JULIUS-MAXIMILIANS-UNIVERSITÄT WÜRZBURG
Fakultät für Mathematik und Informatik

Fluids in Gravitational Fields – Well-Balanced Modifications for Astrophysical Finite-Volume Codes

Dissertation zur Erlangung
des naturwissenschaftlichen Doktorgrades
der Julius-Maximilians-Universität Würzburg

vorgelegt von

Jonas Philipp Berberich
geboren in Wertheim am Main

Würzburg, 2020

Eingereicht am 05.10.2020

Abstract

Stellar structure can – in good approximation – be described as a *hydrostatic state*, which which arises due to a balance between gravitational force and pressure gradient. Hydrostatic states are static solutions of the full compressible Euler system with gravitational source term, which can be used to model the stellar interior. In order to carry out simulations of dynamical processes occurring in stars, it is vital for the numerical method to accurately maintain the hydrostatic state over a long time period. In this thesis we present different methods to modify astrophysical finite volume codes in order to make them *well-balanced*, preventing them from introducing significant discretization errors close to hydrostatic states. Our well-balanced modifications are constructed so that they can meet the requirements for methods applied in the astrophysical context: They can well-balance arbitrary hydrostatic states with any equation of state that is applied to model thermodynamical relations and they are simple to implement in existing astrophysical finite volume codes. One of our well-balanced modifications follows given solutions exactly and can be applied on any grid geometry. The other methods we introduce, which do no require any a priori knowledge, balance local high order approximations of arbitrary hydrostatic states on a Cartesian grid. All of our modifications allow for high order accuracy of the method. The improved accuracy close to hydrostatic states is verified in various numerical experiments.

Zusammenfassung

Die Struktur von Sternen kann in guter Näherung als hydrostatischer Zustand beschrieben werden, der durch ein Gleichgewicht zwischen Gravitationskraft und Druckgradient gegeben ist. Hydrostatische Zustände sind statische Lösungen der vollständigen komprimierbaren Euler-Gleichungen mit Gravitationsquellenterm, die zur Modellierung des Sterninneren verwendet werden können. Um Simulationen dynamischer Prozesse in Sternen durchführen zu können, ist es wichtig, dass die verwendete numerische Methode den hydrostatischen Zustand über einen langen Zeitraum genau aufrechterhalten kann. In dieser Arbeit stellen wir verschiedene Methoden vor, um astrophysikalische Finite-Volumen-Codes so zu modifizieren, dass sie die *well-balancing*-Eigenschaft erhalten, d.h., dass sie keine signifikanten Diskretisierungsfehler nahe hydrostatischer Zustände verursachen. Unsere well-balancing-Modifikationen sind so konstruiert, dass sie die Anforderungen für Methoden erfüllen, die im astrophysikalischen Kontext angewendet werden: Sie können beliebige hydrostatische Zustände mit jeder Zustandsgleichung, die zur Modellierung der thermodynamischen Beziehungen angewendet wird, balancieren und sind einfach in vorhandene astrophysikalische Finite-Volumen-Codes zu implementieren. Eine unserer well-balancing Modifikationen erhält bekannte Lösungen exakt und kann auf jede Gittergeometrie angewendet werden. Die anderen Methoden, für die keine A-priori-Kenntnisse erforderlich sind, balancieren lokale Approximationen beliebiger hydrostatischer Zustände mit hoher Fehlerordnung auf einem kartesischen Gitter. Alle unsere Modifikationen erlauben eine hohe Fehlerordnung der Methode. Die verbesserte Genauigkeit nahe an hydrostatischen Zuständen wird in verschiedenen numerischen Experimenten verifiziert.

Acknowledgements

This thesis is the result of about ten years of studying Maths and Physics at Würzburg University. I can happily say that I enjoyed all of this time as a time of academic learning, personal development, and great experiences. Many different people played important roles in that progress: First of all, I want to thank my advisor Prof. Dr. Christian Klingenberg for leading and supporting me in the last years and, jointly with Prof. Dr. Friedrich Röpke, for granting me the chance to go for a PhD.[1] In this whole time I also received continuous support from Prof. Praveen Chandrashekar.

I thank all my dear colleagues for the great discussions and experiences, including Roger, Jens, Markus, Marlies, Wasilij, Simon, Claudius, Farah, Jayesh, Andrea, Marc, and Sandra from the math-side such as Philipp, Sebastian, Alejandro, Kai, Aron, Leo, Flo, Robert, Hans, and Giovanni from the astrophysics-side.

I am grateful for all the personal support from my parents, my brothers (especially Lukas, who was also a colleague), and my wonderful wife and daughters.

Many people supported me on my way and I thank all of them, whether or whether not I mentioned them explicitly above.

[1] I formally acknowledge the financial support by the University of Würzburg and the Klaus Tschira foundation.

Contents

Chapter 1

Introduction

Hyperbolic partial differential equations Partial differential equations (PDEs) and systems of PDEs are used to model static and dynamical systems in various fields such as science (especially physics), economics, and finance. Their solutions describe the behavior of the system depending on boundary and/or initial conditions. Theory about PDEs is hence a modern area of mathematics which is strongly inspired by real world applications and often has immediate relevance for science and industry.

Time dependent conservative systems in which information is transported with finite velocity[1] are commonly described using hyperbolic equations and systems. Examples are the description of sound waves, advective processes, and traffic flows. Source terms can be added to the system to model the effect of external forces such as given fields of attractive or repulsive forces or other non-conservative effects.

Compressible Euler equations The compressible Euler system is a hyperbolic system that is used to model the density, velocity, and pressure of inviscid compressible fluids. It was introduced by Leonhard Euler in 1757 [59] and is thus amongst the first known system of PDEs that has been written down.[2] Even though the system has been known for a long time, it is still an active area of research in both theoretical and numerical analysis. Recent results regard for example the uniqueness of the system's solutions [1].

Inviscid fluid dynamics are relevant in industrial applications to model fluid flows around or through turbines, wind parks, airfoils, vehicles, and various other objects. In atmospheric physics and astrophysics a gravitational source term is added to the compressible Euler equations so that the atmospheric or stellar structure can be modeled. More physical effects can be coupled to the system via additional source terms or equations or via applying a specific equation of state (EoS) which is used to close the under-determined system.

Numerical methods While mathematical theory helps to provide a fundamental understanding of different aspects such as a solution's existence, uniqueness, or

[1] Physically speaking, the transport of information is always restricted to a finite velocity. However, from the modeling perspective this restriction is neglected in many cases.

[2] The compressible full Euler system, which we consider in this thesis, was introduced later, though.

structure, it can in most cases not provide formulae or analytical tools to actually construct solutions of PDEs for general initial or boundary conditions. For this purpose, numerical methods are developed as instruments to approximate solutions using computers and computing centers. PDEs, which are formulated using partial derivatives of the solution itself, describe continua, whereas computers can only store a finite amount of information and can only conduct a finite number of computations in a given time. Consequently, the problem has to be discretized in some way.

Finite difference approaches are based on discretizing the partial derivative operators such that their evaluation only requires point values of the solution. These point values are then evolved in time. *Finite element* and *smooth particle* methods discretize the matter which is described and evolve these elements in time based on the forces acting on and between them. *Finite volume* (FV) and *Discontinuous Galerkin* (DG) methods discretize space into control volumes and approximate the fluxes of different quantities between them according to the Gauß theorem.[3] These methods are a common choice for the approximation of solutions of hyperbolic systems, since they are by construction conservative and allow for discontinuous solutions thus mirroring two fundamental properties of solutions of hyperbolic systems. The error with which a consistent numerical method approximates solutions of a system of PDEs can be controlled by refining the discretization. However, in the case of long-term simulations in which the system is close to a stationary state, this might not be sufficient and there is the requirement for numerical methods that are free of a discretization error for certain stationary solutions. Structurally interesting and relevant static states can arise in the presence of a source term in particular.

Numerical methods for Euler equations in stellar astrophysical applications In long phases of stellar evolution, the most basic stellar structure can be approximately described as a hydrostatic state, i.e., a static state in which the pressure gradient balances the gravitational force [38]. Dynamical processes such as shell or core convection can then be regarded as – potentially relatively small – deviations from the hydrostatic state. Therefore, numerical methods which are used to simulate processes in the stellar interior are required to have a *well-balanced* property, which means that they are capable of maintaining the hydrostatic state over a long time period. This is especially relevant if slow processes, such as convection at low Mach numbers, are simulated. The time-scale necessary to simulate in order to resolve the dynamics of the flow can become arbitrarily large when the Mach number declines.

Various well-balanced finite volume methods have been developed for compressible Euler equations with gravity source term over the last decades (e.g., [34, 101, 104, 41, 72, 167, 17, 43, 69] and references therein). The majority of them, however, has been constructed to balance certain classes of hydrostatic solutions under the assumption of an ideal gas EoS. For astrophysical simulations, this does not suffice in many cases: The thermodynamical conditions in a star are more extreme than the ones on the earth's surface. This leads to pressure from radiation due to high temperatures, quantum effects due to extreme pressure, and also to relativistic effects due to

[3] For modern arbitrary Lagrangian Eulerian methods based on moving meshes that follow the flow, one could also argue that to some extent this resembles a discretization of matter rather than space.

high microscopic velocities. These effects enter the model via the EoS. An example of an EoS including all of these effects is the Helmholtz EoS (e.g., [160]). Numerically solving this EoS is computationally expensive. Therefore, the thermodynamical relations are often interpolated from a table instead [161]. The hydrostatic profiles, which the simulations are based on, can be obtained in the form of discrete data from a traditionally one-dimensional stellar evolution code and mapped to the multi-dimensional mesh. Well-balanced methods that meet these requirements of stellar astrophysical hydrodynamics simulations have been developed in [92, 94, 15, 78]. In the course of the PhD project presented in this thesis, we developed the Deviation method [14] as an extension to the second order α-β method based on a priori knowledge of the hydrostatic state. The α-β method was introduced in the master thesis [11] and published in [13, 15]. The new method allows for arbitrarily high order accuracy and it can not only balance hydrostatic states but also non-static stationary states (i.e., stationary states with non-vanishing velocity) and even time-dependent solutions provided they are known a priori. Furthermore, the Deviation method can be applied to exactly represent static, non-static stationary, and time-dependent solutions of any hyperbolic system. In the author's understanding while writing this thesis, the Deviation method is the most general well-balanced finite volume method existing for hyperbolic systems assuming a priori knowledge of the balanced solution. As another type of well-balanced method satisfying the above-mentioned requirements we developed the Discretely Well-Balanced and Local Approximation method [12] as high order extension to the discretely well-balanced methods introduced in [92, 94, 78]. The methods [92, 94, 78] are constructed to balance certain classes of hydrostatic states and they balance a second order approximation of the hydrostatic state if the hydrostatic state is not in these classes. The Discretely Well-Balanced method balances a high order approximation to any arbitrary hydrostatic state without restriction. Since it suffers from an increased stencil, the stencil has been localized in the Local Approximation method.

All well-balanced methods introduced in this thesis can be added as modifications to existing finite volume codes with minimal effort. Furthermore, they are flexible in the sense that they can be combined with various numerical flux functions and ODE (ordinary differential equation) solvers for the time evolution. This allow for example the combination with numerical flux functions that are particularly suited for the simulation of low Mach number flows [112, 159, 140, 111, 127, 8, 12] (especially in the combination with an implicit ODE solver), lead to a faithful treatment of the discrete kinetic energy evolution [89, 40, 137, 138, 12], or provably dissipate entropy [142, 40, 137, 138, 12]. A numerical flux function with all of the three aforementioned properties has been developed in the course of this PhD project and published in [12]. It is, however, not discussed in this thesis.

Structure of the thesis The rest of this thesis is structured as follows. In Chapter 2 we discuss some basic theoretical aspects of one-dimensional hyperbolic conservation laws, balance laws, and systems. The compressible Euler equations are introduced and hydrostatic solutions are described. One-dimensional Runge–Kutta finite volume methods are discussed in Chapter 3 starting with the original Godunov method and improving it to a general high order method in the course of the

chapter. In Chapter 4 we introduce our well-balanced methods. First, the concept of hydrostatic reconstruction is discussed. The α-β method is recapitulated and the Deviation method is introduced. Thereafter, the Discretely Well-Balanced and Local Approximation methods are presented. The four well-balanced methods are numerically compared to each other and to a non well-balanced standard method in Section 4.6 to conclude Chapter 4. After discussing the two-dimensional Euler system the RK-FV method is extended to two spatial dimensions in Chapter 5 and the corresponding extensions of our well-balanced methods are shown in Chapter 6. Finally, we summarize and conclude the thesis in Chapter 7.

Chapter 2

Compressible Euler Equations with Gravity Source Term

2.1 One-Dimensional Scalar Conservation Laws

A one-dimensional scalar conservation law can be written in the form

$$\partial_t q(x,t) + \partial_x f(q(x,t)) = 0 \tag{2.1}$$

for the conserved variable $q : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{R}$ and the differentiable flux function $f \in \mathcal{C}^2(\mathbb{R}, \mathbb{R})$.[1] In the following discussion we only consider *convex scalar conservation laws*, which means that we assume the flux f to be a convex ($f''(q) \geq 0$) or concave ($f''(q) \leq 0$) function. For a discussion about the non-convex case the reader is referred to [103]. Equations of this type are relevant for modeling dynamical processes, in which scalar quantities are transported with finite speed. The most basic example is the one-dimensional *advection equation*, which is given by Eq. (2.1) with the flux function $f_{\text{advection}}(q) = cq$, where $c \in \mathbb{R}$ is a constant. This equation describes the transport of q with the constant velocity c. For initial conditions $q(x,0) = q_0(x)$ it is easy to show that the solution at any time $t \geq 0$ is given by $q_0(x - ct)$, i.e., the solution is constant on the *characteristic* $\{(x,t) \in \mathbb{R} \times \mathbb{R}_0^+ \,:\, x = x_0 + ct\}$ for each $x_0 \in \mathbb{R}$. This is visualized in Fig. 2.1.

The *inviscid Burgers' equation* is a simple example for a nonlinear scalar hyperbolic balance law. It is defined by the flux function $f_{\text{Burgers}}(q) = q^2/2$ in Eq. (2.1). The characteristics for the inviscid Burgers' equation get evident when they are rewritten in the *quasi-linear* form

$$\partial_t q(x,t) + f'(q)\partial_x q(x,t) = 0 \tag{2.2}$$

of a scalar conservation law, which is

$$\partial_t q(x,t) + q(x,t)\partial_x q(x,t) = 0 \tag{2.3}$$

[1] Depending on the problem which is modeled, the spatial and temporal domain can be reduced to only a subset of $\mathbb{R}$ or $\mathbb{R}_0^+$ respectively. In that case also boundary conditions have to be provided. Also, the values of q might be restricted to a certain subset of $\mathbb{R}$. However, since this does not add fundamental insight in the following general discussion, we do not consider this case.

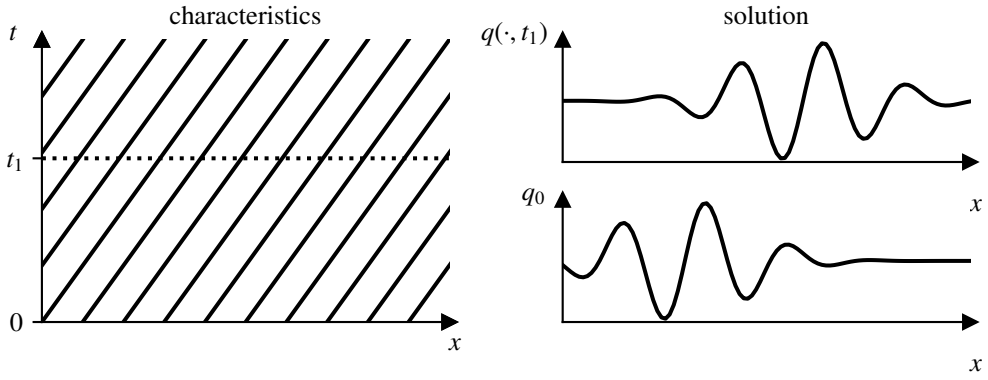


Figure 2.1: Characteristics of the linear advection equations (left panel). The solution (right panels; at two different times) follows the characteristics, which are independent from the solution and have a constant slope. We omit showing values at the axis, since the figures are schematic.

in the case of the inviscid Burgers' equation. The characteristics are thus determined by the condition $x = x_0 + tf'(q_0(x_0))$ for each $x_0 \in \mathbb{R}$. However, different then in the case of the advection equation, this does in general not yield a unique way to identify the value of q at any point (x, t) from the initial condition q_0. There are two situations which require further discussion:

Case (a): There are two or more points x_0 such that $x = x_0 + tf'(q_0(x_0))$ for a given value of $t > 0$, i.e., characteristics are crossing each other. As can be seen in Fig. 2.2, smooth initial data q_0 that are evolved according to Burgers' equation can develop discontinuities at finite time.

The same holds true for other nonlinear scalar conservation laws. From the form of Eqs. (2.1) to (2.3) one might expect that q has to be differentiable in space and time. Differentiable solutions are called *strong solutions*. To further evolve q in time after a jump developed, however, a wider concept of solutions is required. So-called *weak solutions*, which allow for discontinuities, are introduced in Section 2.1.1. In the following, we discuss how to deal with the non-uniqueness of solutions arising from the crossing of characteristics. Basically, the task of choosing a solution boils down to deciding on the velocity s with which a shock travels based on the values q^L and q^R of q directly left and right of the shock's position. Demanding the conservation of q across the shock front yields the *Rankine–Hugoniot* jump condition (e.g., [103])

$$s = \frac{f\left(q^R\right) - f\left(q^L\right)}{q^R - q^L} \tag{2.4}$$

for the scalar conservation law (2.1) (visualized in Fig. 2.2).

Case (b) There is no point x_0 such that $x = x_0 + tf'(q_0(x_0))$. This can happen if there is a jump in q_0 and the characteristics on both sides of the jump move away from each other, i.e., $f'(q^L) < f'(q^R)$ (see left panel of Fig. 2.3). The issue of non-uniqueness also arises in this case: To construct a solution which is defined in $\mathbb{R} \times \mathbb{R}_0^+$ one can define the missing characteristics and the values carried on them in various

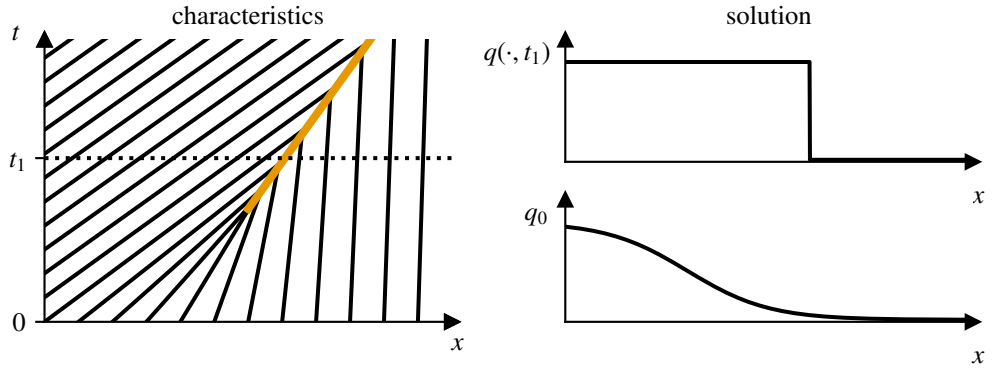


Figure 2.2: Characteristics (left panel) for specific initial conditions q_0 (bottom right panel) for the inviscid Burgers' equation, corresponding to case (a) in the text. The slope of the characteristics depends on the values of the solution q. Due to crossing characteristics (left panel), a discontinuity (top right panel) develops at finite time. The discontinuity travels with the shock speed given by the Rankine–Hugoniot condition (2.4) (visualized as yellow line in the left panel). We omit showing values at the axis, since the figures are schematic.

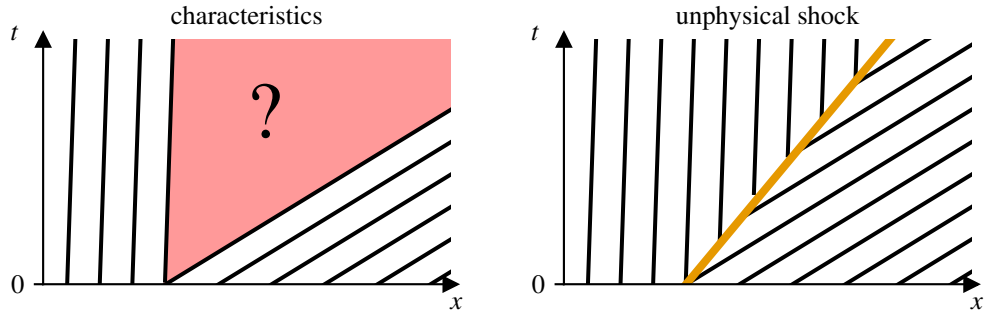


Figure 2.3: Characteristics corresponding to case (b) in the text. There is an area in the (x, t)-plane, for which there are no characteristics given by the initial condition (left panel). Setting the missing characteristics assuming a jump leads to unphysical behavior: There are new characteristics appearing at all times, which means that information is produced (right panel). New information that is produced inside the system is a contradiction to the fundamental second law of thermodynamics, which has to be satisfied in any physical system (e.g., [118]). We omit showing values at the axis, since the figures are schematic.

ways – *independent from the initial condition* q_0. This leads to non-uniqueness. To choose a particular one from a set of solutions, one turns to physics, since there is little mathematical guidance for this.

The zero-viscosity limit There are different approaches for a mechanism to choose a solution that makes sense for the dynamical system which is modeled by the equation. A first idea is considering that the scalar conservation law (2.1) is used to model for example a physical process. In most physical processes some kind of viscosity acts, even though it is often neglected in models when the effect is sufficiently small. These considerations lead to the idea to study the solutions of Eq. (2.1) which can be found in the *zero-viscosity limit*, i.e., solutions determined as the limit of series of solutions q^ϵ of the viscous equation

$$\partial_t q^\epsilon + \partial_x f\left(q^\epsilon\right) = \epsilon \partial_x^2 q^\epsilon \tag{2.5}$$

for $\epsilon \to 0^+$ [103].[2] The viscosity on the right hand side acts such that it smooths out the solution additionally to transporting it on the characteristics[3]. If $\epsilon > 0$, an initial discontinuity is smoothed such that all values between q^L and q^R appear. Consequently, a fan of characteristics arises at the previous shock position if $f'(q^L) < f'(q^R)$. Hence, for the solution to be admissible, we demand that a fan of characteristics arises from a shock with $f'(q^L) < f'(q^R)$ such that the shock vanishes instantly. This is called a *rarefaction wave*. In the case that $f'(q^L) < f'(q^R)$, a self-steepening process due to transport occurs which opposes the effect of viscosity. Hence, the solution of a shock moving with velocity s in case (a) is also consistent with the approach of the zero-viscosity limit. To eventually choose a unique solution in the case of a rarefaction wave, one can assume that characteristics arising from the same point $(t_0, 0)$[4] are self-similar in the sense that $q(x,t) = q_{ss}(x/t)$. Using the conservation law, it is straightforward to derive the condition (e.g., [103])

$$f'(q_{ss}(x/t)) = x/t. \tag{2.6}$$

This allows for explicit computation of the rarefaction. In the example of the inviscid Burgers' equation this yields the solution

$$q(x,t) := \begin{cases} q^L & \text{if } x < tq^L, \\ q^R & \text{if } x > tq^R, \\ x/t & \text{else} \end{cases} \tag{2.7}$$

for the initial data given by

$$q_0(x) := \begin{cases} q^L & \text{if } x < 0, \\ q^R & \text{if } x \geq 0. \end{cases} \tag{2.8}$$

A solution obtained via the zero-viscosity limit approach for the inviscid Burgers' equation is sketched in Fig. 2.4. Collecting the cases, the solution of the scalar

[2] Note that the diffusion equation $\partial_t q = \epsilon \partial_x^2 q$ with a constant $\epsilon > 0$, which models viscous processes, is a parabolic PDE. In parabolic PDEs information travel with infinite velocity. Hence equations containing a viscosity term, such as Eq. (2.5), are not hyperbolic.

[3] Which means that q can change its value on characteristics.

[4] We center the initial shock leading to the rarefaction wave at $x = 0$ just for simplicity.

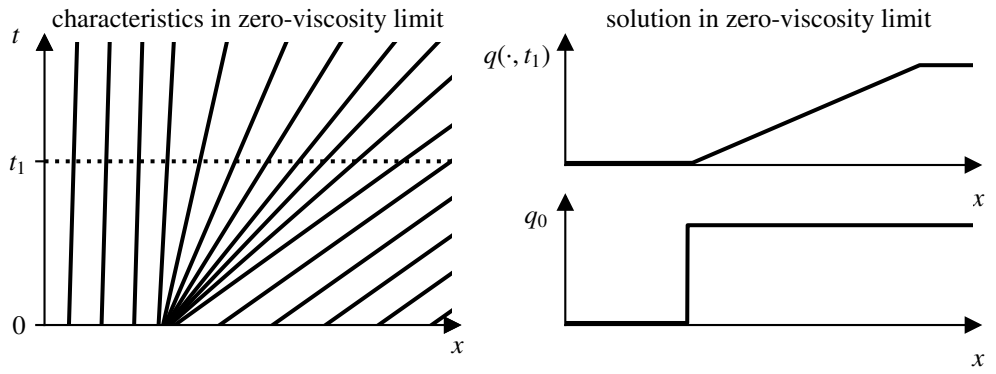


Figure 2.4: Characteristics corresponding to the zero-viscosity limit solution in case (b) in the text (right panel). The corresponding solution at later time (top right panel) is continuous, even for discontinuous initial conditions (bottom right panel). We omit showing values at the axis since the plots are schematic.

conservation law (2.1) is obtained by following characteristics, and, wherever discontinuities appear, shocks and rarefaction waves are constructed using the methods above.

2.1.1 Weak Solutions

As pointed out above, the notion of strong solutions of Eq. (2.1), for which q has to be differentiable, is not sufficient, since it does not allow for shocks. To also admit discontinuities in the solution, we define *weak solutions* as in [103]:

Definition 2.1.1. *The function $q(x,t)$ is a weak solution of the conservation law (2.1) if*

$$\int_{\mathbb{R}_0^+} \int_{\mathbb{R}} q(x,t)\partial_t\varphi(x,t) + f(q(x,t))\partial_x\varphi(x,t)\,dx\,dt = -\int_{\mathbb{R}} q_0(x)\varphi(x,0)\,dx \qquad (2.9)$$

holds for any continuously differentiable test function $\varphi \in \mathcal{C}_0^1\left(\mathbb{R}\times\mathbb{R}_0^+, \mathbb{R}\right)$ with compact support.

This concept of solutions is more general than the classical concept of strong solutions. It can be shown (e.g., [84]) that every strong solution of a conservation law is a weak solution. In the following we always consider weak solutions if not stated explicitly.

2.2 One-Dimensional Systems of Hyperbolic Conservation Laws

After understanding the basic properties of and concepts for one-dimensional scalar conservation laws, we now take the next step and consider the system of $\mathfrak{n}$ one-dimensional hyperbolic conservation laws

$$\partial_t \boldsymbol{q}(x,t) + \partial_x \boldsymbol{f}(\boldsymbol{q}(x,t)) = 0 \qquad (2.10)$$

for the vector(field) of conserved quantities $\boldsymbol{q} : \mathbb{R} \times \mathbb{R}_0^+ \to \mathbb{D}_{\boldsymbol{f}}$ and the flux function $\boldsymbol{f} : \mathcal{C}^2(\mathbb{D}_{\boldsymbol{f}}, \mathbb{R}^{\mathfrak{n}})$, where $\mathbb{D}_{\boldsymbol{f}} \subset \mathbb{R}^{\mathfrak{n}}$ is the set of admissible states.[5] For the system to be *hyperbolic*, the *flux Jacobian*

$$A(\boldsymbol{q}) := \left.\frac{\partial \boldsymbol{f}(\bar{\boldsymbol{q}})}{\partial \bar{\boldsymbol{q}}}\right|_{\bar{\boldsymbol{q}}=\boldsymbol{q}} \tag{2.11}$$

has to be diagonalizable with real eigenvalues. If, additionally, the eigenvalues of A are distinct the system is called *strictly hyperbolic*. As in the scalar case, we understand Eq. (2.10) in the weak sense, i.e., we allow for weak solutions satisfying the weak formulation

$$\int_{\mathbb{R}_0^+}\int_{\mathbb{R}} \boldsymbol{q}(x,t) \odot \partial_t \boldsymbol{\varphi}(x,t) + \boldsymbol{f}(\boldsymbol{q}(x,t)) \odot \partial_x \boldsymbol{\varphi}(x,t)\, dx\, dt = -\int_{\mathbb{R}} \boldsymbol{q}_0(x) \odot \boldsymbol{\varphi}(x,0)\, dx \tag{2.12}$$

of Eq. (2.10) for any test function $\boldsymbol{\varphi} \in C_0^1\left(\mathbb{R} \times \mathbb{R}_0^+, \mathbb{R}^{\mathfrak{n}}\right)$, where the symbol $\odot$ denotes the Hadamard product, i.e., the element-wise product. As in the one-dimensional case, hyperbolic systems model dynamical processes in which certain quantities are transported with finite velocity. However, it is more complicated for systems: In general, the quantities in the entries of the state vector $\boldsymbol{q}$ are not the ones which are transported along a characteristic. Also, the transport velocities are not as evident as in the case of a scalar conservation law. To investigate this, let us consider the quasi-linear form

$$\partial_t \boldsymbol{q}(x,t) + A(\boldsymbol{q}(x,t))\partial_x \boldsymbol{q}(x,t) = 0 \tag{2.13}$$

of Eq. (2.10). In the case that $\boldsymbol{f}$ in Eq. (2.10) is linear, we find that $A \in \mathbb{R}^{\mathfrak{n}\times\mathfrak{n}}$ has constant coefficients. The equation can then be diagonalized to the form

$$\partial_t R^{-1}\boldsymbol{q}(x,t) + \Lambda \partial_x \left(R^{-1}\boldsymbol{q}(x,t)\right) = 0, \tag{2.14}$$

where R is the matrix of right eigenvectors of A and $\Lambda = R^{-1}AR = \operatorname{diag}(\lambda_1, \ldots, \lambda_{\mathfrak{n}})$ is the diagonal matrix with the eigenvalues of A. This decouples the system and Eq. (2.14) consists of $\mathfrak{n}$ independent advection equations such that the k-th equation describes the transport of the k-th component of the *characteristic variables* $\boldsymbol{q}^{\text{char}} := R^{-1}\boldsymbol{q}$ with velocity λ_k.

In the nonlinear case, unfortunately, the situation is not as clear as in the linear case. The matrix of right eigenvectors depends on $\boldsymbol{q}$ and thus, implicitly, on x and t such that it does not commute with the spatial and temporal partial derivation. Moreover, the characteristic variables are not related to the conserved variables $\boldsymbol{q}$ in a linear way. However, even though we cannot construct a general solution formula for any hyperbolic system based on characteristics, we can still gain some understanding regarding the structure of solutions.

[5] The corresponding analogon to Footnote 1 holds here.

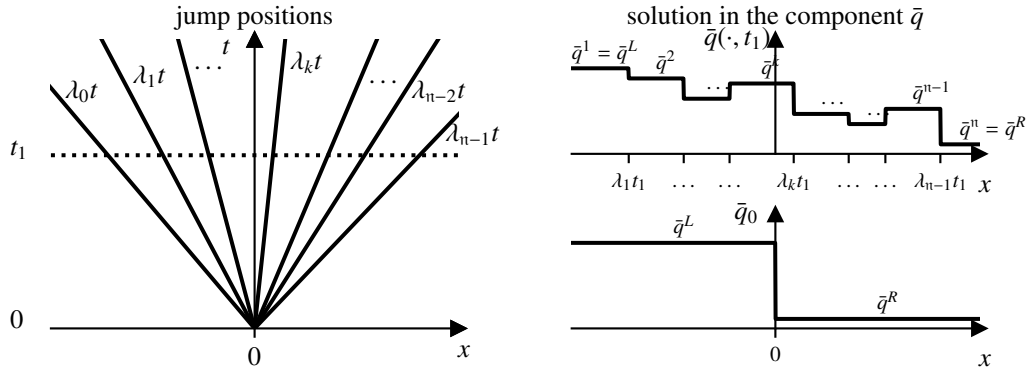


Figure 2.5: Shock positions in the (x,t)-plane (left panel) for Riemann problem initial data (right bottom) for a linear system of hyperbolic balance laws with a flux Jacobian that has $\mathfrak{n}$ mutually different eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_{\mathfrak{n}-1}$. The initial jump dissolves into maximal $\mathfrak{n}$ jumps that travel with the velocities λ_k. In the top left panel the solution in a component $\bar{q}$ of the solution $\boldsymbol{q}$ is shown at time $t_1 > 0$. The values of the intermediate states $\bar{q}^k$ are computed as discussed in the text. This plot is schematic and the jump velocities and values of the intermediate states are chosen randomly.

2.2.1 The Riemann Problem for One-Dimensional Systems of Conservation Laws

Especially later, when we discuss numerical approaches to solve systems of this type, the Riemann problem given by the initial data

$$\boldsymbol{q}_0(x) := \begin{cases} \boldsymbol{q}^L & \text{if } x < 0, \\ \boldsymbol{q}^R & \text{if } x > 0 \end{cases} \tag{2.15}$$

for arbitrary admissible left and right states $\boldsymbol{q}^L$ and $\boldsymbol{q}^R$ will have particular relevance. Remember that there is no necessity to decide on a value for $\boldsymbol{q}_0(0)$, since the hyperbolic conservation law is interpreted in the weak sense and single point values are irrelevant.

In the following we only give a very short discussion on how to approach and solve a Riemann problem, since an in-depth discussion can become quite lengthy and excellent explanations can be found in literature (e.g., [103, 164]). To get a first idea one can, once more, consider the linear system defined by $\boldsymbol{f}(\boldsymbol{q}) = A\boldsymbol{q}$, were the matrix A has constant coefficients. In this case, as discussed above, the system can be diagonalized (otherwise it is not hyperbolic by definition) and the solution of the Riemann problem can be obtained by solving the Riemann problem for each scalar conservation law describing the evolution of $(\boldsymbol{q}^{\text{char}})_k$ as described in Section 2.1. Transforming back to conserved variables $\boldsymbol{q} = R\boldsymbol{q}^{\text{char}}$ yields the solution. The solution consists then of up to $\mathfrak{n}$ jumps moving with velocities λ_k as shown schematically in Fig. 2.5. Between the jump positions $\boldsymbol{q}$ takes up to $\mathfrak{n}-1$ intermediate states $\boldsymbol{q}^k$, $k \in \{1, \ldots, \mathfrak{n}-1\}$ with the relation $\boldsymbol{q}^{k+1} = \boldsymbol{q}^k + a^k \boldsymbol{r}^k$ for $k \in \{0, \ldots, \mathfrak{n}-1\}$, where $a^k \in \mathbb{R}$, $\boldsymbol{r}^k$ is the k-th right eigenvector and we set $\boldsymbol{q}^0 = \boldsymbol{q}^L$ and $\boldsymbol{q}^{\mathfrak{n}} = \boldsymbol{q}^R$. In short, this means that the different states appearing in the solutions are connected along eigenvectors over the jumps.

In principle, the same approach is used to solve the Riemann problem for nonlinear systems of conservation laws. There are three main difficulties compared to the linear case: Firstly, the shock velocities have to be determined using a Rankine–Hugoniot condition as in Section 2.1, since the eigenvalues of $A(\boldsymbol{q}^k)$ and $A(\boldsymbol{q}^{k+1})$ are in general not the same. Secondly, the different intermediate states are not connected via eigenvectors, but via integral curves of the eigenvector fields, since the eigenvectors of $A(\boldsymbol{q})$ depend on the state $\boldsymbol{q}$. This makes it much more challenging to find the correct intermediate states for the solution of the Riemann problem. And finally, as already discussed in the scalar case, some states should actually not be connected by shocks but by rarefaction waves to yield physical solutions which can, e.g., be found in the zero-viscosity limit. To decide, whether states should be connected by shocks or rarefaction waves, it is common to use an entropy inequality as a criterion.

2.2.2 Entropy Conditions

A more general approach than the one given in Section 2.1 to decide on the admissibility of solutions is an *entropy condition* in the form of the inequality

$$\partial_t \eta(\boldsymbol{q}) + \partial_x \psi(\boldsymbol{q}) \leq 0, \tag{2.16}$$

which is added to the hyperbolic conservation law (2.10). The pair (η, ψ) is called *entropy-entropy flux pair*. The *entropy* $\eta : \mathbb{R}^{\mathfrak{n}} \to \mathbb{R}$ is a convex function and together with the *entropy flux* it satisfies the relation

$$\psi'(\boldsymbol{q}) = \eta'(\boldsymbol{q}) f'(\boldsymbol{q}). \tag{2.17}$$

This ensures that the entropy is conserved in regions in which the solution is smooth. At discontinuities the entropy decreases if the solution is obtained via the zero-viscosity limit approach and it increases if there is an unphysical shock. In [103], e.g., the entropy inequality is derived by considering the zero-viscosity limit. Note that, in order to be applicable in the presence of discontinuities at all, Eq. (2.16) has to be interpreted in the weak sense

$$\int_{\mathbb{R}_0^+} \int_{\mathbb{R}} \eta(\boldsymbol{q}(x,t)) \partial_t \varphi(x,t) + \psi(\boldsymbol{q}(x,t)) \partial_x \varphi(x,t) \, dx \, dt \leq - \int_{\mathbb{R}} \eta(\boldsymbol{q}_0(x)) \varphi(x,0) \, dx. \tag{2.18}$$

Here, since Eq. (2.18) is an inequality rather than an equality, we only consider non-negative test-functions $\varphi \in \mathcal{C}_0^1\left(\mathbb{R} \times \mathbb{R}_0^+, \mathbb{R}_0^+\right)$. For scalar conservation laws, an entropy-entropy flux pair can be defined by simply choosing a convex function η and integrating Eq. (2.17) to obtain a suitable entropy flux ψ. For systems with $\mathfrak{n} \geq 2$, the existence of an entropy-entropy flux pair is not guaranteed in general. According to Godunov [76] (see also [150]) any symmetrizable system has an entropy and any system that has a convex entropy can be symmetrized by its Hessian matrix $\eta''(\boldsymbol{q})$ [64].

An entropy condition is a natural condition for the admissibility of solutions of PDEs modeling physical processes, since the second law of thermodynamics

(e.g., [118]) has to be satisfied in any physical process. The second law of thermodynamics states that the total entropy of an isolated system can never increase[6] over time. The Euler system, which is the hyperbolic system mainly considered in this thesis, is symmetrizable and thus has an entropy (see Section 2.3.2).

2.3 Homogeneous Compressible Euler Equations

The one-dimensional compressible Euler equations, which model the flow of compressible, inviscid fluids, are an important example of a strictly hyperbolic system. They are given by Eq. (2.10) with

$$\boldsymbol{q} = \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{f}(\boldsymbol{q}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ (E + p)\, u \end{pmatrix}, \tag{2.19}$$

where ρ, p, and u are the fluid's volumetric density, pressure, and velocity. The volumetric total energy is given by $E = \varepsilon + E_{\text{kin}}$ with the volumetric kinetic energy $E_{\text{kin}} = \frac{1}{2}\rho u^2$ and the volumetric internal energy $\varepsilon = \rho\epsilon$. The specific internal energy ϵ is related to density and pressure via an additional relation which is called *equation of state* (EoS).

2.3.1 Equations of State

An EoS is a relation between the thermodynamical quantities ρ, p, and ϵ, which is given in an explicit or implicit form. Often it is formulated using the gas temperature T, which is also related to ρ, p, and ϵ. In this thesis, whenever we explicitly relate to a quantity that is obtained by directly using the EoS, we write it as a function $\rho_{\text{EoS}}(p, \epsilon)$, $p_{\text{EoS}}(\rho, \epsilon)$, or $\epsilon_{\text{EoS}}(\rho, p)$, respectively.[7] The choice of an EoS is a physical question, or, more precisely, a question regarding the properties of the fluid that shall be modeled. The relation between the thermodynamic quantities is generally obtained via theoretical physical considerations (e.g., [160]). In the following we give two examples of physically relevant EoS.

2.3.1.1 Ideal Gas

A relatively simple yet highly relevant EoS is the ideal gas EoS given by

$$p = \rho R T \qquad \text{and} \qquad \epsilon = \frac{RT}{\gamma - 1} \tag{2.20}$$

with the ratio of specific heats γ and the gas constant R (e.g., [163]). Equation (2.20) is a formulation of the ideal gas law or Boyle–Charles–Avogadro law [20], which

[6] In the conventional physical notation, the sign of the entropy is chosen different to the mathematical convention, so actually the second law of thermodynamics states, that the total entropy of an isolated system can never decrease over time.

[7] This distinction is relevant for certain derivatives and also later in numerical methods, in which the direct evolution of a quantity using the scheme and the quantity computed via the EoS using the other thermodynamical quantities which have been evolved by the scheme in some way may yield different results.

describes the relation between the thermodynamical quantities for a gas that follows the early experimental findings:

- *Boyle's law*: Isotherms are $pV = \text{const.}$, where V is the volume of the gas (e.g., [131])
- *Charles' law*: When the pressure on a sample of a dry gas is held constant, the Kelvin temperature and the volume will be in direct proportion [66]. This means that $V \propto T$.
- *Gay-Lussac's law*: If the volume of a given mass of an ideal gas is kept constant, the pressure varies directly with the temperature, i.e., $p \propto T$ (e.g., [48]).
- *Avogadro's law*: Two samples of an ideal gas at same temperature and pressure contain the same number of molecules [4].

The ideal gas law can also be derived from first principles using kinetic gas theory (e.g., [20]). Explicit evaluation of $\rho_{\text{EoS}}(p, \epsilon)$, $p_{\text{EoS}}(\rho, \epsilon)$, and $\epsilon_{\text{EoS}}(\rho, p)$ is possible using

$$p = (\gamma - 1)\rho\epsilon. \tag{2.21}$$

The ratio of specific heats γ can be related to the number of molecular degrees of freedom $\mathfrak{f}$ by $\gamma = 1 + 2/\mathfrak{f}$ (e.g., [20]). A single atom has three degrees of freedom (translational), and a mono-atomic gas can hence be described using $\gamma = 5/3$. For a diatomic gas it is $\gamma = 7/5 = 1.4$, since there are five degrees of freedom (three translational and two rotational) for this type of molecule. To model air on the earth's surface, which is mainly composed from the diatomic gases N_2 (around 78%) and O_2 (around 21%) [162], we use Eq. (2.21) with $\gamma = 1.4$.

2.3.1.2 Ideal Gas with Radiation Pressure

There are many different scenarios, in which the ideal gas assumptions are violated by real physical gases. In that case other physical effects enter and the description of the thermodynamical relation of the quantities can become quite complicated (e.g., [160]). In numerical experiments in this thesis, we only consider one correction term[8] to the ideal gas EoS: The EoS [38]

$$p = \rho RT + \frac{1}{3}a_{SB}T^4, \qquad \epsilon = \frac{RT}{\gamma - 1} + \frac{a_{SB}}{\rho}T^4, \tag{2.22}$$

where a_{SB} is the *Stefan–Boltzmann constant*, describes a gas which mainly satisfies the ideal gas assumptions but additionally is subject to radiation pressure. The major practical difference between the ideal gas EoS and the EoS (2.22) is that the ideal gas EoS can be evaluated explicitly. In the case of an ideal gas with radiation pressure, to compute the pressure from the specific internal energy or vice versa while knowing the density includes solving an implicit equation for the temperature.

[8] even though our numerical methods are constructed such that they work with any arbitrarily complicated EoS

2.3.2 Variables, Eigenstructure, and Entropy

As pointed out earlier, the eigenstructure of the flux Jacobian is of high importance for solving Riemann problems, since the eigenvalues in the end provide the integral curves that connect states over shocks or rarefactions. For this purpose, let us take a look at the flux Jacobian

$$A(\boldsymbol{q}) = \left.\frac{\partial \boldsymbol{q}^{\text{cons}}}{\partial \boldsymbol{q}^{\text{prim}}}\right|_{\boldsymbol{q}} A^{\text{prim}}(\boldsymbol{q}) \left.\frac{\partial \boldsymbol{q}^{\text{prim}}}{\partial \boldsymbol{q}^{\text{cons}}}\right|_{\boldsymbol{q}} \tag{2.23}$$

of compressible Euler equations for a moment. Here, $\boldsymbol{q}^{\text{cons}} = \boldsymbol{q}$ is the state vector of conserved variables as introduced in Eq. (2.19) and

$$\boldsymbol{q}^{\text{prim}} := \begin{pmatrix} \rho \\ u \\ p \end{pmatrix} = \begin{pmatrix} q_1^{\text{cons}} \\ \frac{q_2^{\text{cons}}}{q_1^{\text{cons}}} \\ p_{\text{EoS}}\left(q_1^{\text{cons}}, \frac{q_3^{\text{cons}}}{q_1^{\text{cons}}} - \frac{1}{2}\left(\frac{q_2^{\text{cons}}}{q_1^{\text{cons}}}\right)^2\right) \end{pmatrix} \tag{2.24}$$

is the state vector of *primitive variables*. The transformation matrices between these two variable systems are

$$\frac{\partial \boldsymbol{q}^{\text{cons}}}{\partial \boldsymbol{q}^{\text{prim}}} = \begin{pmatrix} 1 & 0 & 0 \\ u & \rho & 0 \\ \frac{1}{2}u^2 - \frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \rho}\left(\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon}\right)^{-1} & \rho u & \left(\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon}\right)^{-1} \end{pmatrix}. \tag{2.25}$$

and

$$\frac{\partial \boldsymbol{q}^{\text{prim}}}{\partial \boldsymbol{q}^{\text{cons}}} = \left(\frac{\partial \boldsymbol{q}^{\text{cons}}}{\partial \boldsymbol{q}^{\text{prim}}}\right)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{u}{\rho} & \frac{1}{\rho} & 0 \\ \frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \rho} + \frac{1}{2}\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon}u^2 & -\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon}u & \frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon} \end{pmatrix}. \tag{2.26}$$

We use these since the flux Jacobian takes a simpler form in primitive variables:

$$A^{\text{prim}}(\boldsymbol{q}) = \begin{pmatrix} u & \rho & 0 \\ 0 & u & \frac{1}{\rho} \\ 0 & \rho c^2 & u \end{pmatrix} = R^{\text{prim}}(\boldsymbol{q})\Lambda(\boldsymbol{q})\left(R^{\text{prim}}(\boldsymbol{q})\right)^{-1} \tag{2.27}$$

with the *speed of sound*

$$c = c(\rho,\epsilon) := \sqrt{\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \rho} + \frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon} \cdot \frac{\epsilon + p_{\text{EoS}}(\rho,\epsilon)}{\rho}}. \tag{2.28}$$

The eigensystem of A^{prim} is given by the diagonal matrix of eigenvalues

$$\Lambda(\boldsymbol{q}) = \text{diag}(u, u+c, u-c) \tag{2.29}$$

and the matrix of right eigenvectors

$$R^{\text{prim}}(\boldsymbol{q}) = \begin{pmatrix} 1 & \frac{1}{c^2} & \frac{1}{c^2} \\ 0 & \frac{1}{c\rho} & -\frac{1}{c\rho} \\ 0 & 1 & 1 \end{pmatrix}. \tag{2.30}$$

The right eigenvectors of A^{cons} can then be obtained via

$$R^{\text{cons}}(\boldsymbol{q}) = \left.\frac{\partial \boldsymbol{q}^{\text{cons}}}{\partial \boldsymbol{q}^{\text{prim}}}\right|_{\boldsymbol{q}} R^{\text{prim}}(\boldsymbol{q}). \tag{2.31}$$

With this, solutions of the general Riemann problem (2.15) can be constructed. Physical solutions are be chosen by adding an entropy inequality with the entropy-entropy flux pair

$$\eta = -\frac{\rho s}{\gamma - 1}, \qquad \psi := -\frac{\rho u s}{\gamma - 1}, \tag{2.32}$$

where $s := \ln(p\rho^{-\gamma}) = -(\gamma - 1)\ln(\rho) - \ln(\beta) - \ln(2)$ up to a constant with $\beta := 1/(2RT)$.

2.4 Compressible Euler Equations with Gravitational Source Term

The numerical methods introduced in this thesis are not developed for the conservative homogeneous Euler system Eq. (2.19). Instead, we add the source term

$$\boldsymbol{s} = \begin{pmatrix} 0 \\ \rho g \\ \rho u g \end{pmatrix} \tag{2.33}$$

to the system, in which the gravitational acceleration g is usually defined as the negative spatial derivative of a given gravitational potential $\phi \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$, i.e., $g(x) = -\phi'(x)$. The Euler system with gravitational source term has the form of a hyperbolic balance law

$$\partial_t \boldsymbol{q} + \partial_x \boldsymbol{f} = \boldsymbol{s}, \tag{2.34}$$

which is non-conservative in momentum and total energy.[9] Obviously, this also changes the solutions of Riemann problems (e.g., [22]). The numerical methods we discuss later are nonetheless based on the Riemann problem for homogeneous Euler equations. A new type of stationary solutions is admitted by the system when the gravity source term is added. These solutions are discussed in the following section.

2.4.1 Hydrostatic Solutions

By setting the velocity to $u(x,t) = 0$ for all x and t in the Euler system with gravity (2.34), one obtains $\partial_t \boldsymbol{q} = 0$ and

$$\partial_x p = \rho g. \tag{2.35}$$

[9] The system can be made conservative in total energy by redefining it to $E = \varepsilon + E_{\text{kin}} + \rho\phi$ and setting the third component of the source term to zero. We decided to keep the energy source term to show how it can be treated in the numerical methods we introduce. It is then straightforward to adapt the methods to the case in which the energy source term is included in the total energy as potential energy $\rho\phi$.

Equation (2.35) is called *hydrostatic equation* and a constant-in-time state $\boldsymbol{q}$ (often given in terms of ρ and p) is called *hydrostatic solution* if it satisfies Eq. (2.35) together with the EoS which is used to model the thermodynamical relations. From the modeling point of view, hydrostatic states can play a fundamental role depending on the application. Classical examples are the basic stellar structure and also the structure of a planet's atmosphere, which can in many cases be well approximated using hydrostatic states (e.g., [38]).

Solutions of the ODE[10] (2.35) are in general not unique, since additional degrees of freedom enter the equation via the EoS. However, they exist[11]: Solutions can for example be constructed by choosing a piecewise continuous function ρ and defining the pressure $p(x) = p_0 + \int_{x_0}^{x} \rho(x)g(x)\,dx$. In this example, the hydrostatic solution is chosen by assuming a certain stratification for the density and choosing a tuple (x_0, p_0). In the following we introduce some relevant hydrostatic states using different assumptions:

Isothermal hydrostatic solution Assuming an ideal gas law and a constant temperature $T = T_0$, the following hydrostatic states can be found for a given gravitational potential ϕ:

$$\rho(x) = \rho_0 \exp\left(-\frac{\phi(x)}{RT_0}\right), \qquad p(x) = \rho_0 RT_0 \exp\left(-\frac{\phi(x)}{RT_0}\right), \qquad u \equiv 0, \tag{2.36}$$

where $\rho_0 > 0$ is some constant.

Polytropic hydrostatic solution Polytropic hydrostatic solutions are of the form

$$\theta(x) = 1 - \frac{\nu - 1}{\nu}\phi(x), \quad \rho(x) = \kappa\theta(x)^{\frac{1}{\nu-1}}, \quad p(x) = \kappa\theta(x)^{\frac{\nu}{\nu-1}}, \quad u(x) = 0 \tag{2.37}$$

with constants $\nu > 1$ and $\kappa > 0$. Equation (2.37) describes a hydrostatic state of the one-dimensional compressible Euler system with gravity source term independent from the EoS. They are based on the assumption that the density and pressure stratification are related by the additional condition $p \propto \rho^\nu$. If the thermodynamical properties of the gas are modeled via an ideal gas law the hydrostatic temperature is $T(x) = \theta(x)/R$.

[10] In Eq. (2.35) we use a partial derivative, since the pressure in the Euler system is allowed to depend on time. Using $\partial_t \boldsymbol{q} = 0$, however, lets the temporal derivative vanish and Eq. (2.35) is essentially an ODE.

[11] provided the boundary conditions allow this

Chapter 3

Finite Volume Methods for One-Dimensional Systems of Hyperbolic Balance Laws

The numerical techniques for approximating solutions of hyperbolic balance laws we use in this thesis are based on the *method of lines* (e.g., [146]), an approach in which every dimension but the temporal one is discretized. The discretization yields then a (potentially very large) system of ODEs which can be solved using standard methods such as *Runge–Kutta* (RK) methods (see Section 3.7). It seems a natural choice to discretize in space and evolve the system in time using an ODE solver, since it is usually clear how the spatial domain Ω looks like whereas the temporal domain with $t \geq t_0$ can be unbounded such that it is not a priori clear how to choose the final time of a simulation. Additionally, when we extend the method to multiple spatial dimensions, it is a natural choice to treat all the spatial dimensions in the same manner.

For the spatial discretization we choose a *finite volume* (FV) approach, which is based on averaging the conserved quantities in each cell and approximating the fluxes between the cells. This approach is especially suited for hyperbolic conservation laws, because it leads to conservative methods and allows for discontinuities at the cell interfaces. Thus, it mirrors two of the basic properties of hyperbolic systems. In this chapter, we describe the basic FV discretization for one-dimensional balance laws which is then evolved in time using RK methods in the method of lines framework.

3.1 Godunov's Method

The roots of modern FV methods can be found in Godunov's method ([75], for an English description see [103]). This section aims at illustrating Godunov's approach. Consider the initial value problem given by the one-dimensional system of hyperbolic conservation laws (2.10) in the spatial domain $\Omega := [a, b] \subset \mathbb{R}$ $(a < b)$ with an initial

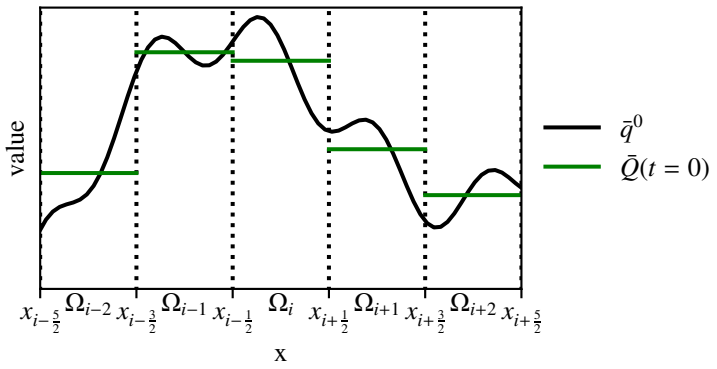


Figure 3.1: To initialize the Godunov method, the initial data given as function $\boldsymbol{q}^0$ are averaged in each cell Ω_k. The piecewise constant function $\boldsymbol{Q}(t=0)$, which is obtained by this averaging process, is then used as initial data in the Godunov algorithm. In this schematic plot only one component of the vector-valued functions is shown. This is indicated by the bar $\bar{\cdot}$.

condition $\boldsymbol{q}_0$, i.e.,

$$\partial_t \boldsymbol{q}(x,t) + \partial_x \boldsymbol{f}(\boldsymbol{q}(x,t)) = 0 \qquad \text{for } x \in \Omega, t > 0, \tag{3.1}$$

$$\boldsymbol{q}(x,0) = \boldsymbol{q}_0(x) \quad \text{for } x \in \Omega. \tag{3.2}$$

Equation (3.1) is interpreted in the weak sense. For simplicity, we neglect the treatment of the boundaries a and b in this section. Boundaries will be discussed in Section 3.8. Since this problem cannot be solved in general, let us find a simple approximation of the initial data, for which we can solve it analytically for a short time interval. For this purpose, we subdivide the domain Ω by choosing points $a = x_{-\frac{1}{2}} < x_{\frac{1}{2}} < \cdots < x_{N+\frac{1}{2}} = b$ and defining sub-intervals $\Omega_i = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ for $i \in \mathcal{I} = \{1, \ldots, N\}$, which we call *cells* in the following.[1] In every cell we average the initial data and define

$$(\hat{\boldsymbol{q}_0})_i := \frac{1}{\Delta x_i} \int_{\Omega_i} \boldsymbol{q}_0(x)\, dx, \tag{3.3}$$

where $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ is the length of the i-th cell. We use these averages to define the piecewise constant approximate initial data

$$\boldsymbol{Q}(x,0) = \hat{\boldsymbol{Q}}_i^0 := (\hat{\boldsymbol{q}_0})_i \qquad \text{for } x \in \Omega_i \tag{3.4}$$

as visualized in Fig. 3.1 and consider the modified initial value problem given by

$$\partial_t \boldsymbol{Q} + \partial_x \boldsymbol{f}(\boldsymbol{Q}) = 0 \quad \text{for } x \in \Omega, t > 0, \tag{3.5}$$

$$\boldsymbol{Q}(x,0) = \hat{\boldsymbol{Q}}_i^0 \text{ for } x \in \Omega_i. \tag{3.6}$$

[1] Note that the cells are, in this definition, not really disjunct but they share common interface points. However, in Section 2.1.1 we have seen that single points are not relevant due to the integral form of the weak solutions. This behavior is mirrored in the FV approach. We could also define the cells as $\Omega_i = \left(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right)$ without further consequences.

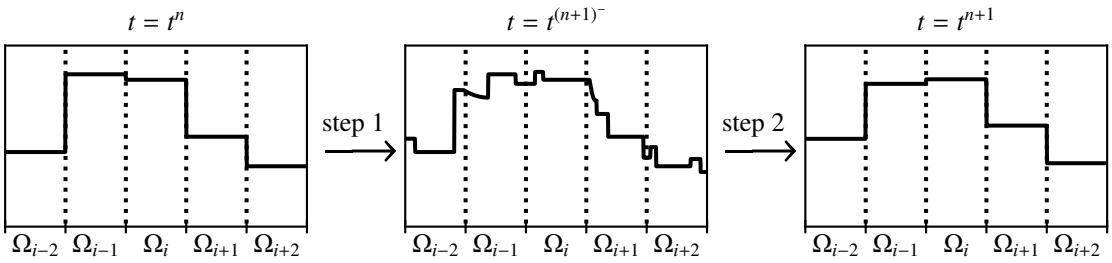


Figure 3.2: Visualization of steps 1 and 2 of Algorithm 3.1.1. In the first step, the Riemann problems at the interfaces of the piecewise constant function (left panel) are solved in order to obtain an exact solution at time t^{n+1} (central panel). In the second step, the exact solution is cell-averaged such that a piecewise constant function is obtained (right panel). The plot is schematic and we omit labeling of the axis for clarity of the visualization.

The piecewise constant data can be evolved in time analytically, since only the Riemann problems at the cell interfaces $x_{i+\frac{1}{2}}$ ($i \in \{1, .., N-1\}$) have to be solved (recall Section 2.2.1).

This approach provides the exact solution until the waves of the Riemann problem's solutions from neighboring interfaces meet. Let us denote the time at which the first waves on the whole domain meet with t^1. Then, this step can be repeated after computing the new cell averages $\boldsymbol{Q}_i^1$ from the exact solution $\boldsymbol{Q}(\cdot, \Delta t^1)$ at time t^1 of the initial value problem in Eqs. (3.5) and (3.6).

The procedure can be formulated as

Algorithm 3.1.1 (Godunov's method)**.** *The Algorithm starts with cell-averages $\hat{\boldsymbol{Q}}_i^0$ for $i \in \mathcal{I}$. An upper index denotes the time step number n, and we set $n = 0$ initially.*

1. *Solve the initial value problem*

$$\partial_t \boldsymbol{Q}(x,t) + \partial_x \boldsymbol{f}(\boldsymbol{Q}(x,t)) = 0 \quad \text{for } (x,t) \in \Omega \times [t^n, t^{n+1}], \tag{3.7}$$

$$\boldsymbol{Q}(x,t^n) = \hat{\boldsymbol{Q}}_i^n \ \text{ for } x \in \Omega_i, i \in \mathcal{I}, \tag{3.8}$$

   where t^{n+1} is chosen such that waves from different interfaces cannot interact.

2. *Compute the new cell averages*

$$\hat{\boldsymbol{Q}}_i^{n+1} := \frac{1}{\Delta x_i} \int_{\Omega_i} \boldsymbol{Q}\left(x, t^{n+1}\right) \, dx \tag{3.9}$$

   from the solution $\boldsymbol{Q}$ of the initial value problem Eqs. (3.7) *and* (3.8)*.*

3. *Check if the desired final time t_{final} is reached, i.e., if $t^n \geq t_{\text{final}}$. If not, increase n by 1 and repeat the three steps.*

Algorithm 3.1.1 is visualized in Fig. 3.2. For the purpose of choosing a suitable time t^{n+1} before solving Eqs. (3.7) and (3.8), we use the knowledge from Section 2.2.1

about the wave structure of the Riemann problem's solution at each interface. A common choice is $t^{n+1} = t^n + \Delta t^n$ with

$$\Delta t^n = \frac{\min\limits_{i\in\mathcal{I}}(\Delta x_i)}{\max\limits_{i\in\mathcal{I}}\left(\left(\left|\lambda\left(\hat{\boldsymbol{Q}}_i^n\right)\right|_{\max}\right)\right)}, \tag{3.10}$$

where

$$|\lambda(\boldsymbol{q})|_{\max} := \max\{|\lambda| \,:\, \lambda \text{ is eigenvalue of } A(\boldsymbol{q})\} \tag{3.11}$$

with the matrix field A being the flux Jacobian of $\boldsymbol{f}$ as defined as in Eq. (2.11). This time interval Δt^n is chosen such that every wave crosses at most half of a cell during this period

3.2 Finite Volume Formulation of Godunov's Method

To proceed towards more recent methods, we reformulate Godunov's method as a first order accurate Runge–Kutta finite volume (RK-FV) method. We use the same spatial discretization as in Section 3.1 to solve the initial value problem (3.1), (3.2) approximately. In this section we cell-average both the initial condition (3.2) and the conservation law (3.1). Applying the fundamental theorem of calculus to Eq. (3.1) yields the time-evolution

$$\frac{d}{dt}\hat{\boldsymbol{q}}_i(t) = -\frac{1}{\Delta x_i}\left[\boldsymbol{f}\left(\boldsymbol{q}\left(x_{i+\frac{1}{2}}, t\right)\right) - \boldsymbol{f}\left(\boldsymbol{q}\left(x_{i-\frac{1}{2}}, t\right)\right)\right] \tag{3.12}$$

for the cell-averaged states

$$\hat{\boldsymbol{q}}_i(t) := \frac{1}{\Delta x_i}\int_{\Omega_i} \boldsymbol{q}(x,t)\,dx, \tag{3.13}$$

which is still an exact statement. As in Section 3.1 we aim to evolve a numerical approximation of the cell-averages in time and the interface values of the solution are thus not known. Instead, we obtain the interface values from the solutions of the interface Riemann problems an write

$$\frac{d}{dt}\hat{\boldsymbol{Q}}_i(t) = -\frac{1}{\Delta x_i}\left[\boldsymbol{F}^{\mathrm{God}}\left(\hat{\boldsymbol{Q}}_i(t), \hat{\boldsymbol{Q}}_{i+1}(t)\right) - \boldsymbol{F}^{\mathrm{God}}\left(\hat{\boldsymbol{Q}}_{i-1}(t), \hat{\boldsymbol{Q}}_i(t)\right)\right], \tag{3.14}$$

with

$$\boldsymbol{F}^{\mathrm{God}}\left(\hat{\boldsymbol{Q}}_i(t), \hat{\boldsymbol{Q}}_{i+1}(t)\right) := \boldsymbol{f}\left(\boldsymbol{Q}^*_{i+\frac{1}{2}}(t)\right), \tag{3.15}$$

where $\boldsymbol{Q}^*_{i+\frac{1}{2}}(t)$ is the solution of the Riemann problem

$$\partial_s \boldsymbol{Q}^{t,*}(x,s) + \partial_x \boldsymbol{f}(\boldsymbol{Q}^{t,*}(x,s)) = 0 \text{ for } (x,s) \in \Omega_i \times [t, t+\tau], \tag{3.16}$$

$$\boldsymbol{Q}^{t,*}(x,t) = \begin{cases} \hat{\boldsymbol{Q}}_i(t) & \text{for } x \le x_{i+\frac{1}{2}} \\ \hat{\boldsymbol{Q}}_{i+1}(t) & \text{for } x > x_{i+\frac{1}{2}} \end{cases}, \tag{3.17}$$

at the $i+\frac{1}{2}$ interface for $s = t+\tau$, i.e., $\boldsymbol{Q}^*_{i+\frac{1}{2}}(t) = \boldsymbol{Q}^{t,*}\left(x_{i+\frac{1}{2}}, t+\tau\right)$, where $\tau > 0$ can be chosen arbitrarily.

Equation (3.14) is still a nonlinear equation. We evolve it in time using small time steps Δt^n, which are determined as in Eq. (3.10).[2] The full scheme reads

$$\hat{\boldsymbol{Q}}_i^0 = \frac{1}{\Delta x_i}\int_{\Omega_i} \boldsymbol{q}_0(x)\,dx \tag{3.18}$$

$$\hat{\boldsymbol{Q}}_i^{n+1} = \hat{\boldsymbol{Q}}_i^n - \frac{\Delta t^n}{\Delta x_i}\left[\boldsymbol{F}^{\mathrm{God}}\left(\hat{\boldsymbol{Q}}_i^n, \hat{\boldsymbol{Q}}_{i+1}^n\right) - \boldsymbol{F}^{\mathrm{God}}\left(\hat{\boldsymbol{Q}}_{i-1}^n, \hat{\boldsymbol{Q}}_i^n\right)\right] \tag{3.19}$$

for all $i \in \mathcal{I}$, $n \in \mathbb{N}$.

Theorem 3.2.1. *The scheme described in Eqs.* (3.18) *and* (3.19) *is equivalent to Algorithm 3.1.1 in the sense that it yields the same values* $\hat{\boldsymbol{Q}}_i^n$ *for all* $i \in \mathcal{I}$, $n \in \mathbb{N}$, *provided that the initial data are the same.*

Proof. This can be shown based on the insight that – after averaging the exact solution of the piecewise constant initial data in the i-th cell – only the flux at the interface states $\boldsymbol{Q}^*_{i-\frac{1}{2}}, \boldsymbol{Q}^*_{i+\frac{1}{2}}$ given by the solution of the Riemann problem is relevant in Godunov's method as described in Algorithm 3.1.1. Instead of a technical full proof we just refer to [163], which is also a resourceful reference for further information about Godunov's method. □

We shortly summarize the previously introduced original Godunov method in the form of an RK-FV scheme: In Eq. (3.14) we approximate Eq. (2.10) by a semi-discrete equation using the *numerical flux function* $\boldsymbol{F}^{\mathrm{God}}$. In Eq. (3.19) we use the forward Euler method (see Section 3.7.1) to evolve the system of ODEs Eq. (3.14) in time. The forward Euler method is the simplest case of *Runge–Kutta* (RK) method. These ingredients of an RK-FV method, numerical flux functions and RK methods, are introduced in the course of this chapter (Sections 3.3 and 3.7). Moreover, we discuss how the spatial order of a finite volume method can be increased (Section 3.4) and how source terms can be added into the numerical method (Section 3.6).

3.3 Numerical Flux Functions

In this and the following section (Sections 3.3 and 3.4) we only consider the semi-discrete equation

$$\frac{d}{dt}\hat{\boldsymbol{Q}}_i(t) = -\frac{1}{\Delta x_i}\left[\boldsymbol{F}\left(\hat{\boldsymbol{Q}}_i(t), \hat{\boldsymbol{Q}}_{i+1}(t)\right) - \boldsymbol{F}\left(\hat{\boldsymbol{Q}}_{i-1}(t), \hat{\boldsymbol{Q}}_i(t)\right)\right], \tag{3.20}$$

which is the same as Eq. (3.14) with a general numerical flux function $\boldsymbol{F}$.

Definition 3.3.1. *A function* $\boldsymbol{F} : \mathbb{D}_{\boldsymbol{f}} \times \mathbb{D}_{\boldsymbol{f}} \to \mathbb{R}^{\mathfrak{n}}$ *is called* numerical flux function consistent with the physical flux $\boldsymbol{f} \in \mathcal{C}^2\left(\mathbb{D}_{\boldsymbol{f}}, \mathbb{R}^{\mathfrak{n}}\right)$ *(recall that* $\mathbb{D}_{\boldsymbol{f}} \subset \mathbb{R}^{\mathfrak{n}}$ *is the set of admissible values for* $\boldsymbol{q}$*) if it satisfies the following properties:*

[2] In Section 3.7.1 we will see that this is the forward Euler method for time stepping.

(i) $\boldsymbol{F}(\boldsymbol{q},\boldsymbol{q}) = \boldsymbol{f}(\boldsymbol{q})$ *for any* $\boldsymbol{q} \in \mathbb{D}_{\boldsymbol{f}}$.

(ii) $\boldsymbol{F}$ *is Lipschitz-continuous in both arguments.*

Consistency of the numerical flux function is important, since it ensures consistency of the semi-discrete scheme Eq. (3.20) with the hyperbolic conservation law Eq. (2.10) (see [103] for reference). We illustrate this statement under the simplifying assumption that the solution which shall be approximated is sufficiently smooth.

