Lecture Notes for the Course
Numerical Methods for Continuum Mechanics 2
Conservation Laws

Walter Zulehner
Institute for Computational Mathematics
Johannes Kepler University Linz

Winter Semester 2012/13

Contents

Chapter 1

Models

1.1 Kinematics

Let $\Omega \subset \mathbb{R}^3$ be an open, bounded and connected set with Lipschitz-continuous boundary $\Gamma = \partial\Omega$. The set Ω is called the reference configuration and describes, e.g., the initial state of a fluid.

A configuration is a sufficiently smooth, orientation preserving and injective mapping

$$\phi\colon \Omega \longrightarrow \mathbb{R}^3.$$

This mapping describes, e.g., the state of the fluid at some later time. The set $\phi(\Omega)$ consists of all points (or particles) x of the form

$$x = \phi(X)$$

with $X \in \Omega$. X are called the material (or Lagrangian) coordinates, x are called the spatial (or Eulerian) coordinates of a particle.

The motion of a fluid is described by a curve

$$t \mapsto \phi_t.$$

Interpretation: The position x of a fluid particle at time t, whose position at time 0 was X, is given by

$$x = \phi_t(X) \equiv \phi(X, t).$$

Then the material (or Lagrangian) velocity of this fluid particle as a function of X and t is given by

$$V_t(X) = V(X, t) = \frac{\partial \phi}{\partial t}(X, t),$$

and the material (or Lagrangian) acceleration is given by

$$A_t(X) = A(X, t) = \frac{\partial^2 \phi}{\partial t^2}(X, t).$$

Observe the following linear relation between velocity and acceleration:

$$A(X,t) = \frac{\partial V}{\partial t}(X,t).$$

In the Eulerian approach the motion of a particle is described by the spatial velocity (field) $v(x,t)$, where $v(x,t)$ is the velocity of that particle, which passes through x at time t, so

$$v_t(x) = v(x,t) = V(X,t) = \frac{\partial \phi}{\partial t}(X,t) \text{ with } x = \phi(X,t).$$

i.e.:

$$v(x,t) = \frac{\partial \phi}{\partial t}(\phi_t^{-1}(x),t).$$

For the spatial acceleration $a(x,t)$ of that particle we obtain:

$$a_t(x) = a(x,t) = A(X,t) = \frac{\partial^2 \phi}{\partial t^2}(X,t) \ \text{ with } x = \phi(X,t).$$

We have for $x = \phi(X,t)$:

$$a(x,t) = \frac{\partial}{\partial t}[v(\phi(X,t),t)] = \frac{\partial v}{\partial t}(x,t) + \sum_i v_i(x,t)\frac{\partial v}{\partial x_i}(x,t).$$

Notation: The differential operator $v \cdot \text{grad} = v \cdot \nabla$, given by

$$(v \cdot \text{grad})f = (v \cdot \nabla)f = \sum_{i=1}^{3} v_i \frac{\partial f}{\partial x_i},$$

is called the convective derivative and the differential operator d/dt, given by

$$\frac{df}{dt} = \dot{f} = \frac{\partial f}{\partial t} + (v \cdot \text{grad})f,$$

is called the total or material derivative.

With these notations the spatial acceleration can be written in the following form:

$$a(x,t) = \frac{dv}{dt}(x,t) = \frac{\partial v}{\partial t}(x,t) + (v(x,t)\cdot \text{grad})v(x,t) = \frac{\partial v}{\partial t}(x,t) + (v(x,t)\cdot\nabla)v(x,t).$$

Observe that this is a nonlinear relation between velocity and acceleration in the Eulerian approach.

Remark: For a given velocity (field) $v(x,t)$ one obtains the trajectories $\phi(X,t)$ of the individual particles as solution of the initial value problem:

$$\begin{aligned} \frac{\partial \phi}{\partial t}(X,t) &= v(\phi(X,t),t), \\ \phi(X,0) &= X. \end{aligned}$$

1.2 Balance Laws

The set

$$\Omega_t = \phi_t(\Omega) = \{\phi(X,t) \bigm| X \in \Omega\}$$

describes the position of all particles from the reference configuration at time t. Let $\omega \subset \overline{\omega} \subset \Omega$, be an open set with Lipschitz-continuous boundary. Then the set ω_t, given by

$$\omega_t = \phi_t(\omega) = \{\phi(X,t) \bigm| X \in \omega\},$$

describes the position of those particles at time t, which were in ω at time $t = 0$.

1.2.1 The Reynolds Transport Theorem

The Reynolds transport theorem describes the rate change of the quantity

$$\mathcal{F}(t) = \int_{\omega_t} F(x,t)\ dx$$

for a given function F of x and t:

Theorem 1.1 (Reynolds transport theorem)**.** *Let ϕ be twice continuously differentiable and F continuously differentiable. Then*

$$\frac{d\mathcal{F}}{dt}(t) = \int_{\omega_t} \left[\frac{\partial F}{\partial t}(x,t) + \operatorname{div}(Fv)(x,t)\right]\ dx = \int_{\omega_t} \left[\frac{dF}{dt}(x,t) + F\ \operatorname{div}(v)(x,t)\right]\ dx.$$

Notation: The following notation was used in the Transport Theorem: $\operatorname{div} G = \nabla \cdot G$, given by

$$\operatorname{div} G = \nabla \cdot G = \sum_{i=1}^{3} \frac{\partial G_i}{\partial x_i}$$

for a continuously differentiable vector-valued function G, is called the divergence of G.

Remark: With the help of Gauss' theorem it follows immediately that

$$\frac{d\mathcal{F}}{dt}(t) = \int_{\omega_t} \frac{\partial F}{\partial t}\ dx + \int_{\partial\omega_t} F\ v \cdot n\ d\sigma.$$

Here $n = n(x,t)$ denotes the outer normal unit vector at a point x on the boundary of ω_t.

1.2.2 Conservation of Mass

Let $\rho(x,t)$ denote the mass density of a body at the position x and time t. The principle of conservation of mass states that no mass will be generated or destroyed, i. e.:

$$\frac{d}{dt}\int_{\omega_t} \rho(x,t)\ dx = 0.$$

Under appropriate smoothness conditions the Transport Theorem implies:

$$\int_{\omega_t} \left[\frac{\partial\rho}{\partial t}(x,t) + \operatorname{div}(\rho v)(x,t)\right]\ dx = 0$$

for all t and all open sets $\omega \subset \overline{\omega} \subset \Omega$ with Lipschitz-continuous boundary. This results in the following differential equation, the so-called equation of continuity (in conservative form):

$$\frac{\partial\rho}{\partial t} + \operatorname{div}(\rho v) = 0. \tag{1.1}$$

1.2.3 Balance of Momentum and Angular Momentum

The total (linear) momentum of all particles in ω_t is given by

$$\int_{\omega_t} \rho(x,t)v(x,t)\ dx.$$

Newton's second law states that the rate of change of the (linear) momentum is equal to the applied forces $F(\omega_t)$, hence

$$\frac{d}{dt}\int_{\omega_t} \rho(x,t)v(x,t)\ dx = F(\omega_t). \tag{1.2}$$

The forces acting on the body can be split into applied body forces $F_V(\omega_t)$ and applied surface forces $F_S(\omega_t)$:

$$F(\omega_t) = F_V(\omega_t) + F_S(\omega_t).$$

If the body forces can be described by a specific force density (force per unit mass) $f(x,t)$, then we obtain the representation

$$F_V(\omega_t) = \int_{\omega_t} \rho(x,t)f(x,t)\ dx.$$

An example of such a force is the force of gravity with $f = (0,0,-g)^T$.

The internal surface forces can be described by a vector $\vec{t}(x,t,n)$ (force per unit area), the so-called Cauchy stress vector:

$$F_S(\omega_t) = \int_{\partial\omega_t} \vec{t}(x,t,n(x,t))\ d\sigma.$$

Summarizing, we obtain the following balance law for the momentum:

$$\frac{d}{dt}\int_{\omega_t} \rho(x,t)v(x,t)\ dx = \int_{\omega_t} \rho(x,t)f(x,t)\ dx + \int_{\partial\omega_t} \vec{t}(x,t,n(x,t))\ d\sigma.$$

The total angular momentum of all particles in ω_t is given by

$$\int_{\omega_t} x \times \rho(x,t)v(x,t)\ dx.$$

Newton's second law states that the rate of change of the angular momentum is equal to the applied torque, so

$$\frac{d}{dt}\int_{\omega_t} x \times \rho(x,t)v(x,t)\ dx = \int_{\omega_t} x \times \rho(x,t)f(x,t)\ dx + \int_{\partial\omega_t} x \times \vec{t}(x,t,n(x,t))\ d\sigma.$$

These two equations are also called equations of motion, in the steady state case, also the equilibrium conditions.

Under reasonable assumptions it can be shown that the stress vector $\vec{t}(x,t,n) = (t_i(x,t,n))_{i=1,2,3}$ can be represented by the so-called Cauchy stress tensor $\sigma = (\sigma_{ij})$ in the following form:

$$t_i(x,t,n) = \sum_j \sigma_{ji}(x,t)\, n_j.$$

Using Gauss' Theorem and the Transport Theorem one obtains for sufficiently smooth functions the following differential equation (in conservative form) from the balance law of momentum:

$$\frac{\partial}{\partial t}(\rho v_i) + \operatorname{div}(\rho v_i v) = \sum_j \frac{\partial \sigma_{ji}}{\partial x_j} + \rho f_i. \tag{1.3}$$

It can be shown that the balance of angular momentum is satisfied if and only if σ is symmetric:

$$\sigma^T = \sigma.$$

Therefore, the balance of momentum can also be written in the following form:

$$\frac{\partial}{\partial t}(\rho v) + \operatorname{div}\rho(v \otimes v) = \operatorname{div}\sigma + \rho f$$

with

$$(v \otimes w)_{ij} = v_i w_j \quad \text{and} \quad \operatorname{div}\sigma = \left(\sum_j \frac{\partial \sigma_{ij}}{\partial x_j}\right)_{i=1,2,3}.$$

1.2.4 Balance of Energy

The total energy of all particles in ω_t is given by

$$\int_{\omega_t} \rho(x,t) e(x,t)\ dx,$$

where $e(x,t)$ the specific energy density of the fluid. The total energy is the sum of the internal energy and the kinetic energy. Hence

$$e = \varepsilon + \frac{1}{2}|v|^2$$

with the specific internal energy density ε.

The law of conservation of energy states that the rate of chance of the total energy is equal to the powers of the volume forces and the surface forces and the amount of transmitted heat:

$$\frac{d}{dt}\int_{\omega_t} \rho(x,t)e(x,t)\ dx = \int_{\omega_t} \rho(x,t)\ f(x,t)\cdot v(x,t)\ dx + \int_{\partial\omega_t} \vec{t}(x,t,n(x))\cdot v(x,t)\ d\sigma + Q(\omega_t).$$

The transmitted heat is given by

$$Q(\omega_t) = \int_{\omega_t} \rho(x,t)\ q(x,t)\ dx + \int_{\partial\omega_t} h(x,t,n(x))\ d\sigma$$

with the specific density $q(x,t)$ of heat sources and the heat flux $h(x,t,n)$ across the the surface. Under reasonable assumptions it can be shown that the heat flux can be represented by a so-called heat flux vector $\vec{q}(x,t)$:

$$h(x,t,n) = -\vec{q}(x,t)\cdot n.$$

Using Gauss' Theorem and the Transport Theorem one obtains for sufficiently smooth functions the following differential equation (in conservative form):

$$\frac{\partial}{\partial t}(\rho e) + \operatorname{div}(\rho e v) = \rho\ (f\cdot v + q) + \operatorname{div}(\sigma v - \vec{q}). \tag{1.4}$$

1.2.5 The Second Law of Thermodynamics

In thermodynamics there is another important state variable, the so-called entropy. The total entropy of all particles in ω_t is given by

$$\int_{\omega_t} \rho(x,t)s(x,t)\ dx,$$

where s denotes the specific entropy density. The Second Law of Thermodynamics states that

$$\frac{d}{dt}\int_{\omega_t} \rho(x,t)s(x,t)\ dx \geq \int_{\omega_t} \frac{\rho(x,t)\ q(x,t)}{T(x,t)}\ dx + \int_{\partial\omega_t} \frac{h(x,t,n(x,t))}{T(x,t)}\ d\sigma,$$

where T denotes the (absolute) temperature.

Using Gauss' Theorem and the Transport Theorem one obtains for sufficiently smooth functions the following differential inequality (in conservative form):

$$\frac{\partial}{\partial t}(\rho s) + \operatorname{div}(\rho s v) \geq \frac{\rho q}{T} - \operatorname{div}\left(\frac{\vec{q}}{T}\right). \tag{1.5}$$

1.3 Constitutive Laws

The equations of motion, the balance laws of mass and energy, and the entropy condition do not yet completely describe the motion of a fluid. These 5 equations (and one inequality) involve so far the 3 components of velocity (describing the motion of the fluid), the 3 thermodynamic variables ρ, T and s, the internal energy ε, the stress tensor σ and the heat flux vector $\vec{q}$. The force density f and the heat source density q are assumed to be given.

Additional information on the fluid is required (constitutive laws). Here we will focus on the following assumptions:

Ideal Fluid

Neglecting the viscosity of the fluid one obtains the following representation of the Cauchy stress tensor

$$\sigma = -p\, I,$$

where $p(x,t)$ denotes the pressure in the fluid at the position x and time t.

No heat flux

The heat flux (by heat conduction) can be neglected:

$$\vec{q} = 0.$$

Perfect gas

Internal energy (caloric equation of state):

$$\varepsilon = C_v\, T.$$

Thermodynamic equation of state:

$$p = \rho\, R\, T.$$

Here R is the specific gas constant and C_v is the specific heat at constant volume. We have $R = C_p - C_v$, where C_p is the specific heat at constant pressure.

Let $\kappa = C_p/C_v > 1$. Then the entropy of a perfect gas is given by

$$s = C_v \ln \frac{p/p_0}{(\rho/\rho_0)^\kappa} + s_0 = C_v \ln \frac{T/T_0}{(\rho/\rho_0)^{\kappa-1}} + s_0.$$

By eliminating T one obtains the relation

$$p = (\kappa - 1)\,\rho\,\varepsilon = (\kappa - 1)\,\left[\rho\, e - \frac{1}{2}\,\rho\, v^2\right] \tag{1.6}$$

and

$$s = C_v \ln \frac{\varepsilon/\varepsilon_0}{(\rho/\rho_0)^{\kappa-1}} + s_0. \tag{1.7}$$

In summary we obtain the following system of differential equations (Euler equations for a compressible inviscid fluid):

$$\begin{aligned}
&\frac{\partial\rho}{\partial t} + \operatorname{div}(\rho v) = 0,\\
&\frac{\partial}{\partial t}(\rho v) + \operatorname{div}\left(\rho v \otimes v + p\, I\right) = \rho f,\\
&\frac{\partial}{\partial t}(\rho e) + \operatorname{div}\left[(\rho e + p) v\right] = \rho\left(f \cdot v + q\right).
\end{aligned}$$

In the case $f = 0$ and $q = 0$ the system reads

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{3} \frac{\partial}{\partial x_j}(\mathbf{f}_j(\mathbf{u})) = 0$$

with

$$\mathbf{u} = \begin{pmatrix} \rho \\ \rho v \\ \rho e \end{pmatrix}, \quad \mathbf{f_j}(\mathbf{u}) = \begin{pmatrix} \rho\, v_j \\ \rho\, v_j\, v + p\, e_j \\ (\rho e + p) v_j \end{pmatrix},$$

where e_j denotes the j-ten canonical unit vector. Such a system is called a system of conservation laws. Observe that p can be written as a function of $\mathbf{u}$, see the definition of $\mathbf{u}$ and (1.6).

If $\mathbf{u}(x,t)$ is a continuously differentiable function, the system of conversation laws is equivalent to the following system of first-order differential equations in quasi-linear form:

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{3} A_j(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x_j} = 0$$

with

$$A_j(\mathbf{u}) = \mathbf{f}_j'(\mathbf{u}).$$

For a complete description additional conditions are necessary:

1. Initial conditions:

$$\mathbf{u}(x,0) = \mathbf{u}_0(x) \quad \text{for all } x \in \Omega.$$

The (pure) initial value problem in the case $\Omega = \mathbb{R}^3$ is called the Cauchy problem.

2. Appropriate boundary conditions in the case $\Omega \neq \mathbb{R}^3$.

From the entropy condition we additionally have in the considered case ($q = 0$ and $\vec{q} = 0$):

$$\frac{\partial}{\partial t}(\rho s) + \operatorname{div}(\rho s v) \geq 0.$$

That is

$$\frac{\partial}{\partial t}U(\mathbf{u}) + \sum_{j=1}^{3} \frac{\partial}{\partial x_j}(F_j(\mathbf{u})) \leq 0$$

with $U = -\rho\, s$ and $F = (F_1, F_2, F_3)^T = -\rho\, s\, v$. Observe that s, U and F can be written as functions of $\mathbf{u}$, see the definition of $\mathbf{u}$ and (1.7).

It can be shown that

$$U'(\mathbf{u})\mathbf{f}_j'(\mathbf{u}) = F_j'(\mathbf{u}) \quad \text{for } j = 1, 2, 3.$$

Therefore, if $\mathbf{u}(x, t)$ is a continuously differentiable function, it follows that

$$\begin{aligned} &\frac{\partial}{\partial t}U(\mathbf{u}) + \sum_{j=1}^{3} \frac{\partial}{\partial x_j}(F_j(\mathbf{u})) \\ = \;& U'(\mathbf{u})\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{3} F_j'(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x_j} = U'(\mathbf{u})\left[\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{3} \mathbf{f}_j'(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x_j}\right] \\ = \;& U'(\mathbf{u})\left[\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{3} \frac{\partial}{\partial x_j}(\mathbf{f}_j(\mathbf{u}))\right]. \end{aligned}$$

So, any smooth solution of the system of conservation laws automatically satisfies the entropy condition with equality.

Remark:

1. More complicated models involve the thermodynamics of real gases, based on a thermodynamic equation of state of the form

   $$p = p(\rho, T),$$

   a caloric equation of state of the form

   $$\varepsilon = \varepsilon(\rho, T)$$

   and a representation of the entropy of the form

   $$s = s(\rho, T).$$

Assume that the caloric equation of state can be used to express T in terms of ε and ρ. Then, from the thermodynamic equation of state we obtain a relation of the form

$$p = p(\rho, \rho\,\varepsilon)$$

and the entropy can be expressed similarly:

$$s = s(\rho, \rho\,\varepsilon)$$

2. In some applications the equation of state reduces to the form

$$p = p(\rho),$$

e.g. for problems in gas dynamics with constant entropy (isentropic flow) and the system consisting of the equation of continuity and the equations of motion suffice to describe the problem.

3. A typical law for modeling the heat flux (heat conduction) is Fourier's law:

$$\vec{q} = -k\,\mathrm{grad}\,T$$

where k denotes the heat conductivity.

4. If viscosity is included in the form

$$\sigma = -pI + 2\mu\,D$$

with

$$D_{ij} = \frac{1}{2}\left[\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right]$$

and the dynamic viscosity μ, the compressible Navier-Stokes equations are obtained.

More generally we will discuss conservation laws of the form

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{d} \frac{\partial}{\partial x_j}(\mathbf{f}_j(\mathbf{u})) = 0$$

with $\mathbf{u} : \Omega \longrightarrow \mathbb{R}^p$, $\Omega \subset \mathbb{R}^d$ open, and continuously differentiable functions $\mathbf{f}_j : D \longrightarrow \mathbb{R}^p$, $D \subset \mathbb{R}^p$ open. The entropy condition is of the general form

$$\frac{\partial}{\partial t} U(\mathbf{u}) + \sum_{j=1}^{d} \frac{\partial}{\partial x_j}(F_j(\mathbf{u})) \leq 0$$

with continuously differentiable functions $U : D \longrightarrow \mathbb{R}$ and $F_j : D \longrightarrow \mathbb{R}$ satisfying the conditions

$$U'(\mathbf{u})\mathbf{f}_j'(\mathbf{u}) = F_j'(\mathbf{u}) \quad \text{for } j = 1, \ldots, d.$$

In Chapter 2 the one-dimensional scalar case is considered ($d = 1$ and $p = 1$). In Chapter 3 we study one-dimensional systems ($d = 1$ and $p > 1$) including the one-dimensional Euler equations ($d = 1$ and $p = 3$: $\rho = \rho(x_1, t)$, $v_1 = v_1(x_1, t)$, $v_2 = v_3 \equiv 0$, $e = e(x_1, t)$). Next, in Chapter 4, multi-dimensional scalar conservation laws are considered ($d > 1$ and $p = 1$). And, finally, multi-dimensional systems ($d > 1$ and $p > 1$) including the full Euler equations ($d = 3$ and $p = 5$) are addressed.

