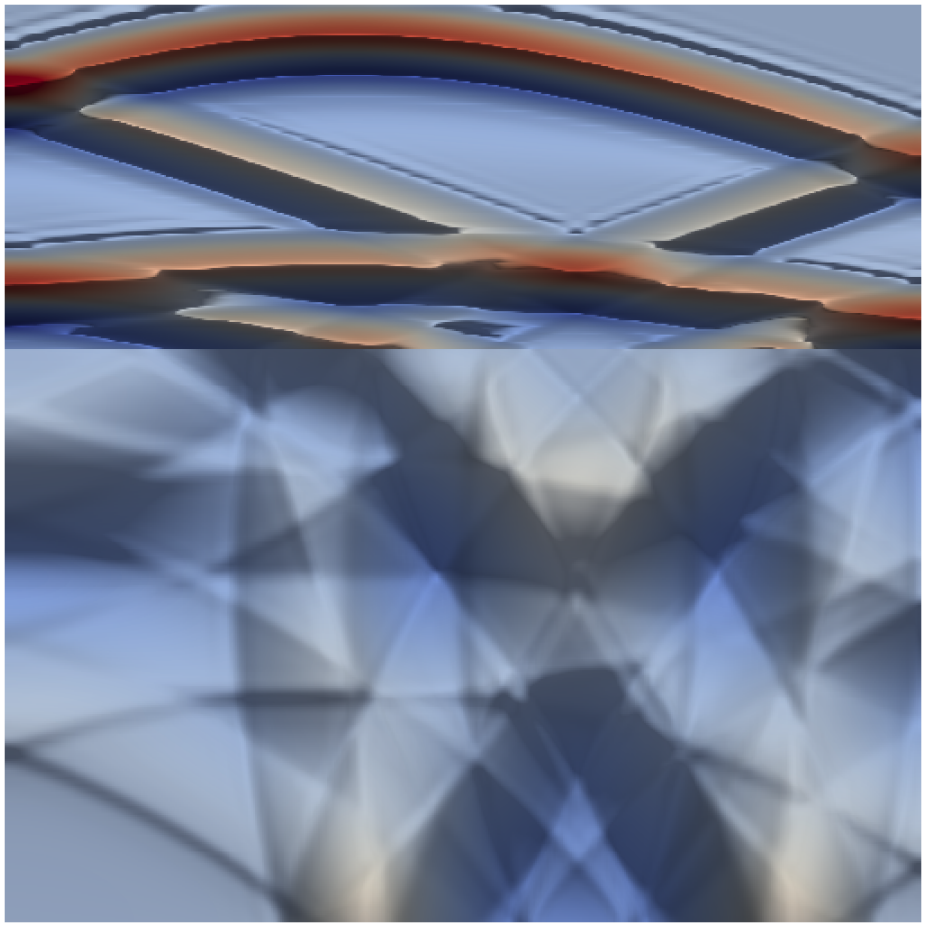
Short Course on

Numerical Solution of Hyperbolic Partial Differential Equations

Peter Bastian
Interdisziplinäres Zentrum für Wissenschaftliches Rechnen
Universität Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg
Peter.Bastian@iwr.uni-heidelberg.de

October 3, 2017



Contents

Chapter 1

Introduction

1.1 Hyperbolicity

In this course we are interested in the numerical solution of first-order hyperbolic partial differential equations (PDEs) which, in their general conservative form, are given by

$$\partial_t u(x,t) + \nabla \cdot F(u(x,t),x,t) + g(u(x,t),x,t) = 0 \quad \text{in } U = \Omega \times \Sigma. \tag{1.1}$$

For the theoretical treatment in this chapter $\Omega = \mathbb{R}^d$, $d \in \mathbb{N}$, is the unbounded spatial domain and $\Sigma = \mathbb{R}^+$ is the unbounded temporal domain. Later, in the practical treatment we will treat finite domains (which adds the difficult point of boundary conditions). Equation (1.1) is supplemented with initial conditions

$$u(x,0) = u_0(x).$$

A *classical solution* of the PDE (1.1) is a vector-valued, differentiable function $u : \Omega \times \Sigma \to \mathbb{R}^m$ with $m \in \mathbb{N}$ that satisfies the partial differential equation (1.1) in every point $(x,t) \in U$. The matrix-valued function $F : \mathbb{R}^m \times \Omega \times \Sigma \to \mathbb{R}^{m \times d}$ with the columns $F(u,x,t) = [F_1(u,x,t), \ldots, F_n(u,x,t)]$ is called *flux function*. Note that the divergence is defined as $\nabla \cdot F(u(x,t),x,t) = \sum_{j=1}^{d} \partial_{x_j} F_j(u(x,t),x,t)$.

Equation (1.1) is said to be in *conservative form* as it arises naturally from the formulation of conservation of mass, momentum and energy. If the flux function is smooth enough, the PDE can be put in its *non-conservative* or *quasi-linear* form which reads

$$\partial_t u(x,t) + \sum_{j=1}^{d} B_j(u(x,t),x,t)\partial_{x_j} u(x,t) + \tilde{g}(u(x,t),x,t) = 0 \quad \text{in } \Omega \times \Sigma. \tag{1.2}$$

The reason is the chain rule

$$\partial_{x_j} F_{i,j}(u(x,t),x,t) = \sum_{k=1}^{m} \frac{\partial F_{i,j}}{\partial u_k}(u(x,t),x,t)\frac{\partial u_k}{\partial x_j}(x,t) + \frac{\partial F_{i,j}}{\partial x_j}(u(x,t),x,t)$$

which shows

$$(B_j(u,x,t))_{i,k} = \frac{\partial F_{i,j}}{\partial u_k}(u,x,t), \qquad \tilde{g}_i(u,x,t) = g_i(u,x,t) + \frac{\partial F_{i,j}}{\partial x_j}(u,x,t).$$

In its most general form equation (1.1) is very difficult to solve. In the following we discuss a number of important special cases:

- The case $m = 1$, i.e. a single component, is called the scalar case whereas $m > 1$ indicates a system of equations.

- The case $d = 1$ is called one-dimensional while $d > 1$ is the multi-dimensional case.

- If the matrix-valued functions B_j are independent of u the PDE is called linear, otherwise if B_j depends on u it is called non-linear. In the linear case, if B_j is also independent of x and t the PDE has constant coefficients otherwise it has variable coefficients. If the dependence on x and t is not continuous we say that the PDE has discontinuous coefficients.

It turns out that many systems of the form (1.2) which are of practical interest satisfy an important property that is essential in the theoretical and numerical treatment.

Definition 1.1 (Hyperbolic First-Order PDE)**.** The system of equations (1.2) is called *hyperbolic* if for each feasible state $u \in \mathbb{R}^m$, $x \in \Omega$, $t \in \Sigma$ and $y \in \mathbb{R}^d$ the $m \times m$ matrix

$$B(u, x, t; y) = \sum_{j=1}^{d} y_j B_j(u, x, t) \tag{1.3}$$

is real diagonalizable, i.e. $B(u, x, t; y)$ has m real eigenvalues $\lambda_1(x, t; y), \ldots, \lambda_m(x, t; y)$ and its corresponding right eigenvectors $r_1(x, t; y), \ldots, r_m(x, t; y)$ form a basis of $\mathbb{R}^m$. In addition there are the special cases:

i) The system is called *symmetric hyperbolic* if $B_j(u, x, t)$ is symmetric for every feasible state $u \in \mathbb{R}^m$, $x \in \Omega$, $t \in \Sigma$ and $j = 1, \ldots, m$.

ii) The system is called *strictly hyperbolic* if all m eigenvalues are distinct for every feasible state $u \in \mathbb{R}^m$, $x \in \Omega$, $t \in \Sigma$. □

Note that the definition of hyperbolicity relies on the non-conservative form.

1.2 Classical Solutions and the Method of Characteristics

In this chapter we turn to classical solutions of the scalar, quasi-linear form.

Theorem 1.2 (Method of Characteristics (nonlinear case))**.** Let $U = \mathbb{R} \times \mathbb{R}^+$ be the upper half plane, i.e. we restrict ourselves to one space dimension $d = 1$. Let $u : U \to \mathbb{R}$ be a classical solution of the quasi-linear first-order partial differential equation

$$\partial_t u(x,t) + v(u(x,t))\partial_x u(x,t) = 0 \quad (x,t) \in U$$

subject to the initial condition

$$u(x,0) = u_0(x).$$

Then $u(x,t)$ is constant along the characteristic curves $X(t;\xi)$, $X : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}$, given by the ordinary differential equation

$$\frac{dX}{dt}(t;\xi) = v(u(X(t;\xi),t)) \ (t > 0), \quad X(0;\xi) = \xi \ (t = 0). \tag{1.4}$$

Moreover, the characteristic curves are straight lines of the form

$$X(t;\xi) = \xi + tv(u_0(\xi)).$$

Proof. Differentiate u along the characteristic curve:

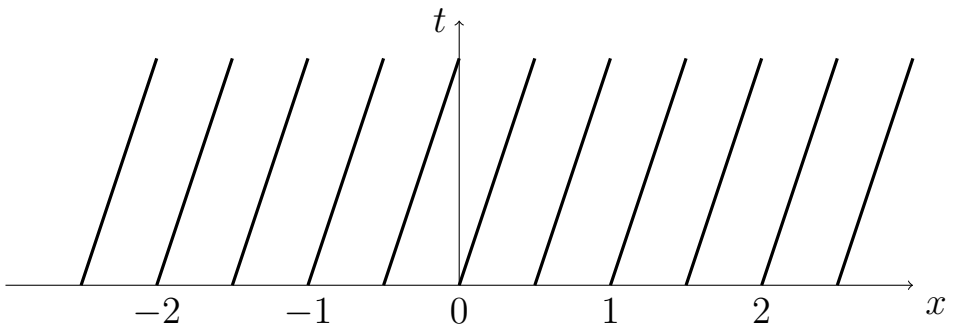
$$\begin{aligned}\frac{d}{dt}u(X(t;\xi),t) &= \frac{\partial u}{\partial t}(X(t;\xi),t) + \frac{\partial u}{\partial x}(X(t;\xi),t)\frac{dX}{dt}(t;\xi) \\ &= \frac{\partial u}{\partial t} + v(u(X(t;\xi),t))\partial_x u(X(t;\xi),t) = 0.\end{aligned}$$

Since the characteristic curve starts at $\xi \in \mathbb{R}$ at time $s = 0$, u has the constant value $u_0(\xi)$ along the characteristic curve and the solution of the ordinary differential equation (1.4) is $X(t;\xi) = \xi + v(u_0(\xi))s$. □

Example 1.3. Let us first consider the simple linear, one-dimensional example

$$\partial_t u(x,t) + a\partial_x u(x,t) = 0, \quad (x,t) \in U = \mathbb{R} \times \mathbb{R}^+$$

with the initial condition $u(x,0) = u_0(x)$ and $a \in \mathbb{R}$. Then the characteristic curves are the straight lines $X(t;\xi) = \xi + at$ which are all parallel and have slope a. This is illustrated below for $a > 0$:



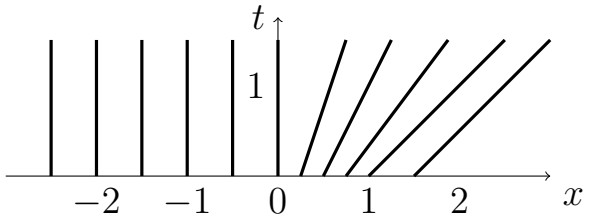
Consequently, the solution is given by $u(x,t) = u_0(x-at)$. It is interesting that this formula remains valid (in the sense of weak solutions introduced below) even if $u_0(x)$ is a discontinuous function which then gives rise to a discontinuous solution of the PDE.

Matters change dramatically when the PDE is nonlinear. Consider the PDE

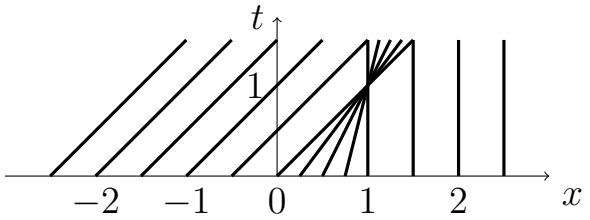
$$\partial_t u(x,t) + u(x,t)\partial_x u(x,t) = 0, \quad (x,t) \in U = \mathbb{R} \times \mathbb{R}^+$$

called the *inviscid Burgers' equation* (check that $\partial_t u(x,t) + \partial_x F(u(x,t)) = 0$ with $F(u) = \frac{1}{2}u^2$ is the conservative form of the equation). According to Theorem 1.2 the characteristic curves are straight lines given by $X(t;\xi) = \xi + u_0(\xi)t$. Depending on the initial condition, the characteristic curves may look as follows:

a)
$$u_0(x) = \begin{cases} 0 & x \le 0 \\ \sin^2\left(\frac{x\pi}{2}\right) & 0 < x < 1 \\ 1 & x \ge 1 \end{cases}$$



b)
$$u_0(x) = \begin{cases} 1 & x \le 0 \\ \sin^2\left(\frac{(1-x)\pi}{2}\right) & 0 < x < 1 \\ 0 & x \ge 1 \end{cases}$$



In the case a) the solution is uniquely defined for all times $t > 0$ and it is given by

$$u(x,t) = \begin{cases} 0 & x \le 0 \\ \Xi(x,t) & 0 < x < 1+t \\ 1 & x \ge 1+t \end{cases}, \quad \frac{x - \Xi(x,t)}{t} = \sin^2\left(\frac{\Xi(x,t)\pi}{2}\right).$$

Such a solution is called a *rarefaction wave*. In the case b), however the characteristic curves are intersecting in large parts of the right half plane. This is interpreted in that way that a classical solution does not exist. Instead the notion of solution has to be extended to so-called weak solutions. It turns out that weak solutions may exhibit discontinuities, also called *shocks*. Note that the problem of intersecting discontinuities occurs despite the fact that the initial condition is continuously differentiable.

Let us now turn to the linear case in more detail. In that case we may also treat the case with a source term.

Theorem 1.4 (Method of Characteristics (linear case))**.** Let $U = \mathbb{R}^d \times \mathbb{R}^+$ be the space-time domain and $u : U \to \mathbb{R}$ a classical solution of the linear first-order partial differential equation in conservative form

$$\partial_t u(x,t) + \nabla \cdot (v(x,t)u(x,t)) = 0 \quad (x,t) \in U \tag{1.5}$$

with $v \in \left[C^1(U)\right]^d$ a continuously differentiable vector field and subject to the initial condition

$$u(x,0) = u_0(x).$$

Then $u(x,t)$ has the value

$$u(X(t;\xi),t) = u_0(\xi) \exp\left(- \int_0^t \nabla \cdot v(X(s;\xi),s) ds \right).$$

along the characteristic curves $X(t;\xi)$, $X : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$, given by the ordinary differential equation

$$\frac{dX}{dt}(t;\xi) = v(X(t;\xi),t) \ (t > 0), \quad X(0;\xi) = \xi \ (t = 0). \tag{1.6}$$

The characteristic curves are not necessarily straight lines. If the characteristic curves do not intersect and if any point $(x,t) \in U$ can be reached from time $t = 0$ then the classical solution exists and is unique.

Proof. Using the fact that u is a classical solution and v is continuously differentiable we can write (1.5) in nonconservative form

$$\partial_t u(x,t) + v(x,t) \cdot \nabla u(x,t) + (\nabla \cdot v(x,t))u(x,t) = 0 \quad (x,t) \in U$$

and using the chain rule and the definition of the characteristic curves we obtain

$$\begin{aligned} \frac{d}{dt} u(X(t;\xi),t) &= \frac{\partial u}{\partial t}(X(t;\xi),t) + v(u(X(t;\xi),t)) \cdot \nabla u(X(t;\xi),t) \\ &= -(\nabla \cdot v(X(t;\xi),t))u(X(t;\xi),t). \end{aligned}$$

This ordinary differential equation has the form

$$\frac{d}{dt} u(X(t;\xi),t) = -a(t)u(X(t;\xi),t) \tag{1.7}$$

with $a(t) = \nabla \cdot v(X(t;\xi),t)$. The main theorem of calculus states that $g(t) = \left(\frac{d}{dt}\int_0^t g(s)ds\right)(t)$. With that we obtain

$$\frac{d}{dt}\left[u_0(\xi) \exp\left(- \int_0^t a(s) ds \right) \right] = u_0(\xi) \exp\left(- \int_0^t a(s) ds \right) (-a(t))$$

which is (1.7). □

Remark 1.5. Clearly, when the velocity field is divergence-free, i.e. $\nabla \cdot v(x,t) = 0$ for all times, then u is constant along characteristics.

1.3 Weak Solutions

The previous section showed that classical solution do not in general exist for all times for nonlinear hyperbolic PDEs. In order to study these PDEs the notion of a "solution" needs to be extended to allow also discontinuous functions. This is done by introducing weak solutions.

Definition 1.6. A weak solution of the general vector-valued first-order PDE

$$\partial_t u(x,t) + \nabla \cdot F(u(x,t)) = 0 \qquad (x,t) \in U = \mathbb{R}^d \times \mathbb{R}^+ \tag{1.8a}$$

$$u(x,0) = u_0(x) \qquad x \in \mathbb{R}^d \tag{1.8b}$$

is a function $u : U \to \mathbb{R}^m$ such that

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^+} u(x,t) \cdot \partial_t \phi(x,t) + F(u(x,t)) : \nabla\phi(x,t)\, dx\, dt + \int_{\mathbb{R}^d} u_0(x) \cdot \phi(x,0)\, dx = 0 \tag{1.9}$$

for all test functions $\phi \in (C_0^1(U))^m$ with

$$C_0^1(U) = \left\{ w \in C^1(U) \,:\, \exists r > 0 \text{ s.t. } \operatorname{supp} w \subset (B_r \cap U) \right\},$$

the support of a function defined as

$$\operatorname{supp} w = \overline{\{(x,t) \in U \,:\, w(x,t) \neq 0\}}$$

and $B_r = \{(x,t) \,:\, \|x\|^2 + t^2 < r^2\}$. Moreover, we introduced the notation $(\nabla\phi(x,t))_{i,j} = \frac{\partial \phi_i}{\partial x_j}(x,t)$ (gradient) and $A : B = \sum_{i=1}^m \sum_{j=1}^d (A)_{ij} B_{ij}$. □

Weak solutions are an extension of the concept of classical solution in the following sense.

Theorem 1.7. Let $u(x,t)$ be a classical solution of (1.8). Then it is also a weak solution.

