Julius-Maximilians-Universität Würzburg
Graduate School of Science and Technology

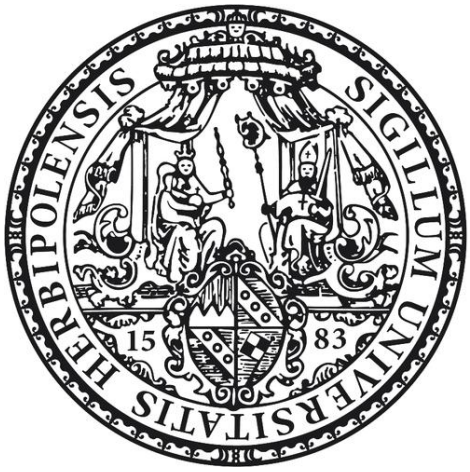
Doctoral thesis / Dissertation

for the doctoral degree / zur Erlangung des Doktorgrads

Doctor rerum naturalium (Dr. rer. nat.)

On Numerical Methods for Astrophysical Applications

Über Numerische Methoden für Astrophysikalische Anwendungen



Submitted by / Vorgelegt von

Markus Zenk

from / aus

Schesslitz

Würzburg, 2017

Submitted on / Eingereicht am: .....

Stamp / Stempel Graduate School

Members of thesis committee / Mitglieder des Promotionskomitees

Chairperson / Vorsitz: .....

1. Reviewer and Examiner/ 1. Gutachter und Prüfer: .....

2. Reviewer and Examiner/ 2. Gutachter und Prüfer: .....

3. Examiner/ 3. Prüfer: .....

Additional Examiners/ Weitere Prüfer: .....

.....

Day of thesis defense/ Tag des Promotionskolloqiums: .....

CURRICULUM VITAE

born on Februar, 07th 1983 in Scheßlitz, single

Education

09/1989 - 07/1994	Elementary School in Scheßlitz
09/1994 - 07/2002	E.T.A.-Hoffmann-Gymnasium in Bamberg graduating with Abitur
04/2004 - 08/2004	Study of Political Science at Otto-Friedrich-University Bamberg
09/2004 - 02/2010	Study of Economathematics at Julius-Maximilians-University Würzburg graduating with Diploma Degree
10/2010 - 08/2013	Study of Mathematics at Julius-Maximilians-University Würzburg graduating with Master Degree
Since 08/2013	PhD Student at Julius-Maximilians-University Würzburg, Mathematical Institute, Chair of Applied Analysis and Fellow at Graduiertenkolleg GRK1147 Theoretical Astrophysics and Particle Physics at Julius-Maximilians-University Würzburg

Bamberg, June 7, 2018 Markus Zenk

Zusammenfassung

Diese Arbeit befasst sich mit der Approximation der Lösungen von Modellen zur Beschreibung des Strömungsverhaltens in Atmosphären. Im Speziellen umfassen die hier behandelten Modelle die kompressiblen Euler Gleichungen der Gasdynamik mit einem Quellterm bezüglich der Gravitation und die Flachwassergleichungen mit einem nicht konstanten Bodenprofil. Verschiedene Methoden wurden bereits entwickelt um die Lösungen dieser Gleichungen zu approximieren. Im Speziellen geht diese Arbeit auf die Approximation von Lösungen nahe des Gleichgewichts und, im Falle der Euler Gleichungen, bei kleinen Mach Zahlen ein. Die meisten numerischen Methoden haben die Eigenschaft, dass die Qualität der Approximation sich mit der Anzahl der Freiheitsgrade verbessert. In der Praxis werden deswegen diese numerischen Methoden auf großen Computern implementiert um eine möglichst hohe Approximationsgüte zu erreichen. Jedoch sind auch manchmal diese großen Maschinen nicht ausreichend, um die gewünschte Qualität zu erreichen. Das Hauptaugenmerk dieser Arbeit ist darauf gerichtet, die Qualität der Approximation bei gleicher Anzahl von Freiheitsgrade zu verbessern.

Diese Arbeit ist im Zusammenhang einer Kollaboration zwischen Prof. Klingenberg des Mathemaitschen Instituts in Würzburg und Prof. Röpke des Astrophysikalischen Instituts in Würzburg entstanden. Das Ziel dieser Kollaboration ist es, Methoden zur Berechnung von stellarer Atmosphären zu entwickeln. In dieser Arbeit werden vor allem zwei Problemstellungen behandelt. Die erste Problemstellung bezieht sich auf die akkurate Approximation des Quellterms, was zu den so genannten well-balanced Schemata führt. Diese erlauben genaue Approximationen von Lösungen nahe des Gleichgewichts. Die zweite Problemstellung bezieht sich auf die Approximation von Strömungen bei kleinen Mach Zahlen. Es ist bekannt, dass Lösungen der kompressiblen Euler Gleichungen zu Lösungen der inkompressiblen Euler Gleichungen konvergieren, wenn die Mach Zahl gegen null geht. Klassische numerische Schemata zeigen ein stark diffusives Verhalten bei kleinen Mach Zahlen. Das hier entwickelte Schema fällt in die Kategorie der asymptotic preserving Schematas, d.h. das numerische Schema ist auf einem diskrete Level kompatibel mit dem auf dem Kontinuum gezeigten verhalten. Zusätzlich wird gezeigt, dass die Diffusion des hier entwickelten Schemas unabhängig von der Mach Zahl ist.

In Kapitel 3 wird ein HLL approximativer Riemann Löser für die Approximation der Lösungen der Flachwassergleichungen mit einem nicht konstanten Bodenprofil angewendet und ein well-balanced Schema entwickelt. Die meisten well-balanced Schemata für die Flachwassergleichungen behandeln nur den Fall eines Fluids im Ruhezustand, die so genannten Lake at Rest Lösungen. Hier wird ein Schema entwickelt, welches sich mit allen Gleichgewichten befasst. Zudem wird eine zweiter Ordnung Methode entwickelt, welche im Gegensatz zu anderen in der Literatur nicht auf einem iterativen Verfahren basiert. Numerische Experimente werden durchgeführt um die Vorteile des neuen Verfahrens zu zeigen.

In Kapitel 4 wird ein Suliciu Relaxations Löser angepasst um die hydrostatischen Gleichgewichte der Euler Gleichungen mit einem Gravitationspotential aufzulösen. Die Gleichungen der hydrostatischen Gleichgewichte sind unterbestimmt und lassen deshalb keine Eindeutigen Lösungen zu. Es wird jedoch gezeigt, dass das neue Schema für eine große Klasse dieser Lösungen die well-balanced Eigenschaft besitzt. Für bestimmte Klassen werden Quadraturformeln zur Approximation des Quellterms entwickelt. Es wird auch gezeigt, dass das Schema robust, d.h. es erhält die Positivität der Masse und Energie, und stabil bezüglich der Entropieungleichung ist. Die numerischen Experimente konzentrieren sich vor allem auf

den Einfluss der Quadraturformeln auf die well-balanced Eigenschaften.

In Kapitel 5 wird ein Suliciu Relaxations Schema angepasst für Simulationen im Bereich kleiner Mach Zahlen. Es wird gezeigt, dass das neue Schema asymptotic preserving und die Diffusion kontrolliert ist. Zudem wird gezeigt, dass das Schema für bestimmte Parameter robust ist. Eine Stabilität wird aus einer Chapman-Enskog Analyse abgeleitet. Resultate numerische Experimente werden gezeigt um die Vorteile des neuen Verfahrens zu zeigen.

In Kapitel 6 werden die Schemata aus den Kapiteln 4 und 5 kombiniert um das Verhalten des numerischen Schemas bei Flüssen mit kleiner Mach Zahl in durch die Gravitation geschichteten Atmosphären zu untersuchen. Es wird gezeigt, dass das Schema well-balanced ist. Die Robustheit und die Stabilität werden analog zu Kapitel 5 behandelt. Auch hier werden numerische Tests durchgeführt. Es zeigt sich, dass das neu entwickelte Schema in der Lage ist, die Dynamiken besser Aufzulösen als vor der Anpassung.

Das Kapitel 7 beschäftigt sich mit der Entwicklung eines multidimensionalen Schemas basierend auf der Suliciu Relaxation. Jedoch ist die Arbeit an diesem Ansatz noch nicht beendet und numerische Resultate können nicht präsentiert werden. Es wird aufgezeigt, wo sich die Schwächen dieses Ansatzes befinden und weiterer Entwicklungsbedarf besteht.

Abstract

This work is concerned with the numerical approximation of solutions to models that are used to describe atmospheric or oceanographic flows. In particular, this work concentrates on the approximation of the Shallow Water equations with bottom topography and the compressible Euler equations with a gravitational potential. Numerous methods have been developed to approximate solutions of these models. Of specific interest here are the approximations of near equilibrium solutions and, in the case of the Euler equations, the low Mach number flow regime. It is inherent in most of the numerical methods that the quality of the approximation increases with the number of degrees of freedom that are used. Therefore, these schemes are often run in parallel on big computers to achieve the best possible approximation. However, even on those big machines, the desired accuracy can not be achieved by the given maximal number of degrees of freedom that these machines allow. The main focus in this work therefore lies in the development of numerical schemes that give better resolution of the resulting dynamics on the same number of degrees of freedom, compared to classical schemes.

This work is the result of a cooperation of Prof. Klingenberg of the Institute of Mathematics in Würzburg and Prof. Röpke of the Astrophysical Institute in Würzburg. The aim of this collaboration is the development of methods to compute stellar atmospheres. Two main challenges are tackled in this work. First, the accurate treatment of source terms in the numerical scheme. This leads to the so called well-balanced schemes. They allow for an accurate approximation of near equilibrium dynamics. The second challenge is the approximation of flows in the low Mach number regime. It is known that the compressible Euler equations tend towards the incompressible Euler equations when the Mach number tends to zero. Classical schemes often show excessive diffusion in that flow regime. The here developed scheme falls into the category of an asymptotic preserving scheme, i.e. the numerical scheme reflects the behavior that is computed on the continuous equations. Moreover, it is shown that the diffusion of the numerical scheme is independent of the Mach number.

In chapter 3, an HLL-type approximate Riemann solver is adapted for simulations of the Shallow Water equations with bottom topography to develop a well-balanced scheme. In the literature, most schemes only tackle the equilibria when the fluid is at rest, the so called Lake at rest solutions. Here a scheme is developed to accurately capture all the equilibria of the Shallow Water equations. Moreover, in contrast to other works, a second order extension is proposed, that does not rely on an iterative scheme inside the reconstruction procedure, leading to a more efficient scheme.

In chapter 4, a Suliciu relaxation scheme is adapted for the resolution of hydrostatic equilibria of the Euler equations with a gravitational potential. The hydrostatic relations are underdetermined and therefore the solutions to that equations are not unique. However, the scheme is shown to be well-balanced for a wide class of hydrostatic equilibria. For specific classes, some quadrature rules are computed to ensure the exact well-balanced property. Moreover, the scheme is shown to be robust, i.e. it preserves the positivity of mass and energy, and stable with respect to the entropy. Numerical results are presented in order to investigate the impact of the different quadrature rules on the well-balanced property.

In chapter 5, a Suliciu relaxation scheme is adapted for the simulations of low Mach number flows. The scheme is shown to be asymptotic preserving and not suffering from excessive diffusion in the low Mach number regime. Moreover, it is shown to be robust

under certain parameter combinations and to be stable from an Chapman-Enskog analysis. Numerical results are presented in order to show the advantages of the new approach.

In chapter 6, the schemes developed in the chapters 4 and 5 are combined in order to investigate the performance of the numerical scheme in the low Mach number regime in a gravitational stratified atmosphere. The scheme is shown the be well-balanced, robust and stable with respect to a Chapman-Enskog analysis. Numerical tests are presented to show the advantage of the newly proposed method over the classical scheme.

In chapter 7, some remarks on an alternative way to tackle multidimensional simulations are presented. However no numerical simulations are performed and it is shown why further research on the suggested approach is necessary.

List of Publications

Desveaux, V.; Zenk, M.; Berthon, C.; Klingenberg, C.: 2014: A well-balanced scheme for the Euler equation with a gravitational potential, J. Fuhrmann et al. (eds.), Finite Volumes for Complex Applications VII - Methods and Theoretical Aspects, Springer Proceedings in Mathematics and Statistics 77

Desveaux, V.; Zenk, M.; Berthon, C.; Klingenberg, C.: 2015: Well balanced schemes to capture non-explicit steady states: Ripa model, Mathematics of Computation, Volume 85, Number 300, July 2016

Desveaux, V. ; Zenk, M. ; Berthon, C. ; Klingenberg, C.: 2016: Well-balanced schemes to capture non-explicit steady states on the Euler equation with a gravity, International Journal for Numerical Methods in Fluids, Volume 81, Issue 2, pp. 104–127, (2016)

Chandrashekar, P. ; Zenk, M.; 2017: Well-Balanced Nodal Discontinuous Galerkin Method for Euler Equations with Gravity, Journal of Scientific Computing, Volume 71, Issue 3, pp. 1062-1093, June 2017

Contents

1 Introduction

The purpose of this chapter is to give a brief overview on the type of equations under consideration. Therefore it is neither complete nor extensive. The following notions, if not mentioned otherwise, can be found in classical textbooks such as [115],[23],[49],[66],[161],[82],[60] and many others.

1.1 Conservation Laws

1.1.1 Existence and uniqueness of solutions

Physical models to describe atmospheric flows can be derived by means of conservation of physical quantities such as mass, momentum and energy. These considerations often lead to hyperbolic partial differential equations (PDE). A hyperbolic PDE may take the following shape

$$u(t,x)_t + \nabla \cdot f(u(t,x)) = 0, \tag{1.1}$$

where $u(t,x) : \mathbb{R} \times \mathbb{R}^n \mapsto \mathbb{R}^m$ gives the vector of conserved quantities and $f = (f_1, ...f_n)^T$ with $\forall_i \;\; f_i : \mathbb{R}^m \mapsto \mathbb{R}^m$ is called the flux function. x denotes the spatial coordinate and t denotes the time. Additionally, the following abbreviations for the partial derivatives are being used: $u_t = \frac{\partial u}{\partial t}$ and $u_{x_i} = \frac{\partial u}{\partial x_i}$ for the partial derivatives with respect to time and the spatial coordinates respectively. In order to call a PDE hyperbolic, certain restrictions on the flux function must be satisfied. They are specified in definition 1.1.1.

Definition 1.1.1. *A system of type (1.1) is called hyperbolic if and only if for all $i \in \{1, ..., n\}$ the matrix $\frac{\partial}{\partial u} f_i(u)$ is diagonalizable with real eigenvalues.*

A critical property of hyperbolic PDEs is that they ensure the conservation of the variables u. This can be seen by integrating (1.1) over a volume $V \in \mathbb{R}^n$ and applying the Gauss theorem on the flux derivative to get

$$\frac{\partial}{\partial t} \int_V u(t,x)dx + \int_{\partial V} \mathbf{n} \cdot f(u(t,x))dx = 0, \tag{1.2}$$

where $\mathbf{n}$ is the outward normal to the boundary of V, i.e. ∂V. Equation (1.2) gives that changes to the volume integral of the conserved quantity are only due to fluxes on the boundary of V, thus ensuring, if the system is closed, i.e. $\int_{\partial V} n \cdot f(u(t,x))dx = 0$, the volume integral of u is invariant in time. Equation (1.2) is also sometimes referred to as the integral form of (1.1). It will be useful to derive the finite volume scheme in section 2.

In practice, the system (1.1) is used to compute the evolution of some initial data. This gives rise to an initial value problem, also called a Cauchy problem, of the type

$$\begin{cases} u(t,x)_t + \nabla \cdot f(u) = 0, \\ u(0,x) = u_0(x). \end{cases} \tag{1.3}$$

When dealing with hyperbolic PDEs it is a classical observation that solutions to (1.3) may develop discontinuities over time, even when the initial data is very smooth, i.e. $u_0 \in \mathbb{C}^\infty$. A classic example here is derived from the inviscid Burgers equation, see [115] and [82] for a detailed analysis. So it may be hard to give meaning to the partial derivatives originally used to describe the evolution of u. To overcome this issue, the concept of weak solutions has been introduced. The idea is that one can get ride of the partial derivatives by putting them via integration by parts onto so called *test functions*, which in turn carry the desired regularity to give a proper meaning to the equation. By multiplying (1.1) with a test function $\phi(t,x) \in \mathbb{C}^1_c([0,T[,V)$ and integrating over a volume V and a time interval $[0,T[$ leads to

$$\int_0^T \int_V \phi_t u + \nabla\phi \cdot f(u) dx dt = -\int_V \phi(0,x)u(0,x)dx, \tag{1.4}$$

which gives rise to the definition 1.1.2.

Definition 1.1.2. *A function $u(t,x)$ is called a weak solution if u satisfies (1.4) for all $\phi(t,x) \in \mathbb{C}^1_c([0,T[,V)$.*

It is not obvious if solutions to (1.4) have anything to do with solutions to (1.3). The derivation of the weak form involves integration by parts which is only true for sufficiently smooth solutions. But the aim is actually to get more control, if the solutions are not smooth. However, for specific systems, the following simplified version of theorem 5.3.1 from [49] connects the concept of weak solutions to the solutions of (1.3).

Theorem 1.1.1 (Weak Strong Uniqueness)**.** *If there exists a solution $u \in \mathbb{C}^1$ to (1.4), then it is the unique weak entropy solution to (1.4) and also a solution to (1.3)*

It is not yet specified what a weak entropy solution is. In general, the entropy is a function of the distribution u and might give additional information or restrictions when computing the evolution of u. The mathematical concept of entropy may be understood as closely related to physical entropy from the second law of thermodynamics that states that the entropy of a closed system is never decreasing and reaches its maximum at its equilibrium state, see for example [102]. While, as will be shown now, the dynamics of a mathematical entropy is just reversed from the physical entropy, the idea of restricting the evolution of a distribution u, or a system, by adding additional information to the system is similar. To shortly review the concept of a mathematical entropy, it is assumed there exists a convex entropy $\psi(u)$ with an associated entropy flux $\Psi(u)$ such that $\psi_u f_u \nabla u = \nabla \cdot \Psi(u)$. The pair (ψ, Ψ) is also called an entropy, entropy-flux pair. In order to derive its dynamics one multiplies (1.1) with ψ_u to get

$$\psi(u)_t + \nabla \cdot \Psi(u) = 0.$$

From this one might conclude that also the entropy is a conserved quantity for the system (1.1). But the previous calculations are only formal, i.e. they are only valid for smooth solutions. A tool often used to determine the dynamics of the entropy is the vanishing viscosity approach, see for example [115] and [49]. Here the the system (1.1) is extended by a parameterized second order term on the right as

$$u_t + \nabla \cdot f(u)_x = \varepsilon \Delta u.$$

Now, again multiplying with ψ_u, integrating this over an arbitrary volume V and a finite time interval $[t_1, t_2]$, after some rearranging, one has

$$\int_{t_1}^{t_2} \int_V \psi_t + \nabla \cdot \Psi(u) dx dt = \int_{t_1}^{t_2} \varepsilon \int_{\partial V} \mathbf{n} \cdot \psi_u \nabla u dx - \varepsilon \int_V \psi_{uu} \sum_{i=1}^{n} u_{x_i}^2 dx dt.$$

To get information about the original equations, one analyses the limit behavior when $\varepsilon \to 0$. The first spatial integral on the right hand side vanishes, since the integrand is bounded, at least if u is smooth on ∂V. Since V is arbitrary, this can be achieved when u only has finitely many discontinuities. The second integral might not be bounded, especially, if u is discontinuous inside V. However, since $u_x^2 > 0$ and ψ is assumed to be convex and therefore $\psi_{uu} > 0$ holds, the term stays positive. Since the volume of integration was arbitrary, it is straightforward to write the differential form of the evolution of the entropy as

$$\psi(u)_t + \nabla \cdot \Psi(u) \leq 0. \tag{1.5}$$

Since one has to deal with discontinuities, a weak form of the last inequality can be derived by the same means as for the conservation law to get

$$\int_0^T \int_V \phi_t \psi + \nabla \phi \cdot \Psi(u) dx dt \leq - \int_V \phi(0, x) \psi(0, x) dx. \tag{1.6}$$

This gives rise to definition 1.1.3.

Definition 1.1.3. *A function $u(t, x)$ is called a weak entropy solution, if u satisfies (1.4) and (1.6) for all $\phi(t, x) \in \mathbb{C}_c^1([0, T[, V)$ and for all convex entropy-entropy flux pairs (ψ, Ψ).*

In general it is an open question if definition 1.1.3 gives a suitable class to search for unique solutions. For the case $m = 1$, Krushkov showed the uniqueness of weak entropy solutions [99]. For the case $n = 1$ and $m > 1$, existence of solutions for specific systems has been shown for example by Glimm [63] or Temple [159]. In the case of the 2-dimensional isentropic Euler equations of gas dynamic, De Lellis and Székelyhidi [113] showed that there exist infinitely many weak entropy solutions. Despite the failure of the definition of weak entropy solutions to generally pick out a unique solution, the concept of weak entropy solutions is still used to design numerical schemes.

1.1.2 Computing Discontinuous Solutions

It has been discussed that discontinuities may appear in the solution of the Cauchy problem (1.3). Now the dynamics of these discontinuities shall be tackled. A way to approach this problem is to analyze the Riemann problem, which consists of a piecewise constant initial condition separated by a discontinuity. For $n = 1$, it is defined as

$$\begin{cases} u_t + f(u)_x = 0, \\ u(0, x) = u_L \text{ if } x < 0, \\ u(0, x) = u_R \text{ if } x > 0. \end{cases} \tag{1.7}$$

Assume for now that there may be a single discontinuity arising from this problem moving with a constant speed s. This means that one is searching for a solution of the following type

$$u(t,x) = \begin{cases} u_L \text{ if } x < st \\ u_R \text{ if } x > st \end{cases} . \tag{1.8}$$

Assume a large enough volume $V = [-dV, dV]$ around 0. Then, from (1.2), one obtains

$$\frac{\partial}{\partial t} \int_V u(t,x)dx = f(u_L) - f(u_R). \tag{1.9}$$

On the other hand, given the assumed structure of the solution (1.8), one can evaluate the integral at a given time t exactly to get

$$\int_V u(t,x)dx = (dV + st)u_L + (dV - st)u_R. \tag{1.10}$$

Differentiating (1.10) with respect to time and using (1.9) one gets that

$$f(u_R) - f(u_L) = s(u_R - u_L). \tag{1.11}$$

The equations (1.11) are the so called Rankine Hugoniot jump conditions. They are a indispensable tool, when dealing with discontinuous solutions.

In the case of $m = 1$, the Riemann problem (1.7) always admits a single shock solution and the shock speed is uniquely determined by the left and right state. However, those solutions might not be entropy weak solutions. If $m > 1$, the left and right side of (1.11) are vectors and therefore the condition is satisfied if and only if the two vectors are linear dependent. If the vectors are linear dependent, one can compute the shock speed as in the case for $m = 1$. If this is not the case, more sophisticated techniques have to be used.

Start to tackle that problem by computing, given a state u_L, which states u_R do satisfy the Rankine-Hugoniot relations. To do this parameterize the state u_R and the shock speed s in (1.7) in the following way

$$\begin{aligned} u_R &= u(\theta, u_L) \text{ with } u(0, u_L) = u_L, \\ s &= s(\theta, u_L, u_R). \end{aligned}$$

Using these in (1.11) and differentiating with respect to θ gives then

$$f_u u(\theta, u_L)_\theta = s(\theta, u_L, u_R)_\theta u(\theta, u_L)_\theta. \tag{1.12}$$

The parameterized curve $u(\theta, u_L)$ is thus tangent to an eigenvector of f_u. The family of points $u(\theta, u_L)$ is also called the Rankine-Hugoniot loci with respect to u_L. However, since f is assumed to be hyperbolic, it admits a basis of m eigenvectors and therefore m different curves of this type exist. In order to find a unique solution, it is worthy taking a step back. It can be seen from (1.12) that the shock speed is somewhat related to the respective eigenvalue to which eigenvector the curve is tangent. Since the system is considered to be hyperbolic, it can be diagonalized to have

$$w_t + diag(\lambda_1(w), ..., \lambda_m(w))w_x = 0,$$

where w are referred to as the characteristic variables and are chosen to be ordered with

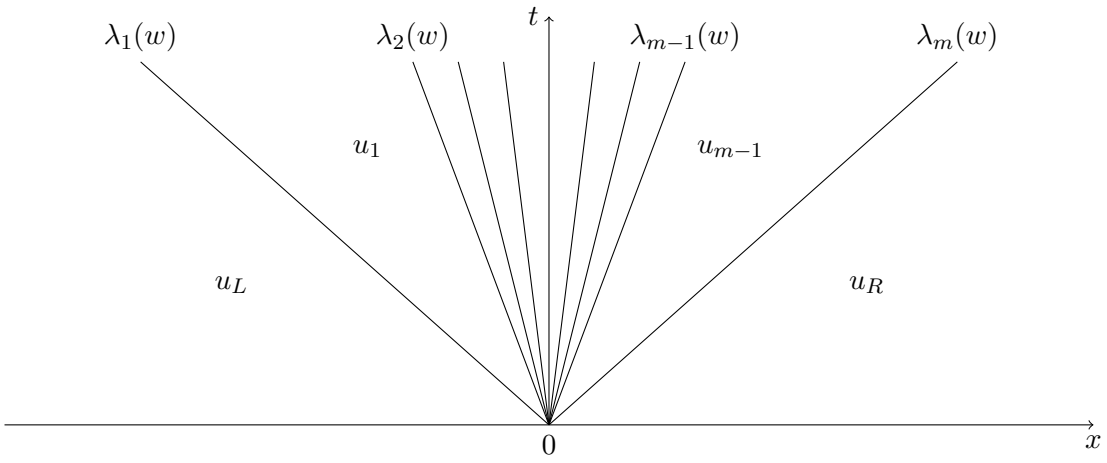


Fig. 1.1: Solution structure to a Riemann problem for a linear system.

respect to their eigenvalues $\lambda_1(w) < .... < \lambda_m(w)$. In this ordering it is assumed, that the eigenvalues are all distinct. Later in this work, there will be systems, where this is not true and eigenvalues will coincide. This is referred to as the resonant case. For now and for the sake of simplicity, resonance is not considered.

If f would be linear, the eigenvalues would not depend on w and one could solve the now m decoupled equations independently. This then leads to a piecewise constant solution with $m + 1$ states, separated by m discontinuities, whose propagation speeds are given by the respective eigenvalue. The states can be ordered from left to right as $u_L, u_1, ..., u_{m-1}, u_R$, see figure 1.1. A straightforward computation shows that u_{i-1} and u_i are connected by a shock curve related to the eigenvector v_i with the respective eigenvalue λ_i.

In the non-linear case, the diagonalized system is not decoupled. The eigenvalues in general depend on w and therefore the equations do not decouple. However, one can learn from the linear case that information in a hyperbolic system is propagated with the speed of the eigenvalues. In order to respect that structure, one computes at u_L the shock curve regarding to the eigenvector with the smallest eigenvalue and the state u_1 must lie on this curve. Then from u_1 one constructs the shock curve with respect to the eigenvector with the second smallest eigenvalue to find suitable locations for u_2 and so on. Denoting the shock curve with respect to the k-th eigenvector originating from u_k as $u(\theta, u_k)_k$, one can state that finding a solution to (1.7) is equivalent to find $(\theta_1, ..., \theta_m)$ such that

$$\begin{cases} u_1 &= u(\theta_1, u_L)_1, \\ u_2 &= u(\theta_2, u_1)_2, \\ \vdots & \\ u_R &= u(\theta_m, u_{m-1})_m. \end{cases} \tag{1.13}$$

It should be remarked that this is a highly non-linear system, where the shock curves u_i might not even be known explicitly. The Riemann problem will be a central building block for the setup of the numerical schemes which are developed in this work, while the idea is

to somehow approximate the exact solution to the Riemann problem to avoid unnecessary computational load.

The existence of solutions to the system (1.13) is discussed and under certain conditions ensured by for example Lax [104] [105] or Smoller [154]. The concept behind the proofs in general rely on the application of the implicit function theorem to ensure the existence of the shock curves where the hyperbolicity of f_u give that the eigenvectors span the whole phase space. However, the application of the implicit function theorem only allows for small variations in the initial data, i.e. $\mid u_L - u_R \mid \leq \varepsilon$. Existence for large data variation may only be shown for specific systems like the isentropic Euler equations, while there are counterexamples where this is not true, see Smoller [153].

Along with the question of existence of solutions to the Riemann problem comes the question of uniqueness. In order to discuss this issue, a description of two different types of shock curves is given.

Definition 1.1.4. *A shock curve $u(\theta, u)_i$ is called genuinely nonlinear, if*

$$\forall u \;\; \nabla\lambda_i(u) \cdot v_i(u) \neq 0 \tag{1.14}$$

and linear degenerate, if

$$\forall u \;\; \nabla\lambda_i(u) \cdot v_i(u) = 0. \tag{1.15}$$

As mentioned before, the entropy plays a critical role in finding solutions to conservation laws. As can be shown, not every discontinuity is admissible due to the entropy condition (1.6). A criterium for admissibility gives the Lax entropy condition.

Definition 1.1.5 (Lax Entropy Condition)**.** *For a genuine non-linear shock curve in the k-th field, a discontinuity is called admissible, if*

$$\lambda(u_{k-1}) > s > \lambda(u_k), \tag{1.16}$$

where s is the speed of the shock.

This gives an additional restriction on which states can be connected by a discontinuity. In order to search for a state that can be connected through a shock wave, one is only allowed to go along the shock curve in the direction of a decreasing eigenvalue. If the connection is along a shock curve in the direction where the eigenvalue increases, all the values along the shock curves are actually realized in the final solution. Those parts are called rarefaction waves and are continuous and non constant parts of the solution the the Riemann problem. It can be shown that across rarefaction waves and linear degenerate discontinuities, which are also called contact discontinuities, entropy is conserved. Entropy will only decrease across shock waves. Equipped with the admissibility condition 1.1.5 it can be shown that for special systems the solution to the Riemann problem exists and is unique, see again Lax [104],[105].

Finally for this section, there should be mentioned another way to solve for the solution to the Riemann problem, which will also be used to derive the numerical schemes.

Definition 1.1.6. *A function $\Phi : \mathbb{R}^m \mapsto \mathbb{R}$ is called a Riemann invariant to the field i, if*

$$\nabla_u \Phi \cdot v_{i,k} = 0, \tag{1.17}$$

for all $v_{i,k}$ such that $f_u v_{i,k} = \lambda_i v_{i,k}$.

Note that in definition 1.1.6, the case of resonance is considered, i.e. there may be more then one eigenvector to a given eigenvalue. Since the shock curves are parallel to the respective eigenvector, the condition (1.17) states that the function Φ is constant along the shock curve.

Along this, the question of how many of these functions Φ can be found for each shock curve arises. Assume, that the multiplicity of the eigenvalue and eigenvector is l. It is straightforward, that one can find $m-l$ vectors w_i, which are orthogonal to $v_{i,k}$. It remains to check weather the given vector fields are integrable, see [23] for a more detailed discussion and the statement of theorem 1.1.2.

Theorem 1.1.2. *To a respective eigenvalue with multiplicity l, there exist at most $m-l$ Riemann invariants, while if $l=1$, then there exist exactly $m-1$ invariants.*

Therefore for the special case of distinct eigenvalues, one has for each field $m-1$ invariants. So for each field one can state $m-1$ equations to determine the relations across the discontinuity. Therefore, to find the solution to the Riemann problem the following system of equations has to be solved.

$$\forall_{k=1}^{m-1}\forall_{i=0}^{m-1}\Phi(u_i)_k = \Phi(u_{i+1})_k. \tag{1.18}$$

These are $(m-1)\times m$ equations, while solving for u_i for $i=1,...,m-1$ give $(m-1)\times m$ unknowns. However, in general (1.18) is again a non-linear system of equations and existence and uniqueness is not obvious from the beginning and have to be proven for the different cases.

1.2 Balance Laws

1.2.1 Physical Source terms

In practice, many physical models which are formulated as conservation laws are extended by external influences. In astrophysical applications they might reach from radiation over chemical reactions up to gravitational acceleration. It is often not possible to include these terms in the flux divergence, so one ends up with a system of the following type

$$u(t,x)_t + \nabla\cdot f(u) = S(u), \tag{1.19}$$

where $u(t,x):\mathbb{R}\times\mathbb{R}^n\mapsto\mathbb{R}^m$ gives the vector of the dependent quantities and $f=(f_1,...f_n)^T$ with $\forall_i\ f_i:\mathbb{R}^m\mapsto\mathbb{R}^m$ is called the flux function and $S:\mathbb{R}^m\mapsto\mathbb{R}^m$ is called the source term.

Consider the case $n=1$ and rewrite system (1.19) with the help of the function $a(x)=x$ by multiply (1.19) with $a(x)_x$ one has

$$\begin{cases} u(t,x)_t + f(u)_x = S(u)a(x)_x, \\ a_t = 0. \end{cases} \tag{1.20}$$

The advantage of this approach now lies in the fact that the system might be rewritten in quasilinear form. Let $\tilde{u}=\begin{pmatrix} u \\ a \end{pmatrix}$, there is

$$\tilde{u}(t,x)_t + \begin{pmatrix} f_u & S(u) \\ 0 & 0 \end{pmatrix} \tilde{u}(t,x)_x = 0. \tag{1.21}$$

Therefore, the source term can be understood as adding another equation and with it a new linear degenerate eigenvalue 0. It can be seen, that hyperbolicity in the sense of definition 1.1.1 is recovered if all the eigenvalues of f_u are non zero. The solution to a Riemann problem now involves dealing with a stationary wave due to the action of the source term. If all other waves are bounded away from 0, the classical techniques of conservation laws can be applied for them. Following [69], across the 0 wave, generalized jump conditions have to be considered using the theory of non-conservative products developed in [128]. The details of this analysis is omitted here.

The existence and uniqueness of solutions to the Cauchy problem for such equations is not so well established as for the conservation laws. One difficulty lies in the various forms the source term can take. Another one lies in the fact that in physical applications, the matrix f_u might admit a 0 eigenvalue. This case is called resonance, since two waves, in this case the one from the source and one from the conservative part, coincide. Two specific issues may arise is this case. First, the geometric multiplicity and the algebraic multiplicity of the eigenvalue might not be the same, and therefore the eigenvectors of the system matrix will not span the whole phase space. Second, even if one has a complete set of eigenvectors, the Rankine Hugoniout loci are now not curves, but hyperplanes in phase space and their intersections are given by curves. It is not clear which state to choose on that curve for a solution and a system dependent analysis is needed. For results on the analysis on those equations the reader is referred to the following publications [120],[127],[163],[2],[112],[122],[84],[85],[3],[65]

Of special interest in this work are the equilibrium solutions to (1.19). Reviewing the second law of thermodynamics, closed systems undergo a change in entropy until they reach their respective equilibrium. Therefore one can expect, that many physical systems are at least close to their equilibrium state. While in the end, the aim is not to compute exactly those equilibria, but time dependent solutions close to those equilibria, it will be crucial to understand the structure of those states. An equilibrium is given as a solution to the balance law (1.19) where the time derivative is set to 0.

Definition 1.2.1. *A distribution $u(t,x)$ is called a steady state for system (1.19), if the following equilibrium condition is satisfied*

$$\nabla \cdot f(u) = S(u). \tag{1.22}$$

In general, those equilibria are therefore again determined by some PDE. The existence and uniqueness of solutions to those equations may also not be obvious. Consider the case of $n = 1$. Then (1.22) can be rewritten in the following form

$$f_u u_x = S(u).$$

Now, if f_u is invertible, the differential equation for u can be written as

$$u_x = f_u^{-1} S(u).$$

However, in general f_u is not invertible. Since f_u is hyperbolic, f_u is not invertible if it admits a 0 eigenvalue. This corresponds to the previous mentioned case of resonance. As it will turn out, the case of resonance will be relevant in section 1.4.

1.2.2 Relaxation Source terms

For the derivation of the numerical schemes, the use of relaxation systems will be crucial. A brief introduction on the structure of those systems shall be given in the following. Also here for brevity, it is assumed that $n = 1$. Following the central publication by Chen, Levermore and Liu [39], relaxation may be found naturally in many physical applications like kinetic theory [33], gases not in thermodynamic equilibrium [98],[166], elasticity with memory [145],[50], multiphase flow and phase transition [144],[64] and linear and nonlinear waves [169]. A relaxation system can be understood as an extension of an underlying PDE to model additional effects not yet captured by the homogeneous model. A relaxation system may take the following shape

$$v_t + g(v)_x = \frac{1}{\varepsilon} R(v), \tag{1.23}$$

where for $M \geq m$ now $v \in \mathbb{R}^M$, $g(v) : \mathbb{R}^M \mapsto \mathbb{R}^M$ and $R(v) : \mathbb{R}^M \mapsto \mathbb{R}^M$. R is called the relaxation source term and $\varepsilon > 0$ is the relaxation parameter. Usually the relaxation source term is structured such that there exists a k dimensional manifold $\mathcal{M} \in \mathbb{R}^M$, where $k < M$, such that $R(v) = 0$ if and only if $v \in \mathcal{M}$. $\mathcal{M}$ is also called the equilibrium manifold of the relaxation system. ε is a parameter which determines the time for the system to reach its equilibrium manifold and is also sometimes called the relaxation time of the system. Assume for now, that ε is small and the time derivative in (1.23) is dominated by the relaxation source term. In order for the manifold $\mathcal{M}$ to be stable under the resulting dynamics it must hold that

$$R_v \mid_{\mathcal{M}} < 0. \tag{1.24}$$

The class of relaxation systems is rich. This work concentrates on a specific type of relaxation system. Consider the following form of system (1.23)

$$\begin{pmatrix} u \\ u_r \end{pmatrix}_t + \begin{pmatrix} f_c(v) \\ f_r(v) \end{pmatrix}_x = \frac{1}{\varepsilon} \begin{pmatrix} 0 \\ r(v) \end{pmatrix}, \tag{1.25}$$

where $u \in \mathbb{R}^m$ as in (1.1) and $u_r \in \mathbb{R}^{M-m}$, $f_c : \mathbb{R}^M \mapsto \mathbb{R}^m$,$f_r : \mathbb{R}^M \mapsto \mathbb{R}^{M-m}$ and $r : \mathbb{R}^M \mapsto \mathbb{R}^{M-m}$. The implicit function theorem gives that for the equilibrium manifold there is $\mathcal{M} \in \mathbb{R}^m$.

As mentioned before, relaxation systems are often used to extend the dynamics of a given PDE to capture additional effects. Therefore it is natural to ask some consistency properties of the relaxation system with respect to the original system.

Definition 1.2.2. *A relaxation system of type (1.25) is called consistent with the system (1.1), if*

$$v \in \mathcal{M} \iff f_c(v) = f(u) \iff r(v) = 0. \tag{1.26}$$

In other words, when the state v is constrained on the equilibrium manifold $\mathcal{M}$, then the dynamics of the system (1.1) are recovered.

Concerning the stability of the relaxation system with respect to the underlying PDE, different criteria are used. A rough criterium is already given in (1.24). It can be considered as a necessary condition, but it can not be sufficient, because it does not incooperate the

flux functions involved. A step towards a more rigorous criterium is performing a Chapman-Enskog analysis of the relaxation system. Since the equilibrium manifold and the relaxation source term may have very complicated structures it is beneficial to make some assumptions on the structure of the relaxation system.

First, it is suitable to define the equilibrium manifold $\mathcal{M}$ as a function of the variables of the reduced system, i.e.

$$\forall v \in \mathcal{M} \;\; \exists Q : \mathbb{R}^m \mapsto \mathbb{R}^{M-m} \;\; \text{s.t.} \;\; Q(u) = u_r.$$

In practice, to find this function Q relates strongly on the choice which parts of the dynamics of the underlying system should be extended. It might be determined by physical concepts or practical reasons. For the following analysis, it is necessary to require some regularity from Q.

The second assumption is on the structure of the source term. Let $r(v)$ be given as

$$r(v) = Q(u) - u_r.$$

This definition immediately satisfies the first stability criterium (1.24) and part of the consistency relation in definition 1.2.2. Now, the first step in the Chapman-Enskog analysis is to expand v in terms of the relaxation parameter ε as

$$v = v_0 + \varepsilon v_1 + \varepsilon^2 v_2 + ...$$

If ε is small, the relaxation system is dominated by the dynamics from the source term and the variable v will tend tend towards the equilibrium manifold $\mathcal{M}$. Therefore the consistency demands to set $v_0 = \begin{pmatrix} u \\ Q(u) \end{pmatrix}$.

Now rewrite the lower part of (1.25) to get

$$u_r = Q(u) - \varepsilon((u_r)_t + f_r(v)_x).$$

Using the expansion and keeping only the first order terms in ε on the right hand side yields

$$u_r = Q(u) - \varepsilon(Q(u)_t + f_r(v_0)_x).$$

Now one multiplies (1.1) by Q_u to get

$$Q(u)_t + Q_u f_u u_x = 0.$$

After further rewriting there is

$$u_r = Q(u) - \varepsilon(f_r(v_0)_x - Q_u f_u u_x).$$

Applying the chain rule for the upper part of the relaxation system (1.25) on the other hand gives

$$u_t + (f_c)_u u_x + (f_c)_{u_r}(u_r)_x = 0.$$

Inserting the expression for u_r then gives

$$u_t + (f_c)_u u_x + (f_c)_{u_r} Q(u)_x = \varepsilon (f_c)_{u_r} (f_r(v_0)_x - Q_u f_u u_x)_x.$$

Since the term $Q(u)$ ensures that the left side is on the equilibrium manifold, the consistency relation from definition 1.2.2 can be used to finally get

$$u_t + f(u)_x = \varepsilon (f_c)_{u_r} (f_r(v_0)_x - Q_u f_u u_x)_x. \tag{1.27}$$

For stability it remains to check if the right hand side gives a stable dissipation.

The here presented version of the Chapman-Enskog analysis might seem unnecessary cumbersome, but as it turns out, all relaxation systems used in this work can be reformulated in the previous described form, so it is worthy, analyzing the approach in this way. It also should be remarked that the right hand side now shares some similarities with the vanishing viscosity approach from the previous section. The vanishing viscosity approach has been successfully used to proof existence and uniqueness of conservation laws. Therefore one might hope that extending a conservation law by the relaxation approach might give some new insights.

The third stability criterium for a relaxation system comes from the consideration of entropy. As in section 1.1, one can start by searching for an entropy, entropy flux pair (ψ, Ψ), such that the relaxation system (1.23) can be reformulated to

$$\psi_t + \Psi_x = \frac{1}{\varepsilon} \psi_v R(v). \tag{1.28}$$

In contrast to the conservation laws, now there is a non-zero right hand side. To have control upon the dissipation of entropy, it is natural to ask the right hand side to be negative. Definition 1.2.3 is given in [39].

Definition 1.2.3. *A twice-differentiable function $\psi : \mathbb{R}^M \mapsto \mathbb{R}$ is said to be an entropy for the system (1.23) provided*

- *$\psi_{v,v} g_v$ is symmetric for all v*
- *$\psi_v R(v) < 0$ for all v*
- *The following are equivalent*
  - *$R(v) = 0$*
  - *$\psi_v R(v) = 0$*
  - *$\exists \bar{\psi}$ s.t. $\psi_v = \bar{\psi}_u$ if $v \in \mathcal{M}$*

The entropy ψ is called convex, if

- *$\psi_{v,v} > 0$ for all v*

The first condition is the Lax entropy condition for conservation laws [105]. It guarantees an existence of an entropy flux Ψ. The second condition can be related to the H theorem of Boltzmann [33]. Equipped with this definition, it is possible to prove the following theorem from [39].

Theorem 1.2.1. *Assume that there exists an entropy by definition 1.2.3 for the system (1.23). Then the local equilibrium approximation*

$$u_t + f_c(u)_x = 0$$

is hyperbolic with the convex entropy entropy-flux pair $(\bar{\psi}(u), \bar{\Psi}(u))$

$$\forall v \in \mathcal{M} \;\; \bar{\psi}(u) = \psi(v) \;\; \textit{and} \;\; \bar{\Psi}(u) = \Psi(v).$$

To put it in other words, if such an entropy for the relaxation system exists, then the dynamics of the relaxation system, if restricted to the equilibrium manifold, are identical with the underlying conservation law. Actually this theorem is even stronger, as it predicts that conservation laws can be derived from such relaxation system. While it is out of the scope of this work to derive conservation laws by the means of relaxation systems, this theorem gives confidence in working with those relaxation systems as approximations to conservation laws.

Several other properties can be shown by assuming the existence of an entropy as given by definition 1.2.3. First, it ensures the positiveness of the diffusion on the right hand side of (1.27), see Theorem 2.2 in [39]. The entropy condition is therefore a stronger stability argument as the Chapman-Enskog analysis. The second property gives a relation between the wave structures of the relaxation system and its equilibrium system.

Theorem 1.2.2 (Interlacing of the eigenvalues [39])**.** *Given the eigenvalues of the relaxation system as* $\Lambda_1 \leq ... \leq \Lambda_M$ *and the eigenvalues of the equilibrium system as* $\lambda_1 \leq ... \leq \lambda_m$ *. Then, if* $v \notin \mathcal{M}$*, then there is*

$$\Lambda_1 < \lambda_1 \leq ... \leq \lambda_m < \Lambda_M \tag{1.29}$$

and if $v \in \mathcal{M}$*, then*

$$\Lambda_1 = \lambda_1 \leq ... \leq \lambda_m = \Lambda_M. \tag{1.30}$$

The inequalities in (1.29) state that the relaxation system propagates information faster than the equilibrium system if not in equilibrium. This is also often referred to as the Whitham- or subcharacteristic-condition. It is therefore a necessary stability criterion, when analyzing relaxation system.

A special case of a relaxation system is the so called Jin-Xin relaxation [88]. It is probably the most widely used relaxation approach for theoretical analysis, as well as for numerical applications. Consider for this again the conservation law of the type

$$u_t + f(u)_x = 0.$$

Multiplying by f_u from the left gives by the chain rule

$$f(u)_t + f_u^2 u_x = 0.$$

A relaxation system can be formed by combining the last two equations, substituting $f(u) = u_r$ and $f_u^2 = c^2$ and adding the relaxation source term to get the following system

$$\begin{pmatrix} u \\ u_r \end{pmatrix}_t + \begin{pmatrix} u_r \\ c^2 u \end{pmatrix}_x = \frac{1}{\varepsilon} \begin{pmatrix} 0 \\ f(u) - u_r \end{pmatrix}. \tag{1.31}$$

In the previous defined terminology one has that $Q(u) = f(u)$, $f_c(u.u_r) = u_r$ and $f_r(u, u_r) = c^2 u$. Therefore the first order approximation to the relaxation system now reads

$$u_t + f(u)_x = \varepsilon(c^2 I - f_u^2)u_{xx}.$$

Stability is therefore assured if $c^2 I > f_u^2$. It is straightforward to see that this relation can be reformulated in terms of the eigenvalues of the flux function f as $c^2 > \lambda_{max}^2$, where λ_{max} is the eigenvalue of f with the largest absolute value. Hence, in this case one can get the subcharacteristic condition directly from the Chapman-Enskog expansion.

The derivation of the relaxation model is only formal and if the solutions of the relaxation model converge to solutions of the underlying PDE is not obvious. If $m = 1$, i.e. for scalar conservation laws, then $M = 2$ and system (1.31) becomes a 2×2 system. In this case, the rigorous investigation of the limit $\varepsilon \to 0$ has been started in [39] and [40] by using compensated compactness techniques, see [158]. Further work on the convergence to the weak solution of the Cauchy problem can for example be found in the work by Natalini [137]. There are also various other contributions to this subject. Here now given is a list, which is neither complete nor exhaustive and the interested reader is referred to these, but also the references therin [126],[138],[121],[103],[45],[174],[36],[100],[119], [91],[90]

1.3 The Inviscid Compressible Euler Equations of Gas Dynamics

The inviscid Euler equations of gas dynamic are a hyperbolic PDE used to describe gas flow. They are a crucial part in models for atmospheres. In 3 space dimensions they are given as

$$\begin{cases} \rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \\ (\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + Ip) = 0, \\ E_t + \nabla \cdot (\mathbf{u}(E + p)) = 0, \end{cases} \tag{1.32}$$

where $\rho(x,t) : \mathbb{R}^3 \times \mathbb{R}^+ \mapsto \mathbb{R}^+$ is the density and $\mathbf{u}(x,t) : \mathbb{R}^3 \times \mathbb{R}^+ \mapsto \mathbb{R}^3$ denotes the velocity of the fluid. $E : \mathbb{R}^3 \times \mathbb{R}^+ \mapsto \mathbb{R}^+$ is the total energy, which is composed of the internal and kinetic energy of the fluid as $E(x,t) = \rho e + \rho \frac{\mathbf{u}^2}{2}$, where $e = e(x,t) : \mathbb{R}^3 \times \mathbb{R} \mapsto \mathbb{R}^+$ is the internal energy density. The system is closed by the pressure p, which is a function of the dependent variables, i.e. $p = p(\rho, E) : \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}^+$.

The system (1.32) is a classic example for conservation laws. It is composed of the conservation of mass ρ, 3 equations for the conservation of the linear momenta $\rho\mathbf{u}$ and the conservation of the total energy E. It can be derived by considerations of the conservation of the previous mentioned quantities under application of Reynolds transport theorem, Newtons laws of motion and the first law of thermodynamics, see for example [58],[115]. Alternatively, one might derive the Euler equations by taking moments of the Boltzmann equations and using Maxwells equation to find the closure of the system, see for example [165].

The eigenvalues λ_i of the flux functions in each direction i can be read as

$$\lambda_i \in \{u_i, u_i + \sqrt{\frac{pp_e - p_\tau}{\rho}}, u_j\} \text{ for } j \neq i, \tag{1.33}$$

where u_i is the $i-th$ component of $\mathbf{u}$. It can be shown, [115], that the eigenvalues u_i are linear degenerate and the others are genuinely nonlinear.

1.3.1 Some Thermodynamic Properties

In general it is assumed that the pressure law satisfies the second law of thermodynamics. Therefore, there exists a specific entropy $\eta(\rho, e) : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$, which satisfies, for some temperature $T(\rho, e) > 0$, the following relation:

$$-Td\eta = de + pd\tau, \tag{1.34}$$

where $\tau = \frac{1}{\rho}$ is called the specific volume. It follows, that the specific entropy satisfies the following equations:

$$\eta(\tau, e)_\tau = -\frac{p(\tau, e)}{T(\tau, e)} < 0 \quad \text{and} \quad \eta(\tau, e)_e = -\frac{1}{T(\tau, e)} < 0. \tag{1.35}$$

In addition, throughout this work, the specify entropy is assumed to be strictly convex. Additionally, to enforce hyperbolicity, a positive acoustic impedance is assumed as

$$\frac{pp_e - p_\tau}{\rho} > 0. \tag{1.36}$$

Due to the hyperbolicity of the system, solutions may become discontinuous after a finite time. As described in section 1.1, an entropy-entropy flux pair has to be chosen to rule out unphysical solutions. Lemma 1.3.1, see also [55], shows that the specific entropy previously defined is the building block for such a pair.

Lemma 1.3.1. *The smooth solutions of (1.32) satisfy the additional conservation laws*

$$(\rho\mathcal{F}(\eta))_t + \nabla \cdot \rho\mathcal{F}(\eta)\mathbf{u} = 0, \tag{1.37}$$

for all smooth functions $\mathcal{F}$.
Moreover, assume

$$\mathcal{F}'(\eta) > 0 \quad \textit{and} \quad \frac{1}{c_p}\mathcal{F}'(\eta) + \mathcal{F}''(\eta) > 0, \tag{1.38}$$

where c_p is the specific heat at constant pressure, defined by

$$c_p = -T\left(\frac{\partial \eta}{\partial T}\right)_p,$$

then $w \mapsto \rho\mathcal{F}(\eta)$ is strictly convex. As a consequence, the pair $(\rho\mathcal{F}(\eta), \rho\mathcal{F}(\eta)\mathbf{u})$ defines a Lax entropy-entropy flux pair for system (1.32). Hence, the weak solutions of (1.32) satisfy in addition:

$$(\rho\mathcal{F}(\eta))_t + \nabla \cdot \rho\mathcal{F}(\eta)\mathbf{u} \leq 0. \tag{1.39}$$

Proof. *First consider smooth solutions of (1.32). From the continuity equation there is*

$$\tau_t + \tau\nabla \cdot \mathbf{u} - \mathbf{u}\nabla\tau = 0, \tag{1.40}$$

and from the equations for momentum and energy there is

$$e_t + \mathbf{u}\nabla e + p\tau\nabla\cdot\mathbf{u} = 0. \tag{1.41}$$

Next, multiplying (1.40) by $-\frac{p}{T}$ *and (1.41) by* $-\frac{1}{T}$ *and using the relations (1.35), it holds that*

$$\partial_\tau\eta\partial_t\tau + \mathbf{u}\partial_\tau\eta\nabla\tau - \tau\partial_\tau\eta\nabla\cdot\mathbf{u} = 0, \tag{1.42}$$

$$\partial_e\eta\partial_t e + \mathbf{u}\partial_e\eta\nabla e + \tau\partial_\tau\eta\nabla\cdot\mathbf{u} = 0. \tag{1.43}$$

The sum of (1.42) and (1.43) easily gives

$$\partial_t\eta + \mathbf{u}\nabla\eta = 0.$$

The result is then achieved by multiplying this relation by $\rho\mathcal{F}'(\eta)$ *and combining it with the continuity equation.*

The establishment of the Lax entropy pair comes from a straightforward study of the Hessian matrix of $\rho\mathcal{F}(\eta)$ *(for instance, see [49, 66, 77, 111] and references therein).*

All the previous calculations only make sense if the dependent variables are in a physical reasonable regime, i.e. $\rho > 0$ and $e > 0$ and therefore there should be defined the set of physical admissible states as

$$\Omega_{Phys} = \{(\rho, \rho\mathbf{u}, E) \in \mathbb{R}^5; \rho > 0, e > 0\}. \tag{1.44}$$

There are different closures for the system (1.32). The most widely known is the ideal gas law, where

$$p = \rho RT, \tag{1.45}$$

and R is the gas constant.

For a polytropic gas, this relation can be rewritten in terms of the internal energy as

$$p = \rho(\gamma - 1)e, \tag{1.46}$$

where γ is the polytropic index. It depends on the ratio of specific heats

$$\gamma = \frac{c_p}{c_v}, \tag{1.47}$$

and c_v is the specific heat at constant temperature. The entropy for an polytropic gas takes therefore the following the form

$$s = \log(\frac{p}{\rho^\gamma}). \tag{1.48}$$

Equation (1.45) is also called an equation of state (EOS). EOSs are used to describe the properties of the fluid through thermodynamical principles. They often depend on the state and composition of the fluid under consideration. The ideal gas law, often used for its simplicity, yet often falls short to capture the properties of fluids in more realistic scenarios. For other more sophisticated EOS especially in the context of astrophysical applications see [131],[59],[160] and references therein.

1.3.2 The Incompressible Limit

An important feature of the system (1.32) can be derived by analyzing the equations at low Mach number. The Mach number is defined as

$$M = \frac{\|\mathbf{u}\|}{\bar{c}}, \tag{1.49}$$

where $\bar{c}$ is the speed of sound. To analyze the behavior of solutions in this regime, the equations first are rewritten in a non-dimentionalized form, see also [9]. To this end, the dependent and independent variables k are rescaled, such that $k = \frac{\hat{k}}{k_{ref}}$, where $\mathbf{u}_{ref} = \frac{x_{ref}}{t_{ref}}$. Rewriting the equations in terms of the non-dimentionalized values $\hat{k}$ and dropping the hats for convenience, the following set of equations is obtained

$$\begin{cases} \rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \\ (\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + I \frac{p}{M^2}) = 0, \\ E_t + \nabla \cdot (\mathbf{u}(E+p)) = 0, \end{cases} \tag{1.50}$$

employed with the definition for the total energy

$$E = \rho e + M^2 \rho \frac{\mathbf{u}^2}{2}. \tag{1.51}$$

Now, one is interested in the limit when $M \to 0$. If one does not consider the energy equation in (1.50), it has been shown in the pioneering work by Klainermann and Majda [93] that the compressible equations tend, under suitable boundary conditions, to its incompressible counterpart, given as

$$\begin{cases} \rho = const, \\ \mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla \bar{p} = 0, \\ \nabla \cdot \mathbf{u} = 0. \end{cases} \tag{1.52}$$

This result is even more astonishing when the role of the pressure is analyzed. Taking the divergence of the velocity equations yields

$$\Delta \bar{p} = -\nabla \cdot (\mathbf{u} \cdot \nabla \mathbf{u}).$$

In contrast to the compressible equations, where the pressure depends locally on the conserved quantities, now the pressure satisfies an elliptic equation depending on the velocity field $\mathbf{u}$. Also other properties like the conservation of internal and kinetic energy can be derived [132]. For the full system given in (1.50), only formal derivations of the incompressible limit behavior exist, see for example [52],[74]. However, these derivations will be crucial to analyze the behavior of the numerical schemes at low Mach numbers. Therefore a short review is presented here.

Consider the system (1.50) and expand the dependent variables in terms of the Mach number

$$\rho = \sum_{i=0}^{\infty} M^i \rho_i \qquad \mathbf{u} = \sum_{i=0}^{\infty} M^i \mathbf{u}_i \qquad p = \sum_{i=0}^{\infty} M^i p_i \qquad e = \sum_{i=0}^{\infty} M^i e_i.$$

When plugged back into system (1.50) and only relations of the same order of Mach number are considered, the pressure satisfies the following relations

$$p = p_0 + M^2 p_2 \qquad \nabla p_0 = \nabla p_1 = 0. \tag{1.53}$$

Under the assumption of for example open boundary conditions, further computations yield

$$\rho = \rho_0 + M\rho_1 \qquad \mathbf{u} = \mathbf{u}_0 + M\mathbf{u}_1 \qquad e = e_0 + Me_1, \tag{1.54}$$

$$\nabla \rho_0 = 0 \qquad \nabla \cdot \mathbf{u}_0 = 0. \tag{1.55}$$

One can understand the derived scalings as necessary conditions to reach the incompressible limit. Analogous to the case of physical admissible states, a set of asymptotic preserving states can be defined as

$$\Omega_{AP} = \{(\rho, (\rho\mathbf{u}), E) \in \mathbb{R}^5; \nabla p_0 = \nabla p_1 = 0, \nabla \rho_0 = 0, \nabla \cdot \mathbf{u}_0 = 0\}. \tag{1.56}$$

The name asymptotic preserving is to be understood in the sense of being compatible with the limit behavior of the system (1.49). Combining the definitions of (1.44) and (1.56), the following set is defined

$$\Omega = \Omega_{Phys} \cap \Omega_{AP}. \tag{1.57}$$

One of the aims of this work is to compute numerical approximations, that respect the physical admissibility as well as the asymptotic scalings of the dependent variables. Or in other words, (1.57) is an invariant set also for the numerical scheme.

1.4 The Euler Equations with a Gravitational Potential

When dealing with atmospheres, the Euler equations (1.32) are equipped with a source term due to the gravitational acceleration of the fluid in the gravitational field of the astrophysical object. The equations then read

$$\begin{cases} \rho_t + \nabla \cdot (\rho\mathbf{u}) = 0, \\ (\rho\mathbf{u})_t + \nabla \cdot (\rho\mathbf{u} \otimes \mathbf{u} + Ip) = -\rho\nabla\Phi, \\ E_t + \nabla \cdot (\mathbf{u}(E+p)) = -\rho\langle u, \nabla\Phi\rangle, \\ \Phi_t = 0, \end{cases} \tag{1.58}$$

where $\Phi(x) : \mathbb{R}^m \mapsto \mathbb{R}^m$ is the gravitational potential and is throughout this work assumed

as given. An entropy-entropy flux pair for this system can easily be found when observing that in the proof of Lemma (1.3.1) the source term does not play any role. Therefore the system (1.58) admits the same entropy as the system (1.32). Moreover, the eigenvalues λ_i of the system in each direction i can be read as

$$\lambda_i \in \{0, u_i, u_i + \sqrt{\frac{pp_e - p_\tau}{\rho}}, u_j\} \text{ for } j \neq i, \tag{1.59}$$

where the additional 0 eigenvalue is due to the source term.

1.4.1 Equilibrium Solutions

Most astrophysical objects spend most of their time close to a steady state. Also the atmosphere of the earth is a system somehow close to an equilibrium state. It is therefore interesting to investigate the equilibrium solutions of (1.58). In order to compute the equilibria, one sets the time derivatives in (1.58) to 0 and gets the following PDE

$$\begin{cases} \nabla \cdot (\rho \mathbf{u}) = 0, \\ \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + Ip) = -\rho \nabla \Phi, \\ \nabla \cdot (\mathbf{u}(E + p)) = -\rho \langle u, \nabla \Phi \rangle. \end{cases} \tag{1.60}$$

An important sub-class of solutions to (1.60) are the hydrostatic equilibria. These are equilibria where the fluid is at rest, i.e. $\mathbf{u} = 0$. When using this relation, one is left with the following PDE

$$\nabla p = -\rho \nabla \Phi. \tag{1.61}$$

Depending on the EOS, in general the system (1.61) is underdetermined. In general, the pressure p is a function of density and temperature, i.e. $p = p(\rho, T)$. However, using the chain rule, one can rewrite the system as

$$p_\rho \nabla \rho + p_T \nabla T = -\rho \nabla \Phi.$$

This is now a system of m equations, while there are $2m$ unknowns, i.e. $\nabla \rho$ and ∇T. Additional assumptions are needed to compute solutions to (1.61). Presented here are 2 different assumptions, that lead to explicit solutions for a hydrostatic equilibrium

- Isothermal Atmosphere for an ideal Gas Law: $p = \rho RT$

$$\begin{cases} T(x) = const, \\ \rho(x) = \rho_0 \exp(-\frac{\Phi(x)}{RT}), \\ p(x) = RT\rho_0 \exp(-\frac{\Phi(x)}{RT}), \end{cases} \tag{1.62}$$

- Polytropic Atmosphere $p = K\rho^\Gamma$ for $\Gamma \in (0,1) \cup (1, \infty)$

$$\begin{cases} \rho(x) = \left(\frac{\Gamma - 1}{\Gamma K}(C - \Phi(x))\right)^{\frac{1}{\Gamma - 1}}, \\ p(x) = K^{\frac{1}{1-\Gamma}} \left(\frac{\Gamma - 1}{\Gamma}(C - \Phi(x))\right)^{\frac{\Gamma}{\Gamma - 1}}, \end{cases} \tag{1.63}$$

where C is just a constant of integration. If an ideal gas law is assumed, and $\Gamma = \gamma$, then the polytropic atmosphere coincides with the isentropic atmosphere. Another way to compute an isentropic atmosphere is given in [89]. Consider the thermodynamic relation (1.34)

$$-Td\eta = de + pd\tau.$$

rewriting leads to

$$dh = d(e + \tau p) = -Td\eta + \tau dp,$$

where h is the specific enthalpy. The isentropic assumption gives $d\eta = 0$ and one can use the last relation in (1.61) to get

$$h + \phi = const. \tag{1.64}$$

Other solutions to the system (1.61) exist, see for example [131],[59],[35] and references therein. One of the main difficulties in computing accurate approximations to the system (1.58) is the lack of knowledge about a general solution to the system (1.61).

As a remark to the computation of the hydrostatic equilibria, a different approach is presented here. Consider the case of one space dimension, following section 1.2.1, the equilibrium equations can be rewritten as

$$f(u)_u u_x = S(u). \tag{1.65}$$

Now, in oder to give an explicit expression of the derivatives, one would like to multiply from the left by f_u^{-1}. Since f is hyperbolic, one has to check the eigenvalues of f to know if the inverse exists. Assuming a polytropic gas law, the eigenvalues λ_i of f are $\lambda_i \in \{u, u \pm \sqrt{\gamma \frac{p}{\rho}}\}$. In fact, the hydrostatic regime is just the case where f is not invertible and one has to deal with the previous mentioned case of resonance. Apart from that, an interesting feature of those general equilibria in one space dimension might be presented. Assume now that u is such that f is invertible. By multiplying f_u^{-1} from the left and some rearranging, the following set of differential equations can be obtained

$$\begin{cases} \rho_x = \rho \frac{\rho \Phi_x}{\rho u^2 - \gamma p}, \\ u_x = -u \frac{\rho \Phi_x}{\rho u^2 - \gamma p}, \\ p_x = \gamma p \frac{\rho \Phi_x}{\rho u^2 - \gamma p}. \end{cases} \tag{1.66}$$

Those equations can be further simplified. From the first equation in (1.60) it can be seen that the velocity is inversely proportional to the density, i.e. $u = \frac{\alpha}{\rho}$. A second property involves the entropy in those equilibria. One can compute the derivative of the entropy to get

$$s_x = \log(\frac{p}{\rho^\gamma})_x = \frac{p_x \rho - \gamma p \rho_x}{\rho p} = 0, \tag{1.67}$$

where in the second step the first and third equation from (1.66) are used. Therefore, the equilibria given by (1.66) are isentropic and the pressure can be related to the density as $p = \beta \rho^\gamma$. Using these relations, one can rewrite the the system (1.66) in to one equation for the density to have

$$\rho_x = \frac{\rho^3 \Phi_x}{\alpha^2 - \gamma\beta\rho^{\gamma+1}}. \tag{1.68}$$

The existence of solutions to this differential strongly depends on the gravitational potential Φ. A critical value for ρ is, when the denominator on the right hand side is 0, i.e. $\bar{\rho} = (\frac{\alpha^2}{\gamma\beta})^{\frac{1}{\gamma+1}}$. By using the definitions for α and β, $\bar{\rho}$ just refers to the sonic point, i.e. $u = \pm\sqrt{\gamma\frac{p}{\rho}}$. Of special interest is, if the solutions tend towards this critical value. It is easy to see that

$$\text{If } \Phi_x > 0 \text{ and } \begin{cases} \rho > \bar{\rho} \text{ then } \rho_x < 0, \\ \rho < \bar{\rho} \text{ then } \rho_x > 0, \end{cases} \text{, and if } \Phi_x < 0 \text{ and } \begin{cases} \rho > \bar{\rho} \text{ then } \rho_x > 0, \\ \rho < \bar{\rho} \text{ then } \rho_x < 0. \end{cases} \tag{1.69}$$

So in the case of $\Phi_x > 0$, the solutions tend towards its critical values $\bar{\rho}$. Away from the critical values $\bar{\rho}$, the solutions to the differential equation (1.68) exists and are unique by the Picard-Lindelöf theorem. It should also be obvious that the interval on which the uniqueness is guaranteed grows, when α^2 is getting smaller. So one can analyze how the solution will behave, when $\alpha \to 0$. To this end, consider this limit in (1.68) to get

$$\rho_x = -\frac{\rho^{2-\gamma}\Phi_x}{\gamma\beta}, \tag{1.70}$$

which gives back the polytropic equilibrium as in (1.63) and especially now the isentropic equilibrium. The non-uniqueness of the hydrostatic equilibrium therefore comes from the fact, that f_u is singular at $u = 0$ and therefore, assume given a solution $\bar{u}_x$ to the system $f_u u_x = S(u)$, there exist infinitely many solutions which can be expressed as $\bar{u}_x + v$, where $v \in \ker f_u$. The kernel of f_u for $u = 0$ is simply the eigenvector of f_u to the eigenvalue u, which in primitive variables reads $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. Therefore the hydrostatic equilibrium equations may be recast in the following form

$$\begin{cases} \rho_x = -\frac{\rho^2\Phi_x}{\gamma p} + \rho\delta, \\ p_x = -\rho\Phi_x, \end{cases} \tag{1.71}$$

where $\delta = \delta(x)$ is some arbitrary function coming from the parameterization of the kernel of f_u. A last observation can be made in the connection between δ and the distribution of entropy in the atmosphere. To this end, insert the equations (1.71) into the derivative for the entropy given in (1.67) to get

$$s_x = -\gamma\delta. \tag{1.72}$$

Now it can be shown, that the first equation in (1.71) is a consequence of the second equation. Use the formulation of the entropy in (1.48) and rewrite for the pressure to get

$$p = \rho^\gamma \exp(s).$$

Taking the derivative with respect to x then gives

$$p_x = \frac{\gamma p}{\rho}(\rho_x + \frac{\rho}{\gamma} s_x).$$

Using this in the second equation of (1.71) gives then the desired result.

It is now trivial to see that $\delta \equiv 0$ gives the isentropic equilibrium. To retrieve the isothermal equilibrium, one can determine δ by setting $p_x = RT\rho_x$ to get $\delta(x) = \frac{1-\gamma}{\gamma}\frac{\Phi_x}{RT}$. In the end, one has the equivalent result to the previous analysis, i.e. that the hydrostatic equilibrium equation is underdetermined. While here the parametrization is related to the entropy distribution, rather then the temperature profile. Furthermore, it should be stressed that the limit of the general equilibria (1.66) is well defined and corresponds to the isentropic equilibrium rather then the whole class of hydrostatic equilibria. However, to discuss the consequences of this result is out of the scope of this work.

1.4.2 Computing the Limit Behavior

Finally, it is reasonable, when trying to compute approximations near hydrostatic equilibria, to analyze the system (1.58) with respect to its behavior at low Mach numbers. A nondimentionalization as for the homogeneous case gives the following set of equations

$$\begin{cases} \rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \\ (\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + I\frac{p}{M^2}) = -\frac{\rho}{Fr^2}\nabla\Phi, \\ E_t + \nabla \cdot (\mathbf{u}(E+p)) = -\rho\frac{M^2}{Fr^2}\langle u, \nabla\Phi\rangle, \\ \Phi_t = 0 \end{cases} \tag{1.73}$$

where

$$Fr = \frac{\|\mathbf{u}\|}{\sqrt{x_{ref} g_{ref}}}, \tag{1.74}$$

is the Froude number and g_{ref} is the characteristic gravitational acceleration. When concerning the limit $M \to 0$, the scaling of the Froude number with respect to the Mach number becomes crucial. To see this, expand the dependent variables in the Mach and Froude number to get

$$\rho = \sum_{i,j=0}^{\infty} M^i Fr^j \rho_{i,j} \quad \mathbf{u} = \sum_{i,j=0}^{\infty} M^i Fr^j \mathbf{u}_{i,j} \quad p = \sum_{i,j=0}^{\infty} M^i Fr^j p_{i,j} \quad e = \sum_{i,j=0}^{\infty} M^i Fr^j e_{i,j}.$$

The interesting part is the balance in the momentum equation. Ordering of the terms in powers of M and Fr and looking at the terms in M^{-2} and Fr^{-2} gives

$$\frac{\nabla p_{0,0}}{M^2} = -\frac{\rho_{0,0}}{Fr^2}\nabla\Phi. \tag{1.75}$$

Assuming $Fr = M^k$, then, if and only if $k = 1$, in the limit of $M \to 0$, the hydrostatic equilibrium is reached as

$$\nabla p_{0,0} = -\rho_{0,0}\nabla\Phi.$$

The same results hold true for the first order terms in the expansion. A divergence constraint on the velocity field can not be reached due to the hydrostatic stratification of the density profile. If $k < 1$, then the terms in Mach number will dominate and the limit is equivalent to the one in the homogeneous case, see (1.56). When $k > 1$, the gravitational forces will dominate in the limit and there is no reasonable in the framework of this model. In the end, the scaling of the Froude and the Mach number should be determined by the properties of the physical object under consideration. For computations of atmospheres, the case of $k = 1$ is the most reasonable because in the limit, one reaches the hydrostatic equilibrium. In this case, one can adjust the definition of the asymptotic preserving set for this system as

$$\Omega_{AP} = \{(\rho, (\rho\mathbf{u}), E) \in \mathbb{R}^5; \nabla p_{0,0} = -\rho_{0,0}\nabla\Phi, \nabla p_{0,1} = -\rho_{0,1}\nabla\Phi, \nabla p_{1,0} = -\rho_{1,0}\nabla\Phi\}, \quad (1.76)$$

or in other words, fluctuations around the hydrostatic equilibrium scale with M^2, i.e. $\nabla p + \rho\nabla\Phi = O(M^2)$.

