A characteristic method for one-dimensional scalar hyperbolic conservation laws with stiff source terms

Citation for published version (APA):

Document status and date:
Published: 01/01/1993

Publisher Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Abstract

In this paper we consider the initial value problem for scalar hyperbolic conservation laws in one space dimension, with (stiff) source terms. The source term admits two equilibrium states as solutions of the underlying characteristic equation, one unstable and one stable. An essential numerical difficulty for this initial value problem is that numerical reaction waves are propagating at non-physical wave speeds. An implicit numerical method is introduced, which produces discontinuities propagating at the correct speed. It is proved that this method preserves monotone profiles and an estimate for the global discretization error is given. The numerical results illustrate the correct behaviour of this method.
1 Introduction

In reactive gas dynamics, chemical reactions between the constituent gases need to be modelled along with the fluid dynamics. Problems of this form arise, for example, in combustion (cf. [1],[6]). We restrict our attention to chemical reactions with infinitely thin reaction zones in one-dimensional inviscid flows; so-called reacting shock waves (cf. [2]). In reacting shock waves, energy release occurs so quickly that molecular diffusion and thermal conductivity are usually unimportant transport mechanisms, and therefore they are ignored. If the effects of walls, heat sources and external forces are also ignored, we obtain essentially the Euler equations of gas dynamics, completed with the continuity equations for the different species. These latter equations include source terms which describe the chemical reactions. The total system of equations is often referred to as the reactive Euler equations. Since the time-scale of the chemical reaction in reacting shock waves is very small compared to the time-scale of the fluid dynamics, the source terms in reacting shock waves are therefore called 'stiff'.

When we attempt to solve the reactive Euler equations numerically, new problems arise that are absent in non-reacting flows. Apart from the increase in the number of equations, the main difficulty stems from the stiffness of the reaction terms. For stiff reactions it is possible to obtain stable numerical solutions that look reasonable and yet are completely wrong, because the discontinuities have the wrong locations. Thus, the numerical reaction waves are propagating at non-physical wave speeds (cf. [2]).

The same essential numerical difficulty of discontinuities travelling at incorrect speeds can be observed in scalar problems (cf. [5]). Therefore, numerical methods for scalar conservation laws with stiff source terms are studied in this paper. Clearly, scalar models are inadequate as a full test problem for any numerical method. However, they do model one essential difficulty encountered in reacting flow problems, namely discontinuities travelling at incorrect speeds. A study of these problems therefore suffices to analyze some of the difficulties that may arise also in more complicated systems (cf. [2]). The main issue in this paper is to obtain a numerical method for scalar conservation laws with (stiff) source terms, which produces discontinuities propagating at the correct speed.

This paper is organized as follows. In the next section one-dimensional scalar conservation laws with source terms are introduced. Furthermore, solutions of these conservation laws along characteristics are considered. In Section 3 an implicit numerical method is described. We study the behaviour of the corresponding numerical solution and we give an estimate for the discretization error. Finally, in Section 4, we present some numerical results, which illustrate the preceding analysis.

2 Scalar Hyperbolic Conservation Laws with Stiff Source Terms

In the following we consider the initial value problem for the scalar hyperbolic conservation law with source term in one space dimension. This initial value problem is given by

\[
\begin{align*}
\frac{\partial}{\partial t} u(x,t) + \frac{\partial}{\partial x} f(u(x,t)) &= q(u(x,t)), \quad \forall x \in \mathbb{R}, \forall t > 0, \quad (2.1a) \\
u(x,0) &= u^0(x), \quad \forall x \in \mathbb{R}, \quad (2.1b)
\end{align*}
\]
where $u : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ is the weak solution of (2.1), $f : \mathbb{R} \to \mathbb{R}$ is the flux-function and $q : \mathbb{R} \to \mathbb{R}$ is the source term. The weak solution $u$ is assumed to be unique. Furthermore, it is assumed that the flux-function $f$ is twice continuously differentiable. Let its first derivative be denoted by $a$, i.e. $a(u) = f'(u)$ with $a : \mathbb{R} \to \mathbb{R}$. The quasi-linear form of (2.1) reads

$$\frac{\partial}{\partial t}u(x,t) + a(u(x,t))\frac{\partial}{\partial x}u(x,t) = q(u(x,t)), \quad \forall x \in \mathbb{R}, \forall t > 0,$$

$u(x,0) = u^0(x), \quad \forall x \in \mathbb{R}.$

In the remainder $\alpha, \beta \in \mathbb{R}$ are given constants such that $0 \leq \alpha < \beta$. The source term $q$ is assumed to satisfy the following three conditions:

$$(2.2a) q \in C^2(\alpha, \beta),$$

$$(2.2b) q(\alpha) = q(\beta) = 0,$$

$$(2.2c) q(u)q''(u) < 0, \quad \forall u \in (\alpha, \beta).$$

Furthermore, it is assumed that the initial data satisfy

$$u^0(x) \in [\alpha, \beta], \quad \forall x \in \mathbb{R}. \quad (2.3)$$

**Example 2.1** A typical example of a scalar conservation law with a source term that satisfies (2.2) is

$$\frac{\partial}{\partial t}u + a(u)\frac{\partial}{\partial x}u = -\mu(u - \alpha)(u - \beta), \quad (2.4a)$$

$$u(x,0) = u^0(x) \quad (2.4b)$$

with $\mu > 0$ (cf. [3]). The source term in (2.4a) is chosen such that all starting values $\alpha < u^0(x) \leq \beta$ are tending to $\beta$ along the characteristic as time evolves. If $u^0(x) = \alpha$, then $u$ remains equal to $\alpha$ along the characteristic for all $t > 0$. Hence, the source term in (2.4a) admits two equilibrium states as solutions of the underlying characteristic equation, one unstable ($u = \alpha$) and the other stable ($u = \beta$).

