A ROBUST SECOND ORDER INVARIANT-DOMAIN PRESERVING APPROXIMATION OF THE COMPRESSIBLE EULER EQUATIONS WITH AN ARBITRARY EQUATION OF STATE

A Thesis

by

BENNETT GILES CLAYTON

Submitted to the Office of Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Bojan Popov
Committee Members,	Jean-Luc Guermond
	Matthias Maier
	Jean Ragusa
Head of Department,	Sarah Witherspoon

May 2023

Major Subject: Mathematics

ABSTRACT

For many physical problems, robust numerical methods for solving the compressible Euler equations are essential. For the Euler equations to accurately describe the fluid behavior a suitable equation of state (EOS), which describes the relationship between the thermodynamic variables, must be chosen. However, a robust numerical method which can handle an arbitrary equation of state has been unavailable.

In this thesis, we present a second order invariant-domain preserving method for the compressible Euler equations with an arbitrary equation of state. The description of the second order method first requires the development of a first order method that preserves certain thermodynamic properties of the fluid. A method which preserves this physical aspect is referred to as an invariant-domain preserving method. The fundamental methodology of the first order method relies on estimating the maximum wave speed of local Riemann problems. For an arbitrary equation of state this estimation can be impossible. We circumvent the issue by extending the system with an interpolatory EOS, and rigorously justify that the use of the max wave speed of this extended problem implies the invariant-domain preserving properties of the method.

Using a higher order graph viscosity cannot guarantee the invariant-domain preserving property. We resolve this issue through the use of quasiconcave limiting on the density and a surrogate entropy. For an arbitrary equation of state, access to the entropy may not be possible, therefore, this surrogate entropy is used to guarantee that a local approximation of the entropy will increase across shock waves. Furthermore, limiting on the surrogate entropy guarantees that the specific internal energy satisfies the invariant domain constraint.

DEDICATION

To my wife, Unkyung, and daughter, Alice Sena.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Bojan Popov, as well as my committee, Dr. Jean-Luc Guermond, Dr. Matthias Maier, and Dr. Jean Ragusa, for helping me throughout this arduous journey. I would not have been able to achieve this work without their support and guidance. I would also like to thank Texas A&M University for giving me the opportunity to pursue this degree.

I would also like to thank my friends and colleagues: Dr. Eric Tovar and Dr. Weston Baines for providing insightful discussion, practicing presentations, and all around general help for a variety of issues.

Lastly, I thank my wife for surviving with me and supporting me along the way. I may never have even tried to enter into the PhD program without her.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Bojan Popov (advisor), Professor Jean-Luc Guermond, and Professor Matthias Maier, of the Mathematics Department and Professor Jean Ragusa of the Department of Nuclear Engineering. In addition, part of the work was done in collaboration with Dr. Eric Tovar of Los Alamos National Lab. All other work conducted for the thesis was completed by the student independently.

Funding Sources

The graduate study was supported under the following grants: The National Science Foundation DMS-2110868, The Air Force Office of Scientific Research, USAF, under the grant/contract number FA9550-18-1-0397, The Army Research Office under grant number W911NF-19-1-0431, and the U.S. Department of Energy by Lawrence Livermore National Laboratory under the contracts, B640889 and B641173.

NOMENCLATURE

DoF(s)	Degree(s) of Freedom
EOS	Equation of State
FEM	Finite Element Method
JWL	Jones-Wilkins-Lee
NASG	Noble-Abel Stiffened Gas
TAMU	Texas A&M University
vdW	van der Waals

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. HYPERBOLIC CONSERVATION LAWS

1.1 Introduction

The partial differential equations (PDEs) known as "conservation laws" are used to model a wide variety of physical phenomena. For example, the shallow water models can be used to predict evacuation zones for a storm surge brought about by a hurricane or due to a tsunami. One can also predict and categorize flood zones for which insurance companies and home-buyers use to assess their respective risk. Another active area of research is in compressible flow. Engineers and scientists are researching supersonic aircraft designs which mitigate the sonic boom effect that happens when the aircraft breaks the sound speed barrier. This research is carried out by numerically solving the compressible Euler or Navier-Stokes equations. Furthermore, the study and design of hypersonic objects is now of increasing importance for both national defense and atmospheric reentry vehicles.

These motivations play a major role in the development of robust numerical methods. These robust methods must maintain several important properties. One, the numerical method should preserve the physical properties of quantity or material being studied. For example, preservation of positive water height in the shallow water equations, positive specific internal energy in compressible Euler equations, and so on. Two, the numerical method should produce a "physically relevant" solution. It is known that there are infinitely many weak solutions to a scalar conservation law. In the case of the compressible Euler equations, the numerical method should be able to exclude non-physical solutions through the enforcement of the minimum principle on the specific entropy. Three, the method should not have any tunable parameters that must be adjusted depending on the initial data or computational domain. Four, the method should be scalable. Real applications often require an extremely fine grid to accurately model the object being studied, or operate over a very large spatial domain as in the case of modeling hurricanes. Therefore, it is necessary to simulate

these experiments on a supercomputer. Hence, scalability implies that the numerical method can be programmed efficiently in parallel. Five, the numerical method should be high order. That is, the method converges to the solution rapidly as the mesh size decreases.

Numerical methods for the compressible Euler equations go all the way back to von Neumann [3] and Lax [4] in the early 1950s with the use of finite difference schemes in one dimension. In 1960, Lax and Wendroff [5] introduced a second order method for systems of conservation laws with artificial viscosity to stabilize the method. Also in 1959, Godunov published a new finite difference method, (which could be regarded as a finite volume method), for hydrodynamics. It uses the solution to local Riemann problems at the interface between cells to define the flux on these interfaces. As the original paper was published in Russian in Matematicheskii Sbornik; an English translation provided in [6]. In 1969, a novel second order accurate method for hydrodynamics was published by MacCormack [7] (the citation provided is a reprint of the original paper). This method is often referred to as the "predictor-corrector" method. The method works by first computing an intermediate solution using a first order forward-differencing method. This solution is then used to recompute the flux at the interfaces. The final solution is then computed using a backward differencing method but with the new fluxes. This results in a second order method. In 1981, a new approach was developed by Steger and Warming [8] which involves splitting the flux vector into two components based on the positivity or negativity of the eigenvalues of the Jacobian matrix. This method is referred to as the flux-vector splitting method, see also, van Leer [9]. Around the same time, an approximate Riemann solver was used to compute the flux on the cell interface by Roe [10]. Further advances in the use of approximate Riemann solvers appear in the HLL scheme [11], and the HLLC scheme [12]. Since then many new ideas using flux-vector splitting and approximate Riemann solvers were developed like the AUSM scheme by Liou; see [13]. We would also like to point out the popular ENO [14] and WENO [15] schemes which work to minimize oscillations near discontinuities. Lastly, a new approach to numerical methods for conservation laws, referred to as, invariant-domain

preserving methods, began in 2016 with Guermond & Popov [16].

The specific focus of this thesis is on the development of a second order accurate invariant-domain preserving method for the compressible Euler equations with an arbitrary or tabulated equation of state (EOS). That is, the method will be shown to always preserve the necessary thermodynamic properties enforced by the equation of state. For example, positive density and positive specific internal energy. We do this by discretizing the compressible Euler equations with $\mathbb{P}_1$ (or $\mathbb{Q}_1$) finite elements and also introduce a "graph viscosity" to the equations which is essential for invariant-domain preserving properties. The numerical update is then written as a convex combination of some "states". These states motivate the definition of the graph viscosity which is determined by the maximum wave speed to local Riemann problems. It is the determination of this maximum wave speed that is the fundamental difficulty when the equation of state provided is tabulated. There are several numerical methods (e.g. Godunov-type methods) which require the solution to the Riemann problem. Several papers have addressed the issue of the solution to the Riemann problem, but only if the EOS is defined analytically (not tabulated), see [17, Sec. 1], [18], and [19]. Another approach for handling an arbitrary EOS is to use an approximate Riemann solver. See the work done in, Dukowicz [20], Roe & Pike [21], Pike [22], and Lee et. al. [23]. However, these methods make no guarantee on preserving invariant-domain properties other than positivity of the density.

Once the maximum wave speeds are determined, we are able to justify that the states belong to a so-called, "invariant set". This invariant set is defined by quasiconcave functionals which express the physical quantities we are trying to preserve. This invariant set is slightly different but inspired from the notion of an invariant region described in Chueh et. al. [24], Hoff [25], and Frid [26]. This invariant set is convex and hence a convex combination of states inside a convex set imply that the updated state also belongs to that set. Thus we guarantee that the physical properties we desire, are held.

The extension to second order is done by defining a so-called "entropy viscosity" which is

small in the regions of smooth flow and large when there are discontinuities. The motivating factor for this viscosity measures the change in entropy. I.e., no change in entropy indicates that the flow is smooth and hence we can take the graph viscosity to be small. This second order method by itself is not invariant-domain preserving and a local "limiting" process based on local bounds is required to preserve the physical properties. Limiting is performed on the density and an entropy. However, for a given EOS, we may not know the specific entropy or the specific entropy may not be a concave function of the specific volume and the specific internal energy. Therefore, we suggest the use of a surrogate entropy which behaves similar to a physical entropy; in that, the surrogate entropy increases across shocks.

In Chapter 2 we go over some of the necessary thermodynamic properties for describing the equation of state as well as, a description of several EOS which we use in our numerical demonstrations. In Chapter 3 we give a brief survey of the finite element method of which we employ in the discretization of the problem. Note, however, that the numerical method described in Chapters 4, 5, and 6 is discretization independent. That is, one could equivalently describe the problem in the finite volume or finite difference context; see Remark 4.0.1. In Chapter 4, we derive the first order invariant-domain preserving method for an arbitrary or tabulated equation of state. The estimation of the maximum wave speed for local Riemann problems is detailed here. Then in Chapter 5 we extend the first order method to second order by using the consistent mass matrix and a smaller graph viscosity. It is known that the use of a graph viscosity which is too small can lead to instabilities in multiple forms. This issue is addressed with the use of quasiconcave limiting described in Chapter 6. Lastly, numerical results are presented in Chapter 7.

1.2 Review of the Theory of Conservation Laws

We begin by reviewing some relevant facts regarding the theory of hyperbolic conservation laws. We are interested in the partial differential equation (PDE),

$$\partial_t \boldsymbol{u} + \nabla \cdot \mathrm{g}(\boldsymbol{u}) = \mathbf{0}, \qquad \text{for } \boldsymbol{x} \in \mathbb{R}^d,\ t > 0, \tag{1.1a}$$

$$\boldsymbol{u}(\boldsymbol{x}, 0) = \boldsymbol{u}_0(\boldsymbol{x}) \in \mathcal{B} \subset \mathbb{R}^m, \tag{1.1b}$$

where d denotes the spatial dimension and $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}, t) = (u_1(\boldsymbol{x}, t), \ldots, u_m(\boldsymbol{x}, t))^\mathsf{T}$ is the unknown vector of conserved quantities with $^\mathsf{T}$ denoting the transpose, and $\mathrm{g} \in C^2(\mathbb{R}^m; \mathbb{R}^{m\times d})$ is the flux. In particular, $\mathrm{g}(\boldsymbol{u}) = (\boldsymbol{g}_1(\boldsymbol{u}), \ldots, \boldsymbol{g}_d(\boldsymbol{u}))$, where $\boldsymbol{g}_k : \mathcal{B} \subset \mathbb{R}^m \to \mathbb{R}^m$ for $k \in \{1 : d\}$ and $\boldsymbol{g}_k(\boldsymbol{u}) = (g_{k1}(\boldsymbol{u}), \ldots, g_{km}(\boldsymbol{u}))^\mathsf{T}$. The set $\mathcal{B}$ is some subset of the phase space and will be explored more in Chapter 4.

The form (1.1) is referred to as the conservative form and can be equivalently written in the quasi-linear form,

$$\partial_t \boldsymbol{u} + \sum_{k=1}^{d} \mathbb{A}_k(\boldsymbol{u}) \partial_{x_k} \boldsymbol{u} = \mathbf{0}, \qquad \text{for } \boldsymbol{x} \in \mathbb{R}^d,\ t > 0. \tag{1.2}$$

where $\mathbb{A}_k(\boldsymbol{u}) = \left(\frac{\partial g_{ki}}{\partial u_j}(\boldsymbol{u})\right)_{1\le i,j\le m}$.

Definition 1.2.1 (Hyperbolic System)**.** Define $\mathbb{A}(\boldsymbol{u}; \boldsymbol{n}) := \sum_{k=1}^{d} \mathbb{A}_k(\boldsymbol{u}) n_k$, where $\boldsymbol{n} := (n_1, \ldots, n_d)^\mathsf{T}$ and $\|\boldsymbol{n}\|_{\ell^d} = 1$. We say that (1.1) is *hyperbolic*, if $\mathbb{A}(\boldsymbol{u}; \boldsymbol{n})$ has m real eigenvalues $\lambda_1(\boldsymbol{u}; \boldsymbol{n}), \cdots, \lambda_m(\boldsymbol{u}; \boldsymbol{n})$, for all $\boldsymbol{u} \in \mathcal{B}$ and $\boldsymbol{n} \in \mathbb{S}^{d-1}$. Furthermore, we say that (1.1) is *strictly hyperbolic* if the eigenvalues are all distinct. □

It is a well known fact that solutions to hyperbolic conservation laws can develop discontinuities in finite time, even if the initial data is smooth. The goal instead, is to find **weak** solutions to (1.1).

Definition 1.2.2 (Weak Solutions)**.** A function, $\boldsymbol{u} \in [L^\infty(\mathbb{R}^d \times (0, \infty))]^m$ is said to be a

weak solution to (1.1) if

$$\int_{\mathbb{R}^d}\int_0^\infty \boldsymbol{u}\partial_t\varphi + \mathrm{g}(\boldsymbol{u})\nabla\varphi \,\mathrm{d}t\,\mathrm{d}\boldsymbol{x} = -\int_{\mathbb{R}^d} \boldsymbol{u}_0(\boldsymbol{x})\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}, \tag{1.3}$$

holds for all $\varphi \in C_c^1(\mathbb{R}^d \times [0,\infty))$, with C_c^1 denoting compactly supported C^1 functions. □

For a system of strictly hyperbolic conservation laws in one dimension, under the assumption that all of the characteristic fields are either genuinely nonlinear or linearly degenerate (see Definition 1.2.5), Glimm [27], proved the existence of weak solutions utilizing a random choice method.

In the coming numerical method, we seek physically relevant weak solutions; that is, solutions in the vanishing viscosity sense. This vanishing viscosity solutions will satisfy the so-called, entropy inequalities, see [28, Chapter 11, Theorem 2].

Definition 1.2.3 (Entropy Solutions)**.** A weak solution $\boldsymbol{u} \in [L^\infty(\mathbb{R}^d \times [0,\infty))]^{d+2}$ to (1.1) is said to be an entropy solution if,

$$\int_{\mathbb{R}^d}\int_0^\infty \eta(\boldsymbol{u})\partial_t\varphi(\boldsymbol{x},t) + \boldsymbol{F}(\boldsymbol{u})\cdot\nabla\varphi(\boldsymbol{x},t)\,\mathrm{d}t\,\mathrm{d}\boldsymbol{x} \geq 0 \tag{1.4}$$

for all $\varphi \in C_c^1(\mathbb{R}\times[0,\infty))$ and all entropy, entropy-flux pairs $(\eta, \boldsymbol{F})$ with η convex. In short, we say that $\partial_t\eta(\boldsymbol{u}) + \nabla\cdot\boldsymbol{F}(\boldsymbol{u}) \leq 0$ in the *weak sense*. □

It is still an open problem on whether the entropy solution will be unique for a hyperbolic system of conservation laws. However, there are results for special cases. For scalar conservation laws, the weak solution only needs to satisfy the so-called Kruzhkov entropies, to ensure uniqueness of the solution, see Kruzhkov [29]. For the Cauchy problem with viscosity, $\partial_t\boldsymbol{u} + \mathbb{A}(\boldsymbol{u})\partial_x\boldsymbol{u} = \varepsilon\partial_{xx}\boldsymbol{u}$, if $\mathbb{A}(\boldsymbol{u})$ is strictly hyperbolic and if the initial data is small enough; that is, $\|\boldsymbol{u}\|_{\mathrm{BV}} < \delta$ for some $\delta > 0$, then there exists a unique solution $\boldsymbol{u}^\varepsilon(x,t)$. This solution, $\boldsymbol{u}^\varepsilon$ is referred to as the viscosity solution. For $\varepsilon \to 0^+$, we have that $\boldsymbol{u}^\varepsilon$ converges to the unique solution to $\partial_t\boldsymbol{u} + \mathbb{A}(\boldsymbol{u})\partial_x\boldsymbol{u} = \boldsymbol{0}$. This solution is referred to as the *vanishing*

viscosity solution. For the general system $\partial_t \boldsymbol{u} + \partial_x \mathrm{g}(\boldsymbol{u}) = \mathbf{0}$, the vanishing viscosity solution is an entropy solution. This solution was shown in Bianchini & Bressan [30] to be the same solution derived by Glimm in [27]. For more on the theory of vanishing viscosity solutions, see [30]. Despite the lack of results for systems in higher dimensions, entropy solutions are still sought after since they exclude many non-physical solutions.

One of the main motivations of our numerical method, the invariant-domain preserving method, is that the solution should reside in region of the phase space for which the physical properties are satisfied. This motivates the definition of an *invariant region*.

Definition 1.2.4 (Invariant Region)**.** A set $\mathcal{B} \subset \mathbb{R}^m$ is said to be an invariant region for the PDE (1.1) if $\boldsymbol{u}(\boldsymbol{x}, 0) \in \mathcal{B}$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and there exists $\delta > 0$ such that a solution $\boldsymbol{u}(\boldsymbol{x}, t)$ exists for all $0 < t < \delta$, and if $\boldsymbol{u}(\boldsymbol{x}, t) \in \mathcal{B}$, then we say that $\mathcal{B}$ is an invariant region. □

Remark 1.2.1. This set $\mathcal{B}$ will be defined in terms of some quasiconcave functionals. That is, let $\{\Psi_i\}_{i=1}^N$ be a collection of quasiconcave functionals, then

$$\mathcal{B} := \{\boldsymbol{u} \in \mathbb{R}^m : \Psi_i(\boldsymbol{u}) > 0,\ i = 1, \ldots, N\}, \tag{1.5}$$

is an invariant region. This definition was originally motivated by Chueh et. al. [24] in the context of convection-diffusion problems and the functionals, $\{\Psi_i\}$ are chosen to be convex, with the constraints $\Psi_i(\boldsymbol{u}) \leq 0$. □

As much of the upcoming material (see Chapter 4) requires analysis of the Riemann problem which is one-dimensional, we switch our focus to conservation laws in one-dimension.

Consider the one-dimensional problem,

$$\partial_t \boldsymbol{u} + \partial_x(\boldsymbol{g}(\boldsymbol{u})) = \mathbf{0}, \qquad \text{for } x \in \mathbb{R},\ t > 0. \tag{1.6}$$

Let $\boldsymbol{A}(\boldsymbol{u})$ be the Jacobian matrix for the flux $\boldsymbol{g}(\boldsymbol{u})$. Assume the eigenvalues of $\boldsymbol{A}(\boldsymbol{u})$ are all real and distinct, $\lambda_1(\boldsymbol{u}) < \cdots < \lambda_m(\boldsymbol{u})$. Let $\boldsymbol{r}_1(\boldsymbol{u}), \ldots, \boldsymbol{r}_m(\boldsymbol{u})$ be the associated right

eigenvectors. Then we have the following definitions,

Definition 1.2.5 (Wave Type Definitions)**.** For the one-dimensional conservation law, (1.6), we have the following definitions,

- We call $(\lambda_i(\boldsymbol{u}), \boldsymbol{r}_i(\boldsymbol{u}))$ the *i-characteristic field.*
- The i-characteristic field is said to be *genuinely nonlinear* if $D\lambda_i(\boldsymbol{u}) \cdot \boldsymbol{r}_i(\boldsymbol{u}) \neq 0$ for all $\boldsymbol{u} \in \mathcal{B}$.
- The i-characteristic field is said to be *linearly degenerate* if, $D\lambda_i(\boldsymbol{u}) \cdot \boldsymbol{r}_i(\boldsymbol{u}) = 0$ for all $\boldsymbol{u} \in \mathcal{B}$.

□

Definition 1.2.6 (k-Riemann invariant)**.** A smooth function $w : \mathcal{B} \to \mathbb{R}$ is said to be a k-Riemann invariant, if $Dw(\boldsymbol{u}) \cdot \boldsymbol{r}_k(\boldsymbol{u}) = 0$ for all $\boldsymbol{u} \in \mathcal{B}$. □

Definition 1.2.7 (Rankine-Hugoniot Conditions)**.** The Rankine-Hugoniot conditions are defined as,

$$S(\boldsymbol{u}_{\mathrm{L}} - \boldsymbol{u}_{\mathrm{R}}) = \boldsymbol{g}(\boldsymbol{u}_{\mathrm{L}}) - \boldsymbol{g}(\boldsymbol{u}_{\mathrm{R}}), \tag{1.7}$$

where S is the instantaneous speed of the discontinuity and $\boldsymbol{u}_{\mathrm{L}}$ and $\boldsymbol{u}_{\mathrm{R}}$ are the instantaneous states to the left and right of the discontinuity. □

1.3 The Riemann Problem*

A large number of numerical methods for conservation laws require, in some way, the concept of a Riemann problem.

Definition 1.3.1. Let $\boldsymbol{g} \in C^1(\mathbb{R}^m; \mathbb{R}^m)$ be the flux, then the Riemann problem is written

* Lemma 1.3.1 and its proof are taken from [2] and are reprinted with permission from [2].

as: Find a self-similar (weak) solution $\boldsymbol{u} \in L^\infty(\mathbb{R}\times(0,\infty);\mathbb{R}^m)\cap C^0((0,\infty);L^1_{\text{loc}}(\mathbb{R};\mathbb{R}^m))$ to

$$\partial_t \boldsymbol{u} + \partial_x \boldsymbol{g}(\boldsymbol{u}) = \boldsymbol{0}, \tag{1.8a}$$

$$\boldsymbol{u}(x,0) = \begin{cases} \boldsymbol{u}_L, & x < 0, \\ \boldsymbol{u}_R, & x > 0, \end{cases} \tag{1.8b}$$

□

We then have a fundamental result for the existence and uniqueness of solutions to the Riemann problem,

Theorem 1.3.1 (Existence and Uniqueness)**.** *Assume the system* (1.8a) *is strictly hyperbolic and the characteristic fields are either genuinely nonlinear or linearly degenerate. If* $\|\boldsymbol{u}_L - \boldsymbol{u}_R\| < \delta$ *for* $\delta > 0$*, is sufficiently small, then there exists a unique self-similar solution to the Riemann problem.*

For the proof of this theorem see, [31, Chapt. I. Theorem 6.1]. The Riemann problem was first studied by Riemann in the context of the isentropic Euler equations in the seminal paper [32]; an English translation can be found in [33, pg. 109]. In particular, the Riemann solution can be constructed using Lax's method, see [34, Sec. 9]. In particular, the solution to the Riemann problem consists of $m+1$ constant states separated by m waves. That is, there exists $2m$ numbers,

$$\lambda_1^- \le \lambda_1^+ \le \lambda_2^- \le \lambda_2^+ \le \cdots \le \lambda_m^- \le \lambda_m^+. \tag{1.9}$$

These $2m$ numbers define $2m+1$ sectors in the (x,t)-plane. In particular, they can be defined through the self-similarity parameter, $\xi := x/t$. That is, ξ belongs to one of the intervals, $(-\infty,\lambda_1^-)$, $(\lambda_1^-,\lambda_1^+)$, $(\lambda_1^+,\lambda_2^-)$, …, $(\lambda_m^-,\lambda_m^+)$, and (λ_m^+,∞). Note, some of these intervals can be empty. If the interval $(\lambda_i^-,\lambda_i^+)$ is non-empty for $i\in\{1:m\}$, then the solution inside this interval must be a rarefaction wave. All other wave solutions are either shocks or contact

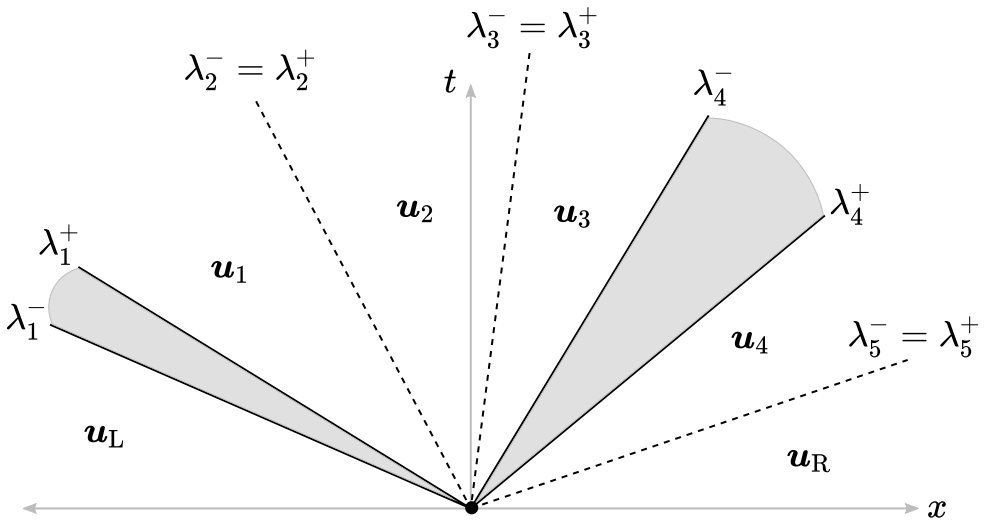

Figure 1.1: An example Riemann fan consisting of 5 waves and 6 constants states

discontinuities. Note that $\boldsymbol{u}(x,t) = \boldsymbol{u}_{\mathrm{L}}$ for $\frac{x}{t} \in (-\infty, \lambda_1^-)$ and $\boldsymbol{u}(x,t) = \boldsymbol{u}_{\mathrm{R}}$ for $\frac{x}{t} \in (\lambda_m^+, \infty)$. See Figure 1.1 for a visual description.

The main motivation for introducing these $2m$ numbers is that the numerical method described in Chapter 4 relies on estimation of the maximum speed of propagation. That is, we need an upper bound on

$$\lambda_{\max} := \max\{|\lambda_1^-|, |\lambda_m^+|\}. \tag{1.10}$$

For the purposes of our numerical method, me modify slightly the definition of the invariant region, see Definition 1.2.4.

Definition 1.3.2 (Invariant Set)**.** We say that $\mathcal{B} \subset \mathbb{R}^m$ is an **invariant set** for (1.1) if for every pair $(\mathbf{U}_L, \mathbf{U}_R) \in \mathcal{B} \times \mathcal{B}$ we have that average of the entropy solution to (1.8a)–(1.8b) with $\boldsymbol{g}(\boldsymbol{u}) := \mathrm{g}(\boldsymbol{u})\boldsymbol{n}$ for any $\boldsymbol{n} \in \mathbb{S}^{d-1}$ over the Riemann fan, remains in $\mathcal{B}$; that is,

$$\frac{1}{t(\lambda_m^+ - \lambda_1^-)} \int_{\lambda_1^- t}^{\lambda_m^+ t} \boldsymbol{u}(x,t)\, \mathrm{d}x \in \mathcal{B} \tag{1.11}$$

for any $t > 0$. □

Lemma 1.3.1. *Let* $\boldsymbol{u} \in L^\infty(\mathbb{R} \times (0,\infty); \mathbb{R}^m) \cap C^0((0,\infty); L^1_{\mathrm{loc}}(\mathbb{R}; \mathbb{R}^m))$ *be a solution to the*

Riemann problem (1.8a)–(1.8b) *with* $\boldsymbol{u}_{\mathrm{L}}, \boldsymbol{u}_{\mathrm{R}} \in \mathcal{B}$. *Then the following holds for all* $t \in (0, \frac{1}{2\lambda_{\max}}]$ *where* $\lambda_{\max}$ *is the maximal wave speed to the Riemann problem defined in* (1.10),

1. $\overline{\boldsymbol{u}}(t) := \int_{-1/2}^{1/2} \boldsymbol{u}(x,t)\,\mathrm{d}x = \frac{1}{2}(\boldsymbol{u}_{\mathrm{L}} + \boldsymbol{u}_{\mathrm{R}}) - t(\boldsymbol{g}(\boldsymbol{u}_{\mathrm{R}}) - \boldsymbol{g}(\boldsymbol{u}_{\mathrm{L}}))$.

2. $\overline{\boldsymbol{u}}(t) \in \mathcal{B}$.

3. *Let* $\Psi \in C^1(\mathcal{B};\mathbb{R})$ *be a quasiconcave functional. Assume that* $\Psi(\boldsymbol{u}(x,t)) \geq 0$, *for a.e.* $x \in \mathbb{R}$ *and all* $t > 0$. *Then* $\Psi(\overline{\boldsymbol{u}}) \geq 0$.

4. *Let* $\Psi \in C^1(\mathcal{B},\mathbb{R})$ *be a concave functional. Assume that* $\Psi(\boldsymbol{u}(x,t)) \geq 0$ *for a.e.* $x \in \mathbb{R}$ *and all* $t > 0$. *Assume that there exists* $\lambda_\flat, \lambda_\sharp \in [-\lambda_{\max}, \lambda_{\max}]$, $\lambda_\flat < \lambda_\sharp$, *so that* $\Psi(\boldsymbol{w}(x,t)) > 0$ *for a.e.* $\frac{x}{t} \in (\lambda_\flat, \lambda_\sharp)$. *Then* $\Psi(\overline{\boldsymbol{w}}(t)) > 0$.

Proof. The proof of this lemma is taken from Clayton et. al. [2, Lemma 3.2]. For the entire proof t is a fixed real number in $(0, \frac{1}{2\lambda_{\max}})$.
(i) Let $u_1, \ldots, u_m$ be the m components of $\boldsymbol{u}$, and let $g_1, \ldots, g_m$ be the m components of the flux $\boldsymbol{g}$. Let $l \in \{1:m\}$. Since $\boldsymbol{u}$ is a weak solution to (1.8a), we have

$$0 = \int_{-\infty}^{\infty}\int_{0}^{\infty} (-u_l \partial_\tau \phi - g_l(\boldsymbol{u})\partial_x \phi)\mathrm{d}\tau\mathrm{d}x - u_{l,\mathrm{L}} \int_{-\infty}^{0} \phi(x,0)\mathrm{d}x - u_{l,\mathrm{R}} \int_{0}^{\infty} \phi(x,0)\mathrm{d}x$$

for all $\phi \in W^{1,\infty}(\mathbb{R}\times[0,\infty);\mathbb{R})$ with compact support in $\mathbb{R}\times[0,\infty)$. Here, $u_{l,\mathrm{Z}}$ is the l-th component of $\boldsymbol{u}_{\mathrm{Z}}$. Now we define a sequence of smooth functions $(\phi_\epsilon)_{\epsilon>0}$ with $\phi_\epsilon(x,\tau) = \phi_{1,\epsilon}(|x|)\phi_{2,\epsilon}(\tau)$,

$$\phi_{1,\epsilon}(x) = \begin{cases} 1 & 0 \leq x \leq \frac{1}{2}, \\ \frac{1}{\epsilon}(-x + \frac{1}{2} + \epsilon) & \frac{1}{2} \leq x \leq \frac{1}{2} + \epsilon, \\ 0 & \frac{1}{2} + \epsilon \leq x, \end{cases} \quad \phi_{2,\epsilon}(\tau) = \begin{cases} 1 & 0 \leq \tau \leq t, \\ \frac{1}{\epsilon}(-\tau + t + \epsilon) & t \leq \tau \leq t + \epsilon, \\ 0 & t + \epsilon \leq \tau. \end{cases}$$

Using that $u_l \in C^0([0,\infty); L^1_{\mathrm{loc}}(\mathbb{R}))$, we infer that $\int_{-\infty}^{\infty}\int_0^\infty -u_l\partial_\tau\phi_\epsilon \mathrm{d}x\mathrm{d}\tau \to \int_{-\frac{1}{2}}^{\frac{1}{2}} u_l(x,t)\mathrm{d}x$ as $\epsilon \to 0$. Likewise, we have $\int_{-\infty}^{\infty}\int_0^\infty -g_l(\boldsymbol{u})\partial_x\phi_\epsilon \mathrm{d}\tau\mathrm{d}x \to \int_0^t (g_l(\boldsymbol{u}_{\mathrm{R}}) - g_l(\boldsymbol{u}_{\mathrm{L}}))\mathrm{d}\tau = (g_l(\boldsymbol{u}_{\mathrm{R}}) -$

$g_l(\boldsymbol{u}_{\mathrm{L}}))t$ as $\epsilon \to 0$. Finally, $-u_{l,\mathrm{L}} \int_{-\infty}^{0} \phi_\epsilon(x,0)\mathrm{d}x - u_{l,\mathrm{R}} \int_{0}^{\infty} \phi_\epsilon(x,0)\mathrm{d}x \to -\frac{1}{2}(u_{l,\mathrm{L}} + u_{l,\mathrm{R}})$ as $\epsilon \to 0$. In conclusion, we have established that

$$0 = \overline{\boldsymbol{u}}(t) + (\boldsymbol{g}(\boldsymbol{u}_{\mathrm{R}}) - \boldsymbol{g}(\boldsymbol{u}_{\mathrm{L}}))t - \tfrac{1}{2}(\boldsymbol{u}_{\mathrm{L}} + \boldsymbol{u}_{\mathrm{R}}). \tag{1.12}$$

(ii) Since $\mathcal{B}$ is convex, $\boldsymbol{u}(x,t) \in \mathcal{B}$ for æ$x \in \mathbb{R}$ and all $t > 0$, and the length of the interval $[-\frac{1}{2}, \frac{1}{2}]$ is 1, we infer that $\overline{\boldsymbol{u}}(t) \in \mathcal{B}$.

(iii) Let $\Psi \in C^1(\mathcal{B}; \mathbb{R})$ be a quasiconcave functional. The quasiconcavity implies that $\Psi(\overline{\boldsymbol{u}}(t)) \geq \operatorname{ess\,inf}_{x \in (-\frac{1}{2}, \frac{1}{2})} \Psi(\boldsymbol{u}(x,t)) \geq 0$.

(iv) Let $\Psi \in C^1(\mathcal{B}; \mathbb{R})$ be a concave functional. Jensen's inequality implies

$$\Psi(\overline{\boldsymbol{u}}(t)) \geq \int_{-\frac{1}{2}}^{\frac{1}{2}} \Psi(\boldsymbol{u}(x,t))\mathrm{d}x \geq \int_{\lambda_\flat t}^{\lambda_\sharp t} \Psi(\boldsymbol{u}(x,t))\mathrm{d}x > 0, \tag{1.13}$$

where we used $-\frac{1}{2} \leq \lambda_\flat t < \lambda_\sharp t \leq \frac{1}{2}$. This concludes the proof. □

Results 1. and 2. in Lemma 1.3.1 are used in Chapter 4 to prove that the numerical method is invariant-domain preserving. Results 3. and 4. are used for the purposes of quasiconcave limiting, see Chapter 6. For more information regarding the existence of self similar solutions to the Riemann problem see, Dafermos [35, Chapter IX].

1.4 The Compressible Euler Equations

The main focus of this thesis is on numerical methods for solving the compressible Euler equations. These equations are a fundamental and important model for simulating fluid flow with very small viscosity (or, equivalently, a very large Reynolds number). They also form the basis for more complicated multi-physics models.

The Euler equations represent the conservation of mass, momentum and total energy;

they are given respectively as,

$$\partial_t \rho(\boldsymbol{x},t) + \nabla \cdot (\rho(\boldsymbol{x},t)\boldsymbol{v}(\boldsymbol{x},t)) = 0, \qquad (\boldsymbol{x},t) \in \mathbb{R}^d \times (0,\infty), \tag{1.14a}$$

$$\partial_t \boldsymbol{m}(\boldsymbol{x},t) + \nabla \cdot (\tfrac{\boldsymbol{m}(\boldsymbol{x},t)}{\rho(\boldsymbol{x},t)} \otimes \boldsymbol{m}(\boldsymbol{x},t) + p(\rho, \mathsf{e}(\boldsymbol{u}))\mathbb{I}_d) = \boldsymbol{0}, \qquad (\boldsymbol{x},t) \in \mathbb{R}^d \times (0,\infty), \tag{1.14b}$$

$$\partial_t E(\boldsymbol{x},t) + \nabla \cdot (\tfrac{\boldsymbol{m}(\boldsymbol{x},t)}{\rho(\boldsymbol{x},t)}(E(\boldsymbol{x},t) + p(\rho, \mathsf{e}(\boldsymbol{u}))) = 0, \qquad (\boldsymbol{x},t) \in \mathbb{R}^d \times (0,\infty), \tag{1.14c}$$

where ρ is the density, $\boldsymbol{m} = (m_1, \ldots, m_d)$ is the momentum, E is the total energy, $\otimes$ is the outer product, d is the spatial dimension (where d is either 1, 2, or 3) and $\boldsymbol{u} = (\rho, \boldsymbol{m}, E)^T$. The pressure mapping $\mathbb{R}^2 \supset \mathcal{A} \ni (\rho, e) \mapsto p(\rho, e) \in \mathbb{R}$ is defined by an *equation of state* on some suitable thermodynamic region, $\mathcal{A}$, for which the equation of state can be inverted as, $e = e(\rho, p)$; see Chapter 2. Furthermore, we refer to the pressure, $p = p(\rho, e)$ as the **oracle** as we make no assumptions on the exact description of the equation of state (EOS). The total energy is, $E = \rho e + \frac{1}{2}\rho\|\boldsymbol{v}\|_{\ell^2}$, where ρe is the internal energy and $\frac{1}{2}\rho\|\boldsymbol{v}\|_{\ell^2}$ is the kinetic energy with $\boldsymbol{v} := \boldsymbol{m}/\rho$ being the velocity and $\|\cdot\|_{\ell^2}$ the usual Euclidean norm. Solving for e, we can also write, $e(\rho, p) = \mathsf{e}(\boldsymbol{u}) := \frac{E}{\rho} - \frac{\|\boldsymbol{m}\|_{\ell^2}^2}{2\rho^2}$. We write the Euler equations in the more general form,

$$\partial_t \boldsymbol{u} + \nabla \cdot \mathbb{f}(\boldsymbol{u}) = \boldsymbol{0}, \tag{1.15}$$

with,

$$\mathbb{f}(\boldsymbol{u}) := \begin{pmatrix} \boldsymbol{m}^\mathsf{T} \\ \frac{\boldsymbol{m}}{\rho} \otimes \boldsymbol{m} + p\mathbb{I}_d \\ \frac{\boldsymbol{m}^\mathsf{T}}{\rho}(E+p) \end{pmatrix}. \tag{1.16}$$

Additionally, for practical purposes, we must numerically solve the Euler equations on a bounded domain $D \subset \mathbb{R}^d$. This requires the implementation of boundary conditions. Implementation of boundary conditions is reserved for Chapter 7.

Remark 1.4.1 (Assumptions on the Oracle)**.** The only assumption we make on the equation of state is that, $p + p_\infty \geq 0$ where p_∞ is a reference pressure state or is the absolute value of the global minimum pressure. If p_∞ is, a priori, not known, then we set $p_\infty = 0$ and require

$p \geq 0$. □

Remark 1.4.2 (Parameters Given by the Oracle)**.** There are certain parameters that the oracle may possess that can be used in the approximation of the problem. We list them here. The *maximum compressibility* (or *covolume*) constant, $b > 0$ which represents the smallest volume the fluid can occupy; that is, $\tau - b > 0$ or $\rho < b^{-1}$. A reference specific internal energy, $q \in \mathbb{R}$, so that $e - q > 0$. A reference pressure p_∞ as described in Remark 1.4.1. The exact use of these parameters will be evident in Chapter 4. □

1.5 The Riemann Problem for the Euler Equations

As the compressible Euler equations are a multi-dimensional system of conservation laws, the notion of a Riemann problem is not necessarily clear. For many numerical methods, the Riemann problem for the Euler equations is defined in terms of a direction, $\boldsymbol{n}$. Let $\boldsymbol{n} \in \mathbb{S}^{d-1}(\mathbf{0}, 1)$ be given, then define the orthonormal basis, $\{\boldsymbol{n}, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_{d-1}\}$. The exact choice of $\{\boldsymbol{t}_i\}_{i=1}^{d-1}$ is not important. With respect to this basis, we write the momentum as, $\boldsymbol{m} = (m, \boldsymbol{m}^\perp)^\mathsf{T}$ where $m := \boldsymbol{m} \cdot \boldsymbol{n}$ and $\boldsymbol{m}^\perp := (\boldsymbol{m} \cdot \boldsymbol{t}_1, \ldots, \boldsymbol{m} \cdot \boldsymbol{t}_{d-1})^\mathsf{T}$. The velocity is defined similarly, $\boldsymbol{v} = (v, \boldsymbol{v}^\perp)^\mathsf{T}$ where $v := m/\rho$ and $\boldsymbol{v}^\perp := \boldsymbol{m}^\perp/\rho$. Thus the Riemann problem in the direction $\boldsymbol{n}$ is defined by,

$$\partial_t \boldsymbol{u} + \partial_x(\mathbb{f}(\boldsymbol{u})\boldsymbol{n}) = \mathbf{0}, \qquad \text{for } x \in \mathbb{R},\ t > 0 \tag{1.17}$$

$$\boldsymbol{u}(x, 0) = \begin{cases} \boldsymbol{u}_\mathrm{L}, & x < 0, \\ \boldsymbol{u}_\mathrm{R}, & x > 0, \end{cases} \tag{1.18}$$

This can be explicitly written as,

$$\partial_t \begin{pmatrix} \rho \\ m \\ \boldsymbol{m}^\perp \\ E \end{pmatrix} + \partial_x \begin{pmatrix} m \\ \frac{1}{\rho} m^2 + p(\boldsymbol{u}) \\ \frac{m}{\rho} \boldsymbol{m}^\perp \\ \frac{m}{\rho}(E + p(\boldsymbol{u})) \end{pmatrix} = \mathbf{0} \tag{1.19}$$

The solution to this Riemann problem is carried out by first solving,

$$\partial_t \begin{pmatrix} \rho \\ m \\ \mathcal{E} \end{pmatrix} + \partial_x \begin{pmatrix} m \\ \frac{1}{\rho} m^2 + p(\boldsymbol{u}) \\ \frac{m}{\rho}(\mathcal{E} + p(\boldsymbol{u})) \end{pmatrix} = \mathbf{0} \tag{1.20}$$

where $\mathcal{E} := E - \frac{\|\boldsymbol{m}^\perp\|_{\ell^2}^2}{2\rho}$. Then $\boldsymbol{m}^\perp$ is found afterwards by solving $\partial_t \boldsymbol{m}^\perp + \partial_x(\frac{m}{\rho}\boldsymbol{m}^\perp) = \mathbf{0}$. Note also that $\rho e = \mathcal{E} - \frac{m^2}{2\rho} = E - \frac{\|\boldsymbol{m}\|_{\ell^2}^2}{2\rho}$; this says that the internal energy does not depend on the change of basis. More on the solution to the Riemann problem will be presented in Chapter 4.

Remark 1.5.1 (Invariant Sets for the Euler Equations)**.** The compressible Euler equations depend on the EOS, therefore, each EOS may define a different invariant set. For example, if the EOS is given by the ideal gas law, $p(\rho, e) = (\gamma - 1)\rho e$, then, a possible invariant set would be,

$$\mathcal{B} := \{\boldsymbol{u} \in \mathbb{R}^{d+2} : \rho > 0,\, \mathsf{e}(\boldsymbol{u}) > 0,\, \mathsf{s}(\boldsymbol{u}) \geq s_{\min}\}. \tag{1.21}$$

Note $\mathsf{s}(\boldsymbol{u}) \geq s_{\min}$ is the minimum principle on the specific entropy and is discussed more in Chapter 2. Furthermore, the sets, $\mathcal{B} := \{\boldsymbol{u} \in \mathbb{R}^{d+2} : \rho > 0,\, \mathsf{e}(\boldsymbol{u}) > 0\}$. and $\mathcal{B} := \{\boldsymbol{u} \in \mathbb{R}^{d+2} : \rho > 0\}$ are also invariant sets.

If the EOS is given by the covolume EOS, $p(\rho, e) = (\gamma - 1)\frac{\rho e}{1 - b\rho}$ then an invariant set involves the maximum compressibility constant, b^{-1} (for $b \neq 0$). That is,

$$\mathcal{B}(b) := \{\boldsymbol{u} \in \mathbb{R}^{d+2} : 0 < \rho < b^{-1},\, \mathsf{e}(\boldsymbol{u}) > 0,\, \mathsf{s}(\boldsymbol{u}) \geq s_{\min}\}. \tag{1.22}$$

As we will see through the course of this thesis, we will be working with the Nobel-Abel Stiffened Gas EOS (see Section 2.2). An invariant set for this EOS is,

$$\mathcal{B}(b, q, p_\infty) := \{\boldsymbol{u} \in \mathbb{R}^{d+2} : 0 < \rho < b^{-1},\, \mathsf{e}(\boldsymbol{u}) - q > p_\infty(\rho^{-1} - b)\}. \tag{1.23}$$

□

Lastly, we would also like to present the Rankine-Hugoniot conditions for the compressible Euler equations for reference.

Definition 1.5.1. The Rankine-Hugoniot conditions for the compressible Euler equations are given by,

$$S(\rho_{\mathrm{L}} - \rho_{\mathrm{R}}) = \rho_{\mathrm{L}} v_{\mathrm{L}} - \rho_{\mathrm{R}} v_{\mathrm{R}}, \tag{1.24a}$$

$$S(\rho_{\mathrm{L}} v_{\mathrm{L}} - \rho_{\mathrm{R}} v_{\mathrm{R}}) = \rho_{\mathrm{L}} v_{\mathrm{L}}^2 + p_{\mathrm{L}} - (\rho_{\mathrm{R}} v_{\mathrm{R}}^2 + p_{\mathrm{R}}), \tag{1.24b}$$

$$S(\mathcal{E}_{\mathrm{L}} - \mathcal{E}_{\mathrm{R}}) = v_{\mathrm{L}}(\mathcal{E}_{\mathrm{L}} + p_{\mathrm{L}}) - v_{\mathrm{R}}(\mathcal{E}_{\mathrm{R}} + p_{\mathrm{R}}). \tag{1.24c}$$

□

2. THE EQUATION OF STATE

In this chapter we cover specific topics pertaining to the many different EOS.

The *equation of state* (EOS) is one of the most important aspects of the Euler equations, it is an equation which relates three thermodynamic quantities; for example, the density, temperature and pressure or volume, specific internal energy, and specific entropy. The choice of the EOS is chosen based on the physical problem of interest. In particular, the EOS models the material. The EOS are often designed to accurately model specific materials under certain conditions. Typically, the Euler equations are numerically solved using the ideal gas law as it is the simplest. In addition, most theoretical results for the Euler equations are proven for the ideal gas law. There are, in fact, hundreds of different EOS and numerically solving the Euler equations with invariant domain preserving properties becomes a highly non-trivial task for each of these equations. To further exacerbate the issue, many industrial and research laboratories rely on EOS which are tabulated; that is, there is no analytic function, but rather, thermodynamic quantities are generated from a database.

We begin by reviewing some essential principles in thermodynamics. The first law of thermodynamics states that the change in energy in a non-adiabatic system; that is, a system for which energy can be transferred through the walls of the system, is equal to the change in heat plus the work done on that system. Letting Q denote the heat, the first law of thermodynamics is written as,

$$\mathrm{d}e = \mathrm{d}Q - p\mathrm{d}V. \tag{2.1}$$

Since the EOS can be written in terms of many different thermodynamic variables, it is often convenient to overload the notation. For example, we may use the same symbol p to write $p = p(\tau, e)$ and $p = p(\rho, s)$. The context should be clear context but comments will be made when necessary to avoid confusion. To further emphasize the variables being used, the notation for partial differentiation is written in the form $\left(\frac{\partial f}{\partial x}\right)_y$, where $\left(\,\cdot\,\right)_y$ emphasizes

that the variable held constant is y. This also displays that f is a function of x and y.

A common assumption for the EOS is the existence of a physical entropy. That is, the assumption that there is a function $s = s(\tau, e)$ which satisfies the second law of thermodynamics

$$T\,\mathrm{d}s = \mathrm{d}e + p\,\mathrm{d}\tau. \tag{2.2}$$

We must assume that the entropy, $s(\tau, e)$, is strictly convex (as a function of (τ, e)). Otherwise, the pressure can be multivalued as a function of τ and s. Furthermore, from Clairaut's theorem we have

$$\frac{\partial}{\partial \tau}\Big(\frac{\partial s}{\partial e}(\tau, e)\Big) = \frac{\partial}{\partial e}\Big(\frac{\partial s}{\partial \tau}(\tau, e)\Big). \tag{2.3}$$

Using appropriate identities for $\left(\frac{\partial s}{\partial e}\right)_\tau$ and $\left(\frac{\partial s}{\partial \tau}\right)_e$, (2.3) gives us the so-called, *Maxwell's relation*, see Callen [36, Chapter 7]. However, the existence of an entropy is not always guaranteed.

Definition 2.0.1 (Incomplete EOS)**.** An *incomplete* equation of state is a relation of the form, $p = p(\rho, e)$. It is called *incomplete* as the equation cannot completely describe all of the thermodynamical properties of the system. □

Definition 2.0.2 (Complete EOS)**.** A *complete* equation of state is one for which a concave entropy, $s = s(\tau, e)$, exists such that $p = p(\tau, s)$. Or equivalently, the mapping $(\tau, e) \mapsto -s(\tau, e)$, is convex. This equation of state completely describes all of the thermodynamic properties of the system. □

So an *incomplete* EOS makes no assumption on the existence of a convex entropy. Any given complete EOS describes a unique incomplete EOS; however; the converse is not true, there can be more than one complete EOS corresponding to the same incomplete EOS. For more on this, see Menikoff & Plohr [37, Sec. II. F.]. Note that the first order method described in Chapter 4 does not require a complete EOS. However, the second order method in some way requires the use of an entropy (which may not exist), this is resolved in Chapter 5 and Chapter 6.

We now review some important facts regarding the entropy. First, concavity of $s(\tau, e)$ requires that

$$\left(\frac{\partial^2 s}{\partial \tau^2}\right)_e < 0, \quad \left(\frac{\partial^2 s}{\partial e^2}\right)_\tau < 0, \quad \text{and} \quad \left(\frac{\partial^2 s}{\partial \tau^2}\right)_e \left(\frac{\partial^2 s}{\partial e^2}\right)_\tau - \left(\frac{\partial^2 s}{\partial \tau \partial e}\right)^2 < 0, \tag{2.4}$$

i.e., the Hermitian matrix of $-s(\tau, e)$ is positive definite.