1.5 The Shallow Water Equations

The Shallow Water equations were first derived by Saint-Venant in 1871 [150]. They can be used to describe flows in which the depth is much smaller compared to the width. They can be derived from the Euler equations with gravity (1.58). Certain assumptions are necessary to do that. First write the Euler equations with gravity such that the gravity is only active in the vertical component and constant. The fluid is considered to be incompressible. Therefore there is

$$\begin{cases} \rho = const, \\ \nabla \cdot \mathbf{u} = 0, \\ u_t + uu_x + vu_y + wu_z + \frac{p_x}{\rho} = 0, \\ v_t + uv_x + vv_y + wv_z + \frac{p_y}{\rho} = 0, \\ w_t + uw_x + vw_y + ww_z + \frac{p_z}{\rho} = -g, \\ P_t + uP_x + vP_y + wP_z = 0. \end{cases} \quad (1.77)$$

The flow is modeled as being enclosed by two boundaries. On the bottom by a topography term $B(x, y)$ and at the top by a free surface $S(t, x, y) = h(t, x, y) + B(x, y)$, where h denotes the height of the fluid. The boundary conditions are formulated as follows

$$\begin{cases} S_t + uS_x + vS_y = w \text{ for } z = h + B, \\ p = 0 \text{ for } z = h + B, \\ \mathbf{u} \cdot \nabla(z - B(x, y)) = 0 \text{ for } z = B. \end{cases} \quad (1.78)$$

The first condition states that the free surface gets advected in the x, y plane by the flow and lifted in the vertical direction by the vertical velocity w. The second condition reflects that there is no pressure at the surface and the third condition gives that the bottom B acts like a solid wall boundary, i.e. at the bottom, the flow is parallel to the bottom. Furthermore, a long-wavelength approximation is used which states, that the resulting waves are much

longer then the depth. Therefore, vertical accelerations can be neglected and the vertical momentum equation from (1.78) becomes

$$p_z = -g\rho. \tag{1.79}$$

Integrating (1.79) from the surface to an arbitrary height $\bar{z}$ by using the zero pressure assumption on the surface gives

$$p(t,x,y,z) = g\rho(h + B - \bar{z}). \tag{1.80}$$

(1.80) is now replacing the pressure equation in (1.78). Using it further in the momenta equations for u and v and neglecting their vertical variations gives then

$$\begin{cases} u_t + uu_x + vu_y + g(h+B)_x = 0, \\ v_t + uv_x + vv_y + g(h+B)_x = 0. \end{cases} \tag{1.81}$$

The next goal is to derive an evolution equation for the fluid height h. To this end, consider the divergence constraint and integrate it in the vertical direction to get

$$\begin{aligned} 0 &= \int_B^{h+b} \nabla \cdot \mathbf{u} dz = \frac{\partial}{\partial x}\int_B^{h+B} u dz + \frac{\partial}{\partial y}\int_B^{h+b} v dz + [w]_B^{h+B} \\ &= (hu)_x + (hv)_y + [w - uz_x - vz_y]_B^{h+B}, \end{aligned}$$

while in the third step it is used, that u and v do not depend on z and zeros have been added with the terms uz_x and vz_y. Now the boundary conditions at the bottom and the surface can be used to further get

$$0 = (hu)_x + (hv)_y + S_t.$$

Since the surface is the sum of the waterheight h and the bottom B, which does not depend on time, one has the final form of the continuity equation

$$h_t + (hu)_x + (hv)_y = 0.$$

Multiplying the momenta equation (1.81) with h and using the continuity equation, one gets the Shallow Water model with bottom topography

$$\begin{cases} h_t + (hu)_x + (hv)_y = 0, \\ (hu)_t + (hu^2 + g\frac{h^2}{2})_x + (hvu)_y = -ghB_x, \\ (hv)_t + (huv)_x + (hv^2 + g\frac{h^2}{2})_y = -ghB_y. \\ B_t = 0 \end{cases} \tag{1.82}$$

This system is hyperbolic with eigenvalues $\lambda_1 \in \{0, u \pm \sqrt{gh}\}$ and $\lambda_2 \in \{0, v \pm \sqrt{gh}\}$ as long as $h > 0$. One might see the similarity to the Euler system (1.58) by stating the pressure as $p = g\frac{h^2}{2}$ and leaving out the equation for energy. In a similar fashion, an entropy, can be defined, see for example [23], as

$$s = \frac{u^2 + v^2}{2} + g(h + B), \tag{1.83}$$

and it can be shown that there holds

$$s_t + (us)_x + (vs)_y \leq 0, \tag{1.84}$$

where the equality holds for smooth solutions. In the case of the shallow water equations, the entropy is referred to as an energy for the system. As well as for the Euler equations, a physical relevant set can be defined as

$$\Omega_{Phys} = \{(h, hu, hv) \in \mathbb{R}^3; h > 0\}. \tag{1.85}$$

One can search also here for the equilibrium solutions of this system. For this, the case of two space dimensions is omitted and only one spatial dimension is considered. Setting the time derivatives to 0 one has after some simplification

$$\begin{cases} (hu)_x = 0, \\ (\frac{u^2}{2} + g(h + B))_x = 0. \end{cases} \tag{1.86}$$

These are referred to as moving equilibria. This class of equilibria is rich and, due to its nonlinearities, delicate to deal with. Of special interest is a subclass of these solutions when the velocity is set to 0. These are called the Lake at Rest solutions and are determined by

$$\begin{cases} u = 0, \\ g(h + B)_x = 0. \end{cases} \tag{1.87}$$

In contrast to the hydrostatic equilibrium relation for the Euler equations, for determining the Lake at Rest solutions, the resonance phenomenon does not occur and one in the end does not have to deal with non-unique solutions. Additionally, the Lake at Rest relations are algebraic equations instead of differential equations, which makes it easier to search for a solution. However, resonance may occur for the moving equilibria (1.86). when the flow changes type between sub- and supercritical, i.e. $u^2 = gh$.

2 Finite Volume Approximations of Hyperbolic PDEs

This chapter is devoted to the derivation of the basic framework to design the numerical scheme used in this work to search for approximations to the solutions of hyperbolic PDEs. A key problem in searching for approximations to the PDEs with the help of a computer is that PDEs are defined on a continuum in space and time. On a computer however, only discrete values can be handled. Therefore, the PDE is discretized, or in other words, approximated by a form which can then be used to compute solutions on a computer. Hence, when designing a numerical scheme, the solution one gets is not a solution to the original PDE, but at best an approximation. Even though, one might hope that when enough information is put into the computer, those approximations will be sufficiently close to the solutions of the underlying PDE. Another maybe even more drastic viewpoint on this issue is the construction of so called modified equations, see for example [80],[167]. The idea is that the numerical scheme actually solves a different PDE to higher accuracy than the underlying PDE.

There are many different ways to discretize a hyperbolic PDE. Two main approaches consist of either projecting the continuum on discrete points, like in finite difference methods, see for example [117],[73],[152], or to project the distributions u on some suitable function spaces, like in finite volume or Galerkin methods, see for example [115],[161],[23],[43],[81]. While discussing the advantages of every approach is out of the scope of this work, the basic framework will be the finite volume approach.

2.1 Finite Volume Approach for Conservation Laws

In order to derive the finite volume scheme for conservation laws, first recast the general form of a conservation law (1.1) in one space dimension equipped with an initial condition as

$$\begin{cases} u_t + f(u)_x = 0, \\ u(0,x) = u_0. \end{cases} \tag{2.1}$$

One is interested to find approximations to the solution to the Cauchy problem (2.1). To this end, first the computational domain $\mathscr{D}$ is divided into N_x finite volumes V_i as

$$\forall_{i=1}^{N_x} \;\; V_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], \tag{2.2}$$

where the volumes are not overlapping, i.e. $V_i \cap V_j = \emptyset$ and the volumes are covering the whole domain, i.e. $\mathscr{D} = \cup_{i=1}^{N_x} V_i$. When integrating the conservation law over a volume V_i and using the divergence theorem one has

$$\int_{V_i} u_t dx + f(u(t, x_{i+\frac{1}{2}})) - f(u(t, x_{i-\frac{1}{2}})) = 0. \tag{2.3}$$

Defining the averaged quantity $U(t)_i = \frac{1}{\Delta_{x_i}} \int_{V_i} u(t,x)dx$, normalized by the size of the volume $\Delta_{x_i} = \int_{V_i} dx$ one can rewrite (2.3) to get

$$U_{i,t} + \frac{1}{\Delta_{x_i}}(f(u(t, x_{i+\frac{1}{2}})) - f(u(t, x_{i-\frac{1}{2}}))) = 0. \tag{2.4}$$

The problem now has been transformed from solving a PDE to solving a set of N_x ODEs. Observe, that in the definition of the fluxes across the boundaries of the volumes V_i, the exact solution u to (2.1) is still needed. However, one does not know the exact solution and therefore one needs to approximate the flux functions. Usually this can be done by setting

$$f(u(t,x)) \approx F(U_1, ..., U_{N_x}, x). \tag{2.5}$$

F is also called the numerical flux function. This can be used in (2.5) to have

$$U_{i,t} + \frac{1}{\Delta_{x_i}}(F_i(U_1, ..., U_{N_x}, x_{i+\frac{1}{2}}) - F_i(U_1, ..., U_{N_x}, x_{i-\frac{1}{2}})) = 0. \tag{2.6}$$

It is very important to understand, that the step from (2.5) to (2.6) is critical. Up to then, only reformulations of the original conservation law have been made. But now the equation has changed by replacing the exact flux f with the numerical flux function F. Therefore, solving the ODEs arising from (2.6) is in general not equivalent to solving the original conservation law (2.1). In the end, if in (2.6) a somehow well suited approximation is used, one might hope to achieve reasonable approximations to solutions of (2.1). Important properties for the choice of the numerical flux function are stated in definition 2.1.1.

Definition 2.1.1. *A numerical flux function $F(U_1, ..., U_{N_x}, x)$ is called consistent with the flux function $f(u)$ if*

$$U_1 = ... = U_{N_x} = u \Rightarrow F(u, ..., u, x) = f(u).$$

A numerical flux functions F_i are called conservative, if

$$F_i(U_1, ..., U_{N_x}, x_{i+\frac{1}{2}}) = F_{i+1}(U_1, ..., U_{N_x}, x_{i+\frac{1}{2}}).$$

The consistency property connects the numerical scheme (2.6) to the underlying PDE (2.1) in a sense, that in the simplest case of $u = const$, the numerical scheme exactly reproduces the dynamics of the PDE. Moreover, consider for theoretical purposes that one might want to know if the approximations produced by the scheme converge to the exact solution. The convergence process might be formulated by looking at the limit $\Delta_{x_i} \to 0$. Then, as long as u is smooth, the averaged values U_i approach the exact value of u at their respective position in space. In order for the numerical approximations to achieve the same limit, the numerical flux function should share the same limit behavior. In order to achieve that some regularity, like for example Lipschitz continuity, is usually asked from the numerical flux function.

The conservation property of a flux function is, not surprisingly, designed to reflect the conservation property of the hyperbolic PDE. To see this, multiply (2.6) with Δ_{x_i} and then

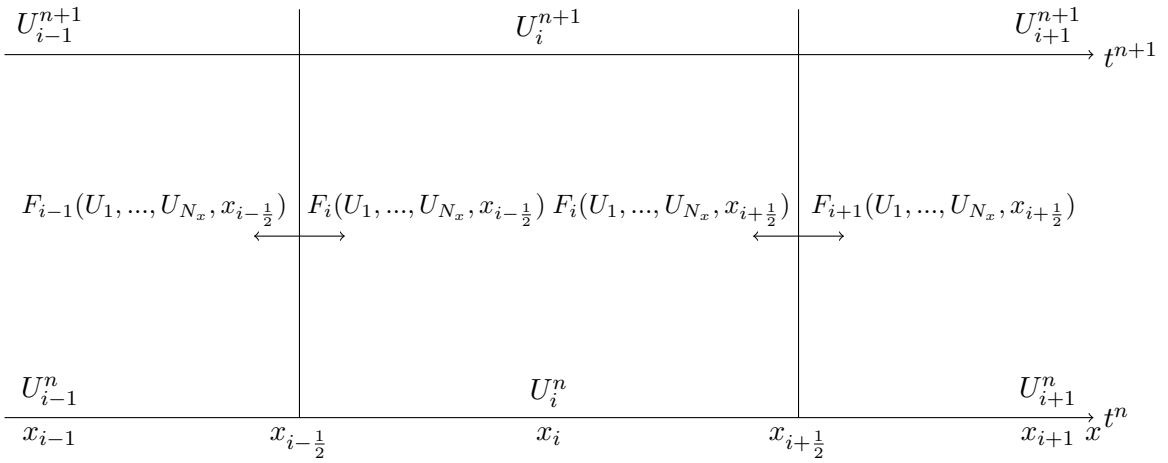


Fig. 2.1: Depiction of a finite volume discretization.

sum over the whole domain $\mathscr{D}$ to get

$$0 = \sum_{i=1}^{N_x} \left(\Delta_{x_i} U_{i,t} + F_i(U_1, ..., U_{N_x}, x_{i+\frac{1}{2}}) - F_i(U_1, ..., U_{N_x}, x_{i-\frac{1}{2}}) \right)$$
$$= \int_{\mathscr{D}} u_t d\mathscr{D} + F_{N_x}(U_1, ..., U_{N_x}, x_{N_x+\frac{1}{2}}) - F_1(U_1, ..., U_{N_x}, x_{1-\frac{1}{2}}).$$

Due to the conservation property, the sum is telescoping and the remaining numerical flux terms are fluxes at the boundary of the domain $\mathscr{D}$. Additionally, the conservation property is also a necessary condition for convergence of a numerical scheme. If a scheme is not in conservation form, the approximations may suggest the wrong propagation of discontinuities, even when reducing the size of the volumes, see [115] for more details.

The formulation for the numerical scheme (2.6) is up to now only discrete in space and the time derivative is still defined on a continuum. A classical way is to discretize time into a sequence of points t_n, where $t_n = n\Delta_t$. Integrating (2.6) over the time interval $[t_n, t_{n+1}]$ and using the divergence theorem one has

$$U_i^{n+1} - U_i^n + \frac{1}{\Delta_{x_i}} \int_{t_n}^{t_{n+1}} F_i(U_1, ..., U_{N_x}, x_{i+\frac{1}{2}}) - F_i(U_1, ..., U_{N_x}, x_{i-\frac{1}{2}}) dt = 0, \tag{2.7}$$

where $U_i^n = U_i(t_n)$, see also figure 2.1. Observe that the numerical flux functions in this formulation are still time dependent. Therefore, another approximation is needed to evaluate the time integral. It is reasonable to assume that all the averages U_i^n are known at the time instance t_n. A straightforward way to approximate the integral is therefore

$$F_i(U_1, ..., U_{N_x}, x_{i+\frac{1}{2}}) \approx F_i^n(U_1^n, ..., U_{N_x}^n, x_{i+\frac{1}{2}}). \tag{2.8}$$

Now the numerical flux functions are not depending on time anymore and the evaluation of the integral is simple. The numerical scheme can be rewritten as

$$U_i^{n+1} = U_i^n - \frac{\Delta_t}{\Delta_{x_i}} \left(F_i^n(U_1^n, ..., U_{N_x}^n, x_{i+\frac{1}{2}}) - F_i^n(U_1^n, ..., U_{N_x}^n, x_{i-\frac{1}{2}}) \right). \tag{2.9}$$

Now all the known quantities are on the right side and the unknowns can be computed explicitly by the formula (2.9). This type of time discretization is also often referred to as a forward Euler time step. Alternatively, the dependence of the flux functions might be chosen to be on the new values U_i^{n+1}, i.e.

$$F_i(U_1, ..., U_{N_x}, x_{i+\frac{1}{2}}) \approx F_i^{n+1}(U_1^{n+1}, ..., U_{N_x}^{n+1}, x_{i+\frac{1}{2}}), \tag{2.10}$$

where this leads to the following numerical scheme

$$U_i^{n+1} + \frac{\Delta_t}{\Delta_{x_i}} \left(F_i^{n+1}(U_1^{n+1}, ..., U_{N_x}^{n+1}, x_{i+\frac{1}{2}}) - F_i^{n+1}(U_1^{n+1}, ..., U_{N_x}^{n+1}, x_{i-\frac{1}{2}}) \right) = U_i^n. \tag{2.11}$$

Again all the unknowns are put on the left hand side of the equation. In contrast to the forward Euler formulation, the unknowns U_i^{n+1} can now in general not be directly computed from the average values at the previous time step and a in general nonlinear system of equations has to be solved. This way of discretizing in time is also referred to as a backward Euler time step.

The time interval of length Δ_t seems arbitrary. However, in order for (2.9) to give reasonable approximations, the time step Δ_t has to satisfy a stability criterium. Consider again the conservation law in (2.1)

$$u_t + f(u)_x = 0. \tag{2.12}$$

Since f is hyperbolic, the matrix T, composed of all the right eigenvectors of f_u, has full rank and one can multiply (2.12) from the left with T^{-1} to get

$$T^{-1}u_t + T^{-1}f_u T T^{-1}u_x = 0. \tag{2.13}$$

Define now the variables $w = T^{-1}u$ to get

$$w_t + Dw_x = 0. \tag{2.14}$$

w are also called the characteristic variables. D is a diagonal matrix as

$$D = diag(\lambda_1, ..., \lambda_n), \tag{2.15}$$

where λ_i are the eigenvalues of f_u. What can be seen from this equation is that the eigenvalues represent velocities. Especially they determine how fast information propagates in the domain. Since the eigenvalues are real and finite, one can conclude that the solution u at a certain point (t, x) is influenced only by a finite set in space at an earlier time $t - \Delta_t$. This finite set is also called the domain of dependence of $u(t, x)$ and is defined as follows

Definition 2.1.2. *The set $\mathscr{D}_{f,\Delta_t}(t, x)$ defined as*

$$\mathscr{D}_{f,\Delta_t}(t, x) = \{\bar{x} \mid u(t - \Delta_t, \bar{x}) \text{ has influence on } u(t, x)\}, \tag{2.16}$$

is called the domain of dependence of (t, x) *under the system (2.12)*

To exactly compute the domain of dependence might be very difficult. One problem is that for a general nonlinear system the eigenvalues depend on the solution, i.e. $\lambda_i = \lambda_i(w)$. So, apart from trivial cases, it is in general hard to get an exact evolution of the eigenvalues and it is almost equivalent to actually solve the system (2.12) exactly. For linear systems things are easier, since the eigenvalues do not depend on the solution and the propagation of information can be determined exactly. For nonlinear systems, estimates on the eigenvalues are needed to get information on the domain of dependence. Assume that for the time interval $[t - \Delta_t, t]$, there exists bound on the eigenvalues as $\lambda_- < \lambda_i < \lambda_+$, then it is easy to see that

$$\mathscr{D}_{f,\Delta_t}(t, x) \subset [x - \lambda_+\Delta_t, x - \lambda_-\Delta_t] \tag{2.17}$$

Now, one can define a domain of dependence also for the time explicit numerical scheme (2.9). Given the formulas from above it is straightforward to see that

$$\mathscr{D}_{F,\Delta_t}(t, x) = [x - \Delta_x, x + \Delta_x]. \tag{2.18}$$

For the numerical scheme to produce reasonable approximations to the PDE, the Domain of dependence of the PDE at a given point (t, x) should always be a subset of the Domain of dependence of the numerical scheme. In other words, for the numerical scheme to compute an approximation at a point (t, x), it should at least have all the information the PDE has to determine the value u at that point. So the stability criterium can be formulated as

$$\mathscr{D}_{f,\Delta_t}(t, x) \subset \mathscr{D}_{F,\Delta_t}(t, x). \tag{2.19}$$

Given the estimate (2.17) and the domain of the dependence of the numerical scheme (2.18), (2.19) can further be rewritten as

$$\frac{\Delta_x}{\Delta_t} C_{CFL} \leq \lambda_{max}, \tag{2.20}$$

where $\lambda_{max} = \max(|\ \lambda_-\ |, |\ \lambda_+\ |)$. C_{CFL} is the Courant-Friedrichs-Lewy-number (CFL)-number and was first determined in [48] as a necessary condition for stability. A classical way to satisfy 2.20 is to choose $C_{CFL} = \frac{1}{2}$.

Regarding the implicit in time discretization (2.11), the domain of dependence for any (t, x) is actually the whole computational domain. Therefore, such a discretization is unconditionally stable. However, as mentioned before, a maybe nonlinear system of equations has to be solved in order to determine the new values U_i^{n+1}, which is, if single time integration steps are compared, more costly then the explicit formulation (2.9). On the other hand, due to the stability of the implicit scheme, larger time steps might be taken. To be more precise, denote by C_E and C_I the computational cost to perform one Euler step with the explicit (2.9) and the implicit integration step (2.11) respectively. Additionally, denote by dt_E and dt_I the time increments that are allowed for the different discretization techniques such that the CFL criterium holds. Even though the implicit time integrations does not need a time step restriction due to stability, as will be explained in section 2.3, the numerical error scales with the time increment. Therefore the time increment is restricted due to accuracy reasons. Now, one can compute the time a computer would need to integrate the initial condition in (2.1) up to a certain time T. For the explicit scheme, this gives $T_E = T\frac{C_E}{dt_E}$ and for the

implicit $T_I = T\frac{C_I}{dt_I}$. Therefore, if dt_I compared to dt_E is large enough, an implicit time discretization can be computationally more efficient then its explicit counter part.

After defining the numerical scheme, one is interested, whether the scheme gives actually reasonable approximations to the underlying conservation law. Moreover, can someone hope for convergence to an exact solution of the conservation law? In general, it is very hard to proof convergence and results depend heavily on the type of PDE under consideration. For systems, there are a few numerical schemes for which convergence could be proven, even when only one space dimension is considered. A fundamental result is due to Glimm [63] by using the random choice method and DiPerna [57] using the front tracking method. For a more comprehensive overview on the topic see for example [82]. Those two methods are somewhat different from the classical finite volume framework presented here and their results are not applicable to the presented framework. A more general but weaker theorem on convergence is given by Lax and Wendroff [106], whereas the version here presented is taken from [115].

Theorem 2.1.1 (Lax-Wendroff Convergence Theorem)**.** *Consider a sequence of grids $(\Delta_x, \Delta_t)_l$ with $\Delta_t \to 0$ where $(\frac{\Delta_x}{\Delta_t})_l$ is fixed and satisfying the CFL restriction (2.20). Denote by U_l the numerical approximations computed with the method (2.9). Suppose U_l converges to a function u. Then $u(t,x)$ is a weak solution to (2.12). Furthermore, if the numerical scheme satisfies the discrete entropy inequality (2.23), then the limiting weak solution $u(t,x)$ also satisfies the entropy inequalities.*

The theorem assumes a convergence of the numerical approximations towards a solution. This convergence needs to be specified in order for the theorem to be true.

Definition 2.1.3. *A series of numerical approximations U_l is said to converge towards a solution u if*

$$\int_0^T \int_{\mathscr{D}} |U_l(t,x) - u(t,x)| dx dt \to 0 \ \text{ for } \ l \to 0, \tag{2.21}$$

and

$$TV(U_l(t,\cdot)) \leq R \ \ \forall t \in [0,T], \tag{2.22}$$

where $TV(\cdot)$ is the total variation of a function and can be defined as $TV(V) = \int_{\mathscr{D}} |v(x)_x| dx$.

Another property mentioned in the theorem is the discrete entropy inequality. Analogue to section 1.1, for some methods a numerical entropy entropy-flux pair might be found such that the method (2.9) rewrites as

$$\psi(U_i^{n+1}) \leq \psi(U_i^n) - \frac{\Delta_t}{\delta_{x_i}} \left(\Psi^n_{i+\frac{1}{2}} - \Psi^n_{i-\frac{1}{2}} \right). \tag{2.23}$$

An alternative formulation of the Lax-Wendroff convergence theorem with rigorous proof can be found in [53]. Keep in mind that theorem (2.1.1) does not ensure convergence. But it might give some confidence in the computed approximations, when the numerical approximations show a reasonable behavior with what is expected from an analysis of the underlying PDE.

2.2 Approximate Riemann Solvers

2.2.1 The Godunov Method

Up to now it has not been specified how to choose the numerical flux functions F_i mentioned in the previous section. Not surprisingly, there are many ways to define such a numerical flux function. A family of methods for defining the numerical flux function is the so called Godunov method [68]. Consider a numerical scheme of the type (2.9)

$$U_i^{n+1} = U_i^n - \frac{\Delta_t}{\Delta_{x_i}} \left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}} \right), \tag{2.24}$$

where the numerical flux functions $F^n_{i+\frac{1}{2}}$ only depends on the values from neighboring cells, i.e.

$$F^n_{i+\frac{1}{2}} = F_{i+\frac{1}{2}}(U_i^n, U_{i+1}^n). \tag{2.25}$$

The numerical approximations U_i^n can be seen as piecewise constant approximations to the exact solution u and can be used to define a global approximation function $U(x) = \sum_{i=1}^{Nx} \chi_{V_i} U_i$, where χ_{V_i} is the characteristic function of the cell V_i. Now, $U(x)$ is a piecewise constant function, where it exhibits discontinuities at the cell interfaces $x_{i+\frac{1}{2}}$, giving rise to Riemann problems as described in section 1.1. They can be reformulated as

$$\begin{aligned} &u_t + f(u)_x = 0, \\ &u_0 = \begin{cases} U_i & \text{for } x < 0 \\ U_{i+1} & \text{for } x > 0 \end{cases} . \end{aligned} \tag{2.26}$$

Let $W_{i+\frac{1}{2}}(t,x)$ denote the solution to the Riemann problem (2.26). Since it can be shown that the solution is selfsimilar, i.e. along rays $\frac{x}{t} = const$ the solution $W_{i+\frac{1}{2}}(t,x)$ is constant, the solution is constant in time at the cell interface. Therefore, a suitable definition for the numerical flux function is

$$F^n_{i+\frac{1}{2}} = f(W_{i+\frac{1}{2}}(t,0)). \tag{2.27}$$

This definition clearly satisfies the consistency property of a numerical flux function given in definition 2.1.1. The flux is well defined apart from the cases when there is a discontinuity at the cell interface. Therefore a more robust definition should be given here. Denote by 0^- the limit $x \to 0$ from the left and by 0^+ the limit $x \to 0$ from the right and rewrite the numerical scheme (2.24) as

$$U_i^{n+1} = U_i^n - \frac{\Delta_t}{\Delta_{x_i}} \left(F^n_{i+\frac{1}{2}^-} - F^n_{i-\frac{1}{2}^+} \right), \tag{2.28}$$

and define the numerical fluxes as

$$\begin{cases} F^n_{i+\frac{1}{2}^-} = f(W_{i+\frac{1}{2}}(t,0^-)), \\ F^n_{i-\frac{1}{2}^+} = f(W_{i-\frac{1}{2}}(t,0^+)). \end{cases} \tag{2.29}$$

These defintions are now well defined and the consistency property still holds. What is

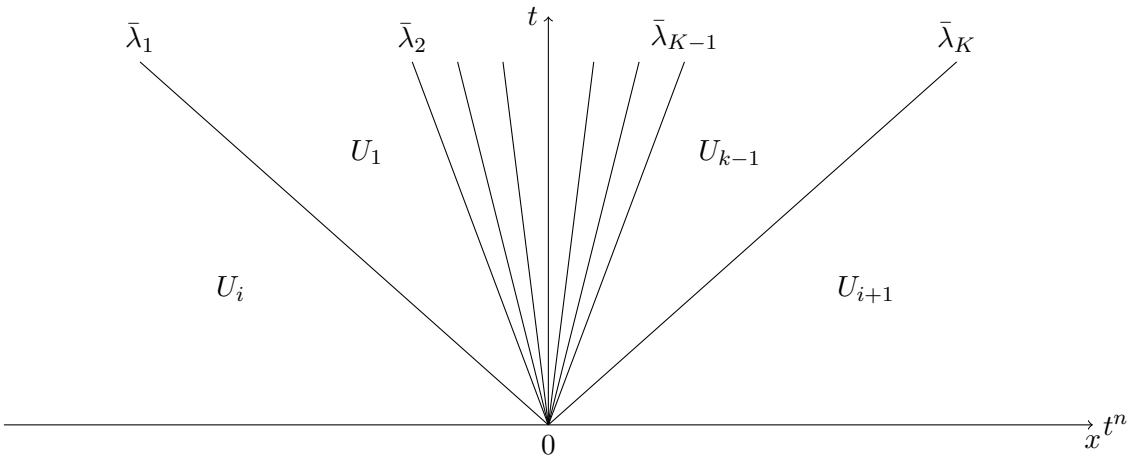


Fig. 2.2: Solution structure of a piecewise constant approximate Riemann solver.

left to check is if the numerical flux functions are still conservative. To this end, consider that the only case where the limits $W_{i+\frac{1}{2}}(t, 0^-)$ and $W_{i+\frac{1}{2}}(t, 0^+)$ are different is when there is a discontinuity at the interface. But then, the Rankine Hugoniout conditions (1.11) must be satisfied. So, for a discontinuity moving with 0 velocity, there must hold $f(u_R) = f(u_L)$ and therefore it holds that

$$f(W_{i+\frac{1}{2}}(t, x^-_{i+\frac{1}{2}})) = f(W_{i+\frac{1}{2}}(t, x^+_{i+\frac{1}{2}})),$$

and thus the numerical flux functions are conservative.

2.2.2 A Model for an Approximate Riemann solver

Now, as has been described in section 1.1, finding the exact solution to the Riemann problem can be hard since in general it involves solving a nonlinear system of equations. Since only the values at the cell interfaces are used, most of the structure of the exact Riemann solution is irrelevant. So the strategy of approximate Riemann solvers is to find an easy way to get at least an approximation to the solution at the interface. There are different strategies to do that. The idea behind the approaches discussed here is the observation, that especially rarefaction waves are hard to compute. In practice it is easier to handle solutions to the Riemann problem, where there are only shocks or contact discontinuities, i.e. constant states separated by discontinuities. Therefore, a framework for approximate Riemann solvers is to find an approximation $\mathcal{W}_{i+\frac{1}{2}}(t, x)$ to the initial value problem (2.26) in the following form

$$\mathcal{W}_{i+\frac{1}{2}}(t, x) = \begin{cases} U_i \text{ if } \frac{x}{t} < \bar{\lambda}_1, \\ w_1 \text{ if } \bar{\lambda}_1 < \frac{x}{t} < \bar{\lambda}_2, \\ \vdots, \\ w_{k-1} \text{ if } \bar{\lambda}_{K-1} < \frac{x}{t} < \bar{\lambda}_K, \\ U_{i+1} \text{ if } \bar{\lambda}_K < \frac{x}{t}, \end{cases} \tag{2.30}$$

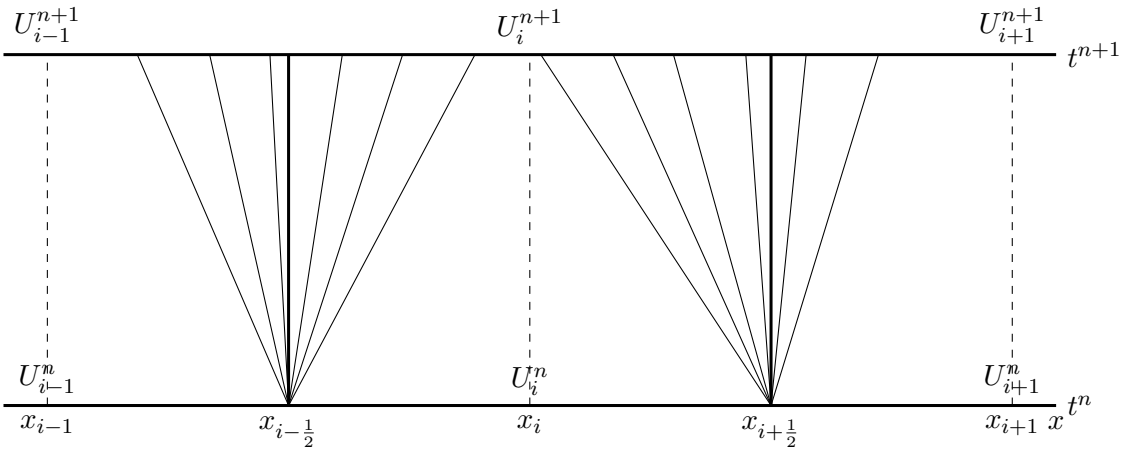


Fig. 2.3: Juxtaposed Riemann problems. The CFl condition ensures, that the waves from neighboring approximate Riemann solvers $\mathcal{W}_{i\pm\frac{1}{2}}(t,x)$ do not interact.

see also figure 2.2 for a graphical depiction. In order to find such a solution, different strategies have been proposed. In the following, the HLL, Roe and Suliciu approximate Riemann solvers are discussed. With the model (2.30), there are some advantages that can be used. As has been mentioned in section 1.3, certain models have restrictions on the domain where the dependent variables are defined, for example the positivity of density and temperature. As in section 1.3 one has for a specific system a set Ω_{Phys} that defines the physical admissible states. Theorem 2.2.1 is helpful in the design of robust approximate Riemann solvers.

Theorem 2.2.1 (Robustness of Approximate Riemann Solvers)**.** *Given the scheme (2.28) and a CFL number of $\frac{1}{2}$, where the numerical flux is defined by the model (2.30). If the set Ω_{Phys} is convex and at each interface the states $w_k \in \Omega_{Phys}$, then $U_i^{n+1} \in \Omega_{Phys}$.*

Proof. *Consider at the cell interfaces $x_{i\pm\frac{1}{2}}$ the two approximate solutions $\mathcal{W}_{i\pm\frac{1}{2}}(t,x)$. From the integral form of the conservation law it is clear that (2.28) can be rewritten as*

$$U_i^{n+1} = U_i^n - \frac{\Delta_t}{\Delta_{x_i}}\left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right) = \frac{1}{\Delta_{x_i}}\int_{x_{i-\frac{1}{2}}}^{x_i} \mathcal{W}_{i-\frac{1}{2}}(x,t^{n+1})dx + \frac{1}{\Delta_{x_i}}\int_{x_i}^{x_i+\frac{1}{2}} \mathcal{W}_{i+\frac{1}{2}}(x,t^{n+1})dx.$$

Since it is assumed, that $\mathcal{W}_{i\pm\frac{1}{2}}(t,x) \in \Omega_{Phys}$ and Ω_{Phys} is convex, due to the convexity of the integrals there is

$$\int_{x_{i-\frac{1}{2}}}^{x_i} \mathcal{W}_{i-\frac{1}{2}}(x,t^{n+1})dx \in \Omega_{Phys} \;\; \textit{and} \;\; \int_{x_i}^{x_i+\frac{1}{2}} \mathcal{W}_{i+\frac{1}{2}}(x,t^{n+1})dx \in \Omega_{Phys},$$

and therefore $U_i^{n+1} \in \Omega_{Phys}$.

The role of the CFL condition is also depicted in figure 2.3. Another important property of a numerical scheme is the entropy stability (2.23). How this stability property can be determined from the framework of an approximate Riemann solver is answered in the next theorem.

Theorem 2.2.2 (Stability of Approximate Riemann Solvers)**.** *Given a numerical scheme as in theorem 2.2.1. Assume additionally, that there exists a convex entropy ψ to the underlying conservation law. Then, if the approximate Riemann solvers $\mathcal{W}_{i\pm\frac{1}{2}}$ satisfy*

$$\begin{aligned}
\int_{x_i}^{x_{i+\frac{1}{2}}} \psi(\mathcal{W}_{i+\frac{1}{2}}(t^{n+1},x))dx \leq & \int_{x_i}^{x_{i+\frac{1}{2}}} \psi(\mathcal{W}_{i+\frac{1}{2}}(t^n,x))dx \\
& - \Delta_t(\Psi(\mathcal{W}_{i+\frac{1}{2}}(t^n,x^-_{i+\frac{1}{2}})) - \Psi(\mathcal{W}_{i+\frac{1}{2}}(t^n,x_i))), \\
\int_{x_{i-\frac{1}{2}}}^{x_i} \psi(\mathcal{W}_{i-\frac{1}{2}}(t^{n+1},x))dx \leq & \int_{x_{i-\frac{1}{2}}}^{x_i} \psi(\mathcal{W}_{i-\frac{1}{2}}(t^n,x))dx \\
& - \Delta_t(\Psi(\mathcal{W}_{i-\frac{1}{2}}(t^n,x_i)) - \Psi(\mathcal{W}_{i-\frac{1}{2}}(t^n,x^+_{i-\frac{1}{2}}))),
\end{aligned} \tag{2.31}$$

then the numerical scheme satisfies the entropy inequality (2.23)

Proof. *Observe, that under the CFL number $\frac{1}{2}$, the values $\mathcal{W}_{i\pm\frac{1}{2}}(t^n,x_i)$ are constant in time. Also due to the selfsimilarity of the approximate Riemann solver $\mathcal{W}$, the values $\mathcal{W}_{i\pm\frac{1}{2}}(t^n,x^{\mp}_{i\pm\frac{1}{2}})$ are constant. Therefore, the numerical entropy fluxes Ψ are well defined, and consistent with exact entropy fluxes if the exact fluxes are evaluated at the respective values. Summing the two conditions in (2.31) gives*

$$\begin{aligned}
\int_{x_{i-\frac{1}{2}}}^{x_i} \psi(\mathcal{W}_{i-\frac{1}{2}}(t^{n+1},x))dx + \int_{x_i}^{x_{i+\frac{1}{2}}} \psi(\mathcal{W}_{i+\frac{1}{2}}(t^{n+1},x))dx & \\
\leq \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \psi(U_i^n)dx - \Delta_t(\Psi(\mathcal{W}_{i+\frac{1}{2}}(t^n,x^-_{i+\frac{1}{2}})) - \Psi(\mathcal{W}_{i-\frac{1}{2}}(t^n,x^+_{i-\frac{1}{2}}))).&
\end{aligned}$$

Furthermore since ψ is convex the Jensen inequality can be applied and the following inequalities hold

$$\begin{aligned}
\int_{x_{i-\frac{1}{2}}}^{x_i} \psi(\mathcal{W}_{i-\frac{1}{2}}(t^{n+1},x))dx + \int_{x_i}^{x_{i+\frac{1}{2}}} \psi(\mathcal{W}_{i+\frac{1}{2}}(t^{n+1},x))dx \geq & \\
\psi(\int_{x_{i-\frac{1}{2}}}^{x_i} \mathcal{W}_{i-\frac{1}{2}}(t^{n+1},x)dx + \int_{x_i}^{x_{i+\frac{1}{2}}} \mathcal{W}_{i+\frac{1}{2}}(t^{n+1},x)dx) = \psi(U_i^{n+1}),&
\end{aligned}$$

which concludes the proof.

It should be remarked that, if for the definition of the numerical fluxes instead of the approximate Riemann solver $\mathcal{W}$ the exact solution W is chosen, the requirements for the theorems 2.2.2 and 2.2.1 are satisfied. Moreover, the usual Harten, Lax and van Leer entropy consistency [78] reads:

$$\frac{1}{\Delta_x}\int_{x_i}^{x_{i+1}} \psi(\mathcal{W}_{i+\frac{1}{2}}(t^{n+1},x))dx \leq \frac{1}{2}(\psi(U_i^n)+\psi(U_{i+1}^n)) - \frac{\Delta_t}{\Delta_x}(\Psi(U_i)-\Psi(U_{i+1})). \quad (2.32)$$

However, (2.32) is a consequence of the conditions (2.31). In fact, the formulations (2.31) will be more convenient to derive in chapter 4.

Remark 2.2.1. *The theorems 2.2.1 and 2.2.2 only apply for the explicit time stepping technique (2.9). If those theorems also hold if an implicit time stepping as (2.11) is chosen is not obvious. Even though implicit time stepping techniques are crucial for the practical relevance of the numerical schemes developed in chapter 5 and chapter 6, to proof the robustness in this case is out of the scope of this work.*

2.2.3 The HLL Approximate Riemann Solver

First, the HLL scheme by Harten, Lax and van Leer [78] is discussed. The idea in the HLL framework is to achieve the model (2.30) by first stating some K wave speeds $\bar{\lambda}_k$. Now one needs to solve for the unknown vectors $U_1,...,U_{K-1}$. To make the model (2.30) consistent with the integral form of the conservation law, choose M large enough and state the following relation

$$\int_{-M}^{M} \mathcal{W}_{HLL}(t,x)dx = \int_{-M}^{M} W(t,x)dx. \quad (2.33)$$

Two things are to be observed in (2.33). First, due to the imposed solution structure (2.30), the integrals can be evaluated exactly. The right hand side rewrites

$$\frac{1}{2M}\int_{-M}^{M} W(t,x)dx = \frac{U_L+U_R}{2} + t(f(U_L)-f(U_R)). \quad (2.34)$$

and the left side rewrites as

$$\int_{-M}^{M} \mathcal{W}_{HLL}(t,x)dx = (M+t\bar{\lambda}_1)w_i + t\sum_{i=1}^{K-1} w_i(\bar{\lambda}_{i+1}-\bar{\lambda}_i) + (M-t\bar{\lambda}_K)w_{i+1}. \quad (2.35)$$

Second, (2.33) gives m equations. In total there are $(K-1)m$ unknowns for the system (2.30). In order to have a well posed system based on only (2.33) it must hold $(K-1)m = m$, i.e $K = 2$. For $K = 2$ one has exactly the HLL scheme as proposed in [78]. There are various extensions of the HLL scheme, most prominent the HLLC approximate Riemann solver, where an additional third wave is modeled to capture the contact discontinuity of the full Euler system (1.32). Then, additional relations to (2.33) have to be imposed in order to get a well posed, hopefully linear, system for the unknowns $U_1,...,U_{K-1}$.

2.2.4 The Suliciu Relaxation Approximate Riemann Solver

Approximate Riemann solvers based on relaxation systems are a relatively recent approach. However, it is gaining popularity since its flexibility has been used by several authors to tackle different problems, see for example (see [4, 10, 15, 18, 23, 24, 25, 37, 38, 46, 47, 61, 88, 116]).

A specific type of approximate Riemann solver is the Suliciu relaxation approach by Coquel and Perthame [47]. It can be derived from the previous mentioned relaxation systems, see section 1.2.2. One seeks to find another system approximating the conservation law and solve the Riemann problem for this system instead to define the numerical fluxes. Different approaches have been made to derive relaxation systems for the definition of numerical fluxes, while best know is probably the Jin-Xin relaxation [88], as also has been introduced in section 1.2.2. As it turns out, the HLL scheme can be reformulated as a variant of the Jin-Xin relaxation [23]. The Jin-Xin relaxation is very general and can be used for any conservation law. In this work, the relaxation approach is used to deal with a specific type of system, namely the compressible Euler equations of gas dynamic. The Suliciu relaxation is derived for specific systems, as for example for the compressible Euler equations, and it is presented here following the notes in [23].

Consider the Euler equations in one space dimension

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + p) = 0, \\ E_t + (u(E+p))_x = 0. \end{cases} \tag{2.36}$$

This system exhibits the full nonlinear dynamics which lead to difficulties when trying to find the solution to the Riemann problem. The idea is to extend the system (2.36) by an additional equation to get a simpler system. To this end, denote by $p' = \frac{\partial p}{\partial \rho}|_{s=const}$ and multiply the first equation in (2.36) by p' to get

$$p_t + u p_x + \rho p' u_x = 0. \tag{2.37}$$

Now, multiply (2.37) by ρ and the continuity equation in (2.36) by p. Adding these equations gives then

$$(\rho p)_t + (\rho u p)_x + \rho^2 p' u_x = 0. \tag{2.38}$$

Adding (2.38) to the system (2.36) is valid for smooth solutions. Keep in mind that the pressure is actually already determined by the conserved variables. However, one wants to resolve the Riemann problem at the cell interfaces and the nonconservative product $\rho^2 p'$ is hard to evaluate across discontinuities. Therefore, an additional degree of freedom is introduced by replacing p with π, $\rho^2 p'$ by some constant c^2 and adding a source term to equation (2.38) to connect the new system to the original equations. The Suliciu relaxation system now reads

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + \pi)_x = 0, \\ E_t + (u(E+\pi))_x = 0, \\ (\rho \pi)_t + (\rho u \pi + c^2 u)_x = \frac{\rho}{\varepsilon}(p - \pi). \end{cases} \tag{2.39}$$

The new variable π is called the relaxation pressure and can be understood as a perturbation of the original pressure p. Lemma 2.2.1 concerns the stability of the relaxation system.

Lemma 2.2.1. *The system (2.39) is a stable first order perturbation of the system (2.36)*

and the subcharacteristic condition reads

$$c^2 > \rho^2 p'. \tag{2.40}$$

Proof. *Performing a Chapman-Enskog expansion as has been introduced in section 1.2.2, the first order perturbation of (2.39) reads*

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + p)_x = \varepsilon(\frac{1}{\rho}(c^2 - \rho^2 p')u_x)_x, \\ E_t + (u(E+p))_x = \varepsilon(\frac{1}{\rho}(c^2 - \rho^2 p')(\frac{u^2}{2})_x)_x. \end{cases} \tag{2.41}$$

Now, it is not clear how the system (2.39) may lead to an easier solution to the Riemann problem, since now there is one more equation and also a source term present in the system. To see the advantage of the relaxation approach, one takes a closer look at the homogeneous part of system (2.39), i.e. setting $\varepsilon = \infty$.

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + \pi)_x = 0, \\ E_t + (u(E+\pi))_x = 0, \\ (\rho\pi)_t + (\rho u\pi + c^2 u)_x = 0. \end{cases} \tag{2.42}$$

Lemma 2.2.2 discusses the algebra and robustness of the system (2.42)

Lemma 2.2.2. *The system (2.42) is hyperbolic and admits only linear degenerate eigenvalues $\lambda \in \{u, u \pm \frac{c}{\rho}\}$ and therefore admits a solution of the form (2.30) as*

$$\mathcal{W}_{SR}(t,x) = \begin{cases} U_L \ \text{ if } \ \frac{x}{t} < u - \frac{c}{\rho}, \\ U_{CL} \ \text{ if } \ u - \frac{c}{\rho} < \frac{x}{t} < u, \\ U_{CR} \ \text{ if } \ u < \frac{x}{t} < u + \frac{c}{\rho}, \\ U_R \ \text{ if } \ u + \frac{c}{\rho} < \frac{x}{t}. \end{cases} \tag{2.43}$$

Moreover, the intermediate states (2.43) can be computed explicitly and are give by

$$\begin{aligned} \pi_C &= \frac{\pi_L + \pi_R}{2} + c\frac{u_L - u_R}{2}, & u_C &= \frac{u_L + u_R}{2} + \frac{\pi_L - \pi_R}{2c}, \\ \frac{1}{\rho_{CR}} &= \frac{1}{\rho_R} + \frac{\pi_R - \pi_C}{c^2}, & \frac{1}{\rho_{CL}} &= \frac{1}{\rho_L} + \frac{\pi_L - \pi_C}{c^2}, \\ e_{CR} &= e_R - \frac{\pi_R^2 - \pi_C^2}{2c^2}, & e_{CL} &= e_L - \frac{\pi_L^2 - \pi_C^2}{2c^2}, \end{aligned} \tag{2.44}$$

and for $c > 0$ large enough, there is $U_{CL}, U_{CR} \in \Omega_{phys}$.

The solution structure (2.43) is also depicted in figure 2.4.

Proof. *The system (2.42) can be diagonalized to get*

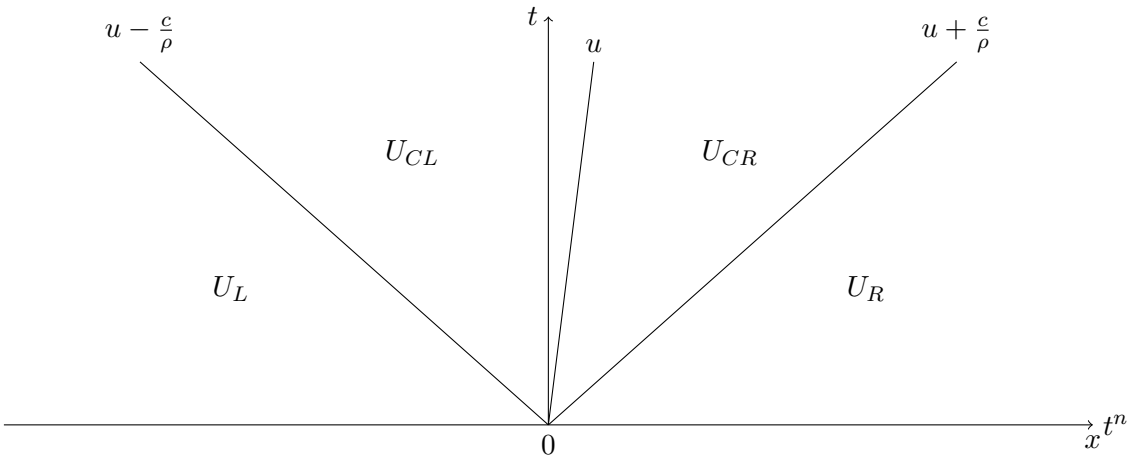


Fig. 2.4: Solution for a Riemann problem for the homogeneous Suliciu relaxation system (2.42).

$$\begin{cases} (\pi + cu)_t + (u + \frac{c}{\rho})(\pi + cu)_x = 0, \\ (\pi - cu)_t + (u - \frac{c}{\rho})(\pi - cu)_x = 0, \\ (\frac{1}{\rho} + \frac{\pi}{c^2})_t + u(\frac{1}{\rho} + \frac{\pi}{c^2})_x = 0, \\ (e - \frac{\pi^2}{2c^2})_t + u(e - \frac{\pi^2}{2c^2})_x = 0. \end{cases} \tag{2.45}$$

and the eigenvalues can be read directly from (2.45). To see that they are linear degenerate, one could compute the eigenvectors and check the condition (1.15). A more direct approach is now used here. From the definition of linear degeneracy (1.15) and the definition 1.1.6 for Riemann invariants, it is clear that an eigenvalue is linear degenerate if and only if it is an Riemann invariant to its field. Moreover, any linear combination of Riemann invariants is again a Riemann invariant. The diagonalized system (2.45) already provides the complete set of Riemann invariants, i.e. all the characteristic variables to the other fields. Therefore, if an eigenvalue λ_i can be expressed as a sum of the characteristics which are not transported with its velocity, the eigenvalue will also be an invariant for its field and therefore linear degenerate. The following relations hold

$$\begin{aligned} u \pm \frac{c}{\rho} &= \pm c(\frac{1}{\rho} + \frac{\pi}{c^2}) \mp \frac{(\pi \mp cu)}{c}, \\ u &= \frac{(\pi + cu)}{c} - \frac{(\pi - cu)}{c}. \end{aligned}$$

From (2.45) the Riemann invariants to a wave with a velocity λ_i, denoted by Φ_{λ_i}, can be computed to be

$$\begin{cases} \Phi_{u \pm \frac{c}{\rho}} = \{\frac{1}{\rho} + \frac{\pi}{c^2}, e - \frac{\pi^2}{2c^2}, \pi \mp cu\}, \\ \Phi_u = \{u, \pi\}. \end{cases} \tag{2.46}$$

Since u and π are invariant across the center wave, the notation $\pi_{CL} = \pi_{CR} = \pi_C$ is

introduced. Computing the intermediate states now involves solving the linear system arising from the invariants described in (1.18) and is left to the reader.

For the positivity of the densities it is sufficient to demand the ordering of the eigenvalues as

$$u - \frac{c}{\rho} < u < u + \frac{c}{\rho}. \tag{2.47}$$

Since the eigenvalues are also Riemann invariants, on can rewrite (2.47) in terms of the intermediate states to get

$$u_C - \frac{c}{\rho_{CL}} < u_C < u_C + \frac{c}{\rho_{CR}}.$$

Subtracting u_C from these inequalities then easily gives $\rho_{CR}, \rho_{CL} > 0$ as long as $c > 0$. For the positivity of the internal energies, one needs to expand the expression in (2.44) to get

$$e_{CR} = e_R + \frac{\frac{1}{8}(\pi_L + \pi_R)^2 - \frac{\pi_R^2}{2}}{c^2} + \frac{(\pi_L + \pi_R)(u_L - u_R)}{4c} + \frac{(u_L - u_R)^2}{8},$$
$$e_{CL} = e_L + \frac{\frac{1}{8}(\pi_L + \pi_R)^2 - \frac{\pi_L^2}{2}}{c^2} + \frac{(\pi_L + \pi_R)(u_L - u_R)}{4c} + \frac{(u_L - u_R)^2}{8}.$$

The positivity of the internal energy follows for c large enough. In fact, multiplying by c^2 gives a second order polynomial in c and the roots can be computed explicitly.

It should be remarked that some physical applications demand not only that $e > 0$, but also that the internal energy stays above a certain threshold e_{min}. As long as $e_{min} < e_R, e_L$, the relations above give an explicit formula of how to choose the relaxation parameter c to ensure this restriction.

After discussing the algebra and the properties of the Suliciu relaxation system, one might be convinced that the homogeneous system has some good properties that might be useful for an approximate Riemann solver. It is now discussed how to use the relaxation system (2.39) to define a numerical scheme. In general, one would like the solutions to the relaxation system to be close to the original system. Formally, this can be achieved by setting ε as small as possible. However, this introduces numerical difficulties since the source term then becomes stiff and very small time steps have to be chosen in order to ensure stability. A common approach is to split the operators in (2.39) and perform two updates on the numerical solution, i.e. an evolution step based on the homogeneous relaxation system (2.42) and a projection step which involves only the source term. Define for this the numerical flux function as in (2.29) to get for the evolution step

$$\bar{U}_i^{n+1} = U_i^n + \frac{\Delta_t}{\Delta_x}(F^+_{i-\frac{1}{2}} - F^-_{i-\frac{1}{2}}). \tag{2.48}$$

The projection step is then implicitly solved

$$U_i^{n+1} = \bar{U}_i^{n+1} + \frac{\rho}{\varepsilon} R(U_i^{n+1}). \tag{2.49}$$

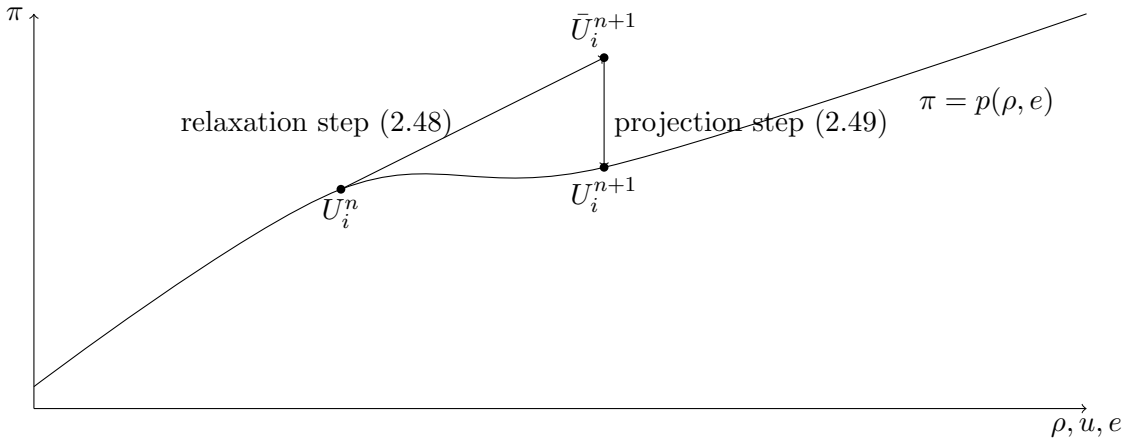


Fig. 2.5: Conceptional drawing of the operator splitting in the relaxation technique in phase space.

The strategy is to solve (2.49) in the limit of $\varepsilon \to 0$, keeping in mind, that one is actually interested in approximating solutions to the Euler equations and not the relaxation system. This procedure is also depicted in figure 2.5. One advantage of this method is, that the relaxation pressure π actually does not need to be computed in the evolution step, because the second step actually gives

$$\pi_i^{n+1} = p_i^{n+1}(\rho_i^{n+1}, T_i^{n+1}).$$

Therefore one gets the relaxation pressure from the conserved quantities at the new time step. So the description of the operator splitting is actually only for theoretical concerns, but in the numerical code neither implicit solvers have to be implemented nor are extra equations needed to describe the evolution of the additional variable.

Since the projection step only acts on the relaxation pressure π, Lemma 2.2.2 together with Theorem 2.2.1 ensure, that the Suliciu relaxation gives a robust numerical scheme. The entropy stability will be discussed in section 4. Even though this concerns the case of the Euler equations with a gravitational potential, the proof can also be directly applied in this case, since it does not involve the gravitational source term.

A general closing remark to the relaxation scheme concerns the efficiency of the scheme. It has been shown, that the solution to the Riemann problem admits an explicit solution and by the use of the projection method, there are never any extra equations introduced in the numerical scheme. Therefore it can be concluded, that the relaxation schemes are competitive if the computational costs are concerned.

2.2.5 The Roe Approximate Riemann Solver

The Roe approximate Riemann solver [147] is maybe most used in practice. Similar to the Suliciu relaxation approach, one seeks to find a simpler system to solve for the Riemann problem. The idea here is to linearize the flux function as $f(u) = Au$ to get the approximating

system

$$u_t + Au_x = 0.$$

The choice of the matrix A is not arbitrary. Roe suggested different criteria this matrix has to satisfy and was able to derive such a matrix. However, the Roe scheme showed problems when dealing with rarefactions, in particular sonic rarefactions. In this case the scheme was shown to violate the discrete entropy inequality (2.23). Moreover, the positivity of density and pressure is also not ensured in its original form. But entropy fixes for the Roe scheme have been proposed and the reader is referred to [115] and [161] and references therein for more details on that subject.

2.3 Higher Order Schemes

Solutions computed by a numerical scheme as (2.24) can in general only serve as approximations to the solutions of a conservation law. Information is lost by projecting data from the continuum onto the finitely many volumes. As already been pointed out by the Lax-Wendroff theorem 2.1.1, one might hope for convergence to an entropy solution of the conservation law. However, in practice convergence is not the primary objective. Decreasing the size of the volumes is costly in terms of computational effort and therefore the task is to set up a numerical scheme that, given a set of volumes, gives the best approximation.

One way to measure the quality of the approximation is to analyze the error that has been made in terms of the volume size and the timesteps. To analyze the error, start with the integral form of a conservation law and integrate in time to get

$$\begin{aligned}
\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(t_{n+1})dx &= \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(t_n)dx - \int_{t_n}^{t_{n+1}} f(u(t, x_{i+\frac{1}{2}}) - f(u(t, x_{i-\frac{1}{2}})dt \\
&= \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(t_n)dx - \Delta_t(f(u(t_n, x_{i+\frac{1}{2}}) - f(u(t_n, x_{i-\frac{1}{2}})) + O(\Delta_t) \\
&= \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(t_n)dx - \Delta_t(F(U^n_{i+1}, U^n_i) - F(U^n_i, U^n_{i-1})) + O(\Delta_t, \Delta_x).
\end{aligned}$$

From the first line to the second line, the approximation given in (2.8) was used. Analogue to that, the time implicit approximation (2.10) might be used. From the second to the third line, the approximation in (2.5) has been used. Assuming smoothness and using Taylor expansions gives for both cases the claimed order of accuracy of the approximation. This short computation shows that the finite volume formulation is a first oder in time and space approximation to the conservation law and therefore, the resulting cell averages are also only first order accurate approximations to the exact solution.

One way to increase the quality of the numerical approximations given fixed volumes V_i and fixed time increments Δ_t is to increase the order of the scheme, i.e. to achieve an error of $O(\Delta_t^p, \Delta_x^p)$ with $p > 1$ in the above calculation. There are various ways in the literature to achieve that, the reader is referred to [115] and [161] for a broader overview on the topic. A general approach is to deal with the order in space and time separately.

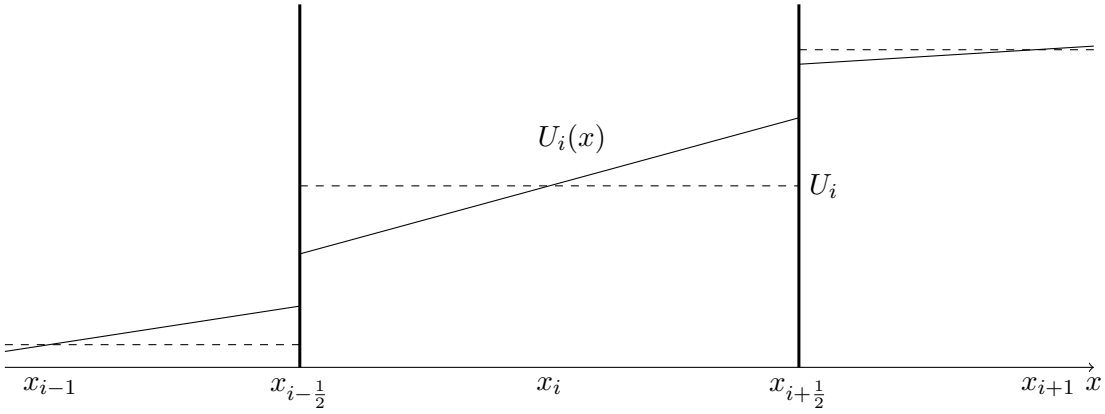


Fig. 2.6: Linear reconstruction of the numerical approximation. Dashed lines show the piecewise constant representation.

2.3.1 Higher Order in Space

For achieving higher order in space, the strategy of the Monotone Upstream-centered Scheme for Conservation Laws (MUSCL), introduced by Van Leer [107],[108],[109]. The idea is to modify the piecewise constant representation of the solution $U(x)$ by piecewise linear functions. More specific, consider the volume V_i, then the numerical approximation reads

$$U_i(x) = U_i + \sigma_i(x - x_i), \tag{2.50}$$

where, when U_i is a vector, the slopes σ_i are also vectors, see also figure 2.6. Observe that the total amount of U_i in the cell V_i does not change after applying the reconstruction (2.50). Therefore this type of reconstruction is also referred to as a conservative reconstruction. Denoting by $U_i(x^-_{i+\frac{1}{2}})$ the limit of the linear function U_i as $x \to x_{i+\frac{1}{2}}$, the numerical scheme then reads

$$U_i^{n+1} = U_i^n - \frac{\Delta_t}{\Delta_{x_i}} \left(F(U_i(x^-_{i+\frac{1}{2}}), U_{i+1}(x^+_{i+\frac{1}{2}})) - F(U_{i-1}(x^-_{i-\frac{1}{2}}), U_i(x^+_{i-\frac{1}{2}})) \right). \tag{2.51}$$

To compute the numerical fluxes, the approximate Riemann solvers derived in section 2.2.2 might be used. In fact, by formulating the numerical fluxes like this already uses a specific interpretation of the linear functions. Recall that the use of the approximate Riemann solvers relied on the fact, that a Riemann problem, i.e. an initial condition with two constant states separated by a discontinuity, could be stated at the cell interfaces. Now, the initial conditions are linear functions, which in turn leads to a so called generalized Riemann problem, see for example [11]. However, solving the generalized Riemann problem can be cumbersome. To avoid this, the linear function is projected onto two constant states inside the cell as

$$\bar{U}_i(x) = \begin{cases} U_i(x^+_{i-\frac{1}{2}}) & \text{if } x < x_i, \\ U_i(x^-_{i+\frac{1}{2}}) & \text{if } x > x_i, \end{cases} \tag{2.52}$$

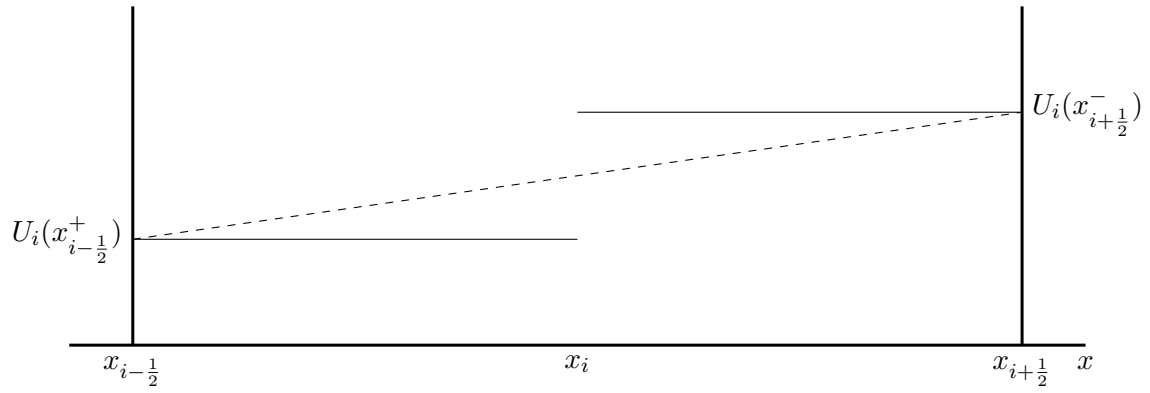


Fig. 2.7: Piecewise constant representation in the case of a conservative linear reconstruction. The dashed line represents the linear reconstruction.

see also figure 2.7. Also here the cell average value of the function does not change. The reconstruction is conservative and with the piecewise constant interpretation leads more naturally to the formulation of the higher order scheme (2.51).

Besides getting a better spatial resolution, these higher order methods require for stability reasons a smaller CFL number. While in the first order case, the CFL could be derived from the waves of the Riemann problem at the interfaces not to pass through the cell center, and therefore resulting in the CFL number $\frac{1}{2}$, now there is a new Riemann problem at the cell center. One usually does not compute the waves coming from the center problem, but assumes that they are somehow bounded by the wave speeds coming from the Riemann problems at the interfaces. Again, in order to avoid wave interactions, the waves are only allowed to travel a quarter of the cell resulting in a CFL of $\frac{1}{4}$.

It should be remarked that it is not necessary to compute the linear functions in (2.50) in terms of the conserved quantities U. Often a reconstruction in characteristic or primitive variables is more convenient for practical applications. Denote this set of variables as $P(U)$. Then a reconstruction can be performed on those variables as

$$P(U)_i(x) = P(U)_i + \sigma_i(x - x_i). \tag{2.53}$$

However, if the reconstruction (2.53) is applied, the conservation property is lost. In the end, the interpretation of piecewise constant data can help here. In [17] it has been found that such a reconstruction can interpreted as projection onto 3 constant states, i.e.

$$\bar{U}_i(x) = \begin{cases} U(P)_i(x^+_{i-\frac{1}{2}}) \text{ if } x < x_{i-\frac{1}{4}}, \\ U_C \text{ if } x_{i-\frac{1}{4}} < x < x_{i+\frac{1}{4}}, \\ U(P)_i(x^-_{i+\frac{1}{2}}) \text{ if } x > x_{i+\frac{1}{4}}, \end{cases} \tag{2.54}$$

where U_C is determined by $U(P)_i(x^+_{i-\frac{1}{2}})$and $U(P)_i(x^-_{i+\frac{1}{2}})$ and is a physical relevant state, if the left and right states are physical relevant, see also figure 2.8. However, the CFL number has to be adjusted in order to ensure stability, i.e. $CFL = \frac{1}{8}$.

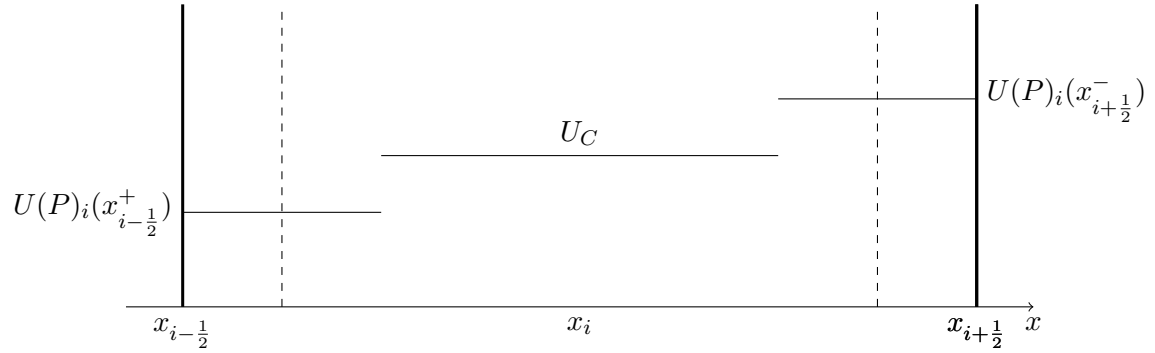


Fig. 2.8: Piecewise constant representation in the case of a non-conservative linear reconstruction. For stability, the waves from the Riemann problem are not allowed to cross the dashed lines in a time step.

Up to now, it has not yet been mentioned, how the slopes σ_i are to be computed. In this work, if not mentioned otherwise, the minmod limiter introduced in [148] is applied. It can be written as

$$\sigma_i = \begin{cases} 0 \text{ if } (U_{i+1} - U_i)(U_i - U_{i-1}) \leq 0, \\ \min(U_{i+1} - U_i, U_i - U_{i-1}) \text{ if } U_{i+1} - U_i > 0, \\ \max(U_{i+1} - U_i, U_i - U_{i-1}) \text{ if } U_{i+1} - U_i < 0. \end{cases} \tag{2.55}$$

If U is a vector, the procedure is understood componentwise. If the minmod procedure is applied to the conservative variables, it can be shown that the total variation of the reconstructed solution is bounded by the piecewise constant solution. The total variation has already been mentioned in the Lax-Wendrof convergence theorem. In fact, in can be shown, that scalar conservation laws satisfy the property that the total variation diminishes. Using this property also for the numerical scheme helps to reduce spurious oscillations, especially near shocks. However, for systems, the TVD property does not hold, but the philosophy is applied to the components.

In the end, there are limits to this TVD procedure. In fact, at local extrema, the minmod procedure is at most first order accurate. This is obvious from the first case in (2.55), since at an local extremum, the slopes have different sings on the left and the right. A more general theorem on this issue is due to Godunov [67], where it is stated, that a linear monotone method is at most first order accurate. Discussing the specifics and the implications of this theorem, however, is out of the scope of this work and the reader is referred to [115].

A final remark on this issue should be devoted to the physical relevance of the reconstructed states. If the minmod limiter is applied to the conservative variables, it is clear, that the interface values lie in the interval of the cell averages, i.e. $U_i(x^+_{i-\frac{1}{2}}), U_i(x^+_{i-\frac{1}{2}}) \in [U_{i-1}, U_i]$. Therefore, if the physical relevant set is convex, so are the reconstructed states in the physical relevant set. With the interpretation of the piecewise constant reconstruction (2.52), it is a straightforward application of the theorem (2.2.1) to ensure under a suitable CFL condition the physical relevance of the updated states. If the minmod limiter is applied to a different set of variables, further limitations may be applied to ensure the physical relevance

of the states at the interface. If they are physical relevant, the projection (2.54) ensures the physical relevance of the new cell average under a suitable CFL condition. For more detailed information on this see again [17].

The method proposed here is a classical approach to achieve a second order accurate in space method. As well there are numerous other ways to achieve second order [115],[161], there are also methods which reconstruct even higher degrees of polynomials inside a cell to achieve higher order of accuracy. Prominent examples of higher then second order accuracy methods are the piecewise-parabolic method [44] and the essentially non-oscillatory (ENO) methods [79],[76] and weighted ENO (WENO) methods [123], see for example [151] for an overview.

2.3.2 Higher Order in Time

Recall, that during the derivation of the finite volume scheme, one arrives at the coupled system of ODEs. Rewriting (2.6) gives the following form

$$\forall_{i=1}^{N_x} \; U_{i,t} + R(U)_i = 0. \tag{2.56}$$

Discretizing a system of ODEs is a classical problem in numerics and numerous methods have been proposed, see for example [28], [101]. In the derivation of the finite volume scheme, two ways of discretizing the system of ODEs have already been used, i.e. the forward Euler and backward Euler method, resulting in

$$U_i^{n+1} = U_i^n - \Delta_t R(U)_i^n, \tag{2.57}$$

and

$$U_i^{n+1} = U_i^n - \Delta_t R(U)_i^{n+1}. \tag{2.58}$$

A widely used class of higher order integration methods are the so called Runge-Kutta methods. The strategy is to combine multiple forward and backward Euler steps. The general form of an Runge-Kutta method, where the right hand side does not explicitly depend on time, can be written as

$$U_i^{n+1} = U_i^n - \Delta_t \sum_{j=1}^{s} b_j k_j, \tag{2.59}$$

where

$$\forall_{j=1}^{s} \; k_j = R(U(t_n) + \Delta_t(\sum_{l=1}^{s} a_{j,l} k_l))_i. \tag{2.60}$$

The coefficients a, b can be visualized in the Butcher tableau

$$\begin{array}{c|cccc} 0 & a_{1,1} & a_{2,1} & \dots & a_{1,s} \\ c_2 & a_{2,1} & a_{2,2} & \dots & a_{2,s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s,1} & a_{s,2} & \dots & a_{s,s} \\ \hline & b_1 & b_2 & \dots & b_s \end{array} .$$

The Butcher table defines the chosen Runge-Kutta method exhaustively and some properties can bee seen directly. For example, a scheme is explicit, if $a_{j,k} = 0$ for $k \geq l$. For the forward and backward Euler methods the tableau reads

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \qquad\qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}.$$

As been mentioned before, those methods are only first order accurate. Two popular second order schemes are the midpoint rule and Heuns method. Their Butcher tableau reads

$$\begin{array}{c|cc} 0 & & \\ \alpha & \alpha & \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array},$$

where for $\alpha = \frac{1}{2}$ one gets the midpoint rule and for $\alpha = 1$, there is Heuns method. The midpoint rule is used in the MUSCL-Hancock approach [110], where an estimate of the cell interface values has to be computed at half time step, to then define the whole time interval by the flux derived at half time step. In this work, if second order explicit time integration is used, a version of Heuns method is used. To discuss the variant, write out Heuns method as

$$U_i^{n+1} = U_i^n - \frac{\Delta_t}{2}(R(U(t_n)) + R(U(t_{n+1}))). \tag{2.61}$$

This method can be decomposed into two forward Euler steps and a recombination as

$$\begin{cases} \bar{U}_i^{n+1} & = U_i^n - \Delta_t \underbrace{\frac{1}{\Delta_{x_i}}(F(U^n_{i+\frac{1}{2}-}, U^n_{i+\frac{1}{2}+}) - F(U^n_{i-\frac{1}{2}-}, U^n_{i-\frac{1}{2}+}))}_{=R(U(t_n))}, \\ \bar{U}_i^{n+2} & = \bar{U}_i^{n+1} - \Delta_t \underbrace{\frac{1}{\Delta_{x_i}}(F(\bar{U}^{n+1}_{i+\frac{1}{2}-}, \bar{U}^{n+1}_{i+\frac{1}{2}+}) - F(\bar{U}^{n+1}_{i-\frac{1}{2}-}, \bar{U}^{n+1}_{i-\frac{1}{2}+}))}_{=R(U(t_{n+1}))}, \\ U_i^{n+1} & = U_i^n + \frac{1}{2}(\bar{U}_i^{n+2} - U_i^n), \end{cases} \tag{2.62}$$

where the values $\bar{U}_i^{n+1}$,$\bar{U}_i^{n+2}$ are just intermediate values that help to define the total update. Now, the forward Euler steps have to satisfy a CFL criterium for stability. This stability depends on the volume size Δ_{x_i}, but also on the states used to derive the numerical fluxes. In the case of approximate Riemann solvers the wave speeds from the solution at the interface are important to determine the maximal time step. Therefore the optimal, i.e. largest, time step one can choose for the forward Euler steps may be different for each stage. On the other hand, the formulation of the standard Runge-Kutta scheme involves only one time increment Δ_t. Now, one might try to guess the total time step Δ_t in such a way, that in both steps the CFL criterium is satisfied, which may lead to cumbersome calculations. A way around that has been proposed in [14]. The modified Heun method reads

$$\begin{cases} \bar{U}_i^{n+1} & = U_i^n - \frac{\Delta_{t_1}}{\Delta_{x_i}}(F(U^n_{i+\frac{1}{2}-}, U^n_{i+\frac{1}{2}+}) - F(U^n_{i-\frac{1}{2}-}, U^n_{i-\frac{1}{2}+})), \\ \bar{U}_i^{n+2} & = \bar{U}_i^{n+1} - \frac{\Delta_{t_2}}{\Delta_{x_i}}(F(\bar{U}^{n+1}_{i+\frac{1}{2}-}, \bar{U}^{n+1}_{i+\frac{1}{2}+}) - F(\bar{U}^{n+1}_{i-\frac{1}{2}-}, \bar{U}^{n+1}_{i-\frac{1}{2}+})), \\ U_i^{n+1} & = U_i^n + \frac{2\Delta_{t_1}\Delta_{t_2}}{(\Delta_{t_1}+\Delta_{t_2})^2}(\bar{U}_i^{n+2} - U_i^n), \end{cases} \tag{2.63}$$

where the corresponding time increments Δ_{t_1} and Δ_{t_2} can be chosen such that they satisfy the CFL restriction for each forward Euler step. The total time increment $\Delta_t = t_{n+1} - t_n$ is given by

$$\Delta_t = \frac{2\Delta_{t_1}\Delta_{t_2}}{\Delta_{t_1} + \Delta_{t_2}}. \tag{2.64}$$

It should be remarked that, since for $\Delta_{t_1}, \Delta_{t_2} > 0$ there is

$$0 < \frac{2\Delta_{t_1}\Delta_{t_2}}{(\Delta_{t_1} + \Delta_{t_2})^2} < 1. \tag{2.65}$$

Therefore, the final update for U_i^{n+1} is a convex combination of the values $\bar{U}_i^{n+2}$ and U_i^n. Hence, if the scheme used to compute the forward Euler updates respects a convex physical relevant set, the total time integration procedure will share the same property.