Let $\hat{\boldsymbol{q}}_i$ for $i \in \mathcal{I}$ be the cell-averages of a solution $\boldsymbol{q} \in \mathcal{C}^1$ of Eq. (2.10) . Let $\boldsymbol{F}$ be a numerical flux function consistent with $\boldsymbol{f}$ as defined in Definition 3.3.1. Then the following holds true

$$\begin{aligned}\left\|\boldsymbol{F}\left(\hat{\boldsymbol{q}}_i,\hat{\boldsymbol{q}}_{i+1}\right)-\boldsymbol{f}\left(\boldsymbol{q}_{i+\frac{1}{2}}\right)\right\| &= \left\|\boldsymbol{F}\left(\hat{\boldsymbol{q}}_i,\hat{\boldsymbol{q}}_{i+1}\right)-\boldsymbol{F}\left(\boldsymbol{q}_{i+\frac{1}{2}},\boldsymbol{q}_{i+\frac{1}{2}}\right)\right\| \\ &\leq \left\|\boldsymbol{F}\left(\hat{\boldsymbol{q}}_i,\hat{\boldsymbol{q}}_{i+1}\right)-\boldsymbol{F}\left(\hat{\boldsymbol{q}}_i,\boldsymbol{q}_{i+\frac{1}{2}}\right)\right\| + \left\|\boldsymbol{F}\left(\hat{\boldsymbol{q}}_i,\boldsymbol{q}_{i+\frac{1}{2}}\right)-\boldsymbol{F}\left(\boldsymbol{q}_{i+\frac{1}{2}},\boldsymbol{q}_{i+\frac{1}{2}}\right)\right\| \\ &\leq C_1\left\|\hat{\boldsymbol{q}}_{i+1}-\boldsymbol{q}_{i+\frac{1}{2}}\right\| + C_2\left\|\boldsymbol{q}_{i+\frac{1}{2}}-\hat{\boldsymbol{q}}_i\right\|, \quad (3.21)\end{aligned}$$

where $C_1, C_2 \in \mathbb{R}^+$ are the Lipschitz-constants for the first and second argument of $\boldsymbol{F}$, respectively, and $\boldsymbol{q}_{i+\frac{1}{2}} := \boldsymbol{q}\left(x_{i+\frac{1}{2}}\right)$. To control the size of the cells we introduce the parameter $h > 0$ such that

$$\Delta x_i \leq h \qquad \forall i \in \mathcal{I}. \quad (3.22)$$

Since $\hat{\boldsymbol{q}}_i = \boldsymbol{q}(x_i) + \mathcal{O}(\Delta x_i) = \boldsymbol{q}(x_i) + \mathcal{O}(h)$ (If $\hat{\boldsymbol{q}} \in \mathcal{C}^2$, it is actually $\hat{\boldsymbol{q}}_i = \boldsymbol{q}(x_i) + \mathcal{O}(\Delta x_i^2)$, because cell-centered evaluation is a one-point Gaussian quadrature; see Section 3.5) and $\boldsymbol{q}(x_i) = \boldsymbol{q}_{i+\frac{1}{2}} + \mathcal{O}(\Delta x_i) = \boldsymbol{q}_{i+\frac{1}{2}} + \mathcal{O}(h)$ according to Taylor's theorem [155] (English translation in [148]), we have

$$\hat{\boldsymbol{q}}_i = \boldsymbol{q}(x_i) + \mathcal{O}(h) = \boldsymbol{q}_{i+\frac{1}{2}} + \mathcal{O}(h) \quad (3.23)$$

and, analogously, $\hat{\boldsymbol{q}}_{i+1} = \boldsymbol{q}_{i+\frac{1}{2}} + \mathcal{O}(h)$. Combining Eqs. (3.21) and (3.23) yields

$$\boldsymbol{F}\left(\hat{\boldsymbol{q}}_i,\hat{\boldsymbol{q}}_{i+1}\right) = \boldsymbol{f}\left(\boldsymbol{q}_{i+\frac{1}{2}}\right) + \mathcal{O}(h). \quad (3.24)$$

Consistency, together with the stability of a numerical method, ensures convergence under refinement of the grid (i.e., $h \to 0$ for the parameter h defined in Eq. (3.22)). This is often named the *fundamental theorem of numerical methods for PDEs*. For a discussion of this topic, including stability, we refer to [103]. Here, we only say that there are different notions of stability and many of them are formulated in terms of norms which are not allowed to increase over time (e.g., [103]) or in terms of the preservation of invariant domains (e.g., [22]). For example, if the initial data of a scalar conservation law are in the domain $[d, e] \subset \mathbb{R}$, then it is $q(\Omega, t) \subset [d, e]$ for any time $t > 0$ and this property shall be reflected by the solution obtained using a numerical method.

In the case of finite volume methods for hyperbolic conservation laws, the *Lax–Wendroff theorem* [100] comes in handy: given a consistent and conservative method

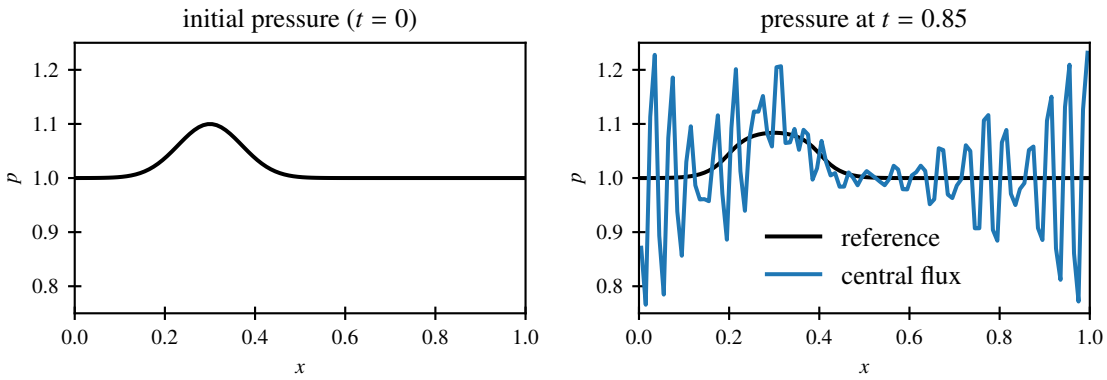


Figure 3.3: Evolution of a pressure perturbation on a constant state with zero velocity. The left panel displays the initial pressure, in the right panel the solution is shown at final time of the test. The first accurate order finite volume method using the central flux introduces oscillations indicating instability. The full description of the test case is given in Appendix A.1.

for a hyperbolic conservation law and convergence under grid refinement, this theorem guarantees that the approximate solutions actually converge towards a weak solution of the conservation law. Collecting these statements: *A stable finite volume method with a consistent numerical flux function leads to an arbitrarily good approximation of a weak solution given that the grid is sufficiently fine.* In this thesis we will not explicitly discuss the different notions of stability, instead we refer to [103].

One example of a numerical flux function that satisfies Definition 3.3.1 has been given in Section 3.2: The Godunov flux $\boldsymbol{F}^{\mathrm{God}}$ seems to be a reasonable choice, since it gives the exact interface flux for the piecewise constant discretized states. However, a Riemann problem has to be solved exactly to determine the value of $\boldsymbol{F}^{\mathrm{God}}$. Depending on the hyperbolic system which is solved numerically, finding the solution of a Riemann problem can be a challenging, and computationally expensive task. Therefore, many different *approximate Riemann solvers* have been developed. In the following we present some popular ones.

3.3.1 Central Flux

Amongst the most obvious choices for a numerical flux function satisfying Definition 3.3.1 is the *central flux*

$$\boldsymbol{F}^{\mathrm{central}}\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right) := \frac{1}{2}\left(\boldsymbol{f}\left(\boldsymbol{q}^L\right) + \boldsymbol{f}\left(\boldsymbol{q}^R\right)\right). \tag{3.25}$$

This numerical flux clearly satisfies Definition 3.3.1 because $\boldsymbol{f}$ is differentiable. However, the numerical method 3.19 with $\boldsymbol{F}^{\mathrm{central}}$ instead of $\boldsymbol{F}^{\mathrm{God}}$ turns out to be unconditionally unstable (e.g., [103]), as illustrated in Fig. 3.3. To regain stability, it is necessary to add some diffusion to the numerical flux.

3.3.2 Lax–Friedrichs and Rusanov Flux

A fast and robust choice for a jump-dependent diffusion is used in the (global) *Lax–Friedrichs flux*

$$\boldsymbol{F}^{\mathrm{LF}}\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right) := \boldsymbol{F}^{\mathrm{central}}\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right) - \frac{|\lambda|_{\max}^{\mathrm{glob}}}{2}\left(\boldsymbol{q}^R - \boldsymbol{q}^L\right), \tag{3.26}$$

where

$$|\lambda|_{\max}^{\mathrm{glob}} = \max_{i \in \mathcal{I}} \left(\left|\lambda\left(\hat{\boldsymbol{Q}}_i\right)\right|_{\max}\right) \tag{3.27}$$

with the largest eigenvalue $\left|\lambda\left(\hat{\boldsymbol{Q}}_i\right)\right|_{\max}$ of the flux Jacobian defined as in Eq. (3.11) evaluated at the state $\hat{\boldsymbol{Q}}_i$. This definition uses a global diffusion coefficient, which makes the method excessively diffusive especially in simulations, in which the wave velocities vary strongly in different regions of the domain. Improving this is quite simple: In the Rusanov [145] (or local Lax–Friedrichs) flux, the same diffusion coefficient is chosen locally

$$\boldsymbol{F}^{\mathrm{Rus}}\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right) := \boldsymbol{F}^{\mathrm{central}}\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right) - \frac{\max\left(\left|\lambda\left(\boldsymbol{q}^L\right)\right|_{\max}, \left|\lambda\left(\boldsymbol{q}^R\right)\right|_{\max}\right)}{2}\left(\boldsymbol{q}^R - \boldsymbol{q}^L\right). \tag{3.28}$$

At each interface, the absolute value of the largest eigenvalue is chosen as diffusion coefficient. With these numerical fluxes (3.20) can be shown to be stable and yield numerical solutions that converge towards the vanishing viscosity solution in the limit $h \to 0$ (e.g., [103]). In Fig. 3.4 one can see for the example of a Riemann problem test case that the Rusanov flux is significantly less diffusive than the Lax–Friedrichs flux.

Therefore, it seems preferable to use the local over the global diffusion coefficient. However, there are cases in which the global Lax-Friedrichs method has advantages. In the context of ideal magnetohydrodynamics, for example, the global Lax–Friedrichs flux applied in the scheme 3.20 exactly preserves a discretization of the divergence of the magnetic field (e.g., [65]).

3.3.3 Roe's Approximate Riemann Solver

The Rusanov flux is a reliable and accurate choice for scalar conservation laws, since it is close to realizing upwinding (see [163] for upwinding). For systems, as in Eq. (3.28), this is not the case: The Rusanov flux does not account for the more complicated wave structure arising in systems. A more accurate method has been developed in [141]. In the so-called *Roe flux*

$$\boldsymbol{F}^{\mathrm{Roe}}\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right) := \boldsymbol{F}^{\mathrm{central}}\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right) - \frac{1}{2} D^{\mathrm{Roe}}\left(\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right)^{\mathrm{Roe}}\right)\left(\boldsymbol{q}^R - \boldsymbol{q}^L\right), \tag{3.29}$$

the diffusion coefficient is not scalar but matrix valued. The diffusion is given by the matrix field

$$D^{\mathrm{Roe}}\left(\boldsymbol{q}\right) := R(\boldsymbol{q})\left|\Lambda\left(\boldsymbol{q}\right)\right| R(\boldsymbol{q})^{-1}, \tag{3.30}$$

where $R(\boldsymbol{q})$ is the matrix of right eigenvectors of the flux Jacobian $A(\boldsymbol{q})$, $\Lambda(\boldsymbol{q}) := R(\boldsymbol{q})^{-1} A(\boldsymbol{q}) R(\boldsymbol{q})$ is the diagonal matrix of the corresponding eigenvalues, and $|\cdot|$ is

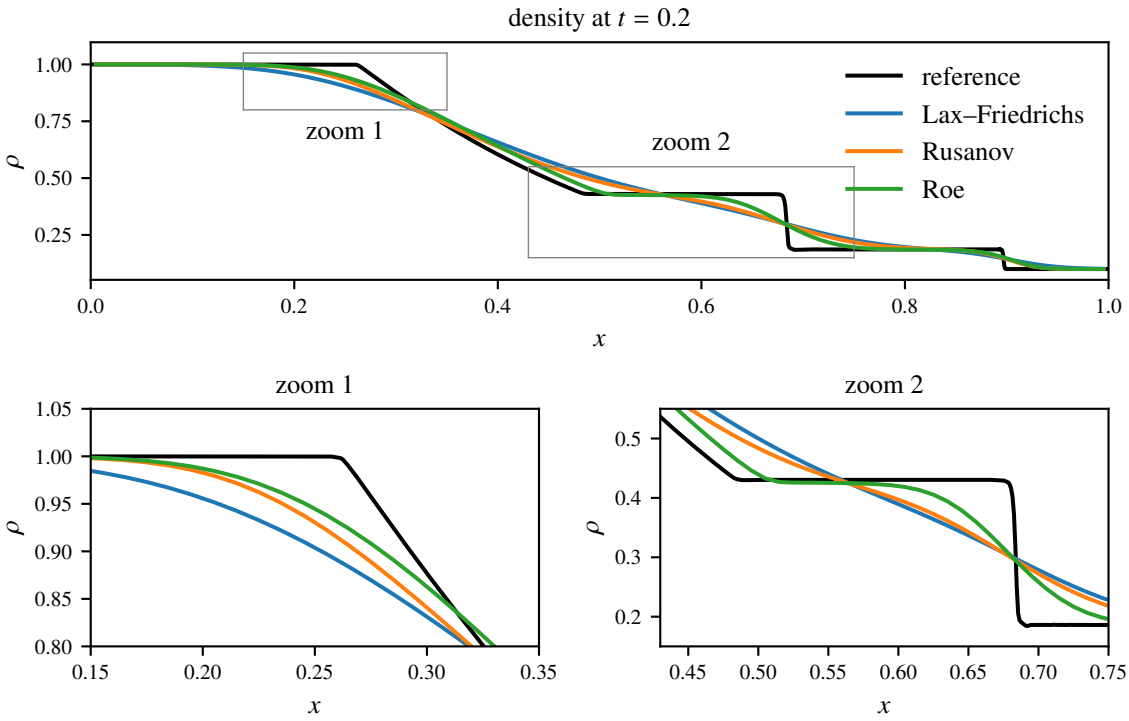


Figure 3.4: Density at final time for the Sod shock tube test case [147] using a first order accurate finite volume method with different numerical flux functions. In the top panel, the full density profile is shown, in the bottom panels certain areas in the (x, ρ)-plane are magnified as indicated by the boxes in the upper panel. Details on the test setup are given in Appendix A.2.

applied component-wise on the diagonal matrix. The *Roe average state* $\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right)^{\text{Roe}}$ is defined such that the condition

$$A\left(\left(\boldsymbol{q}^L, \boldsymbol{q}^R\right)^{\text{Roe}}\right)\left(\boldsymbol{q}^R - \boldsymbol{q}^L\right) = \boldsymbol{f}\left(\boldsymbol{q}^R\right) - \boldsymbol{f}\left(\boldsymbol{q}^L\right) \tag{3.31}$$

holds. This definition leads to a numerical flux well-suited to capture and accurately follow shock fronts as can be seen in Fig. 3.4. The Roe scheme can be interpreted as an approximate Riemann solver that assumes only shocks and no rarefactions in the solution of the Riemann problem, since it is based on linearizing the Riemann problem.

Consequently, entropy violating shocks can arise in the application of Roe's approximate Riemann solver when it is applied to the compressible Euler system (2.19). Corrections to this problem related to Euler equations have, amongst others, been suggested in [124, 125, 151, 152, 73], mainly based on assuring that the entries $|u \pm c|$ in Λ corresponding to the sonic eigenvalues can never completely vanish. Since all of the numerical experiment in this thesis are purely subsonic, we do not apply an entropy fix. The $|u|$ corresponding to the advective eigenvalue should not be modified in the same way, since its potential to become zero is important for the Roe flux's contact property

Definition 3.3.2 (Contact Property)**.** *Consider the density ρ^L (ρ^R) on the left (right) side of a contact discontinuity the constant pressure p. A numerical flux function $\boldsymbol{F}$ for the one-dimensional compressible Euler system* (2.19) *that satisfies the condition*

$$\boldsymbol{F}\left(\boldsymbol{q}^{\text{prim},L} = \left[\rho^L, 0, p\right]^T, \boldsymbol{q}^{\text{prim},R} = \left[\rho^R, 0, p\right]^T\right) = [0, p, 0]^T \tag{3.32}$$

is said to have the contact property*.*

This property ensures the ability of a numerical flux to exactly capture stationary contact discontinuities of the Euler equations. It plays a fundamental role for the well-balanced method we introduce in Section 4.4. There are numerous other numerical flux functions with the contact property, such as the well-known HLLC flux [164]. The Lax–Friedrichs and the Rusanov flux, on the other hand, do not satisfy Definition 3.3.2.

3.4 Higher Order Methods

As we have seen in Section 3.3, the discretization error of the semi-discrete equation (3.20) is of the size $\mathcal{O}(h)$ for sufficiently smooth solutions. In order to get more accurate results, the discrete data can be reconstructed in each cell.

Definition 3.4.1 (Conservative consistent reconstruction)**.** *Let $i \in \mathcal{I}$ be a grid index, $\hat{\boldsymbol{Q}}_j$ cell-averaged states for $j \in \mathcal{I}$, $\mu \in \mathbb{N}$ odd, and $x \in \Omega$. A function*

$$\begin{gathered}\mathcal{R}_i : \mathcal{C}\left(\mathbb{R} \times \left(\mathbb{R}^{\mathfrak{n}}\right)^{\mu}, \mathbb{R}^{\mathfrak{n}}\right), \\ \left(x, \hat{\boldsymbol{Q}}_{i-\frac{\mu-1}{2}}, \ldots, \hat{\boldsymbol{Q}}_{i+\frac{\mu-1}{2}}\right) \mapsto \boldsymbol{Q}_i^{\text{rec}}(x) = \mathcal{R}_i\left(x; \hat{\boldsymbol{Q}}_{i-\frac{\mu-1}{2}}, \ldots, \hat{\boldsymbol{Q}}_{i+\frac{\mu-1}{2}}\right)\end{gathered} \tag{3.33}$$

which satisfies

$$\mathcal{R}_i\left(x;\boldsymbol{q},\ldots,\boldsymbol{q}\right)=\boldsymbol{q} \tag{3.34}$$

is called consistent reconstruction*. A consistent reconstruction is called* conservative *if*

$$\frac{1}{\Delta x_i}\int_{\Omega_i}\mathcal{R}_i\left(x;\hat{\boldsymbol{Q}}_{i-\frac{\mu-1}{2}},\ldots,\hat{\boldsymbol{Q}}_{i+\frac{\mu-1}{2}}\right)dx=\hat{\boldsymbol{Q}}_i. \tag{3.35}$$

In this definition we are still ignoring the domain boundaries. However, note that additional cells on both sides of the domain are in general necessary to define the reconstruction close to the boundaries. This is discussed in Section 3.8.

Notation 3.4.2. *In this thesis we only consider conservative consistent reconstructions. The phrases* reconstruction *or* consistent reconstruction *always indicate a conservative consistent reconstruction in the rest of this thesis.*

Sometimes, we also use the phrase reconstruction *or* conservative consistent reconstruction *to refer to the mapping* $\mathcal{R}_i\left(\cdot;\hat{\boldsymbol{Q}}_{i-\frac{\mu-1}{2}},\ldots,\hat{\boldsymbol{Q}}_{i+\frac{\mu-1}{2}}\right)\mapsto\boldsymbol{Q}_i^{\mathrm{rec}}\in\mathcal{C}\left(\mathbb{R},\mathbb{R}^{\mathfrak{n}}\right)$*. For brevity, we sometimes condense some of the arguments and use the notation* $\mathcal{R}_i\left(x;\left\{\hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}\right)$ *instead of* $\mathcal{R}_i\left(x;\hat{\boldsymbol{Q}}_{i-\frac{\mu-1}{2}},\ldots,\hat{\boldsymbol{Q}}_{i+\frac{\mu-1}{2}}\right)$*, where*

$$\mathcal{S}_i=\left\{i-\frac{\mu-1}{2},\ldots,i+\frac{\mu-1}{2}\right\} \tag{3.36}$$

is the stencil *of the reconstruction.*

The purpose of a reconstruction is to find a (piecewise) approximation of a smooth function $\boldsymbol{q}$ from only knowing the cell-averages $\hat{\boldsymbol{q}}_i$, $i\in\mathcal{I}$. Condition (3.35) ensures that the cell-averages are conserved. The numerical flux function is then applied to the reconstructed interface values instead of to the cell-average values, i.e.,

$$\frac{d}{dt}\hat{\boldsymbol{Q}}_i=-\frac{1}{\Delta x_i}\left[\boldsymbol{F}\left(\boldsymbol{Q}_{i+\frac{1}{2}}^L,\boldsymbol{Q}_{i+\frac{1}{2}}^R\right)-\boldsymbol{F}\left(\boldsymbol{Q}_{i-\frac{1}{2}}^L,\boldsymbol{Q}_{i-\frac{1}{2}}^R\right)\right], \tag{3.37}$$

where

$$\boldsymbol{Q}_{i-\frac{1}{2}}^L:=\boldsymbol{Q}_{i-1}^{\mathrm{rec}}\left(x_{i-\frac{1}{2}}\right),\qquad \boldsymbol{Q}_{i-\frac{1}{2}}^R:=\boldsymbol{Q}_i^{\mathrm{rec}}\left(x_{i-\frac{1}{2}}\right), \tag{3.38}$$

$$\boldsymbol{Q}_{i+\frac{1}{2}}^L:=\boldsymbol{Q}_i^{\mathrm{rec}}\left(x_{i+\frac{1}{2}}\right),\qquad \boldsymbol{Q}_{i+\frac{1}{2}}^R:=\boldsymbol{Q}_{i+1}^{\mathrm{rec}}\left(x_{i+\frac{1}{2}}\right) \tag{3.39}$$

with the functions $\boldsymbol{Q}_j^{\mathrm{rec}}$ $(j\in\mathcal{I})$ defined as in Eq. (3.33).

Definition 3.4.3 (Order of accuracy of a reconstruction)**.** *We say that the conservative consistent reconstruction* $\mathcal{R}$ *is* m*-th order accurate if*

$$\boldsymbol{q}(x)-\mathcal{R}_i\left(x;\hat{\boldsymbol{q}}_{i-\frac{\mu-1}{2}},\ldots,\hat{\boldsymbol{q}}_{i+\frac{\mu-1}{2}}\right)=\mathcal{O}\left(h^m\right)\qquad\text{for}\qquad x\in\Omega_i \tag{3.40}$$

for the cell-averages $\hat{\boldsymbol{q}}_j$ $(j\in\mathcal{I})$ *of any function* $\boldsymbol{q}\in\mathcal{C}^m\left(\bigcup_{j\in\mathcal{S}_i},\mathbb{R}^{\mathfrak{n}}\right)$ *in the cells* Ω_j *with size* $\Delta x_j<h$*.*

Let us call the right-hand side of Eq. (3.37) *residual*

$$\mathcal{L}_i\left(\left\{\hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}, t\right) := -\frac{1}{\Delta x_i}\left[\boldsymbol{F}\left(\boldsymbol{Q}^L_{i+\frac{1}{2}}(t), \boldsymbol{Q}^R_{i+\frac{1}{2}}(t)\right) - \boldsymbol{F}\left(\boldsymbol{Q}^L_{i-\frac{1}{2}}(t), \boldsymbol{Q}^R_{i-\frac{1}{2}}(t)\right)\right]. \tag{3.41}$$

In general, the residual of a semi-discrete scheme for a hyperbolic system is defined such that we can write the scheme in the form

$$\frac{d}{dt}\hat{\boldsymbol{Q}}_i(t) = \mathcal{L}_i\left(\left\{\hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}, t\right). \tag{3.42}$$

Definition 3.4.4. *A semi-discrete scheme is* m-th order accurate *if*

$$\mathcal{L}_i\left(\{\hat{\boldsymbol{q}}_j\}_{j\in\mathcal{S}_i}, t\right) = \frac{d}{dt}\hat{\boldsymbol{q}}_i(t) + \mathcal{O}\left(h^m\right) \tag{3.43}$$

holds true in in every cell Ω_i *with size* $\Delta x_i \leq h$ *for any solution* $\boldsymbol{q} \in \mathcal{C}^m$ *of the hyperbolic system discretized by the scheme. If the full RK-FV scheme is discussed, we say the scheme is* m-th order accurate in space *to refer to the order of accuracy of the semi-discrete scheme.*

Using the Lipschitz-continuity of the consistent numerical flux in the residual $\mathcal{L}$ it is straightforward to show that the order of accuracy of a semi-discrete scheme for a hyperbolic conservation law is given by the order of accuracy of the reconstruction $\mathcal{R}$.

3.4.1 Polynomial Reconstruction

In this and the following section (Section 3.4.2) we discuss the reconstruction techniques only for scalar equations, i.e., we assume $\mathfrak{n} = 1$. All methods naturally extend to $\mathfrak{n} > 1$ by applying them component-wise.

Let m be odd. The simplest way to construct an m-th order accurate reconstruction is via the polynomial ansatz

$$Q_i^{\text{rec},\mathcal{P}m}(x) := \sum_{\kappa=0}^{m-1} a_\kappa (x - x_i)^\kappa \tag{3.44}$$

with $a_0, \ldots, a_{m-1} \in \mathbb{R}$. The coefficients can be determined uniquely by solving

$$\frac{1}{\Delta x_j}\int_{\Omega_j} Q_i^{\text{rec},\mathcal{P}m}(x)\, dx = \hat{Q}_j \tag{3.45}$$

for $j \in \mathcal{S}_i$. The stencil $\mathcal{S}_i$ of this approach is given by Eq. (3.36) with $\mu = m$. The polynomial reconstruction is visualized in Fig. 3.5 for different values of m. This simple approach is for example used in the *piecewise parabolic method* (PPM), originally developed in [46].

For even m, to uniquely determine the coefficients in the same manner, one has to choose an asymmetric stencil. Hence, polynomial reconstruction for even m is uncommon. The only exception is a second order reconstruction realized by a linear

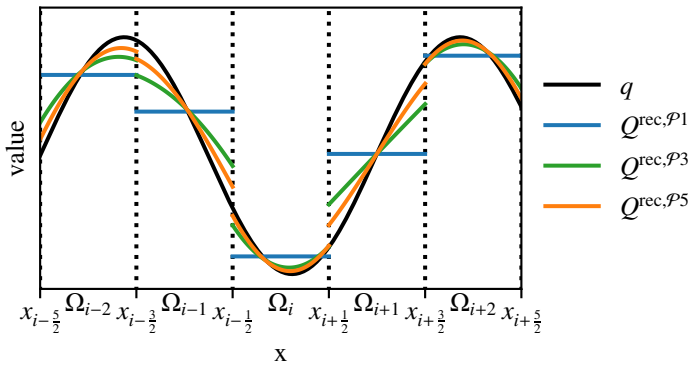


Figure 3.5: Piecewise polynomial reconstructions as introduced in Section 3.4.1 for $m = 1, 3, 5$ applied on a sine function. It gets evident, that using a higher order reconstruction leads to a more accurate approximation of the sine function and reduces the size of the interface jumps.

function $Q_i^{\text{rec},\mathcal{P}2}(x) = a_0 + a_1 x$.[3] This is usually realized by choosing the overall three-point stencil of a parabolic (third order accurate) reconstruction and return a linear combination of the linear functions obtained from the left and right biased two-cells stencils. How to combine them is discussed in the following section.

3.4.2 Limited Reconstruction

Let $m = 2$ and consider the linear[4] reconstruction functions

$$Q_i^{\text{rec},L}(x) := \hat{Q}_i + \sigma_i^L (x - x_i) \quad \text{with} \quad \sigma_i^L := 2\frac{\hat{Q}_i - \hat{Q}_{i-1}}{\Delta x_i + \Delta x_{i-1}} \quad \text{and} \tag{3.46}$$

$$Q_i^{\text{rec},R}(x) := \hat{Q}_i + \sigma_i^R (x - x_i) \quad \text{with} \quad \sigma_i^R := 2\frac{\hat{Q}_{i+1} - \hat{Q}_i}{\Delta x_{i+1} + \Delta x_i} \tag{3.47}$$

consistent with the cell-averages of the left and right biased two-cell stencil, respectively. The obvious question is: Which one of the slopes σ_i^L and σ_i^L shall be chosen for the linear reconstruction $Q_i^{\text{rec},\mathcal{P}2}$? For a moment, let us just use the arithmetic average and define

$$\mathcal{P}2_i\left(x; \hat{Q}_{i-1}, \hat{Q}_i, \hat{Q}_{i+1}\right) := Q_i^{\text{rec},\mathcal{P}2}(x) := \hat{Q}_i + \frac{\sigma_i^L + \sigma_i^R}{2}(x - x_i). \tag{3.48}$$

This choice satisfies Definition 3.4.3 for $m = 2$ and yields correspondingly accurate results on smooth solutions. In the presence of discontinuities, however, artificial extrema arise and the reconstruction introduces spurious oscillations in the numerical

[3] Second order finite volume methods are popular because they are much simpler than higher order methods. They are, for example, easier to extend to multiple spatial dimensions. Additionally, no conversion between cell-average values and cell-centered values is necessary, since theses are second order approximations of each other. This results in an easy conversion between different sets of variables. All topics mentioned here are discussed in subsequent sections and chapters.

[4] *Linear* refers to the functions $Q_i^{\text{rec},L/R} : \Omega_i \to \mathbb{R}$ and not to the reconstruction $\mathcal{R}$ (which is linear in all arguments but the first one for the unlimited reconstruction techniques presented to this point).

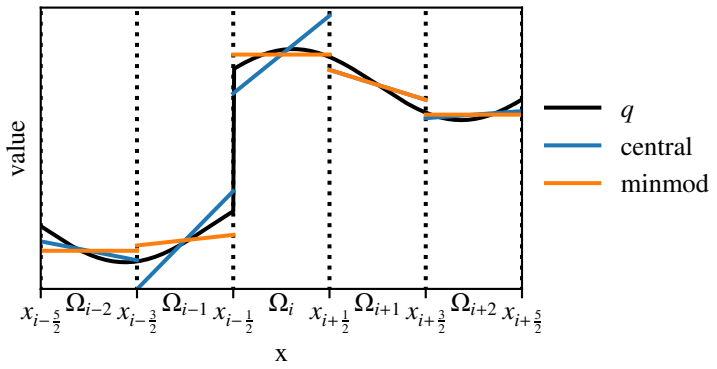


Figure 3.6: Piecewise linear reconstructions of a function with a discontinuity. Choosing the central slope (Eq. (3.48)) introduces new extrema close to the discontinuity. Applying the minmod-limiter to the slope as in Eq. (3.50) leads to a reduction of the slope close to the discontinuity such that no new extrema arise.

solution (see Fig. 3.6) such that the finite volume method is not stable in the presence of discontinuities. Consequently, *slope limiters* have been developed to choose the slope of the linear reconstruction function depending on the sizes of the jumps between the i-th cell and its neighboring cells (e.g., [103]). In this thesis we only present one of them, namely the minmod limiter

$$\operatorname{minmod}(a,b) := \begin{cases} a & \text{if } |a| < |b| \text{ and } ab > 0, \\ b & \text{if } |b| < |a| \text{ and } ab > 0, \\ 0 & \text{if } ab \leq 0, \end{cases} . \tag{3.49}$$

for $a, b \in \mathbb{R}$. We define the minmod-limited linear reconstruction by

$$\begin{aligned} \mathcal{P}2\text{minmod}_i \left(x; \hat{Q}_{i-1}, \hat{Q}_i, \hat{Q}_{i+1}\right) &:= Q_i^{\text{rec},\mathcal{P}2\text{minmod}}(x) \\ &:= \hat{Q}_i + \operatorname{minmod}\left(\sigma_i^L, \sigma_i^R\right)(x - x_i) \end{aligned} \tag{3.50}$$

with σ_i^L, σ_i^R defined in Eqs. (3.46) and (3.47). The minmod limiter (Eq. (3.49)) chooses the more moderate slope to apply it in the linear reconstruction. In case of an extremum, the slope is set to zero. This ensures that no oscillations are introduced as is visualized in Fig. 3.6.

Higher order polynomial reconstruction as introduced in Section 3.4.1 also give rise to spurious oscillations close to discontinuities. Popular approaches to limiting in this case are based on the *essentially non-oscillatory* (ENO) reconstruction strategy ([82, 81]) which is the foundation for a large class of reconstruction methods. In an m-th order accurate ENO scheme, degree $m-1$ polynomials with different stencils (e.g., some with left bias and some with right bias) are reconstructed and compared to each other regarding their total variation. The polynomial which has the lowest total variation (we refer to [103] for total variation) is chosen as reconstruction polynomial. This allows to avoid reconstruction over discontinuities while using a high order reconstruction. This strategy can still introduce spurious oscillations, although they are usually small in practice. Unfortunately, the ENO procedure leads

to a large stencil compared to the simple polynomial reconstruction introduced in Section 3.4.1.

To recover this smaller stencil, *weighted ENO* (WENO) schemes have been introduced in [115, 91] and later improved and extended in [5, 55]. Many modern WENO-type schemes (e.g., [178, 172]) reconstruct polynomials of different order using subsets of the stencil $\mathcal{S}_i = \left\{i - \frac{m-1}{2}, \ldots, i + \frac{m-1}{2}\right\}$. The reconstructed states are then given as a linear combination of all these polynomials using weights which take small values if a discontinuity is in the polynomial's reconstruction stencil. Furthermore, the weights are defined such that the WENO procedure leads to an m-th order accurate method on sufficiently smooth solutions.

Central WENO (CWENO) methods, which have been introduced in [105], are an example for such methods (for further background on CWENO methods the reader is referred to [107, 97, 47]). It can be beneficial that CWENO reconstruction yields a polynomial which is defined in the whole cell and can be extended to neighboring cells.[5] The well-balanced methods we introduce in Sections 4.4 and 4.5 are based on this property. In our numerical experiments in Section 4.6, we use the third order accurate CWENO3 reconstruction introduced in [97], the fifth order accurate CWENO5 reconstruction from [32] and the seventh order accurate CWENO7 reconstruction from [47] for limited high order reconstruction. For limited second order reconstruction we apply the minmod-limited linear reconstruction (3.50). We also apply unlimited reconstruction (as defined in Eq. (3.48) and Section 3.4.1) in order to illustrate the advantage of limiting.

3.4.2.1 Third Order Accurate Central Weighted Essentially Non-Oscillatory Reconstruction

As an example for the CWENO methods applied in the numerical experiments in this thesis, we present the CWENO3 method from [97] in the following. Assume a uniform grid with cell size Δx. For simplicity we define the coordinate $\bar{x} := (x - x_i)/\Delta x$. The construction starts with the parabolic polynomial reconstruction obtained as described in Section 3.4.1:

$$P_{\text{opt}} = a_0 + a_1\bar{x} + a_2\bar{x}^2 \tag{3.51}$$

with

$$a_0 = \hat{Q}_i - \frac{1}{24}\left(\hat{Q}_{i+1} - 2\hat{Q}_i + \hat{Q}_{i-1}\right), \tag{3.52}$$

$$a_1 = \frac{1}{2}\left(\hat{Q}_{i+1} - \hat{Q}_{i-1}\right), \qquad \text{and} \tag{3.53}$$

$$a_2 = \hat{Q}_{i+1} - 2\hat{Q}_i + \hat{Q}_{i-1}. \tag{3.54}$$

Additionally, we define the one-sided linear reconstruction polynomials

$$P_{\text{L}}(\bar{x}) := \hat{Q}_i + \left(\hat{Q}_i - \hat{Q}_{i-1}\right)\bar{x} \qquad \text{and} \tag{3.55}$$

$$P_{\text{R}}(\bar{x}) := \hat{Q}_i + \left(\hat{Q}_{i+1} - \hat{Q}_i\right)\bar{x}. \tag{3.56}$$

[5] which is not provided in the original WENO method [115]

The central parabola P_{C} is defined via

$$P_{\mathrm{C}}(\bar{x}) := \frac{1}{C_{\mathrm{C}}} \left(P_{\mathrm{opt}}(\bar{x}) - C_{\mathrm{L}} P_{\mathrm{L}}(\bar{x}) - C_{\mathrm{R}} P_{\mathrm{R}}(\bar{x}) \right) \tag{3.57}$$

with the linear weights $C_{\mathrm{L}} = C_{\mathrm{R}} = \frac{1}{4}$, $C_{\mathrm{C}} = \frac{1}{2}$ as in [97]. Note, however, that this choice is not unique. The reconstruction polynomial shall then be defined by the linear combination

$$P_{\mathrm{CWENO3}}(\bar{x}) := \omega_{\mathrm{L}} P_{\mathrm{L}}(\bar{x}) + \omega_{\mathrm{C}} P_{\mathrm{C}}(\bar{x}) + \omega_{\mathrm{R}} P_{\mathrm{R}}(\bar{x}). \tag{3.58}$$

The non-linear weights ω_k ($k \in \{\mathrm{L}, \mathrm{C}, \mathrm{R}\}$) have to satisfy $\omega_{\mathrm{L}} + \omega_{\mathrm{C}} + \omega_{\mathrm{R}} = 1$ for consistency. Additionally, $\omega_k \approx C_k$ should hold if the cell-averages $\hat{Q}_j$ ($j \in \mathcal{S}_i$) belong to a smooth solution, since the optimal polynomial P_{opt} shall be recovered in this case by the reconstruction polynomial P_{CWENO3}. In order to avoid spurious oscillations, ω_k has to take a small value if P_k varies significantly stronger in space than at least one of the remaining two polynomials on the right hand side of Eq. (3.58). As in [91, 105, 106, 97], in order to obtain the required properties, we define the non-linear-weights via

$$\omega_k := \frac{\alpha_k}{\sum_{l \in \{\mathrm{L},\mathrm{C},\mathrm{R}\}} \alpha_l} \quad \text{with} \quad \alpha_k := \frac{C_k}{(\varepsilon + IS_k)^p} \quad \text{for} \quad k \in \{\mathrm{L}, \mathrm{C}, \mathrm{R}\} \tag{3.59}$$

with the smoothness indicators

$$IS_k = \sum_{l=1}^{2} \int_{\Omega_k} \left(P_k^{(l)}(\bar{x}) \right)^2 d\bar{x} \qquad \text{for} \qquad k \in \{\mathrm{L}, \mathrm{C}, \mathrm{R}\}. \tag{3.60}$$

Following [105, 97] we choose $p = 2$, which is a value that has been determined empirically to yield good accuracy and stability. As suggested in [2] we use $\varepsilon = K \Delta x^2$ for better convergence in the presence of discontinuities compared to a constant choice of ε. It is also suggested in [2] to choose K dependent on the discrete solution $\hat{Q}_i$ ($i \in \mathcal{I}$) in order to obtain a method that is independent of the scale (e.g., of the physical dimension which is used to describe a quantity). However, in our numerical experiments we choose $K = 1$ to avoid the evaluation of a global parameter. In literature, a fixed value around $\varepsilon \approx 10^{-6}$ is often used (e.g., [105, 106], see discussion in [2]), since it seems to yield sufficiently accurate results in practice. With the choice of $K = 1$ we obtain this value for a grid with 1000 cells in the unit domain $\Omega = [0, 1]$.

3.4.3 Reconstruction Variables

Depending on the properties a method is supposed to have, different sets of variables can be reconstructed. Obviously, one can simply reconstruct conserved variables $\boldsymbol{q}^{\mathrm{cons}} = \boldsymbol{q}$. In the case of the Euler system (2.19) or (2.34), it is also common to reconstruct primitive variables $\boldsymbol{q}^{\mathrm{prim}}$ as defined in Eq. (2.24), especially, if the reconstruction shall preserve positivity of density and pressure (see [132]). *Characteristic variables* $\boldsymbol{q}^{\mathrm{char}} = R(\boldsymbol{q})^{-1} \boldsymbol{q}^{\mathrm{cons}}$ (recall that R is the matrix of right eigenvectors of

the flux Jacobian) can be reconstructed to support accurate shock capturing. Reconstruction of so-called *scaled entropy variables* can be used as a component in the construction of high order (semi-discretely) entropy stable numerical methods (e.g., [137, 138]).

To reconstruct a certain set of variables in the i-th cell, the cell-averaged conserved states $\hat{\boldsymbol{Q}}_j$ for $j \in \mathcal{S}_i$ have to be converted into cell-averaged states $\boldsymbol{q}^{\text{other}}$, where $\boldsymbol{q}^{\text{other}} = \boldsymbol{q}^{\text{prim}}, \boldsymbol{q}^{\text{char}}, \dots$ is the set of variables which is reconstructed. A sufficiently accurate conversion is easy to conduct in first and second order methods: It is

$$\begin{aligned}\hat{\boldsymbol{q}}^{\text{other}} = \widehat{T\left(\boldsymbol{q}^{\text{cons}}\right)\boldsymbol{q}^{\text{cons}}} &= T\left(\boldsymbol{q}^{\text{cons}}(x_i)\right)\boldsymbol{q}^{\text{cons}}(x_i) + \mathcal{O}\left(h^2\right) \\ &= T\left(\hat{\boldsymbol{q}}^{\text{cons}} + \mathcal{O}\left(h^2\right)\right)\left(\hat{\boldsymbol{q}}^{\text{cons}} + \mathcal{O}\left(h^2\right)\right) + \mathcal{O}\left(h^2\right) \\ &= T\left(\hat{\boldsymbol{q}}^{\text{cons}}\right)\hat{\boldsymbol{q}}^{\text{cons}} + \mathcal{O}\left(h^2\right), \quad (3.61)\end{aligned}$$

where $T(\boldsymbol{q}) := \left.\frac{\partial \boldsymbol{q}^{\text{other}}(\boldsymbol{q}^{\text{cons}})}{\partial \boldsymbol{q}^{\text{cons}}}\right|_{\boldsymbol{q}^{\text{cons}}=\boldsymbol{q}}$ is the transformation matrix between the variable systems. In Eq. (3.61) we used that cell-centered evaluation is a second order accurate Gauß quadrature (see Section 3.5) and that T is at least Lipschitz continuous[6]. A second order accurate conversion can thus be conducted by simply transforming the cell-averaged quantities like point values.

For higher order methods, the procedure is more complicated: The states have to be transformed on a number of quadrature points (see next section), which requires reconstruction to obtain sufficiently high order accurate point values. The quadrature yields then the transformed cell-average. In this thesis, however, this kind of variable transform is only applied in first and second order accurate methods.

3.5 Quadrature Rules

In higher order methods, it is often necessary to integrate data in a cell in order to obtain or convert cell-averages. In cases in which the integral cannot be computed exactly, it has to be approximated using a quadrature rule. The so-called *Gaussian quadrature rules* have been first introduced in 1815 by Gauß [71]. In 1826 Jacobi [88] introduced their present-day form based on orthogonal polynomials.

Definition 3.5.1 (Gaussian quadrature rule)**.** *Let* $\xi_1, \dots, \xi_n \in [-1, 1]$ *and* $\omega_1, \dots, \omega_n \in \mathbb{R}$ *with* $n \in \mathbb{N}$*. We call*

$$I_{x\in[-1,1]} : \mathcal{C}\left([-1,1], \mathbb{R}\right) \to \mathbb{R}, \quad (3.62)$$

$$f \mapsto I_{x\in[-1,1]}\left[f(x)\right] := \sum_{i=1}^{n} \omega_i f\left(\xi_i\right), \quad (3.63)$$

a Gaussian quadrature rule *approximating the integral*

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \quad (3.64)$$

[6]For the variable systems mentioned in this section the transformation T is smooth.

if the equality

$$I_{x\in[-1,1]}\left[p(x)\right] = \int_{-1}^{1} p(x)\,\mathrm{d}x \tag{3.65}$$

holds for any polynomial p with $\deg(p) \leq 2n-1$. The ξ_i are then called quadrature points and the ω_i quadrature weights for $i \in \{1,\ldots,n\}$.

A Gaussian quadrature rule is generalized to the interval $[a,b]$ $(a,b \in \mathbb{R}, a<b)$ using

$$I_{x\in\Omega}\left[f(x)\right] := \frac{b-a}{2}\sum_{i=1}^{n}\omega_i f\left(x_i\right) \quad \textit{with} \quad x_i := \frac{a+b}{2} + \frac{b-a}{2}\xi_i. \tag{3.66}$$

The existence of these quadrature rules is for example shown in [139]. It can also be shown that this degree of accuracy is the maximal one for a quadrature formula of the form (3.66) (e.g., [139], again).

Theorem 3.5.2. *The average value $\hat{f}$ of the function $f \in \mathcal{C}^{2n}(\Omega,\mathbb{R})$ over the interval $\Omega \subset \mathbb{R}$ is approximated by an n-point Gaussian quadrature rule with $(2n)$-th order accuracy*

Proof. Let $h := b-a$ be the length and $x_c := \frac{a+b}{2}$ the center of the interval $\Omega = [a,b]$. From Taylor's theorem we know that there exists a polynomial p_f with $\deg(p_f) = 2n-1$ such that we can decompose

$$f(x) = p_f(x) + g(x) \tag{3.67}$$

with $g(x) = \mathcal{O}\left((x-x_c)^{2n}\right)$. Applying an n-point Gaussian quadrature rule to approximate the cell-average yields

$$\hat{f} \approx \frac{1}{h} I_{x\in\Omega}\left[f(x)\right] = \frac{1}{h} I_{x\in\Omega}\left[p_f(x)\right] + \frac{1}{h} I_{x\in\Omega}\left[g(x)\right] = \hat{p_f} + \frac{1}{h} I_{x\in\Omega}\left[g(x)\right], \tag{3.68}$$

because the n-point Gaussian quadrature rule is linear in the argument (obvious from Eq. (3.66)) and exact on the polynomial p_f with $\deg\left(p_f\right) = 2n-1$. Furthermore, it is

$$\begin{aligned}\frac{1}{h} I_{x\in\Omega}\left[g(x)\right] = \frac{1}{2}\sum_{i=1}^{n}\omega_i g\left(x_c + \frac{h}{2}\xi_i\right) = \frac{1}{2}\sum_{i=1}^{n}\omega_i \mathcal{O}\left(\left(\frac{h}{2}\xi_i\right)^{2n}\right) \\ = \frac{1}{2}\sum_{i=1}^{n}\omega_i \mathcal{O}\left(h^{2n}\right) = \mathcal{O}\left(h^{2n}\right).\end{aligned} \tag{3.69}$$

On the other hand, for the exact cell-average we have

$$\hat{f} = \hat{p_f} + \hat{g} \tag{3.70}$$

and

$$\hat{g} = o\left(h^{2n}\right) \tag{3.71}$$

follows from $g(x) = \mathcal{O}\left((x-x_c)^{2n}\right) = \mathcal{O}\left(h^{2n}\right)$ because of

$$\frac{\hat{g}}{h^{2n}} \leq \frac{h\,\max_{x\in\Omega} g(x)}{h^{2n}} \to 0 \qquad \text{for} \qquad h \to 0. \tag{3.72}$$

Collecting Eqs. (3.69) to (3.71) yields

$$\hat{f} = \frac{1}{h} I_{x\in\Omega}\left[f(x)\right] + \mathcal{O}\left(h^{2n}\right). \tag{3.73}$$

□

It follows that a cell-average obtained using an n-point Gaussian quadrature rule is m-th order accurate if $n \geq \frac{m}{2}$. Throughout this thesis, the phrase *quadrature* is used analogously to *Gaussian quadrature*. It is convenient to give quadrature rules for the interval $\Omega = [-1, 1]$, since a common convention helps to uniquely define the numerical values ξ_i and ω_i of normalized quadrature points and weights for certain Gaussian quadrature rules. These quadrature rules are then extended to general intervals as described in Definition 3.5.1. Examples of Gaussian quadrature rules are Gauß-Legendre quadrature rules, which are based on Legendre polynomials (e.g., [139]). In this thesis, we only use Gauß–Legendre quadrature rules. The corresponding quadrature points and weights can be found in [135]. To vector-valued functions the quadrature rules are applied component-wise.

3.6 Source Terms

In order to develop FV methods capable of providing approximate solutions for the compressible Euler system with gravity, we add a source term to the residual (3.41)

$$\begin{aligned}\mathcal{L}_i\left(\left\{\hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}, t\right) :=& \\ -\frac{1}{\Delta x_i}\left[\boldsymbol{F}\left(\boldsymbol{Q}^L_{i+\frac{1}{2}}(t), \boldsymbol{Q}^R_{i+\frac{1}{2}}(t)\right) - \boldsymbol{F}\left(\boldsymbol{Q}^L_{i-\frac{1}{2}}(t), \boldsymbol{Q}^R_{i-\frac{1}{2}}(t)\right)\right] &+ \hat{\boldsymbol{S}}_i\left(\left\{\hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}, t\right).\end{aligned} \tag{3.74}$$

This section is concerned with finding a suitable discretization $\hat{\boldsymbol{S}}_i$ of the cell-averaged source term $\hat{\boldsymbol{s}}_i$ in the i-th cell.[7] Since source terms model non-conservative influences to the hyperbolic system, a source term discretization can be constructed without respecting certain conservation properties. This allows straightforward discretizations of source terms. There is no unique approach, and in this section we just give simple examples of general source term discretizations. The following theorem gives an important assertion.

Theorem 3.6.1. *The order of accuracy of the semi-discrete scheme*

$$\frac{d}{dt}\hat{\boldsymbol{Q}}_i(t) = \mathcal{L}_i\left(\left\{\hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}, t\right) \tag{3.75}$$

with the residual $\mathcal{L}_i$ defined in Eq. (3.74) *is given by the minimum of the order of accuracy of the reconstruction method used to obtain the interface states and the order of accuracy of the source term discretization.*

[7] For brevity, we often write $\hat{\boldsymbol{S}}_i$ or $\hat{\boldsymbol{S}}(t)$ instead of $\hat{\boldsymbol{S}}_i\left(\left\{\hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}, t\right)$ in the following.

Proof. Obviously, the errors of the interface flux approximations and of the source term discretization in the residual simply add up. □

For first and second order accuracy it is sufficient to use the simple source term discretization

$$\hat{\boldsymbol{S}}_i^{\mathrm{cc}}(t) := \hat{\boldsymbol{S}}_i^{\mathrm{cc}}\left(\left\{\hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}, t\right) := \boldsymbol{s}(\hat{\boldsymbol{Q}}_i(t), x_i, t) \tag{3.76}$$

and for m-th order accuracy with $m \geq 3$ odd, one option is to use

$$\hat{\boldsymbol{S}}_i^{\mathrm{quad}}(t) := \hat{\boldsymbol{S}}_i^{\mathrm{quad}}\left(\left\{\hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}, t\right) := \frac{1}{\Delta x} I_{x\in\Omega_i}\left[\boldsymbol{s}\left(\boldsymbol{Q}_i^{\mathrm{rec}}(x,t), x, t\right)\right], \tag{3.77}$$

where $I_{x\in\Omega_i}$ is an at least m-th order accurate quadrature rule as defined in Definition 3.5.1 and $\boldsymbol{Q}_i^{\mathrm{rec}}$ is obtained from an at least m-th order accurate reconstruction.

Theorem 3.6.2. *The source term discretizations* (3.76) *and* (3.77) *are second and m-th ($m \geq 3$ odd) order accurate, respectively, in the sense that*

$$\hat{\boldsymbol{S}}_i^{\mathrm{cc}}\left(\{\hat{\boldsymbol{q}}_i\}, t\right) = \frac{1}{\Delta x_i}\int_{\Omega_i} \boldsymbol{s}(\boldsymbol{q}(x,t), x, t)\, dx + \mathcal{O}\left(h^2\right) \tag{3.78}$$

and

$$\hat{\boldsymbol{S}}_i^{\mathrm{quad}}\left(\{\hat{\boldsymbol{q}}_j\}_{j\in\mathcal{S}_i}, t\right) = \frac{1}{\Delta x_i}\int_{\Omega_i} \boldsymbol{s}(\boldsymbol{q}(x,t), x, t)\, dx + \mathcal{O}\left(h^m\right) \tag{3.79}$$

for sufficiently smooth solutions $\boldsymbol{q}$.

Proof. Let us first consider $\hat{\boldsymbol{S}}_i^{\mathrm{quad}}$. Since both the applied quadrature rule and the reconstruction routine are at least m-th order accurate, we have

$$\begin{aligned}
\hat{\boldsymbol{S}}_i^{\mathrm{quad}}\left(\{\hat{\boldsymbol{q}}_j\}_{j\in\mathcal{S}_i}, t\right) &= \frac{1}{\Delta x_i}\int_{\Omega_i} \boldsymbol{s}(\boldsymbol{q}_i^{\mathrm{rec}}(x,t), x, t)\, dx + \mathcal{O}\left(h^m\right) \\
&= \frac{1}{\Delta x_i}\int_{\Omega_i} \boldsymbol{s}(\boldsymbol{q}(x,t), x, t) + \mathcal{O}\left(h^m\right)\, dx + \mathcal{O}\left(h^m\right) \\
&= \frac{1}{\Delta x_i}\int_{\Omega_i} \boldsymbol{s}(\boldsymbol{q}(x,t), x, t)\, dx + \mathcal{O}\left(h^m\right),
\end{aligned} \tag{3.80}$$

where $\boldsymbol{q}_i^{\mathrm{rec}}(t) = \mathcal{R}_i\left(x; \{\hat{\boldsymbol{q}}_j(t)\}_{j\in\mathcal{S}_i}\right)$ is obtained from cell-averaged values via the reconstruction. Second order accuracy of $\hat{\boldsymbol{S}}_i^{\mathrm{cc}}$ can be derived in the same manner by using linear reconstruction and a one-point Gauß-Legendre quadrature. □

3.7 Runge–Kutta Methods

The semi-discrete scheme (3.75) forms a system of $\mathfrak{n} \times N$ coupled ODEs. It can be evolved in time using standard ODE solvers. A well-known class of ODE solvers is the class of Runge–Kutta (RK) methods (e.g., [19]; for classical reference [27, 28, 90, 29, 126]).

Definition 3.7.1 (Runge–Kutta method)**.** *An* s-stage RK method *to obtain* $\left\{\hat{\boldsymbol{Q}}_i^{n+1}\right\}_{i\in\mathcal{I}}$ *from* $\left\{\hat{\boldsymbol{Q}}_i^{n}\right\}_{i\in\mathcal{I}}$ *can be written as*

$$\hat{\boldsymbol{Q}}_i^{(k)} := \hat{\boldsymbol{Q}}_i^n + \Delta t^n \sum_{l=1}^{s} a_{kl}\mathcal{L}_i\left(\left\{\hat{\boldsymbol{Q}}_j^{(l)}\right\}_{j\in\mathcal{S}_i}, t^n + c_l\Delta t^n\right), \quad k \in \{1,2,\dots,s\} \tag{3.81}$$

$$\hat{\boldsymbol{Q}}_i^{n+1} := \hat{\boldsymbol{Q}}_i^n + \Delta t^n \sum_{l=1}^{s} b_l\mathcal{L}_i\left(\left\{\hat{\boldsymbol{Q}}_j^{(l)}\right\}_{j\in\mathcal{S}_i}, t^n + c_l\Delta t^n\right) \tag{3.82}$$

with $s \in \mathbb{N}$, $A = \{a_{kl}\}_{k,l=1}^{s} \in \mathbb{R}^{s\times s}$, $\boldsymbol{b} = (b_1,\dots,b_s) \in \mathbb{R}^s$, $\boldsymbol{c} = (c_1,\dots,c_s) \in \mathbb{R}^s$, *and the time step size* $\Delta t^n = t^{n+1} - t^n$.

An s-stage RK method is characterized by the quantities A, $\boldsymbol{b}$, and $\boldsymbol{c}$. Often, a *Butcher tableau*

$$\begin{array}{c|c} \boldsymbol{c}^T & A \\ \hline & \boldsymbol{b} \end{array} = \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \tag{3.83}$$

is used to present them. The difference between explicit and implicit RK methods is briefly discussed in the following. Some examples are given. For a deeper discussion we refer the reader to [56] or [30].

3.7.1 Explicit Runge–Kutta Methods

Definition 3.7.2. *An RK method, defined in Definition 3.7.1, is called* explicit *if the matrix* A *is strictly lower triangular.*

The Butcher tableau of an explicit s-stage RK method is given by

$$\begin{array}{c|ccccc} c_1 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} . \tag{3.84}$$

The advantage of explicit RK schemes is that each stage in Eq. (3.81) can be computed independently in every cell. Explicit RK methods are computationally cheap compared to implicit RK methods (i.e., RK methods satisfying Definition 3.7.1 but not Definition 3.7.2), which require the use of iterative solvers (e.g., [30]).

A large variety of explicit RK methods with different properties has been developed in literature, also with focus on FV methods for Euler equations (e.g., [90, 170, 154, 153]). The simplest example of an explicit RK method is given by the forward Euler method, which we already applied in the FV formulation of Godunov's method in Section 3.2. The corresponding Butcher tableau is

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} . \tag{3.85}$$

The forward Euler method (RK1) is first order accurate. In the numerical tests in this thesis we also use the explicit third order RK method (RK3) from [98] given by the Butcher tableau

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ 1 & \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{2} \end{array}, \tag{3.86}$$

the explicit fifth order RK method (RK5) given by the Butcher tableau (e.g., [136])

$$\begin{array}{c|cccccc} 0 & & & & & & \\ \frac{1}{4} & \frac{1}{4} & & & & & \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & & & & \\ \frac{1}{2} & 0 & \frac{-1}{2} & 1 & & & \\ \frac{3}{4} & \frac{3}{16} & 0 & 0 & \frac{9}{16} & & \\ 1 & \frac{-3}{7} & \frac{2}{7} & \frac{12}{7} & \frac{-12}{7} & \frac{8}{7} & \\ \hline & \frac{7}{90} & 0 & \frac{32}{90} & \frac{12}{90} & \frac{32}{90} & \frac{7}{90} \end{array}, \tag{3.87}$$

and the explicit tenth order, 17-stage RK method (RK10) from [62]. The coefficients can also be found in [169].

In the first order finite volume method discussed in Section 3.2 there was a restriction to the size of the time-step given by the reasoning that waves emerging from different interfaces' Riemann problems should not meet. We can write this in the form of a *CFL-condition*

$$\Delta t = c_{\mathrm{CFL}} \min_{i\in\mathcal{I}} \left(\frac{\Delta x_i}{\max\left(|\lambda|^{\max}_{i-\frac{1}{2}}, |\lambda|^{\max}_{i+\frac{1}{2}} \right)} \right) \leq C_{\mathrm{CFL}} \min_{i\in\mathcal{I}} \left(\frac{\Delta x_i}{\max\left(|\lambda|^{\max}_{i-\frac{1}{2}}, |\lambda|^{\max}_{i+\frac{1}{2}} \right)} \right), \tag{3.88}$$

where $|\lambda|^{\max}_{i+\frac{1}{2}}$ denotes the velocity of the fastest wave emerging from the Riemann problem at the $i+\frac{1}{2}$ interface. The maximal value of C_{CFL} depends on the numerical scheme. For the first order scheme from Section 3.2 the time-step is restricted by $C_{\mathrm{CFL}} = \frac{1}{2}$. More general, in a RK-FV method evolved in time using the forward Euler method, Eq. (3.88) with $C_{\mathrm{CFL}} = \frac{1}{2}$ is a necessary condition for stability (e.g., [103]). For other RK-methods, the value for C_{CFL} can be different. For example for the RK3 method we use in this article [98], we have $C_{\mathrm{CFL}} = 1$. For a discussion of stability of semi-discrete schemes evolved in time using RK methods we refer to [90, 108]. The parameter c_{CFL} is included here, since Eq. (3.88) shows a typical way, in which the time-step is actually determined in a numerical code. In the numerical experiments in this thesis, we use $c_{\mathrm{CFL}} = 0.4$ for all methods except RK3, where we choose $c_{\mathrm{CFL}} = 0.9$.

3.8 Boundary Conditions

A hyperbolic PDE on a bounded domain not only requires initial conditions, but also boundary conditions. Since FV methods evolve solutions on a bounded domain, there is the need for numerical boundary conditions which mimic the analytical boundary conditions given for the PDE. There are different techniques to realize this. The most common approaches either define fluxes at the domain boundaries (possibly taking the numerical data inside the domain into account) or add *ghost cells*. Especially for higher order finite volume methods, ghost cell boundaries yield the advantage that the scheme needs not to be modified close to the boundaries. The ghost cell technique is described in the following.

Let us add N_{gc} ghost cells

$$\Omega_0, \dots, \Omega_{-N_{\text{gc}}+1} \qquad \text{and} \qquad \Omega_{N+1}, \dots, \Omega_{N+N_{\text{gc}}}$$

adjacent to each domain boundary. We assign data

$$\hat{\boldsymbol{Q}}_i \qquad \text{for} \qquad i \in \{-N_{\text{gc}}+1, \dots, -0, N+1, \dots, N+N_{\text{gc}}\}$$

to the corresponding ghost cells which are used to compute the interface fluxes on interfaces close to the domain boundary in the same way they are computed inside the domain. The number N_{gc} of ghost cells is chosen depending on the stencil of the reconstruction such that all interface fluxes, including $\boldsymbol{F}_{\frac{1}{2}}$ and $\boldsymbol{F}_{N+\frac{1}{2}}$, can be computed without modification of the scheme. We additionally make the assumption that $N > 2N_{\text{gc}}$, which is no relevant restriction in practice. In the following we describe possible ways to fill the ghost cells with data. For a deeper exploration of this topic we refer to [83].

Periodic boundary conditions To handle periodic boundary conditions, the states in the ghost cells are set to the values

$$\hat{\boldsymbol{Q}}_i = \hat{\boldsymbol{Q}}_{((i-1) \mod N)+1}. \tag{3.89}$$

Dirichlet boundary conditions Dirichlet boundary conditions are usually mimicked in a straightforward way: The prescribed solution is directly written into the ghost cells. However, whereas Dirichlet boundary conditions for a hyperbolic system give the solution exclusively at the boundary ($\boldsymbol{q}(a,t) = \boldsymbol{q}_{\text{lower boundary}}(t)$, $\boldsymbol{q}(b,t) = \boldsymbol{q}_{\text{upper boundary}}(t)$), numerical Dirichlet boundary conditions give information about the solution in the ghost cell layers $\left[x_{-N_{\text{gc}}+\frac{1}{2}}, x_{\frac{1}{2}}\right]$ and $\left[x_{N+\frac{1}{2}}, x_{N+N_{\text{gc}}+\frac{1}{2}}\right]$. In practice, it is in many cases clear how to set these data. Even though in many of our numerical tests it is practical to follow this approach, in general, it is not a suitable choice for hyperbolic systems: Due to their characteristic structure, waves do not always require information from the boundaries, but some waves also travel into the boundaries. Since this is not taken into account, Dirichlet boundary conditions can effectively reflect waves and thus introduce unphysical artifacts.

Wall boundary conditions Wall boundary conditions are used to mimic solid walls which enclose the domain. Their setup depends on the system to be solved. The basic idea is that the states are mirrored at the boundaries. For Euler equations, for example, wall boundaries are given by setting

$$\hat{\boldsymbol{Q}}_{1-\iota} = \begin{pmatrix} \hat{\rho}_\iota \\ -(\hat{\rho u})_\iota \\ \hat{E}_\iota \end{pmatrix} \qquad \text{and} \qquad \hat{\boldsymbol{Q}}_{N+\iota} = \begin{pmatrix} \hat{\rho}_{N+1-\iota} \\ -(\hat{\rho u})_{N+1-\iota} \\ \hat{E}_{N+1-\iota} \end{pmatrix} \tag{3.90}$$

for $\iota \in \{1, \dots, N_{\mathrm{gc}}\}$.

Extrapolation boundary conditions A simple approach for free boundary conditions that support high order accuracy is to extrapolate the reconstructed states from the first and the N-th cell to the ghost cells, i.e., we set

$$\hat{\boldsymbol{Q}}_{1-\iota} = \frac{1}{\Delta x_{1-\iota}} \int_{\Omega_{1-\iota}} \boldsymbol{Q}_1^{\mathrm{rec}}(x)\, dx \quad \text{and} \quad \hat{\boldsymbol{Q}}_{N+\iota} = \frac{1}{\Delta x_{N+\iota}} \int_{\Omega_{N+\iota}} \boldsymbol{Q}_N^{\mathrm{rec}}(x)\, dx \tag{3.91}$$

for $\iota \in \{1, \dots, N_{\mathrm{gc}}\}$. Note that these boundary conditions in practice reflect some waves. For correct outflow boundaries one has to define outflow fluxes at the boundaries (e.g., [83]).

Chapter 4

Well-Balanced Finite Volume Methods in One Spatial Dimension

As described in the previous chapter, FV methods are well-suited to approximate solutions of systems of hyperbolic balance laws. However, in some cases the results can be significantly improved by making the numerical method exact on certain relevant stationary solutions.

The shallow-water system with non-flat bottom topography (see [103]), which models water height and height-averaged velocity for waters like rivers, lakes, and oceans, admit static solutions which describe resting waters with flat surface, the so-called lake-at-rest solutions. Standard FV methods developed using the techniques described in Chapter 3 introduce spatial discretization errors leading to spurious flows. This makes it hard to resolve small perturbations and flows which are dominated by the discretization error. Also, long-time simulations are much less reliable: Even if a fine grid is used, discretization errors add up and eventually corrupt the approximate solution. To cure this, so-called *well-balanced* methods have been developed, i.e., methods that are free of a discretization error on the considered stationary solution. There is a rich literature (e.g., [18, 102, 25, 3], and references therein) on methods which are well-balanced for the lake-at-rest solution including high order methods (e.g., [122, 109, 173]). A third order accurate well-balanced active flux method for shallow water equations has been developed in [6]. Other steady states of the shallow water system include velocities and well-balanced methods capable of exactly preserving these have been developed in [123, 119] and references therein. The importance of numerical methods capable of exactly maintaining non-static stationary states has been pointed out in [176]. High order methods for shallow water equations on non-flat manifolds have been developed in [37]. These methods can take the earth's surface geometry into account and are thus suitable for tsunami simulations. The two-layer and multi-layer shallow-water models have been developed to provide some amount of vertical resolution of the flows. Since they also admit the lake-at-rest solution and other, non-static, stationary solutions, well-balanced methods for these models have been developed, e.g., in [23, 116]. In [166, 52, 24] (and references therein), well-balanced schemes for the related Ripa model have been introduced.

For the compressible Euler system with gravity, hydrostatic states are given via

a differential equation as we have seen in Section 2.4.1. Due to this, hydrostatic solutions are in general not unique, which adds an additional challenge to the development of well-balanced methods for this system compared to the shallow water system and related models, in which the static solutions can be described in the form of an algebraic relation.[1] Therefore, a strategy to choose the hydrostatic solution which shall be well-balanced is required. In the following we try to categorize different well-balanced methods for the Euler system with gravity into three categories, knowing that this distinction is neither rigorous nor unique. However, it can help to gain some understanding regarding the different approaches.

First approach: well-balancing classes of hydrostatic solutions The first – and probably most classical – approach is to choose a certain class of hydrostatic solutions and construct numerical methods that well-balance these exactly. Examples for this approach are [34, 101, 104, 41, 72, 167, 17, 43, 69] (and references therein). Higher order methods of this type are [175, 68]. Most of these methods are constructed to well-balance isothermal (see Eq. (2.36)), polytropic (see Eq. (2.37)), or isentropic (i.e., solutions with constant entropy) hydrostatic solutions of the compressible Euler system with gravity closed by an ideal gas law. A special case is [70], since this method is not designed to balance hydrostatic solutions but a class of stationary solutions based on a balance between gravity and the centrifugal force which appears as a source term if the Euler system is transformed to polar coordinates.

Second approach: well-balancing solutions known a priori While assuming an ideal gas EoS and a certain structure of the hydrostatic solution is suitable for some applications, it is not sufficient for others: In astrophysical simulations of the interior of stellar objects, e.g., complex EoS have to be used including different physical effects from classical, quantum, and relativistic physical theories (see [160]). On the other hand, the underlying hydrostatic state of the star is often known a priori in these simulations. For this type of application, the following approach can be the most suitable: A well-balanced method is constructed that balances each hydrostatic state, that is given to it explicitly, exactly. This approach allows to even balance hydrostatic states that are not given in a closed form but as discrete data[2] and it furthermore admits arbitrary EoS while still guaranteeing exact preservation of the hydrostatic state. Second order methods of this type have been introduced in [13, 15, 156] and higher order methods in [96, 14, 36]. Similar techniques can be found in the context of numerical atmospheric modeling (e.g. [21, 74, 72]). The well-balanced methods from [96, 14] are also capable of exactly preserving a priori known non-static stationary states of the Euler system with gravity. The methods in [14, 36] are developed for general hyperbolic balance laws (one-dimensional in [36], multi-dimensional in [14]) and are thus not restricted to Euler equations. The method [14] is the most general method of this type, since it allows to exactly follow

[1] Developing high order well-balanced methods for the shallow-water system, on the other hand, includes challenges like a proper treatment of wet/dry (non-vacuum/vacuum) fronts inside cells, which are not similarly relevant in the Euler system.

[2] Recall the example from Chapter 1, in which hydrostatic profiles for stars are obtained from stellar evolution codes.

any a priori known solution of any multi-dimensional system of hyperbolic balance laws. This is one of the methods we discuss and test in this thesis in detail. The second order method from [15] is also presented in this thesis, since the idea of [14] originates in this method.

Third approach: well-balancing approximations to hydrostatic solutions If the restrictions of the first approach can not be met but the solution which shall be well-balanced is not known a priori, the third approach can yield the method of choice. Hydrostatic solutions are in some way approximated from the cell-averaged states in each step of the numerical method and this approximation is well-balanced. Well-balanced methods of this type have been developed in [92, 94, 171, 78] (second order approximations) and [63, 12] (high order approximations) for the full Euler system with gravity source term. In this thesis we present two methods from [12]. The well-balanced methods following this approach are usually very flexibility, but the well-balancing is not exact in general. For the methods in [92, 94], however, cases can be identified in which the approximation of the hydrostatic state coincide with the exact hydrostatic state. In this sense, they can also be seen as methods of the first approach. On the other hand, assuming a certain type of hydrostatic state in each cell gives an approximation of the actual hydrostatic state. In that sense all the methods following the first approach can be seen as methods from the third approach. The main difference is then that the methods [92, 94] can be applied for general EoS. Recently, a well-balanced active flux method for a linear system with gravity source term has been developed in [7] which follows the approach described in this paragraph. The hydrostatic states of this system are the same as for the compressible Euler system with gravity.

The scientific contribution of the methods discussed in this thesis In this thesis we discuss four well-balanced methods. The first one, which we refer to as α-β method (Sections 4.2 and 6.1), has already been introduced in the master thesis [11] and published in [13, 15]. It is, to the author's knowledge, the first well-balanced finite volume method for the compressible Euler equations that was capable of balancing any hydrostatic state exactly. Based on the same idea, the Deviation method [14] has been constructed (Sections 4.3 and 6.2). This simple well-balancing modification enables a finite volume scheme to exactly follow any given solution of any multi-dimensional system of hyperbolic conservation or balance laws. Thus, it proposes a unified approach to well-balancing, that can be followed even for newly developed models, since it does not exploit any specific structure besides the finite volume structure. To the author's knowledge, the Deviation method is the most general well-balanced method for hyperbolic systems that is available in literature. The method introduced in [130] by Pareschi exactly balances any stationary solution in any semi-discrete numerical method for a PDE. Our Deviation method can, for hyperbolic systems, additionally follow time-dependent solution. Also, since our method only reconstructs deviations instead of the full solution, it significantly reduces diffusion close the target solution which is chosen to be followed exactly. The α-β and the Deviation method both follow the second approach as discussed above.

In order to well-balance general hydrostatic solutions of the compressible Euler

system without any a priori knowledge regarding their structure, the Discretely Well-Balanced and Local Approximation methods (Sections 4.4, 4.5 and 6.3) have been developed in [12] as methods following the third approach. These methods can be seen as the next generalization step in the series of the methods [92, 93, 94, 78]. The Discretely Well-Balanced and Local Approximation method are high order accurate, such as [78]. The local hydrostatic approximation which is balanced in [78] is based on the assumption of constant entropy, which we overcome in the Discretely Well-Balanced and Local Approximation method.

Techniques to achieve well-balancing Different techniques have been applied to achieve the well-balanced property for a numerical method. Well-balanced methods based on relaxation have been developed, e.g., in [50, 51, 156]. The path-conservative approach to well-balancing has been applied in [129, 35, 70] and references therein.

