Eidgenössische
Technische Hochschule
Zürich

Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

Error Estimators for the Position of Discontinuities in Hyperbolic Conservation Laws with Source Terms which are solved using Operator Splitting

R. Jeltsch and P. Klingenstein

Research Report No. 97-16
October 1997

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Error Estimators for the Position of Discontinuities in Hyperbolic Conservation Laws with Source Terms which are solved using Operator Splitting

R. Jeltsch and P. Klingenstein

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Research Report No. 97-16 October 1997

Abstract

When computing numerical solutions of hyperbolic conservation laws with source terms, one may obtain spurious solutions — these are unphysical solutions that only occur in numerics such as shock waves moving with wrong speeds, cf. [7], [2], [1], [10], [3]. Therefore it is important to know how errors of the location of a discontinuity can be controlled.

To derive appropriate error-estimates and to use them to control such errors, is the aim of our investigations in this paper. We restrict our considerations to numerical solutions which are computed by using a splitting method. In splitting methods, the homogeneous conservation law and an ordinary differential equation (modelling the source term) are solved separately in each time step.

Firstly, we derive error-estimates for the scalar Riemann problem. The analysis shows that the local error of the location of a discontinuity mainly consists of two parts. The first part is introduced by the splitting and the second part is due to smearing of the discontinuity.

Next, these error-estimates are used to construct an adaptation of the step size so that the error of the location of the discontinuity remains sufficiently small. The adaptation is applied to several examples, which are a scalar problem, a simplified combustion model, and the one-dimensional inviscid reacting compressible Euler equations. All the examples show that the adaptation based on the derived error-estimates works well.

The theory can also be extended to planar two-dimensional problems.

Keywords: adaptation, error-estimates, operator splitting, shock location, stiff source terms

Subject Classification: 35L65, 65M15, 76L05, 76M25

1 Introduction

Hyperbolic conservation laws often arise in models of physical processes that ignore the effects of dissipative and dispersive mechanisms. In gas dynamics, for example, hyperbolic conservation laws are obtained if viscous effects and heat conduction are neglected. Source terms arise in various contexts. We are interested in those that are due to the physical model, as they occur in non-equilibrium or in chemically reacting gas dynamics. In the latter context the problem often is stiff, which means that the time scale of the source term is of orders of magnitude different from that of the fluid dynamics.

Solving hyperbolic conservation laws with stiff source terms numerically causes special difficulties. Often one is not interested in resolving the solution profile on the finest scale, but wants to compute the solution on a grid that is appropriate to the fluid dynamics. On the one hand, one has to be careful to handle the stiff source term in a stable manner. On the other hand, this is not sufficient to get a physically correct solution. When computing numerical solutions of hyperbolic conservation laws with source terms, one may obtain spurious solutions — these are unphysical solutions that only occur in numerics such as shock waves moving at wrong speeds. This phenomenon is due to the coupling of the source term and the fluid dynamics.

Wrong propagation speeds have been observed by several authors, e.g. [7], [2], [1], [10]. In [7], LeVeque and Yee investigate a scalar advection equation with a nonlinear source term, which has two stable and one unstable equilibrium state. They observe nonphysical numerical shock speeds for piece-wise constant initial data. Their investigations show that these problems are caused by the smearing of the discontinuity. They conclude that spatial resolution is as important as temporal resolution.

In [2], Colella, Majda and Roytburd consider reacting shock waves. They show for a simplified combustion model and for the Navier-Stokes equations that dynamically stable weak detonations occur in bifurcating wave patterns for sufficiently small heat release or large reaction rate. Similar wave patterns are also obtained in inviscid calculations for relatively large step sizes, where a precursor numerical weak detonation wave is moving at the speed of one grid cell per time step. Those solutions are incorrect as the bifurcating wave patterns vanish for smaller step sizes.

In [1], Berkenbosch derives a condition on the ignition value for the simplified combustion model and the reacting Euler equations with ignition temperature kinetics such, that the numerical solutions exhibit correct wave speeds even for relatively large step sizes.

In [10], Pember considers relaxation problems. He proposes criteria ensuring that the numerical methods do not produce spurious solutions as the relaxation time vanishes. These criteria are, firstly, that the solution has to tend to the solution of the equilibrium equation as the relaxation time vanishes, and secondly, that a certain subcharacteristic condition has to be satisfied.

This paper, a brief version of [4], is concerned with the exactness of numerical locations of discontinuities. Estimates are derived for scalar Riemann problems with stiff source terms, in order to be able to control the error of the location of a discontinuity[1]. This work is done for numerical solutions that are computed by using a splitting method. This means that in each time step the homogeneous conservation law and an ODE modelling the source term are solved separately. Splitting methods are a popular approach to solve conservation laws with source terms. Their advantage is that good numerical methods exist for each of the subproblems. Furthermore, the analysis of wrong propagation speeds is relatively easy in this case. With a Strang splitting, second order accuracy can be achieved. The mentioned numerical difficulties are not due to splitting, but they also occur if other methods are used, cf. [7].

The analysis shows that the local error of the location of a discontinuity mainly consists of two parts — one part that is introduced by the splitting and another part that is due to smearing. Both parts are, apart from the influence of the discretisation errors of the solvers used, of second order in the step size multiplied by the size of the source term.

Based on those error-estimates, an adaptation of the step size is constructed in order to keep the error of the location of the discontinuity sufficiently small. The adaptation is applied to several examples. These are a scalar problem, a simplified combustion model, and the one-dimensional inviscid reacting compressible Euler equations. The adaptation constructed here works well for all these examples. In [4], the theory is also extended to planar two-dimensional problems.

We also applied the error-estimates to the simplified combustion model described in [1], where the condition on the ignition value proposed by Berkenbosch was fulfilled. The results showed relatively large approximated errors for large step sizes, although the numerical solutions exhibited correct wave speeds. Consequently, the error-estimates derived for scalar piece-wise constant problems would have to be improved in order to get good results for numerical solutions computed with large step sizes and showing correct speeds. The step sizes obtained by our adaptation ensure that the wave speed is sufficiently correct, and yet are not unreasonable small. So if one wants to resolve the solution profile and therefore use (locally) small step sizes, the adaptation based on the error-estimates presented here gives satisfactory results.

This paper is organised as follows. Section 2 introduces the problem, and a scalar example shows the difficulties that arise because of smeared-out shock profiles in Section 3. In Section 4, estimators for the error of the shock location are derived, which are used to construct an adaptation of the step size in Subsection 4.7. Finally, in Section 5, three numerical examples are presented.

[1] which includes shocks and contact discontinuities

2 The Problem

Let us consider the scalar equation

$$u_t + f(u)_x = q(u) \tag{1}$$

where $u(x,t)$, $f(u)$, $q(u) \in \mathrm{I\!R}$ and $x \in \mathrm{I\!R}$, $t \geq 0$. We study the Riemann problem on the time interval $[t_n, t_{n+1}]$ with

$$u(x,t_n) = \begin{cases} u_L, & x < \sigma \\ u_R, & x > \sigma \end{cases} . \tag{2}$$

We consider entropy solutions consisting of two states that only depend on the time t. We want these states to be the left and right state of a Riemann problem because we are only interested in the interaction between the source and the shock. Consequently, the solution $u(x,t)$ of (1), (2) should be

$$u(x,t) = \begin{cases} u_L(t), & x < \sigma(t) \\ u_R(t), & x > \sigma(t) \end{cases} , \ t \in [t_n, t_{n+1}] \tag{3}$$

such that the two states are separated by a shock curve $\sigma(t)$. $\sigma(t)$ is determined by integrating the so called jump condition

$$\dot{\sigma}(t) = \frac{f(u_L(t)) - f(u_R(t))}{u_L(t) - u_R(t)} \tag{4}$$

where we assume $u_L(t) \neq u_R(t)$. This jump condition can be derived from the weak formulation of the integral form of (1), which has to be fulfilled by the solution of (1). To ensure that the solution (3) is the entropy satisfying weak solution, we impose the following entropy condition (cf. [6]) on the discontinuity.

Definition 2.1 *$u(x,t)$ defined by (3) is the entropy solution if the discontinuity travelling with speed $\dot{\sigma}(t)$ given by (4) has the property that*

$$\frac{f(u) - f(u_L(t))}{u - u_L(t)} \geq \dot{\sigma}(t) \geq \frac{f(u) - f(u_R(t))}{u - u_R(t)} \tag{5}$$

for all u between $u_L(t)$ and $u_R(t)$.

Remark 2.2 *In fact, one of the two inequalities in (5) with $\dot{\sigma}(t)$ on one side suffices because they are equivalent.*

The exact solution of (1), (2) is given explicitely in Subsection 4.3.

The numerical solution of (1) is computed using a time splitting method, which means that in each time step the homogeneous conservation law

$$u_t + f(u)_x = 0 \tag{6}$$

and the ODE

$$u_t = q(u) \tag{7}$$

are solved separately. To describe the numerical approximations we use the following notations.

The spatial mesh points are denoted by $x_j = j\triangle x$ with $j \in \mathbb{Z}$. $\triangle x$ is the spatial step size. Likewise, t_n with $n \in \mathbb{N} \cup \{0\}$ stands for the discrete time levels. $\triangle t$ is the step size in time so that $t_{n+1} = t_n + \triangle t$. The step sizes $\triangle x$ and $\triangle t$ have to fulfil the CFL-stability-condition. For a three-point-scheme, the CFL-condition reads

$$CFL := |\max\{f'(u)\}| \frac{\triangle t}{\triangle x} < 1, \tag{8}$$

where we introduced the parameter CFL. The discrete numerical solution u_i^n at time t_n can be interpreted as the mean value of the cell i, $[x_i - \frac{1}{2}\triangle x, x_i + \frac{1}{2}\triangle x)$, of an approximate solution $u^n(x)$ to $u(x, t_n)$

$$u_i^n = \frac{1}{\triangle x} \int_{x_i - \frac{1}{2}\triangle x}^{x_i + \frac{1}{2}\triangle x} u^n(x) dx. \tag{9}$$

Furthermore, u^n may denote the discrete numerical solution at time level t_n, $u^n(x)$ or $u(x, t_n)$. u_L^n and u_R^n are the numerical approximations of $u_L(t_n)$ and $u_R(t_n)$, respectively. Finally, σ^n is the location of the discontinuity at time t_n of the numerical solution, and $\sigma(t_n)$ is the location of the discontinuity at time t_n of the exact solution of (1).

The solution operator for the ODE is called $L_q(\triangle t)$ and the one for the homogeneous conservation law $L_f(\triangle t)$, respectively. In the following, we denote solutions of the homogeneous conservation law by v and solutions of the ODE by w instead of u, the notations that were just made for u are used analogously for v and w. This reads as follows:

- Solve $w_t = q(w)$ approximately by $w^{n+1} = L_q(\triangle t) w^n$,
- solve $v_t + f(v)_x = 0$ approximately by $v^{n+1} = L_f(\triangle t) v^n$.

Here each of the solution operators $L_q(\triangle t)$ and $L_f(\triangle t)$ may denote either a discrete or a continuous operator between two function spaces.

$L_q(\triangle t)$ may be the exact or an approximate solution operator with incremental function Φ. Φ depends on the numerical solution at the time levels t_n and/or t_{n+1}, on the time step $\triangle t$ and on the source q. We write shortly

$$w_j^{n+1} = L_q(\triangle t) w^n = w_j^n + \triangle t \Phi(w_j^n). \tag{10}$$

$L_f(\triangle t)$ is assumed to be exact or to be a consistent and conservative solution operator of the form $v_j^{n+1} = v_j^n + \frac{\triangle t}{\triangle x}\left[F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n\right]$ with numerical flux function $F_{j+\frac{1}{2}}$.

