Capturing detonation waves for the reactive Euler equations

Citation for published version (APA):
Berkenbosch, A. C. (1995). *Capturing detonation waves for the reactive Euler equations*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Mathematics and Computer Science]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR445469

DOI:
10.6100/IR445469

Document status and date:
Published: 01/01/1995

Document 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.

Download date: 04. Oct. 2023

CAPTURING DETONATION WAVES FOR THE REACTIVE EULER EQUATIONS

A.C. BERKENBOSCH

CAPTURING DETONATION WAVES FOR THE REACTIVE EULER EQUATIONS

CAPTURING DETONATION WAVES FOR THE REACTIVE EULER EQUATIONS

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de Rector Magnificus, prof.dr. J.H. van Lint, voor een commissie aangewezen door het College van Dekanen in het openbaar te verdedigen op donderdag 5 oktober 1995 om 16.00 uur

door

ADRIAAN CORNELIS BERKENBOSCH

Geboren te Leiden

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. R.M.M. Mattheij
en
prof.dr.ir. A.A. van Steenhoven

Copromotor: dr.ir. J.H.M. ten Thije Boonkkamp

©1995 Æneas, uitgeverij van vakinformatie, Postbus 356, 5680 AJ Best

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Berkenbosch, Adriaan Cornelis

Capturing detonation waves for the reactive Euler
equations / Adriaan Cornelis Berkenbosch. - Best: Æneas,
uitgeverij van vakinformatie
Proefschrift Technische Universiteit Eindhoven. - Met index.
Lit. opg. - Met samenvatting in het Nederlands.
ISBN 90-75365-04-7
NUGI 811
Trefw.: Euler-vergelijkingen.

CONTENTS

Preface ix

1 Introduction 1
- 1.1 Detonation Waves . . . 1
  - 1.1.1 General Remarks . . . 1
  - 1.1.2 The One-Dimensional Model . . . 3
  - 1.1.3 Basic Assumptions in this Thesis . . . 5
- 1.2 Hyperbolic Conservation Laws . . . 5
- 1.3 The Scope of this Thesis . . . 7
- 1.4 The Contents of this Thesis . . . 11

2 The One-Dimensional Reactive Euler Equations 13
- 2.1 The One-Dimensional Conservation Equations for Reacting Gas Flow 13
- 2.2 Derivation of the One-Dimensional Reactive Euler Equations . . . 16
- 2.3 Dimensionless Equations . . . 21
- 2.4 The Reactive Rankine-Hugoniot Equations . . . 24
  - 2.4.1 Derivation of the Reactive Rankine-Hugoniot Equations . . . 24
  - 2.4.2 The Hugoniot Curve . . . 26
  - 2.4.3 Various Types of Processes. . . . 27
- 2.5 The ZND Model for Detonation Waves . . . 30

3 One-Dimensional Hyperbolic Conservation Laws 37
- 3.1 Introduction . . . 37
- 3.2 Weak Solutions . . . 40
- 3.3 The Riemann Problem for Homogeneous Conservation Laws . . . 44
  - 3.3.1 Preliminaries . . . 44
  - 3.3.2 Rarefaction Waves . . . 46

3.3.3 Contact Discontinuities 48
3.3.4 Shock Waves and the Lax Entropy Condition 49
3.3.5 General Solution of the Riemann Problem 50
3.3.6 Riemann Invariants 52
3.4 The Riemann Problem for the Nonreactive Euler equations 53

4 Finite Difference Methods for Conservation Laws 57
4.1 Some Basic Numerical Concepts 58
4.2 Discrete Conservation 64
4.3 Modified Equations 65
4.4 The Numerical Entropy Condition 67
4.5 Godunov-type Methods 69
4.5.1 Introduction 69
4.5.2 Basic Godunov Method 70
4.5.3 Roe's Method 72

5 High Resolution Methods 79
5.1 Monotone and TVD Methods for Scalar Conservation Laws 80
5.2 Flux Limiter Methods for Scalar Conservation Laws 84
5.3 Slope Limiter Methods 89
5.3.1 Introduction 89
5.3.2 Scalar, Linear Conservation Laws 90
5.3.3 Linear Systems of Conservation Laws 92
5.3.4 Nonlinear Systems of Conservation Laws 93

6 Time Splitting Methods 97
6.1 Introduction 98
6.2 The Splitting Error 99
6.3 Numerical Methods Based on Time Splitting 106

