Low Mach number finite volume methods for the acoustic and Euler equations

PhD thesis

Wasilij Barsukow

2018

F

Abstract

Finite volume methods for compressible Euler equations suffer from an excessive diffusion in the limit of low Mach numbers. This PhD thesis explores new approaches to overcome this.

The analysis of a simpler set of equations that also possess a low Mach number limit is found to give valuable insights. These equations are the acoustic equations obtained as a linearization of the Euler equations. For both systems the limit is characterized by a divergencefree velocity. This constraint is nontrivial only in multiple spatial dimensions. As the Jacobians of the acoustic system do not commute, acoustics cannot be reduced to some kind of multi-dimensional advection. Therefore first an exact solution in multiple spatial dimensions is obtained. It is shown that the low Mach number limit can be interpreted as a limit of long times.

It is found that the origin of the inability of a scheme to resolve the low Mach number limit is the lack a discrete counterpart to the limit of long times. Numerical schemes whose discrete stationary states discretize all the analytic stationary states of the PDE are called stationarity preserving. It is shown that for the acoustic equations, stationarity preserving schemes are vorticity preserving and are those that are able to resolve the low Mach limit (low Mach compliant). This establishes a new link between these three concepts.

Stationarity preservation is studied in detail for both dimensionally split and multi-dimensional schemes for linear acoustics. In particular it is explained why the same multi-dimensional stencils appear in literature in very different contexts: These stencils are unique discretizations of the divergence that allow for stabilizing stationarity preserving diffusion.

Stationarity preservation can also be generalized to nonlinear systems such as the Euler equations. Several ways how such numerical schemes can be constructed for the Euler equations are presented. In particular a low Mach compliant numerical scheme is derived that uses a novel construction idea. Its diffusion is chosen such that it depends on the velocity divergence rather than just derivatives of the different velocity components. This is demonstrated to overcome the low Mach number problem. The scheme shows satisfactory results in numerical simulations and has been found to be stable under explicit time integration.

Contents

Acknowledgments

I have experienced a lot of help and support during my dissertation project. Particularly I would like to thank my supervisor Christian Klingenberg for his kind and constant support. When I joined his workgroup I felt welcome and throughout my time in Würzburg I have enjoyed the freedom to question the obvious and to try out ideas with an uncertain outcome. Also, the joint work with Fritz Röpke and Philipp Edelmann was a huge inspiration for me. I have met Philipp during my stay at MPA and have heard from him of the low Mach number problem. Those days, upon me expressing the wish to do a PhD in mathematics, he brought to my attention the research done in Würzburg.

The numerous visitors of our group were a constant enrichment and it was a pleasure to meet them. Among my fellow PhD students I particularly enjoyed and have profited from numerous discussions with Markus Zenk, Marlies Pirner, Simon Markfelder, Jens Klotzky and Jonas Berberich. Also, I am happy that Jonathan Hohm and Lukas Thanhäuser joined our workgroup and I would like to thank them for sharing their ideas with me. I also thank Marlies for carefully reading the thesis draft.

The German National Academic Foundation (Studienstiftung des Deutschen Volkes) kindly supported me throughout the work on my dissertation. I am grateful for their constant financial support and for paying for my numerous conference trips. Invaluable for me was the intense exchange with other students of the foundation. With scholars focusing on ever narrower domains, seemingly what a challenge to bring together people from all the domains of natural sciences, humanities and arts and make them talk to each other. In my experience with the meetings organized by the foundation this never was a challenge at all. And so I am grateful to the foundation of having brought me together with classical philologists, veterinary doctors, linguists, polar researchers, geologists, historians, economists, lawyers, psychologists, biologists, fellow mathematicians and physicists.

Introduction

Ideal hydrodynamics is a theory of conservation laws. The motion of an ideal fluid is entirely determined by mass, momentum and energy conservation. These conservation laws are called the Euler equations. Even for smooth initial data the solution can develop jumps. It can be shown also across these discontinuities the above quantities are conserved. By the Lax-Wendroff theorem, numerical methods that are conservative in the discrete sense are the ones that are able to converge. Thus, conservativity is also the fundamental concept for numerical methods, which are referred to as finite volume methods. The discrete degrees of freedom are volume averages over cells, and their time evolution is determined by fluxes through the boundaries. Thus, whatever flows out of a computational cell enters its neighbouring cells.

The Euler equations form a hyperbolic system of partial differential equations. Information travels with finite speed, that in general depends on the direction. One thus can identify regions that can influence a given location $\mathbf{x}$ at a given time t, and those that are unable to do so. The former are referred to as causally connected to $(t, \mathbf{x})$. This is another fundamental property of the equations that has a counterpart in numerics. For stability of the numerical method, only those discrete values have to be involved in an update procedure that are causally connected to the cell that is being updated. This is referred to as upwinding. Thus a technical aspect of a numerical scheme (stability) can be traced back to its origin in a fundamental property of the equations (causality).

A natural question is what other properties of the equations a numerical scheme should reflect. On the other hand one might try to understand deficiencies of numerical methods by making a link to a violation of some properties of the equations. This thesis shows how both approaches can be fruitfully combined and lead to a better understanding of the so-called low Mach number problem.

In the context of finite volume schemes one faces a dichotomy: certain schemes are observed to have the more numerical artefacts the lower the Mach number of the flow is, and certain schemes are found not to have this property. These latter have only been discovered in the late 1980s and are called low Mach compliant schemes. In all cases the quality of the simulations increases with resolution. Therefore the statement can be rephrased as follows: For a given simulation quality, numerical schemes that are not low Mach compliant require finer and finer grids the lower the Mach number of

the flow is. This is impractical (and at some point prohibitive), such that a complete understanding of low Mach compliant schemes is an avenue towards hugely increased quality of numerical simulations of fluid flow.

Even inside the class of finite volume schemes there is large variety. One of the most prominent members of this class are Godunov schemes. They make use of an exact short-time solution of the equations as a building block for the numerical scheme. Thus Godunov methods inherit important properties of the equations, such as entropy dissipation and stability. Interestingly, the Godunov method fails in capturing the low Mach number flow correctly, i.e. it is not low Mach compliant.

There exist a number of numerical schemes that do not involve the exact solution, but rather some approximation to it (approximate Riemann solvers). Methods can also be derived by replacing certain parts of a given scheme by carefully chosen values in order achieve particular aims. Examples of such an approach are many low Mach compliant methods (low Mach fixes). They lack a derivation from some fundamental basis, and may have or have not an optimal stability range. They also often come with free parameters.

Therefore the puzzling situation is that those finite volume schemes that are derived from fundamental considerations fail to be low Mach compliant, and those that are, lack a satisfactory derivation from first principles.

In order to study this problem, in this thesis another system of equations is considered first. It has a very similar limit of low Mach number, and numerical schemes even show visually similar numerical artefacts. This system is called the acoustic equations, and has the very useful property of being linear. It is obtained as a linearization of the Euler equations, and thus shows that the low Mach number problem is not a prerogative of nonlinear equations or schemes. The starting point of the analysis is a careful study of the numerical artefacts that arise when the acoustic equations are solved using the corresponding Godunov scheme and the upwind/Roe scheme. In this analysis one observes that the numerical solution is the same, whether one changes the Mach number by a given factor, or the simulation time. Indeed for the acoustic equations, and this is shown in detail in this thesis, the limit of long time is the same as the limit of low Mach number.

Thus the artefacts of a scheme that is not low Mach compliant can be entirely understood as its failure to properly resolve those solutions that remain after long times, i.e. stationary states. The stationary states of the acoustic equations are divergenceless; the only stationary states of the upwind/Roe scheme, for instance, are shear flows. These latter are divergenceless, but not all divergenceless flows are shear flows, and this scheme fails to discretize all the stationary states of the equations. To make such notions precise is one of the topics of this thesis. Schemes that discretize all of the analytic stationary states are termed stationarity preserving, and it is shown that for the acoustic equations they are exactly those that are low Mach compliant.

This new approach leads to further new insights. Indeed, stationarity preservation can be shown to be but a mirror image of another concept that has been studied in the literature before: vorticity preservation. For both the Euler and the acoustic equations, vorticity (i.e. the curl of the velocity) is an important concept, and one might wish the numerical scheme to have a discrete counterpart to vorticity, and this discrete counterpart

to fulfill a particular equation that reflects the analytic evolution of vorticity. Such schemes are called vorticity preserving. For the acoustic equations, they can be shown to be stationarity preserving, and thus low Mach compliant. This links three very different concepts.

The methods described in this thesis are not restricted to the equations of linear acoustics. First of all they can be applied to all multi-dimensional hyperbolic linear systems. A typical occurrence of stationary solutions is the balance between a differential operator and a source term. When trying to evolve numerically a setup close to an exact stationary solution one faces the problem that the numerical discretization does not quite match the source. This makes the numerical solution evolve in time, rather than remain stationary. The problem of finding discretizations that achieve stationarity of the discrete solution in the context of source terms is called well-balancing. Ideas of stationarity preservation easily can be applied to such a situation and for the acoustic equations, augmented by a gravity-like source term, they are shown to lead to simple and convincing arguments.

There are several direct consequences of these studies that can be drawn for the Euler equation. First, stationarity preservation gives a necessary condition for low Mach compliance of a scheme for the nonlinear Euler equations. Indeed, such a scheme should be low Mach compliant already in the linearized regime, in which the evolution is governed by the acoustic equations. Such a linearized scheme thus has to be stationarity preserving. The usual approach to the behaviour of nonlinear schemes in the low Mach number limit is via formal asymptotic analysis. These are not rigorous statements, but they can also be applied to linear acoustics. This allows to compare the results to predictions of stationarity preservation, and to judge the quality of arguments that involve asymptotic analysis.

A number of construction principles for stationarity preserving schemes that are developed for linear acoustics can be applied to the Euler equations. In a nutshell, this is possible because via the Leibniz rule a derivative of a (nonlinear) product can be rewritten as terms that involve derivatives of the individual factors, and because discrete counterparts to such Leibniz rules exist. This thesis presents three different construction strategies for schemes for the Euler equations, a list that is by no means exhaustive. It is clear that not every part of an argumentation that is based on linear arguments can be carried over to a nonlinear setting, but these examples show that linear examples can give a good deal of guidance for much more complicated situations. The resulting schemes are meant to exemplify novel ways how numerical schemes for the multi-dimensional Euler equations can be constructed and hopefully contribute to a more thorough exploration of what is possible.

The thesis is structured as follows: After a discussion of the Euler equations in Section 1 and the acoustic equations in Section 2, the concept of stationarity preservation for linear systems is introduced in Section 3. It contains a generalization of the equivalence to vorticity preservation, as well as construction principles for multi-dimensional schemes, that can be formulated generally. In Section 4, numerical schemes for linear acoustics are considered. The concept of stationarity preservation and implications for the limit of low Mach number are discussed. The multi-dimensional Godunov scheme is derived, and it

is shown that it fails to be low Mach compliant. It is shown for both dimensionally split and multi-dimensional schemes how low Mach compliance can be achieved. Particularly in case of the former this has an influence on the stability properties, such that linear stability of a certain class of schemes is studied. Section 5 finally demonstrates how these construction principles can be applied to the Euler equations. It is followed by conclusions and an outlook.

Conventions

The numerical methods in this thesis are considered on equidistant Cartesian grids, mostly in two spatial dimensions. The following notation is used:

Definition 0.1. *i) A d-dimensional* grid *is a tiling of $\mathbb{R}^d$ with countably many polygons/polyhedra, called* cells*. A* rectangular grid *is a tiling that uses only rectangles in a way such that adjacent rectangles always share one full side. A smoothly deformed rectangular grid is called a* structured grid*, and is a generalization of a rectangular grid. A particular rectangular grid whose cells are all congruent is called* Cartesian*. If the rectangles are actually squares, the grid is called* square grid*.*

ii) The sides of the rectangular cells of a d-dimensional Cartesian grid are generically denoted by $\Delta x \equiv \Delta x_1$, $\Delta y \equiv \Delta x_2$, $\Delta z \equiv \Delta x_3$ ($d \leq 3$).

iii) Cells of a d-dimensional rectangular/structured grid are indexed by elements of $\mathbb{Z}^d$.

iv) If not stated differently, q_i^n is the value of the function q in cell $i \in \mathbb{Z}$ of a one-dimensional grid at time step $n \in \mathbb{N}_0$. Analogously, q_{ij}^n is the value of the function q in cell $(i,j) \in \mathbb{Z}^2$ of a two-dimensional grid at time step $n \in \mathbb{N}_0$.

v) Boundaries of a d-dimensional cell of a rectangular grid are indexed by $\left(\mathbb{Z}+\frac{1}{2}\right) \times \mathbb{Z}^{d-1}$, $\mathbb{Z} \times \left(\mathbb{Z}+\frac{1}{2}\right) \times \mathbb{Z}^{d-2}$, …, $\mathbb{Z}^{d-1} \times \left(\mathbb{Z}+\frac{1}{2}\right)$. They are referred to as cell interfaces*.*

Note that in this thesis indices never denote derivatives; the convention of summing over repeated indices is adapted wherever they occur. Also when matrices are specified, in order to improve readability often only the nonvanishing entries are given. The imaginary unit is denoted by $\mathbb{i}$ and the identity map/matrix by $\mathbb{1}$.

In order to cope with the lengthy expressions for numerical schemes, the following notation is used:

Definition 0.2.

$$[q]_{i+\frac{1}{2}} := q_{i+1} - q_i \qquad \{q\}_{i+\frac{1}{2}} := q_{i+1} + q_i$$

$$[q]_{i\pm1} := q_{i+1} - q_{i-1}$$

$$[[q]]_{i\pm\frac{1}{2}} := [q]_{i+\frac{1}{2}} - [q]_{i-\frac{1}{2}} \qquad \{\{q\}\}_{i\pm\frac{1}{2}} := \{q\}_{i+\frac{1}{2}} + \{q\}_{i-\frac{1}{2}}$$

The only nontrivial identity is

$$\{[q]\}_{i\pm\frac{1}{2}} = [q]_{i+\frac{1}{2}} + [q]_{i-\frac{1}{2}} = [q]_{i\pm1}$$

For multiple dimensions the notation is combined, e.g.

$$[[q]_{i\pm1}]_{j\pm1} = q_{i+1,j+1} - q_{i-1,j+1} - q_{i+1,j-1} + q_{i-1,j-1}$$

The brackets for different directions commute.

Chapter 1

Euler equations of hydrodynamics

Contents

1.1 The Euler equations

1.1.1 Introductory remarks

The derivation of the equations of motion of a fluid encounters a number of difficulties: the fluids under consideration always consist of individual interacting particles with their own (atomic/molecular/...) substructure. However, the size of the particles and also the length scales of the inter-particle distances are often much smaller than length scales of interest in an experimental situation. One thus might be tempted to interpret, say, the velocity measured by an experimental device as some spatial average of the particle velocities in some small volume around the point where the measurement was taken. One thus would be able to assign a macroscopic velocity to any point in space by suitably averaging the particle velocities in some volume around it. This is called a continuum description of the fluid. However this transition is highly nontrivial and it is worth being thought over and over again.

A continuum description seems to reflect the wish to concentrate on the relevant features of fluid flow. One thus might be tempted to think that a continuum description is somehow "easier" than the actual discrete situation. This however depends very much on what the definition of "easier" is. Consider as an example a linear chain of point masses m that are coupled by elastic springs (with a stiffness k). The time evolution of

the displacements $x_i(t)$, $i \in \mathbb{Z}$ of all the masses are the "microscopic" description of the chain. By Newton's law one finds

$$m x_i''(t) = k\Big(x_{i+1}(t) - 2x_i(t) + x_{i-1}(t)\Big) \qquad \forall i \in \mathbb{Z} \tag{1.1}$$

Consider now a continuum model of this chain. For example, assume that there is some limiting procedure which leads from (1.1) to

$$\partial_t^2 s(t,x) = \frac{k}{m}\partial_x^2 s(t,x) \tag{1.2}$$

for some function $s : \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}$. This is the wave equation with wave speed $\sqrt{k/m}$. One has replaced a (large but finite) system of ordinary differential equations by a partial differential equation with an uncountably infinite number of degrees of freedom. They can, e.g. be parametrized by the Fourier modes that solve (1.2). The "approximation" of a macroscopic description suddenly turns out to be structurally tremendously more complex, particularly if you imagine a situation where the continuum description is a nonlinear PDE.

Note that it is easy to mix up two interpretations of what a continuum description is. Either you think of it as the equation that governs the behaviour of the discrete system as the microscopic length scales vanish in some carefully chosen limiting procedure, or you consider the continuum limit to be an approximation to the behaviour of the discrete system. The former is what mathematically makes more sense, but the latter is what it actually is used for. Indeed, when somebody designs an airplane and uses the Euler equations (say) to compute the flow of air, he does so because he expects the continuum description to give faithful predictions about the behaviour of air with microscopic length scales about $1\,\mu\mathrm{m} = 10^{-6}\,\mathrm{m}$, and not because the airplane will be flying through air that has undergone a limiting process during which all its microscopic length scales have vanished.

Consider oscillations of the linear chain. Obviously there is a microscopic length scale involved, which is given by the typical distance between two neighbouring point masses. This means that oscillations of the chain cannot have a shorter wave length than twice this distance. This corresponds to some maximum frequency. On the other hand, harmonic waves of *all* frequencies are solutions to the wave equation. Therefore the continuum description is not a faithful approximation of the discrete chain if one is interested in high frequency waves. The *realm of validity* of a continuous model should not be lost out of sight when dealing with it while having some practical applications in mind.

It is by no means clear that a continuum description would be free of pathologies. One might imagine, that for some discrete chain of masses coupled in a very complicated nonlinear manner a similar limiting procedure would lead to a PDE, which has some singularity. For example imagine the following: the PDE has well-behaved solutions only if the initial data are not too oscillatory, and otherwise they blow up very quickly. In particle physics jargon this would be referred to as an ultraviolet catastrophe. One might then argue that perhaps this time the PDE has not "forgotten" everything about the

underlying microscopic length scales – indeed in the discrete setting the highly oscillatory solutions just don't exist because there is some natural maximum frequency. One would need to regularize the continuum model somehow, which might become very complicated.

If all were well with the Euler equations, one might accept the above ideas as potential but irrelevant issues. However since the works of [DLS10, CDLK15, FKKM17] and others it is known that weak solutions to the Euler equations are not unique in multiple spatial dimensions for certain initial data. And although most probably the underlying reason will have an explanation unrelated to what has been said above, it shows once more that the transition from discrete models to their continuous descriptions is fascinatingly complex.

1.1.2 Continuum description of a fluid

Contrary to classical thermodynamics where the gas is uniform throughout the whole volume, in hydrodynamics the state of the gas can vary from location to location. Therefore for the averaging to be a reasonable description of the fluid, it is to happen at scales much larger than the microscopic scales (e.g. the mean free path), but much smaller than the total volume. The transition from a particle description to the Boltzmann equation and from there to the fluid equations is fairly complicated, and its understanding not entirely complete. Therefore it shall not be touched upon in this presentation. The reader interested in more details on this topic is referred e.g. to the review [Vil02] and the references therein.

The state of the fluid is described by specifying first of all the density ρ and velocity $\mathbf{v}$ of the fluid as functions of (d-dimensional) space and time:

$$\rho : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}^+$$
$$\mathbf{v} : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}^d$$

Physically, the density should always remain positive, as it measures mass per unit volume. *Momentum per unit volume* is the product $\rho\mathbf{v}$. Energy that is related to the motion of the gas is called *kinetic energy*, and its amount per unit volume is

$$e_{\text{kin}} = \frac{1}{2}\rho|\mathbf{v}|^2$$

The time evolution of ρ and $\rho\mathbf{v}$ is governed by

$$\partial_t \rho + \nabla \cdot (\rho\mathbf{v}) = 0 \tag{1.3}$$
$$\partial_t(\rho\mathbf{v}) + \nabla \cdot (\rho\mathbf{v} \otimes \mathbf{v} + p \cdot \mathbb{1}) = 0 \tag{1.4}$$

Here p is a new function that is called *pressure* and will be discussed below.

The two equations can be combined into an equation for the velocity $\mathbf{v}$, if all the derivatives exist in the classical sense:

$$\partial_t \mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{v} + \frac{\nabla p}{\rho} = 0$$

Note also the following important identity of vector calculus

$$(\mathbf{v}\cdot\nabla)\mathbf{v} = \frac{1}{2}\nabla(|\mathbf{v}|^2) - \mathbf{v}\times(\nabla\times\mathbf{v})$$

The evolution of the kinetic energy is obtained in a straightforward manner from the Equations (1.3)–(1.4):

$$\partial_t\frac{|\mathbf{v}|^2}{2} = \mathbf{v}\cdot\partial_t\mathbf{v} = -\mathbf{v}\cdot\Big[(\mathbf{v}\cdot\nabla)\mathbf{v}\Big] - \mathbf{v}\cdot\frac{\nabla p}{\rho} = -\frac{1}{2}\mathbf{v}\cdot\nabla(|\mathbf{v}|^2) - \mathbf{v}\cdot\frac{\nabla p}{\rho}$$
$$\partial_t e_{\text{kin}} = \frac{|\mathbf{v}|^2}{2}\partial_t\rho + \rho\partial_t\frac{|\mathbf{v}|^2}{2} = -\frac{|\mathbf{v}|^2}{2}\nabla\cdot(\mathbf{v}\rho) - \frac{1}{2}\rho\mathbf{v}\cdot\nabla(|\mathbf{v}|^2) - \mathbf{v}\cdot\nabla p$$

This gives the evolution equation for the kinetic energy:

$$\partial_t e_{\text{kin}} + \nabla\cdot(\mathbf{v}(e_{\text{kin}}+p)) = p\nabla\cdot\mathbf{v} \tag{1.5}$$

In hydrodynamics, pressure appears in two distinct ways. For the incompressible Euler equations, it is a free variable that, at every time t, has to be found such that the divergence constraint

$$\nabla\cdot\mathbf{v} = 0$$

is fulfilled. It is thus an evolving variable. The **incompressible Euler equations** read

$$\begin{aligned}\partial_t\rho + \nabla\cdot(\rho\mathbf{v}) &= 0\\ \partial_t(\rho\mathbf{v}) + \nabla\cdot(\rho\mathbf{v}\otimes\mathbf{v} + p\cdot\mathbb{1}) &= 0\\ \nabla\cdot\mathbf{v} &= 0\\ \text{unknowns: } \rho, \mathbf{v}, p\end{aligned}$$

For the compressible Euler equations, p is a given functional (a so called *equation of state*) of the state of the gas. For example it may be a function $p(\rho)$ of ρ only. The choice $p(\rho) = K\rho^\gamma$, $K\in\mathbb{R}^+$, $\gamma\in\mathbb{R}$, $\gamma\geq 1$ gives the **isentropic Euler equations**

$$\partial_t\rho + \nabla\cdot(\rho\mathbf{v}) = 0 \tag{1.6}$$
$$\partial_t(\rho\mathbf{v}) + \nabla\cdot(\rho\mathbf{v}\otimes\mathbf{v} + p(\rho)\cdot\mathbb{1}) = 0 \tag{1.7}$$
$$\text{unknowns: } \rho, \rho\mathbf{v}$$

On physical grounds (see e.g. [LL13], Vol. 6) one defines[1] the *temperature* as

$$T = \frac{p}{\rho}$$

[1]For an ideal gas consisting of particles of mass m and with the Boltzmann constant denoted by $k_{\text{B}} = 1.38\cdot 10^{-23}\text{J/K}$ the physically correct formula is $T = \frac{m}{k_{\text{B}}}\frac{p}{\rho}$. Here the temperature is obtained in Kelvin. The definition above amounts to a different choice of units.

This is why an equation of state with $\gamma = 1$ is called *isothermal*. Temperature is related to a form of energy, which is called *internal energy* and its amount per unit volume is denoted by e_{int}.

Finally, the Euler equations for an ideal gas shall be discussed. Ideal gases are also parametrized by a parameter γ, which should not be confused with the exponent γ of the isentropic Euler equations. There are reasons why these two traditionally are given the same symbol, but to begin with they are two different constants because they appear in two different contexts. The formulae for the different properties of an ideal gas can be derived as consequences of some more fundamental definition of what an ideal gas is (see e.g. [LL13], Vol. 6). Here to facilitate the presentation, the ideal gas is *defined* to be a fluid characterized by $\gamma \in \mathbb{R}$, $\gamma > 1$ that fulfills

$$e_{\text{int}} = \frac{p}{\gamma - 1}$$

The total energy is the sum of the internal and the kinetic energies. From First Law of Thermodynamics one obtains ([LL13])

$$\partial_t e_{\text{int}} + \nabla \cdot (\mathbf{v} e_{\text{int}}) + p \nabla \cdot \mathbf{v} = 0$$

and together with (1.5) an equation for the total energy $e = e_{\text{kin}} + e_{\text{int}}$

$$\partial_t e + \nabla \cdot (\mathbf{v}(e + p)) = 0$$

If everything is differentiable in the classical sense, this can be used to derive an evolution equation for p:

$$\partial_t p + \mathbf{v} \cdot \nabla p + \gamma p \nabla \cdot \mathbf{v} = 0$$

Thus the **Euler equations for an ideal gas** read

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{1.8}$$

$$\partial_t (\rho \mathbf{v}) + \nabla \cdot (\rho \mathbf{v} \otimes \mathbf{v} + p(\rho) \cdot \mathbb{1}) = 0 \tag{1.9}$$

$$\partial_t e + \nabla \cdot (\mathbf{v}(e + p)) = 0 \tag{1.10}$$

$$\text{unknowns: } \rho, \rho\mathbf{v}, e$$

Variables $\rho, \rho\mathbf{v}, e$ are called *conservative variables*, the set $\rho, \mathbf{v}, p$ is called *primitive variables*.

Equations (1.8), (1.9), (1.10) are hyperbolic (see e.g. [Tor09]). Small pressure perturbations to a constant background are found to be governed by a linear wave equation with speed

$$c = \sqrt{\frac{\gamma p}{\rho}}$$

which is called *sound speed*. Details of this linearization are given in Section 2.1.1.

In all the equations it is possible to single out a derivative operator

$$\frac{\mathrm{d}}{\mathrm{d}t} := \partial_t + \mathbf{v} \cdot \nabla$$

called the *advective, or comoving derivative*. It describes the time change of a quantity as it moves with the flow. The equations then become

$$\begin{aligned} \frac{\mathrm{d}\rho}{\mathrm{d}t} + \rho \nabla \cdot \mathbf{v} &= 0 \\ \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \frac{\nabla p}{\rho} &= 0 \\ \frac{\mathrm{d}p}{\mathrm{d}t} + \gamma p \nabla \cdot \mathbf{v} &= 0 \end{aligned}$$

The First Law of Thermodynamics expresses energy conservation. The Second Law of Thermodynamics states that a quantity s called *entropy* is a monotone function of time. Whereas in physics literature it is defined to grow, in mathematics it typically is defined to decay in time by inverting the sign. The origin of such a function can be found in statistical mechanics. It is fascinating that macroscopic irreversibility can be found in a system whose microscopic laws are fully time reversible. The same concept was discovered in other domains, for instance by Claude Shannon ([Sha48]) when dealing with information (or uncertainty). Having asked John von Neumann for a better name for his "uncertainty function" he got the following answer (as reported in [TM71]):

> *You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.*

Consider a gas filling a volume, and made of particles that collide randomly with each other. It is highly improbable that they would at some point in time all be in the left half of the volume, and the right half be empty. Seeing a film that shows a gas expand into vacuum you can tell that everything is fine, while seeing a film that shows the opposite you are able to tell that the film must be running backwards.

Making these ideas precise is not easy, and "present day mathematics is unable to prove [Boltzmann's H-theorem] rigorously and in satisfactory generality" [Vil08]. Here again, the presentation will content itself with stating that

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \partial_t s + \mathbf{v} \cdot \nabla s \leq 0$$

for $s = p\rho^{-\gamma}$. By combining with the continuity equation (1.3) one can rewrite this as

$$\frac{\mathrm{d}(\rho s)}{\mathrm{d}t} = \partial_t(\rho s) + \nabla \cdot (\rho \mathbf{v} s) \leq 0$$

As the Euler equations describe an ideal fluid, equality is true away from discontinuities.

The curl $\omega = \nabla \times \mathbf{v}$ of $\mathbf{v}$ is given a particular name: *vorticity*. For the Euler equations (1.8)–(1.10) vorticity fulfills

$$\partial_t \omega + \mathbf{v} \cdot \nabla \omega - (\omega \cdot \nabla)\mathbf{v} + \omega(\nabla \cdot \mathbf{v}) + \frac{\nabla p \times \nabla \rho}{\rho^2} = 0 \tag{1.11}$$

1.2 Low Mach number limit

The limit of low Mach numbers in the context of the Euler equations (1.8)–(1.10) is best explored by introducing a family of solutions, parametrized by a real dimensionless number $\epsilon > 0$, $\epsilon \to 0$.

Definition 1.1 (Asymptotic scaling)**.** *Assume that in the limit $\epsilon \to 0$ a function $f(t, \mathbf{x}; \epsilon)$ can be written as an expansion in ϵ as*

$$f = \epsilon^p (f^{(0)} + f^{(1)}\epsilon + \ldots)$$

with the functions $f^{(i)}$ not depending on ϵ. Then it is said to be asymptotically scaling *as ϵ^p, or $f \in \mathcal{O}(\epsilon^p)$.*

Using some intuition, it is possible to directly insert powers of a real number $\epsilon > 0$ into the equations, with the aim of highlighting a particular limiting regime as $\epsilon \to 0$. Such equations are called *rescaled*. This procedure changes the equations, because they are not invariant under insertion of arbitrary factors. So what do these equations describe? The following derivation aims at making this clear. It considers first just the original equation, and a family of solutions, parametrized by ϵ, that tend to the particular limiting regime of interest. This family of solutions gives rise to a family of equations, parametrized by ϵ, in the following way: To highest order in ϵ, the family of solutions can be rewritten as one solution to a family of different equations. More precisely, the solutions still depend on ϵ, but their leading order scalings have been transferred to the equation, such that it also depends on ϵ now. These equations are the rescaled equations. The way of derivation presented here (and published in [BEK+17]) allows to find the most general form of the rescaled equations without making assumptions that need physical guidance.

Consider thus a family of solutions to the Euler equations, parametrized by $\epsilon > 0$.

i) The local Mach number $M_{\text{loc}}(t, \mathbf{x}) := |\mathbf{v}(t, \mathbf{x})| / \sqrt{\frac{\gamma p(t, \mathbf{x})}{\rho(t, \mathbf{x})}}$ is assumed to be written as an expansion in ϵ and scaling asymptotically as ϵ: $M_{\text{loc}} \in \mathcal{O}(\epsilon)$. This condition makes clear that the limiting regime of interest is that of low Mach number.

ii) Every member of the family shall fulfill the same equation of state

$$e = \frac{p}{\gamma - 1} + \frac{1}{2}\rho |\mathbf{v}|^2$$

iii) Every member of the family shall fulfill the Euler equations.

These requirements replace some physically motivated requirements found in the literature (e.g. in [Kle95, MRE15]) which relate the scalings of the pressure to scalings of ρc^2 rather than $\rho|\mathbf{v}|^2$. It is considered important to make clear which parts of the reasoning actually rely on physical arguments, and which ones are consequences of the equations themselves. The latter evidently can then be reused in a variety of other circumstances where physical intuition might not be available.

Theorem 1.1 (Asymptotic scalings)**.** *The most general asymptotic scalings of the dependent and independent quantities that are consistent with the requirements i) and ii) are*

$$\begin{aligned} \mathbf{x} &= \epsilon^{\mathfrak{a}}\tilde{\mathbf{x}}, & t &= \epsilon^{\mathfrak{b}}\tilde{t} \\ \rho(t,\mathbf{x}) &= \epsilon^{\mathfrak{c}+2-2\mathfrak{d}}\tilde{\rho}(\tilde{t},\tilde{\mathbf{x}}), & \mathbf{v}(t,\mathbf{x}) &= \epsilon^{\mathfrak{d}}\tilde{\mathbf{v}}(\tilde{t},\tilde{\mathbf{x}}) \\ e(t,\mathbf{x}) &= \epsilon^{\mathfrak{c}}\tilde{e}(\tilde{t},\tilde{\mathbf{x}}), & p(t,\mathbf{x}) &= \epsilon^{\mathfrak{c}}\tilde{p}(\tilde{t},\tilde{\mathbf{x}}) \end{aligned}$$

with $\mathfrak{a},\mathfrak{b},\mathfrak{c},\mathfrak{d}$ arbitrary numbers. It is understood that quantities with a tilde are $\mathcal{O}(1)$ when expanded as power series in ϵ.

Proof. Assume general scalings

$$\begin{aligned} \mathbf{x} &= \epsilon^{\mathfrak{a}}\tilde{\mathbf{x}}, & t &= \epsilon^{\mathfrak{b}}\tilde{t} \\ \rho(t,\mathbf{x}) &= \epsilon^{\mathfrak{e}}\tilde{\rho}(\tilde{t},\tilde{\mathbf{x}}), & \mathbf{v}(t,\mathbf{x}) &= \epsilon^{\mathfrak{d}}\tilde{\mathbf{v}}(\tilde{t},\tilde{\mathbf{x}}) \\ e(t,\mathbf{x}) &= \epsilon^{\mathfrak{c}}\tilde{e}(\tilde{t},\tilde{\mathbf{x}}), & p(t,\mathbf{x}) &= \epsilon^{\mathfrak{f}}\tilde{p}(\tilde{t},\tilde{\mathbf{x}}) \end{aligned}$$

Then by computing M_{loc} one finds

$$M_{\text{loc}} = \epsilon^{\mathfrak{d}-\mathfrak{f}/2+\mathfrak{e}/2}|\tilde{\mathbf{v}}|/\sqrt{\frac{\gamma\tilde{p}}{\tilde{\rho}}} = \epsilon^{\mathfrak{d}-\mathfrak{f}/2+\mathfrak{e}/2}\tilde{M}_{\text{loc}}$$

i.e. by condition i) above

$$2\mathfrak{d} - \mathfrak{f} + \mathfrak{e} = 2 \tag{1.12}$$

Inserting the general scalings into the equation of state one is left with

$$\epsilon^{\mathfrak{c}}\tilde{e} = \frac{\epsilon^{\mathfrak{f}}\tilde{p}}{\gamma-1} + \frac{1}{2}\epsilon^{\mathfrak{e}+2\mathfrak{d}}\tilde{\rho}|\tilde{\mathbf{v}}|^2$$

and using (1.12)

$$\epsilon^{\mathfrak{c}}\tilde{e} = \frac{\epsilon^{\mathfrak{f}}\tilde{p}}{\gamma-1} + \frac{1}{2}\epsilon^{2+\mathfrak{f}}\tilde{\rho}|\tilde{\mathbf{v}}|^2$$

Therefore one concludes $\mathfrak{c} = \min(\mathfrak{f},\mathfrak{f}+2) = \mathfrak{f}$. Indeed, $\mathfrak{c}$ being an asymptotic scaling denotes the lowest order of ϵ that appears on the right hand side. Replacing

$$\begin{aligned} \mathfrak{e} &= 2 - 2\mathfrak{d} + \mathfrak{c} \\ \mathfrak{f} &= \mathfrak{c} \end{aligned}$$

proves the assertion. □

An example of such a family of solutions is given by the Gresho vortex setup in (5.15)–(5.16) (page 158).

Inserting the above scalings into the Euler equations yields a system of equations that is fulfilled by quantities with a tilde:

Corollary 1.1 (Rescaled Euler equations)**.** *Inserting the scalings obtained in Theorem 1.1 into the Euler equations yields*

$$\tilde{e} = \frac{\tilde{p}}{\gamma - 1} + \frac{1}{2}\epsilon^2 \tilde{\rho} |\tilde{\mathbf{v}}|^2 \tag{1.13}$$

and

$$\begin{aligned} \epsilon^{\mathfrak{a}-\mathfrak{d}-\mathfrak{b}} \partial_t \tilde{\rho} + \nabla \cdot (\tilde{\rho}\tilde{\mathbf{v}}) &= 0 \\ \epsilon^{\mathfrak{a}-\mathfrak{d}-\mathfrak{b}} \partial_t (\tilde{\rho}\tilde{\mathbf{v}}) + \nabla \cdot \left(\tilde{\rho}\tilde{\mathbf{v}} \otimes \tilde{\mathbf{v}} + \frac{\tilde{p}}{\epsilon^2} \cdot \mathbb{1} \right) &= 0 \\ \epsilon^{\mathfrak{a}-\mathfrak{d}-\mathfrak{b}} \partial_t \tilde{e} + \nabla \cdot (\tilde{\mathbf{v}}(\tilde{e} + \tilde{p})) &= 0 \end{aligned}$$

Observe the fact that the kinetic energy obtains an additional factor of ϵ^2 in the equation of state. This is not in contradiction to condition ii because (1.13) describes the equation of state fulfilled by quantities with a tilde.

The factor in front of the time derivatives is related to the dimensionless Strouhal number

$$Str_{\text{loc}} = \frac{x}{|\mathbf{v}|t} = \frac{\epsilon^{\mathfrak{a}} \tilde{x}}{\epsilon^{\mathfrak{d}} |\tilde{\mathbf{v}}| \cdot \epsilon^{\mathfrak{b}} \tilde{t}}$$

This factor is *not* identical to the Strouhal number, but is just its asymptotic ϵ-scaling. As an additional condition on the family of solutions one is tempted to insist on $Str_{\text{loc}} \in \mathcal{O}(1)$, i.e. $\mathfrak{a} - \mathfrak{d} - \mathfrak{b} = 0$. This corresponds to adapting the time scales to the speed of the fluid (and not to sound wave crossing times). This yields

$$\partial_t \tilde{\rho} + \nabla \cdot (\tilde{\rho}\tilde{\mathbf{v}}) = 0 \tag{1.14}$$

$$\partial_t (\tilde{\rho}\tilde{\mathbf{v}}) + \nabla \cdot \left(\tilde{\rho}\tilde{\mathbf{v}} \otimes \tilde{\mathbf{v}} + \frac{\tilde{p}}{\epsilon^2} \cdot \mathbb{1} \right) = 0 \tag{1.15}$$

$$\partial_t \tilde{e} + \nabla \cdot (\tilde{\mathbf{v}}(\tilde{e} + \tilde{p})) = 0 \tag{1.16}$$

Different ways of decreasing the Mach number (e.g. by decreasing the value of the velocity, or increasing the sound speed instead, or a combination of both) are equivalent and result in the same rescaled equations. This explains why the precise value of $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}$ does not matter for the form of the rescaled equations. For notational convenience the tilde is dropped.

In the limit $\epsilon \to 0$ the solutions to (1.14)–(1.16) tend to solutions of the incompressible Euler equations, see e.g. [Ebi77, KM81, U$^+$86, Asa87, Iso87, KLN91, Sch94, MS01]. The limit can formally be found by expanding all quantities as series in ϵ, e.g. for the pressure this would give

$$p(t, \mathbf{x}) = p^{(0)}(t, \mathbf{x}) + \epsilon p^{(1)}(t, \mathbf{x}) + \epsilon^2 p^{(2)}(t, \mathbf{x}) + \mathcal{O}(\epsilon^3).$$

Inserting these into the above equations, collecting order by order and assuming impermeable boundaries gives (compare e.g. [Kle95, HJL12])

$$p^{(0)} = \text{const}, \tag{1.17}$$
$$p^{(1)} = \text{const}, \tag{1.18}$$
$$(\nabla \cdot \mathbf{v})^{(0)} = 0, \tag{1.19}$$

and

$$\begin{aligned} \partial_t \rho^{(0)} + \mathbf{v}^{(0)} \cdot \nabla \rho^{(0)} &= 0, \\ \partial_t \mathbf{v}^{(0)} + (\mathbf{v}^{(0)} \cdot \nabla)\mathbf{v}^{(0)} + \nabla p^{(2)}/\rho^{(0)} &= 0. \end{aligned}$$

These equations describe incompressible flows. Conditions (1.17), (1.18) and (1.19) are understood to be true at any time. Initial data that fulfill them are called *well-prepared*. Not well-prepared initial data may lead to an incompressible flow as well, but then an initial disturbance is produced.

The equation for the kinetic energy $e_{\text{kin}} = \frac{1}{2}\rho|\mathbf{v}|^2$ can be rewritten as

$$\partial_t e_{\text{kin}} + \nabla \cdot \left[\mathbf{v}\left(e_{\text{kin}} + \frac{p}{\epsilon^2}\right)\right] = \frac{p}{\epsilon^2}\nabla \cdot \mathbf{v}.$$

The source term vanishes for incompressible flows and in this case the kinetic energy becomes a conserved quantity. For compressible flows, this is true in the limit $\epsilon \to 0$ as well, despite of $\frac{\nabla\cdot\mathbf{v}}{\epsilon^2} \notin \mathcal{O}(\epsilon)$. Expanding the quantities and using (1.17) and (1.18) makes the terms proportional to $\frac{1}{\epsilon}$ or $\frac{1}{\epsilon^2}$ cancel and gives

$$\partial_t e_{\text{kin}} + \operatorname{div}\left[\mathbf{v}\left(e_{\text{kin}} + p^{(2)}\right)\right] = p^{(2)}\nabla \cdot \mathbf{v} + \mathcal{O}(\epsilon).$$

Now the source term indeed is $\mathcal{O}(\epsilon)$ and the kinetic energy can be seen to become a conserved quantity in the limit $\epsilon \to 0$ by Equation (1.19).

1.3 Gravity source terms

When gravity is present, the Euler equations (1.8)–(1.10) have to be augmented by source terms involving a vector field $\mathbf{g}$:

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{1.22}$$
$$\partial_t(\rho \mathbf{v}) + \nabla \cdot (\rho \mathbf{v} \otimes \mathbf{v} + p \cdot \mathbb{1}) = \rho \mathbf{g} \tag{1.23}$$
$$\partial_t e + \nabla \cdot [\mathbf{v}(e + p)] = \rho \mathbf{v} \cdot \mathbf{g} \tag{1.24}$$

In presence of an exterior gravitational field $\mathbf{g}$, Equations (1.22), (1.23), (1.24) describe the motion and the steady states of the fluid. In general, $\mathbf{g}$ is a function of space and time.
In certain applications (e.g. when the atmosphere of the Earth is considered), $\mathbf{g}$ is a given function of space only. It may even be chosen constant in certain cases.

In other applications, $\mathbf{g}$ is evolving in time. For example, when the fluid is evolving under the influence of its own gravity (*self-gravity*), system (1.22)–(1.24) has to be augmented by an equation which tells how mass creates its gravity field. From classical Newtonian gravity follows that $\mathbf{g}$ is a gradient field

$$\mathbf{g} = -\nabla\Phi$$

and the *gravitational potential* Φ is a new dependent variable which has to fulfill

$$\Delta\Phi = 4\pi\rho G \tag{1.25}$$

The gravitational constant $G \simeq 6.67 \cdot 10^{-11}$ (in SI units) is a fundamental constant of nature. For notational convenience, it can also be absorbed into a choice of different units (together with 4π).

The system (1.22)–(1.24) together with (1.25) now is closed. It is of mixed type because (1.25) is elliptic: changing the mass distribution, i.e. ρ, in some region immediately implies a change of Φ everywhere and thus immediately changes the motion of the fluid arbitrarily far away. In General Relativity, no information can travel faster than light, and one replaces equation (1.25) by some hyperbolic equation again[2]. It limits the domain of influence of a change of mass distribution to what is reachable by traveling with the speed of light.

In the following $\mathbf{g}$ is assumed to be a given function of space only.

Definition 1.2 (Stationary and static)**.** *A* stationary *solution of any set of evolution equations is characterized by a vanishing time derivative.*
A static *solution of any set of evolution equations that contain a variable "velocity"* $\mathbf{v}$ *is a stationary solution that additionally fulfills* $\mathbf{v} = 0$*.*

The static states (called *hydrostatic equilibria*) of (1.22)–(1.24) are given by

$$\nabla p = \rho\mathbf{g} \tag{1.26}$$

This equation leads to a solution once a relationship between p and ρ is known. In view of $p = \rho T$ this hydrostatic equilibrium can be determined, once the temperature of the gas is given everywhere. One can show that this equilibrium can be unstable. Having buoyantly lighter fluid placed under heavier one, small perturbations will increase and ultimately lead to convective motion and mixing of the two fluids. The reversed situation is a stable one, and therefore small perturbations will not grow in time.

Consider again a family of solutions to (1.22)–(1.24).

[2]Indeed, the Einstein equations are in a sense similar to (1.25). However in General Relativity gravity is not described by just one variable Φ, but by a symmetric 4×4 tensor field. Thus in the end, the equations look very differently and add tremendous complexity. For an introduction see e.g. [MTW17, Str12]

Theorem 1.2 (Rescaled Euler equations with gravity)**.** *With conditions i)–ii) as well as* $Str_{loc} \in \mathcal{O}(1)$ *the rescaled Euler equations are*

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{1.27}$$

$$\partial_t (\rho \mathbf{v}) + \nabla \cdot \left(\rho \mathbf{v} \otimes \mathbf{v} + \frac{p}{\epsilon^2} \cdot \mathbb{1} \right) = \frac{1}{Fr^2} \rho \mathbf{g} \tag{1.28}$$

$$\partial_t e + \nabla \cdot [\mathbf{v}(e+p)] = \frac{\epsilon^2}{Fr^2} \rho \mathbf{v} \cdot \mathbf{g} \tag{1.29}$$

Here $Fr = \epsilon^{\mathfrak{d} - \frac{\mathfrak{a}+\mathfrak{g}}{2}}$.

Proof. By Theorem 1.1 conditions i)–ii) imply

$$\begin{aligned} \mathbf{x} &= \epsilon^{\mathfrak{a}} \tilde{\mathbf{x}}, & t &= \epsilon^{\mathfrak{b}} \tilde{t} \\ \rho(t, \mathbf{x}) &= \epsilon^{\mathfrak{c}+2-2\mathfrak{d}} \tilde{\rho}(\tilde{t}, \tilde{\mathbf{x}}), & \mathbf{v}(t, \mathbf{x}) &= \epsilon^{\mathfrak{d}} \tilde{\mathbf{v}}(\tilde{t}, \tilde{\mathbf{x}}) \\ e(t, \mathbf{x}) &= \epsilon^{\mathfrak{c}} \tilde{e}(\tilde{t}, \tilde{\mathbf{x}}), & p(t, \mathbf{x}) &= \epsilon^{\mathfrak{c}} \tilde{p}(\tilde{t}, \tilde{\mathbf{x}}) \\ \mathbf{g}(\mathbf{x}) &= \epsilon^{\mathfrak{g}} \tilde{\mathbf{g}}(\tilde{\mathbf{x}}) \end{aligned}$$

Inserting this into the Euler equations leads to Equations (1.27), (1.28), (1.29). This proves the assertion. □

Note that Fr only denotes the power of ϵ which appears in the definition of the local Froude number

$$Fr_{\text{loc}} = \frac{|\mathbf{v}|}{\sqrt{|\mathbf{x} \cdot \mathbf{g}|}} = \epsilon^{\mathfrak{d} - \frac{\mathfrak{a}+\mathfrak{g}}{2}} \frac{|\tilde{\mathbf{v}}|}{\sqrt{|\tilde{\mathbf{x}} \cdot \tilde{\mathbf{g}}|}}$$

Colloquially Fr is often referred to as the Froude number as well. Whereas the Mach number squared is the ratio between kinetic and internal energies, the local Froude number squared quantifies the relative magnitudes of kinetic and gravitational potential energies. In the example of a constant gravity field, this gives the appearance of the independent spatial variable in the definition of Fr_{loc} the interpretation of the height measured in the direction opposite to $\mathbf{g}$.

Consider now a setup where the fluid performs motions with a small velocity superposing an ϵ-independent hydrostatic equilibrium. This suggests to actually choose the velocities to decrease as ϵ (i.e. $\mathfrak{d} = 1$) and to let gravity and length scales (as quantities involved in the background equilibrium) become asymptotically constant: $\mathfrak{a} = \mathfrak{g} = 0$:

$$\begin{aligned} \mathbf{x} &= \tilde{\mathbf{x}}, & t &= \epsilon^{\mathfrak{b}} \tilde{t}, \\ \rho(t, \mathbf{x}) &= \epsilon^{\mathfrak{c}} \tilde{\rho}(\tilde{t}, \tilde{\mathbf{x}}), & \mathbf{v}(t, \mathbf{x}) &= \epsilon \tilde{\mathbf{v}}(\tilde{t}, \tilde{\mathbf{x}}), \\ e(t, \mathbf{x}) &= \epsilon^{\mathfrak{c}} \tilde{e}(\tilde{t}, \tilde{\mathbf{x}}), & p(t, \mathbf{x}) &= \epsilon^{\mathfrak{c}} \tilde{p}(\tilde{t}, \tilde{\mathbf{x}}), \\ \mathbf{g}(\mathbf{x}) &= \tilde{\mathbf{g}}(\tilde{\mathbf{x}}), \end{aligned}$$

This still leaves $\mathfrak{c}$ and $\mathfrak{b}$ arbitrary, but already implies the relation

$$Fr = \epsilon$$

Note that this choice did not follow from equations, but has been taken having a particular application in mind. It seems, however, to be the only interesting borderline case. If $Fr = \epsilon^\alpha$, $\alpha < 1$, then gravity becomes irrelevant in the limit. If $Fr = \epsilon^\alpha$ with $\alpha > 1$, then gravity dominates everything else, and it does not seem to be possible to derive any nontrivial limit. The borderline case $Fr = \epsilon$ leads, by formally performing asymptotic expansions on all the quantities, to

$$\nabla p^{(0)} = \rho^{(0)} \mathbf{g}^{(0)} \tag{1.30}$$

This is the equation of hydrostatic equilibrium encountered in (1.26) already, and which is to be seen contrasting (1.17). Observe that the velocity field is not divergenceless any more because this was a direct consequence of (1.17).

Chapter 2

Equations of linear acoustics

Contents

The Euler equations contain the kinematic essence of fluid motion, as they express how the fundamental conservation laws of mechanics apply to continua. Why are they nonlinear? Indeed, linear elasticity, for example, ends up as a linear theory, although it also applies fundamental laws of mechanics to continua. The simplest reason is that in elasticity, the elements are tied to each other, whereas fluids can mix. This means that even the simplest model of fluid motion, that just makes use of the most fundamental conservation laws, is already a system of vast complexity.

This has two consequences. The first consequence is that, at least so far, numerical simulations are the only way to gain insight in realistic flows. Indeed, exact solutions of Euler equations typically have a high degree of symmetry and are mostly of academic interest. The second consequence is that, when it comes to designing numerical methods, the complexity of the Euler equations is again in the way.

In the context of one-dimensional problems, it has been found that, for example, upwinding is the key to stable schemes (see [VL06] for a review). This means that finite difference approximations to spatial derivatives should respect the direction of information propagation that is dictated by the equations. However, upwinding is still not easy to apply to the Euler equations, as, depending on the flow, certain information propagates in one direction, and certain information – in another. This is why, when introducing the concept of upwinding, textbooks (e.g. [LeV02, Tor09]) prefer to focus first on a situation with a unique direction of information propagation: the linear advection

equation

$$\partial_t q(t,x) + c\,\partial_x q(t,x) = 0 \qquad q : \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}, \quad c \in \mathbb{R}$$
$$q(0,x) := q_0(x)$$

This is, so to speak, the basic building block. Although, obviously, the Euler equations are not just advection, and are nonlinear, understanding upwinding for the advection equation is of immense help.

Multi-dimensional problems pose additional challenges. The problem of low Mach number flows is a such – in [Tur87, KLN91], for example, it has been found that numerical schemes seem to split into those whose error increases as the Mach number of the flow is reduced, and those which do not have this problem. Therefore it seems necessary to introduce a way to discriminate between the two, and to understand the origin of the error in this limit – in order to avoid it, of course. There exists a variety of approaches in the literature. They are described in more detail later. The only relevant aspect for the moment is that most of them ([DOR10] is an exception) discuss the Euler equations directly. However, just as the clue to the concept of upwinding lies in the linear advection equation, this thesis wants to argue that the clue to the low Mach number limit is in the *acoustic equations*. In both cases, understanding the linear problem gives such a deep understanding, that the nonlinear case can be fruitfully dealt with.

In a certain way, the acoustic equations play the general role of the building block for the multi-dimensional situation, just as the advection equation plays this role in the one-dimensional case (see also the review [Roe17]). Indeed, it is not just the low Mach number limit that can be studied, but also such concepts as structure-preservation, e.g. with respect to vorticity. The idea of the acoustic equations as a stepping stone to an understanding of numerical methods is picked up in Section 4. This Section deals with the equations themselves: their derivation and origin and their exact solution.

2.1 Properties of the acoustic equations

2.1.1 Linearization

The acoustic equations describe the time evolution of small perturbations to a constant flow. There are several different linearization procedures that lead to essentially the same system of equations.

Linearizing the isentropic Euler equations in d spatial dimensions

$$\partial_t \rho + \operatorname{div}(\rho \mathbf{v}) = 0$$
$$\partial_t(\rho \mathbf{v}) + \operatorname{div}(\rho \mathbf{v} \otimes \mathbf{v} + p\mathbb{1}) = 0$$

with $p(\rho) = K\rho^\gamma$ around the state $(\rho, \mathbf{v}) = (\bar\rho, 0)$ yields

$$\partial_t \rho + \bar\rho \operatorname{div} \mathbf{v} = 0 \tag{2.1}$$
$$\partial_t \mathbf{v} + c^2 \frac{\operatorname{grad} \rho}{\bar\rho} = 0 \tag{2.2}$$

where one defines $c = \sqrt{p'(\bar{\rho})}$. Linearization with respect to a fluid state moving at some constant speed $\mathbf{U}$ can be easily removed or added via a Galilei transform.

The same system can be obtained from the Euler equations endowed with an energy equation

$$\begin{aligned} \partial_t \rho + \operatorname{div}(\rho \mathbf{v}) &= 0 \\ \partial_t(\rho \mathbf{v}) + \operatorname{div}(\rho \mathbf{v} \otimes \mathbf{v} + p\mathbb{1}) &= 0 \\ \partial_t e + \operatorname{div}(\mathbf{v}(e+p)) &= 0 \end{aligned}$$

with $e = \frac{p}{\gamma - 1} + \frac{1}{2}\rho|\mathbf{v}|^2$. Linearization around $(\rho, \mathbf{v}, p) = (\bar{\rho}, 0, \bar{p})$ yields

$$\partial_t \rho + \bar{\rho} \operatorname{div} \mathbf{v} = 0$$

$$\partial_t \mathbf{v} + \frac{\operatorname{grad} p}{\bar{\rho}} = 0 \tag{2.3}$$

$$\partial_t p + \bar{\rho} c^2 \operatorname{div} \mathbf{v} = 0 \tag{2.4}$$

Equations (2.3)–(2.4) are (up to rescaling and renaming) the same as (2.1)–(2.2). Both can be linearly transformed to either

$$\partial_t \mathbf{v} + \operatorname{grad} p = 0 \qquad \mathbf{v} : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}^d \tag{2.5}$$

$$\partial_t p + c^2 \operatorname{div} \mathbf{v} = 0 \qquad c = \text{const} \qquad p : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R} \tag{2.6}$$

or the symmetric version

$$\partial_t \mathbf{v} + c \operatorname{grad} p = 0 \tag{2.7}$$

$$\partial_t p + c \operatorname{div} \mathbf{v} = 0 \qquad c = \text{const} \tag{2.8}$$

The transformation which symmetrizes the Jacobian $J = \begin{pmatrix} 0 & 1 \\ c^2 & 0 \end{pmatrix}$ is

$$S = \begin{pmatrix} 1 & 0 \\ 0 & c \end{pmatrix} \tag{2.9}$$

such that $J = S \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix} S^{-1}$. In multiple spatial dimensions the upper left entry in S has to be replaced by an appropriate block-identity-matrix.

This system is the one to be studied in what follows, sometimes the symmetric version being preferred over (2.5)–(2.6). p will be called pressure and $\mathbf{v}$ the velocity – just to have names. Due to the different linearizations and the symmetrization they are not exactly the physical pressure or velocity any more, but still closely related.

These equations describe the time evolution of small perturbations to a constant state of the fluid, usually referred to as *sound waves*. Therefore they are called the equations of *linear acoustics*. Equations (2.7)–(2.8) have been studied among others in [LS02, MR01, AG15, DOR10, FG17, BK17]. Unless stated differently, the equations are always considered on all $\mathbb{R}^d$.

2.1.2 Vorticity

It should be noted that only the equation for the scalar p is the usual scalar wave equation

$$\partial_t^2 p - c^2 \Delta p = 0 \tag{2.10}$$

whereas $\mathbf{v}$ fulfills

$$\partial_t^2 \mathbf{v} - c^2 \mathrm{grad}\,\mathrm{div}\,\mathbf{v} = 0 \tag{2.11}$$

The identity $\nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot v) - \Delta \mathbf{v}$ links this operator to the vector Laplacian in 3-d. By (2.7)

$$\partial_t (\nabla \times \mathbf{v}) = 0 \tag{2.12}$$

but $\nabla \times \mathbf{v}$ needs not be zero initially. The quantity $\omega := \nabla \times \mathbf{v}$ is called *vorticity* by analogy with the Euler equations.

2.2 Exact solution

Consider the Cauchy problem for the multi-dimensional hyperbolic system

$$\begin{aligned} \partial_t q + (\mathbf{J} \cdot \nabla) q &= 0 \qquad & q : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}^n \\ q(0, \mathbf{x}) &= q_0(\mathbf{x}) \end{aligned} \tag{2.13}$$

where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{J}$ is the vector of the Jacobians into the different directions.

For the symmetrized system (2.7)–(2.8) in 3-d one has $q := (\mathbf{v}, p)$ and

$$\mathbf{J} = \left(\begin{pmatrix} 0 & & & c \\ & 0 & & \\ & & 0 & \\ c & & & 0 \end{pmatrix}, \begin{pmatrix} 0 & & & \\ & 0 & & c \\ & & 0 & \\ & c & & 0 \end{pmatrix}, \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 0 & c \\ & & c & 0 \end{pmatrix} \right) \tag{2.14}$$

The one-dimensional system

$$\partial_t p(t,x) + c \partial_x v(t,x) = 0 \tag{2.15}$$

$$\partial_t v(t,x) + c \partial_x p(t,x) = 0 \tag{2.16}$$

with the initial data ($x \in \mathbb{R}$)

$$p(0,x) = p_0(x) \qquad v(0,x) = v_0(x)$$

can be solved via characteristics, observing that

$$\partial_t (p \pm v) \pm c \partial_x (p \pm v) = 0$$

i.e.

$$p(t,x) = \frac{1}{2}\Big[p_0(x-ct) + p_0(x+ct)\Big] + \frac{1}{2}\Big[v_0(x-ct) - v_0(x+ct)\Big] \tag{2.17}$$

$$v(t,x) = \frac{1}{2}\Big[v_0(x-ct) + v_0(x+ct)\Big] + \frac{1}{2}\Big[p_0(x-ct) - p_0(x+ct)\Big] \tag{2.18}$$

With respect to the numerics of the 1-d system (2.15)–(2.16), one can write down the exact Godunov scheme by solving the 1-d Riemann problem for this system. This scheme is referred to as the upwind/Roe scheme, as it is the linear version of the Roe scheme [Roe81]. The dimensionally split case applied to multiple dimensions is discussed in [GZI+76].

In the one-dimensional case one observes a discrepancy in the required regularity of q_0: the solution in (2.15)–(2.16) has to be differentiable, whereas the solution formula (2.17)–(2.18) does not even require continuity of the initial data. The discrepancy is removed by generalizing the notion of a solution to all objects for which the solution formula makes sense. To say it in the words of Schwartz ([Sch78]),

> *On peut écrire l'expression générale d'une solution de l'équation aux dérivées partielles* $\frac{\partial^2 U}{\partial x^2} - \frac{\partial^2 U}{\partial y^2} = 0$ *sous la forme* $U = f(x+y) + g(x-y)$*; mais une telle fonction* U *ne peut vérifier l'équation aux dérivées partielles que si* f *et* g *sont deux fois dérivables. Dans le cas contraire, on peut convenir de dire que* U *est "solution généralisée" de l'équation.*

An essential ingredient of later discussion is the Riemann Problem, i.e. the Cauchy Problem for discontinuous initial data. Therefore a generalization of the notion of a solution is necessary in order to include discontinuous data. How this generalization is to be chosen depends on the precise shape of the solution formula. Contrary to formulae (2.17)–(2.18), which only contain the *values* of the initial data, the solution formula for (2.7)–(2.8) turns out to contain derivatives of the initial data (Section 2.2.2). This makes it necessary to consider distributional solutions. They are given a brief review in Section 2.2.1 first, which gives the opportunity to fix the notation. The solution is derived in Section 2.2.2, with Theorems 2.6 and 2.8 being the main results. These results have been published in [BK17].

2.2.1 Distributions

In this Section a brief review of definitions and results from the theory of distributions is given. This is done in first place to fix notation that will be used throughout the rest of the thesis and thus many results are stated without proofs. The reader interested in a thorough introduction is, for example, referred to [Sch78], [GS64], [Hör13], [Rud91].

Definition 2.1 (Distribution)**.** *A* distribution *is a continuous linear functional on the set* $D(\mathbb{R}^d)$ *of compactly supported smooth test functions* ψ*. The evaluation of the distribution* f *on a test function* ψ *is denoted by* $\langle f|\psi\rangle \in \mathbb{R}$ *(or* $\mathbb{C}$*). The set of all distributions is denoted by* $D'(\mathbb{R}^d)$*.*

It is possible to show (see e.g. [Rud91]) that a function $h \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$ gives rise to a distribution $\underline{h} \in D'(\mathbb{R}^d)$ in the following way: the action $\langle \underline{h} | \psi \rangle$ of $\underline{h}$ onto any test function $\psi \in D(\mathbb{R}^d)$ is defined as:

$$\langle \underline{h} | \psi \rangle := \int_{\mathbb{R}^d} \mathrm{d}\mathbf{x}\, h(\mathbf{x}) \psi(\mathbf{x}) \tag{2.19}$$

Definition 2.2 (Regular distribution)**.** *Given $h \in L^1_{loc}(\mathbb{R}^d)$, the distribution $\underline{h}$, defined by its action onto a test function $\psi \in D(\mathbb{R}^d)$ as in* (2.19)*, is called* regular distribution*.*

In order to make explicit the independent variable, the notation $\langle \underline{h} | \psi(\mathbf{x}) \rangle$ will be used. Two regular distributions $\underline{h_1}$ and $\underline{h_2}$ are equal, if $h_1 = h_2$ almost everywhere.

Definition 2.3 (Tempered distribution)**.** *The* Schwartz space $S(\mathbb{R}^d)$ *of rapidly decreasing functions f on $\mathbb{R}^d$ is defined as*

$$S(\mathbb{R}^d) := \Big\{ f \in C^\infty(\mathbb{R}^d) : \sup_{\mathbf{x} \in \mathbb{R}^d} |x_1^{a_1} \dots x_d^{a_2} \partial_{x_1}^{b_1} \cdots \partial_{x_d}^{b_d} f| < \infty \\ \forall (a_1, \dots, a_d, b_1, \dots, b_d) \in (\mathbb{N}_0)^{2d} \Big\}$$

The set $S'(\mathbb{R}^d)$ of tempered distributions *is the continuous dual of $S(\mathbb{R}^d)$.*

It is possible to show that the derivative $\nabla_{\mathbf{x}} T$ of a distribution $T \in D'(\mathbb{R}^d)$, defined in the following, is again a distribution (see e.g. [Rud91]).

Definition 2.4. *i) The derivative of a distribution $T \in D'(\mathbb{R}^d)$ is defined as*

$$\langle \nabla_{\mathbf{x}} T | \psi(\mathbf{x}) \rangle := -\langle T | \nabla_{\mathbf{x}} \psi(\mathbf{x}) \rangle \qquad \forall \psi \in D(\mathbb{R}^d)$$

ii) The Fourier transform $\mathbb{F}_{t,\mathbf{x}}$ applied to an integrable function $f : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\hat{f}(\omega, \mathbf{k}) := \mathbb{F}_{t,\mathbf{x}}[f](\omega, \mathbf{k}) := \frac{1}{\sqrt{2\pi}} \int \mathrm{d}t \frac{1}{(2\pi)^{d/2}} \int \mathrm{d}\mathbf{x} \exp(-\mathbb{i}\omega t + \mathbb{i}\mathbf{k} \cdot \mathbf{x}) f(t, \mathbf{x})$$

Generically, $\mathbf{k}$ denotes the dual variable to $\mathbf{x}$ and ω the dual to t. Note the symmetric prefactor convention chosen here, and that ω is used with the reverse sign. Also, generically, the tilde denotes a Fourier transform in the following.

iii) The Fourier transform $\mathbb{F}_{t,\mathbf{x}}[T](\omega, \mathbf{k})$ of a distribution T is defined by

$$\langle \mathbb{F}[T] | \psi \rangle := \langle T | \mathbb{F}[\psi] \rangle \qquad \forall \psi \in D(\mathbb{R}^{d+1})$$

or, making explicit the independent variables,

$$\langle \mathbb{F}_{t,\mathbf{x}}[T](\omega, \mathbf{k}) | \psi(\omega, \mathbf{k}) \rangle := \langle T(t, \mathbf{x}) | \mathbb{F}_{\omega,\mathbf{k}}[\psi](t, \mathbf{x}) \rangle \qquad \forall \psi \in D(\mathbb{R}^{d+1})$$

The class $S(\mathbb{R}^d)$ allows to put the Fourier transform to maximal use:

Theorem 2.1. *The Fourier transform is an automorphism on $S'(\mathbb{R}^d)$.*

For a proof see e.g. [Rud91].
The usual rules of differentiation apply:

Theorem 2.2. *Consider a distribution q and its Fourier transform $\hat{q}$. Then*

i)

$$\nabla_{\mathbf{x}} \mathbb{F}_{\mathbf{k}}^{-1}[\hat{q}(\mathbf{k})] = F_{\mathbf{k}}^{-1}[\mathbb{i}\mathbf{k}\hat{q}(\mathbf{k})] \tag{2.20}$$

ii)

$$\begin{aligned} \mathbb{F}_{\mathbf{k}}^{-1}[\nabla_{\mathbf{k}}\hat{q}] &= -\mathbb{i}\mathbf{x}\mathbb{F}_{\mathbf{k}}^{-1}[\hat{q}] \\ \mathbb{F}_{\mathbf{x}}[\nabla_{\mathbf{x}} q] &= \mathbb{i}\mathbf{k}\mathbb{F}_{\mathbf{x}}[q] \end{aligned} \tag{2.21}$$

Proof. i) For every test function $\psi \in S(\mathbb{R}^d)$

$$\langle \nabla_{\mathbf{x}} \mathbb{F}_{\mathbf{k}}^{-1}[\hat{q}(\mathbf{k})] | \psi(\mathbf{x}) \rangle = -\langle \hat{q}(\mathbf{k}) | \mathbb{F}_{\mathbf{x}}^{-1}[\nabla_{\mathbf{x}}\psi(\mathbf{x})] \rangle = \langle \mathbb{F}_{\mathbf{k}}^{-1}[\mathbb{i}\mathbf{k}\hat{q}(\mathbf{k})] | \psi \rangle$$

ii) Analogously,

$$\begin{aligned} \langle \mathbb{F}_{\mathbf{k}}^{-1}[\nabla_{\mathbf{k}}\hat{q}(\mathbf{k})] | \psi(\mathbf{x}) \rangle &= \langle \nabla_{\mathbf{k}}\hat{q}(\mathbf{k}) | \mathbb{F}_{\mathbf{x}}^{-1}[\psi] \rangle = -\langle \hat{q}(\mathbf{k}) | \nabla_{\mathbf{k}} \mathbb{F}_{\mathbf{x}}^{-1}[\psi] \rangle \\ &= -\langle \hat{q}(\mathbf{k}) | \mathbb{F}_{\mathbf{x}}^{-1}[\mathbb{i}\mathbf{x}\psi] \rangle = \langle -\mathbb{i}\mathbf{x}\mathbb{F}_{\mathbf{k}}^{-1}[\hat{q}(\mathbf{k})] | \psi(\mathbf{x}) \rangle \end{aligned}$$

The other equation is shown by repeating the argumentation for $\mathbb{F}_{\mathbf{x}}[q]$.

