2018

Invariant-region-preserving discontinuous Galerkin methods for systems of hyperbolic conservation laws

Yi Jiang

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Mathematics Commons

Recommended Citation


https://lib.dr.iastate.edu/etd/16599

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Invariant-region-preserving discontinuous Galerkin methods for systems of hyperbolic conservation laws

by

Yi Jiang

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY

Major: Applied Mathematics

Program of Study Committee:
Hailiang Liu, Major Professor
Lisheng Steven Hou
James Rossmanith
Paul Sacks
Sung-Yell Song

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2018

Copyright © Yi Jiang, 2018. All rights reserved.
DEDICATION

To my grandmother Zhifen Jiang.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>x</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xii</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Hyperbolic conservation laws</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Solution features</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Numerical challenges</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Numerical methods</td>
<td>3</td>
</tr>
<tr>
<td>1.2.1 Discontinuous Galerkin (DG) methods</td>
<td>5</td>
</tr>
<tr>
<td>1.2.2 Strong stability preserving (SSP) Runge-Kutta (RK) methods</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Invariant-region-preserving (IRP) schemes</td>
<td>6</td>
</tr>
<tr>
<td>1.3.1 Motivation and related work</td>
<td>6</td>
</tr>
<tr>
<td>1.3.2 Ingredients to enforce the IRP property</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Thesis organization</td>
<td>10</td>
</tr>
<tr>
<td>CHAPTER 2. AN IRP LIMITER TO DG METHODS FOR COMPRESSIBLE EULER EQUA-</td>
<td>12</td>
</tr>
<tr>
<td>TIONS</td>
<td></td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>12</td>
</tr>
<tr>
<td>2.2 The limiter</td>
<td>14</td>
</tr>
<tr>
<td>2.2.1 Averaging is a contraction</td>
<td>14</td>
</tr>
<tr>
<td>2.2.2 Reconstruction</td>
<td>15</td>
</tr>
<tr>
<td>2.2.3 Algorithm</td>
<td>19</td>
</tr>
<tr>
<td>2.3 Numerical tests</td>
<td>20</td>
</tr>
<tr>
<td>2.4 Conclusion and future work</td>
<td>24</td>
</tr>
<tr>
<td>References</td>
<td>24</td>
</tr>
<tr>
<td>CHAPTER 3. AN IRP LIMITER FOR DG SCHEMES TO ISENTROPIC EULER EQUA-</td>
<td>26</td>
</tr>
<tr>
<td>TIONS</td>
<td></td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>26</td>
</tr>
<tr>
<td>Abstract</td>
<td>26</td>
</tr>
</tbody>
</table>
 CHAPTER 4. IRP DG METHODS FOR MULTI-DIMENSIONAL HYPERBOLIC CONSERVATION LAW SYSTEMS, WITH AN APPLICATION TO COMPRESSIBLE EULER EQUATIONS

Abstract

4.1 Introduction

4.2 The IRP limiter

4.2.1 The limiter

4.2.2 Algorithm

4.3 IRP DG schemes

4.3.1 One-dimensional case

4.3.2 Multi-dimensional case

4.4 Application to compressible Euler equations

4.4.1 1D case

4.4.2 2D case

4.5 Numerical tests

4.6 Concluding remarks

Appendix A. Proof of Lemma 4.2.1 for compressible Euler equations.

Appendix B. Proof of Lemma 4.3.2

References

CHAPTER 5. IMPLEMENTATION DETAILS

5.1 Mesh generation and solution space

5.2 Scheme formulation

5.2.1 Local formulation

5.2.2 Global formulation

5.3 DG implementation

5.3.1 Method of line

5.3.2 Time discretization

5.3.3 IRP limiter

CHAPTER 6. SUMMARY AND DISCUSSION
BIBLIOGRAPHY ................................................................. 117

APPENDIX. A MATLAB SCRIPT ........................................... 123
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Accuracy test with $P^2$-DG scheme for Example 1.</td>
<td>22</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Accuracy test with $P^3$-DG scheme for Example 1.</td>
<td>22</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Accuracy of the $L^2$ projection using $P^1$ approximation.</td>
<td>50</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Accuracy of the $L^2$ projection using $P^2$ approximation.</td>
<td>51</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Accuracy test with $P^1$-DG scheme for (3.1.2).</td>
<td>51</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Accuracy test with $P^2$-DG scheme for (3.1.2).</td>
<td>52</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Accuracy test with $P^1$-DG scheme for (3.1.3).</td>
<td>52</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>Accuracy test with $P^2$-DG scheme for (3.1.3).</td>
<td>53</td>
</tr>
<tr>
<td>Table 3.7</td>
<td>Convergence test</td>
<td>53</td>
</tr>
<tr>
<td>Table 3.8</td>
<td>Convergence test</td>
<td>53</td>
</tr>
<tr>
<td>Table 3.9</td>
<td>Choice of $h$ and corresponding parameter values for unbounded $C_4.$</td>
<td>61</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Accuracy of density function in Example 1 using $P^1$-DG scheme.</td>
<td>92</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Accuracy test for density function in Example 1 using $P^2$-DG scheme.</td>
<td>93</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Some parameter values when the limiter is first called in the $P^1$-DG scheme solving Example 1.</td>
<td>93</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Some parameter values when the limiter is first called in the $P^2$-DG scheme solving Example 1.</td>
<td>93</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>Accuracy test for density function in Example 2 using $P^1$-DG scheme.</td>
<td>94</td>
</tr>
<tr>
<td>Table 4.6</td>
<td>Accuracy test for density function in Example 2 using $P^2$-DG scheme.</td>
<td>94</td>
</tr>
<tr>
<td>Table 4.7</td>
<td>Some parameter values when the limiter is first called in the $P^1$-DG scheme solving Example 2.</td>
<td>95</td>
</tr>
</tbody>
</table>
Table 4.8  Some parameter values when the limiter is first called in the $P^2$-DG scheme solving Example 2.
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>Lax shock tube problem. Exact solution (solid line) vs numerical solution (dots); Top: With positive-preserving limiter; Bottom: With IRP limiter</td>
<td>23</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Shu-Osher problem. Exact solution (solid line) vs numerical solution (dots); Left: With positive-preserving limiter; Right: With IRP limiter</td>
<td>23</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Invariant region for p-system</td>
<td>32</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Shock-rarefaction wave problem. Exact solution (solid line) vs numerical solution (dots); Top: DG without limiter; Bottom: DG with limiter</td>
<td>55</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Rarefaction-shock wave problem. Exact solution (solid line) vs numerical solution (dots); Top: DG without limiter; Bottom: DG with limiter</td>
<td>56</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Sod shock tube problem. Exact solution (solid line) vs numerical solution (dots); Left: With positive-preserving limiter; Right: With invariant-region-preserving limiter</td>
<td>96</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>1D double rarefaction problem. Exact solution (solid line) vs numerical solution (dots); Left: with positivity-preserving limiter; Right: with IRP limiter</td>
<td>98</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Configuration 2. Contour plot of numerical solution of density with 30 contour levels. Left: with positive-preserving limiter; Right: with invariant-region-preserving limiter</td>
<td>99</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>Configuration 6. Contour plot of numerical solution of density with 30 contour levels. Left: with positive-preserving limiter; Right: with invariant-region-preserving limiter</td>
<td>100</td>
</tr>
</tbody>
</table>
Figure 4.5  Zoom-in plot of contour plots of configuration 6. Left: with positive-preserving limiter; Right: with invariant-region-preserving limiter. Top: lower-left interface; Bottom: lower-right interface.
ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of pursuing the doctoral degree at Iowa State University.

First and foremost, my sincere appreciation goes to my major Professor, Dr. Hailiang Liu. His pure passion for the math and life has always been infectious. My spirit got refreshed every time after I talked to him no matter how frustrated I might feel before. His two principles in writing, be simple and be clear, always remind me of the essence of doing the math. There have been many ups and downs in the journey of doing the research, but thanks to Dr. Liu’s constant patience, insightful suggestions and warm encouragement, I got to see and enjoy the fun and the beauty in math.

I would like to thank Dr. Paul Sacks and Dr. James Rossmanith for their teaching and support through my graduate study. Their availability for meeting in the office and answering questions have always been helpful and appreciated. Thanks also go to Dr. L. Steven Hou and Dr. Sung-Yell Song for their willingness being my committee members as well as their constructive comments on my work and thesis writing. In addition, I want to especially thank Dr. Glenn Luecke. Without taking his high-performance computing course, I would not be able to finish the numerical implementations (especially for two-dimensional problems) needed in this thesis work as planned.

I want to thank Dr. Heather Bolles, Dr. Elgin Johnston and Dr. Jennifer Newman for their valuable suggestions together with their trust and support in my teaching. Also, I’m indebted to all the staff in the main office of the math department. Special thanks go to Melanie Erickson and Ellen Olson. Words would just fail to express how nice and helpful you’ve been to me.

Last but not the least, I’m indebted to my beloved families and friends for their tremendous support and endless love. I own my most sincere gratitude to my parents, Yuanshou Wang and Min Jiang. It is their unconditional love that supports me to pursue my dream. And my dear husband
Keguo and daughter Grace, I cannot imagine a sweeter school life than having you around. You are the gifts from God.
This thesis is aimed at developing high order invariant-region-preserving (IRP) discontinuous Galerkin (DG) schemes solving hyperbolic conservation law systems. In particular, our focus is on designing an IRP limiter to modify the numerical solution and identifying sufficient conditions for IRP DG schemes by extending Zhang and Shu’s work on positivity-preserving schemes for conservation laws [X. Zhang and C.-W. Shu., Journal of Computational Physics, 2010]. The one-dimensional model problems investigated are the p-system for $2 \times 2$ system and compressible Euler equations for $3 \times 3$ case. The feature that the invariant regions for these systems can be expressed by convex or concave functions of the solution to the system allows us to construct the IRP limiter in an explicit form, which therefore is easy for computer implementations. Rigorous analysis is presented to show that the IRP limiter does not destroy the order of approximation accuracy for smooth solutions, provided that the cell average of numerical solution is away from the boundary of the invariant region. For arbitrarily high order DG schemes solving hyperbolic conservation law systems, sufficient conditions are identified for cell averages to remain in the invariant region. We further extended these results to general multidimensional hyperbolic conservation law systems as long as (i) the system admits a global invariant region and (ii) the corresponding one-dimensional projected system shares the same invariant region. An application of these results has been investigated for two dimensional compressible Euler equations. Numerical tests on model problems have shown that the IRP limiter maintains the order of approximation for smooth solutions and helps damp oscillations near discontinuities. In addition, we designed second and third order IRP DG schemes for the viscous p-system by using the direct DG (DDG) diffusive flux proposed in [H. Liu and J. Yan., Communications in Computational Physics, 2010], considering the p-system and its viscous counterpart share the same invariant region. Numerical tests validate the desired
properties of the IRP limiter for the viscous problem. The convergence of the viscous profiles to the entropy solution is illustrated from a numerical point of view.
CHAPTER 1. INTRODUCTION

We are interested in hyperbolic conservation law systems. Such systems have a wide application in science and engineering such as shallow water equations governing flows in rivers, coastal regions or atmosphere and compressible Euler equations in gas dynamics. Since, in general, the exact solutions of these equations cannot be obtained directly, it is important to construct stable, accurate, and efficient numerical methods to solve the systems and capture the features of the solutions. In this thesis, we focus on the hyperbolic systems of conservation laws that admit global invariant regions and explore the high order accurate invariant-region-preserving (IRP) methods through the design of a new limiter called IRP limiter.

1.1 Hyperbolic conservation laws

A general system of conservation laws is of the form

$$\partial_t w + \sum_{j=1}^{d} \partial_x F_j(w) = 0, \quad x \in \mathbb{R}^d, \quad t > 0,$$

where $w : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^l$ is the unknown vector and $F_j(w) : \mathbb{R}^l \to \mathbb{R}^l$, for $j = 1, \ldots, d$, are flux functions. This system is hyperbolic if for all $w$, any linear combination of $\{\partial F_j / \partial w\}$, the Jacobian matrix of each flux function, has $l$ real eigenvalues and $l$ linearly independent corresponding eigenvectors. See e.g. [12, 17]. We consider the initial value problem for (1.1.1) subject to the initial data $w(x, 0) = w_0(x)$.

1.1.1 Solution features

An important feature of nonlinear systems of the form (1.1.1) is that discontinuities may develop in the solutions in finite time even for smooth initial data. Hence solutions in the classical sense
may fail to exist and we must consider weak solutions that satisfy (1.1.1) in the following sense:

\[
\int_0^\infty \int_{\mathbb{R}^d} \left( w \partial_t \phi dx + \sum_{j=1}^d F_j(w) \partial_{x_j} \phi \right) dx dt + \int_{\mathbb{R}^d} w_0(x) \phi(x,0) dx = 0,
\]

for all test functions \( \phi \in C_0^1(\mathbb{R}^d \times [0, \infty)) \), where \( C_0^1 \) is the space of functions that are continuously differentiable with compact support. However, weak solutions may not be unique. For example, infinitely many weak solutions can be found for Burgers’ equation corresponding to the same initial data as shown in [36]. In order to single out a unique physically relevant solution, i.e, the entropy solution, additional conditions must be imposed.

One way to obtain the entropy solution of the system (1.1.1) is to let \( w = \lim_{\epsilon \to 0} w^\epsilon \) a.e., where \( w^\epsilon \) is the solution of the corresponding viscous equations

\[
\partial_t w^\epsilon + \sum_{j=1}^d \partial_{x_j} F_j(w^\epsilon) = \epsilon \Delta w^\epsilon, \quad x \in \mathbb{R}^d, \quad t > 0.
\]

However, the parabolic system above is not easy to work with. An alternative approach to characterizing the entropy solution relies on the concept of entropy function. A weak solution of (1.1.1) is an entropy solution if \( w \) satisfies the entropy inequality

\[
\partial_t \eta(w) + \sum_{j=1}^d \partial_{x_j} \Psi_j(w) \leq 0,
\]

in the distributional sense, for every convex entropy function \( \eta \) and corresponding entropy fluxes \( \{\Psi_j\}_{j=1}^d \), where \( \nabla_w \eta^\top \frac{\partial F_j}{\partial w} = \nabla_w \Psi_j^\top \), for \( j = 1, \cdots, d \). Note that for smooth \( w \), the entropy pairs satisfy the conservation law

\[
\partial_t \eta(w) + \sum_{j=1}^d \partial_{x_j} \Psi_j(w) = 0, \quad x \in \mathbb{R}^d, \quad t > 0.
\]

As mentioned in [36], (1.1.2) is also useful in numerical analysis since if a discrete form of entropy inequality is found to hold for some numerical methods, it can be shown that the method converges to the entropy solution.

1.1.2 Numerical challenges

Since the flux functions in (1.1.1) are usually nonlinear functions of \( w \), it is impossible to derive exact solutions to such systems in general. Hence there is need to develop numerical methods to
find approximate solutions, while new difficulties show up due to the solution features described above.

For first-order methods, the solutions are usually smeared around discontinuities due to the large amount of “numerical viscosity”, while for regular second-order methods, dispersive effects are produced and lead to oscillations in the solutions. People are then interested in the high resolution methods, which are expected to be high order accurate in space in smooth regions, capture discontinuities in sharp resolution, and also produce solutions free from spurious oscillations near the neighborhood of discontinuities. Besides, it’s not trivial to conclude that the approximating solutions indeed converge or converge to the correct physically-relevant weak solution, hence it calls for the study of the nonlinear stability and discrete form of the entropy condition when designing numerical schemes.

1.2 Numerical methods

A wide variety of numerical methods have been developed to solve the hyperbolic conservation laws in the literature.

The early approaches lie in the family of finite difference (FD) methods, where the spatial and temporal derivatives in the equations are simply approximated by appropriate finite differences. For example, Lax-Friedrichs’ method in [16, 35] is a central method where a spatial average in the previous time step is used in the Euler forward time discretization to achieve conditional stability, while the first-order upwind method [11] is a one-sided method taking advantage of the fact that the information propagates in the direction of characteristic curves. Both of these two methods are first-order accurate and suffer from smearing out the solution near the discontinuities due to the numerical dissipation. The Lax-Wendroff method in [34] is a second-order method derived by cutting off a Taylor expansion and using central differences for spatial derivatives approximations. Some related methods include the Beam-Warming method [1], MacCormack method [40] and Richtmyer method [44]. All these methods are second order accurate and produce oscillations due to dispersive effects. The Lax-Wendroff theorem in [34] says that only a conservative scheme, when converging,
can converge to the correct weak solution. A second-order conservative FD scheme which offers high resolution is a central scheme introduced by Nessyahu and Tadmor in [41], where the authors used the Lax-Friedrichs method as the building block and then used the piecewise-linear MUSCL-type interpolants [49] to compensate the excessive numerical viscosity. This approach was further improved by Kurganov et al. [32, 29, 31, 33] by designing second or third order schemes that admit a semi-discrete form and enjoy smaller numerical dissipations. More extensions of Nessyahu-Tadmor schemes can be found in [39, 28, 2, 30] and the references therein.

Another important class of numerical methods is the group of finite volume (FV) methods, which are based on an integral form of the conservation laws. Rather than point values at grid points that are approximated in FD methods, the integrals of unknowns over each cell volume (cell averages) are evolved at each time step in FV methods. The two categories are usually interlaced, but an important property of FV methods is that they are conservative in nature since the change of the cell average depends only on the fluxes over the cell edges and therefore have more advantages in capturing shock waves accurately. One of the most fundamental methods is Godunov's method, which can be viewed as a generalization of the upwind method for nonlinear systems of conservation laws. The method, proposed by Godunov in [18], uses the piecewise constant function for approximation and evolves the numerical solution by solving Riemann problems at the cell interfaces forward in time exactly. Godunov’s method is only first order, but it can be viewed as a forerunner of a variety of high resolution FV methods. For example, the slope-limiter methods improve the order of approximation by using more accurate representation of the numerical solution such as piecewise linear function used in Monotonic Upstream-centered Scheme for Conservation Laws (MUSCL) [49], quadratics used in piecewise parabolic method (PPM) [10] and high order reconstructions introduced in essentially non-oscillatory (ENO) and weighted essentially non-oscillatory (WENO) methods [22, 45, 46, 38].

Except for FD and FV methods, there are also other families of methods such as finite element methods, spectral methods and etc. In particular, in recent years, a rapid development of high resolution methods for conservation laws has been prompted for discontinuous Galerkin (DG) methods,
which can be classified as finite element methods but using discontinuous basis functions for both solutions and test functions. The DG method was first introduced by Reed and Hill in [43] and developed by Cockburn and Shu with their collaborators in a series of papers [8, 7, 6, 4, 9]. The advantages of DG methods as noted in [5] include that (1) the arbitrary high order of accuracy can be achieved by choosing the degree of the approximating polynomials appropriately; (2) the methods are highly parallelizable; (3) complicated geometries can be handled and treatment required for the boundary conditions is simple; (4) adaptivity strategies can be easily handled. One of the focus of the research in the DG community lies in finding a suitable time discretization to obtain a stable and high order accurate method. A key development is the Runge-Kutta Discontinuous Galerkin (RKDG) method by Cockburn and Shu in [8, 7], where the Strong Stability Preserving (SSP) high order RK explicit time discretization in [19] (termed Total Variation Diminishing (TVD) in [45]) is adopted and contributes to the stability result for the fully discretized schemes.

Our work in this thesis is mainly under the framework of DG methods using SSP RK methods in time discretization, hence we give a brief description of two methods in the following.

1.2.1 Discontinuous Galerkin (DG) methods

To solve (1.1.1) over a computational cell \( K \) with a general DG scheme is to find a vector solution \( \mathbf{w}_h \), whose components are in \( V_h \), a finite element space of piecewise polynomials of degree \( k \) in \( K \), i.e.,

\[
V_h = \{ v \in L^2(\Omega) : v\big|_K \in \mathbb{P}^k(K) \}, \quad \Omega = \bigcup K,
\]

such that

\[
\int_K \partial_t (\mathbf{w}_h) \phi dx - \int_K \mathbf{F}(\mathbf{w}_h) \cdot \nabla_x \phi dx + \sum_{i=1}^Q \int_{e_i^K} \hat{F}(\mathbf{w}^-_h, \mathbf{w}^+_h, \nu^i) \phi ds = 0, \quad \forall \phi \in V_h,
\]

where \( \mathbf{F} = (F_1, \cdots, F_d)^T \), \( e_i^K \) is the \( i \)-th edge (or surface) of \( K \), \( \nu^i \) is the normal vector on \( e_i^K \), \( i = 1, \cdots, Q \), \( \mathbf{w}^-_h \) and \( \mathbf{w}^+_h \) denote the approximations to \( \mathbf{w}_h \) on the edge of \( K \) from interior and exterior of \( K \) respectively, and \( \hat{F} \) is an admissible numerical flux (approximate or exact Riemann solvers). This leads to an ODE system for the unknown coefficients of the numerical solution.
1.2.2 Strong stability preserving (SSP) Runge-Kutta (RK) methods

The SSP methods developed in [19] are to solve ODE systems

$$\frac{d}{dt} \psi = L(\psi), \quad (1.2.1)$$

which are usually obtained after the spatial discretization is applied to the hyperbolic conservation laws. For example, for semi-discrete DG schemes, $\psi$ consists of unknown coefficients of the spatial basis and $L$ is the corresponding spatial operator.

As noted in [19], the essence of explicit SSP methods is that when a time step is restricted appropriately, the methods maintain the strong stability in the same norm as the first order forward-Euler time discretization does, which therefore plays an important role in solving hyperbolic conservation laws especially when the discontinuous solutions are involved.

In this thesis, we will mainly use the third order SSP RK method introduced in [45, 19] for numerical implementations. The method is given by

$$\psi^{(1)} = \psi^n + \Delta t L(\psi^n),$$

$$\psi^{(2)} = \frac{3}{4} \psi^n + \frac{1}{4} \psi^{(1)} + \frac{1}{4} \Delta t L(\psi^{(1)}),$$

$$\psi^{n+1} = \frac{1}{3} \psi^n + \frac{2}{3} \psi^{(2)} + \frac{2}{3} \Delta t L(\psi^{(2)}),$$

where $\psi^n$ represents the solution of the ODE (1.2.1) at $n$-th time step and $\psi^{(i)}$ is the approximation in the $i$-th intermediate stage, $i = 1, 2$.

1.3 Invariant-region-preserving (IRP) schemes

1.3.1 Motivation and related work

It is known that the entropy solution to the initial value problem of the scalar conservation law

$$\partial_t w + \partial_x f(w) = 0, \quad x \in \mathbb{R}, t > 0,$$

$$w(x, 0) = w_0 \quad (1.3.1)$$
satisfies a strict maximum principle

\[ m \leq w(x, t) \leq M, \quad \forall x, t, \]

where \( m = \min_x w_0(x) \) and \( M = \max_x w_0(x) \). This property is also desired in numerical schemes to approximate (1.3.1); in general considerations, quantities such as density, volume ratio, etc. need to stay positive in order to be physically meaningful. For the system of conservation laws, the notion of maximum principle no longer applies. Instead, the notion of invariant region needs to be considered.

The system (1.1.1) is said to have an invariant region \( \Sigma \) if when the initial data is in the region \( \Sigma \), the solution will remain in \( \Sigma \) for all \( t > 0 \). Based on Hoff’s results in [23], the invariant region for hyperbolic conservation law systems is necessarily convex. Hence in this thesis, we consider convex invariant set in phase space. For one dimensional \( 2 \times 2 \) systems, i.e. (1.1.1) with \( d = 1 \), \( l = 2 \), the invariant region can be described by two Riemann invariants, see. e.g., [3, 23, 47, 15]. For general conservation law systems with \( l \geq 3 \), it is an open question to identify a global invariant region. One important example studied in this thesis is the compressible Euler equations of gas dynamics. A natural condition for its solution is the positivity of density and pressure, and the minimum principle for the specific entropy [48]. In the one dimensional case, the Euler equation has the following form

\[ \partial_t w + \partial_x f(w) = 0, \quad t > 0, \quad x \in \mathbb{R}, \quad (1.3.2) \]

with \( w = (\rho, m, E)\top \),

\[ f(w) = (m, \rho u^2 + p, (E + p) u)\top, \quad E = \frac{1}{2} \rho u^2 + \frac{p}{\gamma - 1}, \quad (1.3.3) \]

where \( \rho \) is the density, \( u \) is the velocity, \( m = \rho u \) is the momentum, \( E \) is the total energy, \( p \) is the pressure, and \( \gamma \) is the ratio of specific heats for the gas/fluid (for most gases, \( 1 < \gamma < 3 \)). The corresponding invariant region is the following set

\[ G = \{ w \mid \rho > 0, \quad p > 0, \quad s \geq s_0 \}, \quad (1.3.4) \]
where \( s = \log(p/\rho^\gamma) \) is the specific entropy, and \( s_0 := \inf_x s(w_0(x)) \).

From the numerical perspective, invariant regions are also important for guaranteeing the stability of numerical schemes in the sense that enforcing the solutions to remain in certain physical domains could keep them from blowing up in the algorithms. A more common study in the literature is on positivity-preserving schemes, where, instead of the whole invariant region, the positivity of some physical quantities are considered, such as the water height in shallow water equations and density and pressure in compressible Euler equations. Early work such as [13], [37] and the references therein investigated positivity-preserving schemes which are up to second order only. In [42], third-order positivity-preserving finite volume schemes for Euler equations are constructed based on positivity-preserving properties by the corresponding first-order schemes for both density and pressure. Following [42], positivity-preserving high order DG schemes for compressible Euler equations were first introduced in [57], where the limiter is used to modify the solution polynomials to achieve the positivity requirement and the parameters are derived separately for the density and pressure functions. This limiter is a generalization of the work of same authors in [56] for scalar conservation law, which is shown to maintain the original high order accuracy of the numerical approximation (see [55]). Related developments for shallow water equations can be found in [52, 51, 50]. Another positivity-preserving technique is the flux limiting approach which utilizes a convex combination of the first order monotone flux and the high order flux to achieve both the positivity-preserving property and high order accuracy under certain conditions. See [24, 53] and references therein. As pointed out in [54], such technique has an advantage when implemented in the framework of finite difference method, while one big challenge is to show that the high order accuracy of the solution is not compromised after the flux limiter is applied.

Nevertheless, except for the positivity property, pointwise bounds regarding some other quantities also need to be considered. For example, as noted in [37], enforcing the maximum(or minimum) entropy principle numerically for Euler equations, also a part of the invariant region as mentioned before, is important for convergence studies. It has also been evidenced in [58] that adding this property is helpful for reducing spurious oscillations, compared to the schemes with positivity-
preserving property alone. In fact, as we can see in Euler equations and examples in Section 3.2 and Section 4.2.1, the positivity of the physical quantities are usually already contained in the invariant region, therefore it would be natural to study the numerical schemes that can preserve the whole invariant region as long as it exists.

In [23], Hoff studied the necessary and sufficient conditions for a region to be invariant for (Glimm) solutions of the system of conservation laws and made remarks on such regions for some finite difference schemes including the Lax-Friedrichs scheme, Godunov scheme and one modified Osher scheme. In [14, 15], the IRP property under the Lax-Friedrich schemes was further studied by Frid for nonlinear systems of conservation laws in several space variables with a known invariant region. In the context of continuous finite elements, the IRP property has been studied by Guermond and Popov [21] using the first order approximation to solve general hyperbolic conservation law systems, and then in [20] using the second order approximation with convex limiting to solve compressible Euler equations. However, no higher order IRP numerical schemes have been investigated, to our best knowledge.

The IRP schemes developed in this thesis can be arbitrarily high order accurate and are mainly discussed under the DG framework, although the IRP limiter can be applied to any finite-volume-type schemes. This work can be viewed as an extension of earlier works on positivity-preserving schemes for compressible Euler equations in [42, 57]. In the next subsection, we will explain how the IRP property is enforced in our work by following this research line.

1.3.2 Ingredients to enforce the IRP property

Analogous to the development of positivity-preserving schemes in [57, 58], two main tasks in constructing IRP finite-volume type schemes in this thesis are,

• given \( w^n(x) \in \Sigma \) for all \( x \), how to find sufficient conditions such that the evolved cell average \( \bar{w}_K^{n+1} \in \Sigma_0 \);

• given \( \bar{w}_K^{n+1} \in \Sigma_0 \), how to modify the numerical solution \( w^{n+1}(x) \) such that it lies in \( \Sigma \) for all \( x \in K \);
Here $\Sigma$ is the desired invariant region, $\Sigma_0$ is its interior, $K$ is the computational cell and $n$ indicates $n$-th time step.

The first ingredient we consider is a one-dimensional IRP numerical flux, with which the first-order finite volume type scheme has the IRP property under certain CFL condition. We have proved in Section 4.3.1 that four popular fluxes including the Godunov, Lax-Friedrichs, HLL (Harten-Lax-van Leer) and HLLC (Harten-Lax-van Leer-Contact) fluxes are all IRP fluxes. We then find sufficient conditions to keep the cell averages in $\Sigma$ by repeated convex combinations, in the same way as that has been well established for positivity-preserving schemes, see e.g. [57, 58]. In the present setting, we use first order schemes with an IRP flux which can keep numerical solutions in $\Sigma$ as building blocks, and show that high order spatial discretization with forward Euler can be written as a convex combination of first order IRP schemes, thus will keep $\Sigma$ provided certain sufficient conditions are satisfied. The IRP limiter is then used to enforce the sufficient condition where the numerical solution needs to be within in $\Sigma$ on a test set. Our limiter modifies the polynomial solution through a linear convex combination of original approximation and the cell average as in [57, 58], yet with the main distinction that using a single uniform parameter defined explicitly.

Finally, we should mention that when the high order time discretization is utilized, it will keep the IRP property as long as it can be written as convex combinations of forward Euler, thanks to the convexity of the invariant region. One example of such time discretization is the high order SSP RK methods described in the previous section.

### 1.4 Thesis organization

The thesis is organized as follows. In Chapter 2, an IRP limiter for one-dimensional compressible Euler equations is introduced. By rewriting the convex set $G$ given in (1.3.4) into an equivalent set characterized by only convex or concave functions, we are able to derive the limiter in an explicit form, which is easy for computer implementation. Numerical experiments are implemented to show that the limiter preserves the high order of accuracy in general cases and helps reduce
the oscillations shown in some Riemann solutions. Chapter 3 is devoted to a focus on the one-dimensional $2 \times 2$ system which is in form of (1.1.1). For such a system, the invariant region can be described by two Riemann invariants. In particular, we study the model system, the $p$-system, as well as its corresponding viscous system. A rigorous analysis is given there to prove that the IRP limiter preserves the cell average and does not destroy the high order of accuracy for smooth solutions as long as the cell average stays away from the boundary of the invariant region.

We also identify sufficient conditions for IRP high order DG schemes solving the $p$-system and those for second and third order IRP direct DG schemes solving the corresponding viscous system. Numerical experiments are carried out to illustrate the properties of the limiter. In Chapter 4, we extend our results to the general multi-dimensional hyperbolic conservation law systems assuming (i) it features a global invariant region, which is characterized by convex functions of the unknown vector $\mathbf{w}$, and (ii) the corresponding one-dimensional projected system shares the same invariant region as the full multi-dimensional hyperbolic system does. An explicit IRP limiter is designed and proved to maintain high order accuracy of the approximation in general cases. A generic algorithm incorporating the IRP limiter is presented for high order finite-volume-type schemes. We also identify sufficient conditions under which arbitrarily high order DG schemes feature the desired IRP property for both one and multi-dimensional cases. These general results are further applied to two-dimensional compressible Euler equations with numerical validations. In Chapter 5, we discuss some implementation details for IRP DG schemes with the SSP RK3 method. Finally, we give concluding remarks in Chapter 6.
CHAPTER 2. AN IRP LIMITER TO DG METHODS FOR COMPRESSIBLE EULER EQUATIONS


Yi Jiang and Hailiang Liu

Abstract

We introduce an explicit invariant-region-preserving limiter applied to DG methods for compressible Euler equations. The invariant region considered consists of positivity of density and pressure and a maximum principle of a specific entropy. The modified polynomial by the limiter preserves the cell average, lies entirely within the invariant region and does not destroy the high order of accuracy for smooth solutions, as long as the cell average stays away from the boundary of the invariant region. Numerical tests are presented to illustrate the properties of the limiter. In particular, the tests on Riemann problems show that the limiter helps to damp the oscillations near discontinuities.

2.1 Introduction

We consider the one dimensional version of the compressible Euler equations for the perfect gas in gas dynamics:

\[ \mathbf{w}_t + F(\mathbf{w})_x = 0, \quad t > 0, \quad x \in \mathbb{R}, \]

\[ \mathbf{w} = (\rho, m, E)^	op, \quad F(\mathbf{w}) = (m, \rho u^2 + p, (E + p)u)^	op \]

with

\[ m = \rho u, \quad E = \frac{1}{2} \rho u^2 + \frac{p}{\gamma - 1}, \]
where $\gamma > 0$ is a constant ($\gamma = 1.4$ for the air), $\rho$ is the density, $u$ is the velocity, $m$ is the momentum, $E$ is the total energy and $p$ is the pressure; supplemented by initial data $w_0(x)$. For the associated entropy function $s = \log \left( \frac{p(x)}{\rho^\gamma(x)} \right)$, it is known that

$$A = \{(\rho, m, E)^\top, \rho > 0, \ p > 0, \ s \geq s_0\}$$

for any $s_0 \in \mathbb{R}$ is an invariant region in the sense that if $w_0(x) \in A$, then $w(x, t) \in A$ for all $t > 0$ (see e.g. [9, 3]). At numerical level this set is proved to be invariant by the first order Lax-Friedrichs scheme (see [1]), and by the first order Finite Element method (see [2]), in which a larger class of hyperbolic conservation laws is considered. It is difficult, if not impossible, to preserve such set by a high order numerical method unless some nonlinear limiter is imposed at each step while marching in time. In this work we design such a limiter.

In recent years an interesting mathematical literature has developed devoted to high order maximum-principle-preserving schemes for scalar conservation equations (see [12]) and positivity-preserving schemes for hyperbolic systems including compressible Euler equations (see e.g. [6, 13, 15]). In [6] up to third order positivity-preserving finite volume schemes are constructed based on positivity-preserving properties by the corresponding first order schemes for both density and pressure of one and two dimensional compressible Euler equations. Following [6], positivity-preserving high order DG schemes for compressible Euler equations were first introduced in [13], where the limiter in [12] is generalized. A recent work by Zhang and Shu in [14] introduced a minimum-entropy-principle-preserving limiter for high order schemes to the compressible Euler equation. In their work, the limiter for entropy part is enforced separately from the ones for the density and pressure and is given implicitly with the limiter parameter solved by Newton’s iteration.

For the isentropic gas dynamics, the invariant region is bounded by two global Riemann invariants; for which the authors have designed an explicit limiter in [4] to preserve the underlying invariant region, called an invariant-region-preserving (IRP) limiter. Our goals in this work are to design an IRP limiter for the compressible Euler system (2.1.1) and to rigorously prove that such a limiter does not destroy the high order accuracy in general cases. Our limiter differs from that in [14] in two aspects: (i) it is given in an explicit form; (ii) the scaling reconstruction depends on
a uniform parameter for the whole vector solution polynomial; in addition to the rigorous proof of the preservation of the accuracy by the limiter. As a result, the limiter preserves the positivity of density and pressure and also a maximum principle for the specific entropy [10], with reduced computational costs in numerical implementations.

2.2 The limiter

We construct a novel limiter based on both the cell average (strictly in $A$) and the high order polynomial approximation, which is not entirely in $A$; through a linear convex combination as in [12, 14].