Remark 2.0.1 (Convex Entropy in the Literature)**.** In the literature, concavity of the entropy is often framed as $(\tau, e) \mapsto -s(\tau, e)$ being convex. □

Remark 2.0.2 (Assumptions on the Oracle)**.** Throughout the course of this Thesis, the only assumption we make on the oracle, is that there exists a minimum bound on the pressure, $-p_\infty$, for $p_\infty \geq 0$. Furthermore, we make no assumption on the existence of an entropy for the oracle. Other remarks will be made regarding the oracle when relevant. □

Note that a convex entropy From Equation (2.2), we have the following identities,

$$\left(\frac{\partial s}{\partial e}\right)_\tau = T^{-1}, \quad \text{and} \quad \left(\frac{\partial s}{\partial \tau}\right)_e = pT^{-1}. \tag{2.5}$$

The pressure can then be defined by,

$$p(\tau, e) = \frac{\left(\frac{\partial s}{\partial \tau}\right)_e}{\left(\frac{\partial s}{\partial e}\right)_\tau} \tag{2.6}$$

or by,

$$\frac{\partial s}{\partial \tau} - p(\tau, e)\frac{\partial s}{\partial e} = 0. \tag{2.7}$$

We also have $\left(\frac{\partial e}{\partial \tau}\right)_s = -p(\tau, s)$ and $\left(\frac{\partial e}{\partial s}\right)_\tau = T$ from Equation (2.2). It is also assumed that the EOS must satisfy $\left(\frac{\partial e}{\partial s}\right)_\tau = T > 0$.

2.1 Relevant Thermodynamic Relations

The material speed of sound is defined by,

$$a := \sqrt{\Big(\frac{\partial p}{\partial \rho}\Big)_s} = \sqrt{-\tau^2\Big(\frac{\partial p}{\partial \tau}\Big)_s}. \tag{2.8}$$

As with the pressure, we also leave the independent variables of the speed of sound unspecified; the appropriate variables should be clear from context. Finding the specific entropy, s, is not always an easy task. For an incomplete EOS the sound speed can more easily be computed by applying the chain rule and (2.2), to (2.8) to find,

$$a = \sqrt{-\tau^2\Big[\Big(\frac{\partial p}{\partial \tau}\Big)_e + \Big(\frac{\partial p}{\partial e}\Big)_\tau \Big(\frac{\partial e}{\partial \tau}\Big)_s\Big]} = \sqrt{-\tau^2\Big[\Big(\frac{\partial p}{\partial \tau}\Big)_e - p\Big(\frac{\partial p}{\partial e}\Big)_\tau\Big]}. \tag{2.9}$$

The specific heat of a fluid is defined by $\frac{\mathrm{d}Q}{\mathrm{d}T}$. From the first law of thermodynamics, we have, $\mathrm{d}Q = \mathrm{d}e + p\,\mathrm{d}\tau = T\,\mathrm{d}s$. We then are able to define the specific heat at constant volume and constant pressure using the first law of thermodynamics by,

$$c_v := \Big(\frac{\partial e}{\partial T}\Big)_\tau = T\Big(\frac{\partial s}{\partial T}\Big)_\tau \qquad c_p := T\Big(\frac{\partial s}{\partial T}\Big)_p = \Big(\frac{\partial (e + p\tau)}{\partial T}\Big)_p = \Big(\frac{\partial h}{\partial T}\Big)_p. \tag{2.10}$$

2.2 The Noble-Abel Stiffened Gas EOS

A large portion of the material in this thesis relies on the use of the Noble-Abel Stiffened Gas EOS (NASG EOS). This equation of state was first introduced by Le M'etayer & Saurel in [38] and has been extended to more general form in [39]. The NASG EOS can also be viewed as an extension of the Noble-Abel (or covolume) EOS (see [40] for a more recent discussion and use of the Noble-Abel EOS).

The NASG EOS is defined by the incomplete EOS,

$$p(\tau, e) = (\gamma - 1)\frac{e - q}{\tau - b} - \gamma p_\infty \tag{2.11}$$

where q is some reference internal energy, p_∞ is also a reference pressure state, and b represents the maximum compressibility of the fluid; that is, the fluid cannot be compressed to an infinitely small volume. It can be shown that the specific entropy is given by,

$$s(\tau, e) = \log\Big(\big(e - q - p_\infty(\tau - b)\big)^{\frac{1}{\gamma-1}}(\tau - b)\Big). \tag{2.12}$$

That is, $s(\tau, e)$ satisfies equation (2.7). It is important to note that this equation of state is convex, see [38, Appendix B], which is necessary in the solution to the extended Riemann problem in Section 4.2.

2.3 The van der Waals EOS

The van der Waals thermal EOS is given by,

$$p(\tau, T) := \frac{RT}{\tau - b} - \frac{a}{\tau^2}, \tag{2.13}$$

where b represents the maximum compression of the fluid; that is, $\tau > b$, a is a material dependent constant describing the attractive forces of fluid and should not be confused with the speed of sound, a, and R is the universal gas constant. This EOS was first derived by van der Waals in [41]. In order to find the complete EOS, more information must be provided. In particular, we suggest that the temperature be defined by

$$T = \frac{\gamma - 1}{R}\Big(e + \frac{a}{\tau}\Big). \tag{2.14}$$

The reasoning behind this follows from Maxwell's relation (2.3). The details can be found in [36, Sec. 3.5]. From this definition, we find the caloric EOS to be,

$$p(\tau, e) = (\gamma - 1)\frac{e + a/\tau}{\tau - b} - \frac{a}{\tau^2} \tag{2.15}$$

One also sees that, under the above definitions of T and p, the specific entropy,

$$s(\tau, e) = R \log\left(\left(e + \tfrac{a}{\tau}\right)^{\frac{1}{\gamma-1}}(\tau - b)\right) - s_0, \quad s_0 \in \mathbb{R}, \tag{2.16}$$

satisfies the second law of thermodynamics, (2.2). Using (2.15) and (2.16) we can eliminate e and find the complete EOS for p,

$$p(\tau, s) = \frac{\gamma - 1}{(\tau - b)^{\gamma}} \exp\left(\frac{\gamma - 1}{R}(s - s_0)\right) - \frac{a}{\tau^2}. \tag{2.17}$$

Recall the sound speed is,

$$a^2 = \frac{(\gamma - 1)\gamma\tau^2}{(\tau - b)^{\gamma+1}} \exp\left(\frac{\gamma - 1}{R}(s - s_0)\right) - \frac{2a}{\tau}. \tag{2.18}$$

Using (2.17) we can compute the sound speed in terms of p and ρ.

$$a^2 = \gamma \frac{p + a\rho^2}{\rho(1 - b\rho)} - 2a\rho. \tag{2.19}$$

2.4 The Cubic EOS

There is a general class of EOS referred to as the cubic EOS as the pressure is a cubic function of $\sqrt{T}$. A general from of this EOS is can be defined as,

$$p(\tau, T) = \frac{RT}{\tau - b} - \frac{\alpha(T)}{(\tau - br_1)(\tau - br_2)}, \tag{2.20}$$

where $\alpha(T)$ is referred to as the attractive term. We refer to [42] for a survey of this general class of EOS. Note that the van der Waals EOS for the pressure is recovered if one sets $\alpha(T) = a$ and $r_1 = r_2 = 0$. Next, we suppose that the specific internal energy is given by,

$$e(\tau, T) = c_v T + \frac{\alpha(T) - T\alpha'(T)}{b(r_1 - r_2)} \log\left(\frac{\tau - br_1}{\tau - br_2}\right) + e_0, \tag{2.21}$$

where e_0 is a reference specific internal energy and $r_1 \neq r_2$. This assumption follows from the relation, $\mathrm{d}e = c_v \mathrm{d}T + \left(T\left(\frac{\partial p}{\partial T}\right)_\tau - p\right)\mathrm{d}\tau$.

2.4.1 The Redlich-Kwong EOS

The Redlich-Kwong EOS was first introduced in by Redlich and Kwong in [43]. It can be obtained from the general cubic EOS by setting $\alpha(T) = a/\sqrt{T}$, $r_1 = 0$, and $r_2 = -1$. The parameters a and b are the same parameters from the van der Waals EOS. The equation of state is thus,

$$p(\tau, T) = \frac{RT}{\tau - b} - \frac{a}{\sqrt{T}\tau(\tau + b)}, \tag{2.22a}$$

$$e(\tau, T) = c_v T + \frac{3a}{2b\sqrt{T}} \log\left(\frac{\tau}{\tau + b}\right) + e_0. \tag{2.22b}$$

Note that an explicit expression for $p = p(\tau, e)$ cannot be written. Instead, one must first solve a cubic equation and then use that solution in $p(\tau, T)$ or $e(\tau, T)$ depending on the desired quantity. The two cubic equations to solve for temperature, given p or e, respectively, are,

$$\widetilde{T}^3 - \frac{p(\tau - b)}{R}\widetilde{T} - \frac{a(\tau - b)}{\tau(\tau + b)} = 0, \tag{2.23a}$$

$$\widetilde{T}^3 - \frac{e - e_0}{c_v}\widetilde{T} + \frac{3a}{2bc_v} \log\left(\frac{\tau}{\tau + b}\right) = 0, \tag{2.23b}$$

for $\widetilde{T} := \sqrt{T}$. We can also see that the specific entropy defined by,

$$s(\tau, T) := c_v \log(T) + \frac{a}{2bT^{3/2}} \log\left(\frac{\tau}{\tau + b}\right) + R \log(\tau - b), \tag{2.24}$$

satisfies the 2nd law of thermodynamics, (2.2).

Remark 2.4.1 (Solving the Cubic)**.** For completion, we briefly describe how to compute the roots for a cubic equation of the form, $x^3 + c_1 x + c_0 = 0$. Define $\Delta := \frac{c_1^3}{27} + \frac{c_0^2}{4}$. If $\Delta \geq 0$,

then this cubic equation has one real root,

$$x_1 = \sqrt[3]{-\frac{c_0}{2} + \sqrt{\Delta}} + \sqrt[3]{-\frac{c_0}{2} - \sqrt{\Delta}}. \tag{2.25}$$

If $\Delta < 0$, then there are three real roots and they are defined by,

$$x_k := 2\sqrt{-\frac{c_1}{3}} \cos\Big[\frac{1}{3}\arccos\Big(\frac{3c_0}{2c_1}\sqrt{\frac{-3}{c_1}}\Big) - \frac{2\pi(k-1)}{3}\Big], \quad \text{for } k = 1, 2, 3. \tag{2.26}$$

Note, in both cases, (2.23a)–(2.23b), for $T = 0$, the cubic equation is negative, and since the coefficient of $\widetilde{T}^3$ is positive, we must have at least one positive root. In the event that there are multiple positive roots, then we simply take the largest root. □

Equations of State in $p = p(\rho, e)$ Form	
Ideal	$p = (\gamma - 1)\rho e$ (2.27)
Covolume	$p = (\gamma - 1)\dfrac{\rho e}{1 - b\rho}$ (2.28)
van der Waals	$p = (\gamma - 1)\dfrac{\rho e + a\rho^2}{1 - b\rho} - a\rho^2$ (2.29)
Redlich-Kwong	No explicit formula
Jones-Wilkins-Lee	$p = A\Big(1 - \dfrac{\omega}{R_1}\dfrac{\rho}{\rho_0}\Big)e^{-R_1\frac{\rho_0}{\rho}} + B\Big(1 - \dfrac{\omega}{R_2}\dfrac{\rho}{\rho_0}\Big)e^{-R_2\frac{\rho_0}{\rho}} + \omega\rho e$ (2.30)

Table 2.1: A quick reference for common equations of state in $p = p(\rho, e)$ form. More information on these EOS can be found in Chapter 2.

Sound Speeds for Different EOS in $a = a(\rho, p)$ Form	
Ideal	$$a^2 = \frac{\gamma p}{\rho} \quad (2.31)$$
Covolume	$$a^2 = \frac{\gamma p}{\rho(1-b\rho)} \quad (2.32)$$
van der Waals	$$a^2 = \gamma \frac{p + a\rho^2}{\rho(1-b\rho)} - 2a\rho \quad (2.33)$$
Redlich-Kwong	No explicit formula
Jones-Wilkins-Lee	$$a^2 = \frac{(\omega+1)p}{\rho} + A\Big(\frac{R_1\rho_0}{\rho^2} + \frac{\omega(\omega+1+\rho)}{R_1\rho_0} - 1\Big)\exp\Big(-R_1\frac{\rho_0}{\rho}\Big) + B\Big(\frac{R_2\rho_0}{\rho^2} + \frac{\omega(\omega+1+\rho)}{R_2\rho_0} - 1\Big)\exp\Big(-R_2\frac{\rho_0}{\rho}\Big) \quad (2.34)$$

Table 2.2: A quick reference for common sound speeds for several EOS $a = a(\rho, p)$ form. More information on these EOS can be found in Chapter 2.

2.5 The Jones-Wilkins-Lee EOS

The Jones-Wilkins-Lee (JWL) EOS is an empirical EOS used to model detonation products in multi-material reactive flow. The ideal gas law is often used to model the surrounding ambient fluid. As the purpose of this thesis is only for single material compressible Euler equations, we use the JWL EOS as a demonstration of the numerical method. The original paper for the development of this EOS can be found in [44].

The pressure is defined by,

$$p(\rho, e) := A\Big(1 - \frac{\omega}{R_1}\frac{\rho}{\rho_0}\Big)\exp\Big(-R_1\frac{\rho_0}{\rho}\Big) + B\Big(1 - \frac{\omega}{R_2}\frac{\rho}{\rho_0}\Big)\exp\Big(-R_2\frac{\rho_0}{\rho}\Big) + \omega\rho(e - e_0), \quad (2.35)$$

where A, B, R_1 and R_2 are parameters specific to the model or experiment, ρ_0 and e_0 are

some reference density and specific internal energy, and ω is also a chosen constant depending on the material, and is related to the Grüneisen coefficient.

Remark 2.5.1 (JWL Misinterpretation)**.** There is a lot of confusion surrounding the exact definition of the JWL EOS. In some cases in the literature, the JWL EOS is taken as an isentrope and not an EOS by replacing $e - e_0$ with e_0. The technical report, [45], provides an overview of this confusion. □

2.6 The Mie-Gruneisen EOS

The Mie-Gruneisen EOS is an incomplete EOS defined by,

$$p(\rho, e) := p_{\text{ref}}(\rho) + \rho\Gamma(\rho)(e - e_{\text{ref}}(\rho)) \tag{2.36}$$

where p_{ref} and e_{ref} are some reference pressure and specific internal energy curves, respectively. For example the reference curve could be shock locus, an isotherm, an isentrope and so on. A quick overview of this incomplete EOS as well as a means to *complete* it, is given in [46].

For our numerical demonstrations we use the linear Hugoniot locus, defined by,

$$p_{\text{ref}}(\rho) := P_0 + \rho_0 c_0^2 \frac{1 - \frac{\rho_0}{\rho}}{\left(1 - s(1 - \frac{\rho_0}{\rho})\right)^2}, \tag{2.37a}$$

$$e_{\text{ref}}(\rho) := e_0 + \frac{P_0 + p_{\text{ref}}(\rho)}{2\rho_0}\Big(1 - \frac{\rho_0}{\rho}\Big), \tag{2.37b}$$

where $s > 1$ and ρ_0, c_0, e_0, and P_0 are reference density, sound speed, specific internal energy, and pressure, respectively. This particular equation of state is used for modeling solids under high pressures. More details on this particular equation of state can be found in [47, Sec. 4.4].

The numerical method that we outline in this thesis is actually discretization independent. However, our numerical simulations are computed using the finite element method; in particular, $\mathbb{P}_1$ and $\mathbb{Q}_1$ continuous finite elements. There is a wealth of literature on the finite element method. Some recommended resources for this topic are: Grossman et. al. [48], Larsson & Thomèe [49], Ciarlet [50], and the three volume series on the finite element method by Ern & Guermond [51], [52], and [53].

Therefore, in this chapter, we present some of the fundamentals regarding the continuous finite element method.

3.1 The Continuous Galerkin (cG) Framework

We use a continuous Galerkin method for solving the compressible Euler equations. We start by defining the geometric finite element (using the notation of Ciarlet, [50]) $(\widehat{K}_{\text{geo}}, \widehat{P}_{\text{geo}}, \widehat{\Sigma}_{\text{geo}})$. Here, $\widehat{K}_{\text{geo}}$ is the reference element composed with the vertices $\{\widehat{\boldsymbol{a}}_i\}_{i\in\widehat{n}_{\text{geo}}}$, $\widehat{P}_{\text{geo}}$ is the geometric polynomial space used to construct the geometric mapping, and $\widehat{\Sigma}_{\text{geo}}$ are the nodal Lagrange degrees of freedom (dofs). Let $\{\widehat{\theta}_i\}_{i\in\widehat{n}_{\text{geo}}}$ denote the collection of reference shape functions. That is, $\widehat{\sigma}_i(\widehat{\theta}_j) = \widehat{\theta}_j(\widehat{\boldsymbol{a}}_i) = \delta_{ij}$ for $i, j \in \widehat{n}_{\text{geo}}$ and $\widehat{\sigma}_i \in \widehat{\Sigma}_{\text{geo}}$.

We use this reference geometric finite element to construct a collection of mappings, of which, defines our mesh. Let $\mathcal{T}_h = \{K_i\}_{i\in n_{\text{shape}}}$ denote a sequence of shape regular non-overlapping elements which exactly covers our domain D. Then $\bigcup_{i\in n^{\text{shape}}} K_i = D$. Let (K, P, Σ) be a local finite element where K is polytope with vertices $\{\boldsymbol{a}_i\}_{i\in n_{\text{geo}}}$. Then the affine geometric mapping, $\boldsymbol{T}_K : \widehat{K}_{\text{geo}} \to K$ is defined by,

$$\boldsymbol{T}_K(\widehat{\boldsymbol{x}}) = \sum_{i\in n} \widehat{\theta}_i(\widehat{\boldsymbol{x}})\boldsymbol{a}_i, \qquad \text{for } \widehat{\boldsymbol{x}} \in \widehat{K}_{\text{geo}}. \tag{3.1}$$

Therefore, if our domain, Ω, is polygonal, we can exactly triangulate the domain. That is, $\Omega = \bigcup_{K\in\mathcal{T}_h} K$.

However, since we will be using the first order Lagrange finite element for the approximation space, this means that the geometric and the approximation finite elements will coincide. Nevertheless, we will still make the distinction between the two elements. For the numerical method described in Chapters 4 and 5 we use the following finite element approximation space for the discretization of the PDE,

$$P(\mathcal{T}_h) := \{v \in C^0(\Omega) : v \circ \boldsymbol{T}_K \in \widehat{P},\ \forall K \in \mathcal{T}_h\}, \tag{3.2}$$

where $\mathcal{T}_h$ consists of either simplices or quadrangles, $\widehat{P} = \mathbb{P}_1(\widehat{K})$ or $\widehat{P} = \mathbb{Q}_1(\widehat{K})$ (that is, first order multivariate polynomials defined on the reference element $\widehat{K}$), and $\boldsymbol{T}_K$ is defined in (3.1). Let $\{\boldsymbol{x}_i\}_{i\in\mathcal{V}}$ be the collection of nodes of our mesh $\mathcal{T}_h$, where $\mathcal{V}$ is the index set for the nodes. A basis for $P(\mathcal{T}_h)$ is given by $\operatorname{span}\{\varphi_i(\boldsymbol{x})\}_{i\in\mathcal{V}}$, where

$$\varphi_i(\boldsymbol{x}) = \begin{cases} \widehat{\theta}_{\texttt{j\_dof}^{-1}(i)}(\boldsymbol{T}_K^{-1}(\boldsymbol{x})), & \text{if } \boldsymbol{x} \in K \in \mathcal{T}_h(i) \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

where $\mathcal{T}_h(i) \subset \mathcal{T}_h$ such that $K \in \mathcal{T}_h$ contains the vertex $\{\boldsymbol{x}_i\}$ and $\texttt{j\_dof} : \mathcal{T}_h \times \widehat{\eta}_{\text{geo}} \to \mathcal{V}$ and is defined so that `j_dof` identifies the corresponding global degree of freedom from a cell $K \in \mathcal{T}_h$ and local node $j \in \widehat{\eta}_{\text{geo}}$ from the reference element. Note that $\texttt{j\_dof}^{-1}(i)$ is well defined since the element K is fixed. The φ_i are often referred to as *tent functions* based on their shape in two dimensions, see Figure 3.1. These tent functions share the nice property that $\varphi_i(\boldsymbol{x}) \geq 0$ for all $i \in \mathcal{V}$ and that they form a *partition of unity*; that is, $\sum_{i\in\mathcal{V}} \varphi_i(\boldsymbol{x}) = 1$.

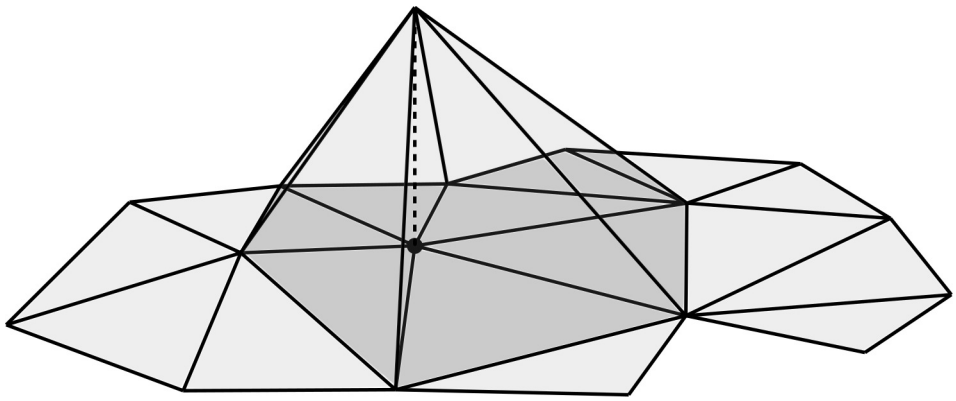

Figure 3.1: A typical $\mathbb{P}_1$ basis function.

3.1.1 Semi-discrete Scheme

Numerically solving the Euler equations is done through the conserved variables: ρ, $\boldsymbol{m}$, and E. The semi-discretization of these variables is given as,

$$\rho_h(\boldsymbol{x},t) := \sum_{i\in\mathcal{V}} \rho_i(t)\varphi_i(\boldsymbol{x}), \quad \boldsymbol{m}_h(\boldsymbol{x},t) = \sum_{i\in\mathcal{V}} \mathsf{M}_i(t)\varphi_i(\boldsymbol{x}), \quad E_h(\boldsymbol{x},t) = \sum_{i\in\mathcal{V}} \mathsf{E}_i(t)\varphi_i(\boldsymbol{x}). \tag{3.4}$$

The unknowns are the $\rho_i(t)$, $\mathsf{M}_i(t)$, and $\mathsf{E}_i(t)$, and the global shape functions $\varphi_i(\boldsymbol{x})$ are the usual tent functions which for a basis for our finite dimensional space $P(\mathcal{T}_h)$. This approximation can be written in the more compact form,

$$\boldsymbol{u}_h(\boldsymbol{x},t) := \sum_{i\in\mathcal{V}} \boldsymbol{u}_i(t)\varphi_i(\boldsymbol{x}), \tag{3.5}$$

where $\boldsymbol{u}_i(t) = (\varrho_i(t), \mathsf{M}_i(t), \mathsf{E}_i(t))^{\mathsf{T}}$. Thus, our problem can be written as,

$$\partial_t \boldsymbol{u}_h + \nabla\cdot \mathbb{f}(\boldsymbol{u}_h) = \boldsymbol{0}. \tag{3.6}$$

3.2 The Fully Discrete Scheme

For computer implementation, it is not feasible to solve the semi-discrete problem. So we further discretize in time. We approximate the time derivative with the forward Euler

method.

Remark 3.2.1 (Time Stepping Methods)**.** In the numerical results the actual time stepping is performed using the strong stability preserving 3rd order Runge-Kutta method (SSP RK3). However, a much more efficient time stepping method has been introduced in Ern & Guermond [54]. It uses an explicit Runge-Kutta method which applies a nonlinear limiting process on the high order update at each stage of the method. This allows for a less restrictive time step while still being conservative and invariant-domain preserving (see Definition 3.2.1). □

Let $t^{n+1} := t^n + \Delta t$ where t^n is the time at the nth time step and Δt is the time step. Let $\boldsymbol{u}_h^n(\boldsymbol{x}) := \boldsymbol{u}_h(\boldsymbol{x}, t^n)$ be the approximate solution at time t^n and $\mathbf{U}_i^n := \boldsymbol{u}_i(t^n)$ be the coefficient of the ith basis function $\varphi_i(\boldsymbol{x})$ at time t^n. That is, $\mathbf{U}_i^n = (\rho_i(t^n), \mathbf{M}_i(t^n), \mathsf{E}_i(t^n))^\mathsf{T} =: (\rho_i^n, \mathbf{M}_i^n, \mathsf{E}_i^n)^\mathsf{T}$, hence

$$\boldsymbol{u}_h^n(\boldsymbol{x}) = \sum_{i \in \mathcal{V}} \mathbf{U}_i^n \varphi_i(\boldsymbol{x}). \tag{3.7}$$

As will be necessary later on, we approximate the flux, $\mathbb{f}(\boldsymbol{u}_h^n)$ by projecting it onto the discrete finite element space. We define this projection $\Pi_h : C^0(\mathbb{R}^{d+2}; \mathbb{R}^{(d+2)\times d}) \to [P(\mathcal{T}_h)]^{(d+2)\times d}$ by,

$$\mathbb{f}(\boldsymbol{u}_h^n) \approx \Pi_h \mathbb{f}(\boldsymbol{u}_h^n) = \sum_{i \in \mathcal{V}} \mathbb{f}(\mathbf{U}_i^n) \varphi_i(\boldsymbol{x}). \tag{3.8}$$

Putting this altogether with the forward Euler method, we arrive at the following numerical method: find $\boldsymbol{u}_h^{n+1}$ such that,

$$\frac{\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n}{\Delta t} + \nabla \cdot (\Pi_h \mathbb{f}(\boldsymbol{u}_h^n)) = \mathbf{0}. \tag{3.9}$$

Solving this equation is done in the weak sense by testing the equation with φ_i for all $i \in \mathcal{V}$. That is, multiply equation (3.9) by φ_i and integrate over D. Doing so gives the

following system of $\text{card}(\mathcal{V})$ equations,

$$\frac{1}{\Delta t_n}\sum_{j\in\mathcal{J}(i)}(\mathbf{U}_j^{n+1}-\mathbf{U}_j^n)m_{ij}+\sum_{j\in\mathcal{J}(i)}\mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}=\mathbf{0},\qquad\text{for all } i\in\mathcal{V}, \tag{3.10}$$

where

$$m_{ij}:=\int_D\varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,d\boldsymbol{x}, \tag{3.11}$$

$$\boldsymbol{c}_{ij}:=\int_D\varphi_i(\boldsymbol{x})\nabla\varphi_j(\boldsymbol{x})\,d\boldsymbol{x}. \tag{3.12}$$

Equation (3.10) is the so-called **Galerkin method**, see Ern & Guermond [52, Sec. 26.1]. Unfortunately, this method can be unstable if any discontinuities develop in the solution, a method for resolving this issue is to introduce an artificial viscosity, see von Neumann & Richtmyer [3].

Remark 3.2.2 (Partition of Unity)**.** Recall that the basis functions $\{\varphi_i\}_{i\in\mathcal{V}}$ form a partition of unity hence $\sum_{j\in\mathcal{V}}m_{ij}=\sum_{j\in\mathcal{J}(i)}m_{ij}=\int_\Omega\varphi_i(\boldsymbol{x})\,d\boldsymbol{x}$. We refer to $m_i:=\int_\Omega\varphi_i(\boldsymbol{x})\,d\boldsymbol{x}$ as the *lumped mass.* Furthermore, from the partition of unity, we have that $\sum_{j\in\mathcal{V}}\boldsymbol{c}_{ij}=\sum_{j\in\mathcal{J}(i)}\boldsymbol{c}_{ij}=0$. □

Remark 3.2.3 (Stability of the Galerkin Approximation)**.** It is well known that the Galerkin approximation is stable as long as $\boldsymbol{u}_0(\boldsymbol{x})$ is smooth and the solution $\boldsymbol{u}(\boldsymbol{x},t)$ remains smooth up to some final time, t_{final}. Furthermore, for $\mathbb{P}_1$ or $\mathbb{Q}_1$ continuous finite elements, we can achieve second order convergence. However, if the solution develops a discontinuity, then the approximate solution produces wild non-physical oscillations. □

The method by which we resolve this issue is to supply a *graph viscosity.* That is, we modify the scheme, (3.10), to be,

$$\frac{1}{\Delta t_n}\sum_{j\in\mathcal{J}(i)}m_{ij}(\mathbf{U}_i^{n+1}-\mathbf{U}_i^n)+\sum_{j\in\mathcal{J}(i)}\mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}-\sum_{j\in\mathcal{J}(i)}d_{ij}^n(\mathbf{U}_j^n-\mathbf{U}_i^n)=\mathbf{0}, \tag{3.13}$$

for all $i \in \mathcal{V}$. We also assume that d_{ij}^n satisfies the following properties,

$$d_{ij}^n \geq 0 \quad \text{for } i \neq j, \quad d_{ij}^n = d_{ji}^n, \quad \text{and } d_{ii} := -\sum_{j \in \mathcal{J}(i)} d_{ij}^n. \tag{3.14}$$

This specific discrete scheme with the graph viscosity is first introduced by Guermond & Popov in [16, Sec. 3.2]. We now introduce the definition of an invariant-domain preserving method.

Definition 3.2.1 (Invariant-Domain Preserving Method)**.** Let $\mathcal{B}$ be a convex invariant set. For $\mathbf{U}_i^0 \in \mathcal{B}$ for all $i \in \mathcal{V}$, if the updated states satisfy $\mathbf{U}_i^n \in \mathcal{B}$ for all $i \in \mathcal{V}$ and $n \in \mathbb{N}$, then we say the numerical method which provides the states $\{\mathbf{U}_i^n\}_{i \in \mathcal{V}}$ is said to be *invariant-domain preserving.* □

4. THE FIRST ORDER APPROXIMATION*

In this chapter, we outline the first order method and prove that it is invariant domain preserving for an arbitrary equation of state. We first approximate the mass matrix, $\{m_{ij}\}_{i,j\in\mathcal{V}}$, by the lumped mass matrix, $\{m_i\}_{i\in\mathcal{V}}$. We can rewrite, (3.13), as an explicit equation,

$$\mathbf{U}_i^{\mathsf{L},n+1} = \mathbf{U}_i^n - \frac{\Delta t}{m_i}\Big(\sum_{j\in\mathcal{G}(i)} \mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij} - \sum_{j\in\mathcal{G}(i)} d_{ij}^n(\mathbf{U}_j^n - \mathbf{U}_i^n)\Big) \quad \text{for all } i\in\mathcal{V}. \tag{4.1}$$

Note we also use $^{\mathsf{L}}$ superscript to signify that this is the low order update. This will be necessary for distinguishing between the high order update in the chapters to come.

Remark 4.0.1 (Discretization Independent Method)**.** Note that the numerical method described in (4.1) can be made discretization independent. The quantities m_i and $\boldsymbol{c}_{ij}$ will vary depending on the specific discretization. More on this can be found in Guermond et. al. [55]. □

4.1 Invariant Domain Preserving Method

With a bit of rearrangement and using the fact that $\sum_{j\in\mathcal{G}(i)}\boldsymbol{c}_{ij} = 0$, we rewrite (4.1) as a convex combination of states under the CFL condition, (4.7),

$$\mathbf{U}_i^{\mathsf{L},n+1} = \Big(1 - \sum_{j\in\mathcal{G}(i)\setminus\{i\}} \frac{2\Delta t_n d_{ij}^{\mathsf{L},n}}{m_i}\Big)\mathbf{U}_i^n + \sum_{j\in\mathcal{G}(i)\setminus\{i\}} \frac{2\Delta t_n d_{ij}^{\mathsf{L},n}}{m_i}\overline{\mathbf{U}}_{ij}^n\Big(\frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2d_{ij}^{\mathsf{L},n}}\Big), \tag{4.2}$$

where

$$\overline{\mathbf{U}}_{ij}^n(t) = \frac{1}{2}(\mathbf{U}_i^n + \mathbf{U}_j^n) - t(\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n))\frac{\boldsymbol{c}_{ij}}{\|\boldsymbol{c}_{ij}\|_{\ell^2}}. \tag{4.3}$$

Throughout this thesis, we refer to $\overline{\mathbf{U}}_{ij}^n$ as the **bar states**.

* A majority of this chapter is a modification of the work done in [1] and is reprinted with permission from [1].

The bar states are a fundamental ingredient in proving that the numerical method is invariant domain preserving (see Definition 3.2.1). Since for a large enough $d_{ij}^{\mathsf{L},n}$, we claim that $\overline{\mathbf{U}}_{ij}^n \in \mathcal{B}(b)$. We define $d_{ij}^{\mathsf{L},n}$ to be,

$$d_{ij}^{\mathsf{L},n} := \max\{\lambda_{\max}(\mathbf{U}_i^n, \mathbf{U}_j^n, \boldsymbol{n}_{ij})\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{\max}(\mathbf{U}_j^n, \mathbf{U}_i^n, \boldsymbol{n}_{ji})\|\boldsymbol{c}_{ji}\|_{\ell^2}\}, \tag{4.4}$$

for $i \neq j$, where $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$ and $\lambda_{\max}(\mathbf{U}_i^n, \mathbf{U}_j^n, \boldsymbol{n}_{ij})$ is the max wave speed to the local Riemann problem,

$$\partial_t \boldsymbol{u} + \partial_x(\mathbb{f}(\boldsymbol{u})\boldsymbol{n}_{ij}) = \mathbf{0}, \qquad \boldsymbol{u}_0(x) = \begin{cases} \mathbf{U}_i^n, & \text{if } x < 0, \\ \mathbf{U}_j^n, & \text{if } x > 0. \end{cases} \tag{4.5}$$

Notice that the bar states, (4.3), are the average of the solution to the Riemann problem, (4.5), at the "artificial" time $t = \frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2d_{ij}^{\mathsf{L},n}} \leq \frac{1}{2\lambda_{\max}}$; see Theorem 1.3.1.

Remark 4.1.1 (Choice of Larger Artificial Viscosity)**.** Notice that, taking larger values of $d_{ij}^{\mathsf{L},n}$; that is, finding an upper estimate on the maximum wave speed, still preserves the desired structure of the bar states. Since if $\widehat{\lambda}_{\max}$ is an upper bound on $\lambda_{\max}$ and $\widehat{d}_{ij}^{\mathsf{L},n}$ is simply $d_{ij}^{\mathsf{L},n}$ with $\lambda_{\max}$ replaced with $\widehat{\lambda}_{\max}$, then

$$t = \frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2\widehat{d}_{ij}^{\mathsf{L},n}} \leq \frac{1}{2\widehat{\lambda}_{\max}} \leq \frac{1}{2\lambda_{\max}} \tag{4.6}$$

Hence the bar states, (4.3), are still the averages of the Riemann solution for (4.5) at a different "fake" time. □

Remark 4.1.2 (Justification for the Flux Approximation)**.** Recall that the approximation of $\mathbb{f}(\boldsymbol{u}_h^n)$ is $\sum_{i\in\mathcal{V}} \mathbb{f}(\mathbf{U}_i^n)\varphi_i(\boldsymbol{x})$ which was defined in (3.8). This approximation was necessary to justify the bar states as being the average of the solution to the Riemann problem (4.5). □

The challenge is now to determine $\lambda_{\max}$ for an arbitrary EOS or a sufficiently close upper bound. This is not a simple task and will be investigated in Chapter 4. Note that, once $d_{ij}^{\mathsf{L},n}$

has been determined, this gives us the following CFL constraint which guarantees a convex combination in of states in (4.2),

$$\Delta t \leq \frac{m_i}{2\sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^{\mathsf{L},n}} = -\frac{m_i}{2d_{ii}^{\mathsf{L},n}}, \quad \text{for all } i \in \mathcal{V}. \tag{4.7}$$

Theorem 4.1.1 (Invariant-Domain Preservation)**.** *[16]] If $d_{ij}^{\mathsf{L},n}$ is defined by* (4.4) *and $\mathbf{U}_i^n \in \mathcal{B}(b)$ for all $i \in \mathcal{V}$, then the update provided by* (4.1) *is invariant domain preserving under the CFL condition, $\Delta t \leq \min_{i\in\mathcal{V}}(-\frac{m_i}{2d_{ii}^{\mathsf{L},n}})$.*

Proof. From the choice of $d_{ij}^{\mathsf{L},n}$ we have from Theorem 1.3.1 that $\overline{\mathbf{U}}_{ij}^n \in \mathcal{B}(b)$ for each $j \in \mathcal{I}(i) \setminus \{i\}$ and from the assumption, $\mathbf{U}_i^n \in \mathcal{B}(b)$. Thus the update provided in (4.2) is a convex combination of states in the convex set $\mathcal{B}(b)$ (under the CFL condition (4.7)). Hence $\mathbf{U}_i^{n+1} \in \mathcal{B}(b)$. □

4.2 Extended Riemann Problem

If the pressure is given by the ideal gas law, then the solution to the Riemann problem an exact self-similar weak solution can be determined. This was originally done by Lax in [34]; for other useful resources regarding the exact solution to this Riemann problem, see [56, Chapter 4], [31, Chapter II, Section 3], and [57, Section 5.6]. However, for an arbitrary or tabulated equations of state, computing the max wave speed, $\lambda_{\max}$, is extremely difficult if not impossible. In order to solve this issue, we propose extending the Riemann problem with an auxiliary equation in terms of a new variable, Γ. This idea is motivated by the paper by Abgrall & Karni [58] in the context of multi-fluids. We also change the EOS given by the oracle to an equation of state based on the Noble-Abel Stiffened-Gas (NASG) EOS which interpolates the left and right pressure, p_L and p_R, respectively. Recall the NASG EOS is defined by $p(\rho, e) := (\gamma - 1)\frac{\rho(e-q)}{1-b\rho} + \gamma p_\infty$. Alternatively, we write this equation of state as, $p + p_\infty = (\gamma - 1)(\frac{\rho(e-q)}{1-b\rho} - p_\infty)$. This method is first introduced in [2] but uses the covolume EOS instead of the NASG EOS.

Definition 4.2.1. The extended Riemann problem is, $\partial_t \widetilde{\boldsymbol{u}} + \partial_x(\widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}})\boldsymbol{n}) = \mathbf{0}$, where

$$\widetilde{\boldsymbol{u}} := (\boldsymbol{u}, \Gamma)^{\mathsf{T}} = \begin{pmatrix} \rho \\ m \\ \boldsymbol{m}^{\perp} \\ \mathcal{E} \\ \Gamma \end{pmatrix}, \qquad \widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}})\boldsymbol{n} := \begin{pmatrix} m \\ \frac{m^2}{\rho} + \widetilde{p}_{\mathsf{nasg}}(\widetilde{\boldsymbol{u}}) \\ \frac{m}{\rho}\boldsymbol{m}^{\perp} \\ \frac{m}{\rho}(\mathcal{E} + \widetilde{p}_{\mathsf{nasg}}(\widetilde{\boldsymbol{u}})) \\ \frac{m}{\rho}\Gamma \end{pmatrix}, \tag{4.8}$$

with left and right data $\widetilde{\mathbf{U}}_{\mathrm{Z}} := (\mathbf{U}_{\mathrm{Z}}, \Gamma_{\mathrm{Z}}) = (\rho_{\mathrm{Z}}, \boldsymbol{m}_{\mathrm{Z}} \cdot \boldsymbol{n}, \boldsymbol{m}_{\mathrm{Z}}^{\perp}, \mathcal{E}_{\mathrm{Z}}, \Gamma_{\mathrm{Z}})^{\mathsf{T}}$ with

$$\Gamma_Z := \rho_Z \left(\frac{p_Z + p_\infty}{\frac{\rho_Z(\mathsf{e}(\mathbf{U}_Z) - q)}{1 - b\rho_Z} - p_\infty} + 1 \right) \tag{4.9}$$

where $Z \in \{L, R\}$ and the pressure is defined by

$$\widetilde{p}_{\mathsf{nasg}}(\widetilde{\boldsymbol{u}}) := \Big(\frac{\Gamma}{\rho} - 1\Big)\Big(\frac{\mathcal{E} - \frac{1}{2}m^2/\rho - \rho q}{1 - b\rho} \ - p_\infty\Big) + \frac{\Gamma}{\rho} p_\infty. \tag{4.10}$$

□

It is often easier to work with the primitive variables, so let $\gamma := \Gamma/\rho$ then $\widetilde{p}_{\mathsf{nasg}}$ can be rewritten as $p_{\mathsf{nasg}}(\rho, e, \gamma) := \widetilde{p}_{\mathsf{nasg}}(\widetilde{\boldsymbol{u}}) = (\gamma - 1)\big(\frac{\rho(e-q)}{1-b\rho} - p_\infty\big) - p_\infty$ which is simply the NASG EOS with a variable γ. The left and right states are then $(\rho_{\mathrm{Z}}, \boldsymbol{v}_{\mathrm{Z}} \cdot \boldsymbol{n}, \boldsymbol{v}_{\mathrm{Z}}^{\perp}, p_{\mathrm{Z}}, \gamma_{\mathrm{Z}})$ where $\gamma_{\mathrm{Z}} := \frac{(p_{\mathrm{Z}} + p_\infty)(1 - b\rho_{\mathrm{Z}})}{\rho_{\mathrm{Z}}(e(\rho_{\mathrm{Z}}, p_{\mathrm{Z}}) - q) - p_\infty(1 - b\rho_{\mathrm{Z}})} + 1$ and $\boldsymbol{v}_{\mathrm{Z}}^{\perp} := \rho_{\mathrm{Z}}^{-1}\boldsymbol{m}_{\mathrm{Z}}^{\perp}$ for $\mathrm{Z} \in \{\mathrm{L}, \mathrm{R}\}$.

Remark 4.2.1 (Interpolating with p_{nasg})**.** The reason for this choice of γ_{Z} is because $p_{\mathsf{nasg}}(\rho_{\mathrm{Z}}, e(\rho_{\mathrm{Z}}, p_{\mathrm{Z}}), \gamma_{\mathrm{Z}}) = p_{\mathrm{Z}}$ for $\mathrm{Z} \in \{\mathrm{L}, \mathrm{R}\}$. That is, p_{nasg} interpolates the left and right pressures. □

Definition 4.2.2 (Extended Invariant Domain)**.** We define the extended invariant domain as,

$$\widetilde{\mathcal{B}}(b, q, p_\infty) := \{\widetilde{\boldsymbol{u}} \in \mathbb{R}^{d+3} : \boldsymbol{u} \in \mathcal{B}(b, q, p_\infty),\ \Gamma > \rho\}. \tag{4.11}$$

to account for the new variable Γ. $\square$

Remark 4.2.2 (Extended Bar States)**.** Notice that $\widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}}_{\mathrm{Z}}) = (\mathbb{f}(\boldsymbol{u}_{\mathrm{Z}}), \boldsymbol{v}_{\mathrm{Z}}\Gamma_{\mathrm{Z}})^{\mathsf{T}}$ since $\widetilde{p}_{\mathsf{nasg}}(\widetilde{\boldsymbol{u}}_{\mathrm{Z}}) = p_{\mathrm{Z}} = p(\boldsymbol{u}_{\mathrm{Z}})$. Let $\overline{\boldsymbol{u}}_{\mathrm{LR}} := \overline{\mathbf{U}}_{ij}^{n}$. Then, the bar state for the extended Riemann problem is,

$$\overline{\widetilde{\boldsymbol{u}}}_{\mathrm{LR}} = \begin{pmatrix} \overline{\boldsymbol{u}}_{\mathrm{LR}} \\ \frac{1}{2}(\Gamma_{\mathrm{L}} + \Gamma_{\mathrm{R}}) - \frac{1}{2\lambda}(\boldsymbol{v}_{\mathrm{R}}\Gamma_{\mathrm{R}} - \boldsymbol{v}_{\mathrm{L}}\Gamma_{\mathrm{L}}) \cdot \boldsymbol{n} \end{pmatrix}, \tag{4.12}$$

where $\lambda = -\frac{m_i}{2d_{ii}^{\mathrm{L},n}}$. Its important to note that the density, momentum, and total energy of the state $\overline{\widetilde{\boldsymbol{u}}}_{\mathrm{LR}}$ is the same as the state $\overline{\boldsymbol{u}}_{\mathrm{LR}}$. Thus, if we can prove positivity of the density and specific internal energy of the extended Riemann problem, then this immediately carries over to the original Riemann problem. This is remark is critical in the justification of the invariant-domain preserving method. $\square$

4.2.1 The Wave Structure

We begin by first deriving the wave structure of this Riemann problem. This is done by computing the Jacobian matrix of $\widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}})\boldsymbol{n}$. However, the computation is simpler if we make a change of variables. Let $\theta : \widetilde{\mathcal{B}} \subset \mathbb{R}^{d+3} \to \widetilde{\mathcal{B}}$ be a smooth diffeomorphism, defined by

$$\theta(\widetilde{\boldsymbol{u}}) = \left(\rho, \frac{m}{\rho}, \frac{\boldsymbol{m}^{\perp}}{\rho}, \mathsf{e}(\boldsymbol{u}), \frac{\Gamma}{\rho}\right)^{\mathsf{T}} =: \widetilde{\boldsymbol{w}}. \tag{4.13}$$

That is, θ maps to the primitive variables, $\theta(\widetilde{\boldsymbol{u}}) = \widetilde{\boldsymbol{w}} = (\rho, v, \boldsymbol{v}^{\perp}, e, \gamma)^{\mathsf{T}}$. Then the extended Riemann problem is formulated as,

$$\partial_t \rho + v\partial_x \rho + \rho \partial_x v = 0, \tag{4.14a}$$

$$\partial_t v + v\partial_x v + \rho^{-1}\partial_x p_{\mathsf{nasg}} = 0, \tag{4.14b}$$

$$\partial_t \boldsymbol{v}^{\perp} + v\partial_x \boldsymbol{v}^{\perp} = 0, \tag{4.14c}$$

$$\partial_t e + \rho^{-1} p_{\mathsf{nasg}} \partial_x v + v\partial_x e = 0, \tag{4.14d}$$

$$\partial_t \gamma + v\partial_x \gamma = 0, \tag{4.14e}$$

Thus, the Jacobian matrix of the system (4.14a)–(4.14e), is,

$$\mathbb{B}(\widetilde{\boldsymbol{w}}) = \begin{pmatrix} v & \rho & \mathbf{0}^{\mathsf{T}} & 0 & 0 \\ \frac{1}{\rho}\frac{\partial p_{\mathsf{nasg}}}{\partial \rho} & v & \mathbf{0}^{\mathsf{T}} & \frac{1}{\rho}\frac{\partial p_{\mathsf{nasg}}}{\partial e} & \frac{1}{\rho}\frac{\partial p_{\mathsf{nasg}}}{\partial \gamma} \\ \mathbf{0} & \mathbf{0} & v\mathbb{I}_{d-1} & \mathbf{0} & \mathbf{0} \\ 0 & \frac{p_{\mathsf{nasg}}}{\rho} & \mathbf{0}^{\mathsf{T}} & v & 0 \\ 0 & 0 & \mathbf{0}^{\mathsf{T}} & 0 & v \end{pmatrix}. \tag{4.15}$$

The eigenvalues of this matrix are, $\mu_1(\widetilde{\boldsymbol{w}}) = v - \widetilde{a}(\widetilde{\boldsymbol{w}})$, $\mu_2(\widetilde{\boldsymbol{w}}) = v$ (with multiplicity $d+1$), and $\mu_3(\widetilde{\boldsymbol{w}}) = v + \widetilde{a}(\widetilde{\boldsymbol{w}})$, where

$$\widetilde{a}(\widetilde{\boldsymbol{w}})^2 = \frac{\gamma(\widetilde{p}_{\mathsf{nasg}}(\theta^{-1}(\widetilde{\boldsymbol{w}})) + p_\infty)}{\rho(1-b\rho)} = \gamma(\gamma-1)\Big(\frac{e-q}{(1-b\rho)^2} - \frac{p_\infty}{\rho(1-b\rho)}\Big). \tag{4.16}$$

Thus, the solution to the Riemann problem is composed of 3 waves. We denote them by the L-wave, C-wave, and R-wave.

Remark 4.2.3 (No Entropy for p_{nasg})**.** The interpolatory pressure, p_{nasg}, is not a real equation of state since it is also a function of the extra variable, γ. Therefore, we have not made a change of variable to the specific entropy which is usually done when computing the eigenvalues of the Jacobian matrix. However, in Section 4.3, the restriction of p_{nasg} to each wave, will define an equation of state. □

Note that the change of variables does not affect the eigenvalues of the Jacobian matrix of the original system. That is, if $\mathbb{A}(\widetilde{\boldsymbol{u}})$ is the Jacobian matrix for the flux, $\widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}})\boldsymbol{n}$, given in (4.8), then the eigenvalues of $\mathbb{A}(\widetilde{\boldsymbol{u}})$ and $\mathbb{B}(\widetilde{\boldsymbol{w}})$ are the same. Furthermore, let $\boldsymbol{r}_i(\widetilde{\boldsymbol{u}})$ denote an eigenvector of $\mathbb{A}(\widetilde{\boldsymbol{u}})$ with corresponding eigenvalue, $\lambda_i(\widetilde{\boldsymbol{u}})$ and $\boldsymbol{s}_i(\widetilde{\boldsymbol{w}})$ an eigenvector of $\mathbb{B}(\widetilde{\boldsymbol{w}})$, with its corresponding eigenvalue, $\mu_i(\widetilde{\boldsymbol{w}})$. Note also, eigenvectors are related by the following identity, $\boldsymbol{r}_i(\widetilde{\boldsymbol{u}}) = (D_{\widetilde{\boldsymbol{u}}}\theta(\widetilde{\boldsymbol{u}}))^{-1}\boldsymbol{s}_i(\theta(\widetilde{\boldsymbol{u}}))$. Then it can be shown that, $D\lambda_i(\widetilde{\boldsymbol{u}})\cdot\boldsymbol{r}_i(\widetilde{\boldsymbol{u}}) = D\mu_i(\widetilde{\boldsymbol{w}})\cdot\boldsymbol{s}_i(\widetilde{\boldsymbol{w}})$. The details on the change of variables regarding the Jacobian matrix can be found in [31, Section 2.1.1]

We now present the left and right eigenvectors for completeness. Note, that the associated eigenvectors are $\boldsymbol{s}_1(\widetilde{\boldsymbol{w}}) = (1, -a/\rho, \mathbf{0}_{d-1}, 0, 0)^\mathsf{T}$, $\boldsymbol{s}_3(\widetilde{\boldsymbol{w}}) = (1, a/\rho, \mathbf{0}_{d-1}, 0, 0)^\mathsf{T}$, and

$$\boldsymbol{s}_2(\widetilde{\boldsymbol{w}}) \in \operatorname{span}\left\{ \begin{pmatrix} 1 \\ 0 \\ \mathbf{0}_{d-1} \\ -\frac{e-q}{\rho(1-b\rho)} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \mathbb{e}_1 \\ 0 \\ 0 \end{pmatrix}, \cdots, \begin{pmatrix} 0 \\ 0 \\ \mathbb{e}_{d-1} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\rho(e-q)}{1-b\rho} - p_\infty \\ 0 \\ \mathbf{0}_{d-1} \\ 0 \\ -\frac{(\gamma-1)(e-q)}{(1-b\rho)^2}. \end{pmatrix} \right\}, \tag{4.17}$$

where $\mathbb{e}_i$ are the standard basis vectors in $\mathbb{R}^{d-1}$. We have chosen specific eigenvectors for $\boldsymbol{s}_1$ and $\boldsymbol{s}_3$ since the dimension of their respective eigenspaces is only one.