As already been mentioned before, explicit time integration is sometimes outperformed by implicit time integration. Also in this work there are simulations where an implicit time integration is crucial to get results in a reasonable amount of time. The method chosen here is an explicit first stage singly diagonally implicit Runge-Kutta (ESDIRK) method. The butcher tableau to this class of methods reads

$$\begin{array}{c|cccc} 0 & 0 & & & \\ c_2 & a_{2,1} & \alpha & & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s,1} & a_{s,2} & \dots & \alpha \\ \hline & b_1 & b_2 & \dots & b_s \end{array} .$$

The first line in the butcher tableau gives that the first stage of the ESDIRK scheme is computed explicitly. The fact that the matrix $a_{l,k}$ has no entries above the diagonal means that the s residuals can be computed one after another and no coupled system has to be solved. The specific method chosen here is the ESDIRK34 scheme, where the first integer denotes the order and the second the number of stages used. For more specifics on the ESDIRK schemes and the implementation see for example [131].

As already mentioned, the drawback of implicit time discretizations is the need to solve maybe nonlinear systems of equations. For simplicity, recall the form of the backward Euler time discretization

$$U_i^{n+1} = U_i^n - \Delta_t R(U)_i^{n+1}. \tag{2.66}$$

Solving this system for the unknowns U_i^{n+1} is equivalent to find the zero of the function Q, where the N_x components of Q are defined as

$$Q(U)_i = U_i^{n+1} - U_i^n + \Delta_t R(U)_i^{n+1}. \tag{2.67}$$

This can be done by the Newton-Raphson method, where the root is found in an iterative way. The new sequence element U_{k+1} is determined by the previous one U_k by solving

$$\frac{\partial}{\partial U} Q(U_k)(U_{k+1} - U_k) = -Q(U_k), \tag{2.68}$$

where $\frac{\partial}{\partial U} Q(U)$ is the Jacobian of Q with respect to the sequence element U_k.

There are many ways to solve the system (2.68). Popular examples are the Krylov-subspace and multigrid methods , see [129] and references therein. As the choice of the right method for a specific problem and the efficiency gains due to for example preconditioning are crucial for the succesful use of such methods, to deal with the specifics of such solvers is out of the scope of this work. In fact, those methods will be used, since the schemes from chapter 5 and chapter 6 are implemented in the Seven-League-Hydro (SLH) code, where the design of that code and the use of these iterative methods is described in [59, 131].

2.4 Finite Volume Approach for Balance Laws

Consider now a balance law of the form

$$u_t + f(u)_x = S(u). \tag{2.69}$$

A common approach to discretize (2.69) is the fractional splitting method. First, the system gets split into the conservative part and the non-homogeneous part as

$$\begin{cases} u_t + f(u)_x = 0, \\ u_t = S(u). \end{cases} \tag{2.70}$$

Now, the two equations are discretized separately as for example

$$\begin{cases} U_{i,t} + \frac{1}{\Delta_{x_i}}(F_i(U_1, ..., U_{N_x}, x_{i+\frac{1}{2}}) - F_i(U_1, ..., U_{N_x}, x_{i-\frac{1}{2}})) = 0, \\ U_{i,t} = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} S(u)dx. \end{cases} \tag{2.71}$$

The two operators are applied in an alternate way to evolve the cell averages U_i. Two well know methods are the Godunov and the Strang splitting, see [115] for more details. However, in some applications the resolution of near equilibrium solutions is of major importance. Schemes that are consistent with an equilibrium of (2.69) are called well-balanced schemes. The next definition helps to specify what is understood by the term consistent.

Definition 2.4.1. *Let u_{eq} be an equilibrium solution to (2.69), i.e. it satisfies*

$$f(u_{eq})_x = S(u_{eq}). \tag{2.72}$$

Consider now a discretization of u_{eq} as U_{eq} and a numerical scheme approximating (2.69) as

$$D_t(U) + D_x(f(U)) = D_S(S(U)). \tag{2.73}$$

A discretization (2.73) is called well-balanced if for the discrete equilibrium U_{eq} the discrete time derivative $D_t(U_{eq})$ vanishes, i.e.

$$D_x(f(U_{eq})) = D_S(S(U_{eq})). \tag{2.74}$$

It is not obvious that the fractional splitting method (2.71) will satisfy the requirement (2.74). In general, it will not. Therefore, if one applies the method (2.71) to a state somehow close to the discrete equilibrium U_{eq}, the numerical scheme will produce unphysical oscillations, see for example [62],[114].

The main issue is that the balance of the flux and the source term is not considered in the splitting approach, since the two parts are computed separately. Another approach to achieve a numerical method satisfying the definition 2.4.1 is to consider the inhomogeneous Riemann problem at the cell interfaces arising from the (maybe modified) complete system (2.69) as suggested by [72],[71],[42],[69],[70]. First, one can rederive the finite volume scheme as in the homogeneous case to arrive at

$$U_{i,t} + \frac{1}{\Delta_{x_i}}(F_i(U_1, ..., U_{N_x}, x_{i+\frac{1}{2}}) - F_i(U_1, ..., U_{N_x}, x_{i-\frac{1}{2}})) = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} S_i(U_1, ..., U_{N_x}, x_i)dx. \tag{2.75}$$

The numerical flux functions F_i and the numerical source term S_i are now evaluated with respect to the inhomogeneous Riemann problem at the cell interfaces. To achieve the resolution of the inhomogeneous Riemann problem, similar to section 1.2, one can multiply the source term by the derivative of the function $a(x) = x$ to arrive at the following form

$$\begin{cases} u_t + f(u)_x = S(u)a_x, \\ a_t = 0. \end{cases} \tag{2.76}$$

However, this work concentrates on the numerical treatment of the Shallow Water equations and the Euler equations of gas dynamic. Both system are already in the form (2.76) when writing

$$\begin{cases} u_t + f(u)_x = S(u)Z_x, \\ Z_t = 0. \end{cases} \tag{2.77}$$

and Z is either the topography term or the gravitational potential. Now, similar to the case of conservation laws, one projects the data onto piecewise constant data. At the cell interface $x_{i+\frac{1}{2}}$, an inhomogeneous Riemann problem arises with respect to the system (2.77) and initial conditions

$$u(0,x) = \begin{cases} U_i^n \text{ if } x < 0, \\ U_i^{n+1} \text{ if } x > 0, \end{cases} \qquad Z(0,x) = \begin{cases} Z_i \text{ if } x < 0, \\ Z_{i+1} \text{ if } x > 0. \end{cases} \tag{2.78}$$

At this point, it should be remarked that the general resolution of the inhomogeneous Riemann problem involving the system (2.77) might be hard due to the non-conservative product $S(u)Z_x$, because at the interface Z as well as $S(u)$ might be discontinuous and in general this product is not defined in this case. A general way to give meaning to such a product can be found in [128]. Here, the authors regularize the non-conservative product by defining a parametrization of a path for $S(u)$ to take through the discontinuity. However, it is shown that the value of the non-conservative product depends on the choice of such a path. While for conservative terms the choice of the path is irrelevant, the authors can not provide a criterium to choose such a path in the non-conservative case. Despite that lack of uniqueness, this theory has been used to develop the so called path-conservative schemes, see for example [143],[142]. The authors use the given degree of freedom to specify a path, which will guarantee the well-balanced property. However, the theory is not complete and there are also practical examples, where path conservative schemes fail to converge or give wrong shock speeds, see [30],[1].

Despite the difficulties in deriving the exact solution to the inhomogeneous Riemann problem, assume that one knows an (approximate) solution $\mathcal{W}(t,x)$ as in section 2.2. Define the numerical fluxes as

$$\begin{cases} F^-_{i+\frac{1}{2}} = f(\mathcal{W}_{i+\frac{1}{2}}(t, x^-_{i+\frac{1}{2}})), \\ F^+_{i+\frac{1}{2}} = f(\mathcal{W}_{i+\frac{1}{2}}(t, x^+_{i+\frac{1}{2}})). \end{cases} \tag{2.79}$$

The numerical scheme can be rewritten in the following form

$$U_{i,t} + \frac{1}{\Delta_{x_i}}(F^+_{i-\frac{1}{2}} - F^-_{i+\frac{1}{2}}) = \frac{1}{\Delta_{x_i}} \int_{x_{i-\frac{1}{2}}}^{x_i} S_i(\mathcal{W}(t,x)_{i-\frac{1}{2}}) Z_x dx + \frac{1}{\Delta_{x_i}} \int_{x_i}^{x_{i+\frac{1}{2}}} S_i(\mathcal{W}(t,x)_{i+\frac{1}{2}}) Z_x dx, \tag{2.80}$$

while $\mathcal{W}(t,x)_{i-\frac{1}{2}}$ denotes the solution to the Riemann problem at the interface $x_{i-\frac{1}{2}}$. Two things are important to realize in the formulation (2.80). First, since Z is constant in each cell, the integrals on the right hand side vanishes and one is left with

$$U_{i,t} + \frac{1}{\Delta_{x_i}}(F^+_{i-\frac{1}{2}} - F^-_{i+\frac{1}{2}}) = 0. \tag{2.81}$$

Second, the scheme is not anymore in conservation form, since, due to the non-conservative wave at the cell interface, $F^-_{i+\frac{1}{2}} \neq F^+_{i+\frac{1}{2}}$. This should not be surprising, since the underlying equations are also not in conservation form and the numerical scheme at this point just reflects that property.

In the end, the form of the numerical scheme (2.81) is beneficial for determining the well-balanced property. Suppose that the initial condition in the Riemann problem (2.78) satisfies some discrete version of an equilibrium to (2.77). Therefore, if the solution to the inhomogeneous Riemann problem also respects this equilibrium relation, the solution can be written as

$$\begin{cases} W(t, x^-_{i+\frac{1}{2}}) &= U_i, \\ W(t, x^+_{i+\frac{1}{2}}) &= U_{i+1}. \end{cases} \tag{2.82}$$

Using the definition of the numerical fluxes (2.79) then gives directly $U_{i,t} = 0$ and therefore this approach satisfies the well-balanced definition 2.4.1. The well-balanced approximate Riemann solver presented in this work all rely on achieving the property (2.82). Moreover, in this section it is not specified if an explicit or an implicit time discretization is chosen. Therefore, if the property (2.82) can be shown for an approximate Riemann solver, it guarantees the well-balanced property for an explicit as well as for an implicit scheme.

2.5 Finite Volume schemes in 2 space dimensions

The final section in this chapter is concerned on how to set up a finite volume scheme in 2 space dimensions. Consider the conservation law

$$u_t + f_{1,x} + f_{2,y} = 0. \tag{2.83}$$

Consider now a cartesian mesh in two space dimensions as

$$V_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]. \tag{2.84}$$

Similar to section 2.4, a popular approach to discretize the system (2.83) is to split the operators, while here it is called a dimensional split.

$$\begin{cases} u_t & +f_{1,x} = 0, \\ u_t & +f_{2,y} = 0. \end{cases} \tag{2.85}$$

Now, one can evolve the data given in each volume $V_{i,j}$ according to a discretization to each system in (2.85). The Godunov and Strang splitting, are again two popular strategies on how to combine the two discretizations. As long as the combination of the two systems is convex, the robustness and stability of a finite volume scheme directly translates from the one dimensional case, since in every substep only one dimensional problems are concerned. However, the method has its drawbacks, especially in the case when the volumes $V_{i,j}$ are not rectangular. Even though this case is not considered in this work, it is desired that the results may also be applicable for more complex meshes.

Alternatively to the splitting approach, the full integral over the volume $V_{i,j}$ is considered to derive the finite volume scheme. With similar approximations of the boundary terms as in section 2.1, one arrives at the following form

$$U_{i,j}^{n+1} = U_i^n - \frac{\Delta_t}{\Delta_{x_i}} \left(F_{1,i+\frac{1}{2},j} - F_{1,i-\frac{1}{2},j} \right) - \frac{\Delta_t}{\Delta_{y_i}} \left(F_{2,i,j+\frac{1}{2}} - F_{2,i,j-\frac{1}{2}} \right). \tag{2.86}$$

The numerical flux functions can be defined by using an approximate Riemann solver as described in section 2.2. Consider for this the 2 Riemann problems

$$\begin{aligned} &u_t + f_{1,x} = 0, \\ &u(0,x,y) = \begin{cases} U_{i-1,j} & \text{if } x < 0, \\ U_{i,j} & \text{if } x > 0, \end{cases} \end{aligned}$$

and

$$\begin{aligned} &u_t + f_{1,y} = 0, \\ &u(0,x,y) = \begin{cases} U_{i,j-1} & \text{if } y < 0, \\ U_{i,j} & \text{if } y > 0, \end{cases} \end{aligned}$$

and denote their solution as $\mathcal{W}_{i-\frac{1}{2},j}$ and $\mathcal{W}_{i,j-\frac{1}{2}}$ respectively. Then the numerical fluxes in (2.86) are defined by

$$\begin{aligned} F_{1,i-\frac{1}{2}^+,j} &= \mathcal{W}_{i-\frac{1}{2},j}(t,0^+,y) \qquad & F_{1,i+\frac{1}{2}^-,j} &= \mathcal{W}_{i+\frac{1}{2},j}(t,0^-,y), \\ F_{2,i,j-\frac{1}{2}^+} &= \mathcal{W}_{i,j-\frac{1}{2}}(t,x,0^+) \qquad & F_{2,i,j+\frac{1}{2}^-} &= \mathcal{W}_{i,j+\frac{1}{2}}(t,x,0^-), \end{aligned} \tag{2.87}$$

and the numerical scheme writes

$$U_{i,j}^{n+1} = U_i^n - \frac{\Delta_t}{\Delta_{x_i}} \left(F_{1,i+\frac{1}{2}^-,j} - F_{1,i-\frac{1}{2}^+,j} \right) - \frac{\Delta_t}{\Delta_{y_i}} \left(F_{2,i,j+\frac{1}{2}^-} - F_{2,i,j-\frac{1}{2}^+} \right). \tag{2.88}$$

The conservation property from definition 2.1.1 directly translates to the scheme (2.88). The robustness and stability of this approach can be derived from the respective one dimensional approaches by adjusting the CFL number. Rewrite (2.88) into the following two step form

$$U_{i,j}^{n+1} = \frac{1}{2} \left(\bar{U}_{\Delta i,j} + \bar{U}_{i,\Delta j} \right), \tag{2.89}$$

where

$$\begin{aligned} \bar{U}_{\Delta i,j} &= U_i^n - 2\frac{\Delta_t}{\Delta_x} \left(F_{1,i+\frac{1}{2}^-,j} - F_{1,i-\frac{1}{2}^+,j} \right), \\ \bar{U}_{i,\Delta j} &= U_i^n - 2\frac{\Delta_t}{\Delta_y} \left(F_{2,i,j+\frac{1}{2}^-} - F_{2,i,j-\frac{1}{2}^+} \right). \end{aligned} \tag{2.90}$$

In other words, the scheme (2.88) can be rewritten as a convex combination of one dimensional schemes of the type (2.89). Therefore, if robustness and stability can be proven for the one dimensional schemes, it holds also for the two dimensional scheme. However observe that the CFL condition has to be adopted in this case to be $\frac{1}{4}$. Following the lines of [16] and [20], these results also apply if a MUSCL approach is used to achieve second order in space and can be extended to unstructured meshes as well, while further adjustment to the CFL condition is needed.

2.6 Boundary Conditions

In contrast to the a theoretical analysis of PDEs, where infinitely large domains are allowed, in numerical applications the domain $\mathcal{D}$ on which the approximations are computed is finite. So one has to deal with the values of the solution on the boundary of $\mathcal{D}$, denoted as $\partial\mathcal{D}$. Therefore the initial value problem (1.3) is extended as follows

$$\begin{cases} u(t,x)_t + \nabla \cdot f(u) = 0, \\ u(0,x) = u_0(x), \\ \forall x \in \partial\mathcal{D}: \quad u(t,x) = \bar{u}(t,x). \end{cases} \tag{2.91}$$

It is clear, that the definition of the boundary conditions can have a significant influence on the solution. To study these effects is out of the scope of this work. Here, some of the classical boundary conditions are used. To this end, for simplicity, consider the case of $n = 1$. The domain $\mathcal{D}$ is then given by the simple interval $\mathcal{D} = [x_L, x_R]$. In this work, three types

of boundary conditions are used. The first are the periodic boundary conditions given as

$$u(t, x_L) = u(t, x_R). \tag{2.92}$$

The second type of boundary conditions are the so called Neuman boundary conditions

$$\begin{cases} u(t, x_L)_x = 0, \\ u(t, x_R)_x = 0. \end{cases} \tag{2.93}$$

The third type of boundary conditions used in this work are the solid wall boundary conditions. They involve the velocity of the fluid, denoted here as $\mathbf{u}$. They are formulated as

$$\begin{cases} \mathbf{u}(t, x_L) = 0, \\ \mathbf{u}(t, x_R) = 0. \end{cases} \tag{2.94}$$

For a general domain the solid wall boundary conditions are formulated as $\mathbf{n} \cdot \mathbf{u} = 0$, where $\mathbf{n}$ is the unit outward normal of $\partial\mathcal{D}$. It basically states that there is no fluid flowing out or into the domain.

The numerical treatment of boundary conditions is usually done by introducing so called ghost cells. These ghost cells are extensions of the computational domain. Consider a family of volumes $V_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ covering $\mathcal{D}$. Denote the cell at the left boundary as $V_1 = [x_L, x_{1+\frac{1}{2}}]$ and the cell at the right boundary as $V_{N_x} = [x_{N_x-\frac{1}{2}}, x_R]$. Denoting the spatial order of the scheme as p, then p ghost cells are imposed as well as on the left as on the right side of the boundary as

$$\begin{cases} \forall_{k=1}^{p} & V_{1-k} & = [x_L - k\Delta_x, x_L - (k-1)\Delta_x], \\ \forall_{k=1}^{p} & V_{N_x+k} & = [x_R + (k-1)\Delta_x, x_R + k\Delta_x]. \end{cases} \tag{2.95}$$

Depending on the type of boundary conditions, numerical approximations $\bar{U}_i$ are imposed on the ghost cells depending on the values U_i inside the computational domain. For periodic boundary conditions there is

$$\begin{cases} \forall_{k=1}^{p} & \bar{U}_{1-k} & = U_{N_x-k+1}, \\ \forall_{k=1}^{p} & \bar{U}_{N_x+k} & = U_k. \end{cases} \tag{2.96}$$

For Neumann boundary conditions the ghost cells are set as

$$\begin{cases} \forall_{k=1}^{p} & \bar{U}_{1-k} & = U_1, \\ \forall_{k=1}^{p} & \bar{U}_{N_x+k} & = U_{N_x}. \end{cases} \tag{2.97}$$

In case of a solid wall boundary, one starts with the same procedure as for the Neumann boundary conditions. But now the velocity components in the ghost cells have to be modified in order to achieve the condition (2.94). Consider for this the velocity component in the cell V_1 as $\mathbf{u}_1$. This velocity can be decomposed into an orthogonal component $\mathbf{u}_1^{\perp}$ and a parallel component $\mathbf{u}_1^{\parallel}$ with respect to the interface. The fluid velocities are then set as

$$\begin{cases} \forall_{k=1}^{p} & \mathbf{u}_{1-k} & = -\mathbf{u}_1^{\perp} + \mathbf{u}_1^{\parallel}, \\ \forall_{k=1}^{p} & \mathbf{u}_{N_x+k} & = -\mathbf{u}_{N_x}^{\perp} + \mathbf{u}_{N_x}^{\parallel}. \end{cases} \tag{2.98}$$

These boundary conditions are then used to determine the fluxes $F^{+}_{1-\frac{1}{2}}$ and $F^{-}_{N_x+\frac{1}{2}}$ according to the chosen numerical method.

It should be remarked, that the extension of these methods to 2 space dimensions can be difficult. Especially if complex geometries are considered. However in this work, only cartesian meshes are considered where the described methods extend intuitively.

3 A Well-Balanced HLL Type Scheme for the Shallow Water Equations

This chapter is concerned with the numerical approximation to the Shallow Water equations introduced in section 1.5. The focus lies especially on the resolution of near equilibrium solutions. The aim is to show that a scheme that satisfies the well-balanced property given in definition 2.4.1 is superior to non well-balanced schemes especially when numerical approximations of near equilibrium solutions are concerned. However, robustness and entropy stability are not considered in this chapter.

The derivation of well-balanced schemes for the Shallow Water equations has been an active area of research since the pioneering work in [72] and [13]. Numerous schemes have been developed to capture the Lake at Rest solutions (1.87), see for example [6],[118],[87],[26], [69], [27],[125],[140],[143],[32],[149],[23]. The Lake at Rest solutions however are only a subclass of the general equilibria (1.86). Further research has been done to achieve the well-balanced property also in the extended case of non zero velocities, see [31],[61],[170],[171], [172],[173],[19],[41]. Especially in [172], the superiority of a well-balanced scheme for the general case over schemes which only consider the Lake at Rest solution has been shown. However, the proposed scheme involves a sophisticated splitting at the interface. Here the aim is to derive such a scheme in a more simpler manner. Even more, especially when higher order extensions are considered, the scheme presented in [172] relies on solving a third order equation by using an iterative Newton method, which for one is costly. Additionally, due to the fact that this third order equations has multiple zeros, the convergence to the right zero is not obvious and therefore clever initial values for the iteration have to be chosen.

In the approach presented here, a HLL type scheme, see section 2.2.2, previous applied to the Euler equations with gravity, see [54], is applied to the Shallow Water equations and adapted to the case of general equilibria. The scheme presented here is close to a numerical scheme derived in [19] for subcritical flow. However, the publication lacks numerical results to show the performance of the scheme. Numerical results are presented at the end of the chapter. Additionally, the technical issues in going higher order are avoided by solving the third order equation exactly and an exhaustive analysis of the roots is presented.

3.1 HLL-type Schemes for the Shallow Water equations

Consider the Shallow water equations as presented in section 1.5 in one space dimension

$$\begin{cases} h_t + (hu)_x = 0, \\ (hu)_t + \left(hu^2 + g\frac{h^2}{2}\right)_x = -ghB_x. \\ B_t = 0 \end{cases} \tag{3.1}$$

The system is hyperbolic with eigenvalues $\lambda_{1,2} = u \pm \sqrt{gh}$. If the eigenvalues $\lambda_{1,2}$ are of different sign, the flow is considered to be subcritical. If the eigenvalues are of same sign,

the flow is called supercritical and if one of the eigenvalues is zero, the flow is called critical. Of special interest are the equilibria of the system. They can be computed by setting the time derivative to 0 to get

$$\begin{cases} hu = const, \\ \frac{u^2}{2} + g(h+B) = const. \end{cases} \tag{3.2}$$

In the following, for brevity, it is sometimes used that $E = \frac{u^2}{2} + g(h+B)$. As has already been mentioned, a subclass of (3.2) are the so called Lake at Rest solutions. They are derived by setting in (3.2) $u = 0$ and one has

$$\begin{cases} u = 0, \\ h + B = const. \end{cases} \tag{3.3}$$

The aim is to derive a well-balanced scheme for the general equilibria (3.2). In fact, for the sake of comparison, a second scheme is derived, which is only consistent with the Lake at Rest solutions.

3.1.1 Choice of the wave speeds

Consider the model of an approximate Riemann solver of the type (2.30). In a first step, the number and the values of the artificial wave speeds have to be determined. Here it is decided to work with two waves modeling the dynamics form the homogeneous part of the system (3.1), denoted by λ_L and λ_R and one wave coming from the source term, denoted by λ_0.

Following the arguments of section 1.2 the presence of the source term leads to a standing wave, i.e. it is set

$$\lambda_0 = 0. \tag{3.4}$$

The choice for the waves λ_L, λ_R follows the classical HLL framework for conservation laws. Compute first

$$\begin{aligned} \lambda_{R,\pm} &= u_R \pm \sqrt{g h_R}, \\ \lambda_{L,\pm} &= u_L \pm \sqrt{g h_L}. \end{aligned}$$

Then the artificial wave speeds are determined by

$$\begin{aligned} \bar{\lambda}_R &= \max\left(\lambda_{L,+}, \lambda_{R,+}\right), \\ \bar{\lambda}_L &= \min\left(\lambda_{L,-}, \lambda_{R,-}\right). \end{aligned} \tag{3.5}$$

However, for stability reasons, the wave speeds are usually rescaled in the following way

$$\lambda_R = \begin{cases} \delta_1 \bar{\lambda}_R, & \text{if } \bar{\lambda}_R > 0, \\ \frac{\bar{\lambda}_R}{\delta_2}, & \text{if } \bar{\lambda}_R < 0, \end{cases} \qquad \text{and} \qquad \lambda_L = \begin{cases} \delta_1 \bar{\lambda}_L & \text{, if } \bar{\lambda}_L < 0, \\ \frac{\bar{\lambda}_L}{\delta_2}, & \text{if } \bar{\lambda}_L > 0, \end{cases} \tag{3.6}$$

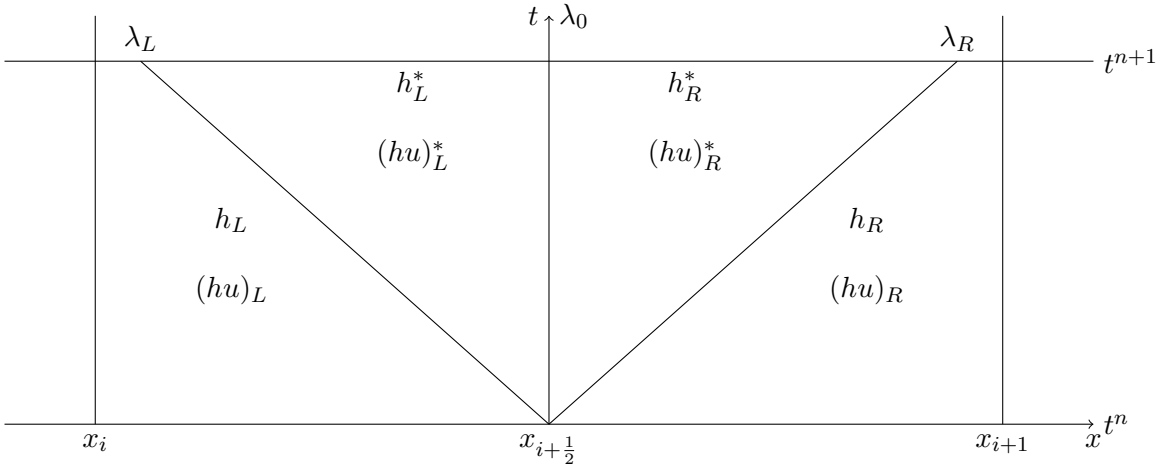


Fig. 3.1: Structure of the HLL type approximate Riemann solver $\mathcal{W}(t, x)$ in the subcritical case.

for some $\delta_1, \delta_2 > 1$. Since the wave speed of the source term is fixed, it is not clear how the waves are ordered. There are three possibilities.

- The subcritical case: $\lambda_L < \lambda_0 < \lambda_R$
- The critical case: $\lambda_L = \lambda_0 < \lambda_R$ or $\lambda_L < \lambda_0 = \lambda_R$
- The supercritical case: $\lambda_0 < \lambda_L < \lambda_R$ or $\lambda_L < \lambda_R < \lambda_0$

Different strategies have to be applied in all three cases. In section 3.1.2 the scheme for the subcritical case is derived. Section 3.1.3 deals with the supercritical case and in section 3.1.4 the critical case is considered.

3.1.2 The HLL-type Model for subcritical Flow

Consider the following model of an approximate Riemann solver of the type (2.30) for the subcritical case

$$\mathcal{W}_{sub}(t, x) = \begin{cases} w_L & \text{if } \frac{x}{t} < \lambda_L, \\ w_L^* & \text{if } \lambda_L < \frac{x}{t} < \lambda_0, \\ w_R^* & \text{if } \lambda_0 < \frac{x}{t} < \lambda_R, \\ w_R & \text{if } \lambda_R < \frac{x}{t}, \end{cases} \tag{3.7}$$

where the vector of dependent variables is $w = (h, hu, B)$. Since the evolution of the bottom topography B is decoupled from the system, one directly has that

$$B_L^* = B_L \qquad B_R^* = B_R. \tag{3.8}$$

Therefore the remaining unknowns are $h_L^*, h_R^*, (hu)_L^*, (hu)_R^*$, see also figure 3.1.

As introduced in section 2.2.2, the strategy to solve for the unknowns is to make use of the consistency relation (2.33), which writes in this case

$$\frac{1}{\Delta x}\int_{x_i}^{x_{i+1}} \mathcal{W}\left(\frac{x}{t^{n+1}}, w_L, w_R\right) dx = \frac{1}{\Delta x}\int_{x_i}^{x_{i+1}} W\left(\frac{x}{t^{n+1}}, w_L, w_R\right) dx, \tag{3.9}$$

where W is the exact solution to the Riemann problem. Following the notions in section 2.2.2, the evaluation of the two components of (3.9) gives for the waterheight

$$\left(\frac{1}{2} + \lambda_L \frac{\Delta t}{\Delta x}\right) h_L - \lambda_L \frac{\Delta t}{\Delta x} h_L^\star + \lambda_R \frac{\Delta t}{\Delta x} h_R^\star + \left(\frac{1}{2} - \lambda_R \frac{\Delta t}{\Delta x}\right) h_R = \frac{h_L + h_R}{2} - \frac{\Delta t}{\Delta x}(h_R u_R - h_L u_L), \tag{3.10}$$

and for the discharge

$$\left(\frac{1}{2} + \lambda_L \frac{\Delta t}{\Delta x}\right) (hu)_L - \lambda_L \frac{\Delta t}{\Delta x} (hu)_L^\star + \lambda_R \frac{\Delta t}{\Delta x} (hu)_R^\star + \left(\frac{1}{2} - \lambda_R \frac{\Delta t}{\Delta x}\right) (hu)_R = \frac{h_L u_L + h_R u_R}{2} - \frac{\Delta t}{\Delta x}(h_R u_R^2 + g\frac{h_R^2}{2} - h_L u_L^2 - g\frac{h_L^2}{2}) - \frac{1}{\Delta x}\int_{x_i}^{x_{i+1}}\int_{t^n}^{t^{n+1}} gh\left(\frac{x}{t}, w_L, w_R\right) B_x dt dx, \tag{3.11}$$

where $h\left(\frac{x}{t}, w_L, w_R\right)$ denotes the exact solution for the waterheight to the Riemann problem. A first step is to approximate the integral in the momentum equation due to the source term by a quadrature

$$\frac{1}{\Delta x}\int_{x_i}^{x_{i+1}}\int_{t^n}^{t^{n+1}} h\left(\frac{x}{t}, w_L, w_R\right) B_x dt dx = \frac{\Delta_t}{\Delta_x}\overline{ghB_x}. \tag{3.12}$$

As it will turn out, the evaluation of this approximation will strongly influence the well-balanced properties. How this quadrature term is evaluated will be discussed in section 3.1.5.

The equations (3.10) and (3.11) only give two relations for the 4 unknowns $h_L^*, h_R^*, (hu)_L^*$ and $(hu)_R^*$ and additional equations have to be imposed. To find the missing two relations, the equilibrium relations are imposed on the λ_0 wave to have

$$(hu)_R^\star = (hu)_L^\star, \tag{3.13}$$

$$\left(\frac{1}{2}(\frac{(hu)_L^*}{h_L^*})^2 + g(h_L^* + B_L)\right) = \left(\frac{1}{2}(\frac{(hu)_R^*}{h_R^*})^2 + g(h_R^* + B_R)\right). \tag{3.14}$$

Observe that equation (3.14) is non-linear. In fact, if equation (3.14) is used to solve for the intermediate states, the roots of a fifth order polynomial have to be found, where in general there is no explicit expression for these roots. However, it has been proven in [19], that, if equation (3.14) is used, under some conditions on the wave speeds λ_L and λ_R, the scheme would be robust and entropy stable. The goal here is to develop a practical scheme

and therefore it is suggested to linearize (3.14) in the following way

$$\left(\frac{1}{2}(\frac{(hu)_L^*}{h_L})^2 + g(h_L^* + B_L)\right) = \left(\frac{1}{2}(\frac{(hu)_R^*}{h_R})^2 + g(h_R^* + B_R)\right). \tag{3.15}$$

The exact impact of the linearization is difficult to analyze. However, we would like to remark that in the case of an equilibrium solution, the solution to $h_{L,R}^*$ should be $h_{L,R}$. In this case, the linearization is actually exact and no error is introduced.

The solutions to the intermediate states can now be found when using (3.10),(3.11),(3.13) and (3.15) and are given by

$$h_L^\star = \frac{\lambda_R h_R - \lambda_L h_L - \lambda_R D + (hu)_L - (hu)_R}{\lambda_R - \lambda_L}, \tag{3.16}$$

$$h_R^\star = \frac{\lambda_R h_R - \lambda_L h_L - \lambda_L D + (hu)_L - (hu)_R}{\lambda_R - \lambda_L}, \tag{3.17}$$

$$(hu)_{L,R}^\star = \frac{\lambda_R h_R u_R - \lambda_L h_L u_L - (h_R u_R^2 + g\frac{h_R^2}{2} - h_L u_L^2 - g\frac{h_L^2}{2}) - \overline{ghB_x}}{\lambda_R - \lambda_L}, \tag{3.18}$$

where

$$D = \frac{(hu)^{*2}}{2g}(\frac{1}{h_L^2} - \frac{1}{h_R^2}) + (B_L - B_R). \tag{3.19}$$

The solutions to the intermediate states are somehow stable as long as $(\lambda_R - \lambda_L)$ is not small or when h_L and h_R are sufficiently large. Both cases relate to a regime with zero or only small waterheight. However, the case of wet/dry areas is not considered in this chapter.

3.1.3 The Supercritical Case

This section is devoted to the derivation of an HLL type scheme in the supercritical case. For symmetry reasons it is enough to analyze the case $\lambda_{L,R} > 0$. The model (3.7) is therefore modified as

$$\mathcal{W}_{sup}(t,x) = \begin{cases} w_L & \text{if } \frac{x}{t} < \lambda_0, \\ w_L^* & \text{if } \lambda_0 < \frac{x}{t} < \lambda_L, \\ w_R^* & \text{if } \lambda_L < \frac{x}{t} < \lambda_R, \\ w_R & \text{if } \lambda_R < \frac{x}{t}. \end{cases} \tag{3.20}$$

As in the subcritical case, the evolution of the bottom topography B is decoupled from the system and one directly has

$$B_L^* = B_L \qquad B_R^* = B_R, \tag{3.21}$$

and the remaining unknowns are $h_L^*, h_R^*, (hu)_L^*, (hu)_R^*$, see also figure 3.2. However, by the model (3.20), the flux for the cell V_i is already determined by the initial condition. To get the flux for the cell V_{i+1} one only has to solve for $h_L^*, (hu)_L^*$.

Similar to the subcritical case, the consistency relations are applied to get for the water-height

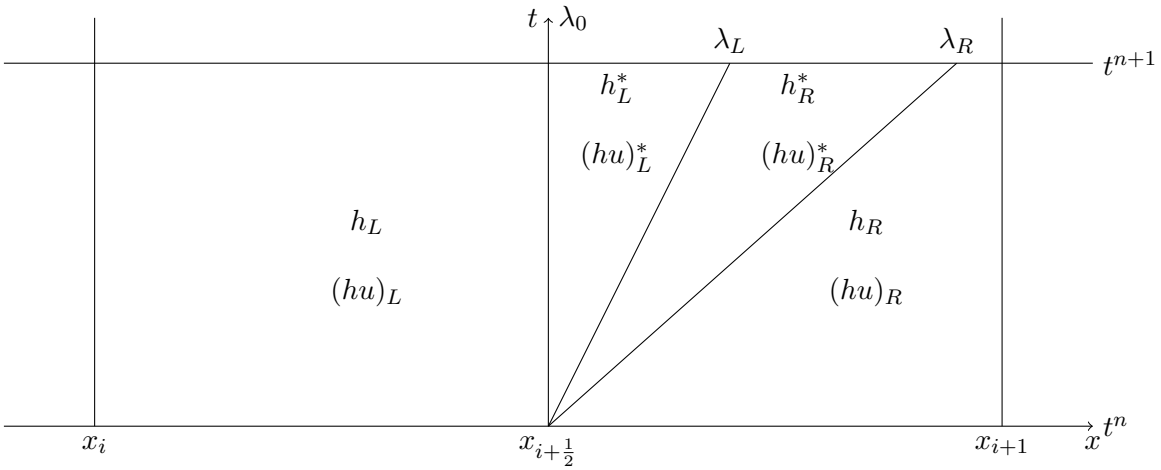


Fig. 3.2: Structure of the HLL type approximate Riemann solver $\mathcal{W}(t,x)$ for supercritical flows.

$$\frac{1}{2}h_L+\lambda_L\frac{\Delta t}{\Delta x}h_L^\star+(\lambda_R-\lambda_L)\,\frac{\Delta t}{\Delta x}h_R^\star+\left(\frac{1}{2}-\lambda_R\frac{\Delta t}{\Delta x}\right)h_R=\frac{h_L+h_R}{2}-\frac{\Delta t}{\Delta x}(h_Ru_R-h_Lu_L), \tag{3.22}$$

and for the discharge

$$\frac{1}{2}(hu)_L+\lambda_L\frac{\Delta t}{\Delta x}(hu)_L^\star+(\lambda_R-\lambda_L)\,\frac{\Delta t}{\Delta x}(hu)_R^\star+\left(\frac{1}{2}-\lambda_R\frac{\Delta t}{\Delta x}\right)(hu)_R=$$
$$\frac{h_Lu_L+h_Ru_R}{2}-\frac{\Delta t}{\Delta x}(h_Ru_R^2+g\frac{h_R^2}{2}-h_Lu_L^2-g\frac{h_L^2}{2})-\frac{\Delta_t}{\Delta x}\overline{ghB_x}, \tag{3.23}$$

where the source term already has been averaged. Two more equations are needed to solve for the unknowns. It is suggested to impose the equilibrium equations across the λ_0 wave. Therefore, the following relations are imposed

$$(hu)_L=(hu)_L^\star, \tag{3.24}$$

$$\left(\frac{1}{2}(\frac{(hu)_L}{h_L})^2+g(h_L+B_L)\right)=\left(\frac{1}{2}(\frac{(hu)_L^*}{h_L^*})^2+g(h_L^*+B_R)\right). \tag{3.25}$$

Observe, that the equilibrium relations (3.24) and (3.25) allow to directly compute the dependent variables on the right side of the cell interface. Equation (3.24) gives directly the discharge. Therefore, equation (3.25) can be rewritten as a third order polynomial in h_L^* to get

$$h_L^{*^3}+\frac{(gB_R-E_L)}{g}h_L^{*^2}+\frac{(hu)_L^2}{2g}=0. \tag{3.26}$$

How to solve for the roots of (3.26) is discussed in detail in section 3.3.

Even though they are not needed for the definition of the numerical fluxes, the solutions to the other unknowns are given for completion. There is for the waterheight

$$h_R^\star = \frac{(\lambda_R h_R - \lambda_L h_L^\star) - (h_R u_R - h_L u_L)}{\lambda_R - \lambda_L}, \tag{3.27}$$

and for the discharge

$$(hu)_R^\star = \frac{\lambda_R (hu)_R - \lambda_L (hu)_L - (h_R u_R^2 + g\frac{h_R^2}{2} - h_L u_L^2 - g\frac{h_L^2}{2}) - \overline{ghB_x}}{\lambda_R - \lambda_L}. \tag{3.28}$$

However, if one would like to do an analysis regarding the robustness and the stability of this approach, following the theorems 2.2.1 and 2.2.2, the values defined in (3.27) and (3.28) are important as well. But this kind of analysis is omitted for the schemes described in this section.

3.1.4 The Critical Case

Finally, consider the transcritical case, i.e. $\lambda_L = \lambda_0 = 0$. Again, for symmetry reasons, the case $\lambda_R = 0$ is omitted here. The model for the HLL type scheme reads then

$$\mathcal{W}_{crit}(t,x) = \begin{cases} w_L & \text{if } \frac{x}{t} < \lambda_0, \\ w^* & \text{if } \lambda_0 < \frac{x}{t} < \lambda_R, \\ w_R & \text{if } \lambda_R < \frac{x}{t}, \end{cases} \tag{3.29}$$

see also figure 3.3. The model (3.29) only admits two unknowns and the application of the consistency relations is sufficient to find the solution. Similar to (3.22) and (3.23) one gets

$$\frac{1}{2}h_L + \lambda_R \frac{\Delta t}{\Delta x} h^\star + \left(\frac{1}{2} - \lambda_R \frac{\Delta t}{\Delta x}\right) h_R = \frac{h_L + h_R}{2} - \frac{\Delta t}{\Delta x}(h_R u_R - h_L u_L), \tag{3.30}$$

and for the discharge

$$\begin{aligned} \frac{1}{2}(hu)_L + \lambda_R \frac{\Delta t}{\Delta x}(hu)^\star + \left(\frac{1}{2} - \lambda_R \frac{\Delta t}{\Delta x}\right)(hu)_R = \\ \frac{h_L u_L + h_R u_R}{2} - \frac{\Delta t}{\Delta x}(h_R u_R^2 + g\frac{h_R^2}{2} - h_L u_L^2 - g\frac{h_L^2}{2}) - \frac{\Delta_t}{\Delta x}\overline{ghB_x}. \end{aligned} \tag{3.31}$$

Therefore, the intermediate states can be directly computed as

$$h^\star = h_R - \frac{h_R u_R - h_L u_L}{\lambda_R}, \tag{3.32}$$

$$(hu)^\star = (hu)_R - \frac{(h_R u_R^2 + g\frac{h_R^2}{2} - h_L u_L^2 - g\frac{h_L^2}{2}) + \overline{ghB_x}}{\lambda_R}. \tag{3.33}$$

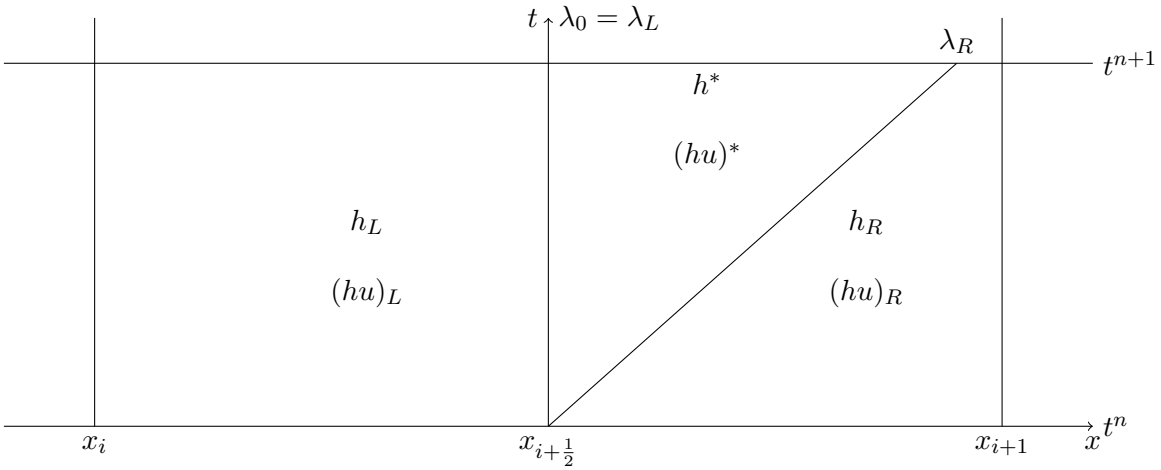


Fig. 3.3: Structure of the HLL type approximate Riemann solver $\mathcal{W}(t, x)$ for critical flows.

3.1.5 The Well-Balanced Source Average

As it turns out, the quadrature (3.12) of the source term will be crucial in order to ensure the well-balanced property given in definition 2.4.1.

In definition 2.4.1, the well-balanced property of a scheme is defined with respect to a discretization of an equilibrium. Accordingly, the data U_i are said to be in equilibrium, when they satisfy

$$\forall_{i=0}^{N_x} \begin{cases} (hu)_i = (hu)_{i+1}, \\ (\frac{u^2}{2} + g(h+B))_i = (\frac{u^2}{2} + g(h+B))_{i+1}, \end{cases} \tag{3.34}$$

and are said to satisfy the Lake at Rest, when there is

$$\forall_{i=0}^{N_x} \begin{cases} u_i = 0, \\ (h+B)_i = (h+B)_{i+1}. \end{cases} \tag{3.35}$$

Observe, that the second equation in (3.34) is a nonlinear function of the dependent variables h and hu. In the numerical experiments there will be initial conditions derived on the continuum. The projection onto the discrete data U_i will be taken as pointwise, i.e. $U_i = u(x_i)$ rather than the average over the volume. However, due to the midpoint rule, the given projection is second order accurate to the cell average value and since only up to second order accurate schemes are considered, no problems are expected from this slightly different discretization.

Following the arguments given in section 2.4, to achieve the well-balanced property it is sufficient to demand that in equilibrium there must hold

$$\mathcal{W}(t, 0^-) = w_L \qquad \text{and} \qquad \mathcal{W}(t, 0^+) = w_R. \tag{3.36}$$

The Lemma 3.1.1 concerns the well-balanced properties of the HLL scheme for subcritical

flows.

Lemma 3.1.1. *Given data that satisfy the relation (3.34), then, if the source quadrature is defined as*

$$\overline{ghB_x} = \frac{(hu)_L + (hu)_R}{2}(u_L - u_R) + \frac{(h_R + h_L)}{2}(\frac{u_R^2}{2} - \frac{u_L^2}{2}) + \frac{g}{2}(B_R - B_L)(h_R + h_L), \quad (3.37)$$

the HLL scheme from section 3.1.2 satisfies the relation (3.36) and is therefore well-balanced.

Given data that satisfy the relation (3.35), then, if the source quadrature is defined as

$$\overline{ghB_x} = \frac{g}{2}(B_R - B_L)(h_R + h_L), \quad (3.38)$$

or by (3.37), the HLL scheme from section 3.1.2 satisfies the relation (3.36) and is therefore well-balanced.

Proof. *The strategy is to first proof the well-balanced property for the discharge and then for the waterheights. Given data in equilibrium as in (3.34) and the source average determined by (3.37), there is*

$$\begin{aligned}(hu)^\star_{L,R} &= \frac{\lambda_R h_R u_R - \lambda_L h_L u_L - (h_R u_R^2 + g\frac{h_R^2}{2} - h_L u_L^2 - g\frac{h_L^2}{2}) - \overline{ghB_x}}{\lambda_R - \lambda_L} \\ &= hu + g\frac{(h_R + h_L)}{2}\frac{h_L - h_R - (\frac{u_R^2}{2g} - \frac{u_L^2}{2g}) - (B_R - B_L)}{\lambda_R - \lambda_L} = hu.\end{aligned}$$

Furthermore, for the waterheights it can be found that

$$D = \frac{(hu)^{*2}}{2g}(\frac{1}{h_L^2} - \frac{1}{h_R^2}) + (B_L - B_R) = \frac{1}{2g}(u_L^2 - u_R^2) + (B_L - B_R) = h_R - h_L,$$

and therefore

$$h_L^\star = \frac{\lambda_R h_R - \lambda_L h_L - \lambda_R D + q_L - q_R}{\lambda_R - \lambda_L} = \frac{\lambda_R h_R - \lambda_L h_L - \lambda_R(h_R - h_L)}{\lambda_R - \lambda_L} = h_L,$$

and

$$h_R^\star = \frac{\lambda_R h_R - \lambda_L h_L + \lambda_L D + q_L - q_R}{\lambda_R - \lambda_L} = \frac{\lambda_R h_R - \lambda_L h_L + \lambda_L(h_R - h_L)}{\lambda_R - \lambda_L} = h_R.$$

If, on the other hand, the data is in an equilibrium as (3.35) and the source average (3.38) is used, then there is

$$(hu)^\star_{L,R} = \frac{\frac{g}{2}(h_R^2 - h_L^2) - \overline{ghB_x}}{\lambda_R - \lambda_L} = \frac{g}{2}(h_R + h_L)\frac{(h_R - h_L) - (B_R - B_L)}{\lambda_R - \lambda_L} = 0.$$

For the waterheights there is

$$D = B_L - B_R,$$

and therefore

$$h_L^\star = \frac{\lambda_R h_R - \lambda_L h_L - \lambda_R(B_L - B_R)}{\lambda_R - \lambda_L} = \frac{(\lambda_R - \lambda_L)h_L}{\lambda_R - \lambda_L} = h_L,$$

and

$$h_R^\star = \frac{\lambda_R h_R - \lambda_L h_L + \lambda_L(B_L - B_R)}{\lambda_R - \lambda_L} = \frac{(\lambda_R - \lambda_L)h_R}{\lambda_R - \lambda_L} = h_R.$$

Finally, it holds that for $u = 0$ the definitions (3.38) and (3.37) are equivalent.

It should be remarked, that the quadrature (3.38) will not give a well-balanced scheme in the case of a general equilibrium. To see this, take data as in (3.34) and use the quadrature (3.38) to determine the discharge to get

$$(hu)^\star_{L,R} = hu + \frac{hu(u_R - u_L) + g(\frac{u_R^2}{2} - \frac{u_L^2}{2})\frac{h_L + h_R}{2}}{\lambda_R - \lambda_L}. \tag{3.39}$$

However, observe that if $\Delta_x \to 0$, then $(hu)^\star_{L,R} \to hu$ and therefore (3.38) is in this limit consistent with the general equilibrium. Therefore, it is expected that with smaller mesh size, the numerical errors will decrease and therefore the general equilibrium gets better resolved when using the quadrature (3.38).

On the other hand, the quadrature (3.37) is not consistent with the case of a flat bottom topography. In this case, the quadrature should go to 0 and the standard HLL scheme should be recovered. This problem is not unique to this approach and suggestions in the literature are that one artificially enforces this consistency by setting the quadrature to 0 depending on some thresholds on $(B_R - B_L)$. This is surely problem dependent and is not considered in this work.

Next the well-balanced properties of the schemes in the critical and supercritical case are discussed.

Lemma 3.1.2. *Given data that satisfy the relation (3.34), then, if the source quadrature is defined by (3.37), the schemes from section 3.1.4 and section 3.1.3 satisfy the relation (3.36) and are therefore well-balanced.*

Proof. *In both cases the limit $\mathcal{W}(t, 0^-) = w_L$ from (3.36) is trivial, since it is imposed in the model. For the model in the critical case (3.29) it remains to check that $h^* = h_R$ and $(hu)^* = (hu)_R$. It holds from (3.32) that*

$$h^\star = h_R - \frac{hu - hu}{\lambda_R} = h_R,$$

and following the proof of lemma 3.1.1 for equation (3.33) there is also

$$(hu)^\star = (hu)_R - \frac{(h_R u_R^2 + g\frac{h_R^2}{2} - h_L u_L^2 - g\frac{h_L^2}{2}) + \overline{ghB_x}}{\lambda_R} = (hu).$$

The model in the supercritical case satisfies condition (3.36) for the discharge (hu) by definition through equation (3.24). For the waterheight one has to realize that, in the case of equilibrium data (3.34), $h_L^* = h_R$ is a solution to (3.25). However, in general there is at most one, but up to three non-complex solutions to (3.25). In section 3.3, the structure of these solutions is analyzed and one can show that, if the data is physical relevant, there is

only one supercritical and only one subcritical solution to (3.25). Choosing in this case the supercritical root gives then the well-balanced result.

Even though the property (3.36) already holds, it should be remarked that, since in the supercritical case there is $h_L^* = h_R$, from (3.27) and (3.28) there is also $h_R^* = h_R$ and $(hu)_R^* = (hu)$.

3.1.6 On the Continuous Transition between the Models

This section concerns the transition of the numerical scheme when the flow changes type between sub- and supercritical. The models in the sections 3.1.2-3.1.4 are distinct and well behaved transition between them is needed to give reasonable approximations.

As has been pointed out in section 2.1, under the condition $CFL < \frac{1}{2}$, computing the update in the cell V_i as

$$U_i^{n+1} = U_i^n + \frac{\Delta_t}{\Delta_{x_i}}(F^+_{i-\frac{1}{2}} - F^-_{i+\frac{1}{2}}), \tag{3.40}$$

is equivalent to use the integral over the approximate Riemann solver at the time t_{n+1} as

$$\Delta_x U_i^{n+1} = \int_{x_{i-\frac{1}{2}}}^{x_i} W_{i-\frac{1}{2}}(t_{n+1}, x)dx + \int_{x_i}^{x_{i+\frac{1}{2}}} W_{i+\frac{1}{2}}(t_{n+1}, x)dx. \tag{3.41}$$

The models (3.7),(3.20) and (3.29) depend on parameterized wave speeds $\lambda_{L,R}$. Let this be denoted by $\mathcal{W}(t, x, \lambda_{L,R})$. What is meant exactly by a continuous transition is made clear in definition 3.1.1.

Definition 3.1.1. *The transition from model A to B is continuous at the interface $x_{i+\frac{1}{2}}$, if*

$$\begin{aligned} \int_{x_i}^{x_{i+\frac{1}{2}}} W_A(t_{n+1}, x, \lambda)dx &\xrightarrow[\lambda\to K]{} \int_{x_i}^{x_{i+\frac{1}{2}}} W_B(t_{n+1}, x, \lambda)dx, \\ \int_{x_{i+\frac{1}{2}}}^{x_{i+1}} W_A(t_{n+1}, x, \lambda)dx &\xrightarrow[\lambda\to K]{} \int_{x_{i+\frac{1}{2}}}^{x_{i+1}} W_B(t_{n+1}, x, \lambda)dx, \end{aligned} \tag{3.42}$$

and is denoted by $\mathcal{W}_A \xrightarrow[\lambda\to K]{} \mathcal{W}_B$.

Theorem 3.1.1. *If the intermediate state (3.26) is bounded in the limit $\lambda_L \to 0$, then there is*

$$W_{sub} \xrightarrow[\lambda_L\to 0^-]{} \mathcal{W}_{crit} \qquad \text{and} \qquad W_{sup} \xrightarrow[\lambda_L\to 0^+]{} \mathcal{W}_{crit}. \tag{3.43}$$

Proof. *First compute the integrals from definition 3.1.1 to get*

$$\begin{aligned} \int_{x_i}^{x_{i+\frac{1}{2}}} W_{sub}(t_{n+1}, x, \lambda)dx &= \frac{\Delta_x}{2}((1-\lambda_L)w_L + \lambda_L w^*_{sub,L}), \\ \int_{x_{i+\frac{1}{2}}}^{x_{i+1}} W_{sub}(t_{n+1}, x, \lambda)dx &= \frac{\Delta_x}{2}((1-\lambda_R)w_R + \lambda_R w^*_{sub,R}), \end{aligned}$$

and

$$\int_{x_i}^{x_{i+\frac{1}{2}}} W_{sup}(t_{n+1},x,\lambda)dx = \frac{\Delta_x}{2} w_L,$$
$$\int_{x_{i+\frac{1}{2}}}^{x_{i+1}} W_{sup}(t_{n+1},x,\lambda)dx = \frac{\Delta_x}{2}((1-\lambda_R)w_R + (\lambda_R-\lambda_L)w^*_{sup,R} + \lambda_L w^*_{sup,L}),$$

and

$$\int_{x_i}^{x_{i+\frac{1}{2}}} W_{crit}(t_{n+1},x,\lambda)dx = \frac{\Delta_x}{2} w_L,$$
$$\int_{x_{i+\frac{1}{2}}}^{x_{i+1}} W_{crit}(t_{n+1},x,\lambda)dx = \frac{\Delta_x}{2}((1-\lambda_R)w_R + \lambda_R w^*_{crit,R}).$$

In order to show $W_{sub} \xrightarrow[\lambda_L\to 0^-]{} \mathcal{W}_{crit}$*, it is sufficient to show that* $w^*_{sub,L}$ *is bounded and that* $w^*_{sub,R} \to w^*_{crit,R}$*. For the subcritical case this is straightforward by (3.18) and (3.16) if wet/dry areas are not considered.*

Moreover, it holds that there is (3.17) $\xrightarrow[\lambda_L\to 0^-]{}$ *(3.32) and (3.18)* $\xrightarrow[\lambda_L\to 0^-]{}$ *(3.33).*

In order to show $W_{sup} \xrightarrow[\lambda_L\to 0^+]{} \mathcal{W}_{crit}$*, it is sufficient to show that* $w^*_{sup,L}$ *is bounded and that* $w^*_{sup,R} \to w^*_{crit,R}$*. The bound on* $w^*_{sup,L}$ *is given by assumption and it also can easily be seen that there is (3.27)* $\xrightarrow[\lambda_L\to 0^+]{}$ *(3.32) and (3.28)* $\xrightarrow[\lambda_L\to 0^+]{}$ *(3.33).*

In section 3.3 it is made clear that the intermediate state (3.26) is in fact bounded by the critical state or the subcritical root. Therefore, theorem 3.1.1 gives confidence in computing transcritical flows.

3.2 Second order Extension

This section is devoted to find a higher order extension to the previous described finite volume scheme. This will work along the lines of section 2.3.1, where it is described to find a linear representation of the dependent variables to get a better estimate for the cell interface values. As also described in section 2.3.1, this can be done in any set of variables. This approach here follows the surface gradient method developed in [175] for the Lake at Rest solutions. The idea is to reconstruct deviations from an equilibrium state. To apply the surface gradient method, define the equilibrium variables $q = hu$ and $E = \frac{q^2}{2h^2} + g(h+B)$. These will be constant in equilibrium and define the surface to which the slopes are computed. The strategy is then to compute the slopes in the equilibrium variables q and E to compute the interface values. Then, the interface values in equilibrium variables have to be projected to the dependent variables h, hu.

The slopes in cell V_i in the equilibrium variables are computed as

$$\begin{aligned} \sigma E_i &= minmod\left(E_{i+1}-E_i, E_i - E_{i-1}\right), \\ \sigma q_i &= minmod\left(q_{i+1}-q_i, q_i - q_{i-1}\right). \end{aligned} \tag{3.44}$$

Therefore, the interface values can be computed as

$$E_{i+1/2^-} = E_i + \frac{\Delta_x}{2}\sigma E_{x,i}, \tag{3.45}$$

$$E_{i-1/2^+} = E_i - \frac{\Delta_x}{2}\sigma E_{x,i}, \tag{3.46}$$

$$q_{i+1/2^-} = q_i + \frac{\Delta_x}{2}\sigma q_{x,i}, \tag{3.47}$$

$$q_{i-1/2^+} = q_i + \frac{\Delta_x}{2}\sigma q_{x,i}. \tag{3.48}$$

The bottom topography is not reconstructed in this approach because

$$\frac{B_i - B_{i-1}}{\Delta_x} = B_x(x_{i-\frac{1}{2}}) + O(\Delta_x^2), \tag{3.49}$$

as long as B is sufficiently smooth. To get the respective values for the dependent variables h and hu one has to solve a third order polynomial in the waterheight as

$$\begin{aligned} P(h_{i+1/2^-}) &= h_{i+1/2^-}^3 + \frac{gB_i - E_{i+1/2^-}}{g} h_{i+1/2^-}^2 + \frac{q_{i+1/2^-}^2}{2g} \stackrel{!}{=} 0, \\ P(h_{i-1/2^+}) &= h_{i-1/2^+}^3 + \frac{gB_i - E_{i-1/2^+}}{g} h_{i-1/2^+}^2 + \frac{q_{i-1/2^+}^2}{2g} \stackrel{!}{=} 0. \end{aligned} \tag{3.50}$$

How to solve for the roots in (3.50) is described in section 3.3. As it turns out, if there are positive roots to (3.50), then there is a sub- and a supercritical one. The strategy also applied in [172] is not to change the type of the flow by the reconstruction, i.e. if the flow is subcritical at the cell center, then also the subcritical roots are taken at the interface and vice versa.

The reconstruction procedure described here does not guarantee that the interface values in the equilibrium variables give real and positive roots. A more detailed analysis may be needed on this subject, but is omitted in this work. Instead, if no positive real roots can be found, it is decided to go back to first order and set $h_{i+1/2^-} = h_{i-1/2^+} = h_i$.

If the interface values for the dependent variables are known, the fluxes can be computed by the models (3.7),(3.20) and (3.29). However, the source averages (3.37) and (3.38) are computed with respect to the cell center values.

Finally, it is straightforward to see that this second order extension gives a well-balanced scheme. In equilibrium the slopes computed in (3.44) are zero and therefore the interface values for the equilibrium variables coincide with the equilibrium variables at the cell center. When computing the dependent variabels again, not changing the type of the flow guarantees that the cell centered value for the waterheight is recovered. This follows again from the fact that there is only one root for for each polynomial in (3.50) in the sub- or the supercritical regime. Having computed the source term quadrature with respect to the cell centered values immediately leads back to lemma 3.1.1.

3.3 Finding the Roots

In the second order extension (3.2) and in the model for supersonic flow (3.20), the roots of a third order polynomial have to be found to compute the waterheight. Defining $q = hu$ and the energy $E = \frac{q^2}{2h^2} + g\,(h + B)$, the polynomial in h reads

$$\begin{aligned} P(h) &= h^3 + \frac{gB - E}{g} h^2 + \frac{q^2}{2g} \\ &= h^3 + a_0 h^2 + a_2. \end{aligned} \tag{3.51}$$

From the fundamental theorem of algebra it is known that the polynomial in (3.51) has either one or three real roots. The purpose of this section is to analyze, which values a_0 and a_2 admit one or three real roots and, if these roots are positive, to which flow regimes these roots belong, i.e. sub- or supercritical, and compute them explicitly.

3.3.1 Structure of P in the Physical relevant Case

From the definition of the energy E, one immediately has two properties of the coefficients:

$$a_0 < 0, \quad a_2 \geq 0. \tag{3.52}$$

The polynomial $P(h)$ is now analyzed for parameters that satisfy (3.52). Since the leading coefficient of $P(h)$ is positive, one has that

$$\lim_{h \to -\infty} P(h) = -\infty \qquad \text{and} \qquad \lim_{h \to +\infty} P(h) = +\infty. \tag{3.53}$$

Now compute the extrema of of $P(h)$. Setting

$$\frac{\partial P(h)}{\partial h} = 3h^2 + 2a_0 h \stackrel{!}{=} 0, \tag{3.54}$$

admits the two solutions

$$\tilde{h}_1 = 0 \qquad \text{and} \qquad \tilde{h}_2 = -\frac{2a_0}{3}, \tag{3.55}$$

and evaluating the second derivative at $\tilde{h}_{1,2}$ gives

$$\frac{\partial^2 P(h)}{\partial h^2}(\tilde{h}_1) = 2a_0 \qquad \text{and} \qquad \frac{\partial^2 P(h)}{\partial h^2}(\tilde{h}_2) = -2a_0. \tag{3.56}$$

Therefore, in the case (3.52), $\tilde{h}_1$ is a local maximum and $\tilde{h}_2$ is a local minimum of $P(h)$.

Now, first analyze the cases where the extrema $\tilde{h}_{1,2}$ are roots of $P(h)$ by computing $P(\tilde{h}_{1,2}) = 0$.

- Case I: $P(\tilde{h}_1) = 0$ gives $a_2 = 0$ and therefore $q = 0$, i.e. a flow at rest. Here the roots are computed directly to be

$$h_{1,2} = 0, \qquad \text{and} \qquad h_3 = -a_0. \tag{3.57}$$

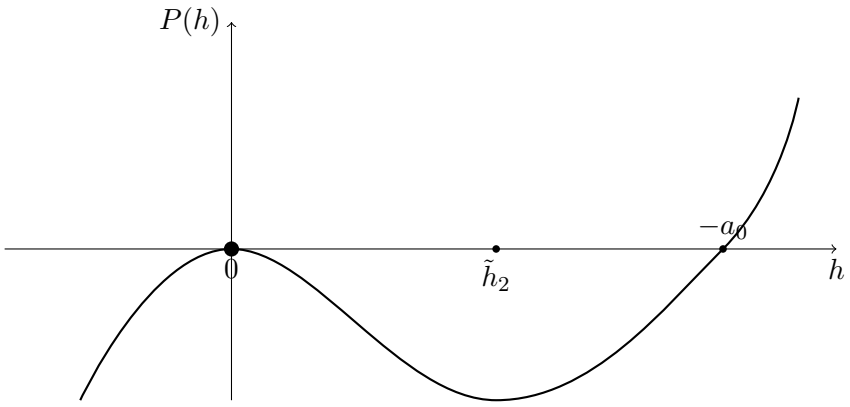


Fig. 3.4: Shape of the polynomial $P(h)$ when there is a double root at zero.

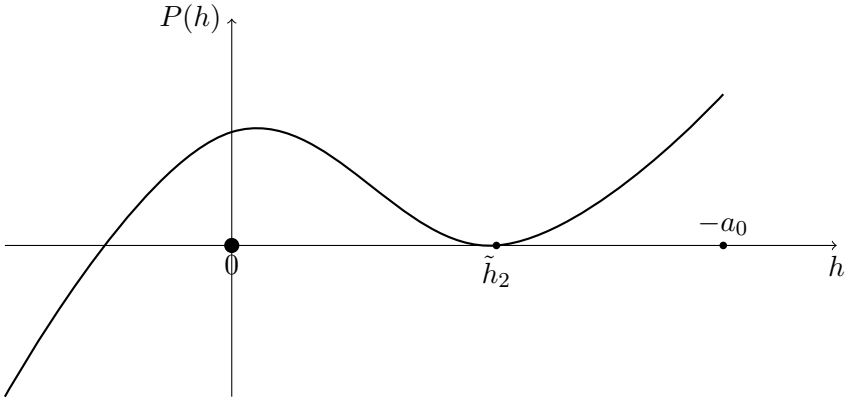


Fig. 3.5: Shape of the polynomial $P(h)$ when there is a double root at $\tilde{h}_2$.

Therefore there is a double root at zero which is also a maximum. From (3.52), the third root is on the positive half axis and the polynomial takes the shape as depicted in figure 3.4. Since the case of wet/dry zones is omitted in this work, the physical relevant root is h_3.

- Case II: $P(\tilde{h}_2) = 0$ gives $a_2 = -\frac{4}{27}a_0^3$. Here there is a double root at the minimum $\tilde{h}_2$ and a simple polynomial division gives that, in the case of (3.52), there is one root on the negative half axis as shown in figure (3.5).

$$h_1 = \frac{1}{3}a_0, \qquad \text{and} \qquad h_{2,3} = -\frac{2a_0}{3}. \tag{3.58}$$

The physical relevant roots are therefore $h_{2,3}$. In fact, these roots correspond to the critical regime since, by the definition of a_0, it holds in this case that $u^2 = gh$.

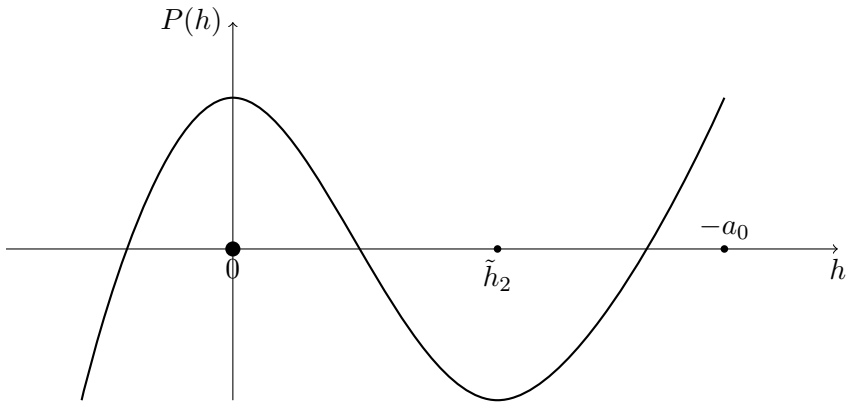


Fig. 3.6: Shape of the polynomial $P(h)$, when there are two physical relevant roots.

Since in the physical relevant case (3.52) $P(0) \geq 0$ and there is no extremum on the negative half axis, apart from case I, there is always an unphysical negative real root. In order to have physical relevant roots, a necessary and sufficient condition is to demand $P(\tilde{h}_2) \leq 0$, see figure 3.6. In turn, this gives an additional restriction to the parameters a_0 and a_2 as

$$a_2 \in [0, -\frac{4}{27}a_0^3]. \tag{3.59}$$

In this case, there are two positive real roots $h_{1,2}$, for which there is

$$h_1 \in (0, \tilde{h}_2) \qquad \text{and} \qquad h_2 > \tilde{h}_2. \tag{3.60}$$

Using again (3.55) it can be computed that for these two roots there is

$$gh_1 < u^2 \qquad \text{and} \qquad gh_2 > u^2. \tag{3.61}$$

Therefore, h_1 corresponds to the supercritical regime and h_2 to the subcritical regime. Moreover, see that, if there is a root in the respective flow regime, it is unique. This is critical for the development of the model (3.20) and the projection to the dependent variables in (3.2).

If $a_2 > -\frac{4}{27}a_0^3$, there is $P(\tilde{h}_2) > 0$ and therefore there are no positive real roots, see figure (3.7). Therefore, in order for the parameters a_0 and a_2 to give physical relevant solutions, the following relations have to be satisfied

$$a_0 < 0 \qquad \text{and} \qquad a_2 \in [0, -\frac{4}{27}a_0^3] \tag{3.62}$$

3.3.2 Computation of the roots

Now the computation of the roots of the polynomial $P(h)$ in the physical relevant scenario (3.62) is discussed. By the substitution $h = t - \frac{a_0}{3}$ one has the following depressed form:

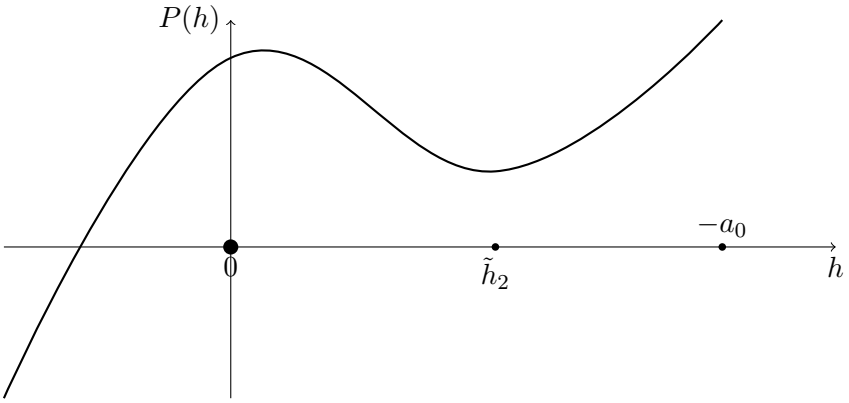


Fig. 3.7: Shape of the polynomial $P(h)$, when there are no physical relevant roots.

$$Q(t) = t^3 - \frac{a_0^2}{3}t + \frac{27a_2 + 2a_0^3}{27}. \tag{3.63}$$

Due to the substitution it is obvious that, if the roots of $Q(t)$ are real, so are the roots of $P(h)$ and vice versa. If the roots of $Q(t)$ are all real, then these roots can be computed as

$$t_1 = -\frac{2}{3}a_0\cos(\frac{\phi}{3}) \quad t_2 = -\frac{2}{3}a_0\cos(\frac{\phi+2\pi}{3}) \quad t_3 = -\frac{2}{3}a_0\cos(\frac{\phi+4\pi}{3}), \tag{3.64}$$

where there is

$$\phi = \arctan(-\frac{3\sqrt{3}\sqrt{-a_2\left(4a_0^3+27a_2\right)}}{2a_0^3+27a_2}), \tag{3.65}$$

see also [83]. This can now be rewritten as roots of $P(h)$ as

$$\begin{gathered} h_1 = -\frac{1}{3}a_0\left(2\cos(\frac{\phi}{3})+1\right), \quad h_2 = -\frac{1}{3}a_0\left(2\cos(\frac{\phi+2\pi}{3})+1\right), \\ h_3 = -\frac{1}{3}a_0\left(2\cos(\frac{\phi+4\pi}{3})+1\right). \end{gathered} \tag{3.66}$$

Since in the physical relevant case all roots are real, these formulas can be applied to compute the roots of $P(h)$. It remains to check, which of the roots in (3.66) correspond to the unphysical, sub- and supercritical root as discussed in section 3.3.1.

Consider the following cases:

- Case I: $a_2 = 0$. This gives $\phi = 0$ and the roots to $P(h)$ from (3.66) read

$$h_1 = -a_0 \qquad h_2 = 0 \quad h_3 = 0 \tag{3.67}$$

, see also figure 3.8.

- Case II: $a_2 = -\frac{4}{27}a_0^3$. This gives $\phi = \pi$ and the roots to $P(h)$ from (3.66) read

$$h_1 = -\frac{2}{3}a_0 \qquad h_2 = \frac{1}{3}a_0 \quad h_3 = -\frac{2}{3}a_0, \tag{3.68}$$

  see also figure 3.9.

