A General Approach to the Discretization of Hyperbolic Conservation Laws on Unstructured Grids

Dissertation
zur Erlangung des Doktorgrades
des Fachbereichs Mathematik
der Universität Hamburg

vorgelegt von

Arne Ahrend

aus Hildesheim

Hamburg
1999

Als Dissertation angenommen vom Fachbereich
Mathematik der Universität Hamburg

auf Grund der Gutachten von Prof. Dr. Thomas Sonar
und Prof. Dr. Rainer Ansorge

Hamburg, den 20. August 1999

Prof. Dr. Hans Daduna
Dekan des Fachbereichs Mathematik

Contents

Introduction **4**

1 Conservation Laws **7**
1.1 Fundamental Principles 7
1.2 Scalar Equations 10
Linear Advection 10
Characteristics 14
Discontinuities and Self Similarity 15
Weak Solutions 17
Diffusion and Entropy 18
Upwinding and the Equal Area Rule 20
Construction Criteria for Weak Solutions 22
Numerical Flux Functions 24
Burgers Equation 27
1.3 Hyperbolic Systems 28
The Covariant Formulation 28
Symmetry and Hyperbolicity 28
The Riemann Problem and Riemann Invariants 31
Numerical Flux Functions 38

2 Discretization **43**
2.1 Data Functionals 43
2.2 Unstructured Grids 47
Construction of Boxes 49
Collocation Grids 49
2.3 Boundary Conditions 51
2.4 Time Integration 52
2.5 The Finite Volume Method 53
2.6 Reconstruction 56
Approximation 58
Stencil Selection 65

Oscillation Indicators 67
2.7 Numerical Divergence Operator 68
Approximation of Linear Functionals 69
The Divergence Functional 74
2.8 The CFL Condition 76

3 The Euler Equations of Gas Dynamics 79
3.1 The Euler Flux Function 79
3.2 Riemann Invariants 82
3.3 Jump Relations 84
3.4 Numerical Flux Functions 85
3.5 Boundary Conditions 88

4 Collocation Schemes 89
4.1 The Lax-Friedrichs Scheme 89
Cartesian Grids 89
Unstructured Grids 92
4.2 Upwinding Schemes 94
Choice of Data Locations 95
4.3 Numerical Experiments 96
2D Shock Tube for the Euler Equations 96
Double Mach Reflection 98

A Measure of d-dimensional simplex 101

B List of Symbols 103

Bibliography 107

Danksagung 113

Abstract 115

Zusammenfassung 117

Lebenslauf 119

Introduction

Numerical simulation of physical, biological, chemical and even financial processes is becoming an increasingly widespread technique and replacing more time and money consuming experimental techniques like the building of models. Next to savings in expenditure this is due to the great flexibility of using a computer that allows rapid adaption of the investigated configuration.

Many natural and technical processes are governed by simple principles, like minimization of convex functionals or conservation of certain quantities. This thesis is devoted to the study of numerical methods to simulate the latter ones. In particular we develop a collocation method for hyperbolic conservation laws which is capable of adequately resolving strong shocks in transonic flow fields. The robustness and accuracy of the method is demonstrated by certain well established test cases for the Euler equations of gas dynamics.

Given the great success of finite volume methods (besides their formal elegance) considering collocation methods today perhaps requires some justification. The finite volume method combines a discrete integral formulation of the conservation principle with a rich geometric data structure. The state of the art in two dimensions is marked by adaptive methods using either cartesian grids or conforming simplex grids with boxes. Simplex grids in particular allow flexible and automatic discretization of complex geometries and we focus our considerations on equally unstructured and flexible grids.

One urgent demand arising from practical applications is the extension of contemporary methods from two to three space dimensions. Unfortunately, the generation of regular simplex grids in higher dimensions is a hard problem. For this reason one might look for ways to weaken the geometrical structure that underpins the finite volume method. A collocation scheme that operates on sets of smoothly scattered points and only requires some information about neighbourhood relations between these points would be easier to implement in higher dimensions than methods that work on tessalations.

In the course of developing such a method one is presented with formidable

technical difficulties. One important property of a program for simulating transonic gas flow, for example, is the capability of handling discontinuities. Finite volume methods achieve this via the solution of locally one dimensional Riemann problems. We have been able to design a similar mechanism for the collocation case by considering edges between neighbouring points and fluxes essentially directed along these edges.

Another motive for considering collocation functionals is the desire to use arbitrary trial spaces for which cell averaging functionals might be too expensive to handle. The polynomial recovery techniques widely used today work quite well, but they are largely based on heuristics. In particular, there is no rigorous theory of oscillation indicators and reconstruction weights available. Only very recently have mathematicians begun to look into these problems systematically. Collecting practical experience with generalizations of the methods that are so well-established today can perhaps help to pave a road towards a scheme that is both computationally efficient and founded on a comprehensive theory.

In chapter one of this thesis we discuss the features of hyperbolic conservation laws in some depth focusing on the topics relevant to schemes with upwind properties. We analyze the flux across discontinuities and classify numerical flux functions by the way they differ from the plain average of the fluxes to the left and right of the discontinuity.

The second chapter is concerned with discretization techniques. It introduces the grids we consider and presents a general framework of discretization which comprises both finite volume and collocation methods. We demonstrate uniform stability of the reconstruction process under similarity transformations of the grid for both collocation and cell averaging functionals.

In the third chapter we review the Euler equations of gas dynamics, perhaps the most important and well-studied hyperbolic system. It is closely related to the discussion in the first chapter, however, placing it here stresses that the discretization techniques developed in the second chapter are not specific to the Euler equations.

The fourth chapter finally contains a few remarks on our early attempts and numerical examples for some commonly accepted test cases for Euler computations generated with the current version of our collocation method.

Chapter 1

Conservation Laws

1.1 Fundamental Principles

If a quantity M is conserved within a region Ω, any change of the amount of M contained in Ω corresponds to transport of M across the boundary of Ω.

Under the assumption that there is a finite upper bound on the velocity with which either the quantity M or the region Ω may move along, the conservation principle has a far reaching immediate consequence:

> During a short interval in time all changes of the amount of M contained in Ω depend only on the distribution of M in a layer about the boundary of Ω. It does not matter how M is distributed deep inside Ω or far away from it.

In order to express the above principle in the language of mathematics, we introduce the following abstractions and definitions:

Definition 1.1. A **control volume** or **cell** is a fixed (not moving) compact polyhedron[1] $\Sigma \subset \mathbb{R}^d$

Polyhedra allow H^1 approximation of smooth geometries, and in order to improve the geometric approximation, it may sometimes be desirable to relax the above definition in the following way: given a "reasonably good" H^k approximation of a sufficiently smooth geometrical object, the surfaces of the control volumes may be modified to yield an H^{k+1} approximation of the object under consideration.

More general notions of a control volume may be envisaged, like that of a connected bounded set whose boundary has a piecewise continuous outer

[1] a non empty connected set with $\overline{\operatorname{int}\Sigma} = \Sigma$ whose boundary is formed by a finite number of subsets of hyperplanes

normal $\vec{n}$. The key point in any such definition is the availability of some variant of the divergence theorem, but there would be no substantial benefit in using those generalizations here.

Let M take values in $\mathbb{R}^s$ and $M(t, \Sigma)$ denote the amount of M in the control volume Σ at time t. Finally assume that the movement of M at time t and point $\vec{x}$ is described by a flux $j(t, \vec{x}) \in \mathbb{R}^{s \times d}$. Now the conservation principle can be stated in integral formulation:

$$M(t, \Sigma) = M(t_0, \Sigma) - \int_{t_0}^{t} \int_{\partial\Sigma} j\vec{n}\, do.$$

Transport out of Σ decreases the components of $M(t, \Sigma)$, but for such transport the integrand on the right hand side is positive, as $\vec{n}$ is the outer normal.

Definition 1.2. A function $u : \mathbb{R} \times \mathbb{R}^d \to S$ is called density of M at the point $\vec{x} \in \mathbb{R}^d$ at time t, if for any control volume $\Sigma \subset \mathbb{R}^d$

$$M(t, \Sigma) = \int_{\Sigma} u(t, \vec{x})\, dV.$$

$S \subset \mathbb{R}^s$ is called state space and $\mathbb{R}^d$ physical space or just space.

Throughout this thesis we will only consider quantities which have densities and speak of the conservation of the abstract quantities and their densities synonymously.

The conservation principle may now be restated in terms of the density:

$$\int_{\Sigma} u(t, \vec{x})\, dV = \int_{\Sigma} u(t_0, \vec{x})\, dV - \int_{t_0}^{t} \int_{\partial\Sigma} j\vec{n}\, do \tag{1.1a}$$

or after differentiation with respect to time

$$\frac{d}{dt} \int_{\Sigma} u(t, \vec{x})\, dV = - \int_{\partial\Sigma} j\vec{n}\, do. \tag{1.1b}$$

If the functions in equations (1.1) are sufficiently smooth, we may swap differentiation with respect to time and spatial integration on the left hand side and use the divergence theorem on the right hand side to obtain[2]

$$\int_{\Sigma} \frac{\partial}{\partial t} u(t, \vec{x})\, dV = - \int_{\Sigma} \mathrm{div} j\, dV.$$

[2]All vector operations are carried out on each state component separately.

For the class of Σ's we are considering this implies that the integrands must match pointwise, hence

$$\frac{\partial}{\partial t}u + \operatorname{div} j = 0. \tag{1.2}$$

Equation (1.2) is called the differential form of the conservation principle. For physical phenomena the flux should not explicitly depend on the time and space coordinates, since the forms of the laws of nature should not hinge on any particular frame of reference imposed by an observer [Ein05]. If the flux thus depends on u alone and is C^1, then $j = \boldsymbol{F} \circ u$ with a (smooth case) flux function $\boldsymbol{F} := (\boldsymbol{F}^1, \ldots, \boldsymbol{F}^d) : S \to \mathbb{R}^{s\times d}$ with $\boldsymbol{F}^k : \mathbb{R}^s \to \mathbb{R}^s$, and equation (1.2) can be stated in quasi-linear form:

$$\frac{\partial}{\partial t}u + \sum_{k=1}^{d} \frac{\partial \boldsymbol{F}^k}{\partial u}\,\frac{\partial u}{\partial x^k} = 0. \tag{1.3}$$

We will assume throughout this thesis that $\boldsymbol{F} \in C^2(S \to \mathbb{R}^{s\times d})$.

The problems we will study in this thesis consist of a conservation law in either of the above forms, a prescribed domain Ω, an initial density distribution $u_0(\vec{x}) = u(t_0, \vec{x})$ on Ω and boundary conditions on the fluxes across the boundary of Ω: for $\vec{x} \in \partial\Omega$

$$\boldsymbol{F}(u(t, \vec{x}))\vec{n} = \boldsymbol{B}(u, t, \vec{x}). \tag{1.4}$$

Restricting $\boldsymbol{F}$ to be dependent on u alone precludes modeling inhomogenities in space, source terms, certain kinds of boundary conditions and diffusive effects. In many applications, however, the flux function is simply a sum of a convective term (depending on u alone), a diffusive term that depends essentially on ∇u alone and source terms which mainly depend on space and time coordinates. Furthermore, these additional terms satisfy certain regularity conditions, cf. [Maj84] for a detailed discussion, and may consequently be regarded as small corrections to the convective term. A flux with a diffusive component has the form

$$\boldsymbol{F}(u) - A\nabla u$$

where $A = A(u, t, \vec{x})$ is always a diagonalizable positive semi-definite matrix. Roughly speaking the term $-A\nabla$ contributes downhill transport, locally levelling u. If A had negative eigenvalues, then there would occur uphill transport, causing small differences to blow up, similar to the behaviour of a "backward heat transport equation". The solution to such a problem exponentially blows up in time, and a numerical scheme simulating it cannot be stable.

Modeling just the convective flux entails a number of interesting difficulties – the spontaneous generation of discontinuities is perhaps the most notable of these – that translate into certain strategies for implementing numerical schemes for conservation laws. In this thesis we shall almost exclusively contemplate convective phenomena.

1.2 Scalar Equations

Linear Advection

Consider an initial scalar density distribution $u_0 : \mathbb{R}^d \to \mathbb{R}$ at time t_0 being shifted by multiples of a constant vector $\vec{\nu} \in \mathbb{R}^d$:

$$u(t, \vec{x}) := u_0(\vec{x} - \vec{\nu}(t - t_0)). \tag{1.5}$$

Obviously u is conserved. Furthermore, if u_0 is a smooth function, $u(t, \cdot)$ will always be smooth, as it is simply a shifted version of u_0. We may thus differentiate with respect to time and space:

$$\left.\frac{\partial}{\partial t}\right|_{t,\vec{x}} u = -\vec{\nu} \cdot \nabla|_{\vec{x}-\vec{\nu}(t-t_0)} u_0$$
$$\nabla|_{t,\vec{x}} u = \nabla|_{\vec{x}-\vec{\nu}(t-t_0)} u_0$$

Taking into account that $\operatorname{div} \vec{\nu} u_0 = \vec{\nu} \cdot \nabla u_0$ we infer that the differential form (1.2) of the conservation law is satisfied, if and only if we define the flux by

$$j := \boldsymbol{F}(u) := u\vec{\nu}^{\mathrm{t}}.$$

We now apply the integral formulation of equations (1.1) of the conservation principle to a certain class of discontinuous functions u_0 in order to derive a definition of the flux which will satisfy (1.5) at discontinuities. While there is no hope of managing a completely random function u_0, the conservation principle may be successfully applied to an initial distribution u_0 whose discontinuities are aligned with the interfaces of suitably chosen control volumes. It turns out that the flux across such a discontinuity is a function of the states at both sides of the discontinuity and the direction normal to it. This function is called the **Riemann solver** for the flux $\boldsymbol{F}$. Later we will develop approximate Riemann solvers for non-linear systems of hyperbolic conservation laws.

Proposition 1.3. Let Σ be a cell. Assume that $u_0 : \mathbb{R}^d \to \mathbb{R}$ is a bounded piecewise continuous function which is continuous inside Σ and can be continuously extended to Σ from the inside and piecewise continuously to $\mathbb{R}^d \setminus (\operatorname{int}\Sigma)$ from the outside. On the boundary of Σ define almost everywhere

$$\begin{aligned} u_i(\vec{x}) &:= \lim_{\vec{y}\to\vec{x}} u_0(\vec{y}) \quad (\vec{y} \in \operatorname{int}\Sigma) \\ u_o(\vec{x}) &:= \lim_{\vec{y}\to\vec{x}} u_0(\vec{y}) \quad (\vec{y} \notin \Sigma). \end{aligned} \tag{1.6}$$

Then the integral formulation (1.1) of the conservation principle is satisfied, if and only if we define the Riemann solver

$$\boldsymbol{F}(u_i, u_o, \vec{n}) := j\vec{n} := \begin{cases} \vec{\nu}\cdot\vec{n}\, u_i & \text{if } \vec{\nu}\cdot\vec{n} \geqslant 0 \\ \vec{\nu}\cdot\vec{n}\, u_o & \text{if } \vec{\nu}\cdot\vec{n} < 0 \end{cases} \tag{1.7}$$

almost everywhere on $\partial\Sigma$.

Before presenting the proof we observe that the definition of u_i and u_o in (1.6) depends on the particular Σ under consideration, but the Riemann solver $\boldsymbol{F}(u_i, u_o, \vec{n})$ in (1.7) does not. On a common interface between any two control volumes Σ and Σ' with outer normals $\vec{n}$ and $\vec{n}'$ respectively one has almost everywhere: $\vec{n} = -\vec{n}'$, $u_i = u'_o$, $u_o = u'_i$ – the inner limit as seen from Σ is the outer limit as seen from Σ' and vice versa – and consequently $\boldsymbol{F}(u_i, u_o, \vec{n}) = j\vec{n} = -j\vec{n}' = -\boldsymbol{F}(u'_i, u'_o, \vec{n}')$.

Proof of proposition 1.3. We first establish by indirect proof that the definition in equation (1.7) is necessary. Let us assume that the Riemann solver could be defined otherwise.

There are then a unit vector $\vec{q} \in \mathbb{R}^d$ and two numbers $u_l, u_r \in \mathbb{R}$ for which a flux different from that of equation (1.7) will satisfy the integral form of the conservation law. Now let

$$u_0(\vec{x}) := \begin{cases} u_l & \text{if } \vec{x}\cdot\vec{q} \leqslant 0 \\ u_r & \text{if } \vec{x}\cdot\vec{q} > 0 \end{cases}.$$

Since the density distribution is known, we may compute the content of an oblique cylinder or prism Θ of height h and cross section – parallel to the hyperplane – A (Volume $V = Ah$) with one end initially on the hyperplane $\vec{x}\cdot\vec{q} = 0$ and outer normal $\vec{q}$ for this end (i.e. Θ lies initially to the left of the hyperplane, see figure 1.1). For $t \in [t_0, t_0 + h/\left|\vec{\nu}\cdot\vec{q}\right|)$ one has

$$M(t,\Theta) = \begin{cases} u_l A h & \text{if } \vec{\nu}\cdot\vec{q} \geqslant 0 \\ [(u_l - u_r)(t - t_0)\,\vec{\nu}\cdot\vec{q} + u_l h]A & \text{if } \vec{\nu}\cdot\vec{q} < 0 \end{cases}$$

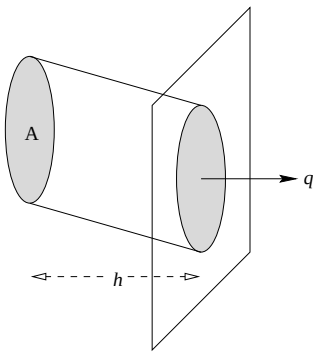

Figure 1.1: The cylinder Θ and the unit vector $\vec{q}$ on the hyperplane separating the states u_l and u_r. If $\vec{\nu} \cdot \vec{q} < 0$, the hyperplane moves left and u_r "enters" Θ, otherwise Θ will always contain only u_l.

which implies

$$\left.\frac{d}{dt}\right|_{t_0+} M(t,\Theta) = \begin{cases} 0 & \text{if } \vec{\nu}\cdot\vec{q} \geqslant 0 \\ (u_l - u_r)\,\vec{\nu}\cdot\vec{q}\,A & \text{if } \vec{\nu}\cdot\vec{q} < 0 \end{cases}.$$

On the other hand all fluxes across the jacket cancel out

$$\int_{\partial\Theta} j\vec{n}\,do = (\boldsymbol{F}(u_l, u_r, \vec{q}) - \boldsymbol{F}(u_l)\vec{q})A$$

and by virtue of the conservation principle in equations (1.1) we obtain a contradiction to our assumption:

$$\boldsymbol{F}(u_l, u_r, \vec{q}), q) = \boldsymbol{F}(u_l)\vec{q} - \frac{1}{A}\frac{d}{dt}\left|_{t_0+} M(t,\Theta) = \begin{cases} \vec{\nu}\cdot\vec{q}\,u_l & \text{if } \vec{\nu}\cdot\vec{q} \geqslant 0 \\ \vec{\nu}\cdot\vec{q}\,u_r & \text{if } \vec{\nu}\cdot\vec{q} < 0 \end{cases}.\right.$$

The proof of sufficiency is straightforward, but technical. By assumption u_0 can be continuously extended to Σ which is compact. Hence u_0 is uniformly continuous on any subset of Σ. Similarly, for a compact set Ξ with $\Sigma \subset \operatorname{int}\Xi$, u_0 is uniformly continuous on the intersection of any continuity component adjacent to Σ with $\Xi \setminus \Sigma$. Subdivide Σ into a finite number of closed convex polyhedra $\Sigma_1, \ldots, \Sigma_N$ and each Σ_k into subsets Θ such that

- each Θ is the intersection of Σ_k and a prism or cylinder with axis parallel to $\vec{\nu}$ and
- any intersections of edges of Σ_k with Θ are aligned with edges of Θ.

Θ is convex and the flux across its jacket – the part of its boundary formed by the jacket of a cylinder – is pointwise zero, because $\vec{\nu} \cdot \vec{n} = 0$ for any normal $\vec{n}$ on the jacket.

Denote by Θ_1 the surface part of Θ for which $\vec{\nu} \cdot \vec{n} < 0$, and by Θ_2 the opposite end. Now by a reasoning similar to that above and the Fubini theorem on integration on product spaces

$$\begin{aligned} M(t,\Theta) &= M(t_0,\Theta) + \int_{\Theta_1}\int_{t_0}^{t} u(\tau,\vec{x})\,d\tau\,do - \int_{\Theta_2}\int_{t_0}^{t} u(\tau,\vec{x})\,d\tau\,do \\ &= M(t_0,\Theta) + \int_{\Theta_1}\int_{t_0}^{t} u_0(\vec{x}-\tau\vec{\nu})\,d\tau\,do - \int_{\Theta_2}\int_{t_0}^{t} u_0(\vec{x}-\tau\vec{\nu})\,d\tau\,do \end{aligned}$$

and uniform continuity permits swapping integration over $\Theta_{1,2}$ and the limes of the difference quotient

$$\begin{aligned} \left.\frac{d}{dt}\right|_{t_0+} M(t,\Theta) &= \int_{\Theta_1} \|\vec{\nu}\| \lim_{\tau\searrow t_0} u_0(\vec{x}-\tau\vec{\nu})\,do - \int_{\Theta_2} \|\vec{\nu}\| \lim_{\tau\searrow t_0} u_0(\vec{x}-\tau\vec{\nu})\,do \\ &= \int_{\Theta_1} \|\vec{\nu}\| \lim_{\tau\searrow 0} u(t_0,\vec{x}-\tau\vec{\nu})\,do - \int_{\Theta_2} \|\vec{\nu}\| \lim_{\tau\searrow 0} u(t_0,\vec{x}-\tau\vec{\nu})\,do \\ &= -\int_{\partial\Theta} \boldsymbol{F}(u_i,u_o,\vec{n})\,do. \end{aligned}$$

Summation over all Θ completes the proof. □

We conclude this paragraph by summarizing (without proof) some algebraic properties of the Riemann solver:

Lemma 1.4. Let $\vec{n} \in \mathbb{R}^d$ be an arbitrary unit vector. The Riemann solver of equation (1.7)

1. is consistent with the flux function in the following way:

$$\left|\boldsymbol{F}(u_i,u_o,\vec{n}) - \frac{\boldsymbol{F}(u_i)+\boldsymbol{F}(u_o)}{2}\vec{n}\right| \le \|\vec{\nu}\|_{\mathbb{R}^d} \frac{|u_i-u_o|}{2},$$

2. may equivalently be written as

$$\boldsymbol{F}(u_i,u_o,\vec{n}) = \frac{\boldsymbol{F}(u_i)+\boldsymbol{F}(u_o)}{2}\vec{n} + |\vec{\nu}\cdot\vec{n}|\,\frac{u_i-u_o}{2}$$

3. and obeys

$$\boldsymbol{F}(u_i,u_o,\vec{n}) = -\boldsymbol{F}(u_o,u_i,-\vec{n}).$$

Characteristics

Direction and velocity of the density profiles propagation in the linear advection equation above were those of the vector $\vec{\nu} \in \mathbb{R}^d$ in

$$\frac{\partial}{\partial t} u + \vec{\nu} \cdot \nabla u = 0.$$

Let us now consider the case of a scalar conservation law with a non linear differentiable flux $j = \boldsymbol{F}(u) \in \mathbb{R}^{1\times d}$ depending on u alone. Let $C \subset \mathbb{R}$ be an interval of length greater than zero.

Definition 1.5. A continuous function $\chi : C \to \mathbb{R}^d$ is called a characteristic curve for the quasi-linear equation (1.3), if for $t_0 \in C$ fixed and any $t \in C$:

$$u(t, \chi(t)) = u(t_0, \chi(t_0)). \tag{1.8}$$

In the scalar case the characteristics are essentially straight lines in the direction of $\boldsymbol{F}'(u)$. This implies that $\boldsymbol{F}' \circ u$ plays the rôle of a characteristic velocity field: in smooth regions of u small local phenomena travel at velocity $\boldsymbol{F}'(u)$.

Lemma 1.6. With the same expressions as in the preceding definition assume that $\chi : C \to \mathbb{R}^d$ is differentiable and u is smooth on an open set containing $\big\{(t, \chi(t)) \in C \times \mathbb{R}^d : t \in C\big\}$. Then the lines defined by

$$\chi^{\mathsf{t}}(t) := \chi^{\mathsf{t}}(t_0) + (t - t_0)\boldsymbol{F}'(u(t_0, \chi(t_0))) \tag{1.9}$$

are characteristics.

Proof. Differentiate (1.8) with respect to time:

$$\begin{aligned}
0 &= \left.\frac{d}{d\tau}\right|_t u(\tau, \chi(\tau)) && \text{by (1.8)} \\
&= \left.\frac{\partial}{\partial\tau}\right|_t u + \left(\left.\frac{d}{d\tau}\right|_t \chi^{\mathsf{t}}\right) \nabla|_{(t,\chi(t))} u && \text{by the chain rule} \\
&= \left(\left.\frac{d}{d\tau}\right|_t \chi^{\mathsf{t}} - \boldsymbol{F}'\Big|_{u(t,\chi(t))}\right) \nabla|_{(t,\chi(t))} u && \text{by (1.3)} \\
&= \left(\left.\frac{d}{d\tau}\right|_t \chi^{\mathsf{t}} - \boldsymbol{F}'\Big|_{u(t_0,\chi(t_0))}\right) \nabla|_{(t,\chi(t))} u && \text{by (1.8)}
\end{aligned}$$

which is satisfied by $\chi^{\mathsf{t}}(t) = \chi^{\mathsf{t}}(t_0) + (t - t_0)\boldsymbol{F}'(u(t_0, \chi(t_0)))$. □

Remark 1.7. One important consequence of lemma 1.6 is that for a Lipschitz continuous flux function $\boldsymbol{F}$ the global Lipschitz constant $L_{\boldsymbol{F}}$ plays the rôle of a maximal signal velocity. On the other hand side, for a conservation law modeling physical phenomena with a finite upper bound on signal velocities, the flux function is Lipschitz continuous. In order to obtain a global Lipschitz constant for nonlinear equations it will generally be necessary to restrict the state of valid states such that an upper bound on $|\boldsymbol{F}'(u)|$ can be found. Regarding the modeled phenomenon this will hopefully only exclude extreme states for which the chosen model fails to be valid anyway or that are unphysical altogether, like particles moving faster than the speed of light.

Discontinuities and Self Similarity

Based on the characteristics we are now in a position to discuss the evolution of a given density profile u_0. It turns out that even for perfectly smooth initial data the evolving profile may develop discontinuities. Therefore we either need to consider weak solutions to the differential form of the conservation law or to abandon the differential form altogether. We define weak solutions in the next paragraph. Until then we use the term "weak solution" in a very loose fashion to denote a function pieced together from fragments of classical smooth solutions.

Compression Waves

By following the characteristics the evolution of the density profile may be constructed from the initial data, as long as the characteristics do not intersect. When they do cross, the classical concept of the solution is no longer valid, as a multi-valued solution would emerge at such a point.

Consider the following one dimensional example with a convex flux function $\boldsymbol{F} \in C^2(\mathbb{R} \to \mathbb{R})$, $\boldsymbol{F}''(u(t_0, x_0)) > 0$ and assume that $u_0 := u(t_0, \cdot) \in C^1(\mathbb{R} \to \mathbb{R})$ with $u_0'(\vec{x}_0) < 0$ and $\boldsymbol{F}'(u(t_0, x_0)) > 0$ and two characteristics χ_0 and χ_h, one passing through (t_0, x_0) and the other through $(t_0, x_0 + h)$:

$$\begin{aligned}\chi_h(t) &= (\vec{x}_0 + h) + (t - t_0)\boldsymbol{F}'(u_0(x_0 + h)) \\ &= (x_0 + h) + (t - t_0)\left[\boldsymbol{F}'(u_0(x_0)) + h\boldsymbol{F}''(u_0(x_0))u_0'(x_0) + hR_h\right]\end{aligned}$$

with $\lim_{h\to 0} R_h = 0$ and $\chi_0(t) = x_0 + (t - t_0)\boldsymbol{F}'(u_0(x_0))$. These will cross at time

$$t = t_0 - \frac{1}{u_0'(x_0)\boldsymbol{F}''(u_0(x_0)) + R_h}.$$

Letting $h \to 0$ we infer that after time

$$t^* := t_0 - \frac{1}{u_0'(x_0)\boldsymbol{F}''(u_0(x_0))} > t_0$$

the classical solution breaks down due to the fact that characteristics have crossed. In the context of weak solutions the multiplicity is removed by inserting one or more discontinuities **(compression waves** or **shocks)** in such a way that the conservation principle in equations (1.1) is satisfied and all characteristics go into the discontinuity.[3]

Once such a discontinuity has formed, it is not possible to tell – based on the weak solution – how long ago that happened. It is evident from this constructive process that the weak solution produces no "new" values, but stays within the range of the initial data. Lax [Lax71] formally shows that the time evolution of a step function with left and right states u_l and u_r respectively takes values between u_l and u_r and its variation is bounded by $|u_l - u_r|$.

Rarefaction Waves

For the configuration above, but this time with u_0 discontinuous at x_0, $u_0(x_0-) < u_0(x_0+)$ and $\boldsymbol{F}'$ monotonely increasing on $[u_0(x_0-), u_0(x_0+)]$, we choose as a weak solution a **rarefaction wave** or **fan** $u(t,x) := u^*$ where u^* is defined by

$$\boldsymbol{F}'(u^*) = \begin{cases} \dfrac{x - x_0}{t - t_0} & \text{for } \boldsymbol{F}'(u_0(x_0-)) < \dfrac{x - x_0}{t - t_0} < \boldsymbol{F}'(u_0(x_0+)) \\ \boldsymbol{F}'(u_0(x)) & \text{otherwise} \end{cases}.$$

This choice may be motivated by the observation that a smeared discontinuity – a large but finite gradient about x_0 – in the initial data u_0 would evolve that way, i.e. be spread further and further. In the case of the crossing characteristics above the smeared part would merely be resharpened.

Contact Discontinuities

If the characteristics are parallel (like in the linear advection case), a discontinuity may still slide along, such a situation is called "contact discontinuity". No characteristics go into a contact discontinuity and none come out of it. This fact is sometimes referred to by the statement that no matter crosses a contact discontinuity.

[3] Some authors refer to the multivalued function – before or without the insertion of any discontinuities – as the compression wave.

Self Similarity

In each of the above cases the weak solution we constructed could be represented by a function depending solely on the quotient of $x - x_0$ and $t - t_0$. The construction of the rarefaction wave followed the characteristics and explicitly ensured that

$$\xi := \frac{x - x_0}{t - t_0} = \boldsymbol{F}'(u(t, x)), \tag{1.10}$$

but also for the shocks and contact discontinuities one has

$$u(t, x) = w\left(\frac{x - x_0}{t - t_0}\right)$$

with $w(\xi) = u_l$ or $w(\xi) = u_r$ depending on ξ such that the integral form of the conservation principle in equations (1.1) is satisfied. Whenever w is differentiable, equation (1.10) implies

$$\left(-\frac{x - x_0}{(t - t_0)^2} + \frac{1}{t - t_0}\boldsymbol{F}'(w(\xi))\right) w'(\xi) = 0$$

and hence equation (1.3) holds.

Weak Solutions

A numerical method operating on the integral formulation (1.1) of the conservation principle approximates the solution based on values from a finite number of given control volumes. While for each of the chosen control volumes the integral form (1.1) is satisfied exactly, the "weakness" of the formulation originates from the fact that we only consider a finite number of **data functionals:** the average of the function on each control volume.

Setting out from the differential form (1.2) of the conservation principle we multiply equation (1.2) by suitable smooth **test functions** and formally integrate by parts to shift differentiation from u to the test function. For any compactly supported $\phi \in C_0^1(\mathbb{R}^d \to \mathbb{R})$ we have

$$\frac{d}{dt}\int_\Omega u\phi \, dV + \int_{\partial\Omega} (\boldsymbol{F} \circ u)\phi\vec{n} \, do - \int_\Omega (\boldsymbol{F} \circ \phi)\nabla\phi \, dV = 0$$

and after scaling Ω until $\operatorname{supp}\phi \subset \operatorname{int}\Omega$ the boundary integral vanishes:

$$\frac{d}{dt}\int_{\mathbb{R}^d} u\phi \, dV - \int_{\mathbb{R}^d} (\boldsymbol{F} \circ \phi)\nabla\phi \, dV = 0. \tag{1.11}$$

Equation (1.11) is called the weak form of the differential equation and its solutions are weak solutions to equation (1.2). Weak solutions are by no means unique. In particular the velocities at which discontinuities propagate are not necessarily uniquely determined. Therefore one has to look for additional criteria to single out the unique weak solution

- that is piecewise a classical smooth solution
- whose discontinuities are aligned with fragments of differentiable manifolds of codimension one in $\mathbb{R}^d$ and move at the desired propagation speeds.

Diffusion and Entropy

As long as the classic smooth solution of the differential form (1.2) of the conservation principle exists, it is invariant under PT-transformations, i.e. changing the signs of time and all space coordinates (parity). In fact, if $u(t, \vec{x})$ satisfies the differential form of the conservation principle (1.2), so does $u(\alpha t, \alpha \vec{x})$ for any fixed $\alpha \in \mathbb{R}$. Shocks, however, violate PT-invariance. After a PT-transformation a shock immediately dissolves into a rarefaction fan (and possibly a whole sequence of rarefaction fans and compression waves), irrespective of how long it existed before. One is therefore interested in additional conditions to break the PT-invariance immanent in (1.2), see for instance [AS97].

Contemplate the following – somewhat crude – mechanical analogue: We have seen that in smooth regions the characteristics are straight lines that converge and break at shocks. In the absence of friction a particle moving along with the flow field would follow a straight path at constant velocity, a characteristic. On passing a shock, however, the particle gets slowed down, so it is reasonable to look for a mechanism providing some kind of friction to dissipate part of the particles kinetic energy into heat, thus increasing the entropy.