A classical way to achieve well-balancing is using the *equilibrium preserving reconstruction* technique. [3] is an early method of this type for the shallow water system (*surface reconstruction*). For Euler equations, this technique has for example been used in [13, 15, 41, 43, 72, 96, 167] (*hydrostatic reconstruction*). Since all of the methods we discuss in detail in this thesis ([15, 14, 16]) are based on equilibrium preserving reconstructions, we explain this concept in the following section.

4.1 Equilibrium Preserving Reconstruction

In this section we discuss the *equilibrium preserving reconstruction* technique, which can be used to construct well-balanced methods. In the context of the lake-at-rest solution of the shallow-water equations, this approach is called *surface reconstruction*, since basically the water surface level is reconstructed instead of the water height. In the context of Euler equations it is usually called *hydrostatic reconstruction* if it is applied to reconstruct deviations from a hydrostatic state. First, in Section 4.1.1 we present the basic principle using the example of a scalar balance law. Then, in Section 4.1.2, we discuss the question, how it can be applied to the one-dimensional compressible Euler system with gravity source term.

4.1.1 The Basic Idea of an Equilibrium Preserving Reconstruction

Let us explain the basic idea for a scalar balance law

$$\partial_t q(x,t) + \partial_x f(q(x,t)) = s(q(x,t),x). \tag{4.1}$$

Assume the balance law in (4.1) has a smooth stationary solution q^{eq}, i.e., the relation

$$\partial_x f(q^{\mathrm{eq}}(x)) = s(q^{\mathrm{eq}}(x),x) \tag{4.2}$$

holds. Now, we want to construct a semi-discrete scheme

$$\frac{d}{dt}\hat{Q}_i = -\frac{1}{\Delta x_i}\left[F\left(Q^L_{i+\frac{1}{2}}(t), Q^R_{i+\frac{1}{2}}(t)\right) - F\left(Q^L_{i-\frac{1}{2}}(t), Q^R_{i-\frac{1}{2}}(t)\right)\right] + \hat{S}_i \tag{4.3}$$

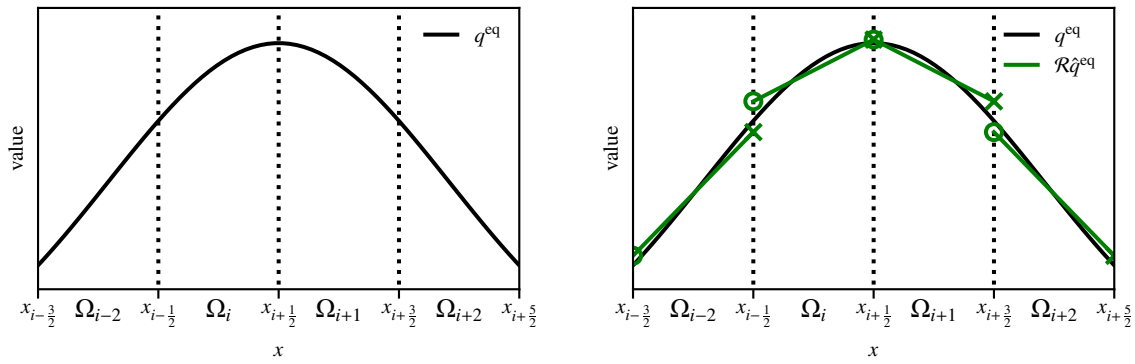


Figure 4.1: The stationary solution q^{eq} (left panel) is reconstructed using a linear reconstruction directly on the cell averages $\hat{q}^{\text{eq}}$ (green lines in the right panel). Green crosses and circles denote interface values left and right of the interface.

that satisfies the discrete equivalent of Eq. (4.2) for the cell-averages $\hat{q}^{\text{eq}}$ of the stationary solution.

Standard reconstruction A standard method can in general not maintain the stationary solution exactly, since the discretization introduces spurious errors. The numerical flux, for example, is in general different from the physical flux. Due to the consistency condition Definition 3.3.1 (i), however, the physical flux is recovered if the interface values, at which the numerical flux function is computed, coincide. This can usually not be achieved with a standard reconstruction. Let us denote this reconstruction with $\mathcal{R}$. The reconstruction $\mathcal{R}\hat{q}^{\text{eq}}$ of the cell-averaged stationary solution is visualized in Fig. 4.1 for the example of the piecewise linear reconstruction defined in Eq. (3.48).

Equilibrium preserving reconstruction An equilibrium preserving reconstruction is designed for this purpose. Let $\mathcal{T} = \mathcal{T}(x) : \mathbb{R} \to \mathbb{R}$ be a potentially space-dependent transformation, that transforms $\boldsymbol{q}^{\text{eq}}$ to a constant function, i.e., $\mathcal{T}\boldsymbol{q}^{\text{eq}} \equiv c$ for some $c \in \mathbb{R}$. The equilibrium preserving reconstruction is then defined as $\mathcal{R}^{\text{eq}} := \mathcal{T}^{-1}\mathcal{R}\mathcal{T}$. In Fig. 4.2, the reconstruction process $\mathcal{R}^{\text{eq}}\hat{q}^{\text{eq}}$ is visualized.[3] Note that the reconstruction of the constant states $\mathcal{T}q^{\text{eq}}$ is exact due to the consistency condition (Eq. (3.34)) for reconstruction methods. This leads to $\mathcal{R}^{\text{eq}}q^{\text{eq}} = q^{\text{eq}}$. Hence, due to $F\left(q^{\text{eq}}\left(x_{i+\frac{1}{2}}\right), q^{\text{eq}}\left(x_{i+\frac{1}{2}}\right)\right) = f\left(q^{\text{eq}}\left(x_{i+\frac{1}{2}}\right)\right)$, the scheme (4.3) reduces to

$$\frac{d}{dt}\hat{q}_i^{\text{eq}} = -\frac{1}{\Delta x_i}\left[f\left(q^{\text{eq}}\left(x_{i+\frac{1}{2}}\right)\right) - f\left(q^{\text{eq}}\left(x_{i-\frac{1}{2}}\right)\right)\right] + \hat{S}_i \tag{4.4}$$

if it is applied to the equilibrium state.

[3] Let us for simplicity of the notation just define $\mathcal{T}\hat{q} := (\hat{\mathcal{T}}q)$ without adding a new notation for the discrete transformation. For an actual application of the method, a different definition has to be used.

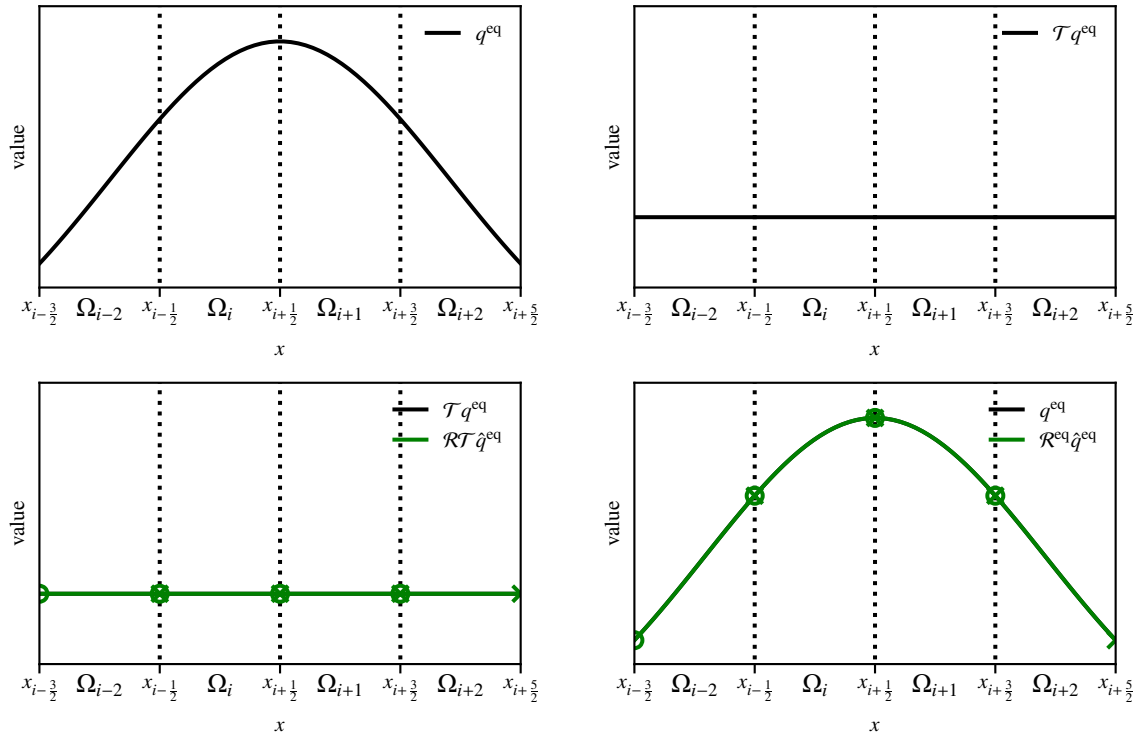


Figure 4.2: The stationary solution q^{eq} (top left panel) is transformed to the equilibrium variable (top right panel). A linear reconstruction is applied in equilibrium variables (green lines in bottom left panel) and transformed back to the standard description (bottom right panel). Green crosses and circles denote interface values left and right of the interface.

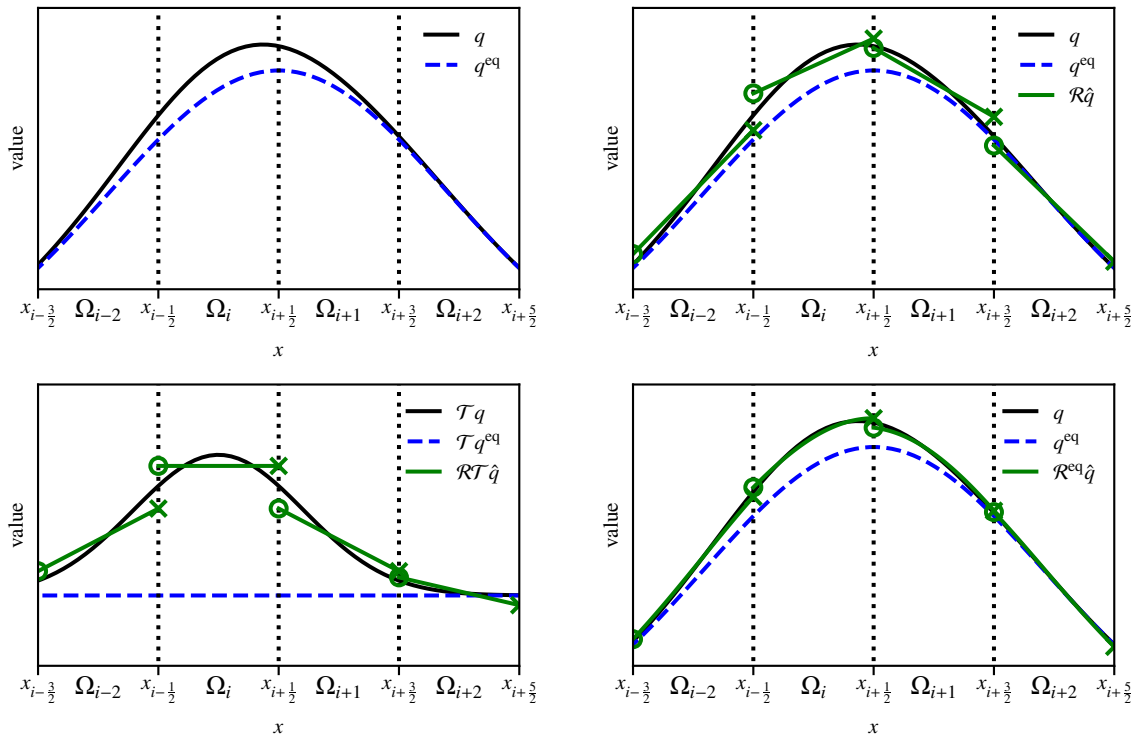


Figure 4.3: Top left: A smooth perturbation is added to the stationary state. The resulting function is just called q for brevity. Top right: Result of a linear reconstruction directly applied to $\hat{q}$. Bottom left: Linear reconstruction applied to the cell-averaged perturbation $\mathcal{T}\hat{q}$. Bottom right: Result of the equilibrium preserving reconstruction $\mathcal{R}^{\text{eq}}\hat{q} = \mathcal{T}^{-1}\mathcal{R}\mathcal{T}\hat{q}$. Green crosses and circles denote interface values left and right of the interface.

Source term The only thing remaining is to define the source term such that it cancels the fluxes at the stationary state, i.e.,

$$\hat{S}\left(\hat{q}^{\text{eq}}\right) = \frac{1}{\Delta x_i}\left[f\left(q^{\text{eq}}\left(x_{i+\frac{1}{2}}\right)\right) - f\left(q^{\text{eq}}\left(x_{i-\frac{1}{2}}\right)\right)\right], \tag{4.5}$$

which leads to

$$\frac{d}{dt}\hat{q}_i^{\text{eq}} = 0 \tag{4.6}$$

for the numerical method. Hence, the method is *well-balanced*, i.e., it maintains the stationary solution without discretization error.

Evolving perturbations The well-balanced property, as described above, does not seem very exciting: For stationary solutions, we design the method to do nothing. To do nothing, a numerical method is not necessary. However, the idea behind well-balancing is different: Often, small perturbations to a stationary solution have to be evolved in time. If a standard method is used, the discretization errors on the stationary state might be larger than the perturbation that is considered. We expect the well-balanced method to be more accurate on perturbations than the standard method. In Fig. 4.3 we demonstrate the difference between the standard

reconstruction and the equilibrium preserving reconstruction, if there is a perturbation on the stationary state. Note, that the equilibrium preserving reconstruction is more accurate on the perturbation. The smaller the (smooth) perturbation is, the stronger this effect can be expected to be.

In Figs. 4.1 to 4.3 we used the linear reconstruction define in Eq. (3.48) and the simple transformation $\mathcal{T}q = q - q^{\mathrm{eq}}$.

4.1.2 Hydrostatic Reconstruction for Euler Equations with Gravity

If this technique shall be applied to the compressible Euler equations with gravity source term in order to balance hydrostatic states, a hydrostatic reconstruction has to be applied at least to the gas pressure. Hydrostatic solutions – speaking in primitive variables – only involve density and pressure, while the velocity is zero anyway. Discontinuities in the density can be admitted in the hydrostatic reconstruction, if the numerical flux which is used in the method satisfies the contact property Definition 3.3.2. To also use numerical fluxes which do not satisfy the contact property, also the density has to be reconstructed in a hydrostatic way. In this thesis we discuss four different well-balanced methods in Sections 4.2 to 4.5. The α-β method (Section 4.2) and the Deviation method (Section 4.3) reconstruct as well pressure as density in a hydrostatic way. The Discretely Well-Balanced (Section 4.4) and Local Approximation method (Section 4.5) only use a hydrostatic reconstruction for the pressure, which means that they rely on the contact property of the numerical flux that is applied.

4.2 The α-β Method

The well-balanced method discussed in this section has been developed in the master thesis [11] and published in [13]. We repeat it here, since the method introduced in Section 4.3 is based on a modification of this approach that allows for a higher versatility and higher order of accuracy. In the discussion of the α-β method in this thesis we add new details such as a formal proof of the second order accuracy of the method.

4.2.1 Description of the α-β Method

Consider a hydrostatic solution $(\rho, u, p) = (\alpha, 0, \beta)$ described by the functions $\alpha : \Omega \to \mathbb{R}^+$ and $\beta \in \mathcal{C}^1(\Omega, \mathbb{R}^+)$, which satisfy the hydrostatic equation (2.35), i.e.,

$$\beta'(x) = \alpha(x)g(x) \qquad \text{or} \qquad g(x) = \frac{\beta'(x)}{\alpha(x)}. \tag{4.7}$$

In the numerical method Eq. (3.75) for the one-dimensional compressible Euler equations with gravity source term (2.34) we apply the following modifications:

Reconstruction For each $x \in \Omega$ we define the transformation

$$\mathcal{T}_x^{\alpha\text{-}\beta} : \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+, \tag{4.8}$$

$$\mathcal{T}_x^{\alpha\text{-}\beta}(\boldsymbol{q}) := \begin{pmatrix} \frac{1}{\alpha(x)} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{\beta(x)} \end{pmatrix} \left.\frac{\partial \boldsymbol{q}^{\text{prim}}}{\partial \boldsymbol{q}^{\text{cons}}}\right|_{\boldsymbol{q}}. \tag{4.9}$$

Let $\mathcal{R}$ be a consistent reconstruction as defined in Definition 3.4.1. The values at the $i+\frac{1}{2}$ interface are then obtained by the reconstruction $\mathcal{R}^{\alpha\text{-}\beta} := \left(\mathcal{T}^{\alpha\text{-}\beta}\right)^{-1} \mathcal{R} \mathcal{T}^{\alpha\text{-}\beta}$, i.e.,

$$\boldsymbol{Q}^L_{i+\frac{1}{2}} := \left(\mathcal{T}^{\alpha\text{-}\beta}_{x_{i+\frac{1}{2}}}\right)^{-1} \left(\mathcal{R}_i \left(x_{i+\frac{1}{2}}; \left\{ \mathcal{T}^{\alpha\text{-}\beta}_{x_j} \left(\hat{\boldsymbol{Q}}_j \right) \right\}_{j \in \mathcal{S}_i} \right) \right) \tag{4.10}$$

$$\boldsymbol{Q}^R_{i+\frac{1}{2}} := \left(\mathcal{T}^{\alpha\text{-}\beta}_{x_{i+\frac{1}{2}}}\right)^{-1} \left(\mathcal{R}_{i+1} \left(x_{i+\frac{1}{2}}; \left\{ \mathcal{T}^{\alpha\text{-}\beta}_{x_j} \left(\hat{\boldsymbol{Q}}_j \right) \right\}_{j \in \mathcal{S}_{i+1}} \right) \right) \tag{4.11}$$

Remark 4.2.1. *The transformation applied to conserved quantities yields*

$$\mathcal{T}_x^{\alpha\text{-}\beta} \left((\rho(x), \rho u(x), E(x))^T \right) = \left(\frac{\rho(x)}{\alpha(x)}, u(x), \frac{p(x)}{\beta(x)} \right)^T. \tag{4.12}$$

The conserved quantities $\boldsymbol{q}^{\text{hs}}$ belonging to the hydrostatic solution $(\rho, u, p) = (\alpha, 0, \beta)$ are mapped to the constant state

$$\mathcal{T}_x^{\alpha\text{-}\beta} \left(\boldsymbol{q}^{\text{hs}}(x) \right) = (1, 0, 1)^T \tag{4.13}$$

by the transformation. Hence, the reconstruction $\left(\mathcal{T}^{\alpha\text{-}\beta}\right)^{-1} \mathcal{R} \mathcal{T}^{\alpha\text{-}\beta}$ is an equilibrium reconstruction as explained in Section 4.1.1.[4]

Source term discretization Using Eq. (4.7), we can write the source term of the momentum equation as

$$s^{\rho u}(x,t) = \frac{\beta'(x)}{\alpha(x)} \rho(x,t). \tag{4.14}$$

To discretize Eq. (4.14) in the i-th cell we choose

$$\hat{S}_i^{\rho u, \alpha\text{-}\beta}(t) := \frac{\beta_{i+\frac{1}{2}} - \beta_{i-\frac{1}{2}}}{\Delta x_i} \frac{\hat{\rho}_i(t)}{\alpha_i} = \frac{1}{\Delta x_i} \int_{\Omega_i} s^{\rho u}(x,t)\, dx + \mathcal{O}\left((\Delta x)^2 \right), \tag{4.15}$$

where α_i, β_i are the cell-centered values of α, β and $\alpha_{i+\frac{1}{2}}, \beta_{i+\frac{1}{2}}$ are the interface values at the $i+\frac{1}{2}$ interface. The approximation of the source term cell average in then

$$\hat{\boldsymbol{S}}_i^{\alpha\text{-}\beta}(t) := \begin{pmatrix} 0 \\ \hat{S}_i^{\rho u, \alpha\text{-}\beta}(t) \\ \frac{\hat{\rho u}_i}{\hat{\rho}_i} \hat{S}_i^{\rho u, \alpha\text{-}\beta}(t) \end{pmatrix}, \tag{4.16}$$

which is a second order accurate discretization of the cell-averaged source term.

[4] In Section 4.1.1, an equilibrium reconstruction has been explained for scalar states. However, the extension to systems seems to be quite obvious.

4.2.2 Properties of the α-β Method

We now state the basic results on the accuracy and well-balanced property of the α-β method described above.

4.2.2.1 Order of Accuracy

Theorem 4.2.2. *Consider the finite volume scheme* (3.75)*, for which the interface values have been obtained with a reconstruction as described in Eqs.* (4.10) *and* (4.11) *and with the source term discretization as defined in Eq.* (4.16) *based on the functions* $\alpha \in \mathcal{C}^2(\Omega, \mathbb{R}^+)$ *and* $\beta \in \mathcal{C}^3(\Omega, \mathbb{R}^+)$*. This semi-discrete scheme is second order accurate in space, if the reconstruction* $\mathcal{R}$ *underlying the hydrostatic reconstruction* $\mathcal{R}^{\alpha\text{-}\beta}$ *is second order accurate.*

Proof. Let $\boldsymbol{q}$ be a sufficiently smooth solution. First, we show that the source term approximation $\hat{\boldsymbol{S}}_i^{\alpha\text{-}\beta}$ is second order accurate. We have

$$\frac{1}{\Delta x_i} \int_{\Omega_i} \hat{s}^{\rho u}(\boldsymbol{q}(x,t), x)\, dx = s^{\rho u}(\boldsymbol{q}(x_i,t), x) + \mathcal{O}\left(h^2\right) \tag{4.17}$$

and

$$\begin{aligned} s^{\rho u}(\boldsymbol{q}(x_i,t), x) = -\rho(x_i,t) g(x_i) &\overset{Eq.\ (4.7)}{=} \rho(x_i,t) \frac{\beta'(x_i)}{\alpha\,(x_i)} \\ &= \hat{\rho}_i(t) \frac{\beta_{i+\frac{1}{2}} - \beta_{i-\frac{1}{2}}}{\alpha_i} + \mathcal{O}\left(h^2\right). \end{aligned} \tag{4.18}$$

Combining Eqs. (4.17) and (4.18) yields

$$\hat{S}_i^{\rho u, \alpha\text{-}\beta}(t) = \frac{1}{\Delta x_i} \int_{\Omega_i} \hat{s}^{\rho u}(\boldsymbol{q}(x,t), x)\, dx + \mathcal{O}\left(h^2\right). \tag{4.19}$$

For the energy source term we have

$$\begin{aligned} \hat{S}_i^{E, \alpha\text{-}\beta}(t) = \frac{\hat{\rho u}_i(t)}{\hat{\rho}_i(t)} \hat{S}_i^{\rho u, \alpha\text{-}\beta}(t) &= \frac{\rho u(x_i,t)}{\rho(x_i,t)} \hat{S}_i^{\rho u, \alpha\text{-}\beta}(t) + \mathcal{O}\left(h^2\right) \\ &\overset{Eq.\ (4.18)}{=} \frac{\rho u(x_i,t)}{\rho(x_i,t)} s^{\rho u}(x_i,t) + \mathcal{O}\left(h^2\right) \\ &= \int_{\Omega_i} \frac{\rho u(x,t)}{\rho(x,t)} s^{\rho u}(x,t)\, dx + \mathcal{O}\left(h^2\right) \\ &= \int_{\Omega_i} s^E(x,t)\, dx + \mathcal{O}\left(h^2\right). \end{aligned} \tag{4.20}$$

Hence, we have $\hat{\boldsymbol{S}}_i^{\alpha\text{-}\beta}(t) = \frac{1}{\Delta x_i} \int_{\Omega_i} \boldsymbol{s}(\boldsymbol{q}(x,t), x)\, dx + \mathcal{O}\left(h^2\right)$. Now, we show that the hydrostatic reconstruction is second order accurate: First, note that

$$\frac{1}{\Delta x_i} \int_{\Omega_i} \mathcal{T}_x^{\alpha\text{-}\beta}\left(\boldsymbol{q}(x,t)\right)\, dx = \mathcal{T}_{x_i}^{\alpha\text{-}\beta}\left(\hat{\boldsymbol{q}}_i(t)\right) + \mathcal{O}\left(h^2\right), \tag{4.21}$$

which is shown in Appendix B.2. The reconstruction $\mathcal{R}$ is assumed to be second order accurate in Theorem 4.2.2. The back-transformation $\left(\mathcal{T}^{\alpha\text{-}\beta}_{x_{i+\frac{1}{2}}}\right)^{-1}$ of the interface values transports the second order error (see Appendix B.2). In total the hydrostatic reconstruction is second order accurate. This makes the method described in Theorem 4.2.2 second order accurate. □

4.2.2.2 Well-Balanced Property

Theorem 4.2.3. *Consider the finite volume scheme* (3.75)*, for which the interface values have been obtained with a reconstruction as described in Eqs.* (4.10) *and* (4.11) *and with the source term discretization as defined in Eq.* (4.16)*. This scheme is well-balanced in the sense that* $\frac{d}{dt}\hat{\boldsymbol{Q}}_i^{\text{hs}} = 0$ *follows for any initial data* $\hat{\boldsymbol{Q}}_i^{\text{hs}}$ *that satisfy* $\mathcal{T}^{\alpha\text{-}\beta}_{x_i}(\hat{\boldsymbol{Q}}_i^{\text{hs}}) = (a,0,a)^T$ *for some constants* $a > 0$ *independent from* i*. In other words, the relations*

$$\frac{\hat{\rho}_i^{\text{hs}}}{\alpha_i} = \frac{p_i^{\text{hs}}}{\beta_i} = \text{const.}, \qquad \hat{\rho u}_i^{\text{hs}} = 0 \tag{4.22}$$

with $p_i^{\text{hs}} := p_{EoS}\left(\hat{\rho}_i^{\text{hs}}, \hat{E}_i^{\text{hs}} - \frac{\left(\hat{\rho u}_i^{\text{hs}}\right)^2}{\hat{\rho}_i^{\text{hs}}}\right)$ *lead to a vanishing residual.*

Proof. Since it is true by construction of the method, Theorem 4.2.3 is straightforward to show: A simple computation can verify that the relations in Eq. (4.22) are equivalent to $\mathcal{T}^{\alpha\text{-}\beta}_{x}(\hat{\boldsymbol{Q}}_i^{\text{hs}}) = (a,0,a)^T$ for some $a > 0$. Using this relation and Eqs. (4.10) and (4.11) we find that

$$\begin{aligned}
\left(\boldsymbol{Q}^{\text{hs}}_{i+\frac{1}{2}}\right)^L &= \left(\mathcal{T}^{\alpha\text{-}\beta}_{x_{i+\frac{1}{2}}}\right)^{-1}\left(\mathcal{R}_i\left(x_{i+\frac{1}{2}}; \left\{\mathcal{T}^{\alpha\text{-}\beta}_{x_j}\left(\hat{\boldsymbol{Q}}_j^{\text{hs}}\right)\right\}_{j\in\mathcal{S}_i}\right)\right)\\
&= \left(\mathcal{T}^{\alpha\text{-}\beta}_{x_{i+\frac{1}{2}}}\right)^{-1}\left(\mathcal{R}_i\left(x_{i+\frac{1}{2}}; \left\{(a,0,a)^T\right\}_{j\in\mathcal{S}_i}\right)\right)\\
&= \left(\mathcal{T}^{\alpha\text{-}\beta}_{x_{i+\frac{1}{2}}}\right)^{-1}\left((a,0,a)^T\right)\\
&= \left(\mathcal{T}^{\alpha\text{-}\beta}_{x_{i+\frac{1}{2}}}\right)^{-1}\left(\mathcal{R}_{i+1}\left(x_{i+\frac{1}{2}}; \left\{(a,0,a)^T\right\}_{j\in\mathcal{S}_{i+1}}\right)\right)\\
&= \left(\mathcal{T}^{\alpha\text{-}\beta}_{x_{i+\frac{1}{2}}}\right)^{-1}\left(\mathcal{R}_{i+1}\left(x_{i+\frac{1}{2}}; \left\{\mathcal{T}^{\alpha\text{-}\beta}_{x_j}\left(\hat{\boldsymbol{Q}}_j^{\text{hs}}\right)\right\}_{j\in\mathcal{S}_{i+1}}\right)\right)\\
&= \left(\boldsymbol{Q}^{\text{hs}}_{i+\frac{1}{2}}\right)^R =: \boldsymbol{Q}^{\text{hs}}_{i+\frac{1}{2}}.
\end{aligned} \tag{4.23}$$

Hence, the numerical flux at the $i+\frac{1}{2}$ interface is

$$\begin{aligned}
\boldsymbol{F}^{\mathrm{hs}}_{i+\frac{1}{2}} &:= \boldsymbol{F}\left(\left(\boldsymbol{Q}^{\mathrm{hs}}_{i+\frac{1}{2}}\right)^{L}, \left(\boldsymbol{Q}^{\mathrm{hs}}_{i+\frac{1}{2}}\right)^{R}\right) \\
&= \boldsymbol{f}\left(\boldsymbol{Q}^{\mathrm{hs}}_{i+\frac{1}{2}}\right) \\
&= \boldsymbol{f}\left(\left(\mathcal{T}^{\alpha\text{-}\beta}_{x_{i+\frac{1}{2}}}\right)^{-1}\left((a,0,a)^{T}\right)\right) \\
&= \boldsymbol{f}\left(\left.\frac{\partial \boldsymbol{q}^{\mathrm{cons}}}{\partial \boldsymbol{q}^{\mathrm{prim}}}\right|_{\boldsymbol{Q}^{\mathrm{hs}}_{i+\frac{1}{2}}}\begin{pmatrix} a\alpha_{i+\frac{1}{2}} \\ 0 \\ a\beta_{i+\frac{1}{2}}\end{pmatrix}\right) = \begin{pmatrix} 0 \\ a\beta_{i+\frac{1}{2}} \\ 0\end{pmatrix}
\end{aligned} \tag{4.24}$$

Equation (4.16) together with Eq. (4.24) yields

$$\begin{aligned}
\hat{\boldsymbol{S}}^{\mathrm{hs}}_{i} := \left.\hat{\boldsymbol{S}}^{\alpha\text{-}\beta}_{i}\right|_{\hat{\boldsymbol{Q}}^{\mathrm{hs}}_{i}} &= \begin{pmatrix} 0 \\ \frac{\beta_{i+\frac{1}{2}}-\beta_{i-\frac{1}{2}}}{\Delta x_i}\frac{\hat{\rho}^{\mathrm{hs}}_{i}}{\alpha_i} \\ 0\end{pmatrix} = \begin{pmatrix} 0 \\ \frac{\beta_{i+\frac{1}{2}}-\beta_{i-\frac{1}{2}}}{\Delta x_i}a \\ 0\end{pmatrix} \\
&= \frac{1}{\Delta x_i}\left(\begin{pmatrix} 0 \\ a\beta_{i+\frac{1}{2}} \\ 0\end{pmatrix} - \begin{pmatrix} 0 \\ a\beta_{i-\frac{1}{2}} \\ 0\end{pmatrix}\right) = \frac{1}{\Delta x_i}\left(\boldsymbol{F}^{\mathrm{hs}}_{i+\frac{1}{2}} - \boldsymbol{F}^{\mathrm{hs}}_{i-\frac{1}{2}}\right),
\end{aligned} \tag{4.25}$$

i.e., $\frac{d}{dt}\hat{\boldsymbol{Q}}^{\mathrm{hs}}_{i} = 0$ from Eq. (3.75). □

4.2.2.3 Scope

Equation (4.22) actually describes a second order discretization to the hydrostatic state. However, since this approximation consists of local discretizations in each cell of a global hydrostatic solution, we will still say that the well-balanced method is exact. An example of a discrete hydrostatic state that has to be seen in a different light is given in Section 4.4.

Remark 4.2.4. *Theorem 4.2.3 is a general result in the sense that we have not made any assumptions on the EoS or on the type of gravitational field. It holds for any consistent numerical flux and reconstruction scheme applied to obtain the cell interface values.*

4.3 The Deviation Method

The method described in this section (and in [14]) has been developed in the effort of extending the α-β method described in Section 4.2 to more general systems. The α-β method uses the structure of the Euler equations (2.34) and their solutions. This specific transformation to hydrostatic variables is only possible due to the positivity of ρ and p. To use a similar hydrostatic reconstruction for example for static states of the ideal MHD equations which involve magnetic fields, it has to be modified: Magnetic fields are not restricted to positive values in all the components. This led to

the idea of using subtraction instead of division in the hydrostatic reconstruction. The following method, which we believe is the most general well-balanced finite volume method in literature that assumes a priori knowledge of the target solution, is the result.

4.3.1 Description of the Deviation Method

Assume the one-dimensional system of hyperbolic balance laws

$$\partial_t \boldsymbol{q}(x,t) + \partial_x \boldsymbol{f}(\boldsymbol{q}(x,t)) = \boldsymbol{s}(\boldsymbol{q}(x,t),x). \tag{4.26}$$

Let $\tilde{\boldsymbol{q}}$ be a given smooth solution of Eq. (4.26), i.e.,

$$\partial_t \tilde{\boldsymbol{q}}(x,t) + \partial_x \boldsymbol{f}(\tilde{\boldsymbol{q}}(x,t)) = \boldsymbol{s}(\tilde{\boldsymbol{q}}(x,t),x) \tag{4.27}$$

holds. The difference of these equations can be written in the form

$$\begin{aligned}\partial_t \Delta\boldsymbol{q}(x,t) + \partial_x \left(\boldsymbol{f}\left(\tilde{\boldsymbol{q}}(x,t) + \Delta\boldsymbol{q}(x,t)\right) - \boldsymbol{f}\left(\tilde{\boldsymbol{q}}(x,t)\right)\right) \\ = \boldsymbol{s}(\tilde{\boldsymbol{q}}(x,t) + \Delta\boldsymbol{q}(x,t),x) - \boldsymbol{s}(\tilde{\boldsymbol{q}}(x,t),x)\end{aligned} \tag{4.28}$$

if we formulate it with respect to the target solution $\tilde{\boldsymbol{q}}$ and the deviation

$$\Delta\boldsymbol{q} := \boldsymbol{q} - \tilde{\boldsymbol{q}}. \tag{4.29}$$

Averaging Eq. (4.28) in Ω_i yields

$$\begin{aligned}\frac{d}{dt}(\Delta\hat{\boldsymbol{q}}_i(t)) = -\frac{1}{\Delta x_i}\Big[&\left(\boldsymbol{f}((\Delta\boldsymbol{q}+\tilde{\boldsymbol{q}})(x_{i+\frac{1}{2}},t)) - \boldsymbol{f}(\tilde{\boldsymbol{q}}(x_{i+\frac{1}{2}},t))\right) \\ &- \left(\boldsymbol{f}((\Delta\boldsymbol{q}+\tilde{\boldsymbol{q}})(x_{i-\frac{1}{2}},t)) - \boldsymbol{f}(\tilde{\boldsymbol{q}}(x_{i-\frac{1}{2}},t))\right)\Big] \\ &+ \frac{1}{\Delta x_i}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \boldsymbol{s}((\Delta\boldsymbol{q}+\tilde{\boldsymbol{q}})(x,t),x,t) - \boldsymbol{s}(\tilde{\boldsymbol{q}}(x,t),x,t)\,dx.\end{aligned} \tag{4.30}$$

The semi-discrete scheme to evolve the approximation $\Delta\hat{\boldsymbol{Q}}_i$ to the cell-average deviation $\Delta\hat{\boldsymbol{q}}_i := \hat{\boldsymbol{q}}_i - \hat{\tilde{\boldsymbol{q}}}_i$ is then obtained using standard discretization techniques. It takes the form

$$\begin{aligned}\frac{d}{dt}\left(\Delta\hat{\boldsymbol{Q}}_i(t)\right) = -\frac{1}{\Delta x_i}\Big[&\Delta\boldsymbol{F}\left(\Delta\boldsymbol{Q}^L_{i+\frac{1}{2}}(t), \Delta\boldsymbol{Q}^R_{i+\frac{1}{2}}(t), \tilde{\boldsymbol{q}}(x_{i+\frac{1}{2}},t)\right) \\ &- \Delta\boldsymbol{F}\left(\Delta\boldsymbol{Q}^L_{i-\frac{1}{2}}(t), \Delta\boldsymbol{Q}^R_{i-\frac{1}{2}}(t), \tilde{\boldsymbol{q}}(x_{i-\frac{1}{2}},t)\right)\Big] \\ &+ \Delta\boldsymbol{S}_i\left((\Delta\boldsymbol{Q})_i^{\mathrm{rec}}, \tilde{\boldsymbol{q}}, t\right)\end{aligned} \tag{4.31}$$

with

$$\Delta\boldsymbol{F}\left(\Delta\boldsymbol{Q}^L, \Delta\boldsymbol{Q}^R, \tilde{\boldsymbol{q}}\right) := \boldsymbol{F}(\Delta\boldsymbol{Q}^L + \tilde{\boldsymbol{q}}, \Delta\boldsymbol{Q}^R + \tilde{\boldsymbol{q}}) - \boldsymbol{f}(\tilde{\boldsymbol{q}}), \tag{4.32}$$

where $\boldsymbol{F}$ is a numerical flux function consistent with $\boldsymbol{f}$ and the functions $(\Delta\boldsymbol{Q})_i^{\mathrm{rec}}$ of reconstructed states are obtained using an m-th order accurate consistent conservative reconstruction on the cell average deviations $\Delta\hat{\boldsymbol{Q}}_i$. The interface values

$\Delta \boldsymbol{Q}^{L/R}$ are computed from $\boldsymbol{Q}_i^{\text{rec}}$. The discretization of the source term difference $\Delta \boldsymbol{S}$ is defined via

$$\Delta \boldsymbol{S}_i(\Delta \boldsymbol{q}, \tilde{\boldsymbol{q}}, t) := \boldsymbol{S}_i(\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}}, t) - \boldsymbol{S}_i(\tilde{\boldsymbol{q}}, t), \tag{4.33}$$

where

$$\boldsymbol{S}_i(\boldsymbol{q}, t) = \frac{1}{\Delta x_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \boldsymbol{s}(\boldsymbol{q}(x,t), x, t)\, dx + \mathcal{O}\left(h^m\right) \tag{4.34}$$

is some m-th order accurate source term discretization.

4.3.2 Properties of the Deviation Method

In the following we discuss some fundamental properties of the Deviation method.

4.3.2.1 Accuracy

Theorem 4.3.1. *The semi-discrete scheme* (4.31) *is consistent and m-th order accurate in space.*

Proof. Let q be a smooth solution of Eq. (4.26), where we denote the deviations from the target state with $\Delta \boldsymbol{q} = \boldsymbol{q} - \tilde{\boldsymbol{q}}$. Since the reconstruction is m-th order accurate we have

$$\Delta \boldsymbol{Q}_i^{\text{rec}}(x) = \mathcal{R}_i\left(x; \{\Delta \hat{\boldsymbol{q}}_j\}_{j \in \mathcal{S}_i}\right) = \Delta \boldsymbol{q}(x) + \mathcal{O}\left(h^m\right) \qquad \text{for} \qquad x \in \Omega_i. \tag{4.35}$$

Consequently, for the interface values we have $\Delta \boldsymbol{Q}_{i+\frac{1}{2}}^{L/R} = \Delta \boldsymbol{q}(x_{i+\frac{1}{2}}) + \mathcal{O}\left(h^m\right)$. Since the numerical flux is Lipschitz continuous and consistent, the following holds:

$$\begin{aligned}
\Delta \boldsymbol{F}\left(\Delta \boldsymbol{Q}_{i+\frac{1}{2}}^{L}, \Delta \boldsymbol{Q}_{i+\frac{1}{2}}^{R}, \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}\right) &= \boldsymbol{F}(\Delta \boldsymbol{Q}_{i+\frac{1}{2}}^{L} + \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}, \Delta \boldsymbol{Q}_{i+\frac{1}{2}}^{R} + \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) - \boldsymbol{f}(\tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) \\
&= \boldsymbol{F}(\Delta \boldsymbol{q}_{i+\frac{1}{2}} + \tilde{\boldsymbol{q}}_{i+\frac{1}{2}} + \mathcal{O}\left(h^m\right), \Delta \boldsymbol{q}_{i+\frac{1}{2}} + \tilde{\boldsymbol{q}}_{i+\frac{1}{2}} + \mathcal{O}\left(h^m\right)) - \boldsymbol{f}(\tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) \\
&= \boldsymbol{F}(\Delta \boldsymbol{q}_{i+\frac{1}{2}} + \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}, \Delta \boldsymbol{q}_{i+\frac{1}{2}} + \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) - \boldsymbol{f}(\tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) + \mathcal{O}\left(h^m\right) \\
&= \boldsymbol{f}(\Delta \boldsymbol{q}_{i+\frac{1}{2}} + \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) - \boldsymbol{f}(\tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) + \mathcal{O}\left(h^m\right).
\end{aligned} \tag{4.36}$$

The source term discretization $\boldsymbol{S}_i$ is m-th order accurate by definition of the method in Section 4.3.1. This leads to

$$\begin{aligned}
\Delta \boldsymbol{S}_i(\Delta \boldsymbol{q}, \tilde{\boldsymbol{q}}, t) &= \boldsymbol{S}_i(\Delta \boldsymbol{Q}_i^{\text{rec}} + \tilde{\boldsymbol{q}}, t) - \boldsymbol{S}_i(\tilde{\boldsymbol{q}}, t) = \boldsymbol{S}_i(\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}} + \mathcal{O}\left(h^m\right), t) - \boldsymbol{S}_i(\tilde{\boldsymbol{q}}, t) \\
&= \frac{1}{\Delta x_i} \int_{\Omega_i} \boldsymbol{s}((\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}})(x,t), x, t) - \boldsymbol{s}(\tilde{\boldsymbol{q}}(x,t), x, t) + \mathcal{O}\left(h^m\right)\, dx + \mathcal{O}\left(h^m\right) \\
&= \frac{1}{\Delta x_i} \int_{\Omega_i} \boldsymbol{s}((\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}})(x,t), x, t) - \boldsymbol{s}(\tilde{\boldsymbol{q}}(x,t), x, t)\, dx + \mathcal{O}\left(h^m\right).
\end{aligned} \tag{4.37}$$

Since the fluxes and source term are approximated to m-th order in space, the semi-discrete scheme (4.31) is m-th order accurate in space. □

Corollary 4.3.2 (Accuracy of the deviation method)**.** *The semi-discrete scheme* (4.31)*, interpreted as a method to evolve $\hat{\boldsymbol{Q}}_i = \hat{\tilde{\boldsymbol{q}}}_i + \Delta \hat{\boldsymbol{Q}}_i$, is m-th order accurate in space.*

Proof. Corollary 4.3.2 follows directly from Theorem 4.3.1 and the relation $\frac{d}{dt}\Delta \hat{\boldsymbol{q}}_i = \frac{d}{dt}\hat{\boldsymbol{q}}_i - \frac{d}{dt}\hat{\tilde{\boldsymbol{q}}}_i$. □

4.3.2.2 Well-Balanced Property

Theorem 4.3.3. *The semi-discrete scheme* (4.31) *maintains the zero state, i.e., the initial conditions* $\Delta \boldsymbol{Q}_i(t=0) = 0$ *for all* $i \in \mathcal{I}$ *lead to* $\Delta \boldsymbol{Q}_i(t) = 0$ *for all times* $t > 0$.

Proof. The theorem can be shown by simple computation: Let $\Delta \hat{\boldsymbol{Q}}_i = 0$ for all $i \in \mathcal{I}$. Due to the consistency of the reconstruction, the reconstructed states

$$\Delta \boldsymbol{Q}_i^{\text{rec}}(x) = \mathcal{R}_i\left(x; \left\{\Delta \hat{\boldsymbol{Q}}_j\right\}_{j\in\mathcal{S}_i}\right) = \mathcal{R}_i\left(x; \{0\}_{j\in\mathcal{S}_i}\right) = 0 \tag{4.38}$$

vanish. This includes $\Delta Q^{L/R}_{i+\frac{1}{2}} = 0$. The flux difference discretization is hence

$$\begin{aligned}\Delta \boldsymbol{F}\left(\Delta \boldsymbol{Q}^L_{i+\frac{1}{2}}, \Delta \boldsymbol{Q}^R_{i+\frac{1}{2}}, \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}\right) &= \boldsymbol{F}(\Delta \boldsymbol{Q}^L_{i+\frac{1}{2}} + \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}, \Delta \boldsymbol{Q}^R_{i+\frac{1}{2}} + \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) - \boldsymbol{f}(\tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) \\ &= \boldsymbol{F}(\tilde{\boldsymbol{q}}_{i+\frac{1}{2}}, \tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) - \boldsymbol{f}(\tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) \\ &\overset{(*)}{=} \boldsymbol{f}(\tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) - \boldsymbol{f}(\tilde{\boldsymbol{q}}_{i+\frac{1}{2}}) = 0,\end{aligned} \tag{4.39}$$

where $(*)$ holds because of the consistency of the numerical flux $\boldsymbol{F}$. We use the notation $\boldsymbol{q}_{i+\frac{1}{2}} := \boldsymbol{q}\left(x_{i+\frac{1}{2}}, t\right)$. For the source term difference we find

$$\Delta \boldsymbol{S}_i\left((\Delta \boldsymbol{Q})_i^{\text{rec}}, \tilde{\boldsymbol{q}}, t\right) = \Delta \boldsymbol{S}_i\left(0, \tilde{\boldsymbol{q}}, t\right) \overset{Eq.\ (4.34)}{=} \boldsymbol{S}_i\left(\tilde{\boldsymbol{q}}, t\right) - \boldsymbol{S}_i\left(\tilde{\boldsymbol{q}}, t\right) = 0. \tag{4.40}$$

Plugging Eqs. (4.39) and (4.40) into Eq. (4.31) yields $\frac{d}{dt}\Delta \hat{\boldsymbol{Q}} = 0$. □

Corollary 4.3.4 (Well-balanced property of the Deviation method)**.** *The semi-discrete scheme Eq.* (4.31)*, interpreted as a method to evolve* $\hat{\boldsymbol{Q}}_i = \hat{\tilde{\boldsymbol{q}}}_i + \Delta \hat{\boldsymbol{Q}}_i$*, is well-balanced in the sense that* $\hat{\boldsymbol{Q}}_i(t=0) = \hat{\tilde{\boldsymbol{q}}}_i(t=0)$ *implies* $\hat{\boldsymbol{Q}}_i(t) = \hat{\tilde{\boldsymbol{q}}}_i(t)$ *for all times* $t > 0$.

Proof. Corollary 4.3.4 follows directly from Theorem 4.3.3. □

4.3.2.3 Scope

Note that the well-balanced property of the Deviation method (Corollary 4.3.4) is even more general than the one for the α-β method (Theorem 4.2.3). For example, the α-β method is constructed to balance hydrostatic solutions of the compressible Euler equations with gravity. The Deviation method is not restricted to hydrostatic solutions. The target solution $\tilde{\boldsymbol{q}}$ can also be a non-static stationary state or a time-dependent solution of Eq. (4.26). In the following we provide some remarks regarding the simplicity and versatility of the Deviation method, which have been published in [14] in the same or a similar form (Remarks 4.3.5-4.3.8).

Remark 4.3.5. *The target solution in the Deviation method can very well be time-dependent. This does not change the consistency, accuracy, or the well-balanced property formulated in Corollaries 4.3.2 and 4.3.4. Thus, we use the phrase "well-balancing" in a wider sense than it is typically used.*

Even more, Eq. (4.26) can describe any hyperbolic balance law, which means, that the Deviation method is not restricted to the compressible Euler equations. Other examples of hyperbolic balance laws, for which it could be used, are ideal MHD equations with gravity or shallow water equations.

If the method is applied to balance a stationary state, it can be structured in a different way:

Remark 4.3.6. *If a stationary solution is chosen as target solution (which is the case for classical well-balancing applications), the time derivative of the target solution vanishes by definition. This leads to* $\frac{d}{dt}\boldsymbol{Q}_i = \frac{d}{dt}(\Delta\boldsymbol{Q}_i)$. *The described method can then also be used to directly evolve the* $\boldsymbol{Q}_i$ *in time instead of* $\Delta\boldsymbol{Q}_i$. *For that the scheme can be adapted by just substituting all* $\Delta\boldsymbol{Q}$ *terms with the corresponding* $\boldsymbol{Q} - \tilde{\boldsymbol{Q}}$ *terms. The reconstruction has then to be applied on the states* $\boldsymbol{Q} - \tilde{\boldsymbol{Q}}$.

In this case, it is easy to see the similarity in the basic idea behind the α-β method and the Deviation method. In both cases, a hydrostatic reconstruction is applied and the source term is defined such that the residual cancels in the hydrostatic/stationary case, which is balanced by the scheme. Also, the source term discretization can be simplified, if the Deviation method is applied for compressible Euler equations with gravity.

Remark 4.3.7. *Note that*

$$\boldsymbol{s}((\tilde{\boldsymbol{q}} + \Delta\boldsymbol{q})(x,t), x, t) - \boldsymbol{s}(\tilde{\boldsymbol{q}}(x,t), x, t) = \boldsymbol{s}(\Delta\boldsymbol{q}(x,t), x, t) \tag{4.41}$$

if the source term $\boldsymbol{s}$ *in Eq.* (4.26) *is linear in the first argument. In the case of linearity of the corresponding source term discretizations* $\boldsymbol{S}_i$, *this relation also holds for the discretized source terms. This is the case for the gravitational source term in Euler or ideal MHD equations and the bottom topography source term in the shallow water equations.*

This further simplifies the implementation of the method to an existing code. Also, the following can be useful in some applications.

Remark 4.3.8. *Our well-balanced method can even be beneficially applied if there is no source term. Applications could include stationary solutions based on vorticity in multi-dimensional simulations. Corresponding numerical tests are presented in Sections 6.4.5 and 6.4.6.*

In Section 4.3.1 we assume the target solution to be smooth. This is not required for the well-balanced property of the method. However, removing the restriction would have some implications.

Remark 4.3.9. *The target solution* $\tilde{\boldsymbol{q}}$ *is assumed to be smooth in Section 4.3.1. Note, that the well-balanced property also holds if the target solution is not smooth. However, for non-smooth target solutions we cannot expect the method to be high order accurate. Since the exact flux of the target solution at the interface is computed, the target solution has to be continuous. Discontinuous target solutions could in principle be used if a numerical two-state flux is applied at the target state interface values instead of the exact flux. This would make the method slower and more diffusive. This modification is not treated in this thesis.*

In Section 3.4.3 we discussed the possibility to reconstruct in different sets of variables in order to obtain different properties for the full scheme. This is also possible for the Deviation method.

Remark 4.3.10. *To reconstruct the deviations from the target solution in another set of variables, one can transform both the cell-averaged target solution* $\hat{\tilde{\boldsymbol{q}}}$ *and the cell-averaged states* $\hat{\boldsymbol{q}} = \hat{\tilde{\boldsymbol{q}}} + (\hat{\Delta \boldsymbol{q}})$ *to the other set of variables* $\hat{\tilde{\boldsymbol{q}}}^{\text{other}}$, $\hat{\boldsymbol{q}}^{\text{other}}$ *using a sufficiently high order accurate variable transform. The cell-averaged deviations* $(\hat{\Delta \boldsymbol{q}})^{\text{other}} = \hat{\boldsymbol{q}}^{\text{other}} - \hat{\tilde{\boldsymbol{q}}}^{\text{other}}$ *in the other set of variables are then reconstructed. The reconstructed states can be transformed back to conserved variables in a point-wise manner.*

This is applied in the tests presented in Sections 6.4.2 and 6.4.7 and in numerical experiments shown in [58].

4.4 The Discretely Well-Balanced Method

While the two methods introduced above in Sections 4.2 and 4.3 are exact on any hydrostatic solution[5], there is still one disadvantage, namely, that the target solution has to be defined before applying the method. This requires a priori knowledge. This knowledge is available in a wide range of practical applications. However, there are also examples, in which the target solution is not clear: Consider for example a simulation of a star (modeled using compressible Euler equations with gravity), which is in a dynamical state and settles in some hydrostatic solution subject to self-gravity. In this case, it might be challenging to apply the α-β or Deviation method. In this section, we present the well-balanced method introduced in [12], which we call Discretely Well-Balanced method in this thesis.

4.4.1 Description of the Discretely Well-Balanced Method

Similar to the Deviation method, the Discretely Well-Balanced method is a modification to the standard method (3.75) that only modifies the reconstruction and uses a certain source term discretization to achieve the well-balanced property. To finally obtain a well-balanced property, we have to assume the numerical flux which is applied satisfies the contact property (Definition 3.3.2). The Discretely Well-Balanced method has been described and published in [12], and some parts of the description in Sections 4.4.1.1 and 4.4.1.2 are similar to the one in [12].

4.4.1.1 Source Term Discretization

Let us define the source term approximation for the momentum and energy equations by

$$S^{\rho u,\mathrm{DWB},i}(x) = \rho_i^{\mathrm{rec}}(x) g_i^{\mathrm{int}}(x) \quad \text{and} \quad S^{E,\mathrm{DWB},i}(x) = (\rho u)_i^{\mathrm{rec}}(x) g_i^{\mathrm{int}}(x), \tag{4.42}$$

[5] As shown above, the Deviation method can also exactly follow non-static solutions

where ρ_i^{rec} and $(\rho u)_i^{\text{rec}}$ are the m-th order accurate CWENO reconstruction polynomials in the i-th cell. g_i^{int} has been interpolated from cell-centered point values using m-th order accurate polynomial interpolation. The local source term approximation is

$$\boldsymbol{S}^{\text{DWB},i}(x) := \begin{pmatrix} 0 \\ S^{\rho u,\text{DWB},i}(x) \\ S^{E,\text{DWB},i}(x) \end{pmatrix} \tag{4.43}$$

for each cell. The global source term approximation is then defined by

$$\boldsymbol{S}^{\text{DWB}}(x) := \begin{pmatrix} 0 \\ S^{\rho u,\text{DWB}}(x) \\ S^{E,\text{DWB}}(x) \end{pmatrix}, \tag{4.44}$$

where

$$S^{\rho u,\text{DWB}}(x) := S^{\rho u,\text{DWB},i}(x), \quad S^{E,\text{DWB}}(x) := S^{E,\text{DWB},i}(x) \quad \text{for } x \in \Omega_i. \tag{4.45}$$

Since the components of $\boldsymbol{S}^{\text{DWB}}$ are piecewise polynomials, the cell-average source term approximation

$$\hat{\boldsymbol{S}}_i^{\text{DWB}} := \frac{1}{\Delta x_i} \int_{\Omega_i} \boldsymbol{S}^{\text{DWB}}(x)\, dx \tag{4.46}$$

can be evaluated exactly.

Remark 4.4.1. *Note that any other high order source term approximation can be used in our algorithm, as long as it provides a function which is defined in the whole cell and that can be consistently extrapolated to neighboring cells.*

4.4.1.2 Local Hydrostatic Reconstruction

The basic idea of the local hydrostatic reconstruction is to reconstruct deviations to a high order accurate local equilibrium profile $\boldsymbol{Q}_i^{\text{hs}}(x)$ consistent with the cell-averaged conserved variables. This idea is similar to the α-β and Deviation method. The basic difference in the Discretely Well-Balanced method is, that the hydrostatic solution which is well-balanced is found by local approximation: To find a cell's hydrostatic profile the source term is integrated to neighboring cells. This procedure will now be described in technical detail.

Step 1 This first step corresponds to the "$\mathcal{T}$"-step in the equilibrium reconstruction procedure described in Section 4.1.

We begin by the construction of the high-order local hydrostatic profile. A one-dimensional hydrostatic equilibrium is by definition given by the solution of the hydrostatic equation

$$\left(p^{\text{hs}}\right)'(x) = \rho^{\text{hs}}(x) g(x). \tag{4.47}$$

We construct the local pressure hydrostatic profile p_i^{hs} within the i-th cell by simply integrating Eq. (4.47)

$$p_i^{\text{hs}}(x) = p_{0,i} + \int_{x_i}^{x} S^{\rho u,\text{DWB}}(\xi)\, d\xi, \tag{4.48}$$

where $p_{0,i}$ is the point value of the pressure at cell center x_i.

The hydrostatic internal energy density $\varepsilon_i^{\mathrm{hs}}(x)$ profile can be computed through the EoS:

$$\varepsilon_i^{\mathrm{hs}}(x) = \varepsilon\left(\rho_i^{\mathrm{hs}}(x), p_i^{\mathrm{hs}}(x)\right). \tag{4.49}$$

In the ideal gas case, the computation is trivial and can be performed explicitly

$$\varepsilon_i^{\mathrm{hs}}(x) = \frac{p_i^{\mathrm{hs}}(x)}{\gamma - 1}. \tag{4.50}$$

The cell center pressure $p_{0,i}$, which anchors the equilibrium pressure profile (4.48) at cell center, is determined by demanding the consistency

$$\hat{\varepsilon}_i^{\mathrm{est}} = \hat{\varepsilon}_i^{\mathrm{hs}}. \tag{4.51}$$

The cell average

$$\hat{\varepsilon}_i^{\mathrm{hs}} = \frac{1}{\Delta x_i} I_{x\in\Omega_i}\left[\varepsilon_i^{\mathrm{hs}}(x)\right] = \frac{1}{\Delta x_i} I_{x\in\Omega_i}\left[\varepsilon\left(\rho_i^{\mathrm{hs}}(x), p_i^{\mathrm{hs}}(x)\right)\right] \tag{4.52}$$

of the hydrostatic internal energy density is computed using the m-th order accurate quadrature rule[6] with the estimate $\hat{\varepsilon}_i^{\mathrm{est}}$ of the cell-averaged internal energy density in cell Ω_i obtained from

$$\hat{\varepsilon}_i^{\mathrm{est}} = \hat{E}_i - \frac{1}{2}\frac{\left(\hat{\rho u}\right)_i^2}{\hat{\rho}_i} \tag{4.53}$$

following [78]. Note that $\hat{\varepsilon}_i^{\mathrm{est}} = \hat{\varepsilon}_i$ without error at any hydrostatic state (since $\hat{\rho u}_i = 0$ in that case).

For general EoS, Eq. (4.51) is a scalar nonlinear equation for the cell center pressure $p_{0,i}$. The general case is omitted at this place for readability and instead discussed in Section 4.4.1.3.

Assuming the ideal gas EoS (2.21), Eq. (4.51) is linear and can be solved analytically to yield

$$p_{0,i} = (\gamma - 1)\hat{\varepsilon}_i - \frac{1}{\Delta x_i} I_{x\in\Omega_i}\left[\int_{x_i}^{x} S^{\rho u,\mathrm{DWB}}(\xi)\, d\xi\right]. \tag{4.54}$$

Now that the pressure at cell center $p_{0,i}$ is fixed, we have fully specified the high-order accurate representation of the equilibrium conserved variables in cell Ω_i:

$$\boldsymbol{Q}_i^{\mathrm{hs}}(x) := \begin{pmatrix} \rho_i^{\mathrm{hs}}(x) \\ 0 \\ \varepsilon_i^{\mathrm{hs}}(x) \end{pmatrix}. \tag{4.55}$$

Our aim is to develop a high-order hydrostatic reconstruction procedure. To this end, we decompose in every cell the solution into a hydrostatic and a (not necessarily small) perturbation part. The idea of this procedure has been explained

[6] For an ideal gas EoS, this cell-average can in principle also be computed using exact integration. However, we apply a quadrature rule since it yields the same oder of accuracy and is necessary for the scheme to be applicable for arbitrary EoS.

in Section 4.1. The hydrostatic part in cell Ω_i is simply given by $\boldsymbol{Q}_i^{\mathrm{hs}}(x)$ of Eq. (4.55). The cell average of the perturbation $\delta\hat{\boldsymbol{Q}}_{ki}$ in cell Ω_k is obtained by taking the difference between the cell average $\hat{\boldsymbol{Q}}_k$ and the cell average of the hydrostatic part $\hat{\boldsymbol{Q}}_i^{\mathrm{hs}}$ in cell Ω_k, i.e.,

$$\delta\hat{\boldsymbol{Q}}_{ki} = \hat{\boldsymbol{Q}}_k - \frac{1}{\Delta x_i} I_{x\in\Omega_k}\left[\boldsymbol{Q}_i^{\mathrm{hs}}(x)\right]. \tag{4.56}$$

for $k \in \mathcal{S}_i$.

Step 2 This step corresponds to the "$\mathcal{R}$"-step in the equilibrium reconstruction procedure described in Section 4.1.

The perturbation part in cell Ω_i is directly reconstructed by applying the standard reconstruction procedure $\mathcal{R}$ to the cell-averaged perturbation

$$\delta\boldsymbol{Q}_i^{\mathrm{rec}}(x) = \mathcal{R}_i\left(x; \left\{\delta\hat{\boldsymbol{Q}}_{ki}\right\}_{k\in S_i}\right). \tag{4.57}$$

This results in an m-th order accurate representation of the perturbation in cell Ω_i.

Step 3 This last step corresponds to the "$\mathcal{T}^{-1}$"-step in the equilibrium reconstruction procedure described in Section 4.1.

The complete hydrostatic reconstruction $\mathcal{R}^{\mathrm{DWB}}$ is obtained by the sum of the (approximate) hydrostatic state and the reconstructed perturbation

$$\boldsymbol{Q}_i^{\mathrm{rec}}(x) = \mathcal{R}_i^{\mathrm{DWB}}\left(x; \{\hat{\boldsymbol{Q}}_k\}_{k\in S_i}\right) := \boldsymbol{Q}_i^{\mathrm{hs}}(x) + \delta\boldsymbol{Q}_i^{\mathrm{rec}}(x). \tag{4.58}$$

This concludes the description of the hydrostatic reconstruction procedure. Replacing only this component in a finite volume method renders it well-balanced for arbitrary hydrostatic solutions in a discrete sense as shown in Theorem 4.4.4.

4.4.1.3 Determining the Cell-Centered Pressure for Arbitrary Equations of State

In this section, we present the details for using the Discretely Well-Balanced scheme with a general EoS. In that case, Eq. (4.51) is not explicitly solvable for the cell center equilibrium pressure $p_{0,i}$ in cell Ω_i. The problem can be rewritten as

$$f(p_{0,i}) = 0, \tag{4.59}$$

where

$$f(p) = \hat{\varepsilon}_i^{\mathrm{est}} - \frac{1}{\Delta x_i} I_{x\in\Omega_i}\left[\varepsilon\left(\rho_i^{\mathrm{hs}}(x),\ p + \int_{x_i}^{x} S^{\rho u,\mathrm{DWB},i}(\xi)\ d\xi\right)\right]. \tag{4.60}$$

We again stress that the equilibrium density ρ_i^{hs} and the gravitational acceleration g_i^{int} are polynomials and, consequently, almost everything can be evaluated analytically in a straightforward manner. Only the EoS conversion to internal energy density given density and pressure $\varepsilon = \varepsilon(\rho, p)$ is in general not explicitly available.

Therefore, solving Eq. (4.59) for $p_{0,i}$ requires some iterative procedure such as Newton's method

$$p_{0,i}^{(k+1)} = p_{0,i}^{(k)} - \frac{f(p_{0,i}^{(k)})}{f'(p_{0,i}^{(k)})}, \quad (k = 0, 1, \dots), \tag{4.61}$$

where the superscript in parenthesis labels the iteration number and the derivative of Eq. (4.60) is given by

$$f'(p) = -\frac{1}{\Delta x_i} I_{x\in\Omega_i} \left[\frac{\partial \varepsilon}{\partial p} \left(\rho_i^{\mathrm{hs}}(x),\ p + \int_{x_i}^{x} S^{\rho u,\mathrm{DWB},i}(\xi)\ d\xi \right) \right]. \tag{4.62}$$

The iteration is started with the pressure

$$p_{0,i}^{(0)} = p(\hat{\rho}_i, \hat{\varepsilon}_i^{\mathrm{est}}) \tag{4.63}$$

computed from the cell-averaged conserved variables as initial guess. It is stopped and the cell-centered pressure $p_{0,i} = p_{0,i}^{(k)}$ is returned by the routine if the condition

$$\left| \frac{f\left(p_{0,i}^{(k)}\right)}{f'\left(p_{0,i}^{(k)}\right)} \right| < \tau \tag{4.64}$$

is met, where we chose $\tau = 10^{-13}$ in the numerical experiments conducted in this thesis.

As is well-known, the global convergence properties of Newton's method are poor. However, it is straightforward to build a robust solver by combining it with, for example, the bisection method (see [49, 133] and references therein for details). Such a modification was not necessary for the presented numerical examples using the ideal gas with radiation pressure EoS.

Simplified approach However, for many applications it might be sufficient to use a simplified approach to determine the value of $p_{0,i}$. Choose

$$\varepsilon_{0,i} := E_i^{\mathrm{rec}}(x_i) - \frac{1}{2} \frac{\left((\rho u)_i^{\mathrm{rec}}(x_i)\right)^2}{\rho_i^{\mathrm{rec}}(x_i)}, \tag{4.65}$$

which is the cell-centered internal energy computed from the CWENO reconstruction polynomials. Then apply the EoS to compute

$$p_{0,i} := p_{\mathrm{EoS}}\left(\rho_i^{\mathrm{rec}}(x_i), \varepsilon_{0,i}\right). \tag{4.66}$$

The resulting method will be referred to as Discretely Well-Balanced (fast computation of p_0) or simply DWB-fast. Note, however, that Corollary 4.3.4, which states the well-balanced property for the Deviation method in the following section, does not hold if this simplified approach for computing the cell-centered pressure is applied.

4.4.2 Properties of the Discretely Well-Balanced Method

Remark 4.4.2. *In the description of the Discretely Well-Balanced method we use a hydrostatic target state* $\boldsymbol{Q}^{\mathrm{hs}}$ *to make it more similar and comparable to the Deviation method (Section 4.3). However, the density reconstruction is not modified, since* $\rho_i^{\mathrm{hs}} = \rho_i^{\mathrm{rec}}$*, and there are no density perturbations by construction. The hydrostatic reconstruction routine returns* ρ^{hs} *which is equal to* ρ^{rec}*. From the implementation point of view, we can thus conclude, that only the reconstruction of the total energy has to be modified.*

4.4.2.1 Accuracy

Theorem 4.4.3. *Consider the semi-discrete scheme Eq.* (3.75) *for compressible Euler equations with gravity* (2.34) *with a numerical flux* $\boldsymbol{F}$*, the hydrostatic reconstruction* $\mathcal{R}^{\mathrm{DWB}}$ *(Eq.* (4.58)*) based on an m-th order accurate spatial reconstruction procedure* $\mathcal{R}$*, and the gravitational source term discretization* $\hat{\boldsymbol{S}}_i^{\mathrm{DWB}}$ *(Eq.* (4.46)*).*

The scheme is consistent and at least m-th order accurate in space for smooth solutions.

Proof. It is straight forward to show that the source term discretization $\hat{\boldsymbol{S}}_i^{\mathrm{DWB}}$ is m-th order accurate: it is obtained by integrating the source term evaluation on m-th order accurate reconstructed states.

The reconstruction routine is actually only modified in the reconstruction of the total energy. The local equilibrium pressure approximation (and thus also the last component in Eq. (4.55)) is an m-th order approximation to a smooth function (a hydrostatic pressure) provided that the density resembles a smooth solution. This suffices to proof the m-th order accuracy of the reconstruction $\mathcal{R}^{\mathrm{DWB}}$ analogously to the corresponding part in the proof of Theorem 4.3.1. □

4.4.2.2 Well-Balanced Property

Theorem 4.4.4. *Consider the semi-discrete scheme Eq.* (3.75) *for the compressible Euler equations with gravity* (2.34) *with a contact property (Definition 3.3.2) fulfilling numerical flux* $\boldsymbol{F}$*, the hydrostatic reconstruction* $\mathcal{R}^{\mathrm{DWB}}$ *(Eq.* (4.58)*) based on an m-th order accurate spatial reconstruction procedure* $\mathcal{R}$*, and the gravitational source term discretization* $\hat{\boldsymbol{S}}_i^{\mathrm{DWB}}$ *(Eq.* (4.46)*). The scheme is well-balanced in the sense that it exactly preserves a discrete hydrostatic equilibrium approximating an arbitrary non-periodic smooth hydrostatic equilibrium to m-th order accuracy.*

Proof. The proof of this theorem consists of two parts. First, we construct a discrete equilibrium $\hat{\boldsymbol{Q}}_i^{\mathrm{hs}}$ $(i \in \mathcal{I})$ approximating an arbitrary (sufficiently smooth) hydrostatic solution $\bar{\boldsymbol{q}}^{\mathrm{hs}}$ with m-th order accuracy such that the cell-centered pressure values obtained from $\hat{\boldsymbol{Q}}_i^{\mathrm{hs}}$ and $\hat{\boldsymbol{Q}}_j^{\mathrm{hs}}$ $(i, j \in \mathcal{I})$ satisfy the condition

$$p_{0,j}^{\mathrm{hs}} = p_{0,i}^{\mathrm{hs}} + \int_{x_i}^{x_j} S^{\rho u,\mathrm{DWB}}(\xi)\, d\xi. \tag{4.67}$$

Second, we show that the just constructed discrete hydrostatic equilibrium is exactly preserved by the scheme.

Part 1: Construction of the discrete hydrostatic solution. Let an arbitrary (but smooth enough) hydrostatic equilibrium be given

$$\bar{\boldsymbol{q}}^{\mathrm{hs,prim}}(x) = [\bar{\rho}^{\mathrm{hs}}(x), 0, \bar{p}^{\mathrm{hs}}(x)]^T \tag{4.68}$$

with gravitational acceleration $g(x)$. The corresponding equilibrium conserved variables are then

$$\bar{\boldsymbol{q}}^{\mathrm{hs}}(x) = [\bar{\rho}^{\mathrm{hs}}(x), 0, \varepsilon(\bar{\rho}^{\mathrm{hs}}(x), \bar{p}^{\mathrm{hs}}(x))]^T. \tag{4.69}$$

We stress that these are exact hydrostatic profiles, i.e., they satisfy Eq. (2.35). Let the density cell averages

$$\hat{\rho}_i^{\mathrm{hs}} := \frac{1}{\Delta x_i} I_{x\in\Omega_i}\left[\bar{\rho}^{\mathrm{hs}}(x)\right]. \tag{4.70}$$

of our discrete hydrostatic solution in every cell be defined by using the m-th order accurate quadrature rule I on the exact hydrostatic density $\bar{\rho}^{\mathrm{hs}}$. By applying the m-th order accurate standard reconstruction procedure $\mathcal{R}$ to the density cell averages $\hat{\rho}_i^{\mathrm{hs}}$ $(i \in \mathcal{I})$ we obtain the m-th order accurate approximation

$$\rho^{\mathrm{hs}}(x) = \mathcal{R}\left(x; \{\hat{\rho}_k\}_{k\in\mathcal{S}_i}\right) \qquad \text{for} \qquad x \in \Omega_i, \tag{4.71}$$

of $\bar{\rho}^{\mathrm{hs}}(x)$ on the whole domain Ω. Let us now choose a particular cell Ω_i and anchor the approximate equilibrium pressure profile at its center by setting $p_{0,i} = \bar{p}^{\mathrm{hs}}(x_i)$ in Eq. (4.48) and defining

$$p^{\mathrm{hs}}(x) = \bar{p}^{\mathrm{hs}}(x_i) + \int_{x_i}^{x} S^{\rho u,\mathrm{DWB}}(\xi)\, d\xi, \tag{4.72}$$

where the source term approximation $S^{\rho u,\mathrm{DWB}}$ is obtained from ρ^{hs} and point values of $g(x)$ as described in Section 4.4.1.1. We emphasize that exact integration is used in the definition of the approximate equilibrium pressure p^{hs}, as discussed in Section 4.4.1.1.