Furthermore, compute the derivative of the roots with respect to a_2 as

$$\begin{aligned} \frac{\partial h_i}{\partial a_2} &= \frac{\partial}{\partial a_2}\left(-\frac{1}{3}a_0\left(2\cos(\frac{\phi(a_2)+2(i-1)\pi}{3})+1\right)\right) \\ &= \frac{4}{9}a_0\left(\sin(\frac{\phi(a_2)+2(i-1)\pi}{3})\frac{\partial\phi(a_2)}{\partial a_2}\right). \end{aligned} \tag{3.69}$$

In the physical relevant case (3.62) it holds that

$$\begin{aligned} \frac{\partial\phi(a_2)}{\partial a_2} &= \frac{\partial}{\partial a_2}\arctan(-\frac{3\sqrt{3}\sqrt{-a_2\left(4a_0^3+27a_2\right)}}{2a_0^3+27a_2}) \\ &= \frac{1}{1+x^2}\frac{12\sqrt{3}a_0^6}{(2a_0^3+27a_2)^2\sqrt{-a_2(4a_0^3+27a_2)}} > 0, \end{aligned} \tag{3.70}$$

where

$$x = -\frac{3\sqrt{3}\sqrt{-a_2\left(4a_0^3+27a_2\right)}}{2a_0^3+27a_2}$$

Since by Case I and Case II, $\phi(0) = 0$ and $\phi(-\frac{2}{27}a_0^3) = \pi$ and with (3.70), there is in the physical relevant case $\phi(a_2) \mapsto [0,\pi]$ monotonically. It remains to check the sign of $\sin(\frac{\phi+2(i-1)\pi}{3})$ in (3.69) for $\phi \in (0,\pi)$. There is

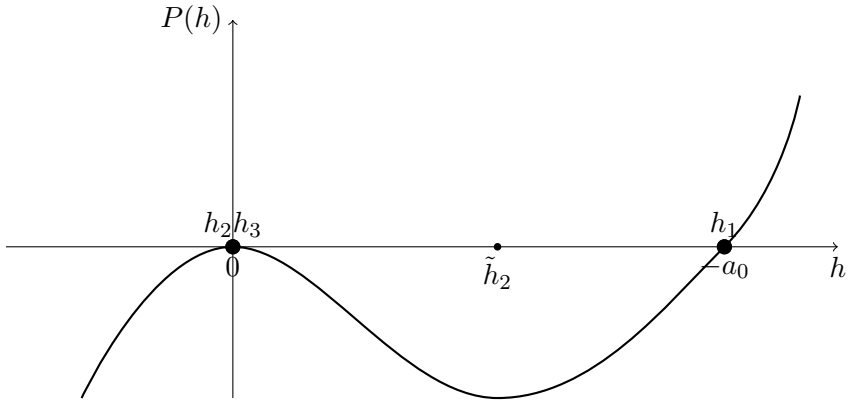


Fig. 3.8: Location of the roots (3.66) in the case $a_2 = 0$

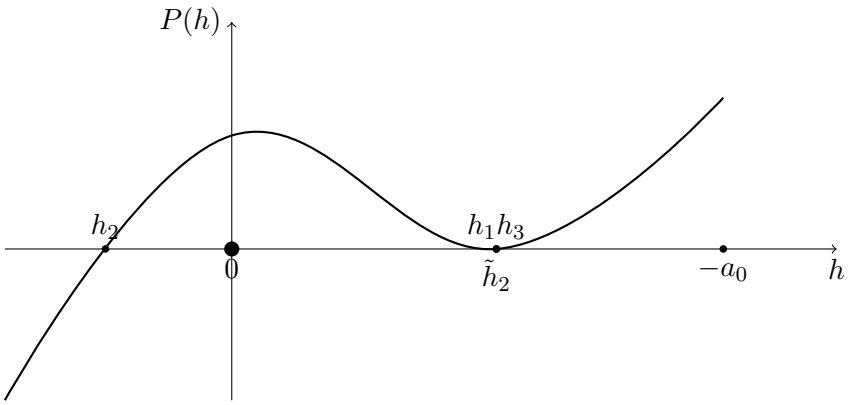


Fig. 3.9: Location of the roots (3.66) in the case $a_2 = -\frac{4}{27}a_0^3$

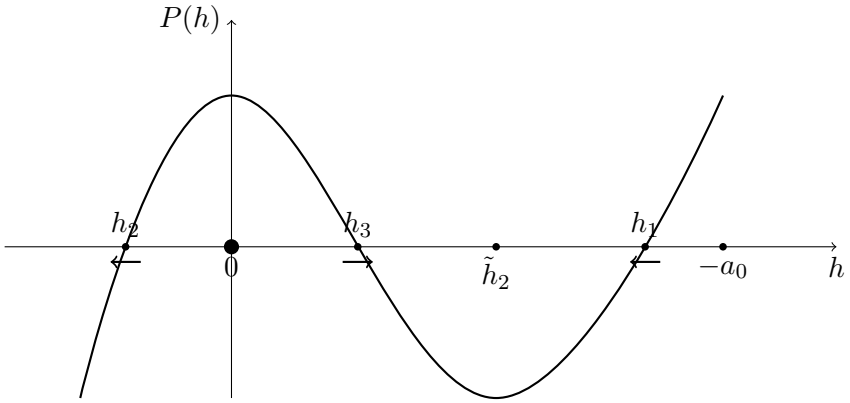


Fig. 3.10: Derivatives of the roots with respect to a_2 in the case $a_2 \in (0, -\frac{4}{27}a_0^3)$

- $i = 1$: $\sin(\frac{\phi}{3}) > 0$ for $\phi \in (0, \pi)$,
- $i = 2$: $\sin(\frac{\phi+2\pi}{3}) > 0$ for $\phi \in (0, \pi)$,
- $i = 3$: $\sin(\frac{\phi+4\pi}{3}) < 0$ for $\phi \in (0, \pi)$,

which gives for the derivatives from (3.69)

$$\frac{\partial h_1}{\partial a_2} < 0 \qquad \frac{\partial h_2}{\partial a_2} < 0 \qquad \frac{\partial h_3}{\partial a_2} > 0, \tag{3.71}$$

see also figure (3.10). Therefore, from (3.61),(3.67),(3.68) and (3.71) it follows that the roots from 3.66 can be associated to the respective flow regime as follows

$$h_1 \text{ :subcritical} \quad , h_2 \text{ :unphysical} \quad , h_3 \text{ :supercritical.}$$

3.3.3 Synopsis

The results from the sections 3.3.1 and 3.3.2 are now repeated in a compact form. Define the equilibrium variables $q = hu$ and $E = \frac{q^2}{2h^2} + g\,(h + B)$. To solve for the waterheight one has to solve $P(h) = 0$, where

$$\begin{aligned} P(h) &= h^3 + \frac{gB - E}{g}h^2 + \frac{q^2}{2g} \\ &= h^3 + a_0 h^2 + a_2. \end{aligned}$$

$P(h)$ only admits physical relevant roots, if

$$a_0 < 0, \quad a_2 \in [0, -\frac{4}{27}a_0^3]. \tag{3.72}$$

They can be computed explicitly to be

$$h_{sub} = -\frac{1}{3}a_0 \left(2\cos(\frac{\phi}{3}) + 1\right) \quad h_{super} = -\frac{1}{3}a_0 \left(2\cos(\frac{\phi + 4\pi}{3}) + 1\right), \tag{3.73}$$

where there is

$$\phi = \arctan(-\frac{3\sqrt{3}\sqrt{-a_2\left(4a_0^3 + 27a_2\right)}}{2a_0^3 + 27a_2}). \tag{3.74}$$

3.4 Numerical Tests

This section is concerned with the practical application of the derived scheme. To this end, first three test cases are considered with respect to the three different flow regimes, i.e. a sub-, trans-, and supercritical equilibrium. These tests have been developed by the author and are not motivated by practical applications, but by the simplicity of the tests with respect to implementation, but also with respect to the scalability to the different flow regimes. After that, some testcases proposed in the literature are considered in order to investigate the practical applicability of the scheme.

In all tests, an equidistant grid is concerned. Denote by D the length of the domain and by N_x the number of cells, then there is $\Delta_x = \frac{D}{N_x}$. Furthermore, the parameters $\delta_{1,2}$ from (3.6) are set to 1.5, and Neumann boundary conditions are imposed.

Besides investigating the performance of the derived scheme, the purpose of this section is also to show the benefits of a well-balanced scheme if near equilibrium solutions are computed. In section 3.1.5, two different quadrature rules for the source term are derived, i.e. the quadrature (3.37) gives a scheme that is consistent with the general equilibria (3.2) and the quadrature (3.38) gives a scheme that is only consistent with the Lake at Rest solutions (3.3). Both quadratures are applied in order to show the superiority of the general quadrature (3.37) over the Lake at Rest quadrature (3.38). The schemes employed with the respective quadratures are noted as HLL_{ME} and HLL_{LR}. Moreover, it shall be shown, that the second order in space extension discussed in section 3.2 for one gives a well-balanced scheme and also gives better approximations as compared to the first order scheme at the same resolution. Whenever the second order in space approach is used, the second order

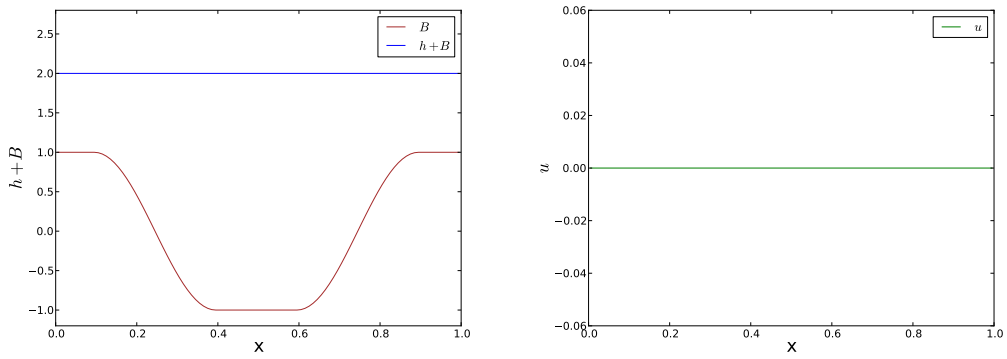


Fig. 3.11: Lake at Rest initial condition. Left: The bottom topography B and the total waterheight $h + B$. Right: Velocity u

time-discretization from [14] discussed in section 2.3.2 is applied. The first and second order schemes are denoted as HLL^{FO} and HLL^{SO} respectively.

3.4.1 Lake at Rest

In this section a Lake at Rest equilibrium is concerned. The domain size D is set to 1 as well as the gravitational constant g is set to 1 and the bottom topography takes the following shape

$$B(x) = \begin{cases} 1 & \text{if} \quad x < x_0, \\ \cos(\frac{x-x_0}{x_1-x_0}\pi) & \text{if} \quad x_0 < x < x_1, \\ -1 & \text{if} \quad x_1 < x < x_2, \\ -\cos(\frac{x-x_2}{x_3-x_2}\pi) & \text{if} \quad x_2 < x < x_3, \\ 1 & \text{if} \quad x > x_3, \end{cases} \tag{3.75}$$

where $x_0 = 0.1$, $x_1 = 0.4$, $x_2 = 0.6$ and $x_3 = 0.9$. The Lake at Rest equilibrium is set up by defining the equilibrium variables as

$$\begin{cases} h(x)_{eq} + B(x) = 2, \\ u(x)_{eq} = 0, \end{cases} \tag{3.76}$$

see also figure 3.11.

At first the equilibrium defined by (3.75) and (3.76) is used as an initial condition and the respective schemes are used to compute the evolution. The expected result is, that the discrete time derivative vanishes for all the schemes and therefore the Lake at Rest equilibrium is preserved up to machine precision. The L^1 error is shown in table 3.1 which shows, that all the schemes show the expected performance.

Next a disturbance on the waterheight is placed on the equilibrium such that

$$h(0,x) - h(x)_{eq} = \begin{cases} \sin(\frac{x-x_1}{x_2-x_1}\pi) & \text{if} \quad x_1 < x < x_2, \\ 0 & \text{else}, \end{cases} \tag{3.77}$$

	HLL^{FO}_{ME}		HLL^{SO}_{ME}		HLL^{FO}_{LR}	
N	h	hu	h	hu	h	hu
100	0.01E-16	4.01E-16	0.01E-16	3.17E-16	0.01E-16	4.01E-16
200	0.01E-16	5.40E-16	0.01E-16	6.06E-16	0.01E-16	5.40E-16
400	0.01E-16	9.00E-16	0.11E-16	9.40E-16	0.01E-16	9.00E-16
800	0.01E-16	7.20E-16	0.24E-16	6.77E-16	0.01E-16	7.20E-16
1600	0.01E-16	8.95E-16	0.24E-16	7.56E-16	0.04E-16	8.95E-16
3200	0.11E-16	9.68E-16	0.47E-16	9.27E-16	0.11E-16	9.68E-16

Table 3.1: L^1 error for the undisturbed Lake at Rest solution given by (3.75) and (3.76).

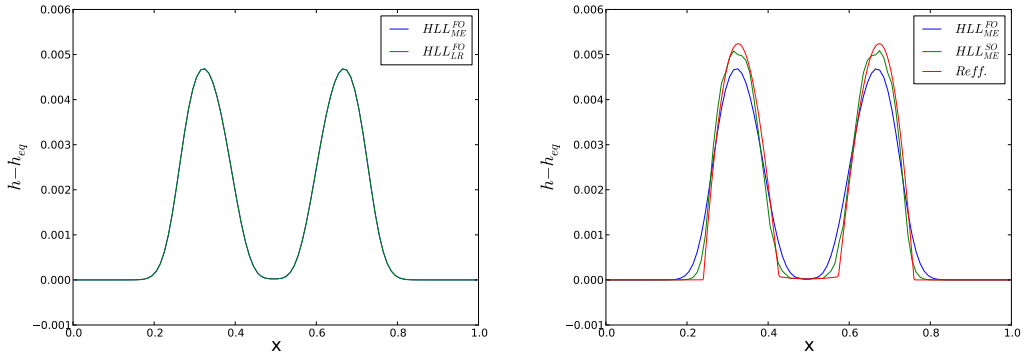


Fig. 3.12: Solutions to the disturbed Lake at Rest after time 0.1. Left: Comparison of the first order schemes with the different quadratures. Right: Comparison of the higher order extension with respect to the first order approximation and a reference solution.

and the evolution of the disturbance is computed on a mesh with size $N_x = 100$. The results are depicted in figure 3.12. The first order schemes with the different quadrature rules give almost identical results, since the quadrature for the general equilibria (3.2) is consistent with the quadrature for the Lake at Rest solutions (3.38) when Lake at Rest solutions are concerned. Moreover a comparison with the first order and second order version are shown with respect to a reference solution computed on a mesh with $N_x = 5400$. The second order extension gives a better resolution of the resulting waves even on the coarser mesh.

3.4.2 Scalable Moving Equilibrium

Now, a general equilibrium is concerned. Again the domain size D is set to 1 as well as the gravitational constant g is set to 1. The equilibrium conditions are given by

$$\begin{cases} h(x)_{eq}u(x)_{eq} & = Cq, \\ \frac{u(x)^2_{eq}}{2} + g(h(x)_{eq} + B(x)) & = Ce, \end{cases} \tag{3.78}$$

where Cq and Ce are constants determining the equilibrium structure. First, rearrange (3.78) in the following way

$$\begin{cases} u(x)_{eq} &= \frac{Cq}{h(x)_{eq}}, \\ B(x) &= \frac{Ce}{g} - (\frac{Cq^2}{2gh(x)_{eq}^2} + h(x)_{eq}). \end{cases} \tag{3.79}$$

Therefore the strategy for this section is to compute the equilibrium solutions first by stating $h(x)_{eq}$, Cq and Ce and then compute $u(x)_{eq}$ and $B(x)$ accordingly. In comparison to [140], where the bottom topography and the constants Cq and Ce are given, this approach has the advantage that the equilibrium solutions are known explicitly and therefore it is easy to check if the equilibrium is physical relevant and if it is sub-, trans- or supercritical.

Here it is decided to parameterize the moving equilibrium as follows

$$\begin{aligned} h(x)_{eq} &= \begin{cases} 3 & \text{if } x < 0.25, \\ 2 + 0.5\cos((x - 0.25)2\pi) & \text{if } 0.25 < x < 0.75, \\ 2 & \text{if } x > 0.75, \end{cases} \\ Ce &= 3. \end{aligned} \tag{3.80}$$

The parameter Cq is used to scale the equilibrium to take all the three flow types. In fact, with the first equation of (3.79), the condition on the left eigenvalue can be rewritten

$$\begin{aligned} u - \sqrt{gh} \leq 0 \Leftrightarrow \left(\frac{Cq^2}{g}\right)^3 \leq h(x)_{eq}, \\ u - \sqrt{gh} \geq 0 \Leftrightarrow \left(\frac{Cq^2}{g}\right)^3 \geq h(x)_{eq}. \end{aligned} \tag{3.81}$$

As the distribution for the waterheight is given by (3.80), the following values for Cq correspond to the respective flow regime

$$Cq = \begin{cases} 1 & \text{for subcritical flow}, \\ 3 & \text{for transcritical flow}, \\ 5 & \text{for supercritical flow}. \end{cases} \tag{3.82}$$

The resulting equilibria are shown in figure 3.13.

In the following, the schemes are tested for the equilibria in all the flow regimes with respect to their well-balanced property as well as the ability to accurately capture small deviations on the respective equilibria, where the deviations are chosen to take the following form

$$h(0,x) - h(x)_{eq} = \begin{cases} 0.1 \times 10^{-7} \sin(\frac{x - x_0}{x_1 - x_0}\pi) & \text{if} \quad x_0 < x < x_1, \\ 0 & \text{else}, \end{cases} \tag{3.83}$$

with $x_0 = 0.45$ and $x_1 = 0.55$.

First, the subcritical equilibrium is concerned. The equilibrium is set as initial condition for the respective schemes and then is numerical integrated. The L^1 errors are shown in table 3.2, where EOC is short for estimated rate of convergence. Both the first order and the second order scheme using the quadrature rule (3.2) are consistent with the equilibrium and

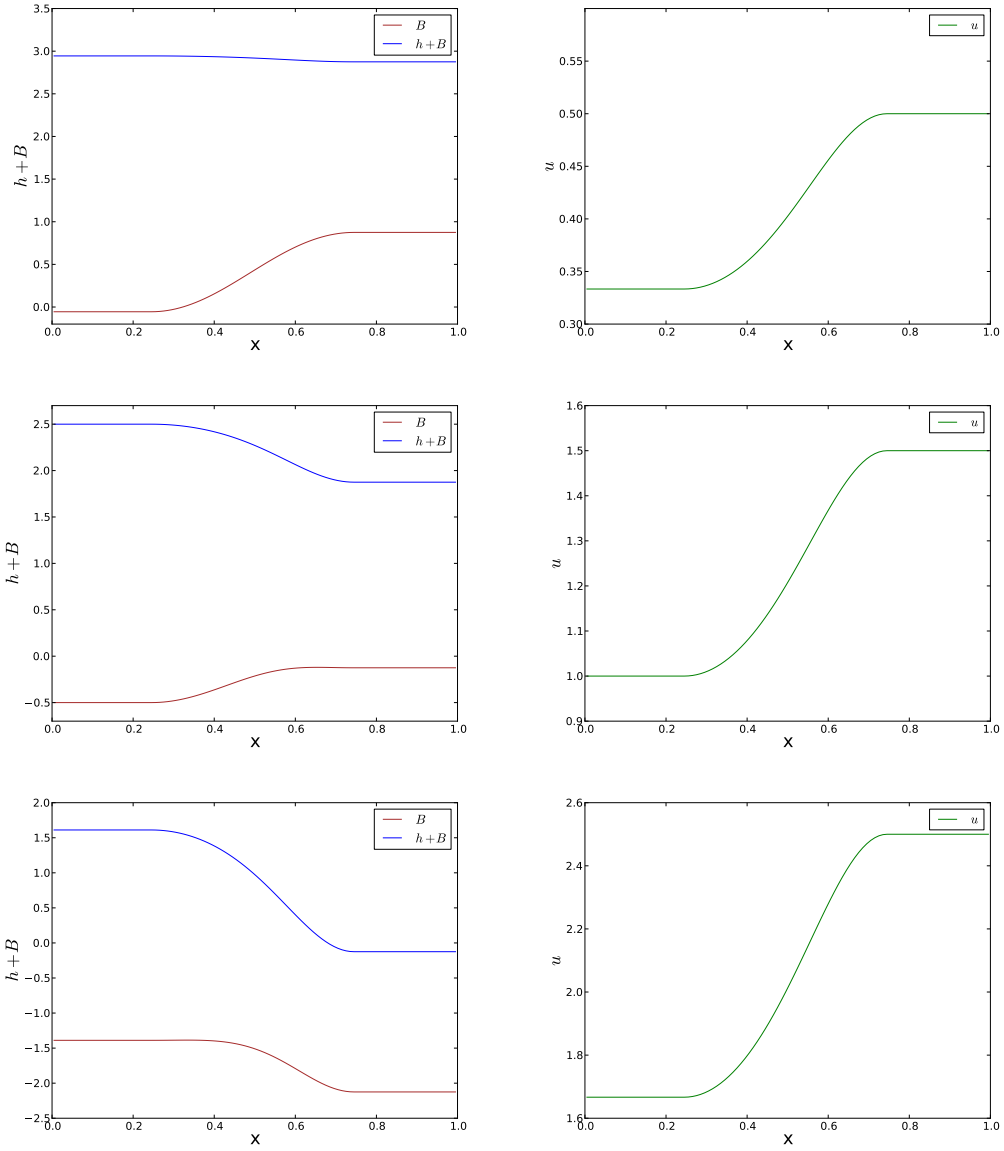


Fig. 3.13: Moving equilibria. From top to bottom are the sub- trans- and supercritical equilibria. Left: The bottom topography B and the total waterheight $h+B$. Right: Velocity u

	HLL^{FO}_{ME}		HLL^{SO}_{ME}		HLL^{FO}_{LR}			
N	h	hu	h	hu	h	EOC	hu	EOC
100	0.01E-16	1.23E-16	0.01E-16	2.79E-16	6.69E-09	-	1.64E-08	-
200	0.01E-16	1.33E-16	0.01E-16	2.69E-16	8.28E-10	3.01	2.08E-09	2.98
400	0.01E-16	1.50E-16	0.78E-16	2.75E-16	1.03E-10	3.01	2.63E-10	2.98
800	0.01E-16	1.50E-16	0.01E-16	0.97E-16	1.29E-11	3.00	3.30E-11	3.00
1600	0.01E-16	1.47E-16	0.01E-16	2.37E-16	1.67E-12	2.95	4.06E-12	3.00
3200	0.01E-16	1.59E-16	0.01E-16	2.36E-16	2.70E-14	5.96	2.99E-13	3.76

Table 3.2: L^1 error for the undisturbed subcritical moving equilibrium at time 0.18

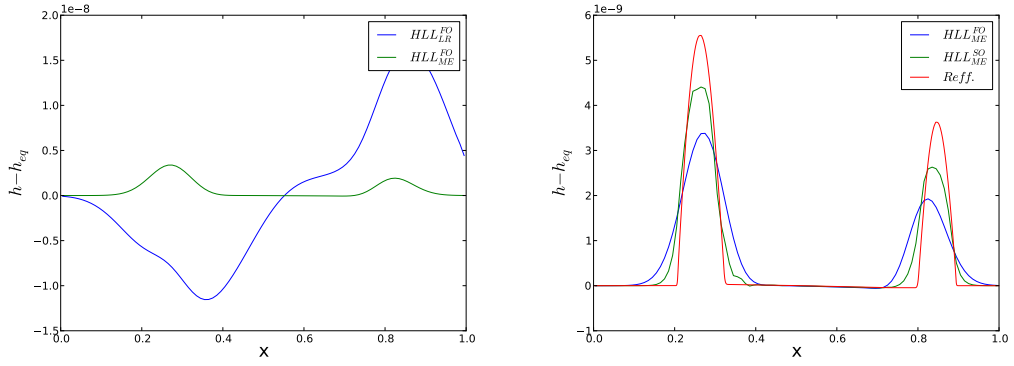


Fig. 3.14: Solutions to the disturbed subcritical moving equilibrium at time 0.18. Left: Comparison of the first order schemes computed with 100 cells. Right: Comparison of the first and second order scheme with respect to a reference solution computed with 5400 cells.

preserve it up to machine precision. The scheme equipped with the quadrature (3.38) is not consistent with that type of equilibrium and therefore introduces numerical errors. However, increasing the resolution decreases the error both in the waterheight and the discharge with third order.

Next, the evolution of a disturbance on the subcritical equilibrium is computed. The results are depicted in figure 3.14. The scheme with the quadrature (3.38) is not able to accurately capture the resulting waves. The numerical solution is dominated by the error coming from the inconsistency with the moving equilibrium. However, it is expected from table 3.2 that, if a higher resolution is chosen, the numerical error will decrease and the resolution of the waves will be better, see also section 3.4.3. Additionally, it is shown that the second order extension decreases the numerical viscosity and gives a better resolution of the dynamics compared to the first order scheme.

The second numerical experiments are devoted to the transcritical equilibrium. As for the subcritical equilibrium, first the undisturbed equilibrium is used as an initial condition and is integrated numerically. The L^1 errors are given in table 3.3. Again, the quadrature consistent only with the Lake at Rest solution is not able to capture the equilibrium exactly and therefore numerical errors are introduced, which again decrease with third order. Also as expected the first and second order scheme using the quadrature for the general equilibria

	HLL^{FO}_{ME}		HLL^{SO}_{ME}		HLL^{FO}_{LR}			
N	h	hu	h	hu	h	EOC	hu	EOC
100	0.01E-16	0.01E-16	0.01E-16	0.01E-16	4.71E-08	-	1.15E-07	-
200	0.01E-16	0.01E-16	0.01E-16	0.01E-16	5.74E-09	3.03	1.40E-08	3.04
400	0.01E-16	0.01E-16	0.01E-16	0.01E-16	7.08E-10	3.02	1.72E-09	3.02
800	0.01E-16	0.01E-16	0.01E-16	0.01E-16	8.79E-11	3.01	2.13E-10	3.01
1600	0.01E-16	0.01E-16	0.01E-16	0.01E-16	1.09E-11	3.01	4.06E-11	2.40
3200	0.01E-16	0.01E-16	0.01E-16	0.01E-16	1.20E-12	3.18	3.43E-12	3.57

Table 3.3: L^1 error for the undisturbed transcritical moving equilibrium at time 0.12.

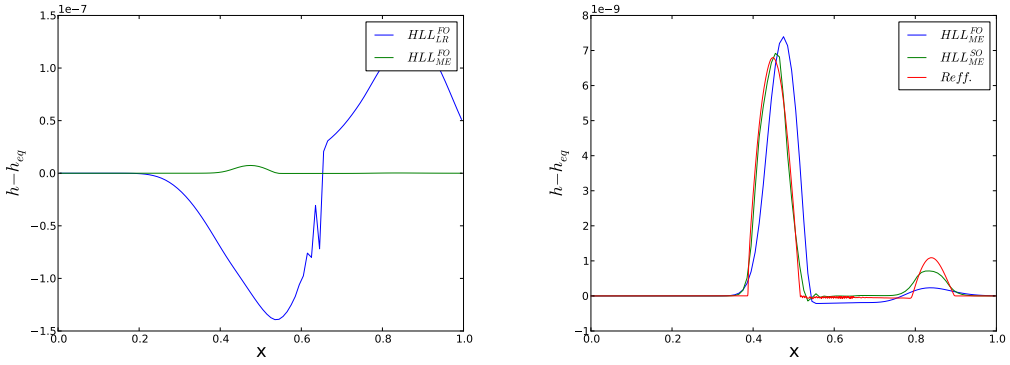


Fig. 3.15: Solutions to the disturbed transcritical moving equilibrium at time 0.12. Left: Comparison of the first order schemes computed with 100 cells. Right: Comparison of the first and second order scheme with respect to a reference solution computed with 5400 cells.

show the desired well-balanced property.

Again, in order to show the applicability of the well-balanced scheme, a small deviation is placed now on the transcritical equilibrium and then integrated numerically, see figure 3.15. Also in this case the general quadrature (3.2) shows a far better behavior to capture the resulting dynamics. Observe now, since the flow changes type from sub- to supercritical as going from left to right where the transcritical point is just where the initial disturbance is placed, one wave moves downstream as the other wave does not move due to the critical flow velocity. Also in this case the second order extension shows the desired properties in giving a sharper resolution of the waves. The oscillation in the first order scheme can be explained by the unphysical diffusion from the inconsistency of the scheme with the underlying PDE.

Finally, the case of a supercritical equilibrium is concerned. In this case it is not necessary to distinguish between the different quadrature rules, since the model derived in section 3.1.3 does not depend on a quadrature to define the numerical fluxes. In table 3.4 are again given the L^1 errors with respect to the equilibrium.

In this regime the HLL approximate Riemann solver is considered to give the most accurate solutions, since the equilibrium conditions are exactly solved across the interface. However, due to the supercritical structure, all the eigenvalues are positive and the finite volume approach itself introduces diffusion. The evolution of a small disturbance is shown in figure

	HLL^{FO}_{ME}		HLL^{SO}_{ME}	
N	h	hu	h	hu
100	0.01E-16	0.01E-16	0.01E-16	0.01E-16
200	0.01E-16	0.01E-16	0.01E-16	0.01E-16
400	0.01E-16	0.01E-16	0.01E-16	0.01E-16
800	0.01E-16	0.01E-16	0.01E-16	0.22E-16
1600	0.01E-16	0.01E-16	0.01E-16	0.14E-16
3200	0.01E-16	0.01E-16	0.01E-16	0.07E-16

Table 3.4: L^1 error for the undisturbed supercritical moving equilibrium at time 0.08

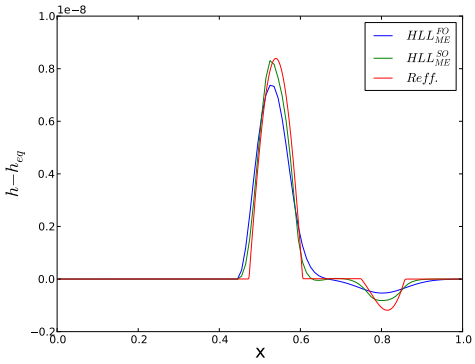


Fig. 3.16: Solutions to the disturbed supercritical moving equilibrium at time 0.08. Comparison of the first and second order scheme computed with 100 cells with respect to a reference solution computed with 5400 cells.

3.16. Due to the supercritical regime, now both waves move to the right, while the left waves only moves very slowly. The second order scheme also gives in this case a better resolution, especially on the left wave. On the other hand, the computation of the right wave seems to suffer strongly from numerical diffusion. Even though the second order scheme gives a better resolution, it is still quite far away from the reference solution.

3.4.3 The Noelle-Shu-Xing Testcases

The next testcase are suggested by [140] to show the superiority of general well-balanced schemes with respect to schemes that are only well-balanced with respect to the Lake at Rest. For all the testcases the size of the domain is set to $D = 25$ and the gravitational constant is $g = 9.812$. In specific two testcases are considered, namely a subcritical and a transcritical one. The subcritical equilibrium is determined by

$$\begin{aligned} Cq &= 4.42, \\ Ce &= 22.06605, \\ B(x) &= \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 < x < 12, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{3.84}$$

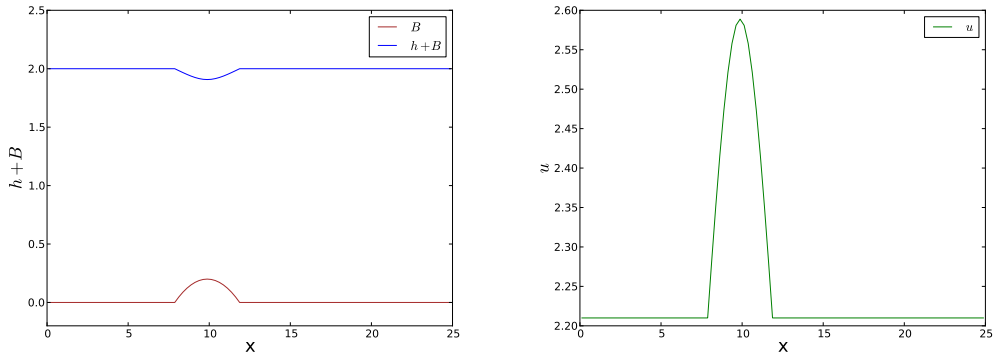


Fig. 3.17: Subcritical equilibrium as suggested in [140]. Left: Bottom topography and total waterheight. Right: velocity.

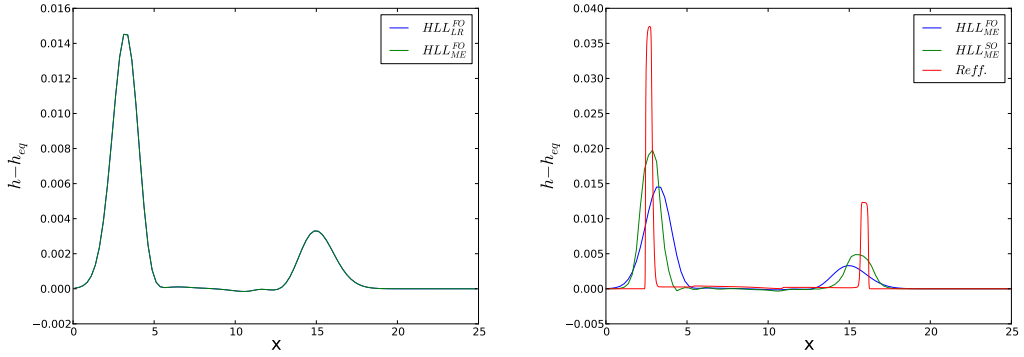


Fig. 3.18: Solutions to the disturbed subcritical moving equilibrium at time 1.5. Left: Comparison of the first order schemes computed with 100 cells. Right: Comparison of the first and second order scheme with respect to a reference solution computed with 5400 cells.

The waterheight can be computed by solving the third order polynomial as explained in section 3.3. The resulting distributions are shown in figure 3.17.

A disturbance is placed left of the bump in the bottom topography as

$$h(0,x) - h(x)_{eq} = \begin{cases} 0.05 & \text{if} \quad 5.75 < x < 6.25, \\ 0 & \text{else}, \end{cases} \tag{3.85}$$

and the resulting waves are computed with the derived schemes, see figure 3.18. In this case, the HLL_{LR} and HLL_{ME} give almost identical results. Therefore, in this case there is not an advantage to consider the general quadrature (3.2). When comparing the different orders, it can bee seen that the resulting waves are captured by both schemes.

In addition to the test proposed by [140], a slight modification is now suggested here. Since the first order schemes are almost identical in the proposed test, it is suggested to consider smaller perturbations, i.e.

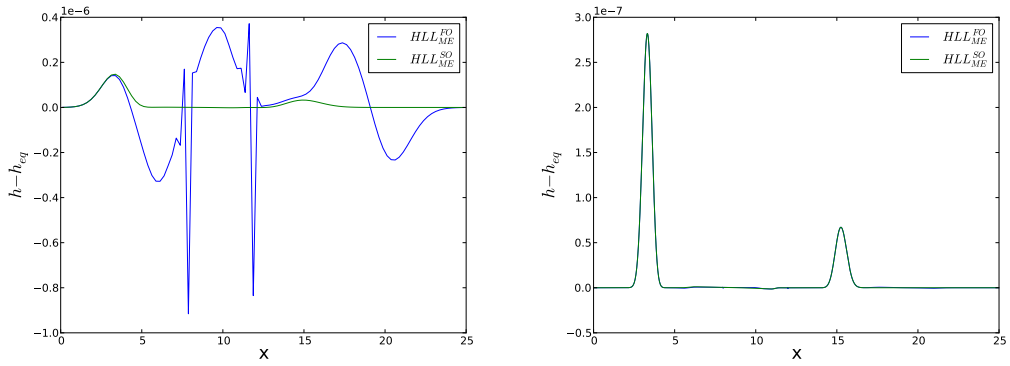


Fig. 3.19: Solutions to the subcritical moving equilibrium with smaller perturbation at time 1.5. Left: Comparison of the first order schemes computed with 100 cells. Right: Comparison of the first order schemes computed with 1000 cells.

$$h(0,x) - h(x)_{eq} = \begin{cases} 0.05 \times 10^{-5} & \text{if} \quad 5.75 < x < 6.25, \\ 0 & \text{else.} \end{cases} \tag{3.86}$$

The results are shown in figure 3.19. When computing these small perturbations on a mesh with 100 cells, the inconsistency of the quadrature (3.38) is starting to influence the numerical results strongly. However, when increasing the resolution the numerical error of the HLL_{LR} scheme falls below the considered dynamics and both schemes again give almost identical results.

The transcritical equilibrium is determined by

$$\begin{aligned} Ce &= 11.7744, \\ Cq &= 1.70507, \\ B(x) &= \begin{cases} 0.2 - 0.05(x-10)^2 & \text{if } 8 < x < 12, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{3.87}$$

The resulting distributions are shown in figure 3.20. The flow is considered to be supercritical left of the bump in the bottom topography , critical at $x = 10$, i.e. the maximum of the bottom topography, and then changes type to subcritical.

Again the perturbation (3.85) is placed on top of the equilibrium and the resulting waves are computed with the derived schemes, see figure 3.21. Here the same results are found as in the subcritical case. For the first order schemes, there is almost no difference between the two variants.

Therefore a smaller perturbation is considered as

$$h(0,x) - h(x)_{eq} = \begin{cases} 0.05 \times 10^{-7} & \text{if} \quad 5.75 < x < 6.25, \\ 0 & \text{else,} \end{cases} \tag{3.88}$$

and again the perturbations are computed as depicted in figure 3.22. The result is similar to the subcritical case. However, it should be remarked that the small perturbations start

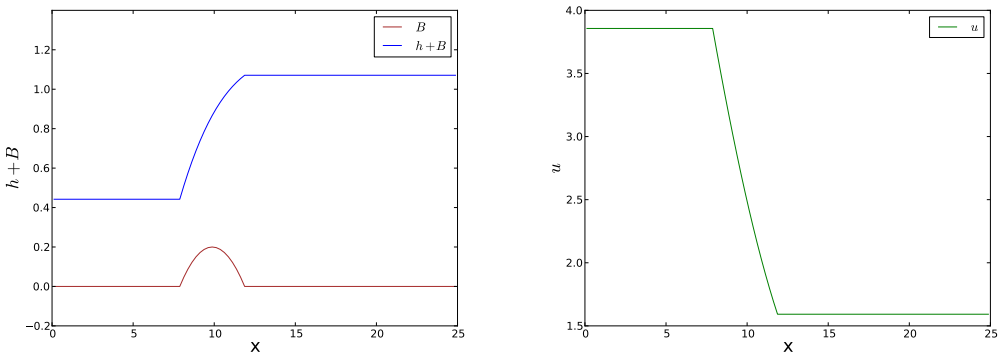


Fig. 3.20: Transcritical equilibrium as suggested in [140]. Left: Bottom topography and total waterheight. Right: velocity.

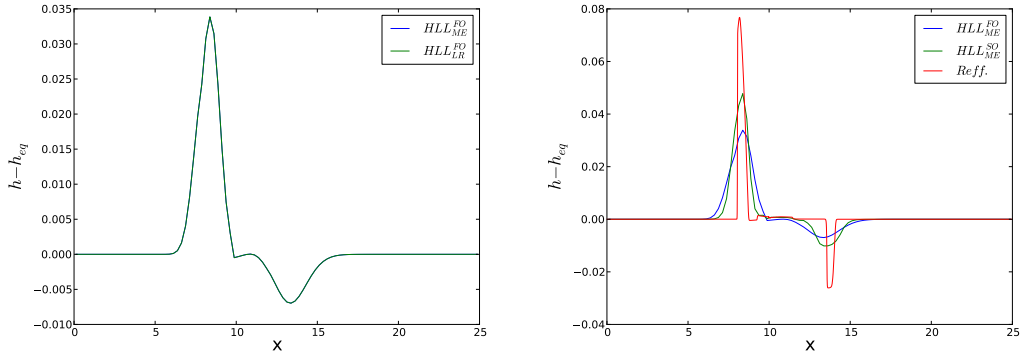


Fig. 3.21: Solutions to the disturbed transcritical moving equilibrium at time 1.5. Left: Comparison of the first order schemes computed with 100 cells. Right: Comparison of the first and second order scheme with respect to a reference solution computed with 5400 cells.

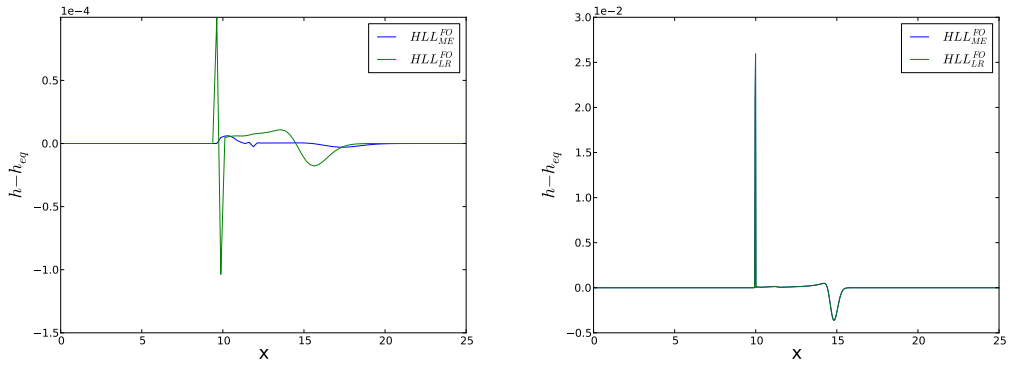


Fig. 3.22: Solutions to the transcritical moving equilibrium with smaller perturbation at time 1.5. Left: Comparison of the first order schemes computed with 100 cells. Right: Comparison of the first order schemes computed with 1000 cells.

forming a quite steep shock at the critical point. If this behavior is physical is questionable. The here derived scheme surely misses robustness and stability properties. It may be interesting how such properties may constrain the numerical scheme and therefore influence the numerical approximations.

4 A Well-Balanced Suliciu Relaxation Scheme for the Euler Equations with Gravity

This section deals with the derivation of a numerical scheme to approximate the solutions of the Euler equations with gravity. The system has been introduced in section 1.4 and reads

$$\begin{cases} \rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \\ (\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + Ip) = -\rho \nabla \Phi, \\ E_t + \nabla \cdot (\mathbf{u}(E+p)) = -\rho \langle u, \nabla \Phi \rangle, \\ \Phi_t = 0, \end{cases} \tag{4.1}$$

where $\rho(\mathbf{x}, t) > 0$ denotes the density, $\mathbf{u}(\mathbf{x}, t) \in \mathbb{R}$ the velocity, $E(\mathbf{x}, t) > 0$ the total energy given by

$$E = \rho e + \frac{1}{2}\rho \mathbf{u}^2,$$

where $e > 0$ is the internal energy, and the function $\Phi : \mathbb{R}^3 \to \mathbb{R}$ is a given smooth gravitational potential. Following section 1.3.1, the pressure is assumed to satisfy some thermodynamical properties and the system admits a family of convex entropies as

$$(\rho \mathcal{F}(\eta))_t + \nabla \cdot \rho \mathcal{F}(\eta) \mathbf{u} \leq 0, \tag{4.2}$$

for η the specific entropy and any convex function $\mathcal{F}$. Moreover, it admits a set of physical relevant states as

$$\Omega_{Phys} = \{(\rho, \rho \mathbf{u}, E) \in \mathbb{R}^5; \rho > 0, e > 0\}. \tag{4.3}$$

Of specific interest are approximations of near hydrostatic equilibrium solutions to (4.1). The hydrostatic equilibria have been derived in (1.61) and are here recast as

$$\begin{cases} \mathbf{u} = 0, \\ \nabla p = -\rho \nabla \Phi. \end{cases} \tag{4.4}$$

As mentioned in section 1.4, system (4.4) is underdetermined, since the pressure depends on the density as well as on the temperature. Therefore, to the best knowledge of the author, additional assumptions have to be made to solve (4.4).

Of specific interest in this chapter are the following classes of solutions to (4.4)

- Isothermal Atmosphere for an ideal Gas Law: $p = \rho RT$

$$\begin{cases} T(x) = const, \\ \rho(x) = \exp(-\frac{\Phi(x)}{RT}), \\ p(x) = RT \exp(-\frac{\Phi(x)}{RT}). \end{cases} \tag{4.5}$$

- Polytropic Atmosphere $p = K\rho^{\Gamma}$ for $\Gamma \in (0,1) \cup (1,\infty)$

$$\begin{cases} \rho(x) = \left(\frac{\Gamma-1}{\Gamma K}(C - \Phi(x))\right)^{\frac{1}{\Gamma-1}}, \\ p(x) = K^{\frac{1}{1-\Gamma}} \left(\frac{\Gamma-1}{\Gamma}(C - \Phi(x))\right)^{\frac{\Gamma}{\Gamma-1}}, \end{cases} \tag{4.6}$$

which have been derived and discussed in section 1.4.

The focus in this chapter is to derive a well-balanced numerical scheme for the hydrostatic equilibria given in (4.5) and (4.6). Moreover, the numerical scheme is designed to give physical relevant and entropy stable approximations of the weak solutions of (4.1).

Numerous techniques were proposed in the literature to derive well-balanced schemes. Most of them concerned the shallow-water equations, see chapter 3. However, in the case of the Euler equations with gravity, the derivation of well-balanced schemes is more delicate since the steady states are in general only given by the underdetermined PDE system (4.4).

A unique approach is given by Cargo and LeRoux [29]. They show that the system (4.1) in the case of one space dimension and a linear gravitational potential can be rewritten into a homogeneous system. Classical methods for conservation laws can now be applied and the well-balanced property comes directly from the consistency of the numerical flux function. However, it seems very hard to extend that approach to more than one space dimension and a nonlinear gravitational potential. Although Cargo-LeRoux's technique was recently revisited in [37], where a suitable relaxation technique was used. Another technique is based on a local hydrostatic reconstruction and is applied [89], where the focus is on preserving the isentropic steady states. Another hydrostatic reconstruction technique is used in [34] to derive a well-balanced scheme for the isothermal and polytropic states.

This chapter are very close to the publication [55], which, following the companion paper [56] devoted to the Ripa model, is a result of the author's collaboration. The strategy is to develop a Suliciu-type relaxation scheme, see section 2.2.4, that is consistent with the hydrostatic equilibria (4.5) and (4.6).

The chapter is organized as follows: section 4.1 is devoted to the derivation of the relaxation model, which is an extension of the work in [56]; section 4.2 concerns the robustness and the well-balanced properties; section 4.3 is devoted to prove that the derived approximate Riemann solver is consistent with the entropy inequalities (4.2) in the sense of Harten, Lax and van Leer [78]; in section 4.4, the Godunov-type scheme associated with the derived approximate solver is presented and in Section 4.5 numerical experiments are performed to investigate the performance of the scheme in practical applications.

4.1 Derivation of the Suliciu Relaxation Model

In order to design a well-balanced approximate Riemann solver, the Suliciu relaxation technique from section 2.2.4 is adopted. The derivation goes along the lines of [56], where a relaxation model was developed in the framework of the the Ripa model. However, regarding the robustness, stability and well-balancedness of the scheme, the Ripa model and the Euler equations with gravity give different challenges.

According to the Suliciu relaxation approach, the pressure p is approximated by a new

variable π governed by the following evolution law:

$$\pi_t + u\pi_x + \frac{c^2}{\rho} u_x = \frac{1}{\varepsilon} \left(p(\tau, e) - \pi \right). \tag{4.7}$$

The relaxation parameter $c > 0$ will be fixed later in order to satisfy some robustness and stability conditions.

Consider now the system (4.1) in one space dimension, where the pressure is approximated by the relation (4.7)

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + \pi)_x = -\rho \Phi_x, \\ E_t + (u(E + \pi))_x = -\rho u \Phi_x, \\ (\rho \pi)_t + (u\pi)_x + c^2 u_x = \frac{\rho}{\varepsilon} \left(p(\tau, e) - \pi \right), \\ \Phi_t = 0. \end{cases} \tag{4.8}$$

According to the strategy developed in section 2.2.4, for the definition of the numerical fluxes it is sufficient to compute the Riemann problem at the cell interfaces according to the system (4.8) in the limit of $\varepsilon \to \infty$, given as

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + \pi)_x = -\rho \Phi_x, \\ E_t + (u(E + \pi))_x = -\rho u \Phi_x, \\ (\rho \pi)_t + (u\pi)_x + c^2 u_x = 0, \\ \Phi_t = 0. \end{cases} \tag{4.9}$$

Lemma 4.1.1. *For $c > 0$, the system (4.9) is hyperbolic with linear degenerate eigenvalues $\lambda_i \in \{0, u, u \pm \frac{c}{\rho}\}$, where $\lambda = u$ has multiplicity 2. The respective Riemann invariants are*

$$\begin{aligned} \Psi_0 &\in \{\rho u, \pi + \frac{c^2}{\rho}, e - \frac{\pi^2}{c^2}, \phi + \frac{u^2}{2} - \frac{c^2}{2\rho^2}\}, \\ \Psi_u &\in \{u, \pi, \Phi\}, \\ \Psi_{u \pm \frac{c}{\rho}} &\in \{u \pm \frac{c}{\rho}, \pi \mp cu, e - \frac{\pi^2}{c^2}, \Phi\}. \end{aligned} \tag{4.10}$$

The proof involves lengthy but straightforward computations and is omitted for brevity, see for example [23] for details. Following lemma 4.1.1, system (4.9) admits a piecewise constant solution to the Riemann problem as suggested in the model (2.30), see also figure 4.1.

However, there are two practical issues associated with the resolution of the Riemann problem. First, the Riemann invariants due to the 0 wave introduce non-linearities and second, the ordering of the waves is not known in advance. These problems can make the process of finding a solution quite cumbersome. However, in [164] a detailed analysis of the solution to (4.9) is presented.

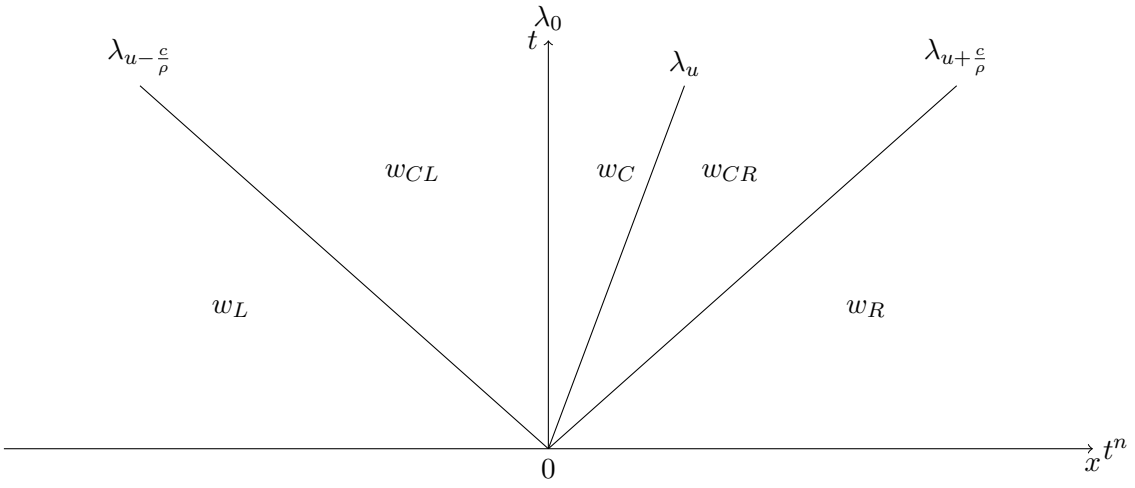


Fig. 4.1: Solution to a Riemann problem for the relaxation system (4.9). A wave with velocity 0 is added due to the source term.

The approach presented here seeks to avoid the technicalities needed to find a solution to (4.9). To this end, in order to fix the issue of the ordering of the eigenvalues, it is suggested to apply an additional relaxation approximation for the gravitational potential Φ by a new variable Z as

$$Z_t + uZ_x = \frac{1}{\varepsilon}(\Phi - Z). \tag{4.11}$$

This now leads to the following relaxation model

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + \pi)_x = -\rho Z_x, \\ E_t + (u(E+\pi))_x = -\rho u Z_x, \\ (\rho\pi)_t + (u\pi)_x + c^2 u_x = 0, \\ Z_t + uZ_x = 0, \end{cases} \tag{4.12}$$

where the relaxation source terms have already been omitted due to section 2.2.4. However, it should be remarked that the full equations (4.7) and (4.11) together with the system (4.12) guarantee for the consistency property from definition 1.2.2.

Lemma 4.1.2. *For $c > 0$, the system (4.12) is hyperbolic with linear degenerate eigenvalues $\lambda_i \in \{u, u \pm \frac{c}{\rho}\}$, where $\lambda = u$ has multiplicity 3. The respective Riemann invariants are*

$$\begin{aligned} \Psi_u &\in \{u\}, \\ \Psi_{u\pm\frac{c}{\rho}} &\in \{u \pm \frac{c}{\rho}, \pi \mp cu, e - \frac{\pi^2}{c^2}, \Phi\}. \end{aligned} \tag{4.13}$$

The proof again is skipped for brevity. According to lemma 4.1.2, the system (4.12) admits a piecewise constant solution as

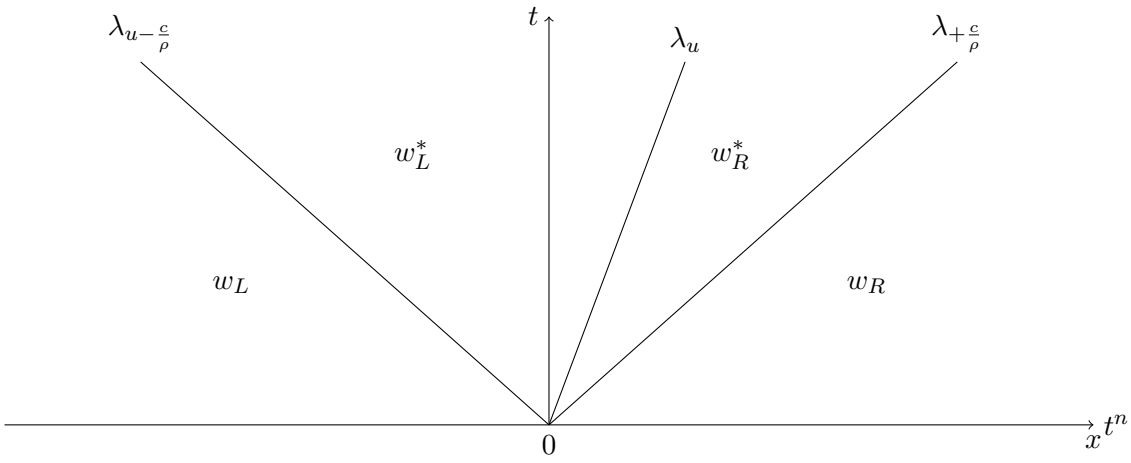


Fig. 4.2: Solution to a Riemann problem for the relaxation system (4.12). The source term is now advected with the fluid velocity u.

$$W_{\mathcal{R}}\left(\frac{x}{t}; W_L, W_R\right) = \begin{cases} W_L & \text{if } x/t < \lambda_{u-\frac{c}{\rho}}, \\ W_L^* & \text{if } \lambda_{u-\frac{c}{\rho}} < x/t < \lambda_u, \\ W_R^* & \text{if } \lambda_u < x/t < \lambda_{u+\frac{c}{\rho}}, \\ W_R & \text{if } \lambda_{u+\frac{c}{\rho}} < x/t. \end{cases} \tag{4.14}$$

, see also figure 4.2.

In contrast to the solution to the system (4.9), as long as $c > 0$, the ordering of the waves is known a priori and the Riemann invariants in (4.14) admit no non-linearities. However, there are ten intermediate states W_L^*, W_R^*, but the system (4.12) only admits nine invariants in (4.13). Therefore, the approximate Riemann solver (4.14) is not uniquely defined. It is suggested to make use of this additional degree of freedom and an additional Riemann invariant is imposed on the wave with speed λ_u. The aim is to derive a well-balanced scheme for the system (4.1). Therefore it is beneficial to impose a discretization of the hydrostatic equilibrium relations (4.4) onto the approximate Riemann solver. On the continuum, the relaxed hydrostatic equilibrium relations are

$$\pi_x = -\rho Z_x. \tag{4.15}$$

As a discretization of (4.15) it is suggested to use

$$\pi_R^\star - \pi_L^\star = -\overline{\rho}(W_L, W_R)(Z_R^\star - Z_L^\star), \tag{4.16}$$

to impose as a Riemann invariant across the wave λ_u. The function $\overline{\rho} : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ denotes a ρ-average function that satisfies the following consistency and symmetry properties:

$$\begin{aligned} \rho_L = \rho_R = \rho \quad \Rightarrow \quad & \overline{\rho}(W_L, W_R) = \rho, \\ & \overline{\rho}(W_R, W_L) = \overline{\rho}(W_L, W_R). \end{aligned} \tag{4.17}$$

Equipped with the Riemann invariants (4.13) and the additional closure relation (4.16),

the intermediate states can be found explicitly.

Lemma 4.1.3. *The Riemann problem associated with the system (4.12) completed by the relation (4.16) admits a unique solution that satisfies the structure (4.14) and the intermediate states W_L^* and W_R^* are defined by*

$$\begin{aligned}
&Z_L^* = Z_L, \qquad Z_R^* = Z_R,\\
&u^* = u_L^* = u_R^* = \frac{1}{2}(u_L + u_R) - \frac{\pi_R - \pi_L + \overline{\rho}(W_L, W_R)(Z_R - Z_L)}{2c},\\
&\pi_L^* = \pi_L + c\frac{u_L - u_R}{2} + \frac{\pi_R - \pi_L + \overline{\rho}(W_L, W_R)(Z_R - Z_L)}{2},\\
&\pi_R^* = \pi_R + c\frac{u_L - u_R}{2} - \frac{\pi_R - \pi_L + \overline{\rho}(W_L, W_R)(Z_R - Z_L)}{2},\\
&\frac{1}{\rho_L^*} = \frac{1}{\rho_L} + \frac{1}{c}(u^* - u_L), \qquad \frac{1}{\rho_R^*} = \frac{1}{\rho_R} + \frac{1}{c}(u_R - u^*),\\
&e_L^* = e_L + \frac{1}{2c^2}\left({\pi_L^*}^2 - {\pi_L}^2\right), \qquad e_R^* = e_R + \frac{1}{2c^2}\left({\pi_R^*}^2 - {\pi_R}^2\right).
\end{aligned} \tag{4.18}$$

Once again, the proof is skipped for brevity.

To artificially impose the closure (4.16) might seem arbitrary. However, in [56] a different relaxation model has been proposed to make this process rigorous. In fact, imposing an additional relaxation process on the source term leads to the following system

$$\begin{cases}
\rho_t + (\rho u)_x = 0,\\
(\rho u)_t + (\rho u^2 + p)_x = -\overline{\rho}(\rho^-, \rho^+) Z_x,\\
E_t + (u(E+p))_x = -\overline{\rho}(\rho^-, \rho^+) u Z_x,\\
(\rho\pi)_t + (u\pi)_x + c^2 u_x = 0,\\
Z_t + u Z_x = 0,\\
\rho_t^- + (u - \frac{c}{\rho} - \delta)\rho_x^- = 0,\\
\rho_t^+ + (u + \frac{c}{\rho} + \delta)\rho_x^+ = 0,
\end{cases} \tag{4.19}$$

for some $\delta > 0$. It is shown, that system (4.19) admits essentially the same solution as system (4.12) with the additional closure (4.16). For more details on that see [56].

4.2 Robustness and Well-Balanced Properties

This section is concerned with the robustness and the well-balanced properties of the approximate Riemann solver defined by (4.12). First, it is shown that the approximate Riemann solver is consistent with the physical relevant set (4.3)

Lemma 4.2.1 (Robustness)**.** *Let be given $W_L, W_R \in \Omega_{Phys}$, defined by (4.3), for the relaxation parameter c large enough, there is $W_L^*, W_R^* \in \Omega_{Phys}$.*

Proof. *The proof follows exactly the lines of the proof of lemma 2.2.2. The positivity of the densities ρ_R^*, ρ_L^* is equivalent, since the source term does not appear in the respective*

Riemann invariants. The formulas for the internal energies rewrite as

$$e_{CR} = e_R - \frac{\pi_R \bar{s} + \frac{\bar{s}^2}{4}}{2c^2} + \frac{(u_L - u_R)(\pi_R - \bar{s})}{2c} + \frac{(u_L - u_R)^2}{8},$$
$$e_{CL} = e_L + \frac{\pi_L \bar{s} + \frac{\bar{s}^2}{4}}{2c^2} + \frac{(u_L - u_R)(\pi_L + \bar{s})}{2c} + \frac{(u_L - u_R)^2}{8},$$

where $\bar{s} = \pi_R - \pi_L + \overline{\rho}(W_L, W_R)(Z_R - Z_L)$. *Since the terms that are independent of* c *are all positive, choosing* c *large enough ensures the positivity.*

Now the well-balanced property of the approximate Riemann solver is discussed.

Lemma 4.2.2 (Well-Balancedness)**.** *Let* w_L *and* w_R *be given in* Ω_{Phys} *such that*

$$\begin{cases} u_L = u_R = 0, \\ p_R - p_L + \overline{\rho}(W_L, W_R)(\Phi_R - \Phi_L) = 0. \end{cases} \tag{4.20}$$

Then the approximate Riemann solver is at rest, i.e. satisfies relation (2.82) and is therefore well-balanced.

Proof. *When looking at the intermediate states (4.18) the proof is straightforward. First, the relations (4.20) give*

$$u^* = 0 \qquad \text{and} \qquad \pi^*_{L,R} = \pi_{L,R}.$$

With these it is then also straightforward to see that $\rho^*_{L,R} = \rho_{L,R}$ *and* $e^*_{L,R} = e_{L,R}$.

The description of a discrete steady state in (4.20) is quite general. If a steady state defined on a continuum satisfies the relation (4.20) strongly depends on the choice of the projection onto the discrete data and the choice of the function $\overline{\rho}$. Lemma 4.2.3 specifies how to choose the function $\overline{\rho}$ such that it gives a well-balanced scheme for the isothermal and polytropic equilibria when the data is projected pointwise onto the cell centers as mentioned in chapter 3. However, the class of hydrostatic equilibria is rich and the general formulation of lemma 4.2.2 admit, that in practical applications other functions for $\overline{\rho}$ may be found to satisfy the respective hydrostatic relation. Additionally, the here presented suggestions for $\overline{\rho}$ are at least second order approximations to any hydrostatic equilibrium and therefore, if the application allows for some errors due to the source term discretization, the averages from lemma 4.2.3 may also be applied.

Lemma 4.2.3. *1. Let* $W_L, W_R \in \Omega_{Phys}$ *and satisfying the isothermal equilibrium relations (4.5), i.e.*

$$\begin{cases} u_L = u_R = 0, \\ \rho_{L,R} = \exp\left(\frac{C - \Phi_{L,R}}{K}\right), \\ p_{L,R} = K \exp\left(\frac{C - \Phi_{L,R}}{K}\right), \end{cases} \tag{4.21}$$

with $K > 0$ *and* $C \in \mathbb{R}$. *Assume that* $\overline{\rho}$ *is defined by*

$$\overline{\rho}(W_L, W_R) = \begin{cases} \dfrac{\rho_R - \rho_L}{\ln(\rho_R) - \ln(\rho_L)} & \text{if } \rho_L \neq \rho_R, \\ \rho_L & \text{if } \rho_L = \rho_R, \end{cases} \tag{4.22}$$

then the approximate Riemann solver satisfies relation (2.82) and is therefore well-balanced for the isothermal equilibria (4.5).

2. *Let $W_L, W_R \in \Omega_{Phys}$ and satisfying the polytropic equilibrium relations (4.6), i.e.*

$$\begin{cases} u_L = u_R = 0, \\ \rho_{L,R} = \left(\frac{\Gamma-1}{\Gamma K}(C - \Phi_{L,R})\right)^{\frac{1}{\Gamma-1}}, \\ p_{L,R} = K^{\frac{1}{1-\Gamma}} \left(\frac{\Gamma-1}{\Gamma}(C - \Phi_{L,R})\right)^{\frac{\Gamma}{\Gamma-1}}, \end{cases} \tag{4.23}$$

with $\Gamma \in (0,1) \cup (1,+\infty)$, $K > 0$ and $C \in \mathbb{R}$. Assume that $\overline{\rho}$ is defined by

$$\overline{\rho}(W_L, W_R) = \begin{cases} \dfrac{\Gamma - 1}{\Gamma} \dfrac{\rho_R^\Gamma - \rho_L^\Gamma}{\rho_R^{\Gamma-1} - \rho_L^{\Gamma-1}} & \text{if } \rho_L \neq \rho_R, \\ \rho_L & \text{if } \rho_L = \rho_R, \end{cases} \tag{4.24}$$

then the approximate Riemann solver satisfies relation (2.82) and is therefore well-balanced for the polytropic equilibria (4.6).

Proof. *From the formulas for the intermediate states (4.18), it is sufficient to proof that*

$$p_R - p_L = -\overline{\rho}(W_L, W_R)(\Phi_R - \Phi_L) \tag{4.25}$$

in the respective cases.

In the isothermal case from (4.23) there is

$$\begin{aligned} \Phi_R - \Phi_L &= K(\ln(\rho_R) - \ln(\rho_L)), \\ p_R - p_L &= K(\rho_R - \rho_L). \end{aligned}$$

Together with the isothermal definition (4.22) of $\overline{\rho}(W_L, W_R)$ gives (4.25) .

In the polytropic case from (4.24) there is

$$\begin{aligned} \Phi_R - \Phi_L &= K\frac{\Gamma}{\Gamma - 1}\left(\rho_R^{\Gamma-1} - \rho_L^{\Gamma-1}\right), \\ p_R - p_L &= K\left(\rho_R^\Gamma - \rho_L^\Gamma\right). \end{aligned}$$

Together with the polytropic definition (4.24) of $\overline{\rho}(W_L, W_R)$ gives (4.25)

In [97] and [12] additional strategies have been proposed to define the quadrature. In specific, by allowing the quadrature also depend on the spatial variable x and giving and parameterizing a specific hydrostatic equilibrium, the well-balanced property can be extended to a wider class of hydrostatic equilibria.

4.3 Consistency with the entropy inequalities

This section concerns the entropy stability of the approximate Riemann solver defined by (4.12). The Euler equations with gravity admit an entropy that is non-increasing over time, see lemma 1.3.1. Since the numerical solutions computed by a finite volume scheme at best

only serve as approximations to the underlying PDE, it is desirable to transfer as many properties of PDE as possible to the numerical scheme to compute relevant approximations. The stability with respect to entropy is a non-linear stability property and therefore important since it respects the full non-linear dynamics that are build into the model.

However, this section is technical and therefore a short outline of the general strategy to proof the entropy stability shall be given in advance; also see [15],[21] for similar arguments. First, denote the dependent variables of the Euler system (4.1) as $U = (\rho, \rho u, E, \Phi)$ and the dependent variables of the relaxation system (4.12) as $W = (\rho, \rho u, E, \pi, Z, \Phi)$. The relaxation system admits an equilibrium manifold $\mathcal{M} \in \mathbb{R}^4$, such that on $\mathcal{M}$ there is $\pi = p(\tau, e)$ and $Z = \phi$, which is further denoted as $W_{eq} = \{W | W \in \mathcal{M}\}$.

Now a quantity $\bar{s} = \bar{s}(W)$ is constructed such, under the dynamics of the relaxation system (4.12), it satisfies the following transport relation

$$\bar{s}_t + u\bar{s}_x = 0. \tag{4.26}$$

Moreover, $\bar{s}$ is a conserved quantity. In the model (4.14) this leads in the case of the Riemann problem to the following distribution of $\bar{s}$.

$$\bar{s}\left(\frac{x}{t}; W_L, W_R\right) = \begin{cases} \bar{s}(W_L) & \text{if } x/t < \lambda_{u-\frac{c}{\rho}}, \\ \bar{s}(W_L) & \text{if } \lambda_{u-\frac{c}{\rho}} < x/t < \lambda_u, \\ \bar{s}(W_R) & \text{if } \lambda_u < x/t < \lambda_{u+\frac{c}{\rho}}, \\ \bar{s}(W_R) & \text{if } \lambda_{u+\frac{c}{\rho}} < x/t. \end{cases} \tag{4.27}$$

Then, the quantity $\bar{s}$ is connected to the specific entropy $s = s(U)$ of the Euler system (4.1) such that

$$\bar{s}(W) \geq s(U) \qquad \text{and} \qquad \bar{s}(W_{eq}) = s(U(W_{eq})), \tag{4.28}$$

i.e. the function $\bar{s}$ reaches its minimum on the equilibrium manifold $\mathcal{M}$ and coincides there with the specific entropy of the Euler system. This reflects the two step behavior of the numerical method when dealing with relaxation schemes. First, the data is evolved due to the homogeneous part of the relaxation system. According to (4.26), this is conservative for the entropy $\bar{s}$. Second, the data is projected to its equilibrium manifold $\mathcal{M}$. During the projection, the quantity $\bar{s}$ is nonincreasing and coincides on the manifold with the entropy s of the original system. Therefore, the evolution of s is bounded by the evolution of $\bar{s}$, which is non-increasing, and therefore s is non-increasing, see also figure 4.3. This gives that the approximate Riemann solver given by (4.12) is consistent with the entropy inequalities (4.2) in the sense of Harten, Lax and van Leer [78].

Now, the reasoning given above is made precise. In order for the quantity $\bar{s}$ to satisfy the transport property (4.26), it is beneficial to make it dependent on functions $I(W)$ and $J(W)$ that satisfy the same transport property. In fact, let $I(W)$ and $J(W)$ be defined as follows

$$I(W) := I(\pi, \tau) = \pi + c^2\tau, \tag{4.29}$$

$$J(W) := J(\pi, e) = e - \frac{\pi^2}{2c^2}, \tag{4.30}$$

then $I(W), J(W)$ are strong Riemann invariants of the system (4.12) in the following

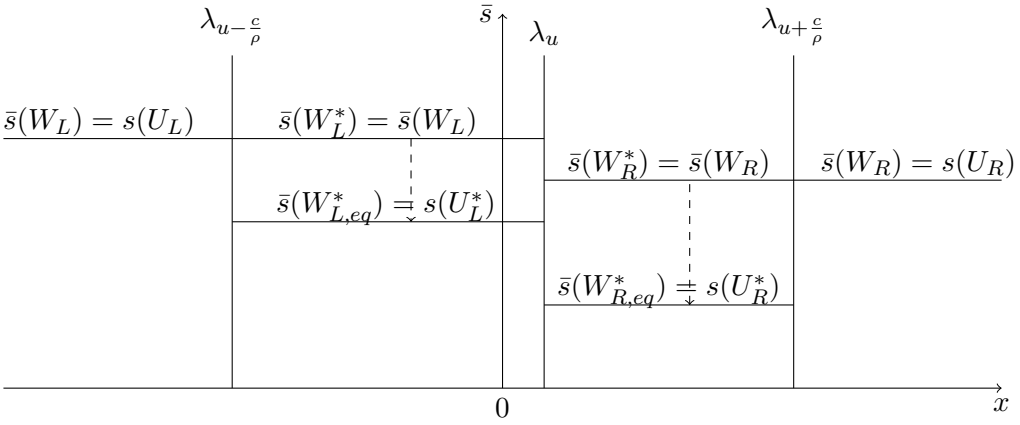


Fig. 4.3: Distribution of $\bar{s}$ at t_{n+1}. The initial condition is on the equilibrium manifold $\mathcal{M}$, therefore $\bar{s}$ and s coincide on the left and right states. The values of $\bar{s}(W_{L,R}^*)$ follow from the transport property. Finally, during the projection step denoted by the dashed lines, $\bar{s}$ is non-increasing.

sense:

Lemma 4.3.1. *The weak solutions of the relaxation model (4.12) satisfy*

$$\partial_t \rho\Psi(I,J) + \partial_x \rho\Psi(I,J)u = 0, \tag{4.31}$$

for all smooth function $\Psi : \mathbb{R}^2 \to \mathbb{R}$.

Proof. *First consider a smooth solution W of the system (4.12). A straightforward computation gives for the internal energy $e = \frac{E}{\rho} - \frac{u^2}{2}$, the specific volume $\tau = \frac{1}{\rho}$ and for the quantity $\frac{\pi^2}{2c^2}$, that there is*

$$\begin{aligned} e_t + \pi\tau u_x + ue_x &= 0, \\ \tau_t + u\tau_x - \tau u_x &= 0, \\ (\frac{\pi^2}{2c^2})_t + \pi\tau u_x + u(\frac{\pi^2}{2c^2})_x &= 0. \end{aligned}$$

From these, it easily holds that

$$\begin{aligned} I_t + uI_x &= 0, \\ J_t + uJ_x &= 0. \end{aligned}$$

Therefore for all smooth functions $\bar{\Psi} : \mathbb{R}^2 \to \mathbb{R}$, the following transport equation is satisfied:

$$\bar{\Psi}_t(I,J) + u\bar{\Psi}_x(I,J) = 0,$$

which gives (4.31) for smooth solutions. To conclude, see that the system (4.12) only has linearly degenerate fields. Therefore the same result also holds true for weak solutions of (4.31).

On the other hand, in order to ensure the consistency property of $\bar{s}$ with respect to s in

(4.28), one would like $\bar{s}$ and s to be of similar form. Since s takes as arguments the specific volume τ and the internal energy e, let $\bar{s}$ be defined as

$$\bar{s} = s(\bar{\tau}, \bar{e}). \tag{4.32}$$

Given the above reasoning, one would like to compute $\bar{\tau} = \bar{\tau}(I, J), \bar{e} = \bar{e}(I, J)$. In fact, given the values I, J and evaluating their definitions (4.29) and (4.30) on the equilibrium manifold gives the two following non-linear equations

$$\begin{aligned} I &= p(\tau, e) + c^2\tau, \\ J &= e - \frac{p(\tau, e)^2}{2c^2}. \end{aligned} \tag{4.33}$$

These two equations may be solved for $\bar{\tau}$ and $\bar{e}$. The following definitions and derivations discuss if the values $\bar{\tau}$ and $\bar{e}$ as solutions to (4.33) are well-defined. Indeed the system (4.33) can be rearranged in the following way

$$\begin{aligned} p(\bar{\tau}, \bar{e}) &= I - c^2\bar{\tau}, \\ \bar{e} &= J + \frac{p(\bar{\tau}, \bar{e})^2}{2c^2}. \end{aligned} \tag{4.34}$$

Using the first in the second equation gives then

$$\bar{e} = J + \frac{(I - c^2\bar{\tau})^2}{2c^2}, \tag{4.35}$$

and again using this in the first equation of (4.34) gives then

$$p(\bar{\tau}, J + \frac{(I - c^2\bar{\tau})^2}{2c^2}) = I - c^2\bar{\tau}. \tag{4.36}$$

Therefore, consider the function $f_{I,J} : \mathbb{R}^+ \to \mathbb{R}$ defined as follows:

$$f_{I,J}(\tau) = \tau p\left(\tau, J + \frac{(I - c^2\tau)^2}{2c^2}\right) + c^2\tau^2 - I\tau. \tag{4.37}$$

Following the above calculations, $\bar{\tau}$ may be defined as a root of the function $f_{I,J}(\tau)$ and $\bar{e}$ is then given by (4.35). If a root of $f_{I,J}(\tau)$ exists is not obvious and depends on the pressure law. Before stating an assumption on the pressure law to ensure the existence of the root, the following set is introduced to specify for which values of (I, J) a solution is searched for

the system (4.33)

$$\mathcal{A} = \Big\{ (I,J) \in \mathbb{R}^2; \quad \exists \tau > 0, \quad \exists e > 0 \text{ such that:}$$
$$I = p(\tau, e) + c^2 \tau, , \tag{4.38}$$
$$J = e - \frac{p(\tau,e)^2}{2c^2}, , \tag{4.39}$$
$$c^2 > p(\tau,e)\,\partial_e p(\tau,e) - \partial_\tau p(\tau,e) \Big\}. \tag{4.40}$$

The inequality (4.40) is the sub-characteristic or Whitham condition [169] and ensures the stability of the relaxation procedure. It imposes that the sound speed $c\tau$ of the system (4.12) has to be greater than the sound speed $\bar{c} = \tau\sqrt{p\partial_e p - \partial_\tau p}$ of the original model (4.1). Moreover, for all $\tau > 0$ and $e > 0$ satisfying (4.40), the definitions (4.29) and (4.30) of I and J imply that the pair $(I(p(\tau,e),\tau), J(p(\tau,e),e))$ belongs to $\mathcal{A}$.