7 Numerical Solution of a Simplified Detonation Model 113
7.1 A Simplified Detonation Model 114
7.1.1 Introduction 114
7.1.2 The Analogue of the Reactive Rankine-Hugoniot Equations . . 115
7.1.3 The Analogue of the ZND Model 117
7.2 Numerical Computation of Strong or CJ Detonation Waves 120
7.3 Detonation Capturing for Stiff Combustion Chemistry 126
7.3.1 The Avoidance of Nonphysical Detonation Waves 126
7.3.2 Numerical Results 134
7.4 A High Resolution Method 136

8 Numerical Solution of the Reactive Euler Equations 141
8.1 Numerical Computation of Strong or CJ Detonation Waves 142
8.2 The Avoidance of Nonphysical Detonation Waves 148

8.3 A High Resolution Method 152
8.4 Some Remarks on Front Tracking 155

Nomenclature **159**

Bibliography **161**

Index **169**

Samenvatting **173**

Curriculum Vitae **175**

PREFACE

Research in combustion science is becoming one of the most challenging fields of advanced engineering. Combustion is a subject that exhibits both fundamental and applied characters and includes theoretical and experimental aspects. Therefore it involves some of the most advanced research fields, e.g. gas dynamics, turbulence, stability, thermochemistry and chemical kinetics. At the same time, it offers stimulating problems in numerical analysis and applied mathematics.

As a matter of fact, the interaction of fluid motion with the thermodynamic changes associated with chemical reactions is a difficult problem, especially when unsteady phenomena have to be analysed. For instance, detailed reaction schemes involve a number of chemical species of the order 10 to 100 and about 30 to 1000 elementary reactions. In addition, many of these reactions represent very rapid relaxation processes so that the reaction system introduces very stiff source terms in the balance equations for the chemical species. As a consequence, fully resolved numerical simulations of multidimensional combustion problems are far from being routine. Nevertheless, in recent years the theoretical basis of combustion phenomena and the capability of numerical prediction of the relevant flow fields have been impressively developed.

A special kind of combustion is called detonation. Detonation waves are high speed combustion waves in which very fast chemical reactions take place. In the fifty years since Zel'dovich, von Neumann and Döring independently formulated treatments of detonation, many workers have contributed to further understanding of the process. In many applications detonation waves are described by a system of first order hyperbolic conservation laws with source terms, the so-called reactive Euler equations.

From a numerical point of view, detonation waves are interesting since they have a discontinuous structure including a strong compression shock. This leads to many well-known numerical problems. However, a new and interesting problem is the treatment of the source terms (the chemical reactions). Especially, since the time scale associated with the flow is several orders of magnitude larger than typical time scales of the chemical reactions in general. Consequently the source terms lead to a stiff system of equations.

In this thesis we like to develop a detonation capturing method, which describes the global behaviour of the detonation wave, without describing all physical details. To this end space grids and time steps are used appropriate for the fluid dynamics but not for the rapid chemical reaction. Certainly, in this situation any information about the detailed structure of the detonation wave is lost. On these coarse grids the detonation wave is represented as a gas dynamic discontinuity of zero width. For these kind of problems it is possible to obtain completely wrong numerical detonation wave speeds and what we should demand is that the method produces a sharp resolution of the detonation wave at the correct location (the numerical detonation wave should propagate with the correct speed).

Although this is a mathematical thesis, I also included some physics in order to present a more complete picture. This thesis studies some problems occurring in the theory of detonation waves and hopefully, it reflects some of my enthusiasm about this subject.

At this place I would like to thank some people for their contribution to my work. First, I would like to thank dr.ir. Jan ten Thije Boonkkamp and dr. Rik Kaasschieter for their stimulating support during the last four years. They advised me and collaborated with me on many occasions. Furthermore, I am grateful to dr.-ing. Rupert Klein from the R.W.T.H. Aachen. He generously shared with me many of his interesting ideas and he taught me many aspects of the theory of detonation waves. His enthusiasm constantly served as a stimulation to me. I also thank prof.dr. Mattheij for the constructive criticism, the helpful comments and the support during my Ph.D. . Naturally, I would like to thank all my friends and colleagues for their support, advice and collaboration. I am especially grateful to Judith. Her patience, care and support were very important to me and helped a lot in writing this thesis. Finally, I thank my roommate Peter Ferket for the endless discussions about mathematics, politics and soccer which made my stay at the T.U.E. much more pleasant and interesting.