2.2.1 Averaging is a contraction

For initial density $\rho_0 > 0$ and pressure $p_0 > 0$, we fix

$$ s_0 = \inf_x \log \left( \frac{p_0(x)}{\rho_0(x)} \right), $$

and define $q = (s_0 - s)\rho$, then the set $A$ is equivalent to the following set:

$$ \Sigma = \{ w : \rho > 0, p > 0, q \leq 0 \}, $$

which is convex due to the concavity of $p$ and convexity of $q$. By using set $\Sigma$ we are able to work out an explicit limiter which has the invariant-region-preserving property. Numerically, the set of admissible states is defined as

$$ \Sigma^\epsilon = \{ w : \rho \geq \epsilon, p \geq \epsilon, q \leq 0 \}, $$

with its interior denoted by

$$ \Sigma_0^\epsilon = \{ w : \rho > \epsilon, p > \epsilon, q < 0 \}, $$

where $\epsilon$ is a small positive number chosen (say as $10^{-13}$ in practice) so that $q$ is well defined.

For any bounded interval $I$ (or bounded domain in multi-dimensional case), we define the average of $w(x)$ by

$$ w = \frac{1}{|I|} \int_I w(x) dx $$
where $|I|$ is the measure of $I$. Such an averaging operator is a contraction:

**Lemma 2.2.1.** Let $\mathbf{w}(x)$ be non-trivial piecewise continuous vector functions. If $\mathbf{w}(x) \in \Sigma^\epsilon$ for all $x \in I$, then $\mathbf{w} \in \Sigma_0^\epsilon$ for any bounded interval $I$.

**Proof.** For the entropy part, since $q$ is convex, using Jensen’s inequality and the assumption, we have

$$q(\bar{\mathbf{w}}) = q\left(\frac{1}{|I|} \int_I \mathbf{w}(x)dx\right) \leq \frac{1}{|I|} \int_I q(\mathbf{w}(x))dx \leq 0. \quad (2.2.6)$$

With this we can show $q(\bar{\mathbf{w}}) < 0$. Otherwise, if $q(\bar{\mathbf{w}}) = 0$, we must have $q(\mathbf{w}(x)) = 0$ for all $x \in I$; that is

$$(s_0 - s(\bar{\mathbf{w}})) \bar{\rho} = (s_0 - s(\mathbf{w}(x))) \rho(x). \quad (2.2.7)$$

By taking average of this relation over $I$ on both sides, we have for $g_1 = s \rho$,

$$s_0 \bar{\rho} - g_1(\bar{\mathbf{w}}) = s_0 \bar{\rho} - \frac{1}{|I|} \int_I g_1(\mathbf{w}(x))dx. \quad (2.2.8)$$

This gives

$$\frac{1}{|I|} \int_I g_1(\mathbf{w}(x))dx = g_1(\bar{\mathbf{w}}). \quad (2.2.9)$$

By taking the Taylor expansion around $\bar{\mathbf{w}}$, we have

$$g_1(\mathbf{w}(x)) = g_1(\bar{\mathbf{w}}) + \nabla_w g_1(\bar{\mathbf{w}}) \cdot \xi + \xi^T H_1 \xi, \quad \forall x \in I, \quad \xi := \mathbf{w}(x) - \bar{\mathbf{w}}, \quad (2.2.10)$$

which upon integration yields $\frac{1}{|I|} \int_I \xi^T H_1 \xi dx = 0$, where $H_1$ is the Hessian matrix of $g_1$. This combined with the strict concavity of $g_1$ ensures that $\mathbf{w}(x) \equiv \bar{\mathbf{w}}$, which contradicts the assumption.

We can show $p(\bar{\mathbf{w}}) > \epsilon$ by a similar contradiction argument. The density part is obvious. \qed

### 2.2.2 Reconstruction

Let $\mathbf{w}_h(x) = (\rho_h(x), m_h(x), E_h(x))^T$ be a vector of polynomials of degree $k$ over an interval $I$, which is a high order approximation to the smooth function $\mathbf{w}(x) = (\rho(x), m(x), E(x))^T \in \Sigma^\epsilon$. We assume that the average $\bar{\mathbf{w}}_h \in \Sigma_0^\epsilon$, but $\mathbf{w}_h(x)$ is not entirely located in $\Sigma^\epsilon$ for $x \in I$, then we can use the average as a reference in the following reconstruction

$$\hat{\mathbf{w}}_h(x) = \theta \mathbf{w}_h(x) + (1 - \theta)\bar{\mathbf{w}}_h, \quad (2.2.11)$$
where
\[ \theta = \min\{1, \theta_1, \theta_2, \theta_3\}, \quad (2.2.12) \]
with
\[ \theta_1 = \frac{\bar{\rho}_h - \epsilon}{\rho_h - \rho_{h,\text{min}}}, \quad \theta_2 = \frac{p(\bar{w}_h) - \epsilon}{p(w_h) - p_{h,\text{min}}}, \quad \theta_3 = \frac{-q(\bar{w}_h)}{q_{h,\text{max}} - q(w_h)} \]
and
\[ \rho_{h,\text{min}} = \min_{x \in I} \rho_h(x), \quad p_{h,\text{min}} = \min_{x \in I} p(w_h(x)), \quad q_{h,\text{max}} = \max_{x \in I} q(w_h(x)). \quad (2.2.13) \]

Note that \( p(\bar{w}_h) > p_{h,\text{min}} \) and \( q(\bar{w}_h) < q_{h,\text{max}} \) due to the concavity of \( p \) and convexity of \( q \). Therefore \( \theta_i \)'s are well-defined and positive, for \( i = 1, 2, 3 \). We can prove that this reconstruction has three desired properties, summarized in the following.

**Theorem 2.2.2.** The reconstructed polynomial \( \tilde{w}_h(x) \) satisfies the following three properties:

(i) the average is preserved, i.e. \( \bar{w}_h = \tilde{w}_h \);

(ii) \( \tilde{w}_h(x) \) lies entirely within invariant region \( \Sigma^\epsilon, \forall x \in I \);

(iii) order of accuracy is maintained, i.e., \( \|\tilde{w}_h - w\|_{\infty} \leq C\|w_h - w\|_{\infty} \), provided \( \|w_h - w\|_{\infty} \) is sufficient small, where \( C \) is a positive constant that only depends on \( \bar{w}_h, w \), and the invariant region \( \Sigma^\epsilon \).

**Proof.**

(i) Since \( 0 < \theta \leq 1 \) is a uniform constant, average preservation is obvious.

(ii) If \( \rho_{h,\text{min}} \geq \epsilon, p_{h,\text{min}} \geq \epsilon \), and \( q_{h,\text{max}} \leq 0 \), then \( \theta = 1 \), no reconstruction is needed. When \( \theta = \theta_1 \), we have
\[ \tilde{\rho}_h(x) = \theta_1 \rho_h(x) + (1 - \theta_1) \bar{\rho}_h \]
\[ = (\tilde{\rho}_h - \epsilon) \frac{\rho_h(x) - \rho_{h,\text{min}}}{\tilde{\rho}_h - \rho_{h,\text{min}}} + \epsilon \geq \epsilon. \quad (2.2.14) \]

Since \( \theta_1 \leq \theta_2 \), we have \( (p(\bar{w}_h) - p_{h,\text{min}})\theta_1 + \epsilon \leq p(\bar{w}_h) \). Therefore, by the concavity of \( p \), we have
\[ p(\tilde{w}_h) \geq \theta_1 p(w_h) + (1 - \theta_1) p(\bar{w}_h) \]
\[ = \theta_1 (p(w_h) - p(\bar{w}_h)) + p(\bar{w}_h) \]
\[ \geq \theta_1 (p(w_h) - p(\bar{w}_h)) + (p(\bar{w}_h) - p_{h,\text{min}})\theta_1 + \epsilon \]
\[ = \theta_1 (p(w_h) - p_{h,\text{min}}) + \epsilon \geq \epsilon. \quad (2.2.15) \]
For entropy part, since \( \theta_1 \leq \theta_3 \), we have \( \theta_1(q_{h, \text{max}} - q(\bar{w}_h)) \leq -q(\bar{w}_h) \). Therefore, by the convexity of \( q \), we have

\[
q(\bar{w}_h) < \theta_1 q(w_h) + (1 - \theta_1)q(\bar{w}_h) = \theta_1(q(w_h) - q(\bar{w}_h)) + q(\bar{w}_h)
\]

\[
\leq \theta_1(q_{h, \text{max}} - q(\bar{w}_h)) + q(\bar{w}_h) \leq 0.
\]

(2.2.16)

In the case that \( \theta = \theta_2 \) or \( \theta = \theta_3 \) the proof is similar.

(iii) We prove for the case \( \theta = \theta_2 \), the other cases are similar. In such case we only need to prove

\[
\|\bar{w}_h - w_h\|_\infty \leq C\|w_h - w\|_\infty,
\]

from which (iii) follows by using the triangle inequality. Here and in what follows \( \|\cdot\|_\infty := \max_{x \in I} |\cdot| \).

We prove (2.2.17) in four steps.

Step 1. From (2.2.11) it follows that

\[
\|\bar{w}_h - w_h\|_\infty = (1 - \theta_2)\|\bar{w}_h - w_h\|_\infty
\]

\[
= \max_{x \in I} |\bar{w}_h - w_h(x)|
\]

\[
= \frac{p(w_h) - p_{h, \text{min}}}{p(\bar{w}_h) - p_{h, \text{min}}} (\epsilon - p_{h, \text{min}}).
\]

(2.2.18)

Step 2. The overshoot estimate. Since \( w(x) \in \Sigma^\epsilon \),

\[
\epsilon - p_{h, \text{min}} \leq \max_x (p(w) - p(w_h)) \leq C_1\|w - w_h\|_\infty, \quad C_1 := \|\nabla p\|_\infty.
\]

(2.2.19)

Step 3. We map \( I \) to \([0, 1]\) by \( \xi = (x - a)/(b - a) \) for \( I = [a, b] \), and let \( l_\alpha(\xi) \) \( (\alpha = 1, \ldots, N) \) be the Lagrange interpolating polynomials at quadrature points \( \hat{\xi}^\alpha \in [0, 1] \) with \( N = k + 1 \), then

\[
w_h(x) - \bar{w}_h = \sum_{\alpha=1}^N (w_h(\hat{x}^\alpha) - \bar{w}_h)l_\alpha(\xi), \quad \text{where} \quad \hat{x}^\alpha = a + (b - a)\hat{\xi}^\alpha.
\]

Hence, we have

\[
\max_{x \in I} |w_h(x)| \leq \max_{\xi \in [0, 1]} \sum_{\alpha=1}^N |l_\alpha(\xi)||\bar{w}_h - w_h(\hat{x}^\alpha)|
\]

\[
\leq C_2 \max_{\alpha} |\bar{w}_h - w_h(\hat{x}^\alpha)|,
\]

where \( C_2 = \Lambda_{k+1}([0, 1]) \leq \sum_{\alpha=1}^N |l_\alpha(\xi)| \) is the Lebesgue constant. Note that

\[
\max_{\alpha} |\bar{w}_h - w_h(\hat{x}^\alpha)| \leq \max_{\alpha} |\bar{p}_h - \rho_h(\hat{x}^\alpha)| + \max_{\alpha} |\bar{m}_h - m_h(\hat{x}^\alpha)| + \max_{\alpha} |\bar{E}_h - E_h(\hat{x}^\alpha)|.
\]

(2.2.21)
Define
\[ \hat{f}_{h,\min} = \min_{\alpha} f(w_h(\hat{x}^{\alpha})), \quad \hat{f}_{h,\max} = \max_{\alpha} f(w_h(\hat{x}^{\alpha})), \]
(2.2.22)
we can show that
\[ \max_{\alpha} |\hat{f}_h - f_h(\hat{x}^{\alpha})| \leq \max\{\hat{f}_h - \hat{f}_{h,\min}, \hat{f}_{h,\max} - \hat{f}_h\} \leq C_3(\hat{f}_h - \hat{f}_{h,\min}), \]
(2.2.23)
where
\[ C_3 = \max \left\{ 1, \frac{1 - \min_{\alpha} \hat{w}_\alpha}{\min_{\alpha} \hat{w}_\alpha} \right\}. \]
(2.2.24)
Here \( f_h = \rho_h, m_h, E_h \). The type of estimates using \( C_2 \) and \( C_3 \) is known, see [11, Lemma 7, Appendix C]), where the proof was accredited to Mark Ainsworth.

Step 4. The above three steps lead to
\[ \|\hat{w}_h - w_h\|_\infty \leq C_1C_2C_3 \frac{B}{p(\bar{w}_h) - p_{h,\min}} \|w - w_h\|_\infty, \]
(2.2.25)
with
\[ B = \rho_h - \rho_{h,\min} + m_h - m_{h,\min} + E_h - \hat{E}_{h,\min}. \]
(2.2.26)
On one hand, we have \( p_{h,\min} \leq \epsilon \) since \( \theta = \theta_2 \leq 1 \), leading to
\[ p(\bar{w}_h) - p_{h,\min} \geq p(\bar{w}_h) - \epsilon; \]
(2.2.27)
On the other hand the assumption \( \theta = \theta_2 \leq \theta_1 \) implies
\[ \rho_h - \rho_{h,\min} \leq \left( \frac{\rho_h - \epsilon}{p(\bar{w}_h) - \epsilon} \right) \cdot (p(\bar{w}_h) - p_{h,\min}). \]
(2.2.28)
By the assumption on the smallness of \( \|w_h - w\|_\infty \) we have
\[ \bar{m}_h - m_{h,\min} \leq 2\|m - m_h\|_\infty + \bar{m} - m_{\min} \]
(2.2.29)
and
\[ \bar{E}_h - \bar{E}_{h,\min} \leq \bar{E} + 1. \]
(2.2.30)
where \( \bar{E} \geq \frac{\epsilon}{\gamma - 1} \) is used. Collecting the above estimates we take
\[ C_4 = \rho_h - \epsilon + 2\|m - m_h\|_\infty + \bar{m} - m_{\min} + \bar{E} + 1 \]
(2.2.31)
\[ p(\bar{w}_h) - \epsilon \]
to conclude the desired estimate in (iii) with \( C = \Pi_{i=1}^{4} C_i \).
2.2.3 Algorithm

Let $w^n_h$ be the numerical solution generated from a high order scheme of an abstract form

$$w^{n+1}_h = \mathcal{L}(w^n_h), \quad (2.2.32)$$

where $w^n_h = w^n_h(x) \in V_h$, which is a finite element space of piecewise polynomials of degree $k$ over each computational cell $I$. Assume $\lambda = \frac{\Delta t}{h}$ is the mesh ratio, where $h$ is the characteristic length of the mesh size.

Provided that scheme (2.2.32) has the following property: there exists $\lambda_0$, and a test set $S_I$ in each computational cell $I$ such that if

$$\lambda \leq \lambda_0 \quad \text{and} \quad w^n_h \in \Sigma^\epsilon, \quad x \in S_I \quad (2.2.33)$$

then

$$\bar{w}^{n+1}_h \in \Sigma_0^\epsilon, \quad (2.2.34)$$

then the IRP limiter can be applied with $I$ replaced by $S_I$ in (2.2.13), i.e.,

$$\rho_{h,\min} = \min_{x \in S_I} \rho_h(x), \quad p_{h,\min} = \min_{x \in S_I} p(w_h(x)), \quad q_{h,\max} = \max_{x \in S_I} q(w_h(x)). \quad (2.2.35)$$

Our algorithm is given as follows:

**Step 1.** Initialization: take the piecewise $L^2$ projection of $w_0$ onto $V_h$, such that

$$\int_I (w^0_h(x) - w_0(x)) \phi(x) dx = 0, \quad \forall \phi \in V_h. \quad (2.2.36)$$

Also from $w_0$, we compute $s_0$ as defined in (2.2.1) to determine the invariant region $\Sigma^\epsilon$.

**Step 2.** Impose the modified limiter (2.2.11), (2.2.12) with (2.2.35) on $w^n_h$ for $n = 0, 1, \cdots$.

**Step 3.** Update by the scheme:

$$w^{n+1}_h = \mathcal{L}(w^n_h). \quad (2.2.37)$$

Return to Step 2.
Remark 2.2.1. Indeed the limiter (2.2.11), (2.2.12) with (2.2.35) can well enhance the efficiency of computation, and we will use this modified IRP limiter in the numerical experiments. Note that with (2.2.35), (i) and (iii) in Theorem 2.2.2 remain valid, and the resulting reconstructed polynomial lies within invariant region $\Sigma^\epsilon$ only for $x \in S_I$.

Remark 2.2.2. Notice that Lemma 2.2.1 ensures that $\hat{w}_h^0$ lies strictly within $\Sigma_0^\epsilon$, therefore the limiter is valid already at the initialization step.

Remark 2.2.3. Some sufficient conditions for (2.2.33) to ensure the cell average propagation property (2.2.34) for the DG method have been obtained for one-dimensional case ([13]), as well as for rectangular meshes ([13, 14]) and triangular meshes ([15]) in two-dimensional cases. For example, the test set $S_I$ and the CFL condition given in [13, Theorem 2.1] is

$$S_I = \{ \hat{x}^\alpha, \alpha = 1, \cdots, N \},$$

which is a set of $N$-point Legendre Gauss-Lobatto quadrature on $I$ with $2N - 3 \geq k$, and

$$\lambda \|(|u| + c)\|_{\infty} \leq \frac{1}{2} \hat{w}_1,$$

where $\hat{w}_1$ is the first Legendre Gauss-Lobatto quadrature weights for the interval $[-\frac{1}{2}, \frac{1}{2}]$ such that $\sum_\alpha ^N \hat{w}_\alpha = 1$.

### 2.3 Numerical tests

We present numerical tests for the IRP limiter applied to a general high order DG scheme with the Lax-Friedrich numerical flux, using a proper time discretization. The semi-discrete DG scheme we take is a closed ODE system of the form

$$\frac{d}{dt} W = L(W),$$

where $W$ consists of the unknown coefficients of the numerical solution in terms of the spatial basis, and $L$ is the corresponding spatial operator.

We consider the following two types of time discretizations.
- The third order SSP Runge-Kutta (RK3) method in [8] reads as
  \[
  \begin{align*}
  W^{(1)} &= W^n + \Delta t L(W^n), \\
  W^{(2)} &= \frac{3}{4} W^n + \frac{1}{4} W^{(1)} + \frac{1}{4} \Delta t L(W^{(1)}), \\
  W^{n+1} &= \frac{1}{3} W^n + \frac{2}{3} W^{(2)} + \frac{2}{3} \Delta t L(W^{(2)}).
  \end{align*}
  \]
  (2.3.2)

- The third order SSP multi-stage (MS) method in [7] reads as
  \[
  W^{n+1} = \frac{16}{27} (W^n + 3\Delta t L(W^n)) + \frac{11}{27} \left( W^{n-3} + \frac{12}{11} \Delta t L(W^{n-3}) \right).
  \]
  (2.3.3)

We apply the limiter at each time stage or each time step.

Remark 2.3.1. In the implementation of the third order SSP multi-step method, we apply SSP RK3 method in the first three time evolutions to obtain the starting values.

In all of the following examples $\gamma = 1.4$ is taken.

**Example 1. Accuracy Test**

We first test the accuracy of the IRP DG scheme. The initial condition is
\[
\begin{align*}
\rho_0(x) &= 1 + \frac{1}{2} \sin(2\pi x), \quad u_0(x) = 1, \quad p_0(x) = 1.
\end{align*}
\]
(2.3.4)

The domain is $[0, 1]$ and the boundary condition is periodic. The exact solution is
\[
\begin{align*}
\rho(x, y, t) &= 1 + \frac{1}{2} \sin(2\pi(x - t)), \quad u(x, t) = 1, \quad p(x, t) = 1.
\end{align*}
\]
(2.3.5)

The results presented in Tables 2.1 and 2.2 show that using IRP limiter does not destroy high order accuracy.

In the following examples, we solve (2.1.1) subject to several different Riemann initial data. We compare the numerical solution obtained from the DG scheme with IRP limiter (2.2.11), (2.2.12) with (2.2.35) and the one obtained from the DG scheme with only positivity-preserving limiter, that is, using $\theta = \min\{1, \theta_1, \theta_2\}$, where $\theta_1$ and $\theta_2$ are defined as in (2.2.12).

**Example 2. Lax Shock Tube Problem**

Consider the Lax initial data:
\[
(\rho, m, E) = \begin{cases} 
  (0.445, 0.311, 8.928), & x < 0, \\
  (0.5, 0, 1.4275), & x > 0,
\end{cases}
\]
(2.3.6)
Table 2.1 Accuracy test with $P^2$-DG scheme for Example 1.

<table>
<thead>
<tr>
<th>$P^2$ DG</th>
<th>SSP RK</th>
<th>SSP multi-step</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$L^\infty$Error</td>
<td>Order</td>
</tr>
<tr>
<td>8</td>
<td>5.43E-04 /</td>
<td>5.77E-04 /</td>
</tr>
<tr>
<td>16</td>
<td>8.98E-05</td>
<td>2.60</td>
</tr>
<tr>
<td>32</td>
<td>1.04E-05</td>
<td>3.11</td>
</tr>
<tr>
<td>64</td>
<td>1.33E-06</td>
<td>2.97</td>
</tr>
<tr>
<td>128</td>
<td>1.67E-07</td>
<td>2.99</td>
</tr>
</tbody>
</table>

Table 2.2 Accuracy test with $P^3$-DG scheme for Example 1.

<table>
<thead>
<tr>
<th>$P^3$ DG</th>
<th>SSP RK</th>
<th>SSP multi-step</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$L^\infty$Error</td>
<td>Order</td>
</tr>
<tr>
<td>8</td>
<td>1.44E-05 /</td>
<td>1.09E-05 /</td>
</tr>
<tr>
<td>16</td>
<td>1.39E-06</td>
<td>3.37</td>
</tr>
<tr>
<td>32</td>
<td>7.06E-08</td>
<td>4.30</td>
</tr>
<tr>
<td>64</td>
<td>6.34E-09</td>
<td>3.48</td>
</tr>
<tr>
<td>128</td>
<td>3.50E-10</td>
<td>4.18</td>
</tr>
</tbody>
</table>

which induces a composite wave, a rarefaction wave followed by a contact discontinuity and then by a shock. We calculate the exact solution by following the formulas given in [5, Section 14.11]. The $P^2$-DG scheme with SSP RK3 method in time discretization is tested on $N = 100$ cells over $x \in [-2, 2]$ at final time $T = 0.5$. From Fig. 2.1, we see that the IRP limiter helps to damp oscillations near the discontinuities.

**Example 3. Shu-Osher Shock Tube Problem**

Consider the Shu-Osher problem:

\[
(\rho, u, p) = \begin{cases} 
(3.857143, 2.629369, 10.3333), & x < -4, \\
(1 + 0.2 \sin 5x, 0, 1), & x \geq -4.
\end{cases}
\]  

(2.3.7)

The $P^2$-DG scheme with SSP RK3 method in time discretization is tested on $N = 100$ cells over $x \in [-5, 5]$ at final time $T = 1.8$. The reference solution is obtained from $P^2$-DG scheme with SSP RK3 method on $N = 2560$ cells. The results presented in Fig. 2.2 show that the shock is captured well.
Figure 2.1  Lax shock tube problem. Exact solution (solid line) vs numerical solution (dots); Top: With positive-preserving limiter; Bottom: With IRP limiter

Figure 2.2  Shu-Osher problem. Exact solution (solid line) vs numerical solution (dots); Left: With positive-preserving limiter; Right: With IRP limiter
2.4 Conclusion and future work

In this work, we introduced a novel IRP limiter for the one-dimensional compressible Euler equations. The limiter is made so that the reconstructed polynomial preserves the cell average, lies entirely within the invariant region and does not destroy the original high order of accuracy for smooth solutions. Moreover, this limiter is explicit and easy for computer implementation. Let us point out that the IRP limiter (2.2.11) may be applied to multi-dimensional compressible Euler equations as well if we replace $I$ in (2.2.13) by multi-dimensional cells or test set in each cell. Implementation details are in a forthcoming paper. Future work would be to investigate IRP limiters for more general hyperbolic systems or specific systems in important applications.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant DMS1312636 and by NSF Grant RNMS (KI-Net) 1107291.

References


CHAPTER 3. AN IRP LIMITER FOR DG SCHEMES TO ISENTROPIC EULER EQUATIONS

A paper accepted for publication in Numerical Methods for Partial Differential Equations.

Yi Jiang and Hailiang Liu

Abstract

In this paper, we introduce an invariant-region-preserving (IRP) limiter for the p-system and the corresponding viscous p-system, both of which share the same invariant region. Rigorous analysis is presented to show that for smooth solutions the order of approximation accuracy is not destroyed by the IRP limiter, provided the cell average stays away from the boundary of the invariant region. Moreover, this limiter is explicit, and easy for computer implementation. A generic algorithm incorporating the IRP limiter is presented for high order finite volume type schemes as long as the evolved cell average of the underlying scheme stays strictly within the invariant region. For any high order discontinuous Galerkin (DG) scheme to the p-system, sufficient conditions are obtained for cell averages to stay in the invariant region. For the viscous p-system, we design both second and third order IRP DG schemes. Numerical experiments are provided to test the proven properties of the IRP limiter and the performance of IRP DG schemes.

3.1 Introduction

We are interested in invariant-region-preserving (IRP) high order numerical approximations of solutions to systems of hyperbolic conservation laws

\[ w_t + F(w)_x = 0, \quad x \in \mathbb{R}, \quad t > 0 \]  

(3.1.1)
with the unknown vector \( w \in \mathbb{R}^m (m \geq 2) \) and the flux function \( F(w) \in \mathbb{R}^m \), subject to initial data \( w(x,0) = w_0(x) \). An invariant region \( \Sigma \) to this system is a convex open set in phase space \( \mathbb{R}^m \) so that if the initial data is in the region \( \Sigma \), then the solution will remain in \( \Sigma \). It is desirable to construct high order numerical schemes solving (3.1.1) that can preserve the whole invariant region \( \Sigma \), which is in general a difficult problem.

For scalar conservation equations the notion of invariant region is closely related to the maximum principle. In the case of nonlinear systems, the notion of maximum principle does not apply and must be replaced by the notion of invariant region. There are models that feature known invariant regions. For example, the invariant region of \( 2 \times 2 \) \((m = 2) \) systems of hyperbolic conservation laws can be described by two Riemann invariants. In this work, we focus on a model system, i.e., the \( p \)-system and its viscous counterpart, although the specific form of the system is not essential for the IRP approach. Other one-dimensional hyperbolic system of conservation laws can be studied along the same lines as long as it admits a convex invariant region.

The initial value problem (IVP) for the \( p \)-system is given by
\[
\begin{align*}
v_t - u_x &= 0, \quad x \in \mathbb{R}, \quad t > 0 \\
u_t + p(v)_x &= 0, \\
v &= v_0 > 0, \quad u = u_0, \quad x \in \mathbb{R}, \quad t = 0,
\end{align*}
\] (3.1.2)
where \( p : \mathbb{R}^+ \rightarrow \mathbb{R}^+ \) satisfies \( p' < 0 \) and \( p'' > 0 \). The choice of \( p = kv^{-\gamma} \) with the adiabatic gas constant \( \gamma > 1 \) and positive constant \( k > 0 \) leads to the isentropic (=constant entropy) gas dynamic equations. These equations represent the conservation of mass and momentum, where \( v \) denotes the specific volume, \( v = \frac{1}{\rho} \) and \( \rho \) is the density, \( u \) denotes the velocity, see e.g. [23]. This system and the counterpart in Eulerian coordinates are called compressible Euler equations.

The entropy solution of the \( p \)-system can be realized as the limit of the following diffusive system
\[
\begin{align*}
v_t - u_x &= \epsilon v_{xx}, \quad x \in \mathbb{R}, \quad t > 0, \\
u_t + p(v)_x &= \epsilon u_{xx}, \\
v &= v_0 > 0, \quad u = u_0, \quad x \in \mathbb{R}.
\end{align*}
\] (3.1.3)
It is well known that these two systems share a common invariant region $\Sigma$, which is a convex open set in the phase space and expressed by two Riemann invariants of the $p$-system. For more general discussion on invariant regions, we refer to [23, 10].

The main objective of this paper is to present both an explicit IRP limiter and several high order IRP schemes to solve the above two systems. The constructed high order numerical schemes can thus preserve the whole invariant region $\Sigma$.

IRP high order methods have seldom been studied from a numerical point of view and the main goal of the present article is to introduce the required tools and explain how the standard approach for preserving the maximum principle of scalar conservation laws should be modified. While the presentation is given for this particular model, it can easily be carried over to other systems equipped with a convex invariant region. For example, the one-dimensional shallow water system, and the isentropic compressible Euler system in Eulerian coordinates (see Section 2 for details of their respective invariant regions). On the other hand, it does not seem easy to extend the analysis to multi-dimensional setting. One technical difficulty is that the invariant region determined by the Riemann invariants does not apply to the multi-dimensional case.

We should also point out the global existence of the right physical solutions to compressible Euler equations is a formidable open problem in general, and the underlying system has been of much academic interest, since the isentropic case allows for a simplified mathematical system with a rich structure. Indeed the global existence has been well established following the ideas of Lax in [13] on entropy restrictions upon bounded solutions and the program initiated by Diperna [5, 4], extended and justified by Chen [1], and completed by Lions et al [15] for all $\gamma \in (0, \infty)$, using tools of the kinetic formulation [16].

Enforcing the IRP property numerically should help damp oscillations in numerical solutions, as evidenced by the maximum-principle-preserving high order finite volume schemes developed in [20, 11] for scalar conservation laws. A key development along this line is the work by Zhang and Shu [28], where the authors constructed a maximum-principle-preserving limiter for finite volume (FV) or discontinuous Galerkin (DG) schemes of arbitrary order for solving scalar conservation laws.
Their limiter is proved to maintain the original high order accuracy of the numerical approximation (see [27]).

For nonlinear systems of conservation laws in several space variables, invariant regions under the Lax-Friedrich schemes were studied by Frid in [6, 7]. A recent study of invariant regions for general nonlinear hyperbolic systems using the first order continuous finite elements was given by Guermond and Popov in [9]. However, there has been little study on the preservation of the invariant region for systems by high order numerical schemes. In the literature, instead of the whole invariant region, positivity of some physical quantities are usually considered. Positivity-preserving finite volume schemes for Euler equations are constructed in [21] based on positivity-preserving properties by the corresponding first order schemes for both density and pressure. Following [21], positivity-preserving high order DG schemes for compressible Euler equations were first introduced in [29], where the limiter introduced in [28] is generalized. We also refer to [26, 25, 24, 31] for more related works about hyperbolic systems of conservation laws including Euler equations, reactive Euler equations, and shallow water equations. A more closely related development is the work by Zhang and Shu [30], where the authors introduced a minimum-entropy-principle-preserving limiter for high order schemes to the compressible Euler equation, while the limiter is implicit in the sense that the limiter parameter is solved by Newton’s iteration. It is explained there how the high order accuracy can be maintained for generic smooth solutions. An explicit IRP limiter for compressible Euler equations was recently proposed by us in [12]. The main distinction between the limiter in [30] and that in [12] is that we give an explicit formula, with a single uniform scaling parameter for the whole vector solution polynomials. This is particularly relevant at reducing computational costs in numerical implementations.

In this paper, we construct an IRP limiter to preserve the whole invariant region $\Sigma$ of the p-system and the viscous p-system. The cell average of numerical approximation polynomials is used as a reference to pull each cell polynomial into the invariant region. The limiter parameter is given explicitly, which is made possible by using the convexity and concavity of the two Riemann invariants, respectively. Such explicit form is quite convenient for computer implementation. More-
over, rigorous analysis is presented to prove that for smooth solutions the high order of accuracy is not destroyed by the limiter in general cases. By general cases we mean that in some rare cases accuracy deterioration can still happen, as shown by example in Appendix B.

Furthermore, we present a generic algorithm incorporating this IRP limiter for high order finite volume type schemes as long as the evolved cell average by the underlying scheme stays strictly in the invariant region. This is true for first order schemes under proper CFL conditions (see, for instance, Theorem 3.4.1). For high order schemes, this may hold true provided solution values on a set of test points (called test set hereafter) stay within the invariant region, in addition to the needed CFL condition. Indeed we are able to obtain such CFL condition and test set for any high order DG schemes to the p-system.

For the viscous p-system we present both second and third order DDG schemes, using the DDG diffusive flux proposed in [18], and prove that the cell average remains strictly within the invariant region under some sufficient conditions. In order to ensure that the diffusive contribution lies strictly in $\Sigma_0$, the interior of $\Sigma$, we use the positivity-preserving results proved by Liu and Yu in [19] for linear Fokker-Planck equations.

The paper is organized as follows. We first review the concept of invariant region for the p-system and show that the averaging operator is a contraction in Section 2. In Section 3 we construct an explicit IRP limiter and prove that the high order of accuracy is not destroyed by such limiter in general cases. Accordingly, a generic IRP algorithm is presented for numerical implementations. In Section 4, we discuss the IRP property for cell averages of any high order finite volume type scheme for the p-system, and second and third order DG schemes for the viscous p-system. Sufficient conditions include a CFL condition and a test set for each particular scheme with forward Euler time-discretization. In Section 5, we present extensive numerical experiments to test the desired properties of the IRP limiter and the performance of the IRP DG schemes. In addition, the convergence of the viscous profiles to the entropy solution is illustrated from a numerical point of view. Concluding remarks are given in Section 6.
### 3.2 Invariant region and averaging

The p-system is strictly hyperbolic and admits two Riemann invariants

\[
r = u - \int_{m}^{v} \sqrt{-p'(\xi)}d\xi, \quad s = u + \int_{m}^{v} \sqrt{-p'(\xi)}d\xi, \quad m = \inf_{x} v_{0}(x) > 0,
\]

which, associated with the eigenvalues \(\pm \sqrt{-p'(v)}\), satisfy

\[
\begin{cases}
    r_{t} + \sqrt{-p'(v)}r_{x} = 0, \\
    s_{t} - \sqrt{-p'(v)}s_{x} = 0.
\end{cases}
\]

In addition we assume that the pressure function \(p\) also satisfies

\[
\int_{0}^{v} \sqrt{-p'(\xi)}d\xi = \infty, \quad \forall v > 0.
\]  

(3.2.1)

This condition is met for \(p(v) = v^{-\gamma}\) with \(\gamma > 1\). Set

\[g(v) = \int_{m}^{v} \sqrt{-p'(\xi)}d\xi,\]

then the assumptions on the pressure implies that \(g\) is increasing, concave and tends asymptotically towards \(v = 0\), i.e.,

\[
g'(v) = \sqrt{-p'(v)} > 0, \quad g''(v) = \frac{-p''(v)}{2\sqrt{-p'(v)}} < 0, \quad g(0) = -\infty.
\]

This shows that

\[
r = u - g(v) \text{ is convex, and } s = u + g(v) \text{ is concave},
\]

(3.2.2)

and two level set curves \(r = r_{0}\) and \(s = s_{0}\) must intersect. We now fix

\[
r_{0} = \sup_{x} r(v_{0}(x), u_{0}(x)), \quad s_{0} = \inf_{x} s(v_{0}(x), u_{0}(x))
\]

(3.2.3)

and define

\[
\Sigma = \{(v, u)\mid r \leq r_{0}, s \geq s_{0}\},
\]
which is known as the invariant region of the p-system, see [23], in the sense that if \((v_0(x), u_0(x))^\top \in \Sigma\), then the entropy solution \((v(x, t), u(x, t))^\top \in \Sigma\).

This implies that for entropy solution, \(v\) remains bounded away from 0 and \(u\) remains bounded in \(L^\infty\)-norm. But no upper bound on \(v\) is available because \(v = \infty\) corresponds to \(\rho = 0\) in Eulerian coordinates, which may actually happen.