Now the following additional assumption is imposed to be satisfied by the pressure law:

Assumption 4.3.1. *Assume the pressure law is such that the function $\tau \mapsto f_{I,J}(\tau)$, defined by (4.37), is strictly convex for all pair (I,J) fixed in $\mathcal{A}$ and admits at least one root $\bar{\tau}$, which is not a minimum.*

Remark 4.3.1. *Such assumptions are satisfied by the ideal gas law, given by*

$$p(\tau,e) = (\gamma - 1)\frac{e}{\tau}.$$

Then the function $f_{I,J}$ writes

$$f_{I,J}(\tau) = \frac{\gamma+1}{2}c^2\tau^2 - \gamma I\tau + (\gamma-1)J + \frac{\gamma-1}{2c^2}I^2,$$

and $f_{I,J}$ is a second-order polynomial with a positive highest degree coefficient and is therefore strictly convex. Moreover, the roots can be computed as

$$\tau_{1,2} = \frac{2\gamma I}{c^2(\gamma+1)} \pm \frac{1}{c^2(\gamma+1)} \left(\underbrace{4\gamma^2 I^2 - 2c^2(\gamma^2-1)(J + \frac{I^2}{2c^2})}_{=K(I,J)} \right)^{\frac{1}{2}},$$

and it holds that with the definitions (4.33) there is

$$K(I,J) = 4\gamma^2 p^2 + 4c^2\gamma p\tau(\gamma-1) + c^4\tau^2(3\gamma^2+1) \overset{\gamma\geq 1}{>} 0.$$

Since there is at least one pressure law satisfying assumption (4.3.1), the next step is to uniquely define the values $\bar{\tau}$ and $\bar{e}$.

Definition 4.3.1. *When the function $f_{I,J}$ admits more than one root within $\mathbb{R}^+$, let by $\overline{\tau}(I,J)$ be defined the largest one and $\bar{e}$ is defined by*

$$\overline{e}(I,J) = J + \frac{(I - c^2\overline{\tau}(I,J))^2}{2c^2}. \tag{4.41}$$

The following result connects Assumption 4.3.1 with the existence and uniqueness of the pair $(\bar{\tau}, \bar{e})$ as well as their consistency.

Lemma 4.3.2. *Let $c > 0$ be given. When they are defined, the functions $\overline{\tau}$ and $\overline{e}$ satisfy for all (I, J) in $\mathcal{A}$:*

$$I = p(\overline{\tau}(I,J), \overline{e}(I,J)) + c^2 \overline{\tau}(I,J), \tag{4.42}$$

$$J = \overline{e}(I,J) - \frac{p(\overline{\tau}(I,J), \overline{e}(I,J))^2}{2c^2}, \tag{4.43}$$

$$c^2 > p\left(\overline{\tau}(I,J), \overline{e}(I,J)\right) \partial_e p\left(\overline{\tau}(I,J), \overline{e}(I,J)\right) - \partial_\tau p\left(\overline{\tau}(I,J), \overline{e}(I,J)\right). \tag{4.44}$$

Moreover, for all $\tau > 0$ and $e > 0$ which satisfy the Whitham condition (4.40), the reals $\overline{\tau}\left(I(p(\tau,e),\tau)), J(p(\tau,e),e)\right)$ and $\overline{e}\left(I(p(\tau,e),tau)), J(p(\tau,e),e)\right)$ are well-defined and satisfy

$$\overline{\tau}\left(I(p(\tau,e),\tau), J(p(\tau,e),e)\right) = \tau, \tag{4.45}$$

$$\overline{e}\left(I(p(\tau,e),\tau), J(p(\tau,e),e)\right) = e. \tag{4.46}$$

Proof. *By definition of $\overline{\tau}$ as root of $f_{I,J}$, one immediately has relation (4.42). Relation (4.43) follows from (4.42) and the definition of $\overline{e}$ given by (4.41).*

To establish inequality (4.44), compute the derivative of $f_{I,J}$:

$$\frac{\partial}{\partial \tau} f_{I,J}(\tau) = \left(p + \tau \partial_\tau p - (I - c^2\tau)\tau \partial_e p + 2c^2\tau - I\right) \left(\tau, J + \frac{(I - c^2\tau)^2}{2c^2}\right).$$

Using relations (4.42) and (4.43) gives

$$\frac{\partial}{\partial \tau} f_{I,J}(\overline{\tau}) = \overline{\tau}\left(c^2 + \partial_\tau p(\overline{\tau}, \overline{e}) - p(\overline{\tau}, \overline{e}) \partial_e p(\overline{\tau}, \overline{e})\right).$$

Since by assumption 4.3.1 $f_{I,J}$ is strictly convex and its root is not a minimum, there is $\frac{\partial}{\partial \tau} f_{I,J}(\overline{\tau}) > 0$ at its largest root and therefore (4.44) holds.

Finally, given $\tau > 0$ and $e > 0$, it is straightforward to get

$$(f_{I,J})\Big|_{\substack{I=I(p(\tau,e),\tau) \\ J=J(p(\tau,e),e)}} (\tau) = 0,$$

and

$$\left(\frac{\partial}{\partial \tau} f_{I,J}\right)\Big|_{\substack{I=I(p(\tau,e),\tau) \\ J=J(p(\tau,e),e)}} (\tau) = \tau(c^2 + \partial_\tau p(\tau,e) - p(\tau,e)\partial_e p(\tau,e) > 0,$$

since the Whitham condition (4.40) is satisfied. The function $f_{I,J}$ being strictly convex, it cannot have more than one root where the derivative is positive. Consequently (4.45) holds. Relation (4.46) comes directly from the definition (4.41) of $\overline{e}$.

It should be remarked that in general $\bar{\tau}(I(W), J(W)) \neq \tau$ and $\bar{e}(I(W), J(W)) \neq e$. The pair $(\bar{\tau}, \bar{e})$ is artificial and only helps to define the bound on the entropy. However, it has been shown that the system (4.33) can be solved for $\bar{\tau}$ and $\bar{e}$. Especially, when $I = I(W_{eq})$ and $J = J(W_{eq})$, then (4.45) and (4.46) give the consistency of the proposed solution strategy.

Defining the following set

$$\mathcal{E} = \left\{ W \in \mathcal{O}; \quad (I(W), J(W)) \in \mathcal{A}, \quad c^2 > p\left(\tau, e\right) \partial_e p\left(\tau, e\right) - \partial_\tau p\left(\tau, e\right) \right\}. \tag{4.47}$$

allows for a more precise definition of the quantity $\bar{s}$ in contrast to (4.32) as

$$\overline{s}(W) = s(\overline{\tau}(I(W), J(W)), \overline{e}(I(W), J(W))), \tag{4.48}$$

since according to Lemma 4.3.2, the quantities $\overline{\tau}(I(W), J(W))$ and $\overline{e}(I(W), J(W))$ are well-defined as soon as W belongs to the set $\mathcal{E}$. Lemma 4.3.2 together with (4.48) gives immediately the consistency property of the function $\bar{s}$ with respect to s, i.e.

$$\bar{s}(\bar{\tau}(I(W_{eq}), J(W_{eq})), \bar{e}(I(W_{eq}), J(W_{eq}))) = \bar{s}(\tau, e) = s. \tag{4.49}$$

What is left to show is that $\bar{s}$ reaches its minimum when the relaxation variables are on the equilibrium manifold. The following notations are introduced for the sake of simplicity

$$\overline{\tau} = \overline{\tau}(I(W), J(W)), \quad \overline{e} = \overline{e}(I(W), J(W)), \quad \overline{p} = p(\overline{\tau}, \overline{e}). \tag{4.50}$$

Lemma 4.3.3. *For all $W \in \mathcal{E}$, there is*

$$\overline{s}(W) \geq s\left(\tau, e\right). \tag{4.51}$$

Proof. *In order to show (4.51) it is sufficient to show*

$$\partial_\pi \overline{s}(W_{eq}) = 0, \tag{4.52}$$

and

$$\partial_{\pi\pi} \overline{s}(W_{eq}) > 0. \tag{4.53}$$

Therefore, first evaluate the first derivative of $\bar{s}$ with respect to π as

$$\partial_\pi \overline{s} = \partial_\tau \overline{s} \partial_\pi \overline{\tau} + \partial_e \overline{s} \partial_\pi \overline{e}. \tag{4.54}$$

To get the derivatives $\partial_\pi \overline{\tau}$ and $\partial_\pi \overline{e}$ first derive the relations (4.42) and (4.43) from lemma 4.3.2 to get

$$\begin{aligned} \partial_\pi I &= \partial_\pi \overline{\tau} \partial_\tau \overline{p} + \partial_\pi \overline{e} \partial_e \overline{p} + c^2 \partial_\pi \overline{\tau}, \\ \partial_\pi J &= \partial_\pi \overline{e}(I, J) - \frac{\bar{p}}{c^2} (\partial_\pi \overline{\tau} \partial_{\bar{\tau}} \bar{p} + \partial_\pi \overline{e} \partial_e \overline{p}). \end{aligned} \tag{4.55}$$

On the other hand, from (4.29) and (4.30) there is

$$\begin{aligned} \partial_\pi I &= 1, \\ \partial_\pi J &= -\frac{\pi}{c^2}. \end{aligned} \tag{4.56}$$

Using (4.55) and (4.56) gives then

$$\begin{aligned}\partial_\pi \overline{\tau} &= \frac{(\overline{p}-\pi)\partial_e \overline{p} - c^2}{c^2\left(\overline{p}\partial_e \overline{p} - \partial_\tau \overline{p} - c^2\right)},\\ \partial_\pi \overline{e} &= \frac{1}{\partial_e \overline{p}}\left(1 - (\partial_\tau \overline{p} + c^2)\partial_\pi \overline{\tau}\right).\end{aligned} \tag{4.57}$$

Therefore (4.54) can be rewritten to get

$$\partial_\pi \overline{s} = \frac{(\overline{p}-\pi)(\partial_\tau \overline{s}\partial_e \overline{p} - \partial_\tau \overline{p}\partial_e \overline{s} - c^2\partial_e \overline{s}) + c^2(\overline{p}\partial_e \overline{s} - \partial_\tau \overline{s})}{c^2(\overline{p}\partial_e \overline{p} - \partial_\tau \overline{p} - c^2)}. \tag{4.58}$$

However, by the definition of the specific entropy (1.35), there is

$$\partial_\tau s = p\partial_e s, \tag{4.59}$$

and the relation (4.58) can be further simplified to get

$$\partial_\pi \overline{s} = \frac{\overline{p}-\pi}{c^2}\partial_e \overline{s}. \tag{4.60}$$

This proofs (4.52), since due to the consistency relations (4.45) and (4.46)

$$\begin{aligned}\overline{p}_{|\pi=p(\tau,e)} &= p\left(\overline{\tau}\left(p(\tau,e) + c^2\tau, e - \frac{p(\tau,e)^2}{2c^2}\right), \overline{e}\left(p(\tau,e) + c^2\tau, e - \frac{p(\tau,e)^2}{2c^2}\right)\right)\\ &= p(\tau,e).\end{aligned}$$

Finally, deriving (4.60) again with respect to π, there is

$$\partial_{\pi\pi}\overline{s} = \frac{\partial_e \overline{s}}{c^2}\left(\partial_e \overline{s}\partial_\tau \overline{p} + (\overline{p}-\pi)\partial_{\tau e}\overline{s}\right)\partial_\pi \overline{\tau} + \frac{1}{c^2}\left(\partial_e \overline{s}\partial_e \overline{p} + (\overline{p}-\pi)\partial_{ee}\overline{s}\right)\partial_\pi \overline{e} - \frac{\partial_e \overline{s}}{c^2}.$$

Use again the relations (4.57) to get

$$\begin{aligned}\partial_{\pi\pi}\overline{s} = \frac{1}{a^4(\overline{p}\partial_e \overline{p} - \partial_\tau \overline{p} - c^2)}\Big((\overline{p}-\pi)^2\left(\partial_e \overline{p}\partial_{\tau e}\overline{s} - \partial_{ee}\overline{s}\partial_\tau \overline{p} - c^2\partial_{ee}\overline{s}\right)\\ + c^2(\overline{p}-\pi)\left(\overline{p}\partial_{ee}\overline{s} - \partial_{\tau e}\overline{s} - \partial_e \overline{p}\partial_e \overline{s}\right) + a^4\partial_e \overline{s}\Big).\end{aligned} \tag{4.61}$$

Furthermore deriving (4.59) with respect to e gives

$$\partial_{\tau e}s = p\partial_{ee}s + \partial_e p\partial_e s,$$

to finally get

$$\partial_{\pi\pi}\overline{s} = \frac{(\overline{p}-\pi)^2}{c^4}\partial_{ee}\overline{s} + \frac{\partial_e \overline{s}}{c^4(\overline{p}\partial_e \overline{p} - \partial_\tau \overline{p} - c^2)}\left((\overline{p}-\pi)\partial_e \overline{p} - c^2\right)^2. \tag{4.62}$$

Since $(\tau, e) \mapsto s(\tau, e)$ is a strictly convex function, $\partial_{ee}\overline{s} > 0$. On the other hand, from the inequalities (1.35), there is $\partial_e \overline{s} < 0$. Since the sub-characteristic condition (4.40) is assumed to hold, there finally is $\partial_{\pi\pi}\overline{s} > 0$, which concludes the proof.

Equipped with the minimization principle (4.51) and the consistency relation (4.49), it is straightforward to see that since $I_{L,R} = I^*_{L,R}$ and $J_{L,R} = J^*_{L,R}$, the relation (4.27) holds true, i.e.

$$\begin{aligned} \bar{s}(\bar{\tau}(I(W_L), J(W_L)), \bar{e}(I(W_L), J(W_L))) &= \bar{s}(\bar{\tau}(I(W^*_L), J(W^*_L)), \bar{e}(I(W^*_L), J(W^*_L))), \\ \bar{s}(\bar{\tau}(I(W_R), J(W_R)), \bar{e}(I(W_R), J(W_R))) &= \bar{s}(\bar{\tau}(I(W^*_R), J(W^*_R)), \bar{e}(I(W^*_R), J(W^*_R))). \end{aligned} \tag{4.63}$$

Moreover, if $W_{L,R}$ and $W^*_{L,R}$ belong to $\mathcal{E}$, then the following relations hold

$$\begin{aligned} \bar{s}(\bar{\tau}(I(W_L), J(W_L)), \bar{e}(I(W_L), J(W_L))) &= s(\tau_L, e_L), \\ \bar{s}(\bar{\tau}(I(W^*_L), J(W^*_L)), \bar{e}(I(W^*_L), J(W^*_L))) &\geq s(\tau^*_L, e^*_L), \\ \bar{s}(\bar{\tau}(I(W^*_R), J(W^*_R)), \bar{e}(I(W^*_R), J(W^*_R))) &\geq s(\tau^*_R, e^*_R), \\ \bar{s}(\bar{\tau}(I(W_R), J(W_R)), \bar{e}(I(W_R), J(W_R))) &= s(\tau_R, e_R). \end{aligned} \tag{4.64}$$

This gives the bound on the entropy for the approximate Riemann solver derived from the relaxation system (4.12) and the following theorem can be proven.

Theorem 4.3.1. *Let W_L and W_R be two states of Ω_{phys}. Consider a smooth function $\mathcal{F}$ such that the hypotheses (1.38) are satisfied. Let $c > 0$ be a parameter such that the following sub-characteristic Whitham conditions hold:*

$$c^2 > p(\tau_L, e_L)\,\partial_e p(\tau_L, e_L) - \partial_\tau p(\tau_L, e_L)\,,, \tag{4.65a}$$
$$c^2 > p(\tau^*_L, e^*_L)\,\partial_e p(\tau^*_L, e^*_L) - \partial_\tau p(\tau_L *, e^*_L)\,,, \tag{4.65b}$$
$$c^2 > p(\tau^*_R, e^*_R)\,\partial_e p(\tau^*_R, e^*_R) - \partial_\tau p(\tau^*_R, e^*_R)\,,, \tag{4.65c}$$
$$c^2 > p(\tau_R, e_R)\,\partial_e p(\tau_R, e_R) - \partial_\tau p(\tau_R, e_R)\,.. \tag{4.65d}$$

Fix $\Delta t > 0$ and $\Delta x > 0$ two constants such that the following (CFL) restriction is satisfied:

$$\frac{\Delta t}{\Delta x} \max\left\{ \left| u_L - \frac{c}{\rho_L} \right|, \left| u_R + \frac{c}{\rho_R} \right| \right\} \leq \frac{1}{2}. \tag{4.66}$$

Moreover, assume that the pressure law satisfies Assumption 4.3.1. Then the approximate Riemann solver (4.14) satisfies the inequalities

$$\begin{aligned} \frac{1}{\Delta x} \int_0^{\Delta x/2} (\rho \mathcal{F}(s)) \left(W_{\mathcal{R}}\left(\frac{x}{\Delta t}; w_L, w_R \right) \right) dx \leq \frac{\rho_R \mathcal{F}(s_R)}{2} \\ - \frac{\Delta t}{\Delta x} \left(\rho_R \mathcal{F}(s_R) u_R - \{\rho \mathcal{F}(s) u\}_{L,R} \right), \end{aligned} \tag{4.67}$$

$$\begin{aligned} \frac{1}{\Delta x} \int_{-\Delta x/2}^{0} (\rho \mathcal{F}(s)) \left(W_{\mathcal{R}}\left(\frac{x}{\Delta t}; w_L, w_R \right) \right) dx \leq \frac{\rho_L \mathcal{F}(s_L)}{2} \\ - \frac{\Delta t}{\Delta x} \left(\{\rho \mathcal{F}(s) u\}_{L,R} - \rho_L \mathcal{F}(s_L) u_L \right), \end{aligned} \tag{4.68}$$

where there is

$$\{\rho\mathcal{F}(s)u\}_{L,R} = \begin{cases} \rho_L\mathcal{F}(s_L)u_L & \text{if } 0 < u_L - \frac{c}{\rho_L}, \\ \rho_L^*\mathcal{F}(s_L)u^* & \text{if } u_L - \frac{c}{\rho_L} < 0 < u^*, \\ \rho_R^*\mathcal{F}(s_R)u^* & \text{if } u^* < 0 < u_R + \frac{c}{\rho_R}, \\ \rho_R\mathcal{F}(s_R)u_R & \text{if } u_R + \frac{c}{\rho_R} < 0. \end{cases} \tag{4.69}$$

Proof. *Consider the weak solutions of the relaxation model (4.12), with an initial condition given by*

$$W(x,0) = \begin{cases} W_{eq}(U_L) & \text{if } x < 0, \\ W_{eq}(U_R) & \text{if } x > 0. \end{cases}$$

The function $W \mapsto \overline{s}(W)$, defined by (4.48), only depends of I and J, so Lemma 4.3.1 ensures that the weak solutions of (4.12) satisfy the additional following conservation law:

$$\partial_t \rho\mathcal{F}(\overline{s}) + \partial_x \rho\mathcal{F}(\overline{s})u = 0.$$

Integrate this equation over $[0, \Delta x/2) \times [0, \Delta t)$ to get

$$\begin{aligned} \int_0^{\Delta x/2} (\rho\mathcal{F}(\overline{s}))\left(W_{\mathcal{R}}\left(\frac{x}{\Delta t}; W_{eq}(U_L), W_{eq}(U_R)\right)\right) dx = \\ \int_0^{\Delta x/2} (\rho\mathcal{F}(\overline{s}))(W(x,0))dx - \Delta t(\rho\mathcal{F}(\overline{s})u)\left(W_{\mathcal{R}}\left(\frac{\Delta x}{2\Delta t}; W_{eq}(U_L), W_{eq}(U_R)\right)\right) \\ + \Delta t(\rho\mathcal{F}(\overline{s})u)\left(W_{\mathcal{R}}\left(0; W_{eq}(U_L), W_{eq}(U_R)\right)\right). \end{aligned} \tag{4.70}$$

Since the state $W_{eq}(U_R)$ is at the relaxation equilibrium, the following sequence of equalities holds for $x \in [0, \Delta x/2)$:

$$(\rho\mathcal{F}(\overline{s}))(W(x,0)) = (\rho\mathcal{F}(\overline{s}))(W_{eq}(U_R)) = (\rho\mathcal{F}(s))(W_{eq}(U_R)) = \rho_R\mathcal{F}(s_R). \tag{4.71}$$

On the other hand, the CFL restriction (4.66) implies for all $x \in [0, \Delta x/2)$:

$$W_{\mathcal{R}}\left(\frac{\Delta x}{2\Delta t}; W_{eq}(U_L), W_{eq}(U_R)\right) = W_{eq}(U_R).$$

As a consequence, the first flux term in (4.70) writes

$$(\rho\mathcal{F}(\overline{s})u)\left(W_{\mathcal{R}}\left(\frac{\Delta x}{2\Delta t}; W_{eq}(U_L), W_{eq}(U_R)\right)\right) = \rho_R\mathcal{F}(s_R)u_R. \tag{4.72}$$

Now, since the relations (4.63) and (4.64) are valid and the function $\mathcal{F}$ is increasing due to (1.38), there is

$$\mathcal{F}(\overline{s})\left(W_{\mathcal{R}}\left(\frac{x}{\Delta t}; W^{eq}(w_L), W^{eq}(w_R)\right)\right) \geq \mathcal{F}(s)\left(w^{eq}\left(\frac{x}{\Delta t}; w_L, w_R\right)\right),$$

which gives the inequality (4.67).

The proof for the inequality (4.68) is similar and therefore omitted.

With the inequalities (4.67) and (4.68), theorem 2.2.2 can be applied to prove the entropy stability of the scheme.

4.4 Definition of the Numerical Scheme

Following the lines of section 2.4, the Euler equations with gravity (4.1) are discretized as given in (2.80), i.e.

$$U_{i,t} + \frac{1}{\Delta_{x_i}}(F^+_{i-\frac{1}{2}} - F^-_{i+\frac{1}{2}}) = \frac{1}{\Delta_{x_i}} \int_{x_{i-\frac{1}{2}}}^{x_i} S_i(\mathcal{W}_{i-\frac{1}{2}}(t,x)) Z_x dx + \frac{1}{\Delta_{x_i}} \int_{x_i}^{x_{i+\frac{1}{2}}} S_i(\mathcal{W}_{i+\frac{1}{2}}(t,x)) Z_x dx. \tag{4.73}$$

However, in section 2.4 it is argued that Z is constant inside each cell V_i and therefore the integrals on the right hand side vanish. Since the source term in the relaxation model (4.12) satisfies a transport relation, it is not constant anymore inside each cell and the right hand side of (4.73) has to be taken into account.

In the model (4.12), the source term is approximated by a quadrature to derive the well-balanced property and advected with the fluid flow. Therefore the integrals in (4.73) are evaluated as

$$\begin{aligned} &\int_{x_{i-\frac{1}{2}}}^{x_i} S_i(\mathcal{W}_{i-\frac{1}{2}}(t,x)) Z_x dx = S_{i,i-\frac{1}{2}} \\ &= \begin{cases} (0,0,0)^T & \text{if} \quad u^*_{i-\frac{1}{2}} \leq 0, \\ \left(0, -\overline{\rho}(\rho_{i-1},\rho_i)(\Phi_i - \Phi_{i-1}), -\overline{\rho}(\rho_{i-1},\rho_i) u^*_{i-\frac{1}{2}} (\Phi_i - \Phi_{i-1})\right)^T & \text{if} \quad u^*_{i-\frac{1}{2}} > 0, \end{cases} \end{aligned} \tag{4.74}$$

$$\begin{aligned} &\int_{x_i}^{x_{i+\frac{1}{2}}} S_i(\mathcal{W}_{i+\frac{1}{2}}(t,x)) Z_x dx = S_{i,i+\frac{1}{2}} \\ &= \begin{cases} (0,0,0)^T & \text{if} \quad u^*_{i+\frac{1}{2}} \geq 0, \\ \left(0, -\overline{\rho}(\rho_i,\rho_{i+1})(\Phi_{i+1} - \Phi_i), -\overline{\rho}(\rho_i,\rho_{i+1}) u^*_{i+\frac{1}{2}} (\Phi_{i+1} - \Phi_i)\right)^T & \text{if} \quad u^*_{i+\frac{1}{2}} < 0. \end{cases} \end{aligned} \tag{4.75}$$

With the definitions (4.74) and (4.75), the scheme (4.73) can be rewritten in the following way

$$U_{i,t} + \frac{1}{\Delta_{x_i}}(G^+_{i-\frac{1}{2}} - G^-_{i+\frac{1}{2}}) = 0, \tag{4.76}$$

with the definitions

$$\begin{aligned} G^+_{i-\frac{1}{2}} &= F^+_{i-\frac{1}{2}} - S_{i,i-\frac{1}{2}}, \\ G^-_{i+\frac{1}{2}} &= F^-_{i+\frac{1}{2}} + S_{i,i+\frac{1}{2}}. \end{aligned} \tag{4.77}$$

Observe that the scheme (4.76) still satisfies the well-balanced property (2.82), since in equilibrium $S_{i,i\pm\frac{1}{2}} = 0$

Next, the extension to a formally second order scheme uses, similar to section 3.2, a variant of the surface gradient method from [175]. The discrete equilibrium preserved by the

numerical scheme writes as follows

$$\begin{cases} u_{i-1} = u_i = u_{i+1} = 0, \\ p_i - p_{i-1} = -\overline{\rho}(\rho_{i-1}, \rho_i)(\Phi_i - \Phi_{i-1}), \\ p_{i+1} - p_i = -\overline{\rho}(\rho_i, \rho_{i+1})(\Phi_{i+1} - \Phi_i). \end{cases} \tag{4.78}$$

In order to apply the surface gradient method, first the following transformation is applied

$$\begin{cases} Q^i_{i-1} &= p_{i-1} - \overline{\rho}(\rho_{i-1}, \rho_i)(\Phi_i - \Phi_{i-1}), \\ Q^i_i &= p_i, \\ Q^i_{i+1} &= p_{i+1} + \overline{\rho}(\rho_i, \rho_{i+1})(\Phi_{i+1} - \Phi_i). \end{cases} \tag{4.79}$$

Q^i reflects the equilibrium relation between the pressure and the gravitational acceleration. However, the projection to the variable Q^i is local, i.e. for every cell V_i the projection will give different states Q^i.

The slopes in the cell V_i are then computed following the lines of section 2.3.1 on the variables ρ, u and Q_i to get the slopes $\sigma_{\rho,i}$,$\sigma_{u,i}$ and $\sigma_{Q^i,i}$. Then, the interface values are computed as

$$\begin{cases} \rho^{\mp}_{i\pm\frac{1}{2}} &= \rho_i + \sigma_{\rho,i}(x_{i\pm\frac{1}{2}} - x_i), \\ u^{\mp}_{i\pm\frac{1}{2}} &= u_i + \sigma_{u,i}(x_{i\pm\frac{1}{2}} - x_i), \\ p^{\mp}_{i\pm\frac{1}{2}} &= p_i + \sigma_{Q,i}(x_{i\pm\frac{1}{2}} - x_i), \end{cases} \tag{4.80}$$

and from (4.80), the dependent variables $\rho, \rho u, E$ can be recovered. Similar as in the case of the shallow water equations, see section 3.2, the projection on the equilibrium variables has the advantage that, if the data is in equilibrium, i.e. the relations (4.78) hold, the slopes $\sigma_{u,i}$ and $\sigma_{Q,i}$ are zero. Therefore, the interface values for the pressure and velocity coincide with the cell centered values. However, due to the reconstruction in ρ, the interface values might not coincide with the cell centered values. Nonetheless, evaluating the quadrature for the source term at the cell centered values as in (4.74) and (4.75) gives that the approximate Riemann solver defined by the relaxation system (4.12) satisfies $u^*_{i\pm\frac{1}{2}} = 0$. Therefore, the densities do not contribute to the flux function and it holds that

$$G^+_{i-\frac{1}{2}} = G^-_{i+\frac{1}{2}}, \tag{4.81}$$

and the well-balanced property is achieved. The robustness of the reconstruction is achieved as soon as the projected variables $Q^i_{i\pm1}$ are bounded from below by zero. Then, if for example a minmod limiter is used to compute the slopes, it holds that $p^{\mp}_{i\pm\frac{1}{2}} \in [p_i; p_{i\pm1}]$ and the interface values are in Ω_{phys}.

The extension to more than one space dimension is straightforward following the lines of section 2.5.

4.5 Numerical results

The aim of this section is to show the applicability of the proposed scheme. Similar to numerical experiments in chapter 3, in all tests, an equidistant grid is concerned. Denote by

D the length of the domain and by N_x the number of cells, then there is $\Delta_x = \frac{D}{N_x}$.

In all the applications, the pressure will be given by an ideal gas law

$$p = (\gamma - 1)\rho e,$$

where the adiabatic coefficient is set to $\gamma = \frac{5}{3}$.

In [55], the presented scheme already has been shown to perform well on testcases proposed in the respective literature. Here it is chosen to put emphasis on the influence of the choice of the different quadratures presented in section 4.2 also with respect to the mesh size. A fractional splitting of the source term for comparison is not considered. It is rather obvious that fractional splitting schemes have strong difficulties achieving the well-balanced property. Concern for this that the numerical fluxes computed without accounting for the source term admits non-zero velocities due to the pressure stratification along the atmosphere. Therefore there are non-zero mass fluxes in the first update. However, the physical source term is zero in the mass component and therefore artificial terms in the source term discretization would be needed to counter the errors introduced by the previous upwind procedure.

The different quadratures to be used by the numerical scheme are given as

$$\begin{aligned}
\bar{\rho}(W_L, W_R)_{ISO} &= \frac{\rho_R - \rho_L}{\ln(\rho_R) - \ln(\rho_L)}, \\
\bar{\rho}(W_L, W_R)_{PG} &= \frac{\Gamma - 1}{\Gamma} \frac{\rho_R^\Gamma - \rho_L^\Gamma}{\rho_R^{\Gamma-1} - \rho_L^{\Gamma-1}}, \\
\bar{\rho}(W_L, W_R)_{PI} &= \frac{1}{R} \frac{p_R - p_L}{\ln(p_R) - \ln(p_L)} \frac{\ln(T_R) - \ln(T_L)}{T_R - T_L}, \\
\bar{\rho}(W_L, W_R)_{AV} &= \frac{\rho_R + \rho_L}{2},
\end{aligned}$$

and the schemes equipped with these quadratures are denoted as $SR_{ISO}, SR_{PG}, SR_{PI}$ and SR_{AV} respectively. As has been shown in the previous section, the first quadrature is consistent with an isothermal atmosphere. The second quadrature is consistent with a polytropic steady state for an arbitrary pressure law. The last one is just a simple average and is used for comparison of the different quadratures. However, for $\Gamma = 2$, the arithmetic average coincides with the polytropic quadrature. For the third quadrature a straightforward computation shows that it is consistent with the polytropic states if an ideal gas law is assumed. It has the advantage, that the parameter Γ does not need to be specified. Moreover, the second and third quadrature coincide with the first in the case of an isothermal atmosphere.

The different schemes are also tested with first and second order accuracy, denoted by SR^{FO} and SR^{SO} respectively. For the second order accuracy in space, the reconstructions presented in section 4.4 are used with the respective quadratures to determine the slope in the pressure. For the second order accuracy in time, the modified Heun method from [14] is used, which is also discussed in section 2.3.2.

	SR^{FO}_{ISO}	SR^{SO}_{ISO}	SR^{FO}_{AV}		SR^{SO}_{AV}	
N	u	u	u	EOC	u	EOC
100	0.61E-16	1.46E-16	1.35E-04	-	1.37E-04	-
200	0.82E-16	1.91E-16	3.41E-05	1.99	3.44E-05	1.99
400	0.40E-16	1.40E-16	8.57E-06	1.99	8.61E-06	1.99
800	2.67E-16	1.12E-16	2.15E-06	2.00	2.15E-06	2.00
1600	5.24E-16	1.69E-15	5.38E-07	2.00	5.38E-07	2.00
3200	5.34E-16	1.29E-15	1.35E-07	1.99	1.36E-07	1.98

Table 4.1: L^1 errors for the undisturbed isothermal equilibrium (4.82)-(4.83) at time 0.2 for the schemes based on two different quadratures and different orders of accuracy.

4.5.1 An Isothermal Atmosphere

The first tests are concerned with approximations close to an isothermal atmosphere. Consider the following setup for an isothermal atmosphere

$$\begin{cases} \Phi(x) &= \cos(2\pi), \\ R &= 1, \\ T &= 1. \end{cases} \tag{4.82}$$

Then the hydrostatic equilibrium reads

$$\begin{cases} u(x)_{eq} &= 0, \\ \rho(x)_{eq} &= \exp(-\cos(2\pi)), \\ p(x)_{eq} &= \exp(-\cos(2\pi)). \end{cases} \tag{4.83}$$

The hydrostatic equilibrium (4.82) and (4.83) is now used as an initial condition for the schemes SR_{ISO} and SR_{AV} and its evolution is computed. The L^1 errors for the velocity are given in table 4.1.

As expected, the schemes based on the isothermal quadrature preserve the isothermal equilibrium up to machine precision. In contrast the schemes SR_{AV} both at first and second order introduce spurious oscillations. Even going higher order in the way suggested here does not reduce the numerical errors. However, also independent of the order of the scheme, the error decreases with second order.

The second order accuracy can be computed analytically for the quadrature. Consider that there is

$$\begin{aligned} \frac{p_{i+1} - p_i}{2\Delta_x} &= \frac{\partial}{\partial x} p_{i+\frac{1}{2}} + O(\Delta_x^2), \\ \frac{\Phi_{i+1} - \Phi_i}{2\Delta_x} &= \frac{\partial}{\partial x} \Phi_{i+\frac{1}{2}} + O(\Delta_x^2), \\ \frac{\rho_{i+1} + \rho_i}{2} &= \rho_{i+\frac{1}{2}} + O(\Delta_x^2). \end{aligned}$$

Combining these gives that

	SR^{FO}_{ISO}	SR^{SO}_{ISO}	SR^{FO}_{AV}		SR^{SO}_{AV}	
N	u	u	u	EOC	u	EOC
100	3.67E-14	1.80E-13	1.23E-11	-	1.80E-13	-
200	3.19E-13	3.08E-13	3.86E-12	-	3.08E-13	-
400	7.53E-13	6.07E-13	1.38E-11	-	6.07E-13	-
800	5.06E-13	2.81E-13	4.45E-12	-	2.82E-13	-
1600	3.27E-12	2.84E-12	4.96E-12	-	2.84E-12	-
3200	1.19E-13	6.72E-13	2.43E-13	-	6.72E-13	-

Table 4.2: L^1 errors for the undisturbed isothermal equilibrium (4.84)-(4.85) at time 0.2 for the schemes based on two different quadratures and different orders of accuracy.

$$p_{i+1} - p_i + \bar{\rho}_{AV}(\Phi_{i+1} - \Phi_i) + O(\Delta_x^2) = \frac{\partial}{\partial x} p_{i+\frac{1}{2}} + \rho_{i+\frac{1}{2}} \frac{\partial}{\partial x} \Phi_{i+\frac{1}{2}}.$$

Therefore, by using the closure (4.16), a second order approximation to all hydrostatic equilibria is enforced to hold on the centered wave. Hence, better approximations are achieved when the mesh size Δ_x is decreased.

Moreover, the quality of the approximation does not only depend on Δ_x, but also on some constant that depends on the structure of the hydrostatic equilibrium. To see this, consider the following isothermal equilibrium, where

$$\begin{cases} \Phi(x) & = gx, \\ g & = 9.81, \\ R & = 8.3144598, \\ T & = 300, \\ C_\rho & = 1.225, \end{cases} \tag{4.84}$$

and the solution is given by

$$\begin{cases} u(x)_{eq} & = 0, \\ \rho(x)_{eq} & = C_\rho \exp(-gx), \\ p(x)_{eq} & = C_\rho RT \exp(-gx), \end{cases} \tag{4.85}$$

and $C_\rho = 1.225$. Again both schemes are used to integrate the hydrostatic equilibrium in time and the errors are given in table 4.2. Now the errors for the non-well-balanced scheme have drastically decreased and are on the order of machine accuracy. Therefore, even if the well-balance property does not hold exactly, one might still hope for a reasonable good approximation if the respective equilibrium shows a suitable scaling.

Next, the numerical approximations of a small disturbance on the isothermal equilibrium (4.82)-(4.83) is concerned. For this, the initial pressure distribution is modified as follows

$$p(0,x) - p(x)_{eq} = \begin{cases} 0.05 \times 10^{-3} \sin(\frac{x-x_0}{x_1-x_0}\pi) & \text{if} \quad x_0 < x < x_1, \\ 0 & \text{else}, \end{cases} \tag{4.86}$$

for $x_0 = 0.45$ and $x_1 = 0.55$.

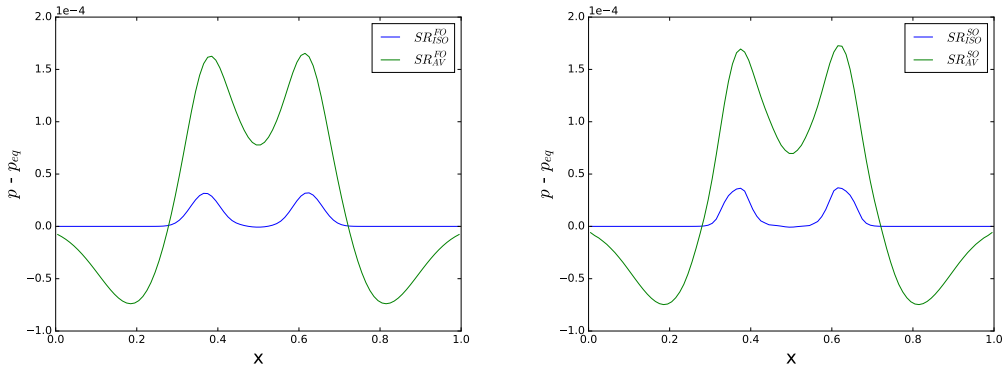


Fig. 4.4: Solutions to the isothermal equilibrium (4.82)-(4.83) with perturbation at time 0.2 computed with 100 cells. Left: Comparison of the first order schemes. Right: Comparison of the second order schemes.

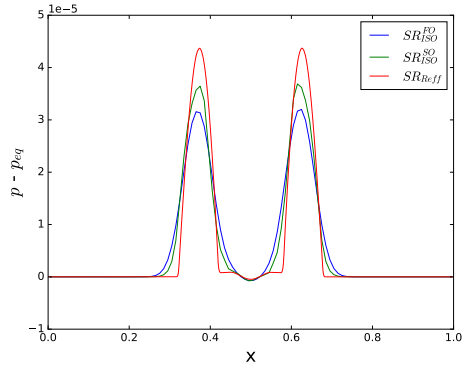


Fig. 4.5: Solutions to the isothermal equilibrium (4.82)-(4.83) with perturbation at time 0.2. The first and second order schemes are computed with 100 cells, the reference solution is computed with 3200 cells.

First, the two schemes SR_{ISO} and SR_{AV} are compared on different orders of accuracy. The results are depicted in figure 4.4. It can be seen that for the scheme SR_{AV} the numerical errors dominate the dynamics and the resulting waves are poorly resolved in the first order case, as well as in the second order case. In contrast to that, the scheme SR_{ISO} shows a good resolution of the waves. Next the results for the scheme SR_{ISO} are compared on different orders with respect to a reference solution computed at high resolution, see figure 4.5. The second order scheme performs better than the first order scheme in capturing the waves resulting from the perturbation, while both schemes seem to be in a good agreement with the reference solution.

	SR_{PG}^{FO}	SR_{PG}^{SO}	SR_{ISO}^{FO}		SR_{ISO}^{SO}	
N	u	u	u	EOC	u	EOC
100	0.73E-16	3.23E-16	1.90E-05	-	1.94E-05	-
200	1.63E-16	2.10E-16	4.80E-06	1.98	4.87E-06	1.99
400	0.95E-16	4.82E-16	1.21E-06	1.99	1.22E-06	2.00
800	1.40E-16	5.80E-16	3.04E-07	1.99	3.05E-07	2.00
1600	1.86E-16	7.31E-16	7.61E-08	2.00	7.63E-08	2.00
3200	1.18E-15	2.66E-15	1.90E-08	2.00	1.91E-08	2.00

Table 4.3: L^1 errors for the undisturbed polytropic equilibrium (4.87)-(4.89) at time 0.2 for the schemes SR_{PG} and SR_{ISO} for different orders of accuracy.

4.5.2 A Polytropic Atmosphere

Consider now a polytropic atmosphere as

$$\begin{cases} \Phi(x) &= \cos(2\pi), \\ p(x) &= K\rho(x)^{\Gamma}, \\ u(x) &= 0. \end{cases} \tag{4.87}$$

Then the solution to (4.87) is given in (4.6) as

$$\begin{cases} \rho(x) &= \left(\frac{\Gamma-1}{\Gamma K}(C-\Phi(x))\right)^{\frac{1}{\Gamma-1}}, \\ p(x) &= K^{\frac{1}{1-\Gamma}}\left(\frac{\Gamma-1}{\Gamma}(C-\Phi(x))\right)^{\frac{\Gamma}{\Gamma-1}}, \end{cases} \tag{4.88}$$

and the coefficients are chosen to be

$$\begin{cases} \Gamma = 3, \\ K = 1.5, \\ C = 2. \end{cases} \tag{4.89}$$

For the numerical tests, periodic boundary conditions are considered and the L^1 errors are given in table 4.3 and table 4.4. The schemes SR_{PG} and SR_{PI} show the expected well-balanced property, while the other schemes introduce numerical errors. However, these errors again decrease with higher resolution with second order of the cell sizes.

Next, as in the case of the isothermal atmosphere, a perturbation on top of the polytropic atmosphere is considered as

$$p(0,x) - p(x)_{eq} = \begin{cases} 0.05 \times 10^{-3} \sin(\frac{x-x_0}{x_1-x_0}\pi) & \text{if} \quad x_0 < x < x_1, \\ 0 & \text{else}, \end{cases} \tag{4.90}$$

for $x_0 = 0.45$ and $x_1 = 0.55$. At first the first order schemes are compared for their performance, see figure 4.6. As expected, the schemes SR_{PI} and SR_{PG} show a better resolution of the waves compared with the other schemes due to their consistency with the polytropic equilibrium. Moreover, the schemes SR_{PI} and SR_{PG} give almost identical results. In figure 4.7, the results for the scheme SR_{PI} are compared at different orders of accuracy with

	SR_{PI}^{FO}	SR_{PI}^{SO}	SR_{AV}^{FO}		SR_{AV}^{SO}	
N	u	u	u	EOC	u	EOC
100	1.03E-16	1.22E-16	9.46E-06	-	9.69E-06	-
200	1.97E-16	3.30E-16	2.40E-06	1.98	2.43E-06	2.00
400	1.77E-16	2.66E-16	6.05E-06	1.99	6.09E-07	2.00
800	2.10E-16	1.50E-15	1.52E-07	2.00	1.52E-07	2.00
1600	2.32E-16	2.34E-15	3.81E-08	2.00	3.81E-08	2.00
3200	4.96E-16	3.27E-15	9.52E-09	2.00	9.54E-09	2.00

Table 4.4: L^1 errors for the undisturbed polytropic equilibrium (4.87)-(4.89) at time 0.2 for the schemes SR_{PI} and SR_{AV} for different orders of accuracy.

respect to a reference solution. Again the second order approach increases the resolution of the waves and both schemes are in good agreement with the reference solution.

4.5.3 General steady state

A last one dimensional test case concerns a general steady state that does not belong to the polytropic family described by (4.6). It is a popular test for determining the behavior of a well-balanced scheme in a general stratified atmosphere and is investigated in [97] and [12]. The gravitational potential is here defined by

$$\Phi(x) = -\sin(2\pi x), \tag{4.91}$$

and the equilibrium is given by the following distributions

$$\begin{cases} \rho(x) &= 3 + 2\sin(2\pi x), \\ u(x) &= 0, \\ p(x) &= 3 + 3\sin(2\pi x) - 0.5\cos(4\pi x). \end{cases} \tag{4.92}$$

Periodic boundary conditions are imposed for the simulations. It is decided to test again all the schemes described above, where for the general polytropic quadrature a value of $\Gamma = 2$ is chosen. The results are given in the table 4.5 and table 4.6. As it turns out, the scheme SR_{AV} is consistent with the proposed equilibrium. Since for $\Gamma = 2$ the quadratures $\bar{\rho}_{PG}$ and $\bar{\rho}_{AV}$ are equivalent, also the scheme SR_{PG} shows a suitable well-balanced property. However, the general equilibrium (4.91)-(4.92) is not a polytropic equilibrium and therefore the scheme SR_{PI} introduces numerical errors.

Again, a small perturbation in pressure on top of the general equilibrium is considered as

$$p(0,x) - p(x)_{eq} = \begin{cases} 0.05 \times 10^{-3} \sin(\frac{x-x_0}{x_1-x_0}\pi) & \text{if} \quad x_0 < x < x_1, \\ 0 & \text{else}, \end{cases} \tag{4.93}$$

for $x_0 = 0.20$ and $x_1 = 0.30$. The first order results are depicted in figure 4.8. As expected, the schemes SR_{AV} and SR_{PG} perform well and no relevant numerical errors due to the stratification of the hydrostatic equilibrium are introduced. In contrast, the schemes SR_{ISO} and SR_{PI} are not able to accurately capture the waves. Moreover, the schemes SR_{AV}

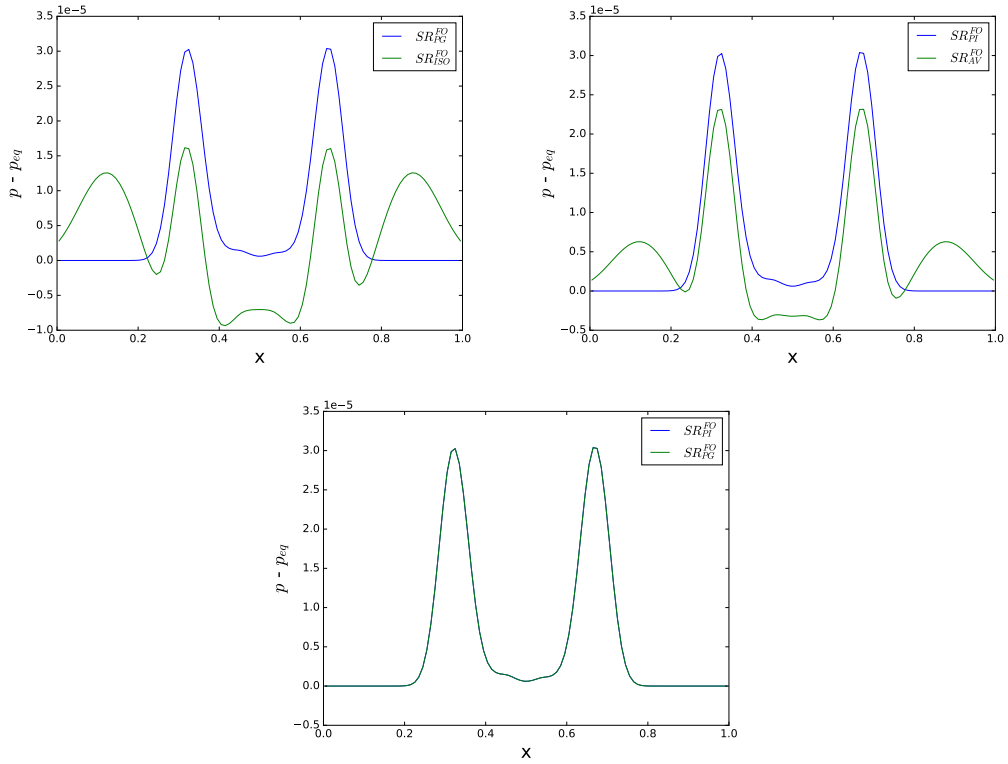


Fig. 4.6: Solutions to the polytropic equilibrium (4.87)-(4.89) with perturbation at time 0.2 computed with 100 cells. Top Left: Comparison of the first order schemes SR_{PG} and SR_{ISO}. Top Right: Comparison of schemes SR_{PI} and SR_{AV}. Bottom: Comparison of schemes SR_{PI} and SR_{PG}.

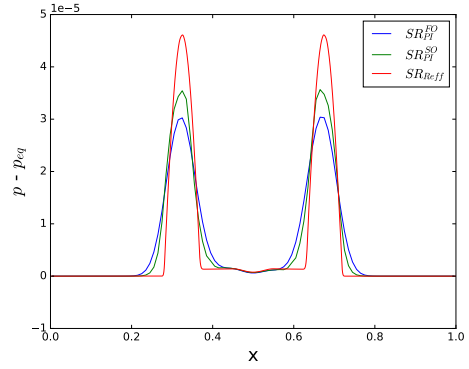


Fig. 4.7: Solutions to the polytropic equilibrium (4.87)-(4.89) with perturbation at time 0.2 comparing the different orders of accuracy for the scheme SR_{PI} with 100 cells. The reference solution is computed with 3200 cells.

	SR_{PG}^{FO}	SR_{PG}^{SO}	SR_{ISO}^{FO}		SR_{ISO}^{SO}	
N	u	u	u	EOC	u	EOC
100	0.76E-16	1.73E-15	4.13E-05	-	4.65E-05	-
200	0.45E-16	8.00E-16	9.97E-06	2.05	1.08E-05	2.11
400	0.60E-16	4.67E-15	2.45E-06	2.02	2.56E-06	2.08
800	2.00E-16	1.40E-15	6.10E-07	2.01	6.21E-07	2.04
1600	1.41E-15	1.31E-14	1.52E-07	2.00	1.53E-07	2.02
3200	1.53E-15	2.81E-14	3.79E-08	2.01	3.81E-08	2.01

Table 4.5: L^1 errors for the undisturbed general equilibrium (4.91)-(4.92) at time 1.0 for the schemes SR_{PG} and SR_{ISO} for different orders of accuracy.

	SR_{AV}^{FO}	SR_{AV}^{SO}	SR_{PI}^{FO}		SR_{PI}^{SO}	
N	u	u	u	EOC	u	EOC
100	0.49E-16	3.53E-16	9.03E-06	-	1.06E-05	-
200	0.83E-16	4.40E-16	2.20E-06	2.04	2.43E-06	2.13
400	1.39E-16	4.90E-15	4.47E-07	2.30	5.76E-07	2.08
800	0.60E-16	6.14E-15	1.36E-07	1.72	1.40E-07	2.04
1600	4.10E-15	1.75E-14	3.40E-08	2.00	3.45E-08	2.02
3200	1.18E-15	1.87E-14	8.50E-09	2.00	8.55E-09	2.01

Table 4.6: L^1 errors for the undisturbed general equilibrium (4.91)-(4.92) at time 1.0 for the schemes SR_{AV} and SR_{PI} for different orders of accuracy.

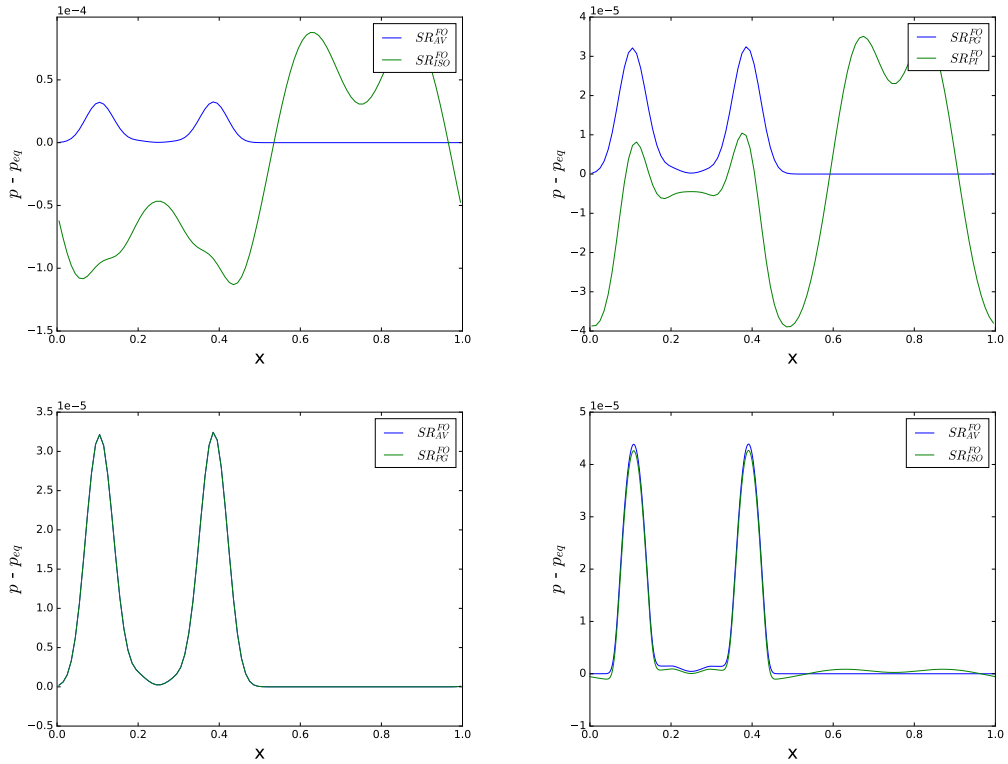


Fig. 4.8: Solutions to the general equilibrium (4.91)-(4.92) with perturbation at time 0.1. Top Left: Comparison of the first order schemes SR_{AV} and SR_{ISO} computed with 100 cells. Top Right: Comparison of schemes SR_{PG} and SR_{PI} computed with 100 cells. Bottom Left : Comparison of schemes SR_{AV} and SR_{PG} computed with 100 cells. Bottom Right: Comparison of schemes SR_{AV} and SR_{ISO} computed with 1000 cells.

and SR_{PG} give almost identical results. As can be seen from the tables 4.5 and 4.6, if a scheme is not exactly well-balanced, the discretization errors are decreasing with second order. Therefore, if the resolution is increased, even a non consistent scheme may be able to capture the resulting dynamics. This is shown on the bottom left of figure 4.8. Here, the consistent scheme SR_{AV} is compared with the non consistent scheme SR_{ISO}, but now on a higher resolution. Both schemes now give comparable results for the approximation of the waves. Finally, also the second order extension is shown to achieve the expected results, see figure 4.9.

4.5.4 An Isothermal Atmosphere in 2 Space Dimensions

The last test in this chapter concerns a two dimensional atmosphere. It is intended to show, that the well-balanced property derived for the one dimensional schemes extends to the two dimensional case naturally, if the approach presented in section 2.5 is used. For this, consider the following isothermal equilibrium

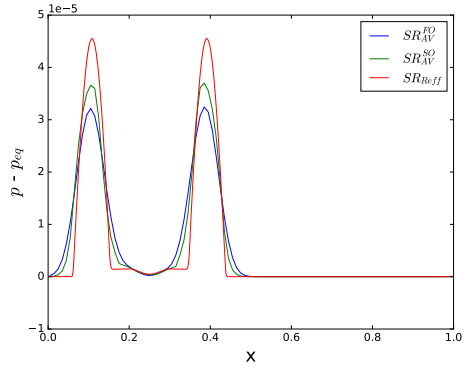


Fig. 4.9: Solutions to the polytropic equilibrium (4.91)-(4.92) with perturbation at time 0.1 comparing the different orders of accuracy for the scheme SR_{AV} with 100 cells. The reference solution is computed with 3200 cells.

$$\begin{cases} \Phi(x,y) & = \cos((x-0.5)\pi)\cos((y-0.5)\pi), \\ R & = 1, \\ T & = 1, \end{cases} \tag{4.94}$$

and the distribution for the dependent variables reads

$$\begin{cases} \mathbf{u}(x,y)_{eq} & = 0, \\ \rho(x,y)_{eq} & = \exp(-\Phi(x,y)), \\ p(x,y)_{eq} & = \exp(-\Phi(x,y)). \end{cases} \tag{4.95}$$

To show the performance in the two dimensional case, a disturbance on the pressure is considered in the following way

$$p(0,x,y) - p(x,y)_{eq} = \begin{cases} 10^{-4}\sin(\frac{r_d - ||\mathbf{x}-c_d||}{2r_d}) & \text{if} \quad ||\mathbf{x}-c_d|| \leq r_d, \\ 0 & \text{otherwise}, \end{cases} \tag{4.96}$$

with $r_d = 0.05$ and $c_d = (0.5, 0.5)^T$. The computational domain is set $D = [0,1] \times [0,1]$ and Neumann boundary conditions are imposed. The scheme SR_{ISO} is applied to compute the evolution of the perturbation, see figure 4.10. The scheme performs as expected and the dynamics of the perturbation is captured accurately without introducing artificial numerical errors.

In this work, the scheme is tested on a cartesian mesh. In [141] it is shown, that the scheme can also be extended to unstructured meshes underlining the flexibility of the presented well-balanced approach.

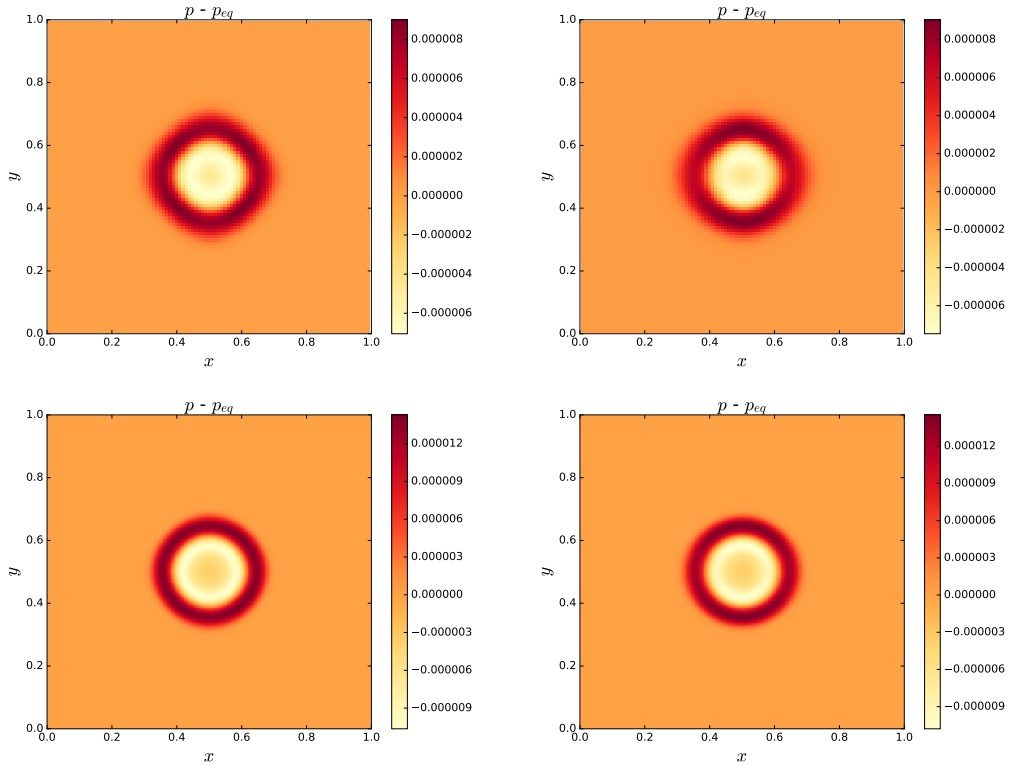


Fig. 4.10: Solutions to the isothermal equilibrium (4.94)-(4.95) with perturbation (4.96) at time 0.1. Top Left: Scheme SR^{FO}_{ISO} computed with 100×100 cells. Top Right: Scheme SR^{SO}_{ISO} computed with 100×100 cells. Bottom Left : Scheme SR^{FO}_{ISO} computed with 400×400 cells. Bottom Right: Scheme SR^{SO}_{ISO} computed with 400×400 cells.

5 A Low Diffusion Suliciu Relaxation Scheme for Low Mach Number Flows

This chapter is concerned with the approximations to the solutions to the compressible Euler equations in the regime of low Mach numbers. It is discussed in section 1.3.2 that, when the Mach number tends to zero, the compressible Euler equations (1.32) reach in the limit the incompressible equations (1.52). In order to find the limit behavior, the Euler equations are non-dimentionalized, which leads to the following system

$$\begin{cases} \rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \\ (\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + I \frac{p}{M^2}) = 0, \\ E_t + \nabla \cdot (\mathbf{u}(E+p)) = 0, \end{cases} \tag{5.1}$$

where the total energy is given by $E = \rho e + M^2 \rho \frac{\mathbf{u}^2}{2}$. In the following, approximations are computed with respect to the non-dimentionalized system (5.1). It admits a set of physical admissible states

$$\Omega_{Phys} = \{(\rho, \rho\mathbf{u}, E) \in \mathbb{R}^5; \rho > 0, e > 0\}, \tag{5.2}$$

and a set of asymptotic preserving states

$$\Omega_{AP} = \{(\rho, (\rho\mathbf{u}), E) \in \mathbb{R}^5; \nabla p_0 = \nabla p_1 = 0, \nabla \rho_0 = 0, \nabla \cdot \mathbf{u}_0 = 0\}. \tag{5.3}$$

As also discussed in section 1.3.2, it is a necessary condition for the limit behavior to the incompressible equations to hold that the dependent variables are in the set (5.3).

There have been different approaches to design low Mach number schemes, which are based on different aspects of the limit to the incompressible equations. The first concept mentioned here are the so called asymptotic preserving schemes. The underlying equations give rise to a limit behavior depending on the Mach number. The numerical scheme in turn should be consistent with that limit behavior. This is depicted in figure 5.1. A widely used approach to deal with this problem is to split the stiff and non-stiff terms in the system (5.1). Then the non-stiff parts are discretized in a time explicit and the stiff parts in a time implicit way to ensure the stability of the scheme. This leads to the IMEX approach and it is used for example in [22],[139],[86],[75] and [51]. The advantage of this splitting approach is that the stiff parts give rise to a linear system, which reduces the computational efforts when solving the coupled system for the variables at the new time step. The stiff part of system (5.1) is strongly related to the pressure term and therefore these splitting approaches often fall in the spirit of Klein [94], where a split of the pressure term into fast and slow fluctuations is proposed. The splitting of the pressure is also critical when deriving the here proposed relaxation scheme.

Another viewpoint on designing low Mach number schemes concerns directly the accuracy of the numerical scheme. For example [162],[168] and [132] point out, that in the low Mach

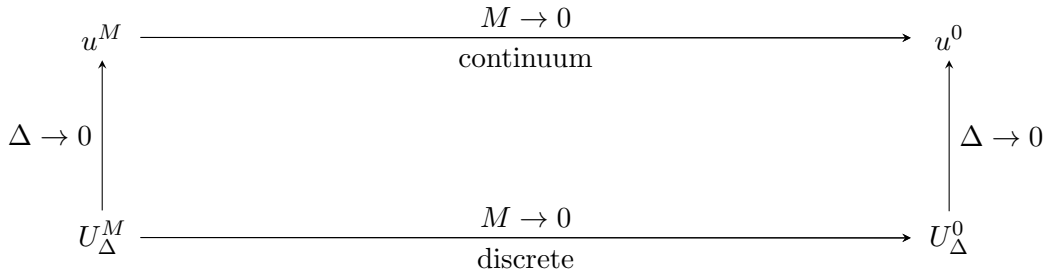


Fig. 5.1: Asymptotic Preserving Diagram: u^M is a solution to (5.1) and u^0 is a solution to (1.52). U_Δ^M and U_Δ^0 are discrete approximations to the respective solutions.

number regime, standard upwind schemes suffer from excessive numerical diffusion. They conclude that with respect to the central flux, the Roe scheme [147] introduces a diffusion which scales as $O(\frac{1}{M})$. In order to get this result, the numerical flux is rewritten in the following form

$$F_{i\pm\frac{1}{2}} = \frac{1}{2}(f(U_{i-1}) + f(U_i)) - \frac{1}{2}D_{i-1,i}(U_i - U_{i-1}). \tag{5.4}$$

The diffusive part $D_{i-1,i}(U_i - U_{i-1})$ then admits entries that scale inversely to the Mach number. To cure this defect, a wide range of preconditioners have been developed to modify the diffusion matrix of the Roe scheme. However, an application of these preconditioning techniques to the relaxation scheme is difficult, since the diffusion matrix $D_{i-1,i}$ can in general only be computed implicitly. For details on the derivation of the diffusion matrix see Appendix B. Instead of computing the scaling of the diffusion from the diffusion matrix, a more direct approach is used here.

Finally, the scaling of the dependent variables in the numerical scheme shall also reflect the scaling in the continuous equations. This has been discussed for example by [52],[74],[136]. As already discussed in section 1.3.2, this gives a constraint on the scaling of the different variables with respect to the Mach number in order to achieve the incompressible limit equations. As have also been put forward by the same authors, standard upwind schemes often violate the scaling of the pressure. In terms of the Roe scheme it has been found that it introduces pressure fluctuations of order $O(M)$ and therefore the incompressible limit might not be achieved. This also leads to excessive diffusion in the low Mach number regime.

To be precise, in this chapter a scheme is developed, for which

$$\Omega = \Omega_{Phys} \cap \Omega_{AP}, \tag{5.5}$$

is an invariant region, while the diffusion is controlled and the asymptotic behavior of the underlying equations is reflected. However, the multidimensional operators used to define the set Ω_{AP} are very hard to satisfy exactly in a numerical approximation. Suitable discrete approximations of these operators are given, when these properties are concerned.

In order to motivate the issue with the standard upwind schemes, the standard Suliciu relaxation model is analyzed in section 5.1 for its low Mach number properties. Then in section 5.2, the modified relaxation scheme is proposed and the low Mach number properties are derived. Finally, in section 5.3, numerical results are given to show the applicability of

the scheme.

5.1 The standard Suliciu relaxation model

In this section, the standard Suliciu relaxation scheme [47],[23],[156],[157] is analyzed for its low Mach properties. The analysis is restricted on the one dimensional fluxes in x direction and is completely analogous for the other directions. The standard Suliciu relaxation approach gives the following system

$$\begin{array}{rcrcl} \rho_t & + & (\rho u)_x & = & 0 \\ (\rho u)_t & + & (\rho u^2 + \frac{\pi}{M^2})_x & = & 0 \\ (\rho v)_t & + & (\rho uv)_x & = & 0, \\ E_t & + & (u(E+\pi))_x & = & 0 \\ (\rho\pi)_t & + & (\rho u\pi + c^2 u)_x & = & 0 \end{array} \tag{5.6}$$

where the relaxation source term has been omitted for brevity.

Lemma 5.1.1. *The system (5.6) is hyperbolic with eigenvalues $\lambda_\pm = u \pm \frac{\rho}{cM}$ and $\lambda_c = u$, where λ_u has multiplicity 3. For a given Riemann problem, it admits an explicit solution.*

Proof. *The proof is straightforward and left to the reader.*

In order to determine the numerical fluxes, the Riemann problem for system (5.6) has to be solved at the cell interfaces. As discussed in section 2.2.4, since the system (5.6) is fully linear degenerate, it gives rise to a solution of the form (2.43), see also figure 2.4. The intermediate states of the solution to the Riemann problem can be computed to be

$$\pi_C = \pi_{L^*} = \pi_{R^*} = \frac{\pi_L + \pi_R}{2} - cM\frac{u_R - u_L}{2} \qquad u_C = u_{L^*} = u_{R^*} = \frac{u_L + u_R}{2} - \frac{\pi_R - \pi_L}{2cM}$$
$$\tau_{L^*} = \tau_L + \frac{\pi_C - \pi_L}{c^2} \quad \tau_{R^*} = \tau_R + \frac{\pi_R - \pi_C}{c^2} \qquad e_{L^*} = e_L - \frac{\pi_L^2 - \pi_C^2}{2c^2} \quad e_{R^*} = e_R - \frac{\pi_R^2 - \pi_C^2}{2c^2}.$$

In order to show the scaling of the intermediate states, make use of the scalings given in (1.53),(1.54) and (1.55) and impose them on the initial conditions to get

$$\begin{array}{rcl} \pi_L = p_0 & \text{and} & \pi_R = p_0 + O(M^2), \\ \tau_L = \tau_0 & \text{and} & \tau_R = \tau_0 + O(M), \\ u_L = u_0 & \text{and} & u_R = u_0 + O(1), \\ v_L = v_0 & \text{and} & v_R = v_0 + O(1), \\ e_L = e_0 & \text{and} & e_R = e_0 + O(M). \end{array} \tag{5.7}$$

With (5.7) one can determine the scaling of the intermediate states to be

$$\begin{aligned} \pi_C &= p_0 + O(M) \quad \text{and} \quad u_C = u_0 + O(1), \\ v_C^L &= v_0 \quad \text{and} \quad v_C^R = v_0 + O(1), \\ \tau_C^L &= \tau_0 + O(M) \quad \text{and} \quad \tau_C^R = \tau_0 + O(M), \\ e_C^L &= e_0 + O(M) \quad \text{and} \quad e_C^R = e_0 + O(M), \end{aligned} \tag{5.8}$$

and therefore there is

$$W_{L^*}, W_{R^*} \notin \Omega_{AP}.$$

The most important factor here is the pressure π_C, which admits variations of the order $O(M)$. Moreover, the failure of preserving the asymptotic behavior of the pressure leads to excessive diffusion. As it is shown in the Appendix B, the diffusion matrix can not be computed explicitly. Therefore it is decided to compute the diffusion with respect to the central flux directly. Using the scalings (5.8) and plug them into the flux function to get

$$\frac{1}{2}(f_L + f_R) - F_{L,R} = \begin{pmatrix} \rho_0 u_0 + O(1) \\ \rho_0 u_0^2 + \frac{p_0}{M^2} + O(1) \\ \rho_0 u_0 v_0 + O(1) \\ u_0(E_0 + p_0) + O(1) \end{pmatrix} - \begin{pmatrix} \rho_0 u_0 + O(1) \\ \rho_0 u_0^2 + \frac{p_0}{M^2} + O(\frac{1}{M}) \\ \rho_0 u_0 v_0 + O(1) \\ u_0(E_0 + p_0) + O(1) \end{pmatrix} = \begin{pmatrix} O(1) \\ O(\frac{1}{M}) \\ O(1) \\ O(1) \end{pmatrix}. \tag{5.9}$$

Therefore excessive diffusion is expected in the momentum orthogonal to the interface.

Remark 5.1.1. *When one is considering only one dimensional flows, due to the divergence constraint, the scaling of the velocities in (5.7) can be rewritten as $u_L = u_0$ and $u_R = u_0 + O(M)$. In this case, $\pi_C = p_0 + O(M^2)$ and therefore the scheme reflects the asymptotic behavior. The problem of low Mach number approximations originates therefore only from multidimensional considerations.*

It is shown that the standard version of the Suliciu relaxation is not suited for approximating low Mach number flows. However, it turns out that the main problem is in the scaling of the intermediate relaxation pressure π_C. Since this relaxation technique relies on controlling a relaxation pressure, it seems natural to search for a different control of the relaxation pressure in order to achieve the asymptotic behavior of the intermediate states. In fact, this is the basis to derive the low diffusion relaxation scheme in the next section.

5.2 An All Mach Number Relaxation Model

To cure the deficiencies of the model presented in section 5.1, a different relaxation model is proposed, capable to accurately capture low Mach number flows. In the the spirit of [95] the pressure is split into a pressure for the slow dynamics and one for the fast acoustics. The pressure term in the momentum equations can be rewritten as follows

$$\frac{p}{M^2} = p + \frac{1 - M^2}{M^2} p. \tag{5.10}$$

Now the idea is to introduce two different relaxation pressures for the right hand side of (5.10) as

$$p + \frac{1 - M^2}{M^2} p = \pi + \frac{1 - M^2}{M^2} \psi. \tag{5.11}$$

In the spirit of the standard Suliciu relaxation, the evolution equations for the relaxation pressures are derived as

$$\pi_t + u\pi_x + \frac{c^2}{\rho} u_x = \frac{1}{\epsilon}(p - \pi), \tag{5.12}$$

$$\psi_t + u\psi_x + \frac{c^2}{\rho} u_x = \frac{1}{\epsilon}(p - \psi). \tag{5.13}$$

In order to complete the fast acoustics, an additional relaxation process has to be introduced for the velocity u, where this is then coupled with the fast acoustic pressure ψ. To this end, the following set of equations is suggested

$$\bar{u}_t + u\bar{u}_x + \frac{\psi_x}{\rho M^4} = \frac{1}{\epsilon}(u - \bar{u}), \tag{5.14}$$

$$\psi_t + u\psi_x + \frac{c^2}{\rho} \bar{u}_x = \frac{1}{\epsilon}(p - \psi), \tag{5.15}$$

which gives rise to the following system

$$\begin{array}{rcccl} \rho_t & + & (\rho u)_x & = & 0 \\ (\rho u)_t & + & (\rho u^2 + \pi + \frac{1-M^2}{M^2}\psi)_x & = & 0 \\ (\rho v)_t & + & (\rho v u)_x & = & 0 \\ E_t & + & (u(E + M^2\pi + (1 - M^2)\psi)_x & = & 0 \\ (\rho\pi)_t & + & (\rho u\pi + c^2 u)_x & = & \frac{\rho}{\epsilon}(p - \pi) \\ (\rho\psi)_t & + & (\rho u\psi + c^2\bar{u})_x & = & \frac{\rho}{\epsilon}(p - \psi) \\ (\rho\bar{u})_t & + & (\rho u\bar{u} + \frac{1}{M^4}\psi)_x & = & \frac{\rho}{\epsilon}(u - \bar{u}) \end{array} . \tag{5.16}$$

To make the scheme adaptive to local flow properties, it is decided to make the splitting of the pressures described in (5.11) dependent on the local Mach number. So, from now on, it is distinguished between a reference Mach number M_{ref}, which is used to rescale the dependent and independent variables, and a local Mach number M_{loc}, which is derived from the local flow properties. Therefore, the following splitting of the pressure is suggested

$$\frac{p}{M_{ref}^2} = \frac{M_{loc}^2}{M_{ref}^2} p + \frac{1 - M_{loc}^2}{M_{ref}^2} p = \frac{M_{loc}^2}{M_{ref}^2} \pi + \frac{1 - M_{loc}^2}{M_{ref}^2} \psi, \tag{5.17}$$

and the evolution for the relaxed velocity $\bar{u}$ is modified to give

$$\bar{u}_t + u\bar{u}_x + \frac{\psi_x}{\rho M_{loc}^2 M_{ref}^2} = \frac{1}{\epsilon}(u - \bar{u}). \tag{5.18}$$

In total, the following relaxation system is now concerned

$$\begin{array}{lcccl}
\rho_t & + & (\rho u)_x & = & 0 \\
(\rho u)_t & + & (\rho u^2 + \frac{M_{loc}^2}{M_{ref}^2}\pi + \frac{1-M_{loc}^2}{M_{ref}^2}\psi)_x & = & 0 \\
(\rho v)_t & + & (\rho v u)_x & = & 0 \\
E_t & + & (u(E + M_{loc}^2\pi + (1-M_{loc}^2)\psi)_x & = & 0 \\
(\rho\pi)_t & + & (\rho u\pi + c^2 u)_x & = & \frac{\rho}{\epsilon}(p-\pi) \\
(\rho\psi)_t & + & (\rho u\psi + c^2\bar{u})_x & = & \frac{\rho}{\epsilon}(p-\psi) \\
(\rho\bar{u})_t & + & (\rho u\bar{u} + \frac{1}{M_{loc}^2 M_{ref}^2}\psi)_x & = & \frac{\rho}{\epsilon}(u-\bar{u}),
\end{array} \tag{5.19}$$

where $E = \rho e + M_{ref}\frac{u^2+v^2}{2}$. Next the robustness, stability and consistency properties of the new relaxation methods are discussed.

