac{n\rho_{xy}}{\rho^2},$$
$$\rho_t = -m_x - n_y,$$
$$m_t = -\left(\hat{\gamma}E_x + \tfrac{3-\gamma}{2}m_x u + \tfrac{3-\gamma}{2}mu_x - \tfrac{\hat{\gamma}}{2}n_x v - \tfrac{\hat{\gamma}}{2}nv_x + m_y v + mv_y\right),$$
$$n_t = -\left(n_x u + nu_x + \hat{\gamma}E_y - \tfrac{\hat{\gamma}}{2}m_y u - \tfrac{\hat{\gamma}}{2}mu_y + \tfrac{3-\gamma}{2}n_y v + \tfrac{3-\gamma}{2}nv_y\right),$$
$$E_t = -\left(\gamma E_x u + \gamma E u_x - \tfrac{\hat{\gamma}}{2}m_x u^2 - \hat{\gamma}muu_x - \tfrac{\hat{\gamma}}{2}m_x v^2 - \hat{\gamma}mvv_x\right)$$
$$\quad -\left(\gamma E_y v + \gamma E v_y - \tfrac{\hat{\gamma}}{2}n_y u^2 - \hat{\gamma}nuu_y - \tfrac{\hat{\gamma}}{2}n_y v^2 - \hat{\gamma}nvv_y\right)$$
$$u_t = \tfrac{m_t}{\rho} - \tfrac{u\rho_t}{\rho},$$
$$v_t = \tfrac{n_t}{\rho} - \tfrac{v\rho_t}{\rho},$$
$$\rho_{tx} = -m_{xx} - n_{xy},$$
$$\rho_{ty} = -m_{xy} - n_{yy},$$

$$
\begin{aligned}
m_{tx} &= -\big(\hat{\gamma}E_{xx} + \tfrac{3-\gamma}{2}m_{xx}u + (3-\gamma)m_x u_x + \tfrac{3-\gamma}{2}mu_{xx} \\
&\quad -\tfrac{\hat{\gamma}}{2}n_{xx}v - \hat{\gamma}n_x v_x - \tfrac{\hat{\gamma}}{2}nv_{xx} + m_{xy}v + m_y v_x + m_x v_y + mv_{xy}\big) \\
m_{ty} &= -\big(\hat{\gamma}E_{xy} + \tfrac{3-\gamma}{2}m_{xy}u + \tfrac{3-\gamma}{2}m_x u_y + \tfrac{3-\gamma}{2}m_y u_x + \tfrac{3-\gamma}{2}mu_{xy} \\
&\quad -\tfrac{\hat{\gamma}}{2}n_{xy}v - \tfrac{\hat{\gamma}}{2}n_x v_y - \tfrac{\hat{\gamma}}{2}n_y v_x - \tfrac{\hat{\gamma}}{2}nv_{xy} + m_{yy}v + 2m_y v_y + mv_{yy}\big) \\
n_{tx} &= -\big(n_{xx}u + 2n_x u_x + nu_{xx} + \hat{\gamma}E_{xy} - \tfrac{\hat{\gamma}}{2}m_{xy}u - \tfrac{\hat{\gamma}}{2}m_y u_x \\
&\quad -\tfrac{\hat{\gamma}}{2}m_x u_y - \tfrac{\hat{\gamma}}{2}mu_{xy} + \tfrac{3-\gamma}{2}n_{xy}v + \tfrac{3-\gamma}{2}n_y v_x + \tfrac{3-\gamma}{2}n_x v_y + \tfrac{3-\gamma}{2}nv_{xy}\big) \\
n_{ty} &= -\big(n_{xy}u + n_x u_y + n_y u_x + nu_{xy} + \hat{\gamma}E_{yy} - \tfrac{\hat{\gamma}}{2}m_{yy}u - \hat{\gamma}m_y u_y - \tfrac{\hat{\gamma}}{2}mu_{yy} \\
&\quad +\tfrac{3-\gamma}{2}n_{yy}v + (3-\gamma)n_y v_y + \tfrac{3-\gamma}{2}nv_{yy}\big) \\
E_{tx} &= -(\gamma E_{xx}u + 2\gamma E_x u_x + \gamma Eu_{xx} - \tfrac{\hat{\gamma}}{2}m_{xx}u^2 - 2\hat{\gamma}m_x uu_x - \hat{\gamma}mu_x^2 \\
&\quad -\hat{\gamma}muu_{xx} - \tfrac{\hat{\gamma}}{2}m_{xx}v^2 - 2\hat{\gamma}m_x vv_x \\
&\quad -\hat{\gamma}mv_x^2 - \hat{\gamma}mvv_{xx} + \gamma E_{xy}v + \gamma E_y v_x + \gamma E_x v_y + \gamma Ev_{xy} - \tfrac{\hat{\gamma}}{2}n_{xy}u^2 - \hat{\gamma}n_y uu_x - \hat{\gamma}n_x uu_y \\
&\quad -\hat{\gamma}nu_x u_y - \hat{\gamma}nuu_{xy} - \tfrac{\hat{\gamma}}{2}n_{xy}v^2 - \hat{\gamma}n_y vv_x - \hat{\gamma}n_x vv_y - \hat{\gamma}nv_x v_y - \hat{\gamma}nvv_{xy}\big) \\
E_{ty} &= -(\gamma E_{yy}v + 2\gamma E_y v_y + \gamma Ev_{yy} - \tfrac{\hat{\gamma}}{2}n_{yy}v^2 - 2\hat{\gamma}n_y vv_y - \hat{\gamma}nv_y^2 \\
&\quad -\hat{\gamma}nvv_{yy} - \tfrac{\hat{\gamma}}{2}n_{yy}u^2 - 2\hat{\gamma}n_y uu_y \\
&\quad -\hat{\gamma}nu_y^2 - \hat{\gamma}nuu_{yy} + \gamma E_{xy}u + \gamma E_x u_y + \gamma E_y u_x + \gamma Eu_{xy} - \tfrac{\hat{\gamma}}{2}m_{xy}v^2 - \hat{\gamma}m_x vv_y - \hat{\gamma}m_y vv_x \\
&\quad -\hat{\gamma}mv_y v_x - \hat{\gamma}mvv_{xy} - \tfrac{\hat{\gamma}}{2}m_{xy}u^2 - \hat{\gamma}m_x uu_y - \hat{\gamma}m_y uu_x - \hat{\gamma}mu_y u_x - \hat{\gamma}muu_{xy}\big) \\
u_{tx} &= \tfrac{m_{tx}}{\rho} - \tfrac{m_x\rho_t}{\rho^2} - \tfrac{m_t\rho_x + m\rho_{tx}}{\rho^2} + \tfrac{2u\rho_x\rho_t}{\rho^2}, \\
u_{ty} &= \tfrac{m_{ty}}{\rho} - \tfrac{m_y\rho_t}{\rho^2} - \tfrac{m_t\rho_y + m\rho_{ty}}{\rho^2} + \tfrac{2u\rho_y\rho_t}{\rho^2}, \\
v_{tx} &= \tfrac{n_{tx}}{\rho} - \tfrac{n_x\rho_t}{\rho^2} - \tfrac{n_t\rho_x + n\rho_{tx}}{\rho^2} + \tfrac{2v\rho_x\rho_t}{\rho^2}, \\
v_{ty} &= \tfrac{n_{ty}}{\rho} - \tfrac{n_y\rho_t}{\rho^2} - \tfrac{n_t\rho_y + n\rho_{ty}}{\rho^2} + \tfrac{2v\rho_y\rho_t}{\rho^2}, \\
\rho_{tt} &= -m_{tx} - n_{ty}, \\
m_{tt} &= -\big(\hat{\gamma}E_{tx} + \tfrac{3-\gamma}{2}m_{tx}u + \tfrac{3-\gamma}{2}m_x u_t + \tfrac{3-\gamma}{2}m_t u_x + \tfrac{3-\gamma}{2}mu_{tx} \\
&\quad -\tfrac{\hat{\gamma}}{2}n_{tx}v - \tfrac{\hat{\gamma}}{2}n_x v_t - \tfrac{\hat{\gamma}}{2}n_t v_x - \tfrac{\hat{\gamma}}{2}nv_{tx} + m_{ty}v + m_y v_t + m_t v_y + mv_{ty}\big) \\
n_{tt} &= -\big(n_{tx}u + n_x u_t + n_t u_x + nu_{tx} + \hat{\gamma}E_{ty} - \tfrac{\hat{\gamma}}{2}m_{ty}u - \tfrac{\hat{\gamma}}{2}m_y u_t - \tfrac{\hat{\gamma}}{2}m_t u_y - \tfrac{\hat{\gamma}}{2}mu_{ty} \\
&\quad +\tfrac{3-\gamma}{2}n_{ty}v + \tfrac{3-\gamma}{2}n_y v_t + \tfrac{3-\gamma}{2}n_t v_y + \tfrac{3-\gamma}{2}nv_{ty}\big) \\
E_{tt} &= -\big(\gamma E_{tx}u + \gamma E_x u_t + \gamma E_t u_x + \gamma Eu_{tx} - \tfrac{\hat{\gamma}}{2}m_{tx}u^2 - \hat{\gamma}m_x uu_t - \hat{\gamma}m_t uu_x - \hat{\gamma}mu_t u_x \\
&\quad -\hat{\gamma}muu_{tx} - \tfrac{\hat{\gamma}}{2}m_{tx}v^2 - \hat{\gamma}m_x vv_t - \hat{\gamma}m_t vv_x - \hat{\gamma}mv_t v_x - \hat{\gamma}mvv_{tx} + \gamma E_{ty}v + \gamma E_y v_t
\end{aligned}
$$

$$
\begin{aligned}
&+\gamma E_t v_y+\gamma E v_{ty}-\tfrac{\hat{\gamma}}{2}n_{ty}u^2-\hat{\gamma}n_y u u_t-\hat{\gamma}n_t u u_y-\hat{\gamma}n u_t u_y-\hat{\gamma}n u u_{ty}-\tfrac{\hat{\gamma}}{2}n_{ty}v^2\\
&-\hat{\gamma}n_y v v_t-\hat{\gamma}n_t v v_y-\hat{\gamma}n v_t v_y-\hat{\gamma}n v v_{ty})\\
u_{tt}=&-\tfrac{2m_t\rho_t}{\rho^2}+\tfrac{m_{tt}}{\rho}+u(\tfrac{2\rho_t^2}{\rho^2}-\tfrac{\rho_{tt}}{\rho}),\\
v_{tt}=&-\tfrac{2n_t\rho_t}{\rho^2}+\tfrac{n_{tt}}{\rho}+v(\tfrac{2\rho_t^2}{\rho^2}-\tfrac{\rho_{tt}}{\rho}),\\
\rho_{ttt}=&-(m_{tt})_x-(n_{tt})_y,\\
m_{ttt}=&-\big(\hat{\gamma}E_{tt}+\tfrac{3-\gamma}{2}m_{tt}u+\tfrac{3-\gamma}{2}mu_{tt}+(3-\gamma)m_t u_t-\tfrac{\hat{\gamma}}{2}n_{tt}v-\tfrac{\hat{\gamma}}{2}nv_{tt}-\hat{\gamma}n_t v_t\big)_x\\
&-(m_{tt}v+mv_{tt}+2m_t v_t)_y\\
n_{ttt}=&-(nu_{tt}+n_{tt}u+2n_t u_t)_x\\
&-\big(\hat{\gamma}E_{tt}-\tfrac{\hat{\gamma}}{2}m_{tt}u-\tfrac{\hat{\gamma}}{2}mu_{tt}-\hat{\gamma}m_t u_t+\tfrac{3-\gamma}{2}n_{tt}v+\tfrac{3-\gamma}{2}nv_{tt}+(3-\gamma)n_t v_t\big)_y\\
E_{ttt}=&-\big(\gamma E_{tt}u+\gamma E u_{tt}+2\gamma E_t u_t-\tfrac{\hat{\gamma}}{2}m_{tt}u^2-\hat{\gamma}m(u_t^2+uu_{tt})-2\hat{\gamma}m_t u u_t\\
&-\tfrac{\hat{\gamma}}{2}m_{tt}v^2-\hat{\gamma}m(v_t^2+vv_{tt})-2\hat{\gamma}m_t v v_t\big)_x\\
&-\big(\gamma E_{tt}v+\gamma E v_{tt}+2\gamma E_t v_t-\tfrac{\hat{\gamma}}{2}n_{tt}u^2-\hat{\gamma}n(u_t^2+uu_{tt})-2\hat{\gamma}n_t u u_t\\
&-\tfrac{\hat{\gamma}}{2}n_{tt}v^2-\hat{\gamma}n(v_t^2+vv_{tt})-2\hat{\gamma}n_t v v_t\big)_y
\end{aligned}
$$

A.3 Maximum-principle-satisfying LWDG schemes for scalar conservation laws in one dimension on nonuniform meshes

We have discussed the bound-preserving LWDG schemes on uniform meshes in Sections 2.2 and 2.3. In this appendix, we show how to extend the technique to nonuniform meshes. For simplicity, we only consider the scalar conservation law in one space dimension, but the same methodology can be adopted to construct bound-preserving schemes for the Euler equations and multi-dimensional spaces.