One common way of doing this is the diffusion-entropy approach: treat (1.2) as the one-sided limit for $\varepsilon \searrow 0$ of

$$\frac{\partial}{\partial t} u^\varepsilon + \operatorname{div} j^\varepsilon = \varepsilon \, \Delta \, u^\varepsilon \tag{1.12}$$

with $j^\varepsilon = \boldsymbol{F}(u^\varepsilon)$. The solutions u^ε of (1.12) are known to be smooth, furthermore this equation is **not** invariant under PT-transformations. Now let $\boldsymbol{E} \in C^2(\mathbb{R} \to \mathbb{R})$ be an arbitrary strictly convex function and $g := \boldsymbol{E}' \boldsymbol{F}'$ with

antiderivative $\boldsymbol{G}$, the entropy flux. If u is differentiable twice, we have by the chain rule

$$\Delta(\boldsymbol{E} \circ u) = (\boldsymbol{E}' \circ u)\,\Delta\, u + (\boldsymbol{E}'' \circ u)\,\|\nabla u\|^2$$

and

$$\begin{aligned}\frac{\partial}{\partial t}(\boldsymbol{E} \circ u^\varepsilon) + \operatorname{div}(\boldsymbol{G} \circ u^\varepsilon) &= (\boldsymbol{E}' \circ u^\varepsilon)\frac{\partial}{\partial t}u^\varepsilon + (\boldsymbol{G}' \circ u^\varepsilon)\nabla u^\varepsilon \\ &= (\boldsymbol{E}' \circ u^\varepsilon)\frac{\partial}{\partial t}u^\varepsilon + (\boldsymbol{E}' \circ u^\varepsilon)(\boldsymbol{F}' \circ u^\varepsilon)\nabla u^\varepsilon \\ &= \varepsilon(\boldsymbol{E}' \circ u^\varepsilon)\,\Delta\, u^\varepsilon \\ &= \varepsilon\left[\Delta(\boldsymbol{E} \circ u^\varepsilon) - (\boldsymbol{E}'' \circ u^\varepsilon)\,\|\nabla u^\varepsilon\|^2\right]\end{aligned}$$

Existence and uniqueness of the limit function for $\varepsilon \searrow 0$ are guaranteed by the following theorem due to Kružkov of which its author claims that it also extends to systems [Kru70]:

Theorem 1.8. *Assume that the initial density function $u_0 \in L^\infty(\mathbb{R}^d \to \mathbb{R})$ and the flux $\boldsymbol{F} \in C^1(\mathbb{R} \to \mathbb{R})$ is Lipschitz continuous. Then the solutions u^ε of (1.12) converge almost everywhere in $[t_0; T] \times \mathbb{R}^d$ to a function $u \in L^\infty(\mathbb{R}^d \to \mathbb{R})$ as $\varepsilon \searrow 0$ and this limit function u is a solution of equations (1.1).*

In regions where $\lim_{\varepsilon \searrow 0} \|u^\varepsilon - u\| = 0$ in a sufficiently strong norm u is differentiable twice, and

$$\frac{\partial}{\partial t}(\boldsymbol{E} \circ u) + \operatorname{div}(\boldsymbol{G} \circ u) = 0.$$

Particularly at discontinuities of u such strong convergence will be unattainable, therefore we multiply (1.12) with a test function $\phi \in C_0^2(\mathbb{R}^{d+1} \to \mathbb{R}_{\geqslant 0})$ and integrate by parts to shift differentiation from u^ε to ϕ. Formally using the product rule

$$\phi\,\Delta\,\boldsymbol{E} = \operatorname{div}[\phi\nabla\boldsymbol{E}] - (\nabla\phi)(\nabla\boldsymbol{E}) = \boldsymbol{E}\,\Delta\,\phi + \operatorname{div}[\nabla(\phi\boldsymbol{E}) - 2\boldsymbol{E}\nabla\phi]$$

we obtain

$$\begin{aligned}&\phi\frac{\partial}{\partial t}u^\varepsilon + \phi\operatorname{div} j^\varepsilon = \varepsilon\phi\,\Delta\, u^\varepsilon = \\ &\underbrace{\varepsilon(\boldsymbol{E} \circ u^\varepsilon)\,\Delta\,\phi}_{=:R_1} + \underbrace{\varepsilon\operatorname{div}[\nabla(\phi(\boldsymbol{E} \circ u^\varepsilon)) - 2(\boldsymbol{E} \circ u^\varepsilon)\nabla\phi]}_{=:R_2} - \underbrace{\varepsilon(\boldsymbol{E}'' \circ u^\varepsilon)\phi\,\|\nabla u^\varepsilon\|^2}_{=:R_3}.\end{aligned}$$

The second term on the right hand side, R_2, permits using the divergence theorem. The ensuing boundary integral is zero, if the compact support of ϕ lies inside the domain of integration. In the course of his proof Kružkov shows that for any compact $K \subset \mathbb{R}^d$ the set $\{u^\varepsilon|_K : \varepsilon > 0\} \subset L^1(K \to \mathbb{R})$ is compact. Hence $\lim_{\varepsilon \searrow 0} R_1 = 0$, but R_3 resists further simplification and we only know $R_3 \geqslant 0$, as $\boldsymbol{E}$ is convex and $\phi \geqslant 0$. Therefore

$$\limsup_{\varepsilon \searrow 0} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \left(\phi \frac{\partial}{\partial t}(\boldsymbol{E} \circ u^\varepsilon) + \phi \operatorname{div}(\boldsymbol{G} \circ u^\varepsilon) \right) dV\, dt \leqslant 0.$$

For this reason we demand that the weak solutions of (1.2) should satisfy the following **entropy condition:** For any convex function $\boldsymbol{E} \in C^2(\mathbb{R} \to \mathbb{R})$ and $\boldsymbol{G}$ as constructed above

$$\frac{\partial}{\partial t}(\boldsymbol{E} \circ u) + \operatorname{div}(\boldsymbol{G} \circ u) \leqslant 0 \tag{1.13}$$

in a distributional sense. This may be generalized to arbitrary convex functions $\boldsymbol{E} : \mathbb{R} \to \mathbb{R}$, such functions are differentiable twice almost everywhere.

Upwinding and the Equal Area Rule

Suppose we knew in advance the velocity at which a discontinuity separating two constant states travels along. We fix a unit vector $\vec{n} \in \mathbb{R}^d$, a number $\alpha \in \mathbb{R}$ and assume that the discontinuity is initially aligned with the hyperplane $\vec{x} \cdot \vec{n} = \alpha$ and travels at velocity $\vec{\nu} \in \mathbb{R}^d$ maintaining constant height.

Denoting as "left" the region $\left\{\vec{x} \in \mathbb{R}^d : \vec{x} \cdot \vec{n} < \alpha\right\}$, as "right" the rest and by u_l and u_r the constant states left and right of the discontinuity respectively, u may be expressed as

$$u(t, \vec{x}) := \begin{cases} u_l & \text{if } (\vec{x} - (t - t_0)\vec{\nu}) \cdot \vec{n} < \alpha \\ u_r & \text{if } (\vec{x} - (t - t_0)\vec{\nu}) \cdot \vec{n} \geqslant \alpha \end{cases},$$

and we conclude by considering a control volume to the left of the discontinuity (the reasoning is quite similar to that in the indirect "necessity" part of the proof of proposition 1.3)

$$\boldsymbol{F}(u_l, u_r, \vec{n}) = \begin{cases} \boldsymbol{F}(u_l)\vec{n} & \text{if } \vec{\nu} \cdot \vec{n} > 0 \\ \boldsymbol{F}(u_l)\vec{n} - (u_l - u_r)\, \vec{\nu} \cdot \vec{n} & \text{if } \vec{\nu} \cdot \vec{n} \leqslant 0 \end{cases} \tag{1.14a}$$

or equivalently for a control volume to the right

$$\boldsymbol{F}(u_l, u_r, \vec{n}) = \begin{cases} \boldsymbol{F}(u_r)\vec{n} + (u_l - u_r)\, \vec{\nu} \cdot \vec{n} & \text{if } \vec{\nu} \cdot \vec{n} > 0 \\ \boldsymbol{F}(u_r)\vec{n} & \text{if } \vec{\nu} \cdot \vec{n} \leqslant 0 \end{cases}. \tag{1.14b}$$

The vector $\vec{n}$ above points from left to right which is in keeping with our use of u_i, u_o and the outer normal $\vec{n}$ pointing from the inside to the outside of a control volume. Putting equations (1.14) together we arrive at

Lemma 1.9 (Rankine-Hugoniot jump condition). The velocity $\vec{\nu}\cdot\vec{n}$ at which a discontinuity of constant height may travel in the direction $\vec{n}$ normal to it is subject to the following Rankine-Hugoniot jump condition:

$$(\boldsymbol{F}(u_l) - \boldsymbol{F}(u_r))\vec{n} = (u_l - u_r)\,\vec{\nu}\cdot\vec{n}. \tag{1.15}$$

The Rankine-Hugoniot jump condition is also a direct consequence of the conservation principle stated in equations (1.1), if we consider a cylinder Σ with axis parallel to $\vec{n}$ centered about the discontinuity and moving along with it. All fluxes across the jacket cancel out and $M(t,\Sigma)$ is constant – we may even let the height of Σ tend to zero. Therefore $(\boldsymbol{F}(u_l) - \vec{\nu}\,u_l)\cdot\vec{n} = (\boldsymbol{F}(u_r) - \vec{\nu}\,u_r)\cdot\vec{n}$.

Using equation (1.15) the flux across the discontinuity (1.14) may be more compactly expressed as

$$\boldsymbol{F}(u_l, u_r, \vec{n}) = \begin{cases} \boldsymbol{F}(u_l)\vec{n} & \text{if } \vec{\nu}\cdot\vec{n} > 0 \\ \boldsymbol{F}(u_r)\vec{n} & \text{if } \vec{\nu}\cdot\vec{n} \leqslant 0 \end{cases}. \tag{1.16}$$

Equation (1.16) is the mathematical expression of the upwinding principle: if the local transport ("wind") is from left to right, evaluate the flux function for the left state, otherwise for the right state, i.e. look against ("up") the wind. We stress the following features of equation (1.16):

1. Only the sign of $\vec{\nu}\cdot\vec{n}$ matters and
2. the Riemann solver does not jump at $\vec{\nu}\cdot\vec{n} = 0$, since the Rankine-Hugoniot condition (1.15) then implies $\boldsymbol{F}(u_l)\vec{n} = \boldsymbol{F}(u_r)\vec{n}$.

Let us investigate the situation a little closer. The problem is essentially one dimensional along $\vec{n}$ and we may think of the discontinuity as a step with lower level at u_b (bottom) and upper level at u_t (top). Now imagine that each point at height u of the step – including those in the vertical part – is advected with its characteristic speed $\boldsymbol{F}'(u)$. The propagation of the step then depends on the behaviour of $\boldsymbol{F}'$ on $[u_b, u_t]$.

If the step function and $\boldsymbol{F}'$ are correspondingly monotone, then the characteristics diverge and we choose a rarefaction fan as weak solution. If they have opposite monotone behaviour, the characteristics converge and we choose a compression wave[4] whose velocity is determined by (1.15). Following the characteristics in this case leads to a multi-valued solution, and the

[4] In this case a single discontinuity is sufficient.

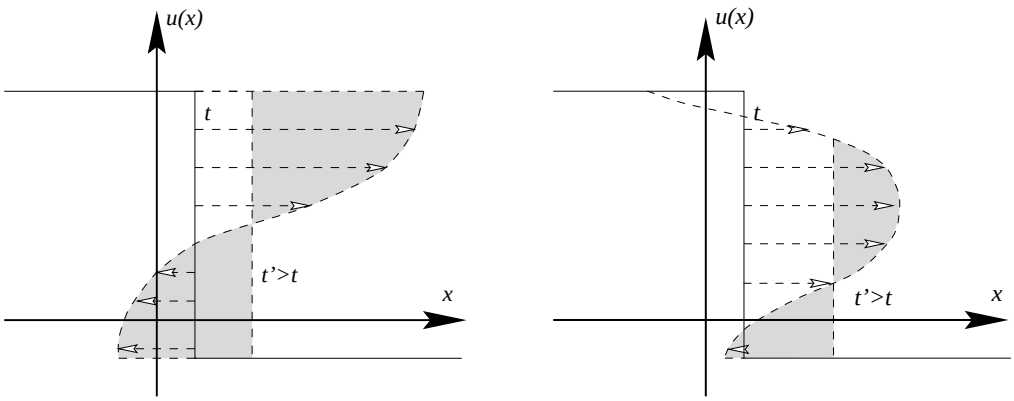

Figure 1.2: Equal area rule and the condition E. In the left part the flux function $\boldsymbol{F}$ is convex on $[u_b, u_t]$, the right part illustrates the desired behaviour of the weak solution for a more general $\boldsymbol{F}$. The shaded areas in each part have equal sizes.

insertion of the discontinuity must remove the multiplicity without violating the conservation principle. An equivalent expression for the conservation principle in terms of this geometric construction is the **equal area rule:**

$$(t - t_0) \int_{u_r}^{u_l} \boldsymbol{F}'(u)\vec{n}\, du = (t - t_0)(u_l - u_r)\, \vec{\nu} \cdot \vec{n} \tag{1.17}$$

which obviously implies (1.15).

However, for a nonconvex flux $\boldsymbol{F}$ with $\boldsymbol{F}'\vec{n}$ having several zeros (sonic points) the actual geometric construction of a suitable weak solution involves several different waves and can be quite complicated (cf. the right part of figure 1.2 for a simple example). Regarding the conservation principle we demand that for each sub-discontinuity whose left and right states are again denoted by u_l and u_r respectively $\vec{\nu} \cdot \vec{n}$ should satisfy equation (1.17) and hence the Rankine-Hugoniot jump condition of equation (1.15). An explicit expression for the desired weak solution in the scalar case is known, we present it in the next paragraph.

Construction Criteria for Weak Solutions

Above all the weak solutions we construct should satisfy the entropy condition (1.13). In this paragraph we present some more explicit consequences of equation (1.13). As noted before, all characteristics should go into a compression wave, and none come out of it, as time advances. This is crucial to the compression waves stability under small perturbations, as the

resharpening effect of the converging characteristics has to restore the discontinuity completely after any mollification. An obvious condition therefore is $\boldsymbol{F}'(u_l)\vec{n} > \vec{\nu}\cdot\vec{n} > \boldsymbol{F}'(u_r)\vec{n}$. Furthermore, after smearing the original discontinuity, it should be recovered without decaying into two (or more) smaller ones. This is expressed by **Oleinik's condition E:**

$$\frac{\boldsymbol{F}(u)-\boldsymbol{F}(u_l)}{u-u_l}\vec{n} \geqslant \vec{\nu}\cdot\vec{n} \geqslant \frac{\boldsymbol{F}(u)-\boldsymbol{F}(u_r)}{u-u_r}\vec{n} \qquad \text{for } u\in(u_b;u_t). \tag{1.18}$$

Take as an example the right part of figure 1.2: propagation of the discontinuity at its full initial height would, by the equal area rule, be slower than the smaller discontinuity with a rarefaction fan at the top end. We reject the first variant on the grounds that it violates (1.18). For a rigorous investigation of shock front stability we refer to Majda [Maj84].

Geometrically speaking, the areas we exchange in such a construction should not only be of equal size, but also as small as possible, because a weak solution constructed this way will not change dramatically, if we mollify the initial data. We demand that Oleinik's condition E (1.18) be satisfied across each discontinuity. Osher [Osh84] proves the following

Proposition 1.10. The (exact) solution $u(t,\vec{x}) = w(\xi)$ with

$$\xi := (\vec{x}-\vec{x}_0)\cdot\vec{n}/(t-t_0)$$

and initial data

$$u_0(\vec{x}) = u(t_0,\vec{x}) = \begin{cases} u_l & \text{if } \vec{x}\cdot\vec{n} \leqslant \vec{x}_0\cdot\vec{n} \\ u_r & \text{if } \vec{x}\cdot\vec{n} > \vec{x}_0\cdot\vec{n} \end{cases}$$

satisfies

$$\boldsymbol{F}(w(\xi))\vec{n} - \xi w(\xi) = \min_{u\in[u_l,u_r]}(\boldsymbol{F}(u)\vec{n}-\xi u) \qquad \text{if } u_l < u_r \tag{1.19a}$$

$$\boldsymbol{F}(w(\xi))\vec{n} - \xi w(\xi) = \max_{u\in[u_r,u_l]}(\boldsymbol{F}(u)\vec{n}-\xi u) \qquad \text{if } u_l > u_r. \tag{1.19b}$$

With this proposition the flux across the discontinuity at $\vec{x}=\vec{x}_0$ may be expressed by the Riemann solver

$$\boldsymbol{F}(u_l,u_r,\vec{n}) = \boldsymbol{F}(w(0)) = \begin{cases} \min\limits_{u\in[u_l,u_r]} \boldsymbol{F}(u)\vec{n} & \text{if } u_l < u_r \\ \max\limits_{u\in[u_r,u_l]} \boldsymbol{F}(u)\vec{n} & \text{if } u_l > u_r \end{cases}. \tag{1.20}$$

Furthermore, whenever w is differentiable, we infer

$$\frac{d}{d\xi}[\boldsymbol{F}(w(\xi))\vec{n} - \xi w(\xi)] = \underbrace{[\boldsymbol{F}'(w(\xi))\vec{n} - \xi]}_{=0 \text{ by equation (1.10)}} w'(\xi) - w(\xi)$$

and hence

$$w = -\frac{d}{d\xi}\left(\min_{u\in[u_l,u_r]}(\boldsymbol{F}(u)\vec{n} - \xi u)\right) \qquad \text{if } u_l < u_r$$
$$w = -\frac{d}{d\xi}\left(\max_{u\in[u_r,u_l]}(\boldsymbol{F}(u)\vec{n} - \xi u)\right) \qquad \text{if } u_l > u_r.$$

Conversely, the minimum or maximum in the preceeding equations are continuous with respect to ξ and differentiable almost everywhere. The jumps in the derivative of the minimum or the maximum are precisely the jumps of w.

Proposition 1.10 completely and explicitly describes the evolution of the desired weak solution to a scalar conservation law, but

- it does not generalize to systems,
- the Riemann solver of equation (1.20) is not continuously differentiable and
- even if we can, the desired weak solution in the case of systems may be expensive to compute.

In the following paragraph we consider approximations to the Riemann solver of equation (1.20) which have these features and discuss some of their properties.

Numerical Flux Functions

For a linear flux function each approximate Riemann solver in this paragraph is equivalent to the Riemann solver considered in lemma 1.4. Harten and Hyman [HH83] give a general survey of construction criteria for numerical flux functions, a compact discussion can also be found in Hirsch [Hir90]. We shall focus our presentation on the systematic addition of "upwinding" to the average of $\boldsymbol{F}(u_l)$ and $\boldsymbol{F}(u_r)$. A straightforward generalization of equation (1.16) estimates the propagation velocity of a discontinuity in the direction normal to it from the Rankine-Hugoniot jump condition. This leads to the numerical flux function of Roe [Roe81a, Roe81b]:

$$\boldsymbol{H}^{\text{Roe}}(u_l, u_r, \vec{n}) := \frac{\boldsymbol{F}(u_l) + \boldsymbol{F}(u_r)}{2}\vec{n} + |A(u_l, u_r)|\,\frac{u_l - u_r}{2} \tag{1.21a}$$

where

$$A(u,u) := \boldsymbol{F}'(u)\vec{n} \qquad \text{and} \tag{1.21b}$$

$$A(u_l,u_r) := \frac{\boldsymbol{F}(u_l)-\boldsymbol{F}(u_r)}{u_l-u_r}\vec{n} \qquad \text{if } u_l \neq u_r. \tag{1.21c}$$

If $\boldsymbol{F}\vec{n}$ is not monotone on $[u_b,u_t] := [\min\{u_l,u_r\},\max\{u_l,u_r\}]$, then the numerical flux function of Roe admits unphysical weak solutions that violate the entropy conditions, if the sonic point contained in $[u_b,u_t]$ is critical. This is mainly due to the fact that the Roe flux might stay too close – compared to equation (1.20) – to the average of left and right flux:

$$\begin{aligned}\left|\operatorname*{ext}_{u\in[u_b,u_t]}\boldsymbol{F}(u)\vec{n}-\frac{\boldsymbol{F}(u_l)+\boldsymbol{F}(u_r)}{2}\vec{n}\right| &\geqslant \left|\operatorname*{ext}_{u\in\{u_l;u_r\}}\boldsymbol{F}(u)\vec{n}-\frac{\boldsymbol{F}(u_l)+\boldsymbol{F}(u_r)}{2}\vec{n}\right| \\ &\geqslant \left|\frac{\boldsymbol{F}(u_l)-\boldsymbol{F}(u_r)}{2}\vec{n}\right| \\ &\geqslant \left|\boldsymbol{H}^{\mathrm{Roe}}(u_l,u_r,\vec{n})-\frac{\boldsymbol{F}(u_l)+\boldsymbol{F}(u_r)}{2}\vec{n}\right|,\end{aligned}$$

where 'ext' is either 'max' or 'min'. One common remedy to protect against unwanted weak solutions is the conspicuous addition of diffusion ("entropy fix") [RP84] (see also [SO89]), as the regularization introduced by smearing the initial discontinuity will force it to decay into several smaller ones, if it fails to be stable under such perturbation. Instead of explicitly detecting the sonic points and adding the necessary entropy fixes we consider a different and slightly more sophisticated generalization of the upwinding principle using the derivative of $\boldsymbol{F}\vec{n}$. In the case of hyperbolic systems it is precisely the diagonalizability of the Jacobi matrix of $\boldsymbol{F}\vec{n}$ which enables us to generalize the upwinding principle to systems. In equation (1.16) we let $\sigma := \operatorname{sign}(\vec{\nu}\cdot\vec{n})$ and rewrite it in the following way:

$$\boldsymbol{F}(u_l,u_r,\vec{n}) = \frac{\boldsymbol{F}(u_l)+\boldsymbol{F}(u_r)}{2}\vec{n} + \underbrace{\left(\frac{\sigma}{u_l-u_r}\int_{u_r}^{u_l}\boldsymbol{F}'(u)\vec{n}\,du\right)}_{\text{upwinding term}}\frac{u_l-u_r}{2} \tag{1.22}$$

Because of the equal area rule (1.17) the upwinding term is nonnegative. In a way, equation (1.16) represents the minimal upwinding necessary to barely advect the discontinuity. If the upwinding term were made smaller, the resulting flux would not even suffice to advect the discontinuity at the velocity determined by the Rankine-Hugoniot jump condition (1.15), this would lead to a pile-up behind the discontinuity, causing its height to rise. Such upwinding deficit is a source of numerical instability. Making the upwinding

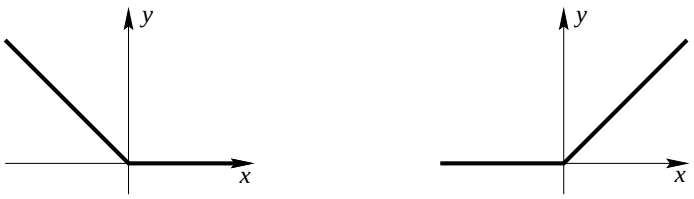

Figure 1.3: Negative part x^- (left) and positive part function x^+ (right).

term larger, on the other hand, increases downhill transport, which is very similar to the effects of diffusion.

In fact, equation (1.20) represents the Riemann solver we seek to approximate and, compared to (1.16), it corresponds to a probably larger upwinding term. Several practically relevant numerical flux functions to approximate the flux across a discontinuity are constructed by slightly increasing the upwinding term of equation (1.22):

$$0 \leqslant \frac{1}{u_l - u_r} \int_{u_r}^{u_l} \sigma \boldsymbol{F}'(u)\vec{n}\, du \leqslant \underbrace{\frac{1}{u_l - u_r} \int_{u_r}^{u_l} |\boldsymbol{F}'(u)\vec{n}|\, du}_{\text{Enquist-Osher term}} \leqslant L_{\boldsymbol{F}}^{\text{local}} \leqslant L_{\boldsymbol{F}}^{\text{global}}.$$

Substituting the Enquist-Osher term for the upwinding term eliminates the dependence of σ and yields the flux function $\boldsymbol{H}^{\text{EO}}$ of Engquist and Osher [EO80, EO81]. It can be stated in three equivalent ways, the third (1.23c) of these is perhaps the most suggestive in terms of implementing upwinding. For ease of presentation we define the positive and negative part functions: for '+' and '−' in '±' separately let $^{\pm} : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$, $x^{\pm} := (|x| \pm x)/2$.

$$\boldsymbol{H}^{\text{EO}}(u_l, u_r, \vec{n}) := \frac{\boldsymbol{F}(u_l) + \boldsymbol{F}(u_r)}{2}\vec{n} + \frac{1}{2}\int_{u_r}^{u_l} |\boldsymbol{F}'(u)\vec{n}|\, du \tag{1.23a}$$

$$= \boldsymbol{F}(u_l)\vec{n} + \int_{u_r}^{u_l} (\boldsymbol{F}'(u)\vec{n})^-\, du \tag{1.23b}$$

$$= \boldsymbol{F}(u_r)\vec{n} + \int_{u_r}^{u_l} (\boldsymbol{F}'(u)\vec{n})^+\, du. \tag{1.23c}$$

$\boldsymbol{H}^{\text{EO}}(\cdot, u_r, \vec{n})$ and $\boldsymbol{H}^{\text{EO}}(u_l, \cdot, \vec{n})$ are both in $C^1(\mathbb{R} \to \mathbb{R})$:

$$\frac{\partial}{\partial u_l}\boldsymbol{H}^{\text{EO}}(u_l, u_r, \vec{n}) = (\boldsymbol{F}'(u_l)\vec{n})^+$$
$$\frac{\partial}{\partial u_r}\boldsymbol{H}^{\text{EO}}(u_l, u_r, \vec{n}) = -(\boldsymbol{F}'(u_r)\vec{n})^-,$$

and they have the same smoothness as the flux $\boldsymbol{F}$ away from zeros of $\boldsymbol{F}'\vec{n}$. The flux function of Engquist and Osher involves at least as much upwinding as (1.20) – like on page 22 u_b and u_t are the minimum and maximum of u_l and u_r respectively:

$$\frac{1}{2}\left(\max_{u\in[u_b,u_t]}\boldsymbol{F}(u)\vec{n}-\min_{u\in[u_b,u_t]}\boldsymbol{F}(u)\vec{n}\right)\leqslant\left|\boldsymbol{H}^{\mathrm{EO}}(u_l,u_r,\vec{n})-\frac{\boldsymbol{F}(u_l)+\boldsymbol{F}(u_r)}{2}\vec{n}\right|,$$

since the term on the right hand side is obviously just half the total variation of $\boldsymbol{F}\vec{n}$ on $[u_b,u_t]$.

Going even further we might replace the upwinding term with a Lipschitz constant $L_{\boldsymbol{F}}$ of the flux function. This leads to the Lax-Friedrichs approximative Riemann solver $\boldsymbol{H}^{\mathrm{LF}}$. Depending on whether $L_{\boldsymbol{F}}$ is chosen to be a local (about u_l and u_r) or the global Lipschitz constant – provided the latter exists – the resulting approximate Riemann solver is called local or global Lax-Friedrichs flux function:

$$\boldsymbol{H}^{\mathrm{LF}}(u_l,u_r,\vec{n}):=\frac{\boldsymbol{F}(u_l)+\boldsymbol{F}(u_r)}{2}\vec{n}+L_{\boldsymbol{F}}\frac{u_l-u_r}{2}. \tag{1.24}$$

The global Lax-Friedrichs flux function has maximal smoothness: the same as that of the original flux function. If we base an estimate of the upwinding term on equations (1.23b) or (1.23c), we obtain expressions like $\boldsymbol{F}(u_l)\vec{n}+L_{\boldsymbol{F}}(u_l-u_r)$ and $\boldsymbol{F}(u_r)\vec{n}+L_{\boldsymbol{F}}(u_l-u_r)$. These are not equivalent to the Lax-Friedrichs flux function and not used.

Burgers Equation

The simplest non-linear scalar conservation law is the Burgers equation. Its flux is defined by

$$\boldsymbol{F}(u):=u^2\vec{\nu}^{\mathrm{t}} \tag{1.25}$$

with a fixed $\vec{\nu}\in\mathbb{R}^d$. The Burgers equation may develop discontinuities during the evolution of an initially smooth profile. The numerical flux function $\boldsymbol{H}^{\mathrm{EO}}$ of Engquist and Osher for the Burgers equation reads

$$\boldsymbol{H}^{\mathrm{EO}}(u_l,u_r,\vec{n})=u_l(u_l\vec{\nu}\cdot\vec{n})^+-u_r(u_r\vec{\nu}\cdot\vec{n})^- \tag{1.26}$$

and the (local) Lax-Friedrichs flux function

$$\boldsymbol{H}^{\mathrm{LF}}(u_l,u_r,\vec{n})=\frac{u_l^2+u_r^2}{2}\vec{\nu}\cdot\vec{n}+(u_l-u_r)\left\|\vec{\nu}\right\|\max\left\{\left|u_l\right|,\left|u_r\right|\right\}. \tag{1.27}$$

In the examples we always use a simple boundary treatment and (approximately) solve a Riemann problem with a prescribed outer state $u=0$ for the Burgers equation.

1.3 Hyperbolic Systems

The Covariant Formulation

For systems of conservation laws each component is in itself essentially a conserved scalar quantity, and the flux function is typically of a very specific form – it is hyperbolic. Our aim in this section is to introduce and motivate the notion of hyperbolicity. The underlying physical principle – symmetry – fits into the general framework of a covariant theory with space and time being treated on equal footing.[5] With a control volume $\Sigma \subset \mathbb{R}^{d+1}$ in space-time and a flux $j : \mathbb{R}^{d+1} \to \mathbb{R}^{s\times(d+1)}$ the covariant conservation principle in integral formulation reads

$$\int_{\partial\Sigma} j\vec{n}\, do = 0$$

and in differential form $\operatorname{Div} j = 0$. If the flux $j = \boldsymbol{F} \circ u$ depends on u alone, we have with the flux function $\boldsymbol{F} = (\boldsymbol{F}^0, \dots, \boldsymbol{F}^d) \in C^1(\mathbb{R}^s \to \mathbb{R}^{s\times(d+1)})$:

$$\sum_{k=0}^{d} \frac{\partial \boldsymbol{F}^k}{\partial u} \frac{\partial u}{\partial x^k} = 0. \tag{1.28}$$

All definitions and statements in this section apply to both covariant and non covariant formulations, in the latter case we take $\boldsymbol{F}^0 := \mathrm{id}_{\mathbb{R}^s}$ and choose vectors from space-time to be space-like, i.e. the time-like component x^0 is zero.

Symmetry and Hyperbolicity

The covariant formulation is homogeneous, this gives us considerable freedom in choosing a specific representation of the density function. A "good" choice may be characterized by demanding that all density components are treated on equal footing.

Definition 1.11. The flux function $\boldsymbol{F}$ in (1.28) is called **symmetric,** if there exists a set of independent variables $u^1, \dots, u^s$ such that the Jacobi matrices $(\partial \boldsymbol{F}^k_i / \partial u^j)_{ij}$ are symmetric.

Assume that $A(u) := \partial \boldsymbol{F}^0 / \partial u$ is always positive definite and define $v := \boldsymbol{F}^0 \circ u$ and $\boldsymbol{G}^k := \boldsymbol{F}^k \circ (\boldsymbol{F}^0)^{-1}$. If the components $\boldsymbol{F}^k$ take values in $\tilde{S}$, then

[5] We shall not go into the details of relativity – the Schwarzschild metric of space-time and such – but merely remove the formal bias between differentiation with respect to time and space immanent in (1.3).

the components of the flux function $\boldsymbol{G}$ are mappings $\boldsymbol{G}^k : \tilde{S} \to \tilde{S}$. By the chain rule we have

$$\frac{\partial \boldsymbol{G}^k}{\partial v} \frac{\partial \boldsymbol{F}^0}{\partial u} = \frac{\partial \boldsymbol{G}^k}{\partial v} A(u) = \frac{\partial \boldsymbol{F}^k}{\partial u}$$

and upon identifying x^0 with time (1.28) takes the following form:

$$\frac{\partial}{\partial t} v + \sum_{k=1}^{d} \frac{\partial \boldsymbol{G}^k}{\partial v} \frac{\partial v}{\partial x^k} = 0$$

which is just the non covariant quasi linear form of the conservation principle. Although we have not yet generalized the upwinding principle to systems, it should be noted that positive definiteness of $A(u)$ implies that such upwinding involves only "past" values of u. Now given a particular system of conservation laws we inquire whether it is – at least formally – derived from a covariant formulation with a symmetric flux function.

Definition 1.12. The system (1.28) of conservation laws is called **symmetrizable,** if there exists for all u a symmetric and positive definite matrix $A(u) \in \mathbb{R}^{s \times s}$ such that for any fixed vector $\vec{n}$ the matrix

$$\frac{\partial \boldsymbol{F}\vec{n}}{\partial u} A(u)$$

is symmetric. In other words, the Jacobi matrices of the flux components are simultaneously symmetrizable.

There is a strong formal relation between symmetrizability and the abstract entropy discussed on page 18: The Hesse matrix of a strictly convex function $\boldsymbol{E} : \mathbb{R}^s \to \mathbb{R}$ is clearly symmetric and positive definite. Conversely the Godunov-Mock theorem states that $\boldsymbol{E}$ is an entropy function, if and only if its Hesse matrix symmetrizes the system, see [GR96] for a proof.

Definition 1.13. A system of conservation laws is called **hyperbolic,** if for any fixed vector $\vec{n}$ from space-time the Jacobi matrix $\partial \boldsymbol{F}\vec{n}/\partial u$ is diagonalizable, i.e. there exist numbers

$$\lambda_1(u, \vec{n}), \lambda_2(u, \vec{n}), \ldots, \lambda_s(u, \vec{n}) \in \mathbb{R}$$

and a regular matrix $R(u, \vec{n}) \in \mathbb{R}^{s \times s}$ such that

$$\frac{\partial \boldsymbol{F}\vec{n}}{\partial u} = R(u, \vec{n})\, D(u, \vec{n})\, R^{-1}(u, \vec{n})$$

with $D(u, \vec{n}) := \operatorname{diag}(\lambda_1(u, \vec{n}), \ldots, \lambda_s(u, \vec{n}))$. The system is called **strictly hyperbolic,** if no two eigenvalues are equal. We shall denote by $r_k(u, \vec{n})$ the right eigenvector corresponding to $\lambda_k(u, \vec{n})$.