In general, all source terms that satisfy (2.2) admit two equilibrium states as solutions of the underlying characteristic equation, one unstable and the other stable.

It is not essential to consider only one interval $[\alpha, \beta]$ in (2.2). Let $I = [u_L, u_R] \subset \mathbb{R}$ be a closed interval and let $u_1, u_2, \ldots, u_M \in I$, with $u_L = u_1 < u_2 < \ldots < u_M = u_R$. If $q$ satisfies (2.2) on each closed subinterval $[u_k, u_{k+1}]$ for $k = 1, 2, \ldots, M - 1$, then the analysis in this paper can be easily extended to the entire interval $I$. In the following example we consider a source term which satisfies (2.2) for two subintervals (cf. [5]).

**Example 2.2** Consider the following initial value problem:

$$\frac{\partial}{\partial t}u + a(u)\frac{\partial}{\partial x}u = -\mu u(u - 1)(u - \frac{1}{2}), \quad (2.5a)$$

$$u(x,0) = \begin{cases} 1, & x \leq x^0, \\ 0, & x > x^0 \end{cases} \quad (2.5b)$$

with $\mu > 0$. Note that the source term $q(u) = -\mu u(u - 1)(u - 1/2)$ satisfies (2.2) on the subintervals $[0,1/2]$ and $[1/2,1]$, respectively. It is easy to see that there are two stable equilibrium states, namely: $u = 0$ and $u = 1$, and there is one unstable equilibrium state, namely $u = 1/2$. 

2
In the following, the coupled systems (2.6)-(2.7), instead of (2.1), will be solved numerically. If the time-scale of (2.7) is very small compared to the time-scale of (2.6), then

\begin{align}
q(u^*(XO,t») > 0, & \quad uO(XO).
\end{align}

\begin{align}
\frac{d}{dt} x^*(XO,t) &= a(u(x^*(XO,t),t)), \quad \forall t > 0, \quad (2.6a) \\
x^*(XO,0) &= xo, \quad (2.6b)
\end{align}

In the remainder of this paper it is assumed that for each \((x,t) \in R \times [0,\infty)\) it is possible to determine a unique \(x^0 \in R\) and a corresponding function \(x^*(x^0,\cdot)\), which satisfies (2.6) such that \(x^*(x^0,t) = x\).

As a result of this assumption \(u\) can be expressed in terms of \(x^0\) and \(t\), and we write

\begin{align}
u(x,t) = u(x^*(x^0,t),t) = u^*(x^0,t).
\end{align}

Consider a point \((x,t) \in R \times [0,\infty)\) and suppose that \(x^*(x^0,t) = x\). It is easy to see that

\begin{align}
\frac{d}{dt} u^*(x^0,t) &= \frac{d}{dt} u(x^*(x^0,t),t) \\
&= \frac{\partial}{\partial t} u(x^*(x^0,t),t) + \frac{\partial}{\partial x} u(x^*(x^0,t),t) \frac{d}{dt} x^*(x^0,t) \\
&= \frac{\partial}{\partial t} u(x,t) + a(u(x,t)) \frac{\partial}{\partial x} u(x,t) \\
&= q(u(x,t)) \\
&= q(u^*(x^0,t)).
\end{align}

Note that

\begin{align}
u^*(x^0,0) = u(x^*(x^0,0),0) = u^0(x^0).
\end{align}

Since \((x,t)\) can be chosen arbitrarily, (2.1) can be replaced by the requirement that for all \(x^0 \in R\), a unique function \(x^*(x^0,\cdot)\) exists, which is defined by (2.6); and that for all \(x^0 \in R\) the following differential equation must hold

\begin{align}
\frac{d}{dt} u^*(x^0,t) &= q(u^*(x^0,t)), \quad \forall t > 0, \quad (2.7a) \\
u^*(x^0,0) &= u^0(x^0). \quad (2.7b)
\end{align}

In the following, the coupled systems (2.6)-(2.7), instead of (2.1), will be solved numerically. If the time-scale of (2.7) is very small compared to the time-scale of (2.6), then
the source term \( q \) is called 'stiff'. For example, the source term in (2.4) is stiff for large \( \mu \). Clearly, an explicit method will be inadequate if the source term is stiff. Therefore, an implicit method will be considered.