Arco Berkenbosch Eindhoven, August 1995.

1

INTRODUCTION

Detonation waves are high speed combustion waves in which very fast chemical reactions take place. In practice the time scales of the chemical reactions are very small compared with the time scale of the flow and in many practical cases we cannot afford a sufficient resolution of the chemistry. The main goal of this thesis is to develop a numerical method that captures the detonation wave correctly, without solving the chemistry in detail.

This chapter is organized as follows. In the first section we describe some typical properties of detonation waves and briefly present the underlying physical model. Furthermore, we present the basic assumptions of this thesis. Section 1.2 has a more general nature. We discuss general hyperbolic conservation laws and especially the possible nonuniqueness of solutions. The main goals of this thesis are described in Section 1.3. Finally, in Section 1.4 we outline briefly the contents of this thesis.

1.1. DETONATION WAVES

1.1.1. General Remarks

The rapid and violent form of combustion called *detonation* differs from other forms in that all the important energy is transferred by strong compression waves, with negligible contributions from other processes like heat conduction, which are important in flames. The leading part of a detonation front is a strong shock wave propagating into the fresh mixture. This shock heats the material by compressing it, thus triggering a chemical reaction, and a balance is attained such that the chemical reaction supports the shock (see Figure 1.1). In this process fuel mass is consumed 10^3 to 10^6 times faster than in a flame, making detonations easily distinguishable from other combustion processes. In a detonation wave the energy is converted very rapidly. For example a serious gaseous detonation converts energy at a rate of 10^{10} Watts per square meter of its detonation front [26].

The most easily measured characteristic property of a detonation is the propagation speed of the front into the explosive. The front of a detonation wave initiated at one end of a large-diameter tube of explosive material is found to approach a nearly plane shape and a constant propagation speed. Thus it seems reasonable to assume that a limiting speed exists, and that in the limit the chemical reaction takes place in a steadily propagating zone in the explosive [26, 66]. A mathematically tractable problem related to physical reality eventually is that of a plane, steady detonation. It is generally assumed that the solution of this problem describes detonation experiments for a situation when the length approaches infinity. Therefore, experiments designed for validation of the theory are usually measurements of a set of long cylinders of different diameters initiated at one end. Subsequently, the results are 'extrapolated' to infinite diameter. Unfortunately, to our knowledge no other problem, not even the seemingly simple problem of a spherically expanding detonation, is studied experimentally in a proper way yet, so a direct validation (without extrapolation to infinite size) is not possible [2, 26, 43, 45].

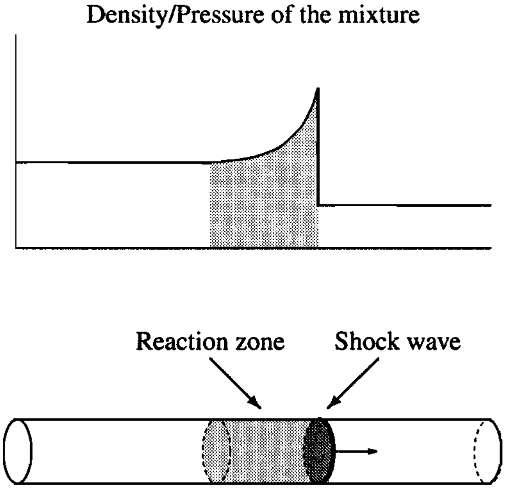

Figure 1.1. Schematic diagram of the detonation wave travelling through a tube and the corresponding density/pressure profile.

A socially and politically sensitive application of detonation research is the safety of nuclear reactors. During an accident thousands of kilograms of hydrogen can be released into the containment atmosphere, from an oxidation reaction of the metal shells of the reactor fuel elements with the cooling water. In this process a huge amount of the highly explosive hydrogen-air-steam mixture is formed. If this detonates, it will blow apart any existing reactor containment. Therefore, realistic computations for the hydrogen-air-steam system are desirable and currently of great interest. For this system detailed reaction schemes involve a number of chemical species of the order 10 to 100 and about 30 to 1000 elementary reactions. In addition, many of these reactions represent very rapid relaxation processes so that the reaction system introduces very stiff source terms in the balance equations for the chemical species. As a consequence, fully resolved numerical simulations of multidimensional detonation problems are far from being routine. All chemical reactions used as examples in this chapter, are taken from the hydrogen-air-steam detonation [66, 71].