5.2.1 Robustness, Stability and Consistency of the Low Mach Number Relaxation Approach

Lemma 5.2.1. *If the relaxation constant c satisfies the subcharacteristic condition $c^2 > \rho^2 p'$, then, to first order of the relaxation parameter ϵ, the relaxation system (5.19) is a stable diffusive approximation of system (5.1).*

Proof. *A stability analysis of the relaxation system by using the Chapman-Enskog analysis as described in section 1.2.2 is performed. From the last 3 equations of system (5.19) there is in terms of $\pi, \psi, \bar{u}$*

$$\begin{aligned}
\pi &= p - \epsilon(\pi_t + u\pi_x + \frac{c^2}{\rho}u_x), \\
\psi &= p - \epsilon(\psi_t + u\psi_x + \frac{c^2}{\rho}u_x), \\
\bar{u} &= u - \epsilon(\bar{u}_t + u\bar{u}_x + \frac{1}{\rho M_{loc}^2 M_{ref}^2}\psi_x).
\end{aligned}$$

The relaxation variables are expanded in terms of the relaxation parameter ϵ in the following way

$$\pi = \pi_0 + \epsilon\pi_1 + h.o.t., \;\; \psi = \psi_0 + \epsilon\psi_1 + h.o.t., \;\; \bar{u} = \bar{u}_0 + \epsilon\bar{u}_1 + h.o.t.,$$

where the equilibrium condition reads

$$\pi_0 = p, \;\; \psi_0 = p, \;\; \bar{u}_0 = u.$$

To first order of the relaxation parameter ϵ there is

$$\pi = p - \epsilon(p_t + up_x + \frac{c^2}{\rho}u_x),$$
$$\psi = p - \epsilon(p_t + up_x + \frac{c^2}{\rho}u_x),$$
$$\bar{u} = u - \epsilon(u_t + uu_x + \frac{c^2}{\rho M_{loc}^2 M_{ref}^2}p_x)).$$

Given that $\frac{\partial p}{\partial \rho}\,|_{s=const} = p'$*, then, from conservation of mass and momentum, there is*

$$p_t + up_x = -\rho p' u_x,$$
$$u_t + uu_x = -\frac{p_x}{\rho M_{ref}^2},$$

and the following first order approximations to the relaxation variables hold

$$\pi = p - \epsilon((\frac{c^2}{\rho} - \rho p')u_x),$$
$$\psi = p - \epsilon((\frac{c^2}{\rho} - \rho p')u_x),$$
$$\bar{u} = u - \epsilon((\frac{1}{M_{loc}^2} - 1)\frac{p_x}{\rho M_{ref}^2}).$$

Now, use these expressions in the momentum and energy equation to get

$$(\rho u)_t + (\rho u^2 + \frac{p}{M_{ref}^2})_x = \epsilon\left(\frac{1}{\rho M_{ref}^2}(c^2 - \rho^2 p')u_x\right)_x,$$
$$E_t + (u(E+p))_x = \epsilon\left(\frac{1}{\rho}(c^2 - \rho^2 p')(\frac{u^2}{2})_x\right)_x,$$

and therefore for stability the following subcharacteristic condition has to hold

$$c^2 > \rho^2 p'.$$

Remark 5.2.1. *In the proof of the stability of the relaxation system, the evolution equation for* $\bar{u}$ *drops out, since* $\bar{u}$ *is not directly present in the fluxes for the conserved variables. This might seem strange, since from the Chapman Enskog expansion, the evolution for* $\bar{u}$ *is not restricted. However, consider the system*

$$
\begin{array}{lcccl}
\rho_t & + & (\rho u)_x & = & 0 \\
(\rho u)_t & + & (\rho u^2 + \frac{M_{loc}^2}{M_{ref}^2}\pi + \frac{1-M_{loc}^2}{M_{ref}^2}\frac{\psi_1+\psi_2}{2})_x & = & 0 \\
(\rho v)_t & + & (\rho v u)_x & = & 0 \\
E_t & + & (u(E + M_{loc}^2\pi + (1-M_{loc}^2)\frac{\psi_1+\psi_2}{2})_x & = & 0 \\
(\rho\pi)_t & + & (\rho u\pi + c^2 u)_x & = & \frac{\rho}{\epsilon}(p-\pi) \\
(\rho\psi_1)_t & + & (\rho u\psi_1 + \frac{c^2}{M_{loc}M_{ref}}\psi_1)_x & = & \frac{\rho}{\epsilon}(p + cM_{loc}M_{ref}u - \psi_1) \\
(\rho\psi_2)_t & + & (\rho u\psi_2 - \frac{c^2}{M_{loc}M_{ref}}\psi_2)_x & = & \frac{\rho}{\epsilon}(p - cM_{loc}M_{ref}u - \psi_2)
\end{array} \quad . \tag{5.20}
$$

A straightforward computation shows that the approximate Riemann solver defined by (5.20) is equivalent to the approximate Riemann solver from (5.19) and the alternative system (5.20) admits the same subcharacteristic condition from the Chapman Enskog analysis, see Appendix A. The choice of the system 5.19 to use for the numerical fluxes is, that it is a little bit easier to handle in the code and also easier to extend to the case, when there is a gravitational source term present, see chapter 6.

To set up the numerical scheme, the solution to the Riemann problem at the cell interface is needed. This is concerned by lemma 5.2.2.

Lemma 5.2.2. *The relaxation system (5.19) is hyperbolic and fully linear degenerate with eigenvalues* $\lambda_{\pm_s} = u \pm \frac{cM_{loc}}{\rho M_{ref}}$, $\lambda_{\pm_f} = u \pm \frac{c}{\rho M_{loc}M_{ref}}$ *and* $\lambda_c = u$, *where* λ_c *has multiplicity* 3. *Moreover, the solution to the Riemann problem is composed of 6 constant states separated by 5 contact discontinuities as in the following way*

$$
W_{\mathcal{R}}(\frac{x}{t}; W_L, W_R) = \begin{cases} W_L \ \textit{if} \ \frac{x}{t} < \lambda_{-_f}, \\ W_{L^*} \ \textit{if} \ \lambda_{-_f} < \frac{x}{t} < \lambda_{-_s}, \\ W_{C^L} \ \textit{if} \ \lambda_{-_s} < \frac{x}{t} < \lambda_c, \\ W_{C^R} \ \textit{if} \ \lambda_c < \frac{x}{t} < \lambda_{+_s}, \\ W_{L^*} \ \textit{if} \ \lambda_{+_s} < \frac{x}{t} < \lambda_{+_f}, \\ W_R \ \textit{if} \ \frac{x}{t} > \lambda_{+_f}. \end{cases} \tag{5.21}
$$

The solution to the states W *can be derived explicitly and is given as follows*

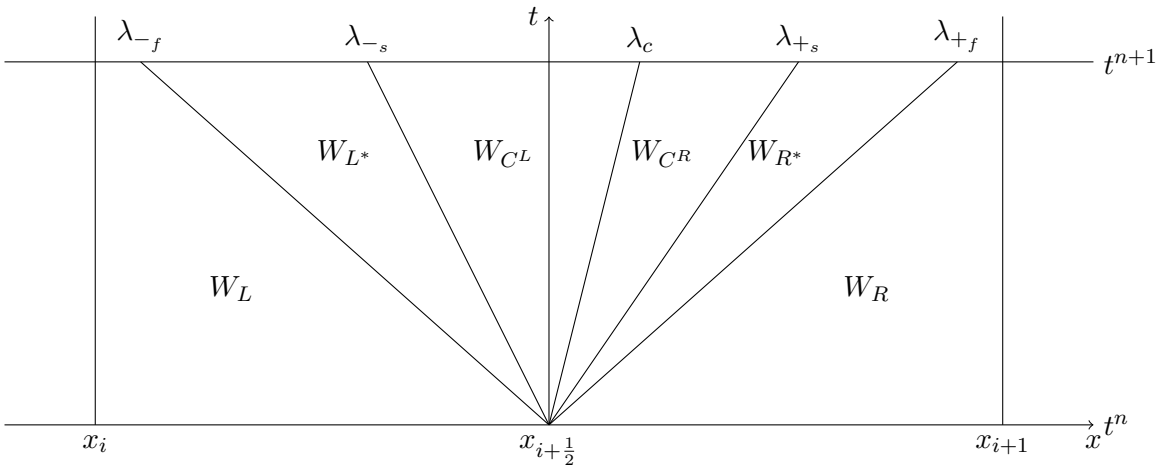


Fig. 5.2: Solution structure to the Riemann problem for the system (5.19)

$$\psi_C = \frac{\psi_L + \psi_R}{2} + cM_{loc}M_{ref}\frac{\bar{u}_L - \bar{u}_R}{2} \qquad \bar{u}_C = \frac{\bar{u}_L + \bar{u}_R}{2} + \frac{\psi_L - \psi_R}{2cM_{loc}M_{ref}},$$

$$\pi_{L^*} = \pi_L + \frac{M_{loc}^2(\psi_C - \psi_L)}{1 + M_{loc}^2} \qquad \pi_{R^*} = \pi_R + \frac{M_{loc}^2(\psi_C - \psi_R)}{1 + M_{loc}^2},$$

$$u_{L^*} = u_L - \frac{M_{loc}(\psi_C - \psi_L)}{cM_{ref}^2(1 + M_{loc}^2)} \qquad u_{R^*} = u_R + \frac{M_{loc}(\psi_C - \psi_R)}{cM_{ref}^2(1 + M_{loc}^2)},$$

$$\pi_C = \frac{\pi_{L^*} + \pi_{R^*}}{2} + \frac{cM_{ref}}{M_{loc}}\frac{(u_{L^*} - u_{R^*})}{2} \qquad u_C = \frac{u_{L^*} + u_{R^*}}{2} + \frac{M_{loc}}{cM_{ref}}\frac{\pi_{L^*} - \pi_{R^*}}{2},$$

$$\tau_{L^*} = \tau_L + \frac{\pi_L - \pi_{L^*}}{c^2} \qquad \tau_{R^*} = \tau_R + \frac{\pi_R - \pi_{R^*}}{c^2},$$

$$\tau_{C^L} = \tau_L + \frac{\pi_L - \pi_C}{c^2} \qquad \tau_{C^R} = \tau_R + \frac{\pi_R - \pi_C}{c^2},$$

$$e_{L^*} = e_L - \frac{M_{loc}^2}{2c^2}(\pi_L^2 - \pi_{L^*}^2 + \frac{1 - M_{loc}^2}{1 + M_{loc}^2}(\psi_L^2 - \psi_C^2)),$$

$$e_{R^*} = e_R - \frac{M_{loc}^2}{2c^2}(\pi_R^2 - \pi_{R^*}^2 + \frac{1 - M_{loc}^2}{1 + M_{loc}^2}(\psi_R^2 - \psi_C^2)),$$

$$e_{C^L} = e_{L^*} - \pi_{L^*}\frac{M_{loc}^2\pi_{L^*} + 2(1 - M_{loc}^2)\psi_C}{2c^2} + \pi_C\frac{M_{loc}^2\pi_C + 2(1 - M_{loc}^2)\psi_C}{2c^2},$$

$$e_{C^R} = e_{R^*} - \pi_{R^*}\frac{M_{loc}^2\pi_{R^*} + 2(1 - M_{loc}^2)\psi_C}{2c^2} + \pi_C\frac{M_{loc}^2\pi_C + 2(1 - M_{loc}^2)\psi_C}{2c^2}.$$

The resulting wave structure is also depicted in figure 5.2.

Proof. *The computation of the eigenvalues is straightforward and omitted for brevity. Now, investigate the Riemann Invariants to the respective fields. Compute the eigenvectors in the primitive variables* $V = \{\rho, u, v, e, \pi, \psi, \bar{u}\}$ *to find for each eigenvalue*

- λ_c: *The eigenvectors read* $\begin{pmatrix} 1\\0\\0\\0\\0\\0\\0 \end{pmatrix}, \begin{pmatrix} 0\\0\\1\\0\\0\\0\\0 \end{pmatrix}, \begin{pmatrix} 0\\0\\0\\0\\1\\0\\0 \end{pmatrix}$, *and the Riemann Invariants are* $\Phi_c = \{u, \pi, \psi, \bar{u}\}$.

- $\lambda_{\pm_s} = u \pm \frac{cM_{loc}}{\rho M_{ref}}$: *The eigenvectors read* $\begin{pmatrix} \rho^2 \\ \pm\frac{cM_{loc}}{M_{ref}} \\ 0 \\ c^2 \\ M_{loc}^2\pi + (1-M_{loc}^2)\psi \\ 0 \\ 0 \end{pmatrix}$, *and the Riemann Invariants are* $\Phi_{\pm_s} = \{u \pm \frac{cM_{loc}}{\rho M_{ref}}, \pi \mp \frac{cM_{ref}}{M_{loc}}u, e - \pi\frac{M_{loc}^2\pi + 2(1-M_{loc}^2)\psi}{2c^2}, v, \psi, \bar{u}\}$.

- $\lambda_{\pm_f}$ *The eigenvector read* $\begin{pmatrix} \rho^2 \\ \pm\frac{c}{M_{ref}M_{loc}} \\ 0 \\ c^2 \\ M_{loc}^2\pi + (1-M_{loc}^2)\psi \\ \frac{\pm c(1+M_{loc}^2)}{M_{loc}^3 M_{ref}} \\ \frac{c^2(1+M_{loc}^2)}{M_{loc}^2} \end{pmatrix}$. *Therefore the Riemann Invariants are* $\Phi_{\pm_f} = \{\pi + \frac{c^2}{\rho}, \psi - \frac{1+M_{loc}^2}{M_{loc}^2}\pi, \psi \mp c\frac{M_{ref}(1+M_{loc})^2}{M_{loc}}u, \psi \mp cM_{loc}M_{ref}\bar{u}, e - \frac{M_{loc^2}}{2c^2}(\pi^2 + \frac{1-M_{loc}^2}{1+M_{loc}^2}\psi^2), v\}$.

The solutions to the intermediate states can be achieved by using the Riemann Invariants and solving the resulting linear system of equations. Moreover, the linear degeneracy of the eigenvalues follows from the fact that all eigenvalues are Riemann invariants for their respective fields.

Now, the robustness of the new relaxation approach is concerned.

Theorem 5.2.1. *[Robustness] Let W_L and W_R be given in Ω_{phys}.*
For $M_{loc} < 1$, $M_{ref} \notin \left[\frac{M_{loc}^2}{2+M_{loc}^2+\sqrt{1-M_{loc}^4}}, \frac{M_{loc}^2}{2+M_{loc}^2-\sqrt{1-M_{loc}^4}}\right]$ and choosing $c > 0$ large enough, the states W as defined in Theorem 5.2.2 also belong to the set Ω_{phys}.

Proof. *The positivity of the density follows from the ordering of the eigenvalues. Consider first the ordering of the centered eigenvalues*

$$u_C - c\frac{M_{loc}}{M_{ref}}\tau_{C^L} < u_C < u_C + c\frac{M_{loc}}{M_{ref}}\tau_{C^R}.$$

By subtraction of u_C there is

$$-c\frac{M_{loc}}{M_{ref}}\tau_{C^L} < 0 < c\frac{M_{loc}}{M_{ref}}\tau_{C^R},$$

and so for $c > 0$, there is $\tau_{C^L}, \tau_{C^R} > 0$. Second, consider the ordering of the left eigenvalues as given by

$$u_{L^*} - c\frac{1}{M_{loc}M_{ref}}\tau_{L^*} < u_{L^*} - c\frac{M_{loc}}{M_{ref}}\tau_{L^*}. \tag{5.22}$$

Again, by subtracting u_{L^*} and some further manipulation there is

$$\tau_{L^*}(1 - M_{loc}^2) > 0. \tag{5.23}$$

For $M_{loc} < 1$ there is $\tau_{L^*} > 0$. By symmetry, the same holds true for τ_{R^*}. For proving the positivity of the internal energies, first take a look at the respective formulas for the left states

$$\begin{aligned} e_{L^*} &= e_L - \frac{M_{loc}^2}{2c^2}(\pi_L^2 - \pi_{L^*}^2 + \frac{1 - M_{loc}^2}{1 + M_{loc}^2}(\psi_L^2 - \psi_C^2)), \\ e_{C^L} &= e_{L^*} - \pi_{L^*}\frac{M_{loc}^2\pi_{L^*} + 2(1 - M_{loc}^2)\psi_C}{2c^2} + \pi_C\frac{M_{loc}^2\pi_C + 2(1 - M_{loc}^2)\psi_C}{2c^2}. \end{aligned} \tag{5.24}$$

Since $e_L > 0$, the positivity of e_{L^*} and e_{C^L} is ensured by choosing c large enough. In order to show that, it is convenient to rewrite π_{L^*}, π_C and ψ_C in terms of c and rewrite for better readability. From Lemma 5.2.2 there is

$$\begin{aligned} \pi_{L^*} &= c\frac{M_{loc}^3 M_{ref}}{1 + M_{loc}^2}\frac{(u_L - u_R)}{2} + l.o.t. = c\theta_1\frac{(u_L - u_R)}{2} + l.o.t., \\ \pi_C &= c\frac{M_{ref}(1 + M_{loc}^2 + M_{loc}^4) - M_{loc}^2}{M_{loc}(1 + M_{loc}^2)}\frac{(u_L - u_R)}{2} + l.o.t. = c\theta_2\frac{(u_L - u_R)}{2} + l.o.t., \\ \psi_C &= cM_{ref}M_{loc}\frac{(u_L - u_R)}{2} + l.o.t. = c\theta_3\frac{(u_L - u_R)}{2} + l.o.t.. \end{aligned} \tag{5.25}$$

First, use (5.25) in (5.24) for e_{L^*} to get

$$e_{L^*} = e_L + \underbrace{M_{loc}^2(\theta_1^2 + \frac{1 - M_{loc}^2}{1 + M_{loc}^2}\theta_2^2)}_{=\theta_4}\frac{(u_L - u_R)^2}{8} + O(\frac{1}{c}).$$

For $M_{loc} < 1$ there is $\theta_4 > 0$ and choosing c large enough gives $e_{L^*} > 0$. Further use (5.25) in (5.24) for e_{C^L} to get

$$e_{L^C} = e_L + \underbrace{\left(M_{loc}^2 \frac{1 - M_{loc}^2}{1 + M_{loc}^2} \theta_3^2 + M_{loc}^2 \theta_2^2 - 2(1 - M_{loc}^2)\theta_3(\theta_1 - \theta_2) \right)}_{:= f(M_{loc}, M_{ref})} \frac{(u_L - u_R)^2}{8} + O(\frac{1}{c}).$$

The sign of $f(M_{loc}, M_{ref})$ is not obvious. First, substitute the definitions for θ_i and simplify to get

$$f(M_{loc}, M_{ref}) = \frac{(3 + 4M_{loc}^2 + 2M_{loc}^4)M_{ref}^2 - M_{ref}(4M_{loc}^2 + 2M_{loc}^4) + M_{loc}^4}{(1 + M_{loc}^2)^2}.$$

Since the denominator is always positive, it is sufficient to consider the numerator as given in the following function

$$\bar{f}(M_{loc}, M_{ref}) = (3 + 4M_{loc}^2 + 2M_{loc}^4)M_{ref}^2 - M_{ref}(4M_{loc}^2 + 2M_{loc}^4) + M_{loc}^4.$$

The function $\bar{f}$ is a convex second order polynomial in M_{ref}. Therefore, one can compute the zeros with respect to M_{ref} to get

$$M_{ref,1,2} = \frac{M_{loc}^2}{2 + M_{loc}^2 \pm \sqrt{1 - M_{loc}^4}}, \tag{5.26}$$

which shows the desired result.

Remark 5.2.2. *The restriction on the choice of the parameter refers to states, where $M_{loc} > M_{ref}$. To see this, take the larger root from (5.26) and compute*

$$\frac{M^2}{2 + M^2 - \sqrt{1 - M^4}} - M = \frac{M(M - 2 - M^2 + \sqrt{1 - M^4})}{2(2 + M^2 - \sqrt{1 - M^4})} < 0,$$

for $0 \leq M \leq 1$.

Remark 5.2.3. *Theorem 5.2.1 together with Theorem 2.2.1 ensures that, if a time explicit time discretization is chosen, the numerical scheme based on the relaxation system 5.19 is robust, i.e. the numerical approximations are all in Ω_{phys}. However, for efficiency reasons, an implicit time discretizations has to be chosen. A robustness result for this case is out of the scope of this work.*

Now the limit behavior $M_{loc} \to 1$ of the system (5.19) is considered. Even though when M_{loc} tends to one for $M_{loc} > M_{ref}$, there is a region where the positivity of the internal energy can not be ensured, the consistency of the new proposed relaxation procedure with respect to the standard Suliciu relaxation system is of interest. The standard Suliciu relaxation system is known to perform reasonable well around shocks. This property should translate to the new relaxation scheme.

Theorem 5.2.2. *(Consistency) For $M_{loc} \to 1$, the numerical scheme based on the system (5.19) goes to the numerical scheme based on the system (5.6).*

Proof. *When $M_{loc} \to 1$, the last two equations in (5.19) do not have any influence one the rest of the system. The upper part of the system (5.19) in turn is identical to the standard relaxation system (5.6).*

5.2.2 Low Mach Number Properties of the New Relaxation Scheme

In this section, the low Mach properties of the scheme derived from the relaxation system (5.19) shall be discussed. Three properties of the low Mach behavior are investigated. The scaling of the intermediate states, the diffusion and the asymptotic preserving property.

First, the scaling of the the intermediate states in the solution of the Riemann problem is discussed. In the following, it is assumed that

$$\forall_{k\in\mathbb{N}} \;\; O(M_{ref}^k) = O(M_{loc}^k) = O(M^k). \tag{5.27}$$

With the formulas for the intermediate states given in Lemma 5.2.2, it is straightforward to give the following results for the scaling of the intermediate states

$$\begin{aligned} \pi_{L^*} &= p_0 + O(M^4) & \qquad \pi_{R^*} &= p_0 + O(M^4), \\ \pi_C &= p_0 + O(1) & \qquad \psi_C &= p_0 + O(M^2). \end{aligned}$$

This seems at a first glance like a step back, because the pressure π_C now scales with $O(1)$, whereas in the standard relaxation π_C scales with $O(M)$. However, consider the now modified momentum equation

$$(\rho u)_t + (\rho u^2 + O(1)\pi + O(M^{-2})\psi)_x = 0.$$

Since π now gets multiplied by a factor of $O(1)$, the scaling of π_C is consistent with the scaling given in (5.3) and therefore, if W_L and W_R are given in Ω_{AP}, then

$$W \in \Omega_{AP}. \tag{5.28}$$

Next, the diffusion of the upwind scheme is discussed. Similar to section 5.1, the diffusion vector is computed as the difference from the central flux and the interface flux.

$$D = \frac{f(u_L) + f(u_R)}{2} - f^*. \tag{5.29}$$

To compute the scaling of the interface flux, it is also necessary to specify the scaling of the other dependent variables. It is straightforward to show that

$$\begin{aligned} u_{L^*} &= u_0 + O(M^2), & \qquad u_{R^*} &= u_0 + O(M^2), \\ \tau_{L^*} &= \tau_0 + O(M), & \qquad \tau_{R^*} &= \tau_0 + O(M), \\ \tau_{C^L} &= \tau_0 + O(M^2), & \qquad \tau_{C^R} &= \tau_0 + O(M^2), \\ \pi_C &= p_0 + O(1), & \qquad u_C &= u_0 + O(1). \end{aligned}$$

Plugging these into the numerical flux function gives

$$\begin{pmatrix} (\rho u)^* \\ (\rho u^2 + \frac{M_{loc}^2}{M_{ref}^2}\pi + \frac{1-M_{ref}^2}{M_{loc}^2}\psi)^* \\ (\rho uv)^* \\ (u(E + M_{loc}^2\pi + (1-M_{loc}^2)\psi)^* \end{pmatrix} = \begin{pmatrix} \rho_0 u_0 + O(1) \\ \rho_0 u_0^2 + \frac{p_0}{M^2} + O(1) \\ \rho_0 u_0 v_0 + O(1) \\ (u_0(E_0 + p_0)) + O(1) \end{pmatrix}.$$

Therefore the scaling of the diffusion vector reads

$$D = \begin{pmatrix} O(1) \\ O(1) \\ O(1) \\ O(1) \end{pmatrix}.$$

This analysis shows in particular, that the diffusion of the scheme is independent of the Mach number.

Lastly for this section, the asymptotic preserving property of the scheme is discussed. To analyze the behavior of the scheme in the limit of $M \to 0$, consider now the full 2D - scheme as defined in section 2.5. Start by considering the equation for the momentum in the direction x_1

$$(\rho u)_{i,j}^{n+1} = (\rho u)_{i,j}^n + \frac{\Delta_t}{\Delta_x}\left(F_{1,i-1/2,j}^{\rho u} - F_{1,i+1/2,j}^{\rho u} + F_{2,i,j-1/2}^{\rho u} - F_{2,i,j+1/2}^{\rho u}\right), \tag{5.30}$$

where

$$\begin{aligned} F_{1,i\pm1/2,j}^{\rho u} &= (\rho u^2)_{i\pm1/2,j} + \pi_{i\pm1/2,j} + \frac{1-M^2}{M^2}\psi_{i\pm1/2,j}, \\ F_{2,i,j\pm1/2}^{\rho u} &= (\rho uv)_{i,j\pm1/2}. \end{aligned}$$

Multiply (5.30) with M^2 to get

$$\begin{aligned} &M^2\left((\rho u)_{i,j}^{n+1} - (\rho u)_{i,j}^n\right) = \\ &M^2\frac{\Delta_t}{\Delta_x}\left((\rho u^2)_{i-1/2,j} + \pi_{i-1/2,j} - (\rho u^2)_{i-1/2,j} - \pi_{i-1/2,j} + (\rho uv)_{i,j-1/2} - (\rho uv)_{i,j+1/2}\right) \\ &\qquad (1-M^2)\frac{\Delta_t}{\Delta_x}\left(\psi_{i-1/2,j} - \psi_{i+1/2,j}\right). \end{aligned} \tag{5.31}$$

By Lemma 5.2.2 it holds true that ρ, u, π and ψ are bounded in M. It is assumed that the discrete time derivative is also bounded in M. Therefore, in the limit $M \to 0$, there is

$$0 = \psi_{i-1/2,j} - \psi_{i+1/2,j}.$$

With the definition of ψ_C there is

$$\psi_{i-1/2,j} = \frac{p_{i-1,j} + p_{i,j}}{2}, \tag{5.32}$$

and therefore the following relations can be derived for the pressure

$$p_{i+1,j} = p_{i-1,j}. \tag{5.33}$$

From symmetry, a similar result from the equation for the y momentum can be derived

$$p_{i,j+1} = p_{i,j-1}. \tag{5.34}$$

To analyze the divergence constraint on the velocity field, consider the Energy equation

$$\begin{aligned} E_{i,j}^{n+1} = E_{i,j}^{n} + \frac{\Delta_t}{\Delta_x}((u(E + M^2\pi + (1-M^2)\psi))_{i-1/2,j} - (u(E + M^2\pi + (1-M^2)\psi))_{i+1/2,j} \\ + (v(E + M^2\pi + (1-M^2)\psi))_{i,j-1/2} - (v(E + M^2\pi + (1-M^2)\psi))_{i,j+1/2}). \end{aligned}$$

In this model there is $E = \rho e + M^2 \frac{u^2+v^2}{2}$. Therefore, when the low Mach number limit is considered, there is $E = (\rho e)$. Hence, it is beneficial to derive the limit of ρe in the states L^*, R^*, C^L, C^R.

$$e_{R^*} = e_R - \frac{M^2}{2c^2}(\pi_R^2 + \frac{1-M^2}{1+M^2}\psi_R^2) + \frac{M^2}{2c^2}(\pi_{R^*}^2 + \frac{1-M^2}{1+M^2}\psi_C^2), \tag{5.35}$$

$$e_{L^*} = e_L - \frac{M^2}{2c^2}(\pi_L^2 + \frac{1-M^2}{1+M^2}\psi_L^2) + \frac{M^2}{2c^2}(\pi_{L^*}^2 + \frac{1-M^2}{1+M^2}\psi_C^2). \tag{5.36}$$

From the formulas given in Lemma 5.2.2 it is straightforward to see that π_{L^*}, π_{R^*} and ψ_C are bounded in M, so that in the low Mach number limit there is

$$e_{R^*} = e_R \qquad\qquad e_{L^*} = e_L. \tag{5.37}$$

For the states in the center consider the following formulas

$$e_{C^R} = e_{R^*} - \pi_{R^*}\frac{M_{loc}^2\pi_{R^*} + 2(1-M_{loc}^2)\psi_C}{2c^2} + \pi_C\frac{M_{loc}^2\pi_C + 2(1-M_{loc}^2)\psi_C}{2c^2}, \tag{5.38}$$

$$e_{C^L} = e_{L^*} - \pi_{L^*}\frac{M_{loc}^2\pi_{L^*} + 2(1-M_{loc}^2)\psi_C}{2c^2} + \pi_C\frac{M_{loc}^2\pi_C + 2(1-M_{loc}^2)\psi_C}{2c^2}. \tag{5.39}$$

In order to take the limit, one has to check first the limit behavior of π_C. Following Lemma 5.2.2, π_C is defined as

$$\pi_C = \frac{\pi_{L^*} + \pi_{R^*}}{2} + c\frac{(u_{L^*} - u_{R^*})}{2}. \tag{5.40}$$

It is straightforward to see that $\lim_{M\to 0}\pi_{R^*,L^*} \to \pi_{R,L}$. Regarding the terms u_{L^*,R^*}, it is useful to write them in terms of the initial conditions to have

$$u_{L^*} = u_L - \frac{1}{cM(1+M^2)}(\frac{p_R - p_L}{2} + cM^2\frac{u_L - u_R}{2}), \tag{5.41}$$

$$u_{R^*} = u_R - \frac{1}{cM(1+M^2)}(\frac{p_L - p_R}{2} + cM^2\frac{u_L - u_R}{2}). \tag{5.42}$$

The initial conditions are in the set Ω_{AP}. Therefore, there is $p_L - p_R = O(M^2)$ and $u_L - u_R = O(1)$. So in the limit there is $u_{L^*,R^*} = u_{L,R}$ and therefore in total it holds

$$\lim_{M\to 0} \pi_C = p_0 + c\frac{(u_L - u_R)}{2}. \tag{5.43}$$

Similar, the limit $\lim_{M\to 0} \psi_C = p_0$ is computed. The limit for the internal energies e_{C^L}, e_{C^R} reads then

$$e_{C^R} = e_R - \frac{p_0(p_0 + c\frac{u_L - u_R}{2})}{c^2} \qquad e_{C^L} = e_L - \frac{p_0(p_0 + c\frac{u_L - u_R}{2})}{c^2}. \tag{5.44}$$

The same computation is performed for the inverse mass fractions. Rearranging the formulas given in Lemma 5.2.2 gives

$$\tau_{C^L} = \tau_L + \frac{1+M^4}{1+M^2}\frac{u_R - u_L}{2c} + \frac{\pi_L - \pi_R}{2c^2} \qquad \tau_{C^R} = \tau_R + \frac{1+M^4}{1+M^2}\frac{u_R - u_L}{2c} + \frac{\pi_R - \pi_L}{2c^2}.$$

When taking the limit of $M \to 0$ and using $p_L - p_R = O(M^2)$ it holds

$$\tau_{C^L} = \tau_L + \frac{u_R - u_L}{2c} \qquad \tau_{C^R} = \tau_R + \frac{u_R - u_L}{2c},$$

so the density reads

$$\rho_{C^L} = \rho_L\frac{2c}{2c + \rho_L(u_R - u_L)} \qquad \rho_{C^R} = \rho_R\frac{2c}{2c + \rho_R(u_R - u_L)}.$$

Therefore, by using $(\rho e)_{L,R} = \frac{\pi_{L,R}}{\gamma - 1}$, the energy in the low Mach number limit is computed to be

$$\rho_{C^L}e_{C^L} = \rho_L e_L\frac{2c}{2c + \rho_L(u_R - u_L)}(1 - \frac{p_0(p_0 + c\frac{u_L - u_R}{2})}{c^2 e_L}),$$

$$\rho_{C^R}e_{C^R} = \rho_R e_R\frac{2c}{2c + \rho_R(u_R - u_L)}(1 - \frac{p_0(p_0 + c\frac{u_L - u_R}{2})}{c^2 e_R}).$$

Now define the following quantities

$$H_{i\pm\frac{1}{2},j} = (\rho e)^*_{i\pm\frac{1}{2},j} + \psi^*_{i\pm\frac{1}{2},j},$$
$$H_{i,j\pm\frac{1}{2}} = (\rho e)^*_{i,j\pm\frac{1}{2}} + \psi^*_{i,j\pm\frac{1}{2}},$$

and rewrite the energy equation in the low Mach number limit in the following way

$$\frac{\pi^{n+1}_{i,j}}{\gamma-1} = \frac{\pi^n_{i,j}}{\gamma-1} + \frac{\Delta_t}{\Delta_x}\Big(\frac{u_{i-1,j}+u_{i,j}}{2}H_{i-\frac{1}{2},j} - \frac{u_{i,j}+u_{i+1,j}}{2}H_{i+\frac{1}{2},j} + \frac{v_{i,j-1}+v_{i,j}}{2}H_{i,j-\frac{1}{2}} - \frac{v_{i,j}+v_{i,j+1}}{2}H_{i,j+\frac{1}{2}}\Big).$$

Under the assumption that $\frac{\pi^{n+1}_{i,j}}{\gamma-1} = \frac{\pi^n_{i,j}}{\gamma-1}$ and some further manipulation, it holds that there is

$$(u_{i+1,j}-u_{i-1,j}) + (v_{i,j+1}-v_{i,j-1}) = u_{i,j}(H_{i-\frac{1}{2},j} - H_{i+\frac{1}{2},j}) + v_{i,j}(H_{i,j-\frac{1}{2}} - H_{i,j+\frac{1}{2}}) + u_{i-1,j}(1-H_{i-\frac{1}{2},j}) + u_{i+1,j}(1-H_{i+\frac{1}{2},j}) + v_{i,j-1}(1-H_{i,j-\frac{1}{2}}) + v_{i,j+1}(1-H_{i,j+\frac{1}{2}}).$$

Assuming sufficient regularity, the respective quantities can be approximated by a Taylor expansion. After some long calculations it can be shown that

$$\frac{u_{i+1,j}-u_{i-1,j}}{2\Delta_x} + \frac{v_{i,j+1}-v_{i,j-1}}{\Delta_x} = O(\Delta_x^2). \tag{5.45}$$

The following theorem combines all the results derived in this section.

Theorem 5.2.3 (Low Mach Properties)**.** *The numerical scheme defined based on the relaxation system (5.19) is asymptotic preserving for $M \to 0$, if W_L and W_R are given in Ω_{phys}, the states $W_{L^*}, W_{R^*}, W_{C^L}, W_{C^R}$ all belong to the set Ω_{AP} and the diffusion is independent of the Mach number.*

5.3 Numerical results

Now, the proposed low Mach number scheme is tested for its practical applicability. In all test cases, an ideal gas law is used with $\gamma = \frac{5}{3}$. Also in every test an equidistant grid is used. The emphasis in these tests lies in the comparison of the new relaxation scheme with respect to the standard Suliciu relaxation scheme. From the Theorem 5.2.2 it is clear that the standard scheme is recovered when choosing $M_{loc} = 1$. This scheme is denoted in the following with $S1$, while the new relaxation scheme is denoted by $S2$.

5.3.1 SOD Shock Tube test

The first test case investigates the capability of the low Mach number scheme to deal with discontinuities. To this end, the Sod shock tube test is concerned, see [155]. The computational domain is $D = [0, 1]$ and the initial conditions are set as

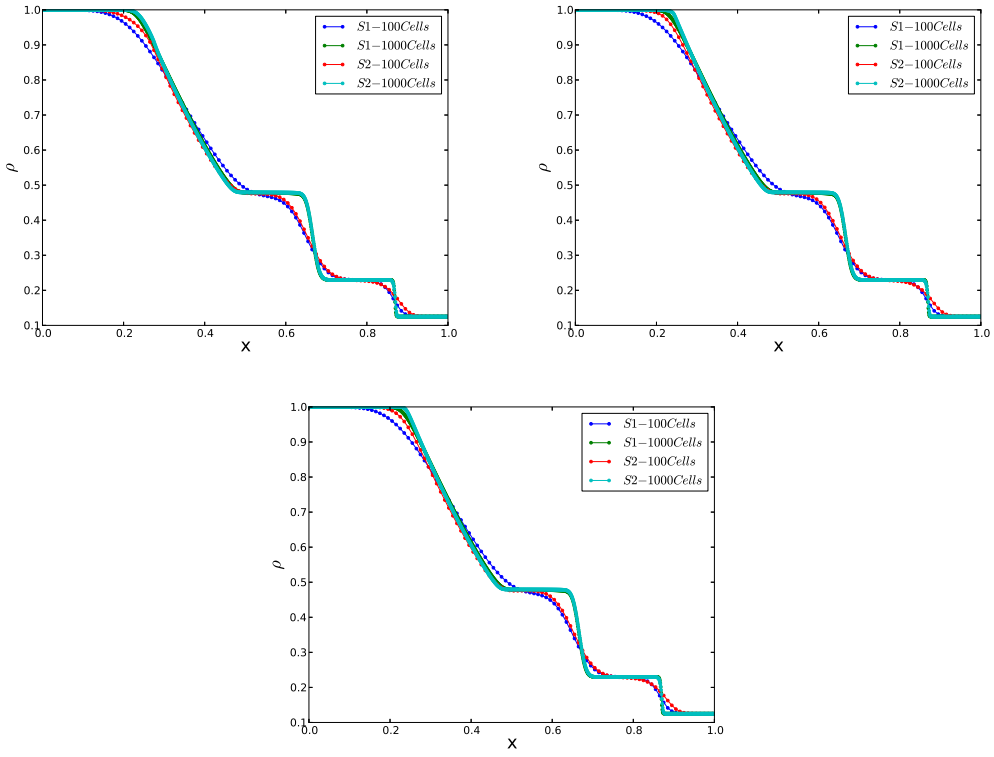


Fig. 5.3: Numerical approximations to the SOD shock tube test for the schemes $S1$ and $S2$ at different Mach numbers and at different resolutions at time 0.2. Top left: $M_{ref} = 10^{-1}$. Top right: $M_{ref} = 10^{-2}$. Bottom center: $M_{ref} = 10^{-3}$

$$\rho(x) = \begin{cases} 1.0 & \text{if} \quad x < 0.5, \\ 0.125 & \text{if} \quad x > 0.5, \end{cases} \tag{5.46}$$

and

$$p(x) = \begin{cases} 1.0 & \text{if} \quad x < 0.5, \\ 0.1 & \text{if} \quad x > 0.5. \end{cases} \tag{5.47}$$

and the velocity is set to zero. Only first order versions of the schemes $S1$ and $S2$ are concerned in order to investigate the influence of the numerical flux function on the approximation. Moreover, in this test case an explicit time integration is performed. In order to perform an explicit time integration for the scheme $S2$, the local Mach number has to be controlled in order for the fast eigenvalues, that scale with $O(\frac{1}{M_{loc}})$, to be bounded. Therefore in the scheme $S2$ it is set $M_{ref} = M_{loc}$. The results are shown in figure 5.3.

When looking at the numerical approximations with 100 cells and comparing them with the solutions on higher resolutions, a similar behavior on all the different Mach numbers can be observed. At first, the low Mach number scheme seems to be more diffusive on the shock around 0.9. Both schemes show a comparable performance on the contact discontinuity at 0.6, while the rarefaction wave is much better captured by the low Mach number scheme.

Moreover, both schemes are in good agreement in all regimes when the resolution is increased. However, a slight discrepancy is observed at the left side of the rarefaction wave, where the low Mach number scheme shows a sharper resolution.

5.3.2 Gresho Vortex test

As has been mentioned in section 5.1, the problems with low Mach number flow only emerge when multidimensional problems are concerned. A classical test case for low Mach number properties is the Gresho vortex. The Gresho vortex is an axisymmetric steady state solution of the compressible Euler equations. Here the modified version from [132] is considered. It is defined in polar coordinates and, since the solution is axisymmetric, only the radial component is specified. Denoting by u_ϕ the angular velocity, it is set as

$$u_\phi(r) = \begin{cases} 5r & \text{if} \quad 0 \leq r \leq 0.2, \\ 2 - 5r & \text{if} \quad 0.2 \leq r \leq 0.4, \\ 0 & \text{if} \quad 0.4 \leq r, \end{cases} \tag{5.48}$$

and the pressure distribution is given by

$$p(r) = p_0 + \begin{cases} \frac{25}{2} r^2 & \text{if} \quad 0 \leq r \leq 0.2, \\ \frac{25}{2} r^2 + 4(1 - 5r - \ln 0.2 + \ln r) & \text{if} \quad 0.2 \leq r \leq 0.4, \\ 4 \ln 2 - 2 & \text{if} \quad 0.4 \leq r, \end{cases} \tag{5.49}$$

where $p_0 = \frac{\rho}{\gamma M_{ref}^2}$. The density ρ is considered as constant and the computational domain is $D = [-1, 1] \times [-1, 1]$. Holding ρ and γ fixed, the reference Mach number M_{ref} is used to scale the vortex to different regimes. The respective schemes are implemented in the SLH code, where the implicit time integration method ESDIRK34 from [92] is used. In order to ensure a suitable convergence behavior for the Newton iteration the slopes are not limited as suggested in section 2.3.1, but are chosen as differentiable functions from the cell centered values from neighboring cells, see also [131]. The simulations are performed on an equidistant grid in both spatial dimensions with $N_x = N_y = 40$ and periodic boundary conditions are imposed. This resolution seems rather low. However, as it has been shown in the well-balanced tests, increasing the resolution will increase the quality of the numerical approximation. It is desired to see how the schemes perform on a not favorable low resolution. The resulting distributions of the relative Mach number, i.e. $M_{rel}(t, x, y) = \frac{M_{loc}(t,x,y)}{M_{loc}(0,x,y)}$, after one rotation for different reference Mach numbers are shown in figure 5.4.

The scheme $S1$ introduces an increasing amount of diffusion with decreasing Mach number; as can be seen in the left row of figure 5.4. In contrast to that, the new relaxation scheme $S2$ preserves the vortex structure on all Mach numbers equally good. This result is expected form the derivations of the numerical diffusion of the upwind schemes. The scheme $S1$ was shown to introduce a diffusion that scales with $O(\frac{1}{M_{ref}})$ in the momentum equations, while the diffusion for the scheme $S2$ is shown to be independent of the Mach number.

Another criterium to check the quality of the numerical approximation is the kinetic energy. Since the vortex is a stationary solution, the kinetic energy also remains constant in the exact solution. The evolution of the total kinetic energy in the computational domain is shown in figure 5.5.

The scheme $S1$ shows an increasing diffusion of the kinetic energy by decreasing Mach

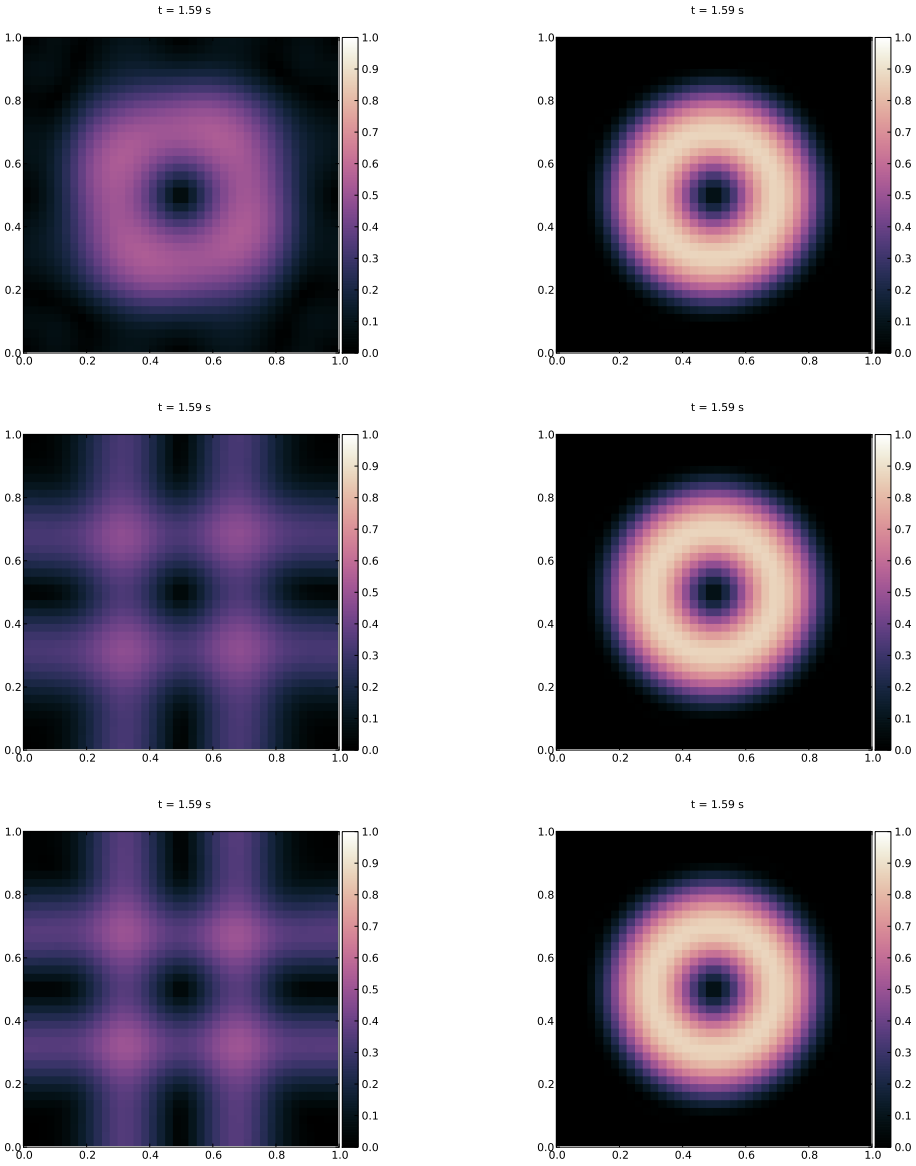


Fig. 5.4: Local relative Mach number for the Gresho Vortex after one rotation. Left: results for the scheme $S1$. Right: results for the scheme $S2$. From top to bottom the reference Mach numbers $M_{ref} = 10^{-2}$, $M_{ref} = 10^{-3}$ and $M_{ref} = 10^{-4}$ are chosen to set up the initial condition.

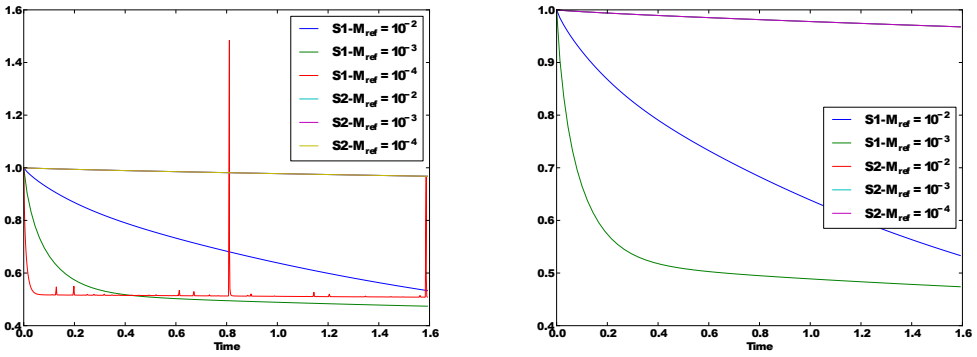


Fig. 5.5: Evolution of the total kinetic energies in the numerical approximation of the Gresho vortex at different Mach numbers for different schemes. Shown is the relative total kinetic energy, i.e. $\frac{tKE(t)}{tKE(0)}$. Left: With the scheme $S1$ at $M_{ref} = 10^{-4}$. Right: Without the scheme $S1$ at $M_{ref} = 10^{-4}$.

number. Even more, for Mach number 10^{-4}, the scheme actually shows also convergence problems and the solution becomes unphysical. On the other, hand the scheme $S2$ shows only a small diffusion of the kinetic energy. Moreover, the evolution of the total relative kinetic energy is almost identical at the different Mach numbers.

5.3.3 Kelvin Helmholtz Instability

The last test case concerns the approximation of a Kelvin-Helmholtz instability. The setup is taken also from [132]. The idea is to introduce a non steady flow problem to further investigate the influence of the numerical diffusion on the quality of the numerical approximations. The Kelvin-Helmholtz instability is a shear instability, where two flow regimes are considered, that are separated by a sharp discontinuity. The flow velocity is parallel to the shear discontinuity, but in opposite direction on either side. For the Euler equations, this is actually a steady solution. However, in numerical applications, due to numerical instabilities, the shear instability is triggered and a complex dynamic behavior of the numerical approximation occurs. However, the triggering of the shear instability due to numerical errors is not beneficial for comparing numerical solutions for different schemes, since these errors are random. Instead, a modification of the classical setup is suggested, such that at first only a specific mode of the instability is excited. This gives the possibility to compare the results for different schemes. Therefore the initial conditions are set as

$$\rho = \begin{cases} \rho_1 - \rho_m \exp(\frac{y-0.25}{L}) & \text{if} \quad 0 \leq y \leq 0.25, \\ \rho_2 + \rho_m \exp(\frac{-y+0.25}{L}) & \text{if} \quad 0.25 \leq y \leq 0.5, \\ \rho_2 + \rho_m \exp(\frac{y-0.75}{L}) & \text{if} \quad 0.5 \leq y \leq 0.75, \\ \rho_1 - \rho_m \exp(\frac{-y+0.75}{L}) & \text{if} \quad 0.75 \leq y \leq 1. \end{cases} \tag{5.50}$$

$$u = \begin{cases} u_1 - u_m \exp(\frac{y-0.25}{L}) & \text{if} \quad 0 \leq y \leq 0.25, \\ u_2 + u_m \exp(\frac{-y+0.25}{L}) & \text{if} \quad 0.25 \leq y \leq 0.5, \\ u_2 + u_m \exp(\frac{y-0.75}{L}) & \text{if} \quad 0.5 \leq y \leq 0.75, \\ u_1 - u_m \exp(\frac{-y+0.75}{L}) & \text{if} \quad 0.75 \leq y \leq 1. \end{cases} \tag{5.51}$$

and $p = 2.5$. The parameters are

$$\begin{aligned} \rho_1 &= 1.0, \qquad \rho_2 = 2.0, \qquad \rho_m = \frac{\rho_1 - \rho_2}{2}, \\ u_1 &= 1.0, \qquad u_2 = 2.0, \qquad u_m = \frac{\rho_1 - \rho_2}{2}, \end{aligned} \tag{5.52}$$

and $L = 0.025$. The computational domain is $D = [0, 1] \times [0, 1]$ and periodic boundary conditions are imposed. The instability is triggered by a perturbation in the vertical velocity as

$$v = 10^{-2} \sin(2\pi x), \tag{5.53}$$

and the simulations are performed with a Mach number of $M_{ref} = 10^{-2}$. The results are depicted in figure 5.6.

Again the scheme $S2$ shows a better resolution of the resulting instabilities. Moreover, the scheme $S1$ showed again convergence problems in the Newton iteration and therefore the instability got triggered at a later time.

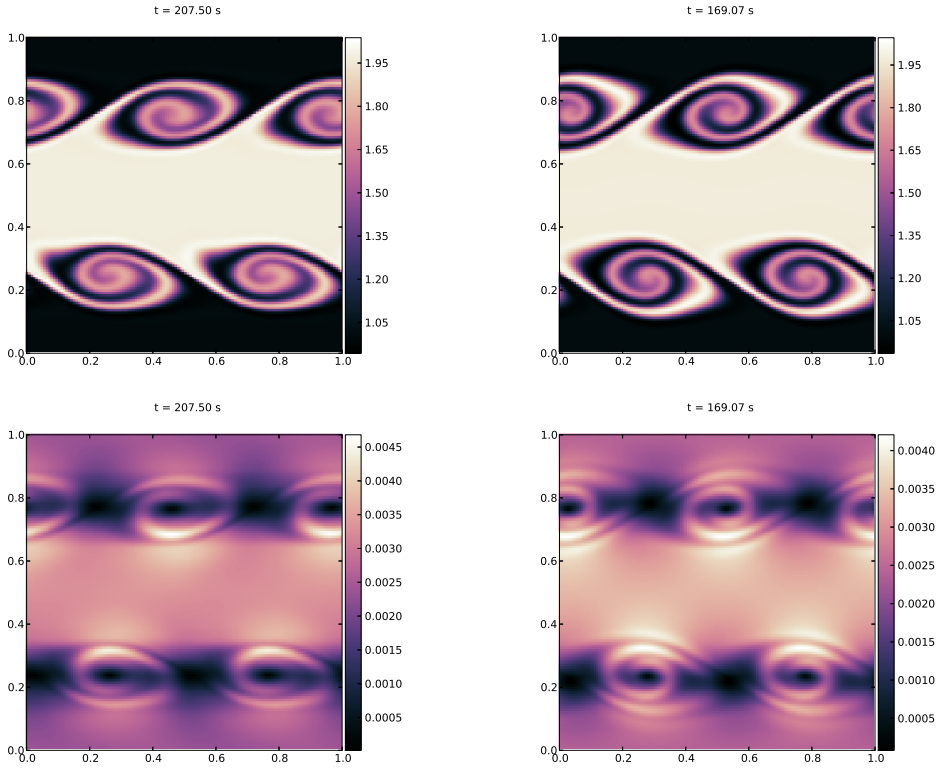


Fig. 5.6: Kelvin-Helmholtz Instability computed with the schemes $S1$ and $S2$ on a 128×128 grid. Left: Scheme $S1$. Right: Scheme $S2$. Top: Density. Bottom: Mach number.

6 A Low Diffusion scheme for the Euler Equations with Gravity

This chapter concerns the numerical approximations of the Euler equations with gravity in the low Mach number flow regime. The equations are discussed in section 1.4 and are given as

$$\begin{cases} \rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \\ (\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + I \frac{p}{M^2}) = -\frac{\rho}{Fr^2} \nabla \Phi, \\ E_t + \nabla \cdot (\mathbf{u}(E+p)) = -\rho \frac{M^2}{Fr^2} \langle u, \nabla \Phi \rangle, \\ \Phi_t = 0. \end{cases} \tag{6.1}$$

The numerical approximation of low Mach number flow regime for the homogeneous Euler equations is discussed in chapter 5. It is found there that the control of the scaling of the variables with respect to the Mach number in the approximate Riemann solver is crucial in order to develop a low diffusive scheme when low Mach number flows are considered. Now, due to the gravitational source term, also the Froude number is influencing the flow. As it is discussed in section 1.4, the limit behavior of the system (6.1) strongly depends on the relative behavior of the Froude and the Mach number. It is found that if and only if $O(M) = O(Fr)$ in the limit of $M \to 0$, the system (6.1) approaches the hydrostatic equilibrium. Atmospheres usually are close to a hydrostatic equilibrium and therefore in the following, when the scalings of the two different non-dimensional variables are considered, the case of

$$O(M) = O(Fr) \tag{6.2}$$

is assumed.

This gives rise to two challenges in the numerical approximation of (6.1) in the low Mach number regime. First, the accurate resolution of the hydrostatic equilibrium, as this is the limit behavior of the system. Second, controlling the diffusion in the low Mach number regime.

The accurate approximation of atmospheres is discussed in chapter 4, where a well-balanced approximate Riemann solver is developed. The approximations of low Mach number flows are considered in chapter 5, where an asymptotic preserving approximate Riemann solver is developed. In both cases, the Suliciu relaxation technique is used and adapted to the respective challenges. Therefore in this section, the techniques form chapter 4 and chapter 5 are combined to tackle the challenges in the approximation of solutions to (6.1) in the low Mach number flow regime, when the scaling (6.2) is assumed.

6.1 Derivation of the Relaxation System

Consider the relaxation system (5.19) developed for low Mach number flows and extend it with the gravitational source term from the system (6.1) to get the following relaxation system.

$$\begin{array}{rcccl}
\rho_t & + & (\rho u)_x & = & 0 \\
(\rho u)_t & + & (\rho u^2 + \frac{M_{loc}^2}{M_{ref}^2}\pi + \frac{1-M_{loc}^2}{M_{ref}^2}\psi)_x & = & -\frac{\rho}{Fr^2}\Phi_x \\
(\rho v)_t & + & (\rho v u)_x & = & 0 \\
E_t & + & (u(E + M_{loc}^2\pi + (1-M_{loc}^2)\psi)_x & = & -\frac{M_{ref}^2}{Fr^2}\rho u \Phi_x, \\
(\rho\pi)_t & + & (\rho u\pi + c^2 u)_x & = & \frac{\rho}{\epsilon}(p-\pi) \\
(\rho\psi)_t & + & (\rho u\psi + c^2\bar{u})_x & = & \frac{\rho}{\epsilon}(p-\psi) \\
(\rho\bar{u})_t & + & (\rho u\bar{u} + \frac{1}{M_{loc}^2 M_{ref}^2}\psi)_x & = & \frac{\rho}{\epsilon}(u-\bar{u})
\end{array} \tag{6.3}$$

where $E = \rho e + M_{ref}\frac{u^2+v^2}{2}$. In a first step, analogous to chapter 4, in order to derive a practical well-balanced scheme, a relaxation on the gravitational source term is introduced as follows

$$\begin{array}{rcccl}
\rho_t & + & (\rho u)_x & = & 0 \\
(\rho u)_t & + & (\rho u^2 + \frac{M_{loc}^2}{M_{ref}^2}\pi + \frac{1-M_{loc}^2}{M_{ref}^2}\psi)_x & = & -\frac{\overline{\rho}(\rho^-,\rho^+)}{Fr^2}Z_x \\
(\rho v)_t & + & (\rho v u)_x & = & 0 \\
E_t & + & (u(E + M_{loc}^2\pi + (1-M_{loc}^2)\psi)_x & = & -\frac{M_{ref}^2}{Fr^2}\overline{\rho}(\rho^-,\rho^+) u Z_x \\
(\rho\pi)_t & + & (\rho u\pi + c^2 u)_x & = & \frac{\rho}{\epsilon}(p-\pi) \\
(\rho\psi)_t & + & (\rho u\psi + c^2\bar{u})_x & = & \frac{\rho}{\epsilon}(p-\psi) \\
(\rho\bar{u})_t & + & (\rho u\bar{u} + \frac{1}{M_{loc}^2 M_{ref}^2}\psi)_x & = & \frac{\rho}{\epsilon}(u-\bar{u}) \\
Z_t & + & u Z_x & = & \frac{1}{\epsilon}(\Phi - Z) \\
\rho_t^- & + & (u - \frac{c^2}{\rho M_{ref}^2 M_{loc}^2} - \delta)\rho_x^- & = & \frac{1}{\epsilon}(\rho - \rho^-) \\
\rho_t^+ & + & (u + \frac{c^2}{\rho M_{ref}^2 M_{loc}^2} + \delta)\rho_x^+ & = & \frac{1}{\epsilon}(\rho - \rho^+)
\end{array} \tag{6.4}$$

The relaxation procedure on the gravitational potential is already explained in chapter 4. There, this additional procedure is shown to lead to a simple solution to the Riemann problem. However, the solution is not unique, since one Riemann invariant is missing. An additional equation has to be imposed on the centered wave in order to achieve the well-balanced property. As it has been shown in [56], the method of imposing this invariant is equivalent to impose an additional relaxation on the density in the source term. This procedure is also followed here and reflected by the last two equations of (6.4), where $\delta > 0$ is a parameter that can be chosen artificially small, such that in practice the CFL criterium is not affected. However, the evolution equations for ρ^- and ρ^+ are purely theoretical and only help to justify the quadrature of the density in the source term. For the sake of simplicity, they are omitted in the following, keeping in mind that they could be added without any problems in any step.

Now, consider the singularity in the source term, when the Froude number goes to zero. By assuming the relative scaling (6.2), it is expected that the solutions to the system (6.1) tend towards the hydrostatic equilibrium given by

$$\frac{p_x}{M_{ref}^2} = -\frac{\rho}{Fr^2}\Phi_x. \tag{6.5}$$

However, in the relaxation model (6.3), due to the splitting of the pressure, the following balance must hold in the equilibrium

$$\frac{M_{loc}^2}{M_{ref}^2}\pi_x + \frac{1-M_{loc}^2}{M_{ref}^2}\psi_x = -\frac{\rho}{Fr^2}\Phi_x. \tag{6.6}$$

It is now proposed to split the singularity in the source term in a similar manner as the pressure. Consider the following splitting

$$\frac{\rho}{Fr^2}\Phi_x = (\frac{Fr_{loc}^2}{Fr_{ref}^2}\rho + \frac{1-Fr_{loc}^2}{Fr_{ref}^2}\rho)\Phi_x. \tag{6.7}$$

Applying again a relaxation on the gravitational potential and the densities gives the following relation to hold in hydrostatic equilibrium

$$\frac{M_{loc}^2}{M_{ref}^2}\pi_x + \frac{1-M_{loc}^2}{M_{ref}^2}\psi_x = -(\frac{Fr_{loc}^2}{Fr_{ref}^2}\bar{\rho}_1 + \frac{1-Fr_{loc}^2}{Fr_{ref}^2}\bar{\rho}_2)Z_x. \tag{6.8}$$

In order to satisfy (6.8) it is decided, to enforce the hydrostatic balance on the terms with the same scaling of the non-dimensional variables. Therefore the relation (6.8) is split into two parts, i.e

$$\begin{aligned} \frac{M_{loc}^2}{M_{ref}^2}\pi_x &= -\frac{Fr_{loc}^2}{Fr_{ref}^2}\bar{\rho}_1 Z_x, \\ \frac{1-M_{loc}^2}{M_{ref}^2}\psi_x &= -\frac{1-Fr_{loc}^2}{Fr_{ref}^2}\bar{\rho}_2 Z_x. \end{aligned} \tag{6.9}$$

It is straightforward to see that if the relations (6.9) hold, so does (6.8). Observe that the density in the source term is also split into two different values. It is not clear at this point if different quadrature rules have to be used to achieve the balance (6.9).

In order to derive a well-balanced approximate Riemann solver, the designed relaxation system has to share the same hydrostatic relation than the original equations. In fact, the relations (6.9) are stronger than the original relation (6.5), while this reformulation is needed due to the splitting of the pressure in order to control the scaling in the low Mach number flow regime. However, the relaxation system (6.4) is not in equilibrium when the relations (6.9) are satisfied. The pressure term in the evolution equation for $\rho\bar{u}$ is not balanced. Hence, also in the relaxation part for the fast acoustics, the gravitational source term has to be considered. It is not desired to change the homogeneous part of the relaxation system since it has been proven beneficial for approximations of the homogeneous equations. So it is needed to derive how to add the source term to the fast acoustics such that the new relaxation system admits the hydrostatic relations (6.9). However, a simple reformulation of the second equation in (6.9) gives

$$\frac{\psi_x}{M_{loc}^2 M_{ref}^2} = -\frac{1-Fr_{loc}^2}{1-M_{loc}^2}\frac{\bar{\rho}_2}{Fr_{ref}^2 M_{loc}^2} Z_x. \tag{6.10}$$

Therefore, if the system (6.4) is extended by the right hand side of (6.10), it shares the hydrostatic relations (6.9). Extending the relaxation system (6.4) by the right hand side of (6.10) and applying the splitting of the source term (6.8) now leads to the following relaxation system

$$\begin{array}{ccccc}
\rho_t & + & (\rho u)_x & = & 0 \\
(\rho u)_t & + & (\rho u^2 + \frac{M_{loc}^2}{M_{ref}^2}\pi + \frac{1-M_{loc}^2}{M_{ref}^2}\psi)_x & = & -(\frac{Fr_{loc}^2}{Fr_{ref}^2}\bar{\rho}_1 + \frac{1-Fr_{loc}^2}{Fr_{ref}^2}\bar{\rho}_2)Z_x \\
(\rho v)_t & + & (\rho v u)_x & = & 0 \\
E_t & + & (u(E + M_{loc}^2\pi + (1-M_{loc}^2)\psi)_x & = & -\frac{M_{ref}^2}{Fr_{ref}^2}(Fr_{loc}^2\bar{\rho}_1 + (1-Fr_{loc}^2)\bar{\rho}_2)uZ_x \\
(\rho\pi)_t & + & (\rho u\pi + c^2 u)_x & = & \frac{\rho}{\epsilon}(p-\pi) \\
(\rho\psi)_t & + & (\rho u\psi + c^2\bar{u})_x & = & \frac{\rho}{\epsilon}(p-\psi) \\
(\rho\bar{u})_t & + & (\rho u\bar{u} + \frac{1}{M_{loc}^2 M_{ref}^2}\psi)_x & = & -\frac{1-Fr_{loc}^2}{1-M_{loc}^2}\frac{\bar{\rho}_2}{Fr_{ref}^2 M_{loc}^2}Z_x + \frac{\rho}{\epsilon}(u-\bar{u}) \\
Z_t & + & uZ_x & = & \frac{1}{\epsilon}(\Phi - Z)
\end{array} \tag{6.11}$$

Since the system (6.11) admits the hydrostatic relations (6.9), which are consistent with hydrostatic equilibrium (6.5), it is expected, that the resulting approximate Riemann solver is well-balanced. However, the well-balanced property is also expected to depend on the choice of ρ_1 and ρ_2. To analyze the influence of these quadratures on the approximate Riemann solver, it is now desired to compute the solution to the Riemann problem.

Lemma 6.1.1. *The system (6.11) is hyperbolic. The eigenvalues* $\lambda_{\pm_s} = u \pm \frac{M_{loc}^2}{M_{ref}^2}\frac{c}{\rho}$, $\lambda_{\pm_f} = u \pm \frac{1}{M_{loc}^2 M_{ref}}\frac{c}{\rho}$ *and* $\lambda_C = u$ *are linear degenerate where* $\lambda_c = u$ *has multiplicity* 4.

Proof. *The proof is straightforward and left to the reader.*

Following Lemma 6.1.1, the approximate Riemann solver defined by the system (6.11) admits the following solution structure

$$W_{\mathcal{R}}(\frac{x}{t}; W_L, W_R) = \begin{cases} W_L \text{ if } \frac{x}{t} < \lambda_{-_f}, \\ W_{L^*} \text{ if } \lambda_{-_f} < \frac{x}{t} < \lambda_{-_s}, \\ W_{C^L} \text{ if } \lambda_{-_s} < \frac{x}{t} < \lambda_c, \\ W_{C^R} \text{ if } \lambda_c < \frac{x}{t} < \lambda_{+_s}, \\ W_{L^*} \text{ if } \lambda_{+_s} < \frac{x}{t} < \lambda_{+_f}, \\ W_R \text{ if } \frac{x}{t} > \lambda_{+_f}, \end{cases} \tag{6.12}$$

also depicted in figure 6.1.

Now the choice of the different quadratures $\bar{\rho}_1$ and $\bar{\rho}_2$ is analyzed. Following the analysis of the derivation of the well-balanced scheme in chapter 4, in order to check if the relaxation system gives a well-balanced approximate Riemann solver, it is sufficient to check the intermediate states for the pressure. After a straightforward computation, it can be found that the relaxation pressures admit the form given in (6.13)

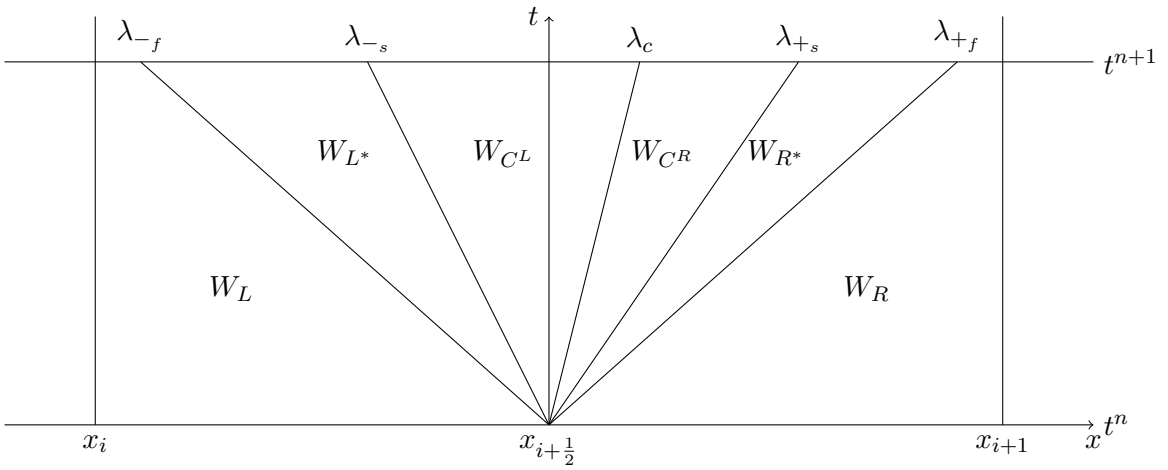


Fig. 6.1: Solution structure to the Riemann problem for the system (6.11)

$$
\begin{aligned}
\psi_{C^L} &= \psi_L + \frac{\psi_R - \psi_L + \frac{(1-Fr^2_{loc})M^2_{ref}}{(1-M^2_{loc})Fr^2_{ref}}\bar{\rho}_2(Z_R - Z_L)}{2} + cM_{loc}M_{ref}\frac{\bar{u}_L - \bar{u}_R}{2},\\
\psi_{C^R} &= \psi_R - \frac{\psi_R - \psi_L + \frac{(1-Fr^2_{loc})M^2_{ref}}{(1-M^2_{loc})Fr^2_{ref}}\bar{\rho}_2(Z_R - Z_L)}{2} + cM_{loc}M_{ref}\frac{\bar{u}_L - \bar{u}_R}{2},\\
\pi_{L^*} &= \pi_L + \frac{M^2_{loc}}{1+M^2_{loc}}(\psi_{CL} - \psi_L),\\
\pi_{R^*} &= \pi_R + \frac{M^2_{loc}}{1+M^2_{loc}}(\psi_{CR} - \psi_R),\\
\pi_{C^L} &= \pi^*_L + \frac{\pi^*_R - \pi^*_L + \frac{Fr^2_{loc}M^2_{ref}}{Fr^2_{ref}M^2_{loc}}\bar{\rho}_1(Z_R - Z_L)}{2} + \frac{cM_{ref}}{M_{loc}}\frac{u^*_L - u^*_R}{2},\\
\pi_{C^R} &= \pi^*_R - \frac{\pi^*_R - \pi^*_L + \frac{Fr^2_{loc}M^2_{ref}}{Fr^2_{ref}M^2_{loc}}\bar{\rho}_1(Z_R - Z_L)}{2} + \frac{cM_{ref}}{M_{loc}}\frac{u^*_L - u^*_R}{2}.
\end{aligned}
\tag{6.13}
$$

Assume now, that the initial condition for the Riemann problem admits a discrete hydrostatic equilibrium of the system (6.1) in the following way

$$
\begin{aligned}
u_L &= u_R = 0,\\
p_R - p_L &= -\frac{M^2_{ref}}{Fr^2_{ref}}\bar{\rho}(\Phi_R - \Phi_L).
\end{aligned}
\tag{6.14}
$$

For the approximate Riemann solver defined by (6.11) to be well-balanced it is necessary to have that from (6.14) there is

$$\begin{aligned}
\psi_{C^L} &= \psi_L = p_L, \\
\psi_{C^R} &= \psi_R = p_R, \\
\pi_{C^L} &= \pi_{L^*} = \pi_L = p_L, \\
\pi_{C^R} &= \pi_{R^*} = \pi_R = p_R.
\end{aligned} \tag{6.15}$$

To satisfy the relations (6.15) given (6.14), from (6.13) it is sufficient for the quadratures $\bar{\rho}_1$ and $\bar{\rho}_2$ to satisfy

$$\begin{aligned}
p_R - p_L &= -\frac{Fr_{loc}^2 M_{ref}^2}{Fr_{ref}^2 M_{loc}^2}\bar{\rho}_1(Z_R - Z_L), \\
p_R - p_L &= -\frac{(1 - Fr_{loc}^2)M_{ref}^2}{(1 - M_{loc}^2)Fr_{ref}^2}\bar{\rho}_2(Z_R - Z_L).
\end{aligned} \tag{6.16}$$

However, from the second relation in (6.14) the quadratures $\bar{\rho}_1$ and $\bar{\rho}_2$ must satisfy

$$\begin{aligned}
\bar{\rho}_1 &= \frac{M_{loc}^2}{Fr_{loc}^2}\bar{\rho}, \\
\bar{\rho}_2 &= \frac{1 - M_{loc}^2}{1 - Fr_{loc}^2}\bar{\rho}.
\end{aligned} \tag{6.17}$$

Therefore the splitting of the source term defined in (6.8) can now be further simplified as

$$(\frac{Fr_{loc}^2}{Fr_{ref}^2}\bar{\rho}_1 + \frac{1 - Fr_{loc}^2}{Fr_{ref}^2}\bar{\rho}_2)Z_x = (\frac{M_{loc}^2}{Fr_{ref}^2}\bar{\rho} + \frac{1 - M_{loc}^2}{Fr_{ref}^2}\bar{\rho})Z_x = \frac{\bar{\rho}}{Fr_{ref}^2}Z_x. \tag{6.18}$$

The relation (6.18) can now be used to further simplify the relaxation system (6.11) to get the new relaxation system

$$\begin{array}{ccccc}
\rho_t & + & (\rho u)_x & = & 0 \\
(\rho u)_t & + & (\rho u^2 + \frac{M_{loc}^2}{M_{ref}^2}\pi + \frac{1-M_{loc}^2}{M_{ref}^2}\psi)_x & = & -\frac{\bar{\rho}}{Fr_{ref}^2}Z_x \\
(\rho v)_t & + & (\rho v u)_x & = & 0 \\
E_t & + & (u(E + M_{loc}^2\pi + (1 - M_{loc}^2)\psi)_x & = & -\frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}uZ_x \\
(\rho\pi)_t & + & (\rho u\pi + c^2 u)_x & = & \frac{\rho}{\epsilon}(p - \pi) \\
(\rho\psi)_t & + & (\rho u\psi + c^2\bar{u})_x & = & \frac{\rho}{\epsilon}(p - \psi) \\
(\rho\bar{u})_t & + & (\rho u\bar{u} + \frac{1}{M_{loc}^2 M_{ref}^2}\psi)_x & = & -\frac{\bar{\rho}}{Fr_{ref}^2 M_{loc}^2}Z_x + \frac{\rho}{\epsilon}(u - \bar{u}) \\
Z_t & + & uZ_x & = & \frac{1}{\epsilon}(\Phi - Z)
\end{array} . \tag{6.19}$$

The relaxation system (6.19) is the final form of the derivations. In the next sections, the numerical properties of the approximate Riemann solver defined from system (6.19) are investigated. It should be remarked that even if the splitting of the source term is not present anymore, it is implicitly built into the system (6.19). In fact, the form (6.19) arises naturally

when considering the well-balanced property.