In what follows we shall use \(\Sigma_0\) to denote \(\Sigma\) without the two boundary sides, i.e.,

\[
\Sigma_0 = \{(v, u)^\top \mid r < r_0, s > s_0\}.
\]

Figure 3.1 Invariant region for p-system

**Lemma 3.2.1.** Let \((v^*, u^*)^\top\) be the intersection of upper and lower boundaries of \(\Sigma\), then

\[
0 < v^* \leq m = \inf v_0(x).
\]  

(3.2.4)

**Proof.** According to the definition of \(r\) and \(s\), we have

\[
\begin{align*}
    r_0 &= \sup_x (u_0(x) - \int_{m}^{v_0(x)} \sqrt{-p'(\xi)}d\xi) = u^* - \int_{m}^{v^*} \sqrt{-p'(\xi)}d\xi, \\
    s_0 &= \inf_x (u_0(x) + \int_{m}^{v_0(x)} \sqrt{-p'(\xi)}d\xi) = u^* + \int_{m}^{v^*} \sqrt{-p'(\xi)}d\xi.
\end{align*}
\]
Hence
\[
\int_{m}^{v^*} \sqrt{-p'(\xi)} d\xi = \frac{s_0 - r_0}{2} \leq \int_{m}^{v_0(x)} \sqrt{-p'(\xi)} d\xi, \quad \forall x.
\]
This implies that \( v^* \leq v_0(x) \) for all \( x \), and \( v^* > 0 \) in light of the assumption (3.2.1). \( \square \)

For any bounded interval \( I \), we define the average of \( w(x) \equiv (v(x), u(x))^\top \) by
\[
\bar{w} = \frac{1}{|I|} \int_I w(x) dx,
\]
where \(|I|\) is the measure of \( I \). Such an averaging operator is a contraction in the sense stated in the following result.

**Lemma 3.2.2.** Let \( w(x) = (v(x), u(x))^\top \), where both \( v(x) \) and \( u(x) \) are non-trivial piecewise smooth functions. If \( w(x) \in \Sigma \) for all \( x \in \mathbb{R} \), then \( \bar{w} \in \Sigma_0 \) for any bounded interval \( I \).

**Proof.** Since \( r \) is convex, according to Jensen’s inequality, we have
\[
r(\bar{w}) = r \left( \frac{1}{|I|} \int_I w(x) \right) \leq \frac{1}{|I|} \int_I r(w(x)) dx \leq r_0.
\]
If \( r(\bar{w}) = r_0 \), then \( r(w(x)) = r_0 \) for all \( x \). That is,
\[
\bar{u} - g(\bar{v}) = u(x) - g(v(x)).
\]
By taking cell average of this relation on both sides, we have
\[
\bar{u} - g(\bar{v}) = \bar{u} - \frac{1}{|I|} \int_I g(v(x)) dx.
\]
This gives
\[
\frac{1}{|I|} \int_I g(v(x)) dx = g(\bar{v}).
\]
Since we have
\[
g(\bar{v}) \geq \frac{1}{|I|} \int_I g(v(x)) dx
\]
 implied by Jensen’s inequality, we conclude that \( v(x) \) must be a constant, so is \( u(x) = r_0 + g(v(x)) \).
Therefore, \( r(\bar{w}) < r_0 \) when \( w(x) \) is non-trivial. The proof of \( s(\bar{w}) > s_0 \) is entirely similar. \( \square \)
Remark 3.2.1. From given initial data in Σ, the above result ensures that its cell average lies strictly within Σ₀. On the other hand, the usual piecewise $L^2$ projection, used as initial data for numerical schemes, may not lie entirely in Σ, but it has the same cell average. Such cell average is the key ingredient in the construction of our IRP limiter.

In passing we present the invariant regions of two other equations, which can be studied along the same lines.

1. The one dimensional isentropic gas dynamic system in Euler coordinates, i.e.

$$\rho_t + (\rho u)_x = 0,$$
$$\rho u_t + (\rho u^2 + p(\rho))_x = 0,$$

with initial condition $(\rho_0(x), m_0(x))$, where $p(\rho) = \rho^\gamma$, $\gamma > 1$ and $m = \rho u$. The two Riemann invariants are

$$r = u + \frac{2\sqrt{\gamma}}{\gamma - 1} \rho^{(\gamma - 1)/2}, \quad s = u - \frac{2\sqrt{\gamma}}{\gamma - 1} \rho^{(\gamma - 1)/2}.$$ 

Hence

$$\Sigma_1 = \{(\rho, m)^\top | \ r \leq r_0, \ s \geq s_0\}$$

(3.2.6)

is the invariant region for system (3.2.5), where $r_0$ and $s_0$ are chosen as

$$r_0 = \sup_x r(\rho_0(x), m_0(x)), \quad s_0 = \inf_x s(\rho_0(x), m_0(x)).$$

2. The one dimensional shallow water equations with a flat bottom topography take the form

$$h_t + (hu)_x = 0,$$
$$hu_t + (hu^2 + \frac{g}{2} h^2)_x = 0,$$

where $h$ is the fluid height, $u$ the fluid velocity, and $g$ is the acceleration due to gravity, subject to initial condition $(h_0(x), m_0(x))$, where $m = hu$. The two Riemann invariants are

$$r = u + 2\sqrt{gh}, \quad s = u - 2\sqrt{gh}.$$
Hence

\[
\Sigma_2 = \{(h, m)^T | \text{ } r \leq r_0, \text{ } s \geq s_0\} 
\]  

(3.2.8)

is the invariant region for system (3.2.7), where \( r_0 \) and \( s_0 \) are chosen as

\[
r_0 = \sup_x r(h_0(x), m_0(x)), \text{ } s_0 = \inf_x s(h_0(x), m_0(x)).
\]

In contrast to the invariant region in Figure 1, which is open and unbounded, both \( \Sigma_1 \) and \( \Sigma_2 \) are closed and bounded regions in terms of \( (\rho, m) \) and \( (h, m) \), respectively.

### 3.3 The IRP limiter

#### 3.3.1 A limiter to enforce the IRP property

Let \( \mathbf{w}_h(x) = (v_h(x), u_h(x))^T \) be a vector of polynomials of degree \( k \) over an interval \( I \), which is a high order approximation to the smooth function \( \mathbf{w}(x) = (v(x), u(x))^T \in \Sigma \). We assume that the average \( \bar{\mathbf{w}}_h \in \Sigma_0 \), but \( \mathbf{w}_h(x) \) is not entirely located in \( \Sigma \). We then seek a modified polynomial using the cell average as a reference to achieve three objectives: (i) the modified polynomial preserves the cell average, (ii) it lies entirely within \( \Sigma \), and (iii) the high order of accuracy is not destroyed.

The modification is through a linear convex combination, which is of the form

\[
\tilde{\mathbf{w}}_h(x) = \theta \mathbf{w}_h(x) + (1 - \theta) \bar{\mathbf{w}}_h,
\]  

(3.3.1)

where \( \theta \in (0, 1] \) is defined by

\[
\theta = \min\{1, \theta_1, \theta_2\},
\]  

(3.3.2)

with

\[
\theta_1 = \frac{r_0 - r(\bar{\mathbf{w}}_h)}{r_{\text{max}} - r(\bar{\mathbf{w}}_h)}, \text{ } \theta_2 = \frac{s(\mathbf{w}_h) - s_0}{s(\bar{\mathbf{w}}_h) - s_{\text{min}}}
\]

and

\[
r_{\text{max}} = \max_{x \in I} r(\mathbf{w}_h(x)), \text{ } s_{\text{min}} = \min_{x \in I} s(\mathbf{w}_h(x)).
\]  

(3.3.3)
Notice that since $\bar{w}_h \in \Sigma_0$, we have $r(\bar{w}_h) < r_0$ and $s(\bar{w}_h) > s_0$. Also $r(\bar{w}_h) < r_{\max}$ and $s(\bar{w}_h) > s_{\min}$ due to the convexity of $r$ and concavity of $s$, respectively. Both $\theta_1$ and $\theta_2$ are well-defined and positive.

We proceed to show (ii) in Lemma 3.3.1, and prove the accuracy-preserving property (iii) rigorously in Lemma 3.3.3.

**Lemma 3.3.1.** If $\bar{w}_h \in \Sigma_0$, then $\tilde{w}_h(x) \in \Sigma$, $\forall x \in I$.

**Proof.** We only need to discuss the case $\theta = \theta_1$ or $\theta_2$. If $\theta = \theta_1$, with the convexity of $r$, we have

$$r(\tilde{w}_h(x)) \leq \theta_1 r(w_h(x)) + (1 - \theta_1)r(\bar{w}_h)$$

$$= (r_0 - r(\bar{w}_h)) \frac{r(w_h(x)) - r(\bar{w}_h)}{r_{\max} - r(\bar{w}_h)} + r(\bar{w}_h) \leq r_0.$$ 

On the other hand, from $\theta_1 \leq \theta_2$ it follows that

$$(s(\bar{w}_h) - s_{\min})\theta_1 \leq s(\bar{w}_h) - s_0. \quad (3.3.4)$$

By the concavity of $s$, we have

$$s(\tilde{w}_h(x)) \geq \theta_1 s(w_h(x)) + (1 - \theta_1)s(\bar{w}_h)$$

$$\geq \theta_1(s_{\min} - s(\bar{w}_h)) + s(\bar{w}_h)$$

$$\geq s_0 - s(\bar{w}_h) + s(\bar{w}_h) = s_0.$$ 

In the case that $\theta = \theta_2$, the proof is entirely similar. \qed

Next we show that this limiter does not destroy the accuracy. We first prepare the following lemma. We adopt the notation $\| \cdot \|_{\infty} \doteq \max_{x \in I} | \cdot |$ in the rest of the paper.

**Lemma 3.3.2.** If $\|w_h - w\|_{\infty}$ is sufficiently small, we have

(i) $\|w_h\|_{\infty} \leq \|w\|_{\infty} + 1$;

(ii) $\frac{v}{\theta} \leq v_h(x) \leq \|v\|_{\infty} + 1$, $\forall x \in I$;

(iii) $\|\nabla r\|_{\infty} \leq \sqrt{1 - p'(v^*/2)}$, $g'(v) \geq \sqrt{-p'(\|v\|_{\infty} + 1)}$. 

Proof. The statements (i) and (ii) follow from the assumption and the fact that \( v^* \leq v(x) \leq \|v\|_\infty \), for all \( x \in I \). And (iii) results from the facts that \( \|\nabla r\|_2 = \sqrt{1-p'(v)} \) is decreasing and \( g'(v) = \sqrt{-p'(v)} \) is decreasing, with respect to \( v \).

Lemma 3.3.3. The reconstructed polynomial preserves high order accuracy, i.e.

\[
\|\tilde{w}_h - w\|_\infty \leq C\|w_h - w\|_\infty,
\]

where \( C > 0 \) depends on \( w \) and \( \Sigma \), also on \( \tilde{w}_h \).

Proof. Consider the case \( \theta = \theta_1 \leq \theta_2 \). We only need to prove

\[
\|\tilde{w}_h - w_h\|_\infty \leq C\|w_h - w\|_\infty,
\]

from which (3.3.5) follows by further using the triangle inequality. We proceed to prove (3.3.6) in several steps.

Step 1. From (3.3.1), it follows that

\[
\|\tilde{w}_h - w_h\|_\infty = (1 - \theta_1)\|w_h - w\|_\infty
= \left(1 - \frac{r_0 - r(\tilde{w}_h)}{r_{\text{max}} - r(\tilde{w}_h)}\right)\|w_h - w\|_\infty
\leq \max_{x \in I} |\tilde{w}_h - w_h(x)| (r_{\text{max}} - r_0).
\]

Step 2. The overshoot estimate. Since \( w(x) \in \Sigma \),

\[
r_{\text{max}} - r_0 \leq \max_{x \in I} (r(w_h(x)) - r(w(x)))
\leq \|\nabla r\|_\infty \|w_h - w\|_\infty \leq C_1\|w_h - w\|_\infty,
\]

where \( C_1 = \sqrt{1-p'(v^*/2)} \) by Lemma 3.3.2.

Step 3. Let \( \hat{x}^\alpha (\alpha = 1, \ldots, k+1) \) be the Gauss quadrature points such that

\[
\tilde{w}_h = \sum_{\alpha=1}^{k+1} \hat{w}_\alpha w_h(\hat{x}^\alpha),
\]

(3.3.7)
with \( \hat{w}_\alpha > 0 \) and \( \sum_{\alpha=1}^{k+1} \hat{w}_\alpha = 1 \). Using these quadrature points, we also have

\[
\mathbf{w}_h(x) - \bar{\mathbf{w}}_h = \sum_{\alpha=1}^{k+1} (\mathbf{w}_h(\hat{x}^\alpha) - \bar{\mathbf{w}}_h) l_\alpha(\xi), \quad \xi = (x-a)/(b-a),
\]

where \( l_\alpha(\xi) \) \((\alpha = 1, \ldots, N)\) are the Lagrange interpolating polynomials. Hence, we have

\[
\max_{x \in I} |\bar{\mathbf{w}}_h - \mathbf{w}_h(x)| \leq \max_{\xi \in [0,1]} \sum_{\alpha=1}^{k+1} |l_\alpha(\xi)||\bar{\mathbf{w}}_h - \mathbf{w}_h(\hat{x}^\alpha)|
\]

\[
\leq C_2 \max_{\alpha} |\bar{\mathbf{w}}_h - \mathbf{w}_h(\hat{x}^\alpha)|,
\]

where \( C_2 = \Lambda_{k+1}([0,1]) \equiv \max_{\xi \in [0,1]} \sum_{\alpha=1}^{k+1} |l_\alpha(\xi)| \) is the Lebesgue constant.

**Step 4.** Change of variables. Since \( g(v) \) is increasing and concave, we have

\[
u_h(x) = \frac{r(\mathbf{w}_h(x)) + s(\mathbf{w}_h(x))}{2}, \quad \bar{\nu}_h(x) = g^{-1}\left(\frac{s(\mathbf{w}_h(x)) - r(\mathbf{w}_h(x))}{2}\right),
\]

\[ar{\mathbf{u}}_h(x) = \frac{r(\bar{\mathbf{w}}_h(x)) + s(\bar{\mathbf{w}}_h(x))}{2}, \quad \bar{\bar{\nu}}_h(x) = g^{-1}\left(\frac{s(\bar{\mathbf{w}}_h(x)) - r(\bar{\mathbf{w}}_h(x))}{2}\right).
\]

If we set

\[
E = \max_{\alpha} |r(\bar{\mathbf{w}}_h) - r(\mathbf{w}_h(\hat{x}^\alpha))| + \max_{\alpha} |s(\bar{\mathbf{w}}_h) - s(\mathbf{w}_h(\hat{x}^\alpha))|,
\]

then

\[
\max_{\alpha} |\bar{\mathbf{w}}_h - \mathbf{w}_h(\hat{x}^\alpha)| \leq \sqrt{\max_{\alpha} |\bar{\mathbf{u}}_h - u_h(\hat{x}^\alpha)|^2 + \max_{\alpha} |\bar{\nu}_h - v_h(\hat{x}^\alpha)|^2}
\]

\[
\leq \frac{1}{2} \sqrt{1 + \|g^{-1}(\cdot)\|_\infty^2} E
\]

\[
\leq C_3 E,
\]

where \( C_3 = \frac{1}{2} \sqrt{1 - \frac{1}{P'(\|v\|_\infty + 1)}} \) by Lemma 3.3.2.

**Step 5.** We are now left to show the uniform bound of

\[
\frac{E}{r_{\max} - r(\bar{\mathbf{w}}_h)} = \max_{\alpha} |r(\bar{\mathbf{w}}_h) - r(\mathbf{w}_h(\hat{x}^\alpha))| + \max_{\alpha} |s(\bar{\mathbf{w}}_h) - s(\mathbf{w}_h(\hat{x}^\alpha))|,
\]

which is equivalent to show the boundedness of

\[
\max \left\{ \frac{\hat{r}_{\max} - r(\bar{\mathbf{w}}_h)}{r_{\max} - r(\bar{\mathbf{w}}_h)}, \frac{\hat{r}(\bar{\mathbf{w}}_h) - \hat{r}_{\min}}{r_{\max} - r(\bar{\mathbf{w}}_h)} \right\} + \max \left\{ \frac{s(\bar{\mathbf{w}}_h) - s_{\min}}{r_{\max} - r(\bar{\mathbf{w}}_h)}, \frac{s(\bar{\mathbf{w}}_h) - s_{\max}}{r_{\max} - r(\bar{\mathbf{w}}_h)} \right\},
\]
where

\[
\hat{r}_{\text{max}} \doteq \max_{\alpha} r(w_h(\hat{x}^\alpha)), \quad \hat{r}_{\text{min}} \doteq \min_{\alpha} r(w_h(\hat{x}^\alpha)),
\]

\[
\hat{s}_{\text{max}} \doteq \max_{\alpha} s(w_h(\hat{x}^\alpha)), \quad \hat{s}_{\text{min}} \doteq \min_{\alpha} s(w_h(\hat{x}^\alpha)).
\]

Recall that \(\theta = \theta_1 \leq \theta_2\), therefore

\[
\frac{r_0 - r(w_h)}{r_{\text{max}} - r(w_h)} \leq \frac{s(w_h) - s_0}{s(w_h) - s_{\text{min}}} \Rightarrow \frac{s(w_h) - s_{\text{min}}}{r_{\text{max}} - r(w_h)} \leq \frac{s(w_h) - s_0}{r_0 - r(w_h)}.
\]

Then

\[
\max \left\{ \frac{\hat{r}_{\text{max}} - r(w_h)}{r_{\text{max}} - r(w_h)}, \frac{r(w_h) - \hat{r}_{\text{min}}}{r_{\text{max}} - r(w_h)} \right\} + \max \left\{ \frac{s(w_h) - \hat{s}_{\text{min}}}{r_{\text{max}} - r(w_h)}, \frac{\hat{s}_{\text{max}} - s(w_h)}{r_{\text{max}} - r(w_h)} \right\} \\
\leq C_4 \left( \max \left\{ 1, \frac{r(w_h) - \hat{r}_{\text{min}}}{r_{\text{max}} - r(w_h)} \right\} + \max \left\{ 1, \frac{\hat{s}_{\text{max}} - s(w_h)}{s(w_h) - s_{\text{min}}} \right\} \right),
\]

where \(C_4 = 2 \max \{1, \frac{s(w_h) - s_0}{r_0 - r(w_h)}\}\). Note that for the case \(\theta = \theta_2\), \(C_4\) needs to be replaced by \(2 \max \{1, \frac{r_0 - r(w_h)}{s(w_h) - s_0}\}\).

Step 6. Note that (3.3.7) when combined with convexity of \(r\) yields

\[
r(w_h) = r \left( \sum_{\alpha=1}^{k+1} \hat{w}_\alpha w_h(\hat{x}^\alpha) \right) \leq \sum_{\alpha=1}^{k+1} \hat{w}_\alpha r(w_h(\hat{x}^\alpha)).
\]

Assume \(\hat{r}_{\text{min}}\) is achieved at \(\hat{x}^1\), then

\[
\hat{w}_1 (r(w_h) - \hat{r}_{\text{min}}) \leq \sum_{\alpha=2}^{k+1} \hat{w}_\alpha (r(w_h(\hat{x}^\alpha)) - r(w_h)) \leq (r_{\text{max}} - r(w_h)) \sum_{\alpha=2}^{k+1} \hat{w}_\alpha
\]

\[
\Rightarrow \frac{r(w_h) - \hat{r}_{\text{min}}}{r_{\text{max}} - r(w_h)} \leq \frac{1 - \min_{\alpha} \hat{w}_\alpha}{\min_{\alpha} \hat{w}_\alpha}.
\]

Similarly, we can show that

\[
\frac{\hat{s}_{\text{max}} - s(w_h)}{s(w_h) - s_{\text{min}}} \leq \frac{1 - \min_{\alpha} \hat{w}_\alpha}{\min_{\alpha} \hat{w}_\alpha},
\]

due to the concavity of \(s\).

The above steps together have verified the claimed estimate (3.3.6), with the bounding constant

\[
C = \prod_{i=1}^{5} C_i, \quad C_4 = 2 \max \left\{ 1, \frac{s(w_h) - s_0}{r_0 - r(w_h)} \right\}, \quad C_5 = 2 \max \left\{ 1, \frac{1 - \min_{\alpha} \hat{w}_\alpha}{\min_{\alpha} \hat{w}_\alpha} \right\}.
\]

(3.3.8)
Remark 3.3.1. It is noticed that when $\bar{w}_h$ is close enough to the boundary of $\Sigma$, the constant $C_4$ can become quite large, indicating the possibility of accuracy deterioration in some cases. In appendix 3.6, we present two examples to show that the magnitude of $C_4$ can be uniformly bounded or quite large.

Remark 3.3.2. Our proof is the first attempt at rigorous justification of the accuracy-preservation by the IRP limiter for system case. Related techniques to Step 3 and Step 6 are also found in the proof of Lemma 7 [27] for scalar conservation laws, where the proof was accredited to Mark Ainsworth. An alternative proof for the accuracy of the limiter is given in [19], in which the authors consider some weighted averages for scalar Fokker-Planck equations.

In summary, we have the following result.

**Theorem 3.3.4.** Let $w_h(x) = (v_h(x), u_h(x))^T$ be a polynomial approximation to the smooth function $w(x) = (v(x), u(x))^T \in \Sigma$, over bounded cell $I$, and $\bar{w}_h \in \Sigma_0$. Define $r_{\text{max}} = \max_{x \in I} r(w_h(x))$, $s_{\text{min}} = \min_{x \in I} s(w_h(x))$. Then the modified polynomial

$$\tilde{w}_h(x) = \theta w_h(x) + (1 - \theta) \bar{w}_h,$$

where

$$\theta_1 = \frac{r_0 - r(\bar{w}_h)}{r_{\text{max}} - r(\bar{w}_h)}, \quad \theta_2 = \frac{s(\bar{w}_h) - s_0}{s(\bar{w}_h) - s_{\text{min}}}.$$ 

satisfies the following three properties:

1. the cell average is preserved, i.e. $\bar{w}_h = \tilde{w}_h$;

2. it entirely lies within invariant region $\Sigma$, i.e. $r(\tilde{w}_h) \leq r_0$, $s(\tilde{w}_h) \geq s_0$;

3. high order of accuracy is maintained, i.e. $\|\tilde{w}_h - w\|_\infty \leq C\|w_h - w\|_\infty$, where $C$ is a positive constant that only depends on $\bar{w}_h, w, \Sigma$.

Remark 3.3.3. The IRP limiter (3.3.1)–(3.3.3) remains valid for $\Sigma_1$ in (3.2.6) and $\Sigma_2$ in (3.2.8). The analysis above and the accuracy results in Theorem 3.3.4 can be easily generalized to both
the one-dimensional isentropic gas dynamic system in Eulerian coordinates (3.2.5) and the shallow water equations (3.2.7).

To show when this simple limiter can be incorporated into an existing high order numerical scheme, we present the following IRP algorithm.

### 3.3.2 Algorithm

Let $w^n_h$ be the numerical solution generated from a high order finite-volume-type scheme of an abstract form

$$w^{n+1}_h = L(w^n_h), \quad (3.3.9)$$

where $w^n_h = w^n_h(x) \in V_h$, and $V_h$ is a finite element space of piecewise polynomials of degree $k$ in each computational cell, i.e.,

$$V_h = \{v : v|_{I_j} \in \mathbb{P}_k(I_j)\}. \quad (3.3.10)$$

Assume $\lambda = \frac{\Delta t}{h}$ is the mesh ratio, where $h = \max_j |I_j|$. Provided that scheme (3.3.9) has the following property: there exists $\lambda_0$, and a test set $S$ such that if

$$\lambda \leq \lambda_0 \quad \text{and} \quad w^n_h(x) \in \Sigma \quad \text{for} \quad x \in S \quad (3.3.11)$$

then

$$w^{n+1}_h \in \Sigma_0; \quad (3.3.12)$$

the limiter can then be applied with $I$ replaced by $S_I := S \cap I_j$ in (3.3.3), i.e.,

$$r_{\text{max}} = \max_{x \in S_I} r(w_h(x)), \quad s_{\text{min}} = \min_{x \in S_I} s(w_h(x)). \quad (3.3.13)$$

The algorithm can be described as follows:

**Step 1.** Initialization: take the piecewise $L^2$ projection of $w_0$ onto $V_h$, such that

$$\int_{I_j} (w^0_h(x) - w_0(x)) \phi(x) dx = 0, \quad \forall \phi \in V_h. \quad (3.3.14)$$
Also from \( w_0 \), we compute \( r_0 \) and \( s_0 \) as defined in (3.2.3) to determine the invariant region \( \Sigma \).

**Step 2.** Impose the modified limiter (3.3.1), (3.3.2) with (3.3.13) on \( w_h^n \) for \( n = 0, 1, \ldots \) to obtain \( \tilde{w}_h^n \).

**Step 3.** Update by the scheme:
\[
    w_h^{n+1} = \mathcal{L}(\tilde{w}_h^n).
\] (3.3.15)

Return to **Step 2**.

**Remark 3.3.4.** With the modification defined in (3.3.13), both (1) and (3) in Theorem 3.3.4 remain valid, and the modified polynomials lie within the invariant region \( \Sigma \) only for \( x \in S_I \). Moreover, the limiter (3.3.1), (3.3.2) with (3.3.13) can enhance the efficiency of computation, and we shall use this modified IRP limiter in our numerical experiments.

### 3.4 IRP high order schemes

In this section we discuss the IRP property of high order schemes for both the p-system (3.1.2) and the viscous p-system (3.1.3). These systems are known to share the same invariant region \( \Sigma \) (see [23]). For simplicity, we will assume periodic boundary conditions from now on.

#### 3.4.1 IRP schemes for (3.1.2)

Rewrite the p-system into a compact form
\[
    w_t + F(w)_x = 0,
\] (3.4.1)

with \( w = (v, u)^T \) and \( F(w) = (-u, p(v))^T \). We begin with the first order finite volume scheme
\[
    w_j^{n+1} = w_j^n - \lambda \left( \hat{F}(w_j^n, w_{j+1}^n) - \hat{F}(w_{j-1}^n, w_j^n) \right),
\] (3.4.2)

with the Lax-Friedrich flux defined by
\[
    \hat{F}(w_i^n, w_{i+1}^n) = \frac{1}{2} \left( F(w_i^n) + F(w_{i+1}^n) - \sigma(w_{i+1}^n - w_i^n) \right), \quad i = j - 1, j,
\] (3.4.3)
where \( w_j^n \) is approximation to the average of \( w \) on \( I_j = [x_{j-1/2}, x_{j+1/2}] \) at time level \( n \), and \( \sigma \) is a positive number dependent upon \( w_i \) for \( l = j-1, j, j+1 \). Given \( w_j^n \in \Sigma \), we want to find a proper \( \sigma \) and a CFL condition so that \( w_j^{n+1} \in \Sigma \).

We can rewrite (3.4.2) as

\[
 w_j^{n+1} = (1 - \lambda \sigma) w_j^n + \lambda \sigma w^*,
\]

where

\[
 w^* = \left( \frac{w_{j-1}^n + w_{j+1}^n}{2} - \frac{F(w_{j+1}^n) - F(w_{j-1}^n)}{2\sigma} \right).
\]

It is known, see [14, Section 14.1] for example, that for \( \sigma \geq \max_{v_j^n, v_{j+1}^n} \sqrt{-p'(v)} \),

\[
 w^* = \frac{1}{2\sigma} \int_{-\sigma}^{\sigma} w(t, t) d\xi,
\]

where \( w(x, t) \) is the exact Riemann solution to (3.4.1) subject to initial data

\[
 w(x, 0) = \begin{cases} 
 w_{j-1}^n, & x < 0, \\
 w_{j+1}^n, & x > 0.
\end{cases}
\]

For \( w_{j+1}^n \in \Sigma \), we have \( w(x, t) \in \Sigma \) since \( \Sigma \) is the invariant region of the p-system. Therefore \( w^* \) lies in \( \Sigma_0 \) by Lemma 3.2.2. Since \( w_j^{n+1} \) is a convex combination of two vectors \( w \in \Sigma \) and \( w^* \in \Sigma_0 \) for \( \lambda \sigma \leq 1 \), we then have \( w_j^{n+1} \in \Sigma_0 \). In summary, we have the following.

**Theorem 3.4.1** (First order scheme). The first order scheme (3.4.2), (3.4.3) with \( \sigma \geq \max_{v_j^n, v_{j+1}^n} \sqrt{-p'(v)} \), preserves the invariant region \( \Sigma \), i.e.,

\[
  \text{if } w_i^n \in \Sigma \text{ for } i = j, j \pm 1, \text{ then } w_j^{n+1} \in \Sigma_0,
\]

under the CFL condition

\[
  \lambda \sigma \leq 1 \quad \text{with} \quad \sigma \geq \max_{v_j^n, v_{j+1}^n} \sqrt{-p'(v)}.
\]

**Remark 3.4.1.** Note that for scheme (3.4.2), the flux parameter \( \sigma \) at each \( x_{j+1/2} \) needs to be dependent on \( w_i^n \) with \( i = j-1, j, j+1, j+2 \), instead of the usual local Lax-Friedrichs flux with
σ depending on \( w^n_i \) with \( i = j, j + 1 \). An alternative proof of the positivity-preserving property was given in the appendix in [21] for compressible Euler equations; their technique when applied to (3.4.2) with the global Lax-Friedrich flux leads to \( \lambda \sigma \leq \frac{1}{2} \). In contrast, the CFL condition (3.4.5) is more relaxed. We also note that using the quadratic structure of the flux function \( F \) in the full compressible Euler equation, Zhang was able to obtain a relaxed CFL in [27] with the local Lax-Friedrich flux by direct verification without reference to the Riemann solution.

We next consider a \((k+1)\)th-order scheme with reconstructed polynomials or approximation polynomials of degree \( k \). With forward Euler time discretization, the cell average evolves by

\[
\bar{w}^{n+1}_j = \bar{w}^n_j - \lambda [\hat{F}(w^{−1/2}_j, w^{+1/2}_j) - \hat{F}(w^{−1/2}_{j−1}, w^{+1/2}_{j+1})],
\]

where \( \bar{w}^n_j \) is the cell average of \( w^n_h \) on \( I_j \) at time level \( n \), \( w^{±1/2}_j \) are approximations to the point value of \( w \) at \( x_{j+1/2} \) and time level \( n \) from the left and from the right respectively. The local Lax-Friedrichs flux is taken with

\[
\sigma \geq \max_{v_{j±1/2}^{−1/2}, v_{j±1/2}^{+1/2}} \sqrt{-p′(v)}.
\]

We consider an \( N \)-point Legendre Gauss-Lobatto quadrature rule on \( I_j \), with quadrature weights \( \hat{ω}_i \) on \([-\frac{1}{2}, \frac{1}{2}]\) such that \( \sum_{i=1}^{N} \hat{ω}_i = 1 \), which is exact for integrals of polynomials of degree up to \( k \), if \( 2N - 3 \geq k \). Denote these quadrature points on \( I_j \) as

\[
S_j^C := \{ \hat{x}^i_j, 1 \leq i \leq N \},
\]

where \( \hat{x}^1_j = x_{j−1/2} \) and \( \hat{x}^N_j = x_{j+1/2} \). The cell average decomposition then takes the form

\[
\bar{w}^n_j = \sum_{i=2}^{N−1} \hat{ω}_i w^n_h(\hat{x}^i_j) + \hat{ω}_1 w^{+1/2}_j \bar{w}^{-1/2}_{j−1} + \hat{ω}_N w^{+1/2}_j \bar{w}^{-1/2}_{j+1},
\]

where it is known that \( \hat{ω}_1 = \hat{ω}_N = 1/(N(N−1)) \). Hence (3.4.6) can be rewritten as a linear convex combination of the form

\[
\bar{w}^{n+1}_j = \sum_{i=2}^{N−1} \hat{ω}_i w^n_h(\hat{x}^i_j) + \hat{ω}_1 K_1 + \hat{ω}_N K_N,
\]
where

\[ K_1 = w_{j-\frac{1}{2}}^+ - \frac{\lambda}{\hat{\omega}_1} \left( \hat{F}(w_{j-\frac{1}{2}}^+, w_{j+\frac{1}{2}}^-) - \hat{F}(w_{j-\frac{1}{2}}^-, w_{j-\frac{1}{2}}^+) \right), \]

\[ K_N = w_{j+\frac{1}{2}}^- - \frac{\lambda}{\hat{\omega}_N} \left( \hat{F}(w_{j+\frac{1}{2}}^-, w_{j+\frac{1}{2}}^+) - \hat{F}(w_{j-\frac{1}{2}}^+, w_{j+\frac{1}{2}}^-) \right) \]

are of the same type as the first order scheme (3.4.2). Such decomposition of (3.4.6), first introduced by Zhang and Shu ([29]) for the compressible Euler equation, suffices for us to conclude the following result.

**Theorem 3.4.2** (High order scheme). A sufficient condition for \( \bar{w}_j^{n+1} \in \Sigma_0 \) by scheme (3.4.6) with (3.4.7) is

\[ w^n_h(x) \in \Sigma \text{ for } x \in S^C_j \]

under the CFL condition

\[ \lambda \sigma \leq \frac{1}{N(N-1)} \text{ with } \sigma \geq \max_{v_{j\pm1/2}^-, v_{j\pm1/2}^+} \sqrt{-p'(v)}, \quad N = \left\lceil \frac{k+3}{2} \right\rceil. \]  

(3.4.10)

**Remark 3.4.2.** For third order schemes, an alternative decomposition of the cell average can be found through Lagrangian interpolation on the test set \( S^C_j = x_j + \frac{\Delta x}{2} \{ -\gamma, 1 \} \):

\[ \bar{w}_j^n = \frac{1+3\gamma}{6(1+\gamma)} w_{j-\frac{1}{2}}^+ + \frac{2}{3(1-\gamma^2)} w_h^n \left( x_j + \frac{\Delta x}{2} \gamma \right) + \frac{1-3\gamma}{6(1-\gamma)} w_{j+\frac{1}{2}}^-, \]  

(3.4.11)

where \( |\gamma| \leq \frac{1}{3} \) to ensure non-negative weights. Theorem 3.4.2 still holds under this choice of test set.

**Remark 3.4.3.** Even though we proved both Theorem 3.4.1 and Theorem 3.4.2 only for the p-system (3.1.2), the results obtained actually hold for all one-dimensional hyperbolic conservation laws (3.1.1) equipped with a convex invariant region, as long as the flux parameter \( \sigma \) is chosen so that

\[ \sigma \geq \lambda_{\text{max}}, \]

where \( \lambda_{\text{max}} \) is the maximum wave speed (largest eigenvalues of the Jacobian matrix of \( F(w) \)). A direct calculation shows that \( \lambda_{\text{max}} = |u| + \sqrt{p'(\rho)} \) for (3.2.5), and \( \lambda_{\text{max}} = |u| + \sqrt{gh} \) for (3.2.7).
3.4.2 IRP DG schemes for (3.1.3)

The p-system with artificial diffusion can be rewritten as

$$w_t + F(w)_x = \varepsilon w_{xx} \quad (3.4.12)$$

with $w = (v, u)^\top$ and $F(w) = (-u, p(v))^\top$. The DG scheme for (3.4.12) is to find $w_h \in V_h$ such that

$$\int_{I_j} (w_h)_t \phi dx = \int_{I_j} F(w_h) \phi_x dx - \hat{F}(w_h, w_h) \phi|_{\partial I_j}$$

$$- \int_{I_j} \epsilon(w_h)_x \phi_x dx + \epsilon \left( (\widehat{w_h})_x \phi + (w_h - \{w_h\}) \phi_x \right)|_{\partial I_j} \quad \forall \phi \in V_h, \quad (3.4.13)$$

where $\hat{F}$ is the local Lax-Friedrich flux for the convection part as defined in (3.4.3), see [3]; while the diffusive numerical flux $(\widehat{w_h})_x$ is chosen, following [18, 17], as

$$(\widehat{w_h})_x = \beta_0 \frac{[w_h]}{\Delta x} + \{ (w_h)_x \} + \beta_1 \Delta x [ (w_h)_{xx} ], \quad (3.4.14)$$

where $[.]$ denotes the jump of the function and $\{::\}$ denotes the average of the function across the cell interface. The flux parameters $\beta_0$ and $\beta_1$ are to be chosen from an appropriate range in order to achieve the desired IRP property.

