# Advanced Numerical Methods for Hyperbolic Equations and Applications

Lecture notes by Michael Dumbser

Laboratory of Applied Mathematics,
University of Trento
Via Mesiano 77, I-38040 Trento, Italy
*E-mail address*: michael.dumbser@ing.unitn.it

# Contents

# Preface

These lecture notes summarize in a very compact form the material presented during the second course of the *Winter School on Advanced Numerical Methods for Free Surface Flows*, held in February 2011 in Trento, Italy. The reader is introduced to the elementary theory and basic concepts of modern numerical methods for the discretization of hyperbolic conservation laws in one and multiple space dimensions. These lecture notes do not claim to be complete and should be used together with additional standard literature available on the topic, such as the books of Toro [**77**], LeVeque [**55**], Hirsch [**43, 44**], Kröner [**51**] and the classical book of Godlewski–Raviart [**36**].

Trento, 13/02/2011  
Prof. Dr.-Ing. Michael Dumbser

# Part 1

# Basic Algorithms for the Discretization of Hyperbolic Conservation Laws

CHAPTER 1

# Linear Hyperbolic Equations

## 1. Linear scalar advection equation

We consider the following partial differential equation (PDE)

$$\frac{\partial q}{\partial t} + a\frac{\partial q}{\partial x} = 0, \tag{1.1}$$

with $q = q\left(x, t\right)$, $x \in \mathbb{R}$, $t \in \mathbb{R}_0^+$, $a \in \mathbb{R}$ and the initial condition (IC)

$$q\left(x, 0\right) = h\left(x\right). \tag{1.2}$$

The exact solution of (1.1) is

$$q\left(x, t\right) = h\left(x - at\right), \tag{1.3}$$

which can be easily proven as follows:

$$q_t = -ah', q_x = h' \Rightarrow -ah' + ah' = 0. \tag{1.4}$$

The solution $q\left(x, t\right)$ consists of the *transport* of the initial condition $h\left(x\right)$, moving with a constant velocity $a$.

### 1.1. Characteristic curves.

The so–called *characteristic curves* or *characteristics* are functions $x\left(t\right)$ that satisfy the following initial value problem of the ordinary differential equation (ODE)

$$\frac{dx}{dt} = a, \tag{1.5}$$

with $x \in \mathbb{R}$, $t \in \mathbb{R}_0^+$, $a \in \mathbb{R}$ and the initial condition $x\left(0\right) = x_0$. The value $x_0$ is the initial position or the so–called *foot of the characteristic* and $a$ denotes the characteristic velocity in (1.5). The solution of the ordinary differential equation (ODE) given by (1.5) is

$$x(t) = x_0 + at, \tag{1.6}$$

plotted in Fig. 1.

The *total derivative* or *material derivative* of $q\left(x, t\right)$ computed along the characteristic curve $x\left(t\right)$ is given by

$$\frac{dq}{dt} = \frac{\partial q}{\partial t}\frac{dt}{dt} + \frac{\partial q}{\partial x}\frac{dx}{dt} = \frac{\partial q}{\partial t} + a\frac{\partial q}{\partial t} = 0. \tag{1.7}$$

Hence, along the characteristic $x\left(t\right) = x_0 + at$ the PDE (1.1) reduces to the trivial ODE

$$\frac{dq}{dt} = 0. \tag{1.8}$$

It follows that the solution $q = q\left(x, t\right)$ remains *constant* along the characteristic curve. Hence, the solution at a general point $x$ in space and at a general time $t$ can be computed by evaluating the initial condition at the *foot of the characteristic* passing through $x$ and $t$:

$$q\left(x, t\right) = q\left(x_0, 0\right) = h\left(x_0\right) = h\left(x - at\right). \tag{1.9}$$

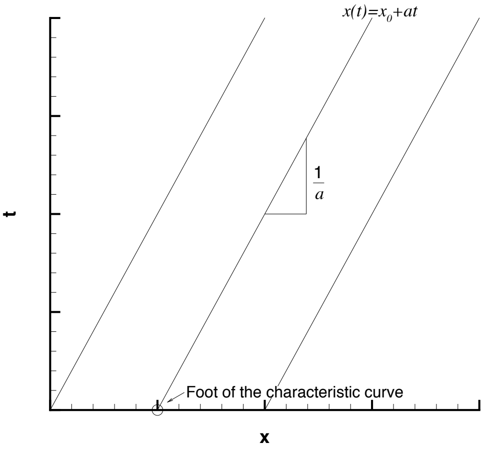

FIGURE 1. The characteristic curves resulting from the initial value problem (1.5).

Note that for the construction of the solution in Eqn. (1.9), the differentiability of the solution $q = q\,(x, t)$ has not been used. In fact, the function $q = q\,(x, t)$ can also be discontinuous and discontinuities will propagate along characteristic curves.

Already with this very simple example, we can see the important difference between the *Eulerian* description of physical processes, where the observer looks at the process from a fixed laboratory frame, and the *Lagrangian* description of physical processes, where the observer moves along characteristic curves. In the Eulerian case, the physics of advection is described by a partial differential equation, here the PDE (1.1). In the Lagrangian case, the physics of advection is described by a system of two ordinary differential equations, namely the ODE (1.5) for the evolution of the observer's position and the ODE (1.8), which describes the temporal evolution of the state variable $q$ in the co–moving frame of the observer. While the Eulerian description uses *partial* derivatives to model the physics, the Lagrangian description uses *total* or *material* derivatives. However, both formulations are equivalent, even in the presence of discontinuities, as proven by Wagner in [**89**].

EXAMPLE 1. *Let us assume a unitary characteristic velocity $a = 1$ in the linear scalar advection equation* (1.1) *and the following initial condition*

$$q\,(x, 0) = h\,(x) = \begin{cases} 1, & \text{if } -1 \leq x \geq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1.10}$$

*Then, the exact solution reproduces the initial profile translated to the right with a velocity of one. The discontinuities remain sharp, since the PDE* (1.1) *does not contain any diffusion terms. The exact solution is plotted in Fig. 2 following the Eulerian description (on the left) and the Lagrangian approach (on the right).*

**1.2. Riemann problem.** The Riemann problem is a particular initial value problem (Cauchy problem) of a PDE. The initial condition consists in this case of two piecewise constant states $q_L$ and $q_R$, separated by a discontinuity:

$$q_t + aq_x = 0, \tag{1.11}$$

$$q\,(x, 0) = h\,(x) = \begin{cases} q_L, & \text{if } x \leq 0, \\ q_R, & \text{if } x > 0, \end{cases} \tag{1.12}$$

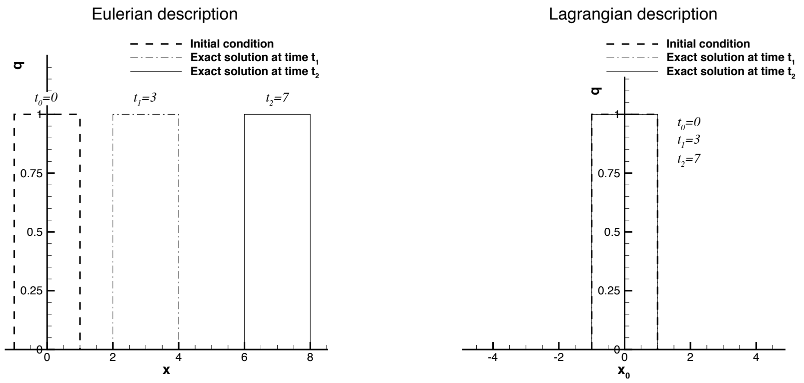

FIGURE 2. The exact solution of the initial value problem (1.10) of the linear scalar advection equation (1.1) with $a = 1$ according to the Eulerian description (on the left) and the Lagrangian description (on the right).

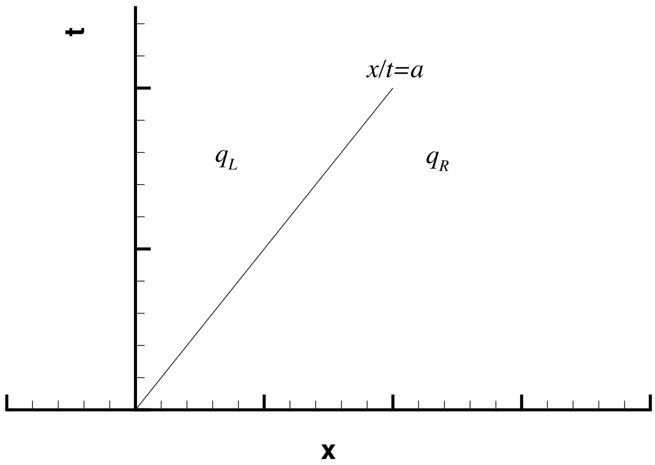

FIGURE 3. Sketch of the exact solution of the Riemann problem for the linear scalar advection equation.

where the notation $q_t$, $q_x$ denotes the partial derivatives of the function $q\left(x,t\right)$ with respect to $t$ and $x$, respectively. The exact solution of the Riemann problem is

$$q\left(x,t\right) = h\left(x - at\right) = \begin{cases} q_L, & \text{if } x - at \leq 0, \\ q_R, & \text{if } x - at > 0, \end{cases} \tag{1.13}$$

that is equivalent for $t > 0$ to

$$q\left(x,t\right) = \begin{cases} q_L, & \text{if } \frac{x}{t} \leq a, \\ q_R, & \text{if } \frac{x}{t} > a. \end{cases} \tag{1.14}$$

An example of an exact solution of the Riemann problem for the linear scalar advection equation is plotted in Fig. 3.

**1.3. Initial boundary value problems.** Let us consider a general initial boundary value problem given by the partial differential equation (PDE) and the corresponding initial condition (IC) and boundary conditions (BC), as follows

$$\begin{array}{ll} \text{PDE:} & q_t + a q_x = 0, \quad x \in [x_L, x_R], t \in \mathbb{R}_0^+, \\ \text{IC:} & q(x,0) = h(x), \\ \text{BC:} & q(x_L, t) = b_L(t), \quad q(x_R, t) = b_R(t). \end{array} \tag{1.15}$$

The exact solution for $a \in \mathbb{R}$ is the following

$$q(x,t) = \begin{cases} h(x - at), & \text{if } \max(x_L, x_L + at) \le x \le \min(x_R, x_R + at), \\ b_L\left(t - \frac{x - x_L}{a}\right), & \text{if } x_L \le x < x_L + at, \\ b_R\left(t - \frac{x_R - x}{a}\right), & \text{if } x_R \ge x > x_R + at. \end{cases} \tag{1.16}$$

## 2. Linear scalar advection reaction equation

We consider a linear scalar advection reaction equation

$$q_t + a q_x = \beta q, \quad \text{with } x \in \mathbb{R}, t \in \mathbb{R}_0^+, \beta \in \mathbb{R}_0^+, \tag{1.17}$$

$$q(x,0) = h(x), \tag{1.18}$$

where $\beta$ in the linear source term is a constant. Solving the linear scalar advection reaction equation (1.17) along a characteristic curve $x = x_0 + at$, given again by the solution of the ODE (1.5) $\frac{dx}{dt} = a$, we obtain

$$\frac{dq}{dt} = \frac{\partial q}{\partial t}\frac{dt}{dt} + \frac{\partial q}{\partial x}\frac{dx}{dt} = q_t + a q_x = \beta q. \tag{1.19}$$

The exact solution along the characteristic curve is the solution of the previous ODE:

$$q(t) = q(0)\, e^{\beta t}. \tag{1.20}$$

Hence, the exact solution of the linear scalar advection reaction equation (1.17) is

$$q(x,t) = h(x - at)\, e^{\beta t}. \tag{1.21}$$

The fact that Eqn. (1.21) is the exact solution of the advection–reaction equation can be proven by computing the derivative in time of Eqn. (1.21)

$$q_t = h\beta e^{\beta t} - a h' e^{\beta t}, \tag{1.22}$$

and the derivative in space

$$q_x = h' e^{\beta t}. \tag{1.23}$$

Replacing these expressions in (1.19) yields

$$q_t + a q_x = h\beta e^{\beta t} - a h' e^{\beta t} + a h' e^{\beta t} = \beta h e^{\beta t} = \beta q. \tag{1.24}$$

## 3. Linear advection with variable coefficient

We consider the following initial value problem

$$q_t + a(x,t)\, q_x = 0, \tag{1.25}$$

$$q(x,0) = h(x). \tag{1.26}$$

The characteristic curves are the solution of the problem

$$\frac{dx}{dt} = a(x(t), t), \tag{1.27}$$

$$x(0) = x_0. \tag{1.28}$$

In this case, the characteristics are in general no longer straight lines, but curves. The formal solution of the ODE (1.27) is

$$x(t) = x_0 + \int_0^t a\left(x(\tau), \tau\right) d\tau, \tag{1.29}$$

and hence the exact solution of the advection equation with variable coefficient (1.25) can be written formally as

$$q(x,t) = h(x_0) = h\left(x(t) - \int_0^t a\left(x(\tau), \tau\right) d\tau\right). \tag{1.30}$$

EXERCISE 1. *Prove that* (1.30) *actually is a solution of* (1.25).

EXAMPLE 2. *We consider the following problem*

$$q_t + cxq_x = 0, \tag{1.31}$$

$$q\left(x,0\right) = h\left(x\right) = \begin{cases} 1, & \text{if } |x| \leq 1, \\ 0, & \text{if } |x| > 1. \end{cases} \tag{1.32}$$

*The corresponding characteristic curves are given by the following ODE*

$$\frac{dx}{dt} = cx, \tag{1.33}$$

$$x\left(0\right) = x_0, \tag{1.34}$$

*the solution of which leads to*

$$x\left(t\right) = x_0 e^{ct}, \quad \Rightarrow \quad x_0 = xe^{-ct}. \tag{1.35}$$

*It follows that the exact solution of the problem* (1.32) *is*

$$q\left(x,t\right) = h(x_0) = h\left(xe^{-ct}\right) = \begin{cases} 1, & \text{if } |xe^{-ct}| \leq 1, \\ 0, & \text{if } |xe^{-ct}| > 1. \end{cases} \tag{1.36}$$

EXAMPLE 3. *Another linear advection equation with variable coefficient is given by*

$$q_t + cxtq_x = 0, \tag{1.37}$$

$$q\left(x,0\right) = h\left(x\right) \begin{cases} 1, & \text{if } |x| \leq 1, \\ 0, & \text{if } |x| > 1. \end{cases} \tag{1.38}$$

*We apply the method of the characteristics, as follows*

$$\frac{dx}{dt} = cxt, \tag{1.39}$$

$$x\left(0\right) = x_0, \tag{1.40}$$

*with the solution*

$$x\left(t\right) = x_0 e^{\frac{ct^2}{2}}, \quad \Rightarrow \quad x_0 = xe^{\frac{-ct^2}{2}}. \tag{1.41}$$

*This leads to the solution*

$$q\left(x,t\right) = h(x_0) = h\left(xe^{-\frac{ct^2}{2}}\right) \begin{cases} 1, & \text{if } \left|xe^{-\frac{ct^2}{2}}\right| \leq 1, \\ 0, & \text{if } \left|xe^{-\frac{ct^2}{2}}\right| > 1. \end{cases} \tag{1.42}$$

## 4. Linear hyperbolic systems

We consider the following system of equations

$$\frac{\partial \vec{Q}}{\partial t} + \underline{\underline{A}} \cdot \frac{\partial \vec{Q}}{\partial x} = \vec{0}, \quad x \in \mathbb{R}, t \in \mathbb{R}_0^+, \underline{\underline{A}} \in \mathbb{R}^{m \times m}, \tag{1.43}$$

where the vector of the unknowns $\vec{Q} = \vec{Q}\,(x,t)$ and the matrix $\underline{\underline{A}}$ are defined as

$$\vec{Q} = \begin{pmatrix} q_1 \\ q_2 \\ \dots \\ q_m \end{pmatrix}, \quad \underline{\underline{A}} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{pmatrix}, \tag{1.44}$$

with the following initial conditions

$$\vec{Q}\,(x,0) = \vec{h}\,(x)\,. \tag{1.45}$$

The system of equations (1.43) is called *hyperbolic* if all the eigenvalues of the matrix $\underline{\underline{A}}$ are real numbers *and* if there exists a complete set of $m$ eigenvectors that are linearly independent. A system (1.43) is *strictly hyperbolic* if all the eigenvalues are real and *distinct.* A system that is strictly hyperbolic is necessarily hyperbolic because distinct eigenvalues automatically imply the existence of a complete set of linearly independent eigenvectors. A system (1.43) is called *weakly hyperbolic* if all the eigenvalues are real, but not all are distinct and there does *not exist* a complete set of linearly independent eigenvectors.

The properties of the right eigenvectors $\vec{r}$ are

$$\underline{\underline{A}} \cdot \vec{r} = \lambda \vec{r}, \tag{1.46}$$

where $\lambda$ denotes the eigenvalues, which are the roots of the characteristic polynomial

$$P\,(\lambda) = \left| \underline{\underline{A}} - \lambda \underline{\underline{I}} \right| = 0, \tag{1.47}$$

with the identity matrix $\underline{\underline{I}}$. The left eigenvectors $\vec{l}$ satisfy the relation

$$\vec{l}^T \cdot \underline{\underline{A}} = \lambda \vec{l}^T. \tag{1.48}$$

The eigenvalues $\lambda$ in (1.46) and (1.48) are the same, which can be seen easily by some simple manipulations:

$$\vec{l}^T \underline{\underline{A}} = \left( \underline{\underline{A}}^T \vec{l} \right)^T, \tag{1.49}$$

hence with (1.48) it follows that

$$\underline{\underline{A}}^T \vec{l} = \lambda \vec{l}, \tag{1.50}$$

whose characteristic polynomial is

$$P(\lambda) = |\underline{\underline{A}}^T - \lambda \underline{\underline{I}}| = |\underline{\underline{A}} - \lambda \underline{\underline{I}}|, \tag{1.51}$$

which is identical to the characteristic polynomial of the right eigenvectors.

We are furthermore interested in a relationship between the left and right eigenvectors. Let us define $\underline{\underline{R}}$, $\underline{\underline{L}}$ and $\underline{\underline{\Lambda}}$, as follows

$$\underline{\underline{R}} = \left( \vec{r}_1, \vec{r}_2, ..., \vec{r}_m \right), \quad \underline{\underline{L}} = \begin{pmatrix} \vec{l}_1 \\ \vec{l}_2 \\ \dots \\ \vec{l}_m \end{pmatrix}, \quad \underline{\underline{\Lambda}} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_m \end{pmatrix}. \tag{1.52}$$

According to the definition of the right eigenvectors (1.46) and by multiplying with the matrix of the left eigenvectors $L$ from the left, we obtain

$$\underline{\underline{L}}\,(\underline{\underline{A}}\,\underline{\underline{R}}) = \underline{\underline{L}}\,(\underline{\underline{R}}\,\underline{\underline{\Lambda}}), \tag{1.53}$$

and similarly for the left eigenvectors,

$$\left(\underline{\underline{L}}\,\underline{\underline{A}}\right)\underline{\underline{R}} = \left(\underline{\underline{\Lambda}}\,\underline{\underline{L}}\right)\underline{\underline{R}}. \tag{1.54}$$

This leads to

$$\underline{\underline{L}}\,\underline{\underline{R}}\,\underline{\underline{\Lambda}} = \underline{\underline{\Lambda}}\,\underline{\underline{L}}\,\underline{\underline{R}}. \tag{1.55}$$

Hence, $\left(\underline{\underline{L}}\,\underline{\underline{R}}\right)$ and $\underline{\underline{\Lambda}}$ are commutative matrices. The only matrix that is commutative with a diagonal matrix with distinct values, is also a *diagonal* matrix. It follows that the right and left eigenvectors are orthogonal, i.e.

$$\vec{l}_i^T \vec{r}_j \begin{cases} \neq 0, & \text{if } i = j, \\ = 0, & \text{if } i \neq j. \end{cases} \tag{1.56}$$

With a proper choice of the scaling factors of the eigenvectors, one can *always* obtain the result that the right and left eigenvectors become *orthonormal*, that is

$$\underline{\underline{L}}\,\underline{\underline{R}} = \underline{\underline{I}}, \quad \text{or} \quad \underline{\underline{L}} = \underline{\underline{R}}^{-1}. \tag{1.57}$$

Finally, from (1.46) and (1.57) we obtain

$$\underline{\underline{A}} = \underline{\underline{R}}\,\underline{\underline{\Lambda}}\,\underline{\underline{R}}^{-1} = \underline{\underline{R}}\,\underline{\underline{\Lambda}}\,\underline{\underline{L}}, \tag{1.58}$$

which is equivalent to

$$\underline{\underline{\Lambda}} = \underline{\underline{R}}^{-1}\underline{\underline{A}}\,\underline{\underline{R}} = \underline{\underline{LAR}}. \tag{1.59}$$

**4.1. Characteristic variables.** To solve the linear hyperbolic system of equations one introduces the so called characteristic variables $\vec{C}$, defined by

$$\vec{Q} = \underline{\underline{R}}\vec{C}, \quad \vec{C} = \underline{\underline{R}}^{-1}\vec{Q}. \tag{1.60}$$

Then, the system (1.43) becomes

$$\frac{\partial}{\partial t}\left(\underline{\underline{R}}\vec{C}\right) + \underline{\underline{A}}\frac{\partial}{\partial x}\left(\underline{\underline{R}}\vec{C}\right) = 0. \tag{1.61}$$

In the case of a linear hyperbolic system with constant coefficients, the matrix $\underline{\underline{A}}$ and thus also $\underline{\underline{R}}$, $\underline{\underline{R}}^{-1}$ and $\underline{\underline{\Lambda}}$ do not depend on $(x, t)$, hence we have

$$\underline{\underline{R}}\frac{\partial \vec{C}}{\partial t} + \underline{\underline{R}}\,\underline{\underline{A}}\frac{\partial \vec{C}}{\partial x} = 0. \tag{1.62}$$

According to (1.59) this leads to a new formulation of the linear system (1.43) written in characteristic variables

$$\frac{\partial \vec{C}}{\partial t} + \underline{\underline{\Lambda}}\frac{\partial \vec{C}}{\partial x} = 0. \tag{1.63}$$

The system (1.63) is a system of $m$ *decoupled* linear scalar advection equations,

$$\frac{\partial c_i}{\partial t} + \lambda_i \frac{\partial c_i}{\partial x} = 0, \quad \text{for } i = 1, .., m, \tag{1.64}$$

which can be easily solved componentwise. The initial condition for the system (1.63) is just obtained from (1.60), as follows

$$\vec{C}\,(x, 0) = \underline{\underline{R}}^{-1}\vec{h}\,(x)\,. \tag{1.65}$$

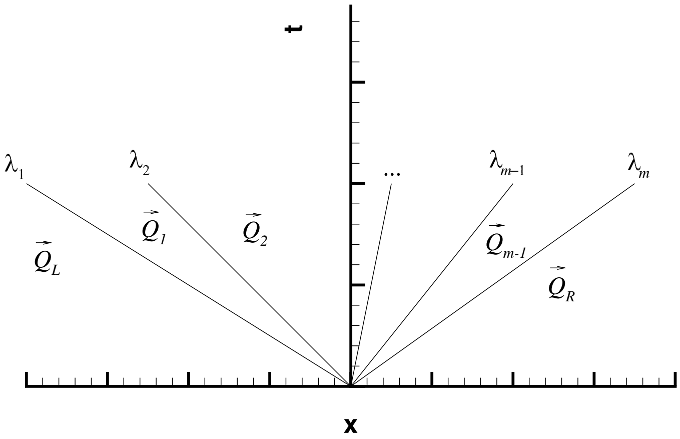

FIGURE 4. Sketch of the exact solution of the Riemann problem for a linear hyperbolic system of equations.

**4.2. Riemann problem.** Now, we consider the Riemann problem

$$\frac{\partial \vec{Q}}{\partial t} + \underline{\underline{A}} \frac{\partial \vec{Q}}{\partial x} = 0, \tag{1.66}$$

$$\vec{Q}\,(x,0) = \vec{h}\,(x) = \begin{cases} \vec{Q}_L, & \text{if } x \leq 0, \\ \vec{Q}_R, & \text{if } x > 0. \end{cases} \tag{1.67}$$

Written in characteristic variables it becomes

$$\frac{\partial \vec{C}}{\partial t} + \underline{\underline{\Lambda}} \frac{\partial \vec{C}}{\partial x} = 0. \tag{1.68}$$

$$\vec{C}\,(x,0) = \begin{cases} \underline{\underline{R}}^{-1}\vec{Q}_L := \vec{C}_L, & \text{if } x \leq 0, \\ \underline{\underline{R}}^{-1}\vec{Q}_R := \vec{C}_R, & \text{if } x > 0. \end{cases} \tag{1.69}$$

Then, the exact solution of the Riemann problem (1.66) written in terms of the characteristic variables is

$$c_i\,(x,t) = \begin{cases} c_i^L, & \text{if } \frac{x}{t} \leq \lambda_i, \\ c_i^R, & \text{if } \frac{x}{t} > \lambda_i, \end{cases} \tag{1.70}$$

or, in more compact form, using the matrix sign function, which is applied componentwise for a diagonal matrix:

$$\vec{C}\,(x,t) = \frac{1}{2}\left(1 + \text{sign}(\underline{\underline{\Lambda}} - \frac{x}{t}\underline{\underline{I}})\right)\vec{C}_L + \frac{1}{2}\left(1 - \text{sign}(\underline{\underline{\Lambda}} - \frac{x}{t}\underline{\underline{I}})\right)\vec{C}_R. \tag{1.71}$$

The exact solution in physical variables results from (1.60). Hence, we obtain the following general expression for the solution in physical variables in terms of the initial condition (1.67):

$$\vec{Q}\,(x,t) = \frac{1}{2}\underline{\underline{R}}\left(\underline{\underline{I}} + \text{sign}(\underline{\underline{\Lambda}} - \frac{x}{t}\underline{\underline{I}})\right)\underline{\underline{R}}^{-1}\vec{Q}_L + \frac{1}{2}\underline{\underline{R}}\left(\underline{\underline{I}} - \text{sign}(\underline{\underline{\Lambda}} - \frac{x}{t}\underline{\underline{I}})\right)\underline{\underline{R}}^{-1}\vec{Q}_R. \tag{1.72}$$

The solution consists of $m+1$ piecewise constant states, separated by the characteristics moving with speeds $\lambda_i$, see Fig. 4.

EXAMPLE 4. *Let us solve the following system of equations*

$$\frac{\partial \vec{Q}}{\partial t} + \underline{\underline{A}} \frac{\partial \vec{Q}}{\partial x} = \vec{0}, \quad \text{with } \underline{\underline{A}} = \begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}, \tag{1.73}$$

$$\vec{Q} = h\left(x\right). \tag{1.74}$$

*The characteristic polynomial results*

$$P\left(\lambda\right) = \left\| \begin{array}{cc} -\lambda & a \\ b & -\lambda \end{array} \right\| = \lambda^2 - ab = 0, \tag{1.75}$$

*and the corresponding eigenvalues are $\lambda_{1,2} = \mp\sqrt{ab}$. Now, we compute the right eigenvectors, solving the following linear systems related to the first and second eigenvalue, respectively*

$$\left( \begin{array}{cc} +\sqrt{ab} & a \\ b & +\sqrt{ab} \end{array} \right) \left( \begin{array}{c} x_1 \\ x_2 \end{array} \right) = \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \tag{1.76}$$

$$\left( \begin{array}{cc} -\sqrt{ab} & a \\ b & -\sqrt{ab} \end{array} \right) \left( \begin{array}{c} x_1 \\ x_2 \end{array} \right) = \left( \begin{array}{c} 0 \\ 0 \end{array} \right). \tag{1.77}$$

*We obtain*

$$\underline{\underline{R}} = \left( \begin{array}{cc} 1 & 1 \\ -\sqrt{\frac{b}{a}} & +\sqrt{\frac{b}{a}} \end{array} \right), \quad \underline{\underline{\Lambda}} = \left( \begin{array}{cc} -\sqrt{ba} & 0 \\ 0 & +\sqrt{ba} \end{array} \right), \quad \underline{\underline{R}}^{-1} = \frac{1}{2} \left( \begin{array}{cc} 1 & -\sqrt{\frac{a}{b}} \\ 1 & +\sqrt{\frac{a}{b}} \end{array} \right). \tag{1.78}$$

Example 5. *We consider a system of three equations, as follows*

$$\frac{\partial \vec{Q}}{\partial t} + \underline{\underline{A}} \frac{\partial \vec{Q}}{\partial x} = \vec{0}, \quad \text{with } \underline{\underline{A}} = \left( \begin{array}{ccc} 1 & 2 & -2 \\ -1 & 1 & 1 \\ -1 & 2 & 0 \end{array} \right), \tag{1.79}$$

$$\vec{Q} = h\left(x\right). \tag{1.80}$$

*We compute the eigenvalues from the characteristic polynomial*

$$P\left(\lambda\right) = \left\| \begin{array}{ccc} 1-\lambda & 2 & -2 \\ -1 & 1-\lambda & 1 \\ -1 & 2 & -\lambda \end{array} \right\| = \left(1-\lambda\right)\left(-\lambda\left(1-\lambda\right)-2\right) - 2\left(\lambda+1\right) - 2\left(-2+\left(1-\lambda\right)\right) = 0. \tag{1.81}$$

*They are $\lambda_{1,2} = \mp 1$ and $\lambda_3 = 2$. Using the first eigenvalue we obtain the following right eigenvector*

$$\left( \begin{array}{ccc} 2 & 2 & -2 \\ -1 & 2 & 1 \\ -1 & 2 & 1 \end{array} \right) \left( \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \right) = \left( \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right), \quad \Rightarrow \quad \vec{r_1} = \left( \begin{array}{c} 1 \\ 0 \\ 1 \end{array} \right). \tag{1.82}$$

*Similarly, for the other eigenvectors we have*

$$\left( \begin{array}{ccc} 0 & 2 & -2 \\ -1 & 0 & 1 \\ -1 & 2 & -1 \end{array} \right) \left( \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \right) = \left( \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right), \quad \Rightarrow \quad \vec{r_2} = \left( \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right), \tag{1.83}$$

$$\left( \begin{array}{ccc} -1 & 2 & -2 \\ -1 & -1 & 1 \\ -1 & 2 & -2 \end{array} \right) \left( \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \right) = \left( \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right), \quad \Rightarrow \quad \vec{r_3} = \left( \begin{array}{c} 0 \\ 1 \\ 1 \end{array} \right). \tag{1.84}$$

*Finally, the matrices $\underline{\underline{R}}$, $\underline{\underline{R}}^{-1}$ and $\underline{\underline{\Lambda}}$ result as*

$$\underline{\underline{R}} = \left( \begin{array}{ccc} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{array} \right), \quad \underline{\underline{\Lambda}} = \left( \begin{array}{ccc} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{array} \right), \quad \underline{\underline{R}}^{-1} = \left( \begin{array}{ccc} 0 & -1 & 1 \\ 1 & 1 & -1 \\ -1 & 0 & 1 \end{array} \right). \tag{1.85}$$

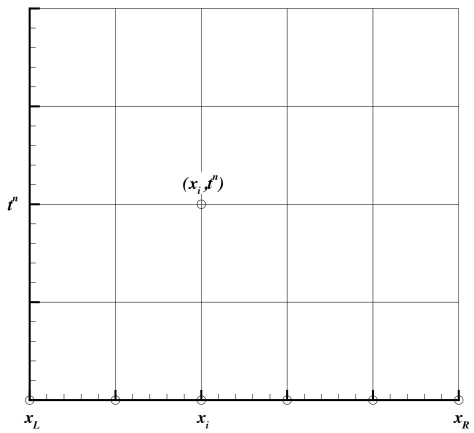

FIGURE 5. The mesh used to compute the numerical solution.

## 5. Numerical methods for the linear scalar advection equation

The aim of this section is to suggest some numerical methods to solve numerically the linear advection equation (PDE), given the initial (IC) and boundary (BC) conditions

$$\begin{array}{ll} \text{PDE:} & q_t + aq_x = 0, \quad x \in [x_L; x_R], t \in \mathbb{R}_0^+, q \in \mathbb{R}, a \in \mathbb{R}, \\ \text{IC:} & q(x,0) = h(x), \\ \text{BC:} & q(x_L,t) = b_L(t), \quad q(x_R,t) = b_R(t). \end{array} \tag{1.86}$$

For the numerical solution of (1.86) we first proceed with the *discretization* of space and time by introducing a *mesh* composed of *discrete points* in space and time as follows:

$$x_i = x_L + i\Delta x, \quad i = 0, 1, ... \tag{1.87}$$

$$t^n = n\Delta t, \quad n = 0, 1, ... \tag{1.88}$$

where $\Delta x$ is the interval length of the mesh or the *mesh spacing* and $\Delta t$ is the time step. The indices $i$ and $n$ identify a node of the mesh in space and time, as depicted in Fig. 5. Let us introduce the short notation for the numerical solution, defined by

$$q_i^n = q(x_i, t^n). \tag{1.89}$$

First, we consider the *finite difference* method to solve (1.86). A simple finite difference applied to the time derivative in (1.86) results to be

$$q_t(x_i, t^n) \approx \frac{q_i^{n+1} - q_i^n}{\Delta t}, \tag{1.90}$$

which is obviously consistent with the *definition* of the derivative

$$q_t(x_i, t^n) = \lim_{\Delta t \to 0} \frac{q(x_i, t^n + \Delta t) - q(x_i, t^n)}{\Delta t}. \tag{1.91}$$

While the continuous derivative (1.91) uses *infinitesimal* differences, the discrete derivative (1.90) uses *finite* differences.

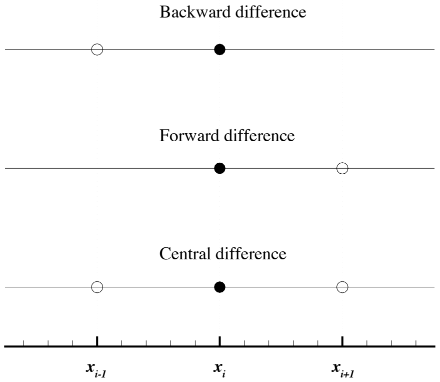

FIGURE 6. The discretization of the derivatives in space following the finite difference approach.

The particular finite difference (1.90) applied to a time derivative is also called *first order Euler* time discretization. To discretize the derivative in space there are at least the following three different possibilities, as sketched in Fig. 6:

- The so called *backward difference*

$$q_x\left(x_i, t^n\right) \approx \frac{q_i^n - q_{i-1}^n}{\Delta x}. \tag{1.92}$$

- The *forward difference*

$$q_x\left(x_i, t^n\right) \approx \frac{q_{i+1}^n - q_i^n}{\Delta x}. \tag{1.93}$$

- The *central difference*

$$q_x\left(x_i, t^n\right) \approx \frac{q_{i+1}^n - q_{i-1}^n}{2\Delta x}. \tag{1.94}$$

**5.1. Some basic explicit finite difference schemes.** It follows that three different numerical methods are possible to solve the linear advection equation (1.86) using the explicit Euler time discretization:

- The first scheme is based on the *backward difference* discretization of the space derivative

$$q_i^{n+1} = q_i^n - a\frac{\Delta t}{\Delta x}\left(q_i^n - q_{i-1}^n\right), \tag{1.95}$$

Note that according to the sign of the constant velocity $a$, we obtain an *upwind* scheme if $a > 0$ and a *downwind* method with $a < 0$.

- The second method applies the *forward difference* discretization in space

$$q_i^{n+1} = q_i^n - a\frac{\Delta t}{\Delta x}\left(q_{i+1}^n - q_i^n\right). \tag{1.96}$$

On the contrary to the approach above, the scheme (1.96) leads to a *downwind* method if $a > 0$ and an *upwind* scheme if $a < 0$.

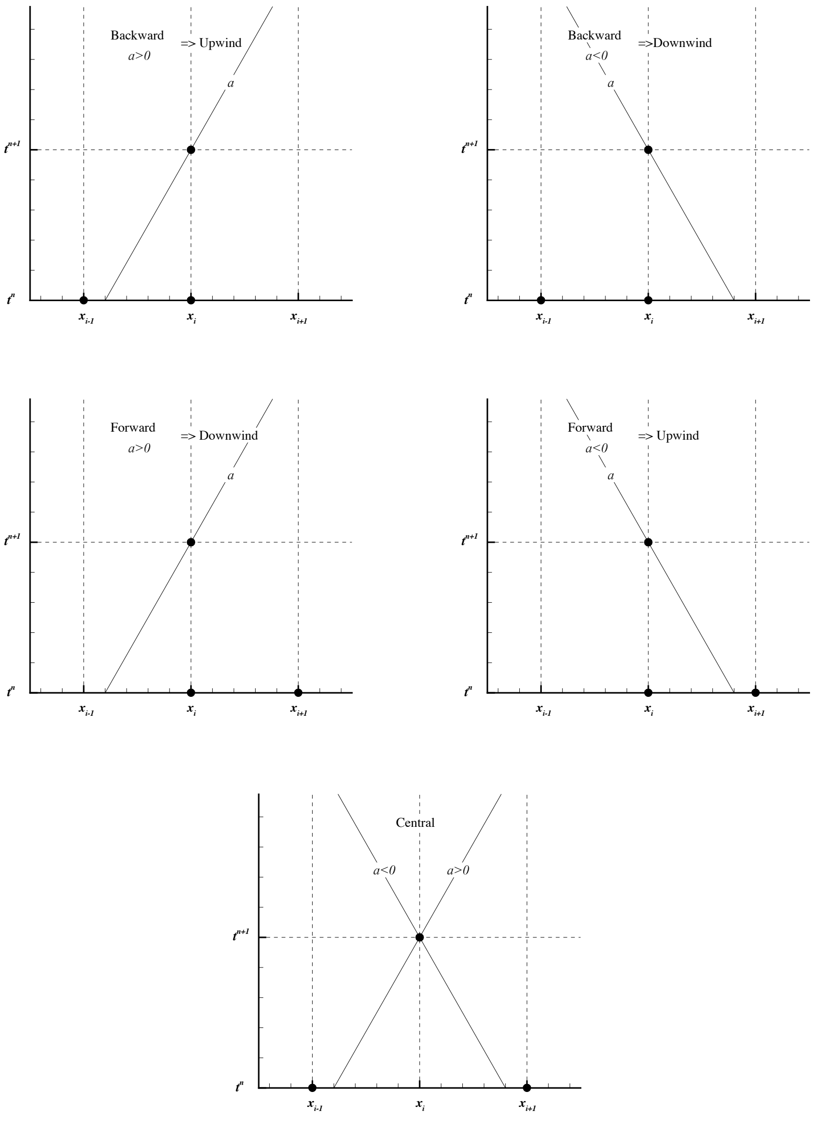

FIGURE 7. Visualization of the numerical methods derived from the finite difference method applying the three different approaches to discretize the space derivatives.