When we speak of a hyperbolic system we implicitly assume that the flux function $\boldsymbol{F}$ is C^1. Keeping $\vec{n}$ fixed we may thus assume that the eigenvalues $\lambda_k(\cdot, \vec{n})$ and the (suitably normalized) eigenvectors $r_k(\cdot, \vec{n})$ are continuous. We want to assume that they are also continuously differentiable and consequently that, as stated on page 9, the flux function is C^2. Sometimes we also rely on an ordering of the eigenvalues in either increasing or decreasing order. While pointwise such an ordering can always be enforced, it might destroy the smoothness of the eigenvalue functions, if some of them change positions relative to the ordering.

Hyperbolicity enables us to generalize many of the results obtained for the scalar case to systems, as we may diagonalize the Jacobi matrix of the flux function in the direction normal to a discontinuity and then operate on each state component separately. The following lemma establishes that hyperbolicity is not just a fortunate technical convenience. In fact, for all practical problems arising from physics an entropy function with a physical meaning can be found. Those systems are consequently symmetrizable.

Lemma 1.14. A symmetrizable system is hyperbolic.

Before proving lemma 1.14 we state as a separate lemma an important property of the symmetrizing matrix A:

Lemma 1.15. Let $X \in \mathbb{R}^{s\times s}$ be an arbitrary matrix and $A \in \mathbb{R}^{s\times s}$ be symmetric and positive definite. Obviously A has a Cholesky decomposition $A = CC^{\mathsf{t}}$ with a (triangular) matrix $C \in \mathbb{R}^{s\times s}$. Then

$$S := XA \text{ is symmetric} \quad \Longleftrightarrow \quad U := C^{-1}XC \text{ is symmetric.}$$

Proof.

$$\begin{array}{ccccc} S = S^{\mathsf{t}} & \Longleftrightarrow & XA = A^{\mathsf{t}}X^{\mathsf{t}} & \Longleftrightarrow & XA = AX^{\mathsf{t}} \\ & \Longleftrightarrow & XCC^{\mathsf{t}} = CC^{\mathsf{t}}X^{\mathsf{t}} & \Longleftrightarrow & C^{-1}XC = C^{\mathsf{t}}X^{\mathsf{t}}C^{-\mathsf{t}} \\ & \Longleftrightarrow & U = U^{\mathsf{t}} & & \end{array}$$

□

Proof of lemma 1.14. Let $J^k := \partial \boldsymbol{F}^k/\partial u \in \mathbb{R}^{s\times s}$ $(k \in \{0, \ldots, d\})$, $\vec{n} = (n^0, \ldots, n^d) \in \mathbb{R}^{d+1}$ be a fixed vector and

$$X := \sum_{k=0}^{d} n^k J^k.$$

We need to diagonalize X. By assumption there exists a positive definite symmetric matrix $A = CC^{\mathsf{t}} \in \mathbb{R}^{s\times s}$ such that all products $J^k A$ are symmetric and so is $S := XA$.

By lemma 1.15 $U := C^{-1}XC$ is symmetric, too. We can therefore find a diagonal matrix D and an orthogonal matrix $Q = Q^{-\mathsf{t}}$ such that $U = QDQ^{\mathsf{t}}$. Now

$$X = CUC^{-1} = (CQ)D(CQ)^{-1}.$$

□

The equivalence in lemma 1.15 represents two different flavors of variable transformation: We may either multiply the J^k's from the right with a symmetric positive definite matrix and thus transform just the independent state variables or multiply them from both right and left with a matrix and its inverse in which case we also transform the dependent variables.

For a hyperbolic system in the form of equation (1.3) we will tacitly assume that the flux components are mappings $\boldsymbol{F}^k : S \to S$ and prefer the simultaneous variable transformation in both domain and range.

The Riemann Problem and Riemann Invariants

The Riemann Problem is an initial value problem with two constant states separated by a hyperplane as initial data. We have so far encountered it several times and turn it now into a formal definition:

Definition 1.16. The Riemann problem for a hyperbolic system (1.3)

$$\frac{\partial}{\partial t}u + \sum_{k=1}^{d} \frac{\partial \boldsymbol{F}^k}{\partial u}\frac{\partial u}{\partial x^k} = 0$$

has initial data

$$u_0(\vec{x}) = u(t_0, \vec{x}) = \begin{cases} u_l & \text{if } \vec{x}\cdot\vec{n} \leqslant \vec{x}_0\cdot\vec{n} \\ u_r & \text{if } \vec{x}\cdot\vec{n} > \vec{x}_0\cdot\vec{n} \end{cases}$$

with a fixed unit vector $\vec{n} \in \mathbb{R}^d$.

It is possible to show that for small discontinuities a solution of the Riemann problem always exists, see [Lax73]. In the scalar case the jump could be of arbitrary size and an explicit formula for the desired weak solution was available. In the case of hyperbolic systems the idea is to define $s-1$

intermediate states $u_{m_1}, \ldots, u_{m_{s-1}} \in S$ and a path $\Gamma : \mathbb{R} \to S$ joining $u_{m_0} := u_l, u_{m_1}, \ldots, u_{m_{s-1}}, u_{m_s} := u_r$ such that two subsequent states $u_{m_{k-1}}$ and u_{m_k} joined by a subpath Γ_k are separated by either a rarefaction wave, a compression wave or a contact discontinuity. Since the right eigenvectors $r_1, \ldots, r_s$ are by the assumption of hyperbolicity linearly independent, the paths whose tangents are these eigenvectors should locally parameterize S, see figure 1.4. The intermediate states are furthermore chosen to be joined by paths whose tangents correspond to subsequently larger eigenvalues, because passing from left to right after a short amount of time has elapsed, one sees first the phenomena corresponding to the smaller eigenvalues. Functions that are constant along those paths (Riemann invariants) may be introduced to compute the intermediate states. For details of the proof we refer to [Lax73].

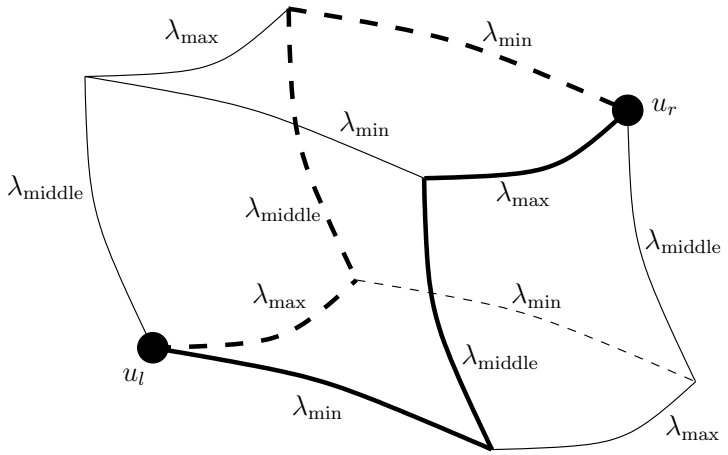

Figure 1.4: Paths in the state space for $s = 3$. The tangents to the subpaths are right eigenvectors of $\partial \boldsymbol{F}\vec{n}/\partial u$ corresponding to the eigenvalues $\lambda_{\min} \leqslant \lambda_{\text{middle}} \leqslant \lambda_{\max}$ respectively. The heavy solid line corresponds to an increasing (physical) ordering of the eigenvalues along the path, the heavy dashed line to a decreasing one.

The simple equation (1.8) on page 14 does not generalize to systems, as $\boldsymbol{F}\vec{n}$ will in general depend on all state components simultaneously, hence the isolines of any state component u^j will not be straight lines. This is a principal difference between the scalar and the vector-valued regime. In just one space dimension or along a prescribed direction $\vec{n}$ one might look for a definition of a k-characteristic in terms of an ordinary differential equation:

$$\frac{d}{dt}\chi_k(t) = \lambda_k(u(t, \chi_k(t)\vec{n}), \vec{n})$$

where $\chi_k : \mathbb{R} \to \mathbb{R}$ and $\lambda_k(u, \vec{n})$ is the k-th eigenvalue of $\partial \boldsymbol{F}\vec{n}/\partial u$. If we ask for the second derivative $d^2\chi_k(t)/dt^2$ to be approximately zero, we are led to consider paths in the state space along which $\lambda_k(u, \vec{n})$ changes slowly. The question whether $\lambda_k(u, \vec{n})$ changes slowly along a path whose tangent is the corresponding k-th right eigenvector $r_k(u, \vec{n})$ leads to a useful classification of the eigenvalues and eigenvectors into linearly degenerate, genuinely nonlinear and other ones.

For a single linear scalar equation the derivative of $\boldsymbol{F}\vec{n}$ with respect to u is a constant in $\mathbb{R}$ and its second derivative is zero. If the latter fails to be identically zero, the equation is not linear. For systems (along the direction $\vec{n}$) the gradient with respect to u of an eigenvalue λ of the Jacobi matrix of $\boldsymbol{F}\vec{n}$ may be orthogonal to the corresponding eigenvector without being zero.

Definition 1.17. Let $\vec{n}$ be a fixed unit vector and $r_k(u) := r_k(u, \vec{n})$ ($k \in \{1, \ldots, s\}$) a right eigenvector of the Jacobi matrix $\partial \boldsymbol{F}\vec{n}/\partial u$ of $\boldsymbol{F}\vec{n}$. A function $\Psi \in C^1(S \to \mathbb{R})$ is called a ***k*-Riemann invariant,** if on all S the function

$$\langle \Psi; r_k \rangle_S := r_k \cdot \nabla_u \Psi = 0. \tag{1.29}$$

The index u to the nabla operator indicates that differentiation is performed with respect to the state variables, not with respect to time or space variables. The scalar product in the tangent space for $\nabla_u \Psi$ and r_k is independent of the particular choice of variables. Let us briefly contemplate why this is so. Introducing the (bijective) coordinate mappings $K : \mathbb{R}^s \supset G \to S$ and $\tilde{K} : \mathbb{R}^s \supset \tilde{G} \to S$ we have the Jacobi matrices

$$\hat{A} := \mathbf{J}(K^{-1} \circ \boldsymbol{F}\vec{n} \circ K) : G \to \mathbb{R}^{s \times s}$$
$$\tilde{A} := \mathbf{J}(\tilde{K}^{-1} \circ \boldsymbol{F}\vec{n} \circ \tilde{K}) : \tilde{G} \to \mathbb{R}^{s \times s}.$$

These satisfy

$$\tilde{A} = [\mathbf{J}(K^{-1} \circ \tilde{K})]^{-1} \hat{A} \, [\mathbf{J}(K^{-1} \circ \tilde{K})],$$

they have thus the same eigenvalues λ and corresponding eigenvectors $\hat{r}$ and $\tilde{r}$ respectively:

$$\lambda \tilde{r} = \tilde{A}\tilde{r} = [\mathbf{J}(K^{-1} \circ \tilde{K})]^{-1} \lambda \hat{r}.$$

Therefore

$$\tilde{r} = [\mathbf{J}(K^{-1} \circ \tilde{K})]^{-1} \hat{r}.$$

With $\hat{\Psi} := \Psi \circ K$ and $\tilde{\Psi} := \Psi \circ \tilde{K}$ we now conclude

$$\hat{r} \cdot \nabla \hat{\Psi} = \tilde{r} \cdot \nabla \tilde{\Psi} = \langle \Psi; r \rangle_S$$

independent of the coordinate mapping.

For any fixed $k \in \{1, \ldots, s\}$ we may expect to find $s-1$ k-Riemann invariants, as there are $s-1$ directions orthogonal to that of r_k available. Altogether there are $s(s-1)$ Riemann invariants which may be used to compute the $(s-1)s$ unknown components of the $s-1$ intermediate states mentioned above.

Next we construct a path whose tangents are k-th right eigenvectors and verify that indeed the k-Riemann invariants are constant along it: define on an open interval $I \subset \mathbb{R}$ about zero $\Gamma_k \in C^1(I \to S)$ by the ordinary differential equation

$$\frac{d}{d\xi}\Gamma_k = r_k \circ \Gamma_k \tag{1.30}$$

$$\Gamma_k(0) = \hat{u}$$

for an arbitrary $\hat{u} \in S$. We have still considerable freedom in scaling the eigenvector r_k on the right hand side. Such scaling corresponds to transforming the parameter ξ. For now we normalize r_k to have Euclidean norm one – thinking of the tangent space as the ordinary $\mathbb{R}^s$ – by replacing it with $r_k / \|r_k\|_{\mathbb{R}^s}$. By the Picard-Lindelöf theorem there exists for this normalization a unique solution to (1.30) on an open interval $I \subset \mathbb{R}$ which contains $\xi = 0$. For a k-Riemann invariant Ψ one has:

$$\frac{d}{d\xi}(\Psi \circ \Gamma_k) = \langle \Psi; r_k \rangle_S \circ \Gamma_k = 0$$

and therefore

$$\Psi(\Gamma_k(\xi)) = \Psi(\Gamma_k(0)) = \Psi(\hat{u}).$$

This proves

Lemma 1.18. A k-Riemann invariant is constant along the path defined by (1.30).

Definition 1.19. Let $\vec{n}$ be a fixed unit vector and $\lambda_k(u) := \lambda_k(u, \vec{n})$ ($k \in \{1, \ldots, s\}$) be an eigenvalue of $\partial \boldsymbol{F} \vec{n} / \partial u$ with corresponding right eigenvector $r_k(u) := r_k(u, \vec{n})$. If λ_k is a k-Riemann invariant (here we ultimately require $\boldsymbol{F} \in C^2$), then r_k and λ_k are called **linearly degenerate.** If always

$$\langle \lambda_k; r_k \rangle_S \neq 0,$$

then r_k and λ_k are called **genuinely nonlinear.**

The structure of solutions to the Riemann problem is quite simple, if the eigenvalues (and eigenvectors) fall into either of these categories: a genuinely nonlinear eigenvector corresponds to a rarefaction fan or a compression wave, a linearly degenerate eigenvector to a contact discontinuity.

Linearly degenerate case

If r_k is linearly degenerate, then λ_k is by lemma 1.18 constant along Γ_k. Furthermore,

$$(\boldsymbol{F} \circ \Gamma_k)\vec{n} - \lambda_k(\hat{u})\Gamma_k = \text{const} \tag{1.31}$$

is a constant, too:

$$\begin{aligned}\frac{d}{d\xi}\left[(\boldsymbol{F} \circ \Gamma_k)\vec{n} - \lambda_k(\hat{u})\Gamma_k\right] &= \left(\left.\frac{\partial \boldsymbol{F}\vec{n}}{\partial u}\right|_{\Gamma_k} - \lambda_k(\hat{u})\right)\frac{d}{d\xi}\Gamma_k \\ &= \left(\left.\frac{\partial \boldsymbol{F}\vec{n}}{\partial u}\right|_{\Gamma_k} - \lambda_k(\hat{u})\right) r_k \circ \Gamma_k \\ &= 0.\end{aligned}$$

Therefore for $\hat{u} := u_l$ and an arbitrary state $u_r \in \Gamma_k(I)$

$$u(t,\vec{x}) := \begin{cases} u_l & \text{if } \dfrac{\vec{x}-\vec{x}_0}{t-t_0}\cdot\vec{n} < \lambda_k(u_l,\vec{n}) \\ \\ u_r & \text{if } \dfrac{\vec{x}-\vec{x}_0}{t-t_0}\cdot\vec{n} \geqslant \lambda_k(u_l,\vec{n}) \end{cases} \tag{1.32}$$

satisfies the conservation principle, if and only if it satisfies the Rankine-Hugoniot jump condition (1.15) in all state components simultaneously

$$[\boldsymbol{F}(u_l) - \boldsymbol{F}(u_r)]\vec{n} = (u_l - u_r)\vec{\nu}\cdot\vec{n}.$$

This is by equation (1.31) clearly the case for $\vec{\nu}\cdot\vec{n} = \lambda_k(u_l,\vec{n}) = \lambda_k(u_r,\vec{n})$.

Genuinely nonlinear case

If r_k is genuinely nonlinear, then we normalize it in a different way by replacing r_k with $r_k/\langle\lambda_k; r_k\rangle_S$:

$$\langle\lambda_k; r_k\rangle_S = 1.$$

The Picard-Lindelöf theorem again ensures existence and uniqueness of a solution to equation (1.30) on an open interval $I \subset \mathbb{R}$ containing $\xi = 0$.

$$\frac{d}{d\xi}(\lambda_k \circ \Gamma_k) = \langle \lambda_k; r_k \rangle_S \circ \Gamma_k = 1$$

and consequently

$$\lambda_k(\Gamma_k(\xi)) = \lambda_k(\Gamma_k(0)) + \xi = \lambda_k(\hat{u}) + \xi \tag{1.33}$$

We let $\hat{u} := u_l$ in equation (1.30),

$$\zeta := \frac{\vec{x} - \vec{x}_0}{t - t_0} \cdot \vec{n} - \lambda_k(u_l, \vec{n}), \tag{1.34a}$$

and define for an arbitrary state $u_r \in \Gamma_k(I \cap \mathbb{R}_{>0})$

$$u(t, \vec{x}) := \begin{cases} u_l & \text{if } \zeta < 0 \\ \Gamma_k(\zeta) & \text{if } 0 \leqslant \zeta \leqslant \lambda_k(u_r, \vec{n}) - \lambda_k(u_l, \vec{n}) \\ u_r & \text{if } \zeta > \lambda_k(u_r, \vec{n}) - \lambda_k(u_l, \vec{n}) \end{cases}. \tag{1.34b}$$

Claim: This u satisfies equation (1.3).

Proof. We have for the middle line of equation (1.34b)

$$\left.\frac{\partial}{\partial t}\right|_{t,\vec{x}} u = -\frac{\vec{x} - \vec{x}_0}{(t - t_0)^2} \cdot \vec{n} \left.\frac{d}{d\xi}\right|_{\zeta} \Gamma_k.$$

On the other hand

$$\nabla|_{t,\vec{x}} u = \frac{\vec{n}}{t - t_0} \left.\frac{d}{d\xi}\right|_{\zeta} \Gamma_k$$

implies

$$\begin{aligned} \left.\frac{\partial \boldsymbol{F}\vec{n}}{\partial u}\right|_{\Gamma(\zeta)} \nabla|_{t,\vec{x}} u &= \left.\frac{\partial \boldsymbol{F}\vec{n}}{\partial u}\right|_{\Gamma(\zeta)} \frac{\vec{n}}{t - t_0} \left.\frac{d}{d\xi}\right|_{\zeta} \Gamma_k \\ &= \frac{\vec{n}}{t - t_0} \left.\frac{\partial \boldsymbol{F}\vec{n}}{\partial u}\right|_{\Gamma(\zeta)} r_k(\Gamma_k(\zeta)) \end{aligned}$$

and since r_k is an eigenvector

$$\begin{aligned} &= \frac{\vec{n}}{t - t_0} \lambda_k(\Gamma_k(\zeta))\, r_k(\Gamma_k(\zeta)) \\ &= \frac{\vec{n}}{t - t_0} \lambda_k(\Gamma_k(\zeta)) \left.\frac{d}{d\xi}\right|_{\zeta} \Gamma_k. \end{aligned}$$

One has by equations (1.33) and (1.34a)

$$= \frac{\vec{n}}{t-t_0} \left(\lambda_k(u_l) + \frac{\vec{x}-\vec{x}_0}{t-t_0} \cdot \vec{n} - \lambda_k(u_l) \right) \left. \frac{d}{d\xi} \right|_\zeta \Gamma_k.$$

Hence

$$\operatorname{div}|_{t,\vec{x}}(\boldsymbol{F} \circ u) = \frac{\vec{x}-\vec{x}_0}{(t-t_0)^2} \cdot \vec{n} \left. \frac{d}{d\xi} \right|_\zeta \Gamma_k = - \left. \frac{\partial}{\partial t} \right|_{t,\vec{x}} u.$$

The top and bottom line of equation (1.34b) represent constant states which trivially satisfy equation (1.3) and fit together continuously with the definition in the middle line. □

For $u_r \in \Gamma_k(I \cap \mathbb{R}_{<0})$ a construction similar to that above will not fit together continuously, but can be interpreted as a multi-valued solution. We need to insert a discontinuity moving at velocity $\vec{\nu}$ instead such that the conservation principle is satisfied, but in contrast to the scalar regime we have to observe the other eigenvalues as well. We now assume that the eigenvalue functions are sorted in either increasing or decreasing order.

Definition 1.20. Define $\sigma_\lambda \in \{-1, 1\}$ by

$$\sigma_\lambda := \begin{cases} +1 & \text{if } \lambda_k \leqslant \lambda_{k+1} \\ -1 & \text{if } \lambda_k \geqslant \lambda_{k+1} \end{cases}$$

for all $k \in \{1, \dots, s-1\}$.

This is equivalent to

$$\lambda_k \leqslant \lambda_{k+\sigma_\lambda} \text{ for all } k \in \{2 - (1+\sigma_\lambda)/2, \dots, s - (1+\sigma_\lambda)/2\}.$$

Lax [Lax73] demands not only that the k-characteristics all go into the discontinuity

$$\cdots \geqslant \lambda_{k+\sigma_\lambda}(u_l) \geqslant \lambda_k(u_l) > \vec{\nu} \cdot \vec{n} > \lambda_k(u_r) \geqslant \lambda_{k-\sigma_\lambda}(u_r) \geqslant \cdots, \tag{1.35a}$$

but also that these are the only characteristics entering the discontinuity, i.e. the $(k-\sigma_\lambda)$-characteristics do not go into the discontinuity from the left and the $(k+\sigma_\lambda)$-characteristics not from the right

$$\cdots \leqslant \lambda_{k-\sigma_\lambda}(u_l) < \vec{\nu} \cdot \vec{n} < \lambda_{k+\sigma_\lambda}(u_r) \leqslant \cdots. \tag{1.35b}$$

These requirements now imply that there is a total of $s+1$ characteristic fields entering the discontinuity.

Definition 1.21. Under the assumption that the eigenvalues $\lambda_1, \ldots, \lambda_s$ can be sorted without destroying their smoothness as functions and that λ_k is genuinely nonlinear we define a ***k*-shock** as a discontinuity satisfying the Rankine-Hugoniot jump condition (1.15) in all state components simultaneously

$$[\boldsymbol{F}(u_l) - \boldsymbol{F}(u_r)]\vec{n} = (u_l - u_r)\vec{\nu} \cdot \vec{n}.$$

and for which equations (1.35) hold.

After eliminating $\vec{\nu} \cdot \vec{n}$ from the Rankine-Hugoniot conditions we obtain $s-1$ compatibility relations. Together with the information from the $s+1$ characteristic fields going into the discontinuity we have $2s$ genuinely nonlinear equations in the $2s$ components of u_l and u_r. Equation (1.35b) thus prevents the generation of an (a priori) overdetermined system.

If we fix just u_l in the Rankine-Hugoniot jump condition, then we obtain s equations in $s+1$ unknowns (u_r and $\vec{\nu} \cdot \vec{n}$). We may hence expect a one parameter family of possible states u_r to which u_l can be connected via a k-shock. A detailed proof of existence of such a family may be found in [GR96].

Numerical Flux Functions

Any function having properties similar to those stated in lemma 1.4 on page 13 for scalar equations is acceptable as an approximate Riemann solver for systems. We restate said properties for the sake of clarity:

Definition 1.22. A numerical flux function is a function $\boldsymbol{H} : S \times S \times \mathbb{R}^d \to \mathbb{R}^s$ which is consistent with the flux function $\boldsymbol{F}$:

$$\left\| \boldsymbol{H}(u_l, u_r, \vec{n}) - \frac{\boldsymbol{F}(u_l) + \boldsymbol{F}(u_r)}{2}\vec{n} \right\|_S \leq \alpha \frac{\|u_l - u_r\|_S}{2}$$

with $\|\vec{n}\| = 1$ and a fixed constant[6] $\alpha \in \mathbb{R}$ which also obeys

$$\boldsymbol{H}(u_l, u_r, \vec{n}) = -\boldsymbol{H}(u_r, u_l, -\vec{n}).$$

Let us first consider the case of a linear hyperbolic system with s equations in just one space dimension:

$$\frac{\partial}{\partial t}u + A\frac{\partial}{\partial x}u = 0 \tag{1.36}$$

[6]We may have to restrict S such that a global Lipschitz constant $L_{\boldsymbol{F}}$ for the flux function exists, cf. the discussion in remark 1.7 on page 15.

with initial condition $u(t_0, x) = u_0(x)$. By the assumption of hyperbolicity the matrix $A \in \mathbb{R}^{s\times s}$ is diagonalizable:

$$A = RDR^{-1} \tag{1.37}$$

with a diagonal matrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_s) \in \mathbb{R}^{s\times s}$ and a regular matrix $R \in \mathbb{R}^{s\times s}$. Multiplying equation (1.36) from the left with R^{-1} and introducing new variables by

$$v := R^{-1}u \qquad \text{and} \qquad v_0 := R^{-1}u_0$$

we obtain the following decoupled system of s scalar equations:

$$\frac{\partial}{\partial t}v + D\frac{\partial}{\partial x}v = 0.$$

The solution of this decoupled system may be stated as

$$v(t,x) = \begin{pmatrix} v_0^1(x - \lambda_1(t-t_0)) \\ \vdots \\ v_0^s(x - \lambda_s(t-t_0)) \end{pmatrix}$$

and denoting the k-th row of R^{-1} by l_k^{t} (this is just the k-th left eigenvector of A) the solution of equation (1.36) as

$$u(t,x) = R \begin{pmatrix} l_1^{\mathsf{t}} u_0(x - \lambda_1(t-t_0)) \\ \vdots \\ l_s^{\mathsf{t}} u_0(x - \lambda_s(t-t_0)) \end{pmatrix}.$$

A Riemann solver for the decoupled system is easily defined by upwinding each state component separately. Transforming back to the form of equation (1.36) we obtain the following definition of positive and negative part and absolute value for diagonalizable matrices.

Definition 1.23. For a diagonal matrix $D := \operatorname{diag}(\lambda_1, \ldots, \lambda_s) \in \mathbb{R}^{s\times s}$ define for '+' and '−' in '±' separately

$$D^{\pm} := \operatorname{diag}(\lambda_1^{\pm}, \ldots, \lambda_s^{\pm}) \qquad \text{and} \qquad |D| := \operatorname{diag}(|\lambda_1|, \ldots, |\lambda_s|).$$

The functions $^{\pm} : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ are as on page 26 defined by $x^{\pm} := (|x| \pm x)/2$. For a diagonalizable matrix $X := RDR^{-1} \in \mathbb{R}^{s\times s}$ (D is of course diagonal) define

$$X^{\pm} := RD^{\pm}R^{-1} \qquad \text{and} \qquad |X| := R\,|D|\,R^{-1}.$$

Obviously $X^{\pm} = (|X| \pm X)/2$ holds. Positive or negative part and absolute value for matrices are well-defined, since diagonalization is unique up to rearranging eigenvalues and eigenvectors. The definition, however, does not hinge on any particular arrangement of these. The flux function of lemma 1.4 on page 13 for linear scalar equations may now be generalized to linear systems ($\boldsymbol{F}(u)\vec{n} = A\vec{n}u$ for any vector $\vec{n} \in \mathbb{R}^d$):

$$\boldsymbol{F}(u_l, u_r, \vec{n}) := \frac{\boldsymbol{F}(u_l) + \boldsymbol{F}(u_r)}{2}\vec{n} + |A\vec{n}|\,\frac{u_l - u_r}{2}.$$

The flux function $\boldsymbol{H}^{\mathrm{OS}}$ of Osher and Solomon [OS82] generalizes the Engquist and Osher flux function for scalar equations to hyperbolic systems by choosing a particular path $\Gamma : \mathbb{R} \to S$ in the state space for which the integrals in equations (1.23) can be easily evaluated. Its principal idea is closely related to the solution of the Riemann problem sketched on page 32.

By convention the path Γ is oriented from u_l towards u_r, we therefore have to swap the integration bounds and the signs in front of the integrals in equations (1.23). Basing our presentation of the flux function of Osher and Solomon on equation (1.23c)

$$\boldsymbol{H}^{\mathrm{EO}}(u_l, u_r, \vec{n}) = \boldsymbol{F}(u_r)\vec{n} - \int_{u_l}^{u_r} (\boldsymbol{F}'(u)\vec{n})^{+}\, du,$$

we suppose that we have already determined appropriate intermediate states and that along the path between any two subsequent states the eigenvectors forming the tangents of the path are either linearly degenerate or genuinely nonlinear. Osher and Solomon now avoid constructing the compression wave, if the eigenvalue at the start of the path is greater than that at the end, but reverse the parameterization of that path component to make the genuinely nonlinear eigenvalue a strictly decreasing function along the path in this case. This simplification avoids the most expensive part of a complete solution of the Riemann problem: the numerical approximation of the one parameter family of states that can be joined via a compression wave.

$$\boldsymbol{H}^{\mathrm{OS}}(u_l, u_r, \vec{n}) := \boldsymbol{F}(u_r)\vec{n} - \sum_{k=1}^{s} \int_{\Gamma_k} \left(\left.\frac{\partial \boldsymbol{F}\vec{n}}{\partial u}\right|_{\tilde{u}} \right)^{+} d\tilde{u}. \tag{1.38}$$

If λ_k is genuinely nonlinear, then it is a strictly monotone function on the path $\Gamma_k : I \to S$ defined by equation (1.30) and possibly reparameterized as explained above. There are four cases: λ_k may change sign from positive to negative or from negative to positive or it may be either non-positive or non-negative throughout. The last two cases are also the only alternatives for a

linearly degenerate eigenvalue which is by lemma 1.18 constant along Γ_k. If λ_k does change sign on Γ_k, then there is exactly one sonic point $n_k \in \Gamma_k(I)$ such that $\lambda_k(n_k) = 0$. All in all, the value of the integral along each path component hinges on the four possible combinations of the sign of λ_k at start and end of the path component (table 1.1).

We denote by $[a_k, b_k] \subset I$ $(a_k < b_k)$ a suitable interval in I and by $s_k = \Gamma_k(a_k)$ ("start") and $e_k = \Gamma_k(b_k)$ ("end") two states joined by the path Γ_k. Obviously $e_k = s_{k+1} = u_{m_k}$ is the k-th intermediate state, $s_1 = u_l$ and $e_s = u_r$. Let us first consider the case that λ_k does not change sign on Γ_k:

$$\begin{aligned}
\int_{\Gamma_k} \left(\left. \frac{\partial \boldsymbol{F}\vec{n}}{\partial u} \right|_{\tilde{u}} \right)^+ d\tilde{u} &= \int_{a_k}^{b_k} \left(\left. \frac{\partial \boldsymbol{F}\vec{n}}{\partial u} \right|_{\Gamma_k(\xi)} \right)^+ r_k(\Gamma_k(\xi))\, d\xi \\
&= \int_{a_k}^{b_k} \lambda_k^+(\Gamma_k(\xi))\, r_k(\Gamma_k(\xi))\, d\xi \\
&= \begin{cases} \displaystyle\int_{a_k}^{b_k} \lambda_k(\Gamma_k(\xi))\, r_k(\Gamma_k(\xi))\, d\xi & \text{if } \lambda_k \geqslant 0 \text{ on } \Gamma_k \\ 0 & \text{if } \lambda_k \leqslant 0 \text{ on } \Gamma_k. \end{cases}
\end{aligned}$$

The case $\lambda_k \geqslant 0$ on Γ_k can be further simplified:

$$\begin{aligned}
\int_{\Gamma_k} \left(\left. \frac{\partial \boldsymbol{F}\vec{n}}{\partial u} \right|_{\tilde{u}} \right)^+ d\tilde{u} &= \int_{a_k}^{b_k} \left. \frac{\partial \boldsymbol{F}\vec{n}}{\partial u} \right|_{\Gamma_k(\xi)} r_k(\Gamma_k(\xi))\, d\xi \\
&= \int_{\Gamma_k} \left. \frac{\partial \boldsymbol{F}\vec{n}}{\partial u} \right|_{\tilde{u}} d\tilde{u} \\
&= \boldsymbol{F}(e_k)\vec{n} - \boldsymbol{F}(s_k)\vec{n}.
\end{aligned}$$

Now we turn to the case that λ_k changes sign at $n_k = \Gamma_k(c_k)$ with $c_k \in (a_k, b_k)$:

$$\begin{aligned}
\int_{\Gamma_k} \left(\left. \frac{\partial \boldsymbol{F}\vec{n}}{\partial u} \right|_{\tilde{u}} \right)^+ d\tilde{u} &= \int_{a_k}^{b_k} \left(\left. \frac{\partial \boldsymbol{F}\vec{n}}{\partial u} \right|_{\Gamma_k(\xi)} \right)^+ r_k(\Gamma_k(\xi))\, d\xi \\
&= \int_{a_k}^{b_k} \lambda_k^+(\Gamma_k(\xi))\, r_k(\Gamma_k(\xi))\, d\xi \\
&= \begin{cases} \displaystyle\int_{c_k}^{b_k} \lambda_k(\Gamma_k(\xi))\, r_k(\Gamma_k(\xi))\, d\xi & \text{if } \lambda_k(\Gamma_k(b_k)) > 0 \\ \displaystyle\int_{a_k}^{c_k} \lambda_k(\Gamma_k(\xi))\, r_k(\Gamma_k(\xi))\, d\xi & \text{if } \lambda_k(\Gamma_k(a_k)) > 0 \end{cases}
\end{aligned}$$

	$\lambda_k(e_k) \geqslant 0$	$\lambda_k(e_k) < 0$
$\lambda_k(s_k) \geqslant 0$	$\boldsymbol{F}(e_k)\vec{n} - \boldsymbol{F}(s_k)\vec{n}$	$\boldsymbol{F}(n_k)\vec{n} - \boldsymbol{F}(s_k)\vec{n}$
$\lambda_k(s_k) < 0$	$\boldsymbol{F}(e_k)\vec{n} - \boldsymbol{F}(n_k)\vec{n}$	0

Table 1.1: The integral $\displaystyle\int_{\Gamma_k} \left(\frac{\partial \boldsymbol{F}\vec{n}}{\partial u}\bigg|_{\tilde{u}} \right)^+ d\tilde{u}$.

and each of these integrals can be evaluated similar to the case $\lambda_k \geqslant 0$ above:

$$\int_{\Gamma_k} \left(\frac{\partial \boldsymbol{F}\vec{n}}{\partial u}\bigg|_{\tilde{u}} \right)^+ d\tilde{u} = \begin{cases} \boldsymbol{F}(e_k)\vec{n} - \boldsymbol{F}(n_k)\vec{n} & \text{if } \lambda_k(e_k) > 0 \\ \\ \boldsymbol{F}(n_k)\vec{n} - \boldsymbol{F}(s_k)\vec{n} & \text{if } \lambda_k(s_k) > 0. \end{cases}$$

The expressions for the integral are summarized in table 1.1. The original version of the flux function of Osher and Solomon chooses the reverse ordering ($\sigma_\lambda = -1$) of the paths joining the intermediate states compared to the "physical" ($\sigma_\lambda = 1$) solution strategy of the Riemann problem: the tangents along the path components joining two intermediate states correspond to subsequently smaller eigenvalues. While Osher and Solomon claim that their ordering improves the numerical behaviour of the flux function, Spekreijse [Spe87] bases his multigrid solver of the Euler equations on the increasing ordering.