As we want to investigate local errors of the location of the discontinuity, which occur in a time step $[t_n, t_{n+1}]$, we assume the numerical solution u_j^n at time level n to fulfil the following properties corresponding to the exact solution.
The first assumption is that

$$\triangle x \sum_{j=J}^{K} u_j^n = \int_{x_J-\frac{1}{2}\triangle x}^{x_K+\frac{1}{2}\triangle x} u(x,t_n)dx \tag{11}$$

and

$$u_L^n = u_L(t_n), \quad u_R^n = u_R(t_n). \tag{12}$$

hold. Additionally, we still denote the solution of the homogeneous conservation law by v instead of u, with initial data $v(x,t_n) = u(x,t_n)$, $v_L(t_n) = u_L(t_n)$ and $v_R(t_n) \equiv u_R(t_n)$.
If $L_f(\triangle t)$ is not the exact solution operator then the numerical solution may be smeared-out. Therefore

(13) we assume that the numerical solution — as the exact solution — takes the constant values u_L^n resp. u_R^n on both sides of some interval $[x_L^n, x_R^n]$ that contains the discontinuity.

The interval $[x_l^n, x_R^n]$ will be referred to as the *region of smearing*.

3 Scalar Example

In this Section we consider an example that agrees with the assumptions made in Section 2. It is Burgers' equation with a piece-wise linear source term. In order to see what influence the source term may have on the part of the error that is due to smearing when the source is getting stiff, the initial data are chosen such that the local splitting error is equal to zero. For relative large step sizes wrong shock speeds are obtained.
The equation reads

$$u_t + (\frac{1}{2}u^2)_x = -\mu(u - a(u)) \tag{14}$$

where $\mu > 0$ and

$$a(u) := \begin{cases} 1, & u \geq 0.25 \\ 0, & u < 0.25 \end{cases} . \tag{15}$$

The initial data define a Riemann problem

$$u(x,0) = \begin{cases} 1, & x < 0 \\ 0, & x > 0 \end{cases} . \tag{16}$$

With the given initial data the source is equal to zero on both sides of the shock and the entropy condition (5) is fulfilled. In conclusion, the exact solution is

$$u(x,t) = u(x - \frac{1}{2}t, 0). \tag{17}$$

Remark 3.1 *In [7], a linear advection equation with a source term $-\mu u(u-1)(u-.5)$ is investigated. It is stated that wrong shock speeds occur because of the influence of the source on intermediate values of a smeared-out shock profile. The source term considered here has similar properties in that it shifts values of u that are greater than or equal to 0.25 towards 1 and values that are less than 0.25 towards 0. The point of discontinuity of q does not cause trouble because if the ODE $\tilde{u}_t = q(\tilde{u})$ has to be solved then $a(\tilde{u})$ only depends on the initial $\tilde{u}$ — and $\tilde{u}$ that are at one time less than 0.25 will forever stay smaller than 0.25. The analogous statement is valid for $\tilde{u}$ that are greater than or equal to 0.25. This will be shown in Subsection 3.1. Consequently, for given initial data the source will never become discontinuous while integrating the ODE.*

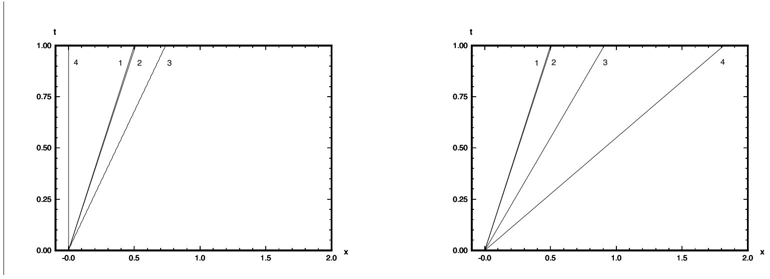
The numerical solution is computed using the Strang splitting. The temporal step size is determined by the CFL-condition. The homogeneous conservation law is solved by using the upwind scheme and the ODE is solved by using the implicit Euler scheme. u_i^{n+1} only depends on the relative time $T_{rel} := \mu \triangle t$ and the initial data for fixed CFL. As the numerical solution does not depend on $\triangle t$ but on T_{rel}, we fix the step sizes in the following computations, and vary only the relative time T_{rel}. We get the following results (see Fig. 1):

- For T_{rel} small the shock speed is correct.
- If we increase T_{rel}, the shock speed becomes wrong.
- If T_{rel} is large enough, two phenomena occur:
  - For ratios of the step sizes $\frac{\triangle t}{\triangle x} < 0.5$ the discontinuity does not move at all.
  - For $\frac{\triangle t}{\triangle x} > 0.5$ the discontinuity moves one grid cell per time step.

3.1 Explanation of the Phenomena

The phenomena shown in the previous subsection for the scalar example occur because of the smearing of the shock profile. To see this, let us look at the single steps of the Strang splitting procedure (see Fig. 2):

Figure 1: Shock curves of the numerical solutions for different T_{rel} where $\triangle t$ and $\triangle x$ are fixed with $\frac{\triangle t}{\triangle x} = 0.45$ resp. $\frac{\triangle t}{\triangle x} = 0.55$, and $\triangle t = 0.01$.



$\frac{\triangle t}{\triangle x} = 0.45$:

$\frac{\triangle t}{\triangle x} = 0.55$:

$1: T_{rel} = 0.00001\,,\ 2: T_{rel} = 0.01\,,\ 3: T_{rel} = 1\,,\ 4: T_{rel} = 100.$

- In the first step $L_q(\frac{1}{2}\triangle t)$ is applied - nothing happens: $\quad q(u_i^n) = 0 \quad \forall i.$
- In the second step the upwind scheme produces one intermediate value. For ratios $\frac{\triangle t}{\triangle x} < 0.5$ this value is less than 0.25 and for $\frac{\triangle t}{\triangle x} > 0.5$ it is greater than 0.25.

  We consider the first case.
- For $\frac{\triangle t}{\triangle x} < 0.5$ the intermediate value is less than 0.25 so that $L_q(\frac{1}{2}\triangle t)$ shifts this value towards zero.

  To understand the influence of the source term, let us consider the ODE

  $$\tilde{u}_t(t) = -\mu(\tilde{u}(t) - a(\tilde{u}(t))), \quad \tilde{u}(0) = u_0. \tag{18}$$

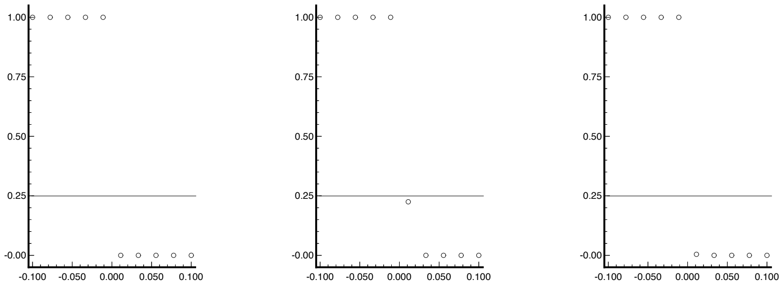
  $\tilde{u}(t)$ is attracted by $a(u_0)$ and $a(\tilde{u}(t)) = a(u_0)$. $\tilde{u}(t)$ reads

  $$\tilde{u}(t) = \begin{cases} u_0 e^{-\mu t} & , u_0 < 0.25 \\ 1 - (1 - u_0)e^{-\mu t} & , u_0 \geq 0.25 \end{cases} . \tag{19}$$

- In the next time step $L_q(\frac{1}{2}\triangle t)$ is applied first. The intermediate value is again shifted towards zero.

If T_{rel} is large enough, the intermediate value is shifted so close to zero that after one time step the discontinuity has not moved at all. The case $\frac{\triangle t}{\triangle x} > 0.5$ behaves analogously.

Figure 2: The three steps of the Strang splitting (over one time step) for $T_{rel} = 100$ and $\frac{\triangle t}{\triangle x} = 0.45$.



4 Estimator for the Error of the Shock Speed

4.1 Smeared-out Shock Profiles and the Equal Area Rule

Firstly, we show that numerically a shock profile has in general to be represented by a smeared solution. Let us therefore consider a homogeneous conservation law. We assume that the mesh points x_j are chosen in such a way that at time t_n the discontinuity lies on a cell boundary. Now we define

$$v_j^n = \frac{1}{\triangle x} \int_{x_j - \frac{1}{2}\triangle x}^{x_j + \frac{1}{2}\triangle x} v(x, t_n) dx. \tag{20}$$

From consistency and conservativity we have

$$\triangle x \sum_{j=J}^{K} v_j^{n+1} = \int_{x_J - \frac{1}{2}\triangle x}^{x_K + \frac{1}{2}\triangle x} v(x, t_{n+1}) dx. \tag{21}$$

Now we can say that, in general, $\sigma(t_{n+1})$ will not coincide with a cell boundary. To see this, one can argue as follows. Assume $\sigma(t_{n+1})$ would coincide with a cell boundary. Then the numerical shock speed would always be an integer multiple of $\frac{\triangle x}{\triangle t}$ even if $\triangle x \rightarrow 0$. Since we usually fix the ratio $\frac{\triangle x}{\triangle t}$, we will in general not have convergence to the correct shock speed.
So if σ^{n+1} does not coincide with a cell boundary there must be at least one value $v_{j_0}^{n+1}$ not equal to v_L^{n+1} or v_R^{n+1} because of (21).

As we have seen, in numerical solutions shocks may be smeared-out. If we want to measure errors of the shock location, we have to define for any numerical solution where this shock location has to be. This will be achieved by using the equal area rule. But before we go into details of the equal area rule, let us introduce some notations and properties concerning the solution of our Riemann problem and the smeared-out discontinuity.

We consider Riemann problems with an entropy solution of the form (3). If the discontinuity is smeared-out in the numerical solution then there are values of u_j^n which are not equal to u_L^n or u_R^n near the discontinuity. For each time level n, we define indices $kl(n)$, $kr(n)$ such that

$$\begin{aligned} &\exists kl(n) \in I : \, x_{kl(n)} < \sigma^n \, \wedge \, u_j^n = u_L^n, \, j \leq kl(n), \, j \in \mathbb{Z} \\ &\exists kr(n) \in I : \, x_{kr(n)} > \sigma^n \, \wedge \, u_j^n = u_R^n, \, j \geq kr(n), \, j \in \mathbb{Z}. \end{aligned} \tag{22}$$

As we do not stress on the boundary treatment, we just consider a sub-domain I of the computational domain with

$$I = \{J, J+1, \ldots, K\} \subseteq \mathbb{Z}.$$

Of course, the computational domain has to be so much larger than I that all values of u_j^n to be inserted into the flux functions $F_{J-\frac{1}{2}}^n$ and $F_{K+\frac{1}{2}}^n$ are known. We demand that the boundaries of the considered domain $I = \{J, J+1, \ldots, K\}$ are away far enough from the discontinuity so that

$$kl(n+1), \, kr(n+1) \in I. \tag{23}$$

Remark that $kl(n+1)$ resp. $kr(n+1)$ differ at least the width of the stencil of $L_f(\triangle t)$ from $kl(n)$ resp. $kr(n)$.