□

The Fourier transform of 1 is (up to factors) the Dirac distribution δ_0:

Definition 2.5 (Dirac distribution)**.** *The Dirac distribution $\delta_{\mathbf{x}'}$, or $\delta_{\mathbf{x}=\mathbf{x}'}$, is defined as $\langle \delta_{\mathbf{x}'} | \psi \rangle := \psi(\mathbf{x}')\ \forall \psi \in D(\mathbb{R}^d)$.*

Distributional solutions to partial differential equations are discussed in e.g. [Joh78].

Definition 2.6 (Distributional solution)**.** *Given a first order linear differential operator $\mathcal{L}$ containing derivatives with respect to $t \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$ and given initial data $q_0 \in D'(\mathbb{R}^d)$, q is called a* distributional solution *of $\mathcal{L}q = 0$ if it holds as an identity in $D'(\mathbb{R}^d)$; i.e. if $\langle \mathcal{L}q | \psi \rangle = 0\ \forall \psi \in D(\mathbb{R}^{d+1})$ and $q|_{t=0} = q_0$.*

Whenever a solution $f \in L^1_{\mathrm{loc}}$ in the sense of functions exists, then $\underbrace{f}$ is a distributional solution. In general in the following the solution will not be a function, but the initial data q_0 will. If the initial data are locally integrable functions, then in the context of a *distributional* initial value problem they are to be interpreted as regular distributions $\underbrace{q_0}$.

The convolution $F * G$ of two distributions F and G can be defined in certain cases. Here only the following definition is needed, and the reader is referred to e.g. [Rud91] for further details.

Definition 2.7 (Convolution)**.** *The convolution $F * G$ of $F, G \in D'(\mathbb{R}^d)$, with at least one of them having compact support, is defined $\forall \psi \in D(\mathbb{R}^d)$ as*

$$\langle (F * G)(\mathbf{x}) | \psi(\mathbf{x}) \rangle = \Big\langle F(\mathbf{x}) | \langle G(\mathbf{y}) | \psi(\mathbf{x}+\mathbf{y}) \rangle \Big\rangle$$

If F and G are regular distributions, i.e. $F = \underline{f}$, $G = \underline{g}$, with f, g having compact support, then

$$\langle \underline{f} * \underline{g} \, | \psi \rangle = \int \mathrm{d}x f(x) \int \mathrm{d}y g(y) \psi(x+y) = \int \mathrm{d}\xi \left(\int \mathrm{d}y f(\xi - y) g(y) \right) \psi(\xi)$$

It can be shown that δ_0 acts as the identity upon convolution, and $\delta_{\mathbf{x}'}$ as translation by $\mathbf{x}'$. For $F, G \in S'(\mathbb{R}^d)$ and at least one of them compactly supported, the product of Fourier transforms is the Fourier transform of the convolution:

$$\mathbb{F}_{\mathbf{x}}[F](\mathbf{k}) \cdot \mathbb{F}_{\mathbf{x}}[G](\mathbf{k}) = \frac{1}{(2\pi)^{d/2}} \mathbb{F}_{\mathbf{x}}[F * G](\mathbf{k})$$

2.2.2 Solution formulae

Definition 2.8 (Evolution operator)**.** *The* evolution operator T_t *maps suitable initial data $q_0(\mathbf{x})$ to the solution of the corresponding Cauchy problem for* (2.13) *(that is assumed to exist and be unique) at time t:*

$$(T_t \, q_0)(t, \mathbf{x}) = q(t, \mathbf{x})$$

Obviously $T_0 = \mathrm{id}$.

Theorem 2.3. *T_t is linear.*

Proof. Consider two initial data q_0 and p_0 and their time evolutions $T_t \, q_0$ and $T_t \, p_0$. Taking $\lambda, \mu \in \mathbb{R}$ consider the time evolution $T_t(\lambda q_0 + \mu p_0)$ of $\lambda q_0 + \mu p_0$. Then, by linearity of (2.13)

$$(\partial_t + \mathbf{J} \cdot \nabla)(T_t(\lambda q_0 + \mu p_0) - \lambda T_t \, q_0) = 0$$

and $T_t(\lambda q_0 + \mu p_0) - \lambda T_t \, q_0$ is a solution of (2.13) with initial data μp_0. Therefore by uniqueness of the solution to the Cauchy problem

$$T_t(\lambda q_0 + \mu p_0) - \lambda T_t \, q_0 = T_t(\mu p_0)$$

If $\mu = 0$, linearity of T_t is shown. Otherwise, again by linearity of (2.13)

$$(\partial_t + \mathbf{J} \cdot \nabla) \left(\frac{1}{\mu} T_t(\mu p_0) \right) = 0$$

$\frac{1}{\mu}T_t(\mu p_0)$ is a solution with initial data p_0, i.e.

$$\frac{1}{\mu}T_t(\mu p_0) = T_t\, p_0$$

and thus on total

$$T_t(\lambda q_0 + \mu p_0) = \lambda T_t\, q_0 + \mu T_t\, p_0$$

□

The very standard procedure of finding a solution to any linear equation such as (2.13) for sufficiently regular initial data $q_0(\mathbf{x})$ is to decompose them into Fourier modes

$$q_0(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \int \mathrm{d}\mathbf{k}\, \hat{q}_0(\mathbf{k}) \exp(\mathbb{i}\mathbf{k}\cdot\mathbf{x})$$

where $d \in \mathbb{N}$ is the dimensionality of the space. The coefficients $\hat{q}_0(\mathbf{k}) = (\hat{\mathbf{v}}_0(\mathbf{k}), \hat{p}_0(\mathbf{k}))$ of this decomposition are the Fourier transform of q_0 and $\mathbf{k}$ here characterizes the mode. The time evolution of any single Fourier mode can be found via the ansatz

$$T_t\Big(\exp(\mathbb{i}\mathbf{k}\cdot\mathbf{x})\Big) = \exp(-\mathbb{i}\omega(\mathbf{k})t + \mathbb{i}\mathbf{k}\cdot\mathbf{x})$$

where the function $\omega(\mathbf{k})$ is to be determined from the equations by inserting the ansatz. The time evolution $q(t,\mathbf{x})$ of the initial data $q_0(\mathbf{x})$ is given by adding all the time evolutions of the individual modes. For the acoustic system (2.7)–(2.8) the solution can then be found to be

$$\begin{aligned} p(t,\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \int \Bigg(& \frac{\hat{p}_0(\mathbf{k}) + \frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|}}{2} \exp(\mathbb{i}\mathbf{k}\cdot\mathbf{x} - \mathbb{i}c|\mathbf{k}|t) \\ & + \frac{\hat{p}_0(\mathbf{k}) - \frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|}}{2} \exp(\mathbb{i}\mathbf{k}\cdot\mathbf{x} + \mathbb{i}c|\mathbf{k}|t) \Bigg) \mathrm{d}\mathbf{k} \end{aligned}$$

$$\begin{aligned} \mathbf{v}(t,\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \int \Bigg(& \frac{\hat{p}_0(\mathbf{k}) + \frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|}}{2} \frac{\mathbf{k}}{|\mathbf{k}|} \exp(\mathbb{i}\mathbf{k}\cdot\mathbf{x} - \mathbb{i}c|\mathbf{k}|t) \\ & + \frac{\frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|} - \hat{p}_0(\mathbf{k})}{2} \frac{\mathbf{k}}{|\mathbf{k}|} \exp(\mathbb{i}\mathbf{k}\cdot\mathbf{x} + \mathbb{i}c|\mathbf{k}|t) \\ & + \left\{ \hat{\mathbf{v}}_0(\mathbf{k}) - \frac{\mathbf{k}}{|\mathbf{k}|} \frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|} \right\} \exp(\mathbb{i}\mathbf{k}\cdot\mathbf{x}) \Bigg) \mathrm{d}\mathbf{k} \end{aligned}$$

An analogous formula is valid in the sense of distributions:

Theorem 2.4. *Given $\hat{q}_0 = (\hat{\mathbf{v}}_0, \hat{p}_0) \in (S'(\mathbb{R}^d))^n$, $n = d+1$, the distributional solution to (2.7)–(2.8) is*

$$p(t,\mathbf{x}) = \mathbb{F}_{\mathbf{k}}^{-1}\left[\frac{\hat{p}_0(\mathbf{k}) + \frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|}}{2}\exp(-\mathbb{i}c|\mathbf{k}|t) + \frac{\hat{p}_0(\mathbf{k}) - \frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|}}{2}\exp(\mathbb{i}c|\mathbf{k}|t)\right](\mathbf{x}) \quad (2.22)$$

$$\begin{aligned}\mathbf{v}(t,\mathbf{x}) = \mathbb{F}_{\mathbf{k}}^{-1}\Bigg[&\frac{\hat{p}_0(\mathbf{k}) + \frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|}}{2}\frac{\mathbf{k}}{|\mathbf{k}|}\exp(-\mathbb{i}c|\mathbf{k}|t) + \frac{\frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|} - \hat{p}_0(\mathbf{k})}{2}\frac{\mathbf{k}}{|\mathbf{k}|}\exp(\mathbb{i}c|\mathbf{k}|t)\\ &+ \left\{\hat{\mathbf{v}}_0(\mathbf{k}) - \frac{\mathbf{k}}{|\mathbf{k}|}\frac{\mathbf{k}\cdot\hat{\mathbf{v}}_0(\mathbf{k})}{|\mathbf{k}|}\right\}\Bigg](\mathbf{x}) \quad (2.23)\end{aligned}$$

Proof. The use of S' makes sure that the Fourier transforms exist according to Theorem 2.1. Denoting the solution $q = (\mathbf{v}, p)$ (with independent variables t, $\mathbf{x}$) and its Fourier transform $\hat{q}$ (with independent variables $\omega, \mathbf{k}$), one has $q = \mathbb{F}^{-1}_{\omega,\mathbf{k}}[\hat{q}]$. q being the distributional solution to $\partial_t q + \mathbf{J}\cdot\nabla q = 0$ means

$$\left\langle (\partial_t + \mathbf{J}\cdot\nabla)\mathbb{F}^{-1}_{\omega,\mathbf{k}}[\hat{q}]\middle|\psi\right\rangle = 0 \qquad \forall\psi \in S(\mathbb{R}^d)$$

which by (2.20) is

$$\left\langle \mathbb{F}^{-1}_{\omega,\mathbf{k}}[\mathbb{i}(-\omega + \mathbf{J}\cdot\mathbf{k})\hat{q}]\middle|\psi\right\rangle = 0$$

This is only true if ω equals to one of the eigenvalues ω_n ($n = 1, \dots d+1$) of $\mathbf{J}\cdot\mathbf{k}$. For the acoustic system (2.7)–(2.8) $\mathbf{J}\cdot\mathbf{k}$ is symmetric and $\omega_n \in \{0, \pm c|\mathbf{k}|\}$.

The matrix $\mathbf{J}\cdot\mathbf{k}$ appears in the study of the Cauchy problem and bicharacteristics (see e.g. [CH62], VI, §3). Hyperbolicity guarantees its real diagonalizability. Choosing orthonormal eigenvectors e_n ($n = 1, \dots, d+1$) which fulfill

$$(\mathbf{J}\cdot\mathbf{k})e_n = \omega_n e_n$$

the vector $\hat{q}$ can be, for every $\mathbf{k}$, decomposed according to the eigenbasis of $\mathbf{J}\cdot\mathbf{k}$:

$$\hat{q}(\mathbf{k}) = \sum_{n=1}^{d+1} e_n(e_n\cdot\hat{q}(\mathbf{k}))$$

Then

$$\sum_{n=1}^{d+1}\left\langle \mathbb{F}^{-1}_{\omega,\mathbf{k}}[\mathbb{i}(-\omega+\omega_n)e_n(e_n\cdot\hat{q}_0(\mathbf{k}))]|\psi\right\rangle = 0$$

Thus, knowing that $\langle xT|\psi(x)\rangle = 0$ is solved by $T = \delta_0$,

$$e_n(e_n\cdot\hat{q}_0(\mathbf{k})) = \delta_{\omega=\omega_n}\hat{q}_{0n}(\mathbf{k})\sqrt{2\pi}$$

with arbitrary distributions $\hat{q}_{0n}(\mathbf{k}) \in S'(\mathbb{R}^d)$. The factor $\sqrt{2\pi}$ has been chosen for convenience. Performing the Fourier transform with respect to ω gives

$$q(t, \mathbf{x}) = \sum_{n=1}^{d+1} \mathbb{F}_{\mathbf{k}}^{-1}[\hat{q}_{0n}(\mathbf{k}) \exp(-\mathbb{i}\omega_n(\mathbf{k})t)]$$

in the sense of distributions. Obviously, $\hat{q}_{0n}(\mathbf{k})$ are related to the initial data:

$$\hat{q}_{0n}(\mathbf{k}) = e_n(e_n \cdot \hat{q}_0(\mathbf{k}))$$

Using this, and computing the eigenvectors explicitly for (2.7)–(2.8) completes the proof. □

Given the Fourier transform of any initial data therefore the solution can easily be constructed. The solution formulae are most conveniently expressed using spherical averages:

Definition 2.9 (Sphere and ball)**.** *Choose $r \in \mathbb{R}^+$ and $d \in \mathbb{N}^+$. Let B_r^d denote the d-ball of radius r:*

$$B_r^d := \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}| \leq r\}$$

and let the sphere S_r^{d-1} denote its boundary.

Definition 2.10 (Radial Dirac distribution and step function)**.** *Choose $r \in \mathbb{R}^+$ and $d \in \mathbb{N}^+$.*

i) The radial Dirac distribution $\delta_{|\mathbf{x}|=r}$ *is defined as*

$$\left\langle \delta_{|\mathbf{x}|=r} | \, \psi(\mathbf{x}) \right\rangle := \int_{S_r^{d-1}} \mathrm{d}\mathbf{x} \, \psi(\mathbf{x}) \qquad \forall \psi \in D(\mathbb{R}^d)$$

ii) In order to restrict a suitable function f onto the ball B_r^d define the following notation (multiplication by a step function)

$$\Theta_{|\mathbf{x}|\leq r} f(\mathbf{x}) := \begin{cases} f(\mathbf{x}) & \mathbf{x} \in B_r^d \\ 0 & else \end{cases}$$

Definition 2.11 (Spherical average)**.** *In three spatial dimensions, the spherical average at a radius r of a distribution T is given by*

$$\frac{1}{4\pi} \frac{\delta_{|\mathbf{x}|=r}}{r^2} * T$$

If T is a regular distribution $T = \underline{f}$, then by Definition 2.7 $\forall \psi \in S(\mathbb{R}^3)$

$$\frac{1}{4\pi}\left\langle \frac{\delta_{|\mathbf{x}|=r}}{r^2} * T \middle| \psi \right\rangle = \frac{1}{4\pi}\frac{1}{r^2}\int_{S_r^2} \mathrm{d}\mathbf{x} \int \mathrm{d}\mathbf{y} f(\mathbf{y})\psi(\mathbf{x}+\mathbf{y})$$
$$= \left\langle \underbrace{\frac{1}{4\pi}\int_{S_1^2} \mathrm{d}\mathbf{y} f(\mathbf{x}+r\mathbf{y})}_{} \middle| \psi(\mathbf{x}) \right\rangle$$

Here, $\int_{S_1^2} \mathrm{d}\mathbf{y}$ denotes an integration over the surface of a 2-sphere of radius 1, i.e. in spherical polar coordinates this amounts to

$$\int_0^\pi \mathrm{d}\vartheta \sin\vartheta \int_0^{2\pi} \mathrm{d}\varphi$$

Such spherical means appear already in the study of the scalar wave equation (see e.g. [Joh78, Eva98]).

Definition 2.12 (Unit normal)**.** *Define the unit normal vector* $\mathbf{n} := \frac{\mathbf{x}}{|\mathbf{x}|}$ *and denote its components by* n_i, $i = 1, 2, 3$.

Theorem 2.5 (Radial Dirac distribution)**.** *The radial derivative of the radial step function* $\Theta_{|\mathbf{x}|\leq r}$ *is closely related to the radial Dirac distribution:*

$$-\delta_{|\mathbf{x}|=r}\mathbf{n} = \nabla \underbrace{\Theta_{|\mathbf{x}|\leq r}}$$

which can be rewritten as

$$-\delta_{|\mathbf{x}|=r} = \partial_r \Theta_{|\mathbf{x}|\leq r}$$

Proof. Recall Definition 2.9 of a ball $B_r^{d+1} = \{\mathbf{x} \in \mathbb{R}^{d+1} : |\mathbf{x}| \leq r\}$. Use Definition 2.10 and Gauss' Theorem, for any $\psi \in D(\mathbb{R}^d)$:

$$-\langle \delta_{|\mathbf{x}|=r}\mathbf{n} | \psi \rangle = -\int_{S_r^d} \mathrm{d}\mathbf{x} \frac{\mathbf{x}}{|\mathbf{x}|} \cdot \psi = -\int_{B_r^{d+1}} \mathrm{d}\mathbf{x}\, \nabla \cdot \psi = -\langle \underbrace{\Theta_{|\mathbf{x}|\leq r}}\, | \nabla\psi \rangle$$
$$= \langle \nabla \underbrace{\Theta_{|\mathbf{x}|\leq r}}\, | \psi \rangle$$

Multiplying through with $\mathbf{n}$ proves the assertion. □

Lemma 2.1 (Fourier transforms)**.** *i) In three spatial dimensions, given* $r \in \mathbb{R}^+$, $\mathbf{k} \in \mathbb{R}^3$,

$$\int_{S_r^2} \mathrm{d}\mathbf{x} \exp(\mathring{\imath}\mathbf{k}\mathbf{x}) = 4\pi r^2 \frac{\sin(|\mathbf{k}|r)}{|\mathbf{k}|r}$$

ii) In three spatial dimensions, the Fourier transform of the radial Dirac distribution $\delta_{|\mathbf{x}|=r}$ *is given by*

$$\mathbb{F}_{\mathbf{x}}[\delta_{|\mathbf{x}|=r}](\mathbf{k}) = \frac{2}{(2\pi)^{1/2}} r^2 \frac{\sin(|\mathbf{k}|r)}{|\mathbf{k}|r}$$

iii) In three spatial dimensions, the Fourier transform of $\Theta_{|\mathbf{x}|\leq r}\frac{1}{|\mathbf{x}|}$ *is given by*

$$\mathbb{F}_{\mathbf{x}}\left[\frac{\Theta_{|\mathbf{x}|\leq r}}{|\mathbf{x}|}\right](\mathbf{k}) = -\frac{2}{(2\pi)^{1/2}}\frac{\cos(|\mathbf{k}|r)-1}{|\mathbf{k}|^2}$$

Proof. i) Integrating in spherical polar coordinates:

$$\int_{S_r^2} \mathrm{d}\mathbf{x}\, \exp(\mathbb{i}\mathbf{k}\mathbf{x}) = r^2\int_0^{\pi}\mathrm{d}\vartheta\, \sin\vartheta \int_0^{2\pi}\mathrm{d}\varphi\, \exp(\mathbb{i}|\mathbf{k}|r\cos\vartheta) = 4\pi r^2\frac{\sin(|\mathbf{k}|r)}{|\mathbf{k}|r}$$

ii) Using i) and Definitions 2.10 and 2.4, for any $\psi \in S(\mathbb{R}^3)$

$$\begin{aligned}\langle \mathbb{F}_{\mathbf{x}}[\delta_{|\mathbf{x}|=r}]|\psi\rangle = \langle \delta_{|\mathbf{x}|=r}|\mathbb{F}_{\mathbf{k}}[\psi]\rangle &= \int_{S_r^2}\mathrm{d}\mathbf{x}\frac{1}{(2\pi)^{3/2}}\mathrm{d}\mathbf{k}\exp(-\mathbb{i}\mathbf{k}\cdot\mathbf{x})\psi(\mathbf{k})\\ &= \Big\langle \frac{1}{(2\pi)^{3/2}}\int_{S_r^2}\mathrm{d}\mathbf{x}\exp(-\mathbb{i}\mathbf{k}\cdot\mathbf{x})\Big|\psi\Big\rangle\\ &= \Big\langle \frac{2}{(2\pi)^{1/2}}r^2\frac{\sin(|\mathbf{k}|r)}{|\mathbf{k}|r}\Big|\psi\Big\rangle\end{aligned}$$

iii) Note that $\Theta_{|\mathbf{x}|\leq r}\frac{1}{|\mathbf{x}|}$ is an L^1_{loc} compactly supported function in three spatial dimensions. Thus,

$$\begin{aligned}\mathbb{F}_{\mathbf{x}}\left[\frac{\Theta_{|\mathbf{x}|\leq \rho}}{|\mathbf{x}|}\right](\mathbf{k}) &= \frac{1}{(2\pi)^{3/2}}\int_0^{\rho}\mathrm{d}r\frac{1}{r}\int_{|\mathbf{x}|=r}\mathrm{d}\mathbf{x}\exp(-\mathbb{i}\mathbf{k}\cdot\mathbf{x})\\ &= \frac{2}{(2\pi)^{1/2}|\mathbf{k}|}\int_0^{\rho}\mathrm{d}r\sin(|\mathbf{k}|r) = -\frac{2}{(2\pi)^{1/2}}\frac{\cos(|\mathbf{k}|\rho)-1}{|\mathbf{k}|^2}\end{aligned}$$

□

Finally, with these results it is possible to derive solution formulae for (2.7)–(2.8).

Theorem 2.6 (Solution formulae)**.** *Consider the distributions*

$$q(t,\mathbf{x}) = (\mathbf{v}(t,\mathbf{x}), p(t,\mathbf{x})) \in (S'(\mathbb{R}^3))^n$$

with

$$p(t,\mathbf{x}) = \underbrace{p_0(\mathbf{x})} - \frac{1}{4\pi}\frac{1}{ct}(\operatorname{div}\ \underbrace{\mathbf{v}_0} * \delta_{|\mathbf{x}|=ct}) - \frac{1}{4\pi}(\operatorname{div}\operatorname{grad}\ \underbrace{p_0} * \underbrace{\frac{\Theta_{|\mathbf{x}|\leq ct}}{|\mathbf{x}|}}) \tag{2.24}$$

$$\mathbf{v}(t,\mathbf{x}) = \underbrace{\mathbf{v}_0(\mathbf{x})} + \frac{1}{4\pi}(\operatorname{grad}\operatorname{div}\ \underbrace{\mathbf{v}_0} * \underbrace{\frac{\Theta_{|\mathbf{x}|\leq ct}}{|\mathbf{x}|}}) - \frac{1}{4\pi}\frac{1}{ct}(\operatorname{grad}\ \underbrace{p_0} * \delta_{|\mathbf{x}|=ct}) \tag{2.25}$$

They are distributional solutions to

$$\partial_t q + \mathbf{J}\cdot\nabla q = 0$$

with $\mathbf{J}$ *given by* (2.14)*,* $d = 3$*,* $n = d + 1$ *and compactly supported*[1] L^1_{loc} *initial data* $\mathbf{v}_0(\mathbf{x}), p_0(\mathbf{x})$ *such that*

$$(\mathbf{v}(0,\mathbf{x}), p(0,\mathbf{x})) = (\underbrace{\mathbf{v}_0(\mathbf{x})}, \underbrace{p_0(\mathbf{x})}) \in (S'(\mathbb{R}^3))^n$$

Proof. Recall the differentiation rule in presence of the Fourier transform as formulated in (2.21). Denote by ∂_i differentiation with respect to the i-the direction and k_i the corresponding component of $\mathbf{k}$.

Inserting the definition of $\hat{p}_0(\mathbf{k})$ and $\hat{\mathbf{v}}_0(\mathbf{k})$ into (2.22)–(2.23) yields

$$\begin{aligned}
\mathbb{F}_{\mathbf{x}}[p(t,\mathbf{x})](\mathbf{k}) &= \mathbb{F}_{\mathbf{x}}[\underbrace{p_0}](\mathbf{k}) \cdot \cos(c|\mathbf{k}|t) - \mathbb{F}_{\mathbf{x}}[\operatorname{div} \underbrace{\mathbf{v}_0}](\mathbf{k}) \cdot \frac{\sin(c|\mathbf{k}|t)}{|\mathbf{k}|} \\
\mathbb{F}_{\mathbf{x}}[\mathbf{v}(t,\mathbf{x})](\mathbf{k}) &= \mathbb{F}_{\mathbf{x}}[\underbrace{\mathbf{v}_0}](\mathbf{k}) - \mathbb{F}_{\mathbf{x}}[\operatorname{grad}\operatorname{div} \underbrace{\mathbf{v}_0}](\mathbf{k}) \frac{\cos(c|\mathbf{k}|t) - 1}{|\mathbf{k}|^2} \\
&\quad - \mathbb{F}_{\mathbf{x}}[\operatorname{grad} \underbrace{p_0}](\mathbf{k}) \frac{\sin(c|\mathbf{k}|t)}{|\mathbf{k}|}
\end{aligned}$$

Now using Lemma 2.1 one rewrites

$$\begin{aligned}
\mathbb{F}_{\mathbf{x}}[p(t,\mathbf{x})](\mathbf{k}) &= \mathbb{F}_{\mathbf{x}}[\underbrace{p_0}](\mathbf{k}) - \mathbb{F}_{\mathbf{x}}[\operatorname{div} \underbrace{\mathbf{v}_0}](\mathbf{k}) \cdot \mathbb{F}_{\mathbf{x}}[\delta_{|\mathbf{x}|=ct}](\mathbf{k}) \frac{\sqrt{2\pi}}{2ct} \\
&\quad - \mathbb{F}_{\mathbf{x}}[\operatorname{div}\operatorname{grad} \underbrace{p_0}](\mathbf{k}) \cdot \mathbb{F}_{\mathbf{x}} \left[\underbrace{\frac{\Theta_{|\mathbf{x}|\le ct}}{|\mathbf{x}|}} \right] (\mathbf{k}) \frac{\sqrt{2\pi}}{2} \\
\mathbb{F}_{\mathbf{x}}[\mathbf{v}(t,\mathbf{x})](\mathbf{k}) &= \mathbb{F}_{\mathbf{x}}[\underbrace{\mathbf{v}_0}](\mathbf{k}) + \mathbb{F}_{\mathbf{x}}[\operatorname{grad}\operatorname{div} \underbrace{\mathbf{v}_0}](\mathbf{k}) \cdot \mathbb{F}_{\mathbf{x}} \left[\underbrace{\frac{\Theta_{|\mathbf{x}|\le ct}}{|\mathbf{x}|}} \right] (\mathbf{k}) \frac{\sqrt{2\pi}}{2} \\
&\quad - \mathbb{F}_{\mathbf{x}}[\operatorname{grad} \underbrace{p_0}](\mathbf{k}) \cdot \mathbb{F}_{\mathbf{x}}[\delta_{|\mathbf{x}|=ct}](\mathbf{k}) \frac{\sqrt{2\pi}}{2ct}
\end{aligned}$$

When rewriting $\cos(c|\mathbf{k}|t) - 1$ as a Fourier transform, $1 = \frac{\mathbf{k}\cdot\mathbf{k}}{|\mathbf{k}|^2}$ has been inserted. As both $\underbrace{\frac{\Theta_{|\mathbf{x}|\le ct}}{|\mathbf{x}|}}$ and $\delta_{|\mathbf{x}|=ct}$ have compact support, the convolutions that involve one of them are well defined (see Definition 2.7). Thus the products above can be converted into Fourier transforms of convolutions, which proves the assertion. □

Corollary 2.1. *If all the derivatives exist in the sense of functions and are integrable, then Equations* (2.24)–(2.25) *become*

$$p(t,\mathbf{x}) = p_0(\mathbf{x}) + \int_0^{ct} \mathrm{d}r\, r \frac{1}{4\pi} \int_{S_1^2} \mathrm{d}\mathbf{y}\ (\operatorname{div}\operatorname{grad} p_0)(\mathbf{x} + r\mathbf{y}) - ct \frac{1}{4\pi} \int_{S_1^2} \mathrm{d}\mathbf{y}\ \operatorname{div} \mathbf{v}_0(\mathbf{x} + ct\mathbf{y}) \tag{2.26}$$

$$\mathbf{v}(t,\mathbf{x}) = \mathbf{v}_0(\mathbf{x}) + \int_0^{ct} \mathrm{d}r\, r \frac{1}{4\pi} \int_{S_1^2} \mathrm{d}\mathbf{y}\ (\operatorname{grad}\operatorname{div} \mathbf{v}_0)(\mathbf{x} + r\mathbf{y}) - ct \frac{1}{4\pi} \int_{S_1^2} \mathrm{d}\mathbf{y}\ (\operatorname{grad} p_0)(\mathbf{x} + ct\mathbf{y}) \tag{2.27}$$

[1] Any other condition that makes a regular distribution be in $S'(\mathbb{R}^d)$ would fit here as well. For finite t, the solution to hyperbolic PDEs only involves a compact subset of the initial data, such that the behaviour of the initial data at spatial infinity can be chosen fairly arbitrary.

Proof. The formulae are transformed into the asserted ones by noting that if f is integrable, then for all $\psi \in D(\mathbb{R}^d)$

$$\frac{1}{4\pi}\left(f * \frac{\Theta_{|\mathbf{x}|\leq ct}}{|\mathbf{x}|}\right) = \frac{1}{4\pi}\int_{|\mathbf{y}|\leq ct} \mathrm{d}\mathbf{y}\frac{1}{|\mathbf{y}|}f(\mathbf{x}-\mathbf{y}) = \int_0^{ct}\mathrm{d}r\, r\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y} f(\mathbf{x}+r\mathbf{y})$$
$$\frac{1}{4\pi}\langle \underbrace{f}_{} * \delta_{|\mathbf{x}|=ct}|\psi\rangle = \left\langle \frac{1}{4\pi}\int_{S_{ct}^2}\mathrm{d}\mathbf{y} f(\mathbf{x}-\mathbf{y})\middle|\psi\right\rangle = \left\langle (ct)^2\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y} f(\mathbf{x}+ct\mathbf{y})\middle|\psi\right\rangle$$

□

Usage of the Helmholtz decomposition allows to write down a scalar wave equation for p and for the curl-free part of $\mathbf{v}$, whereas the time evolution of the curl is given by $\partial_t(\nabla\times\mathbf{v}) = 0$. The solution to the scalar wave equation is well-known ([Joh78]) and the Helmholtz decomposition of the two parts of the velocity conveniently reassembles into (2.26)–(2.27). The above formulae appear without proof in [ER13] where they have been obtained by this analogy with the scalar wave equation. A similar approach is taken in [FG17], again assuming that the solution is smooth enough. It is important to note however, that the initial data onto which the solution formulae are applied in [ER13] are not sufficiently well-behaved for the second derivatives to exist, such that a justification in the sense of distributions was needed. Theorem 2.6 of this thesis now states that (2.24)–(2.25) is the correct distributional solution.

Equation (2.25) makes obvious that any change in time of $\mathbf{v}$ is a gradient. Indeed, the curl must be stationary due to Eq. (2.12).

The spatial derivatives that appear in the solution formulae (2.24)–(2.25) can be transformed into derivatives with respect to r only. The new formulae are more useful in certain situations (as will be seen later), and display interesting properties of the solution that are discussed after stating the Theorem. Here, for notational convenience, the components of $\mathbf{y}\in\mathbb{R}^d$ are denoted by y_i, $i = 1,\dots,d$ and δ_{ij} denotes the Kronecker symbol

$$\delta_{ij} := \begin{cases} 1 & i = j \\ 0 & \text{else} \end{cases}$$

In order to simplify notation the following distributions will be used:

Theorem 2.7. *i) Given a test function $\psi\in D(\mathbb{R}^3)$ the following integral exists*

$$\int_0^{ct}\mathrm{d}r\frac{1}{r^3}\int_{S_r^2}\mathrm{d}\mathbf{y}\left(3\frac{y_iy_j}{|\mathbf{y}|^2}-\delta_{ij}\right)\psi(\mathbf{y}) \tag{2.28}$$

This defines a distribution $\Sigma_{ij}(ct)$ whose action $\langle\Sigma_{ij}(ct)|\psi\rangle$ onto a test function ψ is given by (2.28)*.*

ii) Given a test function $\psi\in D(\mathbb{R}^3)$ the following integrals exists

$$\int_0^{ct}\mathrm{d}r\frac{1}{r}\partial_r\left[\frac{1}{r}\partial_r\left(r\int_{S_r^2}\mathrm{d}\mathbf{y}\frac{y_iy_j}{|\mathbf{y}|^2}\psi(\mathbf{y})\right)-\frac{1}{r}\int_{S_r^2}\mathrm{d}\mathbf{y}\delta_{ij}\psi(\mathbf{y})\right] \tag{2.29}$$

This defines a distribution $\sigma_{ij}(ct)$ whose action $\langle\sigma_{ij}(ct)|\psi\rangle$ onto a test function ψ is given by (2.29).

Proof. One needs to prove the existence of the integrals, because including the origin into the integration domain might potentially be problematic. Therefore, for $\delta > 0$, divide the integration over $[0, ct]$ into two integrals over $[0, \delta]$ and $[\delta, ct]$ and consider $\delta \to 0$.

The reason why the integrals exist is subtle. First of all, without the test function one finds upon explicit computation

$$\int_{S_1^2} \mathrm{d}\mathbf{y}\, y_i y_j = \frac{1}{3}\int_{S_1^2} \mathrm{d}\mathbf{y}\delta_{ij} = \frac{4\pi}{3}\delta_{ij} \tag{2.30}$$

Precisely this combination $3y_iy_j - \delta_{ij}$ appears in (2.28) and (2.28).

i) Recall that $\psi \in C^\infty$, and therefore by the mean value theorem there exists $\xi(r, \mathbf{y})$ such that

$$\psi(r\mathbf{y}) = \psi(0) + r\mathbf{y}\cdot\nabla\psi(\xi\mathbf{y})$$

Then for $\delta > 0$

$$\int_0^\delta \mathrm{d}r\,\frac{1}{r^3}\int_{S_r^2}\mathrm{d}\mathbf{y}\,\left(3\frac{y_iy_j}{|\mathbf{y}|^2} - \delta_{ij}\right)\psi(\mathbf{y}) \overset{S_r^2\leftrightarrow S_1^2}{=} \int_0^\delta \mathrm{d}r\,\frac{1}{r}\int_{S_1^2}\mathrm{d}\mathbf{y}\,(3y_iy_j - \delta_{ij})\psi(r\mathbf{y})$$
$$= \int_0^\delta \mathrm{d}r\,\frac{1}{r}\int_{S_1^2}\mathrm{d}\mathbf{y}\,(3y_iy_j - \delta_{ij})\Big(\psi(0) + r\mathbf{y}\cdot\nabla\psi(\xi\mathbf{y})\Big)$$

$$\left|\int_0^\delta \mathrm{d}r\,\frac{1}{r^3}\int_{S_r^2}\mathrm{d}\mathbf{y}\,\left(3\frac{y_iy_j}{|\mathbf{y}|^2} - \delta_{ij}\right)\psi(\mathbf{y})\right| \overset{(2.30)}{=} \left|\int_0^\delta \mathrm{d}r\,\int_{S_1^2}\mathrm{d}\mathbf{y}\,(3y_iy_j - \delta_{ij})\mathbf{y}\cdot\nabla\psi(\xi\mathbf{y})\right|$$
$$\leq C\delta\|\nabla\psi\|_\infty$$

Therefore the $\frac{1}{r}$-term is harmless. Observe that the presence of the factor 3 is crucial; otherwise the integral would indeed diverge.

ii) By expanding

$$\langle\sigma_{ij}(ct)|\psi\rangle := \int_0^{ct}\mathrm{d}r\frac{1}{r}\partial_r\left[\frac{1}{r}\partial_r\left(r^3\int_{S_1^2}\mathrm{d}\mathbf{y}y_iy_j\psi(r\mathbf{y})\right) - r\int_{S_1^2}\mathrm{d}\mathbf{y}\delta_{ij}\psi(r\mathbf{y})\right]$$
$$= \int_0^{ct}\mathrm{d}r\left[\frac{1}{r}\int_{S_1^2}\mathrm{d}\mathbf{y}\,(3y_iy_j - \delta_{ij})\psi(r\mathbf{y}) + 5\partial_r\int_{S_1^2}\mathrm{d}\mathbf{y}\,(5y_iy_j - \delta_{ij})\psi(r\mathbf{y})\right.$$
$$\left. + r\partial_r^2\int_{S_1^2}\mathrm{d}\mathbf{y}\,y_iy_j\psi(r\mathbf{y})\right]$$

This reduces to the case discussed in i).

□

In the following the convention of summing over repeated indices is adapted.

Lemma 2.2. *i) The distributional analogue to*

$$\partial_r \int_{S_1^2} \mathrm{d}\mathbf{y}\, f(\mathbf{x}+r\mathbf{y}) = \int_{S_1^2} \mathrm{d}\mathbf{y}\, y_i \partial_i f(\mathbf{x}+r\mathbf{y})$$

is

$$\partial_r \left(\underline{f} * \delta_r \frac{1}{r^2} \right) = \partial_i \underline{f}(x) * \delta_r \frac{1}{r^2} n_i$$

ii) The distributional analogue to

$$r^2 \int_{S_1^2} \mathrm{d}\mathbf{y}\, \mathbf{y} \cdot \nabla p_0(\mathbf{x}+r\mathbf{y}) = \int_0^r \mathrm{d}r'\, r'^2 \int_{S_1^2} \mathrm{d}\mathbf{y}\, \nabla \cdot \nabla p_0(\mathbf{x}+r'\mathbf{y})$$

is

$$\partial_i (\underline{f} * \delta_{|\mathbf{x}|=r} n_i) = -\nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}} \underline{f}(x) * \underline{\Theta_{|\mathbf{x}|\leq r}}$$

Proof. i) Consider a test function $\psi \in D(\mathbb{R}^3)$:

$$\langle \frac{\partial}{\partial r} \underline{f} * \delta_r \frac{1}{r^2} | \psi \rangle = \Big\langle \underline{f}(\mathbf{x}) | \frac{\partial}{\partial r} \langle \delta_r(\mathbf{y}) \frac{1}{|\mathbf{y}|^2} | \psi(\mathbf{x}+\mathbf{y}) \rangle \Big\rangle = -\Big\langle \underline{f}(\mathbf{x}) | \int_{S_1^2} \mathrm{d}\mathbf{y} \frac{\partial}{\partial r} \psi(\mathbf{x}+r\mathbf{y}) \Big\rangle$$

$$\begin{aligned}
&= -\langle \underline{f}(\mathbf{x}) | \int_{S_1^2} \mathrm{d}\mathbf{y}\, \mathbf{y} \cdot \nabla_{\mathbf{x}} \psi(\mathbf{x}+r\mathbf{y}) \rangle \\
&= \left\langle \nabla_{\mathbf{x}} \underline{f}(\mathbf{x}) \Big| \int_{S_r^2} \mathrm{d}\mathbf{y} \frac{1}{|\mathbf{y}|^2} \frac{\mathbf{y}}{|\mathbf{y}|} \cdot \psi(\mathbf{x}+\mathbf{y}) \right\rangle \\
&= \left\langle \frac{\partial}{\partial x_i} \underline{f}(x) * \delta_r \frac{1}{r^2} n_i \Big| \psi \right\rangle
\end{aligned}$$

ii) Recall Theorem 2.5 which states that $-\delta_{|\mathbf{x}|=r}\mathbf{n} = \nabla\, \underline{\Theta_{|\mathbf{x}|\leq r}}$. Thus

$$\begin{aligned}
\Big\langle \partial_i (\underline{f} * \delta_{|\mathbf{x}|=r} n_i) \Big| \psi \Big\rangle &= -\Big\langle \underline{f} \Big| \langle \delta_{|\mathbf{x}|=r} n_i | \partial_i \psi \rangle \Big\rangle \\
&= \Big\langle \underline{f} \Big| \Big\langle \partial_i\, \underline{\Theta_{|\mathbf{x}|\leq r}} \, | \partial_i \psi \Big\rangle \Big\rangle = -\Big\langle \underline{f} \Big| \Big\langle \partial_i \partial_i\, \underline{\Theta_{|\mathbf{x}|\leq r}} \, | \psi \Big\rangle \Big\rangle \\
&= -\Big\langle \nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}} \underline{f}(x) * \underline{\Theta_{|\mathbf{x}|\leq r}} \Big| \psi \Big\rangle
\end{aligned}$$

□

Theorem 2.8 (Solution formulae with radial derivatives only)**.** *Consider the setup of Theorem 2.6 and write the components of the initial data* $\mathbf{v}_0$ *as* v_{0i}*,* $i = 1, 2, 3$*. The solution* (2.24)–(2.25) *can be rewritten as*

$$p(t,\mathbf{x}) = \partial_r \left(\frac{1}{4\pi} \frac{\delta_{|\mathbf{x}|=r}}{r} * \underline{p_0} \right) - \frac{1}{r} \partial_r \left(\frac{1}{4\pi} \delta_{|\mathbf{x}|=r} n_i * \underline{v_{0i}} \right) \tag{2.31}$$

$$v_j(t,\mathbf{x}) = \frac{2}{3} \underline{v_{0j}}(\mathbf{x}) - \frac{1}{r} \partial_r \left(\frac{1}{4\pi} \delta_{|\mathbf{x}|=r} n_j * \underline{p_0} \right) + \partial_r \left(\frac{1}{4\pi} \frac{\delta_{|\mathbf{x}|=r}}{r} n_i n_j * \underline{v_{0j}} \right) \tag{2.32}$$

$$- \left(\frac{1}{4\pi} \frac{\delta_{|\mathbf{x}|=r}}{r^2} (\delta_{ij} - 3 n_i n_j) * \underline{v_{0i}} \right) + \frac{1}{4\pi} \Sigma_{ij}(ct) * \underline{v_{0i}} \tag{2.33}$$

It is assumed that r *is set to* ct *after performing the derivatives.*

Equation (2.32)–(2.33) *is equivalent to*

$$v_j(t,\mathbf{x}) = \underline{v_{0j}}(\mathbf{x}) - \frac{1}{r} \partial_r \left(\frac{1}{4\pi} \delta_{|\mathbf{x}|=r} n_j * \underline{p_0} \right) + \frac{1}{4\pi} \sigma_{ij}(ct) * \underline{v_{0i}} \tag{2.34}$$

Note: The convolutions that appear in the above formulae show a particular structure of the solution: The distribution which is convoluted with the initial data carries the name of *Green's function.*

Proof. In order to transfer the r-derivatives in (2.31)–(2.34) into the derivative operators in (2.24)–(2.25) one uses the Gauss theorem for the sphere of radius r. For example, differentiating

$$\partial_r \left(\underline{p_0} * \frac{\delta_{|\mathbf{x}|=r}}{r} \right)$$

with respect to r yields

$$\partial_r^2 \left(\underline{p_0} * \frac{\delta_{|\mathbf{x}|=r}}{r} \right) = \frac{1}{r} \partial_r \left(r^2 \partial_r \left(\underline{p_0} * \frac{\delta_{|\mathbf{x}|=r}}{r^2} \right) \right)$$

by elementary manipulations. According to Lemma 2.2 i), differentiation with respect to r can be replaced by $\mathbf{n} \cdot \nabla$ inside the spherical mean:

$$= \frac{1}{r} \partial_r \left(r^2 \left(\frac{\partial}{\partial x_i} \underline{p_0} * \frac{\delta_{|\mathbf{x}|=r} n_i}{r^2} \right) \right)$$

and by Gauss theorem (Lemma 2.2 ii)) as well as Theorem 2.5

$$= -\frac{1}{r} \partial_r \left(\nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}} \underline{p_0} * \underline{\Theta_{|\mathbf{x}|\le r}} \right) = \frac{1}{r} \left(\nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}} \underline{p_0} * \delta_{|\mathbf{x}|=r} \right)$$

Integrating over r, and evaluating at $r = ct$ yields the sought identity

$$\partial_r \left(\underline{p_0} * \frac{\delta_{|\mathbf{x}|=r}}{r} \right) \Big|_{r=ct} = \underline{p_0} + \nabla \cdot \nabla \underline{p_0} * \underline{\frac{\Theta_{|\mathbf{x}|\le ct}}{|\mathbf{x}|}}$$

In a similar way the equivalence of the other terms can be shown and is omitted here. □

Corollary 2.2. *If all the derivatives exist and the functions are integrable, Equation* (2.31)–(2.33) *amounts to*

$$p(t,\mathbf{x}) = \partial_r\left(r\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\,p_0\right) - \frac{1}{r}\partial_r\left(r^2\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\,\mathbf{y}\cdot\mathbf{v}_0\right) \tag{2.35}$$

$$\begin{aligned}\mathbf{v}(t,\mathbf{x}) = {}& \frac{2}{3}\mathbf{v}_0(\mathbf{x}) - \frac{1}{r}\partial_r\left(r^2\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\,p_0\mathbf{y}\right) + \partial_r\left(r\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\,(\mathbf{v}_0\cdot\mathbf{y})\mathbf{y}\right)\\ & - \frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\;[\mathbf{v}_0 - 3(\mathbf{v}_0\cdot\mathbf{y})\mathbf{y}] - \int_0^{ct}\mathrm{d}r\frac{1}{r}\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\;[\mathbf{v}_0 - 3(\mathbf{v}_0\cdot\mathbf{y})\mathbf{y}]\end{aligned}$$

and Equation (2.34) *becomes*

$$\begin{aligned}\mathbf{v}(t,\mathbf{x}) = {}& \mathbf{v}_0(\mathbf{x}) - \frac{1}{r}\partial_r\left(r^2\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\,p_0\mathbf{y}\right)\\ & + \int_0^{ct}\mathrm{d}r\frac{1}{r}\partial_r\left[\frac{1}{r}\partial_r\left(r^3\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}(\mathbf{v}_0\cdot\mathbf{y})\mathbf{y}\right) - r\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\mathbf{v}_0\right]\end{aligned}$$

Note: Everything (if not stated explicitly) is understood to be evaluated at $\mathbf{x}+r\mathbf{y}$, and wherever it remains, $r = ct$ to be taken at the very end. For example, the term $\partial_r\left(r\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\,p_0\right)$ appearing in (2.35), if fully explicited, reads

$$\partial_r\left(r\frac{1}{4\pi}\int_{S_1^2}\mathrm{d}\mathbf{y}\,p_0(\mathbf{x}+r\mathbf{y})\right)\Bigg|_{r=ct}$$

Note that by Theorem 2.7 the integral

$$\int_0^{ct}\mathrm{d}r\frac{1}{r}\int_{S_1^2}\mathrm{d}\mathbf{y}\;[\mathbf{v}_0 - 3(\mathbf{v}_0\cdot\mathbf{y})\mathbf{y}]$$

is finite for continuous $\mathbf{v}_0$.

2.2.3 The exponential map

Using the exponential map, an analytic solution of

$$\partial_t\begin{pmatrix}\mathbf{v}\\ p\end{pmatrix} + \begin{pmatrix}\mathscr{G}p\\ \mathscr{D}\mathbf{v}\end{pmatrix} = 0$$

with $\mathscr{G}$ and $\mathscr{D}$ (not necessarily commuting) operators is given by

$$\begin{pmatrix}\mathbf{v}\\ p\end{pmatrix} = \exp(-t\mathscr{M})\begin{pmatrix}\mathbf{v}\\ p\end{pmatrix}_0$$

with

$$\mathscr{M} = \begin{pmatrix} 0 & \mathscr{G} \\ \mathscr{D} & 0 \end{pmatrix}$$

One observes that

$$\mathscr{M}^2 = \begin{pmatrix} 0 & \mathscr{G} \\ \mathscr{D} & 0 \end{pmatrix} \begin{pmatrix} 0 & \mathscr{G} \\ \mathscr{D} & 0 \end{pmatrix} = \begin{pmatrix} \mathscr{G}\mathscr{D} & 0 \\ 0 & \mathscr{D}\mathscr{G} \end{pmatrix}$$

and

$$\mathscr{M}^{2m} = \begin{pmatrix} (\mathscr{G}\mathscr{D})^m & 0 \\ 0 & (\mathscr{D}\mathscr{G})^m \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{v} \\ p \end{pmatrix} = \begin{pmatrix} \sum_{m=0}^{\infty} \frac{t^{2m}(\mathscr{G}\mathscr{D})^m}{(2m)!}\mathbf{v}_0 \\ \sum_{m=0}^{\infty} \frac{t^{2m}(\mathscr{D}\mathscr{G})^m}{(2m)!}p_0 \end{pmatrix} - \begin{pmatrix} \sum_{m=0}^{\infty} \frac{t^{2m+1}(\mathscr{G}\mathscr{D})^m}{(2m+1)!}\mathscr{G}p_0 \\ \sum_{m=0}^{\infty} \frac{t^{2m+1}(\mathscr{D}\mathscr{G})^m}{(2m+1)!}\mathscr{D}\mathbf{v}_0 \end{pmatrix} \tag{2.36}$$

In the case considered here, $\mathscr{G} = \text{grad}$ and $\mathscr{D} = \text{div}$. Restricting oneself to analytic solutions, one can show the equivalence to the one derived above by expanding all quantities around $r = 0$, e.g.

$$p_0(\mathbf{x} + r\mathbf{n}) = \sum_{m=0}^{\infty} \frac{\partial_r^m}{m!} p_0(\mathbf{x})$$

and using the following identity

Lemma 2.3. *Denoting by n_j the j-th component of $\mathbf{n}$, the integration of a tensor product of normal vectors over the unit sphere yields*

$$\frac{1}{4\pi}\int_{S_1^2} \mathrm{d}\mathbf{x}\, \underbrace{n_i n_j \cdots n_k n_\ell}_{m \text{ factors}} = \frac{1}{(m+1)!}\underbrace{\delta_{(ij}\cdots\delta_{k\ell)}}_{m/2 \text{ factors}} \qquad (m \text{ even}) \tag{2.37}$$

Here $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & else \end{cases}$ denotes the Kronecker symbol and the round brackets denote a symmetrization, i.e. a sum over all the permutations (without any prefactor included). E.g. for $m = 2$:

$$\delta_{(ij)} = \delta_{ij} + \delta_{ji} = 2\delta_{ij}$$

and thus

$$\frac{1}{4\pi}\int_{S_1^2} \mathrm{d}\mathbf{x} n_i n_j = \frac{1}{3}\delta_{ij}$$

Proof. Integration over an odd number of factors yields zero by considerations of symmetry. In case of an even number, again rotational symmetry enforces the result not to have any preferred directions. It thus has to involve only the identity δ, and the result has to remain unchanged with respect to the exchange of any two indices, which explains the symmetrization. The prefactor can then be determined by evaluating the integral in some special case that is easy to check explicitly, e.g. $\int \mathrm{d}\bar{\Omega}\, n_z n_z \cdots n_z n_z = \frac{1}{m+1}$. □

As an example consider (2.34) with only the initial data in $\mathbf{v}_0$. Here again (upper and lower) indices denote the different components

$$\begin{aligned} v^i(t,\mathbf{x}) &= v_0^i(\mathbf{x}) + \int_0^{ct} \mathrm{d}r \frac{1}{r}\partial_r \left[\frac{1}{r}\partial_r \left(r^3 \int \mathrm{d}\bar{\Omega}(v_0^k n_k) n^i\right) - r\int \mathrm{d}\bar{\Omega} v_0^i\right] \\ &= v_0^i(\mathbf{x}) + \sum_{m=0}^{\infty} \int_0^{ct} \mathrm{d}r \frac{1}{r}\partial_r \left[\frac{1}{r}\partial_r \left(r^{m+3} \int \mathrm{d}\bar{\Omega} \frac{(n^j\partial_j)^m}{m!}(v_0^k n_k) n^i\right)\right. \\ &\qquad \left. - r^{m+1}\int \mathrm{d}\bar{\Omega}\frac{(n^j\partial_j)^m}{m!} v_0^i\right] \\ &= v_0^i(\mathbf{x}) + \sum_{m=0}^{\infty} \frac{m+1}{m}\frac{r^m}{m!}\left[(m+3)\int \mathrm{d}\bar{\Omega} n_k n^i (n^j\partial_j)^m v_0^k - \int \mathrm{d}\bar{\Omega}(n^j\partial_j)^m v_0^i\right] \\ &= v_0^i(\mathbf{x}) + \sum_{m=1}^{\infty} \frac{r^{2m}}{(2m)!}\partial^i(\partial^a\partial_a)^{m-1}\partial_k v_0^k \end{aligned} \tag{2.38}$$

In the last equality the identity (2.37) and

$$(m+3)\int \mathrm{d}\bar{\Omega} n_k n^i (n^j\partial_j)^m v_0^k = \begin{cases} \frac{1}{m+1}(\partial^a\partial_a)^{m/2} v_0^i + \frac{m}{m+1}\partial^i(\partial^a\partial_a)^{\frac{m}{2}-1}\partial_k v_0^k & m \geq 2 \\ v_0^i & m = 0 \end{cases}$$

for m even was used. Equation (2.38) is the one from (2.36) which completes the example. Analogous computations confirm the other parts of the solution.

The approach of writing the solution by employing the exponential map is used in [MR01] as a guideline for numerical solutions, but the authors do not derive the solution formulae.

2.2.4 Properties of the solution

There is a number of striking differences to the one-dimensional case that appear in multiple spatial dimensions.

In one-dimensional problems only the values of the initial data appear in the solution formulae, not their derivatives. This is different in multiple dimensions and can already be observed for the scalar wave equation (as discussed e.g. in [Eva98]). A similar statement is true for the solution to Equations (2.7)–(2.8). (As explained in Section 2.1.2 this system cannot be reduced to scalar wave equations.) (2.24)–(2.25) makes the impression that second derivatives of the initial data need to be computed, but Theorem

2.8 states that the solution can be rewritten into Equation (2.33), which involves only first spatial derivatives.

In one spatial dimension, according to formulae (2.17)–(2.18), the solution at a point x depends only on initial data at points y for which $|y - x| = ct$. This motivates the following (compare e.g. [O'N83], Chapter 14)

Definition 2.13 (Causal structure of spacetime)**.** *The pair* $(t, \mathbf{x}) \in \mathbb{R}_0^+ \times \mathbb{R}^d$ *is called a* spacetime point *and refers to the spatial location* $\mathbf{x}$ *at a time* t*.*
The set $\mathscr{T}_{(t,\mathbf{x})} := \{(s, \mathbf{y}) : |\mathbf{x}-\mathbf{y}| < c(t-s)\}$ *is called* timelike past *of the spacetime point* $(t, \mathbf{x})$*.*
The set $\mathscr{N}_{(t,\mathbf{x})} := \{(s, \mathbf{y}) : |\mathbf{x} - \mathbf{y}| = c(t - s)\}$ *is called* null past *of the spacetime point* $(t, \mathbf{x})$*.*
The union $\mathscr{T}_{(t,\mathbf{x})} \cup \mathscr{N}_{(t,\mathbf{x})}$ *is called* causal past *of* $(t, \mathbf{x})$*.*
The intersection between any of these sets and the initial data slice $\{(s, \mathbf{y}) : s = 0\}$ *defines regions of space onto which the initial data can be restricted. Thus restricted initial data are referred to as* timelike initial data*,* null initial data *and* causal initial data*, respectively.*

Employing the new language one can state that in one spatial dimension the solution to the acoustic equations depends on null initial data only. In multiple spatial dimensions the situation is more complicated. Take first again the example of a scalar wave equation (2.10). Its solution depends on null initial data for odd dimensions $d = 1, 3, 5, \ldots$, whereas in even spatial dimensions $d = 2, 4, \ldots$ it also depends on timelike initial data (see e.g. [Joh78, Eva98]). For the acoustic system (2.7)–(2.8), which involves a scalar as well as a vector wave equation (2.11), the solution depends on timelike initial data even in odd spatial dimensions. Which terms exactly depend on which initial data, however, is not always easy to see. It is obvious, for example, that the pressure in Equation (2.31) only depends on null initial data in both p and $\mathbf{v}$. At the same time, $\mathbf{v}$ depends on null initial data in the pressure and timelike initial data in the velocity. However different ways to rewrite the formulae might obscure these relationships.

2.2.5 The two-dimensional Riemann problem

As an example, a particular feature of the exact solution (2.31)–(2.34), or (2.24)–(2.25) of a two-dimensional Riemann problem (in the x-z-plane for computational convenience) shall be discussed here. The initial velocity shall be $\mathbf{v}_0 = (0, 0, 1)^{\mathrm{T}}$ in the first quadrant and vanish everywhere else (see Fig. 2.1). Also everywhere $p_0 = 0$.

Denote the independent variable $\mathbf{x} =: (x, y, z)$ and the components of $\mathbf{v} =: (v_x, v_y, v_z)$, $\mathbf{v}_0 =: (v_{0x}, v_{0y}, v_{0z})$.

Recall the distribution defined in Theorem 2.7: σ_{ij} acts onto test functions as

$$\langle \sigma_{ij}(ct) | \psi \rangle := \int_0^{ct} \mathrm{d}r \frac{1}{r} \partial_r \left[\frac{1}{r} \partial_r \left(r \int_{S_r^2} \mathrm{d}\mathbf{y} \frac{y_i y_j}{|\mathbf{y}|^2} \psi(\mathbf{y}) \right) - \frac{1}{r} \int_{S_r^2} \mathrm{d}\mathbf{y} \psi(\mathbf{y}) \right]$$

Define the components of $\mathbf{y} =: (y_x, y_y, y_z)$.

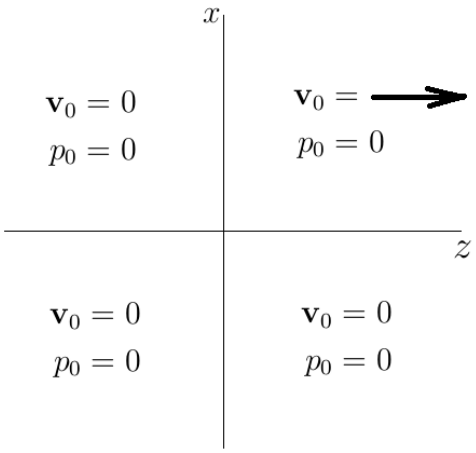

Figure 2.1: Setup of the 2-dimensional Riemann Problem. The only non-vanishing initial datum is the x-velocity in the first quadrant, indicated by the arrow. As the problem is linear its magnitude is of no importance and is chosen to be 1.

Inserting $v_{0z}(\mathbf{x}) = \Theta(x)\Theta(z)$, $v_{0x} = v_{0y} = 0$ into (2.34) gives

$$\begin{aligned}\langle v_x(t,\cdot)|\psi\rangle &= \frac{1}{4\pi}\langle \sigma_{ct}^{zx} * \underline{v_{0z}}\,|\psi\rangle \\ &= \frac{1}{4\pi}\int_0^{ct} \mathrm{d}r \frac{1}{r}\partial_r\left[\frac{1}{r}\partial_r\left(r\int_{S_r^2}\mathrm{d}\mathbf{y}\frac{y_x y_z}{|\mathbf{y}|^2}\int \mathrm{d}\mathbf{x}\,\Theta(x)\Theta(z)\psi(\mathbf{x}+\mathbf{y})\right)\right]\end{aligned}$$

Compute first

$$\begin{aligned}&\int_{S_r^2}\mathrm{d}\mathbf{y}\frac{y_x y_z}{|\mathbf{y}|^2}\int \mathrm{d}\mathbf{x}\,\Theta(x)\Theta(z)\psi(\mathbf{x}+\mathbf{y}) \\ &= \int \mathrm{d}\mathbf{x}\int_{S_r^2}\mathrm{d}\mathbf{y}\frac{y_x y_z}{r^2}\Theta(x-y_x)\Theta(z-y_z)\psi(\mathbf{x})\end{aligned}$$

This defines a regular distribution associated to

$$\int \mathrm{d}\mathbf{y}\frac{y_x y_z}{r^2}\Theta(x-y_x)\Theta(z-y_z)$$

Evaluating the integral for the special case of $x = 0$ one obtains

$$= \int_{-r}^{\min(r,z)}\mathrm{d}y_z\int_{-\sqrt{r^2-y_z^2}}^{0}\mathrm{d}y_x\frac{2y_x y_z}{r^2\sqrt{1-y_x^2-y_z^2}} = \frac{2(r^2-z^2)^{\frac{3}{2}}}{3r}\Theta(r-|z|)$$

Here the first fundamental form of the unit sphere $(1-y_x^2-y_z^2)^{-\frac{1}{2}}$ was used to express the surface integral.

The velocity becomes

$$v_x(t,\mathbf{x}) = \frac{1}{4\pi}\int_z^{ct} \mathrm{d}r \frac{1}{r}\partial_r \left[\frac{1}{r}\partial_r \left(r\frac{2(r^2-z^2)^{3/2}}{3r}\right)\right]$$

and using that for any function f

$$\frac{1}{r}\partial_r f(r^2) = 2f'(r^2)$$

one obtains

$$v_x(t,\mathbf{x}) = \frac{1}{2\pi}\int_z^{ct} \mathrm{d}r (r^2-z^2)^{-1/2} = \frac{1}{2\pi}\mathscr{L}\left(\frac{z}{ct}\right) \tag{2.39}$$

having defined

$$\mathscr{L}(s) := \ln\frac{1+\sqrt{1-s^2}}{s} = -\ln\frac{s}{2} - \frac{s^2}{4} + \mathcal{O}(s^4)$$

One can verify that $\mathrm{e}^{-\mathscr{L}(s)} = \tan\dfrac{\arcsin s}{2}$.

Note that due to the appearance of the factor $\Theta(r-|z|)$ above, $v_x(t,\mathbf{x})$ vanishes outside $|\mathbf{x}| \le ct$ by causality.

Therefore the x-component of the velocity has a logarithmic singularity at the origin, which is the corner of the initial discontinuity of the z-component. Such a behaviour of the solution does not have analogues in one spatial dimension because two different components of the velocity $\mathbf{v}$ are involved. This has already been mentioned in [AG15] in the context of self-similar solutions to Riemann Problems. Here it has been obtained by application of the general formula (2.31)–(2.34) which is not restricted to self-similar time evolution.

The solution obtained so far was restricted to $x = 0$ to simplify the presentation. This was also sufficient in order to study the appearance of a singularity. For Fig. 2.2–2.3 the integrals in (2.34) have been computed in the x-z-plane numerically using standard quadratures. They give an impression of the entire solution of the two-dimensional Riemann Problem. It is not very difficult to obtain analytic expressions in all the plane by slightly adapting the above calculations.

A vector plot of the flow is shown in Fig. 2.3.

2.3 Low Mach number limit

The system (2.5)–(2.6) has a low Mach number limit just as the Euler equations (compare [DOR10] and Section 1.2). One introduces a small parameter $\epsilon \to 0$ and inserts the scaling ϵ^{-2} by analogy with the Euler equations in front of the pressure gradient in (2.3). The system and its symmetrized version read, respectively,

$$\partial_t \mathbf{v} + \frac{1}{\epsilon^2}\operatorname{grad} p = 0 \tag{2.40}$$

$$\partial_t p + c^2 \operatorname{div}\mathbf{v} = 0 \tag{2.41}$$

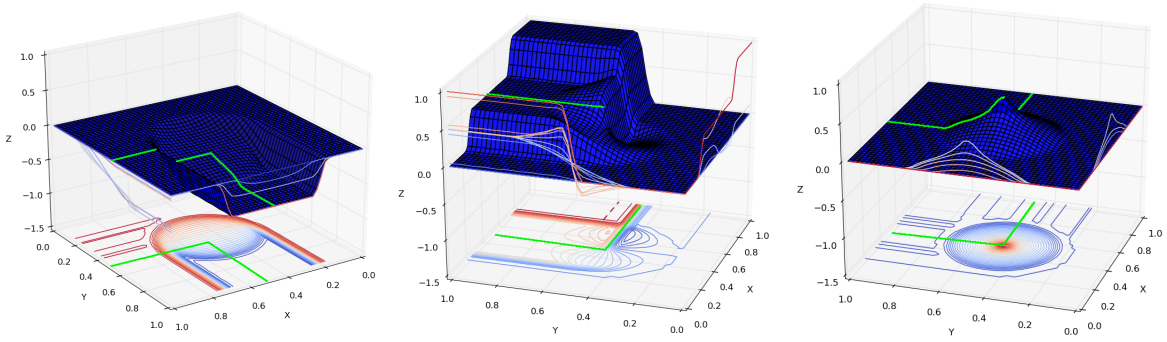

Figure 2.2: Solution of Riemann problem at time $ct = 0.25$. *Left*: Pressure. *Center*: x-velocity. *Right*: y-velocity. The smoothed out discontinuities are due to finite size sampling of the solution. In green the location of the initial discontinuity.