- The last scheme uses the *central difference* approach in space

$$q_i^{n+1} = q_i^n - a\frac{\Delta t}{2\Delta x}\left(q_{i+1}^n - q_{i-1}^n\right). \tag{1.97}$$

A sketch of the upwind or downwind property of the above numerical methods in function of the sign of the propagation speed $a$ is shown in Fig. 7.

As we will see later, not all of the above schemes are useful in practice since not all of the above schemes are stable.

To assure the stability of an explicit method for solving a linear evolutionary partial differential equation a constraint on the time step $\Delta t$ is required. It derives from the Courant-Friedrichs-Lewy (CFL) condition [**18**], that in the one

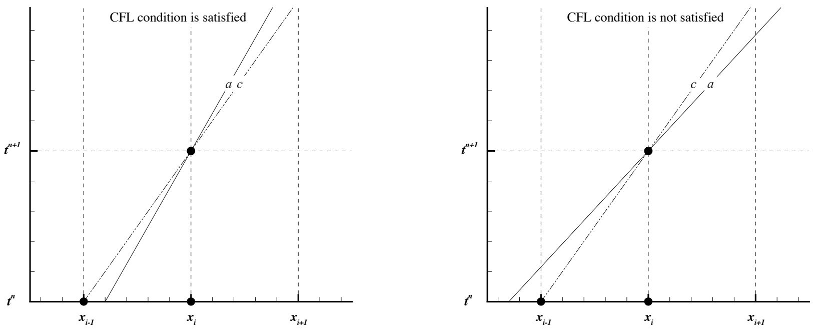

FIGURE 8. The plot of the CFL condition on the x-t plane.

dimensional case is defined by

$$c \leq 1, \quad \text{with } c = a\frac{\Delta t}{\Delta x}, \tag{1.98}$$

where $c$ is the so-called CFL number. The CFL condition leads to a restriction on the maximum admissible numerical time step at fixed grid size. The physical meaning of the CFL number is the measure for the progress of a disturbance over a time step $\Delta t$ related to the grid distance $\Delta x$. If the domain of dependence of the PDE due to the physical celerity is not contained in the domain of dependence of the discrete scheme given by the CFL number, then the numerical scheme will be unstable. This means that numerical information propagates slower than the physical wave, as shown in Fig. 8.

**5.2. The Lax-Friedrichs method.** However, the CFL condition is not sufficient for stability. In fact, via a von Neumann stability analysis shown later in this manuscript, it is possible to prove that *all* the central and downwind methods are unconditionally unstable for any CFL number $c$. Hence, Lax and Friedrichs [**52**] proposed a modification of the original central finite difference method replacing $q_i^n$ in (1.97) by the average of the backward and forward values, as follows

$$q_i^{n+1} = \frac{1}{2}\left(q_{i+1}^n + q_{i-1}^n\right) - a\frac{1}{2}\frac{\Delta t}{\Delta x}\left(q_{i+1}^n - q_{i-1}^n\right). \tag{1.99}$$

It can be proven that the scheme (1.99) is stable.

**5.3. The Lax-Wendroff method.** In this section we present the scheme introduced by Lax and Wendroff [**53**]. Let us consider the Taylor series in time of the function $q_i^{n+1}$ computed at the point $(x_i, t^n)$

$$q_i^{n+1} = q_i^n + \Delta t q_t + \frac{\Delta t^2}{2} q_{tt} + ... \tag{1.100}$$

The derivatives in time are obtained from the original PDE, in this case the linear advection equation (1.86), through the so-called Lax–Wendroff or Cauchy–Kovalewski procedure, as follows

$$q_t = -a q_x, \tag{1.101}$$

$$q_{tt} = -a q_{xt}, \quad q_{xt} = -a q_{xx} \quad \Rightarrow q_{tt} = a^2 q_{xx}. \tag{1.102}$$

Afterwards, the derivatives in space are approximated using the central finite difference approach

$$q_x \approx \frac{q_{i+1}^n - q_{i-1}^n}{2\Delta x}, \quad q_{xx} \approx \frac{q_{i+1}^n - 2q_i^n + q_{i-1}^n}{\Delta x^2}. \tag{1.103}$$

Finally, we substitute the spatial derivatives above into the Taylor expansion and obtain the second order Lax-Wendroff scheme

$$q_i^{n+1} = q_i^n - \frac{a}{2}\frac{\Delta t}{\Delta x}\left(q_{i+1}^n - q_{i-1}^n\right) + \frac{a^2 \Delta t^2}{2\Delta x^2}\left(q_{i+1}^n - 2q_i^n + q_{i-1}^n\right). \tag{1.104}$$

**5.4. The method of lines (MOL).** This technique for solving partial differential equations dates back to the 1960ies [**71**] and consists of a semi-discrete numerical method. It proceeds by first discretizing the spatial derivatives only (e.g. through the backward finite difference method) and leaving the time variable still continuous

$$\frac{\partial q}{\partial t} \approx -a\frac{q_i^n - q_{i-1}^n}{\Delta x}. \tag{1.105}$$

This leads to a system of ordinary differential equations ODE in which the unknown quantities are the $q_i$ at the grid points:

$$\frac{d}{dt}\vec{q}(t) = \vec{f}\left(\vec{q}(t)\right) \quad \text{with } \vec{q}(t) = \begin{pmatrix} q_1(t) \\ q_2(t) \\ ... \\ q_N(t) \end{pmatrix}. \tag{1.106}$$

Here, $N$ denotes the number of the nodes on the mesh. To discretize the system (1.106) in time we can use any numerical method for solving ODE, such as the explicit or implicit first order Euler schemes, the second order accurate semi–implicit Crank-Nicholson method, the explicit second order scheme of Heun, any higher order Runge-Kutta scheme or any method from the family of Adams–Bashforth and Adams–Moulton–type multi–stage methods.

**5.5. Implicit Euler method together with the backward finite difference discretization.** Applying the implicit Euler scheme coupled with the backward finite difference method to (1.106) we obtain the following discrete equation

$$q_i^{n+1} = q_i^n - a\frac{\Delta t}{\Delta x}\left(q_i^{n+1} - q_{i-1}^{n+1}\right). \tag{1.107}$$

This leads to a tridiagonal system of $N$ equations, as follows

$$q_i^{n+1}(1+c) - cq_{i-1}^n = q_i^n, \quad \text{for } i = 1, .., N \tag{1.108}$$

that can be written in matrix form as

$$\begin{pmatrix} c+1 & 0 & 0 & ... & 0 \\ -c & c+1 & 0 & ... & 0 \\ 0 & -c & c+1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & -c & c+1 \end{pmatrix} \begin{pmatrix} q_1^{n+1} \\ q_2^{n+1} \\ q_3^{n+1} \\ ... \\ q_N^{n+1} \end{pmatrix} = \begin{pmatrix} q_1^n + cb_L \\ q_2^n \\ q_3^n \\ ... \\ q_N^n \end{pmatrix}. \tag{1.109}$$

Here, we supposed $a > 0$ to introduce the left boundary condition.

**5.6. The Thomas algorithm.** The Thomas algorithm consists of a simplified form of the Gauss algorithm and is used to solve tridiagonal system of equations

$$\underline{\underline{A}} \cdot \vec{x} = \vec{b}, \tag{1.110}$$

where $\vec{x}$ is the vector of the unknowns. The tridiagonal matrix $\underline{\underline{A}}$ and the right–hand–side vector $\vec{b}$ are given by

$$\underline{\underline{A}} = \begin{pmatrix} d_1 & u_1 & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ l_2 & d_2 & u_2 & 0 & ... & 0 & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 0 & l_{N-1} & d_{N-1} & u_{N-1} \\ 0 & 0 & 0 & 0 & ... & 0 & 0 & l_N & d_N \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_{N-1} \\ b_N \end{pmatrix}, \tag{1.111}$$

where $N$ is the number of unknowns. The first step of the Thomas algorithm consists of a forward elimination:

(1.112)

```
$u_1 = u_1/d_1$
$b_1 = b_1/d_1$
for $i = 2 : N$
  $g = 1/(d_i - u_{i-1}l_i)$
  $u_i = u_i \cdot g$
  $b_i = g \cdot (b_i - l_i b_{i-1})$
end
```

Then, a backward substitution leads to the final solution of the system:

(1.113)

```
$x_N = b_N$
for $i = N - 1 : -1 : 1$
  $x_i = b_i - u_i x_{i+1}$
end
```

**5.7. Implicit Euler method together with the central finite difference discretization.** To solve the advection equation (1.106) we could also use the implicit Euler scheme together with a central finite difference method to compute the derivative in space

$$q_i^{n+1} = q_i^n - \frac{a\Delta t}{2\Delta x}\left(q_{i+1}^{n+1} - q_{i-1}^{n+1}\right), \tag{1.114}$$

which can also be written as

$$\frac{c}{2}q_{i+1}^{n+1} + q_i^{n+1} - \frac{c}{2}q_{i-1}^{n} = q_i^n, \quad \text{for } i = 1,..,N \tag{1.115}$$

In matrix form this results as

$$\underline{\underline{A}} = \begin{pmatrix} 1 & \frac{c}{2} & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ -\frac{c}{2} & 1 & \frac{c}{2} & 0 & ... & 0 & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 0 & -\frac{c}{2} & 1 & \frac{c}{2} \\ 0 & 0 & 0 & 0 & ... & 0 & 0 & 1 & \frac{c}{2} \end{pmatrix} \begin{pmatrix} q_1^{n+1} \\ q_2^{n+1} \\ ... \\ q_{N-1}^{n+1} \\ q_N^{n+1} \end{pmatrix} = \begin{pmatrix} q_1^n + \frac{c}{2}b_L \\ q_2^n \\ ... \\ q_{N-1}^n \\ q_N^n - \frac{c}{2}b_R \end{pmatrix}. \tag{1.116}$$

The linear tridiagonal system of equations above can be solved again using the Thomas algorithm presented in the previous section.

Example 6. *In the following we derive the Beam-Warming method. It is based on the same idea as the Lax-Wendroff scheme, but it uses an upwind stencil, i.e. it uses the points $x_i$, $x_{i-1}$ and $x_{i-2}$, assuming $a > 0$. We compute the finite difference approximations of the first and second space derivative based on these* *three points using a Taylor series expansion in space computed at the point* $(x_i, t^n)$*, as follows:*

$$q_{i-2}^n = q_i^n - 2\Delta x q_x + \frac{4\Delta x^2}{2} q_{xx} - ..., \tag{1.117}$$

$$q_{i-1}^n = q_i^n - \Delta x q_x + \frac{\Delta x^2}{2} q_{xx} - .... \tag{1.118}$$

*From* (1.117) *and* (1.118) *one obtains the finite difference expressions for the first and second spatial derivative as*

$$q_x = \frac{q_{i-2}^n - 4q_{i-1}^n + 3q_i^n}{2\Delta x}, \tag{1.119}$$

$$q_{xx} = \frac{q_{i-2}^n - 2q_{i-1}^n + q_i^n}{\Delta x^2}. \tag{1.120}$$

*Like in the Lax–Wendroff scheme the solution* $q_i^{n+1}$ *at the new time* $t^{n+1}$ *is given by a temporal Taylor series expansion as*

$$q_i^{n+1} = q_i^n + \Delta t q_t + \frac{\Delta t^2}{2} q_{tt} + ... \tag{1.121}$$

*Applying the Cauchy-Kovalewski procedure to the linear advection equation we obtain again*

$$q_t = -a q_x \qquad \text{and} \qquad q_{tt} = +a^2 q_{xx}. \tag{1.122}$$

*Replacing the time derivatives in the time Taylor series by space derivatives and inserting the finite difference approximations shown above leads to the Beam-Warming method*

$$q_i^{n+1} = q_i^n - \frac{a\Delta t}{2\Delta x}\left(q_{i-2}^n - 4q_{i-1}^n + 3q_i^n\right) + \frac{a^2\Delta t^2}{2\Delta x^2}\left(q_{i-2}^n - 2q_{i-1}^n + q_i^n\right). \tag{1.123}$$

Example 7. *The method of Fromm* [**33**] *is based on an arithmetic average of the Lax-Wendroff method and the Beam-Warming scheme. In this approach, the derivatives in space are given by an arithmetic average of the backward finite differences based on the stencil* $x_i$*,* $x_{i-1}$ *and* $x_{i-2}$ *and the central finite difference. We obtain*

$$q_{xx} = \frac{1}{2}\frac{q_{i-2}^n - 2q_{i-1}^n + q_i^n}{\Delta x^2} + \frac{1}{2}\frac{q_{i+1}^n - 2q_i^n + q_{i-1}^n}{\Delta x^2}. \tag{1.124}$$

*and*

$$q_x = \frac{1}{2}\frac{q_{i-2}^n - 4q_{i-1}^n + 3q_i^n}{2\Delta x} + \frac{1}{2}\frac{q_{i+1}^n - q_{i-1}^n}{2\Delta x}. \tag{1.125}$$

*Based on the same ideas as the ones used for deriving the Lax–Wendroff and the Beam–Warming schemes we obtain Fromm's method:*

$$q_i^{n+1} = q_i^n - \frac{a\Delta t}{4\Delta x}\left(q_{i-2}^n - 5q_{i-1}^n + 3q_i^n + q_{i+1}^n\right) + \frac{a^2\Delta t^2}{4\Delta x^2}\left(q_{i-2}^n - q_{i-1}^n - q_i^n + q_{i+1}^n\right). \tag{1.126}$$

*It is very interesting to note that the Fromm scheme is the least dispersive and hence the least oscillatory linear second order scheme.*

CHAPTER 2

# Analysis of Error, Stability and Monotonicity

## 1. Local discretization error

All the explicit finite difference (FD) methods presented in the previous section can be written in the compact form

$$q_i^{n+1} = H\left(q_{i-l}^n, q_{i-l+1}^n, ..., q_{i-1}^n, q_i^n, q_{i+1}^n, ..., q_{i+r-1}^n, q_{i+r}^n\right), \tag{2.1}$$

where the operator $H$ is a real function of $(l+r+1)$ variables and $l$ and $r$ are integer numbers that define the size of the *stencil* (support) of the numerical method.

EXAMPLE 8. *We consider the upwind finite difference method*

$$\begin{array}{ll} q_i^{n+1} = q_i^n - \frac{a\Delta t}{\Delta x}\left(q_i^n - q_{i-1}^n\right), & \text{for } a > 0, \\ q_i^{n+1} = q_i^n - \frac{a\Delta t}{\Delta x}\left(q_{i+1}^n - q_i^n\right), & \text{for } a < 0. \end{array} \tag{2.2}$$

*The two different discretizations above can be written together in the following unified form*

$$q_i^{n+1} = q_i^n - \frac{\Delta t}{\Delta x}\left(\frac{a-|a|}{2}q_{i+1}^n - \frac{a-|a|}{2}q_i^n\right) - \frac{\Delta t}{\Delta x}\left(\frac{a+|a|}{2}q_i^n - \frac{a+|a|}{2}q_{i-1}^n\right), \tag{2.3}$$

*or shorter as*

$$q_i^{n+1} = q_i^n - \frac{\Delta t}{\Delta x}a^-\left(q_{i+1}^n - q_i^n\right) - \frac{\Delta t}{\Delta x}a^+\left(q_i^n - q_{i-1}^n\right), \tag{2.4}$$

*where the $a^\pm$ are defined by $a^\pm = \frac{1}{2}(a \pm |a|)$. Hence, we have*

$$H = H\left(q_{i-1}^n, q_i^n, q_{i+1}^n\right) = \left(1 + \frac{\Delta t}{\Delta x}(a^- - a^+)\right)q_i^n + \frac{\Delta t}{\Delta x}a^+ q_{i-1}^n - \frac{\Delta t}{\Delta x}a^- q_{i+1}^n. \tag{2.5}$$

**1.1. Linear methods.** A numerical method is called *linear* if it can be written in the following form

$$q_i^{n+1} = \sum_{j=-l}^{r} b_j q_{i+j}^n, \tag{2.6}$$

where the coefficients $b_j$ do *not* depend on the solution $q_i^n$, but only on the parameters of the PDE and its discretization, i.e. on the advection speed $a$, on the mesh length $\Delta x$ and on the time step $\Delta t$.

**1.2. Operator form.** It is often useful to define a general numerical method as an operator $L_h$ acting on the discrete solution $q_h = q_h(x_h, t_h)$ in the form

$$L_h\left(q_h\right) = 0. \tag{2.7}$$

Using the notation (2.1) of a general explicit scheme we have

$$L_h\left(q_h\left(x_h, t_h\right)\right) = q_i^{n+1} - H\left(q_{i-l}^n, ..., q_{i-1}^n, q_i^n, q_{i+1}^n, ..., q_{i+r}^n\right) = 0, \tag{2.8}$$

where $x_h = (x_{i-l}, ..., x_{i-1}, x_i, x_{i+1}, ..., x_{i+r})$ and $t_h = \left(t^n, t^{n+1}\right)$ are the discretizations of space and time, respectively, and the numerical solution $q_h\left(x_h, t_h\right)$ is defined in the discrete points $(x,t) \in x_h \times t_h$.

**1.3. Local truncation error.** The local truncation error or the *local discretization error* is due to the truncation of the infinite Taylor series to form the discrete algorithm. It refers to one time step and is defined by

$$\tau = \frac{L_h\left(q\left(x_h, t_h\right)\right)}{\Delta t}, \tag{2.9}$$

where $q\left(x_h, t_h\right)$ is the *exact solution* of the PDE computed at the discrete point $\left(x_h, t_h\right)$. Here, the exact solution $q\left(x, t\right)$ is expanded using a Taylor series in space and time with respect to the point $\left(x_i, t^n\right)$.

Definition 1. *A method is called $k$-th order accurate in space and $n$-th order accurate in time if*

$$\tau = \mathcal{O}(\Delta x^k) + \mathcal{O}(\Delta t^n). \tag{2.10}$$

Example 9. *We consider the upwind finite difference method applied to the linear scalar advection equation assuming $a > 0$. The related numerical operator is given by*

$$L_h\left(q_h\left(x_h, t_h\right)\right) = q_i^{n+1} - q_i^n + \frac{a\Delta t}{\Delta x}\left(q_i^n - q_{i-1}^n\right) = 0. \tag{2.11}$$

*The exact solution is given by the Taylor series expansion as follows:*

$$q_i^{n+1} = q_i^n + \Delta t q_t + \frac{\Delta t^2}{2} q_{tt} + \mathcal{O}(\Delta t^3), \tag{2.12}$$

$$q_{i-1}^n = q_i^n - \Delta x q_x + \frac{\Delta x^2}{2} q_{xx} - \mathcal{O}(\Delta x^3). \tag{2.13}$$

*Now, we compute the local truncation error*

$$\begin{aligned} \tau = \ & \tfrac{1}{\Delta t}\left(q_i^n + \Delta t q_t + \tfrac{\Delta t^2}{2} q_{tt} + \mathcal{O}\left(\Delta t^3\right) - q_i^n\right) + \\ & \tfrac{1}{\Delta t}\tfrac{a\Delta t}{\Delta x}\left(q_i^n - q_i^n + \Delta x q_x - \tfrac{\Delta x^2}{2} q_{xx} - \mathcal{O}\left(\Delta x^3\right)\right), \end{aligned} \tag{2.14}$$

*which results in*

$$\tau = \overbrace{q_t + a q_x}^{=0} + \frac{\Delta t}{2} q_{tt} - \frac{a\Delta x}{2} q_{xx} + \mathcal{O}\left(\Delta x^2\right) + \mathcal{O}\left(\Delta t^2\right). \tag{2.15}$$

*Here, the first term vanishes because the exact solution $q$ obviously must satisfy the PDE exactly. We therefore obtain the final result*

$$\tau = \frac{\Delta t}{2} q_{tt} - \frac{a\Delta x}{2} q_{xx} + \mathcal{O}\left(\Delta x^2\right) + \mathcal{O}\left(\Delta t^2\right) = \mathcal{O}\left(\Delta t\right) + \mathcal{O}\left(\Delta x\right). \tag{2.16}$$

*From the leading error terms (lowest exponents of $\Delta x$ and $\Delta t$ in the above equation) we find that they are proportional to $\Delta x^1$ and $\Delta t^1$, hence the upwind finite difference scheme is of first order of accuracy in space and time.*

Example 10. *In the following, we compute the local truncation error of the central finite difference scheme. The numerical operator is defined by*

$$L_h = q_i^{n+1} - q_i^n + \frac{a\Delta t}{2\Delta x}\left(q_{i+1}^n - q_{i-1}^n\right) = 0. \tag{2.17}$$

*We apply the Taylor series expansion in space to the term $q_{i+1}^n$*

$$q_{i+1}^n = q_i^n + \Delta x q_x + \frac{\Delta x^2}{2} q_{xx} + \frac{\Delta x^3}{6} q_{xxx} + ... \tag{2.18}$$

*The same procedure is carried out for $q_n^{n+1}$ and $q_{i-1}^n$, see Eqns.* (2.12) *and* (2.13)*. Then, the local truncation error related to the central FD method results as*

$$\begin{aligned}\tau = \; & \tfrac{1}{\Delta t}\left(q_i^n + \Delta t q_t + \tfrac{\Delta t^2}{2} q_{tt} + \mathcal{O}\left(\Delta t^3\right) - q_i^n\right) + \\ & \tfrac{1}{\Delta t}\tfrac{a\Delta t}{2\Delta x}\left(+q_i^n + \Delta x q_x + \tfrac{\Delta x^2}{2} q_{xx} + \tfrac{\Delta x^3}{6} q_{xxx} + \mathcal{O}\left(\Delta x^4\right)\right) + \\ & \tfrac{1}{\Delta t}\tfrac{a\Delta t}{2\Delta x}\left(-q_i^n + \Delta x q_x - \tfrac{\Delta x^2}{2} q_{xx} + \tfrac{\Delta x^3}{6} q_{xxx} - \mathcal{O}\left(\Delta x^4\right)\right).\end{aligned} \tag{2.19}$$

*Simplifying the relation above we obtain*

$$\tau = \overbrace{q_t + a q_x}^{=0} + \frac{\Delta t}{2} q_{tt} + \frac{a \Delta x^2}{6} q_{xxx} + \mathcal{O}\left(\Delta x^3\right) + \mathcal{O}\left(\Delta t^2\right) = \mathcal{O}\left(\Delta t\right) + \mathcal{O}\left(\Delta x^2\right). \tag{2.20}$$

*Hence, the central finite difference scheme is of first order of accuracy in time and of second order in space. However, it is not stable.*

Example 11. *In the case of the Lax-Friedrichs scheme* (1.99) *the numerical operator is given by*

$$L_h = q_i^{n+1} - \frac{1}{2}\left(q_{i+1}^n + q_{i-1}^n\right) + \frac{a\Delta t}{2\Delta x}\left(q_{i+1}^n - q_{i-1}^n\right) = 0. \tag{2.21}$$

*We compute the local truncation error using Taylor series expansions of the exact solution in space and time, as in the examples above:*

$$\begin{aligned}\tau = \; & \tfrac{1}{\Delta t}\left(q_i^n + \Delta t q_t + \tfrac{\Delta t^2}{2} q_{tt} + \mathcal{O}\left(\Delta t^3\right)\right) - \\ & \tfrac{1}{\Delta t}\tfrac{1}{2}\left(q_i^n + \Delta x q_x + \tfrac{\Delta x^2}{2} q_{xx} + \mathcal{O}\left(\Delta x^3\right) + q_i^n - \Delta x q_x + \tfrac{\Delta x^2}{2} q_{xx} + \mathcal{O}\left(\Delta x^3\right)\right) + \\ & \tfrac{1}{\Delta t}\tfrac{a\Delta t}{2\Delta x}\left(q_i^n + \Delta x q_x + \tfrac{\Delta x^2}{2} q_{xx} + \mathcal{O}\left(\Delta x^3\right) - q_i^n + \Delta x q_x - \tfrac{\Delta x^2}{2} q_{xx} + \mathcal{O}\left(\Delta x^3\right)\right).\end{aligned} \tag{2.22}$$

*After some manipulations, using the definition of the Courant number $c = a\Delta t/\Delta x$ the local truncation error becomes*

$$\tau = \overbrace{q_t + a q_x}^{=0} + \frac{\Delta t}{2} q_{tt} - \frac{a\Delta x}{2c} q_{xx} + \mathcal{O}\left(\Delta x^2\right) + \mathcal{O}\left(\Delta t^2\right) = \mathcal{O}\left(\Delta t\right) + \mathcal{O}\left(\Delta x\right). \tag{2.23}$$

*This proves that the Lax-Friedrichs scheme is of first order of accuracy in space and in time.*

Exercise 2. *Compute the local truncation error of the Lax–Wendroff scheme, the Beam–Warming method and the scheme of Fromm.*

## 2. The modified equation approach or the method of differential approximation

In general, the numerical solution $q_h$ does *not* satisfy the original PDE

$$\frac{\partial q}{\partial t} + a\frac{\partial q}{\partial x} = 0 \tag{2.24}$$

exactly, but only approximately, i.e. in general we have

$$\frac{\partial q_h}{\partial t} + a\frac{\partial q_h}{\partial x} \neq 0. \tag{2.25}$$

To quantify the errors, we will consider the so–called *modified equation* or *equivalent differential equation*

$$\frac{\partial q_h}{\partial t} + a\frac{\partial q_h}{\partial x} = \sum_{l=2}^{\infty} c_l \frac{\partial^l q_h}{\partial x_l}, \tag{2.26}$$

which is solved *exactly* by the numerical method. A detailed analysis of the modified equation (2.26) gives very important insight into the numerical scheme, in particular concerning its stability, the accuracy and its diffusive and dispersive behaviour. The error coefficients $c_l$ in equation (2.26) are coefficients depending on the advection speed $a$, on the mesh size $\Delta x$ and on the time step.

**2.1. Computation of the modified equation.** The modified equation analysis or the so–called method of differential approximation goes back to the work of the Russian mathematician Y.I. Shokin [**73**] and the US mathematicians Warming and Hyett [**90**]. From the general form of the numerical scheme

$$q_i^{n+1} = H(q_{i-l}^n, ..., q_i^n, ..., q_{i+r}^n), \tag{2.27}$$

the modified equation is computed by expanding the discrete solution $q_h$ at the discrete points $x_h$ and $t_h$ in a Taylor series in space and time with respect to point $x_i$ and $t^n$, similar to the local truncation error analysis, i.e. we obtain

$$\begin{aligned} q + \Delta t q_t + \frac{1}{2}\Delta t^2 q_{tt} + \frac{1}{6}\Delta t^3 q_{ttt} + ... = \\ H(q - l\Delta x q_x + \frac{1}{2}(l\Delta x)^2 q_{xx} - ..., ..., q, ..., q + r\Delta x q_x + \frac{1}{2}(r\Delta x)^2 q_{xx} + ...). \end{aligned} \tag{2.28}$$

Eqn. (2.28) is the so–called Γ–form of the differential approximation of the scheme (2.27), where still higher order time derivatives appear. In order to obtain the final form of the differential approximation, we must then use the Γ–form (2.28) in order to replace time derivatives and mixed space–time derivatives by pure space derivatives using successive differentiation of (2.28) with respect to $x$ and $t$. This procedure, when applied to Eqn. (2.28) is the so–called Warming–Hyett procedure [**90**], which is similar to the Cauchy-Kovalewski procedure, where successive differentiation with respect to $x$ and $t$ is applied to the original PDE. Due to the increased complexity of the PDE (2.28) with respect to the original PDE, the Warming–Hyett procedure is algebraically *much* more complicated than the Cauchy–Kovalewski procedure. To get expressions which can be handled, one must determine a priori the maximum order of $\Delta x$ and $\Delta t$ in the modified equation and all higher order terms are immediately dropped as soon as they appear during the Warming–Hyett procedure. The final result of the method of differential approximation is then the so–called Π–form of the differential approximation (2.26), which contains the original PDE on the left hand side and the error terms involving only pure space derivatives on the right hand side:

$$\frac{\partial q_h}{\partial t} + a\frac{\partial q_h}{\partial x} = \sum_{l=2}^{\infty} c_l \frac{\partial^l q_h}{\partial x_l}.$$

**2.2. Physical interpretation of the error.** The terms on the right–hand side of (2.26) provide in particular a very interesting *physical* interpretation of the error using the harmonic analysis of Fourier applied to (2.26). We consider the evolution of one single Fourier mode (monochromatic wave) as follows:

$$q\left(x,t\right) = q_0 e^{i(kx-\omega t)}, \quad \text{with} \quad k = \frac{2\pi}{\lambda},\ \omega = \frac{2\pi}{T},\ i^2 = -1, \tag{2.29}$$

where $k$ denotes the wave number and $\omega$ is the angular frequency. Its derivatives in space and in time are the following

$$q_t = -i\omega q, \quad q_x = ikq, \quad \frac{\partial^l q}{\partial x^l} = \left(ik\right)^l q. \tag{2.30}$$

Substitution of these relations into (2.26) leads to the so-called *dispersion relation* of the PDE (2.26)

$$-i\omega + aik = \sum_{l=2}^{\infty} c_l \left(ik\right)^l. \tag{2.31}$$

Then, the angular frequency $\omega$ is defined by

$$\omega = ak + i\sum_{l=2}^{\infty} c_l\,(ik)^l\,, \quad \text{with } i^l = \begin{cases} (-1)^m\,, & \text{if } l = 2m, \\ i\,(-1)^m\,, & \text{if } l = 2m+1. \end{cases} \tag{2.32}$$

Hence, we obtain the final expression of the dispersion relation:

$$\omega = ak + i\sum_{m=1}^{\infty} (-1)^m\, c_{2m} k^{2m} - \sum_{m=1}^{\infty} (-1)^m\, c_{2m+1} k^{2m+1}. \tag{2.33}$$

The *exact* solution of (2.26) for one single Fourier mode is given by

$$q\,(x,t) = q_0 e^{ik\cdot\left(x - t\left(a - \overbrace{\sum_{m=1}^{\infty}(-1)^m\, c_{2m+1} k^{2m}}^{\text{dispersion error}}\right)\right)} \cdot e^{\overbrace{t\cdot\sum_{m=1}^{\infty}(-1)^m\, c_{2m} k^{2m}}^{\text{diffusion error}}}\,, \tag{2.34}$$

where the dispersion and diffusion error terms can be clearly identified. Relation (2.34) can be written as

$$q\,(x,t) = q_0 \cdot e^{-dt} \cdot e^{i(k(x - v(k)t))}, \tag{2.35}$$

where $d$ represents the diffusion error and $v(k)$ is the wave speed, which in this case depends on the wavenumber $k$.

The physical phenomenon of *dispersion* means that the wave speed is a function of the wavenumber $k$. An example of a dispersive phenomenon is the formation of a rainbow, which is caused by the separation of white light into different components of different wavelengths (different colors) due to optical dispersion. In our case, the velocity of one single Fourier mode is given by

$$v\,(k) = a - \sum_{m=1}^{\infty} (-1)^m\, c_{2m+1} k^{2m}, \tag{2.36}$$

instead of being

$$v\,(k) = a, \tag{2.37}$$

which would be the correct wave speed for the original linear scalar advection equation (3.1). We note in particular that dispersive errors are introduced only by the coefficients $c_l$ in front of the *odd* derivatives in the errors terms of the equivalent PDE, i.e. by the coefficients $c_3$, $c_5$, etc. In contrast to dispersion, which modifies the wave speed but not the amplitude, the diffusion error causes a reduction or amplification of the amplitude of the wave during its propagation, but does not have any influence on the wave speed. From (2.34) we clearly see that the two effects are separated. The diffusion error is quantified by the term

$$d = -\sum_{m=1}^{\infty} (-1)^m\, c_{2m} k^{2m}, \tag{2.38}$$

which must be positive or zero for a scheme to be stable, i.e. the amplitudes of all the Fourier modes must be *non–increasing.* As we can clearly see from this equation, the diffusion error is only governed by the error coefficients $c_l$ in front of the *even* spatial derivatives, i.e. by the coefficients $c_2$, $c_4$, etc. It is important to note the alternating sign in Eqn. (2.38), which means that for stability we must have $c_2 \geq 0$ but $c_4 \leq 0$ and so on.

EXAMPLE 12. *We consider the backward finite difference (FD) method* (1.95) *applied to the linear advection equation. Let us assume $a > 0$, hence, the* *backward FD scheme is upwind. Using Taylor series expansions of* $q_{i-1}^n$ *and* $q_i^{n+1}$ *we obtain*

(2.39)
$$q_i^n+\Delta t q_t+\frac{\Delta t^2}{2}q_{tt}+\mathcal{O}\left(\Delta t^3\right)=q_i^n-\frac{a\Delta t}{\Delta x}\left(q_i^n-q_i^n+\Delta x q_x-\frac{\Delta x^2}{2}q_{xx}+\mathcal{O}\left(\Delta x^3\right)\right).$$

*Neglecting the terms of higher oder in* (2.39)*, we obtain the* $\Gamma$*–form of the differential approximation*

$$q_t+aq_x=\frac{a\Delta x}{2}q_{xx}-\frac{\Delta t}{2}q_{tt}. \tag{2.40}$$

*To obtain the* $\Pi$*–form* (2.26) *from* (2.40) *we compute the second derivative in time* $q_{tt}$ *via the procedure of Warming-Hyett deriving* (2.40) *in time and in space:*

$$\begin{aligned} q_{tt} &= -aq_{xt}+\tfrac{a\Delta x}{2}q_{xxt}-\tfrac{\Delta t}{2}q_{ttt}\\ q_{tx} &= -aq_{xx}+\tfrac{a\Delta x}{2}q_{xxx}-\tfrac{\Delta t}{2}q_{ttx}. \end{aligned} \tag{2.41}$$

*Substituting* $q_{xt}$ *into the relation for the second derivative in time we can write* $q_{tt}$ *in terms of pure spatial derivatives*

$$q_{tt}=a^2q_{xx}+\mathcal{O}\left(\Delta x,\Delta t\right). \tag{2.42}$$

*Here, we stop at zeroth order terms in the Warming–Hyett procedure since we are only interested in the leading error term of the scheme. If one wanted to obtain also the higher order error terms in the differential approximation of the scheme, then also higher order terms would have to be retained during the Warming–Hyett procedure. We obtain*

$$q_t+aq_{xx}=\frac{a\Delta x}{2}q_{xx}-\frac{a^2\Delta t}{2}q_{xx}=\frac{a\Delta x}{2}\left(1-c\right)q_{xx}, \tag{2.43}$$

*hence the leading error coefficient of the modified equation is*

$$c_2=\frac{a\Delta x}{2}\left(1-c\right),\quad \text{with the Courant number } c=\frac{a\Delta t}{\Delta x}.$$

*For the solution to be stable, the coefficient must be positive, i.e. we obtain the following result:*

$$c_2\geq 0 \quad\Rightarrow\quad c\leq 1, \tag{2.44}$$

*which is the classical Courant–Friedrichs–Lewy stability condition* [**18**]*. We find that the scheme is first order accurate and that the diffusion error is the smaller the larger the Courant number (within the stability limit, of course). For* $c=1$ *the error term vanishes, and one can show that also all the higher order error terms vanish. In this very particular case of* $c=1$*, the upwind scheme reproduces the exact solution of the PDE.*

Example 13. *Now, we study the forward finite difference (FD) method* (1.96) *for the linear scalar advection equation with* $a>0$*, leading to a downwind scheme. Using appropriate Taylor series expansions with respect to the point* $(x_i,t^n)$ *we obtain*

$$q_i^n+\Delta t q_t+\frac{\Delta t^2}{2}q_{tt}=q_i^n-\frac{a\Delta t}{\Delta x}\left(q_i^n+\Delta x q_x-\frac{\Delta x^2}{2}-q_i^n\right), \tag{2.45}$$

*which simplifies to*

$$q_t+aq_x=-\frac{a\Delta x}{2}q_{xx}-\frac{\Delta t}{2}q_{tt}. \tag{2.46}$$

*Applying the procedure of Warming-Hyett, similar to the previous example, we obtain*

$$q_{tt}=a^2q_{xx}+\mathcal{O}\left(\Delta x,\Delta t\right), \tag{2.47}$$

*which substituted into* (2.46) *provides the following* $\Pi$*–form of the differential approximation:*

$$q_t + aq_{xt} = c_2 q_{xx}, \quad \text{with } c_2 = -\frac{a\Delta x}{2}\left(1+c\right). \tag{2.48}$$

*Since for positive a also the Courant number c is always positive, we immediately obtain that*

$$c_2 < 0, \quad \forall c > 0, \tag{2.49}$$

*hence the downwind scheme is unconditionally unstable.*

Example 14. *The aim in the following is to analyze the implicit backward finite difference scheme assuming* $a > 0$*,*

$$q_i^{n+1} = q_i^n - \frac{a\Delta t}{\Delta x}\left(q_i^{n+1} - q_{i-1}^{n+1}\right). \tag{2.50}$$

*The Taylor series expansion in space and time of* $q_{i-1}^{n+1}$ *yields*

$$q_{i-1}^{n+1} = q_i^n - \Delta x q_x + \Delta t q_t - \Delta x \Delta t q_{xt} + \frac{\Delta x^2}{2} q_{xx} + \frac{\Delta t^2}{2} q_{tt}. \tag{2.51}$$

*Substituting the Taylor expansions of* $q_{i-1}^{n+1}$ *and* $q_i^{n+1}$ *into* (2.50) *we obtain*

$$q_t + aq_x = -\frac{\Delta t}{2} q_{tt} - a\Delta t q_{xt} + \frac{a\Delta x}{2} q_{xx}. \tag{2.52}$$

*The Warming-Hyett procedure applied to the relation above provides*

$$q_{tx} = -aq_{xx} + \mathcal{O}\left(\Delta x, \Delta t\right), \qquad q_{tt} = -aq_{xt} + \mathcal{O}\left(\Delta x, \Delta t\right). \tag{2.53}$$

*Finally, the modified equation of the PDE up to first order error terms is given by*

$$q_t + aq_x = c_2 q_{xx}, \quad \text{with } c_2 = \frac{a\Delta x}{2}\left(1+c\right). \tag{2.54}$$

*The positivity condition of the coefficient* $c_2$ *is always satisfied*

$$c_2 > 0, \quad \forall c > 0, \tag{2.55}$$

*hence the scheme is unconditionally stable.*

Example 15. *The explicit central finite difference scheme* (1.97) *is analyzed. Using appropriate Taylor series expansions the* $\Gamma$*–form of the differential approximation results as*

$$q_t + aq_x = -\frac{a\Delta x^2}{6} q_{xxx} - \frac{\Delta t}{2} q_{tt} - \frac{\Delta t^2}{6} q_{ttt}. \tag{2.56}$$