3 The Numerical Solution on a Fixed Spatial Interval

Let \( I = [x_L, x_R] \subset \mathbb{R} \) be a closed fixed spatial interval. We want to approximate the solution of (2.1) for all \( x \in I, t > 0 \). Therefore (2.1) should be completed with some boundary conditions, which are not specified here. For a given time step \( \Delta t \) the discrete time levels \( t_n \) are defined by \( t_n = n\Delta t, n = 0, 1, 2, \ldots \). For a given number of initial spatial mesh points \( (N+1) \), the initial mesh width \( \Delta x \) is computed as \( \Delta x = (x_R - x_L)/N \) and the initial spatial mesh is defined by \( x_i^0 = x_L + i\Delta x, i = 0, 1, \ldots, N \). Later on it will become clear that \( N \) need not be fixed as time increases. The resulting finite difference method produces approximations \( X_i^n \in I \) to the true solution \( x^*(x, t) \) of (2.6) and approximations \( U_i^n \in \mathbb{R} \) to the true solution \( u^*(x, t) = u(x^*(x, t), t) \) of (2.7). Let the numerical initial values be given by

\[
X_i^0 = x_i, \quad i = 0, 1, \ldots, N, \\
U_i^0 = u^0(x_i), \quad i = 0, 1, \ldots, N.
\]

The system (2.6a)-(2.7a) is solved numerically using the *Backward Euler method*. Suppose the numerical approximations \( X_i^n \) and \( U_i^n \) are known. The numerical approximations at time level \( t_{n+1} \) are then defined by

\[
X_i^{n+1} = X_i^n + \Delta t a(U_i^n + 1), \quad i = 0, 1, \ldots, N, \quad (3.1a) \\
U_i^{n+1} = U_i^n + \Delta t q(U_i^n + 1), \quad i = 0, 1, \ldots, N. \quad (3.1b)
\]

If \( U_i^n \in [\alpha, \beta] \), then (2.2) guarantees that (3.1b) has a unique solution in \([\alpha, \beta]\). Therefore, it is required that \( U_i^{n+1} \in [\alpha, \beta] \). This requirement is meaningful, since \( u^0(x) \in [\alpha, \beta] \) for all \( x \in \mathbb{R} \) implies that the true solution remains in the interval \([\alpha, \beta]\). Furthermore, if \( q(u) > 0 \) for all \( u \in (\alpha, \beta) \), then \( u^*(x^0, \cdot) \) is a non-decreasing function. Therefore, \( U_i^{n+1} \) should satisfy

\[
U_i^n \leq U_i^{n+1} \leq \beta.
\]

Analogously, if \( q(u) < 0 \) for all \( u \in (\alpha, \beta) \), then \( u^*(x^0, \cdot) \) is a non-increasing function and (3.1b) yields

\[
\alpha \leq U_i^{n+1} \leq U_i^n.
\]

The nonlinear equation (3.1b) will be solved using Newton iteration. The starting value of the Newton iterations is chosen to be equal to \( \beta \) if \( q(U_i^n) > 0 \), and equal to \( \alpha \) if \( q(U_i^n) < 0 \). This choice guarantees that \( U_i^{n+1} \in [\alpha, \beta] \), whenever \( U_i^n \in [\alpha, \beta] \).

Next the behaviour of the numerical solution is studied and an estimate for the global discretization error is given.

In the remainder of this section \( N \) is assumed to be fixed. For the true solution of (2.6)-(2.7) the following two statements hold for any fixed \( x^0 \in \mathbb{R} \). If \( u_0(x^0) > \alpha \) and \( q(u) > 0 \) for all \( u \in (\alpha, \beta) \), then \( u^*(x^0, t) \uparrow \beta \) as \( t \to \infty \); if \( u_0(x^0) < \beta \) and \( q(u) < 0 \) for all \( u \in (\alpha, \beta) \), then \( u^*(x^0, t) \downarrow \alpha \) as \( t \to \infty \). These properties are shared by the numerical solution.
Theorem 3.1 Let \( i \in \{0, 1, \ldots, N\} \). If \( \alpha < U^*_i < \beta \) and \( q(u) > 0 \) for all \( u \in (\alpha, \beta) \), then \( U^*_i \uparrow \beta \) as \( n \to \infty \). If \( \alpha \leq U^*_i < \beta \) and \( q(u) < 0 \) for all \( u \in (\alpha, \beta) \), then \( U^*_i \downarrow \alpha \) as \( n \to \infty \).

Proof Assume that \( U^*_i > \alpha \) and \( q(u) > 0 \) for all \( u \in (\alpha, \beta) \). If \( U^*_i = \beta \), then it follows from (2.2b) and (3.1b) that \( U^*_i = \beta \) for all \( n \geq 0 \), and the proof is finished. It is now proved by induction that \( \{U^*_i\}_{n=0}^{\infty} \) is an increasing sequence, if \( \alpha < U^*_i < \beta \). Let \( n \geq 0 \) and assume that \( \alpha < U^*_i < \beta \). Let the twice continuously differentiable function \( g : [\alpha, \beta] \to \mathbb{R} \) be defined by \( g(y) = y - \Delta t q(y) - U^*_i \). It follows from (2.2b), (3.1b) and \( q > 0 \) that \( g(\beta) = \beta - U^*_i > 0 \), \( g(U^*_i) = -\Delta t q(U^*_i) < 0 \) and \( g(U^*_i+1) = 0 \). Since it is required that \( U^*_i+1 \in [\alpha, \beta] \) and (2.2) implies that \( g(y) = 0 \) has a unique solution in \([\alpha, \beta]\), it is seen that \( U^*_i < U^*_i+1 < \beta \). The inductive step shows now that \( \{U^*_i\}^{\infty}_{n=0} \) is an increasing sequence with \( \beta \) as the only possible limit. The case \( \alpha < U^*_i < \beta \) and \( q(u) < 0 \) for all \( u \in (\alpha, \beta) \), is proved analogously. \( \square \)