Let us now come back to the equal area rule. With (3), (11), (12), (13), and

$$x_L^n := x_{kl(n)} - \frac{1}{2}\triangle x, \quad x_R^n := x_{kr(n)} + \frac{1}{2}\triangle x$$

we have

$$\begin{aligned} \triangle x \sum_{j=kl(n)}^{kr(n)} u_j^n &= \int_{x_L^n}^{x_R^n} u(x, t_n) \\ &= (\sigma(t_n) - x_L^n) u_L(t_n) + (x_R^n - \sigma(t_n)) u_R(t_n) \\ &= (\sigma(t_n) - x_L^n) u_L^n + (x_R^n - \sigma(t_n)) u_R^n. \end{aligned} \tag{24}$$

This means, that with (24)

$$\triangle x \sum_{j=kl(n)}^{kr(n)} u_j^n = (\sigma^n - x_L^n) u_L^n + (x_R^n - \sigma^n) u_R^n \tag{25}$$

gives σ^n exactly. Notice that only σ^n is unknown so that (25) can be solved for the numerical shock location σ^n. Equation (25) corresponds to the equal area

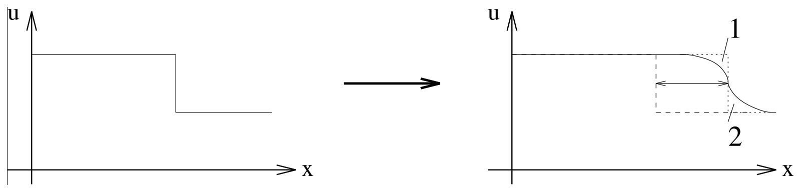


Figure 3: Correct shock position obtained by equal area rule

rule. We define the shock location for any numerical solution by (25). The equal area rule says that the difference area 1, cf. Fig.3, between the smeared-out and the sharp solution profile has to be equal to difference area 2.
If the homogeneous conservation law is solved then, as in (24), it follows with (21) that the equal area rule gives the shock location also at time t_{n+1} exactly:

Lemma 4.1 *Assume that the homogeneous conservation law (6) should be solved for one time step with initial data given by (2) and the entropy condition (5) fulfilled. Let v_j^n be a discrete approximation to these initial data satisfying (11) and $v_L^n = u_L(t_n)$, $v_R^n = u_R(t_n)$. If a conservative scheme with a consistent flux is used to solve (6) numerically then*

- $v_L^{n+1} = v_L^n$ *and* $v_R^{n+1} = v_R^n$,
- *the discontinuity is propagated exactly due to conservation if its location is determined by the equal area rule (25)*
- *thus the correct shock position is obtained.*

For inhomogneous conservation laws the equal area rule means that the production of the source over some temporal and spatial domain should be the same for the numerical smeared-out and the sharp solution. To make this more clear, assume now that the inhomogeneous conservation law $u_t + f(u)_x = q(u)$ should be solved for one time step. Let us therefore consider the integral form of the inhomogeneous conservation law

$$\begin{aligned} \int_{x_J-\frac{1}{2}\triangle x}^{x_K+\frac{1}{2}\triangle x} u(x,t_{n+1})dx &= \int_{x_J-\frac{1}{2}\triangle x}^{x_K+\frac{1}{2}\triangle x} u(x,t_n)dx \\ -\left(\int_{t_n}^{t_{n+1}} f(u(x_K+\frac{1}{2}\triangle x,t))dt \right. &- \left. \int_{t_n}^{t_{n+1}} f(u(x_J-\frac{1}{2}\triangle x,t))dt\right) \\ &+ \int_{t_n}^{t_{n+1}} \int_{x_J-\frac{1}{2}\triangle x}^{x_K+\frac{1}{2}\triangle x} q(u(x,t))dx\,dt. \end{aligned} \tag{26}$$

Here the integration is carried out over one time interval and over the considered spatial computational domain.
From the physical point of view one should demand that the integrals of the fluxes as well as the integral of the source have to be modelled accurately. For the homogeneous conservation law the integrals of the fluxes at the boundary of the considered computational domain I are modelled even exactly by a consistent and conservative numerical scheme. Modelling the integral of the source accurately means that the production of the source over the considered time and spatial domain should be about the same in the numerical and in the exact solution.
The more accurate those integrals are modelled the smaller is the difference

$$\int_{x_J-\frac{1}{2}\triangle x}^{x_K+\frac{1}{2}\triangle x} u(x,t_{n+1}) - \triangle x \sum_{j=J}^{K} u_j^{n+1}. \tag{27}$$

If the numerical shock location is defined by (25) then the error of the numerical shock location depends on (27) and is zero for (27) equal to zero.

4.2 Local Error of the Shock Location

In this section the local error of the shock location is analysed for a scalar Riemann problem. One can interpret the local error of the shock location as coming from two parts: one part introduced by the "splitting" and another occurring because of "smeared-out shock profiles". The numerical solution is computed using the *Strang splitting*

$$u^{n+1} = L_q(\frac{1}{2}\triangle t)L_f(\triangle t)L_q(\frac{1}{2}\triangle t)u^n. \tag{28}$$

Each time one of the solution operators is applied, a change of the shock location is introduced. The following graph illustrates our notations of the intermediate solutions and the parts of the local errors over one time step of the Strang splitting:

$$\begin{tikzcd} u^n \arrow[r, "L_q(\frac{1}{2}\triangle t)"] & u^{n+\frac{1}{2}} \arrow[r, "L_f(\triangle t)"] & \bar{u}^{n+\frac{1}{2}} \arrow[r, "L_q(\frac{1}{2}\triangle t)"] & u^{n+1} \end{tikzcd}$$

$$\begin{array}{ccc} \vdots & \vdots & \vdots \\ \triangle\sigma_1 & \triangle\sigma_f & \triangle\sigma_2 \end{array}$$

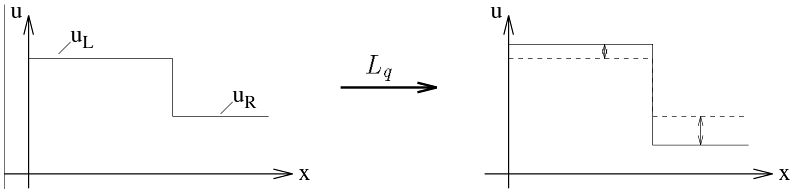
The numerical shock location changes over one time step by

$$\triangle\sigma_{num} := \triangle\sigma_1 + \triangle\sigma_f + \triangle\sigma_2$$

and the change of the shock location of the exact solution is denoted by $\triangle\sigma_{exact}$. So the formula for the local error of the shock location calculates as

$$\mathcal{E} := \triangle\sigma_{num} - \triangle\sigma_{exact} = \triangle\sigma_1 + \triangle\sigma_2 + \triangle\sigma_f - \triangle\sigma_{exact} \tag{29}$$

If there would be no smearing then $\triangle\sigma_1 = \triangle\sigma_2 = 0$. This can be seen as follows. If we apply $L_q(\frac{1}{2}\triangle t)$ to a solution with no smearing then $L_q(\frac{1}{2}\triangle t)$ only changes the quantities $u_L(t)$ and $u_R(t)$. This means that $L_q(\frac{1}{2}\triangle t)$ does not cause a change of the shock location.

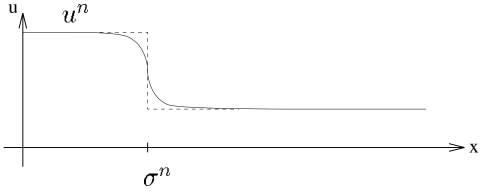


We call

$$(30) \quad \begin{array}{ll} \triangle\sigma_1 + \triangle\sigma_2 & \text{the parts of the error due to smearing} \\ \triangle\sigma_f - \triangle\sigma_{exact} := \mathcal{E}_{spl} & \text{the error due to splitting} \end{array}$$

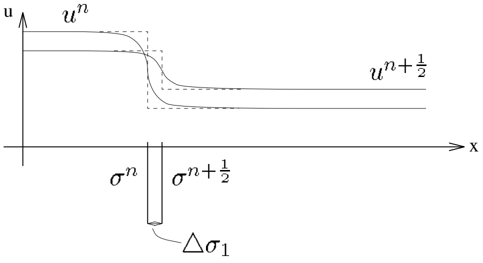
Notice that for exact solution operators $L_f(\triangle t)$ and $L_q(\frac{1}{2}\triangle t)$ the shock profile is not smeared-out. Therefore we also call the error due to splitting $\mathcal{E}_{spl}$ given in (30) the *local splitting error*.
To make it more clear how the changes of the shock location develop if one of the solution operators is applied, we look again at one time step of the Strang splitting procedure.

- We start with the numerical solution u^n which is assumed to have a smeared-out shock profile:



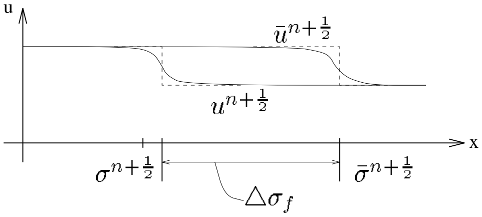
The shock location is determined by the equal area rule.

- Then $L_q(\frac{1}{2}\triangle t)$ is applied. We get $u^{n+\frac{1}{2}}$:



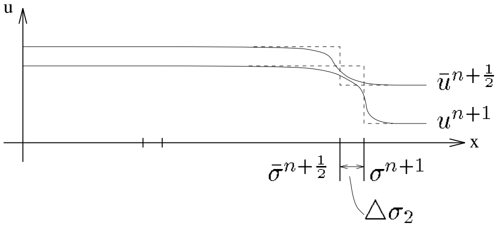
L_q works on the values on the left and on the right sides of the shock *but it also works on the intermediate values.* This causes a change of the shock profile and therefore, in general, a change of the shock location: $\triangle\sigma_1$. Consequently, $\triangle\sigma_1$ arises because of smearing.

- Now $L_f(\triangle t)$ is applied and we get $\bar{u}^{n+\frac{1}{2}}$:



$L_f(\triangle t)$ changes the shock location due to conservation — according to the values $u_L^{n+\frac{1}{2}}$, $u_R^{n+\frac{1}{2}}$: The change of the shock location is denoted by $\triangle\sigma_f$.

- Again $L_q(\frac{1}{2}\triangle t)$ is applied. We get u^{n+1}:



The change of the shock location is denoted by $\triangle\sigma_2$.

4.3 Riemann Problem with Source Term

We consider problem (1), (2) with entropy condition (5) satisfied. We show that the exact solution of this problem can be given explicitly. We need this to compare it to the numerical approximation. We solve equation (1) by integrating along characteristics. One can easily see that all the characteristics starting at time t_n on the left side of the discontinuity σ are curves with the same shape that can be matched on each other by translation in x-direction, and that the same is true for all the characteristics starting on the right side of the discontinuity. Characteristic lines are given by $(x(x_0,t),t)$ and $x(x_0,t)$ has to satisfy

$$(31) \qquad \frac{dx(x_0,t)}{dt} = f'(u(x,t)), \quad x(x_0,t_n) = x_0.$$

Along such a curve $(x(x_0,t),t)$ u fulfils

$$\begin{aligned} \frac{d}{dt}u(x(x_0,t),t) &= u_x(x(x_0,t),t)\frac{dx}{dt} + u_t(x(x_0,t),t) \\ &= q(u(x(x_0,t),t)). \end{aligned} \tag{32}$$

Lemma 4.2 *Let $x(x_0,t)$, $u(x(x_0,t),t)$ with $t \in [t_n, t_{n+1}]$, $x_0 < \sigma$ be a solution of (31), (32) with $x(x_0,t_n) = x_0$. Then for $x_1 < \sigma$*

$$x(x_1,t) = x(x_0,t) + (x_1 - x_0) \tag{33}$$

$$u(x(x_1,t),t) = u(x(x_0,t),t) \tag{34}$$

is also a solution of (31), (32) with

$$x(x_1,t_n) = x_1 \tag{35}$$

The same is valid for x_0, $x_1 > \sigma$.

Proof. (35) is trivially satisfied due to (33), and by (34) we have (32). From (33) it follows that

$$\frac{dx(x_1,t)}{dt} = \frac{dx(x_0,t)}{dt}$$

and hence using (34) again we have (31). □

What we see from Lemma 4.2 is that there exist two states that only depend on the time t

$$u_L(t) := u(x(x_0,t),t) \quad \text{for } x_0 < \sigma \tag{36}$$

and

$$u_R(t) := u(x(x_0,t),t) \quad \text{for } x_0 > \sigma. \tag{37}$$

The two states and the characteristics are obtained from simultaneously solving (31) and (32). But of course we can solve these equations in a much easier fashion.
$u_L(t)$ can be obtained by simply solving

$$\frac{du_L(t)}{dt} = q(u_L(t)) \tag{38}$$

with $u_L(t_n) = u_L$. To compute the characteristics, one has to solve the problem

$$\frac{dx(x_0,t)}{dt} = f'(u_L(t)), \quad x(x_0,t_n) = x_0. \tag{39}$$

One proceeds in the same way for $u_R(t)$.
Now we are ready to state Theorem 4.3.

Theorem 4.3 *Let u_L, u_R and σ be the quantities given in (2) and let $t \in [t_n, t_{n+1}]$. $u_L(t)$ and $u_R(t)$ are supposed to be two solutions of (38) with $u_L(t_n) = u_L$ and $u_R(t_n) = u_R$. Also, let $\sigma(t)$ be the solution of (4) with $\sigma(t_n) = \sigma$. Furthermore assume that the entropy condition (5) is fulfilled.*
Then $u(x,t)$ given by (3) is the weak entropy solution of (1), (2).