Proof. Let $\phi(x,t)$ be a test function such that $\operatorname{supp}(\phi) \subset B_r$ for some $r > 0$ and for any $x \in \mathbb{R}^d$ with $\|x\| < r$ set $T(x) = \sqrt{r^2 - \|x\|^2}$. Using integration by

parts and carefully exploiting that $\phi(x,t) = 0$ for $\|x\| = \sqrt{r^2 - t^2}$ we obtain

$$
\begin{aligned}
0 = \int_{\mathbb{R}^d}\int_{\mathbb{R}^+} \left[\partial_t u(x,t) + \sum_{j=1}^{d} \partial_{x_j} F_j(u(x,t))\right] \cdot \phi(x,t)\, dt\, dx = \\
\int_{\mathbb{R}^d} \sum_{i=1}^{d} \int_{\mathbb{R}^+} \partial_t u_i(x,t)\phi_i(x,t)\, dt\, dx \\
+ \int_{\mathbb{R}^+} \sum_{i=1}^{d}\sum_{j=1}^{d} \int_{\mathbb{R}^d} \partial_{x_j} F_{i,j}(u(x,t))\phi_i(x,t)\, dx\, dt = \\
\int_{\|x\|<r} \sum_{i=1}^{d} \int_{0}^{T(x)} \partial_t u_i(x,t)\phi_i(x,t)\, dt\, dx \\
+ \int_{0}^{r} \sum_{i=1}^{d}\sum_{j=1}^{d} \int_{\|x\|<\sqrt{r^2-t^2}} \partial_{x_j} F_{i,j}(u(x,t))\phi_i(x,t)\, dx\, dt = \\
\int_{\|x\|<r} \sum_{i=1}^{d} \left\{ - \int_{0}^{T(x)} u_i(x,t)\partial_t\phi_i(x,t)\, dt + \left[u_i(x,t)\phi_i(x,t)\right]_0^T(x) \right\} dx \\
- \int_{0}^{r} \sum_{i=1}^{d}\sum_{j=1}^{d} \int_{\|x\|<\sqrt{r^2-t^2}} F_{i,j}(u(x,t))\partial_{x_j}\phi_i(x,t)\, dx\, dt = \\
- \int_{\mathbb{R}^d}\int_{\mathbb{R}^+} u(x,t)\cdot\partial_t\phi(x,t)\, dt\, dx - \int_{\mathbb{R}^d} u_0(x,t)\phi(x,0)\, dx \\
- \int_{\mathbb{R}^+}\int_{\mathbb{R}^d} F(u(x,t) : \nabla\phi(x,t)\, dx\, dt.
\end{aligned}
$$

This is the statement of the theorem. □

Rankine-Hugoniot Condition

The following theorem provides a statement about the discontinuity of a solution of a nonlinear hyperbolic system in one spatial dimension.

Theorem 1.8 (Rankine-Hugoniot)**.** Let u be a weak solution of (1.8) in one spatial dimension, i.e. $U = \mathbb{R}\times\mathbb{R}^+$. Assume there exists a curve $C = \{(x,t) \in U \,:\, x = X_C(t), t \in \mathbb{R}^+\}$ with a differentiable function $X_C(t)$ which divides U

into the parts $U = U^l \cup C \cup U^r$ with $U^l = \{(x,t) \in U : x < X_C(t)\}$ and $U^r = \{(x,t) \in U : x > X_C(t)\}$. In addition assume that

1. u is a classical solution in U^l as well as U^r,
2. u is discontinuous along the curve C and
3. the difference of left and right limit value of u along C is continuous.

Then

$$X_C'(t)\,[\![u]\!](X_C(t),t) = [\![F \circ u]\!](X_C(t),t) \tag{1.10}$$

where we defined the jump $[\![w]\!](x,t) = \lim_{\epsilon\to 0+} w(x-\epsilon,t) - \lim_{\epsilon\to 0+} w(x+\epsilon,t)$.

Proof. Recall Greens' theorem in the (x,t)-plane. Let D be a domain in the (x,t)-plane with boundary Γ. Then for sufficiently smooth functions $v(x,t), w(x,t)$ and $\phi(x,t)$ we have

$$\int_D (\partial_t v + \partial_x w)\phi \, dxdt = -\int_D v\partial_t\phi + w\partial_x\phi \, dxdt + \int_\Gamma (vn_t + wn_x)\phi \, ds$$

where $n = (n_t, n_x)^T$ is the unit outer normal at the boundary Γ.

Now take any test function ϕ and assume D is an open domain in U such that $\operatorname{supp}\phi \subset D$. This ensures that $\phi(x,0) = 0$ and we set $D^l = D \cap U^l$ and $D^r = D \cap U^r$ and $C_D = D \cap C$. Now

$$\begin{aligned}
0 = \int_D u(x,t)\cdot\partial_t\phi(x,t) + F(u(x,t))\cdot\partial_x\phi(x,t)\,dx\,dt = \\
\sum_{i=1}^m \int_{D^l} u_i^l(x,t)\partial_t\phi_i(x,t) + F_i(u^l(x,t))\partial_x\phi_i(x,t)\,dx\,dt \\
+ \sum_{i=1}^m \int_{D^r} u_i^r(x,t)\partial_t\phi_i(x,t) + F_i(u^r(x,t))\partial_x\phi_i(x,t)\,dx\,dt = \\
\sum_{i=1}^m \Big[-\int_{D^l} (\partial_t u_i^l(x,t) + \partial_x F_i(u^l(x,t)))\phi_i(x,t)\,dx\,dt \\
+ \int_{C_D} (u_i^l(X_C(t),t)n_t^l(t) + F_i(u^r(X_C(t),t))n_x^l(t))\phi(X_C(t),t)dt\Big] \\
+ \sum_{i=1}^m \Big[-\int_{D^r} (\partial_t u_i^r(x,t) + \partial_x F_i(u^r(x,t)))\phi_i(x,t)\,dx\,dt \\
+ \int_{C_D} (u_i^r(X_C(t),t)n_t^r(t) + F_i(u^r(X_C(t),t))n_x^r(t))\phi(X_C(t),t)dt\Big] =
\end{aligned}$$

$$\sum_{i=1}^{m} \int_{C_D} \left([\![u_i]\!](X_C(t),t) n_t^l(t) + [\![F_i(u(X_C(t),t))]\!] n_x^l(t) \right) \phi(X_C(t),t) dt$$
$$= \int_{C_D} \left([\![u]\!](X_C(t),t) \frac{X_C'(t)}{\|n^l(t)\|} + [\![F(u(X_C(t),t))]\!] \frac{1}{\|n^l(t)\|} \right) \cdot \phi(X_C(t),t) dt$$
$$= \int_{C_D} \left([\![u]\!](X_C(t),t) X_C'(t) + [\![F(u(X_C(t),t))]\!] \right) \cdot \phi(X_C(t),t) \|n^l(t)\|^{-1} dt.$$

Since the test function is arbitrary (in every component!) and $[\![u]\!]$ is continuous along the curve C the first factor under the integral needs to vanish (fundamental theorem of calculus). □

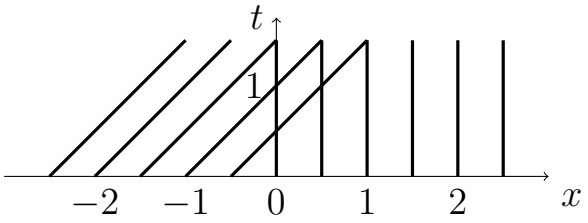
The Rankine-Hugoniot condition allows to compute the propagation speed of a discontinuity (shock). This is illustrated by the following example.

Example 1.9. Let us return to the one-dimensional, inviscid Burgers' equation we already treated in Example 1.3. It reads in conservative form:

$$\partial_t u(x,t) + \partial_x F(u(x,t)) = 0, \quad (x,t) \in U = \mathbb{R} \times \mathbb{R}^+,$$

with $F(u) = u^2/2$. We modify the second initial condition from above in the following discontinuous way:

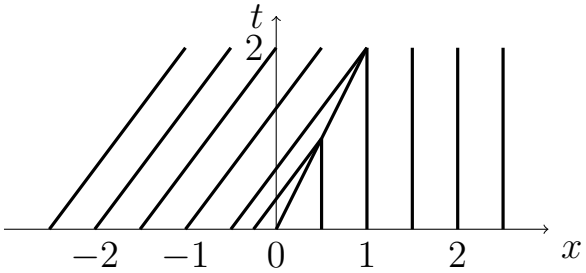
$$u_0(x) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases}$$



Now the characteristics are intersecting in the right half plane and the method of characteristics does not provide a solution there. From the Rankine-Hugoniot condition (1.10) we deduce that a weak solution should satisfy the condition

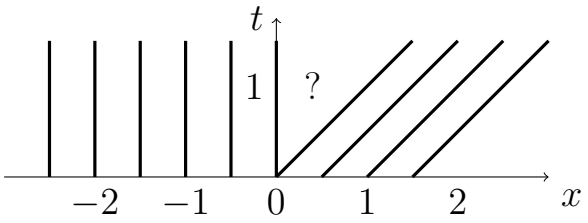
$$X_C'(t)\,(1-0) = \frac{1^2}{2} - \frac{0^2}{2} \qquad \Leftrightarrow \qquad X_C'(t) = \frac{1}{2} \tag{1.11}$$

which corresponds to the following characteristic diagram:



Now let us turn to the following initial condition

$$u_0(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$$



The method of characteristics does not give a value for the solution in the triangular shaped region indicated by the question mark. In fact it turns out that there are infinitely many weak solutions possible in that case.

The first possibility is a shock solution of the form

$$u(x,t) = \begin{cases} 0, & x \leq t/2 \\ 1, & x > t/2. \end{cases} \tag{1.12}$$

The shock speed is again $1/2$ as we have the Rankine-Hugoniot condition $X_C'(t)(0-1) = 0^2/2 - 1^2/2 = 1/2$. A second solution has the form

$$u(x,t) = \begin{cases} 0, & x \leq t/2 \\ x/t, & 0 < x \leq t \\ 1, & x > t. \end{cases} \tag{1.13}$$

and is called a *rarefaction wave*. It is continuous but not differentiable. One can show that a continuous function that is a piecewise classical solution is also a weak solution (combine Theorem 1.7 with Theorem 1.8 where the jump is now zero due to the fact that u is continuous along any curve C). Now the two solutions (1.12) and (1.13) can be combined to an infinity of solutions parametrized by $\gamma \in [0,1]$ with the idea of taking a rarefaction wave up to γ and then having a discontinuity up to the value one. From the Rankine-Hugoniot condition we obtain the shock speed as

$$X_C'(t)(\gamma - 1) = \gamma^2/2 - 1^2/2 \qquad \Leftrightarrow \qquad X_C'(t) = \frac{\gamma^2 - 1}{2(\gamma - 1)} = \frac{1+\gamma}{2},$$

and thus the solution family

$$u(x,t) = \begin{cases} 0, & x \leq 0 \\ x/t, & 0 < x \leq \gamma t \\ \gamma & \gamma t < x \leq \frac{1+\gamma}{2} t \\ 1, & \frac{1+\gamma}{2} t < x \end{cases} \tag{1.14}$$

Note that $(1+\gamma)/2 \geq \gamma$ for $\gamma \in [0,1]$.

Example 1.10. As a second example consider the linear hyperbolic system with $F(u) = Au$ for a real diagonalizable $m \times m$ matrix A. Then the Rankine-Hugoniot condition amounts to

$$X_C'(t)[\![u]\!](X_C(t),t) = [\![Au]\!](X_C(t),t) = A[\![u]\!](X_C(t),t)$$

which means that $X_C'(t)$ must be an eigenvalue of the matrix A and $[\![u]\!]$ is an eigenvector to the corresponding eigenvalue.

Selection Criteria

Example 1.9 shows that weak solutions may not be unique and there might be even infinetely many weak solutions. On the other hand we may assume that the physical problem to be modelled by a hyperbolic PDE, e.g. gas dynamics, has a unique solution. Thus, additional conditions need to be enforced on a weak solution to obtain the physically meaningful weak solution.

It turns out that there is no single selection criterion given the "right" physically meaningful weak solution for any nonlinear hyperbolic system of PDEs. They all agree on certain basic cases but may disagree on more complicated problems. Without going into details (as we will mostly concentrate on linear hyperbolic PDEs later) we state just some of them.

Lax Shock Condition Consider the first order nonlinear strictly hyperbolic PDE in the conservative form (1.8) in one dimension, i.e. $n = 1$ and $m \geq 1$

$$\partial_t u(x,t) + \partial_x F(u(x,t)) = 0.$$

Strict hyperbolicity implies that for every admissible state u the $m \times m$ matrix $\nabla F(u)$ has m distinct and real eigenvalues $\lambda_1(u) < \cdots < \lambda_m(u)$ as well as corresponding left amd right eigenvactors.

Definition 1.11. Let u be a piecewise classical solution with a discontinuity (in possibly every component) along a curve C given by $X_C(t)$ and we denote by u^l and u^r the corresponding left and right limiting values. Then the solution satisfies the Lax shock criterion and is called a k-shock if, for a fixed $1 \leq k \leq m$ and each point $p(t) = (X_C(t), t) \in C$ it satisfies the Rankine-Hugoniot condition as well as the following condition (see [7, Definition 3.24]):

$$\begin{aligned}\lambda_1(u^l(p)) < \cdots < \lambda_{k-1}(u^l(p)) < X_C'(t) < \lambda_k(u^l(p)) \quad &\text{and} \\ \lambda_k(u^r(p)) < X_C'(t) < \lambda_{k+1}(u^r(p)) < \cdots < \lambda_m(u^r(p)),&\end{aligned}$$

i.e. there are exactly $k-1$ eigenvalues at the left state smaller than $X_C'(t)$ and exactly $m-k$ eigenvalues at the right state greater than $X_C'(t)$.

For the scalar case $m = 1$ there are only 1-shocks as there is only one eigenvalue which is $F'(u)$. The Lax shock condition then formally reduces to

$$F'(u^r(p)) < X_C'(t) < F'(u^l(p)).$$

Since $X_C'(t)$ needs to satisfy the Rankine-Hugoniot condition we have

$$F'(u^r(p)) < \frac{F(u^l(p)) - F(u^r(p))}{u^l(p) - u^r(p)} < F'(u^l(p))$$

which means that the flux function needs to be either concave or convex. If $F(u)$ is in addition monotone increasing or decreasing we conclude $u^l > u^r$ which means we can only "jump down" across the shock.

For the inviscid Burger's equation this rules out the shock solution for the case where $u^l < u^r$ in the initial condition and leaves only the rarefaction wave as physically correct solution.

However it is clear that the Lax shock condition cannot be applied for nonconvexe (or nonconcave) flux functions (for example the Buckley-Leverett problem in oil recovery which has an S-shaped flux function). Therefore other criteria are needed.

Viscosity solution This selection criterion is based on the idea that in a physical system, e.g. gas dynamics, some form of disspation is present which ensures the uniqueness of the solution. In mathematical terms this can be put as follows.

Definition 1.12. u is called a *(vanishing) viscosity solution* of the problem (1.8) in one dimension if u can be obtained as the limit $\epsilon \to 0+$ of the parametrized problem

$$\partial_t u^\epsilon(x,t) + \partial_x F(u^\epsilon(x,t)) = \epsilon A \partial_{xx} u^\epsilon(x,t)$$

where A is positive definite matrix.

Making this condition rigorous requires a suitable definition of the limiting procedure. This condition is attractive since some numerical schemes can be shown to implicitly add the right hand side term with ϵ related to the mesh size h and thus for $h \to 0$ it provides the viscosity solution. For additional details and its relation to the entropy condition see [7, 1].

Entropy Condition This condition relies on an additional conservation law derived from the nonlinear strictly hyperbolic PDE in the conservative form (1.8) in one dimension.

Definition 1.13 (Entropy - entropy flux pair)**.** Two functions $U, Q : \mathbb{R}^m \to \mathbb{R}$ are called entropy - entropy flux pair if there holds

$$\nabla Q(u) = \nabla U(u)\ \nabla F(u).$$

Note that $\nabla Q(u) \in \mathbb{R}^{1\times m}$. It can be shown that if u is a classical solution then

$$\partial_t U(u(x,t)) + \partial_x Q(u(x,t)) = 0,$$

see [7, Section 3.6.1].

Definition 1.14. A weak solution of (1.8) in one dimension is said to satisfy the entropy condition if there exists an entropy - entropy flux pair with a convex function $U(u)$ such that

$$-\int_{\mathbb{R}^+}\int_{\mathbb{R}} U(u(x,t))\partial_t\phi(x,t) + Q(u(x,t))\partial_x\phi(x,t)\,dx\,dt \le 0$$

for every C^1 test function with compact support in the upper half plane.

One can show that this condition is equivalent to the Lax shock criterion under certain assumptions, see [7, Theorem 3.37]. On the other hand the entropy condition allows for more rigorous uniqueness and regularity results, see [1].

1.4 One-dimensional Linear Systems

We now turn back to the special case of strong solutions of one-dimensional linear systems with unknown function $u(x,t) = (u_1(x,t), \dots, u_m(x,t))^T$ given by

$$\partial_t u + B\partial_x u = 0 \qquad \text{in } \mathbb{R}^d \times \mathbb{R}^+ \tag{1.15a}$$

$$u(x,0) = u_0(x) \qquad x \in \mathbb{R} \tag{1.15b}$$

where B is a constant $m \times m$ matrix. Note that for B independent of x and t the conservative and nonconservative form are equaivalent. This system can be solved explicitly with the method of characteristics. (Note that Theorem 1.4 only considered the scalar but multi-dimensional case of a linear problem). The importance of the solution shown here is that it plays a crucial role in numerical methods for linear systems also in the multi-dimensional case.