Lemma 4.2.1 (Wave Structure of the Extended Riemann Problem)**.** *The* L- *and* R-*waves are genuinely nonlinear and the C-wave is linearly degenerate. That is,* $D\lambda_i(\widetilde{\boldsymbol{u}}) \cdot \boldsymbol{r}_i(\widetilde{\boldsymbol{u}}) \neq 0$ *for* $i = 1, 3$ *and* $D\lambda_2(\widetilde{\boldsymbol{u}}) \cdot \boldsymbol{r}_2(\widetilde{\boldsymbol{u}}) = 0$ *for all* $\widetilde{\boldsymbol{u}} \in \widetilde{\mathcal{B}}$*.*

Proof. As mentioned above regarding the change of variables, we can equivalently work with eigenpairs $(\mu_i, \boldsymbol{s}_i)$. Then we see that $D\mu_1(\widetilde{\boldsymbol{w}}) \cdot \boldsymbol{s}_1(\widetilde{\boldsymbol{w}}) = -\frac{\partial a}{\partial \rho}(\widetilde{\boldsymbol{w}}) - \frac{a(\widetilde{\boldsymbol{w}})}{\rho} \neq 0$ and $D\mu_3(\widetilde{\boldsymbol{w}}) \cdot \boldsymbol{s}_3(\widetilde{\boldsymbol{w}}) = \frac{\partial a(\widetilde{\boldsymbol{w}})}{\partial \rho} + \frac{a(\widetilde{\boldsymbol{w}})}{\rho} \neq 0$ for all $\widetilde{\boldsymbol{w}} \in \widetilde{\mathcal{B}}$. Similarly, $D\mu_2(\widetilde{\boldsymbol{w}}) = (0, 1, \mathbf{0}_{d-1}, 0, 0)^\mathsf{T}$ and is orthogonal to the space of eigenvectors defined in (4.17). □

Lemma 4.2.2 (Continuity on the Contact)**.** *The pressure and velocity are continuous across the contact.*

Proof. To prove this we show that p and v are 2-Riemann invariants and hence are continuous across the contact. Note the following identity holds for a i-characteristic wave,

$$\begin{aligned} D_{\widetilde{\boldsymbol{u}}} p(\widetilde{\boldsymbol{w}}) \cdot \boldsymbol{r}_i(\widetilde{\boldsymbol{u}}) &= D_{\widetilde{\boldsymbol{u}}} p(\theta(\widetilde{\boldsymbol{u}})) \cdot (D_{\widetilde{\boldsymbol{u}}} \theta(\widetilde{\boldsymbol{u}}))^{-1} \boldsymbol{s}_i(\theta(\widetilde{\boldsymbol{u}})) \\ &= \left((D_{\widetilde{\boldsymbol{w}}} p(\widetilde{\boldsymbol{w}}))^\mathsf{T} D_{\widetilde{\boldsymbol{u}}} \theta(\widetilde{\boldsymbol{u}})\right) \cdot \left((D_{\widetilde{\boldsymbol{u}}} \theta(\widetilde{\boldsymbol{u}}))^{-1} \boldsymbol{s}_i(\widetilde{\boldsymbol{w}})\right) \\ &= D_{\widetilde{\boldsymbol{w}}} p(\widetilde{\boldsymbol{w}}) \cdot \boldsymbol{s}_i(\widetilde{\boldsymbol{w}}) \end{aligned}$$

Therefore, we only need to show that $D_{\widetilde{\boldsymbol{w}}}\boldsymbol{p}(\widetilde{\boldsymbol{w}}) \cdot \boldsymbol{s}_2(\widetilde{\boldsymbol{w}}) = 0$ for $\boldsymbol{p}$ to be a 2-Riemann invariant. Computing the derivative of $\boldsymbol{p}$, we have,

$$D_{\widetilde{\boldsymbol{w}}}\boldsymbol{p}(\widetilde{\boldsymbol{w}}) = \Big(\frac{(\gamma-1)(e-q)}{(1-b\rho)^2}, 0, \mathbf{0}^\mathsf{T}, \frac{(\gamma-1)\rho}{1-b\rho}, \frac{\rho(e-q)}{1-b\rho} - p_\infty\Big). \tag{4.18}$$

Then it is a quick check to see that $D_{\widetilde{\boldsymbol{w}}}\boldsymbol{p}(\widetilde{\boldsymbol{w}}) \cdot \boldsymbol{s}_2(\widetilde{\boldsymbol{w}}) = 0$ for any $\boldsymbol{s}_2(\widetilde{\boldsymbol{w}})$ defined in (4.17). Therefore, $\boldsymbol{p}$ is a 2-Riemann invariant and hence is constant across the contact.

For the velocity, v is automatically a 2-Riemann invariant by Lemma 4.2.1, since $\lambda_2(\widetilde{\boldsymbol{u}}) = v$. Thus the velocity is constant across the contact.

□

4.3 The Solution to the Extended Riemann Problem

We suppose that the solution to the extended Riemann problem must be a self similar solution composed of three waves where the Z-wave ($\mathrm{Z} \in \{\mathrm{L}, \mathrm{R}\}$) is either a shock or an expansion and the C-wave is a contact. Since γ is being transported ($\partial_t\gamma + v\partial_x\gamma = 0$), we propose that the $\gamma = \gamma_\mathrm{L}$ left of the contact and $\gamma = \gamma_\mathrm{R}$ right of the contact, Hence, the solution for γ is,

$$\gamma(x,t) = \begin{cases} \gamma_\mathrm{L}, & \frac{x}{t} < v^*, \\ \gamma_\mathrm{R}, & \frac{x}{t} > v^*, \end{cases} \tag{4.19}$$

where v^* is the speed of the contact. See Figure 4.1 for a visual description of this solution. The solution on each wave can be constructed with the NASG EOS for $\gamma = \gamma_\mathrm{Z}$ for each respective wave. For the sake of completion, we derive the solution to this extended Riemann problem.

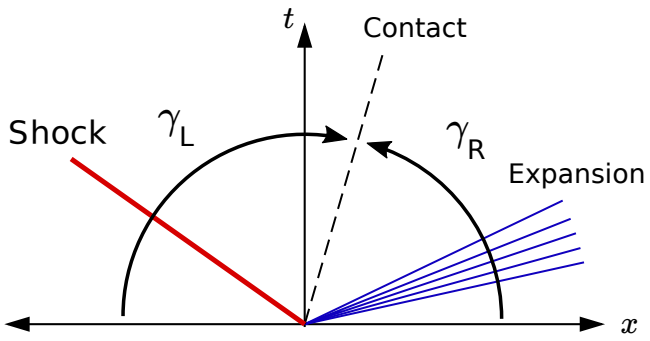

Figure 4.1: An example solution for $\gamma(x,t)$ in the extended Riemann problem

The solution to the Riemann is first constructed for the problem,

$$\partial_t \begin{pmatrix} \rho \\ m \\ \mathcal{E} \\ \Gamma \end{pmatrix} + \partial_x \begin{pmatrix} m \\ \frac{1}{\rho}m^2 + p_{\text{nasg}} \\ \frac{m}{\rho}(\mathcal{E} + p_{\text{nasg}}) \\ \frac{m}{\rho}\Gamma \end{pmatrix}, \tag{4.20}$$

with left and right data, $(\rho_{\mathrm{Z}}, \boldsymbol{m}_{\mathrm{Z}} \cdot \boldsymbol{n}, \mathcal{E}_{\mathrm{Z}}, \Gamma_Z)^{\mathsf{T}}$ for $\mathrm{Z} \in \{\mathrm{L}, \mathrm{R}\}$ where $\mathcal{E} = E - \frac{\|\boldsymbol{m}^{\perp}\|_{\ell^2}^2}{2\rho}$. Then the complete solution is found for $\boldsymbol{m}^{\perp}$ by solving $\partial_t \boldsymbol{m}^{\perp} + \partial_x(v\boldsymbol{m}^{\perp}) = \boldsymbol{0}$, see [16, Sec. 2.5]. However, the solution for $\boldsymbol{m}^{\perp}$ is never needed, as we are only interested in the maximum wave speed to the problem, (4.20). This extended Riemann problem (4.20), is the focus of the remainder of this Chapter.

Remark 4.3.1 (Internal Energy Change of Basis)**.** The internal energy can be written as $\rho e = E - \frac{\|\boldsymbol{m}\|_{\ell^2}^2}{2\rho} = \mathcal{E} - \frac{m^2}{2\rho}$. Hence the internal energy does not depend on the basis, which is what we expect. □

Now let $\boldsymbol{c} := (\rho, v, p, \gamma)^{\mathsf{T}}$ be the primitive state and set $\boldsymbol{c}_{\mathrm{Z}} := (\rho_{\mathrm{Z}}, v_{\mathrm{Z}}, p_{\mathrm{Z}}, \gamma_{\mathrm{Z}})^{\mathsf{T}}$. Recall that $\gamma_{\mathrm{Z}} = \frac{(p_{\mathrm{Z}}+p_\infty)(1-b\rho_{\mathrm{Z}})}{\rho_{\mathrm{Z}}(e_Z - q)} + 1$ and based on the assumption that the oracle provides positive pressure, see Remark 1.4.1, we have that $\min(\gamma_{\mathrm{L}}, \gamma_{\mathrm{R}}) > 1$. Furthermore, notice that the oracle is only invoked to compute the left and right Riemann data p_{L} and p_{R}.

We now define an important function that will appear in the solution to the Riemann problem,

$$f_{\mathrm{Z}}(p) := \begin{cases} f_{\mathrm{Z}}^{\exp}(p) := \frac{2a_{\mathrm{Z}}(1-b\rho_{\mathrm{Z}})}{\gamma_{\mathrm{Z}}-1}\left(\left(\frac{p+p_\infty}{p_{\mathrm{Z}}+p_\infty}\right)^{\frac{\gamma_{\mathrm{Z}}-1}{2\gamma_{\mathrm{Z}}}} - 1\right), & \text{if } -p_\infty \le p < p_{\mathrm{Z}}, \\ f_{\mathrm{Z}}^{\mathrm{shock}}(p) := (p-p_{\mathrm{Z}})\sqrt{\frac{A_{\mathrm{Z}}}{p+p_\infty+B_{\mathrm{Z}}}}, & \text{if } p \ge p_{\mathrm{Z}}. \end{cases} \tag{4.21}$$

where $A_{\mathrm{Z}} := \frac{2(1-b\rho_{\mathrm{Z}})}{(\gamma_{\mathrm{Z}}+1)\rho_Z}$ and $B_{\mathrm{Z}} := \frac{\gamma_{\mathrm{Z}}-1}{\gamma_{\mathrm{Z}}+1}(p_{\mathrm{Z}}+p_\infty)$.

We also introduce the wave speeds for the Riemann problem which are essential for the estimation of the maximum wave speed. They are as follows:

$$\lambda_{\mathrm{L}}^-(p^*) := v_{\mathrm{L}} - a_{\mathrm{L}}\left(1 + \frac{\gamma_{\mathrm{L}}+1}{2\gamma_{\mathrm{L}}}\left(\frac{p^*-p_{\mathrm{L}}}{p_{\mathrm{L}}+p_\infty}\right)_+\right)^{\frac{1}{2}}, \tag{4.22a}$$

$$\lambda_{\mathrm{L}}^+(p^*) := \begin{cases} v_{\mathrm{L}} - f_{\mathrm{L}}(p^*) - a_{\mathrm{L}}\frac{1-b\rho_{\mathrm{L}}}{1-b\rho_{\mathrm{L}}^*}\left(\frac{p^*+p_\infty}{p_{\mathrm{L}}+p_\infty}\right)^{\frac{\gamma_{\mathrm{L}}-1}{2\gamma_{\mathrm{L}}}}, & \text{if } p^* < p_{\mathrm{L}}, \\ \lambda_{\mathrm{L}}^-(p^*), & \text{if } p_{\mathrm{L}} \le p^*, \end{cases} \tag{4.22b}$$

$$\lambda_{\mathrm{R}}^+(p^*) := v_{\mathrm{R}} + a_{\mathrm{R}}\left(1 + \frac{\gamma_{\mathrm{R}}+1}{2\gamma_{\mathrm{R}}}\left(\frac{p^*-p_{\mathrm{R}}}{p_{\mathrm{R}}+p_\infty}\right)_+\right)^{\frac{1}{2}}, \tag{4.22c}$$

$$\lambda_{\mathrm{R}}^-(p^*) := \begin{cases} v_{\mathrm{R}} + f_{\mathrm{R}}(p^*) + a_{\mathrm{R}}\frac{1-b\rho_{\mathrm{R}}}{1-b\rho_{\mathrm{R}}^*}\left(\frac{p^*+p_\infty}{p_{\mathrm{R}}+p_\infty}\right)^{\frac{\gamma_{\mathrm{R}}-1}{2\gamma_{\mathrm{R}}}}, & \text{if } p^* < p_{\mathrm{R}}, \\ \lambda_{\mathrm{R}}^+(p^*), & \text{if } p_{\mathrm{R}} \le p^*, \end{cases} \tag{4.22d}$$

The derivation of these waves speeds is shown in the following sections.

4.3.1 Shock Wave

Assume the L-Wave is a shock. Then the solution with a shock wave must satisfy the Rankine-Hugoniot conditions. Let $\mathcal{S}_{\mathrm{L}}$ denote the speed of the shock and we denote the state across the shock by '$*$L' subscript. Define, $\widehat{v}_{\mathrm{L}} := v_{\mathrm{L}} - \mathcal{S}_{\mathrm{L}}$ and $\widehat{v}_{*\mathrm{L}} := v_{*\mathrm{L}} - \mathcal{S}_{\mathrm{L}}$. Then the

Rankine Hugoniot conditions become,

$$\rho_{\mathrm{L}}\widehat{v}_{\mathrm{L}} = \rho_{*\mathrm{L}}\widehat{v}_{*\mathrm{L}}, \tag{4.23}$$

$$\rho_{\mathrm{L}}\widehat{v}_{\mathrm{L}}^2 + p_{\mathrm{L}} = \rho_{*\mathrm{L}}\widehat{v}_{*\mathrm{L}}^2 + p^*, \tag{4.24}$$

$$\widehat{v}_{\mathrm{L}}(\widehat{E}_{\mathrm{L}} + p_{\mathrm{L}}) = \widehat{v}_{*\mathrm{L}}(\widehat{E}_{*\mathrm{L}} + p^*), \tag{4.25}$$

where $\widehat{E}_{\mathrm{L}} = \rho_{\mathrm{L}}e_{\mathrm{L}} + \frac{1}{2}\rho_{\mathrm{L}}\widehat{v}_{\mathrm{L}}^2$ and $\widehat{E}_{*\mathrm{L}} = \rho_{*\mathrm{L}}e_{*\mathrm{L}} + \frac{1}{2}\rho_{*\mathrm{L}}\widehat{v}_{*\mathrm{L}}^2$. From the Rankine-Hugoniot conditions, one can also derive,

$$e_{*\mathrm{L}} - e_{\mathrm{L}} = \frac{1}{2}(p^* - p_{\mathrm{L}})\Big(\frac{\rho_{*\mathrm{L}} - \rho_{\mathrm{L}}}{\rho_{*\mathrm{L}}\rho_{\mathrm{L}}}\Big), \tag{4.26}$$

see [56, Section 3.1.3]. Note also that this identity is independent of the EOS. The goal is to determine an equation which relates the two unknowns, p^* and $\rho_{*\mathrm{L}}$. So, applying the NASG EOS, we have,

$$\frac{p^* - \gamma_{\mathrm{L}}p_\infty}{\gamma_{\mathrm{L}} - 1}\frac{1 - b\rho_{*\mathrm{L}}}{\rho_{*\mathrm{L}}} - \frac{p_{\mathrm{L}} - \gamma_{\mathrm{L}}p_\infty}{\gamma_{\mathrm{L}} - 1}\frac{1 - b\rho_{\mathrm{L}}}{\rho_{\mathrm{L}}} = \frac{1}{2}(p^* - p_{\mathrm{L}})\Big(\frac{\rho_{*\mathrm{L}} - \rho_{\mathrm{L}}}{\rho_{*\mathrm{L}}\rho_{\mathrm{L}}}\Big). \tag{4.27}$$

Now multiply the equation by $(\gamma_{\mathrm{L}} - 1)\rho_{*\mathrm{L}}$ and expand,

$$\begin{aligned}(p^* + \gamma_{\mathrm{L}}p_\infty) - \rho_{*\mathrm{L}}b(p^* + \gamma_{\mathrm{L}}p_\infty) - (p_{\mathrm{L}} + \gamma_{\mathrm{L}}p_\infty)\frac{\rho_{*\mathrm{L}}}{\rho_{\mathrm{L}}} + \rho_{*\mathrm{L}}b(p_{\mathrm{L}} + \gamma_{\mathrm{L}}p_\infty)& \\ = \frac{\gamma_{\mathrm{L}} - 1}{2}(p^* + p_{\mathrm{L}})\Big(\frac{\rho_{*\mathrm{L}}}{\rho_{\mathrm{L}}} - 1\Big)&\end{aligned} \tag{4.28}$$

We rearrange the equation so that all of the $\rho_{*\mathrm{L}}$ are on one side,

$$\begin{aligned}\rho_{*\mathrm{L}}\Big(b(p_{\mathrm{L}} + \gamma_{\mathrm{L}}p_\infty) - \frac{1}{\rho_{\mathrm{L}}}(p_{\mathrm{L}} + \gamma_{\mathrm{L}}p_\infty) - b(p^* + \gamma_{\mathrm{L}}p_\infty) - \frac{\gamma_{\mathrm{L}} - 1}{2\rho_{\mathrm{L}}}(p^* + p_{\mathrm{L}})\Big)& \\ = -(p^* + \gamma_{\mathrm{L}}p_\infty) - \frac{\gamma_{\mathrm{L}} - 1}{2}(p^* + p_{\mathrm{L}}).&\end{aligned} \tag{4.29}$$

More rewriting yields,

$$\begin{aligned} -\frac{\rho_{*\mathrm{L}} p_{\mathrm{L}}}{2\rho_{\mathrm{L}}}\Big((\gamma_{\mathrm{L}}-1+2b\rho_{\mathrm{L}})\frac{p^*}{p_{\mathrm{L}}}+\frac{2\gamma_{\mathrm{L}}p_\infty}{p_{\mathrm{L}}}+(\gamma_{\mathrm{L}}+1-2b\rho_{\mathrm{L}})\Big) \\ =-\frac{1}{2}(\gamma_{\mathrm{L}}+1)p^*-\gamma_{\mathrm{L}}p_\infty-\frac{1}{2}(\gamma_{\mathrm{L}}-1)p_{\mathrm{L}}. \end{aligned} \tag{4.30}$$

Solving for $\rho_{*\mathrm{L}}$,

$$\rho_{*\mathrm{L}}=\frac{\rho_{\mathrm{L}}((\gamma_{\mathrm{L}}+1)p^*+2\gamma_{\mathrm{L}}p_\infty+(\gamma_{\mathrm{L}}-1)p_{\mathrm{L}})}{p_{\mathrm{L}}\big((\gamma_{\mathrm{L}}-1+2b\rho_{\mathrm{L}})\frac{p^*}{p_{\mathrm{L}}}+\frac{2\gamma_{\mathrm{L}}p_\infty}{p_{\mathrm{L}}}+(\gamma_{\mathrm{L}}+1-2b\rho_{\mathrm{L}})\big)} \tag{4.31}$$

Dividing by ρ_{L} and rewriting, we have,

$$\frac{\rho_{*\mathrm{L}}}{\rho_{\mathrm{L}}}=\left[\frac{\frac{p^*+p_\infty}{p_{\mathrm{L}}+p_\infty}+\frac{\gamma_{\mathrm{L}}-1}{\gamma_{\mathrm{L}}+1}}{\left(\frac{\gamma_{\mathrm{L}}-1+2b\rho_{\mathrm{L}}}{\gamma_{\mathrm{L}}+1}\right)\frac{p^*+p_\infty}{p_{\mathrm{L}}+p_\infty}+\frac{\gamma_{\mathrm{L}}+1-2b\rho_{\mathrm{L}}}{\gamma_{\mathrm{L}}+1}}\right]. \tag{4.32}$$

From (4.24) we have,

$$\rho_{\mathrm{L}}^2\widehat{v}_{\mathrm{L}}^2=-\frac{p^*-p_{\mathrm{L}}}{\frac{1}{\rho_{*\mathrm{L}}}-\frac{1}{\rho_{\mathrm{L}}}}=\rho_{\mathrm{L}}\frac{p^*-p_{\mathrm{L}}}{1-\frac{\rho_{\mathrm{L}}}{\rho_{*\mathrm{L}}}}. \tag{4.33}$$

In addition, using (4.23) and (4.24), we can also derive the identity, $\rho_{\mathrm{L}}\widehat{v}_{\mathrm{L}}=\frac{p^*-p_{\mathrm{L}}}{v_{\mathrm{L}}-v_{*\mathrm{L}}}$ which we can equate with (4.33), to find,

$$v_*=v_{\mathrm{L}}-\sqrt{p^*-p_{\mathrm{L}}}\sqrt{\frac{1}{\rho_{\mathrm{L}}}\Big(1-\frac{\rho_{\mathrm{L}}}{\rho_{*\mathrm{L}}}\Big)}. \tag{4.34}$$

Finally, we substitute (4.32) into (4.34) to find,

$$v^*=v_{\mathrm{L}}-f_{\mathrm{L}}^{\mathrm{shock}}(p^*) \tag{4.35}$$

where $f_{\mathrm{L}}^{\mathrm{shock}}$ is defined in (4.21). Thus the solution across the shock wave is,

$$\boldsymbol{u}_{*\mathrm{L}}:=\left(\frac{\rho_{\mathrm{L}}((\gamma_{\mathrm{L}}+1)p^*+2\gamma_{\mathrm{L}}p_\infty+(\gamma_{\mathrm{L}}-1)p_{\mathrm{L}})}{p_{\mathrm{L}}\big((\gamma_{\mathrm{L}}-1+2b\rho_{\mathrm{L}})\frac{p^*}{p_{\mathrm{L}}}+\frac{2\gamma_{\mathrm{L}}p_\infty}{p_{\mathrm{L}}}+(\gamma_{\mathrm{L}}+1-2b\rho_{\mathrm{L}})\big)},v^*-f_{\mathrm{L}}^{\mathrm{shock}}(p^*),p^*,\gamma_{\mathrm{L}}\right)^{\mathsf{T}}. \tag{4.36}$$

The density across the shock can be written in the following alternative form,

$$\boldsymbol{u}_{*\mathrm{L}} := \left(\frac{\rho_{\mathrm{L}}\left(\frac{p^*}{p_{\mathrm{L}}} + \frac{2\gamma_{\mathrm{L}} p_\infty}{(\gamma_{\mathrm{L}}+1)p_{\mathrm{L}}} + \frac{\gamma_{\mathrm{L}}-1}{\gamma_{\mathrm{L}}+1}\right)}{\frac{\gamma_{\mathrm{L}}-1+2b\rho_{\mathrm{L}}}{\gamma_{\mathrm{L}}+1}\frac{p^*}{p_{\mathrm{L}}} + \frac{2\gamma_{\mathrm{L}} p_\infty}{(\gamma_{\mathrm{L}}+1)p_{\mathrm{L}}} + \frac{\gamma_{\mathrm{L}}+1-2b\rho_{\mathrm{L}}}{\gamma_{\mathrm{L}}+1}}, v^* - f_{\mathrm{L}}^{\mathrm{shock}}(\boldsymbol{p}^*), \boldsymbol{p}^*, \gamma_{\mathrm{L}} \right)^{\mathsf{T}}. \tag{4.37}$$

Next, we derive the shock speed. From (4.33), we can solve for shock speed, $\mathcal{S}_{\mathrm{L}}$,

$$\mathcal{S}_{\mathrm{L}}(\boldsymbol{p}^*) = v_{\mathrm{L}} - \sqrt{\frac{\boldsymbol{p}^* - p_{\mathrm{L}}}{\rho_{\mathrm{L}}(1 - \frac{\rho_{\mathrm{L}}}{\rho_{*\mathrm{L}}})}}. \tag{4.38}$$

Using in (4.32) again and with a lot of simplification, we find,

$$\mathcal{S}_{\mathrm{L}} = v_{\mathrm{L}} - a_{\mathrm{L}}\sqrt{\frac{\gamma_{\mathrm{L}}+1}{2\gamma_{\mathrm{L}}}\left(\frac{\boldsymbol{p}^* - p_{\mathrm{L}}}{p_{\mathrm{L}} + p_\infty}\right) + 1}, \tag{4.39}$$

where $a_{\mathrm{L}} = \sqrt{\frac{\gamma_{\mathrm{L}}(p_{\mathrm{L}}+p_\infty)}{\rho_{\mathrm{L}}(1-b\rho_{\mathrm{L}})}}$ is the sound speed for the NASG EOS. The shock speed is the wave speed for the L-wave. That is, we define $\lambda_1^-(\boldsymbol{p}^*) = \lambda_1^+(\boldsymbol{p}^*) := \mathcal{S}_L$.

The R-wave can be computed similarly. Doing so, we find,

$$v^* = v_{\mathrm{R}} + f_{\mathrm{R}}^{\mathrm{shock}}(\boldsymbol{p}) \tag{4.40}$$

and

$$\mathcal{S}_{\mathrm{R}}(\boldsymbol{p}^*) = v_{\mathrm{R}} + a_{\mathrm{R}}\sqrt{\frac{\gamma_{\mathrm{R}}+1}{2\gamma_{\mathrm{R}}}\left(\frac{\boldsymbol{p}^* - p_{\mathrm{R}}}{p_{\mathrm{R}} + p_\infty}\right) + 1}. \tag{4.41}$$

The wave speed on the R-wave is then $\lambda_{\mathrm{R}}^-(\boldsymbol{p}^*) = \lambda_{\mathrm{R}}^+(\boldsymbol{p}^*) = \mathcal{S}_{\mathrm{R}}$ and the solution across the shock is given by,

$$\boldsymbol{u}_{*\mathrm{R}} := \left(\frac{\rho_{\mathrm{R}}((\gamma_{\mathrm{R}}+1)\boldsymbol{p}^* + 2\gamma_{\mathrm{R}} p_\infty + (\gamma_{\mathrm{R}}-1)p_{\mathrm{R}})}{p_{\mathrm{R}}\left((\gamma_{\mathrm{R}} - 1 + 2b\rho_{\mathrm{R}})\frac{p^*}{p_{\mathrm{R}}} + \frac{2\gamma_{\mathrm{R}} p_\infty}{p_{\mathrm{R}}} + (\gamma_{\mathrm{R}} + 1 - 2b\rho_{\mathrm{R}})\right)}, v^* + f_{\mathrm{R}}^{\mathrm{shock}}(\boldsymbol{p}^*), \boldsymbol{p}^*, \gamma_{\mathrm{R}} \right)^{\mathsf{T}}. \tag{4.42}$$

or in the alternative form,

$$\boldsymbol{u}_{*\mathrm{R}} := \left(\frac{\rho_{\mathrm{R}}\left(\frac{p^*}{p_{\mathrm{R}}} + \frac{2\gamma_{\mathrm{R}} p_\infty}{(\gamma_{\mathrm{R}}+1)p_{\mathrm{R}}} + \frac{\gamma_{\mathrm{R}}-1}{\gamma_{\mathrm{R}}+1}\right)}{\frac{\gamma_{\mathrm{R}}-1+2b\rho_{\mathrm{R}}}{\gamma_{\mathrm{R}}+1}\frac{p^*}{p_{\mathrm{R}}} + \frac{2\gamma_{\mathrm{R}} p_\infty}{(\gamma_{\mathrm{R}}+1)p_{\mathrm{R}}} + \frac{\gamma_{\mathrm{R}}+1-2b\rho_{\mathrm{R}}}{\gamma_{\mathrm{R}}+1}}, v^* + f_{\mathrm{R}}^{\mathrm{shock}}(\boldsymbol{p}^*), \boldsymbol{p}^*, \gamma_{\mathrm{R}} \right)^{\mathsf{T}}. \tag{4.43}$$

4.3.2 Rarefaction Wave

Now assume that the left wave is a rarefaction. Since the left wave uses the NASG EOS for $\gamma = \gamma_L$, we can use the existence of the specific entropy, s, for the NASG EOS. This is due to the NASG EOS being a convex equation of state; see Section 2.2. Since the specific entropy is constant along expansions, we use (2.12) to write the isentropic pressure law for the NASG EOS (with a slight abuse of the notation) as,

$$\boldsymbol{p}(\rho) = C\Big(\frac{\rho}{1-b\rho}\Big)^{\gamma_L} - p_\infty, \tag{4.44}$$

where C is some constant depending on the specific entropy. Thus across the wave, for any state on the expansion connected to the left state, we have the following parametrization of the density,

$$\frac{1}{\rho(\boldsymbol{p})} - b = \Big(\frac{1}{\rho_{\mathrm{L}}} - b\Big)\Big(\frac{p_{\mathrm{L}} + p_\infty}{\boldsymbol{p} + p_\infty}\Big)^{\frac{1}{\gamma_L}} \tag{4.45}$$

Using this parametrization, we can also parametrize the sound speed on the expansion, call it $a_{\mathrm{L}}^{\exp}$. Using (4.45) in the definition of the sound speed, (4.16), we see that,

$$a_{\mathrm{L}}^{\exp}(\boldsymbol{p}) = a_{\mathrm{L}}\frac{\rho_{\mathrm{L}}}{\rho(\boldsymbol{p})}\Big(\frac{\boldsymbol{p}+p_\infty}{p_{\mathrm{L}}+p_\infty}\Big)^{\frac{\gamma_{\mathrm{L}}+1}{2\gamma_{\mathrm{L}}}} = a_{\mathrm{L}}\frac{1-b\rho_{\mathrm{L}}}{1-b\rho(\boldsymbol{p})}\Big(\frac{\boldsymbol{p}+p_\infty}{p_{\mathrm{L}}+p_\infty}\Big)^{\frac{\gamma_{\mathrm{L}}-1}{2\gamma_{\mathrm{L}}}}. \tag{4.46}$$

In order to determine the solution on the expansion we make use of the generalized 1-Riemann invariant on this wave. Recall the definition of the generalized Riemann invariant given in Definition 1.2.6. The corresponding right eigenvector for the 1-characteristic is $\boldsymbol{s}_1 = (\rho, -a_{\mathrm{L}}^{\exp}(\boldsymbol{p}), \mathbf{0}_{d-1}, 0, 0)^{\mathsf{T}}$. The generalized 1-Riemann invariant, W_1, can be found by solving, $D\mathsf{W}_1(\boldsymbol{z}) \cdot \boldsymbol{s}_1 = 0$ where $\boldsymbol{z} = (\rho, v, \boldsymbol{p})$ is the vector of primitive variables. This gives

us the differential equation,

$$\frac{\partial \mathsf{W}_1}{\partial \rho}(\boldsymbol{z}) + \frac{a_{\mathrm{L}}^{\exp}(\boldsymbol{p})}{\rho}\frac{\partial \mathsf{W}_1}{\partial v}(\boldsymbol{z}) = 0. \tag{4.47}$$

A solution to this differential equation is,

$$\mathsf{W}_1(\boldsymbol{z}) = v + \int_{\rho_0}^{\rho} \frac{a_{\mathrm{L}}^{\exp}(\boldsymbol{z})}{\varrho}\,\mathrm{d}\varrho \tag{4.48}$$

Using the sound speed definition, (4.16), and the isentropic pressure law, (4.44), the Riemann invariant becomes,

$$\mathsf{W}_1(\boldsymbol{z}) = v - \sqrt{C\gamma_{\mathrm{L}}}\int_{\rho_0}^{\rho} \frac{\varrho^{\frac{1}{2}(\gamma_{\mathrm{L}}-3)}}{(1-b\varrho)^{\frac{1}{2}(\gamma_{\mathrm{L}}+1)}}\,\mathrm{d}\rho = v - \frac{2\sqrt{C\gamma_{\mathrm{L}}}}{\gamma_{\mathrm{L}}-1}\Big(\frac{\varrho}{1-b\varrho}\Big)^{\frac{1}{2}(\gamma_{\mathrm{L}}-1)}\Big|_{\rho_0}^{\rho} \tag{4.49}$$

Furthermore, the sound speed on the expansion can be written as, $a_{\mathrm{L}}^{\exp}(\boldsymbol{p}(\rho)) = \sqrt{\frac{C\gamma_L\rho^{\gamma_L-1}}{(1-b\rho)^{\gamma_L+1}}}$. Hence the Riemann invariant takes the form,

$$\mathsf{W}_1(\boldsymbol{z}) = v + \frac{2a_{\mathrm{L}}^{\exp}(\boldsymbol{z})}{\gamma_L - 1}(1-b\rho) + \text{const.} \tag{4.50}$$

Note that the 1-Riemann invariant is constant along the 1-wave if the 1-wave is an expansion. Hence, $\mathsf{W}_1(\boldsymbol{z}_L) = \mathsf{W}_1(\boldsymbol{z}_{*L})$. Therefore, we have the relationship for any state, $\boldsymbol{z}$ on the expansion,

$$v + \frac{2a_{\mathrm{L}}^{\exp}(\boldsymbol{z})}{\gamma_{\mathrm{L}}-1}(1-b\rho) = v_{\mathrm{L}} + \frac{2a_{\mathrm{L}}}{\gamma_{\mathrm{L}}-1}(1-b\rho_{\mathrm{L}}) \tag{4.51}$$

We now introduce the self similarity parameter, $\xi := x/t$. Hence, along the rarefaction, we have $\xi = v - a = v - a_{\mathrm{L}}^{\exp}(\boldsymbol{p})$. Using (4.51) we have the equation,

$$\xi(\boldsymbol{p}) = v_{\mathrm{L}} - \frac{2a_{\mathrm{L}}}{\gamma_{\mathrm{L}}-1}(1-b\rho_{\mathrm{L}}) - \frac{2a_{\mathrm{L}}^{\exp}(\boldsymbol{p})}{\gamma_{\mathrm{L}}-1}(1-b\rho(\boldsymbol{p})) - a_{\mathrm{L}}^{\exp}(\boldsymbol{p}). \tag{4.52}$$

Using the definition of $a_{\mathrm{L}}^{\exp}$, (4.46), in (4.52) and with a lot of algebra, we arrive at the

following,

$$\xi_{\mathrm{L}}(p) = v_{\mathrm{L}} - \frac{2a_{\mathrm{L}}(1-b\rho_{\mathrm{L}})}{\gamma_{\mathrm{L}}-1}\Big(\Big(\frac{p+p_\infty}{p_{\mathrm{L}}+p_\infty}\Big)^{\frac{\gamma_{\mathrm{L}}-1}{2\gamma_{\mathrm{L}}}} - 1\Big) - \frac{a_{\mathrm{L}}(1-b\rho_{\mathrm{L}})}{1-b\rho(p)}\Big(\frac{p+p_\infty}{p_{\mathrm{L}}+p_\infty}\Big)^{\frac{\gamma_{\mathrm{L}}-1}{2\gamma_{\mathrm{L}}}}. \tag{4.53}$$

Using the notation from (4.21), we write,

$$\xi_{\mathrm{L}}(p) = v_{\mathrm{L}} - f_{\mathrm{L}}^{\exp}(p) - \frac{a_{\mathrm{L}}(1-b\rho_{\mathrm{L}})}{1-b\rho(p)}\Big(\frac{p+p_\infty}{p_{\mathrm{L}}+p_\infty}\Big)^{\frac{\gamma_{\mathrm{L}}-1}{2\gamma_{\mathrm{L}}}}. \tag{4.54}$$

Notice that $\xi_{\mathrm{L}}(p)$ is a strictly decreasing function for $p \in [p^*, p_{\mathrm{L}}]$. Hence by the inverse function theorem, $\xi(p)$ is invertible, and thus p can be expressed as a decreasing function of ξ for $\xi \in [\xi_{\mathrm{L}}(p_{\mathrm{L}}), \xi_{\mathrm{L}}(p^*)] = [\lambda_{\mathrm{L}}^-(p_{\mathrm{L}}), \lambda_{\mathrm{L}}^+(p^*)]$ where λ_{L}^- and λ_{L}^+ are defined in (4.22). Similarly, if an expansion on the R-wave occurs, then the parametrization of ξ is given by,

$$\xi_{\mathrm{R}}(p) = v_{\mathrm{R}} + f_{\mathrm{R}}^{\exp}(p) + \frac{a_{\mathrm{R}}(1-b\rho_{\mathrm{R}})}{1-b\rho(p)}\Big(\frac{p+p_\infty}{p_{\mathrm{R}}+p_\infty}\Big)^{\frac{\gamma_{\mathrm{R}}-1}{2\gamma_{\mathrm{R}}}}. \tag{4.55}$$

Then it can be seen that $\xi_{\mathrm{R}}(p)$ is a strictly increasing function for $p \in [p^*, p_{\mathrm{R}}]$. And again, $\xi_{\mathrm{R}}(p)$ is invertible, and so we can conclude that $p = p(\xi)$ is a strictly increasing function for $\xi \in [\xi_{\mathrm{R}}(p^*), \xi_{\mathrm{R}}(p_{\mathrm{R}})] = [\lambda_{\mathrm{R}}^-(p^*), \lambda_{\mathrm{R}}^+(p_{\mathrm{R}})]$. Note however, we cannot find an explicit equation for $p(\xi)$.

In short, we can say that the pressure, p, decreases across expansions. Thus the solution for an expansion on the L-wave is,

$$\boldsymbol{u}_{\mathrm{LL}}(\xi) := \begin{pmatrix} \Big(b + \big(\frac{1}{\rho_{\mathrm{L}}} - b\big)\big(\frac{p_{\mathrm{L}}+p_\infty}{p(\xi)+p_\infty}\big)^{\frac{1}{\gamma_{\mathrm{L}}}}\Big)^{-1} \\ v_{\mathrm{L}} - f_{\mathrm{L}}^{\exp}(p(\xi)) \\ p_{\mathsf{nasg}}(\xi) \\ \gamma_{\mathrm{L}} \end{pmatrix}, \tag{4.56}$$

and similarly, the solution on the R-wave is,

$$\boldsymbol{u}_{\mathrm{RR}}(\xi) := \begin{pmatrix} \left(b + \left(\frac{1}{\rho_{\mathrm{R}}} - b\right)\left(\frac{p_{\mathrm{R}}+p_\infty}{p_{\mathsf{nasg}}(\xi)+p_\infty}\right)^{\frac{1}{\gamma_{\mathrm{R}}}}\right)^{-1} \\ v_{\mathrm{R}} + f_{\mathrm{R}}^{\exp}(p(\xi)) \\ p_{\mathsf{nasg}}(\xi) \\ \gamma_{\mathrm{R}} \end{pmatrix}, \tag{4.57}$$

4.4 Connecting the L- and R-waves

Using the fact that the velocity and pressure are constant across the contact, we have that,

$$v^* = v_{\mathrm{L}} - f_{\mathrm{L}}(p^*) = v_{\mathrm{R}} + f_{\mathrm{R}}(p^*), \tag{4.58}$$

which defines the following nonlinear equation to solve for p^*,

$$\varphi(p) := f_{\mathrm{R}}(p) + f_{\mathrm{L}}(p) + v_R - v_L = 0, \qquad \text{for } p \in (-p_\infty, \infty), \tag{4.59}$$

where f_{Z} is defined in (4.21) for $\mathrm{Z} \in \{\mathrm{L}, \mathrm{R}\}$.

Lemma 4.4.1 (Properties of φ)**.** *The function φ defined in* (4.59) *is $C^2((-p_\infty, \infty))$, strictly increasing, concave, and $\varphi'''(p) > 0$ a.e.*

Proof. To see the proof, we must compute the derivative of the functions, $f_{\mathrm{Z}}^{\exp}$ and $f_{\mathrm{Z}}^{\mathrm{shock}}$. Recall, $f_{\mathrm{Z}}^{\exp}(p) = \frac{2a_{\mathrm{Z}}(1-b\rho_{\mathrm{Z}})}{\gamma_{\mathrm{Z}}-1}\left(\left(\frac{p+p_\infty}{p_{\mathrm{Z}}+p_\infty}\right)^{\frac{\gamma_{\mathrm{Z}}-1}{2\gamma_{\mathrm{Z}}}} - 1\right)$ and $f_{\mathrm{Z}}^{\mathrm{shock}} = (p - p_{\mathrm{Z}})\sqrt{\frac{A_{\mathrm{Z}}}{p+p_\infty+B_{\mathrm{Z}}}}$. Then we have,

$$\frac{\mathrm{d}f_{\mathrm{Z}}^{\exp}}{\mathrm{d}p}(p) = \frac{a_{\mathrm{Z}}(1-b\rho_{\mathrm{Z}})}{\gamma_{\mathrm{Z}}(p_{\mathrm{Z}}+p_\infty)}\left(\frac{p+p_\infty}{p_{\mathrm{Z}}+p_\infty}\right)^{-\frac{\gamma_{\mathrm{Z}}+1}{2\gamma_{\mathrm{Z}}}} > 0, \tag{4.60a}$$

$$\frac{\mathrm{d}^2 f_{\mathrm{Z}}^{\exp}}{\mathrm{d}p^2}(p) = -\frac{a_{\mathrm{Z}}(\gamma_{\mathrm{Z}}+1)(1-b\rho_{\mathrm{Z}})}{2\gamma_{\mathrm{Z}}^2(p_{\mathrm{Z}}+p_\infty)^2}\left(\frac{p+p_\infty}{p_{\mathrm{Z}}+p_\infty}\right)^{-\frac{3\gamma_{\mathrm{Z}}+1}{2\gamma_{\mathrm{Z}}}} < 0, \tag{4.60b}$$

$$\frac{\mathrm{d}^3 f_{\mathrm{Z}}^{\exp}}{\mathrm{d}p^3}(p) = \frac{a_{\mathrm{Z}}(3\gamma_{\mathrm{Z}}+1)(\gamma_{\mathrm{Z}}+1)(1-b\rho_{\mathrm{Z}})}{4\gamma_{\mathrm{Z}}^3(p_{\mathrm{Z}}+p_\infty)^3}\left(\frac{p+p_\infty}{p_{\mathrm{Z}}+p_\infty}\right)^{-\frac{5\gamma_{\mathrm{Z}}+1}{2\gamma_{\mathrm{Z}}}} > 0, \tag{4.60c}$$

and

$$\frac{\mathrm{d}f_{\mathrm{Z}}^{\mathrm{shock}}}{\mathrm{d}\rho}(\rho) = \sqrt{\frac{A_{\mathrm{Z}}}{\rho + p_\infty + B_{\mathrm{Z}}}}\Big(1 - \frac{\rho - p_{\mathrm{Z}}}{2(\rho + p_\infty + B_{\mathrm{Z}})}\Big) > 0, \tag{4.61a}$$

$$\frac{\mathrm{d}^2 f_{\mathrm{Z}}^{\mathrm{shock}}}{\mathrm{d}\rho^2}(\rho) = -\frac{\sqrt{A_{\mathrm{Z}}}(\rho + p_\infty + 3(p_{\mathrm{Z}} + p_\infty) + 4B_{\mathrm{Z}})}{4(\rho + p_\infty + B_{\mathrm{Z}})^{5/2}} < 0, \tag{4.61b}$$

$$\frac{\mathrm{d}^3 f_{\mathrm{Z}}^{\mathrm{shock}}}{\mathrm{d}\rho^3}(\rho) = \frac{3\sqrt{A_{\mathrm{Z}}}(\rho + p_\infty + 5(p_{\mathrm{Z}} + p_\infty) + 6B_{\mathrm{Z}})}{8(\rho + p_\infty + B_{\mathrm{Z}})^{7/2}} > 0. \tag{4.61c}$$

Based on the definition of $\varphi(\rho)$, (4.59), we see that φ is monotonically increasing, concave down and $\varphi'''(\rho) > 0$. Lastly, we need to show that $\varphi \in C^2((-p_\infty, \infty))$. As $f_{\mathrm{Z}}^{\mathrm{shock}}$ and $f_{\mathrm{Z}}^{\mathrm{exp}}$ are both C^3 functions, we just need to show that continuity of the derivatives holds at p_{Z}. Using the fact that $\frac{a_{\mathrm{Z}}(1-b\rho_{\mathrm{Z}})}{\gamma_{\mathrm{Z}}(p_{\mathrm{Z}}+p_\infty)} = \sqrt{\frac{A_{\mathrm{Z}}}{p_{\mathrm{Z}}+p_\infty+B_{\mathrm{Z}}}}$, we see that $\frac{\mathrm{d}f_{\mathrm{Z}}^{\mathrm{exp}}}{\mathrm{d}\rho}(p_{\mathrm{Z}}) = \frac{\mathrm{d}f_{\mathrm{Z}}^{\mathrm{shock}}}{\mathrm{d}\rho}(p_{\mathrm{Z}})$, $\frac{\mathrm{d}^2 f_{\mathrm{Z}}^{\mathrm{exp}}}{\mathrm{d}\rho^2}(p_{\mathrm{Z}}) = \frac{\mathrm{d}^2 f_{\mathrm{Z}}^{\mathrm{shock}}}{\mathrm{d}\rho^2}(p_{\mathrm{Z}})$, but

$$\frac{\mathrm{d}^3 f_{\mathrm{Z}}^{\mathrm{exp}}}{\mathrm{d}\rho^3}(p_{\mathrm{Z}}) = \frac{a_{\mathrm{Z}}(3\gamma_{\mathrm{Z}}+1)(\gamma_{\mathrm{Z}}+1)(1-b\rho_{\mathrm{Z}})}{4\gamma_{\mathrm{Z}}^3(p_{\mathrm{Z}}+p_\infty)^3} \neq \frac{9\sqrt{A_{\mathrm{Z}}}}{4(p_{\mathrm{Z}}+p_\infty+B_{\mathrm{Z}})^{5/2}} = \frac{\mathrm{d}^3 f_{\mathrm{Z}}^{\mathrm{shock}}}{\mathrm{d}\rho^3}(p_{\mathrm{Z}}). \tag{4.62}$$

This completes the proof. □

Remark 4.4.1 (The Nonvacuum Condition)**.** Since the function $\varphi(\rho)$ is strictly increasing, a root to the equation (4.59) only exists if $\varphi(-p_\infty) < 0$. This gives us the so-called *nonvacuum condition*

$$v_{\mathrm{R}} - v_{\mathrm{L}} < \frac{2a_{\mathrm{L}}(1-b\rho_{\mathrm{L}})}{\gamma_{\mathrm{L}}-1} + \frac{2a_{\mathrm{R}}(1-b\rho_{\mathrm{R}})}{\gamma_{\mathrm{R}}-1}. \tag{4.63}$$

If $\varphi(-p_\infty) \geq 0$, then we set $\rho^* := -p_\infty$ and the solution of the Riemann problem consists of two expansions into vacuum. □

Lemma 4.4.2 (Unique Root for $\varphi(\rho) = 0$)**.** *If* (4.63) *holds, then* φ *has a unique root,* $\rho^* \in (-p_\infty, \infty)$.

Proof. Note that condition if (4.63) holds, then that implies $\varphi(-p_\infty) < 0$ and since φ is $C^3((-p_\infty, \infty))$ and strictly increasing, this implies the existence of a unique root. □

Remark 4.4.2 (Fast Root Finding Method)**.** Since φ has nice properties from Lemma 4.4.1, we can use the quadratic Newton method to quickly find the root of the equation. This fast estimation method was first introduced by Guermond & Popov in [59]. To avoid the nonlinear solver, we also propose an upper bound on the root in Section 4.6. □

4.5 A Weak Solution to the Extended Riemann Problem

For $\mathrm{Z} \in \{\mathrm{L}, \mathrm{R}\}$, define,

$$\boldsymbol{u}_{\mathrm{Z}}^{*} := \begin{cases} \boldsymbol{u}_{\mathrm{ZZ}}(\xi(p^{*})), & \text{if } p^{*} < p_{\mathrm{Z}}, \\ \boldsymbol{u}_{*\mathrm{Z}}, & \text{if } p_{\mathrm{Z}} \leq p^{*}. \end{cases} \tag{4.64}$$

For reference, $\boldsymbol{u}_{\mathrm{ZZ}}$ is defined in (4.56) and (4.57) and $\boldsymbol{u}_{*\mathrm{Z}}$ is defined in (4.37) and (4.43). We claim that,

$$\widetilde{\boldsymbol{u}}(x,t) := \begin{cases} \boldsymbol{u}_{\mathrm{L}} & \text{if } \frac{x}{t} < \lambda_{\mathrm{L}}^{-}, \\ \boldsymbol{u}_{\mathrm{LL}}(\frac{x}{t}) & \text{if } \lambda_{\mathrm{L}}^{-} \leq \frac{x}{t} < \lambda_{\mathrm{L}}^{+}, \\ \boldsymbol{u}_{\mathrm{L}}^{*} & \text{if } \lambda_{\mathrm{L}}^{+} \leq \frac{x}{t} < v^{*} \\ \boldsymbol{u}_{\mathrm{R}}^{*} & \text{if } v^{*} < \frac{x}{t} < \lambda_{\mathrm{R}}^{-}, \\ \boldsymbol{u}_{\mathrm{RR}}(\frac{x}{t}) & \text{if } \lambda_{\mathrm{R}}^{-} \leq \frac{x}{t} < \lambda_{\mathrm{R}}^{+}, \\ \boldsymbol{u}_{\mathrm{R}} & \text{if } \lambda_{\mathrm{R}}^{+} \leq \frac{x}{t} \end{cases} \tag{4.65}$$

is a weak solution to the Riemann problem.