Notice that the whole solution is known. Since the right hand side of (4) is known, we can just integrate (4) and demand $\sigma(t_n) = \sigma$ to obtain $\sigma(t)$. This means that we can compute the whole solution exactly even with a source term. The following two lemmata contain continuity statements for $u_L(t)$, $u_R(t)$ and $\sigma(t)$.

Lemma 4.4 *Let $q(w)$ be a (n-1)-times continuously differentiable function of w, but at least continuous, with w, $q(w) \in \mathbb{R}$. Suppose $w(t)$ to be a solution of (38), i.e. $w(t)$ satisfies*

$$\frac{dw(t)}{dt} = q(w(t)) \tag{40}$$

Then $w(t)$ is n-times continuously differentiable.

Lemma 4.5 *Let $f(u)$ be n-times and let $q(u)$ be (n-1)-times continuously differentiable, with $n \geq 1$ resp. q at least continuous, with u, $f(u)$, $q(u) \in \mathbb{R}$. Furthermore let $t \in [t_n, t_{n+1}]$ and suppose $u_L(t)$ and $u_R(t)$ to be two solutions of (38) with $u_L(t) \neq u_R(t)$.*
Then $\dot{\sigma}(t)$ defined by (4) is n-times continuously differentiable.

The proofs of Lemmata 4.4 and 4.5 follow by differentiating.

4.4 Local Splitting Error

In this subsection we analyse the local splitting error $\mathcal{E}_{spl} = \triangle\sigma_f - \triangle\sigma_{exact}$. With the abbreviation

$$h(v,w) := \frac{f(v) - f(w)}{v - w} \tag{41}$$

we can write the jump condition (4) as

$$h(u_L(t), u_R(t)) = \dot{\sigma}(t). \tag{42}$$

We also use the abbreviation

$$h_{10}(v,w) := \frac{\partial h(v,w)}{\partial v}, \quad h_{01}(v,w) := \frac{\partial h(v,w)}{\partial w}. \tag{43}$$

The second derivatives are defined analogously, e.g. $h_{20}(v,w)$ means that the function $h(v,w)$ is differentiated twice with respect to the first argument.

Additionally, let $\Phi_L^n := \Phi(u_L^n)$ for the incremental function Φ defined in (10), and let Φ_R^n be defined analogously.
To compute the local splitting error we have to investigate $\triangle\sigma_f$ and $\triangle\sigma_{exact}$. Let us consider $\triangle\sigma_f$ first. $\triangle\sigma_f$ solves the homogeneous equation (6). As the shock speed is constant for the homogeneous conservation law and by Lemma 4.1 it follows that

$$\triangle\sigma_f = \triangle t\, h(u_L^{n+\frac{1}{2}}, u_R^{n+\frac{1}{2}}). \tag{44}$$

On the other hand, the change of the shock location of the exact solution can be written as

$$\triangle\sigma_{exact} = \int_{t_n}^{t_{n+1}} \dot{\sigma}(t)dt = \int_{t_n}^{t_{n+1}} h(u_L(t), u_R(t))dt. \tag{45}$$

Expansion and comparison give the following results.

Theorem 4.6 *Let the conditions of Theorem 4.3 be fulfilled. Furthermore assume f to be three times and q to be at least twice continuously differentiable. Request that $L_f(\triangle t)$ is a consistent and conservative solution operator and let $L_q(\triangle t)$ be either the exact or an approximate solution operator with a local discretisation error of the order $p+1$ with $p \geq 1$. In the last case q has to be p-times continuously differentiable. We assume the numerical solution at time t_n to fulfil (11) and $u_L^n = u_L(t_n)$, $u_R^n = u_R(t_n)$.*
Then the local splitting error $\mathcal{E}_{spl}$ is

$$\begin{aligned} \mathcal{E}_{spl} &= (\frac{1}{2} - \theta)\triangle t^2 \left[h_{10}(u_L^n, u_R^n)\, q(u_L^n) + h_{01}(u_L^n, u_R^n)\, q(u_R^n)\right] \\ &\quad + O(C_{h,2}\triangle t^3) + K_q O(C_{u,p+1} C_{h,u} \triangle t^{p+2}) \end{aligned} \tag{46}$$

$$\begin{aligned} \mathcal{E}_{spl} &= \triangle t^2 \Bigg[h_{10}(u_L^n, u_R^n)\, (\frac{1}{2}\Phi_L^n - \theta q(u_L^n)) \\ &\quad + h_{01}(u_L^n, u_R^n)\, (\frac{1}{2}\Phi_R^n - \theta q(u_R^n))\Bigg] + O(C_{h,2}\triangle t^3) \end{aligned} \tag{47}$$

$$\begin{aligned} \mathcal{E}_{spl} &= -\frac{1}{24}\triangle t^3 \,\frac{d^2 h(u_L^{n+\frac{1}{2}}, u_R^{n+\frac{1}{2}})}{dt^2} \\ &\quad + O(C_{h,3}\triangle t^4) + K_q O(C_{u,p+1} C_{h,u} \triangle t^{p+2}) \end{aligned} \tag{48}$$

where $\theta \in (0,1)$. The constant K_q is given by

$$K_q = \begin{cases} 0 \ , \ L_q \ exact \\ 1 \ , \ else \end{cases} .$$

$\dfrac{d^2 h(u_L^{n+\frac{1}{2}}, u_R^{n+\frac{1}{2}})}{dt^2}$ *means the following: The function $h(u_L(t), u_R(t))$ is differentiated twice with respect to t, where the derivatives of u_L, u_R with respect to t are* *written in terms of* $q(u_L)$, $q(u_R)$. *This gives an expression of* $\dfrac{d^2h(u_L(t),u_R(t))}{dt^2}$ *in terms of* u_L *and* u_R. *Then insert the values* $u_L^{n+\frac{1}{2}}$ *resp.* $u_R^{n+\frac{1}{2}}$.

The constants $C_{u,k}$ *depend on the* k^{th} *derivative of* u *with respect to* t, C_{h,u^M} *depend on mixed derivatives of order* M *of* h *with respect to the first and second argument, and* $C_{h,k}$ *depend on the* k^{th} *derivative of* h *with respect to* t *- in the sense as it is described for* $\frac{d^2h}{dt^2}$.

The error-estimates are the most interesting if the source term is stiff. This means that the time scale of the source term is much faster than the time scale of the fluid dynamics. This again corresponds to the fact that the absolute eigenvalues of the (linearized) source are much larger than the absolute eigenvalues of the (linearized) fluid dynamics. Therefore we may request that the source fulfils

$$q^{(m)} =: \mu\tilde{q}^{(m)} \text{ with } \tilde{q}^{(m)} = O(1),\ \mu \in [0,\infty),\ m = 0,1,2,\ldots \tag{49}$$

$q^{(m)}$ denotes the m^{th} derivative of q with respect to u. Using this assumption we treat all derivatives of $u_L(t)$ and $u_R(t)$ as being of order $O(\mu)$ whereas all the other terms are treated as being of order $O(1)$. Then the constants in Theorem 4.6 can be estimated as

$$\begin{aligned} C_{u,k} &= O(\mu^k),\ k = 1,2,3,p+1 \\ C_{h,u^M} &= O(1) \\ C_{h,3} &= O(\mu^3) \end{aligned}$$

These estimates follow by applying the chain rule for differentiating and, if necessary, induction. Here we used $C_{u,1} = O(\mu)(1 + O(\mu^p \triangle t^p) = O(\mu)$, $p > 0$.

4.5 Influence of Smearing

In this subsection we want to analyse the parts $\triangle\sigma_1$ and $\triangle\sigma_2$ of the local error of the shock location which are due to smearing. The analysis is done for $\triangle\sigma_1$ and can be carried out analogously for $\triangle\sigma_2$.
Remember that the interval $[x_L^n, x_R^n]$ was defined to be the region of smearing and $kl(n)$, $kr(n)$ are given by (22). With $u_j^{n+\frac{1}{2}} = L_q(\frac{1}{2}\triangle t)u_j^n$ it follows that $[x_L^n, x_R^n]$ can also be taken as region of smearing for $u^{n+\frac{1}{2}}$.
Applying the equal area rule (25) to $u_j^{n+\frac{1}{2}}$ and u_j^n, building the difference of these two equations , and using

$$\sigma^{n+\frac{1}{2}} = \sigma^n + \triangle\sigma_1 \tag{50}$$

gives

$$\triangle\sigma_1 = \Big[\triangle x \sum_{j=kl(n)}^{kr(n)} (u_j^{n+\frac{1}{2}} - u_j^n) - (\sigma^n - x_L^n)(u_L^{n+\frac{1}{2}} - u_L^n)$$

$$-(x_R^n - \sigma^n)(u_R^{n+\frac{1}{2}} - u_R^n)\Big] : \Big[u_L^{n+\frac{1}{2}} - u_R^{n+\frac{1}{2}}\Big]. \tag{51}$$

To compute σ^n, the values u_j^n have to be inserted into the equal area rule (25) explicitly. Consequently, we want to write (51) only in terms of u_j^n so that the values $u_j^{n+\frac{1}{2}}$ do not need to be inserted to compute $\triangle\sigma_1$. To do this we denote $u_j(t)$ to be the exact solution of the ODE

$$w_t = q(w), \quad t \in [t_n, t_n + \frac{1}{2}\triangle t] \quad \text{with } w(t_n) = u_j^n \tag{52}$$

so that by means of Taylor series expansion we get

$$u_j(t_n + \frac{1}{2}\triangle t) - u_j^n = \delta(u_j^n) + O(\triangle t^3 \mathcal{C}_{u,3}) \tag{53}$$

with

$$\delta(u) := \frac{\triangle t}{2} q(u) + \frac{\triangle t^2}{8} q'(u) q(u). \tag{54}$$

$\mathcal{C}_{u,3}$ depends on the second derivatives of q with respect to t.
Using the incremental step function Φ, we have

$$u_j^{n+\frac{1}{2}} - u_j^n = \frac{\triangle t}{2}\Phi(u_j^n). \tag{55}$$

We use the abbreviation $\Phi_L^n := \Phi(u_L^n)$, and Φ_R^n defined analogously. Then expansions give the following results by using (53), (55) in (51).

Theorem 4.7 *Let the conditions of Theorem 4.6 be fulfilled. The numerical solution may be smeared-out. Then, with σ^n defined by (25),*

$$\triangle\sigma_1 = \Big[\triangle x \sum_{kl(n)}^{kr(n)} \delta(u_j^n) - (\sigma^n - x_L^n)\delta(u_L^n) - (x_R^n - \sigma^n)\delta(u_R^n)\Big]$$

$$: \Big[u_L^n - u_R^n + \delta(u_L^n) - \delta(u_R^n)\Big]$$

$$+ O((x_R^n - x_L^n)\mathcal{C}_{u,3}\triangle t^3) \; + \; K_q O((x_R^n - x_L^n)\mathcal{C}_{u,p+1}\triangle t^{p+1}) \tag{56}$$

$$\triangle\sigma_1 = \frac{\triangle t}{2}\Big[\triangle x \sum_{kl(n)}^{kr(n)} \Phi(u_j^n) - \Phi_L^n(\sigma^n - x_L^n) - (x_R^n - \sigma^n)\Phi_R^n\Big]$$

$$: \Big[u_L^n - u_R^n + \frac{\triangle t}{2}(\Phi_L^n - \Phi_R^n)\Big] \tag{57}$$

where $\mathcal{C}_{u,p+1}$ is defined analogously to $\mathcal{C}_{u,3}$ and K_q is the same as in Theorem 4.6.
If we replace u_j^n by $\bar{u}_j^{n+\frac{1}{2}}$ we get $\triangle\sigma_2$.