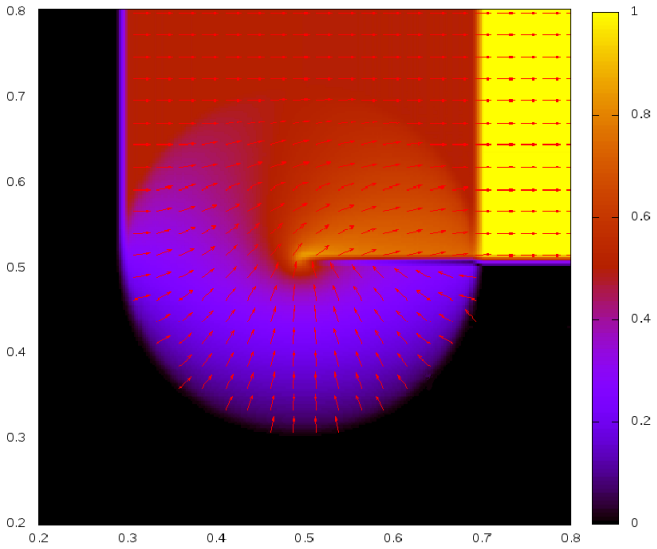

Figure 2.3: Solution of Riemann problem at time $ct = 0.2$. The direction of the velocity $\mathbf{v}(t, \mathbf{x})$ is indicated by the arrows, color coded is the absolute value $|\mathbf{v}|$.

and

$$\partial_t \mathbf{v} + \frac{c}{\epsilon} \operatorname{grad} p = 0$$
$$\partial_t p + \frac{c}{\epsilon} \operatorname{div} \mathbf{v} = 0$$

Regarding the low Mach number limit the non-symmetrized version is more natural.

There exist asymptotic scalings (compare e.g. [GV99], [BEK+17]) of the dependent and independent variables t, $\mathbf{x}$, p, $\mathbf{v}$ and of c which lead from

$$\partial_t \mathbf{v} + \nabla p = 0$$
$$\partial_t \mathbf{v} + c^2 \nabla \cdot \mathbf{v} = 0$$

to the rescaled equations (2.40)–(2.41). These scalings can be easily computed explicitly and are, in their most general form,

$$t \mapsto \epsilon^{\mathfrak{a}} t \qquad \mathbf{v} \mapsto \epsilon^{\mathfrak{c}} \mathbf{v} \tag{2.42}$$

$$\mathbf{x} \mapsto \epsilon^{\mathfrak{b}} \mathbf{x} \qquad p \mapsto \epsilon^{\mathfrak{d}} p$$

$$\mathbf{k} \mapsto \epsilon^{-\mathfrak{b}} \mathbf{k} \qquad c \mapsto \epsilon^{\mathfrak{e}} c \tag{2.43}$$

where the free parameters $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}, \mathfrak{e}$ have to satisfy

$$\mathfrak{d} + \mathfrak{a} - \mathfrak{b} - \mathfrak{c} = -2 \tag{2.44}$$

$$2\mathfrak{e} + \mathfrak{c} - \mathfrak{b} + \mathfrak{a} - \mathfrak{d} = 0 \tag{2.45}$$

Formally, in the limit $\epsilon \to 0$ the solutions to (2.40)–(2.41) have constant pressure and a divergenceless velocity. Consider an expansion of the dependent quantities as power series in ϵ:

$$\mathbf{v} = \mathbf{v}^{(0)} + \epsilon \mathbf{v}^{(1)} + \epsilon^2 \mathbf{v}^{(2)} + \mathcal{O}(\epsilon^3)$$
$$p = p^{(0)} + \epsilon p^{(1)} + \epsilon^2 p^{(2)} + \mathcal{O}(\epsilon^3)$$

Note that they start with a term $\mathcal{O}(\epsilon^0)$ because the leading order scalings have been taken care of in (2.42)–(2.43) already.

Inserting these into (2.40)–(2.41) and collecting order by order yields

$$\nabla p^{(0)} = 0 \qquad \operatorname{div} \mathbf{v}^{(0)} = 0$$
$$\nabla p^{(1)} = 0$$

Here as usual it has been assumed that $\partial_t p^{(0)} = 0$, as the equations are considered on all the $\mathbb{R}^d$.

This can also be shown in more detail. First, there is an alternative interpretation to the limit of low Mach number:

Theorem 2.9 (Limit equivalence)**.** *The limit of low Mach number $\epsilon \to 0$ for* (2.40)–(2.41) *is the same as substituting $\frac{t}{\epsilon}$ for t and letting $\epsilon \to 0$ (long time limit) for*

$$\partial_t \mathbf{v} + \nabla p = 0$$
$$\partial_t p + c^2 \nabla \cdot \mathbf{v} = 0$$

Proof. In Section 2.2, an exact solution to (2.40)–(2.41) is derived by using the Fourier transform. Here a small aspect of this derivation shall be taken as starting point. Recall that any Fourier mode, characterized by its wave vector $\mathbf{k}$, has the form

$$\exp(-\mathbb{i}\omega t + \mathbb{i}\mathbf{k} \cdot \mathbf{x})$$

The time evolution is governed by $\omega \in \mathbb{R}$, which may be a function of $\mathbf{k}$. Denoting by $\mathbf{J}$ the Jacobian of this system, in Section 2.2 it is shown that, up to factors, ω is an eigenvalue of $\mathbf{J} \cdot \mathbf{k}$.

Upon explicit calculation, the eigenvalues of $\mathbf{J}\cdot\mathbf{k}$ for the acoustic equations are 0 and $\pm\frac{c|\mathbf{k}|}{\epsilon}$. Assembling the time evolution of the Fourier modes gives

$$q(t,x) = \hat{q}_0(\mathbf{k})\exp(\mathbb{i}\mathbf{k}\cdot\mathbf{x}) + \hat{q}_\pm(\mathbf{k})\exp\left(\mp\mathbb{i}|\mathbf{k}|\frac{ct}{\epsilon} + \mathbb{i}\mathbf{k}\cdot\mathbf{x}\right)$$

with $\hat{q}_0, \hat{q}_\pm$ following from the initial data. One observes the two possible readings of the time-evolving part: either as the low Mach number limit $\frac{c}{\epsilon}t$ or as the long-time limit $c\frac{t}{\epsilon}$. □

Thus, decreasing ϵ by a factor of 10 and looking at the solution at time $t = 1$, is the same as leaving ϵ as it was, and looking at the solution at time $t = 10$.

The behaviour of numerical schemes in this limit is studied in Section 4.1.1.

Chapter 3

Numerical stationary states for linear multi-dimensional systems

Contents

Convergence of a numerical method means that the discrete solution approaches the exact one as the discretization length (grid spacing, time step, ...) decreases. However, sometimes the discrete solution needs to reflect certain properties of the exact solution already at finite discretization. A prominent example is the positivity of the density and pressure in the context of the Euler equations. Indeed, a violation immediately leads to the impossibility of computing the speed of sound $c = \sqrt{\gamma p/\rho}$. However small the violation of, say, $\rho > 0$, the consequences are dramatic.

The exact solution might exhibit further properties that one would like to see reflected by the discrete solution. For the acoustic equations (2.7)–(2.8), for instance, the vorticity $\omega = \nabla \times \mathbf{v}$ is stationary:

$$\partial_t \omega = 0$$

Obviously, ω cannot be exactly computed just from a set of discrete values of $\mathbf{v}$. One might however still ask, whether there exists any *discretization* of $\nabla \times \mathbf{v}$ that remains stationary upon the numerical evolution of $\mathbf{v}$. The answer is yes: there is a large body of work related to such so-called *vorticity preserving* schemes for the acoustic equations (see Definition 4.1). Examples of schemes have been presented in e.g. [LMMW00, MR01, Sid02, JT06, LFS07, MT11, LR14]; they all have been constructed for particular choices of the discrete vorticity.

This Section presents a more general framework. Instead of focusing on a particular discretization of vorticity, here a scheme is called vorticity preserving, if there exists *any* stationary discrete approximation to $\nabla \times \mathbf{v}$.

There is an alternative interpretation of a *vorticity preserving* scheme, which has been given the name *stationarity preserving* ([Bar17a, Bar17b]). Both concepts are equivalent – a scheme that is vorticity preserving is also stationarity preserving and vice versa. Interestingly, stationarity preservation allows for a new understanding of the behaviour of numerical schemes in the limit of low Mach numbers. This displays the connection between vorticity preserving schemes and low Mach number compliant schemes for the acoustic equations.

These concepts have been developed with a strong focus on the acoustic equations. However, they are of general applicability to all linear hyperbolic systems, and Section 4.8 shows an example how they can even be used for linear systems endowed with source terms. Therefore this Section introduces stationarity preservation for general linear hyperbolic systems in multiple spatial dimensions, and the Sections 4.4.1, 4.5 discuss applications of the framework to the acoustic equations. In particular it is there that the connection to the low Mach number limit is clarified. This Section is largely based on work published in [Bar17a].

3.1 Stationary states

First, in this Section stationary states of both the linear hyperbolic systems and their discretizations are discussed. Nontrivial stationary states are interesting for numerics because they turn out to be the key to understanding many more properties, like the low Mach number limit and vorticity preservation, which are subjects of later sections. The main result for the discrete situation is Theorem 3.3.

3.1.1 Continuous case

This Section deals with stationary states for the general hyperbolic linear $n \times n$ system in d spatial dimensions ($\mathbf{J}$ being a d-dimensional vector of matrices $(J_x, J_y, \ldots)$)

$$\partial_t q + \mathbf{J} \cdot \nabla q = 0 \tag{3.1}$$
$$q : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}^n$$

Although this analysis is inspired by a particular example of such system, namely the acoustic equations discussed in Section 2, stationary states (both at continuous and discrete level) can be fruitfully studied for more general problems, as is done in this Section.

Obviously, data that satisfy $\nabla q = 0$ remain stationary for all times. In general:

Definition 3.1 (Trivial stationary states)**.** *A solution q to* (3.1) *is called a* trivial stationary state *if*

$$\nabla q \in \ker \mathbf{J}$$

i.e.

$$\partial_x q \in \ker J_x \qquad \text{and} \qquad \partial_y q \in \ker J_y \qquad \dots$$

Recall that, given any $0 \neq \mathbf{k} \in \mathbb{R}^d$, hyperbolicity of the system (3.1) guarantees diagonalizability of $\mathbf{J} \cdot \mathbf{k}$ with real eigenvalues. An eigenvalue zero precisely corresponds to the existence of a richer set of stationary states:

Theorem 3.1 (Non-trivial stationary states)**.** *Given* $0 \neq \mathbf{k} \in \mathbb{R}^d$*, if* $\det(\mathbf{k} \cdot \mathbf{J})$ *vanishes for all* $\mathbf{k}$*, then there exist non-trivial stationary states of* (3.1)*.*

Proof. Consider the Fourier transform of (3.1) by inserting

$$q(t, \mathbf{x}) = \hat{q} \exp(-\mathbb{i}\omega t + \mathbb{i}\mathbf{k} \cdot \mathbf{x})$$

to obtain the eigenvalue problem $\omega \hat{q} = \mathbf{J} \cdot \mathbf{k} \hat{q}$. Every Fourier mode evolves in time as $\exp(-\mathbb{i}\omega t)$ with the corresponding eigenvalue ω that depends on $\mathbf{k}$. Initial data that remain stationary have to be such that their time evolution is governed by $\omega = 0$. Their Fourier transform therefore is (parallel to) the eigenvector $\hat{q}_0$ of $\mathbf{J} \cdot \mathbf{k}$ which corresponds to an eigenvalue zero. As $\mathbf{k} \neq 0$, the stationary state is not constant in space, i.e. not trivial. The eigenvalue zero makes the determinant vanish. □

Trivial stationary states typically do not pose challenges to the numerical schemes. The remainder of this thesis therefore focuses its attention onto systems for which $\det(\mathbf{J} \cdot \mathbf{k}) = 0 \ \forall \mathbf{k}$, i.e. those that possess non-trivial stationary states. This is the case for the acoustic equations to be considered later (Corollary 4.1 in Section 4).

Definition 3.2 (Constant of motion)**.** *Given an evolution equation for a function* $q(t, \mathbf{x})$*, a* constant of motion *is a function* Ω *of* q*, such that*

$$\partial_t \Omega = 0$$

is a consequence of the evolution equation for q*.*

Theorem 3.2 (Constant of motion)**.** *The existence of nontrivial stationary states for the system* (3.1) *is equivalent to the existence of a constant of motion (one for each zero eigenvalue), i.e. a linear function* $\Omega q : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}$ *of the solution* q*, which does not evolve in time for any initial data:*

$$\partial_t(\Omega q) = 0$$

Proof. Having only spatial coordinates Fourier transformed, i.e. inserting

$$q(t, \mathbf{x}) = \hat{q}(t) \exp(\mathbb{i}\mathbf{k} \cdot \mathbf{x})$$

yields $\partial_t \hat{q}(t) = -\mathbb{i}(\mathbf{k} \cdot \mathbf{J})\hat{q}(t)$. Assume that $\det(\mathbf{k} \cdot \mathbf{J}) = 0$, such that there exists a left eigenvector $\hat{\Omega}$ that belongs to the eigenvalue zero. Take Ωq to be the inverse Fourier transform of $\hat{\Omega}\hat{q}$. Then $\partial_t(\Omega q) = 0$ for any q. □

An example of such a constant of motion for the acoustic equations is given in Corollary 4.2 of Section 4.

3.1.2 Stationarity preserving schemes

The study of stationary states by means of the Fourier transform in the proof of Theorem 3.1 has a natural equivalent in the discrete sense.

Assume Eqn. (3.1) to be solved numerically on a rectangular d-dimensional grid. Recall Definition 0.1. The cells of the grid shall be indexed by $I \in \mathbb{Z}^d$, with components I_m, $m = 1, \ldots, d$. Again, q_I is the value of q in cell I. Given $\mathbf{k} \in \mathbb{R}^d$, its components are denoted by k_m, $m = 1, \ldots, d$. The Fourier ansatz now reads

$$q_I = \hat{q} \exp\left(-\mathbb{i}\omega t + \mathbb{i} \sum_{m=1}^{d} I_m k_m \Delta x_m \right)$$

Example 3.1. *In 2-d one has $I = (i, j)$ and recall from Definition 0.1 $\Delta x_1 =: \Delta x$, $\Delta x_2 =: \Delta y$. Therefore*

$$\begin{aligned} q_{ij} &= \hat{q} \exp\left(\mathbb{i}[-\omega t + i k_1 \Delta x_1 + j k_2 \Delta x_2]\right) \\ &= \hat{q} \exp\left(\mathbb{i}[-\omega t + i k_x \Delta x + j k_y \Delta y]\right) \end{aligned}$$

◁

Definition 3.3 (Translation factor)**.** *The shift by one cell is conveyed by the* translation factor $t_m := \exp(\mathbb{i} k_m \Delta x_m)$.

This allows to write

$$q_I = \hat{q} \exp\left(-\mathbb{i}\omega t\right) \prod_{m=1}^{d} t_m^{I_m} \tag{3.2}$$

Definition 3.4 (Stencil)**.** *Consider a mapping that assigns to every cell $I \in \mathbb{Z}^d$ of a d-dimensional Cartesian grid with values $\{q_J \in \mathbb{R}^n : J \in \mathbb{Z}^d\}$ a new value Q_I. Assume that this mapping has the property that Q_I depends only on q_I and the values in the neighbours of cell I in a way that is independent of I. Then this mapping is called a* stencil.

Note: In the following, unless stated differently, only *compact stencils* are considered, i.e. stencils such that Q_I depends only on values in finitely many neighbouring cells.

Any linear numerical stencil at cell I can be written as

$$\sum_{S \in [-N,N]^d \subset \mathbb{Z}^d} \alpha_S q_{I+S} \tag{3.3}$$

Example 3.2. *On a two-dimensional grid the central difference in x-direction is*

$$\frac{1}{2}\left(q_{i+1,j} - q_{i-1,j}\right)$$

with $N = 1$ and $\alpha_{1,0} = \frac{1}{2}$, $\alpha_{-1,0} = -\frac{1}{2}$ and all other α vanish. ◁

In general of course, as the object of study are systems of equations, q is a vector, and every α_S is an $n \times n$ matrix.

Applying the Fourier transform to the stencil gives rise to a discrete analogue of the condition $\det(\mathbf{k} \cdot \mathbf{J})$ found in Theorem 3.1. The role of $\mathbf{k} \cdot \mathbf{J}$ is played by the *evolution matrix* $\mathcal{E}$:

Definition 3.5 (Evolution matrix)**.** *The* evolution matrix *associated to the stencil* (3.3) *is the matrix*

$$\mathcal{E} \quad = - \sum_{S \in [-N,N]^d} \mathbb{i} \alpha_S \prod_{m=1}^{d} t_m^{S_m}$$

Definition 3.6 (Stationarity preservation)**.** *A consistent linear scheme is called* stationarity preserving *if it discretizes all nontrivial stationary states of the equation.*

Theorem 3.3 (Stationarity preservation)**.** *A necessary condition for a consistent linear scheme to be* stationarity preserving *is the vanishing of the determinant* $\det \mathcal{E}$ *of its evolution matrix. If* $\mathbf{k} \cdot \mathbf{J}$ *has only one vanishing eigenvalue, then this condition is sufficient. The numerical stationary states are discretizations of the stationary states of the PDE.*

Proof. The stencil (3.3), inserting (3.2), has the Fourier transform

$$\exp\left(-\mathbb{i}\omega t\right) \left(\prod_{m=1}^{d} t_m^{I_m} \right) \sum_{S \in [-N,N]^d} \alpha_S \left(\prod_{m=1}^{d} t_m^{S_m} \right) \hat{q} \tag{3.4}$$

The semidiscrete scheme, assumed to be a consistent discretization of (3.1),

$$\partial_t q_I \quad + \sum_{S \in [-N,N]^d \subset \mathbb{Z}^d} \alpha_S q_{I+S} = 0 \tag{3.5}$$

after the Fourier transform is taken leads to $\omega \hat{q} + \mathcal{E}\hat{q} = 0$. From here the argument is exactly the same as in the proof of Theorem 3.1, and the existence of nontrivial stationary states is characterized by $\det \mathcal{E} = 0$. They are a discretization of the analytical stationarity condition $\mathbf{J} \cdot \nabla q = 0$ by consistency. □

In the following the focus shall lie on the situation when $\ker(\mathbf{J} \cdot \mathbf{k})$ is one-dimensional. Otherwise additionally to the determinant condition of Theorem 3.3 one has to make sure that the nullities (or ranks) of $\mathcal{E}$ and $\mathbf{J} \cdot \mathbf{k}$ coincide.

The numerical stationary states, i.e. those states that are kept exactly[1] stationary by the numerics are given by the eigenvector of $\mathcal{E}$ corresponding to eigenvalue zero.

Analogously to Theorem 3.2, then together with the right eigenvector (characterizing the numerical stationary states) one is given a corresponding left eigenvector which yields a (numerical) constant of motion.

[1] *Exactly* here and in the following means "up to machine precision".

Theorem 3.4. *Any stationarity preserving scheme with the evolution matrix $\mathcal{E}$ gives rise to a numerical constant of motion, whose Fourier transform is given by the left eigenvector belonging to eigenvalue zero of $\mathcal{E}$.*

Proof. Replacing $\mathbf{J} \cdot \mathbf{k}$ by $\mathcal{E}$ in Theorem 3.2 proves the assertion. □

3.2 Multi-dimensional schemes

3.2.1 Stationarity-consistent stencils

Central derivatives have been shown to be stationarity preserving, but they are known to be unstable under forward Euler time integration. Therefore the natural question is whether it is possible to write down a stabilizing diffusion that does not spoil the property of stationarity preservation.

3.2.1.1 Continuous case

Consider, as an example, the linear system (3.1) in $d = 2$ spatial dimensions

$$\partial_t q + J_x \partial_x q + J_y \partial_y q = 0 \tag{3.6}$$

with $q : \mathbb{R}_0^+ \times \mathbb{R}^2 \to \mathbb{R}^n$ and J_x, J_y being $n \times n$ matrices. The stationary solutions are given by

$$J_x \partial_x q + J_y \partial_y q = 0 \tag{3.7}$$

Consider now a numerical scheme for Eqn. (3.6), e.g. a finite volume scheme or a finite difference scheme. Before the concepts can be detailed for the discrete case, it is easier to discuss them in a continuous situation (as is done e.g. in [Sid02, LFS07]), where effects of the numerics are taken into account as a diffusive term, e.g. as

$$\partial_t q + J_x \partial_x q + J_y \partial_y q = D_x \partial_x^2 q + D_y \partial_y^2 q + \bar{D} \partial_x \partial_y q \tag{3.8}$$

with D_x, D_y, $\bar{D}$ matrices. Of course, for stability there are certain conditions that these matrices need to fulfill, which shall not matter for the moment.

Consider now initial data fulfilling (3.7), such that they are preserved exactly in time if the evolution is governed by (3.6). If the initial data are evolved according to (3.8), then their initial evolution $\partial_t q$ is given entirely by the diffusion

$$D_x \partial_x^2 q + D_y \partial_y^2 q + \bar{D} \partial_x \partial_y q$$

In this situation, a *stationarity consistent diffusion* would be a term containing second derivatives, that vanishes whenever $J_x \partial_x q + J_y \partial_y q$ vanishes.

As an example, one could take

$$J_x \partial_x^2 q + J_y \partial_x \partial_y q = \partial_x (J_x \partial_x q + J_y \partial_y q)$$

i.e.

$$D_x = J_x \qquad D_y = 0 \qquad \bar{D} = J_y$$

or anything proportional to it. Observe also the necessary appearance of mixed second derivatives.

3.2.1.2 Discrete case

Consider a linear $n \times n$ system

$$\partial_t q + \mathbf{J} \cdot \nabla q = 0$$

A linear stencil, in general, has the shape (3.3)

$$\sum_{S \in [-N,N]^d \subset \mathbb{Z}^d} \alpha_S q_{I+S}$$

with α_S an $n \times n$ matrix.

Definition 3.7 (Stationarity consistency)**.** *Consider discrete data such that a linear stencil $\mathcal{A}$, evaluated on them, vanishes everywhere on the grid. If a stencil $\mathcal{B}$, that is evaluated on the same data always vanishes as well, then $\mathcal{B}$ is called* stationarity-consistent *with $\mathcal{A}$.*

The vanishing of a stencil can be best explored via the Fourier transform. According to (3.4) the Fourier transform of any stencil $\sum_S \alpha_S q_{I+S}$ is proportional to the Laurent polynomial

$$\sum_S \alpha_S \hat{q} \prod_{m=1}^{d} t_m^{S_m}$$

in the variables $\{t_m\}_{1 \le m \le d}$ and is linear in $\hat{q}$. By trivially factoring out the smallest power this establishes a mapping between stencils in d dimensions and polynomials in d variables.

$\mathcal{B}$ being stationarity-consistent with $\mathcal{A}$ means that the Fourier transform of $\mathcal{B}$ contains the Fourier transform of $\mathcal{A}$ as a factor.

Theorem 3.5 (Stationarity consistency)**.** *Two linear stencils $\mathcal{A}$ and $\mathcal{B}$ are stationarity-consistent to each other if their Fourier transforms $\hat{\mathcal{A}}$, $\hat{\mathcal{B}}$ are related by a factor $\hat{\mathcal{F}}$ that does not depend on $\hat{q}$:*

$$\hat{\mathcal{B}} = \hat{\mathcal{A}} \cdot \hat{\mathcal{F}}$$

Proof. The condition of stationarity-consistency means that the Fourier transform of stencil $\mathcal{B}$ can be written as the Fourier transform of $\mathcal{A}$ times some factor. This factor has to respect that both $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$ are linear in $\hat{q}$. It thus cannot itself depend on $\hat{q}$. □

Note: This factor does not need to be a polynomial in t_x, t_y, ... itself, but can be an arbitrary function. As usual, the product in Fourier space can be rewritten as a convolution in real space. Thus, if $\mathcal{B}$ is stationarity-consistent with $\mathcal{A}$, then there exists a stencil $\mathcal{F}$, such that $\mathcal{B}$ is the convolution of $\mathcal{A}$ and $\mathcal{F}$:

$$\mathcal{B} = \mathcal{A} * \mathcal{F}$$

If its Fourier transform $\hat{\mathcal{F}}$ is not a polynomial, then the stencil $\mathcal{F}$ is not compactly supported, i.e. it is a stencil that involves values of the grid that are arbitrarily far away.

Note: For scalar problems, any two stencils are stationarity-consistent. This is different when systems are considered.

A stationarity-consistent diffusion easily gives rise to a stationarity preserving scheme:

Theorem 3.6. *If $\mathcal{A}$ is a stationarity preserving linear discretization of $\mathbf{J} \cdot \nabla q$, then adding a stationarity-consistent diffusion does not destroy the stationarity preservation property.*

Proof. The semidiscrete scheme that involves only $\mathcal{A}$ reads

$$\partial_t q + \mathcal{A} = 0$$

Write the Fourier transform of $\mathcal{A}$ by linearity in q as $\hat{a}\hat{q}$ with $\hat{a}$ a matrix valued function that depends only on $t_x, t_y, \ldots$. According to Theorem 3.3, the scheme being stationarity preserving implies $\det \hat{a} = 0$.

Consider now a stencil $\mathcal{B}$ that is stationarity-consistent with $\mathcal{A}$. By Theorem 3.5, its Fourier transform $\hat{b}\hat{q}$ is related to the Fourier transform of $\mathcal{A}$ by a factor $\hat{\mathcal{F}}$, that does not depend on $\hat{q}$:

$$\hat{b}\hat{q} = \hat{\mathcal{F}}\hat{a}\hat{q}$$

The condition of stationarity preservation for the new scheme

$$\partial_t q + \mathcal{A} + \mathcal{B} = 0$$

now reads $\det(\hat{a} + \hat{b}) = \det((\mathbb{1} + \hat{\mathcal{F}})\hat{a}) = 0$. Also any eigenvector of $\hat{a}$ that belongs to an eigenvalue zero is a such for the matrix $(\mathbb{1} + \hat{\mathcal{F}})\hat{a}$. This proves the assertion. □

3.2.2 Construction principles

Via the Fourier transform there exists a mapping between linear stencils in d dimensions and Laurent polynomials in d variables. Stationarity consistency deals with (possibly non-polynomial) factors that relate these Fourier transforms. This allows to make a number of general statements about the shape of stationarity consistent stencils.

One typically is interested in finding the continuous differential operator that is formally approximated by a given discrete one (e.g. $\frac{q_{i+1}-q_{i-1}}{2\Delta x}$). This can be done by expanding the discrete operator as Taylor series in Δx, i.e.

$$q_{i+1} = q + \Delta x \cdot \partial_x q + \frac{1}{2}\Delta x^2 \cdot \partial_x^2 q + \ldots$$

Stationarity preservation uses the language of the Fourier transform. The Fourier transform of

$$\sum_{j=-k}^{k} \alpha_j q_{i+j}$$

is proportional to a polynomial in t_x:

$$\sum_{j=-k}^{k} \alpha_j t_x^j = \frac{1}{t_x^k} \sum_{j=0}^{2k} \alpha_{j-k} t_x^j$$

which by the fundamental theorem of algebra is

$$\frac{1}{t_x^k} \prod_{j=1}^{2k} (t_x - s_j)$$

for some set of $s_j \in \mathbb{C}$.

When dealing with stationarity preservation, the discrete differential operators appear via their Fourier transforms. It is useful to be able to say something about the continuous operators that they approximate without having to undo the Fourier transform and expand in Δx. The theorems of the beginning of this Section thus aim at developing a language that relates Fourier transforms of discrete differential operators to the continuous operators.

Lemma 3.1. *The Fourier transform of a stencil that approximates ∂_x^n contains precisely n factors $t_x - 1$.*

Proof. Recall that in Fourier space, the derivative $\partial_x^n q$ becomes $(\mathbb{i}k)^n \hat{q}$. Expand $t_x = \exp(\mathbb{i}\Delta x k)$ in powers of Δx as $\Delta x \to 0$. Then

$$\frac{1}{t_x^k} \prod_{j=1}^{2k} (t_x - s_j) = \frac{1}{1 + \mathcal{O}(\Delta x)} \prod_{j=1}^{2k} (1 - s_j + \mathbb{i}k\Delta x + \mathcal{O}(\Delta x^2))$$

This is only proportional to $(\mathbb{i}k)^n$, if in precisely n of the linear factors $s_j = 1$. □

Lemma 3.2 (Normalization)**.** *Consider the Fourier transform*

$$\frac{1}{t_x^k} \prod_{j=1}^{2k} (t_x - s_j)$$

with $s_j = 1$ for $j = 1, 2, \ldots, n$ and otherwise $s_j \neq 1$. Then, as $\Delta x \to 0$, it approximates $A\Delta x^n \partial_x^n$ with

$$A = \prod_{j=1, s_j \neq 1}^{2k} (1 - s_j)$$

Proof. The Fourier transform approximates an n-th order differential operator by Lemma 3.1 because precisely n values of j with $s_j = 1$. Expand again in powers of Δx:

$$\begin{aligned}\frac{1}{t_x^k}\prod_{j=1}^{2k}(t_x - s_j) &= \frac{1}{1+\mathcal{O}(\Delta x)}\prod_{j=1,s_j=1}^{2k}(\mathbb{i}k\Delta x + \mathcal{O}(\Delta x^2))\prod_{j=1,s_j\neq 1}^{2k}(1 - s_j + \mathcal{O}(\Delta x)) \\ &= (\mathbb{i}k)^n \Delta x^n \prod_{j=1,s_j\neq 1}^{2k}(1-s_j) + \text{higher order terms}\end{aligned}$$

□

Lemma 3.3. *Consider a linear stencil*

$$\sum_{j=-k}^{k}\alpha_j q_{i+j}$$

Its Fourier transform is invariant under the mapping $t_x \mapsto \frac{1}{t_x}$, iff it is symmetric, i.e. if $a_k = \alpha_{-k}\,\forall k$.

Proof. Consider the Fourier transform of a symmetric stencil and apply the mapping $t_x \mapsto \frac{1}{t_x}$ to it.

$$\frac{1}{t_x^k}\sum_{j=0}^{2k}\alpha_{j-k}t_x^j \overset{t_x\mapsto\frac{1}{t_x}}{=} t_x^k\sum_{j=0}^{2k}\alpha_{j-k}\frac{1}{t_x^j} = \frac{1}{t_x^k}\sum_{j=0}^{2k}\alpha_{j-k}t_x^{2k-j} \overset{j\mapsto 2k-j}{=} \frac{1}{t_x^k}\sum_{j=0}^{2k}\alpha_{k-j}t_x^j$$

which is true if $a_k = \alpha_{-k}\,\forall k$. The converse is obtained by reading the equations in the opposite direction. □

Theorem 3.7 (One-dimensional symmetric stencil)**.** *Consider a one-dimensional symmetric linear stencil on cells $\{x_{i-k},\dots,x_i,\dots,x_{i+k}\}$, $k\in\mathbb{N}$ for ∂_x. Its Fourier transform is proportional the Laurent polynomial*

$$\frac{1}{t_x^k}\prod_{j=1}^{2k}(t_x - s_j)$$

with $s_1 = 1$ and $\mathbb{C}\ni s_j\neq 1$ for $j>1$. Moreover, if $s_j\neq -1$, there exists $j'\in[1,2k]$ such that $s_{j'} = \frac{1}{s_j}$. Therefore the Laurent polynomial up to a constant can be written as

$$\frac{1}{t_x^k}(t_x-1)\underbrace{(t_x+1)\cdots(t_x+1)}_{N\text{ times},1\leq N\leq 2k-1,N\text{ odd}}\underbrace{\left(t_x^2 - \frac{r_1^2+1}{r_1}t_x + 1\right)\cdots\left(t_x^2 - \frac{r_M^2+1}{r_M}t_x+1\right)}_{M=\frac{2k-1-N}{2}\text{ times}} \tag{3.9}$$

with $r_j\in\mathbb{C}$ obtained by reordering and selecting from $\{s_j\}$ such that $r_j^2\neq 1\ \forall j = 1,\dots,\frac{2k-1-N}{2}$. N can be freely chosen in its range.

The converse is true as well.

Note: The proportionality factor is obtained using Lemma 3.2.

Proof. i) As the stencil is one-dimensional, its Fourier transform involves a Laurent polynomial in t_x only.

In the case under consideration, by Lemma 3.1 the Fourier transform contains precisely one factor $(t_x - 1)$. Thus, by reordering, $s_1 = 1$ and for $j > 1$, $1 \neq s_j \in \mathbb{C}$.

The symmetry requirement, by Lemma 3.3, means that the stencil is invariant under $t_x \mapsto \frac{1}{t_x}$:

$$\frac{1}{t_x^k}\prod_{j=1}^{2k}(t_x - s_j) \overset{!}{=} t_x^k \prod_{j=1}^{2k}\left(\frac{1}{t_x} - s_j\right) = \frac{1}{t_x^k}\prod_{j=1}^{2k}(1 - s_j t_x)$$

which means that if s_j is a zero, then the polynomial also vanishes at $t_x = \frac{1}{s_j}$. If there exists an $s_j^2 \neq 1$, then there is also a j' such that $s_{j'} = \frac{1}{s_j}$. Therefore there can only be an even number of non-one-non-minus-one zeros of the polynomial. They can be reordered in pairs:

$$(t_x - s_j)\left(t_x - \frac{1}{s_j}\right) = t_x^2 - \left(s_j + \frac{1}{s_j}\right)t_x + 1$$

The degree of the polynomial (up to the prefactor t_x^{-k}) is even as well. Thus there is at least one factor $(t_x + 1)$.

ii) The converse follows from Lemma 3.3, which states that the one-dimensional stencil associated to the Fourier transform (3.9) is symmetric.

□

In general, a polynomial of even degree either approximates a differential operator ∂_x^{2n} of an even order and thus has an even number of $(t_x - 1)$ factors, or it approximates a differential operator of odd order and therefore must have at least one factor $(t_x + 1)$, alongside with an odd number of factors $(t_x - 1)$.

Assume in the following $s_1 = 1$, $s_2 = -1$, and $s_j \neq 1 \, \forall j > 2$ (by reordering). A dimensionally split discretization of the divergence $\partial_x u + \partial_y v$ involving the cells

$$\begin{array}{ccccc} & & x_{i,j+k} & & \\ & & \vdots & & \\ x_{i-k,j} & \cdots & x_{ij} & \cdots & x_{i+k,j} \\ & & \vdots & & \\ & & x_{i,j-k} & & \end{array}$$

has (up to prefactors involving Δx, Δy and normalization constants) the Fourier transform

$$(t_x - 1)\frac{1}{t_x^k}\prod_{j=2}^{2k}(t_x - s_j)\hat{u} + (t_y - 1)\frac{1}{t_y^k}\prod_{j=2}^{2k}(t_y - s_j)\hat{v} \tag{3.10}$$

The highest possible order of derivative that is obtainable on this stencil is $2k$. This is for example ∂_x^{2k} and its Fourier transform is given by

$$\frac{(t_x - 1)^{2k}}{t_x^k} \tag{3.11}$$

The aim is now first to construct a discrete counterpart to

$$\partial_x u + \partial_y v = 0 \qquad \Rightarrow \qquad \partial_x^{2k-1}(\partial_x u + \partial_y v) = 0$$

The focus lies on symmetric divergence stencils. In a one-dimensional situation (up to prefactors) such a stencil reduces to a symmetric discretization of $\partial_x u$. Its Fourier transform by Theorem 3.7 is

$$(t_x - 1)\frac{1}{t_x^k}\prod_{j=2}^{2k}(t_x - s_j)\hat{u}$$

with $s_1 = 1$, $s_2 = -1$, and $s_j \neq 1\,\forall j > 2$.

Theorem 3.8 (Divergence)**.** *Assume a symmetric divergence stencil. In order to obtain a symmetric discretization of $\partial_x^{2k-1}(\partial_x u + \partial_y v)$ that is stationarity consistent with this divergence, in multiple spatial dimensions the stencils have to involve the cells*

$$\begin{array}{ccccc}
x_{i-k,j+k} & \cdots & x_{i,j+k} & \cdots & x_{i+k,j+k} \\
\ddots & & \vdots & & \ddots \\
x_{i-k,j} & \cdots & x_{ij} & \cdots & x_{i+k,j} \\
\ddots & & \vdots & & \ddots \\
x_{i-k,j-k} & \cdots & x_{i,j-k} & \cdots & x_{i+k,j-k}
\end{array}$$

and the divergence has to be discretized such that its Fourier transform is

$$\begin{aligned}
&\frac{t_x^2 - 1}{t_x} \cdot \frac{(t_y + 1)^2}{t_y}\frac{1}{t_x^{k-1}}\prod_{j=3}^{2k}(t_x - s_j)\frac{1}{t_y^{k-1}}\prod_{j=3}^{2k}(t_y - s_j)\hat{u} \\
&\quad + \frac{t_y^2 - 1}{t_y} \cdot \frac{(t_x + 1)^2}{t_x}\frac{1}{t_y^{k-1}}\prod_{j=3}^{2k}(t_y - s_j)\frac{1}{t_x^{k-1}}\prod_{j=3}^{2k}(t_x - s_j)\hat{v}
\end{aligned}$$

up to prefactors involving Δx, Δy and normalization constants.

Proof. In order to obtain a stationarity consistent discretization of $\partial_x^{2k-1}(\partial_x u + \partial_y v)$ one needs to obtain (3.11) (or anything that would reduce to it in a one-dimensional situation) in the first term in (3.10). Thus one needs to multiply (3.10) with

$$\frac{(t_x-1)^{2k-1}}{\prod_{j=2}^{2k}(t_x-s_j)}$$

The denominator comes about because the degree of the polynomial is not to be changed during this process. Obviously, it is not possible to divide (3.10) by this term. Indeed, the second part of the expression (3.10), does not depend on t_x at all, and so it is not possible to divide by $\prod_{j=2}^{2k}(t_x-s_j)$ and still obtain a polynomial in t_x, t_y.

Therefore a more adequate approximation to the divergence is taking into account the perpendicular direction and reads

$$(t_x-1)\frac{1}{t_x^k}\prod_{j=2}^{2k}(t_x-s_j)\prod_{j=2}^{2k}(t_y-s_j)\hat{u}+(t_y-1)\frac{1}{t_y^k}\prod_{j=2}^{2k}(t_y-s_j)\prod_{j=2}^{2k}(t_x-s_j)\hat{v}$$

where two factors have been added to ensure symmetry. They now allow for a division. However, the polynomial $\prod_{j=2}^{2k}(t_y-s_j)$ has an odd degree and cannot be symmetric. It thus contains one linear factor less than needed. It is known that it does not contain and that we cannot add any factors (t_y-1). It is also not possible to add factors (t_y-s) with $s\neq\frac{1}{s}$, because they come in pairs. The only option is to add one factor (t_y+1). Actually $s_2=-1$ (i.e. there is already one factor t_y+1 present) which proves the assertion. □

Turn now to other derivatives, i.e. the discrete counterparts to

$$\partial_x u+\partial_y v=0 \qquad \Rightarrow \qquad \partial_x^{n-1}(\partial_x u+\partial_y v)=0$$

with $2\leq n\leq 2k$.

In one dimension, increasing the order of the differential operator in a stationarity consistent manner is easy: One of the linear factors in

$$\frac{1}{t_x^k}\prod_{j=1}^{2k}(t_x-s_j)$$

with $s_j\neq 1$ has to be divided out and replaced by t_x-1. However, the resulting stencil has to be symmetric. This leads to the following rules

Theorem 3.9 (Replacement rules)**.** *Consider notation as in Theorem 3.8. The order of the differential operator that is discretized by the linear stencil*

$$\frac{1}{t_x^k}\prod_{j=1}^{2k}(t_x-s_j) \tag{3.12}$$

can be increased in a stationarity consistent manner by (repeated) multiplication with

i) $\frac{t_x-1}{t_x+1}$ *if the stencil* (3.12) *contains a factor* (t_x+1)

ii) $\frac{(t_x-1)^2}{(t_x-s)(t_x-1/s)}$ *if the stencil* (3.12) *contains* $(t_x-s)(t_x-1/s)$, $s^2 \neq 1$

The first choice increases the order of differentiation by 1, the second by 2.
Additionally, no change in the order happens upon multiplication with

iii) $\frac{(t_x-r)(t_x+r)}{(t_x+1)^2}$ *if the stencil* (3.12) *contains a factor* $(t_x+1)^2$ *and* $r^2 \neq 1$

iv) $\frac{(t_x-r)(t_x-1/r)}{(t_x-s)(t_x-1/s)}$ *if the stencil* (3.12) *contains* $(t_x-s)(t_x-1/s)$, $s^2 \neq 1$, $r^2 \neq 1$

All the stencils that are thus obtained are symmetric.

Proof. The change of order of differentiation is clear by Lemma 3.1, because every time factors of (t_x-1) are (not) added. The symmetry follows from Theorem 3.7 because the stencils have the form given in (3.9). There it has been shown that symmetry implies the appearance of factors t_x-s, $s^2 \neq 1$ only in pairs $(t_x-s)(t_x-1/s)$. □

These rules allow to subsequently construct derivatives $\partial_x^{n-1}(\partial_x u+\partial_y v)$, $2 \leq n \leq 2k$. The results, in general are not unique.

Observe that the Fourier transforms are products of Laurent polynomials in t_x with Laurent polynomials in t_y. In a sense, the multi-dimensional stencils somehow "consist" of two one-dimensional ones. This can be used in order to simplify notation:

Definition 3.8 (⊗-Notation)**.** *Consider a one-dimensional stencil* $\mathcal{A}_x$ *along the x-direction and a one-dimensional stencil* $\mathcal{A}_y$ *along the y-direction. Denote by* $\hat{\mathcal{A}}_x(t_x)$ *and* $\hat{\mathcal{A}}_y(t_y)$ *their corresponding Fourier transforms. The stencil*

$$\mathcal{A}_x \otimes \mathcal{A}_y$$

is defined by its Fourier transform being $\hat{\mathcal{A}}_x \cdot \hat{\mathcal{A}}_y$.

Example 3.3. *The* ⊗*-notation*

$$(u_{i+1}-u_{i-1}) \otimes (u_{j+1}+2u_j+u_{j-1})$$

abbreviates

$$(u_{i+1,j+1}+2u_{i+1,j}+u_{i+1,j-1})-(u_{i-1,j+1}+2u_{i-1,j}+u_{i-1,j-1})$$

whose Fourier transform is

$$\frac{(t_x-1)(t_x+1)}{t_x} \cdot \frac{(t_y+1)^2}{t_y}$$

◁

This definition naturally extends to higher dimensions.

3.2.2.1 Example 1

Consider the first the smallest one-dimensional symmetric discretization of ∂_x with $k = 1$, i.e. on a stencil that includes three cells. By Theorem 3.7 it follows that its Fourier transform is proportional to

$$\frac{1}{t_x^k}(t_x - 1)(t_x + 1)$$

Together with Lemma 3.2 it means that the stencil is unique.

Then, from Theorem 3.7 it follows that it is unique and its Fourier transform is

$$\frac{(t_x - 1)(t_x + 1)}{2\Delta x t_x}$$

The normalization has been chosen using Lemma 3.2.

By Theorem 3.8, the corresponding divergence is

$$\frac{(t_x + 1)(t_x - 1)}{2\Delta x t_x}\frac{(t_y + 1)^2}{4t_y}\hat{u} + \frac{(t_y + 1)(t_y - 1)}{2\Delta y t_y}\frac{(t_x + 1)^2}{4t_x}\hat{v} \tag{3.13}$$

Upon multiplication with

$$2\frac{t_x - 1}{t_x + 1}$$

(replacement rule i) in Theorem 3.9) this becomes a stationarity consistent discretization of $\Delta x \partial_x(\partial_x u + \partial_y v)$:

$$\frac{(t_x - 1)^2}{\Delta x t_x}\frac{(t_y + 1)^2}{4t_y}\hat{u} + \frac{(t_y + 1)(t_y - 1)}{2\Delta y t_y}\frac{(t_x + 1)(t_x - 1)}{2t_x}\hat{v}$$

This is the Fourier transform of

$$\frac{\{\{[[u]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{4\Delta x} + \frac{[[v]_{i\pm1}]_{j\pm1}}{4\Delta y} \tag{3.14}$$

Note that the choice of divergence is unique (because the stencil size does not allow for additional terms other than those dictated by symmetry).

3.2.2.2 Example 2

Consider the stencil $q_{i-2} - 6q_{i-1} + 6q_{i+1} - q_{i+2}$, i.e. $k = 2$. Its Fourier transform is

$$\frac{(t_x - 1)(t_x + 1)(-1 + 6t_x - t_x^2)}{8\Delta x t_x^2}$$

This polynomial has the form discussed in Theorem 3.7; with the notation introduced there one has $s_1 = 1$, $s_2 = -1$. The polynomial $-1 + 6t_x - t_x^2$ has irrational roots and

its factorization thus is of no use. Obviously both the stencil and the polynomial are symmetric. According to Theorem 3.8, the divergence operator is

$$\begin{aligned}\frac{t_x^2-1}{2\Delta x t_x}\cdot\frac{(t_y+1)^2}{4t_y}\frac{(-1+6t_x-t_x^2)}{4t_x}\frac{(-1+6t_y-t_y^2)}{4t_y}\hat{u}\\ +\frac{t_y^2-1}{2\Delta y t_y}\cdot\frac{(t_x+1)^2}{4t_x}\frac{(-1+6t_y-t_y^2)}{4t_y}\frac{(-1+6t_x-t_x^2)}{4t_x}\hat{v}\end{aligned}\tag{3.15}$$

There is one odd derivative $n=3$, i.e. the discretization of $\partial_x^2(\partial_x u+\partial_y v)$. But it is possible to multiply with

$$4\frac{(t_x-1)^2}{-1+6t_x-t_x^2}$$

according to replacement rule ii). Replacement rule i) cannot be used twice here because there is just one factor (t_x+1).

As for the even derivative $\partial_x(\partial_x u+\partial_y v)$ $(n=2)$, the factor can either be

$$2\frac{t_x-1}{t_x+1}$$

(replacement rule i)) or

$$2\frac{t_x-1}{t_x+1}\cdot\frac{(t_x+1)^2}{-1+6t_x-t_x^2}=2\frac{(t_x-1)(t_x+1)}{-1+6t_x-t_x^2}$$

(replacement rules i) and iii)).

Consider the latter choice. Upon multiplication of (3.15) with this factor, the discrete operator has the Fourier transform

$$\begin{aligned}\frac{(t_x-1)^2}{\Delta x t_x}\frac{(t_x+1)^2}{4t_x}\cdot\frac{(t_y+1)^2}{4t_y}\frac{(-1+6t_y-t_y^2)}{4t_y}\hat{u}\\ +\frac{t_y^2-1}{2\Delta y t_y}\cdot\frac{(t_x+1)^2}{4t_x}\frac{(t_x+1)(t_x-1)}{2t_x}\frac{(-1+6t_y-t_y^2)}{4t_y}\hat{v}\end{aligned}$$

This translates, using Definition 3.8 into

$$\begin{aligned}\frac{u_{i+2}-2u_i+u_{i-2}}{4\Delta x}\otimes\frac{-u_{j+2}+4u_{j+1}+10u_j+4u_{j-1}-u_{j-2}}{16}\\ +\frac{u_{i+2}+2u_{i+1}-2u_{i-1}-u_{i-2}}{8}\otimes\frac{-v_{j+2}+6v_{j+1}-6v_{j-1}+v_{j-2}}{8\Delta y}\end{aligned}$$

The fourth derivative is trivial, as it is the maximal one on this stencil, and the factor that makes it appear is just

$$8\frac{(t_x-1)^3}{(t_x+1)(-1+6t_x-t_x^2)}$$

3.2.2.3 Example 3

Analogously, for the stencil[2]

$$q_{i-2} - 8q_{i-1} + 8q_{i+1} - q_{i+2} \simeq \frac{(t_x+1)(t_x-1)(-1+8t_x-t_x^2)}{t_x^2}$$

the divergence operator becomes

$$\hat{u}\frac{(t_x^2-1)}{2\Delta x t_x}\frac{(t_y+1)^2}{4t_y}\frac{-1+8t_x-t_x^2}{6t_x}\frac{-1+8t_y-t_y^2}{6t_y} + \hat{v}\frac{(t_y^2-1)}{2\Delta y t_y}\frac{(t_x+1)^2}{4t_x}\frac{-1+8t_y-t_y^2}{6t_y}\frac{-1+8t_x-t_x^2}{6t_x}$$

The factors can be chosen as:

$$\begin{aligned}
\partial_x^3 &\qquad 12\frac{(t_x-1)^3}{(t_x+1)(-1+8t_x-t_x^2)}\\
\partial_x^2 &\qquad 6\frac{(t_x-1)^2}{-1+8t_x-t_x^2}\\
\partial_x &\qquad 3\frac{(t_x-1)(t_x+1)}{-1+8t_x-t_x^2}
\end{aligned}$$

Note that

$$\begin{aligned}
-q_{i-2} + 4q_{i-1} - 6q_i + 4q_{i+1} - q_{i+2} &\simeq -\frac{(t_x-1)^4}{t_x^2}\\
-q_{i-2} - 2q_{i-1} + 6q_i - 2q_{i+1} - q_{i+2} &\simeq -\frac{(t_x-1)^2(1+4t_x+t_x^2)}{t_x^2}\\
q_{i-2} - 2q_{i-1} + 2q_{i+1} - q_{i+2} &\simeq \frac{(t_x-1)^3(t_x+1)}{t_x^2}
\end{aligned}$$

3.2.3 Pseudo-inverse

Consider a linear system of equations

$$\partial_t q + J_x\partial_x q + J_y\partial_y q = 0 \qquad\qquad q : \mathbb{R}_0^+ \times \mathbb{R}^2 \to \mathbb{R}^n \tag{3.16}$$

As in Section 3.2.1.1, instead of performing the discrete analysis straight away, first the action of the numerics is mimicked by continuous diffusion. A dimensionally split upwind/Roe scheme, up to terms higher than $\mathcal{O}(\Delta x, \Delta y)$ is

$$\partial_t q + J_x\partial_x q + J_y\partial_y q = \frac{1}{2}\Delta x|J_x|\partial_x^2 q + \frac{1}{2}\Delta y|J_y|\partial_y^2 q \tag{3.17}$$

Postpone for the moment the exact definition of $|J_x|$ and sign J_x. Stationarity preservation will be obtained, if $|J_x|\partial_x^2 q = \text{sign } J_x \cdot J_x\partial_x^2 q$ is replaced by sign $J_x \cdot (J_x\partial_x^2 q + J_y\partial_x\partial_y q)$.

[2]The symbol $A \simeq B$ means that the Fourier transform of A is B, possibly up to numerical prefactors.

Indeed, if (3.16) is stationary, then

$$J_x \partial_x q + J_y \partial_y q = 0$$

and thus also

$$J_x \partial_x^2 q + J_y \partial_x \partial_y q = 0$$

In the discrete sense, via stationarity consistency, this would yield a stationarity preserving scheme by Theorem 3.5 (see also Theorem 3.11 below).

Now the precise definitions of the matrix operations $|J_x|$ and sign J_x shall be discussed.

Definition 3.9 (Absolute value)**.** *Given a diagonalization $A = R\Lambda R^{-1}$ of a real-diagonalizable $n \times n$ matrix A with $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$, the* absolute value $|A|$ *is defined on the eigenvalues $\lambda_1, \ldots, \lambda_n$ as follows*

$$|A| := R \cdot diag(|\lambda_1|, \ldots, |\lambda_n|) \cdot R^{-1}$$

If additionally the inverse A^{-1} exists, then a natural definition of sign A would be

$$\operatorname{sign} A := |A| A^{-1} \qquad \text{if } A \text{ invertible}$$

However, for certain equations, in particular the equations of linear acoustics (2.40)–(2.41) the matrices J_x, J_y are not invertible. This makes some kind of regularization necessary:

Definition 3.10 (Moore-Penrose pseudo-inverse)**.** *i) The* pseudo-inverse $c^{\textcircled{-1}}$ *of a number $c \in \mathbb{R}$ is defined as*

$$c^{\textcircled{-1}} := \begin{cases} \frac{1}{c} & c \neq 0 \\ 0 & c = 0 \end{cases}$$

ii) Given a diagonalizable $n \times n$ matrix $J = R\Lambda R^{-1}$ with real eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$, its Moore-Penrose pseudo-inverse *is defined as*

$$J^{\textcircled{-1}} := R \cdot diag(\lambda_1^{\textcircled{-1}}, \ldots, \lambda_n^{\textcircled{-1}}) \cdot R^{-1}$$

This definition can be extended to non-diagonalizable matrices, but the way the definition is given suits perfectly the purposes of this Section. So far it is unclear what impact the particular choice of regularization will have on the resulting numerical scheme. The Moore-Penrose pseudo-inverse will be used as a hypothesis, and the properties of the scheme obtained have be studied in detail in the end.

Definition 3.11 (Sign)**.** *Given a real-diagonalizable $n \times n$ matrix A, the matrix* sign A *is defined as follows:*

$$\operatorname{sign} A := |A| A^{\textcircled{-1}}$$

Modifying the diffusion of (3.17) as described leads to the following expression:

$$\partial_t q + J_x \partial_x q + J_y \partial_y q = \frac{1}{2}\Delta x \operatorname{sign} J_x (J_x \partial_x^2 q + J_y \partial_x \partial_y q) + \frac{1}{2}\Delta y \operatorname{sign} J_y (J_x \partial_x \partial_y q + J_y \partial_y^2 q)$$

The only missing ingredient is the discrete counterpart to the statement

$$J_x \partial_x q + J_y \partial_y q = 0 \qquad \Rightarrow \qquad J_x \partial_x^2 q + J_y \partial_x \partial_y q = 0$$

Theorem 3.10. *For constant $n \times n$ matrices J_x, J_y and q an n-dimensional vector defined on a rectangular grid, the following discrete statement holds:*

$$\frac{J_x}{8\Delta x}\{\{[q]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + \frac{J_y}{8\Delta y}[\{\{q\}\}_{i\pm\frac{1}{2}}]_{j\pm1} = 0 \quad \Rightarrow \quad \frac{J_x}{8\Delta x}\{\{[[q]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{J_y}{8\Delta y}[[q]_{i\pm1}]_{j\pm1} = 0$$

(The notation is introduced in Definition 0.2.)

Proof. The proof repeats the proof of Theorem 3.8. Applying the Fourier transformation to

$$\frac{J_x}{8\Delta x}\{\{[q]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + \frac{J_y}{8\Delta y}[\{\{q\}\}_{i\pm\frac{1}{2}}]_{j\pm1} \tag{3.18}$$

gives

$$\frac{J_x\,\hat{q}}{\Delta x}\frac{(t_x+1)(t_x-1)}{2t_x}\frac{(t_y+1)^2}{4t_y} + \frac{J_y\,\hat{q}}{\Delta y}\frac{(t_x+1)^2}{4t_x}\frac{(t_y+1)(t_y-1)}{2t_y}$$

Upon multiplication with $\frac{t_x-1}{t_x+1}$ this becomes

$$\frac{J_x\,\hat{q}}{\Delta x}\frac{(t_x-1)^2}{2t_x}\frac{(t_y+1)^2}{4t_y} + \frac{J_y\,\hat{q}}{\Delta y}\frac{(t_x+1)(t_x-1)}{4t_x}\frac{(t_y+1)(t_y-1)}{2t_y}$$

which is the Fourier transformation of

$$\frac{J_x}{8\Delta x}\{\{[[q]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{J_y}{8\Delta y}[[q]_{i\pm1}]_{j\pm1} \tag{3.19}$$

As the two expressions are related by a simple factor, whenever (3.18) vanishes, (3.19) vanishes. □

This allows to formulate the following statement:

Theorem 3.11. *Consider the linear hyperbolic system* (3.16) *in two spatial dimensions. The semidiscrete numerical scheme*

$$\begin{aligned}\partial_t q &+ \frac{J_x}{8\Delta x}\{\{[q]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + \frac{J_y}{8\Delta y}[\{\{q\}\}_{i\pm\frac{1}{2}}]_{j\pm1} \\ &- \operatorname{sign} J_x \left(\frac{J_x}{8\Delta x}\{\{[[q]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{J_y}{8\Delta y}[[q]_{i\pm1}]_{j\pm1}\right) \\ &- \operatorname{sign} J_y \left(\frac{J_x}{8\Delta x}[[q]_{i\pm1}]_{j\pm1} + \frac{J_y}{8\Delta y}[[\{\{q\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}}\right) = 0\end{aligned} \tag{3.20}$$

is stationarity preserving and reduces to the upwind/Roe scheme in one spatial dimension.

Proof. The dimensionally split upwind/Roe scheme for (3.16) reads

$$\partial_t q + \frac{J_x}{2\Delta x}[q]_{i\pm1,j} + \frac{J_y}{2\Delta y}[q]_{i,j\pm1} - \frac{|J_x|}{2\Delta x}[[q]]_{i\pm\frac{1}{2},j} - \frac{|J_y|}{2\Delta y}[[q]]_{j\pm\frac{1}{2}} = 0$$

Replacing

$$\frac{J_x}{2\Delta x}[q]_{i\pm1,j} + \frac{J_y}{2\Delta y}[q]_{i,j\pm1}$$

by (3.18), as well as

$$\frac{|J_x|}{2\Delta x}[[q]]_{i\pm\frac{1}{2},j}$$

by (3.19) (and analogously for the other direction) proves the assertion. The scheme is stationarity preserving by Proposition 3.10, with the discrete stationary states given by

$$\frac{J_x}{8\Delta x}\{\{[q]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + \frac{J_y}{8\Delta y}[\{\{q\}\}_{i\pm\frac{1}{2}}]_{j\pm1} = 0$$

□

Observe that this strategy in principle can be applied to all linear hyperbolic systems of PDEs, although it is unclear whether the scheme thus obtained will always be stable. Additionally, the precise influence of the choice of regularization procedure is unknown. A successful application of this procedure is presented in Section 4.5.2.2. One can try to derive schemes for nonlinear equations along the same lines; this is discussed in Section 5.4.2.

3.2.4 Taylor series and rotationally invariant operators

Consider a discrete divergence $\mathscr{D}$ (e.g. Equation (3.13)) and a stationarity consistent discrete second derivative $\mathscr{S}$ (e.g. Equation (3.14)). Taking up the example of Section 3.2.2.1, to highest order $\mathscr{S} = \partial_x(\mathscr{D}) + \mathcal{O}(\Delta x, \Delta y) = \partial_x(\partial_x u + \partial_y v) + \mathcal{O}(\Delta x, \Delta y)$. How do the higher order terms look like? For example, do they contain higher derivatives of $\partial_x u + \partial_y v$?

Theorem 3.12. *Consider $\mathscr{S}$, a discrete second derivative of a function q, that is stationarity consistent with $\mathscr{D}$, a discrete first derivative of q. Denote their Fourier transforms by $\hat{\mathscr{S}}$, $\hat{\mathscr{D}}$ and the factor that relates them by f:*

$$\hat{\mathscr{S}} = f\hat{\mathscr{D}} \tag{3.21}$$

Assume that f is a function of $t_x = \exp(\mathbb{i}k_x\Delta x)$ only and that it can be expanded as Taylor series in $\mathbb{i}k_x\Delta x$. Denote by a_n the corresponding coefficients:

$$f = \sum_{n=0}^{\infty} a_n \mathbb{i}^n k_x^n \Delta x^n \tag{3.22}$$

Then $\mathscr{S}$ can be written as a function of $\mathscr{D}$ and its derivatives as

$$\mathscr{S} = \sum_{n=0}^{\infty} a_n \Delta x^n \partial_x^n \mathscr{D}$$

Note: An example (it is studied in detail after the proof) is $q = (u, v)$ and

$$\mathscr{D} = \frac{1}{8\Delta x}\{\{[u]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[\{\{v\}\}_{i\pm\frac{1}{2}}]_{j\pm1}$$
$$\mathscr{S} = \frac{1}{8\Delta x}\{\{[[u]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[[v]_{i\pm1}]_{j\pm1}$$

Proof. Make use of the Fourier transform and expand, e.g. in two spatial dimensions:

$$\mathscr{S} = \sum_{\mathbf{k}} \hat{\mathscr{S}} \exp(\mathbb{i}k_x x + \mathbb{i}k_y y) \overset{(3.21)}{=} \sum_{\mathbf{k}} f\hat{\mathscr{D}} \exp(\mathbb{i}k_x x + \mathbb{i}k_y y)$$
$$\overset{(3.22)}{=} \sum_{n=0}^{\infty} a_n \Delta x^n \sum_{\mathbf{k}} (\mathbb{i}k_x)^n \hat{\mathscr{D}} \exp(\mathbb{i}k_x x + \mathbb{i}k_y y) = \sum_{n=0}^{\infty} a_n \Delta x^n \partial_x^n \sum_{\mathbf{k}} \hat{\mathscr{D}} \exp(\mathbb{i}k_x x + \mathbb{i}k_y y)$$

$$= \sum_{n=0}^{\infty} a_n \Delta x^n \partial_x^n \mathscr{D}$$

□

Consider the example of Section 3.2.2.1 and take $\Delta y = \Delta x$ to ease the notation:

$$\begin{aligned}
&\frac{1}{4}\left(\frac{\{\{[[u]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{\Delta x} + \frac{[[v]_{i\pm1}]_{j\pm1}}{\Delta y}\right) = (\partial_x^2 u + \partial_x\partial_y v)\Delta x \\
&+ \frac{(3\partial_x^2\partial_y^2 u + \partial_x^4 u + 2(\partial_x\partial_y^3 v + \partial_x^3\partial_y v))\Delta x^3}{12} \\
&+ \frac{(15\partial_x^2\partial_y^4 u + 15\partial_x^4\partial_y^2 u + 2\partial_x^6 u + 6\partial_x\partial_y^5 v + 20\partial_x^3\partial_y^3 v + 6\partial_x^5\partial_y v)\Delta x^5}{720} \\
&+ \frac{(14\partial_x^2\partial_y^6 u + 35\partial_x^4\partial_y^4 u + 14\partial_x^6\partial_y^2 u + \partial_x^8 u + 4(\partial_x\partial_y^7 v + 7(\partial_x^3\partial_y^5 v + \partial_x^5\partial_y^3 v) + \partial_x^7\partial_y v))\Delta x^7}{20160} \\
&+ \mathcal{O}(\Delta x^9)
\end{aligned}$$

Obviously, the highest order term is just of the form $\partial_x(\partial_x u + \partial_y v)$. It would however be erroneous to expect the higher order terms to be higher derivatives of $\partial_x u + \partial_y v$, because it is not $\partial_x u + \partial_y v$ that is preserved exactly by the numerics – it is rather the discrete divergence

$$\begin{aligned}
\mathscr{D} = &\frac{\{\{[u]_{i\pm1}\}\}_{j\pm\frac{1}{2}}}{8\Delta x} + \frac{[\{\{v\}\}_{i\pm\frac{1}{2}}]_{j\pm1}}{8\Delta y} = (\partial_x u + \partial_y v) \\
&+ \frac{1}{12}\Big(3\partial_x\partial_y^2 u + 2(\partial_x^3 u + \partial_y^3 v) + 3\partial_x^2\partial_y v\Big)\Delta x^2 \\
&+ \frac{1}{240}\Big(5\partial_x\partial_y^4 u + 2(5\partial_x^3\partial_y^2 u + \partial_x^5 u + \partial_y^5 v + 5\partial_x^2\partial_y^3 v) + 5\partial_x^4\partial_y v\Big)\Delta x^4 \\
&+ \frac{1}{10080}\Big(7\partial_x\partial_y^6 u + 35\partial_x^3\partial_y^4 u + 21\partial_x^5\partial_y^2 u + 2\partial_x^7 u + 2\partial_y^7 v + 7(3\partial_x^2\partial_y^5 v + 5\partial_x^4\partial_y^3 v + \partial_x^6\partial_y v)\Big)\Delta x^6 \\
&+ \mathcal{O}(\Delta x^8)
\end{aligned}$$

Indeed, one then finds that

$$\frac{1}{4}\left(\frac{\{\{[[u]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{\Delta x}+\frac{[[v]_{i\pm1}]_{j\pm1}}{\Delta y}\right)$$
$$=\Delta x\,\partial_x\mathscr{D}-\frac{\Delta x^3}{12}\partial_x^3\mathscr{D}+\frac{\Delta x^5}{120}\partial_x^5\mathscr{D}-\frac{17\Delta x^7}{20160}\partial_x^7\mathscr{D}+\mathcal{O}(\Delta x^9)$$

which is, by Theorem 3.12 the formal series of

$$2\frac{\exp(\Delta x\partial_x)-1}{\exp(\Delta x\partial_x)+1}\mathscr{D}$$

because f has been found to be $2\frac{t_x-1}{t_x+1}$ in Section 3.2.2.1.

The stationarity-consistency thus manifests itself in the fact, that every order in the series of $\mathcal{S}$ is a derivative of $\mathscr{D}$. Whenever $\mathscr{D}$ vanishes, the discrete second derivative vanishes as well. This calculation shows that the stencils being "rotationally-invariant", as suggested in [Sid02], is actually not a relevant condition. They might seem so to first order in their expansion in powers of Δx, but they cannot remain so when higher order terms are taken into account.

Chapter 4

Numerical schemes for linear acoustics

Contents

The introduction to Section 2 emphasized the central role of the advection equation for understanding the concept of upwinding. It has been argued that the acoustic equations play a similarly important role when it comes to understanding the behaviour of schemes in the limit of low Mach numbers. This has several reasons.