With forward Euler time discretization of (3.4.13), the cell average evolves by

$$w_{j}^{n+1} = \bar{w}_j^n - \lambda \hat{F}(w_h, w_h) \big|_{\partial I_j} + \epsilon \mu \Delta x (\widehat{w_h})_x |_{\partial I_j}, \quad (3.4.15)$$

where the mesh ratios $\mu = \frac{\Delta t}{\Delta x^2}$ and $\lambda = \frac{\Delta t}{\Delta x}$.

We want to find sufficient conditions on $\lambda, \mu$ and a test set $S$ such that if $w_h^n(x) \in \Sigma$ on $S$, then $w_{j}^{n+1} \in \Sigma_0$ for all $j$. To do so, we rewrite (3.4.15) as

$$w_{j}^{n+1} = \frac{1}{2} C_j + \frac{1}{2} D_j, \quad (3.4.16)$$

where

$$C_j = \bar{w}_j^n - 2\lambda \left( \hat{F}(w_{j+\frac{1}{2}}^{-}, w_{j+\frac{1}{2}}^{+}) - \hat{F}(w_{j-\frac{1}{2}}^{-}, w_{j-\frac{1}{2}}^{+}) \right),$$

$$D_j = \bar{w}_j^n + 2\epsilon \mu \left( \Delta x (w_h^n)_x |_{x_{j+\frac{1}{2}}} - \Delta x (\widehat{w}_h^n)_x |_{x_{j-\frac{1}{2}}} \right).$$
Recall the result obtained in Theorem 3.4.2, we see that for any \( k > 0 \), the sufficient condition for \( C_j \in \Sigma_0 \) is the following:

\[
\mathbf{w}_h^n(x) \in \Sigma \quad \text{for } x \in S_j^C
\]

and the CFL condition

\[
\lambda \sigma \leq \frac{1}{2N(N - 1)} \quad \text{with} \quad \sigma \geq \max_{v_j^\pm 1/2} \sqrt{-p'(v)}, \quad (3.4.17)
\]

where the \( N \geq (3 + k)/2 \) is required to hold. The remaining task is to find sufficient conditions so that \( D_j \in \Sigma_0 \). These together when combined with (3.4.16) will give \( \bar{w}_j^{n+1} \in \Sigma_0 \) as desired.

In order to ensure that \( D_j \in \Sigma_0 \), we follow Liu and Yu in [19] so as to obtain some sufficient conditions for \( k = 1, 2 \), respectively. The results are summarized in the following two lemmas.

**Lemma 3.4.3.** Consider scheme (3.4.15) with \( k = 1 \) and \( \beta_0 \geq 1/2 \). The sufficient condition for \( D_j \in \Sigma_0 \) is

\[
\mathbf{w}_h^n(x) \in \Sigma \quad \text{for } x \in \bigcup_{i=j-1}^{j+1} S_i^D,
\]

where

\[
S_i^D = x_i + \Delta x \frac{\gamma}{2} \{-\gamma, \gamma\}
\]

for \( \gamma \in [-1, 1] \setminus \{0\} \) with

\[
\left| 1 - \frac{1}{\beta_0} \right| \leq |\gamma| \leq 1, \quad (3.4.18)
\]

under the condition \( \mu \leq \frac{1}{4\varepsilon \beta_0} \).

**Lemma 3.4.4.** Consider scheme (3.4.15) with \( k = 2 \) and

\[
\beta_0 \geq 1, \quad \frac{1}{8} \leq \beta_1 \leq \frac{1}{4}, \quad (3.4.19)
\]

The sufficient condition for \( D_j \in \Sigma_0 \) is

\[
\mathbf{w}_h^n(x) \in \Sigma \quad \text{for } x \in \bigcup_{i=j-1}^{j+1} S_i^D,
\]
where

\[ S_i^D = x_i + \frac{\Delta x}{2} \{ -1, \gamma, 1 \} \]

for \( \gamma \in [-1, 1] \) satisfying

\[ |\gamma| < \frac{1}{3} \text{ and } |\gamma| \leq 8\beta_1 - 1 \quad (3.4.20) \]

under the CFL conditions \( \mu \leq \mu_0 \), where

\[ \mu_0 = \frac{1}{12\epsilon} \min \left\{ \frac{1 \pm 3\gamma}{(1 \pm \gamma)^2 - \theta} \right\}, \quad \theta := 2 - 8\beta_1 \in [0, 1]. \quad (3.4.21) \]

For reader’s convenience, we outline the related calculations leading to the above results in appendix 3.6.

Remark 3.4.4. In [19], Liu and Yu proposed a third order maximum-principle-preserving DDG scheme over Cartesian meshes for the linear Fokker-Planck equation

\[ u_t = \nabla_x \cdot (\nabla_x u + \nabla_x V(x) u), \quad (3.4.22) \]

where the potential \( V \) is given, provided the flux parameter \( (\beta_0, \beta_1) \) falls into the range \( \beta_0 \geq 1 \) and \( \beta_1 \in [1/8, 1/4] \). The maximum-principle for (3.4.22) means that if \( u_0 \in [c_1, c_2]e^{-V(x)} \), then \( u(x, t) \in [c_1, c_2]e^{-V(x)} \) for all \( t > 0 \), which implies the usual maximum-principle for diffusion \( u(x, t) \in [c_1, c_2] \) for all \( t > 0 \). Extension to unstructured meshes is non-trivial, we refer to [2] for some recent results in solving diffusion equations over triangular meshes.

Combining (3.4.17), with Lemma 3.4.3 and Lemma 3.4.4, we are able to establish the following result.

**Theorem 3.4.5.** A sufficient condition for \( \bar{w}_j^{n+1} \in \Sigma_0 \) in scheme (3.4.15) is

\[ w_h^n(x) \in \Sigma \quad \text{for} \quad x \in \bigcup_{i=j-1}^{j+1} S_i, \]

under the CFL conditions \( \lambda \leq \lambda_0 \) and \( \mu \leq \mu_0 \), where

(i) for second order scheme with \( \beta_0 \geq \frac{1}{2} \),

\[ S_i = x_i + \frac{\Delta x}{2} \{ -1, 1 \}, \quad \lambda_0 = \frac{1}{4\sigma}, \quad \mu_0 = \frac{1}{4\epsilon \beta_0}; \]
(ii) for third order scheme with $\beta_0 \geq 1$ and $\frac{1}{8} \leq \beta_1 \leq \frac{1}{4}$,

$$S_i = x_i + \frac{\Delta x}{2} \{-1, \gamma, 1\},$$

$$\lambda_0 = \frac{1}{12\sigma}, \quad \mu_0 = \frac{1}{12\epsilon} \min\left\{\frac{1 \pm 3\gamma}{(1 \pm \gamma)\beta_0 - \theta}, 2\right\}, \quad \theta := 2 - 8\beta_1 \in [0, 1],$$

where $|\gamma| < \frac{1}{3}$.

Proof. (i) For the second order scheme with $\beta_0 \geq \frac{1}{2}$, we have $S_j^C = x_j + \frac{\Delta x}{2} \{-1, 1\}$ and $S_j^D = x_j + \frac{\Delta x}{2} \{-\gamma, \gamma\}$, where $\gamma$ satisfies \eqref{eq:gamma}. Therefore, we take $\gamma = 1$ so that $S_j = S_j^C \cup S_j^D = S_j^C$.

(ii) For the third order scheme with $\beta_0 \geq 1$ and $\frac{1}{8} \leq \beta_1 \leq \frac{1}{4}$, $S_j^D = x_j + \frac{\Delta x}{2} \{-1, \gamma, 1\}$, where $\gamma$ satisfies \eqref{eq:gamma}. While for convection part, we consider the alternative decomposition of the cell average \eqref{eq:cell_average_alt} in Remark 3.4.2, so that $S_j = S_j^C = S_j^D$.

The CFL conditions are obtained directly by using \eqref{eq:cfl} and the results stated in two Lemmas above.

Remark 3.4.5. It is known that the high order Strong Stability Preserving (SSP) time discretizations are convex combinations of forward Euler, therefore Theorems 3.4.2 and 3.4.5 remain valid when using high order SSP schemes. In our numerical experiments we will adopt such high order SSP time discretization so as to also achieve high order accuracy in time.

3.5 Numerical examples

In this section, we present numerical examples for testing the IRP limiter and the performance of the IRP DG scheme \eqref{eq:irp_dg} with the local Lax-Friedrich flux \eqref{eq:lax_friedrich}, using the third order SSP Runge-Kutta (RK3) method for time discretization. $\gamma = 1.4$ is taken in all of the examples.

The semi-discrete DG scheme is a closed ODE system

$$\frac{d}{dt} W = L(W),$$
where \( \mathbf{W} \) consists of the unknown coefficients of the spatial basis, and \( L \) is the corresponding spatial operator. The third order SSP RK3 in [8, 22] reads as

\[
\begin{align*}
\mathbf{W}^{(1)} &= \mathbf{W}^n + \Delta t L(\mathbf{W}^n), \\
\mathbf{W}^{(2)} &= \frac{3}{4} \mathbf{W}^n + \frac{1}{4} \mathbf{W}^{(1)} + \frac{1}{4} \Delta t L(\mathbf{W}^{(1)}), \\
\mathbf{W}^{n+1} &= \frac{1}{3} \mathbf{W}^n + \frac{2}{3} \mathbf{W}^{(2)} + \frac{2}{3} \Delta t L(\mathbf{W}^{(2)}).
\end{align*}
\]

We apply the limiter at each time stage in the RK3 method. Notice that (3.4) is a linear convex combination of Euler Forward method, therefore the invariant region is preserved by the full scheme if it is preserved at each time stage.

In the first four examples, the initial condition is chosen as

\[ v_0(x) = 2 - \sin(x), \quad u_0(x) = 1, \quad x \in [0, 2\pi], \]

and the boundary condition is periodic. Using (3.2.3) we obtain \( r_0 = 1 \) and \( s_0 = 1 \), so that the invariant region \( \Sigma = \{(v, u) \mid r(v, u) \leq 1, s(v, u) \geq 1\} \). We fix the initial mesh \( h = \frac{2\pi}{32} \) for all four examples.

**Example 1. Accuracy test of the limiter**

We test the performance of the IRP limiter introduced in Section 3.3 by comparing the accuracy of the \( L^2 \) projection (3.3.14) with and without limiter (3.3.1). From Tables 3.1 and 3.2, we see that the IRP limiter preserves the accuracy of high order approximations well.

<table>
<thead>
<tr>
<th>( P^1 )</th>
<th>Projection without limiter</th>
<th></th>
<th></th>
<th>Projection with limiter</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta x )</td>
<td>( L^\infty ) error</td>
<td>Order</td>
<td>( L^1 ) error</td>
<td>Order</td>
<td>( L^\infty ) error</td>
<td>Order</td>
</tr>
<tr>
<td>( h )</td>
<td>1.34E-04</td>
<td>/</td>
<td>8.53E-05</td>
<td>/</td>
<td>6.70E-04</td>
<td>/</td>
</tr>
<tr>
<td>( h/2 )</td>
<td>3.35E-05</td>
<td>2.00</td>
<td>2.13E-05</td>
<td>2.00</td>
<td>1.67E-04</td>
<td>2.00</td>
</tr>
<tr>
<td>( h/4 )</td>
<td>8.37E-06</td>
<td>2.00</td>
<td>5.33E-06</td>
<td>2.00</td>
<td>4.18E-05</td>
<td>2.00</td>
</tr>
<tr>
<td>( h/8 )</td>
<td>2.09E-06</td>
<td>2.00</td>
<td>1.33E-06</td>
<td>2.00</td>
<td>1.05E-05</td>
<td>2.00</td>
</tr>
<tr>
<td>( h/16 )</td>
<td>5.23E-07</td>
<td>2.00</td>
<td>3.33E-07</td>
<td>2.00</td>
<td>2.62E-06</td>
<td>2.00</td>
</tr>
</tbody>
</table>
Table 3.2 Accuracy of the $L^2$ projection using $P^2$ approximation.

<table>
<thead>
<tr>
<th>$P^2$</th>
<th>Projection without limiter</th>
<th>Projection with limiter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L^\infty$ error</td>
<td>Order</td>
</tr>
<tr>
<td>$\Delta x$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>9.16E-06 / 5.86E-06 /</td>
<td></td>
</tr>
<tr>
<td>h/2</td>
<td>1.15E-06 3.00</td>
<td>7.32E-07 3.00</td>
</tr>
<tr>
<td>h/4</td>
<td>1.44E-07 3.00</td>
<td>9.15E-08 3.00</td>
</tr>
<tr>
<td>h/8</td>
<td>1.80E-08 3.00</td>
<td>1.14E-08 3.00</td>
</tr>
<tr>
<td>h/16</td>
<td>2.25E-09 3.00</td>
<td>1.43E-09 3.00</td>
</tr>
</tbody>
</table>

**Example 2. Accuracy test of the IRP DG-RK3 scheme solving (3.1.2)**

We apply the IRP DG-RK3 scheme to solve (3.1.2), with time step set as $\Delta t = \min(\frac{\Delta x}{3\sigma}, \Delta x^2)$ when using $P^1$ polynomials, and $\Delta t = \frac{\Delta x}{6\sigma}$ when using $P^2$ polynomials, where

$$\sigma = \max_j \max \left\{ \sqrt{-p'(v^-_{j+\frac{1}{2}})}, \sqrt{-p'(v^+_{j+\frac{1}{2}})} \right\}. \quad (3.5.2)$$

The numerical results evaluated at $T = 0.1$ are given in Table 3.3 and Table 3.4, where the reference solution is calculated using the fourth order DG scheme on 4096 meshes. These results show that the IRP DG scheme maintains the optimal order of accuracy in $L^1$ norm. The order of accuracy in $L^\infty$ norm is compromised in some cases.

Table 3.3 Accuracy test with $P^1$-DG scheme for (3.1.2).

<table>
<thead>
<tr>
<th>$P^1$-DG</th>
<th>DG without limiter</th>
<th>DG with limiter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta x$</td>
<td>$L^\infty$ error</td>
<td>Order</td>
</tr>
<tr>
<td>h</td>
<td>1.41E-03 / 1.07E-03 /</td>
<td></td>
</tr>
<tr>
<td>h/2</td>
<td>3.39E-04 2.05</td>
<td>2.70E-04 1.99</td>
</tr>
<tr>
<td>h/4</td>
<td>8.30E-05 2.03</td>
<td>6.59E-05 2.04</td>
</tr>
<tr>
<td>h/8</td>
<td>1.98E-05 2.07</td>
<td>1.53E-05 2.10</td>
</tr>
<tr>
<td>h/16</td>
<td>4.51E-06 2.14</td>
<td>3.27E-06 2.23</td>
</tr>
</tbody>
</table>

**Example 3. Accuracy test of the IRP DG-RK3 scheme solving (3.1.3)**

We apply the IRP DG-RK3 scheme to solve system (3.1.3) with $\epsilon = 0.01$. Flux parameters and
Table 3.4  Accuracy test with $P^2$-DG scheme for (3.1.2).

<table>
<thead>
<tr>
<th>$P^2$-DG</th>
<th>DG without limiter</th>
<th>DG with limiter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta x$</td>
<td>$L^\infty$ error</td>
<td>Order</td>
</tr>
<tr>
<td>h</td>
<td>2.21E-05 / 1.81E-05 /</td>
<td></td>
</tr>
<tr>
<td>h/2</td>
<td>3.14E-06 2.82 2.75E-06 2.72</td>
<td>8.85E-05 1.66 7.54E-06 2.33</td>
</tr>
<tr>
<td>h/4</td>
<td>3.62E-07 3.12 3.13E-07 3.14</td>
<td>1.34E-05 2.72 8.71E-07 3.11</td>
</tr>
<tr>
<td>h/8</td>
<td>3.87E-08 3.23 3.27E-08 3.26</td>
<td>1.50E-06 3.16 9.84E-08 3.15</td>
</tr>
<tr>
<td>h/16</td>
<td>3.25E-09 3.57 2.75E-09 3.58</td>
<td>3.01E-07 2.31 1.16E-08 3.09</td>
</tr>
</tbody>
</table>

the time step are chosen as

$$P^1: \quad \Delta t = \min \left\{ \frac{\Delta x}{4\sigma}, \frac{\Delta x^2}{6\epsilon\beta_0}, \frac{\Delta x^{\frac{3}{2}}}{8} \right\}, \quad \beta_0 = 2,$$

$$P^2: \quad \Delta t = \min \left\{ \frac{\Delta x}{12\sigma}, \frac{\Delta x^2}{12\epsilon(8\beta_1 + \beta_0 - 2)} \right\}, \quad (\beta_0, \beta_1) = (2, \frac{1}{4}),$$

where $\sigma$ is defined in (3.5.2). The reference solution is computed by $P^2$-DG using 4096 meshes. Both errors and orders of numerical solutions at $T = 0.1$ are given in Table 3.5 and Table 3.6, from which we see that our IRP DG-RK3 scheme maintains the desired order of accuracy. A noticeable difference between Example 2 and Example 3 is that the accuracy result is better for the latter.

One observation that might explain this is that in Example 3, the limiter is imposed only in the first several time steps, yet in Example 2, the limiter is called more frequently in time.

Table 3.5  Accuracy test with $P^1$-DG scheme for (3.1.3).

<table>
<thead>
<tr>
<th>$P^1$-DG</th>
<th>DG without limiter</th>
<th>DG with limiter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta x$</td>
<td>$L^\infty$ error</td>
<td>Order</td>
</tr>
<tr>
<td>h</td>
<td>1.27E-03 / 1.03E-03 /</td>
<td></td>
</tr>
<tr>
<td>h/2</td>
<td>2.94E-04 2.11 2.47E-04 2.05</td>
<td>3.28E-04 2.59 2.48E-04 2.07</td>
</tr>
<tr>
<td>h/4</td>
<td>6.63E-05 2.15 5.61E-05 2.14</td>
<td>6.63E-05 2.31 5.61E-05 2.14</td>
</tr>
<tr>
<td>h/8</td>
<td>1.46E-05 2.18 1.19E-05 2.24</td>
<td>1.46E-05 2.18 1.19E-05 2.24</td>
</tr>
<tr>
<td>h/16</td>
<td>2.86E-06 2.35 2.23E-06 2.42</td>
<td>2.86E-06 2.35 2.23E-06 2.42</td>
</tr>
</tbody>
</table>

Example 4. Convergence of viscous profiles to the entropy solution

In this example, we illustrate the convergence of viscous solutions to the entropy solution when taking $\epsilon \sim O(\Delta x^r)$ for some $r > 0$. 
Table 3.6  Accuracy test with $P^2$-DG scheme for (3.1.3).

<table>
<thead>
<tr>
<th>$P^2$-DG</th>
<th>DG without limiter</th>
<th>DG with limiter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta x$</td>
<td>$L^\infty$ error</td>
<td>Order</td>
</tr>
<tr>
<td>h</td>
<td>2.49E-05 /</td>
<td>2.41E-05 /</td>
</tr>
<tr>
<td>h/2</td>
<td>5.26E-06 2.24</td>
<td>4.72E-06 2.35</td>
</tr>
<tr>
<td>h/4</td>
<td>1.27E-06 2.05</td>
<td>9.49E-07 2.32</td>
</tr>
<tr>
<td>h/8</td>
<td>2.37E-07 2.42</td>
<td>1.74E-07 2.44</td>
</tr>
<tr>
<td>h/16</td>
<td>3.64E-08 2.70</td>
<td>2.75E-08 2.66</td>
</tr>
<tr>
<td>h/32</td>
<td>4.70E-09 2.96</td>
<td>3.63E-09 2.92</td>
</tr>
</tbody>
</table>

We apply $P^k$-DG to (3.1.3), with the reference solution obtained from $P^3$-DG scheme to (3.1.2) on 4096 meshes. The time steps are chosen as in Example 3. From Table 3.7 and Table 3.8 we observe the optimal convergence as long as $r \geq k + 1$.

Table 3.7  Convergence test

<table>
<thead>
<tr>
<th>$P^1$-DG</th>
<th>$\epsilon = \Delta x^2$</th>
<th>$\epsilon = \Delta x^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta x$</td>
<td>$L^\infty$ error</td>
<td>Order</td>
</tr>
<tr>
<td>h</td>
<td>9.72E-04 /</td>
<td>4.59E-04 /</td>
</tr>
<tr>
<td>h/2</td>
<td>2.65E-04 1.88</td>
<td>1.30E-04 1.82</td>
</tr>
<tr>
<td>h/4</td>
<td>6.83E-05 1.95</td>
<td>3.65E-05 1.83</td>
</tr>
<tr>
<td>h/8</td>
<td>1.70E-05 2.01</td>
<td>1.02E-05 1.83</td>
</tr>
<tr>
<td>h/16</td>
<td>4.08E-06 2.06</td>
<td>2.96E-06 1.79</td>
</tr>
</tbody>
</table>

Table 3.8  Convergence test

<table>
<thead>
<tr>
<th>$P^2$-DG</th>
<th>$\epsilon = \Delta x^3$</th>
<th>$\epsilon = \Delta x^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta x$</td>
<td>$L^\infty$ error</td>
<td>Order</td>
</tr>
<tr>
<td>h</td>
<td>2.60E-04 /</td>
<td>1.71E-04 /</td>
</tr>
<tr>
<td>h/2</td>
<td>3.24E-05 3.00</td>
<td>2.17E-05 2.98</td>
</tr>
<tr>
<td>h/4</td>
<td>4.03E-06 3.01</td>
<td>2.70E-06 3.01</td>
</tr>
<tr>
<td>h/8</td>
<td>5.01E-07 3.01</td>
<td>3.38E-07 3.00</td>
</tr>
<tr>
<td>h/16</td>
<td>6.19E-08 3.02</td>
<td>4.21E-08 3.00</td>
</tr>
</tbody>
</table>
In the next two examples, we test the IRP DG-RK3 scheme to solve (3.1.2) with Riemann initial data. It is known in [23] that for each given state \( w_l = (v_l, u_l) \), two shock curves are governed by

\[
    u = u_l - \sqrt{(v - v_l)(p(v_l) - p(v))},
\]

and two rarefaction curves are given by

\[
    u = u_l + \int_{v_l}^{v} \sqrt{-p'(y)} dy,
\]

\[
    u = u_l - \int_{v_l}^{v} \sqrt{-p'(y)} dy.
\]

From these curves, we identify two cases for our testing.

**Example 5. Shock-rarefaction wave**

Consider the following Riemann initial data,

\[
    w_0(x) = \begin{cases} 
    (1, 0), & x < 0, \\
    (\frac{1}{4}, 0.1053), & x > 0. 
    \end{cases}
\]

Following the procedure in [23, Section A, Chapter 17] we see that such initial data will induce a composite wave, a back shock wave followed by a front rarefaction wave. Therefore, we solve the equations (3.5.3) and (3.5.5) for the intermediate state \( (\frac{1}{2}, -0.9053) \) and construct the exact solution. The invariant region is given by \( \Sigma = \{(v, u) \mid r(v, u) \leq 0.1053, \ s(v, u) \geq 0.1053\} \). We test the \( P^1\)-DG scheme over \([-1, 1]\) at final time \( T = 0.1 \) on 128 meshes. We see in Figure 3.2 that the oscillations around discontinuities are either damped or oppressed by the IRP limiter.

**Example 6. Rarefaction-shock wave**

Consider the following Riemann initial data,

\[
    w_0(x) = \begin{cases} 
    (1, 0), & x < 0, \\
    (2, -0.3509), & x > 0. 
    \end{cases}
\]

A similar checking tells that such initial data will induce a composite wave, a back rarefaction wave followed by a front shock wave. From (3.5.3) and (3.5.4) we find the intermediate state \( (1.2, 0.2118) \)
Figure 3.2  Shock-rarefaction wave problem. Exact solution (solid line) vs numerical solution (dots); Top: DG without limiter; Bottom: DG with limiter.
and construct the exact solution. The invariant region is given by $\Sigma = \{(v, u) \mid r(v, u) \leq 0, s(v, u) \geq 0\}$. We test the $P^1$-DG scheme over $[-1, 1]$ at final time $T = 0.1$ on 128 meshes. It shows from Figure 3.3 that oscillations occur near discontinuities. However, after the limiter is applied, the oscillations are reduced.

Figure 3.3 Rarefaction-shock wave problem. Exact solution (solid line) vs numerical solution (dots); Top: DG without limiter; Bottom: DG with limiter.

Remark 3.5.1. The IRP limiter presented in this work is a mild limiter. As we can see in Example 5 and 6 that not all oscillations can be completely damped, even though the invariant region is preserved. For big oscillations, some stronger limiters might be needed. For example, the weighted essentially non-oscillatory (WENO) limiter developed by Zhong and Shu in [32].
3.6 Conclusions remarks

In this paper, we introduced an explicit IRP limiter for the p-system, including the isentropic gas dynamic equation, and the viscous p-system. The limiter itself is shown to preserve the accuracy of high order approximation in general cases through both rigorous analysis and numerical tests. We also specify sufficient conditions, including CFL conditions and test sets, for high order IRP DG schemes with Euler forward time discretization. Numerical tests on such schemes with RK3 time discretization confirm the desired properties. An interesting observation is that the IRP DG scheme solving the viscous p-system is more accurate than the one solving the p-system. For the latter the IRP limiter is called much more frequently, indicating a possible error accumulation through the evolution in time.

Acknowledgements

The authors would like to thank Professor Chi-Wang Shu for helpful discussions on the subtlety of the limiter in the system case, which led to our construction of two examples in Appendix B. This research was supported by the National Science Foundation under Grant DMS1312636.

Appendix A. Proofs of Lemma 3.4.3 and 3.4.4

Proof of Lemma 3.4.3: Set

\[ p_j(\xi) = w^n_h(x_j + \frac{\Delta x}{2} \xi) = w^n_h(x)|_{I_j}, \quad \xi \in [-1, 1], \]

then for any \( \gamma \in [-1, 1] \setminus \{0\} \), we have

\[ p_j(\xi) = \frac{1}{2} \left( 1 - \frac{\xi}{\gamma} \right) p_j(\gamma) + \frac{1}{2} \left( 1 + \frac{\xi}{\gamma} \right) p_j(-\gamma), \]

from which the cell average can be expressed as \( \bar{w}_j^n = \frac{1}{2}p_j(-\gamma) + \frac{1}{2}p_j(\gamma) \). A direct calculation of \( D_j \) in (3.4.16) using flux (3.4.14) gives

\[
D_j = 2\epsilon \mu \alpha_1 p_{j+1}(-\gamma) + 2\epsilon \mu p_{j+1}(\gamma) + 2\epsilon \mu \alpha_2 p_{j-1}(-\gamma) + 2\epsilon \mu \alpha_1 p_{j-1}(\gamma) \\
+ \left( \frac{1}{2} - 2\epsilon \mu (\alpha_1 + \alpha_2) \right) p_j(-\gamma) + \left( \frac{1}{2} - 2\epsilon \mu (\alpha_1 + \alpha_2) \right) p_j(\gamma),
\]  (3.6.1)
where
\[
\alpha_1 = \frac{\beta_0}{2} + \frac{\beta_0 - 1}{2\gamma}, \quad \alpha_2 = \frac{\beta_0}{2} - \frac{\beta_0 - 1}{2\gamma}
\]
and both are non-negative with \(\beta_0 \geq \frac{1}{2}\) and (3.4.18). Note that \(\alpha_1 + \alpha_2 = \beta_0\). If \(\mu \leq \frac{1}{4\epsilon\beta_0}\), (3.6.1) is a convex combination of vector values over set \(S_i^D(i = j-1, j, j+1)\), therefore \(D_j \in \Sigma\) due to the convexity of \(\Sigma\). Moreover, for non-trivial polynomials, a similar argument to the proof of Lemma 3.2.2 shows that \(D_j \in \Sigma_0\). \(\Box\)

**Proof of Lemma 3.4.4:** Using the same notation for \(p_j(\xi)\) as in the proof of Lemma 3.4.3, we see that in the case of \(p_j(\xi) \in \mathbb{P}^2\), for any \(\gamma \in (-1, 1)\),
\[
p_j(\xi) = \frac{(\xi - 1)(\xi - \gamma)}{2(1 + \gamma)} p_j(-1) + \frac{(\xi - 1)(\xi + 1)}{(\gamma - 1)(\gamma + 1)} p_j(\gamma) + \frac{(\xi + 1)(\xi - \gamma)}{2(1 - \gamma)} p_j(1).
\]
Then the cell average is
\[
\bar{w}_j^\gamma = c_1 p_j(-1) + c_2 p_j(\gamma) + c_3 p_j(1),
\]
where
\[
c_1 = \frac{1 + 3\gamma}{6(1 + \gamma)}, \quad c_2 = \frac{2}{3(1 - \gamma^2)}, \quad c_3 = \frac{1 - 3\gamma}{6(1 - \gamma)}
\]
are all non-negative for \(|\gamma| \leq \frac{1}{3}\).

A direct calculation gives \(D_j\) as
\[
D_j = \alpha_4 p_j(-1) + \alpha_5 p_j(\gamma) + \alpha_6 p_j(1)
\]
\[
+ 2\epsilon\mu\alpha_1(\gamma)p_{j+1}(-1) + 2\epsilon\mu\alpha_2(\gamma)p_{j+1}(\gamma) + 2\epsilon\mu\alpha_3(\gamma)p_{j+1}(1)
\]
\[
+ 2\epsilon\mu\alpha_3(-\gamma)p_{j-1}(-1) + 2\epsilon\mu\alpha_2(-\gamma)p_{j-1}(\gamma) + 2\epsilon\mu\alpha_1(-\gamma)p_{j-1}(1)
\]
with
\[
\alpha_1(\gamma) = \beta_0 + \frac{8\beta_1 - 3 - \gamma}{2(\gamma + 1)}, \quad \alpha_2(\gamma) = \frac{8\beta_1 - 2}{\gamma^2 - 1}, \quad \alpha_3(\gamma) = \frac{8\beta_1 - 1 - \gamma}{2(1 - \gamma)},
\]
\[
\alpha_4 = \frac{1 + 3\gamma}{6(1 + \gamma)} - 2\epsilon\mu(\alpha_3(-\gamma) + \alpha_1(\gamma)),
\]
\[
\alpha_5 = \frac{2}{3(1 - \gamma^2)} - 2\epsilon\mu(\alpha_2(-\gamma) + \alpha_2(\gamma)),
\]
\[
\alpha_6 = \frac{1 - 3\gamma}{6(1 - \gamma)} - 2\epsilon\mu(\alpha_1(-\gamma) + \alpha_3(\gamma)).
\]
Note that (3.6.2) is a convex combination of vector values over set $S^D_j$ if $\alpha_i \geq 0$ for $i = 1, \cdots, 6$.
This is guaranteed by (3.4.19) and $|\gamma| \leq 8\beta_1 - 1$, together with $\mu \leq \mu_0$ given in (3.4.21).

Appendix B. Is $C_4$ uniformly bounded?

In this appendix, we give two examples on the magnitude of $C_4$, which is defined in (3.3.8) in the proof of Lemma 3.3.3.

Example 1 ( $C_4$ is uniformly bounded)
Consider $v(x) = 1 + h^2 x^2$, $u(x) = 1$ for $x \in [0, 1]$, where $0 < h \ll 1$. For $\gamma = 3$, we have
\[
    r(v(x), u(x)) = 1 + \sqrt{3} \left( -1 + \frac{1}{1 + h^2 x^2} \right),
    \quad s(v(x), u(x)) = 1 - \sqrt{3} \left( -1 + \frac{1}{1 + h^2 x^2} \right),
\]
and $r_0 = s_0 = 1$.

Now, consider linear interpolation of $v(x)$ and $u(x)$ over $[0, 1]$ as follows
\[
    \tilde{v}(x) = 1 - \frac{1}{8} h^2 + \frac{3}{4} h^2 x,
    \quad \tilde{u}(x) = 1,
\]
where the interpolation points are chosen as $\frac{1}{4}$ and $\frac{1}{2}$. The corresponding cell averages of two polynomials are
\[
    \bar{\tilde{v}} = 1 + \frac{1}{4} h^2,
    \quad \bar{\tilde{u}} = 1.
\]

Also, we can find
\[
    r(\tilde{v}(x), \tilde{u}(x)) = 1 + \sqrt{3} \left( -1 + \frac{1}{1 - \frac{h^2}{8} + \frac{3h^2}{4} x} \right),
    \quad s(\tilde{v}(x), \tilde{u}(x)) = 1 - \sqrt{3} \left( -1 + \frac{1}{1 - \frac{h^2}{8} + \frac{3h^2}{4} x} \right).
\]
We can verify that $r(\tilde{v}(x), \tilde{u}(x)) > 1$, $s(\tilde{v}(x), \tilde{u}(x)) < 1$ for $x < \frac{1}{6}$; namely, $q(x) = (\tilde{v}(x), \tilde{u}(x))$ lies outside of invariant region.
Let \( \bar{r} = r(\bar{v}, \bar{u}) \), \( \bar{s} = s(\bar{v}, \bar{u}) \), then we have

\[
\begin{align*}
r_0 - \bar{r} &= -\sqrt{3} \left( -1 + \frac{1}{1 + \frac{h^2}{4}} \right), \\
\bar{s} - s_0 &= -\sqrt{3} \left( -1 + \frac{1}{1 + \frac{h^2}{4}} \right),
\end{align*}
\]

which indicates \( C_4 = 1 \).

**Example 2.** \((C_4 = O(h^{-2}))\)

Consider \( v(x) = \frac{2\sqrt{3}}{2\sqrt{3} + (1 + h^2)(x-1)} \), \( u(x) = (1 - h^2)x + \frac{h^2-1}{2} \) for \( x \in [0,1] \), where \( 0 < h \ll 1 \). For \( \gamma = 3 \), we have

\[
\begin{align*}
r(v(x), u(x)) &= -\frac{3 - h^2}{2} + 2x, \\
s(v(x), u(x)) &= \frac{1 + 3h^2}{2} - 2h^2 x,
\end{align*}
\]

and \( r_0 = s_0 = \frac{1-h^2}{2} \).

Now, consider linear interpolation of \( v(x) \) and \( u(x) \) over \([0,1]\) as follows

\[
\begin{align*}
\tilde{v}(x) &= 2\sqrt{3} \left( \frac{1}{1 - 2\sqrt{3} + h^2} - \frac{4}{3 - 4\sqrt{3} + 3h^2} \right) + \frac{-8\sqrt{3}(1 + h^2)x}{(-3 + 4\sqrt{3} - 3h^2)(-1 + 2\sqrt{3} - h^2)} \\
\tilde{u}(x) &= \frac{h^2 - 1}{2} + (1 - h^2)x,
\end{align*}
\]

where the interpolation points are chosen as \( \frac{1}{4} \) and \( \frac{1}{2} \). The corresponding averages of two polynomials are

\[
\tilde{v} = \frac{2\sqrt{3}}{-1 + 2\sqrt{3} - h^2}, \quad \tilde{u} = 0.
\]

Let \( \bar{r} = r(\tilde{v}, \tilde{u}), \bar{s} = s(\tilde{v}, \tilde{u}) \), then we have

\[
\begin{align*}
\bar{r} &= -\frac{1 - h^2}{2}, \\
\bar{s} &= \frac{1 + h^2}{2}
\end{align*}
\]

and

\[
\begin{align*}
r_0 - \bar{r} &= 1, \\
\bar{s} - s_0 &= h^2.
\end{align*}
\]

In the following table, we show that with certain choices of \( h \), we have \( \theta_2 < \theta_1 < 1 \), which indicates \( C_4 = \frac{1}{h^2} \), where \( \theta_1 \) and \( \theta_2 \) are as defined in \((3.3.2)\),
Table 3.9  Choice of $h$ and corresponding parameter values for unbounded $C_4$. 