6.2 Stability and Robustness of the Relaxation Scheme

In this section, the stability, the algebra and the robustness of the relaxation scheme are discussed. The results are very similar to the results in the chapter 4 and chapter 5. However, they are repeated here for completeness. First, the stability of the relaxation system is discussed.

Theorem 6.2.1. *[Stability] The relaxation system (6.19) is a stable diffusive approximation of (6.1), provided the following subcharacteristic condition holds*

$$c^2 > \rho^2 p'. \tag{6.20}$$

Proof. *The relaxation in the variables π and ψ is equivalent to the relaxation in chapter 5, where it is found that to first order of the relaxation parameter there is*

$$\begin{aligned} \pi &= p - \epsilon((\frac{c^2}{\rho} - \rho p')u_x), \\ \psi &= p - \epsilon((\frac{c^2}{\rho} - \rho p')u_x). \end{aligned} \tag{6.21}$$

The relaxation parameter $\bar{u}$ does not appear in the flux for the original variables and therefore can be omitted in the analysis. The justification for this is given in Appendix A.

For the relaxation of the gravitational potential it can be found that to first order of the relaxation parameter ϵ there is

$$Z = \Phi - \epsilon u \Phi_x. \tag{6.22}$$

Using (6.21) and (6.22) in (6.19) gives

$$\begin{array}{ccccc} \rho_t & + & (\rho u)_x & = & 0 \\ (\rho u)_t & + & \frac{p_x}{M_{ref}^2} & = & -\frac{\bar{\rho}}{Fr_{ref}^2}\Phi_x + \epsilon\frac{\bar{\rho}}{Fr_{ref}^2}(u\Phi_x)_x + \epsilon\left(\frac{1}{\rho M_{ref}^2}(c^2 - \rho^2 p')u_x\right)_x \\ (\rho v)_t & + & (\rho v u)_x & = & 0 \\ E_t & + & (u(E+p))_x & = & -\frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}u\Phi_x + \epsilon\frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}u(u\Phi_x)_x + \epsilon\left(\frac{1}{\rho}(c^2 - \rho^2 p')(\frac{u^2}{2})_x\right)_x \end{array},$$

which gives the subcharacteristic condition (6.20).

The next lemma concerns the basic properties of the system (6.19) and the structure of the solution to the Riemann problem.

Lemma 6.2.1. *The system (6.19) is hyperbolic. The eigenvalues $\lambda_{\pm_s} = u \pm \frac{M_{loc}^2}{M_{ref}^2}\frac{c}{\rho}$, $\lambda_{\pm_f} = u \pm \frac{1}{M_{loc}^2 M_{ref}}\frac{c}{\rho}$ and $\lambda_C = u$ are linear degenerate where $\lambda_c = u$ has multiplicity 4. Moreover, the solution to the Riemann problem admits the following structure*

$$W_{\mathcal{R}}(\frac{x}{t};W_L,W_R)=\begin{cases}W_L \;\; if \;\; \frac{x}{t}<\lambda_{-_f},\\ W_{L^*} \;\; if \;\; \lambda_{-_f}<\frac{x}{t}<\lambda_{-_s},\\ W_{C^L} \;\; if \;\; \lambda_{-_s}<\frac{x}{t}<\lambda_c,\\ W_{C^R} \;\; if \;\; \lambda_c<\frac{x}{t}<\lambda_{+_s},\\ W_{L^*} \;\; if \;\; \lambda_{+_s}<\frac{x}{t}<\lambda_{+_f},\\ W_R \;\; if \;\; \frac{x}{t}>\lambda_{+_f},\end{cases} \tag{6.23}$$

where the solutions to the intermediate states can be computed explicitly. There is for ψ

$$\begin{cases}\psi_{CL} &= \psi_L+\frac{\psi_R-\psi_L+\frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}(Z_R-Z_L)}{2}+cM_{loc}M_{ref}\frac{\bar{u}_L-\bar{u}_R}{2},\\ \psi_{CR} &= \psi_R-\frac{\psi_R-\psi_L+\frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}(Z_R-Z_L)}{2}+cM_{loc}M_{ref}\frac{\bar{u}_L-\bar{u}_R}{2}.\end{cases} \tag{6.24}$$

For the relaxation pressure π *it holds that*

$$\begin{cases}\pi_{L^*} &= \pi_L+\frac{M_{loc}^2}{1+M_{loc}^2}(\psi_{CL}-\psi_L),\\ \pi_{R^*} &= \pi_R+\frac{M_{loc}^2}{1+M_{loc}^2}(\psi_{CR}-\psi_R),\\ \pi_{C^L} &= \pi_L^*+\frac{\pi_R^*-\pi_L^*+\frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}(Z_R-Z_L)}{2}+\frac{cM_{ref}}{M_{loc}}\frac{u_L^*-u_R^*}{2},\\ \pi_{C^R} &= \pi_R^*-\frac{\pi_R^*-\pi_L^*+\frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}(Z_R-Z_L)}{2}+\frac{cM_{ref}}{M_{loc}}\frac{u_L^*-u_R^*}{2}.\end{cases} \tag{6.25}$$

For the velocity u *there is*

$$\begin{cases}u_C &= \frac{u_L^*+u_R^*}{2}-\frac{M_{loc}}{cM_{ref}}\frac{\pi_R^*-\pi_L^*+\frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}(Z_R-Z_L)}{2},\\ u_{L^*} &= u_L-\frac{M_{loc}}{cM_{ref}^2(1+M_{loc}^2)}(\psi_{CL}-\psi_L),\\ u_{R^*} &= u_R+\frac{M_{loc}}{cM_{ref}^2(1+M_{loc}^2)}(\psi_{CR}-\psi_R).\end{cases} \tag{6.26}$$

For the inverse mass fractions τ *one can therefore find that*

$$\begin{cases}\tau_{L^*} &= \tau_L+\frac{\pi_L-\pi_{L^*}}{c^2},\\ \tau_{R^*} &= \tau_R+\frac{\pi_R-\pi_{R^*}}{c^2},\\ \tau_{C^L} &= \tau_L+\frac{\pi_L-\pi_C}{c^2},\\ \tau_{C^R} &= \tau_R+\frac{\pi_R-\pi_C}{c^2}.\end{cases} \tag{6.27}$$

And for the internal energies e *there is*

$$\begin{cases}e_{L^*} &= e_L-\frac{M_{loc}^2}{2c^2}(\pi_L-\pi_{L^*}+\frac{1-M_{loc}^2}{1+M_{loc}^2}(\psi_L-\psi_{CL})),\\ e_{R^*} &= e_R-\frac{M_{loc}^2}{2c^2}(\pi_R-\pi_{R^*}+\frac{1-M_{loc}^2}{1+M_{loc}^2}(\psi_R-\psi_{CR})),\\ e_{C^L} &= e_{L^*}-\pi_{L^*}\frac{M_{loc}^2\pi_{L^*}+2(1-M_{loc}^2)\psi_C}{2c^2}+\pi_C\frac{M_{loc}^2\pi_C+2(1-M_{loc}^2)\psi_C}{2c^2},\\ e_{C^R} &= e_{R^*}-\pi_{R^*}\frac{M_{loc}^2\pi_{R^*}+2(1-M_{loc}^2)\psi_C}{2c^2}+\pi_C\frac{M_{loc}^2\pi_C+2(1-M_{loc}^2)\psi_C}{2c^2}.\end{cases} \tag{6.28}$$

Proof. *The computation of the eigenvalues is straightforward and omitted for brevity. In order to determine the solution to the intermediate states, the eigenvectors to the respective eigenvalues have to be computed in order to determine the Riemann Invariants. In primitive variables* $(\rho, u, v, \pi, e, \bar{u}, \psi, Z)$ *there is*

- λ_c*: The eigenvectors read* $\begin{pmatrix} 1\\0\\0\\0\\0\\0\\0\\0 \end{pmatrix}, \begin{pmatrix} 0\\0\\1\\0\\0\\0\\0\\0 \end{pmatrix}, \begin{pmatrix} 0\\0\\0\\0\\1\\0\\0\\0 \end{pmatrix}, \begin{pmatrix} 0\\0\\0\\-\bar{\rho}\frac{M_{ref}^2}{Fr_{ref}^2}\\0\\-\bar{\rho}\frac{M_{ref}^2}{Fr_{ref}^2}\\1 \end{pmatrix}.$

  Therefore the Riemann Invariants are $\Phi_c = \{u, \bar{u}, \psi + \frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}Z, \pi + \frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}Z\}$.

- $\lambda_{\pm_s} = u \pm \frac{cM_{loc}}{\rho M_{ref}}$*: The eigenvectors read* $\begin{pmatrix} \rho^2\\ \pm\frac{cM_{loc}}{M_{ref}}\\ 0\\ c^2\\ M_{loc}^2\pi + (1-M_{loc}^2)\psi\\ 0\\ 0\\ 0 \end{pmatrix}.$

  Therefore the Riemann Invariants are
  $\Phi_{\pm_s} = \{u \pm \frac{cM_{loc}}{\rho M_{ref}}, \pi \mp \frac{cM_{ref}}{M_{loc}}u, e - \pi\frac{M_{loc}^2\pi + 2(1-M_{loc}^2)\psi}{2c^2}, \psi, \bar{u}, v, Z\}$.

- $\lambda_{\pm_f}$ *The eigenvectors read* $\begin{pmatrix} \rho^2\\ \pm\frac{c}{M_{ref}M_{loc}}\\ 0\\ c^2\\ M_{loc}^2\pi + (1-M_{loc}^2)\psi\\ \frac{\pm c(1+M_{loc}^2)}{M_{loc}^3 M_{ref}}\\ \frac{c^2(1+M_{loc}^2)}{M_{loc}^2}\\ 0 \end{pmatrix}.$

  Therefore the Riemann Invariants are

$$\Phi_{\pm_f} = \{\pi + \frac{c^2}{\rho}, \psi - \frac{1+M_{loc}^2}{M_{loc}^2}\pi, \psi \mp c\frac{M_{ref}(1+M_{loc})^2}{M_{loc}}u, \psi \mp cM_{loc}M_{ref}\bar{u},$$
$$e - \frac{M_{loc^2}}{2c^2}(\pi^2 + \frac{1-M_{loc}^2}{1+M_{loc}^2}\psi^2), v, Z\}. \quad (6.29)$$

The eigenvalues are linear degenerate, because they are Riemann invariants for their respective field. The solution to the intermediate states can be computed by using the given Riemann invariants.

Consider now the robustness of the approximate Riemann solver (6.23).

Theorem 6.2.2 (Robustness)**.** *Let $W_L, W_R \in \Omega_{phys}$.*
For $M_{loc} < 1$, $M_{ref} \notin \left[\frac{M_{loc}^2}{2+M_{loc}^2+\sqrt{1-M_{loc}^4}}, \frac{M_{loc}^2}{2+M_{loc}^2-\sqrt{1-M_{loc}^4}}\right]$ and choosing $c > 0$ large enough, the states W defined in lemma 6.2.1 also belong to the set Ω_{phys}.

Proof. *The positivity of the density follows from the ordering of the eigenvalues. Since the eigenvalues do not depend on the source term, the proof is analogous to the proof of Theorem 5.2.1.*

The proof of the positivity for the internal energies is also similar to the proof of Theorem 5.2.1, as soon as it is computed that also here it holds that

$$e_{L^*} = e_L + M_{loc}^2(\theta_1^2 + \frac{1-M_{loc}^2}{1+M_{loc}^2}\theta_2^2)\frac{(u_L-u_R)^2}{8} + O(\frac{1}{c}),$$
$$e_{L^C} = e_L + \left(M_{loc}^2\frac{1-M_{loc}^2}{1+M_{loc}^2}\theta_3^2 + M_{loc}^2\theta_2^2 - 2(1-M_{loc}^2)\theta_3(\theta_1-\theta_2)\right)\frac{(u_L-u_R)^2}{8} + O(\frac{1}{c}).$$

Finally, as in chapter 5, the consistency of the newly proposed relaxation scheme with the scheme developed in chapter 4 is considered.

Theorem 6.2.3. *(Consistency) For $M_{loc} \to 1$, the numerical scheme based on the system (6.19) goes to the numerical scheme based on the system (4.12).*

Proof. *When $M_{loc} \to 1$, the last two equations in (6.19) do not have any influence one the rest of the system. The upper part of the system (5.19) in turn is identical to the non-dimensional form of the relaxation system (4.12).*

6.3 Well-Balanced and Asymptotic Preserving Properties

This section is concerned with the well-balanced and asymptotic preserving properties of the proposed relaxation scheme. Following the notions of chapter 5, the numerical scheme for the non-dimentionalized Euler equations (6.1) is called asymptotic preserving if it gives a consistent discretization of the limit equations of (6.1) when $M, Fr \to 0$. Under the assumption (6.2), the limit equations are the hydrostatic equilibrium equations. Therefore in order to show that the scheme is asymptotic preserving, it is sufficient to show its well-balanced property. However, in (1.76) a set of asymptotic preserving states is defined for the system (6.1). As it is shown in chapter 5, preserving the scaling of the dependent variables is crucial in order to control the diffusion in the low Mach number regime. Therefore the scaling of the intermediate states will also be considered in this section.

First, the well balanced property of the approximate Riemann solver is considered.

Theorem 6.3.1 (Well-Balancedness)**.** *Let W_L and W_R be given in Ω_{phys} such that*

$$\begin{cases} u_L = u_R = 0, \\ p_R - p_L + \frac{M_{ref}^2}{Fr_{ref}^2}\overline{\rho}(W_L, W_R)(\Phi_R - \Phi_L) = 0. \end{cases} \tag{6.30}$$

Then the approximate Riemann solver is at rest, i.e. satisfies relation (2.82) and is therefore well-balanced.

Proof. *When looking at the intermediate states defined in Lemma 6.2.1 the proof is straightforward. At first the relations (6.30) give that* $\psi_{C^L} = p_L$ *and* $\psi_{C^R} = p_R$*. Therefore there is* $\pi_L^* = p_L$*,* $\pi_R^* = p_R$ *and* $u_L^* = u_R^* = 0$*. With these and again with (6.30), it further holds that* $\pi_{C^L} = p_L$*,* $\pi_{C^R} = p_R$ *and* $u_C = 0$*. The respective relations for* τ *and* e *follow from the properties of* π *and* ψ*.*

Remark 6.3.1. *The quadrature* $\overline{\rho}(W_L, W_R)$ *is general and in chapter 4 it is discussed how to choose this quadrature in order to exactly preserve certain classes of hydrostatic equilibria. The results form Lemma 4.2.3 can be directly applied also in this case.*

Since the scheme is now shown to be consistent with the hydrostatic equilibrium, the next step is to investigate the scaling of the dependent variables in the low Mach number regimes, which is the subject of the next Theorem.

Theorem 6.3.2 (Preservation of Scaling)**.** *Let* W_L *and* W_R *be given in* Ω_{AP}*, then the states* $W_{L^*}, W_{R^*}, W_{C^L}, W_{C^R}$ *all belong to the set* Ω_{AP}*.*

Proof. *Since the asymptotic preserving set in this case only gives restrictions on the pressure, the only thing to check is the scaling of* π *and* ψ*. From lemma 6.2.1 there is for* ψ

$$\psi_{CL} = \psi_L + \underbrace{\frac{\psi_R - \psi_L + \frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}(Z_R - Z_L)}{2}}_{=O(M^2)} + cM_{loc}M_{ref}\frac{\bar{u}_L - \bar{u}_R}{2} = p_L + O(M^2),$$

$$\psi_{CR} = \psi_R - \underbrace{\frac{\psi_R - \psi_L + \frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}(Z_R - Z_L)}{2}}_{=O(M^2)} + cM_{loc}M_{ref}\frac{\bar{u}_L - \bar{u}_R}{2} = p_R + O(M^2),$$

and therefore for the relaxation pressure π *it holds that*

$$\begin{aligned}
\pi_{L^*} &= \pi_L + \frac{M_{loc}^2}{1 + M_{loc}^2}\underbrace{(\psi_{CL} - \psi_L)}_{=O(M^2)} = p_L + O(M^4),\\
\pi_{R^*} &= \pi_R + \frac{M_{loc}^2}{1 + M_{loc}^2}\underbrace{(\psi_{CR} - \psi_R)}_{=O(M^2)} = p_R + O(M^4),\\
\pi_{C^L} &= \pi_L^* + \underbrace{\frac{\pi_R^* - \pi_L^* + \frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}(Z_R - Z_L)}{2}}_{=O(M^2)} + \frac{cM_{ref}}{M_{loc}}\frac{u_L^* - u_R^*}{2} = p_L + O(1),\\
\pi_{C^R} &= \pi_R^* - \underbrace{\frac{\pi_R^* - \pi_L^* + \frac{M_{ref}^2}{Fr_{ref}^2}\bar{\rho}(Z_R - Z_L)}{2}}_{=O(M^2)} + \frac{cM_{ref}}{M_{loc}}\frac{u_L^* - u_R^*}{2} = p_R + O(1).
\end{aligned} \tag{6.31}$$

Since π is in the momentum equation of relaxation system multiplied by a factor of $O(1)$, the derived scaling is consistent and therefore concludes the proof.

6.4 Definition of the numerical scheme

The numerical scheme is defined completely analogous to section 4.4. Observe that the source term has to be included into the flux function since the gravitational potential is advected with the fluid flow. Therefore the source term is not concentrated on the cell interface anymore and has a non zero contribution to the volume integral. Also keep in mind that in chapter 4 only time explicit discretization of (4.76) are concerned. In this chapter, for efficiency, an implicit time discretization is considered.

6.5 Numerical Results

In this section the practical performance of the proposed relaxation scheme is investigated. Following Theorem (6.2.3), to test the effect of the introduced splitting, two different schemes are compared. The scheme $S1$ is denoted as the numerical scheme resulting from the system (6.19), when the local Mach number is set to 1, i.e. $M_{loc} = 1$. The scheme denoted as $S2$ results from evaluating the parameter M_{loc} from the local flow properties. In every test an ideal gas law is used with $\gamma = \frac{5}{3}$. For efficiency, in every test in this chapter, only implicit time discretizations are used. The scheme is implemented in the SLH code, where the ESDIRK43 is chosen as the time integrator. Finally, equidistant grids are used.

6.5.1 Vortex in a Gravitational Field

The first test case concerns vortices in a gravitational field. In chapter 5, the Gresho vortex is used to show the advantage of the low Mach relaxation scheme. Here, it is desired to adapt a similar test case to the Euler equations with gravity. In the Appendix C, stationary axisymmetric vortices are derived. They are used as an initial condition and then numerically integrated for one rotation. The grid resolution is chosen as $N_x = N_y = 40$ and the computational domain is set $D = [0, 1] \times [0, 1]$. Additionally, in all tests the case of $M_{ref} = Fr_{ref}$ is considered. Since the solution is axisymmetric, it is given in polar coordinates where the center is set at $(0.5, 0.5)$. The gravitational potential is set as

$$\Phi(r) = r^2.$$

In the Appendix C, two different versions of such vortices are derived. In both cases it is critical to determine a velocity profile in dependence on the distance to the center. The geometric parameters r_i are set as

$$r_0 = 0, \qquad r_1 = 0.2, \qquad r_2 = 0.4. \tag{6.32}$$

First, a vortex on top of a fixed density distribution is derived. In the here assumed scaling of the Mach and the Froude number, the solutions are given by

$$\rho(r) = \begin{cases} \exp(-\frac{\Phi(r)}{RT_1}) & \text{if} \quad r \leq r_2, \\ \exp(-\frac{\Phi(r_2)}{R}(\frac{1}{T_1} - \frac{1}{T_2}))\exp(-\frac{\Phi(r)}{RT_2}) & \text{if} \quad r > r_2. \end{cases} \tag{6.33}$$

$$p(r) = \begin{cases} RT_1 \exp(-\frac{\Phi(r)}{RT_1}) + M^2 \exp(h(r)) + C & \text{if} \quad r \leq r_2, \\ RT_2\rho(r) & \text{if} \quad r > r_2. \end{cases} \tag{6.34}$$

$$v_\phi = \sqrt{r \exp((\frac{\Phi(r)}{RT_1} + h(r)))h(r)_r}, \tag{6.35}$$

with $C = -M^2 \exp(h(0))$ and $T_2 = \frac{p(r_2)}{R\rho(r_2)}$.

The parameter function $h(r)$ is chosen as a piecewise quadratic function, with

$$h(r) = \begin{cases} a_{0,1}r + \frac{a_{1,1}}{2}r^2 - C_1 & \text{if } r \leq r_1, \\ a_{0,2}r + \frac{a_{1,2}}{2}r^2 - C_1 + C_2 - C_3 & \text{if } r_1 \leq r \leq r_2, \\ C_4 - C_1 + C_2 - C_3 & \text{if } r_2 \leq r. \end{cases} \tag{6.36}$$

Following Appendix C, the vortex is completely defined by setting the parameters as

$$h(0) = 0, \quad \bar{h} = 1, \quad R = 8.3144598 \quad \rho(0) = 1, \quad p(0) = \frac{\rho(0)\pi^2}{71\frac{7}{8}M_{ref}^2}. \tag{6.37}$$

The second family of vortices is derived by considering a constant temperature throughout the entire computational domain. The distributions of the dependent variables are given as

$$\begin{aligned} \rho(r) &= \exp(\frac{M^2}{RT}F(r)), \\ p(r) &= RT\rho, \end{aligned} \tag{6.38}$$

where

$$F'(r) = f(r) = \frac{v_\phi^2}{r} - \frac{\Phi_r}{Fr^2}. \tag{6.39}$$

The primitve of the term $\frac{v_\phi^2}{r}$ is computed as

$$\int_0^r \frac{v_\phi^2}{2}dr = \begin{cases} a_{0,1}^2 \log r + 2a_{0,1}a_{1,1}r + \frac{a_{1,1}^2}{2}r^2 - C_1 & \text{if } r \leq r_1, \\ a_{0,2}^2 \log r + 2a_{0,2}a_{1,2}r + \frac{a_{1,2}^2}{2}r^2 - C_1 + C_2 - C_3 & \text{if } r_1 \leq r \leq r_2, \\ C_4 - C_1 + C_2 - C_3 & \text{if } r_2 \leq r. \end{cases} \tag{6.40}$$

Following Appendix C, the solution is completely determined by setting the following parameters

$$R = 8.3144598, \quad \bar{v} = 1, \quad \rho(0) = 1, \quad p(0) = \frac{\rho(0)\pi^2}{15\frac{5}{8}M_{ref}^2}. \tag{6.41}$$

For both families, if the angular velocity v_ϕ tends to zero, the vortices tend towards an isothermal hydrostatic equilibrium. The vortices (6.33) - (6.35) are derived in the spirit of the original Gresho vortex, where, on top of a constant density profile, the centrifugal

forces emerging from the velocity profile are balanced by a pressure gradient, which results from an increase in temperature in the outer parts of the vortex. Here the constant density is substituted by a density distribution with respect to an isothermal equilibrium. The centrifugal forces are then balanced by an increase in temperature in the outer parts of the vortex. However, when the solution further than r_2 is considered, an isothermal equilibrium solution has to be taken, where the temperature is different from the temperature that is used to determine the density profile inside the vortex. This is due to the increase in temperature through the vortex and the temperature outside the vortex is just the temperature at the boundary of the vortex. The families (6.38)- (6.40) are computed by considering a fixed temperature in all the computational domain. The formulas are simpler as for the other family. However, when approximating flows in a gravitational field, in chapter 4 it is found that the density distribution is critical for the quality of numerical approximations. Only for specific density distributions the quadrature of the source term proposed in (4.16) can be exact. If a given density distribution is not captured exactly by the quadrature, numerical errors are introduced by the scheme. The density distribution in (6.38) is not of an isothermal type. It is expected that numerical errors due to the source term discretization will influence the numerical integration. The density profile (6.33) is of an isothermal type. However, there is a change in temperature at the boundary of the vortex. This may also lead to inconsistencies in the approximation of the source term. The numerical results for the scheme $S2$ are depicted in figure 6.2.

In both cases, the scheme is able to capture the vortex dynamics. However, in contrast to the Gresho vortex test in chapter 5, the diffusion still depends on the Mach number. Although not as strong as compared to the standard upwind scheme. It is conjectured that this behavior is related to the quadrature of the source term. As it has been discussed, the quadrature is not consistent with any of the density distributions. Although it is suspected that the errors are smaller in the case of the vortex families (6.33) - (6.35). Also in figure 6.2 it can be seen, that the diffusion of the vortex is stronger in the case of the families (6.38)-(6.40).

Additionally, in figure 6.3 the evolution of the relative total kinetic energies is plotted in the different regimes for the different schemes. It can be seen that for both vortex families the scheme $S2$ performs significantly better then the scheme $S1$ when the results at the same Mach number is compared. The scheme $S1$ again shows convergence problems in the Newton iteration and results can not be shown in every regime. Also the same dependence of the type of vortex with respect to the diffusion can be seen as in figure 6.2.

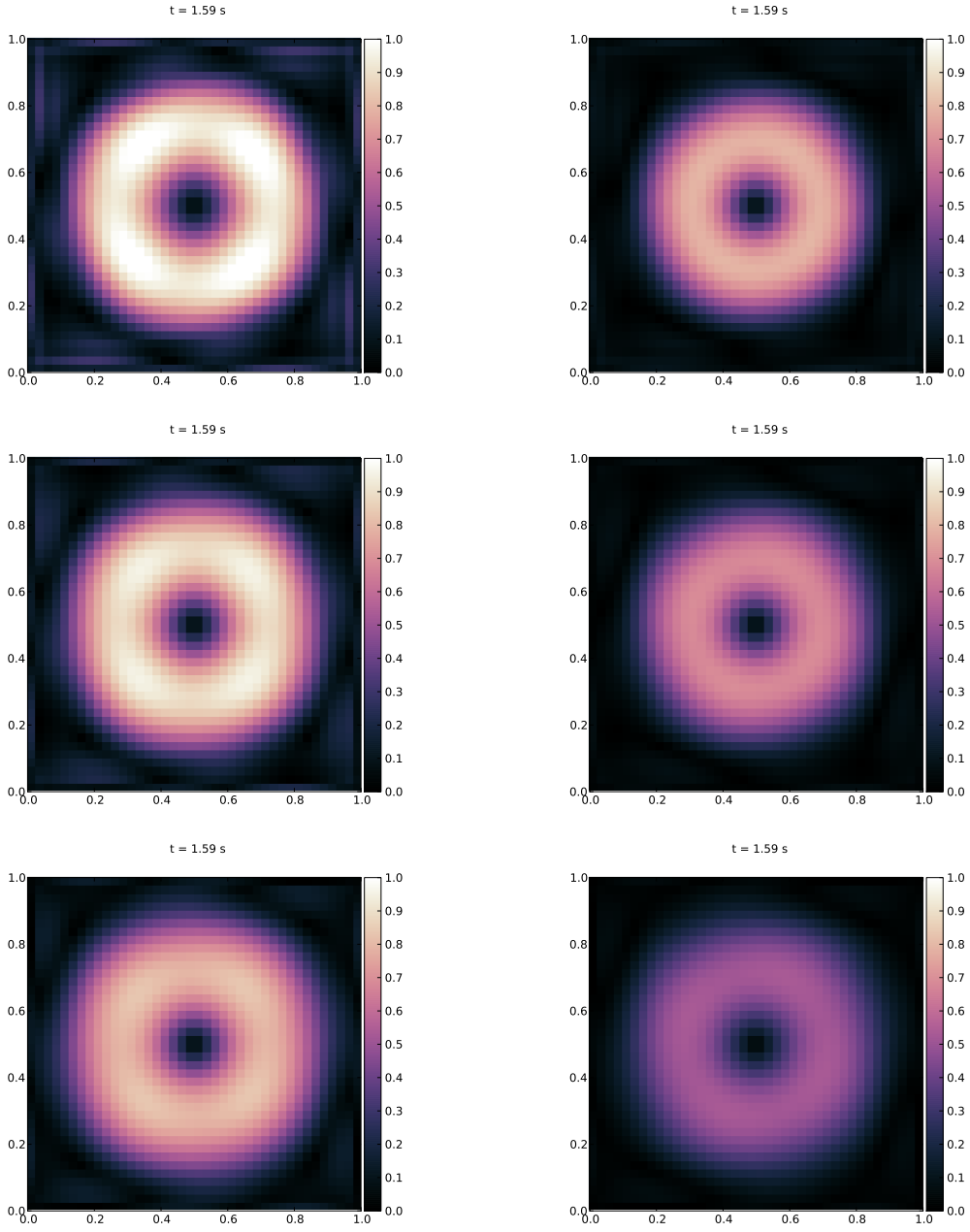


Fig. 6.2: Mach numbers of the vortices after one rotation computed by the scheme $S2$. Left: Vortices of the type (6.33) - (6.35). Right: Vortices of the type (6.38)- (6.40). From top to bottom the reference Mach number is chosen as $M_{ref} = 10^{-2}$,$M_{ref} = 10^{-3}$ and $M_{ref} = 10^{-4}$ respectively.

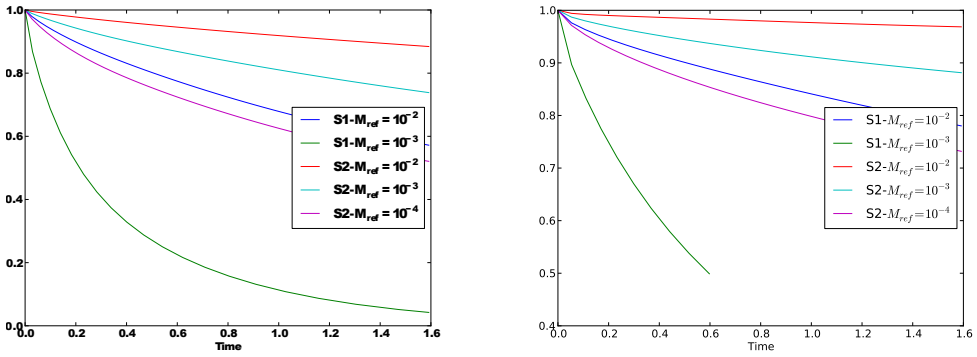


Fig. 6.3: Relative total kinetic energies of the vortices after one rotation computed by the schemes $S1$ and $S2$ at different regimes. Left: Vortices of the type (6.33) - (6.35). Right: Vortices of the type (6.38)- (6.40).

6.5.2 Rise of a Hot Bubble

The next test case is suggested in [130]. The computational domain is set as $D = [0km, 10km] \times [0km, 15km]$. In x-direction periodic boundary conditions and in y-direction solid wall boundary conditions are considered. The gravitational potential is set as

$$\Phi(y) = gy, \tag{6.42}$$

where $g = 9.81 \frac{m}{s^2}$, i.e. the gravitational acceleration is along the y-axis. The stratification of the atmosphere is defined in terms of the potential temperature θ. The potential temperature is the temperature a fluid parcel would get, if it would be brought adiabatically to a standard pressure p_0. θ is defined by the following relation

$$\theta = T(\frac{p_0}{p})^{\frac{R}{c_p}}, \tag{6.43}$$

where p_0 is a reference pressure and is set as $p_0 = 1bar$. The potential temperature is considered as constant throughout the atmosphere as $\theta_{eq} = 300K$. Rearranging (6.43) by using the relations $\gamma = \frac{c_p}{c_v}$ and $c_p - c_v = R$, the density and the pressure are connected in the atmosphere by the following relation

$$p_{eq} = \rho_{eq}^{\gamma}(R\theta_{eq})^{\gamma} p_0^{\gamma\frac{R}{c_P}}. \tag{6.44}$$

Therefore the atmosphere is of the polytropic type given in (4.6). In particular, due to the polytropic coefficient, it is also isentropic. The reference pressure p_0 is imposed at the bottom of the computational domain. Therefore the density at the bottom satisfies

$$\rho_{eq}(x, 0) = \frac{p_0}{\theta_{eq} R}. \tag{6.45}$$

The solution can then be integrated over the whole domain following the formulas in (4.6). The aim is to compute approximations to non-stationary solutions. To this end, a small disturbance in the potential temperature on the atmosphere is considered. It is

modeled as an axisymmetric bubble with the center $x_c = 5.0km$ and $y_c = 2.75km$. The distance to the center r is computed by

$$r = \left(\frac{x - x_c}{r_0}\right)^2 + \left(\frac{y - y_c}{r_0}\right)^2, \tag{6.46}$$

where $r_0 = 2.5$ is a scaling factor. The perturbations is now set as

$$\theta - \theta_{eq} = \begin{cases} \theta_0 \cos^2(\frac{\pi r}{2}) & \text{if} \quad r \leq 1, \\ 0 & \text{else,} \end{cases} \tag{6.47}$$

for $\theta_0 = 6.6K$. For the non-dimentionalization, a reference Mach number of $M_{ref} = 10^{-4}$ is chosen.

The evolution of the perturbation is now integrated by the schemes $S1$ and $S2$ up to the time $t = 800s$. It is hard to compute the exact evolution of the perturbation. Moreover, in contrast to the one dimensional tests in chapter 4, multidimensional effects influence the solution and it is not clear in advance how to judge the quality of the numerical approximation. Therefore it is chosen to compare the different schemes also on different mesh sizes. Assuming a convergent behavior of the numerical schemes, it is hoped that this gives a deeper intuition on the expected structure of the solution.

Three different resolutions are considered

	N_x	N_y
Low Res.	40	60
Mid. Res.	80	120
High Res.	120	180

, resulting in an equidistant cartesian mesh.

It is decided to analyze three different variables in this test. First, the resulting local Mach numbers are investigated in order to determine whether the numerical approximations are in the low Mach number regime. This is depicted in figure 6.4. After that, the resulting density distributions, see figure 6.5, and temperature distributions, see figure 6.6, are discussed.

From figure 6.4, it can be seen, that in all the simulations the maximum local Mach number is about 10^{-2}. Concluding from the previous simulations, it is expected that the scheme $S1$ will be more diffusive then the scheme $S2$ in this regime. Although the effect is maybe not too strong. Now the quality of the numerical approximation is analyzed by investigating the results shown in figure 6.5 and figure 6.6. In both variables it can be seen, that the numerical approximations show a convergent behavior when the resolution is increased. Therefore the solutions at the highest resolution are taken as reference solutions to which the other approximations can be compared. The scheme $S2$ shows definitely a less diffusive behavior. Especially the vortices at the edges of the rising bubble are better captured by the scheme $S2$. Indeed, increasing the resolution results in better approximations for the scheme $S1$.

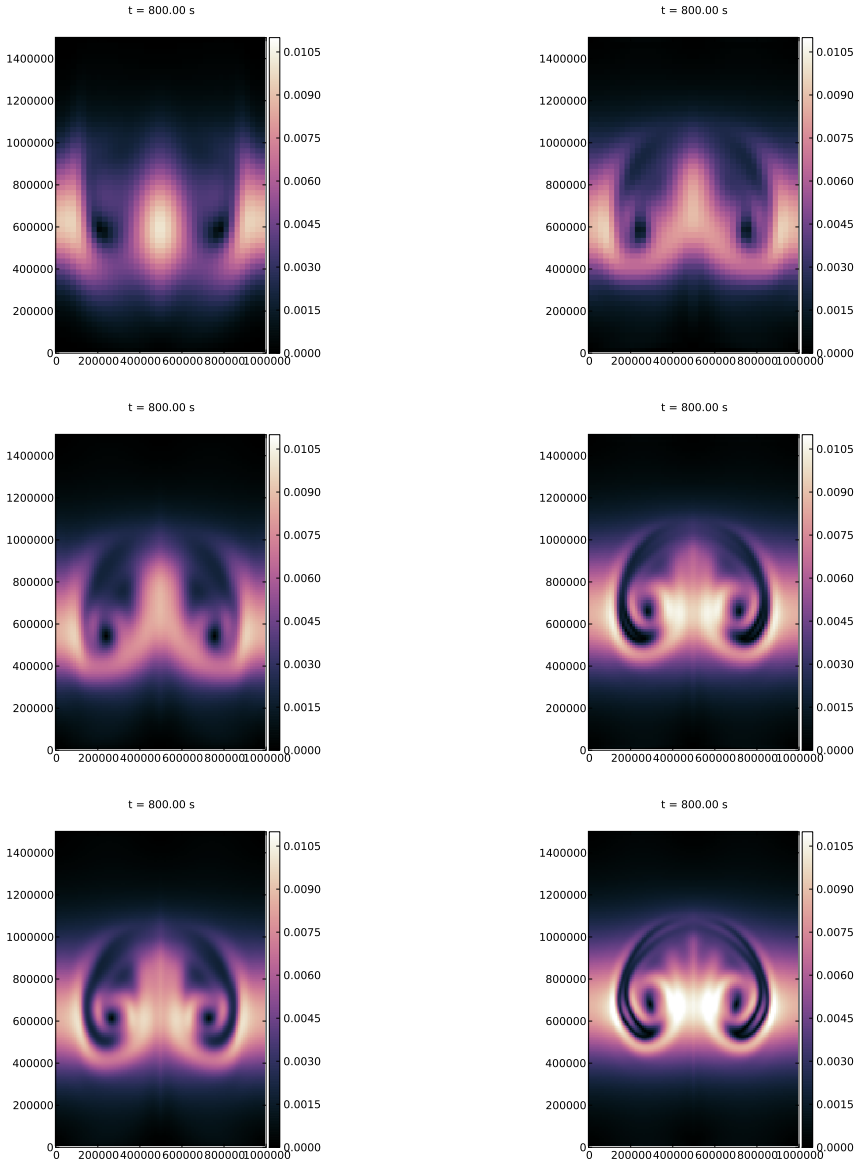


Fig. 6.4: Mach number in the rise of a hot bubble test at $t = 800s$. Left: Scheme $S1$. Right: Scheme $S2$. The resolution increases from top to bottom.

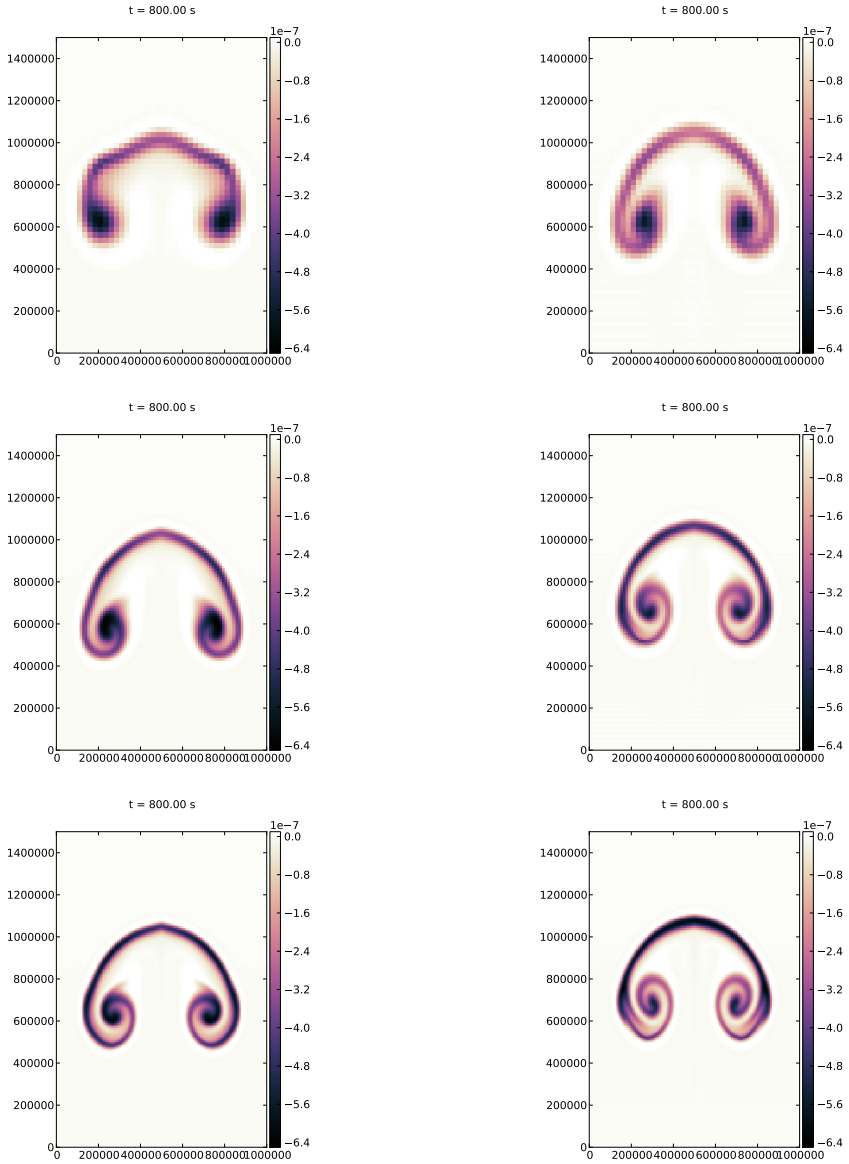


Fig. 6.5: Fluctuations in density with respect to the background atmosphere in the rise of a hot bubble test at $t = 800s$. Left: Scheme $S1$. Right: Scheme $S2$. The resolution increases from top to bottom.

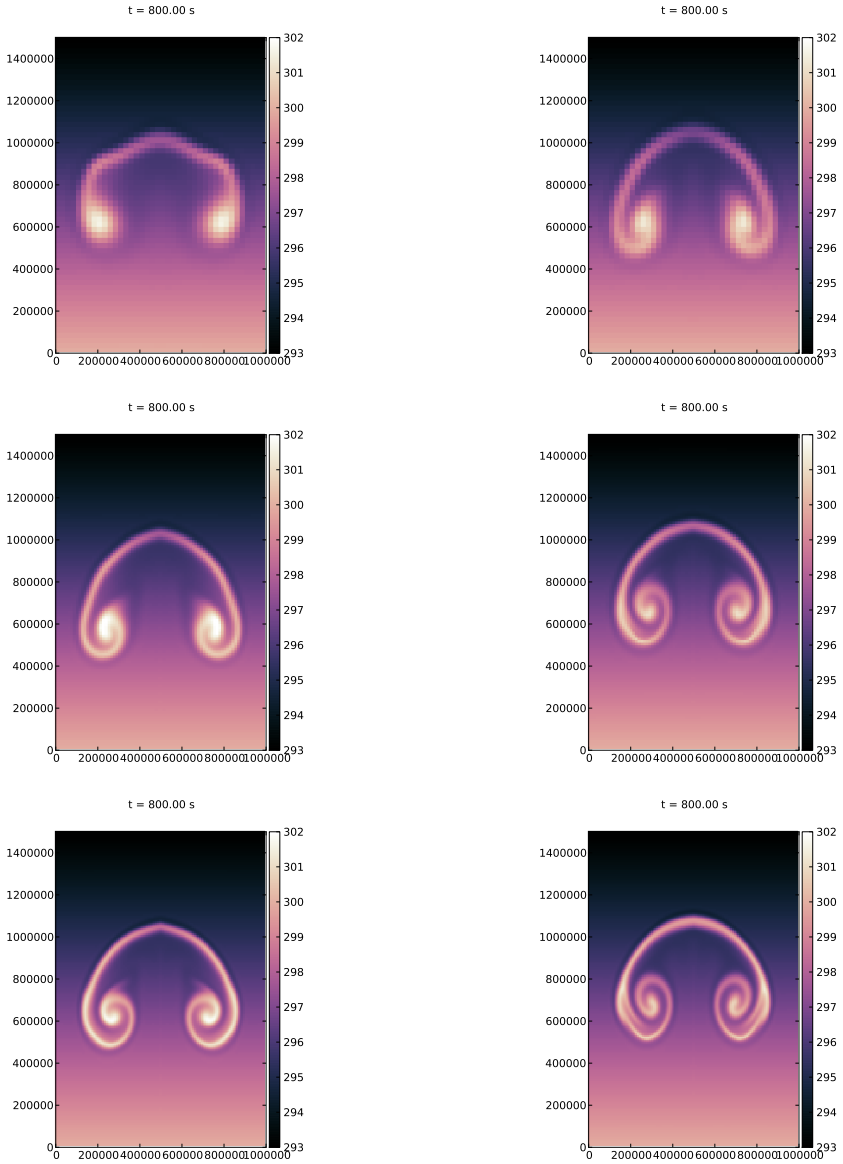


Fig. 6.6: Temperature in the rise of a hot bubble test at $t = 800s$. Left: Scheme $S1$. Right: Scheme $S2$. The resolution increases from top to bottom.

6.5.3 Hot and Cold Bubbles

The next test case is suggested in [146]. The atmosphere is the same as in the test case in section 6.5.2. However, now two perturbations are considered. A large hot bubble rising and a small cold bubble falling in the atmosphere. Both bubbles are initialized as follows

$$\Delta\theta_i = \begin{cases} \theta_i \cos^2(\frac{\pi r}{2}) & \text{if} \quad r \leq r_i, \\ \theta_i \exp(-\frac{(r-r_i)^2}{\sigma^2}) & \text{else,} \end{cases} \tag{6.48}$$

where

$$r = (x - x_i)^2 + (y - y_i)^2 \,, \tag{6.49}$$

for $i \in \{cold, hot\}$. The total potential temperature is then given by

$$\theta - \theta_{eq} = \Delta\theta_{hot} + \Delta\theta_{cold}. \tag{6.50}$$

The parameters for the two bubbles are given as follows

$$\begin{array}{llll} r_{hot} = 150m & x_{hot} = 500m & y_{hot} = 300m & \Delta\theta_{hot} = 0.5K \,, \\ r_{cold} = 0m & x_{cold} = 560m & y_{cold} = 640m & \Delta\theta_{cold} = -0.15K \,, \end{array}$$

while for both bubbles there is $\sigma = 50m$. The computational domain is set as $D = [0m, 1000m] \times [0m, 1000m]$ and the boundary conditions are the same as in the test from section 6.5.2. Again the numerical approximations are computed at different resolutions in order to investigate the qualitative behavior of the numerical schemes. Three different resolutions are considered

	N_x	N_y
Low Res.	50	50
Mid Res.	100	100
High Res.	150	150

, resulting in an equidistant cartesian mesh.

Again the two different schemes $S1$ and $S2$ are compared for their performance. The initial condition is integrated up to time $t = 600s$. First analyze the resulting Mach number in this test case. From figure 6.7 it can be seen that the Mach numbers in that test case do not exceed the regime of 10^{-3}. Therefore it is expected that there is a significant difference in the performance of both schemes. In fact, when analyzing the density fluctuations and temperature distributions from figure 6.8 and figure 6.9 respectively it can be concluded that the scheme $S2$ performs significantly better in capturing the fine structures of the solution. Just when the highest resolution is chosen for the scheme $S1$, the results are comparable to the results computed with the scheme $S2$ on the lowest resolution.

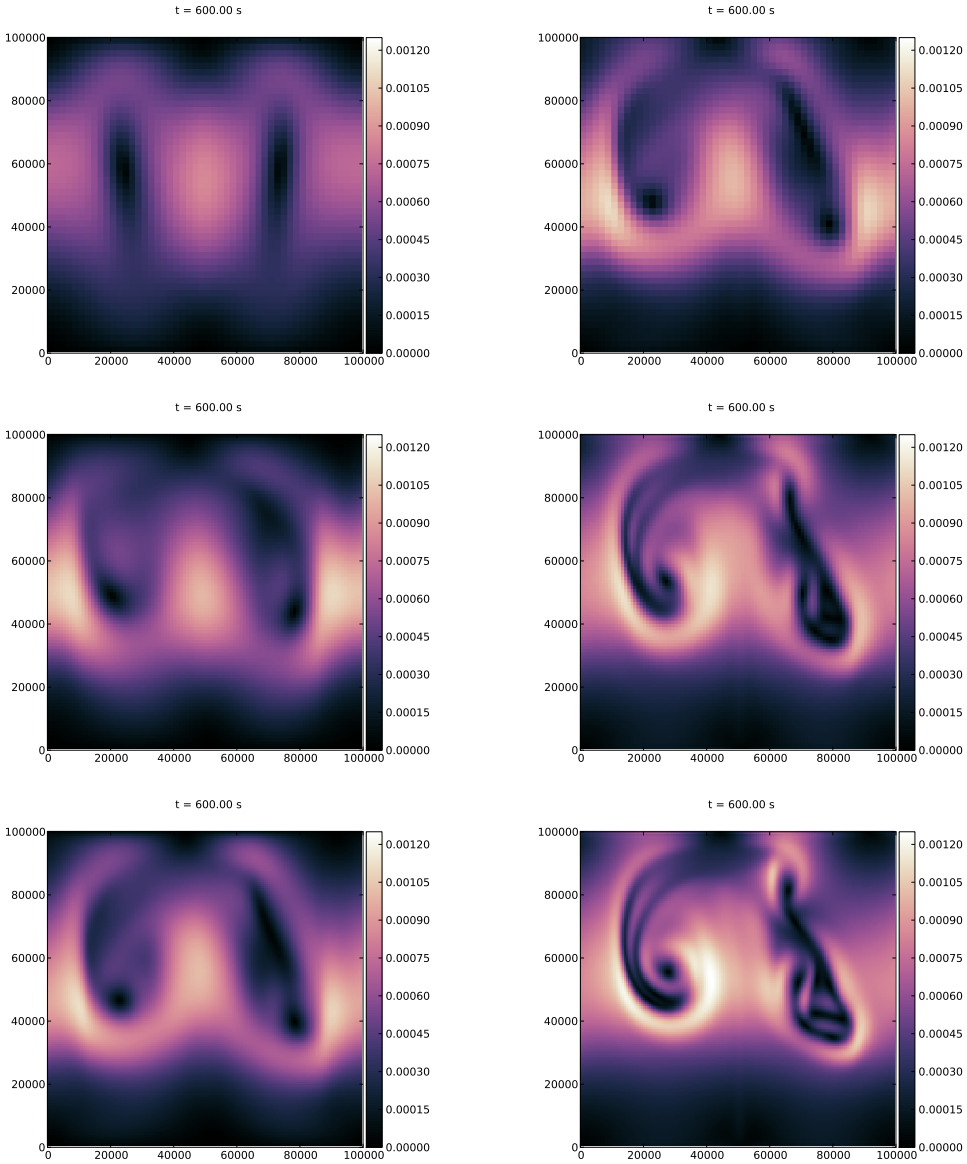


Fig. 6.7: Mach number in the hot and cold bubble test case at $t = 600s$. Left: Scheme $S1$. Right: Scheme $S2$. The resolution increases from top to bottom.

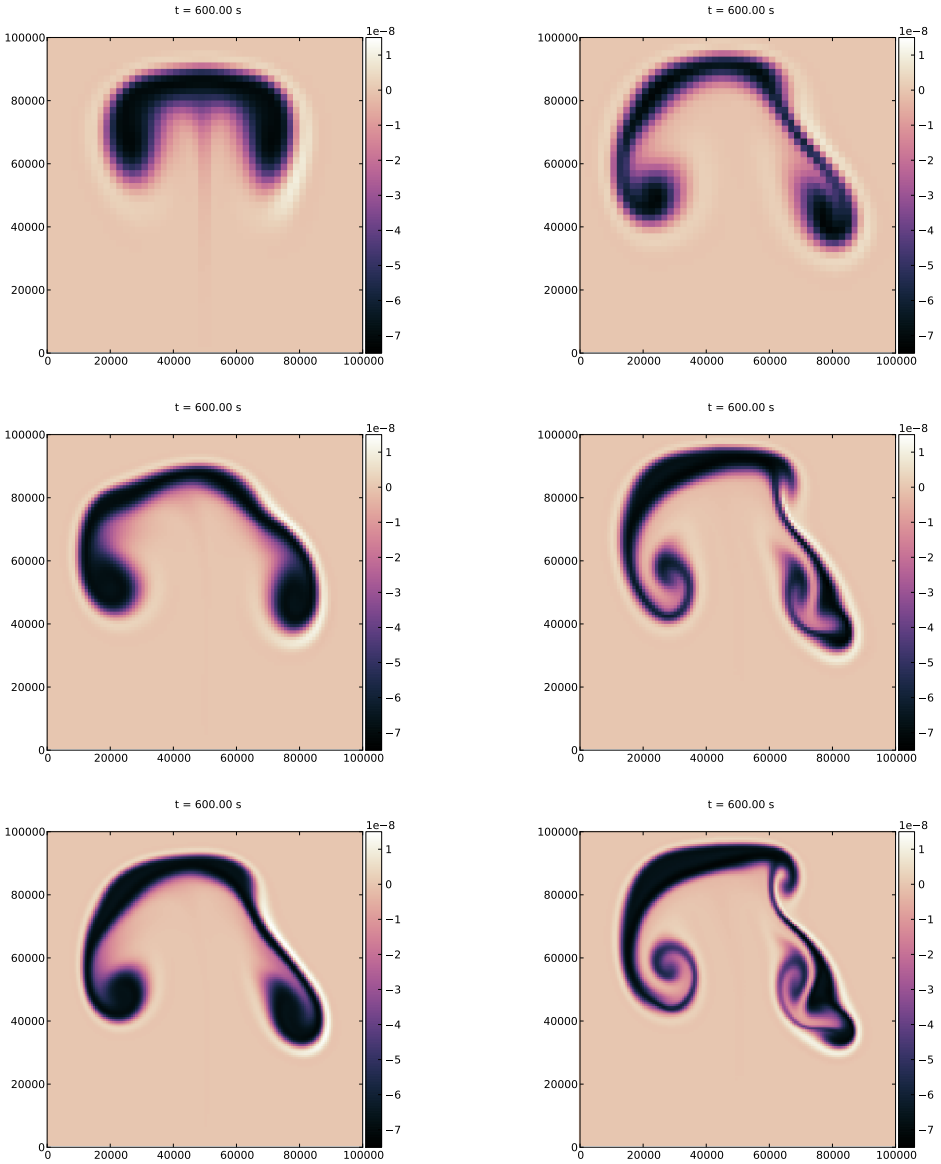


Fig. 6.8: Density fluctuations with respect to the background atmosphere in the hot and cold bubble test case at $t = 600s$. Left: Scheme $S1$. Right: Scheme $S2$. The resolution increases from top to bottom.

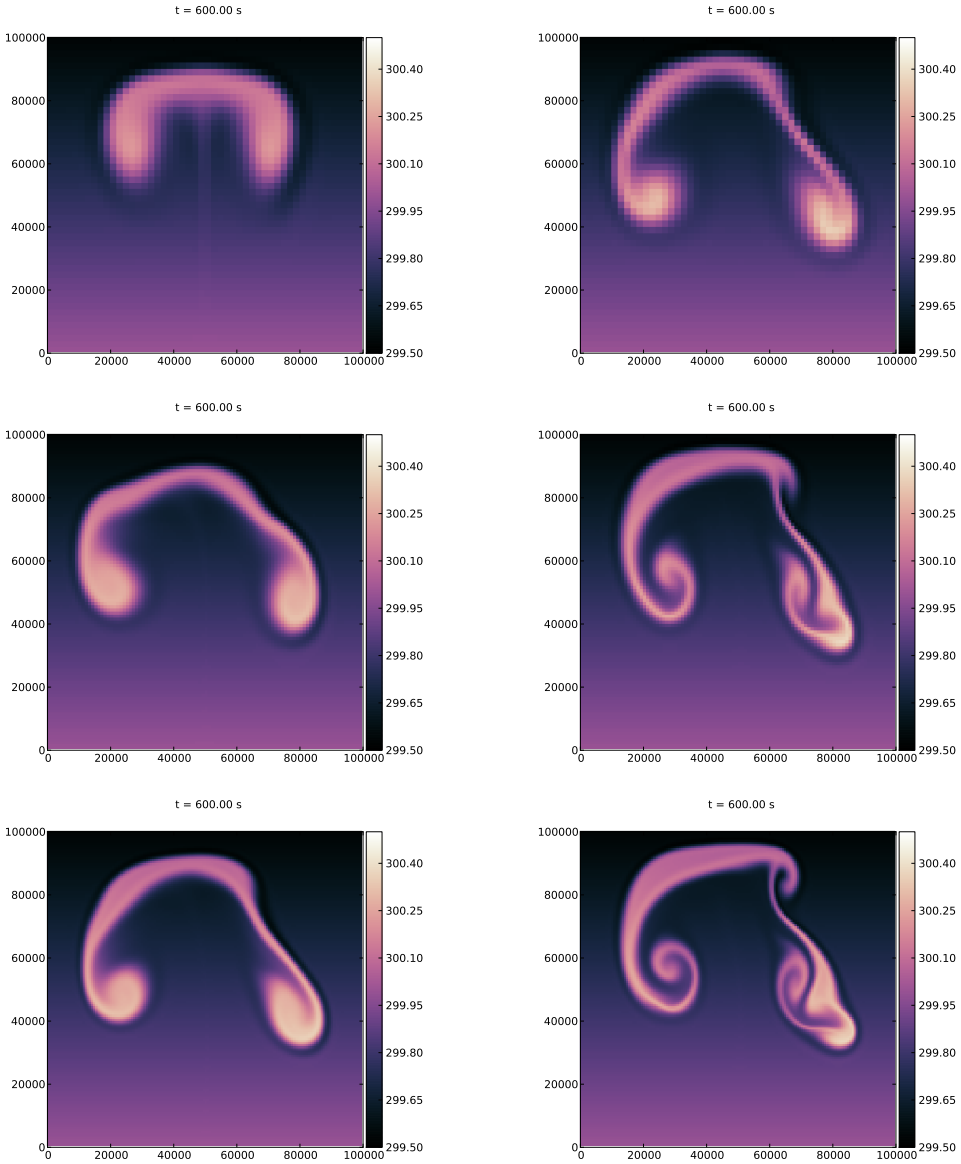


Fig. 6.9: Temperature in the hot and cold bubble test case at $t = 600s$. Left: Scheme $S1$. Right: Scheme $S2$. The resolution increases from top to bottom.

7 Towards a Multidimensional Relaxation Scheme

This chapter is concerned with a genuinely multidimensional approach for defining the numerical fluxes. As has been pointed out in chapter 5, the excessive diffusivity of standard upwind schemes only occurs when multidimensional flows are considered. Therefore the idea is to design an upwind scheme that incorporates the multidimensionality of the problem from the start.

Many different numerical schemes have been proposed that incooperate the upwind mechanism also at the cell vertices. For some examples of publications concerning multidimensional numerical fluxes see [133],[134],[135],[5],[124],[7],[8],[96]. However, the here proposed Ansatz is not completed and there are some issues which have to be tackled in order to make the it ready for numerical tests. In the end, this chapter is considered as a possibility for the author to contribute some thoughts on the problem which may lead to further research in the future.

When dealing with a multidimensional cartesian mesh, it is suggested in section 2.5 to consider the one dimensional fluxes across the boundaries of the cell to determine the numerical fluxes. The idea is that the one dimensional fluxes can be computed by solutions to the Riemann problem. However, doing so is neglecting that in the corners of the cells, there may be a complex contribution due to the influence of 4 different states instead of two. This leads to the consideration of the 2-dimensional Riemann problems at the corners of the cell, see figure 7.1.

Finding the solution to the two dimensional Riemann problem is very hard. In fact, it is not clear, if a unique entropy solution exists in the class of weak solutions. However, approximate Riemann solvers have been successfully used to define the numerical fluxes in the one-dimensional case. Therefore one might hope to extend those techniques to the multidimensional case. This work uses heavily the Suliciu relaxation. In the following it is thought to extend this special relaxation technique to two space dimensions.

7.1 Standard Suliciu Relaxation for the Full Euler System

First consider the full Euler system in two space dimensions as

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2 + p)_x + (\rho uv)_y = 0, \\ (\rho v)_t + (\rho uv)_x + (\rho v^2 + p)_y = 0, \\ E_t + (u(E+p))_x + (v(E+p))_y = 0. \end{cases} \tag{7.1}$$

Next apply the standard Suliciu relaxation technique and rewrite the system in primitive variables to get

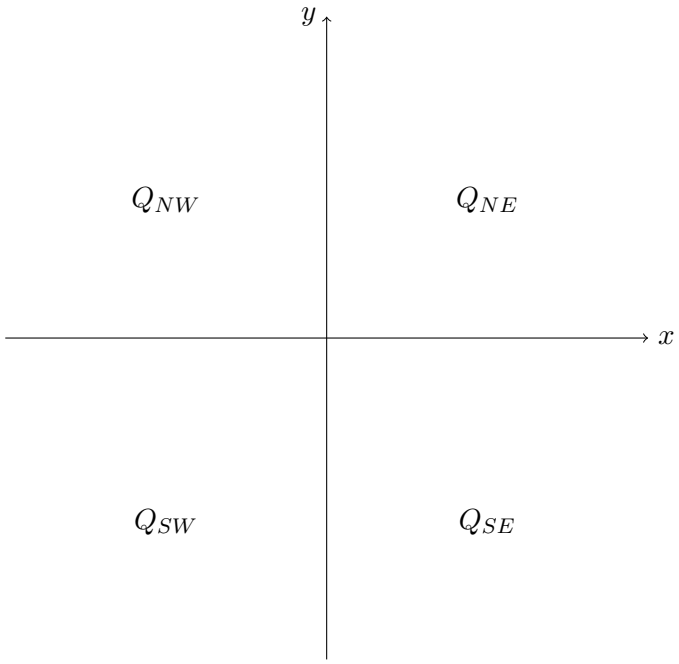


Fig. 7.1: Conceptional drawing of a 4-fielded two dimensional Riemann problem

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2 + \pi)_x + (\rho uv)_y = 0, \\ (\rho v)_t + (\rho uv)_x + (\rho v^2 + \pi)_y = 0, \\ E_t + (u(E+\pi))_x + (v(E+\pi))_y = 0, \\ (\rho\pi) + (\rho u\pi + c^2 u)_x + (\rho v\pi + c^2 v)_y = \frac{\rho}{\epsilon}(p - \pi). \end{cases} \tag{7.2}$$

The system (7.2) can also be put in quasilinear form as

$$Q_t + AQ_x + BQ_y = \frac{1}{\epsilon}R. \tag{7.3}$$

A critical feature of the one dimensional Suliciu relaxation system is that it can be diagonalized. Then, the Riemann problem can be solved by considering the transport of the characteristic variables. However it can be verified, that the matrices A and B in (7.3) do not commute, i.e. $AB \neq BA$. Therefore, both operators can not be diagonalized simultaneously. The Ansatz here is now to find a relaxation system, where the operators commute, i.e. the system can be diagonalized also in the fully two dimensional case.

7.2 Commutative Suliciu Relaxation for the 2-dimensional Euler System

The key idea to derive the extended Suliciu relaxation is to introduce an extended relaxation pressure $\overline{\pi}$ in the form of a tensor as follows

$$\overline{\pi} = \begin{pmatrix} \pi & \psi \\ \psi & \pi \end{pmatrix}, \tag{7.4}$$

such that the equations in conservative form now read

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2 + \pi)_x + (\rho uv + \psi)_y = 0, \\ (\rho v)_t + (\rho uv + \psi)_x + (\rho v^2 + \pi)_y = 0, \\ E_t + (u(E+\pi) + v\psi)_x + (v(E+\pi) + u\psi)_y = 0, \\ (\rho\pi)_t + (\rho u\pi + c^2 u)_x + (\rho v\pi + c^2 v)_y = \frac{\rho}{\epsilon}(p - \pi), \\ (\rho\psi)_t + (\rho u\psi + c^2 v)_x + (\rho v\psi + c^2 u)_y = \frac{\rho}{\epsilon}(D - \psi). \end{cases} \tag{7.5}$$

Therefore the new relaxation variable ψ can be understood as a perturbation of a constant. In specific, the constant D is chosen as zero. The the system (7.5) can also be put in quasilinear from as

$$Q_t + AQ_x + BQ_y = \frac{1}{\epsilon} R. \tag{7.6}$$

A straightforward computation shows, that the matrices A and B now commute. Therefore the homogeneous system part of (7.5) can be put into diagonal form where the characteristic variables read

$$\begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \end{pmatrix} = \begin{pmatrix} e - \frac{\pi^2+\psi^2}{2c^2} \\ \frac{\pi}{c^2} + \frac{1}{\rho} \\ \frac{\pi+\psi}{2c} - \frac{u+v}{2} \\ \frac{\pi-\psi}{2c} + \frac{v-u}{2} \\ \frac{\pi-\psi}{2c} + \frac{u-v}{2} \\ \frac{\pi+\psi}{2c} + \frac{u+v}{2} \end{pmatrix}. \tag{7.7}$$

The diagonalized system now reads

$$\begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \end{pmatrix}_t + \begin{pmatrix} u & 0 & 0 & 0 & 0 & 0 \\ 0 & u & 0 & 0 & 0 & 0 \\ 0 & 0 & u - \frac{c}{\rho} & 0 & 0 & 0 \\ 0 & 0 & 0 & u - \frac{c}{\rho} & 0 & 0 \\ 0 & 0 & 0 & 0 & u + \frac{c}{\rho} & 0 \\ 0 & 0 & 0 & 0 & 0 & u + \frac{c}{\rho} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \end{pmatrix}_x$$
$$+ \begin{pmatrix} v & 0 & 0 & 0 & 0 & 0 \\ 0 & v & 0 & 0 & 0 & 0 \\ 0 & 0 & v - \frac{c}{\rho} & 0 & 0 & 0 \\ 0 & 0 & 0 & v + \frac{c}{\rho} & 0 & 0 \\ 0 & 0 & 0 & 0 & v - \frac{c}{\rho} & 0 \\ 0 & 0 & 0 & 0 & 0 & v + \frac{c}{\rho} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \end{pmatrix}_y = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{7.8}$$

An interesting property of the system (7.5) is that the conservation of the angular momentum is not lost due to the introduction of the new relaxation variable. To show that,

first define the angular momentum $w = u \wedge \mathbf{x}$, where $\mathbf{x} := (x, y)^T$. Compute first the wedge product with the momentum equations in the system (7.1) to get that

$$(\rho w)_t + \nabla \cdot (\rho \mathbf{u} \times \mathbf{u}) + curl(p\mathbf{x}) = 0. \tag{7.9}$$

Now compute the wedge product of $\mathbf{x}$ with the momentum equations in the new multi-d relaxation system (7.5) to get

$$(\rho w)_t + \nabla \cdot (\rho \mathbf{u} \times \mathbf{u}) + curl(p\mathbf{x}) + curl(\psi \mathbf{y}) = 0, \tag{7.10}$$

where $\mathbf{y} := (y, x)^T$. Therefore, also in the new relaxation approach, the evolution equation for the angular momentum can be put into conservative form. Moreover, there is no relaxation source term on the right hand side. Therefore the new relaxation system allows for conservation of the angular momentum.

As for the previous relaxation systems, a Chapman Enskog stability analysis can be performed. After a straightforward computation the momentum equations read

$$\begin{aligned}
(\rho u)_t + (\rho u^2 + p)_x + (\rho uv)_y &= \epsilon(((\frac{c^2}{\rho} - \rho p')u_x)_x + ((\frac{c^2}{\rho} - \rho p')v_y)_x + (\frac{c^2}{\rho}v_x)_y + (\frac{c^2}{\rho}u_y)_x), \\
(\rho v)_t + (\rho uv)_x + (\rho v^2 + p)_y &= \epsilon(((\frac{c^2}{\rho} - \rho p')u_x)_y + ((\frac{c^2}{\rho} - \rho p')v_y)_y + (\frac{c^2}{\rho}v_x)_x + (\frac{c^2}{\rho}u_y)_y),
\end{aligned} \tag{7.11}$$

and for the energy equations it holds that

$$\begin{aligned}
E_t + (u(E+p))_x + (v(E+p))_y = \epsilon((\frac{c^2}{\rho} - \rho p')((\frac{u^2}{2})_x + uv_y)_x + (\frac{c^2}{\rho} - \rho p')((\frac{v^2}{2})_x + vu_x)_y) \\
+ \epsilon(\frac{c^2}{\rho}((\frac{v^2}{2})_x + vu_y)_x + \frac{c^2}{\rho}((\frac{u^2}{2})_y + uv_x)_y).
\end{aligned} \tag{7.12}$$

Therefore the subcharacteristic condition reads

$$c^2 > \rho^2 p' \tag{7.13}$$

One can hope to get a stable numerical scheme when defining the numerical fluxes by the new relaxation system (7.5). However, even though the system is now diagonalizable, the solution to the Riemann problem is far from straightforward. The problem is that the transport equations for the characteristic variables in (7.8) are not decoupled. The Suliciu relaxation system only admits a linear degenerate system and not a linear system. Therefore the solution to the Riemann problem can take a very complex shape. To see this, numerical experiments are performed on the homogeneous part of the new relaxation system (7.5), where for the definition of the numerical fluxes the approach proposed in section 2.5 is used. Various two dimensional Riemann problems are used as initial conditions. Without specifying them here, some solutions are shown in figure 7.2.

From these simulations it can be conjectured that the solution to the two dimensional Rie-

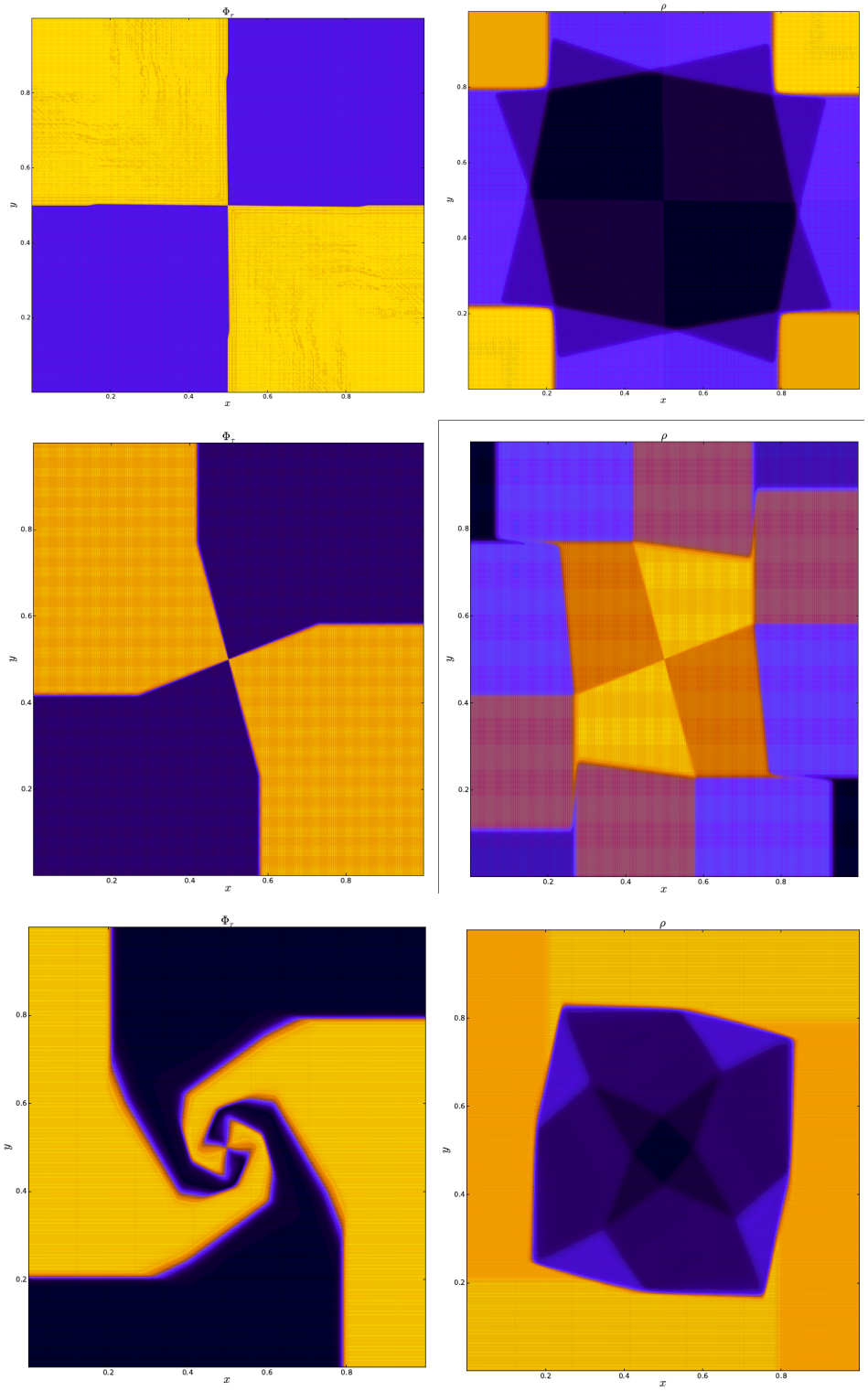


Fig. 7.2: From top to bottom: Different numerical approximations to the solution of the two dimensional Riemann problem for the homogeneous part of system (7.5). Left: Distribution of the characteristic variable $\Phi_\tau = \phi_2$. Right: Distribution of the density ρ to the respective Riemann problem.

mann problem admits a piecewise constant solution. However, the structure of the solution might still be very complicated and varies strongly for different initial conditions. Therefore, up to this point it seems impractical to try to compute these solutions, since computing all these different structures is time consuming. However, a crucial point that gives rise to this complexity is the dependence of the eigenvalues on the characteristic variables. Should this Ansatz be topic of further research, this deficiency has to be resolved in order to arrive at a practical numerical scheme.

8 Conclusion and Outlook

In this work, the numerical approximation of models for atmospheric fluid flows were considered. The focus is on two challenges in the approximations of these models. First, the numerical treatment of source terms due to a varying bottom topography in the Shallow Water case, or due to gravity in the compressible Euler equations. The second challenge is the accurate approximation of low Mach number flows in the compressible Euler equations. The aim is to develop numerical schemes that give accurate approximations of the respective flows even on a coarse mesh.