A second important property is monotonicity preservation. If \( u^0 \) is monotone initially, then \( u(\cdot, t) \) is monotone for all \( t > 0 \). The following theorem shows that in this case the numerical solution also remains monotone at all time levels.

Theorem 3.2 If \( U^*_{i-1} \geq U^*_i \) for \( i = 1, \ldots, N \), then \( U^*_{i-1} \geq U^*_i \) for \( i = 1, \ldots, N \) and for all \( n \geq 0 \).

Proof The theorem is proved by induction. Let \( n \geq 0 \) and assume that \( U^*_i \geq U^*_{i-1} \) for \( i = 1, \ldots, N \). If \( U^*_i = \beta \), then (2.2b) and (3.1b) show that \( U^*_{i-1} = \beta \) and therefore \( \beta = U^*_i = U^*_{i-1} \). \( U^*_i = \alpha \), then (2.2b) and (3.1b) show that \( U^*_i = \alpha \) and therefore \( U^*_i \geq U^*_{i-1} = \alpha \). Hence, it may be assumed that \( \beta > U^*_i \geq U^*_{i-1} = \alpha \). Let the twice continuously differentiable function \( h : [\alpha, \beta] \to \mathbb{R} \) be defined by \( h(y) = y - \Delta t q(y) \). Since the source term \( q \) fulfils (2.2c), it is easy to see that \( h''(y) \neq 0 \) for all \( y \in (\alpha, \beta) \). Furthermore, it follows from (2.2b) and (3.1b) that \( h(\alpha) = \alpha, h(\beta) = \beta, h(U^*_i) = U^*_i = U^*_i = U^*_{i-1} \). Thus, \( \beta > U^*_i \geq U^*_{i-1} \) implies that \( h(\beta) > h(U^*_i) \geq h(U^*_{i-1}) > h(\alpha) \). Subsequently, using \( h''(y) \neq 0 \), it can be shown that \( U^*_{i-1} \geq U^*_i \). Since \( i \) was chosen arbitrarily, we have \( U^*_{i-1} \geq U^*_i \) for \( i = 1, \ldots, N \). The inductive step completes the proof. \( \square \)

An analogous proof shows that if \( U^*_i \geq U^*_{i-1} \) for \( i = 1, \ldots, N \), then \( U^*_{i-1} \geq U^*_i \) for \( i = 1, \ldots, N \) and for all \( n \geq 0 \).

Next the behaviour of the global discretization error of the numerical method (3.1) is studied. The discretization errors introduced by (3.1b) are called u-discretization errors and the discretization errors introduced by (3.1a) are called x-discretization errors. Since (3.1b) can be solved independently from (3.1a), the global discretization error introduced by (3.1b) is studied first. Let the global u-discretization error \( (Eu)^n_i \) at the point \( (x^0_i, t_n) \) be defined by

\[
(Eu)^n_i = U^*_i - u^*(x^0_i, t_n)
\]

for \( i = 0, 1, \ldots, N \) and \( n \geq 0 \). In order to derive an expression for \( (Eu)^n_i \), it is useful to define the local u-discretization error \( (Du)^n_i \). This local u-discretization error is defined
\[(Du)_i^n = \frac{1}{\Delta t} \{ u^*(x_{i+1}, t_{n+1}) - u^*(x_i, t_n) \} - q(u^*(x_{i+1}, t_{n+1})) \tag{3.3}\]

Since \(u^*(x_i, \cdot) \in C^2([0, \infty))\) (see (2.7)), a straightforward Taylor series expansion of \(u^*(x_i, \cdot)\) around \(u^*(x_i, t_{n+1})\) yields

\[(Du)_i^n = -\frac{1}{2} \Delta t \frac{d^2}{dt^2} u^*(x_i, \tau_n)\]

for some arbitrary point \(\tau_n \in (t_n, t_{n+1})\). It follows from the chain rule and (2.7) that

\[(Du)_i^n = -\frac{1}{2} \Delta t q'(u^*(x_i, \tau_n)) q(u^*(x_i, \tau_n)). \tag{3.4}\]

Finally, let \(Q\) be defined by

\[Q = \sup_{u \in (\alpha, \beta)} |q''(u)|.\]

We are now able to prove the following lemma.