Lemma 4.5.1 (Upper Bound on the Density)**.** *For $b > 0$, if $\widetilde{\boldsymbol{u}}_L, \widetilde{\boldsymbol{u}}_R \in \widetilde{\mathcal{B}}(b, q, p_\infty)$, then the solution to the extended Riemann problem, $\widetilde{\boldsymbol{u}}(x,t)$ given in* (4.65) *satisfies, $\rho < b^{-1}$.*

Proof. If the Z-wave is an expansion, then the parametrization of the density as a function of $\boldsymbol{\rho}$, seen in (4.56) or (4.57) is an increasing function of $\boldsymbol{\rho}$. As seen in Section 4.3.2, we know that the pressure decreases across an expansion hence, the density decreases across an expansion. Thus, $\rho(x,t) \leq \rho_{\mathrm{Z}} < b^{-1}$ on an expansion wave.

Now assume that the Z-wave is a shock. From (4.37) or (4.43) we see that

$$\rho_{*\mathrm{Z}} = \frac{\rho_{\mathrm{Z}}\left(\frac{p^*}{p_{\mathrm{Z}}} + \frac{2\gamma_{\mathrm{Z}} p_\infty}{(\gamma_{\mathrm{Z}}+1)p_{\mathrm{Z}}} + \frac{\gamma_{\mathrm{Z}}-1}{\gamma_{\mathrm{Z}}+1}\right)}{\frac{\gamma_{\mathrm{Z}}-1+2b\rho_{\mathrm{Z}}}{\gamma_{\mathrm{Z}}+1}\frac{p^*}{p_{\mathrm{Z}}} + \frac{2\gamma_{\mathrm{Z}} p_\infty}{(\gamma_{\mathrm{Z}}+1)p_{\mathrm{Z}}} + \frac{\gamma_{\mathrm{Z}}+1-2b\rho_{\mathrm{Z}}}{\gamma_{\mathrm{Z}}+1}}. \tag{4.66}$$

Using that $\gamma_{\mathrm{Z}} > 1$ and $\rho_{\mathrm{Z}} < b^{-1}$, we can see through elementary calculus that $\rho_{*\mathrm{Z}}$ is an increasing function of p^*. Taking $p^* \to \infty$, we see that

$$\lim_{p^*\to\infty} \rho_{*\mathrm{Z}}(p^*) = \frac{\gamma_{\mathrm{Z}}+1}{\gamma_{\mathrm{Z}}-1+2b\rho_{\mathrm{Z}}}\rho_{\mathrm{Z}}. \tag{4.67}$$

Thus, the upper bound is now seen as a decreasing function of γ_{Z}. Taking $\gamma_{\mathrm{Z}} \to 1^+$, we find that,

$$\lim_{\gamma_{\mathrm{Z}}\to 1^+} \frac{\gamma_{\mathrm{Z}}+1}{\gamma_{\mathrm{Z}}-1+2b\rho_{\mathrm{Z}}}\rho_{\mathrm{Z}} = b^{-1}. \tag{4.68}$$

Since $\gamma_{\mathrm{Z}} > 1$, we have that $\rho_{*\mathrm{Z}} < b^{-1}$ and therefore, $\rho < b^{-1}$ in the solution to the extended Riemann problem for a.e. $x \in \mathbb{R}$ and $t > 0$. □

Lemma 4.5.2 (Weak Solution to the Extended Riemann Problem)**.** *Assume the nonvacuum condition holds,* (4.63)*. Then* $\widetilde{\boldsymbol{u}}(x,t)$ *defined by* (4.65)*, is a weak solution to the extended Riemann problem,* (4.20)*. Moreover,* $\widetilde{\boldsymbol{u}}(x,t) \in \widetilde{\mathcal{B}}(b,q,p_\infty)$.

Proof. In the domain $\{x < v^*t\}$, we have that $\gamma = \gamma_{\mathrm{L}}$ and hence $\Gamma = \gamma_{\mathrm{L}}\rho$. Therefore, the last equation in (4.20), $\partial_t\Gamma + \partial_x(v\Gamma)$ is equivalent to the conservation of mass. Then by construction, the first three equations in (4.20) hold are satisfied in the weak sense since with the EOS defined by $p = (\gamma_{\mathrm{L}}-1)\frac{\rho(e-q)}{1-b\rho} - \gamma_{\mathrm{L}}p_\infty$.

The result is the same in the domain $\{x > v^*t\}$ but with $\gamma = \gamma_{\mathrm{R}}$. In order to complete the proof, we need the two states, $\widetilde{\boldsymbol{u}}^*_{\mathrm{L}} = (\rho_{*\mathrm{L}}, v^*_{\mathrm{L}}, p^*_{\mathrm{L}}, \gamma^*_{\mathrm{R}})$ and $\widetilde{\boldsymbol{u}}^*_{\mathrm{R}} = (\rho_{*\mathrm{R}}, v^*_{\mathrm{R}}, p^*_{\mathrm{R}}, \gamma^*_{\mathrm{R}})$ on the left and right of the contact, $\{x = v^*t\}$ to satisfy the Rankine-Hugoniot condition. Since the nonvacuum condition holds, we have that $v^*_{\mathrm{L}} = v^*_{\mathrm{R}} = v^*$ and $p^*_{\mathrm{L}} = p^*_{\mathrm{R}} = p^*$. The Rankine-Hugoniot conditions are immediately satisfied.

Since there is no vacuum state, the density remains strictly positive and by Lemma 4.5.1.

By the definition of p^* as the root of (4.59), we have that $p^* > -p_\infty$ in the nonvacuum case. In particular, in the case of a double expansion, we always have that $p > -p_\infty$. From the NASG EOS, we have that $(\gamma_Z - 1)\left(\frac{\rho(e-q)}{1-b\rho} - p_\infty\right) = p + p_\infty > 0$ on either the left or right of the contact. This implies the invariant-domain constraint that $e - q > p_\infty(\tau - b)$, since $\gamma_Z - 1 > 0$. Lastly, $\Gamma > \rho$ is equivalent to $\gamma_L > 1$ left of the contact and $\gamma_R > 1$ right of the contact. These again both hold since the Riemann data satisfy $\boldsymbol{u}_Z \in \mathcal{B}(b, q, p_\infty)$ for $Z \in \{L, R\}$. Therefore, we conclude that $\widetilde{\boldsymbol{u}}(\boldsymbol{x}, t) \in \widetilde{\mathcal{B}}(b, q, p_\infty)$. □

This brings us to the main result from Clayton et. al. [2], which has been modified to hold for the NASG interpolatory EOS.

Theorem 4.5.1. *(i) Let $\mathbf{U}_i^n, \mathbf{U}_j^n \in \mathcal{B}(b, q, p_\infty)$ where $\mathcal{B}(b, q, p_\infty)$ is defined in (1.23). Let p^* be the root of the equation $\varphi(p) = 0$ defined in (4.59) and let $\widehat{p}^*$ be any upper bound on p^*; i.e., $\widehat{p}^* \geq p^*$. Let,*

$$\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) := \max(-\lambda_L^-(\widehat{p}^*), \lambda_R^+(\widehat{p}^*)) \tag{4.69a}$$

$$d_{ij}^{L,n} := \max(\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \widehat{\lambda}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}). \tag{4.69b}$$

Let $\overline{\mathbf{U}}_{ij}^n(t_0)$ be defined in (4.3) with $t_0 = \frac{\|\boldsymbol{c}_{ij}\|}{2d_{ij}^{L,n}}$. Then $\overline{\mathbf{U}}_{ij}^n(t_0) \in \mathcal{B}(b, q, p_\infty)$.

(ii) Let $i \in \mathcal{V}$ and $\mathbf{U}_j^n \in \mathcal{B}(b, q, p_\infty)$ for all $j \in \mathcal{I}(i)$. Let $d_{ij}^{L,n}$ be defined by (4.69b) and assume Δt is small enough so that $\Delta t \sum_{j \in \mathcal{I}(i)\setminus\{i\}} \frac{2d_{ij}^{L,n}}{m_i} \leq 1$. If $\mathbf{U}_i^{L,n+1}$ is defined by (4.1), then $\mathbf{U}_i^{L,n+1} \in \text{conv}\{\overline{\mathbf{U}}_{ij}^n : j \in \mathcal{I}(i)\} \subset \mathcal{B}(b, q, p_\infty)$.

Proof. First notice that $-\lambda_L^-(p)$ and $\lambda_R^+(p)$ are both increasing functions of p. Therefore, for $\widehat{p}^* \geq p^*$, we have that $\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n) \geq \max(-\lambda_L^-(p^*), \lambda_R^+(p^*))$. This implies that,

$$\left(0, \frac{1}{2\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)}\right] \subset \left(0, \frac{1}{2\max(-\lambda_L^-(p^*), \lambda_R^+(p^*))}\right]. \tag{4.70}$$

Thus by definition of $d_{ij}^{L,n}$ and t_0, we have that $t_0 \in (0, 1/(2\widehat{\lambda}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n))]$. In order to apply Lemma 1.3.1, we use $\boldsymbol{g}(\boldsymbol{u}) := \widetilde{\mathbb{f}}(\widetilde{\boldsymbol{u}})\boldsymbol{n}_{ij}$ with the left and right Riemann data $\widetilde{\mathbf{U}}_i^n$ and

$\widetilde{\mathbf{U}}_j^n$, respectively. Then we must show that the solution $\widetilde{\boldsymbol{u}}(x,t) \in \widetilde{\mathcal{B}}(b,q,p_\infty)$ for a.e. $x \in \mathbb{R}$ and all $t > 0$.

Note, the solution given in (4.65) satisfies $\rho > 0$ in every case and by Lemma 4.5.1, $\rho < b^{-1}$ (if $b > 0$). For the constraint, $e - q > p_\infty(\tau - b)$ recall that $e - q = \frac{p+\gamma p_\infty}{\gamma-1}(\tau - b)$. By construction of the solution, $\boldsymbol{p}^* \in (-p_\infty, \infty)$ and hence $e - q > p_\infty(\tau - b)$. Lastly, since $\gamma(x,t) = \gamma_{\mathrm{L}}$ if $x < v^*t$, $\gamma(x,t) = \gamma_{\mathrm{R}}$ if $x > v^*t$ and $\gamma_{\mathrm{L}}, \gamma_{\mathrm{R}} > 1$, this implies that $\Gamma(x,t) > \rho(x,t)$ for a.e. $x \in \mathbb{R}$ and all $t > 0$. Therefore, $\widetilde{\boldsymbol{u}} \in \widetilde{\mathcal{B}}(b,q,p_\infty)$.

From Lemma 4.5.1, 2., we conclude that $\overline{\widetilde{\mathbf{U}}}_{ij}^n(t_0) \in \widetilde{\mathcal{B}}(b,q,p_\infty)$. But from (4.12), the density, momentum and total energy are the same for $\overline{\mathbf{U}}_{ij}^n(t_0)$ and $\overline{\widetilde{\mathbf{U}}}_{ij}^n(t_0)$. So, defining, $\widetilde{\Psi}_1(\widetilde{\boldsymbol{u}}) := \rho$, $\widetilde{\Psi}_2(\widetilde{\boldsymbol{u}}) := 1 - b\rho$, and $\widetilde{\Psi}_3(\widetilde{\boldsymbol{u}}) := \mathsf{e}(\boldsymbol{u}) - q - p_\infty(\rho^{-1} - b)$. We have that $\Psi_l(\overline{\mathbf{U}}_{ij}^n(t_0)) = \widetilde{\Psi}_l\big(\overline{\widetilde{\mathbf{U}}}_{ij}^n(t_0)\big) > 0$ for all $l = 1,2,3$. Therefore, we conclude that $\overline{\mathbf{U}}_{ij}^n(t_0) \in \mathcal{B}(b,q,p_\infty)$. □

4.5.1 Weak solution with Vacuum state

In the case that the nonvacuum condition, (4.63), fails, then $\boldsymbol{p}^* := -p_\infty$ and the velocity in Riemann fan is no longer continuous. That is,

$$v_{\mathrm{L}}^* := v_{\mathrm{L}} - f_{\mathrm{L}}^{\exp}(-p_\infty) = v_{\mathrm{L}} + \frac{2a_{\mathrm{L}}(1 - b\rho_{\mathrm{L}})}{\gamma_{\mathrm{L}} - 1}, \tag{4.71a}$$

$$v_{\mathrm{R}}^* := v_{\mathrm{R}} + f_{\mathrm{R}}^{\exp}(-p_\infty) = v_{\mathrm{R}} - \frac{2a_{\mathrm{R}}(1 - b\rho_{\mathrm{R}})}{\gamma_{\mathrm{R}} - 1}. \tag{4.71b}$$

Define $\widetilde{\boldsymbol{u}}_{\mathrm{Z}}^* := \widetilde{\boldsymbol{u}}_{\mathrm{ZZ}}(v_{\mathrm{Z}}^*) = (0, v_{\mathrm{Z}}^*, -p_\infty, \gamma_{\mathrm{Z}})^{\mathsf{T}}$ for $\mathrm{Z} \in \{\mathrm{L}, \mathrm{R}\}$. Then we claim that the solution to the extended Riemann problem in the presence of vacuum is,

$$\widetilde{\boldsymbol{u}}(x,t) := \begin{cases} \widetilde{\boldsymbol{u}}_{\mathrm{L}} & \text{if } \frac{x}{t} < v_{\mathrm{L}} - a_{\mathrm{L}}, \\ \widetilde{\boldsymbol{u}}_{\mathrm{LL}}(\frac{x}{t}) & \text{if } \lambda_{\mathrm{L}}^- \le \frac{x}{t} < v_{\mathrm{L}}^*, \\ \frac{v_{\mathrm{R}}^* - \frac{x}{t}}{v_{\mathrm{R}}^* - v_{\mathrm{L}}^*}\widetilde{\boldsymbol{u}}_{\mathrm{L}}^* + \frac{\frac{x}{t} - v_{\mathrm{L}}^*}{v_{\mathrm{R}}^* - v_{\mathrm{L}}^*}\widetilde{\boldsymbol{u}}_{\mathrm{R}}^* & \text{if } v_{\mathrm{L}}^* \le \frac{x}{t} < v_{\mathrm{R}}^* \\ \widetilde{\boldsymbol{u}}_{\mathrm{RR}}(\frac{x}{t}) & \text{if } v_{\mathrm{R}}^* \le \frac{x}{t} < v_{\mathrm{R}} + a_{\mathrm{R}}, \\ \widetilde{\boldsymbol{u}}_{\mathrm{R}} & \text{if } v_{\mathrm{R}} + a_{\mathrm{R}} \le \frac{x}{t} \end{cases} \tag{4.72}$$

Lemma 4.5.3 (Weak Solution with Vacuum)**.** *Assume that the nonvacuum condition* (4.63) *fails, i.e.,* $\boldsymbol{p}^* = 0$*. Then the following is true*

1. *The solution* $\widetilde{\boldsymbol{u}}$ *given by* (4.72) *is a weak solution to* (4.20)*.*

2. *If* $\frac{x}{t} \in (v_L - a_L, v_L^*) \cup (v_R^*, v_R + a_R)$*, then* $\widetilde{\boldsymbol{u}} \in \widetilde{\mathcal{B}}(b, q, p_\infty)$*.*

3. $\widetilde{\boldsymbol{u}} \in \overline{\widetilde{\mathcal{B}}}(b, q, p_\infty)$*; that is,* $\widetilde{\boldsymbol{u}}$ *is in the closure of* $\widetilde{\mathcal{B}}(b, q, p_\infty)$*.*

Proof. 1. We have already established that $\widetilde{\boldsymbol{u}}$ is a weak solution to (4.20) from Lemma 4.5.2 in the region $\{x < v_{\mathrm{L}}^* t\} \cup \{v_{\mathrm{R}}^* t < x\}$. In the region $\{v_{\mathrm{L}}^* t < x < v_{\mathrm{R}}^* t\}$, the solution defined by (4.72) is a weak solution since all of the conservative variables are all zero. We only need to show to show that $\widetilde{\boldsymbol{u}}$ is continuous on the line $\{x = v_{\mathrm{L}}^* t\}$ and $\{x = v_{\mathrm{R}}^* t\}$. From the definition of $\xi_{\mathrm{L}}(\boldsymbol{p})$ given in (4.54), we see that $\xi_{\mathrm{L}}(-p_\infty) = v_{\mathrm{L}} - f_{\mathrm{L}}^{\exp}(-p_\infty) = v_{\mathrm{L}}^*$ therefore, $p(v_{\mathrm{L}}^*) = -p_\infty$. Hence $\lim_{\xi \uparrow v_{\mathrm{L}}^*} \widetilde{\boldsymbol{u}}_{\mathrm{LL}}(\xi) = (0, v_{\mathrm{L}}^*, -p_\infty, \gamma_{\mathrm{L}})^{\mathsf{T}}$. Similarly, taking $\lim_{\xi \downarrow v_{\mathrm{L}}^*} \frac{v_{\mathrm{R}}^* - \frac{x}{t}}{v_{\mathrm{R}}^* - v_{\mathrm{L}}^*} \widetilde{\boldsymbol{u}}_{\mathrm{L}}^* + \frac{\frac{x}{t} - v_{\mathrm{L}}^*}{v_{\mathrm{R}}^* - v_{\mathrm{L}}^*} \widetilde{\boldsymbol{u}}_{\mathrm{R}}^* = (0, v_{\mathrm{L}}^*, -p_\infty, \gamma_{\mathrm{L}})^{\mathsf{T}}$. Hence the solution is continuous on the line $\{x = v_{\mathrm{L}}^* t\}$. A similar proof holds for the line $\{x = v_{\mathrm{R}}^* t\}$.

2. This proof follows identically to what is given in the proof of Lemma 4.5.2.

3. We only need to show that the state in the region $\{v_{\mathrm{L}}^* t < x < v_{\mathrm{R}}^* t\}$ lies in $\overline{\widetilde{\mathcal{B}}}(b, q, p_\infty)$. First note that, $\frac{v_{\mathrm{R}}^* - \frac{x}{t}}{v_{\mathrm{R}}^* - v_{\mathrm{L}}^*} \Gamma_{\mathrm{L}}^* + \frac{\frac{x}{t} - v_{\mathrm{L}}^*}{v_{\mathrm{R}}^* - v_{\mathrm{L}}^*} \Gamma_{\mathrm{R}}^* = 0$ for $\frac{x}{t} \in (v_{\mathrm{L}}^*, v_{\mathrm{R}}^*)$ since $\rho = 0$. Hence $\Gamma(x, t) \geq \rho(x, t)$ for a.e. $x \in \mathbb{R}$ and all $t > 0$.

In the case that $p_\infty > 0$, then $p = -p_\infty$ implies that $\frac{\rho(e-q)}{1-b\rho} = p_\infty$. In particular, as $\rho \downarrow 0$, we have that $e \to \infty$. Nevertheless, we have that, $\widetilde{\boldsymbol{u}} \in \overline{\widetilde{\mathcal{B}}}(b, q, p_\infty)$.

For $p_\infty = 0$, we see that $e \downarrow q$ as $\rho \downarrow 0$. Hence $\widetilde{\boldsymbol{u}} \in \{\widetilde{\boldsymbol{w}} \in \mathbb{R}^{d+3} : \rho \geq 0,\ 1 - b\rho > 0,\ e - q \geq 0,\ \Gamma \geq \rho\} \subset \overline{\widetilde{\mathcal{B}}}(b, q, p_\infty)$. □

4.6 Upper Bound on the Max Wave Speed

We now have the following important result regarding the max wave speed.

Theorem 4.6.1 (Upper Bound on the Max Wave Speed)**.** *Let* $\widehat{\lambda}_{\max} := \lambda_{\max}(\widehat{\boldsymbol{p}}^*)$*, where* $\widehat{\boldsymbol{p}}^*$ *is any upper bound on* $\boldsymbol{p}^*$*, where* $\boldsymbol{p}^*$ *is the root of equation* (4.59)*. Then* $\widehat{\lambda}_{\max} \geq \lambda_{\max}(\boldsymbol{p}^*)$

and the update (4.1) *is invariant domain preserving with* $\widehat{\lambda}$.

Recall that the equation to find p^* requires the solution of a nonlinear equation. This requires the use of a root finding method such as the Newton-Raphson method or the Newton-Secant method. One can also use a quadratic Newton method since $\varphi'''(p) > 0$ for all $p \geq -p_\infty$, details on this can be found in [59]. Unfortunately, despite this, it is still expensive to solve this nonlinear equation. We can alternatively use an upper estimate, $\widehat{p}^*$ on p^* to compute the max wave speed since the invariant-domain preserving properties remain unchanged. In this section, we outline the details for the computation of a close upper bound.

We begin by first introducing the *double-rarefaction* approximation of the function $\varphi(p)$,

$$\begin{aligned}\varphi_{RR}(p) := \frac{2a_{\mathrm{L}}(1-b\rho_{\mathrm{L}})}{\gamma_{\mathrm{L}}-1}\left(\left(\frac{p+p_\infty}{p_{\mathrm{L}}+p_\infty}\right)^{\frac{\gamma_{\mathrm{L}}-1}{2\gamma_{\mathrm{L}}}}-1\right) \\ + \frac{2a_{\mathrm{R}}(1-b\rho_{\mathrm{R}})}{\gamma_{\mathrm{R}}-1}\left(\left(\frac{p+p_\infty}{p_{\mathrm{R}}+p_\infty}\right)^{\frac{\gamma_{\mathrm{R}}-1}{2\gamma_{\mathrm{R}}}}-1\right)+v_{\mathrm{R}}-v_{\mathrm{L}},\end{aligned} \tag{4.73}$$

Lemma 4.6.1 (Approximation of φ)**.** *If* $\max(\gamma_{\mathrm{L}},\gamma_{\mathrm{R}}) \leq \frac{5}{3}$, *then* $\varphi_{RR}(p) \leq \varphi(p)$ *for all* $p \geq -p_\infty$. *In particular,* $\varphi_{RR}(p) = \varphi(p)$ *for* $p \in [-p_\infty, \min(p_{\mathrm{L}}, p_{\mathrm{R}})]$.

Proof. The proof for the case when $p_\infty = 0$ can be seen in [59, Theorem 4.1]. For $p_\infty \neq 0$, the proof is just a translation of the original. □

Note that the current approximation with φ_{RR} fails for $\gamma_{\mathrm{Z}} > \frac{5}{3}$. To address this we have the following lemma,

Lemma 4.6.2 (Approximation of $f_{\mathrm{Z}}^{\mathrm{shock}}(p)$)**.** *For all* $p > p_{\mathrm{Z}}$, *we have that* $f^{\mathrm{shock}}(p) \geq c(\gamma_Z) f^{\mathrm{exp}}(p)$, *where*

$$c(\gamma_Z) := \begin{cases} 1, & \text{if } 1 < \gamma_Z \leq \frac{5}{3} \\ \left(\frac{1}{2}+\frac{4}{3(\gamma_Z+1)}\right)^{\frac{1}{2}}, & \text{if } \frac{5}{3} \leq \gamma_Z \leq 3, \\ \left(\frac{1}{2}+\frac{2}{\gamma_Z-1}3^{\frac{4-2\gamma_Z}{\gamma_Z-1}}\right)^{\frac{1}{2}}, & \text{if } 3 \leq \gamma_Z. \end{cases} \tag{4.74}$$

Proof. The proof can be found in the supplementary material to [2] at [local/web 319KB] □

Notice in passing that $\gamma_{\mathrm{Z}} \mapsto c(\gamma_{\mathrm{Z}})$ is continuous and $c(\gamma_{\mathrm{Z}}) \in (\frac{1}{\sqrt{2}}, 1]$. Lemma 4.6.2 enables us to get an alternative double-rarefaction approximation of the $\varphi(\boldsymbol{p})$ as we will see shortly. To simplify upcoming notation we define the following **subscripts**. Let

$$\min := \begin{cases} \mathrm{L}, & \text{if } p_{\mathrm{L}} \leq p_{\mathrm{R}}, \\ \mathrm{R}, & \text{if } p_{\mathrm{L}} > p_{\mathrm{R}}, \end{cases} \qquad \max := \begin{cases} \mathrm{R}, & \text{if } p_{\mathrm{L}} \leq p_{\mathrm{R}}, \\ \mathrm{L}, & \text{if } p_{\mathrm{L}} > p_{\mathrm{R}}. \end{cases} \tag{4.75}$$

With this notation we have that $p_{\min} = \min(p_{\mathrm{L}}, p_{\mathrm{R}})$ and $p_{\max} = \max(p_{\mathrm{L}}, p_{\mathrm{R}})$. Or in other words, $\min = \arg\min_{\mathrm{Z}\in\{\mathrm{L},\mathrm{R}\}}(p_{\mathrm{Z}})$ and $\max = \arg\max_{\mathrm{Z}\in\{\mathrm{L},\mathrm{R}\}}(p_{\mathrm{Z}})$. Similarly, we define the following subscripts in relation to γ_{Z},

$$m := \begin{cases} \mathrm{L}, & \text{if } \gamma_{\mathrm{L}} \leq \gamma_{\mathrm{R}}, \\ \mathrm{R}, & \text{if } \gamma_{\mathrm{L}} > \gamma_{\mathrm{R}}, \end{cases} \qquad M := \begin{cases} \mathrm{R}, & \text{if } \gamma_{\mathrm{L}} \leq \gamma_{\mathrm{R}}, \\ \mathrm{L}, & \text{if } \gamma_{\mathrm{L}} > \gamma_{\mathrm{R}}. \end{cases} \tag{4.76}$$

So $\gamma_m = \min(\gamma_{\mathrm{L}}, \gamma_{\mathrm{R}})$ and $\gamma_M = \max(\gamma_{\mathrm{L}}, \gamma_{\mathrm{R}})$. It is important to notice that the subscripts m and min do not have to coincide (similarly for M and max). It is worth emphasizing that "min" and "max" are **subscripts** representing either L or R.

We are now ready to define the approximation $\widehat{p}^*$ for p^*.

4.6.1 Case 0: Vacuum

In the case of vacuum, i.e., $v_{\mathrm{R}} - v_{\mathrm{L}} \geq \frac{2a_{\mathrm{L}}(1-b\rho_{\mathrm{L}})}{\gamma_{\mathrm{L}}-1} + \frac{2a_{\mathrm{R}}(1-b\rho_{\mathrm{R}})}{\gamma_{\mathrm{R}}-1}$, there is no root $\varphi(\boldsymbol{p}) = 0$ and hence $p^* := 0$. This implies that $\lambda_1^-(0)v_{\mathrm{L}} - a_{\mathrm{L}}$ and $\lambda_3^+(0) = v_{\mathrm{R}} + a_{\mathrm{R}}$, which is what we expect as vacuum only occurs for a double expansion. Hence, we simply set $\widehat{p}^* = 0$.

4.6.2 Case 1: $p^* > 0$ and $\varphi(p_{\min}) > 0$

In this case the solution to the Riemann problem is composed of two expansions and therefore, the equation to solve is $\varphi_{RR}(\boldsymbol{p}) = 0$. However, for $\gamma_{\mathrm{L}} \neq \gamma_{\mathrm{R}}$, the equation is nonlinear. Fortunately for us, we only need the maximum wave speed, therefore an estimate

for p^* is not necessary since, $\lambda_1^-(p^*) = v_{\rm L} - a_{\rm L}$ and $\lambda_3^+(p^*) = v_{\rm R} + a_{\rm R}$. If one does need an approximation to p^*, then one can solve the equation $\widehat{\varphi}(p) = 0$ where,

$$\widehat{\varphi}_{RR}(p) := \alpha_{\rm L}\left(\left(\frac{p+p_\infty}{p_{\rm L}+p_\infty}\right)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + \alpha_{\rm R}\left(\left(\frac{p+p_\infty}{p_{\rm R}+p_\infty}\right)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + v_{\rm R} - v_{\rm L}. \tag{4.77}$$

Note that $\widehat{\varphi}_{RR}(p) \le \varphi_{RR}(p) = \varphi(p)$ for all $p \in [-p_\infty, p_{\min}]$. The associated root of $\widehat{\varphi}(p) = 0$ is,

$$\widetilde{p}^* = \left(\frac{\max(\alpha_{\rm R} + \alpha_{\rm L} - (v_{\rm R} - v_{\rm L}), 0)}{\alpha_{\rm R}(p_{\rm R}+p_\infty)^{-\frac{\gamma_M-1}{2\gamma_M}} + \alpha_{\rm L}(p_{\rm L}+p_\infty)^{-\frac{\gamma_M-1}{2\gamma_M}}}\right)^{\frac{2\gamma_M}{\gamma_M-1}} - p_\infty \tag{4.78}$$

Note that $\widetilde{p}^* \ge p^*$, but by assumption, we also have that $p_{\min} \ge p^*$. Therefore, we define the upper estimate by $\widehat{p}^* := \min(\widetilde{p}^*, p_{\min})$.

4.6.3 Case 2: $\varphi(p_{\min}) \le 0 \le \varphi(p_{\max})$

In this case, we have that $p^* \in [p_{\min}, p_{\max}]$ and therefore, the "min" wave is a shock; that is $f_{\min}(p) = f_{\min}^{\rm shock}(p)$ and the "max" wave is an expansion; i.e., $f_{\max}(p) = f_{\max}^{\rm exp}(p)$. Therefore, using Lemma 4.6.2 we approximate $f_{\min}^{\rm shock}$ by, $c(\gamma_{\min}) f_{\min}^{\rm exp}(p) \le f_{\min}^{\rm shock}(p)$ for $p \in [p_{\min}, p_{\max}]$ and leave $f_{\max}^{\rm exp}$ unchanged. That is, the lower bounding function is,

$$\widehat{\varphi}_{RR}(p) := \widehat{\alpha}_{\min}\left(\left(\frac{p+p_\infty}{p_{\min}+p_\infty}\right)^{\frac{\gamma_{\min}-1}{2\gamma_{\min}}} - 1\right) + \alpha_{\max}\left(\left(\frac{p+p_\infty}{p_{\max}+p_\infty}\right)^{\frac{\gamma_{\max}-1}{2\gamma_{\max}}} - 1\right) + v_{\rm R} - v_{\rm L}. \tag{4.79}$$

Unfortunately, this is still a nonlinear equation. So we need to coarsen our approximation a bit more. Two do this we declare two sub cases.

4.6.3.1 *Case 2a:* $\gamma_{\min} = \gamma_m$

Define,

$$\widehat{\varphi}_1(p) := \widehat{\alpha}_{\min}\left(\delta\Big(\frac{p+p_\infty}{p_{\min}+p_\infty}\Big)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + \alpha_{\max}\left(\Big(\frac{p+p_\infty}{p_{\max}+p_\infty}\Big)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + v_{\mathrm{R}} - v_{\mathrm{L}}, \tag{4.80a}$$

$$\widehat{\varphi}_2(p) := \widehat{\alpha}_{\min}\left(\Big(\frac{p+p_\infty}{p_{\min}+p_\infty}\Big)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\right) + \alpha_{\max}\left(\delta\Big(\frac{p+p_\infty}{p_{\max}+p_\infty}\Big)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\right) + v_{\mathrm{R}} - v_{\mathrm{L}}, \tag{4.80b}$$

where $\delta := \big(\frac{p_{\min}+p_\infty}{p_{\max}+p_\infty}\big)^{\frac{\gamma_M-\gamma_m}{2\gamma_m\gamma_M}}$. Observe that,

$$\begin{aligned}
\Big(\frac{p+p_\infty}{p_{\min}+p_\infty}\Big)^{\frac{\gamma_m-1}{2\gamma_m}} &\geq \Big(\frac{p+p_\infty}{p_{\max}+p_\infty}\Big)^{\frac{\gamma_M-1}{2\gamma_M}}\Big(\frac{p_{\max}+p_\infty}{p_{\min}+p_\infty}\Big)^{\frac{\gamma_m-1}{2\gamma_m}} \\
&= (p+p_\infty)^{\frac{\gamma_M-1}{2\gamma_M}}(p_{\max}+p_\infty)^{\frac{\gamma_m-1}{2\gamma_m}-\frac{\gamma_M-1}{2\gamma_M}}(p_{\min}+p_\infty)^{-\frac{\gamma_m-1}{2\gamma_m}} \\
&= \Big(\frac{p+p_\infty}{p_{\min}+p_\infty}\Big)^{\frac{\gamma_M-1}{2\gamma_M}}\Big(\frac{p_{\min}+p_\infty}{p_{\max}+p_\infty}\Big)^{\frac{\gamma_M-\gamma_m}{2\gamma_m\gamma_M}}.
\end{aligned}$$

Therefore, $\widehat{\varphi}_1(p) \leq \widehat{\varphi}(p)$ for $p \in [p_{\min}, p_{\max}]$. We can apply the same reasoning to show $\widehat{\varphi}_2(p) \leq \widehat{\varphi}(p)$. Thus $\max(\widehat{\varphi}_1(p), \widehat{\varphi}_2(p)) \leq \widehat{\varphi}(p)$ for $p \in [p_{\min}, p_{\max}]$. The roots of (4.80a) and (4.80b) are,

$$\widetilde{p}_1^* = \left(\frac{\widehat{\alpha}_{\min} + \alpha_{\max} - (v_{\mathrm{R}} - v_{\mathrm{L}})}{\delta\widehat{\alpha}_{\min}(p_{\min}+p_\infty)^{-\frac{\gamma_M-1}{2\gamma_M}} + \alpha_{\max}(p_{\max}+p_\infty)^{-\frac{\gamma_M-1}{2\gamma_M}}}\right)^{\frac{2\gamma_M}{\gamma_M-1}} - p_\infty \tag{4.81a}$$

$$\widetilde{p}_2^* = \left(\frac{\widehat{\alpha}_{\min} + \alpha_{\max} - (v_{\mathrm{R}} - v_{\mathrm{L}})}{\widehat{\alpha}_{\min}(p_{\min}+p_\infty)^{-\frac{\gamma_m-1}{2\gamma_m}} + \delta\alpha_{\max}(p_{\max}+p_\infty)^{-\frac{\gamma_m-1}{2\gamma_m}}}\right)^{\frac{2\gamma_m}{\gamma_m-1}} - p_\infty \tag{4.81b}$$

Thus, our approximation for p^* is $\widehat{p}^* := \min(\widetilde{p}_1^*, \widetilde{p}_2^*, p_{\max})$.

4.6.4 Case 2b: $\gamma_{\min} = \gamma_M$

This case is very similar to Case 2a. Since $\gamma_{\min} = \gamma_M$ and $\gamma_{\max} = \gamma_M$ we have that the following lower bounds on $\widehat{\varphi}$,

$$\widehat{\varphi}_1(p) := \widehat{\alpha}_{\min}\left(\left(\frac{p+p_\infty}{p_{\min}+p_\infty}\right)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\right) + \alpha_{\max}\left(\left(\frac{p+p_\infty}{p_{\max}+p_\infty}\right)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\right) + v_{\mathrm{R}} - v_{\mathrm{L}}, \tag{4.82a}$$

$$\widehat{\varphi}_2(p) := \widehat{\alpha}_{\min}\left(\left(\frac{p+p_\infty}{p_{\min}+p_\infty}\right)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + \alpha_{\max}\left(\left(\frac{p+p_\infty}{p_{\max}+p_\infty}\right)^{\frac{\gamma_M-1}{2\gamma_M}} - 1\right) + v_{\mathrm{R}} - v_{\mathrm{L}}, \tag{4.82b}$$

and they have the corresponding roots,

$$\widetilde{p}_1^* = \left(\frac{\widehat{\alpha}_{\min} + \alpha_{\max} - (v_{\mathrm{R}} - v_{\mathrm{L}})}{\widehat{\alpha}_{\min}(p_{\min}+p_\infty)^{-\frac{\gamma_m-1}{2\gamma_m}} + \alpha_{\max}(p_{\max}+p_\infty)^{-\frac{\gamma_m-1}{2\gamma_m}}}\right)^{\frac{2\gamma_m}{\gamma_m-1}} - p_\infty, \tag{4.83a}$$

$$\widetilde{p}_2^* = \left(\frac{\widehat{\alpha}_{\min} + \alpha_{\max} - (v_{\mathrm{R}} - v_{\mathrm{L}})}{\widehat{\alpha}_{\min}(p_{\min}+p_\infty)^{-\frac{\gamma_M-1}{2\gamma_M}} + \alpha_{\max}(p_{\max}+p_\infty)^{-\frac{\gamma_M-1}{2\gamma_M}}}\right)^{\frac{2\gamma_M}{\gamma_M-1}} - p_\infty. \tag{4.83b}$$

Therefore our upper bound on p^* is $\widehat{p}^* := \min(\widetilde{p}_1^*, \widetilde{p}_2^*, p_{\max})$.

4.6.4.1 Case 3: $\varphi(p_{\max}) < 0$

In this case, we have $f_{\min}^{\mathrm{shock}}(p) \geq c(\gamma_{\min}) f_{\min}^{\exp}$ and $f_{\max}^{\mathrm{shock}}(p) \geq c(\gamma_{\max}) f_{\max}^{\exp}$ for all $p \geq p_{\max}$. Therefore, a lower approximation of φ is,

$$\widehat{\varphi}_{RR}(p) := \widehat{\alpha}_{\min}\left(\left(\frac{p+p_\infty}{p_{\min}+p_\infty}\right)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\right) + \widehat{\alpha}_{\max}\left(\left(\frac{p+p_\infty}{p_{\max}+p_\infty}\right)^{\frac{\gamma_m-1}{2\gamma_m}} - 1\right) + v_{\mathrm{R}} - v_{\mathrm{L}}, \tag{4.84}$$

with the corresponding root,

$$\widetilde{p}_1^* = \left(\frac{\widehat{\alpha}_{\min} + \widehat{\alpha}_{\max} - (v_{\mathrm{R}} - v_{\mathrm{L}})}{\widehat{\alpha}_{\min}(p_{\min}+p_\infty)^{-\frac{\gamma_m-1}{2\gamma_m}} + \widehat{\alpha}_{\max}(p_{\max}+p_\infty)^{-\frac{\gamma_m-1}{2\gamma_m}}}\right)^{\frac{2\gamma_m}{\gamma_m-1}} - p_\infty. \tag{4.85}$$

Alternatively, we can make use of the fact that φ is composed of the two shock curves, $f_{\mathrm{L}}^{\mathrm{shock}}$ and $f_{\mathrm{R}}^{\mathrm{shock}}$ and derive an alternative lower bound. For $p \in (p_{\max}, \infty)$ we have that $B_{\mathrm{Z}} \leq B_{\mathrm{Z}} p p_{\max}^{-1}$. Therefore,

$$f_{\mathrm{Z}}^{\mathrm{shock}}(p) = (p - p_{\mathrm{Z}})\sqrt{\frac{A_{\mathrm{Z}}}{p + B_{\mathrm{Z}}}} \geq \frac{p - p_{\mathrm{Z}}}{\sqrt{p}}\sqrt{\frac{A_{\mathrm{Z}}}{1 + B_{\mathrm{Z}} p_{\max}^{-1}}} \tag{4.86}$$

and hence we can define the following lower bound on φ,

$$\widehat{\varphi}_{SS}(p) := \frac{p - p_{\mathrm{L}}}{\sqrt{p}}\sqrt{\frac{A_{\mathrm{L}}}{1 + B_{\mathrm{L}} p_{\max}^{-1}}} + \frac{p - p_{\mathrm{R}}}{\sqrt{p}}\sqrt{\frac{A_{\mathrm{R}}}{1 + B_{\mathrm{R}} p_{\max}^{-1}}} + v_{\mathrm{R}} - v_{\mathrm{L}}. \tag{4.87}$$

Solving $\widehat{\varphi}_{SS}(p) = 0$ is equivalent to solving the quadratic equation,

$$(x_{\mathrm{L}} + x_{\mathrm{R}})p + (v_{\mathrm{R}} - v_{\mathrm{L}})\sqrt{p} - \Big(p_{\mathrm{L}} x_{\mathrm{L}} + p_{\mathrm{R}} x_{\mathrm{L}}\Big) = 0, \tag{4.88}$$

where $x_{\mathrm{Z}} = \sqrt{\frac{A_{\mathrm{Z}}}{1 + B_{\mathrm{Z}} p_{\max}^{-1}}}$. Define $a := x_{\mathrm{L}} + x_{\mathrm{R}}$, $b := v_{\mathrm{R}} - v_{\mathrm{L}}$, $c := -p_{\mathrm{L}} x_{\mathrm{L}} - p_{\mathrm{R}} x_{\mathrm{R}}$. Then the root of this equation is,

$$\widetilde{p}_2^* = \Big(\frac{-b + \sqrt{b^2 - 4ac}}{2a}\Big)^2 \tag{4.89}$$

Therefore, the upper bound on p^* is $\widehat{p}^* := \min(\widetilde{p}_1^*, \widetilde{p}_2^*)$.

This completes all possible cases for determining $\widehat{p}^*$ such that $\widehat{p}^* \geq p^*$. The psuedocode for computing the maximum wave speed with $\widehat{p}^*$ is given in Algorithm 1.

5. HIGH ORDER APPROXIMATION OF THE EULER EQUATIONS WITH TABULATED EOS*

In this chapter, we develop a second order method for the Euler equations with a tabulated EOS. This builds on the first order method described in Chapter 4. The foundations of this high order method are based on Guermond et. al. [60] and we will extend the ideas presented there to hold for a tabulated or arbitrary equation of state. This extension is taken from Clayton et. al. [1].

5.1 The Use of the Consistent Mass Matrix

In the low order method we used the *lumped* mass matrix, $(m_i)_{i\in\mathcal{V}}$ as opposed to the consistent mass matrix, $(m_{ij})_{i,j\in\mathcal{V}}$, in order to prove the method was invariant-domain preserving. Furthermore, it was shown in [61], that the use of the consistent mass matrix always violates the maximum principle in a scalar conservation law. However, it is necessary to use the consistent mass matrix for the high order method as using the lumped mass matrix increases dispersion errors. This can be seen in [62]. Additionally, the use of the consistent mass matrix can provide superconvergence effects; see [63]. See also [64]. The high order update is,

$$\frac{1}{\Delta t_n}\sum_{j\in\mathcal{G}(i)} m_{ij}(\mathbf{U}_j^{n+1}-\mathbf{U}_j^n)+\sum_{j\in\mathcal{G}(i)}\mathbb{f}(\mathbf{U}_j^n)\boldsymbol{c}_{ij}-d_{ij}^{\mathsf{H},n}(\mathbf{U}_j^n-\mathbf{U}_i^n)=\mathbf{0}, \tag{5.1}$$

where we use a higher order graph viscosity, $d_{ij}^{\mathsf{H},n}$. The choice of this graph viscosity is discussed in Section 5.3. A problem with the use of the consistent mass matrix is that it requires numerical matrix solvers to determine the solution. This can increase the computational time even with the use of a preconditioner. We avoid the use of matrix solvers, by using a Neumann series expansion.

* A majority of this chapter is a modification of the work done in [1] and is reprinted with permission from [1].

Proposition 5.1.1 (Neumann Series Expansion). *Let $(X, \|\cdot\|)$ be a Banach space and $T : X \to X$ a bounded linear operator. If $\|T\| < 1$, then*

$$(I - T)^{-1} = \sum_{n=0}^{\infty} T^n \tag{5.2}$$

where I is the identity map and $T^0 = I$.

The application of the Neumann series to (5.1) can be seen in [65, Sec. 3.4]. We review this application of the Neumann series expansion. Rewrite (5.1) as,

$$\sum_{j \in \mathcal{G}(i)} \frac{m_{ij}}{m_j} \frac{m_j}{\Delta t_n} (\mathbf{U}_j^{n+1} - \mathbf{U}_j^n) + \sum_{j \in \mathcal{G}(i)} \mathbb{f}(\mathbf{U}_j^n) \boldsymbol{c}_{ij} - d_{ij}^{\mathsf{H},n} (\mathbf{U}_j^n - \mathbf{U}_i^n) = \mathbf{0}, \tag{5.3}$$

Let $\mathbb{M}$ be the matrix with entries $(m_{ij}/m_j)_{i,j \in \mathcal{V}}$. Then the solution, $\mathbf{U}_i^{n+1}$, can be found inverting $\mathbb{M}$. By Proposition 5.1.1, we have $\mathbb{M}^{-1} = (\mathbb{I} - (\mathbb{I} - \mathbb{M}))^{-1} = \sum_{n=0}^{\infty} (\mathbb{I} - \mathbb{M})^n$, since. This series converges if $\|\mathbb{I} - \mathbb{M}\| < 1$. Note for small enough mesh size, h, we have that $\sup_{\|\boldsymbol{x}\|_{\ell^\infty} = 1} \|(\mathbb{I} - \mathbb{M})\boldsymbol{x}\|_{\ell^\infty} < 1$. We approximate this Neumann series expansion by taking only the first two terms of the summation; that is, $\mathbb{M}^{-1} \approx \mathbb{I} + \mathbb{B}$ where $\mathbb{B} = (b_{ij})_{i,j \in \mathcal{V}}$ with $b_{ij} := \delta_{ij} - \frac{m_{ij}}{m_j}$. Next, let $I := \operatorname{card}(\mathcal{V})$ and $\mathsf{R} \in \mathbb{R}^I$. Then we have the following,

$$(\mathbb{M}^{-1}\mathsf{R})_i \approx (\mathsf{R} + \mathbb{B}\mathsf{R})_i = \mathsf{R}_i + \sum_{j \in \mathcal{G}(i)} b_{ij} \mathsf{R}_j = \mathsf{R}_i + \sum_{j \in \mathcal{G}(i)} (b_{ij} \mathsf{R}_j - b_{ji} \mathsf{R}_i), \tag{5.4}$$

Notice that $\sum_{j \in \mathcal{G}(i)} b_{ji} \mathsf{R}_i = 0$ since $\sum_{j \in \mathcal{G}(i)} m_{ji} = m_i$ implies $\sum_{j \in \mathcal{G}(i)} b_{ji} = 0$. Applying this to (5.3) the second order update is given by,

$$\frac{m_i}{\Delta t} (\mathbf{U}_i^{\mathsf{H},n+1} - \mathbf{U}_i^n) = \mathbf{R}_i^n + \sum_{j \in \mathcal{G}(i)} (b_{ij} \mathbf{R}_j^n - b_{ji} \mathbf{R}_i^n), \tag{5.5a}$$

$$\text{where} \quad \mathbf{R}_i^n := \sum_{j \in \mathcal{G}(i)} \big(- \mathbb{f}(\mathbf{U}_j^n) \boldsymbol{c}_{ij} + d_{ij}^{\mathsf{H},n} (\mathbf{U}_j^n - \mathbf{U}_i^n) \big) \tag{5.5b}$$

A more generalized version of the Neumann series approximation is proved by Guermond

& Pasquetti in [62, Sec. 3.1] in the context of transport type problems.

5.2 A Review of the Entropy Solution

In order to define a higher order graph viscosity, we are motivated by the method of the "entropy viscosity commutator" introduced in Guermond et. al. [60, Section 3.4]. The idea is to measure the discrete relative error in the entropy commutator, $\nabla\cdot\boldsymbol{F}(\boldsymbol{u})-(D\eta(\boldsymbol{u}))^{\mathsf{T}}\nabla\cdot\mathbb{f}(\boldsymbol{u})$ where $(\eta,\boldsymbol{F})$ is the entropy pair, see Theorem 5.2.1. We first review some results regarding entropy solutions.

Definition 5.2.1 (Entropy, Entropy-Flux Pair)**.** We say that $(\eta(\boldsymbol{u}),\boldsymbol{F}(\boldsymbol{u}))$ is an *entropy, entropy-flux pair* for the Euler equations, (1.14) if

$$\nabla\cdot\boldsymbol{F}(\boldsymbol{u})=(D\eta(\boldsymbol{u}))^{\mathsf{T}}\nabla\cdot\mathbb{f}(\boldsymbol{u}),\quad\forall\boldsymbol{u}\in\mathcal{B}(b). \tag{5.6}$$

□

Definition 5.2.2. A weak solution $\boldsymbol{u}\in[L^{\infty}(\mathbb{R}^d\times\mathbb{R}_+)]^m$ to the conservation law,

$$\partial_t\boldsymbol{u}+\nabla\cdot\mathbb{f}(\boldsymbol{u})=\boldsymbol{0}, \tag{5.7}$$

for $\mathbb{g}\in C^1(\mathbb{R}^m;\mathbb{R}^{m\times d})$ is said to be an *entropy solution* if,

$$\partial_t\eta(\boldsymbol{u})+\nabla\cdot\boldsymbol{F}(\boldsymbol{u})\geq 0, \tag{5.8}$$

holds in the sense of distributions for every concave entropy $\eta:\mathbb{R}^m\to\mathbb{R}$ and associated entropy-flux, $\boldsymbol{F}:\mathbb{R}^m\to\mathbb{R}^d$. □

Theorem 5.2.1 (Smooth Entropy Solutions)**.** *Every smooth solution to the compressible Euler equations,* (1.14)*, is also an entropy solution.*

Remark 5.2.1 (Mathematical Entropy)**.** Definition 5.2.2 can be equivalently framed by assuming $\partial_t\eta(\boldsymbol{u})+\nabla\cdot\boldsymbol{F}(\boldsymbol{u})\leq 0$ holds in the sense of distributions for η convex. See [66].

For numerical purposes, the distinction is not necessary as we will be measuring the discrete error of the quantity, $\nabla \cdot \boldsymbol{F}(\boldsymbol{u}) - (D\eta(\boldsymbol{u}))^{\mathsf{T}} \nabla \cdot \mathbb{f}(\boldsymbol{u})$. $\square$

For the compressible Euler equations with ideal gas law, it is unknown if there is a characterization of all the entropy pairs $(\eta, \boldsymbol{F})$ like the class of Kruzhkov entropies for scalar conservation laws, see Kruzhkov [29]. However, we will exploit a class of known entropy, entropy-flux pairs. They are the so-called *generalized entropy pairs*, defined by $\eta(\boldsymbol{u}) = -\rho f(\mathsf{s}(\boldsymbol{u}))$ where $f \in C^2(\mathbb{R})$, $\boldsymbol{F}(\boldsymbol{u}) = \frac{\boldsymbol{m}}{\rho}\eta(\boldsymbol{u})$, where $f'(s) > 0$, $f'(s)c_p^{-1} - f''(s) > 0$ for all $(\rho, e) \in \mathcal{A}$. In this case η is convex (see [67] and [66]).