We first introduce a direct extension of the maximum-principle-satisfying LWDG from uniform meshes, which is simple and efficient but has constraints on mesh sizes, i.e. $\frac{1}{2} < \frac{\Delta x_{j+1}}{\Delta x_j} < 2, \forall j$. Another way of extension is based on the composite Gauss-Lobatto rule, as used in [11], which removes the constraints on meshes but is less efficient. In practice, we recommend to combine both in the way that the composite Gauss-Lobatto rule is only used on the cells where it is necessary, i.e. the cells that violate $\frac{1}{2} < \frac{\Delta x_{j+1}}{\Delta x_j} < 2$.

A.3.1 A direct extension of the maximum-principle-satisfying LWDG scheme from uniform meshes

We define the DDG flux on nonuniform meshes as

$$\widehat{u}_{x_{j+\frac{1}{2}}}^{\mathrm{DDG}} = \beta_{0,j+\frac{1}{2}} \frac{[u]_{j+\frac{1}{2}}}{\Delta x_{j+\frac{1}{2}}} + \{u_x\}_{j+\frac{1}{2}} + \beta_{1,j+\frac{1}{2}} \Delta x_{j+\frac{1}{2}} [u_{xx}]_{j+\frac{1}{2}} \tag{A.1}$$

where $\Delta x_{j+\frac{1}{2}} = \min\{\Delta x_j, \Delta x_{j+1}\}$ and $\beta_{0,j+\frac{1}{2}}, \beta_{1,j+\frac{1}{2}}, j = 1, 2, \ldots, N$ are penalty parameters satisfying (A.2) for the purpose of maximum-principle-preserving.

$$\begin{aligned} &\frac{1}{8}\max\{\frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}}, \frac{\Delta x_{j+1}}{\Delta x_{j+\frac{1}{2}}}\} < \beta_{1,j+\frac{1}{2}} < \frac{1}{4}\min\{\frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}}, \frac{\Delta x_{j+1}}{\Delta x_{j+\frac{1}{2}}}\}, \quad \forall j, \\ &\beta_{0,j+\frac{1}{2}} > \max\{\frac{3}{2}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j} - 4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}^2}{\Delta x_j^2}, \frac{3}{2}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}} - 4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}^2}{\Delta x_{j+1}^2}\}, \quad \forall j \end{aligned} \tag{A.2}$$

Note that to make sense of (A.2), the nonuniform meshes must have a mild change in mesh size, i.e. $\frac{1}{2} < \frac{\Delta x_{j+1}}{\Delta x_j} < 2, \forall j$.

Similar to (2.13), we have the expansion of the DDG flux on nonuniform meshes.

Lemma A.3.1. *For $u \in V$, the DDG flux $\widehat{u}_{x_{j+\frac{1}{2}}}^{DDG}$ defined in* (A.1) *can be expanded*

on nonuniform meshes as

$$\begin{aligned}\widehat{u}_{x_{j+\frac{1}{2}}}^{DDG} =&(\frac{1}{2\Delta x_j}-\frac{4\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}}{\Delta x_j^2})u^+_{j-\frac{1}{2}}+(-\frac{2}{\Delta x_j}+\frac{8\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}}{\Delta x_j^2})u_j\\&+(-\frac{\beta_{0,j+\frac{1}{2}}}{\Delta x_{j+\frac{1}{2}}}+\frac{3}{2\Delta x_j}-\frac{4\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}}{\Delta x_j^2})u^-_{j+\frac{1}{2}}\\&+(\frac{\beta_{0,j+\frac{1}{2}}}{\Delta x_{j+\frac{1}{2}}}-\frac{3}{2\Delta x_{j+1}}+\frac{4\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2})u^+_{j+\frac{1}{2}}\\&+(\frac{2}{\Delta x_{j+1}}-\frac{8\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2})u_{j+1}+(-\frac{1}{2\Delta x_{j+1}}+\frac{4\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2})u^-_{j+\frac{3}{2}}\end{aligned} \tag{A.3}$$

The proof follows from direct computation and the fact that u is piecewise quadratic.

We now state the main result.

Theorem A.3.2. *Given $m \le u^n \le M$ and the DDG flux* (A.1) *with parameters* (A.2)*, the cell averages $\bar{u}_j^{n+1}, j = 1, 2, \ldots, N$ of the solution of scheme* (2.8) *are bounded between m and M under the CFL condition* (A.4)*:*

$$\Delta t \le \min\{q_1, q_2, \ldots, q_{10}\}, \tag{A.4}$$

where $q_1 = \frac{\hat{\omega}_1}{2M_1}\min_j \Delta x_j$, $q_2 = \min_j\{\frac{4\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}-\frac{1}{2}\Delta x_j}{5(M-m)M_2+\frac{4}{3}M_1}\}$, $q_3 = \min_j\{\frac{4\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}-\frac{1}{2}\Delta x_{j+1}}{5(M-m)M_2+\frac{4}{3}M_1}\}$,

$q_4 = \min_j\{\frac{2\Delta x_j-8\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}}{20(M-m)M_2+\frac{8}{3}M_1}\}$, $q_5 = \min_j\{\frac{2\Delta x_{j+1}-8\beta_{1,j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}}{20(M-m)M_2+\frac{8}{3}M_1}\}$,

$q_6 = \min_j\{\frac{\beta_{0,j+\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}}-\frac{3}{2}+4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}}{15(M-m)M_2+\frac{4}{3}M_1}\Delta x_j\}$,

$q_7 = \min_j\{\frac{\beta_{0,j+\frac{1}{2}}\frac{\Delta x_{j+1}}{\Delta x_{j+\frac{1}{2}}}-\frac{3}{2}+4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}}}{15(M-m)M_2+\frac{4}{3}M_1}\Delta x_{j+1}\}$,

$q_8 = \frac{1}{M_1}\min_j\left(\frac{2\hat{\omega}_1\Delta x_j^2}{3(\beta_{0,j-\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j-\frac{1}{2}}}-\frac{3}{2}+4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j})+3(4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}-\frac{1}{2})}\right)^{\frac{1}{2}}$,

$$q_9 = \frac{1}{M_1}\min_j\left(\frac{2\hat{\omega}_{N_q}\Delta x_j^2}{3(2-8\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j})+3(2-8\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j})}\right)^{\frac{1}{2}},$$

$$q_{10} = \frac{1}{M_1}\min_j\left(\frac{2\hat{\omega}_{2N_q-1}\Delta x_j^2}{3(\beta_{0,j+\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}}-\frac{3}{2}+4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j})+3(4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j}-\frac{1}{2})}\right)^{\frac{1}{2}}.$$

Proof. We have exactly the same results as in (2.17), (2.18), and (2.19), except that the coefficients in (2.19) are now

$$\begin{aligned}
z_1 =& \frac{\lambda_j^2}{4}{f'^2}^-_{j-\frac{1}{2}}\left((4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2}-\frac{1}{2}\frac{\Delta x_j}{\Delta x_{j-1}})+\Delta t{f''}^-_{j-\frac{1}{2}}u^-_{x_{j-\frac{1}{2}}}\frac{\Delta x_j}{\Delta x_{j-1}}\right.\\
&\left.+\frac{4\lambda_j}{3}{f'}^-_{j-\frac{1}{2}}\frac{\Delta x_j^2}{\Delta x_{j-1}^2}\right)+\frac{\lambda_j^2}{4}{f'^2}^+_{j-\frac{1}{2}}(4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2}-\frac{1}{2}\frac{\Delta x_j}{\Delta x_{j-1}}),\\
z_2 =& \frac{\lambda_j^2}{4}{f'^2}^-_{j-\frac{1}{2}}\left((2\frac{\Delta x_j}{\Delta x_{j-1}}-8\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2})-4\Delta t{f''}^-_{j-\frac{1}{2}}u^-_{x_{j-\frac{1}{2}}}\frac{\Delta x_j}{\Delta x_{j-1}}\right.\\
&\left.-\frac{8\lambda_j}{3}{f'}^-_{j-\frac{1}{2}}\frac{\Delta x_j^2}{\Delta x_{j-1}^2}\right)+\frac{\lambda_j^2}{4}{f'^2}^+_{j-\frac{1}{2}}(2\frac{\Delta x_j}{\Delta x_{j-1}}-8\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2})\\
z_3 =& \frac{\lambda_j^2}{4}{f'^2}^-_{j-\frac{1}{2}}\left((\beta_{0,j-\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j-\frac{1}{2}}}-\frac{3}{2}\frac{\Delta x_j}{\Delta x_{j-1}}+4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2})\right.\\
&\left.+3\Delta t{f''}^-_{j-\frac{1}{2}}u^-_{x_{j-\frac{1}{2}}}\frac{\Delta x_j}{\Delta x_{j-1}}+\frac{4\lambda_j}{3}{f'}^-_{j-\frac{1}{2}}\frac{\Delta x_j^2}{\Delta x_{j-1}^2}\right)\\
&+\frac{\lambda_j^2}{4}{f'^2}^+_{j-\frac{1}{2}}(\beta_{0,j-\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j-\frac{1}{2}}}-\frac{3}{2}\frac{\Delta x_j}{\Delta x_{j-1}}+4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2})\\
z_4 =& \frac{1}{2}\hat{\omega}_1-\frac{\lambda_j^2}{4}{f'^2}^-_{j-\frac{1}{2}}(\beta_{0,j-\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j-\frac{1}{2}}}-\frac{3}{2}+4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j})\\
&-\frac{\lambda_j^2}{4}{f'^2}^+_{j-\frac{1}{2}}\left((\beta_{0,j-\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j-\frac{1}{2}}}-\frac{3}{2}+4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j})+3\Delta t{f''}^+_{j-\frac{1}{2}}u^+_{x_{j-\frac{1}{2}}}-\frac{4\lambda_j}{3}{f'}^+_{j-\frac{1}{2}}\right)\\
&-\frac{\lambda_j^2}{4}{f'^2}^-_{j+\frac{1}{2}}\left((4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}-\frac{1}{2})+\Delta t{f''}^-_{j+\frac{1}{2}}u^-_{x_{j+\frac{1}{2}}}+\frac{4\lambda_j}{3}{f'}^-_{j+\frac{1}{2}}\right)\\
&-\frac{\lambda_j^2}{4}{f'^2}^+_{j+\frac{1}{2}}(4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}-\frac{1}{2})
\end{aligned}$$