<table>
<thead>
<tr>
<th>$h$</th>
<th>$r_{\text{max}}$</th>
<th>$r_0$</th>
<th>$\bar{r}$</th>
<th>$\theta_1$</th>
<th>$s_{\text{min}}$</th>
<th>$s_0$</th>
<th>$\bar{s}$</th>
<th>$\theta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>3.831</td>
<td>0.375</td>
<td>-0.625</td>
<td>0.224</td>
<td>-3.081</td>
<td>0.375</td>
<td>0.625</td>
<td>0.068</td>
</tr>
<tr>
<td>0.1</td>
<td>1.310</td>
<td>0.495</td>
<td>-0.505</td>
<td>0.551</td>
<td>-0.320</td>
<td>0.495</td>
<td>0.505</td>
<td>0.012</td>
</tr>
<tr>
<td>0.01</td>
<td>1.27823</td>
<td>0.49995</td>
<td>-0.50005</td>
<td>0.562234</td>
<td>-0.27833</td>
<td>0.49995</td>
<td>0.50005</td>
<td>0.00013</td>
</tr>
</tbody>
</table>

References


CHAPTER 4. IRP DG METHODS FOR MULTI-DIMENSIONAL HYPERBOLIC CONSERVATION LAW SYSTEMS, WITH AN APPLICATION TO COMpressible EULER EQUATIONS

A paper accepted for publication in Journal of Computational Physics.

Yi Jiang and Hailiang Liu

Abstract

An invariant-region-preserving (IRP) limiter for multi-dimensional hyperbolic conservation law systems is introduced, as long as the system admits a global invariant region which is a convex set in the phase space. It is shown that the order of approximation accuracy is not destroyed by the IRP limiter, provided the cell average is away from the boundary of the convex set. Moreover, this limiter is explicit, and easy for computer implementation. A generic algorithm incorporating the IRP limiter is presented for high order finite volume type schemes. For arbitrarily high order discontinuous Galerkin (DG) schemes to hyperbolic conservation law systems, sufficient conditions are obtained for cell averages to remain in the invariant region provided the projected one-dimensional system shares the same invariant region as the full multi-dimensional hyperbolic system does. The general results are then applied to both one and two dimensional compressible Euler equations so to obtain high order IRP DG schemes. Numerical experiments are provided to validate the proven properties of the IRP limiter and the performance of IRP DG schemes for compressible Euler equations.
4.1 Introduction

The multi-dimensional hyperbolic conservation law systems are given by

\[ \partial_t w + \sum_{j=1}^{d} \partial_{x_j} F_j(w) = 0, \quad x \in \mathbb{R}^d, \quad t > 0 \]  

(4.1.1)

with the unknown vector \( w \in \mathbb{R}^l \) and the flux function \( F_j(w) \in \mathbb{R}^l \) for \( j = 1, \ldots, d \). We consider the initial value problem for system (4.1.1) with the initial data \( w(x, 0) = w_0(x) \). For simplicity, we take periodic or compactly supported boundary conditions.

It is well known that entropy inequalities should be considered for general hyperbolic conservation laws so to single out the physically relevant solution among many weak solutions (see, e.g., [19]). In application problems, the pointwise range of solutions may be known from physical considerations, instead of total entropy. For scalar conservation laws, the entropy solution satisfies a strict maximum principle. For hyperbolic conservation law systems, the notion of maximum principle does not apply and must be replaced by the notion of invariant region. To solve a conservation law system with possibly discontinuous solutions, one naturally studies the invariant-region-preserving (IRP) property of the numerical schemes.

In this paper, we are interested in constructing IRP high order accurate schemes for solving (4.1.1) with an application to compressible Euler equations. The invariant region \( \Sigma \) to this system is meant to be a convex set in phase space \( \mathbb{R}^l \) so that if the initial data is in the region \( \Sigma \), then the solution will remain in \( \Sigma \) for all \( t > 0 \). It is highly desirable to construct high order numerical schemes solving (4.1.1) that can preserve the entire invariant region \( \Sigma \), which is in general a difficult problem. In this article, we will discuss the IRP property of arbitrarily high order schemes on shape-regular meshes, following the discontinuous Galerkin (DG) framework developed by Cockburn and Shu [4, 5, 6].

There are models that feature known invariant regions. For example, the invariant region of one-dimensional \( 2 \times 2 \) (\( l = 2 \)) systems of hyperbolic conservation laws can be described by two Riemann invariants, see e.g., [3, 14, 27, 9]. For general conservation law systems with \( l \geq 3 \), it is an open question to identify a global invariant region. When considering the compressible Euler
equations of gas dynamics, a natural condition for the solution is positivity of density and pressure, and the minimum principle for the specific entropy [28]. In the one dimensional case, the Euler equation has the following form

\[ \partial_t w + \partial_x f(w) = 0, \quad t > 0, \ x \in \mathbb{R}, \quad (4.1.2) \]

with \( w = (\rho, m, E)^\top \),

\[ f(w) = (m, \rho u^2 + p, (E + p)u)^\top, \quad E = \frac{1}{2} \rho u^2 + \frac{p}{\gamma - 1}, \quad (4.1.3) \]

where \( \rho \) is the density, \( u \) is the velocity, \( m = \rho u \) is the momentum, \( E \) is the total energy, \( p \) is the pressure, and \( \gamma \) is the ratio of specific heats for the gas/fluid (for most gases, \( 1 < \gamma < 3 \)). The corresponding invariant region is the following set

\[ G = \{ w \mid \rho > 0, \ p > 0, \ s \geq s_0 \}, \]

where \( s = \log(p/\rho^\gamma) \) is the specific entropy, and \( s_0 := \inf_x s(w_0(x)) \).

For nonlinear systems of conservation laws in several space variables with a known invariant region, the IRP property under the Lax-Friedrich schemes was studied by Frid in [8, 9]. For the compressible Euler equations, the first order finite volume schemes including Godunov and Lax-Friedrichs schemes are shown to preserve the minimum entropy property [28]. Further second-order limitation techniques for enforcing the specific entropy bound were proposed in [17], where it was reported that enforcing the minimum entropy principle numerically might damp oscillations in numerical solutions. To have the specific entropy well-defined, one would have to guarantee the positivity of density and pressure of the numerical solution, which can be done for a high order finite volume or a DG scheme following [23, 24, 32, 34]. The main idea of positivity-preserving techniques for high order DG schemes in [32, 34] is to find a sufficient condition to preserve the positivity of the cell averages by repeated convex combinations, combined with a conservative limiter which can enforce the sufficient condition without destroying accuracy for smooth solutions, as shown in [31] for scalar conservation laws. In the context of continuous finite elements, the IRP property has been studied by Guermond and Popov [11] using the first order approximation to solve general
hyperbolic conservation law systems, and then in [10] using the second order approximation with convex limiting to solve compressible Euler equations.

A more closely related development is the work by Zhang and Shu [33], where the authors extended the positivity-preserving high order schemes for compressible Euler equations to preserve the entire $G$, while their limiter for enforcing the lower bound of $s$ is implicit with the limiter parameter solved by the Newton iteration. In [15], we introduced an explicit limiter based on a simple observation that the convex set $G$ can be rewritten as

$$
\Sigma = \{ \mathbf{w} \mid \rho > 0, \quad p > 0, \quad q \leq 0 \},
$$

where $q = (s_0 - s)\rho$. Note that $s$ is quasi-concave, but $q$ is convex; actually the fact that $-\rho s$ has a positive definite Hessian matrix can be derived from a general result by Harten in [12]. Such a reformulation allows us to construct a new IRP limiter in [15] for compressible Euler equations. The limiter modifies the polynomial solution still through a linear convex combination as in [33], yet the limiter parameter is defined explicitly due to the convexity of $q$, and concavity of $p$. The question of particular interest is whether it is still high order accurate. In [15], the IRP limiter was proved to maintain same high order accuracy if the cell average is away from the boundary of the convex set. While the bound preserving limiter [33] for the entropy function was shown to be high order accurate provided the second order derivative of the entropy function for numerical solutions does not vanish.

The work [15] was built upon [16], where we introduced an explicit IRP limiter for DG methods solving the isentropic gas dynamic system (with or without viscosity). Again both convexity and concavity of two Riemann invariants play an essential role in the construction of the explicit IRP limiter there. We observe that the ideas for both the explicit IRP limiters and the high order IRP schemes studied in [15, 16] can be readily extended to general hyperbolic conservation law systems (4.1.1) as long as (i) it features a global invariant region

$$
\Sigma = \{ \mathbf{w} \mid U(\mathbf{w}) \leq 0 \},
$$
where $U$ is convex, and (ii) the corresponding one-dimensional projected system

$$\partial_t w + \partial_\eta \left( \sum_{j=1}^d F_j(w) \nu_j \right) = 0, \quad \eta \in \mathbb{R},$$

where $\nu = (\nu_1, \cdots, \nu_d)$ is any unit vector, shares the same invariant region $\Sigma$. The later assumption is needed in order to obtain an IRP scheme. These observations led to the present work on high order IRP schemes for general conservation law systems (4.1.1). This work may also be seen as to some degree an extension of the earlier works on positivity-preserving schemes for compressible Euler equations. The present emphasis is on the notion of invariant regions.

In this paper we have the following objectives:

(i) to design an explicit IRP limiter, which can be shown to maintain high order accuracy of the approximation;

(ii) to identify sufficient conditions under which arbitrarily high order DG schemes feature the desired IRP property for both one and multi-dimensional hyperbolic conservation law systems;

(iii) to apply the general result in (ii) to two-dimensional compressible Euler equations, with numerical validations.

As for (i), our limiter for preserving $\Sigma$ in (4.1.5) is of the following form

$$\tilde{w}_h(x) = \theta w_h(x) + (1 - \theta) \bar{w}_h, \quad \theta = \min \left\{ 1, \frac{U(\bar{w}_h)}{U(w_h) - U_{\text{max}}^h} \right\},$$

where $w_h(x)$ is a polynomial over domain $K$, its average $\bar{w}_h$ lies in the interior of $\Sigma$, and $U_{\text{max}}^h = \max_{x \in K} U(w_h(x))$. This reconstructed polynomial is shown to maintain the same order of accuracy as $w_h(x)$, provided $U(w_h(x))$ is not close to zero.

As for (ii), we present our analysis for general conservation law systems in order to illustrate the ways in which the IRP property can be ensured for high order finite volume type schemes. The first ingredient is a one-dimensional IRP numerical flux, such as Godunov, Lax-Friedrichs, and Harten-Lax-van Leer [13], with which the first order finite volume scheme has the IRP property under certain CFL condition. This allows us to further find a sufficient condition to keep the cell averages in $\Sigma$ by repeated convex combinations, in the same way as that has been well established for positivity-preserving schemes, see e.g. [32, 33]. In the present setting, we use first order schemes...
with an IRP flux which can keep numerical solutions in $\Sigma$ as building blocks, and show that high order spatial discretization with forward Euler can be written as a convex combination of first order IRP schemes, thus will keep $\Sigma$ provided certain sufficient conditions are satisfied. The IRP limiter is then used to enforce the sufficient condition.

For multi-dimensional conservation law systems we will show that repeated convex combinations can still be achieved, as long as a positive convex decomposition for the cell averages is available, using the Gauss-Lobatto quadrature points on cell interfaces, and some interior points chosen so that the decomposition weights are strictly positive. The numerical solutions need to be within in $\Sigma$ only on a test set consisting of points also used for the cell average decomposition.

As for (iii), we first show that the projected system of the two dimensional Euler systems indeed share the same invariant region $\Sigma$ in (4.1.4). The CFL conditions for the IRP DG schemes on rectangular and triangular meshes are derived, respectively, from our general result for multi-dimensional hyperbolic conservation law systems, while using the test sets identified already in [32] and [34].

Finally, we should mention that in our analysis we only show the ways of numerically preserving $\Sigma$ for the forward Euler time discretization, yet the high order SSP time discretizations ([26]) will keep the validity of our results since they are convex combinations of forward Euler.

This paper is organized as follows: in Sect. 2 we present an explicit invariant-region-preserving (IRP) limiter and prove that for smooth solutions the order of approximation accuracy is not destroyed by the IRP limiter in general cases, followed by a generic IRP algorithm for high order schemes. Then, in Sect. 3, we first show all three popular numerical fluxes can be made an IRP flux, and then identify sufficient conditions including both a test set and the CFL condition, to obtain IRP DG schemes in one and higher space dimensions for arbitrary shape-regular meshes. Sect. 4 is devoted to an application to high order DG schemes for two dimensional compressible Euler equations. In Sect. 5, we present extensive numerical tests. Some concluding remarks are given in Sect. 6. In Appendix A, we prove Lemma 2.1 for the compressible Euler equations for
which the pressure is not strictly concave. In Appendix B, we present the detailed proof of Lemma 3.2 which states that HLLC flux is an IRP flux.

### 4.2 The IRP limiter

For the general multi-dimensional system of conservation laws (4.1.1), we assume it admits an invariant region $\Sigma$, which is a convex set in the phase space $\mathbb{R}^l$, characterized by

$$\Sigma = \{ w \mid U(w) \leq 0 \}$$  \hspace{1cm} (4.2.1)

with $U$ being convex. In what follows, we use

$$\Sigma_0 = \{ w \mid U(w) < 0 \}$$

to denote the interior of $\Sigma$.

For any bounded domain $K$, we define the average of $w(x)$ by

$$\bar{w} = \frac{1}{|K|} \int_K w(x)dx,$$

where $|K|$ is the measure of $K$. The following lemma shows that such an averaging operator is a contraction, which enables us to use the cell average as a reference to construct the IRP limiter.

**Lemma 4.2.1.** Let $w(x)$ be non-trivial piecewise continuous vector functions. If $w(x) \in \Sigma$ for all $x \in K \subset \mathbb{R}^d$, and $U$ is strictly convex, then $\bar{w} \in \Sigma_0$ for any bounded domain $K$.

**Proof.** Since $U$ is convex, using Jensen’s inequality and the assumption, we have

$$U(\bar{w}) = U\left( \frac{1}{|K|} \int_K w(x)dx \right) \leq \frac{1}{|K|} \int_K U(w(x))dx \leq 0.$$

With this, we can show $U(\bar{w}) < 0$. Otherwise if $U(\bar{w}) = 0$, we must have $U(w(x)) = 0$ for almost all $x \in K$; that is

$$U(\bar{w}) = U(w(x)) \quad a.e. \text{ in } K.$$
This, upon taking cell average on both sides, gives

\[
U(\bar{w}) = \frac{1}{|K|} \int_K U(w(x))dx.
\]

By taking the Taylor expansion around \(\bar{w}\), we have

\[
U(w(x)) = U(\bar{w}) + \nabla_w U(\bar{w}) \cdot \xi + \xi^T H \xi, \quad \forall x \in K, \quad \xi := w(x) - \bar{w},
\]

which upon integration yields \(\frac{1}{|K|} \int_K \xi^T H \xi dx = 0\), where \(H\) is the Hessian matrix of \(U\). This when combined with the strict convexity of \(U\) ensures that \(w(x) \equiv \bar{w}\) almost everywhere, which contradicts the assumption. \(\square\)

### 4.2.1 The limiter

Let \(w_h(x)\) be a sequence of vector polynomials over \(K\), which is a high order accurate approximation to the smooth function \(w(x) \in \Sigma\). We assume \(\bar{w}_h \in \Sigma_0\), but \(w_h(x)\) is not entirely located in \(\Sigma\), then we can modify the polynomial \(w_h(x)\) with reference to \(\bar{w}_h\) through a linear convex combination:

\[
\tilde{w}_h(x) = \theta w_h(x) + (1 - \theta) \bar{w}_h, \quad (4.2.2)
\]

where \(\theta \in (0, 1]\) is defined by

\[
\theta = \min\{1, \theta_1\}, \quad (4.2.3)
\]

where

\[
\theta_1 = \frac{U(\bar{w}_h)}{U(\bar{w}_h) - U_{\text{max}}^h}, \quad (4.2.4)
\]

with

\[
U_{\text{max}}^h = \max_{x \in K} U(w_h(x)) > 0. \quad (4.2.5)
\]

Notice that since \(\bar{w}_h \in \Sigma_0\), we have \(U(\bar{w}_h) < 0\). Also \(U(\bar{w}_h) < U_{\text{max}}^h\). Therefore, \(\theta_1\) is well-defined and positive.

As for the above limiter, we have the following conclusion.
Theorem 4.2.2. If \( \mathbf{w}_h \in \Sigma_0 \), then \( \mathbf{w}_h(x) \in \Sigma \), for all \( x \in K \). Moreover, the reconstructed polynomial preserves high order accuracy, i.e., if \( \| \mathbf{w}_h - \mathbf{w} \|_\infty \leq 1 \), then

\[
\| \mathbf{\tilde{w}}_h - \mathbf{w} \|_\infty \leq \frac{C}{|U(\mathbf{\tilde{w}}_h)|} \| \mathbf{w}_h - \mathbf{w} \|_\infty,
\]

where \( C > 0 \) depends on \( \mathbf{w} \) and \( \Sigma \).

Proof. The claim that the constructed polynomial lies within \( \Sigma \) is implied by the definition of \( \theta \). In fact, for the case \( \theta = \theta_1 \) with the convexity of \( U \), we have

\[
U(\mathbf{\tilde{w}}_h(x)) \leq \theta U(\mathbf{w}_h(x)) + (1 - \theta)U(\mathbf{\tilde{w}}_h) \\
\leq \theta_1 U_{\text{max}}^h + (1 - \theta_1)U(\mathbf{\tilde{w}}_h) = 0.
\]

For the accuracy estimate, we consider the case when \( \theta \neq 1 \). We only need to prove

\[
\| \mathbf{\tilde{w}}_h - \mathbf{w}_h \|_\infty \leq \frac{C}{|U(\mathbf{\tilde{w}}_h)|} \| \mathbf{w}_h - \mathbf{w} \|_\infty,
\] (4.2.6)

from which the conclusion follows by using the triangle inequality. From the reconstruction, it follows that

\[
\| \mathbf{\tilde{w}}_h - \mathbf{w}_h \|_\infty = (1 - \theta)\| \mathbf{\tilde{w}}_h - \mathbf{w}_h \|_\infty \\
= \frac{\| \mathbf{\tilde{w}}_h - \mathbf{w}_h \|_\infty}{U_{\text{max}}^h} U^h_{\text{max}}.
\]

Since \( U(\mathbf{w}) \leq 0 \) for \( \mathbf{w}(x) \in \Sigma \), we have

\[
U_{\text{max}}^h \leq \max_{x \in K} (U(\mathbf{w}_h)) - U(\mathbf{w}) \leq \| \nabla U \|_\infty \| \mathbf{w} - \mathbf{w}_h \|_\infty.
\]

Also, since \( U_{\text{max}}^h > 0 \) when \( \theta < 1 \), we have

\[
U_{\text{max}}^h - U(\mathbf{\tilde{w}}_h) > -U(\mathbf{\tilde{w}}_h) > 0.
\]

According to the assumption that \( \mathbf{w}_h \) is an approximation to \( \mathbf{w} \), we have

\[
\| \mathbf{\tilde{w}}_h - \mathbf{w}_h \|_\infty = \| \mathbf{\tilde{w}}_h - \mathbf{\tilde{w}} + \mathbf{\tilde{w}} - \mathbf{w} + \mathbf{w} - \mathbf{w}_h \|_\infty \leq 2\| \mathbf{w} - \mathbf{w}_h \|_\infty + \| \mathbf{w} - \mathbf{\tilde{w}} \|_\infty
\]

Therefore, we arrive at (4.2.6) with \( C \) given by
\[ C = 2\|\nabla U\|_\infty (1 + \|w\|_\infty), \]

which is positive and finite, depending only on \(w\), as well as the invariant region \(\Sigma\) through \(U\).

\textit{Remark 4.2.1.} We would like to point out that when \(\bar{w}_h\) is close enough to the boundary of \(\Sigma\), the factor \(C/|U(\bar{w}_h)|\) can become large, indicating the possibility of accuracy deterioration in some cases.

In practice, \(\Sigma\) is usually given by several pieces of convex functions in the form of

\[ \Sigma = \bigcap_{i=1}^{M} \{w| U_i(w) \leq 0\}. \]

Then the limiter parameter given in (4.2.4) needs to be modified as

\[ \theta = \min\{1, \theta_1, \ldots, \theta_M\}, \quad (4.2.7) \]

where

\[ \theta_i = \frac{U_i(\bar{w}_h)}{U_i(\bar{w}_h) - U^\text{max}_{i,h}} \quad U^\text{max}_{i,h} = \max_{x \in K} U_i(w_h(x)). \quad (4.2.8) \]

It can be shown that Theorem 4.2.2 remains valid for the general case as such.

Here we present two such examples for a convex invariant region.

\textbf{Example 1.} The first example is the one dimensional isentropic gas dynamic system in Euler coordinates, i.e. the system (4.1.2) with \(w = (\rho, m)\) and \(f(w) = (\rho u, \rho u^2 + p(\rho))\), where \(p(\rho) = \rho^\gamma, \gamma > 1\) and \(m = \rho u\). The corresponding invariant region is given by

\[ \Sigma = \{(\rho, m)^\top | r \leq r_0, s \geq s_0\}, \]

where \(r_0 = \sup_x r(\rho_0(x), m_0(x))\), \(s_0 = \inf_x s(\rho_0(x), m_0(x))\) and

\[ r = u + \frac{2\sqrt{\gamma}}{\gamma - 1} \rho \frac{\gamma - 1}{2}, \quad s = u - \frac{2\sqrt{\gamma}}{\gamma - 1} \rho \frac{\gamma - 1}{2}, \]
are two Riemann invariants. We point out that $\Sigma$ is a closed domain in $\{(\rho, m)^{\top} \mid \rho \geq 0\}$. 

Note for the pressure-less Euler equation, i.e. the system (4.1.2) with $w = (\rho, m)^{\top}$ and $f(w) = (\rho u, \rho u^2)^{\top}$, where $m = \rho u$, two Riemann invariants $r$ and $s$ become identical. In such case, the invariant region is

$$
\Sigma = \{ (\rho, m)^{\top} \mid \rho > 0, \ s_0 \rho \leq m \leq r_0 \rho \}, \quad (4.2.9)
$$

where $r_0 = \sup_x (u_0(x))$ and $s_0 = \inf_x (u_0(x))$.

**Example 2.** Another example is the compressible Euler equations, for which an invariant region is

$$
\Sigma = \{ (\rho, m, E)^{\top} \mid \rho > 0, \ p > 0, \ q \leq 0 \},
$$

with $p$ and $q$ defined by

$$
p = (\gamma - 1)(E - \frac{1}{2} \rho |u|^2), \quad q = (s_0 - s) \rho, \ \gamma > 1,
$$

where $u$ is the velocity, $m = \rho u$, $s = \log \left( \frac{p(x)}{\rho^\gamma(x)} \right)$, $s_0 = \inf_x \log \left( \frac{p_0(x)}{\rho_0^\gamma(x)} \right)$, and $(\rho_0, m_0, E_0)^{\top}$ is the initial data.

**Remark 4.2.2.** When $U$ is not strictly convex, the result in Lemma 4.2.1 may still hold true. The proof needs to be modified based on further details of $U$. For example, for the compressible Euler equation, $p$ is concave but not strictly concave. We present an illustrative proof in Appendix A.

**4.2.2 Algorithm**

Let $w_h^n$ be the numerical solution at $n$-th time step generated from a high order scheme of an abstract form

$$
w_h^{n+1} = \mathcal{L}(w_h^n), \quad (4.2.10)
$$
starting with initial data \( w_h^0 \), where \( w_h^n = w_h^n(x) \in V_h \), and \( V_h \) is a finite element space of piecewise polynomials of degree \( k \) in each computational cell \( K \), i.e.,

\[
V_h = \{ v : v|_K \in \mathbb{P}^k(K) \}.
\]

Assume \( \lambda = \frac{\Delta t}{|K|} \) is the mesh ratio. The IRP algorithm can be stated as follows:

**Algorithm 4.2.1.** Provided that scheme (4.2.10) has the following property: there exists \( \lambda_0 \), and a test set \( S \) such that if

\[
\lambda \leq \lambda_0 \quad \text{and} \quad w_h^n(x) \in \Sigma \quad \text{for} \quad x \in S
\]

then

\[
w_h^{n+1} \in \Sigma_0;
\]

then the IRP limiter can be applied, with \( K \) replaced by \( S_K := S \cap K \) in (4.2.8), i.e.,

\[
U_h^{\max} = \max_{x \in S_K} U(w_h(x)), \quad (4.2.11)
\]

through the following algorithm:

**Step 1:** Initialization: take the piecewise \( L^2 \) projection of \( w_0 \) onto \( V_h \), such that

\[
\langle w_h^0 - w_0, \phi \rangle = 0, \quad \forall \phi \in V_h.
\]

**Step 2:** Imposing the modified limiter (4.2.2), (4.2.3) with (4.2.11) on \( w_h^n \) for \( n = 0, 1, \ldots \) to obtain \( \tilde{w}_h^n \).

**Step 3:** Update by the scheme:

\[
w_h^{n+1} = L(\tilde{w}_h^n).
\]

Return to **Step 2**.
Remark 4.2.3. For given initial data $w_0 \in \Sigma$, its cell average lies strictly within $\Sigma_0$. On the other hand, $w^0_h$ may not lie entirely in $\Sigma$, but it has the same cell average as the initial data due to the $L^2$ projection. Therefore, the IRP limiter can already be applied to $w^0_h$ (included in Step 2 in the algorithm).

4.3 IRP DG schemes

In this section, we discuss some sufficient conditions for high order DG schemes solving the general conservation laws to be invariant-region-preserving.

4.3.1 One-dimensional case

We begin with the one-dimensional system of conservation laws of the form

$$\partial_t w + \partial_x f(w) = 0,$$

(4.3.1)

where $f$ is a smooth vector flux function. A first order finite volume scheme on a cell $I_j = [x_{j-1/2}, x_{j+1/2}]$ takes the following form

$$w^{n+1}_j = w^n_j - \lambda \left( \hat{f}_{j+1/2} - \hat{f}_{j-1/2} \right),$$

(4.3.2)

where $w^n_j$ is the approximation to the average of $w(x)$ on $I_j = [x_{j-1/2}, x_{j+1/2}]$ at $n$-th time level $t^n$. $\hat{f}_{j+1/2}$ is a single-valued numerical flux at the element interface, depending on the values of numerical solution from both sides

$$\hat{f}_{j+1/2} = \hat{f}(w^n_j, w^n_{j+1}).$$

In general, $\hat{f}_{j+1/2}$ is derived from some Riemann solvers (exact or approximate).

Definition 4.3.1. A consistent numerical flux $\hat{f}_{j+1/2}$ is called an IRP flux for (4.3.1) if there exists $c_0$, such that for $\sigma \lambda \leq c_0$, $w^n_j, w^n_{j\pm 1} \in \Sigma$ implies $w^{n+1}_j \in \Sigma_0$, where $\sigma$ is the global maximum wave speed of the system (4.3.1).
For scalar conservation laws, the invariant region is simply an interval ensured by the maximum principle. It is known that the monotone flux is maximum-principle-preserving, see e.g. [31], therefore it is also the IRP flux. For systems, most popular numerical fluxes rely on Riemann solvers, which exactly compute or approximate the solution of the Riemann problem, i.e., (4.3.1) with initial data,

\[ w(x, 0) = \begin{cases} 
    w_l, & x < 0 \\
    w_r, & x > 0 .
\end{cases} \] (4.3.3)

The solution of the Riemann problem is self-similar. Assume that the Riemann solver also has some self-similar structure and is denoted by \( R(\xi; w_l, w_r) \) with \( \xi = \frac{\xi}{t} \). Let \( \sigma_l \) and \( \sigma_r \) be the leftmost and rightmost wave speed such that \( R = w_l \) for \( \xi \leq \sigma_l \) and \( R = w_r \) for \( \xi \geq \sigma_r \). Let \( S_l = \min\{\sigma_l, 0\} \) and \( S_r = \max\{\sigma_r, 0\} \). Integration of (4.3.1) over \([S_l t, S_r t] \times [0, t]\), divided by \((S_r - S_l)t\), leads to the following identity

\[
\frac{1}{S_r - S_l} \int_{S_l}^{S_r} R(\xi; w_l, w_r) d\xi = \frac{S_r w_r - S_l w_l}{S_r - S_l} - \frac{f_r - f_l}{S_r - S_l},
\] (4.3.4)

where \( f_r = f(w_r) \) and \( f_l = f(w_l) \). This identity is useful in finding sufficient conditions for each of the following numerical fluxes to be an IRP flux.

1. Godunov flux:

\[
\hat{f}(w_l, w_r) = f(R(0; w_l, w_r));
\] (4.3.5)

2. Lax-Friedrich flux:

\[
\hat{f}(w_l, w_r) = \frac{1}{2} (f(w_l) + f(w_r) - \sigma(w_r - w_l));
\] (4.3.6)

3. HLL flux [13]:

\[
\hat{f}(w_l, w_r) = \begin{cases} 
    f(w_l), & \text{if } 0 \leq \sigma_l, \\
    \frac{\sigma_r f(w_l) - \sigma_l f(w_r) + \sigma_l \sigma_r (w_r - w_l)}{\sigma_r - \sigma_l}, & \text{if } \sigma_l \leq 0 \leq \sigma_r, \\
    f(w_r), & \text{if } 0 \geq \sigma_r.
\end{cases}
\] (4.3.7)
Lemma 4.3.1.

(i) For $c_0 = 1$, both Godunov flux and Lax-Friedrich flux are IRP fluxes for (4.3.1);

(ii) For $c_0 = \frac{1}{2}$, the HLL flux is an IRP flux for (4.3.1).

Proof. (i) With the Godunov flux in (4.3.2), $w_{j}^{n+1}$ can be viewed as the cell average of the exact Riemann solution at $t^{n+1}$ when $\lambda \sigma \leq 1$. See [21, Section 13.2]. Since the exact solution lies in $\Sigma$, then according to Lemma 4.2.1, we have $w_{j}^{n+1} \in \Sigma_0$ if $\lambda \sigma \leq 1$.

When the Lax-Friedrich flux (4.3.6) is used, the update $w_{j}^{n+1}$ in (4.3.2) can be rewritten as

$$w_{j}^{n+1} = (1 - \lambda \sigma) w_{j}^{n} + \lambda \sigma w^*, \tag{4.3.8}$$

where

$$w^* = \left( \frac{w_{j-1}^{n} + w_{j+1}^{n}}{2} - \frac{f(w_{j+1}^{n}) - f(w_{j-1}^{n})}{2 \sigma} \right).$$

From (4.3.4) it follows that

$$w^* = \frac{1}{2 \sigma} \int_{-\sigma}^{\sigma} R(\xi; w_{j-1}^{n}, w_{j+1}^{n}) d\xi.$$

For $w_{j+1}^{n} \in \Sigma$, we have $R(\xi; w_{j-1}^{n}, w_{j+1}^{n}) \in \Sigma$. Therefore $w^*$ lies in $\Sigma_0$ by Lemma 4.2.1. Since $w_{j}^{n+1}$ is a convex combination of two vectors: $w_{j}^{n} \in \Sigma$ and $w^* \in \Sigma_0$ for $\lambda \sigma \leq 1$, we then have $w_{j}^{n+1} \in \Sigma_0$.

(ii) For the HLL flux (4.3.7), the evolved cell average $w_{j}^{n+1}$ can be rewritten as

$$w_{j}^{n+1} = (1 - \theta_1 - \theta_2) w_{j}^{n} + \theta_1 \hat{w}^1 + \theta_2 \hat{w}^2, \tag{4.3.8}$$

with

$$\hat{w}^1 = \frac{b_1 w_{j+1}^{n} - a_1 w_{j}^{n}}{b_1 - a_1} - \frac{f(w_{j+1}^{n}) - f(w_{j}^{n})}{b_1 - a_1},$$

$$\hat{w}^2 = \frac{b_2 w_{j}^{n} - a_2 w_{j-1}^{n}}{b_2 - a_2} - \frac{f(w_{j}^{n}) - f(w_{j-1}^{n})}{b_2 - a_2},$$

where $\theta_1 = -\lambda a_1$, $\theta_2 = \lambda b_2$ with

$$b_1 = \max\{\sigma_{j+\frac{1}{2},r}, 0\}, \quad a_1 = \min\{\sigma_{j+\frac{1}{2},l}, 0\},$$

$$b_2 = \max\{\sigma_{j-\frac{1}{2},r}, 0\}, \quad a_2 = \min\{\sigma_{j-\frac{1}{2},l}, 0\},$$
and \( \sigma_{j + \frac{1}{2}, l} \) and \( \sigma_{j + \frac{1}{2}, r} \) are the leftmost and rightmost wave speeds at \( x_{j + \frac{1}{2}} \). Notice that both \( \hat{w}^1 \) and \( \hat{w}^2 \) are in the form of (4.3.4), the cell average of some exact Riemann solutions, hence they both lie in \( \Sigma_0 \) by Lemma 4.2.1. Therefore \( \lambda \sigma \leq \frac{1}{2} \) is a sufficient condition for \( w_j^{n+1} \) in (4.3.8) to be in \( \Sigma_0 \).

\[ \text{Remark 4.3.1.} \text{ Notice that the local Lax-Friedrich flux is a special case of HLL flux, where } \]
\[ \sigma_{j + \frac{1}{2}, r} = -\sigma_{j + \frac{1}{2}, l} = \max_{w^j, w^{j+1}} |\partial_w f(\cdot)|. \]

Hence the local Lax-Friedrich flux is an IRP flux when \( \lambda \sigma \leq \frac{1}{2} \). Here we use \( |\partial_w f| \) as a notation to denote the absolute value of eigenvalues of Jacobian matrix \( \partial_w f \).

The HLLC approximate Riemann solver as a three wave model was proposed by Toro et al. [29] as a modification of the HLL flux whereby the missing contact and shear waves in the Euler equations are restored. The HLLC flux is given by

\[
\hat{f}(w_l, w_r) = \begin{cases} 
    f(w_l), & \text{if } 0 \leq \sigma_l, \\
    f_{sl}, & \text{if } \sigma_l \leq 0 \leq \sigma_s, \\
    f_{sr}, & \text{if } \sigma_s \leq 0 \leq \sigma_r, \\
    f(w_r), & \text{if } 0 \geq \sigma_r,
\end{cases}
\]

where \( \sigma_s \) is the speed of middle wave, and the intermediate fluxes are given by

\[
f_{sl} = f(w_l) + \sigma_l (w_{sl} - w_l), \quad f_{sr} = f(w_r) + \sigma_r (w_{sr} - w_r),
\]

and \( w_{sl}, w_{sr} \) are two intermediate states determined by integral averages of the Riemann solution

\[
w_{sl} = \frac{1}{\sigma_s - \sigma_l} \int_{\sigma_l}^{\sigma_s} w(\xi t, t) d\xi, \quad w_{sr} = \frac{1}{\sigma_r - \sigma_s} \int_{\sigma_s}^{\sigma_r} w(\xi t, t) d\xi.
\]

The two intermediate fluxes are related by

\[
f_{sr} = f_{sl} + \sigma_s (w_{sr} - w_{sl}). \quad (4.3.9)
\]

Note that there are more unknowns than equations and some extra conditions need to be imposed in order to determine the intermediate fluxes, see [29] for two versions of the HLLC flux for the
compressible Euler equation. For general 1D hyperbolic conservation law systems, the following result for the HLLC flux to be an IRP flux is proved in Appendix B.

Lemma 4.3.2. For \( c_0 = \frac{1}{2} \), the HLLC flux is an IRP flux.