**Lemma 3.3** Let \(q(u) > 0\) for all \(u \in (\alpha, \beta)\). Furthermore, let \(\epsilon = -q'(\beta)/(2Q)\). If, for some \(i \in \{0, 1, \ldots, N\}\) and \(n \geq 1\), \(u^*(x_i^n, t_n), U^a_i^n \in [\beta - \epsilon, \beta]\), then

\[|\langle Eu \rangle_i^n| \leq \frac{1}{1 + \Delta tc} \left( |\langle Eu \rangle_i^{n-1}| + \Delta t |(Du)_i^{n-1}| \right), \tag{3.5}\]

where \(c = -q'(\beta)/2 > 0\).

**Proof** From (3.1b) and (3.3) it follows that

\[\langle Eu \rangle_i^{n-1} = \langle Eu \rangle_i^n - \Delta t \{ q(U^a_i^n) - q(u^*(x_i^n, t_n)) \} + \Delta t (Du)_i^{n-1}.\]

It follows from the mean value theorem that

\[q(U^a_i^n) - q(u^*(x_i^n, t_n)) = q'(u^*(x_i^n, t_n) + \theta_n (Eu)_i^n) (Eu)_i^n\]

for some \(\theta_n \in (0, 1)\). A second application of the mean value theorem gives

\[q(U^a_i^n) - q(u^*(x_i^n, t_n)) = \left\{ q'(\beta) + c_1 (u^*(x_i^n, t_n) - \beta) + c_2 (U^a_i^n - \beta) \right\} (Eu)_i^n,\]

where \(c_1 = (1 - \theta_n)q''(\mu_n)\) and \(c_2 = \theta_n q''(\mu_n)\) for some \(\mu_n \in (\alpha, \beta)\). Note that \(c_1 < 0\) and \(c_2 < 0\). Since \(u^*(x_i^n, t_n), U^a_i^n \in [\beta - \epsilon, \beta]\) and \(0 < \epsilon = -q'(\beta)/2Q\) it follows that

\[-q'(\beta) - c_1 (u^*(x_i^n, t_n) - \beta) - c_2 (U^a_i^n - \beta) \geq -q'(\beta) + q''(\mu_n) \epsilon > 0.\]

It is easy to see that \(-q'(\beta) + q''(\mu_n) \epsilon \geq -q'(\beta) - Q \epsilon = -q'(\beta)/2 = c > 0\). Using the triangle inequality we deduce

\[(1 + \Delta tc) |\langle Eu \rangle_i^n| \leq |\langle Eu \rangle_i^{n-1}| + \Delta t |(Du)_i^{n-1}|.\]

The result (3.5) follows directly from the above inequality. \(\square\)

The lemma above is used to prove the following theorem.
Theorem 3.4 Let \( q(u) > 0 \) for all \( u \in (\alpha, \beta) \). Furthermore, let \( \varepsilon = -q'(\beta)/(2Q) \). If \( j \geq 0 \) is chosen such that for some \( i \in \{0, 1, \ldots, N\} \), \( u^*(x^0_i, t_j), U^*_i \in [\beta - \varepsilon, \beta] \), then

\[
\begin{align*}
\|E(u)\|_n &= \left( \frac{1}{1 + \Delta t^2} \right)^{n+1-j} \left\{ \|E(u)\|_{n-1}^j + \Delta t\|D(u)\|_{n-1}^j \right\} + \Delta t q(\beta - \varepsilon) \tag{3.7}
\end{align*}
\]

for \( n \geq j \), where \( c = -q'(\beta)/2 > 0 \).

Proof Note that (3.6) follows directly from the proof of Theorem 3.1. Therefore, it remains to prove inequality (3.7). Let \( i \in \{0, 1, \ldots, N\} \) and \( n \geq j \). Note that Lemma 3.3 holds with \( c = -q'(\beta)/2 > 0 \). After iterating (3.5) we obtain

\[
\begin{align*}
\|E(u)\|_n^j &\leq \left( \frac{1}{1 + \Delta t} \right)^{n+1-j} \left\{ \|E(u)\|_{n-1}^j + \Delta t\|D(u)\|_{n-1}^j \right\} + \Delta t \sum_{l=j}^{n-1} \left( \frac{1}{1 + \Delta t} \right)^{n-l} \|D(u)\|_l^j.
\end{align*}
\]

It is easy to see that if \( u \in [\beta - \varepsilon, \beta] \), then \( q'(u) = q'(\beta) + (u - \beta)q''(\mu) \leq q'(\beta) + \varepsilon Q = q'(\beta)/2 < 0 \) for some \( \mu \in (u, \beta) \). Using this together with \( q''(u) \leq 0 \) for all \( u \in [\alpha, \beta] \) and (3.4), we deduce

\[
\begin{align*}
\|D(u)\|_l^j &\leq \frac{1}{2} \Delta t |q'(u^*(x^0_i, t_{l+1}))| q(u^*(x^0_i, t_l)) \tag{3.8}
\end{align*}
\]

for all \( l \geq j \). It is straightforward that (3.8) implies that \( \|D(u)\|_1^j \leq -\Delta t q'(\beta)q(\beta - \varepsilon)/2 \). It follows from \( c > 0 \), and subsequently, \( 1/(1 + \Delta t) < 1 \), that inequality (3.7) holds. This completes the proof. \( \square \)

Results that are analogous to Lemma 3.3 and Theorem 3.4 hold if \( q(u) < 0 \) for all \( u \in (\alpha, \beta) \).