$$
\begin{aligned}
z_5 =& \frac{1}{2}\hat{\omega}_N - \frac{\lambda_j^2}{4} f'^{2-}_{j-\frac{1}{2}} (2 - 8\beta_{1,j-\frac{1}{2}} \frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j}) \\
& - \frac{\lambda_j^2}{4} f'^{2+}_{j-\frac{1}{2}} \left((2 - 8\beta_{1,j-\frac{1}{2}} \frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j}) - 4\Delta t f''^{+}_{j-\frac{1}{2}} u_{x^+_{j-\frac{1}{2}}} + \frac{8\lambda_j}{3} f'^{+}_{j-\frac{1}{2}} \right) \\
& - \frac{\lambda_j^2}{4} f'^{2-}_{j+\frac{1}{2}} \left((2 - 8\beta_{1,j+\frac{1}{2}} \frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}) - 4\Delta t f''^{-}_{j+\frac{1}{2}} u_{x^-_{j+\frac{1}{2}}} - \frac{8\lambda_j}{3} f'^{-}_{j+\frac{1}{2}} \right) \\
& - \frac{\lambda_j^2}{4} f'^{2+}_{j+\frac{1}{2}} (2 - 8\beta_{1,j+\frac{1}{2}} \frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}) \\
z_6 =& \frac{1}{2}\hat{\omega}_{2N_q-1} - \frac{\lambda_j^2}{4} f'^{2-}_{j-\frac{1}{2}} (4\beta_{1,j-\frac{1}{2}} \frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j} - \frac{1}{2}) \\
& - \frac{\lambda_j^2}{4} f'^{2+}_{j-\frac{1}{2}} \left((4\beta_{1,j-\frac{1}{2}} \frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j} - \frac{1}{2}) + \Delta t f''^{+}_{j-\frac{1}{2}} u_{x^+_{j-\frac{1}{2}}} - \frac{4\lambda_j}{3} f'^{+}_{j-\frac{1}{2}} \right) \\
& - \frac{\lambda_j^2}{4} f'^{2-}_{j+\frac{1}{2}} \left((\beta_{0,j+\frac{1}{2}} \frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}} - \frac{3}{2} + 4\beta_{1,j+\frac{1}{2}} \frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}) + 3\Delta t f''^{-}_{j+\frac{1}{2}} u_{x^-_{j+\frac{1}{2}}} + \frac{4\lambda_j}{3} f'^{-}_{j+\frac{1}{2}} \right) \\
& - \frac{\lambda_j^2}{4} f'^{2+}_{j+\frac{1}{2}} (\beta_{0,j+\frac{1}{2}} \frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}} - \frac{3}{2} + 4\beta_{1,j+\frac{1}{2}} \frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}) \\
z_7 =& \frac{\lambda_j^2}{4} f'^{2-}_{j+\frac{1}{2}} (\beta_{0,j+\frac{1}{2}} \frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}} - \frac{3}{2} \frac{\Delta x_j}{\Delta x_{j+1}} + 4\beta_{1,j+\frac{1}{2}} \frac{\Delta x_j \Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2}) \\
& + \frac{\lambda_j^2}{4} f'^{2+}_{j+\frac{1}{2}} \left((\beta_{0,j+\frac{1}{2}} \frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}} - \frac{3}{2} \frac{\Delta x_j}{\Delta x_{j+1}} + 4\beta_{1,j+\frac{1}{2}} \frac{\Delta x_j \Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2}) \right. \\
& \left. + 3\Delta t f''^{+}_{j+\frac{1}{2}} u_{x^+_{j+\frac{1}{2}}} \frac{\Delta x_j}{\Delta x_{j+1}} - \frac{4\lambda_j}{3} f'^{+}_{j+\frac{1}{2}} \frac{\Delta x_j^2}{\Delta x_{j+1}^2} \right) \\
z_8 =& \frac{\lambda_j^2}{4} f'^{2-}_{j+\frac{1}{2}} (2\frac{\Delta x_j}{\Delta x_{j+1}} - 8\beta_{1,j+\frac{1}{2}} \frac{\Delta x_j \Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2}) \\
& + \frac{\lambda_j^2}{4} f'^{2+}_{j+\frac{1}{2}} \left((2\frac{\Delta x_j}{\Delta x_{j+1}} - 8\beta_{1,j+\frac{1}{2}} \frac{\Delta x_j \Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2}) - 4\Delta t f''^{+}_{j+\frac{1}{2}} u_{x^+_{j+\frac{1}{2}}} \frac{\Delta x_j}{\Delta x_{j+1}} \right. \\
& \left. + \frac{8\lambda_j}{3} f'^{+}_{j+\frac{1}{2}} \frac{\Delta x_j^2}{\Delta x_{j+1}^2} \right) \\
z_9 =& \frac{\lambda_j^2}{4} f'^{2-}_{j+\frac{1}{2}} (4\beta_{1,j+\frac{1}{2}} \frac{\Delta x_j \Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2} - \frac{1}{2} \frac{\Delta x_j}{\Delta x_{j+1}}) \\
& + \frac{\lambda_j^2}{4} f'^{2+}_{j+\frac{1}{2}} \left((4\beta_{1,j+\frac{1}{2}} \frac{\Delta x_j \Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2} - \frac{1}{2} \frac{\Delta x_j}{\Delta x_{j+1}}) + \Delta t f''^{+}_{j+\frac{1}{2}} u_{x^+_{j+\frac{1}{2}}} \frac{\Delta x_j}{\Delta x_{j+1}} \right. \\
& \left. - \frac{4\lambda_j}{3} f'^{+}_{j+\frac{1}{2}} \frac{\Delta x_j^2}{\Delta x_{j+1}^2} \right)
\end{aligned}
$$

It can be verified that

$$\frac{1}{2}\sum_{\gamma=2}^{N_q-1}\hat{\omega}_\gamma+\frac{1}{2}\sum_{\gamma=N_q+1}^{2N_q-2}\hat{\omega}_\gamma+z_1+z_2+\ldots+z_9=\frac{1}{2}$$

and

$$\begin{aligned}
z_1 \geq & \frac{\lambda_j^2}{4}f'^{2-}_{j-\frac{1}{2}}\left((4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2}-\frac{1}{2}\frac{\Delta x_j}{\Delta x_{j-1}})-5(M-m)M_2\frac{\Delta t}{\Delta x_{j-1}}\frac{\Delta x_j}{\Delta x_{j-1}}\right.\\
& \left.-\frac{4\lambda_j}{3}M_1\frac{\Delta x_j^2}{\Delta x_{j-1}^2}\right)+\frac{\lambda_j^2}{4}f'^{2+}_{j-\frac{1}{2}}(4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2}-\frac{1}{2}\frac{\Delta x_j}{\Delta x_{j-1}})\geq 0,\\
z_2 \geq & \frac{\lambda_j^2}{4}f'^{2-}_{j-\frac{1}{2}}\left((2\frac{\Delta x_j}{\Delta x_{j-1}}-8\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2})-20(M-m)M_2\frac{\Delta t}{\Delta x_{j-1}}\frac{\Delta x_j}{\Delta x_{j-1}}\right.\\
& \left.-\frac{8\lambda_j}{3}M_1\frac{\Delta x_j^2}{\Delta x_{j-1}^2}\right)+\frac{\lambda_j^2}{4}f'^{2+}_{j-\frac{1}{2}}(2\frac{\Delta x_j}{\Delta x_{j-1}}-8\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2})\geq 0,\\
z_3 \geq & \frac{\lambda_j^2}{4}f'^{2-}_{j-\frac{1}{2}}\left((\beta_{0,j-\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j-\frac{1}{2}}}-\frac{3}{2}\frac{\Delta x_j}{\Delta x_{j-1}}+4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2})\right.\\
& \left.-15(M-m)M_2\frac{\Delta t}{\Delta x_{j-1}}\frac{\Delta x_j}{\Delta x_{j-1}}-\frac{4\lambda_j}{3}M_1\frac{\Delta x_j^2}{\Delta x_{j-1}^2}\right)\\
& +\frac{\lambda_j^2}{4}f'^{2+}_{j-\frac{1}{2}}(\beta_{0,j-\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j-\frac{1}{2}}}-\frac{3}{2}\frac{\Delta x_j}{\Delta x_{j-1}}+4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}\Delta x_j}{\Delta x_{j-1}^2})\geq 0,\\
z_4 \geq & \frac{1}{2}\hat{\omega}_1-\frac{\lambda_j^2}{4}M_1^2(\beta_{0,j-\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j-\frac{1}{2}}}-\frac{3}{2}+4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j})\\
& -\frac{\lambda_j^2}{4}M_1^2\left((\beta_{0,j-\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j-\frac{1}{2}}}-\frac{3}{2}+4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j})+15(M-m)M_2\lambda_j+\frac{4\lambda_j}{3}M_1\right)\\
& -\frac{\lambda_j^2}{4}M_1^2\left((4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}-\frac{1}{2})+5(M-m)M_2\lambda_j+\frac{4\lambda_j}{3}M_1\right)\\
& -\frac{\lambda_j^2}{4}M_1^2(4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}-\frac{1}{2})\geq 0,
\end{aligned}$$

$$
\begin{aligned}
z_5 \geq & \frac{1}{2}\hat{\omega}_N - \frac{\lambda_j^2}{4}M_1^2(2 - 8\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j}) \\
& - \frac{\lambda_j^2}{4}M_1^2\left((2 - 8\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j}) + 20(M-m)M_2\lambda_j + \frac{8\lambda_j}{3}M_1\right) \\
& - \frac{\lambda_j^2}{4}M_1^2\left((2 - 8\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}) + 20(M-m)M_2\lambda_j + \frac{8\lambda_j}{3}M_1\right) \\
& - \frac{\lambda_j^2}{4}M_1^2(2 - 8\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}) \geq 0, \\
z_6 \geq & \frac{1}{2}\hat{\omega}_{2N_q-1} - \frac{\lambda_j^2}{4}M_1^2(4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j} - \frac{1}{2}) \\
& - \frac{\lambda_j^2}{4}M_1^2\left((4\beta_{1,j-\frac{1}{2}}\frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j} - \frac{1}{2}) + 5(M-m)M_2\lambda_j + \frac{4\lambda_j}{3}M_1\right) \\
& - \frac{\lambda_j^2}{4}M_1^2\left((\beta_{0,j+\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}} - \frac{3}{2} + 4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}) + 15(M-m)M_2\lambda_j + \frac{4\lambda_j}{3}M_1\right) \\
& - \frac{\lambda_j^2}{4}M_1^2(\beta_{0,j+\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}} - \frac{3}{2} + 4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}) \geq 0, \\
z_7 \geq & \frac{\lambda_j^2}{4}f'^{2-}_{j+\frac{1}{2}}(\beta_{0,j+\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}} - \frac{3}{2}\frac{\Delta x_j}{\Delta x_{j+1}} + 4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_j\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2}) \\
& + \frac{\lambda_j^2}{4}f'^{2+}_{j+\frac{1}{2}}\left((\beta_{0,j+\frac{1}{2}}\frac{\Delta x_j}{\Delta x_{j+\frac{1}{2}}} - \frac{3}{2}\frac{\Delta x_j}{\Delta x_{j+1}} + 4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_j\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2})\right. \\
& \left. -15(M-m)M_2\frac{\Delta t}{\Delta x_{j+1}}\frac{\Delta x_j}{\Delta x_{j+1}} - \frac{4\lambda_j}{3}M_1\frac{\Delta x_j^2}{\Delta x_{j+1}^2}\right) \geq 0, \\
z_8 \geq & \frac{\lambda_j^2}{4}f'^{2-}_{j+\frac{1}{2}}(2\frac{\Delta x_j}{\Delta x_{j+1}} - 8\beta_{1,j+\frac{1}{2}}\frac{\Delta x_j\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2}) \\
& + \frac{\lambda_j^2}{4}f'^{2+}_{j+\frac{1}{2}}\left((2\frac{\Delta x_j}{\Delta x_{j+1}} - 8\beta_{1,j+\frac{1}{2}}\frac{\Delta x_j\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2})\right. \\
& \left. -20(M-m)M_2\frac{\Delta t}{\Delta x_{j+1}}\frac{\Delta x_j}{\Delta x_{j+1}} - \frac{8\lambda_j}{3}M_1\frac{\Delta x_j^2}{\Delta x_{j+1}^2}\right) \geq 0, \\
z_9 \geq & \frac{\lambda_j^2}{4}f'^{2-}_{j+\frac{1}{2}}(4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_j\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2} - \frac{1}{2}\frac{\Delta x_j}{\Delta x_{j+1}}) \\
& + \frac{\lambda_j^2}{4}f'^{2+}_{j+\frac{1}{2}}\left((4\beta_{1,j+\frac{1}{2}}\frac{\Delta x_j\Delta x_{j+\frac{1}{2}}}{\Delta x_{j+1}^2} - \frac{1}{2}\frac{\Delta x_j}{\Delta x_{j+1}})\right. \\
& \left. -5(M-m)M_2\frac{\Delta t}{\Delta x_{j+1}}\frac{\Delta x_j}{\Delta x_{j+1}} - \frac{4\lambda_j}{3}M_1\frac{\Delta x_j^2}{\Delta x_{j+1}^2}\right) \geq 0,
\end{aligned}
$$

under the CFL condition (A.4).