Then it is clear that the above $p_i^{\mathrm{hs}}(x)$ is an m-th order approximation of $p^{\mathrm{hs}}(x)$ for any $x \in \Omega$. With the local equilibrium density and pressure profile available, we readily obtain the internal energy density using the EoS as in Eq. (4.49). Applying the quadrature rule I as in Eq. (4.52), we obtain the cell-averaged internal energy density within cell Ω_i. Note that the so obtained cell-averaged conserved variables within the i-th cell are m-th order accurate approximation of the exact cell-averaged equilibrium conserved variables, i.e.,

$$\hat{\bar{\boldsymbol{q}}}_i^{\mathrm{hs}} = \frac{1}{\Delta x_i}\int_{\Omega_i} \bar{\boldsymbol{q}}^{\mathrm{hs}}(x)dx = [\hat{\rho}_i^{\mathrm{hs}}, 0, \hat{\varepsilon}_i^{\mathrm{hs}}]^T + \mathcal{O}\left(h^m\right) = \hat{\boldsymbol{Q}}_i^{\mathrm{hs}} + \mathcal{O}\left(h^m\right). \tag{4.73}$$

These are the discrete equilibrium cell-averaged conserved variables within this particular i-th cell obtained from $\boldsymbol{q}^{\mathrm{hs}}(x)$. The only thing that remains to be shown in this first part of the proof is that the cell-centered pressure values $p_{0,i}$ and $p_{0,j}$ obtained from $\boldsymbol{Q}_i^{\mathrm{hs}}$ and $\boldsymbol{Q}_j^{\mathrm{hs}}$ for any $i, j \in \mathcal{I}$ satisfy Eq. (4.67). For this we show that the cell-centered pressure $p_{0,i} > 0$ $(i \in \mathcal{I})$ in Eq. (4.48) determined by the relations (4.51) and (4.52) exists and is unique.

Existence: We only have to show that the discrete hydrostatic pressure approximation p^{hs} is positive. Since the actual pressure p is assumed to be positive in the Euler equations, the domain Ω is compact, and the pressure is continuous in the hydrostatic state, there is a minimal pressure value $p_{\min}$. The discrete hydrostatic pressure approximation has an error $\|p^{\text{hs}} - p\|_{l^1} = \mathcal{O}(h^m)$. Consequently, for sufficiently small values of h we have $\|p^{\text{hs}} - p\|_{l^1} < p_{\min}$ which implies $p^{\text{hs}} > 0$.

Uniqueness: For the ideal gas EoS uniqueness is clear because of Eq. (4.54). For any other EoS it can be shown as follows: Notice that the derivative of internal energy density with respect to pressure at constant density in Eq. (4.62) is a fundamental EoS property. This expression is related to the so-called Grüneisen coefficient

$$\Gamma = \left(\frac{\partial p_{\text{EoS}}}{\partial \varepsilon} \right)_{\rho}, \tag{4.74}$$

which measures the spacing of the isentropes in the p-V-plane ($V = 1/\rho$ is the so-called specific volume). The Grüneisen coefficient is a characteristic EoS variable and it is – for actual physical EoS – positive away from phase transitions (e.g., [117]). So, if we assume that the quadrature weights are positive, the function's derivative (4.62) will always keep the same sign away from a phase transition. Therefore, the function whose root we are seeking (Eq. (4.60)) is a strictly monotone function in the pressure variable and the root is unique.

Part 2: Well-balanced property for the discrete hydrostatic state. Due to $\left(\hat{\rho u}\right)_i^{\text{hs}} = 0$ and Eq. (4.53) we have $\hat{E}_i^{\text{hs}} = \hat{\varepsilon}_i^{\text{hs}}$. Consistency of the reconstruction in Eq. (4.58) ensures $\delta E_i(x) = 0$ for all i. Using Eq. (4.67) we have

$$\begin{aligned} p_{i+1}^{\text{hs}}(x_{i+\frac{1}{2}}) &= p_{0,i+1} + \int_{x_{i+1}}^{x_{i+\frac{1}{2}}} S^{\rho u,\text{DWB}}(\xi)\, d\xi \\ &= p_{0,i} + \int_{x_i}^{x_{i+1}} S^{\rho u,\text{DWB}}(\xi)\, d\xi + \int_{x_{i+1}}^{x_{i+\frac{1}{2}}} S^{\rho u,\text{DWB}}(\xi)\, d\xi \\ &= p_{0,i} + \int_{x_i}^{x_{i+\frac{1}{2}}} S^{\rho u,\text{DWB}}(\xi)\, d\xi = p_i^{\text{hs}}(x_{i+\frac{1}{2}}). \end{aligned} \tag{4.75}$$

Because of $\delta E_i(x) = 0$ and $\rho u = 0$ for all $i \in \mathcal{I}$ this directly yields

$$\begin{aligned} E_{i+1}^{\text{rec}}\left(x_{i+\frac{1}{2}}\right) &= \delta E_{i+1}^{\text{rec}}\left(x_{i+\frac{1}{2}}\right) + \varepsilon_{\text{EoS}}\left(\rho_{i+1}^{\text{hs}}\left(x_{i+\frac{1}{2}}\right), p^{\text{hs}}\left(x_{i+\frac{1}{2}}\right)\right) \\ &= \varepsilon_{\text{EoS}}\left(\rho_{i+1}^{\text{hs}}\left(x_{i+\frac{1}{2}}\right), p^{\text{hs}}\left(x_{i+\frac{1}{2}}\right)\right) \quad \text{and} \end{aligned} \tag{4.76}$$

$$\begin{aligned} E_{i}^{\text{rec}}\left(x_{i+\frac{1}{2}}\right) &= \delta E_{i}^{\text{rec}}\left(x_{i+\frac{1}{2}}\right) + \varepsilon_{\text{EoS}}\left(\rho_{i}^{\text{hs}}\left(x_{i+\frac{1}{2}}\right), p^{\text{hs}}\left(x_{i+\frac{1}{2}}\right)\right) \\ &= \varepsilon_{\text{EoS}}\left(\rho_{i}^{\text{hs}}\left(x_{i+\frac{1}{2}}\right), p^{\text{hs}}\left(x_{i+\frac{1}{2}}\right)\right), \end{aligned} \tag{4.77}$$

where density has a lower index because it is discontinuous, i.e., E_{i+1}^{rec} and E_i^{rec} do in general not coincide at the interface. However, since the EoS evaluation yields a unique result as discussed above, the interface pressure values obtained from E_{i+1}^{rec}

and E_i^{rec} are

$$p_{i+\frac{1}{2}}^R = p_{\mathrm{EoS}}\left(\rho_{i+1}^{\mathrm{hs}}\left(x_{i+\frac{1}{2}}\right), E_{i+1}^{\mathrm{rec}}\left(x_{i+\frac{1}{2}}\right)\right) = p^{\mathrm{hs}}\left(x_{i+\frac{1}{2}}\right) \quad \text{and} \tag{4.78}$$

$$p_{i+\frac{1}{2}}^L = p_{\mathrm{EoS}}\left(\rho_i^{\mathrm{hs}}\left(x_{i+\frac{1}{2}}\right), E_i^{\mathrm{rec}}\left(x_{i+\frac{1}{2}}\right)\right) = p^{\mathrm{hs}}\left(x_{i+\frac{1}{2}}\right). \tag{4.79}$$

Hence, the interface pressure values in the computation of the numerical flux coincide and take the value $p_{i+\frac{1}{2}}^{L/R} = p^{\mathrm{hs}}\left(x_{i+\frac{1}{2}}\right)$. Since we assumed that the numerical flux has the contact property, it is

$$\begin{aligned}&\frac{1}{\Delta x_i}\left(F^{\rho u}\left(\boldsymbol{Q}_i^{\mathrm{rec}}\left(x_{i+\frac{1}{2}}\right), \boldsymbol{Q}_{i+1}^{\mathrm{rec}}\left(x_{i+\frac{1}{2}}\right)\right) - F^{\rho u}\left(\boldsymbol{Q}_{i-1}^{\mathrm{rec}}\left(x_{i+\frac{1}{2}}\right), \boldsymbol{Q}_i^{\mathrm{rec}}\left(x_{i+\frac{1}{2}}\right)\right)\right)\\ &\quad = \frac{1}{\Delta x_i}\left(p^{\mathrm{hs}}\left(x_{i+\frac{1}{2}}\right) - p^{\mathrm{hs}}\left(x_{i+\frac{1}{2}}\right)\right) = \frac{1}{\Delta x_i}\int_{\Omega_i} S^{\rho u,\mathrm{DWB}}(x)\,dx = \hat{S}_i^{\rho u,\mathrm{DWB}}\end{aligned} \tag{4.80}$$

in the momentum equation. Furthermore, the density and energy fluxes vanish due to the contact property, the common interface pressure on both sides, and the zero velocity (Definition 3.3.2). □

Remark 4.4.5. *Periodic hydrostatic solutions have been excluded in Theorem 4.4.4 since the construction of the discrete approximation that is used in the proof can fail for periodic boundary conditions. Periodic hydrostatic states are anyway academic problems, since they cannot appear in real physical situations. However, the Discretely Well-Balanced method can still be beneficially applied to periodic hydrostatic states as will be demonstrated in Sections 4.6.1 and 4.6.3.*

4.4.2.3 Scope

The Discretely Well-Balanced method balances high order approximations of any hydrostatic state. Hence, the method is general since it is not restricted to certain classes of hydrostatic solutions. Also, unlike in the α-β and Deviation methods, no a priori knowledge of the hydrostatic solution is required.

Note, that the Discretely Well-Balanced method does not balance exact solutions of the hydrostatic equations. Instead it balances a high order approximation satisfying Eq. (4.67). Provided an arbitrary hydrostatic solution $\bar{\boldsymbol{q}}^{\mathrm{hs}}$, an approximate hydrostatic solution, which approximated $\bar{\boldsymbol{q}}^{\mathrm{hs}}$ with m-th order accuracy and is well-balanced exactly by the method can be constructed as described in the proof of Theorem 4.4.4. Since, in this construction, the source term is integrated from one point in the domain to obtain the approximate hydrostatic pressure, the difference $|\hat{E}_j - \hat{E}_j^{\mathrm{mod}}|$ tends to increase, the further x_j is away from x_i. This is a significant difference to the discrete hydrostatic state which is well-balanced in the α-β scheme (see Section 4.2.2.3). However, when the grid is refined, the approximation converges to the exact hydrostatic solution.

There are different ways to make use of the Discretely Well-Balanced method. If the simulation starts at some hydrostatic state, the procedure described above can be used to modify the initial energy, such that the Discretely Well-Balanced method can balance the hydrostatic state exactly. However, in this case, it is also possible to

simply apply the α-β or Deviation method, to balance the exact hydrostatic state. The advantage of the Discretely Well-Balanced method is that it can also be used if there is no knowledge about the hydrostatic state. In simulations, e.g., in which the initial conditions describe some dynamical state and the solution is expected to settle at some unknown hydrostatic state, the α-β and Deviation methods cannot be easily applied. The Discretely Well-Balanced method on the other hand, is able to balance an approximation to any hydrostatic state appearing in the simulation.

In simulations of stellar objects, the Euler equations are often coupled to a Poisson equation to compute the gravitational potential, which is called self-gravity. Since this is usually treated in an operator-split way, any method for compressible Euler equations with gravity can be still applied together with a Poisson solver. This means, that the gravitational potential changes in time.[7] In this case, using the α-β or Deviation method to balance a fixed hydrostatic state can even lead to inconsistency, since the solution's dependency of the gravitational potential is not taken into account correctly. The Discretely Well-Balanced method, however, can always be used to achieve more accurate results for nearly hydrostatic flows.

4.4.2.4 Stencil

In the following we determine the stencil of the modified reconstruction in the Discretely Well-Balanced method. The description of the stencil and the corresponding figure (Fig. 4.4) have been published in [12] in a similar form. Assume m to be odd[8]. To update the cell-average values $\hat{\boldsymbol{Q}}_i$, a standard m-th order method requires $\hat{\boldsymbol{Q}}_{i-\frac{m+1}{2}}, \dots, \hat{\boldsymbol{Q}}_{i+\frac{m+1}{2}}$. This includes $\frac{m-1}{2}$ cells in each direction for the reconstruction and one for the flux computations from the reconstructed values in the $i-1, i$, and $i+1$ cell.

The Discretely Well-Balanced method proposed above increases the stencil in the following way: The transformation to local hydrostatic variables requires the values of $S^{\rho u,\mathrm{DWB}}$ in each cell in the reconstruction stencil. This adds $\frac{m-1}{2}$ cells in each direction to the stencil. In total, to update the cell-average values $\hat{\boldsymbol{Q}}_i$, the methods require the values $\hat{\boldsymbol{Q}}_{i-m}, \dots, \hat{\boldsymbol{Q}}_{i+m}$. The stencil (of the reconstruction, not the total stencil) is visualized in Fig. 4.4.

Depending on the application (especially in parallel computing using a domain decomposition), this increased stencil can lead to a considerable increase in computation time and memory. As a possible solution to this problem, we propose modified methods in Section 4.5. These modified methods only use the stencil of the m-th order accurate standard scheme.

[7] Let us assume that the change in the gravitational potential is slow compared to the advective and sonic time scale. Then it still makes sense to well-balance quasi-static states, i.e., states close to hydrostatic states that have a time-dependency on a long time scale.

[8] We use m-th order CWENO methods in the Discretely Well-Balanced method. These are usually constructed for odd m

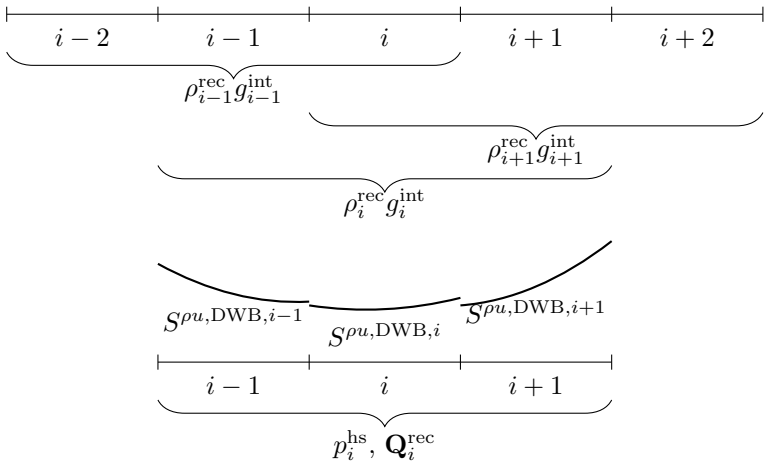


Figure 4.4: Stencil of the reconstruction in the i-th cell of the third order accurate Discretely Well-Balanced method. The local hydrostatic reconstruction which yields $\boldsymbol{Q}^{\mathrm{rec}}_i$ requires the source term approximations $S^{\rho u,\mathrm{DWB},i-1}$, $S^{\rho u,\mathrm{DWB},i}$, and $S^{\rho u,\mathrm{DWB},i+1}$ in the $i-1$, i, and $i+1$ cell respectively (shown at the bottom of the figure). Each of these source term approximation has a stencil involving one neighboring cell per dimension. The total stencil to determine $\boldsymbol{Q}^{\mathrm{rec}}_i$ thus involves five cells. This figure has, in a similar form, been shown in [12].

4.5 The Local Approximation Method

The reason for the increased stencil in the Discretely Well-Balanced method is that the source term has to be discretized in each cell of the CWENO stencil. To avoid this, the following modification is proposed in [12]. Parts of the text in Section 4.5 are similar to the text in our original article [12].

4.5.1 Description of the Local Approximation Method

To compute the hydrostatic pressure with respect to the i-th cell, we only use the source term discretization from the i-th cell. This definition is extended to the whole domain in the trivial way without using additional information. Consequently, there is no unique source term discretization. Instead it depends on the cell in which we aim to reconstruct.

To achieve this, we only have to modify Eq. (4.48) in the Discretely Well-Balanced method by using $S^{\rho u,\mathrm{DWB},i}$ instead of $S^{\rho u,\mathrm{DWB}}$:

$$p^{\mathrm{hs}}_i(x) = p_{0,i} + \int_{x_i}^{x} S^{\rho u,\mathrm{DWB},i}(\xi)\, d\xi. \tag{4.81}$$

Thus, instead of using the globally defined momentum source term approximation $S^{\rho u,\mathrm{DWB},i}$, we extrapolate the source term polynomial from the i-th cell to the neighboring cells. This only affects the reconstruction of the energy deviations. The rest of the method remains unmodified. We refer to the resulting method as *Local Approximation* (LA) method. The hydrostatic reconstruction $\mathcal{R}^{\mathrm{DWB}}$ with the modification described in Eq. (4.81) defines the corresponding reconstruction $\mathcal{R}^{\mathrm{LA}}$. If the sim-

plified approach to computing the cell-centered pressure is applied, we refer to the method as LA-fast.

4.5.2 Properties of the Local Approximation Method

In the following we discuss the accuracy and the well-balancing property of the Local Approximation method.

4.5.2.1 Accuracy

Theorem 4.5.1. *Consider the semi-discrete scheme Eq.* (3.75) *for compressible Euler equations with gravity* (2.34) *with a numerical flux* $\boldsymbol{F}$*, the hydrostatic reconstruction* $\mathcal{R}^{\mathrm{LA}}$ *(Section 4.5.1) based on an* m*-th order accurate spatial reconstruction procedure* $\mathcal{R}$*, and the gravitational source term discretization* $\hat{\boldsymbol{S}}_i^{\mathrm{DWB}}$ *(Eq.* (4.46)*).*

The scheme is consistent and at least m*-th order accurate in space for smooth solutions.*

Proof. This can be shown analogously to Theorem 4.4.3 with the only difference that the local hydrostatic pressure approximation used in the reconstruction is now actually smooth. □

4.5.2.2 Well-Balanced Property

For the Local Approximation method described in Section 4.5.1, there is no globally defined hydrostatic pressure function. Consequently, there is in general no well-defined cell-to-cell relation like Eq. (4.67), which is balanced to machine precision.

Whether the method actually succeeds in significantly reducing the discretization error at hydrostatic solutions has to be tested in numerical experiments.

4.5.2.3 Stencil

In the Local Approximation method, the reconstruction routine only requires the local hydrostatic pressure polynomial from the i-th cell (different from the Discretely Well-Balanced method). The stencil of the method is the same as the stencil of the standard method of the same formal order of accuracy as visualized in Fig. 4.5.

4.5.2.4 Scope

Similar to the Discretely Well-Balanced method, the Local Approximation method can be applied in high order simulations in which no other well-balancing methods for Euler with gravity are suitable, i.e., if the hydrostatic state that shall be balanced does not belong to a specific class and there is no a priori knowledge of it. Compared to the Discretely Well-Balanced method, the Local Approximation method has a significantly reduced stencil, which makes in favorable for example in parallel computing applications using domain decomposition.

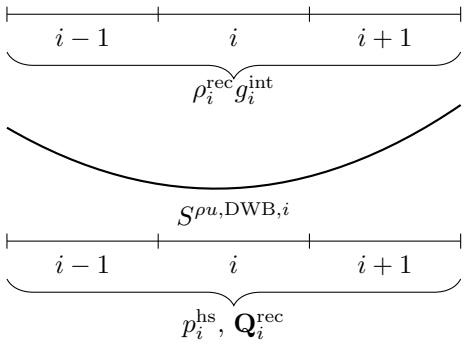


Figure 4.5: Stencil of the reconstruction in the i-th cell for third order accurate Local Approximation method. Different from the Discretely Well-Balanced method (see stencil in Fig. 4.4), the local hydrostatic reconstruction only requires the source term approximation computed in the i-th cell to compute $\boldsymbol{Q}_i^{\mathrm{rec}}$. The source term approximation $S^{\rho u,\mathrm{DWB},i}$ is for this purpose extrapolated to the neighboring cells. Thus the total stencil of the reconstruction only includes three cells, equivalent to a non-well-balanced standard method. This figure has, in a similar form, been shown in [12].

4.6 Numerical Tests

In the following we verify the accuracy and well-balancedness of the introduced methods numerically.

All the numerical experiments in this section are conducted using a custom Python code. Unless stated explicitly, we use Roe's approximate Riemann solver. The ratio of specific heats is set to $\gamma = 1.4$, the gas constant to $R = 1$, and the Stefan–Boltzmann constant to $a_{\mathrm{SB}} = 3$ if not stated explicitly. We use the first, second, third, fifth, and seventh order accurate reconstruction routines introduced in Section 3.4, possibly modified by the well-balanced methods as described in Sections 4.2 to 4.5. The source term is evaluated using Eq. (3.76) or Eq. (3.77) with the corresponding order of accuracy if it is not defined by the applied well-balanced method. To evolve the semi-discrete schemes in time we rely on the explicit Runge–Kutta methods as described in Section 3.7.1: In the first order method we use the forward Euler method (Eq. (3.85)), in the third, fifth, and seventh order method we use RK3 [98], RK5 (Eq. (3.87)), and RK10 [62], respectively. The size of the time step is chosen as described in Section 3.7.1. The domain is $\Omega = [0, 1]$, if not stated explicitly. Most of the test cases presented in the following have been conducted in the same or a similar form in at least one of the articles [13, 15, 14, 12]. Here, we collect the test cases and apply all well-balanced methods on the same setups as far as possible.

4.6.1 Isothermal Hydrostatic State

We set the isothermal hydrostatic state described in Eq. (2.36) as initial condition in the domain $[0, 1]$. The speed of sound for an isothermal hydrostatic solution can

be computed via $c = \sqrt{\gamma p/\rho} = \sqrt{\gamma}$. The sound crossing time is then given by

$$\tau := \int_\Omega \frac{1}{c} dx, \tag{4.82}$$

which yields $\tau = \sqrt{1/\gamma}$. We run the tests to the final time $t = 2\tau \approx 1.7$. In our tests, we use the linear gravitational potential $\phi(x) = 5x$, i.e., constant gravity $g = -5$, and the periodic gravitational potential $\phi(x) = 2 + 2\sin(2\pi x)$. In the tests with the linear gravitational potential, Dirichlet boundary conditions are applied. In the tests with the periodic gravitational potential we use periodic boundary conditions.

The numerical errors of the exactly well-balanced methods on a 128 cells grid are presented in Tables 4.1 and 4.2. This includes the Discretely Well-Balanced method on the discrete hydrostatic solution (dhs) as described in Section 4.4.2.3. The α-β and Discretely Well-Balanced method preserve the hydrostatic state to machine precision, the Deviation method even introduces no error at all. Convergence rates for the third, fifth, and seventh order accurate standard, Discretely Well-Balanced, and Local Approximation method can be seen in Tables 4.3 and 4.4. The convergence rates for the standard method are as expected. For the Discretely Well-Balanced and Local Approximation method, the order of accuracy is improved compared to the standard method. Also, the errors are smaller.

4.6.2 Polytropic Hydrostatic State

In the next test we consider polytropic hydrostatic states as described in Eq. (2.37). The speed of sound for this hydrostatic state is given by

$$c(x) = \sqrt{\frac{\gamma p(x)}{\rho(x)}} = \sqrt{\gamma \left(1 + \frac{1-\nu}{\nu}\phi(x)\right)} \tag{4.83}$$

assuming $\phi(x) < \frac{\nu}{\nu-1}$. We choose the polytropic exponent $\nu = 1.2$ and an ideal gas EoS with $\gamma = 1.4$. The domain, boundary conditions, gravitational potentials, and numerical methods are the same as in the previous test (Section 4.6.1). The final time is $t = 2\tau$, which is $t \approx 2.40$ for the constant gravity and $t \approx 1.87$ for the periodic gravity. The L^1-errors for the simulations with the exact well-balanced methods (again including the Discretely Well-Balanced method on a discrete hydrostatic state (dhs) as described in Section 4.4.2.3) on a grid with 128 cells are presented in Table 4.5 for the constant gravity case and in Table 4.6 for the periodic gravity case. The well-balanced errors introduce no discretization error on the hydrostatic state. In Tables 4.7 and 4.8, L^1-errors and convergence rates are shown for the Discretely Well-Balanced and Local Approximation method on the exact hydrostatic state with constant and periodic gravity respectively. Again, the results are comparable to the results from the test on the isothermal hydrostatic state: The well-balanced methods lead to a significantly increased accuracy and improved order of convergence compared to the standard methods.

Table 4.1: L^1-errors for an isothermal hydrostatic solution of the Euler equations with constant gravity after two sound crossing times computed using different methods at a 128 cells grid. The setup is described in Section 4.6.1.

first order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	4.89e-03	2.24e-03	2.58e-02
α-β WB	128	0.00e+00	2.31e-17	0.00e+00
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

second order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	1.18e-04	4.35e-05	6.43e-04
α-β WB	128	9.07e-16	6.20e-16	7.86e-16
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

third order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	2.68e-06	1.15e-06	1.38e-05
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	6.58e-16	2.83e-16	2.36e-15

fifth order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	2.39e-09	8.14e-10	1.12e-08
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	3.17e-16	2.72e-16	1.20e-16

seventh order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	8.20e-13	3.44e-13	4.15e-12
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	1.54e-16	2.63e-16	1.10e-16

Table 4.2: L^1-errors for an isothermal hydrostatic solution of the Euler equations with periodic gravity after two sound crossing times computed using different methods at a 128 cells grid. The setup is described in Section 4.6.1.

first order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	2.01e-02	6.07e-03	1.35e-01
α-β WB	128	7.21e-18	1.92e-17	1.47e-17
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

second order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	1.28e-03	5.71e-04	7.97e-03
α-β WB	128	1.60e-15	6.37e-17	2.45e-15
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

third order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	2.80e-04	1.74e-04	1.58e-03
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	1.85e-15	1.42e-16	5.22e-15

fifth order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	2.78e-06	8.99e-07	7.13e-06
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	2.24e-16	1.44e-16	4.53e-16

seventh order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	1.11e-05	5.90e-06	3.20e-05
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

Table 4.3: L^1-errors and convergence rates for an isothermal hydrostatic solution of the Euler equations with constant gravity after two sound crossing times computed using different methods. The setup is described in Section 4.6.1.

third order methods, constant gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	2.68e-06	–	1.15e-06	–	1.38e-05	–
	256	3.36e-07	3.0	1.42e-07	3.0	1.70e-06	3.0
	512	4.21e-08	3.0	1.77e-08	3.0	2.11e-07	3.0
DWB	128	1.15e-08	–	4.45e-10	–	1.75e-08	–
	256	7.13e-10	4.0	2.70e-11	4.0	1.06e-09	4.0
	512	4.44e-11	4.0	1.66e-12	4.0	6.58e-11	4.0
LA	128	9.69e-09	–	4.89e-09	–	5.21e-08	–
	256	2.62e-10	5.2	1.54e-10	5.0	1.45e-09	5.2
	512	8.45e-12	5.0	4.91e-12	5.0	3.48e-11	5.4

fifth order methods, constant gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	7.63e-08	–	2.48e-08	–	3.43e-07	–
	128	2.39e-09	5.0	8.14e-10	4.9	1.12e-08	4.9
	256	7.46e-11	5.0	2.59e-11	5.0	3.56e-10	5.0
DWB	64	5.77e-10	–	1.67e-11	–	1.05e-09	–
	128	9.11e-12	6.0	2.81e-13	5.9	1.61e-11	6.0
	256	1.46e-13	6.0	5.00e-15	5.8	2.43e-13	6.1
LA	64	2.42e-10	–	6.82e-11	–	5.48e-10	–
	128	2.52e-12	6.6	6.58e-13	6.7	6.87e-12	6.3
	256	2.49e-14	6.7	5.90e-15	6.8	6.68e-14	6.7

seventh order methods, constant gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	16	1.76e-06	–	6.66e-07	–	8.58e-06	–
	32	1.27e-08	7.1	5.29e-09	7.0	6.53e-08	7.0
	64	1.03e-10	6.9	4.34e-11	6.9	5.28e-10	7.0
DWB	16	3.83e-08	–	9.45e-09	–	7.05e-08	–
	32	1.75e-10	7.8	1.86e-11	9.0	2.91e-10	7.9
	64	7.18e-13	7.9	7.39e-14	8.0	1.16e-12	8.0
LA	16	3.13e-09	–	2.12e-10	–	4.98e-09	–
	32	1.53e-11	7.7	1.51e-12	7.1	3.75e-11	7.1
	64	5.64e-14	8.1	4.32e-15	8.5	1.18e-13	8.3

Table 4.4: L^1-errors and convergence rates for an isothermal hydrostatic solution of the Euler equations with periodic gravity after two sound crossing times computed using different methods. The setup is described in Section 4.6.1.

third order methods, periodic gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	256	4.02e-05	–	2.55e-05	–	2.19e-04	–
	512	5.23e-06	2.9	3.26e-06	3.0	2.92e-05	2.9
	1024	6.56e-07	3.0	4.77e-07	2.8	3.71e-06	3.0
DWB	256	1.57e-07	–	6.25e-08	–	4.08e-07	–
	512	9.57e-09	4.0	3.94e-09	4.0	2.48e-08	4.0
	1024	5.89e-10	4.0	2.44e-10	4.0	1.52e-09	4.0
LA	256	4.35e-07	–	1.39e-07	–	2.20e-06	–
	512	1.38e-08	5.0	5.05e-09	4.8	7.40e-08	4.9
	1024	4.37e-10	5.0	1.82e-10	4.8	2.54e-09	4.9

fifth order methods, periodic gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	2.78e-06	–	8.99e-07	–	7.13e-06	–
	256	8.95e-08	5.0	3.21e-08	4.8	2.35e-07	4.9
	512	2.81e-09	5.0	1.04e-09	4.9	7.45e-09	5.0
DWB	128	1.47e-08	–	6.76e-09	–	4.78e-08	–
	256	2.44e-10	5.9	1.16e-10	5.9	8.01e-10	5.9
	512	3.87e-12	6.0	1.83e-12	6.0	1.27e-11	6.0
LA	128	5.45e-08	–	9.11e-09	–	9.90e-08	–
	256	4.89e-10	6.8	9.83e-11	6.5	1.05e-09	6.6
	512	4.28e-12	6.8	9.65e-13	6.7	1.05e-11	6.6

seventh order methods, periodic gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	32	4.33e-04	–	9.11e-05	–	1.15e-03	–
	64	5.51e-06	6.3	2.93e-06	5.0	1.59e-05	6.2
	128	5.24e-08	6.7	2.53e-08	6.9	2.11e-07	6.2
DWB	32	1.43e-04	–	1.43e-04	–	5.02e-04	–
	64	7.03e-08	11.0	4.04e-08	11.8	2.25e-07	11.1
	128	2.72e-10	8.0	1.35e-10	8.2	8.32e-10	8.1
LA	32	7.27e-05	–	2.69e-05	–	2.75e-04	–
	64	1.19e-07	9.3	4.48e-08	9.2	2.60e-07	10.0
	128	1.24e-10	9.9	4.56e-11	9.9	3.19e-10	9.7

Table 4.5: L^1-errors for a polytropic hydrostatic solution of the Euler equations with constant gravity after two sound crossing times computed using different methods at a 128 cells grid. The setup is described in Section 4.6.2.

first order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	9.09e-03	2.44e-03	3.91e-02
α-β WB	128	0.00e+00	1.67e-17	0.00e+00
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

second order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	1.52e-04	4.67e-05	8.66e-04
α-β WB	128	1.43e-15	4.07e-16	9.91e-16
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

third order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	4.64e-06	1.73e-06	2.25e-05
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	2.48e-16	3.45e-16	6.44e-16

fifth order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	1.32e-09	3.68e-10	5.79e-09
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	4.11e-16	3.74e-16	3.17e-16

seventh order methods, constant gravity

method	N	ρ error	ρu error	E error
Standard	128	8.59e-13	3.63e-13	3.13e-12
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	3.40e-16	9.14e-16	2.97e-16

Table 4.6: L^1-errors for a polytropic hydrostatic solution of the Euler equations with periodic gravity after two sound crossing times computed using different methods at a 128 cells grid. The setup is described in Section 4.6.2.

first order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	1.40e-02	3.76e-03	4.66e-02
α-β WB	128	1.40e-17	1.70e-16	1.94e-17
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

second order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	3.01e-04	9.54e-05	1.15e-03
α-β WB	128	4.66e-15	6.28e-16	6.04e-15
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

third order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	2.67e-05	2.85e-05	1.29e-04
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	2.51e-15	1.97e-16	7.59e-15

fifth order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	6.05e-08	2.09e-09	1.00e-07
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	2.50e-15	1.98e-15	2.40e-15

seventh order methods, periodic gravity

method	N	ρ error	ρu error	E error
Standard	128	2.01e-10	2.05e-10	5.83e-10
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00
DWB (dhs)	128	6.00e-16	1.47e-15	7.23e-16

Table 4.7: L^1-errors and convergence rates for a polytropic hydrostatic solution of the Euler equations with constant gravity after two sound crossing times computed using different methods. The setup is described in Section 4.6.2.

third order methods, constant gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	4.64e-06	–	1.73e-06	–	2.25e-05	–
	256	5.84e-07	3.0	2.18e-07	3.0	2.82e-06	3.0
	512	7.32e-08	3.0	2.75e-08	3.0	3.53e-07	3.0
DWB	128	6.98e-09	–	7.98e-10	–	8.87e-09	–
	256	4.41e-10	4.0	5.05e-11	4.0	5.46e-10	4.0
	512	2.77e-11	4.0	3.16e-12	4.0	3.36e-11	4.0
LA	128	8.63e-09	–	4.79e-09	–	4.87e-08	–
	256	2.81e-10	4.9	1.58e-10	4.9	1.53e-09	5.0
	512	1.01e-11	4.8	5.16e-12	4.9	4.63e-11	5.0

fifth order methods, constant gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	4.03e-08	–	1.10e-08	–	1.77e-07	–
	128	1.32e-09	4.9	3.68e-10	4.9	5.79e-09	4.9
	256	4.20e-11	5.0	1.19e-11	5.0	1.85e-10	5.0
DWB	64	5.35e-11	–	6.13e-12	–	7.88e-11	–
	128	8.68e-13	5.9	1.04e-13	5.9	1.27e-12	6.0
	256	1.33e-14	6.0	3.47e-15	4.9	2.01e-14	6.0
LA	64	8.29e-11	–	2.34e-11	–	2.32e-10	–
	128	7.53e-13	6.8	2.18e-13	6.7	2.13e-12	6.8
	256	4.21e-15	7.5	1.93e-15	6.8	7.46e-15	8.2

seventh order methods, constant gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	9.54e-11	–	4.24e-11	–	3.74e-10	–
	128	8.59e-13	6.8	3.63e-13	6.9	3.13e-12	6.9
	256	6.75e-15	7.0	2.81e-15	7.0	2.16e-14	7.2
DWB	64	5.78e-14	–	3.14e-14	–	7.25e-14	–
	128	4.38e-16	7.0	5.98e-16	5.7	6.26e-16	6.9
	256	6.74e-17	2.7	3.86e-16	0.6	4.16e-17	3.9
LA	64	1.30e-16	–	2.49e-16	–	1.02e-16	–
	128	5.49e-17	1.2	3.36e-16	-0.4	4.81e-17	1.1
	256	8.93e-17	-0.7	4.22e-16	-0.3	5.00e-17	-0.1

Table 4.8: L^1-errors and convergence rates for a polytropic hydrostatic solution of the Euler equations with periodic gravity after two sound crossing times computed using different methods. The setup is described in Section 4.6.2.

third order methods, periodic gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	2.67e-05	–	2.85e-05	–	1.29e-04	–
	256	3.28e-06	3.0	3.89e-06	2.9	1.58e-05	3.0
	512	4.01e-07	3.0	5.03e-07	3.0	1.98e-06	3.0
DWB	128	1.04e-08	–	2.28e-08	–	3.27e-08	–
	256	5.15e-10	4.3	1.43e-09	4.0	1.61e-09	4.3
	512	2.64e-11	4.3	8.72e-11	4.0	8.17e-11	4.3
LA	128	1.13e-07	–	5.36e-08	–	2.46e-07	–
	256	3.68e-09	4.9	2.26e-09	4.6	8.15e-09	4.9
	512	1.15e-10	5.0	1.09e-10	4.4	2.58e-10	5.0

fifth order methods, periodic gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	1.88e-06	–	9.23e-08	–	3.09e-06	–
	128	6.05e-08	5.0	2.09e-09	5.5	1.00e-07	4.9
	256	1.91e-09	5.0	7.11e-11	4.9	3.18e-09	5.0
DWB	64	2.46e-09	–	5.42e-09	–	7.21e-09	–
	128	3.92e-11	6.0	8.82e-11	5.9	1.15e-10	6.0
	256	6.18e-13	6.0	1.39e-12	6.0	1.81e-12	6.0
LA	64	1.08e-08	–	3.55e-09	–	1.86e-08	–
	128	4.65e-11	7.9	5.17e-11	6.1	8.13e-11	7.8
	256	2.72e-13	7.4	7.84e-13	6.0	9.01e-13	6.5

seventh order methods, periodic gravity

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	2.59e-08	–	2.09e-08	–	6.75e-08	–
	128	2.01e-10	7.0	2.05e-10	6.7	5.83e-10	6.9
	256	1.61e-12	7.0	1.79e-12	6.8	4.90e-12	6.9
DWB	64	2.89e-11	–	5.08e-11	–	8.76e-11	–
	128	1.71e-12	4.1	1.97e-12	4.7	5.35e-12	4.0
	256	2.97e-15	9.2	1.88e-15	10.0	4.48e-15	10.2
LA	64	1.93e-10	–	2.58e-11	–	2.15e-10	–
	128	3.84e-13	9.0	9.41e-14	8.1	3.87e-13	9.1
	256	9.69e-16	8.6	1.57e-15	5.9	3.45e-15	6.8

4.6.3 Isothermal Hydrostatic State with Perturbation

In order to verify the accuracy of our numerical methods for small perturbations to hydrostatic states we consider the perturbation

$$\rho(x) = \tilde{\rho}(x), \quad u(x) = \tilde{u}(x), \quad p(x) = \tilde{p}(x) + \eta \exp\left(-100\left(x - \frac{1}{2}\right)^2\right) \tag{4.84}$$

on the isothermal hydrostatic state $(\tilde{\rho}, \tilde{u}, \tilde{p})$ from the test case in Section 4.6.1 with the periodic gravitational potential $\phi(x) = \sin(2\pi x)$. We evolve the solution to the final time $t = 0.5$. The errors and convergence rates of the standard and well-balanced methods – compared to a reference solution obtained from seventh order standard method on a grid with 2024 cells – are shown in Tables 4.9 and 4.10 for a large perturbation of $\eta = 10^{-1}$ and in Tables 4.11 and 4.12 for a small perturbation of $\eta = 10^{-5}$. The convergence rates are as expected. While there is no significant difference between the accuracy of the standard method and the well-balanced methods for the large perturbation, the well-balanced methods show an improved capability to resolve the small perturbation compared to the standard method. For the small perturbation the Discretely Well-Balanced and Local Approximation method show an increased order of convergence. In Figs. 4.6 to 4.10 the density deviations from the isothermal hydrostatic state at final time are visualized for the different methods and different grid sizes. All of the well-balanced methods capture the perturbation more accurately than the standard method. The exact well-balanced methods (α-β and Deviation) are accurate even at low resolutions. The approximate well-balanced methods (Discretely Well-Balanced and Local Approximation) are less accurate than the exact well-balanced methods. However, compared to the standard method, they improve the result significantly.

4.6.4 Riemann Problem on an Isothermal Hydrostatic State

For the next test we use the initial data given by

$$\rho(x) := \begin{cases} ac\exp(-a\phi(x)) & \text{if } x < x_0, \\ b\exp(-b\phi(x)) & \text{if } x \geq x_0, \end{cases} \tag{4.85}$$

$$p(x) := \begin{cases} c\exp(-a\phi(x)) & \text{if } x < x_0, \\ \exp(-b\phi(x)) & \text{if } x \geq x_0. \end{cases} \tag{4.86}$$

This test has been conducted for the Discretely Well-Balanced and Local Approximation method in [12]. The description of the test case and analysis are similar to the one in this original publication. Equations (4.85) and (4.86) describe a piecewise isothermal hydrostatic solution with a jump discontinuity at $x = x_0$, which gives rise to all three waves of the Euler equations; the parameters are chosen as $x_0 = 0.125, a = 0.5, b = 1, c = 2$. An ideal gas EoS with $\gamma = 1.4$ is applied. We set these initial data on the domain $[0, 0.25]$ and evolve them to the final time $t = 0.02$ using our formally second, third, fifth, and seventh order accurate standard and well-balanced methods on a grid with 128 cells and Dirichlet boundary conditions. Additionally, we run the same tests using unlimited reconstruction (Section 3.4.1) instead of minmod or CWENO reconstruction. As a reference solution to compute the

Table 4.9: L^1-errors and convergence rates for a perturbation on an isothermal hydrostatic solution of the Euler equations with periodic gravity after one quarter of the sound crossing time computed using different methods. The setup is described in Section 4.6.3.

first order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	1.09e-02	–	1.83e-02	–	3.65e-02	–
	256	5.52e-03	1.0	9.26e-03	1.0	1.85e-02	1.0
	512	2.79e-03	1.0	4.67e-03	1.0	9.29e-03	1.0
α-β WB	128	1.64e-03	–	1.98e-03	–	5.74e-03	–
	256	8.81e-04	0.9	1.07e-03	0.9	3.09e-03	0.9
	512	4.57e-04	0.9	5.58e-04	0.9	1.60e-03	0.9
Deviation WB	128	1.65e-03	–	2.02e-03	–	5.85e-03	–
	256	8.83e-04	0.9	1.09e-03	0.9	3.16e-03	0.9
	512	4.59e-04	0.9	5.71e-04	0.9	1.64e-03	0.9

second order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	1.32e-04	–	1.01e-04	–	4.65e-04	–
	256	3.29e-05	2.0	2.42e-05	2.1	1.16e-04	2.0
	512	8.20e-06	2.0	5.93e-06	2.0	2.88e-05	2.0
α-β WB	128	7.76e-05	–	9.17e-05	–	2.77e-04	–
	256	1.94e-05	2.0	2.29e-05	2.0	6.91e-05	2.0
	512	4.86e-06	2.0	5.74e-06	2.0	1.73e-05	2.0
Deviation WB	128	7.99e-05	–	9.49e-05	–	2.83e-04	–
	256	1.98e-05	2.0	2.37e-05	2.0	7.00e-05	2.0
	512	4.95e-06	2.0	5.92e-06	2.0	1.75e-05	2.0

third order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	1.22e-04	–	1.77e-04	–	4.39e-04	–
	256	2.23e-05	2.5	3.15e-05	2.5	7.92e-05	2.5
	512	3.99e-06	2.5	5.22e-06	2.6	1.46e-05	2.4
Deviation WB	128	1.10e-04	–	1.33e-04	–	3.88e-04	–
	256	2.10e-05	2.4	2.55e-05	2.4	7.41e-05	2.4
	512	3.58e-06	2.6	4.36e-06	2.5	1.26e-05	2.6
DWB	128	1.09e-04	–	1.28e-04	–	3.78e-04	–
	256	2.15e-05	2.3	2.55e-05	2.3	7.48e-05	2.3
	512	3.47e-06	2.6	4.60e-06	2.5	1.36e-05	2.5
LA	128	1.08e-04	–	1.28e-04	–	3.76e-04	–
	256	2.14e-05	2.3	2.54e-05	2.3	7.47e-05	2.3
	512	3.47e-06	2.6	4.60e-06	2.5	1.36e-05	2.5

Table 4.10: L^1-errors and convergence rates for a perturbation on an isothermal hydrostatic solution of the Euler equations with periodic gravity after one quarter of the sound crossing time computed using different methods. The setup is described in Section 4.6.3.

fifth order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	5.15e-07	–	6.37e-07	–	1.84e-06	–
	256	1.72e-08	4.9	2.10e-08	4.9	6.11e-08	4.9
	512	5.48e-10	5.0	6.68e-10	5.0	1.95e-09	5.0
Deviation WB	128	6.65e-07	–	8.08e-07	–	2.36e-06	–
	256	2.24e-08	4.9	2.73e-08	4.9	7.97e-08	4.9
	512	7.21e-10	5.0	8.80e-10	5.0	2.57e-09	5.0
DWB	128	7.89e-07	–	9.40e-07	–	2.80e-06	–
	256	2.80e-08	4.8	3.30e-08	4.8	9.92e-08	4.8
	512	9.47e-10	4.9	1.12e-09	4.9	3.36e-09	4.9
LA	128	7.89e-07	–	9.40e-07	–	2.80e-06	–
	256	2.80e-08	4.8	3.30e-08	4.8	9.92e-08	4.8
	512	9.47e-10	4.9	1.12e-09	4.9	3.36e-09	4.9

seventh order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	2.02e-08	–	2.46e-08	–	7.11e-08	–
	256	2.03e-10	6.6	2.46e-10	6.6	7.15e-10	6.6
	512	1.85e-12	6.8	2.25e-12	6.8	6.49e-12	6.8
Deviation WB	128	3.24e-08	–	3.88e-08	–	1.14e-07	–
	256	4.08e-10	6.3	4.88e-10	6.3	1.45e-09	6.3
	512	4.36e-12	6.5	5.21e-12	6.6	1.54e-11	6.6
DWB	128	2.82e-08	–	3.37e-08	–	9.95e-08	–
	256	3.37e-10	6.4	4.04e-10	6.4	1.20e-09	6.4
	512	3.68e-12	6.5	4.40e-12	6.5	1.30e-11	6.5
LA	128	2.82e-08	–	3.37e-08	–	9.95e-08	–
	256	3.37e-10	6.4	4.04e-10	6.4	1.20e-09	6.4
	512	3.68e-12	6.5	4.40e-12	6.5	1.30e-11	6.5

Table 4.11: L^1-errors and convergence rates for a perturbation on an isothermal hydrostatic solution of the Euler equations with periodic gravity after one quarter of the sound crossing time computed using different methods. The setup is described in Section 4.6.3.

first order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	1.15e-02	–	1.73e-02	–	3.70e-02	–
	256	5.84e-03	1.0	8.74e-03	1.0	1.88e-02	1.0
	512	2.94e-03	1.0	4.40e-03	1.0	9.45e-03	1.0
α-β WB	128	1.61e-07	–	1.92e-07	–	5.54e-07	–
	256	8.58e-08	0.9	1.03e-07	0.9	2.96e-07	0.9
	512	4.44e-08	1.0	5.35e-08	0.9	1.53e-07	0.9
Deviation WB	128	1.62e-07	–	1.96e-07	–	5.66e-07	–
	256	8.67e-08	0.9	1.06e-07	0.9	3.04e-07	0.9
	512	4.49e-08	0.9	5.48e-08	0.9	1.58e-07	0.9

second order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	1.02e-04	–	4.25e-05	–	3.52e-04	–
	256	2.59e-05	2.0	7.71e-06	2.5	8.94e-05	2.0
	512	6.51e-06	2.0	1.61e-06	2.3	2.25e-05	2.0
α-β WB	128	7.74e-09	–	8.83e-09	–	2.68e-08	–
	256	1.93e-09	2.0	2.21e-09	2.0	6.70e-09	2.0
	512	4.84e-10	2.0	5.53e-10	2.0	1.68e-09	2.0
Deviation WB	128	7.84e-09	–	9.10e-09	–	2.72e-08	–
	256	1.94e-09	2.0	2.27e-09	2.0	6.75e-09	2.0

third order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	6.23e-06	–	2.05e-05	–	1.38e-05	–
	256	7.63e-07	3.0	2.58e-06	3.0	1.65e-06	3.1
	512	9.44e-08	3.0	3.23e-07	3.0	2.02e-07	3.0
Deviation WB	128	1.70e-09	–	1.99e-09	–	5.98e-09	–
	256	2.16e-10	3.0	2.52e-10	3.0	7.58e-10	3.0
	512	2.71e-11	3.0	3.16e-11	3.0	9.52e-11	3.0
DWB	128	7.44e-08	–	6.27e-08	–	2.49e-07	–
	256	4.67e-09	4.0	3.95e-09	4.0	1.57e-08	4.0
	512	2.94e-10	4.0	2.51e-10	4.0	9.85e-10	4.0
LA	128	6.64e-08	–	4.86e-08	–	2.21e-07	–
	256	4.19e-09	4.0	3.06e-09	4.0	1.39e-08	4.0
	512	2.64e-10	4.0	1.96e-10	4.0	8.75e-10	4.0

Table 4.12: L^1-errors and convergence rates for a perturbation on an isothermal hydrostatic solution of the Euler equations with periodic gravity after one quarter sound crossing time computed using different methods. The setup is described in Section 4.6.3.

fifth order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	2.08e-08	–	3.69e-08	–	6.61e-08	–
	256	6.36e-10	5.0	1.17e-09	5.0	2.01e-09	5.0
	512	1.96e-11	5.0	3.69e-11	5.0	6.21e-11	5.0
Deviation WB	128	1.99e-11	–	2.38e-11	–	6.89e-11	–
	256	6.31e-13	5.0	7.55e-13	5.0	2.19e-12	5.0
	512	2.00e-14	5.0	2.39e-14	5.0	6.84e-14	5.0
DWB	128	3.22e-10	–	9.66e-11	–	1.10e-09	–
	256	5.15e-12	6.0	1.77e-12	5.8	1.75e-11	6.0
	512	8.42e-14	5.9	3.64e-14	5.6	2.85e-13	5.9
LA	128	1.93e-10	–	8.04e-11	–	6.55e-10	–
	256	3.13e-12	5.9	1.54e-12	5.7	1.06e-11	6.0
	512	5.35e-14	5.9	3.33e-14	5.5	1.79e-13	5.9

seventh order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	128	7.05e-11	–	9.87e-11	–	2.95e-10	–
	256	5.36e-13	7.0	7.94e-13	7.0	2.29e-12	7.0
	512	6.29e-15	6.4	6.59e-15	6.9	1.87e-14	6.9
Deviation WB	128	3.79e-13	–	4.35e-13	–	1.31e-12	–
	256	3.40e-15	6.8	3.76e-15	6.9	1.02e-14	7.0
	512	1.97e-15	0.8	9.09e-16	2.0	2.70e-15	1.9
DWB	128	1.44e-12	–	6.95e-13	–	5.02e-12	–
	256	6.82e-15	7.7	4.16e-15	7.4	2.18e-14	7.8
	512	4.34e-15	0.7	1.59e-15	1.4	4.52e-15	2.3
LA	128	1.01e-12	–	9.10e-13	–	3.28e-12	–
	256	5.51e-15	7.5	4.85e-15	7.6	1.67e-14	7.6
	512	4.03e-15	0.5	1.51e-15	1.7	4.70e-15	1.8

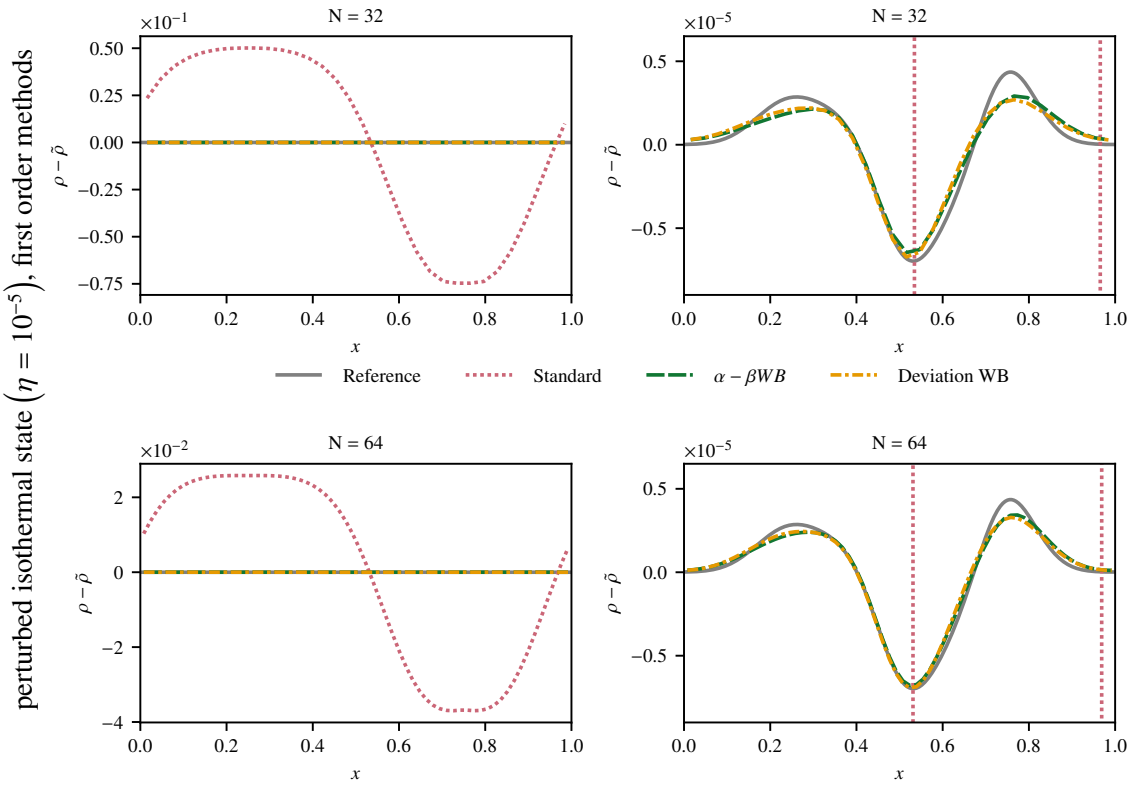


Figure 4.6: Density deviations from the isothermal hydrostatic state for the test with a perturbation on the hydrostatic state described in Section 4.6.3. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, in simulations with the first order accurate methods . On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in the y-axis of the left panels.

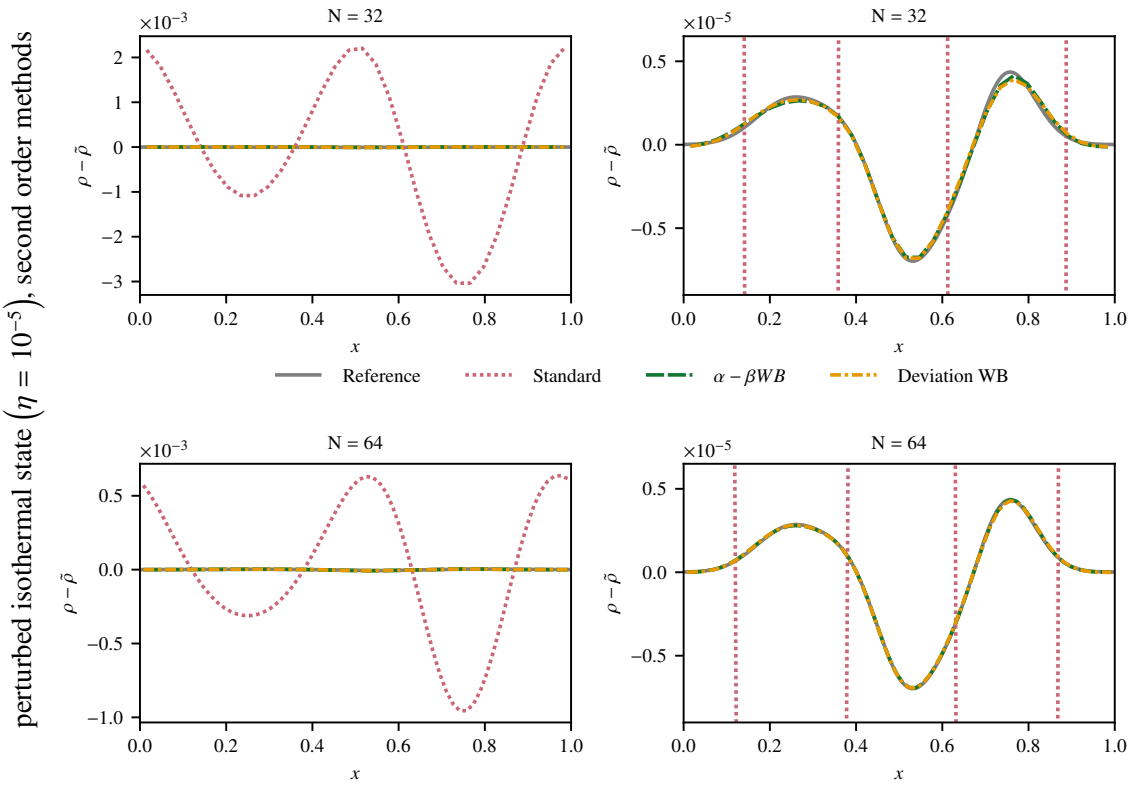


Figure 4.7: Density deviations from the isothermal hydrostatic state for the test with a perturbation on the hydrostatic state described in Section 4.6.3. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, in simulations with the second order accurate methods . On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in the y-axis of the left panels.

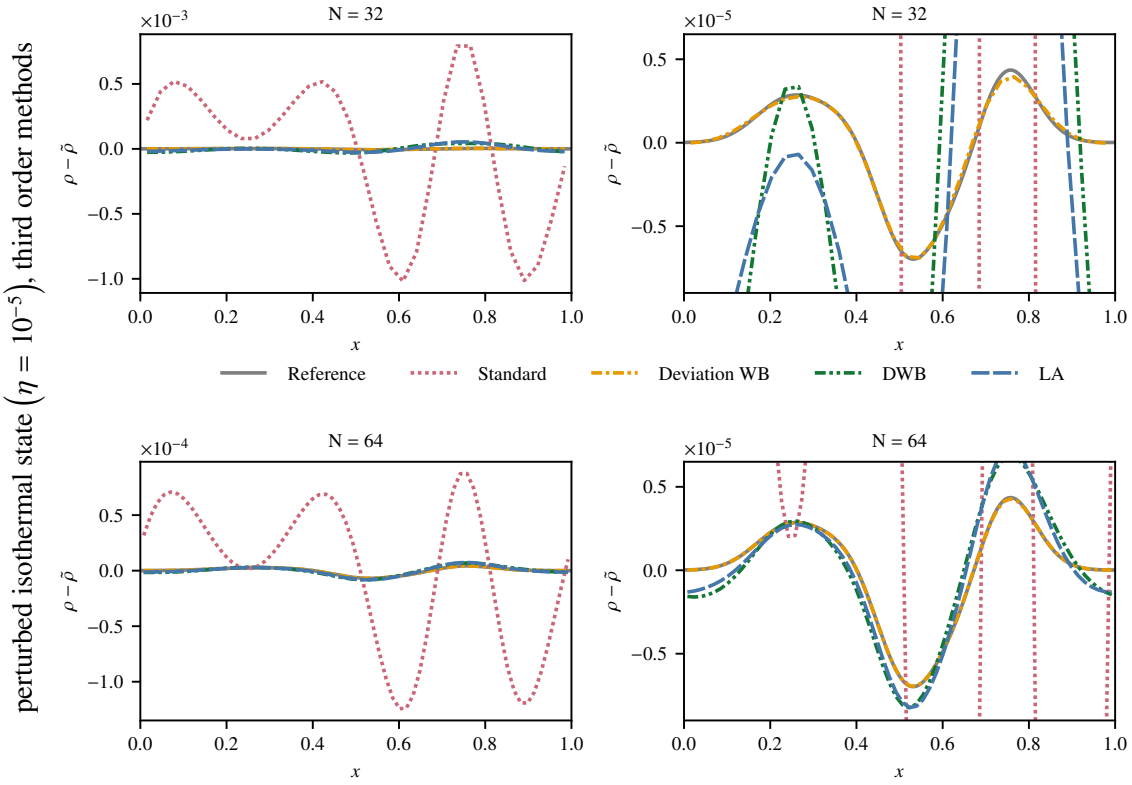


Figure 4.8: Density deviations from the isothermal hydrostatic state for the test with a perturbation on the hydrostatic state described in Section 4.6.3. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, in simulations with the third order accurate methods . On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in the y-axis of the left panels.

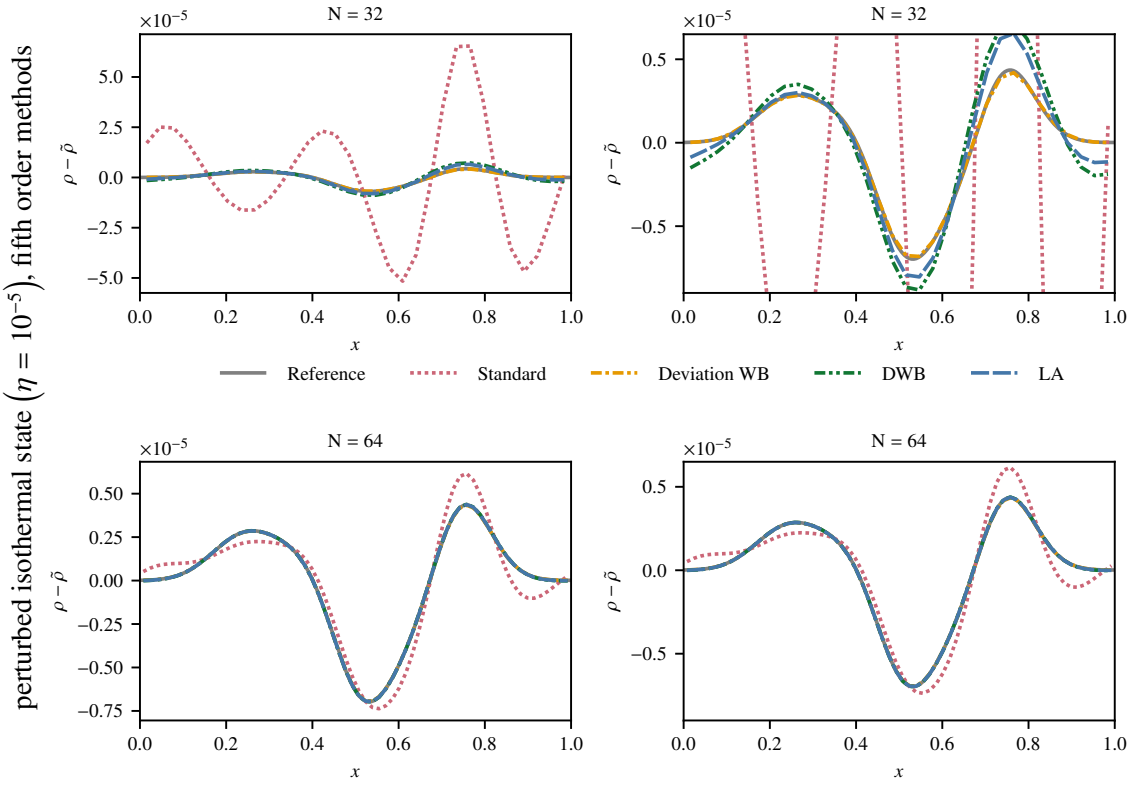


Figure 4.9: Density deviations from the isothermal hydrostatic state for the test with a perturbation on the hydrostatic state described in Section 4.6.3. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, in simulations with the fifth order accurate methods . On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in the y-axis of the left panels.

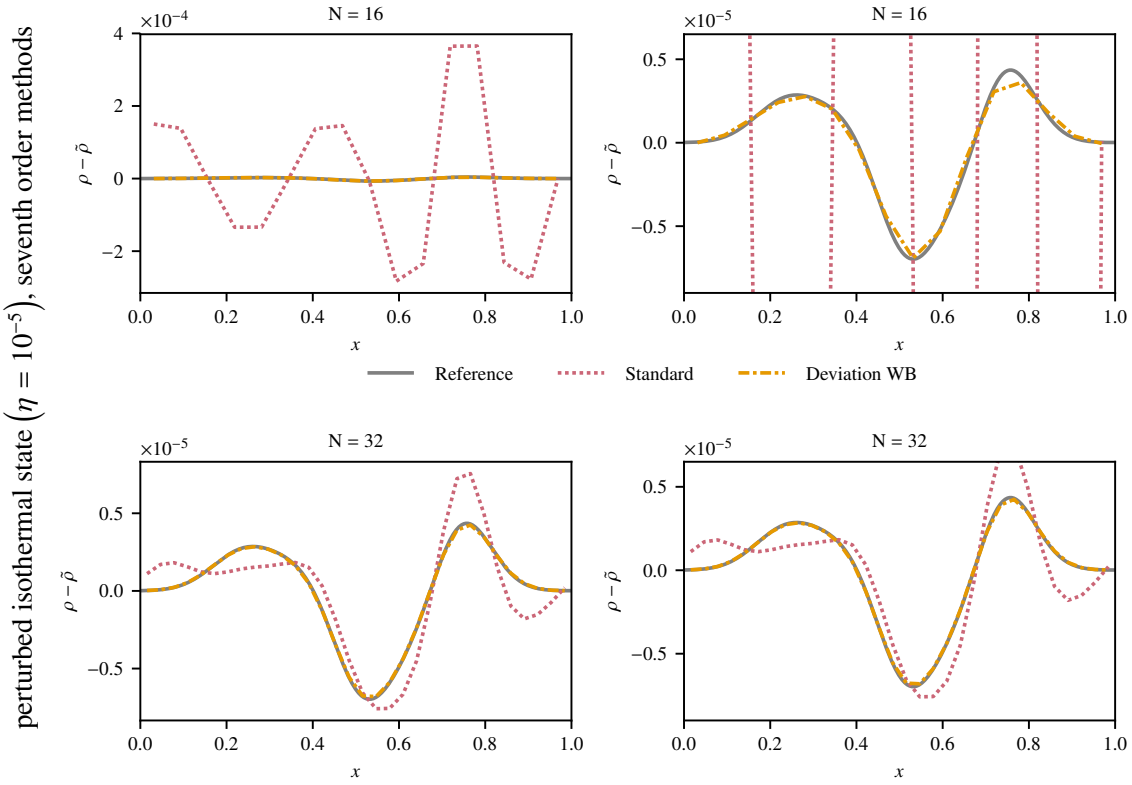


Figure 4.10: Density deviations from the isothermal hydrostatic state for the test with a perturbation on the hydrostatic state described in Section 4.6.3. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, in simulations with the seventh order accurate methods . On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in the y-axis of the left panels.

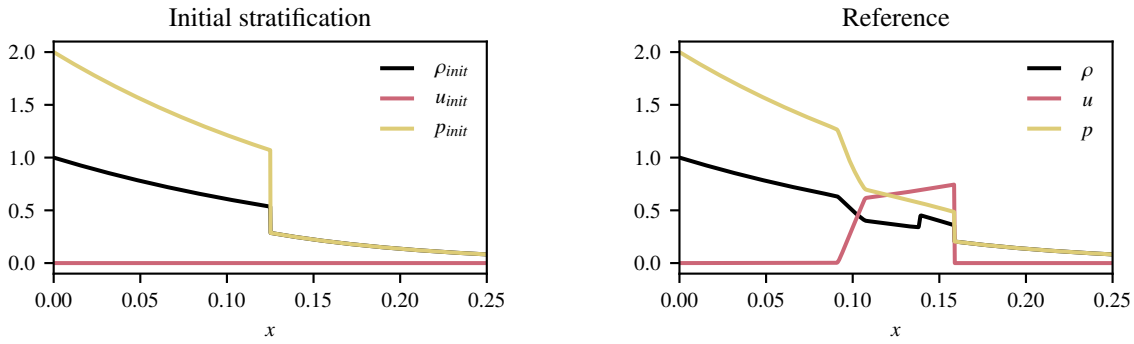


Figure 4.11: Initial data (left panel) and reference solution at final time (right panel) for the tests performed in Section 4.6.4 are shown. The reference solution has been computed with a standard first order accurate method on a grid with 32768 cells.

error we use a numerical solution obtained using a standard first order method with 32768 cells. The initial data and reference solution are visualized in Fig. 4.11. In Fig. 4.12 we see the numerical results at final time for the Discretely Well-Balanced methods. The simulations using unlimited polynomial reconstruction and interpolation introduce spurious oscillations. When limiting (minmod or CWENO) is used, no oscillations are visible. Using any of the other methods with the same formal order of accuracy leads to visually very similar results, hence we omit showing them for brevity.

To give quantitative results, we also compute the total variation of the solution at final time for all methods. The total variation of a quantity $\alpha = \rho, \rho u, E$ of a numerical solution is defined by

$$\mathrm{TV}(\alpha) := \sum_{i=1}^{N} |\alpha_i - \alpha_{i-1}|. \tag{4.87}$$

In Table 4.13 we present the difference in total variation relative to the total variation of the reference solution

$$\theta(\alpha) := \frac{\mathrm{TV}(\alpha)}{\mathrm{TV}(\alpha_{\mathrm{ref}})} - 1. \tag{4.88}$$

A negative value of θ indicates that the total variation is smaller than in the reference solution. A positive value of θ means that there are additional oscillations. Since the reference solution is computed with a first order method on a fine grid, it can be expected to be accurate and free of oscillations. In Tables 4.13 and 4.14, the θ values for different methods with and without limiting are presented alongside the L_1 errors. All unlimited methods introduce spurious oscillations. The methods using limiting lead to a decrease in total variation in conserved variables compared to the reference solution. The only exceptions are the seventh order accurate Discretely Well-Balanced and Local Approximation method, for which the θ value signals minor (compared to the oscillations in the solutions obtained using the unlimited methods) oscillations in the momentum. Overall, this indicates that the limited well-balanced methods are as robust as the limited standard methods.