*From the Warming-Hyett procedure we obtain the higher order time derivatives as follows:*

$$\begin{aligned} q_{ttt} &= -a^3 q_{xxx} + \mathcal{O}\left(\Delta x, \Delta t\right), \\ q_{tt} &= a^2 q_{xx} + \mathcal{O}\left(\Delta x, \Delta t\right). \end{aligned} \tag{2.57}$$

*The modified equation of the PDE results as*

$$q_t + aq_x = c_2 q_{xx} + c_3 q_{xxx}, \quad \text{with } c_2 = -c\frac{a\Delta x}{2}, \quad c_3 = \frac{a\Delta x^2}{6}\left(c^2 - 1\right). \tag{2.58}$$

*For the coefficient* $c_2$ *we obtain*

$$c_2 < 0, \quad \forall c > 0, \tag{2.59}$$

*hence the explicit central finite difference scheme is unconditionally unstable.*

EXERCISE 3. *Prove that the equivalent differential equation for the Lax–Wendroff method applied to the linear scalar advection equation is up to third order error terms given by*

$$q_t + aq_x = \frac{1}{6}a\Delta x^2(c^2-1)q_{xxx} + \frac{1}{8}a\Delta x^3 c(c^2-1)q_{xxxx}. \tag{2.60}$$

## 3. The von Neumann stability analysis

The stability of numerical schemes can also be investigated by carrying out the so–called *von Neumann* stability analysis. The basic idea of this procedure is to assume that the discrete solution can be developed in a Fourier series, instead of the Taylor expansions used in the method of differential approximation. The numerical solution is defined by a component of the Fourier series, as follows

$$q_j^n = \hat{Q}^n e^{i\theta j}, \quad \text{with } i^2 = -1, \quad \theta = k\Delta x = 2\pi\frac{\Delta x}{\lambda_k}, \tag{2.61}$$

where the integer $j$ denotes the spatial index (to avoid confusion with the imaginary unit $i$), $\hat{Q}^n$ is the amplitude at time $t^n$, $0 \le \theta \le \pi$ is an angle, $k$ is the wave number and $\lambda_k \ge 2\Delta x$ is the associated wave length. According to the Shannon theorem the smallest wavelength of a periodic function that can still be represented on a discrete mesh of size $\Delta x$ is $\lambda_{\min} = 2\Delta x$. Substituting the definition (2.61) in the numerical scheme we obtain the amplification factor $G$ given by

$$G = \frac{\hat{Q}^{n+1}}{\hat{Q}^n}, \quad \text{with } G \in \mathbb{C}. \tag{2.62}$$

It defines the ratio of the wave amplitudes at the new and the old time, respectively. In order to obtain a *stable* scheme, the wave amplitude must be *non–increasing*, i.e. the following condition must hold for stability:

$$|G| \le 1, \tag{2.63}$$

which is equivalent to

$$|G|^2 = \Re(G)^2 + \Im(G)^2 \le 1. \tag{2.64}$$

EXAMPLE 16. *We propose an application of the von Neumann stability analysis to the explicit backward finite difference method* (1.95)*, that is upwind in the case of* $a > 0$*. First, we replace the function* $q_j^n$ *with the Fourier mode* (2.61)

$$\hat{Q}^{n+1}e^{i\theta j} = \hat{Q}^n e^{i\theta j} - c\hat{Q}^n e^{i\theta j} + c\hat{Q}^n e^{i\theta(j-1)}. \tag{2.65}$$

*Subsequently, we compute the amplification factor* $G$

$$G = \frac{\hat{Q}^{n+1}}{\hat{Q}^n} = 1 - c + ce^{-i\theta} = 1 - c + c(\cos\theta - i\sin\theta), \tag{2.66}$$

*and the square of its absolute value is given by the sum of the square of the real and the imaginary part of G, i.e. for stability we must have*

$$|G|^2 = (1 - c + c\cos\theta)^2 + c^2\sin^2\theta = 1 - 2c(1-c)(1-\cos\theta) \le 1. \tag{2.67}$$

*In order to verify the stability criterion* (2.63) *it is equivalent to write*

$$\overbrace{2c}^{>0}\,(1-c)\,\overbrace{(1-\cos\theta)}^{\ge 0} \ge 0, \tag{2.68}$$

*which becomes a condition on the CFL number*

$$c \le 1. \tag{2.69}$$

*This result is again the famous CFL condition for the Courant number* [**18**] *and is identical with the result obtained by the method of differential approximation shown previously. This condition guarantees that small perturbations do not grow in time.*

EXAMPLE 17. *In this example we carry out the von Neumann stability analysis of the explicit central finite difference scheme* (1.97) *proving that it is unconditionally unstable. Substitution of the Fourier ansatz* (2.61) *into the numerical scheme* (1.97) *yields*

$$\hat{Q}^{n+1}e^{i\theta j} = \hat{Q}^{n}e^{i\theta j} - \hat{Q}^{n}\frac{c}{2}\left(e^{i\theta(j+1)} - e^{i\theta(j-1)}\right), \tag{2.70}$$

*hence the amplification factor $G$ becomes*

$$G = \frac{\hat{Q}^{n+1}}{\hat{Q}^{n}} = 1 - \frac{c}{2}\left(e^{i\theta} - e^{-i\theta}\right) = \left(1 - c\,i\sin\theta\right). \tag{2.71}$$

*We obtain the square of the absolute value of $G$ as*

$$|G|^2 = \Re(G)^2 + \Im(G)^2 = 1 + c^2\sin^2\theta \geq 1, \quad \forall c > 0. \tag{2.72}$$

*From* (2.72) *we can conclude that the stability condition* (2.63) *is never satisfied, hence the explicit central finite difference method is unconditionally unstable.*

EXAMPLE 18. *We consider the explicit forward finite difference scheme* (1.96) *and assume $a > 0$. Then, the numerical method is downwind. Using the von Neumann method we obtain*

$$\hat{Q}^{n+1}e^{i\theta j} = \hat{Q}^{n}e^{i\theta j} - c\hat{Q}^{n}\left(e^{i\theta(j+1)} - e^{i\theta j}\right). \tag{2.73}$$

*The amplification factor $G$ is then given by*

$$G = \frac{\hat{Q}^{n+1}}{\hat{Q}^{n}} = 1 - c(e^{i\theta} - 1) = 1 + c - c(\cos\theta + i\sin\theta). \tag{2.74}$$

*The stability criterion for the square of the absolute value of the amplification factor is then*

$$|G|^2 = \left((1+c) - c\cos\theta\right)^2 + c^2\sin^2\theta = 1 + 2c\,(1+c)\,(1-\cos\theta) \leq 1. \tag{2.75}$$

*It is obvious that the stability criterion* (2.63) *is not satisfied*

$$\overbrace{2c}^{>0}\,\overbrace{(1+c)}^{>0}\,\overbrace{(1-\cos\theta)}^{\geq 0} \leq 0. \tag{2.76}$$

*The explicit downwind finite difference scheme is therefore unconditionally unstable.*

EXAMPLE 19. *We analyze the stability of the Lax-Wendroff scheme* (1.104). *Substituting the Fourier ansatz* (2.61) *into* (1.104) *we obtain*

$$\hat{Q}^{n+1}e^{i\theta j} = \hat{Q}^{n}e^{i\theta j} - \frac{c}{2}\hat{Q}^{n}\left(e^{i\theta(j+1)} - e^{i\theta(j-1)}\right) + \frac{c^2}{2}\hat{Q}^{n}\left(e^{i\theta(j+1)} - 2e^{i\theta j} + e^{i\theta(j-1)}\right). \tag{2.77}$$

*To verify the stability of the method we compute again the amplification factor, which reduces to*

$$G = \left(1 - c^2\right) + c^2\cos\theta - i\,c\sin\theta, \tag{2.78}$$

*and the square of its absolute value must satisfy*

$$|G|^2 = 1 + c^2\left(1 - \cos^2\theta\right) - 2c^2\left(1 - \cos\theta\right) + c^4\left(1 - \cos\theta\right)^2 \leq 1, \tag{2.79}$$

*that is equivalent to*

$$1 + \overbrace{c^2\left(1 - \cos\theta\right)^2}^{\geq 0}\left(c^2 - 1\right) \leq 1. \tag{2.80}$$

*The stability condition for the Lax-Wendroff scheme* (1.104) *is verified when*

$$c \leq 1, \tag{2.81}$$

*i.e. the scheme is stable under the classical CFL condition.*

Exercise 4. *Prove that the Beam–Warming method for the linear scalar advection equation with* $a > 0$ *is stable up to Courant number two. Hint: the square of the absolute value of the amplification factor should result as* $|G|^2 = 1 + \alpha(1 - \cos\theta)^2$ *with* $\alpha = c^4 - 4c^3 + 5c^2 - 2c = c(c-2)(c-1)^2$.

## 4. Monotonicity and the Godunov theorem

We consider a numerical method written in the compact form

$$q_i^{n+1} = H\left(q_{i-l}^n, q_{i-l+1}^n, ..., q_{i-1}^n, q_i^n, q_{i+1}^n, ..., q_{i+r-1}^n, q_{i+r}^n\right), \tag{2.82}$$

where $H$ can also be a *nonlinear* operator. The scheme is called monotone if $H$ is a non-decreasing function of all its arguments, i.e. when

$$\frac{\partial H}{\partial q_k^n} \geq 0, \quad i - l \leq k \leq i + r. \tag{2.83}$$

From the definition above it follows that a linear method given by (2.6) is monotone if and only if all its coefficients $b_k$ in (2.6) are non-negative. We therefore must have

$$b_k \geq 0, \quad \forall -l \leq k \leq +r. \tag{2.84}$$

Example 20. *We consider the backward finite difference scheme* (1.96), *which is upwind for* $a > 0$*:*

$$q_i^{n+1} = H\left(q_{i-1}^n, q_i^n\right) = q_i^n\left(1 - c\right) + q_{i-1}^n c. \tag{2.85}$$

*According to definition* (2.83) *we differentiate the function* $H$ *with respect to its two arguments* $q_{i-1}^n$ *and* $q_i^n$ *and obtain the following criteria for monotonicity:*

$$\frac{\partial H}{\partial q_{i-1}^n} = c \geq 0, \quad \frac{\partial H}{\partial q_i^n} = 1 - c \geq 0. \tag{2.86}$$

*While* $c \geq 0$ *is always verified, the condition* $1 - c \geq 0$ *requires* $c \leq 1$ *for monotonicity, which is the same condition required for the stability of the scheme, see* (2.69)*. In other words: if the CFL condition is satisfied, the function* $H$ *is always non– decreasing and thus the explicit upwind backward FD method is monotone.*

Lemma 1. *(Roe) A linear method written in form* (2.6) *is of order* $p > 0$, *in space and time in the sense of the local discretization error*

$$\tau = \frac{L_h}{\Delta t} = \mathcal{O}\left(\Delta x^p, \Delta t^p\right), \tag{2.87}$$

*if and only if*

$$\sum_{k=-l}^{r} b_k k^m = \left(-c\right)^m, \quad \forall m = 0, 1, ..., p \quad \text{with } c = \frac{a\Delta t}{\Delta x},\ 0^0 := 1. \tag{2.88}$$

Proof. The operator of a general linear scheme is given by

$$L_h\left(q_h\left(x_h, t_h\right)\right) = q_i^{n+1} - \sum_{k=-l}^{r} b_k q_{i+k}^n = 0. \tag{2.89}$$

We compute the local truncation error of the method and replace the expressions $q_i^{n+1}$ and $q_{i+k}^n$ in (2.89) with appropriate Taylor series expansions in time and space, computed with respect to the point $(x_i, t^n)$:

$$\tau = \frac{L_h}{\Delta t} = \frac{1}{\Delta t}\left[q_i^n + \sum_{m=1}^{\infty} \frac{\Delta t^m}{m!}\frac{\partial^m q}{\partial t^m} - \sum_{k=-l}^{r} b_k\left(q_i^n + \sum_{m=1}^{\infty}\frac{\left(k\Delta x\right)^m}{m!}\frac{\partial^m q}{\partial x^m}\right)\right], \tag{2.90}$$

which is equivalent to

$$\tau = \frac{1}{\Delta t}\left[q_i^n - \sum_{k=-l}^{r} b_k q_i^n + \sum_{m=1}^{\infty}\frac{\Delta t^m}{m!}\frac{\partial^m q}{\partial t^m} - \sum_{k=-l}^{r} b_k \sum_{m=1}^{\infty}\frac{(k\Delta x)^m}{m!}\frac{\partial^m q}{\partial x^m}\right]. \tag{2.91}$$

Since by hypothesis the local truncation error must be of order $p$ in space and in time, i.e.

$$\tau = \mathcal{O}\left(\Delta x^p, \Delta t^p\right), \tag{2.92}$$

all terms of order $p$ or lower must vanish in (2.91). For the leading error term we obtain the condition

$$q_i^n\left(1 - \sum_{k=-l}^{r} b_k\right) = 0. \tag{2.93}$$

This is the case $m = 0$ in Eqn. (2.88), which thus has been proven. For the higher order terms we have the conditions

$$\sum_{m=1}^{p}\left(\frac{\Delta t^m}{m!}\frac{\partial^m q}{\partial t^m} - \sum_{k=-l}^{r} b_k\frac{(k\Delta x)^m}{m!}\frac{\partial^m q}{\partial x^m}\right) = 0, \quad \text{for } 0 < m \leq p. \tag{2.94}$$

To simplify this expression we apply the Cauchy-Kovalewski procedure to the linear advection equation obtaining the following general result (the proof is very simple using complete induction):

$$\frac{\partial^m q}{\partial t^m} = (-a)^m\frac{\partial^m q}{\partial x^m}. \tag{2.95}$$

Substituting the time derivatives (2.95) into (2.94) we obtain the conditions

$$\sum_{m=1}^{p}\left(\frac{\Delta t^m}{m!}(-a)^m\frac{\partial^m q}{\partial x^m} - \sum_{k=-l}^{r} b_k\frac{(k\Delta x)^m}{m!}\frac{\partial^m q}{\partial x^m}\right) = 0, \quad \text{for } 0 < m \leq p, \tag{2.96}$$

which after factorization yields

$$\sum_{m=1}^{p}\frac{\Delta x^m}{m!}\frac{\partial^m q}{\partial x^m}\left[\frac{\Delta t^m\,(-a)^m}{\Delta x^m} - \sum_{k=-l}^{r} b_k k^m\right] = 0 \tag{2.97}$$

The first term in the square brackets can be simplified using the definition of the Courant number $c$. Since this relation must hold for any value of the spatial derivatives, the term in the square brackets must vanish, i.e. we have the condition

$$(-c)^m - \sum_{k=-l}^{r} b_k k^m = 0, \quad \text{for } 0 < m \leq p, \tag{2.98}$$

which completes the proof of the lemma.

□

THEOREM 1. *(Godunov, 1959). For the linear scalar advection equation* $q_t + aq_x = 0$ *there exists* ***no*** *monotone linear method of the form* (2.6) *of order of accuracy greater or equal than two.*

PROOF. To prove the theorem, we restrict ourselves to the second order case. Let us define coefficients $s_m$ as follows:

$$s_m = \sum_{k=-l}^{r} k^m b_k. \tag{2.99}$$

Using a linear method of second order we need to determine the coefficients $s_0$, $s_1$, $s_2$. According to the Lemma of Roe (2.88), they result as $s_0 = 1$, $s_1 = -c$, $s_2 = c^2$. We therefore have

$$s_2 = \sum_{k=-l}^{r} k^2 b_k = \sum_{k=-l}^{r} (k+c)^2 b_k - 2c \overbrace{\sum_{k=-l}^{r} k b_k}^{s_1=-c} - c^2 \overbrace{\sum_{k=-l}^{r} b_k}^{s_0=1} = c^2. \tag{2.100}$$

Using the expressions for $s_0$ and $s_1$ according to the lemma of Roe, this is equivalent to

$$s_2 = \sum_{k=-l}^{r} (k+c)^2 b_k + 2c^2 - c^2 = c^2, \tag{2.101}$$

and hence

$$\sum_{k=-l}^{r} \overbrace{(k+c)^2}^{>0} b_k = 0. \tag{2.102}$$

Following the definition of a monotone linear method (2.84) the coefficients $b_k$ have to be non–decreasing, i.e. $b_k \geq 0$. Hence, the only possibility to satisfy the condition (2.102) preserving the monotonicity implies to impose

$$b_k = 0, \quad \forall k = -l, ..., r, \tag{2.103}$$

but this leads to a contradiction because according to the Roe lemma we must have for zeroth order consistency

$$s_0 = \sum_{k=-l}^{r} b_k = 1, \tag{2.104}$$

hence the $b_k$ can not be all zero! □

CHAPTER 3

# Nonlinear Hyperbolic Equations

## 1. Nonlinear scalar hyperbolic equations

We consider the following first order nonlinear scalar equation

$$\frac{\partial q}{\partial t}+\frac{\partial}{\partial x}f\left(q\right)=0, \tag{3.1}$$

where $q$ denotes the conservative variable and $f\left(q\right)$ is the nonlinear flux. The equation (3.1) is called *conservation law.* Classical examples from physics are the conservation of mass, momentum and energy. If the conservative variable $q$ and the flux $f$ are sufficiently regular, the derivatives in (3.1) are defined and the PDE can be written also in the so–called *quasi-linear* form using the chain rule:

$$\frac{\partial q}{\partial t}+a\left(q\right)\frac{\partial q}{\partial x}=0,\quad \text{with } a\left(q\right)=\frac{df}{dq}=f'\left(q\right). \tag{3.2}$$

In (3.2) the term $a\left(q\right)$ is the characteristic velocity, generally function of $q$. An important property of $a(q)$ is its monotonicity, for which there are three cases:

- Convex flux, when $a\left(q\right)$ is monotone and increasing

$$\frac{d}{dq}a\left(q\right)=a'\left(q\right)=f''\left(q\right)>0,\quad \forall q. \tag{3.3}$$

- Concave flux, when $a\left(q\right)$ is monotone and decreasing

$$\frac{d}{dq}a\left(q\right)=a'\left(q\right)=f''\left(q\right)<0,\quad \forall q. \tag{3.4}$$

- Non-convex and non-concave flux if

$$\exists q:\frac{d}{dq}a\left(q\right)=a'\left(q\right)=f''\left(q\right)=0. \tag{3.5}$$

Example 21. *We consider the nonlinear Burgers equation, defined by*

$$\frac{\partial q}{\partial t}+\frac{\partial}{\partial x}\left(\frac{1}{2}q^2\right)=0,\quad f(q)=\frac{1}{2}q^2. \tag{3.6}$$

*We compute the characteristic velocity $a(q)$ and its derivative*

$$a\left(q\right)=\frac{df}{dq}=q,\quad a'=f''=1>0,\forall q. \tag{3.7}$$

*We find that the flux of the Burgers equation is convex. This means that larger values of $q$ propagate with higher velocity than the small ones.*

Example 22. *The following PDE represents a simple model for traffic flow*

$$\frac{\partial q}{\partial t}+\frac{\partial}{\partial x}f\left(q\right)=0,\quad f(q)=u_{\max}\left(1-\frac{q}{q_{\max}}\right)q, \tag{3.8}$$

*where $q$ denotes the vehicle density per length unit, $u_{\max}$ is the maximum velocity (speed limit) and $q_{\max}$ is the maximum vehicle density. The term $a(q)$ and its* *derivative $a'$ are given by*

$$a\left(q\right)=f'\left(q\right)=u_{\max}\left(1-\frac{2q}{q_{\max}}\right),\quad a'=f''=-\frac{2u_{\max}}{q_{\max}}<0,\quad\forall q. \tag{3.9}$$

*The flux is concave. Here, the larger values of $q$ propagate at slower velocities than the smaller values of $q$.*

Example 23. *In this case we consider a PDE with cubic flux*

$$\frac{\partial q}{\partial t}+\frac{\partial}{\partial x}f\left(q\right)=0,\quad f(q)=q^3. \tag{3.10}$$

*We obtain*

$$a\left(q\right)=f'\left(q\right)=3q^2,\quad a'=f''=6q\begin{cases}<0, & \text{if } q<0,\\ =0, & \text{if } q=0,\\ >0, & \text{if } q>0.\end{cases} \tag{3.11}$$

*Hence, the cubic flux is neither convex nor concave.*

**1.1. Solution along the characteristics.** We consider the initial value problem

$$\frac{\partial q}{\partial t}+\frac{\partial f}{\partial x}=0,\qquad q\left(x,0\right)=h\left(x\right), \tag{3.12}$$

which is written in quasi-linear form as

$$\frac{\partial q}{\partial t}+a\left(q\right)\frac{\partial q}{\partial x}=0,\qquad q\left(x,0\right)=h\left(x\right). \tag{3.13}$$

In the nonlinear case, the characteristic curves are defined by the ODE

$$\frac{dx}{dt}=a\left(q\left(x(t),t\right)\right),\qquad x\left(0\right)=x_0. \tag{3.14}$$

Now we compute the material derivative of the variable $q$ along the characteristic

$$\frac{dq}{dt}=\frac{\partial q}{\partial t}\frac{dt}{dt}+\frac{\partial q}{\partial x}\frac{dx}{dt}=\frac{\partial q}{\partial t}+a\left(q\right)\frac{\partial q}{\partial x}=0. \tag{3.15}$$

We find that even in the nonlinear case, the solution remains *constant* along the characteristic curves, hence also the characteristic velocity $a\left(q\right)$ is constant along the characteristic. It follows that the characteristic curves are also *straight lines* even in the *nonlinear* case. Starting from the initial condition in (3.13) we obtain the equation of the characteristic curves as

$$x=x_0+a\left(h\left(x_0\right)\right)t, \tag{3.16}$$

which is a nonlinear scalar algebraic equation for the foot of the characteristic $x_0$ in terms of $x$ and $t$. The solution of the nonlinear PDE (3.13) is then given by the initial condition evaluated at the foot of the characteristic $x_0$ as

$$q\left(x,t\right)=h\left(x_0\right)=h\left(x-a\left(h\left(x_0\right)\right)t\right). \tag{3.17}$$

To verify this solution we compute the derivatives of $q\left(x,t\right)$ in time and in space

$$\frac{\partial q}{\partial t}=h'\left(x_0\right)\frac{\partial x_0}{\partial t},\qquad\frac{\partial q}{\partial x}=h'\left(x_0\right)\frac{\partial x_0}{\partial x}. \tag{3.18}$$

Similarly, we differentiate (3.16)

$$\begin{array}{l}\frac{\partial x}{\partial t}=\frac{\partial x_0}{\partial t}+a+a'h'\frac{\partial x_0}{\partial t}t=0\Rightarrow\frac{\partial x_0}{\partial t}=\frac{-a}{1+a'h't},\\ \frac{\partial x}{\partial x}=\frac{\partial x_0}{\partial x}+a+a'h'\frac{\partial x_0}{\partial x}t=1\Rightarrow\frac{\partial x_0}{\partial x}=\frac{1}{1+a'h't}.\end{array} \tag{3.19}$$

Substituting into (3.15) we verify that (3.17) is the solution of the PDE

$$\frac{\partial q}{\partial t}+a\left(q\right)\frac{\partial q}{\partial x}=-\frac{ah'}{1+a'h't}+\frac{ah'}{1+a'h't}=0. \tag{3.20}$$

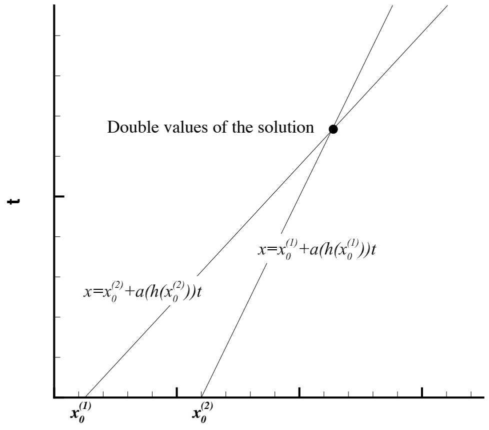
Double values of the solution
t
$x=x_0^{(1)}+a(h(x_0^{(1)}))t$
$x=x_0^{(2)}+a(h(x_0^{(2)}))t$
$x_0^{(1)}$
$x_0^{(2)}$

Figure 1. The intersection of two characteristic curves leads to a multi–valued solution in the intersection point.

Note that the solution of (3.12) is formally derived as in the linear case using the characteristic curves, but it is *implicit*. Then, it necessary to implement an iterative scheme, such the Newton or bisection method, to compute $x_0$ from $x$ and $t$ solving the following equation

$$g\left(x_0\right)=x_0-x+a\left(h\left(x_0\right)\right)t=0. \tag{3.21}$$

Moreover, the method of characteristics breaks down when two characteristic curves intersect, which we will see later in the section of shock waves. Since each characteristic curve transports a constant solution, at the intersection point there would be two different values of $q$ defined, according to the different values transported by each characteristic. In this case we would have a so–called multi–valued solution at the intersection point $(x,t)$, which is sketched in Fig. 1.

**1.2. Rarefaction waves.** As an example, we consider the following initial value problem

$$\begin{aligned} &\text{PDE:} && q_t+f_x=0, \qquad f''>0,\\ &\text{IC:} && q\left(x,0\right)=h\left(x\right)=\begin{cases} q_L, & \text{if } x<x_L,\\ q_L+\frac{q_R-q_L}{x_R-x_L}\left(x-x_L\right), & \text{if } x_L\le x\le x_R,\\ q_R & \text{if } x>x_R, \end{cases} \end{aligned} \tag{3.22}$$

with $a\left(q_L\right)\le a\left(q_R\right)$. The initial condition of the problem (3.22) consists of two piecewise constant states and of a piecewise linear distribution of the variable $q$ between the points $x_L$ and $x_R$. The solution of (3.22) is given by a rarefaction wave, plotted in Fig. 2. Its borders are defined by the two particular characteristic curves

$$\begin{aligned} x&=x_L+a\left(q_L\right)t,\\ x&=x_R+a\left(q_R\right)t, \end{aligned} \tag{3.23}$$

which are called the *head* and the *tail* of the rarefaction wave with the foot points located at $x_L$ and $x_R$, respectively. The characteristics associated with the tail and the head of the rarefaction divide the half-plane $x-t$ into three regions $R_0$, $R_1$ and $R_2$.

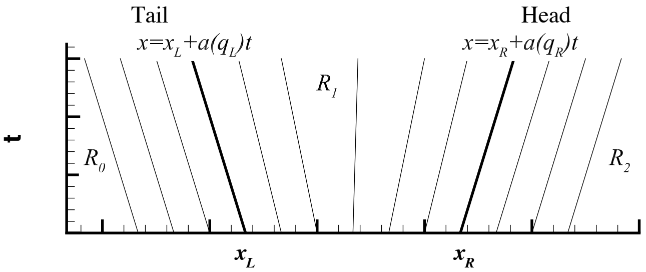

FIGURE 2. The solution of the initial value problem (3.22).

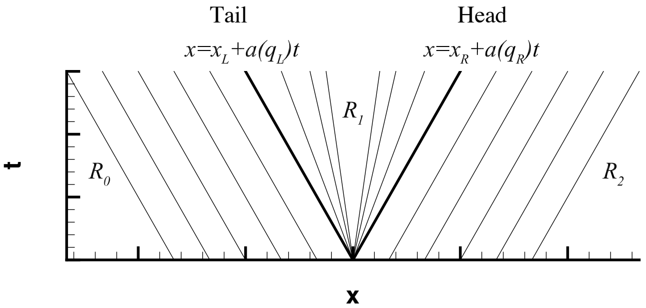

FIGURE 3. The solution of the Riemann problem (3.24).

Now, we consider the Riemann problem

$$q_t + f_x = 0, \quad f'' > 0, \qquad q(x,0) = h(x) = \begin{cases} q_L, & \text{if } x < x_0, \\ q_R, & \text{if } x \geq x_0, \end{cases} \tag{3.24}$$

with $a(q_L) \leq a(q_R)$. This problem can be considered as the limit case of the previous one (3.22) with $(x_L, x_R) \to x_0$. Thus, we find self-similar solutions $q(x,t)$ depending only on one variable $\xi = \frac{x-x_0}{t}$ so that

$$q(x,t) = q(\xi) = q\left(\frac{x-x_0}{t}\right). \tag{3.25}$$

We compute the derivatives in space and in time of the unknown $q(x,t)$

$$\frac{\partial q}{\partial t} = -\frac{x-x_0}{t^2}q', \quad \frac{\partial q}{\partial x} = \frac{1}{t}q', \tag{3.26}$$

where $q'$ means the derivative with respect to $\xi$. Substitution into (3.22) yields

$$\left(-\frac{x-x_0}{t^2} + a(q(\xi))\frac{1}{t}\right)Q' = 0, \quad \Rightarrow \quad a(q(\xi)) = \frac{x-x_0}{t} = \xi. \tag{3.27}$$

For the Riemann problem (3.24), the foot points of the head and the tail of the rarefaction wave coincide with the initial location of the discontinuity $x_0$, as depicted in Fig. 3. The complete solution of the Riemann problem is given by

$$q(x,t) = \begin{cases} q_L, & \text{if } \frac{x-x_0}{t} < a(q_L), \\ \text{root of } g(q) = a(q) - \frac{x-x_0}{t} = 0, & \text{if } a(q_L) \leq \frac{x-x_0}{t} \leq a(q_R), \\ q_R, & \text{if } \frac{x-x_0}{t} > a(q_R), \end{cases} \tag{3.28}$$

where the intermediate states between the head and the tail of the rarefaction wave are computed by solving the nonlinear scalar algebraic equation

$$a(q) = \xi, \tag{3.29}$$

which can be solved numerically by using for example a Newton–Raphson method or the bisection algorithm.

**1.3. Shock waves.** As a motivation for the introduction of shock waves, first introduced into the theory of nonlinear partial differential equations by Bernhard Riemann in [**67, 68**], we solve following initial value problem for the Burgers equation (3.6) written in quasi–linear form

$$\begin{aligned} &\text{PDE:} \quad \tfrac{\partial q}{\partial t} + q\tfrac{\partial q}{\partial x} = 0, \\ &\text{IC:} \qquad q\left(x,0\right) = h\left(x\right) = \begin{cases} 1, & \text{if } x < 0, \\ 1-x, & \text{if } 0 \le x \ge 1, \\ 0, & \text{if } x \ge 1. \end{cases} \end{aligned} \tag{3.30}$$

The characteristic curves divide the domain in three regions, i.e. $R_0$, $R_1$ and $R_2$, as shown in Fig. 4. The lines $c_0$ and $c_1$ are defined by

$$\begin{aligned} c_0 : \quad & x = 0 + 1 \cdot t = t, \\ c_1 \quad & x = 1 + 0 \cdot t = 1. \end{aligned} \tag{3.31}$$

In the regions $R_0$ and $R_2$ all the characteristic curves are parallel to the line $c_0$ and $c_1$, respectively. In the middle (region $R_1$) the characteristics converge on the intersection point $P$ and the foot of the characteristic is computed from (3.16) as

$$x = x_0 + a(h(x_0))t = x_0 + (1-x_0)t = x_0\left(1-t\right) + t. \tag{3.32}$$

Hence, the foot of the characteristic in terms of $x$ and $t$ is given by

$$x_0 = \frac{x-t}{1-t}, \tag{3.33}$$

and thus the solution along the characteristic curve in the region $R_1$ is

$$q\left(x,t\right) = h\left(x_0\right) = 1 - x_0 = 1 - \frac{x-t}{1-t}. \tag{3.34}$$

To conclude, we summarize the complete solution until the intersection time $t = 1$:

$$q\left(x,t\right) = \begin{cases} 1, & \text{if } x < t, \\ \frac{1-x}{1-t}, & \text{if } t \le x \le 1, \\ 0, & \text{if } x > 1. \end{cases} \tag{3.35}$$

Fig. 5 shows the evolution in time of the solution of (3.30). Note that at time $t \geq 1$ the solution $q$ has multiple values for $x \geq 1$. The problem of multi–valued solutions has been solved by Bernhard Riemann already in 1860 [**67, 68**] by the introduction of the concept of *shock waves*, i.e. a *discontinuity* in the solution that guarantees the conservation of the variable $q$ through the condition $A_1 = A_2$. The terms $A_1$ and $A_2$ are plotted in Fig. 6 and denote the area given by the intersection between the multi–valued solution (solid line) and the shock profile assumed by Riemann (dashed line). The discontinuity is localized so that the solution is *conservative.*

**1.4. Conservation laws in integral form.** We analyze a general conservative PDE (3.1). Let us define a control volume in space $I = [x_L, x_R]$ and integrate the PDE over this spatial control volume:

$$\int_{x_L}^{x_R} \frac{\partial q}{\partial t} dx + \int_{x_L}^{x_R} \frac{\partial}{\partial x} f\left(q\right) dx = 0. \tag{3.36}$$

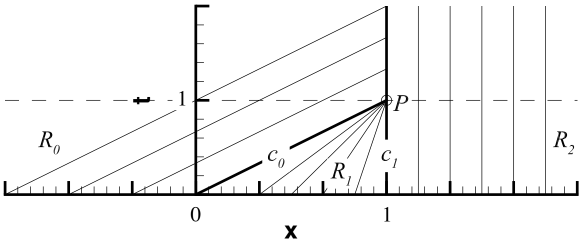

FIGURE 4. The characteristic curve corresponding to the problem (3.30) that define three regions $R_0$, $R_1$ and $R_2$.

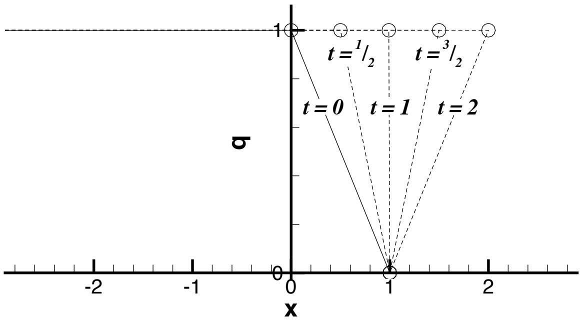

FIGURE 5. The solution of (3.30) at times $t = \frac{1}{2}, 1, \frac{3}{2}, 2$.

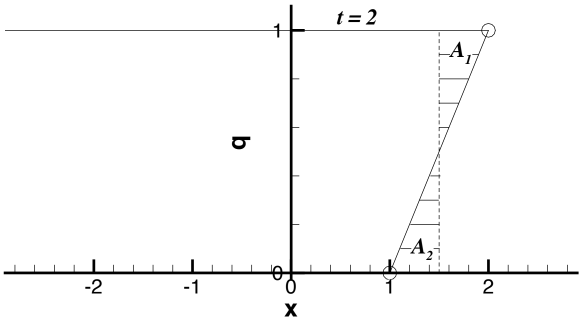

FIGURE 6. The solution at times $t = 2$ that presents a shock wave with multiple values of $q$. Assuming the equality of $A_1$ and $A_2$ the shock wave can be localized so that the solution is conservative.

Integrating the second term by parts we obtain

$$\frac{\partial}{\partial t}\int_{x_L}^{x_R} q\left(x,t\right)dx + f\left(q\left(x_R,t\right)\right) - f\left(q\left(x_L,t\right)\right) = 0. \tag{3.37}$$

This means that the temporal variation of the conservative variable $q\left(x,t\right)$ inside the spatial interval $I = [x_L; x_R]$ is only due to the difference of the fluxes on the boundaries of the control volume $I$.

When the control volume is defined in space *and* in time $V = [x_L; x_R] \times [t_1; t_2]$, we obtain

$$\int_{t_1}^{t_2} \int_{x_L}^{x_R} \frac{\partial q}{\partial t} dx dt + \int_{t_1}^{t_2} \int_{x_L}^{x_R} \frac{\partial}{\partial x} f\left(q\right) dx dt = 0, \tag{3.38}$$

that is equivalent to

$$\int_{x_L}^{x_R} q\left(x, t_2\right) dx = \int_{x_L}^{x_R} q\left(x, t_1\right) dx - \int_{t_1}^{t_2} \left(f\left(q\left(x_R, t\right)\right) - f\left(q\left(x_L, t\right)\right)\right) dt. \tag{3.39}$$

Thus, the conservative quantity at time $t_2$ is equal to the conservative quantity at time $t_1$ subtracted by the difference of the integrals of the fluxes in time computed on the spatial boundaries of the control volume.

**1.5. Rankine-Hugoniot conditions.** The Rankine–Hugoniot conditions relate the jump of the conserved quantities across a shock wave with the propagation velocity $s$ of the discontinuity. They are derived from integral conservation over a control volume that includes the shock.

Let us assume that the discontinuity travels with a velocity $s$ and that it is an isolated wave that divides two piecewise constant states on the left $q_L$ and on the right $q_R$ of the shock. Then, the PDE in integral form over a space–time control volume $V = [x_L; x_R] \times [t_1; t_2]$ with $x_L \leq x_d \leq x_R$ and $x_L \leq x_d + s\left(t_2 - t_1\right)$ is given by

$$\int_{x_L}^{x_R} \left(q\left(x, t_2\right) - q\left(x, t_1\right)\right) dx + \left(\int_{t_1}^{t_2} \left(f\left(q\left(x_R, t\right)\right) - f\left(q\left(x_L, t\right)\right)\right) dt\right) = 0, \tag{3.40}$$

where $x_d$ is the location of the shock wave at time $t_1$. Evaluation of the integrals yields

$$\begin{aligned} q_L\left(x_d + s\left(t_2 - t_1\right)\right) + q_R\left(x_R - x_d - s\left(t_2 - t_1\right)\right) - \\ -q_L(x_d - x_L) - q_R(x_R - x_d) + \left(f\left(q_R\right) - f\left(q_L\right)\right)\left(t_2 - t_1\right) = 0. \end{aligned} \tag{3.41}$$

This leads to the Rankine-Hugoniot relations

$$s\left(q_R - q_L\right) = f\left(q_R\right) - f\left(q_L\right), \tag{3.42}$$

which relate the jump of the conservative quantity $q$ with the speed of the shock wave $s$ and the jump of the flux over the discontinuity.