Chapter 2

One-Dimensional Scalar Conservation Laws

2.1 Weak Solutions

Consider the following initial value problem of a one-dimensional scalar conversation law (differential equation in conservative form):

Find a function $u : \mathbb{R} \times [0, \infty) \longrightarrow \mathbb{R}$ such that

$$\begin{aligned} u_t + f(u)_x &= 0 \qquad x \in \mathbb{R},\ t > 0, && (2.1)\\ u(x, 0) &= u_0(x) \quad x \in \mathbb{R}, && (2.2) \end{aligned}$$

where the continuously differentiable flux $f : \mathbb{R} \longrightarrow \mathbb{R}$ and the initial value $u_0 : \mathbb{R} \longrightarrow \mathbb{R}$ are given.

Using the chain rule one obtains the differential equation in quasi-linear form for continuously differentiable solutions u:

$$u_t + f'(u)u_x = 0.$$

Example:

1. The linear wave equation:

   $$u_t + a\, u_x = 0$$

   where a is a given constant. Here the flux is linear: $f(u) = a\, u$.

2. Burgers' equation

   $$u_t + \left(\frac{1}{2}u^2\right)_x = 0$$

   or in quasi-linear form

   $$u_t + u\, u_x = 0$$

   with the non-linear flux $f(u) = u^2/2$.

Let $f \in C^1(\mathbb{R})$ and let $u \in C^1(\mathbb{R} \times (0, \infty)) \cap C(\mathbb{R} \times [0, \infty))$ be a (classical) solution of (2.1), (2.2). With the help of characteristic curves the one-dimensional scalar conservation laws are easy to analyze:

Definition 2.1. *A curve Γ_{x_0}, parameterized by $(\gamma(t), t)$, $t \in [0, \infty)$, is called a characteristic curve if and only if*

$$\begin{aligned} \gamma'(t) &= f'(u(\gamma(t), t)) \quad t \in (0, \infty), \\ \gamma(0) &= x_0. \end{aligned}$$

The solution u is constant along a characteristic curve:

$$\frac{d}{dt} u(\gamma(t), t) = u_t(\gamma(t), t) + u_x(\gamma(t), t)\gamma'(t) = u_t + f'(u)u_x = 0.$$

Hence

$$u(\gamma(t), t) = u(\gamma(0), 0) = u_0(x_0).$$

Therefore:

$$\gamma'(t) = f'(u(\gamma(t), t)) = f'(u_0(x_0)) = \text{constant}.$$

In summary one obtains:

Theorem 2.1. *Let $f \in C^1(\mathbb{R})$ and let $u \in C^1(\mathbb{R} \times (0, \infty)) \cap C(\mathbb{R} \times [0, \infty))$ be a solution of (2.1), (2.2). Then*

1. *All characteristic curves are straight lines.*

2. *The solution u is constant along a characteristic curve.*

Example:

1. The linear wave equation $u_t + a\, u_x = 0$: The characteristic curves are straight lines with slope $1/a$ in the x-t-diagram, given by

   $$x - a\, t = x_0.$$

   Hence we have:

   $$u(x, t) = u_0(x - a\, t).$$

   This setting for u would also make sense for discontinuous initial values like

   $$u_0(x) = \begin{cases} 1 \text{ for } x < 0, \\ 0 \text{ for } x \geq 0, \end{cases}$$

   although, in this case, u is not a smooth solution.

2. For Burgers' equation $u_t + u\, u_x = 0$ with a smooth initial value satisfying the property

$$u_0(x) = \begin{cases} 1 \text{ for } x < -1, \\ 0 \text{ for } x \geq 0 \end{cases}$$

it is easy to show that characteristic curves intersect. So, C^1-solutions exist only for a finite time interval. What happens afterwards?

This discussion shows that the concept of a solution has to be reconsidered. For that the so-called weak form of a conservation law is introduced:

Let $\varphi \in C_0^\infty(\mathbb{R} \times [0, \infty))$, i.e.: φ is infinitely many times differentiable and has a compact support in $\mathbb{R} \times [0, \infty)$. Moreover, let $f \in C^1(\mathbb{R})$ and let $u \in C^1(\mathbb{R} \times (0, \infty)) \cap C(\mathbb{R} \times [0, \infty))$ be a classical solution of (2.1), (2.2). By multiplying (2.1) by φ and integrating over $\mathbb{R} \times [0, \infty)$ one obtains

$$\int_{-\infty}^{\infty} \int_0^\infty u_t\, \varphi \; dt \; dx + \int_{-\infty}^{\infty} \int_0^\infty f(u)_x\, \varphi \; dt \; dx = 0.$$

By integration by parts it follows that

$$\int_{-\infty}^{\infty} \left[u\, \varphi \Big|_{t=0}^{t=\infty} - \int_0^\infty u\, \varphi_t \; dt \right] dx \;+$$
$$+ \int_0^\infty \left[f(u)\, \varphi \Big|_{x=-\infty}^{x=\infty} - \int_{-\infty}^{\infty} f(u)\, \varphi_x \; dt \right] dx = 0.$$

Hence

$$\int_{-\infty}^{\infty} \int_0^\infty [u\varphi_t + f(u)\varphi_x] \; dt \; dx + \int_{-\infty}^{\infty} u_0(x)\varphi(x, 0) \; dx = 0.$$

This equation motivates the following definition of a weak solution:

Definition 2.2. *Let $u_0 \in L^\infty_{loc}(\mathbb{R})$. A function $u \in L^\infty_{loc}(\mathbb{R} \times [0, \infty))$ is called a weak solution (a solution in the sense of distributions) of the Cauchy problem (2.1), (2.2) if and only if*

$$\int_{-\infty}^{\infty} \int_0^\infty [u\varphi_t + f(u)\varphi_x] \; dt \; dx + \int_{-\infty}^{\infty} u_0(x)\varphi(x, 0) \; dx = 0$$

for all $\varphi \in C_0^\infty(\mathbb{R} \times [0, \infty))$.

Let $u : \mathbb{R} \times [0, \infty) \longrightarrow \mathbb{R}$ be a weak solution of (2.1), (2.2), which is piecewise smooth in the following sense: There is a smooth curve Σ in $\mathbb{R} \times [0, \infty)$, parameterized by $(\sigma(t), t)$, $t \in [0, \infty)$, which divides the set $Q = \mathbb{R} \times (0, \infty)$ into two parts Q_L and Q_R, and there are functions $u_L \in C^1(\overline{Q}_L)$ and $u_R \in C^1(\overline{Q}_R)$ with

$$u(x, t) = \begin{cases} u_L(x, t) \text{ for } (x, t) \in Q_L, \\ u_R(x, t) \text{ for } (x, t) \in Q_R. \end{cases}$$

Let $\varphi \in C_0^\infty(\mathbb{R} \times [0,\infty))$ be an arbitrary test function. Then

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_0^{\infty} [u\varphi_t + f(u)\varphi_x]\ dt + \int_{-\infty}^{\infty} u_0(x)\varphi(x,0)\ dx \\
&= \int_{Q_L} [u_L\varphi_t + f(u_L)\varphi_x]\ dt\ dx + \int_{Q_R} [u_R\varphi_t + f(u_R)\varphi_x]\ dt\ dx + \int_{-\infty}^{\infty} u_0(x)\varphi(x,0)\ dx \\
&= \int_\Sigma [\nu_t\, u_L + \nu_x\, f(u_L)]\varphi\ ds - \int_{-\infty}^{\sigma(0)} u_L(x,0)\varphi(x,0)\ dx - \int_{Q_L} [(u_L)_t + f(u_L)_x]\varphi\ dt\ dx \\
&\quad - \int_\Sigma [\nu_t\, u_R + \nu_x\, f(u_R)]\varphi\ ds - \int_{\sigma(0)}^{\infty} u_R(x,0)\varphi(x,0)\ dx - \int_{Q_R} [(u_R)_t + f(u_R)_x]\varphi\ dt\ dx \\
&\quad + \int_{-\infty}^{\infty} u_0(x)\varphi(x,0)\ dx \\
&= \int_\Sigma [\nu_t\,(u_L - u_R) + \nu_x\,(f(u_L) - f(u_R))]\varphi\ ds \\
&\quad + \int_{-\infty}^{\sigma(0)} [u_0(x) - u_L(x,0)]\varphi(x,0)\ dx + \int_{\sigma(0)}^{\infty} [u_0(x) - u_R(x,0)]\varphi(x,0)\ dx \\
&\quad - \int_{Q_L} [(u_L)_t + f(u_L)_x]\varphi\ dt\ dx - \int_{Q_R} [(u_R)_t + f(u_R)_x]\varphi\ dt\ dx.
\end{aligned}$$

Here ν denotes the outer normal vector to the boundary Σ of Q_L, given by

$$\nu = \begin{pmatrix} \nu_x \\ \nu_t \end{pmatrix} = \frac{1}{\sqrt{1+\sigma'(t)^2}} \begin{pmatrix} 1 \\ -\sigma'(t) \end{pmatrix}.$$

So u is a weak solution if and only if it is a classical solution to (2.1) in Q_L and Q_R satisfying the initial condition (2.2) pointwise on $(-\infty,\sigma(0))$ and $(\sigma(0),\infty)$, and

$$\nu_t\,(u_L - u_R) + \nu_x\,(f(u_L) - f(u_R)) = 0 \quad \text{on } \Sigma,$$

i.e.:

$$\sigma'(t)(u_R - u_L)\Big|_{(x,t)=(\sigma(t),t)} = (f(u_R) - f(u_L))\Big|_{(x,t)=(\sigma(t),t)}$$

or, shortly,

$$s\,[u] = [f(u)],$$

with

$$s = \sigma'(t), \quad [u] = u_R - u_L, \quad [f(u)] = f(u_R) - f(u_L).$$

Here s is the speed of propagation of the discontinuity and [.] denotes the jump across Σ. This condition is called the Rankine-Hugoniot jump condition.

The statement can be easily extended to more general piecewise smooth functions. A function $u : \mathbb{R} \times [0,\infty) \longrightarrow \mathbb{R}$ is called piecewise smooth if there is a finite number of smooth (C^1-)curves Σ_r, $r = 1, \ldots$ in $\mathbb{R} \times [0,\infty)$ such that u is smooth (C^1) outside these curves, and one-sided limits $u^\pm(x,t)$ exist for all points $(x,t) \in \Sigma_r$.

Theorem 2.2. *Let $u : \mathbb{R} \times [0, \infty) \longrightarrow \mathbb{R}$ be a piecewise smooth function. Then u is a weak solution to the Cauchy problem (2.1), (2.2) if and only if*

1. *u is a smooth solution in each domain where u is smooth and*

2. *u satisfies the condition*

$$s[u] = [f(u)]$$

for each point on each curve of discontinuity, where $(1, -s)^T$ is a normal vector to the curve of discontinuity in the considered point.

Example:

1. For the linear wave equation $u_t + a\,u_x = 0$ it follows for a possible point $\sigma(t)$ of discontinuity:

$$s\,(u_R - u_L) = a\,(u_R - u_L),$$

hence

$$s = a,$$

i.e.: discontinuities of a piecewise smooth weak solution are possible only along a characteristic curve.

2. For the speed of propagation of a discontinuity with Burgers' equation the Rankine-Hugoniot condition implies:

$$s(u_R - u_L) = \frac{1}{2}(u_R^2 - u_L^2) = \frac{1}{2}(u_R + u_L)(u_R - u_L),$$

hence

$$s = \frac{1}{2}(u_L + u_R).$$

For the special initial value

$$u_0(x) = \begin{cases} 1 \text{ for } x < 0, \\ 0 \text{ for } x \geq 0 \end{cases}$$

one obtains a piecewise constant weak solution with $u_L = 1$ and $u_R = 0$, separated by the curve of discontinuity $x = t/2$ ($s = 1/2$).

2.2 Entropy Solutions

The concept of a weak solution, however, does not necessarily guarantee uniqueness of the solution.

Example: For Burgers' equation with initial value

$$u_0(x) = \begin{cases} 0 \text{ for } x < 0, \\ 1 \text{ for } x \geq 0 \end{cases}$$

a smooth solution cannot be determined with the help of the characteristic curves for $0 < x < t$. The characteristic curves, which are determined by the initial values, do not completely fill the domain $\mathbb{R} \times [0, \infty)$. This can be used to construct several weak solutions.

Analogous to above one obtains, e.g., a piecewise constant weak solution

$$u(x,t) = \begin{cases} 0 \text{ for } x < t/2, \\ 1 \text{ for } x \geq t/2. \end{cases} \tag{2.3}$$

Besides this discontinuous solution (and many more discontinuous solutions) there is also a piecewise smooth and for $t > 0$ continuous (and, therefore, a weak) solution:

$$u(x,t) = \begin{cases} 0 & \text{for } x < 0 \\ \dfrac{x}{t} & \text{for } 0 \leq x < t, \\ 1 & \text{for } x \geq t. \end{cases}$$

So the question is: Which one is the "right" solution?

For motivation consider the system of conservation laws in gas dynamics (say in the one-dimensional case), which are of the form

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0.$$

The Second Law of Thermodynamics implies the existence of the entropy s, which satisfies a differential inequality

$$U(\mathbf{u})_t + F(\mathbf{u})_x \leq 0$$

with $U(\mathbf{u}) = -\rho\, s$ and $F(\mathbf{u}) = -\rho\, v\, s$. It can be shown that $U \in C^2$ and $U''(\mathbf{u}) > 0$ which implies that U is a convex function in $\mathbf{u}$, i.e.:

$$U(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \alpha U(\mathbf{u}) + (1-\alpha)U(\mathbf{v}) \quad \text{for all } \mathbf{u}, \mathbf{v} \text{ and all } \alpha \in [0,1].$$

Moreover the following compatibility condition can be shown:

$$U'(\mathbf{u})\mathbf{f}'(\mathbf{u}) = F'(\mathbf{u}).$$

This condition guarantees that

$$U(\mathbf{u})_t + F(\mathbf{u})_x = 0,$$

for continuously differentiable solutions of the conservation law. So the entropy condition is satisfied automatically for smooth solutions $\mathbf{u}$.

These properties motivate the following definition:

Definition 2.3. *Let $f \in C^1(\mathbb{R})$. Functions $U, F \in C^1(\mathbb{R})$ are called an entropy pair if and only if*

1. *U is convex and*
2. *$U'(u)f'(u) = F'(u)$ for all $u \in \mathbb{R}$.*

For the scalar case the existence of an entropy is trivial: Each convex function $U(u) \in C^1(\mathbb{R})$ is an entropy: Take for $F(u)$ a primitive of $U'(u)f'(u)$.

Example: For Burgers' equation and the strictly convex function $U(u) = u^2/2$ the entropy flux $F(u)$ has to be a primitive of $U'(u)f'(u) = u^2$, so we can choose

$$U(u) = \frac{1}{2}u^2, \quad F(u) = \frac{1}{3}u^3$$

as an entropy pair with a strictly convex entropy function.

Analogous to the weak formulation of the conservation law the classical differential inequality

$$U(u)_t + F(u)_x \leq 0$$

implies the condition

$$\int_{-\infty}^{\infty}\int_0^{\infty} [U(u)\varphi_t + F(u)\varphi_x] \; dt \; dx + \int_{-\infty}^{\infty} U(u_0(x))\,\varphi(x,0) \; dx \geq 0$$

for all test functions $\varphi \in C_0^\infty(\mathbb{R} \times [0,\infty))$ with $\varphi \geq 0$. This leads to the following definition:

Definition 2.4. *A weak solution $u \in L^\infty_{loc}(\mathbb{R} \times [0,\infty))$ of the Cauchy problem (2.1), (2.2) is called an entropy solution if and only if*

$$\int_{-\infty}^{\infty}\int_0^{\infty} [U(u)\varphi_t + F(u)\varphi_x] \; dt \; dx + \int_{-\infty}^{\infty} U(u_0(x))\,\varphi(x,0) \; dx \geq 0$$

for all entropy pairs (U, F) and all test functions $\varphi \in C_0^\infty(\mathbb{R} \times [0,\infty))$ with $\varphi \geq 0$.

For piecewise smooth entropy solution u one obtains analogously to the Rankine-Hugoniot jump condition the following jump inequality:

$$s[U(u)] \geq [F(u)].$$

Example: For Burgers' equation and the entropy pair $U(u) = u^2/2$, $F(u) = u^3/3$ the jump inequality becomes

$$\frac{1}{2}s(u_R^2 - u_L^2) \geq \frac{1}{3}(u_R^3 - u_L^3)$$

with

$$s = \frac{1}{2}(u_L + u_R).$$

This is equivalent to

$$(u_L - u_R)^3 \geq 0.$$

Therefore, for entropy solutions a discontinuity is allowed only if

$$u_L > u_R.$$

This excludes the weak solution (2.3). The continuous weak solution is, of course, an entropy solution.

Remark:

1. In the scalar one-dimensional case one can show the following result: Assume that $f \in C^1(\mathbb{R})$ is strictly convex and u is a piecewise smooth weak solution of (2.1), (2.2) that satisfies the weak entropy condition for one entropy pair with a strictly convex entropy function $U(u)$. Then u satisfies the weak entropy condition for all entropy pairs.

2. In the one-dimensional case one can show the existence and uniqueness of an entropy solution in appropriate function spaces under appropriate assumptions. For example, there is a unique entropy solution $u \in L^\infty(\mathbb{R} \times (0,T))$ for $f \in C^1(\mathbb{R})$ and $u_0 \in L^\infty(\mathbb{R})$. The existence is shown by the vanishing viscosity method. Consider the following perturbed problem:

$$\begin{aligned} u_t + f(u)_x - \varepsilon\, u_{xx} &= 0 \qquad && x \in \mathbb{R},\ t > 0, \\ u(x,0) &= u_0(x) && x \in \mathbb{R}. \end{aligned}$$

One can show that a solution $u_\varepsilon(x,t)$ exists for all $\varepsilon > 0$, and that the limit $u(x,t) = \lim_{\varepsilon \downarrow 0} u_\varepsilon(x,t)$ is an entropy solution.

We will consider entropy solutions as the physically relevant solutions.

Remark:

1. One could instead use directly the vanishing viscosity method to define the limit function as the physically relevant solution.

2. Another possible criterion is the so-called Lax entropy condition (for convex scalar fluxes):

$$f'(u_L) > s > f'(u_R),$$

where s is the speed of propagation of the discontinuity, given by the Rankine-Hugoniot condition:

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

That means that the characteristic curves enter the curve of discontinuity, they are not allowed to originate from a point of the curve of discontinuity.

3. Under appropriate assumptions (but not always) these different concepts of physically relevant solutions coincide.

2.3 Conservative Finite Difference Methods

Let u be a smooth solution of the conservation law

$$u_t + f(u)_x = 0. \tag{2.4}$$

Let $\Delta x > 0$ be a given spatial mesh size and $\Delta t > 0$ be a given time step. Set

$$x_j = j\Delta x,\ x_{j+\frac{1}{2}} = \left(j + \frac{1}{2}\right)\Delta x,\ t_n = n\Delta t.$$

If (2.4) is integrated over the set $[x_{j-1/2}, x_{j+1/2}] \times [t_n, t_{n+1}]$, one obtains:

$$\begin{aligned} 0 &= \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \int_{t_n}^{t_{n+1}} [u_t + f(u)_x]\ dt\ dx \\ &= \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t_{n+1})\ dx - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t_n)\ dx \\ &\quad + \int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}}, t))\ dt - \int_{t_n}^{t_{n+1}} f(u(x_{j-\frac{1}{2}}, t))\ dt. \end{aligned}$$

If divided by $\Delta x\, \Delta t$, we obtain:

$$\begin{aligned} 0 &= \frac{1}{\Delta t}\left[\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t_{n+1})\ dx - \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t_n)\ dx\right] \\ &\quad + \frac{1}{\Delta x}\left[\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}}, t))\ dt - \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j-\frac{1}{2}}, t))\ dt\right]. \end{aligned}$$

This identity is the starting point for computing approximations of the spatial average values

$$u_j^{n+1} \approx \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t_{n+1})\ dx$$

from known approximations

$$u_j^n \approx \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t_n)\ dx.$$

For this one needs appropriate approximations of the temporal average values of the fluxes at the boundary points of the cell $[x_{j-1/2}, x_{j+1/2}]$:

$$\begin{aligned} \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j-\frac{1}{2}}, t))\ dt &\approx g_{j-\frac{1}{2}}^n = g(\ldots, u_{j-1}^n, u_j^n, \ldots), \\ \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}}, t))\ dt &\approx g_{j+\frac{1}{2}}^n = g(\ldots, u_j^n, u_{j+1}^n, \ldots). \end{aligned}$$

Here we allow g to be continuous function in $2q$ arguments.

This motivates the following class of finite difference methods:

Definition 2.5. *A method of the form*

$$\frac{1}{\Delta t}\left(u_j^{n+1}-u_j^n\right)+\frac{1}{\Delta x}\left(g_{j+1/2}^n-g_{j-1/2}^n\right) \;=\; 0$$

with

$$\begin{aligned} g_{j-\frac{1}{2}}^n &= g(\ldots,u_{j-1}^n,u_j^n,\ldots)\\ g_{j+\frac{1}{2}}^n &= g(\ldots,u_j^n,u_{j+1}^n,\ldots) \end{aligned}$$

is called conservative. g is called the numerical flux.