First of all, the acoustic equations (2.7)–(2.8) are sufficiently complicated, such that from learning how to deal with the new features one may hope to learn lessons for similar features of the Euler equations. Indeed, despite being linear, the acoustic equations contain the differential operators grad and div, which in multiple spatial dimensions are different. This leads to the fact that the x- and y-Jacobians of the acoustic equations do not commute and thus are not simultaneously diagonalizable. For this reason, linear acoustics cannot be reduced to some kind of multi-dimensional advection. This also becomes obvious in light of the exact solution obtained in Section 2.2 which involves

a Mach cone. Therefore the acoustic system in multiple spatial dimensions contains features not present in advective problems (see also the review [Roe17]).

Secondly, the acoustic system is sufficiently simple to be tractable both analytically and numerically. The exact solution, although lengthy, can be computed to full extent (see Section 2.2) and linear numerical schemes for the acoustic equations can be studied with the methods of Section 3. Indeed, the acoustic equations fit into the framework of Theorem 3.1 of Section 3.1.1:

Corollary 4.1. *The acoustic equations* (2.7)–(2.8) *allow for non-trivial stationary states given by*

$$\nabla p = 0 \qquad \textit{and} \qquad \nabla \cdot \mathbf{v} = 0$$

Proof. The proof follows from Theorem 3.1. The eigenvector corresponding to the zero-eigenvalue of

$$\mathbf{J} \cdot \mathbf{k} = \begin{pmatrix} 0 & 0 & \frac{c}{\epsilon} k_x \\ 0 & 0 & \frac{c}{\epsilon} k_y \\ \frac{c}{\epsilon} k_x & \frac{c}{\epsilon} k_y & 0 \end{pmatrix}$$

is

$$\hat{q}_0 = (-k_y, k_x, 0)^{\mathrm{T}}$$

$(\hat{u}, \hat{v}, \hat{p})^{\mathrm{T}}$ is only a multiple of $\hat{q}_0$ if $\hat{p} = 0$ and $k_x \hat{u} + \hat{v} k_y = 0$. These are the Fourier transforms of $\nabla p = 0$, $\operatorname{div} \mathbf{v} = 0$. □

Trivial stationary states of (2.7)–(2.8) are shear flows $\nabla p = 0$, $\partial_x u = 0 = \partial_y v$.
From Theorem 3.2 follows the

Corollary 4.2. *The acoustic equations* (2.7)–(2.8) *have a constant of motion* $\omega = \nabla \times \mathbf{v}$, *i.e.*

$$\partial_t \omega = 0 \tag{4.1}$$

Its Fourier transform is $-k_y \hat{u} + k_x \hat{v}$.

Proof. The left eigenvector corresponding to the zero eigenvalue of $\mathbf{J} \cdot \mathbf{k}$ is $(-k_y, k_x, 0)$ such that

$$(-k_y, k_x, 0) \partial_t \begin{pmatrix} \hat{u} \\ \hat{v} \\ \hat{p} \end{pmatrix} = 0$$

This means that $k_x \hat{v} - k_y \hat{u}$ is stationary, which is the Fourier transform of $\omega = \partial_x v - \partial_y u$, and thus $\partial_t \omega = 0$. □

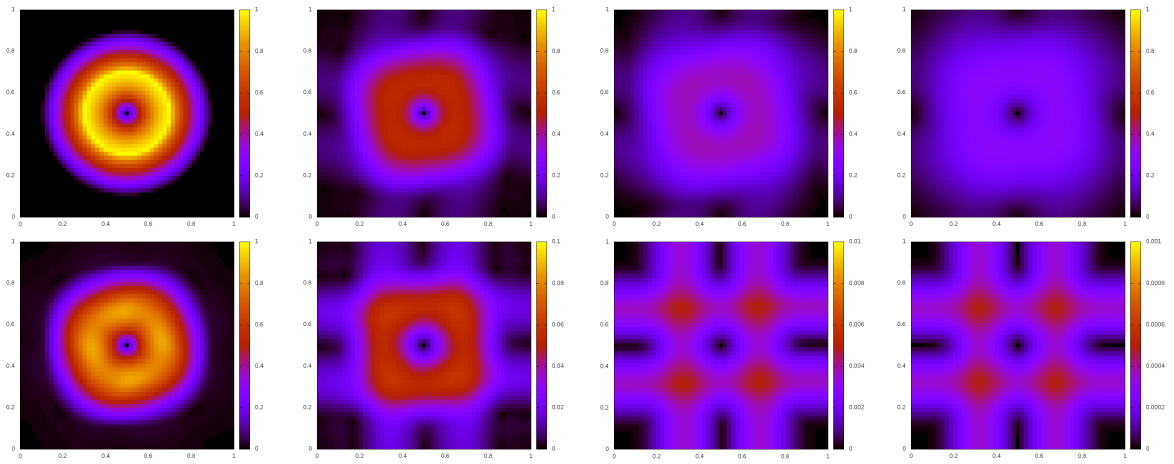

Figure 4.1: Simulation results for a vortex setup for $t = 0, 1, 2, 3$ (from left to right). Colour coded is $\sqrt{u^2 + v^2}$ *Top row*: Euler equations. *Bottom row*: Acoustic equations.

ω is the vorticity introduced in Section 2.1.2. Equation (4.1) also immediately follows from the application of the curl operator to Equation (2.5). However the language of the Fourier transform is more useful when the discrete situation is considered.

A hint towards a possible understanding of the low Mach number problem with just the acoustic system is the appearance of visually similar artefacts between the acoustic and Euler equations, when the low Mach limit is approached with, say, the upwind/Roe scheme (see Figure 4.1).

To present an explanation of these artefacts for the acoustics equations, the reasons why they appear and different ways how to avoid them is the aim of this Section. Ideas already present in the literature are found to fit well into the new framework and are given a new interpretation. Some of them seem to find better foundation in the framework of stationarity preservation. Additionally there appears a connection to vorticity preserving schemes.

4.1 Low Mach number limit

It is an experimental fact that there are schemes whose numerical results deteriorate as the Mach number decreases, and those whose numerical error does not increase in the limit. The aim of this Section is to present a consistent picture of why this happens, at least for the acoustic equations. The concept of stationarity preservation is shown to be a fruitful and consistent way to understand the behaviour of schemes for low Mach numbers. It unveils a connection between vorticity preserving schemes and schemes that are able to resolve the low Mach number limit. This work has been published in [Bar17a, Bar17b].

4.1.1 Connection to stationarity preservation

Recall the acoustic equations (2.40)–(2.41):

$$\partial_t \mathbf{v} + \frac{\nabla p}{\epsilon^2} = 0$$
$$\partial_t p + c^2 \nabla \cdot \mathbf{v} = 0$$

Recall also that the limit $\epsilon \to 0$ of this family of equations has been given the name of the low Mach number limit by analogy with the Euler equations, but that for the acoustic equations it can also be rewritten as the limit of long time (see Section 2.3).

The existence of a stationary vorticity for the acoustic equations has been given particular attention in [TF04], [MT11], [MR01] and others. The class of schemes that possess a stationary discrete counterpart has been given a name:

Definition 4.1 (Vorticity preserving)**.** *A consistent scheme for* (2.7)–(2.8) *is called* vorticity preserving*, if there exists a discretization of the vorticity that remains unchanged during the time evolution.*

Recall the definition of the evolution matrix for linear schemes (Definition 3.5). By Theorem 3.4 and Corollary 4.2, having a vanishing eigenvalue of the evolution matrix yields the numerical stationary states as the right eigenvector and a numerical vorticity operator as the left eigenvector (as Fourier transforms). Conversely, if no eigenvalue of $\mathcal{E}$ vanishes, there is no discrete analogue of the vorticity that would remain stationary:

Corollary 4.3. *For the acoustic equations* (2.7)–(2.8)*, a scheme is vorticity preserving iff it is stationarity preserving.*

Proof. As Theorem 3.4 shows, the existence of nontrivial stationary states gives rise to a constant of motion for every eigenvalue zero. □

It is known (see [DOR10] among others) that the Roe scheme displays artefacts in the limit $\epsilon \to 0$ when applied to the equations (2.40)–(2.41). With the concept of stationarity preservation these artefacts can be given a completely new interpretation.

Theorem 2.9 states that the limit $\epsilon \to 0$ of Equations (2.40)–(2.41) can be understood as the long time limit $t \mapsto \frac{t}{\epsilon}$, $\epsilon \to 0$ of

$$\partial_t \mathbf{v} + \nabla p = 0$$
$$\partial_t p + c^2 \nabla \cdot \mathbf{v} = 0$$

The reason for that is the appearance of $\frac{c}{\epsilon}$ and t only inside the combination $\frac{ct}{\epsilon}$. This is also true at discrete level:

Definition 4.2 (Low Mach compliant)**.** *A numerical scheme for* (2.40)–(2.41) *is called* low Mach compliant *if in the limit $\epsilon \to 0$ it has solutions that discretize all the analytic solutions given by the limit equations*

$$\nabla p = 0 \qquad\qquad \nabla \cdot \mathbf{v} = 0$$

of (2.40)–(2.41)*.*

Theorem 4.1. *Consider consistent and von Neumann stable linear numerical schemes for* (2.40)–(2.41)*. Additionally, the eigenvalues of their evolution matrices* $\mathcal{E}$ *shall be linear in* c*. Vorticity preserving schemes, that fulfill these conditions, are low Mach compliant.*

Proof. The proof consists of two parts:

i) An analogous result to that of Theorem 2.9 is shown. The time evolution of the discrete Fourier modes, according to Equation (3.5) is given by the eigenvalues of the evolution matrix $\mathcal{E}$. By assumption, they are linear in c. In general, the eigenvalue will have the form $c|\mathbf{k}|f(\mathbf{k}, \Delta x)$, with f an arbitrary function (which does not depend on c).

One now needs to show that the eigenvalues actually are linear in $\frac{c}{\epsilon}$. There exist asymptotic scalings of the dependent and independent variables that lead from the system (2.5)–(2.6) (which does not contain ϵ) to its rescaled version (2.40)–(2.41). They are calculated explicitly in Section 2.3.

$\mathcal{E}$ having an eigenvalue $c|\mathbf{k}| \cdot f$ implies that there is a quantity $\hat{q}$ that satisfies an equation of the form

$$\partial_t \hat{q} + c|\mathbf{k}| f \hat{q} = 0 \tag{4.2}$$

Rescaling according to (2.42)–(2.43), and using (2.44)–(2.45) forces the rescaled form of Equation (4.2) to become

$$\partial_t \hat{q} + \frac{c}{\epsilon}|\mathbf{k}| f \hat{q} = 0$$

Therefore, the eigenvalues of $\mathcal{E}$ are linear in $\frac{c}{\epsilon}$. Therefore in the time evolution of any non-stationary Fourier mode only the combination $\frac{ct}{\epsilon}$ appears. This is in full analogy to the continuous case (Theorem 2.9).

ii) In order to study the limit of low Mach numbers one looks at the long time evolution of the numerical scheme for the non-rescaled equations

$$\begin{aligned} \partial_t \mathbf{v} + \nabla p &= 0 \\ \partial_t \mathbf{v} + c^2 \nabla \cdot \mathbf{v} &= 0 \end{aligned}$$

The statement of von Neumann stability is that every Fourier mode is either stationary, or decaying. Thus, as the numerical scheme is stable by assumption, then after long times only stationary Fourier modes will have survived. After long times, or equivalently for low Mach numbers, the numerical solution will be approaching one of the numerical stationary states of the scheme. As the scheme is assumed to be vorticity preserving, or equivalently stationarity preserving (by Corollary 4.3), its stationary states are a discretization of the analytic stationary states (by Theorem 3.3), which by Theorem 2.9 are the limit equations for low Mach number.

□

If the scheme is not stationarity preserving, it will fail to have the right limit as $\epsilon \to 0$, as its limit equations do not discretize all of the limit equations of the underlying PDE. This is discussed for the upwind/Roe scheme in Section 4.4.2.

In fact, by considering the physical dimension that an eigenvalue of the evolution matrix $\mathcal{E}$ must have, one observes that linearity in c is the only way for it to obtain the correct units. Therefore this assumption is actually always true.

This result is reminiscent of the concept of *asymptotic preserving* (see e.g. [Jin99]). A comparative discussion of the relation between stationarity preservation and asymptotic preserving schemes for the low Mach number limit is found in Section 4.6.

4.1.2 Construction principles for low Mach number schemes

Not all numerical schemes for the acoustic equations are stationarity preserving. In particular, the upwind/Roe scheme is not (this is proven in Section 4.4.1). Interestingly, the upwind/Roe scheme is the Godunov scheme for linear acoustics in one spatial dimension: It is the scheme that one obtains upon constant reconstruction in every cell, exact evolution, and subsequent averaging over cells. It feels very disturbing that such a fundamental scheme may be failing so badly to capture important aspects of the exact solution.

On the other hand in the literature there exists a number of schemes that are low Mach compliant. Many of them are obtained by modifying the upwind/Roe scheme (*low Mach fixes*). They typically involve free parameters or even functions. The low Mach fixes are mostly formulated for schemes for the Euler equations (and as such are discussed in Section 5), but they often can be reinterpreted for the acoustic equations. In a number of such fixes, e.g. in [TD08, Rie11, LG13, OSB+16], the ϵ-scaling of one entry of a diffusion matrix is identified as causing problems in the low Mach number regime, and it is suggested to multiply it with some function that modifies this scaling. The precise shape of this function is free. Such schemes indisputably manage to produce good results in the limit of low Mach numbers and can be shown to be stationarity preserving. The amount of arbitrariness in the choice of *low Mach fixes* might be unsatisfying, though.

Here is the dilemma: A scheme derived from seemingly fundamental principles does not capture the low Mach number limit, and those schemes that do, seem to do so only using some *fix*. The Godunov scheme contains the three steps *reconstruction*, *evolution*, *averaging*. The result of these three steps in one spatial dimension is the upwind/Roe scheme that is not suitable for low Mach numbers. Which one of these steps is in conflict with the low Mach number limit then?

There is a simple answer, that unfortunately turns out to be wrong. The upwind/Roe scheme is indeed the Godunov scheme for linear acoustics, but only for one spatial dimension. The limit equations for $\epsilon \to 0$ in 1-d are just

$$\partial_x v = 0 \qquad\qquad \partial_x p = 0$$

This is just a constant flow with constant pressure, and any such flow is well resolved by the upwind/Roe scheme. There is no *low Mach number problem* in 1 spatial dimension.

It is only in *multiple sptial dimensions* that $\partial_x v = 0$ becomes the much less trivial condition

$$\nabla \cdot \mathbf{v} = 0$$

In short: the low Mach number limit is a multi-dimensional phenomenon (as has been also emphasized in [Del10]). The upwind/Roe scheme is the Godunov scheme only in one spatial dimension. The obvious conjecture thus is that maybe it would just suffice to derive the Godunov scheme in multiple spatial dimensions, and that it would then automatically have the right low Mach number limit.

Unfortunately, this is not true. The 2-dimensional Godunov scheme for linear acoustics is derived in Section 4.2, and it turns out not to be stationarity preserving. In [LMMW00, FG17] other schemes are derived that are inspired by the multi-dimensional exact evolution operator for linear acoustics. None of them is stationarity preserving. This is an astonishing result: it is not sufficient to use the exact evolution operator in order to obtain a scheme with good behaviour in the low Mach number limit! Which part of the *reconstruction-evolution-averaging* procedure has to be modified in order to improve the scheme, and how, is to the author's knowledge an open question, although problems with the Godunov scheme have been noticed as early as in [GM04].

Thus so far one has to try to use the condition of stationarity preservation somehow constructively. One way would be to take a large set of schemes with a number of free parameters (as is done in [LR14] for a class of Lax-Wendroff schemes) and try to adjust the parameters such that the condition of stationarity preservation is met. This is performed in Section 4.4.1 for dimensionally split schemes. The results allow one to rediscover many low Mach number schemes and fixes present in the literature. This also shows that stationarity preservation indeed is a powerful tool to study the low Mach number limit. However, contrary to the Godunov scheme, such schemes are not necessarily stable (under explicit time discretization). A stability analysis for a broad class of dimensionally split schemes for the acoustic equations is presented in Section 4.3.

There are further ways to construct stationarity preserving schemes, which have already been discussed in Chapter 3 for general linear systems. The application to the acoustic equations is discussed in Section 4.5. They lead to multi-dimensional schemes, i.e. schemes that cannot be obtained by a dimensionally split approach. These construction principles in practice seem to lead directly to stable schemes. It is also possible to construct stationarity preserving schemes of higher order (Section 4.7).

4.2 The multidimensional Godunov scheme

4.2.1 Historical overview

The knowledge of the exact solution makes it possible to derive a Godunov scheme. This is similar in spirit to an idea by Gelfand mentioned in [GZI+76, God97] (a translated version is [God08]), but the absence of a(n accessible) published work or details on the

procedure makes it hard to compare it to the approach taken here. From the scarce account it seems that an exact solution as presented here was not used and that the present approach is more general, in the sense that it might also be used to derive other solvers.

Interestingly, the aim of the computation performed back then was completely opposite. After Godunov's method has been successfully used for one-dimensional problems, an extension to multiple dimensions was needed. The complexity of a multi-dimensional Riemann solver for the Euler equations was obvious. Citing from [God08] ([GZI+76], p. 56):

> *When we implemented the 2-D approach based on the solutions of the Riemann problem with arbitrary initial conditions, the first question concerned the construction of such solutions. In the 2-D case the rectangular grid cells can neighbor not only on each other but also on the nodes where four cells meet. If one constructs a 2-D scheme analogously to the 1-D, one should have analytic solutions of hydrodynamical equations with four discontinuities of initial data at one point. We did not have such solutions, even now they do not exist, at least for general initial data. We had the audacity to suggest to use only classical solutions of the Riemann problem combined from the plane waves and describing initial Riemann problem placed on the edges of the neighboring cells. We ignored interaction of four cells having a common node. Implementing this approach we abandoned a clear physical interpretation on which the construction of the 1-D scheme was based. Obviosly, there were a lot of discussions about the suggestion mentioned above.*

The multi-dimensional calculation was done in 1956:

> *At the same time, for the acoustic waves propagating in a medium at rest, K. V. Brushlinskii, based on Gelfands suggestion, constructed the solution of the problem using an interaction of all cells adjoint to one node with the help of the Sobolevs method of functionally invariant solutions. This solution was used in a numerical scheme completely analogous to the 1-D one. [...] To our surprise and satisfaction, we did not discover any essential differences. After that, only the rough model was employed.*

Therefore Godunov and his team were content finding that the development of multi-dimensional schemes was not worth doing. This is probably due to them having high Mach phenomena in mind. Indeed, recall that the existence of schemes that do not deteriorate in the limit of low Mach numbers was not noticed until works like [Tur87, KLN91, WS95, Kle95]. The remedy until then was to make the grid finer, and generally the acceptance of diffusivity of schemes must have been much higher in view of the little computing power available in the late 80s, not to mention the 50s. Today it is obvious that the class of multi-dimensional schemes is not restricted to ones obtained by solving multi-dimensional Riemann problems, and that this class does indeed contain very useful schemes (e.g. stationarity preserving ones, as discussed in Section 4.5 below).

The exact evolution operator for linear acoustics (2.8)–(2.7) already appears in [ER13, Roe17], albeit without the justification as distributional solution. It has been taken as inspiration for a new kind of numerical schemes in [ER13]: the *active flux* method, which contains additional, pointwise degrees of freedom that are evolved in time exactly. A finite volume scheme of the usual kind, as derived in [FG17], takes only an approximation of the exact solution into account. Among others, [LMMW00] considered the bicharacteristics relation ([CH62]) in order to derive schemes which incorporate multidimensional information. However, the bicharacteristics do not allow to write down the time evolution of initial data directly and thus again, the scheme only uses an approximation of the exact relation.

The conceptually simplest finite volume is a Godunov scheme with the Riemann Problem as a building block. [AG15, LS02] studied the solutions to multi-dimensional Riemann Problems for linear acoustics using a self-similarity ansatz. However, no numerical scheme has been derived there. The purpose of the derivation presented here is twofold. First, it is derived in order to verify that it lacks stationarity preservation and that multi-dimensionality alone does not solve the low Mach number problem. Second, the derivation demonstrates the usage of the exact evolution operator. One might in future imagine numerical schemes that do not rely on Riemann Problems, but still use the exact evolution operator on a different kind of reconstruction.

This Section is largely based on work published in [BK17].

4.2.2 Procedure

In this Section the aim is to derive a two-dimensional finite volume scheme, which updates the numerical solution q_{ij}^n in a Cartesian cell $\mathcal{C}_{ij} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ at a time t^n to a new solution q_{ij}^{n+1} at time $t^{n+1} = t^n + \Delta t$ using q_{ij}^n and information from the neighbours of $\mathcal{C}_{ij}$. The grid is taken equidistant, and notation of Definition 0.1 is used.

The knowledge of the exact solution makes it possible to derive a Godunov scheme via the procedure of *reconstruction-evolution-averaging* ([Tor09, LeV02]). As it is formulated, the derivation encounters technical difficulties, because the solution is needed at all points inside the cell. This means an evaluation of the integrals (2.31)–(2.34), or (2.24)–(2.25), for basically all $\mathbf{x}$ with no direct simplification. It afterwards undergoes the possibly nontrivial task of being integrated over the cell.

Employing the structure of the conservation law allows to rewrite the volume integral into a time integral over the boundary, which is a huge simplification. Now the exact solution is only needed along the boundary of the cell, and one of the components of $\mathbf{x}$ is zero. Still however one needs to evaluate the solution formulae at a continuous set of $\mathbf{x}$ values.

In the following it is shown that for linear systems a Godunov scheme can be written down using just *one* evaluation of the solution formula at a single point in space by suitably modifying the initial data.

Consider the general linear $n \times n$ hyperbolic system (2.13) in d spatial dimensions

$$\partial_t q + (\mathbf{J} \cdot \nabla) q = 0$$

with initial data

$$q(0, \mathbf{x}) = q_0(\mathbf{x})$$

Recall the Definition 2.8 of the time evolution operator T_t: $(T_t\, q_0)(t, \mathbf{x})$ satisfies (2.13) with (at $t = 0$) initial data $q_0(\mathbf{x})$.

Definition 4.3 (Sliding average)**.** *Define the sliding average operator A in two spatial dimensions by its action onto a function $q : \mathbb{R}^d \to \mathbb{R}^n$ as*

$$(Aq)(\mathbf{x}) := \frac{1}{\Delta x \Delta y} \int\limits_{\left[-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right] \times \left[-\frac{\Delta y}{2}, \frac{\Delta y}{2}\right]} \mathrm{d}\mathbf{s} \quad q(\mathbf{x} + \mathbf{s})$$

The objective is to construct a Godunov scheme by introducing a reconstruction $q_0(\mathbf{x})$ using the discrete values $\{q_{ij}^n\}$ in the cells and computing its exact time evolution. The reconstruction needs to be conservative, i.e. $(Aq_0)(\mathbf{x}_{ij}) = q_{ij}$. The easiest choice is a piecewise constant reconstruction

$$q_0(\mathbf{x}) := q_{ij} \qquad \text{if} \qquad \mathbf{x} \in \mathcal{C}_{ij}$$

It is shown in Fig. 4.2 (left) and obviously is locally integrable.

The Godunov procedure *reconstruction-evolution-averaging* can be written as

$$q_{ij}^{n+1} = (A\, T_{\Delta t}\, q_0)(\mathbf{x}_{ij})$$

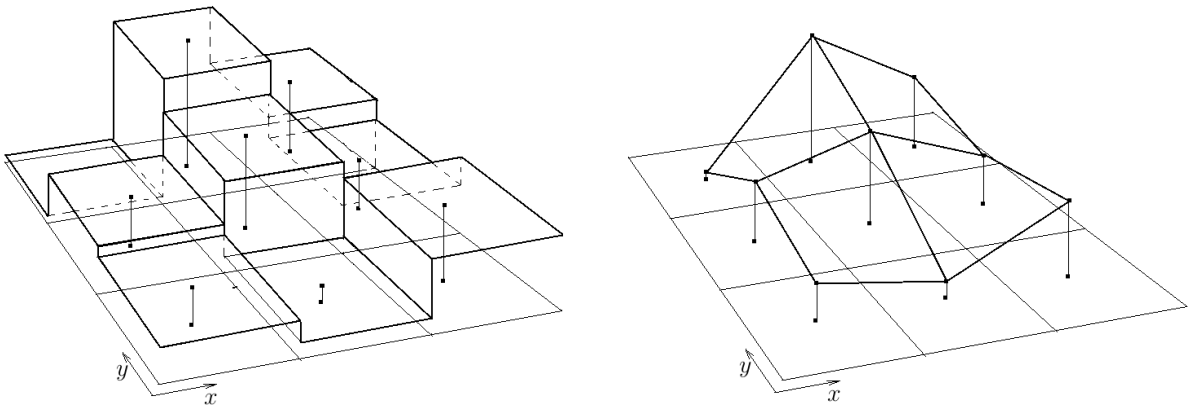

Figure 4.2: *Left*: Piecewise constant reconstruction. *Right*: Application of the sliding average to the same data amounts to a bilinear interpolation of the values q_{ij} interpreted as point values at $\mathbf{x}_{ij}$.

Lemma 4.1. *Provided all expressions exist, the two operators commute:*

$$A\, T_{\Delta t}\, q_0 \Big|_{\mathbf{x}_{ij}} = \quad T_{\Delta t}\, A\, q_0 \Big|_{\mathbf{x}_{ij}}$$

Proof. By linearity of $T_{\Delta t}$ (Theorem 2.3),

$$\begin{aligned}(A\,T_{\Delta t}\,q_0)(\mathbf{x}) &= \frac{1}{\Delta x \Delta y} \int\limits_{\left[-\frac{\Delta x}{2},\frac{\Delta x}{2}\right]\times\left[-\frac{\Delta y}{2},\frac{\Delta y}{2}\right]} \mathrm{d}\mathbf{s} \qquad (T_{\Delta t}\,q_0)(\mathbf{x}+\mathbf{s}) \\ &= \frac{1}{\Delta x \Delta y} T_{\Delta t} \int\limits_{\left[-\frac{\Delta x}{2},\frac{\Delta x}{2}\right]\times\left[-\frac{\Delta y}{2},\frac{\Delta y}{2}\right]} \mathrm{d}\mathbf{s} \qquad q_0(\mathbf{x}+\mathbf{s}) \\ &= T_{\Delta t}\,(Aq_0)(\mathbf{x})\end{aligned}$$

□

In short, for linear systems the last two steps of *reconstruction-evolution-averaging* can be turned around to be *reconstruction-averaging-evolution* which tremendously simplifies the derivation: It suffices to find the solution of (2.13) at $\mathbf{x}_{ij}$ taking the sliding-averaged initial data Aq_0. The sliding average of a piecewise constant reconstruction on a 2-d grid amounts to a bilinear interpolation of the values q_{ij} taken at points $\mathbf{x}_{ij}$ (see Fig. 4.2, right).

Example 4.1 (Godunov scheme for linear advection)**.** *Consider linear advection in 1-d*

$$\begin{aligned}\partial_t q + c\partial_x q = 0 \qquad c > 0 \\ q : \mathbb{R}^0_+ \times \mathbb{R} \to \mathbb{R}\end{aligned}$$

and piecewise constant initial data $q_0(x) = q_i$ if $x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. Application of the sliding average for $x \in [x_{i-1}, x_i]$:

$$\begin{aligned}(Aq_0)(x) &= \frac{1}{\Delta x}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \mathrm{d}s\, q_0(x+s) = \frac{1}{\Delta x}\int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} \mathrm{d}s\, q_0(s) \\ &= \frac{1}{\Delta x}\left(x_{i-\frac{1}{2}} - x + \frac{\Delta x}{2}\right) q_{i-1} + \frac{1}{\Delta x}\left(x + \frac{\Delta x}{2} - x_{i-\frac{1}{2}}\right) q_i\end{aligned}$$

The exact evolution operator at x_i evaluates the sliding averaged initial data at $x_i - c\Delta t$ if $c\Delta t < \Delta x$:

$$T_{\Delta t}(Aq_0)(x_i) = q_i - \frac{c\Delta t}{\Delta x}(q_i - q_{i-1})$$

This is the usual upwind scheme. ◁

4.2.3 Finite volume scheme

Performing the evaluation of the exact solution formulae as outlined in Section 4.2.2 is straightforward: In every one of the four quadrants all the derivatives of the initial data exist. At the locations where the quadrants meet, the initial data are continuous with, in general, discontinuous first derivatives. However, the derivatives are continuous in

r-direction, as the kinks are all oriented towards the location $\mathbf{x}_{ij}$ (see Fig. 4.2, right). Thus the radial derivatives never lead to the appearance of actual distributions and the solution is a function. The reason for the different behaviour as compared to the Riemann Problem in Section 2.2.5 is that here the evolution operator is applied onto sliding-averaged discontinuities which are continuous. This goes back to the *averaging*-step of the Godunov procedure.

After carefully collecting the different terms one obtains the following numerical scheme (the notation is introduced on page 13)

$$\begin{aligned}
u^{n+1} &= u_{ij}^n - \frac{c\Delta t}{2\epsilon\Delta x}\left([p]_{i\pm1,j} - [[u]]_{i\pm\frac12,j}\right) \\
&\quad - \frac12\frac{(c\Delta t)^2}{\epsilon^2\Delta x\Delta y}\left(-\frac{1}{2\pi}[[[[u]]_{i\pm\frac12}]]_{j\pm\frac12} - \frac14[[v]_{i\pm1}]_{j\pm1} + \frac14[[[p]_{i\pm1}]]_{j\pm\frac12}\right) \qquad (4.3)\\
v^{n+1} &= v_{ij}^n - \frac{c\Delta t}{2\epsilon\Delta y}\left([p]_{i,j\pm1} - [[v]]_{i,j\pm\frac12}\right) \\
&\quad - \frac12\frac{(c\Delta t)^2}{\epsilon^2\Delta x\Delta y}\left(-\frac{1}{2\pi}[[[[v]]_{i\pm\frac12}]]_{j\pm\frac12} - \frac14[[u]_{i\pm1}]_{j\pm1} + \frac14[[[p]]_{i\pm\frac12}]_{j\pm1}\right) \\
p^{n+1} &= p_{ij} - \frac{c\Delta t}{2\epsilon\Delta x}\left([u]_{i\pm1,j} - [[p]]_{i\pm\frac12,j}\right) - \frac{c\Delta t}{2\epsilon\Delta y}\left([v]_{i,j\pm1} - [[p]]_{i,j\pm\frac12}\right) \\
&\quad - \frac12\frac{(c\Delta t)^2}{\epsilon^2\Delta x\Delta y}\left(\frac14[[[u]_{i\pm1}]]_{j\pm\frac12} + \frac14[[[v]]_{i\pm\frac12}]_{j\pm1} - 2\cdot\frac{1}{2\pi}[[[[p]]_{i\pm\frac12}]]_{j\pm\frac12}\right) \qquad (4.4)
\end{aligned}$$

This scheme is conservative because it is a Godunov scheme, and can be written as

$$q^{n+1} = q^n - \frac{\Delta t}{\Delta x}\left(f^{(x)}_{i+\frac12,j} - f^{(x)}_{i-\frac12,j}\right) - \frac{\Delta t}{\Delta y}\left(f^{(y)}_{i,j+\frac12} - f^{(y)}_{i,j-\frac12}\right)$$

One can identify the x-flux through the boundary located at $x_{i+\frac12}$:

$$\begin{aligned}
f^{(x)}_{i+\frac12} &= \frac12\frac{c}{\epsilon}\begin{pmatrix} \{p\}_{i+\frac12,j} - [u]_{i+\frac12,j} \\ 0 \\ \{u\}_{i+\frac12,j} - [p]_{i+\frac12,j} \end{pmatrix} \\
&\quad + \frac12\frac{c\Delta t}{\epsilon\Delta y}\cdot\frac{c}{\epsilon}\begin{pmatrix} -\frac{1}{2\pi}[[[u]_{i+\frac12}]]_{j\pm\frac12} - \frac14[\{v\}_{i+\frac12}]_{j\pm1} + \frac14[[\{p\}_{i+\frac12}]]_{j\pm\frac12} \\ 0 \\ \frac14[[v]_{i+\frac12}]_{j\pm1} - \frac{1}{2\pi}[[[p]_{i+\frac12}]]_{j\pm\frac12} \end{pmatrix} \qquad (4.5)
\end{aligned}$$

The corresponding perpendicular flux is its symmetric analogue. The first bracket is the flux obtained in a dimensionally split situation.

The appearance of prefactors which contain π in schemes derived using the exact multi-dimensional evolution operators has already been noticed in [LMMW00], but none of the schemes mentioned therein matches the one presented here.

For better comparison to other schemes, below the scheme (4.3)–(4.4) is given in the variables prior to symmetrization, i.e. such that it is a numerical approximation to (2.40)–(2.41). This is achieved by applying the transformation (2.9) or, which is equivalent, by replacing $p \mapsto \frac{p}{c\epsilon}$:

$$
\begin{aligned}
u^{n+1} &= u_{ij} - \frac{\Delta t}{2\Delta x}\left(\frac{1}{\epsilon^2}[p]_{i\pm1,j} - \frac{c}{\epsilon}[[u]]_{i\pm\frac{1}{2},j}\right) \\
&\quad - \frac{1}{2}\frac{\Delta t}{\Delta x}\frac{c\Delta t}{\epsilon\Delta y}\left(-\frac{1}{2\pi}\frac{c}{\epsilon}[[[[u]]_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}} - \frac{1}{4}\frac{c}{\epsilon}[[v]_{i\pm1}]_{j\pm1} + \frac{1}{4}\frac{1}{\epsilon^2}[[[p]_{i\pm1}]]_{j\pm\frac{1}{2}}\right) \quad (4.6)\\
v^{n+1} &= v_{ij} - \frac{\Delta t}{2\Delta y}\left(\frac{1}{\epsilon^2}[p]_{i,j\pm1} - \frac{c}{\epsilon}[[v]]_{i,j\pm\frac{1}{2}}\right) \\
&\quad - \frac{1}{2}\frac{\Delta t}{\Delta y}\frac{c\Delta t}{\epsilon\Delta x}\left(-\frac{1}{2\pi}\frac{c}{\epsilon}[[[[v]]_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}} - \frac{1}{4}\frac{c}{\epsilon}[[u]_{i\pm1}]_{j\pm1} + \frac{1}{4}\frac{1}{\epsilon^2}[[[p]]_{i\pm\frac{1}{2}}]_{j\pm1}\right) \\
p^{n+1} &= p_{ij} - \frac{\Delta t}{2\Delta x}\left(c^2[u]_{i\pm1,j} - \frac{c}{\epsilon}[[p]]_{i\pm\frac{1}{2},j}\right) - \frac{\Delta t}{2\Delta y}\left(c^2[v]_{i,j\pm1} - \frac{c}{\epsilon}[[p]]_{i,j\pm\frac{1}{2}}\right) \\
&\quad - \frac{1}{2}\frac{c\Delta t^2}{\epsilon\Delta x\Delta y}\left(\frac{1}{4}c^2[[[u]_{i\pm1}]]_{j\pm\frac{1}{2}} + \frac{1}{4}c^2[[[v]]_{i\pm\frac{1}{2}}]_{j\pm1} - 2\cdot\frac{1}{2\pi}\frac{c}{\epsilon}[[[[p]]_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}}\right) \quad (4.7)
\end{aligned}
$$

Dimensionally split schemes in two spatial dimensions have a stability condition ([GZI+76], Eq. 8.15, p. 63)

$$c\Delta t < \frac{1}{\frac{1}{\Delta x} + \frac{1}{\Delta y}}$$

which for square grids gives a maximum CFL number of 0.5. As the present scheme is an exact multidimensional Godunov scheme it is stable up to the physical CFL number.

4.2.4 Stability and numerical examples

The scheme (4.3)–(4.4) is applied to two test cases. The first one is the Riemann Problem considered analytically in Section 2.2.5. The second is a test of the low Mach number abilities of the scheme.

4.2.4.1 Riemann Problem

The initial setup is that of Section 2.2.5 (Fig. 2.1); it is solved on a square grid of 101×101 cells on a domain that is large enough such that the disturbance produced by the corner has not reached the boundaries for $t = 0.25$. Here, $c = \epsilon = 1$.

The results are shown in Fig. 4.3.

In Fig. 4.4 the y-component of the velocity obtained with the numerical scheme is compared to the analytic solution (2.39) found in Section 2.2.5.

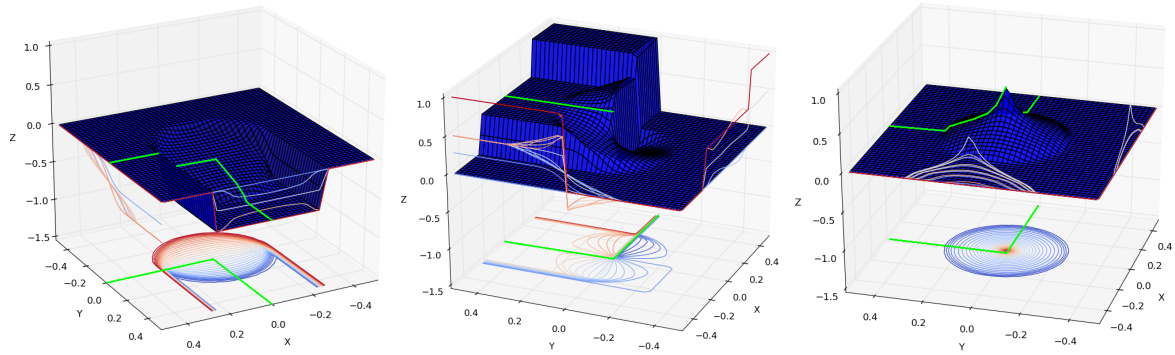

Figure 4.3: Solution of Riemann problem at time $ct = 0.25$ using scheme (4.3)–(4.4). *Left*: Pressure. *Center*: x-velocity. *Right*: y-velocity. Compare the images to Fig. 2.2. The sharpness of the discontinuities is due to the particular choice of the CFL number being very close to 1.

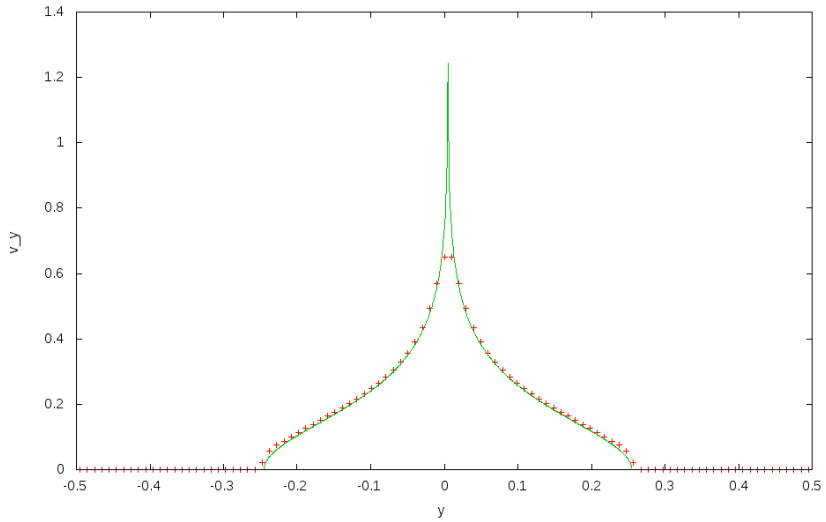

Figure 4.4: The y-component of the velocity obtained by the numerical scheme (4.3)–(4.4) is shown together with the analytic solution (2.39). Although the latter has only been computed along one particular axis, the solution is symmetric around the location of the corner in the initial data because of self-similarity and scale invariance.

4.2.4.2 Low Mach number vortex

The second test shows the properties of the scheme in the limit $\epsilon \to 0$. The setup is that of a stationary, divergencefree velocity field and constant pressure:

$$p_0(\mathbf{x}) = 1$$
$$\mathbf{v}(\mathbf{x}) = \mathbf{e}_\varphi \begin{cases} \frac{r}{d} & r < d \\ 2 - \frac{r}{d} & d \le r < 2d \\ 0 & \text{else} \end{cases}$$

The velocity thus has a compact support, which is entirely contained in the computational domain, discretized by 51×51 square cells. Here $c = 1$ and $d = 0.2$. Zero-gradient

boundaries are enforced.

Fig. 4.5 shows the error norm at time $t = 1$ for different CFL numbers; results for the scheme (4.3)–(4.4) are shown as solid lines. As has been stated earlier, it is stable until a CFL number of 1, which is confirmed by a rapid increase of error beyond this value. Additionally, the error drops significantly when the CFL number approaches 1 from below. This drop is more and more abrupt the lower ϵ is.

The dimensionally split solver is known to display artefacts in the limit $\epsilon \to 0$ (see e.g. [GV99], [Bar17a]). Results obtained with this scheme are shown in the same Figure by dashed lines. For small CFL numbers, the flux (4.5) approaches the dimensionally split case, which is confirmed experimentally. For the dimensionally split scheme the error does not depend on the CFL number. Also the small stability region CFL < 0.5, as well as the growth of the error for decreasing ϵ are prominent in the Figure.

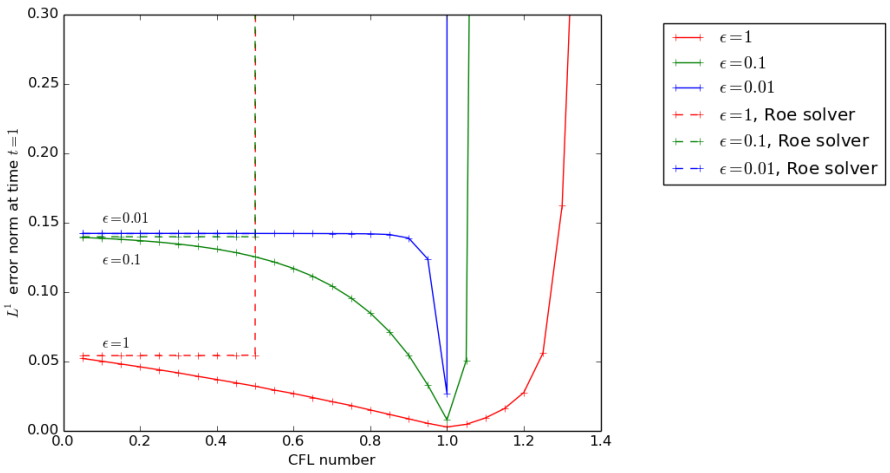

Figure 4.5: Solid lines show L^1 error norms of the numerical solution at time $t = 1$ for the scheme (4.3)–(4.4) as a function of the CFL number and for three choices of ϵ. Dashed lines, for comparison, display the same for the dimensionally split scheme.

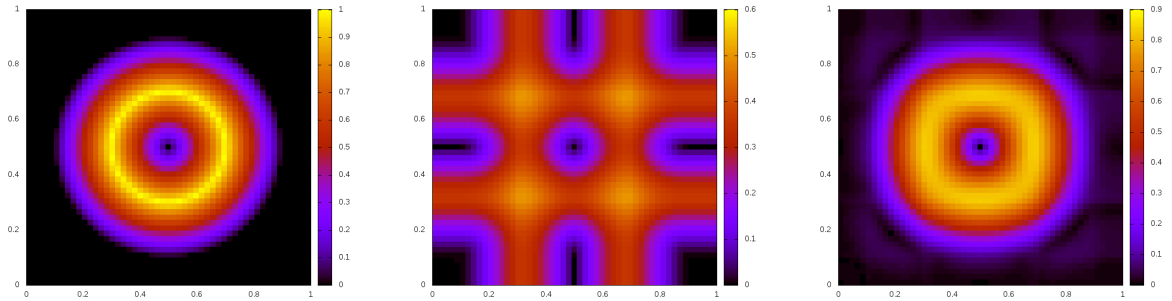

Figure 4.6: Solution of the vortex initial data at time $ct = 1$ using scheme (4.3)–(4.4) for $\epsilon = 10^{-2}$. The quantity shown in color is the magnitude of the velocity. *Left*: Exact solution = initial data. *Center*: CFL = 0.8. *Right*: CFL = 1.

This growth is equally observed for the scheme presented here. Therefore it is not suitable for the low Mach number regime.

At interesting feature however is that the choice CFL = 1 results in a considerable improvement, as is shown in Fig. 4.3 for the case $\epsilon = 10^{-2}$. Although not useful for actual computations, it is instructive to understand this behaviour. A reason can be found by considering the modified equation for the scheme (4.6)–(4.7) (here it is of advantage to use the scheme in its non-symmetrized shape). The modified equation reads

$$\begin{aligned}
\frac{u^{n+1}-u^n}{\Delta t} + \frac{1}{\epsilon^2}\partial_x p &= \Delta x \frac{c}{2\epsilon}\partial_x(\partial_x u + \text{CFL}\ \partial_y v) + \mathcal{O}(\Delta x^2) \\
\frac{v^{n+1}-v^n}{\Delta t} + \frac{1}{\epsilon^2}\partial_y p &= \Delta x \frac{c}{2\epsilon}\partial_y(\text{CFL}\ \partial_x u + \partial_y v) + \mathcal{O}(\Delta x^2) \\
\frac{p^{n+1}-p^n}{\Delta t} + c^2(\partial_x u + \partial_y v) &= \Delta x \frac{c}{2\epsilon}(\partial_x^2 p + \partial_y^2 p) + \mathcal{O}(\Delta x^2)
\end{aligned}$$

Here, for simplicity, $\Delta x = \Delta y$ has been used. Usually, the low Mach number artefacts are attributed to the diffusion terms scaling as $\mathcal{O}(\epsilon^{-1})$. One observes that here choosing a CFL number of 1 makes the velocity diffusion become a gradient of the divergence. In the limit $\epsilon \to 0$ the divergence becomes $\mathcal{O}(\epsilon)$, and thus the highest order terms in the modified equation become $\mathcal{O}(1)$. The low Mach number artefacts are then entirely due to contributions from the $\mathcal{O}(\Delta x^2)$ terms.

4.3 Stability of one-dimensional schemes

The upwind/Roe scheme, being the Godunov scheme for linear acoustics in one spatial dimension, is stable under explicit time integration, provided the CFL number is less than unity. This is also the physical stability condition. A central scheme, for example, is unstable under first order explicit time integration. There are many schemes that are stable, but under a restricted CFL condition. For example, a dimensionally split scheme that is stable for CFL < 1 in 1-d, is only stable for CFL < 0.5 in two spatial dimensions. When it comes to dimensionally split schemes, the modifications of the diffusion matrix that are needed in order to enforce stationarity preservation at the same time affect the stability of the scheme. Indeed, for example, the central scheme is stationarity preserving, but cannot be used with an explicit time integrator. In this Section the stability of a particular class of dimensionally split schemes is studied theoretically in one spatial dimension. To extend the analysis to two spatial dimensions with the same methods seems, generally, an unfeasible task. Here one has to resort to numerical experiments.

These methods can also be used to perform a *linear stability* analysis of numerical schemes for the Euler equations. Therefore the presentation of the methods is partly adapted to this task. The Section is based on work partly published in [BEK+17].

4.3.1 General procedure in one spatial dimension

4.3.1.1 Homogeneous systems of equations

Consider a one-dimensional linear hyperbolic system of 2 equations:

$$\partial_t q + J\partial_x q = 0 \qquad\qquad q : \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}^2$$

This is a one-dimensional analogue of schemes considered in Section 3. J is the 2×2 Jacobian matrix.

Assume an explicit numerical scheme for this system on a uniform grid of spacing Δx to be of the form

$$\Delta x \frac{q_i^{n+1} - q_i^n}{\Delta t} + \frac{1}{2} J(q_{i+1}^n - q_{i-1}^n) - \frac{1}{2} D(q_{i+1}^n - 2q_i^n + q_{i-1}^n) = 0 \tag{4.8}$$

with D a constant 2×2 matrix (*diffusion matrix*).

Then, analogously to Section 3 in order to perform a von Neumann analysis consider a Fourier mode with wave vector k

$$q_i^n = \hat{q} \exp(-\mathbb{i}\omega n \Delta t + \mathbb{i} \Delta x i k)$$

For simplicity, define $\beta := \Delta x \cdot k$. Then inserting this ansatz into (4.8) yields

$$\Delta x \frac{\exp(-\mathbb{i}\omega \Delta t) - 1}{\Delta t} + J \mathbb{i} \sin\beta - D(\cos\beta - 1) = 0$$

Also define

$$\nu := \frac{\Delta t}{\Delta x}$$

The Fourier mode stays bounded if $|\exp(-\mathbb{i}\omega\Delta t)| \leq 1$.

Definition 4.4 (Amplification matrix)**.** *Define the* amplification matrix $\mathcal{A}$

$$\mathcal{A} = \mathbb{1} - \nu\Big(J\mathbb{i}\sin\beta + D(1 - \cos\beta)\Big)$$

Definition 4.5 (Stability)**.** *Call $\lambda \in \mathbb{C}$ an eigenvalue of $\mathcal{A}$. A linear scheme is called* (von Neumann) stable *if* $\max|\lambda| \leq 1$.

4.3.2 Scalar upwinding

The simplest choice of a diffusion matrix D is a scalar, i.e. $D = d \cdot \mathrm{id}$. This case is considered to briefly demonstrate how the stability analysis is performed.

Theorem 4.2 (Scalar stability)**.** *If the diffusion matrix is a scalar d (times the identity matrix), and $\bar{\lambda}$ denotes an eigenvalue of J, then for stability*

$$\boxed{d \geq |\bar{\lambda}|}$$

Proof. The amplification matrix becomes

$$\mathrm{id}\,(1 - \nu d(1 - \cos\beta)) - \nu J \cdot \mathbb{i}\sin\beta$$

and from

$$\begin{aligned} 0 &= \det\left(\mathrm{id}\left(1-\nu d(1-\cos\beta)\right) - \nu J\cdot \mathbb{i}\sin\beta + \lambda \mathrm{id}\right) \\ &= \det\left(\mathrm{id}\frac{1+\lambda-\nu d(1-\cos\beta)}{\nu\mathbb{i}\sin\beta} - J\right)(\nu\mathbb{i}\sin\beta)^{\text{some power}} \end{aligned}$$

one observes that the eigenvalue λ of the amplification matrix, with $\bar{\lambda}$ the eigenvalue of J, is given as

$$\lambda = \nu\mathbb{i}\sin\beta\bar{\lambda} + \nu d(1-\cos\beta) - 1 \tag{4.9}$$

Assuming both $\bar{\lambda}$ and d to be real, one has

$$\begin{aligned} 1 &\overset{!}{>} |\lambda|^2 = (\nu d(1-\cos\beta)-1)^2 + (\nu\sin\beta\bar{\lambda})^2 \\ 0 &\overset{!}{>} \nu d^2(1-\cos\beta)^2 - 2d(1-\cos\beta) + \nu\sin^2\beta\bar{\lambda}^2 \\ \nu &\overset{!}{<} \frac{2d(1-\cos\beta)}{d^2(1-\cos\beta)^2+\sin^2\beta\bar{\lambda}^2} \\ \nu &\overset{!}{<} \frac{2d}{d^2(1-\cos\beta)+(1+\cos\beta)\bar{\lambda}^2} = \frac{2d}{d^2+\bar{\lambda}^2+\cos\beta(\bar{\lambda}^2-d^2)} =: G(\cos\beta) \end{aligned}$$

The denominator is zero at $\cos\beta = \dfrac{d^2+\bar{\lambda}^2}{d^2-\bar{\lambda}^2} = 1 + 2\dfrac{\bar{\lambda}^2}{d^2-\bar{\lambda}^2}$. Thus if $0 < |d| < |\bar{\lambda}|$, there is a singularity inside $[-1,1]$ and the G-image of the interval $[-1,1]$ is unbounded to both sides of the real line. Under the condition, that $|d| \geq |\bar{\lambda}|$, G is (in $[-1,1]$) a monotone function of $\cos\beta$ (its derivative does not change sign), therefore it attains its maximum at the boundary $\cos\beta = \pm 1$:

$$\nu \overset{!}{<} \min\left(\frac{d}{\bar{\lambda}^2}, \frac{1}{d}\right) \overset{d>|\bar{\lambda}|}{=} \frac{1}{d}$$

Thus for stability, d must be positive and

$$d \geq |\bar{\lambda}|$$

□

Scalar diffusion therefore has to be chosen positive and such that it exceeds the absolute value of every eigenvalue of the Jacobian.

4.3.3 Equal diagonal entries

In order to adapt the studies to the acoustic system, the Jacobian is assumed to have the following shape:

$$J = \begin{pmatrix} a & a_{12} \\ a_{21} & a \end{pmatrix} \tag{4.10}$$

with a, a_{12}, a_{21} real numbers.

For simplicity, consider first a diffusion matrix with equal entries on the diagonal

$$D = \begin{pmatrix} d & d_{12} \\ d_{21} & d \end{pmatrix} \tag{4.11}$$

Theorem 4.3. *Consider the Jacobian* (4.10) *and the diffusion matrix* (4.11)*. Define*

$$A = -a_{12}a_{21}\sin^2\beta + d_{12}d_{21}(1-\cos\beta)^2 \tag{4.12}$$
$$B = (a_{12}d_{21} + d_{12}a_{21})(1-\cos\beta)\sin\beta \tag{4.13}$$

The stability condition amounts to

$$\nu \overset{!}{<} 2\frac{d(1-\cos\beta) \mp \sqrt{\frac{\sqrt{A^2+B^2}+A}{2}}}{\left(d(1-\cos\beta) \mp \sqrt{\frac{\sqrt{A^2+B^2}+A}{2}}\right)^2 + \left(a\sin\beta \mp sgn\,(B)\sqrt{\frac{\sqrt{A^2+B^2}-A}{2}}\right)^2} \tag{4.14}$$

which has to be true for all β.

Proof. The amplification matrix is

$$\begin{pmatrix} 1-\nu(a\mathrm{i}\sin\beta + d(1-\cos\beta)) & -\nu(a_{12}\mathrm{i}\sin\beta + d_{12}(1-\cos\beta)) \\ -\nu(a_{21}\mathrm{i}\sin\beta + d_{21}(1-\cos\beta)) & 1-\nu(a\mathrm{i}\sin\beta + d(1-\cos\beta)) \end{pmatrix}$$

Its eigenvalue λ fulfills

$$(1-\nu d(1-\cos\beta) - \nu a\mathrm{i}\sin\beta - \lambda)^2 = \nu^2[a_{12}\mathrm{i}\sin\beta + d_{12}(1-\cos\beta)][a_{21}\mathrm{i}\sin\beta + d_{21}(1-\cos\beta)]$$

$$=: \nu^2(A + B\mathrm{i}) \tag{4.15}$$

$$1-\nu d(1-\cos\beta) - \nu a\mathrm{i}\sin\beta \pm \nu\sqrt{A+B\mathrm{i}} = \lambda$$

The square root of a complex number can be rewritten as

$$\sqrt{A+B\mathrm{i}} = \sqrt{\frac{\sqrt{A^2+B^2}+A}{2}} + \mathrm{i}\,\mathrm{sgn}(B)\cdot\sqrt{\frac{\sqrt{A^2+B^2}-A}{2}}$$

Therefore

$$|\lambda|^2 = \left(1 - \nu d(1-\cos\beta) \pm \nu\sqrt{\frac{\sqrt{A^2+B^2}+A}{2}}\right)^2 + \left(-\nu a \sin\beta \pm \operatorname{sgn}(B)\nu\sqrt{\frac{\sqrt{A^2+B^2}-A}{2}}\right)^2$$

$$= \left(-1 + \nu d(1-\cos\beta) \mp \nu\sqrt{\frac{\sqrt{A^2+B^2}+A}{2}}\right)^2 + \left(\nu a \sin\beta \mp \operatorname{sgn}(B)\nu\sqrt{\frac{\sqrt{A^2+B^2}-A}{2}}\right)^2$$

$$= 1 + \nu^2\left(d(1-\cos\beta) \mp \sqrt{\frac{\sqrt{A^2+B^2}+A}{2}}\right)^2 - 2\nu\left(d(1-\cos\beta) \mp \sqrt{\frac{\sqrt{A^2+B^2}+A}{2}}\right)$$

$$+ \nu^2\left(a\sin\beta \mp \operatorname{sgn}(B)\sqrt{\frac{\sqrt{A^2+B^2}-A}{2}}\right)^2 \overset{!}{<} 1$$

Solving for ν proves the assertion. □

Corollary 4.4 (upwind/Roe scheme)**.** *Consider the rescaled system of acoustic equations with additionally an advective velocity: Take $a_{12} = \frac{1}{\epsilon^2}$, $a_{21} = c^2$ and the diagonal value a of the Jacobian v with ϵ, c, v real, positive numbers, i.e*

$$\begin{pmatrix} v & \frac{1}{\epsilon^2} \\ c^2 & v \end{pmatrix} \tag{4.16}$$

Consider first $d_{12} = d_{21} = 0$ and $d = \left|\frac{c}{\epsilon} + v\right|$. Then the scheme is stable if

$$\nu \overset{!}{<} \frac{1}{\left|\frac{c}{\epsilon} + v\right|}$$

Proof. Inserting the given parameters yields

$$A = -\frac{c^2}{\epsilon^2}\sin^2\beta < 0$$

$$B = 0$$

$$\nu \overset{!}{<} 2\frac{d(1-\cos\beta) \mp \sqrt{\frac{|A|+A}{2}}}{\left(d(1-\cos\beta) \mp \sqrt{\frac{|A|+A}{2}}\right)^2 + \left(v\sin\beta \mp \operatorname{sgn}(B)\sqrt{\frac{|A|-A}{2}}\right)^2}$$

$$= 2\frac{d(1-\cos\beta)}{d^2(1-\cos\beta)^2 + \left(v \mp \operatorname{sgn}(B)\frac{c}{\epsilon}\right)^2\sin^2\beta} = \frac{2d}{d^2(1-\cos\beta) + \left(v \mp \operatorname{sgn}(B)\frac{c}{\epsilon}\right)^2(1+\cos\beta)}$$

Results of scalar stability (Theorem 4.2) can thus be used here and yield $d \geq \left|\frac{c}{\epsilon} + v\right|$, and in case of equality one obtains the assertion. □

Using the scheme as suggested in [MRE15]

$$P = \begin{pmatrix} 1 & -\frac{\delta}{c\epsilon} \\ c\delta\epsilon & 1 \end{pmatrix}$$

leads to a different form of the upwinding matrix:

$$D = \frac{1}{c\epsilon\sqrt{c^2(1+\delta^2) - \delta^2\epsilon^2 v^2}} \begin{pmatrix} c^3 & \frac{c^2\delta + c\epsilon v - \delta\epsilon^2 v^2}{\epsilon} \\ c^2\epsilon(-c^2\delta + c\epsilon v + \delta\epsilon^2 v^2) & c^3 \end{pmatrix} \tag{4.17}$$

Define $\tau := \sqrt{c^2(1+\delta^2) - \delta^2\epsilon^2 v^2}$. For the Jacobian (4.16), one can investigate the limit of small ϵ.

Corollary 4.5. *Consider the Jacobian* (4.16) *and the upwinding matrix* (4.17) *in the limit* $\epsilon \to 0$. *If* $\delta \in \mathcal{O}(1)$ *and the scheme is stable, then the stability condition is* $\nu_{max} \in \mathcal{O}\left(\frac{\epsilon}{c}\right)$; *if* $\delta \in \mathcal{O}(1/\epsilon)$ *and the scheme is stable, then the stability condition is* $\nu_{max} \in \mathcal{O}\left(\frac{\epsilon^2}{c}\right)$.

Proof. For the components of the upwinding matrix one has (having in mind the two cases $\delta \in \mathcal{O}(\frac{1}{\epsilon})$ and $\delta \in \mathcal{O}(1)$):

$$\tau \sim c\sqrt{1+\delta^2}$$

$$\bar{d} \sim \frac{c}{\sqrt{1+\delta^2}\epsilon} \quad d_{12} \sim \frac{\delta}{\sqrt{1+\delta^2}\epsilon^2} \quad d_{21} \sim -\frac{c^2\delta}{\sqrt{1+\delta^2}}$$

Therefore

$$\mathcal{A} = -\frac{c^2}{\epsilon^2}\left(\sin^2\beta + \frac{\delta^2}{1+\delta^2}(1-\cos\beta)^2\right)$$
$$\mathcal{B} = \frac{2cv}{\sqrt{1+\delta^2}\epsilon}(1-\cos\beta)\sin\beta$$

where due to a lot of cancellations the exact values were used for $\mathcal{B}$. Whereas in both cases $\mathcal{A} \in \mathcal{O}(1/\epsilon^2)$, one has

$$\mathcal{B} \in \mathcal{O}(1/\epsilon) \qquad \text{if } \delta \in \mathcal{O}(1)$$
$$\mathcal{B} \in \mathcal{O}(1) \qquad \text{if } \delta \in \mathcal{O}(1/\epsilon)$$

As, $|\mathcal{A}| = -\mathcal{A}$, $\sqrt{\mathcal{A}^2+\mathcal{B}^2} - \mathcal{A} \sim 2|\mathcal{A}|$ and

$$\sqrt{\mathcal{A}^2+\mathcal{B}^2} + \mathcal{A} \sim |\mathcal{A}|\frac{\mathcal{B}^2}{2\mathcal{A}^2} = \frac{\mathcal{B}^2}{2|\mathcal{A}|} \in \begin{cases} \mathcal{O}(1) & \text{if } \delta \in \mathcal{O}(1) \\ \mathcal{O}(\epsilon^2) & \text{if } \delta \in \mathcal{O}(1/\epsilon) \end{cases}$$

The term $\sqrt{\frac{\sqrt{\mathcal{A}^2+\mathcal{B}^2}+\mathcal{A}}{2}}$ will be compared to

$$\bar{d}(1-\cos\beta) \in \begin{cases} \mathcal{O}(1/\epsilon) & \text{if } \delta \in \mathcal{O}(1) \\ \mathcal{O}(1) & \text{if } \delta \in \mathcal{O}(1/\epsilon) \end{cases}$$

and the latter wins in both cases. Therefore

$$\begin{aligned} \nu &< 2\frac{d(1-\cos\beta)}{d^2(1-\cos\beta)^2 + \left(c\sin\beta \mp \operatorname{sgn}(B)\sqrt{|A|}\right)^2} \\ &\sim 2\frac{\frac{c}{\sqrt{1+\delta^2}\epsilon}(1-\cos\beta)}{\frac{c^2}{(1+\delta^2)\epsilon^2}(1-\cos\beta)^2 + \frac{c^2}{\epsilon^2}\left|\sin^2\beta - \frac{\delta^2}{1+\delta^2}(1-\cos\beta)^2\right|} \\ &= \frac{\epsilon}{c}\frac{\frac{2}{\sqrt{1+\delta^2}}}{\frac{1}{1+\delta^2}(1-\cos\beta) + \left|(1+\cos\beta) - \frac{\delta^2}{1+\delta^2}(1-\cos\beta)\right|} \\ &= \frac{\epsilon}{c}\frac{2\sqrt{1+\delta^2}}{1-\cos\beta+|1+(1+2\delta^2)\cos\beta|} \end{aligned}$$

Now a minimum over all $\beta \in [0, 2\pi)$ has to be performed in order to obtain the global maximum value of ν. If $\delta \in \mathcal{O}(1)$ (in particular one might be interested to recover for $\delta = 0$ the usual Roe scheme) then

$$\nu_{\max} \sim \frac{\epsilon}{c}$$

if a suitable minimizing $\cos\beta_{\min}$ exists, which is $\mathcal{O}(1)$ (trivially the case for $\delta = 0$).

However if $\delta \in \mathcal{O}(1/\epsilon)$, then $|\cos\beta_{\min}| = 1$ and

$$\nu_{\max} \sim \frac{\epsilon}{c}\frac{\sqrt{1+\delta^2}}{\delta^2} \in \mathcal{O}\left(\frac{\epsilon^2}{c}\right)$$

□

4.3.4 Arbitrary diagonal entries

For the Jacobian (4.10) consider now an arbitrary upwinding matrix

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix} \tag{4.20}$$

Theorem 4.4. *Given the diffusion matrix* (4.20) *for the Jacobian* (4.10)*, the stability condition from Proposition 4.3 can be used, if the following substitutions are made:*

$$\begin{aligned} d &\mapsto \bar{d} := \frac{d_{11}+d_{22}}{2} \\ A &\mapsto \tilde{A} = \left(\frac{d_{22}-d_{11}}{2}\right)^2 (1-\cos\beta)^2 + A \end{aligned}$$

Proof. The amplification matrix is then given by

$$\begin{pmatrix} 1-\nu(a\mathbb{i}\sin\beta + d_{11}(1-\cos\beta)) & -\nu(a_{12}\mathbb{i}\sin\beta + d_{12}(1-\cos\beta)) \\ -\nu(a_{21}\mathbb{i}\sin\beta + d_{21}(1-\cos\beta)) & 1-\nu(a\mathbb{i}\sin\beta + d_{22}(1-\cos\beta)) \end{pmatrix}$$

and one finds for its eigenvalue λ

$$\begin{aligned}&[1-\nu(a\mathbb{i}\sin\beta + d_{11}(1-\cos\beta)) - \lambda][1-\nu(a\mathbb{i}\sin\beta + d_{22}(1-\cos\beta)) - \lambda] \\ &= \nu^2[a_{12}\mathbb{i}\sin\beta + d_{12}(1-\cos\beta)][a_{21}\mathbb{i}\sin\beta + d_{21}(1-\cos\beta)] =: \nu^2(A + B\mathbb{i})\end{aligned}$$

Note that A and B are the same as in (4.12) ans (4.13). Define for the moment

$$1-\nu(a\mathbb{i}\sin\beta + d_{ii}(1-\cos\beta)) =: \alpha_i$$

Then the above equation reads

$$\begin{aligned}(\alpha_1-\lambda)(\alpha_2-\lambda) &= \nu^2(A+B\mathbb{i}) \\ \lambda^2 - \lambda(\alpha_1+\alpha_2) + \alpha_1\alpha_2 &= \nu^2(A+B\mathbb{i}) \\ \left(\lambda - \frac{\alpha_1+\alpha_2}{2}\right)^2 - \frac{(\alpha_1+\alpha_2)^2}{4} + \alpha_1\alpha_2 &= \nu^2(A+B\mathbb{i}) \\ \left(\lambda - \frac{\alpha_1+\alpha_2}{2}\right)^2 &= \nu^2(A+B\mathbb{i}) + \left(\frac{\alpha_1-\alpha_2}{2}\right)^2\end{aligned}$$

Now

$$\begin{aligned}\frac{\alpha_1+\alpha_2}{2} &= 1-\nu\left(a\mathbb{i}\sin\beta + \frac{d_{11}+d_{22}}{2}(1-\cos\beta)\right) \\ \frac{\alpha_1-\alpha_2}{2} &= \nu\frac{d_{22}-d_{11}}{2}(1-\cos\beta) \in \mathbb{R}\end{aligned}$$

Therefore formula (4.14) can be reused according to the substitution rules of the assertion. □

Corollary 4.6. *Consider the Jacobian* (4.16) *and the following upwinding matrix:*

$$D = \begin{pmatrix} |v| & \frac{1}{\epsilon^2} \\ -c^2 & \frac{2c}{\epsilon} \end{pmatrix}$$

Such a scheme is stable in the limit $\epsilon \to 0$, or alternatively if $v = 0$, with $\nu \in \mathcal{O}(\epsilon)$.