Example 24. *We consider the Burgers equation* (3.6) *in conservative form given by the flux* $f\left(q\right) = \frac{1}{2}q^2$. *The relations of Rankine-Hugoniot provide the shock speed* $s$, *as follows*

$$\left(\frac{1}{2}q_R^2 - \frac{1}{2}q_L^2\right) = s\left(q_R - q_L\right). \tag{3.43}$$

*Simplifying we obtain*

$$s = \frac{1}{2}\left(q_R + q_L\right). \tag{3.44}$$

**1.6. Riemann problem with shock wave.** We consider the initial value problem

$$\begin{aligned} &\text{PDE:} \quad \tfrac{\partial q}{\partial t} + \tfrac{\partial}{\partial x} f(q) = 0, \\ &\text{IC:} \quad\; q\left(x, 0\right) = h\left(x\right) = \begin{cases} q_L, & \text{if } x < x_0, \\ q_R, & \text{if } x \geq x_0, \end{cases} \end{aligned} \tag{3.45}$$

where we suppose $a\left(q_L\right) > a\left(q_R\right)$. The solution of this initial value problem is given by

$$q\left(x, t\right) = \begin{cases} q_L, & \text{if } \frac{x - x_0}{t} < s, \\ q_R, & \text{if } \frac{x - x_0}{t} \geq s, \end{cases} \tag{3.46}$$

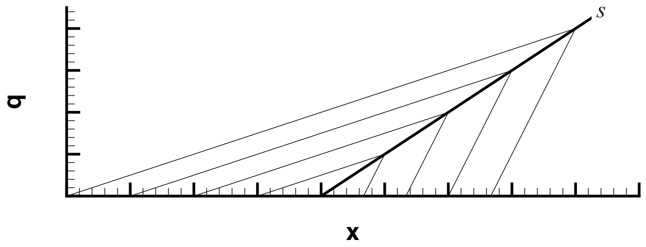

FIGURE 7. The solution of the initial value problem that presents a shock that propagates in space with a celerity $s$.

where the shock speed $s$ is determined by the use of the Rankine–Hugoniot conditions (3.42). The characteristic curves intersect and lead to a discontinuity (shock wave), traveling with speed $s$, as shown in Fig. 7.

**1.7. Non-uniqueness of solutions and the Lax entropy condition.** Solving nonlinear hyperbolic conservation laws using the weak (integral) formulation allows us to admit discontinuities in the solution, thus extending the set of possible solutions even if they are not sufficiently smooth. However, this does not guarantee uniqueness. It is necessary to introduce further conditions that lead to a unique solution starting from an initial value problem.

EXAMPLE 25. *We solve the following initial value problem for the Burgers equation* (3.6)*:*

$$\begin{aligned} &\text{PDE:} \quad \frac{\partial q}{\partial t} + q\frac{\partial q}{\partial x} = 0, \\ &\text{IC:} \qquad q\left(x,0\right) = h\left(x\right) = \left\{ \begin{array}{ll} 0, & \text{if } x < 0, \\ 1, & \text{if } x \geq 0, \end{array} \right. \end{aligned} \tag{3.47}$$

*We have two possible solution alternatives, namely a rarefaction wave or a shock wave.*

*In the first case, we assume a rarefaction wave solution, plotted in Fig. 8 and given by*

$$q\left(x,t\right) = \left\{ \begin{array}{ll} 0, & \text{if } \frac{x}{t} < 0, \\ \frac{x}{t}, & \text{if } 0 \leq \frac{x}{t} \leq 1, \\ 1, & \text{if } \frac{x}{t} > 1. \end{array} \right. \tag{3.48}$$

*This first solution is a classical solution, based on the theory of characteristics and therefore it is also a weak solution since all classical solutions that satisfy the PDE in its strong differential form also satisfy the PDE in its weak integral form.*

*In the second case, the solution contains a discontinuity (rarefaction shock), as plotted in Fig. 9. It is given by*

$$q\left(x,t\right) = \left\{ \begin{array}{ll} 0, & \text{if } \frac{x}{t} < s, \\ 1, & \text{if } \frac{x}{t} \geq s, \end{array} \right. \qquad \text{with } s = \frac{1}{2}\left(q_L + q_R\right) = \frac{1}{2}. \tag{3.49}$$

*The second solution, obtained by using the Rankine–Hugoniot conditions, satisfies the integral form of the conservation law and therefore is also a weak solution of our conservation law. This can be proven by using the integral form of the conservation law* (3.39) *over the space–time control volume* $[x_L; x_R] \times [t_1; t_2]$*, see Fig. 10. With* $s = 1/2$ *we obtain*

$$x_R - st_2 - \left(x_R - st_1\right) + \frac{1}{2}\left(t_2 - t_1\right) = 0. \tag{3.50}$$

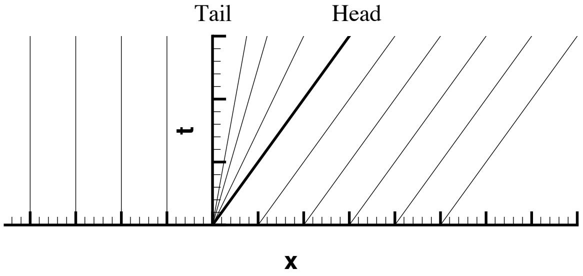

FIGURE 8. The solution of the Burgers equation including the rarefaction fan.

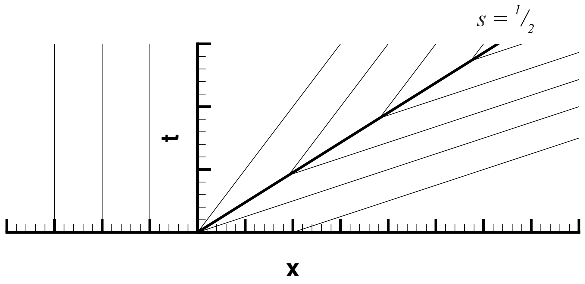

FIGURE 9. The solution of the Burgers equation including the rarefaction shock.

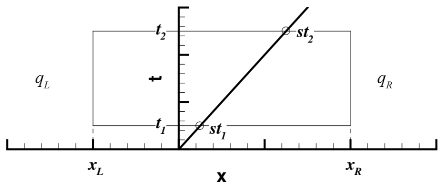

FIGURE 10. The space-time control volume used for the integral form (3.39) of the conservation law.

*This proves indeed that also the rarefaction shock solution satisfies the integral form of the conservation law and hence is a possible weak solution of the problem. Nevertheless, according to the entropy criterion of Lax it is not an admissible solution from the physical point of view, since shock waves are only produced if the characteristic curves enter the shock wave. In contrast to this, as plotted in Fig. 9, for the rarefaction shock solution, the characteristic curves leave the shock wave.*

Since hyperbolic conservation laws allow more than one solution if we also admit weak solutions, we need an additional criterion to select the physically correct one. A criterion introduced by Lax is the so–called Lax entropy condition

$$a\left(q_L\right) > s > a\left(q_R\right), \tag{3.51}$$

where $s$ is the velocity of the shock obtained from the Rankine–Hugoniot condition. The Lax entropy condition requires that the characteristic curves enter the shock wave from both sides, or in other words the shock is *compressed* from both sides (left and right). When the criterion (3.51) is not satisfied, then the unique solution that is physically admissible is the rarefaction wave.

## 2. Nonlinear hyperbolic systems

We consider the following system of hyperbolic conservation laws

$$\frac{\partial \vec{Q}}{\partial t} + \frac{\partial}{\partial x}\vec{f}\left(\vec{Q}\right) = 0, \quad \text{with } \vec{Q} \in \Omega_Q \subset \mathbb{R}^m, \vec{f} \in \mathbb{R}^m, \tag{3.52}$$

in which the PDE are nonlinear and the terms $\vec{Q}$ and $\vec{f}(\vec{Q})$ are the vector of the conserved quantities and the flux vector, respectively.

$$\vec{Q} = (q_1, q_2, ..., q_m)^T, \quad \vec{f}\left(\vec{Q}\right) = (f_1, f_2, ..., f_m)^T. \tag{3.53}$$

In quasi-linear form the expression above results in

$$\frac{\partial \vec{Q}}{\partial t} + \underline{\underline{A}}\left(\vec{Q}\right)\frac{\partial \vec{Q}}{\partial x} = 0, \quad \text{with} \quad \underline{\underline{A}}\left(\vec{Q}\right) = \frac{\partial \vec{f}}{\partial \vec{Q}}, \tag{3.54}$$

where the matrix $\underline{\underline{A}}(\vec{Q})$ is the Jacobian matrix of the flux vector. The system (3.52) is called *hyperbolic* if the Jacobian matrix of the flux has $m$ real eigenvalues and a complete set of linearly independent eigenvectors for all states $\vec{Q} \in \Omega_Q$.

Similarly to (3.39) we define a space-time control volume $V = [x_L; x_R] \times [t_1, t_2]$ and write the nonlinear system (3.52) in integral form

$$\int_{x_L}^{x_R} \vec{Q}\left(x, t_2\right) dx = \int_{x_L}^{x_R} \vec{Q}\left(x, t_1\right) dx + \int_{t_1}^{t_2} \left(\vec{f}\left(\vec{Q}\left(x_R, t\right) - \vec{f}\left(\vec{Q}\left(x_L, t\right)\right)\right)\right) dt. \tag{3.55}$$

The corresponding Rankine-Hugoniot conditions at a shock wave propagating with speed $s$ are given by

$$\vec{f}\left(\vec{Q}\left(x_R, t\right)\right) - \vec{f}\left(\vec{Q}\left(x_L, t\right)\right) = s\left(\vec{Q}\left(x_R, t\right) - \vec{Q}\left(x_L, t\right)\right). \tag{3.56}$$

To understand the nature of the characteristic curves we analyze the $i^{th}$ eigenvalue $\lambda_i$ of the matrix $\underline{\underline{A}}$ and the associated eigenvector $\vec{r}_i$. The characteristic field is defined to be *linearly degenerate* if

$$\nabla\lambda_i\left(\vec{Q}\right) \cdot \vec{r}_i\left(\vec{Q}\right) = 0, \quad \forall \vec{Q} \in \Omega_Q. \tag{3.57}$$

Otherwise, it is *genuinely nonlinear* when

$$\nabla\lambda_i\left(\vec{Q}\right) \cdot \vec{r}_i\left(\vec{Q}\right) \neq 0, \quad \forall \vec{Q} \in \Omega_Q. \tag{3.58}$$

Here, the gradient $\nabla\lambda$ means the gradient with respect to the vector of conserved quantities, i.e. we use the notation

$$\nabla\lambda = \frac{\partial \lambda}{\partial \vec{Q}}. \tag{3.59}$$

The phase-space or state-space $\Omega_Q$ is the space of the vectors

$$\vec{Q} = (q_1, q_2, ..., q_m)^T \in \Omega_{\vec{Q}} \subseteq \mathbb{R}^m. \tag{3.60}$$

Example 26. *We consider the equations of isentropic gasdynamics given by*

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho u \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} \rho u \\ \rho u^2 + p \end{pmatrix} = 0, \tag{3.61}$$

*where the variables $\rho$, $u$ and $p$ denote the density, the velocity in $x$–direction and the pressure, respectively. The system* (3.61) *is not closed because the number of* *equations is not sufficient to determine the flow field. Hence, a closure relation is needed, called the equation of state (EOS). A very simple choice consists of the isentropic EOS, given by*

$$p = k\rho^{\gamma}, \quad \text{with } \gamma = \frac{c_P}{c_V}, \quad \gamma \in \mathbb{R}, k \in \mathbb{R}, \tag{3.62}$$

*where $c_V$ and $c_P$ are the heat capacity at constant volume and the heat capacity at constant pressure, respectively. Replacing* (3.62) *into* (3.61) *and writing the system in terms of the conservative variables $\vec{Q} = (q_1, q_2)^T = (\rho, \rho u)^T$ we obtain*

$$\frac{\partial}{\partial t}\begin{pmatrix} q_1 \\ q_2 \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} q_2 \\ \frac{q_2^2}{q_1} + kq_1^{\gamma} \end{pmatrix} = 0. \tag{3.63}$$

*It follows that the Jacobian matrix of the flux vector of isentropic gasdynamics results as*

$$\underline{\underline{A}}\left(\vec{Q}\right) = \frac{\partial \vec{f}}{\partial \vec{Q}} = \begin{pmatrix} 0 & 1 \\ C^2 - \frac{q_2^2}{q_1^2} & 2\frac{q_2}{q_1} \end{pmatrix}, \tag{3.64}$$

*where c is the sound speed, given by*

$$C^2 = \frac{dp}{d\rho} = \gamma k q_1^{\gamma-1} = \frac{\gamma p}{\rho}. \tag{3.65}$$

*The matrix $\underline{\underline{A}}$ can be written in physical variables, i.e. primitive variables, as follows*

$$\underline{\underline{A}} = \begin{pmatrix} 0 & 1 \\ C^2 - u^2 & 2u \end{pmatrix}. \tag{3.66}$$

*Now, we compute the eigenvalues from the characteristic polynomial*

$$\left\|\underline{\underline{A}} - \lambda\underline{\underline{I}}\right\| = \lambda^2 - 2u\lambda - C^2 + u^2 = 0, \tag{3.67}$$

*the roots of which are $\lambda_{1,2} = u \pm C$. Using the first eigenvalue we obtain the following right eigenvector*

$$\begin{pmatrix} C-u & 1 \\ C^2 - u^2 & 2u + C - u \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Rightarrow \vec{r}_1 = \begin{pmatrix} 1 \\ u - C \end{pmatrix}. \tag{3.68}$$

*Similarly, the right eigenvector associated with the second eigenvalue is*

$$\begin{pmatrix} -C-u & 1 \\ C^2 - u^2 & u - C \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Rightarrow \vec{r}_1 = \begin{pmatrix} 1 \\ u + C \end{pmatrix}. \tag{3.69}$$

*In matrix form we have*

$$\underline{\underline{R}} = \begin{pmatrix} 1 & 1 \\ u - C & u + C \end{pmatrix}, \quad \underline{\underline{\Lambda}} = \begin{pmatrix} u - C & 0 \\ 0 & u + C \end{pmatrix}, \quad \underline{\underline{R}}^{-1} = \frac{1}{2C}\begin{pmatrix} C + u & -1 \\ C - u & 1 \end{pmatrix}. \tag{3.70}$$

Example 27. *We consider the shallow water model, that consists of a set of hyperbolic PDE*

$$\frac{\partial}{\partial t}\begin{pmatrix} h \\ hu \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} hu \\ hu^2 + \bar{p} \end{pmatrix} = 0, \tag{3.71}$$

*where the variables h and u are the water depth and the velocity in x–direction, respectively. The term $\bar{p}$ denotes the averaged pressure force along the water depth. Since the shallow water equations are derived from depth–integrating the incompressible Navier-Stokes equations assuming that the horizontal length scale is much larger than the vertical one, the vertical pressure is assumed to be hydrostatic. Then, the average pressure force (normalized with the fluid density ρ) is given by*

$$\bar{p} = \frac{1}{2}gh^2, \tag{3.72}$$

*where $g$ is the gravity acceleration along the vertical direction. We write the system* (3.71) *in terms of the conservative variables $\vec{Q} = (q_1, q_2)^T = (h, hu)^T$, as follows*

$$\frac{\partial}{\partial t}\begin{pmatrix} q_1 \\ q_2 \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} q_2 \\ \frac{q_2^2}{q_1} + \frac{1}{2} g q_1^2 \end{pmatrix} = 0. \tag{3.73}$$

*The Jacobian matrix $\underline{\underline{A}}$ is given by*

$$\underline{\underline{A}}\left(\vec{Q}\right) = \frac{\partial \vec{f}}{\partial \vec{Q}} = \begin{pmatrix} 0 & 1 \\ -\frac{q_2^2}{q_1^2} + C^2 & 2\frac{q_2}{q_1} \end{pmatrix}, \tag{3.74}$$

*where $C$ is the celerity of the gravity waves that propagate on the water surface and is equal to $C = \sqrt{gh}$. Similarly, in primitive variables the Jacobian matrix is written as*

$$\underline{\underline{A}} = \begin{pmatrix} 0 & 1 \\ C^2 - u^2 & 2u \end{pmatrix}. \tag{3.75}$$

*Note that there is a very strong similarity between the system of isentropic compressible gasdynamics* (3.61) *and the shallow water model. Hence, we can write by analogy as follows*

$$\underline{\underline{R}} = \begin{pmatrix} 1 & 1 \\ u - C & u + C \end{pmatrix}, \quad \underline{\underline{\Lambda}} = \begin{pmatrix} u - C & 0 \\ 0 & u + C \end{pmatrix}, \quad \underline{\underline{R}}^{-1} = \frac{1}{2C}\begin{pmatrix} C + u & -1 \\ C - u & 1 \end{pmatrix}. \tag{3.76}$$

## 3. The exact solution of the Riemann problem for isothermal gasdynamics

For the Godunov flux [**37**] described later in this manuscript we need the *exact solution* of the local Riemann problem at element interfaces. We already know how to solve the Riemann problem in the case of general linear hyperbolic systems and for general scalar nonlinear hyperbolic equations. In this section and the following one, we will give a strategy for the solution of the Riemann problem for two particular nonlinear hyperbolic systems, namely the governing equations of isothermal compressible gasdynamics augmented by the transport of a passive scalar and the shallow water equations. We restrict ourselves to these two particular systems since the general nonlinear case is very difficult to handle. For many important but rather complicated hyperbolic systems the exact solution of the Riemann problem is still the topic of current research.

In this section, we consider the isothermal compressible gasdynamics, augmented by the transport equation of a passive scalar, such as small dust particles. The governing PDE system consists of the conservation equations of mass and momentum for the fluid as well as the mass conservation equation for the passive scalar given by

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) &= 0, \\ \frac{\partial \rho u}{\partial t} + \frac{\partial}{\partial x}(\rho u^2 + p) &= 0, \\ \frac{\partial \rho \psi}{\partial t} + \frac{\partial}{\partial x}(\rho u \psi) &= 0. \end{aligned} \tag{3.77}$$

where $\rho$ is the mass density of the fluid, $u$ is the velocity, $\psi$ is the concentration of the passive scalar and $p = a^2 \rho$ is the pressure according to the isothermal hypothesis. Recall that the ideal gas law is $p/\rho = RT$ and for the isothermal case the square of the sound speed is hence $a^2 = RT$, where $R$ is the specific gas constant and $T$ is the temperature. The nonlinear PDE system can be written as

$$\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x} f(Q) = 0, \qquad Q, f \in \mathbb{R}^3, \quad t \in \mathbb{R}_0^+, \quad x \in \mathbb{R},$$

with the vector of conserved variables $Q$ and the flux vector $f$ being defined as

$$Q = \begin{pmatrix} \rho \\ \rho u \\ \rho\psi \end{pmatrix} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}, f = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u\psi \end{pmatrix} = \begin{pmatrix} q_2 \\ q_2^2/q_1 + a^2 q_1 \\ q_2 q_3/q_1 \end{pmatrix}. \tag{3.78}$$

The Jacobian matrix $A$ of the flux $f$ with respect to the vector of conserved variables $Q$ is

$$A = \begin{pmatrix} 0 & 1 & 0 \\ -q_2^2/q_1^2 + a^2 & 2q_2/q_1 & 0 \\ -q_2 q_3/q_1^2 & q_3/q_1 & q_2/q_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ a^2 - u^2 & 2u & 0 \\ -u\psi & \psi & u \end{pmatrix}. \tag{3.79}$$

For (3.78) the diagonal matrix of eigenvalues is $\Lambda = \text{diag}(u - a, u, u + a)$ and the corresponding matrix of right eigenvectors $R$ and its inverse $R^{-1}$ are given by

$$R = \begin{pmatrix} 1 & 0 & 1 \\ u - a & 0 & u + a \\ \psi & 1 & \psi \end{pmatrix}, \quad R^{-1} = \frac{1}{2a} \begin{pmatrix} a + u & -1 & 0 \\ -2a\psi & 0 & 2a \\ a - u & 1 & 0 \end{pmatrix}. \tag{3.80}$$

For the first and the third eigenvalue we find that the associated characteristic fields are *genuinely nonlinear* since

$$\nabla\lambda_{1,3} = \nabla(u \mp a) = \frac{\partial}{\partial Q}(q_2/q_1 \mp a) = (-q_2/q_1^2, 1/q_1, 0)^T = (-u/\rho, 1/\rho, 0)^T. \tag{3.81}$$

and hence

$$\nabla\lambda_{1,3} \cdot \vec{r}_{1,3} = (-u/\rho, 1/\rho, 0)^T \cdot (1, u \mp a, 0)^T = \mp a/\rho \neq 0, \quad \forall Q \in \Omega_Q. \tag{3.82}$$

The second characteristic field, associated to $\lambda_2 = u$ is linearly degenerate, which can be seen as follows:

$$\nabla\lambda_2 = \nabla u = \frac{\partial}{\partial Q}(q_2/q_1) = (-q_2/q_1^2, 1/q_1, 0)^T = (-u/\rho, 1/\rho, 0)^T. \tag{3.83}$$

$$\nabla\lambda_2 \cdot \vec{r}_2 = (-u/\rho, 1/\rho, 0)^T \cdot (0, 0, 1)^T = 0, \quad \forall Q \in \Omega_Q. \tag{3.84}$$

The general solution of the Riemann problem for the nonlinear PDE system (3.78) consists of three waves associated with the three eigenvalues of the PDE. The intermediate wave, associated to $\lambda_2$ is always a so–called *contact discontinuity*, due to the linear degeneracy of the second characteristic field. The nonlinear waves associated with the first and the third eigenvalue can be either a shock wave, in the case of compression of the characteristics, or a centered rarefaction wave, according to the Lax entropy condition. A sketch for a typical configuration can be found in Fig. 11.

**3.1. Shock waves and Rankine–Hugoniot relations.** The general Rankine–Hugoniot conditions for a system of conservation laws read

$$s\,(Q_R - Q_L) = (f(Q_R) - f(Q_L)). \tag{3.85}$$

Since the algebraic manipulations can be performed much easier for a steady shock wave, we use the principle of Galilean invariance of Newtonian mechanics which states that the governing PDE remain unchanged under a linear transformation of the velocity of the form

$$\hat{u} = u - s. \tag{3.86}$$

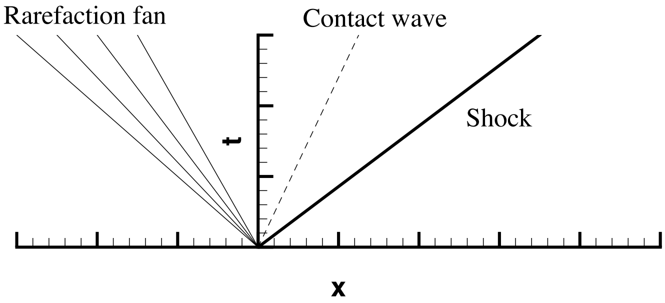

FIGURE 11. The plot for a typical solution of the Riemann problem for the nonlinear PDE system (3.78).

Here, $u$ is the fluid velocity in the fixed laboratory frame, $s$ is the (constant) velocity of an observer that moves with the speed of the shock wave $s$ and $\hat{u}$ is the transformed velocity in the co–moving frame. Note that in particular we have $\hat{s} = 0$, since the shock wave is steady in the co–moving frame. With the use of the principle of Galilean invariance, the Rankine–Hugoniot relations written in the co–moving frame reduce to

$$0 \cdot (\hat{Q}_R - \hat{Q}_L) = f(\hat{Q}_R) - f(\hat{Q}_L), \qquad \Rightarrow \qquad f(\hat{Q}_R) = f(\hat{Q}_L). \tag{3.87}$$

In isothermal gasdynamics this leads to

$$\begin{aligned} &\rho_R \hat{u}_R = \rho_L \hat{u}_L := M, \\ &\rho_R \hat{u}_R^2 + a^2 \rho_R = \rho_L \hat{u}_L^2 + a^2 \rho_L, \\ &\rho_R \hat{u}_R \psi_R = \rho_L \hat{u}_L \psi_L. \end{aligned} \tag{3.88}$$

Performing algebraic manipulations on the equations above we obtain

$$\begin{aligned} &\psi_R = \psi_L, \\ &M = -a^2 \tfrac{\rho_R - \rho_L}{\hat{u}_R - \hat{u}_L}, \\ &\hat{u}_R = \tfrac{M}{\rho_R}, \quad \hat{u}_L = \tfrac{M}{\rho_L}. \end{aligned} \tag{3.89}$$

Finally, the transformed velocity referred to the states on the left and on the right of the shock discontinuity is given by

$$\hat{u}_R = -a\sqrt{\frac{\rho_L}{\rho_R}}, \quad \hat{u}_L = -a\sqrt{\frac{\rho_R}{\rho_L}}. \tag{3.90}$$

Then, the velocity jump across the shock in the fixed laboratory frame results

$$\hat{u}_R - \hat{u}_L = u_R - u_L = a\frac{\rho_R - \rho_L}{\sqrt{\rho_R \rho_L}}, \tag{3.91}$$

and the speed of the shock wave is equal to

$$s = u_R - \hat{u}_R = u_R + a\sqrt{\frac{\rho_L}{\rho_R}}. \tag{3.92}$$

The Lax entropy condition (3.51) applied to isothermal gasdynamics requires that the eigenvalue $\lambda_3 \left(Q_{L,R}\right) = u_{L,R} + a$ satisfies the following relation

$$\lambda_3 \left(Q_L\right) > s > \lambda_3 \left(Q_R\right), \tag{3.93}$$

that is equivalent to

$$\overbrace{u_R - a\frac{\rho_R - \rho_L}{\sqrt{\rho_R \rho_L}}}^{=u_L} + a > u_R + a\sqrt{\frac{\rho_L}{\rho_R}} > u_R + a, \quad \Rightarrow \quad \rho_L > \rho_R. \tag{3.94}$$

The Lax entropy condition therefore requires that the shock wave produces a compression in the fluid, i.e. an *increase* of density across the shock wave.

**3.2. Rarefaction waves and Riemann invariants.** For a general quasi-linear hyperbolic system

$$\frac{\partial Q}{\partial t} + A\left(Q\right)\frac{\partial Q}{\partial x} = 0, \tag{3.95}$$

we seek self-similar solutions so that

$$Q\left(x,t\right) = v\left(\xi\right), \quad \text{with } \xi = \frac{x}{t}. \tag{3.96}$$

In terms of the self-similar variable $v$ the PDE results

$$\frac{\partial v}{\partial t} + A\left(v\right)\frac{\partial v}{\partial x} = 0. \tag{3.97}$$

Using the chain–rule, we obtain

$$\frac{dv}{d\xi}\frac{\partial \xi}{\partial t} + A\left(v\right)\frac{dv}{d\xi}\frac{\partial \xi}{\partial x} = -v'\frac{x}{t^2} + A\left(v\right)\frac{v'}{t} = \left(A\left(v\right) - \xi I\right)v' = 0. \tag{3.98}$$

The trivial solution of (3.98) is $v' = 0$. Non–trivial solutions are obtained for the case where $\xi$ is an eigenvalue of $A$ and $v'$ is the associated eigenvector of $A$, i.e. when

$$\xi = \lambda_k\left(v\left(\xi\right)\right), \qquad \text{and} \qquad v' = r_k\left(v\left(\xi\right)\right)\cdot\alpha\left(\xi\right). \tag{3.99}$$

Here, $\alpha(\xi)$ is an appropriate scaling factor, since the eigenvectors of $A$ are determined only up to a multiplicative constant.

Following the method of characteristics, a characteristic curve $C_k$ is defined by the ODE

$$\frac{dx}{dt} = \lambda_k\left(Q\left(x,t\right)\right) = \lambda_k\left(v\left(\xi\right)\right), \tag{3.100}$$

where the index $k$ denotes the number of the characteristic field under consideration. If we fix the value of $\xi$, the variable $v\left(\xi\right)$ and the eigenvalue $\lambda_k$ are constants. In the case of a simple centered rarefaction wave, the solution is thus constant along the characteristic $C_k$, which therefore is a straight line, as in the linear and the nonlinear scalar case and as plotted in Fig. 12.

We consider the first characteristic field ($\lambda_1 = u - a$) for isothermal gasdynamics. Let us introduce a change of coordinates $\xi \to \zeta$ so that $\alpha := 1$. Then, the system is given by

$$\frac{dv\left(\zeta\right)}{d\zeta} = r_1\left(v\left(\zeta\right)\right), \quad \Rightarrow \quad \frac{d}{d\zeta}\left(\begin{array}{c}\rho\\ \rho u\\ \rho\psi\end{array}\right) = \left(\begin{array}{c}1\\ u-a\\ \psi\end{array}\right). \tag{3.101}$$

This leads to the following relations

$$\frac{d\rho}{d\zeta} = 1, \quad \frac{d}{d\zeta}\left(\rho u\right) = u\frac{d\rho}{d\zeta} + \rho\frac{du}{d\zeta} = u - a, \quad \frac{d}{d\zeta}\left(\rho\psi\right) = \psi\frac{d\rho}{d\zeta} + \rho\frac{d\psi}{d\zeta} = \psi. \tag{3.102}$$

Hence, we integrate across the first wave and obtain

$$\begin{array}{l}\int_{\rho_L}^{\rho} d\rho = \int_0^1 d\zeta, \quad \Rightarrow \quad \rho\left(\zeta\right) = \rho_L + \zeta,\\ \int_{u_L}^{u} du = \int_0^1 -\frac{a}{\rho(\zeta)}d\zeta = \int_0^1 -\frac{a}{\rho_L+\zeta}d\zeta, \quad \Rightarrow \quad u\left(\zeta\right) = u_L + a\ln\frac{\rho_L}{\rho(\zeta)},\\ \int_{\psi_L}^{\psi} d\psi = 0, \quad \Rightarrow \quad \psi = \psi_L = const,\end{array} \tag{3.103}$$

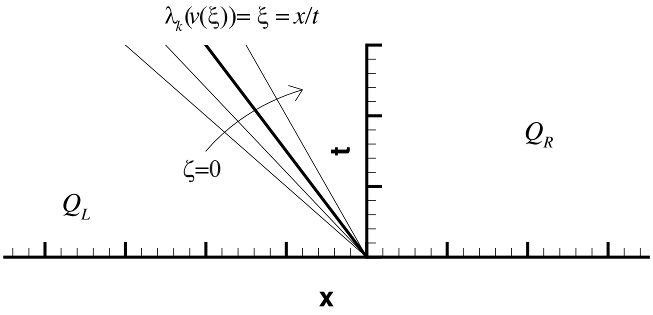

FIGURE 12. Sketch of an isolated centered rarefaction fan, for which the characteristic curves consist of straight lines.

where $L$ indicates the left state. From (3.103) we can derive the so–called *Riemann invariants* for the system of isothermal gasdynamics. They consist of the following relations that hold true across the wave structure

$$u + a\ln\rho = u_L + a\ln\rho_L := w_1^1 = const., \quad \psi = \psi_L := w_1^2 = const. \tag{3.104}$$

Here, the notation $w_k^j$ denotes the $j$–th Riemann invariant associated with the $k$–th characteristic field. To compute the solution $Q = (\rho, \rho u, \rho\psi)$ in a particular point $\xi$ along a characteristic curve we use the equation $\xi = \lambda_1 = u - a$, see Fig. 12. Using (3.104) one derives the primitive variables

$$u = a + \xi, \quad \rho = \rho_L e^{\frac{u_L - a - \xi}{a}}, \quad \psi = \psi_L. \tag{3.105}$$

We now consider the contact discontinuity with $\lambda_2 = u$, obtaining the following system of ODE across the wave structure:

$$\frac{d}{d\zeta}\begin{pmatrix} \rho \\ \rho u \\ \rho\psi \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \tag{3.106}$$

which correspond to

$$\frac{d\rho}{d\zeta} = 0, \quad \frac{d}{d\zeta}(\rho u) = u\frac{d\rho}{d\zeta} + \rho\frac{du}{d\zeta} = 0, \quad \frac{d}{d\zeta}(\rho\psi) = \psi\frac{d\rho}{d\zeta} + \rho\frac{d\psi}{d\zeta} = 1. \tag{3.107}$$

With easy algebraic manipulations the Riemann invariants are defined by

$$\rho = \rho_L := w_2^1 = const., \quad u = u_L := w_2^2 = const. \tag{3.108}$$

This result is also consistent with the Rankine–Hugoniot relations. Let us consider a Riemann problem with the initial states, such that $u_L = u_R$ and $\rho_L = \rho_R$. We verify that the Rankine–Hugoniot conditions (3.56) are satisfied:

$$u\begin{pmatrix} \rho_R - \rho_L \\ \rho_R u_R - \rho_L u_L \\ \rho_R\psi_R - \rho_L\psi_L \end{pmatrix} = \begin{pmatrix} \rho_R u_R \\ \rho_R u_R^2 + a\rho_R^2 \\ \rho_R u_R \psi_R \end{pmatrix} - \begin{pmatrix} \rho_L u_L \\ \rho_L u_L + a\rho_L^2 \\ \rho_L u_L \psi_L \end{pmatrix}. \tag{3.109}$$

We found that the velocity $u$ is a Riemann invariant for the second characteristic field, which is identical with the eigenvalue $\lambda_2 = u$. We will analyze this feature in

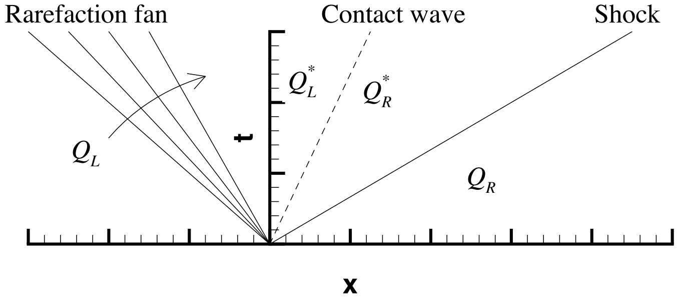

FIGURE 13. The plot for the Riemann solution.

a more general setting in the following. In general, a Riemann invariant $w_k(Q) = w_k(v(\xi))$ associated with the characteristic field $k$ satisfies

$$\frac{d}{d\xi}w_k = 0, \tag{3.110}$$

hence, with the chain rule we obtain

$$\frac{\partial w_k}{\partial Q} \cdot \frac{dQ}{d\xi} = 0, \qquad \Rightarrow \qquad \nabla w_k \cdot v' = 0, \tag{3.111}$$

since $dQ/d\xi = v'$. Since $v' = r_k(v(\xi)) \cdot \alpha(\xi)$, it follows that

$$\nabla w_k \cdot r_k = 0. \tag{3.112}$$

Comparing with the definition of linear degenerate fields (3.57), it results that the eigenvalue $\lambda_k$ is a Riemann invariant, i.e. $w_k = \lambda_k$, if the field $k$ is linearly degenerate.

**3.3. Solution of the Riemann problem.** We seek the exact solution of the Riemann problem for isothermal gasdynamics. A typical wave structure is sketched in Fig. 13.

Across the rarefaction fan the Riemann invariants are constant. Therefore, we use the relations (3.104) to compute the state $Q_L^*$ in the star region on the left of the contact discontinuity as

$$u_L^* + a\ln\rho_L^* = u_L + a\ln\rho_L, \quad \psi_L^* = \psi_L, \tag{3.113}$$

that leads to

$$\rho_L^* = \rho_L e^{\frac{u_L - u_L^*}{a}}, \quad \psi_L^* = \psi_L. \tag{3.114}$$

Similarly, across the contact wave the density and the velocity are defined by

$$\rho_R^* = \rho_L = \rho^*, \quad u_R^* = u_L = u^*. \tag{3.115}$$

The Rankine–Hugoniot conditions (3.56) across the shock give

$$u_R^* = u_R + a\frac{\rho_R^* - \rho_R}{\sqrt{\rho_R^*\rho_R}}, \quad s = u_R + a\sqrt{\frac{\rho_R^*}{\rho_R}}, \quad \psi_R^* = \psi_R. \tag{3.116}$$

Through algebraic manipulations we obtain the following relations

$$u_L + a\ln\frac{\rho_L}{\rho^*} = u_R + a\ln\frac{\rho^* - \rho_R}{\sqrt{\rho^*\rho_R}}, \tag{3.117}$$

$$g\left(\rho^*\right) = a \ln \frac{\rho^*}{\rho_L} + a \ln \frac{\rho^* - \rho_R}{\sqrt{\rho^* \rho_R}} + u_R - u_L = 0. \tag{3.118}$$

Note that (3.117) and (3.118) hold true only when the exact solution follows the configuration plotted in Fig. 13. For a general wave structure we have to find the zero of the following nonlinear function in terms of the density variable $\rho^*$

$$g\left(\rho^*\right) = \phi_L + \phi_R + u_R - u_L = 0, \tag{3.119}$$

where the value of $\phi_{L,R}$ depends on the nature of the nonlinear wave (shock or rarefaction):

$$\phi_L = \begin{cases} a \ln \frac{\rho^*}{\rho_L}, & \text{if } \rho^* \leq \rho_L, \quad \text{(rarefaction)} \\ a \frac{\rho^* - \rho_L}{\sqrt{\rho^* \rho_L}}, & \text{if } \rho^* > \rho_L, \quad \text{(shock)} \end{cases} \tag{3.120}$$

$$\phi_R = \begin{cases} a \ln \frac{\rho^*}{\rho_R}, & \text{if } \rho^* \leq \rho_R, \quad \text{(rarefaction)} \\ a \frac{\rho^* - \rho_R}{\sqrt{\rho^* \rho_R}}, & \text{if } \rho^* > \rho_R. \quad \text{(shock)} \end{cases} \tag{3.121}$$

Solving the equation (3.119) requires an iterative procedure. A wide used and proper choice consists of the Newton method, that assures quick convergence to the sought solution $\rho^*$.