Remark: Let $\theta \in [0,1]$ be a given parameter. The (average value of the) flux can be approximated more generally in the following way:

$$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j\pm\frac{1}{2}}))\;dt \approx (1-\theta)g_{j\pm\frac{1}{2}}^n+\theta g_{j\pm\frac{1}{2}}^{n+1}.$$

This leads in the case $\theta \neq 0$ to implicit conservative methods of the form

$$\frac{1}{\Delta t}\left(u_j^{n+1}-u_j^n\right)+\frac{1}{\Delta x}\left[(1-\theta)(g_{j+1/2}^n-g_{j-1/2}^n)+\theta(g_{j+1/2}^{n+1}-g_{j-1/2}^{n+1})\right] \;=\; 0.$$

Definition 2.5 corresponds to the case $\theta = 0$ of explicit methods only.

Observe that, for a conservative method, the approximation of the temporal average values at the right boundary of the cell $[x_{j-1/2}, x_{j+1/2}]$ and at the left boundary of the neighboring cell $[x_{j+1/2}, x_{j+3/2}]$ coincide.

If u is a constant function then, of course, we have

$$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u)\;dt = f(u).$$

This implies a natural minimal requirement for the numerical flux:

Definition 2.6. *A numerical flux g is called consistent with the flux f if and only if*

$$g(u,\ldots,u) = f(u) \quad \textit{for all } u \in \mathbb{R}.$$

A first attempt to construct a conservative method for the example of the linear wave equation

$$u_t + a\,u_x = 0$$

is based on the use of the forward difference quotient for the time derivative and the central difference quotient for the spatial derivative:

$$\frac{1}{\Delta t}(u_j^{n+1}-u_j^n)+\frac{a}{2\Delta x}(u_{j+1}^n-u_{j-1}^n) = 0. \tag{2.5}$$

The method is an explicit one-step method: Starting from the initial values

$$u_j^0 = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x)\ dx$$

approximation at time t_{n+1} can be computed from the approximations at the previous time t_n by a simple evaluation of the expressions.

The method is conservative with the numerical flux

$$g(u, v) = \frac{a}{2}(u + v).$$

The method is consistent:

$$g(u, u) = \frac{a}{2}(u + u) = au = f(u).$$

However, it is never stable, and, therefore, useless.

The Lax-Friedrichs method

A first attempt to stabilize the method from above leads the so-called Lax-Friedrichs method: The value u_j^n is replaced by the mean value of the two spatial neighbors u_{j-1}^n and u_{j+1}^n. Therefore, the Lax-Friedrichs method for a general conservation law

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

reads

$$\frac{1}{\Delta t}\left(u_j^{n+1} - \frac{1}{2}(u_{j-1}^n + u_{j+1}^n)\right) + \frac{1}{2\Delta x}[f(u_{j+1}^n) - f(u_{j-1}^n)] = 0,$$

or, equivalently,

$$\frac{1}{\Delta t}(u_j^{n+1} - u_j^n) + \frac{1}{2\Delta x}[f(u_{j+1}^n) - f(u_{j-1}^n)] = \frac{1}{2\Delta t}(u_{j+1}^n - 2u_j^n + u_{j-1}^n).$$

The method is conservative with the consistent numerical flux

$$g_{LF}(u, v) = \frac{1}{2}[f(u) + f(v)] - \frac{1}{2\lambda}(v - u),$$

where $\lambda = \Delta t/\Delta x$.

It is intuitively clear this method is more stable. The additional term on the right hand side can be interpreted as the discretization of a diffusion term of the form $(\Delta x)/(2\lambda)\ u_{xx}$.

The Lax-Wendroff method

This method is based on a Taylor expansion in time up to second-order terms. For a smooth solution we have:

$$u_t = -f(u)_x, \quad u_{tt} = -[f(u)_x]_t = -[f(u)_t]_x = -[f'(u)u_t]_x = [f'(u)f(u)_x]_x$$

Therefore, for a Taylor expansion at the point (x_j, t_n) it follows that:

$$\begin{aligned} u(x_j, t_n + \Delta t) &= u + u_t \; \Delta t + \frac{1}{2} u_{tt} \; \Delta t^2 + O(\Delta t^3) \\ &= u - f(u)_x \; \Delta t + \frac{1}{2}[f'(u)f(u)_x]_x \; \Delta t^2 + O(\Delta t^3). \end{aligned}$$

The derivatives are approximated by central difference quotients:

$$f(u)_x(x_j, t_n) \approx \frac{1}{2\Delta x}(f^n_{j+1} - f^n_{j-1})$$

and

$$\begin{aligned} [f'(u)f(u)_x]_x(x_j, t_n) &\approx \frac{1}{\Delta x}\left[f'(u)f(u)_x\Big|_{(x_{j+\frac{1}{2}}, t_n)} - f'(u)f(u)_x\Big|_{(x_{j-\frac{1}{2}}, t_n)} \right] \\ &\approx \frac{1}{\Delta x^2}\left[a^n_{j+1/2}\,(f^n_{j+1} - f^n_j) - a^n_{j-1/2}\,(f^n_j - f^n_{j-1}) \right]. \end{aligned}$$

where $a^n_{k+1/2}$ denotes a suitable approximation of $f'(u(x_{k+1/2}, t))$. Then one obtains:

$$u^{n+1}_j = u^n_j - \frac{\lambda}{2}(f^n_{j+1} - f^n_{j-1}) + \frac{\lambda^2}{2}\left[a^n_{j+\frac{1}{2}}(f^n_{j+1} - f^n_j) - a^n_{j-\frac{1}{2}}(f^n_j - f^n_{j-1}). \right]$$

Possible choices for $a^n_{k+1/2}$:

$$a^n_{k+1/2} = f'((u^n_k + u^n_{k+1})/2)$$

or

$$a^n_{k+\frac{1}{2}} = a(u^n_k, u^n_{k+1}) \quad \text{with} \quad a(u,v) = \frac{f(v) - f(u)}{v - u}.$$

We will concentrate on the second choice which avoids the calculation of the derivative of f: The Lax-Wendroff method is conservative with the consistent numerical flux

$$g_{LW}(u, v) = \frac{1}{2}[f(u) + f(v)] - \frac{1}{2\lambda}\nu(u,v)^2(v - u)$$

where

$$\nu(u, v) = \lambda \frac{f(v) - f(u)}{v - u}.$$

It can also be written in the form

$$\frac{1}{\Delta t}(u_j^{n+1} - u_j^n) + \frac{1}{2\Delta x}[f(u_{j+1}^n) - f(u_{j-1}^n)]$$
$$= \frac{1}{2\Delta t}\left[(\nu_{j+\frac{1}{2}}^n)^2(u_{j+1}^n - u_j^n) - (\nu_{j-\frac{1}{2}}^n)^2(u_j^n - u_{j-1}^n)\right].$$

Analogous to the Lax-Friedrichs method there is a stabilizing term on the right hand side which can be interpreted as a discretization of a diffusion term $(\nu^2\,\Delta x)/(2\lambda)\;u_{xx}$.

For a linear flux $f(u) = a\,u$ we have

$$\nu(u,v) = \nu = a\,\lambda,$$

$\nu = \lambda\,a$ is called the Courant number or CFL number (named after Courant, Friedrichs, Lewy).

The Courant-Isaacson-Rees method

Firstly this method will be derived for the linear wave equation

$$u_t + a\,u_x = 0.$$

As it was discussed earlier the solution u is constant along characteristic curves. Hence

$$u_j^{n+1} = u(x_j - a\,\Delta t, t_n)$$

Since, in general $x_j - a\,\Delta t$ is not a grid point the value of $\overline{u}$ at this point is approximated by a linear interpolation of the values at the neighboring grid points:

$$u(x_j - a\,\Delta t, t_n) \approx \begin{cases} \dfrac{a\Delta t}{\Delta x}u(x_{j-1},t_n) + \left(1 - \dfrac{a\Delta t}{\Delta x}\right)u(x_j,t_n) & \text{if } a \geq 0, \\ \left(1 + \dfrac{a\Delta t}{\Delta x}\right)u(x_j,t_n) - \dfrac{a\Delta t}{\Delta x}u(x_{j+1},t_n) & \text{if } a < 0. \end{cases}$$

This leads to the following method

$$u_j^{n+1} = \begin{cases} u_j^n - \lambda a(u_j^n - u_{j-1}^n) & \text{if } a \geq 0, \\ u_j^n - \lambda a(u_{j+1}^n - u_j^n) & \text{if } a < 0. \end{cases}$$

This corresponds to the approximation of $(au)_x$ by a backward difference quotient relative to the direction a.

With the notation

$$a^+ = \max(a,0), \quad a^- = \min(a,0)$$

the Courant-Isaacson-Rees method can be put into the following compact form:

$$u_j^{n+1} = u_j^n - \lambda[a^+(u_j^n - u_{j-1}^n) + a^-(u_{j+1}^n - u_j^n)].$$

It is conversative with the consistent numerical flux

$$g_{CIR}(u,v) = a^+u + a^-v = \frac{a}{2}(u+v) - \frac{|a|}{2}(v-u).$$

It can also be written in the form

$$\frac{1}{\Delta t}(u_j^{n+1} - u_j^n) + \frac{a}{2\Delta x}[u_{j+1}^n - u_{j-1}^n] = \frac{|a|}{2\Delta x}(u_{j+1}^n - 2u_j^n - u_{j-1}^n).$$

The stabilizing term on the right hand side can be interpreted as the discretization of the diffusion term $(\Delta x|a|)/2\ u_{xx}$.

2.4 Godunov's Method

The Courant-Isaacson-Rees method can be easily extended to conservation laws in quasi-linear form by replacing a by $f'(u_j^n)$:

$$u_j^{n+1} = \begin{cases} u_j^n - \lambda f'(u_j^n)(u_j^n - u_{j-1}^n) & \text{if } f'(u_j^n) \geq 0, \\ u_j^n - \lambda f'(u_j^n)(u_{j+1}^n - u_j^n) & \text{if } f'(u_j^n) < 0. \end{cases}$$

Example: This method is applied to Burgers equation, then:

$$u_j^{n+1} = u_j^n - \lambda u_j^n (u_j^n - u_{j-1}^n).$$

For the initial values

$$u_0(x) = \begin{cases} 1 \text{ for } x < 0, \\ 0 \text{ for } x \geq 0 \end{cases}$$

the following approximations are obtained:

$$u_j^0 = \begin{cases} 1 \text{ for } j < 0, \\ 0 \text{ for } j \geq 0, \end{cases}$$

i.e.:

$$u_j^{n+1} = u_j^n.$$

One obtains a stationary solution with the wrong shock speed $s = 0$.

A successful extension to the nonlinear case leads to Godunov's method. In preparation for the construction of this method basic facts about the so-called Riemann problem are needed:

The Riemann problem

Let $u_L, u_R \in \mathbb{R}$ be given constants. The Cauchy problem

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

with initial condition

$$u(x,0) = \begin{cases} u_L & \text{for } x < 0, \\ u_R & \text{for } x > 0 \end{cases}$$

is called a Riemann problem.

For the linear wave equation $u_t + a\, u_x = 0$ one obtains the following weak solution of the Riemann problem:

$$u(x,t) = \begin{cases} u_L & \text{for } x/t < a, \\ u_R & \text{for } x/t > a. \end{cases}$$

In this case we have $F'(u) = U'(u) f'(u) = aU'(u)$, which implies $F(u) = a\,U(u) + b$. Therefore, $[F(u)] = a[U(u)] = s[U(u)]$. So, the entropy condition is also satisfied. Hence u is an entropy solution.

It can be written in the following form:

$$u(x,t) = u^*(\frac{x}{t}; u_L, u_R)$$

with

$$u^*(\xi; u_L, u_R) = \begin{cases} u_L & \text{for } \xi < a, \\ u_R & \text{for } \xi > a. \end{cases}$$

This solution is called a **contact discontinuity**.

Next we consider the non-linear case with $f \in C^2(\mathbb{R})$ with $f'' > 0$. A piecewise constant weak solution to this Riemann problem is given by

$$u(x,t) = \begin{cases} u_L & \text{for } x/t < s, \\ u_R & \text{for } x/t > s \end{cases}$$

with

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

This solution is called a shock wave. It can be written in the following form:

$$u(x,t) = u_S^*(\frac{x}{t}; u_L, u_R)$$

with

$$u_S^*(\xi; u_L, u_R) = \begin{cases} u_L & \text{for } \xi < s, \\ u_R & \text{for } \xi > s. \end{cases}$$

In the case $u_L > u_R$ it is called a **compression shock**, in the case $u_L < u_R$ it is called an **expansion shock**. Similar to Burgers equation it can be shown that only the compression shock satisfies the entropy condition.

In the case $u_L \leq u_R$ a continuous solution of the Riemann problem can be derived with the ansatz

$$u(x,t) = w(\frac{x}{t}).$$

Since

$$u_t = -w'\frac{x}{t^2}, \quad u_x = w'\frac{1}{t},$$

u is a solution to the conservation law if and only if

$$-w'\frac{x}{t^2} + f'(w)\,w'\frac{1}{t} = \frac{w'}{t}\left[f'(w) - \frac{x}{t}\right] = 0.$$

So, u is a solution if $w(\xi)$ solves the algebraic equation

$$f'(w(\xi)) = \xi \quad \text{for all } \xi.$$

From this one obtains the following solution

$$u(x,t) = \begin{cases} u_L & \text{for } \frac{x}{t} \leq f'(u_L), \\ w(\frac{x}{t}) & \text{for } f'(u_L) < \frac{x}{t} < f'(u_R), \\ u_R & \text{for } \frac{x}{t} \geq f'(u_R). \end{cases}$$

It is easy to see that u is continuous: E.g., for $x/t = f'(u_L)$ it follows that

$$f'(w(\frac{x}{t})) = \frac{x}{t} = f'(u_L)).$$

Since f' is strictly monotone, this implies

$$w(\frac{x}{t}) = u_L.$$

This continuous and piecewise smooth solution is, of course, also an entropy solution. It is called a **rarefaction wave**. It can be written in the form:

$$u(x,t) = u_V^*(\frac{x}{t}; u_L, u_R)$$

with

$$u_V^*(\xi; u_L, u_R) = \begin{cases} u_L & \text{for } \xi \leq f'(u_L), \\ w(\xi) & \text{for } f'(u_L) < \xi < f'(u_R), \\ u_R & \text{for } \xi \geq f'(u_R). \end{cases}$$

Summarizing one obtains (relatively easily) an entropy solution of the form $u^*(x/t; u_L, u_R)$ for the Riemann problem, given by

$$u^*(\xi; u_L, u_R) = \begin{cases} u_S^*(\xi; u_L, u_R) & \text{for } u_L > u_R, \\ u_V^*(\xi; u_L, u_R) & \text{for } u_L \leq u_R. \end{cases}$$

So, solving a Riemann problem requires only the solution of an algebraic equation in the case $u_L \leq u_R$.

Godunov's method

The method can be subdivided into three steps:

1. Reconstruction: From the values u_j^n a piecewise constant function is constructed by
$$v(x,t_n) = u_j^n \quad \text{for } x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}).$$

2. Exact solution of the Cauchy problem
$$\begin{aligned} u_t + f(u)_x &= 0 \quad x \in \mathbb{R},\ t > t_n, \\ u(x,t_n) &= v(x,t_n). \end{aligned}$$
Let $v(x,t)$ denote this solution. It can be represented by the solutions of local Riemann problems in the intervals (x_j, x_{j+1}): Since
$$\frac{f(u_j^n) - f(u_{j+1}^n)}{u_j^n - u_{j+1}^n} = f'(\overline{u}_j^n)$$
for values $\overline{u}_j^n$ between u_j^n and u_{j+1}^n, the domains of influence do not overlap as long as
$$|f'(u)| \le \frac{\Delta x/2}{\Delta t} \quad \text{for all } u \in [\min_k u_k^n, \max_k u_k^n],$$
i.e.
$$\frac{\Delta t}{\Delta x} \sup_u |f'(u)| \le \frac{1}{2}.$$
This condition is called the CFL condition (named after Courant, Friedrichs, Lewy). It is a condition on the time step in dependence of the spatial mesh size.

   From this we obtain for $v(x,t)$:
$$v(x,t) = u^* \left(\frac{x - x_{j+\frac{1}{2}}}{t - t_n}; u_j^n, u_{j+1}^n \right) \quad \text{for } x \in (x_j, x_{j+1}),\ t \in [t_n, t_{n+1}].$$

3. Averaging: New values u_j^{n+1} are obtained by averaging:
$$u_j^{n+1} = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x, t_{n+1})\ dx.$$

Godunov's method is conservative: By integrating over the set $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \times [t_n, t_{n+1}]$ one obtains:
$$\begin{aligned} &\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x,t_{n+1})\ dx - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x,t_n)\ dx \\ &\qquad + \int_{t_n}^{t_{n+1}} f(v(x_{j+\frac{1}{2}}, t))\ dt - \int_{t_n}^{t_{n+1}} f(v(x_{j-\frac{1}{2}}, t))\ dt = 0. \end{aligned}$$

From the setting from above and the properties of $v(x,t)$ it follows that:

$$\Delta x(u_j^{n+1} - u_j^n) + \Delta t \left[f(u^*(0; u_j^n, u_{j+1}^n)) - f(u^*(0; u_{j-1}^n, u_j^n)) \right] = 0.$$

Hence Godunov's method is conservative with the consistent numerical flux

$$g_G(u,v) = f(u^*(0; u, v)).$$

Example: For the linear wave equation $u_t + a\, u_x = 0$ one obtains

$$g_G(u,v) = f(u^*(0;u,v)) = \begin{cases} au & \text{for } 0 < a, \\ av & \text{for } 0 > a. \end{cases}$$

So, in the linear case Godunov's method agrees with the Courant-Isaacson-Rees method.

Example: Under the assumptions $f \in C^2(\mathbb{R})$ with $f'' > 0$ one obtains for the general conservation law $u_t + f(u)_x = 0$ from the analysis of the corresponding Riemann problem:

$$g_G(u,v) = \begin{cases} f(u_S^*(0;u,v)) \quad \text{for } u > v & = \begin{cases} f(u) & \text{for } f(u) > f(v) \\ f(v) & \text{for } f(u) \le f(v) \end{cases} \\ \\ f(u_V^*(0;u,v)) \quad \text{for } u \le v & = \begin{cases} f(u) & \text{for } 0 \le f'(u) \\ f(w(0)) & \text{for } f'(u) < 0 < f'(v) \\ f(v) & \text{for } f'(v) \le 0 \end{cases} \end{cases}$$

$$= \begin{cases} f(u) & \text{for } u > v \text{ and } f(u) > f(v), \\ f(v) & \text{for } u > v \text{ and } f(u) \le f(v), \\ f(u) & \text{for } u_s \le u \le v, \\ f(u_s) & \text{for } u < u_s < v, \\ f(v) & \text{for } u \le v \le u_s, \end{cases}$$

where $u_s = w(0)$ is the so-called sonic point, given by the algebraic equation

$$f'(u_s) = 0.$$

2.5 Roe's Approximate Riemann Solver

Godunov's method requires the exact solution of the Riemann problem. The solution of a Riemann problem requires the solution of an algebraic problem. In order to save computing time the exact solution could be replaced by an approximation. One possible way leads to Roe's method:

For approximating the solution of the Riemann problem

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

with initial values

$$u(x,0) = \begin{cases} u_L & \text{for } x < 0, \\ u_R & \text{for } x > 0, \end{cases}$$

the original conservation law is replaced by a conservation law

$$u_t + \hat{f}(u)_x = 0$$

with a linear flux $\hat{f}$ of the form

$$\hat{f}(u) = \hat{a}(u_L, u_R)\, u \ . \tag{2.6}$$

This leads to the following modification of Godunov's method:

1. Reconstruction like in Godunov's method.

2. Exact solution of the Cauchy problem

$$\begin{aligned} u_t + \hat{f}(u)_x &= 0 \quad x \in \mathbb{R},\ t > t_n, \\ u(x, t_n) &= v(x, t_n). \end{aligned}$$

3. Averaging:

$$u_j^{n+1} = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \hat{v}(x, t_{n+1})\ dx.$$

Like in Godunov's method we obtain the representation:

$$\frac{1}{\Delta t}(u_j^{n+1} - u_j^n) + \frac{1}{\Delta x}\left[\hat{f}(\hat{u}^*(0; u_j^n, u_{j+1}^n)) - \hat{f}(\hat{u}^*(0; u_{j-1}^n, u_j^n))\right] = 0.$$

It is easy to see that the method is conservative. However, the numerical flux $\hat{f}(\hat{u}^*(0; u, v))$ is not necessarily consistent.