In chapter 3, a well-balanced scheme based on the HLL approximate Riemann solver approach has been derived. The resulting scheme is able to accurately capture all the one dimensional equilibrium solutions of the Shallow Water equations. It has been shown that a specific challenge in preserving all the equilibria is the case of transcritical flows. Here the waves from the homogeneous part are in resonance with the standing wave coming from the source term. Different models are developed to deal with the different flow regimes. Moreover, it is shown, that these models blend into each other when the type of flow changes. Additionally, a formally second order extension is suggested to enhance the accuracy of the scheme. In the literature, often an iterative algorithm has to be used to compute the second order extension. This is circumvented by solving for the roots directly. Numerical experiments are conducted to show the advantages of this well-balanced approach on coarse grids. However, the scheme is not proven to be robust, i.e. it might give unphysical solutions when wet and dry areas are considered. A stability analysis is also missing for this model. However, for small perturbations of these equilibrias, the scheme seems to perform as expected.

In chapter 4, a well-balanced scheme based on the Suliciu relaxation technique for the hydrostatic equilibrium solutions of the Euler equations with gravity is developed. A major challenge is that the hydrostatic equilibrium equations are underdetermined and therefore do not admit unique solutions. It is shown that the scheme is able to accurately capture certain classes of the hydrostatic equilibria. It is shown that the well-balanced property strongly depends on the choice of a quadrature rule. Here, quadrature formulas for specific classes of equilibria are derived. However, the scheme admits more flexibility since it does not demand a priori a specific description of the equilibrium. In fact, the scheme allows for the use of different quadratures which are not derived in this work to seek for accurate approximations of specific atmospheres. The scheme is further shown to be robust and entropy stable. Numerical experiments are then conducted in order to investigate the properties of the scheme. Here it is decided to put emphasis on the influence of the choice of the quadrature rule. The scheme is tested on different model atmospheres. It is shown that the consistency with a certain class of hydrostatic equilibria is in general of major importance to achieve accurate solutions on coarse grids. However, there might be special equilibria where different quadratures give similar results. It is also shown that the scheme naturally extends to two space dimensions as well as a formally second order extension is presented. In [141] it is shown that the scheme actually can be extended to unstructured meshes as well. Furthermore it

is found there that the numerical scheme also performs better than a fractional splitting method when the flow is far away from equilibrium. This is thought to be an effect of including the source term into the upwinding process. An interesting question is how to tackle the conservation of the total energy throughout the numerical simulation. Moreover, in [141] the scheme is implemented in a moving mesh code. However, for the simulations the mesh has to be held fixed. An extension to a moving mesh can be an interesting challenge.

In chapter 5, a low diffusive scheme based on the Suliciu relaxation technique for the compressible Euler equations is developed. It is shown that the standard upwind scheme is not able accurately capture flows at low Mach numbers. Three different aspects of this behavior are considered, namely the analysis of the scaling of the intermediate states, the analysis of the diffusion and asymptotic preserving properties. It is shown that the adapted relaxation scheme is compatible with the low Mach number behavior in all of these three aspects. Moreover, it is shown that the scheme is robust and stable with respect to a Chapman-Enskog analysis. The stronger entropy stability as for the well-balanced scheme from chapter 4 could not be shown. Even more, the robustness results only apply to time explicit time discretizations. However, for reasons of efficiency only implicit time discretizations are feasible in practical applications. How to transfer the stability properties from the explicit to the implicit time stepping techniques is subject to further research. Numerical tests are performed in order to show the applicability of the new scheme. At first, a shock tube test is considered to investigate how the new relaxation scheme handles discontinuities. It is found that the new relaxation schemes shows a slightly more diffusive behavior at shocks, while the performance at contact discontinuities is comparable with the standard Suliciu relaxation approach, while a slightly less diffusive behavior can be observed at rarefaction waves. Following this, a Gresho vortex test is considered. The new scheme shows good performance on the vortex test. Moreover, it is shown that the diffusion is independent of the Mach number, making the scheme ready to perform simulations on even smaller Mach numbers than considered here. Also a Kelvin-Helmholtz instability is considered, where the new scheme shows a better performance than the standard relaxation scheme.

In chapter 6 the schemes from chapter 4 and chapter 5 are combined in order to develop a scheme to approximate low Mach number flows in a stratified atmosphere. It is shown that all the properties from the previous developed schemes transfer to the hybrid scheme except the entropy stability. Similar to chapter 5 only a stability with respect to a Chapman-Enskog analysis can be shown. Numerical tests are performed to show the practical applicability of the scheme. First, the Gresho vortex for the homogeneous Euler equations is extended to the Euler equations with gravity. Two different versions of this vortex are tested. Similar to chapter 5, the new relaxation scheme shows a less diffusive behavior in the vortex test. However, the diffusion seems to depend on the stratification of the density. It is conjectured that the source quadrature chosen to approximate the source term as suggested in chapter 4 may influences the diffusivity. A further investigation on the choice of the quadrature on the diffusivity in these tests is considered as an interesting subject of future research. Next, two tests are performed that consider the rising and falling of hot and cold bubbles in an isentropic atmosphere. It is shown that the new scheme outperforms the standard relaxation scheme. The difference of the two schemes is even stronger the lower the Mach number in the fluid flow is considered. The disadvantage of these tests is that an explicit solution is not known. For the vortex tests the solutions is known explicitly, since these vortices are time independent. Further research may involve the derivation of time dependent solutions to the Euler equations with gravity to achieve a better justification of the numerical results.

Chapter 7 contains some remarks on an attempt to derive a multidimensional relaxation scheme. Some stability properties are shown as well as a conservation property for the angular momentum. The Ansatz for the new relaxation scheme is that the operators in the different spatial directions commute and the system can be diagonalized. However, it is investigated by some numerical experiments that even though the relaxation scheme admits a piecewise constant solution, the coupled transport of the characteristic variables give a complicated solution structure. For now, it seems impractical to exactly compute the solution in a numerical scheme to define the numerical fluxes. Further research is needed to achieve a simpler solution structure.

Appendix

Appendix A. Analysis of the alternative Relaxation System

Consider the alternative relaxation system proposed in section 5.2.1.

$$\begin{array}{rcccl}
\rho_t & + & (\rho u)_x & = & 0 \\
(\rho u)_t & + & (\rho u^2 + \frac{M_{loc}^2}{M_{ref}^2}\pi + \frac{1-M_{loc}^2}{M_{ref}^2}\frac{\psi_1+\psi_2}{2})_x & = & 0 \\
(\rho v)_t & + & (\rho v u)_x & = & 0 \\
E_t & + & (u(E + M_{loc}^2\pi + (1-M_{loc}^2)\frac{\psi_1+\psi_2}{2})_x & = & 0 \\
(\rho\pi)_t & + & (\rho u\pi + c^2 u)_x & = & \frac{\rho}{\epsilon}(p-\pi) \\
(\rho\psi_1)_t & + & (\rho u\psi_1 + \frac{c^2}{M_{loc}M_{ref}}\psi_1)_x & = & \frac{\rho}{\epsilon}(p + cM_{loc}M_{ref}u - \psi_1) \\
(\rho\psi_2)_t & + & (\rho u\psi_2 - \frac{c^2}{M_{loc}M_{ref}}\psi_2)_x & = & \frac{\rho}{\epsilon}(p - cM_{loc}M_{ref}u - \psi_2)
\end{array} . \tag{A.1}$$

Compute the Chapman Enskog expansion for a stability analysis of system (A.1). From the last three equations of (A.1), there is for π, ψ_1, ψ_2

$$\begin{array}{rcccl}
\pi & = & p & - & \epsilon(\pi_t + u\pi_x + \frac{\alpha^2}{\rho}u_x) \\
\psi_1 & = & p + cM_{loc}M_{ref}u & - & \epsilon(\psi_{1,t} + (u + \frac{c}{\rho M_{loc}M_{ref}})\psi_{1,x}). \\
\psi_2 & = & p - cM_{loc}M_{ref}u & - & \epsilon(\psi_{2,t} + (u - \frac{c}{\rho M_{loc}M_{ref}})\psi_{2,x})
\end{array} \tag{A.2}$$

For the relaxation pressures the following expansions are considered

$$\begin{aligned}
\pi &= \pi_0 + \epsilon\pi_1 + h.o.t., \\
\psi_1 &= \psi_{1,0} + \epsilon\psi_{1,1} + h.o.t., \\
\psi_2 &= \psi_{2,0} + \epsilon\psi_{2,1} + h.o.t.,
\end{aligned} \tag{A.3}$$

where the equilibrium conditions read

$$\begin{aligned}
\pi_0 &= p, \\
\psi_{1,0} &= p + cM_{loc}M_{ref}u, \\
\psi_{2,0} &= p - cM_{loc}M_{ref}u.
\end{aligned} \tag{A.4}$$

Using (A.3) and (A.4) in (A.2), to first oder of the relaxation parameter ϵ there is

$$\begin{array}{rcccl}
\pi & = & p & - & \epsilon(p_t + up_x + \frac{\alpha^2}{\rho}u_x) \\
\psi_1 & = & p + cM_{loc}M_{ref}u & - & \epsilon(p_t + (u + \frac{c}{\rho M_{loc}M_{ref}})p_x + cM_{loc}M_{ref}(u_t + (u + \frac{c}{\rho M_{loc}M_{ref}})u_x)). \\
\psi_2 & = & p - cM_{loc}M_{ref}u & - & \epsilon(p_t + (u - \frac{c}{\rho M_{loc}M_{ref}})p_x - cM_{loc}M_{ref}(u_t + (u - \frac{c}{\rho M_{loc}M_{ref}})u_x))
\end{array} \tag{A.5}$$

Using the conservation of mass and momentum of the original Euler system it can be derived that

$$\begin{aligned} \frac{\partial p}{\partial \rho}\,|_{s=const} &= p', \\ p_t + up_x &= -\rho p' u_x, \\ u_t + uu_x &= -\frac{p_x}{\rho M_{ref}^2}. \end{aligned} \tag{A.6}$$

Using (A.6) in (A.5) further gives

$$\begin{array}{rcccl} \pi & = & p & - & \epsilon((\frac{c^2}{\rho} - \rho p')u_x) \\ \psi_1 & = & p + cM_{loc}M_{ref}u & - & \epsilon((\frac{c^2}{\rho} - \rho p')u_x + p_x \frac{c}{\rho M_{ref}}(\frac{1}{M_{loc}} - M_{loc})). \\ \psi_2 & = & p - cM_{loc}M_{ref}u & - & \epsilon((\frac{c^2}{\rho} - \rho p')u_x - p_x \frac{c}{\rho M_{ref}}(\frac{1}{M_{loc}} - M_{loc})) \end{array} \tag{A.7}$$

From the last two equations of (A.7) there is

$$\frac{\psi_1 + \psi_2}{2} = p - \epsilon((\frac{c^2}{\rho} - \rho p')u_x). \tag{A.8}$$

Using (A.8) and the first equation of (A.7) in the momentum equation of the relaxation system (A.1), it holds that

$$(\rho u)_t + (\rho u^2 + \frac{p}{M_{ref}^2})_x = \epsilon \left(\frac{1}{\rho M_{ref}^2}(c^2 - \rho^2 p')u_x \right)_x. \tag{A.9}$$

Also replacing the relaxation pressures in the energy equation gives

$$E_t + (u(E+p))_x = \epsilon \left(\frac{1}{\rho}(c^2 - \rho^2 p')uu_x \right)_x. \tag{A.10}$$

Therefore the relaxation system (A.1) admits the same subcharacteristic condition as system (5.19), i.e.

$$c^2 > \rho^2 p'. \tag{A.11}$$

Now concern the equivalence of the numerical flux function. It is straightforward to see that the homogeneous part of (A.1) admits the same waves as (5.19) . Moreover, the relaxation variables ψ_1 and ψ_2 satisfy a simple transport equation. Therefore it holds that

$$\begin{aligned} \psi_{1_L} &= \psi_{1_{L^*}} = \psi_{1_{C^L}} = \psi_{1_{C^R}} = \psi_{1_{R^*}} = p_L + cM_{Loc}M_{ref}u_L, \\ \psi_{1_R} &= p_R + cM_{Loc}M_{ref}u_R, \\ \psi_{2_L} &= p_L - cM_{Loc}M_{ref}u_L, \\ \psi_{2_{L^*}} &= \psi_{2_{C^L}} = \psi_{2_{C^R}} = \psi_{2_{R^*}} = \psi_{1_R} = p_R - cM_{Loc}M_{ref}u_R. \end{aligned} \tag{A.12}$$

Consider now the following quantity

$$\bar{\psi} = \frac{\psi_1 + \psi_2}{2}. \tag{A.13}$$

From (A.12) it is straightforward to compute that

$$\begin{aligned} \bar{\psi}_L &= p_L, \\ \bar{\psi}_{L^*} = \bar{\psi}_{C^L} = \bar{\psi}_{C^R} = \bar{\psi}_{R^*} &= \frac{p_L + p_R}{2} + cM_{Loc}M_{ref}(u_L - u_R), \\ \bar{\psi}_R &= p_R. \end{aligned} \tag{A.14}$$

From Theorem 5.2.2 on the intermediate states of the relaxation system (5.19) it can be seen that $\bar{\psi} = \psi$ and the numerical fluxes resulting from both systems are equivalent. Therefore the stability analysis presented here extends to the system (5.19), while in this stability analysis, all the equations take part in deriving the subcharacteristic condition.

Appendix B. Diffusion Matrix of the Suliciu Relaxation Scheme

Here the derivation of the diffusive form for Suliciu Relaxation scheme is concerned. The aim is to represent the interface flux f^* in the following form

$$f^* = \frac{f_i(U_i) + f_{i+1}(U_{i+1})}{2} - D(U_{i+1} - U_i), \tag{B.1}$$

where D is called the diffusion matrix for the numerical scheme. This can be done for the Roe scheme. The aim is to show, why this is more difficult for the Suliciu relaxation scheme.

In the derivation of the diffusive form , first some steps are taken, which are the same for the Roe and the Suliciu relaxation scheme. Assume that the approximate Riemann solver can be written in the form given in (2.30)

$$\mathcal{W}(t,x) = \begin{cases} W_L \text{ if } \frac{x}{t} < \bar{\lambda}_1, \\ W_1 \text{ if } \bar{\lambda}_1 < \frac{x}{t} < \bar{\lambda}_2, \\ \vdots, \\ W_{k-1} \text{ if } \bar{\lambda}_{K-1} < \frac{x}{t} < \bar{\lambda}_K, \\ U_R \text{ if } \bar{\lambda}_K < \frac{x}{t}. \end{cases} \tag{B.2}$$

Both the Roe and the Suliciu relaxation scheme satisfy this form. The idea in the Roe scheme is to linearize the conservation law in the following way

$$u_t + A\,|_{Roe}\, u_x = 0, \tag{B.3}$$

where $A\,|_{Roe} = f'(u\,|_{Roe})$. So A is the flux Jacobian evaluated at a specific value $u\,|_{Roe}$, which depends on the left and right states of the Riemann problem, denoting this by $A = A(u_L, u_R)$. The value $u\,|_{Roe}$ is chosen in such a way that the following properties hold

- Conservation property: $A(u_L, u_R)(u_R - u_L) = f(u_R) - f(u_L)$
- Consistency: $A(u,u) = f(u)$
- $A(u_L, u_R)$ is hyperbolic

and the Riemann problem is then solved exactly for the modified system (B.3). The similar property holds true for the relaxation procedure, where a modified system is proposed for which the Riemann problem can be solved exactly. So in both cases, the discontinuities in the model (B.2) satisfy the Rankine Hugoniot condition and it holds that

$$\lambda_k(u_k - u_{k-1}) = f(u_k) - f(u_{k-1}). \tag{B.4}$$

Summing up the Rankine Hugoniout relations once from the left and from the right up to the interface value gives the following formulas

$$\begin{aligned} f^* &= f(u_L) + \sum_{\lambda_k<0} \lambda_i(u_k - u_{k-1}), \\ f^* &= f(u_R) - \sum_{\lambda_k>0} \lambda_i(u_k - u_{k-1}). \end{aligned} \tag{B.5}$$

Summing and averaging these two equations gives the following form

$$f^* = \frac{f(u_L) + f(u_R)}{2} - \frac{1}{2} \sum_k \mid \lambda_k \mid (u_k - u_{k-1}). \tag{B.6}$$

From here on, the derivations for the Roe scheme and the Suliciu relaxation are different.

Roe Scheme

In the case of a Roe scheme, the system under consideration is linear and one can express the state differences on the right side as a scalar multiple of the corresponding eigenvectors. With r_k being the respective eigenvector, there is

$$u_k - u_{k-1} = \alpha_k r_k. \tag{B.7}$$

Inserting (B.7) in (B.6) gives

$$f^* = \frac{f(u_L) + f(u_R)}{2} - \frac{1}{2} \sum_k \mid \lambda_k \mid \alpha_k r_k. \tag{B.8}$$

Since the matrix A is hyperbolic, it admits a full set of eigenvectors and (B.8) can be expanded by cleverly multiplying with an identity to get

$$f^* = \frac{f(u_L) + f(u_R)}{2} - \frac{1}{2} \sum_k R \Lambda R^{-1} \alpha_k r_k, \tag{B.9}$$

where R is the matrix with all the eigenvectors of A and Λ is a diagonal matrix with the absolute values of the eigenvalues. Using (B.7) again, it is clear that the sum telescopes and the interface flux can be put into the form (B.1) as

$$f^* = \frac{f(u_L) + f(u_R)}{2} - \frac{1}{2} R \Lambda R^{-1} (u_R - u_L). \tag{B.10}$$

Suliciu Relaxation

Since the Suliciu relaxation gives not for a linear, but a linear degenerate system, expressing the jump at the discontinuities by some multiple of an eigenvector as in (B.7) is not necessarily true. Recall that the shock curves needed to determine the intermediate states are tangent to the eigenvectors. However, in the linear case the eigenvectors do not depend on the solution, resulting in linear shock curves in phase space. In contrast, in the case of the Suliciu relaxation, one first has to show, if the respective eigenvector is in fact constant along a shock curve. Computing the eigenvectors to the flux Jacobian of the Suliciu relaxation gives that

$$\lambda = u \pm \frac{c}{\rho} \qquad \text{the eigenvectors read} \qquad \begin{pmatrix} 1 \\ u \pm \frac{c}{\rho} \\ \frac{\pi \pm cu + \rho(e + \frac{u^2}{2})}{c^2 + \pi\rho} \\ \pi + \frac{c^2}{\rho} \end{pmatrix}, \tag{B.11}$$

and for

$$\lambda = u \qquad \text{the eigenvectors read} \qquad \begin{pmatrix} 1 \\ u \\ 0 \\ \pi \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}. \tag{B.12}$$

For the eigenvalue u, the vales u and π are actually Riemann invariants. Therefore, for this discontinuity, the respective eigenvectors are constant along the shock curve and the shock curve is a linear function in phase space. The jump in the dependent variables can be expressed as a scalar multiple of the eigenvectors evaluated at u_C and π_C.

For discontinuities moving with the eigenvalue $u \pm \frac{c}{\rho}$, the values $u \pm \frac{c}{\rho}$ and $\pi + \frac{c^2}{\rho}$ are Riemann invariants for the respective discontinuities. However, the entries in the energy component $InE := \frac{\pi \pm cu + \rho(e + \frac{u^2}{2})}{c^2 + \pi\rho}$ can not be expressed in terms of Riemann invariants. Therefore the shock curve is in general not linear in phase space. However, lets write the shock curves from the states $u_{R,L}$, denoted as $\eta_{R,L}$, as a parameterized curve in phase space as

$$\eta_{R,L} = \eta(\theta, u_{R,L})_{R,L}. \tag{B.13}$$

Since the shock curve is tangent to its eigenvector and the eigenvector is constant in all but one component it can be rewritten as

$$\eta\theta_{R,L} = U_{R,L} + \begin{pmatrix} \theta \\ \theta(u_{R,L} \pm \frac{c}{\rho_{R,L}}) \\ g(\theta) \\ \theta(\pi_{R,L} + \frac{c^2}{\rho_{R,L}}) \end{pmatrix}. \tag{B.14}$$

Therefore the shock curve is linear in all but one component. In the end, the aim is to express the jump in the dependent variables in terms of some vector. The eigenvectors are a good choice to express these jumps since, if the system is hyperbolic, they span the whole phase space and the matrix R^{-1} in (B.9) is well defined. Since the shock curve is only in one component not linear, the jump might still be expressed in terms of the eigenvector. To see this, rescale the parameter θ such that

$$\eta(0, u_{R,L})_{R,L} = u_{R,L} \qquad \text{and} \qquad \eta(1, u_{R,L})_{R,L} = u_{C^R, C^L}. \tag{B.15}$$

With this the following relations hold

$$u_R - u_{C^R} = \eta(0, u_R)_R - \eta(1, u_R)_R \qquad \text{and} \qquad u_L - u_{C^L} = \eta(0, u_L)_L - \eta(1, u_L)_L. \tag{B.16}$$

The jump in the dependent variables is now expressed as a difference of a vector valued function. If η would be a scalar function, the mean value theorem would connect these differences to the derivative of η with respect to θ. The derivative in turn is a scalar multiple of the respective eigenvector. However, all but one component of η is linear and the mean value theorem can be applied to the non-linear component to get

$$\begin{aligned} \exists\theta_1 \quad & \eta(0,u_R)_R - \eta(1,u_R)_R = \nabla_\theta \eta(\theta_1,u_R)_R, \\ \exists\theta_2 \quad & \eta(0,u_L)_L - \eta(1,u_L)_L = \nabla_\theta \eta(\theta_2,u_L)_L. \end{aligned} \tag{B.17}$$

Moreover, since the shock curve η is an integral curve of the respective eigenvector, the following relations hold

$$\nabla_\theta \eta(\theta_1,u_R)_R = r_+(\eta(\theta_1,u_R)_R) \quad \text{and} \quad \nabla_\theta \eta(\theta_2,u_L)_L = r_-(\eta(\theta_2,u_L)_L), \tag{B.18}$$

where $r_\pm$ are scalar multiples of the eigenvectors in (B.11). Therefore the jump in the dependent variables can be expressed as

$$u_R - u_{C^R} = \alpha_R \begin{pmatrix} 1 \\ u_R + \frac{c}{\rho_R} \\ InE(\eta(\theta_1,u_R)_R) \\ \pi_R + \frac{c^2}{\rho_R} \end{pmatrix} \quad \text{and} \quad u_L - u_{C^L} = \alpha_L \begin{pmatrix} 1 \\ u_L + \frac{c}{\rho_L} \\ InE(\eta(\theta_2,u_L)_L) \\ \pi_L + \frac{c^2}{\rho_L} \end{pmatrix}. \tag{B.19}$$

This implies that the jumps at all the discontinuities can be expressed by scalar multiples of the respective eigenvectors, where the state on which the eigenvector is evaluated might only be know implicitly. In practice, the values $InE(\eta(\theta_1,u_R)_R)$ and $InE(\eta(\theta_2,u_L)_L)$ can be computed numerically as

$$InE(\eta(\theta_1,u_R)_R) = \frac{InE(u_{C^R}) - InE(u_R)}{E_{C^R} - E_R}. \tag{B.20}$$

After all this, it is shown that the reformulation from (B.6) to (B.8) can also be done in the case of the Suliciu relaxation, i.e. it holds that

$$u_k - u_{k-1} = \alpha_k r_k, \tag{B.21}$$

where the vectors are defined by the above calculations. Therefore, inserting (B.21) in (B.6) gives

$$f^* = \frac{f(u_L) + f(u_R)}{2} - \frac{1}{2} \sum_k \mid \lambda_k \mid \alpha_k r_k. \tag{B.22}$$

Now define the matrix of eigenvectors R as follows

$$R = \begin{pmatrix} 1 & 0 & 1 & 1 \\ u_C & 0 & u_C - \frac{c}{\rho_{C^L}} & u_C + \frac{c}{\rho_{C^R}} \\ 0 & 1 & InE_L & InE_R \\ \pi_C & 0 & \pi_C + \frac{c^2}{\rho_{C^L}} & \pi_C + \frac{c^2}{\rho_{C^R}} \end{pmatrix}, \tag{B.23}$$

which is composed of the respective eigenvectors evaluated at the respective states. In order to mimic the step (B.8) to (B.9), it remains to check if R^{-1} is well-defined. The Suliciu relaxation system is hyperbolic and therefore the eigenvectors are linear independent if evaluated at the same state. However, the eigenvectors in (B.23) are evaluated at different

states and therefore it has to be checked directly if R is invertible.

$$\det(R) = \begin{vmatrix} 1 & 0 & 1 & 1 \\ u_C & 0 & u_C - \frac{c}{\rho_{C^L}} & u_C + \frac{c}{\rho_{C^R}} \\ 0 & 1 & InE_L & InE_R \\ \pi_C & 0 & \pi_C + \frac{c^2}{\rho_{C^L}} & \pi_C + \frac{c^2}{\rho_{C^R}} \end{vmatrix} = \frac{2c^3}{\rho_{C^L}\rho_{C^R}}. \tag{B.24}$$

Therefore if $c, \rho_{C^L}, \rho_{C^R} \neq 0$, R is invertible.

Next define the diagonal matrix of the eigenvalues Λ as

$$\Lambda = \begin{pmatrix} \mid u_C \mid & 0 & 0 & 0 \\ 0 & \mid u_C \mid & 0 & 0 \\ 0 & 0 & \mid u_C - \frac{c}{\rho_{C^L}} \mid & 0 \\ 0 & 0 & 0 & \mid u_C + \frac{c}{\rho_{C^L}} \mid \end{pmatrix}. \tag{B.25}$$

With this, the interface flux can be further rewritten into the form

$$f^* = \frac{f(u_L) + f(u_R)}{2} - \frac{1}{2}\sum_k R\Lambda R^{-1}\alpha_k r_k. \tag{B.26}$$

Using again (B.21) and realizing that the sum telescopes, the interface flux for the Suliciu relaxation can be rewritten as

$$f^* = \frac{f(u_L) + f(u_R)}{2} - \frac{1}{2}R\Lambda R^{-1}(U_R - U_L), \tag{B.27}$$

with the respective definitions of the matrices.

Since the derivation of the diffusive form of the numerical flux involves the use of the mean value theorem, a similar analysis of the diffusion matrix as in the case of the Roe scheme as in [162],[168] and [132] will be very difficult.

Appendix C. Vortices in a Gravitational Field

The aim is to derive stationary vortices in a gravitational field. Consider the Euler equations in polar coordinates with an axisymmetric gravitational potential $\Phi(r)$

$$\begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix}_t + \frac{1}{r}\begin{pmatrix} r\rho u \\ r(\rho u^2 + \frac{p}{M^2}) \\ r\rho uv \\ ru(E+p) \end{pmatrix}_r + \frac{1}{r}\begin{pmatrix} \rho v \\ \rho uv \\ (\rho v^2 + \frac{p}{M^2}) \\ v(E+p) \end{pmatrix}_\phi = \begin{pmatrix} 0 \\ \frac{\rho v^2 + p}{r} - \rho\frac{\Phi_r}{Fr^2} \\ \frac{\rho vu}{-r} \\ -\rho u^2 \frac{M^2}{Fr^2}\Phi_r \end{pmatrix}.$$

The vortices are considered to have the following properties:

- Axisymmetric : $U_\phi = 0$,
- Stationary : $U_t = 0$,
- No flow along the radius : $u = 0$.

Therefore the above equations reduce to the following form

$$\frac{1}{rM^2}(rp)_r = \frac{\rho v^2 + \frac{p}{M^2}}{r} - \rho\frac{\Phi_r}{Fr^2},$$

and after further simplification there is

$$\frac{p_r}{M^2} = \rho\frac{v^2}{r} - \rho\frac{\Phi_r}{Fr^2}. \tag{C.1}$$

In the following, two different ways to integrate (C.1) are suggested. One on top of a fixed density distribution and one on top of a fixed temperature distribution.

Vortices on top of a fixed density distribution

Make the assumption for the density to have a distribution according to an isothermal atmosphere. In the spirit of the original Gresho vortex, the additional pressure gradient to balance the centrifugal force is computed by adjusting the temperature profile accordingly.

In the case of an ideal gas law, an isothermal distribution of the density is given by

$$\rho = exp(-\frac{M^2}{Fr^2}\frac{\Phi(r)}{RT}).$$

Therefore (C.1) can be rewritten as

$$p_r = M^2 exp(-\frac{M^2}{Fr^2}\frac{\Phi(r)}{RT})(\frac{v^2}{r} - \frac{\Phi_r}{Fr^2}). \tag{C.2}$$

Now, make the Ansatz for the pressure distribution and a function $h(r)$, such that

$$p(r) = RT\exp(-\frac{M^2}{Fr^2}\frac{\Phi(r)}{RT}) + M^2\exp(\frac{M^2}{Fr^2}h(r)) + C, \tag{C.3}$$

with the convenient choice for C to be

$$C = -M^2\exp(\frac{M^2}{Fr^2}h(0)). \tag{C.4}$$

Take the derivative of (C.3) to get

$$p(r)_r = \exp(-\frac{M^2}{Fr^2}\frac{\Phi(r)}{RT})(-\frac{M^2}{Fr^2}\Phi(r)_r) + M^2\exp(\frac{M^2}{Fr^2}h(r))\frac{M^2}{Fr^2}h(r)_r. \tag{C.5}$$

Using (C.2) and (C.5) gives then a relation of the angular velocity with respect to the function $h(r)$ as

$$v = \pm\frac{M}{Fr}\sqrt{r\exp(\frac{M^2}{Fr^2}(\frac{\Phi(r)}{RT} + h(r)))h(r)_r}. \tag{C.6}$$

Therefore, the function h has to satisfy the following compatibility condition

$$h_r \geq 0. \tag{C.7}$$

Next the choice of a velocity profile with respect to the function $h(r)$ is discussed. Choose h_r as a piecewise linear continuous function, i.e.

$$h(r)_r = \begin{cases} 0 \text{ if } r \leq r_0, \\ a_{0,1} + a_{1,1}r \text{ if } r_0 \leq r \leq r_1, \\ a_{0,2} + a_{1,2}r \text{ if } r_1 \leq r \leq r_2, \\ 0 \text{ if } r_2 \leq r, \end{cases} \tag{C.8}$$

and determine the parameters by defining

$$h(r_1)_r = \bar{h} \geq 0. \tag{C.9}$$

Therefore there is

$$a_{0,1} = -\frac{r_0\bar{h}}{r_1 - r_0} \qquad a_{1,1} = \frac{\bar{h}}{r_1 - r_0},$$
$$a_{0,2} = -\frac{r_2\bar{h}}{r_1 - r_2} \qquad a_{1,2} = \frac{\bar{h}}{r_1 - r_2}.$$

To compute the density and pressure distribution, the integral of $h(r)$ has to be evaluated. For this the primitive is given by

$$h(r) = a_{0,i}r + \frac{a_{1,i}}{2}r^2 + C, \tag{C.10}$$

and the definite integral is given in the following form

$$h(r) = \begin{cases} 0 \text{ if } r \leq r_0, \\ a_{0,1}r + \frac{a_{1,1}}{2}r^2 - C_1 \text{ if } r_0 \leq r \leq r_1, \\ a_{0,2}r + \frac{a_{1,2}}{2}r^2 - C_1 + C_2 - C_3 \text{ if } r_1 \leq r \leq r_2, \\ C_4 - C_1 + C_2 - C_3 \text{ if } r_2 \leq r, \end{cases} \tag{C.11}$$

where

$$C_1 = a_{0,1}r_0 + \frac{a_{1,1}}{2}r_0^2 \qquad C_2 = a_{0,1}r_1 + \frac{a_{1,1}}{2}r_1^2,$$
$$C_3 = a_{0,2}r_1 + \frac{a_{1,2}}{2}r_1^2 \qquad C_4 = a_{0,2}r_2 + \frac{a_{1,2}}{2}r_2^2.$$

Finally, the above definitions give a non-isothermal atmosphere if $r > r_2$. To cure this, compute the temperature from the above definitions at r_2 and integrate the isothermal atmosphere with respect to this temperature. Therefore define the inner temperature, with which the density profile is integrated up to r_2 by T_1, and the temperature from there on as T_2. The vortex is determined as

$$\rho(r) = \begin{cases} \exp(-\frac{M^2}{Fr^2}\frac{\Phi(r)}{RT_1}) & \text{if} \quad r \le r_2, \\ \exp(-\frac{M^2}{Fr^2}\frac{\Phi(r_2)}{R}(\frac{1}{T_1} - \frac{1}{T_2}))\exp(-\frac{M^2}{Fr^2}\frac{\Phi(r)}{RT_2}) & \text{if} \quad r > r_2. \end{cases} \tag{C.12}$$

$$p(r) = \begin{cases} RT_1\exp(-\frac{M^2}{Fr^2}\frac{\Phi(r)}{RT_1}) + M^2\exp(\frac{M^2}{Fr^2}h(r)) + C & \text{if} \quad r \le r_2, \\ RT_2\rho(r) & \text{if} \quad r > r_2. \end{cases} \tag{C.13}$$

$$v = \frac{M}{Fr}\sqrt{r\exp(\frac{M^2}{Fr^2}(\frac{\Phi(r)}{RT_1} + h(r)))h(r)_r}, \tag{C.14}$$

with $C = -M^2\exp(\frac{M^2}{Fr^2}h(0))$ and $T_2 = \frac{p(r_2)}{R\rho(r_2)}$.

Vortices on top of a fixed temperature distribution

Now make the assumption that p and ρ are proportional all through the vortex, i.e.

$$p = \rho RT, \tag{C.15}$$

where the temperature T is independent of r. Use (C.15) in (C.1) leads to the following differential equation

$$\rho_r = \rho\frac{M^2}{RT}(\frac{v^2}{r} - \frac{\Phi_r}{Fr^2}). \tag{C.16}$$

(C.16) can be solved to take the following form

$$\rho(r) = \exp(\frac{M^2}{RT}F(r)), \tag{C.17}$$

where

$$F'(r) = f(r) = \frac{v^2}{r} - \frac{\Phi_r}{Fr^2}. \tag{C.18}$$

What is left to define is a velocity profile that is easily integrable. The velocity can be chosen independently from the potential Φ, as long as $\lim_{r\to 0}\frac{v(r)^2}{r} \le C$. It is suggested to choose the velocity profile as a piecewise linear continuous function, i.e.

$$v(r) = \begin{cases} 0 \text{ if } r \leq r_0, \\ a_{0,1} + a_{1,1} r \text{ if } r_0 \leq r \leq r_1, \\ a_{0,2} + a_{1,2} r \text{ if } r_1 \leq r \leq r_2, \\ 0 \text{ if } r_2 \leq r. \end{cases} \tag{C.19}$$

The parameters can be determined by defining

$$v(r_1) = \bar{v}, \tag{C.20}$$

and therefore there is

$$\begin{aligned} a_{0,1} &= -\frac{r_0 \bar{v}}{r_1 - r_0} \qquad a_{1,1} = \frac{\bar{v}}{r_1 - r_0}, \\ a_{0,2} &= -\frac{r_2 \bar{v}}{r_1 - r_2} \qquad a_{1,2} = \frac{\bar{v}}{r_1 - r_2}. \end{aligned} \tag{C.21}$$

To compute the density and pressure distribution, the integral of $\frac{v^2}{r}$ has to be evaluated. For this first write the piecewise definition

$$\frac{v^2}{r} = \frac{a_{0,i}^2}{r} + 2 a_{0,i} a_{1,i} + a_{1,i}^2 r, \tag{C.22}$$

and the primitive is therefore given by

$$\int \frac{v^2}{r} = a_{0,i}^2 \log r + 2 a_{0,i} a_{1,i} r + \frac{a_{1,i}^2}{2} r^2. \tag{C.23}$$

In total there is

$$\int_0^r \frac{v^2}{2} dr = \begin{cases} 0 \text{ if } r \leq r_0, \\ a_{0,1}^2 \log r + 2 a_{0,1} a_{1,1} r + \frac{a_{1,1}^2}{2} r^2 - C_1 \text{ if } r_0 \leq r \leq r_1, \\ a_{0,2}^2 \log r + 2 a_{0,2} a_{1,2} r + \frac{a_{1,2}^2}{2} r^2 - C_1 + C_2 - C_3 \text{ if } r_1 \leq r \leq r_2, \\ C_4 - C_1 + C_2 - C_3 \text{ if } r_2 \leq r, \end{cases} \tag{C.24}$$

where the constants of integration are given by

$$\begin{aligned} C_1 &= a_{0,1}^2 \log r_0 + 2 a_{0,1} a_{1,1} r_0 + \frac{a_{1,1}^2}{2} r_0^2, \\ C_2 &= a_{0,1}^2 \log r_1 + 2 a_{0,1} a_{1,1} r_1 + \frac{a_{1,1}^2}{2} r_1^2, \\ C_3 &= a_{0,2}^2 \log r_1 + 2 a_{0,2} a_{1,2} r_1 + \frac{a_{1,2}^2}{2} r_1^2, \\ C_4 &= a_{0,2}^2 \log r_1 + 2 a_{0,2} a_{1,2} r_2 + \frac{a_{1,2}^2}{2} r_2^2. \end{aligned}$$

List of Figures

Bibliography

1 Abgrall, R. ; Karni, S.: A comment on the computation of non-conservative products. In: *Journal of Computational Physics* 229 (2010), p. 2759–2763

2 Amadori, D. ; Gosse, L. ; Guerra, G.: Global BV entropy solutions and uniqueness for hyperbolic systems of balance laws. In: *Arch Ration. Mech. Anal.* 162 (2002), p. 327–366

3 Amadori, D. ; Gosse, L. ; Guerra, G.: Godunov-type approximation for a general resonant balance law with large data. In: *J. Diff. Eq* 198 (2004), p. 233–274

4 Ambroso, A. ; Chalons, C. ; Coquel, F. ; Galié, T.: Relaxation and numerical approximation of a two-fluid two-pressure diphasic model. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 43 (2009), Nr. 06, p. 1063–1097

5 Arun, K. R. ; Lukacova-Medvidova, M.: A Characteristics based genuinely multidimensional discrete kinetic scheme for the Euler equations. In: *J. Sci. Comp.* 55 (2013), p. 40–64

6 Audusse, E. ; Bouchut, F. ; Bristeau, M. O. ; Klein, R. ; Perthame, B.: A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. In: *SIAM J. Sci.Comp* 25 (2004), p. 2050–2065

7 Balsara, D. S.: Multidimensional HLLE Riemann solver: Application to Euler and magnetohydrodynamic flows. In: *Journal of Computational Physics* 229 (2010), p. 1970–1993

8 Balsara, D. S.: A two-dimensional HLLC Riemann solver for conservation law: : Application to Euler and magnetohydrodynamic flows. In: *Journal of Computational Physics* 231 (2012), p. 7476–7503

9 Barsukow, W. ; Edelmann, P. ; Klingenberg, C. ; Miczek, F. ; Röpke, F: A numerical scheme for the compressible low-Mach number regime of ideal fluid dynamics. In: *Journal of Scientific Computing* accepted (2017)

10 Baudin, M. ; Berthon, C. ; Coquel, F. ; Masson, R. ; Tran, Q. H.: A relaxation method for two-phase flow models with hydrodynamic closure law. In: *Numerische Mathematik* 99 (2005), Nr. 3, p. 411–440

11 Ben-Artzi, M. ; Falcovitz, J.: A Second Order Godunov-Type Scheme for Compressible Fluid Dynamics. In: *J . Comput. Phys.* 55 (1984), p. 1–32

12 Berberich, J. ; Chandrashekar, P. ; Klingenberg, C.: A general well-balanced finite volume scheme for Euler equations with gravity. In: *to appear in Springer Proceedings in Mathematics and Statistics of the International Conference on Hyperbolic Problems: Theory, Numeric and Applications in Aachen 2016* (2017)

13 Bermudez, A. ; Vazquez, M.E.: Upwind methods for hyperbolic conservation laws with source terms. In: *Computers & Fluids* 23 (1994), Nr. 8, p. 1049–1071

14 Berthon, C.: Stability of the MUSCL schemes for the Euler equations. In: *Comm. Math. Sciences* 3 (2005), p. 133–158

15 Berthon, C.: Numerical approximations of the 10-moment Gaussian closure. In: *Mathematics of Computation* 75 (2006), Nr. 256, p. 1809–1832. – ISSN 0025–5718

16 Berthon, C.: Robustness of MUSCL schemes for 2D unstructered meshes. In: *Journal of Computational Physics* 218 (2006), p. 495–509

17 Berthon, C.: Why the MUSCL-Hancock scheme is L1-stable. In: *Numer. Math.* (2006)

18 Berthon, C. ; Breuss, M. ; Titeux, M.-O.: A relaxation scheme for the approximation of the pressureless Euler equations. In: *Numerical Methods for Partial Differential Equations* 22 (2006), Nr. 2, p. 484–505

19 Berthon, C. ; Chalons, C.: A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations. In: *Math. of Comput.* 85 (2016), p. 1281–1307

20 Berthon, C. ; Coudière, Y. ; Desveaux, V.: Second-order MUSCL schemes based on dual mesh gradient reconstruction (DMGR). In: *Math. Model. Numer. Anal.* 48 (2014), p. 583–602

21 Berthon, C. ; Dubroca, B. ; Sangam, A.: A Local Entropy Minimum Principle for Deriving Entropy Preserving Schemes. In: *SIAM Journal on Numerical Analysis* 50 (2012), Nr. 2, p. 468–491

22 Bispen, G. ; Arun, K.R. ; Lukáčová-Medvid´ová, M. ; Noelle, S.: IMEX large time step finite volume methods for low Froude number shallow water flows. In: *Communications in Computational Physics* 16 (2014), August, Nr. 2, p. 307–347

23 Bouchut, F.: *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources.* Basel : Birkhäuser Verlag, 2004 (Frontiers in Mathematics). – viii+135 S. DOI 10.1007/b93802. `http://dx.doi.org/10.1007/b93802`. – ISBN 3–7643–6665–6

24 Bouchut, F. ; Klingenberg, C. ; Waagan, K.: A multiwave approximate Riemann solver for ideal MHD based on relaxation. I: theoretical framework. In: *Numerische Mathematik* 108 (2007), Nr. 1, p. 7–42

25 Bouchut, F. ; Klingenberg, C. ; Waagan, K.: A multiwave approximate Riemann solver for ideal MHD based on relaxation II: numerical implementation with 3 and 5 waves. In: *Numer. Math.* 115 (2010), Nr. 4, p. 647–679. `http://dx.doi.org/10.1007/s00211-010-0289-4`. – DOI 10.1007/s00211–010–0289–4. – ISSN 0029–599X

26 Bryson, S. ; Epshteyn, Y. ; Kurganov, A. ; Petrova, G.: Well-balanced positivity preserving central-upwind scheme on triangular grids for the Saint-Venant system. In: *ESAIM Math. Model. Numer. Anal.* 45 (2011), p. 423–446

27 Buffard, T. ; ´et, T. G. ; Hérard, J.M.: A naive Godunov scheme to solve shallow water equations. In: *CR Acad. Sci. Paris* 326 (1998), p. 385–390

28 Butcher, J.: *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, 2008

29 Cargo, P. ; Le Roux, A.-Y.: Un schéma équilibre adapté au modèle d'atmosphère avec termes de gravité. In: *Comptes rendus de l'Académie des sciences. Série 1, Mathématique* 318 (1994), Nr. 1, p. 73–76

30 Castro, M. J. ; LeFloch, P. G. ; Munoz-Ruiz, Maria L. ; Pares, Carlos: Why many theories of shock waves are necessary. Convergence error in formally path-consistent schemes. In: *Journal of Computational Physics* 227 (2008), p. 8107–8129

31 Castro, M. J. ; Pardo, A. ; Parés, C.: Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique. In: *Mathematical Models and Methods in Applied Sciences* 17 (2007), p. 2065–2113

32 Castro, M.J. ; Gallardo, J.M. ; López-García, J.A. ; Parés, C.: Well-balanced high order extensions of Godunov's method for semilinear balance laws. In: *SIAM J. Num. Anal.* 46 (2008), p. 1012–1039

33 Cerecignani, C.: *The Boltzmann Equation and Its Applications*. Springer-Verlag, New York, 1988

34 Chadrashekar, P. ; Klingenberg, C.: A second order well-balanced finite volume method for the Euler equations with a gravity. In: *SIAM Journal on Scientific Computing* 37 (2015), Nr. 3, p. 382–402

35 Chadrashekar, P. ; Zenk, M.: Well-balanced nodal discontinuous Galerkin method for Euler equations with gravity. In: *Journal of Scientific Computing* (2017). `http://dx.doi.org/doi:10.1007/s10915-016-0339-x`. – DOI doi:10.1007/s10915–016–0339–x

36 Chalabi, A.: Convergence of relaxation schemes for hyperbolic conservation laws with stiff source terms. In: *Math. Comp* 68 (1999), p. 955–970

37 Chalons, C. ; Coquel, F. ; Godlewski, E. ; Raviart, P.-A. ; Seguin, N.: Godunov-type schemes for hyperbolic systems with parameter-dependent source. The case of Euler system with friction. In: *Math. Models Methods Appl. Sci.* 20 (2010), Nr. 11, p. 2109–2166. `http://dx.doi.org/10.1142/S021820251000488X`. – DOI 10.1142/S021820251000488X. – ISSN 0218–2025

38 Chalons, C. ; Coquel, F. ; Marmignon, C.: Well-balanced time implicit formulation of relaxation schemes for the Euler equations. In: *SIAM Journal on Scientific Computing* 30 (2008), Nr. 1, p. 394–415

39 Chen, G.-Q. ; Levermore, C.-D. ; Liu, T.-P.: Hyperbolic Conservation Laws with stiff Relaxation Terms and Entropy. In: *Pure and Applied Math.* 47 (1994), p. 787–830

40 Chen, G.-Q. ; Liu, T.-P.: Zero relaxation and dissipation limits for hyperbolic conservation laws. In: *Comm. Pure Appl. Math* 46 (1993), p. 755–830

41 Cheng, Y. ; Kurganov, A.: Moving-Water Equilibria Preserving Central-Upwind Schemes for the Shallow Water Equations. In: *Communications in Mathematical Sciences* 14 (2016), p. 1643–1663

42 Chinnaya, A. ; LeRoux, A.Y. ; Seguin, N.: A well-balanced numerical scheme for the approximation of the shallow-water equations with topography: the resonance phenomenon. In: *Int. J. Finite Vol.* 1 (2004), p. 1–33

43 Cockburn, B. ; Karniadakis, G. E. ; Shu, C.-W.: *Discontinuous Galerkin methods. Theory, computation and applications.* Lecture Notes in Computational Science and Engineering,11, Springer-Verlag, Berlin, 2000

44 Colella, P. ; Woodward, P.: The piecewise-parabolic method (PPM) for gas-dynamical simu- lations. In: *J. Comput. Phys.* 54 (1984), p. 174–201

45 Collet, J. F. ; Rascle, M.: Convergence of the relaxation approximation to a scalar non- linear hyperbolic equation arising in chromatography. In: *Angew. Math. Phys.* 47 (1996), p. 400–409

46 Coquel, F. ; Hérard, J.-M. ; Saleh, K. ; Seguin, N.: A robust entropy-satisfying finite volume scheme for the isentropic Baer-Nunziato model. In: *ESAIM Math. Model. Numer. Anal.* 48 (2014), Nr. 1, p. 165–206. `http://dx.doi.org/10.1051/m2an/2013101`. – DOI 10.1051/m2an/2013101. – ISSN 0764–583X

47 Coquel, F. ; Perthame, B.: Relaxation of Energy and Approximate Riemann Solvers for General Pressure Laws in Fluid Dynamics. In: *SIAM Journal on Numerical Analysis* 35 (1998), Nr. 6, p. 2223–2249. – ISSN 0036–1429

48 Courant, R. ; Friedrichs, K. ; Lewy, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. In: *Mathematische Annalen* 100 (1928), p. 32–74

49 Dafermos, C. M.: *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Bd. 325: *Hyperbolic conservation laws in continuum physics*. Third. Berlin : Springer-Verlag, 2010. – xxxvi+708 S. DOI 10.1007/978–3–642–04048–1. `http://dx.doi.org/10.1007/978-3-642-04048-1`. – ISBN 978–3–642–04047–4

50 Dafermos, C.M.: Contemporary Issues in the Dynamic Behaviour of Continous Media. In: *LCDS Lecture Notes* 85 (1985)

51 Degond, P. ; Tang, M.: All Speed Scheme for the Low Mach Number Limit of the Isentropic Euler Equations. In: *Communications in Computational Physics* 10 (2011), Nr. 1, p. 1–31

52 Dellacherie, S.: Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. In: *Journal of Computational Physics* 229 (2010), p. 978–1016

53 Desveaux, V.: *Contribution a l'approximation numerique des systemes hyperboliques*, University of Nantes, PhD thesis, 2014

54 Desveaux, V. ; Zenk, M. ; Berthon, C. ; Klingenberg, C.: A Well-balanced scheme for the Euler equation with a gravitational potential. In: *Finite Volumes for complex Application VII, Springer* (2014), p. 217–226

55 Desveaux, V. ; Zenk, M. ; Berthon, C. ; Klingenberg, C.: Well-balanced schemes to capture non-explicit steady states on the Euler equation with a gravity. In: *International Journal for Numerical Methods in Fluids* 81 (2016), p. 104–127

56 Desveaux, V. ; Zenk, M. ; Berthon, C. ; Klingenberg, C.: Well-balanced schemes to capture non-explicit steady states: Ripa model. In: *Math. of Comput.* 85 (2016), p. 1571–1602

57 DiPerna, R. J.: Global existence of solutions to nonlinear systems of conservation laws. In: *J. Diff. Eq.* 20 (1976), p. 187–212

58 Eck, C. ; Garcke, H. ; Knabner, P.: *Mathematische Modellierung.* Springer, 2008

59 Edelmann, P. V. F.: *Coupling of Nuclear Reaction Networks and Hydrodynamics for Application in Stellar Astrophysics*, Technische Universitat Munchen, PhD thesis, 2013

60 Evans, L. C.: *Partial differential equations.* American Mathematical Society, 1998 (Graduate studies in mathematics)

61 F. Bouchut, T. M.: A subsonic-well-balanced reconstruction scheme for shallow water flows. In: *SIAM Journal on Numerical Analysis* 48 (2010), Nr. 5, p. 1733–1758

62 Gallouet, T. ; Hérard, Jean-Marc ; Hurisse, O. ; LeRoux, A.: Well-balanced schemes versus fractional step method for hyperbolic systems with source terms. In: *CALCOLO* 43 (2006), p. 217–251

63 Glimm, J.: Solutions in the large for nonlinear hyperbolic systems of equations. In: *Comm. Pure Appl Math* 18 (1965), p. 697–715

64 Glimm, J.: The continous structure of discontinuities. In: *Lecture Notes in Physics* 344 (1986), p. 177–186

65 Goatin, P. ; LeFloch, P.G.: The Riemann problem for a class of resonant hyperbolic systems of balance laws. In: *Annales de l'Institut Henri Poincare (C) Non Linear Analysis* 21 (2004), p. 881–902

66 Godlewski, E. ; Raviart, P.-A.: *Applied Mathematical Sciences.* Bd. 118: *Numerical approximation of hyperbolic systems of conservation laws*. New York : Springer-Verlag, 1996

67 Godunov, S. K.: *Different Methods for Shock Waves*, Moscow State University, PhD thesis, 1954

68 Godunov, S.K.: A Difference Scheme for Numerical Solution of Discontinuous Solution of Hydrodynamic Equations. In: *Math. Sbornik* 47 (1959), p. 271–306. – translated US Joint Publ. Res. Service, JPRS 7226, 1969

69 Gosse, L.: A Well-Balanced Flux-Vector Splitting Scheme Designed for Hyperbolic Systems of Conservation Laws with Source Terms. In: *Computers and Mathematics with Applications* 39 (2000), p. 135–159

70 Gosse, L.: A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms. In: *Math. Models Methods Appl. Sci.* 11 (2001), p. 339

71 Greenberg, J. ; LeRoux, A. ; Baraille, R. ; Noussair, A.: Analysis and Approximation of Conservation Laws with Source Terms. In: *SIAM J. Numer. Anal.* 34 (1997), p. 1980–2007

72 Greenberg, J. M. ; Leroux, A.-Y.: A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. In: *SIAM Journal on Numerical Analysis* 33 (1996), Nr. 1, p. 1–16

73 Grossmann, C. ; Roos, H.-G. ; Stynes, M.: *Numerical Treatment of Partial Differential Equations*. Springer Science & Business Media, 2007

74 Guillard, H. ; Viozat, C: On the bahaviour of upwind schemes in the low Mach number limit. In: *Computers & Fluids* 28 (1999), p. 63–.86

75 Haack, J. ; Jin, S. ; Liu, J.: An all-speed asymptotic-preserving method for the isentropic Euler and Navier-Stokes equations. In: *Communications in Computational Physics* 12 (2012), Nr. 4, p. 955–980

76 Harten, A. ; Engquist, B. ; Osher, S. ; Chakravarthy, S.: Uniformly high order accurate essentially nonoscillatory schemes, III. In: *J. Comput. Phys.* 71 (1987), p. 231–303

77 Harten, A. ; Lax, P. D. ; Levermore, C. D. ; Morokoff, W. J.: Convex entropies and hyperbolicity for general Euler equations. In: *SIAM J. Numer. Anal.* 35 (1998), Nr. 6, p. 2117–2127. http://dx.doi.org/10.1137/S0036142997316700. – DOI 10.1137/S0036142997316700. – ISSN 0036–1429

78 Harten, A. ; Lax, P.D. ; Van Leer, B.: On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. In: *SIAM review* 25 (1983), p. 35–61. – ISSN 0036–1445

79 Harten, A. ; Osher, S.: Uniformly high-order accurate nonoscillatory schemes. I. In: *SIAM J. Numer. Anal.* 24 (1987), p. 279–309

80 Hedestrom, G.: Models of difference schemes for $u_t + u_x = 0$ by partial differential equations. In: *Math. Comp.* 29 (1975), p. 969–977

81 Hesthaven, J.S. ; Warburton, T.: *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Springer Texts in Applied Mathematics 54. Springer Verlag, New York, 2008

82 Holden, H. ; Risebro, N. H. ; Antman (Hrsg.) ; Marsden (Hrsg.) ; Sirovich (Hrsg.): *Applied Mathematical Sciences*. Bd. 152: *Front Tracking of Hyperbolic Conservation Laws*. Springer, 2011

83 Holz, M. ; Wille, D.: *Repetitorium der linearen Algebra. Teil 2.* Binomi Verlag, Auflage: 2. (2006)

84 Isaacson, E. ; Temple, B.: Nonlinear resonance in systems of conservation laws. In: *SIAM J. Appl. Math* 52 (1992), p. 1260–1278

85 Isaacson, E. ; Temple, B.: Convergence of a 2x2 Godunov method for a general resonant nonlinear balance law. In: *SIAM J. Appl. Math* 55 (1995), p. 625–640

86 Jin, S.: Efficient Asymptotic-Preserving (AP) Schemes For Some Multiscale Kinetic Equations. In: *SIAM Journal of Scientific Computing* 21 (1999), Nr. 2, p. 441–454

87 Jin, S.: A steady-state capturing method for hyperbolic systems with geometrical source terms. In: *M2AN Math. Model. Numer. Anal.* 35 (2001), p. 631–645

88 Jin, S. ; Xin, Z.: The Relaxation Scheme for Systems of Conservation Laws in Arbitrary Space Dimension. In: *Comm. Pure Appl. Math.* 45 (1995), p. 235–276

89 Käppeli, R. ; Mishra, S.: Well-balanced schemes for the Euler equations with gravitation. In: *Journal of Computational Physics* 259 (2014), Nr. 0, p. 199 – 219. `http://dx.doi.org/http://dx.doi.org/10.1016/j.jcp.2013.11.028`. – DOI http://dx.doi.org/10.1016/j.jcp.2013.11.028. – ISSN 0021–9991

90 Karlsen, K. H. ; Klingenberg, C. ; Risebro, N. H.: A relaxation scheme for conservation laws with a discontinous coefficient. In: *Math. Comp* 73 (2003), p. 1235–1259

91 Katsoulakis, M. A. ; Tzavaras., A. E.: Contractive relaxation systems and the scalar multi- dimensional conservation law. In: *Comm. Partial Differential Equations* 22 (1997), p. 195–233

92 Kennedy, C. A. ; Carpenter, M. H.: Additive Runge-Kutta Schemes for Convection-Diffusion-Reaction Equation. In: *Tech. rep., NASA Technical Memorandum* (2001)

93 Klainerman, S. ; Majda, A.: Compressible and incompressible fluids. In: *Comm. Pure Appl Math* 35 (1982), p. 629–651

94 Klein, R.: Semi-implicit extension of a godunov-type scheme based on low mach number asymptotics I: One-dimensional flow. In: *Journal of Computational Physics* 121 (1995), October, Nr. 2, p. 213–237

95 Klein, R. ; Botta, N. ; Schneider, T. ; Munz, C.D. ; Roller, S. ; Meister, A. ; Sonar, L. Hoffmann T.: Asymptotic adaptive methods for multi-scale problems in fluid mechanics. In: *Journal of Engineering Mathematics* 39 (2001), Nr. 1, p. 261–343

96 Klingenberg, C.: Two dimensional Riemann problems and its applications. In: *Notes on Num. Fluid Mech.* 20 (1988)

97 Klingenberg, C. ; Thomann, A.: On computing compressible Euler equations with gravity. In: *to appear in Springer Proceedings in Mathematics and Statistics of the International Conference on Hyperbolic Problems: Theory, Numeric and Applications in Aachen 2016* (2017)

98 Kreizer, J.: *Statistical Thermodynamics of Nonequilibrium Processes.* Springer-Verlag, New York, 1987

99 Krushkov, S. N.: First Order quasilinear equations with several space variables. In: *Math. USSR. Sb.* 10 (1970), p. 217–243

100 Kurganov, A. ; Tadmor., E.: Stiff systems of hyperbolic conservation laws: convergence and error estimates. In: *SIAM J. Math. Anal.* 28 (1997), p. 1446–1456

101 Lambert, J.D.: *Numerical Methods for Ordinary Differential Systems: The Initial Value Problem.* John Wiley & Sons, Inc, 1991

102 Landau, L. D. ; Lifshitz, E.M.: *Statistical Physics, Course of Theoretical Physics, Volume 5.* ELSEVIER, 1980

103 Lattanzio, C. ; Marcati, P.: The zero relaxation limit for 2x2 hyperbolic systems. In: *Nonlinear Anal* 38 (1999), p. 375–389

104 Lax, P.D.: Hyperbolic systems of conservation laws, II. In: *Comm. Pure Appl Math.* 10 (1957), p. 537–566

105 Lax, P.D.: Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves. In: *SIAM Regional Conference Series in Applied Mathematics* 11 (1972)

106 Lax, P.D. ; Wendroff, B.: Systems of conservation laws. In: *Comm. Pure Appl Math* 13 (1960), p. 217–237

107 Leer, B. van: Towards the Ultimate Conservative Difference Scheme 111. Upstream-Centered Finite Difference Schemes for Ideal Compressible Flow. In: *J .Comput. Phys.* 23 (1977), p. 263–275

108 Leer, B. van: Towards the Ultimate Conservative Difference Scheme IV. A New Approach to Numerical Convection. In: *J . Comput. Phys* 23 (1977), p. 276–299

109 Leer, B. van: Towards the Ultimate Conservative Difference Scheme V. A Second Order Sequel to Godunov's Method. In: *J . Comput. Phys* 32 (1979), p. 101–136

110 Leer, B. van: On the Relation Between the Upwind-Differencing Schemes of Godunov, Enquist-Osher and Roe. In: *SIAM J. Sci. Stat. Comput.* 5 (1985), p. 1–20

111 LeFloch, P. G.: *Hyperbolic systems of conservation laws.* Basel : Birkhäuser Verlag, 2002 (Lectures in Mathematics ETH Zürich). – x+294 S. DOI 10.1007/978–3–0348–8150–0. `http://dx.doi.org/10.1007/978-3-0348-8150-0`. – ISBN 3–7643–6687–7. – The theory of classical and nonclassical shock waves

112 LeFloch, P.G. ; Thanh, M.D.: The Riemann problem for fluid flows in a nozzle with discontinuous cross section. In: *Comm. Math. Sci.* 1 (2003), p. 763–797

113 Lellis, C. D. ; Székelyhidi, L.: On Admissibility Criteria for Weak Solutions of the Euler Equations. In: *Arch. Rational Mech. Anal.* 195 (2010), p. 225–260

114 LeVeque, R.: Balancing Source Terms and Flux Gradients in High-Resolution Godunov Methods: The Quasi-Steady Wave-Propagation Algorithm. In: *Journal of Computational Physics* 146 (1998), p. 346–365

115 LeVeque, R. J.: *Finite volume methods for hyperbolic problems.* Cambridge : Cambridge University Press, 2002 (Cambridge Texts in Applied Mathematics). – xx+558 S. DOI 10.1017/CBO9780511791253. http://dx.doi.org/10.1017/CBO9780511791253. – ISBN 0–521–81087–6; 0–521–00924–3

116 LeVeque, R. J. ; Pelanti, M.: A class of approximate Riemann solvers and their relation to relaxation schemes. In: *Journal of Computational Physics* 172 (2001), Nr. 2, p. 572–591

117 LeVeque, R.J.: *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady State and Time Dependent Problems.* Society for Industrial and Applied Mathematics (SIAM), 2007

118 Liang, Q. ; Marche, F.: Numerical resolution of well-balanced shallow water equations with complex source terms. In: *Advances in water resources* 32 (2009), Nr. 6, p. 873–884

119 Liu, H. ; Warnecke., G.: Convergence rates for relaxation schemes approximating conserva- tion laws. In: *SIAM J. Numer. Anal.* 37 (2000), p. 1316–1337

120 Liu, T.-P.: Quasilinear hyperbolic systems. In: *Comm. Math. Phys.* 68 (1979), p. 141–172

121 Liu, T.-P.: Hyperbolic Conservation Laws with Relaxation. In: *Comm. Math. Phys.* 108 (1987), p. 153–175

122 Liu, T.-P.: Nonlinear resonance for quasilinear hyperbolic equations. In: *J. Math. Phys.* 28 (1987), p. 2593–2602

123 Liu, X.-D. ; Osher, S. ; Chan, T.: Weighted essentially non-oscillatory schemes. In: *J. Comput. Phys.* 115 (1994), p. 200–212

124 Lukacova-Medvidova, M. ; Saibertova-Zatocilova, J.: Finite Volume Schemes for Multi-Dimensional Hyperbolic Systems Based on the Use of Bicharacteristics. In: *Applications of Mathematics* 51 (2006), p. 205–228

125 Luna, T. M. ; Diaz, M. J. C. ; Parés, C. ; Nieto, E. F.: On a shallow water model for the simulation of turbidity currents. In: *Commun. Comput. Phys.* 6 (2009), p. 848–882

126 Luo, T. ; Natalini, R. ; Yang, T.: Global BV solutions to a p-system with relaxation. In: *J. Differential Equations* 1 (2000), p. 174–198

127 Mascia, C. ; Terracina, A.: Large-Time Behaviour for conservation laws with source in a bounded domain. In: *Journal of Differential Equations* 159 (1999), p. 485–514

128 Maso, G. D. ; Floch, P.G. L. ; Murat, F.: Definition and weak stability of nonconservative products. In: *J. Math. Pures Appl.* 74 (1995), p. 483–548

129 Meister, A.: *Numerik linearer Gleichungssysteme.* Springer Fachmedien Wiesbaden, 2015

130 Mendez-Nunez, L. R. ; Carroll, J. J.: Application of the MacCormack scheme to atmospheric nonhydrostatic models. In: *Mon. Wea. Rev.* 122 (1994), p. 984–1000

131 Miczek, F.: *Simulation of low Mach number astrophysical flows*, München, Technische Universität München, PhD thesis, 2013

132 Miczek, F. ; Röpke, F. K. ; Edelmann, P. V. F.: New numerical solver for flows at various Mach numbers. In: *Astronomy and Astrophysics* 576 (2015)

133 Mishra, S. ; Tadmor, E.: Constraint preserving schemes using potential-based fluxes. I. Multidimensional transport equations. In: *Communications in Computational Physics* 9 (2010), p. 688–710

134 Mishra, S. ; Tadmor, E.: Constraint preserving schemes using potential-based fluxes. II. Genuinely multi-dimensional schemes for systems of conservation laws. In: *SIAM Journal on Numerical Analysis* 49 (2011), p. 1023–1045

135 Mishra, S. ; Tadmor, E.: Constraint preserving schemes using potential-based fluxes. III. Genuinely multi-dimensional schemes for the MHD equations. In: *Mathematical Modeling and Numerical Analysis* 46 (2012), p. 661–680

136 Murrone, H. Guillard A.: On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes. In: *Computers & Fluids* 33 (2004), p. 655–675

137 Natalini, R.: Convergence to equilibrium for the relaxation approximations of conservation laws. In: *Comm. Pure Appl Math* 8 (1998), p. 795–823

138 Natalini, R. ; Terracina, A.: Convergence of a relaxation approximation to a boundary value problem for conservation laws. In: *Comm. Partial Differential Equations* 26 (2001), p. 1235–1252

139 Noelle, S. ; Bispen, G. ; Arun, K.R. ; Lukáčová-Medvid´ová, M. ; Munz, C.D.: An Asymptotic Preserving all Mach Number Scheme for the Euler Equations of Gas Dynamics. In: *SIAM Journal of Scientific Computing* 36 (2014), Nr. 6, p. 989–1024

140 Noelle, S. ; Pankratz, N. ; Puppo, G. ; Natvig, J.R.: Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. In: *J. Comp. Phys.* 474-499 (2006), p. 213

141 Ohlmann, S.: *Hydrodynamics of the Common Envelope Phase in Binary Stellar Evolution*, Universität Heidelberg, PhD thesis, 2016

142 Pares, C.: Numerical methods for nonconservative hyperbolic systems: a theoretical framework. In: *SIAM J. Numer. Anal.* 44 (2006), p. 300–321

143 Pares, C. ; Castro, M.: On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water system. In: *Math. Model. Numer. Anal.* 38 (2004), p. 821–852

144 Rabie, R.L. ; Fowles, G.R. ; Fickett, W.: The polyymorphic Detonation. In: *Physics of Fluids* 22 (1979), p. 422–435

145 Renardy, M. ; Hrusa, W. ; Nohel, J.: Mathematical problems in Viscoelasity. In: *Pitman Monographs and Surveys in Pure and Applied Mathematics* 35 (1987)

146 Robert, A.: Bubble convection experiments with a semi-implicit formulation of the Euler equations. In: *J. Atmos. Sci.* 50 (1993), p. 1865–1873

147 Roe, P.L.: Approximate Riemann solvers, parameter vectors, and difference schemes. In: *J. Comput. Phys.* 43 (1981), p. 357–372

148 Roe, P.L.: Characteristic-based schemes for the Euler equations. In: *Annu. Rev. Fluid Mech.* 18 (1986), p. 337–365

149 Russo, G. ; Khe, A.: High order well-balanced schemes based on numerical reconstruction of the equilibrium variables. In: *Proceedings "WASCOM 2009" 15th Conference on Waves and Stability in Continuous Media* (2010). – 230241, World Sci. Publ., Hackensack, NJ,

150 Saint-Venant, B.: Theorie du mouvement non permanent des eaux, avec application aux crues des rivieres er a l'introduction des marees dans leurs lits. In: *Comptes rendus de Seances de l'Academie des Sciences* 73 (1871), p. 237–240

151 Shu, C.-W.: Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In: *ICASE Report No. 97-65, NASA/CR-97-206253, NASA Langley Research Center* (1997)

152 Smith, G.D.: *Numerical solution of partial differential equations: finite difference methods.* Oxford University Press., 1985

153 Smoller, J.: On the solution of the Riemann problem with general step data for an extended class of hyperbolic systems. In: *Mich. Math. J.* 16 (1969), p. 201–210

154 Smoller, J.: *Shock Waves and Reaction-Diffusion Equations.* Springer, 1983

155 Sod, G.A.: A Survey of Several Finite Difference Methods for Systems of Nonlinear Hyperbolic Conservation Laws. In: *J. Comp. Phys.* 27 (1978), p. 1–31

156 Suliciu, I.: On modelling phase transitions by means of rate-type constitutive equations, shock wave structure. In: *International Journal of Engineering Science* 28 (1990), p. 829–841

157 Suliciu, I.: Some stability-instability problems in phase transitions modelled by piecewise linear elastic or viscoelastic constitutive equations. In: *International Journal of Engineering Science* 30 (1992), p. 483–494

158 Tartar, L.: Compensated compactness and applications to partial differential equations. In: *Nonlinear analysis and mechanics: Heriot-Watt symposium (R.J. Knops, eds.), Research notes in mathematics* 4 (1979), p. 136–212

159 Temple, J.B.: *Solutions in the Large of Some Nonlinear Hyperbolic Conservation Laws of Gas Dynamics*, Univ. of Michigan, PhD thesis, 1980

160 Timmes, F. X. ; Swesty, F. D.: The accuracy, consistency, and speed of an electron-positron equation of state based on table interpolation of the Helmholtz free energy. In: *Astrophysical Journal, Supplement Series* 126 (2000), p. 501–516

161 Toro, E. F.: *Riemann solvers and numerical methods for fluid dynamics.* Third. Berlin : Springer-Verlag, 2009 DOI 10.1007/b79761. http://dx.doi.org/10.1007/b79761. – A practical introduction

162 Turkel, E.: Preconditioned Methods for Solving the Incompressible and Low Speed Compressible Equations. In: *Journal of Computational Physics* 72 (1987)

163 Vasseur, A.: Well-posedness of scalar conservation laws with singular sources. In: *Methods Appl Anal* 9 (2002), p. 291–312

164 Vides, J. ; Braconnier, B. ; Audit, E. ; Berthon, C. ; Nkonga, B.: A Godunov-Type Solver for the Numerical Approximation of Gravitational Flows. In: *Comm. Comput. Physics* 15 (2014), p. 46–75

165 Villani, C.: A review of mathematical topics in collisional kinetic theory. In: *Handbook of mathematical fluid dynamics* 1 (2002), p. 71–305

166 Vincenti, W.G. ; Kruger, C. H.: *Introduction to Physical Gasdynamics.* R.E. Krieger Publication Co., 1982

167 Warming, R. ; Hyett: The modified equation approach to the stability and accuracy analysis of finite-difference methods. In: *J. Comput. Phys.* 14 (1974), p. 159–179

168 Weiss, J.M. ; Smith, W.A.: Preconditioning applied to variable and constant density flows. In: *AIAA Journal* 33 (1995), Nr. 11, p. 2050–2057

169 Whitham, G. B.: *Linear and nonlinear waves.* Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1974. – xvi+636 S. – Pure and Applied Mathematics

170 Xing, Y.: Exactly well-balanced discontinuous Galerkin methods for the shallow water equations with moving water equilibrium. In: *J. Comput. Phys.* 257 (2014), p. 536–553

171 Xing, Y. ; Shu, C.-W.: High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source term. In: *J. Comput. Phys.* 214 (2006), p. 567–598

172 Xing, Y. ; Shu, C.-W. ; Noelle, S.: On the advantage of well-balanced schemes for moving-water equilibria of the shallow water equations. In: *J. Sci. Comput.* 48 (2011), p. 339–349

173 Xing, Y. ; Zhang, X. ; Shu, C.-W.: Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equation. In: *Adv. Water Resour.* 33 (2010), p. 1476–1493

174 Yong, W. A.: A difference scheme for a stiff system of conservation laws. In: *Proc .Roy. Soc. Edinburgh* 128 (1998), p. 1403–1414

175 Zhou, J.G. ; Causon, D.M. ; Mingham, C.G. ; Ingram, D.M.: The Surface Gradient Method for the Treatment of Source Terms in the Shallow-Water Equations. In: *Journal of Computational Physics* 168 (2001), p. 1–25

Acknowledgments

Even though I am at the center of the development of this thesis, I am more than glad that there were many other people involved that accompanied me during the past years. At first I would like to thank my supervisor Prof. Klingenberg for giving me the opportunity to take part at the scientific process, for introducing me to so many researchers all across the world and having an open ear for my mathematical and sometimes not so mathematical problems. Secondly I thank Prof. Röpke as this work is motivated from a collaboration between him and Prof. Klingenberg. Moreover Prof. Röpke patiently answered a lot of my questions regarding the physical processes involved in the simulations and making me aware of the practical problems of a high performance scheme. Next I thank Prof. Berthon from the University of Nantes. He was invaluable to me when investigating the relaxation schemes. I learned a lot from him about the technical details as well as practical issues of these schemes. Also I thank him for the passion and patience he brought into the collaboration. Closely related to Prof. Berthon is his former PhD student Dr. Desveaux. The work on the well-balanced scheme started for me with his visit to Würzburg. I thank him for being a very calm, open and cooperative person. The times we worked on this topic were always very productive. Next I would like to thank Prof. Chandrashekar from TIFR in Bangalore. Not only during my two months stay in that institute, but also during his multiple visits to Würzburg, he introduced me to many practical issues regarding the derivation and implementation of numerical schemes. Moreover I am glad that he opened to me the chance to contribute to the developed well-balanced Discontinous Galerkin approach. Also I thank him and Dr. Ray for being welcoming hosts in Bangalore. Also I have to thank at this point the DAAD program A New Passage to India, which funded my extended stay in Bangalore. Furthermore I would like to thank Dr. Edelmann from the work group of Prof. Röpke. He was crucial in implementing the developed low Mach number schemes into the SLH code. Moreover, his practical experience with low Mach number schemes in stratified atmospheres were more than inspiring for the derivation of the relaxation scheme. Special thanks also go to Dr. Ohlmann, with whom I gladly shared an office for more than two years. Besides having some good times, he also explained to me a lot of the physics that was suddenly invading my professional life. Moreover, I am grateful that we could extend the well-balanced relaxation scheme to work in the AREPO code. Without him actually implementing the solver, this would have never been done. Furthermore I thank Prof. Kurganov from Tulane University for the joint work on the well-balanced scheme for the Shallow Water equations. Especially by discussing the numerical results with him sharpened my mind for judging the quality of a numerical scheme. Moreover I thank all the members of the work groups of Prof. Klingenberg and Prof. Röpke for creating a comfortable and productive work environment. Special thanks also go to the GRK 1147, Theoretische Astrophysik und Teilchenphysik headed by Prof. Porod for kindly accepting me as a fellow and funding the first part of my working period. The second part was kindly supported by the SPPEXA, SPP 1648 program funded by the DFG.

Also important for me during the last years were my friends and family. Here I would like

to thank Elena for helping me correcting this thesis and for strengthen me both physically and mentally during the writing process. Furthermore I like to thank Carlos, Corni, Elena, Eule, Hannsi, Markus and Phil from Eulenvolk and Malasaners for sharing some nice times in playing music together. At last I like to thank my family Christina, Franz, Marianne and Marlene for always supporting my decisions in life and standing with me especially when things are not looking so nice.

Affidavit

I hereby confirm that my thesis entitled On Numerical Methods for Astrophysical Applications is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Bamberg, June 7, 2018 Markus Zenk

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation Über Numerische Methoden für Astrophysikalische Anwendungen eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Bamberg, 7. Juni 2018 Markus Zenk