Since II can be written as a half of a convex combination of point values of u^n, we still have $\frac{1}{2}m \le \text{II} \le \frac{1}{2}M$ as before, which implies $m \le \bar{u}_j^{n+1} \le M, j = 1, 2, \ldots, N$ $\square$

A.3.2 A maximum-principle-satisfying scheme on arbitrary nonuniform meshes

To construct the maximum-principle-satisfying scheme on arbitrary nonuniform meshes, we shall first introduce the composite quadrature rule to be used. Define $\Delta x_{j+\frac{1}{2}} = \frac{1}{3}\min\{\Delta x_j, \Delta x_{j+1}\}$ and denote by $\tilde{u}_j^1 = u(x_{j-\frac{1}{2}} - \Delta x_{j-\frac{1}{2}}), \tilde{u}_j^2 = u(x_{j-\frac{1}{2}} - \frac{1}{2}\Delta x_{j-\frac{1}{2}}), \tilde{u}_j^3 = u(x_{j-\frac{1}{2}} + \frac{1}{2}\Delta x_{j-\frac{1}{2}}), \tilde{u}_j^4 = u(x_{j-\frac{1}{2}} + \Delta x_{j-\frac{1}{2}}), \tilde{u}_j^5 = u(x_{j+\frac{1}{2}} - \Delta x_{j+\frac{1}{2}}), \tilde{u}_j^6 = u(x_{j+\frac{1}{2}} - \frac{1}{2}\Delta x_{j+\frac{1}{2}}), \tilde{u}_j^7 = u(x_{j+\frac{1}{2}} + \frac{1}{2}\Delta x_{j+\frac{1}{2}}), \tilde{u}_j^8 = u(x_{j+\frac{1}{2}} + \Delta x_{j+\frac{1}{2}})$, for the cell I_j.

We adopt the composite Gauss-Lobatto rule as follows: The interval I_j is divided into three subintervals, i.e. $I_j = [x_{j-\frac{1}{2}}, x_{j-\frac{1}{2}} + \Delta x_{j-\frac{1}{2}}] \cup [x_{j-\frac{1}{2}} + \Delta x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}} - \Delta x_{j+\frac{1}{2}}] \cup [x_{j+\frac{1}{2}} - \Delta x_{j+\frac{1}{2}}, x_{j+\frac{1}{2}}]$, and each subinterval is assigned with the $2N_q - 1$ Gauss-Lobatto quadrature rule, which results in the quadrature points $\{\tilde{x}_1^j, \tilde{x}_2^j, \ldots, \tilde{x}_{6N_q-5}^j\}$ and quadrature weights $\{\tilde{\omega}_1^j, \omega_2^j, \ldots, \tilde{\omega}_{6N_q-5}^j\}$ on the interval I_j as follows,

$$\tilde{x}_\alpha^j = \begin{cases} x_{j-\frac{1}{2}} + \left(\frac{\hat{x}_\alpha + 1}{2}\right)\Delta x_{j-\frac{1}{2}} & \alpha = 1, 2, \ldots, 2N_q - 1, \\ x_{j-\frac{1}{2}} + \Delta x_{j-\frac{1}{2}} + \left(\frac{\hat{x}_{\alpha-2N_q+2}+1}{2}\right)\left(\Delta x_j - \Delta x_{j-\frac{1}{2}} - \Delta x_{j+\frac{1}{2}}\right) & \alpha = 2N_q, \ldots, 4N_q - 3, \\ x_{j+\frac{1}{2}} - \Delta x_{j+\frac{1}{2}} + \left(\frac{\hat{x}_{\alpha-4N_q+4}+1}{2}\right)\Delta x_{j+\frac{1}{2}} & \alpha = 4N_q - 2, \ldots, 6N_q - 5, \end{cases}$$

and

$$\tilde{\omega}_\alpha^j = \begin{cases} \frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j}\hat{\omega}_\alpha & \alpha = 1, 2, \ldots, 2N_q - 2 \\ \frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j}\hat{\omega}_{2N_q-1} + \left(1 - \frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j} - \frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}\right)\hat{\omega}_1 & \alpha = 2N_q - 1, \\ \left(1 - \frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j} - \frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}\right)\hat{\omega}_{\alpha-2N_q+2} & \alpha = 2N_q, \ldots, 4N_q - 4, \\ \left(1 - \frac{\Delta x_{j-\frac{1}{2}}}{\Delta x_j} - \frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}\right)\hat{\omega}_{2N_q-1} + \frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}\hat{\omega}_1 & \alpha = 4N_q - 3, \\ \frac{\Delta x_{j+\frac{1}{2}}}{\Delta x_j}\hat{\omega}_{\alpha-4N_q+4} & \alpha = 4N_q - 2, \ldots, 6N_q - 5, \end{cases}$$

respectively, where $\{\hat{x}_\alpha, \alpha = 1, 2, \ldots, 2N_q - 1\}$ and $\{\hat{\omega}_\alpha, \alpha = 1, 2, \ldots, 2N_q - 1\}$ are the Gauss-Lobatto points on $[-1, 1]$ and weights satisfying $\sum_{\alpha=1}^{2N_q-1} \hat{\omega}_\alpha = 1$.

We redefine the DDG flux on nonuniform meshes:

$$\widehat{u_x}_{j+\frac{1}{2}}^{\text{DDG}} = \beta_0 \frac{[u]_{j+\frac{1}{2}}}{\Delta x_{j+\frac{1}{2}}} + \{u_x\}_{j+\frac{1}{2}} + \beta_1 \Delta x_{j+\frac{1}{2}} [u_{xx}]_{j+\frac{1}{2}}, \tag{A.5}$$

where β_0, β_1 are penalty parameters satisfying $\frac{1}{8} < \beta_1 < \frac{1}{4}, \beta_0 > \frac{3}{2} - 4\beta_1, j = 1, 2, \ldots, N$ as in the uniform meshes.

Similarly, we have the expansion of DDG fluxes for $u \in V$.

$$\begin{aligned} \widehat{u_x}_{j+\frac{1}{2}}^{\text{DDG}} =& \frac{1}{\Delta x_{j+\frac{1}{2}}} \Big((\frac{1}{2} - 4\beta_1)\tilde{u}^5 + (8\beta_1 - 2)\tilde{u}^6 + (-\beta_0 + \frac{3}{2} - 4\beta_1)u^-_{j+\frac{1}{2}} \\ & + (\beta_0 - \frac{3}{2} + 4\beta_1)u^+_{j+\frac{1}{2}} + (2 - 8\beta_1)\tilde{u}^7 + (4\beta_1 - \frac{1}{2})\tilde{u}^8 \Big) \end{aligned}$$

and

$$\begin{aligned} \widehat{u_x}_{j-\frac{1}{2}}^{\text{DDG}} =& \frac{1}{\Delta x_{j-\frac{1}{2}}} \Big((\frac{1}{2} - 4\beta_1)\tilde{u}^1 + (8\beta_1 - 2)\tilde{u}^2 + (-\beta_0 + \frac{3}{2} - 4\beta_1)u^-_{j-\frac{1}{2}} \\ & + (\beta_0 - \frac{3}{2} + 4\beta_1)u^+_{j-\frac{1}{2}} + (2 - 8\beta_1)\tilde{u}^3 + (4\beta_1 - \frac{1}{2})\tilde{u}^4 \Big) \end{aligned}$$

The main result is as follows,

Theorem A.3.3. *Given $m \le u^n \le M$ and the DDG flux* (A.5)*, the cell averages $\bar{u}_j^{n+1}, j = 1, 2, \ldots, N$ of the solution of scheme* (2.8) *are bounded between m and M under the CFL condition* (A.6)*.*

$$\Delta t \le \min\{q_1, q_2, \ldots, q_7\}, \tag{A.6}$$

where $q_1 = \frac{\hat{\omega}_1}{2M_1} \min_j \Delta x_j$, $q_2 = \frac{4\beta_1 - \frac{1}{2}}{5(M-m)M_2 + \frac{4}{3}M_1} \min_j \Delta x_{j+\frac{1}{2}}$, $q_3 = \frac{2 - 8\beta_1}{20(M-m)M_2 + \frac{8}{3}M_1} \min_j \Delta x_{j+\frac{1}{2}}$,
$q_4 = \frac{\beta_0 - \frac{3}{2} + 4\beta_1}{15(M-m)M_2 + \frac{4}{3}M_1} \min_j \Delta x_{j+\frac{1}{2}}$, $q_5 = \left(\frac{2\hat{\omega}_1}{3M_1^2(\beta_0 - \frac{3}{2} + 4\beta_1)}\right)^{\frac{1}{2}} \min_j \Delta x_{j+\frac{1}{2}}$,
$q_6 = \left(\frac{2\hat{\omega}_{N_q}}{3M_1^2(2 - 8\beta_1)}\right)^{\frac{1}{2}} \min_j \Delta x_{j+\frac{1}{2}}$, $q_7 = \left(\frac{4\hat{\omega}_1}{3M_1^2(4\beta_1 - \frac{1}{2})}\right)^{\frac{1}{2}} \min_j \Delta x_{j+\frac{1}{2}}$.