In order to make clear the effect of stiff source terms on the errors of the shock location, condition (49) was introduced in Subsection 4.4. If we assume (49) here we can argue in a similar manner to get

$$\mathcal{C}_{u,k} = O(\mu^k),\ k = 3, p+1.$$

4.6 Estimators

In general, the structure of the solution is not as simple as in the analysis of the scalar Riemann problem. In the case where a solution has a shock that is not interacting with another shock in the considered time interval, we have to define a region of smearing of the discontinuity in order to approximate the local error of the shock location. This region of smearing at time t_n is denoted by $[x_{kl(n)} - \frac{1}{2}\triangle x, x_{kr(n)} + \frac{1}{2}\triangle x]$ where the indices $kl(n)$, $kr(n)$ are in the general case no longer given by (22), but have to be defined suitably for each specific problem. Next, use this region of smearing to approximate the values u_l^n, u_r^n on the left and on the right sides of the discontinuity. One way to proceed is to approximate the values u_l^n, u_r^n by the values of the solution at the boundary of the region of smearing: $u_l^n := u_{kl(n)}^n$ and $u_r^n := u_{kr(n)}^n$. Now those values can be inserted into the error-formula which gives us approximated local errors of the shock location.
All the numerical experiments shown in this thesis, except of the scalar examples where the solution is piece-wise constant, use this approach.
All estimators neglect the higher order terms in the error-formulae.

Application to Systems

To apply the estimators for the scalar equation to a system of hyperbolic conservation laws with source terms, define each of the parts of the local error of the location of a discontinuity for each equation of the system separately. Consequently, the local error of the location of the discontinuity can be computed for each of the conservative (we call them so although there is a source present) variables. So the results of the scalar case can be applied to each single equation of the system by taking into account that one works with vectors now.
We consider numerical approximations of exact solutions having a discontinuity between two states connected by a p-shock that satisfies the Rankine-Hugoniot jump conditions and the entropy condition. This discontinuity should not interact with another one in the considered time interval. One way to approximate the local error of the location of the discontinuity in such a case is analogous to the scalar case: Define a region of smearing $[x_{kl} - \frac{1}{2}\triangle x, x_{kr} + \frac{1}{2}\triangle x]$ through

one of the conservative variables (depending on the problem) and use this to approximate the vectors U_L, U_R on the left and on the right side of the discontinuity. For each of the equations of the system insert those values into the equation for the local errors of a scalar equation.

To compute the local splitting error, the Rankine-Hugoniot jump conditions are used. Due to numerical inaccuracies and depending on how the values U_L, U_R are chosen in the numerical solution, the jump conditions may not be fulfilled exactly. We mention one possible way to proceed.

Approximate U_L, U_R by $U_L := U_{kl}$ and $U_R := U_{kr}$ even if those values do not fulfil the jump conditions exactly. As a consequence of the analysis being done for a piece-wise constant scalar solution, the local error of the location of the discontinuity is approximated in a rather rough way. Therefore it need not to be required that values U_L, U_R fulfilling the jump conditions exactly are used — they are an approximation, anyway.

In the numerical experiments of this paper only this approach is used.

4.7 Adaptation

In Subsections 4.4 and 4.5 we computed local errors of the location of the discontinuity. We want to use these error-estimates to construct an adaptation of the step size in order to keep the relative global errors sufficiently small. Therefore we have to find an upper bound $\mathcal{B}$ for the local error to indicate if the mesh has to be refined or not.

Let us first investigate the influence of $\mathcal{B}$ on the relative global error. We denote the local error of the location of the discontinuity of the n^{th} time step for the moment by $\mathcal{E}(n)$. Our aim is to ensure that after N time steps the relative global error E_{rel} is smaller than a certain positive number. So if we demand the absolute values of the local errors $|\mathcal{E}(n)|$ to be smaller than a certain upper bound $\mathcal{B}$

$$|\mathcal{E}(n)| < \mathcal{B},$$

it follows that

$$\Rightarrow \sum_{n=1}^{N} |\mathcal{E}(n)| < N\,\mathcal{B}.$$

We define an averaged speed $\bar{\bar{\sigma}}$ by $|\sigma^N - \sigma^0| =: \bar{\bar{\sigma}}\, N \,\triangle t_{max}$, where $\triangle t_{max}$ is the largest step size of the N time steps. Then the relative global error is bounded from above by

$$E_{rel} = \frac{|\sum\limits_{n=1}^{N} \mathcal{E}(n)|}{|\sigma^N - \sigma^0|} \leq \frac{\sum\limits_{n=1}^{N} |\mathcal{E}(n)|}{|\sigma^N - \sigma^0|} < \frac{\mathcal{B}}{\bar{\bar{\sigma}}\,\triangle t_{max}}\,. \tag{58}$$

If

$$\mathcal{B} \leq \bar{\bar{\sigma}}\,\triangle t \cdot tol \leq \bar{\bar{\sigma}}\,\triangle t_{max} \cdot tol$$

with a certain positive constant tol, then

$$E_{rel} < tol\,.$$

Notice that in general, $\triangle t$ may change from time step to time step.

Remark 4.8 *Of course, one would like to have $\mathcal{B}$ automatically chosen during the computations. But how to automate the adaptation will not be a topic of this paper, we will just show that the adaptation works well for appropriate bounds $\mathcal{B}$.*

The adaptation works as follows: The estimated local error of the shock location $\mathcal{E}_{est}$ is expected to be smaller than the upper bound $\mathcal{B}$. If this assumption is not satisfied, the step size $\triangle x$ is bisected, $\triangle t$ is computed in dependence of $\triangle x$, and then the last step is repeated. Which exact value for $\mathcal{B}$ is going to be used in an adaptation, is decided here on test computations.

5 Examples

5.1 Scalar Example

In this subsection the local errors of the shock location are computed for the scalar example presented in Section 3. The exact errors are compared to those approximated with (46) and (56) by neglecting the higher order terms. Furthermore, the adaptation described in Subsection 4.7 is tested.
In this example, given by (14), (15), (16), the local error of the shock location (29) computes as

$$(59) \qquad \mathcal{E} = \triangle\sigma_{num} - \triangle\sigma_{exact} = \triangle x \sum_i (u_i^{n+1} - u_i^n) - \frac{1}{2}\triangle t.$$

The local errors of the shock location that are computed by using the error-formulae (46) and (56), reduce in the considered problem to

$$(60) \qquad \mathcal{E} = \triangle\sigma_1 + \triangle\sigma_2$$

because with the special choice of the source term we have $q(u_L^n) = q(u_R^n) = 0$ for any n and therefore $\mathcal{E}_{spl} = 0$.
The relative (global) error after N time steps is defined by

$$(61) \qquad E_{rel} := \frac{|\sigma^N - \sigma(t_N)|}{|\sigma(t_N)|}.$$

Table 1 lists the relative (global) errors E_{rel} and the relative differences between the estimated and the exact local errors in dependence of the relative time T_{rel}.

Table 1: Relative time T_{rel}, relative (global) error E_{rel}, relative maxima of the (absolute) differences between the estimated and the exact local errors. The relative time T_{rel} is varied while the step sizes $\triangle t$ and $\triangle x$ are fixed with $CFL = .55$. The solutions are computed up to time $T = t_N = 1$.

T_{rel}	E_{rel}	Difference
.1-2	.1594-2	.4168-7
.2-2	.3178-2	.1667-6
.3-2	.4762-2	.3751-6
.4-2	.6330-2	.6670-6
.5-2	.8068-2	.1042-5
.6-2	.9297-2	.1501-5
.7-2	.1102-1	.2043-5
.8-2	.1255-1	.2669-5
.9-2	.1385-1	.3379-5
1.0-2	.1571-1	.4172-5

The notation $.1 + 1 := .1 \cdot 10^1$ *is used.*

The relative differences of the errors are the quotients of the maxima of the (absolute) differences between the estimated and the exact local errors and the maxima of the (absolute) local errors. The maxima are taken over the N computed time steps up to time $T = t_N = 1$.

As expected, are the relative differences between the estimated and the exact local errors sufficiently small. E_{rel} can now be used to decide which value T_{rel} should not exceed.

Adaptation

To make the adaptation work, we have to determine an appropriate bound $\mathcal{B}$ as described in Subsection 4.7. If we choose *tol* $= .01$ then Table 1 shows that $\max(T_{rel})$ should not exceed 0.007. Based on test computations, we define the upper bound for the local error to be

$$\mathcal{B} := 0.5 \cdot \triangle t \cdot c_0 \cdot 0.01\,, \qquad c_0 := \begin{cases} 1, & b \leq 3 \\ b, & 4 \leq b \leq 8 \\ b+1, & 9 \leq b \end{cases}$$

where b denotes the number of bisections. Notice that the factor 0.5 is the shock speed of the exact solution. $c_0 \cdot 0.01$ corresponds to the constant *tol*.

If during the adaptation estimated local error of the shock location $\mathcal{E}_{est}$ is not smaller than the upper bound $\mathcal{B}$, then the step sizes $\triangle t$ and $\triangle x$ are bisected so that the ratio $\frac{\triangle t}{\triangle x}$ stays constant. We computed the solution up to time $T = 1$ for $\mu = 10$. Table 2 lists the numerical results.

Each of the resulting step sizes is not larger than twice the smallest of them, and the relative error remains less than 1.3%. Also, one can see that the resulting T_{rel} is smaller than 0.01 and thus gives sufficiently correct solutions, see Fig. 1. This shows that the adaptation works well for this example.

Table 2: Number of bisections b and time steps N, the resulting smallest $\triangle t$, and the relative error of the shock location for various step sizes $\triangle t^0$ at time $t = 0$. The solutions are computed up to time $T = 1$ for $\mu = 10$ and $CFL = .55$.

$\triangle t^0$	b	N	$\triangle t$	E_{rel}
0.01	4	1601	.625000-3	.100392-1
0.10	8	2554	.396250-3	.676327-2
0.20	8	1278	.761250-3	.129742-1
1.00	11	2020	.488281-3	.945125-2
2.00	12	1901	.488281-3	.128107-1

The following notation is used: $.10 + 1 := .10 \cdot 10^1$

5.2 The Combustion Model

We consider a simplified model for the inviscid reacting compressible Euler equations in one space dimension. This model is a (2×2) - system, given by Burgers' equation coupled to a chemical kinetics equation. It has analogues to the structure of reacting shock profiles (see [2] or [8] for more details). In [2] it is recommended to use these simplified model equations for testing and developing numerical codes that have to handle shock phenomena in reacting gases.
The equations for the simplified combustion model are given by

$$u_t + (\frac{1}{2}u^2 - q_0 Z)_x = 0 \tag{62}$$

$$Z_x = \Phi(u) Z \tag{63}$$

with initial data

$$u(x,0) = \begin{cases} 1.0, & x < 0 \\ -0.7, & x > 0 \end{cases} . \tag{64}$$

We use ignition temperature kinetics with $\Phi(u)$ given by

$$\Phi(u) = \begin{cases} 1, & u \geq 0 \\ 0, & u < 0 \end{cases} .$$

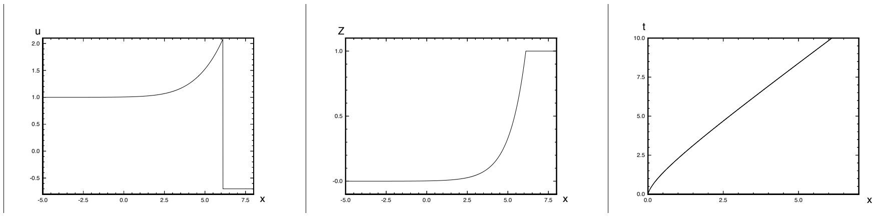
Z is the mass fraction of unburned gas with $\lim_{x \to \infty} Z(x,t) = 1$. The x-coordinate in (63) represents the distance from the reaction zone. u can be interpreted as a lumped variable with some features of pressure or temperature, cf. [2]. We set $q_0 = 0.935$ where q_0 is the heat release.
The data given by (64) initiate a shock wave that raises u above 0 which causes the reaction to start. In the reaction zone that follows the mass fraction of unburnt gas decreases as long as the gas is getting burnt. As time passes a

combustion spike develops and a travelling wave solution evolves. The solution structure shows a reaction rate that is zero ahead of the shock and finite behind it. This structure is analogous to the solution structure of the reacting Euler equations considered in Chapter 5.3.