Chapter 2

Discretization

2.1 Data Functionals

In order to treat problems from infinite dimensional spaces (like differential equations on Banach spaces) with the aid of a computer, they have to be discretized, i.e. transformed into an algebraic relation in finite dimensional vector spaces like $\mathbb{R}^n$. In fact, even the real numbers themselves are represented by a finite set of machine numbers. The truncation error introduced by replacing a real number with its machine representation has to be observed when designing numerical algorithms such as solvers for linear equations, iteration schemes for nonlinear equations, etc. Let us review in this section the main strategy for discretizing a partial differential equation and its associated boundary conditions. In the absence of a closed formula for the solution one has almost no alternative to choosing a suitable **finite set of data functionals** and to approximate the solution of the original problem based on the finitely many values of the data functionals. There is a very strong interplay between the choice of the data functionals and the data structure representing the geometry of the problem under consideration.

The question whether the partial differential equation has a (unique) solution at all is outside the discretization procedure, but it has to be taken into account when discussing convergence of the numerical method:

1. Do the approximations obtained from the finitely many values of the data functionals converge to a limit function as more and more data functionals from a prescribed sequence are considered and

2. is this limit function a meaningful solution of the original problem modeled by the partial differential equation?

For general hyperbolic systems these are, unfortunately, open questions. Regularizing the system with a Laplace term on the right hand side one may apply the Kružkov existence and uniqueness theorem (generalized to systems) [Kru70]. For certain classes of numerical schemes and particular systems the theory of measure valued solutions and the existence and uniqueness theorem of DiPerna [DiP85] may be used to obtain convergence results within an integral norm for the numerical solutions, too.

We restrict ourselves to the following notion of a computational domain:

Definition 2.1. The **computational domain** is a compact subset $\Omega \subset \mathbb{R}^d$ with $\Omega = \overline{\operatorname{int} \Omega}$ and a piecewise smooth boundary $\partial\Omega$.

For hyperbolic conservation laws (1.1) a natural approach consists in subdividing the computational domain Ω into finitely many cells such that the cells completely[1] cover the computational domain and their interiors are mutually disjoint. Starting from the values of the data functionals the density distribution within each control volume is locally approximated by a polynomial. The integral of the flux across the boundary of each control volume is approximated by a quadrature rule with positive weights from the values of a numerical flux function evaluated at the quadrature points. For control volume interfaces on the boundary of the computational domain some prescribed boundary conditions are required. The approximations of the integrals are then used to update the values of the data functionals and the procedure can be iterated. In a very straightforward way this can be turned into an **explicit time stepping** scheme which is the path we shall follow. Using an implicit method typically leads to a sparse large system of (linearized) equations which has to be solved iteratively for each time step, see [Mei96] for a compact survey of the iterative solvers applied to the Navier-Stokes equations.

The control volumes are the core of finite volume methods which we will very briefly describe below, but they are not well suited for developing a collocation method. While we like to think of the integral formulation (1.1) of the conservation principle as the more "genuine" expression of the laws of nature, the differential form (1.2) provides a more convenient starting point for mathematical manipulation. In either case we will discretize the equation in space and time separately and consider only data functionals δ which commute with the time derivative, i.e. they operate only on the space

[1] We allow exceptions near the boundary.

variables of a function $u = u(t, \vec{x})$:[2]

$$\delta \frac{\partial}{\partial t} u = \frac{d}{dt} \delta u. \tag{2.1}$$

In the case of systems we apply the data functionals to each component separately. The data functionals we will consider are either **cell averaging functionals**

$$\delta_\Sigma u := \frac{1}{|\Sigma|} \int_\Sigma u \, dV$$

where necessarily $|\Sigma| > 0$, **collocation functionals**

$$\delta_{\{\vec{x}\}} u := u(\cdot, \vec{x})$$

or convex combinations of collocation functionals

$$\delta_{\{\vec{x}_1, \ldots, \vec{x}_n\}} u := \sum_{k=1}^{n} w_k \, u(\cdot, \vec{x}_k).$$

We choose not to include the weights w_k in the notation of the data functional, since we will mostly use the simple average $w_k := 1/n$.

For the support Σ of the functional δ we introduce the term **data location.** This is either a cell $\Sigma \subset \mathbb{R}^d$ which has $|\Sigma| > 0$ or a finite subset of $\mathbb{R}^d$ for convex combinations of collocation functionals. The barycentre of a data location Σ is obtained by applying the data functional to the components of the identity function $\mathrm{id}_{\mathbb{R}^d}$ on $\mathbb{R}^d$. It is suggestive to denote the identity on $\mathbb{R}^d$ by the vector $\vec{x}$ itself:

$$\text{barycentre}\, \Sigma = \delta_\Sigma \vec{x}.$$

These types of data functionals have norm $\|\delta\| = 1$, if the underlying function space is equipped with the supremum[3] norm: Obviously

$$|\delta u| \leqslant \|u\|_\infty$$

and equality holds, if u is constant. We denote by $BL^\infty(\Omega \to \mathbb{R})$ the space of bounded measurable functions with the supremum norm from Ω to $\mathbb{R}$. Formally one may obtain a collocation functional as the limit of cell averaging functionals. If $(\Sigma_k)_{k \in \mathbb{N}}$ is a sequence of cells in $\mathbb{R}^d$ with

$$\Sigma_{k+1} \subset \Sigma_k \qquad \text{and} \qquad \lim_{k \to \infty} \operatorname{diam} \Sigma_k = 0,$$

[2] Strictly speaking, δ is a mapping which assigns to a time t a linear functional $\delta(t)$.

[3] In the case of cell averaging the essential supremum would be sufficient.

then by the Cantor intersection theorem[4]

$$\bigcap_{k\in\mathbb{N}} \Sigma_k = \{\vec{x}\}$$

for some $\vec{x} \in \mathbb{R}^d$. If the function f is continuous at $\vec{x}$, then

$$\lim_{k\to\infty} \delta_{\Sigma_k} f = \delta_{\{\vec{x}\}} f.$$

Conversely the cell averaging functionals may be obtained as convolutions of characteristic functions with collocation functionals:

$$\delta_\Sigma f = \frac{1}{|\Sigma|}\int_\Sigma f\, dV = \frac{1}{|\Sigma|}\chi_\Sigma * \delta_{\vec{x}} f.$$

Applying these data functionals to the differential form of the conservation law gives:

$$\frac{d}{dt}\delta u = -\delta \operatorname{div}(\boldsymbol{F}\circ u). \tag{2.2a}$$

Assembling the finitely many data functionals δ into a vector $\boldsymbol{\Lambda}$ (we will subsequently write $\delta \in \boldsymbol{\Lambda}$ to refer to a particular component of that vector) we obtain

$$\frac{d}{dt}\boldsymbol{\Lambda} u = -\boldsymbol{\Lambda} \operatorname{div}(\boldsymbol{F}\circ u). \tag{2.2b}$$

The general structure of the flow solver can now be sketched in the following way:

1. Based on the values of the data vector $\boldsymbol{\Lambda}u$ compute an approximation to the exact solution via a suitable **reconstruction** procedure. We denote the reconstruction process by $\mathcal{R}$.

2. Supply a numerical approximation $\widetilde{\nabla\cdot}$ to the divergence operator.

3. Approximate the flow field by a (numerical flux) function $\boldsymbol{H}$ based on the reconstruction $\mathcal{R}\,\boldsymbol{\Lambda}u$. This includes using the prescribed boundary flux where appropriate.

4. Approximate the right hand side of (2.2b) by

$$-\boldsymbol{\Lambda}\,\widetilde{\nabla\cdot}(\boldsymbol{H}\circ\mathcal{R}\,\boldsymbol{\Lambda}u).$$

[4]Cells are by definition closed and not empty.

5. Use a numerical solver for systems of ordinary differential equations to integrate

$$\frac{d}{dt}\mathbf{\Lambda} u = -\mathbf{\Lambda}\,\widetilde{\nabla}\cdot(\boldsymbol{H}\circ\mathcal{R}\,\mathbf{\Lambda} u) \tag{2.2c}$$

in time.

We summarize the result of the above approximations as **numerical time stepping operator:**

$$u(t+\Delta t,\cdot) = \mathrm{TS}(u,t,\Delta t). \tag{2.2d}$$

Regarding the numerical scheme equation (2.2c) **defines** the way updates in time to the data functionals are computed. As a formal consequence of equation (2.2b) it is certainly only an approximation. Bearing in mind that we are dealing with numerical approximations anyway, we follow the common practice in literature and stick to the '=' sign throughout. We shall subsequently use the symbol u to denote the density function as well as an approximation to it obtained via a recovery procedure from the discrete values of data functionals: $u \approx \mathcal{R}\,\mathbf{\Lambda} u$.

2.2 Unstructured Grids

A fairly general data structure for discretizing the computational domain Ω is a conforming subdivision into simplices. A set of simplices is called **conforming,** if the intersection of any two simplices is a common sub-simplex of the two. To avoid ambiguities we give the following

Definition 2.2. A **simplex** Σ is the closed convex hull of $d+1$ points $V := \{\vec{v}_0,\ldots,\vec{v}_d\} \subset \mathbb{R}^d$, where V is not subset of a hyperplane of $\mathbb{R}^d$. Any closed convex hull of a subset of V is called a **sub-simplex** of Σ (this includes the empty set). The elements of V are the extreme points of Σ and called **vertices**, sub-simplices formed by two vertices are called **edges** and sub-simplices formed by d vertices are called **faces** of Σ.

Simplices having precisely one common face are called **neighbours.** Faces which belong to precisely one simplex are called **boundary faces.** Vertices of these are **boundary vertices** and the corresponding simplices are **boundary simplices.**

Definition 2.3. A conforming collection $\mathcal{G}$ of simplices $\Sigma \subset \mathbb{R}^d$ is called a **primary unstructured grid** for the computational domain $\Omega \subset \mathbb{R}^d$, if the

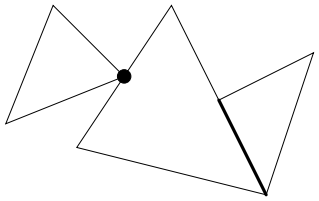

Figure 2.1: Nonconforming triangles in two dimensions.

boundary faces of the simplices constitute a proper polyhedral approximation to the boundary of Ω, i.e. all boundary vertices of the grid lie on the boundary of Ω and no vertices lie outside Ω.

The following information is required to perform computations on a primary unstructured grid:

1. An array for the d coordinates of each vertex,
2. an array for the $d+1$ indices of the vertices of each simplex (the boundary simplices should be the first ones in the array),
3. an array for the $d+1$ indices of the neighbours of each simplex (boundary simplices have less neighbours which may be indicated by filling the unused space with negative entries, furthermore, minus one minus such a negative value could point to an entry in an array of boundary faces[5]) and
4. an array for the d indices of the vertices $\vec{v}_1, \dots \vec{v}_d$ of each boundary face. These should be ordered such that the vector defined by the formal determinant

$$\begin{vmatrix} \vec{e}_1 & (\vec{v}_2 - \vec{v}_1) \cdot \vec{e}_1 & \dots & (\vec{v}_d - \vec{v}_1) \cdot \vec{e}_1 \\ \vdots & \vdots & & \vdots \\ \vec{e}_d & (\vec{v}_2 - \vec{v}_1) \cdot \vec{e}_d & \dots & (\vec{v}_d - \vec{v}_1) \cdot \vec{e}_d \end{vmatrix}$$

   points out of Ω.

For computational purposes the grids should satisfy certain regularity conditions: the density of the vertices should vary smoothly over the computational domain and the simplices should not be degenerate. As a measure of degeneracy we introduce the following function which is invariant under similarity transformations

$$\min_{\Sigma \in \mathcal{G}} \frac{|\Sigma|}{|\partial\Sigma| \operatorname{diam} \Sigma}.$$

[5]We use the C-convention: the first element in an array has index zero.

It should be **uniformly** bounded away from zero for all grids under consideration. For a single equilateral simplex ($\mathcal{G} = \{\Sigma\}$) in d dimensions this fraction takes the value[6]

$$\frac{1}{d(d+1)}\sqrt{\frac{d}{2(d+1)}}.$$

Construction of Boxes

One is often interested in control volumes which are more ball-shaped than simplices (and still tessalate the computational domain, of course). Boxes are polyhedra constructed on top of a given primary unstructured grid $\mathcal{G}$. Each simplex is first barycentrically subdivided into smaller simplices. The (conforming) union of all smaller simplices containing a fixed vertex of the unstructured grid is called the **box** for that vertex. The grid composed of the boxes is sometimes referred to as the **secondary grid.**

Definition 2.4. A **barycentric** subdivision of a simplex $\Sigma \subset \mathbb{R}^d$ is a conforming subdivision of Σ into $(d+1)!$ smaller simplices by uniquely assigning a smaller simplex to each of the $(d+1)!$ possible permutations π of the integers $0, \ldots, d$ in the following way: Letting $\{\vec{v}_{\Sigma,0}, \ldots, \vec{v}_{\Sigma,d}\}$ denote the vertices of Σ define

$$\Sigma_\pi := \overline{\text{co}} \left\{ \frac{1}{k+1} \sum_{j=0}^{k} \vec{v}_{\Sigma,\pi(j)} : k = 0, \ldots, d \right\}$$

(see figure 2.2). The **box** $B_{\vec{v}}$ for the vertex $\vec{v}$ is then defined as

$$B_{\vec{v}} := \bigcup \big\{ \Sigma_\pi : \Sigma \in \mathcal{G} \text{ and } \vec{v}_{\Sigma,\pi(0)} = \vec{v} \big\}.$$

A **box grid** is the set of all boxes.

Two boxes $B_{\vec{v}}$ and $B_{\vec{v}'}$ are neighbours, if and only if $\vec{v}$ and $\vec{v}'$ are vertices of a common simplex, i.e. joined by an edge. Hence the intersection of the neighbours $B_{\vec{v}}$ and $B_{\vec{v}'}$ contains faces of some Σ_π and thus has codimension one.

Collocation Grids

Collocation schemes basically operate on a set of points smoothly distributed over the computational domain. Neighbourhood relations between the points

[6] A proof may be found in the appendix.

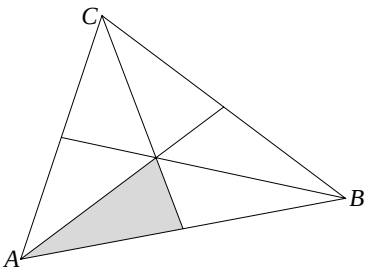

Figure 2.2: Barycentric subdivision of a triangle with vertex indices $\{A, B, C\} = \{0, 1, 2\}$. The shaded triangle corresponds to the permutation $\pi(A) = 0$, $\pi(B) = 1$ and $\pi(C) = 2$.

are established via **edges** connecting points. In order to compute the outer normal on the boundary it is convenient to store the same boundary information as for the primary grids, i.e. **boundary faces.** Edges of boundary faces are **boundary edges.** We obtain the following data structure for edge grids:

1. An array for the d coordinates of each point,
2. an array for the two indices of the endpoints of each edge (boundary edges should be the first ones in the array)
3. an array for the d indices of vertices that form a boundary face.

Similar to the construction of boxes for primary unstructured grids one may specify an alternate set of data locations for the collocation functionals based on the edge grid. We will always observe the following conditions when choosing such alternate data locations:

- There is a one to one correspondence between the data locations and the points in the edge grid (like there is between boxes and vertices in a primary grid) and
- the neighbourhood relations of the edge grid remain meaningful for the alternate data locations.

Definition 2.5. A **collocation grid** is the set of all data locations. The plain term **grid** refers to either primary unstructured grids, box grids or collocation grids. These are all sets of their respective data locations.

2.3 Boundary Conditions

We demand that the computational domain has to accurately approximate the geometry of the modeled phenomenon. Physical boundaries, such as surfaces of solid bodies, have to be reproduced by the polyhedral boundary of the grid. If such surfaces happen to have corners, the grid, too, should have corners accordingly. Smooth curved surfaces can only be approximated up to a certain degree. We allow the computational grid to cover less (or more) than the actual domain, but insist that the boundary of the grid approximates the smooth parts of the computational domain with high order of accuracy. We shall use the symbol Ω to denote both the computational domain and the union of the data locations in the grid:

$$\Omega \approx \bigcup_{\Sigma \in \mathcal{G}} \Sigma.$$

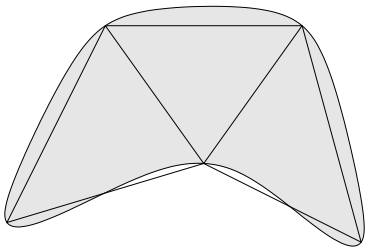

Figure 2.3: Triangulation of a computational domain.

There are now two kinds of boundary conditions:

- Physical boundary conditions which arise from the modeled phenomenon, like external forces and surfaces of solid bodies and
- numerical boundary conditions that have no correspondence in the modeled problem, but originate from restricting the computational domain to a compact set.

The physical boundary conditions have to be applied on the boundary of the grid instead of the computational domain. The error introduced by this geometrical discretization is of order $\mathcal{O}(h^2)$ (h is a characteristic local length of the grid) for a polygonal approximation of a smooth curved surface. While the treatment of the physical boundary conditions models the laws of nature, the numerical boundary conditions have to **define** a flux across the boundary based on the inner state near the boundary and possibly some prescribed information about the outer state. One may employ the information available

from the eigenvalues of $(\partial \boldsymbol{F}\vec{n}/\partial u)(u_i)$ to decide how many characteristics (in the direction of the outer normal $\vec{n}$) enter into the computational domain and use as little of a prescribed outer state as possible.

2.4 Time Integration

Having completed the approximations mentioned above, equation (2.2c) finally implies (we drop the u's) the following system of ordinary differential equations:

$$\frac{d}{dt}\boldsymbol{\Lambda} = \mathcal{L}(t, \boldsymbol{\Lambda}) \tag{2.3}$$

For purely convective phenomena $\boldsymbol{F}$ depends on u alone, so inside the computational domain Ω the right hand side of (1.2) does not explicitly depend on the time t, yet this may be the case for the boundary conditions. We use the explicit TVB Runge-Kutta time stepping schemes as developed by Shu and Osher [SO88, SO89] to integrate

$$\boldsymbol{\Lambda}(t + \Delta t) = \boldsymbol{\Lambda}(t) + \int_t^{t+\Delta t} \mathcal{L}(\tau, \boldsymbol{\Lambda}(\tau))\, d\tau.$$

These schemes have favorable numerical properties, since the involved integration weights are non-negative. For ease of presentation we let n denote the approximation order of the scheme and

$$\begin{aligned} \boldsymbol{\Lambda}^{(0)} &:= \boldsymbol{\Lambda}(t), \\ \boldsymbol{\Lambda}(t + \Delta t) &:= \boldsymbol{\Lambda}^{(n)} \text{ and} \\ \mathcal{L}^{(s)} &:= \mathcal{L}(t + \tau^{(s)}\Delta t, \boldsymbol{\Lambda}^{(s)}). \end{aligned}$$

For all schemes up to third order the τ's are defined by $\tau^{(0)} := 0$, $\tau^{(1)} := 1$, $\tau^{(2)} := 1/2$ and the initial stage is simply first order forward Euler:

$$\boldsymbol{\Lambda}^{(1)} := \boldsymbol{\Lambda}^{(0)} + \Delta t\, \mathcal{L}^{(0)}. \tag{2.4}$$

The second order scheme reads:

$$\boldsymbol{\Lambda}^{(2)} := \frac{1}{2}\boldsymbol{\Lambda}^{(0)} + \frac{1}{2}\boldsymbol{\Lambda}^{(1)} + \frac{1}{2}\Delta t\, \mathcal{L}^{(1)} \tag{2.5}$$

and the third order scheme:

$$\boldsymbol{\Lambda}^{(2)} := \frac{3}{4}\boldsymbol{\Lambda}^{(0)} + \frac{1}{4}\boldsymbol{\Lambda}^{(1)} + \frac{1}{4}\Delta t\, \mathcal{L}^{(1)} \tag{2.6}$$

$$\boldsymbol{\Lambda}^{(3)} := \frac{1}{3}\boldsymbol{\Lambda}^{(0)} + \frac{2}{3}\boldsymbol{\Lambda}^{(2)} + \frac{2}{3}\Delta t\, \mathcal{L}^{(2)}. \tag{2.7}$$

Shu and Osher also present an explicit fourth order scheme, which fails to have only non-negative weights, and suggest an (expensive) dual problem strategy to avoid the numerical instability related to the use of negative weights. We do not use this fourth order scheme.

It should be noted that the third order scheme, like the second order scheme, requires only one additional temporary copy of the data vector.

2.5 The Finite Volume Method

The finite volume method discretizes the integral form of the conservation principle. It is most frequently used on box grids, but success has as well been reported for primary grid variants. We let u_i and u_o denote the inner and outer limit of u. It is taken separately for each component as in equation (1.6) on page 11. Having tessalated the computational domain into cells, the data functionals are the corresponding averaging functionals:

$$\frac{d}{dt}\delta_\Sigma u = -\frac{1}{|\Sigma|}\int_{\partial\Sigma} \boldsymbol{F}(u_i, u_o, \vec{n})\, do. \tag{2.8}$$

Historically this discretization was perhaps first proposed by Godunov in [God59] where exact solutions of the Riemann problems at the cell interfaces were attempted. Such exact solutions are time consuming and the numerical properties of the exact Riemann flux are not very favourable, it is for instance not necessarily a smooth function. Furthermore, the integration process is effectively a projection onto a piecewise constant function and discards all information about the behaviour of u inside the cell Σ (except the average value, of course). While use of a recovery procedure may give a higher order approximation to u, replacing the integrand on the right hand side of equation (2.8) with a numerical flux function $\boldsymbol{H}$ (denoting the prescribed boundary flux $\boldsymbol{B}$ for faces on the boundary) does not increase the discretization error:

$$\frac{d}{dt}\delta_\Sigma u = -\frac{1}{|\Sigma|}\int_{\partial\Sigma} \boldsymbol{H}(u_i, u_o, \vec{n})\, do. \tag{2.9}$$

Finite volume schemes enjoy a **geometric conservation property:** Since the control volumes tessalate the domain, the fluxes across interfaces inside Ω cancel out and we have

$$\sum_{\Sigma\in\mathcal{G}} \int_{\partial\Sigma} \boldsymbol{H}(u_i, u_o, \vec{n})\, do = \int_{\partial\Omega} \boldsymbol{B}(u, t, \vec{x})\, do. \tag{2.10a}$$

This implies

$$\frac{d}{dt}\delta_\Omega u = \frac{1}{|\Omega|}\sum_{\Sigma\in\mathcal{G}}|\Sigma|\,\frac{d}{dt}\delta_\Sigma u = -\frac{1}{|\Omega|}\int_{\partial\Omega}\boldsymbol{B}(u,t,\vec{x})\,do. \tag{2.10b}$$

In one space dimension the integral on the right hand side of equation (2.8) reduces to a difference of two fluxes:

$$\frac{d}{dt}\delta_\Sigma u = -\frac{\boldsymbol{H}(u_i^{(r)},u_o^{(r)},1)-\boldsymbol{H}(u_i^{(l)},u_o^{(l)},-1)}{|\Sigma|}$$

with a superscript (r) referring to the right endpoint of Σ and (l) to the left. For equations in more than one space dimension the integration in equation (2.8) has to be carried out numerically via a suitable (typically Gaussian) quadrature rule $\mathrm{Q}_{\partial\Sigma}$ with **positive weights.** In this setting the requirement of positive weights has an evident physical interpretation, since a negative weight would reverse the local transport obtained as the approximate solution of a physically relevant Riemann problem. The finite volume method is now obtained by defining the time derivative of the data functionals in terms of the unstructured grid

$$\frac{d}{dt}\delta_\Sigma u := -\frac{1}{|\Sigma|}\,\mathrm{Q}_{\partial\Sigma}\,\boldsymbol{H}(u_i,u_o,\vec{n}) \tag{2.11}$$

for each $\Sigma\in\mathcal{G}$ with the prescribed boundary flux $\boldsymbol{B}$ being used instead of $\boldsymbol{H}$ where appropriate. Since the boundaries of the control volumes are composed of fragments of hyperplanes, the numerical integration should be carried out on each such fragment separately. The quadrature rule for each fragment which is not on the boundary of Ω is then used for both adjacent control volumes. The numerical integrations inside Ω now cancel out in exactly the same way as did the integrations in equation (2.10a):

$$\begin{aligned}\frac{d}{dt}\delta_\Omega u &= -\frac{1}{|\Omega|}\sum_{\Sigma\in\mathcal{G}}\mathrm{Q}_{\partial\Sigma}\,\boldsymbol{H}(u_i,u_o,\vec{n})\\ &= -\frac{1}{|\Omega|}\,\mathrm{Q}_{\partial\Omega}\,\boldsymbol{B}(u,t,\vec{x}).\end{aligned} \tag{2.12}$$

In two space dimensions the cell boundaries are polygons. The points and weights of the gaussian quadrature rule for up to five points can be computed exactly and are given in table 2.1 on page 55 for the unit interval.

n	points	weights
2	$\gamma_{1,2} = \frac{1}{2} \pm \frac{1}{6}\sqrt{3}$	$w_{1,2} = \frac{1}{2}$
3	$\gamma_{1,3} = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{3}{5}}$ $\gamma_2 = \frac{1}{2}$	$w_{1,3} = \frac{5}{18}$ $w_2 = \frac{4}{9}$
4	$\gamma_{1,4} = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{3}{7} + \frac{2}{7}\sqrt{\frac{6}{5}}}$ $\gamma_{2,3} = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{3}{7} - \frac{2}{7}\sqrt{\frac{6}{5}}}$	$w_{1,4} = \frac{1}{4} - \frac{1}{12}\sqrt{\frac{5}{6}}$ $w_{2,3} = \frac{1}{4} + \frac{1}{12}\sqrt{\frac{5}{6}}$
5	$\gamma_{1,5} = \frac{1}{2} \pm \frac{1}{6}\sqrt{5 + 2\sqrt{\frac{10}{7}}}$ $\gamma_{2,4} = \frac{1}{2} \pm \frac{1}{6}\sqrt{5 - 2\sqrt{\frac{10}{7}}}$ $\gamma_3 = \frac{1}{2}$	$w_{1,5} = \frac{7}{900}\left(23 - \frac{13}{14}\sqrt{70}\right)$ $w_{2,4} = \frac{7}{900}\left(23 + \frac{13}{14}\sqrt{70}\right)$ $w_3 = \frac{64}{225}$

Table 2.1: Gaussian quadrature on $[0, 1]$

2.6 Reconstruction

The data functionals representing u are commonly visualized as a piecewise constant function

$$(\mathcal{R}^{\text{trivial}}\,\mathbf{\Lambda} u)|_{\Sigma} := \delta_{\Sigma} u. \tag{2.13}$$

If we are dealing with a cell based scheme, this defines the reconstruction **inside** all cells in $\mathcal{G}$, but the reconstructions obtained for two adjacent cells may not coincide on the intersection of the two cells. For a collocation scheme we extend each point value of u locally to a neighbourhood, such that these neighbourhoods again tessalate the domain, but we do not specify the neighbourhoods exactly. Even if the underlying function u is smooth, its approximation by a piecewise constant function obviously is not. The fact that the reconstruction is well defined only inside the cells or neighbourhoods, not on their boundaries, is not a problem when using supremum norms, since the reconstructed function will be a piecewise polynomial.

The discontinuities that are not present in the underlying function, but only introduced in the course of the discretization are of magnitude $\mathcal{O}(h)$ (h denotes a characteristic local length, for instance the maximal diameter of nearby cells) and decrease – slowly – when the grid is refined. For these discontinuities it would be perfectly acceptable to use the smooth case flux function at a suitably chosen intermediate state instead of the numerical Riemann solver for inner and outer state. The problem is to decide, based on the values of the data functionals, whether the unknown underlying function is locally or one sided smooth. In the spirit of van Leer [vL74, vL77, vL79] we treat the higher order approximation of each component of the density distribution from its data functional values as a purely approximation theoretical problem which is decoupled completely from the physical stage of computing the fluxes. A suitable recovery procedure[7]

$$\mathcal{R} : \mathbb{R}^{\mathcal{G}} \to \left(\Pi^{q}(\mathbb{R}^{d} \to \mathbb{R})\right)^{\mathcal{G}}$$

has to

- interpolate the values of the data functionals on each data location:

$$\delta_{\Sigma}(\mathcal{R}\,\mathbf{\Lambda} u) = \delta_{\Sigma} u, \tag{2.14}$$

- approximate any smooth function with high order of accuracy,

[7] We only consider polynomial recovery up to a fixed degree.

- locally take values within the range of those of the data functionals and in particular
- avoid the Gibbs phenomenon (i.e. not oscillate near discontinuities).

Consequently it will not operate on arbitrary input in always exactly the same way, but operation will be modified depending on the data: **recovery is a non-linear process.** As a general rule we demand that these modifications depend smoothly on the input data, no digital switching should occur.

For each $\Sigma \in \mathcal{G}$ the interpolation requirement (2.14) defines a hyperplane in $\Pi^q(\mathbb{R}^d \to \mathbb{R})$. Typical recovery strategies

- choose a compact convex subset of that hyperplane in a data independent (i.e. linear) fashion and then
- pick one element of that set such that variation or oscillation is small. (Recently even schemes choosing several elements and using a different one for each part of the numerical boundary integration have been proposed [HS99].)

The following methods represent extreme cases within this framework: One may either choose the convex hull of

1. (WENO) a large number of high order interpolants or
2. (limiting) just one high order and the locally constant interpolant.

At the time of this writing WENO ("weighted essentially non oscillatory") schemes are the more commonly used and a live topic of research. They trace their origins back to Harten et al. [HO87] and [HEOC87]. The idea of weighting instead of digital candidate selection was introduced in [LOC94]. Friedrich [Fri98] has constructed a scheme with quadratic polynomials on box grids, much of the analytic background can be found in [Son97b]. In his thesis Hempel [Hem99] proposes an interesting limiting strategy and successfully applies this strategy to the Euler equations using a dynamically adapted unstructured grid with boxes. Sonar [Son97a, Son98] investigates the general structure from the viewpoint of optimal recovery.

As WENO schemes always form a convex combination of higher order approximations, they never loose formal order of accuracy. This may turn into a drawback, if there is no stencil available for which the data vector permits a reasonable high order approximation. Ollivier-Gooch [OG97] suggests a very promising strategy for switching the reconstruction degree according to the quality of the local data.

Approximation

In order to compute a polynomial interpolant for the data location $\Sigma \in \mathcal{G}$ we need to consider the values of other data functionals as well. The space $\Pi^q(\mathbb{R}^d \to \mathbb{R})$ of polynomials in d variables up to degree q has dimension

$$\dim \Pi^q(\mathbb{R}^d \to \mathbb{R}) = \binom{d+q}{q}.$$

Let us denote the kernel in $\Pi^q(\mathbb{R}^d \to \mathbb{R})$ of δ by

$$\delta^\perp := \big\{\pi \in \Pi^q(\mathbb{R}^d \to \mathbb{R}) : \delta\pi = 0\big\}. \tag{2.15}$$

It is a hyperplane of $\Pi^q(\mathbb{R}^d \to \mathbb{R})$ and one has

$$\dim \delta^\perp = \binom{d+q}{q} - 1.$$

We want to establish a stability result for the interpolation or least-squares approximation problem of functions from the values of certain data functionals. The stability should be uniform under (isotropic) scaling of the grid (reflection, translation and rotation present no difficulty). We need to show that the operation of the data functionals can, in a way, be reversed and that this reversal is uniformly stable, if the grid is refined.