If the numerical flux is redefined by

$$g_{Roe}(u, v) = \hat{f}(\hat{u}^*(0; u, v)) - \hat{f}(v) + f(v),$$

a consistent flux is obtained. The method from above is recovered if

$$\hat{f}(u_R) - f(u_R) = \hat{f}(u_L) - f(u_L),$$

or, equivalently:

$$\hat{f}(u_R) - \hat{f}(u_L) = f(u_R) - f(u_L).$$

Together with the linearity of the numerical flux (2.6) this leads to the requirement

$$\hat{a}(u_L, u_R) = \frac{f(u_R) - f(u_L)}{u_R - u_L} \quad \text{for } u_L \neq u_R.$$

The solution of the Riemann problems for the Cauchy problem

$$u_t + \hat{f}(u)_x = 0$$

with initial condition

$$u(x,0) = \begin{cases} u_L & \text{for } x < 0, \\ u_R & \text{for } x > 0 \end{cases}$$

is given by

$$\hat{u}^*(\xi; u_L, u_R) = \begin{cases} u_L & \text{for } \xi < \hat{a}(u_L, u_R), \\ u_R & \text{for } \xi > \hat{a}(u_L, u_R). \end{cases}$$

Therefore, we obtain for the numerical flux

$$\begin{aligned} g_{Roe}(u,v) &= \hat{a}(u,v)u^*(0;u,v) - \hat{a}(u,v)v + f(v) \\ &= \begin{cases} f(u) & \text{for } 0 < \hat{a}(u,v), \\ f(v) & \text{for } 0 > \hat{a}(u,v) \end{cases} \\ &= \frac{1}{2}\left[f(u) + f(v)\right] - \frac{1}{2}|\hat{a}(u,v)|(v-u). \end{aligned}$$

This numerical flux agrees with the numerical flux of Godunov's method except for the case

$$u < u_s < v.$$

In order to avoid non-physical solutions in this case (the sonic case) the following modification is usually proposed for g_{Roe}:

$$\widetilde{g}_{Roe}(u,v) = \frac{1}{2}\left[f(u) + f(v)\right] - \frac{1}{2}Q_\delta(\hat{a}(u,v))(v-u)$$

with

$$Q_\delta(x) = \begin{cases} \dfrac{x^2}{2\delta} + \dfrac{\delta}{2} & \text{for } |x| \le \delta, \\ |x| & \text{for } |x| > \delta. \end{cases}$$

2.6 The Enquist-Osher Method

Motivation: The numerical flux of the Courant-Isaacson-Rees method for the linear wave equation

$$u_t + a\,u_x = 0$$

is given by

$$g(u,v) = a^+\,u + a^-\,v.$$

The numerical flux is consistent, since

$$f(w) = a\,w = a^+\,w + a^- w.$$

With

$$f^+(w) = a^+\, w \quad \text{and} \quad f^-(w) = a^-\, w$$

it follows that:

$$f(w) = f^+(w) + f^-(w)$$

and

$$g(u,v) = f^+(u) + f^-(v).$$

A similar splitting of the flux (flux vector splitting) in a positive and a negative part is also possible for non-linear flux functions:

Starting point is the representation

$$\begin{aligned} f(w) &= f(0) + \int_0^w f'(s)\; ds \\ &= f(0) + \int_0^w \left[f'(s)^+ + f'(s)^-\right]\; ds. \end{aligned}$$

With

$$f^+(w) = f(0) + \int_0^w f'(s)^+\; ds \quad \text{and} \quad f^-(w) = \int_0^w f'(s)^-\; ds$$

we obviously obtain

$$f(w) = f^+(w) + f^-(w).$$

The resulting numerical flux leads to the Enquist-Osher method:

$$\begin{aligned} g_{EO}(u,v) &= f^+(u) + f^-(v) \\ &= f(0) + \int_0^u f'(s)^+\; ds + \int_0^v f'(s)^-\; ds \\ &= f(0) + \frac{1}{2}\int_0^u (f'(s) + |f'(s)|)\; ds + \frac{1}{2}\int_0^v (f'(s) - |f'(s)|)\; ds \\ &= f(0) + \frac{1}{2}[f(u) - f(0)] + \frac{1}{2}\int_0^u |f'(s)|\; ds + \frac{1}{2}[f(v) - f(0)] - \frac{1}{2}\int_0^v |f'(s)|\; ds \\ &= \frac{1}{2}\left[f(u) + f(v)\right] - \frac{1}{2}\int_u^v |f'(s)|\; ds. \end{aligned}$$

This representation shows that Roe's method is obtained from the Enquist-Osher method by using an appropriate approximation for the integral term.

For the case $f'' > 0$ the numerical flux of the Enquist-Osher method can be represented in a much simpler way:

Let u_s be the sonic point. Then

$$g_{EO}(u,v) = \begin{cases} f(u) & \text{for } u_s \le u, v, \\ f(v) & \text{for } u, v \le u_s, \\ f(u_s) & \text{for } u < u_s < v, \\ f(u) + f(v) - f(u_s) & \text{for } v < u_s < u. \end{cases}$$

2.7 Convergence Analysis for Smooth Solutions

We restrict ourselves to the Cauchy problem of homogeneous linear differential equations with constant coefficients:

$$u_t = P\left(\frac{\partial}{\partial x}\right)u, \qquad (x,t) \in \mathbb{R} \times (0,T), \tag{2.7}$$

$$u(x,0) = u_0(x), \qquad x \in \mathbb{R}, \tag{2.8}$$

where $P(x)$ is a polynomial with $P(0) = 0$.

Notation: For a sequence $v = (v_j)_{j\in\mathbb{Z}}$ the following shift operators are introduced:

$$S_+ v_j = v_{j+1}, \quad S_- v_j = v_{j-1}$$

With these notations finite difference methods can often be represented in the form:

$$u_j^{n+1} = Q(S_-, S_+)u_j^n \tag{2.9}$$

with an appropriate polynomial $Q(x_-, x_+)$ and initial values u_j^0, $j \in \mathbb{Z}$.

Example: The linear wave equation is of the form

$$u_t = P\left(\frac{\partial}{\partial x}\right)u \quad \text{with } P(x) = -a\,x.$$

The Courant-Isaacson-Rees method for $a > 0$ can be written as:

$$u_j^{n+1} = Q(S_-, S_+)u_j^n \quad \text{with } Q(x_-, x_+) = (1-\lambda a) + \lambda a x_-.$$

A finite difference method of the form (2.9) generates a sequence $v = (v_j)_{j\in\mathbb{Z}}$ of approximations at each time step. In the following the symbol $\|.\|$ denotes a norm in the space of such sequences. An important example is the discrete L^2-norm, given by:

$$\|v\|_{\ell_2(\mathbb{Z})} = \left(\sum_{j\in\mathbb{Z}} |v_j|^2 \Delta x\right)^{\frac{1}{2}}.$$

Let $l_2(\mathbb{Z})$ denote the space of sequences with finite discrete L^2-norm.

Definition 2.7. *Let u be an exact (smooth) solution to (2.7), (2.8).*

1. *Then (τ_j^n), given by*

$$\begin{aligned} \tau_j^{n+1} &= \frac{1}{\Delta t}\left[u(x_j, t_{n+1}) - Q(S_-, S_+)u(x_j, t_n)\right], \quad \textit{for } j \in \mathbb{Z},\ n = 0, 1, \ldots, \\ \tau_j^0 &= u_j^0 - u(x_j, 0), \quad \textit{for } j \in \mathbb{Z}, \end{aligned}$$

is called the local truncation error of the finite difference method (2.9).

2. *The finite difference method is called consistent if and only if*

$$\max_n \|\tau^n\| \to 0 \quad \textit{for } \Delta t,\ \Delta x \to 0.$$

with $\tau^n = (\tau_j^n)_{j\in\mathbb{Z}}$.

3. *The finite difference method is called consistent of the order* (p,q) *if and only if*

$$\max_n \|\tau^n\| = O(\Delta x^p) + O(\Delta t^q).$$

Occasionally the consistency or a certain consistency order cannot be shown for arbitrary step sizes $(\Delta x, \Delta t)$ but only under additionally conditions of the form $(\Delta x, \Delta t) \in U$, where $U \subset (0,\infty)^2$ is a suitable set with accumulation point $(0,0)$.

For smooth solutions the consistency order is typically determined by Taylor expansions.

Example: For the Courant-Isaacson-Rees method for the linear wave equation $u_t + a\,u_x = 0$ with $a > 0$ we have:

$$\begin{aligned}
&\frac{1}{\Delta t}[u(x,t+\Delta t) - u(x,t)] + \frac{a}{\Delta x}[u(x,t) - u(x-\Delta x,t)] = \\
&\quad = \frac{1}{\Delta t}[u + u_t\Delta t + \frac{1}{2}u_{tt}\Delta t^2 + O(\Delta t^3) - u] \\
&\quad\quad - \frac{a}{\Delta x}[u - u + u_x\Delta x - \frac{1}{2}u_{xx}\Delta x^2 - O(\Delta x^3)] \\
&\quad = u_t - au_x + \frac{1}{2}\Delta t u_{tt} - \frac{1}{2}au_{xx}\Delta x + O(\Delta t^2) + O(\Delta x^2) = O(\Delta t) + O(\Delta x)
\end{aligned}$$

From this pointwise analysis of the local truncation error one can easily derive estimates, e.g., in the discrete L^2-norm. So the method is consistent of order $(1,1)$ for smooth solutions u with appropriate integrability conditions.

The ultimate goal is to verify the convergence of the approximate solutions to the exact solutions:

Definition 2.8. *Let u be the exact (smooth) solution to (2.7), (2.8).*

1. *Then* (e_j^n), *given by*

$$e_j^n = u_j^n - u(x_j, t_n)$$

is called the discretization error of the method (2.9).

2. *The finite difference method is called convergent if and only if*

$$\max_n \|e^n\| \to 0 \quad \textit{for } \Delta t,\ \Delta x \to 0$$

with $e^n = (e_j^n)_{j\in\mathbb{Z}}$.

3. *The finite difference method is called convergent of order* (p, q) *if and only if*

$$\max_n \|e^n\| = O(\Delta x^p) + O(\Delta t^q).$$

Similar to the consistency, typically the convergence or the convergence order cannot be shown for arbitrary step sizes $(\Delta x, \Delta t)$ but only under additional assumptions of the form $(\Delta x, \Delta t) \in U$, where $U \subset (0, \infty)^2$ is a suitable set with accumulation point $(0, 0)$.

Consistency does not suffice to prove convergence. Additionally the method has to be stable:

Definition 2.9. *The finite difference method (2.9) is called stable for step sizes* $(\Delta x, \Delta t)$ *from some set* U *if and only if there is a constant* C *such that*

$$\|Q^n\| \leq C$$

for all $(\Delta x,\ \Delta t) \in U$ *and* $n \in \mathbb{N}$.

If the finite difference method is stable, it follows that the approximate solutions are uniformly bounded:

$$\|u^n\| \leq C\,\|u^0\|$$

for all $(\Delta x, \Delta t) \in U$ and n with $n \in \mathbb{N}$.

Theorem 2.3 (Lax)**.** *Assume that the finite difference method is consistent (of order* (p, q)*) and stable. Then the method is convergent (of order* (p, q)*).*

Proof. We have:

$$\begin{aligned} u_j^{n+1} &= Qu_j^n \\ u(x_j, t_{n+1}) &= Qu(x_j, t_n) + \tau_j^{n+1}\Delta t. \end{aligned}$$

By subtraction it follows:

$$e^{n+1} = Qe^n - \tau^{n+1}\Delta t.$$

From this recursion one obtains:

$$e^n = Q^n\tau^0 - \Delta t \sum_{m=0}^{n} Q^m \tau^{n-m}.$$

The stability ensures

$$\|Q^n\| \leq C.$$

and, therefore,

$$\|e^n\| \leq C(1 + n\Delta t)C \max_{m \leq n} \|\tau^m\| \leq C(1 + T) \max_{m \leq n} \|\tau^m\|,$$

from which the rest follows immediately. □

For the discrete L^2-norm the stability is usually analyzed with the help of the Fourier transform:

Theorem 2.4. *The finite difference method is stable with respect the discrete L^2-norm if and only if*

$$|\rho(\xi)| \leq 1 \quad \textit{for all } \xi \in [0, 2\pi),$$

where

$$\rho(\xi) = Q(e^{-i\xi}, e^{i\xi})$$

is the so-called symbol (or amplification factor) of the finite difference method.

Proof. The Fourier transform $\mathcal{F}\colon l_2(\mathbb{Z}) \to L^2(0, 2\pi)$ is given by

$$\mathcal{F}u(\xi) = \hat{u}(\xi) = \frac{\Delta x}{\sqrt{2\pi}} \sum_{j \in \mathbb{Z}} u_j e^{-ij\xi}, \quad \xi \in [0, 2\pi).$$

The inverse of $\mathcal{F}$ is given by

$$u_j = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \hat{u}(\xi) e^{ij\xi} \, d\xi, \quad j \in \mathbb{Z}.$$

Parseval's identity holds:

$$\|u\|_{\ell_2(\mathbb{Z})} = \|\hat{u}\|_{L^2(0,2\pi)}$$

with

$$\|\hat{u}\|_{L^2(0,2\pi)} = \left(\int_0^{2\pi} |\hat{u}(\xi)|^2 \, d\xi \right)^{\frac{1}{2}}.$$

One easily see that

$$\widehat{S_+ u}(\xi) = e^{i\xi} \, \hat{u}(\xi), \qquad \widehat{S_- u}(\xi) = e^{-i\xi} \, \hat{u}(\xi).$$

Hence

$$\mathcal{F}[Q(S_-, S_+)u](\xi) = Q(e^{-i\xi}, e^{i\xi}) \, \hat{u}(\xi) = \rho(\xi) \, \hat{u}(\xi).$$

Using Parseval's identity one obtains

$$\begin{aligned} \|Q^n\|_{\ell_2(\mathbb{Z})} &= \sup_{u \neq 0} \frac{\|Q^n u\|_{\ell_2(\mathbb{Z})}}{\|u\|_{\ell_2(\mathbb{Z})}} = \sup_{u \neq 0} \frac{\|\widehat{Q^n u}\|_{L^2(0,2\pi)}}{\|\hat{u}\|_{L^2(0,2\pi)}} \\ &= \sup_{\hat{u} \neq 0} \frac{\left(\int_0^{2\pi} |\rho(\xi)^n \hat{u}(\xi)|^2 \, d\xi \right)^{\frac{1}{2}}}{\left(\int_0^{2\pi} |\hat{u}(\xi)|^2 \, d\xi \right)^{\frac{1}{2}}} = \max_{\xi \in [0, 2\pi)} |\rho(\xi)|^n, \end{aligned}$$

which completely the proof. □

Example: For the Courant-Isaacson-Rees method applied to the linear wave equation $u_t + a\,u_x = 0$ one obtains the following amplification factor:

$$\rho(\xi) = 1 - \lambda a + \lambda a e^{-i\xi} = 1 - \lambda a + \lambda a \cos(\xi) - i\lambda a \sin(\xi).$$

Therefore:

$$\begin{aligned} |\rho(\xi)|^2 &= [(1-\lambda a) + \lambda a \cos\xi]^2 + [\lambda a \sin\xi]^2 \\ &= (1-\lambda a)^2 + 2(1-\lambda a)\lambda a \cos\xi + \lambda^2 a^2 \\ &= 1 - 2(1-\lambda a)\lambda a (1-\cos\xi). \end{aligned}$$

The stability condition

$$|\rho(\xi)|^2 \leq 1, \quad \text{for all } \xi \in [0, 2\pi)$$

is obviously satisfied if and only if

$$a\lambda = a\frac{\Delta t}{\Delta x} \leq 1.$$

Observe that $a\lambda$ is the so-called CFL number (or Courant number).

Remark: All methods discussed so far for the linear wave equation $u_t + a\,u_x = 0$ are of the form

$$u_j^{n+1} = H(u_{j-1}^n, u_j^n, u_{j+1}^n).$$

A necessary condition for the convergence of such a method is the so-called CFL condition: The domain of dependency of the finite difference method must include the domain of dependency of the differential equation. The domain of dependency of the finite difference method in a grid point (x_j, t_n) is that interval of all grid points at initial time which the value of the approximate solution at (x_j, t_n) (at (x, t)) depends on. The domain of dependency of the differential equation in a point (x, t) is that point at initial time which the value of the solution at (x, t) depends on.

This minimal requirement on a finite difference method already leads to the necessary condition

$$a\frac{\Delta t}{\Delta x} \leq 1.$$

The analysis from above shows that this condition is also sufficient for the Courant-Isaacson-Rees method. A method for which the CFL condition is sufficient for stability is called optimally stable.

Remark: The presented convergence analysis can easily be extended to inhomogeneous problems and implicit methods.

So far only the case of linear differential equations with constant coefficients were discussed, which was particularly helpful for studying the stability by Fourier analysis. For sufficiently smooth solutions the convergence analysis can be extended to

- linear differential equations with variable coefficients (by localization),
- non-linear differential equations (by linearization).

2.8 Convergence Analysis for Weak Solutions

For solving the Cauchy problem

$$\begin{aligned} u_t + f(u)_x &= 0, \qquad (x,t) \in \mathbb{R} \times (0,\infty), \\ u(x,0) &= u_0(x), \qquad x \in \mathbb{R} \end{aligned}$$

with flux $f \in C(\mathbb{R})$ we consider conservative methods

$$u_j^{n+1} = u_j^n - \lambda(g_{j+\frac{1}{2}}^n - g_{j-\frac{1}{2}}^n),$$

where

$$g_{j+\frac{1}{2}} = g(u_j^n, u_{j+1}^n)$$

with a consistent numerical flux $g \in C(\mathbb{R} \times \mathbb{R})$

$$g(u,u) = f(u).$$

The convergence analysis consists of two parts: First it will be shown that the approximate solutions converge towards a weak solution, if they converge at all. Then it will be shown that there exists at least a convergent sub-sequence.

Let (Δx_m) and (Δt_m) be sequences of step sizes. In the following $u_m\colon \mathbb{R} \times [0,\infty) \to \mathbb{R}$ denotes the piecewise constant reconstruction from the values u_j^n obtained by the conservative method with step sizes Δx_m and Δt_m:

$$u_m(x,t) = u_j^n, \quad \text{for } x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}),\ t \in [t_n, t_{n+1}).$$

For the initial approximations the average values of the initial data are chosen:

$$u_j^0 = \frac{1}{\Delta x_m} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x)\ dx.$$

The following theorem on the consistency of conservative methods is of central importance:

Theorem 2.5 (Lax-Wendroff). *Let $u_0 \in L^\infty(\mathbb{R})$ and let (Δx_m) and (Δt_m) be sequences of step sizes approaching 0. Assume that the sequence (u_m) of approximate solutions satisfy the following conditions: There is a constant C with*

$$\|u_m\|_{L^\infty(\mathbb{R}\times(0,\infty))} \le C \quad \textit{for all } m$$

and

$$u_m(x,t) \to u(x,t) \quad \textit{almost everywhere (in short, a.e.) in } \mathbb{R} \times (0,\infty).$$

Then u is a weak solution of the Cauchy problem.

Proof. To simplify the notation the index m will be omitted in Δx_m, Δt_m.

Let $\varphi \in C_0^\infty(\mathbb{R} \times [0,\infty))$ and set

$$\varphi_j^n = \frac{1}{\Delta t}\frac{1}{\Delta x}\int_{t_n}^{t_{n+1}}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \varphi(x,t)\; dx\; dt.$$

By multiplying with $\varphi_j^n \Delta x$ and adding over j and n one obtains for a conservative method:

$$\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty}\left[(u_j^{n+1}-u_j^n)+\lambda(g_{j+\frac{1}{2}}^n - g_{j-\frac{1}{2}}^n)\right]\varphi_j^n \Delta x = 0.$$

By summation by parts it follows

$$\sum_{n=0}^{\infty}(u_j^{n+1}-u_j^n)\varphi_j^n = -\sum_{n=1}^{\infty} u_j^n(\varphi_j^n - \varphi_j^{n-1}) - u_j^0\varphi_j^0$$

and

$$\sum_{j=-\infty}^{\infty}(g_{j+\frac{1}{2}}^n - g_{j-\frac{1}{2}}^n)\varphi_j^n = -\sum_{j=-\infty}^{\infty} g_{j+\frac{1}{2}}^n(\varphi_{j+1}^n - \varphi_j^n).$$

It remains to show:

$$\sum_{n=1}^{\infty}\sum_{j=-\infty}^{\infty} u_j^n(\varphi_j^n - \varphi_j^{n-1})\Delta x \longrightarrow \int_0^\infty\int_{-\infty}^\infty u\,\varphi_t\; dx\; dt \tag{2.10}$$

$$\sum_{j=-\infty}^{\infty} u_j^0\varphi_j^0\Delta x \longrightarrow \int_{-\infty}^\infty u_0(x)\,\varphi(x,0)\; dx \tag{2.11}$$

$$\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty} g_{j+\frac{1}{2}}^n(\varphi_{j+1}^n - \varphi_j^n)\Delta t \longrightarrow \int_0^\infty\int_{-\infty}^\infty f(u)\,\varphi_x\; dx\; dt \tag{2.12}$$

Proof of (2.10):

$$\sum_{n=1}^{\infty}\sum_{j=-\infty}^{\infty} u_j^n(\varphi_j^n - \varphi_j^{n-1})\Delta x = \sum_{n=1}^{\infty}\sum_{j=-\infty}^{\infty} u_j^n\frac{\varphi_j^n - \varphi_j^{n-1}}{\Delta t}\Delta x\Delta t = \int_{\Delta t_m}^\infty\int_{-\infty}^\infty u_m(\Delta_t\varphi)_m\; dx\; dt$$

with the piecewise constant function $(\Delta_t\varphi)_m$, given by

$$(\Delta_t\varphi)_m = \frac{\varphi_j^n - \varphi_j^{n-1}}{\Delta t} \quad \text{for } x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}),\ t \in [t_n, t_{n+1}).$$

From the assumptions on the sequence u_m and the smoothness of φ it follows:

$$u_m(x,t)\,(\Delta_t\varphi)_m(x,t) \longrightarrow u(x,t)\,\varphi_t(x,t) \quad \text{a.e. in } \mathbb{R}\times(0,\infty)$$

and

$$\|u_m\,(\Delta_t\varphi)_m\|_{L^\infty(\mathbb{R}\times(0,\infty))} \leq C\|\varphi_t\|_{L^\infty(\mathbb{R}\times(0,\infty))}.$$

Therefore, it follows

$$\int_0^\infty \int_{-\infty}^\infty u_m(\Delta_t\varphi)_m \; dx \; dt \longrightarrow \int_0^\infty \int_{-\infty}^\infty u\,\varphi_t \; dx \; dt$$

by Lebesgue's Theorem.