The detonation wave has the internal structure of an ordinary shock wave followed by a reaction zone. The shock wave has to satisfy the Rankine-Hugoniot jump condition. This jump condition states that the speed of the combustion wave with the left and right states given by (64) is $\dot{\sigma}(0) = 0.15$ at time $t = 0$. The shock wave raises u from $u_+ = -0.7$ to a left state u_{peak}. If the travelling wave solution has evolved with the fixed speed $\dot{\sigma}(t) = 0.7$ then by again using the jump condition $\dot{\sigma}(u_{peak} - u_+) = (\frac{1}{2}u_{peak}^2 - \frac{1}{2}u_+^2)$ we have for the left state $u_{peak} = 2.1$.

Figure 4 shows a reference solution.

Figure 4: Reference solution of the simplified combustion model showing u, Z at time $T = 10$ and the shock curve. The step sizes are $\triangle t = 0.001$ with $\frac{\triangle t}{\triangle x} = .45$.



The numerical solution is computed using the Strang splitting where the equations

$$u_t + (\frac{1}{2}u^2)_x = 0 \tag{65}$$

$$Z_x = \Phi(u)Z \tag{66}$$

$$u_t = q_0\Phi(u)Z \tag{67}$$

are solved separately in each time step. (65) is solved with an upwind scheme, c.f. [12]. (66) is solved by trapezoidal approximation of the integral in the exact solution formula, c.f. [2],

$$Z_i = Z_{i+1}e^{-\frac{\triangle x}{2}(\Phi(u_{j-1})+\Phi(u_j))}, i = J, J+1, \ldots, K; \quad Z_K = 1.$$

Here the superscripts denoting the time level were omitted. To solve (67) with fixed Z we use

$$u_t = \begin{cases} const > 0, & \text{for the initial data } u_0 \geq 0 \\ 0, & \text{else,} \end{cases}$$

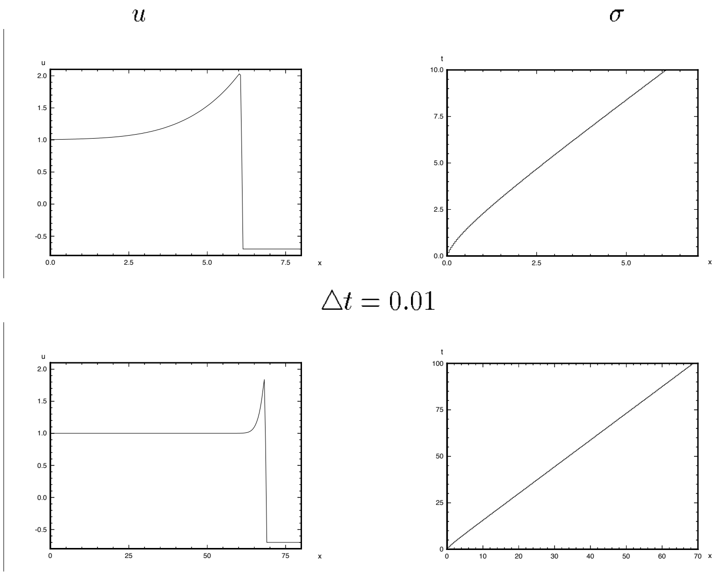
which means that $\Phi(u)$ remains constant. Using this result we solve equation (67) exactly. We solve (66) first and then (67), and combining the solution operators for these two equations in this way gives $L_q(\frac{1}{2}\triangle t)$.
The numerical solution shows the following behaviour of the combustion wave speeds for different step sizes $\triangle t$ and $\triangle x$ with $\frac{\triangle t}{\triangle x} = 0.45$, see Fig. 5:

- For $\triangle t$ small the shock speed is correct.
- If we increase $\triangle t$, the shock speed becomes slower than the correct speed.
- If $\triangle t$ is large enough, the shock speed remains unchanged: $\dot{\sigma}(t) \equiv \dot{\sigma}(0) = 0.15$.

This phenomenon occurs because the combustion spike of the solution decreases when the step size is increased.

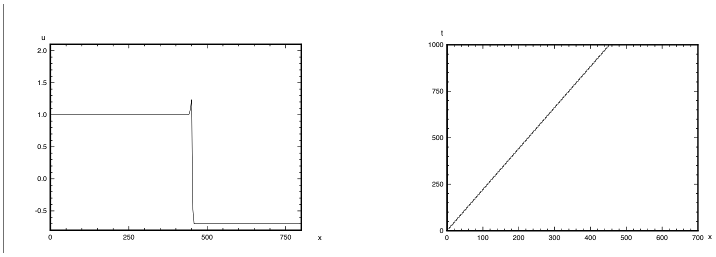
Remark 5.1 *In [2] another kind of spurious solution for the same example is obtained for large step sizes. Other solution operators and somewhat different data are used. These spurious solutions exhibit a bifurcating wave pattern as it also appears for the reacting Euler equations, see Chapter 5.3.*

Figure 5: Numerical solution of the simplified combustion model showing u after 1000 time steps and the shock curve for different $\triangle t$.

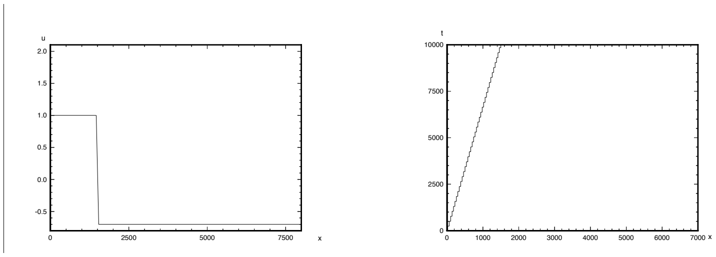


$\triangle t = 0.01$

$\triangle t = 0.1$



$\triangle t = 1$



$\triangle t = 10$

Errors of the Shock Location

The solution profile is not piece-wise constant so that this error can only be approximated. The solution exhibits a shock wave where the right state equals the initially given right state and where the left state corresponds to the combustion peak. The variable Z begins to decrease just after the shock wave. To approximate $\triangle\sigma_1$ and $\triangle\sigma_2$ we have to define $x_{kl(n)}$, $x_{kr(n)}$ for the region of the smearing. This is done by prescribing $kl(n)$ and $kr(n)$ as

$$kl(n) \; : \; u^n_{kl(n)} = \max_{i\in I} u^n_i \qquad \text{and} \qquad kr(n) := \min\{ i \in I \,|\, u^n_i = -0.7 \,\}.$$

We approximate Z by $Z \equiv 1$ over the region of the smearing $[x_{kl(n)}, x_{kr(n)}]$. Finally, we estimate the local error of the shock location by

$$\mathcal{E}_{est} = |\triangle\tilde{\sigma}_1 + \triangle\tilde{\sigma}_2| + \frac{1}{4} q_0 \triangle t^2 \tag{68}$$

where $\triangle\tilde{\sigma}_1$ resp. $\triangle\tilde{\sigma}_2$ denote the approximations to $\triangle\sigma_1$ resp. $\triangle\sigma_2$.

Adaptation

Now we want to test if an adaptation of the step size based on $\mathcal{E}_{est}$ (68) gives satisfactory results. The aim is to keep the relative global errors of the shock location sufficiently small. The adaptation works as described in Subsection 4.7. If $\mathcal{E}_{est}$ is not smaller than a certain upper bound $\mathcal{B}$, the step sizes $\triangle t$ and $\triangle x$ are bisected so that the ratio $\frac{\triangle t}{\triangle x}$ stays constant, and then the last time step is repeated.

Based on test computations, we set the upper bound for the local error to be

$$\mathcal{B} := 0.01 \cdot \triangle t.$$

Furthermore, we estimate the relative (global) error of the shock location by

$$\tilde{E}_{rel} = \frac{\sum_{n=1}^{N} \mathcal{E}_{est}(n)}{|\sigma^N|}$$

where $\mathcal{E}_{est}(n)$ denotes the approximate local error of the n^{th} time step.
The solutions are computed up to time $T = t_N = 10$. Table 3 lists the numerical results.

Table 3: Number of bisections b and time steps N, the resulting smallest $\triangle t$, and $\tilde{E}_{rel}$ for various step sizes $\triangle t^0$ at time $t = 0$. The solutions are computed up to time $T = t_N = 10$ and $\frac{\triangle t}{\triangle x} = .25$.

$\triangle t^0$	b	N	$\triangle t$	$\tilde{E}_{rel}$
0.1	5	3199	.312500-2	.5507-2
1.0	8	2560	.390625-2	.6950-2
2.0	9	2560	.390625-2	.6950-2
5.0	11	4090	.244141-2	.4328-2
10.0	12	4090	.244141-2	.4328-2
20.0	13	4090	.244141-2	.4328-2

The following notation is used: $.10 + 1 := .10 \cdot 10^1$

Each of the resulting step sizes is smaller than twice the smallest of them, and the approximate relative errors remain less than 0.7%. Also, one can see that the resulting step sizes yield sufficiently correct solutions, see Fig. 5.

5.3 The Reacting Euler Equations

In this subsection the adaptation is applied to the one-dimensional inviscid reacting compressible Euler equations. The solution of those equations shows a detonation wave where the chemical reaction is taking place very much faster than the fluid flows. Numerical solutions of those equations exhibit an unphysical bifurcating wave pattern for large step sizes, where a precursor weak detonation wave is moving with a speed of one grid cell per time step, cf. [2].
In our model equations for a reacting mixture the following simplifying assumptions are made, see [1], [11]: There are only two species present, unburnt gas and burnt gas. The unburnt gas is converted to burnt gas by a one-step irreversible chemical reaction. Furthermore, the specific heats at constant pressure are assumed to be equal and constant and the gas mixtures should behave like ideal gases with the same gas constant γ. Finally, effects of diffusion are ignored.

Under these assumptions the model is described by the inviscid reacting compressible Euler equations [2]:

$$U_t + F(U)_x = Q(U)\,, \tag{69}$$

where

$$U = \begin{pmatrix} \rho \\ \rho u \\ \rho E \\ \rho Z \end{pmatrix} \quad F(U) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u E + u p \\ \rho u Z \end{pmatrix} \quad Q(U) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -\rho K(T) Z \end{pmatrix}.$$

Here we have density ρ, fluid velocity u, total energy $E = e + \frac{u^2}{2} + q_0 Z$, mass fraction of unburnt gas Z, constant amount of heat released per unit mass by the chemical reaction q_0, specific internal energy e, pressure $p = (\gamma - 1)\rho e$, temperature $T = \frac{p}{\rho \frac{R}{M}}$, Boltzmann's gas constant R, molecular weight M, specific heat at constant density c_ρ, and gas constant $\gamma = \frac{R}{c_\rho} + 1$.
We use ignition temperature kinetics such that

$$K(T) = \begin{cases} K_0, & T \geq T_0 \\ 0, & T < T_0\,, \end{cases} \tag{70}$$

with T_0 the ignition temperature and K_0 the reaction rate.
Of special interest is a detonation that spontaneously emerges from the process of combustion itself. It belongs in a series of important cases to the Chapman-Jouguet (C-J) point [2][5]. The initial data are piece-wise constant, defining a C-J detonation:

$$\begin{array}{ll} (p_0, \rho_0, u_0, Z_0)\ ,\ x > 0 & \text{pre-shock state} \\ (p_1, \rho_1, u_1, Z_1)\ ,\ x \leq 0 & \text{post-shock state.} \end{array} \tag{71}$$

The pre-shock state corresponds to 25% ozone and 75% oxygen at roughly room temperature. The data are:

variables	pre-shock state	post-shock state
$p\,[\frac{g}{cm\,sec^2}]$	$0.8321 \cdot 10^6$	$7.9434 \cdot 10^6$
$\rho\,[\frac{g}{cm^3}]$	$1.2001 \cdot 10^{-3}$	$1.9690 \cdot 10^{-3}$
$u\,[\frac{cm}{sec}]$	$0.$	$4.8057 \cdot 10^4$
$q_0 Z\,[\frac{cm^2}{sec^2}]$	$6.9283 \cdot 10^9$	$0.$

Furthermore we have $R = 8.3143 \cdot 10^7 \frac{cm^2\,g}{sec^2\,{}^oK\,mol}$, $M = 36\,\frac{g}{mol}$, $\gamma = 1.4$, $T_0 = 500^oK$, and $K_0 = 0.582458 \cdot 10^{10} \frac{1}{s}$.