The simplest non-trivial entropy pair is when $f(s) = 1$; that is, $\eta(\boldsymbol{u}) = -\rho$ and $\boldsymbol{F}(\boldsymbol{u}) = \boldsymbol{m}$. However, the entropy, η is not strictly convex. For $f(s) = s$, $\eta(\boldsymbol{u}) = -\rho\mathsf{s}(\boldsymbol{u})$ which is the mathematical entropy (the negative of the physical entropy). Another example is the so-called *Harten entropy* given by $\eta(\boldsymbol{u}) = \rho \exp(\frac{\gamma-1}{\gamma+\alpha}\mathsf{s}(\boldsymbol{u}))$ for $\alpha > 0$. This was shown for the ideal EOS in [67]; however, it also holds for the covolume and NASG as $c_p = \frac{\gamma}{\gamma-1}$ for each of these EOS. It should be noted that these examples are not entropy pairs for every equation of state. Notably, if the equation of state is non-convex, then the entropies proposed are not convex.

If the interpolatory EOS is covolume, then the specific entropy is given by, $\mathsf{s}(\boldsymbol{u}) := \log\big((\mathsf{e}(\boldsymbol{u})^{\frac{1}{\gamma-1}}(\rho^{-1} - b)\big)$. Then the Harten entropy can be written in the form,

$$\eta(\boldsymbol{u}) = \Big(\frac{\rho^{\alpha+1}\mathsf{e}(\boldsymbol{u})}{(1-b\rho)^{1-\gamma}}\Big)^{\frac{1}{\gamma+\alpha}}. \tag{5.9}$$

Note that a nice choice for the Harten entropy is to set $\alpha = 1$, since $\rho^2\mathsf{e}(\boldsymbol{u}) = \rho E - \frac{1}{2}\|\boldsymbol{m}\|_{\ell^2}^2$, which is a quadratic function of the conserved variables.

5.3 The Entropy Viscosity Method

Motivated by notion of the entropy commutator, $\nabla\cdot\boldsymbol{F}(\boldsymbol{u})-(D\eta(\boldsymbol{u}))^{\mathsf{T}}\nabla\cdot\mathbb{f}(\boldsymbol{u})$, we propose to find a local entropy $\eta^{n,i}$ at every node $i\in\mathcal{V}$ and every time t_n. We define the local flux,

$$\mathbb{f}^{n,i}(\boldsymbol{u}) := \begin{pmatrix} \boldsymbol{m} \\ \boldsymbol{v}\otimes\boldsymbol{m} + p^{n,i}_{\mathsf{nasg}}(\boldsymbol{u})\mathbb{I}_d \\ \boldsymbol{v}(E+p^{n,i}_{\mathsf{nasg}}(\boldsymbol{u})) \end{pmatrix} \tag{5.10}$$

where $p^{n,i}_{\mathsf{nasg}}(\boldsymbol{u}) := (\gamma_i^{\min,n}-1)\Big(\frac{\rho(\mathsf{e}(\boldsymbol{u})-q)}{1-b\rho}-p_\infty\Big)-p_\infty$, $\gamma_i^{\min,n} := \min_{j\in\mathcal{J}(i)}\gamma_j^n$, and

$$\gamma_j^n := \frac{(p_j^n+p_\infty)(1-b\varrho_j^n)}{\varrho_j^n(\mathsf{e}(\mathbf{U}_j^n)-q)-(1-b\varrho_j^n)p_\infty}+1. \tag{5.11}$$

Note that, $p^{n,i}_{\mathsf{nasg}}$ is the NASG EOS with $\gamma=\gamma_i^{\min,n}$. Then two possible entropy pairs for the Euler equations with flux $\mathbb{f}^{n,i}$ are,

$$\eta_1^{i,n}(\boldsymbol{u}) := -\rho\log\left((e(\boldsymbol{u})-q-p_\infty(\rho^{-1}-b))^{\frac{1}{\gamma_i^{\min,n}-1}}(\rho^{-1}-b)\right)-\frac{\rho}{\rho_i^n}\eta^{i,n}_{\mathrm{ref},1}, \tag{5.12a}$$

$$\boldsymbol{F}_1^{n,i}(\boldsymbol{u}) := \boldsymbol{v}\eta_1^{n,i}(\boldsymbol{u}), \tag{5.12b}$$

where $\eta^{i,n}_{\mathrm{ref},1} := -\rho_i^n\log\left((e_i^n-q-p_\infty((\rho_i^n)^{-1}-b))^{\frac{1}{\gamma_i^{\min,n}-1}}((\rho_i^n)^{-1}-b)\right)$ and

$$\eta_2^{n,i}(\boldsymbol{u}) := \Big(\frac{\rho^{\alpha+1}(\mathsf{e}(\boldsymbol{u})-q)-p_\infty\rho^\alpha(1-b\rho)}{(1-b\rho)^{1-\gamma_i^{\min,n}}}\Big)^{\frac{1}{\gamma_i^{\min,n}+\alpha}}-\frac{\rho}{\rho_i^n}\eta^{n,i}_{\mathrm{ref},2}, \tag{5.13a}$$

$$\boldsymbol{F}_2^{n,i}(\boldsymbol{u}) := \boldsymbol{v}\eta_2^{n,i}(\boldsymbol{u}), \tag{5.13b}$$

where $\eta^{n,i}_{\mathrm{ref},2} = \Big(\frac{(\rho_i^n)^{\alpha+1}(\mathsf{e}(\mathbf{U}_i^n)-q)-p_\infty(\rho_i^n)^\alpha(1-b\rho_i^n)}{(1-b\rho_i^n)^{1-\gamma_i^{\min,n}}}\Big)^{\frac{1}{\gamma_i^{\min,n}+\alpha}}$, for $\alpha>0$. Note that each entropy has been shifted by $\rho\eta^{n,i}_{\mathrm{ref},k}/\rho_i^n$ so that $\eta_k^{n,i}(\mathbf{U}_i^n)=0$ for $k=1,2$. This shift is valid since $\eta(\boldsymbol{u})=\rho(f(\mathsf{s}(\boldsymbol{u}))-c)$ is also an entropy for any constant c with the corresponding entropy-flux, $\boldsymbol{F}(\boldsymbol{u}) := \boldsymbol{m}(f(\mathsf{s}(\boldsymbol{u}))-c)$.

Remark 5.3.1 (Remarks on $\eta^{n,i}$ and $\gamma_i^{\min,n}$)**.** Note that $\eta_2^{n,i}$ is the (shifted) Harten entropy. The choice of $\alpha = 1$ is also convenient as $\rho^2(\mathsf{e}(\boldsymbol{u})-q)-p_\infty\rho(1-b\rho) = \rho E - \frac{1}{2}\|\boldsymbol{m}\|_{\ell^2}^2 - p_\infty\rho(1-b\rho)$ which is a quadratic function of the conserved variables; this simplifies the computation of $D\eta^{n,i}$.

In the definition of the entropy pairs, (5.12) and (5.13), we could have alternatively used $\gamma_i^{\max,n}$ or simply γ_i^n, as each of these choices will recover the expected Harten entropy if the pressure is given by the ideal, covolume, or NASG equations of state. We choose $\gamma_i^{\min,n}$ as it is already necessary to compute for the limiting of the surrogate physical entropy, see Theorem 6.4.3. □

We may select either of the two entropy, entropy-flux pairs, so we drop the subscript notation for $\eta_k^{i,n}$. With this collection of entropy pairs, $\{(\eta^{n,i}, \boldsymbol{F}^{n,i})\}_{i\in\mathcal{V}}$ for $n \in \mathbb{N}$, we can measure the local error in the entropy viscosity commutator by approximating

$$\int_\Omega \big((\nabla\cdot\boldsymbol{F}^{n,i}(\boldsymbol{u}) - (D\eta^{n,i}(\boldsymbol{u}))^\mathsf{T}\nabla\cdot\mathbb{f}^{n,i}(\boldsymbol{u})\big)\varphi_i(\boldsymbol{x})\,d\boldsymbol{x} \tag{5.14}$$

with our numerical solution $\boldsymbol{u}_h^n = \sum_{i\in\mathcal{V}}\mathbf{U}_i^n\varphi_i$. As mentioned in Section 3.2, the divergence of the flux is approximated by, $\nabla\cdot\mathbb{f}^{n,i}(\boldsymbol{u}_h^n) \approx \sum_{i\in\mathcal{V}}\mathbb{f}^{n,i}(\mathbf{U}_i^n)\nabla\varphi_i$. Similarly, for the entropy flux, $\nabla\cdot\boldsymbol{F}^{n,i}(\boldsymbol{u}_h^n) \approx \sum_{i\in\mathcal{V}}\boldsymbol{F}^{n,i}(\mathbf{U}_i^n)\cdot\nabla\varphi_i$. The discretization of (5.14) is thus given by,

$$N_i^n := \sum_{j\in\mathcal{J}(i)}(\boldsymbol{F}^{n,i}(\mathbf{U}_j^n) - (D\eta(\mathbf{U}_i^n))^\mathsf{T}\mathbb{f}^{n,i}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij}. \tag{5.15}$$

We now define the *entropy residual* as

$$R_i^n := \frac{|N_i^n|}{D_i^n + \varepsilon\max_{j\in\mathcal{V}}D_j^n + \epsilon}, \quad \varepsilon = 10^{-1}, \quad \epsilon = 10^{-14}, \tag{5.16}$$

$$D_i^n := \sum_{j\in\mathcal{J}(i)}|\boldsymbol{F}^{n,i}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}| + \sum_{j\in\mathcal{J}(i)}|(D\eta^{n,i}(\mathbf{U}_i^n))^\mathsf{T}\mathbb{f}(\mathbf{U}_i^n)\cdot\boldsymbol{c}_{ij}|. \tag{5.17}$$

The parameter ϵ is a machine precision parameter used to avoid division by zero.

Remark 5.3.2 (Concave/Convex Entropy)**.** Note that there is no assumption made on the concavity or convexity of $\eta^{n,i}(\boldsymbol{u})$ need not be a concave entropy. We only need that $(\eta, \boldsymbol{F})$ be an entropy, entropy-flux pair so as to measure the smoothness of the solution. Therefore a non-convex entropy is valid in the construction of the entropy residual. □

The high order viscosity is then defined as,

$$d_{ij}^{\mathsf{H},n} := \max(R_i^n, R_j^n) d_{ij}^{\mathsf{L},n}. \tag{5.18}$$

Another alternative for the definition can be,

$$d_{ij}^{\mathsf{H},n} := \frac{R_i^n + R_j^n}{2} d_{ij}^{\mathsf{L},n}. \tag{5.19}$$

Remark 5.3.3 (Thresholding Function)**.** One can further emphasize the result of the residual. That is, define the function,

$$\psi(x) := \frac{4x_0^3 - (x + x_0)(x - 2x_0)\big((x - 2x_0) - \mathrm{ReLU}(x - 2x_0)\big)}{4x_0^3} \tag{5.20}$$

where $\mathrm{ReLU}(x) = (x + |x|)/2$ and $x_0 \in (0, 0.5]$ parameter. Note that ψ satisfies, $\psi(0) = 0$, $\psi(x_0) = \frac{1}{2}$, and $\psi(x) = 1$ for all $x \in [2x_0, 1]$. The fixed point, $\psi(x^*) = x^*$ is computed by $x^* = x_0(\frac{3}{2} - \frac{1}{2}\sqrt{9 - 16x_0})$. Thus we can alternatively define the high order viscosity as,

$$d_{ij}^{\mathsf{H},n} = \max(\psi(R_i^n), \psi(R_j^n)) d_{ij}^{\mathsf{L},n}. \tag{5.21}$$

or using an average of the residuals,

$$d_{ij}^{\mathsf{H},n} := \tfrac{1}{2}(\psi(R_i^n) + \psi(R_j^n)) d_{ij}^{\mathsf{L},n}. \tag{5.22}$$

□

To conclude, the high order viscosity, $d_{ij}^{\mathsf{H},n}$ detects whether the solution is not smooth and if so, takes on values close to the low order artificial viscosity.

6. QUASICONCAVE LIMITING*

Since the second order method is **not** invariant-domain preserving, we must perform a limiting process on the solution to keep it in the invariant domain. The concept of limiting was originally developed by Boris & Book, named the Flux Corrected Transport (FCT) method. This development took course over the three papers [68, 69, 70] and was also extended to multiple dimensions Zalesak in [71]. Numerous work has been done on this methodology; for more resources, see [72].

For this chapter, the limiting is done with respect to some local bounds (see Section 6.1) that are satisfied by the first order method. This limiting process is performed on quasi-concave functionals which was first developed by Guermond et. al. in [60]. We justify in Section 6.3 that limiting with respect to these local bounds preserves the invariant domain properties.

6.1 Local Bounds

Since the high order update is known to violate the invariant domain properties, we need to limit our solution so that it is no longer non-physical. For example, if $\rho_i^{\mathsf{H},n+1} < 0$, then we would like to somehow, pull this high order solution closer to the low order solution. That is, a mapping, $\mathcal{L} : \rho_i^{\mathsf{H},n+1} \mapsto \rho_i^{n+1}$, such that $\rho_i^{n+1} > 0$. The challenge is to do this globally, efficiently and only when necessary. The technique we employ is based on local bounds that are satisfied by the low order solution, which consequently preserve the invariant-domain properties.

Theorem 6.1.1. *The low order update,* $\mathbf{U}_i^{\mathsf{L},n+1}$*, defined in* (4.2)*, satisfies the following local*

* A majority of this chapter is a modification of the work done in [1] and is reprinted with permission from [1].

bounds,

$$\rho_i^{\min,n} := \min_{j\in\mathcal{J}(i)}(\overline{\rho}_{ij}^n) \le \rho_i^{\mathsf{L},n+1} \le \max_{j\in\mathcal{J}(i)}(\overline{\rho}_{ij}^n) =: \rho_i^{\max,n} \tag{6.1}$$

$$\mathsf{E}_i^{\min,n} := \min_{j\in\mathcal{J}(i)}(\overline{\mathsf{E}}_{ij}^n) \le \mathsf{E}_i^{\mathsf{L},n+1} \le \max_{j\in\mathcal{J}(i)}(\overline{\mathsf{E}}_{ij}^n) =: \mathsf{E}_i^{\max,n} \tag{6.2}$$

$$\Upsilon_i^{\min} := \min_{j\in\mathcal{J}(i)}(\rho\mathsf{e})(\overline{\mathbf{U}}_{ij}^n) \le (\rho\mathsf{e})(\mathbf{U}_i^{\mathsf{L},n+1}) \tag{6.3}$$

for all $i \in \mathcal{V}$.

Proof. The result for (6.1)–(6.2) immediately follows from the definition of the low order states as a convex combination of bar state and that $\overline{\rho}_{ij}^n > 0$ and $\overline{\mathsf{E}}_{ij}^n > 0$ for all $i \in \mathcal{V}$ and $j \in \mathcal{J}(i)$. The internal energy minimum, (6.3), is due to the fact that internal energy is a concave function of of the conservative variables, therefore,

$$(\rho\mathsf{e})(\mathbf{U}_i^{\mathsf{L},n+1}) = (\rho\mathsf{e})\Big(\sum_{j\in\mathcal{J}(i)}\alpha_j\overline{\mathbf{U}}_{ij}^n\Big) \ge \sum_{j\in\mathcal{J}(i)}\alpha_j(\rho\mathsf{e})(\overline{\mathbf{U}}_{ij}^n) \ge \min_{j\in\mathcal{J}(i)}(\rho\mathsf{e})(\overline{\mathbf{U}}_{ij}^n), \tag{6.4}$$

where α_j are the coefficients of the bar states shown in (4.2). □

Remark 6.1.1 (Local Upper Bound on the Density)**.** Note from Theorem 6.1.1, if the oracle has a maximum compression constant, b^{-1}; for example, the van der Waals or covolume EOS, then we also have that $0 < 1 - b\rho_i^{\max,n} \le 1 - b\rho_i^{\mathsf{L},n+1} \le 1 - b\rho_i^{\min,n}$. This follows since $\mathcal{B}(b)$ is an invariant domain. □

6.1.1 Relaxation on the Density Bounds

When performing limiting on the density as described in Section 6.2 or Section 6.3.4 the high order solution can actually be reduced to first order. The methodology for correcting this issue to loosen the local bounds (6.1). That is, the solution after the limiting process can violate these bounds. This relaxation of the bounds is done so that the error is second order and still preserves $\rho > 0$ and $1 - b\rho > 0$.

First, we have the following lemma in regards to the maximum value of ρ in the solution

to the Riemann problem, (4.8).

Lemma 6.1.1 (Maximum Density Bound)**.** *The following is true.*

1. *The density in the solution the extended Riemann problem* (4.8) *satisfies the following*

$$\rho \leq \max_{Z\in\{i,j\}} \frac{1}{\tau_Z^\infty} = \max_{Z\in\{i,j\}} \frac{(\gamma_Z+1)\rho_Z}{(\gamma_Z-1)+2b\rho_Z} \tag{6.5}$$

2. *Under the* CFL *condition stated in Theorem 4.1.1, the low order update satisfies*

$$\rho_i^{\mathsf{L},n+1} \leq \frac{(\gamma_i^{\min,n}+1)\rho_i^{\max,n}}{(\gamma_i^{\min,n}-1)+2b\rho_i^{\max,n}} \tag{6.6}$$

Proof. 1. Assume the Z wave is an expansion, then the density decreases across the expansion, hence $\rho \leq \rho_{\mathrm{Z}}$. If the Z wave is a shock, then from Lemma 6.4.1, we know that $\tau \in (\tau_{\mathrm{Z}}^\infty, \tau_{\mathrm{Z}}]$. Hence $\rho \in (\frac{1}{\tau_{\mathrm{Z}}}, \frac{1}{\tau_{\mathrm{Z}}^\infty}]$. This completes the proof.

2. From (6.1) we have that

$$\rho_i^{\mathsf{L},n+1} \leq \max_{j\in\mathcal{G}(i)} \overline{\rho}_{ij}^n \leq \max_{j\in\mathcal{G}(i)} \frac{(\gamma_j+1)\rho_j}{(\gamma_j-1)+2b\rho_j} \tag{6.7}$$

where we have also applied (6.5) to $\overline{\rho}_{ij}^n$ as the average of the solution to the Riemann problem must also satisfy that inequality. Since $\frac{(\gamma+1)\rho}{(\gamma-1)+2b\rho}$ is an increasing function of ρ and a decreasing function of γ, we conclude that,

$$\rho_i^{\mathsf{L},n+1} \leq \frac{(\gamma_i^{\min,n}+1)\rho_i^{\max,n}}{(\gamma_i^{\min,n}-1)+2b\rho_i^{\max,n}}. \tag{6.8}$$

□

Next we introduce an approximation of the local curvature of the density which will also

be used in the relaxation. For each $i \in \mathcal{V}$, define the following,

$$\Delta^2 \rho_i^n := \frac{\sum_{j \in \mathcal{J}(i)\setminus\{i\}} \beta_{ij}(\rho_i^n - \rho_j^n)}{\sum_{j \in \mathcal{J}(i)\setminus\{i\}} \beta_{ij}} \tag{6.9a}$$

$$\overline{\Delta^2 \rho_i^n} := \frac{1}{2\mathrm{card}(\mathcal{J}(i))} \sum_{j \in \mathcal{J}(i)\setminus\{i\}} \Big(\frac{1}{2}\Delta^2 \rho_i^n + \frac{1}{2}\Delta^2 \rho_j^n \Big), \tag{6.9b}$$

where $\beta_{ij} := \int_D \nabla\varphi_i \cdot \nabla\varphi_j \, \mathrm{d}\boldsymbol{x}$ are the stiffness coefficients of the Laplace operator. Recall that φ_i are the global shape functions; see Chapter 3. These definitions are well defined since from the partition of unity property, we have that $\sum_{j \in \mathcal{J}(i)\setminus\{i\}} \beta_{ij} = -\beta_{ii} = -\int_D \|\nabla\varphi_i\|_{\ell^2}^2 \, \mathrm{d}\boldsymbol{x} \neq 0$. Notice that $\overline{\Delta^2 \rho_i^n}$ is an estimate of the local curvature in a neighborhood of node i.

The relaxation on the density is now defined as,

$$\widetilde{\rho_i^{\min,n}} := \max((1 - r_h)\rho_i^{\min,n}, \rho_i^{\min,n} - \overline{\Delta^2 \rho_i^n}), \tag{6.10a}$$

$$\widetilde{\rho_i^{\max,n}} := \min\Big((1 + r_h)\rho_i^{\max,n}, \rho_i^{\max,n} + \overline{\Delta^2 \rho_i^n}, \tfrac{(\gamma+1)\rho_i^{\max,n}}{\gamma - 1 + 2b\rho_i^{\max,n}} \Big), \tag{6.10b}$$

where $r_h := (m_i/|D|)^{1.5/d}$.

6.2 The Flux Corrected Transport Method

To demonstrate the FCT method, we apply it to the update on the density, ρ. Let the low order and high order update be given as $\rho_i^{\mathsf{L},n+1}$ and $\rho_i^{\mathsf{H},n+1}$, respectively. These two quantities are computed using first and second order methods described in Chapters 4 and 5, respectively. Note that the high order update is related to the low order update by,

$$\rho_i^{\mathsf{H},n+1} = \rho_i^{\mathsf{L},n+1} + \sum_{j \in \mathcal{J}(i)} \mathsf{P}_{ij}^{\rho}. \tag{6.11}$$

From Section 6.1 we know that the first order update satisfies, $\rho_i^{\min} \leq \rho_i^{\mathsf{L},n+1} \leq \rho_i^{\max}$ and the second order update does not necessarily satisfy these local bounds. The FCT method begins by splitting $\sum_{j \in \mathcal{J}(i)} \mathsf{P}_{ij}^{\rho}$ into the negative and positive parts. That is, let $\mathcal{J}(i)^+ := \{j \in \mathcal{J}(i) : \mathsf{P}_{ij}^{\rho} > 0\}$ and $\mathcal{J}(i)^- := \{j \in \mathcal{J}(i) : \mathsf{P}_{ij}^{\rho} < 0\}$. Then the goal is to find $\ell_i^+, \ell_i^- \in [0, 1]$

so that new update defined by,

$$\rho_i^{n+1} = \rho_i^{\mathsf{L},n+1} + \ell_i^+ \sum_{j\in\mathcal{J}(i)^+} \mathsf{P}_{ij}^\rho + \ell_i^- \sum_{j\in\mathcal{J}(i)^-} \mathsf{P}_{ij}^\rho, \tag{6.12}$$

satisfies, $\rho_i^{\min} \leq \rho_i^{n+1} \leq \rho_i^{\max}$.

Note that for any $\ell_i^+, \ell_i^- \in [0,1]$ the new update satisfies,

$$\rho_i^{\mathsf{L},n+1} + \ell_i^- \sum_{j\in\mathcal{J}(i)^-} \mathsf{P}_{ij}^\rho \leq \rho_i^{n+1} \leq \rho_i^{\mathsf{L},n+1} + \ell_i^+ \sum_{j\in\mathcal{J}(i)^+} \mathsf{P}_{ij}^\rho. \tag{6.13}$$

Thus under the choice of,

$$\ell_i^- := \min\Big(\frac{\rho_i^{\min} - \rho_i^{\mathsf{L},n+1}}{\sum_{j\in\mathcal{J}(i)^-} P_{ij}^\rho}, 1\Big) \quad \text{and} \quad \ell_i^+ := \min\Big(\frac{\rho_i^{\max} - \rho_i^{\mathsf{L},n+1}}{\sum_{j\in\mathcal{J}(i)^+} P_{ij}^\rho}, 1\Big), \tag{6.14}$$

we have that $\rho_i^{\min,n} \leq \rho_i^{n+1} \leq \rho_i^{\max,n}$. For more recent work in the FCT literature, in the context of the compressible Euler equations, see [73].

Remark 6.2.1 (Limitations of the FCT Method)**.** The FCT method is inherently linear. This is perfectly fine for preserving positivity of the density; however, if one attempts to apply it to the specific internal energy, $\mathsf{e}(\mathbf{U}_i)$, then difficulties immediately arise as e is a nonlinear functional. □

6.3 Quasiconcave Limiting

In this section we discuss a new method introduced in [60] which allows us to perform limiting on quasiconcave functionals. This methodology allows us to address the issue of maintaining positivity of the internal energy.

6.3.1 Quasiconcave Functionals

Definition 6.3.1 (Quasiconcavity)**.** For $\mathcal{C}$ a convex set, $\Phi : \mathcal{C} \to \mathbb{R}$ is said to be quasiconcave if $L_\lambda(\Psi) := \{\mathbf{U} \in \mathcal{C} \,:\, \Psi(\mathbf{U}) \geq \lambda\}$ is a convex set for any $\lambda \in \mathbb{R}$. That is, the upper level sets are convex. □

As might be inferred from the name "quasiconcave", a functional which is quasiconcave may not be concave. For example, if $\mathcal{C} = [-1, 1]$ then $f(x) = x^3$ is certainly not concave, but it is quasiconcave since for any $\lambda \in \text{Range}(f)$, we have, $L_\lambda(f) = [\sqrt[3]{\lambda}, 1]$ which is a convex set. However, every concave functional is also quasiconcave.

Lemma 6.3.1 (Concave Implies Quasiconcave)**.** *Let $\mathcal{C} \subset \mathbb{R}^m$ be convex and $\Psi : \mathcal{C} \to \mathbb{R}$ be a concave functional, then Ψ is quasiconcave.*

Proof. Assume that λ is in the range of Ψ; that is, $L_\lambda(\Psi) \neq \emptyset$, otherwise the result is vacuous. Let $\mathbf{U}_1, \mathbf{U}_2 \in L_\lambda(\Psi)$ and $0 \leq t \leq 1$. Then since Ψ is concave, we have,

$$\Psi(t\mathbf{U}_1 + (1-t)\mathbf{U}_2) \geq t\Psi(\mathbf{U}_1) + (1-t)\Psi(\mathbf{U}_2) \geq \lambda. \tag{6.15}$$

Therefore $t\mathbf{U}_1 + (1-t)\mathbf{U}_2 \in L_\lambda(\Psi)$ for all $t \in [0, 1]$ hence $L_\lambda(\Psi)$ is convex and so Ψ is quasiconcave. □

6.3.2 The Abstract Scheme

Similar to the FCT method, we estimate the difference in the low and high order methods, $\mathbf{U}_i^{\mathsf{H},n+1} - \mathbf{U}_i^{\mathsf{L},n+1}$. Specifically, we look at the difference, $\sum_{j \in \mathcal{J}(i)} m_{ij}(\mathbf{U}_j^{\mathsf{H},n+1} - \mathbf{U}_j^n) - (\mathbf{U}_i^{\mathsf{L},n+1} - \mathbf{U}_i^n)$. Substituting in the definition of the first and second order updates, (4.1) and (5.5), respectively, we have

$$\sum_{j \in \mathcal{J}(i)} m_{ij}(\mathbf{U}_j^{\mathsf{H},n+1} - \mathbf{U}_j^n) - m_i(\mathbf{U}_i^{\mathsf{L},n+1} - \mathbf{U}_i^n) = \Delta t_n \sum_{j \in \mathcal{J}(i)} (d_{ij}^{\mathsf{H},n} - d_{ij}^{\mathsf{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n). \tag{6.16}$$

Now, add and subtract $m_i(\mathbf{U}_i^{\mathsf{H},n+1} - \mathbf{U}_i^n)$ and refactor as,

$$\begin{aligned} &m_i(\mathbf{U}_i^{\mathsf{H},n+1} - \mathbf{U}_i^{\mathsf{L},n+1}) - m_i(\mathbf{U}_i^{\mathsf{H},n+1} - \mathbf{U}_i^n) + \sum_{j \in \mathcal{J}(i)} m_{ij}(\mathbf{U}_j^{\mathsf{H},n+1} - \mathbf{U}_j^n) \\ &\quad = \Delta t_n \sum_{j \in \mathcal{J}(i)} (d_{ij}^{\mathsf{H},n} - d_{ij}^{\mathsf{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n). \end{aligned} \tag{6.17}$$

This can all be combined into a nice expression by defining $\Delta_{ij} := m_i\delta_{ij} - m_{ij}$ where $\delta_{ij} = 1$ if $i = j$ and 0 if $i \neq j$. Thus we have,

$$m_i(\mathbf{U}_i^{\mathsf{H},n+1} - \mathbf{U}_i^{\mathsf{L},n+1}) = \sum_{j\in\mathcal{G}(i)} \Delta_{ij}(\mathbf{U}_j^{\mathsf{H},n+1} - \mathbf{U}_j^n) + \Delta t_n(d_{ij}^{\mathsf{H},n} - d_{ij}^{\mathsf{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n). \tag{6.18}$$

It is important to note that the right hand side of this equation can be expressed as a *skew-symmetric* matrix. Using that $\sum_{j\in\mathcal{G}(i)} \Delta_{ij} = 0$, we can write the following,

$$m_i(\mathbf{U}_i^{\mathsf{H},n+1} - \mathbf{U}_i^{\mathsf{L},n+1}) = \sum_{j\in\mathcal{G}(i)} \mathbf{A}_{ij}^n \tag{6.19a}$$

$$\mathbf{A}_{ij}^n := \Delta_{ij}(\mathbf{U}_j^{\mathsf{H},n+1} - \mathbf{U}_j^n - (\mathbf{U}_i^{\mathsf{H},n+1} - \mathbf{U}_i^n)) + \Delta t_n(d_{ij}^{\mathsf{H},n} - d_{ij}^{\mathsf{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n), \tag{6.19b}$$

from which it is readily seen to be skew-symmetric ($\mathbf{A}_{ij}^n = -\mathbf{A}_{ji}^n$).

Lemma 6.3.2 (Consistent Conservation)**.** *The total mass of the high order scheme is the same as the low order scheme; that is,*

$$\sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{\mathsf{H},n+1} = \sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^{\mathsf{L},n+1}. \tag{6.20}$$

Proof. The result is proven if we can show that

$$\sum_{i\in\mathcal{V}} m_i(\mathbf{U}_i^{\mathsf{H},n+1} - \mathbf{U}_i^{\mathsf{L},n+1}) = \sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{G}(i)} \mathbf{A}_{ij}^n = 0. \tag{6.21}$$

Using that $\mathbf{A}_{ij}^n$ is skew-symmetric, we have, $\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{G}(i)} \mathbf{A}_{ij}^n = \sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{G}(i)} -\mathbf{A}_{ji}^n$. Using Fubini's theorem, we find that, $\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{G}(i)} \mathbf{A}_{ij}^n + \sum_{j\in\mathcal{V}}\sum_{i\in\mathcal{G}(j)} \mathbf{A}_{ji}^n = 0$. But the indices are arbitrary hence, $2\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{G}(i)} \mathbf{A}_{ij}^n = 0$ and the result is proven. □

6.3.3 The Limiter

The idea now, is to introduce a symmetric matrix (ℓ_{ij}) which provides an "intermediate" solution between $\mathbf{U}_i^{\mathsf{H},n+1}$ and $\mathbf{U}_i^{\mathsf{L},n+1}$. In particular we define,

$$\mathbf{U}_i^{n+1} := \mathbf{U}_i^{\mathsf{L},n+1} + \frac{1}{m_i} \sum_{j \in \mathcal{J}(i)} \ell_{ij} \mathbf{A}_{ij}^n. \tag{6.22}$$

where $0 \leq \ell_{ij} \leq 1$ for all $i, j \in \mathcal{V}$ and (ℓ_{ij}) is symmetric. The reason for enforcing symmetry on (ℓ_{ij}), is that $\ell_{ij}\mathbf{A}_{ij}^n$ is still skew-symmetric, hence the total mass of $\mathbf{U}_i^{n+1}$ is the same as $\mathbf{U}_i^{\mathsf{L},n+1}$. Note, if $\ell_{ij} = 0$ for all $j \in \mathcal{J}(i)$, then $\mathbf{U}_i^{n+1} = \mathbf{U}_i^{\mathsf{L},n+1}$ and similarly, if $\ell_{ij} = 1$ for all $j \in \mathcal{J}(i)$, then $\mathbf{U}_i^{n+1} = \mathbf{U}_i^{\mathsf{H},n+1}$. All of the discussion so far on the limiter is still a part of the FCT method as described in, [71].

We now move towards the quasiconcave limiting theory as described in [60, Sec. 4.2]. Introduce $\{\lambda_j^i\}_{j \in \mathcal{J}(i) \setminus \{i\}}$ such that $\sum_{j \in \mathcal{J}(i) \setminus \{i\}} \lambda_j^i = 1$ and $\lambda_j^i > 0$ for all $j \in \mathcal{J}(i)$. The superscript ^{i} is just used to indicate the dependence on i. Then, (6.22) can be written as,

$$\mathbf{U}_i^{n+1} = \sum_{j \in \mathcal{J}(i) \setminus \{i\}} \lambda_j^i \Big(\mathbf{U}_i^n + \ell_{ij} \mathbf{P}_{ij}^n \Big), \quad \text{where } \mathbf{P}_{ij}^n := \frac{1}{m_i \lambda_j^i} \mathbf{A}_{ij}^n. \tag{6.23}$$

Note that $j = i$ is omitted since $\mathbf{A}_{ii}^n = \mathbf{0}$ for all $i \in \mathcal{V}$ since $\mathbf{A}_{ij}^n$ is skew-symmetric. For the remainder of this thesis, we set $\lambda_j^i = \frac{1}{\operatorname{card}(\mathcal{J}(i))-1}$.

Recall that we would like to preserve some physical quantity represented by a quasiconcave functional, Ψ. Therefore, the goal is to determine ℓ_{ij} so that $\Psi(\mathbf{U}_i^{n+1}) \geq 0$. The following two lemmas set the stage for the upcoming numerical algorithms.

Lemma 6.3.3. *Let $\Psi : \mathcal{C} \to \mathbb{R}$ be a quasiconcave functional. Assume $\ell_{ij} \in [0,1]$ are such that $\Psi(\mathbf{U}_i^{\mathsf{L},n} + \ell_{ij}\mathbf{P}_{ij}) \geq 0$ for all $j \in \mathcal{J}(i) \setminus \{i\}$, then*

$$\Psi\Big(\sum_{j \in \mathcal{J}(i) \setminus \{i\}} \lambda_j^i (\mathbf{U}_i^{\mathsf{L},n} + \ell_{ij} \mathbf{P}_{ij}) \Big) \geq 0. \tag{6.24}$$

Proof. The proof of this follows from the definition of quasiconcavity. Note that $\mathbf{U}_i^{\mathsf{L},n+1} + \ell_{ij}\mathbf{P}_{ij} \in L_0(\Psi)$ where $L_0(\Psi) = \{\mathbf{U} \in \mathcal{C} : \Psi(\mathbf{U}) \geq 0\}$. Since $L_0(\Psi)$ is a convex set, then the convex combination defined in (6.23) must also belong to $L_0(\Psi)$, hence the result. □

Lemma 6.3.4 (Symmetrization of ℓ_{ij})**.** *Define,*

$$\ell_j^i := \begin{cases} 1 & \text{if } \Psi(\mathbf{U}_i^{\mathsf{L},n+1} + \mathbf{P}_{ij}) \geq 0 \\ \max\{\ell \in [0,1] : \Psi(\mathbf{U}_i^{\mathsf{L},n+1} + \ell\mathbf{P}_{ij}) \geq 0\} & \text{otherwise} \end{cases} \tag{6.25}$$

for all $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$. Then the following holds,

1. $\Psi(\mathbf{U}_i^{\mathsf{L},n+1} + \ell\mathbf{P}_{ij}) \geq 0$ *for all* $\ell \in [0, \ell_j^i]$.
2. *Let* $\ell_{ij} := \min\{\ell_j^i, \ell_i^j\}$, *then* $\Psi(\mathbf{U}_i^{\mathsf{L},n+1} + \ell_{ij}\mathbf{P}_{ij}) \geq 0$ *for all* $i \in \mathcal{V}$ *and all* $j \in \mathcal{I}(i)$.

Remark 6.3.1 (Comments on the Limiter ℓ_{ij})**.** From Lemma 6.3.3 we see that the determination of ℓ_{ij} only needs to be found for each $j \in \mathcal{I}(i)$ independently. From Lemma 6.3.4 (i) we see that smaller values of the limiter do not violate the non-negativity of the quasiconcave functional. This is intuitively expected, since we are "pulling" our solution closer to the invariant-domain preserving update, $\mathbf{U}_i^{\mathsf{L},n+1}$. Lastly, from Lemma 6.3.4 (ii) we see that (ℓ_{ij}) is symmetric. □

6.3.4 Limiting on the Density

Limiting on the density is performed exactly as done in [60, Sec 4.4]. We define, $\Psi_+^\rho(\mathbf{U}) := \rho - \rho_i^{\min,n}$ and $\Psi_-^\rho(\mathbf{U}) := \rho_i^{\max,n} - \rho$. The limiter is then,

$$\ell_j^{i,\rho} := \begin{cases} \min\left(\frac{|\rho_i^{\min,n} - \rho_i^{\mathsf{L},n+1}|}{|\mathsf{P}_{ij}^\rho| + \epsilon_i}, 1\right) & \text{if } \rho_i^{\mathsf{L},n+1} + \mathsf{P}_{ij}^\rho < \rho_i^{\min,n}, \\ 1 & \rho_i^{\min,n} \leq \text{ if } \rho_i^{\mathsf{L},n+1} + \mathsf{P}_{ij}^\rho \leq \rho_i^{\max,n}, \\ \min\left(\frac{|\rho_i^{\max,n} - \rho_i^{\mathsf{L},n+1}|}{|\mathsf{P}_{ij}^\rho| + \epsilon_i}, 1\right) & \text{if } \rho_i^{\max,n} < \rho_i^{\mathsf{L},n+1} + \mathsf{P}_{ij}^\rho, \end{cases} \tag{6.26}$$

where $\epsilon_i := \epsilon\rho_i^{\max,n}$ with $\epsilon := 10^{-14}$. It bears repeating that, limiting on the density also implies preservation of the maximum compressibility. That is, the high order solution satisfies $1 - b\rho_i^{n+1} > 0$.

6.3.5 Limiting on the Internal Energy

In order to secure that our high order method is invariant domain preserving, we would like to enforce limiting on the internal energy. However, this is not necessary to do as limiting on the surrogate entropy will enforce this condition, see Section 6.4. Still, we briefly recount the method for limiting the internal energy. That is, we require $\mathsf{e}(\mathbf{U}_i^{n+1}) > 0$. The details provided in this section are not new and are outlined in [60, Section 4.5]. Define $\Psi_i^e(\mathbf{U}) := (\rho\mathsf{e})(\mathbf{U}) - \Upsilon_i^{\min} = \mathsf{E} - \frac{\|\mathbf{M}\|_{\ell^2}^2}{2\rho} - \Upsilon_i^{\min}$. The goal is to find the largest $t_0 \in [0, \ell_j^{i,\rho}]$ such that $\Psi^e(\mathbf{U}_i^{\mathsf{L},n+1} + t_0\mathbf{P}_{ij}) \geq 0$ for all $j \in \mathcal{J}(i)$. This can equivalently be solved using with the functional, $\psi_i^e(\mathbf{U}) : \{\mathbf{U} \in \mathbb{R}^{d+2} \mid \rho > 0\}$ where,

$$\psi_i^e(\mathbf{U}) := \rho\Psi_i^e(\mathbf{U}) = \rho\mathsf{E} - \tfrac{1}{2}\|\mathbf{M}\|_{\ell^2} - \rho\Upsilon_i^{\min}. \tag{6.27}$$

The advantage of working with this functional is that it is quadratic with respect to the conserved variables.

The derivatives of this functional are,

$$D\psi_i^e(\mathbf{U}) = \begin{pmatrix} \mathsf{E} - \Upsilon_i^{\min} \\ -\mathbf{M} \\ \rho \end{pmatrix} \qquad \mathbb{H}\psi_i^e(\mathbf{U}) = \begin{pmatrix} 0 & \mathbf{0}^\mathsf{T} & 1 \\ 0 & -\mathbb{I}_d & 0 \\ 1 & \mathbf{0}^\mathsf{T} & 0 \end{pmatrix}. \tag{6.28}$$

Note that $\psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1} + t\mathbf{P}_{ij})$ is a quadratic function of t; in particular, set $\psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1} + t\mathbf{P}_{ij}) = at^2 + bt + c$ where $a = \frac{1}{2}\mathbf{P}_{ij}^\mathsf{T}\mathbb{H}\psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1})\mathbf{P}_{ij}$, $b = D\psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1}) \cdot \mathbf{P}_{ij}$, and $c = \psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1})$. Let t_0 be the smallest positive root of $\psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1} + t\mathbf{P}_{ij}) = 0$ and set $t_0 = 1$ if no such roots exist. We define,

$$\ell_j^{i,e} = \min(t_0, \ell_j^{i,\rho}). \tag{6.29}$$

Lemma 6.3.5. *If $\ell \in [0, \ell_j^{i,e}]$, then $\Psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij}) \geq 0$.*

Proof. Since $\psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij})$ is a quadratic function with $\psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1}) > 0$. By definition of t_0, $\psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, \ell_j^{i,e}]$, otherwise, if there existed $0 < \widetilde{\ell} < \ell_j^{i,e}$, such that $\psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1} + \widetilde{\ell} \mathbf{P}_{ij}$, then t_0 would not be the smallest root. Lastly, $\rho_i^{\mathsf{L},n+1} + \ell P_{ij}^{\rho} > 0$ for all $\ell \in [0, \ell_j^{i,e}] \subset [0, \ell_j^{i,\rho}]$. Hence, we must have that $\Psi_i^e(\mathbf{U}_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, \ell_j^{i,e}]$. $\square$

6.4 The Entropy Surrogate

It is well known that the Euler equations with the ideal EOS preserve the minimum principle on the physical entropy, $s(\boldsymbol{x}, t)$, see [74].

Theorem 6.4.1 (Minimum Principle on the Specific Entropy)**.** *Let $\boldsymbol{u}$ be an entropy solution to the Euler equations supplied with the ideal EOS, then the specific entropy satisfies*

$$\operatorname*{ess\,inf}_{\|\boldsymbol{x}\|_{\ell^2} \leq R} s(\boldsymbol{x}, t) \geq \operatorname*{ess\,inf}_{\|\boldsymbol{x}\|_{\ell^2} \leq R + t\boldsymbol{v}_{\max}} s(\boldsymbol{x}, 0) \tag{6.30}$$

where $R > 0$, and $\boldsymbol{v}_{\max} := \max_{(\boldsymbol{x},t) \in D \times [0,t]} \|\boldsymbol{v}(\boldsymbol{x}, t)\|_{\ell^2}$

However, as mentioned in Chapter 2, we may not have access to the physical entropy for an arbitrary or tabulated EOS. The novel idea of this section which was originally proposed in Clayton et. al. [1], is the introduction of a local surrogate entropy functional, which maintains some similar properties to a physical entropy. In particular, the surrogate entropy functional increases across shocks, see Theorem 6.4.1.

In this section we outline some known facts regarding the physical entropy and approximate these principles with a surrogate physical entropy for our tabulated EOS. This is all done in the context of the Riemann problem as that is where our application lies.

6.4.1 The Entropy for the NASG EOS

Recall, the NASG EOS is given by $p(\rho, e) = (\gamma - 1)\frac{\rho(e-q)}{1-b\rho} - \gamma p_\infty$. The specific entropy is, $s(\tau, e) = \log\Big(\big(e - q - p_\infty(\tau - b)\big)^{\frac{1}{\gamma-1}}(\tau - b)\Big)$. Sometimes, we may abuse the notation and write $s = s(\boldsymbol{u})$. We will also need to work with the variables, ρ and e. So we may express

the entropy as $s(\rho^{-1}, e)$. Note that one can express the physical entropy in an equivalent form, that is,

$$\exp((\gamma-1)s(\tau,e)) = (e - q - p_\infty(\tau - b))(\tau - b)^{\gamma-1}. \tag{6.31}$$

We now define $S(\boldsymbol{u};\gamma) = \exp((\gamma-1)s(\rho^{-1},e))$; that is,

$$S(\boldsymbol{u};\gamma) := \big(\rho(e-q) - p_\infty(1-b\rho)\big)\frac{(1-b\rho)^{\gamma-1}}{\rho^\gamma}. \tag{6.32}$$

In order to preserve some notion of a minimum principle on the specific entropy, we will introduce a local functional Ψ_i^s which behaves similar to a physical entropy.

Theorem 6.4.2 (Equivalence of the Minimum Principle on the Specific Entropy)**.** *If the oracle for the Riemann problem 1.8a–1.8b is given by the NASG EOS,* $p(\boldsymbol{u}) := (\gamma-1)\frac{\rho(\mathsf{e}(\boldsymbol{u})-q)}{1-b\rho} - \gamma p_\infty$*. Then, the minimum principle on the specific entropy in the Riemann problem is equivalent to*

$$\Psi^s(\boldsymbol{u}(x,t)) \geq \min(\Psi^s(\boldsymbol{u}_{\mathrm{L}}), \Psi^s(\boldsymbol{u}_{\mathrm{R}})) \tag{6.33}$$

for all $(x,t) \in \mathbb{R} \times [0,\infty)$.

Proof. First assume that the Riemann solution contains no vacuum states. From the definition of the specific entropy, we have the expression,

$$S(\boldsymbol{u}) = \exp((\gamma-1)s(\boldsymbol{u})) = \big(\rho(\mathsf{e}(\boldsymbol{u}) - q) - p_\infty(1-b\rho)\big)\frac{(1-b\rho)^{\gamma-1}}{\rho^\gamma} \tag{6.34}$$

Define, $S^{\min} := \min\{\exp((\gamma-1)s(\boldsymbol{u}_{\mathrm{L}})), \exp((\gamma-1)s(\boldsymbol{u}_{\mathrm{R}}))\}$. Let $\boldsymbol{u}$ be on a Z-wave, $\mathrm{Z} \in \{\mathrm{L},\mathrm{R}\}$. If the Z-wave is an expansion, then the specific entropy is constant which implies that $S(\boldsymbol{u}_{\mathrm{Z}}) = \exp((\gamma-1)s(\boldsymbol{u}_{\mathrm{Z}})) = S(\boldsymbol{u}(x,t))$ for all (x,t) belonging to the expansion wave. If the Z-wave is a shock, then the entropy increases, hence, $\exp((\gamma-1)s(\boldsymbol{u}_{\mathrm{Z}})) \leq \exp((\gamma-1)s(\boldsymbol{u}(x,t))) = S(\boldsymbol{u}(x,t))$ for all (x,t) on the shock wave. We conclude that $S(\boldsymbol{u}(x,t)) \geq S^{\min}$ for any (x,t) on a Z-wave. By continuity, this inequality also holds across the contact, hence $\Psi^s(\boldsymbol{u}(x,t)) \geq 0$ for all $(x,t) \in \mathbb{R} \times [0,\infty)$.

In the case of vacuum ($\rho = 0$), we have that $\Psi^s(\boldsymbol{u}(x,t)) = 0$ for (x,t) in the vacuum region. Note also that

$$S^{\min} = \min_{\mathrm{Z}\in\{\mathrm{L},\mathrm{R}\}} \big(\rho_{\mathrm{Z}}(e_{\mathrm{Z}} - q) - p_\infty(1 - b\rho_{\mathrm{Z}})\big)\frac{(1-b\rho_{\mathrm{Z}})^{\gamma-1}}{\rho_{\mathrm{Z}}^{\gamma}}, \tag{6.35}$$

and therefore $\min(\Psi^s(\boldsymbol{u}_{\mathrm{L}}), \Psi^s(\boldsymbol{u}_{\mathrm{R}})) = 0$. Hence the result also holds for vacuum states. □

We now prove a result regarding the behavior of these kind of functionals across shock waves.

Lemma 6.4.1 (Behavior across Shocks)**.** *For all $i \in \mathcal{V}$, assume $\mathbf{U}_i^n \in \mathcal{B}(b)$. For all $i \in \mathcal{V}$ and $j \in \mathcal{I}(i)$ let $\gamma_{ij} \in [1, \min(\gamma_i^n, \gamma_j^n)]$. Given left and right data $(\rho_i^n, \boldsymbol{m}_i^n, \mathcal{E}_i^n, \Gamma_i^n)^\mathsf{T}$ and $(\rho_j^n, \boldsymbol{m}_j^n, \mathcal{E}_j^n, \Gamma_j^n)^\mathsf{T}$, respectively, the functional,*

$$\Psi_{ij}^s(\boldsymbol{u}) := \rho(e(\boldsymbol{u}) - q) - p_\infty(1 - b\rho) - S_{ij}^{\min}\rho^{\gamma_{ij}}(1-b\rho)^{1-\gamma_{ij}}, \tag{6.36}$$

with $S_{ij}^{\min} := \min(S(\mathbf{U}_i^n; \gamma_{ij}), S(\mathbf{U}_j^n; \gamma_{ij}))$, increases across shock in the solution to the extended Riemann problem (if a shock wave exists).