1.1.2. The One-Dimensional Model

The one-dimensional model involves simple considerations of the shock front and a reaction zone travelling with the same speed behind it. The propagation speed is supersonic relative to the gas ahead of the front. Furthermore, the speed depends on the nature of the fuel and the composition of the detonable medium, but generally lies in the range 1.5 to 4.0 km/s. The original model proposed by Chapman and Jouguet, to be called the *CJ model* from now on, incorporates an energy input term in the conventional Rankine-Hugoniot analysis [26, 85]. It is related to the propagation speed of the front, through the assumption that the composition of the products corresponds to equilibrium at the temperature and density in the flow behind it. The model is commonly used to produce reliable predictions of global propagation speeds, measured under conditions in which thermal and viscous losses are negligible.

There have been frequent attempts to verify the CJ model experimentally, the majority comparing the propagation speed of detonations to the theoretical value. In Table 1.2 we compare the propagation speeds measured in a 100 mm diameter tube to the speeds predicted by the CJ model. The measured speeds are consistently lower than the theoretical values [66]. This might be expected if losses to the wall are taken into account. The CJ model gives no insight into the internal structure of detonation waves. Independently from each other, Zel'dovich, von Neumann and Döring developed a model which explains the internal structure of detonation waves, the so-called *ZND model*. The original ZND model is based on the concept of the leading shock wave producing a flow of the required density and temperature to trigger exothermic reactions at a short distance behind the shock, depending on the losses in this region (see Figure 1.1) [26, 66, 85].

Mixture	Observed speed (m/s)	Theoretical speed (m/s)
$4H_2 + O_2$	3344	3425
$3H_2 + O_2$	3156	3197
$2H_2 + O_2$	2825	2853
$H_2 + O_2$	2320	2333
$H_2 + 2O_2$	1909	1941
$H_2 + 3O_2$	1691	1759

Table 1.2. Comparison of the experimental propagation speed of the detonation front and the propagation speed of the detonation front predicted by the CJ model [22].

The ZND theory does not give any clues concerning stability of such travelling waves. As it turns out these waves are, in many cases, dynamically unstable; small perturbations of the wave structure grow in time, causing the original planar front to wrinkle and the originally constant wave speed to change with time. Phenomena of detonation instabilities have attracted a substantial experimental and theoretical effort. For a brief review or a reader's guide to some parts of this material, we refer to [75]. In this thesis we assume that a detonation wave propagates and actually survives for some time. This assumption is crucial and it is closely related to the question of stability [10, 11, 38].

It appears from Figure 1.1 that the main exothermic reaction starts immediately behind the shock; however, this is not very realistic. In practice it takes some time between on the one hand the shock compression and on the other hand the onset of exothermic chemical reactions. This time is called the *ignition delay time* t_{ign}. The exothermic reactions are generally associated with the recombination of radicals and atomic species. Moreover, in general there are three different types of reaction steps involved in the chemical reaction. First, there are initiating steps such as dissociation of molecular species, e.g.

$$H_2 + M \rightarrow 2H + M,$$

where M is some particle. The initiating steps are followed by a set of reactions such as chain-branching steps in which the reaction results in the production of a larger number of active species, e.g.

$$H + O_2 \rightarrow OH + O.$$

There are also recombination steps responsible for the major portion of heat release, in which radicals and atomic species are recombinated, e.g.

$$H + OH + M \rightarrow H_2O + M.$$

Hence, during the ignition delay time t_{ign}, initiating and chain-branching steps are performed. Experiments show that t_{ign} is $1 - 10\,\mu s$ at the temperature and the density of the flow behind a shock propagating with a speed that is typical for a detonation wave. In turn, these times suggest that the separation of the leading shock and the main reaction zone in gaseous mixtures is typically in the range of $1 - 10$ mm. The exponential dependence of the different reaction rates on the temperature of the gas mixture is generally expressed in terms of the *activation energy* (roughly speaking the energy that is necessary to initiate the reaction). This activation energy is high for initiating steps such as dissociation, lower for the attack of radicals on molecular species and approaches zero for recombination steps. More realistic expressions, representing the ignition delay time t_{ign} for chain reactions, generally involved in detonations, have been introduced in the model in order to estimate the separation between the shock and the reaction zone. In Table 1.3 we summarize the previous results.