Numerical Solutions

The numerical solution is computed using the Strang splitting. The homogeneous conservation law is solved using a flux-difference splitting where an approximate Riemann solver introduced by Pandolfi [9] is implemented. This solver transforms the variables $(\rho, \rho u, \rho E_0)$ into the variables speed of sound, fluid velocity, and entropy (a, u, s). Here we use the notation $E_0 = e + \frac{u^2}{2}$. The acoustic waves (1,3) are assumed to be locally isentropic.
The ODE $U_t = Q(U)$ is solved by reducing the problem to the scalar equation

$$(\rho q_0 Z)_t = -K(T)(\rho q_0 Z). \qquad (72)$$

Different ODE-solvers are used in order to compare the local errors of the location of the discontinuity. The ODE-solvers are:

1. Exact integration of the linear ODE, where the temperature is held pointwise fixed.
2. Semi-implicit Euler scheme
3. Explicit Euler scheme
4. Explicit Runge-Kutta method of 2. order

The numerical solutions we show here are computed using ODE-solver 1 (exact integration of the linear ODE). The spatial step size $\triangle x$ is varied

$$\triangle x = \alpha R_0\,, \quad R_0 = 5.347 \cdot 10^{-6}\,, \quad \alpha = 0.001, \ldots, 100000$$

and in each time step the step size in time $\triangle t$ is determined by the CFL-condition with CFL-number 0.8. There is no other stability condition for $\triangle t$.

When time evolves, a travelling wave profile with constant wave speed $\dot{\sigma}$ is being built up. The speed of this wave is sonic relative to the gas flow behind it, which means that

$$\dot{\sigma} = a_1 + u_1 = \sqrt{\gamma \frac{p_1}{\rho_1}} + u_1 = 1.2321 \cdot 10^5.$$

We briefly describe the structure of the physical solution, cf. [5]. The actual shock wave in the combustible starting mixture is the front of a detonation wave. In this wave the gas is compressed and warmed up. The state immediately behind this shock wave corresponds to the peaks one can see in the solution profiles of numerical solutions (here shown for pressure and density, see Fig. 6) that are computed using small step sizes. The chemical reaction starts in the compressed gas. Heat is released, the gas expands, and its pressure decreases. This is taking place until the combustion is completed and all of the reaction heat is released.
For large step sizes the solution exhibits totally unphysical bifurcating wave patterns with precursor numerical weak detonations (compare to [2]). All chemical

energy is released too soon in this precursor detonation wave. The slower moving trailing wave profile is an ordinary fluid dynamic shock. The numerical weak detonation wave is always moving at the speed of one mesh point per time step. Figure 6 shows the reference solution for $\alpha = 0.01$ after 0, 2000, ... 10000 time steps.

Figure 6: Numerical solution of the reacting Euler equations showing density ρ, pressure p and chemical energy $q_0 Z$ for $\alpha = 0.01$ after 0, 2000, ... 10000 time steps.

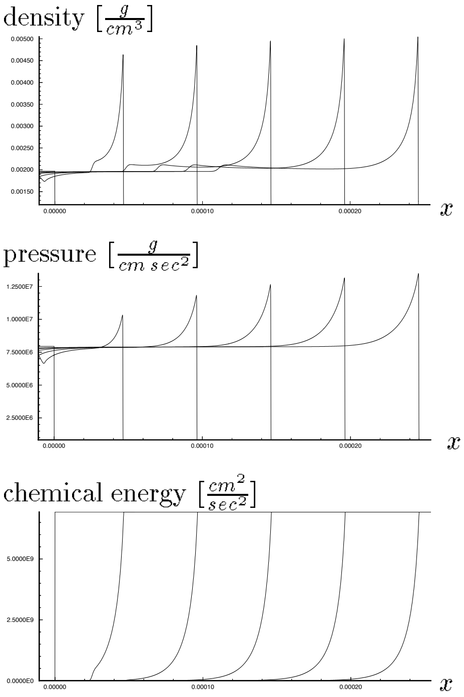
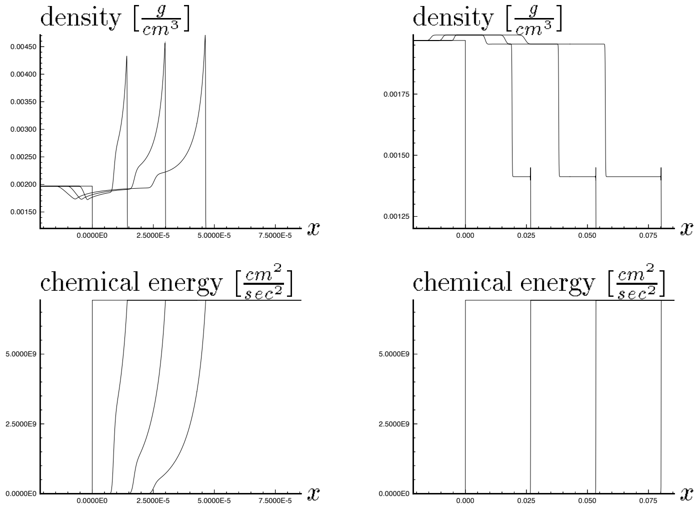


Figure 7 shows the density ρ and chemical energy $q_0 Z$ of the numerical solutions for small and large step sizes after 0, 500, 1000 and 1500 time steps. Notice that $\alpha = 0.01$ corresponds to the reference solution shown in Fig. 6. For small step sizes the solution profiles are still built up at these times, but for larger step sizes this is already done. Nonetheless, the typical behaviour of the solution computed with different step sizes can clearly be seen. Because of this and because adaptation — which is the aim of our investigations — starts with the first time step, we show here the numerical results at rather early times.

Next, numerical wave speeds are investigated. The numerical solution shows

Figure 7: Numerical solution of the reacting Euler equations showing density ρ and chemical energy $q_0 Z$ for small and large α after 0, 500, 1000 and 1500 time steps. The corresponding times are for $\alpha = 0.01$: 0, 1.6-10, 3.3-10, 4.9-10, and for $\alpha = 10$: 0, 1.7-7, 3.5-7, 5.2-7



$\alpha = 0.01$ $\qquad$ $\alpha = 10$

the following behaviour: The shock speed of the precursor detonation wave is equal to $\frac{\triangle x}{\triangle t}$ for large time and space steps. Numerical wave speeds are listed in Table 4. For large step sizes the numerical wave speed is determined by the formula

$$\dot{\sigma} = \frac{x(1500) - x(1000)}{t_{1500} - t_{1000}},$$

where $x(n)$ denotes the location of the precursor detonation wave at time t_n.

Remark 5.2 *For $\alpha = 0.01$ and $\alpha = 0.1$ the solution profile is still being built up between t_{1000} and t_{1500}. Therefore an additional averaged wave speed is computed for 9500 and 10000 time steps, where for $\alpha = 0.1$ CFL=0.4 is used. The speed of the detonation wave of the exact solution is $\dot{\sigma}_{exact} = 1.2321 \cdot 10^5$, cf. p. 29.*

Remark 5.3 *The half-reaction length $L_{1/2}$ is the distance for the half completion of the reaction starting from the front of the detonation. It is often used as a reference length-scale for detonation problems. The number of points in the half-reaction zone is approximated here by the ratio of $L_{1/2}/\triangle x$, where $L_{1/2}$ is estimated from the plots of the numerical solutions. Therefore, if the reaction*

Table 4: Numerical wave speeds $\dot{\sigma}$ compared to the ratio of the step sizes $\frac{\triangle x}{\triangle t}$ for various α. Furthermore, the approximate number of mesh points in the half-reaction zone $L_{1/2}$ is shown.

α	$x(1500)$ $-x(1000)$	t_{1500} $-t_{1000}$	$\dot{\sigma}$	$\frac{\triangle x}{\triangle t}$	Pts/$L_{1/2}$
0.01**	1.6500-5	1.6314-10	1.0114+5	1.6570+5	
0.1	1.6700-4	1.3934-9	1.1985+5	1.9331+5	
1	1.7700-3	1.4364-8	1.2322+5	1.8742+5	1.5
10	2.6735-2	1.7355-7	1.5405+5	1.5405+5	0.5
100	2.6735-1	1.7355-6	1.5405+5	1.5405+5	0*
100000	2.6735+2	1.7355-3	1.5405+5	1.5405+5	0*
α	$x(10000)$ $-x(9500)$	t_{10000} $-t_{95000}$	$\dot{\sigma}$	$\frac{\triangle x}{\triangle t}$	Pts/$L_{1/2}$
0.01**	1.6624-5	1.3991-10	1.1882+5	1.9129+5	80
0.1	0.8400-4	0.6828-9	1.2302+5	3.9166+5	9.5

The following notation is used: $.10 + 1 := .10 \cdot 10^1$
0*: see Remark 5.3.
**: see Remark 5.2.

is completed in at least one time step, no points lie in the reaction zone, which could therefore not be recovered.

Errors of the Shock Location

The local splitting error given by (46) and (48), respectively, is approximated for each of the single equations of the system, i.e. for $i = 1, \ldots, 4$. The approximate local splitting errors for these four equations are denoted by $\tilde{\mathcal{E}}^1_{spl}, \ldots, \tilde{\mathcal{E}}^4_{spl}$.
To determine the values U_L, U_R, kl and kr to be inserted into the error formulae, we proceed as described in Chapter 4.6. Considering the density (it could also be some other quantity showing the combustion spike), we approximate the region of smearing $[x_{kl(n)} - \frac{1}{2}\triangle x, x_{kr(n)} + \frac{1}{2}\triangle x]$ in such a way that just the (smeared-out) detonation wave is captured. That is we set $kr(n)$ to be the smallest index such that $u^n_j = u^n_R$, $j \geq kr(n)$ with $u := u_1 = \rho$. $kl(n) < kr(n)$ is defined to be the largest index with $u^n_{kl(n)} \geq u^n_{kl(n)-1}$ and the temperature greater than the ignition temperature. Then we set $u^n_{iL} := u^n_i|_{kl(n)}$, $i = 1, \ldots, 4$, where $u^n_i|_{kl(n)}$ is the value of u^n_i at location $x_{kl(n)}$. u^n_{iR} is defined analogously. For $\bar{u}_i^{n+\frac{1}{2}}$ we proceed in the same way. Then the captured wave is the one exhibited by solutions computed with small step sizes, and in solutions computed with large step sizes it is the one moving at the speed of one mesh cell per time step. The so-defined region of smearing corresponds to that discontinuity where the source

term actually works.

In the numerical computations the local errors of the location of the discontinuity are estimated by adding the absolute values of these parts. Remember that the local errors are scalar variables. Computing these errors for each of the variables u_i we have:

$$\begin{aligned}
\mathcal{E}^1_{est} &= 0 \\
\mathcal{E}^2_{est} &= |\tilde{\mathcal{E}}^2_{spl}| \\
\mathcal{E}^3_{est} &= |\tilde{\mathcal{E}}^3_{spl}| \\
\mathcal{E}^4_{est} &= |\tilde{\mathcal{E}}^4_{spl}| + |\triangle\tilde{\sigma}^4_1| + |\triangle\tilde{\sigma}^4_2|.
\end{aligned}$$

Notice that $\mathcal{E}^1_{est}$ is zero because of the simplifying assumptions that were made for the scalar analysis and carried over to the case of a system.

To begin, numerical results for solutions computed up to 1500 time steps were compared to computations up to 100 time steps. As they yield similar errors, all computations shown in this section will be done just up to 100 time steps. $\triangle\sigma_1$ resp. $\triangle\sigma_2$ are approximated by using (56) and setting $\frac{\partial}{\partial T}K(T) \equiv 0$.

The maxima of the local errors $\mathcal{E}^2_{est}$, $\mathcal{E}^3_{est}$ and $\mathcal{E}^4_{est}$ are shown in Figure 8. The results depicted in Fig. 8 show — as expected — that for small α the local errors are approximately the same — independent of the ODE-solver. The left figures plot the errors where $\tilde{\mathcal{E}}_{spl}$ is based on (46), while it is based on (48) in the figures on the right.