The hyperbolicity of the system according to Definition 1.1 implies that B is real diagonalizable, i.e. B has m real eigenvalues $\lambda_1, \dots, \lambda_m$ and a corresponding set of right eigenvectors $r_1, \dots, r_m$ that form a basis of $\mathbb{R}^m$. Now from $Br_j = \lambda_j r_j$, $1 \le j \le m$, we can conclude that $BR = RD$ where $R = [r_1, \dots, r_m]$ and

$D = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$. Since R is invertible we have $R^{-1}BR = D$. Using the transformation $u = Rw$ we can transform the system (1.15) into the equivalent system

$$\partial_t w + D\partial_x w = 0 \qquad \text{in } \mathbb{R}^d \times \mathbb{R}^+$$

(insert $u = Rw$ and multiply with R^{-1} from the left). In the transformed system all components decouple and each component w_j can be solved independently using the method of characteristics which gives

$$w_j(x,t) = (w_0)_j(x - \lambda_j t), \qquad 1 \le j \le m,$$

where $w_0(x) = R^{-1}u_0(x)$ is the transformation of the initial condition. Each component of the solution of the original system is then a linear combination of these "simple" waves scaling the corresponding eigenvector:

$$u(x,t) = Rw(x,t) = R\sum_{j=1}^{m} e_j w_j(x,t) = \sum_{j=1}^{m} r_j w_j(x,t). \tag{1.16}$$

1.5 Connection to Second-order Hyperbolic Equations

A second-order linear hyperbolic partial differential equations in n space dimensions has the form

$$\partial_{tt}u(x,t) = \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}\partial_{x_i}\partial_{x_j}u(x,t) \qquad \text{in } \mathbb{R}^d \times \mathbb{R}^+ \tag{1.17}$$

with the matrix $A = (a_{ij})_{i,j=1}^d$ being positive definite. An example is the wave equation where $A = I$. Our aim is now to establish a connection to first-order hyperbolic systems.

Define $v : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^{n+1}$ as $v = (v_1, \ldots, v_n, v_{n+1})^T = (\partial_{x_1}u, \ldots, \partial_{x_n}u, \partial_t u)^T$ with $m = n + 1$ components. Then the second-order scalar equation (1.17) is equivalent to the following system of m equations:

$$\sum_{i=1}^{d} a_{ij}\partial_t v_i(x,t) - \sum_{i=1}^{d} a_{ij}\partial_{x_i}v_{n+1}(x,t) = 0 \qquad (j = 1, \ldots, n),$$

$$\partial_t v_{n+1}(x,t) - \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}\partial_{x_i}v_j(x,t) = 0.$$

Here the first n equations are a consequence of the n identities $\partial_t\partial_{x_i}u = \partial_{x_i}\partial_t u$ and the fact that the columns of A are linearly independent. The last equation

is our second-order hyperbolic PDE. Now this system can be written in matrix form as

$$B_0 \partial_t v(x,t) + \sum_{i=1}^{n} B_i \partial_{x_i} v(x,t) = 0. \tag{1.18}$$

with the matrices

$$B_0 = \begin{pmatrix} a_{11} & \dots & a_{1n} & 0 \\ \vdots & & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}, \qquad B_i = \begin{pmatrix} 0 & \dots & 0 & -a_{i1} \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & -a_{in} \\ -a_{i1} & \dots & -a_{in} & 0 \end{pmatrix}.$$

Since A is symmetric positive definite, B_0 is symmetric positive definite and obviously the B_i are symmetric. The system (1.18) is not in standard form unless $B_0 = I$. But we may transform (1.18) to a system in standard form as follows. Since B_0 is symmetric positive definite there exists an orthogonal matrix Q such that $Q^T B_0 Q = D = \operatorname{diag}(\mu_1, \dots, \mu_m)$ with $\mu_k > 0$. Moreover, since $\mu_k > 0$ we may define the matrices $D^{1/2} = \operatorname{diag}(\mu_1^{1/2}, \dots, \mu_m^{1/2})$ and $B_0^{1/2} = Q^T D^{1/2} Q$ such that $B_0 = B_0^{1/2} B_0^{1/2}$. Applying the transformation $w = B_0^{1/2} u$ the system (1.18) is equivalent to

$$\partial_t w + \sum_{i=1}^{d} B_0^{-1/2} B_i B_0^{-1/2} \partial_{x_i} w = 0 \qquad \text{in } \mathbb{R}^d \times \mathbb{R}^+.$$

Now observe

$$\tilde{B}(y) = \sum_{i=1}^{d} y_i B_0^{-1/2} B_i B_0^{-1/2} = B_0^{-1/2} \left(\sum_{i=1}^{d} y_j B_j \right) B_0^{-1/2} = \left(B_0^{-1/2} \right)^T B(y) B_0^{-1/2},$$

so $\tilde{B}(y)$ is symmetric if and only if $B(y)$ is symmetric. Since $B(y)$ is real and symmetric (the B_i are real and symmetric) it is diagonalizable (i.e. has a full set of right eigenvectors) and so is $\tilde{B}(y)$.

The rigidity theorem of Sylvester states that the signs of positive and negative eigenvalues (and therefore also the number of zero eigenvalues) of a real symmetric matrix A do not change under a transformation $S^T A S$ with any regular matrix S. One may also check that

$$B(y) = \sum_{i=1} y_i B_i = \begin{pmatrix} 0 & b(y) \\ b^T(y) & 0 \end{pmatrix}$$

has exactly two nonzero eigenvalues $\pm \|b(y)\|$ when $y \neq 0$ and all other eigenvalues are zero.

Chapter 2

Examples

2.1 Linear Transport

Scalar, multi-dimensional case We turn back to the linear transport equation

$$\partial_t u(x,t) + \nabla \cdot (v(x,t)u(x,t)) = 0, \qquad (x,t) \in U = \Omega \times \Sigma \tag{2.1}$$

now in a finite spatial domain $\Omega \subset \mathbb{R}^d$ and time interval $\Sigma = (t_0, t_0 + T)$. We assume that $v \in C^1(U)$ is a given continuously differentiable vector field. The method of characteristics from Theorem 1.4 can be extended to this case and this sheds light on the choice of boundary conditions. In the case of a finite domain the a characteristic curve starting at $(\xi, \tau) \in U$ and defined by

$$\frac{dX}{dt}(t;(\xi,\tau)) = -v(X(t;(\xi,\tau)),t) \ \ (t < \tau), \quad X(\tau;(\xi,\tau)) = \xi \ \ (t = \tau), \tag{2.2}$$

can be traced back either until $t = t_0$ or it stops early at $t > t_0$ and a point $x \in \Gamma^-(t) \subseteq \partial\Omega$. In the first case the value in (ξ, τ) depends on the inital condition

$$u(x, t_0) = u_0(x), \qquad (x \in \Omega)\,.$$

In the second case the value at (ξ, τ) is determined from a boundary condition

$$u(x,t) = g(x,t), \qquad ((x,t) \in \Gamma^-(t) \times \Sigma)\,.$$

From the construction it follows that $v(x,t) \cdot n(x) < 0$, with $n(x)$ the unit outer normal vector at $x \in \partial\Omega$, must hold. Otherwise the characteristic would not trace back to x. From this we conclude that boundary conditions can only be prescribed at the *inflow boundary*

$$\Gamma^-(t) = \{x \in \partial\Omega \,:\, v(x,t) \cdot n(x) < 0\}\,.$$

On all other points of the boundary no conditions can be prescribed.

Example 2.1. As an example consider the transport of a contaminant in soil shown in Figure 2.1. The top image shows the magnitude of the velocity field computed as the solution of the groundwater flow equation

$$\nabla v = f, \qquad v = \lambda \nabla u,$$

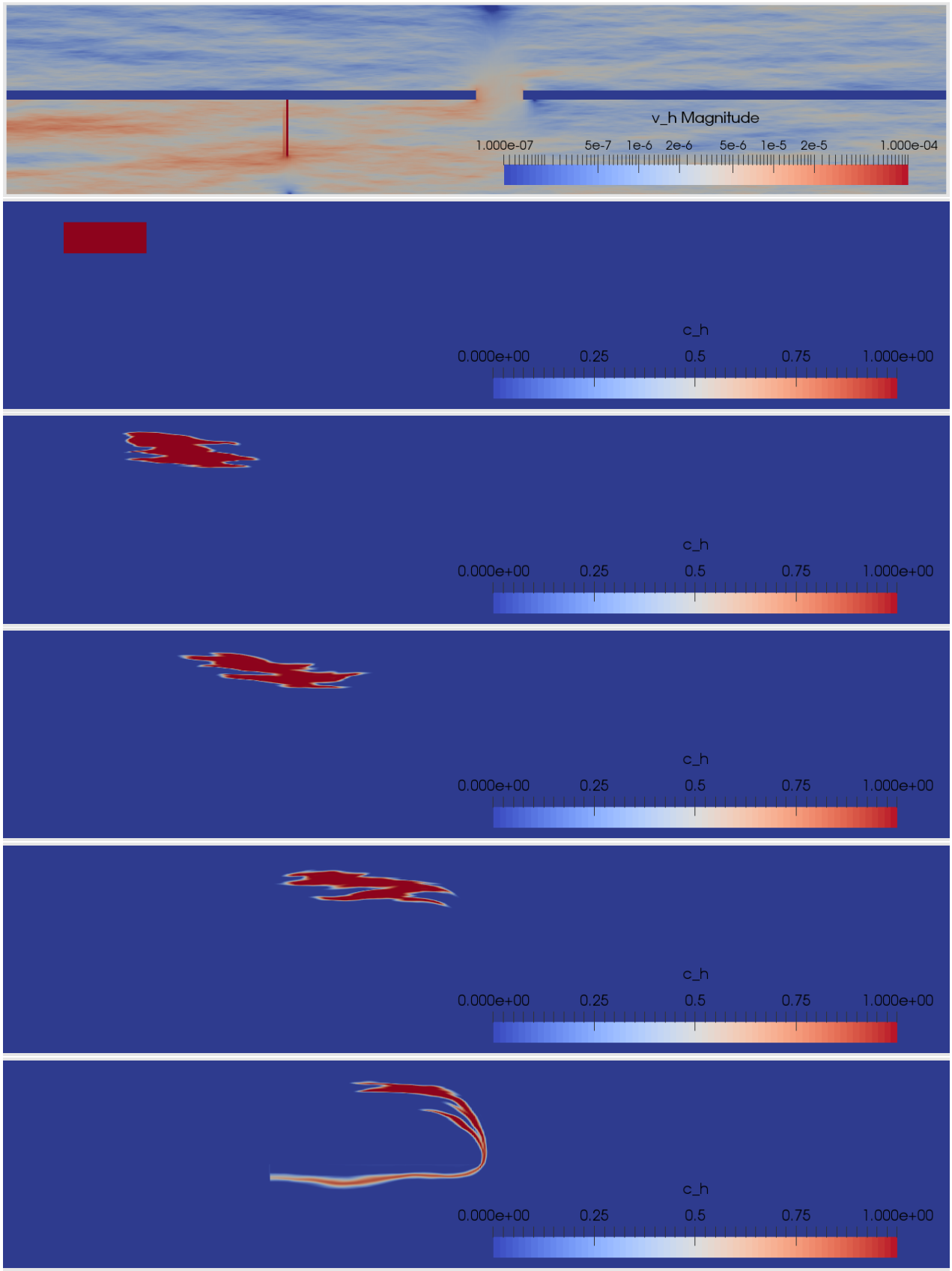


Figure 2.1: Transport of a contaminant in heterogeneous soil.

with heterogeneous mobility $\lambda(x)$. A contaminant is placed in the upper left corner (shown in the second image) while a pumping well is installed under the impermeable layer. The contaminant plume modelled by a first-order linear transport moves and is deformed until it is extracted in the well. Observe the complicated structure of the plume and the thin filament structure.

One-dimensional Systems In order to elaborate further on the aspect of boundary conditions consider the one-dimensional linear systems with m components $u(x,t) = (u_1(x,t), \dots, u_m(x,t))^T$ *in a finite domain*:

$$\partial_t u + B\partial_x u = 0, \qquad \text{in } \Omega \times \Sigma,$$

where B is a constant $m \times m$ matrix, $\Omega = (a,b)$ and $\Sigma = (t_0, t_0 + T)$. Following Section 1.4 this system is equivalent to the transformed system

$$\partial_t w + D\partial_x w = 0, \qquad \text{in } \Omega \times \Sigma,$$

where $D = \operatorname{diag}(\lambda_1, \dots, \lambda_m)$ are the eigenvalues of B. According to the considerations above, the boundary conditions for this system are

$$w_i(x,t) = \begin{cases} g_i(a,t) & \lambda_i > 0 \wedge x = a \\ g_i(b,t) & \lambda_i < 0 \wedge x = b \end{cases}. \tag{2.3}$$

Thus, the choice of boundary conditions depends on the eigenvalues of the matrix B. This indicates that even in this very simple case that boundary conditions for hyperbolic systems are a complicated issue and often omitted in the discussion.

2.2 Euler Equations of Gas Dynamics

The Euler equations of gas dynamics describe the movement of fluid neglecting internal friction and constitute one of the most famous examples of a nonlinear hyperbolic system of partial differential equations. They consist of

$$\partial_t \rho + \nabla \cdot (\rho v) = 0, \qquad \text{(conservation of mass)} \tag{2.4a}$$

$$\partial_t (\rho v) + \nabla \cdot (\rho v v^T + pI) = f, \qquad \text{(conservation of momentum)} \tag{2.4b}$$

$$\partial_t e + \nabla \cdot ((e+p)v) = w, \qquad \text{(conservation of energy)} \tag{2.4c}$$

together with the thermodynamical relation

$$p = p(\rho, e) = (\gamma - 1)(e - \rho \|v\|^2/2) \tag{2.5}$$

in the space time domain $U = \mathbb{R}^d \times \mathbb{R}^+$ with appropriate initial conditions. Here $v(x,t) : U \to \mathbb{R}^d$ is the fluid velocity, $\rho(x,t) : U \to \mathbb{R}^+$ is its density,

$e(x,t) : U \to \mathbb{R}$ is the total energy being the sum of internal energy and kinetic energy and $p(x,t) : U \to \mathbb{R}$ is the pressure. The functions f and w denote the external forces and the energy source term. Equation (2.5) is a consequence of the equation of state $u = p/((\gamma-1)\rho)$ and the definition of total energy. The constant γ is the adiabatic exponent and depends on the type of gas. For more details, see [6, § 14.4]. Pressure is considered a dependent variable in (2.4) which can be eliminated using (2.5) resulting in a system of $m = n+2$ equations for the m unknown functions ρ, v_1, …, v_n and e ($m = 5$ in $n = 3$ space dimensions). It is interesting to note that we can combine all the equations (2.4) into a single equation for the unknown vector function $w = (\rho, \rho v, e)^T$:

$$\partial_t w(x,t) + \nabla \cdot F(w(x,t)) = g \tag{2.6}$$

with the flux function

$$F(w) = \begin{pmatrix} w_2 & w_3 & w_4 \\ \frac{w_2^2}{w_1} + p(w) & \frac{w_2 w_3}{w_1} & \frac{w_2 w_4}{w_1} \\ \frac{w_3 w_2}{w_1} & \frac{w_3^2}{w_1} + p(w) & \frac{w_3 w_4}{w_1} \\ \frac{w_4 w_2}{w_1} & \frac{w_4 w_3}{w_1} & \frac{w_4^2}{w_1} + p(w) \\ \frac{(w_5+p(w))w_2}{w_1} & \frac{(w_5+p(w))w_3}{w_1} & \frac{(w_5+p(w))w_4}{w_1} \end{pmatrix} \tag{2.7}$$

and

$$p(w) = (\gamma-1)\left(w_5 - \frac{w_2^2 + w_3^2 + w_4^2}{2w_1}\right).$$

Using the chain rule we obtain the form

$$\partial_t w(x,t) + \sum_{j=1}^{d} \nabla F_j(w) \partial_{x_j} w(x,t) = g \tag{2.8}$$

where $F(w) = [F_1(w), \ldots, F_n(w)]$ columnwise and $(\nabla F_j(w))_{ik} = \frac{\partial F_{ij}}{\partial w_k}(w)$ is an $m \times m$ matrix. Hyperbolicity then requires that

$$B_{\text{Euler}}(w; y) = \sum_{j=1}^{d} y_j \nabla F_j(w) \tag{2.9}$$

is real diagonalizable for any w and $y \in \mathbb{R}^d$.

Homogeneous Functions

The following class of functions plays an important role in the design of numerical methods.

Definition 2.2 (Homogeneous functions)**.** A function $f : \mathbb{R}^m \to \mathbb{R}^d$ is called (positive) homogeneous of degree $r \in \mathbb{N}$ if

$$f(\alpha w) = \alpha^r f(w) \qquad w \neq 0, \mathbb{R} \ni \alpha > 0 \, . \tag{2.10}$$

Theorem 2.3 (Euler homogeneous function theorem)**.** Let $f : \mathbb{R}^m \to \mathbb{R}^d$, $w \to f(w)$, be a homogeneous function of degree $r \in \mathbb{N}$. Then

$$f(w) = \frac{1}{r} \nabla f(w) w \, . \tag{2.11}$$

(Note for $n = 1$: $f(w) = \frac{1}{r} \nabla f(w) \cdot w$ since $\nabla f(w)$ is a column vector by definition).

Proof. Differentiate each component with respect to α on both sides of (2.10) and use the chain rule:

$$\begin{aligned} \frac{d}{d\alpha} f_i(\alpha w) &= \sum_{j=1}^{d} \frac{df_i}{du_j}(\alpha w) \frac{d(\alpha w_j)}{d\alpha} = \sum_{j=1}^{d} \frac{df_i}{du_j}(\alpha w) w_j = (\nabla f(\alpha w) w)_i \\ &= \frac{d}{d\alpha}(\alpha^r f_i(w)) = r \alpha^{r-1} f_i(w) \end{aligned}$$

and therefore $\nabla f(\alpha w) w = r \alpha^{r-1} f(w)$ for $\alpha > 0$. Setting $\alpha = 1$ proves the result. □

Example 2.4. Check that the columns of the Euler flux $F_j(w) : \mathbb{R}^m \to \mathbb{R}^m$ are homogeneous functions of degree 1, see also [2]. □

2.3 Acoustic Wave Equation

Sound waves are small variations in pressure (and correspondingly density) that move through a fluid (there are also waves in solids). In order to derive an equation for the propagation of these variations we start with the Euler equations (2.4). We write all quantities as a constant background value (indicated by the bar) plus a small variation depending on space and time (indicated by the tilde):

$$\rho = \bar{\rho} + \tilde{\rho}, \qquad p = \bar{p} + \tilde{p}, \qquad v = \bar{v} + \tilde{v}.$$

The background velocity is actually assumed to be zero, $\bar{v} = 0$, and the temperature T of the gas is assumed to be constant throughout the domain. From the ideal gas law (which replaces (2.5)) we get $p = c^2 \rho$ with $c = \sqrt{\bar{R}T}$ the speed of sound and therefore $p = c^2 \rho = c^2(\bar{\rho} + \tilde{\rho}) = c^2 \bar{\rho} + c^2 \tilde{\rho} = \bar{p} + \tilde{p}$.