Proof. We have exactly the same results as in (2.17) and (2.18), but now II is expanded differently:

$$\begin{aligned}
\mathrm{II} =& \frac{1}{2} \sum_{\gamma=2}^{N_q-1} \hat{\omega}_\gamma^j u^\gamma + \frac{1}{2} \sum_{\gamma=N_q+1}^{2N_q-2} \hat{\omega}_\gamma^j u^\gamma + \frac{1}{2} \sum_{\gamma=2N_q}^{4N_q-4} \hat{\omega}_\gamma^j u^\gamma + \frac{1}{2} \sum_{\gamma=4N_q-2}^{5N_q-5} \hat{\omega}_\gamma^j u^\gamma + \frac{1}{2} \sum_{\gamma=5N_q-3}^{6N_q-5} \hat{\omega}_\gamma^j u^\gamma \\
& + z_1 \tilde{u}^1 + z_2 \tilde{u}^2 + z_3 u^-_{j-\frac{1}{2}} + z_4 u^+_{j-\frac{1}{2}} + z_5 \tilde{u}^3 + z_6 \tilde{u}^4 \\
& + z_7 \tilde{u}^5 + z_8 \tilde{u}^6 + z_9 u^-_{j+\frac{1}{2}} + z_{10} u^+_{j+\frac{1}{2}} + z_{11} \tilde{u}^7 + z_{12} \tilde{u}^8,
\end{aligned}$$

where

$$
\begin{aligned}
z_1 =& \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{-}_{j-\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) + \Delta t {f''}^{-}_{j-\frac{1}{2}} {u_x}^{-}_{j-\frac{1}{2}} + \frac{4}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'}^{-}_{j-\frac{1}{2}} \right) \\
&+ \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{+}_{j-\frac{1}{2}} (4\beta_1 - \frac{1}{2}), \\
z_2 =& \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{-}_{j-\frac{1}{2}} \left((2 - 8\beta_1) - 4\Delta t {f''}^{-}_{j-\frac{1}{2}} {u_x}^{-}_{j-\frac{1}{2}} - \frac{8}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'}^{-}_{j-\frac{1}{2}} \right) \\
&+ \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{+}_{j-\frac{1}{2}} (2 - 8\beta_1) \\
z_3 =& \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{-}_{j-\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 3\Delta t {f''}^{-}_{j-\frac{1}{2}} {u_x}^{-}_{j-\frac{1}{2}} + \frac{4}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'}^{-}_{j-\frac{1}{2}} \right) \\
&+ \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{+}_{j-\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \\
z_4 =& \frac{1}{2}\hat{\omega}^{j}_{1} - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{-}_{j-\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \\
&- \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{+}_{j-\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 3\Delta t {f''}^{+}_{j-\frac{1}{2}} {u_x}^{+}_{j-\frac{1}{2}} - \frac{4}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'}^{+}_{j-\frac{1}{2}} \right) \\
z_5 =& \frac{1}{2}\hat{\omega}^{j}_{N_q} - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{-}_{j-\frac{1}{2}} (2 - 8\beta_1) \\
&- \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{+}_{j-\frac{1}{2}} \left((2 - 8\beta_1) - 4\Delta t {f''}^{+}_{j-\frac{1}{2}} {u_x}^{+}_{j-\frac{1}{2}} + \frac{8}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'}^{+}_{j-\frac{1}{2}} \right) \\
z_6 =& \frac{1}{2}\hat{\omega}^{j}_{2N_q-1} - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{-}_{j-\frac{1}{2}} (4\beta_1 - \frac{1}{2}) \\
&- \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'^{2}}^{+}_{j-\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) + \Delta t {f''}^{+}_{j-\frac{1}{2}} {u_x}^{+}_{j-\frac{1}{2}} - \frac{4}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} {f'}^{+}_{j-\frac{1}{2}} \right)
\end{aligned}
$$

$$
\begin{aligned}
z_7 =& \frac{1}{2}\hat{\omega}^j_{4N_q-3} - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{-}_{j+\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) + \Delta t {f''}^{-}_{j+\frac{1}{2}} {u_x}^{-}_{j+\frac{1}{2}} + \frac{4}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'}^{-}_{j+\frac{1}{2}} \right) \\
& - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{+}_{j+\frac{1}{2}} (4\beta_1 - \frac{1}{2}) \\
z_8 =& \frac{1}{2}\hat{\omega}^j_{5N_q-4} - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{-}_{j+\frac{1}{2}} \left((2 - 8\beta_1) - 4\Delta t {f''}^{-}_{j+\frac{1}{2}} {u_x}^{-}_{j+\frac{1}{2}} - \frac{8}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'}^{-}_{j+\frac{1}{2}} \right) \\
& - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{+}_{j+\frac{1}{2}} (2 - 8\beta_1) \\
z_9 =& \frac{1}{2}\hat{\omega}^j_{6N_q-5} - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{-}_{j+\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 3\Delta t {f''}^{-}_{j+\frac{1}{2}} {u_x}^{-}_{j+\frac{1}{2}} + \frac{4}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'}^{-}_{j+\frac{1}{2}} \right) \\
& - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{+}_{j+\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \\
z_{10} =& \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{-}_{j+\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \\
& + \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{+}_{j+\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 3\Delta t {f''}^{+}_{j+\frac{1}{2}} {u_x}^{+}_{j+\frac{1}{2}} - \frac{4}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'}^{+}_{j+\frac{1}{2}} \right) \\
z_{11} =& \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{-}_{j+\frac{1}{2}} (2 - 8\beta_1) \\
& + \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{+}_{j+\frac{1}{2}} \left((2 - 8\beta_1) - 4\Delta t {f''}^{+}_{j+\frac{1}{2}} {u_x}^{+}_{j+\frac{1}{2}} + \frac{8}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'}^{+}_{j+\frac{1}{2}} \right) \\
z_{12} =& \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{-}_{j+\frac{1}{2}} (4\beta_1 - \frac{1}{2}) \\
& + \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'^{2}}^{+}_{j+\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) + \Delta t {f''}^{+}_{j+\frac{1}{2}} {u_x}^{+}_{j+\frac{1}{2}} - \frac{4}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} {f'}^{+}_{j+\frac{1}{2}} \right)
\end{aligned}
$$

One can verify that

$$
\begin{aligned}
& \frac{1}{2}\sum_{\gamma=2}^{N_q-1}\hat{\omega}^j_\gamma + \frac{1}{2}\sum_{\gamma=N_q+1}^{2N_q-2}\hat{\omega}^j_\gamma + \frac{1}{2}\sum_{\gamma=2N_q}^{4N_q-4}\hat{\omega}^j_\gamma + \frac{1}{2}\sum_{\gamma=4N_q-2}^{5N_q-5}\hat{\omega}^j_\gamma + \frac{1}{2}\sum_{\gamma=5N_q-3}^{6N_q-5}\hat{\omega}^j_\gamma \\
& + z_1 + z_2 + z_3 + z_4 + z_5 + z_6 + z_7 + z_8 + z_9 + z_{10} + z_{11} + z_{12} = \frac{1}{2},
\end{aligned}
$$

and

$$
\begin{aligned}
z_1 \geq & \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} f'^{2-}_{j-\frac{1}{2}} \left((4\beta_1 - \frac{1}{2}) - 5(M-m)M_2 \frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} - \frac{4}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1 \right) \\
& + \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} f'^{2+}_{j-\frac{1}{2}} (4\beta_1 - \frac{1}{2}) \geq 0, \\
z_2 \geq & \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} f'^{2-}_{j-\frac{1}{2}} \left((2 - 8\beta_1) - 20(M-m)M_2 \frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} - \frac{8}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1 \right) \\
& + \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} f'^{2+}_{j-\frac{1}{2}} (2 - 8\beta_1) \geq 0, \\
z_3 \geq & \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} f'^{2-}_{j-\frac{1}{2}} \left((\beta_0 - \frac{3}{2} + 4\beta_1) - 15(M-m)M_2 \frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} - \frac{4}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1 \right) \\
& + \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} f'^{2+}_{j-\frac{1}{2}} (\beta_0 - \frac{3}{2} + 4\beta_1) \geq 0, \\
z_4 \geq & \frac{1}{2}\hat{\omega}_1^j - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1^2 (\beta_0 - \frac{3}{2} + 4\beta_1) \\
& - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1^2 \left((\beta_0 - \frac{3}{2} + 4\beta_1) + 15(M-m)M_2 \frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} + \frac{4}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1 \right) \geq 0, \\
z_5 \geq & \frac{1}{2}\hat{\omega}_{N_q}^j - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1^2 (2 - 8\beta_1) \\
& - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1^2 \left((2 - 8\beta_1) + 20(M-m)M_2 \frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} + \frac{8}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1 \right) \geq 0, \\
z_6 \geq & \frac{1}{2}\hat{\omega}_{2N_q-1}^j - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1^2 (4\beta_1 - \frac{1}{2}) \\
& - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1^2 \left((4\beta_1 - \frac{1}{2}) + 5(M-m)M_2 \frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} + \frac{4}{3}\frac{\Delta t}{\Delta x_{j-\frac{1}{2}}} M_1 \right) \geq 0, \\
z_7 \geq & \frac{1}{2}\hat{\omega}_{4N_q-3}^j - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} M_1^2 \left((4\beta_1 - \frac{1}{2}) + 5(M-m)M_2 \frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} + \frac{4}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} M_1 \right) \\
& - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} M_1^2 (4\beta_1 - \frac{1}{2}) \geq 0, \\
z_8 \geq & \frac{1}{2}\hat{\omega}_{5N_q-4}^j - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} M_1^2 \left((2 - 8\beta_1) + 20(M-m)M_2 \frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} + \frac{8}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} M_1 \right) \\
& - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} M_1^2 (2 - 8\beta_1) \geq 0,
\end{aligned}
$$

$$
\begin{aligned}
z_9 \geq & \frac{1}{2}\hat{\omega}^j_{6N_q-5} - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}M_1^2\left((\beta_0 - \frac{3}{2} + 4\beta_1) + 15(M-m)M_2\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} + \frac{4}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}M_1\right) \\
& - \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}M_1^2(\beta_0 - \frac{3}{2} + 4\beta_1) \geq 0, \\
z_{10} \geq & \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}f'^{2-}_{j+\frac{1}{2}}(\beta_0 - \frac{3}{2} + 4\beta_1) \\
& + \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}f'^{2+}_{j+\frac{1}{2}}\left((\beta_0 - \frac{3}{2} + 4\beta_1) - 15(M-m)M_2\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} - \frac{4}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}M_1\right) \geq 0, \\
z_{11} \geq & \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}f'^{2-}_{j+\frac{1}{2}}(2 - 8\beta_1) \\
& + \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}f'^{2+}_{j+\frac{1}{2}}\left((2 - 8\beta_1) - 20(M-m)M_2\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} - \frac{8}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}M_1\right) \geq 0, \\
z_{12} \geq & \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}f'^{2-}_{j+\frac{1}{2}}(4\beta_1 - \frac{1}{2}) \\
& + \frac{\lambda_j}{4}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}f'^{2+}_{j+\frac{1}{2}}\left((4\beta_1 - \frac{1}{2}) - 5(M-m)M_2\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}} - \frac{4}{3}\frac{\Delta t}{\Delta x_{j+\frac{1}{2}}}M_1\right) \geq 0,
\end{aligned}
$$

under the CFL condition (A.6).

Therefore, we have $m \leq \bar{u}_j^{n+1} \leq M, j = 1, 2, \ldots, N$ following the same arguments as before. □

A.3.3 Numerical tests on nonuniform meshes

We demonstrate the accuracy and effectiveness of the maximum-principle-satisfying algorithm established in Section A.3.1 and Section A.3.2 on nonuniform meshes.

Example A.3.1. *We solve the linear equation* $u_t + u_x = 0$ *in the domain* $\Omega = [-1, 1]$ *with periodic boundary conditions and discontinuous initial condition*

$$
u_0(x) = \begin{cases} 1, & -1 \leq x \leq 0, \\ -1, & 0 \leq x \leq 1. \end{cases}
$$

and take the terminal time $T = 100$ *to show the effect of the maximum-principle-preserving.*

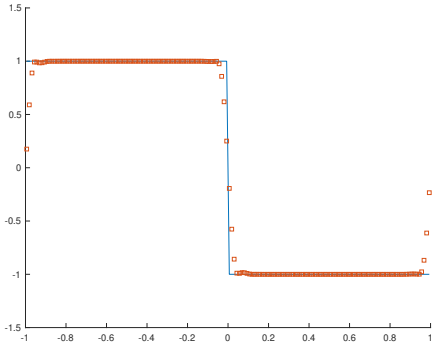
We solve the Burgers' equation $u_t + \left(\frac{u^2}{2}\right)_x = 0$ *in the domain* $\Omega = [0, 2\pi]$ *with initial condition* $u_0(x) = \frac{1}{2} + \sin(x)$ *and periodic boundary conditions, and take the terminal time* $T = 0.3$ *to show the accuracy.*

For the algorithm established in Section A.3.1, we generate the nonuniform meshes by adding uniformly distributed perturbation within $[-0.1\Delta x, 0.1\Delta x]$ on the inner nodes of the uniform mesh. For the algorithm established in Section A.3.2, we generate the nonuniform meshes by adding uniformly distributed perturbation within $[-0.3\Delta x, 0.3\Delta x]$ on the inner nodes of the uniform mesh.