Proof. To simplify the notation, we omit the supscript ^{n}. The solution to the extended Riemann problem, $(\rho, m, \mathcal{E}, \Gamma)^\mathsf{T}(x,t)$ is described in Chapter 4. Recall that $\gamma(x,t) = \gamma_i$ for $x/t < v^*$ and $\gamma(x,t) = \gamma_j$ for $x/t > v^*$ where v^* is the speed of the contact. Let $\mathrm{Z} \in \{i, j\}$ and assume that the Z-wave is a shock wave. Let $\boldsymbol{u}_{\mathrm{Z}} \in \mathcal{B}(b)$ be the state before the shock and $\boldsymbol{u} \in \mathcal{B}(b)$ be an arbitrary state connected to $\boldsymbol{u}_{\mathrm{Z}}$ through a shock wave. Since we are concerned with the Z-wave, $\gamma = \gamma_{\mathrm{Z}}$ is constant, and therefore the interpolatory pressure $\boldsymbol{p}_{\mathsf{nasg}}$ is an EOS. We abuse the notation by writing, $\boldsymbol{p}_{\mathsf{nasg}}(\tau, e) = \boldsymbol{p}_{\mathsf{nasg}}(\tau^{-1}, e, \gamma_{\mathrm{Z}})$ to simplify the notation. From the Rankine-Hugoniot equations, we have the following relationship involving only the thermodynamic quantities, see [31, Chapter III, Lemma 2.2].

$$\mathsf{e}(\boldsymbol{u}) - e_{\mathrm{Z}} + \frac{1}{2}(\boldsymbol{p}_{\mathsf{nasg}}(\tau, \mathsf{e}(\boldsymbol{u})) + \boldsymbol{p}_{\mathsf{nasg}}(\tau_{\mathrm{Z}}, e_{\mathrm{Z}}))(\tau - \tau_{\mathrm{Z}}) = 0 \tag{6.37}$$

Substituting in the definition of p_{nasg}; that is, $p_{\mathsf{nasg}} = (\gamma_{\mathrm{Z}} - 1)\frac{\mathsf{e}(\boldsymbol{u})-q}{\tau-b} - \gamma p_\infty$, we find,

$$\mathsf{e}(\boldsymbol{u}) - q = (e_{\mathrm{Z}} - q)\frac{1 - \frac{(\gamma_{\mathrm{Z}}-1)(\tau-\tau_{\mathrm{Z}})}{2(\tau_{\mathrm{Z}}-b)}}{1 + \frac{(\gamma_{\mathrm{Z}}-1)(\tau-\tau_{\mathrm{Z}})}{2(\tau-b)}} =: r(\tau). \tag{6.38}$$

As we can see, $\mathsf{e}(\boldsymbol{u}) - q$ is only a function of τ along the shock curve. For $\mathsf{e}(\boldsymbol{u})$ to be well defined, we must have that $\tau \in (\tau_{\mathrm{Z}}^\infty, \infty)$ where

$$\tau_{\mathrm{Z}}^\infty := \frac{(\gamma_{\mathrm{Z}} - 1)\tau_{\mathrm{Z}} + 2b}{\gamma_{\mathrm{Z}} + 1}. \tag{6.39}$$

Furthermore, $\mathsf{e}(\boldsymbol{u}) > 0$, for $\tau \in (\tau_{\mathrm{Z}}^\infty, \tau_{\mathrm{Z}}^0)$ where $\tau_{\mathrm{Z}}^0 := \frac{(\gamma_{\mathrm{Z}}+1)\tau_{\mathrm{Z}}-2b}{\gamma_{\mathrm{Z}}-1}$. Since $\mathbf{U}_{\mathrm{Z}} \in \mathcal{B}(b)$ we have that $b < \tau_{\mathrm{Z}}^\infty < \tau < \tau_{\mathrm{Z}}^0$.

The goal now is to show that

$$\mathcal{B}(b) \ni \boldsymbol{u} \mapsto \rho(e(\boldsymbol{u}) - q) - p_\infty(1 - b\rho) - c\rho^\gamma(1 - b\rho)^{1-\gamma} \tag{6.40}$$

is an increasing function across the shock wave for all $\gamma \in (1, \gamma_{\mathrm{Z}}]$ and $c \in (0, ((e_{\mathrm{Z}} - q) - p_\infty(\tau_{\mathrm{Z}} - b))(\tau_Z - b)^{\gamma-1}]$. This will complete our proof since, setting $\gamma = \gamma_{ij} \leq \gamma_{\mathrm{Z}}$ and $c := S_{ij}^{\min}$ we see that

$$\begin{aligned} 0 \leq S_{ij}^{\min} &\leq \big(\rho_{\mathrm{Z}}(e_{\mathrm{Z}} - q) - p_\infty(1 - b\rho_{\mathrm{Z}})\big)\frac{(1 - b\rho_{\mathrm{Z}})^{\gamma_{ij}-1}}{\rho_{\mathrm{Z}}^{\gamma_{ij}}} \\ &= ((e_{\mathrm{Z}} - q) - p_\infty(\tau_{\mathrm{Z}} - b))(\tau_Z - b)^{\gamma-1}. \end{aligned} \tag{6.41}$$

Now define, $q(\tau) := (\gamma_{\mathrm{Z}} - 1)\frac{r(\tau)}{\tau-b} - \gamma_{\mathrm{Z}} p_\infty$ which is the pressure along the shock curve. Then $q'(\tau) = \frac{-4(e_{\mathrm{Z}}-q)\gamma_{\mathrm{Z}}(\gamma_{\mathrm{Z}}-1)}{(\tau+\tau_{\mathrm{Z}}-2b+\gamma_{\mathrm{Z}}(\tau-\tau_{\mathrm{Z}}))^2} < 0$. So q is a decreasing function of τ; i.e., the pressure is a strictly monotonic function of ρ along the shock curve. That is, for $\rho \in [\rho_{\mathrm{Z}}, b^{-1})$. In particular, the pressure is finite only in the range, $\tau \in (\tau_{\mathrm{Z}}^\infty, \tau_{\mathrm{Z}}]$.

Next note from equations (6.38) and (6.40), it is equivalent to prove that the following

function is a nonnegative decreasing function on shock curves,

$$(\tau_{\mathrm{Z}}^{\infty}, \tau_{\mathrm{Z}}] \ni \tau \mapsto g(\tau) := \tau^{-1} r(\tau) - p_{\infty}(1 - b\tau^{-1}) - c\tau^{-1}(\tau - b)^{1-\gamma}. \tag{6.42}$$

We will use the fact that the physical entropy, $s(\tau, e) = \log\Big(\big(e - q - p_{\infty}(\tau - b)\big)^{\frac{1}{\gamma_{\mathrm{Z}}-1}}(\tau - b)\Big)$, increases across the shock curves. That is, $s(\tau, r(\tau))$ is a decreasing function for $\tau \in (\tau_{\mathrm{Z}}^{\infty}, \tau_{\mathrm{Z}}]$. Furthermore, this also implies that $(\tau_{\mathrm{Z}}^{\infty}, \mathrm{Z}] \ni \tau \mapsto \varsigma(\tau) := \exp((\gamma_{\mathrm{Z}} - 1)s(\tau, r(\tau))) = (r(\tau) - p_{\infty}(\tau - b))(\tau - b)^{\gamma_{\mathrm{Z}}-1}$ is a decreasing function. We will use this fact to prove that $g(\tau)$ is a decreasing function. So notice that

$$g(\tau) = \varsigma(\tau)\tau^{-1}(\tau - b)^{1-\gamma_{\mathrm{Z}}} - c\tau^{-1}(\tau - b)^{1-\gamma}. \tag{6.43}$$

We further simplify the calculus by defining $\widetilde{\varsigma}(\tau) = \varsigma(\tau)(\tau - b)^{\gamma-\gamma_{\mathrm{Z}}}$. Then,

$$g(\tau) = \frac{(\tau - b)^{1-\gamma}}{\tau}(\widetilde{\varsigma}(\tau) - c). \tag{6.44}$$

Computing the derivative we find,

$$g'(\tau) = \frac{(\tau - b)^{1-\gamma}}{\tau}\widetilde{\varsigma}'(\tau) - \frac{(\gamma - 1)\tau + (\tau - b)}{\tau^2(\tau - b)^{\gamma}}(\widetilde{\varsigma}(\tau) - c), \tag{6.45}$$

where

$$\widetilde{\varsigma}'(\tau) = (\tau - b)^{\gamma-\gamma_{\mathrm{Z}}}\varsigma'(\tau) + (\gamma - \gamma_{\mathrm{Z}})(\tau - b)^{\gamma-\gamma_{\mathrm{Z}}-1}\varsigma(\tau). \tag{6.46}$$

Note that $\widetilde{\varsigma}'(\tau) \leq 0$, since $\varsigma'(t) \leq 0$ and $\varsigma(\tau) \geq 0$ for all $\tau \in (\tau_{\mathrm{Z}}^{\infty}, \tau_{\mathrm{Z}}]$ and $\gamma \leq \gamma_{\mathrm{Z}}$. Also $\inf_{\tau\in(b,\tau_{\mathrm{Z}}]} \widetilde{\varsigma}(\tau) = \widetilde{\varsigma}(\tau_{\mathrm{Z}}) = e_{\mathrm{Z}}(\tau_{\mathrm{Z}} - b)^{\gamma-1} \geq c$ since $\widetilde{\varsigma}$ is a decreasing function. Thus for $\gamma = \gamma_{ij}$ and $c = S_{ij}^{\min}$ we have that $g(\tau) \geq 0$ and $g'(\tau) \leq 0$. Therefore the mapping $[\rho_{\mathrm{Z}}, \frac{1}{\tau_{\mathrm{Z}}^{\infty}}) \ni \rho \mapsto \rho\mathsf{e}(\boldsymbol{u}) - p_{\infty}(1 - b\rho) - c\rho^{\gamma}(1 - b\rho)^{1-\gamma}$ is a nonnegative increasing function along the shock curves. □

We now present one of the main theorems of this chapter.

Theorem 6.4.3 (Surrogate Entropy)**.** *Let*

$$\gamma_i^{\min,n} := \min_{j\in\mathcal{J}(i)} \gamma_j^n, \qquad S_i^{\min,n} := \min\big(\min_{j\in\mathcal{J}(i)} S(\mathbf{U}_j^n; \gamma_i^{\min,n}), \min_{j\in\mathcal{J}(i)} S(\overline{\mathbf{U}}_{ij}^n; \gamma_i^{\min,n})\big), \tag{6.47a}$$

$$\Psi_i^s(\boldsymbol{u}) := \rho(\mathsf{e}(\boldsymbol{u}) - q) - p_\infty(1-b\rho) - S_i^{\min,n}\rho^{\gamma_i^{\min,n}}(1-b\rho)^{1-\gamma_i^{\min,n}}. \tag{6.47b}$$

Then the following hold for all $i \in \mathcal{V}$:

1. $\Psi_i^s : \mathcal{B}(b) \to \mathbb{R}$ *is a concave functional.*

2. *For $\mathbf{U}_i^{\mathsf{L},n+1}$ defined in* (4.1)*, we have that $\Psi_i^s(\mathbf{U}_i^{\mathsf{L},n+1}) \geq 0$ under the* CFL *constraint, $\tau \sum_{j\in\mathcal{J}(i)\setminus\{i\}} \frac{2d_{ij}^{\mathsf{L},n}}{m_i} \leq 1$.*

3. *Consider the extended Riemann problem 4.20 with left data, $(\rho_i^n, \boldsymbol{m}_i^n \cdot \boldsymbol{n}_{ij}, \mathcal{E}_i^n, \Gamma_i^n)$ and right data, $(\rho_j^n, \boldsymbol{m}_j^n \cdot \boldsymbol{n}_{ij}, \mathcal{E}_j^n, \Gamma_j^n)$. If the solution to the extended Riemann problem has a shock wave, then $\Psi_i^s(\boldsymbol{u})$ increases across the shocks.*

Proof. 1. The function, $\boldsymbol{u} \mapsto \rho(\mathsf{e}(\boldsymbol{u}) - q)$ is a concave functional and $p_\infty(1-b\rho)$ is a linear function of $\boldsymbol{u}$. Then notice that $f(x) := x^{\gamma_i^{\min,n}}(1-bx)^{1-\gamma_i^{\min,n}} = x(\frac{1}{x} - b)^{\gamma_i^{\min,n}}$. A quick exercise in calculus shows that $f(x)$ is a convex for $\gamma_i^{\min,n}$, hence $-S_i^{\min,n}\rho^{\gamma_i^{\min,n}}(1-b\rho)^{1-\gamma_i^{\min,n}}$ is a concave functional since $S_i^{\min,n} \geq 0$. Since the sum of concave functionals is concave, we have that $\Psi_i^s(\boldsymbol{u})$ is concave.

2. From Theorem 4.1.1, under the CFL condition, $\Delta t \leq -\frac{m_i}{2d_{ij}^{\mathsf{L},n}}$ we have that $\mathbf{U}_i^{\mathsf{L},n+1}$ is in the convex hull of the bar states $\{\overline{\mathbf{U}}_{ij}^n\}_{j\in\mathcal{J}(i)}$ and since Ψ_i^s is a concave functional we have that $\Psi_i^s(\mathbf{U}_i^{\mathsf{L},n+1}) \geq \min_{j\in\mathcal{J}(i)} \Psi_i^s(\overline{\mathbf{U}}_{ij}^n)$. To see this is nonnegative, consider,

$$\begin{aligned}
\Psi_i^s(\overline{\mathbf{U}}_{ij}^n) &= \rho(\mathsf{e}(\overline{\mathbf{U}}_{ij}^n) - q) - p_\infty(1-b\overline{\rho}_{ij}^n) - S_i^{\min,n}(\overline{\rho}_{ij}^n)^{\gamma_i^{\min,n}}(1-b\overline{\rho}_{ij}^n)^{1-\gamma_i^{\min,n}} \\
&= \big(S(\overline{\mathbf{U}}_{ij}^n; \gamma_i^{\min,n}) - S_i^{\min,n}\big)(\overline{\rho}_{ij}^n)^{\gamma_i^{\min,n}}(1-b\overline{\rho}_{ij}^n)^{1-\gamma_i^{\min,n}},
\end{aligned}$$

where $S(\mathbf{U};\gamma)$ is defined in (6.32). By definition of $S_i^{\min,n}$ we have that $\Psi_i^s(\overline{\mathbf{U}}_{ij}^n) \geq 0$, hence $\Psi_i^s(\mathbf{U}_i^{\mathsf{L},n+1}) \geq 0$.

3. This immediately follows from, Lemma 6.4.1. □

6.4.2 Limiting on the Surrogate Entropy

Let $\mathcal{A}(b) := \{\boldsymbol{u} \in \mathbb{R}^{d+2} : 0 < 1 - b\rho < 1\}$. From the limiting on the density described in Section 6.3.4, we have $\ell_j^{i,\rho} \in [0,1]$. This implies that the $\rho_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij} \in \mathcal{A}(b)$ for all $j \in \mathcal{I}(i) \setminus \{i\}$ and $\ell \in [0, \ell_j^{i,\rho}]$. Therefore, we can perform the quasiconcave limiting of Ψ_i^s on $\mathcal{A}(b)$ since Ψ_i^s is concave and $\mathcal{A}(b)$ is a convex set. In particular, we seek the largest $\ell_0 \in [0, \ell_{ij}^{\rho}]$ such that $\Psi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell_0 \mathbf{P}_{ij}) \geq 0$. If $\Psi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell_{ij}^{\rho} \mathbf{P}_{ij}) \geq 0$, then we set $\ell_0 := \ell_{ij}^{\rho}$. If this is not the case, then we must solve $h(\ell) := \Psi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij}) = 0$. In particular, since $h(0) = \Psi_i^s(\mathbf{U}_i^{\mathsf{L},n+1}) \geq 0$, and Ψ_i^s is continuous and concave, there exists at least one solution. If Ψ_i^s is strictly concave, then the solution set is a singleton; otherwise, the solution set is connected.

Remark 6.4.1. Note that $\Psi_i^n(\mathbf{U}_i^{\mathsf{L},n+1}) = 0$ occurs when $\gamma_i^n = \gamma_i^{\min,n}$. □

Remark 6.4.2 (Specific Internal Energy)**.** It is not necessary to perform any limiting on the internal energy. To see this, let $\mathbf{U}_i^{n+1}$ to be the final update after limiting on the surrogate entropy. Then $\Psi_i^s(\mathbf{U}_i^{n+1}) \geq 0$ implies the following,

$$\begin{aligned} e(\mathbf{U}_i^{n+1}) - q &\geq p_\infty(\tau_i^{n+1} - b) + S_i^{\min,n} \rho^{\gamma_i^{\min,n}} (1 - b\rho)^{1-\gamma_i^{\min,n}} \\ &\geq p_\infty(\tau_i^{n+1} - b). \end{aligned} \tag{6.48}$$

This is the invariant domain property that we expect when interpolating with the NASG EOS. As a reminder, if b, q, and p_∞ are unknown then we take $b = q = p_\infty = 0$ which results in positivity of the specific internal energy. □

6.5 The Quadratic Newton Method

For computational efficiency, the limiting is not actually performed on Ψ_i^s but rather, $\Phi_i^s(\boldsymbol{u}) := \rho \Psi_i^s(\boldsymbol{u})$ as this functional has some nicer properties. This is valid, as the solution sets for $\Psi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij}) = 0$ and $\Phi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij}) = 0$ are identical since $\rho_i^{\mathsf{L},n+1} + \ell \mathsf{P}_{ij}^{\rho} > 0$ for all $\ell \in [0, \ell_{ij}^{\rho}]$.

Lemma 6.5.1 (Sign of $f'''(\ell)$)**.** *Let* $\boldsymbol{u}_0 \in \mathcal{A}(b) := \{\boldsymbol{u} \in \mathbb{R}^{d+2} : 0 < 1 - b\rho < 1\}$ *and* $\boldsymbol{p} = (p_1, \dots, p_{d+2})^\mathsf{T} \in \mathbb{R}^{d+2}$*. Let* $\ell_0 \in [0, 1]$ *such that* $\boldsymbol{u}_0 + \ell_0 \boldsymbol{p} \in \mathcal{A}(b)$*. Define* $f(\ell) : [0, \ell_0] \ni \ell \mapsto f(\ell) := \Phi_i^s(\boldsymbol{u}_0 + \ell \boldsymbol{p})$*. Then the sign of* $f'''(\ell)$ *is constant over* $[0, \ell_0]$*.*

Proof. Using that $E = \rho e + \frac{\|\boldsymbol{m}\|_{\ell^2}^2}{2\rho}$, we write,

$$\Phi_i^s(\boldsymbol{u}) = \rho \Psi_i^s(\boldsymbol{u}) = \rho E - \tfrac{1}{2}\|\boldsymbol{m}\|_{\ell^2}^2 - p_\infty \rho(1 - b\rho) - S_{\min} \rho^{\gamma+1}(1 - b\rho)^{1-\gamma}. \tag{6.49}$$

Note that $f(\ell)$ is well defined for all $\ell \in [0, \ell_0]$ since $\boldsymbol{u}_0 + \ell_0 \boldsymbol{p} \in \mathcal{A}(b)$ and $\mathcal{A}(b)$ is a convex set. Hence $\boldsymbol{u}_0 + \ell \boldsymbol{p} \in \mathcal{A}(b)$ for all $\ell \in [0, \ell_0]$.

Next notice that $\rho E - \frac{1}{2}\|\boldsymbol{m}\|_{\ell^2}^2 - p_\infty \rho(1 - b\rho)$ is only quadratic in ρ, hence only the last term $S_i^{\min,n} \rho^{\gamma+1}(1 - b\rho)^{1-\gamma}$ remains when computing the third derivative. In particular, we have that,

$$f'(\ell) = D\Phi_i^s(\boldsymbol{u} + \ell \boldsymbol{p})\boldsymbol{p} \quad \text{and} \quad f''(\ell) = \boldsymbol{p}^\mathsf{T} \mathbb{H}\Phi_i^s(\boldsymbol{u} + \ell \boldsymbol{p})\boldsymbol{p}. \tag{6.50}$$

In particular, note that the Hessian of Φ_i^s is,

$$\mathbb{H}\Phi_i^s = \begin{bmatrix} 2p_\infty b - S_i^{\min,n} \frac{\rho^{\gamma-1}}{(1-b\rho)^{\gamma+1}} \big(\gamma(\gamma+1) - 2b\rho(\gamma + 1 - b\rho)\big) & \mathbf{0}^\mathsf{T} & 1 \\ \mathbf{0} & -\mathbb{I}_d & \mathbf{0} \\ 1 & \mathbf{0}^\mathsf{T} & 0 \end{bmatrix}. \tag{6.51}$$

Therefore, when computing the $f'''(\ell)$ the only term that remains when applying the chain rule is $\partial_\rho^3 \Phi_i^s$. Thus,

$$f'''(\ell) = (p_1)^3 \frac{\partial \Phi_i^s}{\partial \rho}(\boldsymbol{u} + \ell \boldsymbol{p}) = -(p_1)^3 S_i^{\min,n} \frac{\gamma(\gamma^2 - 1)(\rho + \ell p_1)^{\gamma-2}}{(1 - b(\rho + \ell p_1))^{\gamma+2}}. \tag{6.52}$$

Since $\boldsymbol{u} + \ell \boldsymbol{p} \in \mathcal{A}(b)$, for all $\ell \in [0, \ell_0]$ we see that $\operatorname{sgn}(f'''(\ell))$ has the same sign as $-(p_1)^3$; i.e., $\operatorname{sgn}(f'''(\ell))$ is constant. □

6.5.1 Review of Divided Differences

For the sake of clarity, we briefly review the notation and some results in regards to the Newton divided differences. Let $f \in C^n([a,b])$, then for $x_0, x_1 \in [a,b]$ with $x_0 \neq x_1$, we define the following,

$$f[x_0] := f(x_0), \quad f[x_0, x_1] := \frac{f[x_1] - f[x_0]}{x_1 - x_0}, \quad f[x_0, x_0] := f'(x_0). \tag{6.53}$$

So in general for $\{x_i\}_{i=1}^n \subset [a,b]$ we have,

$$f[x_0, \ldots, x_n] = \begin{cases} \frac{1}{n!} f^{(n)}(x_0), & \text{if } x_0 = x_1 = \cdots = x_n, \\ \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}, & \text{otherwise.} \end{cases} \tag{6.54}$$

In addition, notice that $f[x_0, \ldots, x_n]$ is unchanged by permutation. That is, given a permutation, $\sigma \in S(n)$, we have that $f[x_{\sigma(0)}, \ldots, x_{\sigma(n)}] = f[x_0, \ldots, x_n]$.

6.5.2 The Quadratic Newton Method

The quadratic Newton method was first introduced by Guermond & Popov in [59] and has been adapted for finding the root of $\Phi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij}) = 0$. We Using the divided difference notation, we define the "left" and "right" interpolating polynomials,

$$P_{\mathrm{L}}(\ell) := f(\ell_{\mathrm{L}}) + f'(\ell_{\mathrm{L}})(\ell - \ell_{\mathrm{L}}) + f[\ell_{\mathrm{L}}, \ell_{\mathrm{L}}, \ell_{\mathrm{R}}](\ell - \ell_{\mathrm{L}})^2 \tag{6.55a}$$

$$P_{\mathrm{R}}(\ell) := f(\ell_{\mathrm{R}}) + f'(\ell_{\mathrm{R}})(\ell - \ell_{\mathrm{R}}) + f[\ell_{\mathrm{L}}, \ell_{\mathrm{R}}, \ell_{\mathrm{R}}](\ell - \ell_{\mathrm{R}})^2 \tag{6.55b}$$

For P_{L}, we see that $P_{\mathrm{L}}(\ell_{\mathrm{L}}) = f(\ell_{\mathrm{L}})$, $P_{\mathrm{L}}(\ell_{\mathrm{R}}) = f(\ell_{\mathrm{R}})$, and $P'_{\mathrm{L}}(\ell_{\mathrm{L}}) = f'(\ell_{\mathrm{L}})$. Similarly for P_{R} we have $P_{\mathrm{R}}(\ell_{\mathrm{L}}) = f(\ell_{\mathrm{L}})$, $P_{\mathrm{R}}(\ell_{\mathrm{R}}) = f(\ell_{\mathrm{R}})$, and $P'_{\mathrm{R}}(\ell_{\mathrm{R}}) = f'(\ell_{\mathrm{R}})$.

Let ℓ^* denote the root $\Phi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell \mathbf{P}_{ij}) = 0$. For the quadratic Newton method, we will show that we can always find $\widetilde{\ell} \leq \ell^*$ for which $\widetilde{\ell} \uparrow \ell^*$. Hence, we guarantee that $\Phi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \widetilde{\ell} \mathbf{P}_{ij}) > 0$ at each step in the quadratic Newton method. Then we have the

following, Lemma 6.5.2, in regards to the approximation of $f(\ell)$ with $P_{\mathrm{L}}(\ell)$ and $P_{\mathrm{R}}(\ell)$. The pseudo-code for the quadratic Newton method is given in Algorithm 2 in Appendix B.

Lemma 6.5.2 (Interpolation Properties of P_{L} and P_{R})**.** *The following holds true:*

1. *The polynomials $P_L(\ell)$ and $P_R(\ell)$ bound the function $f(\ell)$ in the following sense:*

$$\min\big(P_L(\ell), P_R(\ell)\big) < f(\ell) < \max\big(P_L(\ell), P_R(\ell)\big), \qquad \forall \ell \in (\ell_L, \ell_R). \tag{6.56}$$

2. *$P_L(\ell)$ and $P_R(\ell)$ each have a unique zero over the interval (ℓ_L, ℓ_R) respectively given by*

$$\ell^L(\ell_L, \ell_R) := \begin{cases} \ell_L - \frac{2f(\ell_L)}{f'(\ell_L)+\sqrt{f'(\ell_L)^2-4f(\ell_L)f[\ell_L,\ell_L,\ell_R]}}, & \text{if } f[\ell_L,\ell_L,\ell_R] < 0, \\ \ell_L - \frac{2f(\ell_L)}{f'(\ell_L)-\sqrt{f'(\ell_L)^2-4f(\ell_L)f[\ell_L,\ell_L,\ell_R]}}, & \text{if } f[\ell_L,\ell_L,\ell_R] \geq 0, \end{cases} \tag{6.57a}$$

$$\ell^R(\ell_L, \ell_R) := \begin{cases} \ell_R - \frac{2f(\ell_R)}{f'(\ell_R)+\sqrt{f'(\ell_R)^2-4f(\ell_R)f[\ell_L,\ell_R,\ell_R]}}, & \text{if } f[\ell_L,\ell_R,\ell_R] < 0, \\ \ell_R - \frac{2f(\ell_R)}{f'(\ell_R)-\sqrt{f'(\ell_R)^2-4f(\ell_R)f[\ell_L,\ell_R,\ell_R]}}, & \text{if } f[\ell_L,\ell_R,\ell_R] \geq 0. \end{cases} \tag{6.57b}$$

3. *Properties 1. and 2. imply that*

$$\min\big(\ell^L(\ell_L, \ell_R), \ell^R(\ell_L, \ell_R)\big) < \ell^* < \max\big(\ell^L(\ell_L, \ell_R), \ell^R(\ell_L, \ell_R)\big). \tag{6.58}$$

Proof. Without loss of generality, let $P_{\mathrm{L}} \leq P_{\mathrm{R}}$ and $f'''(\ell) > 0$. Then, with a hefty amount of rewriting, one can show that,

$$f(\ell) - P_{\mathrm{L}}(\ell) = f[\ell_{\mathrm{R}}, \ell_{\mathrm{L}}, \ell_{\mathrm{L}}, \ell](\ell_{\mathrm{R}} - \ell)(\ell - \ell_{\mathrm{L}})^2 \tag{6.59}$$

1. Notice that, $f[\ell_{\mathrm{R}}, \ell_{\mathrm{L}}, \ell_{\mathrm{L}}, \ell_{\mathrm{R}}] = f'''(\xi) > 0$ for some $\xi \in (\ell_{\mathrm{L}}, \ell_{\mathrm{R}})$. Then we want to show

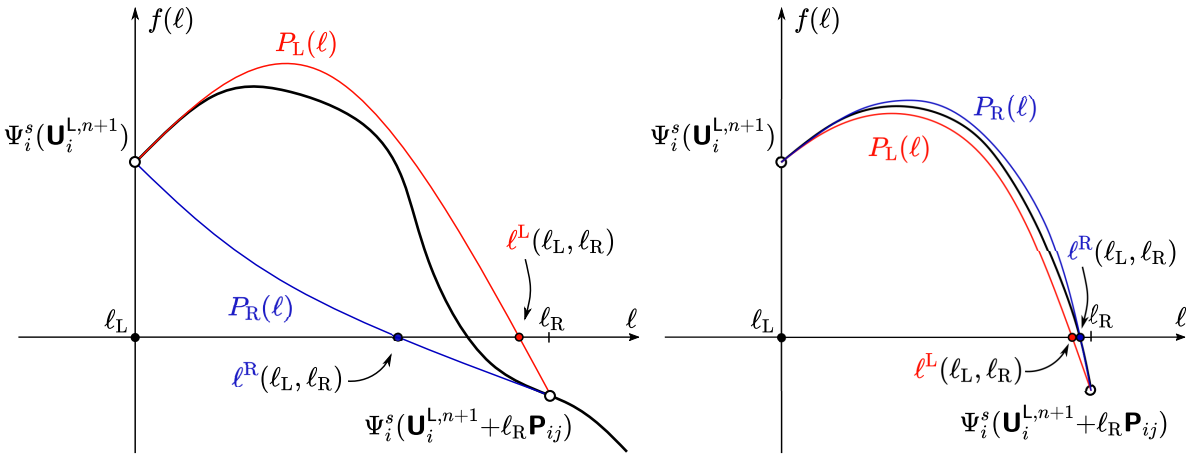

Figure 6.1: Two example visual descriptions of limiting using the quadratic newton method. Left: quasiconcave function that is not concave. Right: concave function.

that $f[\ell_{\mathrm{R}},\ell_{\mathrm{L}},\ell_{\mathrm{L}},\ell] > 0$ for $\ell \in [\ell_{\mathrm{L}},\ell_{\mathrm{R}})$. Also note that,

$$f[\ell_{\mathrm{R}},\ell_{\mathrm{L}},\ell_{\mathrm{L}},\ell] = \frac{f[\ell_{\mathrm{L}},\ell_{\mathrm{L}},\ell] - f[\ell_{\mathrm{R}},\ell_{\mathrm{L}},\ell_{\mathrm{L}}]}{\ell - \ell_{\mathrm{R}}} = \frac{f[\ell_{\mathrm{L}},\ell_{\mathrm{L}},\ell_{\mathrm{R}}] - f[\ell,\ell_{\mathrm{L}},\ell_{\mathrm{L}}]}{\ell_{\mathrm{R}} - \ell}. \tag{6.60}$$

Since $f'''(\ell) > 0$, then $f''(\ell)$ is an increasing function, and therefore, $f[\ell_{\mathrm{L}},\ell_{\mathrm{L}},\ell_{\mathrm{R}}] \geq f[\ell,\ell_{\mathrm{L}},\ell_{\mathrm{L}}]$ for all $\ell \in [\ell_{\mathrm{L}},\ell_{\mathrm{R}})$. Therefore, $f[\ell_{\mathrm{R}},\ell_{\mathrm{L}},\ell_{\mathrm{L}},\ell] \geq 0$ for all $\ell \in [\ell_{\mathrm{L}},\ell_{\mathrm{R}}]$. Dropping this right hand side this implies that, $f(\ell) > P_{\mathrm{L}}(\ell)$. Showing that $P_{\mathrm{R}}(\ell) > f(\ell)$ is similar, and if $f'''(\ell) < 0$ then the P_{R} is the lower bound and P_{L} is the upper bound.

2. Note the formula is simply the quadratic formula but rationalizing the numerator to avoid division by zero in the degenerate case. Since both quadratic functions interpolate a positive and a negative values, the function must have a single root. The root which lies in the interval $(\ell_{\mathrm{L}},\ell_{\mathrm{R}})$ depends on the concavity or convexity of the quadratic function. Hence the result.

3. This follows immediately from 1. and 2. □

6.6 Relaxation on the Surrogate Entropy

We also perform relaxation on the surrogate entropy exactly as described in [60, Sec. 4.7.2]. Specifically, we relax the bound, $S_i^{\min,n}$. Specifically, we define the new bound,

$$\widetilde{S_i^{\min,n}} := \max((1 - r_h)S_i^{\min,n}, S_i^{\min,n} - \Delta S_i^n) \tag{6.61}$$

where $\Delta S_i^n := \max_{j \in \mathcal{G}(i) \setminus \{i\}}(\frac{1}{2}(S_i^n + S_j^n) - S_i^{\min,n})$ and $r_h := (m_i/|D|)^{1.5/d}$. This type of relaxation was originally done by Khobalatte & Perthame in [75, Sec. 3.3] in order to guarantee second order convergence of their numerical method. Since the entropy is constant in regions of smooth flow, we loosen the bound on the order of $\mathcal{O}(h^2)$. However, for the surrogate entropy, we cannot say that $S(\boldsymbol{u}; \gamma_i^{\min,n})$ will be constant in the region of smooth flow and in fact it may even decrease. If there is a shock wave then the relaxation is irrelevant as the first order scheme is used based on the entropy viscosity method outlined in Chapter 5.

7. NUMERICAL RESULTS*

In this chapter we present a variety of numerical illustrations to demonstrate the robust effectiveness of the proposed methods. This chapter is broken up into two parts one-dimensional and two-dimensional problems. When relevant, first order and second order solutions will be compared.

The numerical results for the one-dimensional problems have been performed on an in-house code written in Fortran originally developed by Dr. Jean-Luc Guermond and uses $\mathbb{P}_1$ finite elements. This code base will be referred to as the `TAMU` code. The two-dimensional problems have been performed on a software named `Ryujin` which can be found at `https://github.com/conservation-laws/ryujin`. `Ryujin` is an efficient multithreaded massively parallel C++ code which uses the `deal.II` finite element library, [76]. The design of the efficient and parallel algorithms for `Ryujin` is done by Maier & Kronbichler in [65]. The two-dimensional simulations were performed on the Texas A&M Mathematics department's cluster, nicknamed Whistler, which used anywhere between 10 and 32 compute nodes.

The time step for the `TAMU` code was computed by the following,

$$\Delta t := \mathrm{CFL} \min_{i\in\mathcal{V}} \Big(- \frac{m_i}{2d_{ii}^{\mathsf{L},n}} \Big). \tag{7.1}$$

In the 2D simulations, the boundary conditions we enforce are Dirichlet, slip, and outflow. For the slip condition, we want to enforce $\boldsymbol{v} \cdot \boldsymbol{n} = 0$. This is achieved by redefining the momentum on the slip boundary with,

$$\mathbf{M}_i^{n+1} := \mathbf{M}_i^{n+1} - (\mathbf{M}_i^{n+1} \cdot \boldsymbol{n}_i)\boldsymbol{n}_i. \tag{7.2}$$

where $i \in \mathcal{V}$ is the index of a node on the boundary, ∂D, and $\boldsymbol{n}_i$ is the outward normal

* A majority of the simulations presented in this section are also reported in [2] and [1] and is reprinted with permission from [2] and [1].

at node $\boldsymbol{x}_i$. Since the domain is polygonal, we define $\boldsymbol{n}_i = \frac{1}{2}(\boldsymbol{n}_i^- + \boldsymbol{n}_i^+)$ where $\boldsymbol{n}_i^-$ and $\boldsymbol{n}_i^+$ are the outward normals on the boundary edges which connect to $\boldsymbol{x}_i$. For the outflow condition, we use an approximate solution to the Riemann problem, the details of which can be found in Appendix B.3. This is by no means a perfect outflow implementation, but in certain circumstances, it can work properly. It should be noted that the solution can blow up on the boundary with this implementation. However, effective implementations of outflow boundary conditions are not the focus of this thesis nor on the specific numerical demonstrations.

Lastly, the Schlieren plots mentioned in the 2D section are computed with the following formula,

$$\exp\Big(-\beta\frac{r_i^n - \min_{j\in\mathcal{V}} r_j^n}{\max_{j\in\mathcal{V}} r_j^n - \min_{j\in\mathcal{V}} r_j^n}\Big), \qquad \text{where} \qquad r_i^n := \frac{1}{m_i}\|\sum_{j\in\mathcal{J}(i)} \boldsymbol{c}_{ij}\rho_j^n\|_{\ell^2}. \tag{7.3}$$

We use $\beta = 15$ for all 2D simulations. Generally speaking, this formula plots the gradient of the pressure and is constructed so as to mimic the photography style known as "Schlieren photography."

7.1 Convergence Tests

Let $(\rho_h(t), \boldsymbol{m}_h(t), E_h(t))$ denote the approximate solution at time t. We define the consolidated error indicator as the sum of the relative errors for the density, momentum, and total energy; that is,

$$\delta_q(t) := \frac{\|\rho_h(t) - \rho(t)\|_{L^q(D)}}{\|\rho(t)\|_{L^q(D)}} + \frac{\|\boldsymbol{m}_h(t) - \boldsymbol{m}(t)\|_{\boldsymbol{L}^q(D)}}{\|\boldsymbol{m}(t)\|_{\boldsymbol{L}^q(D)}} + \frac{\|E_h(t) - E(t)\|_{L^q(D)}}{\|E(t)\|_{L^q(D)}}, \tag{7.4}$$

for $q \in [1, \infty]$.

7.1.1 The Fan-Jump-Fan Composite Wave

We begin by verifying that our numerical method converges. In particular, we show convergence for a Riemann problem with the van der Waals equation of state. Recall the

van der Waals equation of state is, $p(\rho, e) = (\gamma - 1)\frac{\rho e + a\rho^2}{1 - b\rho} - a\rho^2$. We use $\gamma = 1.02$, $a = 1$, and $b = 1$. The left and right states are chosen to be,

$$(\rho_{\mathrm{L}}, v_{\mathrm{L}}, p_{\mathrm{L}}) := (0.10, -0.475504638574729, 0.022084258693080), \tag{7.5}$$

$$(\rho_{\mathrm{R}}, v_{\mathrm{R}}, p_{\mathrm{R}}) := (0.39, -0.121375781741349, 0.039073167077590), \tag{7.6}$$

and the computational domain is $D = (-1, 1)$ and we use CFL $= 0.5$. The solution to this Riemann problem is a fan-jump-fan composite wave. That is, the solution is a single wave (3-wave) and is composed of an expansion, immediately followed by a shock and into another expansion. The construction of the exact solution to this problem can be found in the supplementary material of Clayton et. al. [2] at [local/web 319KB]. The exact solution to this problem has also been constructed in [77] and [78]; similar constructions can also be seen in [79].

Convergence rates are reported in Table 7.1 and plots of the approximate and exact solutions are shown in Figure 7.1 for the density and pressure.

#dof	$\delta_1(t)$	rate	$\delta_2(t)$	rate
101	2.14E-01	–	2.67E-01	–
201	1.44E-01	0.58	2.07E-01	0.37
401	9.40E-02	0.62	1.58E-01	0.39
801	5.96E-02	0.66	1.20E-01	0.40
1601	3.66E-02	0.70	8.96E-02	0.42
3201	2.18E-02	0.75	6.66E-02	0.43
6401	1.27E-02	0.78	4.93E-02	0.43
12801	7.26E-03	0.81	3.66E-02	0.43
25601	4.09E-03	0.83	2.72E-02	0.43

Table 7.1: Consolidated errors and convergence rates for the fan-jump-fan composite wave. Solution computed at $t = 5.0$. Reprinted with permission from [2].

Remark 7.1.1 (Existence of Composite Waves)**.** For the van der Waals EOS, the composite

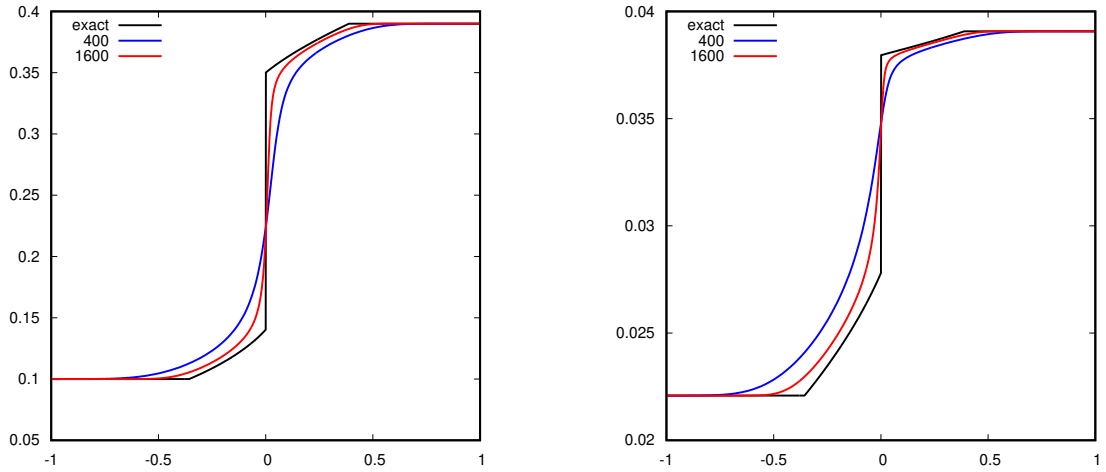

Figure 7.1: Comparison of the exact solution for density (left) and pressure (right) for the fan-jump-fan composite wave. Approximate solutions are computed using 400 and 1600 DoFs and the corresponding mesh sizes are $h = 0.005$ and $h = 0.00125$, respectively.

waves can only exist if $\gamma < 1.06$. This condition implies there is a region in the thermodynamic phase space where the isentropes are non-convex in the (p, τ) diagram. Hence composite waves can emerge. This has been shown in [80, Section 3.1], [78], and [79, Section 6.3]; more information on composite waves can also be found in [37]. □

7.1.2 Smooth Wave with Various EOS

To further demonstrate the robustness, we show convergence rates for the first and second order methods of a smooth travelling wave. The exact solution for the smooth wave is,

$$\rho(x,t) = \begin{cases} \rho_0 + 2^6(x_1 - x_0)^{-6}(x - v_0 t - x_0)^3(x_1 - x + v_0 t)^3 & \text{if } x_0 \le x - v_0 t \le x_1, \\ \rho_0 & \text{otherwise,} \end{cases} \tag{7.7a}$$

$$v(x,t) = v_0, \qquad p(x,t) = p_0, \tag{7.7b}$$

where x_0 and x_1 are arbitrary constants with $x_0 < x_1$. The constants, ρ_0, v_0, and p_0 are chosen depending on the EOS being used. For our simulations we choose the computational domain to be $D = (0, 1)$ with $x_0 = 0.1$ and $x_1 = 0.3$. All tests were performed using

CFL = 0.1. Note that the solution to this problem is independent of the EOS since the exact solution has constant pressure.

7.1.2.1 Ideal EOS

This problem has been performed before in [60] but we repeat it here for comparison. We the parameters, $\gamma = 1.4$, $\rho_0 = v_0 = p_0 = 1$, and for the interpolation parameters, we set $b = q = p_\infty = 0$. The final time is run to $t = 0.6$.

7.1.2.2 Van der Waals EOS

We set the parameters for the van der Waals EOS to be, $a = 1$, $b = 0.075$, $\gamma = 1.4$ and $\rho_0 = v_0 = p_0 = 1$. The interpolation parameters are set to be $b = 0.075$ and $q = p_\infty = 0$. The final time is run to $T = 0.6$.

7.1.2.3 Jones-Wilkins-Lee EOS

We set the parameters for the Jones-Wilkins-Lee EOS to be, $A = 1$, $B = -1$, $R_1 = 2$, $R_2 = \omega = \rho_0 = v_0 = p_0 = 1$. The interpolation parameters are set to be $b = q = p_\infty = 0$ and the final time is run to $t = 0.6$.

7.1.2.4 Mie-Gruneisen EOS

We set the parameters for the Mie-Gruneisen EOS to be, $\widetilde{\rho}_0 = 2790$, $c_0 = 5330$, $s = 1.34$, $P_0 = 0$, $e_0 = 0$, $\Gamma_0 = 2.00$, $\rho_0 = 3500$, $v_0 = 1 \times 10^4$ and $p_0 = 1 \times 10^{11}$. The interpolation parameters are set to $b = q = p_\infty = 0$ and the final run time is $t = 6 \times 10^{-5}$.

The convergence results for the second order method are reported in Table 7.2.

7.1.3 The Isentropic Vortex with van der Waals EOS

To demonstrate the convergence of the high and low order methods in 2D, we run a simulation for the isentropic vortex using the van der Waals equation of state. The exact

	Ideal		VdW		JWL		MG	
$\lvert\mathcal{V}\rvert$	$\delta_\infty(t_{\text{final}})$		$\delta_\infty(t_{\text{final}})$		$\delta_\infty(t_{\text{final}})$		$\delta_\infty(t_{\text{final}})$	
101	1.94e-02	–	1.24e-01	–	7.93e-02	–	1.24e-05	–
201	4.03e-03	2.27	6.24e-03	4.30	2.53e-02	1.65	2.56e-06	2.28
401	7.91e-04	2.35	9.92e-04	2.65	3.61e-03	2.81	5.03e-07	2.35
801	1.44e-04	2.46	1.75e-04	2.51	1.31e-04	4.78	9.17e-08	2.46
1601	2.75e-05	2.39	3.29e-05	2.41	2.51e-05	2.38	1.75e-08	2.39
3201	5.18e-06	2.41	6.17e-06	2.41	4.73e-06	2.41	3.29e-09	2.41
6401	9.69e-07	2.42	1.16e-06	2.42	8.87e-07	2.42	6.22e-10	2.41

Table 7.2: $\delta_\infty(t_{\text{final}})$ error defined in equation (7.4) and corresponding convergence rates with various EOS for the one-dimensional smooth traveling wave problem with exact solution (7.7) under uniform refinement of the interval $D = (0, 1)$. Reprinted with permission from [1].

solution is given by,

$$\rho(\boldsymbol{x}, t) = \left[\frac{3C}{8a} - \frac{1}{2}\sqrt{\frac{9C^2}{16a^2} + \frac{2}{a}\big(F + \frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2\big)}\right]^2, \qquad \boldsymbol{x} \in \mathbb{R}^2, \quad t > 0, \tag{7.8a}$$

$$\boldsymbol{v}(\boldsymbol{x}, t) = \boldsymbol{v}_\infty + \psi(\overline{\boldsymbol{x}})\big(-\bar{x}_2, \bar{x}_1\big), \qquad \boldsymbol{x} \in \mathbb{R}^2, \quad t > 0, \tag{7.8b}$$

$$p(\boldsymbol{x}, t) = C(\gamma - 1)\rho(\boldsymbol{x}, t)^\gamma - a\rho(\boldsymbol{x}, t)^2, \qquad \boldsymbol{x} \in \mathbb{R}^2, \quad t > 0. \tag{7.8c}$$

where $\overline{\boldsymbol{x}} := \boldsymbol{x} - \boldsymbol{x}^0 - \boldsymbol{v}_\infty t = (\overline{x}_1, \overline{x}_2)$, $C := (p_\infty + a\rho_\infty^2)/\rho_\infty^{3/2}$, $F := -a\rho_\infty - 3p_\infty/\rho_\infty$. Here ρ_∞ and p_∞ are the density and pressure in the far field, and $\psi(\boldsymbol{x}) := \frac{\beta}{2\pi}\exp\big(\frac{1}{2}(1 - \frac{\|\boldsymbol{x}\|_{\ell^2}^2}{r_0^2})\big)$. We set the far field conditions to $\rho_\infty = 0.1$, $p_\infty = 1$ and $\boldsymbol{v}_\infty = (1, 1)$. We also set $\gamma = \frac{3}{2}$ and $a = 1$. This gives $C = \frac{101}{\sqrt{10}}$ and $F = -\frac{301}{10}$. The rest of the constants are set as follows: $\boldsymbol{x}^0 = (-1, -1)$, $r_0 = 1$, $\beta = 20$. The derivation of the exact solution is shown in Appendix A.1 for the ideal and van der Waals EOS.

The numerical simulation is performed on the square domain, $D = (-5, 5) \times (-5, 5)$. We impose Dirichlet boundary conditions on all sides of D, since the solution decays rapidly to the far field state, $(\rho_\infty, \boldsymbol{v}_\infty, p_\infty)$, away from $\boldsymbol{x}^0$. Convergence results are reported in Table 7.3.

$\lvert\mathcal{V}\rvert$	$\delta_1(t_{\text{final}})$		$\delta_2(t_{\text{final}})$		$\delta_\infty(t_{\text{final}})$	
289	1.17e-01	–	2.01e-01	–	6.82e-01	–
1089	1.18e-02	3.46	2.65e-02	3.06	1.05e-01	2.82
4225	7.92e-04	3.98	1.96e-03	3.84	7.87e-03	3.82
16641	5.57e-05	3.87	1.32e-04	3.93	5.50e-04	3.88
66049	5.07e-06	3.48	1.20e-05	3.48	7.79e-05	2.83
263169	7.55e-07	2.76	2.25e-06	2.42	2.03e-05	1.95
1050625	1.64e-07	2.20	5.51e-07	2.04	5.52e-06	1.88
4198401	4.08e-08	2.01	1.38e-07	2.00	1.51e-06	1.87

Table 7.3: The consolidated error defined in equation (7.4) and convergence rates for the isentropic vortex problem with the Van der Waals EOS. The exact solution is given in (7.8). Reprinted with permission from [1].

7.2 The Two-Expansion Wave Speed Estimate

As mentioned in Chapter 4 the computation of the max wave speed for a local Riemann problem is extremely difficult if the EOS is complicated. A common heuristic approach is to use the so-called *two-expansion approximation* (or *two-rarefaction approximation*) for the wave speed. That is, an approximate wave speed is defined by,

$$\lambda^{\text{exp}} := \max\{|v_{\text{L}} - a_{\text{L}}|, |v_{\text{R}} + a_{\text{R}}|\}. \tag{7.9}$$

Note that a_{L} and a_{R} are the material sound speed for the oracle and should not be confused with the sound speed used in the NASG EOS used in the interpolation in Chapter 4. We display several test problems for which this heuristic estimation is not robust. It can lead to an overestimation or underestimation of the artificial viscosity which leads immediate problems.

The test problems in this section all use the van der Waals EOS with $\gamma = 1.02$, $a = 1$, and $b = 1$ as in Section 7.1.1.

7.2.1 Underestimation of Max Wave Speed: Test 1

In this problem we show that the two-expansion approximation leads to an underestimation of the viscosity which leads to non-physical results. The Riemann data is,

$$(\rho_{\mathrm{L}}, v_{\mathrm{L}}, p_{\mathrm{L}}) := (0.2450, 0, 2.9123894332846005 \times 10^{-2}), \tag{7.10}$$

$$(\rho_{\mathrm{R}}, v_{\mathrm{R}}, p_{\mathrm{R}}) := (0.1225, 0, 2.0685894810791836 \times 10^{-2}), \tag{7.11}$$

with corresponding sound speeds $(a_{\mathrm{L}}, a_{\mathrm{R}}) \approx (0.00399, 0.306)$. The computational domain is $D = (-0.5, 1)$ and we use CFL $= 0.5$. The simulation is run up to $t = 1.25$. The results for the low order method using Algorithm 1 for computing an upper bound on the max wave speed are seen in Figure 7.2. When we use $\lambda^{\exp}$ for the max wave speed, the computation generates complex sound speed after a few time steps. A comparison of these methods can be seen in Figure 7.3.

7.2.2 Underestimation of Max Wave Speed: Test 2

For the second test, we use the following Riemann data,

$$(\rho_{\mathrm{L}}, v_{\mathrm{L}}, p_{\mathrm{L}}) := (2.5 \times 10^{-1}, 0, 3 \times 10^{-2}), \tag{7.12}$$

$$(\rho_{\mathrm{R}}, v_{\mathrm{R}}, p_{\mathrm{R}}) := (4.9 \times 10^{-5}, 0, 5 \times 10^{-8}), \tag{7.13}$$

The computational domain is $D = (-0.5, 1)$ and we use CFL $= 0.5$. The final time is $t = 0.4$. As before, using the estimation of $\widehat{p}^*$ from Section 4.6 we have the physically relevant solution shown in Figure 7.4. When we use $\lambda^{\exp}$ as the estimate, the simulation immediately crashes, generating negative specific internal energy; this crash occurs no matter the CFL number.