## 4. The exact solution of the Riemann problem for the shallow water equations

We consider the shallow water equations that describe the evolution of a homogeneous, incompressible free surface flow under the condition that the horizontal length scale is much greater in comparison with the vertical one. This leads to neglecting the vertical accelerations of the fluid and to a hydrostatic pressure distribution along the water depth. Then, the mathematical model is strongly simplified and results as

$$\begin{aligned} \frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) &= 0, \\ \frac{\partial hu}{\partial t} + \frac{\partial}{\partial x}(hu^2 + \bar{p}) &= 0, \end{aligned} \tag{3.122}$$

where $h$ is the depth of the fluid, $u$ is the velocity and $\bar{p}$ is the pressure according to the assumption of hydrostatic distribution, $\bar{p} = \frac{1}{2}gh^2$. The system of equations can be written in matrix form

$$\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x} f(Q) = 0, \qquad Q \in \Omega_Q \subset \mathbb{R}^2, \quad t \in \mathbb{R}_0^+, \quad x \in \mathbb{R}.$$

where the vector of conserved variables $Q$ and the flux vector $f$ are defined as

$$Q = \begin{pmatrix} h \\ hu \end{pmatrix} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}, f = \begin{pmatrix} hu \\ hu^2 + \bar{p} \end{pmatrix} = \begin{pmatrix} q_2 \\ q_2^2/q_1 + \frac{1}{2} g q_1^2 \end{pmatrix}. \tag{3.123}$$

The Jacobian matrix $A$ of the flux $f$ with respect to the vector of conserved variables $Q$ is given by

$$A = \begin{pmatrix} 0 & 1 \\ -q_2^2/q_1^2 + a^2 & 2q_2/q_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ a^2 - u^2 & 2u \end{pmatrix}, \tag{3.124}$$

where $a = \sqrt{gh}$ denotes celerity of the gravity waves that propagate on the water surface. As computed in example 27, the eigenvalues are given by $\lambda_{1,2} = u \mp a$ and the corresponding matrix of right eigenvectors $R$ and its inverse $R^{-1}$ are the following

$$R = \begin{pmatrix} 1 & 1 \\ u - a & u + a \end{pmatrix}, \quad R^{-1} = \frac{1}{2a} \begin{pmatrix} a + u & -1 \\ a - u & 1 \end{pmatrix}. \tag{3.125}$$

We first compute the nature of the characteristic fields and find for both eigenvalues

$$\begin{aligned} \nabla\lambda_1 &= \tfrac{\partial}{\partial Q}\left(\tfrac{q_2}{q_1} - \sqrt{gq_1}\right) = \left(-\tfrac{q_2}{q_1^2} - \tfrac{1}{2}\sqrt{\tfrac{g}{q_1}}, \tfrac{1}{q_1}\right)^T = \left(-\tfrac{u}{h} - \tfrac{1}{2}\sqrt{\tfrac{g}{h}}, \tfrac{1}{h}\right)^T, \\ \nabla\lambda_2 &= \tfrac{\partial}{\partial Q}\left(\tfrac{q_2}{q_1} + \sqrt{gq_1}\right) = \left(-\tfrac{q_2}{q_1^2} + \tfrac{1}{2}\sqrt{\tfrac{g}{q_1}}, \tfrac{1}{q_1}\right)^T = \left(-\tfrac{u}{h} + \tfrac{1}{2}\sqrt{\tfrac{g}{h}}, \tfrac{1}{h}\right)^T, \end{aligned} \tag{3.126}$$

that the fields are *genuinely nonlinear*

$$\begin{aligned} \nabla\lambda_1 \cdot \vec{r}_1 &= \left(-\tfrac{u}{h} - \tfrac{1}{2}\sqrt{\tfrac{g}{h}}, \tfrac{1}{h}\right)^T \cdot \left(1, u - \sqrt{gh}\right)^T = -\tfrac{3}{2}\sqrt{\tfrac{g}{h}} \neq 0, \quad \forall Q, \\ \nabla\lambda_2 \cdot \vec{r}_2 &= \left(-\tfrac{u}{h} + \tfrac{1}{2}\sqrt{\tfrac{g}{h}}, \tfrac{1}{h}\right)^T \cdot \left(1, u + \sqrt{gh}\right)^T = +\tfrac{3}{2}\sqrt{\tfrac{g}{h}} \neq 0, \quad \forall Q. \end{aligned} \tag{3.127}$$

The solution of the Riemann problem using the shallow water model produces two waves, namely the rarefaction fan and the shock. Since only one vertical level $h$ is computed, the shallow water equations can not directly encompass any physical quantity that varies with height. Hence, they are especially suitable to model phenomena which have very large length scales compared to the depth $h$, such as rivers and tides.

**4.1. Shock waves and Rankine–Hugoniot conditions.** To achieve the relation associated with the shock we use the Rankine–Hugoniot conditions. We apply the principle of Galilean invariance of Newtonian mechanics and observe the shock wave in a moving frame, such that the linear transformation of the velocity is given by the relation (3.86). Then, the Rankine–Hugoniot conditions (3.56) for the shallow water equations become

$$\begin{aligned} h_R\hat{u}_R &= h_L\hat{u}_L := \mu, \\ h_R\hat{u}_R^2 + \tfrac{1}{2}gh_R^2 &= h_L\hat{u}_L^2 + \tfrac{1}{2}gh_L^2. \end{aligned} \tag{3.128}$$

With algebraic manipulations we obtain

$$\begin{aligned} \hat{u}_R = \tfrac{\mu}{h_R}, \quad \hat{u}_L = \tfrac{\mu}{h_L}, \\ \mu\left(\hat{u}_R - \hat{u}_L\right) = \tfrac{1}{2}g\left(h_L^2 - h_R^2\right), \quad \Rightarrow \quad \mu = -\tfrac{1}{2}g\tfrac{h_R^2 - h_L^2}{\hat{u}_R - \hat{u}_L}, \end{aligned} \tag{3.129}$$

that provide the velocity jump across the shock wave expressed in terms of the initial water depths $h_L$, $h_R$

$$\hat{u}_R - \hat{u}_L = u_R - u_L = \sqrt{\frac{1}{2}g\frac{h_L + h_R}{h_R h_L}}\left(h_R - h_L\right), \tag{3.130}$$

and the celerity

$$s = u_R - \hat{u}_R = u_R + \sqrt{\frac{1}{2}g\frac{h_L}{h_R}\left(h_L + h_R\right)}. \tag{3.131}$$

To satisfy the Lax entropy condition (3.51) for the shallow water equations we need that

$$u_L + \sqrt{gh_L} > s > u_R + \sqrt{gh_R}, \quad \Rightarrow \quad h_L > h_R. \tag{3.132}$$

Then, the Lax entropy condition requires that the water depth increases across the shock wave.

**4.2. Rarefaction waves and Riemann invariants.** We consider the self–similar solutions associated with (3.122). In the case of the shallow water model, the relations (3.98) lead to the following system

$$\frac{d}{d\zeta}\left(\begin{array}{c} h \\ hu \end{array}\right) = \left(\begin{array}{c} 1 \\ \lambda_k \end{array}\right), \tag{3.133}$$

with $\alpha := 1$ and $\xi \to \zeta$. Let us assume the solution structure plotted in Fig. 14 with a rarefaction wave on the left and a shock on the right. For the first eigenvalue $\lambda_1 = \zeta = u - a$, with $a = \sqrt{gh}$ we have

$$\frac{dh}{d\zeta} = 1, \quad \frac{d}{d\zeta}\left(hu\right) = u\frac{dh}{d\zeta} + h\frac{du}{d\zeta} = u + h\frac{du}{d\zeta} = u - \sqrt{gh}. \tag{3.134}$$

Hence, we integrate across the first wave structure

(3.135)

$$\begin{aligned} &\textstyle\int_{h_L}^{h} dh = \int_0^{\zeta} d\zeta \quad \Rightarrow \quad h\left(\zeta\right) = h_L + \zeta, \\ &\textstyle\int_{u_L}^{u} du = -\int_0^{\zeta} d\zeta \sqrt{\frac{g}{h}} = -\int_0^{\zeta} d\zeta \sqrt{\frac{g}{h_L+\zeta}} \quad \Rightarrow \quad u = u_L - \left(2\sqrt{gh\left(\zeta\right)} - 2\sqrt{gh_L}\right), \end{aligned}$$

where the index $L$ denotes the left state $Q_L = (h_L, h_L u_L)$ of the Riemann problem. The relations above lead to the Riemann invariant that holds true across the wave structure given by a rarefaction fan

$$u^* + 2\sqrt{gh^*} = u_L + 2\sqrt{gh_L}. \tag{3.136}$$

**4.3. Solution of the Riemann problem.** The exact Riemann solution for the shallow water equations is computed using the Riemann invariants and the Rankine–Hugoniot conditions. We assume a typical solution for the PDE as plotted in 14. It consists of a left rarefaction wave and a right shock. The region inside the wave structures, the so-called star region, is defined by a state $Q^* = (h^*, u^*)$ that is unknown.

The Riemann invariants for the shallow water equations across the left rarefaction fan give

$$u^* + 2\sqrt{gh^*} = u_L + 2\sqrt{gh_L}. \tag{3.137}$$

The Rankine-Hugoniot conditions provide the jump in the velocity field across the shock wave, as follows

$$u_R - u^* = -\sqrt{\frac{1}{2}g\frac{h^* + h_R}{h^* h_R}}\left(h^* - h_R\right). \tag{3.138}$$

With algebraic manipulations the relations (3.137) and (3.138) reduce to the following function $g$ expressed in terms of $h^*$

$$g\left(h^*\right) = 2\left(\sqrt{gh^*} - \sqrt{gh_L}\right) + \sqrt{\frac{1}{2}g\frac{h^* + h_R}{h^* h_R}}\left(h^* - h_R\right) + u_R - u = 0, \tag{3.139}$$

the root of which provides the state in the star region. Nevertheless, this solution is true only for the wave structure initially assumed. In the general case the state $Q^*$ is the root of

$$g\left(h^*\right) = \Phi_L + \Phi_R + u_R - u_L = 0, \tag{3.140}$$

where the function $\Phi_{L,R}$ is defined by

$$\Phi_{L,R} = \begin{cases} 2\left(\sqrt{gh^*} - \sqrt{gh_{L,R}}\right), & \text{if } h^* \le h_{L,R}, \\ \sqrt{\frac{1}{2}g\frac{h^*+h_{L,R}}{h^* h_{L,R}}}\left(h^* - h_{L,R}\right), & \text{if } h^* > h_{L,R}. \end{cases} \tag{3.141}$$

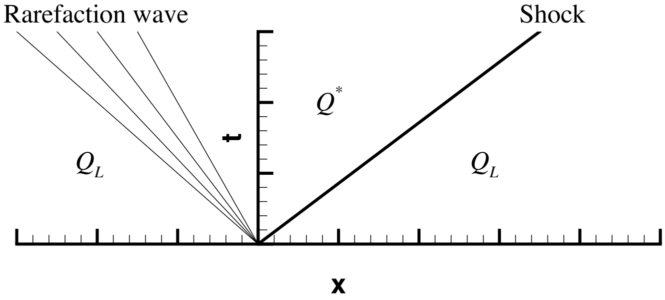

FIGURE 14. The plot for a typical solution of the Riemann problem for the swallow water equations.

CHAPTER 4

# The finite volume method

We consider the following initial–boundary value problem given by

$$
\begin{array}{lll}
 & \text{PDE:} & \frac{\partial}{\partial t}Q + \frac{\partial}{\partial x}f\left(Q\right) = 0, \quad x \in \left[x_L, x_R\right], Q \in \Omega_Q \subset \mathbb{R}^m, \\
(4.1) & \text{IC:} & Q\left(x, 0\right) = h\left(x\right), \\
 & \text{BC:} & Q\left(x_L, t\right) = B_L\left(t\right), \quad Q\left(x_R, t\right) = B_R\left(t\right),
\end{array}
$$

where $Q$ is the vector of the conservative variables, $f(Q)$ is the flux vector, $h$ is the initial condition and $B_{L,R}$ are the boundary conditions on the left and the right. To define the finite volume method we use the integral formulation of the conservation law. In Fig. 1 we show the discretization of the computational domain in space and time. The spatial control volumes are centered in $x_i$ and are defined by $I_i = [x_{i-\frac{1}{2}}; x_{i+\frac{1}{2}}]$, where the boundaries of the interval are equal to $x_{i\mp\frac{1}{2}} = x_i \mp \frac{\Delta x}{2}$, with $\Delta x$ the mesh size. For an equidistant grid, $\Delta x$ is given by $\Delta x = \frac{x_R - x_L}{i_{\max}}$. The temporal control volume is $T_n = \left[t^n; t^{n+1}\right]$, with the time step $\Delta t = t^{n+1} - t^n$. Integrating the PDE (4.1) over the space-time control volume $I_i \times T_n$, using the integral form 3.39, we obtain:

(4.2)

$$
\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} Q\left(x, t^{n+1}\right) dx = \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} Q\left(x, t^n\right) dx - \int\limits_{t^n}^{t^{n+1}} \left(f\left(Q(x_{i+\frac{1}{2}}, t)\right) - f\left(Q(x_{i-\frac{1}{2}}, t)\right)\right) dt.
$$

Since the spatial domain is divided into finite control volumes or cells, we can define the cell-average value computed at time $t^n$ as

$$
Q_i^n = \frac{1}{\Delta x} \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} Q\left(x, t^n\right) dx. \tag{4.3}
$$

Similarly, the time-averaged flux calculated at the cell interface $x_{i+\frac{1}{2}}$ is given by

$$
f_{i+\frac{1}{2}} = \frac{1}{\Delta t} \int\limits_{t^n}^{t^{n+1}} f\left(Q(x_{i+\frac{1}{2}}, t^n)\right) dt. \tag{4.4}
$$

Substituting the definitions (4.3) and (4.4) into (4.2) we can therefore derive the following *exact* relation, which is nothing else than a reformulation of the principle of integral conservation:

$$
Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}\right). \tag{4.5}
$$

So far, no numerical scheme has been introduced yet and the exact solution of (4.1) satisfies also (4.5) exactly. However, we will now use the formulation (4.5) in order to construct a discrete method for the solution of (4.1). Since the scheme is based on the integral conservation over finite control volumes, it is called a *finite volume* scheme. Remember that the finite volume (FV) approach is philosophically different from a finite difference (FD) scheme. In the FD framework we use the *strong differential* form of the PDE and approximate derivatives at discrete points $x_i$, $t^n$. In the FV framework, we discretize the *weak integral* form of the PDE. The main ingredient here is to define the numerical flux at the cell interfaces $f_{i\pm\frac{1}{2}}$ as function of the cell averages $Q_i^n$ and $Q_{i\pm1}^n$, since in the finite volume framework only the cell–averaged solution is known. This means that the finite volume scheme does not provide the discrete value of $Q$ at any spatial point $x_i$, but its integral average over the spatial control volume. However, in order to compute the flux at the element interfaces, we need to define the values of $Q$ at the interfaces $x_{i\pm\frac{1}{2}}$. This procedure is called *reconstruction step.* In the simplest of all cases, the solution $Q_i^n$ is supposed to be *piecewise constant* with value $Q_i^n$ within each cell. This piece–wise constant solution inside each cell leads to discontinuities at the cell interfaces $x_{i\pm1}$, see Fig. 2, since the numerical solution has two values at the interface:

$$Q^-_{i+\frac{1}{2}} = Q_i^n, \qquad \text{and} \qquad Q^+_{i+\frac{1}{2}} = Q_{i+1}^n, \tag{4.6}$$

where $Q^-_{i+\frac{1}{2}}$ denotes the solution at the left of the interface and $Q^+_{i+\frac{1}{2}}$ denotes the solution at the right of the interface $x_{i+\frac{1}{2}}$. To resolve these discontinuities, we need a so–called *numerical flux* that has to be a function of both states $Q^-_{i+\frac{1}{2}}$ and $Q^+_{i+\frac{1}{2}}$ on the left and on the right of the interface:

$$f_{i+\frac{1}{2}} = f_{i+\frac{1}{2}}\left(Q^-_{i+\frac{1}{2}}, Q^+_{i+\frac{1}{2}}\right) = f_{i+\frac{1}{2}}\left(Q_i^n, Q_{i+1}^n\right). \tag{4.7}$$

## 1. Basic numerical fluxes

The choice of the numerical flux can be for example:

- The central flux that is given by

$$f^c_{i+\frac{1}{2}} = \frac{1}{2}\left(f\left(Q_{i+1}^n\right) + f\left(Q_i^n\right)\right). \tag{4.8}$$

- The Lax-Friedrichs flux, defined as follows

$$f^{LF}_{i+\frac{1}{2}} = \frac{1}{2}\left(f\left(Q_{i+1}^n\right) + f\left(Q_i^n\right)\right) - \frac{1}{2}\frac{\Delta x}{\Delta t}\left(Q_{i+1}^n - Q_i^n\right). \tag{4.9}$$

- The flux of Lax–Wendroff in two possible versions. The first one consists of

$$f^{LW1}_{i+\frac{1}{2}} = \frac{1}{2}\left(f\left(Q_{i+1}^n\right) + f\left(Q_i^n\right)\right) - \frac{1}{2}\frac{\Delta t}{\Delta x}A^2_{i+\frac{1}{2}}\left(Q_{i+1}^n - Q_i^n\right), \tag{4.10}$$

where the term $A_{i+\frac{1}{2}}$ is an averaged Jacobian matrix defined at $x_{i+\frac{1}{2}}$, computed for example by the arithmetic average of the two Jacobians computed with the left and the right state, respectively. The second Lax–Wendroff–type flux is given by the following two–stage procedure

$$f^{LW2}_{i+\frac{1}{2}} = f\left(Q^{LW}_{i+\frac{1}{2}}\right), \tag{4.11}$$

where the state $Q^{LW}_{i+\frac{1}{2}}$ is computed from

$$Q^{LW}_{i+\frac{1}{2}} = \frac{1}{2}\left(Q_{i+1}^n + Q_i^n\right) - \frac{1}{2}\frac{\Delta t}{\Delta x}\left(f\left(Q_{i+1}^n\right) - f\left(Q_i^n\right)\right). \tag{4.12}$$

- Another possible choice consists of the FORCE flux, introduced by Toro and Billet [**80**]

$$f^{FO}_{i+\frac{1}{2}} = \frac{1}{2}\left(f^{LF}_{i+\frac{1}{2}} + f^{LW}_{i+\frac{1}{2}}\right). \tag{4.13}$$

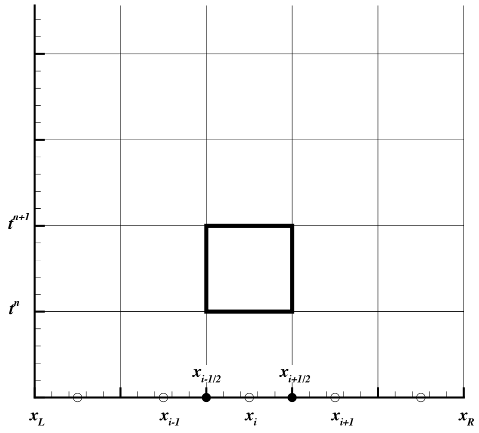

FIGURE 1. The space-time discretization and the space-time control volume $I_i \times T_n$ (bold line) referred to the finite volume method.

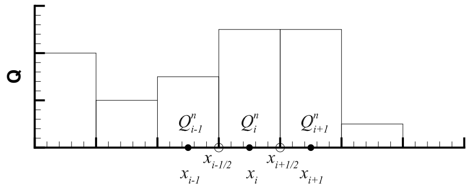

FIGURE 2. The assumption of a piece–wise constant solution using first order accurate finite volume schemes. Note the discontinuities at the element interfaces.

- The most accurate flux, but also the most complex one is the Godunov flux (1959), [**37**].

$$f^{GO}_{i+\frac{1}{2}} = f\left(Q^{GO}_{i+\frac{1}{2}}\right), \tag{4.14}$$

where the so–called Godunov state $Q^{GO}_{i+\frac{1}{2}} = Q^{RP}(0,t)$ is the *exact* solution of a local Riemann problem defined at the cell interface. The local Riemann problem is given by the initial value problem

$$\begin{array}{ll} \text{PDE:} & Q_t + f_x = 0, \\ \text{IC:} & Q\left(x,0\right) = \left\{ \begin{array}{ll} Q^n_i, & \text{if } x < 0, \\ Q^n_{i+1}, & \text{if } x \geq 0, \end{array} \right. \end{array} \tag{4.15}$$

and the exact solution of (4.15) is denoted by $Q^{RP}(x,t)$.

## 2. Properties of finite volume schemes

**2.1. Exact conservation.** A very important feature of finite volume methods is that they are *exactly* conservative, since the variation in time of a quantity, whose evolution obeys a conservation law (4.1), is only given by the exchange through the element interfaces. Hence, a method is conservative when it can be written as follows

$$Q^{n+1}_i = Q^n_i - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}\right), \tag{4.16}$$

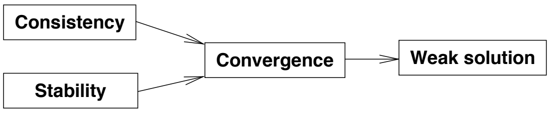

FIGURE 3. Convergence to the weak solution according to the Lax-Wendroff theorem (1960), which requires consistency and stability of the conservative numerical scheme.

with $f_{i+\frac{1}{2}} = f_{i+\frac{1}{2}}\left(Q_{i-l}^n, Q_{i-l+1}^n, ..., Q_i^n, Q_{i+1}^n, ..., Q_{i+r}^n\right)$ and $r, l \geq 0$. The finite volume scheme is exactly conservative by construction, which is easy to prove at the aid of the telescopic property of the resulting sum:

$$\begin{aligned}\sum_{i=1}^{i_{\max}} \Delta x Q_i^{n+1} &= \left(\sum_{i=1}^{i_{\max}} \Delta x Q_i^n\right) - \Delta t \left(\sum_{i=1}^{i_{\max}} f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}\right) = \\ &\left(\sum_{i=1}^{i_{\max}} \Delta x Q_i^n\right) - \Delta t \left(\overbrace{f_{i_{\max}+\frac{1}{2}}}^{\text{right BC}} - \overbrace{f_{1-\frac{1}{2}}}^{\text{left BC}}\right).\end{aligned} \tag{4.17}$$

This exact conservation property is crucial to solve nonlinear hyperbolic conservation laws with discontinuous solutions. According to the theorem of Hou and Le Floch (1994) [**45**] only a conservative method (4.16) is able to capture shock waves properly, with the right velocity $s$ and the right states $Q_L$, $Q_R$. As proven by Hou and Le Floch, a non-conservative scheme converges to the *wrong solution* with wrong shock speed and the wrong states at the shock wave.

**2.2. Consistency with the PDE.** A conservative method (4.16) is consistent if

$$f_{i+\frac{1}{2}}\left(Q, Q, ..., Q\right) = f\left(Q\right). \tag{4.18}$$

**2.3. Convergence to weak solutions.** The Lax–Wendroff theorem (1960) [**53**] states that if the numerical solution of a conservative method (4.16) converges, then it converges to a weak solution of the conservation law. For convergence, the method needs to be stable and consistent, see the sketch of Fig. (3).

**2.4. Monotonicity.** Regarding the property of monotonicity, a necessary condition so that the finite volume scheme is monotone for a scalar PDE $q_t + f_x = 0$ is given by

$$\frac{\partial}{\partial q_i^n} f_{i+\frac{1}{2}}\left(q_i^n, q_{i+1}^n\right) \geq 0, \quad \text{and} \quad \frac{\partial}{\partial q_{i+1}^n} f_{i+\frac{1}{2}}\left(q_i^n, q_{i+1}^n\right) \leq 0. \tag{4.19}$$

Hence, the numerical flux $f_{i+\frac{1}{2}}$ has to be a non-decreasing function of the first argument $q_i^n$ and a non-increasing function of the second one $q_{i+1}^n$. This is proved using the operator $H$ given by

$$q_i^{n+1} = H\left(q_{i-1}^n, q_i^n, q_{i+1}^n\right) = q_i^n - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}\right). \tag{4.20}$$

The general monotonicity condition (2.83) requires that the numerical flux satisfies the following relations:

$$\frac{\partial H}{\partial q_{i-1}^n} = \frac{\Delta t}{\Delta x}\frac{\partial f_{i-\frac{1}{2}}}{\partial q_{i-1}^n} \geq 0, \qquad \Rightarrow \qquad \frac{\partial f_{i-\frac{1}{2}}}{\partial q_{i-1}^n} \geq 0, \tag{4.21}$$

and

$$\frac{\partial H}{\partial q_{i+1}^n} = -\frac{\Delta t}{\Delta x}\frac{\partial f_{i+\frac{1}{2}}}{\partial q_{i+1}^n} \geq 0, \qquad \Rightarrow \qquad \frac{\partial f_{i+\frac{1}{2}}}{\partial q_{i+1}^n} \leq 0. \tag{4.22}$$

Example 28. *The aim of this example is to show that the finite volume scheme with the central flux and the central finite difference method coincide in the one-dimensional case. We consider the linear scalar advection equation* (3.1)*, with a constant positive velocity* $a \in \mathbb{R}^+$*. Using the central flux* (4.8) *in the finite volume scheme we obtain*

$$q_i^{n+1} = q_i^n - \frac{\Delta t}{2\Delta x}\left(aq_{i+1}^n + aq_i^n - aq_{i-1}^n - aq_i^n\right) = q_i^n - \frac{a\Delta t}{2\Delta x}\left(q_{i+1}^n - q_{i-1}^n\right). \tag{4.23}$$

*The numerical scheme derived above is the same obtained from the finite difference approach with the central flux* (1.97)*.*

*Moreover, according to the definition* (4.20) *we have*

$$q_i^{n+1} := H\left(q_{i-1}^n, q_i^n, q_{i+1}^n\right), \tag{4.24}$$

*and the necessary condition for the monotonicity is* ***not*** *satisfied, as can be easily seen below:*

$$\frac{\partial H}{\partial q_i^n} = 1 > 0, \quad \frac{\partial H}{\partial q_{i-1}^n} = \frac{a\Delta t}{2\Delta x} > 0, \quad \frac{\partial H}{\partial q_{i+1}^n} = -\frac{a\Delta t}{2\Delta x} < 0. \tag{4.25}$$

*Since the partial derivative of $H$ with respect to $q_{i+1}^n$ is always negative, the central scheme is* ***not*** *monotone.*

Example 29. *We consider the Lax-Friedrichs flux* (4.9) *in the finite volume case and prove the monotonicity of the resulting scheme, applied to the linear scalar advection equation. The scheme is*

$$\begin{aligned} q_i^{n+1} = q_i^n - \frac{\Delta t}{\Delta x}\Bigg[&\left(\frac{1}{2}\left(aq_i^n + aq_{i+1}^n\right) - \frac{1}{2}\frac{\Delta x}{\Delta t}\left(q_{i+1}^n - q_i^n\right)\right) - \\ &\left(\frac{1}{2}\left(aq_i^n + aq_{i-1}^n\right) - \frac{1}{2}\frac{\Delta x}{\Delta t}\left(q_i^n - q_{i-1}^n\right)\right)\Bigg]. \end{aligned} \tag{4.26}$$

*Simplifying we obtain*

$$q_i^{n+1} = \frac{1}{2}\left(q_{i+1}^n + q_{i-1}^n\right) - \frac{a\Delta t}{2\Delta x}\left(q_{i+1}^n - q_{i-1}^n\right) := H\left(q_{i-1}^n, q_{i+1}^n\right), \tag{4.27}$$

*which coincides with the Lax–Friedrichs finite difference scheme. For monotonicity we must have*

$$\frac{\partial H}{\partial q_{i-1}^n} = \frac{1}{2}(1+c) > 0, \quad \frac{\partial H}{\partial q_{i+1}^n} = \frac{1}{2}(1-c) > 0, \quad \forall 0 < c < 1, \tag{4.28}$$

*with the Courant number $c > 0$. We can conclude that the necessary condition for the monotonicity is satisfied under CFL condition $0 < c \leq 1$.*

Example 30. *For the Godunov flux* [**37**] *we first have to determine the exact solution of the local Riemann problem*

$$\left.\begin{aligned} &q_t + aq_x = 0, & (PDE) \\ &q(x,0) = \begin{cases} q_i^n, & x \leq 0, \\ q_{i+1}^n, & x > 0. \end{cases} & (IC) \end{aligned}\right\} (RP) \tag{4.29}$$

*at the element interface $x_{i+\frac{1}{2}}$, which means that we must find the self–similar solution of* (4.29) *at position $\xi = x/t = 0$. The exact solution of* (4.29) *in the entire $x - t$ half–plane is given by*

$$q(x,t) = \begin{cases} q_i^n, & x/t \leq a, \\ q_{i+1}^n, & x/t > a, \end{cases} \tag{4.30}$$

*and the solution at $\xi = 0$ is*

$$q^{GO}_{i+\frac{1}{2}} = q(x/t = 0) = \begin{cases} q^n_i, & 0 \leq a, \\ q^n_{i+1}, & 0 > a. \end{cases} \tag{4.31}$$

*For $a > 0$ we obtain*

$$q^{n+1}_i = q^n_i - \frac{\Delta t}{\Delta x}(aq^n_i - aq^n_{i-1}), \tag{4.32}$$

*which is identical with the upwind finite difference scheme. It is easy to prove that the upwind scheme is monotone under CFL condition since for $a > 0$ we get*

$$\frac{\partial H}{\partial q^n_i} = 1 - c > 0, \quad \forall 0 < c < 1, \qquad \frac{\partial H}{\partial q^n_{i+1}} = c > 0. \tag{4.33}$$

*For negative speed $a$ we obtain the corresponding upwind forward difference and monotonicity can be proven under the same condition.*

## 3. Approximate Riemann solvers

In general, computing the exact solutions of all the local Riemann problems arising at the element interfaces for the Godunov flux is very complex and time–consuming. Hence, a lot of effort has been put into the research on *approximate* Riemann solvers, which still take into account part of the physical wave structure of the Riemann problem at the element interface, but without aiming at its exact solution. Some popular approximate solvers are presented very briefly in the following. For details, the reader is referred to the excellent overview given in the textbook [**79**].

**3.1. Approximate-state Riemann solvers.** The approximate-state Riemann solvers provide an *approximate Godunov state* $Q^{AG}_{i+\frac{1}{2}}$ at the cell interface in order to compute the numerical flux $f_{i+\frac{1}{2}}$ following the original Godunov approach (4.14)

$$f_{i+\frac{1}{2}} = f_{i+\frac{1}{2}}\left(Q^n_i, Q^n_{i+1}\right) = f\left(Q^{AG}_{i+\frac{1}{2}}\right). \tag{4.34}$$

The approximation carried out to compute the state $Q^{AG}_{i+\frac{1}{2}}$ is not based on the exact resolution of the Riemann problem as proposed by Godunov, but one assumes *a priori* a *fixed wave pattern*, even if this is physically not the correct one. This simplification leads often (but not always) to a direct analytical calculation of the state $Q^*$

3.1.1. *The two-rarefaction approximate–state Riemann solver (TRRS).* The so-called two-rarefaction approximate–state Riemann solver supposes *a priori* that the solution of the Riemann problem consists of two rarefaction fans, as shown in Fig. 4. Here, we consider the particular case of isothermal gasdynamics. From the Riemann invariants we found that the velocity is constant across a contact discontinuity so that $u^*_L = u^*_R = u^*$. In particular, for isothermal gasdynamics the contact wave is in density equilibrium $\rho^*_L = \rho^*_R = \rho^*$. Using the Riemann invariants associated with the two rarefaction waves we obtain the following relation in $\rho^*$

$$a \ln \frac{\rho^*}{\rho_L} + a \ln \frac{\rho^*}{\rho_R} + u_R - u_L = 0, \tag{4.35}$$

the root of which provides the state $Q^* = (\rho^*, u^*)$ in the star region

$$\rho^* = \sqrt{\rho_L \rho_R \exp\left(\frac{u_L - u_R}{a}\right)}, \quad u^* = u_L + a \ln \frac{\rho_L}{\rho^*}. \tag{4.36}$$

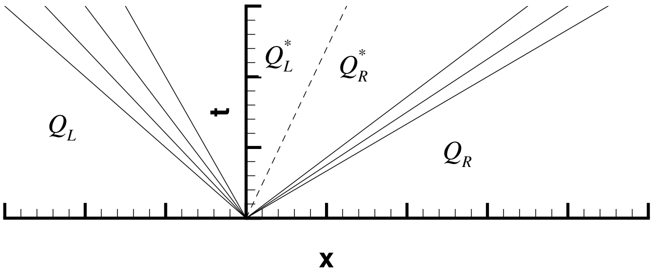

FIGURE 4. A solution of the Riemann problem that consists of two rarefaction waves and a contact discontinuity.

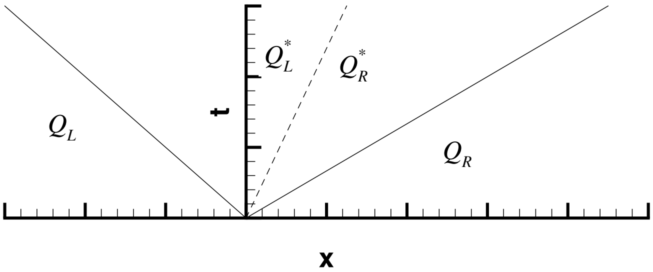

FIGURE 5. A solution of the Riemann problem that consists of two shock waves and a contact discontinuity.

Once the state $Q^*$ is known, one has to compute the approximate state $Q^{AG}_{i+\frac{1}{2}}$ as using the exact Riemann solver, but assuming the fixed structure with two rarefaction waves. Note that the two-rarefaction approximate state Riemann solver is exact when the exact solution actually consists only of two rarefaction waves separated by a contact discontinuity. The two–rarefaction approximate–state Riemann solver is consistent with the entropy condition, as the exact Riemann solver.

3.1.2. *The two-shock approximate–state Riemann solver (TSRS).* In this section we consider the two-shock approximate–state Riemann solver for isothermal gasdynamics. It assumes a simplified solution of the Riemann problem that consists of two shock waves. Fig. 5 shows the fixed solution that we presuppose. At the contact discontinuity we have $u^*_L = u^*_R = u^*$ and $\rho^*_L = \rho^*_R = \rho^*$. These relations have to be combined with the Rankine-Hugoniot conditions which are valid across the shock waves and we obtain

$$a\frac{\rho^* - \rho_L}{\sqrt{\rho^* \rho_L}} + a\frac{\rho^* - \rho_R}{\sqrt{\rho^* \rho_R}} + \overbrace{u_R - u_L}^{=\Delta u} = 0, \tag{4.37}$$

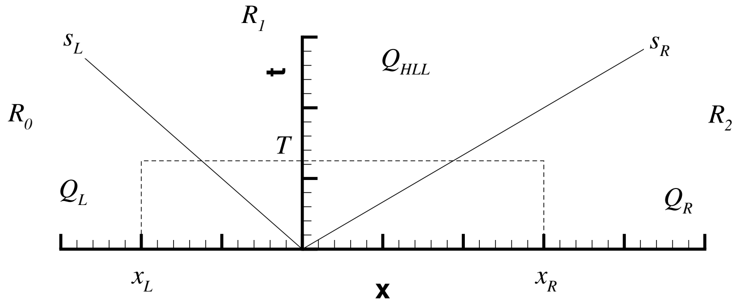

FIGURE 6. The solution assumed by the HLL method.

that is equivalent to

$$\rho^* + \frac{\Delta u}{a}\frac{\sqrt{\rho_R\rho_L}}{\sqrt{\rho_R}+\sqrt{\rho_L}}\sqrt{\rho^*} - \sqrt{\rho_L\rho_R} = 0, \tag{4.38}$$

Then, the state $Q^*$ is given by

$$\sqrt{\rho^*} = -\frac{1}{2a}\Delta u\frac{\sqrt{\rho_L\rho_R}}{\sqrt{\rho_L}+\sqrt{\rho_R}} + \frac{1}{2}\sqrt{\Delta u^2\frac{\rho_L\rho_R}{a^2\left(\sqrt{\rho_L}+\sqrt{\rho_R}\right)^2} + 4\sqrt{\rho_L\rho_R}}, \tag{4.39}$$

and

$$u^* = u_L - a\frac{\rho^* - \rho_L}{\sqrt{\rho^*\rho_L}}. \tag{4.40}$$

The computation of the approximate state $Q^{AG}_{i+\frac{1}{2}}$ occurs as with the exact Riemann solver, but assuming a fixed structure of the wave pattern of the solution consisting of two shock waves. If the exact solution actually contains two shock waves and a contact discontinuity, this approximate state Riemann solver is exact. Unfortunately, on the contrary of the exact Riemann solver and the two–rarefaction approximate–state Riemann solver, the two–shock approximate–state Riemann solver is *not* consistent with the entropy condition. This means that unphysical solutions such as rarefaction shocks are admitted by the method.

**3.2. The method of Harten-Lax-Leer (HLL).** Harten, Lax and van Leer (1983) [**41**] proposed an approximate Riemann solver assuming a wave structure that is still simpler than the solution of the Riemann problem and only includes the fastest waves, as plotted in Fig. 6. In the HLL approach the solution consists of three regions $R_0$, $R_1$, $R_2$ with piece-wise constant states. In $R_0$ we have the initial state on the left $Q_L$, and in $R_2$ we have the initial right state $Q_R$. In the region $R_1$ we assume the intermediate state $Q_{\text{HLL}}$, which is an approximate one and is still to be defined. It is computed starting from the PDE

$$\frac{\partial Q}{\partial t} + \frac{\partial f}{\partial x} = 0, \tag{4.41}$$

and integrating over the space-time control volume $[x_L; x_R] \times [0; T]$, sketched in dashed line in Fig. 6

$$\int\limits_{x_L}^{x_R} \left(Q\left(x,T\right) - Q\left(x,0\right)\right)dx + \int_0^T \left(f\left(Q\left(x_R,t\right)\right) - f\left(Q\left(x_L,t\right)\right)\right)dt = 0, \tag{4.42}$$

which results in

$$Q_L s_L + Q_{\text{HLL}} \left(s_R - s_L\right) - Q_R s_R + f_R - f_L = 0, \tag{4.43}$$

with $f_L = f\left(Q_L\right)$ and $f_R = f\left(Q_R\right)$. Finally, the approximate state $Q_{\text{HLL}}$ is given by

$$Q_{\text{HLL}} = \frac{s_R Q_R - s_L Q_L + f_L - f_R}{s_R - s_L}. \tag{4.44}$$

Note that the HLL method is *not* an approximate state Riemann solver,

$$f^{\text{HLL}}_{i+\frac{1}{2}} \neq f\left(Q_{\text{HLL}}\right)! \tag{4.45}$$

Indeed, it does not provide a state in order to compute the numerical flux as physical flux of that state. To compute the numerical flux using the HLL solver we have to apply *again* the PDE in integral form over the smaller space-time control volume $[0; x_R] \times [0; T]$ as follows

$$\int_0^{x_R} \left(Q\left(x, T\right) - Q\left(x, 0\right)\right) dx + \int_0^T \left(f\left(Q\left(x_R, t\right)\right) - f\left(Q\left(0, t\right)\right)\right) dt = 0, \tag{4.46}$$

that leads to

$$Q_{\text{HLL}}\, s_R T + Q_R \left(x_R - s_R T\right) - Q_R x_R + T f_R - T f^{\text{HLL}}_{i+\frac{1}{2}} = 0. \tag{4.47}$$

We finally obtain the numerical flux at the cell interface as

$$f^{\text{HLL}}_{i+\frac{1}{2}} = f_R + s_R \left(Q_{\text{HLL}} - Q_R\right). \tag{4.48}$$

Substituting (4.44) into the relation above we obtain the numerical flux $f^{\text{HLL}}_{i+\frac{1}{2}}$

$$f^{\text{HLL}}_{i+\frac{1}{2}} = \frac{s_R f_L - s_L f_R + s_L s_R \left(Q_R - Q_L\right)}{s_R - s_L}. \tag{4.49}$$

However, in the HLL method we still need two estimates for the two wave speeds $s_L$ and $s_R$. A simple choice has been proposed by Davis, as follows

$$s_L = \min\left(0, u_L - a, u_R - a\right), \quad s_R = \max\left(0, u_L + a, u_R + a\right). \tag{4.50}$$

Another option is based on the assumption of a single wave speed estimate

$$s_L = -s_{\max}, \quad s_R = +s_{\max}, \quad s_{\max} = \max\left(\left|u_L \pm a\right|, \left|u_R \pm a\right|\right), \tag{4.51}$$

that leads to the so-called Rusanov flux or local Lax–Friedrichs flux

$$f^R_{i+\frac{1}{2}} = \frac{s_{\max} f_L + s_{\max} f_R - s^2_{\max}\left(Q_R - Q_L\right)}{2 s_{\max}} = \frac{1}{2}\left(f_L + f_R\right) - \frac{1}{2} s_{\max}\left(Q_R - Q_L\right). \tag{4.52}$$

Note that using the following assumption

$$s_L = -\frac{\Delta x}{\Delta t}, \quad s_R = +\frac{\Delta x}{\Delta t}, \tag{4.53}$$

the HLL method becomes the classical Lax-Friedrichs flux (4.9).