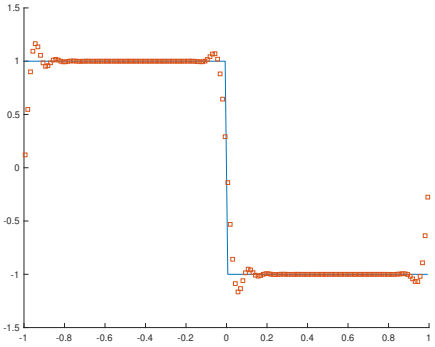
The results are given in Table A.1 and Figure A.1, from which we can observe the third order accuracy and maximum-principle-preserving effect.

	Algorithm A.3.1				Algorithm A.3.2			
N	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
20	9.33E-04	–	1.50E-03	–	1.19E-03	–	2.37E-03	–
40	1.15E-04	3.02	2.38E-04	2.65	1.44E-04	3.05	3.61E-04	2.72
80	1.41E-05	3.03	4.09E-05	2.54	1.90E-05	2.92	8.56E-05	2.08
160	1.73E-06	3.03	5.56E-06	2.88	2.01E-06	3.24	9.33E-06	3.20
320	2.11E-07	3.03	8.14E-07	2.77	2.72E-07	2.89	1.69E-06	2.46
640	2.59E-08	3.03	1.04E-07	2.96	3.23E-08	3.07	1.93E-07	3.13

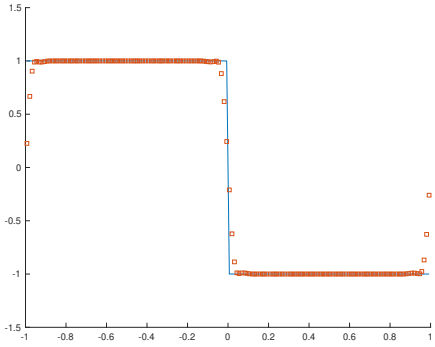
Table A.1: Results of Example A.3.1, Burgers' equation at $T = 0.3$



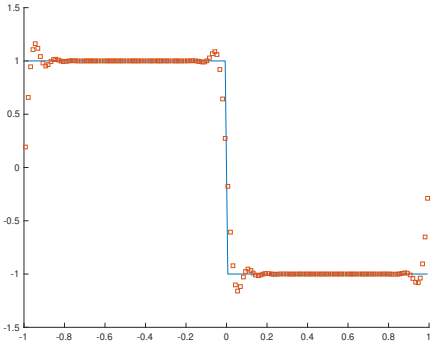
(a) Algorithm A.3.1 with limiter



(b) Algorithm A.3.1 without limiter



(c) Algorithm A.3.2 with limiter



(d) Algorithm A.3.2 without limiter

Figure A.1: Results of Example A.3.1 with discontinuous initial condition at $T = 100$. $N = 160$. Solid line: exact solution; Squares: numerical solution (cell averages).

Appendix B

Appendix for Chapter 3

B.1 The positivity of solution at downwind points in one space dimension

In this appendix, we prove that the solutions of schemes proposed in Section 3.2 are nonnegative at downwind points under certain mesh size conditions, provided the positivity of the boundary condition and source term.

Theorem B.1.1. *For the problem* (3.1) *with* $a(x) > 0$ *and* $f, u(0) \geq 0$*, the solution of the scheme* (3.8) *satisfies* $u^-_{j+\frac{1}{2}} \geq 0, j = 1, 2, \ldots, N$*, for* $k = 1, 2, 3, \ldots$ *if* $\lambda = 0$*, and for* $k = 1, 2$ *if* $\lambda \Delta x_j \leq 2 \min_{x \in I_j} a(x), j = 1, 2, \ldots, N$.

Proof. If $\lambda = 0$, we take the test function $w = 1$ in the scheme (3.8) to yield the equations

$$a(x_{j+\frac{1}{2}}) u^-_{j+\frac{1}{2}} = a(x_{j-\frac{1}{2}}) u^-_{j-\frac{1}{2}} + \fint_{I_j} f dx, j = 1, 2, \ldots, N,$$

satisfied by the solution. Since $a(x) > 0, f(x) \geq 0$ on Ω, and $u^-_{\frac{1}{2}} = u(0) \geq 0$, by induction, we have $u^-_{j+\frac{1}{2}} \geq 0, j = 1, 2, \ldots, N$.

If $\lambda > 0$ and $k = 1$, it is easy to check that the test function $\xi(x) = \frac{1}{a(x_{j+\frac{1}{2}})} - \frac{2\lambda(x_{j+\frac{1}{2}} - x)}{a(x_{j+\frac{1}{2}})(2a(x_j) + \lambda \Delta x_j)} \in P^1(I_j)$ satisfies

$$-\fint_{I_j} \left(a(x) v \xi_x - \lambda v \xi \right) dx + a(x_{j+\frac{1}{2}}) v^-_{j+\frac{1}{2}} \xi^-_{j+\frac{1}{2}} = v^-_{j+\frac{1}{2}}, \tag{B.1}$$

for all $v \in P^1(I_j)$. Moreover, $\xi(x_{j+\frac{1}{2}}) = \frac{1}{a(x_{j+\frac{1}{2}})} > 0$ and $\xi(x_{j-\frac{1}{2}}) = \frac{2a(x_j) - \lambda \Delta x_j}{a(x_{j+\frac{1}{2}})(2a(x_j) + \lambda \Delta x_j)} \geq 0$ if $\lambda \Delta x_j \leq 2 \min_{x \in I_j} a(x)$, which implies that $\xi(x) \geq 0$ on I_j. Therefore, by taking

the test function $w = \xi$ (extends to zero outside I_j) in the scheme (3.8), we have

$$u^-_{j+\frac{1}{2}} = a(x_{j-\frac{1}{2}})u^-_{j-\frac{1}{2}}\xi^+_{j-\frac{1}{2}} + \int_{I_j} f\xi dx, \tag{B.2}$$

which implies $u^-_{j+\frac{1}{2}} \geq 0$ if $u^-_{j-\frac{1}{2}} \geq 0$. Since $u^-_{\frac{1}{2}} = u(0) \geq 0$, by induction, we have $u^-_{j+\frac{1}{2}} \geq 0$ for $j = 1, 2, \ldots, N$.

If $\lambda > 0$ and $k = 2$, one can check that $\xi(x) = \xi_1 L_1(x) + \xi_2 L_2(x) + \xi_3 L_3(x)$ satisfies the equation (B.1) for all $v \in P^2(I_j)$, where $L_1(x), L_2(x), L_3(x)$ are the Lagrange basis at $\{\hat{x}_1, \hat{x}_2, x_{j+\frac{1}{2}}\}$ with $L_1(\hat{x}_1) = 1, L_2(\hat{x}_2) = 1, L_3(x_{j+\frac{1}{2}}) = 1$, and

$$\begin{aligned}
\xi_1 &= \frac{2\sqrt{3}a(\hat{x}_1)(2\sqrt{3}a(\hat{x}_2) - \lambda\Delta x_j)}{a(x_{j+\frac{1}{2}})\left(12a(\hat{x}_1)a(\hat{x}_2) + 3a(\hat{x}_1)\lambda\Delta x_j + 3a(\hat{x}_2)\lambda\Delta x_j + \lambda^2\Delta x_j^2\right)}, \\
\xi_2 &= \frac{2\sqrt{3}a(\hat{x}_2)(2\sqrt{3}a(\hat{x}_1) + \lambda\Delta x_j)}{a(x_{j+\frac{1}{2}})\left(12a(\hat{x}_1)a(\hat{x}_2) + 3a(\hat{x}_1)\lambda\Delta x_j + 3a(\hat{x}_2)\lambda\Delta x_j + \lambda^2\Delta x_j^2\right)}, \\
\xi_3 &= \frac{1}{a(x_{j+\frac{1}{2}})}.
\end{aligned}$$

Moreover, if $\lambda\Delta x_j \leq 2\min_{x\in I_j} a(x)$, we have $\xi(\hat{x}_1) = \xi_1 \geq 0, \xi(\hat{x}_2) = \xi_2 \geq 0$ and

$$\xi(x_{j-\frac{1}{2}}) = \frac{12a(\hat{x}_1)a(\hat{x}_2) - 3a(\hat{x}_1)\lambda\Delta x_j - 3a(\hat{x}_2)\lambda\Delta x_j + \lambda^2\Delta x_j^2}{a(x_{j+\frac{1}{2}})\left(12a(\hat{x}_1)a(\hat{x}_2) + 3a(\hat{x}_1)\lambda\Delta x_j + 3a(\hat{x}_2)\lambda\Delta x_j + \lambda^2\Delta x_j^2\right)} \geq 0.$$

Therefore, follow the same lines as in the case $k = 1$, we obtain $u^-_{j+\frac{1}{2}} \geq 0$ for $j = 1, 2, \ldots, N$. □

Almost the same arguments can be used to prove a similar theorem for the scheme (3.12) with $k = 1$ and scheme (3.13) with $k = 2$, except that the positivity of ξ at the midpoint need to be checked due to the quadrature rules adopted on the right hand side the schemes. The theorem is stated as follows and the proof is omitted.

Theorem B.1.2. *For the problem* (3.2) *with* $a(u) \geq c > 0$, *and* $f, u(0), \lambda \geq 0$, *the*

solutions of the scheme (3.12) *with* $k = 1$ *and scheme* (3.13) *with* $k = 2$ *satisfy* $u^-_{j+\frac{1}{2}} \geq 0, j = 1, 2, \ldots, N$ *if* $\lambda \Delta x_j \leq 2c, j = 1, 2, \ldots, N$.

Remark B.1.1. *For the time-dependent linear problem* $u_t + (a(x)u)_x = 0$*, the backward Euler time discretization approach yields the stationary equations* $(a(x)u^n)_x + \Delta t^{-1}u^n = \Delta t^{-1}u^{n-1}, n = 1, 2, \ldots, \frac{T}{\Delta t}$*. Therefore, the backward Euler discontinuous Galerkin scheme* (3.8) *is positivity-preserving under the CFL condition* $\min_{x\in I_j} a(x)\frac{\Delta t}{\Delta x_j} \geq \frac{1}{2}, \forall j$*, if the positivity-preserving limiter is not applied until the computation of* u^n *is completed at the time level* n*. This result can be viewed as an extension of the theoretical result of positivity-preserving backward Euler discontinuous Galerkin method for* $u_t + u_x = 0$ *analyzed in [55].*

B.2 Investigation of the schemes (3.7) and (3.8) for some special $a(x)$

The unmodulated P^k-DG schemes (3.7) for the equation (3.1) could result in negative cell averages in the solution for some special $a(x)$. For instance, one can take $a(x) = 1 + x, a(x) = 1 + x^2, a(x) = 1 + x^3, a(x) = 1 + x^4, a(x) = 1 + x^5$ in the unmodulated P^1, P^2, P^3, P^4, P^5-DG schemes, respectively, for some particular λ.

More precisely, for the test function $\xi \in P^k([0, h])$, s.t.

$$-\int_0^h (a(x)v\xi_x - \lambda v\xi)\, dx + a(h)v(h)\xi(h) = \frac{1}{h}\int_0^h v dx, \quad \forall v \in P^k([0, h]),$$

where $a(x) = 1 + x^k$, $k = 1, 2, 3, 4, 5$, $\xi(h)$ is strictly negative for sufficiently small h, though $\lim_{h\to 0} \xi(hx) = 1 - x \geq 0$ for $x \in [0, 1]$.

One can check that, if $\lambda = 0$:

- For $k = 1$, $\lim_{h\to 0} \frac{\xi(h)}{h} = -\frac{1}{6}$.
- For $k = 2$, $\lim_{h\to 0} \frac{\xi(h)}{h^2} = -\frac{1}{30}$.
- For $k = 3$, $\lim_{h\to 0} \frac{\xi(h)}{h^3} = -\frac{1}{140}$.
- For $k = 4$, $\lim_{h\to 0} \frac{\xi(h)}{h^4} = -\frac{1}{630}$.
- For $k = 5$, $\lim_{h\to 0} \frac{\xi(h)}{h^5} = -\frac{1}{2772}$.