Proof. Stability follows from Theorem 4.4. One finds

$$\begin{aligned}d &= \frac{|v| + \frac{2c}{\epsilon}}{2} \\ A &= -2\frac{c^2}{\epsilon^2}(1-\cos\beta) + \left(\frac{\frac{2c}{\epsilon} - |v|}{2}\right)^2 (1-\cos\beta)^2 \\ &= -2\frac{c^2}{\epsilon^2}(1-\cos\beta) + \left(\frac{c^2}{\epsilon^2} - 2\frac{c}{\epsilon}\frac{|v|}{2} + \frac{|v|^2}{4}\right)(1-\cos\beta)^2 \\ &= -\frac{c^2}{\epsilon^2}\sin^2\beta + |v|\cdot\mathcal{O}(\epsilon^{-1}) \\ B &= 0\end{aligned}$$

In the limit $\epsilon \to 0$, or for $v = 0$, the stability condition amounts to

$$\nu \overset{!}{<} 2\frac{d(1-\cos\beta)}{d^2(1-\cos\beta)^2 + \sin^2\beta\left(v \mp \operatorname{sgn}(B)\frac{c}{\epsilon}\right)^2} \simeq \frac{\epsilon}{c}$$

□

This method is investigated further in Section 4.4.1.

If the upwinding matrix D shall scale in ϵ elementwise as the Jacobian of linear acoustics, one can investigate the highest order of ϵ appearing in the CFL condition in quite some generality.

Theorem 4.5. *Take, in* (4.10) *and* (4.20)*, a, a_{21}, d_{11}, d_{22} and d_{21} to be asymptotically constant ($\mathcal{O}(1)$) and*

$$a_{12} := \frac{\tilde{a}_{12}}{\epsilon^2} \qquad\qquad d_{12} := \frac{\tilde{d}_{12}}{\epsilon^2} + \mathcal{O}\left(\frac{1}{\epsilon}\right)$$

Assume additionally that $\tilde{a}_{12}d_{21} + \tilde{d}_{12}a_{21} = 0$, $\tilde{d}_{12}d_{21} < 0$ and $\tilde{a}_{12}a_{21} > 0$. Then ν_{max}, if it exists, is $\nu_{max} \in \mathcal{O}(\epsilon^2)$.

Proof.

$$\begin{aligned}
A &= -a_{12}a_{21}\sin^2\beta + d_{12}d_{21}(1-\cos\beta)^2 + \left(\frac{d_{22}-d_{11}}{2}\right)^2(1-\cos\beta)^2 \\
&= \frac{-\tilde{a}_{12}a_{21}\sin^2\beta + \tilde{d}_{12}d_{21}(1-\cos\beta)^2}{\epsilon^2} + \mathcal{O}\left(\frac{1}{\epsilon}\right) \\
B &= (a_{12}d_{21} + d_{12}a_{21})(1-\cos\beta)\sin\beta = \frac{\tilde{a}_{12}d_{21} + \tilde{d}_{12}a_{21}}{\epsilon^2}(1-\cos\beta)\sin\beta + \mathcal{O}\left(\frac{1}{\epsilon}\right) \\
d &:= \frac{d_{11}+d_{22}}{2} \in \mathcal{O}(1)
\end{aligned}$$

With $\tilde{d}_{12}d_{21} < 0$ and $\tilde{a}_{12}a_{21} > 0$, then $|A| + A \in \mathcal{O}\left(\frac{1}{\epsilon}\right)$ and $\sqrt{A^2+B^2} + A \sim \frac{B^2}{2|A|} \in \mathcal{O}(1)$.

Then

$$\nu_{\max} \to 2\frac{d(1-\cos\beta) + \mathcal{O}(1)}{|A|} \in \mathcal{O}(\epsilon^2)$$

□

In such a case, therefore, if there is stability, then an upwinding matrix that scales as the Jacobian and has its diagonal values equal must lead to a CFL condition scaling as ϵ^2.

There is another interesting situation that can be studied with this theory. Consider a numerical scheme whose amplification matrix has eigenvalues that scale $\mathcal{O}(1/\epsilon)$. Does this imply a CFL condition $\nu \in \mathcal{O}(\epsilon)$?

In general, the answer is no:

Theorem 4.6. *If the eigenvalues of the amplification matrix scale $\mathcal{O}(1/\epsilon)$, then the CFl condition of the scheme can be $\nu \in \mathcal{O}(\epsilon^2)$.*

Proof. This is shown by studying an example. Consider an amplification matrix which has the following eigenvalue

$$\lambda = \frac{A + \mathbb{i}B}{\epsilon}\nu + 1 + (C + D\mathbb{i})\nu$$

with A, B, C, D real numbers. Its absolute value, imposed to be less than one is

$$\begin{aligned}
\left(\frac{A}{\epsilon}\nu + 1 + C\nu\right)^2 + \left(\frac{B}{\epsilon} + D\right)^2 \nu^2 < 1 \\
2\left(\frac{A}{\epsilon} + C\right) + \left(\frac{A}{\epsilon} + C\right)^2 \nu + \left(\frac{B}{\epsilon} + D\right)^2 \nu < 0 \\
\nu < \frac{-2\left(\frac{A}{\epsilon} + C\right)}{\left(\frac{A}{\epsilon} + C\right)^2 + \left(\frac{B}{\epsilon} + D\right)^2} \\
= \frac{-2\left(A\epsilon + C\epsilon^2\right)}{\left(A + \epsilon C\right)^2 + \left(B + \epsilon D\right)^2}
\end{aligned}$$

Now if $A = 0$, it turns out that $\nu \in \mathcal{O}(\epsilon^2)$. □

In [BM05], for a particular numerical scheme the eigenvalues of the amplification matrix are found to scale $\mathcal{O}(1/\epsilon^2)$. The conclusion drawn is that stability is only possible if $\nu \in \mathcal{O}(\epsilon^2)$, which is confirmed by experiments. However, as the example above shows, things might be more complicated.

4.3.5 Amplification matrices with decomposing eigenspace

Consider the following 3×3 linear hyperbolic system

$$\partial_t \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} + \begin{pmatrix} a & a_{12} & 0 \\ 0 & a & a_{23} \\ 0 & a_{32} & a \end{pmatrix} \partial_x \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} = 0 \tag{4.21}$$

which shall be solved with a time-explicit scheme of the form (4.8) with

$$D = \begin{pmatrix} d_1 & d_{12} & d_{13} \\ 0 & d_2 & d_{23} \\ 0 & d_{32} & d_2 \end{pmatrix} \tag{4.22}$$

Observe that though the diagonal elements of the Jacobian are the same, there are two different values appearing on the diagonal of D.

Theorem 4.7. *The stability of a scheme for* (4.21) *with the diffusion matrix* (4.22) *is governed by the stability of the truncated system*

$$J' = \begin{pmatrix} a & a_{23} \\ a_{32} & a \end{pmatrix} \qquad D' = \begin{pmatrix} d_2 & d_{23} \\ d_{32} & d_2 \end{pmatrix}$$

and the condition $d_1 \geq |a|$.

Proof. The amplification matrix is

$$D = \begin{pmatrix} 1-\nu(a\mathbb{i}\sin\beta + d_1(1-\cos\beta)) & -\nu(a_{12}\mathbb{i}\sin\beta + d_{12}(1-\cos\beta)) & -\nu(d_{13}(1-\cos\beta)) \\ 0 & 1-\nu(a\mathbb{i}\sin\beta + d_2(1-\cos\beta)) & -\nu(a_{23}\mathbb{i}\sin\beta + d_{23}(1-\cos\beta)) \\ 0 & -\nu(a_{32}\mathbb{i}\sin\beta + d_{32}(1-\cos\beta)) & 1-\nu(a\mathbb{i}\sin\beta + d_2(1-\cos\beta)) \end{pmatrix}$$

Its eigenvalues λ fulfill

$$\begin{aligned}&[1-\nu(a\mathbb{i}\sin\beta + d_1(1-\cos\beta)) - \lambda][1-\nu(a\mathbb{i}\sin\beta + d_2(1-\cos\beta)) - \lambda]^2 \\ &= \nu^2[a_{32}\mathbb{i}\sin\beta + d_{32}(1-\cos\beta)][a_{23}\mathbb{i}\sin\beta + d_{23}(1-\cos\beta)][1-\nu(a\mathbb{i}\sin\beta + d_1(1-\cos\beta)) - \lambda]\end{aligned}$$

which factorizes into

$$1-\nu(a\mathbb{i}\sin\beta + d_1(1-\cos\beta)) = \lambda \tag{4.23}$$

and

$$[1-\nu(a\mathbb{i}\sin\beta + d_2(1-\cos\beta)) - \lambda]^2 \quad = \nu^2[a_{32}\mathbb{i}\sin\beta + d_{32}(1-\cos\beta)][a_{23}\mathbb{i}\sin\beta + d_{23}(1-\cos\beta)] \tag{4.24}$$

Equation (4.23) is easily recognized as a 1-dimensional stability result just as in Equation (4.9). A such leads to the stability condition $d_1 \geq |a|$ and – alone – it would give the stability condition $\nu < \frac{1}{d_1}$.

Equation (4.24) is just Equation (4.15) for the truncated matrices. This proves the assertion. □

In primitive variables the Jacobian and the upwinding matrix, when calculated with the scheme from [MRE15, BEK+17], are

$$J = \left(\begin{array}{c|cc} v & \rho & 0 \\ \hline 0 & v & \frac{1}{\rho\epsilon^2} \\ 0 & \rho c^2 & v \end{array}\right) \qquad D = \left(\begin{array}{c|cc} |v| & \frac{\rho(-c^2\delta + c\epsilon v + \delta\epsilon^2 v^2)}{c\tau} & -\frac{v}{c^2} + \frac{1}{\epsilon\tau} \\ \hline 0 & \frac{c^2}{\epsilon\tau} & \frac{c^2\delta + c\epsilon v - \delta\epsilon^2 v^2}{\rho c\epsilon^2\tau} \\ 0 & \frac{\rho c(-c^2\delta + c\epsilon v + \delta\epsilon^2 v^2)}{\tau} & \frac{c^2}{\epsilon\tau} \end{array}\right)$$

The condition $d_1 \geq |a|$ is fulfilled and gives the necessary condition $\nu < \frac{1}{|v|}$.

The relevant submatrices for Equation (4.24) have already been highlighted. They turn out to be (but for a factor of ρ) identical to those used in (4.16) and (4.17). The relevant conditions have thus already been discussed and carry over to this case without modification.

4.4 Dimensionally split schemes

It should seem surprising that taking into account (direction by direction) only one-dimensional information allows an efficient numerical solution of systems of PDEs in multiple spatial dimensions – if this approach were not ubiquitously in use. Such an approach is called a *dimensionally split* scheme.

Definition 4.6 (Dimensionally split scheme)**.** *Consider a linear $n \times n$ system in e.g. 2 spatial dimensions,*

$$\partial_t q + J_x \partial_x q + J_y \partial_y q = 0 \qquad\qquad q : \mathbb{R}_0^+ \times \mathbb{R}^2 \to \mathbb{R}^n$$

and two semi-discrete one-dimensional schemes

$$\partial_t q_{ij} + \mathcal{A}_x = 0 \qquad\qquad \partial_t q_{ij} + \mathcal{A}_y = 0$$

that discretize, respectively,

$$\partial_t q + J_x \partial_x q = 0 \qquad\qquad \partial_t q + J_y \partial_y q = 0$$

Here $\mathcal{A}_x$ and $\mathcal{A}_y$ are some stencils. The semi-discrete dimensionally split scheme *is given by*

$$\partial_t q + \mathcal{A}_x + \mathcal{A}_y = 0$$

The advantages of simple implementation are so overwhelming that anybody who is considering different schemes is confronted with the need to explain himself. A comparative discussion is postponed until Section 4.5, while this Section deals with dimensionally split schemes, that are stationarity preserving.

Schemes for the acoustic equations split up into those that can compute the limit of low Mach numbers, and those who cannot. As has been mentioned in Section 4.1.2, there is no construction principle for the former known that rests on as firm a basis as the Godunov scheme. Dimensionally split schemes that are able to resolve this limit have free parameters that are introduced by an educated guess, but still ad hoc. An a priori unknown stability has to be checked later either theoretically or experimentally. In particular this second constraint is of a practical relevance that can hardly be underestimated. This Section demonstrates that still, there exist dimensionally split schemes for the acoustic equations that both perform well in the limit of low Mach numbers and are stable under explicit time integration.

4.4.1 Stationarity preservation

For the purpose of the illustration of stationarity preserving properties of schemes in this Section ϵ is treated as some finite parameter. Recall Theorem 4.1 that interprets the limit $\epsilon \to 0$ as the limit of long time.

Consider a centered scheme with numerical diffusion for the acoustic system (2.40)–(2.41), which has the general shape

$$\begin{aligned}\partial_t q + \frac{1}{2\Delta x}\Big(J_x(q_{i+1,j} - q_{i-1,j}) - D_x(q_{i+1,j} - 2q_{ij} + q_{i-1,j})\Big) \\ + \frac{1}{2\Delta y}\Big(J_y(q_{i,j+1} - q_{i,j-1}) - D_y(q_{i,j+1} - 2q_{ij} + q_{i,j-1})\Big) = 0\end{aligned} \tag{4.25}$$

A dimensionally split scheme typically has the following general form of the diffusion matrices D_x, D_y:

$$D_x = \begin{pmatrix} a_1 & 0 & a_2 \\ 0 & 0 & 0 \\ a_3 & 0 & a_4 \end{pmatrix} \qquad D_y = \begin{pmatrix} 0 & 0 & 0 \\ 0 & a_1 & a_2 \\ 0 & a_3 & a_4 \end{pmatrix} \tag{4.26}$$

Dimensionally split schemes under certain conditions can be stationarity preserving:

Theorem 4.8 (Stationarity preserving dimensionally split schemes)**.** *The dimensionally split scheme* (4.25) *with* (4.26) *is stationarity preserving if* $a_1 = 0$*. The stationary states fulfill* $p = \text{const}$ *and*

$$\frac{[u]_{i\pm1,j}}{2\Delta x} + \frac{[v]_{i,j\pm1}}{2\Delta y} - \frac{a_3}{c^2}\left(\frac{[[u]]_{i\pm\frac{1}{2},j}}{2\Delta x} + \frac{[[v]]_{i,j\pm\frac{1}{2}}}{2\Delta y}\right) = 0 \tag{4.27}$$

which is a discretization of $\operatorname{div}\mathbf{v} = 0$*.*

Proof. The evolution matrix is easily found to be

$$\mathcal{E} = \mathbb{i}\begin{pmatrix} -\frac{a_1(t_x - 2 + \frac{1}{t_x})}{2\Delta x} & 0 & -\frac{a_2(t_x - 2 + \frac{1}{t_x})}{2\Delta x} + \frac{(t_x - \frac{1}{t_x})}{2\Delta x \epsilon^2} \\ 0 & -\frac{a_1(t_y - 2 + \frac{1}{t_y})}{2\Delta y} & -\frac{a_2(t_y - 2 + \frac{1}{t_y})}{2\Delta y} + \frac{(t_y - \frac{1}{t_y})}{2\Delta y \epsilon^2} \\ -\frac{a_3(t_x - 2 + \frac{1}{t_x})}{2\Delta x} + \frac{c^2(t_x - \frac{1}{t_x})}{2\Delta x} & -\frac{a_3(t_y - 2 + \frac{1}{t_y})}{2\Delta y} + \frac{c^2(t_y - \frac{1}{t_y})}{2\Delta y} & -\frac{a_4(t_x - 2 + \frac{1}{t_x})}{2\Delta x} - \frac{a_4(t_y - 2 + \frac{1}{t_y})}{2\Delta y} \end{pmatrix}$$

whose determinant is only zero (independently of $\mathbf{k}$), if $a_1 = 0$ as can be shown upon direct computation. In this case the corresponding eigenvector is

$$\begin{pmatrix} \frac{a_3(t_y - 2 + t_y^{-1})}{2\Delta y} - \frac{c^2(t_y - t_y^{-1})}{2\Delta y} \\ -\frac{a_3(t_x - 2 + t_x^{-1})}{2\Delta x} + \frac{c^2(t_x - t_x^{-1})}{2\Delta x} \\ 0 \end{pmatrix} \tag{4.28}$$

which amounts, by inverting the Fourier transform, to the given discrete divergence operator and $p = \text{const}$. □

Numerical data that exactly satisfy (4.27) remain unchanged during the evolution (up to machine error). The discrete operator (4.27) is a first order discretization of $\partial_x u + \partial_y v$, if $a_3 \neq 0$. Choosing both $a_1 = 0$ and $a_3 = 0$ in (4.27) and (4.26) makes all the spatial operators reduce to central differences:

Corollary 4.7. *A scheme for the system* (2.40)–(2.41) *whose spatial derivatives are discretized by central differences in two spatial dimensions is stationarity preserving.*

Choosing $a_3 = 0$ means that the discrete divergence operator which is exactly preserved during the time evolution, is a central one. Together with $a_1 = 0$, however, this would mean that there is no diffusion on the velocity variables at all. In practice this is often not desirable as then the scheme will not be stable upon usage of an explicit time integrator (e.g. forward Euler). One might wonder whether there exists a discrete velocity diffusion such that the resulting scheme would keep the central divergence exactly stationary, as this would lead to the stationary states being discretized to higher order. One is thus led to the question of finding a diffusion stencil that vanishes whenever a given divergence stencil does. This problem is solved in Section 3.2 and applied to the acoustic equations in Section 4.5.

In the literature, there already exist several strategies that have been developed in order to cope with the low Mach number problems. They can now be understood in the light of the new arguments that employ the idea of a stationarity preserving scheme. Adapting the matrices to the case of acoustic equations yields the following selection of diffusion matrices:

1. Method from [BEK+17]: $D_x = \begin{pmatrix} 0 & 0 & \frac{1}{\epsilon^2} \\ 0 & 0 & 0 \\ -c^2 & 0 & 0 \end{pmatrix}$
2. Method from [DOR10]: $D_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -c^2 & 0 & \frac{2c}{\epsilon} \end{pmatrix}$
3. The method investigated in Corollary 4.6: $D_x = \begin{pmatrix} 0 & 0 & \frac{1}{\epsilon^2} \\ 0 & 0 & 0 \\ -c^2 & 0 & \frac{2c}{\epsilon} \end{pmatrix}$

Note how all of them have $a_1 = 0$, and are thus stationarity preserving. Experimental results obtained with method no. 3 of the above list are shown in Fig. 4.9–4.7. This method has been found to be stable under explicit time integration in multiple spatial dimensions experimentally. Its one-dimensional stability is studied in Corollary 4.6.

4.4.2 The upwind/Roe scheme

The upwind, or Roe, scheme has $D_x = |J_x|$, $D_y = |J_y|$, with the absolute value being defined on the eigenvalues. This gives

$$D_x = \begin{pmatrix} \frac{c}{\epsilon} & & \\ & 0 & \\ & & \frac{c}{\epsilon} \end{pmatrix} \qquad D_y = \begin{pmatrix} 0 & & \\ & \frac{c}{\epsilon} & \\ & & \frac{c}{\epsilon} \end{pmatrix}$$

which is of the form (4.26), but violates the condition $a_1 = 0$ found in Theorem 4.8.

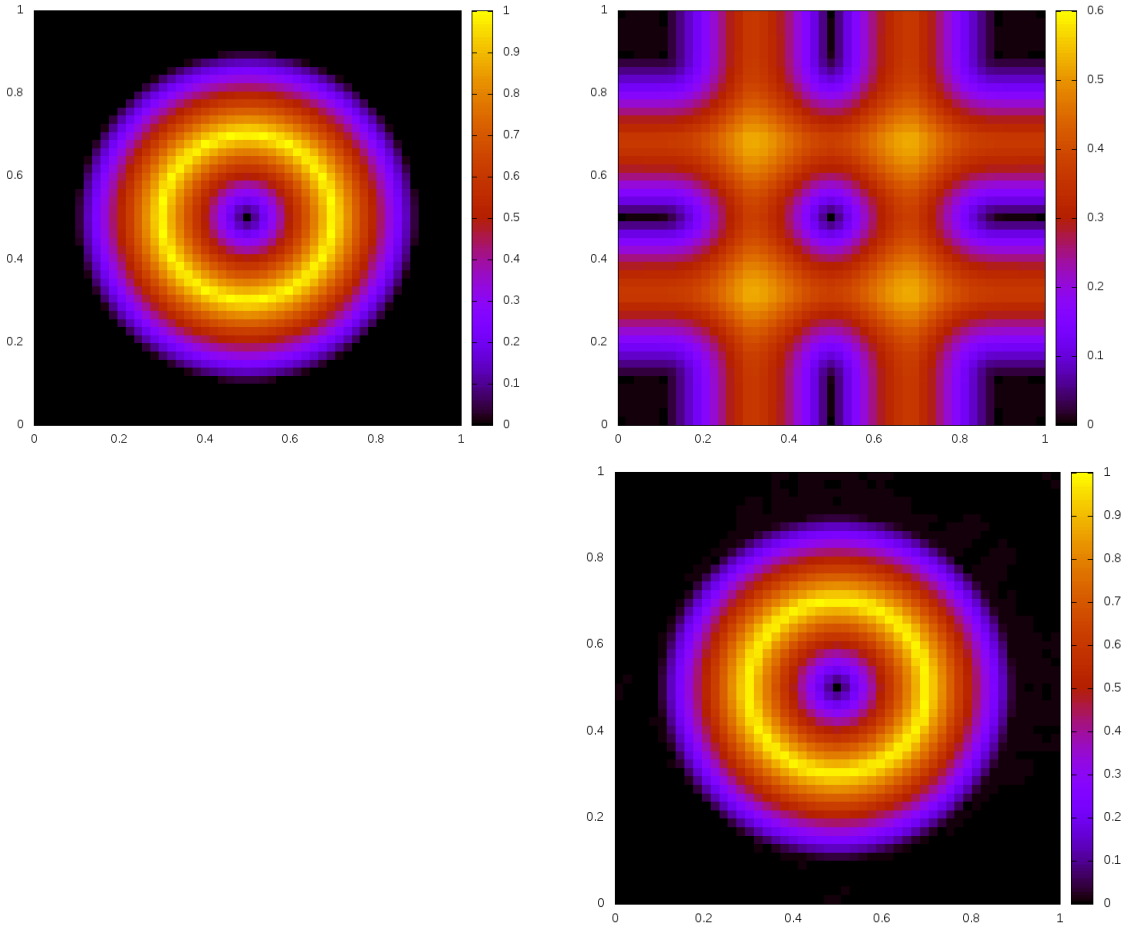

Figure 4.7: *Top left*: Initial setup of a stationary divergence-free vortex. *Top right*: Solution at $t = 10$ with the upwind (Roe) scheme. *Bottom*: Solution at $t = 10$ with solver 3 from the above list. $\sqrt{u^2 + v^2}$ is colour coded; all simulations performed on a 50×50 grid with forward Euler; all methods are of first order in space and time. Observe the improved quality when using a stationarity preserving scheme.

Corollary 4.8. *The Roe scheme for the system* (2.40)–(2.41) *in two spatial dimensions is not stationarity preserving.*

This can be observed in the experiment. Consider the following vortex setup that is similar to the Gresho vortex ([GC90], see also Equations (5.15)–(5.16), page 158). Denoting by $\mathbf{e}_\varphi$ the unit vector in φ-direction and with $r = \sqrt{x^2 + y^2}$ in two spatial dimensions

$$\mathbf{v} = \mathbf{e}_\varphi \cdot \begin{cases} 5r & r < 0.2 \\ 2 - 5r & r < 0.4 \\ 0 & \text{else} \end{cases}$$

$$p = p_c = \text{const}$$

with a constant pressure p_c. As $\mathbf{v}$ is divergenceless, this is a stationary solution of the acoustic equations (2.40)–(2.41).

The numerical time evolution of this setup is shown in Figures 4.8–4.10. From an initial state (that is derived from the analytic stationary solution) one observes the numerical solution to move over to some other stationary solution. Initially, $\partial_x u + \partial_y v = 0$, but $\partial_x u \neq 0$ in general. One observes however in Figure 4.8 that the Roe scheme diffuses away $\partial_x u$ exponentially in time, until it reaches values comparable with machine precision. The initial velocities, shown in the left column of Fig. 4.10, are modified such that after long times only a shear flow is left over. The only states that the scheme is able to keep stationary, are trivial ones. By stability, the scheme is diffusing away all the others. On the other hand, a stationarity preserving scheme would keep stationary also discrete versions of vortical, and in general of all divergenceless flows.

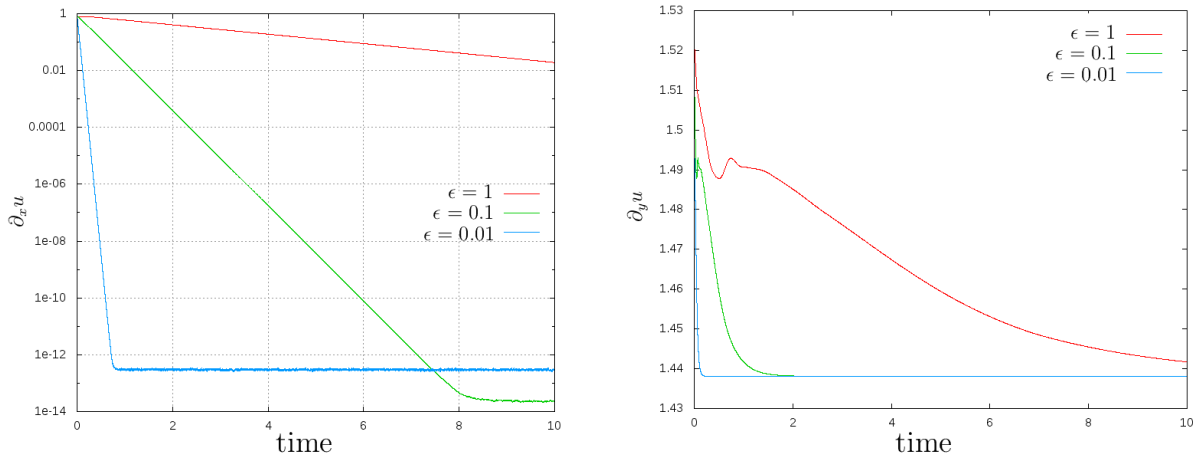

Figure 4.8: *Left*: Decay of $\|\partial_x u\|_{L^1} \sim \exp\left(-\frac{t}{\epsilon}\right)$. *Right*: $\|\partial_y u\|_{L^1}$. Both figures were measured for a stationary vortex setup in a simulation using the Roe solver for the acoustic equations and show curves for values $\epsilon = 1, 0.1, 0.01$. Note the very different vertical axis scalings in the two plots: whereas $\partial_x u \neq 0$ does not comply with stationarity for the Roe scheme, after some transients the scheme settles down on a shear flow ($\partial_y u \neq 0$) that is not significantly different from the initial data.

For the above example the discrete vorticity that is preserved exactly during the time evolution is, by (4.28) (the notation having been introduced on page 13)

$$\frac{[v]_{i\pm1,j}}{2\Delta x} - \frac{[u]_{i,j\pm1}}{2\Delta y} + \frac{a_3}{c^2}\left(\frac{[[u]]_{i,j\pm\frac{1}{2}}}{2\Delta y} - \frac{[[v]]_{i\pm\frac{1}{2},j}}{2\Delta x}\right) = \partial_x v - \partial_y u + \mathcal{O}(\Delta x, \Delta y)$$

In [MR01] the Appendix deals with the preservation of a specific discrete vorticity stencil for a certain family of schemes, and with the production rate in case of non-preservation. However this analysis, as well as the treatment of the acoustic equations in [LFS07], *assume* a vorticity stencil. There might still exist some other vorticity stencil that is exactly preserved. Therefore a more adequate procedure would be to first check

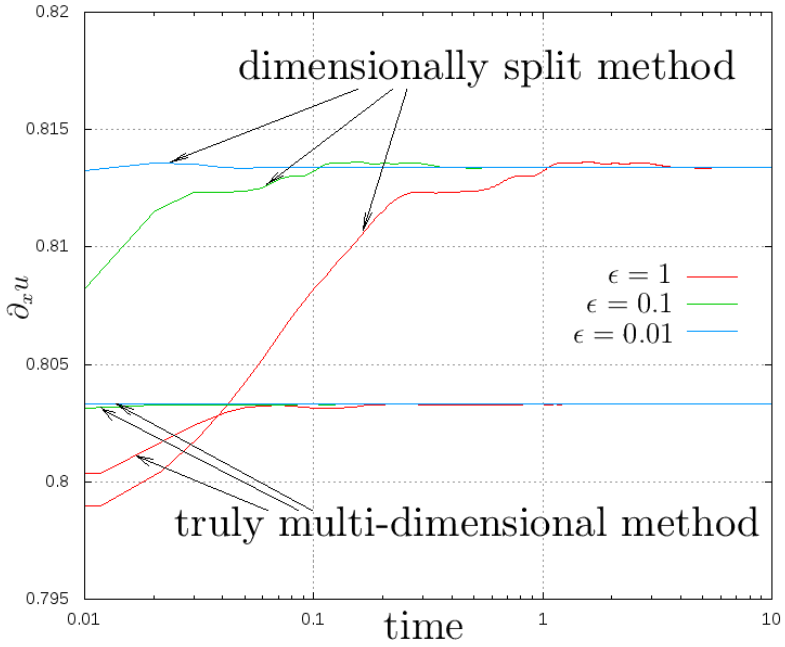

Figure 4.9: Time evolution of $\|\partial_x u\|_{L^1}$. Figure measured for a stationary vortex setup in a simulation using solver 3 mentioned above ("dimensionally split method") and the multidimensional solver ("truly multi-dimensional method") presented in Section 4.5 (Equations (4.32)) for the acoustic equations and show curves for values $\epsilon = 1, 0.1, 0.01$. (Note the scalings of both axes.) Observe the absence of diffusion (contrary to Fig. 4.8, left) and the improved quality of the simulation upon usage of a truly multi-dimensional solver.

(via the determinant of the evolution matrix) the existence of *any* preserved vorticity stencil and to find its shape by evaluating the eigenvector corresponding to the vanishing eigenvalue. Only if the evolution matrix does not contain any vanishing eigenvalues can one claim that there is no preserved discrete vorticity. The condition of Theorem 2 formulated in [LFS07] therefore is sufficient, but not necessary for vorticity preservation.

Next we consider the vortex setup as in Figures 4.8–4.10 for the limit of $\epsilon \to 0$. Again, the upwind/Roe scheme is considered, which has only trivial stationary states. The initial data of the vortex are a discrete version of an analytically stationary solution and have been thus obtained from a divergence-free solution. The upwind/Roe scheme, since it is stable, keeps certain states exactly stationary and diffuses everything else away with time. This diffusion time scales with ϵ, because the non-zero eigenvalues of the evolution matrix scale with $1/\epsilon$. After long time therefore one is left with a numerical stationary state of the scheme.

If the set of numerical stationary states, however, consists only of trivial ones (as it is the case for the Roe scheme), then this numerical solution will have lost all resemblance to the analytic one. In this example the vortex is diffused away and a shear flow left over. Therefore the observed "low Mach number artefacts" are entirely due to the scheme's stationary solutions not being discretizations of all the analytic ones.

The low Mach number limit $\epsilon \to 0$ for the acoustic equations makes the scheme

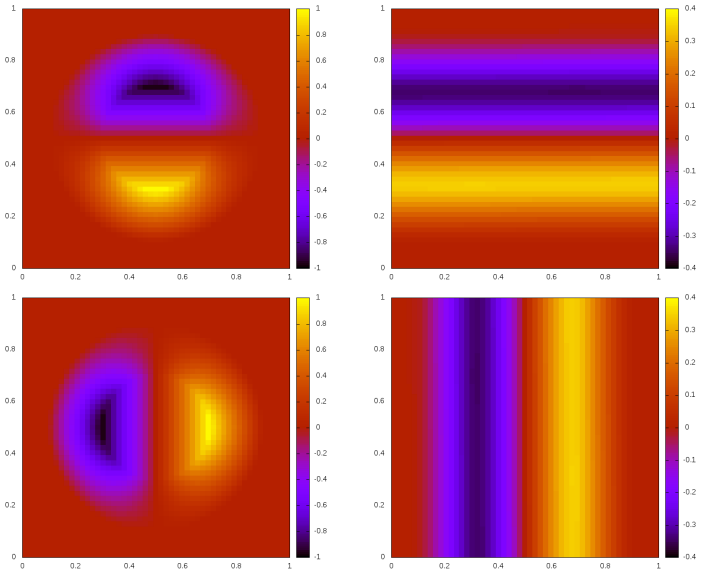

Figure 4.10: Simulation results for $\epsilon = 10^{-3}$ of a vortex setup with the Roe scheme. *Left* are results at time $t = 0$, *right* – at $t = 0.3$. *Top row*: u (colour coded), *bottom row*: v (colour coded). The Roe scheme fails to keep the setup stationary and transitions to a trivial stationary state (shear flow).

attain a numerical stationary state on time scales $\mathcal{O}(\epsilon)$. In order to improve the quality of the numerical solution therefore one needs to choose a scheme which has nontrivial numerical stationary states that capture the rich set of nontrivial stationary states, i.e. a stationarity preserving scheme.

Reducing Δx does not really solve the problem, because then the diffusion time will be longer, but the stationary states will still not be any more similar to the analytic ones. By the equivalence between the low Mach number and the long time limit, this also means that the numerical limit states will not be any closer to the analytic limit states as $\epsilon \to 0$.

4.5 Multi-dimensional schemes

Dimensionally split schemes for the acoustic equations can be stationarity preserving. Do they have any shortcomings? Is it possible to achieve more by studying schemes that cannot be written in a dimensionally split manner? This thesis wants to argue that yes, there exist differences among schemes that are stationarity preserving, and that in order to overcome certain restrictions of the dimensionally split schemes one has to resort to multi-dimensional schemes.

Additionally, it turns out that it is in a sense easier to construct stable multi-dimensional schemes than dimensionally split ones. In Section 4.4 certain entries of the diffusion matrix were modified in order to achieve stationarity preservation, while the scheme still remained dimensionally split. This had strong impact on the stability properties of the scheme. In fact, stable schemes were basically singled out by trial and

error. There is an alternative strategy, which in practice leads to stable schemes straight away.

As result of Theorem 4.8 in Section 4.4.1, in the velocity equation, for a dimensionally split scheme the prefactor of the velocity diffusion has to vanish: denoting the velocity components in 2-d by u and v,

$$\partial_t \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} \frac{1}{2\Delta x}[p]_{i\pm1,j} \\ \frac{1}{2\Delta y}[p]_{i,j\pm1} \end{pmatrix} = a_1 \begin{pmatrix} \frac{1}{2\Delta x}[[u]]_{i\pm\frac{1}{2},j} \\ \frac{1}{2\Delta y}[[v]]_{i,j\pm\frac{1}{2}} \end{pmatrix} + a_2 \begin{pmatrix} \frac{1}{2\Delta x}[[p]]_{i\pm\frac{1}{2},j} \\ \frac{1}{2\Delta y}[[p]]_{i,j\pm\frac{1}{2}} \end{pmatrix}$$

is only part of a stationarity preserving dimensionally split scheme, if $a_1 = 0$. This is, loosely speaking, because a stationary state is characterized by vanishing divergence $\partial_x u + \partial_y v = 0$, which does not constrain in any way the second derivatives $\partial_x^2 u$ or $\partial_y^2 v$. However, if one would be able to make $\partial_x(\partial_x u + \partial_y v)$ out of $\partial_x^2 u$, then these second derivatives would vanish along with the divergence. This would however imply the introduction of a numerical discretization of the mixed derivative $\partial_x \partial_y v$. This is not possible with dimensionally split schemes.

Multi-dimensional schemes that introduce such discretizations have been found to have a certain beauty and also measurable advantages (i.e. in terms of error) over stationarity preserving dimensionally split schemes. The following construction strategy has proven itself very successful. One starts with a one-dimensional stable scheme, e.g. the upwind/Roe scheme. It is first extended to multiple spatial dimensions in a dimensionally split manner. Then one finds terms that have to be added to the velocity diffusion in order to make appear only derivatives of the *divergence*, rather than just derivatives of the different velocity components. How this can be done in the discrete setting is shown in Section 3.2. It is also shown in Theorem 3.6 that such a scheme is stationarity preserving.

The new terms that have to be added to the dimensionally split scheme are all discretizations of mixed derivatives. Therefore these schemes, when applied to a one-dimensional situation, reduce to the one-dimensional scheme that one has started out with. One thus can hope to pass on its stability properties to the multi-dimensional stationarity preserving scheme. Indeed, experimentally the schemes that have been constructed in such way all have turned out to be stable under explicit time integration. Starting with a one-dimensional scheme that is stable under CFL < 1, its dimensionally split extension to two spatial dimensions is only stable under the condition CFL < 0.5. The multi-dimensional extensions presented in this thesis have been found to continue to be stable under CFL < 1 even in multiple spatial dimensions.

It seems that in order to fine-tune the properties of a scheme this construction principle can be fruitfully used in a variety of situations, out of which this thesis presents three. They are discussed in this Section, as well as in the subsequent Sections 4.7 and 4.8, respectively.

4.5.1 Stationarity-consistent divergence

This Section focuses on first order schemes in two spatial dimensions. The extension to higher order is performed in Section 4.7; the extension to higher dimensions can be

performed analogously and is omitted. For first order schemes, the stencils of all the discrete operators involve only the cell itself and its eight neighbours.

Definition 4.7 (Moore stencil)**.** *The* Moore neighbourhood *of a two-dimensional cell* (i, j) *is the cell itself and the 8 cells*

$(i-1, j+1)$	$(i, j+1)$	$(i+1, j+1)$
$(i-1, j)$		$(i+1, j)$
$(i-1, j-1)$	$(i, j-1)$	$(i+1, j-1)$

A Moore stencil *at cell* (i, j) *is a stencil involving only cells from the Moore neighbourhood of cell* (i, j)*.*

The construction idea plays on a discrete counterpart to the statement

$$\partial_x u + \partial_y v = 0 \qquad \Rightarrow \qquad \partial_x^2 u + \partial_x \partial_y v = 0 \tag{4.29}$$

This is the case considered in Section 3.2.2. If Equation (4.29) is true at discrete level for some choice of discretizations, then the resulting scheme is stationarity preserving by Theorem 3.6.

Often in numerics one has to refrain from the wish of obtaining an exact equality, and has to content oneself with *equality up to terms* $\mathcal{O}(\Delta x^r)$. The order $r \in \mathbb{N}$ is then taken as a measure of the quality of approximation. This thinking prevails throughout numerical analysis: indeed, differential operators cannot be, in general, exactly reproduced in the discrete, but only up to some error. This thinking is not sufficient for the purposes of this Section. If (4.29) is true only up to an error, then the scheme will not be stationarity preserving. There is no such thing as *stationarity preserving up to an error*! The diffusion of the multi-dimensional Godunov scheme (4.6)–(4.7), for example, is to first order a derivative of the divergence for certain choices of the CFL number. However, this is not true for all orders (compare Section 3.2.4), and indeed the scheme is not stationarity preserving and does not perform well in the low Mach number limit.

Corollary 4.9. *There is no non-zero Moore stencil of second derivatives of* u *and* v *which is stationarity-consistent with the central divergence.*

Proof. This is a direct consequence of Theorem 3.8. Using the notation established there, for the Moore stencil one has to consider the case $k = 1$. □

Note. This explains why the authors in [JT06], [TF04] found that "the choice of central differences turned out to be not very fruitful".

Modifying the divergence stencil allows to find a stationarity-consistent diffusion.

Corollary 4.10. *The only symmetric divergence discretization on a Moore stencil that allows for a non-zero stationarity-consistent diffusion is*

$$\frac{\{\{[u]_{i\pm1}\}\}_{j\pm\frac{1}{2}}}{8\Delta x} + \frac{[\{\{v\}\}_{i\pm\frac{1}{2}}]_{j\pm1}}{8\Delta y} \tag{4.30}$$

The linear stationarity-consistent stencil of second derivatives associated to the divergence (4.30) *is*

$$\frac{1}{4}c_1\left(\frac{\{\{[[u]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{\Delta x}+\frac{[[v]_{i\pm1}]_{j\pm1}}{\Delta y}\right)+\frac{1}{4}c_2\left(\frac{[[u]_{i\pm1}]_{j\pm1}}{\Delta x}+\frac{[[\{\{v\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}}}{\Delta y}\right) \quad (4.31)$$

with arbitrary parameters c_1, c_2.

Proof. This immediately follows from the example of Section 3.2.2.1. □

This analysis easily generalizes to any number of spatial dimensions.

The divergence stencil (4.30) is – by equivalence of stationarity and vorticity preservation – the "extended operator" in [TF04], [JT06] and has also has been suggested in [MT11] for the system wave equation. The above proof shows that there is actually no other choice among directionally unbiased stencils defined on a 3×3 grid, i.e. among symmetric Moore stencils. In [Sid02] some non-standard finite difference methods are introduced in the context of steady Euler equations. They are reminiscent of the stencils above for the linearized Euler equations. An overview of methods that appear in the literature is presented in Table 4.1.

Source	**Year**	**divergence**	**velocity diffusion**	**keyword**
[LMMW00]	2000	(4.30)	(4.31) (LW)	bicharacteristics
[MR01]	2001	(4.30)	(4.31) (and LW)	vorticity
[Sid02]	2002	(4.30)	(4.31)	*factorizable*
[JT06]	2006	(4.30)	(4.31)	*flux distribution*
[LFS07]	2007	see [MR01]	see [MR01]	vorticity
[MT11]	2011	(4.30)	(4.31)	*potential-based*
[LR14]	2014	unconstrained	(4.31)	vorticity

Table 4.1: Selection of the literature involving stationarity preserving multi-dimensional schemes for the equations (2.7)–(2.8) of linear acoustics. Observe that the discrete divergence and the discrete second derivative of the velocity are the same nearly everywhere. LW denotes the Lax-Wendroff scheme, in the sense that in schemes that are thus marked, the same velocity diffusion is built into a multi-dimensional version of the (second order) Lax-Wendroff scheme. The keyword-column shows that these schemes have been obtained by pursuing very different approaches. Some of the authors have given particular names to these approaches, which are mentioned in the keyword-column. In all cases the reader is referred to the original publications for details.

Note that still it is possible that the schemes that contain this stencil differ – they might treat the pressure differently, or have different order (e.g. compare the scheme in [MR01] to [JT06]).

4.5.2 Construction principles for stationarity preserving multi-dimensional schemes

So far, the focus was lying on the discrete second derivatives of the velocity components. In the previous Section, they have been chosen in such a way that they vanish whenever

the discrete divergence does. Moreover, it has been shown that there is just one choice of this discrete divergence, if one wants to obtain the highest possible order of discretization that is possible on a 3×3 stencil.

How do these stencils assemble into a scheme for the acoustic equations? This Section presents two possible ways.

4.5.2.1 Extension from the upwind/Roe scheme

As has been discussed in the introduction of Section 4.5, modifications of existing schemes that make them stationarity preserving suffer from the challenge not to spoil their stability properties. In the context of dimensionally split schemes this turned out to be very difficult. For multi-dimensional schemes it seems a bit easier. The strategy shall be exemplified here for the case of a first order scheme that is obtained as a multi-dimensional version of the upwind/Roe scheme. Start out from the dimensionally split upwind/Roe scheme for the equations (2.40)–(2.41):

$$\partial_t \begin{pmatrix} u \\ v \\ p \end{pmatrix} + \frac{1}{2\Delta x} \begin{pmatrix} & & \frac{1}{\epsilon^2} \\ & 0 & \\ c^2 & & \end{pmatrix} \begin{bmatrix} u \\ v \\ p \end{bmatrix}_{i\pm1,j} + \frac{1}{2\Delta y} \begin{pmatrix} 0 & & \\ & & \frac{1}{\epsilon^2} \\ & c^2 & \end{pmatrix} \begin{bmatrix} u \\ v \\ p \end{bmatrix}_{i,j\pm1}$$
$$- \frac{1}{2\Delta x} \begin{pmatrix} \frac{c}{\epsilon} & & \\ & 0 & \\ & & \frac{c}{\epsilon} \end{pmatrix} \left[\!\!\left[\begin{bmatrix} u \\ v \\ p \end{bmatrix} \right]\!\!\right]_{i\pm\frac{1}{2},j} - \frac{1}{2\Delta y} \begin{pmatrix} 0 & & \\ & \frac{c}{\epsilon} & \\ & & \frac{c}{\epsilon} \end{pmatrix} \left[\!\!\left[\begin{bmatrix} u \\ v \\ p \end{bmatrix} \right]\!\!\right]_{i,j\pm\frac{1}{2}} = 0$$

or

$$\partial_t \begin{pmatrix} u \\ v \\ p \end{pmatrix} + \begin{pmatrix} \frac{1}{2\Delta x}\frac{1}{\epsilon^2}[p]_{i\pm1,j} \\ \frac{1}{2\Delta y}\frac{1}{\epsilon^2}[p]_{i,j\pm1} \\ c^2\left(\frac{1}{2\Delta x}[u]_{i\pm1,j} + \frac{1}{2\Delta y}[v]_{i,j\pm1}\right) \end{pmatrix} - \frac{c}{\epsilon} \begin{pmatrix} \frac{1}{2\Delta x}[[u]]_{i\pm\frac{1}{2},j} \\ \frac{1}{2\Delta y}[[v]]_{i,j\pm\frac{1}{2}} \\ \frac{1}{2\Delta x}[[p]]_{i\pm\frac{1}{2},j} + \frac{1}{2\Delta y}[[p]]_{i,j\pm\frac{1}{2}} \end{pmatrix} = 0$$

Obvious is the appearance of the central divergence

$$\frac{1}{2\Delta x}[u]_{i\pm1,j} + \frac{1}{2\Delta y}[v]_{i,j\pm1}$$

which has to be replaced by the divergence found in (3.13) (Section 3.2.2.1).

Only now does there a stationarity consistent diffusion exist. Obviously, the term

$$\frac{[[u]]_{i\pm\frac{1}{2},j}}{\Delta x}$$

has to be modified to $\frac{\{\{[[u]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{4\Delta x}$ and augmented by the mixed derivative

$$\frac{[[v]_{i\pm1}]_{j\pm1}}{4\Delta y}$$

The same procedure applies to the second derivative $\frac{[[v]]_{i,j\pm\frac{1}{2}}}{\Delta y}$ which becomes

$$\frac{[[u]_{i\pm1}]_{j\pm1}}{4\Delta x} + \frac{[[\{\{v\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}}}{4\Delta y}$$

The stencils for p are updated to those appearing for the velocity components, i.e.

$$\frac{[p]_{i\pm1,j}}{2\Delta x} \mapsto \frac{\{\{[p]_{i\pm1}\}\}_{j\pm\frac{1}{2}}}{8\Delta x}$$
$$\frac{[[p]]_{i\pm\frac{1}{2},j}}{\Delta x} \mapsto \frac{\{\{[[p]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{4\Delta x}$$

and analogously for the other direction, but no new terms are added. In principle, stationarity preservation does not dictate any conditions on the shape of the discrete derivatives of the pressure.

The scheme becomes

$$\partial_t \begin{pmatrix} u \\ v \\ p \end{pmatrix} + \begin{pmatrix} \frac{1}{8\Delta x}\frac{1}{\epsilon^2}\{\{[p]_{i\pm1}\}\}_{j\pm\frac{1}{2}} \\ \frac{1}{8\Delta y}\frac{1}{\epsilon^2}[\{\{p\}\}_{i\pm\frac{1}{2}}]_{j\pm1} \\ c^2\left(\frac{1}{8\Delta x}\{\{[u]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[\{\{v\}\}_{i\pm\frac{1}{2}}]_{j\pm1}\right) \end{pmatrix} - \frac{c}{\epsilon}\begin{pmatrix} \frac{1}{8\Delta x}\{\{[[u]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[[v]_{i\pm1}]_{j\pm1} \\ \frac{1}{8\Delta x}[[u]_{i\pm1}]_{j\pm1} + \frac{1}{8\Delta y}[[\{\{v\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}} \\ \frac{1}{8\Delta x}\{\{[[p]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[[\{\{p\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}} \end{pmatrix} = 0 \tag{4.32}$$

This scheme is conservative, as it can be rewritten as, with $q = (u, v, p)$

$$\partial_t q + \frac{[f^x]_{i\pm\frac{1}{2},j}}{\Delta x} + \frac{[f^y]_{i,j\pm\frac{1}{2}}}{\Delta y} = 0$$

The flux, in x-direction for example, is then given by

$$(f^x)_{i+\frac{1}{2},j} = \begin{pmatrix} \frac{1}{8}\frac{1}{\epsilon^2}\{\{\{p\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \\ 0 \\ c^2\frac{1}{8}\{\{\{u\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \end{pmatrix} - \frac{c}{\epsilon}\begin{pmatrix} \frac{1}{8}\{\{[u]_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \\ \frac{1}{8}[\{u\}_{i+\frac{1}{2}}]_{j\pm1} \\ \frac{1}{8}\{\{[p]_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \end{pmatrix}$$

This scheme reduces to the upwind (Roe) scheme if restricted to one spatial dimension. Experimentally, it shows stability up to a CFL number of 1 (rather than 0.5 that is found for dimensionally split schemes). It is stationarity preserving by construction. It is similar to the one presented in [JT06, MT11].

4.5.2.2 Pseudo-inverse

The construction strategy of Section 3.2.3 shall be applied to the acoustic equations (2.40)–(2.41). One has

$$J_x = \begin{pmatrix} & 0 & \frac{1}{\epsilon^2} \\ c^2 & & \end{pmatrix} \qquad J_y = \begin{pmatrix} 0 & & \\ & & \frac{1}{\epsilon^2} \\ & c^2 & \end{pmatrix} \tag{4.33}$$

and thus obviously one of the eigenvalues is zero and the matrices are not invertible. This makes some kind of regularization necessary and the Moore-Penrose pseudo-inverse according to Definition 3.10 shall be used for the moment. This gives

For J_x, J_y as in (4.33) one computes

$$\operatorname{sign} J_x \cdot J_y = \begin{pmatrix} 0 & \frac{c}{\epsilon} & \\ & 0 & \\ & & 0 \end{pmatrix} \qquad \operatorname{sign} J_y \cdot J_x = \begin{pmatrix} 0 & & \\ \frac{c}{\epsilon} & 0 & \\ & & 0 \end{pmatrix} \tag{4.34}$$

$$|J_x| = \begin{pmatrix} \frac{c}{\epsilon} & & \\ & 0 & \\ & & \frac{c}{\epsilon} \end{pmatrix} \qquad |J_y| = \begin{pmatrix} 0 & & \\ & \frac{c}{\epsilon} & \\ & & \frac{c}{\epsilon} \end{pmatrix} \tag{4.35}$$

Corollary 4.11. *For the acoustic equations* (2.40)–(2.41)*, the scheme* (3.20) *is the one obtained in Equation* (4.32).

Proof. Using (4.34)–(4.35) one rewrites Equation (3.20) as

$$\begin{aligned} &\partial_t q + \frac{J_x}{8\Delta x}\{\{[q]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + \frac{J_y}{8\Delta y}[\{\{q\}\}_{i\pm\frac{1}{2}}]_{j\pm1} \\ &\quad - \frac{1}{8\Delta x}\frac{c}{\epsilon}\begin{pmatrix} \{\{[[u]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \\ [[u]_{i\pm1}]_{j\pm1} \\ \{\{[[p]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \end{pmatrix} - \frac{1}{8\Delta y}\frac{c}{\epsilon}\begin{pmatrix} [[v]_{i\pm1}]_{j\pm1} \\ [[\{\{v\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}} \\ [[\{\{p\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}} \end{pmatrix} = 0 \end{aligned}$$

This is the numerical scheme that has been obtained in Equation (4.32). □

This justifies the choice of regularization via the Moore-Penrose pseudo-inverse. An approach for nonlinear equations along the same lines is discussed in Section 5.4.2.

4.5.3 Numerical examples

Results of a simulation of a divergence-free vortex with the scheme (4.32) can be seen in Fig. 4.9 and 4.11, and there is evidence for a slight superiority of results obtained with this multi-dimensional scheme as compared to the dimensionally split method presented in Fig. 4.7.

Figure 4.12 shows experimental evidence that it is necessary to use the multi-dimensional operators as they are given in Equation (4.32). This exemplifies that it is not sufficient to find *some* discretizations of $\operatorname{div}\mathbf{v}$ and $\operatorname{grad}\operatorname{div}\mathbf{v}$, but they need to be stationarity consistent. By Corollary 4.10 the only way to obtain a central discretization of the divergence together with a stationarity consistent diffusion is to use multi-dimensional operators.

The multi-dimensional Riemann Problem as discussed in Section 2.2.5 is shown in Figure 4.13, which is to be compared to Figure 2.3 (page 53).

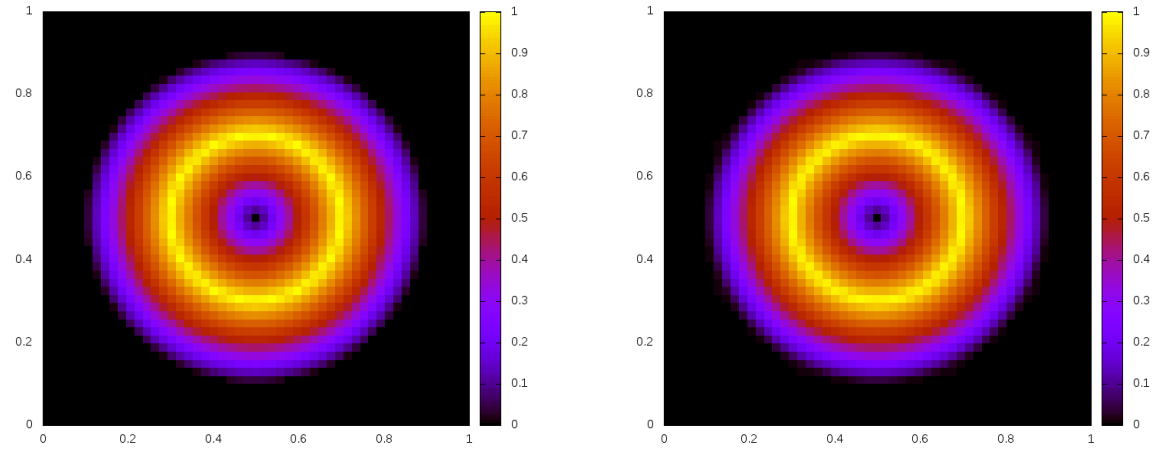

Figure 4.11: *Left*: Initial setup of a stationary divergence-free vortex as in Fig. 4.7. *Right*: Solution at $t = 10$ with the truly multi-dimensional solver (4.32). $\sqrt{u^2 + v^2}$ is colour coded; all simulations performed on a 50×50 grid with forward Euler, therefore the method is of first order in space and time.

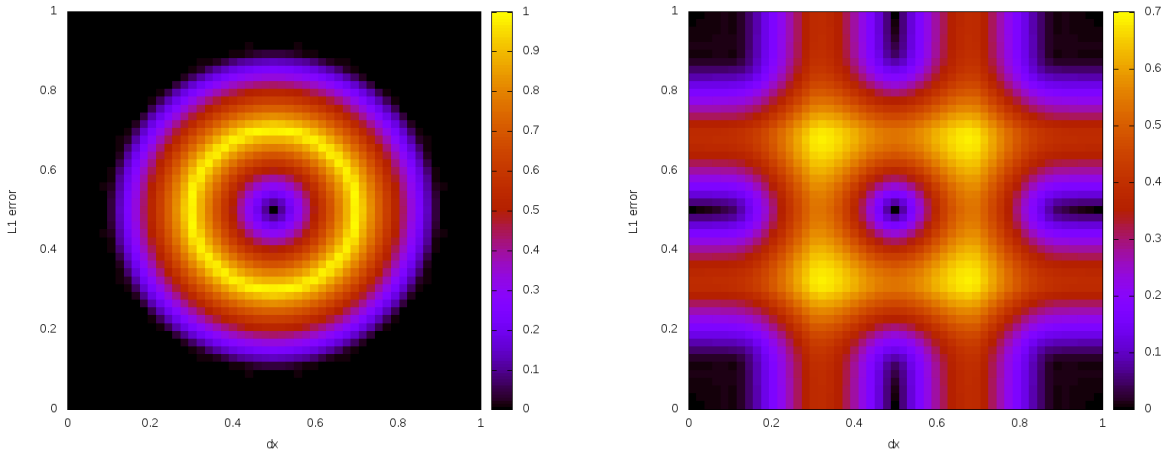

Figure 4.12: Influence of the details of the discretization on stationarity preservation. The setup is that of Figure 4.11 and the result is shown at $t = 500$ (which corresponds to $\epsilon = 0.005$). For results using the Roe scheme see Figure 4.7. Here, in all cases the diffusion remains a discretization of $\operatorname{grad} \operatorname{div} \mathbf{v}$, but the way this discretization is chosen shows drastic effect on stationarity preservation. *Left*: Full implementation according to Equation (4.32). *Right*: Replacing just the multi-dimensional second derivatives $\frac{1}{4}\{\{[[\cdot]]\}\}$ by simple second derivatives $[[\cdot]]$ fails to be stationarity preserving.

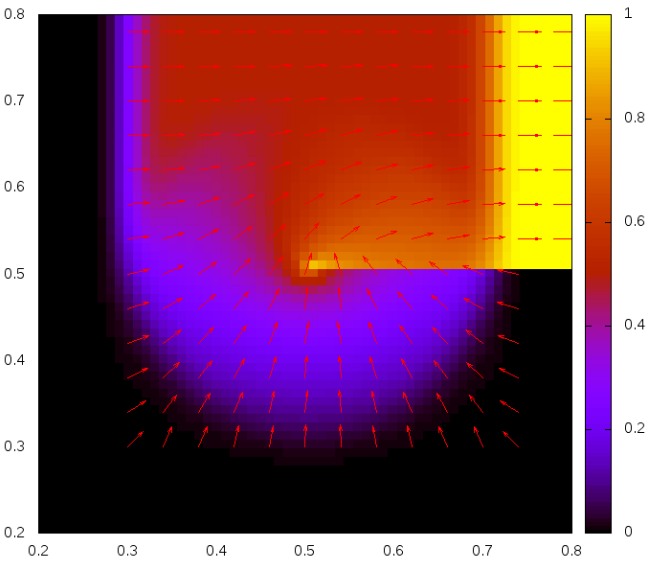

Figure 4.13: Solution of the Riemann Problem discussed in Section 2.2.5. The simulation has been performed with the scheme (4.32) on a 100×100 grid, and the results are shown at $t = 0.2$. The direction of the velocity $\mathbf{v}(t, \mathbf{x})$ is indicated by the arrows, color coded is the absolute value $|\mathbf{v}|$. Compare this to Figure 2.3. Commit hash: 556d8ff (see Appendix for information).

4.6 Asymptotic analysis

Numerical schemes discussed in the previous section are stationarity preserving by construction. They thus are also able to resolve the low Mach number limit by Theorem 4.1. It is instructive though to perform an asymptotic analysis for these multi-dimensional schemes. Indeed they have been constructed starting from the dimensionally split upwind/Roe scheme, which is known to fail in this limit. The question is thus to understand how the multi-dimensional scheme manages to overcome the difficulties of the dimensionally split scheme from a different viewpoint.

Consider an expansion in powers of ϵ of every dependent quantity as in Section 2.3:

$$\begin{aligned} u &= u^{(0)} + \epsilon u^{(1)} + \epsilon^2 u^{(2)} + \mathcal{O}(\epsilon^3) \\ v &= v^{(0)} + \epsilon v^{(1)} + \epsilon^2 v^{(2)} + \mathcal{O}(\epsilon^3) \\ p &= p^{(0)} + \epsilon p^{(1)} + \epsilon^2 p^{(2)} + \mathcal{O}(\epsilon^3) \end{aligned}$$

First, the dimensionally split upwind/Roe scheme is to be analyzed:

$$\begin{aligned} \partial_t \begin{pmatrix} u \\ v \\ p \end{pmatrix} + \begin{pmatrix} \frac{1}{2\Delta x}\frac{1}{\epsilon^2}[p]_{i\pm1,j} \\ \frac{1}{2\Delta y}\frac{1}{\epsilon^2}[p]_{i,j\pm1} \\ c^2\left(\frac{1}{2\Delta x}[u]_{i\pm1,j} + \frac{1}{2\Delta y}[v]_{i,j\pm1}\right) \end{pmatrix} \\ - \frac{c}{\epsilon}\begin{pmatrix} \frac{1}{2\Delta x}[[u]]_{i\pm\frac{1}{2},j} \\ \frac{1}{2\Delta y}[[v]]_{i,j\pm\frac{1}{2}} \\ \frac{1}{2\Delta x}[[p]]_{i\pm\frac{1}{2},j} + \frac{1}{2\Delta y}[[p]]_{i,j\pm\frac{1}{2}} \end{pmatrix} = 0 \end{aligned}$$

Insert the expansions and collect order by order in ϵ:

$$\begin{aligned} &\mathcal{O}(\epsilon^{-2}): \\ 0 &= [p^{(0)}]_{i\pm1,j} \qquad (4.36) \\ 0 &= [p^{(0)}]_{i,j\pm1} \\ &\mathcal{O}(\epsilon^{-1}): \\ 0 &= [p^{(1)}]_{i\pm1,j} - c[[u^{(0)}]]_{i\pm\frac{1}{2},j} \qquad (4.37) \\ 0 &= [p^{(1)}]_{i,j\pm1} - c[[v^{(0)}]]_{i,j\pm\frac{1}{2}} \qquad (4.38) \\ 0 &= \frac{1}{2\Delta x}[[p^{(0)}]]_{i\pm\frac{1}{2},j} + \frac{1}{2\Delta y}[[p^{(0)}]]_{i,j\pm\frac{1}{2}} \\ &\mathcal{O}(\epsilon^{0}): \\ 0 &= \left(\frac{1}{2\Delta x}[u^{(0)}]_{i\pm1,j} + \frac{1}{2\Delta y}[v^{(0)}]_{i,j\pm1}\right) - \frac{1}{c}\left(\frac{1}{2\Delta x}[[p^{(1)}]]_{i\pm\frac{1}{2},j} + \frac{1}{2\Delta y}[[p^{(1)}]]_{i,j\pm\frac{1}{2}}\right) \qquad (4.39) \end{aligned}$$

As usual for open boundaries, $\partial_t p^{(0)} = 0$ is assumed (see e.g. [HJL12]).

Theorem 4.9. *If $t_x \neq -1$, $t_y \neq -1$ Equations (4.36)–(4.39) are only solved by u and v satisfying*

$$[u^{(0)}]_{i\pm\frac{1}{2},j} = 0 \qquad (4.40)$$

$$[v^{(0)}]_{i,j\pm\frac{1}{2}} = 0 \qquad (4.41)$$

which is not a discretization of $\operatorname{div} \mathbf{v}^{(0)} = 0$.

Proof. Consider the Fourier transforms of (4.37)–(4.38)

$$0 = \frac{(t_x-1)(t_x+1)}{t_x}\hat{p}^{(1)} - c\frac{(t_x-1)^2}{t_x}\hat{u}^{(0)}$$
$$0 = \frac{(t_y-1)(t_y+1)}{t_y}\hat{p}^{(1)} - c\frac{(t_y-1)^2}{t_y}\hat{v}^{(0)}$$

Multiplying the first with $\frac{(t_y-1)(t_y+1)}{t_y}$ and the second with $\frac{(t_x-1)(t_x+1)}{t_x}$ and subtracting yields

$$0 = (t_y+1)(t_x-1)\hat{u}^{(0)} - (t_x+1)(t_y-1)\hat{v}^{(0)} \tag{4.42}$$

On the other hand, Equations (4.37)–(4.38) can be combined into

$$0 = \frac{(t_x-1)^2}{2\Delta x t_x}\hat{p}^{(1)} + \frac{(t_y-1)^2}{2\Delta y t_y}\hat{p}^{(1)} - c\left(\frac{(t_x-1)^3}{2\Delta x(t_x+1)t_x}\hat{u}^{(0)} + \frac{(t_y-1)^3}{2\Delta y(t_y+1)t_y}\hat{v}^{(0)}\right)$$

Using the Fourier transform of Equation (4.39)

$$0 = \left(\frac{(t_x+1)(t_x-1)}{2\Delta x t_x}\hat{u}^{(0)} + \frac{(t_y-1)(t_y+1)}{2\Delta y t_y}\hat{v}^{(0)}\right) - \frac{1}{c}\left(\frac{(t_x-1)^2}{2\Delta x t_x}\hat{p}^{(1)} + \frac{(t_y-1)^2}{2\Delta y t_y}\hat{p}^{(1)}\right)$$

one obtains

$$\frac{1}{\Delta x}(t_y+1)(t_x-1)\hat{u}^{(0)} + \frac{1}{\Delta y}(t_x+1)(t_y-1)\hat{v}^{(0)} = 0$$

Using (4.42):

$$(t_y+1)(t_x-1)\hat{u}^{(0)} = 0$$
$$(t_x+1)(t_y-1)\hat{v}^{(0)} = 0$$

i.e. by undoing the Fourier transform

$$[u^{(0)}]_{i\pm\frac{1}{2},j} = 0$$
$$[v^{(0)}]_{i,j\pm\frac{1}{2}} = 0$$

□

Note: This proof essentially shows that the only stationary states of the Roe scheme are trivial shear flows, thus repeating results of Corollary 4.8. However the proof using asymptotic analysis is longer and more involved. One again is left with the result that in the limit, the discrete equations do not discretize *all* of the analytic limit equations.