Adaptation

We show the results of the adaptation for the errors including a splitting error based on (46) as well as those based on (48). In both cases we proceed in the same way.

For the errors based on (46) our aim is $\tilde{E}_{rel} < 1\%$, and, based on test computations, the upper bound for the local error is set to

$$\mathcal{B} := 2 \cdot 10^4 \cdot \Delta t,$$

whereas for errors based on (48) we choose

$$\mathcal{B} := 5 \cdot 10^3 \cdot \Delta t.$$

Now we want to test if an adaptation based on $\mathcal{E}_{est}$ gives satisfactory results. The adaptation is carried out just for one ODE-solver because we expect all the various cases to behave analogously. We used ODE-solver 1 (exact integration of the linear ODE). To approximate the local errors of the location of the discontinuity, we assume q_4' to be piece-wise constant. The adaptation is not tested for unreasonable large step sizes, that is $\alpha > 10$.

This adaptation works as follows: The approximated local error of the location of the discontinuity $\mathcal{E}^4_{est}$ is expected to be smaller than a certain upper bound

B. If this assumption is not satisfied, the step size $\triangle x$ is bisected and $\triangle t$ is computed via CFL-condition. Then we start again with the initial data. This seems to be optimal as we observed that bisections took place just up to the fifth time step.

In a next run, the numerical solution is computed up to a fixed time $T = .5 \cdot 10^{-10}$. This is done because the solution profile is being built up at the beginning of the computation and because therefore the shock speed is not constant. These computations give similar results as the first run. Furthermore, the approximate relative global errors of the location of the discontinuity

$$\tilde{E}^i_{rel} = \frac{\sum_{n=1}^N \mathcal{E}^i_{est}(n)}{|\sigma^N - \sigma^0|}$$

are computed. $\mathcal{E}^i_{est}(n)$ denotes $\mathcal{E}^i_{est}$ of the n^{th} time step. Table 5 shows these results. It lists the various α^0 at time $t = 0$, the resulting smallest α, and the relative global errors $\tilde{E}^2_{rel}$, $\tilde{E}^3_{rel}$, and $\tilde{E}^4_{rel}$ belonging to the second, third, and fourth of the equations' systems.

The results of the adaptation based on the approximate local errors of the shock location which use the sharper estimation of the local splitting error (48) compared to the adaptation based on errors using (46) show that each of these two adaptations works as well as the other.

6 Conclusions

Based on scalar one-dimensional Riemann problems, an estimator for the error of the location of discontinuities has been derived. It can be used for adaptive choice of the step sizes, and the considered examples show that such an adaptation works well.

Of course, the adaptation does not work fully automatically yet. This could be the aim of further investigations. Furthermore, we would appreciate if the error-estimates could be extended to not piece-wise constant solutions.

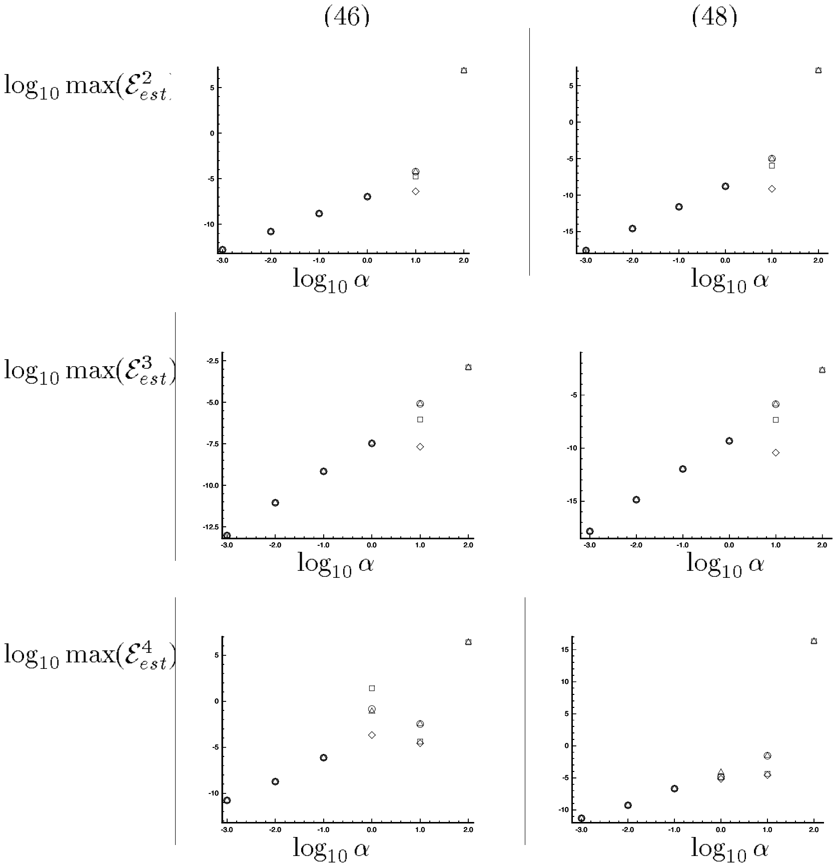
In [4], the theory presented in this paper is extended to planar two-dimensional problems. This is possible if the tangent to the discontinuity in the solution exists — as then the one-dimensional estimates are applied orthogonal to the tangent.

Table 5: Upper part: Results of the adaptation based on (46). The relative errors $\tilde{E}^4_{rel}$ remain less than 1%. Lower part: Results of the adaptation based on the sharper estimate (48). The relative errors $\tilde{E}^4_{rel}$ remain less than .61%. All solutions are computed up to time $T = .5 \cdot 10^{-10}$. The resulting step sizes are all about the same size.

α^0	α	$\tilde{E}^2_{rel}$	$\tilde{E}^3_{rel}$	$\tilde{E}^4_{rel}$
.10-2	.1000000-2	.5160-4	.3434-4	.1140-3
.10-1	.1000000-1	.4894-3	.2984-3	.2655-2
.50-1	.2500000-1	.1143-2	.6320-3	.1278-1
.10+0	.2500000-1	.1143-2	.6320-3	.1278-1
.25+0	.1562500-1	.7459-3	.4362-3	.5648-2
.40+0	.2500000-1	.1143-2	.6320-3	.1278-1
.50+0	.1562500-1	.7459-3	.4362-3	.5648-2
.60+0	.1875000-1	.8874-3	.5093-3	.7644-2
.10+1	.1562500-1	.7459-3	.4362-3	.5648-2
.50+1	.1953125-1	.9183-3	.5244-3	.8200-2
.10+2	.1953125-1	.9183-3	.5244-3	.8200-2
.10-2	.1000000-2	.8144-9	.5419-9	.7056-4
.10-1	.1000000-1	.7702-7	.4690-7	.1670-2
.50-1	.2500000-1	.4494-6	.2477-6	.7854-2
.10+0	.2500000-1	.4494-6	.2477-6	.7854-2
.25+0	.1562500-1	.1834-6	.1070-6	.3569-2
.40+0	.2500000-1	.4494-6	.2477-6	.7854-2
.50+0	.1562500-1	.1834-6	.1070-6	.3569-2
.60+0	.1875000-1	.2618-6	.1498-6	.4769-2
.10+1	.1562500-1	.1834-6	.1070-6	.3569-2
.50+1	.1953125-1	.2820-6	.1606-6	.5104-2
.10+2	.1953125-1	.2820-6	.1606-6	.5104-2

The following notation is used: $.10 + 1 := .10 \cdot 10^1$

Figure 8: Maxima of the estimated local errors of the location of the discontinuity comparing results for the different ODE-solvers. $\tilde{\mathcal{E}}_{spl}$ is based on (46) for the left figures and on (48) for the right figures.



□: Exact solution formula
△: Euler semi-implicit
◇: Euler explicit
○: Runge Kutta 2. order

References

[1] A. C. Berkenbosch. *Capturing Detonation Waves for the Reactive Euler Equations.* Proefschrift Technische Universiteit Eindhoven, 1995.

[2] P. Colella, A. Majda, and V. Roytburd. Theoretical and numerical structure for reacting shock waves. *SIAM J. Sci. Stat. Comput.*, 4:1059–1080, 1986.

[3] D. F. Griffiths, A. M. Stuart, and H. C. Yee. Numerical wave propagation in an advection equation with a nonlinear source term source terms. *SIAM J. Numer. Anal.*, 29:1244 — 1260, 1992.

[4] P. Klingenstein. *Nonlinear Hyperbolic Conservation Laws with Source Terms — Errors of the Shock Location.* Dissertation ETH, No. 12019, 1997.

[5] L. D. Landau and E. M. Lifschitz. *Hydrodynamik, Lehrbuch der Theoretischen Physik*, volume 6. Akademie-Verlag, Berlin, 1974.

[6] R. J. LeVeque. *Numerical Methods for Conservation Laws.* Lectures in Mathematics, ETH Zürich, Birkhäuser, 1992.

[7] R. J. LeVeque and H. C. Yee. A study of numerical methods for hyperbolic conservation laws with stiff source terms. *J. Comput. Phys.*, 86:187 – 210, 1990.

[8] A. Majda. A qualitative model for dynamic combustion. *SIAM J. Appl. Math.*, 41:50 – 93, 1981.

[9] M. Pandolfi. A contribution to the numerical prediction of unsteady flows. *AIAA Journal*, 22(5):602, 1984.

[10] R. B. Pember. Numerical methods for hyperbolic conseervation laws with stiff relaxation, i. spurious solutions. *SIAM J. Appl. Math.*, 53:1293 – 1330, 1993.

[11] J. Tegnér. Properties of detonation waves. 1992.

[12] H. C. Yee. *A class of high-resolution explicit and implicit shock-capturing methods.* NASA Technical Memorandum 101088, 1989.

Research Reports

No.	Authors	Title
97-16	R. Jeltsch, P. Klingenstein	Error Estimators for the Position of Discontinuities in Hyperbolic Conservation Laws with Source Terms which are solved using Operator Splitting
97-15	C. Lage, C. Schwab	Wavelet Galerkin Algorithms for Boundary Integral Equations
97-14	D. Schötzau, C. Schwab, R. Stenberg	Mixed *hp* - FEM on anisotropic meshes II: Hanging nodes and tensor products of boundary layer meshes
97-13	J. Maurer	The Method of Transport for mixed hyperbolic - parabolic systems
97-12	M. Fey, R. Jeltsch, J. Maurer, A.-T. Morel	The method of transport for nonlinear systems of hyperbolic conservation laws in several space dimensions
97-11	K. Gerdes	A summary of infinite element formulations for exterior Helmholtz problems
97-10	R. Jeltsch, R.A. Renaut, J.H. Smit	An Accuracy Barrier for Stable Three-Time-Level Difference Schemes for Hyperbolic Equations
97-09	K. Gerdes, A.M. Matache, C. Schwab	Analysis of membrane locking in *hp* FEM for a cylindrical shell
97-08	T. Gutzmer	Error Estimates for Reconstruction using Thin Plate Spline Interpolants
97-07	J.M. Melenk	Operator Adapted Spectral Element Methods. I. Harmonic and Generalized Harmonic Polynomials
97-06	C. Lage, C. Schwab	Two Notes on the Implementation of Wavelet Galerkin Boundary Element Methods
97-05	J.M. Melenk, C. Schwab	An *hp* Finite Element Method for convection-diffusion problems
97-04	J.M. Melenk, C. Schwab	*hp* FEM for Reaction-Diffusion Equations. II. Regularity Theory
97-03	J.M. Melenk, C. Schwab	*hp* FEM for Reaction-Diffusion Equations. I: Robust Exponentiel Convergence
97-02	D. Schötzau, C. Schwab	Mixed *hp*-FEM on anisotropic meshes
97-01	R. Sperb	Extension of two inequalities of Payne
96-22	R. Bodenmann, A.-T. Morel	Stability analysis for the method of transport
96-21	K. Gerdes	Solution of the $3D$-Helmholtz equation in exterior domains of arbitrary shape using HP-finite infinite elements