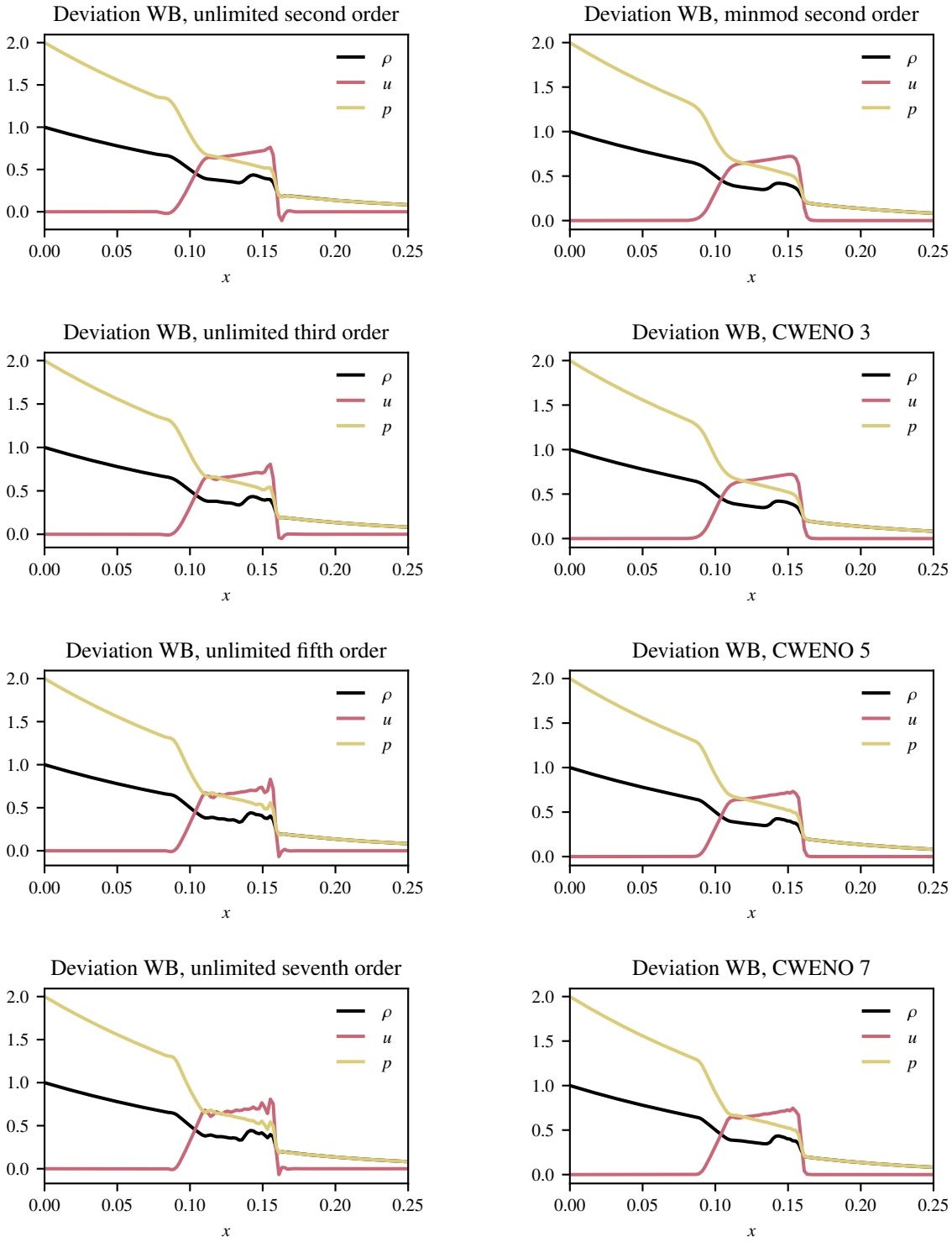


Figure 4.12: Simulation results for the tests performed in Section 4.6.4. The formally second, third, fifth and seventh order accurate Deviation methods are used with and without limiting on 128 cells grid.

Table 4.13: Errors and total variation for the robustness test from Section 4.6.4 at final time $t = 0.02$. The formally second and third order accurate standard and well-balanced methods are used with and without limiting on a 128 cells grid. The oscillation indicator θ is defined in Eq. (4.88).

second order methods

method	ρ error	$\theta(\rho)$	ρu error	$\theta(\rho u)$	E error	$\theta(E)$
unlimited						
Standard	6.85e-04	1.66e-03	8.11e-04	1.34e-01	2.57e-03	1.94e-02
α-β WB	6.61e-04	-1.21e-02	8.04e-04	1.32e-01	2.55e-03	1.36e-02
Deviation WB	6.85e-04	1.66e-03	8.11e-04	1.34e-01	2.57e-03	1.94e-02
minmod						
Standard	8.39e-04	-5.80e-02	9.94e-04	-5.07e-02	3.08e-03	-1.00e-03
α-β WB	8.19e-04	-5.86e-02	9.89e-04	-5.33e-02	3.07e-03	-1.01e-03
Deviation WB	8.31e-04	-5.92e-02	9.87e-04	-5.59e-02	3.09e-03	-1.00e-03

third order methods

method	ρ error	$\theta(\rho)$	ρu error	$\theta(\rho u)$	E error	$\theta(E)$
unlimited						
Standard	6.65e-04	1.08e-02	8.01e-04	2.10e-01	2.52e-03	4.59e-02
Deviation WB	6.65e-04	1.08e-02	8.01e-04	2.10e-01	2.52e-03	4.59e-02
DWB	6.65e-04	1.09e-02	8.02e-04	2.11e-01	2.52e-03	4.61e-02
LA	6.65e-04	1.09e-02	8.02e-04	2.11e-01	2.52e-03	4.61e-02
CWENO 3						
Standard	7.62e-04	-5.50e-02	8.88e-04	-4.38e-02	2.66e-03	-1.01e-03
Deviation WB	7.54e-04	-5.63e-02	8.84e-04	-5.18e-02	2.67e-03	-1.01e-03
DWB	7.66e-04	-5.28e-02	8.92e-04	-4.31e-02	2.71e-03	-1.01e-03
LA	7.65e-04	-5.28e-02	8.91e-04	-4.32e-02	2.71e-03	-1.01e-03

Table 4.14: Errors and total variation for the robustness test from Section 4.6.4 at final time $t = 0.02$. The formally fifth and seventh order accurate standard and well-balanced methods are used with and without limiting on a 128 cells grid. The oscillation indicator θ is defined in Eq. (4.88).

fifth order methods

method	ρ error	$\theta(\rho)$	ρu error	$\theta(\rho u)$	E error	$\theta(E)$
unlimited						
Standard	5.47e-04	9.45e-02	6.63e-04	4.43e-01	2.13e-03	1.18e-01
Deviation WB	5.47e-04	9.45e-02	6.63e-04	4.43e-01	2.13e-03	1.18e-01
DWB	5.47e-04	9.43e-02	6.64e-04	4.44e-01	2.13e-03	1.19e-01
LA	5.47e-04	9.44e-02	6.63e-04	4.44e-01	2.13e-03	1.19e-01
CWENO 5						
Standard	5.71e-04	-5.00e-02	6.78e-04	-4.12e-02	2.04e-03	-1.01e-03
Deviation WB	5.71e-04	-4.92e-02	6.68e-04	-4.09e-02	2.07e-03	-1.01e-03
DWB	5.59e-04	-4.68e-02	6.47e-04	-3.97e-02	2.03e-03	-1.01e-03
LA	5.59e-04	-4.69e-02	6.47e-04	-4.00e-02	2.03e-03	-1.01e-03

seventh order methods

method	ρ error	$\theta(\rho)$	ρu error	$\theta(\rho u)$	E error	$\theta(E)$
unlimited						
Standard	5.60e-04	1.46e-01	7.21e-04	6.22e-01	2.18e-03	1.84e-01
Deviation WB	5.60e-04	1.46e-01	7.21e-04	6.22e-01	2.18e-03	1.84e-01
DWB	5.60e-04	1.46e-01	7.22e-04	6.23e-01	2.18e-03	1.84e-01
LA	5.60e-04	1.46e-01	7.22e-04	6.23e-01	2.18e-03	1.84e-01
CWENO 7						
Standard	4.99e-04	-2.82e-02	5.57e-04	-1.01e-03	1.73e-03	-1.01e-03
Deviation WB	4.88e-04	-2.75e-02	5.51e-04	-8.11e-07	1.73e-03	-1.01e-03
DWB	4.86e-04	-2.59e-02	5.32e-04	1.86e-03	1.72e-03	-9.15e-04
LA	4.87e-04	-2.57e-02	5.35e-04	3.59e-03	1.73e-03	-1.01e-03

4.6.5 Ideal Gas with Radiation Pressure: Polytropic Hydrostatic State

In astrophysical applications the EoS can be different from the ideal gas EoS. In fact, in many cases the conversion between internal energy and pressure (while knowing density) cannot be computed in an explicit way. In this and the following section we consider a gas which is subject to the EoS for an ideal gas with radiation pressure (see Section 2.3.1.2). The conversion is then given as an implicit relation, which we solve using Newton's method (e.g., [49, 133]). While this does not pose any issue for the Deviation and α-β methods, it affects the performance of the Discretely Well-Balanced and Local Approximation methods significantly, as we will see in the following tests. This is due to the fact that the hydrostatic pressure integrals used to find a local approximation of hydrostatic states are now approximated via numerical quadratures instead of exact integration. As in the ideal gas case before, we also use $\gamma = 1.4$ in the EoS for an ideal gas with radiation pressure and for the Stefan–Boltzmann constant we use $a_{SB} = 3$.

In this test, we once more consider the polytropic hydrostatic state given in Section 2.4.1 on the periodic gravitational potential $\phi(x) = \sin(2\pi x)$. In the derivation of this hydrostatic state the EoS is not explicitly used, such that it is a solution for any arbitrary EoS. The speed of sound for an ideal gas with radiation pressure is given by

$$c = \sqrt{\frac{\Gamma_1 p}{\rho}}, \quad \text{where} \quad \Gamma_1 = \beta + \frac{(4-3\beta)^2(\gamma-1)}{\beta + 12(\gamma-1)(1-\beta)} \quad \text{with} \quad \beta = \frac{\rho T}{p} \tag{4.89}$$

and we use Newton's method (e.g., [49, 133]) to compute T and thus p from conserved variables (see Section 2.3.1.2). The sound crossing time for this polytropic setup, computed from the discretized initial data (on 128 cells) using Eq. (4.82) and Eq. (4.89), is $\tau \approx 0.7$. We run the test to a final time of $t = 10\tau$ using the standard and well-balanced methods with different orders of accuracy. The L^1 errors for the exactly well-balanced methods (α-β and Deviation) together with the standard method for a resolution of $N = 128$ are given in Table 4.15. L^1-errors and convergence rates for the approximate well-balanced methods together with the standard method are given in Table 4.16. For the approximate well-balanced methods, we use the two different approaches of computing the cell-centered hydrostatic pressure introduced in Section 4.4.1.3. In the methods called DWB and LA, the iterative approach of computing the cell-centered pressure is applied, which means that the DWB method satisfies Theorem 4.4.4. In the methods DWB-fast and LA-fast, we approximate the cell-centered hydrostatic pressure using the EoS on the cell-centered point values of conserved variables obtained from the reconstruction polynomial. This approach has also been described in Section 4.4.1.3. Using any of the well-balanced methods significantly improves the result compared to the standard method. However, there is no increased rate of convergence (which we saw in the case of an ideal gas EoS for the DWB and LA method). It gets evident that the Local Approximation method with the fast computation of the cell-centered pressure is at least as accurate as the other approximate well-balanced methods. Since this method has a smaller stencil than the Discretely Well-Balanced methods and

Table 4.15: L^1-errors for a polytropic hydrostatic solution of the Euler equations with the EoS for an ideal gas with radiation pressure after ten sound crossing times computed using different methods at a 128 cells grid. The setup is described in Section 4.6.5.

first order methods

method	N	ρ error	ρu error	E error
Standard	128	2.00e-01	1.27e-02	4.49e-01
α-β WB	128	8.93e-15	1.47e-15	3.79e-14
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

second order methods

method	N	ρ error	ρu error	E error
Standard	128	1.55e-02	7.47e-04	1.48e-02
α-β WB	128	3.21e-14	2.84e-15	1.28e-13
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

third order methods

method	N	ρ error	ρu error	E error
Standard	128	1.03e-02	1.27e-04	3.91e-03
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

fifth order methods

method	N	ρ error	ρu error	E error
Standard	128	1.98e-06	4.37e-08	2.39e-06
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

seventh order methods

method	N	ρ error	ρu error	E error
Standard	128	4.84e-08	5.15e-10	1.76e-08
Deviation WB	128	0.00e+00	0.00e+00	0.00e+00

saves computational cost compared to the methods with iterative computation of the cell-centered pressure, we consider it to be the favorable method amongst the approximate well-balanced methods for a non-ideal EoS. Hence, in the following test, we include the LA-fast method as the only representative from the class of approximate well-balanced methods.

4.6.6 Ideal Gas with Radiation Pressure: Polytropic Hydrostatic State with Perturbation

To the setup from Section 4.6.5 we add the perturbation

$$\rho(x) = \tilde{\rho}(x), \quad u(x) = \tilde{u}(x), \quad p(x) = \tilde{p}(x) + \eta \exp\left(-100\left(x - \frac{1}{2}\right)^2\right), \tag{4.90}$$

Table 4.16: L^1-errors and convergence rates for a polytropic hydrostatic solution of the Euler equations with the EoS for an ideal gas with radiation pressure after ten sound crossing times computed using different formally third and fifth order accurate methods. The setup is described in Section 4.6.5.

third order methods

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	32	1.16e-01	–	5.44e-03	2.0	1.18e-01	2.1
	64	3.73e-02	1.6	8.53e-04	2.7	2.20e-02	2.4
	128	1.03e-02	1.9	1.27e-04	2.7	3.91e-03	2.5
DWB	32	2.01e-02	–	5.61e-04	–	1.97e-02	–
	64	5.45e-03	1.9	1.00e-04	2.5	3.36e-03	2.6
	128	1.27e-03	2.1	1.55e-05	2.7	5.43e-04	2.6
DWB-fast	32	2.67e-02	–	7.73e-04	–	2.86e-02	–
	64	7.42e-03	1.8	1.37e-04	2.5	4.54e-03	2.7
	128	1.75e-03	2.1	2.10e-05	2.7	7.29e-04	2.6
LA	32	1.04e-02	–	3.05e-04	–	9.88e-03	–
	64	2.20e-03	2.2	3.56e-05	3.1	1.21e-03	3.0
	128	3.92e-04	2.5	4.25e-06	3.1	1.26e-04	3.3
LA-fast	32	1.04e-02	–	3.04e-04	–	9.86e-03	–
	64	2.20e-03	2.2	3.56e-05	3.1	1.21e-03	3.0
	128	3.92e-04	2.5	4.25e-06	3.1	1.26e-04	3.3

fifth order methods

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	32	1.73e-03	–	4.33e-05	–	2.03e-03	–
	64	6.10e-05	4.8	1.39e-06	5.0	7.39e-05	4.8
	128	1.98e-06	4.9	4.37e-08	5.0	2.39e-06	4.9
DWB	32	5.66e-05	–	1.65e-06	–	7.06e-05	–
	64	2.00e-06	4.8	5.71e-08	4.9	2.56e-06	4.8
	128	6.42e-08	5.0	1.68e-09	5.1	8.34e-08	4.9
DWB-fast	32	2.00e-04	–	5.22e-06	–	2.67e-04	–
	64	7.08e-06	4.8	1.65e-07	5.0	8.19e-06	5.0
	128	2.02e-07	5.1	5.26e-09	5.0	2.55e-07	5.0
LA	32	6.03e-05	–	1.71e-06	–	7.52e-05	–
	64	2.01e-06	4.9	5.79e-08	4.9	2.58e-06	4.9
	128	6.43e-08	5.0	1.69e-09	5.1	8.36e-08	4.9
LA-fast	32	8.22e-05	–	2.16e-06	–	1.08e-04	–
	64	2.26e-06	5.2	3.27e-08	6.0	2.80e-06	5.3
	128	5.62e-08	5.3	2.20e-09	3.9	7.23e-08	5.3

where the functions with $\tilde{\cdot}$[9] describe the polytropic hydrostatic state from Section 4.6.5. This test, using the same methods as in Section 4.6.5, is run to the final time of $t = 0.25\tau$ at different grid resolutions for a large ($\eta = 0.1$) and a small ($\eta = 10^{-5}$) perturbation. The corresponding L^1-errors and convergence rates are shown in Tables 4.17 and 4.18 for the large perturbation and in Tables 4.19 and 4.20 for the small perturbation. The density deviation from the hydrostatic background is visualized in Figs. 4.13 to 4.17 for the different methods for $\eta = 0.1$. All well-balanced methods show an improved capability to capture the deviations from the hydrostatic state compared to the standard method. The exact well-balanced methods (α-β WB and Deviation WB) are even more accurate than the approximate well-balanced method (LA-fast). Surprisingly, the density and energy errors in Table 4.19 are exactly the same for the first and second order accurate α-β and Deviation method and they display second order convergence rates. Presumably, in these tests there is some dominant second order error in the setup. For example, the initial states are given in pressure and density and then converted to total energy. This conversion is realized in a second order accurate manner by simply converting cell averages. For the ideal gas EoS, the hydrostatic energy and pressure are related linearly. Since this is not the case for ideal gas with radiation pressure, this conversion error in the initial conditions can appear here. For the higher order methods, the initial energy is computed using a sufficiently high order accurate Gaussian quadrature from the given pressure and density profile.

[9] We use this notation because the $\tilde{\cdot}$ functions also describe the target solution handed to the Deviation method and the α-β method

Table 4.17: L^1-errors and convergence rates for a large perturbation on a polytropic hydrostatic solution of the Euler equations with periodic gravity and the EoS from Section 2.3.1.2 after one quarter of the sound crossing time computed using different methods. The setup is described in Section 4.6.6.

first order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	1.57e-02	–	2.60e-02	–	9.59e-02	–
	128	8.27e-03	0.9	1.36e-02	0.9	4.98e-02	0.9
	256	4.25e-03	1.0	6.99e-03	1.0	2.54e-02	1.0
α-β WB	64	2.51e-04	–	5.66e-07	–	8.52e-04	–
	128	6.32e-05	2.0	3.31e-07	0.8	2.15e-04	2.0
	256	1.59e-05	2.0	1.81e-07	0.9	5.40e-05	2.0
Deviation WB	64	2.51e-04	–	4.70e-07	–	8.52e-04	–
	128	6.32e-05	2.0	2.77e-07	0.8	2.15e-04	2.0
	256	1.59e-05	2.0	1.52e-07	0.9	5.40e-05	2.0

second order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	2.07e-03	–	1.36e-03	–	5.39e-03	–
	128	5.62e-04	1.9	3.55e-04	1.9	1.37e-03	2.0
	256	1.48e-04	1.9	8.99e-05	2.0	3.47e-04	2.0
α-β WB	64	2.50e-04	–	8.54e-08	–	8.51e-04	–
	128	6.31e-05	2.0	2.66e-08	1.7	2.14e-04	2.0
	256	1.58e-05	2.0	7.27e-09	1.9	5.38e-05	2.0
Deviation WB	64	2.50e-04	–	7.83e-08	–	8.51e-04	–
	128	6.31e-05	2.0	2.25e-08	1.8	2.14e-04	2.0
	256	1.58e-05	2.0	6.07e-09	1.9	5.38e-05	2.0

third order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	1.36e-03	–	8.82e-04	–	1.66e-03	–
	128	3.29e-04	2.0	1.43e-04	2.6	2.70e-04	2.6
	256	7.62e-05	2.1	2.19e-05	2.7	4.07e-05	2.7
Deviation WB	64	1.26e-08	–	1.49e-08	–	4.73e-08	–
	128	1.65e-09	2.9	1.96e-09	2.9	6.19e-09	2.9
	256	2.11e-10	3.0	2.52e-10	3.0	7.94e-10	3.0
LA-fast	64	6.44e-05	–	3.48e-05	–	5.99e-05	–
	128	1.26e-05	2.3	3.81e-06	3.2	9.11e-06	2.7
	256	2.16e-06	2.6	3.89e-07	3.3	1.26e-06	2.9

Table 4.18: L^1-errors and convergence rates for a large perturbation on a polytropic hydrostatic solution of the Euler equations with periodic gravity and the EoS from Section 2.3.1.2 after one quarter of the sound crossing time computed using different methods. The setup is described in Section 4.6.6.

fifth order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	32	6.29e-05	–	7.24e-05	–	1.07e-04	–
	64	2.23e-06	4.8	2.71e-06	4.7	3.94e-06	4.8
	128	7.11e-08	5.0	9.03e-08	4.9	1.25e-07	5.0
Deviation WB	32	2.55e-08	–	2.84e-08	–	8.97e-08	–
	64	1.54e-09	4.0	1.79e-09	4.0	5.54e-09	4.0
	128	9.01e-11	4.1	1.07e-10	4.1	3.30e-10	4.1
LA-fast	32	2.72e-06	–	2.81e-06	–	4.82e-06	–
	64	7.86e-08	5.1	9.74e-08	4.8	1.38e-07	5.1
	128	2.28e-09	5.1	3.20e-09	4.9	4.02e-09	5.1

seventh order methods, $\eta = 1e-1$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	32	8.52e-06	–	7.65e-06	–	1.45e-05	–
	64	1.22e-07	6.1	7.84e-08	6.6	2.02e-07	6.2
	128	1.77e-09	6.1	7.70e-10	6.7	2.03e-09	6.6
Deviation WB	32	4.77e-09	–	5.66e-09	–	1.73e-08	–
	64	3.07e-10	4.0	3.59e-10	4.0	1.09e-09	4.0
	128	4.03e-11	2.9	4.82e-11	2.9	1.45e-10	2.9

Table 4.19: L^1-errors and convergence rates for a small perturbation on a polytropic hydrostatic solution of the Euler equations with periodic gravity and the EoS from Section 2.3.1.2 after one quarter of the sound crossing time computed using different methods. The setup is described in Section 4.6.6.

first order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	1.57e-02	–	2.60e-02	–	9.59e-02	–
	128	8.27e-03	0.9	1.36e-02	0.9	4.98e-02	0.9
	256	4.25e-03	1.0	6.99e-03	1.0	2.54e-02	1.0
α-β WB	64	2.50e-04	–	5.66e-11	–	8.51e-04	–
	128	6.31e-05	2.0	3.31e-11	0.8	2.14e-04	2.0
	256	1.58e-05	2.0	1.81e-11	0.9	5.38e-05	2.0
Deviation WB	64	2.50e-04	–	4.70e-11	–	8.51e-04	–
	128	6.31e-05	2.0	2.77e-11	0.8	2.14e-04	2.0
	256	1.58e-05	2.0	1.52e-11	0.9	5.38e-05	2.0

second order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	2.07e-03	–	1.36e-03	–	5.39e-03	–
	128	5.62e-04	1.9	3.55e-04	1.9	1.37e-03	2.0
	256	1.48e-04	1.9	8.99e-05	2.0	3.47e-04	2.0
α-β WB	64	2.50e-04	–	8.54e-12	–	8.51e-04	–
	128	6.31e-05	2.0	2.66e-12	1.7	2.14e-04	2.0
	256	1.58e-05	2.0	7.26e-13	1.9	5.38e-05	2.0
Deviation WB	64	2.50e-04	–	7.83e-12	–	8.51e-04	–
	128	6.31e-05	2.0	2.25e-12	1.8	2.14e-04	2.0
	256	1.58e-05	2.0	6.06e-13	1.9	5.38e-05	2.0

third order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	64	1.36e-03	–	8.82e-04	–	1.66e-03	–
	128	3.29e-04	2.0	1.43e-04	2.6	2.70e-04	2.6
	256	7.62e-05	2.1	2.19e-05	2.7	4.07e-05	2.7
Deviation WB	64	1.26e-12	–	1.49e-12	–	4.73e-12	–
	128	1.64e-13	2.9	1.97e-13	2.9	6.19e-13	2.9
	256	2.14e-14	2.9	2.55e-14	2.9	7.97e-14	3.0
LA-fast	64	6.44e-05	–	3.48e-05	–	5.99e-05	–
	128	1.26e-05	2.3	3.81e-06	3.2	9.11e-06	2.7
	256	2.16e-06	2.6	3.89e-07	3.3	1.26e-06	2.9

Table 4.20: L^1-errors and convergence rates for a small perturbation on a polytropic hydrostatic solution of the Euler equations with periodic gravity and the EoS from Section 2.3.1.2 after one quarter of the sound crossing time computed using different methods. The setup is described in Section 4.6.6.

fifth order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	32	6.29e-05	–	7.24e-05	–	1.07e-04	–
	64	2.23e-06	4.8	2.71e-06	4.7	3.94e-06	4.8
	128	7.11e-08	5.0	9.03e-08	4.9	1.25e-07	5.0
Deviation WB	32	2.55e-12	–	2.84e-12	–	8.97e-12	–
	64	1.54e-13	4.1	1.79e-13	4.0	5.53e-13	4.0
	128	9.23e-15	4.1	1.09e-14	4.0	3.31e-14	4.1
LA-fast	32	2.72e-06	–	2.80e-06	–	4.84e-06	–
	64	7.86e-08	5.1	9.73e-08	4.8	1.38e-07	5.1
	128	2.29e-09	5.1	3.18e-09	4.9	4.10e-09	5.1

seventh order methods, $\eta = 1e-5$

method	N	ρ error	ρ rate	ρu error	ρu rate	E error	E rate
Standard	32	8.52e-06	–	7.65e-06	–	1.45e-05	–
	64	1.22e-07	6.1	7.83e-08	6.6	2.02e-07	6.2
	128	1.77e-09	6.1	7.52e-10	6.7	2.03e-09	6.6
Deviation WB	32	4.77e-13	–	5.66e-13	–	1.73e-12	–
	64	3.02e-14	4.0	3.62e-14	4.0	1.09e-13	4.0
	128	4.47e-15	2.8	5.03e-15	2.8	1.47e-14	2.9

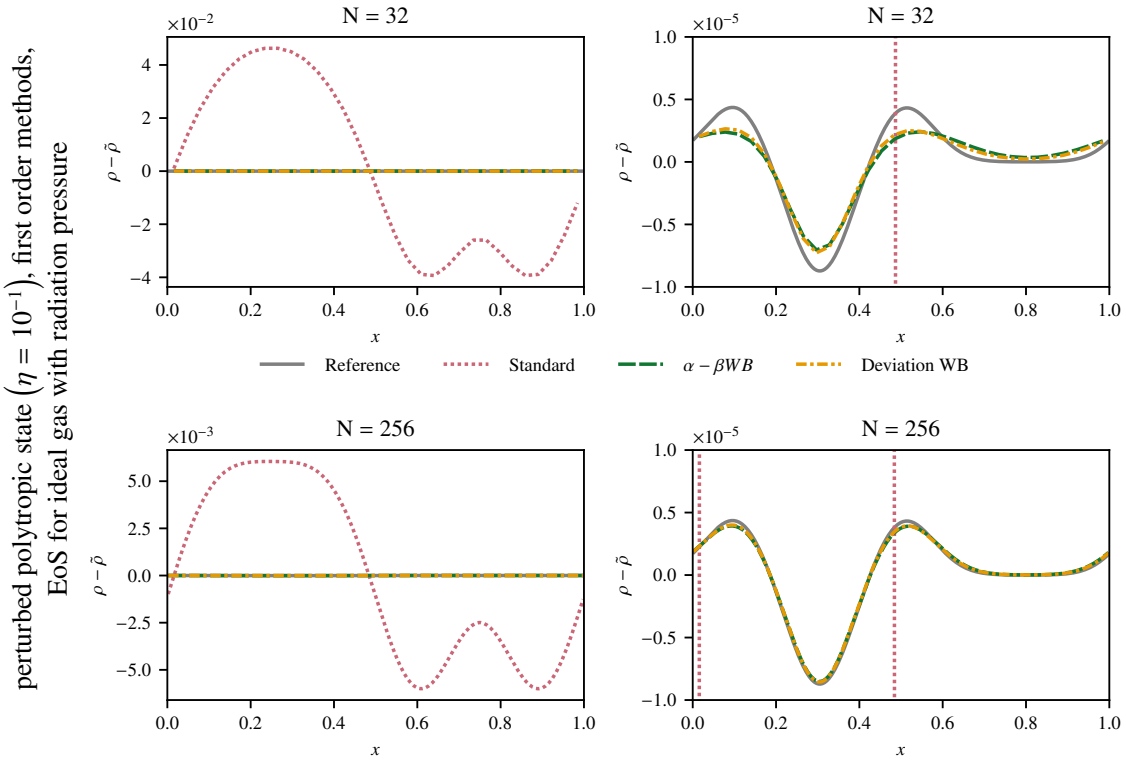


Figure 4.13: Density deviations from the polytropic hydrostatic (EoS for ideal gas with radiation pressure) state for the test with a perturbation on the hydrostatic state described in Section 4.6.6. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, from simulations using the first order accurate methods. On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in to the data from the left panels.

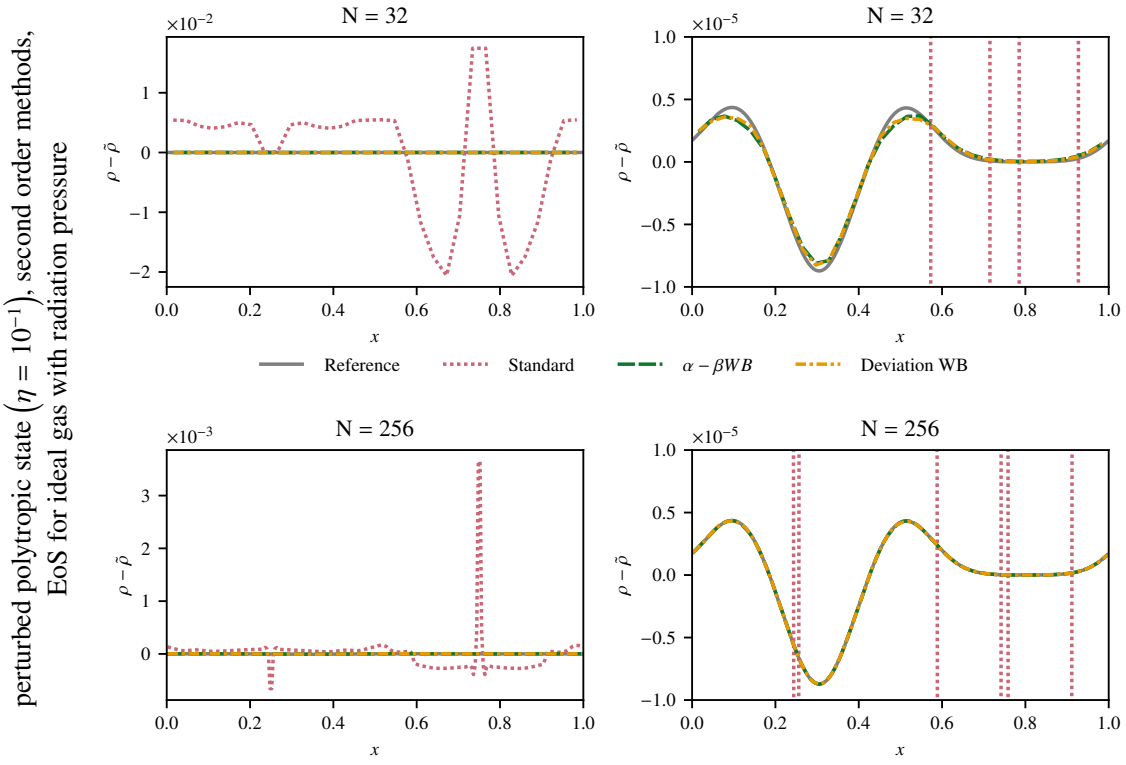


Figure 4.14: Density deviations from the polytropic hydrostatic (EoS for ideal gas with radiation pressure) state for the test with a perturbation on the hydrostatic state described in Section 4.6.6. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, from simulations using the second order accurate methods. On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in to the data from the left panels.

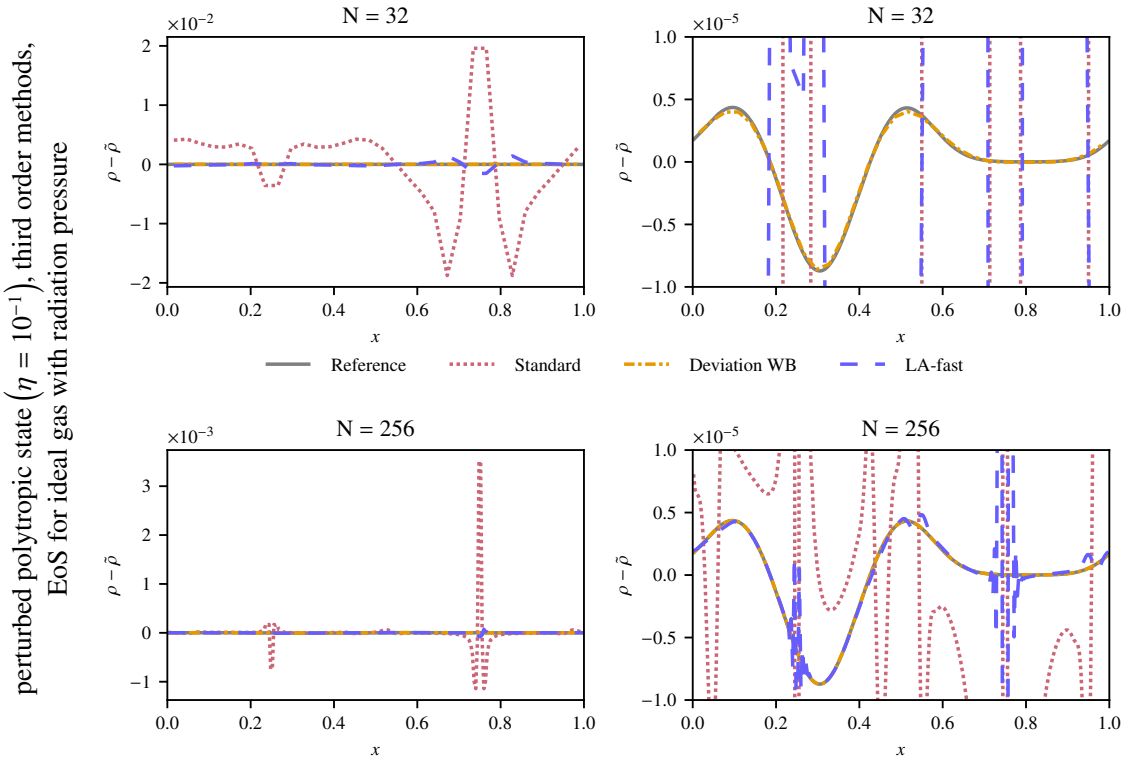


Figure 4.15: Density deviations from the polytropic hydrostatic (EoS for ideal gas with radiation pressure) state for the test with a perturbation on the hydrostatic state described in Section 4.6.6. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, from simulations using the third order accurate methods. On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in to the data from the left panels.

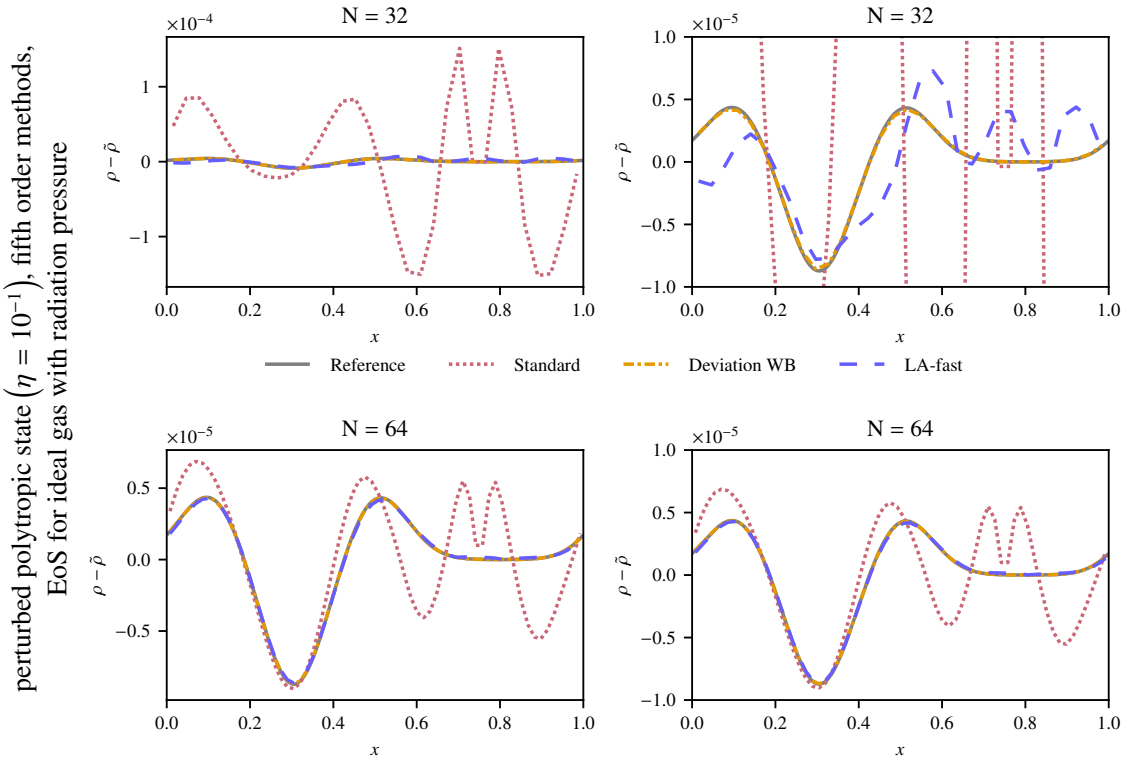


Figure 4.16: Density deviations from the polytropic hydrostatic (EoS for ideal gas with radiation pressure) state for the test with a perturbation on the hydrostatic state described in Section 4.6.6. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, from simulations using the fifth order accurate methods. On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in to the data from the left panels.

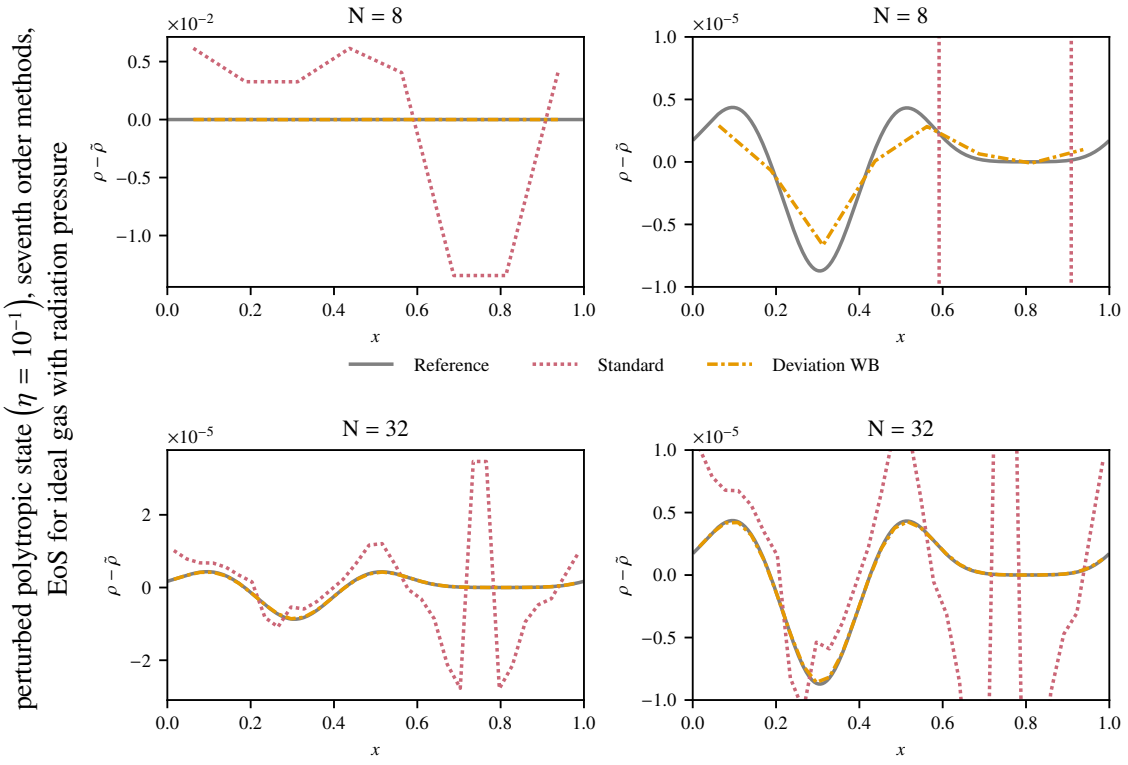


Figure 4.17: Density deviations from the polytropic hydrostatic (EoS for ideal gas with radiation pressure) state for the test with a perturbation on the hydrostatic state described in Section 4.6.6. The snapshot is taken at $t = 0.25\tau$, i.e., after one quarter of the sound crossing time, from simulations using the seventh order accurate methods. On the top panels a coarser grid is used, on the bottom panels a finer grid is used. The right panels show a zoom in to the data from the left panels.

Chapter 5

Finite Volume Methods for Multi-Dimensional Systems of Hyperbolic Balance Laws

In the rest of this thesis, we aim to generalize the numerical methods introduced before for the application on two-dimensional compressible Euler equations with gravity source term. Using these techniques, the methods can in principle also be extended to three-dimensional Euler equations. However, for brevity, and for the reason that we only show two-dimensional simulations in the end, we decided not to present the three-dimensional methods. First, let us introduce and discuss the hyperbolic balance laws we are considering in two spatial dimensions.

5.1 Multi-Dimensional Hyperbolic Balance Laws

For one-dimensional systems of conservation laws, the characteristic structure can be understood by diagonalizing the system. The characteristic variables follow the characteristics. Techniques and approaches to deal with those situations, in which characteristics cross or in which there are no unique characteristics, have been discussed in Chapter 2. For two-dimensional systems

$$\partial_t \boldsymbol{q} + \nabla \cdot \mathcal{F} = 0 \tag{5.1}$$

with the flux tensor $\mathcal{F} = (\boldsymbol{f}_x, \boldsymbol{f}_y)$ the situation is less clear. We cannot base our theory on a diagonal form of the system, since the flux Jacobians

$$A_x(\boldsymbol{q}) := \left.\frac{\partial \boldsymbol{f}_x(\bar{\boldsymbol{q}})}{\partial \bar{\boldsymbol{q}}}\right|_{\bar{\boldsymbol{q}}=\boldsymbol{q}} \qquad \text{and} \qquad A_y(\boldsymbol{q}) := \left.\frac{\partial \boldsymbol{f}_y(\bar{\boldsymbol{q}})}{\partial \bar{\boldsymbol{q}}}\right|_{\bar{\boldsymbol{q}}=\boldsymbol{q}} \tag{5.2}$$

can in general not be diagonalized simultaneously (i.e., the diagonalizing matrices for A_x and A_y are in general different) at a certain state $\boldsymbol{q}$: For the system to be hyperbolic it is only required that the matrix $\boldsymbol{n} \cdot (A_x, A_y)^T$ is diagonalizable with real eigenvalues for any $\boldsymbol{n} \in \mathbb{R}^2 \setminus \{0\}$ (e.g., [103]). To be simultaneously diagonalizable, the matrices A_x and A_y have to commute, i.e., $A_x A_y = A_y A_x$.

Additionally, characteristics have more freedom to move in different directions. Consider the linear acoustics system

$$\partial_t \boldsymbol{q} + A_x \partial_x \boldsymbol{q} + B_y \partial_y \boldsymbol{q} = 0 \tag{5.3}$$

with

$$A_x = \begin{pmatrix} 0 & K_0 & 0 \\ 1/\rho_0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad A_y = \begin{pmatrix} 0 & 0 & K_0 \\ 0 & 0 & 0 \\ 1/\rho_0 & 0 & 0 \end{pmatrix}. \tag{5.4}$$

Independent of the choice of $\boldsymbol{n}$, the matrix $\boldsymbol{n}\cdot(A_x, A_y)^T$ has the eigenvalues $-c_0, 0, +c_0$, where $c_0 = \sqrt{K_0/\rho_0}$ is the speed of sound.[1] Information from one point now travels in every direction with the velocity c_0 (and additionally with velocity 0). Vice-versa, infinitely many characteristics from all directions reach a given point at a given time. For non-linear systems like the two-dimensional compressible Euler equations, the situation is even more complicated.

5.1.1 Two-Dimensional Compressible Euler Equations with Gravitational Source Term

Let $\Omega \subset \mathbb{R}^2$ be the spatial domain of interest. The two-dimensional compressible Euler system with gravity source term is given in the form of a two-dimensional hyperbolic balance law

$$\partial_t \boldsymbol{q} + \nabla \cdot \mathcal{F} = \boldsymbol{s} \tag{5.5}$$

with the state vector

$$\boldsymbol{q} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix}, \tag{5.6}$$

the flux tensor $\mathcal{F} = (\boldsymbol{f}_x, \boldsymbol{f}_y)$, where

$$\boldsymbol{f}_x = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u v \\ (E+p)\,u \end{pmatrix}, \quad \boldsymbol{f}_y = \begin{pmatrix} \rho v \\ \rho u v \\ \rho v^2 + p \\ (E+p)\,v \end{pmatrix}, \tag{5.7}$$

and the gravitational source term

$$\boldsymbol{s} = \begin{pmatrix} 0 \\ \rho g_x \\ \rho g_y \\ \rho \boldsymbol{v} \cdot \boldsymbol{g} \end{pmatrix}. \tag{5.8}$$

The vector-valued gravitational acceleration $\boldsymbol{g} = (g_x, g_y)^T$ is the negative gradient of the gravitational potential $\phi \in \mathcal{C}^1(\Omega, \mathbb{R})$, i.e., $\boldsymbol{g} = -\nabla\phi$. For this thesis we assume

[1] Note that this does not imply that the matrices are simultaneously diagonalizable. In fact, they are not.

ϕ and hence also $\boldsymbol{g}$ to be constant in time, just as in the one-dimensional case. The volumetric total energy is $E = \varepsilon + E_{\text{kin}}$ with $E_{\text{kin}} = \frac{1}{2}\rho|\boldsymbol{v}|^2$, where $\boldsymbol{v} = (u, v)^T$ is the fluid velocity. The volumetric internal energy ε is related to density ρ and pressure p via an EoS as discussed in Section 2.3.1.

The transformations between primitive variables $\boldsymbol{q}^{\text{prim}} = (\rho, u, v, p)^T$ and conserved variables $\boldsymbol{q}^{\text{cons}} = \boldsymbol{q}$ are

$$\frac{\partial \boldsymbol{q}^{\text{cons}}}{\partial \boldsymbol{q}^{\text{prim}}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ u & \rho & 0 & 0 \\ v & 0 & \rho & 0 \\ \frac{1}{2}(u^2+v^2) - \frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \rho}\left(\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon}\right)^{-1} & \rho u & \rho v & \left(\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon}\right)^{-1} \end{pmatrix} \tag{5.9}$$

and

$$\begin{aligned} \frac{\partial \boldsymbol{q}^{\text{prim}}}{\partial \boldsymbol{q}^{\text{cons}}} &= \left(\frac{\partial \boldsymbol{q}^{\text{cons}}}{\partial \boldsymbol{q}^{\text{prim}}}\right)^{-1} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{u}{\rho} & \frac{1}{\rho} & 0 & 0 \\ -\frac{v}{\rho} & 0 & \frac{1}{\rho} & 0 \\ \frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \rho} + \frac{1}{2}\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon}(u^2+v^2) & -\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon}u & -\frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon}v & \frac{\partial p_{\text{EoS}}(\rho,\epsilon)}{\partial \epsilon} \end{pmatrix}. \end{aligned} \tag{5.10}$$

In primitive variables, the flux Jacobians take the form

$$A_x^{\text{prim}}(\boldsymbol{q}) = \begin{pmatrix} u & \rho & 0 & 0 \\ 0 & u & 0 & \frac{1}{\rho} \\ 0 & 0 & u & 0 \\ 0 & \rho c^2 & 0 & u \end{pmatrix}, \qquad A_y^{\text{prim}}(\boldsymbol{q}) = \begin{pmatrix} v & 0 & \rho & 0 \\ 0 & v & 0 & 0 \\ 0 & 0 & v & \frac{1}{\rho} \\ 0 & 0 & \rho c^2 & v \end{pmatrix} \tag{5.11}$$

with the matrices of right eigenvectors

$$R_x^{\text{prim}}(\boldsymbol{q}) = \begin{pmatrix} 0 & 1 & \frac{1}{c^2} & \frac{1}{c^2} \\ 0 & 0 & \frac{1}{c\rho} & -\frac{1}{c\rho} \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \qquad R_y^{\text{prim}}(\boldsymbol{q}) = \begin{pmatrix} 0 & 1 & \frac{1}{c^2} & \frac{1}{c^2} \\ 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{c\rho} & -\frac{1}{c\rho} \\ 0 & 0 & 1 & 1 \end{pmatrix} \tag{5.12}$$

and the diagonal matrices

$$\Lambda_x(\boldsymbol{q}) = \text{diag}(u, u, u+c, u-c), \qquad \Lambda_y(\boldsymbol{q}) = \text{diag}(v, v, v+c, v-c) \tag{5.13}$$

of eigenvalues, where c is the speed of sound defined in Eq. (2.28). The flux Jacobians and matrices of right eigenvectors in conserved variables can be obtained via transformation as discussed in Section 2.3.2.

5.1.2 Hydrostatic Solutions

The multi-dimensional hydrostatic equation takes the form

$$\nabla p = \rho \boldsymbol{g}. \tag{5.14}$$

As in the one-dimensional case (Eq. (2.35)), this partial differential equation is obtained by setting all time-derivatives and the velocity to zero and hence describes (hydro)static solutions. The hydrostatic solutions have to additionally satisfy the EoS (see Section 2.3.1) used to relate the thermodynamical quantities. In principle, Eq. (5.14) admits genuinely multi-dimensional stratifications of density and pressure. However, the physical world tends to produce symmetric solutions: The gravitational potential ϕ has, in many cases, spherical symmetry, since it is given as the solution of the Poisson-equation $\nabla \cdot (\nabla\phi) = 4\pi G\rho$ (G is the universal gravitational constant) and the density distribution of gravitating objects such as planets or stars is often close to spherical symmetry. In other applications the gravitational acceleration $\boldsymbol{g}$ is chosen as a constant, since only a part of an atmosphere at a planet's surface or a part of the stellar interior is simulated, in which the gravitational acceleration does not significantly depend on the coordinates. In both of the described cases, the problem of solving the hydrostatic equation reduces to a one-dimensional problem, e.g., in radial direction or the direction of the (constant) gravitational acceleration. As examples for two-dimensional hydrostatic states, we hence do not introduce new types of hydrostatic solutions. Instead, we give formulae to extend the one-dimensional hydrostatic states introduced in Section 2.4.1 to two spatial dimensions according to the description above.

Radial symmetry Assume the gravitational potential to be spherically symmetric, i.e., $\phi(\boldsymbol{x}) = \phi^{1\mathrm{d}}(\|\boldsymbol{x}\|_2)$. The same holds then for the corresponding hydrostatic states:

$$\rho(\boldsymbol{x}) = \rho^{1\mathrm{d}}(\|\boldsymbol{x}\|_2), \qquad p(\boldsymbol{x}) = p^{1\mathrm{d}}(\|\boldsymbol{x}\|_2), \tag{5.15}$$

where $\rho^{1\mathrm{d}}$ and $p^{1\mathrm{d}}$ describe a one-dimensional hydrostatic state belonging to the gravitational potential $\phi^{1\mathrm{d}}$, such as the ones introduced in Section 2.4.1.

Constant gravitational acceleration Assume a constant gravitational acceleration $\boldsymbol{g} = (g_x, g_y)^T \in \mathbb{R}^2$. In this case, the gravitational potential is $\phi(\boldsymbol{x}) = C - \boldsymbol{x} \cdot \boldsymbol{g}$ for an arbitrary constant $C \in \mathbb{R}$ and the hydrostatic states are

$$\rho(\boldsymbol{x}) = \rho^{1\mathrm{d}}\left(C - \frac{\boldsymbol{x} \cdot \boldsymbol{g}}{\|\boldsymbol{g}\|_2}\right), \qquad p(\boldsymbol{x}) = p^{1\mathrm{d}}\left(C - \frac{\boldsymbol{x} \cdot \boldsymbol{g}}{\|\boldsymbol{g}\|_2}\right), \tag{5.16}$$

where $\rho^{1\mathrm{d}}$ and $p^{1\mathrm{d}}$ describe a one-dimensional hydrostatic state such as the ones introduced in Section 2.4.1 belonging to the one-dimensional gravitational potential given by $\phi(x) = C - |\boldsymbol{g}|x$.

5.2 About Two-Dimensional Runge–Kutta Finite Volume Methods

Let us divide the two-dimensional domain $\Omega \subset \mathbb{R}^2$ into a finite number of cells.[2] Reconsider the Godunov method introduced in Section 3.1. The basic idea was

[2]Spatial discretization of a domain will be discussed below in Section 5.3.

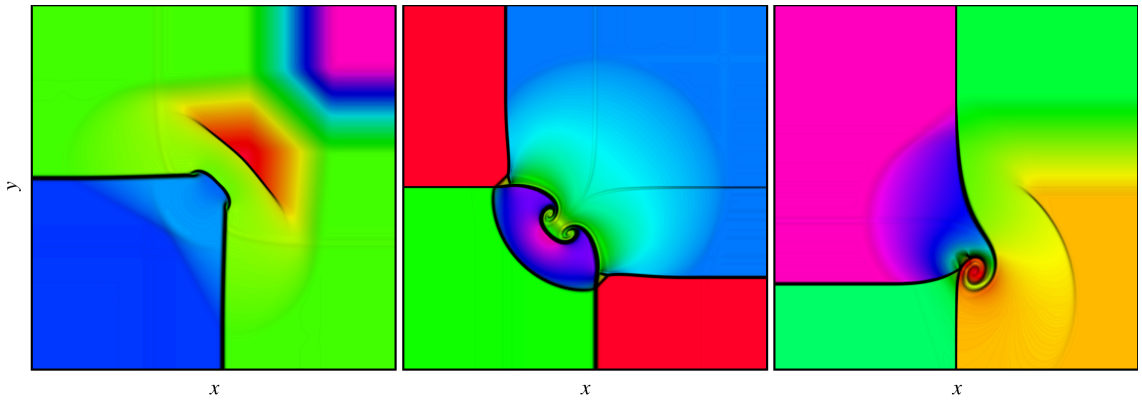


Figure 5.1: Density of the solution of three different four-state Riemann problems for the two-dimensional homogeneous compressible Euler system. Close to the boundaries the solution of the two-state Riemann problems is given by the solution of the one-dimensional Riemann problem. In the central region, however, complex structures appear. The solutions have been obtained applying an RK-FV method based on approximate Riemann solvers for the one-dimensional Riemann problem. These figures provide a qualitative visualization, for quantitative information and details on the test setups see Appendix A.3.

that the given initial data are simplified in each time-step such that they are piecewise constant and can, for a short time, be evolved exactly. This exact evolution was possible, since the two-state Riemann-problems appearing at the cell interfaces could be solved analytically using some additional assumptions such as an entropy condition and self-similarity. In one spatial dimension, it is clear that the solution of a Riemann problem consists of a certain number of states connected by shocks or rarefaction waves. In two spatial dimensions, it is unfortunately not that simple. At the points, at which interfaces meet, Riemann-problems involving more than two states have to be solved. These solutions can have a rich spatial structure, as is visualized in Fig. 5.1. It is thus, compared to the one-dimensional case, much more challenging to find approximate Riemann solvers approximating solutions of two-dimensional Riemann problems.

The RK-FV approach we follow in this thesis avoids solving multi-dimensional Riemann problems, since the applied numerical fluxes are only evaluated at two states (on both sides of an interface) and thus only approximate solutions of the one-dimensional Riemann problem are used. As we can see in Fig. 5.1, it is yet possible to resolve two-dimensional structures. However, it is noteworthy that the numerical methods which are based on numerical two-state fluxes often violate multi-dimensional involutions (e.g., the $\nabla \cdot \boldsymbol{B} = 0$ constraint in compressible ideal magnetohydrodynamics (MHD) equations [60]) or multi-dimensional conditions appearing in some asymptotic limit (e.g., the low Mach limit of compressible Euler equations [79]). In these cases there are solution strategies either adding some multi-dimensional components to the numerical scheme (e.g., constrained transport for MHD [60, 165, 144]) or applying corrections to the numerical fluxes (e.g., low Mach fixes for Euler equations [168, 140, 159, 128, 113, 16]). To accurately capture multi-dimensional structures in the first place, there are also finite volume-type

methods which do not rely on the simplification discussed above like the multi-dimensional active flux method (e.g., [61, 9, 7]) and there are also numerical fluxes using information from a multi-dimensional stencil (e.g., [80, 33, 134]).

However, one important property of our well-balanced methods is that they can easily be added to most existing finite volume codes without restructuring. Since the majority of finite volume codes relies on two-state Riemann solvers, we utilize this structure to extend our well-balanced methods to two spatial dimensions. In the rest of this chapter we only focus on this type of two-dimensional finite volume methods.

5.3 Grids

The basic discretization technique in the context of finite volume methods, is the discretization of space or space-time into a grid. In the context of RK-FV methods, a grid only has to discretize space, hence we focus on spatial discretization in this section. Examples for fully-discrete methods using space-time grids are [44, 149, 68]. We adapt the definition for a grid, which is given in Section 3.1 of [135].

Definition 5.3.1 (Grid)**.** *Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain, i.e., Ω is an open bounded connected subset of $\mathbb{R}^2$ and $\bar{\Omega}$ is the union of a finite number of polygons. As grid $\mathfrak{G}$ for Ω we define a finite decomposition of Ω with the following properties:*

1. *Each $K \in \mathfrak{G}$ is a polygon with $\mathring{K} \neq \emptyset$,*
2. *$\bigcup_{K\in\mathfrak{G}} K = \bar{\Omega}$,*
3. *$\mathring{K}_1 \cap \mathring{K}_2 = \emptyset$ for each pair of distinct $K_1, K_2 \in \mathfrak{G}$.*

In the following, we often denote the grid with $\mathfrak{G} = \{\Omega_i\}_{i\in\mathcal{I}}$, where $\mathcal{I}$ is a suitable set of indices, e.g., $\mathcal{I} = \{1, ..., N\}$ for the number of cells $N \in \mathbb{N}$. An important relation between the cells of a grid is the neighborhood relation:

Definition 5.3.2 (Neighboring cells)**.** *Let $\Omega_i, \Omega_j \in \mathfrak{G}$ $(i, j \in \mathcal{I})$ be cells in the grid $\mathfrak{G}$ for the domain $\Omega \subset \mathbb{R}^2$. We define the common face $\partial\Omega_{ij} = \Omega_i \cap \Omega_j$. If $\delta\Omega_{ij}$ is a manifold of co-dimension 1, we call Ω_j a* neighbor *of Ω_i. The set of indices of neighbors of the cell Ω_i is denoted by*

$$\mathcal{N}_i := \{j \in \mathcal{I} \,:\, \dim(\partial\Omega_{ij}) = d - 1\}\,. \tag{5.17}$$

The neighborhood relation is symmetric, i.e., $i \in \mathcal{N}_j$ is equivalent to $j \in \mathcal{N}_i$ for $i, j \in \mathcal{I}$.

5.3.1 Curvilinear Grids

The presumably simplest grid is the uniform Cartesian grid.

Definition 5.3.3 (Cartesian grid and uniform Cartesian grid)**.** *Consider the rectangle $\Omega := [\xi_{\min}, \xi_{\max}] \times [\eta_{\min}, \eta_{\max}] \subset \mathbb{R}^2$ with $\xi_{\min}, \xi_{\max}, \eta_{\min}, \eta_{\max} \in \mathbb{R}$ and $\xi_{\min} <$*

$\xi_{\max}, \eta_{\min} < \eta_{\max}$. *A grid* $\mathfrak{G} := \{\Omega_{ij}\}_{(i,j)\in\mathcal{I}}$ *(*$\mathcal{I} = \{1,\dots,N_\xi\} \times \{1,\dots,N_\eta\}$ *with* $N_\xi, N_\eta \in \mathbb{N}$*) defined by the cells*

$$\Omega_{ij} := \left[\xi_{i-\frac{1}{2}}, \xi_{i+\frac{1}{2}}\right] \times \left[\eta_{j-\frac{1}{2}}, \eta_{j+\frac{1}{2}}\right] \tag{5.18}$$

with $\xi_{i-\frac{1}{2}} < \xi_{i+\frac{1}{2}}, \eta_{j-\frac{1}{2}} < \eta_{j+\frac{1}{2}}$ *for all* $(i,j) \in \mathcal{I}$ *and* $\xi_{-\frac{1}{2}} = \xi_{\min}, \xi_{N_\xi+\frac{1}{2}} = \xi_{\max}, \eta_{-\frac{1}{2}} = \eta_{\min}, \eta_{N_\eta+\frac{1}{2}} = \eta_{\max}$ *is called* Cartesian grid *of* Ω*. If additionally*

$$\xi_{i+\frac{1}{2}} := \xi_{\min} + i\Delta\xi, \qquad \Delta\xi := \frac{\xi_{\max} - \xi_{\min}}{N_\xi}, \qquad \text{for } i \in \{0,\dots,N_\xi\}, \tag{5.19}$$

$$\eta_{j+\frac{1}{2}} := \eta_{\min} + j\Delta\eta, \qquad \Delta\eta := \frac{\eta_{\max} - \eta_{\min}}{N_\eta}, \qquad \text{for } j \in \{0,\dots,N_\eta\}, \tag{5.20}$$

$\mathfrak{G}$ *is called* uniform Cartesian grid *of* Ω.

Note that we use a different indexing for grid cells compared to the the general grids introduced above. This is useful in applications, since the indexing with two indices mirrors the representation of the cells in a two-dimensional computational array. This structure can be transferred to more general body-fitted grids:

Definition 5.3.4 (Curvilinear grid)**.** *Let* $\Omega^{\text{comp}} \subset \mathbb{R}^2$ *be a rectangle,* $\Omega \subset \mathbb{R}^2$ *an open bounded connected domain and* $\boldsymbol{x} : \Omega^{\text{comp}} \to \Omega, \boldsymbol{\xi} \mapsto \boldsymbol{x}(\boldsymbol{\xi})$ *a* C^1*-diffeomorphism. Let* $\mathfrak{G}_{\text{comp}}$ *be a uniform Cartesian grid for* Ω^{comp} *with the notations from Definition 5.3.3. Furthermore, let* $\mathfrak{G} := \{\Omega_{ij}\}_{(i,j)\in\mathcal{I}}$ *be a grid for* Ω *such that the cells* Ω_{ij} *are convex quadrilaterals defined by the four corners*

$$\boldsymbol{x}_{i\pm\frac{1}{2},j\pm\frac{1}{2}} := \boldsymbol{x}\left(\boldsymbol{\xi}_{i\pm\frac{1}{2},j\pm\frac{1}{2}}\right) \qquad \text{for} \qquad (i,j) \in \mathcal{I} \tag{5.21}$$

with

$$\boldsymbol{\xi}_{i+\frac{1}{2},j+\frac{1}{2}} := \left(\xi_{i+\frac{1}{2}}, \eta_{j+\frac{1}{2}}\right)^T \quad \text{for} \quad (i,j) \in \{0,\dots,N_\xi\} \times \{0,\dots,N_\eta\}. \tag{5.22}$$

Then we say $\mathfrak{G}$ *is a* curvilinear *grid. The domain* Ω^{comp} *is called* computational domain*, whereas* Ω *is called* physical domain*.*

We refer to the coordinates $\boldsymbol{\xi}$ as *computational coordinates* and to $\boldsymbol{x}$ as *physical coordinates*. Definition 5.3.4 defines the grid in a vertex-based approach. The cell-centers can then be constructed from the vertices (i.e., the corners). The simplest approach for that is using the arithmetic average. An alternative approach is to transform the cell-centers using the diffeomorphism $\boldsymbol{x}$ and construct the vertices from the cell-centers. Note, that for high order methods, a high order accurate mapping of all quadrature points at the interfaces and in the domain is necessary (e.g., [45, 174]) and that the mapping has to be sufficiently smooth.

The main advantage of curvilinear grids is that body-fitted coordinates, suitable for the problem which shall be simulated, can be used while keeping the array-like structure of a Cartesian grid. Therefore, curvilinear grids are also called *structured grids* opposing the phrase *unstructured grids*, which is used for all grids that are not structured. In structured grids, it is easy to identify the neighboring cells.

Remark 5.3.5. *In a two-dimensional structured grid, the set of indices of neighboring cells for the cell Ω_{ij} is*

$$\mathcal{N}_{ij} = \{(i', j') \in \mathcal{I} : |i - i'| + |j - j'| = 1\}. \tag{5.23}$$

In the description of methods below, some methods are described for general grids. In that case, the indexing of cells from unstructured grids is used. This still includes structured grids, since the cells on a structured grid can also be numbered using only one index.

5.3.1.1 Transformation to a Curvilinear Grid

In the description and implementation of finite volume methods on curvilinear grids, there are some relevant relations between the two coordinate systems $\boldsymbol{x}$ and $\boldsymbol{\xi}$ and we derive them in this section. The differentials are related by

$$\mathrm{d}\boldsymbol{x} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\xi}} \mathrm{d}\boldsymbol{\xi}, \qquad \mathrm{d}\boldsymbol{\xi} = \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{x}} \mathrm{d}\boldsymbol{x} \tag{5.24}$$

with the local Jacobian matrices

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\xi}} := \begin{pmatrix} \partial_\xi x & \partial_\eta x \\ \partial_\xi y & \partial_\eta y \end{pmatrix} \quad \text{and} \quad \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{x}} := \begin{pmatrix} \partial_x \xi & \partial_y \xi \\ \partial_x \eta & \partial_y \eta \end{pmatrix}. \tag{5.25}$$

Hence, using Cramer's rule, the relation

$$\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{x}} = \left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\xi}} \right)^{-1} = \frac{1}{J} \begin{pmatrix} \partial_\eta y & -\partial_\eta x \\ -\partial_\xi y & \partial_\xi x \end{pmatrix} \tag{5.26}$$

follows with the Jacobian determinant

$$J = \partial_\xi x \, \partial_\eta y - \partial_\eta x \, \partial_\xi y. \tag{5.27}$$

This yields the following identities

$$J \partial_x \xi = \partial_\eta y, \qquad J \partial_y \xi = -\partial_\eta x, \qquad J \partial_x \eta = -\partial_\xi y, \qquad J \partial_y \eta = \partial_\xi x. \tag{5.28}$$

Notation 5.3.6. *Obviously, the Jacobian determinant of the transformation has a spatial dependency which can be expressed as dependency of $\boldsymbol{\xi}$ or $\boldsymbol{x}$. For convenience, we sometimes switch between the notations $J(\boldsymbol{\xi})$ and $J(\boldsymbol{x})$, where $J(\boldsymbol{\xi}) := J|_{\boldsymbol{\xi}}$ and $J(\boldsymbol{x}) := J|_{\boldsymbol{\xi}(\boldsymbol{x})}$, provided that it is clear from the notation, which coordinates we use in the argument.*

Let $\psi : \Omega \to \mathbb{R}$ be a sufficiently smooth function. The cell-average in the ij-th cell is then

$$\frac{1}{V_{ij}} \int_{\Omega_{ij}} \psi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \approx \frac{1}{V_{ij}} \int_{\Omega_{ij}^{\mathrm{comp}}} J(\boldsymbol{\xi}) \psi(\boldsymbol{x}(\boldsymbol{\xi})) \, \mathrm{d}\boldsymbol{\xi}. \tag{5.29}$$

Note that this is only an approximate relation. If the grid is constructed such that the mapping $\boldsymbol{\xi} \mapsto \boldsymbol{x}(\boldsymbol{\xi})$ correctly maps $\boldsymbol{x}\left(\Omega_{ij}^{\mathrm{comp}}\right) = \Omega_{ij}$, the relation is exact. However, in many cases (and also in Definition 5.3.4), only vertices or cell-centers are mapped and the grid is then constructed. Thus, Eq. (5.29) is not exact in general.

5.3.2 Examples of Curvilinear Grids

In this section we present four different two-dimensional curvilinear grids that we use in numerical tests in this thesis. The grids are visualized in Fig. 5.2. A Cartesian grid (see Definition 5.3.3) is a trivial curvilinear grid, e.g., with $\boldsymbol{x}(\boldsymbol{\xi}) = \boldsymbol{\xi}$. A Cartesian grid with $N_x = N_y = 20$ is shown in Fig. 5.2a. A polar grid (shown in Fig. 5.2b) is presumably amongst the most common non-trivial curvilinear grids. It is discretely spherically symmetric. The center has to be avoided, since the mapping is singular for zero radius. The volume of the cells increases from smaller to higher radii. An attempt to construct a grid which is suitable for simulation setups with radial symmetry while still including the center can be found in the cubed sphere grid suggested by [31] and shown in Fig. 5.2c.

5.4 Numerical Fluxes

In two spatial dimensions, the flux is given in the form of a tensor $\mathcal{F} = (\boldsymbol{f}_1, \boldsymbol{f}_2) = (\boldsymbol{f}_x, \boldsymbol{f}_y)$. The flux over an interface between two cells is then $\mathcal{F} \cdot \boldsymbol{n} = n_x \boldsymbol{f}_x + n_y \boldsymbol{f}_y$, where $\boldsymbol{n}$ is the normalized normal vector of the interface in the considered direction. Using the same structure, we can generalize Definition 3.3.1 to two spatial dimensions. A numerical two-state flux function that is consistent with the directional flux $\boldsymbol{n} \cdot \mathcal{F}$ is denoted with $\boldsymbol{F}(\cdot, \cdot, \boldsymbol{n})$ in the following. We now present a simple way to practically extend the numerical flux functions introduced in Section 3.3 to two spatial dimensions in a numerical code. This approach requires the hyperbolic system – ignoring possible source terms – to be spatially isotropic.

Let Θ be a rotation matrix that rotates $\boldsymbol{n}$ to $\boldsymbol{e}_1 = (1, 0)^T$. Construct a bock-diagonal matrix $\tilde{\Theta}$ such that Θ is applied on each vector-valued variable in the state vectors and identity is applied to scalar variables. For the example of Euler equations ($\boldsymbol{q} = (\rho, \rho\boldsymbol{v}, E)$) this is

$$\tilde{\Theta} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \Theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.30}$$

with the inverse matrix

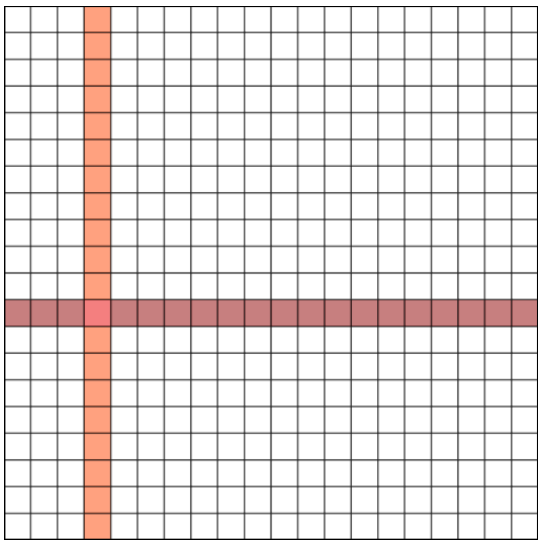
$$\tilde{\Theta}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \Theta^{-1} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.31}$$

The numerical flux in direction $\boldsymbol{n}$ is then defined by

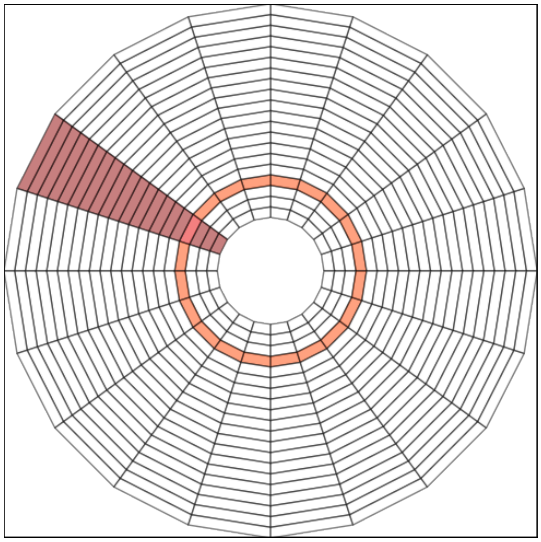
$$\boldsymbol{F}(\boldsymbol{q}^L, \boldsymbol{q}^R, \boldsymbol{n}) := \tilde{\Theta}^{-1} \boldsymbol{F}_x(\tilde{\Theta}\boldsymbol{q}^L, \tilde{\Theta}\boldsymbol{q}^R). \tag{5.32}$$

This approach is equivalent to directly approximating solutions of the one-dimensional Riemann problem over the interface in $\boldsymbol{n}$-direction.[3] Since the compressible Euler system is spatially isotropic, we apply this approach to evaluate numerical fluxes in the numerical experiments in Section 6.4.