Linearizing mass and momentum equations around the background state, dropping all higher-order terms in fluctuations (note especially that $\tilde{v}\tilde{v}^T$ can be

dropped) and assuming *constant background pressure* results (without external sources) in

$$\partial_t \tilde{\rho} + \nabla \cdot (\bar{\rho}\tilde{v}) = 0, \qquad \text{(conservation of mass)} \tag{2.12a}$$
$$\partial_t (\bar{\rho}\tilde{v}) + \nabla \tilde{p} = 0, \qquad \text{(conservation of momentum).} \tag{2.12b}$$

Nonconservative Form of Linear Acoustics Using $\tilde{\rho} = \tilde{p}/c^2$ and assuming that c is constant throughout the domain the density variation is eliminated and we obtain the equations of linear acoustics:

$$\partial_t \tilde{p} + c^2 \bar{\rho} \nabla \cdot \tilde{v} = 0, \tag{2.13a}$$
$$\bar{\rho} \partial_t \tilde{v} + \nabla \tilde{p} = 0. \tag{2.13b}$$

Taking the temporal derivative of the first equation and taking the divergence to the second, the velocity variation can be eliminated from this system and we obtain the classical *wave equation*:

$$\partial_t^2 \tilde{p} - c^2 \Delta \tilde{p} = 0 \tag{2.14}$$

which is second-order hyperbolic. In the analysis of the wave equation, (2.14) is often reduced to a first order system by setting $u = \partial_t \tilde{p}$ and $w = -\nabla \tilde{p}$. Together with the identities $\partial_{x_i} \partial_t \tilde{p} = \partial_t \partial_{x_i} \tilde{p}$ we obtain the system

$$\partial_t u + c^2 \nabla \cdot w = 0,$$
$$\partial_t w + \nabla u = 0,$$

which is equivalent to (2.13) (simply use the transformation $w = \bar{\rho}\tilde{v}$). It should be noted that it is the first order system that is derived from the physics and not the scalar second order wave equation, see also [6, § 2.7].

Conservative Form of Linear Acoustics We now consider the case that the speed of sound c is *piecewise constant* in fixed subdomains (e.g. due to temperature variations). Equation (2.12) is still valid in this case since only $\bar{p}$ being constant has been assumed. From $c^2\rho = p \Leftrightarrow c^2\bar{\rho} + c^2\tilde{\rho} = \bar{p} + \tilde{p}$ observe $\bar{p} = c^2\bar{\rho}$ and $\tilde{p} = c^2\tilde{\rho}$. Furthermore, from integration by parts of (2.12a) we conclude that $\bar{\rho}\tilde{v} \cdot n$ is continuous at subdomain boundaries. From integration by parts of the components of (2.12b) it follows that $c^2\tilde{\rho} = \tilde{p}$ is continuous. We conclude

$$c^2\bar{\rho} = \bar{p} = const \qquad \Rightarrow \bar{\rho} \text{ is piecewise constant,}$$
$$c^2\tilde{\rho} = \tilde{p} \text{ continuous} \qquad \Rightarrow \tilde{\rho} \text{ is discontinuous.}$$

In case of varying speed of sound it is then more appropriate to use the conservative variables $(\tilde{\rho}, \bar{\rho}\tilde{v}) = (\tilde{\rho}, \tilde{q})$ resulting in the system

$$\partial_t \tilde{\rho} + \nabla \cdot \tilde{q} = 0, \tag{2.15a}$$
$$\partial_t \tilde{q} + \nabla(c^2 \tilde{\rho}) = 0, \tag{2.15b}$$

together with the interface conditions

$$\lim_{\epsilon \to 0+} \left(c^2 \tilde{\rho}\right)(x - \epsilon n) = \lim_{\epsilon \to 0+} \left(c^2 \tilde{\rho}\right)(x + \epsilon n) \qquad x \in \Gamma, \tag{2.15c}$$
$$\lim_{\epsilon \to 0+} \tilde{q}(x - \epsilon n) \cdot n = \lim_{\epsilon \to 0+} \tilde{q}(x + \epsilon n) \cdot n \qquad x \in \Gamma, \tag{2.15d}$$

where Γ doneotes the subdomain boundaries where $c(x)$ is discontinuous.

Example 2.5. An example for linear acoustics is shown in Figure 2.2. An initial density peak produces a circular wave which is reflected at the boundaries as well as at the internal boundary where the speed of sound changes to a lower value in the upper right segment.

Hyperbolicity of Linear Acoustics It remains to show the hyperbolicity of the linear acoustics system. We do this by explicitly calculating eigenvalues and eigenvectors for the system (2.15) in three space dimensions. Setting $u = (\tilde{\rho}, \tilde{q}_1, \tilde{q}_2, \tilde{q}_3)$ system (2.15) can be written as

$$\partial_t u + \sum_{j=1}^{3} B_j \partial_{x_j} u = 0$$

with

$$B_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ c^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \; B_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ c^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \; B_3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ c^2 & 0 & 0 & 0 \end{pmatrix}.$$

For any $y \in \mathbb{R}^3$ we therefore have

$$B(y) = \sum_{j=1}^{3} y_j B_j = \begin{pmatrix} 0 & y_1 & y_2 & y_3 \\ c^2 y_1 & 0 & 0 & 0 \\ c^2 y_2 & 0 & 0 & 0 \\ c^2 y_3 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & y^T \\ c^2 y & 0 \end{pmatrix}.$$

In this form, the linear acoustic system can be considered with $y \in \mathbb{R}^d$ in any dimension.

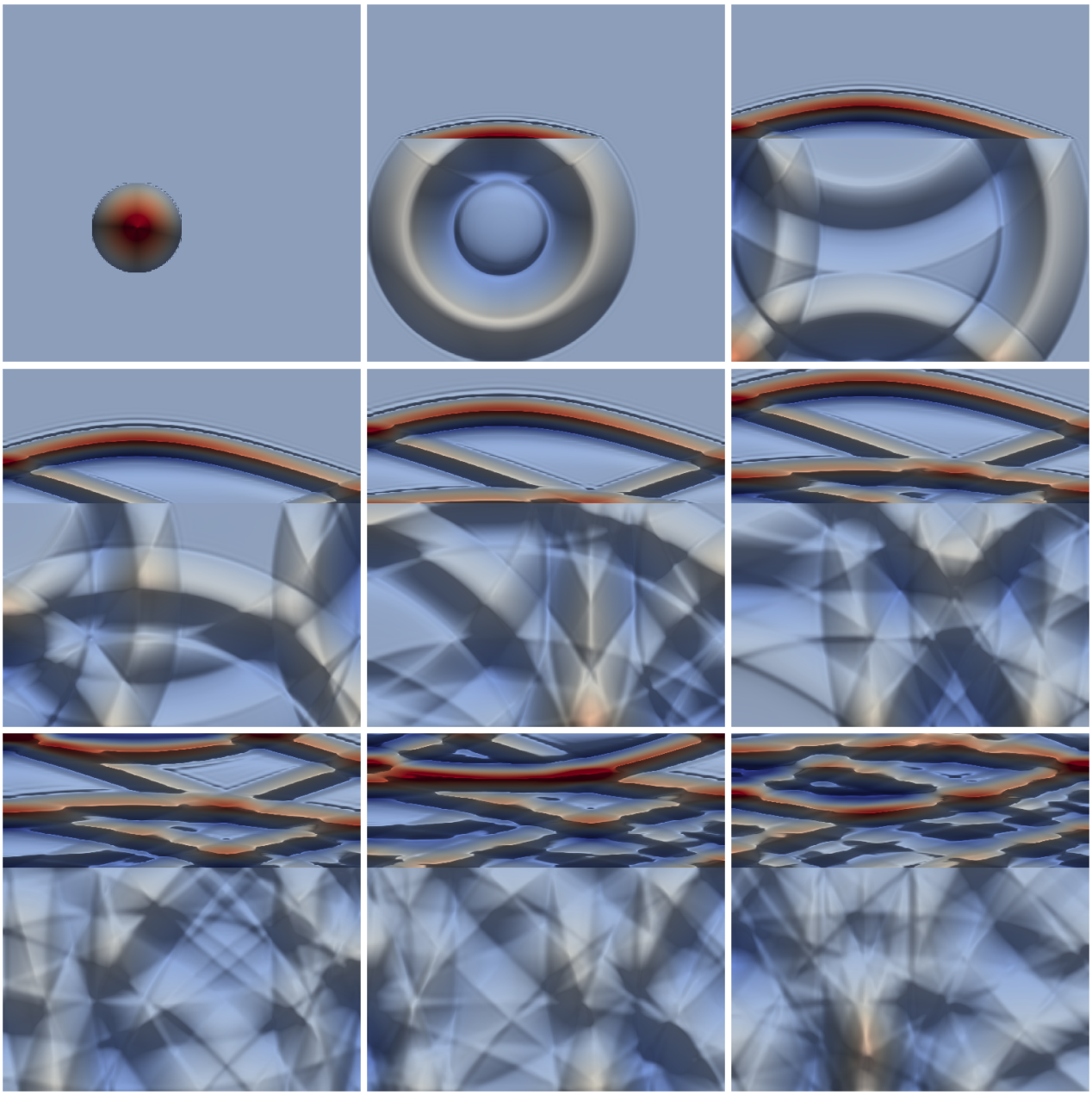


Figure 2.2: Acoustic wave propagation in a heterogeneous medium with reflective boundary conditions. Fully third-order discontinuous Galerkin scheme computed on a 128×128 mesh. Speed of sound in the upper part of the domain is one third of that in the lower part. Time sequence goes from top left to bottom right.

With the transformation matrix $T = \mathrm{diag}(1/c, 1, \ldots, 1)$ we see that $B(y)$ is similar to the symmetric matrix

$$\bar{B}(y) = T^{-1}B(y)T = \begin{pmatrix} 0 & cy^T \\ cy & 0 \end{pmatrix}.$$

$\bar{B}(y)$ is diagonalizable with eigenvalues

$$\lambda_1 = c\|y\|, \qquad \lambda_2 = -c\|y\|, \qquad \lambda_n = 0, n > 2. \tag{2.16}$$

and corresponding eigenvectors

$$r_1 = \begin{pmatrix} \|y\| \\ y \end{pmatrix}, \qquad r_2 = \begin{pmatrix} -\|y\| \\ y \end{pmatrix}, \qquad r_n = \begin{pmatrix} 0 \\ s \end{pmatrix}, s \cdot y = 0. \tag{2.17}$$

Note that the linear acoustics system in dimension d has $d+1$ components and $\lambda = 0$ is a $d-1$-fold eigenvalue. The eigenspace of $\bar{B}(y)$ corresponding to $\lambda = 0$ consists of all vectors $r = (0, s^T)^T$ with $s \cdot y = 0$ which has dimension $d-1$. Thus there is a full set of eigenvectors. In detail we have the following eigenvectors for dimension one

$$r_1 = \begin{pmatrix} y \\ y \end{pmatrix}, \qquad\qquad r_2 = \begin{pmatrix} -y \\ y \end{pmatrix}$$

and two

$$r_1 = \begin{pmatrix} \|y\| \\ y_1 \\ y_2 \end{pmatrix}, \qquad r_2 = \begin{pmatrix} -\|y\| \\ y_1 \\ y_2 \end{pmatrix}, \qquad r_3 = \begin{pmatrix} 0 \\ -y_2 \\ y_1 \end{pmatrix}$$

and three

$$r_1 = \begin{pmatrix} \|y\| \\ y \end{pmatrix}, \quad r_2 = \begin{pmatrix} -\|y\| \\ y \end{pmatrix}, \quad r_3 = \begin{pmatrix} 0 \\ s \end{pmatrix}, \quad r_4 = \begin{pmatrix} 0 \\ y \times s \end{pmatrix}$$

where $s = (y_2 - y_3, y_3 - y_1, y_1 - y_2)^T$ if $\|s\| > \epsilon$ and $s = (y_2 + y_3, y_3 - y_1, -(y_1 + y_2))^T$ else.

Now we need eigenvalues and eigenvectors of $B(y)$. From

$$\bar{B}(y)r_i = \lambda_i r_i \quad \Leftrightarrow \quad T^{-1}B(y)Tr_i = \lambda_i T^{-1}Tr_i \quad \Leftrightarrow \quad B(y)Tr_i = \lambda_i Tr_i$$

we observe that $B(y)$ has the same eigenvalues as $\bar{B}(y)$ and eigenvectors are transformed by T.

Waves in Solids Solid bodies are also able to support a propagation of waves, an example being earthquakes. In the one-dimensional situation we may imagine a string of beads connected by springs with each other. One type of wave consists of small displacements of a bead in the direction of the string resulting in displacements of the neighbouring beads. This type of wave is called a compression wave or P-wave and it is similar to the sound waves in a gas. Another type of wave results from displacements of a bead in a direction perpendicular to the string which also results in the propagation of a wave in the direction of the string. This is called S-wave which usually travels slower than a P-wave. In the one-dimensional situation both types of waves are described by the one-dimensional wave equation $\partial_t^2 u - c^2 \partial_x^2 u = 0$ (A derivation of the P-wave is in [3, § 17.2] and the S-wave can be found in [8, § 176]). In a multi-dimensional solid both types of waves interact and more complicated equations result (see [6, § 2.12] for some discussion). At the surface or at internal boundaries surface waves can be observed.

2.4 Maxwell's Equations

The Maxwell system is given by

$$\partial_t D - \nabla \times H = -J, \qquad \text{(Ampère)} \tag{2.18a}$$
$$\partial_t B + \nabla \times E = 0, \qquad \text{(Faraday)} \tag{2.18b}$$
$$\nabla \cdot D = \rho, \qquad \text{(Gauß)} \tag{2.18c}$$
$$\nabla \cdot B = 0, \qquad \text{(Gauß for magnetism)} \tag{2.18d}$$

together with the constitutive laws

$$D = \epsilon E, \qquad (\ \epsilon.\ \text{permittivity}) \tag{2.19a}$$
$$B = \mu H, \qquad (\ \mu\text{: permeability}) \tag{2.19b}$$
$$J = \sigma E + j, \qquad (\ \sigma\text{: conductivity}). \tag{2.19c}$$

The following vector fields in $\mathbb{R}^3$ need to be determined:

Symbol	Name	Unit
B	magnetic flux density	$\frac{Vs}{m^2}$
H	magnetic field intensity	$\frac{A}{m}$
E	electric field intensity	$\frac{V}{m}$
D	displacement current density	$\frac{As}{m^2}$

whereas the scalar charge density ρ and the current density j are prescribed.

The conditions (2.18c) and (2.18c) are needed only for the initial condition. The evolution in time is described by (2.18b) and (2.18a) only, see [5].

Since D and B are conserved quantities we formulate equations (2.18b) and (2.18a) in terms of D and B using the constitutive equations:

$$\partial_t D - \nabla \times \left(\frac{1}{\mu} B\right) + \frac{\sigma}{\epsilon} D = -j\,, \tag{2.20a}$$

$$\partial_t B + \nabla \times \left(\frac{1}{\epsilon} D\right) = 0\,. \tag{2.20b}$$

Writing out the curl operator $\nabla\times$ and defining the six component vector $u = (D_1, D_2, D_3, B_1, B_2, B_3)^T$ we obtain the linear hyperbolic system

$$\partial_t u + \sum_{j=1}^{3} B_j \partial_{x_j} u + Cu = q \tag{2.21}$$

with

$$B_1 = \left(\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/\mu \\ 0 & 0 & 0 & 0 & -1/\mu & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1/\epsilon & 0 & 0 & 0 \\ 0 & 1/\epsilon & 0 & 0 & 0 & 0 \end{array}\right), \quad B_2 = \left(\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & -1/\mu \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/\mu & 0 & 0 \\ \hline 0 & 0 & 1/\epsilon & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1/\epsilon & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

$$B_3 = \left(\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 1/\mu & 0 \\ 0 & 0 & 0 & -1/\mu & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -1/\epsilon & 0 & 0 & 0 & 0 \\ 1/\epsilon & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right), \quad C = \left(\begin{array}{ccc|ccc} \sigma/\epsilon & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma/\epsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma/\epsilon & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

and $q = (-j_1, -j_2, -j_3, 0, 0, 0)^T$. Hyperbolicity is obtained from the matrix

$$B_{\text{Maxwell}}(y) = \sum_{j=1}^{3} y_j B_j = \left(\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & y_3/\mu & -y_2/\mu \\ 0 & 0 & 0 & -y_3/\mu & 0 & y_1/\mu \\ 0 & 0 & 0 & y_2/\mu & -y_1/\mu & 0 \\ \hline 0 & -y_3/\epsilon & y_2/\epsilon & 0 & 0 & 0 \\ y_3/\epsilon & 0 & -y_1/\epsilon & 0 & 0 & 0 \\ -y_2/\epsilon & y_1/\epsilon & 0 & 0 & 0 & 0 \end{array}\right). \tag{2.22}$$

Using the diagonal transformation matrix

$$T = \operatorname{diag}(\sqrt{1/\epsilon}, \sqrt{1/\epsilon}, \sqrt{1/\epsilon}, \sqrt{1/\mu}, \sqrt{1/\mu}, \sqrt{1/\mu})$$

we obtain the similarity transformation

$$TB_{\text{Maxwell}}(y)T^{-1} = \frac{1}{\sqrt{\epsilon\mu}} \left(\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & y_3 & -y_2 \\ 0 & 0 & 0 & -y_3 & 0 & y_1 \\ 0 & 0 & 0 & y_2 & -y_1 & 0 \\ \hline 0 & -y_3 & y_2 & 0 & 0 & 0 \\ y_3 & 0 & -y_1 & 0 & 0 & 0 \\ -y_2 & y_1 & 0 & 0 & 0 & 0 \end{array}\right). \tag{2.23}$$

Thus $B_{\text{Maxwell}}(y)$ is similar to a real symmetric matrix from which the set of eigenvalues and eigenvectors can be determined. It turns out that the eigenvalues of $B_{\text{Maxwell}}(y)$ are 0, $c\|y\|$ and $c\|y\|$ each with multiplicity 2 and $c = 1/\sqrt{\epsilon\mu}$ the speed of light.