Reaction step	Activation energy	Heat release
initiation	high	negligible
chain-branching	low	low
recombination	negligible	high

Table 1.3. Different types of reaction steps and the corresponding activation energy and heat release.

1.1.3. Basic Assumptions in this Thesis

In practice the time scales of the chemical reactions are very small compared to the time scale of the fluid dynamics. Therefore, a sufficient resolution of the chemistry cannot be afforded in general. In this thesis we will study and develop numerical methods for space grids and time steps appropriate to resolve the fluid dynamics but not for the rapid chemical reactions. Especially we are interested in the global behaviour of the detonation wave, so quantities of interest are the propagation speed of the front, the pressure behind the reaction zone, etc. The three basic assumptions we make are that

(i) molecular diffusion, thermal conduction and viscous effects are ignored;

(ii) the detonation wave is one-dimensional;

(iii) the gas is a binary mixture in which only one chemical reaction takes place.

Assumption (i) is physically realistic, since detonation waves are propagating so fast that molecular diffusion, thermal conduction and viscous effects are usually unimportant, and therefore they may be ignored [26, 85].

One-dimensional models (assumption (ii)) allow for a convenient framework for discussing the properties of the detonation front itself and can be developed further to indicate the mechanisms governing detonation limits. However, they fail to give any guidance on the complex interactions of detonations with confinement. In order to understand these interactions, the multi-dimensional nature of detonations must be considered. However, most of the numerical methods currently in use are heavily based on one-dimensional methods, generalized by "dimensional splitting" or similar techniques. For this reason it is useful to study and develop one-dimensional numerical methods.

Finally, in practical applications assumption (iii) is unrealistic. As noted before, in general a lot of chemical species and a lot of chemical reactions are involved. In order to keep our problem mathematically tractable we restrict ourselves to binary mixtures. Recall that we are not interested in a sufficient resolution of the chemistry. Moreover, global properties, like the propagation speed of the detonation wave, are assumed to be completely determined by the initial chemical composition, pressure and temperature of the mixture. So, we consider the one-step reaction in which a reactant $\mathcal{R}$ (unburnt gas) is converted into a product $\mathcal{P}$ (burnt gas). For this one-step reaction we use a relatively high activation energy and a large heat release (see Table 1.3).

Although physically not very realistic, the one-dimensional model studied in this thesis is interesting for testing and analysing numerical methods. Clearly, the model is inadequate as a full test problem for any numerical method for detonations. However, a study of this problem suffices to analyse some of the difficulties that may arise in the more complicated detonation models.

1.2. HYPERBOLIC CONSERVATION LAWS

Many (practical) problems in science and engineering involve conserved quantities and lead to *hyperbolic conservation laws*. Hyperbolic conservation laws can be formulated as a system of first order partial differential equations. The Euler equations of gas dynamics are an important example of hyperbolic conservation laws [59]. Moreover, one-dimensional detonation waves are in general described by the Euler equations of gas dynamics, completed by the balance equations for the various species (see (2.21) below). These latter equations include source terms describing the chemical reactions. The complete system of equations is called the reactive Euler equations and is a system of first order hyperbolic conservation laws with source terms. Other examples arise in meteorology and astrophysics.

Exact solutions of conservation laws are hard to find in general, and therefore we have to devise and study numerical methods to approximate their solutions. Of course the same is true, more generally, for any nonlinear partial differential equation, and to some extent the general theory of numerical methods for nonlinear PDEs applies to systems of conservation laws also. However, apart from practical considerations, there are two other reasons for studying numerical methods for hyperbolic conservation laws in particular. First, there are special difficulties associated with solving hyperbolic conservation laws (e.g. shock formation) which must be dealt with carefully when developing numerical methods. Methods based on naive finite difference approximations may behave well for smooth solutions, but can give poor results when discontinuities are present. Secondly, a great deal is known about the mathematical structure of these equations and their solutions [49, 79]. This theory can be used to develop special methods to overcome some of the numerical difficulties arising from a more naive approach.