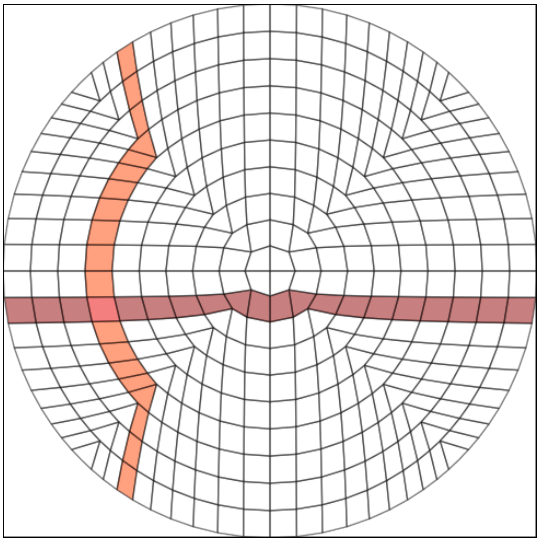
[3]Note, that it is not a priori clear, that the two-dimensional extension of the solution of a one-dimensional Riemann problem is the only admissible solution of the two-dimensional extension of the Riemann problem. In fact, this is not the case [1]. However, it is in line with the current state-of-art of numerical methods to choose this particular solution.



(a) Cartesian grid



(b) Polar grid



(c) Cubed sphere grid

Figure 5.2: Some two-dimensional curvilinear grids which are used in this thesis. Each of the grids has 20×20 grid cells. The same row and line in the corresponding computational grid is colorized for every shown grid. These figures are taken from the author's master thesis [11].

5.5 Multi-Dimensional Quadrature

In one spatial dimension, the numerical flux at the interface approximates the exact flux at a single point. In two spatial dimensions, on the other hand, the interface flux has to be integrated over the interface between two neighboring cells. Since the numerical fluxes are in general highly non-linear, the exact computation of these integrals is tedious if possible at all. However, since the flux is approximated any way, it is sufficient to use quadrature rules to approximate the interface flux integral. At some points we also require quadrature over the cells to approximate cell-average values (e.g., for setting initial values and computing source terms). In the end this means, that quadrature rules for one and two spatial dimensions are required. Two-dimensional quadrature (e.g., [135]) is in general a significantly more delicate problem than one-dimensional quadrature (Section 3.5). In the numerical tests in this thesis, however, two-dimensional quadrature rules are only applied on Cartesian grids such that all integration domains are lines and rectangles. Hence, for us it is sufficient to use the following straightforward extension of the one-dimensional quadrature rules.

Theorem 5.5.1 (Multi-dimensional extension of Gaussian quadrature)**.** *Let $\Omega = [a,b] \times [c,d] \subset \mathbb{R}^2$ with $a < b, c < d$ and $a, b, c, d \in \mathbb{R}$ be a rectangle with $(b-a) < h, (d-c) < h$ for some $h > 0$. Let $g \in \mathcal{C}^{2n}(\Omega, \mathbb{R})$ be a scalar valued function on Ω for some $n \in \mathbb{N}$. Furthermore, let ξ_i for $i = 1, \dots, n$ be the normalized quadrature points and ω_i the weights of a one-dimensional Gaussian quadrature rule as defined in Definition 3.5.1. Then it is*

$$\frac{1}{|\Omega|} \int_\Omega g(x,y)\, d\boldsymbol{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i \omega_j g\left(x_i, y_j\right) + \mathcal{O}\left(h^{2n}\right) \tag{5.33}$$

with

$$x_i = \frac{a+b}{2} + \frac{b-a}{2}\xi_i, \qquad y_i = \frac{c+d}{2} + \frac{d-c}{2}\xi_i \tag{5.34}$$

for $i = 1, \dots, n$.

Proof. Based on Fubini's theorem, the one-dimensional quadrature rule can be applied in the different directions sequentially. Because of Theorem 3.5.2 this yields Eq. (5.33). □

5.6 Reconstruction

We have seen that on a Cartesian grid quadrature rules can be extended to two spatial dimensions by applying them in the different directions sequentially. In the case of reconstruction, however, a truly multi-dimensional approach is in general required, at least for high order methods (higher than second order). Multi-dimensional reconstruction (especially limiting) is an active field of research (e.g., [45, 53, 86, 110, 177]). In this thesis, we only discuss the methods that are actually applied in the numerical tests: Linear reconstruction on curvilinear grids and parabolic reconstruction on two-dimensional uniform Cartesian grids.

5.6.1 Linear Reconstruction

In a second order accurate method, the mid-point quadrature rule is sufficient to approximate the interface flux integral. The numerical flux has therefore only to be evaluated at one point per interface in a second order accurate method. This leads to a decoupling in the reconstruction. The values at the $(i + \frac{1}{2}, j)$-interface, for example, are obtained from reconstructing the cell-average values in ξ-direction in the (i, j)-th and $(i + 1, j)$-th cell. Since this is in the core a one-dimensional procedure, this approach is called *spatial splitting*.

5.6.2 Parabolic Reconstruction

To construct a third order accurate finite volume method, we use genuinely multi-dimensional reconstruction. There are two reasons for that:

1. To approximate the flux integral to third order accuracy, it is not sufficient to only use the cell-centered quadrature point. More quadrature points are required. Hence, the multi-dimensional structure of the reconstruction is relevant in the approximation of the flux integral.

2. To obtain a third order accurate reconstruction, at least a parabola is required. Because of the mixed terms in multi-dimensional parabola, a two-dimensional parabola can not be uniquely determined by combining the information from the one-dimensional parabolas in both coordinate directions. Consequently, a genuinely multi-dimensional stencil is necessary to reconstruct a multi-dimensional parabola.

In literature, there are also high order reconstruction techniques for finite volume methods that are based on spatial splitting (e.g., [26]). However, since these methods require multiple reconstruction steps and are less commonly used than genuinely multi-dimensional reconstruction, we follow the latter approach in this thesis. In the following we give an example how an (unlimited) parabolic reconstruction on a uniform Cartesian grid can be obtained. For polynomial reconstruction on an unstructured grid we refer to [10]. First, let us adapt the notation from [106] for readability. This notation is illustrated in Fig. 5.3. In the following we mix this compass direction notation from [106] with the structured grid indexing notation in order to find a good compromise between readability and compactness of the presentation.

Let us reconstruct from the cell-averages of a scalar function q. The coefficients a_0, a_x, a_y, a_{xx}, a_{xy}, and a_{yy} of the reconstruction parabola

$$P(\bar{\boldsymbol{x}}) = a_0 + a_x\bar{x} + a_y\bar{y} + a_{xx}\bar{x}^2 + a_{xy}\bar{x}\bar{y} + a_{yy}\bar{y}^2, \tag{5.35}$$

where the coordinates

$$\bar{\boldsymbol{x}} := \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} := \begin{pmatrix} \frac{x - x_{\mathrm{C}}}{\Delta x} \\ \frac{y - y_{\mathrm{C}}}{\Delta y} \end{pmatrix} \tag{5.36}$$

$j+1$	NW	N	NE
j	W	C	E
$j-1$	SW	S	SE
	$i-1$	i	$i+1$

Figure 5.3: Notation we use to describe the two-dimensional parabolic reconstruction on a uniform Cartesian grid. Instead of the usual indices $(i,j) \in \mathcal{I}$ we use compass directions.

are used for simplicity and brevity, are obtained using a least squares optimization restricted by conservation, i.e.,

$$\min_{a_0,a_x,a_y,a_{xx},a_{xy},a_{yy}\in\mathbb{R}} \left(\sum_{(k,l)\in\mathcal{S}_{ij}\setminus(i,j)} \left(\int_{\Omega_{kl}} P(\bar{\boldsymbol{x}})\, d\bar{\boldsymbol{x}} - \hat{q}_{kl} \right)^2 \right) \tag{5.37}$$

$$\text{with the constraint } \int_{\Omega_{ij}} P(\bar{\boldsymbol{x}})\, d\bar{\boldsymbol{x}} = \hat{q}_{ij}, \tag{5.38}$$

where the stencil is $\mathcal{S}_{ij} = \{i-1, i, i+1\} \times \{j-1, j, j+1\}$. Note that this approach is not the only one leading to a correct reconstruction parabola, since there are nine cell-averages given in the stencil to determine only six coefficients. Solving[4] Eqs. (5.37) and (5.38) for the coefficients yields

$$a_0 = (132\hat{q}_{\mathrm{C}} - \hat{q}_{\mathrm{E}} - \hat{q}_{\mathrm{N}} - 2\hat{q}_{\mathrm{NE}} - 2\hat{q}_{\mathrm{NW}} - \hat{q}_{\mathrm{S}} - 2\hat{q}_{\mathrm{SE}} - 2\hat{q}_{\mathrm{SW}} - \hat{q}_{\mathrm{W}})/120, \tag{5.39}$$

$$a_x = (\hat{q}_{\mathrm{E}} + \hat{q}_{\mathrm{NE}} - \hat{q}_{\mathrm{NW}} + \hat{q}_{\mathrm{SE}} - \hat{q}_{\mathrm{SW}} - \hat{q}_{\mathrm{W}})/6, \tag{5.40}$$

$$a_y = (\hat{q}_{\mathrm{N}} + \hat{q}_{\mathrm{NE}} + \hat{q}_{\mathrm{NW}} - \hat{q}_{\mathrm{S}} - \hat{q}_{\mathrm{SE}} - \hat{q}_{\mathrm{SW}})/6, \tag{5.41}$$

$$a_{xx} = (-6\hat{q}_{\mathrm{C}} + 3\hat{q}_{\mathrm{E}} - 2\hat{q}_{\mathrm{N}} + \hat{q}_{\mathrm{NE}} + \hat{q}_{\mathrm{NW}} - 2\hat{q}_{\mathrm{S}} + \hat{q}_{\mathrm{SE}} + \hat{q}_{\mathrm{SW}} + 3\hat{q}_{\mathrm{W}})/10, \tag{5.42}$$

$$a_{xy} = (\hat{q}_{\mathrm{NE}} - \hat{q}_{\mathrm{NW}} - \hat{q}_{\mathrm{SE}} + \hat{q}_{\mathrm{SW}})/4, \tag{5.43}$$

$$a_{yy} = (-6\hat{q}_{\mathrm{C}} - 2\hat{q}_{\mathrm{E}} + 3\hat{q}_{\mathrm{N}} + \hat{q}_{\mathrm{NE}} + \hat{q}_{\mathrm{NW}} + 3\hat{q}_{\mathrm{S}} + \hat{q}_{\mathrm{SE}} + \hat{q}_{\mathrm{SW}} - 2\hat{q}_{\mathrm{W}})/10. \tag{5.44}$$

The reconstruction

$$Q_{ij}^{\mathrm{rec},\mathcal{P}3}(\boldsymbol{x}) = \mathcal{P}3_{ij}\left(\boldsymbol{x}; \{\hat{q}_{kl}\}_{(k,l)\in\mathcal{S}_{ij}}\right) := P(\bar{\boldsymbol{x}}(\boldsymbol{x})) \tag{5.45}$$

correctly retrieves the original parabola whenever q is parabolic. However, similar to the one-dimensional polynomial reconstruction described in Section 3.4.1, this reconstruction routine introduces spurious oscillations if q is not smooth enough, thus possibly violating the invariant domain preservation necessary for stability [22]. In order to avoid these oscillations, in our numerical experiments in Section 6.4 we apply the limited CWENO3 reconstruction to obtain a third order accurate method.

[4] We used the technical computing system Mathematica [87] for this.

5.6.2.1 Third Order Accurate Central Weighted Essentially Non-Oscillatory Reconstruction

In the following, let us denote the reconstruction polynomial obtained via the unlimited reconstruction described above as P_{opt} and refer to it as *optimal polynomial*. As for all CWENO methods, the reconstruction polynomial for the two-dimensional CWENO3 method introduced in [106] is obtained as linear combination of different polynomials of different orders. As in [106] we define the linear reconstruction polynomials

$$P_{\text{NE}}\left(\bar{\boldsymbol{x}}\right) := \hat{q}_C + \left(\hat{q}_{\text{E}} - \hat{q}_{\text{C}}\right)\bar{x} + \left(\hat{q}_{\text{N}} - \hat{q}_{\text{C}}\right)\bar{y}, \tag{5.46}$$

$$P_{\text{NW}}(\bar{\boldsymbol{x}}) := \hat{q}_C - \left(\hat{q}_{\text{W}} - \hat{q}_{\text{C}}\right)\bar{x} + \left(\hat{q}_{\text{N}} - \hat{q}_{\text{C}}\right)\bar{y}, \tag{5.47}$$

$$P_{\text{SW}}\left(\bar{\boldsymbol{x}}\right) := \hat{q}_C - \left(\hat{q}_{\text{W}} - \hat{q}_{\text{C}}\right)\bar{x} - \left(\hat{q}_{\text{S}} - \hat{q}_{\text{C}}\right)\bar{y}, \tag{5.48}$$

$$P_{\text{SE}}\left(\bar{\boldsymbol{x}}\right) := \hat{q}_C + \left(\hat{q}_{\text{E}} - \hat{q}_{\text{C}}\right)\bar{x} - \left(\hat{q}_{\text{S}} - \hat{q}_{\text{C}}\right)\bar{y} \tag{5.49}$$

and the central polynomial

$$P_{\text{C}}(\bar{\boldsymbol{x}}) := \frac{1}{C_{\text{C}}}\left(P_{\text{opt}}(\bar{\boldsymbol{x}}) - \sum_{k\in\mathfrak{L}} C_k P_k(\bar{\boldsymbol{x}})\right), \tag{5.50}$$

where the linear weights $C_{\text{C}} = \frac{1}{2}$ and $C_k = \frac{1}{8}$ for $k \in \mathfrak{L} := \{\text{NE}, \text{NW}, \text{SW}, \text{SE}\}$ are chosen such that the reconstruction routine in the end is actually third order accurate on smooth solutions [106]. The final CWENO3 reconstruction polynomial is then obtained as linear combination

$$P_{\text{CWENO3}}(\bar{\boldsymbol{x}}) := \sum_{k\in\mathfrak{L}\cup\{C\}} \omega_k P_k(\bar{\boldsymbol{x}}). \tag{5.51}$$

The rest of the construction is similar to the one in Section 3.4.2.1. In order to recover the optimal polynomial $P_{\text{CWENO3}} \approx P_{\text{opt}}$ we require $\omega_k \approx C_k$ for all nonlinear weights ω_k ($k \in \mathfrak{L}\cup\{C\}$) on smooth solutions. On the other hand, the weights have to be chosen such that ω_k takes a small value if P_k varies strongly thus indicating a discontinuity. For consistency, we require the relation $\sum_{k\in\mathfrak{L}\cup\{C\}}\omega_k = 1$. In order to satisfy all of these requirements, [106], for example, generalizes the one-dimensional approach from [91, 105] and defines the non-linear weights by

$$\omega_k := \frac{\alpha_k}{\sum_{l\in\mathfrak{L}\cup\{C\}}\alpha_l} \quad \text{with} \quad \alpha_k := \frac{C_k}{(\varepsilon + IS_k)^p} \quad \text{for} \quad k \in \mathfrak{L}\cup\{C\} \tag{5.52}$$

using the *smoothness indicators*

$$IS_k := \sum_{|\kappa|=1,2}\int_{\Omega_{\text{C}}} \left(D^\kappa P_k(\bar{\boldsymbol{x}})\right)^2 d\bar{\boldsymbol{x}}. \tag{5.53}$$

The constants p and ε are not derived from assumptions in a mathematical way, but the values $p = 2$ and $\epsilon = 10^{-6}$ that yield accurate and robust results in practice are determined empirically in [91, 105] for one-dimensional methods and also applied in [106] for the two-dimensional method. However, different from this we follow the suggestion from [97] in this thesis and use $\varepsilon = \Delta x\Delta y$ as a natural extension to the choice in one spatial dimension (see discussion in Section 3.4.2.1). The reconstruction routine is then

$$Q_{ij}^{\text{rec,CWENO3}}(\boldsymbol{x}) = \text{CWENO3}_{ij}\left(\boldsymbol{x}; \{\hat{q}_{kl}\}_{(k,l)\in\mathcal{S}_{ij}}\right) := P_{\text{CWENO3}}(\bar{\boldsymbol{x}}(\boldsymbol{x})). \tag{5.54}$$

5.7 Source Terms

In our one-dimensional FV method we used a simple source term discretization (see Eq. (3.76)) which is straight forward to extend to multi-dimensional methods: Let $\boldsymbol{x}_i$ be the cell-center of the cell Ω_i, $i \in \mathcal{I}$.[5] The second order accurate source term (3.76) is then extended to two spatial dimensions by

$$\hat{\boldsymbol{S}}_i^{\mathrm{cc}}(t) := \boldsymbol{s}\left(\hat{\boldsymbol{Q}}_i(t), \boldsymbol{x}_i, t\right). \tag{5.55}$$

In order to construct an m-th order accurate source term discretization, let $\boldsymbol{Q}_i^{\mathrm{rec}}$ be the reconstructed state (m-th order accurate) in the cell Ω_i for $i \in \mathcal{I}$.

$$\hat{\boldsymbol{S}}_i^{\mathrm{quad}}(t) := \frac{1}{|\Omega_i|} I_{\boldsymbol{x}\in\Omega_i}\left[\boldsymbol{s}\left(\boldsymbol{Q}_i^{\mathrm{rec}}(\boldsymbol{x}, t), \boldsymbol{x}, t\right)\right] \tag{5.56}$$

is then an m-th order accurate source term discretization provided that I is an at least m-th order accurate quadrature rule. In the case of a Cartesian grid, we can use the quadrature rules and reconstruction techniques discussed in Sections 5.5 and 6.3.1.2 to develop the source term discretization (5.56). We state the accuracy of these discretizations without proof, since it can be shown easily similar to the accuracy of the one-dimensional source term discretizations in Section 4.4.1.1.

5.8 A High Order Two-Dimensional Runge–Kutta Finite Volume Method

In this section we present a basic arbitrary order RK-FV method for two-dimensional hyperbolic balance laws using the techniques established in the previous sections. This description is similar to the corresponding description in [14].

Consider the two-dimensional system of hyperbolic balance laws

$$\partial_t \boldsymbol{q}(\boldsymbol{x}, t) + \nabla \cdot \mathcal{F}(\boldsymbol{q}(\boldsymbol{x}, t)) = \boldsymbol{s}(\boldsymbol{q}(\boldsymbol{x}, t), \boldsymbol{x}, t). \tag{5.57}$$

Using any grid (grids are discussed in Section 5.3) we divide the domain into N control volumes (i.e., cells). For the i-th control volume Ω_i ($i \in \mathcal{I} := \{1, \ldots, N\}$), we define the cell-average

$$\hat{\boldsymbol{q}}_i(t) := \frac{1}{V_i} \int_{\Omega_i} \boldsymbol{q}(\boldsymbol{x}, t) d\boldsymbol{x}, \tag{5.58}$$

where $V_i = |\Omega_i|$ is the volume of Ω_i. Integrating Eq. (5.57) over Ω_i and applying the divergence theorem yields the evolution equation

$$\frac{d}{dt}\hat{\boldsymbol{q}}_i(t) + \frac{1}{V_i} \int_{\partial\Omega_i} \mathcal{F}(\boldsymbol{q}(\boldsymbol{x}, t)) \cdot \boldsymbol{n}(\boldsymbol{x}) d\sigma = \frac{1}{V_i} \int_{\Omega_i} \boldsymbol{s}(\boldsymbol{q}(\boldsymbol{x}, t), \boldsymbol{x}) d\boldsymbol{x} \tag{5.59}$$

[5] For Cartesian grids it is clear what this means. For non-Cartesian grids, the cell-center can be defined in different ways and the source term discretization discussed here might only be first order accurate depending on the choice of $\boldsymbol{x}_i$.

for the cell-averaged state $\hat{\boldsymbol{q}}_i$. Using the notation $\mathcal{N}_i$ from Section 5.3 we can state the equivalent formulation

$$\frac{d}{dt}\hat{\boldsymbol{q}}_i(t) = -\frac{1}{V_i}\sum_{k\in\mathcal{N}_i}\int_{\partial\Omega_{ik}} \mathcal{F}(\boldsymbol{q}(\boldsymbol{x},t))\cdot\boldsymbol{n}(\boldsymbol{x})d\sigma + \frac{1}{V_i}\int_{\Omega_i}\boldsymbol{s}(\boldsymbol{q}(\boldsymbol{x},t),\boldsymbol{x},t)d\boldsymbol{x}. \tag{5.60}$$

For the discretization of the interface fluxes we use a numerical flux function $\boldsymbol{F}(\cdot,\cdot,\boldsymbol{n})$ consistent with $\boldsymbol{n}\cdot\mathcal{F}$ (see Section 5.4). We apply this discretization to Eq. (5.60) and obtain

$$\begin{aligned}\frac{d}{dt}\hat{\boldsymbol{Q}}_i(t) = -\frac{1}{V_i}\sum_{k\in\mathcal{N}_i}\int_{\partial\Omega_{ik}} \boldsymbol{F}\left(\boldsymbol{Q}_i^{\mathrm{rec}}(\boldsymbol{x},t),\boldsymbol{Q}_k^{\mathrm{rec}}(\boldsymbol{x},t),\boldsymbol{n}(\boldsymbol{x})\right)d\sigma \\ + \frac{1}{V_i}\int_{\Omega_i}\boldsymbol{s}\left(\boldsymbol{Q}_i^{\mathrm{rec}}(\boldsymbol{x},t),\boldsymbol{x},t\right)d\boldsymbol{x},\end{aligned} \tag{5.61}$$

where the reconstructed functions $\boldsymbol{Q}_i^{\mathrm{rec}}, \boldsymbol{Q}_k^{\mathrm{rec}}$ are obtained using a consistent conservative reconstruction routine on the cell average values $\hat{\boldsymbol{Q}}$ of the approximate solution. In the next step we use numerical quadrature rules (Sections 3.5 and 5.5) for the interface flux integral and a discretization of the source term integral (5.7). The semi-discrete method is then

$$\frac{d}{dt}\hat{\boldsymbol{Q}}_i(t) = -\frac{1}{V_i}\left(\sum_{k\in\mathcal{N}_i} I_{\boldsymbol{x}\in\partial\Omega_{ik}}\left[\boldsymbol{F}\left(\boldsymbol{Q}_i^{\mathrm{rec}}(\cdot,t),\boldsymbol{Q}_k^{\mathrm{rec}}(\cdot,t),\boldsymbol{n}\right)\right]\right) + \hat{\boldsymbol{S}}_i. \tag{5.62}$$

The semi-discrete scheme (5.62) is m-th order accurate if the applied reconstruction routine, interface flux quadrature, and source term discretization are all at least m-th order accurate. It is then evolved in time using an at least m-th order accurate RK method (see Section 3.7) to obtain an m-th order accurate fully discrete scheme.

5.9 A Runge–Kutta Finite Volume Method on a Curvilinear Grid

While we discussed the general form of a RK-FV method in the previous section, we restrict to the special case of a RK-FV method on a curvilinear (structured) grid in this section to introduce a more specific notation. Consider a curvilinear mapping $\boldsymbol{\xi}\mapsto\boldsymbol{x}(\boldsymbol{\xi})$ as introduced in Section 5.3.1. The grid vertices are defined by mapping

$$\boldsymbol{x}_{i+\frac{1}{2},j+\frac{1}{2}} := \boldsymbol{x}\left(\boldsymbol{\xi}_{i+\frac{1}{2},j+\frac{1}{2}}\right). \tag{5.63}$$

The edge-centered values are given by the arithmetic mean, i.e.,

$$\boldsymbol{x}_{i+\frac{1}{2},j} := \frac{1}{2}\left(\boldsymbol{x}_{i+\frac{1}{2},j+\frac{1}{2}} + \boldsymbol{x}_{i+\frac{1}{2},j-\frac{1}{2}}\right), \quad \boldsymbol{x}_{i,j+\frac{1}{2}} := \frac{1}{2}\left(\boldsymbol{x}_{i+\frac{1}{2},j+\frac{1}{2}} + \boldsymbol{x}_{i-\frac{1}{2},j+\frac{1}{2}}\right). \tag{5.64}$$

We assume w.l.o.g. that $\Delta\xi = \Delta\eta = 1$ on the whole computational grid. The metric terms, i.e., the derivatives of the coordinate transform, have to be discretized carefully: Using the actual analytical derivative of the coordinate mapping can introduce

spurious source term (e.g., [158, 157]). In order to avoid this, the discretization of the metric terms has to satisfy the relation [95]

$$(\boldsymbol{x}_\eta)_{i+\frac{1}{2},j} - (\boldsymbol{x}_\eta)_{i-\frac{1}{2},j} = (\boldsymbol{x}_\xi)_{i,j+\frac{1}{2}} - (\boldsymbol{x}_\xi)_{i,j-\frac{1}{2}}, \tag{5.65}$$

which can be seen as a discrete version of Schwartz's theorem (stating symmetry of second derivatives of $\mathcal{C}^2$-functions). We choose the following discretization

$$(\boldsymbol{x}_\xi)_{i,j+\frac{1}{2}} := \begin{pmatrix}(x_\xi)_{i,j+\frac{1}{2}}\\(y_\xi)_{i,j+\frac{1}{2}}\end{pmatrix} := \begin{pmatrix}x_{i+\frac{1}{2},j+\frac{1}{2}} - x_{i-\frac{1}{2},j+\frac{1}{2}}\\y_{i+\frac{1}{2},j+\frac{1}{2}} - y_{i-\frac{1}{2},j+\frac{1}{2}}\end{pmatrix}, \tag{5.66}$$

$$(\boldsymbol{x}_\eta)_{i+\frac{1}{2},j} := \begin{pmatrix}(x_\eta)_{i+\frac{1}{2},j}\\(y_\eta)_{i+\frac{1}{2},j}\end{pmatrix} := \begin{pmatrix}x_{i+\frac{1}{2},j+\frac{1}{2}} - x_{i+\frac{1}{2},j-\frac{1}{2}}\\y_{i+\frac{1}{2},j+\frac{1}{2}} - y_{i+\frac{1}{2},j-\frac{1}{2}}\end{pmatrix}, \tag{5.67}$$

which satisfies Eq. (5.65). The interface areas are then computed as

$$A_{i+\frac{1}{2},j} := \left\|(\boldsymbol{x}_\eta)_{i+\frac{1}{2},j}\right\|_2, \qquad A_{i,j+\frac{1}{2}} := \left\|(\boldsymbol{x}_\xi)_{i,j+\frac{1}{2}}\right\|_2 \tag{5.68}$$

and the unit normal vectors are

$$\boldsymbol{n}_{i+\frac{1}{2},j} := \begin{pmatrix}(n_x)_{i+\frac{1}{2},j}\\(n_y)_{i+\frac{1}{2},j}\end{pmatrix} := \frac{1}{A_{i+\frac{1}{2},j}}\begin{pmatrix}(y_\eta)_{i+\frac{1}{2},j}\\-(x_\eta)_{i+\frac{1}{2},j}\end{pmatrix}, \tag{5.69}$$

$$\boldsymbol{n}_{i,j+\frac{1}{2}} := \begin{pmatrix}(n_x)_{i,j+\frac{1}{2}}\\(n_y)_{i,j+\frac{1}{2}}\end{pmatrix} := \frac{1}{A_{i,j+\frac{1}{2}}}\begin{pmatrix}(y_\xi)_{i,j+\frac{1}{2}}\\-(x_\xi)_{i,j+\frac{1}{2}}\end{pmatrix}. \tag{5.70}$$

With this, we approximate the interface fluxes between two neighboring cells to second order via

$$A_{i+\frac{1}{2},j}\boldsymbol{F}_{i+\frac{1}{2},j} := A_{i+\frac{1}{2},j}\boldsymbol{F}\left(\boldsymbol{Q}^-_{i+\frac{1}{2},j}, \boldsymbol{Q}^+_{i+\frac{1}{2},j}, \boldsymbol{n}_{i+\frac{1}{2},j}\right) \tag{5.71}$$

$$\approx \int_{\partial\Omega_{i+\frac{1}{2},j}} \boldsymbol{n}(\boldsymbol{x})\cdot\hat{\mathcal{F}}\left(\boldsymbol{q}(\boldsymbol{x})\right)\,d\boldsymbol{x}, \tag{5.72}$$

$$A_{i,j+\frac{1}{2}}\boldsymbol{F}_{i,j+\frac{1}{2}} := A_{i,j+\frac{1}{2}}\boldsymbol{F}\left(\boldsymbol{Q}^-_{i,j+\frac{1}{2}}, \boldsymbol{Q}^+_{i,j+\frac{1}{2}}, \boldsymbol{n}_{i,j+\frac{1}{2}}\right) \tag{5.73}$$

$$\approx \int_{\partial\Omega_{i,j+\frac{1}{2}}} \boldsymbol{n}(\boldsymbol{x})\cdot\hat{\mathcal{F}}\left(\boldsymbol{q}(\boldsymbol{x})\right)\,d\boldsymbol{x}, \tag{5.74}$$

where the interface values $\boldsymbol{Q}^\pm_{i+\frac{1}{2},j}, \boldsymbol{Q}^\pm_{i,j+\frac{1}{2}}$ are obtained using a linear consistent reconstruction as described in Section 5.6.1 on the computational grid. The semi-discrete scheme

$$\begin{aligned}\frac{d}{dt}\hat{\boldsymbol{Q}}_{ij}(t) = -\frac{1}{V_{ij}}\Big[\;& A_{i+\frac{1}{2},j}\boldsymbol{F}_{i+\frac{1}{2},j} - A_{i-\frac{1}{2},j}\boldsymbol{F}_{i-\frac{1}{2},j}\\ &+A_{i,j+\frac{1}{2}}\boldsymbol{F}_{i,j+\frac{1}{2}} - A_{i,j-\frac{1}{2}}\boldsymbol{F}_{i,j-\frac{1}{2}}\Big] + \hat{\boldsymbol{S}}_{ij},\end{aligned} \tag{5.75}$$

where the source term is evaluated at the cell-average state (in the manner of Eq. (3.76)) using the gravity at the cell-center

$$\boldsymbol{x}_{ij} = \frac{1}{4}\left(\boldsymbol{x}_{i+\frac{1}{2},j} + \boldsymbol{x}_{i-\frac{1}{2},j} + \boldsymbol{x}_{i,j+\frac{1}{2}} + \boldsymbol{x}_{i,j-\frac{1}{2}}\right) \tag{5.76}$$

is then evolved in time using an RK-Method. Trying other choices than (5.76) for the cell-centered coordinates in numerical experiments, we could not see a significant difference. For example, we tested center-of-mass coordinates (i.e., cell-averaged coordinates), cell-center coordinates obtained using the mapping on the Cartesian cell-center coordinates, and interface squared weighted cell-centers, which have been suggested for triangular grids in [121] with the aim to reduce the *skewness* of a grid (see [121] for skewness). On a Cartesian grid, all of these choices yield the same cell-center coordinates and the scheme is second order accurate. Thus we use the RK3 method to evolve it in time. Also, since it yields second order convergence on genuinely curvilinear grids in the numerical tests in [15], we refer to this method as second order method in the following for simplicity.

5.10 Boundary Conditions

Let us assume a structured grid that discretizes the domain Ω using the cells Ω_{ij} with $(i,j) \in \mathcal{I}$, $\mathcal{I} = \mathcal{I}_\xi \times \mathcal{I}_\eta := \{1,\dots,N_\xi\} \times \{1,\dots,N_\eta\}$, which simplifies the treatment of boundaries significantly. Boundary conditions for unstructured meshes are for example discussed in [19]. The sets of indices $\mathcal{G}_\chi := \{-N_{\text{gc}}+1,\dots,0,N_\chi+1,\dots,N_\chi+N_{\text{gc}}\}$ $(\chi = \xi,\eta)$ are used to describe the ghost cells

$$\Omega_{ij}, \qquad \text{where} \qquad (i,j) \in \mathcal{G} := (\mathcal{G}_\xi \times \mathcal{G}_\eta) \cup (\mathcal{G}_\xi \times \mathcal{I}_\eta) \cup (\mathcal{I}_\xi \times \mathcal{G}_\eta)\,. \tag{5.77}$$

We assume that $N_\xi, N_\eta > 2N_{\text{gc}}$. In the following, we discuss how the values $\hat{\boldsymbol{Q}}_{ij}$ for $(i,j) \in \mathcal{G}$ can be set prior to each reconstruction step. This is an extension of the discussion in Section 3.8, in which boundary conditions for the one-dimensional RK-FV scheme have been presented.

Periodic boundary conditions The periodic boundary conditions can be extended to two spatial dimensions in a straightforward way by setting

$$\hat{\boldsymbol{Q}}_{ij} = \hat{\boldsymbol{Q}}_{((i-1) \mod N_\xi)+1,\,((j-1) \mod N_\eta)+1} \qquad \text{for} \qquad (i,j) \in \mathcal{G}. \tag{5.78}$$

Dirichlet boundary conditions As in the one-dimensional case, Dirichlet boundary conditions can be achieved by simply setting all the cell-averages in the ghost cells according to some given function. Depending on the function, a sufficiently high order accurate quadrature rule can be applied instead of exact integration to compute the cell-averaged values.

Wall boundary conditions In Section 3.8 we described wall-boundary conditions for one-dimensional FV methods and provided the particular example of compressible Euler equations. Now we describe wall boundary conditions for the example of two-dimensional compressible Euler equations. For brevity, we describe it only on the interfaces in ξ-direction. Let $\boldsymbol{n}_{i+\frac{1}{2},j}$ be the normal vector at the interface between the cells Ω_{ij} and $\Omega_{i+1,j}$. Since our approach to Curvilinear grids uses approximations that limit their accuracy anyway (see Section 5.9) it is sufficient to only use the normal vector at the center of the interface. For higher order methods

a more sophisticated approach is necessary. On Cartesian grids, however, the normal vector is constant on the whole interface such that there is no restriction to the order of accuracy of the method introduced by this assumption. The velocity $\hat{\boldsymbol{v}}_{kl}$ is decomposed into the perpendicular and parallel velocity components with respect to the interface $\delta\Omega_{i+\frac{1}{2},j}$ via

$$\hat{\boldsymbol{v}}_{kl}^{\perp,i+\frac{1}{2},j} := \left(\boldsymbol{n}_{i+\frac{1}{2},j} \cdot \hat{\boldsymbol{v}}_k\right) \boldsymbol{n}_{i+\frac{1}{2},j} \qquad \hat{\boldsymbol{v}}_{kl}^{\parallel,i+\frac{1}{2},j} := \hat{\boldsymbol{v}}_{kl} - \hat{\boldsymbol{v}}_{kl}^{\perp,i+\frac{1}{2},j}. \tag{5.79}$$

The values in the ghost cells in ξ-direction are then set as

$$\hat{\boldsymbol{Q}}_{1-\iota,j} = \begin{pmatrix} \hat{\rho}_{\iota,j} \\ \hat{\boldsymbol{v}}_{\iota,j}^{\parallel,\frac{1}{2},j} - \hat{\boldsymbol{v}}_{\iota,j}^{\perp,\frac{1}{2},j} \\ \hat{E}_{\iota,j} \end{pmatrix}, \quad \text{and} \quad \hat{\boldsymbol{Q}}_{N_\xi+\iota,j} = \begin{pmatrix} \hat{\rho}_{N_\xi+1-\iota,j} \\ \hat{\boldsymbol{v}}_{N_\xi+1-\iota,j}^{\parallel,\frac{1}{2},j} - \hat{\boldsymbol{v}}_{N_\xi+1-\iota,j}^{\perp,N_\xi+\frac{1}{2},j} \\ \hat{E}_{N_\xi+1-\iota,j} \end{pmatrix} \tag{5.80}$$

for $\iota \in \{1, \dots, N_{\mathrm{gc}}\}$ and $j \in \mathcal{I}_\eta \cup \mathcal{G}_\eta$. The corresponding procedure is applied in η-direction. The corner ghost cell values $\hat{\boldsymbol{Q}}_{ij}$ for $(i,j) \in \mathcal{G}_\xi \times \mathcal{G}_\eta$ are automatically set correctly by applying wall boundary condition in both coordinate directions separately. Note that the application of the wall boundary conditions in ξ and η-direction commutes.

Extrapolation boundary conditions The two-dimensional extrapolation boundary conditions are a straightforward extension of the one-dimensional extrapolation boundary conditions described in Section 3.8, hence we omit showing them for the cells with indices in $\mathcal{I}_\xi \times \mathcal{G}_\eta$ and $\mathcal{G}_\xi \times \mathcal{I}_\eta$. Of particular interest are only the ghost cells in the corners: For $(i,j) \in \{-N_{\mathrm{gc}}+1, \dots, 0\} \times \{-N_{\mathrm{gc}}+1, \dots, 0\}$, for example, we set

$$\hat{\boldsymbol{Q}}_{ij} = \frac{1}{|\Omega_{ij}|} \int_{\Omega_{ij}} \hat{\boldsymbol{Q}}_{1,1}^{\mathrm{rec}}(\boldsymbol{x})\, d\boldsymbol{x} \tag{5.81}$$

and correspondingly in the remaining three corners. In a second order method using a dimension-by-dimension reconstruction approach, the ghost cells in the corners of the domain are neither used in the reconstruction process nor – at least in the methods in this thesis – in the source term discretization. Hence, these corner ghost cell values are only relevant in higher order methods.

Chapter 6

Multi-Dimensional Well-Balanced Finite Volume Methods

In this chapter we present multi-dimensional extensions of the well-balanced methods introduced in Chapter 4. These extensions are straightforward for the α-β method Section 6.1 and the Deviation method Section 6.2, since the target solution which is balanced is assumed to be given. The α-β method is extended to two spatial dimensions using a spatial splitting approach. This is sufficient to obtain second order accuracy. For the Deviation method we use a genuinely multi-dimensional hydrostatic reconstruction and source term discretization. Together with interface flux quadrature and high order Runge–Kutta methods arbitrarily high order of accuracy can be achieved.

In the case of the approximately well-balanced methods, the situation is different. The target solution is not given a priori. Instead, an approximately hydrostatic solution close to the current discrete density and pressure data is constructed in each cell and well-balanced by the method. Extending this to multi-d includes finding a multi-dimensional discrete approximation to a hydrostatic state. In one spatial dimension, the technique to find discrete hydrostatic states is integrating the source term approximation to neighboring cells to find an approximation to a hydrostatic pressure. This is in general not equally simple in two spatial dimensions. The reason is that in multi-d, the integrals connecting different hydrostatic pressure points are in general path dependent, if applied on the source term approximation. This way, it is not possible to construct a global approximation of source term and hydrostatic pressure which is well-balanced exactly.

Instead, we only extend the Local Approximation method to two spatial dimensions. As in the one-dimensional method, no theorem can be stated about well-balancing. However, numerical tests justify this extension, since it leads to increased convergence rates towards hydrostatic states and improved accuracy close to hydrostatic states.

6.1 The α-β Method

In this section we present a two-dimensional extension of the α-β method previously presented for one spatial dimension in Section 4.2. This extension has been published

in [15].

6.1.1 Description of the Two-Dimensional α-β Method

Representation of the hydrostatic solution Let $\Omega \subset \mathbb{R}^2$ be a domain suitable for the discretization with a structured curvilinear grid (Section 5.3.1). Consider a hydrostatic solution $(\rho, u, p) = (\alpha, 0, \beta)$ described by the functions $\alpha : \Omega \to \mathbb{R}^+$ and $\beta \in \mathcal{C}^1(\Omega, \mathbb{R}^+)$, which satisfy the hydrostatic equation (5.14). Then we have

$$\nabla\beta(\boldsymbol{x}) = \alpha(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x}), \qquad \text{i.e.,} \qquad \boldsymbol{g}(\boldsymbol{x}) = \frac{\nabla\beta(\boldsymbol{x})}{\alpha(\boldsymbol{x})}. \tag{6.1}$$

Analogously to the one-dimensional version of the method, we only modify two key components of the finite volume method (5.75) for the two-dimensional compressible Euler equations with gravity source term (5.5).

Reconstruction For each $\boldsymbol{x} \in \Omega$ we define the transformation

$$\mathcal{T}_{\boldsymbol{x}}^{\alpha\text{-}\beta} : \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+, \tag{6.2}$$

$$\mathcal{T}_{\boldsymbol{x}}^{\alpha\text{-}\beta}(\boldsymbol{q}) := \begin{pmatrix} \frac{1}{\alpha(\boldsymbol{x})} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{\beta(\boldsymbol{x})} \end{pmatrix} \left.\frac{\partial \boldsymbol{q}^{\text{prim}}}{\partial \boldsymbol{q}^{\text{cons}}}\right|_{\boldsymbol{q}} \boldsymbol{q}. \tag{6.3}$$

We extend the hydrostatic reconstruction $\mathcal{R}^{\alpha\text{-}\beta}$ presented in Section 4.2 to two spatial dimensions by applying it in each direction separately, i.e.,

$$\boldsymbol{Q}^-_{i+\frac{1}{2},j} := \left(\mathcal{T}^{\alpha\text{-}\beta}_{\boldsymbol{x}_{i+\frac{1}{2},j}}\right)^{-1} \left(\mathcal{R}_{ij}\left(\boldsymbol{x}_{i+\frac{1}{2},j}; \left\{\mathcal{T}^{\alpha\text{-}\beta}_{\boldsymbol{x}_{kj}}\left(\hat{\boldsymbol{Q}}_{kj}\right)\right\}_{(k,j)\in\mathcal{S}_{ij}}\right)\right), \tag{6.4}$$

$$\boldsymbol{Q}^+_{i+\frac{1}{2},j} := \left(\mathcal{T}^{\alpha\text{-}\beta}_{\boldsymbol{x}_{i+\frac{1}{2},j}}\right)^{-1} \left(\mathcal{R}_{i+1,j}\left(\boldsymbol{x}_{i+\frac{1}{2},j}; \left\{\mathcal{T}^{\alpha\text{-}\beta}_{\boldsymbol{x}_{kj}}\left(\hat{\boldsymbol{Q}}_{kj}\right)\right\}_{(k,j)\in\mathcal{S}_{i+1,j}}\right)\right), \tag{6.5}$$

$$\boldsymbol{Q}^-_{i,j+\frac{1}{2}} := \left(\mathcal{T}^{\alpha\text{-}\beta}_{\boldsymbol{x}_{i,j+\frac{1}{2}}}\right)^{-1} \left(\mathcal{R}_{ij}\left(\boldsymbol{x}_{i,j+\frac{1}{2}}; \left\{\mathcal{T}^{\alpha\text{-}\beta}_{\boldsymbol{x}_{ik}}\left(\hat{\boldsymbol{Q}}_{ik}\right)\right\}_{(i,k)\in\mathcal{S}_{ij}}\right)\right), \tag{6.6}$$

$$\boldsymbol{Q}^+_{i,j+\frac{1}{2}} := \left(\mathcal{T}^{\alpha\text{-}\beta}_{\boldsymbol{x}_{i,j+\frac{1}{2}}}\right)^{-1} \left(\mathcal{R}_{i,j+1}\left(\boldsymbol{x}_{i,j+\frac{1}{2}}; \left\{\mathcal{T}^{\alpha\text{-}\beta}_{\boldsymbol{x}_{ik}}\left(\hat{\boldsymbol{Q}}_{ik}\right)\right\}_{(i,k)\in\mathcal{S}_{i,j+1}}\right)\right). \tag{6.7}$$

Note that for $\mathcal{R}$ we still use a one-dimensional reconstruction. In practice, $\mathcal{R}$ will be a constant or limited linear reconstruction as presented in Section 3.4. Higher order reconstruction can not further increase the order of accuracy of the method but it can increase the computational effort.

Source term Using Eq. (6.1), we can write the source term of the momentum equation as

$$\boldsymbol{s}^{\rho\boldsymbol{v}}(\boldsymbol{x}, t) = \frac{\nabla\beta(\boldsymbol{x})}{\alpha(\boldsymbol{x})}\rho(\boldsymbol{x}, t). \tag{6.8}$$

To discretize the cell average of Eq. (6.8) in the ij-th cell, we approximate

$$\frac{1}{V_{ij}} \int_{\Omega_{ij}} \boldsymbol{s}^{\rho \boldsymbol{v}} \, d\boldsymbol{x} \approx \frac{1}{V_{ij}} \int_{\Omega_{ij}^{\text{comp}}} J \boldsymbol{s}^{\rho \boldsymbol{v}} \, d\boldsymbol{\xi} = -\frac{1}{V_{ij}} \frac{\hat{\rho}_{ij}}{\alpha_{ij}} \left(J \nabla \beta \right)_{ij} + \mathcal{O}\left(h^2\right) \tag{6.9}$$

with

$$\begin{aligned} \left(J \nabla \beta \right)_{ij} &:= \begin{pmatrix} \left(J \partial_x \beta \right)_{ij} \\ \left(J \partial_y \beta \right)_{ij} \end{pmatrix} := \\ &\begin{pmatrix} (y_\eta)_{i+\frac{1}{2},j} \beta_{i+\frac{1}{2},j} - (y_\eta)_{i-\frac{1}{2},j} \beta_{i-\frac{1}{2},j} - (y_\xi)_{i,j+\frac{1}{2}} \beta_{i,j+\frac{1}{2}} + (y_\xi)_{i,j-\frac{1}{2}} \beta_{i,j-\frac{1}{2}} \\ -(x_\eta)_{i+\frac{1}{2},j} \beta_{i+\frac{1}{2},j} - (x_\eta)_{i-\frac{1}{2},j} \beta_{i-\frac{1}{2},j} + (x_\xi)_{i,j+\frac{1}{2}} \beta_{i,j+\frac{1}{2}} - (x_\xi)_{i,j-\frac{1}{2}} \beta_{i,j-\frac{1}{2}} \end{pmatrix} \end{aligned} \tag{6.10}$$

based on

$$\begin{aligned} J \nabla \beta = J \begin{pmatrix} \partial_x \beta \\ \partial_y \beta \end{pmatrix} &= J \begin{pmatrix} (\partial_x \xi)(\partial_\xi \beta) + (\partial_x \eta)(\partial_\eta \beta) \\ (\partial_y \xi)(\partial_\xi \beta) + (\partial_y \eta)(\partial_\eta \beta) \end{pmatrix} \\ &= \begin{pmatrix} (\partial_\eta y)(\partial_\xi \beta) - (\partial_\xi y)(\partial_\eta \beta) \\ -(\partial_\eta x)(\partial_\xi \beta) + (\partial_\xi x)(\partial_\eta \beta) \end{pmatrix} = \begin{pmatrix} \partial_\xi((\partial_\eta y)\beta) - \partial_\eta((\partial_\xi y)\beta) \\ -\partial_\xi((\partial_\eta x)\beta) + \partial_\eta((\partial_\xi x)\beta) \end{pmatrix}, \end{aligned} \tag{6.11}$$

where we assumed w.l.o.g. that $\Delta\xi = \Delta\eta = 1$ and hence $|\Omega_{ij}^{\text{comp}}| = \Delta\xi\Delta\eta = 1$. The notations and some of the relations we used have been established in Section 5.3.1.1. This leads to the source term discretization

$$\hat{\boldsymbol{S}}_{ij}^{\alpha\text{-}\beta}(t) := \begin{pmatrix} 0 \\ \hat{S}_{ij}^{\rho u, \alpha\text{-}\beta}(t) \\ \hat{S}_{ij}^{\rho v, \alpha\text{-}\beta}(t) \\ \frac{\hat{\rho \boldsymbol{v}}_{ij}}{\hat{\rho}_{ij}} \cdot \hat{\boldsymbol{S}}_{ij}^{\rho \boldsymbol{v}, \alpha\text{-}\beta}(t) \end{pmatrix} \tag{6.12}$$

with

$$\hat{\boldsymbol{S}}_{ij}^{\rho \boldsymbol{v}, \alpha\text{-}\beta}(t) := \begin{pmatrix} \hat{S}_{ij}^{\rho u, \alpha\text{-}\beta}(t) \\ \hat{S}_{ij}^{\rho v, \alpha\text{-}\beta}(t) \end{pmatrix} := -\frac{1}{V_{ij}} \frac{\hat{\rho}_{ij}(t)}{\alpha_{ij}} (J \nabla \beta)_{ij}, \tag{6.13}$$

which is a second order accurate discretization of the cell-averaged source term.

6.1.2 Properties of the α-β Method

The main properties of the one-dimensional α-β method (see Section 4.2.2) also hold in two spatial dimensions, as we show in the following.

6.1.2.1 Accuracy

For the following statement we assume that the mapping for the curvilinear grid is sufficiently smooth and the cell-centered values are given such that they yield a second order accurate standard scheme in the description in Section 5.9.

Theorem 6.1.1. *Consider the semi-discrete finite volume scheme Eq.* (5.75)*, in which the interface states have been obtained using the reconstruction procedure as described in Eqs.* (6.4) *to* (6.7) *and with the source term discretization as defined in Eq.* (6.12) *based on the functions* $\alpha \in \mathcal{C}^2\left(\Omega, \mathbb{R}^+\right)$ *and* $\beta \in \mathcal{C}^3\left(\Omega, \mathbb{R}^+\right)$*. This semi-discrete scheme is second order accurate in space, if the reconstruction* $\mathcal{R}$ *underlying the hydrostatic reconstruction in Eqs.* (6.4) *to* (6.7) *is second order accurate.*

Proof. Let $\boldsymbol{q}$ be a smooth solution. First, we show that the source term approximation $\hat{\boldsymbol{S}}_{ij}^{\alpha\text{-}\beta}$ is second order accurate. $(J\nabla\beta)_{ij}$ as defined in Eq. (6.10) is a discretization of $(J\nabla\beta)(\boldsymbol{x}_{ij})$ realized by central differences. Hence, it is

$$(J\nabla\beta)_{ij} = (J\nabla\beta)(\boldsymbol{x}_{ij}) + \mathcal{O}\left(h^2\right). \tag{6.14}$$

Using the approximations shown in Appendix B.1 this includes

$$\begin{aligned} V_{ij}\hat{\boldsymbol{S}}_{ij}^{\rho\boldsymbol{v},\alpha\text{-}\beta} = -\frac{\hat{\rho}_{ij}}{\alpha_{ij}}\left(J\nabla\beta\right)_{ij} &= -J(\boldsymbol{x}_{ij})\frac{\rho(\boldsymbol{x}_{ij})}{\alpha(\boldsymbol{x}_{ij})}\nabla\beta(\boldsymbol{x}_{ij}) + \mathcal{O}\left(h^2\right) \\ &= J\boldsymbol{s}^{\rho\boldsymbol{v}}\left(\boldsymbol{q}\left(\boldsymbol{x}_{ij}\right), \boldsymbol{x}_{ij}\right) + \mathcal{O}\left(h^2\right). \end{aligned} \tag{6.15}$$

Cell-centered evaluation approximates integrals to second order, hence we have

$$\hat{\boldsymbol{S}}_{ij}^{\rho\boldsymbol{v},\alpha\text{-}\beta} = \frac{1}{V_{ij}}\int_{\Omega_{ij}} \boldsymbol{s}^{\rho\boldsymbol{v}}\, d\boldsymbol{x} + \mathcal{O}\left(h^2\right). \tag{6.16}$$

The energy source term discretization $\hat{\boldsymbol{S}}_{ij}^{E,\alpha\text{-}\beta}$ approximates $\hat{\boldsymbol{s}}_{ij}^{E}$ to second order because of Eq. (6.15) and Appendix B.1.

Now, we show that the hydrostatic reconstruction is second order accurate: First, note that

$$\begin{aligned} \frac{1}{V_{ij}}\int_{\Omega_{ij}} \mathcal{T}_{\boldsymbol{x}_{ij}}^{\alpha\text{-}\beta}\left(\boldsymbol{q}(\boldsymbol{x},t)\right)\, d\boldsymbol{x} &\overset{(*)}{=} \mathcal{T}_{\boldsymbol{x}_{ij}}^{\alpha\text{-}\beta}\left(\boldsymbol{q}(\boldsymbol{x}_{ij},t)\right) + \mathcal{O}\left(h^2\right) \\ &= \mathcal{T}_{\boldsymbol{x}_{ij}}^{\alpha\text{-}\beta}\left(\hat{\boldsymbol{q}}_{ij}(t) + \mathcal{O}\left(h^2\right)\right) + \mathcal{O}\left(h^2\right) \\ &\overset{(**)}{=} \mathcal{T}_{\boldsymbol{x}_{ij}}^{\alpha\text{-}\beta}\left(\hat{\boldsymbol{q}}_{ij}(t)\right) + \mathcal{O}\left(h^2\right), \end{aligned} \tag{6.17}$$

where $(*)$ and $(**)$ follow from the approximation in Appendix B.1. The reconstruction $\mathcal{R}$ is assumed to be second order accurate in the theorem. The back-transformations $\left(\mathcal{T}_{\boldsymbol{x}_{i+\frac{1}{2},j}}^{\alpha\text{-}\beta}\right)^{-1}$ and $\left(\mathcal{T}_{\boldsymbol{x}_{i,j+\frac{1}{2}}}^{\alpha\text{-}\beta}\right)^{-1}$ of the interface centered values transport the second order error as in the one-dimensional case. In total, the hydrostatic reconstruction is second order accurate. This makes the method described in Theorem 6.1.1 second order accurate. □

6.1.2.2 Well-Balanced Property

Theorem 6.1.2. *Consider the semi-discrete finite volume scheme Eq.* (5.75)*, for which the interface values have been obtained with the hydrostatic reconstruction described in Eqs.* (6.4) *to* (6.7) *and with the source term discretization defined in Eq.* (6.12)*. This scheme is well-balanced in the sense that* $\partial_t\hat{\boldsymbol{Q}}_{ij}^{\mathrm{hs}} = 0$ *for initial conditions that satisfy* $\mathcal{T}_{\boldsymbol{x}_{ij}}^{\alpha\text{-}\beta}(\hat{\boldsymbol{Q}}_{ij}^{\mathrm{hs}}) = (a,0,0,a)^T$ *for some constant* $a > 0$ *independent from* i *and* j*. In other words, the relations*

$$\frac{\hat{\rho}_{ij}^{\mathrm{hs}}}{\alpha_{ij}} = \frac{p_{ij}^{\mathrm{hs}}}{\beta_{ij}} = \text{const.}, \qquad \hat{\rho u}_{ij}^{\mathrm{hs}} = 0, \qquad \hat{\rho v}_{ij}^{\mathrm{hs}} = 0 \tag{6.18}$$

with $p_{ij}^{\mathrm{hs}} := p_{EoS}\left(\hat{\rho}_{ij}^{\mathrm{hs}}, \hat{E}_{ij}^{\mathrm{hs}} - \frac{\left(\hat{\rho u}_{ij}^{\mathrm{hs}}\right)^2 + \left(\hat{\rho v}_{ij}^{\mathrm{hs}}\right)^2}{\hat{\rho}_{ij}^{\mathrm{hs}}}\right)$ *lead to a vanishing residual.*

Proof. This proof is in large parts analogous to the one-dimensional case (proof of Theorem 4.2.3). Hence, we omit some of the computations for brevity. Plugging $\hat{\boldsymbol{Q}}_{ij}^{\text{hs}}$ into the hydrostatic reconstruction leads to

$$\left(\boldsymbol{Q}_{i,j+\frac{1}{2}}^{\text{hs}}\right)^{L} = \left(\boldsymbol{Q}_{i,j+\frac{1}{2}}^{\text{hs}}\right)^{R} =: \boldsymbol{Q}_{i+\frac{1}{2},j}^{\text{hs}}, \tag{6.19}$$

$$\left(\boldsymbol{Q}_{i,j+\frac{1}{2}}^{\text{hs}}\right)^{L} = \left(\boldsymbol{Q}_{i,j+\frac{1}{2}}^{\text{hs}}\right)^{R} =: \boldsymbol{Q}_{i,j+\frac{1}{2}}^{\text{hs}}. \tag{6.20}$$

The numerical fluxes at the $i+\frac{1}{2}, j$ and $i+\frac{1}{2}, j$ interfaces are then

$$\begin{aligned}\boldsymbol{F}_{i+\frac{1}{2},j}^{\text{hs}} &:= \boldsymbol{F}\left(\left(\boldsymbol{Q}_{i+\frac{1}{2},j}^{\text{hs}}\right)^{L}, \left(\boldsymbol{Q}_{i+\frac{1}{2},j}^{\text{hs}}\right)^{R}, \boldsymbol{n}_{i+\frac{1}{2},j}\right) \\ &= \boldsymbol{n}_{i+\frac{1}{2},j} \cdot \mathcal{F}\left(\boldsymbol{Q}_{i+\frac{1}{2},j}^{\text{hs}}\right) = \begin{pmatrix} 0 \\ a(n_x)_{i+\frac{1}{2},j}\beta_{i+\frac{1}{2},j} \\ a(n_y)_{i+\frac{1}{2},j}\beta_{i+\frac{1}{2},j} \\ 0 \end{pmatrix}, \end{aligned} \tag{6.21}$$

$$\boldsymbol{F}_{i,j+\frac{1}{2}}^{\text{hs}} := \boldsymbol{F}\left(\left(\boldsymbol{Q}_{i,j+\frac{1}{2}}^{\text{hs}}\right)^{L}, \left(\boldsymbol{Q}_{i,j+\frac{1}{2}}^{\text{hs}}\right)^{R}, \boldsymbol{n}_{i,j+\frac{1}{2}}\right) = \begin{pmatrix} 0 \\ a(n_x)_{i,j+\frac{1}{2}}\beta_{i,j+\frac{1}{2}} \\ a(n_y)_{i,j+\frac{1}{2}}\beta_{i,j+\frac{1}{2}} \\ 0 \end{pmatrix}. \tag{6.22}$$

Equation (4.16) together with Eqs. (6.21) and (6.22) yields

$$\begin{aligned} -V_{ij}\hat{\boldsymbol{S}}_{ij}^{\alpha\text{-}\beta,\text{hs}} &:= \hat{\boldsymbol{S}}_{ij}^{\alpha\text{-}\beta}\Big|_{\hat{\boldsymbol{Q}}_{ij}^{\text{hs}}} = \frac{\hat{\rho}_{ij}^{\text{hs}}}{\alpha_{ij}} \begin{pmatrix} 0 \\ (J\partial_x\beta)_{ij} \\ (J\partial_y\beta)_{ij} \\ 0 \end{pmatrix} \\ &= a \begin{pmatrix} 0 \\ (y_\eta)_{i+\frac{1}{2},j}\beta_{i+\frac{1}{2},j} - (y_\eta)_{i-\frac{1}{2},j}\beta_{i-\frac{1}{2},j} - (y_\xi)_{i,j+\frac{1}{2}}\beta_{i,j+\frac{1}{2}} + (y_\xi)_{i,j-\frac{1}{2}}\beta_{i,j-\frac{1}{2}} \\ -(x_\eta)_{i+\frac{1}{2},j}\beta_{i+\frac{1}{2},j} - (x_\eta)_{i-\frac{1}{2},j}\beta_{i-\frac{1}{2},j} + (x_\xi)_{i,j+\frac{1}{2}}\beta_{i,j+\frac{1}{2}} - (x_\xi)_{i,j-\frac{1}{2}}\beta_{i,j-\frac{1}{2}} \\ 0 \end{pmatrix} \\ &= \left(A_{i+\frac{1}{2},j}\boldsymbol{F}_{i+\frac{1}{2},j}^{\text{hs}} - A_{i-\frac{1}{2},j}\boldsymbol{F}_{i-\frac{1}{2},j}^{\text{hs}} + A_{i,j+\frac{1}{2}}\boldsymbol{F}_{i,j+\frac{1}{2}}^{\text{hs}} - A_{i,j-\frac{1}{2}}\boldsymbol{F}_{i,j-\frac{1}{2}}^{\text{hs}}\right), \end{aligned} \tag{6.23}$$

i.e., $\frac{d}{dt}\hat{\boldsymbol{Q}}_{ij}^{\text{hs}} = 0$ from Eq. (5.75). □

6.2 The Deviation Method

The Deviation method can be extended to two or three spatial dimensions. In Section 6.2.1 we present the two-dimensional method. However, further extending it to three spatial dimensions is simple. The three-dimensional method has been introduced in [14]. Since the target solution is assumed to be given, the well-balanced property can hold for genuinely multi-dimensional methods, as we show in Section 6.2.2.2. The multi-dimensional Deviation method maintains its high order accuracy (Section 6.2.2.1). Different from the α-β method, the Deviation method uses

genuinely multi-dimensional reconstruction. Also, the scope of application is larger than the one from the α-β method, as already discussed in the one-dimensional case (Section 4.3.2.3).

6.2.1 Description of the Two-Dimensional Deviation Method

Consider the general two-dimensional system of hyperbolic balance laws (5.57). For the given smooth solution $\tilde{\boldsymbol{q}}$ of Eq. (5.57), which we call *target solution* in the following, it is

$$\partial_t \tilde{\boldsymbol{q}}(\boldsymbol{x},t) + \nabla \boldsymbol{f}(\tilde{\boldsymbol{q}}(\boldsymbol{x},t)) = \boldsymbol{s}(\tilde{\boldsymbol{q}}(\boldsymbol{x},t),\boldsymbol{x},t). \tag{6.24}$$

We subtract Eq. (6.24) from Eq. (5.57) and rewrite it in the form

$$\begin{aligned}\partial_t \Delta\boldsymbol{q}(\boldsymbol{x},t) + \nabla\cdot\left(\mathcal{F}\left(\tilde{\boldsymbol{q}}(\boldsymbol{x},t)+\Delta\boldsymbol{q}(\boldsymbol{x},t)\right) - \mathcal{F}\left(\tilde{\boldsymbol{q}}(\boldsymbol{x},t)\right)\right)& \\ = \boldsymbol{s}(\tilde{\boldsymbol{q}}(\boldsymbol{x},t)+\Delta\boldsymbol{q}(\boldsymbol{x},t),\boldsymbol{x},t) - \boldsymbol{s}(\tilde{\boldsymbol{q}}(\boldsymbol{x},t),\boldsymbol{x},t)&\end{aligned} \tag{6.25}$$

with the deviation

$$\Delta\boldsymbol{q} := \boldsymbol{q} - \tilde{\boldsymbol{q}} \tag{6.26}$$

from the target solution $\tilde{\boldsymbol{q}}$. From this point on we follow the construction of the general finite volume method in Section 5.8, but instead of Eq. (5.57) we discretize Eq. (6.25). Cell-averaging Eq. (6.25) over Ω_i[1] yields

$$\begin{aligned}\frac{d}{dt}(\Delta\hat{\boldsymbol{q}}_i(t)) = &-\frac{1}{V_i}\sum_{k\in\mathcal{N}_i}\int_{\partial\Omega_{ik}}\left(\mathcal{F}((\tilde{\boldsymbol{q}}+\Delta\boldsymbol{q})(\boldsymbol{x},t)) - \mathcal{F}(\tilde{\boldsymbol{q}}(\boldsymbol{x},t))\right)\cdot\boldsymbol{n}(\boldsymbol{x})d\sigma \\ &+\frac{1}{V_i}\int_{\Omega_i}\boldsymbol{s}((\tilde{\boldsymbol{q}}+\Delta\boldsymbol{q})(\boldsymbol{x},t),\boldsymbol{x},t) - \boldsymbol{s}(\tilde{\boldsymbol{q}}(\boldsymbol{x},t),\boldsymbol{x},t)d\boldsymbol{x},\end{aligned} \tag{6.27}$$

where we use the notations from Section 5.8 and $\Delta\hat{\boldsymbol{q}}_i := \hat{\boldsymbol{q}}_i - \hat{\tilde{\boldsymbol{q}}}_i$. Using numerical fluxes and quadrature yields the semi-discrete scheme

$$\begin{aligned}\frac{d}{dt}(\Delta\hat{\boldsymbol{Q}}_i(t)) = &-\frac{1}{V_i}\sum_{k\in\mathcal{N}_i} I_{\boldsymbol{x}\in\partial\Omega_{ij}}\left[\Delta\boldsymbol{F}_{\boldsymbol{n}}\left(\Delta\boldsymbol{Q}_i^{\mathrm{rec}}(\cdot,t),\Delta\boldsymbol{Q}_k^{\mathrm{rec}}(\cdot,t),\tilde{\boldsymbol{q}}(\cdot,t)\right)\right] \\ &+\frac{1}{V_i}I_{\boldsymbol{x}\in\Omega_i}\left[\boldsymbol{s}((\Delta\boldsymbol{Q}_i^{\mathrm{rec}}+\tilde{\boldsymbol{q}})(\cdot,t),\cdot,t)\right] - \frac{1}{V_i}I_{\boldsymbol{x}\in\Omega_i}\left[\boldsymbol{s}(\tilde{\boldsymbol{q}}(\cdot,t),\cdot,t)\right],\end{aligned} \tag{6.28}$$

where the reconstructed functions $\Delta\boldsymbol{Q}_i^{\mathrm{rec}}, \Delta\boldsymbol{Q}_k^{\mathrm{rec}}$ are obtained using a consistent conservative reconstruction routine on the cell average values $\Delta\hat{\boldsymbol{Q}}$ and the numerical flux difference $\Delta\boldsymbol{F}_{\boldsymbol{n}}$ is defined as

$$\Delta\boldsymbol{F}_{\boldsymbol{n}}\left(\Delta\boldsymbol{Q}^L,\Delta\boldsymbol{Q}^R,\tilde{\boldsymbol{q}}\right) := \boldsymbol{F}(\Delta\boldsymbol{Q}^L+\tilde{\boldsymbol{q}},\Delta\boldsymbol{Q}^R+\tilde{\boldsymbol{q}},\boldsymbol{n}) - \boldsymbol{n}\cdot\mathcal{F}(\tilde{\boldsymbol{q}}). \tag{6.29}$$

6.2.2 Properties of the Deviation Method

In the following we discuss the main properties of the two-dimensional Deviation method.

[1] Note that, different from the previous section, we only use one cell-index again. The reason is that this formulation also allows for unstructured grids.

6.2.2.1 Accuracy

The simple modification applied in the Deviation method in order to make a standard method well-balanced allows for high order accuracy.

Theorem 6.2.1. *The semi-discrete scheme Eq.* (6.28) *is consistent with Eq.* (6.25) *and m-th order accurate in space provided that the reconstruction routine and the quadrature rules are at least m-th order accurate in space.*

Proof. Let q be a smooth solution of Eq. (5.57), where we denote the deviations from the target state with $\Delta \boldsymbol{q} = \boldsymbol{q} - \tilde{\boldsymbol{q}}$. Since the reconstruction is m-th order accurate we have

$$\Delta \boldsymbol{Q}_i^{\mathrm{rec}}(\boldsymbol{x}) = \mathcal{R}_i\left(\boldsymbol{x}; \{\Delta \hat{\boldsymbol{q}}_j\}_{j\in\mathcal{S}_i}\right) = \Delta \boldsymbol{q}(\boldsymbol{x}) + \mathcal{O}\left(h^m\right) \qquad \text{for} \qquad x \in \Omega_i. \tag{6.30}$$

Since the numerical flux $\boldsymbol{F}_{\boldsymbol{n}}$ is Lipschitz continuous and consistent, the relation

$$\begin{aligned}
&\Delta \boldsymbol{F}_{\boldsymbol{n}}\left(\Delta \boldsymbol{Q}_i^{\mathrm{rec}}(\boldsymbol{x},t), \Delta \boldsymbol{Q}_k^{\mathrm{rec}}(\boldsymbol{x},t), \tilde{\boldsymbol{q}}(\boldsymbol{x},t)\right) \\
&\quad = \boldsymbol{F}\left(\Delta \boldsymbol{Q}_i^{\mathrm{rec}}(\boldsymbol{x},t) + \tilde{\boldsymbol{q}}(\boldsymbol{x},t), \Delta \boldsymbol{Q}_k^{\mathrm{rec}}(\boldsymbol{x},t) + \tilde{\boldsymbol{q}}(\boldsymbol{x},t), \boldsymbol{n}\right) - \boldsymbol{n}\cdot\mathcal{F}\left(\tilde{\boldsymbol{q}}(\boldsymbol{x},t)\right) \\
&= \boldsymbol{F}\left((\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}})(\boldsymbol{x},t) + \mathcal{O}\left(h^m\right), (\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}})(\boldsymbol{x},t) + \mathcal{O}\left(h^m\right), \boldsymbol{n}\right) - \boldsymbol{n}\cdot\mathcal{F}\left(\tilde{\boldsymbol{q}}(\boldsymbol{x},t)\right) \\
&\quad = \boldsymbol{n}\cdot\mathcal{F}\left((\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}})(\boldsymbol{x},t)\right) - \boldsymbol{n}\cdot\mathcal{F}\left(\tilde{\boldsymbol{q}}(\boldsymbol{x},t)\right) + \mathcal{O}\left(h^m\right)
\end{aligned} \tag{6.31}$$

holds for any $\boldsymbol{x} \in \partial\Omega_{ij}$. Hence, the interface flux approximation is also m-th order accurate

$$\begin{aligned}
&I_{\boldsymbol{x}\in\partial\Omega_{ij}}\left[\Delta \boldsymbol{F}_{\boldsymbol{n}}\left(\Delta \boldsymbol{Q}_i^{\mathrm{rec}}(\cdot,t), \Delta \boldsymbol{Q}_k^{\mathrm{rec}}(\cdot,t), \tilde{\boldsymbol{q}}(\cdot,t)\right)\right] \\
&\qquad = I_{\boldsymbol{x}\in\partial\Omega_{ij}}\left[\boldsymbol{n}\cdot\mathcal{F}\left((\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}})(\cdot,t)\right) - \boldsymbol{n}\cdot\mathcal{F}\left(\tilde{\boldsymbol{q}}(\cdot,t)\right) + \mathcal{O}\left(h^m\right)\right] \\
&\quad = \int_{\boldsymbol{x}\in\partial\Omega_{ij}} \boldsymbol{n}\cdot\mathcal{F}\left((\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}})(\boldsymbol{x},t)\right) - \boldsymbol{n}\cdot\mathcal{F}\left(\tilde{\boldsymbol{q}}(\boldsymbol{x},t)\right) + \mathcal{O}\left(h^m\right)\ d\boldsymbol{x} + \mathcal{O}\left(h^m\right) \\
&\qquad = \int_{\boldsymbol{x}\in\partial\Omega_{ij}} \boldsymbol{n}\cdot\mathcal{F}\left((\Delta \boldsymbol{q} + \tilde{\boldsymbol{q}})(\boldsymbol{x},t)\right) - \boldsymbol{n}\cdot\mathcal{F}\left(\tilde{\boldsymbol{q}}(\boldsymbol{x},t)\right)\ d\boldsymbol{x} + \mathcal{O}\left(h^m\right).
\end{aligned} \tag{6.32}$$

Since also the source term is discretized using the m-th order reconstructed functions and an m-th order accurate quadrature rule, the method is m-th order accurate and consistent with Eq. (6.25). □

In this theorem we have shown that the Deviation method is consistent with the modified PDE (6.25). Since this modified PDE is equivalent to the original hyperbolic balance law Eq. (5.57), the following corollary holds.