For any bounded linear operator $L : V \to W$ between the Banach spaces V and W the operator norm of L is defined as

$$\|L\| := \sup_{x \in V\setminus\{0\}} \frac{\|Lx\|_W}{\|x\|_V}.$$

The norm of the inverse of L can be defined, even if that inverse does not exist, in which case the norm will be infinite:

$$\big\|L^{-1}\big\| := \sup_{x \in V\setminus\{0\}} \frac{\|x\|_V}{\|Lx\|_W}.$$

To estimate the worst case error propagation in the solution of the linear problem $Lx = y$ one estimates the worst case error in $\tilde{y} := LL^{-1}y$ which leads to the usual definition of the condition of L:

$$\operatorname{cond} L := \big\|L\big\|\,\big\|L^{-1}\big\|.$$

The following definition follows the pattern of the definition of $\|L^{-1}\|$ above:

Definition 2.6. A **stencil** $\mathcal{S}$ for Σ is a subset of $\mathcal{G}$ which contains Σ. It is convenient to define the symbol $\mathcal{S}_\Sigma := \mathcal{S} \setminus \{\Sigma\}$. For the region covered by the cells in $\mathcal{S}$ we use the symbol

$$\Sigma_\mathcal{S} := \bigcup_{\Theta \in \mathcal{S}} \Theta \subset \Omega.$$

We define the **condition** of $\mathcal{S}$ as

$$\operatorname{cond} \mathcal{S} := \sup_{\pi \in \delta_\Sigma^\perp \setminus \{0\}} \frac{\|\pi\|_{\infty,\Sigma_\mathcal{S}}}{\|(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma}\|_\infty}.$$

If the denominator happens to reach zero, the condition is infinite. More formally – and less intuitively – the definition could be written down as

$$\operatorname{cond} \mathcal{S} := \sup \left\{ \alpha \in \mathbb{R}_{\geqslant 0} : \alpha \, \|(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma}\|_\infty \leqslant \|\pi\|_{\infty,\Sigma_\mathcal{S}} \text{ for all } \pi \in \delta_\Sigma^\perp \right\}.$$

The stencil $\mathcal{S}$ is called **admissible,** if its condition stays below a prescribed constant $M_{\text{cond}} \in \mathbb{R}_{>1}$.

A large condition number indicates that a stencil is not well-suited for interpolation or least-squares approximation. We reject any computational grid which fails to provide sufficiently many admissible stencils.

Lemma 2.7. The condition of $\mathcal{S}$ is invariant under regular affine transformations of $\mathbb{R}^d$.

Proof. Let $A : \mathbb{R}^d \to \mathbb{R}^d$ be a regular affine transformation with inverse A^{-1}. We have $\delta_\Sigma(\pi \circ A) = \delta_{A(\Sigma)}\pi$ and hence $\pi \in \delta_\Sigma^\perp \iff \tilde{\pi} := \pi \circ A^{-1} \in \delta_{A(\Sigma)}^\perp$. Therefore

$$\sup_{\pi \in \delta_\Sigma^\perp \setminus \{0\}} \frac{\|\pi\|_{\infty,\Sigma_\mathcal{S}}}{\|(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma}\|_\infty} = \sup_{\tilde{\pi} \in \delta_{A(\Sigma)}^\perp \setminus \{0\}} \frac{\|\tilde{\pi}\|_{\infty,A(\Sigma_\mathcal{S})}}{\left\|(\delta_{A(\Theta)} \tilde{\pi})_{\Theta \in \mathcal{S}_\Sigma}\right\|_\infty},$$

since $\|\pi\|_{\infty,\Sigma_\mathcal{S}} = \|\tilde{\pi}\|_{\infty,A(\Sigma_\mathcal{S})}$ and $\delta_\Theta \pi = \delta_{A(\Theta)} \tilde{\pi}$ for all $\Theta \in \mathcal{S}_\Sigma$. □

The previous lemma is the key to the stability of polynomial approximation. It applies to general regular affine transformations and essentially establishes uniform stability of polynomial approximation, even under anisotropic grid refinement. The proof relies on the fact that tranformed versions of polynomials are again polynomials of the same degree, i.e. polynomial spaces are invariant under regular affine transformations. In other trial spaces one would probably have to transform the trial space along with the grid in order to obtain condition numbers independent of the grids meshsize.

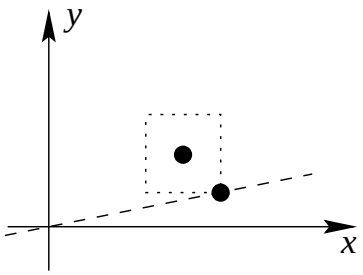

Figure 2.4: Best approximation from a linear subspace with respect to the maximum norm. The norm of the proximum may be greater than the norm of the element being approximated.

On an admissible stencil $\mathcal{S}$ we now have as an immediate consequence of definition 2.6 for any $\pi \in \delta_\Sigma^\perp$

$$\|\pi\|_{\infty,\Sigma_\mathcal{S}} \leqslant M_{\text{cond}} \|(\delta_\Theta \pi)_{\Theta\in\mathcal{S}_\Sigma}\|_\infty \tag{2.16}$$

and for any $u \in BL^\infty(\Omega \to \mathbb{R})$ by the triangle inequality

$$\|\pi\|_{\infty,\Sigma_\mathcal{S}} \leqslant M_{\text{cond}} \left(\|(\delta_\Theta u - \delta_\Sigma u)_{\Theta\in\mathcal{S}_\Sigma}\|_\infty + \|(\delta_\Theta(\pi - u) + \delta_\Sigma u)_{\Theta\in\mathcal{S}_\Sigma}\|_\infty\right).$$

If we can find a polynomial $\pi \in \delta^\perp$ which interpolates $(\delta_\Theta u - \delta_\Sigma u)_{\Theta\in\mathcal{S}_\Sigma}$ exactly, the last term vanishes. Otherwise π is chosen to minimize the residual $\|(\delta_\Theta(\pi - u) + \delta_\Sigma u)_{\Theta\in\mathcal{S}_\Sigma}\|_\infty$. Since the maximum norm is not strictly convex, this kind of best approximation may admit infinitely many solutions, furthermore these solutions are hard to find. An upper bound for the minimal value of

$$\|(\delta_\Theta(\pi - u) + \delta_\Sigma u)_{\Theta\in\mathcal{S}_\Sigma}\|_\infty$$

can be computed by considering the residual obtained for the particular choice $\pi := 0$. That value is $\|(\delta_\Theta u - \delta_\Sigma u)_{\Theta\in\mathcal{S}_\Sigma}\|_\infty$ and gives us the following estimate for a best approximation with respect to the maximum norm:

$$\|\pi\|_{\infty,\Sigma_\mathcal{S}} \leqslant 2M_{\text{cond}} \|(\delta_\Theta u - \delta_\Sigma u)_{\Theta\in\mathcal{S}_\Sigma}\|_\infty.$$

For any stencil $\mathcal{S}$ on which we wish to compute a high order interpolant $\mathcal{S}_\Sigma$ must have at least $\dim \delta_\Sigma^\perp$ elements, since otherwise by the well known kernel-image theorem some nonzero polynomials will have all zero functional values on $\mathcal{S}_\Sigma$. Such a stencil has therefore an inifinite condition number and fails to be admissible.

We attempt to solve a suitable least-squares approximation problem instead of the difficult maximum norm approximation. Introducing a scalar

product on $\mathbb{R}^{\mathcal{S}_\Sigma}$ by

$$\langle \alpha; \beta \rangle_{\mathbb{R}^{\mathcal{S}_\Sigma}} := \alpha^{\mathsf{t}} G \beta$$

with a symmetric and positive definite Gram matrix $G \in \mathbb{R}^{\mathcal{S}_\Sigma \times \mathcal{S}_\Sigma}$ and the Euclidean G-norm

$$\|\alpha\|_G := \sqrt{\langle \alpha; \alpha \rangle_{\mathbb{R}^{\mathcal{S}_\Sigma}}}$$

we consider the following least-squares approximation problem: Find $\pi \in \delta_\Sigma^\perp$ such that

$$\left\| (\delta_\Theta \pi - (\delta_\Theta - \delta_\Sigma) u)_{\Theta \in \mathcal{S}_\Sigma} \right\|_G^2 \to \min .$$

In order to estimate the maximum norm in terms of the G-norm we need

$$G_\infty := \sup_{\alpha \in \mathbb{R}^{\mathcal{S}_\Sigma} \setminus \{0\}} \frac{\|\alpha\|_\infty}{\|\alpha\|_G} \qquad \text{and} \qquad G^\infty := \sup_{\alpha \in \mathbb{R}^{\mathcal{S}_\Sigma} \setminus \{0\}} \frac{\|\alpha\|_G}{\|\alpha\|_\infty}.$$

These suprema are always finite, since $\mathbb{R}^{\mathcal{S}_\Sigma}$ is finite dimensional.

Lemma 2.8. The polynomial least-squares approximation problem on an admissible stencil $\mathcal{S}$ is well posed.

Proof. Let $\pi \in \delta_\Sigma^\perp$ denote the polynomial obtained as the solution of the least-squares approximation problem. The data vector $(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma}$ of the least-squares solution and the residual are G-orthogonal. Hence

$$\|(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma}\|_G^2 + \|(\delta_\Theta(\pi - u) + \delta_\Sigma u)_{\Theta \in \mathcal{S}_\Sigma}\|_G^2 = \|(\delta_\Theta u - \delta_\Sigma u)_{\Theta \in \mathcal{S}_\Sigma}\|_G^2$$

which implies

$$\|(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma}\|_G \leqslant \|(\delta_\Theta u - \delta_\Sigma u)_{\Theta \in \mathcal{S}_\Sigma}\|_G .$$

From equation (2.16) and the definition of G_∞ we infer

$$\begin{aligned} \|\pi\|_{\infty, \Sigma_\mathcal{S}} &\leqslant M_{\mathrm{cond}} G_\infty \|(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma}\|_G \\ &\leqslant M_{\mathrm{cond}} G_\infty \|(\delta_\Theta u - \delta_\Sigma u)_{\Theta \in \mathcal{S}_\Sigma}\|_G . \end{aligned} \tag{2.17}$$

On the other hand

$$\frac{1}{G^\infty} \|(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma})\|_G \leqslant \|(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma})\|_\infty \leqslant \|\pi\|_{\infty, \Sigma_\mathcal{S}} ,$$

since $\|\delta_\Theta\| \leqslant 1$. Now

$$\|(\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma})\|_G \leqslant G^\infty \|\pi\|_{\infty, \Sigma_\mathcal{S}} \leqslant M_{\mathrm{cond}} G_\infty G^\infty \|(\delta_\Theta u - \delta_\Sigma u)_{\Theta \in \mathcal{S}_\Sigma}\|_G .$$

The approximation problem thus has the condition bound $M_{\mathrm{cond}} G_\infty G^\infty$. □

Lemma 2.9. The projection $P : BL^\infty(\Sigma_\mathcal{S} \to \mathbb{R}) \to \Pi^q(\Sigma_\mathcal{S} \to \mathbb{R})$ obtained by polynomial least-squares approximation of the values of the data functionals $(\delta_\Theta)_{\Theta\in\mathcal{S}}$ on an admissible stencil $\mathcal{S}$ has the norm

$$\|P\| \leqslant 1 + 2M_{\text{cond}}G_\infty G^\infty. \tag{2.18}$$

Proof. From equation (2.17) follows

$$\begin{aligned}\|Pu - \delta_\Sigma u\|_{\infty,\Sigma_\mathcal{S}} &\leqslant M_{\text{cond}}G_\infty G^\infty \|(\delta_\Theta u - \delta_\Sigma u)_{\Theta\in\mathcal{S}_\Sigma}\|_\infty \\ &\leqslant M_{\text{cond}}G_\infty G^\infty \|u - \delta_\Sigma u\|_{\infty,\Sigma_\mathcal{S}},\end{aligned}$$

since $\|\delta_\Theta\| \leqslant 1$. Furthermore

$$\|u - \delta_\Sigma u\|_{\infty,\Sigma_\mathcal{S}} \leqslant \|u\|_{\infty,\Sigma_\mathcal{S}} + |\delta_\Sigma u| \leqslant 2\|u\|_{\infty,\Sigma_\mathcal{S}} \tag{2.19}$$

and

$$\begin{aligned}\|Pu\|_{\infty,\Sigma_\mathcal{S}} &\leqslant \|Pu - \delta_\Sigma u\|_{\infty,\Sigma_\mathcal{S}} + |\delta_\Sigma u| \\ &\leqslant M_{\text{cond}}G_\infty G^\infty \|u - \delta_\Sigma u\|_{\infty,\Sigma_\mathcal{S}} + \|u\|_{\infty,\Sigma_\mathcal{S}} \\ &\leqslant (1 + 2M_{\text{cond}}G_\infty G^\infty)\|u\|_{\infty,\Sigma_\mathcal{S}}.\end{aligned}$$

□

In practice, the Gram matrix G is always diagonal and merely provides geometric weighting of the value of a particular δ_Θ ($\Theta \in \mathcal{S}_\Sigma$) in terms of powers of

$$\frac{\|(\delta_\Theta - \delta_\Sigma)\vec{x}\|_{\mathbb{R}^d}}{\operatorname{diam}\Sigma_\mathcal{S}}.$$

This kind of matrix is invariant under translations, (isotropic) scaling, reflection and rotations. Under these transformations G_∞ and G^∞ are therefore uniformly bounded for all stencils in all grids under consideration and we have

Theorem 2.10. *Polynomial least-squares approximation on admissible stencils is uniformly stable under similarity transformations of the grid.*

The following simple lemma is crucial for proving local convergence of the approximating polynomial to smooth functions.

Lemma 2.11. Let $P : BL^\infty(\Sigma_\mathcal{S} \to \mathbb{R}) \to \Pi^q(\Sigma_\mathcal{S} \to \mathbb{R})$ be a bounded projection with norm $\|P\|$. Then we have for any polynomial $\pi \in \Pi^q(\Sigma_\mathcal{S} \to \mathbb{R})$:

$$\|u - Pu\|_{\infty,\Sigma_\mathcal{S}} \leqslant (1 + \|P\|)\|u - \pi\|_{\infty,\Sigma_\mathcal{S}}.$$

Proof. By the triangle inequality

$$\begin{aligned}\|u - Pu\|_{\infty,\Sigma_{\mathcal{S}}} &\leqslant \|u - \pi\|_{\infty,\Sigma_{\mathcal{S}}} + \|\pi - Pu\|_{\infty,\Sigma_{\mathcal{S}}} \\ &\leqslant \|u - \pi\|_{\infty,\Sigma_{\mathcal{S}}} + \|P\pi - Pu\|_{\infty,\Sigma_{\mathcal{S}}} \\ &\leqslant \|u - \pi\|_{\infty,\Sigma_{\mathcal{S}}} + \|P\| \, \|u - \pi\|_{\infty,\Sigma_{\mathcal{S}}} .\end{aligned}$$

□

We like to think of differentiability in terms of an approximation property:

Definition 2.12. A function $u \in L^\infty(\Omega \to \mathbb{R})$ is said to be differentiable q times at $\vec{x}_0 \in \overline{\operatorname{int}\Omega} = \Omega$, if and only if there are a polynomial $\pi \in \Pi^q(\Omega \to \mathbb{R})$ and residual functions R_h such that

$$u(\vec{x}) = \pi(\vec{x}) + h^q R_h(\vec{x})$$

for all $\vec{x} \in B_h(\vec{x}_0) \cap \Omega$ and $\lim_{h\to 0} \sup\{|R_h(\vec{x})| : \vec{x} \in B_h(\vec{x}_0) \cap \Omega\} = 0$. If u is differentiable q times at $\vec{x}_0$, then the polynomial π is determined uniquely and called the **Taylor polynomial** $T_{\vec{x}_0}u$ of degree q for u at $\vec{x}_0$. The space of all functions u differentiable q times for which the Taylor mapping $T_\bullet u : \Omega \to \Pi^q(\Omega \to \mathbb{R})$ is continuous is denoted by $C^q(\Omega \to \mathbb{R})$.

This definition includes the boundary of the compact set Ω. Therefore for $u \in C^q(\Omega \to \mathbb{R})$ boundedness of the Taylor mapping does not have to be assumed separately, but follows from the compactness of Ω. For a function $u \in C^{q+1}(\Omega \to \mathbb{R})$ one may write u as the sum of its Taylor polynomial of degree q at $\vec{x}_0$ and a residual bounded above by $C_{u,\Omega}h^{q+1}$ where $C_{u,\Omega} \in \mathbb{R}$ depends on u and Ω alone, not on $\vec{x}_0$.

Choosing $\vec{x}_0$ as the barycentre of Σ and scaling the grid until $\Sigma_{\mathcal{S}} \subset B_h(\vec{x}_0)$ we may use the Taylor polynomial of degree q of $u \in C^{q+1}(\Omega \to \mathbb{R})$ in lemma 2.11 to obtain the following convergence estimate:

$$\|u - Pu\|_{\infty,\Sigma_{\mathcal{S}}} \leqslant (1 + \|P\|) \, \|u - T_{\vec{x}_0}u\|_{\infty,\Sigma_{\mathcal{S}}} \leqslant C_{u,\Omega}(1 + \|P\|)h^{q+1}.$$

We now turn to the actual computation. Restricting ourselves to a particular choice of basis functions we obtain an algebraic expression for the least squares approximation problem. The condition of this algebraic problem cannot be better than that of the original problem. The proof of lemma 2.7 suggests making the basis functions invariant under affine transformations of $\mathbb{R}^d$. This could be achieved by using barycentric coordinates. We are, however, not concerned with anisotropic grid refinement and restrict our theory to similarity transformations by choosing the basis as functions of

$$\frac{(1 - \delta_\Sigma)\vec{x}}{\operatorname{diam}\Sigma_{\mathcal{S}}},$$

that is: we shift the barycentre of Σ to the origin and scale by the diameter of the stencil. This transformation makes the basis functions invariant under translation, rotation and scaling of the grid. Letting

$$\mathcal{B} := (\pi_1, \dots, \pi_{\dim \delta_\Sigma^\perp})$$

denote a basis of $\delta_\Sigma^\perp$ the problem can be stated in matrix form:

$$\begin{aligned} A &:= (\delta_\Theta \pi)_{\Theta \in \mathcal{S}_\Sigma, \pi \in \mathcal{B}} \in \mathbb{R}^{\mathcal{S}_\Sigma \times \mathcal{B}} \\ \theta &:= ((\delta_\Theta - \delta_\Sigma) u)_{\Theta \in \mathcal{S}_\Sigma} \in \mathbb{R}^{\mathcal{S}_\Sigma}. \end{aligned}$$

This gives us the algebraic least-squares problem $\|A\xi - \theta\|_G^2 \to \min$. It can be solved uniquely, since for an admissible stencil $\mathcal{S}$ the matrix A has maximum (column) rank $\dim \delta_\Sigma^\perp$. The approximating polynomial then is

$$\mathcal{B}\xi + \delta_\Sigma u,$$

i.e. with $\xi := (\xi_1, \dots, \xi_{\dim \delta_\Sigma^\perp})^\mathsf{t}$ we have $\delta_\Sigma u + \sum_{k=1}^{\dim \delta_\Sigma^\perp} \xi_k \pi_k$ as solution. Using either the normal equations

$$A^\mathsf{t} G A \xi = A^\mathsf{t} G \theta \qquad \Longleftrightarrow \qquad \xi = (A^\mathsf{t} G A)^{-1} A^\mathsf{t} G \theta$$

or the Cholesky decomposition $G = C^\mathsf{t} C$ and the QU-decomposition of CA

$$CA = Q \begin{pmatrix} U \\ 0 \end{pmatrix}$$

with an orthogonal matrix $Q = Q^{-\mathsf{t}} \in \mathbb{R}^{\mathcal{S}_\Sigma \times \mathcal{S}_\Sigma}$ and a regular upper triangular matrix $U \in \mathbb{R}^{\mathcal{B} \times \mathcal{B}}$ we obtain

$$\xi = \left(U^{-1} | 0\right) Q^\mathsf{t} C \theta.$$

One verifies immediately that the solutions of both the normal equations and the QU-process agree:

$$\begin{aligned} (A^\mathsf{t} G A)^{-1} A^\mathsf{t} G &= [(CA)^\mathsf{t} (CA)]^{-1} (CA)^\mathsf{t} C \\ &= \left[\left(U^\mathsf{t} | 0\right) Q^\mathsf{t} Q \begin{pmatrix} U \\ 0 \end{pmatrix} \right]^{-1} \left(U^\mathsf{t} | 0\right) Q^\mathsf{t} C \\ &= U^{-1} U^{-\mathsf{t}} \left(U^\mathsf{t} | 0\right) Q^\mathsf{t} C = \left(U^{-1} | 0\right) Q^\mathsf{t} C. \end{aligned}$$

If the basis $\mathcal{B}$ consists of homogeneous polynomials, then the QU-decomposition via Householder transformations is uniformly stable even without scaling the basis functions by $\operatorname{diam} \Sigma_\mathcal{S}$.

Stencil Selection

Let us now describe the actual strategy for selecting reconstruction stencils $\mathcal{S}$ for a data location Σ. In addition to the algebraic admissibility condition discussed in the previous section there are a few more or less heuristic conditions. The general idea is most conveniently described for box grids in two space dimensions.

The stencils should be local and not contain holes. In order to perform linear recovery for a cell Σ we need to include at least two additional cells. These should both be neighbours of Σ and of themselves, see the left part of figure 2.5. Stencil sets of this kind contain candidates which avoid interpolating across a possible discontinuity near Σ. It appears reasonable that the region $\Sigma_{\mathcal{S}}$ covered by the stencil and the cell is not spread unnecessarily. Furthermore, cells whose barycentres lie almost on a straight line would form an ill-conditioned stencil, since linear interpolation in two variables cannot be performed on a line. In the example in figure 2.5 left there are six one sided stencils for linear interpolation and a central stencil for linear approximation available.

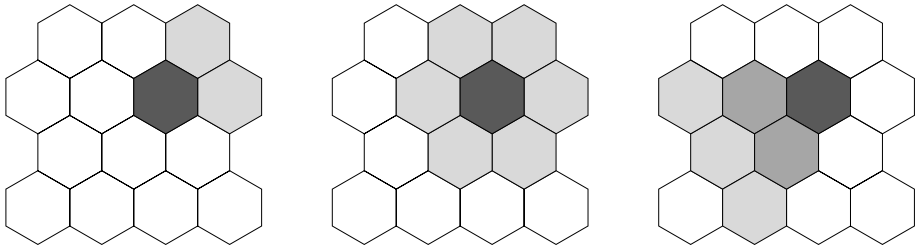

Figure 2.5: Different kinds of stencils for regular box grids in two space dimensions. Left: stencil for linear interpolation. Middle: central stencil for linear or quadratic approximation. Right: One sided stencil for quadratic interpolation. The reconstruction is performed for the dark shaded box, other boxes in the stencil are shaded more lightly, depending on the neighbourhood level. Boxes outside the stencil are left white.

For quadratic recovery we need five additional cells. The central stencil in the middle part of figure 2.5 allows a quadratic least-squares approximation to be computed. In order to avoid interpolation or approximation based on data from both sides of a discontinuity we need to consider one sided stencils as well. These will involve less direct neighbours of Σ, and more remote neighbours instead. We might consider the six alternatives of omitting one light shaded direct neighbour from the middle part of figure 2.5 and further variants using subsequently less direct neighbours. Numerical experience

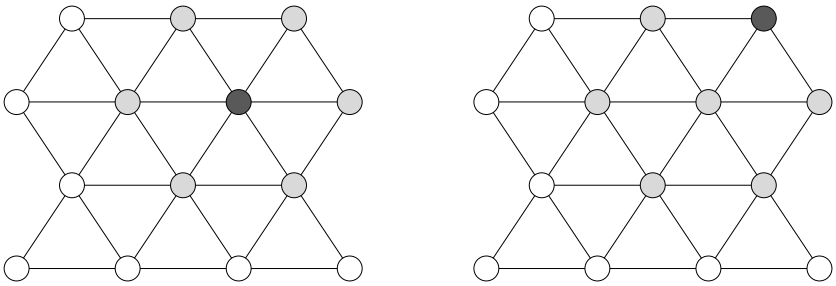

Figure 2.6: Different kinds of reconstruction stencils for collocation grids in two space dimensions. Left: central stencil for linear approximation. Right: one sided stencil for linear approximation. Reconstruction is performed for the dark shaded vertex, other vertices in the stencil are shaded more lightly and points outside the stencil are left white.

indicates that it is necessary to consider stencils involving only two direct neighbours, similar to the right part of figure 2.5, in order to stay clear of discontinuities and obtain non-oscillating interpolants. The convex hull of the barycentres of boxes in a stencil should not contain barycentres of boxes outside the stencil. Using the primary triangulation data structure these types of stencils can be computed very efficiently.

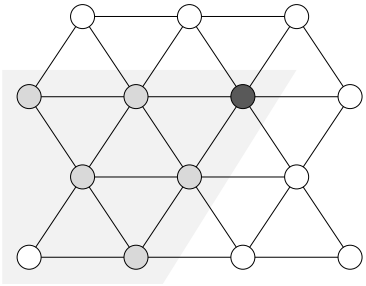

Figure 2.7: One sided stencil consisting of direct and second neighbours for quadratic interpolation on a collocation grid. The points in the stencil should lie within a sufficiently narrow cone.

In the case of a collocation grid stencils similar to the left part of figure 2.5 should be precomputed and stored. The geometric principles for constructing recovery stencils on collocation grids are quite close to those for stencils on box grids, however, collocation stencils cannot be said to be free of holes. Their construction requires some effort, as there is no underlying simplex structure available. Fortunately, it turns out that linear recovery can be performed using just precomputed central stencils. Figure 2.6 shows the

central stencil for the dark shaded vertex in its left part. This stencil is also used as a one sided stencil for the dark vertex in the right part of figure 2.6.

Quadratic recovery, however, requires more one sided stencils which contain fewer direct neighbours than the right part of figure 2.6. To collect sufficiently many points we have to consider second and possibly third level neighbours inside a cone about the dark shaded vertex. Different stencils would be obtained by rotating the cone about its apex.

Oscillation Indicators

WENO schemes form a convex combination of the approximations computed on the various stencils for any given data location. Each candidate is assigned a weight according to its "oscillation" via an oscillation indicator. Choice of an oscillation indicator is largely heuristic: the variation of the reconstructed function should be small. While the gradient is the only degree of freedom one has when reconstructing linearly, it is not clear how second derivatives of quadratic approximations should be treated. It appears that second derivatives are too sensitive to changes in the input data to be of practical value for determining reconstruction weights. Ignoring a candidates local curvature is certainly not very satisfactory, but seems to work in practice. We shall make use of the following simple oscillation indicator:

$$\omega_\Sigma(u) := \sqrt{\delta_\Sigma\left(\|\nabla u\|^2\right)}. \tag{2.20a}$$

Letting $\mathcal{M}_\Sigma$ denote the set of all stencils considered for the data location Σ and $\pi_\mathcal{S}$ the polynomial obtained for the particular stencil $\mathcal{S} \in \mathcal{M}_\Sigma$ we compute the weight $w_\mathcal{S}$ as

$$w_\mathcal{S} := \frac{1}{\varepsilon + (\omega_\Sigma(\pi_\mathcal{S}))^\alpha} > 0 \tag{2.20b}$$

where α is typically 4 or 8 and $\varepsilon \approx 10^{-15}$ a small number to avoid division by zero. The reconstruction function is then computed as

$$\frac{\sum\limits_{\mathcal{S}\in\mathcal{M}_\Sigma} w_\mathcal{S}\,\pi_\mathcal{S}}{\sum\limits_{\mathcal{S}\in\mathcal{M}_\Sigma} w_\mathcal{S}}. \tag{2.20c}$$

Numerical experience indicates that the quality of the solutions improves, if approximations on central stencils are preferred over those on one sided stencils. This can be achieved via additional stencil type weights $g_\mathcal{S}$ which

are about 10 to 50 for central stencils and about one for one sided stencils. Equation (2.20b) would thus be modified:

$$w_{\mathcal{S}} := \frac{g_{\mathcal{S}}}{\varepsilon + (\omega_\Sigma(\pi_{\mathcal{S}}))^\alpha}.$$

An oscillation indicator like that of equation (2.20a) is not well suited for quadratic recovery from collocation functionals for single points, since accidental measurement of oscillation near the zero of a candidates gradient would give this candidate an extremely large weight (about $1/\varepsilon$) and produce a reconstruction function almost identical to this candidate. For quadratic recovery from collocation values it is therefore mandatory to use convex combinations of several collocation functionals as input data.

For the local approximation order of the reconstruction function of a smooth function in $C^{q+1}(\Omega \to \mathbb{R})$ we obtain

$$\left\| u - \frac{\sum_{\mathcal{S}\in\mathcal{M}_\Sigma} w_{\mathcal{S}}\, \pi_{\mathcal{S}}}{\sum_{\mathcal{S}\in\mathcal{M}_\Sigma} w_{\mathcal{S}}} \right\|_{\infty,\Sigma_{\mathcal{S}}} = \left\| \frac{\sum_{\mathcal{S}\in\mathcal{M}_\Sigma} w_{\mathcal{S}}\, (u - \pi_{\mathcal{S}})}{\sum_{\mathcal{S}\in\mathcal{M}_\Sigma} w_{\mathcal{S}}} \right\|_{\infty,\Sigma_{\mathcal{S}}} \leqslant \max_{\mathcal{S}\in\mathcal{M}_\Sigma} \|u - \pi_{\mathcal{S}}\|_{\infty,\Sigma_{\mathcal{S}}}$$
$$\leqslant C_{u,\Omega} \left(1 + \max_{\mathcal{S}\in\mathcal{M}_\Sigma} \|P_{\mathcal{S}}\|\right) h^{q+1}$$

where $P_{\mathcal{S}}$ denotes the projection mapping obtained for the data location Σ on the stencil $\mathcal{S}$ and $C_{u,\Omega} \in \mathbb{R}$ is a constant depending on u and Ω alone. By lemma 2.9 and theorem 2.10 $\|P_{\mathcal{S}}\|$ is uniformly bounded for all stencils in all grids under consideration.

2.7 Numerical Divergence Operator

For cell based schemes the numerical divergence approximation consists in the application of a quadrature rule to the flux across cell interfaces. The normal directions on the cell interfaces are dictated by the computational grid, the choice of quadrature points is defined by a quadrature rule, see figure 2.8.

In order to generalize this concept to collocation methods as well we assign to each data location Σ a set of **evaluation points** $\vec{x} \in \mathbb{R}^d$ and **evaluation directions** $\vec{n} \in \mathbb{R}^d$ with $\|\vec{n}\| = 1$, see figure 2.10. These evaluation points and directions now define a set $\mathcal{F}$ of new data functionals

$$\delta\vec{g} = \vec{g}(\vec{x}) \cdot \vec{n}.$$

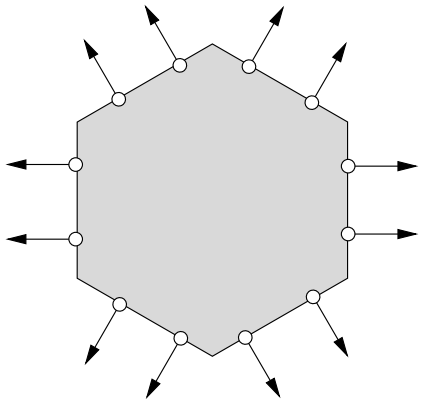

Figure 2.8: Quadrature points and outer normal directions for the boundary integral of the flux across cell interfaces.

from which we seek to approximate a **feature functional** Φ (the divergence) of an unknown vector valued function.

This problem can be solved by first computing an interpolant or approximation for the given input data and then evaluating the divergence of that function. Uniform stability of this approximation process under translation, rotation and scaling of the grid is basically established by the general theory that led to theorem 2.10, but there are few technical catches.

Approximation of Linear Functionals

Even if the data functionals do not contain enough information to compute a polynomial approximation, it may still be possible to evaluate the feature functional. This happens, if the feature functional does not depend on what is missing. We denote by

$$\Phi^\perp := \left\{ \vec{\pi} \in \Pi^f(\mathbb{R}^d \to \mathbb{R}^d) : \Phi\vec{\pi} = 0 \right\}$$

the kernel of the feature functional, by

$$\delta^\perp := \left\{ \vec{\pi} \in \Pi^f(\mathbb{R}^d \to \mathbb{R}^d) : \delta\vec{\pi} = 0 \right\}$$

again the kernel of a data functional and by

$$\ker \mathcal{F} := \bigcap_{\delta \in \mathcal{F}} \delta^\perp$$

the intersection of the kernels of the data functionals in $\mathcal{F}$. The feature functional Φ can be approximated from the values of the data functionals, if

$$\ker \mathcal{F} \subset \Phi^\perp.$$

An optimal choice of data functionals has maximal $\ker \mathcal{F}$, since we then accumulate exactly the information required for approximating the feature functional, like the construction of Gaussian quadrature formulae does for integration. In this respect the finite volume method represents an example of an almost optimal choice. By the divergence theorem the cell average of a functions divergence can be computed from the outer normal component of the function values on the cell boundary and the Gauss quadrature ensures maximum order of accuracy for each boundary part.

The proper polynomial trial space for the approximation of Φ from the functionals in $\mathcal{F}$ is

$$\mathcal{F}^\perp := \Pi^f(\mathbb{R}^d \to \mathbb{R}^d)/\ker \mathcal{F}.$$

We let Υ denote the closed convex hull of the evaluation points for the functionals in $\mathcal{F}$ and introduce the following norm on $\Pi(\Upsilon \to \mathbb{R}^d)$:

$$\|\vec{\pi}\|_{\Upsilon,\infty} := \sup_{\vec{x}\in\Upsilon} \|\vec{\pi}(\vec{x})\|_{\mathbb{R}^d}$$

and the norm relative to $\mathcal{F}^\perp$:

$$\|\vec{\pi}\|_{\mathcal{F}^\perp} := \inf_{\vec{\omega}\in\ker\mathcal{F}} \|\vec{\pi} - \vec{\omega}\|_{\Upsilon,\infty}.$$

The functionals in $\mathcal{F}$ now have norm one, since

$$\delta\vec{\pi} \leqslant \|\vec{\pi}(\vec{x})\|_{\mathbb{R}^d} \leqslant \|\vec{\pi}\|_{\Upsilon,\infty}$$

and equality holds, if $\vec{\pi} = \vec{n}$ on all Υ. It should be noted that these constant functions are not in $\ker \mathcal{F}$, due to the form of the (new) data functionals. The condition of $\mathcal{F}$ can be defined as

$$\operatorname{cond} \mathcal{F} := \sup_{\vec{\pi}\in\Pi^f(\mathbb{R}^d\to\mathbb{R}^d)\setminus\ker\mathcal{F}} \frac{\|\vec{\pi}\|_{\mathcal{F}^\perp}}{\|(\delta\vec{\pi})_{\delta\in\mathcal{F}}\|_\infty}.$$

Next we need to specify a reasonably large class of transformations of $\mathbb{R}^d$ under which the condition of $\mathcal{F}$ is invariant. We do not attempt to cover general affine transformations, since stretching leads to a number of problems connected to the treatment of the evaluation directions: If we stretch figure 2.8 horizontally, the hexagon would (seen from a distance) approximate a rectangle and all outer normal vectors would tend to a horizental position. They would thus not approximate the outer normals of the rectangle.