Since u_m and $(\Delta_t\varphi)_m$ are bounded and $(\Delta_t\varphi)_m$ has a compact support, we obtain

$$\int_0^{\Delta t_m} \int_{-\infty}^\infty u_m(\Delta_t\varphi)_m \; dx \; dt \to 0,$$

which implies (2.10).
The proof of (2.11) is completely analogous and is, therefore, omitted.
Proof of (2.12):

$$\begin{aligned}&\sum_{n=0}^\infty \sum_{j=-\infty}^\infty g^n_{j+\frac{1}{2}}(\varphi^n_{j+1}-\varphi^n_j)\Delta t\\ &\quad=\sum_{n=0}^\infty \sum_{j=-\infty}^\infty g^n_{j+\frac{1}{2}}\frac{\varphi^n_{j+1}-\varphi^n_j}{\Delta x}\Delta x\Delta t = \int_0^\infty \int_{-\infty}^\infty g_m\,(\Delta_x\varphi)_m \; dx \; dt\end{aligned}$$

with the piecewise constant functions g_m and $(\Delta_t\varphi)_m$, given by

$$\begin{aligned} g_m(x,t) &= g^n_{j+\frac{1}{2}} \quad \text{for } x\in(x_{j-\frac{1}{2}},x_{j+\frac{1}{2}}),\ t\in[t_n,t_{n+1})\\ &= g(u^n_j,u^n_{j+1}) = g(u_m(x,t),u_m(x+\Delta x_m,t))\end{aligned}$$

and

$$(\Delta_x\varphi)_m = \frac{\varphi^n_j-\varphi^{n-1}_j}{\Delta x} \quad \text{for } x\in(x_{j-\frac{1}{2}},x_{j+\frac{1}{2}}),\ t\in[t_n,t_{n+1}).$$

For each compact set $K\subset\mathbb{R}\times[0,\infty)$ we have

$$u_m(.+\Delta x_m,.)\to u \quad \text{in } L^1(K).$$

(This follows from the estimate

$$\begin{aligned}&\int_K |u_m(x+\Delta x_m,t)-u(x,t)| \; dx \; dt \leq\\ &\quad\leq \int_K |u_m(x+\Delta x_m,t)-u(x+\Delta x_m,t)| \; dx \; dt + \int_K |u(x+\Delta x_m,t)-u(x,t)| \; dx \; dt\end{aligned}$$

The second term on the right hand side converges towards 0 because of the continuity in mean of functions in L^1. For the first term on the right hand side we have:

$$\begin{aligned}\int_K &|u_m(x+\Delta x_m,t)-u(x+\Delta x_m,t)|\;dx\;dt = \\ &= \int_{\Delta x_m+K} |u_m(x,t)-u(x,t)|\;dx\;dt \\ &\le \int_K |u_m(x,t)-u(x,t)|\;dx\;dt + \int_{(\Delta x_m+K)-K} |u_m(x,t)-u(x,t)|\;dx\;dt\end{aligned}$$

The first integral on the right hand side converges towards 0 because of Lebesgue's Theorem, the second integral converges towards 0 since the function is bounded and the measure of the domain of integration approaches 0.)

The $L^1(K)$-convergence implies the existence of a sub-sequence which converges a.e. in K:

$$u_{m_l}(x+\Delta x_m,t)\to u(x,t) \quad \text{a.e. in } \mathbb{R}\times(0,\infty).$$

Since g is continuous it follows that

$$g_{m_l}(x,t)=g(u_{m_l}(x,t),u_{m_l}(x+\Delta x_{m_l},t))\to g(u(x,t),u(x,t))=f(u(x,t)) \\ \text{a.e. in } \mathbb{R}\times(0,\infty).$$

Together with the smoothness of φ it follows that

$$g_{m_l}(x,t)\,(\Delta_x\varphi)_x(x,t)\to f(u(x,t))\,\varphi_x(x,t) \quad \text{a.e. in } \mathbb{R}\times(0,\infty)$$

and

$$\|g_{m_l}\,(\Delta_x\varphi)_x\|_{L^\infty}\le \max_{|u|,|v|\le C}|g(u,v)|\;\|\varphi_x\|_{L^\infty(\mathbb{R})}$$

This implies (2.12) by using Lebesgue's Theorem. □

The Theorem of Lax-Wendroff is a statement on the consistency of the method. In order to obtain convergence a stability result is required which guarantees the existence of at least a convergent sub-sequence. This is true if the approximate solutions are contained in a compact set. It is reasonable to study this question in the set $L^1_{loc}(\mathbb{R}\times(0,\infty))$ of locally integrable functions on $\mathbb{R}\times(0,\infty)$.

The Theorem of Kolmogorov gives a characterization of compact sets in $L^1(K)$, if $K\subset\mathbb{R}^d$ is a compact set:

Theorem 2.6. *A subset $\mathcal{M}\subset L^1(K)$ is pre-compact if and only if there is a constant C with*

$$\|v\|_{L^1(K)}\le C \quad \textit{for all } v\in\mathcal{M},$$

and, for each $\varepsilon>0$, there is a $\delta>0$ with

$$\int_{K_h}|v(x+h)-v(x)|\;dx\le\varepsilon \quad \textit{for all } h,\ \|h\|\le\delta,\ v\in\mathcal{M},$$

where $K_h=\{x\in K \Big|\ [x,x+h]\subset K\}$.

From the Theorem of Kolmogorov one easily obtains a characterization of compact sets in $L^1_{loc}(\Omega)$ for open sets $\Omega \subset \mathbb{R}^d$: A subset $\mathcal{M} \subset L^1_{loc}(\Omega)$ is pre-compact if and only if

$$\mathcal{M}\Big|_K = \{v\Big|_K : v \in \mathcal{M}\}$$

is pre-compact in $L^1(K)$ for all compact subset $K \subset \Omega$.

For verifying the conditions from above the concept of total variation is needed:

Definition 2.10. 1. *For a function $v\colon [a,b] \to \mathbb{R}$ the total variation of v is given by:*

$$TV_{[a,b]}(v) = \sup_{a=y_0<\ldots<y_l=b} \sum_{k=0}^{l-1} |v(y_{k+1}) - v(y_k)|.$$

If $TV_{[a,b]}(v)$ is finite, the function is said be to of bounded variation. The set of all functions on $[a,b]$ of bounded variation is denoted by $BV([a,b])$.

2. *For a function $v\colon \mathbb{R} \longrightarrow \mathbb{R}$ the total variation is given by*

$$TV(v) = \sup_{-\infty<y_0<\ldots<y_l<\infty} \sum_{k=0}^{l-1} |v(y_{k+1}) - v(y_k)|.$$

If $TV(v)$ is finite, the function is said be to of bounded variation. The set of all functions on $\mathbb{R}$ of bounded variation is denoted by $BV(\mathbb{R})$.

In the following a few properties of the total variation are summarized:

1. If the function $v\colon [a,b] \to \mathbb{R}$ is monotone, then:

$$TV_{[a,b]}(v) = |v(b) - v(a)|$$

2. If the function $v\colon \mathbb{R} \to \mathbb{R}$ is the piecewise constant reconstruction from the sequence $(v_j)_{j\in\mathbb{Z}}$

$$v(x) = v_j \quad \text{for all } x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}),$$

then

$$TV(v) = \sum_{j=-\infty}^{\infty} |v_{j+1} - v_j|.$$

3. If $v \in C^1([a,b])$ or $v \in C^1(\mathbb{R})$, respectively, then

$$TV_{[a,b]}(v) = \int_a^b |v'(x)|\, dx \quad \text{or} \quad TV(v) = \int_{-\infty}^{\infty} |v'(x)|\, dx,$$

respectively.

4. If the function $v\colon \mathbb{R} \to \mathbb{R}$ is measurable, then
$$TV(v) = \sup_h \frac{1}{|h|} \int_{-\infty}^{\infty} |v(x+h) - v(x)| \, dx.$$

The following stability statement holds:

Theorem 2.7 (TV-stability)**.** *Let (Δx_m) and (Δt_m) be sequences of step sizes approaching 0. Assume that, for the sequence (u_m) of approximate solutions, there exist constants C_1 and C_2 with*
$$\|u_m\|_{L^\infty(\mathbb{R}\times(0,\infty))} \le C_1 \quad \textit{for all } m$$
and
$$TV(u_m(.,t_n)) \le C_2 \quad \textit{for all } m, n.$$
Furthermore, assume that the numerical flux g satisfies the Lipschitz condition
$$|g(v_1,v_2) - g(w_1,w_2)| \le L(|v_1 - w_1| + |v_2 - w_2|) \quad \textit{for all } v_i, w_i \textit{ with } |v_i|, |w_i| \le C_1.$$
Then there exist a function $u \in L^\infty(\mathbb{R}\times(0,\infty))$ and a subsequence (u_{m_l}) with
$$u_{m_l} \to u \quad \textit{in } L^1_{loc}(\mathbb{R}\times(0,\infty)).$$

Proof. It suffices to show that the sequence (u_m) is contained in a compact subset of $L^1_{loc}(\mathbb{R}\times(0,\infty))$. For each compact subset $K \subset \mathbb{R}\times(0,\infty)$, the two conditions of the Theorem of Kolmogorov must be verified. The first condition is trivially satisfied.

Let $K \subset \mathbb{R}\times[0,T]$ and $h = (h_x, h_t)$. Then
$$\begin{aligned}
&\int_{K_h} |u_m(x+h_x, t+h_t) - u_m(x,t)| \, dx \, dt \le \\
&\quad \le \int_{K_h} |u_m(x+h_x, t+h_t) - u_m(x+h_x,t)| \, dx \, dt + \int_{K_h} |u_m(x+h_x,t) - u_m(x,t)| \, dx \, dt \\
&\quad \le \int_{[0,T]_{h_t}} \int_{\mathbb{R}} |u_m(x, t+h_t) - u_m(x,t)| \, dx \, dt + \int_{[0,T]} \int_{\mathbb{R}} |u_m(x+h_x,t) - u_m(x,t)| \, dx \, dt \\
&\quad = \sum_{j=-\infty}^{\infty} \int_{[0,T]_{h_t}} |u_m(x_j, t+h_t) - u_m(x_j,t)| \, dt \, \Delta x \\
&\qquad + \sum_n \int_{\mathbb{R}} |u_m(x+h_x, t_n) - u_m(x,t_n)| \, dx \, \Delta t.
\end{aligned}$$

If $N_t \in \mathbb{N}$ is chosen such that $|h_t|/N_t < \Delta t$, then we obtain from the triangle inequality:
$$\begin{aligned}
&\int_{[0,T]_{h_t}} |u_m(x_j, t+h_t) - u_m(x_j,t)| \, dt \le \\
&\quad \le N_t \int_{[0,T]_{h_t/N_t}} |u_m(x_j, t + \frac{h_t}{N_t}) - u_m(x_j,t)| \, dt \\
&\quad \le N_t \sum_n \frac{|h_t|}{N_t} |u_j^{n+1} - u_j^n| = |h_t| \sum_n |u_j^{n+1} - u_j^n|.
\end{aligned}$$

If $N_x \in \mathbb{N}$ is chosen such that $|h_x|/N_x < \Delta x$, then we obtain from the triangle inequality:

$$\begin{aligned}
&\int_{\mathbb{R}} |u_m(x+h_x,t_n) - u_m(x,t_n)| \; dx \leq \\
&\quad \leq N_x \int_{\mathbb{R}} |u_m(x+\frac{h_x}{N_x},t_n) - u_m(x,t_n)| \; dx \\
&\quad \leq N_x \sum_{j=-\infty}^{\infty} \frac{|h_x|}{N_x} |u_{j+1}^n - u_j^n| = |h_x| \sum_{j=-\infty}^{\infty} |u_{j+1}^n - u_j^n|.
\end{aligned}$$

Therefore, the following estimate holds:

$$\begin{aligned}
&\int_{K_h} |u_m(x+h_x,t+h_t) - u_m(x,t)| \; dx \; dt \\
&\quad \leq |h_t|\Delta x \sum_n \sum_{j=-\infty}^{\infty} |u_j^{n+1} - u_j^n| + |h_x|\Delta t \sum_n \sum_{j=-\infty}^{\infty} |u_{j+1}^n - u_j^n|.
\end{aligned}$$

In order to estimate the first term on the right hand side, the Lipschitz condition for g is used:

$$\Delta x |u_j^{n+1} - u_j^n| = \Delta t |g(u_j^n, u_{j+1}^n) - g(u_{j-1}^n, u_j^n)| \leq \Delta t L \left[|u_j^n - u_{j-1}^n| + |u_{j+1}^n - u_j^n)| \right].$$

In summary we obtain

$$\begin{aligned}
&\int_{K_h} |u_m(x+h_x,t+h_t) - u_m(x,t)| \; dx \; dt \\
&\quad \leq |h_t| \, \Delta t \sum_n 2L \; TV(u_m(.,t_n)) + |h_x| \, \Delta t \sum_n TV(u_m(.,t_n)) \leq 2T \, L \; C_2 \, |h_t| + T \, C_2 \, |h_x|,
\end{aligned}$$

which implies the second condition of the Theorem of Kolmogorov. □

So, finally, the following convergence result holds:

Theorem 2.8. *Assume the conditions of Theorem 2.7. Then there exists a subsequence* (u_{m_l}) *with*

$$u_{m_l} \to u \quad \text{in } L^1_{loc}(\mathbb{R} \times (0,\infty)),$$

and $u \in L^\infty(\mathbb{R} \times (0,\infty))$ *is a weak solution.*

Remark: If the uniqueness of a weak solution is guaranteed, then the convergence of the whole sequence of approximate solutions towards the unique weak solution follows under the assumptions of Theorem 2.7.

Remark: The statements from above can be easily extended to entropy solutions: An entropy solution additionally satisfies the weak form of the entropy inequality:

$$U(u)_t + F(u)_x \leq 0,$$

where (U, F) is an entropy pair. If the approximate solutions satisfy a discrete entropy inequality of the form

$$U(u_j^{n+1}) - U(u_j^n) + \lambda \left[G_{j+\frac{1}{2}}^n - G_{j-\frac{1}{2}}^n \right] \leq 0,$$

where

$$G_{j+\frac{1}{2}}^n = G(u_j^n, u_j^{n+1})$$

with a numerical entropy flux G, which satisfies the consistency condition

$$G(u, u) = F(u),$$

then the limit u additionally satisfies the entropy inequality. The proof is completely analogous to the proof of the Theorem of Lax-Wendroff.

As already stated, for $f \in C^1(\mathbb{R})$ and $u_0 \in L^\infty(\mathbb{R})$, there exists a unique entropy solution of the Cauchy problem

$$\begin{aligned} u_t + f(u)_x &= 0, \quad x \in \mathbb{R},\ t > 0 \\ u(x, 0) &= u_0(x), \quad x \in \mathbb{R}. \end{aligned}$$

Moreover, the following properties can be shown:
If u and v are the entropy solutions associated to the initial values u_0 and v_0, respectively, then we have

$$\text{if } u_0(x) \geq v_0(x) \text{ a.e. in } \mathbb{R}, \text{ then } u(t) \geq v(t) \text{ a.e. in } \mathbb{R}.$$

If $u_0 \in BV(\mathbb{R})$, then $u(., t) \in BV(\mathbb{R})$ and

$$\text{if } t \geq s, \text{ then } TV(u(., t)) \leq TV(u(., s)).$$

These properties motivate the next two classes of methods.

2.9 Monotone Methods

A conservative method with numerical flux g is of the form

$$u_j^{n+1} = H(u_{j-l}^n, \dots, u_{j+l}^n)$$

with

$$H(u_{j-l}^n, \dots, u_{j+l}^n) = u_j^n - \lambda \left[g(u_{j-l+1}^n, \dots, u_{j+l}^n) - g(u_{j-l}^n, \dots, u_{j+l-1}^n) \right]$$

and $\lambda = \Delta t / \Delta x$.

Definition 2.11. *A method of the form*

$$u_j^{n+1} = H(u_{j-l}^n, \dots, u_{j+l}^n)$$

is called monotone if and only if H is monotonically increasing with respect to each argument.

Example: For the Lax-Friedrichs method we have:

$$H(u_{j-1}, u_j, u_{j+1}) = \frac{1}{2}(u_{j-1} + u_{j+1}) - \frac{\lambda}{2}[f(u_{j+1}) - f(u_{j-1})].$$

If the CFL condition

$$\lambda \sup_u |f'(u)| \leq 1$$

is satisfied then

$$\begin{aligned} \frac{\partial H}{\partial u_{j-1}} &= \frac{1}{2}\left[1 + \lambda f'(u_{j-1})\right] \geq 0, \\ \frac{\partial H}{\partial u_j} &= 0, \\ \frac{\partial H}{\partial u_{j+1}} &= \frac{1}{2}\left[1 - \lambda f'(u_{j+1})\right] \geq 0. \end{aligned}$$

Hence, the Lax-Friedrichs method is monotone if the CFL condition mentioned above is satisfied.

Godunov's method is monotone (under a suitable CFL condition) since each of the three steps are monotone.

The following theorem can be shown

Theorem 2.9. *For consistent, conservative and monotone methods we have:*

1. $\|u^{n+1}\|_{\ell^\infty} \leq \|u^n\|_{\ell^\infty}$ *for all* $n \geq 0$.
2. $TV(u^{n+1}) \leq TV(u^n)$ *for all* $n \geq 0$.
3. *A discrete entropy inequality is satisfied for the entropy pairs* $(|u-k|, \operatorname{sign}(u-k)(f(u)-f(k)))$, $k \in \mathbb{R}$.

This implies immediately:

Theorem 2.10. *For monotone, consistent and conservative methods we have: Let* $u_0 \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$ *and let* (Δt_m), (Δx_m) *be sequences of step sizes approaching 0. Then there exists a sub-sequence* (u_{m_l}) *with*

$$u_{m_l} \to u \quad \text{in } L^1_{loc}(\mathbb{R} \times (0, \infty))$$

and u *satisfies the entropy condition for all entropy pairs* $(|u-k|, \operatorname{sign}(u-k)(f(u)-f(k))$, $k \in \mathbb{R}$.

Unfortunately we also have

Theorem 2.11. *A monotone, consistent and conservative method is at most of order 1 (except for trivial cases).*

2.10 TVD Methods

Definition 2.12. *A method is called TVD (total variation diminishing), if and only if*

$$TV(u^{n+1}) \leq TV(u^n) \quad \textit{for all } n \geq 0.$$

The following theorem provides a simple criterion for being TVD.