Since the intermediate waves are ignored, the solution of the HLL Riemann solver produces excessive smearing of contact waves (and of shear–waves in multi–dimensional computations). Some modifications have been proposed in order to improve the HLL scheme. One is the HLLE scheme that uses a different estimate of the velocities $s_L$ and $s_R$ following the Einfeldt approach, for details see [**29**]. The HLLEM method assumes no longer a constant intermediate state, but a *piecewise linear* distribution. This approach allows to resolve contact discontinuities according to Einfeld et al. [**30**]. Among all the possible HLL schemes, perhaps the most commonly used approximate Riemann solver of the HLL type is the HLLC method proposed by Toro et al. [**82**], which captures steady contact waves exactly.

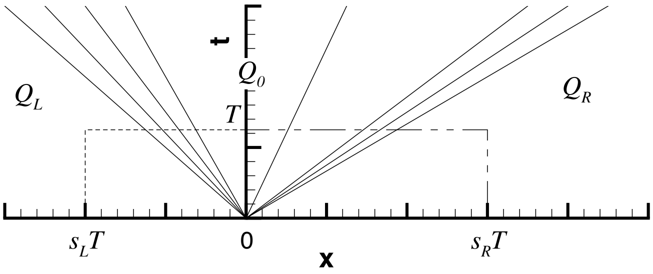

FIGURE 7. The space-time control volumes used to construct the Roe flux.

**3.3. The Roe method.** This approximate Riemann solver [**69**] is based on Roe's ingenious idea to apply a proper linearisation of the nonlinear Riemann problem (1.66) - (1.67). Using the quasi–linear form (3.95) of the PDE, Roe proposed to use some approximate Jacobian matrix in terms of the initial data $Q_L$ and $Q_R$, such that

$$\tilde{A} = \tilde{A}(Q_L, Q_R). \tag{4.54}$$

The matrix $\tilde{A}$ is called the *Roe matrix.* In order to obtain the Roe flux, a *linearized* Riemann problem with initial data $Q_L$ and $Q_R$ and fixed system matrix $\tilde{A}$ is considered:

$$\begin{array}{l} \frac{\partial \bar{Q}}{\partial t} + \tilde{A}\frac{\partial \bar{Q}}{\partial x} = 0, \\ \bar{Q}\,(x,0) = \left\{ \begin{array}{ll} Q_L, & \text{if } x < 0 \\ Q_R, & \text{if } x \geq 0. \end{array} \right. \end{array} \tag{4.55}$$

For the Roe matrix $\tilde{A}$ the following properties are enforced:

- Hyperbolicity. It is required that the Roe matrix $\tilde{A}$ has $m$ real eigenvalues and $m$ linear independent eigenvectors.
- Consistency of $\tilde{A}$ with the Jacobian matrix $A(Q)$, that is

$$\tilde{A}\,(Q,Q) = A\,(Q)\,. \tag{4.56}$$

- Conservation.

$$\tilde{A}\,(Q_L,Q_R) \cdot (Q_R - Q_L) = f\,(Q_R) - f\,(Q_L)\,. \tag{4.57}$$

The intercell flux is given by a similar procedure used in the HLL approach. We consider the original PDE (4.41) in conservative form and integrate twice, first over the space-time control volume on the left $[s_LT; 0] \times [0; T]$, then over the space-time control volume on the right $[0; s_RT] \times [0; T]$, see Fig. 7. We obtain

$$\int_{s_LT}^{0} Q\,(x,T)\,dx - (0 - s_LT)\,Q_L + Tf_{i+\frac{1}{2}} - Tf_L = 0, \tag{4.58}$$

$$\int_{0}^{s_RT} Q\,(x,T)\,dx - (s_RT - 0)\,Q_R + Tf_R - Tf_{i+\frac{1}{2}} = 0. \tag{4.59}$$

In order to determine the integrals in the above expressions, the exact solution $\bar{Q}$ of the linearized Riemann problem (4.55) is used. The integration over the left and the right control volumes gives

$$\int_{s_L T}^{0} \bar{Q}(x,T)\,dx - (0 - s_L T)\,Q_L + T\bar{F}\left(\bar{Q}_0\right) - T\bar{F}_L = 0, \tag{4.60}$$

$$\int_{0}^{s_R T} \bar{Q}(x,T)\,dx - (s_R T - 0)\,Q_R + T\bar{F}_R - T\bar{F}\left(\bar{Q}_0\right) = 0. \tag{4.61}$$

with $\bar{F}_{L,R} = \tilde{A}\cdot\bar{Q}_{L,R}$ and $\bar{Q}_0 = \bar{Q}(0,T)$. Note that $\bar{Q}_0$ is the exact solution of the linear Riemann problem (4.55) for $\frac{x}{t} = 0$. We now introduce the following hypothesis:

$$\int_{s_L T}^{0} Q(x,T)\,dx = \int_{s_L T}^{0} \bar{Q}(x,T)\,dx, \qquad \int_{0}^{s_R T} Q(x,T)\,dx = \int_{0}^{s_R T} \bar{Q}(x,T)\,dx, \tag{4.62}$$

Then, substituting into (4.60) and (4.61) we have

$$Tf_R - Tf_{i+\frac{1}{2}} - T\bar{F}_R + T\bar{F}\left(\bar{Q}_0\right) = 0, \quad \Rightarrow \quad f_{i+\frac{1}{2}} = f_R + \tilde{A}\left(\bar{Q}_0 - Q_R\right), \tag{4.63}$$

$$Tf_{i+\frac{1}{2}} - Tf_L - T\bar{F}\left(\bar{Q}_0 + T\bar{F}_L\right) = 0, \quad \Rightarrow \quad f_{i+\frac{1}{2}} = f_L + \tilde{A}\left(\bar{Q}_0 - Q_L\right). \tag{4.64}$$

Combining (4.63) with (4.64) we obtain the Roe flux as

$$f_{i+\frac{1}{2}} = \frac{1}{2}\left(f_R + f_L\right) + \tilde{A}\bar{Q}_0 - \frac{1}{2}\tilde{A}\left(Q_R + Q_L\right). \tag{4.65}$$

The exact solution of the linearized Riemann problem (4.55) at $x/t = 0$ denoted by $\bar{Q}_0$ is given according to (1.72) by

$$\bar{Q}_0 = \frac{1}{2}\tilde{R}\left(I + \operatorname{sign}\left(\tilde{\Lambda}\right)\right)\tilde{R}^{-1}Q_L + \frac{1}{2}\tilde{R}\left(1 - \operatorname{sign}(\tilde{\Lambda})\right)\tilde{R}^{-1}Q_R, \tag{4.66}$$

For the Roe flux we need the product $\tilde{A}\bar{Q}_0$, which is

$$\tilde{A}\bar{Q}_0 = \tilde{R}\tilde{\Lambda}\tilde{R}^{-1}\bar{Q}_0 = \frac{1}{2}\tilde{R}\left(\tilde{\Lambda} + \tilde{\Lambda}\operatorname{sign}(\tilde{\Lambda})\right)\tilde{R}^{-1}Q_L + \frac{1}{2}\tilde{R}\left(\tilde{\Lambda} - \tilde{\Lambda}\operatorname{sign}(\tilde{\Lambda})\right)\tilde{R}^{-1}Q_R. \tag{4.67}$$

With the definition of the absolute value $|\tilde{\Lambda}| = \tilde{\Lambda}\operatorname{sign}(\tilde{\Lambda})$ we have

$$\tilde{A}\bar{Q}_0 = \frac{1}{2}\tilde{R}\left(\tilde{\Lambda} + \left|\tilde{\Lambda}\right|\right)\tilde{R}^{-1}Q_L + \frac{1}{2}\tilde{R}\left(\tilde{\Lambda} - \left|\tilde{\Lambda}\right|\right)\tilde{R}^{-1}Q_R, \tag{4.68}$$

which can be written more compactly using the notation $|A| = R|\Lambda|R^{-1}$ as

$$\tilde{A}\bar{Q}_0 = \frac{1}{2}\left(\tilde{A} + |A|\right)Q_L + \frac{1}{2}\left(\tilde{A} - |A|\right)Q_R, \tag{4.69}$$

or

$$\tilde{A}\bar{Q}_0 = \frac{1}{2}\tilde{A}\left(Q_R + Q_L\right) - \frac{1}{2}\left|\tilde{A}\right|\left(Q_R - Q_L\right), \tag{4.70}$$

Then, the Roe flux finally results as

$$f^{\text{Roe}}_{i+\frac{1}{2}} = \frac{1}{2}\left(f\left(Q_R\right) + f\left(Q_L\right)\right) - \frac{1}{2}\left|\tilde{A}\right|\left(Q_R - Q_L\right). \tag{4.71}$$

This flux still requires the calculation of the Roe matrix $\tilde{A}$, which can be very complex. One usually applies the property (4.57) of matrix $\tilde{A}$ directly and uses a proper substitution of variables (parameter vector) in order to compute the so–called Roe state $\tilde{Q}$. Then, the Roe matrix $\tilde{A}$ is computed as the Jacobian of the PDE, evaluated at the Roe state

$$\tilde{A}\left(Q_L, Q_R\right) = A\left(\tilde{Q}\left(Q_L, Q_R\right)\right), \tag{4.72}$$

with $\tilde{Q}(Q,Q) = Q$. This assures that also the properties of system hyperbolicity and consistency are satisfied. Unfortunately, the Roe method is *not* entropy satisfying. Harten introduced a technique, the so–called Harten entropy–fix, see [**79**] for details, to avoid this problem. An important feature of Roe's scheme is that it resolves exactly all stationary discontinuities.

EXAMPLE 31. *We use the property* (4.57) *to compute the Roe state* $\tilde{Q}$ *for isothermal gasdynamics assuming that* $\tilde{A} = A\left(\tilde{Q}(Q_L, Q_R)\right)$

(4.73)
$$\begin{pmatrix} 0 & 1 & 0 \\ a^2 - \tilde{u}^2 & 2\tilde{u} & 0 \\ -\tilde{u}\tilde{\psi} & \tilde{\psi} & \tilde{u} \end{pmatrix} \begin{pmatrix} \rho_R - \rho_L \\ \rho_R u_R - \rho_L u_L \\ \rho_R \psi_R - \rho_L \psi_L \end{pmatrix} = \begin{pmatrix} \rho_R u_R - \rho_L u_L \\ \rho_R u_R^2 + a^2 \rho_R - \rho_L u_L^2 - a^2 \rho_L \\ \rho_R u_R \psi_R - \rho_L u_L \psi_L \end{pmatrix}.$$

*Solving the system above we obtain the Roe state for isothermal gasdynamics*

$$(4.74)\qquad (\tilde{\rho} = \sqrt{\rho_L \rho_R}), \quad \tilde{u} = \frac{u_L\sqrt{\rho_L} + u_R\sqrt{\rho_R}}{\sqrt{\rho_L} + \sqrt{\rho_R}}, \quad \tilde{\psi} = \frac{\psi_L\sqrt{\rho_L} + \psi_R\sqrt{\rho_R}}{\sqrt{\rho_L} + \sqrt{\rho_R}}.$$

As an alternative, Toumi (1992) [**86**] proposed to construct the Roe matrix $\tilde{A}$ by using an integral (weak) formulation of the conservation property (4.57), as follows

$$(4.75)\qquad \tilde{A}(Q_R - Q_L) = \int_{Q_L}^{Q_R} A(Q)\, dQ = \int_{Q_L}^{Q_R} \frac{\partial f}{\partial Q} dQ = \int_{f_L}^{f_R} df = f_R - f_L.$$

Following a particular integration path, namely the simple segment–path

$$(4.76)\qquad Q(s) = Q_L + s(Q_R - Q_L), \quad 0 \le s \le 1,$$

we obtain

$$(4.77)\qquad \tilde{A}(Q_R - Q_L) = \int_{Q_L}^{Q_R} A(Q)\, dQ = \left(\int_0^1 A(Q(s))\, ds\right)(Q_R - Q_L),$$

that is equivalent to

$$(4.78)\qquad \tilde{A} = \int_0^1 A(Q(s))\, ds.$$

With this weak formulation based on the segment path, the resulting Roe–type flux becomes

$$(4.79)\qquad f_{i+\frac{1}{2}}^{\mathrm{Roe}'} = \frac{1}{2}(f(Q_R) + f(Q_L)) - \frac{1}{2}\left|\int_0^1 A(Q(s))\, ds\right|(Q_R - Q_L).$$

The integral in (4.79) can be numerically approximated using any classical quadrature formula, such as the trapezoidal rule, or standard Newton-Cotes or Gauss formulae. The purely numerical computation of the Roe matrix has first been proposed in [**23, 28, 27, 8, 7**]. Note that the Roe matrix (4.78) in integral form satisfies the consistency (4.56) and the conservation (4.57) properties by construction, but the hyperbolicity of the system (the first condition for the Roe matrix) is *not* generally guaranteed. Moreover, also this version of the Roe method is *not* consistent with the entropy condition.

**3.4. The Osher method.** Osher and Solomon (1982) [**62**] proposed the following formulation of the numerical flux

$$(4.80)\qquad f_{i+\frac{1}{2}}^{OS} = \frac{1}{2}(f(Q_R) + f(Q_L)) - \frac{1}{2}\int_{Q_L}^{Q_R} |A(Q)|\, dQ.$$

Note that the scheme above is differentiable due to its integral formulation. The original method of Osher and Solomon for nonlinear hyperbolic systems is very

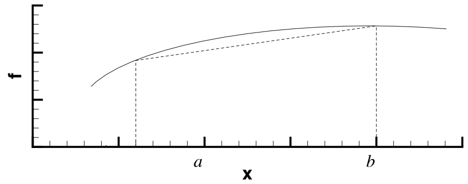

FIGURE 8. The approximation (dashed line) for the integral of a function (solid line) through the trapezoidal rule.

complex, since the exact computation of the integral for the dissipation term is not a trivial task. For more details and test problems see [**79**]. In [**27**], we first proposed a simple segment–path defined by (4.76) to calculate the integral in (4.80), as follows

$$f^{OS}_{i+\frac{1}{2}} = \frac{1}{2}\left(f\left(Q_R\right) + f\left(Q_L\right)\right) - \frac{1}{2}\left(\int_0^1 \left|A\left(Q\left(s\right)\right)\right| ds\right)\left(Q_R - Q_L\right). \tag{4.81}$$

Compare the new formulation of the Osher–Solomon flux (4.81) with the Roe flux given by (4.79).

It is to remark that the method of Osher and Solomon is consistent with the entropy principle and solves the contact waves exactly. The hyperbolicity of the Jacobian matrix $A\left(Q\left(s\right)\right)$ is guaranteed. Moreover, the implementation is straightforward when a numerical quadrature formula for the integral approximation is used. A choice consists of the trapezoidal rule, given by

$$\int_a^b f\left(x\right) dx \approx \frac{1}{2}\left(b-a\right)\left(f\left(a\right) + f\left(b\right)\right). \tag{4.82}$$

A sketch of the basic idea for the integral approximation is plotted in Fig. 8. With the trapezoidal rule the Osher–Solomon–type flux results in

$$f^{OS}_{i+\frac{1}{2}} \approx \frac{1}{2}\left(f\left(Q_R\right) + f\left(Q_L\right)\right) - \frac{1}{2}\frac{\left|A\left(Q_L\right)\right| + \left|A\left(Q_R\right)\right|}{2}\left(Q_R - Q_L\right). \tag{4.83}$$

Similarly, for the Roe flux (4.79) one would have

$$f^{R}_{i+\frac{1}{2}} \approx \frac{1}{2}\left(f\left(Q_R\right) + f\left(Q_L\right)\right) - \frac{1}{2}\left|\frac{A\left(Q_L\right) + A\left(Q_R\right)}{2}\right|\left(Q_R - Q_L\right). \tag{4.84}$$

## 4. Total Variation Diminishing (TVD) Schemes

Due to the Godunov theorem treated in Section 4, there exists no *linear* method of order of accuracy greater or equal than two that also assures the *monotonicity* of the scheme. To overcome the constraint expressed in this theorem we use the fact that the theorem only applies to linear methods. To construct monotone methods of second order of accuracy, we therefore have to develop *nonlinear* methods.

### 4.1. Definitions and the TVD criterion of Harten.

The total variation (TV) of a discrete function $q^n_h(x_h)$ is defined as

$$TV(q^n_h) = \sum_{i=-\infty}^{\infty} |q^n_{i+1} - q^n_i|, \tag{4.85}$$

with $q^n_i \rightarrow const.$ for $i \rightarrow \pm\infty$. A numerical method

$$q^{n+1}_i = H(q^n_{i-l}, ...q^n_i, ...q^n_{i+r}) \tag{4.86}$$

is called TVD if

$$\mathrm{TV}(q^{n+1}_h) \leq \mathrm{TV}(q^n_h). \tag{4.87}$$

Theorem 2. *(Harten 1983). The set of monotone methods $S_{\mathrm{mon}}$ is a subset of TVD methods $S_{\mathrm{TVD}}$.*

$$S_{\mathrm{mon}} \subset S_{\mathrm{TVD}}. \tag{4.88}$$

Proof. For the proof see [**39**]. □

Theorem 3. *(Harten 1983). A numerical method of the form*

$$q_i^{n+1} = q_i^n - C_{i-\frac{1}{2}} \Delta q_{i-\frac{1}{2}}^n + D_{i+\frac{1}{2}} \Delta q_{i+\frac{1}{2}}^n, \tag{4.89}$$

*with $\Delta q_{i-\frac{1}{2}}^n = q_i^n - q_{i-1}^n$ and $\Delta q_{i+\frac{1}{2}}^n = q_{i+1}^n - q_i^n$ is TVD if the following sufficient condition is satisfied:*

$$C_{i+\frac{1}{2}} \geq 0, \quad D_{i+\frac{1}{2}} \geq 0, \quad 0 \leq C_{i+\frac{1}{2}} + D_{i+\frac{1}{2}} \leq 1. \tag{4.90}$$

Proof. The numerical solution at grid points $x_i$ and $x_{i+1}$ at the new time $t^{n+1}$ is given by

$$\begin{aligned} q_i^{n+1} &= q_i^n - C_{i-\frac{1}{2}}(q_i^n - q_{i-1}^n) + D_{i+\frac{1}{2}}(q_{i+1}^n - q_i^n), \\ q_{i+1}^{n+1} &= q_{i+1}^n - C_{i+\frac{1}{2}}(q_{i+1}^n - q_i^n) + D_{i+\frac{3}{2}}(q_{i+2}^n - q_{i+1}^n), \end{aligned}$$

hence

$$q_{i+1}^{n+1} - q_i^{n+1} = (q_{i+1}^n - q_i^n)(1 - C_{i+\frac{1}{2}} - D_{i+\frac{1}{2}}) + D_{i+\frac{3}{2}}(q_{i+2}^n - q_{i+1}^n) + C_{i-\frac{1}{2}}(q_i^n - q_{i-1}^n). \tag{4.91}$$

Using the triangle inequality $|a+b| \leq |a| + |b|$ and $|ab| = |a||b|$ we obtain

$$|q_{i+1}^{n+1} - q_i^{n+1}| \leq |q_{i+1}^n - q_i^n||1 - C_{i+\frac{1}{2}} - D_{i+\frac{1}{2}}| + |D_{i+\frac{3}{2}}||q_{i+2}^n - q_{i+1}^n| + |C_{i-\frac{1}{2}}||q_i^n - q_{i-1}^n|. \tag{4.92}$$

Using the hypothesis of the theorem (4.90) we have

$$\begin{aligned} |q_{i+1}^{n+1} - q_i^{n+1}| \leq |q_{i+1}^n - q_i^n| + C_{i-\frac{1}{2}}|q_i^n - q_{i-1}^n| - C_{i+\frac{1}{2}}|q_{i+1}^n - q_i^n| + \\ D_{i+\frac{3}{2}}|q_{i+2}^n - q_{i+1}^n| - D_{i+\frac{1}{2}}|q_{i+1}^n - q_i^n|. \end{aligned} \tag{4.93}$$

With this intermediate result we have for the total variation at time $t^{n+1}$

$$\begin{aligned} \mathrm{TV}(q_h^{n+1}) = \sum_i |q_{i+1}^{n+1} - q_i^{n+1}| \leq \sum_i |q_{i+1}^n - q_i^n| + \\ \overbrace{\sum_i \Big( C_{i-\frac{1}{2}}|q_i^n - q_{i-1}^n| - C_{i+\frac{1}{2}}|q_{i+1}^n - q_i^n| \Big)}^{=0} + \\ \overbrace{\sum_i \Big( D_{i+\frac{3}{2}}|q_{i+2}^n - q_{i+1}^n| - D_{i+\frac{1}{2}}|q_{i+1}^n - q_i^n| \Big)}^{=0}. \end{aligned} \tag{4.94}$$

The last two sums vanish due to their telescopic property, hence we obtain

$$\mathrm{TV}(q_h^{n+1}) \leq \mathrm{TV}(q_h^n), \tag{4.95}$$

which concludes the proof.

□

**4.2. Higher order spatial reconstruction and the method of Kolgan.** The basic idea consists of changing of the data approximation inside each cell from a piece–wise constant distribution to piece–wise linear data. Since only cell averaged quantities are available, the slopes of the piecewise linear approximation have to be computed appropriately from the cell average of the cell and its neighbors. A sketch of piecewise constant and piecewise linear data is depicted in Fig. 9. Inside each cell we suppose a linear polynomial $P_i^n$ for the state vector $Q$ at time

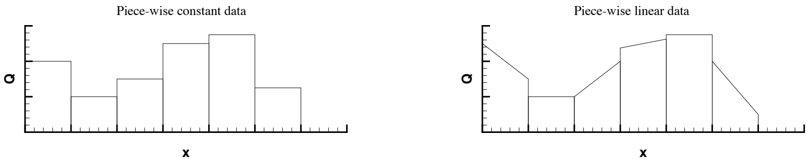

FIGURE 9. Piece–wise constant data (left) and piece-wise linear data (right). Note the discontinuities at the element interfaces.

$t^n$, such that

$$P_i^n(x) = a + b(x - x_i), \tag{4.96}$$

where the index $i$ denotes the mesh element and $n$ is the time step number. The coefficients $a$ and $b$ are computed by using the principle of conservation. We first use a right–sided reconstruction stencil composed of the element itself and its right neighbor to determine a right reconstruction polynomial $P_i^R$, as follows

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} P_i^R(x)\,dx = Q_i^n, \quad \Rightarrow \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} a + b(x - x_i)\,dx = Q_i^n \Delta x, \tag{4.97}$$

$$\frac{1}{\Delta x}\int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} P_i^R(x)\,dx = Q_{i+1}^n, \quad \Rightarrow \int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} a + b(x - x_i)\,dx = Q_{i+1}^n \Delta x. \tag{4.98}$$

Integration yields

$$\left[ax + \frac{b}{2}(x - x_i)^2\right]_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} = Q_i^n \Delta x, \quad \Rightarrow a = Q_i^n, \tag{4.99}$$

$$\left[ax + \frac{b}{2}(x - x_i)^2\right]_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} = Q_{i+1}^n \Delta x, \quad \Rightarrow b = \frac{Q_{i+1}^n - Q_i^n}{\Delta x} = \frac{\Delta Q_{i+\frac{1}{2}}^n}{\Delta x}. \tag{4.100}$$

Similarly, imposing the conservation at the left side of the cell interface we compute the coefficients $a$ and $b$

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} P_i^L(x)\,dx = Q_i^n, \quad \Rightarrow a = Q_i^n, \tag{4.101}$$

$$\frac{1}{\Delta x}\int_{x_{i-\frac{3}{2}}}^{x_{i-\frac{1}{2}}} P_i^L(x)\,dx = Q_{i-1}^n, \quad \Rightarrow b = \frac{Q_i^n - Q_{i-1}^n}{\Delta x} = \frac{\Delta Q_{i-\frac{1}{2}}^n}{\Delta x}. \tag{4.102}$$

Since we have two spatial reconstruction polynomials available, one from the left and one from the right, we now can make the scheme *nonlinear* by letting the actual choice of which reconstruction polynomial we finally retain depend on the data. One option that satisfies the maximum principle (no generation of new extrema) consists of

$$\Delta Q_i^n = \text{minmod}\left(\Delta Q_{i-\frac{1}{2}}^n, \Delta Q_{i+\frac{1}{2}}^n\right), \tag{4.103}$$

with $\Delta Q_{i-\frac{1}{2}}^n = Q_i^n - Q_{i-1}^n$ and $\Delta Q_{i+\frac{1}{2}}^n = Q_{i+1}^n - Q_i^n$, where the so–called minmod function is defined by

$$\text{minmod}(a, b) = \begin{cases} 0, & \text{if } ab \le 0, \\ a, & \text{if } |a| < |b| \\ b, & \text{if } |a| \ge |b|. \end{cases} \tag{4.104}$$

Then, the linear polynomial for the cell $i$ becomes

$$P_i^n(x) = Q_i^n + \frac{\Delta Q_i^n}{\Delta x}(x - x_i), \tag{4.105}$$

Once the polynomial is defined, we can compute the states at the cell interfaces $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$

$$Q_{i-\frac{1}{2}}^{n,+} = Q_i^n + \frac{\Delta Q_i^n}{\Delta x}\left(x_i - \frac{\Delta x}{2} - x_i\right) = Q_i^n - \frac{1}{2}\Delta Q_i^n, \tag{4.106}$$

$$Q_{i+\frac{1}{2}}^{n,-} = Q_i^n + \frac{\Delta Q_i^n}{\Delta x}\left(x_i + \frac{\Delta x}{2} - x_i\right) = Q_i^n + \frac{1}{2}\Delta Q_i^n. \tag{4.107}$$

The numerical flux is computed including the values at the interfaces from the polynomials $P_i^n(x)$ and $P_{i+1}^n(x)$

$$f_{i+\frac{1}{2}} = f_{i+\frac{1}{2}}\left(Q_{i+\frac{1}{2}}^{n,-}, Q_{i+\frac{1}{2}}^{n,+}\right), \tag{4.108}$$

with $Q_{i+\frac{1}{2}}^{n,+} = Q_{i+1}^n - \frac{1}{2}\Delta Q_{i+1}^n$ and $Q_{i+\frac{1}{2}}^{n,-} = Q_{i+1}^n + \frac{1}{2}\Delta Q_i^n$.

Kolgan (1972) [**50**] proposed the following scheme

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2}}\left(Q_{i+\frac{1}{2}}^{n,-}, Q_{i+\frac{1}{2}}^{n,+}\right) - f_{i-\frac{1}{2}}\left(Q_{i-\frac{1}{2}}^{n,-}, Q_{i-\frac{1}{2}}^{n,+}\right)\right). \tag{4.109}$$

To compute the numerical flux any Riemann solver can be used. This method is of second order of accuracy in space and of first order in time. It is linearly unconditionally unstable, but it is nonlinearly stable for a Courant number $c < \frac{1}{2}$.

**4.3. Second order in space and in time: the MUSCL method of van Leer (1979).** An significant improvement in accuracy and stability is given by the MUSCL method of van Leer [**87**], who proposed to use a linear polynomial in space *and in time*, as follows

$$P_i^n(x,t) = Q_i^n + \frac{\Delta Q_i^n}{\Delta x}(x - x_i) + \partial_t Q_i (t - t^n), \tag{4.110}$$

where the term $Q_t$ indicates the derivative in time of the variable $Q$. It can be computed using the PDE and the resulting spatial derivative of the flux is approximated by a central finite difference with respect to the cell center $x_i$ by using the boundary–extrapolated values from *within* the cell:

$$\partial_t Q_i = \frac{\partial Q}{\partial t} = -\frac{\partial f}{\partial x} \approx -\frac{f\left(Q_{i+\frac{1}{2}}^{n,-}\right) - f\left(Q_{i-\frac{1}{2}}^{n,+}\right)}{\Delta x} = \frac{f\left(Q_{i-\frac{1}{2}}^{n,+}\right) - f\left(Q_{i+\frac{1}{2}}^{n,-}\right)}{\Delta x}. \tag{4.111}$$

Substituting into (4.110) we have

$$P_i^n(x,t) = Q_i^n + \frac{\Delta Q_i^n}{\Delta x}(x - x_i) + \frac{f\left(Q_{i-\frac{1}{2}}^{n,+}\right) - f\left(Q_{i+\frac{1}{2}}^{n,-}\right)}{\Delta x}(t - t^n). \tag{4.112}$$

Since the numerical flux is defined by

$$f_{i+\frac{1}{2}} = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f\left(Q(x_{i+\frac{1}{2}}, t)\right) dt, \tag{4.113}$$

we have to compute the time integral above numerically. To achieve the second order of accuracy we have two possibilities

- The trapezoidal rule

$$\int_{t^n}^{t^{n+1}} f\left(Q(x_{i+\frac{1}{2}}, t)\right) dt \approx \frac{\Delta t}{2}\left(f\left(Q\left(x_{i+\frac{1}{2}}, t^n\right)\right) + f\left(Q\left(x_{i+\frac{1}{2}}, t^{n+1}\right)\right)\right). \tag{4.114}$$

- The mid–point rule

$$\int_{t^n}^{t^{n+1}} f\left(Q(x_{i+\frac{1}{2}},t)\right) dt \approx \Delta t f\left(Q(x_{i+\frac{1}{2}},t^{n+\frac{1}{2}})\right) = \Delta t f\left(Q_{i+\frac{1}{2}}^{n+\frac{1}{2}}\right), \tag{4.115}$$

Note that the mid–point integration is cheaper than the trapezoidal rule, hence it is used in the following.

Finally, the MUSCL (monotone upwind scheme for conservation laws) method of van Leer is given by

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2}}\left(Q_{i+\frac{1}{2}}^{n+\frac{1}{2},-},Q_{i+\frac{1}{2}}^{n+\frac{1}{2},+}\right) - f_{i-\frac{1}{2}}\left(Q_{i-\frac{1}{2}}^{n+\frac{1}{2},-},Q_{i-\frac{1}{2}}^{n+\frac{1}{2},+}\right)\right). \tag{4.116}$$

It is second order accurate in space and time and is linearly stable up to Courant number one.

## 5. The MUSCL method on two–dimensional Cartesian grids

We consider the hyperbolic PDE

$$\frac{\partial Q}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0. \tag{4.117}$$

A finite volume scheme based on rectangular spatial control volumes $T_{ij} = [x_{i-\frac{1}{2}}; x_{i+\frac{1}{2}}] \times [y_{i-\frac{1}{2}}; y_{i+\frac{1}{2}}]$ becomes

$$Q_{ij}^{n+1} = Q_{ij}^n - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j}\right) - \frac{\Delta t}{\Delta y}\left(g_{i,j+\frac{1}{2}} - g_{i,j-\frac{1}{2}}\right). \tag{4.118}$$

In order to compute the numerical fluxes across the element interfaces using the second order MUSCL scheme one proceeds as in one space dimension, obtaining:

(1) TVD reconstruction of the gradients (slopes) based on the solution at time $t^n$:

$$\begin{aligned}\Delta Q_{ij}^x &= \text{minmod}(Q_{i+1,j}^n - Q_{i,j}^n, Q_{i,j}^n - Q_{i-1,j}^n),\\ \Delta Q_{ij}^y &= \text{minmod}(Q_{i,j+1}^n - Q_{i,j}^n, Q_{i,j}^n - Q_{i,j-1}^n).\end{aligned} \tag{4.119}$$

(2) The space–time polynomial in each cell $T_{ij}$ which is valid for the current time step is written in terms of a space–time Taylor series expanded about $x_i$, $y_i$ and $t^n$ as

$$Q_{ij}(x,y,t) = Q_{ij}^n + \Delta Q_{ij}^x/\Delta x(x-x_i) + \Delta Q_{ij}^y/\Delta y(y-y_i) + \partial_t Q_{ij}(t-t^n). \tag{4.120}$$

(3) The time derivative in the Taylor series is computed based on the PDE as

$$\partial_t Q_{ij} = -f_x - g_y. \tag{4.121}$$

We now use a central finite difference with respect to the element barycenter and based on the reconstructed values *from within* the element $T_{ij}$ to approximate the spatial derivatives of the fluxes. This yields

$$\partial_t Q_{ij} \approx -\frac{f(Q_{i+\frac{1}{2},j}^{n,-}) - f(Q_{i-\frac{1}{2},j}^{n,+})}{\Delta x} - \frac{g(Q_{i,j+\frac{1}{2}}^{n,-}) - g(Q_{i,j-\frac{1}{2}}^{n,+})}{\Delta y}, \tag{4.122}$$

with the boundary extrapolated values at the old time $t^{n+1}$

$$Q_{i\pm\frac{1}{2},j}^{n,\mp} = Q_{ij}^n \pm \frac{1}{2}\Delta Q_{ij}^x, \qquad Q_{i,j\pm\frac{1}{2}}^{n,\mp} = Q_{ij}^n \pm \frac{1}{2}\Delta Q_{ij}^y. \tag{4.123}$$

(4) Compute the boundary extrapolated values at the half time level $t^{n+\frac{1}{2}}$ as

$$Q_{i\pm\frac{1}{2},j}^{n+\frac{1}{2},\mp} = Q_{ij}^n \pm \frac{1}{2}\Delta Q_{ij}^x + \frac{1}{2}\partial_t Q_{ij}, \qquad Q_{i,j\pm\frac{1}{2}}^{n+\frac{1}{2},\mp} = Q_{ij}^n \pm \frac{1}{2}\Delta Q_{ij}^y + \frac{1}{2}\partial_t Q_{ij}. \tag{4.124}$$

(5) Insert these values into the numerical fluxes of the finite volume scheme (4.118) and update the solution to the new time $t^{n+1}$.

## 6. Multidimensional finite volume schemes on general meshes

The sketch of the general multidimensional case of the finite volume method for arbitrary unstructured meshes is given in the following. We consider the multi–dimensional nonlinear system of hyperbolic conservation laws

$$\frac{\partial Q}{\partial t}+\frac{\partial f}{\partial x}+\frac{\partial g}{\partial y}+\frac{\partial h}{\partial z}=0, \tag{4.125}$$

where $Q, f, g, h \in \mathbb{R}^m$ and $\vec{x}=(x,y,z)\in\mathbb{R}^3$. The PDE (4.125) in compact form results

$$\frac{\partial Q}{\partial t}+\mathrm{div}F=0, \quad \text{or} \quad \frac{\partial Q}{\partial t}+\nabla\cdot F=0, \tag{4.126}$$

where $F$ denotes the flux tensor $F=(f,g,h)$. The PDE integrated over a general space-time control volume $T_i\times\left[t^n,t^{n+1}\right]$ gives the integral form of (4.126)

$$\int_{T_i}\int_{t^n}^{t^{n+1}}\frac{\partial Q}{\partial t}dtdV+\int_{T_i}\int_{t^n}^{t^{n+1}}\nabla\cdot FdtdV=0. \tag{4.127}$$

The volume integral is converted to a surface integral applying the Gauss–Ostrogradski theorem

$$\int_{T_i}\left(Q\left(\vec{x},t^{n+1}\right)-Q\left(\vec{x},t^n\right)\right)dV+\int_{t^n}^{t^{n+1}}\int_{\partial T_i}F\cdot\vec{n}\,dtdS=0, \tag{4.128}$$

where $\partial T_i$ is the boundary of the $i^{th}$ control volume and $\vec{n}_j$ denotes the outward pointing unit normal vector to the surface $\partial T_i$. We introduce the following definitions

(4.129)

$$Q_i^n=\frac{1}{|V_i|}\int_{T_i}Q\left(\vec{x},t^n\right)dV, \quad F_{i,j+\frac{1}{2}}\cdot\vec{n}_j=\frac{1}{\Delta t}\frac{1}{|\partial T_i|}\int_{t^n}^{t^{n+1}}\int_{\partial T_i}F\cdot\vec{n}_jdtdV=0,$$

where the index $j$ indicates the $j^{th}$ side of the control volume $T_i$, which is supposed to have a polygonal shape, i.e. the boundary is composed of piecewise linear faces, as plotted in Fig. 10, and $J$ is the number of faces. We obtain the following *exact* relation

$$Q_i^{n+1}=Q_i^n-\frac{\Delta t}{|T_i|}\sum_{j=1}^{J}\left|\partial T_j\right|F_{i,j+\frac{1}{2}}\cdot\vec{n}_j. \tag{4.130}$$

Inserting any classical numerical flux into the term $F_{i,j+\frac{1}{2}}\cdot\vec{n}_j$ will then give a finite volume scheme on general polygonal control volumes.

## 7. Rotational invariance

Newtonian mechanics is invariant with respect to a rotation of the coordinate system. As a consequence we expect that also the governing PDE of Newtonian mechanics are rotationally invariant. Let the angle of rotation be $\alpha$, then the vector which defines the first axis of the new coordinate system is defined as

$$\vec{n}=\left(\begin{array}{c}\cos\alpha\\ \sin\alpha\end{array}\right). \tag{4.131}$$

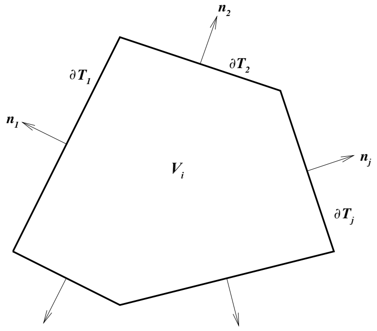

FIGURE 10. A polygonal element of volume $V_i$ and total surface $\partial T_i$.

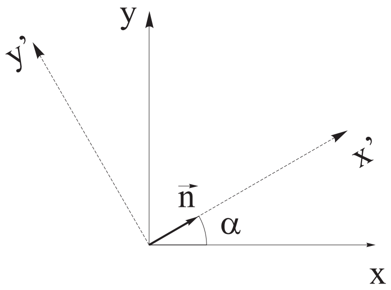

FIGURE 11. Rotation of the coordinate system by an angle $\alpha$.