Remark 4.3.2. We want to point out that Algorithm 4.2.1 can still be applied to weakly hyperbolic conservation laws. A canonical example is the pressure-less Euler system, due to the formation of vacuum and/or delta-shock formation in the density, care is needed in the choice of the numerical flux. The Godunov flux derived in [1] was used in the DG scheme given in [30]. From [30, Lemma 4.2] we see that the Godunov flux is indeed an IRP flux for \( c_0 = \frac{1}{2} \) with \( \sigma = \max\{|s_0|, |r_0|\} \) to preserve \( \Sigma \) as defined in (4.2.9). Our explicit IRP limiter can of course be used as an alternative to the special limiter constructed in [30] in order to fulfill the two requirements: \( \rho \) is positive and the velocity \( u = m/\rho \) satisfies a maximum principle.

For a \((k + 1)\)th-order scheme with reconstructed polynomials or approximation polynomials of degree \( k \), with forward Euler time discretization, the cell average evolves by

\[
\tilde{w}_j^{n+1} = \tilde{w}_j^n - \lambda [\tilde{f}(w_{j+\frac{1}{2}}^-; w_{j+\frac{1}{2}}^+) - \tilde{f}(w_{j-\frac{1}{2}}^-; w_{j-\frac{1}{2}}^+)],
\]

(4.3.10)

where \( \tilde{w}_j^n \) is the cell average of \( w_h^j \) on \( I_j \) at time level \( n \), \( w_{j+\frac{1}{2}}^\pm \) are approximations to the point value of \( w \) at \( x_{j+1/2} \) at time level \( n \) from the left and the right cells, respectively.

We consider an \( N \)-point Legendre Gauss-Lobatto quadrature rule on \( I_j \), with quadrature weights \( \hat{\omega}_i \) on \([\frac{-1}{2}, \frac{1}{2}]\) such that \( \sum_{i=1}^N \hat{\omega}_i = 1 \), which is exact for integrals of polynomials of degree up to \( k \), if \( 2N - 3 \geq k \). Denote these quadrature points on \( I_j \) as

\[
S_j := \{\hat{x}_i^j, 1 \leq i \leq N\},
\]

where \( \hat{x}_1^j = x_{j-1/2} \) and \( \hat{x}_N^j = x_{j+1/2} \). The cell average decomposition then takes the form

\[
\tilde{w}_j^h = \sum_{i=2}^{N-1} \hat{\omega}_i w^n_h(\hat{x}_i^j) + \hat{\omega}_1 w_{j-\frac{1}{2}}^+ + \hat{\omega}_N w_{j+\frac{1}{2}}^-,
\]

(4.3.11)
where it is known that $\hat{\omega}_1 = \hat{\omega}_N = 1/(N(N - 1))$. Hence (4.3.10) can be rewritten as a linear convex combination of the form

$$\bar{w}_{n+1}^j = \sum_{i=2}^{N-1} \hat{\omega}_i \bar{w}_{h}^i(\hat{x}_j^i) + \hat{\omega}_1 H_1 + \hat{\omega}_N H_N,$$

(4.3.12)

where

$$H_1 = w_{j-\frac{1}{2}}^+ - \frac{\lambda}{\hat{\omega}_1} \left( \hat{f}(w_{j-\frac{1}{2}}^+, w_{j+\frac{1}{2}}^-) - \hat{f}(w_{j-\frac{1}{2}}^-, w_{j+\frac{1}{2}}^+) \right),$$

$$H_N = w_{j+\frac{1}{2}}^- - \frac{\lambda}{\hat{\omega}_N} \left( \hat{f}(w_{j+\frac{1}{2}}^-, w_{j+\frac{1}{2}}^+) - \hat{f}(w_{j-\frac{1}{2}}^+, w_{j+\frac{1}{2}}^-) \right)$$

are of the same type as the first order scheme (4.3.2). The decomposition of (4.3.12) is first introduced by Zhang and Shu ([32]) for the compressible Euler equation and it suffices for us to conclude the following result.

Theorem 4.3.3 (High order scheme). A sufficient condition for $\bar{w}_{n+1}^j \in \Sigma_0$ by scheme (4.3.10) with an IRP flux is

$$w_{h}^i(x) \in \Sigma \text{ for } x \in S_j$$

under the CFL condition

$$\sigma \lambda \leq \frac{1}{N(N - 1)} c_0 \text{ with } N = \left\lceil \frac{k + 3}{2} \right\rceil,$$

(4.3.13)

where $\sigma$ is the global maximum of wave speed, $c_0$ is dependent on the IRP flux and $k$ is the degree of approximation polynomials.

4.3.2 Multi-dimensional case

To solve (4.1.1) over a computational cell $K$, we consider a high order DG scheme. That is, to find $w_h \in V_h$ such that

$$\int_K \partial_t (w_h) \phi dx - \int_K F(w_h) \cdot \nabla \phi dx + \sum_{i=1}^{Q} \int_{e_i^K} \check{F}(w_{h}^-, w_{h}^+, \nu^i) \phi ds = 0, \quad \forall \phi \in V_h,$$

(4.3.14)

where $F = (F_1, \ldots, F_d)^T$, $e_i^K$ is the $i$-th edge (or surface) of $K$, $\nu^i$ is the normal vector on $e_i^K$, $w_{h}^-$ and $w_{h}^+$ denote the approximation to $w_h$ on the edge of $K$ from interior and exterior of $K$ respectively, and $\check{F}$ is an admissible numerical flux.
Taking \( \phi = 1/|K| \) in this scheme, where \(|K|\) is the area (or volume) of the element, and evaluate the interface integral with an appropriate quadrature rule, we see that cell averages are actually evolved, when using the Euler-forward for time discretization, by

\[
\bar{w}_K^{n+1} = \bar{w}_K^n - \frac{\Delta t}{|K|} \sum_{i=1}^{Q} \sum_{\beta=1}^{L} w_{\beta} \hat{F}(w_{i,\beta}^{w_K}, w_{i,\beta}^{w_{K_i}}, \nu^i)|e_i^K|, \tag{4.3.15}
\]

where

\[
\bar{w}_K^n = \frac{1}{|K|} \int_K w_\alpha^K dK,
\]

\( w_K^{w_i} \) and \( w_K^{w_{K_i}} \) are approximations at \( \beta \)-th quadrature point to solution values on \( e_i^K \) from cell \( K \) and \( K_i \) respectively, \( w_{\beta} \) are corresponding quadrature weights, and the number of quadrature points \( L \) is chosen to achieve the desired accuracy.

Following [24, 32, 34] in the study of positivity-preserving schemes, we seek to rewrite the evolved cell average in (4.3.15) into a linear convex combination of terms which can then be shown to lie strictly within \( \Sigma \).

Assume one can construct an exact decomposition of the cell average:

\[
\bar{w}_K^n = \sum_{\alpha=1}^{P} c_{\alpha} w_K^{\alpha} + \sum_{i=1}^{L} \sum_{\beta=1}^{d_{\beta}} d_{\beta} w_{i,\beta}^{w_K}, \tag{4.3.16}
\]

where \( w_K^{\alpha} \) are approximations to solution values at some interior points in \( K \), \( c_{\alpha} \) and \( d_{\beta} \) are positive weights satisfying

\[
\sum_{\alpha=1}^{P} c_{\alpha} + Q \sum_{\beta=1}^{d_{\beta}} d_{\beta} = 1. \tag{4.3.17}
\]

This allows for the following reformulation

\[
\bar{w}_K^{n+1} = \sum_{\alpha=1}^{P} c_{\alpha} w_K^{\alpha} + \sum_{i=1}^{L} \sum_{\beta=1}^{d_{\beta}} d_{\beta} w_{i,\beta}^{w_K} - \frac{\Delta t}{|K|} \sum_{i=1}^{Q} \sum_{\beta=1}^{L} w_{\beta} \hat{F}(w_{i,\beta}^{w_K}, w_{i,\beta}^{w_{K_i}}, \nu^i)|e_i^K| = \sum_{\alpha=1}^{P} c_{\alpha} w_K^{\alpha} + \sum_{i=1}^{L} \sum_{\beta=1}^{d_{\beta}} H_{i,\beta}, \tag{4.3.18}
\]

where

\[
H_{i,\beta} = w_{i,\beta}^{w_K} - \frac{\Delta t |e_i^K| w_{\beta}}{|K| d_{\beta}} \left( \hat{F}(w_{i,\beta}^{w_K}, w_{i,\beta}^{w_{K_i}}, \nu^i) - \hat{F}(w_{i,\beta}^{w_K}, w_{i,\beta}^{w_{K_i}}, \nu^i) \right), \quad i = 1, \ldots, Q - 1,
\]

\[
H_{Q,\beta} = w_{Q,\beta}^{w_K} - \frac{\Delta t w_{\beta}}{|K| d_{\beta}} \left( \hat{F}(w_{Q,\beta}^{w_K}, w_{Q,\beta}^{w_{K_i}}, \nu^Q)|e_Q^K| + \sum_{i=1}^{Q-1} \hat{F}(w_{i,\beta}^{w_K}, w_{i,\beta}^{w_{K_i}}, \nu^i)|e_i^K| \right).
\]
Furthermore, $H_{Q,\beta}$ can be rewritten as

$$H_{Q,\beta} = \sum_{i=1}^{Q} \frac{|e_i^Q|}{|\partial K|} \left( w_{K}^{Q,\beta} - \frac{\Delta t w_{\beta}}{|K|d_{\beta}} |\partial K| \left( \hat{F}(w_{K}^{Q,\beta}, w_{K}^{Q,\beta}, \nu^i) - \hat{F}(w_{K}^{Q,\beta}, w_{K}^{Q,\beta}, \nu^j) \right) \right)$$

$$+ \frac{|e_K^Q|}{|\partial K|} \left( w_{K}^{Q,\beta} - \frac{\Delta t w_{\beta}}{|K|d_{\beta}} |\partial K| \left( \hat{F}(w_{K}^{Q,\beta}, w_{Q}^{\beta}, \nu^Q) - \hat{F}(w_{K}^{Q,\beta}, w_{Q}^{\beta}, \nu^Q) \right) \right),$$

where $|\partial K| = \sum_{i=1}^{Q} |e_i^K|$. Here we have used the identity

$$\sum_{i=1}^{Q} |e_i^K| \hat{F}(w_{K}^{Q,\beta}, w_{K}^{Q,\beta}, \nu^i) = \int_{\partial K} F(w_{K}^{Q,\beta}) \cdot \nu dS = \int_{K} \text{div}(F(w_{K}^{Q,\beta})) dx = 0,$$

where $\nu|e_i^K = \nu^i$. At this point it is clear that (4.3.18) is a linear convex combination of interior point values $w_{K}^{\alpha}$, and quantities of the form

$$w^* - c \left( \hat{F}(w^*, w_\nu), \nu - \hat{F}(w^*, w_\nu) \right),$$

where $c$ is a constant and $\nu$ is a unit vector. Note that (4.3.20) can be viewed as obtained from a formal first order scheme to one-dimensional system

$$\partial_t w + \partial_x (F \cdot \nu) = 0.$$  

Here $F \cdot \nu := \sum_{i=1}^{d} F_i \cdot \nu_i$ is a vector flux. Therefore, the cell average $w^{n+1}_{K}$ in (4.3.18) can be shown located in the invariant region under some CFL conditions, as long as we can show the system (4.3.21), also known as projected equations in [7], admits the same invariant region $\Sigma$ for all vectors $\nu \in \{\nu^i\}_{i=1}^{Q}$. We thus have proved the following result.

**Theorem 4.3.4.** Suppose there exists a positive quadrature rule such that (4.3.16) holds. If for $\nu \in \{\nu^i\}_{i=1}^{Q},$ (4.3.21) admits the same invariant region $\Sigma$, then a sufficient condition for $w^{n+1}_{K} \in \Sigma_0$ by scheme (4.3.15) with an IRP flux is

$$w^h_n(x) \in \Sigma \text{ for } x \in S_K,$$

under the CFL condition

$$\sigma \frac{\Delta t}{|K|} \leq \min_{\beta} \frac{d_{\beta}}{|\partial K| w_{\beta}^c} c_0,$$
where $\sigma = \max_i \sigma_i$, $\sigma_i = \max |\partial_w(F \cdot \nu^i)|$, $S_K$ is the set of quadrature points over $K$, $d_\beta$ are positive weights such that (4.3.16) is held and $w_\beta$ are quadrature weights used in (4.3.15), and $c_0 = 1$ or $1/2$ depending on the IRP flux used.

4.4 Application to compressible Euler equations

In this section, we apply the obtained results to the DG schemes for solving the compressible Euler equations.

4.4.1 1D case

We briefly review the IRP limiter first introduced in [15] for the one dimensional compressible Euler equations of the form (4.3.1) with

$$\mathbf{w} = (\rho, m, E)^\top, \quad f(\mathbf{w}) = (m, \rho u^2 + p, (E + p)u)^\top,$$

(4.4.1)

where $\rho$ is the density, $m = \rho u$ is the momentum, $E$ is the total energy, and $p$ is the pressure satisfying

$$E = \frac{1}{2} \rho u^2 + \frac{p}{\gamma - 1}, \quad \gamma > 1.$$

It is known that the system (4.3.1) with (4.4.1) has an invariant region:

$$\Sigma = \{ \mathbf{w} \mid \rho > 0, \ p > 0, \ q \leq 0 \},$$

(4.4.2)

where $q = (s_0 - s)\rho$,

$$s = \log \left( \frac{p(x)}{\rho^\gamma(x)} \right), \quad s_0 = \inf x \log \left( \frac{p_0(x)}{\rho_0^\gamma(x)} \right).$$

Here $\rho_0$ and $p_0$ are obtained from the given initial data $\mathbf{w}_0 = (\rho_0, m_0, E_0)^\top$.

In numerical simulations, we use the modified set of admissible states defined as

$$\Sigma' = \{ \mathbf{w} \mid \rho \geq \epsilon, \ p \geq \epsilon, \ q \leq 0 \},$$
and its interior defined as
\[ \Sigma_0^\epsilon = \{ w \mid \rho > \epsilon, p > \epsilon, q < 0 \}, \]
where \( \epsilon \) is a small positive number as the desired lower bound for density and pressure.

The one dimensional limiter introduced in Section 2 can be applied such that the modified polynomial \((4.2.2)\) lies entirely in \( \Sigma_0^\epsilon \) and is still a high order approximation to \( w(x) \), if \( \theta \) is chosen as
\[ \theta = \min \{ 1, \theta_1, \theta_2, \theta_3 \}, \tag{4.4.3} \]
where
\[ \theta_1 = \frac{\bar{\rho}_h - \epsilon}{\rho_h - \rho_{h,\text{min}}}, \quad \theta_2 = \frac{p(\bar{w}_h) - \epsilon}{p(\bar{w}_h) - p_{h,\text{min}}}, \quad \theta_3 = \frac{q(\bar{w}_h)}{q(\bar{w}_h) - q_{h,\text{max}}} \tag{4.4.4} \]
with
\[ \rho_{h,\text{min}} = \min_{x \in K} \rho_h(x), \quad p_{h,\text{min}} = \min_{x \in K} p(w_h(x)), \quad q_{h,\text{max}} = \max_{x \in K} q(w_h(x)), \tag{4.4.5} \]
where \( K \) may be chosen as \( S_j(j = 1, \cdots, N) \) in one dimensional case, dictated by Theorem 4.3.3.

Also in the CFL condition given in \((4.3.13)\), \( \sigma = \| |u| + c\|_{\infty} \), where \( c = \sqrt{\frac{\gamma p}{\rho}} \) is the sound speed.

### 4.4.2 2D case

Consider the two dimensional compressible Euler equations of the form
\[ \partial_t w + \nabla \cdot F = 0, \quad \tag{4.4.6} \]
with \( w = (\rho, m, n, E)^\top, \quad F(w) = (F_1(w), F_2(w)) \), where
\[
F_1(w) = (m, \rho u^2 + p, \rho uv, (E + p)u)^\top, \quad F_2(w) = (n, \rho uv, \rho v^2 + p, (E + p)v)^\top
\]
\[ m = \rho u, \quad n = \rho v, \quad E = \frac{1}{2} \rho u^2 + \frac{1}{2} \rho v^2 + \frac{p}{\gamma - 1}. \tag{4.4.7} \]

In order to apply Theorem 4.3.4, we first show that \((4.3.21)\) admits the same invariant region \( \Sigma \). The invariant region for weak solutions to hyperbolic conservation law systems is in general an open
problem due to lack of the well-posedness result. As observed by Lax in [18], Glimm’s solutions for one-dimensional systems satisfy all relevant entropy conditions, therefore, in [14] necessary and sufficient conditions for a region to be invariant for (Glimm) solutions of the system of conservation laws are given. We shall verify Hoff’s conditions, for which the building blocks are Riemann solutions, as needed in the present situation with $\Sigma$.

**Lemma 4.4.1.** Let $\mathbf{v}$ be any unit vector, then system (4.3.21) with $w = (\rho, m, n, E)\top$ and $F(w) = (F_1(w), F_2(w))$, where $F_1$ and $F_2$ are given in (4.4.7), has the same invariant region

$$\Sigma = \{w \mid \rho > 0, p > 0, q \leq 0\}.$$ 

**Proof.** Let $u^N = (u, v) \cdot \mathbf{v}$ and $u^T = (u, v) \cdot \mathbf{v}^\perp$. System (4.3.21) can be rewritten as

\begin{align*}
\partial_t \rho + \partial_\eta (\rho u^N) &= 0, \\
\partial_t (\rho u^N) + \partial_\eta (\rho (u^N)^2 + p) &= 0, \\
\partial_t (\rho u^T) + \partial_\eta (\rho u^N u^T) &= 0, \\
\partial_t E + \partial_\eta ((E + p) u^N) &= 0,
\end{align*}

(4.4.9)

where

$$p = (\gamma - 1)(E - \frac{1}{2} \rho (u^N)^2 - \frac{1}{2} \rho (u^T)^2)$$

as deduced from $p = (\gamma - 1)(E - \frac{1}{2} \rho u^2 - \frac{1}{2} \rho v^2)$. These equations using the primitive variables: $U = (\rho, u^N, u^T, p)\top$ may be written as

$$\partial_t U + A(U) \partial_\eta U = 0,$$

where

$$A = \begin{pmatrix}
  u^N & \rho & 0 & 0 \\
  0 & u^N & 0 & 1/\rho \\
  0 & 0 & u^N & 0 \\
  0 & \gamma p & 0 & u^N
\end{pmatrix}.$$
Its eigenvalues are $u^N - c, u^N, u^N + c$, where the speed of sound is $c = \sqrt{\gamma p/\rho}$. The associated left eigenvectors are

$$l^1 = (0, -\rho/2c, 0, 1/(2c^2)), \quad l^2 = (1, 0, 0, -1/c^2),$$
$$l^3 = (1, 0, 1, -1/c^2), \quad l^4 = (0, \rho/(2c), 0, 1/(2c^2)).$$

Consider $\partial \Sigma = \{ w \mid \rho > 0, p > 0, q = 0 \}$, its normal direction is

$$\vec{n} = (\gamma, 0, 0, -\gamma/c^2) = \gamma l^2,$$

and $\Sigma$ is convex. This meets the sufficient and necessary conditions given in [14, Corollary 3.3] for the intersection with a half space, therefore

$$\Sigma = \{(\rho, m, n, E)^\top \mid \rho > 0, p > 0, q \leq 0\}$$

is an invariant region for (4.4.9). 

We next identify the test sets $S_K$ as required in Theorem 4.3.4. In fact, the existing results on test sets for positivity-preserving DG schemes established in [32, 34] can still be used for the IRP DG schemes presented in this work. For approximation polynomials of degree $k$, we discuss two kinds of meshes in the following.

(i) For rectangular mesh $K = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, the test set $S_K$ in [32, Theorem 3.1] is

$$S_K = (S_i^x \times \hat{S}_j^y) \cup (\hat{S}_i^x \times S_j^y),$$

where

$$S_i^x = \{x_\beta, \beta = 1, \cdots, L\}, \quad S_j^y = \{y_\beta, \beta = 1, \cdots, L\}$$

are the Gauss quadrature points on $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and $[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ respectively and $L$ is chosen such that the quadrature rule is exact for single variable polynomials of degree $2k + 1$, and

$$\hat{S}_i^x = \{\hat{x}_\alpha, \alpha = 1, \cdots, N\}, \quad \hat{S}_j^y = \{\hat{y}_\alpha, \alpha = 1, \cdots, N\}$$
are the Gauss-Lobatto quadrature points on \([x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]\) and \([y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]\) respectively and \(N\) is chosen such that \(2N - 3 \geq k\).

In the form of decomposition (4.3.16) satisfying (4.3.17), the cell average \(\bar{w}_n^K\) is then given by

\[
\bar{w}_n^K = \sum_{\alpha=2}^{N-1} \sum_{\beta=1}^{L} \frac{1}{2} w_\beta \hat{w}_\alpha \left( w_h (x_\beta, y^\alpha_j) + w_h (\hat{x}_\beta, y^\alpha_j) \right) + \sum_{l=1}^{4} \sum_{\beta=1}^{L} \frac{1}{2} w_\beta \hat{w}_1 w_{l,\beta}^{K},
\]

where \(w_K^{l,\beta}\) are values of approximation polynomial \(w_h\) in \(K\) at \(\beta\)-th quadrature points on the \(l\)-th edge of \(K\) and \(\hat{w}_1\) is the weight of first Gauss-Lobatto quadrature point and is equal to \(\hat{w}_N\). Then according to Theorem 4.3.4, we find the CFL condition for IRP DG schemes on rectangular meshes is

\[
\sigma \frac{\Delta t}{\Delta x \Delta y} \leq \frac{\hat{w}_1 c_0}{4(\Delta x + \Delta y)}, \tag{4.4.10}
\]

where \(\sigma = \max\{||u| + c|_{\infty}, ||v| + c|_{\infty}\}\) and \(c_0\) is 1 or \(\frac{1}{2}\) depending on the IRP flux used.

**Remark 4.4.1.** For rectangular meshes, one could use a simple dimension by dimension decomposition for \(\bar{w}_n^{n+1}_K\), leading to a less restricted CFL condition on the time step, which is

\[
\Delta t = \frac{\hat{w}_1 c_0}{\sigma_1 \Delta x + \sigma_2 \Delta y}, \tag{4.4.11}
\]

as obtained in [32, (3.12)].

(ii) For triangular meshes, the authors in [34, (3.3)] introduced the following set of quadrature points, denoted by barycentric coordinates, as

\[
S_K = \left\{ \left( \frac{1}{2} + v^\beta, (\frac{1}{2} + \hat{u}^\alpha)(\frac{1}{2} - v^\beta), (\frac{1}{2} - \hat{u}^\alpha)(\frac{1}{2} - v^\beta) \right), \right. \\
\left. \left( \frac{1}{2} - \hat{u}^\alpha)(\frac{1}{2} - v^\beta), \frac{1}{2} + v^\beta, (\frac{1}{2} + \hat{u}^\alpha)(\frac{1}{2} - v^\beta) \right), \right. \\
\left. \left( \frac{1}{2} + \hat{u}^\alpha)(\frac{1}{2} - v^\beta), (\frac{1}{2} - \hat{u}^\alpha)(\frac{1}{2} - v^\beta), \frac{1}{2} + v^\beta \right) \right\},
\]

where \(\{\hat{u}^\alpha, \alpha = 1, \cdots, N\}\) and \(\{v^\beta, \beta = 1, \cdots, k+1\}\) are the Gauss-Lobatto quadrature points and Gauss quadrature points on \([\frac{-1}{2}, \frac{1}{2}]\) respectively and \(N\) is chosen such that \(2N - 3 \geq k\), where \(k\) is the degree of approximation polynomials.
We now show that this set can be used as a test set for the IRP DG schemes in Theorem 4.3.4. Based on this test set, a decomposition of the cell average that is of form (4.3.16) satisfying (4.3.17) was given in [34, (3.5)]:

\[
\bar{w}_K^n = \sum_{a=1}^{P} c_a w_{K}^{a} + \sum_{i=1}^{k+1} \sum_{\beta=1}^{2} \frac{2}{3} w_{\beta} \hat{w}_{1} w_{K}^{i,\beta},
\]

where \(P = 3(N-2)(k+1)\), \(w_{K}^{a}\) are approximations to solution values at quadrature points in the interior of \(K\), \(c_a\) are the corresponding weights, \(w_{K}^{i,\beta}\) are approximations at \(\beta\)-th quadrature point on \(e_{K}^i\) from cell \(K\), \(w_{\beta}\) are weights of Gauss quadrature points and \(\hat{w}_{1}\) is the weight of first Gauss-Lobatto quadrature point. Then, according to Theorem 4.3.4, we find that the CFL condition for IRP DG schemes on triangular meshes is

\[
\sigma \frac{\Delta t}{|K|} \leq \frac{2}{3|\partial K|} \hat{w}_{1} c_0,
\]  

(4.4.12)

where \(\sigma = \|\sqrt{u^2 + v^2} + c\|_{\infty}\) and \(c_0\) is \(\frac{1}{2}\) or 1 depending on the used IRP flux. Here we have used the fact that the eigenvalues for the Jacobian matrix \(\partial_u (F(w) \cdot \nu)\) are \(\{u \cdot \nu - c, u \cdot \nu, u \cdot \nu + c, u \cdot \nu\}\), where \(u = (u, v)\), and \(\nu\) is the unit vector, see [25].

With the test sets and CFL conditions given above, the IRP limiter (4.4.3) with (4.4.4) can be applied, where now \(x \in K \subset \mathbb{R}^2\).

### 4.5 Numerical tests

In this section, we present numerical examples to test the performance of IRP DG schemes presented in previous sections. Unless it is stated specifically, the IRP flux used is the local Lax-Friedrichs flux. \(\gamma = 1.4\) is taken for all of the examples.

The semi-discrete DG scheme is a closed ODE system

\[
\frac{d}{dt} W = L(W),
\]

where \(W\) consists of the unknown coefficients of the spatial basis, and \(L\) is the corresponding spatial operator.
For the time discretization we use the third order SSP Runge-Kutta (RK3) method introduced in [26]:

\[
W^{(1)} = W^n + \Delta t L(W^n), \\
W^{(2)} = \frac{3}{4} W^n + \frac{1}{4} W^{(1)} + \frac{1}{4} \Delta t L(W^{(1)}), \\
W^{n+1} = \frac{1}{3} W^n + \frac{2}{3} W^{(2)} + \frac{2}{3} \Delta t L(W^{(2)}).
\] (4.5.1)

We apply the IRP limiter at each time stage in the RK3 method with \(\epsilon = 10^{-13}\) for all of the examples. Notice that (4.5.1) is a linear convex combination of Euler forward method, therefore the invariant region is preserved by the full scheme if it is preserved at each time stage. Unless specified otherwise, in 1D problems, we use the time step \(\Delta t = \frac{\Delta x}{4\sigma}\) for \(P_1\)-DG scheme and \(\Delta t = \frac{\Delta x}{12\sigma}\) for \(P^2\)-DG scheme, where \(\sigma\) is the global maximum wave speed. In 2D problems, we use the time step \(\Delta t = \frac{1}{4\eta}\) for \(P_1\)-DG scheme and \(\Delta t = \frac{1}{12\eta}\), where \(\eta = \frac{\sigma_1}{\Delta x} + \frac{\sigma_2}{\Delta y}\), \(\sigma_1\) and \(\sigma_2\) are global maximum wave speeds along \(x\) direction and \(y\) direction, respectively.

**Example 1. 1D accuracy tests**

We test the accuracy of the IRP DG scheme solving the one-dimensional compressible Euler equations with periodic boundary conditions. The initial condition is

\[
\rho_0(x) = 1 + 0.5 \sin(2\pi x), \quad u_0(x) = 1, \quad p_0(x) = 1
\]

in domain \([0, 1]\). The exact solution is

\[
\rho(x, y, t) = 1 + 0.5 \sin(2\pi(x - t)), \quad u(x, t) = 1, \quad p(x, t) = 1.
\]

The final time is taken as \(T = 0.1\). Listed in Table 4.1 and Table 4.2 are the errors and orders of accuracy of the DG method for density with and without the IRP limiter, respectively. For the \(P^1\)-DG scheme, we observe the desired order of accuracy for our IRP DG method, justifying that the IRP limiter does not destroy the accuracy for smooth solutions. We also see that the errors with or without the limiter are comparable. While for \(P^2\)-DG scheme with IRP limiter, the third order of accuracy is observed before the mesh is refined to \(N = 512\). The reason for the order loss
with finer meshes could be that the DG polynomials in the intermediate stages in the Runge-Kutta method are already not of the desired order, then the accuracy of modified polynomials using the IRP limiter is also affected. Such phenomenon has also been observed and discussed in [33], in which the SSP multi-step time discretization is used to avoid such issue. We also observe that if we halve the time step for \( P^2 \)-DG scheme with the IRP limiter, then the desired order of accuracy can be recovered.

In order to see that the IRP limiter is indeed turned on, we present Table 4.3 (for \( P^1 \)-DG scheme) and Table 4.4 (for \( P^2 \)-DG scheme) showing some information when the IRP limiter is called for the first time in the code. Here \( N \) is the number of meshes used in spatial discretization, \( j \) denotes that the limiter is applied in the \( j \)-th cell, \( q_{\text{max}} \) is defined in (4.4.5) and \( \theta \) is the limiter parameter given in (4.4.3). For the \( P^3 \)-DG, numerical solutions remain in the invariant region, while no IRP limiter is called.

### Table 4.1 Accuracy of density function in Example 1 using \( P^1 \)-DG scheme.

<table>
<thead>
<tr>
<th>( P^1 ) DG</th>
<th>Without limiter</th>
<th>With IRP limiter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( L^\infty )Error</td>
<td>Order</td>
</tr>
<tr>
<td>16</td>
<td>4.97E-03         /</td>
<td>1.79E-03        /</td>
</tr>
<tr>
<td>32</td>
<td>1.27E-03         1.98</td>
<td>4.43E-04        2.02</td>
</tr>
<tr>
<td>64</td>
<td>3.18E-04         1.99</td>
<td>1.10E-04        2.00</td>
</tr>
<tr>
<td>128</td>
<td>7.96E-05         2.00</td>
<td>2.76E-05        2.00</td>
</tr>
<tr>
<td>256</td>
<td>1.99E-05         2.00</td>
<td>6.89E-06        2.00</td>
</tr>
<tr>
<td>512</td>
<td>4.99E-06         2.00</td>
<td>1.72E-06        2.00</td>
</tr>
<tr>
<td>1024</td>
<td>1.25E-06         2.00</td>
<td>4.30E-07        2.00</td>
</tr>
</tbody>
</table>

### Example 2. 2D accuracy tests

We test the accuracy of the IRP DG scheme on a low density problem in two dimensional case. The initial condition is

\[
\rho_0(x, y) = 1 + 0.99 \sin(x + y), \quad u_0(x, y) = 1, \quad v_0(x, y) = 1, \quad p_0(x, y) = 1
\]
Table 4.2 Accuracy test for density function in Example 1 using $P^2$-DG scheme.

<table>
<thead>
<tr>
<th>$P^2$ DG</th>
<th>Without limiter</th>
<th>With IRP limiter</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>$L^\infty$ Error</td>
<td>Order</td>
</tr>
<tr>
<td>16</td>
<td>3.20E-04 /</td>
<td>1.34E-04 /</td>
</tr>
<tr>
<td>32</td>
<td>4.35E-05 2.88</td>
<td>1.78E-05 2.92</td>
</tr>
<tr>
<td>64</td>
<td>5.54E-06 2.97</td>
<td>2.25E-06 2.98</td>
</tr>
<tr>
<td>128</td>
<td>6.94E-07 3.00</td>
<td>2.83E-07 2.99</td>
</tr>
<tr>
<td>256</td>
<td>8.68E-08 3.00</td>
<td>3.54E-08 3.00</td>
</tr>
<tr>
<td>512</td>
<td>1.08E-08 3.00</td>
<td>4.42E-09 3.00</td>
</tr>
<tr>
<td>1024</td>
<td>1.35E-09 3.00</td>
<td>5.53E-10 3.00</td>
</tr>
</tbody>
</table>

Table 4.3 Some parameter values when the limiter is first called in the $P^1$-DG scheme solving Example 1.

<table>
<thead>
<tr>
<th>N</th>
<th>j</th>
<th>$q_{\text{max}}$</th>
<th>$q(\bar{w}_j)$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>4</td>
<td>4.48E-03</td>
<td>-2.29E-02</td>
<td>0.8360</td>
</tr>
<tr>
<td>32</td>
<td>8</td>
<td>1.02E-03</td>
<td>-5.75E-03</td>
<td>0.8491</td>
</tr>
<tr>
<td>64</td>
<td>16</td>
<td>2.49E-04</td>
<td>-1.44E-03</td>
<td>0.8525</td>
</tr>
<tr>
<td>128</td>
<td>32</td>
<td>6.18E-05</td>
<td>-3.60E-04</td>
<td>0.8534</td>
</tr>
<tr>
<td>256</td>
<td>64</td>
<td>1.54E-05</td>
<td>-9.00E-05</td>
<td>0.8536</td>
</tr>
<tr>
<td>512</td>
<td>128</td>
<td>3.86E-06</td>
<td>-2.25E-05</td>
<td>0.8537</td>
</tr>
<tr>
<td>1024</td>
<td>256</td>
<td>9.64E-07</td>
<td>-5.62E-06</td>
<td>0.8537</td>
</tr>
</tbody>
</table>

Table 4.4 Some parameter values when the limiter is first called in the $P^2$-DG scheme solving Example 1.