Theorem 2.12. *A method of the form*

$$u_j^{n+1} = u_j^n - C_{j-\frac{1}{2}}^n (u_j^n - u_{j-1}^n) + D_{j+\frac{1}{2}}^n (u_{j+1}^n - u_j^n)$$

is TVD, if

$$C_{j+\frac{1}{2}}^n \geq 0, \; D_{j+\frac{1}{2}}^n \geq 0, \; C_{j+\frac{1}{2}}^n + D_{j+\frac{1}{2}}^n \leq 1.$$

Proof. By subtraction one obtains

$$\begin{aligned} u_{j+1}^{n+1} - u_j^{n+1} &= u_{j+1}^n - C_{j+\frac{1}{2}}^n (u_{j+1}^n - u_j^n) + D_{j+\frac{3}{2}}^n (u_{j+2}^n - u_{j+1}^n) \\ &\quad - u_j^n + C_{j-\frac{1}{2}}^n (u_j^n - u_{j-1}^n) - D_{j+\frac{1}{2}}^n (u_{j+1}^n - u_j^n) \\ &= (1 - C_{j+\frac{1}{2}}^n - D_{j+\frac{1}{2}}^n)(u_{j+1}^n - u_j^n) + C_{j-\frac{1}{2}}^n (u_j^n - u_{j-1}^n) + D_{j+\frac{3}{2}}^n (u_{j+2}^n - u_{j+1}^n). \end{aligned}$$

This implies

$$|u_{j+1}^{n+1} - u_j^{n+1}| \leq (1 - C_{j+\frac{1}{2}}^n - D_{j+\frac{1}{2}}^n)|u_{j+1}^n - u_j^n| + C_{j-\frac{1}{2}}^n |u_j^n - u_{j-1}^n| + D_{j+\frac{3}{2}}^n |u_{j+2}^n - u_{j+1}^n|$$

and, therefore, we obtain by summation

$$\begin{aligned} TV(u^{n+1}) &\leq \sum_j (1 - C_{j+\frac{1}{2}}^n - D_{j+\frac{1}{2}}^n)|u_{j+1}^n - u_j^n| \\ &\quad + \sum_j C_{j-\frac{1}{2}}^n |u_j^n - u_{j-1}^n| + \sum_j D_{j+\frac{3}{2}}^n |u_{j+2}^n - u_{j+1}^n| \\ &= \sum_j (1 - C_{j+\frac{1}{2}}^n - D_{j+\frac{1}{2}}^n)|u_{j+1}^n - u_j^n| \\ &\quad + \sum_j C_{j+\frac{1}{2}}^n |u_{j+1}^n - u_j^n| + \sum_j D_{j+\frac{1}{2}}^n |u_{j+1}^n - u_j^n| \\ &= \sum_j |u_{j+1}^n - u_j^n| = TV(u^n). \end{aligned}$$

□

Remark: The coefficients $C_{j-\frac{1}{2}}^n$ and $D_{j+\frac{1}{2}}^n$ may depend on the approximate solution, e.g.:

$$C_{j-\frac{1}{2}}^n = C(\ldots, u_{j-1}^n, u_j^n, \ldots), \quad D_{j+\frac{1}{2}}^n = D(\ldots, u_j^n, u_{j+1}^n, \ldots).$$

Flux Limiters

For the example of the linear wave equation

$$u_t + a\,u_x = 0 \quad \text{with } a > 0$$

a technique is introduced how to combine a low-order stable method with a higher-order non-TVD method in order to produce a TVD method of higher order.

Starting point is the Lax-Wendroff method, a method of consistency order $(2,2)$:

$$u_j^{n+1} = u_j^n - \frac{\nu}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{\nu^2}{2}\left[(u_{j+1}^n - u_j^n) - (u_j^n - u_{j-1}^n)\right].$$

The second term is interpreted as a diffusion term which stabilizes the method. The next representation allows a different interpretation:

$$u_j^{n+1} = u_j^n - \nu(u_j^n - u_{j-1}^n) - \frac{(1-\nu)\nu}{2}\left[(u_{j+1}^n - u_j^n) - (u_j^n - u_{j-1}^n)\right].$$

This time the second term is interpreted as an anti-diffusion term which reduces the high diffusive behavior of the Courant-Isaacson-Rees method. Numerical experiments reveal a oscillatory behavior in the neighborhood of discontinuities, which shows that in certain situations the anti-diffusive term is too high. Therefore, it is recommended to control the anti-diffusive term by introducing suitable parameters:

$$u_j^{n+1} = u_j^n - \nu(u_j^n - u_{j-1}^n) - \frac{(1-\nu)\nu}{2}\left[\phi_j \cdot (u_{j+1}^n - u_j^n) - \phi_{j-1} \cdot (u_j^n - u_{j-1}^n)\right].$$

This corresponds to a certain combination of the numerical flux g^L of the Courant-Isaacson-Rees method and the numerical flux g^H of the Lax-Wendroff method:

$$g_{j+\frac{1}{2}} = g_{j+\frac{1}{2}}^L + \phi_j\,(g_{j+\frac{1}{2}}^H - g_{j+\frac{1}{2}}^L) = (1-\phi_j)\,g_{j+\frac{1}{2}}^L + \phi_j\,g_{j+\frac{1}{2}}^H.$$

By setting

$$\phi_j = \phi(r_j) \quad \text{with } r_j = \frac{u_j - u_{j-1}}{u_{j+1} - u_j}$$

we obtain the following method

$$\begin{aligned} u_j^{n+1} &= u_j^n - \nu(u_j^n - u_{j-1}^n) - \\ &\quad - \frac{(1-\nu)\nu}{2}\left[\phi(r_j)(u_{j+1}^n - u_j^n) - \phi(r_{j-1})(u_j^n - u_{j-1}^n)\right] \\ &= u_j^n - \nu\left\{1 + \frac{1}{2}(1-\nu)\left[\frac{\phi(r_j)}{r_j} - \phi(r_{j-1})\right]\right\}(u_j^n - u_{j-1}^n) \\ &= u_j^n - C_{j-\frac{1}{2}}^n(u_j^n - u_{j-1}^n) + D_{j+\frac{1}{2}}^n(u_{j+1}^n - u_j^n) \end{aligned} \tag{2.13}$$

with

$$C_{j-\frac{1}{2}}^n = \nu\left\{1 + \frac{1}{2}(1-\nu)\left[\frac{\phi(r_j)}{r_j} - \phi(r_{j-1})\right]\right\} \quad \text{and} \quad D_{j+\frac{1}{2}}^n = 0.$$

The conditions of Theorem 2.12 reduce to the single condition

$$0 \leq C^n_{j+\frac{1}{2}} \leq 1.$$

If

$$\left|\frac{\phi(r)}{r} - \phi(s)\right| \leq 2 \quad \text{for all } r \neq 0, s, \tag{2.14}$$

then this condition is satisfied for all Courant numbers $\nu \in (0,1]$. Under the natural assumptions that

$$\phi(r) = 0 \quad \text{for all } r \leq 0 \qquad \text{and} \qquad \phi(r) \geq 0 \quad \text{for all } r \geq 0$$

Condition (2.14) is satisfied, if and only if

$$\phi(r) \leq \min(2, 2r).$$

In summary we obtain:

Theorem 2.13. *Assume that $\phi\colon \mathbb{R} \longrightarrow \mathbb{R}$ is Lipschitz continuous and $0 \leq \phi(r) \leq \min(2,2r)$. Then the limited Lax-Wendroff method (2.13) is TVD.*

For a suitable choice of ϕ one obtains consistency order 2:

Theorem 2.14. *Let $u(x,t)$ be a smooth solution of*

$$u_t + a\,u_x = 0 \qquad \textit{where } a > 0.$$

Assume that $\phi\colon \mathbb{R} \longrightarrow \mathbb{R}$ is Lipschitz continuous with

$$\phi(1) = 1.$$

Then the limited Lax-Wendroff method (2.13) has consistency order 2 in so-called non-critical) points, i.e. in points $(\hat{x},\hat{t})$ with $u_x(\hat{x},\hat{t}) \neq 0$.

Proof. The limited Lax-Wendroff method is of the form:

$$u^{n+1}_j = H(u^n_{j-2}, u^n_{j-1}, u^n_j, u^n_{j+1})$$

with

$$\begin{aligned} &H(u^n_{j-2}, u^n_{j-1}, u^n_j, u^n_{j+1}) = \\ &\qquad u^n_j - \nu(u^n_j - u^n_{j-1}) - \frac{(1-\nu)\nu}{2}\left[\phi(r^n_j)(u^n_{j+1} - u^n_j) - \phi(r^n_{j-1})(u^n_j - u^n_{j-1})\right] \end{aligned}$$

We have

$$H(u^n_{j-2}, u^n_{j-1}, u^n_j, u^n_{j+1}) = H_{LW}(u^n_{j-1}, u^n_j, u^n_{j+1}) - \frac{(1-\nu)\nu}{2} R(u^n_{j-2}, u^n_{j-1}, u^n_j, u^n_{j+1})$$

with

$$H_{LW}(,u_{j-1}^n,u_j^n,u_{j+1}^n) = u_j^n - \nu(u_j^n - u_{j-1}^n) - \frac{(1-\nu)\nu}{2}\left[(u_{j+1}^n - u_j^n) - (u_j^n - u_{j-1}^n)\right]$$

and

$$R(u_{j-2}^n,u_{j-1}^n,u_j^n,u_{j+1}^n) = [\phi(r_j^n)-1](u_{j+1}^n - u_j^n) - [\phi(r_{j-1}^n)-1](u_j^n - u_{j-1}^n).$$

The (un-limited) Lax-Wendroff method has consistency order 2. Therefore, it suffice to study the term $R(u_{j-2}^n,u_{j-1}^n,u_j^n,u_{j+1}^n)$. We have

$$R = [\phi(r_+)-1][u(x+\Delta x,t)-u(x,t)] - [\phi(r_-)-1][u(x,t)-u(x-\Delta x,t)]$$

with

$$r_+ = \frac{u(x,t)-u(x-\Delta x,t)}{u(x+\Delta x,t)-u(x,t)} \quad \text{and} \quad r_- = \frac{u(x-\Delta x,t)-u(x-2\Delta x,t)}{u(x,t)-u(x-\Delta x,t)}.$$

By Taylor expansion it follows that:

$$\begin{aligned} R &= [\phi(r_+)-1]\left[u_x\Delta x + \frac{1}{2}u_{xx}\Delta x^2 + O(\Delta x^3)\right] \\ &\quad - [\phi(r_-)-1]\left[u_x\Delta x - \frac{1}{2}u_{xx}\Delta x^2 + O(\Delta x^3)\right] \\ &= [\phi(r_+)-\phi(r_-)]\; u_x\Delta x + [\phi(r_+)+\phi(r_-)-2]\; u_{xx}\Delta x^2 + O(\Delta x^3) \end{aligned}$$

Moreover,

$$r_+ = 1 - \frac{u_{xx}}{u_x}\Delta x + O(\Delta x^2) \quad \text{and} \quad r_- = 1 - \frac{u_{xx}}{u_x}\Delta x + O(\Delta x^2)$$

which implies

$$\phi(r_+) - \phi(r_-) = O(r_+ - r_-) = O(\Delta x^2)$$

because ϕ is Lipschitz continuous and

$$\phi(r_+)+\phi(r_-)-2 = \phi(r_+)-1+\phi(r_-)-1 = O(r_+-1)+O(r_--1) = O(\Delta x).$$

Hence: $R = O(\Delta x^3)$, which completes the proof. □

Remark:

Setting $\phi(r) = 0$ leads to the Courant-Isaacson-Rees method, which is TVD but not of second order.

Setting $\phi(r) = 1$ leads to the Lax-Wendroff method, which is of second order but not TVD.

Setting $\phi(r) = r$ leads to the so-call Warming-Beam method, which is of second order but not TVD.

Each setting of the form

$$\phi(r) = \theta(r)\,1 + (1-\theta(r))\,r$$

leads to a method of order 2, if θ is Lipschitz continuous. Numerical experiments show that methods with $0 \le \theta(r) \le 1$ leads to good results. Flux limiters of this form which lead to TVD methods lie between the so-called "super-bee"-limiter by Roe

$$\phi(r) = \max(0, \min(1, 2r), \min(r, 2))$$

and the "min-mod"-limiter

$$\phi(r) = \max(0, \min(r, 1)).$$

Remark: Another strategy to construct higher-order methods is the following variant of Godunov's method: Replace the first step by

1'. Reconstruction of a function $v(., t_n)$ from the values u_j^n as a piecewise linear function:

$$v(x, t_n) = u_j^n + s_j^n (x - x_j)$$

e.g. with

$$s_j^n = \frac{1}{\Delta x}(u_{j+1}^n - u_j^n).$$

For the linear wave equation

$$u_t + a\,u_x = 0 \qquad \text{with } a > 0$$

this leads to the Lax-Wendroff method again, which is of second order but not TVD. In order to obtain stability so-called slope limiters $\psi(d_1, d_2)$ are introduced which redefine the slope

$$s_j^n = \frac{1}{\Delta x}\psi(u_j^n - u_{j-1}^n, u_{j+1}^n - u_j^n)$$

and whose job is to guarantee TV-stability, e.g., by a TVD method:

$$TV(v(., t_n)) \le TV(u^n) = \sum_j |u_{j+1}^n - u_j^n|.$$

The next two steps in Godunov's method (exact computation of the entropy solution with initial value $v(., t_n)$ and averaging) do not increase the total variation.

Remark: One can also use the piecewise linear reconstruction in combination with a method of the form

$$u_j^{n+1} = u_j^n - \lambda\left[g(u_j^n, u_{j+1}^n) - g(u_{j-1}^n, u_j^n)\right]$$

(typically a monotone method of first order) in order to obtain a method of second order. One simply replaces the arguments u_j^n and u_{j+1}^n of the numerical flux $g(u_j^n, u_{j+1}^n)$ by the left-sided and right-sided limits of the piecewise linear reconstruction at the point

$x_{j+1/2}$. The resulting methods are of MUSCL-type (monotone upstream-centered scheme for conservation laws):

$$u_j^{n+1} = u_j^n - \lambda \left[g(u^L_{j+\frac{1}{2}}, u^R_{j+\frac{1}{2}}) - g(u^L_{j-\frac{1}{2}}, u^R_{j-\frac{1}{2}}) \right]$$

with

$$u^L_{j+\frac{1}{2}} = \lim_{x \to x_{j+\frac{1}{2}}-} v(x, t_n), \quad u^R_{j+\frac{1}{2}} = \lim_{x \to x_{j+\frac{1}{2}}+} v(x, t_n).$$

Remark: A further extension of these techniques is based on more general piecewise polynomial reconstructions. This leads, among others, to the so-called ENO methods (essentially non-oscillatory schemes).

Chapter 3

Multi-Dimensional Scalar Conservation Laws

3.1 Finite Volume Methods

We consider the Cauchy problem for a multi-dimensional scalar conservation law of the form

$$\begin{aligned} u_t + \operatorname{div} f(u) &= 0 \\ u(x,0) &= u_0(x) \end{aligned}$$

with a flux function $f\colon \mathbb{R} \longrightarrow \mathbb{R}^d$.

Let $\mathcal{T}_h$ be a subdivision of $\mathbb{R}^d$ into non-overlapping polygonal (polyhedral) cells (elements, finite volumes):

$$\mathcal{T}_h = \{T_j \;\Big|\; T_j \text{ is a polyhedron, } j \in J\}$$

with

1. $\mathbb{R}^d = \bigcup_{j \in J} T_j$
2. $T_i \cap T_j$ is either empty, a common vertex, a common edge, (or a common face) for all $i \neq j$.

By integrating over $T_j \times [t_n, t_{n+1}]$ and dividing by the area (volume) $|T_j|$ of T_j and $\Delta t = t_{n+1} - t_n$ one obtains:

$$\frac{1}{\Delta t}\left[\frac{1}{|T_j|}\int_{T_j} u(x,t_{n+1})\;dx - \frac{1}{|T_j|}\int_{T_j} u(x,t_n)\;dx\right] + \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \frac{1}{|T_j|}\int_{\partial T_j} f(u)\cdot n\;dS = 0,$$

where n denotes the outward unit normal vector.

Let $N(j)$ denote the set of indices of the cells T_k, which share a common edge (face) with the cell T_j. For each $k \in N(j)$ let S_{jk} be this common edge (face). Then

$$\partial T_j = \bigcup_{k \in N(j)} S_{jk}$$

and

$$\int_{\partial T_j} f(u) \cdot n \; dS = \sum_{j \in N(j)} \int_{Sjk} f(u) \cdot n_{jk} \; dS,$$

where n_{jk} is the outward unit normal vector for S_{jk}.

Assume that approximate solutions of the averaged exact solution over the cells T_j are available at time t_n:

$$u_j^n \approx \frac{1}{|T_j|} \int_{T_j} u(x, t_n) \; dx.$$

In order to compute approximate solutions at time t_{n+1}

$$u_j^{n+1} \approx \frac{1}{|T_j|} \int_{T_j} u(x, t_{n+1}) \; dx,$$

approximations of the time- and space-averaged fluxes over the edges (faces) are needed:

$$\frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \frac{1}{|S_{jk}|} \int_{S_{jk}} f(u) \cdot n_{jk} \; dS \approx g_{jk}^n = g_{jk}(u_j^n, u_k^n, n_{jk}),$$

where $|S_{jk}|$ is the length (area) of the edge (face) of S_{jk}. The quantity $g_{jk}(v, w, n)$ is called the numerical flux.

This leads to the following general form of an explicit finite volume method:

$$\begin{aligned} u_j^{n+1} &= u_j^n - \frac{\Delta t}{|T_j|} \sum_{k \in N(j)} g_{jk}(u_j^n, u_k^n, n_{jk}) |S_{jk}| \\ &= H_j(u^n) \end{aligned} \tag{3.1}$$

We assume the following properties:

1. The method is conservative: $g_{kj}(v, w, n) = -g_{jk}(w, v, -n)$.
2. The numerical flux is consistent: $g_{jk}(u, u, n) = f(u) \cdot n$.

The methods developed for the one-dimensional case can be easily extended to the multi-dimensional case:

The Lax-Friedrichs method:

$$g_{jk}(v, w, n) = \frac{1}{2} \left[f(v) \cdot n + f(w) \cdot n - \frac{1}{\lambda_{jk}} (w - v) \right].$$

Upwind methods

For linear fluxes $f(u) = a\,u$ with $a \in \mathbf{R}^d$ the Courant-Isaacson-Rees methods leads to

$$g_{jk} = (a \cdot n)^+ v + (a \cdot n)^- w.$$

For homogeneous fluxes $f(\alpha u) = \alpha f(u)$ it follows that

$$f(u) = a(u)\,u \quad \text{with } a(u) = f'(u).$$

This motivates the following extensions for an upwind method:

$$g_{jk}(v, w, n) = \left(a\left(\frac{v+w}{2}\right) \cdot n\right)^+ v + \left(a\left(\frac{v+w}{2}\right) \cdot n\right)^- w$$

(partial upwind after Vijayasundaram) or

$$g_{jk}(v, w, n) = (a(v) \cdot n)^+ v + (a(w) \cdot n)^- w$$

(full upwind after Steger-Warming).

3.2 Discontinuous Galerkin Space Discretization

Let $\mathcal{T}_h$ be a polygonal (polyhedral) subdivision of $\mathbb{R}^d$. By multiply the conservation law with a piecewise continuous test function $v(x)$ and integrating over a cell $T_j \in \mathcal{T}_h$ one obtains

$$\frac{d}{dt}\int_{T_j} u(x,t)\,v(x)\;dx + \int_{T_j} \operatorname{div} f(u(x,t))\,v(x)\;dx = 0.$$

Integration by parts leads to

$$\frac{d}{dt}\int_{T_j} u(x,t)\,v(x)\;dx + \int_{\partial T_j} f(u(x,t)) \cdot n(x)\,v(x)\;dS - \int_{T_j} f(u(x,t)) \cdot \operatorname{grad} v(x)\;dx = 0.$$

Now we replace the exact solution $u(.,t) \in V = L^\infty(\mathbb{R})$ by an approximate solution $u_h(t) \in V_h$ with

$$V_h \subset \{v \in L^\infty(\mathbb{R}) : v|_T \in P_k \text{ for all } T \in \mathcal{T}_h\}.$$

The test functions are also restricted to the set V_h:

$$\frac{d}{dt}\int_{T_j} u_h(x,t)\,v_h(x)\;dx + \int_{\partial T_j} f(u_h(x,t)) \cdot n(x)\,v_h(x)\;dS - \int_{T_j} f(u_h(x,t)) \cdot \operatorname{grad} v_h(x)\;dx = 0.$$

The normal flux $f(u_h(x,t)) \cdot n(x)$ on the part S_{jk} of ∂T_j is replaced by a numerical flux $g_{jk}(u_h(x,t)^-, u_h(x,t)^+, n(x))$, where $u_h(x,t)^-$ and $u_h(x,t)^+$ denote the one-sided limits of $u_h(x,t)$ from the interior and the exterior of the cell T_j, respectively:

$$\begin{aligned} \frac{d}{dt}\int_{T_j} u_h(x,t)\,v_h(x)\;dx \quad &+ \sum_{k \in N(j)} \int_{S_{jk}} g_{jk}(u_h(x,t)^-, u_h(x,t)^+, n_{jk})\,v_h(x)\;dS \\ &- \int_{T_j} f(u_h(x,t)) \cdot \operatorname{grad} v_h(x)\;dx \;=\; 0. \end{aligned}$$

Finally, the integrals are replaced by suitable quadrature rules.

This spatial semi-discretization technique results in a system of ordinary differential equations, which is typically discretized in time by a suitable Runge-Kutta method.

Example: For

$$V_h = \{v \in L^\infty(\mathbb{R}) : v|_T \in P_0 \text{ for all } T \in \mathcal{T}_h\}$$

the method reduces to

$$|T_j| \frac{d}{dt} u_j(t) + \sum_{k \in N(j)} |S_{jk}| \, g_{jk}(u_j(t), u_k(t), n_{jk}) = 0.$$

with the notation $u_j(t) = u_h(x,t)$ for $x \in T_j$. The finite volume method of the previous section corresponds to the explicit Euler method for discretizing in time.

3.3 Measure-Valued Solutions

The important concepts of monotone methods and TVD methods from the one-dimensional case can be easily extended.