The vector $\vec{x}$ of the old coordinates $x$ and $y$ is given in terms of the vector $\vec{x}'$ of the new coordinates $x'$ and $y'$ as

$$\vec{x} = M\vec{x}', \qquad \text{with } M = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}. \tag{4.132}$$

For a sketch of the coordinate rotation see 11. The same transformation applies to all vector quantities, such as velocity $\vec{v} = M\vec{v}'$, while the scalars (density, pressure, energy) do *not* change with the transformation. Let $T$ be the transformation matrix which expresses the coordinate transformation of the vector of conserved quantities. For scalar quantities, $T$ contains a row of zeros with a value of one on the diagonal, whereas for vector quantities, $T$ contains a double row of zeros with the transformation matrix $M$ on the corresponding diagonal block. We then have

$$Q = TQ', \qquad \text{and} \qquad Q' = T^{-1}Q = T^TQ. \tag{4.133}$$

DEFINITION 2. *A system of PDE of the form*

$$\frac{\partial Q}{\partial t} + \nabla \cdot F(Q) = 0, \qquad \text{with } F(Q) = (f, g) \tag{4.134}$$

is said to be rotationally invariant if

$$F(Q) \cdot \vec{n} = Tf(T^{-1}Q). \tag{4.135}$$

EXERCISE 5. *Prove that the Euler equations of compressible gasdynamics and the shallow water equations are rotationally invariant.*

## 8. The wave–propagation form of finite volume schemes

This particular formulation of conservative finite volume schemes is due to LeVeque [**55**] and gives an interesting alternative interpretation of the Riemann problems at the element interfaces. Given an explicit conservative method

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}\right), \tag{4.136}$$

with the Roe flux defined as

$$f_{i+\frac{1}{2}} = \frac{1}{2}\left(f(Q_{i+1}^n) + f(Q_i^n)\right) - \frac{1}{2}|\tilde{A}_{i+\frac{1}{2}}|\left(Q_{i+1}^n - Q_i^n\right), \tag{4.137}$$

we have by adding and subtracting $f_i = f(Q_i^n)$

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(f_{i+\frac{1}{2}} - f_i - (f_{i-\frac{1}{2}} - f_i)\right). \tag{4.138}$$

The right flux difference can be written as

$$f_{i+\frac{1}{2}} - f_i = \frac{1}{2}\left(f(Q_{i+1}^n) - f(Q_i^n)\right) - \frac{1}{2}|\tilde{A}_{i+\frac{1}{2}}|\left(Q_{i+1}^n - Q_i^n\right) \tag{4.139}$$

and using the property of the Roe matrix we have

$$f_{i+\frac{1}{2}} - f_i = \frac{1}{2}\tilde{A}_{i+\frac{1}{2}}\left(Q_{i+1}^n - Q_i^n\right) - \frac{1}{2}|\tilde{A}_{i+\frac{1}{2}}|\left(Q_{i+1}^n - Q_i^n\right), \tag{4.140}$$

or simply

$$f_{i+\frac{1}{2}} - f_i = \frac{1}{2}\left(\tilde{A}_{i+\frac{1}{2}} - |\tilde{A}_{i+\frac{1}{2}}|\right)\left(Q_{i+1}^n - Q_i^n\right) = \tilde{A}_{i+\frac{1}{2}}^-\left(Q_{i+1}^n - Q_i^n\right) := D_{i+\frac{1}{2}}^-. \tag{4.141}$$

A similar manipulation can be also performed on the left hand side, obtaining:

$$f_{i-\frac{1}{2}} - f_i = -\frac{1}{2}\left(\tilde{A}_{i-\frac{1}{2}} + |\tilde{A}_{i-\frac{1}{2}}|\right)\left(Q_i^n - Q_{i-1}^n\right) = \tilde{A}_{i-\frac{1}{2}}^+\left(Q_i^n - Q_{i-1}^n\right) := D_{i-\frac{1}{2}}^+. \tag{4.142}$$

Inserting in the finite volume method we obtain the wave propagation form of LeVeque

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(D_{i+\frac{1}{2}}^- + D_{i-\frac{1}{2}}^+\right), \tag{4.143}$$

or with $\Delta Q_{i-\frac{1}{2}} = Q_i^n - Q_{i-1}^n$ and $\Delta Q_{i+\frac{1}{2}} = Q_{i+1}^n - Q_i^n$ we obtain the new cell average at time $t^{n+1}$ in terms of the right–moving waves that enter from the left and the left–moving waves that enter from the right, both expressed as a product of the positive or negative part of the Roe–matrix times the jump of the vector of conserved quantities at the interface:

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(\tilde{A}_{i+\frac{1}{2}}^-\Delta Q_{i+\frac{1}{2}}^n + \tilde{A}_{i-\frac{1}{2}}^+\Delta Q_{i-\frac{1}{2}}^n\right). \tag{4.144}$$

A graphical interpretation of this scheme is sketched in Fig. 12. Remember that in the method of Roe, the continuous rarefaction waves are replaced by rarefaction shocks. In fact, also the original method of Godunov [**37**] was expressed in the form illustrated by Fig. 12 and *not* in the usual conservative flux divergence form.

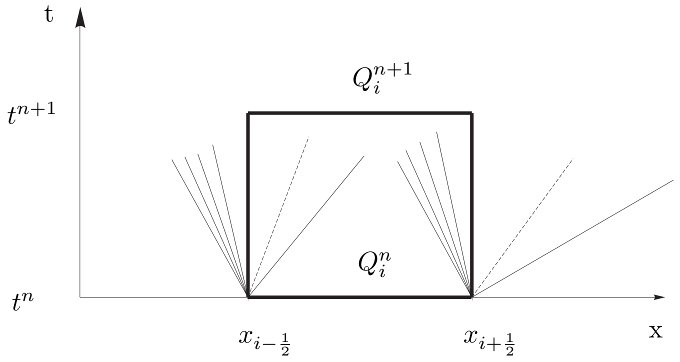

FIGURE 12. Graphical interpretation of the wave–propagation algorithm of LeVeque. The cell–average $Q_i$ is only modified by waves that *enter* the control volume from the boundaries.

# Part 2

# Advanced Algorithms for the Discretization of Hyperbolic PDE

CHAPTER 5

# Path–conservative finite volume schemes

In the previous part of this manuscript we have dealt with *conservative* hyperbolic systems. Actually for many physically relevant processes, it is possible to write the problem in conservative form, see for example the Euler or Navier–Stokes equations of compressible gas dynamics, the equations of magneto–hydrodynamics (MHD) or the equations governing nonlinear elasticity. However, there are also many physically relevant systems that can *not* be written in conservation form, in particular problems arising in multi–phase and multi–fluid flow physics. We therefore have to enrich our set of numerical techniques with schemes that are also able to handle more general hyperbolic systems written in non–conservative form

$$\frac{\partial Q}{\partial t} + A(Q)\frac{\partial Q}{\partial x} = 0, \qquad A \in \mathbb{R}^{m\times m}, Q \in \mathbb{R}^m, x \in \mathbb{R}, t \in \mathbb{R}_0^+. \tag{5.1}$$

## 1. Some basic aspects of the theory of non–conservative systems

For classical (strong) solutions where $Q$ is differentiable there are no particular problems with PDE of the kind (5.1), since the derivatives are defined everywhere and the method of characteristics can be applied. The main problem arises with non–conservative PDE of the type (5.1) in the presence of shock waves (discontinuities), where the differential form of the equation does not hold any more. In the conservative case, the solution to the problem was given by the Rankine–Hugoniot conditions, that were derived from an integral form of the conservation law. For non–conservative systems, the theory of shock waves is much more complex and far less developed than for the conservative case. The main contribution to the theory of shock waves for non–conservative PDE is due to an article by Dal Maso, Le Floch and Murat [**58**], and whenever we refer to this theory in the following we will call it the DLM theory. According to the DLM theory, in the case of a non–conservative PDE, the jump relations across a discontinuity depend on the chosen *integration path* that is used to connect the left and the right state at the discontinuity in phase–space. Different integration paths result in different jump relations.

Definition 3. *A path $\Psi$ in phase–space $\Omega \subset \mathbb{R}^m$ is a Lipschitz continuous function of a parameter $0 \le s \le 1$*

$$\Psi : [0;1] \times \Omega \times \Omega \to \Omega, \tag{5.2}$$

*that must satisfy the following properties:*

$$\Psi(0, Q_L, Q_R) = Q_L, \qquad \Psi(1, Q_L, Q_R) = Q_R, \qquad \forall Q_L, Q_R \in \Omega, \tag{5.3}$$

$$|\frac{\partial \Psi}{\partial s}(s, Q_L, Q_R)| \le k|Q_R - Q_L|, \qquad \forall Q_L, Q_R \in \mathcal{B} \subset \Omega, \tag{5.4}$$

$$\begin{aligned} |\frac{\partial \Psi}{\partial s}(s, Q_L, Q_R) - \frac{\partial \Psi}{\partial s}(s, Q_l, Q_r)| \le K(|Q_L - Q_l| + |Q_R - Q_r|), \\ \forall Q_L, Q_R, Q_l, Q_r \in \mathcal{B} \subset \Omega, \end{aligned} \tag{5.5}$$

*with $\mathcal{B} \subset \Omega$ being an arbitrary bounded subset of the phase–space $\Omega$.*

Example 32. *The simplest possible integration path is obviously the straight line segment, i.e.*

$$\Psi = \Psi(s, Q_L, Q_R) = Q_L + s(Q_R - Q_L), \qquad 0 \leq s \leq 1. \tag{5.6}$$

Exercise 6. *Prove that the segment path* (5.6) *satisfies all the three properties given in* (5.3)-(5.5)*.*

According to the DLM theory [**58**] weak solutions of (5.1) can be interpreted as a Borel measure and the classical Rankine–Hugoniot jump–relations are generalized to

$$\sigma(Q_R - Q_L) = \int\limits_0^1 A(\Psi(s, Q_L, Q_R))\frac{\partial \Psi}{\partial s} ds = 0, \tag{5.7}$$

where $\sigma$ is the speed of the discontinuity between the two states $Q_L$ and $Q_R$. According to the DLM theory, a piecewise $\mathcal{C}^1$ function $Q$ is a *weak solution* of (5.1) if and only if it is a *classical* solution where it is $\mathcal{C}^1$ and if it satisfies (5.7) along the discontinuities.

It is easy to prove that in the case when $A(Q)$ is the Jacobian of a flux $f(Q)$, i.e. $A(Q) = \partial f(Q)/\partial Q$, the extended Rankine–Hugoniot conditions reduce to the classical ones, *independent* of the particular choice of the path:

$$\sigma(Q_R - Q_L) = \int\limits_0^1 A(\Psi)\frac{\partial \Psi}{\partial s} ds = \int\limits_0^1 \frac{\partial f(\Psi)}{\partial \Psi}\frac{\partial \Psi}{\partial s} ds = \int\limits_{f(Q_L)}^{f(Q_R)} df = f(Q_R) - f(Q_R). \tag{5.8}$$

## 2. Path–conservative finite volume schemes

The first special case of a path–conservative scheme has been proposed by Toumi in [**86**], where he proposed a weak formulation of the method of Roe based on the integration along a particular path. The framework of path–conservative schemes has then been successively extended and analyzed by Parés [**63**] and Castro et al. [**9, 10**] in the finite volume context and by Rhebergen et al. [**66**] in the discontinuous Galerkin finite element context. The first better than second order accurate path–conservative schemes on general unstructured meshes in two and three space dimensions have been published in [**21, 23, 28**].

A general path–conservative finite volume scheme of first order of accuracy is written in the following way, which is similar to the wave–propagation formulation of LeVeque [**55**]:

$$Q_i^{n+1} = Q_i^n - \frac{\Delta t}{\Delta x}\left(D^-_{i+\frac{1}{2}} + D^+_{i-\frac{1}{2}}\right), \tag{5.9}$$

with the jump terms on the boundaries

$$D^-_{i+\frac{1}{2}} = C^-_{i+\frac{1}{2}}(Q^n_{i+1} - Q^n_i), \qquad \text{and} \qquad D^+_{i-\frac{1}{2}} = C^+_{i-\frac{1}{2}}(Q^n_i - Q^n_{i-1}). \tag{5.10}$$

The actual choice of the numerical scheme lies in the particular choice of the matrices $C^-_{i+\frac{1}{2}} = C^-_{i+\frac{1}{2}}(Q^n_i, Q^n_{i+1})$ and $C^+_{i-\frac{1}{2}} = C^+_{i-\frac{1}{2}}(Q^n_{i-1}, Q^n_i)$, which are both functions of the left and the right state at the cell interface. For any path–conservative scheme, the matrices $C^\pm_{i+\frac{1}{2}}$ must have the following property at the interface:

$$C^-_{i+\frac{1}{2}} + C^+_{i+\frac{1}{2}} = \tilde{A}_{i+\frac{1}{2}}, \tag{5.11}$$

where $\tilde{A}_{i+\frac{1}{2}} = \tilde{A}(Q^n_i, Q^n_{i+1})$ is the generalized Roe matrix with the following three features:

First, $\tilde{A}_{i+\frac{1}{2}}$ has to satisfy the generalized Rankine–Hugoniot conditions:

$$\tilde{A}(Q_L, Q_R)(Q_R - Q_L) = \int_0^1 A(\Psi(s, Q_L, Q_R))\frac{\partial \Psi}{\partial s} ds. \tag{5.12}$$

From (5.12) it is obvious that for a conservative system, where $A(Q)$ is the Jacobian of a flux $f(Q)$, the generalized Roe property (5.12) reduces to the classical one

$$\tilde{A}(Q_L, Q_R)(Q_R - Q_L) = f(Q_R) - f(Q_L). \tag{5.13}$$

Second, the generalized Roe matrix $\tilde{A}(Q_L, Q_R)$ must be hyperbolic, i.e. it must have real eigenvalues and a full set of linearly independent eigenvectors. Third, the generalized Roe matrix must satisfy the consistency condition

$$\tilde{A}(Q, Q) = A(Q). \tag{5.14}$$

With these properties, it is then easy to prove the following

Theorem 4. *A path–conservative finite volume scheme of the form* (5.9)-(5.12) *is exactly conservative in the case where the system matrix $A(Q)$ is the Jacobian of a flux $f(Q)$, i.e. when*

$$A(Q) = \frac{\partial f(Q)}{\partial Q}. \tag{5.15}$$

Proof. Assuming no contributions from the boundaries for simplicity, the conserved quantity $Q$ at the new time $t^{n+1}$ is

(5.16)

$$\begin{aligned}\sum_i \Delta x Q_i^{n+1} = \sum_i \Delta x Q_i^n - \Delta t \sum_i \left( C^-_{i+\frac{1}{2}}(Q^n_{i+1} - Q^n_i) + C^+_{i-\frac{1}{2}}(Q^n_i - Q^n_{i-1}) \right) = \\ \sum_i \Delta x Q_i^n - \Delta t \sum_i \left( C^-_{i-\frac{1}{2}} + C^+_{i-\frac{1}{2}} \right)(Q^n_i - Q^n_{i-1}) = \\ \sum_i \Delta x Q_i^n - \Delta t \sum_i \tilde{A}_{i-\frac{1}{2}}(Q^n_i - Q^n_{i-1}) = \\ \sum_i \Delta x Q_i^n - \Delta t \sum_i \left( f(Q^n_i) - f(Q^n_{i-1}) \right) = \sum_i \Delta x Q_i^n,\end{aligned}$$

since the generalized Rankine–Hugoniot conditions reduce to the classical ones (5.13) in this case and since the sum vanishes due to its telescopic property. □

A particularly interesting general family of path–conservative schemes can be written in the following form:

$$D^{\pm}_{i+\frac{1}{2}} = \frac{1}{2}\left( \tilde{A}_{i+\frac{1}{2}} \pm \Theta_{i+\frac{1}{2}} \right)\left( Q^n_{i+1} - Q^n_i \right), \tag{5.17}$$

where $\tilde{A}_{i+\frac{1}{2}}$ is the generalized Roe matrix and $\Theta_{i+\frac{1}{2}} > 0$ is the dissipation matrix of the scheme.

Exercise 7. *Prove that the family of path–conservative schemes given by expression* (5.17) *can be written in the classical flux–conservative form with numerical fluxes*

$$f_{i+\frac{1}{2}} = D^-_{i+\frac{1}{2}} + f(Q^n_i), \qquad \text{and} \qquad f_{i-\frac{1}{2}} = f(Q^n_i) - D^+_{i-\frac{1}{2}}, \tag{5.18}$$

*for the conservative case $A(Q) = \partial f(Q)/\partial Q$.*

With different choices of $\Theta_{i+\frac{1}{2}}$ we can reproduce several classical schemes:

- The Lax–Friedrichs method.

$$\Theta^{LF}_{i+\frac{1}{2}} = \frac{\Delta x}{\Delta t}\mathbf{I}. \tag{5.19}$$

- A Lax–Wendroff–type method.

$$\Theta^{LW}_{i+\frac{1}{2}} = \frac{\Delta t}{\Delta x}\tilde{A}^2_{i+\frac{1}{2}}. \tag{5.20}$$

- The FORCE scheme [**80, 81**] and its extensions to non–conservative hyperbolic systems [**23, 8, 7**].

$$\Theta^{FO}_{i+\frac{1}{2}} = \frac{1}{2}\left(\Theta^{LF}_{i+\frac{1}{2}} + \Theta^{LW}_{i+\frac{1}{2}}\right) = \frac{1}{2}\left(\frac{\Delta x}{\Delta t}\mathbf{I} + \frac{\Delta t}{\Delta x}\tilde{A}^2_{i+\frac{1}{2}}\right). \tag{5.21}$$

- The method of Rusanov [**70**].

$$\Theta_{i+\frac{1}{2}} = \max(|\Lambda(Q^n_i)|, |\Lambda(Q^n_{i+1})|)\mathbf{I}. \tag{5.22}$$

- The method of Roe [**69**].

$$\Theta_{i+\frac{1}{2}} = \left|\tilde{A}_{i+\frac{1}{2}}\right|. \tag{5.23}$$

For the segment path (5.6), this is identical with

$$\Theta_{i+\frac{1}{2}} = \left|\tilde{A}_{i+\frac{1}{2}}\right| = \left|\int_0^1 A(\Psi(Q^n_i, Q^n_{i+1}, s))ds\right|. \tag{5.24}$$

The proof is left to the reader as exercise.

- It has been shown in [**28**] that for the case of a segment path (5.6) the method of Osher and Solomon [**62**] takes the form

$$\Theta_{i+\frac{1}{2}} = \int_0^1 \left|A(\Psi(Q^n_i, Q^n_{i+1}, s))\right| ds, \tag{5.25}$$

which is very similar to the method of Roe, apart that the integration and the matrix absolute value operator are exchanged.

In the expressions above, **I** denotes the identity matrix.

## 3. Application to the shallow water equations with variable bottom

The shallow water equations with spatially variable but temporally fixed bottom topography read as

$$\begin{aligned} \frac{\partial h}{\partial t} + \frac{\partial hu}{\partial x} &= 0, \\ \frac{\partial hu}{\partial t} + \frac{\partial}{\partial x}\left(hu^2 + \frac{1}{2}gh^2\right) &= -gh\frac{\partial b}{\partial x}, \\ \frac{\partial b}{\partial t} &= 0. \end{aligned} \tag{5.26}$$

The right hand side due to the spatially variable bottom can not be cast into conservation form, but the system (5.27) can be written easily under the general non–conservative form (5.1) with

$$A(Q) = \begin{pmatrix} 0 & 1 & 0 \\ c^2 - u^2 & 2u & c^2 \\ 0 & 0 & 0 \end{pmatrix}, \tag{5.27}$$

with the wave speed $c^2 = gh$.

**3.1. Eigenstructure.** The eigenvalue matrix is $\Lambda = \mathrm{diag}(u-c, 0, u+c)$ and the corresponding right eigenvector matrix and its inverse are

$$R = \begin{pmatrix} 1 & 1 & 1 \\ u-c & 0 & u+c \\ 0 & \frac{u^2}{c^2}-1 & 0 \end{pmatrix}, \qquad R^{-1} = \frac{1}{2}\begin{pmatrix} \frac{c+u}{c} & -\frac{1}{c} & \frac{c}{c-u} \\ 0 & 0 & \frac{c^2}{u^2-c^2} \\ \frac{c-u}{c} & \frac{1}{c} & \frac{c}{c+u} \end{pmatrix}. \tag{5.28}$$

**3.2. Well–balancing.** A very simple but physically very important steady state solution of (5.27) is the so–called still water or lake–at–rest solution, [**54**]:

$$h + b = const. \qquad u = 0. \tag{5.29}$$

It simply states that if the free surface $\eta = h + b$ is constant and the velocity is zero, the solution must remain constant in time. Although this seems quite obvious from a physical point of view, a lot of research has been dedicated to this topic in the last decades in order to design numerical methods which satisfy the so-called C–property, see [**38, 54, 91, 88, 59, 60, 34, 3, 9, 63**] for more details. Castro et al. [**9**] have proven that path–conservative Roe–type schemes are *exactly* well–balanced for system (5.27) if the straight–line segment path is used. Also Osher–type path–conservative schemes [**28**] are exactly well–balanced for (5.27), and the other schemes of the family (5.17) can be made exactly well–balanced when using segment integration paths and an appropriate modification of the matrix $\Theta_{i+\frac{1}{2}}$, see [**8**].

Exercise 8. *Write a MATLAB program that implements the family of path–conservative methods* (5.17) *for the shallow water equations with variable bottom* (5.27)*. For the Osher and the Roe method, use a* numerical integration *of the integrals along the segment path. Try different quadrature formulae, such as the trapezoidal rule, the Simpson rule or a three–point Gaussian quadrature rule with quadrature points $s_1 = 1/2 - \sqrt{15}/10$, $s_2 = 1/2$, $s_3 = 1/2 + \sqrt{15}/10$ and associated weights $\omega_1 = 5/18$, $\omega_2 = 8/18$, $\omega_3 = 5/18$. Solve the Riemann problem*

$$Q(x,0) = \begin{cases} Q_L, & \text{if} \quad x < 0, \\ Q_R, & \text{if} \quad x > 0, \end{cases} \tag{5.30}$$

*with the following initial conditions:*

- $Q_L = (2,0,0)^T$, *and* $Q_R = (1,0,1)^T$,
- $Q_L = (1,0,0)^T$, *and* $Q_R = (1e-7,0,0)^T$,
- $Q_L = (1.46184,0,0)^T$, *and* $Q_R = (0.30873,0,0.2)^T$.

CHAPTER 6

# Higher order WENO schemes

In the first part of this manuscript we have studied (and proven) the theorem of Godunov, which is a fundamental theorem for the construction of numerical methods for hyperbolic PDE. According to the Godunov theorem, it is not possible to devise any linear numerical scheme that is better than first order accurate and monotone. The only way to circumvent the theorem is the design of *nonlinear* scheme, since in this case the hypothesis of the theorem do not apply. The design of second order accurate and monotone schemes was possible by the use of *slope limiters*, which are nonlinear as they depend on the actual numerical solution of the problem. In this chapter, we study one possibility to construct even higher order accurate numerical methods, which can be achieved using nonlinear ENO [**40**] or WENO [**48**] interpolation.

## 1. Pointwise WENO reconstruction

In a finite volume scheme, we need to compute fluxes across the element interfaces. For this purpose, numerical flux functions are used, as presented in the first part of the manuscript. These numerical flux functions need two point values of the numerical solution at the element interface, one extrapolated to the interface from the left and another value extrapolated to the interface from the right. As a consequence, the WENO method of Jiang and Shu [**48**] produces a high order accurate point–wise reconstruction of the solution at the element interfaces $x_{i\pm\frac{1}{2}}$. The general ideas of the WENO scheme [**48**] are the following.

In order to obtain a $k$–th order accurate WENO method, called WENO $k$ in the following, we need piecewise reconstruction polynomials of degree $M = k-1$ for each cell $T_i = [x_{i-\frac{1}{2}}; x_{i+\frac{1}{2}}]$. To calculate the unknown coefficients of the reconstruction polynomials from the known cell averages $Q_j^n$ one needs a *reconstruction stencil* or *stencil*

$$\mathcal{S}_i^M = \bigcup_{j=i-e}^{i+e} T_j, \tag{6.1}$$

composed of $k = 2e + 1$ elements, where $e$ is the extension of the stencil to the left and the right. The reconstruction stencil must always include the element $T_i$ itself. The resulting reconstruction polynomial has $k$ coefficients and is of degree $M = k - 1$. According to the relative position of the stencil elements with respect to the cell $T_i$ for which we want to do the reconstruction, a stencil is called *centered*, *left–sided* or *right–sided*. The reconstruction polynomial $P_i^M(x, t^n)$ of degree $M$ is obtained from the known cell averages $Q_j^n$ by requiring *integral conservation*, i.e. we require that

$$\frac{1}{\Delta x} \int\limits_{T_j} P_i^M(x, t^n) dx = Q_j^n, \qquad \forall T_j \in \mathcal{S}_i^M. \tag{6.2}$$

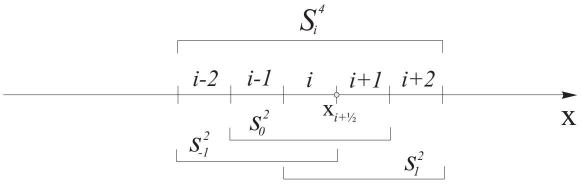

FIGURE 1. Sketch of the WENO5 reconstruction procedure for interface $x_{i+\frac{1}{2}}$. The big stencil $S_4$ needed for the reconstruction of a fourth degree polynomial is divided into three smaller sub–stencils, on each of which a piecewise second degree polynomial is reconstructed. The point value $Q^n_{i+\frac{1}{2}}$ at the interface $x_{i+\frac{1}{2}}$ is then given by a suitable non–linear combination of the reconstruction polynomials obtained on the sub–stencils.

For the WENO method of order $k$ in one space dimension, we need *one big central* reconstruction stencil $S_i^M$ of $k = M + 1$ elements and $M/2 + 1$ small sub–stencils $s_m^{M/2}$ composed of $M/2 + 1$ elements to reconstruct several lower order polynomials of degree $M/2$. Here $m$ is the stencil–shift with respect to the central element $T_i$. The *linear* WENO reconstruction at the element interface $x_{i+\frac{1}{2}}$ is then given as a *linear combination* of the lower order reconstruction polynomials obtained from the substencils $s_r^{M/2}$ using the same integral conservation principle of (6.2), where the linear weights $\lambda$ are chosen in such a way that the linear combination of the lower order polynomials is *identical* with the one obtained via the reconstruction polynomial on the big stencil $S_i^M$. The weights $\lambda$ obviously depend on the position $x$ for which the reconstruction is to be done and for consistency reasons the weights must always sum to one. Furthermore, the weights $\lambda$ should be positive and must not depend on the solution $Q_j^n$.

EXAMPLE 33. *The third order accurate WENO3 method uses one big central stencil composed of the three elements $S_i^2 = \{T_{i-1}, T_i, T_{i+1}\}$ and two one–sided substencils $s_{-1}^1 = \{T_{i-1}, T_i\}$ and $s_1^1 = \{T_i, T_{i+1}\}$. Using the integral conservation principle, we obtain the second order reconstruction polynomial on the big stencil $S_i^2$ as*

(6.3)
$$P_i^2(x) = \frac{1}{2}\,Q^n_{i-1} + \frac{5}{6}\,Q^n_i - \frac{1}{6}\,Q^n_{i+1} + \left(Q^n_i - Q^n_{i-1}\right)\xi + \left(\frac{1}{2}\,Q^n_{i-1} - Q^n_i + \frac{1}{2}\,Q^n_{i+1}\right)\xi^2,$$

*with $x = x_{i-\frac{1}{2}} + \xi\Delta x$. For the two first order polynomials on the sub–stencils we obtain*

$$p_{-1}^1(x) = \frac{1}{2}\,Q^n_i + \frac{1}{2}\,Q^n_{i-1} + \left(Q^n_i - Q^n_{i-1}\right)\xi, \tag{6.4}$$

*and*

$$p_1^1(x) = \frac{3}{2}\,Q^n_i - \frac{1}{2}\,Q^n_{i+1} + \left(Q^n_{i+1} - Q^n_i\right)\xi. \tag{6.5}$$

*The two conditions for obtaining the linear weights $\lambda_{-1}$ and $\lambda_1$ are then*

$$\lambda_{-1} + \lambda_1 = 1, \tag{6.6}$$

*and*

$$\lambda_{-1}p_{-1}^1(x_{i+\frac{1}{2}}) + \lambda_1 p_1^1(x_{i+\frac{1}{2}}) = P_i^2(x_{i+\frac{1}{2}}). \tag{6.7}$$

*The resulting linear weights are $\lambda_{-1} = 1/3$ and $\lambda_{+1} = 2/3$.*

EXERCISE 9. *Compute the linear weights $\lambda$ for the WENO5 method when performing the reconstruction for the point $x_{i+\frac{1}{2}}$. This exercise can be either solved by hand or more conveniently at the aid of a computer algebra system such as MAPLE.*

In order to make the WENO scheme *nonlinear*, i.e. data–dependent, the reconstruction at point $x_{i+\frac{1}{2}}$ is obtained by using a *nonlinear* combination of the lower order reconstruction polynomials of the sub–stencils by substituting the linear weights with *nonlinear weights* $\omega$, which are defined as

$$\omega_s = \frac{\tilde{\omega}_s}{\sum\limits_s \tilde{\omega}_s}, \qquad \tilde{\omega}_s = \frac{\lambda_s}{(\sigma_s + \epsilon)^r}. \tag{6.8}$$

Here, $\sigma_s$ denotes the so–called smoothness or oscillation indicator defined later, $\epsilon$ is a small number to avoid division by zero and $r$ is an exponent for which Jiang and Shu always choose $r = 2$. For the smoothness indicator Jiang and Shu propose

$$\sigma_s = \sum_{l=1}^{M/2} \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} h^{2l-1} \left( \frac{\partial^l}{\partial x^l} p_s^{M/2}(x) \right)^2 dx, \tag{6.9}$$

where the term $h^{2m-1}$ is used to remove scaling effects from the derivatives.

## 2. Polynomial WENO reconstruction

The original pointwise WENO reconstruction of Jiang and Shu [**48**] described for 1D in the previous section is rather difficult to generalize to unstructured triangular and tetrahedral meshes in two and three space dimensions because of the need to determine the optimal linear weights, see [**46, 72, 93**]. Therefore, we present a different idea in this section which can be extended very easily to the unstructured case. The alternative reconstruction procedure described here for the one-dimensional case follows directly from the general guidelines given in [**24, 25**] for unstructured triangular and tetrahedral meshes in two and three space dimensions. The polynomial WENO reconstruction operator of this section produces *entire polynomials* $w_i(x)$, as the ENO approach proposed by Harten *et al.* in [**40**]. However, we formally write our method like a WENO scheme [**48, 57**] with a particularly simple choice for the linear weights. The most important difference of our approach compared to the classical WENO schemes of Jiang and Shu [**48**] described in the next section is that standard WENO methods reconstruct *point values* at the Gaussian integration points instead of an entire polynomial valid inside each control volume $T_i = [x_{i-\frac{1}{2}}; x_{i+\frac{1}{2}}]$.
Reconstruction is again done for each element on a reconstruction stencil $\mathcal{S}_i^s$, which is given by the following union of the element $T_i$ and its neighbors $T_j$,

$$\mathcal{S}_i^s = \bigcup_{j=i+s-e}^{i+s+e} T_j, \tag{6.10}$$

where $s$ is the stencil shift with respect to the central element $T_i$ and $e$ is the spatial extension of the stencil to the left and the right. A central reconstruction stencil is given by $s = 0$, an entirely left-sided stencil is given by $s = -e$ and an entirely right-sided stencil is given by $s = e$. In our approach, we always will use *three fixed* reconstruction stencils $\mathcal{S}_i^0$, $\mathcal{S}_i^{-e}$ and $\mathcal{S}_i^e$.
Given the cell average data $Q_i^n$ in all elements $Q_i$ we are looking for a spatial

reconstruction polynomial obtained from $\mathcal{S}_i^s$ at time $t^n$ of the form

$$w_i^s(\xi, t^n) = \sum_{l=0}^{M} \phi_l(\xi)\hat{w}_l^s(t^n) := \phi_l(\xi)\hat{w}_l^s(t^n), \tag{6.11}$$

with the spatial reconstruction basis functions $\phi_l(\xi)$. The $\phi_l(\xi)$ span the space of piecewise polynomials of degree $M$, and as basis functions one can use for example monomials, or, better, rescaled Legendre polynomials on the unit interval, which form an orthogonal basis. For each element $T_i$ we use a reference coordinate $0 \leq \xi \leq 1$ given by $x = x_{i-\frac{1}{2}} + \xi\Delta x$. In the following, we will use standard tensor index notation, implying summation over indices appearing twice. The number of polynomial coefficients (degrees of freedom) is $k = M+1$, where $M$ is the degree of the reconstruction polynomial and $k$ is the spatial order of accuracy of the scheme in space. To compute the reconstruction polynomial $w_i(\xi, t^n)$ valid for element $T_i$ we require again *integral conservation* for all elements $T_j$ inside the stencil $\mathcal{S}_i^s$, i.e.

$$\int\limits_{T_j} w_i^s(\xi, t^n) d\xi = \int\limits_{T_j} \phi_l(\xi) d\xi \cdot \hat{w}_l^s(t^n) = Q_j^n, \qquad \forall T_j \in \mathcal{S}_i^s. \tag{6.12}$$

Equation (6.12) yields a linear equation system for the unknown coefficients $\hat{w}_l^s(t^n)$ of the reconstruction polynomial on stencil $\mathcal{S}_i^s$ that can be easily solved.

To obtain the final non-oscillatory reconstruction polynomials for each $T_i$ at time $t^n$, we finally construct a data-dependent nonlinear combination of the polynomials $w_i^0(\xi, t^n)$, $w_i^{-k}(\xi, t^n)$ and $w_i^k(\xi, t^n)$ obtained from the central, left-sided and right-sided stencils as follows:

$$w_i(\xi, t^n) = \hat{w}_l(t^n)\phi_l(\xi), \tag{6.13}$$

with

$$\hat{w}_l(t^n) = \omega_0\, \hat{w}_l^0(t^n) + \omega_{-k}\, \hat{w}_l^{-k}(t^n) + \omega_k\, \hat{w}_l^k(t^n). \tag{6.14}$$

The nonlinear weights $\omega_s$ are given by the relations

$$\omega_s = \frac{\tilde{\omega}_s}{\tilde{\omega}_0 + \tilde{\omega}_{-k} + \tilde{\omega}_k}, \quad \tilde{\omega}_s = \frac{\lambda_s}{(\sigma_s + \epsilon)^r}. \tag{6.15}$$

The oscillation indicators $\sigma_s$ are computed as for pointwise WENO reconstructions:

$$\sigma_s = \sum_{l=1}^{M} \int\limits_0^1 \left( \frac{\partial^l}{\partial \xi^l} w_i^s(\xi, t^n) \right)^2 d\xi. \tag{6.16}$$

The parameters $\epsilon$ and $r$ are constants for which one typically chooses $\epsilon = 10^{-14}$ and $r = 8$. For the linear weights $\lambda_s$ one chooses $\lambda_{-k} = \lambda_k = 1$ and a very large linear weight $\lambda_0$ on the central stencil, typically $\lambda_0 = 10^5$. It has been shown previously [**48, 57**] that the numerical results are quite insensitive to the WENO parameters $\epsilon$ and $r$ and also with respect to the linear weight on the central stencil $\lambda_0$, see [**24**]. The proposed reconstruction usually uses the accurate and linearly stable central stencil reconstruction in those regions of $\Omega$ where the solution is smooth because of the large linear weight $\lambda_0$. However, due to the strongly nonlinear dependence of the weights $\omega_s$ on the oscillation indicators $\sigma_s$, in the presence of discontinuities the smoother left- or right-sided stencils are preferred, as for standard ENO and WENO methods. For the nonlinear scalar case, the reconstruction operator described above can be directly applied to the cell averages $Q_i^n$ of the conserved quantity $u$. For nonlinear hyperbolic systems, the reconstruction should be done in *characteristic variables* [**40**] in order to avoid spurious oscillations that may appear when applying ENO or WENO reconstruction operators component-wise to nonlinear hyperbolic systems. The result of the reconstruction procedure is a non-oscillatory spatial polynomial $w_i(\xi, t^n)$ defined at time $t^n$ inside each spatial element $T_i$.

The advantage of the polynomial WENO reconstruction is its straight–forward extension to general unstructured meshes. The drawback is that at a given order of accuracy $k$ the total stencil needed for the reconstruction is *bigger* than the one of the classical pointwise WENO scheme presented before.

Example 34. *For the third order polynomial WENO reconstruction described in this section, we use the rescaled Legendre polynomials up to degree two as reconstruction basis functions:*

$$\phi_0(\xi) = 1, \qquad \phi_1(\xi) = 2\xi - 1, \qquad \phi_2(\xi) = 1 - 6\xi + 6\xi^2. \tag{6.17}$$

*It can be easily shown that these functions are orthogonal on the unit interval. We obtain the following expansion coefficients for the left–sided stencil*

$$\hat{w}_0^{-1} = Q_i^n, \quad \hat{w}_1^{-1} = \frac{1}{4}Q_{i-2}^n - Q_{i-1}^n + \frac{3}{4}Q_i^n, \quad \hat{w}_2^{-1} = \frac{1}{12}Q_{i-2}^n - \frac{1}{6}Q_{i-1}^n + \frac{1}{12}Q_i^n, \tag{6.18}$$

*for the central stencil*

$$\hat{w}_0^0 = Q_i^n, \quad \hat{w}_1^0 = -\frac{1}{4}Q_{i-1}^n + \frac{1}{4}Q_{i+1}^n, \quad \hat{w}_2^0 = \frac{1}{12}Q_{i-1}^n - \frac{1}{6}Q_i^n + \frac{1}{12}Q_{i+1}^n, \tag{6.19}$$

*and for the right–sided stencil*

$$\hat{w}_0^{+1} = Q_i^n, \quad \hat{w}_1^{+1} = -\frac{3}{4}Q_i^n + Q_{i+1}^n - \frac{1}{4}Q_{i+2}^n, \quad \hat{w}_2^{+1} = \frac{1}{12}Q_i^n - \frac{1}{6}Q_{i+1}^n + \frac{1}{12}Q_{i+2}^n. \tag{6.20}$$

*The oscillation indicator is given by*

$$\sigma_s = 156 \left(\hat{w}_2^s\right)^2 + 4 \left(\hat{w}_1^s\right)^2. \tag{6.21}$$

CHAPTER 7

# The Discontinuous Galerkin finite element method

In high order finite volume schemes, only the cell averages $Q_i^n$ are stored for each element and are evolved in time. To achieve higher order of accuracy in space the high order polynomials must be obtained from the cell–averages via a reconstruction procedure or recovery procedure, see the previous chapter. In the discontinuous Galerkin finite element framework, the coefficients of higher order polynomials are *directly* evolved in time for each cell, *without* the need of using a reconstruction operator. This feature of DG schemes is in common with the classical finite element method (FEM). However, unlike in classical finite elements, the numerical solution given by a DG scheme is *discontinuous at element interfaces* and this discontinuity is resolved by the use of a *numerical flux function*, which is a common feature with shock–capturing finite volume schemes.