Now consider analogously the multi-dimensional scheme (4.32). Collecting order by order in ϵ, and assuming as usual $\partial_t p^{(0)} = 0$ again yields:

$$\mathcal{O}(\epsilon^{-2}):$$

$$0 = \{\{[p^{(0)}]_{i\pm1}\}\}_{j\pm\frac{1}{2}} \tag{4.43}$$

$$0 = [\{\{p^{(0)}\}\}_{i\pm\frac{1}{2}}]_{j\pm1} \tag{4.44}$$

$$\mathcal{O}(\epsilon^{-1}):$$

$$0 = \frac{1}{8\Delta x}\{\{[p^{(1)}]_{i\pm1}\}\}_{j\pm\frac{1}{2}} - c\left(\frac{1}{8\Delta x}\{\{[[u^{(0)}]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[[v^{(0)}]_{i\pm1}]_{j\pm1}\right) \tag{4.45}$$

$$0 = \frac{1}{8\Delta y}[\{\{p^{(1)}\}\}_{i\pm\frac{1}{2}}]_{j\pm1} - c\left(\frac{1}{8\Delta x}[[u^{(0)}]_{i\pm1}]_{j\pm1} + \frac{1}{8\Delta y}[[\{\{v^{(0)}\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}}\right)$$

$$0 = \frac{1}{8\Delta x}\{\{[[p^{(0)}]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[[\{\{p^{(0)}\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}} \tag{4.46}$$

$$\mathcal{O}(\epsilon^{0}):$$

$$0 = \frac{1}{8\Delta x}\{\{[u^{(0)}]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[\{\{v^{(0)}\}\}_{i\pm\frac{1}{2}}]_{j\pm1} - \frac{1}{c}\left(\frac{1}{8\Delta x}\{\{[[p^{(1)}]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[[\{\{p^{(1)}\}\}_{i\pm\frac{1}{2}}]]_{j\pm\frac{1}{2}}\right) \tag{4.47}$$

Now the solutions to these equations have to be studied.

Theorem 4.10. *Assuming $t_x \neq -1$, $t_y \neq -1$, the solutions to* (4.43)–(4.47) *fulfill*

$$0 = [p^{(0)}]_{i+\frac{1}{2},j} = [p^{(0)}]_{i,j+\frac{1}{2}} \tag{4.48}$$

$$0 = [p^{(1)}]_{i+\frac{1}{2},j} = [p^{(1)}]_{i,j+\frac{1}{2}} \tag{4.49}$$

$$0 = \frac{1}{8\Delta x}\{\{[u^{(0)}]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[\{\{v^{(0)}\}\}_{i\pm\frac{1}{2}}]_{j\pm1}$$

Proof. From (4.43) follows (4.48) by multiplication with $\frac{1}{(t_x+1)(t_y+1)^2}$, analogously for (4.44) and (4.49). Multiplying (4.43) by $\frac{2}{8\Delta x}\frac{t_x-1}{t_x+1}$, (4.44) by $\frac{2}{8\Delta y}\frac{t_y-1}{t_y+1}$ and adding up yields (4.46). The rest of the proof follows from the stationarity consistency shown in Corollary 4.10. □

In fact, the system (4.43)–(4.47) is not closed: the other $\mathcal{O}(1)$-equations contain $\partial_t u^{(0)}$ and $\partial_t v^{(0)}$. However they also involve $p^{(2)}$. Thus every order in fact couples to higher orders. Recall that in the proof of Theorem 4.9 there appear certain equations that restrict the limit solutions. In order to show that this never can happen for the scheme (4.32) one would thus need to consider all orders of the expansion. This does not seem feasible. Therefore asymptotic analysis either has to restrict its attention to a carefully chosen, but not closed, subsystem of equations, or becomes a very difficult task. Fortunately, the alternative way of proving low Mach compliance for linear acoustics via stationarity preservation is much easier. This line of thought is taken up again in Section 5.2.1.

Also, one observes that it is not sufficient to find the limit equations – one also needs to study their solution space. Proving low Mach compliance via asymptotic analysis is not any simpler than proving stationarity preservation. Observe for instance that the concept of stationarity consistency has been used in the proof. If the divergence

discretization in (4.32) would have been chosen differently, then the diffusion would not be stationarity consistent any longer and the scheme would fail to be stationarity preserving. It would however be very complicated to show this via asymptotic analysis.

Insisting on the discrete limit equations to be consistent discretizations of the analytic limit equations entails a certain danger of wrong argumentation. Indeed, if expanded as power series in Δx and Δy both Equations (4.36)–(4.39) and (4.43)–(4.47) are of the form

$$\begin{aligned} \nabla p^{(1)} &= \mathcal{O}(\Delta x, \Delta y) \\ \operatorname{div} \mathbf{v}^{(0)} &= \mathcal{O}(\Delta x, \Delta y) \end{aligned}$$

and seem to be consistent discretizations of the limit equations of the PDE. This is true! However, in one of the cases (dimensionally split Roe/upwind scheme (4.36)–(4.39)), the discrete solutions turn out to *additionally* fulfill (4.40)–(4.41). It is these additional constraints that reduce the space of limit solutions. In view of stationarity preservation this is clear as well: the whole theory of stationarity preservation is based on the observation that certain methods do not keep stationary a discretization of *all* the analytic stationary states. Therefore a phrasing of a definition that avoids this confusion is

Definition 4.8 (Asymptotic preserving)**.** *A numerical scheme is called* asymptotic preserving *if the solutions to its discrete limit equations (obtained via formal asymptotic analysis) are discretizations of all solutions to the limit PDE.*

Observe that this definition leaves much less room for interpretations than that found in [Jin99]. Inspired by results for other equations and other limits one might expect from the definition given in [Jin99] that a scheme that is not asymptotic preserving would violate the limit equations. For the limit of low Mach number this is not the case. A scheme that is not stationarity-preserving discretizes only a *subset* of the limit equations. This reduction of the solution space is the essence of artefacts at low Mach number, not any hypothetical violation of the limit equations. This also explains the particular phrasing of Theorem 4.2.

There is another reasoning that might seem attractive but does not work; it shall be briefly described now. Finding the discrete solutions may be very difficult. In the above examples the Fourier transform was of great help, but it might not be available in other circumstances. One might wonder whether studying the leading order diffusion, or in general just the modified equation of the discrete limit equations might be sufficient. This is not the case! Asymptotic analysis of numerical schemes in the low Mach number limit has to involve the fully discrete spatial operators. This is because the same is true for the concept of stationarity preservation: There is no way to decide whether a numerical scheme is stationarity preserving by just looking at the modified equation (or the leading order diffusion) of the scheme, as stationarity preservation (see Section 3.2.4) is involving all orders of an expansion in Δx, Δy. Things might seem to work out at first order, but they might fail at the next order. In view of the complexity of such expansions (exemplified in Section 3.2.4) studying ever higher orders does not seem viable.

4.7 Stationarity preserving schemes of higher order

4.7.1 Second order Godunov schemes for linear advection in one spatial dimension

Theorem 4.11. *For the linear advection $\partial_t q + c\partial_x q = 0$ consider the reconstruction $\tilde{q}(x)$ in cell i given by*

$$\tilde{q}(x) = q_i + \frac{x - x_i}{\Delta x}\sigma_i$$

with the slope σ_i any function of the neighbouring values of q. Then the Godunov scheme reads

$$\frac{q_i^{n+1} - q_i^n}{\Delta t} + \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{\Delta x} = 0$$

with the numerical flux

$$\begin{aligned} f_{i+\frac{1}{2}} = \frac{c}{2}\left(q_i + q_{i+1} + \frac{1}{2}(\sigma_i - \sigma_{i+1})\right) - \frac{|c|}{2}\left(q_{i+1} - q_i - \frac{1}{2}(\sigma_i + \sigma_{i+1})\right) \\ + \frac{c\Delta t}{4\Delta x}\Big(- c(\sigma_i + \sigma_{i+1}) + |c|(\sigma_{i+1} - \sigma_i)\Big) \end{aligned}$$

Proof. Consider the case $c > 0$ first and compute the flux $f_{i+\frac{1}{2}}$. Then the reconstruction reads

$$\tilde{q}(x) = q_i + \frac{x - x_i}{\Delta x}\sigma_i$$

and

$$f^+_{i+\frac{1}{2}} = c\frac{1}{\Delta t}\int_0^{\Delta t} \mathrm{d}t\, \tilde{q}(x_{i+\frac{1}{2}} - ct) = c\left(q_i + \frac{\sigma_i}{2\Delta x}(\Delta x - c\Delta t)\right)$$

Otherwise, if $c < 0$, then

$$\begin{aligned} \tilde{q}(x) &= q_{i+1} + \frac{x - x_{i+1}}{\Delta x}\sigma_{i+1} \\ f^-_{i+\frac{1}{2}} &= c\frac{1}{\Delta t}\int_0^{\Delta t} \mathrm{d}t\, \tilde{q}(x_{i+\frac{1}{2}} - ct) = c\left(q_{i+1} - \frac{\sigma_{i+1}}{2\Delta x}(\Delta x + c\Delta t)\right) \end{aligned}$$

The complete flux is given by

$$\begin{aligned} f_{i+\frac{1}{2}} &= \frac{c + |c|}{2} f^+_{i+\frac{1}{2}} + \frac{c - |c|}{2} f^-_{i+\frac{1}{2}} \\ &= \frac{c + |c|}{2}\left(q_i + \frac{\sigma_i}{2\Delta x}(\Delta x - c\Delta t)\right) + \frac{c - |c|}{2}\left(q_{i+1} - \frac{\sigma_{i+1}}{2\Delta x}(\Delta x + c\Delta t)\right) \\ &= \frac{c}{2}\left(q_i + q_{i+1} + \frac{1}{2}(\sigma_i - \sigma_{i+1})\right) - \frac{|c|}{2}\left(q_{i+1} - q_i - \frac{1}{2}(\sigma_i + \sigma_{i+1})\right) \\ &\quad + \frac{c\Delta t}{4\Delta x}\Big(- c(\sigma_i + \sigma_{i+1}) + |c|(\sigma_{i+1} - \sigma_i)\Big) \end{aligned}$$

□

Theorem 4.12. *Inserting the Fromm method in the scheme of Theorem 4.11*

$$\sigma_i = \frac{q_{i+1} - q_{i-1}}{2}$$

yields a second order scheme.

Proof.

$$\begin{aligned} f_{i+\frac{1}{2}} &= \frac{c}{8}\left((-q_{i-1} + 5q_i + 5q_{i+1} - q_{i+2}) + \frac{c\Delta t}{\Delta x}(q_{i-1} + q_i - q_{i+1} - q_{i+2})\right) \\ &\quad - \frac{|c|}{8}\left((q_{i-1} - 3q_i + 3q_{i+1} - q_{i+2}) + \frac{c\Delta t}{\Delta x}(-q_{i-1} + q_i + q_{i+1} - q_{i+2})\right) \\ &= \frac{c}{8}(-q_{i-1} + 5q_i + 5q_{i+1} - q_{i+2}) - \frac{|c|}{8}(q_{i-1} - 3q_i + 3q_{i+1} - q_{i+2}) \\ &\quad + \frac{c}{8}\frac{c\Delta t}{\Delta x}(q_{i-1} + q_i - q_{i+1} - q_{i+2}) - \frac{|c|}{8}\frac{c\Delta t}{\Delta x}(-q_{i-1} + q_i + q_{i+1} - q_{i+2}) \end{aligned}$$

For the flux difference one obtains

$$\begin{aligned} \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{\Delta x} &= \frac{c}{8\Delta x}(q_{i-2} - 6q_{i-1} + 6q_{i+1} - q_{i+2}) + \frac{|c|}{8\Delta x}(q_{i-2} - 4q_{i-1} + 6q_i - 4q_{i+1} + q_{i+2}) \\ &\quad + \frac{\Delta t c}{\Delta x}\Big(\frac{c}{8\Delta x}(-q_{i-2} + 2q_i - q_{i+2}) + \frac{|c|}{8\Delta x}(-q_{i-2} + 2q_{i-1} - 2q_{i+1} + q_{i+2})\Big) \qquad (4.50) \\ &\simeq \left(c\partial_x q - \frac{1}{12}c\partial_x^3 q\Delta x^2\right) + \frac{1}{8}|c|\partial_x^4 q\Delta x^3 \\ &\quad + \frac{\Delta t c}{\Delta x}\left(-\frac{c}{2}\Delta x\left(\partial_x^2 q + \frac{1}{3}\partial_x^4 q\Delta x^2\right) + \frac{|c|}{2}\Delta x^2\left(\frac{1}{2}\partial_x^3 q + \frac{1}{3}\partial_x^5 q\Delta x^2\right)\right) + \mathcal{O}(\Delta x^4) \end{aligned}$$

whereas

$$\frac{q_i^{n+1} - q_i^n}{\Delta t} = \partial_t q + \frac{1}{2}\Delta t\partial_t^2 q + \frac{1}{6}\Delta t^2\partial_t^3 q + \frac{1}{24}\Delta t^3\partial_t^4 q + \mathcal{O}(\Delta t^4)$$

and using the equation $\partial_t q = -c\partial_x q$

$$= -c\partial_x q + \frac{1}{2}\Delta t c^2\partial_x^2 q - \frac{1}{6}\Delta t^2 c^3\partial_x^3 q + \frac{1}{24}\Delta t^3 c^4\partial_x^4 q + \mathcal{O}(\Delta t^4)$$

Adding up this gives

$$\frac{q_i^{n+1} - q_i^n}{\Delta t} + \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{\Delta x} = \frac{1}{4}\Delta x^2\left(\frac{\Delta t|c|}{\Delta x} - \frac{1}{3}\right)c\partial_x^3 q - \frac{1}{6}\Delta t^2 c^3\partial_x^3 q + \mathcal{O}(\Delta x^3, \Delta t^3)$$

which is the diffusion of a second-order scheme. □

One of the stencils of this scheme is used as an example in Section 3.2.2.2.
There are several details to note:

- The scheme only becomes second order if all terms $\mathcal{O}(\Delta t, \Delta x)$ cancel out. The time derivative $\frac{q_i^{n+1}-q_i^n}{\Delta t}$ contains the contribution $\frac{1}{2}\Delta t c^2 \partial_x^2 q$. In order for it to cancel, the spatial derivatives have to contain a term $-\frac{c\Delta t}{\Delta x} \cdot \frac{1}{2}\Delta x c \partial_x^2 q$, which is the case in the above example.

- On the given stencil, the highest order approximation to a first derivative is

$$\frac{q_{i-2} - 8q_{i-1} + 8q_{i+1} - q_{i+2}}{12\Delta x} = \partial_x q + \mathcal{O}(\Delta x^5) \tag{4.51}$$

  However in the scheme above, this derivative is discretized by

$$\frac{q_{i-2} - 6q_{i-1} + 6q_{i+1} - q_{i+2}}{8\Delta x} = \partial_x q - \frac{1}{12}\partial_x^3 q \Delta x^2 + \mathcal{O}(\Delta x^3)$$

  i.e. it has the same order as the overall scheme.

  For first order schemes the situation is different: they can be written as a central difference (which discretizes ∂_x and is second order) and a diffusion with an error of first order. The multi-dimensional scheme in Equation (4.32) changed the diffusion such that the central, i.e. the high order differences discretize the stationary states. They are thus discretized to a higher order than the actual order of the scheme. An analogous "improved order" for stationary states can be achieved for the second order scheme, if the derivative ∂_x is discretized by, say, the highest order stencil (4.51). This can be achieved by choosing the slope σ_i in a different way.

In general, the reconstruction can depend on the sign of c.

Theorem 4.13. *With the notation of Theorem 4.11 and $\xi, \xi' \in \mathbb{R}$, choosing*

$$\sigma_i = \begin{cases} -\frac{1}{2}(1+\xi)q_{i+1} + \xi q_i + \frac{1}{2}(1-\xi)q_{i-1} & c > 0 \\ -\frac{1}{2}(1+\xi')q_{i+1} + \xi' q_i + \frac{1}{2}(1-\xi')q_{i-1} & c < 0 \end{cases}$$

yields a second order scheme, which uses (4.51) *to discretize ∂_x if*

$$\xi = -\xi' = -\frac{1}{3}$$

Proof. The flux difference $\frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{\Delta x}$ contains the term

$$\frac{c}{8\Delta x}\Big((1+\xi)q_{i-2} + (-6-3\xi-\xi')q_{i-1} + (3\xi+3\xi')q_i + (6-\xi-3\xi')q_{i+1} + (\xi'-1)q_{i+2}\Big)$$

$$= c\partial_x q - \frac{1}{24}c\partial_x^3 q \Delta x^2 \left(2 + 3\xi - 3\xi'\right) + \frac{1}{16}c\partial_x^4 q \Delta x^3 (\xi + \xi') + \mathcal{O}(\Delta x^4)$$

This becomes the high order approximation (4.51), if

$$\xi = -\xi' = -\frac{1}{3}$$

One obtains the flux

$$\begin{aligned} f_{i+\frac{1}{2}} = {} & c\frac{-q_{i-1} + 7q_i + 7q_{i+1} - q_{i+2}}{12} - |c|\frac{q_{i-1} - 3q_i + 3q_{i+1} - q_{i+2}}{12} \\ & + \frac{c\Delta t}{\Delta x}\left(c\frac{q_{i-1} + 3q_i - 3q_{i+1} - q_{i+2}}{12} - |c|\frac{-q_{i-1} + q_i + q_{i+1} - q_{i+2}}{12}\right) \end{aligned}$$

and the flux difference

$$\begin{aligned} \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{\Delta x} = {} & c\frac{q_{i-2} - 8q_{i-1} + 8q_{i+1} - q_{i+2}}{12\Delta x} - |c|\frac{-q_{i-2} + 4q_{i-1} - 6q_i + 4q_{i+1} - q_{i+2}}{12\Delta x} \\ & + \frac{c\Delta t}{\Delta x}\left(c\frac{-q_{i-2} - 2q_{i-1} + 6q_i - 2q_{i+1} - q_{i+2}}{12\Delta x} - |c|\frac{q_{i-2} - 2q_{i-1} + 2q_{i+1} - q_{i+2}}{12\Delta x}\right) \end{aligned} \tag{4.52}$$

$$\begin{aligned} = {} & c\left(\partial_x q + \mathcal{O}(\Delta x^4)\right) - |c|\left(-\frac{1}{12}\partial_x^4 q\Delta x^3 + \mathcal{O}(\Delta x^5)\right) \\ & + \frac{c\Delta t}{\Delta x}\left(c\left(-\frac{1}{2}\partial_x^2 q\Delta x - \frac{1}{8}\partial_x^4 q\Delta x^3 + \mathcal{O}(\Delta x^5)\right) - |c|\left(-\frac{1}{6}\partial_x^3 q\Delta x^2 + \mathcal{O}(\Delta x^4)\right)\right) \end{aligned} \tag{4.53}$$

□

4.7.2 Extension to the acoustic system

As has been said, contrary to (4.50), (4.52) can be extended to a stationarity preserving scheme for linear acoustics in such a way that stationary states will be discretized to higher order than the overall order of the scheme. Therefore the focus lies on (4.52) in the following.

The strategy is to first obtain a higher order scheme for linear acoustics in one-dimension (Section 4.7.2.1). Then this scheme is extended to multiple spatial dimensions in a dimensionally split way. The deficiencies of such an approach are discussed in Section 4.7.2.2 and a stationarity preserving extension to multiple spatial dimensions presented.

4.7.2.1 Higher order schemes in one spatial dimension

A strategy to extend the higher order scheme (4.53) to *systems* in one spatial dimension is to diagonalize it and repeat the reasoning for every component. For the general linear system

$$\partial_t q + J\partial_x q = 0 \qquad\qquad q : \mathbb{R}_0^+ \times \mathbb{R} \to \mathbb{R}^n$$

this amounts to formally replacing c by J, and $|c|$ by $|J|$:

$$\begin{aligned}
&\frac{q_i^{n+1}-q_i^n}{\Delta t}+J\frac{q_{i-2}-8q_{i-1}+8q_{i+1}-q_{i+2}}{12\Delta x}-|J|\frac{-q_{i-2}+4q_{i-1}-6q_i+4q_{i+1}-q_{i+2}}{12\Delta x}\\
&\quad+\frac{J\Delta t}{\Delta x}\left(J\frac{-q_{i-2}-2q_{i-1}+6q_i-2q_{i+1}-q_{i+2}}{12\Delta x}-|J|\frac{q_{i-2}-2q_{i-1}+2q_{i+1}-q_{i+2}}{12\Delta x}\right) \qquad (4.54)\\
&\quad=J\left(\partial_x q+\mathcal{O}(\Delta x^4)\right)-|J|\left(-\frac{1}{12}\partial_x^4 q\Delta x^3+\mathcal{O}(\Delta x^5)\right)\\
&\quad+\frac{J\Delta t}{\Delta x}\left(J\left(-\frac{1}{2}\partial_x^2 q\Delta x-\frac{1}{8}\partial_x^4 q\Delta x^3+\mathcal{O}(\Delta x^5)\right)-|J|\left(-\frac{1}{6}\partial_x^3 q\Delta x^2+\mathcal{O}(\Delta x^4)\right)\right)
\end{aligned}$$

The multi-dimensional extension is to be constructed in such a way that it reduces to (4.54) when applied to a one-dimensional situation.

For the acoustic system take

$$J=\begin{pmatrix}0 & \frac{1}{\epsilon^2}\\ c^2 & \end{pmatrix}\qquad\qquad |J|=\begin{pmatrix}\frac{c}{\epsilon} & \\ & \frac{c}{\epsilon}\end{pmatrix}$$

and $q=(u,v,p)$.

Expanding again in powers of Δx gives

$$\begin{aligned}
&\frac{1}{\Delta t}\begin{pmatrix}u_{ij}^{n+1}-u_{ij}^n\\ v_{ij}^{n+1}-v_{ij}^n\\ p_{ij}^{n+1}-p_{ij}^n\end{pmatrix}+\left[\begin{pmatrix}\frac{1}{\epsilon^2}\partial_x p\\ 0\\ c^2\partial_x u\end{pmatrix}_{+\mathcal{O}(\Delta x^4)}\right]+\frac{\Delta x^3}{12}\frac{c}{\epsilon}\left[\begin{pmatrix}\partial_x^4 u\\ 0\\ \partial_x^4 p\end{pmatrix}_{+\mathcal{O}(\Delta x^2)}\right]\\
&\quad+\frac{\Delta t}{\Delta x}\left(-\frac{\Delta x}{2}\frac{c^2}{\epsilon^2}\left[\begin{pmatrix}\partial_x^2 u\\ 0\\ \partial_x^2 p\end{pmatrix}_{+\mathcal{O}(\Delta x^2)}\right]+\frac{\Delta x^2}{6}\frac{c}{\epsilon}\left[\begin{pmatrix}\frac{1}{\epsilon^2}\partial_x^3 p\\ 0\\ c^2\partial_x^3 u\end{pmatrix}_{+\mathcal{O}(\Delta x^2)}\right]\right)=0 \qquad (4.55)
\end{aligned}$$

where from each of the four terms of the flux difference only the highest powers have been kept. Note that the different terms contain different powers of Δx, Δt. However, only the highest term of the expansion has been kept. This might seem inconsistent, if one would aim at collecting terms of equal powers in Δx. This is not the purpose, though. The highest order terms are kept as reminiscences of the discrete operators appearing in (4.54) in order to simplify notation. Just as in 3.2.1.1 and 4.5.2.2, the strategy is first to determine the correct multi-dimensional extension of the higher derivatives at continuous level, and then to determine their stationarity consistent discretizations. Here one will go back to the discrete operators appearing in (4.54) in order to make sure that the multi-dimensional extensions reduce to the operators of (4.54) when the scheme is applied to a one-dimensional situation.

4.7.2.2 Higher order schemes in multiple spatial dimensions

The subsequent extension to multiple spatial dimensions can in principle be performed in a dimensionally split manner, although this will not yet lead to a stationarity preserving scheme.

Consider the linear system

$$\partial_t q + J_x \partial_x q + J_y \partial_y q = 0$$

with, for linear acoustics

$$J_x = \begin{pmatrix} & \frac{1}{\epsilon^2} \\ c^2 & \end{pmatrix} \qquad J_y = \begin{pmatrix} 0 & \\ & \frac{1}{\epsilon^2} \\ c^2 & \end{pmatrix}$$

Then the dimensionally split scheme according to Definition 4.6 derived from Equation (4.55) reads

$$\begin{aligned}
&\frac{q_{ij}^{n+1} - q_{ij}^n}{\Delta t} + J_x \frac{q_{i-2,j} - 8q_{i-1,j} + 8q_{i+1,j} - q_{i+2,j}}{12\Delta x} - |J_x| \frac{-q_{i-2,j} + 4q_{i-1,j} - 6q_{ij} + 4q_{i+1,j} - q_{i+2,j}}{12\Delta x} \\
&+ \frac{J_x \Delta t}{\Delta x} \left(J_x \frac{-q_{i-2,j} - 2q_{i-1,j} + 6q_{ij} - 2q_{i+1,j} - q_{i+2,j}}{12\Delta x} - |J_x| \frac{q_{i-2,j} - 2q_{i-1,j} + 2q_{i+1,j} - q_{i+2,j}}{12\Delta x} \right) \\
&+ J_y \frac{q_{i,j-2} - 8q_{i,j-1} + 8q_{i,j+1} - q_{i,j+2}}{12\Delta y} - |J_y| \frac{-q_{i,j-2} + 4q_{i,j-1} - 6q_{ij} + 4q_{i,j+1} - q_{i,j+2}}{12\Delta y} \\
&+ \frac{J_y \Delta t}{\Delta y} \left(J_y \frac{-q_{i,j-2} - 2q_{i,j-1} + 6q_{ij} - 2q_{i,j+1} - q_{i,j+2}}{12\Delta y} - |J_y| \frac{q_{i,j-2} - 2q_{i,j-1} + 2q_{i,j+1} - q_{i,j+2}}{12\Delta y} \right)
\end{aligned} \tag{4.56}$$

In the following $\Delta x = \Delta y$ is assumed on several occasions to simplify life; in principle it is possible to keep track of their respective occurrences.

Keeping only the highest powers of every term the dimensionally split scheme reads:

$$\begin{aligned}
&\frac{1}{\Delta t} \begin{pmatrix} u_{ij}^{n+1} - u_{ij}^n \\ v_{ij}^{n+1} - v_{ij}^n \\ p_{ij}^{n+1} - p_{ij}^n \end{pmatrix} + \begin{pmatrix} \frac{1}{\epsilon^2} \partial_x p \\ \frac{1}{\epsilon^2} \partial_y p \\ c^2(\partial_x u + \partial_y v) \end{pmatrix} + \frac{\Delta x^3}{12} \frac{c}{\epsilon} \begin{pmatrix} \partial_x^4 u \\ \partial_y^4 v \\ \partial_x^4 p + \partial_y^4 p \end{pmatrix} \\
&\qquad + \frac{\Delta t}{\Delta x} \left(-\frac{\Delta x}{2} \frac{c^2}{\epsilon^2} \begin{pmatrix} \partial_x^2 u \\ \partial_y^2 v \\ \partial_x^2 p + \partial_y^2 p \end{pmatrix} + \frac{\Delta x^2}{6} \frac{c}{\epsilon} \begin{pmatrix} \frac{1}{\epsilon^2} \partial_x^3 p \\ \partial_y^3 p \\ c^2(\partial_x^3 u + \partial_y^3 v) \end{pmatrix} \right) = 0
\end{aligned} \tag{4.57}$$

Extending the one-dimensional scheme (4.55) to a two-dimensional one in a dimensionally split manner turns out to spoil its order:

Theorem 4.14. *The numerical scheme* (4.57) *is not of second order in space and time.*

Proof. The only term that is $\mathcal{O}(\Delta t)$ is

$$-\frac{1}{2} \Delta t \frac{c^2}{\epsilon^2} \begin{pmatrix} \partial_x^2 u \\ \partial_y^2 v \\ \partial_x^2 p + \partial_y^2 p \end{pmatrix}$$

In 1-d, this term is precisely canceled by a term contained in the discrete time derivative. However, in multi-d

$$\begin{aligned}\frac{1}{\Delta t}\begin{pmatrix} u_{ij}^{n+1}-u_{ij}^{n} \\ v_{ij}^{n+1}-v_{ij}^{n} \\ p_{ij}^{n+1}-p_{ij}^{n}\end{pmatrix} &= \begin{pmatrix} \partial_t u+\frac{1}{2}\Delta t\partial_t^2 u \\ \partial_t v+\frac{1}{2}\Delta t\partial_t^2 v \\ \partial_t p+\frac{1}{2}\Delta t\partial_t^2 p\end{pmatrix}+\mathcal{O}(\Delta t^2) \\ &= \begin{pmatrix}\partial_t u \\ \partial_t v \\ \partial_t p\end{pmatrix}+\frac{1}{2}\Delta t\frac{c^2}{\epsilon^2}\begin{pmatrix}\partial_x(\partial_x u+\partial_y v) \\ \partial_y(\partial_x u+\partial_y v) \\ \partial_x^2 p+\partial_y^2 p\end{pmatrix}+\mathcal{O}(\Delta t^2) \quad (4.58)\end{aligned}$$

□

This can be generalized to all linear systems:

Theorem 4.15. *Consider the linear system*

$$\partial_t q + J_x\partial_x q + J_y\partial_y = 0$$

In order for a linear scheme to achieve second order accuracy the spatial derivatives need to cancel

$$\frac{1}{2}\Delta t\{J_x, J_y\}\partial_x\partial_y q$$

which comes from the discretization of the time derivative.

Proof. For this system, the calculation in (4.58) can be repeated:

$$\begin{aligned}\frac{q_{ij}^{n+1}-q_{ij}^{n}}{\Delta t} &= \partial_t q+\frac{1}{2}\Delta t\partial_t^2 q+\mathcal{O}(\Delta t^2) \\ &= \partial_t q+\frac{1}{2}\Delta t(J_x\partial_x(J_x\partial_x q+J_y\partial_y q)+J_y\partial_y(J_x\partial_x q+J_y\partial_y q))+\mathcal{O}(\Delta t^2) \\ &= \partial_t q+\frac{1}{2}\Delta t\Big(J_x^2\partial_x^2 q+\{J_x,J_y\}\partial_x\partial_y q+J_y^2\partial_y^2 q\Big)+\mathcal{O}(\Delta t^2)\end{aligned}$$

with the anti-commutator $\{J_x, J_y\} = J_xJ_y + J_yJ_x$. Thus, the term that cannot be canceled in a dimensionally split framework, generally, is

$$\frac{1}{2}\Delta t\{J_x, J_y\}\partial_x\partial_y q$$

□

The strategy of replacing $\partial_x^2 u$ by $\partial_x(\partial_x u+\partial_y v)$ in previous sections has been motivated by arguments of stationarity preservation. Here already for the purpose of the desired accuracy one needs a multi-dimensional contribution. [LR14] list several attempts of an extension of the (second order) Lax-Wendroff scheme to multiple dimensions, and observe the same necessity of having a multi-dimensional contribution.

One might come up with some discretization of the mixed derivative that takes care of these additional terms that appear in multiple dimensions. However, stationarity preservation puts stricter conditions on the shape of the discrete operators. Indeed, not only does one need to ensure that there are no error terms $\mathcal{O}(\Delta x)$ and $\mathcal{O}(\Delta t)$, but one needs more. A way to obtain stationarity preservation is to make appear only the divergence of $\mathbf{v}$ and its derivative. It is reassuring that this is in agreement with the requirements set by the order of the scheme.

The extension step is not unique. Preference is given to covariant operators, i.e. operators that can be written as several applications of grad and div , where div is only allowed to act on vectors and grad on scalars. For the fourth order derivative of u, for example, it is clear that one has to act first with the divergence, then with the gradient, then with the divergence again, and then again with the gradient:

$$\partial_x^2 u \mapsto \text{grad}\,(\text{div grad}\,)(\partial_x u + \partial_y v) = \begin{pmatrix} \partial_x \\ \partial_x \end{pmatrix} (\partial_x^2 + \partial_y^2)(\partial_x u + \partial_y v)$$

The upgrade to multi-dimensional operators can be chosen as

$$\begin{aligned} \frac{1}{\Delta t} \begin{pmatrix} u_{ij}^{n+1} - u_{ij}^n \\ v_{ij}^{n+1} - v_{ij}^n \\ p_{ij}^{n+1} - p_{ij}^n \end{pmatrix} + \begin{pmatrix} \frac{1}{\epsilon^2}\partial_x p \\ \partial_y p \\ c^2(\partial_x u + \partial_y v) \end{pmatrix} + \frac{\Delta x^3}{12}\frac{c}{\epsilon} \begin{pmatrix} \partial_x(\partial_x^2 + \partial_y^2)(\partial_x u + \partial_y v) \\ \partial_y(\partial_x^2 + \partial_y^2)(\partial_x u + \partial_y v) \\ (\partial_x^2 + \partial_y^2)^2 p \end{pmatrix} \\ + \frac{\Delta t}{\Delta x} \left[-\frac{\Delta x}{2}\frac{c^2}{\epsilon^2} \begin{pmatrix} \partial_x(\partial_x u + \partial_y v) \\ \partial_y(\partial_x u + \partial_y v) \\ (\partial_x^2 + \partial_y^2) p \end{pmatrix} + \frac{\Delta x^2}{6}\frac{c}{\epsilon} \begin{pmatrix} \frac{1}{\epsilon^2}\partial_x(\partial_x^2 + \partial_y^2) p \\ \frac{1}{\epsilon^2}\partial_y(\partial_x^2 + \partial_y^2) p \\ c^2(\partial_x^2 + \partial_y^2)(\partial_x u + \partial_y v) \end{pmatrix} \right] = 0 \end{aligned} \tag{4.59}$$

Recall that the continuous operators are placeholders for the discrete operators. Stationarity preservation will only be obtained if

$$\begin{aligned} \partial_x u + \partial_y v = 0 \quad &\Rightarrow \quad \partial_x(\partial_x u + \partial_y v) = 0 \\ &\Rightarrow \quad (\partial_x^2 + \partial_y^2)(\partial_x u + \partial_y v) = 0 \\ &\Rightarrow \quad \partial_x(\partial_x^2 + \partial_y^2)(\partial_x u + \partial_y v) = 0 \end{aligned}$$

is true at discrete level. The one-dimensional operators shall be those that appear in (4.55), and the stationarity consistent extension of such operators to multi-dimensional ones is described in Section 3.2. The Theorems from this Section can thus be used here.

The starting point are the discrete operators in Equation (4.52). The most prominent role is played by the discrete divergence whose dimensionally split version from (4.56) is

$$\frac{u_{i-2,j} - 8u_{i-1,j} + 8u_{i+1,j} - u_{i+2,j}}{12\Delta x} + \frac{v_{i,j-2} - 8v_{i,j-1} + 8v_{i,j+1} - v_{i,j+2}}{12\Delta y}$$

with the Fourier transform

$$\frac{(t_x - 1)(t_x + 1)}{2\Delta x t_x}\frac{-1 + 8t_x - t_x^2}{6t_x}\hat{u} + \frac{(t_y - 1)(t_y + 1)}{2\Delta y t_y}\frac{-1 + 8t_y - t_y^2}{6t_y}\hat{v}$$

By Theorem 3.8 and the example of Section 3.2.2.3 the multi-dimensional extension of the divergence is

$$\begin{aligned}\frac{\hat{u}}{\Delta x}\frac{(t_x-1)(t_x+1)}{2t_x}\frac{(-1+8t_x-t_x^2)}{6t_x}\cdot\frac{(t_y+1)^2}{4t_y}\frac{(-1+8t_y-t_y^2)}{6t_y}\\ +\frac{\hat{v}}{\Delta y}\frac{(t_x+1)^2}{4t_x}\frac{(-1+8t_x-t_x^2)}{6t_x}\cdot\frac{(t_y-1)(t_y+1)}{2t_y}\frac{(-1+8t_y-t_y^2)}{6t_y}\end{aligned} \tag{4.60}$$

allows to find stationarity consistent discrete derivatives of the divergence.

Upon multiplication of (4.60) with $\Delta x\partial_x \simeq 2\dfrac{(t_x-1)(1+4t_x+t_x^2)}{(t_x+1)(-1+8t_x-t_x^2)}$ one obtains the Fourier transform of a stencil containing second derivatives:

$$\begin{aligned}\Delta x\partial_x(\partial_x u+\partial_y v) \simeq \frac{\hat{u}}{\Delta x}\frac{(t_x-1)^2}{t_x}\frac{1+4t_x+t_x^2}{6t_x}\cdot\frac{(t_y+1)^2}{4t_y}\frac{-1+8t_y-t_y^2}{6t_y}\\ +\frac{\hat{v}}{\Delta y}\frac{(t_x+1)(t_x-1)}{2t_x}\frac{1+4t_x+t_x^2}{6t_x}\cdot\frac{(t_y-1)(t_y+1)}{2t_y}\frac{-1+8t_y-t_y^2}{6t_y}\end{aligned}$$

Factor $\Delta x^2\partial_x^2+\Delta y^2\partial_y^2 \simeq 6\dfrac{(t_x-1)^2}{-1+8t_x-t_x^2}+6\dfrac{(t_y-1)^2}{-1+8t_y-t_y^2}$ gives third derivatives:

$$\begin{aligned}(\Delta x^2\partial_x^2+\Delta y^2\partial_y^2)(\partial_x u+\partial_x v) \simeq \frac{\hat{u}}{\Delta x}\left(\frac{(t_x-1)^3(t_x+1)}{2t_x^2}\cdot\frac{(t_y+1)^2}{4t_y}\frac{-1+8t_y-t_y^2}{6t_y}+\frac{(t_x-1)(t_x+1)}{2t_x}\frac{-1+8t_x-t_x^2}{6t_x}\cdot\frac{(t_y+1)^2(t_y-1)^2}{4t_y^2}\right)\\ +\frac{\hat{v}}{\Delta y}\left(\frac{(t_x+1)^2(t_x-1)^2}{4t_x^2}\cdot\frac{(t_y-1)(t_y+1)}{2t_y}\frac{-1+8t_y-t_y^2}{6t_y}+\frac{(t_x+1)^2}{4t_x}\frac{-1+8t_x-t_x^2}{6t_x}\cdot\frac{(t_y-1)^3(t_y+1)}{2t_y^2}\right)\end{aligned}$$

Finally, the factor $\Delta x\partial_x(\Delta x^2\partial_x^2+\Delta y^2\partial_y^2) \simeq 12\dfrac{(t_x-1)^3}{(t_x+1)(-1+8t_x-t_x^2)}+18\dfrac{(t_x-1)(t_x+1)}{-1+8t_x-t_x^2}\cdot\dfrac{(t_y-1)^2}{-1+8t_y-t_y^2}$ gives fourth derivatives:

$$\begin{aligned}\Delta x\partial_x(\Delta x^2\partial_x^2+\Delta y^2\partial_y^2)(\partial_x u+\partial_y v) \simeq \frac{\hat{u}}{\Delta x}\left(\frac{(t_x-1)^4}{t_x^2}\cdot\frac{(t_y+1)^2}{4t_y}\frac{-1+8t_y-t_y^2}{6t_y}+\frac{(t_x-1)^2(t_x+1)^2}{4t_x^2}\cdot\frac{(t_y+1)^2(t_y-1)^2}{4t_y^2}\right)\\ +\frac{\hat{v}}{\Delta y}\left(\frac{(t_x+1)(t_x-1)^3}{2t_x^2}\cdot\frac{(t_y-1)(t_y+1)}{2t_y}\frac{-1+8t_y-t_y^2}{6t_y}+\frac{(t_x+1)^3(t_x-1)}{8t_x^2}\cdot\frac{(t_y-1)^3(t_y+1)}{2t_y^2}\right)\end{aligned}$$

This allows to replace the terms in (4.59) by their discretizations. The scheme becomes a lengthy expression. The relevant quantity for the implementation is the intercell x-flux $f^x_{i+\frac{1}{2},j}$ from which all other fluxes can be derived by suitable permutations of directions and variables. Also,

$$\mathbb{F}\left[\frac{f^x_{i+\frac{1}{2},j} - f^x_{i-\frac{1}{2},j}}{\Delta x}\right] = \frac{1}{\Delta x}\mathbb{F}\left[f^x_{i+\frac{1}{2},j}\right]\left(1 - \frac{1}{t_x}\right)$$

such that

$$\mathbb{F}\left[f^x_{i+\frac{1}{2},j}\right] = \mathbb{F}\left[\frac{f^x_{i+\frac{1}{2},j} - f^x_{i-\frac{1}{2},j}}{\Delta x}\right]\frac{\Delta x\, t_x}{t_x - 1}$$

Note that the transition from the complete terms as in (4.59) to formulae for the x-flux does not only imply multiplication with $\frac{\Delta x\, t_x}{t_x - 1}$, but also that only half of the terms have to be taken into account, preferably those beginning with a ∂_x-derivative.

This greatly simplifies the expressions. Additionally, the $\otimes$-notation from Definition 3.8 is used and the following notation is introduced:

Definition 4.9. *Define the two averaging notations*

$$\mu(t_y) := \frac{(t_y+1)^2}{4t_y}\frac{-1 + 8t_y - t_y^2}{6t_y}$$
$$\langle q\rangle_j := \frac{-q_{j-2} + 6q_{j-1} + 14q_j + 6q_{j+1} - q_{j+2}}{24}$$

The fluxes corresponding to each of the four terms are given on page 135. Experimentally, the scheme is found to be stable up to a CFL $= 1$.

$$
\begin{pmatrix} \frac{1}{\epsilon^2}\partial_x p \\ \partial_y p \\ c^2(\partial_x u + \partial_y v) \end{pmatrix} \quad : \quad f^x_{i+\frac{1}{2},j} \simeq \begin{pmatrix} \frac{1}{\epsilon^2}\hat{p}\frac{(t_x+1)}{2}\frac{-1+8t_x-t_x^2}{6t_x} \cdot \frac{(t_y+1)^2}{4t_y}\frac{-1+8t_y-t_y^2}{6t_y} \\ 0 \\ c^2\hat{u}\frac{(t_x+1)}{2}\frac{-1+8t_x-t_x^2}{6t_x} \cdot \frac{(t_y+1)^2}{4t_y}\frac{-1+8t_y-t_y^2}{6t_y} \end{pmatrix}
$$

$$
\simeq J_x \frac{-q_{i-1}+7q_i+7q_{i+1}-q_{i+2}}{12} \otimes \langle q \rangle_j
$$

$$
-\frac{\Delta x}{2}\frac{\Delta t}{\Delta x}\frac{c^2}{\epsilon^2}\begin{pmatrix} \partial_x(\partial_x u + \partial_y v) \\ \partial_y(\partial_x u + \partial_y v) \\ (\partial_x^2+\partial_y^2)p \end{pmatrix} \quad : \quad f^x_{i+\frac{1}{2},j} \simeq -\frac{1}{2}\frac{\Delta t}{\Delta x}\frac{c^2}{\epsilon^2}\begin{pmatrix} \hat{u}(t_x-1)\frac{1+4t_x+t_x^2}{6t_x} \cdot \mu(t_y) + \hat{v}\frac{(t_x+1)}{2}\frac{1+4t_x+t_x^2}{6t_x} \cdot \frac{(t_y-1)(t_y+1)}{2t_y}\frac{-1+8t_y-t_y^2}{6t_y} \\ 0 \\ \hat{p}(t_x-1)\frac{1+4t_x+t_x^2}{6t_x} \cdot \mu(t_y) \end{pmatrix}
$$

$$
\simeq -\frac{1}{2}\frac{\Delta t}{\Delta x}\frac{c^2}{\epsilon^2}\begin{pmatrix} \frac{-u_{i-1}-3u_i+3u_{i+1}+u_{i+2}}{6} \otimes \langle u \rangle_j + \frac{v_{i-1}+5v_i+5v_{i+1}+v_{i+2}}{12} \otimes \frac{v_{j-2}-8v_{j-1}+8v_{j+1}-v_{j+2}}{12} \\ 0 \\ \frac{-p_{i-1}-3p_i+3p_{i+1}+p_{i+2}}{6} \otimes \langle p \rangle_j \end{pmatrix}
$$

$$
\Delta t\frac{\Delta x}{6}\frac{c}{\epsilon}\begin{pmatrix} \frac{1}{\epsilon^2}\partial_x(\partial_x^2+\partial_y^2)p \\ \frac{1}{\epsilon^2}\partial_y(\partial_x^2+\partial_y^2)p \\ c^2(\partial_x^2+\partial_y^2)(\partial_x u + \partial_y v) \end{pmatrix} \quad : \quad f^x_{i+\frac{1}{2},j} \simeq \Delta t\frac{1}{6\Delta x}\frac{c}{\epsilon}\begin{pmatrix} \frac{1}{\epsilon^2}\hat{p}\left(\frac{(t_x-1)^2(t_x+1)}{2t_x} \cdot \mu(t_y) + \frac{(t_x+1)}{2}\frac{-1+8t_x-t_x^2}{6t_x} \cdot \frac{(t_y+1)^2(t_y-1)^2}{4t_y^2}\right) \\ 0 \\ c^2\left(\hat{u}\frac{(t_x-1)^2(t_x+1)}{2t_x} \cdot \mu(t_y) + \hat{v}\frac{(t_x+1)^2(t_x-1)}{4t_x} \cdot \frac{(t_y-1)(t_y+1)}{2t_y}\frac{-1+8t_y-t_y^2}{6t_y}\right) \end{pmatrix}
$$

$$
\simeq \Delta t\frac{1}{6\Delta x}\frac{c}{\epsilon}\begin{pmatrix} \frac{1}{\epsilon^2}\left(\frac{p_{i-1}-p_i-p_{i+1}+p_{i+2}}{2} \otimes \langle p \rangle_j + \frac{-p_{i-1}+7p_i+7p_{i+1}-p_{i+2}}{12} \otimes \frac{p_{j-2}-2p_j+p_{j+2}}{4}\right) \\ 0 \\ c^2\left(\frac{u_{i-1}-u_i-u_{i+1}+u_{i+2}}{2} \otimes \langle u \rangle_j + \frac{-v_{i-1}-v_i+v_{i+1}+v_{i+2}}{4} \otimes \frac{v_{j-2}-8v_{j-1}+8v_{j+1}-v_{j+2}}{4}\right) \end{pmatrix}
$$

$$
\frac{\Delta x^3}{12}\frac{c}{\epsilon}\begin{pmatrix} \partial_x(\partial_x^2+\partial_y^2)(\partial_x u + \partial_y v) \\ \partial_y(\partial_x^2+\partial_y^2)(\partial_x u + \partial_y v) \\ (\partial_x^2+\partial_y^2)(\partial_x^2+\partial_y^2)p \end{pmatrix} \quad : \quad f^x_{i+\frac{1}{2},j} \simeq \frac{1}{12}\frac{c}{\epsilon}\begin{pmatrix} \hat{u}\left(\frac{(t_x-1)^3}{t_x} \cdot \mu(t_y) + \frac{(t_x-1)(t_x+1)^2}{4t_x} \cdot \frac{(t_y+1)^2(t_y-1)^2}{4t_y^2}\right) + \hat{v}\left(\frac{(t_x+1)(t_x-1)^2}{2t_x} \cdot \frac{(t_y-1)(t_y+1)}{2t_y}\frac{-1+8t_y-t_y^2}{6t_y} + \frac{(t_x+1)^3}{8t_x} \cdot \frac{(t_y-1)^3(t_y+1)}{2t_y^2}\right) \\ 0 \\ \hat{p}\left(\frac{(t_x-1)^3}{t_x} \cdot \mu(t_y) + \frac{(t_x-1)(t_x+1)^2}{4t_x} \cdot \frac{(t_y+1)^2(t_y-1)^2}{4t_y^2}\right) \end{pmatrix}
$$

$$
\simeq \frac{1}{12}\frac{c}{\epsilon}\begin{pmatrix} (-u_{i-1}+3u_i-3u_{i+1}+u_{i+2}) \otimes \langle u \rangle_j + \frac{-u_{i-1}-u_i+u_{i+1}+u_{i+2}}{4} \otimes \frac{u_{j-2}-2u_j+u_{j+2}}{4} \\ 0 \\ (-p_{i-1}+3p_i-3p_{i+1}+p_{i+2}) \otimes \langle p \rangle_j + \frac{-p_{i-1}-p_i+p_{i+1}+p_{i+2}}{4} \otimes \frac{p_{j-2}-2p_j+p_{j+2}}{4} \end{pmatrix}
$$

$$
+ \frac{1}{12}\frac{c}{\epsilon}\begin{pmatrix} \frac{v_{i-1}-v_i-v_{i+1}+v_{i+2}}{2} \otimes \frac{v_{j-2}-8v_{j-1}+8v_{j+1}-v_{j+2}}{12} + \frac{v_{i-1}+3v_i+3v_{i+1}+v_{i+2}}{8} \otimes \frac{-v_{j-2}+2v_{j-1}-2v_{j+1}+v_{j+2}}{8} \\ 0 \\ 0 \end{pmatrix}
$$

4.7.3 Numerical results

Results of a simulation of a divergence-free vortex with the scheme presented on page 135 can be seen in Fig. 4.14. There is no visible difference to the solution shown in Figure 4.11, because both schemes are stationarity preserving. However, the measured error (shown in Fig 4.15) clearly shows the higher order of the scheme. The error is measured against the initial data. Despite the fact that the stationary divergence is discretized to much higher order, the error against the initial data is still the overall error of the scheme. A numerical solution of the Riemann Problem discussed in Section 2.2.5 is shown in Figure 4.16.

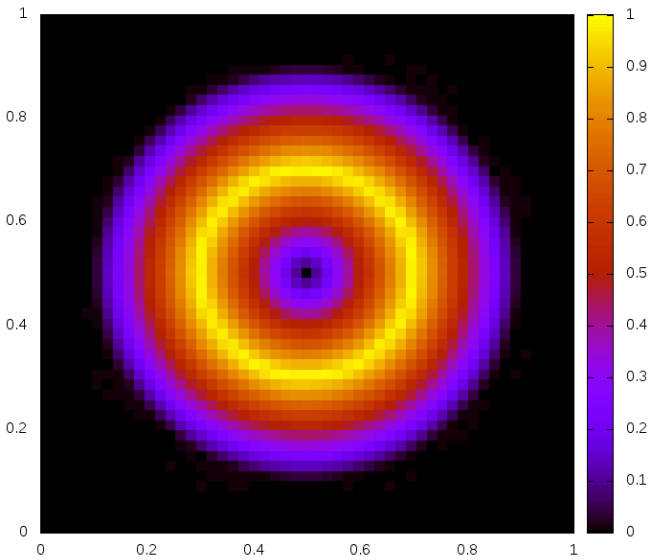

Figure 4.14: Solution of a stationary divergence-free vortex (as in Fig. 4.7) at $t = 10$ with the second order scheme presented on page 135. $\sqrt{u^2 + v^2}$ is colour coded; simulation performed on a 50×50 grid. Commit hash: e87856b.

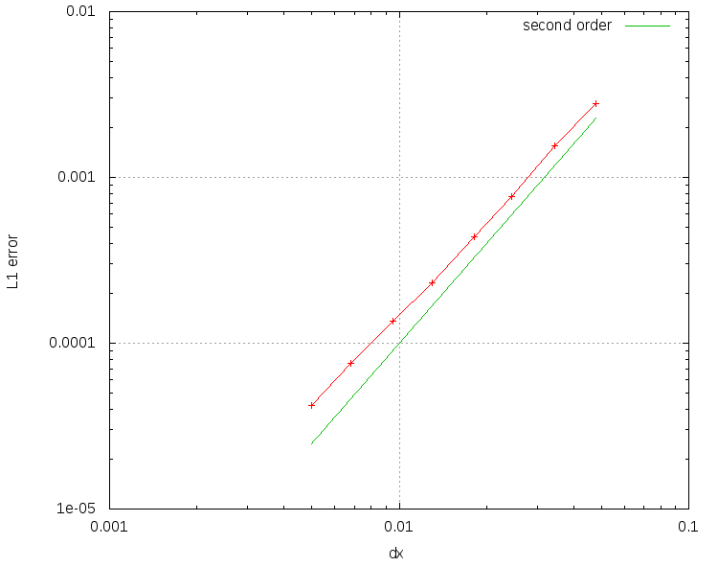

Figure 4.15: Convergence test of the second order scheme presented on page 135. The setup is that of Figure 4.14, and the error measured against the initial data. As a function of the linear cell size Δx it shows the correct behaviour. Commit hash: 291b4e3.

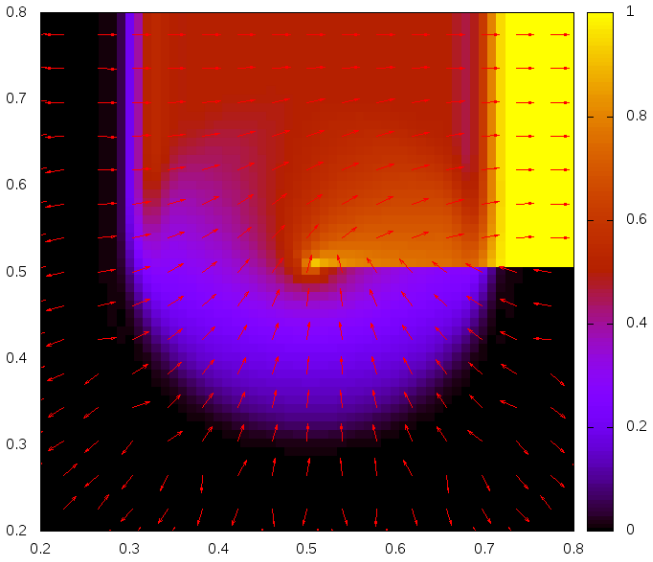

Figure 4.16: Solution of the Riemann Problem discussed in Section 2.2.5. The simulation has been performed with the second order scheme presented on page 135 on a 100×100 grid, and the results are shown at $t = 0.2$. The direction of the velocity $\mathbf{v}(t, \mathbf{x})$ is indicated by the arrows, color coded is the absolute value $|\mathbf{v}|$. As the scheme does not have a limiter one observes small overshoots at the discontinuities (e.g. at $x \simeq 0.7$). Compare this to Figures 2.3 and 4.13. Commit hash: 18aab08.

4.8 Stationarity preserving schemes for gravity-like source terms

The strategies that lead to stationarity preserving multi-dimensional schemes can be also applied to entirely different situations. Therefore this Section is a digression on a different type of a linear hyperbolic system that contains source terms. Consider the following equations:

$$\begin{aligned} \partial_t \rho + \partial_x u + \partial_y v &= 0 \qquad (4.61)\\ \partial_t u + \partial_x p &= \rho g_x \\ \partial_t v + \partial_y p &= \rho g_y \\ \partial_t p + c^2(\partial_x u + \partial_y v) &= 0 \qquad (4.62) \end{aligned}$$

The Jacobians are

$$J_x = \begin{pmatrix} 0 & 1 & & \\ & 0 & & 1 \\ & & 0 & \\ & c^2 & & 0 \end{pmatrix} \qquad J_y = \begin{pmatrix} 0 & & 1 & \\ & 0 & & \\ & & 0 & 1 \\ & & c^2 & 0 \end{pmatrix} \qquad (4.63)$$

Defining S as

$$S := \begin{pmatrix} 0 & & & \\ g_x & 0 & & \\ g_y & & 0 & \\ & & & 0 \end{pmatrix} \qquad (4.64)$$

and $q := (\rho, u, v, p)$ one can rewrite the equations as

$$\partial_t q + J_x \partial_x q + J_y \partial_y q = Sq$$

These equations can be obtained as the linearization of the Euler equations with gravity (1.22)–(1.24). Obviously, both for (1.22)–(1.24) and (4.61)–(4.62) there exists a stationary state with the flux difference being balanced by the source. Such setups are called *equilibria*.

By analogy with Theorem 3.1, the nontrivial stationary states are governed by $\mathbf{J} \cdot \mathbf{k} + \mathbb{i} S$:

Theorem 4.16. *Given $0 \neq \mathbf{k} \in \mathbb{R}^d$, if $\det(\mathbf{k} \cdot \mathbf{J} + \mathbb{i} S)$ vanishes for all $\mathbf{k}$, then there exist non-trivial stationary states of $\partial_t q + J_x \partial_x q + J_y \partial_y q = Sq$ with $q : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}^n$.*

Proof. Consider the Fourier transform as

$$q(t, \mathbf{x}) = \hat{q} \exp(-\mathbb{i}\omega t + \mathbb{i}\mathbf{k} \cdot \mathbf{x})$$

to obtain the eigenvalue problem $\omega q = (\mathbf{J} \cdot \mathbf{k} + \mathbb{i} S) q$. From here the result follows from the proof of Theorem 3.1. □

One finds upon explicit calculation that for the choice (4.63) and (4.64), indeed $\det(\mathbf{k}\cdot\mathbf{J} + \mathbb{i}S) = 0$. The stationary states are governed by

$$\partial_x u + \partial_x v = 0 \qquad \partial_x p = \rho g_x \qquad \partial_y p = \rho g_y$$

In order to concentrate on the new feature of a source term, the Equations (4.61)–(4.62) are first considered in one spatial dimension:

$$\partial_t \rho + \partial_x u = 0 \tag{4.65}$$
$$\partial_t u + \partial_x p = \rho g$$
$$\partial_t p + c^2 \partial_x u = 0 \tag{4.66}$$

with

$$J := \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & c^2 & 0 \end{pmatrix} \qquad |J| = \begin{pmatrix} 0 & & \frac{1}{c} \\ & c & \\ & & c \end{pmatrix} \qquad S := \begin{pmatrix} 0 & 0 & 0 \\ g & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{4.67}$$

Again, one easily confirms that $\det(Jk + \mathbb{i}S) = 0 \ \forall k \in \mathbb{R}$.

4.8.1 Cell-centered source term evaluation

Numerical schemes typically have problems maintaining equilibria numerically because the way the fluxes are computed often does not fit the discretization of the source. Consider as an example the upwind/Roe solver for Equations (4.65)–(4.66) with the simplest possible evaluation of the source term:

$$\partial_t \begin{pmatrix} \rho \\ u \\ p \end{pmatrix}_i + \frac{1}{2\Delta x} J \begin{bmatrix} \rho \\ u \\ p \end{bmatrix}_{i\pm 1} - \frac{1}{2\Delta x} |J| \left[\!\!\left[\begin{matrix} \rho \\ u \\ p \end{matrix} \right]\!\!\right]_{i\pm\frac{1}{2}} = \begin{pmatrix} 0 \\ \rho g \\ 0 \end{pmatrix}_i$$

An example of a simulation with this scheme is shown in Figure 4.17.

Adjusting the discretizations of the fluxes and of the source leads to schemes that are able to maintain (certain) equilibria: such schemes are called *well-balanced*. There is a rich literature and a number of different approaches to this in the context of the Euler equations: e.g. [Gos01, KM14, CK15] and many others. For linear systems such as (4.61)–(4.62), this phenomenon can also be studied using the concept of stationarity preservation. This allows to give the observed numerical artefacts a new interpretation (which is summarized in Definition 4.10) and also shows a way how to construct well-balanced schemes.

Theorem 4.17. *The upwind/Roe solver with cell-centered discretization of gravity in 1-d is stationarity preserving.*

Proof. The evolution matrix for the scheme (4.68) is

$$\frac{1}{2\Delta x} J \frac{t_x^2 - 1}{t_x} - \frac{1}{2\Delta x} |J| \frac{(t_x - 1)^2}{t_x} - S \tag{4.68}$$

with J, $|J|$ and S given by (4.67). Upon explicit computation its determinant is found to vanish which is the condition of Theorem 3.3. □

The Fourier mode that is kept stationary is parallel to the eigenvector

$$\left(1, \frac{\Delta x g}{2c}, \frac{\Delta x g (t_x + 1)}{2(t_x - 1)}\right)^{\mathrm{T}}$$

In particular, a stationary state always has a non-vanishing and spatially non-constant velocity

$$v = \frac{\Delta x g}{2c} \rho + \text{const} \tag{4.69}$$

This can be seen in the experiment, evolved until the stationary state is reached (Figure 4.17).

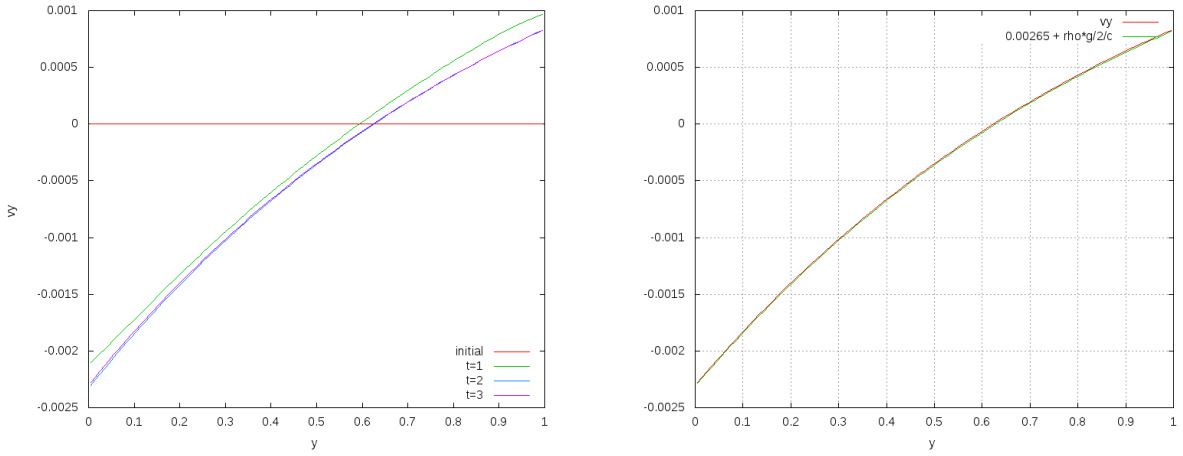

Figure 4.17: Simulation results with the scheme (4.68) on 100 cells with $g = -1$. The initial state is that of a polytropic equilibrium $p = K\rho^\gamma$ with $K = 1$, $\gamma = 1.4$ and zero velocity. *Left*: Evolution of the velocity u for $t = 0, 1, 2, 3$. One observes the stationarization of the numerical results on a spatially non-constant velocity. *Right*: Comparison between the numerical stationary state and the theoretical prediction (4.69) (with the constant having been found experimentally).

Colloquially, one would call the scheme (4.68) not well-balanced. In the language of the methods developed in Section 3 this can be given a new interpretation. It might have been suspected that the observed artefacts are due to the scheme not being stationarity preserving. This is not the case: according to Theorem 4.17 the scheme discretizes all the stationary states of the PDE. Therefore the situation is better than with the limit of low Mach numbers. The visually unsatisfactory behaviour is related to the

discrete stationary states: whereas analytically, $v = 0$, the discrete stationary states have nonvanishing velocity. Therefore a definition of *well-balanced* for this particular situation can be given as follows:

Definition 4.10 (Well-balanced)**.** *A linear scheme for the equations* (4.65)–(4.66) *is* well-balanced*, if it is stationarity preserving and the discrete stationary states are characterized by a spatially constant velocity v.*

Note the condition of *spatially constant* velocity, rather than a vanishing. It is impossible to guarantee vanishing velocity: the state of constant density, velocity and pressure is a trivial stationary state of numerical schemes, and the sum of any two stationary solutions is stationary as well.

4.8.2 Well-balanced diffusion

The challenge is to find a scheme for Equations (4.65)–(4.66) with particular discrete stationary states. This fits exactly into the framework of Section 3.2. There, a discrete divergence was given, and a stationarity-consistent diffusion was constructed. Here, if the central discretization of $\partial_x p$ shall be stationary, one needs to find a diffusion that originates from the discrete stationary state

$$\frac{1}{2\Delta x}[p]_{i\pm1} - s_i = 0$$

The source term has been given the name s generically. In the case that is considered, actually $s = \rho g$, but this is unimportant for the discussion. The Fourier transform reads

$$\frac{(t_x - 1)(t_x + 1)}{2\Delta x t_x}\hat{p} - \hat{s} = 0$$

Obviously this does not allow multiplication with $2\frac{t_x-1}{t_x+1}$ which would lead to the Fourier transform of a second derivative. But it is easy to find the right discretization:

Theorem 4.18 (Well-balanced diffusion)**.** *The scheme*

$$\partial_t \begin{pmatrix} \rho \\ v \\ p \end{pmatrix}_i + \frac{1}{2\Delta x}\begin{pmatrix} [u]_{i\pm1} \\ [p]_{i\pm1} \\ c^2[u]_{i\pm1} \end{pmatrix} - \frac{1}{2\Delta x}\begin{pmatrix} \frac{1}{c}\left([[p]]_{i\pm\frac{1}{2}} - \frac{\Delta x}{2}[s]_{i\pm1}\right) \\ c[[u]]_{i\pm\frac{1}{2}} \\ c\left([[p]]_{i\pm\frac{1}{2}} - \frac{\Delta x}{2}[s]_{i\pm1}\right) \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{4}\{\{s\}\}_{i\pm\frac{1}{2}} \\ 0 \end{pmatrix} \tag{4.70}$$

is well-balanced. The stationary states are characterized by $u = \text{const}$ and

$$\frac{1}{2\Delta x}[p]_{i\pm1} - \frac{1}{4}\{\{s\}\}_{i\pm\frac{1}{2}} = 0 \tag{4.71}$$

Proof. One can compute the kernel of the evolution matrix of this scheme in order to prove the assertion. However one can also note that the Fourier transform of the asserted stationary state

$$\frac{(t_x-1)(t_x+1)}{2\Delta x t_x}\hat{p} - \frac{(t_x+1)^2}{4t_x}\hat{s}$$

upon multiplication with $2\frac{t_x-1}{t_x+1}$ yields

$$\frac{(t_x-1)^2}{\Delta x t_x}\hat{p} - \frac{(t_x-1)(t_x+1)}{2t_x}\hat{s}$$

This is the Fourier transform of

$$\frac{1}{\Delta x}[[p]]_{i\pm\frac{1}{2}} - \frac{1}{2}[s]_{i\pm1}$$

which, by setting $u = \text{const}$ is the only other term that appears in the scheme.