If g_{jk} is the numerical flux of a monotone one-dimensional method for all $j, k \in J$ then the finite volume method is monotone under an appropriate CFL condition.

Unfortunately, we have:

Theorem 3.1. *A TVD method of the form*

$$\begin{aligned} u^{n+1}_{j_1,j_2} = u^n_{j_1,j_2} \quad &- \quad \lambda_x \left[g_1(u_{j_1-l_1+1,j_2-l_2}, \ldots, u_{j_1+l_1,j_2+l_2}) - g_1(u_{j_1-l_1,j_2-l_2}, \ldots, u_{j_1+l_1-1,j_2+l_2}) \right] \\ &- \quad \lambda_y \left[g_2(u_{j_1-l_1,j_2-l_2+1}, \ldots, u_{j_1+l_1,j_2+l_2}) - g_2(u_{j_1-l_1,j_2-l_2}, \ldots, u_{j_1+l_1,j_2+l_2-1}) \right] \end{aligned}$$

is at most of order 1 except for trivial cases.

Observe that finite volume methods based on one-dimensional TVD methods are, in general, not TVD.

Because of Theorem 3.1 it seems to be reasonable to relax the conditions on the sequence (u_m) of approximate solutions. We keep the L^∞-stability but try to do without the TV-stability: There exists a constant C such that

$$\|u_m\|_{L^\infty(\mathbb{R}^d)} \le C.$$

Then the compactness argument is no longer true in $L^1(\mathbb{R}^r \times [0,\infty))$. However, there is a compactness argument in a weaker topology: Since

$$L^\infty(\mathbb{R}^d \times [0,\infty)) = \left(L^1(\mathbb{R}^d \times [0,\infty)) \right)^*,$$

the sequence (u_m) is a bounded sequence in the dual space of $L^1(\mathbb{R}^d \times [0,\infty))$.

We have the following important result:

Theorem: Let X be a separable Banach space. Then bounded sets in the dual space X^* are pre-compact in the weak-$*$ topology.

This means here: There is a sub-sequence (u_{m_l}) and a function $u \in L^\infty(\mathbb{R} \times [0,\infty))$ with

$$\int_0^\infty \int_{\mathbb{R}^d} u_{m_l}\, v\; dx\; dt \to \int_0^\infty \int_{\mathbb{R}^d} u\, v\; dx\; dt$$

for all $v \in L^1(\mathbb{R}^d \times [0,\infty))$.

It remains to check whether u is a weak solution. We start in the same way as in the proof of the Theorem of Lax-Wendroff. For simplicity we consider the one-dimensional case only and we consider only test functions $\varphi \in C_0^\infty(\mathbb{R} \times (0,\infty))$ instead of the more general test functions $\varphi \in C_0^\infty(\mathbb{R} \times [0,\infty))$. Then

$$\begin{aligned} 0 &= \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \left[(u_j^{n+1} - u_j^n)\Delta x + (g^n_{j+\frac{1}{2}} - g^n_{j-\frac{1}{2}})\Delta t \right] \varphi_j^n \\ &= \underbrace{-\sum_{n=1}^{\infty} \sum_{-\infty}^{\infty} u_j^n \frac{\varphi_j^n - \varphi_j^{n-1}}{\Delta t} \Delta x \Delta t}_{=I} \underbrace{- \sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} g^n_{j+\frac{1}{2}} \frac{\varphi^n_{j+1} - \varphi^n_j}{\Delta x} \Delta x \Delta t.}_{=II} \end{aligned}$$

We have

$$I = \int_{\Delta t_m}^\infty \int_{-\infty}^\infty u_m (\Delta_t \varphi)_m\; dx\; dt \;\to\; \int_0^\infty \int_{-\infty}^\infty u\, \varphi_t\; dx\; dt,$$

since

$$u_m\,(\Delta_t\varphi)_m = \underbrace{u_m\,\varphi_t}_{\overset{*}{\rightharpoonup}\, u\,\varphi_t} + \underbrace{u_m}_{|.| \le C} \underbrace{[(\Delta_t\varphi)_m - \varphi_t]}_{\to 0} \overset{*}{\rightharpoonup} u\,\varphi_t.$$

Moreover, we have

$$II = \underbrace{\sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} \left[g^n_{j+\frac{1}{2}} - f(u_j^n) \right] \frac{\varphi^n_{j+1} - \varphi^n_j}{\Delta x} \Delta x \Delta t}_{=III} + \underbrace{\sum_{n=0}^{\infty} \sum_{j=-\infty}^{\infty} f(u_j^n) \frac{\varphi^n_{j+1} - \varphi^n_j}{\Delta x} \Delta x \Delta t}_{=IV}$$

The convergence of the term III towards 0 requires some technical conditions which are not further discussed here.

For the last term IV we obtain:

$$IV = \int_0^\infty \int_{-\infty}^\infty f(u_m)\,\varphi_x\; dx\; dt + \int_0^\infty \int_{-\infty}^\infty f(u_m)\,[(\Delta_x\varphi)_m - \varphi_x]\; dx\; dt$$

The convergence of the second term towards 0 follows from the boundedness of $f(u_m)$ ((u_m) is bounded, f is continuous) and the convergence of $(\Delta_x\varphi)_m$ towards φ_x.

That leaves the question open, whether $f(u_m)\,\varphi_x$ converges towards $f(u)\,\varphi_x$ in the weak-$*$ topology:

$$u_m \overset{*}{\rightharpoonup} u \;\overset{?}{\Rightarrow}\; f(u_m) \overset{*}{\rightharpoonup} f(u)$$

for continuous functions f. Unfortunately, this is wrong in general. However, it can be shown that there is a sub-sequence (u_{m_l}) with:

$$f(u_{m_l}(x,t)) \overset{*}{\rightharpoonup} \int_{\mathbb{R}} f(\lambda)\ d\nu_{(x,t)}(\lambda) = \langle \nu_{(x,t)}, f \rangle,$$

where $\nu_{(x,t)}$ is a probability measure on $\mathbb{R}$ with respect to the σ-algebra of the Borel sets, for each $(x,t) \in \mathbb{R} \times (0,\infty)$.

Let $Prob(\mathbb{R})$ be the set of probability measures on $\mathbb{R}$ with respect to the σ-algebra of the Borel sets. The measure-valued function $\nu \colon \mathbb{R} \times (0,\infty) \longrightarrow Prob(\mathbb{R})$, $(x,t) \mapsto \nu_{(x,t)}$ is called the Young-measure.

Therefore we have:

$$\int_0^\infty \int_{-\infty}^\infty f(u_m)\, \varphi_x\ dx\ dt \to \int_0^\infty \int_{-\infty}^\infty \langle \nu_{(x,t)}, f \rangle \varphi_x\ dx\ dt.$$

This implies for the special case $f = id$:

$$\int_0^\infty \int_{-\infty}^\infty u_m\, \varphi_x\ dx\ dt \to \int_0^\infty \int_{-\infty}^\infty \langle \nu_{(x,t)}, id \rangle \varphi_x\ dx\ dt.$$

In summary, we obtain:

$$\int_0^\infty \int_{-\infty}^\infty \langle \nu_{(x,t)}, id \rangle_t\ dx\ dt + \int_0^\infty \int_{-\infty}^\infty \langle \nu_{(x,t)}, f \rangle_x\ dx\ dt = 0$$

for all $\varphi \in C_0^\infty(\mathbb{R} \times (0,\infty))$.

This motivates the following concept of a solution:

Definition 3.1. *A function $\nu \colon \mathbb{R} \times [0,\infty) \longrightarrow Prob(\mathbb{R})$ is called a measure-valued solution of the conservation law if and only if (in the distributional sense)*

$$\langle \nu_{(x,t)}, id \rangle_t + \langle \nu_{(x,t)}, f \rangle_x = 0 \qquad (x,t) \in \mathbb{R} \times (0,\infty).$$

Remark: It is easy to see that, for two measure-valued solutions $\nu^{[1)}$ and $\nu^{(2)}$, also any convex combination $\alpha\nu^{(1)} + (1-\alpha)\nu^{(2)}$ is a measure-valued solution. So, surprisingly, the non-linear conservation law has become a convex-linear problem.

So the existence of an L^∞-bounded sequence of approximate solutions guarantees the existence of a measure-valued solution of the conservation law. If

$$\nu_{(x,t)} = \delta_{u(x,t)},$$

where δ_v is the Dirac measure centered at v, then:

$$\int_{\mathbb{R}} f(\lambda)\ d\nu_{(x,t)} = \int_{\mathbb{R}} f(\lambda)\ d\delta_{(x,t)} = f(u(x,t)).$$

This implies that u is a weak solution. So the existence of a weak solution can be rephrased as: Is the Young measure a Dirac measure?

Starting from the initial condition

$$\nu_{(x,0)} = \delta_{u_0(x)}$$

and using a discrete entropy condition one can show, that, under appropriate assumptions,

$$\nu_{(x,t)} = \delta_{u(x,t)}. \tag{3.2}$$

Remark: Property (3.2) also implies

$$u_m \to u \qquad \text{in } L^1_{loc}(\mathbb{R} \times [0, \infty).$$

Chapter 4

One-Dimensional Systems of Conservation Laws

4.1 Hyperbolic Systems

We consider the Cauchy problem for a one-dimensional system of conservation laws of the form

$$\begin{aligned} u_t + f(u)_x &= 0 \\ u(x,0) &= u_0(x) \end{aligned}$$

with a flux function $f\colon D \longrightarrow \mathbb{R}^p$, $D \subset \mathbb{R}^p$ open. An important example are the one-dimensional Euler equations:

Example: The one-dimensional Euler equations for a perfect gas are of this form with

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} \rho \\ \rho v \\ \rho e \end{pmatrix}, \qquad f(u) = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ (\rho e + p) v \end{pmatrix}$$

and $p = (\kappa - 1)\rho\varepsilon = (\kappa - 1)(\rho e - \rho v^2/2)$, where $\rho > 0$ and $\varepsilon > 0$. Hence

$$f(u) = \begin{pmatrix} u_2 \\ \dfrac{3-\kappa}{2}\,\dfrac{u_2^2}{u_1} + (\kappa - 1)\,u_3 \\ \kappa\,\dfrac{u_2 u_3}{u_1} - \dfrac{\kappa - 1}{2}\,\dfrac{u_2^3}{u_1^2} \end{pmatrix}$$

for $u \in D = \{(u_1, u_2, u_3)^T : u_1 > 0,\ u_3 - u_2^2/(2u_1) > 0\}$.

For smooth solutions the system of conservation laws can also be written in a quasi-linear form

$$u_t + A(u) u_x = 0 \quad \text{with } A(u) = f'(u).$$

Definition 4.1. *Let $f \in C^1(D, \mathbb{R}^p)$. The system $u_t + f(u)_x = 0$ is called (strictly) hyperbolic, if and only if $f'(u)$ has p (distinct) real eigenvalues and p linearly independent corresponding eigenvectors for all $u \in D$.*

Example: Let A be a constant $p \times p$ matrix. Consider the Cauchy problem

$$\begin{aligned} u_t + Au_x &= 0, \\ u(x,0) &= u_0(x). \end{aligned}$$

This problem is called well-posed if and only if, for each $T \geq 0$, there is a constant C_T such that

$$\|u(.,t)\|_{L^2(\mathbb{R})} \leq C_T \|u_0\|_{L^2(\mathbb{R})} \quad \text{for all } t \leq T.$$

It can be shown that the problem is well-posed if and only if A has p real eigenvalues and p linearly independent corresponding eigenvectors, i.e., if the problem is hyperbolic.

Example: The one-dimensional Euler equations for a perfect gas are a hyperbolic system. The eigenvalues of $A(u) = f'(u)$ are given by $\lambda_1(u) = v - c$, $\lambda_2(u) = v$ and $\lambda_3(u) = v - c$ with $c = \sqrt{\kappa p/\rho}$, the so-called speed of sound.

Characteristic curves

Assume that $f'(u) = A(u)$ has p real eigenvalues

$$\lambda_1(u) \leq \lambda_2(u) \leq \ldots \leq \lambda_p(u)$$

with p linearly independent corresponding (right) eigenvectors $r_k(u)$, $k = 1, \ldots, p$, which build the matrix $R(u) = (r_1(u), \ldots, r_p(u))$. Then

$$R(u)^{-1}A(u)R(u) = \Lambda(u) = \operatorname{diag}(\lambda_1(u), \ldots, \lambda_p(u)).$$

The eigenvector r_k can be interpreted as a mapping from D to $\mathbb{R}^p$, i.e. a vector field, and is called the k-th characteristic field.

Definition 4.2. *Let u be a smooth solution of the hyperbolic system. A curve in $\mathbb{R} \times [0, \infty)$, parameterized by $(\gamma(t), t), t \in [0, \tau]$, is called a k-characteristic curve if and only if*

$$\begin{aligned} \gamma'(t) &= \lambda_k(u(\gamma(t), t)), \\ \gamma(0) &= x_0. \end{aligned}$$

Consider a hyperbolic system in quasi-linear form:

$$u_t + A(u)u_x = 0.$$

Since $R(u)^{-1}A(u)R(u) = \Lambda(u)$, the system can be written in the following equivalent form:

$$u_t + R(u)\Lambda(u)R(u)^{-1}u_x = 0,$$

or, by multiplying with $L(u) = R(u)^{-1}$ from the left:

$$L(u)u_t + \Lambda(u)L(u)u_x = 0$$

Let $l_k^T(u)$ denote the k-th row of $L(u)$. Then $l_k(u)^T$ is a left eigenvector of $A(u)$ for the eigenvalue $\lambda_k(u)$. Along a k-characteristic curve the hyperbolic system has the following form:

$$\begin{aligned} 0 &= l_k(u)^T u_t + \lambda_k(u) l_k(u)^T u_x = l_k(u)^T(u_t + \lambda_k(u)u_x) = l_k(u)^T(u_t + \gamma' u_x) \\ &= l_k(u(\gamma(t),t))^T \frac{d}{dt} u(\gamma(t),t). \end{aligned}$$

So the system of partial differential equations reduces to an ordinary differential equations along a k-characteristic curve.

Example: If A is constant and if the new variables $v = Lu$ are introduced, then it immediately follows that the k-th component $v_k = l_k^T u$ is constant along a k-characteristic curve. The characteristic curves are straight lines. With this information a solution at (x,t) can easily be constructed from the initial data at the points $x - \lambda_k t$, $k = 1, \ldots, p$:

$$v_k(x,t) = v_k(x - \lambda_k t, 0) = l_k^T u_0(x - \lambda_k t).$$

Hence

$$u(x,t) = Rv(x,t) = \sum_{k=1}^{m} v_k(x,t) r_k = \sum_{k=1}^{m} l_k^T u_0(x - \lambda_k t) r_k.$$

The concepts of weak solutions, entropy solutions and measure-valued solutions introduced for scalar conservation laws can directly be extended to systems of conservation laws.

Some important differences to the scalar case are shortly discussed now:

1. The Rankine-Hugoniot jump condition has exactly the same form as in the scalar case:

   $$s\,(u_R - u_L) = f(u_R) - f(u_L),$$

   which leads to the following condition in the linear case $f(u) = Au$:

   $$s\,(u_R - u_L) = A(u_R - u_L).$$

   So s, the speed of propagation of the discontinuity, must be an eigenvalue of A, the jump $u_R - u_L$ must be a corresponding right eigenvector of A.

2. The existence of an entropy function $U(u)$ is not guaranteed for systems, since the compatibility condition

   $$U'(u)f'(u) = F'(u)$$

   is an over-determined system of differential equations.

3. L^∞ estimates of the solutions are, in general, not known for systems.

As in the scalar case the solution of Riemann problems play an essential role in the construction of Godunov's method and its variants, for systems of conservation laws.

4.2 The Riemann Problem

Linear Systems with Constant Coefficients

For the special initial condition

$$u(x,0) = u_0(x) = \begin{cases} u_L & \text{for } x < 0, \\ u_R & \text{for } x > 0 \end{cases}$$

one obtains for $v_k(x,t) = l_k^T u(x,t)$:

$$v_k(x,t) = l_k^T u_0(x - \lambda_k t) = \begin{cases} l_k^T u_L & \text{for } x < \lambda_k t, \\ l_k^T u_R & \text{for } x > \lambda_k t. \end{cases}$$

With the notation

$$\alpha_k = l_k^T u_L, \quad \beta_k = l_k^T u_R$$

it follows for the case $\lambda_l < x/t < \lambda_{l+1}$ with $\lambda_0 = -\infty$ and $\lambda_{p+1} = +\infty$:

$$u(x,t) = \sum_{k=1}^{p} v_k(x,t) r_k = \sum_{k=1}^{l} \beta_k r_k + \sum_{k=l+1}^{p} \alpha_k r_k = w_l,$$

hence

$$u(x,t) = u^*(x/t; u_L, u_R)$$

with

$$u^*(x/t; u_L, u_R) = \begin{cases} w_0 = u_L & \text{for } x/t < \lambda_1, \\ w_1 & \text{for } \lambda_1 < x/t < \lambda_2, \\ \vdots & \\ w_{p-1} & \text{for } \lambda_{p-1} < x/t < \lambda_p, \\ w_p = u_R & \text{for } \lambda_p < x/t. \end{cases}$$

Nonlinear Systems

It is reasonable to use the same ansatz

$$u(x,t) = v\left(\frac{x}{t}\right)$$

also for nonlinear hyperbolic systems

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

If v is smooth, then u is a solution if and only if

$$v'\left(-\frac{x}{t^2}\right) + f'(v)v'\,\frac{1}{t} = 0,$$

hence

$$f'(v(\xi))v'(\xi) = \xi\, v'(\xi)$$

with $\xi = x/t$. So, either $v'(\xi) = 0$ or

$$\begin{aligned} v'(\xi) &= \alpha(\xi)\, r_k(v(\xi)), \\ \xi &= \lambda_k(v(\xi)) \end{aligned}$$

for some $\alpha(\xi) \neq 0$. Then

$$\nabla\lambda_k(v(\xi))^T r_k(v(\xi)) = \frac{1}{\alpha(\xi)} \neq 0$$

for some $k \in \{1, \dots, p\}$. This motivates the following definition:

Definition 4.3. *1. The k-th characteristic field r_k is called genuinely nonlinear if and only if*

$$\nabla\lambda_k(u)^T r_k(u) \neq 0 \quad \textit{for all } u \in D.$$

2. The k-th characteristic field r_k is called linearly degenerate, if and only if

$$\nabla\lambda_k(u)^T r_k(u) = 0 \quad \textit{for all } u \in D.$$

Remark: If r_k is genuinely nonlinear we always obtain

$$\nabla\lambda_k(u)^T r_k(u) = 1 \qquad \text{for all } u \in D$$

by an appropriate scaling of the eigenvector.

With the help of these concepts the following three kinds of special solutions of Riemann problems can be constructed:

k-Rarefaction Waves

Assume that r_k is genuinely nonlinear, let v_k be the solution of the initial value problem

$$\begin{aligned} v_k'(\xi) &= r_k(v_k(\xi)), \quad \xi \geq 0 \\ v_k(0) &= u_L \end{aligned}$$

and set

$$u_R = v_k(\xi_R)$$

for some $\xi_R \geq 0$. It is easy to see that

$$u(x,t) = \begin{cases} u_L & \text{for } \dfrac{x}{t} < \lambda_k(u_L) \\ v_k\left(\dfrac{x}{t} - \lambda_k(u_L)\right) & \text{for } \lambda_k(u_L) \leq \dfrac{x}{t} \leq \lambda_k(u_R) \\ u_R & \text{for } \lambda_k(u_R) < \dfrac{x}{t} \end{cases}$$

is a continuous and piecewise smooth solution and, therefore, a weak entropy solution of the Riemann problem

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

Because

$$\frac{d}{d\xi}\left[\lambda_k(v_k(\xi)) - \xi\right] = \nabla\lambda_k(v_k(\xi)) \cdot v_k'(\xi) - 1 = 0$$

the function $\lambda_k(v_k(\xi))$ is monotonically increasing in ξ, which guarantees that $\lambda_k(u_L) \leq \lambda_k(u_R)$. So $u(x,t)$ is well-defined. Moreover, we have

$$\lambda_k(v_k(\xi)) - \xi = \lambda_k(u_L),$$

which implies:

$$u_t + f'(u)u_x = \frac{1}{t} r_k(v_k(\xi))\left[-\xi - \lambda_k(u_L) + \lambda_k(v_k(\xi))\right] = 0$$

with $\xi = x/t - \lambda_k(u_L)$. In particular, we have

$$\lambda_k(u_R) - \xi_R = \lambda_k(u_L),$$

which guarantees the continuity of $u(x,t)$ at $x/t = \lambda_k(u_R)$. The continuity of $u(x,t)$ at $x/t = \lambda_k(u_L)$ is trivial.