One can also check that, if $\lambda = \frac{1}{2}$:

- For $k = 1$, $\lim_{h\to 0} \frac{\xi(h)}{h} = -\frac{1}{12}$.
- For $k = 2$, $\lim_{h\to 0} \frac{\xi(h)}{h^2} = -\frac{7}{240}$.
- For $k = 3$, $\lim_{h\to 0} \frac{\xi(h)}{h^3} = -\frac{47}{6720}$.
- For $k = 4$, $\lim_{h\to 0} \frac{\xi(h)}{h^4} = -\frac{383}{241920}$.
- For $k = 5$, $\lim_{h\to 0} \frac{\xi(h)}{h^5} = -\frac{349}{967680}$.

Therefore, we can construct proper source term $f(x) \geq 0$ with large values around $x = h$, such that the average of the solution on the cell $[0, h]$ is negative.

However, using the positivity-preserving scheme defined in (3.8), the above problems are resolved. One can check that, for the test function $\xi \in P^k([0, h])$, s.t.

$$-\fint_0^h (a(x)v\xi_x - \lambda v\xi)\, dx + a(h)v(h)\xi(h) = \frac{1}{h}\int_0^h v dx, \quad \forall v \in P^k([0, h]),$$

where $a(x) = 1 + x^k$, $k = 1, 2, 3, 4, 5$, we still have $\lim_{h\to 0} \xi(hx) = 1 - x$ but now $\xi(h) = 0$ in all those cases.

Appendix C

Appendix for Chapter 5

C.1 A comparison of operations in LCD-WENO and RI-WENO for one dimensional shallow water equations

We analyze and compare the floating point operations in the local characteristic decomposition based WENO (LCD-WENO) and Riemann invariants based WENO (RI-WENO) algorithms for one dimensional shallow water equations in Table C.1. From comparison, it is clear that, RI-WENO exempts the computations at steps 1

steps	LCD-WENO	RI-WENO
1	$\mathbf{u} = \frac{1}{2}(\mathbf{u}_j + \mathbf{u}_{j+1})$, or Roe's average.	None
2	$\mathbf{R}(\mathbf{u}) = \begin{bmatrix} 1 & 1 \\ u-\sqrt{gh} & u+\sqrt{gh} \end{bmatrix}$, $\mathbf{R}^{-1}(\mathbf{u}) = \begin{bmatrix} \frac{1}{2}+\frac{u}{2\sqrt{gh}} & -\frac{1}{2\sqrt{gh}} \\ \frac{1}{2}-\frac{u}{2\sqrt{gh}} & \frac{1}{2\sqrt{gh}} \end{bmatrix}$.	$w_1 = u + 2\sqrt{gh}$, $w_2 = u - 2\sqrt{gh}$.
3	$\mathbf{v}_i = \mathbf{R}^{-1}\mathbf{u}_i$, $i = j-r+1, \ldots, j+r$.	None
4	$\mathbf{v}^-_{j+\frac{1}{2}} = \text{weno}(\mathbf{v}_{j-r+1}, \ldots, \mathbf{v}_{j+r-1})$, $\mathbf{v}^+_{j+\frac{1}{2}} = \text{weno}(\mathbf{v}_{j+r}, \ldots, \mathbf{v}_{j-r+2})$,	$\mathbf{w}^-_{j+\frac{1}{2}} = \text{weno}(\mathbf{w}_{j-r+1}, \ldots, \mathbf{w}_{j+r-1})$, $\mathbf{w}^+_{j+\frac{1}{2}} = \text{weno}(\mathbf{w}_{j+r}, \ldots, \mathbf{w}_{j-r+2})$
5	$\mathbf{u}^\pm_{j+\frac{1}{2}} = \mathbf{R}\mathbf{v}^\pm_{j+\frac{1}{2}}$	$\mathbf{u}^\pm_{j+\frac{1}{2}} = \mathbf{u}\left(\mathbf{w}^\pm_{j+\frac{1}{2}}\right)$
6	$\hat{\mathbf{f}}(\mathbf{u}^-_{j+\frac{1}{2}}, \mathbf{u}^+_{j+\frac{1}{2}}, \cdots)$	$\hat{\mathbf{f}}(\mathbf{u}^-_{j+\frac{1}{2}}, \mathbf{u}^+_{j+\frac{1}{2}}, \cdots)$

Table C.1: Comparison of operations in LCD-WENO and RI-WENO algorithms for one dimensional shallow water equations

and 3, saves computational costs at step 2, and has exactly the same costs at steps 4, 5 and 6. (At step 5, both algorithms use 4 multiplications and two additions, due to the relation $h = c\,(w_1 - w_2)^2, u = \frac{1}{2}(w_1 + w_2), hu = h * u$, where $c = \frac{1}{16g}$.)

C.2 The definition of $G(\cdot)$ and computation of $G^{-1}(\cdot)$ in Example 5.4.5

Let

$$g(u) = \frac{1}{(1+\frac{2+\alpha}{2\alpha})u + \frac{2+\alpha}{2\alpha}u^{-1} - \sqrt{\left(\frac{2+\alpha}{2\alpha}\right)^2 u^2 + \left(\frac{2+\alpha}{2\alpha}\right)^2 u^{-2} + \frac{8+8\alpha-2\alpha^2}{4\alpha^2}}}, \quad u > 0,$$

then

$$\begin{aligned} G(u) &= \int_1^u g(y)dy \\ &= \frac{1}{16(1+\alpha)}\left(-\alpha\log 16 + (8+4\alpha)\log u + 4\alpha\log(1+u^2) + 2\alpha\log\left(\frac{\alpha-\alpha u^2+t}{-\alpha+\alpha u^2+t}\right)\right. \\ &+ (\alpha+2)\log\left(\frac{-\alpha^2-4\alpha-4+(\alpha^2-4\alpha-4)u^2+(2+\alpha)t}{\alpha^2-4\alpha-4+(-\alpha^2-4\alpha-4)u^2+(2+\alpha)t}\right) \\ &\left.+(\alpha+2)\log\left(\frac{-\alpha^2+4\alpha+4+(\alpha^2+4\alpha+4)u^2+(2+\alpha)t}{\alpha^2+4\alpha+4+(-\alpha^2+4\alpha+4)u^2+(2+\alpha)t}\right)\right), \end{aligned}$$

where $t = \sqrt{\alpha^2(u^2-1)^2 + (4\alpha+4)(u^2+1)^2}$.

Note that $G(u)$ is a log-like monotone increasing concave function with $\left(\frac{2+\alpha}{2+2\alpha}\right)u^{-1} < g(u) < u^{-1}$ for $u \in (0,\infty)$, and $\lim_{u\to 0^+}\frac{g(u)}{\left(\frac{2+\alpha}{2+2\alpha}\right)u^{-1}} = 1$, $\lim_{u\to\infty}\frac{g(u)}{u^{-1}} = 1$, thus one can compute $G^{-1}(\log\frac{1}{q})$ by solving u from the equation $G(u) + \log q = 0$ based on the Newton iteration.

Bibliography

[1] *Ocean State Center for Advanced Resources*. `https://docs.ccv.brown.edu/oscar/`.

[2] V. G. Babskii, M. Y. Zhukov, and V. I. Yudovich, *Mathematical theory of electrophoresis*, Consultants Bureau, New York, 1989.

[3] D. S. Balsara and C.-W. Shu, *Monotonicity preserving weighted essentially non-oscillatory schemes with increasingly high order of accuracy*, Journal of Computational Physics, 160 (2000), pp. 405–452.

[4] R. Borges, M. Carmona, B. Costa, and W. S. Don, *An improved weighted essentially non-oscillatory scheme for hyperbolic conservation laws*, Journal of Computational Physics, 227 (2008), pp. 3191–3211.

[5] Y. Brenier and S. Osher, *The discrete one-sided Lipschitz condition for convex scalar conservation laws*, SIAM Journal on Numerical Analysis, 25 (1988), pp. 8–23.

[6] R. L. Burden, J. D. Faires, and A. M. Burden, *Numerical analysis*, Cengage learning, 2015.

[7] S. Busto, S. Chiocchetti, M. Dumbser, E. Gaburro, and I. Peshkov, *High order ADER schemes for continuum mechanics*, Frontiers in Physics, 8 (2020), p. 32.

[8] E. Carlini, R. Ferretti, and G. Russo, *A weighted essentially nonoscillatory, large time-step scheme for Hamilton–Jacobi equations*, SIAM Journal on Scientific Computing, 27 (2005), pp. 1071–1091.

[9] M. Castro, B. Costa, and W. S. Don, *High order weighted essentially non-oscillatory WENO-Z schemes for hyperbolic conservation laws*, Journal of Computational Physics, 230 (2011), pp. 1766–1792.

[10] 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 (2017), pp. 427–461.

[11] Z. Chen, H. Huang, and J. Yan, *Third order maximum-principle-satisfying direct discontinuous Galerkin methods for time dependent convection diffusion equations on unstructured triangular meshes*, Journal of Computational Physics, 308 (2016), pp. 198–217.

[12] N. Chuenjarern, Z. Xu, and Y. Yang, *High-order bound-preserving discontinuous Galerkin methods for compressible miscible displacements in porous media on triangular meshes*, Journal of Computational Physics, 378 (2019), pp. 110–128.

[13] 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 (1990), pp. 545–581.

[14] 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 (1989), pp. 90–113.

[15] B. Cockburn, M. Luskin, C.-W. Shu, and E. Süli, *Enhanced accuracy by post-processing for finite element methods for hyperbolic equations*, Mathematics of Computation, 72 (2003), pp. 577–606.

[16] 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 (1989), pp. 411–435.

[17] ——, *The Runge-Kutta local projection-discontinuous-Galerkin finite element method for scalar conservation laws*, ESAIM: Mathematical Modelling and Numerical Analysis, 25 (1991), pp. 337–361.

[18] ——, *The Runge–Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems*, Journal of Computational Physics, 141 (1998), pp. 199–224.

[19] C. M. Dafermos, *Hyperbolic conservation laws in continuum physics*, Series: Grundlehren der mathematischen Wissenschaften 325, Springer-Verlag, Berlin, 2005.

[20] M. Dumbser, C. Enaux, and E. F. Toro, *Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws*, Journal of Computational Physics, 227 (2008), pp. 3971–4001.

[21] M. Dumbser, F. Fambri, M. Tavelli, M. Bader, and T. Weinzierl, *Efficient implementation of ADER discontinuous Galerkin schemes for a scalable hyperbolic PDE engine*, Axioms, 7 (2018), p. 63.

[22] M. Dumbser and M. Käser, *Arbitrary high order non-oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems*, Journal of Computational Physics, 221 (2007), pp. 693–723.

[23] M. Dumbser and C.-D. Munz, *Building blocks for arbitrary high order discontinuous Galerkin schemes*, Journal of Scientific Computing, 27 (2006), pp. 215–230.

[24] W. Fiveland, *Discrete-ordinates solutions of the radiative transport equation for rectangular enclosures*, (1984).

[25] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu, *High order strong stability preserving time discretizations*, Journal of Scientific Computing, 38 (2009), pp. 251–289.

[26] ——, *Strong stability preserving Runge-Kutta and multistep time discretizations*, World Scientific, 2011.

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

[28] H. Guo, W. Feng, Z. Xu, and Y. Yang, *Conservative numerical methods for the reinterpreted discrete fracture model on non-conforming meshes and their applications in contaminant transportation in fractured porous media*, Advances in Water Resources, 153 (2021), p. 103951.

[29] H. Guo and Y. Yang, *Bound-preserving discontinuous Galerkin method for compressible miscible displacement in porous media*, SIAM Journal on Scientific Computing, 39 (2017), pp. A1969–A1990.