Chapter 3

Low-order Finite Volume Methods

3.1 Basic Method for Scalar Linear Transport

We start with the scalar linear model problem

$$\partial_t u(x,t) + \nabla \cdot (\beta(x,t)u(x,t)) = f(x,t), \qquad (x,t) \in U = \Omega \times \Sigma, \tag{3.1a}$$
$$u(x,t) = g(x,t), \qquad (x,t) \in \Gamma^-(t) \times \Sigma, \tag{3.1b}$$
$$u(x,t_0) = u_0(x), \qquad x \in \Omega. \tag{3.1c}$$

Here we denote the velocity field by β since the letter v will be reserved for functions.

Notation for Meshes

Numerical methods are based on a decomposition of the finite domain $\Omega \subset \mathbb{R}^d$ into a mesh $\mathcal{E}_h$ into open domains $e \in \mathcal{E}_h$ also called *cells* or *elements*:

$$\bigcup_{e \in \mathcal{E}_h} \overline{e} = \overline{\Omega}, \qquad \forall e \neq e' : \; e \cap e' = \emptyset.$$

Here we assume for simplicity that the domain Ω is a polyhedron and the elements each are the images of a reference element $\hat{E}$ under a map $\mu_e : \hat{E} \to \overline{e}$ where the reference element is either the unit simplex or unit cube in n dimensions and the map μ_e is linear or multi-linear, respectively.

The diameter of $e \in \mathcal{E}_h$ is h_e and n_e is its unit outer normal vector. An example of a mesh consisting of four triangular elements is shown in figure 3.1. Finite volume methods for first-order hyperbolic PDEs allow for very general meshes, e.g. meshes need not be conforming as the one shown in figure 3.1.

An intersection $f = \overline{e^-} \cap \overline{e^+}$ of codimension 1 (e.g. a surface when $n = 3$) of two elements $\overline{e^-}, \overline{e^+} \in \mathcal{E}_h$ is called an interior intersection and the set of all interior intersections is collected in the set $\mathcal{F}_h^i$. The intersection of an element with the boundary, $f = \overline{e} \cap \partial\Omega$, is a boundary intersection and all such boundary intersections are collected in the set $\mathcal{F}_h^{\partial\Omega}$ which is further partioned into intersections $\mathcal{F}_h^-(t)$ with the inflow boundary $\Gamma^-(t)$ and its complement $\mathcal{F}_h^{0+}(t) = \mathcal{F}_h^{\partial\Omega} \setminus \mathcal{F}_h^-(t)$. These sets depend on time if the velocity field $\beta(x,t)$ depends on time but for ease of writing we will omit this dependence in the

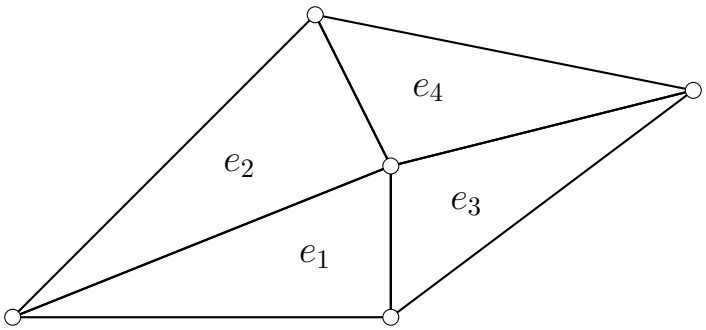


Figure 3.1: A triangular mesh.

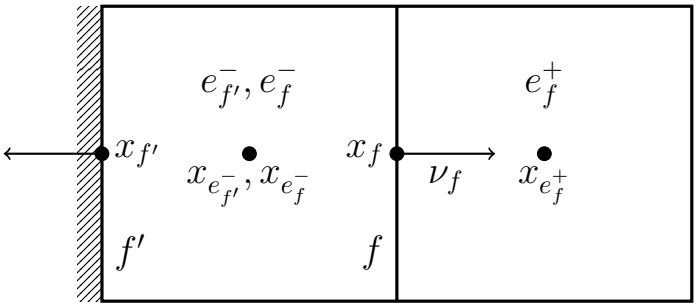


Figure 3.2: Notation for interior and boundary intersections.

following. Finally, we denote by $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^{\partial\Omega}$ the set of all intersections and by $\mathcal{F}_h(e) = \{f \in \mathcal{F}_h \,:\, e = e^-(f) \vee e = e^+(f)\}$ the set of intersections of element e.

The diameter of $f \in \mathcal{F}_h$ is h_f. With each $f \in \mathcal{F}_h^i$ we associate a unit normal vector n_f oriented from element $e^-(f)$ to $e^+(f)$ when $f = \overline{e^-(f)} \cap \overline{e^+(f)}$ (just select one element to be the minus side). With each $f \in \mathcal{F}_h^{\partial\Omega}$ we choose n_f as the unit outer normal to the domain and denote by $e^-(f)$ the element where $f = \overline{e^-(f)} \cap \partial\Omega$. This notation is illustrated in Figure 3.2.

Space Discretization

The finite volume method is based on the space of piecewise constant functions on the mesh $\mathcal{E}_h$ given by

$$V_h^0 = \left\{ v \in L^2(\Omega) : \forall e \in \mathcal{E}_h, v|_e = \text{const} \right\}. \tag{3.2}$$

A function $v \in V_h^0$ is two-valued on an interior face $f \in \mathcal{F}_h^i$ and for $x \in f$ we denote by $v^-(x)$ the restriction from $e^-(f)$ and by $v^+(x)$ the restriction from $e^+(f)$. For any point $x \in f \in \mathcal{F}_h^i$ we define the jump

$$[\![v]\!](x) = v^-(x) - v^+(x) \tag{3.3}$$

and the average

$$\{v\}(x) = \frac{1}{2}v^-(x) - \frac{1}{2}v^+(x). \tag{3.4}$$

In order to solve the time-dependent problem (3.1) we follow the *method of lines* paradigm: First discretize in space leaving the time variable continuous. Then solve the resulting system of ordinary equations by standard methods. In this approach the following notation is quite useful. For any function $u(x,t)$ in space and time consider $u(t) = u(t,\cdot) \in V$ to be a function in some functions space V and $u : \Sigma \to V$ a map associating such a function with every $t \in \Sigma$ such that $u(t)(x) = u(x,t)$ holds.

Now for any function $u(t)$ multiply equation (3.1a) by a test function $v \in V_h^0$ and integrate:

$$\begin{aligned}
&\int_\Omega (\partial_t u(x,t) + \nabla\cdot(\beta(x,t)u(x,t)))v(x)\,dx \\
&\qquad = \sum_{e\in\mathcal{E}_h}\left\{\int_e (\partial_t u(t))v\,dx + \int_e \nabla\cdot(\beta(t)u(t))v\,dx\right\} \\
&\qquad = \sum_{e\in\mathcal{E}_h}\left\{d_t\int_e u(t)v\,dx + \int_{\partial e}\beta(t)\cdot n_e u(t)v\,ds\right\} \\
&\qquad = d_t\int_\Omega u(t)v\,dx + \sum_{f\in\mathcal{F}_h^{0+}}\int_f \beta(t)\cdot n_f u(t)v\,ds \\
&\qquad\quad + \sum_{f\in\mathcal{F}_h^-}\int_f \beta(t)\cdot n_f g(t)v\,ds + \sum_{f\in\mathcal{F}_h^i}\int_f [\![\beta(t)\cdot n_f u(t)v]\!]\,ds.
\end{aligned}$$

The jump term in the last sum is due to the fact that normal directions for the two elements $e^-(f)$ and $e^+(f)$ of an intersection are the negative of each other. From physical reasoning it is clear that a flux normal to an intersection should be the same from both sides, since otherwise the conserved quantity would be lost at the interface. Therefore we introduce a numerical flux $\Phi(u,\beta_n)$ to be evaluated on interior intersections replacing the normal flux $\beta(t)\cdot n_f u(t)$. Two choices will be considered:

$$\Phi_C(u,\beta_n)(x) = \beta_n(x)\frac{u^-(x)+u^+(x)}{2}, \qquad \text{(central)} \tag{3.5}$$

$$\Phi_U(u,\beta_n)(x) = \max(0,\beta_n(x))u^-(x) + \min(0,\beta_n(x))u^+(x), \quad \text{(upwind)}. \tag{3.6}$$

The semi-discretized scheme then reads as follows. Find $u_h : \Sigma \to V_h^0$ such that

$$\begin{aligned} & d_t \int_\Omega u_h(t) v_h \, dx + \sum_{f \in \mathcal{F}_h^{0+}} \int_f \beta(t) \cdot n_f u_h(t) v_h \, ds \\ & \quad + \sum_{f \in \mathcal{F}_h^-} \int_f \beta(t) \cdot n_f g(t) v_h \, ds + \sum_{f \in \mathcal{F}_h^i} \int_f \Phi(u_h(t), \beta(t) \cdot n_f) [\![v_h]\!] \, ds \\ & = \int_\Omega f(x,t) v_h(x) \, dx \qquad \forall v_h \in V_h^0, t \in \Sigma \end{aligned} \tag{3.7}$$

and any of the two numerical fluxes introduced above. It turns out, however, that the scheme using the central flux does not behave very well numerically. Depending on the discretization in time (to be introduced below) the numerical method may become unconditionally unstable.

Remark 3.1. An essential assumption for the numerical fluxes introduced above is that $\beta(t,x) \cdot n_f$ is continuous when traversing with the point x from $e^-(f)$ to $e^+(f)$ across the intersection f. This requirement may not be trivial to satisfy when the velocity β is computed numerically. A sufficient condition in this case is $\beta(t) \in H(\mathrm{div}, \Omega)$.

Remark 3.2. The upwind flux can equivalently be written as

$$\Phi_U(u, \beta_n)(x) = \Phi_C(u, \beta_n)(x) + \frac{|\beta_n|}{2}(u^-(x) - u^+(x)). \tag{3.8}$$

This allows two interpretations of the upwind flux:

I) The method of characteristics implies that if $\beta(x,t) \cdot n_f \geq 0$ for a point $x \in f$ then the value of u should be determined from u in $e^-(f)$.

II) The equivalent version (3.8) suggests that the upwind flux is the (unstable) central flux plus a stabilization term. Setting $v_h = u_h(t)$ as a test function we obtain

$$\int_f \frac{|\beta \cdot n_f|}{2} [\![u_h(x,t)]\!]^2 \, ds > 0$$

which is favourable in the analysis of the scheme.

Before proceeding with the time discretization we rewrite the scheme in a more compact form using the notation

$$(v, w)_\omega = \int_\omega v \cdot w \, dx$$

for the L^2 scalar product of two (later possibly vector-valued) functions. With that we define the time-dependent bilinear form for the upwind variant

$$b_{\mathrm{FVU}}(u,v;t) = \sum_{f\in\mathcal{F}_h^{0+}} (\beta(t)\cdot n_f\, u, v)_f + \sum_{f\in\mathcal{F}_h^{i}} (\Phi_U(u,\beta(t)\cdot n_f), [\![v]\!])_f\,, \tag{3.9}$$

the time-dependent bilinear form for the central variant:

$$b_{\mathrm{FVC}}(u,v;t) = \sum_{f\in\mathcal{F}_h^{0+}} (\beta(t)\cdot n_f\, u, v)_f + \sum_{f\in\mathcal{F}_h^{i}} (\Phi_C(u,\beta(t)\cdot n_f), [\![v]\!])_f \tag{3.10}$$

and the right hand side functional:

$$r(v;t) = (f(t),v)_\Omega - \sum_{f\in\mathcal{F}_h^{-}} (\beta(t)\cdot n_f\, g(t), v)_f \tag{3.11}$$

Now the compact form of the semi-discrete scheme (3.7) for the upind variant is written as follows. Find $u_h : \Sigma \to V_h^0$ s. t.:

$$d_t\,(u_h(t),v_h)_\Omega + b_{\mathrm{FVU}}(u_h(t),v_h;t) = r(v_h;t) \qquad \forall v_h \in V_h^0, t\in\Sigma. \tag{3.12}$$

The variant for the central flux is defined in the same way.

Time Discretization

It remains to employ a time discretization. Subdivide the time interval $\Sigma = (t_0, t_0^T)$ into M subintervals:

$$t_0 = t^0 < t^1 < \ldots < t^M = t_0 + T, \quad \sigma_k = (t^{k-1}, t^k), \quad \Delta t_k = t^k - t^{k-1} \tag{3.13}$$

and denote by $u_h^k \in V_h^0$ the approximation of $u_h(t^k)$. Here we only employ the simplest two methods where the time derivative is approximated by a difference quotient:

$$d_t\,(u_h(t),v_h)_\Omega|_{t^{k-1}} = \frac{(u_h(t^k),v_h)_\Omega - (u_h(t^{k-1}),v_h)_\Omega}{\Delta t_k} + O(\Delta t_k).$$

The *explicit Euler finite volume* scheme reads as follows. Let u_h^0 be a projection of the initial condition $u_0(x)$. For $k = 1,\ldots,M$ determine u_h^k from

$$(u_h^k, v_h)_\Omega = (u_h^{k-1}, v_h)_\Omega - \Delta t_k b_{\mathrm{FVU}}(u_h^{k-1}, v_h; t^{k-1}) + \Delta t_k r(v_h; t^{k-1})\ \forall v_h \in V_h^0. \tag{3.14}$$

The *implicit Euler finite volume scheme* reads as follows. Let u_h^0 be a projection of the initial condition $u_0(x)$. For $k = 1,\ldots,M$ determine u_h^k from

$$(u_h^k, v_h)_\Omega + \Delta t_k b_{\mathrm{FVU}}(u_h^k, v_h; t^k) = (u_h^{k-1}, v_h)_\Omega + \Delta t_k r(v_h; t^k)\ \forall v_h \in V_h^0. \tag{3.15}$$

The corresponding variants for the central flux can be defined in the same way.

Implementation

In order to realize the schemes (3.14) and (3.14) in the computer one needs to insert a basis representation of the space V_h^0. The natural basis functions are given by

$$\varphi_e(x) = \begin{cases} 1 & x \in e \\ 0 & \text{else} \end{cases}, \quad e \in \mathcal{E}_h\,, \qquad \Phi_h = \{\varphi_e \,:\, e \in \mathcal{E}_h\}. \tag{3.16}$$

With that we can represent the solution at time t^k by

$$u_h^k = \sum_{e' \in \mathcal{E}_h} z_{e'}^k \phi_{e'} \tag{3.17}$$

and due to linearity it suffices to test in (3.14) and (3.14) with the finitely many basis functions $\phi_e \in \Phi_h$.

We first treat the explicit scheme. Inserting the basis representation into (3.14) results in

$$\begin{aligned}\left(\sum_{e' \in \mathcal{E}_h} z_{e'}^k \varphi_{e'}, \varphi_e\right)_\Omega &= \left(\sum_{e' \in \mathcal{E}_h} z_{e'}^{k-1} \varphi_{e'}, \varphi_e\right)_\Omega \\ &\quad - \Delta t_k b_{\mathrm{FVU}}\left(\sum_{e' \in \mathcal{E}_h} z_{e'}^{k-1} \varphi_{e'}, \varphi_e; t^{k-1}\right) + \Delta t_k r(\varphi_e; t^{k-1}) \quad \forall e \in \mathcal{E}_h\end{aligned}$$

which upon using linearity in the first argument results in one linear equation per element:

$$\begin{aligned}\sum_{e' \in \mathcal{E}_h} z_{e'}^k \left(\varphi_{e'}, \varphi_e\right)_\Omega = \sum_{e' \in \mathcal{E}_h} z_{e'}^{k-1} \left(\varphi_{e'}, \varphi_e\right)_\Omega - \Delta t_k \sum_{e' \in \mathcal{E}_h} z_{e'}^{k-1} b_{\mathrm{FVU}}\left(\varphi_{e'}, \varphi_e; t^{k-1}\right) \\ + \Delta t_k r(\varphi_e; t^{k-1}) \qquad \forall e \in \mathcal{E}_h.\end{aligned} \tag{3.18}$$

Evaluating the integrals gives

$$\begin{aligned}(\varphi_{e'}, \varphi_e)_\Omega &= \begin{cases} |e| & e = e' \\ 0 & e \neq e' \end{cases}, \\ r(\varphi_e; t) &= \int_e f(t)\, dx - \sum_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^-} \int_f \beta(t) \cdot n_f g(t)\, ds \\ &\approx f(x_e, t)|e| - \sum_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^-} \beta(x_f, t) \cdot n_f g(x_f, t)|f|.\end{aligned}$$

Here, we applied the midpoint rule for quadrature, defined the element center x_e, the face center x_f and introduced the short hand notation $|\omega| = (1,1)_\omega$. And for the remaining integral we obtain

$$\begin{aligned} b_{\mathrm{FVU}}(\varphi_{e'}, \varphi_e; t) &= \sum_{f \in \mathcal{F}_h^{0+}} (\beta(t) \cdot n_f \, \varphi_{e'}, \varphi_e)_f + \sum_{f \in \mathcal{F}_h^i} (\Phi_U(\varphi_{e'}, \beta(t) \cdot n_f), [\![\varphi_e]\!])_f \\ &= \begin{cases} \sum\limits_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^{0+}} \int\limits_f \beta(t) \cdot n_f \, ds + \sum\limits_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^i} \int\limits_f |\beta(t) \cdot n_f| \, ds & e = e' \\ - \sum\limits_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h(e')} \int\limits_f |\beta(t) \cdot n_f| \, ds & e \neq e' \end{cases} \\ &\approx \begin{cases} \sum\limits_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^{0+}} \beta(x_f, t) \cdot n_f |f| + \sum\limits_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^i} |\beta(x_f, t) \cdot n_f| |f| & e = e' \\ - \sum\limits_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h(e')} |\beta(x_f, t) \cdot n_f| |f| & e \neq e' \end{cases} . \end{aligned}$$

Inserting these expressions into (3.18) now results in

$$\begin{aligned} z_e^k = z_e^{k-1} &\left(1 - \Delta t_k \sum_{f \in \mathcal{F}_h(e), \beta(x_f, t^{k-1}) \cdot n_f \geq 0} \frac{|\beta(x_f, t^{k-1}) \cdot n_f| |f|}{|e|} \right) \\ &+ \Delta t_k \sum_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^i, \beta(x_f, t^{k-1}) \cdot n_f < 0} \frac{|\beta(x_f, t^{k-1}) \cdot n_f| |f|}{|e|} z_{\mathrm{nb}(e,f)}^{k-1} \\ &+ \Delta t_k \sum_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^-} \frac{|\beta(x_f, t^{k-1}) \cdot n_f| |f|}{|e|} g(x_f, t^{k-1}) + \Delta t_k f(x_e, t^{k-1}) \end{aligned} \tag{3.19}$$

where we defined the neighbor of e over intersection $f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^i$:

$$\mathrm{nb}(e, f) = \begin{cases} e^+(f) & e = e^-(f) \\ e^-(f) & e = e^+(f) \end{cases} .$$

A similar expression can be developed for the implicit scheme:

$$\begin{aligned} z_e^k &\left(1 + \Delta t_k \sum_{f \in \mathcal{F}_h(e), \beta(x_f, t^k) \cdot n_f \geq 0} \frac{|\beta(x_f, t) \cdot n_f| |f|}{|e|} \right) \\ &\quad - \Delta t_k \sum_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^i, \beta(x_f, t^k) \cdot n_f < 0} \frac{|\beta(x_f, t) \cdot n_f| |f|}{|e|} z_{\mathrm{nb}(e,f)}^k \\ &= z_e^{k-1} + \Delta t_k \sum_{f \in \mathcal{F}_h(e) \cap \mathcal{F}_h^-} \frac{|\beta(x_f, t^k) \cdot n_f| |f|}{|e|} g(x_f, t^k) + \Delta t_k f(x_e, t^k). \end{aligned} \tag{3.20}$$

We leave the development of corresponding expressions for the central flux to the reader.