So, for each state u_L, there is a one-parameter set of states u_R, described by the so-called k-rarefaction curve $\mathcal{R}_k(u_L)$, parameterized by $\xi \mapsto v_k(\xi)$, $\xi \geq 0$, for which the k-rarefaction wave (k-simple wave) defined above is the solution of the Riemann problem. Moreover, $r_k(u_L)$ is a tangent vector to the curve $\mathcal{R}_k(u_L)$ at u_L.

k-Shock Waves

A piecewise constant function of the form

$$u(x,t) = \begin{cases} u_L \text{ for } \dfrac{x}{t} < s, \\ u_R \text{ for } \dfrac{x}{t} > s \end{cases}$$

is a weak solution, if and only if the Rankine-Hugoniot jump condition

$$s\,(u_R - u_L) = f(u_R) - f(u_L)$$

is satisfied.

For a given state u_L the Rankine-Hugoniot set is the set of all states u_R such that there exists an $s(u_L, u_R)$ with

$$s(u_L, u_R)\,(u_R - u_L) = f(u_R) - f(u_L).$$

It can be shown that the Rankine-Hugoniot set of u_L can be locally described by p smooth curves $\mathcal{S}_k(u_L)$, parameterized by $\xi \mapsto v_k(\xi)$ for some smooth function v_k, and $r_k(u_L)$ is a tangent vector to the curve $\mathcal{S}_k(u_L)$ at u_L.

If r_k is genuinely nonlinear, the piecewise constant discontinuous solution described above is called a k-shock wave. It is an entropy solution for small values of the parameter ξ if and only if $\xi \leq 0$. Then the solution is called an admissible k-shock wave.

k-Contact Discontinuities

Assume that r_k linearly degenerate, let v_k be the solution of the problem

$$\begin{aligned} v_k'(\xi) &= r_k(v_k(\xi)), \quad \xi \in \mathbb{R} \\ v_k(0) &= u_L \end{aligned}$$

and set

$$u_R = v_k(\xi_R)$$

for some $\xi_R \in \mathbb{R}$. Then

$$\frac{d}{d\xi}\left(\lambda_k(v_k(\xi))\right) = \nabla\lambda_k(v_k)^T r_k(v_k)\rangle = 0.$$

Hence

$$\lambda_k(v_k(\xi)) = \lambda_k(u_L) = \lambda_k(u_R).$$

The piecewise constant solution

$$u(x,t) = \begin{cases} u_L \text{ for } \dfrac{x}{t} < \lambda_k(u_L), \\ u_R \text{ for } \dfrac{x}{t} > \lambda_k(u_L) \end{cases}$$

satisfies the Rankine-Hugoniot jump condition: Because of

$$\begin{aligned} \frac{d}{d\xi}\left[f(v_k(\xi)) - \lambda_k(v_k(\xi))v_k(\xi)\right] &= f'(v_k(\xi))v_k'(\xi) - \lambda_k(v_k(\xi))v_k'(\xi) \\ &= \lambda_k(v_k)r_k(v_k) - \lambda_k(v_k)r_k(v_k) = 0, \end{aligned}$$

it follows that

$$f(u_L) - \lambda_k(u_L)u_L = f(u_R) - \lambda_k(u_R)u_R$$

and, therefore,

$$s\,(u_R - u_L) = f(u_R) - f(u_L)$$

with $s = \lambda_k(u_L) = \lambda_k(u_R)$. Obviously, in this case the curve $\mathcal{S}_k(u_L)$ can be parameterized by $\xi \mapsto v_k(\xi)$, $\xi \in \mathbb{R}$.

Let (U, F) be an entropy pair. Then

$$\begin{aligned}\frac{d}{d\xi}\left[F(v_k(\xi)) - \lambda_k(v_k(\xi))U(v_k(\xi))\right] &= F'(v_k(\xi))v_k'(\xi) - \lambda_k(v_k(\xi))U'(v_k(\xi))v_k'(\xi) \\ &= U'(v_k(\xi))f'(v_k(\xi))v_k'(\xi) - \lambda_k(v_k(\xi))U'(v_k(\xi))v_k'(\xi) \\ &= U'(v_k(\xi))\left[\lambda_k(v_k)r_k(v_k) - \lambda_k(v_k)r_k(v_k)\right] = 0.\end{aligned}$$

it follows that

$$F(u_L) - \lambda_k(u_L)U(u_L) = F(u_R) - \lambda_k(u_R)U(u_R)$$

and, therefore,

$$s\,(U(u_R) - U(u_L)) = F(u_R) - F(u_L)$$

with $s = \lambda_k(u_R) = \lambda_k(u_R)$. So the piecewise constant solution described above is also an entropy solution. It is called a k-contact discontinuity.

Example:

1. The linear wave equation $u_t + a\,u_x =$ is linear with the constant eigenvalue $\lambda_1 \equiv a$ and the corresponding characteristic field $r_1 \equiv 1$, which is, of course, linearly degenerate. The discontinuous solutions of the Riemann problem are contact discontinuities.

2. Burgers' equation $u_t + (u^2/2)_x = 0$ has the eigenvalue $\lambda_1(u) = u$ with corresponding constant characteristic field $r_1 \equiv 1$. Therefore

   $$\lambda_1'(u)r_1 = 1.$$

   So, the characteristic field is genuinely nonlinear. As already discussed we obtain rarefaction waves and shock waves as solutions of the Riemann problem.

3. The one-dimensional Euler equations have eigenvalues $\lambda_1(u) = v - c$, $\lambda_2(u) = v$ and $\lambda_3(u) = v + c$ with $c = \sqrt{\kappa p/\rho}$. The 1-characteristic field r_1 and the 3-characteristic field r_3 are genuinely nonlinear, the 2-characteristic field r_2 is linearly degenerate. So, particular solutions of Riemann problems are 1-shock waves, 1-rarefaction waves, 2-contact discontinuities, 3-shock waves, and 3-rarefaction waves.

Riemann Invariants

An important concept for computing especially k-rarefaction waves and k-contact discontinuities is the Riemann invariant:

Definition 4.4. *A function $w\colon D \longrightarrow \mathbb{R}$ is called a k-Riemann invariant, if and only if*

$$\nabla w(u)^T r_k(u) = 0 \quad \textit{for all } u \in D.$$

For a linearly degenerate k-characteristic field r_k λ_k is a k-Riemann invariant.

Let w be a k-Riemann invariant and let $v : \mathbb{R} \longrightarrow \mathbb{R}^p$ be a curve with

$$v'(\xi) = r_k(v(\xi)). \tag{4.1}$$

Then w is constant along this curve:

$$\frac{d}{dt} w(v(\xi)) = \nabla w(v(\xi))^T v'(\xi) = \nabla w(v(\xi))^T r_k(\xi) = 0.$$

It can be shown that there exist locally $p-1$ k-Riemann invariants whose gradients are linearly independent.

Example: For the one-dimensional Euler equations for a perfect gas we have the following pairs ($p-1=2$) of Riemann invariants: $v+\ell$ and s are 1-Riemann invariants whose gradients are linearly independent, v and p are 2-Riemann invariants whose gradients are linearly independent, and $v-\ell$ and s are 3-Riemann invariants whose gradients are linearly independent, where ℓ is given by

$$\ell = \frac{2c}{\kappa - 1}.$$

The computation of a k-rarefaction or a k-contact discontinuity wave requires a curve v_k satisfying (4.1). Since v_k must be constant for each of the $p-1$ k-Riemann invariants, the computation of v_k is a purely algebraic problem. The computation of a k-shock wave requires the solution of the Rankine-Hugoniot jump conditions, which is also a purely algebraic problem.

The following general statement on the solution of the Riemann problem for systems of conservation laws can be shown:

Theorem 4.1. *Assume that, for all $k = 1, \dots, p$ the k-th characteristic field is either genuinely nonlinear or linearly degenerate. Then, for each state u_L, there is a neighborhood $\mathcal{U}(u_L)$ of u_L such that, for each state $u_R \in \mathcal{U}(u_L)$, the Riemann problem has a weak solution, which consists of at most $p+1$ constant states, separated either by rarefaction waves, admissible shock waves or contact discontinuities.*

Example: The solution of the Riemann problem for the one-dimensional Euler equations for a perfect gas with

$$u_L = \begin{pmatrix} \rho_L \\ v_L \\ p_L \end{pmatrix} \qquad u_R = \begin{pmatrix} \rho_R \\ v_R \\ p_R \end{pmatrix},$$

where $v_L = v_R = 0$ and $p_L > p_R$ (shock tube problem) consists of 4 constant states

$$u_L, \qquad u_L^* = \begin{pmatrix} \rho_L^* \\ v^* \\ p^* \end{pmatrix}, \qquad u_R^* = \begin{pmatrix} \rho_R^* \\ v^* \\ p^* \end{pmatrix}, \qquad u_R.$$

The state u_L and u_L^* are separated by a 1-rarefaction wave ($\lambda_1 = v - c$), the states u_L^* and u_R^* are separated by a 2-contact discontinuity ($\lambda_2 = v$), and the states u_R^* and u_R are separated by a 3-shock wave ($\lambda_1 = v + c$). The values ρ_L^*, ρ_R^*, v^* and p^* follow from the solution of a purely algebraic system of equations.

4.3 Conservative Methods for Systems of Conservation Laws

The Lax-Friedrichs method and the Lax-Wendroff method can directly be extended to systems.

On the basis of the solution of Riemann problems Godunov's method can also be formulated for systems.

For linear fluxes $f(u) = A\,u$ with constant coefficients we obtain the following numerical flux for Godunov's method:

$$g_G(v,w) = f(u^*(0;v,w)) = A\,u^*(0;v,w).$$

Assume that (for simplicity) $\lambda_l < 0 < \lambda_{l+1}$ for some $l \in \{0,1,\ldots,p\}$ with $\lambda_0 = -\infty$ and $\lambda_{p+1} = +\infty$. Then

$$u(0;u_L,u_R) = w_l = \sum_{k=1}^{l} \beta_k r_k + \sum_{k=l+1}^{p} \alpha_k r_k$$

with

$$\alpha_k = l_k^T u_L, \quad \beta_k = l_k^T u_R.$$

Therefore,

$$\begin{aligned} A w_l &= \sum_{k=1}^{l} \beta_k \lambda_k r_k + \sum_{k=l+1}^{p} \alpha_k \lambda_k r_k \\ &= \sum_{k=1}^{p} \beta_k \lambda_k^- r_k + \sum_{k=1}^{p} \alpha_k \lambda_k^+ r_k \\ &= \sum_{k=1}^{p} \lambda_k^- r_k l_k^T u_R + \sum_{k=1}^{p} \lambda_k^+ r_k l_k^T u_L = A^- u_R + A^+ u_L \end{aligned}$$

with

$$A^\pm = R\Lambda^\pm R^{-1}, \quad \Lambda^\pm = \operatorname{diag}(\lambda_1^\pm,\ldots,\lambda_p^\pm).$$

So

$$g_G(v,w) = A^+ v + A^- w.$$

Observe

$$A^- + A^+ = R\Lambda R^{-1} = A, \quad \Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_p).$$

Possible extensions for nonlinear flux vectors of the form $f(u) = A(u)u$ leads to partial upwinding:

$$g(v,w) = A(\frac{v+w}{2})^+ v + A(\frac{v+w}{2})^- w$$

or full upwinding:

$$g(v,w) = A(v)^+ v + A(w)^- w.$$

Roe's approximative Riemann solver can also be formulated. As in the linear case the flux is linearized:

$$f(u) \approx \hat{f}(u) = A(u_L, u_R)\, u$$

where the matrix $A(u_L, u_R)$ has to satisfy three conditions:

1. $A(u_L, u_R)\,(u_L - u_R) = f(u_L) - f(u_R)$.
2. The linearized problem $u_t + A(u_L, u_R)\, u_x = 0$ is hyperbolic.
3. $A(u, u) = f'(u)$.

The construction of a Roe matrix $A(u_L, u_R)$ with these properties is nontrivial for systems. A Roe matrix for the Euler equation is known.

The extension of these methods to multidimensional systems of the form

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{d} \frac{\partial}{\partial x_i} f_i(u) = 0$$

is straight forward. Such a nonlinear system of conservation laws is called hyperbolic if and only if

$$A(u, \omega) = \sum_{i=1}^{d} \omega_i A_i(u) \quad \text{with } A_i(u) = f_i'(u)$$

has p real eigenvalues $\lambda_k(u, \omega)$ and p corresponding linearly independent eigenvectors $r_k(u, \omega)$ for all $u \in D$ and $\omega \in \mathbb{R}^p$.

For example, a finite volume method has the form

$$u_j^{n+1} \;=\; u_j^n - \frac{\Delta t}{|T_j|} \sum_{k \in N(j)} g_{jk}(u_j^n, u_k^n, n_{jk})|S_{jk}|,$$

where the numerical flux $g_{jk}(u_j^n, u_k^n, n_{jk})$ is an approximation of the normal component of the physical flux $f(u, n) = \sum_{i=1}^{d} n_i f_i(u)$ on the edge (face) S_{jk}. The corresponding Jacobian is given by

$$A(u, n) = \sum_{i=1}^{d} n_i A_i(u) \quad \text{with } A_i(u) = f_i'(u).$$

The one-dimensional methods discussed above applied to the flux $f(u, n)$ with Jacobian $A(u, n)$ can be used for constructing appropriate numerical fluxes.

Chapter 5

An Introduction to Boundary Conditions

Linear one-dimensional scalar conservation laws with constant coefficients

The linear wave equation on a bounded interval, say $(0,1)$, with the usual initial condition, given by

$$\begin{aligned} u_t + a\,u_x &= 0 \quad x \in (0,1),\ t > 0 \\ u(x,0) &= u_0(x) \quad x \in (0,1), \end{aligned}$$

has a unique solution if the following boundary conditions are prescribed:

$$u(0,t) = g(t) \quad \text{for all } t > 0, \quad \text{in the case } a > 0,$$

$$u(1,t) = g(t) \quad \text{for all } t > 0, \quad \text{in the case } a < 0.$$

This easily follows by considering the characteristic curves.

Linear multidimensional scalar conservation laws with constant coefficients

The extension to the multidimensional case

$$\begin{aligned} \frac{\partial u}{\partial t} + \sum_{i=1}^{d} a_i \frac{\partial u}{\partial x_i} &= 0 \quad x \in \Omega,\ t > 0 \\ u(x,0) &= u_0(x) \quad x \in (0,1) \end{aligned}$$

is straight forward and leads to the following boundary condition:

$$u(x,t) = g(x,t) \quad x \in \Gamma_-,\ t > 0,$$

where Γ_- is given by

$$\Gamma_- = \{x \in \partial\Omega : a \cdot n(x) < 0\}.$$

Linear one-dimensional systems of conservation laws with constant coefficients

Consider a hyperbolic system of conservation laws with initial condition, given by

$$\begin{aligned} u_t + A\,u_x &= 0 \quad x \in (0,1),\ t > 0 \\ u(x,0) &= u_0(x) \quad x \in (0,1). \end{aligned}$$

If the (characteristic) variables $v = Lu = R^{-1}u$ are introduced, we obtain p decoupled conservation laws

$$\frac{\partial v_i}{\partial t} + \lambda_i \frac{\partial v_i}{\partial x} = 0.$$

From the discussion before it is clear that the solution is uniquely determined if the following boundary conditions are prescribed:

$$\begin{aligned} v_-(0,t) &= g_0(t) \quad t > 0, \\ v_+(1,t) &= g_1(t) \quad t > 0. \end{aligned}$$

Here v_- and v_+ denote the vectors consisting of those components of v with indices i for which $\lambda_i < 0$ and $\lambda_i > 0$, respectively. So the values of those components are prescribed for which the corresponding characteristic curve is ingoing.

More generally, it is reasonable to prescribe the values of the ingoing characteristic variables in terms of the outgoing characteristic variables. This leads to the following boundary conditions of a more general form:

$$\begin{aligned} v_-(0,t) &= S_0(t)\,v_+(0,t) + g_0(t) \quad t > 0, \\ v_+(1,t) &= S_1(t)\,v_-(0,t) + g_1(t) \quad t > 0. \end{aligned}$$

Next we consider boundary conditions for the original variables u of the form

$$\begin{aligned} B_0 u(0,t) &= g_0(t) \quad t > 0, \\ B_1 u(1,t) &= g_1(t) \quad t > 0. \end{aligned}$$

Let R_- and R_+ be the matrices of all column vectors of R which correspond to negative eigenvalues and positive eigenvalues of A, respectively. Then it is immediately clear that these boundary conditions can be transformed in the boundary conditions for the characteristic variables if $B_0 R_-$ and $B_1 R_+$ are nonsingular square matrices.

For one-dimensional linear systems with constant coefficients it can be shown that the initial-boundary value problems discussed above are well-posed.

It is immediately clear how to extend the boundary conditions to the general multidimensional case:

General multidimensional systems of conservation laws

Consider a general system of conservation laws on a domain $\Omega \subset \mathbb{R}^d$ with initial conditions, given by

$$\begin{aligned} \frac{\partial u}{\partial t} + \sum_{i=1}^{d} \frac{\partial}{\partial x_i} f_i(u) &= 0 \quad x \in \Omega,\ t > 0 \\ u(x,0) &= u_0(x) \quad x \in (0,1). \end{aligned}$$

Let $x \in \Gamma = \partial\Omega$. Then the number of ingoing characteristic curves is given by the number of negative eigenvalues $\lambda_k(u(x,t), n(x))$ of the matrix $A(u(x,t), n(x))$. Therefore, we need the corresponding number of boundary data counted componentwise.

Example: The multidimensional Euler equations for a perfect gas are a system of 5 conservation laws. The system is hyperbolic. The eigenvalues of $A(u,n)$ are $\lambda_1 = v \cdot n - c$, $\lambda_2 = \lambda_3 = \lambda_4 = v \cdot n$ and $\lambda_5 = v \cdot n + c$ with $c = \sqrt{\kappa p/\rho}$. Several important cases are discussed now:

1. At an inlet, i.e. $v \cdot n < 0$, with a subsonic flow, i.e. $|v \cdot n| < c$ we have 4 negative eigenvalues and 1 positive eigenvalue. One could, e.g., prescribe values for ρ and v.

2. At an inlet, i.e. $v \cdot n < 0$, with a supersonic flow, i.e. $|v \cdot n| > c$ we have 5 negative eigenvalues. One prescribes the whole vector u, e.g., in terms of prescribed values for ρ, v and p.

3. At an outlet, i.e. $v \cdot n > 0$, with a subsonic flow, i.e. $|v \cdot n| < c$ we have 1 negative eigenvalue and 4 positive eigenvalues. One could, e.g., prescribe values for p.

4. At an outlet, i.e. $v \cdot n > 0$, with a supersonic flow, i.e. $|v \cdot n| > c$ we have no negative eigenvalue. No boundary conditions are allowed.

5. At a fixed wall, i.e. $v \cdot n = 0$, we have 1 negative eigenvalue and 1 positive eigenvalue. One usually prescribes the condition $v \cdot n = 0$.

WARNING: It is highly non-trivial how to guarantee well-posedness of the multidimensional initial-boundary value problem for systems, even for linear systems with constant coefficients.

Numerical treatment of boundary conditions

Consider a finite volume method. The numerical fluxes at interior edges (faces) can be computed from the two adjacent cells by one of the discussed methods. It remains to determine the numerical fluxes at boundary edges (faces). The boundary conditions provide, in general, only partial information on the values of the variables at the boundary (only one boundary data per ingoing characteristic curve). Boundary data per outgoing characteristic curve is provided by the values from the interior of the domain. Therefore, it is reasonable to use extrapolation from interior values to obtain approximate boundary data for each outgoing characteristic curve.

Example:

1. At an inlet with a subsonic flow we have 1 positive eigenvalue. One could, e.g., compute values for p by extrapolation.

2. At an inlet with a supersonic flow everything is prescribed.

3. At an outlet with a subsonic flow we have 4 positive eigenvalues. One could, e.g., compute values for ρ and v by extrapolation.

4. At an outlet with a supersonic flow we have 5 positive eigenvalues. All values are computed by extrapolation.

5. At a fixed wall we have 1 positive eigenvalue. One could, e.g., compute p by extrapolation.

Bibliography

[1] Miloslav Feistauer. *Mathematical methods in fluid dynamics.* London: Longman Scientific & Technical. New York: Wiley, 1993.

[2] Edwige Godlewski and Pierre-Arnaud Raviart. *Hyperbolic systems of conservation laws.* Mathématiques & Applications. 3-4. Paris: Ellipses, 1991.

[3] Edwige Godlewski and Pierre-Arnaud Raviart. *Numerical approximation of hyperbolic systems of conservation laws.* Applied Mathematical Sciences. 118. New York, NY: Springer, 1996.

[4] Heinz-Otto Kreiss and Jens Lorenz. *Initial-boundary value problems and the Navier-Stokes equations.* Pure and Applied Mathematics, 136. Boston, MA: Academic Press, Inc., 1989.

[5] Kröner, Dietmar. *Numerical schemes for conservation laws.* Wiley-Teubner Series Advances in Numerical Mathematics. Chichester: Wiley. Stuttgart: Teubner, 1997.

[6] Randall J. LeVeque. *Numerical methods for conservation laws.* Lectures in Mathematics, ETH Zürich. Basel: Birkhäuser Verlag, 1990.