Inequality (3.7) clearly illustrates that the global \( u \)-discretization error made in the first \( (j - 1) \) steps is damped out as time evolves. Since \( u^*(x^0_i, t_l) \uparrow \beta \) as \( l \to \infty \) and \( q' \) is bounded, it follows from (3.8) that \( \|D(u)\|_l^j \to 0 \) as \( l \to \infty \). Hence, in practice \( \|D(u)\|_l^j \) will be much smaller than \( -\Delta t q'(\beta)q(\beta - \varepsilon)/2 \) for large \( l \). Therefore, in general, the global \( u \)-discretization error will be much smaller than the estimate given in (3.7). From (3.7) we deduce

\[
\begin{align*}
\|E(u)\|_n^j &\leq e^{-\Delta t/\Delta \alpha \alpha(t_n - t_j)} \left\{ \|E(u)\|_{n-1}^j + \Delta t\|D(u)\|_{n-1}^j \right\} + \Delta t q(\beta - \varepsilon).
\end{align*}
\]

Hence, if \( t_j = j\Delta t \) is bounded (as \( \Delta t \to 0 \)), and \( n\Delta t \) is fixed, then it follows from the inequality above that the global \( u \)-discretization error is estimated by the quantity \( |(E(u)\|_{n-1}^j|exp(-\varepsilon(t_n - t_j)) \right\) as \( \Delta t \to 0 \). Note that this latter estimate is very small for stiff source terms. In all model problems that we considered, numerical computations illustrate that \( t_j \) is bounded. Note that \( t_j \) is very small for stiff problems.

Similarly to (3.2) the global \( x \)-discretization error \( (Ex)^n_i \) at the point \( (x^0_i, t_n) \) is defined by

\[
\begin{align*}
(Ex)^n_i = X^*_i - x^*(x^0_i, t_n)
\end{align*}
\]

for \( i = 0, 1, \ldots, N \) and \( n \geq 0 \). We define the local \( x \)-discretization error \( (Dx)^n_i \) as

\[
\begin{align*}
(Dx)^n_i = \frac{1}{\Delta t} \left\{ x^*(x^0_i, t_{n+1}) - x^*(x^0_i, t_n) \right\} - a(u^*(x^0_i, t_{n+1})).
\end{align*}
\]
Since \( x^*(x_0^0, \cdot) \in C^2([0, \infty)) \), a straightforward Taylor series expansion of \( x^*(x_i^0, \cdot) \) around \( x^*(x_0^0, t_{n+1}) \) yields

\[
(Dx)_i^n = -\frac{1}{2} \Delta t \frac{d^2}{dt^2} x^*(x_0^0, \tau_n)
\]

for some arbitrary point \( \tau_n \in (t_n, t_{n+1}) \). Finally, let \( A \) and \( R \) be defined by

\[
A = \sup_{u \in (\alpha, \beta)} |a'(u)|,
\]

\[
R = \max_{x \in [x_L, x_R]} \sup_{t \geq 0} |\frac{d^2}{dt^2} x^*(x, t)|.
\]

The following theorem gives an estimate for \( |(Ex)_i^n| \).

**Theorem 3.5** Let \( q(u) > 0 \) for all \( u \in (\alpha, \beta) \). Furthermore, let \( \varepsilon = -q'(\beta)/(2Q) \). If \( j \geq 0 \) is chosen such that for some \( i \in \{0, 1, \ldots, N\} \), \( u^*(x_i^0, t_j), U^*_i \in [\beta - \varepsilon, \beta] \), then

\[
|(Ex)_i^n| \leq |(Ex)_i^{n-1}| + \frac{A}{c} \left\{ |(Eu)_i^{n-1}| + \Delta t |(Du)_i^{n-1}| \right\} + C_n(\varepsilon) \Delta t,
\]

where \( c = -q'(\beta)/2 > 0 \) and \( C_n(\varepsilon) = t_n(R/2 + Aq(\beta - \varepsilon)) > 0 \) for all \( n \geq j \).

**Proof** From (3.1a) and (3.10) it follows that

\[
(Ex)_i^{n-1} = (Ex)_i^n - \Delta t \left\{ a(U_i^n) - a(u^*(x_i^0, t_n)) \right\} + \Delta t (Dx)_i^{n-1}.
\]

It follows from the mean value theorem that

\[
a(U_i^n) - a(u^*(x_i^0, t_n)) = (Eu)_i^n a'(u^*(x_i^0, t_n)) + \theta_n(Eu)_i^n
\]

for some \( \theta_n \in (0, 1) \). Using the triangle inequality we deduce

\[
|(Ex)_i^n| \leq |(Ex)_i^{n-1}| + A \Delta t |(Eu)_i^n| + \Delta t |(Dx)_i^{n-1}|.
\]

After iterating the above inequality we obtain

\[
|(Ex)_i^n| \leq |(Ex)_i^{n-1}| + A \Delta t \sum_{l=j}^n |(Eu)_i^l| + \Delta t \sum_{l=j}^{n-1} |(Dx)_i^l|.
\]