[30] W. Guo, J.-M. Qiu, and J. Qiu, *A new Lax–Wendroff discontinuous Galerkin method with superconvergence*, Journal of Scientific Computing, 65 (2015), pp. 299–326.

[31] A. Harten, *High resolution schemes for hyperbolic conservation laws*, Journal of Computational Physics, 49 (1983), pp. 357–393.

[32] ——, *On a class of high resolution total-variation-stable finite-difference schemes*, SIAM Journal on Numerical Analysis, 21 (1984), pp. 1–23.

[33] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, *Uniformly high order accurate essentially non-oscillatory schemes, III*, Journal of Computational physics, 71 (1987), pp. 231–303.

[34] A. K. Henrick, T. D. Aslam, and J. M. Powers, *Mapped weighted essentially non-oscillatory schemes: Achieving optimal order near critical points*, Journal of Computational Physics, 207 (2005), pp. 542–567.

[35] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, Journal of Computational Physics, 126 (1996), pp. 202–228.

[36] Y. Jiang, C.-W. Shu, and M. Zhang, *An alternative formulation of finite difference weighted ENO schemes with Lax–Wendroff time discretization for conservation laws*, SIAM Journal on Scientific Computing, 35 (2013), pp. A1137–A1160.

[37] K. D. Lathrop and B. G. Carlson, *Discrete ordinates angular quadrature of the neutron transport equation*, tech. rep., Los Alamos Scientific Lab., N. Mex., 1964.

[38] J. E. Lavery, *Solution of steady-state one-dimensional conservation laws by mathematical programming*, SIAM Journal on Numerical Analysis, 26 (1989), pp. 1081–1089.

[39] P. D. Lax and B. Wendroff, *Systems of conservation laws*, Communications on Pure and Applied Mathematics, 13 (1960), pp. 217–237.

[40] 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, (1974), pp. 89–123.

[41] R. J. LeVeque, *Numerical methods for conservation laws*, vol. 214, Birkhäuser, 1992.

[42] H. Li, S. Xie, and X. Zhang, *A high order accurate bound-preserving compact finite difference scheme for scalar convection diffusion equations*, SIAM Journal on Numerical Analysis, 56 (2018), pp. 3308–3345.

[43] J. Li and Z. Du, *A two-stage fourth order time-accurate discretization for Lax–Wendroff type flow solvers I. hyperbolic conservation laws*, SIAM Journal on Scientific Computing, 38 (2016), pp. A3046–A3069.

[44] M. Li, Y. Cheng, J. Shen, and X. Zhang, *A bound-preserving high order scheme for variable density incompressible Navier-Stokes equations*, Journal of Computational Physics, 425 (2021), p. 109906.

[45] M. Li, P. Guyenne, F. Li, and L. Xu, *A positivity-preserving well-balanced central discontinuous Galerkin method for the nonlinear shallow water equations*, Journal of Scientific Computing, 71 (2017), pp. 994–1034.

[46] D. Ling, J. Cheng, and C.-W. Shu, *Conservative high order positivity-preserving discontinuous Galerkin methods for linear hyperbolic and radiative transfer equations*, Journal of Scientific Computing, 77 (2018), pp. 1801–1831.

[47] H. Liu and J. Yan, *The direct discontinuous Galerkin (DDG) methods for diffusion problems*, SIAM Journal on Numerical Analysis, 47 (2009), pp. 675–698.

[48] X.-D. Liu, S. Osher, and T. Chan, *Weighted essentially non-oscillatory schemes*, Journal of Computational Physics, 115 (1994), pp. 200–212.

[49] X.-D. Liu and E. Tadmor, *Third order nonoscillatory central scheme for hyperbolic conservation laws*, Numerische Mathematik, 79 (1998), pp. 397–425.

[50] S. A. Moe, J. A. Rossmanith, and D. C. Seal, *Positivity-preserving discontinuous Galerkin methods with Lax–Wendroff time discretizations*, Journal of Scientific Computing, 71 (2017), pp. 44–70.

[51] C. Munz, *On the comparison and construction of two-step schemes for the Euler equations*, Notes on Numerical Fluid Mechanics, 14 (1986), pp. 195–217.

[52] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, Journal of Computational Physics, 87 (1990), pp. 408–463.

[53] L. Pan, K. Xu, Q. Li, and J. Li, *An efficient and accurate two-stage fourth-order gas-kinetic scheme for the Euler and Navier–Stokes equations*, Journal of Computational Physics, 326 (2016), pp. 197–221.

[54] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM Journal on Numerical Analysis, 28 (1991), pp. 133–140.

[55] T. Qin and C.-W. Shu, *Implicit positivity-preserving high-order discontinuous Galerkin methods for conservation laws*, SIAM Journal on Scientific Computing, 40 (2018), pp. A81–A107.

[56] J. Qiu, *Hermite WENO schemes with Lax-Wendroff type time discretizations for Hamilton-Jacobi equations*, Journal of Computational Mathematics, (2007), pp. 131–144.

[57] ——, *A numerical comparison of the Lax–Wendroff discontinuous Galerkin method based on different numerical fluxes*, Journal of Scientific Computing, 30 (2007), pp. 345–367.

[58] 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 (2005), pp. 4528–4543.

[59] J. Qiu and C.-W. Shu, *On the construction, comparison, and local characteristic decomposition for high-order central WENO schemes*, Journal of Computational Physics, 183 (2002), pp. 187–209.

[60] ——, *Finite difference WENO schemes with Lax–Wendroff-type time discretizations*, SIAM Journal on Scientific Computing, 24 (2003), pp. 2185–2198.

[61] ——, *Hermite WENO schemes and their application as limiters for Runge–Kutta discontinuous Galerkin method: one-dimensional case*, Journal of Computational Physics, 193 (2004), pp. 115–135.

[62] W. H. Reed and T. R. Hill, *Triangular mesh methods for the neutron transport equation*, tech. rep., Los Alamos Scientific Lab., N. Mex.(USA), 1973.

[63] F. Renac, *Stationary discrete shock profiles for scalar conservation laws with a discontinuous Galerkin method*, SIAM Journal on Numerical Analysis, 53 (2015), pp. 1690–1715.

[64] G. R. Richter, *An optimal-order error estimate for the discontinuous Galerkin method*, Mathematics of Computation, 50 (1988), pp. 75–88.

[65] P. L. Roe, *Approximate Riemann solvers, parameter vectors, and difference schemes*, Journal of Computational Physics, 27 (1978), pp. 1–31.

[66] D. C. Seal, Q. Tang, Z. Xu, and A. J. Christlieb, *An explicit high-order single-stage single-step positivity-preserving finite difference WENO method for the compressible Euler equations*, Journal of Scientific Computing, 68 (2016), pp. 171–190.

[67] L. I. Sedov, *Similarity and dimensional methods in mechanics*, Academic Press, New York, 1959.

[68] J. Shi, C. Hu, and C.-W. Shu, *A technique of treating negative weights in WENO schemes*, Journal of Computational Physics, 175 (2002), pp. 108–127.

[69] C.-W. Shu, *Tvb uniformly high-order schemes for conservation laws*, Mathematics of Computation, 49 (1987), pp. 105–121.

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

[71] ——, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, B. Cockburn, C. Johnson, C.-W. Shu and E. Tadmor (Editor: A. Quarteroni), Lecture Notes in Mathematics, volume 1697, Springer, Berlin, (1998), pp. 325–432.

[72] ——, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes*, Acta Numerica, 29 (2020), pp. 701–762.

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

[74] ——, *Efficient implementation of essentially non-oscillatory shock-capturing schemes, II*, Journal of Computational Physics, 83 (1989), pp. 32–78.

[75] J. Smoller, *Shock waves and reaction—diffusion equations*, Springer-Verlag, New York, 1983.

[76] R. J. Spiteri and S. J. Ruuth, *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM Journal on Numerical Analysis, 40 (2002), pp. 469–491.

[77] V. A. Titarev and E. F. Toro, *ADER: Arbitrary high order Godunov approach*, Journal of Scientific Computing, 17 (2002), pp. 609–618.

[78] E. F. Toro, *Riemann solvers and numerical methods for fluid dynamics: A practical introduction*, Springer Science & Business Media, 2013.

[79] J. Von Neumann and R. D. Richtmyer, *A method for the numerical calculation of hydrodynamic shocks*, Journal of Applied Physics, 21 (1950), pp. 232–237.

[80] C. Wang, X. Zhang, C.-W. Shu, and J. Ning, *Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations*, Journal of Computational Physics, 231 (2012), pp. 653–665.

[81] Y. Xing and X. Zhang, *Positivity-preserving well-balanced discontinuous Galerkin methods for the shallow water equations on unstructured triangular meshes*, Journal of Scientific Computing, 57 (2013), pp. 19–41.

[82] Y. Xing, X. Zhang, and C.-W. Shu, *Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations*, Advances in Water Resources, 33 (2010), pp. 1476–1493.

[83] T. Xiong, J.-M. Qiu, and Z. Xu, *High order maximum-principle-preserving discontinuous Galerkin method for convection-diffusion equations*, SIAM Journal on Scientific Computing, 37 (2015), pp. A583–A608.

[84] Z. Xu and C.-W. Shu, *Local characteristic decomposition free high order finite difference WENO schemes for hyperbolic systems endowed with a coordinate system of Riemann invariants*, submitted.

[85] ——, *On the conservation property of positivity-preserving discontinuous Galerkin methods for stationary hyperbolic equations*, submitted.

[86] ——, *High order conservative positivity-preserving discontinuous Galerkin method for stationary hyperbolic equations*, Journal of Computational Physics, 466 (2022), p. 111410.

[87] ——, *Third order maximum-principle-satisfying and positivity-preserving Lax-Wendroff discontinuous Galerkin methods for hyperbolic conservation laws*, Journal of Computational Physics, 470 (2022), p. 111591.

[88] Z. Xu, Y. Yang, and H. Guo, *High-order bound-preserving discontinuous Galerkin methods for wormhole propagation on triangular meshes*, Journal of Computational Physics, 390 (2019), pp. 323–341.

[89] Y. Yang, D. Wei, and C.-W. Shu, *Discontinuous Galerkin method for Krause's consensus models and pressureless Euler equations*, Journal of Computational Physics, 252 (2013), pp. 109–127.

[90] D. Yuan, J. Cheng, and C.-W. Shu, *High order positivity-preserving discontinuous Galerkin methods for radiative transfer equations*, SIAM Journal on Scientific Computing, 38 (2016), pp. A2987–A3019.

[91] M. Zhang, J. Cheng, and J. Qiu, *High order positivity-preserving discontinuous Galerkin schemes for radiative transfer equations on triangular meshes*, Journal of Computational Physics, 397 (2019), p. 108811.

[92] X. Zhang, *On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier–Stokes equations*, Journal of Computational Physics, 328 (2017), pp. 301–343.

[93] X. Zhang and C.-W. Shu, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, Journal of Computational Physics, 229 (2010), pp. 3091–3120.

[94] ——, *On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, Journal of Computational Physics, 229 (2010), pp. 8918–8934.

[95] ——, *Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms*, Journal of Computational Physics, 230 (2011), pp. 1238–1248.

[96] 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 (2013), pp. 295–316.

[97] X. Zhao, Y. Yang, and C. E. Seyler, *A positivity-preserving semi-implicit discontinuous Galerkin scheme for solving extended magnetohydrodynamics equations*, Journal of Computational Physics, 278 (2014), pp. 400–415.

[98] F. Zheng, C.-W. Shu, and J. Qiu, *A high order conservative finite difference scheme for compressible two-medium flows*, Journal of Computational Physics, 445 (2021), p. 110597.

[99] J. Zhu and C.-W. Shu, *A new type of multi-resolution WENO schemes with increasingly higher order of accuracy*, Journal of Computational Physics, 375 (2018), pp. 659–683.