Corollary 6.2.2 (Accuracy of the deviation method)**.** *The semi-discrete scheme Eq.* (6.28)*, interpreted as a method to evolve $\hat{\boldsymbol{Q}}_i = \hat{\tilde{\boldsymbol{q}}}_i + \Delta\hat{\boldsymbol{Q}}_i$, is m-th order accurate in space and consistent with Eq.* (5.57)*.*

Proof. Corollary 6.2.2 follows directly from Theorem 6.2.1 and the relation $\frac{d}{dt}\Delta\hat{\boldsymbol{q}}_i = \frac{d}{dt}\hat{\boldsymbol{q}}_i - \frac{d}{dt}\hat{\tilde{\boldsymbol{q}}}_i$. □

6.2.2.2 Well-Balanced Property

In this section we show the well-balanced property of our method. The formulation of Theorem 6.2.3, its proof, and Corollary 6.2.4 have in a similar form been published in our original article [14].

Theorem 6.2.3. *The modified Runge–Kutta finite volume method introduced in Section 6.2.1 satisfies the following property: If*

$$\Delta \boldsymbol{Q}_i = 0 \qquad \forall i \in \{1, \dots, N\} \tag{6.33}$$

at initial time, then this holds for all $t > 0$.

Proof. Let $\Delta \boldsymbol{Q}_i = 0$ for all $i \in I$. The consistency of the applied reconstruction leads to $\Delta \boldsymbol{Q}_i^{\mathrm{rec}} \equiv 0$ at all flux quadrature points. The flux consistency then yields

$$\begin{aligned} \Delta \boldsymbol{F}_{\boldsymbol{n}} \left(\Delta \boldsymbol{Q}_i^{\mathrm{rec}}(\boldsymbol{x}), \Delta \boldsymbol{Q}_j^{\mathrm{rec}}(\boldsymbol{x}), \tilde{\boldsymbol{q}} \right) &= \Delta \boldsymbol{F}_{\boldsymbol{n}} \left(0, 0, \tilde{\boldsymbol{q}} \right) \\ &= \boldsymbol{F}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{q}}, \boldsymbol{n}) - \boldsymbol{n} \cdot \mathcal{F}(\tilde{\boldsymbol{q}}) = \boldsymbol{n} \cdot \mathcal{F}(\tilde{\boldsymbol{q}}) - \boldsymbol{n} \cdot \mathcal{F}(\tilde{\boldsymbol{q}}) = 0 \end{aligned} \tag{6.34}$$

for any quadrature point $\boldsymbol{x}$ on $\partial \Omega_{ij}$. Hence, discrete interface flux is

$$I_{\boldsymbol{x} \in \partial \Omega_{ij}} \left[\Delta \boldsymbol{F}_{\boldsymbol{n}} \left(\Delta \boldsymbol{Q}_i^{\mathrm{rec}}(\cdot, t), \Delta \boldsymbol{Q}_k^{\mathrm{rec}}(\cdot, t), \tilde{\boldsymbol{q}}(\cdot, t) \right) \right] = I_{\boldsymbol{x} \in \partial \Omega_{ij}} \left[0 \right] = 0. \tag{6.35}$$

Now, consider the contribution from the source term: With $\Delta \boldsymbol{Q}_i = 0$ the source term discretization in Eq. (6.28) reduces to

$$\begin{aligned} I_{\boldsymbol{x} \in \Omega_i} \left[\boldsymbol{s}((\Delta \boldsymbol{Q}_i^{\mathrm{rec}} + \tilde{\boldsymbol{q}})(\cdot, t), \cdot, t) \right] - I_{\boldsymbol{x} \in \Omega_i} \left[\boldsymbol{s}(\tilde{\boldsymbol{q}}(\cdot, t), \cdot, t) \right] \\ = I_{\boldsymbol{x} \in \Omega_i} \left[\boldsymbol{s}(\tilde{\boldsymbol{q}}(\cdot, t), \cdot, t) \right] - I_{\boldsymbol{x} \in \Omega_i} \left[\boldsymbol{s}(\tilde{\boldsymbol{q}}(\cdot, t), \cdot, t) \right] = 0. \end{aligned} \tag{6.36}$$

We have shown that the right hand side in Eq. (6.28) vanishes and thus the initial data $\Delta \boldsymbol{Q}_i = 0$ are conserved for all time. □

The formulation of the well-balanced property of the Deviation method in Theorem 6.2.3 is different from the one for the α-β method in Theorem 6.1.2. To make it comparable we reformulate the theorem in an obvious corollary.

Corollary 6.2.4. *If the initial condition $\boldsymbol{Q}_i(t = 0)$, $i = 1, \dots, N$, equals the cell averages of the target solution $\tilde{\boldsymbol{Q}}_i(t = 0)$, $i = 1, \dots, N$, the computed solution equals the target solution for all time.*

6.2.2.3 Scope

The properties and scope of the method which we discussed for the one-dimensional Deviation method in Section 4.3.2.3 can be transferred to the two-dimensional case. This includes...

- ... the possibility to apply the modification to any hyperbolic balance law.
- ... the capability to well-balance time-dependent solutions (Remark 4.3.5).

- ... the simplified implementation if only stationary solutions shall be balanced (Remark 4.3.6).
- ... the simplification for linear source terms (Remark 4.3.7).
- ... the applicability for homogeneous hyperbolic conservation laws (Remark 4.3.8).
- ... the well-balanced property for non-smooth target solutions (Remark 4.3.9).
- ... the possibility to reconstruct in a different set of variables instead of the conserved variables (Remark 4.3.10).

6.3 The Local Approximation Method

In order to define the local hydrostatic pressure approximation in the multi-dimensional extension of the Discretely Well-Balanced or Local Approximation method, the momentum source term approximation $\boldsymbol{S}^{\rho\boldsymbol{v},\mathrm{LA},ij}$ as defined below is integrated from one point to another. In general $\nabla \times \boldsymbol{S}^{\rho\boldsymbol{v},\mathrm{LA},ij} \neq 0$ holds in our construction and this integral is path dependent. It is hence not possible to find a continuous discrete approximation of a multi-dimensional hydrostatic pressure using local pressure polynomials, which would be required in order for a theorem similar to Theorem 4.4.4 to hold.

Since the Local Approximation method shows even better convergence than the Discretely Well-Balanced method in the one-dimensional tests in Section 4.6 and does not suffer from an increased stencil (Section 4.4.2.4), it seems preferable over the Discretely Well-Balanced method for a multi-dimensional method. We develop the method only on a Cartesian mesh, further extension to arbitrary grids can be conducted in future work.

6.3.1 Description of the Local Approximation Method

The two-dimensional Local Approximation method is a natural extension of the one-dimensional method introduced in Section 4.5. The description of the method is similar – and in parts identical – to the description in our original article [12].

6.3.1.1 Source Term Discretization

Let us define the source term approximation

$$\boldsymbol{S}^{\rho\boldsymbol{v},\mathrm{LA},ij}(\boldsymbol{x}) := \begin{pmatrix} S^{\rho u,\mathrm{LA},ij}(\boldsymbol{x}) \\ S^{\rho v,\mathrm{LA},ij}(\boldsymbol{x}) \end{pmatrix} := \begin{pmatrix} \rho_{ij}^{\mathrm{rec}}(\boldsymbol{x})(g_x)_{ij}^{\mathrm{int}}(\boldsymbol{x}) \\ \rho_{ij}^{\mathrm{rec}}(\boldsymbol{x})(g_y)_{ij}^{\mathrm{int}}(\boldsymbol{x}) \end{pmatrix}, \tag{6.37}$$

$$S^{E,\mathrm{LA},ij}(x) := (\rho u)_{ij}^{\mathrm{rec}}(\boldsymbol{x})(g_x)_{ij}^{\mathrm{int}}(\boldsymbol{x}) + (\rho v)_{ij}^{\mathrm{rec}}(\boldsymbol{x})(g_y)_{ij}^{\mathrm{int}}(\boldsymbol{x}), \tag{6.38}$$

where ρ_{ij}^{rec} and $(\rho\boldsymbol{v})_{ij}^{\mathrm{rec}} := \left((\rho u)_{ij}^{\mathrm{rec}}, (\rho v)_{ij}^{\mathrm{rec}}\right)^T$ are m-th order accurate CWENO reconstruction polynomials in the ij-th cell. $\boldsymbol{g}_{ij}^{\mathrm{int}}$ is an m-the order accurate interpolation polynomial from the cell-centered point values of $\boldsymbol{g}$. CWENO interpolation can be used if $\boldsymbol{g}$ is not smooth. Note, that the statement from Remark 4.4.1 is also valid

in the two-dimensional case. Due to the polynomial character of $S^{\rho u,\mathrm{LA},ij}$, $S^{\rho v,\mathrm{LA},ij}$, and $S^{E,\mathrm{LA},ij}$ the source term integrals can be computed explicitly. The cell-averaged source term used in the finite volume method in the i-th cell is hence computed as

$$\hat{\boldsymbol{S}}_{ij}^{\mathrm{LA}} := \frac{1}{|\Omega_{ij}|} \int_{\Omega_{ij}} \begin{pmatrix} 0 \\ S^{\rho u,\mathrm{LA},ij}(\boldsymbol{x}) \\ S^{\rho v,\mathrm{LA},ij}(\boldsymbol{x}) \\ S^{E,\mathrm{LA},ij}(\boldsymbol{x}) \end{pmatrix} d\boldsymbol{x}. \tag{6.39}$$

6.3.1.2 Reconstruction

As in the one-dimensional method, prior to the reconstruction itself we construct a local approximation to the hydrostatic pressure in the cell Ω_{ij}. Using the local hydrostatic density $\rho_{ij}^{\mathrm{hs}} := \rho_{ij}^{\mathrm{rec}}$ we obtain the local hydrostatic pressure in the cell Ω_{ij} by

$$p_{ij}^{\mathrm{hs}}(\boldsymbol{x}) := p_{0,ij} + \int_0^1 \boldsymbol{S}^{\boldsymbol{\rho v},\mathrm{LA},ij}(\boldsymbol{x}_{ij} + (\boldsymbol{x} - \boldsymbol{x}_{ij})t) \cdot (\boldsymbol{x} - \boldsymbol{x}_{ij})\, dt. \tag{6.40}$$

To obtain the cell-centered pressure value $p_{0,ij}$ we solve the equation

$$\hat{\varepsilon}_{ij}^{\mathrm{est}} = \frac{1}{|\Omega_{ij}|} I_{\boldsymbol{x}\in\Omega_{ij}} \left[\varepsilon_{\mathrm{EoS}} \left(\rho_{ij}^{\mathrm{hs}}(\boldsymbol{x}), p_{ij}^{\mathrm{hs}}(\boldsymbol{x}) \right) \right], \tag{6.41}$$

where the cell-averaged internal energy density is estimated by

$$\hat{\varepsilon}_{ij}^{\mathrm{est}} := \hat{E}_{ij} - \frac{\left(\hat{\rho u}\right)_{ij}^2 + \left(\hat{\rho v}\right)_{ij}^2}{2\hat{\rho}_{ij}}. \tag{6.42}$$

On hydrostatic solutions, it is $\hat{\varepsilon}_{ij}^{\mathrm{est}} = \hat{\varepsilon}_{ij}$, since the momentum term vanishes in that case. As in the one-dimensional case, Eq. (6.41) can be solved explicitly if an ideal gas EoS is used to close the Euler system. For general EoS we refer to Section 4.4.1.3.

Now that the pressure at cell center $p_{0,ij}$ is fixed, we have fully specified the high-order accurate representation of the equilibrium conserved variables in cell Ω_{ij}:

$$\boldsymbol{Q}_{ij}^{\mathrm{hs}}(\boldsymbol{x}) = \begin{pmatrix} \rho_{ij}^{\mathrm{hs}}(\boldsymbol{x}) \\ 0 \\ 0 \\ \varepsilon_{ij}^{\mathrm{hs}}(\boldsymbol{x}) \end{pmatrix}, \tag{6.43}$$

where

$$\varepsilon_{ij}^{\mathrm{hs}}(\boldsymbol{x}) = \varepsilon \left(\rho_{ij}^{\mathrm{hs}}(\boldsymbol{x}), p_{ij}^{\mathrm{hs}}(\boldsymbol{x}) \right). \tag{6.44}$$

Next, we develop the high-order equilibrium preserving reconstruction procedure. To this end, as in e.g. [15, 78], we decompose in every cell the solution into an equilibrium and a (possibly large) perturbation part. The equilibrium part in cell Ω_{ij} is simply given by $\boldsymbol{Q}_{ij}^{\mathrm{hs}}(x)$ of Eq. (6.43) above. The perturbation part in cell Ω_{ij} is obtained by applying the standard reconstruction procedure $\mathcal{R}$ to the cell-averaged equilibrium perturbation

$$\delta\boldsymbol{Q}_{ij}^{\mathrm{rec}}(\boldsymbol{x}) = \mathcal{R}\left(\boldsymbol{x}; \left\{ \hat{\boldsymbol{Q}}_{kl} - \frac{1}{|\Omega_{kl}|} I_{\boldsymbol{x}\in\Omega_{kl}} \left[\hat{\boldsymbol{Q}}_{ij}^{\mathrm{hs}}(\boldsymbol{x}) \right] \right\}_{(k,l)\in\mathcal{S}_{ij}} \right). \tag{6.45}$$

We note that the cell average of the equilibrium perturbation in cell Ω_{kl} is obtained by taking the difference between the cell average $\hat{\boldsymbol{Q}}_{kl}$ and the cell average of the equilibrium $\hat{\boldsymbol{Q}}^{\mathrm{hs}}_{ij}$ in cell Ω_{kl}.

The complete equilibrium preserving reconstruction $\mathcal{R}^{\mathrm{LA}}$ is then obtained by the sum of the equilibrium and perturbation reconstruction

$$\boldsymbol{Q}^{\mathrm{rec}}_{ij}(\boldsymbol{x}) = \mathcal{R}^{\mathrm{LA}}\left(\boldsymbol{x}; \left\{\hat{\boldsymbol{Q}}_{kl}\right\}_{(k,l)\in S_{ij}}\right) := \boldsymbol{Q}^{\mathrm{hs}}_{ij}(\boldsymbol{x}) + \delta\boldsymbol{Q}^{\mathrm{rec}}_{ij}(\boldsymbol{x}). \tag{6.46}$$

Remark 6.3.1. *The approximate well-balanced method presented in this section can be extended to three spatial dimensions without further complications.*

6.3.2 Properties of the Local Approximation Method

The fundamental properties of the two-dimensional Local Approximation method are similar to the one of the one-dimensional method, which are discussed in Section 4.5.2

6.3.2.1 Accuracy

The well-balancing modification introduced in Section 6.3.1 does not diminish the order of accuracy of the finite volume method.

Theorem 6.3.2. *Consider the semi-discrete scheme Eq.* (5.62) *on a Cartesian grid for compressible Euler equations with gravity* (5.5) *with a numerical flux* $\boldsymbol{F}$*, the hydrostatic reconstruction* $\mathcal{R}^{\mathrm{LA}}$ *(Eq.* (6.46)*) based on an* m*-th order accurate spatial reconstruction procedure* $\mathcal{R}$*, and the gravitational source term discretization* $\hat{\boldsymbol{S}}^{\mathrm{LA}}_{ij}$ *given in Eq.* (6.39)*. The scheme is consistent and at least* m*-th order accurate in space for smooth solutions.*

Proof. The argumentation is the same as in the one-dimensional case (proof of Theorem 4.5.1): The local pressure approximation is smooth, and hence the m-th order accurate reconstruction of the deviations with respect to Eq. (6.43) yields an m-th order accurate reconstruction of the conserved variables. The source term discretization is m-th order accurate by construction. □

6.3.2.2 Well-Balanced Property

As in the one-dimensional case, it is not possible to show a theorem similar to Theorem 4.4.4 for the Local Approximation method. This has been explained at the beginning of Section 6.3. Hence, we have to validate the increased accuracy of this method close to hydrostatic solutions in numerical experiments.

6.4 Numerical Tests

In the two-dimensional numerical experiments, we use the same Python code as for the tests in Section 4.6. The only exception is the experiment in Section 6.4.7: There we apply the astrophysical finite volume code SLH, which is for example described in

[120, 57, 143]. Both, SLH and the Python code, use the same spatial discretization technique for curvilinear grids that is described in Section 5.9. A third order spatial discretization on a Cartesian grid is achieved as described in Section 5.8 in the Python code. The source term discretization (5.56) is applied in the third order method and the one given in Eq. (3.76) for first and second order methods. Of course, the source term discretization is modified in case of the α-β method according to the description given in Section 6.1. The numerical fluxes and time-stepping routines are chosen as in Section 4.6 if not stated explicitly.

6.4.1 Two-Dimensional Polytrope

In the first two-dimensional test in this thesis, we apply our well-balanced methods on a two-dimensional polytrope. The two-dimensional polytrope is a hydrostatic stratification given by [78]

$$\tilde{\rho}(\boldsymbol{x}) := \frac{\sin\left(\sqrt{2\pi}\,|\boldsymbol{x}|\right)}{\sqrt{2\pi}\,|\boldsymbol{x}|}, \quad \tilde{p}(\boldsymbol{x}) := \tilde{\rho}(\boldsymbol{x})^{\gamma}, \quad \boldsymbol{g} := \nabla\phi(\boldsymbol{x}), \quad \phi(\boldsymbol{x}) := -2\frac{\sin\left(\sqrt{2\pi}\,|\boldsymbol{x}|\right)}{\sqrt{2\pi}\,|\boldsymbol{x}|} \tag{6.47}$$

with $\gamma = 2$ (also in the ideal gas EoS that is applied) and describes an adiabatic gaseous sphere held together by self-gravitation. The functions ρ and ϕ are extended to $\boldsymbol{x} = 0$ continuously. This configuration has been used in [12] to test the Local Approximation method. It is set on Cartesian, polar, and cubed sphere meshes discretizing the domain $[-0.5, 0.5]^2$. In the case of the polar and cubed sphere grid, the domain is reduced accordingly. We use our first and second accurate standard, α-β, and Deviation method. On the Cartesian grid we also use the third order accurate standard, Deviation, and Local Approximation method. We evolve the polytrope up to the final time $t = 10\tau$. The L^1-errors for the exact well-balanced methods (α-β and Deviation) are shown in Table 6.1 for simulations on a 64×64 cells grid. All the errors are on machine precision for the α-β method. For the Deviation method there is no error at all. In Table 6.2 the L^1-errors and convergence rates are presented for the third order accurate standard and Local Approximation method. Using the Local Approximation method significantly improves the result compared to the standard method. Moreover, an increased order of accuracy is observed for the Local Approximation method, similar to our observations in the one-dimensional numerical experiments.

6.4.1.1 Perturbation on Two-Dimensional Polytrope

In order to study the accuracy of our two-dimensional numerical methods on a perturbation from the hydrostatic state, we add a perturbation to the polytrope introduced in Section 6.4.1. The initial pressure is perturbed in the following way [78, 12]

$$p_{\text{pert}}(\boldsymbol{x}) := \left(1 + A\,\exp\left(-\frac{\|\boldsymbol{x}\|_2^2}{0.05^2}\right)\right)\tilde{p}(\boldsymbol{x}). \tag{6.48}$$

As in [78, 12], we use the perturbation amplitude $A = 10^{-8}$. The spatial domain and numerical methods are the same as in Section 6.4.1. The only difference is that

Table 6.1: L^1-errors for the 2-d polytrope test case described in Section 6.4.1 after ten sound crossing times. The exact well-balanced methods have been used on a grid with 64×64 cells. Due to the symmetry of the test problem and grids the momentum errors are the same for ρu and ρv.

first order methods

method	mesh	ρ error	$\rho u/\rho v$ error	E error
Standard	Cartesian	7.17e-02	4.26e-03	3.21e-02
	Polar	1.95e-02	2.48e-03	7.15e-03
	Cubed sphere	6.36e-02	2.86e-03	2.63e-02
α-β WB	Cartesian	0.00e+00	0.00e+00	0.00e+00
	Polar	6.33e-15	4.14e-15	8.38e-16
	Cubed sphere	3.92e-15	7.69e-16	6.57e-16
Deviation WB	Cartesian	0.00e+00	0.00e+00	0.00e+00
	Polar	0.00e+00	0.00e+00	0.00e+00
	Cubed sphere	0.00e+00	0.00e+00	0.00e+00

second order methods

method	mesh	ρ error	$\rho u/\rho v$ error	E error
Standard	Cartesian	1.17e-03	4.57e-04	3.86e-04
	Polar	1.44e-04	2.17e-04	2.43e-04
	Cubed sphere	1.83e-03	1.22e-03	6.64e-04
α-β WB	Cartesian	5.01e-16	5.59e-16	8.51e-16
	Polar	4.90e-16	5.19e-15	1.59e-15
	Cubed sphere	5.51e-16	1.20e-15	1.40e-15
Deviation WB	Cartesian	0.00e+00	0.00e+00	0.00e+00
	Polar	0.00e+00	0.00e+00	0.00e+00
	Cubed sphere	0.00e+00	0.00e+00	0.00e+00

third order methods

method	mesh	ρ error	$\rho u/\rho v$ error	E error
Standard	Cartesian	3.55e-05	5.62e-06	1.57e-05
Deviation WB	Cartesian	0.00e+00	0.00e+00	0.00e+00

Table 6.2: L^1-errors and convergence rates for the 2-d polytrope test case described in Section 6.4.1 after ten sound crossing times. The third order accurate standard and Local Approximation methods have been used on different grid sizes. Due to the symmetry of the test problem the momentum errors and rates are the same for ρu and ρv.

third order methods (Cartesian mesh)

method	N	ρ error	ρ rate	$\rho u/\rho v$ error	$\rho u/\rho v$ rate	E error	E rate
Standard	16	2.12e-03	–	2.12e-04	–	9.83e-04	–
	32	2.79e-04	2.9	4.15e-05	2.4	1.26e-04	3.0
	64	3.55e-05	3.0	5.62e-06	2.9	1.57e-05	3.0
	128	4.47e-06	3.0	6.64e-07	3.1	1.96e-06	3.0
LA	16	9.33e-06	–	9.36e-07	–	4.80e-06	–
	32	2.99e-07	5.0	3.24e-08	4.9	1.57e-07	4.9
	64	9.54e-09	5.0	1.22e-09	4.7	5.19e-09	4.9
	128	3.08e-10	5.0	4.61e-11	4.7	1.80e-10	4.8

the polar grid is not used, since the initial perturbation is defined at the center of the domain, which cannot be included in a polar mesh. For all the simulations we use resolution of 128×128 cells. The final time is reduced to $t = 0.25\tau$, such that the perturbation cannot reach the boundary. As reference solutions to compare the results to we use simulations obtained with the third order accurate Deviation method on a 256×256 Cartesian mesh and the second order accurate Deviation method on a 256×256 cubed sphere mesh. Usually, one would prefer to use a standard method to provide a reference solution. However, the resolution necessary to provide a sufficiently accurate reference solution with the non-well-balanced standard method cannot be easily used due to a limit in a computer's RAM. The Deviation method has been analytically shown to be consistent and the accuracy has been verified in several tests in [14]. This allows us to use it as a method to obtain a reference solution on a sufficiently fine grid. The pressure deviations from the hydrostatic background at final time are visualized in Figs. 6.1 to 6.3 for the formally first, second, and third order accurate standard and well-balanced methods. While the standard method fails to resolve the perturbation in each of the tests due to a dominant error on the hydrostatic background, all well-balanced methods capture the perturbation accurately. The first and second order well-balanced methods are, as expected, more diffusive than the reference solution. The third order accurate Deviation method is in perfect agreement with the reference solution, while the Local Approximation method still shows a small error due to not being exactly well-balanced. However, the Local Approximation method's error is three orders of magnitude smaller than the standard method's error.

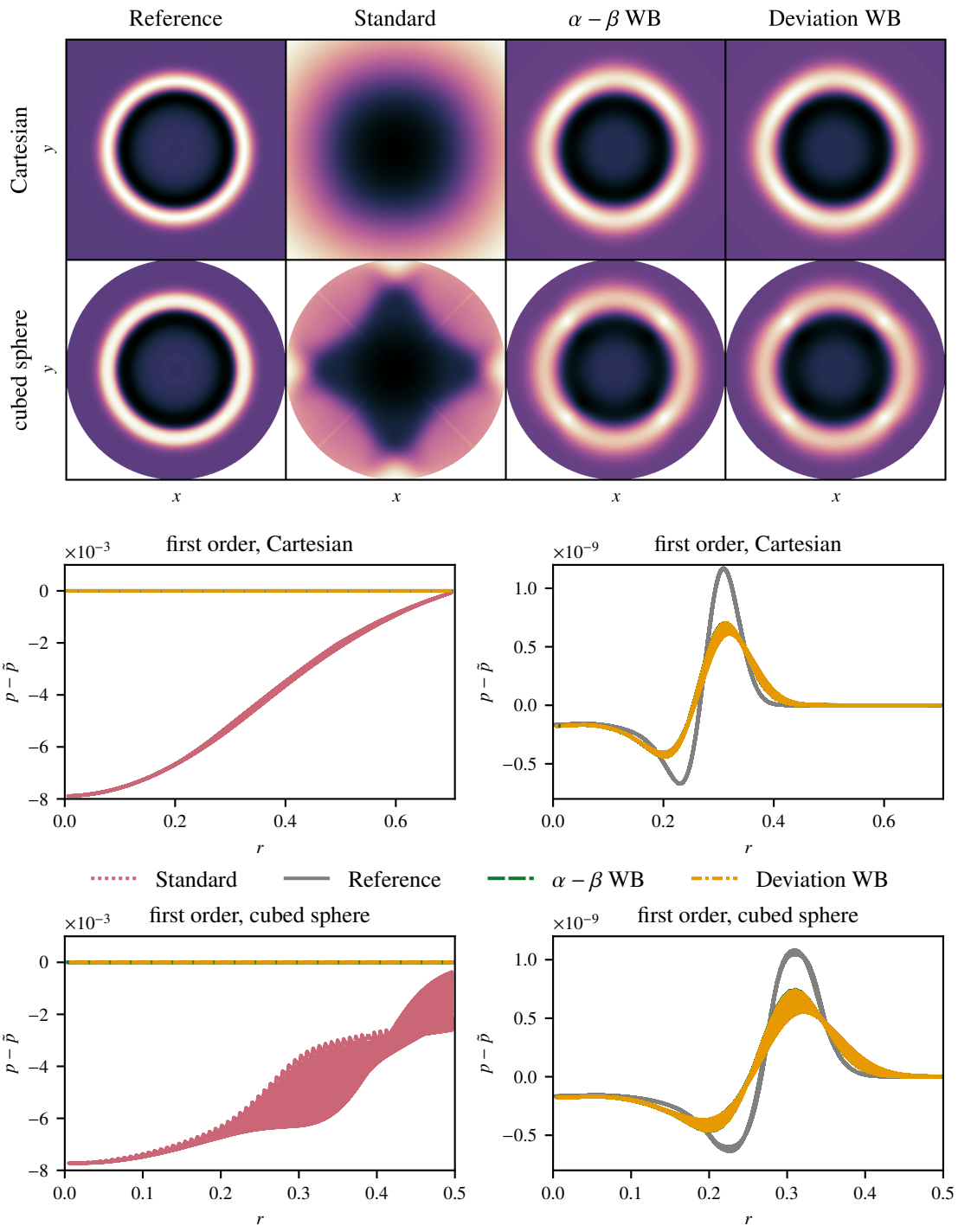


Figure 6.1: Perturbation on the two-dimensional polytrope as described in Section 6.4.1.1 after 0.25 sound crossing times using the different formally first order accurate methods on a Cartesian and cubed sphere grid with 128×128 cells. Top: Density plots of the pressure deviations from the hydrostatic background (Different grids in different rows, different methods in different columns. The x and y-values range from -0.5 to 0.5. Ranges of the colormaps for the standard methods are -8e-3 to 1e-3. For all other methods the range is from -8e-10 to 1.2e-9.). The four bottom panels show scatter plots of the same quantity over the radius. The line for the α-β method is hidden behind the line for the Deviation method.

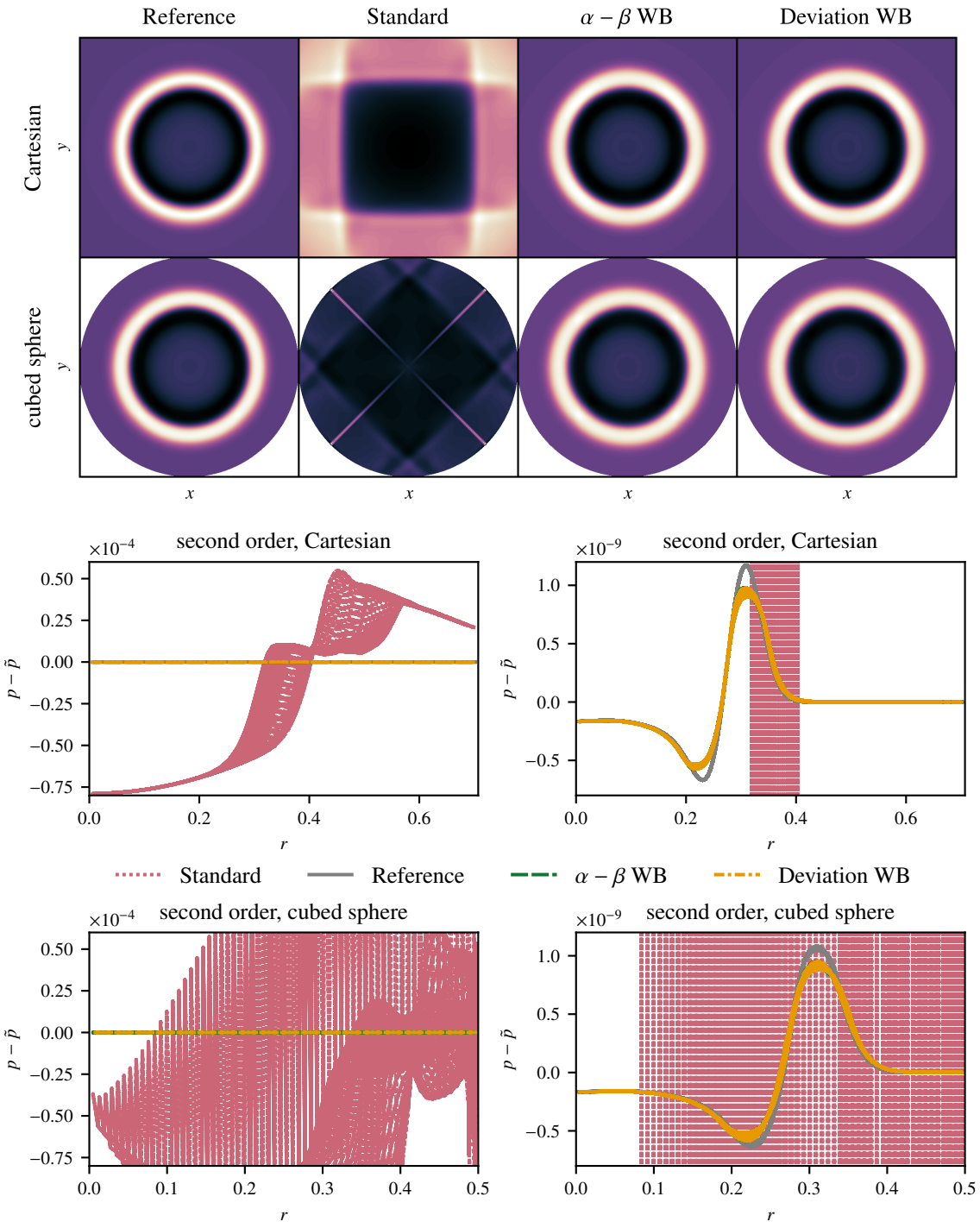


Figure 6.2: Perturbation on the two-dimensional polytrope as described in Section 6.4.1.1 after 0.25 sound crossing times using the different formally second order accurate methods on a Cartesian and cubed sphere grid with 128×128 cells. Top: Density plots of the pressure deviations from the hydrostatic background (Different grids in different rows, different methods in different columns. The x and y-values range from -0.5 to 0.5. Ranges of the colormaps for the standard methods are -8e-5 to 6e-5 for the Cartesian grid and -1.5e-4 to 6e-4 for the cubed sphere grid. For all other methods the range is from -8e-10 to 1.2e-9.). The four bottom panels show scatter plots of the same quantity over the radius. The line for the α-β method is hidden behind the line for the Deviation method.

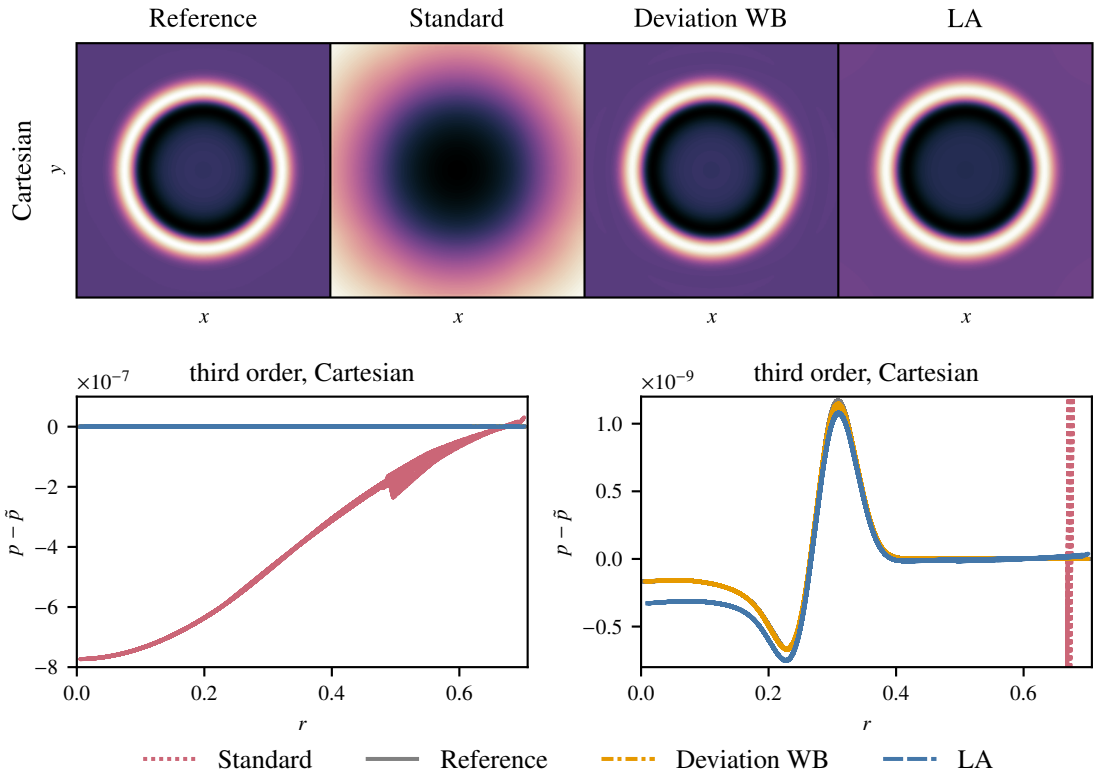


Figure 6.3: Perturbation on the two-dimensional polytrope as described in Section 6.4.1.1 after 0.25 sound crossing times using the different formally third order accurate methods on a Cartesian grid with 128×128 cells. Top: Density plots of the pressure deviations from the hydrostatic background (different methods in different panels. The x and y-values range from -0.5 to 0.5. Ranges of the colormaps for the standard methods are -8e-7 to 1e-7. For all other methods the range is from -8e-10 to 1.2e-9.). The two bottom panels show scatter plots of the same quantity over the radius. The line for the reference solution is hidden behind the line for the Deviation method.

6.4.2 Radial Rayleigh–Taylor Instabilities

In this test, we use a piecewise isothermal hydrostatic state in the two-dimensional gravitational potential

$$\phi(\boldsymbol{x}) := -20\frac{\sin\left(\sqrt{2\pi}\,|\boldsymbol{x}|\right)}{\sqrt{2\pi}\,|\boldsymbol{x}|} \tag{6.49}$$

and the gravitational acceleration $\boldsymbol{g} = -\nabla\phi(\boldsymbol{x})$. The initial data are given by

$$\rho(\boldsymbol{x}) := \begin{cases} \tilde{\rho}_{\text{in}}(\boldsymbol{x}) & \text{if } r(\boldsymbol{x}) < (1+\eta\cos(20\varphi(\boldsymbol{x})))r_0, \\ \tilde{\rho}_{\text{out}}(\boldsymbol{x}) & \text{else,} \end{cases} \tag{6.50}$$

$$p(\boldsymbol{x}) := \begin{cases} \tilde{p}_{\text{in}}(\boldsymbol{x}) & \text{if } r(\boldsymbol{x}) < r_0, \\ \tilde{p}_{\text{out}}(\boldsymbol{x}) & \text{else,} \end{cases} \tag{6.51}$$

$$\boldsymbol{v}(\boldsymbol{x}) := \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{6.52}$$

in the domain $\Omega = [0, 0.5]^2$, where

$$\tilde{\rho}_{\text{in}}(\boldsymbol{x}) := ac\exp\left(-a\phi(\boldsymbol{x})\right), \qquad \tilde{\rho}_{\text{out}}(\boldsymbol{x}) := b\exp\left(-b\phi(\boldsymbol{x})\right), \tag{6.53}$$

$$\tilde{p}_{\text{in}}(\boldsymbol{x}) := c\exp\left(-a\phi(\boldsymbol{x})\right), \qquad \tilde{p}_{\text{out}}(\boldsymbol{x}) := \exp\left(-b\phi(\boldsymbol{x})\right), \tag{6.54}$$

$c = \exp\left((a-b)\phi\left((r_0, 0)^T\right)\right)$, and $r(\boldsymbol{x}), \varphi(\boldsymbol{x})$ are radius and angle of $\boldsymbol{x}$ with respect to $(0, 0)$ and the x-axis. With this choice of c the pressure is continuous, whereas there is a discontinuity in the density (for $a \neq b$). Choosing $b > a$ means that there is a dense fluid on top[2] of a light fluid. Hence, Rayleigh–Taylor instabilities are expected to develop [39]. To break the radial symmetry of the setup for the instabilities to develop we set the parameter $\eta = 0.005$. In a similar configuration, this test case has been applied in [12] to test the Local Approximation method and the description of the test case is in parts similar to the one in this article.

For the parameters in the setup we choose $r_0 = 0.2$ and $(a, b) = (1, 2)$ and evolve the solution up to the final time $t = 0.6$ using different methods on different grids. At the $x = 0$ and $y = 0$ boundaries we use wall-boundary conditions, since these are consistent with the symmetry of the problem. At the outer boundaries we extrapolate $(\rho - \tilde{\rho}_{\text{out}},\ \rho u,\ \rho v,\ E - (\gamma-1)\tilde{p}_{\text{out}})^T$ in order to properly treat the hydrostatic solution at the boundary. For the α-β method we treat the outer boundaries in a slightly different way: We extrapolate the hydrostatic variables $(\rho/\alpha, u, v, p/\beta)$, where we set $\alpha = \tilde{\rho}_{\text{out}}, \beta = \tilde{p}_{\text{out}}$ for the well-balancing routine. In the Deviation method we choose $(\tilde{\rho}_{\text{out}}, 0, 0, (\gamma-1)\tilde{p}_{\text{out}})$ as target solution in conserved variables. In the second order standard and Deviation method we reconstruct primitive variables (see Remark 4.3.10). The results for simulations with the third order accurate standard, Deviation, and Local Approximation method on a 128×128 Cartesian grid are visualized in Fig. 6.4. The simulation using the standard methods crashes at $t \approx 0.4682$ due to negative pressure. The simulations using the well-balanced methods reach the final time. As can be seen in Fig. 6.5, the well-balanced methods help to preserve the hydrostatic states in the regions that are free of dynamical mixing processes, whereas using the standard method leads to a significant error at the

[2] "up" means the direction opposing the direction of the gravitational acceleration

outer boundary, which causes the simulation to crash. The Local Approximation method seems to be significantly more diffusive than the Deviation method (see Fig. 6.4). In Fig. 6.6, results of simulations using the second order accurate α-β and Deviation methods on 128×128 cubed sphere and polar grids are presented. Both well-balanced methods succeed on these curvilinear meshes and resolve the mixing processes well.

6.4.3 Keplerian Disk

In this section we present a test case that has been published in a similar form in [14] and that does not include a hydrostatic solution. Instead, a non-static stationary solution – the Keplerian disk – is considered, in which gravity opposes centrifugal force[3]. To the author's knowledge, in literature there are only the well-balanced method proposed in [67] and the Deviation method capable of balancing the Keplerian disk solution. The scheme in [67] has been developed for this particular purpose. The Deviation method can balance this stationary state since it can balance any stationary state of a hyperbolic system. Consider the stationary solution ([67])

$$\tilde{\rho} \equiv 1, \quad \tilde{u}(\boldsymbol{x}) = -\sin(\varphi(\boldsymbol{x}))\sqrt{\frac{Gm_S}{r(\boldsymbol{x})}}, \quad \tilde{v}(\boldsymbol{x}) = \cos(\varphi(\boldsymbol{x}))\sqrt{\frac{Gm_S}{r(\boldsymbol{x})}}, \quad \tilde{p} \equiv 1 \tag{6.55}$$

of the compressible Euler equations (5.5) with the gravitational potential $\phi(\boldsymbol{x}) = -\frac{Gm_s}{r(\boldsymbol{x})}$ and $r(\boldsymbol{x}) = \sqrt{x^2+y^2}$, $\varphi(\boldsymbol{x}) = \arctan(\frac{y}{x})$, $G = m_s = 1$. We use the initial conditions

$$\rho(x,y) = \begin{cases} 2 & \text{if } (x+1.5)^2 + y^2 < 0.15^2 \\ \tilde{\rho} & \text{else} \end{cases} \tag{6.56}$$

and $u = \tilde{u}, v = \tilde{v}, p = \tilde{p}$ on the domain $\Omega = [-2,2] \times [-2,2]$. Note that there is a singularity in the velocities at $\boldsymbol{x} = 0$, which we avoid by choosing the numerical grids accordingly. The second order standard and Deviation methods are applied on a polar grid with 32×256 cells and a Cartesian grid with 128×128 cells. In the Cartesian grid we take away the center with $r(\boldsymbol{x}) < 1$ and use Dirichlet boundary conditions. This is achieved by setting the initial conditions on the whole grid, but never update the cells with $r(\boldsymbol{x}) < 1$. Hence, these cells are treated as ghost cells. We apply Dirichlet boundary conditions at all boundaries except the boundaries in φ-direction on the polar grid, which are treated with polar boundary conditions. Since there is a discontinuity in the initial setup, we apply the minmod slope limiter. The density at time $t = 2.5$ for each simulation and the exact solution is shown in Fig. 6.7. It gets evident that in the simulation using the standard method the spot of increased density falls into the inner boundary on both grids. The Deviation method has no error on the stationary solution such that the spot of increased density is advected on the correct radius. It is noteworthy that this result can even be obtained on a Cartesian grid, which is generally not well-suited to evolve spherically symmetric setups.

[3] Centrifugal force appears as a source term in the momentum equation, when the Euler equations are rewritten in polar coordinates (e.g. [67]).

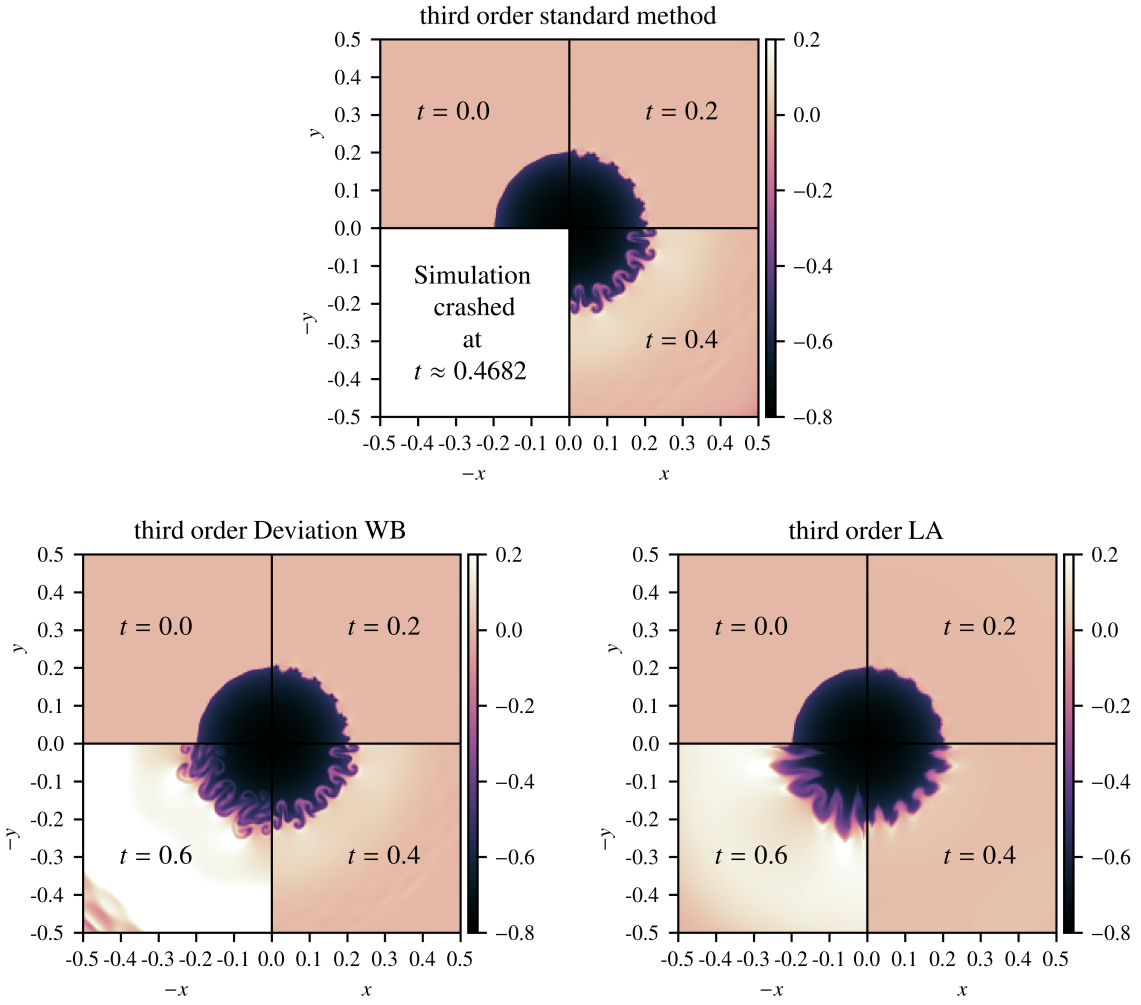


Figure 6.4: Snapshots of the relative density deviations $(\rho - \tilde{\rho}_{\text{out}})/\tilde{\rho}_{\text{out}}$ from the outer density stratification $\tilde{\rho}_{\text{out}}$ (Eq. (6.50)) at different times are shown for the radial Rayleigh–Taylor instability test case from Section 6.4.2. Different third order accurate methods are used in different panels on a 128×128 Cartesian grid. The simulation with the standard method crashes at $t \approx 0.4682$. The simulations with the Deviation and Local Approximation method reach the final time. The Local Approximation method is significantly more diffusive than the Deviation method.

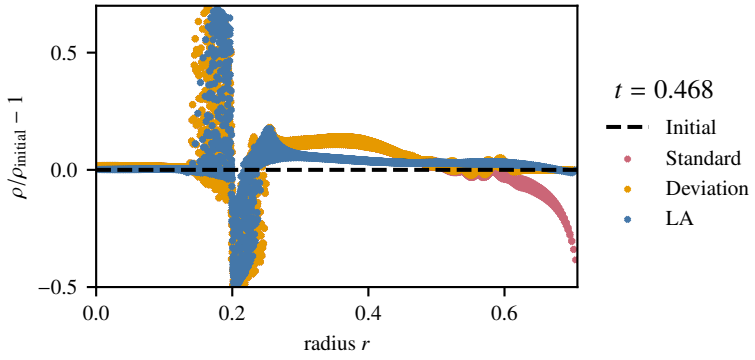


Figure 6.5: Scatter plot of the relative density deviations from the initial density over radius for the tests presented in Fig. 6.4 and discussed in Section 6.4.2 at time $t = 0.468$. This is shortly before the simulation with the standard method crashes. A significant error in the hydrostatic background gets evident at the outer boundary. The well-balanced methods lead to preservation of the initial hydrostatic state in the regions in which there are no mixing processes taking place.

6.4.4 Two-Dimensional Euler Wave in a Gravitational Field

The experiments presented in this section have been presented in our original research article [14] in a similar and in parts identical form. To demonstrate that we can follow time-dependent solutions exactly with the Deviation method, we use a problem from [175] and [42] which involves a known exact solution of the two-dimensional Euler equations with gravity given by

$$\tilde{\rho}(t,x,y) = 1 + \frac{1}{5}\sin(\pi(x+y-t(u_0+v_0))), \tag{6.57}$$

$$\tilde{u}(t,x,y) = u_0, \qquad \tilde{v}(t,x,y) = v_0, \tag{6.58}$$

$$\tilde{p}(t,x,y) = p_0 + t(u_0+v_0) - x - y + \frac{1}{5\pi}\cos(\pi(x+y-t(u_0+v_0))). \tag{6.59}$$

The gravitational potential is $\phi(\boldsymbol{x}) = x + y$, the EoS is the ideal gas EoS. In accordance to [175] and [42] we choose $u_0 = v_0 = 1$, $p_0 = 4.5$ on the domain $\Omega = [0,1]^2$. We use the first, second, and third order accurate Deviation method to evolve the initial data with $t = 0$ to a final time $t = 0.1$ on a 64×64 Cartesian grid and the second order Deviation method on a 64×64 polar grid. The L^1-error in every component of the state vector is exactly zero in each of the tests. We omit showing a table with these values since it does not provide additional insight.

Next, we are going to verify the order of accuracy for perturbations to time-dependent target solutions, if the Deviation method is used. For this we use the initial setup from Eqs. (6.57) to (6.59) and add a pressure perturbation:

$$\rho(t=0,x,y) = \tilde{\rho}(t=0,x,y), \tag{6.60}$$

$$u(t=0,x,y) = \tilde{u}(t=0,x,y), \quad v(t=0,x,y) = \tilde{v}(t=0,x,y), \tag{6.61}$$

$$p(t=0,x,y) = \tilde{p}(t=0,x,y) + \eta\exp\left(-100\left(\left(x-\frac{1}{2}\right)^2 + \left(y-\frac{1}{2}\right)^2\right)\right). \tag{6.62}$$

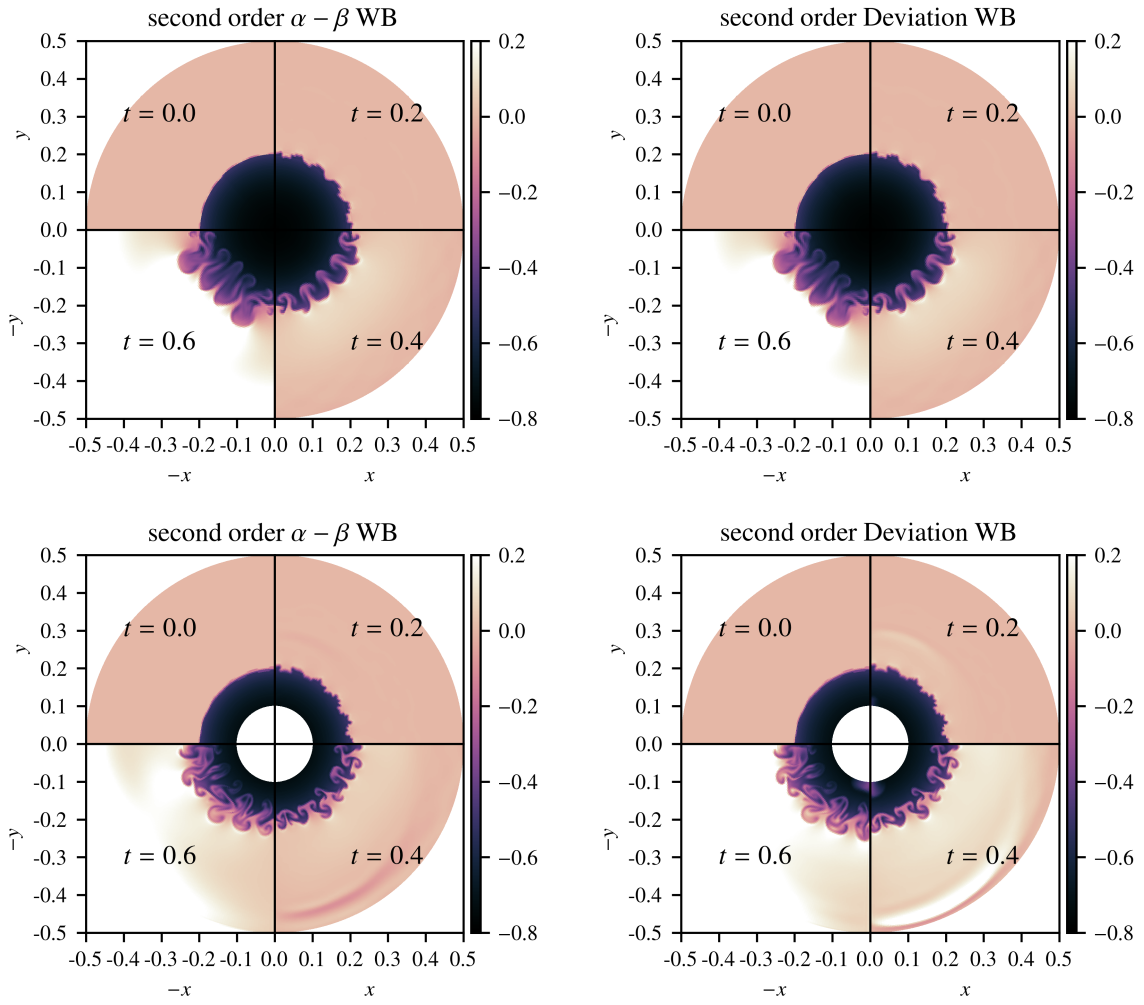


Figure 6.6: Snapshots of the relative density deviations $(\rho - \tilde{\rho}_{\text{out}})/\tilde{\rho}_{\text{out}}$ from the outer density stratification $\tilde{\rho}_{\text{out}}$ (Eq. (6.50)) at different times are shown for the radial Rayleigh–Taylor instability test case from Section 6.4.2. Different second order accurate methods are used in different columns (Left: α-β method. Right: Deviation method), different 128×128 grids are used in different rows (Top: Cubed sphere. Bottom: Polar).

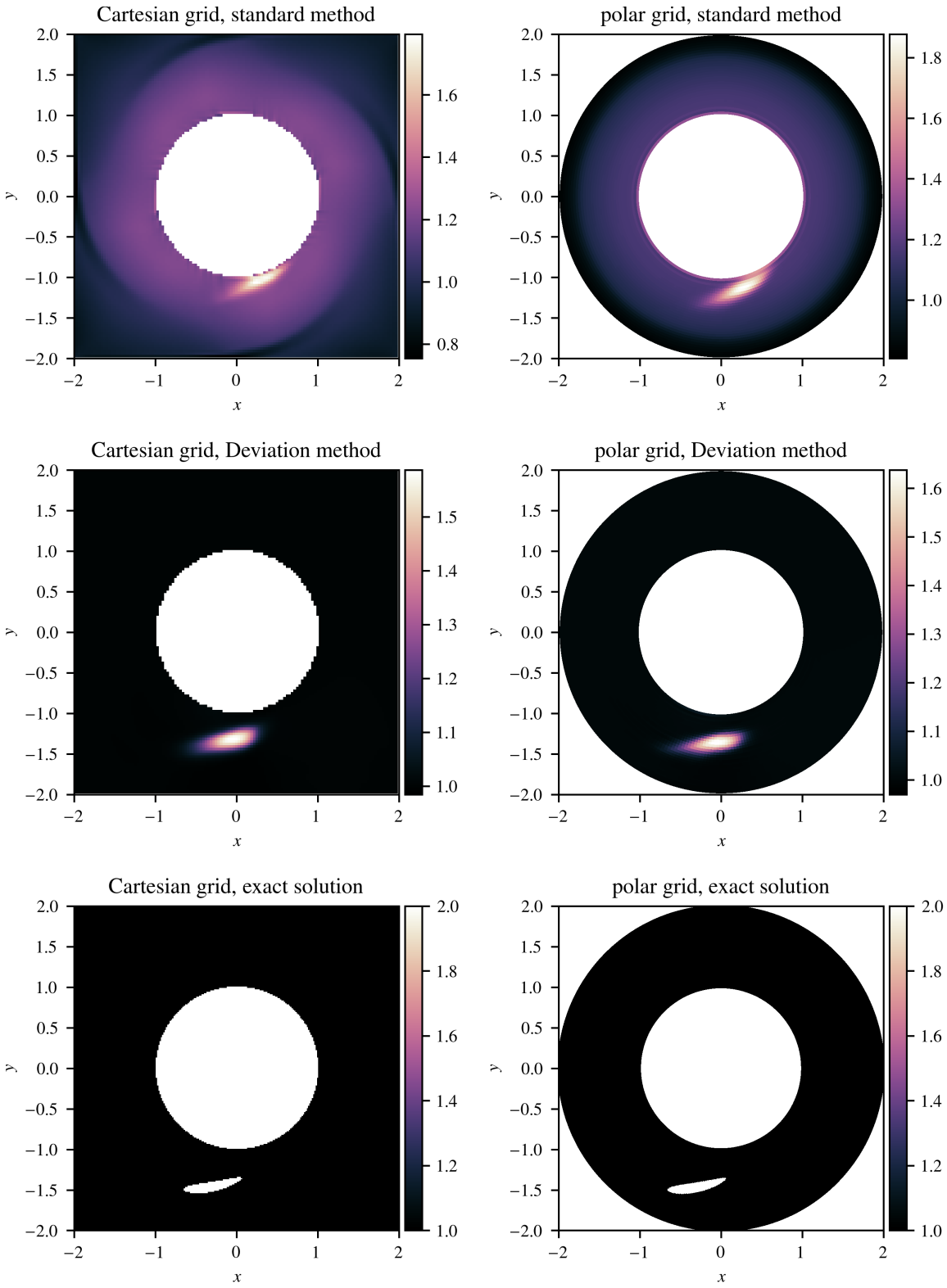


Figure 6.7: Density at final time for the Keplerian disk test case described in Section 6.4.3. A 128×128 Cartesian grid, from which the center has been cut out as described in Section 6.4.3, is used in the left panels, a 32×256 polar grid has been used in the right panels. The top row shows results from simulation with the second order accurate standard method, in the middle row the Deviation method has been used. The bottom row shows the exact solution on the different grids.

Table 6.3: L^1-errors and convergence rates for a pressure perturbation ($\eta = 0.1$) on the wave in a gravitational field solution of the two-dimensional Euler equations after time $t = 0.1$. The third order standard and Deviation method are used. The setup is described in Section 6.4.4.

third order standard method

N	ρ error	ρ rate	ρu error	ρu rate	ρv error	ρv rate	E error	E rate
64	3.77e-05	–	5.00e-05	–	5.00e-05	–	3.96e-04	–
128	5.61e-06	2.7	7.34e-06	2.8	7.34e-06	2.8	5.81e-05	2.8
256	7.18e-07	3.0	9.52e-07	2.9	9.52e-07	2.9	7.50e-06	3.0
512	8.03e-08	3.2	1.08e-07	3.1	1.08e-07	3.1	8.50e-07	3.1

third order Deviation method

N	ρ error	ρ rate	ρu error	ρu rate	ρv error	ρv rate	E error	E rate
64	2.65e-05	–	4.08e-05	–	4.08e-05	–	3.84e-04	–
128	4.17e-06	2.7	6.17e-06	2.7	6.17e-06	2.7	5.65e-05	2.8
256	5.41e-07	2.9	8.07e-07	2.9	8.07e-07	2.9	7.30e-06	3.0
512	6.26e-08	3.1	9.36e-08	3.1	9.36e-08	3.1	8.30e-07	3.1

Table 6.4: L^1-errors and convergence rates for a small pressure perturbation ($\eta = 10^{-5}$) on the wave in a gravitational field solution of the two-dimensional Euler equations after time $t = 0.1$. The third order Deviation method is used. The setup is described in Section 6.4.4.

Third order Deviation method

N	ρ error	ρ rate	ρu error	ρu rate	ρv error	ρv rate	E error	E rate
64	5.66e-09	–	1.07e-08	–	1.07e-08	–	9.38e-08	–
128	9.39e-10	2.6	1.73e-09	2.6	1.73e-09	2.6	1.49e-08	2.7
256	1.23e-10	2.9	2.27e-10	2.9	2.27e-10	2.9	1.93e-09	2.9
512	1.37e-11	3.2	2.56e-11	3.1	2.56e-11	3.1	2.17e-10	3.2

We evolve these initial data to time $t = 0.1$ using the third order standard and Deviation method with $\eta = 0.1$. The L^1 errors and corresponding convergence rates are presented in Table 6.3. As reference solution for determining the error we use a numerically approximated solution computed using the third order standard method on a 1024^2 grid. In this test we use exact boundary conditions for the standard method, which means that we evaluate the states in the ghost cells at any time from Eqs. (6.57) to (6.59). We see third order convergence for both methods. However, there seems to be no significant benefit from using the Deviation method in this test. The choice of $\eta = 0.1$ leads to a large discretization error in the perturbation which seems to dominate the total error. Choosing a large perturbation in this test was necessary since we use a solution computed from the standard method as a reference solution to compute the errors. For smaller perturbations, the standard method fails to provide a sufficiently accurate reference solution. To yet show the improved accuracy of the Deviation method, we add a convergence test with a small perturbation of $\eta = 10^{-5}$ for which a sufficiently accurate reference solution is produced using the third order Deviation method on a 1024^2 grid. The errors and convergence rates for the third order accurate Deviation method are shown in Table 6.4. It gets evident that the Deviation method is capable of accurately resolving small perturbations, since the errors are much smaller than the size of the perturbation – even on the coarsest grid.

6.4.5 Double Gresho Vortex

In this test, we use a vortex for homogeneous two-dimensional Euler equations which has first been introduced in [77]. The exact setup is taken from our original article [14] and the description below is similar to the one in this article. The pressure and the velocity in angular direction of this vortex in dependence of the distance r to the center are given by

$$\hat{u}(r) = \begin{cases} 5r, & \text{if } 0 \leq r < 0.2, \\ 2 - 5r, & \text{if } 0.2 \leq r < 0.4, \\ 0, & \text{if } 0.4 \leq r, \end{cases} \tag{6.63}$$

$$\hat{p}(r) = \begin{cases} 5 + \frac{25}{2} r^2, & \text{if } 0 \leq r < 0.2, \\ 9 - 4\ln(0.2) + \frac{25}{2} r^2 - 20r + 4\ln(r), & \text{if } 0.2 \leq r < 0.4, \\ 3 + 4\ln(2), & \text{if } 0.4 \leq r. \end{cases} \tag{6.64}$$

The radial velocity is zero and the density is $\rho \equiv 1$. In our test we set up the domain $[0, 1] \times [0, 2]$ with two Gresho vortices centered at $(0.5, 0.5)$ and $(0.5, 1.5)$ respectively. The vortices are advected with the velocity $\boldsymbol{v}_0 = (u_0, v_0)^T = (0.2, 0.4)^T$ and the boundaries are periodic. At time $t = 5$, the exact solution of this initial data equals the initial setup. We apply the formally second order accurate Deviation method with Roe's approximate Riemann solver on a 64×128 grid to evolve the initial condition up to final time $t = 5$. Only the vortex initially (and finally) centered at $(0.5, 0.5)$ is included in the target solution. The result is illustrated in Fig. 6.8. It gets evident that the vortex included in the target solution is preserved while the other vortex loses velocity due to diffusion.

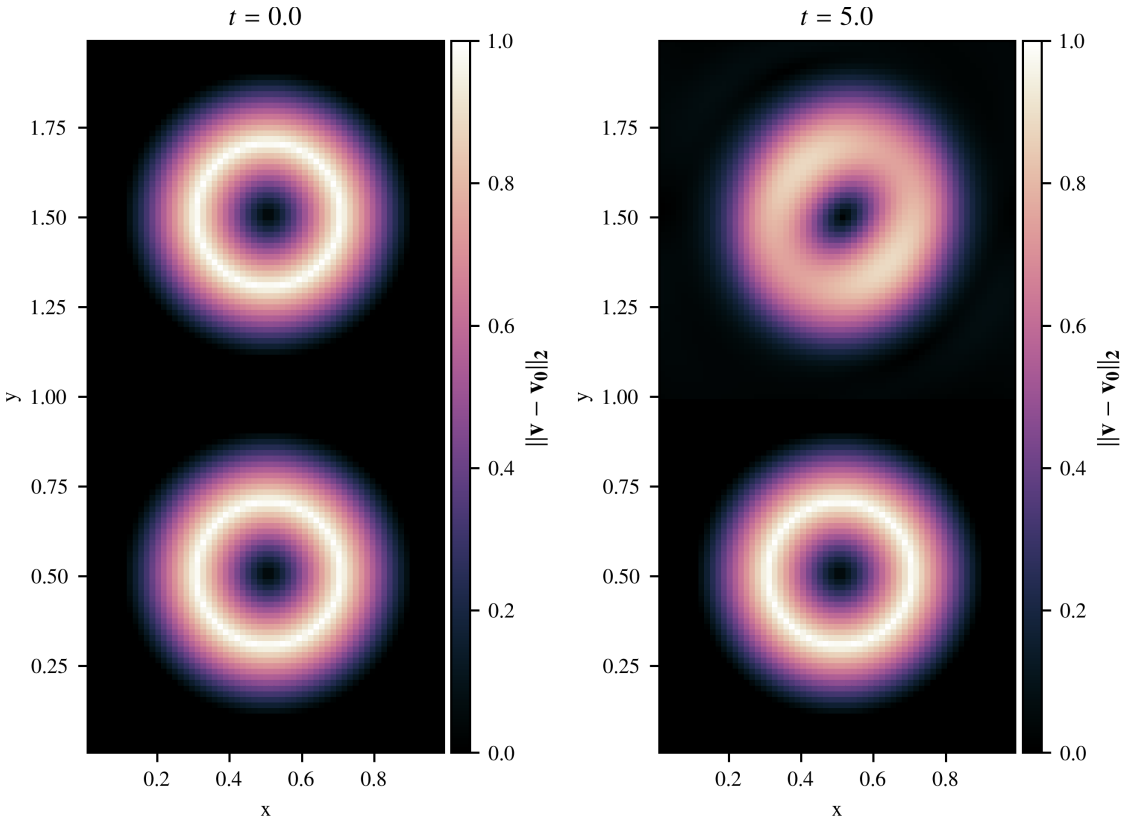


Figure 6.8: Illustration for the double Gresho vortex test from Section 6.4.5. The absolute velocity after subtraction of the constant advection velocity $\boldsymbol{v}_0$ is shown for the initial (left panel) and final (right panel) time. The vortex which is included in the target solution (bottom vortex in both panels) is preserved, whereas the other one is diffused and deformed.

6.4.6 Testing the Deviation Method on Ideal Magnetohydrodynamics Equations

The two-dimensional compressible ideal magnetohydrodynamics (MHD) equations which model the conservation of mass, momentum, magnetic field, and energy are given by

$$\partial_t \boldsymbol{q} + \partial_x \boldsymbol{f}_x + \partial_y \boldsymbol{f}_y = 0. \tag{6.65}$$

The conserved variables

$$\boldsymbol{q} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ B_x \\ B_y \\ E \end{pmatrix} \tag{6.66}$$

are evolved according to the fluxes

$$\boldsymbol{f}_x = \begin{pmatrix} \rho u \\ \rho u^2 + p + \frac{1}{2}(B_y^2 - B_x^2) \\ \rho u v - B_x B_y \\ 0 \\ B_y u - v B_x \\ u(E + p + \frac{1}{2}B_y^2 - \frac{1}{2}B_x^2) - v B_x B_y \end{pmatrix}, \tag{6.67}$$

$$\boldsymbol{f}_y = \begin{pmatrix} \rho v \\ \rho u v - B_x B_y \\ \rho v^2 + p + \frac{1}{2}(B_x^2 - B_y^2) \\ B_x v - u B_y \\ 0 \\ v(E + p + \frac{1}{2}B_x^2 - \frac{1}{2}B_y^2) - u B_x B_y \end{pmatrix}, \tag{6.68}$$

where B_x, B_y are the x- and y-component of the magnetic field. The total energy is $E = \rho\varepsilon + \frac{1}{2}\rho|\boldsymbol{v}|^2 + \frac{1}{2}(B_x^2 + B_y^2)$. All other quantities are defined as in the Euler equations and we apply an ideal gas EoS with the same parameters as for the Euler equations in order to relate the thermodynamical quantities.

Sometimes, two-dimensional compressible ideal MHD equations are defined such that they include ρw and B_z – i.e., the velocity and magnetic field components in the third spatial dimension – due to the genuine three-dimensional interactions between velocity and magnetic field. In our tests we set ρw and B_z to zero and we can thus omit the corresponding equations. The tests with the MHD equations we describe in the following are taken from our article [14] and the description is in parts similar or identical.

We consider an exact solution of the homogeneous two-dimensional ideal MHD

equations given by

$$\bar{x} = x - tu_0, \qquad \bar{y} = y - tv_0, \qquad r^2 = \bar{x}^2 + \bar{y}^2, \tag{6.69}$$

$$u = u_0 - k_p e^{\frac{1-r^2}{2}} \hat{y}, \qquad v = v_0 + k_p e^{\frac{1-r^2}{2}} \hat{x}, \qquad \rho = 1, \tag{6.70}$$

$$B_x = -m_p e^{\frac{1-r^2}{2}} \hat{y}, \qquad B_y = m_p e^{\frac{1-r^2}{2}} \hat{x}, \qquad p = 1 + \left(\frac{m_p^2}{2}(1-r^2) - \frac{k_p^2}{2} \right) e^{1-r^2}. \tag{6.71}$$

The vortex described by these formulae is advected through the domain $\Omega = [-5, 5] \times [-5, 5]$ with the velocity (u_0, v_0). One vortex turnover-time is $t_{\text{turnover}} = \frac{2\pi}{\sqrt{e} k_p} \approx \frac{3.81}{k_p}$. Note that in this section we only present numerical experiments with the Deviation method, since the other well-balanced methods presented in this thesis can only be applied to balance hydrostatic states of the Euler equations. As numerical flux function the Rusanov flux (Section 3.3.2) is used in all tests with the MHD system, since it is universally applicable for any hyperbolic system. Extrapolation boundary conditions are applied in order to allow for high order accuracy.