Many difficulties are caused by the fact that hyperbolic conservation laws may have discontinuous solutions. For instance detonation waves have a discontinuous structure including a strong leading shock front (see Figure 1.1). Obviously, discontinuous solutions do not satisfy the PDE in the classical sense at all points, since the derivatives are not defined at the discontinuities. Therefore, we have to define what is meant by a solution in this case. The general form of a conservation law is an integral equation and the PDE is derived by imposing additional smoothness assumptions on the solution. The crucial fact is that the integral equations are valid even for discontinuous solutions. In order to allow discontinuous solutions of the PDE also, we consider *weak solutions*. It can be shown that a function is a solution of the original integral equation if and only if it is a weak solution of the PDE [79].

Unfortunately, weak solutions may be nonunique for a given set of initial data. If our conservation law models the real world, then there clearly exists one "physically relevant" weak solution and we must have some mechanism to characterize it. Hyperbolic conservation laws often arise in models of physical processes ignoring the effects of various dissipative mechanisms. In more accurate models, these mechanisms make their appearance felt by the presence of higher order derivatives in the equation multiplied by small coefficients, e.g. relating to small viscous effects. In the limit, i.e. when these coefficients approach zero, the solution of the higher order equation should converge to the solution of the system of first order conservation laws. Hence the unique, physically relevant weak solution is, roughly speaking, defined as the stable limit of this vanishing viscosity mechanism [47, 49, 79].

The vanishing viscosity method has some direct use in the analysis of conservation laws, but is clearly not optimal since it requires studying more complicated systems of equations. This is precisely what we tried to avoid by introducing the inviscid equations in the first place. For this reason we like to find other conditions which can be imposed directly on weak solutions of hyperbolic conservation laws in order to pick out the physically relevant solution.

In nonreactive gas dynamics the entropy is constant along particle paths in smooth flows. The entropy jumps to a higher value as the gas crosses a shock discontinuity. It follows from the second law of thermodynamics that the entropy never jumps to a lower value. This gives the extra condition, the so-called *entropy condition*, which picks out the correct weak solution in gas dynamics. For the ZND model for detonation waves, the relevant weak solution is characterized by Jouguet's rule, which says that the flow in a front attached frame of reference is supersonic in the unburnt gases ahead of the wave and subsonic or sonic in the burnt gases behind the wave.

For general hyperbolic conservation laws it is often possible to impose an extra condition upon the solution, such that a physically relevant solution is obtained. As in gas dynamics this condition is called the entropy condition.

Armed with the notion of weak solutions and an appropriate entropy condition we can define mathematically a unique solution to the system of conservation laws that is the physically correct inviscid limit. The entropy condition is very important, since it turns out that numerical solutions may approximate nonphysical weak solutions.

1.3. THE SCOPE OF THIS THESIS

The main objective of this thesis is to develop a reliable numerical method to approximate the ZND solution of the one-dimensional reactive Euler equations. As remarked before, the reactive Euler equations are a system of hyperbolic conservation laws with source terms. Many methods have been derived for one-dimensional hyperbolic conservation laws, for example finite difference methods, streamline diffusion finite element methods [42], spectral viscosity methods [90], front tracking methods [13, 76, 77, 87] and Glimm's method [79, 80]. In this thesis only finite difference methods will be considered.

When solving the reactive Euler equations numerically, we encounter problems which are absent in nonreacting flows. Apart from the increase in the number of equations, one of the main numerical problems occurs in the treatment of the source terms (the chemical reactions). Modern numerical methods, based on the solution of local Riemann problems, have been very successful in nonreacting fluid dynamics (see Chapter 4 and 5). However, generalizing these methods to reacting flows is very complicated since it involves the solution of local Riemann problems for conservation laws with source terms [3]. The problem is even more complicated by the kind of reactive flow problems posed by detonation waves. For detonation waves, the time scale associated to the flow is several orders of magnitude larger than typical time scales of the chemical reactions, in general. Consequently the source terms lead to a stiff system of equations.

A natural and relatively easy way to solve conservation laws with source terms is a *time splitting method* [84]. At each time level in a time splitting methods we alternate between solving the homogeneous conservation law without source term (i.e. the fluid dynamics) and solving the conservation laws without convection (i.e. the chemistry), giving a system of (stiff) ordinary differential equations (see Chapter 6). Time splitting methods are especially interesting for detonation waves, since we can deal with the fluid dynamics and the chemistry in a different way. By decomposing the problem we can use a high quality method for the homogeneous conservation laws and for the (stiff) ordinary differential equations.