7.2.3 Overestimation of Max Wave Speed: Test 3

For the third test, we demonstrate that $\lambda^{\exp}$ can be a gross overestimate of the max wave speed resulting in a dramatic increase in computational time. This is due a decrease in the

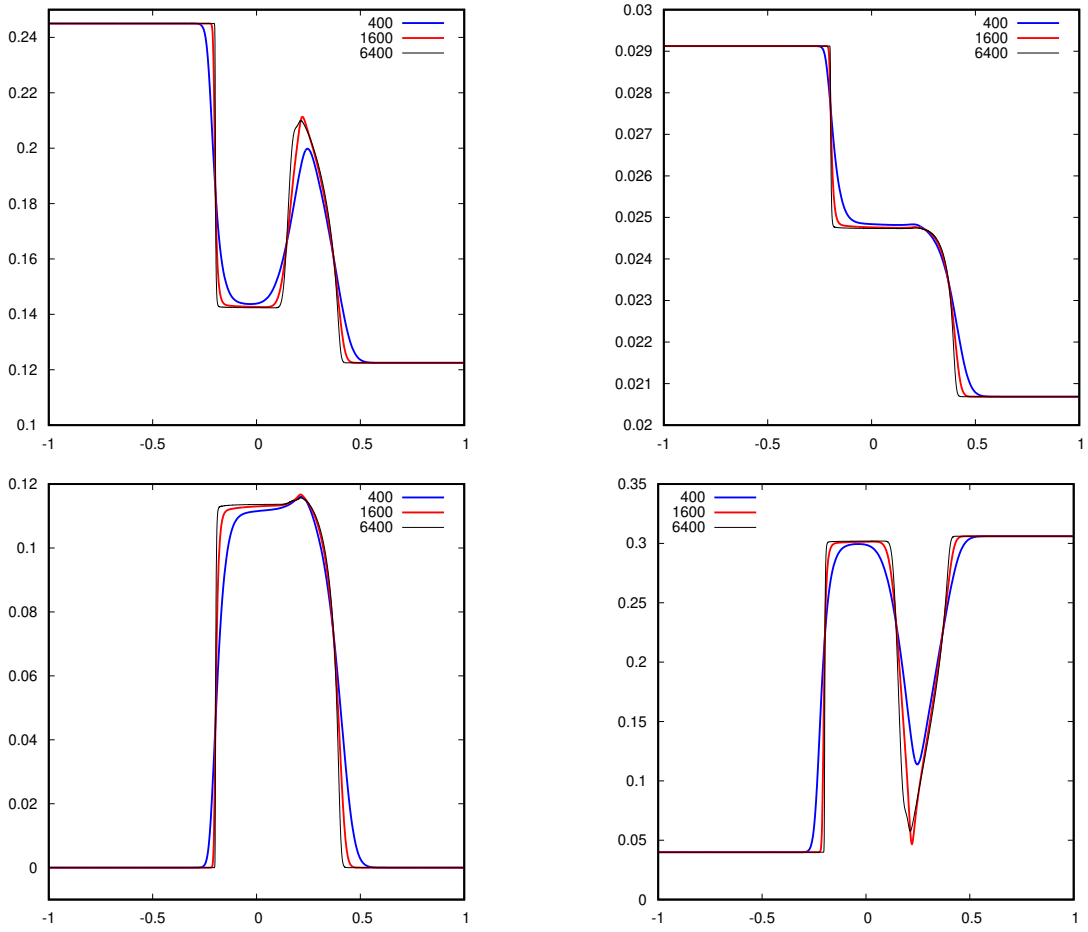

Figure 7.2: Plots for Test 1 of the underestimation problem using $\widehat{p}^*$ for computing the maximum wave speed. (Top left): density, (top right): pressure, (bottom left): velocity, (bottom right): sound speed.

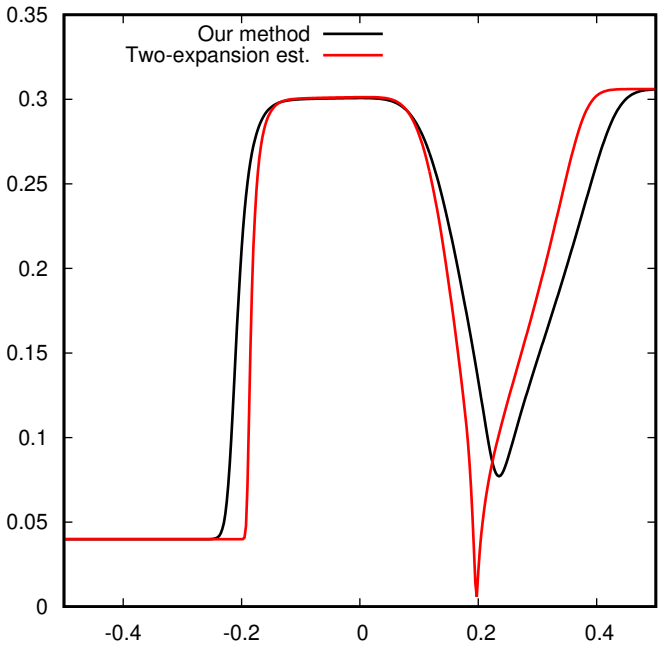

Figure 7.3: Comparison of the sound speed for our method versus the two expansion approximation. This test was run with 800 DoFs up to time $t = 1.2$. The simulation using the two-expansion estimate immediately crashes after $t = 1.2$ generating complex sound speed.

possible range of time step sizes. The Riemann data is,

$$(\rho_{\mathrm{L}}, v_{\mathrm{L}}, p_{\mathrm{L}}) := (0.9932, 3, 2), \tag{7.14}$$

$$(\rho_{\mathrm{R}}, v_{\mathrm{R}}, p_{\mathrm{R}}) := (0.9500, -3, 2). \tag{7.15}$$

The corresponding sound speeds are $(a_{\mathrm{L}}, a_{\mathrm{R}}) \approx (21.2, 7.77)$. The computational domain is $D = (-1.7, 1)$ and the simulation is run to $t_{\mathrm{final}} = 0.005$. When using λ^{exp} we find that the simulation requires a CFL of at most 0.12 in order to maintain positive internal energy. The maximal CFL number we were able to use, using the wave speed, $\widehat{\lambda}_{\max} = \lambda_{\max}(\widehat{p}^*)$, computed by Algorithm 1, while still preserving positive internal energy and real sound speed was 1.42. R (The sound speed and internal energy are checked at every node and at every time step.) Plots of the solution are shown in Figure 7.5. In this case, using the method with λ^{exp} results in a computational time approximately 12 times greater than the method using $\lambda_{\max}(\widehat{p}^*)$.

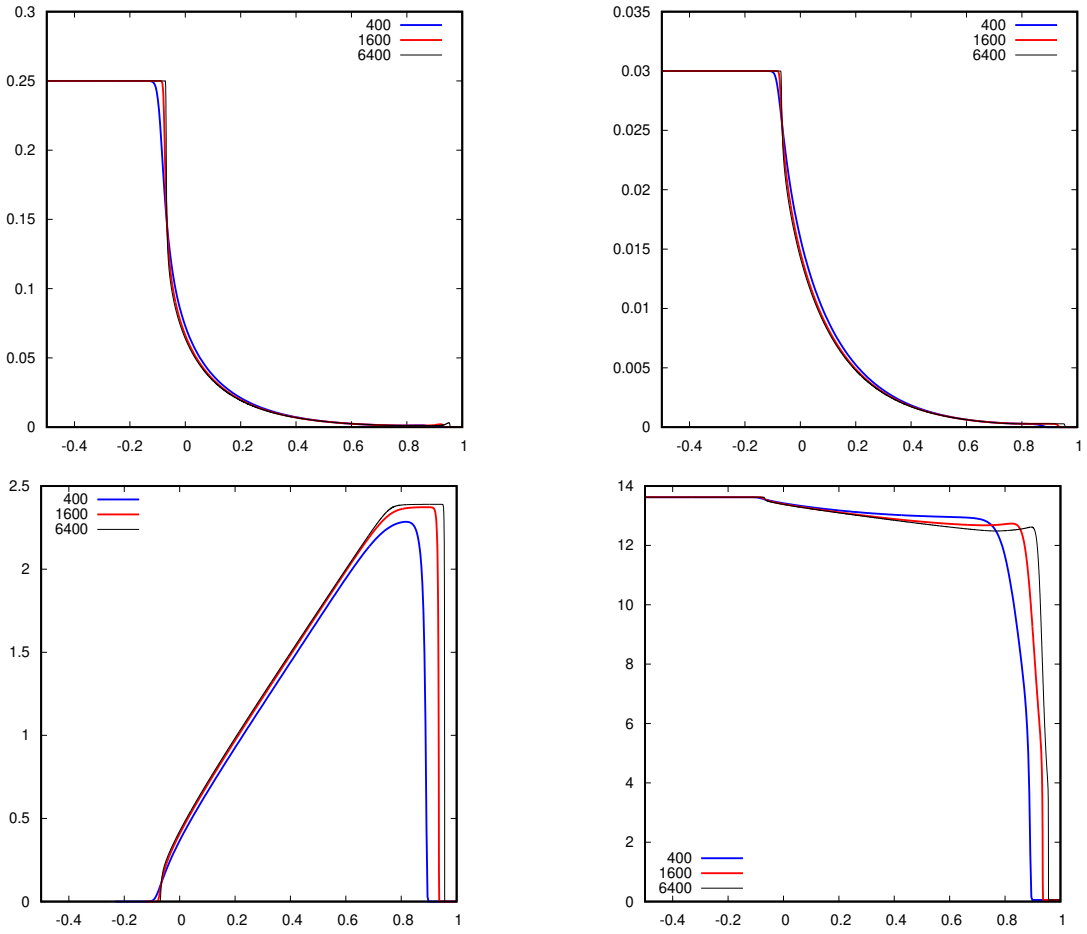

Figure 7.4: Plots for Test 2 of the underestimation problem using $\widehat{p}^*$ for computing the maximum wave speed. (Top left): density, (top right): pressure, (bottom left): velocity, (bottom right): specific internal energy. Final time is $t = 0.4$.

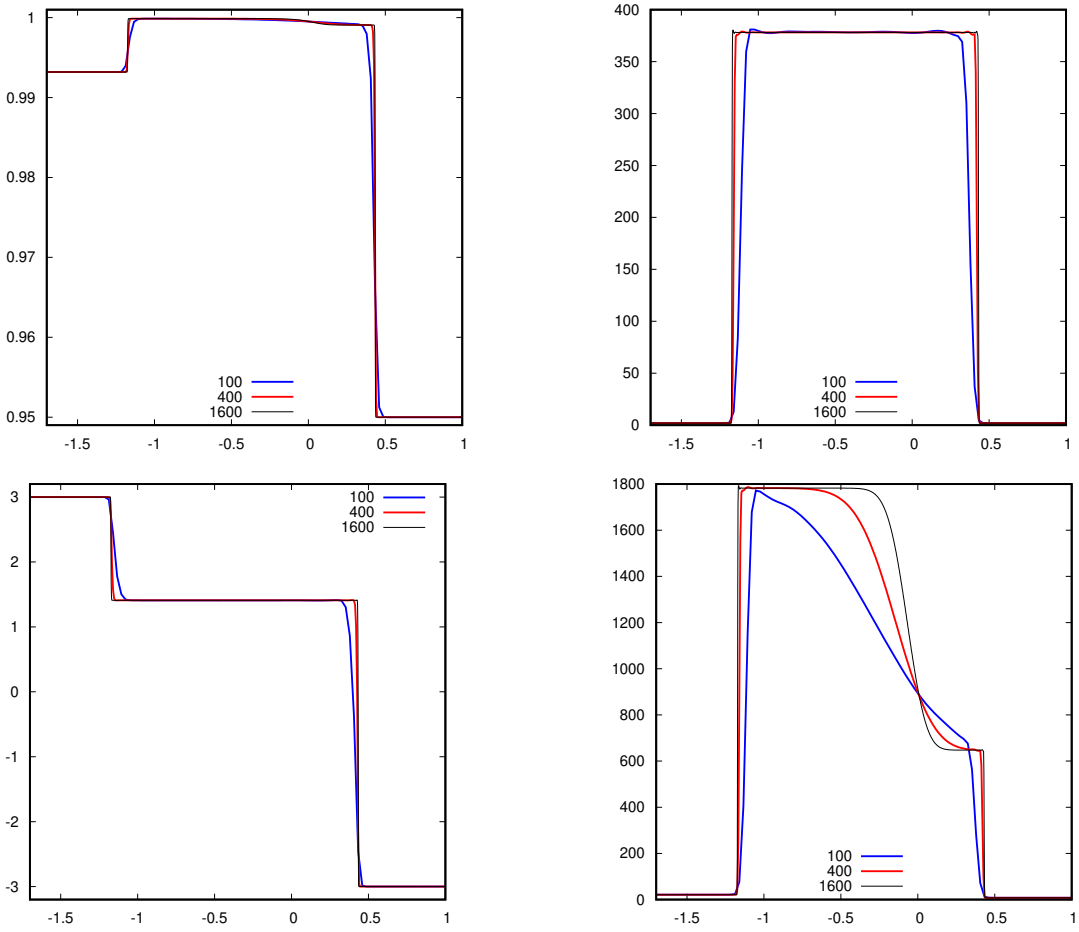

Figure 7.5: Plots for Test 3 of the overestimation problem using $\widehat{p}^*$ for computing the maximum wave speed. From left to right: density, pressure, velocity, sound speed. Final time is $t = 0.005$ and using CFL $= 1.42$. Figures were similar when using $\lambda^{\exp}$.

7.3 The SESAME Database

In order to demonstrate the robustness of our method, we use the SESAME database. The SESAME database is a tabulated EOS developed by the Mechanics of Materials Equation of State group (T-1) at Los Alamos National Lab [81]. The SESAME database houses data on a large variety of materials from argon to cesium. The construction of such a database relies on experimental data and EOS models fitted to the material. However, such details are beyond the scope of this thesis. Access to this database can be acquired by contacting sesame@lanl.gov. The database alone is not usable by itself, we need a way to interface with it. This is done with EOSPAC6 [82]; all versions of this software can also be found at https://laws.lanl.gov/projects/data/eos/eospacReleases.php.

The general process by which EOSPAC6 functions, is that we provide a material identification number and "table type". The "table type" indicates the thermodynamic relationship; that is, it specifies the two thermodynamic input quantities and the corresponding output quantity. For our purposes, we provide $\{(\rho_i^n, e_i^n)\}_{i\in\mathcal{V}}$ and EOSPAC6 returns $\{p_i^n\}_{i\in\mathcal{V}}$. Recall for density, pressure and specific internal energy in the compressible Euler equations are, kg/m^3, Pa, and J/kg, respectively. However, EOSPAC6 works with the following respective units, Mg/m^3, GPa, and MJ/kg. A full description of the functionality of EOSPAC6 can be found in the user manual in Pimentel [83]. Additionally, the list of the available materials in the SESAME database with their corresponding identification numbers are found on the last pages of [81].

We conduct a series of tests utilizing the SESAME database. The Riemann problems all use $x = 0$ as the initial discontinuity. Furthermore, we offer comparisons of the different values of $\widehat{\lambda}_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i, \mathbf{U}_j)$. As $\widehat{\lambda}_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i, \mathbf{U}_j)$ depends on the Riemann problem we use the definition,

$$\widetilde{\lambda}_i^{\max} := \max_{j\in\mathcal{G}(i)\setminus\{i\}} \widehat{\lambda}_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i, \mathbf{U}_j), \tag{7.16}$$

in order to plot the maximum wave speed.

7.3.1 Expansion-Contact-Shock Comparison

For the first problem, we run a Riemann problem using dry air (material id: 5030), aluminum (material ID: 3720), vanadium (material ID: 2552), and sulfur hexaflouride (material ID: 7010) with data which generates a wave profile similar to the Sod shock tube. The Riemann data is,

$$(\rho_{\mathrm{L}}, v_{\mathrm{L}}, e_{\mathrm{L}}) := (0.01\,\mathrm{Mg/m^3}, 0\,\mathrm{m/s}, 4000\,\mathrm{MJ/kg}), \tag{7.17}$$

$$(\rho_{\mathrm{R}}, v_{\mathrm{R}}, e_{\mathrm{R}}) := (0.003\,\mathrm{Mg/m^3}, 0\,\mathrm{m/s}, 3400\,\mathrm{MJ/kg}). \tag{7.18}$$

The problem is simulated on $D = (-1\,\mathrm{m}, 1\,\mathrm{m})$ to a final time of $t = 1.2 \times 10^{-5}\,\mathrm{s}$ and CFL $= 0.5$. See Figure 7.6 for the results.

7.4 Benchmark Configurations

7.4.1 EOS Comparison in a Riemann Problem

In this simulation we compare the different effects an EOS has the solution to a Riemann problem. The initial discontinuity is at $x = 0$, the Riemann data we use is,

$$(\rho_{\mathrm{L}}, v_{\mathrm{L}}, p_{\mathrm{L}}) := (1, 1, 2), \tag{7.19}$$

$$(\rho_{\mathrm{R}}, v_{\mathrm{R}}, p_{\mathrm{R}}) := (1, -1, 1), \tag{7.20}$$

on $D = (-0.5, 0.5)$ with CFL $= 0.5$ to the final time of $t_{\mathrm{final}} = 0.1$. We compare results with the ideal, covolume, van der Waals and EOS. The parameters we use are, $\gamma = 1.4$, $a = 0.5$, $b = 0.5$, $R = 0.4$, and $c_V = 1$ for each of the associated EOS. The results are shown in Figures 7.7.

7.4.2 The Woodward-Colella Blast Wave

A common test problem in the literature is the Woodward-Colella blast wave which was originally proposed in [84]. The initial configuration is composed of three constant states

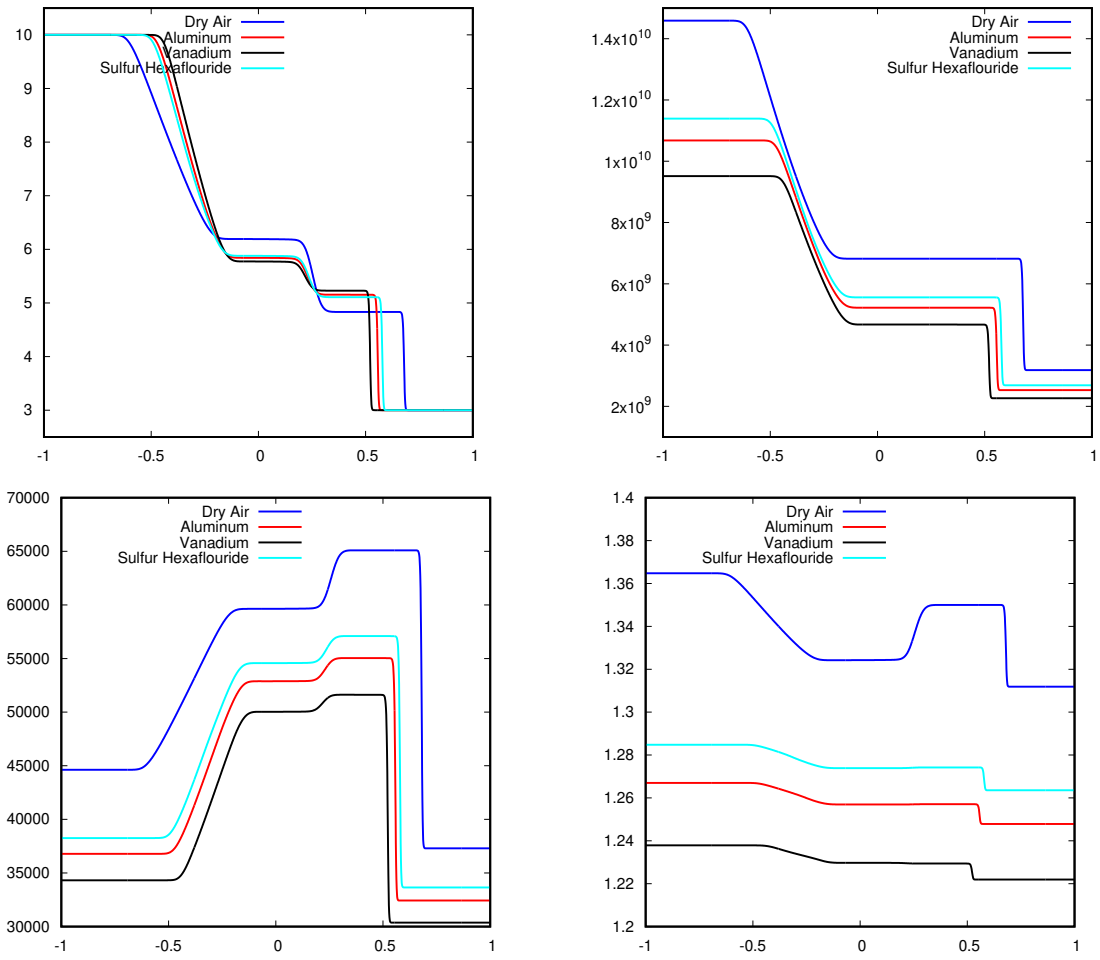

Figure 7.6: Comparison of ρ (top left), p (top right), max wave speed, $\widetilde{\lambda}^{\max}$ (bottom left), and γ (bottom right), for the various materials at the final time, $t = 1.2 \times 10^{-5}$ s.

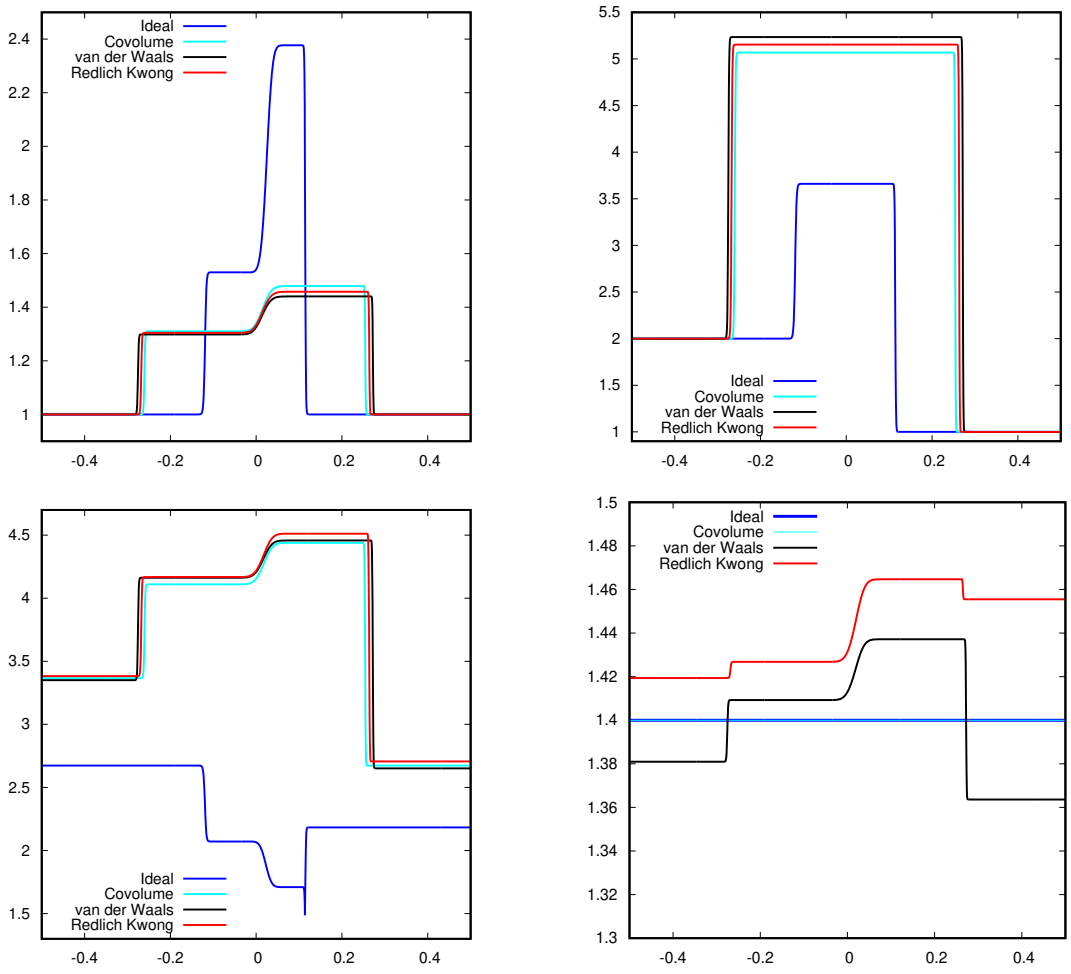

Figure 7.7: Comparison of the density (top left), pressure (top right), max wave speed (bottom left), and γ (bottom right), for the various different EOS at the final time, $t = 0.1$.

which result in the collision of two waves. This collision produces an emergent complex structure. This test problem is normally performed with the ideal EOS; we instead perform two experiments, both using the Jones-Wilkins-Lee EOS. The first experiment uses the parameters proposed in [23, Section 5.1] and the second experiment uses the parameters in [44, Table 2. "HMX"], see Table 7.4 for the values.

The computational domain is $D = (0, 1)$ and the initial state is,

$$(\rho_0(x), v_0(x), p_0(x)) = \begin{cases} (1, 0, 10^3) & \text{if } x \in [0, 0.1], \\ (1, 0, 10^{-2}) & \text{if } x \in (0.1, 0.9), \\ (1, 0, 10^2) & \text{if } x \in [0.9, 1]. \end{cases} \tag{7.21}$$

and the Jones-Wilkins-Lee parameters for the two experiments are given in Table 7.4.

	A	B	R_1	R_2	ω	ρ_0	t_{final}
Case 1	6.321×10^2	-4.472	11.3	1.13	0.8938	1	0.038
Case 2	7.7828×10^{11}	7.071428×10^9	4.2	1.00	0.3000	1891	0.038

Table 7.4: JWL parameters for Woodward-Colella interacting blast wave benchmark. Reprinted with permission from [1].

We use CFL $= 0.9$. The results are recorded in Figure 7.8 and Figure 7.9.

7.4.3 Shock Collision with Triangular Obstacle

The next benchmark problem is referred to in the literature as the Schardin's problem, see Schardin [85] for the original physical experiment. This problem is for studying the effects of a shock wave colliding with a triangular obstacle. This physical experiment has been recreated in [86] with a detailed description of the experimental setup.

For our numerical demonstration we use the setup by Toro et. al. in [23, Sec. 5.4]. The computational domain is $D = (-0.65, 0.5) \times (-0.5, 0.5) \setminus K$, where K is the triangle formed

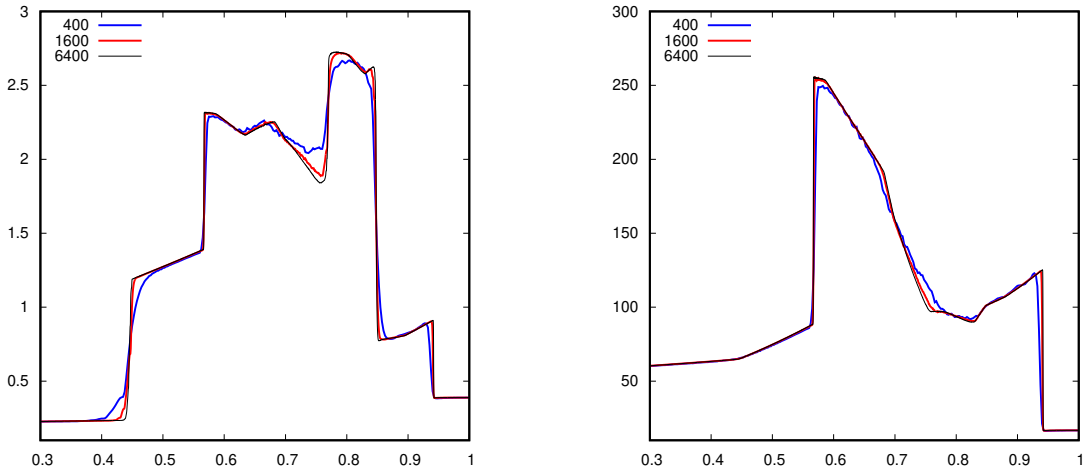

Figure 7.8: Case 1 of the Woodward-Colella blast wave with the JWL EOS. (Left) density, (right) pressure.

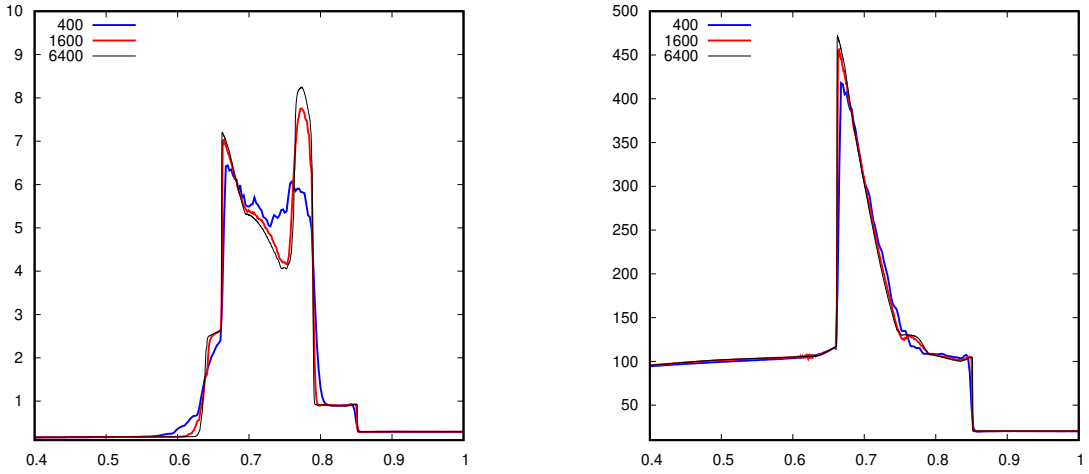

Figure 7.9: Case 2 of the Woodward-Colella blast wave with the JWL EOS. (Left) density, (right) pressure.

by the vertices $(-0.2, 0)$, $(0.1, 1/6)$, and $(0.1, -1/6)$. We enforce Dirichlet conditions on the left boundary, dynamic outflow conditions on the right boundary and slip conditions on the remaining boundaries. Note however that the simulation is stopped before the wave reaches the outflow boundary so one could alternatively enforce Dirichlet on the right boundary. The EOS is van der Waals with $\gamma = 864.7/577.8$, $a = 0.14$ and $b = 3.258 \times 10^{-5}$. For the

interpolation parameters, we use the given b and set $q = p_\infty = 0$.

The ambient state in front of the shock is set to $\boldsymbol{u}_{\mathrm{R}} = (1.225, 0, 0, 101325)^{\mathsf{T}}$. We define the Mach speed of the shock to be $M_S := 1.3$ and then compute the state across the shock using the Rankine-Hugoniot conditions, (1.24). The complete initial state is given by,

$$(\rho_0(\boldsymbol{x}), \boldsymbol{v}_0(\boldsymbol{x}), p_0(\boldsymbol{x})) := \begin{cases} (1.82039, 148.597, 0, 185145), & \text{if } x \leq -0.55, \\ (1.225, 0, 0, 101325), & \text{if } x > -0.55. \end{cases} \tag{7.22}$$

Schlieren plots of the solution can be seen Figure 7.10.

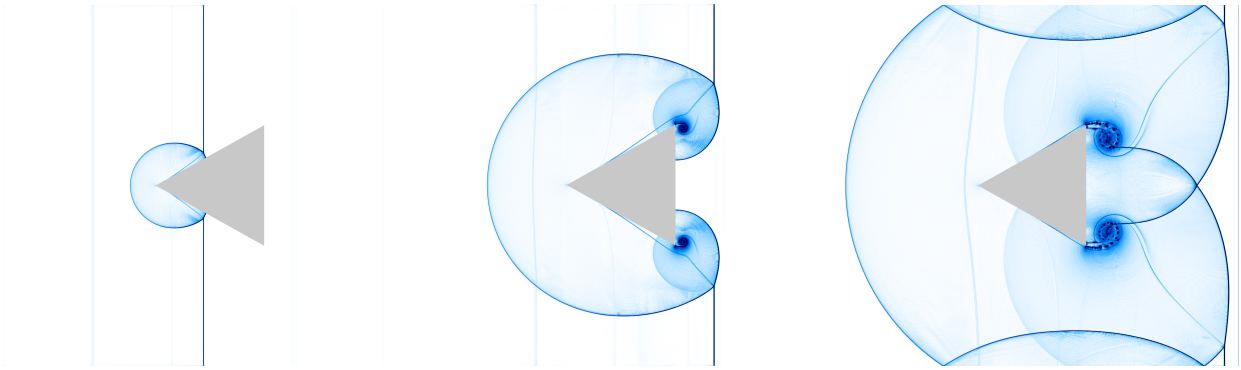

Figure 7.10: Schlieren plot of a shock wave interacting with a triangular obstacle at $t = 1\,\mathrm{ms}$, $1.6\,\mathrm{ms}$, and $2.2\,\mathrm{ms}$. Reprinted with permission from [1].

7.4.4 Shock Bubble Interaction

The next test problem we simulate is the shock-bubble interaction. Much like the Schardin problem done in Section 7.4.3, we again simulate a shock colliding with an obstacle. This time, the obstacle is not a solid boundary but a bubble. The density of the bubble. This problem is normally performed in the context of multi-fluids, where the bubble is a separate material from the ambient fluid. The ambient fluid is typically chosen to be air and the bubble to be helium. More information on the physical experiments can be found

in [87]. We also point out the paper, by Quirk & Karni [88], where they provide comparison between the numerical and physical experiments.

In our case the bubble only differs from the ambient fluid by a difference in density. We use the setup given Wang & Li in [89, Sec. 5.2.2]. The domain is $D = (0,3) \times (0,1)$. The left boundary is Dirichlet, the right boundary is dynamic outflow, and the top and bottom boundaries are slip. Let $\mathfrak{B}$ denote the bubble centered at $(0.5, 0.5)$ with radius 0.25. We use the Jones-Wilkins-Lee EOS with the following parameters,

$$A = 8.545 \times 10^{11},\ B = 2.05 \times 10^{10},\ R_1 = 4.6, R_2 = 1.35,\ \omega = 0.25,\ \rho_0 = 1.84 \times 10^3. \quad (7.23)$$

The primitive states for the ambient fluid and the bubble are,

$$(\rho_{\mathrm{R}}, \boldsymbol{v}_{\mathrm{R}}, p_{\mathrm{R}}) = (1000, 0, 0, 5 \times 10^{10}) \quad (7.24a)$$

$$(\rho_{\mathfrak{B}}, \boldsymbol{v}_{\mathfrak{B}}, p_{\mathfrak{B}}) = (2000, 0, 0, 5 \times 10^{10}) \quad (7.24b)$$

We prescribe the pressure across the shock wave to be $p_{\mathrm{L}} = 4.369 \times 10^{11}$ and determine the remaining variables ρ_{L} and $\boldsymbol{v}_{\mathrm{L}}$ through the Rankine-Hugoniot conditions, (1.24). The primitive initial state is,

$$(\rho_0(\boldsymbol{x}), \boldsymbol{v}_0(\boldsymbol{x}), p_0(\boldsymbol{x})) := \begin{cases} (3778.85, 16867.6, 0, 4.369 \times 10^{11}), & \text{if } x < 0.05, \\ (1000, 0, 0, 5 \times 10^{10}), & \text{if } x \geq 0.5 \text{ and } \boldsymbol{x} \notin \mathfrak{B}, \\ (2000, 0, 0, 5 \times 10^{10}), & \text{if } \boldsymbol{x} \in \mathfrak{B}. \end{cases} \quad (7.25)$$

A diagram of this initial state is shown in Figure 7.11. The simulation is run to final time $t_{\text{final}} = 100\,\mu\text{s}$ with CFL $= 0.5$. Schlieren plots of the density are shown in Figures 7.12, 7.13, and 7.13 at times $t = 40\,\mu\text{s}$, $70\,\mu\text{s}$, and $100\,\mu\text{s}$, respectively. The simulation was run with 50,348,033 $\mathbb{Q}_1$ nodes. Our numerical results match well with experiment results shown in [89].

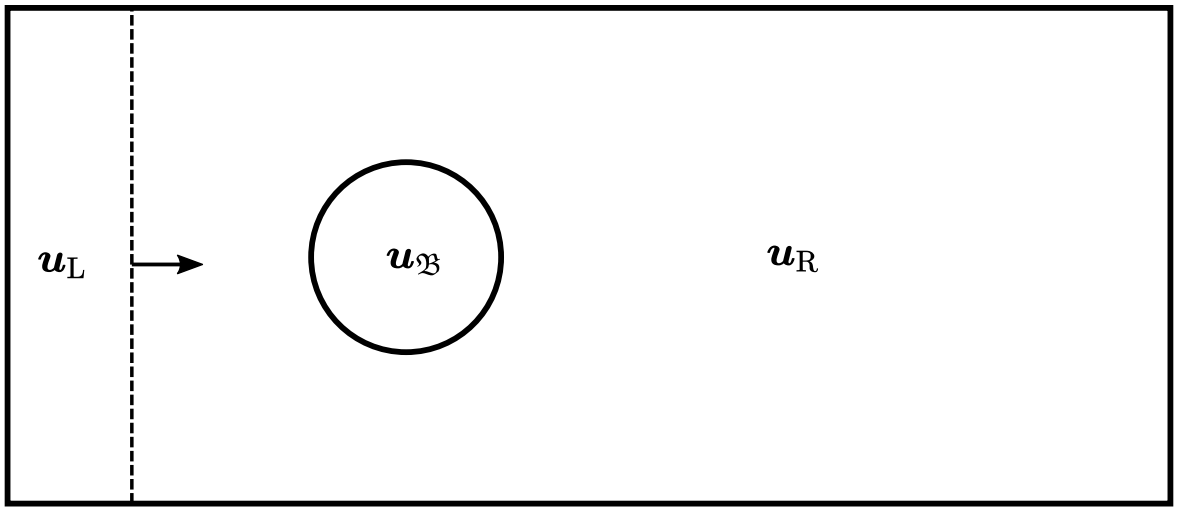

Figure 7.11: Visual description of the initial state for the shock bubble interaction.

7.4.5 Shock Diffraction

This problem involves shock diffracting around a 90° wedge. Physical experiments for a plethora of geometric wedges can be found in [90]. This problem has been studied extensively in the literature for an ideal polytropic gas, see the collection of posters from the 18th International Symposium on Shock Waves (ISSW) [91]. However the results for a non-ideal fluid are much more sparse. Most notably this problem has been simulated for Bethe-Zel'dovich-Thompson (BZT) fluids with the van der Waals equation of state in [92, Problem TD3].

Remark 7.4.1 (BZT Fluids)**.** The Bethe-Zel'dovich-Thompson (BZT) fluids are fluids for which there exists a thermodynamic region where the fundamental derivative,

$$\mathcal{G} := -\frac{\tau}{2}\Big[\Big(\frac{\partial^2 p}{\partial \tau^2}\Big)_s \Big/ \Big(\frac{\partial p}{\partial \tau}\Big)_s\Big] = 1 + \frac{\rho}{a}\Big(\frac{\partial a}{\partial \rho}\Big)_s = \frac{\tau^3}{2a^2}\Big(\frac{\partial^2 p}{\partial \tau^2}\Big)_s, \tag{7.26}$$

is negative. Fluids that exhibit a negative fundamental region behave in an unusual way. For one, the speed of sound increases as the density decreases. This causes the existence of expansion shocks, also called composite waves, where a single wave consists of an expansion

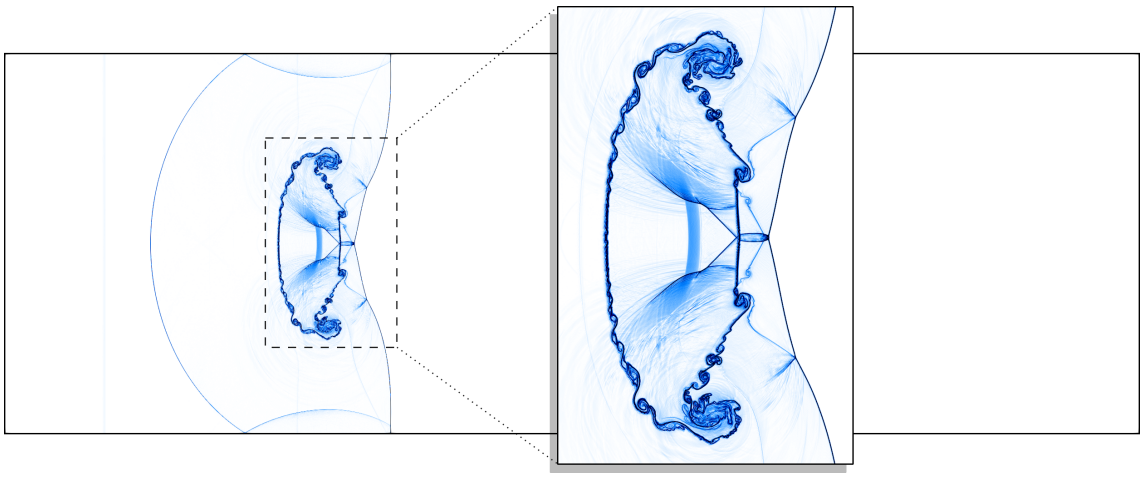

Figure 7.12: Schlieren plots for the shock-bubble interaction benchmark for $t = 40\,\mu$s. Reprinted with permission from [1].

immediately followed by a shock or vice-versa. For more information see [93] and [94]. □

All problems are simulated on the domain $\Omega = [0,2]^2 \backslash ([0,0.5) \times [0,1))$. The first test is run with the van der Waals equation of state where $a = 2.2\,\text{Pa m}^6\,\text{kg}^{-2}$, $b = 7.25 \times 10^{-4}\,\text{m}^3\,\text{kg}^{-1}$, and $\gamma = 1.0125$. We follow the same setup of Brown & Argrow [92, Problem TD3]; however, they use the nondimensionalized forms and we instead work with dimensional variables. It should also be noted that, despite their claim, the left and right states of Problem TD3 do not satisfy the Rankine-Hugoniot conditions. We choose the right state and the left pressure, and derive the remaining state variables through the Rankine-Hugoniot conditions (1.24). The initial condition is,

$$\begin{pmatrix} \rho_0(\boldsymbol{x}) \\ v_0^1(\boldsymbol{x}) \\ v_0^2(\boldsymbol{x}) \\ p_0(\boldsymbol{x})) \end{pmatrix} := \begin{cases} (373.108\,\text{kg/m}^3, 18.4156\,\text{m/s}, 0\,\text{m/s}, 156\,568\,\text{Pa})^\mathsf{T}, & x < 0.5 \\ (128.736\,\text{kg/m}^3, 0\,\text{m/s}, 0\,\text{m/s}, 89\,910.6\,\text{Pa})^\mathsf{T}, & x \geq 0.5. \end{cases} \tag{7.27}$$

Schlieren plots of the solution are shown in Figure 7.15.

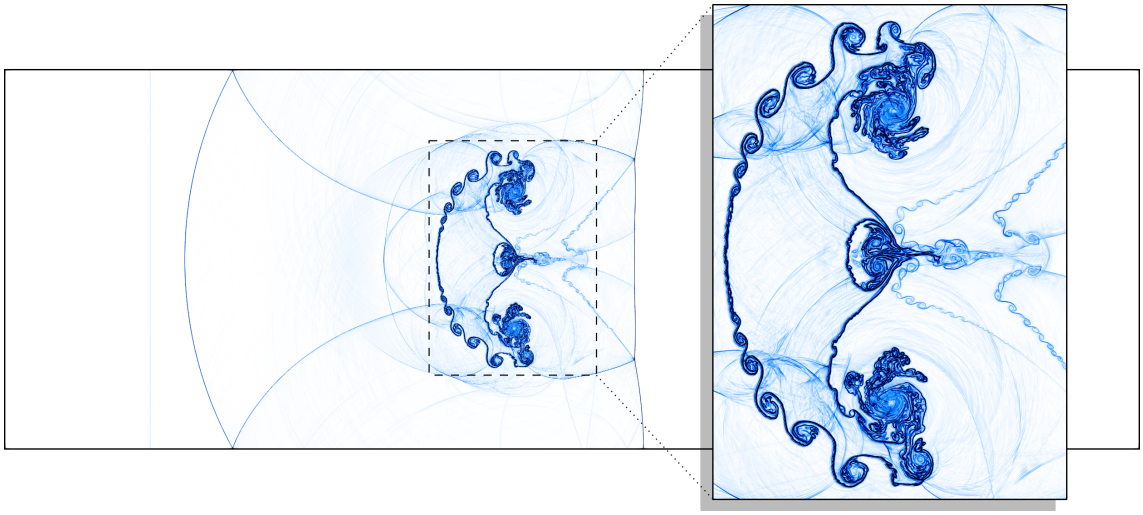

Figure 7.13: Schlieren plots for the shock-bubble interaction benchmark for $t = 70\,\mu$s. Reprinted with permission from [1].

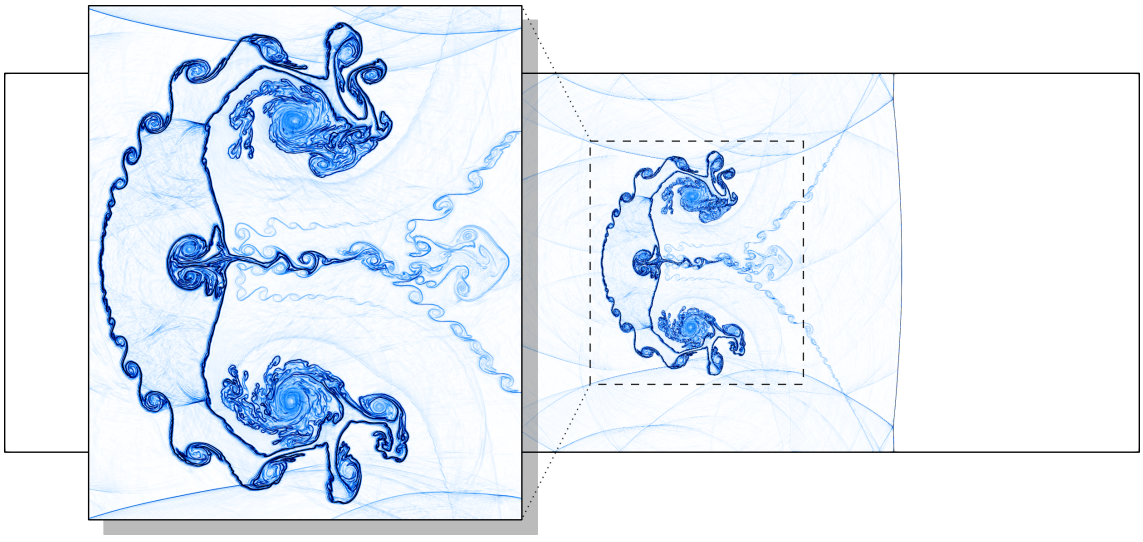

Figure 7.14: Schlieren plots for the shock-bubble interaction benchmark for $t = 100\,\mu$s (bottom). [1].

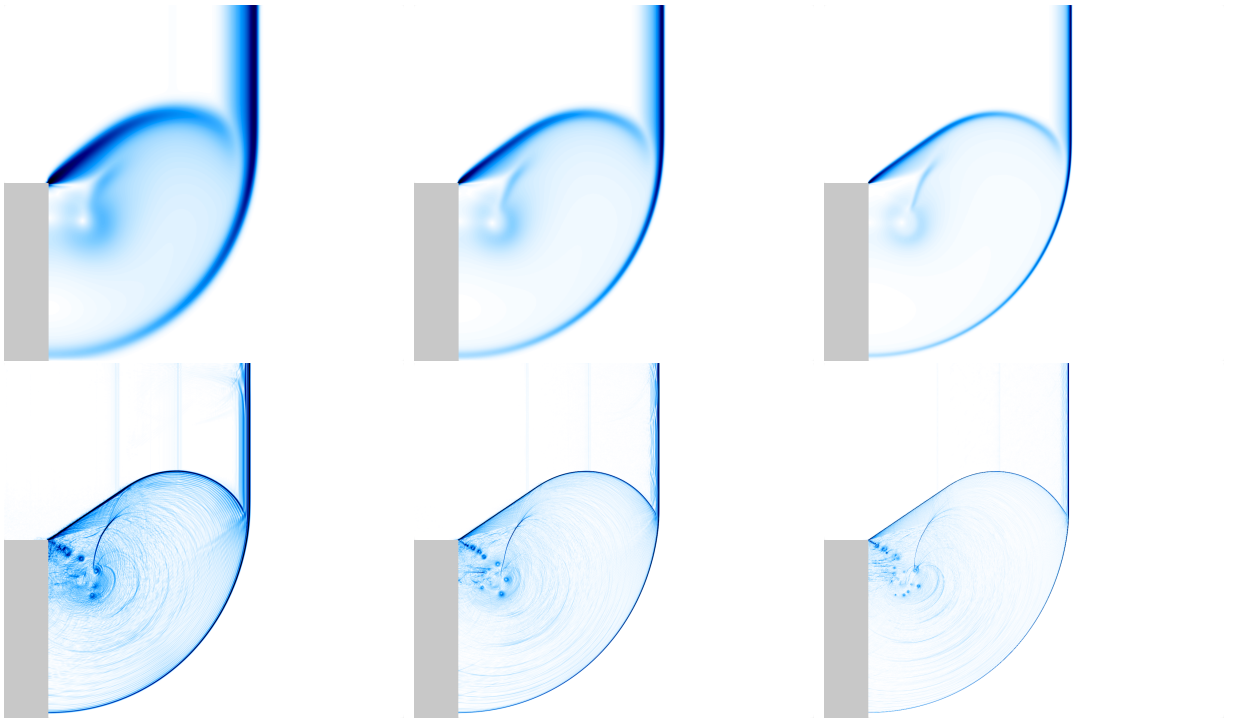

Figure 7.15: Comparison of numerical Schlieren plots for first-order accurate (top row) and second-order accurate (bottom row) solutions at time $t_{\text{final}} = 0.02\,\text{s}$. The mesh resolution increases from left to right as follows: 1,116,289, 4,460,801, and 17,834,497 $\mathbb{Q}_1$-nodes. Reprinted with permission from [1].