3.2 Stability

When we assume that $\nabla \cdot \beta(t) = 0$, $g(x,t) = 0$ and $f(x,t) = 0$ the method of characteristics shows that the solution should, for all times $t > t_0$ be in the range $[\min_x u_0(x), \max_x u_0(x)]$. Stability in the maximum norm means that the numerical solution satisfies the same condition. In this section we carry out the stability analysis for the explicit and implicit upwind finite volume schemes.

Explicit Scheme

Taking the equations of (3.19) for all elements $e \in \mathcal{E}_h$, assuming $g(x,t) = 0$, $f(x,t) = 0$ and arranging the equations in matrix form results in

$$z^k = A_{\text{EUFV}} z^{k-1}. \tag{3.21}$$

By taking the maximum norm on both sides

$$\|z^k\|_\infty = \|A_{\text{EUFV}} z^{k-1}\|_\infty \leq \|A_{\text{EUFV}}\|_\infty \|z^{k-1}\|_\infty \tag{3.22}$$

we conclude that the explicit upwind finite volume scheme satisfies a maximum principle if $\|A_{\text{EUFV}}\|_\infty \leq 1$ where the row sum norm is defined by

$$\|A\|_\infty = \max_i \sum_j |a_{ij}| \ .$$

Further observe that $\nabla \cdot \beta(t) = 0$ implies $\int_{\partial e} \beta(t) \cdot n_e \, ds = 0$. From this one may conclude:

$$\sum_{f \in \mathcal{F}_h(e)} \int_f \beta(x,t) \cdot n_e \, ds = 0. \tag{3.23}$$

A necessary condition for a maximum principle is that $a_{ij} \geq 0$ (otherwise, if $a_{ij} < 0$ take $z_j = 1$ and $z_i = 0$ for $i \neq j$ and we get $(Az)_i < 0$ which violates the maximum principle). For our matrix given by (3.19) this implies for every element e:

$$\begin{aligned} 1 - \Delta t_k \sum_{f \in \mathcal{F}_h(e), \beta(x_f, t^{k-1}) \cdot n_f \geq 0} \frac{|\beta(x_f, t^{k-1}) \cdot n_f||f|}{|e|} \geq 0 \\ \Leftrightarrow \sum_{f \in \mathcal{F}_h(e), \beta(x_f, t^{k-1}) \cdot n_f \geq 0} |\beta(x_f, t^{k-1}) \cdot n_f||f| \leq \frac{|e|}{\Delta t_k}. \end{aligned} \tag{3.24}$$

This is the famous CFL-condition named after Courant Friedrichs and Levy. It can always be satisfied by taking Δt_k small enough. For a structured equidistant

mesh with mesh size h we have $|e| = h^d$ and $f = h^{n-1}$ and we obtain

$$\sum_{f\in\mathcal{F}_h(e),\beta(x_f,t^{k-1})\cdot n_f\geq 0} |\beta(x_f,t^{k-1})\cdot n_f| \leq \frac{h}{\Delta t_k}.$$

If the CFL-condition (3.24) is satisfied we obtain for the row sum norm (observe the signs in (3.19)!):

$$\begin{aligned}
\|A_{\mathrm{EUFV}}\|_\infty &= \max_{e\in\mathcal{E}_h}\Bigg(1 - \Delta t_k \sum_{f\in\mathcal{F}_h(e),\beta(x_f,t^{k-1})\cdot n_f\geq 0} \frac{|\beta(x_f,t^{k-1})\cdot n_f||f|}{|e|} \\
&\quad + \Delta t_k \sum_{f\in\mathcal{F}_h(e)\cap\mathcal{F}_h^i,\beta(x_f,t^{k-1})\cdot n_f<0} \frac{|\beta(x_f,t^{k-1})\cdot n_f||f|}{|e|}\Bigg) \\
&= \max_{e\in\mathcal{E}_h}\Bigg(1 - \frac{\Delta t_k}{|e|}\Bigg[\sum_{f\in\mathcal{F}_h(e),\beta(x_f,t^{k-1})\cdot n_f\geq 0} \beta(x_f,t^{k-1})\cdot n_f|f| \\
&\quad + \sum_{f\in\mathcal{F}_h(e)\cap\mathcal{F}_h^i,\beta(x_f,t^{k-1})\cdot n_f<0} \beta(x_f,t^{k-1})\cdot n_f|f|\Bigg]\Bigg) \\
&= 1
\end{aligned} \tag{3.25}$$

where we have exploited (3.23) (well this requires actually that quadrature is accurate enough for the given velocity field).

Remark 3.3. The CFL-condition (3.24) is sharp! If it is not satisfied the solution will blow up exponentially.

Implicit Scheme

For the implicit upwind finite volume scheme (3.20) we obtain under the assumptions above an evolution of the form

$$A_{\mathrm{IUFV}} z^k = z^{k-1}.$$

Solving for z^k and taking the maximum norm we obtain

$$\|z^k\|_\infty = \|A_{\mathrm{IUFV}}^{-1} z^{k-1}\|_\infty \leq \|A_{\mathrm{IUFV}}^{-1}\|_\infty \|z^{k-1}\|_\infty .$$

$\|A_{\mathrm{IUFV}}^{-1}\|_\infty = 1$ can be established with M-matrix theory (see [4]) and does not require any condition on Δt_k. One can immediately check from (3.20) that the sign condition for M-matrices is satisfied. Moreover, A_{IUFV} is weakly diagonally dominant due to (3.23). Thus the implicit upwind finite volume scheme is unconditionally stable in the maximum norm!

Remark 3.4 (Stability for the central flux)**.** It turns out that the explicit finite volume scheme with central flux is unconditionally unstable. The implicit scheme with central flux is stable when the time step is *large enough* which is odd because then accuracy is harmed. This explains that this scheme is not used in practice.

3.3 Numerical Results

Figure 3.3 gives numerical results for the proposed schemes. The domain is $\Omega = (0,1)^2$ and the velocity field constant at an angle 30° with $\|\beta\| = 1$. The initial condition is discontinuous and is shown in the top row of images. At the left boundary (part of the the inflow) a Dirichlet condition is prescribed. The mesh is quadrilateral with an equidistant size $h = 1/100$.

The left column shows the explicit upwind finite volume scheme operating close to the stability limit at $\Delta t = 1/200$. The middle column shows the implicit upwind finite volume scheme at the same time step while the right column shows the implicit scheme operating at the much larger time step $\Delta t = 1/20$.

The true solution is discontinuous with the initial condition just moving to right and up while a wedge-formed by the boundary condition comes in from the left. All schemes show an excessive smearing of the front with more smearing exhibited by the implicit scheme at the same time step. The implicit scheme is unconditionally stable but the smearing is very pronounced at the large time step.

3.4 Numerical Diffusion

The excessive smearing of the simple upwind scheme will be explained in the following. For that we consider the simple one-dimensional equation in the unbounded domain

$$\begin{aligned} \partial_t u(x,t) + a\partial_x u(x,t) &= 0, && (\text{in } \mathbb{R}\times\mathbb{R}^+) \\ u(x,0) &= u_0, && (t=0) \end{aligned}$$

with smooth initial condition $u_0(x)$ and $a > 0$. The upper half plane is discretized in space with constant size h in x-direction and Δt in t-direction. The value $u_i^k \approx u(x_i, t^k)$ approximates $u(x,t)$ in the center of cell i at time t^k. The explicit and implicit upwind finite volume schemes are then given by

$$\frac{u_i^k - u_i^{k-1}}{\Delta t} + a\frac{u_i^{k-1} - u_{i-1}^{k-1}}{h} = 0, \qquad \text{(explicit scheme)} \tag{3.26a}$$

$$\frac{u_i^k - u_i^{k-1}}{\Delta t} + a\frac{u_i^{k} - u_{i-1}^{k}}{h} = 0, \qquad \text{(implicit scheme).} \tag{3.26b}$$

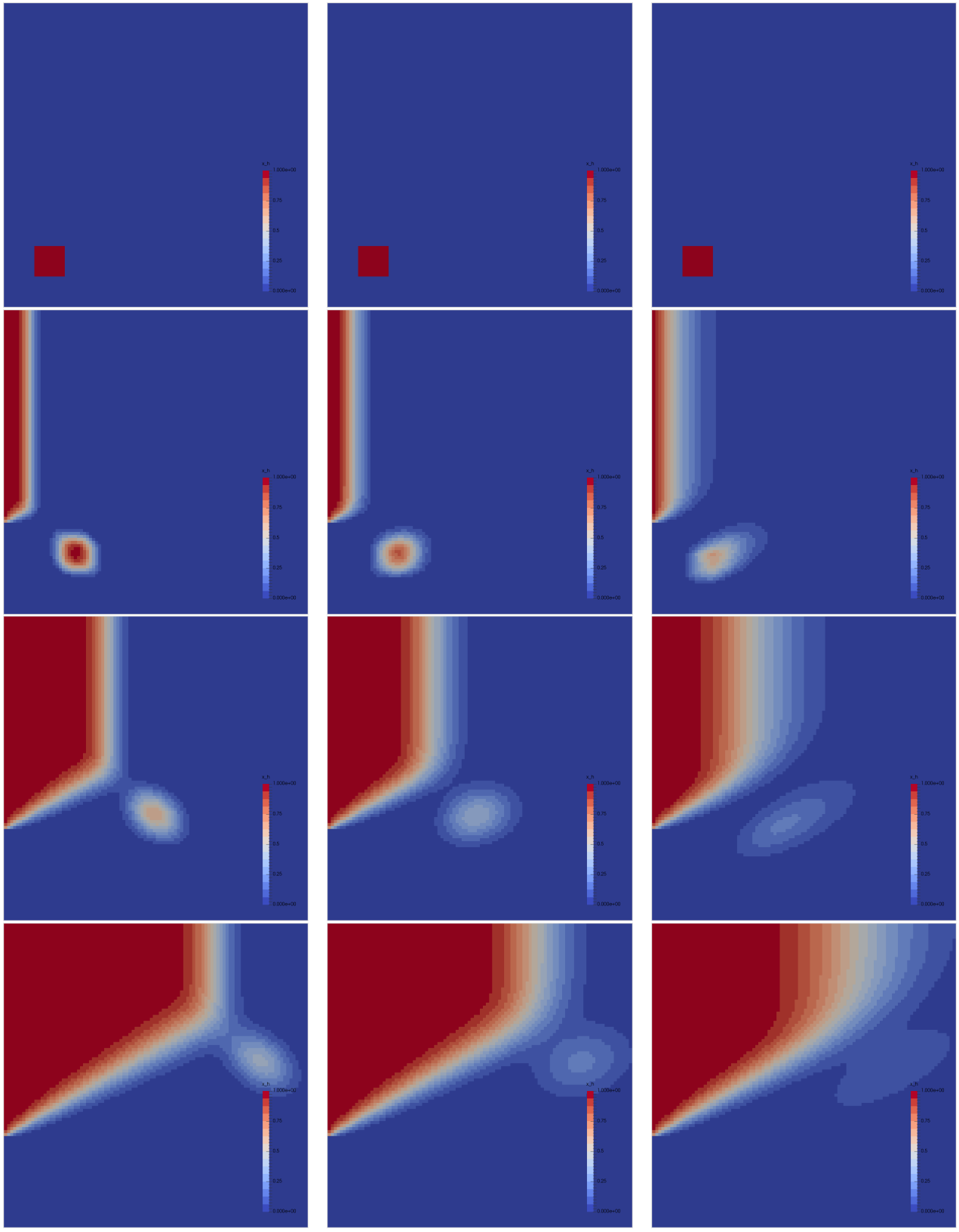


Figure 3.3: Results for a model problem with discontinuous initial condition, $\beta = (\cos(\pi 30/180), \sin(\pi 30/180))^T$, $h = 1/100$. First column: explicit scheme with $\Delta t = 1/200$, runtime 2.8s, middle column: implicit scheme with $\Delta t = 1/200$, runtime 6.3s right column: implicit scheme with $\Delta t = 1/20$, runtime 0.9s.

Using Taylor expansion observe the exact, smooth solution satisfies

$$\begin{aligned}\frac{u(x,t)-u(x,t-\Delta t)}{\Delta t} &= \frac{\partial u}{\partial t}(x,t) - \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(x,t) + \mathcal{O}(\Delta t^2),\\ \frac{u(x,t)-u(x-h,t)}{h} &= \frac{\partial u}{\partial x}(x,t) - \frac{h}{2}\frac{\partial^2 u}{\partial x^2}(x,t) + \mathcal{O}(h^2)\,.\end{aligned}$$

Moreover, for u smooth enough we have:

$$\begin{aligned}\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial t} = 0 \begin{cases} \Rightarrow \frac{\partial^2 u}{\partial t^2} + a\frac{\partial^2 u}{\partial x \partial t} &= 0 \\ \Rightarrow \frac{\partial^2 u}{\partial t \partial x} + a\frac{\partial^2 u}{\partial x^2} &= 0 \end{cases}\\ \Rightarrow \frac{\partial^2 u}{\partial t^2} - a^2\frac{\partial^2 u}{\partial x^2} = 0 \quad \Leftrightarrow \quad \frac{\partial^2 u}{\partial t^2} = a^2\frac{\partial^2 u}{\partial x^2}\,.\end{aligned}$$

Now we combine this to get for the implicit scheme

$$\begin{aligned}&\frac{u(x,t)-u(x,t-\Delta t)}{\Delta t} + a\frac{u(x,t)-u(x-h,t)}{h} =\\ &= \left(\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}\right)\Big|_{(x,t)} - \left(\frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2} + \frac{ah}{2}\frac{\partial^2 u}{\partial x^2}\right)\Big|_{(x,t)} + \mathcal{O}(\Delta t^2 + h^2)\\ &= \left(\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}\right)\Big|_{(x,t)} - \left(\frac{a^2\Delta t + ah}{2}\frac{\partial^2 u}{\partial x^2}\right)\Big|_{(x,t)} + \mathcal{O}(\Delta t^2 + h^2)\,.\end{aligned}$$

We may interpret this result as follows:

- Inserting the exact solution into the difference equation (3.26b) has a leading order error term that has the form of a diffusion term with diffusion coefficient $D_{\text{imp}}(h,\Delta t) = \frac{a^2\Delta t + ah}{2}$. The consistency order is therefore $O(\Delta t + h)$ and the scheme is first order accurate for smooth solutions.
- Alternatively, we may interpret the scheme (3.26b) as second-order accurate discretization of the second-order parabolic PDE

  $$\partial_t u + a\partial_x u - D_{\text{imp}}(h,\Delta t)\partial_{xx} u = 0$$

  of convection-diffusion type. This explains the diffusive character of the numerical solutions.

The same analysis can be carried out for the explicit scheme. Using the expansion point $(x, t-\Delta t)$ in the Taylor expansion we obtain

$$\begin{aligned}\frac{u(x,t)-u(x,t-\Delta t)}{\Delta t} &= \frac{\partial u}{\partial t}(x,t-\Delta t) + \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(x,t-\Delta t) + \mathcal{O}(\Delta t^2),\\ \frac{u(x,t-\Delta t)-u(x-h,t-\Delta t)}{h} &= \frac{\partial u}{\partial x}(x,t-\Delta t) - \frac{h}{2}\frac{\partial^2 u}{\partial x^2}(x,t-\Delta t) + \mathcal{O}(h^2),\end{aligned}$$

and we combine this to

$$\begin{aligned}
&\frac{u(x,t)-u(x,t-\Delta t)}{\Delta t}+a\frac{u(x,t)-u(x-h,t)}{h}= \\
&\quad=\left(\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}\right)\Big|_{(x,t-\Delta t)}+\left(\frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}-\frac{ah}{2}\frac{\partial^2 u}{\partial x^2}\right)\Big|_{(x,t-\Delta t)}+\mathcal{O}(\Delta t^2+h^2) \\
&\quad=\left(\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}\right)\Big|_{(x,t-\Delta t)}-\left(\frac{ah-a^2\Delta t}{2}\frac{\partial^2 u}{\partial x^2}\right)\Big|_{(x,t-\Delta t)}+\mathcal{O}(\Delta t^2+h^2)\,.
\end{aligned}$$

Again, the scheme may be interpreted as a second-order accurate discretization of a convection-diffusion equation but now with the diffusion coefficient $D_{\exp}(h,\Delta t)=\frac{a(h-a\Delta t)}{2}$ which is smaller than in the implicit case. The CFL condition ensures that $a\Delta t\le h$ and therefore $D_{\exp}\ge 0$. In case $h=a\Delta t$ the leading order error term vanishes and in fact the scheme becomes exact. Unfortunately this does only hold for this simple equation in one space dimension when using equistant mesh in space and time.

3.5 One-dimensional Linear Systems

In this section we consider the one-dimensional problem with m components

$$\begin{aligned}
\partial_t u(x,t)+\partial_x F(u(x,t))&=0, && (\text{in } \Omega\times\Sigma=(a,b)\times\mathbb{R}^+)\\
u(x,0)&=u_0, && (t=0),
\end{aligned}$$

with

$$F(u)=Bu,$$

$B\in\mathbb{R}^{m\times m}$ a constant and real diagonalizable matrix, and appropriate boundary conditions, cf. Section 2.1 (we will comment on appropriate boundary conditions below).