In this thesis, we are rather interested in the global behaviour of the detonation wave and not so much in physical details. Moreover, since we consider a regime where detonation combustion occurs on length scales that are much smaller than the characteristic geometrical length scale, the detonation wave can be interpreted as a gas dynamic discontinuity of zero width. We like to develop a detonation capturing method, which produces a sharp representation of the detonation wave at the correct location. To this end, we study and develop numerical methods for space grids and time steps appropriate for the fluid dynamics but not for the rapid chemical reaction. Moreover, if the global behaviour is correct, adaptive refinement of the spatial mesh may be used to describe any details of the detonation wave [10]. However, in this thesis we do not pay any attention to adaptive refinement and restrict ourselves to the global behaviour of detonation waves.

The goal of this thesis is to study some important aspects of simulating the global behaviour of detonation waves numerically. Therefore, we may like to study the following

(i) numerical methods for homogeneous hyperbolic conservation laws;

(ii) numerical methods for (stiff) ordinary differential equations;

(iii) numerical methods based on time splitting;

(iv) capturing one-dimensional detonation waves for stiff combustion chemistry.

We briefly comment the points above.

(i) Numerical methods for homogeneous hyperbolic conservation laws.
When solving a hyperbolic conservation law numerically, we expect a finite difference discretization to be inappropriate near discontinuities (where the differential equation does not hold). Indeed, if we compute discontinuous solutions by standard methods for smooth solutions, we typically obtain poor numerical results (see Chapter 4). If we use a first order method, the results turn out to be smeared in regions near the discontinuities. On the other hand, standard second order methods introduce dispersive effects around the discontinuity leading to large oscillations.

Since the PDEs make sense away from discontinuities, one possible approach is to combine a standard finite difference method in smooth regions with some explicit procedure for tracking the location of the discontinuities. This approach is usually called front tracking [13, 76, 77, 87]. In one space dimension it is often a viable approach. In more space dimensions the discontinuities typically lie along curves (in 2D) or surfaces (in 3D) and in realistic problems there may be many surfaces that interact in complicated ways as time evolves. Although front tracking is still possible, it becomes much more complicated and will not be discussed in this thesis.

Ideally, we would like to have a numerical method which will produce sharp approximations of discontinuous solutions automatically, without explicit tracking. Methods that attempt to do this are called *shock capturing methods.* An example of a shock capturing method is the *high resolution method* [5, 28, 32, 34, 56]. The main goal in point (i) is to develop and study high resolution methods and to elaborate many theoretical aspects (see Chapter 4 and Chapter 5).

(ii) Numerical methods for (stiff) ordinary differential equations.
In this thesis we do not want to go in details of the chemistry involved, so we do not pay attention to solving stiff ODEs numerically. As noted before, in general the time scale of the chemical reaction is very small and an explicit numerical method would imply a severe time step restriction. Therefore, we should solve the ODE by an implicit numerical method. On the other hand, since we are not interested in chemical details, the ODE may often be simplified such that it can be solved exactly. In the latter case we do not need a numerical method at all and point (ii) may be omitted.

(iii) Numerical methods based on time splitting.
In splitting methods we alternatingly solve the fluid dynamics and the chemistry in each time step. At first glance it may appear to be less satisfactory than a standard method, since in practice the fluid dynamics and chemistry are strongly coupled. Therefore, the main question is how well the splitting method approximates the exact solution at each time level. The main goal of point (iii) is to develop a splitting method which is easy to implement and has a sufficient order of accuracy. Furthermore, we like to study how splitting methods may be used in developing numerical methods for hyperbolic conservation laws with source terms (see Chapter 6).

(iv) Capturing one-dimensional detonation waves for stiff combustion chemistry.
Obviously this is the main objective of this thesis and therefore the most important aspect. We like to develop and analyse a numerical method with the following properties:

- at least second order accuracy in smooth parts of the flow;
- a sharp resolution of the discontinuities without excessive smearing or oscillations;
- numerically stable with the usual time step and mesh width restrictions;
- approximating the physically relevant weak solution.