For this reason we shall only consider similarity transformations of the grid. The evaluation directions remain invariant under translations and scaling of the grid, but are rotated along with it. Invariance of the condition of $\mathcal{F}$ under these transformations can now be established similar to lemma 2.7 by considering the following transformations of $\vec{\pi}$:

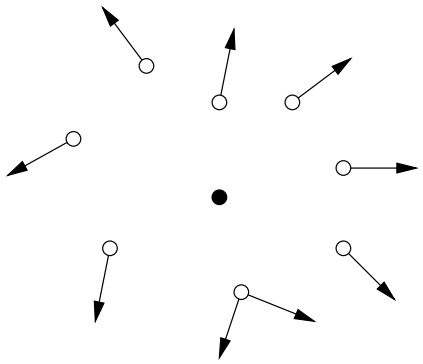

Figure 2.9: Stencil for approximating the feature functional at a certain data location. It consists of evaluation points (white circles) and evaluation directions (arrows). Evaluation points with up to d evaluation directions may coincide.

- $\vec{\pi} \circ A^{-1}$ for translations and isotropic scaling operations A and
- $A \circ \vec{\pi} \circ A^{-1}$ for orthogonal transformations A of the grid.

If $\vec{\pi}$ is in $\ker \mathcal{F}$, then the transformation of $\vec{\pi}$ is in the kernel of the accordingly transformed data functionals. The norm of the data functionals does not change under the transformation. We may hence expect that the condition of the vector valued approximation problem is uniformly bounded under these kinds of transformations.

Next we consider again a least-squares approximation problem. We introduce on $\mathbb{R}^{\mathcal{F}}$ a scalar product

$$\langle \alpha; \beta \rangle_{\mathbb{R}^{\mathcal{F}}} := \alpha^{\mathsf{t}} G \beta$$

with a symmetric and positive definite Gram matrix G, the Euclidean G-norm

$$\|\alpha\|_G := \sqrt{\langle \alpha; \alpha \rangle_{\mathbb{R}^{\mathcal{F}}}}$$

and solve the following problem: Find $\vec{\pi} \in \mathcal{F}^{\perp}$ such that

$$\left\| (\delta\vec{\pi} - \delta\vec{g})_{\delta \in \mathcal{F}} \right\|_G^2 \to \min .$$

The Gram matrix G is again simply a diagonal matrix which provides geometric weighting in terms of powers of

$$\frac{\|\vec{x} - \text{barycentre}\, \Upsilon\|_{\mathbb{R}^d}}{\text{diam}\, \Upsilon}.$$

Choosing a basis

$$\mathcal{B} = (\vec{\pi}_1, \ldots, \vec{\pi}_{\dim \mathcal{F}^\perp})$$

of $\mathcal{F}^\perp$ we obtain the matrix form $\|A\xi - \theta\|_G^2 \to \min$ of the least-squares problem with

$$A := (\delta\vec{\pi})_{\delta\in\mathcal{F},\vec{\pi}\in\mathcal{B}} \in \mathbb{R}^{\mathcal{F}\times\mathcal{B}}$$
$$\theta := (\delta\vec{g})_{\delta\in\mathcal{F}} \in \mathbb{R}^{\mathcal{F}}.$$

The approximating vector valued polynomial is

$$\mathcal{B}\xi = \sum_{k=1}^{\dim \mathcal{F}^\perp} \xi_k \vec{\pi}_k$$

and the feature functional can be evaluated as

$$\Phi\mathcal{B}\xi = \sum_{k=1}^{\dim \mathcal{F}^\perp} \xi_k \Phi\vec{\pi}_k.$$

We are ultimately interested in evaluating the feature functional, and not in the approximating polynomial as such. This now suggests computing a vector of feature weights $\phi \in \mathbb{R}^{\mathcal{F}}$ which gives

$$\Phi\mathcal{B}\xi = \langle\phi;\theta\rangle_{\mathbb{R}^{\mathcal{F}}}$$

by simply forming the scalar product of the vector of the feature weights with the input data vector. Since we have

$$A\xi \approx \theta,$$

we would expect

$$\Phi\mathcal{B}\xi \approx \langle\phi; A\xi\rangle_{\mathbb{R}^{\mathcal{F}}}.$$

This leads us to looking for ϕ as a solution of

$$\Phi\mathcal{B} \approx \phi^{\mathsf{t}} GA.$$

The last relation is not overdetermined and can be solved exactly (A has maximal rank), but not necessarily uniquely. In fact, we have obtained the **dual problem** of the originial least-squares approximation problem: Find $\phi \in \mathbb{R}^{\mathcal{F}}$ such that

$$\phi^{\mathsf{t}} GA = \Phi\mathcal{B} \text{ and } \|\phi\|_G \to \min. \tag{2.21}$$

We observe that the dual problem is independent of the particular choice of basis polynomials. Changing the basis amounts to multiplying the last equation from the right with a regular $\mathbb{R}^{\mathcal{B}\times\mathcal{B}}$ matrix and does not alter ϕ. The dual problem is most conveniently solved using the QU-decomposition of CA (as usual we denote the Cholesky decomposition of the Gram matrix by $G = C^{\mathsf{t}}C$)

$$CA = Q\begin{pmatrix} U \\ \hline 0 \end{pmatrix}$$

with an orthogonal matrix $Q = Q^{-\mathsf{t}} \in \mathbb{R}^{\mathcal{F}\times\mathcal{F}}$ and a regular upper triangular matrix $U \in \mathbb{R}^{\mathcal{B}\times\mathcal{B}}$. We obtain

$$\phi^{\mathsf{t}} := \Phi\mathcal{B}\left(U^{-1}|0\right)Q^{\mathsf{t}}C^{-\mathsf{t}}. \tag{2.22}$$

Lemma 2.13. The least-squares approximation to the feature functional Φ obtained by evaluating the feature functional for a least-squares approximation of the input data values is equivalent to taking the scalar product of the vector of feature weights obtained from the dual problem and the input data vector.

Proof. We need to show that

$$\langle\phi;\theta\rangle_{\mathbb{R}^{\mathcal{F}}} = \Phi\mathcal{B}\xi$$

for all $\theta \in \mathbb{R}^{\mathcal{F}}$. As ξ depends on θ via a least-squares problem, this is equivalent to

$$\langle\phi;\theta\rangle_{\mathbb{R}^{\mathcal{F}}} = \Phi\mathcal{B}\left(U^{-1}|0\right)Q^{\mathsf{t}}C\theta.$$

The latter is indeed the case, since we have by the definition (2.22) of ϕ^{t}

$$\langle\phi;\theta\rangle_{\mathbb{R}^{\mathcal{F}}} = \phi^{\mathsf{t}}G\theta = \Phi\mathcal{B}\left(U^{-1}|0\right)Q^{\mathsf{t}}C^{-\mathsf{t}}C^{\mathsf{t}}C\theta.$$

□

Based on lemma 2.13 we can precompute the required vectors of feature weights during the preprocessing stage. They only have to be recomputed, if the grid is changed. Furthermore, the tasks of generating a divergence formula and the actual divergence computation are now completely separated. It is convenient to store $\phi^{\mathsf{t}}G = \Phi\mathcal{B}\left(U^{-1}|0\right)Q^{\mathsf{t}}C$.

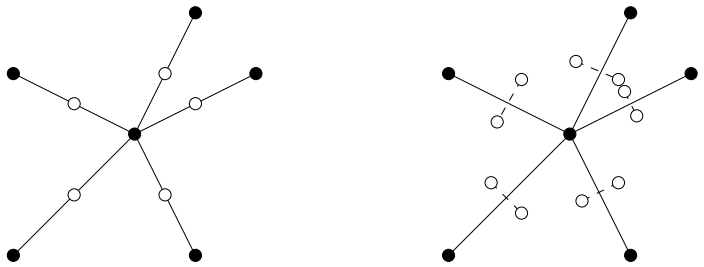

Figure 2.10: Two kinds of divergence stencils. The evaluation directions are parallel to the edges joining the dark vertices (barycentres of data locations). In the right part two evaluation points (white circles) per edge are considered. They are chosen on the mid-perpendicular of each edge, at a fixed portion (one fifth) of the edges length from the midpoint. Each evaluation point is affiliated to two dark vertices and the evaluation directions for both dark vertices are collinear.

The Divergence Functional

Choosing the basis $\mathcal{B}$ of $\mathcal{F}^\perp$ as functions of

$$\frac{\vec{x} - \text{barycentre}\,\Upsilon}{\text{diam}\,\Upsilon}$$

we obtain for the divergence of the basis functions:

$$\text{div}\,\mathcal{B} = \frac{1}{\text{diam}\,\Upsilon}\left(\text{div}\,\vec{\pi}\left(\frac{\vec{x} - \text{barycentre}\,\Upsilon}{\text{diam}\,\Upsilon}\right)\right)_{\vec{\pi}\in\mathcal{B}}.$$

While the condition of the interpolation matrix $A := (\delta\vec{\pi})_{\delta\in\mathcal{F},\vec{\pi}\in\mathcal{B}}$ can be arranged to be invariant under scaling of the grid, we loose one order of accuracy via the right hand side of the dual problem (2.21), as should be expected for first derivatives. For a first order approximation to the divergence operator we choose as in [AHS99]

$$\Pi^1(\Upsilon \to \mathbb{R}^d)$$

as trial space. It has $\dim \Pi^1(\Upsilon \to \mathbb{R}^d) = d(d+1)$. If all evaluation directions all lead radially away from the barycentre of the data location for which the divergence is to be computed (see figure 2.10 left), then the "eddy" functions (taking the barycentre as the origin)

$$\vec{x} \mapsto (\vec{x}\cdot\vec{e}_j)\vec{e}_k - (\vec{x}\cdot\vec{e}_k)\vec{e}_j \text{ for } j \neq k \in \{1,\ldots,d\}$$

are in $\ker\mathcal{F}$. For this kind of stencil we can (and indeed have to) reduce the dimension of the trial space by $d-1$, the number of independent "eddy"

functions. The reduced trial space has dimension $d^2 + 1$. If there are not enough neighbours available in the collocation grid for interpolation, one has two options:

- One might insert extra edges to increase the number of neighbours or
- increase the number of evaluation points (figure 2.10 right) or evaluation directions without changing the number of neighbours.

It should be noted that this problem is not present for the finite volume method, as the boundary integral can be formally approximated with even one evaluation point.

One extreme case of divergence stencil selection would be the choice of d (linearly independent) evaluation directions per d coinciding evaluation points. The vector of feature weights will then contain d weights for such an evaluation point: one per direction. These weights depend linearly on the evaluation directions for a point. Consequently there exists a basis of evaluation directions for which $d-1$ of the weights vanish. For the direction of the non vanishing weight a numerical flux function should be computed and used as input datum for the divergence formula.

It is, however, not clear, whether a **global** choice of evaluation points and evaluation directions for collocation schemes is possible, such that there is for each evaluation point only one non-zero weight and the evaluation directions for the non-zero weights for the adjacent data locations are linearly dependent. In order to keep the cost of geometric preprocessing in reasonable bounds, we simply choose evaluation points and directions according to the edges, as in figure 2.10.

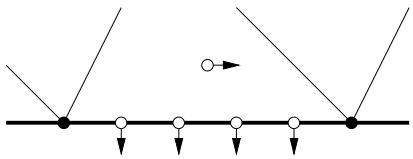

Figure 2.11: Part of divergence stencils for boundary edges. No evaluation point outside the computational domain is chosen. Instead we choose further points (white circles) on the boundary edge itself and outer normals as evaluation directions. The left two evaluation points on the boundary are affiliated to the left dark vertex, the right ones to the right dark vertex. The upper evaluation point is affiliated to both dark vertices.

Higher Order Divergence

A higher order polynomial approximation requires substantially more data,

$$\dim \Pi^f(\mathbb{R}^d \to \mathbb{R}^d) = d\binom{d+f}{f}.$$

Using many levels of neighbours for this computation gave unsatisfactory results and does not seem plausible in terms of the underlying physical transport mechanism. We therefore suggest a higher order divergence formula based on the quasi linear form of the conservation law

$$\operatorname{div}(\boldsymbol{F} \circ u) = \frac{\partial \boldsymbol{F}}{\partial u} \nabla u$$

in smooth regions of the flow. The two divergence formulae for first and higher order approximation can be blended by a simple limiting strategy.

If u is locally C^1 about a data location, then the two values of u considered for an evaluation point differ about $\mathcal{O}(h)$ and so do the values of ∇u computed from the reconstruction on the cell itself and its neighbours. The following smoothness indicator

$$w := \sum (u_i - u_o)^2 + \|\nabla u_i - \nabla u_o\|^2$$

where the sum is taken over all evaluation points will thus be of order $\mathcal{O}(h^2)$ for smooth parts of the flow and large for regions where u fails to be C^1. Thus

$$\left(1 - \frac{w^2}{h^2}\right) \widetilde{\nabla \cdot}(\boldsymbol{F} \circ u) + \frac{w^2}{h^2} \frac{\partial \boldsymbol{F}}{\partial u} \nabla u$$

where $\widetilde{\nabla \cdot}$ represents a stable first order approximation could be used as an asymptotically high order divergence formula. In our numerical experiments, however, this switching strategy did not alter the solutions noticably.

2.8 The CFL Condition

For a one dimensional scheme of Godunov type the waves entering a cell of length h from one end should not be allowed to travel across the whole cell and leave it at the opposite end within a single time step. This leads to a time step constraint of

$$\Delta t < \frac{h}{L_{\boldsymbol{F}}}.$$

In several space dimensions waves entering a cell near a corner across different faces would start interacting very early, but numerical experience indicates that this is not a principal source of trouble. Let us therefore focus on the issue of information propagation in the scheme. For a scalar equation information clearly travels along characteristics at speed $L_{\boldsymbol{F}}$. If the physically relevant part of the numerical scheme, this is the discrete divergence operator for the time step, does not process all of the physically relevant information, then arbitrarily changing the initial values outside the numerical domain of dependence changes the behaviour of the exact solution, but leaves the approximation produced by the scheme unchanged. The scheme can therefore not converge to the correct solution. In fact, a time step chosen too large often manifests itself in numerical instability. Let us clarify the point for the classical Lax-Friedrichs scheme on a grid of uniform mesh size h in one space dimension:

$$\mathrm{TS}(u,t,\Delta t)(x_k) = \\ \frac{u(t,x_{k+1})+u(t,x_{k-1})}{2} - \frac{\Delta t}{2h}\left[\boldsymbol{F}(u(t,x_{k+1})) - \boldsymbol{F}(u(t,x_{k-1}))\right]$$

Here the numerical signal velocity is $h/\Delta t$, as information from neighbours to both sides having distance h enters the numerical divergence approximation, whereas the physical signal velocity is the Lipschitz constant $L_{\boldsymbol{F}}$ of the flux. We demand that the numerical signal velocity be greater than the physical signal velocity:

$$\frac{h}{\Delta t} > L_{\boldsymbol{F}}.$$

This reasoning, originally presented by Courant, Friedrichs and Levy in [CFL28], leads us to the following procedure for computing a legal time step size Δt:

Estimate for each data functional δ a local Lipschitz constant $L_{\boldsymbol{F}}$ of the flux function supplying δu as the argument and a local characteristic length h as half the diameter of the divergence stencil. Define with an arbitrary positive constant $\mathrm{CFL} \in \mathbb{R}_{>0}$:

$$\Delta t := \mathrm{CFL} \min_{\delta \in \boldsymbol{\Lambda}} \frac{h}{L_{\boldsymbol{F}}}.$$

The constant CFL is called the **CFL number.** Constraints of the time step size may be expressed as restrictions on CFL. For the Lax-Friedrichs scheme, for instance, we demand $\mathrm{CFL} < 1$.

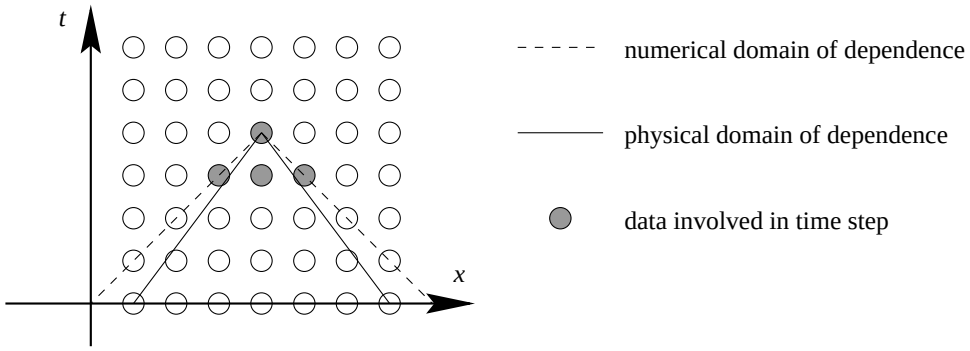

Figure 2.12: The CFL condition: Numerical and physical signal velocity.

For an implicit scheme the numerical signal velocity is infinite, as the implicit time step will link events throughout the grid together. Implicit schemes have much less severe time step constraints in terms of stability and allow considerably larger time steps. These are, however still restricted by the necessity to compute accurate approximations.

One might object that a scheme including substantially more than the physically relevant information to compute the updates in a time step would for reasons similar to those above fail to converge to the correct solution: changing the data outside the physical but within the numerical domain of dependence will change the numerical solution, while the exact solution at a certain point remains unchanged. The major deception about this argumentation is that it postulates propagation along characteristics for the numerical scheme, too. However, the way the "extra" information is used by the scheme is certainly subject to a consistency condition. It will probably slow down the rate of convergence, but as long as the scheme remains stable and consistent not destroy it.

Chapter 3

The Euler Equations of Gas Dynamics

3.1 The Euler Flux Function

The Euler Equations represent the conservation of mass, momentum and energy respectively. They describe the motion of a compressible liquid in the absence of inner friction. In this respect they form a limit case of the system of Navier-Stokes equations. We denote by ρ the density, $\vec{\mathrm{v}}$ the transport velocity, E the specific energy, i.e. energy per mass and by p the pressure. The flux function for the Euler equations reads

$$\boldsymbol{F} = \rho \begin{pmatrix} 1 \\ \vec{\mathrm{v}} \\ \mathrm{E} \end{pmatrix} \vec{\mathrm{v}}^{\mathrm{t}} + \mathrm{p} \begin{pmatrix} 0 \\ \boldsymbol{I} \\ \vec{\mathrm{v}}^{\mathrm{t}} \end{pmatrix} \tag{3.1}$$

(the middle line of equation (3.1) is a shorthand notation for d lines) with an equation of state which for an ideal gas takes the form

$$\mathrm{p} = (\kappa - 1)\rho \left(\mathrm{E} - \tfrac{1}{2} \left\|\vec{\mathrm{v}}\right\|^2\right). \tag{3.2}$$

For an ideal diatomic gas the constant κ takes the value $\kappa = 7/5 = 1.4$. We also introduce the constant $\tilde{\kappa} := \kappa - 1$. The set $S \subset \mathbb{R}^s$ of valid physical states with $s = d + 2$ is restricted to $\rho > 0$ and $\mathrm{p} > 0$. The system of the Euler equations now reads:

$$\frac{\partial}{\partial t} \begin{pmatrix} \rho \\ \rho\vec{\mathrm{v}} \\ \rho\mathrm{E} \end{pmatrix} + \operatorname{div} \left[\rho \begin{pmatrix} 1 \\ \vec{\mathrm{v}} \\ \mathrm{E} \end{pmatrix} \vec{\mathrm{v}}^{\mathrm{t}} + \mathrm{p} \begin{pmatrix} 0 \\ \boldsymbol{I} \\ \vec{\mathrm{v}}^{\mathrm{t}} \end{pmatrix} \right] = 0. \tag{3.3}$$

The quantities differentiated with respect to time

$$u := \begin{pmatrix} \rho \\ \rho\vec{\mathrm{v}} \\ \rho\mathrm{E} \end{pmatrix}$$

are called **conservative variables:** density (of mass), density of momentum and density of energy. We define the enthalpy H, the (unscaled) thermodynamic entropy Z and the speed of sound a as

$$\begin{aligned} \mathrm{H} &:= \mathrm{E} + \frac{\mathrm{p}}{\rho} \\ \mathrm{Z} &:= \ln \frac{\mathrm{p}}{\rho^\kappa} \\ \mathrm{a} &:= \sqrt{\frac{\kappa \mathrm{p}}{\rho}}. \end{aligned} \tag{3.4a}$$

The term "unscaled" indicates that we have omitted the specific heat capacity at constant volume from the definition of the thermodynamic entropy. The thermodynamic entropy is a concave function. Compared to the discussion of the abstract entropy on page 18 this corresponds to a different sign convention, i.e. the thermodynamic entropy will increase across a discontinuity. The following relations are useful:

$$\begin{aligned} \mathrm{H} &= \kappa\mathrm{E} - \tfrac{1}{2}\tilde{\kappa} \left\|\vec{\mathrm{v}}\right\|^2 \\ \tilde{\kappa}\mathrm{H} &= \mathrm{a}^2 + \tfrac{1}{2}\tilde{\kappa} \left\|\vec{\mathrm{v}}\right\|^2 \\ \nabla_u \mathrm{p} &= \tilde{\kappa} \left(\tfrac{1}{2} \left\|\vec{\mathrm{v}}\right\|^2, \quad -\vec{\mathrm{v}}^{\mathsf{t}}, \quad 1\right) \\ \nabla_u \mathrm{Z} &= \frac{1}{\mathrm{p}} \left(\tfrac{1}{2}\tilde{\kappa} \left\|\vec{\mathrm{v}}\right\|^2 - \mathrm{a}^2, \quad -\tilde{\kappa}\vec{\mathrm{v}}^{\mathsf{t}}, \quad \tilde{\kappa}\right) \\ \nabla_u \mathrm{a} &= \frac{1}{2\rho\mathrm{a}} \left(\tfrac{1}{2}\kappa\tilde{\kappa} \left\|\vec{\mathrm{v}}\right\|^2 - \mathrm{a}^2, \quad -\kappa\tilde{\kappa}\vec{\mathrm{v}}^{\mathsf{t}}, \quad \kappa\tilde{\kappa}\right). \end{aligned} \tag{3.4b}$$

The index u to the nabla operator indicates that differentiation is performed with respect to the conservative variables Along a fixed unit vector $\vec{n} \in \mathbb{R}^d$ we define

$$\mathrm{v}_{\vec{n}} := \vec{\mathrm{v}} \cdot \vec{n}.$$

After multiplication with this vector $\vec{n}$ equation (3.1) takes the form

$$\boldsymbol{F}\vec{n} = \begin{pmatrix} \rho\mathrm{v}_{\vec{n}} \\ \rho\mathrm{v}_{\vec{n}}\vec{\mathrm{v}} + \mathrm{p}\vec{n} \\ \rho\mathrm{H}\mathrm{v}_{\vec{n}} \end{pmatrix} \tag{3.5}$$

Obviously $\boldsymbol{F}\vec{n}$ is invariant under rotations of the physical space, i.e. for a rotational matrix $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{s\times s}$ with $B := \operatorname{diag}(1, A, 1)$ one has

$$B^{-1}\boldsymbol{F}(Bu)A\vec{n} = \boldsymbol{F}(u)\vec{n}.$$

The Jacobi matrix of $\boldsymbol{F}\vec{n}$ with respect to the conservative variables is:

$$\frac{\partial \boldsymbol{F}\vec{n}}{\partial(\rho, \rho\vec{\mathrm{v}}, \rho\mathrm{E})} = \begin{pmatrix} 0 & \vec{n}^{\mathsf{t}} & 0 \\ \frac{1}{2}\tilde{\kappa}\left\|\vec{\mathrm{v}}\right\|^2 \vec{n} - \mathrm{v}_{\vec{n}}\vec{\mathrm{v}} & \mathrm{v}_{\vec{n}}\boldsymbol{I} + \vec{\mathrm{v}}\otimes\vec{n} - \tilde{\kappa}\vec{n}\otimes\vec{\mathrm{v}} & \tilde{\kappa}\vec{n} \\ \left(\tilde{\kappa}\left\|\vec{\mathrm{v}}\right\|^2 - \kappa\mathrm{E}\right)\mathrm{v}_{\vec{n}} & \left(\kappa\mathrm{E} - \frac{1}{2}\tilde{\kappa}\left\|\vec{\mathrm{v}}\right\|^2\right)\vec{n}^{\mathsf{t}} - \tilde{\kappa}\mathrm{v}_{\vec{n}}\vec{\mathrm{v}}^{\mathsf{t}} & \kappa\mathrm{v}_{\vec{n}} \end{pmatrix}.$$

It has the eigenvalues ($\sigma_\lambda \in \{-1, 1\}$ as in definition 1.20 on page 37)

$$\begin{aligned} \lambda_1(u,\vec{n}) &= \mathrm{v}_{\vec{n}} - \sigma_\lambda \mathrm{a} \\ \lambda_2(u,\vec{n}) = \cdots = \lambda_{s-1}(u,\vec{n}) &= \mathrm{v}_{\vec{n}} \\ \lambda_s(u,\vec{n}) &= \mathrm{v}_{\vec{n}} + \sigma_\lambda \mathrm{a} \end{aligned} \tag{3.6a}$$

and corresponding right eigenvectors

$$\begin{aligned} r_1(u,\vec{n}) &= \begin{pmatrix} 1 \\ \vec{\mathrm{v}} - \sigma_\lambda \mathrm{a}\vec{n} \\ \mathrm{H} - \sigma_\lambda \mathrm{a}\mathrm{v}_{\vec{n}} \end{pmatrix} \\ r_2(u,\vec{n}) &= \begin{pmatrix} 1 \\ \vec{\mathrm{v}} \\ \frac{1}{2}\left\|\vec{\mathrm{v}}\right\|^2 \end{pmatrix} \\ r_{2+j}(u,\vec{n}) &= \begin{pmatrix} 0 \\ \vec{e}_j \\ \vec{\mathrm{v}}\cdot\vec{e}_j \end{pmatrix} \text{ for } j \in \{1,\ldots,d-1\} \\ r_s(u,\vec{n}) &= \begin{pmatrix} 1 \\ \vec{\mathrm{v}} + \sigma_\lambda \mathrm{a}\vec{n} \\ \mathrm{H} + \sigma_\lambda \mathrm{a}\mathrm{v}_{\vec{n}} \end{pmatrix} \end{aligned} \tag{3.6b}$$

Here $\vec{e}_0 := \vec{n}, \vec{e}_1, \ldots, \vec{e}_{d-1}$ denotes an orthonormal basis of $\mathbb{R}^d$. We define the matrix

$$R(u,\vec{n}) := (r_1(u,\vec{n}), \ldots, r_s(u,\vec{n})).$$

The inverse of this matrix is given by

$$R^{-1}(u,\vec{n}) = \frac{1}{2\mathrm{a}^2}\begin{pmatrix} \frac{1}{2}\tilde{\kappa}\left\|\vec{\mathrm{v}}\right\|^2 + \sigma_\lambda \mathrm{a}\mathrm{v}_{\vec{n}} & -\sigma_\lambda \mathrm{a}\vec{n}^{\mathsf{t}} - \tilde{\kappa}\vec{\mathrm{v}}^{\mathsf{t}} & \tilde{\kappa} \\ 2\mathrm{a}^2 - \tilde{\kappa}\left\|\vec{\mathrm{v}}\right\|^2 & 2\tilde{\kappa}\vec{\mathrm{v}}^{\mathsf{t}} & -2\tilde{\kappa} \\ -2\mathrm{a}^2\vec{\mathrm{v}}\cdot\vec{e}_1 & 2\mathrm{a}^2\vec{e}_1^{\mathsf{t}} & 0 \\ \vdots & \vdots & \vdots \\ -2\mathrm{a}^2\vec{\mathrm{v}}\cdot\vec{e}_{d-1} & 2\mathrm{a}^2\vec{e}_{d-1}^{\mathsf{t}} & 0 \\ \frac{1}{2}\tilde{\kappa}\left\|\vec{\mathrm{v}}\right\|^2 - \sigma_\lambda \mathrm{a}\mathrm{v}_{\vec{n}} & \sigma_\lambda \mathrm{a}\vec{n}^{\mathsf{t}} - \tilde{\kappa}\vec{\mathrm{v}}^{\mathsf{t}} & \tilde{\kappa} \end{pmatrix}.$$

Linear independence of the right eigenvectors is therefore assured and we have the following

Lemma 3.1. The Euler equations (3.3) are hyperbolic. In one space dimension they are even strictly hyperbolic.

3.2 Riemann Invariants

The eigenvalues and eigenvectors originating from the Euler equations (3.3) are either genuinely nonlinear or linearly degenerate. Specifically:

Lemma 3.2. The eigenvalues λ_1 and λ_s are genuinely nonlinear, $\lambda_2, \ldots, \lambda_{s-1}$ are linearly degenerate.

Proof. We consider first $\lambda_2 = \cdots = \lambda_{s-1} = \mathrm{v}_{\vec{n}}$:

$$\nabla_u \mathrm{v}_{\vec{n}} = \nabla_u \lambda_2 = \cdots = \nabla_u \lambda_{s-1} = \frac{1}{\rho}\left(-\mathrm{v}_{\vec{n}}, \quad \vec{n}^{\mathsf{t}}, \quad 0\right) \tag{3.7}$$

is clearly orthogonal on $r_2, \ldots, r_{s-1}$. For λ_s we have:

$$\nabla_u(\mathrm{v}_{\vec{n}} + \sigma_\lambda \mathrm{a}) = \frac{1}{\rho}\left(-\mathrm{v}_{\vec{n}}, \quad \vec{n}^{\mathsf{t}}, \quad 0\right) + \sigma_\lambda \frac{\kappa\tilde{\kappa}}{2\rho\mathrm{a}}\left(\frac{1}{2}\left\|\vec{\mathrm{v}}\right\|^2 - \frac{\mathrm{a}^2}{\kappa\tilde{\kappa}}, \quad -\vec{\mathrm{v}}^{\mathsf{t}}, \quad 1\right)$$

and

$$\sigma_\lambda \frac{\kappa\tilde{\kappa}}{2\rho\mathrm{a}}\left(\frac{1}{2}\left\|\vec{\mathrm{v}}\right\|^2 - \frac{\mathrm{a}^2}{\kappa\tilde{\kappa}}, \quad -\vec{\mathrm{v}}^{\mathsf{t}}, \quad 1\right)\begin{pmatrix} 1 \\ \vec{\mathrm{v}} + \sigma_\lambda \mathrm{a}\vec{n} \\ \mathrm{H} + \sigma_\lambda \mathrm{a}\mathrm{v}_{\vec{n}} \end{pmatrix}$$
$$+\frac{1}{\rho}\left(-\mathrm{v}_{\vec{n}}, \quad \vec{n}^{\mathsf{t}}, \quad 0\right)\begin{pmatrix} 1 \\ \vec{\mathrm{v}} + \sigma_\lambda \mathrm{a}\vec{n} \\ \mathrm{H} + \sigma_\lambda \mathrm{a}\mathrm{v}_{\vec{n}} \end{pmatrix} = \sigma_\lambda \frac{\mathrm{a}}{2\rho}(\kappa + 1) \neq 0.$$

The eigenvalue λ_1 takes the rôle of λ_s, if we reverse the ordering by changing the sign of σ_λ. □

The Riemann invariants corresponding to the eigenvalues are summarized in table 3.1 on page 83. Using the relations from equation (3.4b) they are easily verified.

$\lambda_1 = \mathrm{v}_{\vec{n}} - \sigma_\lambda \mathrm{a}$	$\Psi_1^{(1)} = \mathrm{v}_{\vec{n}} + \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}$ $\Psi_1^{(1+j)} = \vec{\mathrm{v}} \cdot \vec{e}_j$ for $j \in \{1, \ldots, d-1\}$ $\Psi_1^{(s-1)} = \mathrm{Z}$
$\lambda_2 = \mathrm{v}_{\vec{n}}$	$\Psi_2^{(1)} = \mathrm{v}_{\vec{n}}$ $\Psi_2^{(1+j)} = \vec{\mathrm{v}} \cdot \vec{e}_j$ for $j \in \{1, \ldots, d-1\}$ $\Psi_2^{(s-1)} = \mathrm{p}$
$\lambda_k = \mathrm{v}_{\vec{n}}$ for $k \in \{3, \ldots, s-1\}$	$\Psi_k^{(1)} = \mathrm{v}_{\vec{n}}$ $\Psi_k^{(1+j)} = \vec{\mathrm{v}} \cdot \vec{e}_j$ for $j \in \{1, \ldots, d-1\} \setminus \{k-2\}$ $\Psi_k^{(k-1)} = \rho$ $\Psi_k^{(s-1)} = \mathrm{p}$
$\lambda_s = \mathrm{v}_{\vec{n}} + \sigma_\lambda \mathrm{a}$	$\Psi_s^{(1)} = \mathrm{v}_{\vec{n}} - \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}$ $\Psi_s^{(1+j)} = \vec{\mathrm{v}} \cdot \vec{e}_j$ for $j \in \{1, \ldots, d-1\}$ $\Psi_s^{(s-1)} = \mathrm{Z}$

Table 3.1: Riemann invariants for the Euler equations

3.3 Jump Relations

Certain properties of the solution to the Riemann problem for the Euler equations can be deduced directly from the directed flux of equation (3.5). Since each of the conservative variables is conserved in its own right, we gather from the Rankine-Hugoniot jump condition (1.15) for a discontinuity aligned with a hyperplane orthogonal to a fixed unit vector $\vec{n} \in \mathbb{R}^d$ and moving at velocity $\vec{\nu} = \|\vec{\nu}\|\,\vec{n}$ and $\vec{e}_1, \ldots, \vec{e}_{d-1}$ orthogonal to $\vec{n}$:

$$\rho^{(l)}\big(\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|\big) = \rho^{(r)}\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\big) \tag{3.8a}$$

$$\rho^{(l)}\big(\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|\big)^2 + \mathrm{p}^{(l)} = \rho^{(r)}\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\big)^2 + \mathrm{p}^{(r)} \tag{3.8b}$$

$$\rho^{(l)}\big(\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|\big)\,\vec{\mathrm{v}}^{(l)} \cdot \vec{e}_j = \rho^{(r)}\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\big)\,\vec{\mathrm{v}}^{(r)} \cdot \vec{e}_j \tag{3.8c}$$

$$\rho^{(l)}\big(\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|\big)\,\mathrm{H}^{(l)} = \rho^{(r)}\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\big)\,\mathrm{H}^{(r)}. \tag{3.8d}$$

If $\mathrm{v}_{\vec{n}}^{(l)} = \|\vec{\nu}\|$, then by equation (3.8a) $\mathrm{v}_{\vec{n}}^{(r)} = \|\vec{\nu}\|$ and by (3.8b) $\mathrm{p}^{(l)} = \mathrm{p}^{(r)}$. This situation is a contact discontinuity. Since the velocity components orthogonal to the hyperplane equal the speed of the discontinuity, no fluid particle can pass.