□

Note: The new diffusion is conservative:

$$[[p]]_{i\pm\frac{1}{2}} - \frac{\Delta x}{2}[\rho g_x]_{i\pm1} = \left[[p] - \frac{\Delta x}{2}\{\rho g_x\}\right]_{i\pm\frac{1}{2}}$$

The very same method follows if one applies the well-balancing strategy described in [LeV98], which introduces additional jumps at cell centers.

The stationary Fourier mode is proportional to $\left(1, 0, \frac{\Delta x g(t_x+1)}{2(t_x-1)}\right)^{\mathrm{T}}$.

This can be seen in the experiment (Figure 4.18).

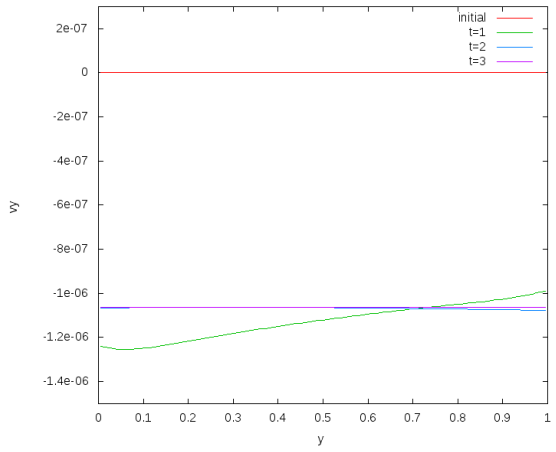

Figure 4.18: Evolution of the velocity u with the scheme (4.70) and the setup of Figure 4.17. Observe how the stationary state of the simulation now has a spatially constant velocity. When comparing to Figure 4.17 mind the scale of the velocity axis.

Note that the stationary velocity is not zero. This is because the initial data do not exactly fulfill (4.71). After some time the other Fourier modes have been diffused away which has led to a small change in u. However the stationary velocity u is spatially constant.

In order to construct numerical schemes for the multi-dimensional system (4.61)–(4.61) the strategy presented here (which takes care of the balance $\nabla p = \rho g$) needs to be combined with ideas of Section 4.5 (which take care of $\nabla \cdot \mathbf{v}$).

Chapter 5

Numerical schemes for the Euler equations

Contents

Consider a system of conservation laws in $d = 2$ spatial dimensions

$$\partial_t q + \partial_x f^x(q) + \partial_y f^y(q) = 0 \qquad q : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}^n \qquad f^x, f^y : \mathbb{R}^n \to \mathbb{R}^n \tag{5.1}$$

with q being the vector of conserved quantities and $\mathbf{f} = (f^x, f^y)$ the flux in x-, y-direction, respectively. In the case of the Euler equations $q = (\rho, \rho\mathbf{v}, e)$. The corresponding equations for three spatial dimensions can be found by direct analogy.

Finite volume schemes interpret the discrete degrees of freedom as cell averages. Integrating (5.1) over one computational cell $\mathcal{C}_{ij} \subset \mathbb{R}^d$ and a time interval $[t^n, t^{n+1}]$ yields

$$\int_{\mathcal{C}_{ij}} \mathrm{d}\mathbf{x}\, q(t, \mathbf{x}) \Big|_{t^n}^{t^{n+1}} + \int_{t^n}^{t^{n+1}} \mathrm{d}t \int_{\partial\mathcal{C}_{ij}} \mathrm{d}s\, \mathbf{f} \cdot \mathbf{n} = 0$$

Here $\mathbf{n}$ is the outward normal onto $\partial\mathcal{C}_{ij}$. The semi-discrete finite volume scheme is obtained with the identification

$$q_{ij}^n := \frac{1}{|\mathcal{C}_{ij}|} \int_{\mathcal{C}_{ij}} q(t^n, \mathbf{x}) \mathrm{d}\mathbf{x}$$

On Cartesian grids, for example, $\mathcal{C}_{ij} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, and the above formula becomes

$$\frac{q_{ij}^{n+1} - q_{ij}^{n}}{\Delta t} + \frac{f^x_{i+\frac{1}{2},j} - f^x_{i-\frac{1}{2},j}}{\Delta x} + \frac{f^y_{i,j+\frac{1}{2}} - f^y_{i,j-\frac{1}{2}}}{\Delta y} = 0$$

having defined

$$f^x_{i+\frac{1}{2},j} := \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \mathrm{d}t \frac{1}{\Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \mathrm{d}y\, f^x(q(t, x_{i+\frac{1}{2}}, y)) \tag{5.2}$$

and analogously the flux $f^y_{i,j+\frac{1}{2}}$ in the other direction.

Obviously, in order to compute the flux $f^x_{i+\frac{1}{2},j}$ through an interface according to formula (5.2), one needs to know the exact solution q along the interface. In the discrete setting, however, only the cell averages q^n_{ij} are known. The essence of the numerical method thus lies in the choice of an approximation to Equation (5.2).

Quite generally, one thus has to accept that the discrete solution is only representing the exact solution up to some error, and this is the case for all flow regimes, be it of high Mach number or low Mach number. However, e.g. in [Tur87, KLN91] it has been noticed that there exists a dichotomy of numerical schemes with respect to their behaviour for *low* Mach number flow: Schemes have been discovered (more or less by accident) whose performance in the low Mach number regime was *much* better than that of schemes known so far, although they had the same order of convergence and the simulations were performed on the same grid. It has become obvious that numerical schemes split up into those for which the quality of the numerical solution deteriorates in the low Mach number limit, and those for which the quality of the results in the limit stays roughly the same. To the former group belong many popular "standard" finite volume schemes. The simulation quality of the latter can only be reached, if the simulations are run at much finer grids. The lower the Mach numbers of the flow, the finer has the grid to be, which is impractical. The quality deterioration is known as the *low Mach number problem* of numerical schemes.

There exists a rich literature devoted to this subject: e.g. [Kle95, DJY07, CDK12, HJL12, CGK13, NBA+14, DLV17, BLMY17, Tur87, KLN91, GV99, GM04, BM05, LG08, TD08, TMD+08, EL98, Lio06, RB09, Rie10, DOR10, Del10, Rie11, LG13, MRE15, OSB+16, DJOR16, BEK+17] and many others.

There are certain situations which generically seem to have less issues with the low Mach limit. These can be staggered grids and, as has been discovered only recently, triangular/tetrahedral grids (of otherwise arbitrary shape) ([DOR10, Gui09]). This does not mean, though, that they are prone to the appearance of low Mach number artefacts. Taking the triangular grids as example, [DOR10] shows that numerical simulations do not suffer from the described quality deterioration when the Roe solver is used, but the low Mach number problem reappears for the HLLC solver.

There exist a number of approaches that explain the deterioration of the quality of numerical solution. A widely used methodology is to consider a formal asymptotic

expansion of the scheme in the limit $\epsilon \to 0$. The scalings thus obtained are compared to those known for the analytic solution, and discrepancies are made responsible for the observed quality deterioration. As with everything that is preceded by the word "formal", formal asymptotic analysis might be unable to explain all experimental findings. In practice however, asymptotic analysis is hugely successful in predicting which scheme will manage to resolve the low Mach number limit. In Sections 5.1.2 and 5.2.1 it is discussed in detail and results of the analysis are shown.

A nonlinear numerical scheme for the Euler equations can be linearized around the state of constant density and pressure and of zero velocity (*static state*, see Definition 1.2). It then becomes a numerical scheme for the equations of linear acoustics (2.40)–(2.41). A necessary condition for the nonlinear scheme to perform well in the low Mach number limit $\epsilon \to 0$ is that its linearized version is, by Theorem 4.1, stationarity preserving. Thus, arguments that were given for linear acoustics actually allow to also make statements about schemes for the Euler equations. Often, purely linear arguments are sufficient to identify schemes that fail. A discussion of these ideas is subject of Section 5.2.2.

Definition 5.1 (Low Mach compliant)**.** *A numerical scheme for* (1.14)–(1.16) *is called* low Mach compliant *if in the limit* $\epsilon \to 0$ *it has solutions that discretize all the analytic solutions given by the limit equations* (1.17)–(1.19)*.*

A lot of work in the literature has been devoted to the question how to modify a scheme such that it *becomes* low Mach compliant, often starting from the Roe scheme. A new modification strategy, introduced in [MRE15, BEK+17] is presented and analyzed in Section 5.3.2. It is constructed such that it can be used even when solving the Euler equations with gravity (Equations (1.27)–(1.29)).

One might also wonder what construction principles would automatically incorporate low Mach compliance, such that the schemes would not need to be *fixed* afterwards. In the context of the Euler equations it is to be understood as follows: Often, one-dimensional schemes are extended to multiple spatial dimensions in a dimensionally split way. This extension is found to violate the ability of the scheme to resolve the low Mach number limit, and needs to be *fixed*. The question is how to extend a one-dimensional scheme to multiple dimensions in such a way that it becomes low Mach compliant straight away. This thesis presents a number of approaches to incorporate ideas of stationarity preservation into the nonlinear setting (Section 5.4) and a novel kind of low Mach compliant scheme (Section 5.5). The latter is inspired by the multi-dimensional stationarity preserving scheme of Section 4.5.2.

For the Euler equations, the evolution of vorticity is much more complicated than in the case of linear acoustics. In general, it is not stationary, and therefore also any discrete vorticity should evolve in time. Focus, for example, on the advective term $(\mathbf{v} \cdot \nabla)\omega$ in (1.11). There exist a lot of ways how such an advection could be discretized, and so far it does not seem obvious which one should be used. Would one prefer certain types of vorticity advection over others? Suggestions with particular examples of – very different – schemes can be found in [Sid02, JT06, LFS07]. The relation to low Mach compliant schemes is subject of future work.

As was said above, the essence of low Mach number is the dichotomy of schemes: those whose solutions deteriorate in the limit, and those whose solutions do not. It seems possible to theoretically capture this dichotomy. Still, two different schemes might give very different solutions, even if they are both low Mach compliant (for instance, they might have different orders of convergence). How to measure and decide which of them is to be preferred over the other is not subject of this discussion, and may also depend on the particular application and setup.

5.1 The Roe scheme and the low Mach number problem

5.1.1 The Roe scheme

One choice of the numerical flux is the Roe method [Roe81], e.g. in x-direction:

$$f^x_{i+\frac{1}{2},j} = \frac{1}{2}\left[f^x\left(q_{ij}\right) + f^x\left(q_{i+1,j}\right) - |J|(q_{i+\frac{1}{2},j})\left(q_{i+1,j} - q_{ij}\right)\right] \tag{5.3}$$

with the *Roe matrix* $|J|(q_{i+\frac{1}{2},j})$ being evaluated with the average $q_{i+\frac{1}{2},j}$. This average can be chosen such that ([Roe81, LeV02], see Theorem 5.6 below)

$$f^x\left(q_{i+1,j}\right) - f^x\left(q_{ij}\right) = J(q_{i+\frac{1}{2},j})(q_{i+1,j} - q_{ij})$$

This scheme (in one spatial dimension) reduces to the upwind scheme if applied to linear systems. It is experimentally found not to be able to resolve low Mach number flows.

The central flux on the other hand,

$$f^x_{i+\frac{1}{2},j} = \frac{1}{2}\Big[f^x\left(q_{ij}\right) + f^x\left(q_{i+1,j}\right)\Big]$$

is unstable under explicit time integration, but yields satisfactory results in the limit of low Mach numbers, if stabilized by implicit time integration ([MRE15]). This has led many authors to careful analysis of the matrix $|J|(q_{i+\frac{1}{2},j})$ in order to understand the origin of the quality deterioration for low Mach numbers.

5.1.2 Asymptotic analysis

Obviously, the Roe scheme cannot be low Mach compliant, because upon linearization around the static state it reduces to the upwind/Roe scheme for linear acoustics, which has been shown not to be stationarity preserving in Corollary 4.8. However, for reference purposes here the formal asymptotic expansion of the scheme is stated. One expands every quantity in the numerical scheme as a power series in ϵ and collects terms of equal order.

Define the notation

$$\mathbf{v} = (u, v)$$
$$|J|_{i+\frac{1}{2},j} := |J|(q_{i+\frac{1}{2},j})$$

The flux difference in x-direction reads

$$\frac{f^x_{i+\frac{1}{2},j} - f^x_{i-\frac{1}{2},j}}{\Delta x} = \frac{1}{2\Delta x}\left(\begin{bmatrix} \rho u \\ \rho u^2 \\ \rho uv \\ u(e+p) \end{bmatrix}_{i+\frac{1}{2},j} + \frac{1}{\epsilon^2} \begin{bmatrix} 0 \\ p \\ 0 \\ 0 \end{bmatrix}_{i+\frac{1}{2},j} - \left[|J|_{\cdot,j} \begin{bmatrix} \rho \\ \rho v \\ \rho w \\ e \end{bmatrix}_{\cdot,j} \right]_{i\pm\frac{1}{2}} \right)$$

The Roe matrix, e.g. if $0 < u < \frac{c}{\epsilon}$, can be written as

$$|J| = \frac{1}{\epsilon} \begin{pmatrix} 0 & 0 & 0 & \frac{\gamma-1}{c} \\ -cu & c & 0 & \frac{2u(\gamma-1)}{c} \\ 0 & 0 & 0 & \frac{(\gamma-1)v}{c} \\ 0 & 0 & 0 & c \end{pmatrix} + \begin{pmatrix} u & 0 & 0 & u\frac{1-\gamma}{c^2} \\ u^2 & 0 & 0 & -\frac{u^2(\gamma-1)}{c^2} \\ 0 & 0 & u & -\frac{(\gamma-1)uv}{c^2} \\ 0 & 0 & 0 & 0 \end{pmatrix} + \mathcal{O}(\epsilon) \tag{5.4}$$
$$=: \frac{1}{\epsilon}\mathcal{A}_{-1} + \mathcal{A}_0 + \mathcal{O}(\epsilon)$$

Taking now the y-direction into account as well, one can collect order by order in ϵ (compare [GV99, GM04]):

$$\mathcal{O}\left(\frac{1}{\epsilon^2}\right) : p^{(0)}_{i+1,j} - p^{(0)}_{i-1,j} = 0 \tag{5.5}$$

$$\mathcal{O}\left(\frac{1}{\epsilon}\right) : \begin{pmatrix} 0 \\ p^{(1)}_{i+1,j} - p^{(1)}_{i-1,j} \\ p^{(1)}_{i,j+1} - p^{(1)}_{i,j-1} \\ 0 \end{pmatrix} + \begin{pmatrix} [\frac{\gamma-1}{c^{(0)}}[e^{(0)}]]_{i\pm\frac{1}{2},j} \\ -[c^{(0)}u^{(0)}[\rho^{(0)}]]_{i\pm\frac{1}{2},j} + [c^{(0)}[\rho^{(0)}u^{(0)}]]_{i\pm\frac{1}{2},j} + [\frac{2u^{(0)}(\gamma-1)}{c^{(0)}}[e^{(0)}]]_{i\pm\frac{1}{2},j} \\ [\frac{(\gamma-1)v^{(0)}}{c^{(0)}}[e^{(0)}]]_{i\pm\frac{1}{2},j} \\ [c^{(0)}[e^{(0)}]]_{i\pm\frac{1}{2},j} \end{pmatrix} + \text{perp. terms} = 0$$

Additionally one has

$$\mathcal{O}(1) : \frac{1}{2\Delta x}\left(u^{(0)}_{i+1,j} - u^{(0)}_{i-1,j}\right) + \frac{1}{2\Delta y}\left(v^{(0)}_{i,j+1} - v^{(0)}_{i,j-1}\right) - [c^{(0)}[p^{(1)}]]_{i\pm\frac{1}{2}} - [c^{(1)}[p^{(0)}]]_{i\pm\frac{1}{2}} + \text{perp. terms} = 0$$

In the literature (e.g. in [GV99] and many others) often the discussion stops at the $\mathcal{O}(\epsilon^{-1})$ equations, and the inability of the Roe scheme to resolve the low Mach number limit is attributed to the discrete second derivatives that follow $p^{(1)}_{i+1,j} - p^{(1)}_{i-1,j}$. To show the lack of low Mach compliance one needs to show that a vanishing of these terms violates the continuous limit equations. This might make a careful study of further orders necessary, but it might indeed suffice to show a contradiction at some order in ϵ, when showing that a numerical scheme is *not* low Mach compliant.

5.2 Implications from linear acoustics

5.2.1 Asymptotic analysis

The limit equations of a numerical scheme for both linear acoustics and the Euler equations are supposed to be a discretization of $\operatorname{div} \mathbf{v} = 0$. Schemes that lack low Mach compliance in the limit discretize only a subspace of all the divergenceless flows. This is well understood using the concept of stationarity preservation – they do not discretize all the limit equations. This can also be shown using asymptotic analysis. For the acoustic equations it is performed for the upwind/Roe scheme in Theorem 4.9: additionally to the divergence constraint there appear equations that restrict the limit to only shear flows. For a low Mach compliant scheme on the other hand one has to show that none of the equations that appear in an expansion in powers of ϵ restricts the set of divergenceless flows to only a subset. Using asymptotic analysis, this turns out to be very difficult, as the equations of one order generally are coupled to higher orders. Recall how the argumentation for the multi-dimensional scheme in Section 4.6 relies on deliberately ignoring certain equations. By ignoring certain equations one is left with the discomfort of an incomplete argument. Considering every order in the expansion at the same time seems unfeasible.

The following theorem relates to the observation that often, e.g. in [GV99, BEK+17] the asymptotic analysis is restricted to a consideration of only the equations to order $\mathcal{O}(\epsilon^{-2}$ and $\mathcal{O}(\epsilon^{-1})$. This might already suffice for proving a violation of low Mach compliance; otherwise it is necessary to consider higher, and in fact all orders.

Theorem 5.1. *To prove that the discretizations of the Euler or acoustics equations are low Mach compliant one has to study further equations in the asymptotic expansion than those to orders $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-1})$.*

Proof. It is sufficient to show the statement for the acoustic equations, as they govern the linearized regime of the Euler equations.

Consider the following three schemes. The equation to order $\mathcal{O}(\epsilon^{-1})$ contained in the upwind/Roe scheme (Equation (4.37)) is:

$$0 = [p^{(1)}]_{i\pm1,j} - c[[u^{(0)}]]_{i\pm\frac{1}{2},j}$$

The Roe scheme is not low Mach compliant, and (e.g. in [GV99]) this is attributed to the presence of the non-vanishing right hand side.

In Section 4.6 the multi-dimensional scheme (4.32) formally contains the Equation (4.45)

$$0 = \frac{1}{8\Delta x}\{\{[p^{(1)}]_{i\pm1}\}\}_{j\pm\frac{1}{2}} - c\left(\frac{1}{8\Delta x}\{\{[[u^{(0)}]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + \frac{1}{8\Delta y}[[v^{(0)}]_{i\pm1}]_{j\pm1}\right) \quad (5.6)$$

This scheme is stationarity preserving and low Mach compliant.

The modification of the multi-dimensional scheme (4.32) that is shown in Figure 4.12 (right) contains

$$0 = \frac{1}{8\Delta x}\{\{[p^{(1)}]_{i\pm1}\}\}_{j\pm\frac{1}{2}} - c\left(\frac{1}{2\Delta x}[[u^{(0)}]]_{i\pm\frac{1}{2},j} + \frac{1}{2\Delta y}[[v^{(0)}]_{i\pm1}]_{j\pm1}\right) \tag{5.7}$$

It is clear by Corollary 4.10 (and obvious in Figure 4.12) that this latter scheme is not stationarity preserving, and thus not low Mach compliant.

The argument why Equation (5.7) implies that the corresponding scheme is not low Mach compliant, and that at the same time Equation (5.6) implies low Mach compliance requires information about the behaviour of the right hand sides. A careful analysis of the argumentation provided in Section 4.6 shows that the interplay with the divergence appearing in one of the $\mathcal{O}(1)$ equations is responsible for low Mach compliance in case of Equation (5.6). It thus does not suffice to only look at the $\mathcal{O}(\epsilon^{-1})$ and $\mathcal{O}(\epsilon^{-2})$. □

When showing low Mach compliance with only the pressure gradients, a clear sign that something must be missing is that the discrete divergence constraint is absent from the argumentations – although it is an equally important part of the limit equations. Stationarity preservation, on the other hand, involves all the limit equations on an equal footing. This is another instance where the study of numerical schemes for linear acoustics is of huge value: stationarity preservation is an alternative way of proving low Mach compliance, which allows to detect deficiencies in arguments based entirely on asymptotic analysis.

Unfortunately, stationarity preservation cannot easily be checked in the nonlinear setting, even if one can think of a extending its definition to nonlinear schemes (see Section 5.4). Therefore asymptotic analysis seems inevitable. Arguments have to rely on a careful choice of equations and a particular way of argumentation. The only reason why it is used in the following sections is that it is hugely successful in selecting low Mach compliant schemes in practice. This however also may be due to it having been applied so far to only a particular kind of schemes. Therefore here care is taken to explicitly pinpointing deficiencies in argumentations based on asymptotic analysis. Maybe in future one will be able to fill the gaps and will understand why asymptotic analysis is successful in practice.

5.2.2 Linearization

Upon linearization around a static state $\rho = \text{const}$, $p = \text{const}$, $\mathbf{v} = 0$ any numerical scheme for the Euler equations becomes a numerical scheme for the acoustic equations (2.40)–(2.41). A necessary condition for the nonlinear scheme to be low Mach compliant is that the linearization is stationarity preserving. Denote by q the vector of conserved quantities.

Definition 5.2 (Roe-type scheme)**.** *A dimensionally split finite volume scheme with fluxes*

$$f^x_{i+\frac{1}{2},j} = \frac{1}{2}\{f(q)\}_{i+\frac{1}{2},j} - \frac{1}{2}D_x[q]_{i+\frac{1}{2},j}$$
$$f^y_{i,j+\frac{1}{2}} = \frac{1}{2}\{f(q)\}_{i,j+\frac{1}{2}} - \frac{1}{2}D_y[q]_{i,j+\frac{1}{2}}$$

with diffusion matrices D_x, D_y is called Roe-type scheme.

Roe-type schemes scheme amounts to a central flux and a diffusion. An example of such a scheme is the Roe scheme with $D_x = |J_x|$.

A linearization of such schemes around the static state $\rho = \text{const}$, $p = \text{const}$, $\mathbf{v} = 0$ makes appear a dimensionally split numerical scheme for the acoustic equations. Conditions on this scheme to be stationarity preserving (formulated in Theorem 4.8) are necessary conditions on the Roe-type scheme to be low Mach compliant.

Theorem 5.2. *A low Mach compliant Roe-type scheme for* (1.14)–(1.16) *has diffusion matrices that contain scaling $\mathcal{O}(\epsilon^{-2})$ in the energy-column, and otherwise scale as $\mathcal{O}(1)$ as $\epsilon \to 0$.*

Proof. Linearization amounts to keeping the diffusion matrices constant. Theorem 4.8 requires a vanishing diagonal entry in the momentum-column. It thus remains to be shown that if it is $\mathcal{O}(1)$ before linearization, then it is proportional to $\mathbf{v}$ and vanishes upon linearization around a static state.

Call the diagonal matrix element in the x-momentum-row A. This element multiplies a jump in the x-momentum and the difference between two of those, after dividing by Δx, or Δy, contributes to the time derivative of the x-momentum again. Symbolically:

$$\partial_t(\rho u) \simeq \frac{1}{2\Delta x}[A[\rho u]]_{i\pm\frac{1}{2}} \tag{5.8}$$

A obviously has units of a velocity. The only two velocities available are $\mathbf{v}$ and c, because the Mach number and the Strouhal number are the only non-dimensional numbers that can be obtained in the setting of the Euler equations, and the Strouhal number involves independent variables that cannot appear in the (constant) linearizations of the diffusion matrices.

Theorem 1.1 establishes the most general scalings that lead from the Euler equations (1.8)–(1.10) to the rescaled equations (1.14)–(1.16) that make appear ϵ. Applying those scalings to Equation (5.8) means that if A scales as c

$$\partial_t(\rho u) \simeq \frac{1}{2\Delta x}\left[\frac{A}{\epsilon}[\rho u]\right]_{i\pm\frac{1}{2}}$$

and thus is not $\mathcal{O}(1)$. If, however, A scales as $\mathbf{v}$ upon rescaling no powers of ϵ appear:

$$\partial_t(\rho u) \simeq \frac{1}{2\Delta x}\left[A[\rho u]\right]_{i\pm\frac{1}{2}}$$

One thus concludes that if $A \in \mathcal{O}(1)$ then it is proportional to $\mathbf{v}$ and vanishes upon linearization around the static state.

The other entries of the diffusion matrix can be treated in the same manner. □

This argument only provides a necessary condition for a nonlinear scheme to be low Mach compliant. A number of low Mach fixes have been suggested in the literature, and they all can be checked to fulfill this condition (see e.g. [LG13]). In view of the variety of different suggestions one might wonder whether maybe this condition also is sufficient. Indeed, pretty much every modification that respects the conditions of Theorem 5.2 seems to lead to a low Mach compliant scheme in practice – as long as it leads to a stable scheme.

Consider a dimensionally split numerical scheme for the Euler equations in two spatial dimensions

$$\partial_t q + [f^x(q)]_{i\pm1,j} + [f^y]_{i,j\pm1} - [D_x[q]]_{i\pm\frac{1}{2}} - [D_y[q]]_{i,j\pm\frac{1}{2}} = 0$$

with $q = (\rho, \rho u, \rho v, e)$, f^x, f^y as in (1.14)–(1.16). D_x, D_y are diffusion matrices which in the limit $\epsilon \to 0$ scale as

$$\begin{pmatrix} \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(\epsilon^{-2}) \\ \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(\epsilon^{-2}) \\ \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(\epsilon^{-2}) \\ \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(\epsilon^{-2}) \end{pmatrix} \tag{5.9}$$

Additionally, one assumes that $\partial_t p^{(0)} = 0$. This depends on the boundary conditions, and may also depend on certain properties of the scheme (see [GV99]). Once constance this is established, the equations that govern $p^{(1)}$ necessarily are just the same ones, because there are no terms $\mathcal{O}(\epsilon^{-1})$. This is normally (see e.g. [GV99]) taken as sufficient an argument in favor of low Mach compliance of the scheme. It is at the same time absolutely unclear whether the discrete velocities in the limit discretize all of the divergenceless flows, or not. This is why this argument, although appearing often in literature, seems incomplete. If one tries to study, say the $\mathcal{O}(1)$ equations, clearly the pressure $p^{(2)}$ would appear, and so on, with every order coupling to the next higher. Another obvious shortcoming of such an analysis has been mentioned in [Del10] – although the low Mach number problem appears only in multiple spatial dimensions, it is not obvious how this plays a role in this argumentation.

A simplified view why the energy column of diffusion matrices is allowed to have terms scaling as $\mathcal{O}(\epsilon^{-2})$ is the following. Consider well-prepared initial data in the pressure, i.e. assume that the numerical data satisfy $\nabla p \in \mathcal{O}(\epsilon^2)$. As the diffusion matrices multiply jumps, the matrix elements scaling as $\mathcal{O}(\epsilon^{-2})$ hit the jump in the energy (or pressure) which is $\mathcal{O}(\epsilon^2)$. On total this gives a diffusion term that scales $\mathcal{O}(1)$. However there are several inconsistencies in this argumentation. First of all, as the divergence constraint has not been mentioned, this cannot be a complete argument. Also it is not clear, what happens after the first time step.

5.3 Dimensionally split low Mach compliant schemes

5.3.1 Low Mach number modifications

There exist a variety of modifications of the Roe scheme that improve the behaviour of numerical schemes in the limit of low Mach number. Focusing on those that can be integrated explicitly in time, they are presented in [Tur87, KLN91, GV99, LG08, TD08, TMD+08, Rie11, LG13, MRE15, OSB+16, DJOR16] and others. As such modifications change the diffusion of the scheme, its stability properties might be modified. In order to perform a linear stability analysis, results of Section 4.3.5 can be useful.

Here the following type of modification shall be analyzed in detail ([Tur87, BEK+17]). Replacing $|J_x|$ in (5.3) by $D_x = P_x^{-1}|P_x J_x|$ with P_x from [WS95]

$$P_x = \begin{pmatrix} 1 & 0 & 0 & \frac{\mu^2-1}{c^2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \mu^2 \end{pmatrix} \tag{5.10}$$

$$D_x = \frac{1}{\epsilon^2\sqrt{4c^2+u^2}} \begin{pmatrix} 0 & 0 & 0 & 2(\gamma-1) \\ 0 & 0 & 0 & 3(\gamma-1)u \\ 0 & 0 & 0 & 2(\gamma-1)v \\ 0 & 0 & 0 & 2c^2 \end{pmatrix} + \mathcal{O}(1)$$

given here in the basis of primitive variables. The parameter μ is given by

$$\mu = \min[1, \max(M_{\text{loc}}, M_{\text{cut}})]$$

M_{loc} is the local Mach number and M_{cut} avoids singularity of the matrix. Its value should be chosen smaller than the smallest expected relevant Mach numbers of the flow. For historical reasons, this modification has been called *preconditioning*. Here this name will not be used as it can be misleading. For high Mach numbers, the absolute value becomes the identity, and the scheme reverts back to the usual Roe scheme.

The performance of this scheme in the context of homogeneous equations (1.14)–(1.16) has been analyzed, among others, in [GV99].

This scheme satisfies the necessary condition of Theorem 5.2.

5.3.2 Low Mach number modifications in presence of gravity

When dealing with the homogeneous Euler equations, in the limit $\epsilon \to 0$ the pressure is constant up to perturbations $\mathcal{O}(\epsilon^2)$. This was the reason why terms $\mathcal{O}(\epsilon^{-2})$ do not pose a "problem" if they appear in the energy column of a diffusion matrix. Once the Euler equations are augmented by gravity source terms as in (1.27)–(1.29) (with $Fr = \epsilon$), the pressure (and also the energy) is not spatially constant, but varies according to (1.30). Keeping the terms $\mathcal{O}(\epsilon^{-2})$ or $\mathcal{O}(\epsilon^{-1})$ in the density and energy rows now changes the asymptotic behaviour of the scheme in the combined limit $\epsilon \to 0$, $Fr = \epsilon$. This Section is based on work published in [BEK+17, BEKR17].

Theorem 5.3. *Consider a Roe-type scheme for the Euler equations with gravity* (1.27)–(1.29)

$$\partial_t q_{ij} + \frac{1}{\Delta x}[f^x(q)]_{i\pm1,j} + \frac{1}{\Delta y}[f^y]_{i,j\pm1} - \frac{1}{2\Delta x}[D_x[q]]_{i\pm\frac{1}{2},j} - \frac{1}{2\Delta y}[D_y[q]]_{i,j\pm\frac{1}{2}} = s(q_{ij})$$

with $q = (\rho, \rho u, \rho v, e)$, f^x, f^y *as in* (1.27)–(1.29) *and* $s(q)$ *the gravity source*

$$s = (0, \rho g_x, \rho g_y, \rho \mathbf{v} \cdot \mathbf{g})$$

and $g_x \neq 0$, $g_y \neq 0$.

D_x, D_y *are diffusion matrices which in the limit* $\epsilon \to 0$ *scale as in Equation* (5.9) *with the diagonal elements in the energy-row denoted by* d_x^e, d_y^e *satisfying*

$$\partial_x(d_x^e \rho g_x) + \partial_y(d_y^e \rho g_y) \neq 0 \tag{5.11}$$

Then the scheme in the limit $\epsilon \to 0$, $Fr = \epsilon$ *formally has solutions which fulfill*

$$\frac{1}{2\Delta x}[d_x^e[p^{(0)}]]_{i\pm\frac{1}{2},j} + \frac{1}{2\Delta y}[d_y^e[p^{(0)}]]_{i,j\pm\frac{1}{2}} = 0 \tag{5.12}$$

which is not a discretization of the limit equation (1.30).

Proof. The assertion immediately follows from the energy row by considering orders $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-1})$. The terms $(\dots)$ are the corresponding entries of the matrix, and (5.12) discretizes

$$\Delta x \partial_x(d_x^e \partial_x p^{(0)} + \Delta y \partial_y(d_y^e \partial_y p^{(0)}) = 0$$

which by condition (5.11) is not consistent with Equation (1.30). □

Consider replacing the diffusion matrix by $P^{-1}|PJ|$ as in Equation 5.10. Additionally to the wrong limit of Theorem 5.3, applying this scheme to any stationary initial data which contain a non-constant pressure gradient, introduces terms which scale with the inverse of M_{cut} (see e.g. [Mic13, BEK+17]). Therefore the numerical errors strongly depend on an arbitrary parameter, and are unacceptable for low values of M_{cut}. In practice, the simulations crash after very short times.

To correct the behaviour of schemes that use P_x as given in (5.10), in [MRE15] a different matrix P_x has been suggested. In entropy variables it takes the form

$$P_{x,\text{entr}} = \begin{pmatrix} 1 & \delta & 0 & 0 \\ -\delta & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

with $\delta = \frac{1}{\min(1,\max(M_{\text{loc}},M_{\text{cut}}))} - 1$. In primitive variables it is

$$P_{x,\text{prim}} = \begin{pmatrix} 1 & \frac{\rho\delta\epsilon}{c} & 0 & 0 \\ 0 & 1 & 0 & -\frac{\delta}{\rho c\epsilon} \\ 0 & 0 & 1 & 0 \\ 0 & \rho c\delta\epsilon & 0 & 1 \end{pmatrix} \tag{5.13}$$

The definition of δ ensures that the scheme reverts back to the original Roe scheme when the local Mach number reaches 1.

In x-direction, the diffusion matrix then reads

$$D_x = \frac{1}{\epsilon^2} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \gamma - 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \mathcal{O}(1). \tag{5.14}$$

Observe that now the $\mathcal{O}(1/\epsilon^2)$ terms have disappeared from both the pressure and the density row. Thus the arguments of Theorem 5.3 cannot make this scheme fail in the combined limit $\epsilon \to 0$, $Fr \in \mathcal{O}(\epsilon)$, although as usual a satisfactory proof cannot be given by means of asymptotic analysis. The equation at order $\mathcal{O}(\epsilon^{-2})$, e.g. in x-direction, becomes

$$\begin{aligned} \frac{1}{2\Delta x}\Big([p^{(0)}]_{i\pm1} - (\gamma-1)[[e^{(0)}]]_{i\pm\frac{1}{2}}\Big) &= (\rho g_x)_i^{(0)} \\ \frac{p_i - p_{i-1}}{2\Delta x} &= (\rho g_x)_i^{(0)} \end{aligned}$$

which is a consistent discretization of the limit equation (1.30). Observe that this scheme can still be applied to the homogeneous Euler equations as no new inverse powers of ϵ has been added. Contrary to D_x as given in (5.10), additionally the new matrix in (5.14) has a finite limit for $M_{\text{cut}} \to 0$, although this parameter is still needed in the definition of P_x. Linear stability of this scheme has been studied in Section 4.3.5. There, a scaling of the CFL number with ϵ^2 rather than ϵ has been found, which can be confirmed experimentally. The original scheme (5.10) is known to perform in the same way ([BM05]). Experimentally, it has been found that integrating the scheme by a third-order Runge-Kutta scheme has a stabilizing influence, such that in practice the CFL condition is not found to be as strict.

Note that the correct asymptotic scaling in general does not guarantee that for finite resolution the numerical solution will be close to the analytical one. The situation is somewhat easier in the homogeneous case, where any solution to the equation $\nabla p^{(0)} = 0$ can be represented exactly on a numerical grid. A discrete version of an exact solution to $\nabla p^{(0)} = \rho^{(0)}\mathbf{g}^{(0)}$ with $\mathbf{v} = 0$ will in general not remain stationary in a numerical simulation because of the mismatch between the exact derivative and the way its numerical approximation is obtained from adjacent cell averages. This behaviour has been

briefly mentioned in Section 4.8 already. The correct asymptotics however means that the errors do not increase without bound in the limit of small Mach numbers. If, in addition to the correct scaling in the vicinity of a hydrostatic equilibrium, one wishes to be able to maintain the equilibrium itself exactly stationary (up to machine precision), a specific discretization (well-balancing) of the source term is necessary (e.g. [CK15]). This, however, is not subject of this thesis.

As a test an isothermal hydrostatic equilibrium of an ideal gas with $\gamma = 1.4$ is chosen, given by (1.26) and $p/\rho = 1$. The exact solution for this setup is stationary. The test is performed in one spatial dimension with gravity pointing towards negative values of the spatial coordinate with a computational domain of $[0, 1]$. It is discretized with $N \in \{50, 100, 200, 400\}$ cells. A Runge-Kutta scheme of 3^{rd} order and a piecewise constant reconstruction are used. The values in the ghost cells are fixed to their initial values. The numerical results are displayed in Fig. 5.1, where the temperature p/ρ, is shown as a function of time at a fixed position of $x = 0.85$. An oscillation is observed, which does not change significantly with time, but whose amplitude decreases with spatial resolution. Its origin is the fact that the numerical discretization of the pressure gradient is not exactly balanced by the numerical treatment of the source term, which is evaluated in a cell-centered manner. The residual acts as a perpetual excitation of the atmosphere to which it answers by oscillation. The characteristic frequency is of the order of the Brunt-Väisälä frequency of this setup. As expected, the numerical errors in the approximation of the gradient and therefore the perturbations decrease with the resolution.

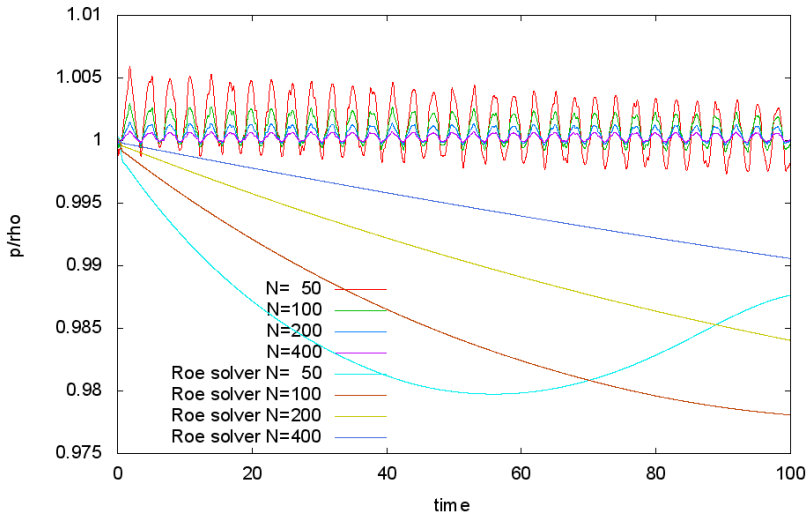

Figure 5.1: Numerical evolution for the initial data of an isothermal, stationary atmosphere. p/ρ, which is proportional to the temperature, is shown as a function of time at a fixed location of $x = 0.85$. For comparison, additionally to the results computed with the new scheme, the evolution of the same setup with the Roe solver is shown.

For comparison, the Figure shows the time evolution of the same initial data of the isothermal equilibrium with the unmodified Roe solver. It suffers from a similar kind of

mismatch between the numerical approximation to the gradient and the source term, but the perturbations experience the strong diffusion which damps oscillations and manifests itself in a rapid decrease of the temperature. In principle, again the deviations can be controlled by increasing the resolution although the diffusive character persists.

The scheme given by the diffusion matrix (5.13) can also be used to numerically solve the homogeneous Euler equations. Consider the Gresho vortex [GC90]. This is an example of a stationary, incompressible rotating flow around the origin in two spatial dimensions. Denoting by $\mathbf{e}_\varphi$ the unit vector in φ-direction and with $r = \sqrt{x^2 + y^2}$

$$\mathbf{v} = \mathbf{e}_\varphi \cdot \begin{cases} 5r & r < 0.2 \\ 2 - 5r & r < 0.4 \\ 0 & \text{else} \end{cases} \tag{5.15}$$

$$p = \begin{cases} p_\mathrm{c} + \frac{25}{2} r^2 & r < 0.2 \\ p_\mathrm{c} + 4\ln(5r) + 4 - 20r + \frac{25}{2} r^2 & r < 0.4 \\ p_\mathrm{c} + 4\ln 2 - 2 & \text{else} \end{cases} \tag{5.16}$$

with the uniform density $\rho = 1$ and the pressure in the vortex center $p_\mathrm{c} = \frac{1}{\gamma \epsilon^2} - \frac{1}{2}$.

In the compressible setting the flow can be endowed with different maximum Mach numbers by varying the parameter ϵ in the value of the central pressure. Therefore this is an example of a family of solutions, parametrized by a real number ϵ, such that M_loc scales asymptotically as ϵ in the limit $\epsilon \to 0$. Such families of solutions are described in Section 1.2. One observes for example that $p = \frac{1}{\epsilon^2}(\tilde{p}^{(0)} + \epsilon \tilde{p}^{(1)} + \ldots)$. Using notation of Section 1.2, the asymptotic scalings here correspond to the following choice of the parameters:

$$\mathfrak{a} = 0 \qquad \mathfrak{b} = 0 \qquad \mathfrak{d} = 0 \qquad \mathfrak{c} = -2$$

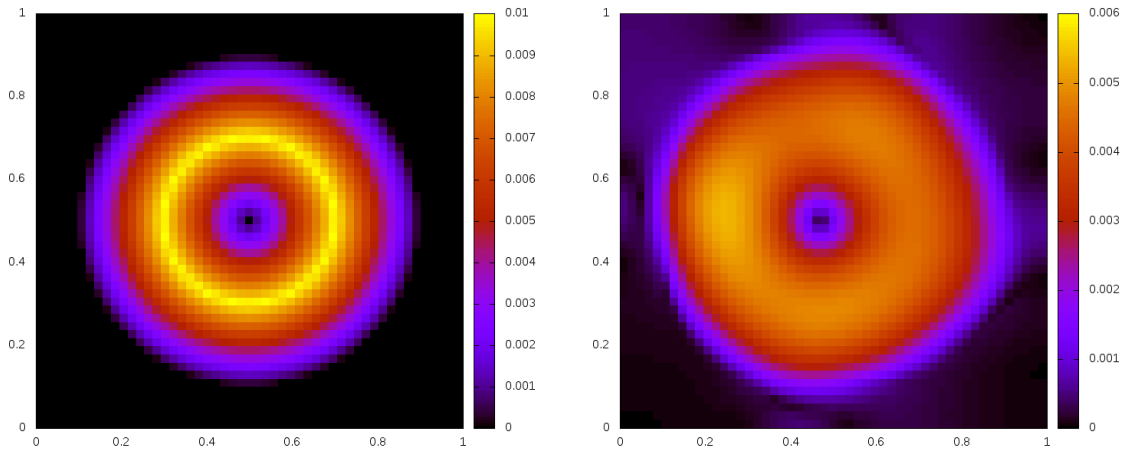

Figure 5.2: Numerical simulation using the scheme described in Section 5.3.2. The initial data are given in Equations (5.15)–(5.16) with $\epsilon = 10^{-2}$. The homogeneous Euler equations (1.14)–(1.15) with $\gamma = 1.4$ are solved on a square 50×50 grid with a Runge-Kutta scheme of third order. Colour coded is the Mach number. *Left*: Initial setup. *Right*: Simulations result at $t = 2$. Commit hash: 5498c5d.

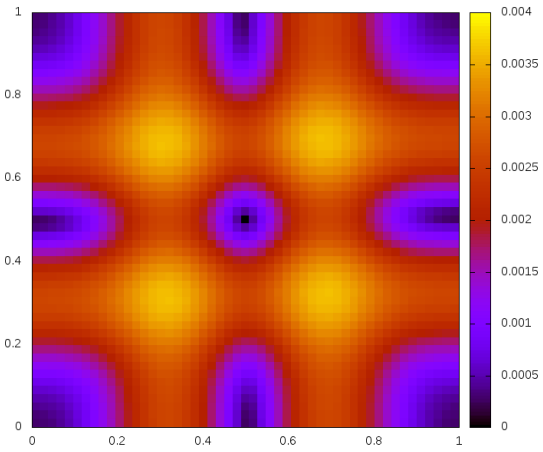

Figure 5.3: Numerical simulation using the Roe scheme (Equation (5.3)). The initial data are given in Equations (5.15)–(5.16) and shown in Figure 5.2 (left). The Euler equations (1.14)–(1.16) with $\gamma = 1.4$ and $\epsilon = 10^{-2}$ are solved on a square 50×50 grid and the result shown at $t = 2$. Colour coded is the Mach number. *Top left*: Initial setup. Commit hash: b428053.

5.4 Stationarity preserving schemes

Stationarity preservation is very efficient in explaining various properties of numerical schemes for linear acoustics (Section 4.5). It is tempting to try to generalize this notion to nonlinear equations. Linearity is not a prerequisite for the existence of stationary states. For linear acoustics the Fourier transform is used to study (numerical) stationary states, which is not directly available for nonlinear problems. It thus might at first be unclear whether any result at all can be reused.

In fact, as is demonstrated below, a number of results carry over to the nonlinear setting. However, clearly, the available tools shrink when dealing with nonlinear equations. *Checking* whether a given nonlinear scheme is stationarity preserving is not possible with the methods presented so far, and seems a very complicated task. This thesis is unable to cover all issues and to construct a theory for the nonlinear case that would be as complete as the one for the linear case is. Carefully tailoring the setup, such that results from Section 4.5 (that were formulated for linear acoustics) can be used, allows to *construct* non-linear schemes. Several nonlinear schemes are constructed in Sections 5.4.1 and 5.4.2, and a novel kind of low Mach number modification is shown in Section 5.5. They demonstrate new construction principles and show satisfactory experimental results, although further investigations are necessary. Hopefully they pave the way towards a deeper understanding of multi-dimensional finite volume schemes for the Euler equations.

The schemes in this Section are based on the following observation, expressed again with continuous operators: if stationary states of a nonlinear PDE

$$\partial_t q + \partial_x f^x(q) + \partial_y f^y(q) = 0 \qquad \begin{aligned} q &: \mathbb{R}_0^+ \times \mathbb{R}^2 \to \mathbb{R}^n \\ f^x, f^y &: \mathbb{R}^n \to \mathbb{R}^n \end{aligned} \tag{5.17}$$

are governed by

$$\partial_x f^x(q) + \partial_y f^y(q) = 0$$

then adding a diffusion of the type

$$\partial_x^2 f^x(q) + \partial_x \partial_y f^y(q) = 0$$

does not change the stationary states.

By Corollary 4.10 and Section 3.2.2.1, stationarity consistency allows to find a discrete counterpart to such a statement: vanishing flux divergence

$$\{\{[f^x(q)]_{i\pm1}\}\}_{j\pm\frac{1}{2}} + [\{\{f^y(q)\}\}_{i\pm\frac{1}{2}}]_{j\pm1} = 0$$

implies vanishing of

$$\{\{[[f^x(q)]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} + [[f^y(q)]_{i\pm1}]_{j\pm1} = 0$$

Note that the way f^x and f^y depend on q does not enter! They can well be nonlinear functions.

This motivates the following definition:

Definition 5.3 (Stationarity preserving)**.** *A consistent numerical scheme for* (5.17) *is called* stationarity preserving *if its stationary states discretize all the analytic stationary states.*

5.4.1 Scalar-vector systems

The idea of the multi-dimensional scheme (4.32) is to make the diffusion of the velocity $\mathbf{v}$ be proportional to grad div $\mathbf{v}$, and to take the diffusion of p proportional to div grad p. This means that the diffusion of a vectorial quantity is a vector (a gradient even), and the diffusion of a scalar quantity a scalar. Also it is in a sense diagonal, as the diffusion of $\mathbf{v}$ is a second derivative of $\mathbf{v}$, and the diffusion of p is a second derivative of p.

Consider the following generalization of the acoustic system in two spatial dimensions:

Definition 5.4 (Scalar-vector system)**.** *The following system of conservation laws in two spatial dimensions is called* scalar-vector system

$$\begin{aligned}\partial_t p + \partial_x f^x + \partial_y f^y &= 0\\ \partial_t u + \partial_x \pi^{xx} + \partial_y \pi^{yx} &= 0\\ \partial_t v + \partial_x \pi^{xy} + \partial_y \pi^{yy} &= 0\end{aligned}$$

or in shorter notation (with $\mathbf{v} = (u, v)^T$*):*

$$\partial_t p + \operatorname{div} \mathbf{f} = 0 \tag{5.18}$$

$$\partial_t \mathbf{v} + \operatorname{div} \pi = 0 \tag{5.19}$$

where $\mathbf{f} = (f^x, f^y)^T$ *and* $\pi = \begin{pmatrix} \pi^{xx}, \pi^{xy} \\ \pi^{yx}, \pi^{yy} \end{pmatrix}$ *are some given, not necessarily linear, functions of* p*,* u *and* v*.*

It describes the time evolution of a scalar quantity $p : \mathbb{R}_0^+ \times \mathbb{R}^2 \to \mathbb{R}$ *and a vector-valued quantity* $\mathbf{v} : \mathbb{R}_0^+ \times \mathbb{R}^2 \to \mathbb{R}^2$*.*

The acoustic system is an example of such a scalar-vector system and is obtained by choosing

$$f^x = c^2 u \qquad f^y = c^2 v$$

$$\pi^{xx} = \pi^{yy} = \frac{p}{\epsilon^2} \qquad \pi^{xy} = \pi^{yx} = 0$$

Isentropic hydrodynamics (1.6)–(1.7) can be found by taking

$$\begin{aligned}\mathbf{f} &= \rho \mathbf{v}\\ \pi &= \rho \mathbf{v} \otimes \mathbf{v} + p \mathbb{1}\end{aligned}$$

The natural stationarity preserving diffusion would be

$$\begin{aligned}\partial_t p + \partial_x f^x + \partial_y f^y &= c_1 \Delta x (\partial_x^2 \pi^{xx} + \partial_x \partial_y (\pi^{xy} + \pi^{yx}) + \partial_y^2 \pi^{yy})\\ \partial_t u + \partial_x \pi^{xx} + \partial_y \pi^{yx} &= c_2 \Delta x \partial_x (\partial_x f^x + \partial_y f^y)\\ \partial_t v + \partial_x \pi^{xy} + \partial_y \pi^{yy} &= c_2 \Delta x \partial_y (\partial_x f^x + \partial_y f^y)\end{aligned}$$

It computes the scalar second derivative by acting with div div onto π. The vectorial second derivative is obtained by computing grad div $\mathbf{f}$. For the acoustic system this procedure reduces to the aforementioned operators div grad and grad div.

Observe that this diffusion (once it has been reproduced in a discrete setting) will lead to a nonlinear stationarity preserving scheme.

Theorem 5.4 (Stationarity preserving scheme for scalar-vector system)**.** *The following scheme for Equations* (5.18)–(5.19) *is stationarity preserving:*

$$\begin{aligned}
&\partial_t p + \frac{1}{8\Delta x}\{\{[f^x]_{i\pm1}\}\}_{j\pm\frac12} + \frac{1}{8\Delta y}[\{\{f^y\}\}_{i\pm\frac12}]_{j\pm1} \qquad (5.20)\\
&\quad = c_1\left(\frac{1}{4\Delta x}\{\{[[\pi^{xx}]]_{i\pm\frac12}\}\}_{j\pm\frac12} + \frac{1}{4\Delta y}[[\pi^{yx}]_{j\pm1}]_{i\pm1} + \frac{1}{4\Delta x}[[\pi^{xy}]_{i\pm1}]_{j\pm1} + \frac{1}{4\Delta y}\{\{[[\pi^{yy}]]_{j\pm\frac12}\}\}_{i\pm\frac12}\right)
\end{aligned}$$

$$\begin{aligned}
&\partial_t u + \frac{1}{8\Delta x}\{\{[\pi^{xx}]_{i\pm1}\}\}_{j\pm\frac12} + \frac{1}{8\Delta y}\{\{[\pi^{yx}]_{j\pm1}\}\}_{i\pm\frac12}\\
&\quad = c_2\left(\frac{1}{4\Delta x}\{\{[[f^x]]_{i\pm\frac12}\}\}_{j\pm\frac12} + \frac{1}{4\Delta y}[[f^y]_{i\pm1}]_{j\pm1}\right)\\
&\partial_t v + \frac{1}{8\Delta x}\{\{[\pi^{xy}]_{i\pm1}\}\}_{j\pm\frac12} + \frac{1}{8\Delta y}\{\{[\pi^{yy}]_{j\pm1}\}\}_{i\pm\frac12}\\
&\quad = c_2\left(\frac{1}{4\Delta x}[[f^x]_{i\pm1}]_{j\pm1} + \frac{1}{4\Delta y}[[\{\{f^y\}\}_{i\pm\frac12}]]_{j\pm\frac12}\right) \qquad (5.21)
\end{aligned}$$

with arbitrary c_1, c_2 (that are allowed to be function of p, u, v).

Proof. As the structure of Equations (5.20)–(5.21) is that of divergence operators, the discrete identity of Corollary 4.10 can be applied three times. □

The coefficients c_1 and c_2 do not follow from these considerations and have to be found from further conditions.

Obviously the operator $\frac14\{\{\cdot\}\}_{j\pm\frac12}$ becomes the identity when one-dimensional problems in x-direction are treated. Therefore Eqns. (5.20)–(5.21) become

$$\partial_t p + \frac{1}{2\Delta x}[f^x]_{i\pm1} = c_1 \cdot \frac{1}{\Delta x}[[\pi^{xx}]]_{i\pm\frac12} \qquad (5.22)$$

$$\partial_t u + \frac{1}{2\Delta x}[\pi^{xx}]_{i\pm1} = c_2 \cdot \frac{1}{\Delta x}[[f^x]]_{i\pm\frac12} \qquad (5.23)$$

Theorem 5.5 (Stability)**.** *In one spatial dimension, consider the scheme* (5.22)–(5.23) *upon linearization*

$$\begin{aligned}
f^x &= a_1 u + a_2 p\\
\pi^{xx} &= a_3 u + a_4 p
\end{aligned}$$

with arbitrary constants a_1, a_2, a_3, a_4, $a_4 \neq 0$, $a_1 \neq 0$. If $a_3 = a_2 =: a$ and $|a| < \sqrt{a_1 a_4}$, it reduces to the upwind/Roe scheme if

$$c_1 = \frac12\sqrt{\frac{a_1}{a_4}} \qquad\qquad c_2 = \frac12\sqrt{\frac{a_4}{a_1}}$$

Proof. With the linearization the scheme (5.22)–(5.23) becomes

$$\partial_t p + \frac{1}{2\Delta x}[a_1 u + a_2 p]_{i\pm 1} = c_1 \cdot \frac{1}{\Delta x}\left(a_3 [[u]]_{i\pm\frac{1}{2}} + a_4 [[p]]_{i\pm\frac{1}{2}}\right)$$
$$\partial_t u + \frac{1}{2\Delta x}[a_3 u + a_4 p]_{i\pm 1} = c_2 \cdot \frac{1}{\Delta x}\left(a_1 [[u]]_{i\pm\frac{1}{2}} + a_2 [[p]]_{i\pm\frac{1}{2}}\right)$$

or

$$\partial_t \begin{pmatrix} u \\ p \end{pmatrix} + \frac{1}{2\Delta x}\left[\begin{pmatrix} a_3 & a_4 \\ a_1 & a_2 \end{pmatrix}\begin{bmatrix} u \\ p \end{bmatrix}_{i\pm 1} - \begin{pmatrix} 2c_2a_1 & 2c_2a_2 \\ 2c_1a_3 & 2c_1a_4 \end{pmatrix}\left[\!\!\left[\begin{matrix} u \\ p \end{matrix} \right]\!\!\right]_{i\pm\frac{1}{2}}\right] = 0$$

Define

$$J := \begin{pmatrix} a_3 & a_4 \\ a_1 & a_2 \end{pmatrix}$$

In order to obtain the upwind/Roe scheme one needs to choose c_1 and c_2 as functions of a_1, a_2, a_3, a_4 such that the matrix

$$D := \begin{pmatrix} 2c_2a_1 & 2c_2a_2 \\ 2c_1a_3 & 2c_1a_4 \end{pmatrix}$$

becomes the diffusion matrix of the upwind/Roe scheme, i.e. $D = |J|$. This means that the two matrices need to be simultaneously diagonalizable and that they commute. The commutator vanishes under the conditions

$$\begin{aligned} c_1 a_3 a_4 &= c_2 a_1 a_2 \\ c_1 a_4^2 + c_2(a_2 a_3 - a_1 a_4 - a_2^2) &= 0 \\ c_2 a_1^2 + c_1(a_2 a_3 - a_3^2 - a_4 a_1) &= 0 \end{aligned}$$

These are three linear equations for the two variables c_1, c_2. If $a_4 \neq 0$, $a_1 \neq 0$ then they only have nontrivial solutions if

$$(a_1 a_4 - a_2 a_3)(a_3 - a_2) = 0$$

Clearly, then there also exists an infinity of solutions.

If $a_3 = a_2 =: a$, then c_1 can be taken arbitrary and

$$c_2 = c_1 \frac{a_4}{a_1}$$

One can explicitly diagonalize J and D in this case, i.e. there exists an invertible matrix R with:

$$R^{-1} J R = \begin{pmatrix} a - \sqrt{a_1 a_4} & \\ & a + \sqrt{a_1 a_4} \end{pmatrix}$$
$$R^{-1} D R = 2c_1\sqrt{\frac{a_4}{a_1}}\begin{pmatrix} \sqrt{a_1 a_4} - a & 0 \\ 0 & \sqrt{a_1 a_4} + a \end{pmatrix}$$

If $|a| < \sqrt{a_1 a_4}$, then

$$R^{-1}|J|R = \begin{pmatrix} \sqrt{a_1 a_4} - a & \\ & a + \sqrt{a_1 a_4} \end{pmatrix}$$

This is obtained by choosing

$$c_1 = \frac{1}{2}\sqrt{\frac{a_1}{a_4}}$$

The other case $\sqrt{a_1 a_4} < a$ is not reachable by any choice of c_1. Indeed, one would need

$$2c_1\sqrt{\frac{a_4}{a_1}}(\sqrt{a_1 a_4} - a) = a - \sqrt{a_1 a_4}$$
$$2c_1\sqrt{\frac{a_4}{a_1}}(\sqrt{a_1 a_4} + a) = a + \sqrt{a_1 a_4}$$

which is only solved by $c_1 = c_2 = 0$. □

Thinking of linearized Euler (linear acoustics and linear advection)

$$\partial_t p + U\partial_x p + c^2 \partial_x u = 0$$
$$\partial_t u + U\partial_x u + \frac{\partial_x p}{\epsilon^2} = 0$$

one would have

$$a_3 = a_2 = a = U \qquad a_1 = c^2 \qquad a_4 = \frac{1}{\epsilon^2}$$

For linearized Euler the case $|a| < \sqrt{a_1 a_4} = \frac{c}{\epsilon}$ corresponds to a subsonic situation. From the theorem then follows the choice

$$c_1 = \frac{1}{2}c\epsilon \qquad c_2 = \frac{1}{2}\frac{1}{c\epsilon}$$

The supersonic case $\sqrt{a_1 a_4} < a$ is not reachable by any choice of c_1.

Numerical results for the scheme (5.20)–(5.21) are shown in Figure 5.4. Experimentally, it is found to be linearly stable in the subsonic regime, but seems to suffer from a different kind of instability that only appears at much later times. Therefore the scheme cannot be satisfactorily used for simulations in its current form, but there is hope that this instability will be understood and cured in future. On the other hand, Figure 5.4 shows that stationarity preservation even in the nonlinear case seems to lead to a low Mach compliant scheme. This is reassuring and subject of future work.

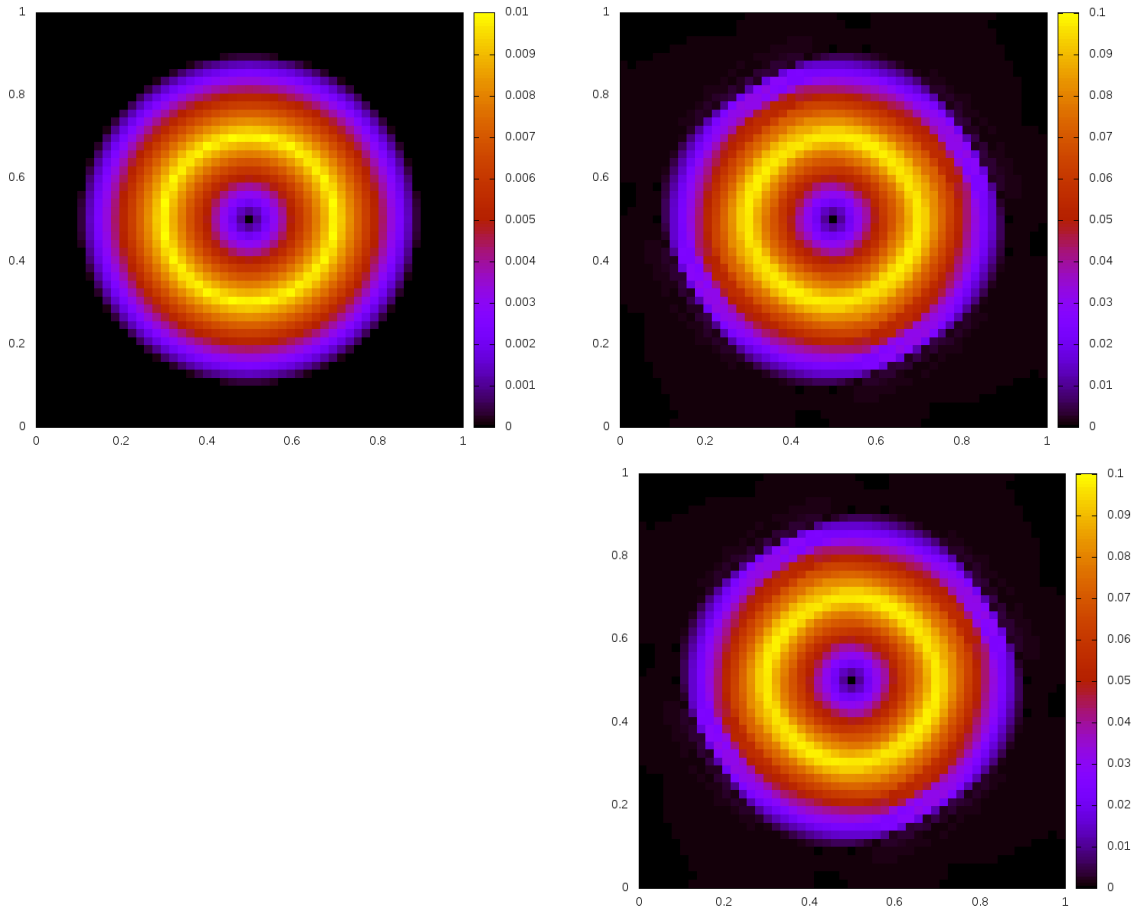

Figure 5.4: Numerical simulation using the scheme described in Theorem 5.4. The initial data are chosen to be an isentropic version of the Gresho vortex given in (5.15)–(5.16). The isentropic Euler equations (1.6)–(1.7) with $K = 1$ and $\gamma = 1.4$ are solved on a square 50×50 grid. Colour coded is the Mach number. *Top left*: Initial setup. *Top right*: Simulation result at $t = 2$ for $\epsilon = 10^{-1}$. *Bottom right*: Simulation result at $t = 2$ for $\epsilon = 10^{-2}$. Commit hashes: 8e290b1 and 3d52689.

5.4.2 Scheme using the pseudo-inverse

In Sections 3.2.3 and 4.5.2.2 a multi-dimensional scheme is derived by extending the one-dimensional Roe scheme to multiple dimensions in a stationarity consistent manner. This does not only involve computing the usual absolute value $|J_x|$ of the Jacobian J_x, but also its sign $|J_x|J_x^{-1}$. However, J_x is in general not invertible. Definition 3.11 thus sets $\operatorname{sign} J_x := |J_x|J_x^{\textcircled{-1}}$ with $J_x^{\textcircled{-1}}$ being the Moore-Penrose pseudo-inverse (Definition 3.10). The strategy of Section 3.2.3 – to be extended to nonlinear systems here – does not involve any free parameters. This is in contrast to the scheme presented in Section 5.4.1.

Consider the hyperbolic system

$$\partial_t q + \partial_x f^x(q) + \partial_y f^y(q) = 0 \tag{5.24}$$

and the one-dimensional Roe solver defined by its numerical flux

$$f^x_{i+\frac{1}{2},j} = \frac{1}{2}\{f^x(q\}_{i+\frac{1}{2},j} - \frac{1}{2}|J_x|(q_{i+\frac{1}{2},j})[q]_{i+\frac{1}{2},j} \tag{5.25}$$

This scheme needs to be extended to multiple spatial dimensions, and this shall be done using a stationarity consistent diffusion. Mimic the scheme again by continuous operators, before dealing with the discrete case. The stationary states of (5.24) are given by

$$\partial_x f^x(q) + \partial_y f^y(q) = 0 \tag{5.26}$$

The Roe flux (5.25) leads to a dimensionally split scheme that has this shape:

$$\partial_t q + \partial_x f^x(q) + \partial_y f^y(q) = (\ldots)\partial_x\big(|J_x|\partial_x q\big) + (\ldots)\partial_y\big(|J_y|\partial_y q\big)$$

Recall that the multi-dimensional extension needs to have a diffusion that vanishes whenever the flux divergence (5.26) vanishes. To find out what (5.26) implies for q is very difficult: this would involve solving the stationary Euler equations. If one manages to express $\partial_x q$ by some function of the fluxes, things would become easier. And indeed, obviously

$$\partial_x f^x = J_x \partial_x q \qquad\qquad \partial_x f^y = J_y \partial_y q \tag{5.27}$$

and thus (ignore the question of the invertibility for the moment, it is discussed below in detail)

$$\partial_t q + \partial_x f^x(q) + \partial_y f^y(q) = (\ldots)\partial_x(|J_x|J_x^{-1}\partial_x f^x) + (\ldots)\partial_y(|J_y|J_y^{-1}\partial_y f^y)$$

This easily has a stationarity consistent extension to multiple spatial dimensions:

$$\begin{aligned}\partial_t q + \partial_x f^x(q) + \partial_y f^y(q) = (\ldots)\partial_x\Big(|J_x|J_x^{-1}(\partial_x f^x + \partial_y f^y)\Big)\\ + (\ldots)\partial_y\Big(|J_y|J_y^{-1}(\partial_x f^x + \partial_y f^y)\Big)\end{aligned} \tag{5.28}$$

Discrete counterparts to (5.27) and (5.28) need to be found, and the inverse of J_x, J_y needs to be treated. This latter issue appeared earlier again and one can hope to be able to regularize it by means of the Moore-Penrose pseudo-inverse.