Let us introduce the following notation for meshes in one space dimension. The interval (a,b) is partitioned into elements (subintervals) $\mathcal{E}_h=\{e_1,\dots,e_{N_h}\}$ given by

$$a=x_0<x_1<\ldots<x_{\ell-1}<x_\ell<\ldots<x_{N_h}=b,\qquad e_\ell=(x_{\ell-1},x_\ell)\,.$$

Discrete functions may be discontinuous at the interior points $\mathcal{F}_h^i=\{x_1,\dots,x_{N_h-1}\}$ and we have, as before

$$w^-(x_\ell)=\lim_{\epsilon\to 0+}w(x-\epsilon),\quad w^+(x_\ell)=\lim_{\epsilon\to 0+}w(x+\epsilon),\quad [\![w]\!](x_\ell)=w^-(x_\ell)-w^+(x_\ell).$$

In order to derive the finite volume scheme we multiply the equation by a m-component test function $v \in \left(V_h^0\right)^m$ and integrate:

$$\begin{aligned}
&\int_\Omega [\partial_t u(x,t) + \partial_x F(u(x,t))] \cdot v(x)\, dx = \\
&\quad = d_t(u,v)_\Omega + \sum_{\ell=1}^{N_h} \int_{x_{\ell-1}}^{\ell} \sum_{i=1}^{m} \partial_x F_i(u(x,t)) v_i(x)\, dx \\
&\quad = d_t(u,v)_\Omega + \sum_{\ell=1}^{N_h} \sum_{i=1}^{m} v_i\left(\frac{x_{\ell-1}+x_\ell}{2}\right)\left(F_i(u^-(x_\ell,t)) - F_i(u^+(x_{\ell-1},t))\right) \\
&\quad = d_t(u,v)_\Omega + \sum_{\ell=1}^{N_h-1} [\![F(u(x_\ell,t)) \cdot v(x_\ell)]\!] + F(u(b,t)) \cdot v(b) - F(u(a,t)) \cdot v(a)\,.
\end{aligned}$$

From this we arrive at the semi-discrete scheme by introducing a numerical flux function $\Phi(u)$ at internal interfaces. From Section 1.4 we recall that hyperbolicity implies that $B = RDR^{-1}$ with $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ and regular R consisting columnwise of the eigenvectors $r_1, \ldots, r_m$. $w = R^{-1}u$ transforms a state u to characteristic variables in which the system is diagonal and where upwinding can be done in the usual way depending on the sign of the eigenvalues. Therefore we introduce the matrices

$$\begin{aligned}
D^+ &= \operatorname{diag}(\max(0,\lambda_1), \ldots, \max(0,\lambda_m)), \\
D^- &= \operatorname{diag}(\min(0,\lambda_1), \ldots, \min(0,\lambda_m)),
\end{aligned}$$

and

$$B^+ = RD^+R^{-1}, \qquad B^- = RD^-R^{-1}, \qquad B = B^+ + B^-\,. \tag{3.27}$$

With this we define the numerical flux at an interior point $x \in \mathcal{F}_h^i$ as

$$\Phi_U(u)(x) = B^+u^-(x) + B^-u^+(x)\,. \tag{3.28}$$

The upwind semi-discrete scheme for one-dimensional linear systems then reads as follows. Find $u_h : \Sigma \to (V_h^0)^m$ s. t.:

$$\begin{aligned}
d_t(u_h(t),v)_\Omega + \sum_{\ell=1}^{N_h-1} &\left(B^+u_h^-(x_\ell,t) + B^-u^+(x_\ell,t)\right) \cdot [\![v(x_\ell)]\!] \\
&+ \left(B^+u_h(b,t) + B^-g(b,t)\right) \cdot v(b) \\
&- \left(B^+g(b,t) + B^-u_h(a,t)\right) \cdot v(a) = 0 \qquad \forall v \in (V_h^0)^m\,.
\end{aligned}$$

Remark 3.5.

- This method is called *flux vector splitting* method.
- The boundary conditions are defined with respect to the characteristic variables $w = R^{-1}u$. At $x = a$ only $w_i(a)$ where $\lambda_i > 0$ can be prescribed. At $x = b$ only $w_i(b)$ where $\lambda_i < 0$ can be prescribed.
- Setting $g = 0$ at the boundary is called *absorbing boundary condition*. By determining g from the inside state in an appropriate way one can achieve *reflecting boundary conditions*.
- The definition of the system upwind flux (3.28) coincides with (3.6) for the scalar case $m = 1$.

3.6 Riemann Solvers

Constant Coefficient Case

The upwind flux (3.28) can be interpreted with the help of the solution of the following so-called *Riemann problem*:

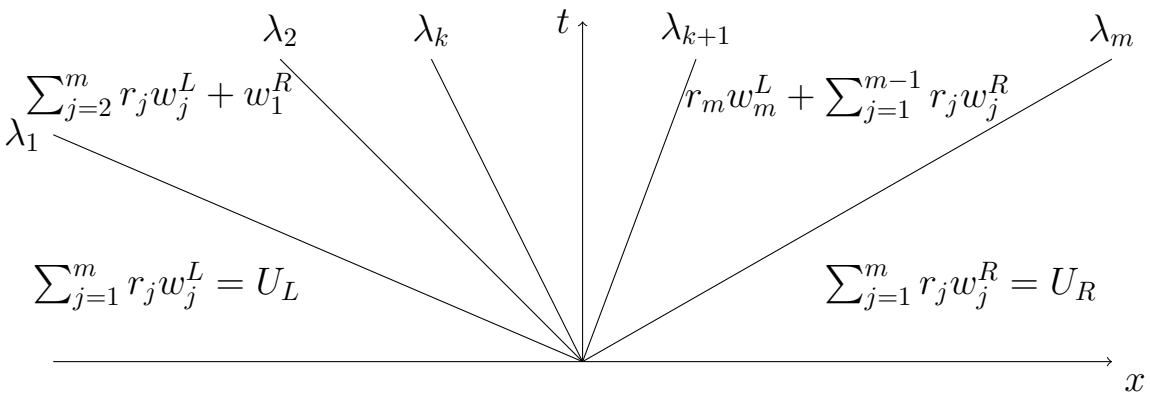
$$\begin{aligned} \partial_t u(x,t) + \partial_x(Bu(x,t)) &= 0, && (\text{in } \mathbb{R} \times \mathbb{R}^+) \\ u(x,0) &= \begin{cases} U_L & x \le 0 \\ U_R & x > 0 \end{cases}, && (t = 0)\,. \end{aligned}$$

Riemann problems are characterized by an initial condition with two constant states and a discontinuity at $x = 0$. Here we assume that B is a constant matrix.

The solution of this Riemann problem can be constructed according to the discussion in Section 1.4. First, transform the left and right states to characteristic variables: $W_L = R^{-1}U_L$ and $W_R = R^{-1}U_R$. Let the eigenvalues be sorted with k eigenvalues negative and $m - k$ eigenvalues non-negative (there may be zero eigenvalues):

$$\lambda_1 \le \lambda_2 \le \ldots \le \lambda_k < 0 \le \lambda_{k+1} \le \ldots \le \lambda_m\,.$$

Then the solution is piecewise constant in space time cones as follows:



The solution along the line $(0,t)$, $t>0$ is given as follows:

$$u(0,t) = \sum_{\lambda_j<0} r_j w_j^R + \sum_{\lambda_j=0} r_j w_j^L + \sum_{\lambda_j>0} r_j w_j^L = \sum_{j=1}^{k} r_j w_j^R + \sum_{j=k+1}^{m} r_j w_j^L$$

and the corresponding flux along the line $(0,t)$, $t>0$ is then

$$\begin{aligned} F(u(0,t)) &= Bu(0,t) = RDR^{-1}\left(\sum_{j=1}^{k} r_j w_j^R + \sum_{j=k+1}^{m} r_j w_j^L\right) \\ &= RDR^{-1}Rw^* = RDw^* = R(D^- + D^+)w^* \\ &= RD^-w^* + RD^+w^* = RD^-W_R + RD^+W_L \\ &= RD^-R^{-1}U_R + RD^+R^{-1}U_L \\ &= B^-U_R + B^+U_L \end{aligned} \tag{3.29}$$

where we used

$$w_j^* = \begin{cases} w_j^R & j \le k \\ w_j^L & j > k \end{cases}.$$

This shows that *the upwind flux may be interpreted as the flux evaluated for the solution of a Riemann problem located at the interface.* It turns out this construction principle is the key to define appropriate numerical fluxes also for nonlinear systems of hyperbolic PDEs such as the Euler equations.

Discontinuous Coefficient Case

The coefficient matrix B may depend on position x. If this dependence is smooth one may put the hyperbolic system in nonconservative form an proceed as shown above. The case of discontinuous coefficient $B(x)$ deserves more thought. Consider the following one-dimensional Riemann problem

$$\partial_t u(x,t) + \partial_x(B(x)u(x,t)) = 0, \qquad (\text{in } \mathbb{R}\times\mathbb{R}^+) \tag{3.30a}$$

$$u(x,0) = \begin{cases} U_L & x \le 0 \\ U_R & x > 0 \end{cases}, \qquad (t=0)\,, \tag{3.30b}$$

$$B(x) = \begin{cases} B_L & x \le 0 \\ B_R & x > 0 \end{cases}. \tag{3.30c}$$

Scalar Case For simplicity let us start with a single component $m=1$. In order to determine what happens at the interface $x=0$ we consider problem

(3.30a) as two problems with an interface condition:

$$\partial_t u_L(x,t) + \partial_x(B_L u_L(x,t)) = 0, \qquad (\text{in } \mathbb{R}^- \times \mathbb{R}^+) \tag{3.31a}$$
$$u_L(x,0) = U_L, \tag{3.31b}$$
$$\partial_t u_R(x,t) + \partial_x(B_R u_R(x,t)) = 0, \qquad (\text{in } \mathbb{R}^+ \times \mathbb{R}^+) \tag{3.31c}$$
$$u_R(x,0) = U_R, \tag{3.31d}$$
$$B_L u_L(0,t) = B_R u_R(0,t) \qquad (\text{flux continuity}). \tag{3.31e}$$

For arbitrary initial states flux continuity demands that B_L and B_R have the same sign: $B_L B_R > 0$. Then system (3.31a) can solved by the method of characteristics. Assume e.g. that $B_L, B_R > 0$, then

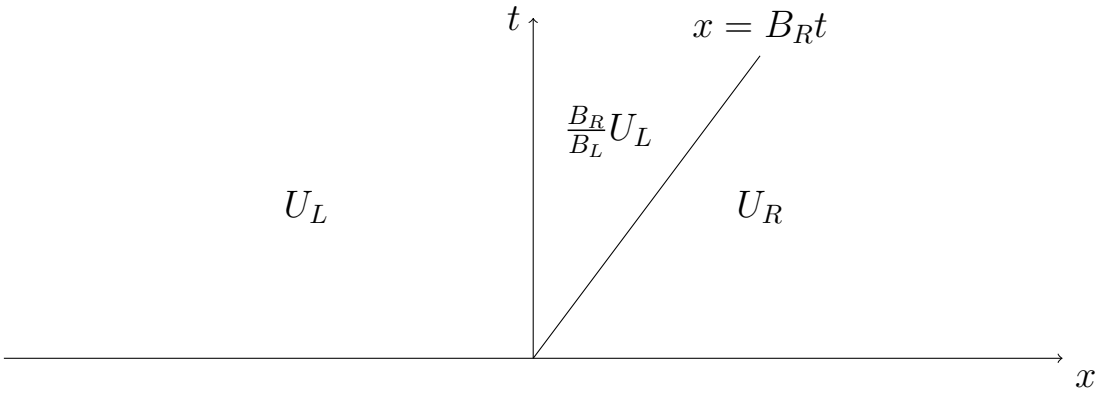
i) $x = 0$ is an outflow boundary for the left domain and $u_L(x,t) = U_L$ for $x \leq 0$.

ii) $x = 0$ is an inflow boundary for the right domain. Flux continuity demands $B_L U_L = B_R u_R(0,t)$ and we get the boundary condition $u_R(0,t) = \frac{B_R}{B_L} U_L$.

iii) By the method of characteristic we obtain in the right domain:

$$u_R(x,t) = \begin{cases} \frac{B_R}{B_L} U_L & x - B_R t \leq 0 \\ U_R & x - B_R t > 0 \end{cases}.$$

In the (x,t)-diagram this is:



System Case This is treated in the same way. However, since waves are going both ways across the interface the states left and right of the interface are determined by the solution of a linear system.

We define the states to the left and right of the interface

$$U_L^* = \lim_{x \to 0-} u_L(x,t), \qquad U_R^* = \lim_{x \to 0+} u_R(x,t), \qquad (\text{for any } t > 0).$$

Due to hyperbolicity, B_L and B_R are diagonalizable with eigenvalues λ_i^L, λ_i^R and eigenvectors r_i^L, r_i^R. The matrices R_L, R_R are formed by the eigenvectors and the diagonal matrices D_L, D_R contain the corresponding eigenvalues. As above we set $B_L^\pm = R_L D_L^\pm R_L^{-1}$, $B_R^\pm = R_R D_R^\pm R_R^{-1}$. By the transformation to characteristic variables we obtain the following representation of the interface states:

$$U_L^* = \sum_{\{i\,:\,\lambda_i^L \geq 0\}} r_i^L (R_L^{-1} U_L)_i + \sum_{\{i\,:\,\lambda_i^L < 0\}} r_i^L \alpha_i, \tag{3.32}$$

$$U_R^* = \sum_{\{i\,:\,\lambda_i^R \leq 0\}} r_i^R (R_R^{-1} U_R)_i + \sum_{\{i\,:\,\lambda_i^R > 0\}} r_i^R \alpha_i. \tag{3.33}$$

The first sum takes into account the waves that reach the boundary from within in the respective domain. The second part describes the contribution coming from the boundary (the minus sign in the second line becomes obvious below). As a first assumption we put

$$\{i\,:\,\lambda_i^L < 0\} = \{i\,:\,\lambda_i^R < 0\} \quad \wedge \quad \{i\,:\,\lambda_i^L > 0\} = \{i\,:\,\lambda_i^R > 0\}, \tag{3.34}$$

i.e. the number of positive (negative) eigenvalues to the left and right coincides (and therefore also the number of zero eigenvalues) and positive and negative eigenvalues are numbered in the same way.

In order to determine the coefficients $\alpha \in \mathbb{R}^{I^*}$, $I^* = \{i\,:\,\lambda_i^L \neq 0\} \subseteq I = \{1, \ldots, m\}$ we exploit flux continuity $B_L U_L^* = B_R U_R^*$. Further notation is needed to handle the case of zero eigenvalues when $m^* = |I^*| < m$. We introduce the "picking-out-matrix" $P \in \mathbb{R}^{I^* \times I}$ defined by

$$(Px)_j = (x)_j \qquad \forall j \in I^*.$$

Observing,

$$B_L U_L^* = \sum_{\{i\,:\,\lambda_i^L \geq 0\}} B_L r_i^L (R_L^{-1} U_L)_i + \sum_{\{i\,:\,\lambda_i^L < 0\}} B_L r_i^L \alpha_i = B_L^+ U_L + R_L D_L^- P^T \alpha,$$

$$B_R U_R^* = \sum_{\{i\,:\,\lambda_i^R \leq 0\}} B_R r_i^R (R_R^{-1} U_R)_i + \sum_{\{i\,:\,\lambda_i^R > 0\}} B_R r_i^R \alpha_i = B_R^- U_R + R_R D_R^+ P^T \alpha.$$

we obtain

$$(R_R D_R^+ - R_L D_L^-) P^T \alpha = S\alpha = B_L^+ U_L - B_R^- U_R. \tag{3.35}$$

The linear system (3.35) has a unique solution if $S \in \mathbb{R}^{I \times I^*}$ has rank m^* and

$$\begin{aligned} \operatorname{span}\left\{r_i^R : \lambda_i^R > 0\right\} + \operatorname{span}\left\{r_i^L : \lambda_i^R < 0\right\} = \\ \operatorname{span}\left\{r_i^L : \lambda_i^R > 0\right\} + \operatorname{span}\left\{r_i^R : \lambda_i^R < 0\right\} \end{aligned} \tag{3.36}$$

and is then given by

$$\alpha = \left(S^T S\right)^{-1} S^T \left(B_L^+ U_L - B_R^- U_R\right). \tag{3.37}$$

The flux can then be computed from either side of the interface, e.g. from the left:

$$\begin{aligned} \hat{F}(U_L, U_R) &= B_L U_L^* = B_L^+ U_L + R_L D_L^- P^T \alpha \\ &= B_L^+ U_L + R_L D_L^- P^T \left(S^T S\right)^{-1} S^T \left(B_L^+ U_L - B_R^- U_R\right) \end{aligned} \tag{3.38}$$

For comparison consider the case of constant coefficients in this framework. Flux continuity then becomes

$$\begin{aligned} & BU_L^* = BU_R^* \\ \Leftrightarrow \quad & \sum_{\{i\,:\,\lambda_i>0\}} r_i \lambda_i (R^{-1}U_L)_i + \sum_{\{i\,:\,\lambda_i<0\}} r_i \lambda_i \alpha_i = \sum_{\{i\,:\,\lambda_i<0\}} r_i \lambda_i (R^{-1}U_R)_i + \sum_{\{i\,:\,\lambda_i>0\}} r_i \lambda_i \alpha_i \end{aligned}$$

Since the r_i are linearly independent we must have

$$\alpha_i = (R^{-1}U_R)_i \text{ for } \lambda_i < 0, \quad \alpha_i = (R^{-1}U_L)_i \text{ for } \lambda_i > 0.$$

Inserting into one of both sides yields

$$\begin{aligned} \hat{F}(U_L, U_R) = BU_L^* &= \sum_{\{i\,:\,\lambda_i>0\}} r_i \lambda_i (R^{-1}U_L)_i + \sum_{\{i\,:\,\lambda_i<0\}} r_i \lambda_i \alpha_i \\ &= \sum_{\{i\,:\,\lambda_i>0\}} r_i \lambda_i (R^{-1}U_L)_i + \sum_{\{i\,:\,\lambda_i<0\}} r_i \lambda_i (R^{-1}U_R)_i \\ &= B^+ U_L + B^- U_R. \end{aligned}$$

Chapter 4

Higher-order Discontinuous Galerkin Methods

In this section we present a numerical method to solve the original problem (1.1) which is repeated for convenience here. Let $u : \Omega \times \Sigma \to \mathbb{R}^m$ be the solution of the hyperbolic first-order system

$$\partial_t u(x,t) + \nabla \cdot F(u(x,t),x,t) = f(u(x,t),x,t), \qquad \text{in } U = \Omega \times \Sigma, \tag{4.1a}$$

$$u(x,t) = u_0(x), \qquad \text{at } t = 0, \tag{4.1b}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded domain, $\Sigma = (t_0, t_0+T)$ is a time interval of interest and $F(u,x,t) = [F_1(u,x,t), \ldots, F_n(u,x,t)]$ is the matrix valued flux function with columns $F_j(u,x,t)$.