If $\mathrm{v}_{\vec{n}}^{(l)} \neq \|\vec{\nu}\|$, then by equation (3.8a) $\mathrm{v}_{\vec{n}}^{(r)} \neq \|\vec{\nu}\|$ and by equation (3.8c) $\vec{\mathrm{v}}^{(l)} \cdot \vec{e}_j = \vec{\mathrm{v}}^{(r)} \cdot \vec{e}_j$. Similarly by equation (3.8d) $\mathrm{H}^{(l)} = \mathrm{H}^{(r)}$. Let us now assume that $\mathrm{v}_{\vec{n}}^{(l)} \neq \mathrm{v}_{\vec{n}}^{(r)}$, since otherwise there is really no discontinuity at all.

$$\mathrm{H}^{(l)} - \frac{1}{2}\big(\vec{\mathrm{v}}^{(l)} \cdot \vec{e}_j\big)^2 = \mathrm{H}^{(r)} - \frac{1}{2}(\vec{\mathrm{v}}^{(r)} \cdot \vec{e}_j)^2$$

gives the following definition of a^* in terms of either left or right state:

$$\frac{(\mathrm{a}^{(l)})^2}{\kappa - 1} + \frac{1}{2}\big(\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|\big)^2 = \frac{(\mathrm{a}^{(r)})^2}{\kappa - 1} + \frac{1}{2}\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\big)^2 =: \frac{1}{2}\,\frac{\kappa + 1}{\kappa - 1}(\mathrm{a}^*)^2. \tag{3.9}$$

Dividing equation (3.8b) by equation (3.8a)

$$\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\| + \frac{\mathrm{p}^{(l)}}{\rho^{(l)}\big(\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|\big)} = \mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\| + \frac{\mathrm{p}^{(r)}}{\rho^{(r)}\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\big)}$$

and therefore

$$\mathrm{v}_{\vec{n}}^{(l)} - \mathrm{v}_{\vec{n}}^{(r)} = \frac{\big(\mathrm{a}^{(r)}\big)^2}{\kappa\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\big)} - \frac{\big(\mathrm{a}^{(l)}\big)^2}{\kappa\big(\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|\big)}.$$

Use of equation (3.9) to eliminate $\big(\mathrm{a}^{(l)}\big)^2$ and $\big(\mathrm{a}^{(r)}\big)^2$ finally yields

$$\big(\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|\big)\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\big) = (\mathrm{a}^*)^2.$$

Introducing the Mach numbers (relative to $\|\vec{\nu}\|$)

$$\mathrm{Ma}_l := \frac{\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|}{\mathrm{a}^{(l)}} \qquad \mathrm{Ma}_r := \frac{\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|}{\mathrm{a}^{(r)}}$$
$$\mathrm{Ma}_l^* := \frac{\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|}{\mathrm{a}^*} \qquad \mathrm{Ma}_r^* := \frac{\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|}{\mathrm{a}^*}$$

we conclude

$$\mathrm{Ma}_l^* \mathrm{Ma}_r^* = 1 \tag{3.10a}$$

and from equation (3.9) (Ma refers to either Ma_l or Ma_r, Ma^* to Ma_l^*or Ma_r^*)

$$\big(\mathrm{Ma}^*\big)^2 = \frac{(\kappa+1)\mathrm{Ma}^2}{(\kappa-1)\mathrm{Ma}^2+2}. \tag{3.10b}$$

From equation (3.8a) we now infer the jump relation for the density

$$\frac{\rho^{(l)}}{\rho^{(r)}} = \frac{\mathrm{Ma}_r^*}{\mathrm{Ma}_l^*} = \big(\mathrm{Ma}_r^*\big)^2 = \frac{(\kappa+1)\big(\mathrm{Ma}_r\big)^2}{(\kappa-1)\big(\mathrm{Ma}_r\big)^2+2}, \tag{3.11a}$$

for the pressure from equation (3.8b)

$$\begin{aligned}
\mathrm{p}^{(l)} - \mathrm{p}^{(r)} &= \rho^{(l)}\big(\mathrm{v}_{\vec{n}}^{(l)} - \|\vec{\nu}\|\,\big)^2 - \rho^{(r)}\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\,\big)^2 \\
&= \rho^{(r)}\big(\mathrm{v}_{\vec{n}}^{(r)} - \|\vec{\nu}\|\,\big)\big(\mathrm{v}_{\vec{n}}^{(l)} - \mathrm{v}_{\vec{n}}^{(r)}\big) \qquad \text{by equation (3.8a)} \\
&= (\mathrm{a}^*)^2\rho^{(r)}\left(1 - \big(\mathrm{Ma}_r^*\big)^2\right)
\end{aligned}$$

and

$$\frac{\mathrm{p}^{(l)}}{\mathrm{p}^{(r)}} = 1 + \frac{2\kappa}{\kappa+1}\left(1 - \big(\mathrm{Ma}_r\big)^2\right). \tag{3.11b}$$

3.4 Numerical Flux Functions

For the numerical flux function $\boldsymbol{H}^{\mathrm{OS}}$ of Osher and Solomon the integration in equation (1.38) is simplified considerably by the fact that $s-2$ eigenvalues of the Jacobi matrix of the directed Euler flux in equation (3.5) are linearly degenerate. Furthermore, these $s-2$ eigenvalues are equal. The remaining two eigenvalues are genuinely nonlinear, one of them is greater and the other smaller than the degenerate ones.

The path components whose tangents are the linearly degenerate eigenvectors do not contain any sonic points, and the sum over the flux integrals

along these components reduces to the difference of the fluxes for just two intermediate states (equation (3.12d) contains the definition of $\mathrm{v}_{\vec{n}}^{(m)}$):

$$\sum_{k=2}^{s-1} \int_{\Gamma_k} \left(\frac{\partial \boldsymbol{F}\vec{n}}{\partial u}\bigg|_{\tilde{u}} \right)^+ d\tilde{u} = \begin{cases} \boldsymbol{F}(e_{s-1})\vec{n} - \boldsymbol{F}(s_2)\vec{n} & \text{if } \mathrm{v}_{\vec{n}}^{(m)} > 0 \\ 0 & \text{if } \mathrm{v}_{\vec{n}}^{(m)} \leqslant 0. \end{cases}$$

Therefore we need to determine two intermediate states $u^{(1)} = s_2$ and $u^{(2)} = e_{s-1}$. Using superscripts (l), (1), (2) and (r) to indicate affiliation to u_l, $u^{(1)}$, $u^{(2)}$ and u_r respectively, we infer for Γ_1 from the Riemann invariants listed in table 3.1 on page 83

$$\mathrm{v}_{\vec{n}}^{(1)} + \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}^{(1)} = \mathrm{v}_{\vec{n}}^{(l)} + \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}^{(l)} =: \Psi^{(l)} \tag{3.12a}$$

$$\vec{\mathrm{v}}^{(1)} \cdot \vec{e}_j = \vec{\mathrm{v}}^{(l)} \cdot \vec{e}_j \text{ for } j \in \{1, \ldots, d-1\} \tag{3.12b}$$

$$\mathrm{Z}^{(1)} = \mathrm{Z}^{(l)}, \tag{3.12c}$$

for the middle paths $\Gamma_2, \ldots, \Gamma_{s-1}$ (ρ and $\vec{\mathrm{v}} \cdot \vec{e}_j$ are not needed)

$$\mathrm{v}_{\vec{n}}^{(1)} = \mathrm{v}_{\vec{n}}^{(2)} =: \mathrm{v}_{\vec{n}}^{(m)} \tag{3.12d}$$

$$\mathrm{p}^{(1)} = \mathrm{p}^{(2)} \tag{3.12e}$$

and for the final path component Γ_s

$$\mathrm{v}_{\vec{n}}^{(2)} - \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}^{(2)} = \mathrm{v}_{\vec{n}}^{(r)} - \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}^{(r)} =: \Psi^{(r)} \tag{3.12f}$$

$$\vec{\mathrm{v}}^{(2)} \cdot \vec{e}_j = \vec{\mathrm{v}}^{(r)} \cdot \vec{e}_j \text{ for } j \in \{1, \ldots, d-1\} \tag{3.12g}$$

$$\mathrm{Z}^{(2)} = \mathrm{Z}^{(r)}. \tag{3.12h}$$

From the definition of Z we infer

$$\kappa^\kappa \exp(\mathrm{Z}) = \mathrm{p}^{-\tilde{\kappa}} \mathrm{a}^{2\kappa} \qquad \text{and} \qquad \mathrm{a} = \sqrt{\kappa \mathrm{p}^{\tilde{\kappa}/\kappa}} \exp\left(\frac{\mathrm{Z}}{2\kappa}\right).$$

Equations (3.12e), (3.12c) and (3.12h) now imply

$$\frac{\mathrm{a}^{(2)}}{\mathrm{a}^{(1)}} = \exp\left(\frac{\mathrm{Z}^{(2)} - \mathrm{Z}^{(1)}}{2\kappa}\right) = \exp\left(\frac{\mathrm{Z}^{(r)} - \mathrm{Z}^{(l)}}{2\kappa}\right) =: \alpha. \tag{3.12i}$$

Together with equation (3.12a), its counterpart (3.12f) and equation (3.12d) we obtain from equation (3.12i) a system of three linear equations

$$\begin{aligned} \mathrm{v}_{\vec{n}}^{(m)} + \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}^{(1)} &= \Psi^{(l)} \\ \mathrm{v}_{\vec{n}}^{(m)} - \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}^{(2)} &= \Psi^{(r)} \\ \mathrm{a}^{(2)} &= \alpha \mathrm{a}^{(1)} \end{aligned}$$

which gives us

$$\begin{aligned}
\mathrm{a}^{(1)} &= \sigma_\lambda \frac{\tilde{\kappa}}{2} \frac{\Psi^{(l)} - \Psi^{(r)}}{1+\alpha} \\
\mathrm{a}^{(2)} &= \sigma_\lambda \alpha \frac{\tilde{\kappa}}{2} \frac{\Psi^{(l)} - \Psi^{(r)}}{1+\alpha} \\
\mathrm{v}_{\vec{n}}^{(m)} &= \frac{\alpha \Psi^{(l)} + \Psi^{(r)}}{1+\alpha}.
\end{aligned}$$

This solution is physically meaningful, only if $\sigma_\lambda(\Psi^{(l)} - \Psi^{(r)}) > 0$. Having thus computed the velocity component parallel to $\vec{n}$ and the speeds of sound for the intermediate states, we need to decide whether Γ_1 and Γ_s contain sonic points. That is the case for Γ_1, if and only if

$$\left(\mathrm{v}_{\vec{n}}^{(l)} - \sigma_\lambda \mathrm{a}^{(l)}\right)\left(\mathrm{v}_{\vec{n}}^{(m)} - \sigma_\lambda \mathrm{a}^{(1)}\right) < 0$$

and for Γ_s, if and only if

$$\left(\mathrm{v}_{\vec{n}}^{(r)} + \sigma_\lambda \mathrm{a}^{(r)}\right)\left(\mathrm{v}_{\vec{n}}^{(m)} + \sigma_\lambda \mathrm{a}^{(2)}\right) < 0.$$

A sonic point n_k is in either case characterized by $\lambda_k(n_k) = 0$ ($k \in \{1, s\}$). We denote by v_* the velocity component parallel to $\vec{n}$ and by a_* the speed of sound for the sonic point. A superscript (l) refers to Γ_1, (r) to Γ_s. Together with equations (3.12a) and (3.12f) we obtain the following linear equations:

$$\begin{aligned}
0 &= \mathrm{v}_*^{(l)} - \sigma_\lambda \mathrm{a}_*^{(l)} & \qquad 0 &= \mathrm{v}_*^{(r)} + \sigma_\lambda \mathrm{a}_*^{(r)} \\
\Psi^{(l)} &= \mathrm{v}_*^{(l)} + \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}_*^{(l)} & \qquad \Psi^{(r)} &= \mathrm{v}_*^{(r)} - \sigma_\lambda \frac{2}{\tilde{\kappa}} \mathrm{a}_*^{(r)}.
\end{aligned}$$

They admit the solutions (${\sigma_\lambda}^2 = 1$)

$$\begin{aligned}
\mathrm{v}_*^{(l)} &= \frac{\kappa - 1}{\kappa + 1} \Psi^{(l)} & \qquad \mathrm{v}_*^{(r)} &= \frac{\kappa - 1}{\kappa + 1} \Psi^{(r)} \\
\mathrm{a}_*^{(l)} &= \sigma_\lambda \frac{\kappa - 1}{\kappa + 1} \Psi^{(l)} & \qquad \mathrm{a}_*^{(r)} &= -\sigma_\lambda \frac{\kappa - 1}{\kappa + 1} \Psi^{(r)}.
\end{aligned}$$

By equation (3.12c) the entropy is constant on Γ_1, by its companion (3.12h) on Γ_s. The velocity components orthogonal to $\vec{n}$ are similarly constant on Γ_1 by equation (3.12b) and by equation (3.12g) on Γ_s. At this stage we have thus at our disposal speed of sound, transport velocity and entropy for any intermediate or sonic state. These variables $(\mathrm{a}, \vec{\mathrm{v}}, \mathrm{Z})$ are called **characteristic variables.** In order to evaluate equation (3.5) for a certain state we

compute density, pressure and enthalpy as follows:

$$\rho = \exp\left(\frac{\ln(\mathrm{a}^2/\kappa) - \mathrm{Z}}{\tilde{\kappa}}\right)$$
$$\mathrm{p} = \frac{\mathrm{a}^2\rho}{\kappa}$$
$$\mathrm{H} = \frac{\mathrm{a}^2}{\tilde{\kappa}} + \frac{1}{2}\left\|\vec{\mathrm{v}}\right\|^2.$$

We may optimize the computation of $\left\|\vec{\mathrm{v}}\right\|^2$, if we observe that for two vectors $\vec{\nu}$ and $\vec{\nu}'$ differing only in their components parallel to $\vec{n}$

$$\left\|\vec{\nu}'\right\|^2 = \left\|\vec{\nu}\right\|^2 - (\vec{\nu}_{\vec{n}})^2 + (\vec{\nu}'_{\vec{n}})^2.$$

Evaluation of the integral in equation (1.38) is now straightforward. The thesis of Spekreijse [Spe87] and the book by Toro [Tor97] contain tables listing the explicit expressions for the flux function of Osher and Solomon in the case of the Euler equations. We do not reproduce those tables here.

3.5 Boundary Conditions

We use the following two kinds of boundary conditions:

- fixed wall and
- moving shock.

The fixed wall boundary condition is characterized by the fact that no particle can leave or enter across this part of the boundary. Consequently the velocity component orthogonal to the boundary vanishes. Enforcing $\mathrm{v}_{\vec{n}} = 0$ in equation (3.5) gives:

$$\boldsymbol{B} = \begin{pmatrix} 0 \\ \mathrm{p}\vec{n} \\ 0 \end{pmatrix}. \tag{3.13}$$

The moving shock boundary condition simulates the movement of a supersonic shock along the boundary. The shock is represented by a hyperplane with normal $\vec{n}$ separating two states u_l and u_r, such that $\vec{n}$ points from the region of u_l to that of u_r. One has to specify the normal $\vec{n}$ to the hyperplane, the velocity $\vec{\nu} = (\vec{\nu}\cdot\vec{n})\vec{n}$ at which the hyperplane shall move and the state to the right of the hyperplane u_r. The state u_l to the left of the hyperplane can then be computed in terms of the **primitive variables** $(\rho, \vec{\mathrm{v}}, \mathrm{p})$ using equations (3.11). The outer state is either u_l or u_r depending on t and $\vec{x}$, and an approximate Riemann problem can be solved in the usual fashion.

Chapter 4

Collocation Schemes

In this chapter we sketch some of our early attempts at developing a collocation method for unstructured grids. These were based on the classical Lax-Friedrichs scheme and isotropic regularization of the solution with a Laplace term. We found these too dissipative and considered anisotropic regularization instead. One such method that works quite well on Cartesian grids uses a digital filtering strategy at local extrema. In the unstructured case using a recovery procedure as a means of analyzing the local data and the downhill transport mechanism immanent in numerical flux functions proved a cheap and superior alternative. We furnish some numerical examples to demonstrate the capabilities of such a scheme.

4.1 The Lax-Friedrichs Scheme

Cartesian Grids

The second order central finite difference scheme on a grid of constant mesh size h in one space dimension

$$\mathrm{TS}(u,t,\Delta t)(x_k) = u(t,x_k) - \frac{\Delta t}{2h}\left[\boldsymbol{F}(u(t,x_{k+1})) - \boldsymbol{F}(u(t,x_{k-1}))\right] \qquad (4.1)$$

is known to be unstable. This is easily seen by a von Neumann error analysis considering a linear flux function.

The von Neumann error analysis considers the evolution of a spatially periodic disturbation of the data for a linear equation, by looking at each term of its Fourier expansion separately. If u has the spatial period L, then

its Fourier expansion can at least formally be written as

$$u(j\Delta t, kh) = \sum_{J\in\mathbb{Z}} \sum_{K\in\mathbb{Z}} \alpha_{J,K} \xi_J^j \exp\left(\frac{2\pi i K k h}{L}\right)$$

where $\alpha_{J,K}$ are the Fourier coefficients and $\xi_J \in \mathbb{C}$ is the expansion in time. Suppressing the indices a single Fourier term now has the form

$$\xi^j \exp(i\alpha k) \text{ with } \alpha \in \mathbb{R}. \tag{4.2}$$

With the linear flux $\boldsymbol{F}(u) = au$ ($a \in \mathbb{R}$ fixed) we infer

$$\xi^{j+1} \exp(i\alpha k) = \xi^j \exp(i\alpha k) - \frac{a\Delta t}{2h}\left[\xi^j \exp(i\alpha(k+1)) - \xi^j \exp(i\alpha(k-1))\right]$$

and hence

$$\begin{aligned} \xi &= 1 - \frac{a\Delta t}{2h}[\exp(i\alpha) - \exp(-i\alpha)] \\ &= 1 - \frac{ia\Delta t}{h}\sin(\alpha) \\ |\xi|^2 &= 1 + \left(\frac{a\Delta t}{h}\sin(\alpha)\right)^2 \geqslant 1. \end{aligned}$$

The amplitude of each error mode therefore increases exponentially in time which is precisely the alleged instability. The classic Lax-Friedrichs scheme removes this instability by replacing the term $u(t, x_k)$ on the right hand side of equation (4.1) with the average of $u(t, x_{k+1})$ and $u(t, x_{k-1})$:

$$\begin{aligned} &\mathrm{TS}(u, t, \Delta t)(x_k) \\ &\quad = \frac{u(t, x_{k+1}) + u(t, x_{k-1})}{2} - \frac{\Delta t}{2h}\left[\boldsymbol{F}(u(t, x_{k+1})) - \boldsymbol{F}(u(t, x_{k-1}))\right] \\ &\quad = u(t, x_k) - \frac{\Delta t}{2h}\left[\left(\boldsymbol{F}(u(t, x_{k+1})) + \frac{h}{\Delta t}(u(t, x_k) - u(t, x_{k+1}))\right)\right. \\ &\qquad\qquad \left. - \left(\boldsymbol{F}(u(t, x_{k-1})) + \frac{h}{\Delta t}(u(t, x_{k-1}) - u(t, x_k))\right)\right]. \end{aligned} \tag{4.3}$$

A von Neumann error analysis for the linearized version of equation (4.3) yields

$$\begin{aligned} \xi^{j+1} \exp(i\alpha k) = \xi^j &\frac{\exp(i\alpha(k+1)) + \exp(i\alpha(k-1))}{2} \\ &- \frac{a\Delta t}{2h}\left[\xi^j \exp(i\alpha(k+1)) - \xi^j \exp(i\alpha(k-1))\right] \end{aligned}$$

and thus

$$\begin{aligned}\xi &= \cos(\alpha) - \frac{ia\Delta t}{h}\sin(\alpha)\\ |\xi|^2 &= \cos^2(\alpha) + \left[\frac{a\Delta t}{h}\sin(\alpha)\right]^2\\ &= 1 + \left[\left(\frac{a\Delta t}{h}\right)^2 - 1\right]\sin^2(\alpha)\end{aligned}$$

which is less than one, if and only if

$$\Delta t < \frac{h}{|a|}.$$

From this point of view the CFL time step constraint [CFL28] appears as a condition ensuring exponential decay in time of spatially periodic error modes. If we substitute $h/\Delta t = L_{\boldsymbol{F}}/\mathrm{CFL}$ in equation (4.3) and then drop this occurence of CFL, we get back to the numerical Lax-Friedrichs flux function:

$$\begin{aligned}&\boldsymbol{H}^{\mathrm{LF}}(u(t,x_k),u(t,x_{k+1}),+1) + \boldsymbol{H}^{\mathrm{LF}}(u(t,x_k),u(t,x_{k-1}),-1)\\ &\quad\approx \frac{1}{2}\left(\boldsymbol{F}(u(t,x_k)) + \boldsymbol{F}(u(t,x_{k+1})) + \frac{h}{\Delta t}(u(t,x_k) - u(t,x_{k+1}))\right)\\ &\qquad - \frac{1}{2}\left(\boldsymbol{F}(u(t,x_{k-1})) + \boldsymbol{F}(u(t,x_k)) + \frac{h}{\Delta t}(u(t,x_{k-1}) - u(t,x_k))\right)\end{aligned}$$

This allows us to interprete the Lax-Friedrichs scheme as a finite volume scheme with piecewise constant reconstruction on cells of length h centered about each grid point x_k. The Lax-Friedrichs scheme of equation (4.3) has a local truncation error

$$L_{\Delta t}(t,x) := \frac{1}{\Delta t}[\mathrm{TS}(u,t,\Delta t)(x) - u(t+\Delta t,x)]$$

that is linear, i.e. assuming smoothness and replacing u with its Taylor expansion we obtain for the Lax-Friedrichs scheme $L_{\Delta t} = \mathcal{O}(\Delta t)$. However, for the modified equation

$$\frac{\partial}{\partial t}u + a\frac{\partial}{\partial x}u = b\frac{\partial^2}{\partial x^2}u \tag{4.4a}$$

with

$$b := \frac{\Delta t}{2}\left(\frac{h^2}{(\Delta t)^2} - a^2\right) \tag{4.4b}$$

the local truncation error is quadratic and the CFL restriction $\Delta t < h/|a|$ implies that we are not simulating an ill-posed backward heat transport equation. In smooth regions of the flow we obtain from equations (4.4)

$$\frac{\partial^2}{\partial t^2}u - a^2\frac{\partial^2}{\partial x^2}u = b\left(\frac{\partial^3}{\partial x^2\partial t}u - a\frac{\partial^3}{\partial x^3}u\right)$$

and

$$\begin{aligned} L_{\Delta t}(t,x) = &\left(\frac{h^2}{2\Delta t} - \frac{a^2\Delta t}{2} - b\right)\frac{\partial^2}{\partial x^2}u \\ &+ \frac{b\Delta t}{2}\left(\frac{\partial^3}{\partial x^2\partial t}u - a\frac{\partial^3}{\partial x^3}u\right) + \mathcal{O}((\Delta t)^2) \end{aligned}$$

With the choice of b of (4.4b) we have indeed $L_{\Delta t} = \mathcal{O}((\Delta t)^2)$. We now seek to stabilize a central scheme on smoothly, but otherwise arbitrarily scattered collocation points by adding diffusion in the spirit of the classical Lax-Friedrich scheme.

Unstructured Grids

We noted in [AFHS] that the amount of artificial diffusion in equations (4.4) depends nonlinearly on the local scale h. The time step size is determined essentially by the smallest h in the grid and leads to unacceptable high diffusion in regions where the grid is coarse. Furthermore, if the CFL number is halved, the diffusion per time step approximately doubles:

$$b = \frac{\Delta t}{2}\left(\frac{1}{\mathrm{CFL}^2} - 1\right)a^2.$$

To render diffusion independent of the CFL number we use a discrete dissipative model proportional to the chosen CFL number (and drop the -1):

$$\mathrm{CFL}\frac{\Delta t}{2}\frac{h^2}{(\Delta t)^2}.$$

We still need to define a "typical" scale h for each stencil. To this end we introduce the local scale

$$h_{\mathcal{S}_\Sigma} := \frac{\operatorname{diam}\Sigma_{\mathcal{S}}}{2}$$

and the global scale

$$h_{\mathcal{G}} := \min_{\Sigma\in\mathcal{G}} h_\Sigma.$$

The dissipative model can now be stated as

$$\tilde{b} := \mathrm{CFL}\frac{\Delta t}{2}\frac{h_{\mathcal{S}_\Sigma} h_{\mathcal{G}}}{(\Delta t)^2}.$$

In our calculations we also replaced the division by two with a division by four. Thus the modified diffusive equation we simulated was

$$\frac{\partial}{\partial t}u + \mathrm{div}\left(\boldsymbol{F}\circ u - \mathrm{CFL}\frac{\Delta t}{4}\frac{h_{\mathcal{S}_\Sigma} h_{\mathcal{G}}}{(\Delta t)^2}\nabla u\right) = 0.$$

The evaluation points were chosen to be all collocation points in a central stencil and two evaluation directions per evaluation point were used, i.e. we incorporated the full vector valued flux as obtained by computing the flux function for the collocation point value of u, no reconstruction procedure was applied.

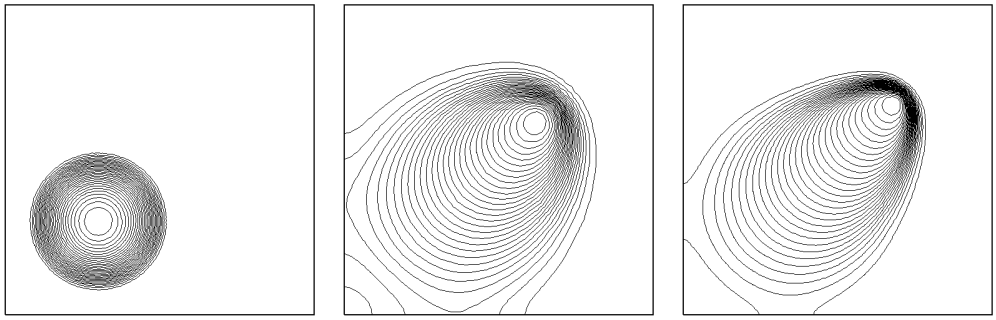

Figure 4.1: Solutions to the Burgers equation generated with the artificial diffusion model on the unit square $[0,1]^2 \subset \mathbb{R}^2$. Left initial condition, middle 2013 points and right 7889 points. The simulated equation is in either case $\boldsymbol{F}(u) = u^2(1,1)^{\mathsf{t}}$ and the final time is $T = 0.4$. The initial condition consists in a Gauss bell shaped function on a circle of radius 0.25 centered about $(0.3, 0.3)$ with height 1. Riemann problems with outer state $u = 0$ are solved on the boundary. The upper parts of the bell are advected faster than the lower parts and start to overtake at some point. At this stage a discontinuity is formed. On finer grids the dissipative effect decreases, but the artificial diffusion necessary for stabilizing the scheme makes the solutions too dissipative.

As an alternative we investigated a filtering strategy in the spirit of [ELS89]. Since numerical instability manifests itself in the generation of noise, i.e. extrema, we checked for local extrema after each time step and applied a local smoothing procedure. If a local extremum had been generated

at a data location Σ, we computed the average of the data functional values on a stencil consisting of Σ and its neighbours

$$\operatorname{avg}_{\mathcal{S}} u := \frac{1}{\operatorname{card} \mathcal{S}} \sum_{\Theta \in \mathcal{S}} \delta_{\Sigma} u$$

and then **scaled** the data functional values towards their average by replacing the stored value of $\delta_{\Theta} u$ with

$$\alpha \delta_{\Theta} u + (1 - \alpha) \operatorname{avg}_{\mathcal{S}} u$$

for each $\Theta \in \mathcal{S}$. Choosing $\alpha \approx 0.9$ proved sufficient to stabilize the scheme, but gave a rather noisy solution, smaller values would diminish the noise, but increase diffusive smearing.

4.2 Upwinding Schemes

Anisotropic data dependent diffusion can also very efficiently be incorporated into the scheme via numerical flux functions. It proved superior to our attempts of isotropic regularization without explicit reconstruction step. It should be noted that anisotropic regularization necessarily involves some kind of data analysis and that linear or higher order reconstruction can be regarded as a means of analyzing the local data. Let us illustrate this idea very loosely: The integral formulation of a conservation law with a diffusive term reads

$$\frac{d}{dt} \int_{\Sigma} u \, dV = - \int_{\partial \Sigma} (\boldsymbol{F} \circ u - b \nabla u) \vec{n} \, do. \tag{4.5}$$

If we replace $\nabla u \vec{n}$ with a suitable difference quotient

$$\nabla u \vec{n} \approx \frac{u(\vec{x}_{\Sigma}) - u(\vec{x}'_{\Sigma})}{\|\vec{x}_{\Sigma} - \vec{x}'_{\Sigma}\|}$$

where $\vec{x}_{\Sigma}$ and $\vec{x}'_{\Sigma}$ represent two points inside and outside the cell Σ respectively on a line orthogonal to $\partial \Sigma$ with

$$\|\vec{x}_{\Sigma} - \vec{x}'_{\Sigma}\| \approx h,$$

we can interprete a numerical flux function $\boldsymbol{H}$ as an approximation to the diffusive flux of equation (4.5) with b suitably chosen. The numerical flux (u_i and u_o are again the inner and outer limit at the cell boundary)

$$\boldsymbol{H}(u_i, u_o, \vec{n}) = \frac{\boldsymbol{F}(u_i) + \boldsymbol{F}(u_o)}{2} \vec{n} + A \frac{u_i - u_o}{2}$$

with a suitable upwinding term $A \in [0, L_{\boldsymbol{F}}]$ approximates the integrand in (4.5), if we identify $(\boldsymbol{F} \circ u)\vec{n}$ with $(\boldsymbol{F}(u_i) + \boldsymbol{F}(u_o))\vec{n}/2$ and observe

$$\begin{aligned} A\frac{u_i - u_o}{2} &= A\frac{\|\vec{x}_\Sigma - \vec{x}'_\Sigma\|}{2} \frac{u_i - u_o}{u(\vec{x}_\Sigma) - u(\vec{x}'_\Sigma)} \frac{u(\vec{x}_\Sigma) - u(\vec{x}'_\Sigma)}{\|\vec{x}_\Sigma - \vec{x}'_\Sigma\|} \\ &\approx \underbrace{A\frac{\|\vec{x}_\Sigma - \vec{x}'_\Sigma\|}{2} \frac{u_i - u_o}{u(\vec{x}_\Sigma) - u(\vec{x}'_\Sigma)}}_{=:b} \nabla u\vec{n}. \end{aligned}$$

The reconstruction process of degree q will generally ensure that

$$\frac{u_i - u_o}{u(\vec{x}_\Sigma) - u(\vec{x}'_\Sigma)} = \begin{cases} 1 + \mathcal{O}(h) \geqslant 0 & \text{at discontinuities} \\ \mathcal{O}(h^q) & \text{in smooth regions of the flow.} \end{cases}$$

The artificial diffusion contributed by approximating the integrand in equation (4.5) with a numerical flux function therefore decreases very rapidly in smooth regions of the flow. Near discontinuities it constitutes the required downhill transport necessary to ensure proper physical transport. The argumentation is of course flawed, if $u(\vec{x}_\Sigma) = u(\vec{x}'_\Sigma)$.

The results of the numerical flux computation are used as input data for approximating the divergence functional. In this context we use only one evaluation direction per evaluation point.

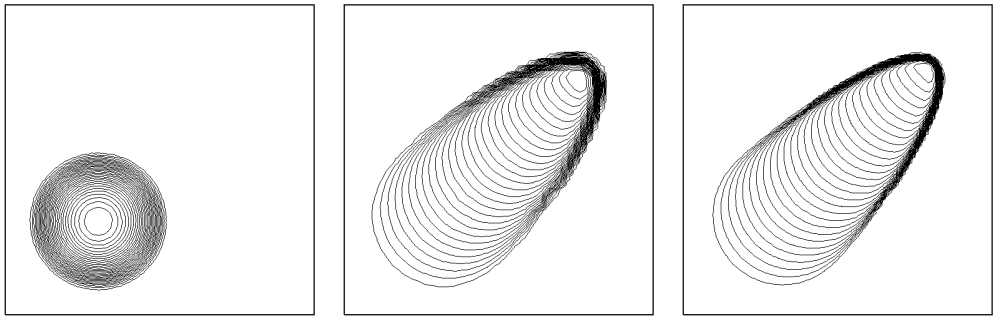

Figure 4.2: Solutions to the Burgers equation generated with the upwinding method. Middle 2013 points and right 7889 points. The setup is identical to figure 4.1.