<table>
<thead>
<tr>
<th>N</th>
<th>j</th>
<th>$q_{\text{max}}$</th>
<th>$q(\bar{w}_j)$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>5</td>
<td>6.25E-05</td>
<td>-1.61E-02</td>
<td>0.996141</td>
</tr>
<tr>
<td>32</td>
<td>9</td>
<td>3.91E-06</td>
<td>-4.07E-03</td>
<td>0.999040</td>
</tr>
<tr>
<td>64</td>
<td>17</td>
<td>2.44E-07</td>
<td>-1.02E-03</td>
<td>0.999760</td>
</tr>
<tr>
<td>128</td>
<td>33</td>
<td>1.53E-08</td>
<td>-2.55E-04</td>
<td>0.999940</td>
</tr>
<tr>
<td>256</td>
<td>65</td>
<td>9.54E-10</td>
<td>-6.37E-05</td>
<td>0.999985</td>
</tr>
<tr>
<td>512</td>
<td>129</td>
<td>5.96E-11</td>
<td>-1.59E-05</td>
<td>0.999996</td>
</tr>
<tr>
<td>1024</td>
<td>257</td>
<td>3.73E-12</td>
<td>-3.98E-06</td>
<td>0.999999</td>
</tr>
</tbody>
</table>
in domain \([0, 2\pi]\) with periodic boundary conditions. The exact solution is

\[
\rho(x, y, t) = 1 + 0.99 \sin(x + y - 2t),
\]

\[
u(x, y, t) = 1, v(x, y, t) = 1, p(x, y, t) = 1.
\]

The final time is \(T = 0.1\). From Table 4.5 and Table 4.6 we can see close to \(k + 1\)-th order of accuracy of the IRP DG scheme with polynomials of degree \(k\). Listed in Table 4.7 (for \(P^1\)-DG scheme) and Table 4.8 (for \(P^2\)-DG scheme) are values of \(\theta\) and related indicators when the IRP limiter is called in the code for the first time. \((i, j)\) denotes that the limiter is used in the \((i, j)\)-th cell.

| \(P^1\) DG | Without limiter | | | | | With IRP limiter | | | |
|---|---|---|---|---|---|---|---|---|
| \(N \times N\) | \(L^\infty\)Error | Order | \(L^1\)Error | Order | \(L^\infty\)Error | Order | \(L^1\)Error | Order |
| 32 \times 32 | 8.96E-03 | / | 2.12E-03 | / | 1.25E-02 | / | 2.45E-03 | / |
| 64 \times 64 | 2.18E-03 | 2.04 | 5.09E-04 | 2.06 | 2.88E-03 | 2.11 | 5.60E-04 | 2.13 |
| 128 \times 128 | 4.96E-04 | 2.14 | 1.24E-04 | 2.04 | 7.43E-04 | 1.95 | 1.32E-04 | 2.09 |
| 256 \times 256 | 1.23E-04 | 2.02 | 3.05E-05 | 2.02 | 1.86E-04 | 1.99 | 3.17E-05 | 2.06 |
| 512 \times 512 | 3.07E-05 | 2.00 | 7.58E-06 | 2.01 | 4.86E-05 | 1.94 | 7.76E-06 | 2.03 |

| \(P^2\) DG | Without limiter | | | | | With IRP limiter | | | |
|---|---|---|---|---|---|---|---|---|
| \(N\) | \(L^\infty\)Error | Order | \(L^1\)Error | Order | \(L^\infty\)Error | Order | \(L^1\)Error | Order |
| 16 | 2.84E-03 | / | 6.37E-04 | / | 3.66E-03 | / | 7.00E-04 | / |
| 32 | 4.04E-04 | 2.82 | 7.92E-05 | 3.01 | 4.04E-04 | 3.18 | 8.59E-05 | 3.03 |
| 64 | 5.11E-05 | 2.98 | 9.83E-06 | 3.01 | 5.11E-05 | 2.98 | 1.00E-05 | 3.10 |
| 128 | 6.39E-06 | 3.00 | 1.22E-06 | 3.01 | 6.39E-06 | 2.99 | 1.23E-06 | 3.03 |
| 256 | 8.06E-07 | 2.99 | 1.51E-07 | 3.01 | 8.06E-07 | 2.97 | 1.52E-07 | 3.02 |

In the following examples, we test the IRP DG schemes solving 1D and 2D Riemann problems, respectively. We compare the results between those with the IRP limiter (4.4.3) and those with only positivity-preserving limiter; that is, using \(\theta = \min\{1, \theta_1, \theta_2\}\), where \(\theta_1\) and \(\theta_2\) are defined as in (4.4.4).
Table 4.7  Some parameter values when the limiter is first called in the $P^1$-DG scheme solving Example 2.

<table>
<thead>
<tr>
<th>$N \times N$</th>
<th>(i,j)</th>
<th>$q_{max}$</th>
<th>$q(\bar{w}_j)$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$32 \times 32$</td>
<td>$(7,1)$</td>
<td>1.95E-02</td>
<td>-3.08E-02</td>
<td>0.6119</td>
</tr>
<tr>
<td>$64 \times 64$</td>
<td>$(15,1)$</td>
<td>4.93E-03</td>
<td>-7.77E-03</td>
<td>0.6120</td>
</tr>
<tr>
<td>$128 \times 128$</td>
<td>$(31,1)$</td>
<td>1.23E-03</td>
<td>-1.95E-03</td>
<td>0.6120</td>
</tr>
<tr>
<td>$256 \times 256$</td>
<td>$(63,1)$</td>
<td>3.09E-04</td>
<td>-4.87E-04</td>
<td>0.6120</td>
</tr>
<tr>
<td>$512 \times 512$</td>
<td>$(127,1)$</td>
<td>7.72E-05</td>
<td>-1.22E-04</td>
<td>0.6120</td>
</tr>
</tbody>
</table>

Table 4.8  Some parameter values when the limiter is first called in the $P^2$-DG scheme solving Example 2.

<table>
<thead>
<tr>
<th>$N \times N$</th>
<th>(i,j)</th>
<th>$q_{max}$</th>
<th>$q(\bar{w}_j)$</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$16 \times 16$</td>
<td>$(3,1)$</td>
<td>1.98E-03</td>
<td>-0.12</td>
<td>0.9836</td>
</tr>
<tr>
<td>$32 \times 32$</td>
<td>$(7,1)$</td>
<td>8.26E-05</td>
<td>-3.08E-02</td>
<td>0.9973</td>
</tr>
<tr>
<td>$64 \times 64$</td>
<td>$(17,1)$</td>
<td>1.52E-06</td>
<td>-7.60E-03</td>
<td>0.9998</td>
</tr>
<tr>
<td>$128 \times 128$</td>
<td>$(33,1)$</td>
<td>2.88E-07</td>
<td>-1.84E-03</td>
<td>0.9998</td>
</tr>
<tr>
<td>$256 \times 256$</td>
<td>$(65,1)$</td>
<td>2.69E-08</td>
<td>-4.55E-04</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

**Example 3. 1D Sod tube problem**

Consider the Sod initial data:

$$(\rho, m, E) = \begin{cases} 
(1, 0, 2.5), & x < 0, \\
(0.125, 0, 0.25), & x \geq 0.
\end{cases}$$

The exact solution, which can be obtained by using the formula given in [22, section 14.11,14.12], consists of a composite wave, that is, a rarefaction wave followed by a contact discontinuity and then by a shock. The numerical solution obtained from the $P^2$-DG scheme on 200 cells at final time $T = 0.16$ is displayed in Figure 4.1, from which we can see that the IRP limiter helps to reduce the oscillations near the interface between the rarefaction waves and the contact discontinuity. It also helps to damp the overshoots and the undershoots. For example, with positivity-preserving limiter alone, the maximum and minimum of velocity solution is 1.004 and −0.177, respectively; while by using the IRP limiter, the maximum and minimum of velocity become 0.998 and −0.171, respectively.
Figure 4.1  Sod shock tube problem. Exact solution (solid line) vs numerical solution (dots); Left: With positive-preserving limiter; Right: With invariant-region-preserving limiter
Example 4. 1D double rarefaction problem

We consider the one dimensional Riemann problem with initial condition

\[
(\rho, u, p) = \begin{cases} 
(1, -12, 1), & x < 0, \\
(1, 12, 1), & x \geq 0.
\end{cases}
\]

The exact solution consists of two double rarefaction waves moving in opposite directions, which results in the creation of a vacuum in the center of the domain. For this problem, we use the global Lax-Friedrich flux and a reduced time step in order to avoid the blow-ups in the computation. The numerical solution obtained from the \( P^2 \)-DG scheme on 400 cells is displayed at the final time \( T = 0.3 \) in Figure 4.2, from which we see that both numerical schemes capture the vacuum region well. Moreover, the IRP DG scheme can help to damp the overshoots near the top of the rarefactions. Although we use \( \Delta t = \frac{\Delta x}{200} \) to obtain results in Figure 4.2, it's been noticed that the time step required to obtain a reasonable solution by the IRP DG scheme (\( \Delta t = \frac{\Delta x}{115} \)) is less restricted than the one by the scheme using only positivity-preserving limiter (\( \Delta t = \frac{\Delta x}{195} \)).

Example 5. 2D Riemann problem

For two dimensional Euler equations, there are nineteen configurations of Riemann solutions that have been studied in [20, 2]. In this example, we test only two configurations.

The first configuration (Configuration 2 in [20]) has the initial condition as

\[
(\rho, u, v, p) = \begin{cases} 
(1, 0, 0, 1), & (x, y) \in (0.5, 1) \times (0.5, 1) \\
(0.5197, -0.7259, 0, 0.4), & (x, y) \in (0, 0.5) \times (0.5, 1) \\
(1, -0.7259, -0.7259, 1), & (x, y) \in (0, 0.5) \times (0, 0.5) \\
(0.5197, 0, -0.7259, 0.4), & (x, y) \in (0.5, 1) \times (0, 0.5)
\end{cases}.
\]

The solution consists of four rarefaction waves. From the contour plots in Figure 4.3, we can see the IRP limiter helps to make the solution smoother.
Figure 4.2 1D double rarefaction problem. Exact solution (solid line) vs numerical solution (dots); Left: with positivity-preserving limiter; Right: with IRP limiter
The second configuration (Configuration 6 in [20]) has the initial condition as

$$(\rho, u, v, p) = \begin{cases} (1, 0.75, -0.5, 1), & (x, y) \in (0.5, 1) \times (0.5, 1) \\ (2, 0.75, 0.5, 1), & (x, y) \in (0, 0.5) \times (0.5, 1) \\ (1, -0.75, 0.5, 1), & (x, y) \in (0, 0.5) \times (0, 0.5) \\ (3, -0.75, -0.5, 1), & (x, y) \in (0.5, 1) \times (0, 0.5) \end{cases}.$$ 

The solution consists of four two-dimensional slip lines. See Figure 4.4. We zoom in the plot near the lower left and lower right interface between two constant states and observe that the IRP limiter helps to damp some of the oscillations. See Figure 4.5.

![Figure 4.3 Configuration 2. Contour plot of numerical solution of density with 30 contour levels. Left: with positive-preserving limiter; Right: with invariant-region-preserving limiter](image)

Remark 4.5.1. From the examples above, we can see that the IRP limiter presented in this work is still a mild limiter, especially for two dimensional case, and oscillations may not be completely damped even the invariant region has been preserved. For stronger oscillations, some stricter limiters may be needed.

4.6 Concluding remarks

In this paper we investigate invariant-region-preserving (IRP) DG schemes for multi-dimensional hyperbolic conservation law systems, with an application to the compressible Euler equations.
Figure 4.4  Configuration 6. Contour plot of numerical solution of density with 30 contour levels. Left: with positive-preserving limiter; Right: with invariant-region-preserving limiter

Figure 4.5  Zoom-in plot of contour plots of configuration 6. Left: with positive-preserving limiter; Right: with invariant-region-preserving limiter. Top: lower-left interface; Bottom: lower-right interface.
Assume that the underlying system admits a global invariant region which is a convex set in the phase space, an explicit IRP limiter is implemented in such a way that the cell averages remain in the invariant region for the entire simulation, which adds a degree of robustness to our IRP DG schemes. We rigorously prove that the invariant region limiter maintains both conservation and high order accuracy. The loss of accuracy might occur when the cell average is close to the boundary of the convex set. A generic algorithm incorporating the IRP limiter is presented for high order finite volume type schemes, and sufficient conditions are further identified if we assume the projected one-dimensional system shares the same invariant region as the full multi-dimensional hyperbolic system. We then apply the results to both one and two dimensional compressible Euler equations so to obtain high order IRP DG schemes. We demonstrate the effectiveness and efficiency of the IRP DG schemes on one- and two-dimensional compressible Euler equations. High-order accuracy is retained after applying the IRP limiter to a set of test problems, while some oscillations in the numerical solution are damped by the limiter as desired.

Acknowledgments

We would like to thank the associate editor and two reviewers for many constructive comments that improved the presentation of the paper.

Appendix A. Proof of Lemma 4.2.1 for compressible Euler equations.

In this appendix, we show that for the compressible Euler equations, where the pressure function is concave but not strictly concave, Lemma 4.2.1 still holds. We consider the one-dimensional case, where the invariant region is given in (4.4.2).

Proof. Since $p$ is concave, using Jensen’s inequality and the assumption, we have

$$p(\bar{w}) = p\left(\frac{1}{|K|} \int_K w(x) dx\right) \geq \frac{1}{|K|} \int_K p(w(x)) dx \geq 0.$$
With this, we can show $p(\bar{w}) < 0$. Otherwise if $p(\bar{w}) = 0$, we must have $p(w(x)) = 0$ for all $x \in K$; that is

$$p(\bar{w}) = p(w). \quad (4.6.1)$$

Upon taking cell average of this relation on both sides, we have

$$p(\bar{w}) = \frac{1}{|K|} \int_K p(w(x))dx.$$ 

By taking the Taylor expansion around $\bar{w}$, we have

$$p(w(x)) = p(\bar{w}) + \nabla_w p(\bar{w}) \cdot \xi + \xi^\top H \xi, \quad \forall x \in I, \quad \xi := w(x) - \bar{w},$$

which upon integration yields $\frac{1}{|K|} \int_K \xi^\top H \xi dx = 0$, where $H$ is the Hessian matrix of $p$:

$$H = (\gamma - 1) \begin{pmatrix} -\frac{m^2}{\rho^3} & \frac{m}{\rho^2} & 0 \\ \frac{m}{\rho^2} & -\frac{1}{\rho} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$ 

Since $p$ is a concave function of $w = (\rho, m, E)^\top$, then $H$ is semi-definite and we have $\xi^\top H \xi \equiv 0$. Therefore, $\xi$ must be in the eigenvalue space corresponding to the zero eigenvalue, that is $\xi = c_1 v_1 + c_2 v_2$, where $v_1 = (0, 0, 1)^\top$ and $v_2 = (\frac{p(x)}{m(x)}, 1, 0)^\top$. Hence $m(x)$ and $E(x)$ must be constants, so is $\rho(x)$ following from (4.6.1), which contradicts the assumption.

Appendix B. Proof of Lemma 4.3.2

In this appendix, we present the proof for Lemma 4.3.2.

**Proof.** When the HLLC flux is used in (4.3.2), there are sixteen different cases in total. Among them, four cases have been included in the HLL flux, so we only need to verify the other twelve cases.

For each case, we rewrite $w_j^{n+1}$ in (4.3.2) as a convex linear combination of some terms that can be shown in the invariant region. Here we use $\sigma_{k,l}$, $\sigma_{k,r}$ and $\sigma_{k,*}$ to denote the leftmost, rightmost
and middle wave speeds at \( x_k \) for \( k = j \pm 1 \), and \((w_{sl})_k\) and \((w_{sr})_k\) to denote the two intermediate states at \( n - th \) time step corresponding to \( x_k \) for \( k = j \pm \frac{1}{2} \).

**Case 1:** If \( \sigma_{j-\frac{1}{2},l} \geq 0 \) and \( \sigma_{j+\frac{1}{2},l} \leq 0 \leq \sigma_{j+\frac{1}{2},r} \), then

\[
\hat{f}_{j-\frac{1}{2}} = f(w_{j-1}^n), \quad \hat{f}_{j+\frac{1}{2}} = (f_{sr})_{j+\frac{1}{2}} = f(w_j^n) + \sigma_{j+\frac{1}{2},l} \left( (w_{sl})_{j+\frac{1}{2}} - w_j^n \right),
\]

and

\[
w_j^{n+1} = w_j^n - \lambda((f_{sr})_{j+\frac{1}{2}} - f(w_{j-1}^n))
= \left(1 + \lambda \sigma_{j+\frac{1}{2},l} - \lambda \sigma_{j-\frac{1}{2},r} \right) w_j^n + \left( -\lambda \sigma_{j+\frac{1}{2},l} \right) (w_{sl})_{j+\frac{1}{2}} + \lambda \sigma_{j-\frac{1}{2},r} \hat{w},
\]

where

\[
\hat{w} = \frac{\sigma_{j-\frac{1}{2},l} w_{j-1}^n - 0 \cdot w_{j-1}^n}{\sigma_{j-\frac{1}{2}} - 0} - \frac{f(w_j^n) - f(w_{j-1}^n)}{\sigma_{j-\frac{1}{2},r} - 0}.
\] (4.6.2)

**Case 2:** If \( \sigma_{j-\frac{1}{2},l} \geq 0 \) and \( \sigma_{j+\frac{1}{2},s} \leq 0 \leq \sigma_{j+\frac{1}{2},r} \), then

\[
\hat{f}_{j-\frac{1}{2}} = f(w_{j-1}^n), \quad \hat{f}_{j+\frac{1}{2}} = (f_{sr})_{j+\frac{1}{2}} = (f_{sl})_{j+\frac{1}{2}} + \sigma_{j+\frac{1}{2},s} \left( (w_{sr})_{j+\frac{1}{2}} - (w_{sl})_{j+\frac{1}{2}} \right),
\]

and

\[
w_j^{n+1} = w_j^n - \lambda((f_{sr})_{j+\frac{1}{2}} - f(w_{j-1}^n))
= w_j^n - \lambda((f_{sl})_{j+\frac{1}{2}} - f(w_{j-1}^n)) - \lambda \sigma_{j+\frac{1}{2},s} \left. \left. \left( (w_{sr})_{j+\frac{1}{2}} - (w_{sl})_{j+\frac{1}{2}} \right) \right) \right)
= \left(1 + \lambda \sigma_{j+\frac{1}{2},l} - \lambda \sigma_{j-\frac{1}{2},r} \right) w_j^n + \lambda \sigma_{j-\frac{1}{2},r} \hat{w}
+ \lambda \left( \sigma_{j+\frac{1}{2},s} - \sigma_{j+\frac{1}{2},l} \right) (w_{sl})_{j+\frac{1}{2}} + \left( -\lambda \sigma_{j+\frac{1}{2},s} \right) (w_{sr})_{j+\frac{1}{2}},
\]

where \( \hat{w} \) is given by (4.6.2).

**Case 3:** If \( \sigma_{j-\frac{1}{2},s} \leq 0 \leq \sigma_{j-\frac{1}{2},r} \) and \( \sigma_{j+\frac{1}{2},r} \leq 0 \), then

\[
\hat{f}_{j-\frac{1}{2}} = (f_{sr})_{j-\frac{1}{2}} = f(w_j^n) + \sigma_{j-\frac{1}{2},r} \left( (w_{sr})_{j-\frac{1}{2}} - w_j^n \right), \quad \hat{f}_{j+\frac{1}{2}} = f(w_{j+1}^n),
\]
and

$$w_{j}^{n+1} = w_{j}^{n} - \lambda \left( f(w_{j+1}^{n}) - (f_{sr})_{j - \frac{1}{2}} \right)$$

$$= \lambda \left( \sigma_{j + \frac{1}{2}, r} - \sigma_{j + \frac{1}{2}, l} \right) \hat{w} + \left( -\lambda \sigma_{j + \frac{1}{2}, r} \right) w_{j+1}^{n}$$

$$+ \left( 1 - \lambda \sigma_{j - \frac{1}{2}, r} + \lambda \sigma_{j + \frac{1}{2}, l} \right) w_{j}^{n} + \lambda \sigma_{j - \frac{1}{2}, r} (w_{sr})_{j - \frac{1}{2}},$$

where

$$\hat{w} = \frac{\sigma_{j + \frac{1}{2}, r} w_{j+1}^{n} - \sigma_{j + \frac{1}{2}, l} w_{j}^{n}}{\sigma_{j + \frac{1}{2}, r} - \sigma_{j + \frac{1}{2}, l}} - \frac{f(w_{j+1}^{n}) - f(w_{j}^{n})}{\sigma_{j + \frac{1}{2}, r} - \sigma_{j + \frac{1}{2}, l}}.$$

**Case 4:** If $\sigma_{j - \frac{1}{2}, l} \leq 0 \leq \sigma_{j - \frac{1}{2}, s}$ and $\sigma_{j + \frac{1}{2}, r} \leq 0$, then

$$\hat{f}_{j - \frac{1}{2}} = (f_{st})_{j - \frac{1}{2}} = (f_{sr})_{j - \frac{1}{2}} - \sigma_{j - \frac{1}{2}, s} \left( (w_{sr})_{j - \frac{1}{2}} - (w_{st})_{j - \frac{1}{2}} \right), \quad \hat{f}_{j + \frac{1}{2}} = f(w_{j+1}^{n}),$$

and

$$w_{j+1}^{n} = w_{j}^{n} - \lambda \left( f(w_{j+1}^{n}) - (f_{st})_{j - \frac{1}{2}} \right)$$

$$= w_{j}^{n} - \lambda \left( f(w_{j+1}^{n}) - (f_{sr})_{j - \frac{1}{2}} \right) - \lambda \sigma_{j - \frac{1}{2}, s} \left( (w_{sr})_{j - \frac{1}{2}} - (w_{st})_{j - \frac{1}{2}} \right)$$

$$= \lambda \left( \sigma_{j + \frac{1}{2}, r} - \sigma_{j + \frac{1}{2}, l} \right) \hat{w} + \left( -\lambda \sigma_{j + \frac{1}{2}, r} \right) w_{j+1}^{n} + \lambda \sigma_{j - \frac{1}{2}, s} (w_{sr})_{j - \frac{1}{2}},$$

where $\hat{w}$ is given by (4.6.3).

**Case 5:** If $\sigma_{j - \frac{1}{2}, l} \leq 0 \leq \sigma_{j - \frac{1}{2}, s}$ and $\sigma_{j + \frac{1}{2}, l} \leq 0 \leq \sigma_{j + \frac{1}{2}, s}$, then

$$\hat{f}_{j - \frac{1}{2}} = (f_{st})_{j - \frac{1}{2}} = (f_{sr})_{j - \frac{1}{2}} - \sigma_{j - \frac{1}{2}, s} \left( (w_{sr})_{j - \frac{1}{2}} - (w_{st})_{j - \frac{1}{2}} \right),$$

$$\hat{f}_{j + \frac{1}{2}} = (f_{st})_{j + \frac{1}{2}} = f(w_{j}^{n}) + \sigma_{j + \frac{1}{2}, l} (w_{st})_{j + \frac{1}{2}}$$

and

$$w_{j+1}^{n} = w_{j}^{n} - \lambda \left( f(w_{j+1}^{n}) - (f_{st})_{j - \frac{1}{2}} \right)$$

$$= \left( 1 + \sigma_{j + \frac{1}{2}, l} - \sigma_{j + \frac{1}{2}, r} \right) w_{j}^{n} + \left( -\lambda \sigma_{j + \frac{1}{2}, l} \right) (w_{st})_{j + \frac{1}{2}},$$

$$+ \lambda \left( \sigma_{j + \frac{1}{2}, r} - \sigma_{j + \frac{1}{2}, s} \right) (w_{sr})_{j - \frac{1}{2}} + \lambda \sigma_{j - \frac{1}{2}, s} (w_{st})_{j - \frac{1}{2}}.$$
Case 6: If \( \sigma_{j-\frac{1}{2},*} \leq 0 \leq \sigma_{j-\frac{1}{2},r} \) and \( \sigma_{j+\frac{1}{2},*} \leq 0 \leq \sigma_{j+\frac{1}{2},r} \), then

\[
\hat{f}_{j-\frac{1}{2}} = (f_{sr})_{j-\frac{1}{2}} = f(w_j^n) + \sigma_{j-\frac{1}{2},r} \left( (w_{sr})_{j-\frac{1}{2}} - w_j^n \right),
\]
\[
\hat{f}_{j+\frac{1}{2}} = (f_{sr})_{j+\frac{1}{2}} = (f_{st})_{j+\frac{1}{2}} + \sigma_{j+\frac{1}{2},*} \left( (w_{sr})_{j+\frac{1}{2}} - (w_{st})_{j+\frac{1}{2}} \right),
\]

then

\[
w_j^{n+1} = w_j^n - \lambda \left( (f_{sr})_{j+\frac{1}{2}} - (f_{sr})_{j-\frac{1}{2}} \right)
= \left( 1 + \lambda \sigma_{j+\frac{1}{2},l} - \lambda \sigma_{j-\frac{1}{2},r} \right) w_j^n + \left( -\sigma_{j+\frac{1}{2},*} \right) (w_{sr})_{j+\frac{1}{2}}
+ \lambda \left( \sigma_{j+\frac{1}{2},*} - \sigma_{j+\frac{1}{2},r} \right) (w_{st})_{j+\frac{1}{2}} + \lambda \sigma_{j-\frac{1}{2},r} (w_{sr})_{j-\frac{1}{2}}.
\]

Case 7: If \( \sigma_{j-\frac{1}{2},*} \leq 0 \leq \sigma_{j-\frac{1}{2},r} \) and \( \sigma_{j+\frac{1}{2},l} \leq 0 \leq \sigma_{j+\frac{1}{2},r} \), then

\[
\hat{f}_{j-\frac{1}{2}} = (f_{sr})_{j-\frac{1}{2}} = f(w_j^n) + \sigma_{j-\frac{1}{2},r} \left( (w_{sr})_{j-\frac{1}{2}} - w_j^n \right),
\]
\[
\hat{f}_{j+\frac{1}{2}} = (f_{sr})_{j+\frac{1}{2}} = (f_{st})_{j+\frac{1}{2}} + \sigma_{j+\frac{1}{2},l} \left( (w_{sr})_{j+\frac{1}{2}} - w_j^n \right),
\]

and

\[
w_j^{n+1} = w_j^n - \lambda \left( (f_{sr})_{j+\frac{1}{2}} - (f_{sr})_{j-\frac{1}{2}} \right)
= \left( 1 + \lambda \sigma_{j+\frac{1}{2},l} - \lambda \sigma_{j-\frac{1}{2},r} \right) w_j^n + \left( -\lambda \sigma_{j+\frac{1}{2},*} \right) (w_{sr})_{j+\frac{1}{2}}
+ \lambda \sigma_{j-\frac{1}{2},l} (w_{sr})_{j-\frac{1}{2}}.
\]

Case 8: If \( \sigma_{j-\frac{1}{2},l} \leq 0 \leq \sigma_{j-\frac{1}{2},*} \) and \( \sigma_{j+\frac{1}{2},*} \leq 0 \leq \sigma_{j+\frac{1}{2},r} \), then

\[
\hat{f}_{j-\frac{1}{2}} = (f_{sr})_{j-\frac{1}{2}} = (f_{sr})_{j-\frac{1}{2}} - \sigma_{j-\frac{1}{2},*} \left( (w_{sr})_{j-\frac{1}{2}} - (w_{st})_{j-\frac{1}{2}} \right),
\]
\[
\hat{f}_{j+\frac{1}{2}} = (f_{sr})_{j+\frac{1}{2}} = (f_{st})_{j+\frac{1}{2}} + \sigma_{j+\frac{1}{2},*} \left( (w_{sr})_{j+\frac{1}{2}} - (w_{st})_{j+\frac{1}{2}} \right),
\]

and

\[
w_j^{n+1} = w_j^n - \lambda \left( (f_{sr})_{j+\frac{1}{2}} - (f_{sr})_{j-\frac{1}{2}} \right)
= \left( 1 + \lambda \sigma_{j+\frac{1}{2},l} - \lambda \sigma_{j-\frac{1}{2},r} \right) w_j^n + \lambda \sigma_{j-\frac{1}{2},*} (w_{sr})_{j-\frac{1}{2}}
+ \left( \lambda \sigma_{j+\frac{1}{2},*} - \lambda \sigma_{j+\frac{1}{2},*} \right) (w_{sr})_{j+\frac{1}{2}}.
\]
Case 9: If $\sigma_{j+\frac{1}{2},r} \leq 0 \leq \sigma_{j-\frac{1}{2},r}$ and $0 \leq \sigma_{j+\frac{1}{2},l}$, then

$$
\hat{f}_{j-\frac{1}{2}} = (f_{sr})_{j-\frac{1}{2}} = f(w^n_j) + \sigma_{j-\frac{1}{2},r} \left((w_{sr})_{j-\frac{1}{2}} - w^n_j\right), \quad \hat{f}_{j+\frac{1}{2}} = f(w^n_j),
$$

and

$$
w^{n+1}_j = w^n_j - \lambda \left(f(w^n_j) - (f_{sr})_{j-\frac{1}{2}}\right) = \left(1 - \lambda \sigma_{j-\frac{1}{2},r}\right) w^n_j + \lambda \sigma_{j-\frac{1}{2},r} (w_{sr})_{j-\frac{1}{2}}.
$$

Case 10: If $\sigma_{j+\frac{1}{2},l} \leq 0 \leq \sigma_{j-\frac{1}{2},r}$ and $0 \leq \sigma_{j+\frac{1}{2},r}$, then

$$
\hat{f}_{j-\frac{1}{2}} = (f_{sr})_{j-\frac{1}{2}} = (f_{sr})_{j-\frac{1}{2}} - \sigma_{j-\frac{1}{2},r} \left((w_{sr})_{j-\frac{1}{2}} - (w_{sl})_{j-\frac{1}{2}}\right), \quad \hat{f}_{j+\frac{1}{2}} = f(w^n_j),
$$

and

$$
w^{n+1}_j = w^n_j - \lambda \left(f(w^n_j) - (f_{sr})_{j-\frac{1}{2}}\right) = \left(1 - \lambda \sigma_{j-\frac{1}{2},r}\right) w^n_j + \lambda \sigma_{j-\frac{1}{2},r} (w_{sr})_{j-\frac{1}{2}}.
$$

Case 11: If $\sigma_{j+\frac{1}{2},r} \leq 0$ and $\sigma_{j+\frac{1}{2},l} \leq 0 \leq \sigma_{j-\frac{1}{2},r}$, then

$$
\hat{f}_{j-\frac{1}{2}} = f(w^n_j), \quad \hat{f}_{j+\frac{1}{2}} = (f_{sr})_{j+\frac{1}{2}} = f(w^n_j) + \sigma_{j+\frac{1}{2},l} \left((w_{sl})_{j+\frac{1}{2}} - w^n_j\right),
$$

and

$$
w^{n+1}_j = w^n_j - \lambda \left(f_{sr})_{j+\frac{1}{2}} - f(w^n_j)\right) = \left(1 - \lambda \sigma_{j+\frac{1}{2},l}\right) w^n_j + \left(-\lambda \sigma_{j+\frac{1}{2},l}\right) (w_{sl})_{j+\frac{1}{2}}.
$$

Case 12: If $\sigma_{j-\frac{1}{2},r} \leq 0$ and $\sigma_{j+\frac{1}{2},r} \leq 0 \leq \sigma_{j+\frac{1}{2},l}$, then

$$
\hat{f}_{j-\frac{1}{2}} = f(w^n_j), \quad \hat{f}_{j+\frac{1}{2}} = (f_{sr})_{j+\frac{1}{2}} = (f_{sr})_{j+\frac{1}{2}} + \sigma_{j+\frac{1}{2},l} \left((w_{sr})_{j+\frac{1}{2}} - (w_{sl})_{j+\frac{1}{2}}\right),
$$

and

$$
w^{n+1}_j = w^n_j - \lambda \left(f_{sr})_{j+\frac{1}{2}} - f(w^n_j)\right) = \left(1 - \lambda \sigma_{j+\frac{1}{2},l}\right) w^n_j + \left(\lambda \sigma_{j+\frac{1}{2},l} - \lambda \sigma_{j+\frac{1}{2},r}\right) (w_{sl})_{j+\frac{1}{2}} + \left(-\lambda \sigma_{j+\frac{1}{2},r}\right) (w_{sr})_{j+\frac{1}{2}}.
For $w^n_j, w^n_{j\pm1} \in \Sigma$, the intermediate states $(w_{st})_k$ and $(w_{sl})_k$, $k = j \pm \frac{1}{2}$ are in $\Sigma$. Also notice that $\hat{w}$ in Case 1-4 are all in the form of (4.3.4), the cell average of some exact Riemann solutions, and therefore lie in $\Sigma_0$ by Lemma 4.2.1. Hence $\lambda \sigma \leq \frac{1}{2}$ is a sufficient condition for $w^{n+1}_j$ defined in (4.3.2) to be in $\Sigma_0$.

References


CHAPTER 5. IMPLEMENTATION DETAILS

In this chapter, we discuss some implementation details of IRP DG schemes with the SSP RK3 method solving one-dimensional hyperbolic conservation law systems. For reader’s convenience, we present a Matlab script for solving an accuracy test problem of the one-dimensional compressible Euler equations in the appendix.

We consider the one-dimensional problem

\[ \partial_t w + \partial_x f(w) = 0, \quad t > 0, x \in \mathbb{R}, \]  

subject to the initial condition \( w(x,0) = w_0(x) \) and periodic boundary conditions, where \( w \in \mathbb{R}^l \).

Assume the system admits an invariant region \( \Sigma \) given by

\[ \Sigma = \bigcap_{i=1}^{M} \{ w \mid U_i(w) \leq 0 \}. \]

5.1 Mesh generation and solution space

Consider a partition of the domain \([a,b]\) as below

\[ a = x_\frac{1}{2} < x_\frac{3}{2} < \cdots < x_{N-\frac{1}{2}} < x_{N+\frac{1}{2}} = b. \]

Let \( I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \) be the \( j \)-th computational cell, \( j = 1, \cdots, N \). For the sake of simplicity, we assume the uniform mesh size, denoted by \( \Delta x = \frac{b-a}{N} \).

The solution space is formed by vectors whose components are in \( V_h = \{ v : v|_{I_j} \in \mathbb{P}^k(I_j) \} \), where \( \mathbb{P}^k(I_j) \) is the space of polynomials of degree \( k \) on \( I_j \).
5.2 Scheme formulation

5.2.1 Local formulation

The DG method for solving (5.0.1) is to find $w_h$ in the solution space such that

$$\int_{I_j} (w_h)_t \psi dx - \int_{I_j} f(w_h) \psi x dx + \hat{f}(w_h^+, w_h^-) \psi \big|_{\partial I_j} = 0, \quad \forall \psi \in V_h,$$

for all $1 \leq j \leq N$, where $\psi \big|_{I_j} \equiv \psi(x_{j+\frac{1}{2}}) - \psi(x_{j-\frac{1}{2}})$, $\hat{f}$ is the numerical flux and $w_h^\pm$ are the approximations to function values at the interface using the information from right (+) and left (−) neighboring cells respectively. To enforce the periodic boundary condition, we define the numerical flux at the boundary using $w_h(x_{\frac{1}{2}}) = w_h(x_{N+\frac{1}{2}})$, $w_h(x_{N+\frac{1}{2}}) = w_h(x_{\frac{1}{2}})$.

5.2.2 Global formulation

Summation of the local formulation over all cells gives

$$\langle (w_h)_t, \psi \rangle + \langle \mathcal{L}(w_h), \psi \rangle = 0, \quad \forall \psi \in V_h,$$

where

$$\langle (w_h)_t, \psi \rangle = \sum_{j=1}^{N} \int_{I_j} (w_h)_t \psi dx,$$

$$\langle \mathcal{L}(w_h), \psi \rangle = -\sum_{j=1}^{N} \int_{I_j} f(w_h) \psi x dx - \sum_{j=1}^{N} \hat{f}(w_h^-, w_h^+) \cdot [\psi]_{j-\frac{1}{2}},$$

and $[\cdot]$ denotes the jump of the function across the cell interface.

5.3 DG implementation

5.3.1 Method of line

We express the solution $w_h(x, t)$ on each cell $I_j$ using the normalized Legendre polynomials $\{\phi_i(\xi)\}_{i=0}^{k}$ on the reference interval $[-1, 1]$ as the basis functions:

$$w_h(x, t) \big|_{I_j} = \sum_{i=0}^{k} W^i_j(t) \phi_i(\xi),$$

(5.3.1)
where $x = x_j + \frac{\Delta x}{2} \xi, \phi_0(\xi) = \sqrt{\frac{1}{2}}, \phi_1(\xi) = \sqrt{\frac{3}{2}} \xi, \phi_2(\xi) = \sqrt{\frac{5}{8}} (3\xi^2 - 1), \phi_3(\xi) = \sqrt{\frac{7}{16}} (5\xi^3 - 3\xi), \ldots$.

By substituting (5.3.1) and taking $\psi = \phi_l(\xi)$, (5.2.1) becomes

$$
\frac{d}{dt} W_l^j(t) = -\frac{2}{\Delta x} \left( -\int_{-1}^{1} f(\sum_{i=0}^{k} W_i^j(t)\phi_i(\xi))\phi'_l(\xi) d\xi + \hat{f}_{j+\frac{1}{2}}(1) - \hat{f}_{j-\frac{1}{2}}(-1) \right)
$$

(5.3.2)

for $1 \leq j \leq N$, where $\hat{f}_{j+\frac{1}{2}} = \hat{f}(\sum_{i=0}^{k} W_i^j(t)\phi_i(1), \sum_{i=0}^{k} W_i^{i+1}(t)\phi_i(-1))$ and the integrals need to be evaluated numerically, for example, using Gauss-Legendre quadrature. Combining (5.3.2) for $l = 0, \ldots, k$ over all cells, we obtain an ODE system

$$
\frac{d}{dt} Y = \mathcal{L}(Y),
$$

(5.3.3)

where $Y = ((W_0^0)^\top, \ldots, (W_k^0)^\top, \ldots, (W_0^k)^\top, \ldots, (W_k^k)^\top)$ and $\mathcal{L}(Y)$ is obtained correspondingly.