Note that Theorem 3.4 may be applied to all \( |(Eu)_i^l| \) with \( l \geq j \), thus

\[
A \Delta t \sum_{l=j}^n |(Eu)_i^l| \leq \frac{A}{c} \left\{ |(Eu)_i^{j-1}| + \Delta t |(Du)_i^{j-1}| \right\} + \Delta t A \Delta t n q(\beta - \varepsilon).
\]

Furthermore, using (3.11), which implies that \( |(Dx)_i^l| \leq \Delta t R/2 \), we derive (3.12). This completes the proof. \( \Box \)

Note that after iterating inequality (3.13) and using that \( |(Dx)_i^0| \leq \Delta t R/2 \), we see that

\[
|(Ex)_i^n| \leq A \Delta t \sum_{l=1}^n |(Eu)_i^l| + \Delta t \frac{t_n R}{2}
\]

for \( i = 0, 1, \ldots, N \) and \( n \geq 0 \). Furthermore, it should be noted that for linear problems \( a(u) \) is constant for all \( u \), which implies that \( A = 0 \) and \( R = 0 \). Hence, for linear problems \( (Ex)_i^0 = 0 \) for \( i = 0, 1, \ldots, N \) and \( n \geq 0 \). This was expected, since (2.6) is solved exactly.
4 Numerical Results and Conclusions

In this section results are presented for the numerical method described in Section 3. Two conservation laws with a linear convective part and two conservation laws with a nonlinear convective part are considered.

There are other scalar model problems that correspond more closely to the one-dimensional reactive Euler equations (cf. [3]). However, these problems give a numerical behaviour similar to the examples presented in this section.

The numerical solutions are computed on the fixed closed spatial interval \([0, 1] \subset \mathbb{R}\). Mesh points (and corresponding \(U^*\)-values) are removed, when they become less than 0 or larger than 1. Furthermore, new mesh points are added, when the distance towards the closest boundary of the interval becomes too large. The corresponding \(U^*\)-values are derived from boundary conditions.

All numerical results presented here, illustrate the correct behaviour of method (3.1) around discontinuities. The solution is always monotone initially. The results show that the numerical solution remains monotone. This was proved in Theorem 3.2.

Firstly, two conservation laws with a linear convective part are considered (see Examples 2.1 and 2.2). For such linear problems (2.6) is solved exactly. Hence, it is expected that the discontinuities have the correct locations, since no errors are introduced in the computation of the characteristics. Furthermore, note that in linear problems the number of spatial mesh points remains constant, since the number of mesh points that is removed in one time step is equal to the number of added mesh points.

**Example 4.1** Consider the initial value problem (2.4) with \(\alpha = 0, \beta = 1, \mu = 100\) and a smooth initial profile that satisfies (2.3), i.e.

\[
\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} u = -100u(u - 1), \tag{4.1a}
\]

\[
u(x, 0) = \begin{cases} \frac{16x^3 - 12x^2 + 1}{4}, & 0 \leq x \leq \frac{1}{2}, \\ 0, & \frac{1}{2} < x \leq 1. \end{cases} \tag{4.1b}
\]

Note that the propagation speed is positive. Therefore, this initial value problem is completed with an extra condition at the left boundary, namely

\[
u(0, t) = 1, \quad \forall t \geq 0. \tag{4.1c}
\]

In Figure 1 the numerical results are compared to the exact solution of (4.1). This exact solution is derived by solving the corresponding ordinary differential equation along the characteristics. It is seen that the numerical discontinuity has the correct location.

Since the exact solution of (4.1) is known, we are able to compute the local \(u\)-discretization error exactly (see (3.3)). Next, we illustrate for (4.1) the remark made in Section 3 that in practice the local \(u\)-discretization error will be much smaller than the estimate \(-\Delta t q'(\beta)q(\beta - \varepsilon)/2\) used in Theorem 3.4. It is obvious that \(\varepsilon = 1/4\) and \(c = 50\). Assume that \(\Delta t = 0.01\), then we derive \(-\Delta t q'(\beta)q(\beta - \varepsilon)/2 = 9.375\). Let \(X^*_t = 0.38\), and thus \(U^*_t = 0.145152\). A straightforward computation shows that \(j = 2\) and that, for example, \(|(Du)_t|^\alpha = 7.051 \cdot 10^{-5}\), which is much smaller than 9.375.
Figure 1: Numerical solution ('o'-marks), exact solution (solid line) and initial solution (dashed line) of (4.1) with $\Delta t = 0.01$ and $\Delta x = 0.02$.