REFERENCES

[1] B. Clayton, J.-L. Guermond, M. Maier, B. Popov, and E. J. Tovar, "Robust second-order approximation of the compressible euler equations with an arbitrary equation of state," Journal of Computational Physics, p. 111926, 2023.

[2] B. Clayton, J.-L. Guermond, and B. Popov, "Invariant domain-preserving approximations for the Euler equations with tabulated equation of state," SIAM Journal on Scientific Computing, vol. 44, no. 1, pp. A444–A470, 2022.

[3] J. von Neumann and R. D. Richtmyer, "A method for the numerical calculation of hydrodynamic shocks," Journal of applied physics, vol. 21, no. 3, pp. 232–237, 1950.

[4] P. D. Lax, "Weak solutions of nonlinear hyperbolic equations and their numerical computation," Communications on pure and applied mathematics, vol. 7, no. 1, pp. 159–193, 1954.

[5] P. Lax and B. Wendroff, "Systems of conservation laws," Communications on Pure and Applied Mathematics, vol. 13, no. 2, pp. 217–237, 1960.

[6] S. K. Godunov and I. Bohachevsky, "Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics," Matematičeskij sbornik, vol. 47, no. 3, pp. 271–306, 1959.

[7] R. W. MacCormack, "The effect of viscosity in hypervelocity impact cratering," Journal of spacecraft and rockets, vol. 40, no. 5, pp. 757–763, 2003.

[8] J. L. Steger and R. Warming, "Flux vector splitting of the inviscid gasdynamic equations with application to finite-difference methods," Journal of computational physics, vol. 40, no. 2, pp. 263–293, 1981.

[9] B. Van Leer, "Flux-vector splitting for the euler equations," in Eighth International Conference on Numerical Methods in Fluid Dynamics: Proceedings of the Conference, Rheinisch-Westfälische Technische Hochschule Aachen, Germany, June 28–July 2, 1982, pp. 507–512, Springer, 2005.

[10] P. L. Roe, "Approximate riemann solvers, parameter vectors, and difference schemes," Journal of computational physics, vol. 43, no. 2, pp. 357–372, 1981.

[11] A. Harten, P. D. Lax, and B. V. Leer, "On upstream differencing and godunov-type schemes for hyperbolic conservation laws," SIAM Review, vol. 25, no. 1, pp. 35–61, 1983.

[12] E. F. Toro, M. Spruce, and W. Speares, "Restoration of the contact surface in the hll-riemann solver," Shock waves, vol. 4, pp. 25–34, 1994.

[13] M.-S. Liou, "The evolution of ausm schemes," Defence Science Journal, vol. 60, no. 6, 2010.

[14] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, "Uniformly high order accurate essentially non-oscillatory schemes, iii," Journal of computational physics, vol. 131, no. 1, pp. 3–47, 1997.

[15] G.-S. Jiang and C.-W. Shu, "Efficient implementation of weighted eno schemes," Journal of computational physics, vol. 126, no. 1, pp. 202–228, 1996.

[16] J.-L. Guermond and B. Popov, "Invariant domains and first-order continuous finite element approximation for hyperbolic systems," SIAM Journal on Numerical Analysis, vol. 54, no. 4, pp. 2466–2489, 2016.

[17] P. Colella and H. M. Glaz, "Efficient solution algorithms for the Riemann problem for real gases," J. Comput. Phys., vol. 59, no. 2, pp. 264–289, 1985.

[18] M. J. Ivings, D. M. Causon, and E. F. Toro, "On Riemann solvers for compressible liquids," Internat. J. Numer. Methods Fluids, vol. 28, no. 3, pp. 395–418, 1998.

[19] L. Quartapelle, L. Castelletti, A. Guardone, and G. Quaranta, "Solution of the Riemann problem of classical gasdynamics," J. Comput. Phys., vol. 190, no. 1, pp. 118–140, 2003.

[20] J. K. Dukowicz, "A general, noniterative Riemann solver for Godunov's method," Journal of Computational Physics, vol. 61, no. 1, pp. 119–137, 1985.

[21] P. L. Roe and J. Pike, "Efficient construction and utilisation of approximate riemann solutions," in Proceedings of the Sixth International Symposium on Computing Methods

in Applied Sciences and Engineering, VI, (Netherlands), pp. 499–518, North-Holland Publishing Co., 1985.

[22] J. Pike, "Riemann solvers for perfect and near-perfect gases," AIAA Journal, vol. 31, no. 10, pp. 1801–1808, 1993.

[23] E. F. Toro, C. E. Castro, and B. J. Lee, "A novel numerical flux for the 3d euler equations with general equation of state," Journal of Computational Physics, vol. 303, pp. 80–94, 2015.

[24] K. N. Chueh, C. C. Conley, and J. A. Smoller, "Positively invariant regions for systems of nonlinear diffusion equations," Indiana University Mathematics Journal, vol. 26, no. 2, pp. 373–392, 1977.

[25] D. Hoff, "Invariant regions for systems of conservation laws," Transactions of the American Mathematical Society, vol. 289, no. 2, pp. 591–610, 1985.

[26] H. Frid, "Maps of convex sets and invariant regions for finite-difference systems of conservation laws," Archive for rational mechanics and analysis, vol. 160, pp. 245–269, 2001.

[27] J. Glimm, "Solutions in the large for nonlinear hyperbolic systems of equations," Communications on pure and applied mathematics, vol. 18, no. 4, pp. 697–715, 1965.

[28] L. C. Evans, Partial differential equations, vol. 19. American Mathematical Society, 2022.

[29] S. N. Kružkov, "First order quasilinear equations in several independent variables," Mathematics of the USSR-Sbornik, vol. 10, pp. 217–243, feb 1970.

[30] S. Bianchini and A. Bressan, "Vanishing viscosity solutions of nonlinear hyperbolic systems," Annals of mathematics, pp. 223–342, 2005.

[31] E. Godlewski and P. Raviart, Numerical Approximation of Hyperbolic Systems of Conservation Laws. Applied Mathematical Sciences, Springer New York, 2021.

[32] B. Riemann, Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite, vol. 8. Verlag der Dieterichschen Buchhandlung, 1860.

[33] J. N. Johnson and R. Chéret, Classic papers in shock compression science. Springer Science & Business Media, 2012.

[34] P. D. Lax, "Hyperbolic systems of conservation laws ii," Communications on pure and applied mathematics, vol. 10, no. 4, pp. 537–566, 1957.

[35] C. M. Dafermos, Hyperbolic conservation laws in continuum physics, vol. 3. Springer, 2005.

[36] H. Callen, Thermodynamics and an introduction to thermostatistics, 2nd ed. John Wiley & Sons, New York, 1985. Second edition.

[37] R. Menikoff and B. J. Plohr, "The riemann problem for fluid flow of real materials," Reviews of modern physics, vol. 61, no. 1, p. 75, 1989.

[38] O. Le Métayer and R. Saurel, "The noble-abel stiffened-gas equation of state," Physics of Fluids, vol. 28, no. 4, p. 046102, 2016.

[39] A. Chiapolino and R. Saurel, "Extended noble–abel stiffened-gas equation of state for sub-and-supercritical liquid-gas systems far from the critical point," Fluids, vol. 3, no. 3, p. 48, 2018.

[40] I. A. Johnston, "The noble-abel equation of state: Thermodynamic derivations for ballistics modelling," tech. rep., Defence Science and Technology Organisation Edinburgh (Australia) Weapons ..., 2005.

[41] J. D. Van der Waals, Over de Continuiteit van den Gas-en Vloeistoftoestand, vol. 1. Sijthoff, 1873.

[42] J. O. Valderrama, "The state of the cubic equations of state," Industrial & engineering chemistry research, vol. 42, no. 8, pp. 1603–1618, 2003.

[43] O. Redlich and J. N. Kwong, "On the thermodynamics of solutions. v. an equation of state. fugacities of gaseous solutions.," Chemical reviews, vol. 44, no. 1, pp. 233–244, 1949.

[44] E. L. Lee, H. C. Hornig, and J. W. Kury, "Adiabatic expansion of high explosive detonation products," tech. rep., Univ. of California Radiation Lab. at Livermore, 1968.

[45] S. B. Segletes, "An examination of the jwl equation of state," tech. rep., Army Research Lab Aberdeen Proving Ground, 2018.

[46] R. Menikoff, "Complete mie-gruneisen equation of state (update)," tech. rep., Los Alamos National Lab, 3 2016.

[47] R. Menikoff, "Empirical equations of state for solids," ShockWave Science and Technology Reference Library, pp. 143–188, 2007.

[48] C. Grossmann, H.-G. Roos, and M. Stynes, Numerical treatment of partial differential equations, vol. 154. Springer, 2007.

[49] S. Larsson and V. Thomée, Partial differential equations with numerical methods, vol. 45. Springer, 2003.

[50] P. G. Ciarlet, The Finite Element Method for Elliptic Problems. Society for Industrial and Applied Mathematics, 2002.

[51] A. Ern and J.-L. Guermond, Finite Elements I: Approximation and Interpolation, vol. 72. Springer, 2021.

[52] A. Ern and J.-L. Guermond, Finite Elements II: Galerkin Approximation, Elliptic and Mixed PDEs, vol. 73. Springer, 2021.

[53] A. Ern and J.-L. Guermond, Finite Elements III: First-Order and Time-Dependent PDEs, vol. 74. Springer, 2021.

[54] A. Ern and J.-L. Guermond, "Invariant-domain-preserving high-order time stepping: I. explicit runge–kutta schemes," SIAM Journal on Scientific Computing, vol. 44, no. 5, pp. A3366–A3392, 2022.

[55] J.-L. Guermond, B. Popov, and I. Tomas, "Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems," Computer Methods in Applied Mechanics and Engineering, vol. 347, pp. 143–175, 2019.

[56] E. F. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics. Springer-Verlag Berlin Heidelberg, 2009.

[57] H. Holden and N. H. Risebro, Front tracking for hyperbolic conservation laws, vol. 152.

Springer, 2015.

[58] R. Abgrall and S. Karni, "Computations of compressible multifluids," J. Comput. Phys., vol. 169, no. 2, pp. 594–623, 2001.

[59] J.-L. Guermond and B. Popov, "Fast estimation of the maximum wave speed in the riemann problem for the euler equations," arXiv preprint arXiv:1511.02756, 2015.

[60] J. L. Guermond, M. Nazarov, B. Popov, and I. Tomas, "Second-order invariant domain preserving approximation of the euler equations using convex limiting," SIAM Journal on Scientific Computing, vol. 40, pp. A3211–A3239, 2018.

[61] J.-L. Guermond, B. Popov, and Y. Yang, "The effect of the consistent mass matrix on the maximum-principle for scalar conservation equations," Journal of Scientific Computing, vol. 70, pp. 1358–1366, 2017.

[62] J.-L. Guermond and R. Pasquetti, "A correction technique for the dispersive effects of mass lumping for transport problems," Computer Methods in Applied Mechanics and Engineering, vol. 253, pp. 186–198, 2013.

[63] M. A. Christon, M. J. Martinez, and T. E. Voth, "Generalized Fourier analyses of the advection-diffusion equation-part I: one-dimensional domains," International Journal for Numerical Methods in Fluids, vol. 45, no. 8, pp. 839–887, 2004.

[64] T. Thompson, "A discrete commutator theory for the consistency and phase error analysis of semi-discrete c0 finite element approximations to the linear transport equation," Journal of Computational and Applied Mathematics, vol. 303, pp. 229–248, 2016.

[65] M. Maier and M. Kronbichler, "Efficient parallel 3d computation of the compressible Euler equations with an invariant-domain preserving second-order finite-element scheme," ACM Transactions on Parallel Computing, vol. 8, no. 3, pp. 16:1–30, 2021.

[66] A. Harten, P. D. Lax, C. D. Levermore, and W. J. Morokoff, "Convex entropies and hyperbolicity for general euler equations," SIAM Journal on Numerical Analysis, vol. 35, no. 6, pp. 2117–2127, 1998.

[67] A. Harten, "On the symmetric form of systems of conservation laws with entropy,"

Journal of Computational Physics, vol. 49, no. 1, pp. 151–164, 1983.

[68] J. P. Boris and D. L. Book, "Flux-corrected transport. i. shasta, a fluid transport algorithm that works," Journal of Computational Physics, vol. 11, no. 1, pp. 38–69, 1973.

[69] D. Book, J. Boris, and K. Hain, "Flux-corrected transport ii: Generalizations of the method," Journal of Computational Physics, vol. 18, no. 3, pp. 248–283, 1975.

[70] J. Boris and D. Book, "Flux-corrected transport. iii. minimal-error fct algorithms," Journal of Computational Physics, vol. 20, no. 4, pp. 397–431, 1976.

[71] S. T. Zalesak, "Fully multidimensional flux-corrected transport algorithms for fluids," Journal of computational physics, vol. 31, no. 3, pp. 335–362, 1979.

[72] S. T. Zalesak, The design of Flux-Corrected Transport (FCT) algorithms for structured grids. Springer, 2012.

[73] C. Lohmann and D. Kuzmin, "Synchronized flux limiting for gas dynamics variables," Journal of Computational Physics, vol. 326, pp. 973–990, 2016.

[74] E. Tadmor, "A minimum entropy principle in the gas dynamics equations," Applied Numerical Mathematics, vol. 2, no. 3, pp. 211–219, 1986. Special Issue in Honor of Milt Rose's Sixtieth Birthday.

[75] B. Khobalatte and B. Perthame, "Maximum principle on the entropy and second-order kinetic schemes," Mathematics of Computation, vol. 62, no. 205, pp. 119–131, 1994.

[76] D. Arndt, W. Bangerth, M. Feder, M. Fehling, R. Gassmöller, T. Heister, L. Heltai, M. Kronbichler, M. Maier, P. Munch, J.-P. Pelteret, S. Sticko, B. Turcksin, and D. Wells, "The `deal.II` library, version 9.4," Journal of Numerical Mathematics, vol. 30, no. 3, pp. 231–246, 2022.

[77] M. S. Cramer and R. Sen, "Exact solutions for sonic shocks in van der waals gases," The Physics of fluids, vol. 30, no. 2, pp. 377–385, 1987.

[78] G. Lai, "Interactions of composite waves of the two-dimensional full euler equations for van der waals gases," SIAM Journal on Mathematical Analysis, vol. 50, no. 4, pp. 3535–3597, 2018.

[79] M. Fossati and L. Quartapelle, "The riemann problem for hyperbolic equations under a nonconvex flux with two inflection points," 2014.

[80] P. A. Thompson and K. C. Lambrakis, "Negative shock waves," Journal of Fluid Mechanics, vol. 60, no. 1, p. 187–208, 1973.

[81] S. P. Lyon, "SESAME: The Los Alamos National Laboratory equation of state database," Los Alamos National Laboratory report LA-UR-92-3407, 1992.

[82] K. Thompson, "EOSPAC6." `https://github.com/KineticTheory/eospac6`. Accessed: April 25, 2023.

[83] D. A. Pimentel, "EOSPAC user's manual: Version 6.4." `https://www.osti.gov/biblio/1489917`, 2019.

[84] P. Woodward and P. Colella, "The numerical simulation of two-dimensional fluid flow with strong shocks," Journal of computational physics, vol. 54, no. 1, pp. 115–173, 1984.

[85] P. H. Schardin, "High frequency cinematography in the shock tube," The Journal of Photographic Science, vol. 5, no. 2, pp. 17–19, 1957.

[86] S.-M. Chang and K.-S. Chang, "On the shock–vortex interaction in schardin's problem," Shock Waves, vol. 10, no. 5, pp. 333–343, 2000.

[87] J.-F. Haas and B. Sturtevant, "Interaction of weak shock waves with cylindrical and spherical gas inhomogeneities," Journal of Fluid Mechanics, vol. 181, pp. 41–76, 1987.

[88] J. J. Quirk and S. Karni, "On the dynamics of a shock-bubble interaction," Journal of Fluid Mechanics, vol. 318, pp. 129–163, 1996.

[89] Y. Wang and J. Li, "Stiffened gas approximation and grp resolution for fluid flows of real materials," arXiv preprint, arXiv:2108.13780, 2021.

[90] T. V. Bazhenova, L. G. Gvozdeva, and M. A. Nettleton, "Unsteady interactions of shock waves," Aerospace ScL, vol. 21, pp. 249–331, 1984.

[91] K. Takayama and O. Inoue, "Shock wave diffraction over a 90 degree sharp corner posters presented at 18th issw," Shock Waves, vol. 1, pp. 301–312, 1991.

[92] B. P. Brown and B. M. Argrow, "Nonclassical dense gas flows for simple geometries,"

AIAA Journal, vol. 36, pp. 1842–1847, 1998.

[93] P. A. Thompson, "A fundamental derivative in gasdynamics," Physics of Fluids, vol. 14, pp. 1843–1849, 1971.

[94] P. Colonna, N. Nannan, A. Guardone, and T. Van der Stelt, "On the computation of the fundamental derivative of gas dynamics using equations of state," Fluid phase equilibria, vol. 286, no. 1, pp. 43–54, 2009.

APPENDIX A

Derivation of Exact Solutions for the Euler Equations

We detail the derivation of several exact solutions for the Euler equations.

A.1 Isentropic Vortex*

Theorem A.1.1. *Consider the van der Waals equation of state* (2.15) *with* $a > 0$, $b := 0$, *and* $\gamma := \frac{3}{2}$ *or* $\gamma := 2$. *Let* $\boldsymbol{x}^0 \in \mathbb{R}^2$, $\beta > 0$, $r_0 > 0$. *Let* $\rho_\infty > 0$, $\boldsymbol{v}_\infty \in \mathbb{R}^2$, $p_\infty > 0$ *and assume that*

$$p_\infty > \tfrac{1}{3} a \rho_\infty^2, \qquad a\rho_\infty + \frac{3p_\infty}{\rho_\infty} > \frac{\beta^2 e^1}{8 r_0^2 \pi^2} \tag{A.1}$$

if $\gamma = \frac{3}{2}$. *The following density, velocity, and pressure fields solve the compressible Euler equations* (1.14a)–(1.14c) *with the van der Waals equation of state:*

$$\rho(\boldsymbol{x}, t) := \begin{cases} \left(\frac{3C}{4a} - \frac{1}{2} \sqrt{\frac{9C^2}{4a^2} + \frac{2}{a}\left(F + \frac{1}{2r_0^2} \psi(\overline{\boldsymbol{x}})^2 \right)} \right)^2, & \text{if } \gamma = \frac{3}{2}, \\ \rho_\infty - \frac{\rho_\infty^2}{4 p_\infty r_0^2} \psi(\overline{\boldsymbol{x}}), & \text{if } \gamma = 2, \end{cases} \tag{A.2a}$$

$$\boldsymbol{v}(\boldsymbol{x}, t) := \boldsymbol{v}_\infty + \psi(\overline{\boldsymbol{x}})(-\bar{x}_2, \bar{x}_1)^{\mathsf{T}}, \tag{A.2b}$$

$$p(\boldsymbol{x}, t) := C \rho(\boldsymbol{x}, t)^\gamma - a \rho(\boldsymbol{x}, t)^2, \tag{A.2c}$$

with $\psi(\overline{\boldsymbol{x}}) := \frac{\beta}{2\pi} \exp(\frac{1}{2}(1 - \frac{1}{r_0^2} \|\overline{\boldsymbol{x}}\|_{\ell^2}^2))$, $(\bar{x}_1, \bar{x}_2) = \overline{\boldsymbol{x}} := \boldsymbol{x} - \boldsymbol{x}^0 - \boldsymbol{v}_\infty t$, $C = (p_\infty + a\rho_\infty^2)/\rho_\infty^{3/2}$, *and* $F = -a\rho_\infty - 3p_\infty/\rho_\infty$.

Proof. The derivation of the isentropic vortex begins with the additional assumption that the velocity field is divergence free. That is, $\nabla \cdot \boldsymbol{v} = 0$. Under this assumption the Euler equations take the following simplified form:

This theorem and proof are reprinted with permission from [1].

$$\partial_t \rho(\boldsymbol{x},t) + \boldsymbol{v}(\boldsymbol{x},t)\cdot\nabla\rho(\boldsymbol{x},t) = 0, \qquad \boldsymbol{x}\in\mathbb{R}^2,\, t>0, \tag{A.3}$$

$$\partial_t \boldsymbol{v}(\boldsymbol{x},t) + (\boldsymbol{v}(\boldsymbol{x},t)\cdot\nabla)\boldsymbol{v}(\boldsymbol{x},t) = -\frac{1}{\rho(\boldsymbol{x},t)}\nabla p(\boldsymbol{x},t), \qquad \boldsymbol{x}\in\mathbb{R}^2,\, t>0 \tag{A.4}$$

$$\partial_t e(\boldsymbol{x},t) + \boldsymbol{v}(\boldsymbol{x},t)\cdot\nabla e(\boldsymbol{x},t) = 0, \qquad \boldsymbol{x}\in\mathbb{R}^2,\, t>0, \tag{A.5}$$

with $\boldsymbol{x} := (x_1, x_2)$, boundary conditions, $(\rho_\infty, \boldsymbol{v}_\infty := (v_{1,\infty}, v_{2,\infty})^\mathsf{T}, p_\infty)$ and yet to be determined initial conditions $(\rho_0(\boldsymbol{x}), \boldsymbol{v}_0(\boldsymbol{x}), p_0(\boldsymbol{x}))$. To keep things general, we make no assumption on the equation of state for $p = p(\rho, e)$.

We write the solution as a perturbation of the far-field state; i.e. we define $\boldsymbol{v} := \boldsymbol{v}_\infty + \delta\boldsymbol{v}$ with

$$\delta\boldsymbol{v}(\boldsymbol{x},t) := \begin{pmatrix} \partial_{x_2}\psi(\boldsymbol{x}-\boldsymbol{x}^0-\boldsymbol{v}_\infty t) \\ -\partial_{x_1}\psi(\boldsymbol{x}-\boldsymbol{x}^0-\boldsymbol{v}_\infty t) \end{pmatrix}, \tag{A.6}$$

with the stream function $\psi(\boldsymbol{x}) := \frac{\beta}{2\pi}\exp(\frac{1}{2}(1-\frac{\|\boldsymbol{x}\|_{\ell^2}^2}{r_0^2}))$. Here $\boldsymbol{x}^0 := (x_1^0, x_2^0) \in \mathbb{R}^2$, β, and r_0 are free parameters. To further simplify notation, define $(\bar{x}_1, \bar{x}_2) = \overline{\boldsymbol{x}} := \boldsymbol{x} - \boldsymbol{x}^0 - \boldsymbol{v}_\infty t$ and $r^2 := \|\overline{\boldsymbol{x}}\|_{\ell^2}^2$. Note the following identities which will be used later on:

$$\partial_{x_i}\psi(\overline{\boldsymbol{x}}) = -\frac{\bar{x}_i}{r_0^2}\psi(\overline{\boldsymbol{x}}), \tag{A.7}$$

$$\partial_{x_i x_j}\psi(\overline{\boldsymbol{x}}) = \frac{1}{r_0^2}\Big(-\delta_{ij} + \frac{\bar{x}_i\bar{x}_j}{r_0^2}\Big)\psi(\overline{\boldsymbol{x}}), \tag{A.8}$$

$$\partial_{tx_i}\psi(\overline{\boldsymbol{x}}) = \frac{1}{r_0^2}\Big(v_{i,\infty} - \frac{\bar{x}_i\boldsymbol{v}_\infty\cdot\overline{\boldsymbol{x}}}{r_0^2}\Big)\psi(\overline{\boldsymbol{x}}), \tag{A.9}$$

where δ_{ij} is the Kronecker symbol and $i, j \in \{1, 2\}$.

Using that $\boldsymbol{v} = \boldsymbol{v}_\infty + \delta\boldsymbol{v}$, the left hand side of (A.4) becomes,

$$\partial_t\boldsymbol{v} + \boldsymbol{v}\cdot\nabla\boldsymbol{v} = \partial_t(\delta\boldsymbol{v}) + (\boldsymbol{v}_\infty\cdot\nabla)\delta\boldsymbol{v} + (\delta\boldsymbol{v}\cdot\nabla)\delta\boldsymbol{v}. \tag{A.10}$$

From the definition of $\delta\boldsymbol{v}$ and the identities (A.7), (A.8) and (A.9), we have,

$$(\delta\boldsymbol{v}\cdot\nabla)\delta\boldsymbol{v} = \begin{bmatrix} (\partial_{x_2}\psi)(\partial^2_{x_1x_2}\psi) - (\partial_{x_1}\psi)(\partial^2_{x_2}\psi) \\ -(\partial_{x_2}\psi)(\partial^2_{x_1}\psi) + (\partial_{x_1}\psi)(\partial^2_{x_1x_2}\psi) \end{bmatrix} = -\frac{\overline{\boldsymbol{x}}}{r_0^4}\psi(\overline{\boldsymbol{x}})^2 \tag{A.11}$$

$$(\boldsymbol{v}_\infty\cdot\nabla)\delta\boldsymbol{v} = \frac{1}{r_0^2}\Big(- \begin{bmatrix} v_{2,\infty} \\ -v_{1,\infty} \end{bmatrix} + \frac{\boldsymbol{v}_\infty\cdot\overline{\boldsymbol{x}}}{r_0^2}\begin{bmatrix} \bar{x}_2 \\ -\bar{x}_1 \end{bmatrix}\Big)\psi(\overline{\boldsymbol{x}}) \tag{A.12}$$

$$\partial_t(\delta\boldsymbol{v}) = \frac{1}{r_0^2}\Big(\begin{bmatrix} v_{2,\infty} \\ -v_{1,\infty} \end{bmatrix} - \frac{\boldsymbol{v}_\infty\cdot\overline{\boldsymbol{x}}}{r_0^2}\begin{bmatrix} \bar{x}_2 \\ -\bar{x}_1 \end{bmatrix}\Big)\psi(\overline{\boldsymbol{x}}) \tag{A.13}$$

Thus equation (A.4) becomes $-\frac{\overline{\boldsymbol{x}}}{r_0^4}\psi(\overline{\boldsymbol{x}})^2 = -\frac{1}{\rho}\nabla p$. This identity is furthermore written as,

$$-\frac{1}{2r_0^2}\nabla(\psi(\overline{\boldsymbol{x}})^2) = \frac{1}{\rho(t,\boldsymbol{x})}\nabla p(\rho(t,\boldsymbol{x})). \tag{A.14}$$

Up to this point, we have not made any assumption on the equation of state. We recover the well known isentropic vortex solution if we assume the pressure is given by the ideal gas law; i.e. $p(\rho) = C\rho^\gamma$ for the isentropic flow where $C = p_\infty/\rho_\infty^\gamma$. We now proceed with the van der Waals equation of state. For isentropic flows we have

$$p(\rho) = \frac{C\rho^\gamma}{(1-b\rho)^\gamma} - a\rho^2, \tag{A.15}$$

where C is some constant. (Note, we work with an arbitrary b to keep things general in the beginning.) Following the same process as in the ideal gas case, we compute the indefinite integral, $\int \frac{1}{\rho}\partial_{x_i}p(\rho)\,dx_i$:

$$
\begin{aligned}
-\frac{1}{2r_0^2}\int \partial_{x_i}\psi(\overline{\boldsymbol{x}})^2\,dx_i &= \int \frac{1}{\rho}\partial_{x_i}p(\rho)\,dx_i = \frac{p(\rho)}{\rho} + \int \frac{p(\rho)}{\rho^2}\partial_{x_i}\rho\,dx_i \\
&= \frac{p(\rho)}{\rho} + \int \Big(\frac{C\rho^{\gamma-2}}{(1-b\rho)^\gamma} - a\Big)\rho_{x_i}dx_i = \frac{p(\rho)}{\rho} + \int \frac{\partial}{\partial x_i}\Big[\frac{C}{\gamma-1}\Big(\frac{\rho}{1-b\rho}\Big)^{\gamma-1} - a\rho\Big]\,dx_i \\
&= \frac{C\rho^{\gamma-1}(\gamma-b\rho)}{(\gamma-1)(1-b\rho)^\gamma} - 2a\rho + F.
\end{aligned}
$$

Hence, $\rho(\overline{\boldsymbol{x}})$ can be found by solving the equation,

$$
-\frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2 = \frac{C\rho^{\gamma-1}(\gamma-b\rho)}{(\gamma-1)(1-b\rho)^\gamma} - 2a\rho + F. \tag{A.16}
$$

We have two immediate cases for solutions that can be found explicitly.

Case 1: $\gamma = 3/2$ and $b = 0$: In this case, (A.16) becomes a quadratic equation for $\sqrt{\rho}$,

$$
\rho - \frac{3C}{2a}\sqrt{\rho} - \frac{1}{2a}\Big(F + \frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2\Big) = 0. \tag{A.17}
$$

The constants C and F are determined by applying the far field condition to (A.15) and (A.17):

$$
C = \frac{p_\infty + a\rho_\infty^2}{\rho_\infty^{3/2}} \quad \text{and} \quad F = -a\rho_\infty - \frac{3p_\infty}{\rho_\infty}. \tag{A.18}
$$

However, care must be taken in the choice of p_∞ and ρ_∞ so that the sound speed remains real. Recall that the sound speed for the van der Waals EOS is,

$$
c(\rho,p) = \sqrt{\gamma\frac{p+a\rho^2}{\rho(1-b\rho)} - 2a\rho}. \tag{A.19}
$$

The hypothesis $p_\infty > \frac{1}{3}a\rho_\infty^2$ guarantees that $c(\rho_\infty,p_\infty)^2 > 0$.

The physical root for equation (A.17) is,

$$
\sqrt{\rho} = \tfrac{3C}{4a} - \tfrac{1}{2}\sqrt{\tfrac{9C^2}{4a^2} + \tfrac{2}{a}\big(F + \tfrac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2\big)}. \tag{A.20}
$$

Furthermore, for the root to be real we require that $-F > \frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}})^2$ for all $\overline{\boldsymbol{x}} \in \mathbb{R}^2$. In particular,

$$a\rho_\infty + \frac{3p_\infty}{\rho_\infty} > \frac{\beta^2 e^1}{8r_0^2\pi^2}. \tag{A.21}$$

Lastly, we must justify that the system remains hyperbolic; that is, the sound speed is real for all $(\boldsymbol{x}, t) \in \mathbb{R}^2 \times [0, \infty)$. Since the flow is isentropic, the sound speed for the van der Waals EOS (with $\gamma = 3/2$ and $b = 0$) is, $f(\rho) := c(p(\rho), \rho)^2 = \frac{3}{2}C\sqrt{\rho} - 2a\rho$. Note that $\lim_{\rho\to 0^+} f(\rho) = 0$, $f'(\rho) = \frac{3C}{4\sqrt{\rho}} - 2a$, and $\lim_{\rho\to 0^+} f'(\rho) = \infty$. Therefore, $f(\rho)$ has a maximum at $\rho = \left(\frac{3C}{8a}\right)^2$ and hence $f(\rho) > 0$ for $\rho \in (0, \left(\frac{3C}{8a}\right)^2)$. From the definition of ρ, (A.2a), we see that $0 < \rho < \left(\frac{3C}{4a}\right)^2$. Thus the sound speed is always real.

Case 2: $\gamma = 2$ and $b = 0$: For these choices of parameters, (A.16) becomes,

$$2(C - a)\rho + F + \frac{1}{2r_0^2}\psi(\overline{\boldsymbol{x}}) = 0. \tag{A.22}$$

Using the far field boundary conditions for (A.2c) and (A.22) we find that $C = \frac{p_\infty}{\rho_\infty^2} + a$ and $F = -2p_\infty/\rho_\infty$, respectively. Solving for ρ in (A.22) we have,

$$\rho = \rho_\infty - \frac{\rho_\infty^2}{4p_\infty r_0^2}\psi(\overline{\boldsymbol{x}}). \tag{A.23}$$

Note the sound speed is $c(p(\rho), \rho)^2 = 2(C - a)\rho = \frac{2p_\infty}{\rho_\infty}\rho > 0$. □

APPENDIX B

Numerical Algorithms

B.1 Upper Bound on Maximum Wave Speed

In this section we provide the algorithm which computes an upper bound to the max wave speed. The details are discussed in Section 4.6.

Algorithm 1 Computing $\lambda_{\max}(\widehat{\boldsymbol{p}}^*)$

Require: $\boldsymbol{u}_{\mathrm{L}}$, $\boldsymbol{u}_{\mathrm{R}}$, $\boldsymbol{n}_{\mathrm{LR}}$, p_L, p_R
  compute $\varphi(p_{\min})$, $\varphi(p_{\max})$ from (4.59)
  if $0 \le \varphi(p_{\min})$ **then**
    compute $\widetilde{p}^*$ from (4.78)
    $\widehat{p}^* = \max(p_{\min}, \widetilde{p}^*)$ ▷ One may also set $\widehat{p}^* = 0$
  else if $0 \le \varphi(p_{\max})$ **then**
    if $\gamma_{\min} = \gamma_m$ **then**
      compute $\widetilde{p}_1^*$ and $\widetilde{p}_2^*$ from (4.81a) and (4.81b) resp.
    else
      compute $\widetilde{p}_1^*$ and $\widetilde{p}_2^*$ from (4.83a) and (4.83b) resp.
    end if
    $\widehat{p}^* = \min\{p_{\max}, \widetilde{p}_1^*, \widetilde{p}_1^*\}$
  else
    compute $\widetilde{p}_1^*$ and $\widetilde{p}_2^*$ from (4.85) and (4.89) resp.
    $\widehat{p}^* = \min\{\widetilde{p}_1^*, \widetilde{p}_2^*\}$
  end if
  return $\lambda_{\max}(\widehat{\boldsymbol{p}}^*) := \max\{-\lambda_L^-(\widehat{\boldsymbol{p}}^*), \lambda_R^+(\widehat{\boldsymbol{p}}^*)\}$

B.2 The Quadratic Newton Method for Limiting the Surrogate Entropy

The algorithm for computing the limiting Φ_i^s is given Algorithm 2.

Algorithm 2 Limiting Φ_i^s with the Quadratic Newton Method

Require: $\mathbf{U}_i^{\mathsf{L},n+1}$, $\mathbf{P}_{ij}$, $\ell_j^{i,\rho}$, ϵ

set $\ell_{\mathrm{L}}^0 = 0$ and $\ell_{\mathrm{R}}^0 = \ell_j^{i,\rho}$
compute $f(\ell_{\mathrm{Z}}^0) = \Phi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell_{\mathrm{Z}}^0 \mathbf{P}_{ij})$ for $\mathrm{Z} = \mathrm{L}$ and $\mathrm{Z} = \mathrm{R}$.
compute $f'(\ell_{\mathrm{Z}}^0) = D\Phi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell_{\mathrm{Z}}^0 \mathbf{P}_{ij}) \cdot \mathbf{P}_{ij}$ for $\mathrm{Z} = \mathrm{L}$ and $\mathrm{Z} = \mathrm{R}$.
compute $f[\ell_{\mathrm{L}}^0, \ell_{\mathrm{L}}^0, \ell_{\mathrm{R}}^0]$ and $f[\ell_{\mathrm{L}}^0, \ell_{\mathrm{R}}^0, \ell_{\mathrm{R}}^0]$.
compute $\ell^{\mathrm{L}}(\ell_{\mathrm{L}}^0, \ell_{\mathrm{R}}^0)$ and $\ell^{\mathrm{R}}(\ell_{\mathrm{L}}^0, \ell_{\mathrm{R}}^0)$ from (6.57a) and (6.57b), resp.
set $\ell_{\mathrm{L}} = \min\{\ell^{\mathrm{L}}(\ell_{\mathrm{L}}^0, \ell_{\mathrm{R}}^0), \ell^{\mathrm{R}}(\ell_{\mathrm{L}}^0, \ell_{\mathrm{R}}^0)\}$ and $\ell_{\mathrm{R}} = \max\{\ell^{\mathrm{L}}(\ell_{\mathrm{L}}^0, \ell_{\mathrm{R}}^0), \ell^{\mathrm{R}}(\ell_{\mathrm{L}}^0, \ell_{\mathrm{R}}^0)\}$
while $|\ell_{\mathrm{R}} - \ell_{\mathrm{L}}|/|\ell_{\mathrm{L}}| > \epsilon$ **do**
    compute $f(\ell_{\mathrm{Z}}) = \Phi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell_{\mathrm{Z}} \mathbf{P}_{ij})$ for $\mathrm{Z} = \mathrm{L}$ and $\mathrm{Z} = \mathrm{R}$.
    compute $f'(\ell_{\mathrm{Z}}) = D\Phi_i^s(\mathbf{U}_i^{\mathsf{L},n+1} + \ell_{\mathrm{Z}} \mathbf{P}_{ij}) \cdot \mathbf{P}_{ij}$ for $\mathrm{Z} = \mathrm{L}$ and $\mathrm{Z} = \mathrm{R}$.
    compute $f[\ell_{\mathrm{L}}, \ell_{\mathrm{L}}, \ell_{\mathrm{R}}]$ and $f[\ell_{\mathrm{L}}, \ell_{\mathrm{R}}, \ell_{\mathrm{R}}]$.
    compute $\ell^{\mathrm{L}}(\ell_{\mathrm{L}}, \ell_{\mathrm{R}})$ and $\ell^{\mathrm{R}}(\ell_{\mathrm{L}}, \ell_{\mathrm{R}})$ from (6.57a) and (6.57b), resp.
    set $\ell_{\mathrm{L}} = \min\{\ell^{\mathrm{L}}(\ell_{\mathrm{L}}, \ell_{\mathrm{R}}), \ell^{\mathrm{R}}(\ell_{\mathrm{L}}, \ell_{\mathrm{R}})\}$ and $\ell_{\mathrm{R}} = \max\{\ell^{\mathrm{L}}(\ell_{\mathrm{L}}, \ell_{\mathrm{R}}), \ell^{\mathrm{R}}(\ell_{\mathrm{L}}, \ell_{\mathrm{R}})\}$
end while
set $\ell_j^{i,s} := \min\{\ell_{\mathrm{L}}, \ell_{\mathrm{R}}\}$

B.3 Approximate Godunov-type Solver

There are several methods for handling outflow boundary conditions in the literature. One way to do this is to exactly solve the Riemann problem at $x = 0$, which requires another state on the interface. This method is referred to as the Godunov method, see Godunov [6] which is used in a vast amount of numerical methods for conservation laws. Due to the costly nature of solving the Riemann problem, we instead propose an approximate solution at $x = 0$.

B.3.1 The Method

In order to solve the extended Riemann problem for the Euler equations we must compute p^* by solving $\varphi(p^*) = 0$, see (4.59). The velocity, v^* is then computed by $v^* = v_{\mathrm{L}} - f_{\mathrm{L}}(p^*) = v_{\mathrm{R}} + f_{\mathrm{R}}(p^*)$. We instead propose to use $\widehat{p}^* \geq p^*$ computed using Algorithm 1. Then, we define the velocity in the star domain by $\widehat{v}^*_{\mathrm{L}} = v_{\mathrm{L}} - f_{\mathrm{L}}(\widehat{p}^*)$ and $\widehat{v}^*_{\mathrm{R}} = v_{\mathrm{R}} + f_{\mathrm{R}}(\widehat{p}^*)$, and assuming $\widehat{p}^* \neq p^*$ then $v^*_{\mathrm{L}} \neq v^*_{\mathrm{R}}$. For a state across a shock wave, we compute the density, $\widehat{\rho}_{*\mathrm{L}}$ or $\widehat{\rho}_{*\mathrm{R}}$ from (4.37) or (4.43); respectively, using $\widehat{p}^*$ instead of p^*. In this case, we have that $\widehat{\rho}_{*\mathrm{Z}} \geq \rho_{*\mathrm{Z}}$ for $\mathrm{Z} \in \{\mathrm{L}, \mathrm{R}\}$.

Lemma B.3.1. *If $\widehat{p}^* \geq p^*$, then $\widehat{v}^*_L \leq \widehat{v}^*_R$ where $\widehat{v}^*_L := v_L - f_L(\widehat{p}^*)$ and $\hat{v}^*_R := v_R + f_R(\widehat{p}^*)$.*

Proof. The proof follows immediately from the fact that $\varphi(p)$ is monotonically increasing (see Lemma 4.4.1); that is,

$$\varphi(\widehat{p}^*) = f_{\mathrm{R}}(\widehat{p}^*) + f_{\mathrm{L}}(\widehat{p}^*) + v_{\mathrm{R}} - v_{\mathrm{L}} \geq \varphi(p^*) = 0. \tag{B.1}$$

Hence $\widehat{v}^*_{\mathrm{R}} \geq \widehat{v}^*_{\mathrm{L}}$. □

B.3.2 The Algorithm

The choice of solution at $x = 0$ is best described through diagrams as the exact solution can be seen in the algorithm. See Figure B.1 for several different wave profiles that deter-

mine different solutions at $x = 0$ and Algorithm 3 for the full context. Since we need the approximate solution at $x = 0$, we must look at the wave structure.

Algorithm 3 Approximate Riemann Solution

Require: $\boldsymbol{u}_{\mathrm{L}}$, $\boldsymbol{u}_{\mathrm{R}}$, $\boldsymbol{n}_{\mathrm{LR}}$, p_{L}, p_{R}
compute $\widehat{p}^*$ from Algorithm 1
compute $\varphi(p_{\mathrm{L}})$ and $\varphi(p_{\mathrm{R}})$
if $\varphi(p_{\mathrm{L}}) \geq 0$ and $\varphi(p_{\mathrm{R}}) \geq 0$ **then**
    compute $\lambda_{\mathrm{L}}^- := v_{\mathrm{L}} - a_{\mathrm{L}}$ and $\lambda_{\mathrm{R}}^+ := v_{\mathrm{R}} + a_{\mathrm{R}}$
else if $\varphi(p_{\mathrm{L}}) \geq 0$ and $\varphi(p_{\mathrm{R}}) < 0$ **then**
    compute $\lambda_{\mathrm{L}}^- := v_{\mathrm{L}} - a_{\mathrm{L}}$ and $\lambda_{\mathrm{R}}^+ := \mathcal{S}_{\mathrm{R}}(\widehat{p}^*)$ from (4.41)
else if $\varphi(p_{\mathrm{L}}) < 0$ and $\varphi(p_{\mathrm{R}}) \geq 0$ **then**
    compute $\lambda_{\mathrm{L}}^- := \mathcal{S}_{\mathrm{L}}(\widehat{p}^*)$ and $\lambda_{\mathrm{R}}^+ := v_{\mathrm{R}} + a_{\mathrm{R}}$ from (4.38)
else
    compute $\lambda_{\mathrm{L}}^- := \mathcal{S}_{\mathrm{L}}(\widehat{p}^*)$ and $\lambda_{\mathrm{R}}^+ := \mathcal{S}_{\mathrm{R}}(\widehat{p}^*)$ from (4.38) and (4.41)
end if
▷ Continued on next page

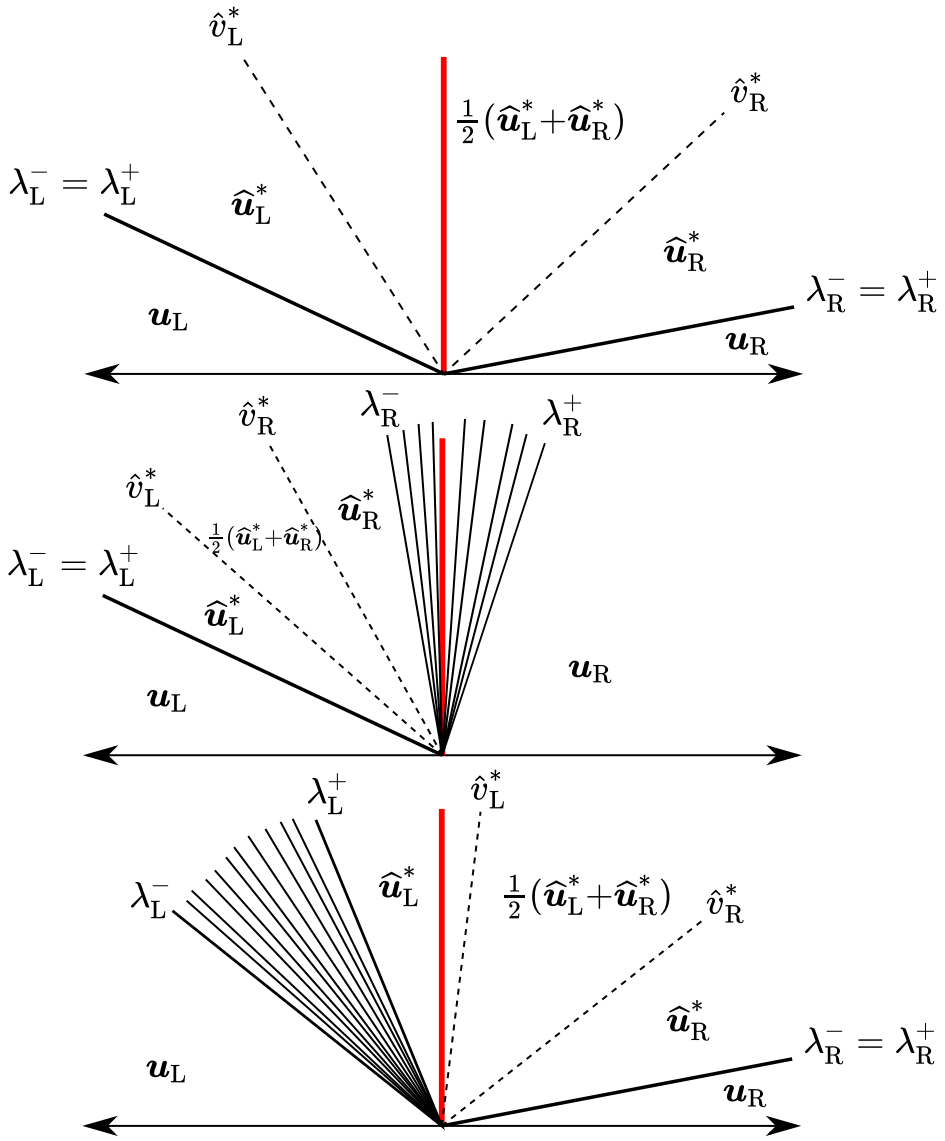

Figure B.1: Figures describing different solutions of $\boldsymbol{u}(0,t)$. (Top): the solution in this case is $\boldsymbol{u}(0,t) = \frac{1}{2}(\widehat{\boldsymbol{u}}^*_{\mathrm{L}} + \widehat{\boldsymbol{u}}^*_{\mathrm{R}})$. (Middle): $\boldsymbol{u}(0,t)$ is the solution on the expansion wave which connects the two states, $\boldsymbol{u}_{\mathrm{R}}$ to $\widehat{\boldsymbol{u}}^*_{\mathrm{R}}$ at $x/t = 0$. (Bottom): $\boldsymbol{u}(0,t) = \widehat{\boldsymbol{u}}^*_{\mathrm{L}}$.

if $\lambda_{\mathrm{L}}^{-} \geq 0$ **then**
    set $\boldsymbol{u} := \boldsymbol{u}_{\mathrm{L}}$
else if $\lambda_{\mathrm{R}}^{+} \leq 0$ **then**
    set $\boldsymbol{u} := \boldsymbol{u}_{\mathrm{R}}$
else
    compute $\widehat{v}_{\mathrm{L}}^{*}$ and $\widehat{v}_{\mathrm{R}}^{*}$ from Lemma B.3.1
    if $\widehat{v}_{\mathrm{L}}^{*} > 0$ **then**
        compute $\widehat{\rho}_{\mathrm{L}}^{*}$ from (4.37)
        if $\varphi(p_{\mathrm{L}}) < 0$ **then**
            set $\boldsymbol{u} := (\widehat{\rho}_{\mathrm{L}}^{*}, \widehat{v}_{\mathrm{L}}^{*}, \boldsymbol{v}_{\mathrm{L}}^{\perp}, \widehat{p}^{*})$
        else
            compute $\widehat{a}_{\mathrm{L}}^{*}$
            if $\widehat{v}_{\mathrm{L}}^{*} - \widehat{a}_{\mathrm{L}}^{*} > 0$ **then**
                set Riemann fan solution
            else
                set $\boldsymbol{u} := (\widehat{\rho}_{\mathrm{L}}^{*}, \widehat{v}_{\mathrm{L}}^{*}, \boldsymbol{v}_{\mathrm{L}}^{\perp}, \widehat{\boldsymbol{p}}^{*})$
            end if
        end if
    else if $\widehat{v}_{\mathrm{R}}^{*} < 0$ **then**
        compute $\widehat{\rho}_{\mathrm{R}}^{*}$ from (4.43)
        if $\varphi(p_{\mathrm{R}}) < 0$ **then**
            set $\boldsymbol{u} := (\widehat{\rho}_{\mathrm{R}}^{*}, \widehat{v}_{\mathrm{R}}^{*}, \boldsymbol{v}_{\mathrm{R}}^{\perp}, \widehat{\boldsymbol{p}}^{*})$
        else
            compute $\widehat{a}_{\mathrm{R}}^{*}$
            if $\widehat{v}_{\mathrm{R}}^{*} + \widehat{a}_{\mathrm{R}}^{*} < 0$ **then**
                set Riemann fan solution
            else
                set $\boldsymbol{u} := (\widehat{\rho}_{\mathrm{R}}^{*}, \widehat{v}_{\mathrm{R}}^{*}, \boldsymbol{v}_{\mathrm{R}}^{\perp}, \widehat{\boldsymbol{p}}^{*})$
            end if
        end if
    else
        if $\widehat{v}_{\mathrm{L}}^{*} + \widehat{v}_{\mathrm{R}}^{*} < 0$ **then**
            set $\boldsymbol{u} := (\frac{1}{2}(\widehat{\rho}_{\mathrm{L}}^{*} + \widehat{\rho}_{\mathrm{R}}^{*}), \frac{1}{2}(\widehat{v}_{\mathrm{L}}^{*} + \widehat{v}_{\mathrm{R}}^{*}), \boldsymbol{v}_{\mathrm{L}}^{\perp}, \widehat{\boldsymbol{p}}^{*})$
        else
            set $\boldsymbol{u} := (\frac{1}{2}(\widehat{\rho}_{\mathrm{L}}^{*} + \widehat{\rho}_{\mathrm{R}}^{*}), \frac{1}{2}(\widehat{v}_{\mathrm{L}}^{*} + \widehat{v}_{\mathrm{R}}^{*}), \boldsymbol{v}_{\mathrm{R}}^{\perp}, \widehat{\boldsymbol{p}}^{*})$
        end if
    end if
end if
return $\boldsymbol{u}$