4.1 Space Discretization with Discontinuous Galerkin

For any test function v being piecewise smooth on the mesh $\mathcal{E}_h$ there holds

$$\begin{aligned}
&\int_\Omega \left[\partial_t u + \sum_{j=1}^{d} \partial_{x_j} F_j(u,x,t)\right] \cdot v \, dx = \\
&\quad = d_t(u,v)_\Omega + \sum_{e\in\mathcal{E}_h} \sum_{j=1}^{d} \sum_{i=1}^{m} \int_e \left(\partial_{x_j} F_{i,j}(u,x,t)\right) v_i \, dx \\
&\quad = d_t(u,v)_\Omega + \sum_{e\in\mathcal{E}_h} \sum_{j=1}^{d} \sum_{i=1}^{m} \Big[- \int_e F_{i,j}(u,x,t)\, \partial_{x_j} v_i \, dx \\
&\qquad + \int_{\partial e} F_{i,j}(u,s,t) v_i n_j \, ds \Big] \qquad (4.2) \\
&\quad = d_t(u,v)_\Omega + \sum_{e\in\mathcal{E}_h} \Big[- \int_e F(u,x,t) : \nabla v \, dx + \int_{\partial e} (F(u,s,t)n) \cdot v \, ds \Big] \\
&\quad = d_t(u,v)_\Omega - \sum_{e\in\mathcal{E}_h} \int_e F(u,x,t) : \nabla v \, dx \\
&\qquad + \sum_{f\in\mathcal{F}_h^i} \int_f [\![(F(u,s,t)n) \cdot v]\!] \, ds + \sum_{f\in\mathcal{F}_h^{\partial\Omega}} \int_f (F(u,s,t)n) \cdot v \, ds \, .
\end{aligned}$$

Next step is to construct a numerical flux function. Here we only consider the linear constant coefficient case $F_j(u) = B_j u$. Then the normal flux is

$$F(u,x,t)n = \sum_{j=1}^{d} F_j(u)n_j = \sum_{j=1}^{d}(B_j u)n_j = \left(\sum_{j=1}^{d} n_j B_j\right) u = B_n u\,. \tag{4.3}$$

Due to hyperbolicity the matrix $B_n = \left(\sum_{j=1}^{d} n_j B_j\right)$ is real diagonalizable for all $n \in \mathbb{R}^d$ and we may use the numerical flux function (3.28) based on flux vector splitting:

$$\Phi_U(u, B_n)(x) = B_n^+ u^-(x) + B_n^- u^+(x). \tag{4.4}$$

In order to achive higher order we employ a finite element space with higher-order polynomials:

$$V_h^q = \left\{v \in L^2(\Omega) \,:\, v|_e = p \circ \mu_e^{-1}, p \in \mathbb{P}^{q,d}\right\} \tag{4.5}$$

where the differentiable and invertible map

$$\mu_e : \hat{E} \to e$$

maps the reference element $\hat{E}$ to an element $e \in \mathcal{E}_h$ and the multivariate polynomials of degree q in d space dimensions are given by

$$\mathbb{P}^{q,d} = \left\{ \begin{array}{l} \left\{p \,:\, p(x_1,\ldots,x_d) = \displaystyle\sum_{0\le\|\alpha\|_1\le q} c_\alpha x_1^{\alpha_1}\cdot\ldots\cdot x_d^{\alpha_d}\right\} \quad (\hat{E} \text{ simplex}), \\ \left\{p \,:\, p(x_1,\ldots,x_d) = \displaystyle\sum_{0\le\|\alpha\|_\infty\le q} c_\alpha x_1^{\alpha_1}\cdot\ldots\cdot x_d^{\alpha_d}\right\} \quad (\hat{E} \text{ cube}), \end{array} \right.$$

depending on the type of element.

The upwind semi-discrete scheme for multi-dimensional linear hyperbolic systems then reads as follows. Find $u_h : \Sigma \to (V_h^q)^m$ s. t.:

$$d_t(u,v)_\Omega + b_{DG}(u(t),v) = l_{DG}(v)\,, \tag{4.6}$$

where the DG spatial bilinear form is given by

$$\begin{aligned} b_{DG}(u,v) = &-\sum_{e\in\mathcal{E}_h}\int_e F(u,x,t) : \nabla v\, dx \\ &+\sum_{f\in\mathcal{F}_h^i}\int_f (B_n^+ u^-(s,t) + B_n^- u^+(s,t))\cdot[\![v]\!]\, ds \\ &+\sum_{f\in\mathcal{F}_h^{\partial\Omega}}\int_f (B_n^+ u^-(s,t))\cdot v\, ds \end{aligned} \tag{4.7}$$

and the right hand side by

$$l_{DG} = (f(t), v)_\Omega - \sum_{f \in \mathcal{F}_h^{\partial\Omega}} \int_f (B_n^- g(s,t)) \cdot v \, ds \,. \tag{4.8}$$

4.2 Runge-Kutta Methods

The Runge-Kutta method for solving an ordinary differential equation coming from semi-discretizing a partial differential equation reads in Shu-Osher form as follows:

1. $u_h^{(0)} = u_h^k$.
2. For $i = 1, \ldots, s \in \mathbb{N}$, find $u_h^{(i)} \in (V_h^q)^m$:

$$\sum_{j=0}^{s} \Big[a_{ij} \left(u_h^{(j)}, v \right)_\Omega + b_{ij} \Delta t^k \left(b_{DG} \left(u_h^{(j)}, v \right) - l_{DG}(v) \right) \Big] = 0 \qquad \forall v \in V_h(t^{k+1})$$

3. $u_h^{k+1} = u_h^{(s)}$.

An s-stage scheme is given by the parameters

$$A = \begin{bmatrix} a_{10} & \ldots & a_{1s} \\ \vdots & & \vdots \\ a_{s0} & \ldots & a_{ss} \end{bmatrix}, \quad B = \begin{bmatrix} b_{10} & \ldots & b_{1s} \\ \vdots & & \vdots \\ b_{s0} & \ldots & b_{ss} \end{bmatrix}, \quad d = (d_0, \ldots, d_s)^T \,.$$

Since we want to solve at most systems of the size of the stationary problem we restrict ourselves to either explicit or diagonally implicit schemes. Some examples are given by the following list.

- One step θ scheme:

$$A = \begin{bmatrix} -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1-\theta & \theta \end{bmatrix}, \quad d = (0,1)^T \,.$$

  Explicit/implicit Euler ($\theta \in \{0, 1\}$), Crank-Nicolson ($\theta = 1/2$).

- Strong stability preserving second order explicit method (Heun):

$$A = \begin{bmatrix} -1 & 1 & 0 \\ -1/2 & -1/2 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix}, \quad d = (0,1,1)^T \,.$$

- Alexander's two-stage second order strongly S-stable scheme:

$$A = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & \alpha & 0 \\ 0 & 1-\alpha & \alpha \end{bmatrix}, \quad d = (0, \alpha, 1)^T$$

  with $\alpha = 1 - \sqrt{2}/2$.

- Fractional step θ, three stage second order strongly A-stable method:

$$A = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} \theta\theta' & 2\theta^2 & 0 & 0 \\ 0 & 2\theta\theta' & 2\theta^2 & 0 \\ 0 & 0 & \theta\theta' & 2\theta^2 \end{bmatrix}, \quad d = (0, \theta, 1-\theta, 1)^T$$

  with $\theta = 1 - \sqrt{2}/2$, $\theta' = 1 - 2\theta = \sqrt{2} - 1$.

- Third order SSP Runge-Kutta method:

$$A = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -3/4 & -1/4 & 1 & 0 \\ -1/3 & 0 & -2/3 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 2/3 & 0 \end{bmatrix}, \quad d = (0, 1, 1/2, 1)^T$$

4.3 Numerical Results

Figure 4.1 shows results for the same problem treated in Section 3.3 using explicit time discretization. We compare formally first, second and third order accurate schemes using different mesh as well as time step sizes which lead approximately to the same total computation time. Clearly the higher order schemes outperform the lower order schemes but all higher order schemes exhibit unphysical oscillations. We do not treat limiter methods here to enforce a maximim principle. These methods remove a lot of the sharpness of the higher order schemes close to the discontinuity. If the unphysical oscillations do not hurt, just accept them. Of course, when solving nonlinear problems limiters are needed to converge towards to enforce a selection principle.

Figure 4.2 shows results for using implicit time discretizations. We compare increasing the order (in space and time) while keeping the spatial and temporal mesh size constant. The results illustrate that the schemes are stable and the accuracy improves as the order is increased. However, the same comment applies with respect to unphysical oscillations.

Implicit versus Explicit

The question of explicit vs. implicit is not easy to answer. In general, explicit methods are to be prefered with hyperbolic problems as the ordinary differential equations arising after semi-discretization are typically not stiff. However they might become so when the data varies highly, such as the velocity magnitude in a porous medium flow problem with wells. Another application of implicit methods is with quasi-stationary problems.

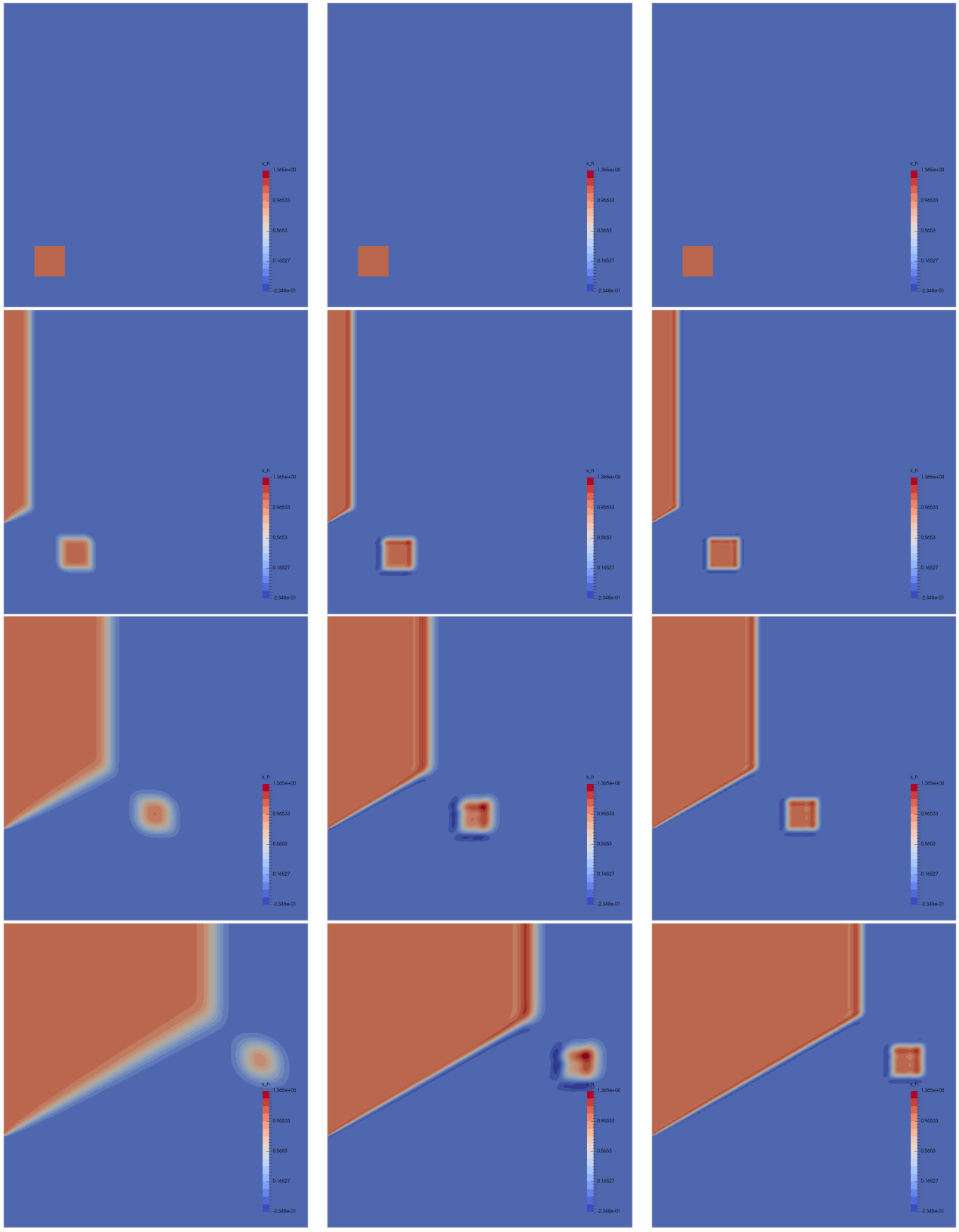


Figure 4.1: Results for a model problem with discontinuous initial condition using explicit higher order methods, $\beta = (\cos(\pi 30/180), \sin(\pi 30/180))^T$. First column: order 1, $h = 1/400$, runtime 160s, middle column: order 2, $h = 1/200$, runtime 55s, right column: order 3, $h = 1/100$, $\Delta t = 1/400$, runtime 34s.

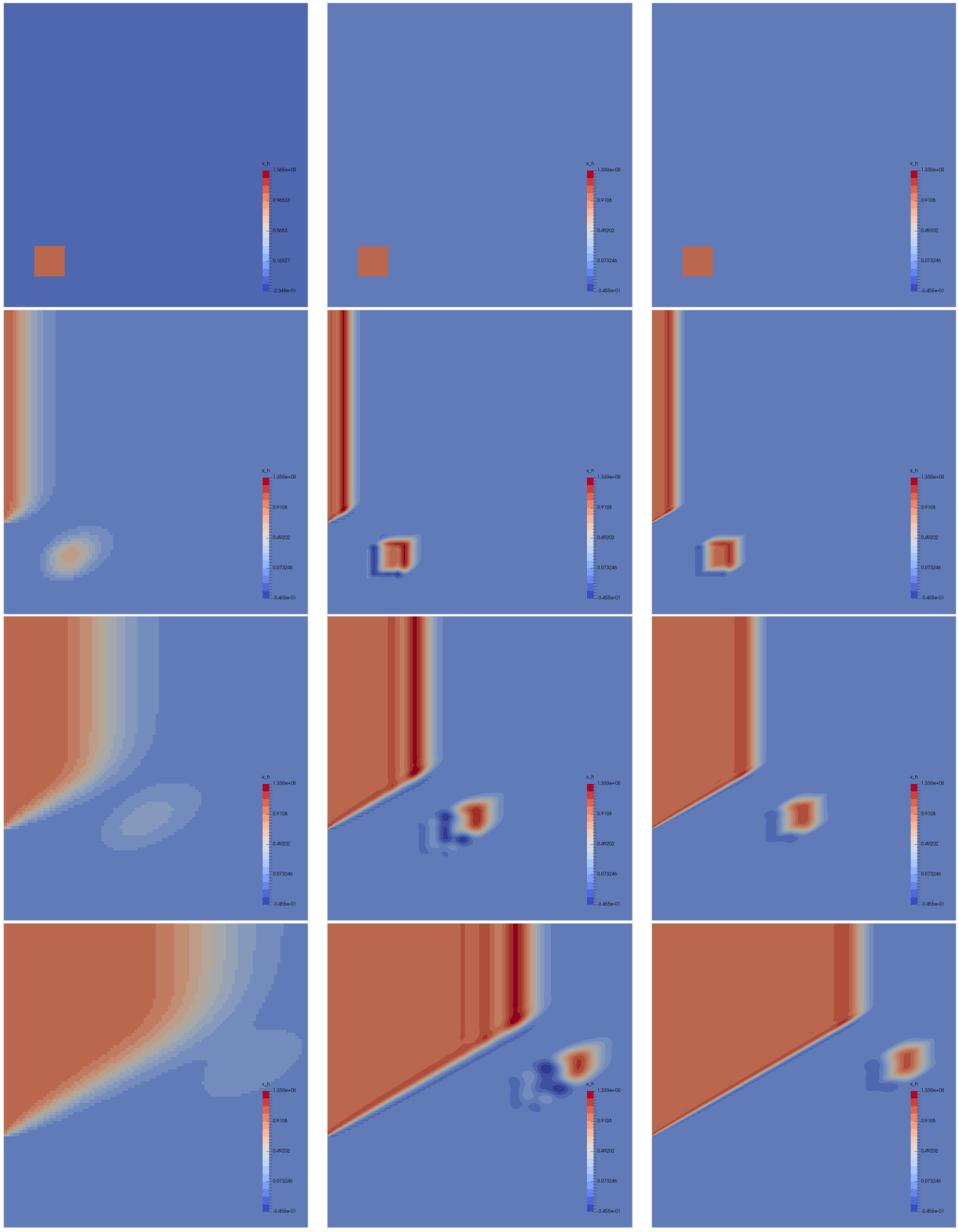


Figure 4.2: Results for a model problem with discontinuous initial condition using implicit higher order methods using $h = 1/100$, $\Delta t = 1/40$, $\beta = (\cos(\pi 30/180), \sin(\pi 30/180))^T$, $h = 1/100$. First column: order 1, runtime 1s, middle column: order 2, runtime 10s, right column: order 3, runtime 53s.

Bibliography

[1] Timothy Barth and Mario Ohlberger. *Finite Volume Methods: Foundation and Analysis*. John Wiley & Sons, Ltd, 2004.

[2] V. Dolejší and M. Feistauer. A semi-implicit discontinuous galerkin finite element method for the numerical solution of inviscid compressible flow. *Journal of Computational Physics*, 198(2):727 – 746, 2004.

[3] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations*. Cambridge University Press, 1996.

[4] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner, 1986. `http://www.mis.mpg.de/preprints/ln/lecturenote-2805.pdf`.

[5] J. Jin. *The Finite Element Method in Electromagnetics*. John Wiley & Sons, 2. edition, 2002.

[6] R. J. Leveque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[7] M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*, volume 13 of *Texts in Applied Mathematics*. Springer, 1993.

[8] W. I. Smirnow. *Lehrgang der höheren Mathematik - Teil II*. VEB Verlag der deutschen Wissenschaften, 15. edition, 1981.