Choice of Data Locations

Because of the diffusive effect immanent in the solution to local Riemann problems, too, positive divergence formulae have to be used: the numerical fluxes directed away from a data location should have a positive weight in the

formula for approximating the flux divergence at that point. Otherwise one would locally simulate the behaviour of a "backward heat transport equation" and suffer numerical instability. We have still some freedom in the exact choice of the collocation functionals which we now use to obtain such positive formulae:

> As collocation functionals we choose the average of the pointwise collocation at each evaluation point in the divergence stencil.

The positivity of a weight w is checked in the following way: Letting $\vec{x}$ denote an evaluation point in the divergence stencil, $\vec{n}$ the evaluation direction for this point and $\vec{x}_0$ the average of the evaluation points we demand

$$\vec{n} \cdot (\vec{x} - \vec{x}_0) \operatorname{sign}(w) \geqslant 0. \tag{4.6}$$

Up to linear reconstruction this choice of the collocation functionals is equivalent to collocation about the barycentre of the divergence stencil.

4.3 Numerical Experiments

Finally we present numerical results for the Euler equations obtained with the upwind collocation method.

2D Shock Tube for the Euler Equations

In this section we present results for the two dimensional shock tube problem with parameters suggested by Lax and Sod. The shock tube has length one and initially a diaphragm at $x = 0.5$ separating gas in different states to its left and right. At time $t_0 = 0$ the diaphragm is removed and the gas starts to mix.

	Lax ($T = 0.1445$)		Sod ($T = 0.18$)	
	left	right	left	right
ρ	0.445	0.5	1	0.125
$\vec{\mathrm{v}}$	$0.698\vec{e}_0$	0	0	0
p	3.528	0.571	1	0.1

Table 4.1: Initial configuration at $t_0 = 0$ for Lax and Sod shock tube problems.

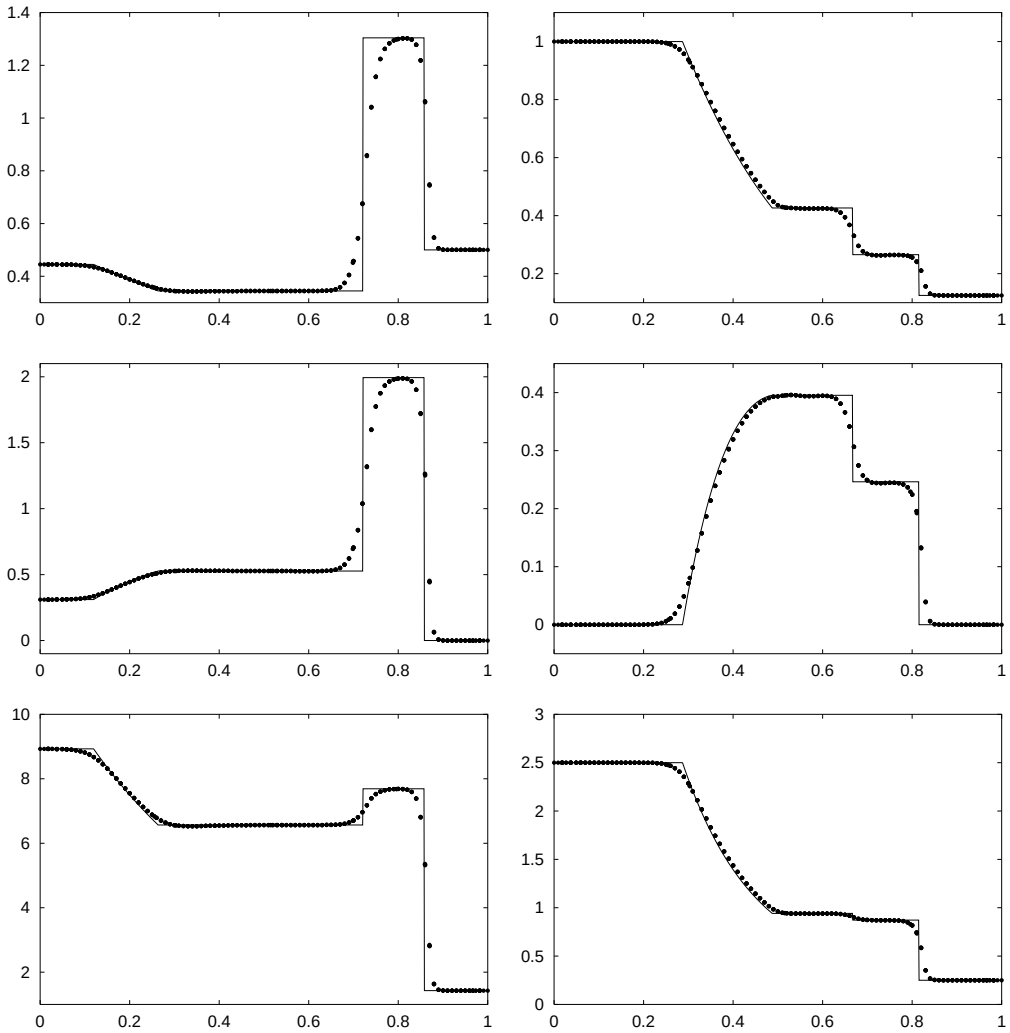

Figure 4.3: Solutions to the shock tube problem plotted along the x-axis. Left part Lax configuration, right part Sod configuration. From top to bottom: density ρ, density of momentum $\rho\vec{\mathrm{v}} \cdot \vec{e}_0$ and density of energy ρE. The average time step for the Lax problem was $\Delta t = 0.0013$ and for the Sod problem $\Delta t = 0.003$ resulting from a CFL number of 0.7 in both cases. The solid line indicates the exact solution.

The computational grid we used covered the region $[0,1]\times[-0.05,0.05] \subset \mathbb{R}^2$. It consisted of 1309 points which corresponds to roughly 100 points spread along the x-axis. We show the results in figure 4.3. The jump of the density for the contact disontinuity in the Lax case is smeared over some ten cells, in the Sod case over six cells. This is comparable to the results presented by Botta [Bot95] for the van Leer limiter.

On the top and bottom part of the grids margin we applied the fixed wall boundary condition and supplied the constant states from the initial condition on the left and right part of the boundary. No particular measures were taken to keep the flow inside the tube one dimensional. We show the iso-lines of the density in figures 4.4 and 4.5.

Figure 4.4: Iso-lines of the density for the Lax shock tube problem.

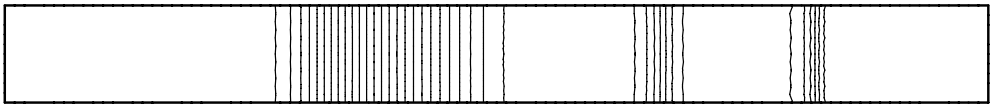

Figure 4.5: Iso-lines of the density for the Sod shock tube problem.

Double Mach Reflection

The double Mach reflection of a strong shock can be realized experimentally by driving a shock down a tube that contains a wedge. Our configuration follows the one presented in [WC84]: a fast planar shock making an angle of $60°$ with the x-axis meets a fixed wall. This setup is commonly regarded as a challenging test case.

The computational domain we used was trapezoidal with corners at $(0,0)$, $(3,0)$, $(3,3/4)$ and $(\sqrt{3}/4,3/4)$. The fixed wall lies at the bottom of the computational domain and stretches from $x = 1/6$ to the right. The undisturbed resting gas to the right of the shock has density $\rho = 1.4$ and pressure $\mathrm{p} = 1$. Its speed of sound therefore is $\mathrm{a} = 1$. The shock moves at Mach 10, its velocity vector is $(5\sqrt{3},-5)$. Initially $(t = 0)$ the shock front touches the leftmost point of the wall.

All parts of the domains boundary that are not formed by the wall are treated with the moving shock boundary condition. Based on the prescribed values for the shocks propagation speed and the state of the gas ahead of the

shock the state of the gas behind the shock can be computed using the jump relations for the Euler equations. On the boundary the appropriate one of the two states is used as outer state and a Riemann problem can be solved in the usual way. In figure 4.6 we show the iso-lines of the computed solution for three subsequently refined grids.

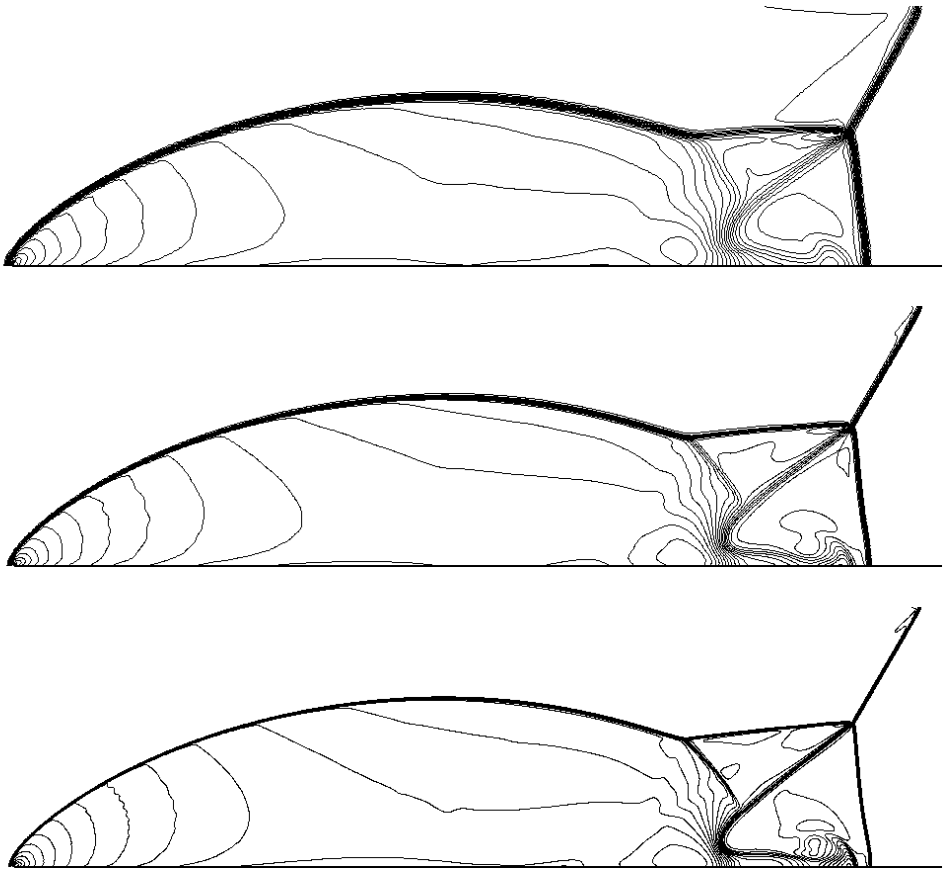

Figure 4.6: Iso-lines of density for the double Mach reflection at $T = 0.2$. Top 25831 points, middle 102629 points and bottom 409037 points. The bottom line of each picture indicates the position of the wall.

Appendix A

Measure of an equilateral d-dimensional simplex

Lemma A.1. The Jordan measure (volume) of an equilateral simplex Σ of diameter a in d dimensions is

$$|\Sigma| = \frac{1}{d!\sqrt{d+1}} \left(\frac{a}{\sqrt{2}}\right)^d.$$

Proof. The simplex $F_d(l) := \left\{(x_1, \ldots, x_d) \in \mathbb{R}^d : x_k \geqslant 0 \text{ and } \sum_{k=1}^d x_k \leqslant l\right\}$ has volume

$$|F_d(l)| = \frac{l^d}{d!}$$

as is easily seen by induction: For $d = 1$ one has $|F_1(l)| = l$ and for $d \geqslant 2$ $|F_d(l)| = |F_{d-1}(l)|\, l/d$ by the *"base times height divided by d"* rule. All edges of the sub-simplex $\Sigma_{d-1}(l)$ of $F_d(l)$ formed by all vertices save $\vec{v} = 0$ have length $a = l\sqrt{2}$ and this is also the diameter of $\Sigma_{d-1}(l)$. The height of $F_d(l)$ on this sub simplex is formed by the vector $(1, \ldots, 1)l/d \in \mathbb{R}^d$ which has length $l/\sqrt{d}$. Now

$$|\Sigma_{d-1}(l)| = \frac{d\,|F_d(l)|}{l/\sqrt{d}} = \frac{l^{d-1}}{(d-1)!\sqrt{d}}$$

and

$$|\Sigma_d(l)| = \frac{l^d}{d!\sqrt{d+1}} = \frac{a^d}{d!\sqrt{2^d(d+1)}}.$$

□

Lemma A.2. For the simplex Σ of the preceding lemma we have

$$\frac{|\Sigma|}{|\partial\Sigma|\,\mathrm{diam}(\Sigma)} = \frac{1}{d(d+1)}\sqrt{\frac{d}{2(d+1)}}.$$

Proof. $\Sigma = \Sigma_d(l)$ has $d+1$ faces $\Sigma_{d-1}(l)$ and therefore

$$\frac{|\Sigma|}{|\partial\Sigma|\,\mathrm{diam}(\Sigma)} = \frac{\frac{1}{d!\sqrt{d+1}}\left(\frac{a}{\sqrt{2}}\right)^d}{\frac{d+1}{(d-1)!\sqrt{d}}\left(\frac{a}{\sqrt{2}}\right)^{d-1} a} = \frac{1}{d(d+1)}\sqrt{\frac{d}{2(d+1)}}.$$

□

Appendix B

List of Symbols

$\overline{A}$	closure of the set A
$\overline{\mathrm{co}}\, A$	closed convex hull of A
$\lvert x\rvert$	absolute value of $x \in \mathbb{R}$
$\lvert X\rvert$	absolute value of the diagonalizable matrix $X \in \mathbb{R}^{n\times n}$
$\lvert\Sigma\rvert$	Jordan measure (volume) of $\Sigma \subset \mathbb{R}^d$
$(A \to B)$, B^A	mappings from A to B
$\lVert\cdot\rVert$	Euclidean norm in $\mathbb{R}^d$
$\lVert\cdot\rVert_G$	Euclidean G-norm in $\mathbb{R}^n$
$\lVert\cdot\rVert_\infty$	maximum norm in $\mathbb{R}^n$
$\lVert\cdot\rVert_\infty$	supremum norm in BL^∞
$\lVert\cdot\rVert_{\infty,\Omega}$	supremum norm in $BL^\infty(\Omega \to \mathbb{R})$
$\lVert\cdot\rVert_\infty$	supremum norm in Π
$\lVert\cdot\rVert_{\infty,\Omega}$	supremum norm in $\Pi(\Omega \to \mathbb{R})$
$\langle\cdot;\cdot\rangle_S$	Riemann form in the tangent space of S
$\otimes$	tensor product in $\mathbb{R}^d$, $\vec{a} \otimes \vec{b} = \vec{a}\vec{b}^{\mathrm{t}}$
$\widetilde{\nabla\cdot}$	numerical divergence operator
$\widetilde{\nabla}$	numerical gradient operator
∇_u	gradient with respect to conserved quantities
a	speed of sound
$\mathcal{B}$	basis of polynomial functions
BL^∞	space of bounded Lebesgue measurable functions with the supremum norm
barycentre	barycentre of a set in $\mathbb{R}^d$
CFL	CFL number
C^q	space of q times continuously differentiable functions
Δ	Laplace operator
Δt	time step size

δ	linear data functional
δ_Σ	linear functional for the data location Σ
$\delta_\Sigma \vec{x}$	barycentre of the data location Σ
$\delta^\perp$	kernel of δ
Div	divergence operator for the space and time variables
d	space dimension
diag$(\cdots)$	diagonal matrix of the arguments
diam	diameter of a set
div	divergence operator for the space variables
do	measure for integration over a surface
dV	measure for integration over a volume
ε	small positive number, typically 10^{-15}
E	specific energy
e_k	intermediate state
$\vec{e}_0, \ldots, \vec{e}_{d-1}$	orthonormal basis of $\mathbb{R}^d$, frequently $\vec{e}_0 = \vec{n}$
ext	extremum, either maximum or minimum
Φ	linear feature functional
ϕ	vector of feature weights
ϕ	test function
$\mathcal{F}$	set of linear functionals
$\mathcal{F}^\perp$	quotient space with respect to $\ker \mathcal{F}$
$\boldsymbol{F}(\cdot)$	flux function
$\boldsymbol{F}(\cdot,\cdot,\cdot)$	Riemann solver
Γ	path in the state space S
$\mathcal{G}$	computational grid
G	Gram matrix (symmetric and positive definite)
$\boldsymbol{H}$	approximate Riemann solver aka. numerical flux function
h	local scale
H	specific enthalpy
$\boldsymbol{I}$	identity matrix
id	identity mapping
int	open interior of a set
$\mathbf{J}$	Jacobi matrix
κ	adiabatic constant
$\tilde{\kappa}$	adiabatic constant minus one
ker	kernel of linear functionals
$\ker \mathcal{F}$	intersection of kernels of the functionals in $\mathcal{F}$
$\boldsymbol{\Lambda}$	vector of datafunctionals for a scheme
λ	eigenvalue
L^p	space of Lebesgue measurable functions with finite p-norm

L^∞	space of essentially bounded Lebesgue measurable functions with the essential supremum norm
$L_\bullet$	Lipschitz constant of a function
l^{t}	left eigenvector in $\mathbb{R}^s$
$\vec{n}$	unit vector in $\mathbb{R}^d$
$\vec{n}\,do$	measure for integration over a surface with respect to the outer normal
n_k	sonic state
Ω	a compact set in $\mathbb{R}^d$ with $\overline{\operatorname{int}\Omega} = \Omega$
Ω	union of data locations in a grid
ω_Σ	oscillation indicator for the data location Σ
$\mathcal{O}$	asymptotic order: $f = \mathcal{O}(g)$ means that $\limsup_{x\to x_0} \lvert f(x)/g(x)\rvert$ is finite
Π	polynomial function space
Π^q	polynomial function space up to degree q
π	permutation of the numbers $\{0,\ldots,d\}$
π	polynomial taking real values
$\vec{\pi}$	vector valued polynomial
Ψ	Riemann invariant
P	linear projection
p	pressure
ρ	density of gas
Q	orthogonal matrix, $Q = Q^{-\mathsf{t}}$
Q	numerical integration
$\mathcal{R}$	reconstruction operator (typically nonlinear)
r	right eigenvector in $\mathbb{R}^s$
Σ	data location
$\Sigma_{\mathcal{S}}$	union of data locations in a stencil
σ_λ	arrangement of eigenvalues, $+1$ for increasing order, -1 for decreasing order
$\mathcal{S}$	stencil
$\mathcal{S}_\Sigma$	stencil without the data location Σ
S	state space
s	state space dimension
sign	$+1$ for nonnegative numbers, -1 for negative numbers
s_k	intermediate state
supp	closed hull of the set where a function does not vanish
$T_{\vec{x}_0} u$	Taylor polynomial at $\vec{x}_0$ of u (typically of degree q)
T	final time in numerical experiments
TS	numerical time stepping operator

$^{\mathsf{t}}$	transposition, formal swapping of rows and columns
u	conserved quantities, components are referred to with superscripts
u_0	initial state density distribution
Θ	data location
$\vec{\mathrm{v}}$	velocity vector for gas flow
$\mathrm{v}_{\vec{n}}$	salar product $\vec{\mathrm{v}} \cdot \vec{n}$
Z	thermodynamic entropy

Bibliography

[AFHS] Arne Ahrend, Jiri Fürst, Daniel Hempel, and Thomas Sonar, *On meshless collocation approximations of conservation laws*, submitted.

[AHS99] Arne Ahrend, Daniel Hempel, and Thomas Sonar, *Preliminary results on hp-cloud methods for conservation laws*, ZAMM **79** (1999), no. 79 supplement 3.

[AS97] Rainer Ansorge and Thomas Sonar, *Informationsverlust, abstrakte Entropie und die numerische Beschreibung des zweiten Hauptsatzes der Thermodynamik*, ZAMM **77** (1997), no. 11, 803–821.

[Bot95] Nicola Botta, *Numerical investigations of two-dimensional euler flow: cylinder at transonic speed*, Ph.D. thesis, ETH Zürich, 1995.

[CFL28] R. Courant, K. O. Friedrichs, and H. Lewy, *Über die partiellen Differenzengleichungen der mathematischen Physik*, Math. Annalen **100** (1928), 32–74.

[DiP85] Ronald J. DiPerna, *Measure-valued solutions to conservation laws*, Arch. Rat. Mech. Anal. **88** (1985), 223–270.

[DO95] C. Armando Duarte and J. T. Oden, *hp clouds – a meshless method to solve boundary value problems*, Tech. Report 95-05, Texas Institute for Computational and Applied Mathematics, 1995.

[Dua95] C. Armando Duarte, *A review of some meshless methods to solve partial differential equations*, Tech. Report 95-06, Texas Institute for Computational and Applied Mathematics, 1995.

[Ein05] Albert Einstein, *Zur Elektrodynamik bewegter Körper*, Annalen der Physik **17** (1905), 891–921.

[ELS89] Björn Engquist, Per Lötstedt, and Björn Sjögreen, *Nonlinear filters for efficient shock computation*, Math. Comp. **52** (1989), 509–537.

[EO80] Björn Engquist and Stanley Osher, *Stable and entropy satisfying approximations for transonic flow calculations*, Math. Comp. **34** (1980), no. 149, 45–75.

[EO81] Björn Engquist and Stanley Osher, *One-sided difference approximations for nonlinear conservation laws*, Math. Comp. **36** (1981), no. 154, 321–351.

[FP97] Joel H. Ferziger and Milovan Perić, *Computational methods for fluid dynamics*, Springer, 1997.

[Fri98] Oliver Friedrich, *Weighted essentially non-oscillatory schemes for the interpolation of mean values on unstructured grids*, J. Comp. Physics **144** (1998), 194–212.

[God59] S. K. Godunov, *A difference scheme for numerical computation of discontinuous solutions of hydrodynamic equations*, Math. USSR Sbornik **47** (1959), 271–306.

[GR96] Edwin Godlewski and Pierre-Arnaud Raviart, *Numerical approximation of hyperbolic systems of conservation laws*, Applied Mathematical Sciences, no. 118, Springer, 1996.

[Gut98] Tim Gutzmer, *AMCIT – a new method for mesh adaptation when solving time-dependent conservation laws*, Ph.D. thesis, ETH Zürich, 1998.

[Hem99] Daniel Hempel, *Rekonstruktionsverfahren auf unstrukturierten Gittern zur numerischen Simulation von Erhaltungsprinzipien*, Ph.D. thesis, Universität Hamburg, 1999.

[HEOC87] Ami Harten, Björn Engquist, Stanley Osher, and S. R. Chakravarthy, *Uniformly high order accurate essentially non-oscillatory schemes III*, J. Comp. Physics **71** (1987), 231–303.

[HH83] Ami Harten and J. M. Hyman, *Self adjusting grid methods for one-dimensional hyperbolic conservation laws*, J. Comp. Physics **50** (1983), 235–269.

[Hir90] Charles Hirsch, *Numerical computation of internal and external flows*, vol. 2, John Wiley & Sons, 1990.

[HO87] Ami Harten and Stanley Osher, *Uniformly high order accurate essentially non-oscillatory schemes I*, SIAM J. Numer. Anal. **24** (1987), 279–309.

[HS99] Changqing Hu and Chi-Wang Shu, *Weighted essentially non-oscillatory schemes on triangular meshes*, J. Comp. Physics **150** (1999), no. 1, 97–127.

[Kru70] S. N. Kružkov, *First order quasilinear equations in several independent variables*, Math. USSR Sbornik **10** (1970), 217–243.

[Lax71] Peter D. Lax, *Shock waves and entropy*, Contributions to Nonlinear Functional Analysis (E. H. Zarontonello, ed.), Academic Press, 1971.

[Lax73] Peter D. Lax, *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, Regional conference series in applied mathematics, 1973.

[LeV90] Randall J. LeVeque, *Numerical methods for conservation laws*, Birkhäuser Verlag, Basel, Boston, Berlin, 1990.

[LOC94] X.-D. Liu, Stanley Osher, and T. Chan, *Weighted essentially non-oscillatory schemes*, J. Comp. Physics **115** (1994), 200–212.

[Maj84] Andrew Majda, *Compressible fluid flow and systems of conservation laws in several space variables*, Applied Mathematical Sciences, vol. 53, Springer, 1984.

[Mei96] Andreas Meister, *Zur zeitgenauen numerischen Simulation reibungsbehafteter, kompressibler, turbulenter Strömungsfelder mit einer impliziten Finite-Volumen-Methode vom Box-Typ*, Ph.D. thesis, TH Darmstadt, 1996.

[OG97] Carl Ollivier-Gooch, *Quasi-ENO schemes for unstructured meshes based on unlimited data-dependent least-squares reconstruction*, J. Comp. Physics **133** (1997), 6–17.

[OS82] Stanley Osher and F. Solomon, *Upwind difference schemes for hyperbolic systems of conservation laws*, J. Comp. Physics **36** (1982), no. 158, 339–374.

[Osh84] Stanley Osher, *Riemann solvers, the entropy condition and difference approximations*, SIAM J. Numer. Anal. **21** (1984), 217–235.

[Roe81a] P. L. Roe, *Approximate Riemann solvers, parameter vectors and difference schemes*, J. Comp. Physics **43** (1981), 357–372.

[Roe81b] P. L. Roe, *The use of the Riemann problem in finite difference schemes*, Lecture Notes in Physics, vol. 141, Springer, 1981, pp. 354–359.

[RP84] P. L. Roe and J. Pike, *Efficient construction and utilization of approximate Riemann solutions*, Computing methods in applied sciences and Engineering (R. Glowinski and J. L. Lions, eds.), North Holland, 1984.

[Smo83] Joel Smoller, *Shock waves and reaction-diffusion equations*, Grundlehren der mathematischen Wissenschaften, no. 258, Springer, 1983.

[SO88] Chi-Wang Shu and Stanley Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comp. Physics **77** (1988), 439–471.

[SO89] Chi-Wang Shu and Stanley Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes II*, J. Comp. Physics **83** (1989), 32–78.

[Son97a] Thomas Sonar, *Mehrdimensionale ENO-Verfahren*, B. G. Teubner, 1997.

[Son97b] Thomas Sonar, *On the construction of essentially non-oscillatory finite volume approximations to hyperbolic conservation laws on general triangulations: Polynomial recovery, accuracy and stencil selection*, Comp. Meth. Appl. Mech. Engrg. **140** (1997), 157–181.

[Son98] Thomas Sonar, *On families of pointwise optimal finite volume ENO approximations*, SIAM J. Numer. Anal. **35** (1998), no. 6, 2350–2369.

[Spe87] Stephanus Petrus Spekreijse, *Multigrid solution of the steady Euler equations*, Ph.D. thesis, Centrum voor Wiskunde en Informatica Amsterdam, 1987.

[Tor97] Eleuterio F. Toro, *Riemann solvers and numerical methods for fluid dynamics*, Springer, 1997.

[vL74] Bram van Leer, *Towards the ultimate conservative difference scheme II. monotonicity and conservation combined in a second order scheme*, J. Comp. Physics **14** (1974), 361–370.

[vL77] Bram van Leer, *Towards the ultimate conservative difference scheme III. Upstream centered finite difference schemes for ideal compressible flow*, J. Comp. Physics **23** (1977), 263–275.

[vL79] Bram van Leer, *Towards the ultimate conservative difference scheme V. A second order sequel to Godunovs method*, J. Comp. Physics **32** (1979), 101–136.

[WC84] Paul Woodward and Phillip Colella, *The numerical simulation of two-dimensional fluid flow with strong shocks*, J. Comp. Physics **54** (1984), 115–173.

Danksagung

Ganz besonders möchte ich mich bei Professor Dr. Thomas Sonar bedanken, der mich nach meinem Diplom in seine Arbeitsgruppe nach Hamburg eingeladen hat. Während der Zeit in Hamburg hat er mir ermöglicht, in großer Freiheit den mich interessierenden Fragestellungen nachzugehen, wobei mir sein Rat und seine Kritik stets von großer Hilfe waren.

Diesen Kontakt vermittelt zu haben ist das Verdienst von Jürgen Dehnhardt, der nicht nur meine Diplomarbeit in Hannover betreute, sondern mich auch in seinem Seminar über Hyperbolische Differentialgleichungen in dieses faszinierende Forschungsgebiet eingeführt hat.

In der Anfangsphase dieses Projektes haben Daniel Hempel und ich intensiv zusammengearbeitet und bis heute gemeinsam ein Büro bewohnt. Ihm verdanke ich viele nützliche Anregungen (oft spät abends bei gemeinsamen Arbeitssitzungen in etlichen Pizzerien) und eine sehr erfrischende Zeit.

Die von mir verwendeten Gitter sind mit Programmen von Oliver Friedrich erzeugt, und bei der Darstellung der Ergebnisse war sein Programm „vis2d“ von unschätzbarem Wert.

Diese Arbeit wurde in den verschiedensten Stadien ihrer Entstehung von Daniel Hempel und Oliver Friedrich gelesen und kommentiert. Ihre Kritik hat viel zur Lesbarkeit des Textes beigetragen. Für sämtliche verbleibenden Mängel bin ich selbstverständlich alleine verantwortlich.

Freundlicherweise hat Professor Dr. Rainer Ansorge das Korreferat für meine Dissertation übernommen.

Meiner Mutter Barbara Ahrend und Frau Christine Färber möchte ich für ihre große Unterstützung danken.

Arne Ahrend

Abstract

We present a collocation method which can easily be generalized to higher dimensions for solving hyperbolic conservation laws on unstructured grids in two space dimensions. In particular, our method does not require a tessalation of the computational domain, but only a cloud of smoothly distributed points and neighbourhood relations between points.

Point-based schemes are normally implemented as finite difference schemes on Cartesian grids, and the approximate solution given by point values is generally regularized in a way similar to the classical Lax-Friedrichs scheme. Those grids are not very well-suited for modeling complex geometries.

For the collocation scheme presented in this thesis numerical flux functions can be employed for the approximate solution of Riemann problems in suitably chosen directions. The weights for computing the divergence for the collocation functionals can be obtained efficiently as solutions of dual problems during the preprocessing stage. Choosing the collocation functionals as convex combinations of the point evaluations also used for divergence approximation proves crucial to stabilizing the scheme.

Combined with WENO reconstruction known from finite volume methods, it is possible to obtain numerical solutions and computing times comparable to finite volume methods. We demonstrate uniform stability of this kind of reconstruction under similarity transformations of the computational domain and classify our method within a general theory comprising also the finite volume methods.

The thesis closes with numerical test cases. These demonstrate that the presented method is stable and capable of resolving discontinuities with high accuracy.

Zusammenfassung

In dieser Arbeit wird ein Kollokationsverfahren zur Lösung hyperbolischer Erhaltungsgleichungen auf unstrukturierten Gittern in zwei Raumdimensionen entwickelt, welches sich leicht auf höhere Dimensionen übertragen lässt. Insbesondere setzt unsere Methode keine Tessalation des Rechengebietes voraus, sondern verwendet lediglich eine gleichmäßig verteilte Punktwolke und Nachbarschaftsrelationen zwischen Punkten.

Punktbasierte Verfahren werden typischerweise als Finite-Differenzen-Verfahren auf kartesischen Gittern implementiert, wobei in der Regel eine an das klassische Lax-Friedrichs-Verfahren angelehnte Regularisierung der punktweise gegebenen Näherungslösung vorgenommen wird. Derartige Gitter eignen sich nur bedingt zur Modellierung komplexer Geometrien.

Für das in dieser Arbeit vorgestellte Kollokationsverfahren konnten numerische Flussfunktionen zur näherungsweisen Lösung von Riemann-Problemen in geeignet gewählten Raumrichtungen eingesetzt werden. Die zur Berechnung der Divergenzen für die einzelnen Kollokationsfunktionale erforderlichen Gewichte lassen sich als Lösungen dualer Probleme effizient in der Vorbereitungsphase des Programmes bestimmen. Als entscheidend zur Stabilisierung des Verfahrens erweist sich die Wahl der Kollokationsfunktionale als Konvexkombination der auch für die Divergenzapproximation verwendeten Punktauswertungen.

Zusammen mit einer aus dem Bereich der Finite-Volumen-Verfahren bekannten WENO-Rekonstruktion lassen sich in zweidimensionalen Testfällen Lösungen und Rechenzeiten, die denjenigen etablierter Finite-Volumen-Methoden vergleichbar sind, erreichen. Wir zeigen, dass diese Art der Rekonstruktion sowohl für Kollokations- als auch für Zellmittelungsfunktionale gleichmäßig stabil unter Ähnlichkeitstransformationen des Rechengebietes ist und ordnen unser Verfahren in eine allgemeine Theorie, welche auch die Finite-Volumen-Verfahren umfasst, ein.

Den Abschluss der Arbeit bilden numerische Testfälle. Diese demonstrieren, dass die vorgestellte Methode stabil und geeignet ist, Unstetigkeiten mit hoher Genauigkeit aufzulösen.

Lebenslauf

Arne Ahrend
geboren am 20. Mai 1969 in Hildesheim

Schulbesuch

August 75 – Juli 79	Grundschule II Sarstedt
August 79 – Juli 81	Orientierungsstufe Sarstedt
August 81 – Juni 88	Gymnasium Sarstedt
Juni 88	Abitur

Zivildienst

June 88 – October 89	St. Nicolai Kirchengemeinde Sarstedt

Universität

Oktober 89 – Juli 93	Grundstudium Mathematik und Physik an der Universität Hannover
Oktober 93 – Juli 94	Auslandsstudium am King's College London
Oktober 94 – Juli 96	Hauptstudium Mathematik an der Universität Hannover
April 92	Vordiplom
Juli 96	Diplom im Studiengang Mathematik
Oktober 96 – August 99	Wissenschaftlicher Mitarbeiter des Instituts für angewandte Mathematik der Universität Hamburg Dissertation und Übungsgruppenleiter an der Technischen Universität Hamburg-Harburg