The DG method was invented by Reed and Hill [**65**] for the solution of neutron transport equations and was put on a solid mathematical basis by Cockburn and Shu in a well–known series of papers [**14, 13, 12, 11, 16**]. The literature on DG finite element schemes is rapidly growing at the moment since the method has several advantages with respect to classical finite volume schemes: The DG method reaches arbitrary order of accuracy in space on general unstructured meshes and allows easily for *hp*–adaptation, i.e. it allows for refinement and recoarsening of the mesh and for a dynamical adaptation of the polynomial degree of the numerical solution. For the DG method, Jiang and Shu have found a very elegant and general proof of *nonlinear stability* in $L_2$–norm, see [**47**]. In the following we give a brief sketch of the method for the one–dimensional scalar case and refer to the literature for details, see for example the nice review paper by Cockburn and Shu [**17**].

## 1. General algorithm

The governing PDE

$$q_t + f(q)_x = 0 \tag{7.1}$$

is multiplied by a test function $\phi = \phi(x)$ from the space of piecewise polynomials of degree $N$ and is integrated over a spatial control volume $T_i = [x_{i-\frac{1}{2}}; x_{i+\frac{1}{2}}]$. Integration by parts yields

$$\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi q_t \, dx + \phi^-_{i+\frac{1}{2}} f_{i+\frac{1}{2}} - \phi^+_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi_x f(q) dx = 0, \tag{7.2}$$

where $f_{i+\frac{1}{2}} = f_{i+\frac{1}{2}}(q^-_{i+\frac{1}{2}}, q^+_{i+\frac{1}{2}})$ is a classical numerical flux function evaluated at the boundary–extrapolated values of the solution $q^\pm_{i+\frac{1}{2}}$ at the left and the right side of the element interface $x_{i+\frac{1}{2}}$, as used in the finite volume context. Likewise, $\phi^\pm_{i+\frac{1}{2}}$ is the boundary–extrapolated value of the test function. For the numerical solution

$q_h = q(x,t)$ we make the following ansatz:

$$q_h(x,t) = \sum_{l=0}^{N} \hat{u}_l(t)\phi_l(x). \tag{7.3}$$

Inserting Eqn. (7.3) into (7.2) and using the same test functions $\phi$ as the ones used for the ansatz (7.3), we obtain for each element $T_i$ a system of $N+1$ ordinary differential equations for the unknown coefficients $\hat{u}_l(t)$ of the numerical solution $q_h$. This ODE system of the so–called *semi–discrete* form of the DG scheme is then typically integrated using a TVD Runge–Kutta scheme in time, see [**16, 17**]. However, also other time–discretizations are possible, such as ADER or Lax–Wendroff–type one–step time discretizations [**64, 26**]. It is very important to underline that explicit DG finite element schemes require a *very severe* CFL–type restriction of the time step. An estimate for the maximum Courant number of common RK–DG schemes is

$$\text{CFL} < \frac{1}{2N+1}. \tag{7.4}$$

For an extension of the DG method to diffusion equations and PDE with higher order derivatives, which is not as straightforward as it seems, see [**4, 15, 92, 56, 42, 35, 1**].

## 2. The TVB limiter of Cockburn and Shu

A still open topic of research is the construction of appropriate *limiters* for high order DG finite element schemes. In [**13, 12**] Cockburn and Shu proposed a TVB (totally variation bounded) limiter, which depends on a constant $K > 0$, to preserve local extrema. They introduced a *modified minmod function*

$$\mathrm{m}(a,b,c) = \begin{cases} a & \text{if } |a| < K\Delta x^2 \\ \mathrm{minmod}(a,b,c), & \text{else,} \end{cases} \tag{7.5}$$

with the classical minmod function being defined as

$$\mathrm{minmod}(a,b,c) = \begin{cases} \min(a,b,c) & \text{if } \mathrm{sign}(a) = \mathrm{sign}(b) = \mathrm{sign}(c), \\ 0 & \text{else.} \end{cases} \tag{7.6}$$

The TVB limiter of Cockburn and Shu is then applied as follows: Compute the $L_2$–projection of the numerical solution $q_h$ into the space of piecewise polynomials of degree one. The result of this projection is called $v_h$. For hierarchical orthogonal basis functions such as the Legendre polynomials in 1D or the orthogonal basis functions on general elements in multiple dimensions defined in [**49, 19**], this projection is simply done by setting the higher order coefficients to zero. Then, compute

$$u^-_{i+\frac{1}{2}} = \bar{v}_i + m(v^-_{i+\frac{1}{2}} - \bar{v}_i, \bar{v}_i - \bar{v}_{i-1}, \bar{v}_{i+1} - \bar{v}_i), \tag{7.7}$$

and

$$u^+_{i-\frac{1}{2}} = \bar{v}_i - m(\bar{v}_i - v^+_{i-\frac{1}{2}}, \bar{v}_i - \bar{v}_{i-1}, \bar{v}_{i+1} - \bar{v}_i), \tag{7.8}$$

where the bar denotes the cell average value. In the case of hierarchical orthogonal basis functions, such as the Legendre polynomials, the cell average is directly given as the first expansion coefficient, associated with the piecewise constant term. If $u^-_{i+\frac{1}{2}} = v^-_{i+\frac{1}{2}}$ and $u^+_{i-\frac{1}{2}} = v^+_{i-\frac{1}{2}}$, then no limitation is applied to the numerical solution. Otherwise, set in each element

$$q_h(x) = \bar{v}_i + (x - x_i) \cdot m(v_x, \frac{\bar{v}_i - \bar{v}_{i-1}}{\Delta x/2}, \frac{\bar{v}_{i+1} - \bar{v}_i}{\Delta x/2}). \tag{7.9}$$

## 3. Nonlinear $L_2$–stability

As we have seen in the first part of this manuscript, high order linear Finite Difference schemes produce spurious oscillations near discontinuities. The reason can be traced back to Godunov's theorem [**37**], which states that there is no monotone linear difference scheme of order greater than one. Linear means in this case that the coefficients of the scheme are independent of the numerical solution.

Obviously, also the DG method is not able to circumvent the theorem of Godunov, but it has been found in literature that even the linear version of the Discontinuous Galerkin finite element method without using limiters exhibits a peculiar robustness in the presence of discontinuities. This was already mentioned by Atkins and Shu [**2**]. To shed some more light onto this interesting fact, we recall in this section the analysis carried out by Jiang and Shu [**47**] and its essential results. Jiang and Shu were able to prove rigorously a discrete cell entropy inequality for the square entropy for the Discontinuous Galerkin method when applied to scalar nonlinear hyperbolic conservation laws. This is a very remarkable result, since it was derived only under the assumption of the existence of a monotone (entropy–satisfying) numerical flux and the analysis is valid even without the use of limiters. The analysis was extended by Yan and Shu for local Discontinuous Galerkin schemes for PDEs with higher order derivatives [**92**].

**3.1. Generalities and definitions.** We start with some general definitions and lemmata. The nonlinear scalar hyperbolic conservation laws under consideration have the form

$$q_t + f(q)_x = 0. \tag{7.10}$$

Definition 4. *The pair of functions* $(U(q), F(q))$ *is called an entropy pair to the hyperbolic conservation law* (7.10)*, if the entropy* $U$ *fulfills*

$$U'' > 0 \tag{7.11}$$

*and the entropy flux* $F$ *satisfies*

$$F'(q) = U'(q)f'(q). \tag{7.12}$$

The superscript $'$ denotes differentiation with respect to $q$.

Lemma 2. *If* $q$ *is a continuously differentiable solution of the conservation law* (7.10) *then any entropy pair* $(U(q), F(q))$ *fulfills the equation*

$$U(q)_t + F(q)_x = 0. \tag{7.13}$$

Proof. If $q$ is a continuously differentiable solution of (7.10), then

$$q_t + f'q_x = 0. \tag{7.14}$$

Multiplying by $U'$

$$U'q_t + U'f'q_x = 0 \tag{7.15}$$

and using (7.12) one obtains

$$U_t + F'q_x = U_t + F_x = 0. \tag{7.16}$$

□

Lemma 3. *If* $q$ *is a weak entropy solution* [**51**] *of the conservation law* (7.10) *then any entropy pair* $(U(q), F(q))$ *fulfills the inequality*

$$U(q)_t + F(q)_x \le 0. \tag{7.17}$$

Proof. The proof can be found in [**51**] □

Definition 5. *We introduce the square entropy*

$$U_2 = \frac{q^2}{2} \tag{7.18}$$

*and its associated entropy flux*

$$F_2 = qf(q) - \int f(q)dq. \tag{7.19}$$

Lemma 4. *The function pair $(U_2, F_2)$ is an entropy pair.*

Proof. The first property of Definition 4 follows directly as

$$U_2''(q) = 1 > 0. \tag{7.20}$$

For the flux $F_2$ we have

$$F_2'(q) = f(q) + qf'(q) - f(q) \tag{7.21}$$

and with $U_2'(q) = q$ follows the second property:

$$F_2'(q) = U_2'(q)f'(q). \tag{7.22}$$

□

Definition 6. *A Lipschitz continuous function $f_{i+\frac{1}{2}} = f_{i+\frac{1}{2}}(q^-, q^+)$ of two states $q^-$ and $q^+$ is called an e-flux* [**61**] *for the conservation law* (7.10)*, if it has the following properties:*

$$f_{i+\frac{1}{2}}(q, q) = f(q). \tag{7.23}$$

$$\int\limits_{q^-}^{q^+} \left(f(q) - f_{i+\frac{1}{2}}(q^-, q^+)\right) dq \geq 0. \tag{7.24}$$

**3.2. Discrete cell entropy inequality.** Approximating the solution of (7.10) by $q \in V_h$ in a discrete function space $V_h \subset L^2$ and multiplying (7.10) by test functions $\Phi \in V_h$ from the same function space and integrating by parts, one gets the following variational formulation

$$\int_{T_i} q_t \Phi \; dx + \Phi^-_{i+\frac{1}{2}} f_{i+\frac{1}{2}} - \Phi^+_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - \int_{T_i} \frac{\partial}{\partial x} \Phi f(q) \; dx. \tag{7.25}$$

In (7.25) the $T_i = [x_{i-\frac{1}{2}}; x_{i+\frac{1}{2}}]$ are the subintervals into which the computational domain is divided, and $f_{i+\frac{1}{2}} = f_{i+\frac{1}{2}}(q^-_{i+\frac{1}{2}}, q^+_{i+\frac{1}{2}})$ is an e-flux between the element $T_i$ and its right neighbor $T_{i+1}$ and $f_{i-\frac{1}{2}} = f_{i+\frac{1}{2}}(q^-_{i-\frac{1}{2}}, q^+_{i-\frac{1}{2}})$ denotes an e-flux between $T_i$ and its left neighbor $T_{i-1}$. $\Phi^-_{i+\frac{1}{2}} = \Phi(x^-_{i+\frac{1}{2}})$ and $\Phi^+_{i-\frac{1}{2}} = \Phi(x^+_{i-\frac{1}{2}})$ denote the values of the test functions inside $T_i$ at the interface with the neighbor elements.

Theorem 5. *The numerical solution $q \in V_h$ of the Discontinuous Galerkin scheme* (7.25) *fulfills the discrete cell–entropy inequality*

$$\int_{T_i} \left(\frac{q^2}{2}\right)_t dx + \hat{F}_{i+\frac{1}{2}} - \hat{F}_{i-\frac{1}{2}} \leq 0. \tag{7.26}$$

*for the square entropy.*

Proof. As $q$ and $\Phi$ are from the same function space, we can also use $q$ as a test function instead of $\Phi$. We get

$$\int_{T_i} \left(\frac{q^2}{2}\right)_t dx + q^-_{i+\frac{1}{2}} f_{i+\frac{1}{2}} - q^+_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - \int_{T_i} q_x f(q) \; dx. \tag{7.27}$$

If we rewrite the last term as

$$\int_{T_i} f(q)q_x \; dx = \int\limits_{q^+_{i-\frac{1}{2}}}^{q^-_{i+\frac{1}{2}}} f(q)dq = g(q^-_{i+\frac{1}{2}}) - g(q^+_{i-\frac{1}{2}}) \tag{7.28}$$

with the definition

$$g(q) = \int f(q)dq, \tag{7.29}$$

eqn. (7.27) takes the form

$$\int_{T_i} \left(\frac{q^2}{2}\right)_t dx + \left(q^-_{i+\frac{1}{2}} f_{i+\frac{1}{2}} - g(q^-_{i+\frac{1}{2}})\right) - \left(q^+_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - g(q^+_{i-\frac{1}{2}})\right) = 0. \tag{7.30}$$

Without changing anything, this can be rewritten as

$$\begin{aligned} \int_{T_i} \left(\frac{q^2}{2}\right)_t dx \quad &+ \quad \left(q^-_{i+\frac{1}{2}} f_{i+\frac{1}{2}} - g(q^-_{i+\frac{1}{2}})\right) - \left(q^-_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - g(q^-_{i-\frac{1}{2}})\right) \\ &- \quad \left(q^+_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - g(q^+_{i-\frac{1}{2}})\right) + \left(q^-_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - g(q^-_{i-\frac{1}{2}})\right) = 0. \end{aligned} \tag{7.31}$$

Eqn. (7.31) finally can be rewritten more compactly as

$$\int_{T_i} \left(\frac{q^2}{2}\right)_t dx + \hat{F}_{i+\frac{1}{2}} - \hat{F}_{i-\frac{1}{2}} + \hat{R}_{i-\frac{1}{2}} = 0, \tag{7.32}$$

with the definitions

$$\hat{F}_{i+\frac{1}{2}} = q^-_{i+\frac{1}{2}} f_{i+\frac{1}{2}} - g(q^-_{i+\frac{1}{2}}), \tag{7.33}$$

$$\hat{F}_{i-\frac{1}{2}} = q^-_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - g(q^-_{i-\frac{1}{2}}), \tag{7.34}$$

$$\hat{R}_{i-\frac{1}{2}} = -\left(q^+_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - g(q^+_{i-\frac{1}{2}})\right) + \left(q^-_{i-\frac{1}{2}} f_{i-\frac{1}{2}} - g(q^-_{i-\frac{1}{2}})\right). \tag{7.35}$$

The terms $\hat{F}_{i+\frac{1}{2}}$ and $\hat{F}_{i-\frac{1}{2}}$ are discrete entropy fluxes that are consistent with the continuous entropy flux of the square entropy (7.19).

The term $\hat{R}_{i-\frac{1}{2}}$ can be written more compactly as

$$\hat{R}_{i-\frac{1}{2}} = \int\limits_{q^-_{i-\frac{1}{2}}}^{q^+_{i-\frac{1}{2}}} \left(f(q) - f_{i+\frac{1}{2}}\left(q^-_{i-\frac{1}{2}}, q^+_{i-\frac{1}{2}}\right)\right) dq. \tag{7.36}$$

Because of property (7.24) of the monotone flux this term is always non-negative,

$$\hat{R}_{i-\frac{1}{2}} \geq 0, \tag{7.37}$$

so from (7.32) and (7.37) finally follows

$$\int_{T_i} \left(\frac{q^2}{2}\right)_t dx + \hat{F}_{i+\frac{1}{2}} - \hat{F}_{i-\frac{1}{2}} \leq 0, \tag{7.38}$$

which concludes the proof. □

COROLLARY 1. *The semi-discrete Discontinuous Galerkin scheme* (7.25) *is $L^2$ stable.*

Proof. Summing up (7.38) over all elements $T_i$ in the domain $\Omega$ and imposing either periodic boundary conditions or zero fluxes at the domain boundary $\partial\Omega$ yields

$$\int_{\Omega} \left(\frac{q^2}{2}\right)_t dx \leq 0. \tag{7.39}$$

□

The remarkable result is that the cell entropy inequality (7.38) and the resulting $L^2$ stability (7.39) follow for *arbitrary* high order semi-discrete DG schemes and for *any* nonlinear scalar hyperbolic conservation law, only under the assumption of an e-flux $f_{i+\frac{1}{2}}$ and supposing that the test and approximation spaces in the variational formulation are the same. The proof of Jiang and Shu is short and elegant and it is not necessary to introduce limiters in order to obtain the cell entropy inequality. The amount of numerical dissipation which is introduced into the numerical solution by the DG scheme is controlled by the term $\hat{R}_{i-\frac{1}{2}}$ which is intrinsically linked to the jump in the variable $q$ at the cell interface. The dissipation becomes the greater the higher the jump.

CHAPTER 8

# One–step time discretization based on a local space–time discontinuous Galerkin scheme

In this last chapter we present a one–step time discretization for high order finite volume and discontinuous Galerkin finite element methods first proposed in [**22**]. Again, we limit us to the scalar one–dimensional case. Here, we also consider nonlinear algebraic source terms, hence the governing PDE reads

$$q_t + f(q)_x = S(q). \tag{8.1}$$

The time discretization presented in the following is to a certain extent similar to the one–step time discretization of the MUSCL scheme of van Leer [**87**] and the one used in the ENO method of Harten et al. [**40**]. In those approaches, the high order polynomials $w(x, t^n)$ are *locally evolved in time* (without coupling to the neighbor elements) using the governing partial differential equation. The result is a space–time polynomial $q(x, t)$ for each element. In the MUSCL method this evolution is done using a first order Taylor series in time, where the first time derivative is obtained by the PDE (8.1) and the spatial derivative of the flux is approximated by a central finite difference using the reconstruction polynomial within the element. For the ENO scheme [**40**] the space–time polynomial is obtained by a higher order Taylor series, where also the higher order time derivatives are replaced by pure space derivatives using the Cauchy–Kovalewski procedure. Another one–step time–discretization is automatically achieved within the ADER schemes of Titarev and Toro [**76, 85, 75, 84, 78**] who used the solution of the *generalized Riemann problem* (GRP) [**31, 6, 32, 5, 83**] at the cell interface as a building block for obtaining high order of accuracy in time. Also the ADER approach is essentially based on the use of the Cauchy–Kovalewski procedure. Unfortunately, this procedure becomes very cumbersome for classical nonlinear hyperbolic systems, see [**25**] for the 3D compressible Euler equations and [**74**] for the 2D magnetohydrodynamics (MHD) equations and is almost impossible to do for the general nonlinear hyperbolic systems case. Furthermore, it is not applicable to the case of stiff source terms. Therefore, we show here a different approach that is based on a *weak* formulation of the governing PDE in space–time which is very general and also applicable to stiff source terms.

Final aim of the approach described in the following is to solve an *element–local* Cauchy problem with initial data $w(x, t^n)$. The initial data stems either from a high order WENO reconstruction, or is directly available as $w(x, t^n) = q_h(x, t^n)$ from the discontinuous Galerkin finite element framework. For the solution of the element–local Cauchy problem, we take the PDE (8.1), multiply it with a space–time test function $\theta_k = \theta_k(x, t)$ and integrate over the space–time control volume $[x_{i-\frac{1}{2}}; x_{i+\frac{1}{2}}] \times [t^n; t^{n+1}]$:

$$\int\limits_{t^n}^{t^{n+1}} \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \theta_k \left(q_t + f(q)_x - S(q)\right) dx\, dt = 0. \tag{8.2}$$

Integration by parts of the time–derivative term yields

$$\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \big(\theta_k(x,t^{n+1})q(x,t^{n+1}) - \theta_k(x,t^n)w(x,t^n)\big)\,dx - \int\limits_{t^n}^{t^{n+1}} \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial}{\partial t}\theta_k q\,dx\,dt + \int\limits_{t^n}^{t^{n+1}} \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \theta_k\left(f_x - S\right)\,dx\,dt = 0. \tag{8.3}$$

Here, we have introduced the initial condition $w(x,t^n)$ at the initial time $t^n$. Introduction of the two operators

$$[f,g]_{T_i}^{t} = \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x,t)g(x,t)dx, \quad \text{and} \quad \langle f,g\rangle_{T_i} = \int\limits_{t^n}^{t^{n+1}} \int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x,t)g(x,t)\,dx\,dt = 0. \tag{8.4}$$

yields the following short–hand notation of Eqn. (8.3):

$$[\theta_k, q]_{T_i}^{t^{n+1}} - \left\langle \frac{\partial}{\partial t}\theta_k, q\right\rangle_{T_i} + \langle \theta_k, f_x - S\rangle_{T_i} = [\theta_k, w]_{T_i}^{t^n}. \tag{8.5}$$

Using the same basis functions $\theta_k$ to expand the solution $q$, the flux $f$ and the source $S$ in (8.5)

$$q_h(x,t) = \sum_l \theta_l(x,t)\hat{q}_l := \theta_l\hat{q}_l, \qquad f_h(x,t) = \theta_l\hat{f}_l, \qquad S_h(x,t) = \theta_l\hat{S}_l \tag{8.6}$$

and insertion into (8.5) leads to

$$[\theta_k,\theta_l]_{T_i}^{t^{n+1}}\,\hat{q}_l - \left\langle \frac{\partial}{\partial t}\theta_k, \theta_l\right\rangle_{T_i}\hat{q}_l + \left\langle \theta_k, \frac{\partial}{\partial x}\theta_l\right\rangle_{T_i}\hat{f}_l - \langle\theta_k,\theta_l\rangle_{T_i}\hat{S}_l = [\theta_k,\phi]_{T_i}^{t^n}\,\hat{w}_l, \tag{8.7}$$

or in even more compact matrix notation

$$K_{kl}^1\hat{q}_l + K_{kl}^{\xi}\hat{f}_l - M_{kl}\hat{S}_l = F_{kl}^0\hat{w}_l, \tag{8.8}$$

with the matrix $K_{kl}^1 = F_{kl}^1 - K_{kl}^{\tau}$, the spatial stiffness matrix $K_{kl}^{\xi}$, the mass matrix $M_{kl}$ and the temporal flux matrix at time $t^n$ $F_{kl}^0$ being defined as

$$K_{kl}^1 = [\theta_k,\theta_l]_{T_i}^{t^{n+1}} - \left\langle \frac{\partial}{\partial t}\theta_k, \theta_l\right\rangle_{T_i}, \quad K_{kl}^{\xi} = \left\langle \theta_k, \frac{\partial}{\partial x}\theta_l\right\rangle_{T_i}, \quad M_{kl} = \langle\theta_k,\theta_l\rangle_{T_i}, \quad F_{kl}^0 = [\theta_k,\phi]_{T_i}^{t^n}. \tag{8.9}$$

If we choose a *nodal basis* $\theta_k$, then the degrees of freedom for the flux and the source can be simply evaluated as

$$\hat{f}_l = f(\hat{q}_l), \qquad \text{and} \qquad \hat{S}_l = S(\hat{q}_l). \tag{8.10}$$

If we use a *modal basis*, then these coefficients have to be computed via appropriate $L_2$–projection. Eqn. (8.8) constitutes a *small element–local* nonlinear algebraic equation system which can be easily solved by the following iteration procedure proposed in [**20**],

$$K_{kl}^1\hat{q}_l^{(m+1)} - M_{kl}\hat{S}_l^{(m+1)} = F_{kl}^0\hat{w}_l - +K_{kl}^{\xi}\hat{f}_l^{(m)}, \tag{8.11}$$

which for the linear homogeneous case can be shown to converge to the exact solution in at most $M+1$ steps, where $M$ is the polynomial degree of the basis functions $\theta_k$. The result of the local space–time DG method is a space–time polynomial for each cell which can be directly used within a high–order finite volume or DG finite element scheme.

# Bibliography

[1] D.N. Arnold, F. Brezzi, B. Cockburn, and L.D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM Journal on Numerical Analysis*, 39:1749–1779, 2002.

[2] H. Atkins and C.W. Shu. Quadrature-free implementation of the discontinuous Galerkin method for hyperbolic equations. *AIAA Journal*, 36:775–782, 1998.

[3] E. Audusse, F. Bouchut, M.O. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing*, 25:2050–2065, 2004.

[4] F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *Journal of Computational Physics*, 131:267–279, 1997.

[5] M. Ben-Artzi and J. Falcovitz. A second-order godunov-type scheme for compressible fluid dynamics. *Journal of Computational Physics*, 55:1–32, 1984.

[6] A. Bourgeade, P. LeFloch, and P.A. Raviart. An asymptotic expansion for the solution of the generalized riemann problem. Part II: application to the gas dynamics equations. *Annales de l'institut Henri Poincaré (C) Analyse non linéaire*, 6:437–480, 1989.

[7] A. Canestrelli, M. Dumbser, A. Siviglia, and E.F. Toro. Well-balanced high-order centered schemes on unstructured meshes for shallow water equations with fixed and mobile bed. *Advances in Water Resources*, 33:291–303, 2010.

[8] A. Canestrelli, A. Siviglia, M. Dumbser, and E.F. Toro. A well-balanced high order centered scheme for nonconservative systems: Application to shallow water flows with fix and mobile bed. *Advances in Water Resources*, 32:834–844, 2009.

[9] M.J. Castro, J.M. Gallardo, and C. Parés. High-order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. applications to shallow-water systems. *Mathematics of Computation*, 75:1103–1134, 2006.

[10] M.J. Castro, P.G. LeFloch, M.L. Muñoz-Ruiz, and C. Parés. Why many theories of shock waves are necessary: Convergence error in formally path-consistent schemes. *Journal of Computational Physics*, 227:8107–8129, 2008.

[11] B. Cockburn, S. Hou, and C. W. Shu. The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case. *Mathematics of Computation*, 54:545–581, 1990.

[12] B. Cockburn, S. Y. Lin, and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one dimensional systems. *Journal of Computational Physics*, 84:90–113, 1989.

[13] B. Cockburn and C. W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework. *Mathematics of Computation*, 52:411–435, 1989.

[14] B. Cockburn and C. W. Shu. The Runge-Kutta local projection P1-Discontinuous Galerkin finite element method for scalar conservation laws. *Mathematical Modelling and Numerical Analysis*, 25:337–361, 1991.

[15] B. Cockburn and C. W. Shu. The local discontinuous Galerkin method for time-dependent convection diffusion systems. *SIAM Journal on Numerical Analysis*, 35:2440–2463, 1998.

[16] B. Cockburn and C. W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems. *Journal of Computational Physics*, 141:199–224, 1998.

[17] B. Cockburn and C. W. Shu. Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. *Journal of Scientific Computing*, 16:173–261, 2001.

[18] R. Courant, K. Friedrichs, and H. Lewy. über die partiellen differenzengleichungen der mathematischen physik. *Mathematische Annalen*, 100:32–74, 1928.

[19] M. Dubiner. Spectral methods on triangles and other domains. *Journal of Scientific Computing*, 6:345–390, 1991.

[20] M. Dumbser, D. Balsara, E.F. Toro, and C.D. Munz. A unified framework for the construction of one-step finite-volume and discontinuous Galerkin schemes. *Journal of Computational Physics*, 227:8209–8253, 2008.

[21] M. Dumbser, M. Castro, C. Parés, and E.F. Toro. ADER schemes on unstructured meshes for nonconservative hyperbolic systems: Applications to geophysical flows. *Computers and Fluids*, 38:1731–1748, 2009.

[22] M. Dumbser, C. Enaux, and E.F. Toro. Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws. *Journal of Computational Physics*, 227:3971–4001, 2008.

[23] M. Dumbser, A. Hidalgo, M. Castro, C. Parés, and E.F. Toro. FORCE schemes on unstructured meshes II: Non–conservative hyperbolic systems. *Computer Methods in Applied Mechanics and Engineering*, 2010.

[24] M. Dumbser and M. Käser. Arbitrary high order non-oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems. *Journal of Computational Physics*, 221:693–723, 2007.

[25] M. Dumbser, M. Käser, V.A Titarev, and E.F. Toro. Quadrature-free non-oscillatory finite volume schemes on unstructured meshes for nonlinear hyperbolic systems. *Journal of Computational Physics*, 226:204–243, 2007.

[26] M. Dumbser and C.D. Munz. Building blocks for arbitrary high order discontinuous Galerkin schemes. *Journal of Scientific Computing*, 27:215–230, 2006.

[27] M. Dumbser and E.F. Toro. On universal Osher–type schemes for general nonlinear hyperbolic conservation laws. *Communications in Computational Physics*. in press.

[28] M. Dumbser and E.F. Toro. A simple extension of the Osher Riemann solver to non-conservative hyperbolic systems. *Journal of Scientific Computing*. in press, DOI: 10.1007/s10915-010-9400-3.

[29] B. Einfeldt. On godunov-type methods for gas dynamics. *SIAM Journal on Numerical Analysis*, 25:294–318, 1988.

[30] B. Einfeldt, C. D. Munz, P. L. Roe, and B. Sjögreen. On godunov-type methods near low densities. *Journal of Computational Physics*, 92:273–295, 1991.

[31] P. Le Floch and P.A. Raviart. An asymptotic expansion for the solution of the generalized riemann problem. Part I: General theory. *Annales de l'institut Henri Poincaré (C) Analyse non linéaire*, 5:179–207, 1988.

[32] P. Le Floch and L. Tatsien. A global asymptotic expansion for the solution of the generalized riemann problem. *Annales de l'institut Henri Poincaré (C) Analyse non linéaire*, 3:321–340, 1991.

[33] J.E. Fromm. A method for reducing dispersion in convective difference schemes. *Journal of Computational Physics*, 3:176–189, 1968.

[34] P. Garcia-Navarro and M.E. Vázquez-Cendón. On numerical treatment of the source terms in the shallow water equations. *Computers & Fluids*, 29:951–979, 2000.

[35] G. Gassner, F. Lörcher, and C.D. Munz. A contribution to the construction of diffusion fluxes for finite volume and discontinuous Galerkin schemes. *Journal of Computational Physics*, 224:1049–1063, 2007.

[36] E. Godlewski and P. A. Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws.* Springer, 1996.

[37] S.K. Godunov. Finite difference methods for the computation of discontinuous solutions of the equations of fluid dynamics. *Mathematics of the USSR: Sbornik*, 47:271–306, 1959.

[38] J. M. Greenberg and A. Y. Le Roux. A wellbalanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM Journal on Numerical Analysis*, 33:1–16, 1996.

[39] A. Harten. High Resolution Schemes for Hyperbolic Conservation Laws. *J. Comput. Phys.*, 49:357–393, 1983.

[40] A. Harten, B. Engquist, S. Osher, and S. Chakravarthy. Uniformly high order essentially non-oscillatory schemes, III. *Journal of Computational Physics*, 71:231–303, 1987.

[41] A. Harten, P.D. Lax, and B. van Leer. On upstream differencing and godunov-type schemes for hyperbolic conservation laws. *SIAM Review*, 25(1):35–61, 1983.

[42] R. Hartmann and P. Houston. Symmetric interior penalty DG methods for the compressible navier–stokes equations I: Method formulation. *Int. J. Num. Anal. Model.*, 3:1–20, 2006.

[43] C. Hirsch. *Numerical Computation of Internal and External Flows Vol I: Fundamentals of Numerical Discretisation.* Wiley, 1988.

[44] C. Hirsch. *Numerical Computation of Internal and External Flows Vol II: Computational Methods for Inviscid and Viscous Flow.* Wiley, 1988.

[45] T.Y. Hou and P.G. LeFloch. Why nonconservative schemes converge to wrong solutions: error analysis. *Mathematics of Computation*, 62:497–530, 1994.

[46] C. Hu and C.W. Shu. Weighted essentially non-oscillatory schemes on triangular meshes. *Journal of Computational Physics*, 150:97–127, 1999.

[47] G. Jiang and C.W. Shu. On a cell entropy inequality for discontinuous Galerkin methods. *Mathematics of Computation*, 62:531–538, 1994.

[48] G.-S. Jiang and C.W. Shu. Efficient implementation of weighted ENO schemes. *Journal of Computational Physics*, 126:202–228, 1996.

[49] G. E. Karniadakis and S. J. Sherwin. *Spectral/hp Element Methods in CFD*. Oxford University Press, 1999.

[50] V. P. Kolgan. Application of the minimum-derivative principle in the construction of finite-difference schemes for numerical analysis of discontinuous solutions in gas dynamics. *Transactions of the Central Aerohydrodynamics Institute*, 3(6):68–77, 1972. in Russian.

[51] D. Kröner. *Numerical Schemes for Conservation Laws*. Wiley - Teubner, 1997.

[52] P.D. Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Communications in Pure and Applied Mathematics*, 7:159–193, 1954.

[53] P.D. Lax and B. Wendroff. Systems of conservation laws. *Communications in Pure and Applied Mathematics*, 13:217–237, 1960.

[54] R. J. LeVeque. Balancing source terms and flux gradients in highresolution Godunov methods. *Journal of Computational Physics*, 146:346–365, 1998.

[55] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[56] D. Levy, C. W. Shu, and J. Yan. Local discontinuous Galerkin methods for nonlinear dispersive equations. *Journal of Computational Physics*, 196:751–772, 2004.

[57] X.D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *Journal of Computational physics*, 115:200–212, 1994.

[58] G. Dal Maso, P.G. LeFloch, and F. Murat. Definition and weak stability of nonconservative products. *J. Math. Pures Appl.*, 74:483–548, 1995.

[59] S. Noelle, N. Pankratz, G. Puppo, and J.R. Natvig. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *Journal of Computational Physics*, 213:474–499, 2006.

[60] S. Noelle, Y.L. Xing, and C.W. Shu. High-order well-balanced finite volume WENO schemes for shallow water equation with moving water. *Journal of Computational Physics*, 226:29–58, 2007.

[61] S. Osher. Riemann solvers, the entropy condition and difference approximations. *SIAM Journal on Numerical Analysis*, 21:217–235, 1984.

[62] S. Osher and F. Solomon. Upwind difference schemes for hyperbolic conservation laws. *Math. Comput.*, 38:339–374, 1982.

[63] C. Parés. Numerical methods for nonconservative hyperbolic systems: a theoretical framework. *SIAM Journal on Numerical Analysis*, 44:300–321, 2006.

[64] J. Qiu, M. Dumbser, and C.W. Shu. The discontinuous Galerkin method with Lax-Wendroff type time discretizations. *Computer Methods in Applied Mechanics and Engineering*, 194:4528–4543, 2005.

[65] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.

[66] S. Rhebergen, O. Bokhove, and J.J.W. van der Vegt. Discontinuous Galerkin finite element methods for hyperbolic nonconservative partial differential equations. *Journal of Computational Physics*, 227:1887–1922, 2008.

[67] B. Riemann. Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite. *Göttinger Nachrichten*, 19, 1859.

[68] B. Riemann. Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite. *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen*, 8:43–65, 1860.

[69] P.L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43:357–372, 1981.

[70] V. V. Rusanov. Calculation of Interaction of Non–Steady Shock Waves with Obstacles. *J. Comput. Math. Phys. USSR*, 1:267–279, 1961.

[71] E.N. Sarmin and L.A. Chudov. On the stability of the numerical integration of systems of ordinary differential equations arising in the use of the straight line method. *USSR Computational Mathematics and Mathematical Physics*, 3:1537–1543, 1963.

[72] J. Shi, C. Hu, and C.W. Shu. A technique of treating negative weights in WENO schemes. *Journal of Computational Physics*, 175:108–127, 2002.

[73] Y.I. Shokin. *The method of differential approximation*. Springer Verlag, 1983.

[74] A. Taube, M. Dumbser, D. Balsara, and C.D. Munz. Arbitrary high order discontinuous Galerkin schemes for the magnetohydrodynamic equations. *Journal of Scientific Computing*, 30:441–464, 2007.

[75] V.A. Titarev and E.F. Toro. ADER: Arbitrary high order Godunov approach. *Journal of Scientific Computing*, 17(1-4):609–618, December 2002.

[76] V.A. Titarev and E.F. Toro. ADER schemes for three-dimensional nonlinear hyperbolic systems. *Journal of Computational Physics*, 204:715–736, 2005.

[77] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics, Second Edition.* Springer–Verlag, 1999.

[78] E. F. Toro and V. A. Titarev. Derivative Riemann solvers for systems of conservation laws and ADER methods. *Journal of Computational Physics*, 212(1):150–165, 2006.

[79] E.F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer, third edition, 2009.

[80] E.F. Toro and S. J. Billet. Centered TVD schemes for hyperbolic conservation laws. *IMA Journal of Numerical Analysis*, 20:44–79, 2000.

[81] E.F. Toro, A. Hidalgo, and M. Dumbser. FORCE schemes on unstructured meshes I: Conservative hyperbolic systems. *Journal of Computational Physics*, 228:3368–3389, 2009.

[82] E.F. Toro, M. Spruce, and W. Speares. Restoration of the contact surface in the Harten-Lax-van Leer Riemann solver. *Journal of Shock Waves*, 4:25–34, 1994.

[83] E.F. Toro and V. A. Titarev. Solution of the generalized Riemann problem for advection-reaction equations. *Proc. Roy. Soc. London*, pages 271–281, 2002.

[84] E.F. Toro and V.A. Titarev. Very high order godunov-type schemes for nonlinear scalar conservation laws. In *Proceedings of ECCOMAS CFD Conference 2001*. ECCOMAS CFD Conference, 2001.

[85] E.F. Toro and V.A. Titarev. ADER schemes for scalar hyperbolic conservation laws with source terms in three space dimensions. *Journal of Computational Physics*, 202:196–215, 2005.

[86] I. Toumi. A weak formulation of Roe's approximate Riemann solver. *Journal of Computational Physics*, 102:360–373, 1992.

[87] B. van Leer. Towards the ultimate conservative difference scheme V: A second order sequel to Godunov's method. *Journal of Computational Physics*, 32:101–136, 1979.

[88] M.E. Vázquez-Cendón. Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry. *Journal of Computational Physics*, 148:497–526, 1999.

[89] D.H. Wagner. Equivalence of euler and lagrangian equations of gas dynamics for weak solutions. *Journal of Differential Equations*, 68:118–136, 1987.

[90] R.F. Warming and B.J. Hyett. The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of Computational Physics*, 14:159–179, 1974.

[91] K. Xu. A well-balanced gas-kinetic scheme for the shallow-water equations with source terms. *Journal of Computational Physics*, 178:533–562, 2002.

[92] J. Yan and C.W. Shu. A local discontinuous Galerkin method for KdV-type equations. *SIAM Journal on Numerical Analysis*, 40:769–791, 2002.

[93] Y.T. Zhang and C.W. Shu. Third order WENO scheme on three dimensional tetrahedral meshes. *Communications in Computational Physics*, 5:836–848, 2009.