Turn first to the discrete version of (5.27). For upwinding in the supersonic case when all the eigenvalues of $|J_x|(q_{i+\frac{1}{2},j})$ are positive, one needs to have (in one spatial dimension again)

$$\begin{aligned} f^x(q_{ij}) \stackrel{!}{=} f^x_{i+\frac{1}{2},j} &= \frac{1}{2}(f^x(q_{i+1,j}) + f^x(q_{ij})) - \frac{1}{2}J_x(q_{i+\frac{1}{2},j})[q]_{i+\frac{1}{2},j}\\ [f^x(q)]_{i+\frac{1}{2},j} &= J_x(q_{i+\frac{1}{2},j})[q]_{i+\frac{1}{2},j}\end{aligned} \tag{5.29}$$

Roe ([Roe81]) explicitly constructs an averaging procedure for the Euler equations, with which $q_{i+\frac{1}{2},j}$ can be computed from q_{ij} and $q_{i+1,j}$ such that (5.29) is true. This equation is a discrete counterpart to (5.27).

Theorem 5.6 (Roe average)**.** *Given*

$$q_R = (\rho_R, \rho_R u_R, \rho_R v_R, e_R) \qquad q_L = (\rho_L, \rho_L u_L, \rho_L v_L, e_L)$$
$$\mathbf{v}_R := (u_R, v_R) \qquad \mathbf{v}_L := (u_L, v_L)$$

define the average $\bar{a}$ *of any two quantities* a_R *and* a_L *by*

$$\bar{a} := \frac{a_R\sqrt{\rho_R} + a_L\sqrt{\rho_L}}{\sqrt{\rho_R} + \sqrt{\rho_L}}$$

Also define the enthalpy

$$h_R := \frac{\gamma e_R}{\rho_R} - (\gamma - 1)\frac{|\mathbf{v}_R|^2}{2}$$

and analogously h_L*.*

Consider the fluxes f^x*,* f^y *of the Euler equations as in* (1.8)–(1.10) *and* $J_x = \nabla_q f^x$*. Then*

$$f^x(q_R) - f^x(q_L) = J_x(\bar{q})(q_R - q_L)$$

is identically (i.e. for all q_R*,* q_L*) true if*

$$\bar{q} = \left(\bar{\rho}, \bar{\rho}\bar{u}, \bar{\rho}\bar{u}, \frac{\bar{\rho}\bar{h}}{\gamma} + \frac{\rho|\bar{\mathbf{v}}|^2}{2}\frac{\gamma - 1}{\gamma}\right)$$

Proof. See e.g. [Roe81, LeV02]. □

Equation (5.25) would thus be rewritten as

$$\begin{aligned} f^x_{i+\frac{1}{2},j} &= \frac{1}{2}\{f^x(q)\}_{i+\frac{1}{2},j} - \frac{1}{2}|J_x|(q_{i+\frac{1}{2},j})J_x^{-1}(q_{i+\frac{1}{2},j})[f(q)]_{i+\frac{1}{2},j} \\ &= \frac{1}{2}\{f_x(q)\}_{i+\frac{1}{2},j} - \frac{1}{2}\operatorname{sign} J_x(q_{i+\frac{1}{2},j})[f^x(q)]_{i+\frac{1}{2},j} \end{aligned}$$

were it not again for the fact that J_x is not invertible. The eigenvalues of $\bar{J}_x$ vanish at the sonic point and at zero velocity (this latter being the relevant case for low Mach number flow).

The following Theorem establishes a regularization:

Theorem 5.7. *Consider* $q_{i+\frac{1}{2},j}$ *as in Theorem 5.6 and the Moore-Penrose pseudo-inverse* $J_x^{\textcircled{-1}}$ *of* J_x *as in Definition 3.10. Then for subsonic flow*

$$|J_x|(q_{i+\frac{1}{2},j})[q]_{i+\frac{1}{2},j} = |J_x|(q_{i+\frac{1}{2},j})J_x^{\textcircled{-1}}(q_{i+\frac{1}{2},j})[f^x(q)]_{i+\frac{1}{2},j}$$

Proof. J_x has eigenvalues $\{u \pm c, u\}$. If $\bar{u}$, associated to $q_{i+\frac{1}{2},j}$ as in Theorem 5.6 does not vanish, then J_x is invertible for subsonic flow, and the pseudo-inverse is the usual inverse by construction. If $\bar{u} = 0$ then both sides of the equation exist, but need not be equal in principle. From $\bar{u} = 0$ follows

$$\sqrt{\rho_i}u_i + \sqrt{\rho_{i+1}}u_{i+1} = 0$$

Using this allows to explicitly compute both sides in order to verify the equality. The computation is lengthy and is omitted here to save a tree. □

It is not obvious whether this finding is pure luck, or whether this has a deep reason. Note that one can easily compute, for the special case $u_i = u_{i+1} = \bar{u} = 0$ that

$$J_x^{\textcircled{-1}}(q_{i+\frac{1}{2},j})[f^x(q)]_{i+\frac{1}{2},j} = \begin{pmatrix} 0 \\ \cdots \\ 0 \\ 0 \end{pmatrix}$$

such that in general

$$J_x^{\textcircled{-1}}(q_{i+\frac{1}{2},j})[f^x(q)]_{i+\frac{1}{2},j} \neq [q]_{i+\frac{1}{2},j}$$

The last ingredient is a discrete counterpart to Equation (5.28). In particular one needs to ensure that a discrete version of

$$\partial_x \Big(\operatorname{sign} J_x (\partial_x f^x + \partial_y f^y) \Big) \tag{5.30}$$

vanishes whenever a discrete version of

$$\partial_x f^x + \partial_y f^y$$

does. So far only a stationarity consistent discretization of

$$\partial_x^2 f^x + \partial_x \partial_y f^y$$

is available by Corollary 4.10. In (5.30), J_x is a complicated nonlinear function of q. However, by the Leibniz rule

$$\partial_x \Big(\operatorname{sign} J_x (\partial_x f^x + \partial_y f^y) \Big) = \operatorname{sign} J_x (\partial_x^2 f^x + \partial_x \partial_y f^y) \Big) + (\partial_x \operatorname{sign} J_x) \cdot (\partial_x f^x + \partial_y f^y)$$

This has a discrete counterpart:

Lemma 5.1 (Discrete Leibniz rule)**.**

$$\Big[A_{\cdot,j} \Big(\frac{[\{\{u\}\}_{j\pm\frac{1}{2}}]}{4} + \frac{\{[v]_{j\pm1}\}}{4} \Big) \Big]_{i\pm\frac{1}{2}} =$$
$$\frac{1}{2} \{A\}_{i\pm\frac{1}{2},j} \Big(\frac{[[\{\{u\}\}_{j\pm\frac{1}{2}}]]_{i\pm\frac{1}{2}}}{4} + \frac{[[v]_{i\pm1}]_{j\pm1}}{4} \Big) + [A]_{i\pm\frac{1}{2},j} \Big(\frac{[\{\{u\}\}_{j\pm\frac{1}{4}}]_{i\pm1}}{8} + \frac{[\{\{v\}\}_{i\pm\frac{1}{4}}]_{j\pm1}}{8} \Big) \tag{5.31}$$

is a discrete version of the statement

$$\partial_x \Big(A(\partial_x u + \partial_y v) \Big) = A(\partial_x^2 u + \partial_x \partial_y v) + (\partial_x A)(\partial_x u + \partial_y v)$$

Proof. Expand (by dividing into two halves)

$$\begin{aligned}
&\Big[A_{\cdot,j} \Big[\{\{u\}\}_{j\pm\frac{1}{2}} \Big] \Big]_{i\pm\frac{1}{2}} + \Big[A_{\cdot,j} \Big\{ [v]_{j\pm1} \Big\} \Big]_{i\pm\frac{1}{2}} \\
&= \frac{1}{2} A_{i+\frac{1}{2},j} \Big[\{\{u\}\}_{j\pm\frac{1}{2}} \Big]_{i+\frac{1}{2}} + \frac{1}{2} A_{i+\frac{1}{2},j} \Big\{ [v]_{j\pm1} \Big\}_{i+\frac{1}{2}} - \frac{1}{2} A_{i-\frac{1}{2},j} \Big[\{\{u\}\}_{j\pm\frac{1}{2}} \Big]_{i-\frac{1}{2}} - \frac{1}{2} A_{i-\frac{1}{2},j} \Big\{ [v]_{j\pm1} \Big\}_{i-\frac{1}{2}} \\
&\quad + \frac{1}{2} A_{i+\frac{1}{2},j} \Big[\{\{u\}\}_{j\pm\frac{1}{2}} \Big]_{i+\frac{1}{2}} + \frac{1}{2} A_{i+\frac{1}{2},j} \Big\{ [v]_{j\pm1} \Big\}_{i+\frac{1}{2}} - \frac{1}{2} A_{i-\frac{1}{2},j} \Big[\{\{u\}\}_{j\pm\frac{1}{2}} \Big]_{i-\frac{1}{2}} - \frac{1}{2} A_{i-\frac{1}{2},j} \Big\{ [v]_{j\pm1} \Big\}_{i-\frac{1}{2}}
\end{aligned}$$

and add zero:

$$
\begin{aligned}
&= \tfrac{1}{2}A_{i+\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i+\frac{1}{2}} + \tfrac{1}{2}A_{i+\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i+\frac{1}{2}} - \tfrac{1}{2}A_{i-\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i-\frac{1}{2}} - \tfrac{1}{2}A_{i-\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i-\frac{1}{2}} \\
&\quad + \boxed{\tfrac{1}{2}A_{i-\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i+\frac{1}{2}} - \tfrac{1}{2}A_{i-\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i+\frac{1}{2}}} + \boxed{\tfrac{1}{2}A_{i+\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i-\frac{1}{2}} - \tfrac{1}{2}A_{i+\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i-\frac{1}{2}}} \\
&\quad + \tfrac{1}{2}A_{i+\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i+\frac{1}{2}} + \tfrac{1}{2}A_{i+\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i+\frac{1}{2}} - \tfrac{1}{2}A_{i-\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i-\frac{1}{2}} - \tfrac{1}{2}A_{i-\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i-\frac{1}{2}} \\
&\quad + \boxed{\tfrac{1}{2}A_{i-\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i+\frac{1}{2}} - \tfrac{1}{2}A_{i-\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i+\frac{1}{2}}} + \boxed{\tfrac{1}{2}A_{i+\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i-\frac{1}{2}} - \tfrac{1}{2}A_{i+\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i-\frac{1}{2}}} \\
&= \tfrac{1}{2}\{A\}_{i\pm\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i+\frac{1}{2}} - \tfrac{1}{2}\{A\}_{i\pm\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i-\frac{1}{2}} + \tfrac{1}{2}[A]_{i\pm\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i+\frac{1}{2}} + \tfrac{1}{2}[A]_{i\pm\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i-\frac{1}{2}} \\
&\quad + \tfrac{1}{2}\{A\}_{i\pm\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i+\frac{1}{2}} - \tfrac{1}{2}\{A\}_{i\pm\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i-\frac{1}{2}} + \tfrac{1}{2}[A]_{i\pm\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i+\frac{1}{2}} + \tfrac{1}{2}[A]_{i\pm\frac{1}{2},j}\left\{[v]_{j\pm1}\right\}_{i-\frac{1}{2}} \\
&= \tfrac{1}{2}\{A\}_{i\pm\frac{1}{2},j}\left[\!\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]\!\right]_{i\pm\frac{1}{2}} + \tfrac{1}{2}\{A\}_{i\pm\frac{1}{2},j}\left[[v]_{j\pm1}\right]_{i\pm1} + \tfrac{1}{2}[A]_{i\pm\frac{1}{2},j}\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i\pm1} + \tfrac{1}{2}[A]_{i\pm\frac{1}{2},j}\left\{\left\{[v]_{j\pm1}\right\}\right\}_{i\pm\frac{1}{2}} \\
&= \tfrac{1}{2}\{A\}_{i\pm\frac{1}{2},j}\left(\left[\!\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]\!\right]_{i\pm\frac{1}{2}} + \left[[v]_{j\pm1}\right]_{i\pm1}\right) + \tfrac{1}{2}[A]_{i\pm\frac{1}{2},j}\left(\left[\{\{u\}\}_{j\pm\frac{1}{2}}\right]_{i\pm1} + \left\{\left\{[v]_{j\pm1}\right\}\right\}_{i\pm\frac{1}{2}}\right)
\end{aligned}
$$

which proves the statement. Note that the average $\frac{1}{4}\{\{\cdot\}\}_{j\pm\frac{1}{2}}$ is a crucial ingredient of the proof.

□

Thus, although Fourier transform methods cannot be directly applied in the nonlinear setting, nonlinearity can be handled in such cases via a discrete Leibniz rule!

Now all the ingredients are in place to write down the scheme:

Theorem 5.8. *Consider the Euler equations* (1.8)–(1.10) *in two spatial dimensions. Denote by* $q_{i+\frac{1}{2},j}$ *the Roe average as in Theorem 5.6 and let* $\operatorname{sign} J_x = |J_x|J_x^{\textcircled{-1}}$. *The scheme obtained from the following numerical flux*

$$
f^x_{i+\frac{1}{2},j} = \frac{1}{2}\left\{\frac{1}{4}\{\{f^x(q)\}\}_{j\pm\frac{1}{2}}\right\}_{i+\frac{1}{2}} - \frac{1}{2}\operatorname{sign} J_x(q_{i+\frac{1}{2},j})\left(\left[\frac{1}{4}\{\{f^x(q)\}\}_{j\pm\frac{1}{4}}\right]_{i+\frac{1}{2}} + \frac{1}{4}\{[f^y(q)]_{j\pm1}\}_{i+\frac{1}{2}}\right)
$$

(with the flux in y-direction given by appropriate rotation) is stationarity preserving with the discrete stationary states given by

$$
\frac{1}{2}\left[\frac{1}{4}\{\{f^x(q)\}\}_{j\pm\frac{1}{2}}\right]_{i\pm1} + \frac{1}{2}\left[\frac{1}{4}\{\{f^y(q)\}\}_{i\pm\frac{1}{2}}\right]_{j\pm1} = 0
$$

It reduces to the Roe scheme when applied to a one-dimensional situation.

Proof. The flux difference reads

$$
[f^x]_{i\pm\frac{1}{2},j} = \frac{1}{2}\left[\frac{1}{4}\{\{f^x(q)\}\}_{j\pm\frac{1}{2}}\right]_{i\pm1} - \frac{1}{2}\left[\operatorname{sign} J_x(q)\left(\left[\frac{1}{4}\{\{f^x(q)\}\}_{j\pm\frac{1}{4}}\right] + \frac{1}{4}\{[f^y(q)]_{j\pm1}\}\right)\right]_{i\pm\frac{1}{2}}
$$

Therefore stationarity preservation follows from Lemma 5.1. Applied to a one-dimensional situation the one-dimensional Roe scheme follows by Theorem 5.7. □

Numerical results for this scheme are shown in Figure 5.5. Experimentally again stationarity preservation seems to lead to a low Mach compliant scheme. In the linearized regime around a static state the scheme reduces to the multi-dimensional scheme (4.32) which is stable. The presence of the sign function seems to lead to artefacts that spoil stability for very long times. This is partly due to the fact that errors at the level of machine precision get amplified. When entering the sign function, two numbers -10^{-16} and 10^{-16} which are very close get torn apart to become ± 1. This is an issue related to the practical implementation of the scheme on real computers, and does not touch the theory above. As a remedy it has been found that in practice it helps to replace the sign function by a continuous function $\operatorname{sign}_\delta$

$$\operatorname{sign}_\delta(x) := \begin{cases} 1 & x \geq \delta \\ -2\left(\frac{a}{\delta}\right)^3 + 3\left(\frac{a}{\delta}\right)^2 & 0 \leq x < \delta \\ -2\left(\frac{a}{\delta}\right)^3 - 3\left(\frac{a}{\delta}\right)^2 & -\delta \leq x < 0 \\ -1 & x < -\delta \end{cases} \tag{5.32}$$

Results of both computations are shown in Figure 5.5. The relation between stationarity preservation and low Mach compliance in the nonlinear case are subject of future work, for which the results obtained with the schemes of Section 5.4.1 and 5.4.2 are an inspiration.

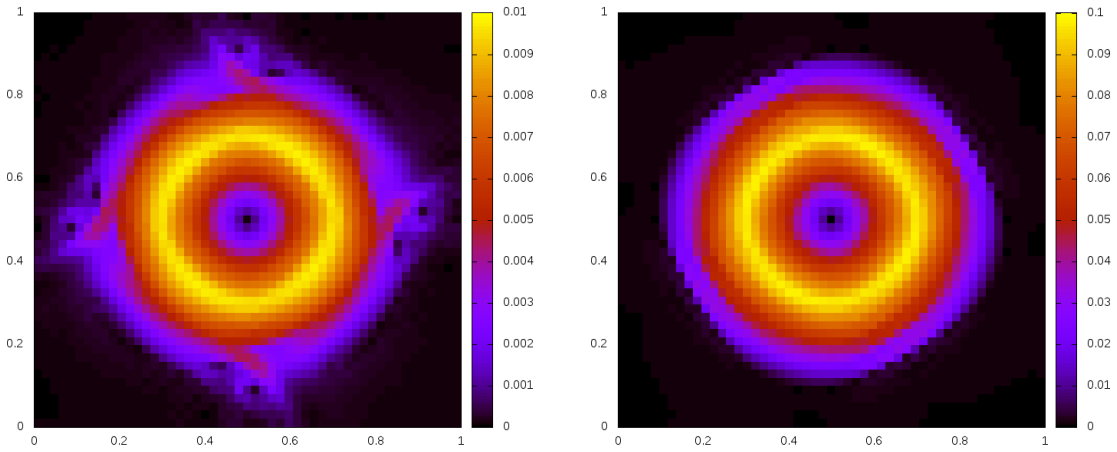

Figure 5.5: Numerical simulation using the scheme described in Theorem 5.8. The initial data are given in Equations (5.15)–(5.16) and shown in Figure 5.2 (left). The Euler equations (1.8)–(1.10) with $\gamma = 1.4$ and $\epsilon = 10^{-2}$ are solved on a square 50×50 grid. Colour coded is the Mach number and the result is shown at $t = 2$. *Left*: Using the usual sign function. *Right*: The sign function smoothed out according to (5.32) with $\delta = 10^{-7}$. Commit hashes: 87ce579 and 331d2ac.

5.5 Low Mach number scheme

Recall Theorem 4.10 for linear acoustics. For low Mach compliance, it is sufficient for the scheme (4.32) to be stationarity preserving, as Theorem 4.1 then establishes its good behaviour in the limit of low Mach number. Theorem 4.10 shows the same via asymptotic analysis. In the nonlinear case the Fourier transform that was used to establish stationarity preservation is unavailable. However, asymptotic analysis is still viable.

What is the essential reason for the scheme (4.32) to be low Mach compliant? Consider one of the equations, obtained formally by an asymptotic analysis, e.g. Equation (4.45):

$$0 = \frac{1}{8\Delta x}\{\{[p^{(1)}]_{i\pm1}\}\}_{j\pm\frac{1}{2}} - c\left(\frac{1}{8\Delta x}\{\{[[u^{(0)}]]_{i\pm\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}\boxed{+\frac{1}{8\Delta y}[[v^{(0)}]_{i\pm1}]_{j\pm1}}\right)$$

It is a discretization of

$$0 = \partial_x p^{(1)} - \frac{\Delta x}{2}c\left(\partial_x^2 u^{(0)}\boxed{+\partial_x\partial_y v^{(0)}}\right)$$

The boxed term is the one that is essentially different from the corresponding equations for the upwind/Roe scheme. Recall that the PDE requires $\partial_x p^{(1)} \in \mathcal{O}(\epsilon)$. So far, low Mach fixes started out with the upwind/Roe scheme and have focused on appending a factor of ϵ to the second derivative of u and thus in a sense removing it. The multi-dimensional scheme (4.32) follows a different strategy – it *adds* another term of the same kind. The sum of the two, however, is $\mathcal{O}(\epsilon)$, and this strategy in the end therefore achieves the same. The advantage of this approach is that it does not need any adjustable parameters.

So far, low Mach fixes generically were derived for the Euler equations, and had an immediate interpretation for acoustic equations. The multi-dimensional scheme (4.32) has been derived for the acoustic equations directly. The aim of this Section is to extend it in some sense to the Euler equations. Whereas for the acoustic equations, stationarity preservation, vorticity preservation and low Mach compliance are equivalent, this is not the case for the Euler equations. Therefore here, for the Euler equations, the focus shall lie on low Mach compliance only. As is shown in Section 5.1.2, it is just the same second derivative of u that appears in the asymptotic analysis of the Roe scheme for the Euler equations. The idea thus is to add a cross-derivative of v in order to balance it.

As usual, mimic the numerical scheme first by continuous diffusion operators. Recall that the Roe matrix (Equation (5.4)) in the limit $\epsilon \to 0$ is

$$|J_x| = \frac{1}{\epsilon}\begin{pmatrix} 0 & 0 & 0 & \frac{\gamma-1}{c} \\ -cu & c & 0 & \frac{2u(\gamma-1)}{c} \\ 0 & 0 & 0 & \frac{(\gamma-1)v}{c} \\ 0 & 0 & 0 & c \end{pmatrix} + \mathcal{O}(1) \tag{5.33}$$

The Roe scheme (omitting the flux divergence) thus contains, to highest order in ϵ, the following diffusion

$$\begin{aligned}\partial_t(\rho u) &\simeq \partial_x\left(-\frac{cu}{\epsilon}\partial_x\rho + \frac{c}{\epsilon}\partial_x(\rho u)\right) + \partial_y\left(-\frac{cv}{\epsilon}\partial_y\rho + \frac{c}{\epsilon}\partial_y(\rho v)\right)\\ &= \partial_x\left(\frac{c\rho}{\epsilon}\partial_x u\right) + \partial_y\left(\frac{c\rho}{\epsilon}\partial_y v\right)\end{aligned}$$

Here it has been used that $-\frac{cu}{\epsilon}\partial_x\rho + \frac{c}{\epsilon}\partial_x(\rho u) = \frac{c\rho}{\epsilon}\partial_x u$.

Observe the presence of second derivatives of u and v. In order to make appear the divergence $\partial_x u + \partial_y v$ one has to augment this expression by cross-derivatives

$$\partial_t(\rho u) \simeq \partial_x\left(\frac{c\rho}{\epsilon}(\partial_x u + \partial_y v)\right) + \partial_y\left(\frac{c\rho}{\epsilon}(\partial_x u + \partial_y v)\right)$$

and going back to conservative variables, rewrite this into

$$= \partial_x\left(-\frac{cu}{\epsilon}\partial_x\rho + \frac{c}{\epsilon}\partial_x(\rho u) - \frac{cv}{\epsilon}\partial_y\rho + \frac{c}{\epsilon}\partial_y(\rho v)\right) + \partial_y\Big(\dots\Big) \tag{5.34}$$

By analogy with the scheme (4.32) in two spatial dimensions, and based on the Roe scheme, one might consider the following multi-dimensional scheme for the Euler equations:

$$f^x_{i+\frac{1}{2}} = \frac{1}{2}\frac{\{\{\{f(q)\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{4} - \frac{1}{2}|J_x|(q_{i+\frac{1}{2},j})\frac{\{\{[q]\}\}_{j\pm\frac{1}{2}}}{4} - \frac{1}{2}K_x\frac{[\{q\}_{i+\frac{1}{2}}]_{j\pm1}}{4} \tag{5.35}$$

with

$$K_x = \begin{pmatrix} 0 & & & \\ -\frac{cv}{\epsilon} & 0 & \frac{c}{\epsilon} & \\ & & 0 & \\ & & & 0 \end{pmatrix} \tag{5.36}$$

It contains discretizations of (5.34). However, recall that when deriving the multi-dimensional scheme (4.32) for linear acoustics, the discrete derivatives have been chosen carefully such that a relation of the type

$$\partial_x u + \partial_y v = 0 \qquad \Rightarrow \qquad \partial_x^2 u + \partial_x\partial_y v = 0$$

remains exactly true at discrete level. In the nonlinear case one has to find a discrete counterpart to (5.34) such that this diffusion vanishes whenever $\partial_x u + \partial_y v$ does. This is not guaranteed by the flux in Equation (5.35), which is due to the nonlinear terms. Such discrete relations however can also be proven in the nonlinear setting if one manages to find a suitable discrete Leibniz rule. Recall the relation (5.31), given here again:

$$\left[A_{\cdot,j}\left(\left[\frac{1}{4}\{\{u\}\}_{j+\frac{1}{4}}\right] + \frac{1}{4}\{[v]_{j\pm1}\}\right)\right]_{i\pm\frac{1}{2}} =$$

$$\frac{1}{2}\{A\}_{i\pm\frac{1}{2},j}\left(\left[\left[\frac{1}{4}\{\{u\}\}_{j\pm\frac{1}{2}}\right]\right]_{i\pm\frac{1}{2}} + \frac{1}{4}[[v]_{i\pm1}]_{j\pm1}\right) + \frac{1}{2}[A]_{i\pm\frac{1}{2},j}\left(\left[\frac{1}{4}\{\{u\}\}_{j\pm\frac{1}{4}}\right]_{i\pm1} + \left[\frac{1}{4}\{\{v\}\}_{i\pm\frac{1}{4}}\right]_{j\pm1}\right)$$

This is a discrete counterpart to

$$\partial_x(A(\partial_x u + \partial_y v)) = A(\partial_x^2 u + \partial_x \partial_y v) + \partial_x A(\partial_x u + \partial_y v)$$

and nowhere does this relation depend on the properties of A. Indeed, it might well be a nonlinear function of the other variables.

Lemma 5.2 (Discrete Leibniz rules)**.** *The following discrete Leibniz rules are discretizations of* $\partial_x(AB) = \partial_x A \cdot B + A \cdot \partial_x B$*:*

i)

$$[AB]_{i+\frac{1}{2}} = \frac{\{A\}_{i+\frac{1}{2}}}{2}[B]_{i+\frac{1}{2}} + [A]_{i+\frac{1}{2}}\frac{\{B\}_{i+\frac{1}{2}}}{2}$$

ii)

$$\frac{\{A\}_{i+\frac{1}{2}}}{2}\frac{\{B\}_{i+\frac{1}{2}}}{2} - \frac{\{A\}_{i-\frac{1}{2}}}{2}\frac{\{B\}_{i-\frac{1}{2}}}{2} = \left[\frac{\{A\}}{2}\frac{\{B\}}{2}\right]_{i\pm\frac{1}{2}} = \frac{\{\{A\}\}_{i\pm\frac{1}{2}}}{4}\frac{[B]_{i\pm1}}{2} + \frac{[A]_{i\pm1}}{2}\frac{\{\{B\}\}_{i\pm\frac{1}{2}}}{4}$$

Proof. Contrary to finding, proving these Leibniz rules is very easy by direct computation. E.g.

$$\begin{aligned}
&\frac{\{A\}_{i+\frac{1}{2}}}{2}[B]_{i+\frac{1}{2}} + [A]_{i+\frac{1}{2}}\frac{\{B\}_{i+\frac{1}{2}}}{2} \\
&= \frac{A_{i+1}+A_i}{2}(B_{i+1}-B_i) + (A_{i+1}-A_i)\frac{B_{i+1}+B_i}{2} \\
&= A_{i+1}B_{i+1} - A_iB_i
\end{aligned}$$

The other relation is obtained in just the same way. □

Recall the dimensionally split Roe scheme for the Euler equations in two spatial dimensions and denote by q the vector of conserved quantities. The flux is given by Equation (5.3)

$$f^x_{i+\frac{1}{2}} = \frac{1}{2}\{f(q)\}_{i+\frac{1}{2},j} - \frac{1}{2}|J_x|(q_{i+\frac{1}{2},j})[q]_{i+\frac{1}{2},j}$$

Theorem 5.9. *Consider the x-flux of the dimensionally split Roe scheme and the following replacements:*

i) The averaged energy flux $\frac{1}{2}\{u(e+p)\}_{i+\frac{1}{2},j}$ *appearing in* $\frac{1}{2}\{f(q)\}_{i+\frac{1}{2},j}$ *is to be replaced by*

$$\frac{1}{2}\left\{\frac{\{\{u\}\}_{j\pm\frac{1}{2}}}{4}\right\}_{i+\frac{1}{2}} \cdot \frac{1}{2}\left\{\frac{\{\{e+p\}\}_{j\pm\frac{1}{2}}}{4}\right\}_{i+\frac{1}{2}}$$

ii) In the jump $[q]$ *that is multiplied by* $|J_x|$ *the density jump is to be chosen as*

$$\frac{\{\{[\rho]_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{4}$$

and the x-momentum jump as

$$\left[\frac{\{\{\rho\}\}_{j\pm\frac{1}{2}}}{4}\cdot\frac{\{\{u\}\}_{j\pm\frac{1}{2}}}{4}\right]_{i+\frac{1}{2}}$$

Also, similarly to (5.35)*, the following term is to be added to the flux:*

$$-\frac{1}{2}K_x\cdot(\textit{discretization of }\partial_y q)$$

with K_x*, to lowest power in* ϵ*, is given by Equation* (5.36)*. The discretization of* $\partial_y q$ *is to be chosen such that* $\partial_y\rho$ *is discretized as*

$$\frac{[\{\rho\}_{i+\frac{1}{2}}]_{j\pm1}}{4}$$

and $\partial_y(\rho v)$ *discretized as*

$$\left[\frac{\{\{\rho\}_{i+\frac{1}{2}}\}}{4}\cdot\frac{\{\{v\}_{i+\frac{1}{2}}\}}{4}\right]_{j\pm\frac{1}{2}}$$

Both $|J_x|$ *and* K_x *are functions of the dependent variables* q*. They are to be evaluated at averaged states in such a way, that the terms* $\frac{cu}{\epsilon}$*,* $\frac{cv}{\epsilon}$ *in* (5.33) *and* (5.36) *in the discrete setting are*

$$\frac{\langle c\rangle_{i+\frac{1}{2},j}\frac{\{\{\{u\}_{i+\frac{1}{2}}\}\}_{j+\frac{1}{2}}}{8}}{\epsilon}\qquad\qquad\frac{\langle c\rangle_{i+\frac{1}{2},j}\frac{\{\{\{v\}_{i+\frac{1}{2}}\}\}_{j+\frac{1}{2}}}{8}}{\epsilon}$$

where $\langle c\rangle_{i+\frac{1}{2},j}$ *denotes any kind of average of* c*. The only condition is that all appearances of* c *in those matrices are discretized using the same average.*

Analogous conditions apply to the y-direction, and it is assumed that the numerical flux in y-direction is obtained by rotating the setup.

A finite volume scheme with such a numerical flux is low Mach compliant. Formally, in the limit the solutions are characterized by

$$p=\text{const}$$
$$\frac{\{[u]_{i\pm1}\}\}_{j\pm\frac{1}{2}}}{8}+\frac{[\{\{v\}\}_{i\pm\frac{1}{2}}]_{j\pm1}}{8}=0$$

Proof. The diffusion, given by the two terms involving $|J_x|$ (to highest order in ϵ) and K_x, by construction is

$$-\frac{\langle c\rangle_{i+\frac{1}{2},j}\frac{\{\{\{u\}_{i+\frac{1}{2}}\}\}_{j+\frac{1}{2}}}{8}}{\epsilon}\frac{\{\{[\rho]_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{4}+\frac{\langle c\rangle_{i+\frac{1}{2},j}}{\epsilon}\left[\frac{\{\{\rho\}\}_{j\pm\frac{1}{2}}}{4}\cdot\frac{\{\{u\}\}_{j\pm\frac{1}{2}}}{4}\right]_{i+\frac{1}{2}}$$
$$-\frac{\langle c\rangle_{i+\frac{1}{2},j}\frac{\{\{\{v\}_{i+\frac{1}{2}}\}\}_{j+\frac{1}{2}}}{8}}{\epsilon}\frac{[\{\rho\}_{i+\frac{1}{2}}]_{j\pm1}}{4}+\frac{\langle c\rangle_{i+\frac{1}{2},j}}{\epsilon}\left[\frac{\{\{\rho\}_{i+\frac{1}{2}}\}}{4}\cdot\frac{\{\{v\}_{i+\frac{1}{2}}\}}{4}\right]_{j\pm\frac{1}{2}}$$

which is a discrete analogue to (5.34).

With the two discrete Leibniz rules of Lemma 5.2 this becomes

$$\frac{\langle c\rangle_{i+\frac{1}{2},j}}{\epsilon}\frac{\{\{\{\rho\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{8}\left(\frac{\{[u]_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}}{4}+\frac{[\{v\}_{i+\frac{1}{2}}]_{j\pm1}}{4}\right)$$

Upon computing the flux difference $[f^x]_{i+\frac{1}{2},j}$ this becomes

$$\left[\frac{\langle c\rangle_{\cdot,j}}{\epsilon}\frac{\{\{\{\rho\}\}\}_{j\pm\frac{1}{2}}}{8}\left(\frac{\{[u]\}\}_{j\pm\frac{1}{2}}}{4}+\frac{[\{v\}]_{j\pm1}}{4}\right)\right]_{i\pm\frac{1}{2}} \tag{5.37}$$

This is exactly what enters (5.31). Therefore the expression in (5.37) vanishes whenever the divergence

$$\frac{\{[u]_{i\pm1}\}\}_{j\pm\frac{1}{2}}}{8}+\frac{[\{\{v\}\}_{i\pm\frac{1}{2}}]_{j\pm1}}{8} \tag{5.38}$$

does. It thus remains to be shown that this is the divergence that appears in the limit $\epsilon\to0$.

As the diffusion does not involve any powers less than ϵ^{-1}, by performing the asymptotic analysis on the momentum equations will yield constancy of $p^{(0)}$, just as in (5.5). Inserting this into the energy equation and making use of condition i) makes appear the discrete divergence (5.38).

Obviously, the same arguments apply in the y-direction as well. □

Corollary 5.1. *A scheme for the Euler equations* (1.8)–(1.10) *that fulfills all the conditions of Theorem 5.9 in two spatial dimensions is given by the following flux:*

$$f^x_{i+\frac{1}{2}}=\begin{pmatrix}\{\{\{\rho u\}\}_{j\pm\frac{1}{2}})\}_{i+\frac{1}{2}}/8\\ \{\{\{\rho u^2+p\}\}_{j\pm\frac{1}{2}})\}_{i+\frac{1}{2}}/8\\ \{\{\{\rho uv\}\}_{j\pm\frac{1}{2}})\}_{i+\frac{1}{2}}/8\\ \{\{\{u\}\}_{j\pm\frac{1}{2}}\}_{i+\frac{1}{2}}\cdot\{\{\{e+p\}\}_{j\pm\frac{1}{2}}\}_{i+\frac{1}{2}}/64\end{pmatrix}$$
$$-|J_x|\begin{pmatrix}\{\{[\rho]_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}/4\\ [\{\{\rho\}\}_{j\pm\frac{1}{2}}\cdot\{\{u\}\}_{j\pm\frac{1}{2}}]_{i+\frac{1}{2}}/16\\ \{\{[\rho v]_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}/4\\ \{\{[e]_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}/4\end{pmatrix}-K_x\begin{pmatrix}[\{\rho\}_{i+\frac{1}{2}}]_{j\pm1}/4\\ [\{\rho u\}_{i+\frac{1}{2}}]_{j\pm1}/4\\ [\{\{\rho\}_{i+\frac{1}{2}}\}\cdot\{\{v\}_{i+\frac{1}{2}}\}]_{j\pm\frac{1}{2}}/16\\ [\{e\}_{i+\frac{1}{2}}]_{j\pm1}/4\end{pmatrix}$$

The numerical flux in y-direction is obtained by an appropriate rotation. Applied to a one-dimensional situation this flux reduces to the Roe flux.

$|J_x|$ *is given as in the dimensionally split Roe scheme, and* K_x *is given by* (5.36)*. The average on which* $|J_x|$ *is evaluated, is, in conservative variables, chosen as*

$$\begin{aligned}
\rho_{i+\frac{1}{2},j} &= \{\{\{\rho\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \\
(\rho u)_{i+\frac{1}{2},j} &= \rho_{i+\frac{1}{2},j} \cdot \frac{\{\{\{u\}_{i+\frac{1}{2}}\}\}_{j+\frac{1}{2}}}{8} \\
(\rho v)_{i+\frac{1}{2},j} &= \{\{\{\rho v\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \\
e_{i+\frac{1}{2},j} &= \{\{\{e\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}
\end{aligned}$$

K_x *is evaluated in the state given by*

$$\begin{aligned}
\rho_{i+\frac{1}{2},j} &= \{\{\{\rho\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \\
(\rho u)_{i+\frac{1}{2},j} &= \{\{\{\rho u\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}} \\
(\rho v)_{i+\frac{1}{2},j} &= \rho_{i+\frac{1}{2},j} \cdot \frac{\{\{\{v\}_{i+\frac{1}{2}}\}\}_{j+\frac{1}{2}}}{8} \\
e_{i+\frac{1}{2},j} &= \{\{\{e\}_{i+\frac{1}{2}}\}\}_{j\pm\frac{1}{2}}
\end{aligned}$$

Note that low Mach compliance deals with only those terms in $|J_x|$ that are $\mathcal{O}(\epsilon^{-1})$. Also K_x can only be specified to this order in ϵ, although it is possible that it contains further terms $\mathcal{O}(1)$. However, they are unimportant for the low Mach number limit and have to be derived using some other conditions. This means that Theorem 5.9 describes a broad class of schemes. They are characterized by the fact that low Mach compliance is achieved by using derivatives of the divergence, rather than derivatives of some components of $\mathbf{v}$. But they still leave plenty of choice, and the scheme in Corollary 5.1 is just one example of such a scheme.

This modification shows very satisfactory results in practice, along with good stability properties of the scheme. The scheme is experimentally found to remain stable up to a CFL number of 1 even in two spatial dimensions. However, as it is not (known to be) stationarity preserving, the numerical results, shown in Figure 5.6, are not of such a good quality as those of Figures 5.4 and 5.5. Nevertheless they seem to demonstrate low Mach compliance, as the numerical errors do not depend on ϵ. Additionally, it shows results obtained with the flux of Equation (5.35), which is not a carefully chosen discretization. However, still the results are indistinguishable. The underlying reasons are subject of future work.

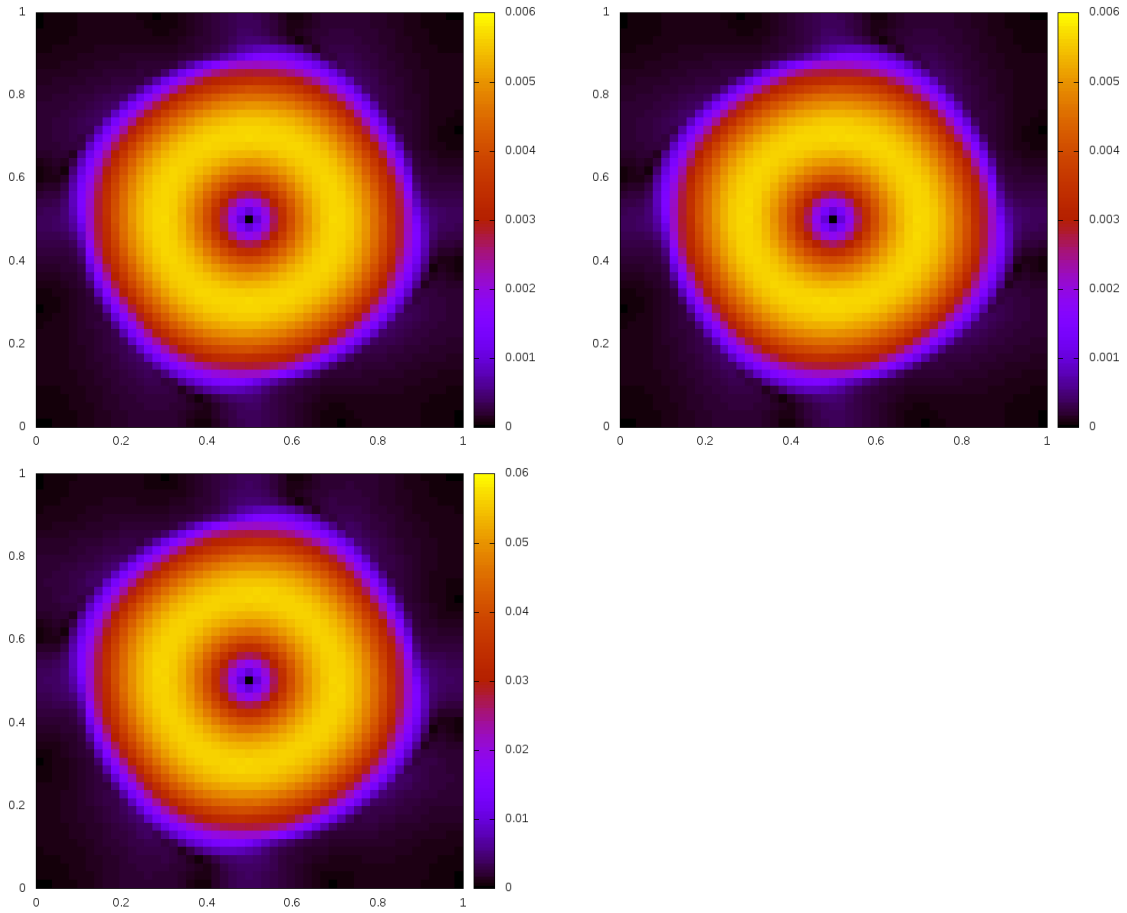

Figure 5.6: Numerical simulation using the scheme described in Corollary 5.1. The initial data are given in Equations (5.15)–(5.16) and shown in Figure 5.2 (left). The Euler equations (1.8)–(1.10) with $\gamma = 1.4$ are solved on a square 50×50 grid. Colour coded is the Mach number and the result is shown at $t = 2$; when comparing to other images of this Section note the different colour scale. *Top left*: Full implementation for $\epsilon = 10^{-2}$. *Top right*: Simplified implementation according to Equation (5.35) for $\epsilon = 10^{-2}$. *Bottom*: Full implementation for $\epsilon = 10^{-1}$. Commit hashes: 506b22e, 9eaae49 and ea7ac58.

Chapter 6

Summary and outlook

For the Euler equations in the low Mach number regime one is facing a dichotomy of numerical schemes: for a given quality of simulation, certain schemes need finer and finer grids the lower the Mach number of the flow is. Among these schemes are such prominent ones as the Godunov or the Roe scheme. Other schemes are able to capture low Mach number flow on a fixed grid. This thesis presents several new approaches for the second kind of scheme.

It seems to be possible to isolate the problem by studying a simpler set of equations. The linearization of the Euler equations around a static state yields the acoustic equations. It is shown in this thesis that the low Mach number limit of the acoustic equations is equivalent to their long time limit. In the low Mach number limit numerical schemes for the acoustic equations are found to display a similar kind of dichotomy as the schemes for the Euler equations.

The analysis numerical schemes for the acoustic equations originates in detailed studies of the numerical solution obtained with the upwind/Roe scheme. It is found, and shown in detail analytically, that for long times the states on which the scheme stationarizes are only a restricted subset of all the analytic stationary states. Whereas the stationary states of the acoustic equations are characterized by a divergencefree velocity ($\partial_x u + \partial_y v = 0$), the numerical stationary states of the Roe scheme are only those that discretize the divergencefree velocity and additionally $\partial_x u = 0$ and $\partial_x v = 0$. The upwind/Roe scheme thus only becomes stationary for shear flows. The ability of a numerical scheme to discretize all the analytic stationary states is termed stationarity preservation. It is shown that the equivalence of the long time and the low Mach number limit has a discrete counterpart. Thus, stationarity preservation allows to single out precisely those schemes that are low Mach compliant, i.e. which discretize all the analytic limit equations for $\epsilon \to 0$.

For the acoustic equations, vorticity is stationary. One can show that a stationarity preserving scheme is vorticity preserving, i.e. there exist a discretization of the vorticity

that remains stationary. Vorticity preserving schemes have been previously studied in literature, motivated by reproducing at discrete level this feature of the equations. The analysis presented here makes a new link between vorticity preserving schemes and low Mach number compliant schemes.

This gives an exhaustive picture of what happens to numerical schemes for linear acoustics in the limit of low Mach number. The tools developed in this thesis allow to efficiently study given schemes for linear acoustics, but can also be applied constructively. A success of the theory is the clarification of the role played by multi-dimensional schemes. A particular multi-dimensional stencil appears in a number of numerical schemes throughout the literature. They usually focus on vorticity preserving schemes for linear acoustics, and in view of the different constructions it might appear surprising that they obtain the same, or very similar, stencils. Stationarity preservation allows to show that this stencil unique in the following sense. Numerical schemes introduce diffusion for stability, and it may spoil the stationary states. The aforementioned stencils appear if one requires the diffusion not to change the stationary states given by a central difference discretization of the PDE. It can be shown that this is only possible with a multi-dimensional scheme that involves all the 8 neighbours of a cell, i.e. also the corner cells. By studying the modified equation associated to this multi-dimensional scheme, the velocity diffusion is to leading order found to be $\operatorname{grad}\operatorname{div}\mathbf{v}$, such that it vanishes whenever $\operatorname{div}\mathbf{v} = 0$. However, to higher orders the operators involved cease to be "rotationally invariant". Although there exist lots of discretizations of $\operatorname{grad}\operatorname{div}\mathbf{v}$ and $\operatorname{div}\mathbf{v}$, obtaining a discrete counterpart to

$$\operatorname{div}\mathbf{v} = 0 \qquad \Rightarrow \qquad \operatorname{grad}\operatorname{div}\mathbf{v} = 0$$

uniquely specifies the aforementioned discrete operators. This is the reason why the corresponding stencils reappear in the literature several times.

In practice, such multi-dimensional schemes show superior behaviour when compared to most dimensionally split ones. They also have better stability properties, with the stability domain extending up to $\textsc{cfl} = 1$. They are conceptually pleasing as they do not introduce any ad hoc parameters, but reflect intrinsic properties of the PDE. Not all multi-dimensional schemes, however, are stationarity preserving (and thus low Mach compliant). This thesis presents a derivation of the two-dimensional Godunov scheme for linear acoustics, which is not found to be stationarity preserving. Its behaviour in the low Mach limit is in large parts similar to that of the dimensionally split upwind/Roe scheme, although it naturally has the larger stability domain with $\textsc{cfl} < 1$.

The thesis presents several ways how the concepts can be extended, and how they allow to construct new schemes. For instance, it is shown how the aforementioned multi-dimensional stationarity preserving scheme can be extended to second order. This extension is not unique and can be performed in such a way that the scheme meets additional requirements. A fascinating question remains, namely how to derive limiters for higher order schemes, that do not spoil stationarity preservation.

The behaviour of numerical methods for other linear systems can also be fruitfully studied using the concept of stationarity preservation. As far as possible, the statements concerning stationarity preservation are formulated for general linear hyperbolic

systems. The application is exemplified in this thesis for linearized Euler equations with gravity. The stationary states are governed by a balance between the source term and the flux divergence. An analytic stationary solution, when inserted into the simulations typically does not remain stationary. The reason is that the way the flux difference is obtained in a numerical method usually does not match the discretization of the source term. There exist several strategies to improve this behaviour, that commonly are referred to as well-balancing. Stationarity preservation allows to understand the origins of unsatisfactory behaviour of the simulation, and to find precise solutions. Even here stationarity preservation yield new insight, despite having been applied to the simplest well-balancing problem.

The extension of stationarity preservation to nonlinear systems is more subtle. Clearly, the linearized regime is covered by results of linear acoustics. Thus, necessary conditions for low Mach compliance are available. Also many of the construction principles, that are initially developed for linear schemes, can be extended to the nonlinear case. This might seem surprising, in view of the essential role that the Fourier transform plays in the proofs. The main reasons that the linear results can be reused is that the Euler equations are quasilinear, and the non-linear products can be treated with discrete counterparts of the Leibniz rule. Clearly, unlike in the linear case, not all questions can be answered to full extent. In particular it remains unclear how for a given scheme to check stationarity preservation. But the different ways of construction of stationarity preserving schemes for the Euler equations that are presented in this thesis show that it is possible to tackle the nonlinear problem. The nonlinear schemes derived in this thesis are not meant to be an exhaustive presentation. However, the fact that they seem to be low Mach compliant encourages future work, devoted to a deeper understanding of the underlying structure.

The stationarity preserving schemes for the Euler equations that are presented in this thesis do not show satisfactory stability properties under explicit time integration. They are stable in the static linearized regime, because in this case they all reduce to a stable multi-dimensional scheme for linear acoustics. To understand reasons for the appearance of the instability is subject of future work.

On the other hand, the multi-dimensional low Mach compliant scheme that has been constructed for the Euler equations is found to perform very well experimentally. When applied to flow of low Mach number it is not found to show artefacts and has a large stability domain which in practice is found to extend up to $\textsc{cfl} = 1$. The way it achieves low Mach compliance, to the author's knowledge, is novel in the realm of the Euler equations. Instead of removing terms that violate the asymptotic scalings, another such term is added. In total they combine to a term that brings another factor of ϵ by exploiting the divergence constraint in the limit. This seems a very elegant way of obtaining low Mach compliance.

The low Mach number limit, stationarity and vorticity preservation are all linked for the acoustic equations. It is not immediately clear what consequences this has for the nonlinear case. Whereas stationarity preservation as a concept can easily be extended to nonlinear equations along the lines of this thesis, vorticity preservation is more involved. Vorticity is not stationary at continuous level. On what grounds to base a discretization of its evolution equation is unclear. It remains to be investigated what connection there

might be between the choice of discrete vorticity evolution and the properties of the scheme in the low Mach number limit.

As the derivation of the low Mach compliant multi-dimensional scheme for the Euler equations focuses on the low Mach number limit, and in particular on scalings $\mathcal{O}(\epsilon^{-1})$, there is ample possibility to improve this scheme further. Low Mach compliance does not provide any conditions for terms that scale as ϵ^0 or higher. They have to come from further considerations, be it vorticity transport, or something else. To investigate the possible extensions of this scheme is subject of future work.

The predictive power of stationarity preservation for acoustics, and the first results for the Euler equations seem to indicate that this work will help to advance numerical methods for the Euler equations in multiple spatial dimensions.

Appendix

The begemot code

BEGEMOT is a conservative code that allows to solve a variety of equations using finite volume methods on structured and unstructured grids. It is written in Java, and due to its object-oriented design it allows quick implementation of new features. In particular inheritance allows to reuse existing parts of the code, thus reducing redundancy. It is freely available from

`https://bitbucket.org/sturzhang/begemot`

The code is under `git` version control ([CS14]). For a number of Figures in this thesis commit hashes (in short form) are given: A *commit* is the state of the code as it was at a certain moment in time. With the hash it is possible to retrieve such a previous version of the code, compile and run it again. This allows to obtain the entire dataset that was used for the plot.

Contents (detailed)

List of definitions

Bibliography

[AG15] Debora Amadori and Laurent Gosse. *Error Estimates for Well-Balanced Schemes on Simple Balance Laws: One-Dimensional Position-Dependent Models*. Springer, 2015.

[Asa87] Kiyoshi Asano. On the incompressible limit of the compressible euler equation. *Japan Journal of Applied Mathematics*, 4(3):455–488, 1987.

[Bar17a] W Barsukow. Stationarity preserving schemes for multi-dimensional linear systems. *submitted*, 2017.

[Bar17b] Wasilij Barsukow. Stationarity and vorticity preservation for the linearized euler equations in multiple spatial dimensions. In *International Conference on Finite Volumes for Complex Applications*, pages 449–456. Springer, 2017.

[BEK+17] Wasilij Barsukow, Philipp VF Edelmann, Christian Klingenberg, Fabian Miczek, and Friedrich K Röpke. A numerical scheme for the compressible low-mach number regime of ideal fluid dynamics. *Journal of Scientific Computing*, 72(2):623–646, 2017.

[BEKR17] Wasilij Barsukow, Philipp VF Edelmann, Christian Klingenberg, and Friedrich K Röpke. A low-mach roe-type solver for the euler equations allowing for gravity source terms. *ESAIM: Proceedings and Surveys*, 58:27–39, 2017.

[BK17] W Barsukow and C Klingenberg. Exact solution and a truly multidimensional godunov scheme for the acoustic equations. *submitted*, 2017.

[BLMY17] Georgij Bispen, Mária Lukáčová-Medvid'ová, and Leonid Yelash. Asymptotic preserving imex finite volume schemes for low mach number euler equations with gravitation. *Journal of Computational Physics*, 335:222–248, 2017.

[BM05] Philipp Birken and Andreas Meister. Stability of preconditioned finite volume schemes at low mach numbers. *BIT Numerical Mathematics*, 45(3):463–480, 2005.

[CDK12] Floraine Cordier, Pierre Degond, and Anela Kumbaro. An asymptotic-preserving all-speed scheme for the euler and navier-stokes equations. *Journal of Computational Physics*, 231(17):5685–5704, 2012.

[CDLK15] Elisabetta Chiodaroli, Camillo De Lellis, and Ondřej Kreml. Global ill-posedness of the isentropic system of gas dynamics. *Communications on Pure and Applied Mathematics*, 68(7):1157–1190, 2015.

[CGK13] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. Large time step and asymptotic preserving numerical schemes for the gas dynamics equations with source terms. *SIAM Journal on Scientific Computing*, 35(6):A2874–A2902, 2013.

[CH62] R Courant and D Hilbert. Methods of mathematical physics. vol. ii: Partial differential equations.(vol. ii by r. courant.). *Interscience, New York*, 1962.

[CK15] Praveen Chandrashekar and Christian Klingenberg. A second order well-balanced finite volume scheme for euler equations with gravity. *SIAM Journal on Scientific Computing*, 37(3):B382–B402, 2015.

[CS14] Scott Chacon and Ben Straub. *Pro git*. Apress, 2014.

[Del10] Stéphane Dellacherie. Analysis of godunov type schemes applied to the compressible euler system at low mach number. *Journal of Computational Physics*, 229(4):978–1016, 2010.

[DJOR16] Stéphane Dellacherie, Jonathan Jung, Pascal Omnes, and P-A Raviart. Construction of modified godunov-type schemes accurate at any mach number for the compressible euler system. *Mathematical Models and Methods in Applied Sciences*, 26(13):2525–2615, 2016.

[DJY07] Pierre Degond, S Jin, and J Yuming. Mach-number uniform asymptotic-preserving gauge schemes for compressible flows. *Bulletin-Institute of Mathematics Academia Sinica*, 2(4):851, 2007.

[DLS10] Camillo De Lellis and László Székelyhidi. On admissibility criteria for weak solutions of the euler equations. *Archive for rational mechanics and analysis*, 195(1):225–260, 2010.

[DLV17] Giacomo Dimarco, Raphaël Loubère, and Marie-Hélène Vignal. Study of a new asymptotic preserving scheme for the euler system in the low mach number limit. *SIAM Journal on Scientific Computing*, 39(5):A2099–A2128, 2017.

[DOR10] Stéphane Dellacherie, Pascal Omnes, and Felix Rieper. The influence of cell geometry on the godunov scheme applied to the linear wave equation. *Journal of Computational Physics*, 229(14):5315–5338, 2010.

[Ebi77] David G Ebin. The motion of slightly compressible fluids viewed as a motion with strong constraining force. *Annals of mathematics*, pages 141–200, 1977.

[EL98] Jack R Edwards and Meng-Sing Liou. Low-diffusion flux-splitting methods for flows at all speeds. *AIAA journal*, 36(9):1610–1617, 1998.

[ER13] Timothy A Eymann and Philip L Roe. Multidimensional active flux schemes. In *21st AIAA computational fluid dynamics conference*, 2013.

[Eva98] Lawrence C Evans. Partial differential equations. *Graduate Studies in Mathematics*, 19, 1998.

[FG17] Emmanuel Franck and Laurent Gosse. Stability of a kirchhoff-roe scheme for multi-dimensional linearized euler systems. *preprint*, 2017.

[FKKM17] Eduard Feireisl, Christian Klingenberg, Ondřej Kreml, and Simon Markfelder. On oscillatory solutions to the complete euler system. *arXiv preprint arXiv:1710.10918*, 2017.

[GC90] Philip M Gresho and Stevens 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.

[GM04] Hervé Guillard and Angelo Murrone. On the behavior of upwind schemes in the low mach number limit: Ii. godunov type schemes. *Computers & fluids*, 33(4):655–675, 2004.

[God97] Sergey Konstantinovich Godunov. *Vospominaniya o raznostnyh shemah: doklad na mezhdunarodnom simpoziume "Metod Godunova v gazovoy dinamike", Michigan 1997*. Nauchnaya Kniga, 1997.

[God08] Sergei K Godunov. Reminiscences about numerical schemes. *arXiv preprint arXiv:0810.0649*, 2008.

[Gos01] Laurent Gosse. A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms. *Mathematical Models and Methods in Applied Sciences*, 11(02):339–365, 2001.

[GS64] IM Gelfand and GE Shilov. Generalized functions. vol. 1. properties and operations. translated from the russian by eugene saletan, 1964.

[Gui09] Hervé Guillard. On the behavior of upwind schemes in the low mach number limit. iv: P0 approximation on triangular and tetrahedral cells. *Computers & Fluids*, 38(10):1969–1972, 2009.

[GV99] Hervé Guillard and Cécile Viozat. On the behaviour of upwind schemes in the low mach number limit. *Computers & fluids*, 28(1):63–86, 1999.

[GZI+76] So K Godunov, AV Zabrodin, M Ia Ivanov, AN Kraiko, and GP Prokopov. Numerical solution of multidimensional problems of gas dynamics. *Moscow Izdatel Nauka*, 1, 1976.

[HJL12] Jeffrey Haack, Shi Jin, and Jian-Guo Liu. An all-speed asymptotic-preserving method for the isentropic euler and navier-stokes equations. *preprint*, 2012.

[Hör13] Lars Hörmander. *Linear partial differential operators*, volume 116. Springer, 2013.

[Iso87] Hiroshi Isozaki. Wave operators and the incompressible limit of the compressible euler equation. *Communications in mathematical physics*, 110(3):519–524, 1987.

[Jin99] Shi Jin. Efficient asymptotic-preserving (ap) schemes for some multiscale kinetic equations. *SIAM Journal on Scientific Computing*, 21(2):441–454, 1999.

[Joh78] Fritz John. Partial differential equations. *Applied Mathematical Sciences*, 1, 1978.

[JT06] Rolf Jeltsch and Manuel Torrilhon. On curl-preserving finite volume discretizations for shallow water equations. *BIT Numerical Mathematics*, 46(1):35–53, 2006.

[Kle95] R Klein. Semi-implicit extension of a godunov-type scheme based on low mach number asymptotics i: One-dimensional flow. *Journal of Computational Physics*, 121(2):213–237, 1995.

[KLN91] H-O Kreiss, J Lorenz, and MJ Naughton. Convergence of the solutions of the compressible to the solutions of the incompressible navier-stokes equations. *Advances in Applied Mathematics*, 12(2):187–214, 1991.

[KM81] Sergiu Klainerman and Andrew Majda. Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids. *Communications on Pure and Applied Mathematics*, 34(4):481–524, 1981.

[KM14] Roger Käppeli and S Mishra. Well-balanced schemes for the euler equations with gravitation. *Journal of Computational Physics*, 259:199–219, 2014.

[LeV98] Randall 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.

[LeV02] Randall J LeVeque. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge university press, 2002.

[LFS07] Alain Lerat, Fabrice Falissard, and Jacques Sidès. Vorticity-preserving schemes for the compressible euler equations. *Journal of Computational Physics*, 225(1):635–651, 2007.

[LG08] Xue-song Li and Chun-wei 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.

[LG13] Xue-song Li and Chun-wei Gu. Mechanism of roe-type schemes for all-speed flows and its application. *Computers & Fluids*, 86:56–70, 2013.

[Lio06] Meng-Sing Liou. A sequel to ausm, part ii: Ausm+-up for all speeds. *Journal of computational physics*, 214(1):137–170, 2006.

[LL13] Lev Davidovich Landau and Evgenii Mikhailovich Lifshitz. *Course of theoretical physics*. Elsevier, 2013.

[LMMW00] Maria Lukáčová-Medvidová, K Morton, and Gerald Warnecke. Evolution galerkin methods for hyperbolic systems in two space dimensions. *Mathematics of Computation of the American Mathematical Society*, 69(232):1355–1384, 2000.

[LR14] TB Lung and PL Roe. Toward a reduction of mesh imprinting. *International Journal for Numerical Methods in Fluids*, 76(7):450–470, 2014.

[LS02] Tatsian Li and Wancheng Sheng. The general riemann problem for the linearized system of two-dimensional isentropic flow in gas dynamics. *Journal of mathematical analysis and applications*, 276(2):598–610, 2002.

[Mic13] Fabian Miczek. *Simulation of low Mach number astrophysical flows*. PhD thesis, Technische Universität München, Dissertation, 2013.

[MR01] KW Morton and Philip L Roe. Vorticity-preserving lax–wendroff-type schemes for the system wave equation. *SIAM Journal on Scientific Computing*, 23(1):170–192, 2001.

[MRE15] F Miczek, FK Röpke, and PVF Edelmann. New numerical solver for flows at various mach numbers. *Astronomy & Astrophysics*, 576:A50, 2015.

[MS01] Guy Métivier and Steve Schochet. The incompressible limit of the non-isentropic euler equations. *Archive for rational mechanics and analysis*, 158(1):61–90, 2001.

[MT11] Siddhartha Mishra and Eitan Tadmor. Constraint preserving schemes using potential-based fluxes. ii. genuinely multidimensional systems of conservation laws. *SIAM Journal on Numerical Analysis*, 49(3):1023–1045, 2011.

[MTW17] Charles W Misner, Kip S Thorne, and John Archibald Wheeler. *Gravitation*. Princeton University Press, 2017.

[NBA+14] Sebastian Noelle, Georgij Bispen, Koottungal Revi Arun, Maria Lukacova-Medvidova, and C-D Munz. A weakly asymptotic preserving low mach number scheme for the euler equations of gas dynamics. *SIAM Journal on Scientific Computing*, 36(6):B989–B1024, 2014.

[O'N83] Barrett O'Neill. *Semi-Riemannian Geometry With Applications to Relativity, 103*, volume 103. Academic press, 1983.

[OSB+16] Kai Oßwald, Alexander Siegmund, Philipp Birken, Volker Hannemann, and Andreas 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.

[RB09] Felix Rieper and Georg Bader. The influence of cell geometry on the accuracy of upwind schemes in the low mach number regime. *Journal of Computational Physics*, 228(8):2918–2933, 2009.

[Rie10] Felix Rieper. On the dissipation mechanism of upwind-schemes in the low mach number regime: A comparison between roe and hll. *Journal of Computational Physics*, 229(2):221–232, 2010.

[Rie11] Felix Rieper. A low-mach number fix for roes approximate riemann solver. *Journal of Computational Physics*, 230(13):5263–5287, 2011.

[Roe81] Philip L Roe. Approximate riemann solvers, parameter vectors, and difference schemes. *Journal of computational physics*, 43(2):357–372, 1981.

[Roe17] Philip Roe. Multidimensional upwinding. *Handbook of Numerical Analysis*, 18:53–80, 2017.

[Rud91] Walter Rudin. Functional analysis. international series in pure and applied mathematics, 1991.

[Sch78] Laurent Schwartz. *Théorie des distributions*. Hermann Paris, 1978.

[Sch94] Steven Schochet. Fast singular limits of hyperbolic pdes. *Journal of differential equations*, 114(2):476–512, 1994.

[Sha48] Claude E Shannon. A mathematical theory of communication, part i, part ii. *Bell Syst. Tech. J.*, 27:623–656, 1948.

[Sid02] David Sidilkover. Factorizable schemes for the equations of fluid flow. *Applied numerical mathematics*, 41(3):423–436, 2002.

[Str12] Norbert Straumann. *General relativity and relativistic astrophysics*. Springer Science & Business Media, 2012.

[TD08] BJR 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.

[TF04] Manuel Torrilhon and Michael Fey. Constraint-preserving upwind methods for multidimensional advection equations. *SIAM journal on numerical analysis*, 42(4):1694–1728, 2004.

[TM71] Myron Tribus and Edward C McIrvine. Energy and information. *Scientific American*, 225(3):179–190, 1971.

[TMD+08] Ben Thornber, Andrew Mosedale, Dimitris Drikakis, David Youngs, and Robin JR Williams. An improved reconstruction method for compressible flows with low mach number features. *Journal of computational Physics*, 227(10):4873–4894, 2008.

[Tor09] Eleuterio F Toro. *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*. Springer Science & Business Media, 2009.

[Tur87] Eli Turkel. Preconditioned methods for solving the incompressible and low speed compressible equations. *Journal of computational physics*, 72(2):277–298, 1987.

[U+86] Seiji Ukai et al. The incompressible limit and the initial layer of the compressible euler equation. *Journal of Mathematics of Kyoto University*, 26(2):323–331, 1986.

[Vil02] Cédric Villani. A review of mathematical topics in collisional kinetic theory. *Handbook of mathematical fluid dynamics*, 1(71-305):3–8, 2002.

[Vil08] Cédric Villani. H-theorem and beyond: Boltzmann's entropy in today's mathematics. *Boltzmanns Legacy*, pages 129–143, 2008.

[VL06] Bram Van Leer. Upwind and high-resolution methods for compressible flow: From donor cell to residual-distribution schemes. *Communications in Computational Physics*, 1(192-206):138, 2006.

[WS95] Jonathan M Weiss and Wayne A Smith. Preconditioning applied to variable and constant density flows. *AIAA journal*, 33(11):2050–2057, 1995.