6.4.6.1 Long Time Evolution

In a first test we set the parameters to $m_p = k_p = 0.1$, $u_0 = v_0 = 0$ and run the test up to $t = 100 t_{\text{turnover}}$ on a 32×32 grid. We use the Deviation method and the target solution equals the initial data. The numerical error at final time compared to the initial setup is exactly zero in all conserved variables.

6.4.6.2 Order of Accuracy

In a second test with the MHD vortex described in Eqs. (6.69) to (6.71), we are interested to see if the Deviation method converges as expected, even if the target solution deviates from the actual solution over time. For that we set $m_p = k_p = 0.1$, $u_0 = v_0 = 0$ in the initial condition. As target solution we use the same vortex but with $u_0 = v_0 = 1$. In Table 6.5 the L^1 errors and rates at final time $t = 0.2$ are presented for the formally first, second, and third order accurate Deviation method. We omitted the errors for ρv and B_y. Due to the symmetry of the setup these errors equal the errors in ρu and B_x respectively. We see that even if the target solution moves away from the actual solution over time the method is still consistent and displays the expected order of accuracy.

6.4.6.3 Numerical Target Solution

To show the versatility of the Deviation method, in the following test case it is applied with a time-dependent target solution that is not given in the form of a function but in the form of discrete data. For this purpose, the MHD vortex described in Eqs. (6.69) to (6.71) with the parameters $k_p = m_p = 0.1$ and $u_0 = v_0 = 0.1$ is evolved up to the final time $t_{\text{final}} = 5$ using the third order standard method on a 128×128 Cartesian mesh. All these parameters are chosen as in [14]. The resulting approximate solution is used as target solution in the Deviation method. For this purpose the date are mapped on coarser grid. The cell-averaged values on the

Table 6.5: L^1-errors and convergence rates for the stationary MHD vortex test case described in Section 6.4.6.2 after time $t = 0.2$. The Deviation method is applied with a target solution that deviates from the actual solution over time.

first order Deviation method

N	ρ error	ρ rate	ρu error	ρu rate	B_x error	B_x rate	E error	E rate
32	2.10e-03	–	1.62e-02	–	1.26e-02	–	8.81e-02	–
64	8.31e-04	1.3	8.28e-03	1.0	6.90e-03	0.9	4.70e-02	0.9
128	3.52e-04	1.2	4.15e-03	1.0	3.57e-03	1.0	2.42e-02	1.0
256	1.59e-04	1.1	2.08e-03	1.0	1.81e-03	1.0	1.23e-02	1.0
512	7.53e-05	1.1	1.04e-03	1.0	9.12e-04	1.0	6.17e-03	1.0

second order Deviation method

N	ρ error	ρ rate	ρu error	ρu rate	B_x error	B_x rate	E error	E rate
32	2.67e-03	–	5.93e-03	–	5.95e-03	–	2.01e-02	–
64	1.18e-03	1.2	2.13e-03	1.5	2.13e-03	1.5	6.36e-03	1.7
128	4.57e-04	1.4	6.36e-04	1.7	6.37e-04	1.7	1.87e-03	1.8
256	1.52e-04	1.6	1.76e-04	1.9	1.76e-04	1.9	5.11e-04	1.9
512	4.49e-05	1.8	4.63e-05	1.9	4.63e-05	1.9	1.34e-04	1.9

third order Deviation method

N	ρ error	ρ rate	ρu error	ρu rate	B_x error	B_x rate	E error	E rate
32	1.91e-04	–	1.07e-03	–	1.07e-03	–	4.43e-03	–
64	1.73e-05	3.5	1.40e-04	2.9	1.39e-04	2.9	6.38e-04	2.8
128	1.67e-06	3.4	1.75e-05	3.0	1.75e-05	3.0	8.46e-05	2.9
256	1.82e-07	3.2	2.19e-06	3.0	2.19e-06	3.0	1.08e-05	3.0
512	2.16e-08	3.1	2.73e-07	3.0	2.73e-07	3.0	1.36e-06	3.0

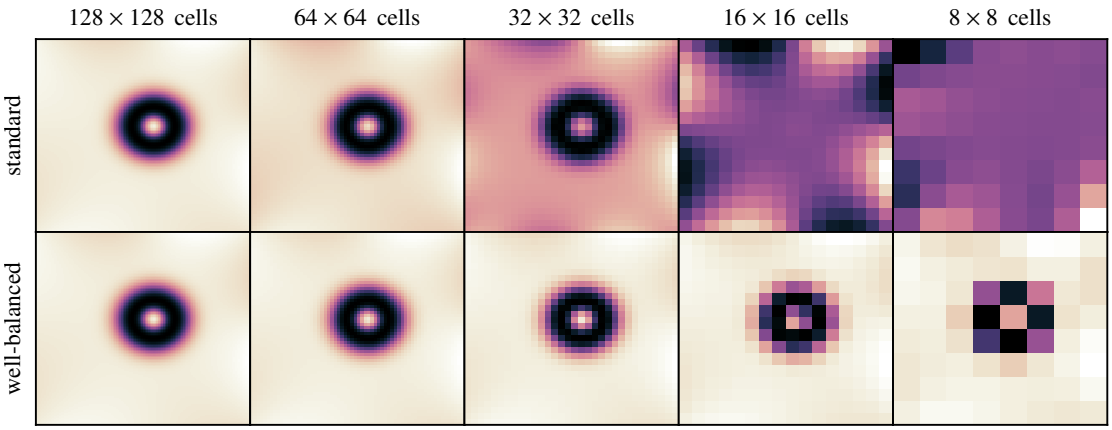


Figure 6.9: Pressure of the moving stationary MHD vortex as described in Section 6.4.6.3 at the final time. The target solution for the Deviation method is the numerical solution computed with the standard method on a 128×128 cells grid (upper left panel). In the upper panels the standard method is used, in the lower panels the Deviation method is used. Different columns correspond to different resolutions.

coarser grid are obtained by averaging the cell-averages from the fine grid. Third order accurate reconstruction is the applied in order to obtain sufficiently accurate cell-centered values and the values at interface quadrature points are then interpolated with a third order accurate interpolation from the cell-centered point values. At the end, a third order accurate interpolation in time is applied in order to obtain all the necessary cell-average and point values at the correct time in each step of the Deviation method. Note that this procedure relies on properties of the Cartesian grid we use. For a more general description the reader is referred to [14].

The Deviation method with the target solution obtained as described above and the standard method are used to evolve the initial data on different Cartesian grids (128×128, 64×64, 32×32, 16×16, 8×8) to the final time t_{final}. The pressure at final time for each of the simulations is shown in Fig. 6.9. All methods use CWENO3 reconstruction and parabolic extrapolation boundary conditions. On the 128×128 grid the solutions for the Deviation and standard method are exactly the same, since the solution from the standard method is used as target solution in the well-balanced method. For smaller resolutions, the standard method is too diffusive to resolve the vortex. The quality of the results obtained with the Deviation method is the same for all resolutions, since all of them use the same 128×128 simulation as target solution. This is a proof of concept that the Deviation method can be applied in combination with a target solution that has been obtained numerically.

6.4.7 Convection in a Stellar Shell

In this section we present a numerical experiment with astrophysical relevance. In a two-dimensional model of a star density, pressure, and temperature are stratified such, that there are three hydrostatic regions at different radial coordinate intervals $RI = [0, r_{I,II}]$, $RII = [r_{I,II}, r_{II,III}]$, and $RIII = [r_{II,III}, r_{\max}]$, where

$r_{\max} = 3/2RT_0/|g|$, $r_{I,II} = r_{\max}/3$, and $r_{II,III} = 2r_{\max}/3$. The center and the outermost shell (RI and $RIII$) are stable with respect to convection, which is ensured by setting the temperature gradient in these regions to $\partial_r T_I, \partial_r T_{III} = 0$. The temperature gradient in the region RII is set to $\partial_r T_{II} = \frac{\gamma-1}{R\gamma}|g|$, which makes the hydrostatic solution in this region marginally stable with respect to convection, since we choose an ideal gas law to describe the thermodynamical relations. The stability of hydrostatic solutions is determined using the square N^2_{BVF} of the Brunt–Väisälä frequency N_{BVF}. This quantity is defined for example in [15]. The sign of N_{BVF} indicates the convective stability of a hydrostatic stratification: If it is positive, the stratification is stable with respect to convection. If it is negative, the stratification is unstable with respect to convection which means that convection can be expected as soon as the hydrostatic stratification is disturbed. $N^2_{\text{BVF}} = 0$ means that the stratification is marginally stable. On the bottom of the marginally stable region, the gas is heated by adding an external energy source term. From the physical perspective, convection can be expected in RII.

The stratification is defined by the temperature gradient

$$\partial_r T(r) := \partial_r T_I + \frac{1}{2}\left(1 + \sin\left[\frac{\pi}{2}\,\eta(r, K, r_{I,II})\right]\right)(\partial_r T_I - \partial_r T_{II,III}(r))\,, \tag{6.72}$$

where

$$\partial_r T_{II,III}(r) := \partial_r T_{II} + \frac{1}{2}\left(1 + \sin\left[\frac{\pi}{2}\,\eta(r, K, r_{II,III})\right]\right)(\partial_r T_{II} - \partial_r T_{III}) \tag{6.73}$$

and

$$\eta(r, K, r_0) = \begin{cases} -1 \text{ if } K(r - r_0) < -1, \\ 1 \text{ if } K(r - r_0) > 1, \\ K(r - r_0) \text{ else .} \end{cases} \tag{6.74}$$

Equations (6.72) and (6.73) smoothly connect the temperature gradients $\partial_r T_I, \partial_r T_{II}$, and $\partial_r T_{III}$. The temperature gradient $\partial_r T$ is then integrated to obtain the temperature profile

$$T(r) := T_0 + \int_0^r \partial_r T(\tau)\, d\tau. \tag{6.75}$$

To find the pressure stratification, we plug the ideal gas EoS into the hydrostatic equation which yields

$$\partial_r p(r) = \frac{p(r)}{RT(r)} g. \tag{6.76}$$

This ODE is solved numerically using the DOPRI5 solver, a fifth order accurate explicit RK method introduced in [54], using the starting point $(r = 0, p = p_0)$. The hydrostatic density can then simply be computed from the pressure and temperature using the ideal gas EoS. The parameters we use in this setup are

$$T_0 = 1e8, \quad p_0 = 1.2e17, \quad g = -1000, \quad K = 15/r_{\max}, \quad R = 8.31446261815324e7. \tag{6.77}$$

An additional energy source term is added to the Euler system to add internal energy in the marginally stable region RII such that convection can develop. Hence, we solve the system

$$\partial_t \boldsymbol{q} + \nabla \cdot \mathcal{F} = \boldsymbol{s} + \boldsymbol{s}_{\text{heating}}, \tag{6.78}$$

where $\boldsymbol{q}$, $\mathcal{F}$, and $\boldsymbol{s}$ are the state, flux-tensor, and the gravity source term of the two-dimensional Euler equations as given in Eqs. (5.6) to (5.8). The heating term $\boldsymbol{s}_{\text{heating}} = \left(0, 0, 0, s^E_{\text{heating}}\right)$ is defined via

$$s^E_{\text{heating}} = \bar{d\varepsilon} \exp\left(-\left(\frac{r - r_0}{\kappa}\right)^2\right), \tag{6.79}$$

where we choose the parameters

$$\bar{d\varepsilon} = 3000 \qquad r_0 = 10^7 \qquad \kappa = \frac{r_0}{10}. \tag{6.80}$$

We apply the second order accurate standard, α-β, and Deviation method on a 144×144 cubed sphere grid. In the standard and Deviation method, primitive variables are reconstructed (see Remark 4.3.10). The heating source term is discretized via cell centered evaluation. As numerical flux function we choose the AUSM$^+$-up solver introduced in [114]. This numerical flux function is designed to accurately simulate flows at low Mach numbers. Instead of our custom Python code we use the Seven-League-Hydro (SLH) code described in [120, 57, 143], for example. The semi-discrete scheme is evolved in time using the implicit Runge–Kutta method ESDIRK23 [85]. The simulation is run to the final time $t = 350000$ with the well-balanced methods but only to $t \approx 20000$ with the standard method, since this reduced time is enough to show the unphysical behavior in that case.

The local Mach number at final time is shown in Fig. 6.10. In the top left panel the value of N^2_{BVF} is shown to illustrate the position of the three regions. In the simulation with the standard method we see turbulent patterns in the center of the domain, which is determined to be stable by theory. Also, the maximal Mach number of about 0.1 is quite high. In the simulations using the α-β and the Deviation method there are, in accordance with the theoretical expectation, only turbulent patterns in the marginally stable region. This test is a proof of concept that our well-balanced methods in combination with a low Mach number compliant numerical flux function can be applied to simulate turbulent stellar convection at low Mach numbers. A similar but more sophisticated test setup, in which the Helmholtz EoS (e.g., [160]) is applied, can be found in [58].

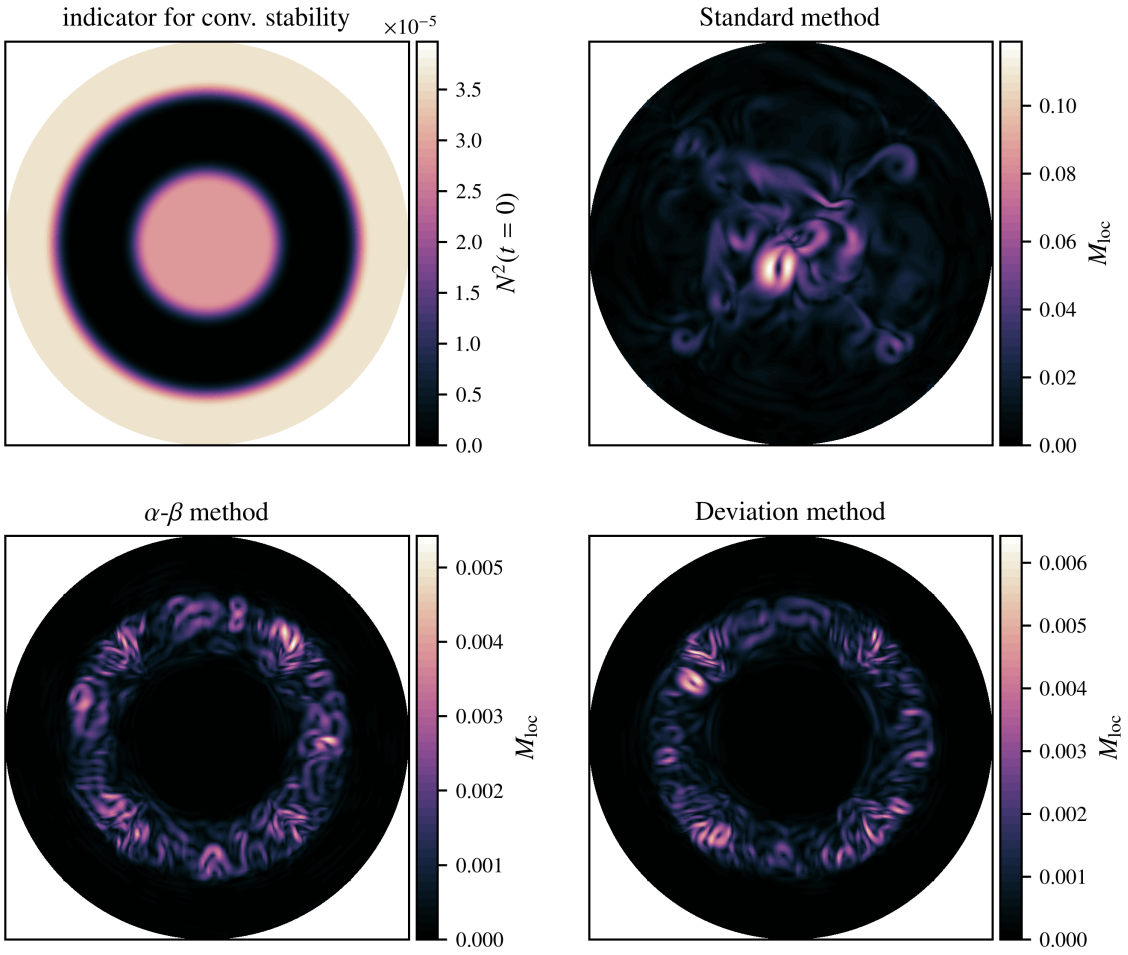


Figure 6.10: Local Mach number at final time for the convection setup described in Section 6.4.7 for different methods. In the top left panel, the square of the Brunt Väisälä frequency at initial time is shown. In the regions with positive values the hydrostatic state is stable with respect to convection. In the region with zero value, the setup is marginally stable with respect to convection. The x-coordinate increases to the right, the y-coordinate to the top. The domain is $[-r_{\max}, r_{\max}]^2$ with $r_{\max}$ as given in the text.

Chapter 7

Conclusions and Outlook

In this thesis, we presented three different high order well-balanced methods. The Deviation method (Section 4.3) (which is based on the idea of the α-β method already introduced in [11] and described in more detail in Sections 4.2 and 6.1) was described in Sections 4.3 and 6.2. The Deviation method is general in the sense that it can be applied to exactly follow any known solution of any hyperbolic system with or without a source term. The Discretely Well-Balanced method, introduced in Section 4.4, is constructed to balance hydrostatic states of the compressible Euler system with gravitational source term without any a priori knowledge or assumptions. For this purpose, a local approximation of the hydrostatic state is balanced exactly. However, the Discretely Well-Balanced method suffers from an increased stencil which can be especially problematic in the context of parallel high performance computing based on domain decomposition. In order to reduce the stencil, the Local Approximation method has been developed based on the Discretely Well-Balanced method and a localization of the hydrostatic pressure approximation. Whereas no well-balanced property could be shown for the latter scheme, it yields similar or even better results in numerical experiments compared to the Discretely Well-Balanced method. The improved accuracy close to hydrostatic states and the order of accuracy have been verified analytically and numerically for all schemes discussed in this thesis.

The Deviation and Local Approximation method have been extended to two spatial dimensions in Chapter 6 and improved accuracy close to hydrostatic solutions was made evident in numerical experiments in Section 6.4. Numerical experiments with non-static stationary and even time-dependent solutions of the Euler equations with and without gravity source term and the equations of homogeneous ideal magnetohydrodynamics verified the versatility of the Deviation method.

In Section 6.4.7 we presented an application of the α-β and Deviation method, in which convection in a stellar shell has been simulated at low Mach numbers on a cubed sphere grid using the astrophysical code SLH. The simulation was based on the combination of the well-balanced methods with a low Mach number compliant numerical flux function and fully implicit time stepping. The cubed sphere grid allows an adaption to the spherical shape of the star with a structured grid without avoiding the center as it would be necessary in a polar grid. Our well-balanced methods are especially suitable for astrophysical simulations, since they are not

restricted to an ideal gas EoS. Also, they have been constructed in a way that they can be implemented in existing finite volume codes with minimal effort. They are furthermore flexible in the sense that they can be combined with arbitrary numerical fluxes and ODE solvers for time stepping.

However, the combination of well-balancing and low Mach number compliant numerical fluxes remains an open area of research. In the low Mach convection test case we presented in this thesis, a numerical flux was applied that explicitly adds pressure diffusion to stabilize the simulation. As most low Mach number compliant numerical fluxes, it has been developed for the homogeneous compressible Euler system without source term. To further improve the capability for simulations of compressible convection at low Mach numbers, a better understanding of the stability of the combination of well-balancing and the application of low Mach number compliant numerical fluxes is required.

Appendix A

Details on Some Test Setups

A.1 Test Shown in Figure 3.3

The initial data for the test setup corresponding to Fig. 3.3 are given by

$$\rho(x,0) := 1, \quad u(x,0) := 0, \quad \text{and} \quad p(x,0) := 1 + \exp\left(-100\,(x-0.3)^2\right) \quad \text{for} \quad x \in \Omega \tag{A.1}$$

on the domain $\Omega = [0,1]$ with periodic boundary conditions. Using the compressible Euler equations (2.19) with an ideal gas law (2.21), where $R = 1$ and $\gamma = 1.4$, the solution is evolved to the final time $t = 0.85$. The reference solution (black line in the right panel of Fig. 3.3) is obtained using a finite volume method with the CWENO7 reconstruction from [47], the Roe flux [141], and the explicit RK10 method from [62] on a grid with 1000 cells. The blue line in Fig. 3.3 is obtained using a finite volume method with constant reconstruction, the central flux described in Section 3.3.1, and explicit forward Euler time stepping.

A.2 Test Shown in Figure 3.4

The initial condition for this classical shock tube problem for the compressible Euler equations (2.19) with the ideal gas law (2.21), where $R = 1$ and $\gamma = 1.4$, is given by [147]

$$\boldsymbol{q}(x,0) := \begin{cases} \boldsymbol{q}^L & \text{for } x < \frac{1}{2} \\ \boldsymbol{q}^R & \text{for } x \geq \frac{1}{2} \end{cases}, \tag{A.2}$$

where

$$\boldsymbol{q}^L = \begin{pmatrix} \rho^L \\ (\rho u)^L \\ E^L \end{pmatrix} := \begin{pmatrix} 1 \\ 0 \\ \frac{1}{\gamma-1} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{q}^R = \begin{pmatrix} \rho^R \\ (\rho u)^R \\ E^R \end{pmatrix} := \begin{pmatrix} 0.1 \\ 0 \\ \frac{0.125}{\gamma-1} \end{pmatrix}. \tag{A.3}$$

The initial data are evolved to the final time $t = 0.2$ on the domain $\Omega = [0,1]$ with Dirichlet boundary conditions as described in Section 3.8. The reference solution in Fig. 3.4 is obtained using a finite volume method with the CWENO7 reconstruction

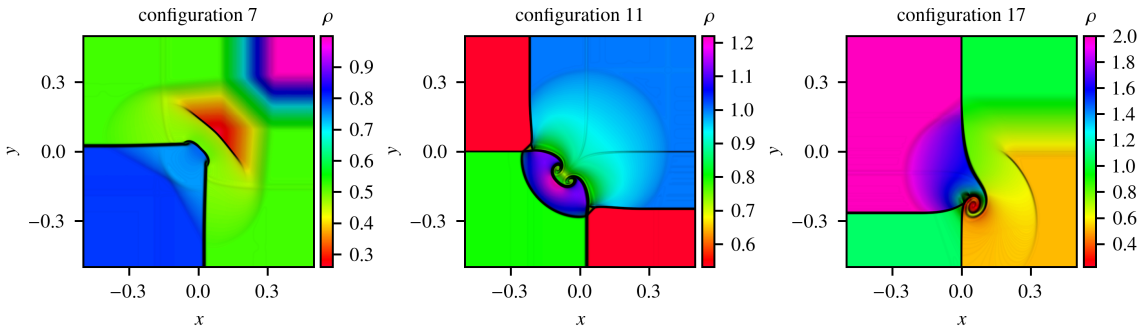


Figure A.1: Figure shown in Fig. 5.1 including labels. The initial data and final time for the Riemann problems are given as configurations 7, 11, and 17 in [99]. The density at final time is shown in each plot. In order to emphasize the interfaces, 1000 thin light gray contour lines, uniformly covering the range of density values, have been added. They are allowed to accumulate to darker and finally black lines.

from [47], the Roe flux [141], and the explicit RK10 method from [62] on a grid with 1000 cells. The other solutions which are shown in Fig. 3.4 are obtained using a finite volume method on a 100 cells grid with constant reconstruction, forward Euler time stepping, and different numerical flux functions as displayed in the legend of the figure.

A.3 Test Shown in Figure 5.1

Two-dimensional Riemann problems for the two-dimensional homogeneous Euler system are simulated using the domain, initial data, and final time from the configurations 7, 11, and 17 in [99]. The density at final time for each problem is shown in Figs. 5.1 and A.1. The simulations are conducted using our second order standard finite volume method on a 600×600 cells Cartesian grid. The Roe flux and minmod limited linear reconstruction are applied. The semi-discrete scheme is evolved with the explicit RK3 ODE solver. The boundaries are extrapolated using constant extrapolation.

Appendix B

Some Second Order Accurate Products and Conversions

B.1 Products

Let $q, r \in \mathcal{C}^2(\Omega, \mathbb{R})$ be functions defined on the convex compact domain $\Omega \subset \mathbb{R}^{\mathfrak{d}}$ ($\mathfrak{d} = 1, 2$) with $\|\boldsymbol{x} - \boldsymbol{y}\| < h$ for any $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and some $h > 0$.

$$\hat{q} = \frac{1}{|\Omega|} \int_{\Omega} q(\boldsymbol{x})\, d\boldsymbol{x}, \qquad \text{and} \qquad \hat{r} = \frac{1}{|\Omega|} \int_{\Omega} r(\boldsymbol{x})\, d\boldsymbol{x} \tag{B.1}$$

are be the averages of q and r over Ω. Let $\boldsymbol{x}_0 \in \Omega$ be the quadrature point of a one-point Gaussian quadrature rule (Sections 3.5 and 5.5), i.e., it is

$$\hat{f} = f(\boldsymbol{x}_0) + \mathcal{O}\left(h^2\right) \tag{B.2}$$

for any function $f \in \mathcal{C}(\Omega, \mathbb{R})$. In the case $\mathfrak{d} = 1$, this is $\boldsymbol{x}_0 = x_0 = (a + b)/2$ for $\Omega = [a, b]$. In the case $\mathfrak{d} = 2$ and if the cell is from a Cartesian grid ($\Omega = [a, b] \times [c, d]$), it is $\boldsymbol{x}_0 = ((a + b)/2, (c + d)/2)^T$.

The following holds true:

(a) $\hat{q}r(\boldsymbol{x}_0) \overset{(B.2)}{=} (q(\boldsymbol{x}_0) + \mathcal{O}(h^2))r(\boldsymbol{x}_0) = q(\boldsymbol{x}_0)r(\boldsymbol{x}_0) + \mathcal{O}(h^2)$

(b) $\hat{q}r(\boldsymbol{x}_0) \overset{(a)}{=} q(\boldsymbol{x}_0)r(\boldsymbol{x}_0) + \mathcal{O}(h^2) \overset{(B.2)}{=} \widehat{(qr)} + \mathcal{O}(h^2)$

(c) $\hat{q}\hat{r} \overset{(B.2)}{=} (q(\boldsymbol{x}_0) + \mathcal{O}(h^2))\,(r(\boldsymbol{x}_0) + \mathcal{O}(h^2)) = q(\boldsymbol{x}_0)r(\boldsymbol{x}_0) + \mathcal{O}(h^2)$

Collecting (a)-(b), all the products

$$\hat{q}\hat{r} \approx \hat{q}r(\boldsymbol{x}_0) \approx q(\boldsymbol{x}_0)\hat{r} \approx q(\boldsymbol{x}_0)r(\boldsymbol{x}_0) \tag{B.3}$$

approximate each other with second order accuracy. Let $r > 0$ on the whole domain Ω. Then, according to the chain rule, $1/r \in \mathcal{C}^2$ and, applying (a)-(c), the quotients

$$\frac{\hat{q}}{\hat{r}} \approx \frac{\hat{q}}{r(\boldsymbol{x}_0)} \approx \frac{q(\boldsymbol{x}_0)}{\hat{r}} \approx \frac{q(\boldsymbol{x}_0)}{r(\boldsymbol{x}_0)} \tag{B.4}$$

approximate each other with second order accuracy. Let $\boldsymbol{q} \in \mathcal{C}(\Omega, \mathbb{R}^{\mathfrak{n}})$ with some $\mathfrak{n} \in \mathbb{N}, \mathfrak{n} > 1$ be vector-valued. Applying above approximations component-wise yields the second order accurate approximations

$$\hat{\boldsymbol{q}}\hat{r} \approx \hat{\boldsymbol{q}} r(\boldsymbol{x}_0) \approx \boldsymbol{q}(\boldsymbol{x}_0)\hat{r} \approx \boldsymbol{q}(\boldsymbol{x}_0) r(\boldsymbol{x}_0) \tag{B.5}$$

and

$$\frac{\hat{\boldsymbol{q}}}{\hat{r}} \approx \frac{\hat{\boldsymbol{q}}}{r(\boldsymbol{x}_0)} \approx \frac{\boldsymbol{q}(\boldsymbol{x}_0)}{\hat{r}} \approx \frac{\boldsymbol{q}(\boldsymbol{x}_0)}{r(\boldsymbol{x}_0)}. \tag{B.6}$$

Note that multiplication with a constant is just a special case of the approximations discussed above.

B.2 Conversions

For Ω, $\boldsymbol{x}_0$, and $\hat{\cdot}$ we use the notation from Appendix B.1. Let $\boldsymbol{q}^{\text{cons}} = (\rho, \rho u, \rho v, E) \in \mathcal{C}^2(\Omega, \mathbb{R}^+ \times \mathbb{R}^2 \times \mathbb{R}^+)$ be a field of states given in conserved variables, and $\boldsymbol{q}^{\text{prim}} = (\rho, u, v, p) \in \mathcal{C}^2(\Omega, \mathbb{R}^+ \times \mathbb{R}^2 \times \mathbb{R}^+)$ the corresponding field of states in primitive variables such that

$$\boldsymbol{q}^{\text{prim}} = T(\boldsymbol{q}^{\text{cons}})\boldsymbol{q}^{\text{cons}}, \tag{B.7}$$

where

$$T^{\text{prim}}_{\text{cons}}(\bar{\boldsymbol{q}}) := \left.\frac{\partial \boldsymbol{q}^{\text{prim}}\left(\boldsymbol{q}^{\text{cons}}\right)}{\partial \boldsymbol{q}^{\text{cons}}}\right|_{\boldsymbol{q}^{\text{cons}}=\bar{\boldsymbol{q}}} \quad \text{and} \quad T^{\text{cons}}_{\text{prim}}(\bar{\boldsymbol{q}}) := \left.\frac{\partial \boldsymbol{q}^{\text{cons}}\left(\boldsymbol{q}^{\text{prim}}\right)}{\partial \boldsymbol{q}^{\text{prim}}}\right|_{\boldsymbol{q}^{\text{prim}}=\bar{\boldsymbol{q}}} \tag{B.8}$$

are the transformations between the variable systems as given in Section 5.1.1. Furthermore, for the EoS we assume that it is sufficiently smooth such that

$$\varepsilon_{\text{EoS}}\left(\rho + \mathcal{O}\left(h^2\right), p + \mathcal{O}\left(h^2\right)\right) = \varepsilon_{\text{EoS}}(\rho, p) + \mathcal{O}\left(h^2\right), \tag{B.9}$$

$$p_{\text{EoS}}\left(\rho + \mathcal{O}\left(h^2\right), \varepsilon + \mathcal{O}\left(h^2\right)\right) = p_{\text{EoS}}(\rho, \varepsilon) + \mathcal{O}\left(h^2\right),$$

$$\rho_{\text{EoS}}\left(p + \mathcal{O}\left(h^2\right), \varepsilon + \mathcal{O}\left(h^2\right)\right) = \rho_{\text{EoS}}(p, \varepsilon) + \mathcal{O}\left(h^2\right). \tag{B.10}$$

Then the following statements hold true:

(a) A straightforward computation shows that

$$T^{\text{prim}}_{\text{cons}}\left(\boldsymbol{q} + \mathcal{O}\left(h^2\right)\right) = T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}) + \mathcal{O}\left(h^2\right) \quad \text{and} \tag{B.11}$$

$$T^{\text{cons}}_{\text{prim}}\left(\boldsymbol{q} + \mathcal{O}\left(h^2\right)\right) = T^{\text{cons}}_{\text{prim}}(\boldsymbol{q}) + \mathcal{O}\left(h^2\right) \tag{B.12}$$

(b) It is also easy to show via direct computation that

$$\left(T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}) + \mathcal{O}\left(h^2\right)\right)\boldsymbol{q} = T^{\text{prim}}_{\text{cons}}(\boldsymbol{q})\boldsymbol{q} + \mathcal{O}\left(h^2\right) \tag{B.13}$$

$$T^{\text{prim}}_{\text{cons}}(\boldsymbol{q})\left(\boldsymbol{q} + \mathcal{O}\left(h^2\right)\right) = T^{\text{prim}}_{\text{cons}}(\boldsymbol{q})\boldsymbol{q} + \mathcal{O}\left(h^2\right) \tag{B.14}$$

and consequently

$$\left(T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}) + \mathcal{O}\left(h^2\right)\right)\left(\boldsymbol{q} + \mathcal{O}\left(h^2\right)\right) = T^{\text{prim}}_{\text{cons}}(\boldsymbol{q})\boldsymbol{q} + \mathcal{O}\left(h^2\right). \tag{B.15}$$

The corresponding holds for the transformation $T^{\text{cons}}_{\text{prim}}$.

(c) Combining (a),(b), and Eq. (B.2) yields

$$T^{\text{prim}}_{\text{cons}}(\hat{\boldsymbol{q}})\hat{\boldsymbol{q}} = T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}(\boldsymbol{x}_0))\boldsymbol{q}(\boldsymbol{x}_0) + \mathcal{O}\left(h^2\right), \tag{B.16}$$

$$T^{\text{cons}}_{\text{prim}}(\hat{\boldsymbol{q}})\hat{\boldsymbol{q}} = T^{\text{cons}}_{\text{prim}}(\boldsymbol{q}(\boldsymbol{x}_0))\boldsymbol{q}(\boldsymbol{x}_0) + \mathcal{O}\left(h^2\right). \tag{B.17}$$

(d) Converting cell-averages is thus second order accurate:

$$\begin{aligned}\hat{\boldsymbol{q}}^{\text{prim}} \overset{Eq.\ (B.2)}{=} \boldsymbol{q}^{\text{prim}}(\boldsymbol{x}_0) + \mathcal{O}\left(h^2\right) &= T^{\text{prim}}_{\text{cons}}\left(\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0)\right)\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0) + \mathcal{O}\left(h^2\right) \\ &\overset{(c)}{=} T^{\text{prim}}_{\text{cons}}\left(\hat{\boldsymbol{q}}^{\text{cons}}\right)\hat{\boldsymbol{q}}^{\text{cons}} + \mathcal{O}\left(h^2\right)\end{aligned} \tag{B.18}$$

and vice-versa

$$\hat{\boldsymbol{q}}^{\text{cons}} = T^{\text{cons}}_{\text{prim}}\left(\hat{\boldsymbol{q}}^{\text{prim}}\right)\hat{\boldsymbol{q}}^{\text{prim}} + \mathcal{O}\left(h^2\right) \tag{B.19}$$

(e) Let $\alpha, \beta > 0$ and let

$$T_{\alpha,\beta}\boldsymbol{q}^{\text{prim}} = T_{\alpha,\beta}(\rho, u, v, p)^T := \left(\frac{\rho}{\alpha}, u, v, \frac{p}{\beta}\right)^T =: \boldsymbol{q}^{\alpha,\beta} \tag{B.20}$$

define the linear operator $T_{\alpha,\beta}$ with inverse $T^{-1}_{\alpha,\beta}$ (Existence is clear because of $\det\left(T_{\alpha,\beta}\right) = \alpha\beta > 0$). Then the following can be shown by straightforward computations

$$\boldsymbol{q}^{\bar{\alpha},\bar{\beta}} = \boldsymbol{q}^{\alpha,\beta} + \mathcal{O}\left(h^2\right), \tag{B.21}$$

$$T_{\alpha,\beta}\left(\boldsymbol{q}^{\text{prim}} + \mathcal{O}\left(h^2\right)\right) = \boldsymbol{q}^{\alpha,\beta} + \mathcal{O}\left(h^2\right), \tag{B.22}$$

$$T_{\bar{\alpha},\bar{\beta}}\left(\boldsymbol{q}^{\text{prim}} + \mathcal{O}\left(h^2\right)\right) = \boldsymbol{q}^{\alpha,\beta} + \mathcal{O}\left(h^2\right), \tag{B.23}$$

where $\bar{\alpha}, \bar{\beta} > 0$ with $\bar{\alpha} = \alpha + \mathcal{O}\left(h^2\right)$ and $\bar{\beta} = \beta + \mathcal{O}\left(h^2\right)$. The corresponding relations hold for the inverse $T^{-1}_{\alpha,\beta}$.

(f) From (d) and (e) it follows that

$$T_{\bar{\alpha},\bar{\beta}}\left(T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}^{\text{cons}}) + \mathcal{O}\left(h^2\right)\right)\left(\boldsymbol{q}^{\text{cons}} + \mathcal{O}\left(h^2\right)\right) = T_{\alpha,\beta}T^{\text{prim}}_{\text{cons}}\left(\boldsymbol{q}^{\text{cons}}\right)\boldsymbol{q}^{\text{cons}} + \mathcal{O}\left(h^2\right) \tag{B.24}$$

(g) Now let $\alpha, \beta \in \mathcal{C}^2\left(\Omega, \mathbb{R}^+\right)$. Then it follows from Eq. (B.2) and (e) that

$$\begin{aligned}\widehat{\left(T_{\alpha,\beta}\hat{\boldsymbol{q}}^{\text{prim}}\right)} \approx T_{\hat{\alpha},\hat{\beta}}\hat{\boldsymbol{q}}^{\text{prim}} &\approx T_{\hat{\alpha},\hat{\beta}}\boldsymbol{q}^{\text{prim}}(\boldsymbol{x}_0) \\ &\approx T_{\alpha(\boldsymbol{x}_0),\beta(\boldsymbol{x}_0)}\hat{\boldsymbol{q}}^{\text{prim}} \approx T_{\alpha(\boldsymbol{x}_0),\beta(\boldsymbol{x}_0)}\boldsymbol{q}^{\text{prim}}(\boldsymbol{x}_0)\end{aligned} \tag{B.25}$$

approximate each other with second order accuracy and correspondingly for the inverse. Similarly, additionally applying (f) we get that

$$\begin{aligned}\widehat{\left(T_{\alpha,\beta}T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}^{\text{cons}})\boldsymbol{q}^{\text{cons}}\right)} &\approx \widehat{\left(T_{\alpha,\beta}T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}^{\text{cons}})\right)}\hat{\boldsymbol{q}}^{\text{cons}} \\ &\approx \widehat{\left(T_{\alpha,\beta}T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}^{\text{cons}})\right)}\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0) \approx T_{\hat{\alpha},\hat{\beta}}\widehat{\left(T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}^{\text{cons}})\boldsymbol{q}^{\text{cons}}\right)} \\ &\approx T_{\hat{\alpha},\hat{\beta}}T^{\text{prim}}_{\text{cons}}(\hat{\boldsymbol{q}}^{\text{cons}})\hat{\boldsymbol{q}}^{\text{cons}} \approx T_{\alpha(\boldsymbol{x}_0),\beta(\boldsymbol{x}_0)}T^{\text{prim}}_{\text{cons}}(\hat{\boldsymbol{q}}^{\text{cons}})\hat{\boldsymbol{q}}^{\text{cons}} \\ &\approx T_{\hat{\alpha},\hat{\beta}}T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0))\hat{\boldsymbol{q}}^{\text{cons}} \approx T_{\hat{\alpha},\hat{\beta}}T^{\text{prim}}_{\text{cons}}(\hat{\boldsymbol{q}}^{\text{cons}})\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0) \\ &\approx T_{\alpha(\boldsymbol{x}_0),\beta(\boldsymbol{x}_0)}T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0))\hat{\boldsymbol{q}}^{\text{cons}} \approx T_{\hat{\alpha},\hat{\beta}}T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0))\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0) \\ &\approx T_{\alpha(\boldsymbol{x}_0),\beta(\boldsymbol{x}_0)}T^{\text{prim}}_{\text{cons}}(\hat{\boldsymbol{q}}^{\text{cons}})\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0) \approx T_{\alpha(\boldsymbol{x}_0),\beta(\boldsymbol{x}_0)}T^{\text{prim}}_{\text{cons}}(\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0))\boldsymbol{q}^{\text{cons}}(\boldsymbol{x}_0)\end{aligned} \tag{B.26}$$

approximate each other with second order accuracy. The corresponding relations hold for the back-transformation $T^{\text{cons}}_{\text{prim}} T^{-1}_{\alpha,\beta}$.

All of the assertions (a)-(g) obviously also hold in the one-dimensional case, in which there is only on velocity and momentum component in the state vectors.

Bibliography

[1] H. Al Baba, C. Klingenberg, O. Kreml, V. Mácha, and S. Markfelder. Nonuniqueness of admissible weak solution to the riemann problem for the full euler system in two dimensions. *SIAM Journal on Mathematical Analysis*, 52(2):1729–1760, 2020.

[2] F. Aràndiga, A. Baeza, A. Belda, and P. Mulet. Analysis of WENO schemes for full and global accuracy. *SIAM Journal on Numerical Analysis*, 49(2):893–915, 2011.

[3] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing*, 25(6):2050–2065, 2004.

[4] A. Avagadro. Essai d'une maniere de determiner les masses relatives des molecules elementaires des corps, et les proportions selon lesquelles elles entrent dans ces combinaisons. *J. Physique*, 73:58–76, 1811.

[5] D. S. Balsara and C.-W. Shu. Monotonicity preserving weighted essentially non-oscillatory schemes with increasingly high order of accuracy. *Journal of Computational Physics*, 160(2):405–452, 2000.

[6] W. Barsukow and J. P. Berberich. A well-balanced active flux scheme for the shallow water equations with wetting and drying. *Submitted to Journal on Scientific Computing*, 2020.

[7] W. Barsukow, J. P. Berberich, and C. Klingenberg. On the active flux scheme for hyperbolic pdes with source terms. *Submitted to SIAM Journal on Scientific Computing*, 2020.

[8] W. Barsukow, P. V. F. Edelmann, C. Klingenberg, F. Miczek, and F. K. Röpke. A numerical scheme for the compressible low-Mach number regime of ideal fluid dynamics. *Journal of Scientific Computing*, pages 1–24, 2017.

[9] W. Barsukow, J. Hohm, C. Klingenberg, and P. L. Roe. The active flux scheme on cartesian grids and its low mach number limit. *Journal of Scientific Computing*, 81(1):594–622, 2019.

[10] T. Barth. Recent developments in high order k-exact reconstruction on unstructured meshes. In *31st Aerospace Sciences Meeting*, page 668, 1993.

[11] J. P. Berberich. A well-balanced low Mach numerical scheme to solve Euler equations with gravity for astrophysical applications. Master's thesis, Universität Würzburg, 2016.

[12] J. P. Berberich, P. Chandrashekar, R. Käppeli, and C. Klingenberg. High order discretely well-balanced methods for arbitrary hydrostatic atmospheres. *Submitted to Communications in Computational Physics*, 2020.

[13] J. P. Berberich, P. Chandrashekar, and C. Klingenberg. A general well-balanced finite volume scheme for Euler equations with gravity. In C. Klingenberg and M. Westdickenberg, editors, *Theory, Numerics and Applications of Hyperbolic Problems I, Springer Proceedings in Mathematics & Statistics 236*, pages 151–163, 2018.

[14] J. P. Berberich, P. Chandrashekar, and C. Klingenberg. High order well-balanced finite volume methods for multi-dimensional systems of hyperbolic balance laws. *arXiv preprint arXiv:1903.05154*, 2019.

[15] J. P. Berberich, P. Chandrashekar, C. Klingenberg, and F. K. Röpke. Second order finite volume scheme for Euler equations with gravity which is well-balanced for general equations of state and grid systems. *Communications in Computational Physics*, 26:599–630, 2019.

[16] J. P. Berberich and C. Klingenberg. Entropy stable numerical fluxes for compressible Euler equations which are suitable for all Mach numbers. *Accepted for publication in: SEMA SIMAI Series: Numerical methods for hyperbolic problems Numhyp 2019*, 2020.

[17] A. Bermúdez, X. López, and M. E. Vázquez-Cendón. Finite volume methods for multi-component Euler equations with source terms. *Computers & Fluids*, 156:113–134, 2017.

[18] A. Bermudez and M. E. Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8):1049–1071, 1994.

[19] J. Blazek. *Computational fluid dynamics: principles and applications*. Butterworth-Heinemann, 2015.

[20] L. Boltzmann. *Lectures on Gas Theory. Translated by Stephen G. Brush.* University of California Press, 1964.

[21] N. Botta, R. Klein, S. Langenberg, and S. Lützenkirchen. Well balanced finite volume methods for nearly hydrostatic flows. *Journal of Computational Physics*, 196(2):539–565, 2004.

[22] F. Bouchut. *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws*. Birkhäuser, Basel, 2004.

[23] F. Bouchut and V. Zeitlin. A robust well-balanced scheme for multi-layer shallow water equations. 2010.

[24] J. Britton and Y. Xing. High order still-water and moving-water equilibria preserving discontinuous Galerkin methods for the Ripa model. *Journal of Scientific Computing*, 82(2):30, 2020.

[25] P. Brufau, M. Vázquez-Cendón, and P. García-Navarro. A numerical model for the flooding and drying of irregular domains. *International Journal for Numerical Methods in Fluids*, 39(3):247–275, 2002.

[26] P. Buchmüller and C. Helzel. Improved accuracy of high-order WENO finite volume methods on Cartesian grids. *Journal of Scientific Computing*, 61(2):343–368, 2014.

[27] J. C. Butcher. Coefficients for the study of runge-kutta integration processes. *Journal of the Australian Mathematical Society*, 3(02):185–201, 1963.

[28] J. C. Butcher. Implicit runge-kutta processes. *Mathematics of Computation*, 18(85):50–64, 1964.

[29] J. C. Butcher. *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods*. Wiley-Interscience, 1987.

[30] J. C. Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

[31] D. A. Calhoun, C. Helzel, and R. J. Leveque. Logically rectangular grids and finite volume methods for PDEs in circular and spherical domains. *SIAM Review*, 50:723–752, Jan. 2008.

[32] G. Capdeville. A central WENO scheme for solving hyperbolic conservation laws on non-uniform meshes. *Journal of Computational Physics*, 227(5):2977–3014, 2008.

[33] G. Capdeville. A high-order multi-dimensional hll-riemann solver for non-linear euler equations. *Journal of Computational Physics*, 230(8):2915–2951, 2011.

[34] P. Cargo and A. LeRoux. A well balanced scheme for a model of atmosphere with gravity. *COMPTES RENDUS DE L ACADEMIE DES SCIENCES SERIE I-MATHEMATIQUE*, 318(1):73–76, 1994.

[35] M. Castro, J. M. Gallardo, J. A. López-GarcÍa, and C. Parés. Well-balanced high order extensions of Godunov's method for semilinear balance laws. *SIAM Journal on Numerical Analysis*, 46(2):1012–1039, 2008.

[36] M. J. Castro and C. Parés. Well-balanced high-order finite volume methods for systems of balance laws. *Journal of Scientific Computing*, 82(2):48, 2020.

[37] M. J. Castro and M. Semplice. Third-and fourth-order well-balanced schemes for the shallow water equations based on the CWENO reconstruction. *International Journal for Numerical Methods in Fluids*, 2018.

[38] S. Chandrasekhar. *An introduction to the study of stellar structure*, volume 2. Courier Corporation, 1958.

[39] S. Chandrasekhar. *Hydrodynamic and Hydromagnetic Stability*. Clarendon Press, Oxford, 1961.

[40] P. Chandrashekar. Kinetic energy preserving and entropy stable finite volume schemes for compressible Euler and Navier-Stokes equations. *Communications in Computational Physics*, 14(5):1252–1286, 2013.

[41] P. Chandrashekar and C. Klingenberg. A second order well-balanced finite volume scheme for Euler equations with gravity. *SIAM Journal on Scientific Computing*, 37(3):B382–B402, 2015.

[42] P. Chandrashekar and M. Zenk. Well-balanced nodal discontinuous Galerkin method for Euler equations with gravity. *Journal of Scientific Computing*, pages 1–32, 2017.

[43] A. Chertock, S. Cui, A. Kurganov, Ş. N. Özcan, and E. Tadmor. Well-balanced schemes for the Euler equations with gravitation: Conservative formulation using global fluxes. *Journal of Computational Physics*, 2018.

[44] E. T. Chung and B. Engquist. Convergence analysis of fully discrete finite volume methods for maxwell's equations in nonhomogeneous media. *SIAM Journal on Numerical Analysis*, 43(1):303–317, 2005.

[45] P. Colella, M. R. Dorr, J. A. Hittinger, and D. F. Martin. High-order, finite-volume methods in mapped coordinates. *Journal of Computational Physics*, 230(8):2952–2976, 2011.

[46] P. Colella and P. R. Woodward. The Piecewise Parabolic Method (PPM) for gas-dynamical simulations. *Journal of Computational Physics*, 54:174–201, Sept. 1984.

[47] I. Cravero, G. Puppo, M. Semplice, and G. Visconti. CWENO: uniformly accurate reconstructions for balance laws. *Mathematics of Computation*, 87(312):1689–1719, 2018.

[48] M. P. Crosland. The origins of Gay-Lussac's law of combining volumes of gases. *Annals of science*, 17(1):1–26, 1961.

[49] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics, Jan. 1996.

[50] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. A well-balanced scheme for the Euler equation with a gravitational potential. In *Finite Volumes for Complex Applications VII-Methods and Theoretical Aspects*, pages 217–226. Springer, 2014.

[51] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. A well-balanced scheme to capture non-explicit steady states in the Euler equations with gravity. *International Journal for Numerical Methods in Fluids*, 81(2):104–127, 2016.

[52] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. Well-balanced schemes to capture non-explicit steady states: Ripa model. *Mathematics of Computation*, 85(300):1571–1602, 2016.

[53] S. Diot, S. Clain, and R. Loubère. A high-order finite volume method for hyperbolic systems: Multi-dimensional Optimal Order Detection (MOOD). *J. Comput. Phys*, 230:4028–4050, 2011.

[54] J. R. Dormand and P. J. Prince. A family of embedded Runge-Kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

[55] M. Dumbser, M. Käser, V. A. Titarev, and E. F. Toro. Quadrature-free non-oscillatory finite volume schemes on unstructured meshes for nonlinear hyperbolic systems. *Journal of Computational Physics*, 226(1):204–243, 2007.

[56] D. R. Durran. *Numerical Methods for Fluid Dynamics*. Springer, 2010.

[57] P. V. F. Edelmann. *Coupling of Nuclear Reaction Networks and Hydrodynamics for Application in Stellar Astrophysics*. Dissertation, Technische Universität München, 2014.

[58] P. V. F. Edelmann, L. Horst, J. P. Berberich, R. Andrassy, J. Higl, C. Klingenberg, and F. Röpke. Well-balanced treatment of gravity in astrophysical hydrodynamics codes. *In Preparation*, 2020.

[59] L. Euler. Principes généraux du mouvement des fluides. *Mémoires de l'Académie des Sciences de Berlin*, pages 274–315, 1757.

[60] C. R. Evans and J. F. Hawley. Simulation of magnetohydrodynamic flows-a constrained transport method. *The Astrophysical Journal*, 332:659–677, 1988.

[61] T. A. Eymann and P. L. Roe. Multidimensional active flux schemes. In *21st AIAA computational fluid dynamics conference*, page 2940, 2013.

[62] T. Feagin. A tenth-order Runge–Kutta method with error estimate. In *Proceedings of the IAENG Conference on Scientific Computing*, 2007.

[63] E. Franck and L. S. Mendoza. Finite volume scheme with local high order discretization of the hydrostatic equilibrium for the euler equations with external forces. *Journal of Scientific Computing*, 69(1):314–354, 2016.

[64] K. O. Friedrichs and P. D. Lax. Systems of conservation equations with a convex extension. *Proceedings of the National Academy of Sciences*, 68(8):1686–1688, 1971.

[65] F. G. Fuchs, S. Mishra, and N. H. Risebro. Splitting based finite volume schemes for ideal MHD equations. *Journal of Computational Physics*, 228(3):641–660, 2009.

[66] P. Fullick. *Physics*. Heinemann, 1994.

[67] E. Gaburro. *Well balanced Arbitrary-Lagrangian-Eulerian Finite Volume schemes on moving nonconforming meshes for non-conservative Hyperbolic systems*. PhD thesis, University of Trento, 2018.

[68] E. Gaburro. A unified framework for the solution of hyperbolic PDE systems using high order direct Arbitrary-Lagrangian-Eulerian schemes on moving unstructured meshes with topology change. *Archives of Computational Methods in Engineering*, 2020.

[69] E. Gaburro, M. J. Castro, and M. Dumbser. Well-balanced Arbitrary-Lagrangian-Eulerian finite volume schemes on moving nonconforming meshes for the Euler equations of gas dynamics with gravity. *Monthly Notices of the Royal Astronomical Society*, 477(2):2251–2275, 2018.

[70] E. Gaburro, M. Dumbser, and M. J. Castro. Direct Arbitrary-Lagrangian-Eulerian finite volume schemes on moving nonconforming unstructured meshes. *Computers & Fluids*, 159:254–275, 2017.

[71] C. F. Gauss. *Methodus nova integralium valores per approximationem inveniendi*. apvd Henricvm Dieterich, 1815.

[72] D. Ghosh and E. M. Constantinescu. Well-balanced, conservative finite difference algorithm for atmospheric flows. *AIAA Journal*, 2016.

[73] T. Gimse and N. H. Risebro. Solution of the cauchy problem for a conservation law with a discontinuous flux function. *SIAM Journal on Mathematical Analysis*, 23(3):635–648, 1992.

[74] F. X. Giraldo and M. Restelli. A study of spectral element and discontinuous Galerkin methods for the Navier–Stokes equations in nonhydrostatic mesoscale atmospheric modeling: Equation sets and test cases. *Journal of Computational Physics*, 227(8):3849–3877, 2008.

[75] S. K. Godunov. Finite difference method for numerical computation of discontinous solution of the equations of fluid dynamics. *Matematicheskii Sbornik*, 47:271, 1959.

[76] S. K. Godunov. The problem of a generalized solution in the theory of quasilinear equations and in gas dynamics. *Russian Mathematical Surveys*, 17(3):145, 1962.

[77] P. M. Gresho and S. T. Chan. On the theory of semi-implicit projection methods for viscous incompressible flow and its implementation via a finite element method that also introduces a nearly consistent mass matrix. part

2: Implementation. *International Journal for Numerical Methods in Fluids*, 11(5):621–659, 1990.

[78] L. Grosheintz-Laval and R. Käppeli. High-order well-balanced finite volume schemes for the Euler equations with gravitation. *Journal of Computational Physics*, 378:324–343, 2019.

[79] H. Guillard and C. Viozat. On the behaviour of upwind schemes in the low mach number limit. *Computers & Fluids*, 28(1):63 – 86, 1999.

[80] V. Guinot. An approximate two-dimensional riemann solver for hyperbolic systems of conservation laws. *Journal of Computational Physics*, 205(1):292–314, 2005.

[81] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy. Uniformly high order accurate essentially non-oscillatory schemes, iii. In *Upwind and high-resolution schemes*, pages 218–290. Springer, 1987.

[82] A. Harten and S. Osher. Uniformly high-order accurate nonoscillatory schemes. i. *SIAM Journal on Numerical Analysis*, 24(2):279–309, 1987.

[83] C. Hirsch. *Numerical computation of internal and external flows: The fundamentals of computational fluid dynamics*. Butterworth-Heinemann, 2007.

[84] H. Holden and N. H. Risebro. *Front tracking for hyperbolic conservation laws*, volume 152. Springer, 2015.

[85] M. Hosea and L. Shampine. Analysis and implementation of tr-bdf2. *Applied Numerical Mathematics*, 20(1-2):21 – 37, 1996. Method of Lines for Time-Dependent Problems.

[86] S. Ii, B. Xie, and F. Xiao. An interface capturing method with a continuous function: The THINC method on unstructured triangular and tetrahedral meshes. *Journal of Computational Physics*, 259:260–269, 2014.

[87] W. R. Inc. Mathematica, version 11.0. Champaign, IL, 2016.

[88] C. G. J. Jacobi. Ueber Gauß neue Methode, die Werthe der Integrale näherungsweise zu finden. *Journal für die reine und angewandte Mathematik*, 1826(1):301–308, 1826.

[89] A. Jameson. Formulation of kinetic energy preserving conservative schemes for gas dynamics and direct numerical simulation of one-dimensional viscous compressible flow in a shock tube using entropy and kinetic energy preserving schemes. *Journal of Scientific Computing*, 34(2):188–208, 2008.

[90] A. Jameson, W. Schmidt, and E. Turkel. Numerical solution of the Euler equations by finite volume methods using Runge Kutta time stepping schemes. June 1981.

[91] G.-S. Jiang and C.-W. Shu. Efficient implementation of weighted ENO schemes. *Journal of computational physics*, 126(1):202–228, 1996.

[92] R. Käppeli and S. Mishra. Well-balanced schemes for the Euler equations with gravitation. *Journal of Computational Physics*, 259:199–219, 2014.

[93] R. Käppeli and S. Mishra. *Well-balanced Schemes for Gravitationally Stratified Media*, volume 498 of *Astronomical Society of the Pacific Conference Series*, page 210. 2015.

[94] R. Käppeli and S. Mishra. A well-balanced finite volume scheme for the Euler equations with gravitation-the exact preservation of hydrostatic equilibrium with arbitrary entropy stratification. *Astronomy & Astrophysics*, 587:A94, 2016.

[95] K. Kifonidis and E. Müller. On multigrid solution of the implicit equations of hydrodynamics. experiments for the compressible Euler equations in general coordinates. *A&A*, 544:A47, Aug. 2012.

[96] C. Klingenberg, G. Puppo, and M. Semplice. Arbitrary order finite volume well-balanced schemes for the Euler equations with gravity. *SIAM Journal on Scientific Computing*, 41(2):A695–A721, 2019.

[97] O. Kolb. On the full and global accuracy of a compact third order WENO scheme. *SIAM Journal on Numerical Analysis*, 52(5):2335–2355, 2014.

[98] J. F. B. M. Kraaijevanger. Contractivity of Runge–Kutta methods. *BIT Numerical Mathematics*, 31(3):482–528, 1991.

[99] A. Kurganov and E. Tadmor. Solution of two-dimensional Riemann problems for gas dynamics without Riemann problem solvers. *Numerical Methods for Partial Differential Equations. An International Journal*, 18(5):584–608, 2002.

[100] P. Lax and B. Wendroff. Systems of conservation laws. comm. pure appl. math., 13. 1960.

[101] R. LeVeque and D. Bale. Wave propagation methods for conservation laws with source terms. *Birkhauser Basel*, pages pp. 609–618, 1999.

[102] R. J. LeVeque. Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. *Journal of computational physics*, 146(1):346–365, 1998.

[103] R. J. LeVeque. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge university press, 2002.

[104] R. J. LeVeque, D. L. George, and M. J. Berger. Tsunami modelling with adaptively refined finite volume methods. *Acta Numerica*, 20:211–289, 2011.

[105] D. Levy, G. Puppo, and G. Russo. Central WENO schemes for hyperbolic systems of conservation laws. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 33(3):547–571, 1999.

[106] D. Levy, G. Puppo, and G. Russo. Compact central WENO schemes for multidimensional conservation laws. *SIAM Journal on Scientific Computing*, 22(2):656–672, 2000.

[107] D. Levy, G. Puppo, and G. Russo. On the behavior of the total variation in CWENO methods for conservation laws. *Applied Numerical Mathematics*, 33(1-4):407–414, 2000.

[108] D. Levy and E. Tadmor. From semidiscrete to fully discrete: Stability of runge–kutta schemes by the energy method. *SIAM review*, 40(1):40–73, 1998.

[109] G. Li, V. Caleffi, and Z. Qi. A well-balanced finite difference WENO scheme for shallow water flow model. *Applied Mathematics and Computation*, 265:1–16, 2015.

[110] W. Li. Accuracy preserving limiters for high-order finite volume methods. In *Efficient Implementation of High-Order Accurate Numerical Methods on Unstructured Grids*, pages 37–91. Springer, 2014.

[111] X.-s. Li. Uniform algorithm for all-speed shock-capturing schemes. *International Journal of Computational Fluid Dynamics*, 28(6-10):329–338, 2014.

[112] X.-s. Li and C.-w. Gu. An all-speed roe-type scheme and its asymptotic analysis of low mach number behaviour. *Journal of Computational Physics*, 227(10):5144–5159, 2008.

[113] X.-s. Li, X.-d. Ren, and C.-w. Gu. An improved roe scheme for all mach-number flows simultaneously curing known problems. *arXiv preprint arXiv:1711.09272*, 2017.

[114] M.-S. Liou. A sequel to ausm, part ii: Ausm$^+$-up for all speeds. *Journal of Computational Physics*, 214(1):137–170, 2006.

[115] X.-D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *Journal of Computational Physics*, 115:200–212, Nov. 1994.

[116] X. Lu, B. Dong, B. Mao, and X. Zhang. A robust and well-balanced numerical model for solving the two-layer shallow water equations over uneven topography. *Comptes Rendus Mécanique*, 343(7-8):429–442, 2015.

[117] R. Menikoff and B. J. Plohr. The Riemann problem for fluid flow of real materials. *Reviews of Modern Physics*, 61(1):75–130, Jan. 1989.

[118] D. Meschede. *Gerthsen physik*. Springer-Verlag, 2015.

[119] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography. *Computers & Mathematics with Applications*, 72(3):568–593, 2016.

[120] F. Miczek. *Simulation of low Mach number astrophysical flows*. Dissertation, Technische Universität München, 2013.

[121] H. Nishikawa. A face-area-weighted centroid formula for reducing grid skewness and improving convergence of edge-based solver on highly-skewed simplex grids. In *AIAA Scitech 2020 Forum*, page 1786, 2020.

[122] S. Noelle, N. Pankratz, G. Puppo, and J. R. Natvig. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *Journal of Computational Physics*, 213(2):474–499, 2006.

[123] S. Noelle, Y. Xing, and C.-W. Shu. High-order well-balanced finite volume WENO schemes for shallow water equation with moving water. *Journal of Computational Physics*, 226(1):29–58, 2007.

[124] S. Osher. Riemann solvers, the entropy condition, and difference. *SIAM Journal on Numerical Analysis*, 21(2):217–235, 1984.

[125] S. Osher and S. Chakravarthy. High resolution schemes and the entropy condition. *SIAM Journal on Numerical Analysis*, 21(5):955–984, 1984.

[126] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *Journal of Computational Physics*, 79:12–49, Nov. 1988.

[127] K. Oßwald, A. Siegmund, P. Birken, V. Hannemann, and A. Meister. L2roe: a low dissipation version of roe's approximate riemann solver for low mach numbers. *International Journal for Numerical Methods in Fluids*, 2015. fld.4175.

[128] K. Oßwald, A. Siegmund, P. Birken, V. Hannemann, and A. Meister. L2roe: a low dissipation version of roe's approximate riemann solver for low mach numbers. *International Journal for Numerical Methods in Fluids*, 81(2):71–86, 2016.

[129] C. Parés. Numerical methods for nonconservative hyperbolic systems: a theoretical framework. *SIAM Journal on Numerical Analysis*, 44(1):300–321, 2006.

[130] L. Pareschi and T. Rey. Residual equilibrium schemes for time dependent partial differential equations. *Computers & Fluids*, 156:329–342, 2017.

[131] W. Pauli. Pauli lectures on physics: thermodynamics and the kinetic theory of gas, vol. 3, 1973.

[132] B. Perthame and C.-W. Shu. On positivity preserving finite volume schemes for compressible euler equations. Technical report, Institute for Computer Applications in Science and Engineering Hampton VA, 1993.

[133] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*, volume 3. Cambridge University Press, 2007.

[134] F. Qu, D. Sun, and J. Bai. A new genuinely two-dimensional riemann solver for multidimensional euler and navier–stokes equations. *Computer Physics Communications*, 243:1–11, 2019.

[135] A. Quarteroni and A. Valli. *Numerical approximation of partial differential equations*, volume 23. Springer Science & Business Media, 2008.

[136] F. Rabiei and F. Ismail. Fifth-order improved Runge–Kutta method for solving ordinary differential equation. *Australian Journal of Basic and Applied Sciences*, 6(3):97–105, 2012.

[137] D. Ray and P. Chandrashekar. Entropy stable schemes for compressible Euler equations. *Int. J. Numer. Anal. Model. Ser. B*, 4(4):335–352, 2013.

[138] D. Ray, P. Chandrashekar, U. S. Fjordholm, and S. Mishra. Entropy stable scheme on two-dimensional unstructured grids for Euler equations. *Communications in Computational Physics*, 19(5):1111–1140, 2016.

[139] T. Richter and T. Wick. *Einführung in die Numerische Mathematik: Begriffe, Konzepte und zahlreiche Anwendungsbeispiele*. Springer-Verlag, 2017.

[140] F. Rieper. A low-Mach number fix for Roe's approximate Riemann solver. *Journal of Computational Physics*, 230(13):5263–5287, 2011.

[141] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43(2):357 – 372, 1981.

[142] P. L. Roe. Affordable, entropy consistent flux functions. In *Eleventh International Conference on Hyperbolic Problems: Theory, Numerics and Applications, Lyon*, 2006.

[143] F. Röpke, J. Berberich, S. Jones, A. Botto Poala, P. Edelmann, A. Michel, and L. Horst. Towards multidimensional hydrodynamic simulations of stars. In *NIC Symposium 2018*, number FZJ-2018-02959. John von Neumann-Institut für Computing, 2018.

[144] J. A. Rossmanith. An unstaggered, high-resolution constrained transport method for magnetohydrodynamic flows. *SIAM Journal on Scientific Computing*, 28(5):1766–1797, 2006.

[145] V. V. Rusanov. Calculation of intersection of non-steady shock waves with obstacles. *J. Comput. Math. Phys. USSR*, 1:267–279, 1961.

[146] W. E. Schiesser. *The numerical method of lines: integration of partial differential equations*. Elsevier, 2012.

[147] G. A. Sod. A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *Journal of computational physics*, 27(1):1–31, 1978.

[148] D. J. Struik. *A source book in mathematics, 1200-1800*, volume 11. Harvard University Press, 1969.

[149] P. K. Subbareddy and G. V. Candler. A fully discrete, kinetic energy consistent finite-volume scheme for compressible flows. *Journal of Computational Physics*, 228(5):1347–1364, 2009.

[150] E. Tadmor. Entropy functions for symmetric systems of conservation laws. 1983.

[151] E. Tadmor. Numerical viscosity and the entropy condition for conservative difference schemes. *Mathematics of Computation*, 43(168):369–381, 1984.

[152] E. Tadmor. The numerical viscosity of entropy stable schemes for systems of conservation laws. i. *Mathematics of Computation*, 49(179):91–103, 1987.

[153] C.-H. Tai, J.-H. Sheu, and P.-Y. Tzeng. Improvement of explicit multistage schemes for central spatial discretization. *AIAA journal*, 34(1):185–188, 1996.

[154] C.-H. Tai, J.-H. Sheu, and B. Van Leer. Optimal multistage schemes for euler equations with residual smoothing. *AIAA journal*, 33(6):1008–1016, 1995.

[155] B. Taylor. *Methodus incrementorum directa & inversa.* typis Pearsonianis: prostant apud Gul. Innys ad Insignia Principis in Coemeterio Paulino, 1715.

[156] A. Thomann, M. Zenk, and C. Klingenberg. A second-order positivity-preserving well-balanced finite volume scheme for Euler equations with gravity for arbitrary hydrostatic equilibria. *International Journal for Numerical Methods in Fluids*, 89(11):465–482, 2019.

[157] J. F. Thompson, Z. U. Warsi, and C. W. Mastin. *Numerical grid generation: foundations and applications*, volume 45. North-holland Amsterdam, 1985.

[158] J. F. Thompson, Z. U. A. Warsi, and C. W. Mastin. Boundary-fitted coordinate systems for numerical solution of partial differential equations – a review. *Journal of Computational Physics*, 47:1, July 1982.

[159] B. Thornber and D. Drikakis. Numerical dissipation of upwind schemes in low mach flow. *International journal for numerical methods in fluids*, 56(8):1535–1541, 2008.

[160] F. X. Timmes and D. Arnett. The accuracy, consistency, and speed of five equations of state for stellar hydrodynamics. *ApJS*, 125:277–294, Nov. 1999.

[161] F. X. Timmes and F. D. Swesty. The accuracy, consistency, and speed of an electron-positron equation of state based on table interpolation of the Helmholtz free energy. *ApJS*, 126:501–516, Feb. 2000.

[162] A. Tokunaga and A. N. Cox. Allen's astrophysical quantities. *AN Cox (Springer)*, page 143, 2000.

[163] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction*. Springer, Berlin Heidelberg, 2009.

[164] E. F. Toro, M. Spruce, and W. Speares. Restoration of the contact surface in the HLL-Riemann solver. *Shock waves*, 4(1):25–34, 1994.

[165] G. Tóth. The $\nabla.B = 0$ Constraint in Shock-Capturing Magnetohydrodynamics Codes. *Journal of Computational Physics*, 161:605–652, July 2000.

[166] R. Touma and C. Klingenberg. Well-balanced central finite volume methods for the Ripa system. *Applied Numerical Mathematics*, 97:42–68, 2015.

[167] R. Touma, U. Koley, and C. Klingenberg. Well-balanced unstaggered central schemes for the Euler equations with gravitation. *SIAM Journal on Scientific Computing*, 38(5):B773–B807, 2016.

[168] E. Turkel. Preconditioned methods for solving the incompressible and low speed compressible equations. *Journal of computational physics*, 72(2):277–298, 1987.

[169] University of Houston, Clear Lake. High-order explicit Runge–Kutta methods, url: `https://sce.uhcl.edu/rungekutta/`, accessed August 24, 2020. Coefficients of the RK10 method introduced in [62].

[170] B. Va, C.-H. Tai, and K. Powell. Design of optimally smoothing multi-stage schemes for the Euler equations. In *9th Computational Fluid Dynamics Conference*, page 1933, 1981.

[171] D. Varma and P. Chandrashekar. A second-order, discretely well-balanced finite volume scheme for Euler equations with gravity. *Computers & Fluids*, 2019.

[172] Y. Wang and J. Zhu. A new type of increasingly high-order multi-resolution trigonometric WENO schemes for hyperbolic conservation laws and highly oscillatory problems. *Computers & Fluids*, 200:104448, 2020.

[173] Z. Wang, J. Zhu, and N. Zhao. A new fifth-order finite difference well-balanced multi-resolution WENO scheme for solving shallow water equations. *Computers & Mathematics with Applications*, 80(5):1387–1404, 2020.

[174] A. Wongwathanarat, H. Grimm-Strele, and E. Müller. Apsara: A multi-dimensional unsplit fourth-order explicit eulerian hydrodynamics code for arbitrary curvilinear grids. *Astronomy & Astrophysics*, 595:A41, 2016.

[175] Y. Xing and C.-W. Shu. High order well-balanced WENO scheme for the gas dynamics equations under gravitational fields. *Journal of Scientific Computing*, 54(2-3):645–662, 2013.

[176] Y. Xing, C.-W. Shu, and S. Noelle. On the advantage of well-balanced schemes for moving-water equilibria of the shallow water equations. *Journal of scientific computing*, 48(1-3):339–349, 2011.

[177] L. Zhang, B. Ai, and Z. Chen. New multi-dimensional limiter for finite volume discretizations on unstructured meshes. In *Asia-Pacific International Symposium on Aerospace Technology*, pages 630–642. Springer, 2018.

[178] J. Zhu and C.-W. Shu. A new type of multi-resolution WENO schemes with increasingly higher order of accuracy. *Journal of Computational Physics*, 375:659–683, 2018.