Example 4.2 In this second example (2.5) is considered, for which the occurrence of non-physical wave-speeds has been shown in [5]. We consider (2.5) on the interval $[0,1]$ with $\mu = 1000$, i.e.

\[
\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = -1000u(u-1)(u-\frac{1}{2}), \tag{4.2a}
\]
\[
u(x,0) = \begin{cases} 
1, & 0 \leq x \leq \frac{1}{2}, \\
0, & \frac{1}{2} < x \leq 1.
\end{cases} \tag{4.2b}
\]

Since all initial values are in equilibrium, i.e. $q(u(x,0)) = 0$, it is easy to see that the exact propagation speed of the shock wave is equal to 1 (the source term remains 0 for all time levels). Hence, this initial value problem should be completed with a left boundary condition, given by

\[
u(0,t) = 1, \quad \forall t \geq 0. \tag{4.2c}
\]

Since the source term is equal to 0 during all numerical calculations, (2.6) and (2.7) are solved exactly. Again the numerical results (see Figure 2) are compared to the exact solution, which illustrates the correct location of the discontinuity.
solution for $t = 0.1$

- 0.5
- 0.5
- 0.5
- 0.5
- 0.5
---
- 0.5
- 0.5
- 0.5
- 0.5
- 0.5

solution for $t = 0.2$

- 0.5
- 0.5
- 0.5
- 0.5
- 0.5

solution for $t = 0.3$

- 0.5
- 0.5
- 0.5
- 0.5
- 0.5
---
- 0.5
- 0.5
- 0.5
- 0.5
- 0.5

Figure 2: Numerical solution ('o'-marks), exact solution (solid line) and initial solution (dashed line) of (4.2) with $\Delta t = 0.01$ and $\Delta x = 0.02$.

Secondly, two conservation laws with a nonlinear convective part are considered. The convective part of (2.1a) is chosen equal to be Burgers' equation, i.e. $f(u) = u^2/2$. In both examples the source term is equal to the source term in (2.5a).

A new problem arises, that we have not mentioned before. Suppose that the initial condition is such that the solution of Burgers' equation without a source term is a simple wave. In this case the characteristics around an initial shock have diverging directions. If, for example, $u^*(x^0_i, 0) < u^*(x^0_{i+1}, 0)$, then the quantity $X^n_{i+1} - X^n_i$ can become very large as $n \to \infty$. Hence, at a certain time level, it is possible that a large area exists where the numerical solution is totally unknown. Certainly, this is not a desirable situation. Therefore, when $(X^n_{i+1} - X^n_i) > 2\Delta x$ at some time $t_n$, then new points are added between $X^n_i$ and $X^n_{i+1}$. The values at time $t_n$ of the numerical solution at these new points are computed by linear interpolation. Hence, it is possible that the number of spatial mesh points $(N + 1)$ changes during the numerical calculations. This is illustrated by the following example.

**Example 4.3** Consider the following initial value problem:

\[
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = -500u(u - 1)(u - \frac{1}{2}),
\]

(4.3a)

\[
u(x, 0) = \begin{cases} 
0, & 0 \leq x \leq \frac{1}{2}, \\
1, & \frac{1}{2} < x \leq 1.
\end{cases}
\]

(4.3b)
Clearly, no extra boundary conditions are necessary. Furthermore, it can be shown that a shock wave is created moving with propagation speed equal to $1/2$. Therefore, the numerical results are compared to the exact solution of

$$\frac{\partial}{\partial t} u + \frac{1}{2} \frac{\partial}{\partial x} u = 0$$

with initial condition

$$u(x, 0) = \begin{cases} 0, & 0 \leq x \leq \frac{1}{2}, \\ 1, & \frac{1}{2} < x \leq 1. \end{cases}$$

It is seen in Figure 3, that the numerical propagation speed is equal to the exact propagation speed.

![Numerical solution and initial solution](image)

It is seen in Figure 3, that the numerical propagation speed is equal to the exact propagation speed.

Another problem that arises in the nonlinear case is that characteristics can intersect. It is required that $X_{i+1}^n < X_i^* + 1$ for all possible $i$ and $n \geq 0$, since otherwise a multivalued compression wave is computed (cf. [4]). Therefore, if $X_{i+1}^n - X_i^*$ is negative or nearly equal to 0, then both $X_i^*$ and $X_{i+1}^*(n)$ (and the corresponding $U^*$-values) are deleted. The discontinuity is not smeared out as time evolves, since no interpolation techniques are used. Again, it is possible that the number of spatial mesh points changes. The intersecting of characteristics is illustrated by the following example.
Example 4.4 Consider the following initial value problem:

\[
\frac{\partial}{\partial t} u + u \frac{\partial}{\partial x} u = -200u(u - 1)(u - \frac{1}{2}), \quad (4.6a)
\]

\[
u(x,0) = 1 - x, \quad 0 \leq x \leq 1. \quad (4.6b)
\]

This initial value problem is completed with a left boundary condition,

\[
u(0,t) = 1, \quad \forall t \geq 0. \quad (4.6c)
\]

Note that from a certain time all values are in equilibrium and no new intermediate states \(0 < u < 1\) are created. Therefore, the exact propagation speed of the shock wave becomes equal to 1/2. Hence the solution of (4.6) is compared to the solution of (4.4) with initial condition

\[
u(x,0) = \begin{cases} 
1, & 0 \leq x \leq \frac{1}{2}, \\
0, & \frac{1}{2} < x \leq 1.
\end{cases} \quad (4.7)
\]

Figure 4 shows that the numerical propagation speed is equal to the exact propagation speed.

Figure 4: Numerical solution ('o'-marks) and initial solution (dashed line) of (4.6) with \(\Delta t = 0.01\) and \(\Delta x = 0.02\); exact solution (solid line) of (4.4) with initial condition (4.7).
References