### 5.3.2 Time discretization

Recall the SSP RK3 method solving (5.3.3):

$$
Y^{(1)} = Y^n + \Delta t \mathcal{L}(Y^n),
$$

$$
Y^{(2)} = \frac{3}{4} Y^n + \frac{1}{4} Y^{(1)} + \frac{1}{4} \Delta t \mathcal{L}(Y^{(1)}),
$$

$$
Y^{n+1} = \frac{1}{3} Y^n + \frac{2}{3} Y^{(2)} + \frac{2}{3} \Delta t \mathcal{L}(Y^{(2)}),
$$

where $Y^n$ represents the solution of the ODE (5.3.3) at $n$-th time step and $Y^{(i)}$ is the approximation in the $i$-th intermediate stage, $i = 1, 2$. The time step size $\Delta t$ is determined such that the CFL condition given in (4.3.3) is satisfied. The initial data $Y^0$ is computed by taking the $L^2$ projection of $w_0$ onto $V_h$.

### 5.3.3 IRP limiter

The IRP limiter is applied after the initialization step and after each stage of the SSP RK3 method.
Consider the numerical solution at $n$-th time step is given by

$$w^n_h(x)|_{I_j} = \sum_{i=0}^{k} W_{j,i}^{i,n} \phi_i(\xi), \quad \xi \in [-1,1].$$

The cell average of $w^n_h(x)$ over $I_j$ is

$$\bar{w}^n_j = \frac{1}{2} \int_{-1}^{1} \sum_{i=0}^{k} W_{j,i}^{i,n} \phi_i(\xi) d\xi = \frac{1}{\sqrt{2}} \int_{-1}^{1} \sum_{i=0}^{k} W_{j,i}^{i,n} \phi_i(\xi) \phi_0(\xi) d\xi = \frac{1}{\sqrt{2}} W_{j,0}^{0,n} = W_{j,0}^{0,n} \phi_0(1).$$

Therefore, the modified solution can be written as

$$\tilde{w}^n_h(x)|_{I_j} = \theta w^n_h(x)|_{I_j} + (1 - \theta) \bar{w}^n_j = \theta \sum_{i=0}^{k} W_{j,i}^{i,n} \phi_i(\xi) + (1 - \theta) W_{j,0}^{0,n} \phi_0(\xi) = W_{j,0}^{0,n} + \theta \sum_{i=1}^{k} W_{j,i}^{i,n} \phi_i(\xi).$$

This indicates that in the implementation, we will keep $W_{j,0}^{0,n}$ unchanged and multiply $W_{j,i}^{i,n}$ by $\theta$ for $i = 1, \ldots, k$ to enforce the IRP property on the numerical solution at $n$-th time step.

Next, we compute the limiter parameter $\theta$ for $\tilde{w}^n_h(x)$ on cell $I_j$ by following the definition given in Section 4.2.1 and the algorithm in Section 4.2.2:

$$\theta = \min\{1, \theta_1, \ldots, \theta_M\}, \quad (5.3.4)$$

where

$$\theta_i = \frac{U_i(\tilde{w}^n_j)}{U_i(w^n_h(x)) - U_{i,j}^{\max}}, \quad U_{i,j}^{\max} = \max_{x \in S_K} U_i(w_h(x)|_{I_j}), \quad (5.3.5)$$

where the test set $S_K = x_j + \frac{\Delta x}{2} \{-1,1\}$ for $P^1$-DG and $S_j = x_j + \frac{\Delta x}{2} \{-1,0,1\}$ for $P^2$-DG and $P^3$-DG. Notice that $\theta$ is a local parameter which depends on the polynomial in the cell $I_j$. 
CHAPTER 6. SUMMARY AND DISCUSSION

In this thesis, we explored the IRP numerical methods solving hyperbolic systems of conservation laws. Through the study of model systems, the p-system and compressible Euler equations, we construct IRP limiters by taking advantage of the convexity and concavity of the functions characterizing the invariant regions. An important feature of the IRP limiter proved in this work is that it will not destroy the order of accuracy assuming that the cell average stays away from the boundary of the invariant region. Moreover, our limiter involves a single uniform scaling parameter, given in an explicit formula, for the whole vector solution polynomials. This is particularly convenient for numerical computations. A generic algorithm incorporating the IRP limiter is presented for high order finite volume type schemes. For any high order DG schemes to the model systems, sufficient conditions are obtained for cell averages to stay in the invariant region. We later extended the results to the general systems of hyperbolic conservation laws as long as (i) the system admits an invariant region that can be expressed by convex functions and (ii) the corresponding projected one-dimensional systems share the same invariant region as the full multidimensional hyperbolic system does. The general results are then applied to two dimensional compressible Euler equations. Numerical experiments on model systems show promising results that the IRP limiter remains the high order of accuracy for smooth solutions and help damp oscillations near discontinuities. The IRP DG schemes also capture the solution features (shock waves, low-density region, etc.) in different Riemann problems well.

In addition, we studied the viscous p-system since it shares the same invariant region with the p-system. Second and third order IRP DG schemes have been developed for this system. An interesting observation from the numerical experiments is that the IRP DG scheme solving the viscous p-system is more accurate than the one solving the p-system. One possible reason is that
the IRP limiter is called much more frequently in the latter system, which indicates a possible error accumulation through the evolution in time.
BIBLIOGRAPHY


differential equations by finite differences. *Communications on Pure and Applied Mathematics*,

of grundlehren der mathematischen wissenschaften [fundamental principles of mathematical
sciences].


on pure and applied Mathematics*, 7(2):345–392.


APPENDIX. A MATLAB SCRIPT

In this appendix, we list the matlab code of IRP DG scheme with SSP RK3 method for solving an accuracy test problem of the one-dimensional compressible Euler equations.

Listing .1 Main program

```matlab
% This code is to implement the IRP DG scheme with SSP RK3 method
% solving 1D Compressible Euler Equations

\frac{\partial w}{\partial t} + \nabla f(w) = 0

% where

w = (\rho, m, E)', \quad f(w) = (m, \rho u^2 + p, (E + p) u)'

\text{with}

u = m/\rho, \quad p = (\gamma - 1)(E - 0.5 m^2/\rho)

% Periodic boundary conditions

% domain: [a, b];
% N = # of cells;
% NumL = # of loops in space;
% deg: degree of polynomial: 1-3;
% s: switch of the limiter; 0: without limiter; 1: with IRP limiter;
% R0,M0,E0: coefficients regarding the basis function phi0 for density,
% momentum and energy function respectively;
% R1,M1,E1: coefficients regarding the basis function phi1 for density,
```
% momentum and energy function respectively;
% R2, M2, E2: coefficients regarding the basis function phi2 for density,
% momentum and energy function respectively;
% R3, M3, E3: coefficients regarding the basis function phi3 for density,
% momentum and energy function respectively;

clear all

global a b gamma xhat what rho0 pr0 q0 s0 Tolerance
a = 0;
b = 1;
gamma = 1.4;
Tolerance = 1e-13;

% Settings
NumL = 3;
tmax = 0.1;
deg = 1;

ER0r = zeros(NumL,1);
ER0m = zeros(NumL,1);
ER0E = zeros(NumL,1);
ER1r = zeros(NumL,1);
ER1m = zeros(NumL,1);
ER1E = zeros(NumL,1);
% Exact solution
rho = @(x,t) 1+0.5*sin(2*pi*(x-t));
u = @(x,t) x-x+t-t+1;
p = @(x,t) x-x+t-t+1;
m = @(x,t) rho(x,t).*u(x,t);
E = @(x,t) p(x,t)/(gamma-1)+0.5*(m(x,t).^2)/rho(x,t);

% Invariant Region: (x,y,z)-->(rho,m,E)
s = 1;
rh0 = Tolerance;
pr0 = Tolerance;
s0 = log(1/(1.5^gamma));
q0 = 0;

rh = @(x) x;
pr = @(x,y,z) (gamma-1)*(z-0.5*y.^2/x);
q = @(x,y,z) x.*(log(x.^gamma/pr(x,y,z))+s0);

% Initial condition
rho0 = @(x) rho(x,0);
u0 = @(x) u(x,0);
p0 = @(x) p(x,0);
m0 = @(x) m(x,0);
e0 = @(x) E(x,0);

% Gauss Lobatto pts and weights: xhat(1)=1; 2N-3= deg
[xhat,what,~] = lglnodes(deg);
% Gauss quadrature pts and weights
[xgq, wgq] = lgwt(5, -1, 1);

% Legendre Basis
phi0 = @(x) x - x + sqrt(1/2);
phi1 = @(x) sqrt(3/2) * x;
phi2 = @(x) sqrt(5/2) * (3 * x.^2 - 1/2);
phi3 = @(x) sqrt(7/2) * (5 * x.^3 - 3 * x/2);

for nl = 1:NumL
    N = 8 * 2^(nl - 1);
    dx = (b - a) / N;
    x = a : dx : b;

    % Index setting for periodic boundary condition
    lb = zeros(N, 1);
    for j = 2:N
        lb(j) = j - 1;
    end
    lb(1) = N;

    [R0, R1, R2, R3] = Initialization(deg, rho0, x, dx, N);
    [M0, M1, M2, M3] = Initialization(deg, m0, x, dx, N);
    [E0, E1, E2, E3] = Initialization(deg, e0, x, dx, N);

    A = [R0, M0, E0, R1, M1, E1, R2, M2, E2, R3, M3, E3];
if ( s == 1 )
    A = ReconVec(A);
endif

t = 0;
while t<tmax
    % Wavespeed calculation
    for j=1:N
        temp = lb(j);
        rhomi = [A(temp,1), A(temp,4), A(temp,7), A(temp,10)]*
            [phi0(1); phi1(1); phi2(1); phi3(1)];
        rhopl = [A(j,1), A(j,4), A(j,7), A(j,10)]*
            [phi0(-1); phi1(-1); phi2(-1); phi3(-1)];
        mmi = [A(temp,2), A(temp,5), A(temp,8), A(temp,11)]*
            [phi0(1); phi1(1); phi2(1); phi3(1)];
        mpl = [A(j,2), A(j,5), A(j,8), A(j,11)]*
            [phi0(-1); phi1(-1); phi2(-1); phi3(-1)];
        Emi = [A(temp,3), A(temp,6), A(temp,9), A(temp,12)]*
            [phi0(1); phi1(1); phi2(1); phi3(1)];
        Epl = [A(j,3), A(j,6), A(j,9), A(j,12)]*
            [phi0(-1); phi1(-1); phi2(-1); phi3(-1)];
        upl = mpl/rhopl;
        umi = mmi/rhomi;
        ppl = (gamma-1)*(Epl-0.5*mpl^2/rhopl);
        pmi = (gamma-1)*(Emi-0.5*mmi^2/rhomi);
        cpl = sqrt(gamma*ppl/rhopl);
        cmi = sqrt(gamma*pmi/rhomi);
rst = 0.5*d0*(rhomi+rhopl);
mst = 0.5*d0*(mmi+mpl);
Est = 0.5*d0*(Emi+Epl);
ust = mst/rst;
pst = pr(rst,mst,Est);
cst = sqrt(gamma*pst/rst);
ws1 = max(upl+cpl, ust+cst);
ws2 = min(umi–cmi, ust–cst);
sig = max(abs(ws1), abs(ws2));
end

if (deg ==1)
    dt = dx/(4*sig);
end

if (deg ==2)
    dt = dx/(12*sig);
end

if (deg ==3)
    dt = (dx^(5.0/3.0))/(24.0*sig);
end

if (tmax – t < dt)
    dt = tmax–t;
end

A = RK3(deg,A,dx,dt,s);

t = t + dt;
end

%%% Error and order calculation

%%% Inf norm

ERV0 = zeros(3,1);
EXV0 = zeros(3,1);
for j = 1:N
  for i = 1:5
    xi = xgq(i);
    fnw1 = A(j,1)*phi0(xi)+A(j,4)*phi1(xi)+A(j,7)*phi2(xi)+A(j,10)*phi3(xi);
    fnw2 = A(j,2)*phi0(xi)+A(j,5)*phi1(xi)+A(j,8)*phi2(xi)+A(j,11)*phi3(xi);
    fnw3 = A(j,3)*phi0(xi)+A(j,6)*phi1(xi)+A(j,9)*phi2(xi)+A(j,12)*phi3(xi);
    fw1 = rho(x(j)+0.5*dx+0.5*dx*xi,tmax);
    fw2 = m(x(j)+0.5*dx+0.5*dx*xi,tmax);
    fw3 = E(x(j)+0.5*dx+0.5*dx*xi,tmax);
    ERV0(1) = max(abs(fnw1-fw1),ERV0(1));
    ERV0(2) = max(abs(fnw2-fw2),ERV0(2));
    ERV0(3) = max(abs(fnw3-fw3),ERV0(3));
    EXV0(1) = max(abs(fw1),EXV0(1));
    EXV0(2) = max(abs(fw2),EXV0(2));
    EXV0(3) = max(abs(fw3),EXV0(3));
  end
end

ER0r(nl) = ERV0(1)/EXV0(1);
ER0m(nl) = ERV0(2)/EXV0(2);
ER0E(nl) = ERV0(3)/EXV0(3);
% $L1$ norm

ERV1 = zeros(3,1);
EXV1 = zeros(3,1);
for j=1:N
    for i=1:5
        xi = xgq(i);
        fnw1 = A(j,1) * phi0(xi) + A(j,4) * phi1(xi) + A(j,7) * phi2(xi) + A(j,10) * phi3(xi);
        fnw2 = A(j,2) * phi0(xi) + A(j,5) * phi1(xi) + A(j,8) * phi2(xi) + A(j,11) * phi3(xi);
        fnw3 = A(j,3) * phi0(xi) + A(j,6) * phi1(xi) + A(j,9) * phi2(xi) + A(j,12) * phi3(xi);
        fw1 = rho(x(j) + 0.5*dx + 0.5*dx*x_i, tmax);
        fw2 = m(x(j) + 0.5*dx + 0.5*dx*x_i, tmax);
        fw3 = E(x(j) + 0.5*dx + 0.5*dx*x_i, tmax);
        ERV1(1) = wgq(i) * abs(fnw1 - fw1) + ERV1(1);
        ERV1(2) = wgq(i) * abs(fnw2 - fw2) + ERV1(2);
        ERV1(3) = wgq(i) * abs(fnw3 - fw3) + ERV1(3);
        EXV1(1) = wgq(i) * abs(fw1) + EXV1(1);
        EXV1(2) = wgq(i) * abs(fw2) + EXV1(2);
        EXV1(3) = wgq(i) * abs(fw3) + EXV1(3);
    end
end

ER1r(nl) = ERV1(1)/EXV1(1);
ER1m(nl) = ERV1(2)/EXV1(2);
ER1E(nl) = ERV1(3)/EXV1(3);
Orp0r = zeros(NumL, 1);
Orp0m = zeros(NumL, 1);
Orp0E = zeros(NumL, 1);
Orp1r = zeros(NumL, 1);
Orp1m = zeros(NumL, 1);
Orp1E = zeros(NumL, 1);

for nl = 2:NumL
    Orp0r(nl) = log2(E0r(nl-1)/E0r(nl));
    Orp0m(nl) = log2(E0m(nl-1)/E0m(nl));
    Orp0E(nl) = log2(E0E(nl-1)/E0E(nl));
    Orp1r(nl) = log2(E1r(nl-1)/E1r(nl));
    Orp1m(nl) = log2(E1m(nl-1)/E1m(nl));
    Orp1E(nl) = log2(E1E(nl-1)/E1E(nl));
end

disp('L\_infinity \_norm')
fprintf('\%12.3e
', E0r)
fprintf('\%12.3e
', Orp0r)
disp('L\_1 \_norm')
fprintf('\%12.3e\n', E1r)
fprintf('\%12.3e\n', Orp1r)

Listing .2 Initialization
function [W0,W1,W2,W3] = Initialization(deg,f,x,h,N)

% Legendre Basis
phi0 = @(x) x-x+sqrt(1./2);
phi1 = @(x) sqrt(3./2)*x;
phi2 = @(x) sqrt(5./2)*(3*x.^2/2-1/2);
phi3 = @(x) sqrt(7/2)*(5*x.^3/2-3*x/2);

W2 = zeros(N,1);
W3 = zeros(N,1);

if (deg == 1)
    W0 = L2Proj(f,phi0,x,h,N);
    W1 = L2Proj(f,phi1,x,h,N);
else
    
    if (deg == 2)
        W0 = L2Proj(f, phi0, x, h, N);
        W1 = L2Proj(f, phi1, x, h, N);
        W2 = L2Proj(f, phi2, x, h, N);
    else
        W0 = L2Proj(f, phi0, x, h, N);
        W1 = L2Proj(f, phi1, x, h, N);
        W2 = L2Proj(f, phi2, x, h, N);
        W3 = L2Proj(f, phi3, x, h, N);
    end
end

end

Listing 3 L2 projection

% INPUT:
% f: function to be projected;
% g: basis function;
% x: grid points (vector);
% h: mesh size;
% N: number of cells;

% OUTPUT:
% F: Nx1 vector, the j-th component is the coefficient regarding the
function F=L2Proj(f,g,x,h,N)

F = zeros(N,1);
for j=1:N
    fun = @(y) f(x(j)+h/2+h*y/2).*g(y);
    F(j) = NumInt(fun,10);
end
end

Listing 4  SSP RK3

function C=RK3(deg,A,h,k,s)
\[ [N, \cdot] = \text{size}(A) \; ; \]
\[ B = \text{Loperator}(A, h) \; ; \]
\[ A_1 = \text{zeros}(N, 12) \; ; \]
\[ A_2 = \text{zeros}(N, 12) \; ; \]
\[ C = \text{zeros}(N, 12) \; ; \]

\[
\text{if} \ (\text{deg} == 1) \\
\quad A_1(:, 1:6) = A(:, 1:6) + k \cdot B(:, 1:6) \\
\text{else} \\
\quad \text{if} \ (\text{deg} == 2) \\
\quad \quad A_1(:, 1:9) = A(:, 1:9) + k \cdot B(:, 1:9) \\
\quad \text{else} \\
\quad \quad \text{if} \ (\text{deg} == 3) \\
\quad \quad \quad A_1 = A + k \cdot B \\
\quad \quad \end{end} \\
\quad \end{end} \\
\end{if} \\
\if \ (s == 1) \\
\quad A_1 = \text{ReconVec}(A_1) \\
\end{if} \\
B = \text{Loperator}(A_1, h) \\
\if \ (\text{deg} == 1) \\
\quad A_2(:, 1:6) = (3/4) \cdot A(:, 1:6) + (1/4) \cdot A_1(:, 1:6) + (1/4) \cdot k \cdot B(:, 1:6) \\
\text{else}
if (deg==2) 
    A2(:,1:9) = (3/4)*A(:,1:9) + (1/4)*A1(:,1:9) + (1/4)*k*B(:,1:9);
else 
    if (deg==3) 
        A2 = (3/4)*A + (1/4)*A1 + (1/4)*k*B;
    end 
end 
end

if (s==1) 
    A2 = ReconVec(A2); 
end 
B = Loperator(A2,h); 

if (deg==1) 
    C(:,1:6) = (1/3)*A(:,1:6) + (2/3)*A2(:,1:6) + (2/3)*k*B(:,1:6); 
else 
    if (deg==2) 
        C(:,1:9) = (1/3)*A(:,1:9) + (2/3)*A2(:,1:9) + (2/3)*k*B(:,1:9); 
    else 
        if (deg==3) 
            C = (1/3)*A + (2/3)*A2 + (2/3)*k*B; 
        end 
    end 
end 
end 

if (s==1)
C = ReconVec(C);
end

end

Listing 5  Right-hand side term calculation

function B = Loperator(A, h)

    global gamma

    fun1 = @(m) m;
    fun2 = @(rho, u, p) rho.*(u.^2)+p;
    fun3 = @(u, p, E) (E+p).*u;

    [N, ~] = size(A);
    B = zeros(N, 12);
    IR0 = zeros(N, 1);
    IR1 = zeros(N, 1);
    IR2 = zeros(N, 1);
    IR3 = zeros(N, 1);
    IM0 = zeros(N, 1);
IM1 = zeros(N,1);
IM2 = zeros(N,1);
IM3 = zeros(N,1);
IE0 = zeros(N,1);
IE1 = zeros(N,1);
IE2 = zeros(N,1);
IE3 = zeros(N,1);

% Legendre basis functions
phi0 = @(x) x-x+sqrt(1./2);
phi1 = @(x) sqrt(3./2)*x;
phi2 = @(x) sqrt(5./2)*(3*x.^2/2-1/2);
phi3 = @(x) sqrt(7/2)*(5*x.^3/2-3*x/2);

% Derivative of basis functions
dphi0 = @(x) 0;
dphi1 = @(x) sqrt(3/2);
dphi2 = @(x) sqrt(5/2)*(3*x);
dphi3 = @(x) sqrt(7/2)*(15*x.^2/2-3/2);

% Index for periodic boundary setting
rb = zeros(N,1);
for j=1:N-1
    rb(j) = j+1;
end
rb(N) = 1;
\% Integrate \( f(w) \ast d\phi \) over \([-1,1]\)

\[
\text{for } j=1:N \\
\quad \text{frho} = @(x) A(j,1) \ast \phi_0(x)+A(j,4) \ast \phi_1(x)+A(j,7) \ast \phi_2(x)+A(j,10) \ast \phi_3(x) ; \\
\quad \text{fm} = @(x) A(j,2) \ast \phi_0(x)+A(j,5) \ast \phi_1(x)+A(j,8) \ast \phi_2(x)+A(j,11) \ast \phi_3(x) ; \\
\quad \text{fE} = @(x) A(j,3) \ast \phi_0(x)+A(j,6) \ast \phi_1(x)+A(j,9) \ast \phi_2(x)+A(j,12) \ast \phi_3(x) ; \\
\quad \text{fu} = @(x) \text{fm}(x) \ast \frac{1}{\text{frho}(x)} ; \\
\quad \text{fp} = @(x) (\gamma -1) \ast (\text{fE}(x) - 0.5 \ast \text{fm}(x)^2 / \text{frho}(x)) ; \\
\quad \text{funR0} = @(x) \text{dphi0}(x) \ast \text{fun1}(\text{fm}(x)) ; \\
\quad \text{funR1} = @(x) \text{dphi1}(x) \ast \text{fun1}(\text{fm}(x)) ; \\
\quad \text{funR2} = @(x) \text{dphi2}(x) \ast \text{fun1}(\text{fm}(x)) ; \\
\quad \text{funR3} = @(x) \text{dphi3}(x) \ast \text{fun1}(\text{fm}(x)) ; \\
\quad \text{funM0} = @(x) \text{dphi0}(x) \ast \text{fun2}(\text{frho}(x), \text{fu}(x), \text{fp}(x)) ; \\
\quad \text{funM1} = @(x) \text{dphi1}(x) \ast \text{fun2}(\text{frho}(x), \text{fu}(x), \text{fp}(x)) ; \\
\quad \text{funM2} = @(x) \text{dphi2}(x) \ast \text{fun2}(\text{frho}(x), \text{fu}(x), \text{fp}(x)) ; \\
\quad \text{funM3} = @(x) \text{dphi3}(x) \ast \text{fun2}(\text{frho}(x), \text{fu}(x), \text{fp}(x)) ; \\
\quad \text{funE0} = @(x) \text{dphi0}(x) \ast \text{fun3}(\text{fu}(x), \text{fp}(x), \text{fE}(x)) ; \\
\quad \text{funE1} = @(x) \text{dphi1}(x) \ast \text{fun3}(\text{fu}(x), \text{fp}(x), \text{fE}(x)) ; \\
\quad \text{funE2} = @(x) \text{dphi2}(x) \ast \text{fun3}(\text{fu}(x), \text{fp}(x), \text{fE}(x)) ; \\
\quad \text{funE3} = @(x) \text{dphi3}(x) \ast \text{fun3}(\text{fu}(x), \text{fp}(x), \text{fE}(x)) ; \\
\quad \text{IR0}(j) = \text{NumInt}((\text{funR0}(x), 10)) ; \\
\quad \text{IR1}(j) = \text{NumInt}((\text{funR1}(x), 10)) ; \\
\quad \text{IR2}(j) = \text{NumInt}((\text{funR2}(x), 10)) ; \\
\quad \text{IR3}(j) = \text{NumInt}((\text{funR3}(x), 10)) ; \\
\quad \text{IM0}(j) = \text{NumInt}((\text{funM0}(x), 10)) ; \\
\quad \text{IM1}(j) = \text{NumInt}((\text{funM1}(x), 10)) ; \\
\quad \text{IM2}(j) = \text{NumInt}((\text{funM2}(x), 10)) ; \\
\quad \text{IM3}(j) = \text{NumInt}((\text{funM3}(x), 10)) ;
IE0(j) = NumInt(funE0,10);
IE1(j) = NumInt(funE1,10);
IE2(j) = NumInt(funE2,10);
IE3(j) = NumInt(funE3,10);
end

%% Flux: FRflux(j) is numerical flux at x(j), i.e. x_{j-1/2}
%% FRflux(1)=FRflux(N+1)
[FRflux,FMflux,FEflux]=Flux(A);

for j=1:N
    temp = rb(j);
    B(j,1) = (-2/h)*(-IR0(j)+phi0(1)*FRflux(temp)-phi0(-1)*FRflux(j));
    B(j,4) = (-2/h)*(-IR1(j)+phi1(1)*FRflux(temp)-phi1(-1)*FRflux(j));
    B(j,7) = (-2/h)*(-IR2(j)+phi2(1)*FRflux(temp)-phi2(-1)*FRflux(j));
    B(j,10) = (-2/h)*(-IR3(j)+phi3(1)*FRflux(temp)-phi3(-1)*FRflux(j));
    B(j,2) = (-2/h)*(-IM0(j)+phi0(1)*FMflux(temp)-phi0(-1)*FMflux(j));
    B(j,5) = (-2/h)*(-IM1(j)+phi1(1)*FMflux(temp)-phi1(-1)*FMflux(j));
    B(j,8) = (-2/h)*(-IM2(j)+phi2(1)*FMflux(temp)-phi2(-1)*FMflux(j));
    B(j,11) = (-2/h)*(-IM3(j)+phi3(1)*FMflux(temp)-phi3(-1)*FMflux(j));
    B(j,3) = (-2/h)*(-IE0(j)+phi0(1)*FEflux(temp)-phi0(-1)*FEflux(j));
    B(j,6) = (-2/h)*(-IE1(j)+phi1(1)*FEflux(temp)-phi1(-1)*FEflux(j));
    B(j,9) = (-2/h)*(-IE2(j)+phi2(1)*FEflux(temp)-phi2(-1)*FEflux(j));
    B(j,12) = (-2/h)*(-IE3(j)+phi3(1)*FEflux(temp)-phi3(-1)*FEflux(j));
end
end
Listing 6  Numerical flux

function [FRflux, FMflux, FEflux] = Flux(A)

    % INPUT:
    % A: Nx12 matrix of coefficients of numerical solution;
    % OUTPUT:
    % FRflux: Nx1 vector, the j-th component is the numerical flux at the %
    % left endpoint of the j-th cell, corresponding to the 1st equation;  %
    % FMflux: Nx1 vector, the j-th component is the numerical flux at the %
    % left endpoint of the j-th cell, corresponding to the 2nd equation;  %
    % FEflux: Nx1 vector, the j-th component is the numerical flux at the %
    % left endpoint of the j-th cell, corresponding to the 3rd equation;  %

    global gamma
    fun1 = @(m) m;
    fun2 = @(rho,u,p) rho.*(u.^2)+p;
    fun3 = @(u,p,E) (E+p).*u;
    pr = @(x,y,z) (gamma-1)*(z-0.5*y.^2/x);
    [N,~] = size(A);
    FRflux = zeros(N,1);
    FMflux = zeros(N,1);
    FEflux = zeros(N,1);

    % Legendre Basis
    phi0 = @(x) x-x+sqrt(1./2);
\phi_1(x) = \sqrt{\frac{3}{2}}x;
\phi_2(x) = \sqrt{\frac{5}{2}}(3x^2/2 - 1/2);
\phi_3(x) = \sqrt{\frac{7}{2}}(5x^3/2 - 3x/2);

\% Index for periodic boundary setting
lb = zeros(N,1);
for j = 2:N
    lb(j) = j - 1;
end
lb(1) = N;

for j = 1:N
    temp = lb(j);
    rhomi = [A(temp,1), A(temp,4), A(temp,7), A(temp,10)]*[\phi_0(1); phi_1(1); phi_2(1); phi_3(1)];
    rhopl = [A(j,1), A(j,4), A(j,7), A(j,10)]*[\phi_0(-1); phi_1(-1); phi_2(-1); phi_3(-1)];
    mmi = [A(temp,2), A(temp,5), A(temp,8), A(temp,11)]*[\phi_0(1); phi_1(1); phi_2(1); phi_3(1)];
    mpl = [A(j,2), A(j,5), A(j,8), A(j,11)]*[\phi_0(-1); phi_1(-1); phi_2(-1); phi_3(-1)];
    Emi = [A(temp,3), A(temp,6), A(temp,9), A(temp,12)]*[
            \phi_0(1); phi_1(1); phi_2(1); phi_3(1)];
    Epl = [A(j,3), A(j,6), A(j,9), A(j,12)]*[\phi_0(-1); phi_1(-1); phi_2(-1); phi_3(-1)];
    upl = mpl/rhopl;
    umi = mmi/rhomi;
\[ ppl = (\gamma - 1)(Epl - 0.5 \times mpl^2 / \rho_{pl}); \]
\[ pmi = (\gamma - 1)(Emi - 0.5 \times mmi^2 / \rho_{mi}); \]
\[ cpl = \sqrt{\gamma \times ppl / \rho_{pl}}; \]
\[ cmi = \sqrt{\gamma \times pmi / \rho_{mi}}; \]
\[ rst = 0.5 d0 \times (\rho_{mi} + \rho_{pl}); \]
\[ mst = 0.5 d0 \times (mmi + mpl); \]
\[ Est = 0.5 d0 \times (Emi + Epl); \]
\[ ust = mst / rst; \]
\[ pst = pr(rst, mst, Est); \]
\[ cst = \sqrt{\gamma \times pst / rst}; \]
\[ ws1 = \max(upl + cpl, ust + cst); \]
\[ ws2 = \min(umi - cmi, ust - cst); \]
\[ \sigma = \max(abs(ws1), abs(ws2)); \]
\[ FRflux(j) = (\text{fun1}(mpl) + \text{fun1}(mmi) - \sigma \times (\rho_{pl} - \rho_{mi}))/2; \]
\[ FMflux(j) = (\text{fun2}(\rho_{pl}, upl, ppl) + \text{fun2}(\rho_{mi}, umi, pmi) - \sigma \times (mpl - mmi))/2; \]
\[ FEflux(j) = (\text{fun3}(upl, ppl, Epl) + \text{fun3}(umi, pmi, Emi) - \sigma \times (Epl - Emi))/2; \]

end

end

Listing .7 Polynomial modification

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
% INPUT:
% %
% A: matrix of coefficients of numerical solution;
% %
% OUTPUT:
% %
% B: matrix of coefficients of modified numerical solution;
% %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function B=ReconVec(A)

  global xhat rh0 pr0 q0 s0 gamma
  [N,~] = size(A);
  B = zeros(N,12);
  theta = zeros(N,1);

  % Lengendre Basis
  phi0 = @(x) x-x+sqrt(1./2);
  phi1 = @(x) sqrt(3./2)*x;
  phi2 = @(x) sqrt(5./2)*(3*x.^2/2-1/2);
  phi3 = @(x) sqrt(7/2)*(5*x.^3/2-3*x./2);

  % Invariant region functions: (x,y,z)->(rho,m,E)
  rh = @(x) x;
  pr = @(x,y,z) (gamma-1)*(z-0.5*y.^2/x);
  q = @(x,y,z) x.*(log(x.^gamma/pr(x,y,z))+s0);

  rhat = zeros(length(xhat),1);
  mhat = zeros(length(xhat),1);
  Ehat = zeros(length(xhat),1);
  phat = zeros(length(xhat),1);
  qhat = zeros(length(xhat),1);

  for j=1:N
    fr = @(x) A(j,1)*phi0(x)+A(j,4)*phi1(x)+A(j,7)*phi2(x)+A(j,10)*phi3(x);
    fm = @(x) A(j,2)*phi0(x)+A(j,5)*phi1(x)+A(j,8)*phi2(x)+A(j,11)*phi3(x);
\[ f_E = @(x) A(j,3) \phi_0(x) + A(j,6) \phi_1(x) + A(j,9) \phi_2(x) + A(j,12) \phi_3(x); \]
\[ f_p = @(x) (\text{gamma} - 1)(f_E(x) - 0.5 \times f_m(x)^2 / f_r(x)); \]

**for** alpha = 1:length(xhat)

\[ rhat(\alpha) = f_r(xhat(\alpha)); \]
\[ mhat(\alpha) = f_m(xhat(\alpha)); \]
\[ Ehat(\alpha) = f_E(xhat(\alpha)); \]
\[ phat(\alpha) = f_p(xhat(\alpha)); \]
\[ qhat(\alpha) = q(rhat(\alpha), mhat(\alpha), Ehat(\alpha)); \]

**end**

\[ rmin = \text{min}(rhat); \]
\[ pmin = \text{min}(phat); \]
\[ qmax = \text{max}(qhat); \]
\[ rbar = r(A(j,1)/\text{sqrt}(2)); \]
\[ pbar = p(A(j,1)/\text{sqrt}(2), A(j,2)/\text{sqrt}(2), A(j,3)/\text{sqrt}(2)); \]
\[ qbar = q(A(j,1)/\text{sqrt}(2), A(j,2)/\text{sqrt}(2), A(j,3)/\text{sqrt}(2)); \]
\[ \theta(j) = \text{ReconPara}(rh0, pr0, q0, rmin, pmin, qmax, rbar, pbar, qbar); \]

\[ B(j,1) = A(j,1); \]
\[ B(j,2) = A(j,2); \]
\[ B(j,3) = A(j,3); \]
\[ B(j,4) = \theta(j) \times A(j,4); \]
\[ B(j,5) = \theta(j) \times A(j,5); \]
\[ B(j,6) = \theta(j) \times A(j,6); \]
\[ B(j,7) = \theta(j) \times A(j,7); \]
\[ B(j,8) = \theta(j) \times A(j,8); \]
\[ B(j,9) = \theta(j) \times A(j,9); \]
\[ B(j,10) = \theta(j) \times A(j,10); \]
\[ B(j,11) = \theta(j) \times A(j,11); \]
\[ B(j,12) = \theta(j) \times A(j,12); \]

end

end

Listing 8 Limit parameter calculation

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% INPUT:
% r0 : lower bound of the density function;
% p0 : lower bound of the pressure function;
% q0 : upper bound of the entropy function;
% rmin: minimum of the numerical density over the test set;
% pmin: minimum of the numerical pressure over the test set;
% qmax: maximum of the numerical entropy over the test set;
% rbar: density evaluated using the cell average;
% pbar: pressure evaluated using the cell average;
% qbar: entropy evaluated using the cell average;
% OUTPUT:
% y: the parameter used in the IRP limiter;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function y=ReconPara(r0,p0,q0,rmin,pmin,qmax,rbar,pbar,qbar)

    y1 = (rbar-r0)/(rbar-rmin);
    y2 = (pbar-p0)/(pbar-pmin);
    y3 = (q0-qbar)/(qmax-qbar);
    y = min(abs(y1),abs(y2));
    y = min(y,abs(y3));
\begin{verbatim}
y = \min(y, 1);
end
\end{verbatim}