Similar to nonreacting flows, methods with the accuracy and resolution properties as above are called high resolution methods. The first three properties are fulfilled relatively easy. The first two are fulfilled by using a second order time splitting method in combination with a high resolution method for the homogeneous conservation law and a second order method for the ODE. Some typical results are given in Figure 1.4. If the chemistry is solved by an appropriate stiff ODE solver, then the third property is also fulfilled. The difficulties entirely arise from satisfying the last property. For fast reactions the numerical detonation wave may be completely wrong, since the reaction takes place at the wrong location [16, 58]. In fact the numerical solution approximates a wrong (nonphysical) weak solution. The computed solution consists of a fake detonation wave, followed by a slower propagating fluid dynamical shock wave (see the wrong solution in Figure 1.4).

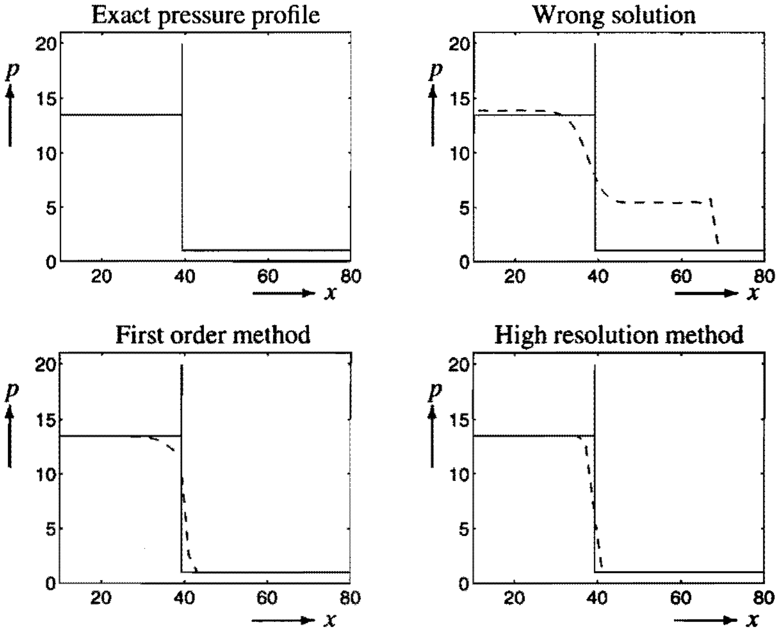

Figure 1.4. The exact detonation wave predicted by the ZND model (solid line) and the numerical detonation wave (dashed line). The numerical results are obtained with methods described in Chapter 8.

We emphasize that we are interested in capturing detonation waves on coarse meshes. Hence, we like to develop a method producing the correct detonation speed, even if the reaction zone merely occupies a tiny fraction of the numerical mesh width. Certainly, in this situation the inner structure of the detonations is not represented and what we should demand is that the method produces a sharp representation of the detonation wave at the correct location (see Chapter 7 and Chapter 8).

1.4. THE CONTENTS OF THIS THESIS

In this section we briefly outline the contents of this thesis.

In Chapter 2 we present the general system of equations modelling a one-dimensional reacting gas flow and describe the assumptions made to obtain the reactive Euler equations. Furthermore we derive the reactive Rankine-Hugoniot equations linking the upstream and downstream conditions of travelling combustion waves. The Rankine-Hugoniot equations play an important role in the mathematical formulation of the ZND model. The ZND model assumes that the detonation wave consists of a leading shock wave producing a flow with a temperature high enough to initiate a chemical reaction. This leading shock wave is followed by a reaction zone propagating with the same speed (see Figure 1.1). For the ZND model the only relevant detonation waves turn out to be characterized by Jouguet's rule, which says that the flow in a front attached frame of reference is supersonic in the unburnt gas ahead of the wave and subsonic or sonic in the burnt gas behind the wave. In the former case, one speaks of a strong or overdriven detonation wave, while a wave with sonic outflow is called a Chapman-Jouguet detonation wave. Hence, Jouguet's rule is used to characterize the physically relevant weak solution (see Section 1.2).

Chapter 3 has a more general setting. We introduce one-dimensional hyperbolic conservation laws and define weak solutions. These weak solutions are not unique in general and the entropy condition is presented in order to define the unique physically relevant weak solution. Finally, the Riemann problem is introduced, i.e. a hyperbolic conservation law with initial data consisting of two constant states. The Riemann problem plays an important role in many numerical methods and is therefore studied in detail in this chapter. We obtain the analytical solution of the Riemann problem for the nonreactive Euler equations. This solution is used later on to solve the reactive Euler equations numerically.

In this thesis we solve the reactive Euler equations with a time splitting method. In this splitting method we alternatingly solve the fluid dynamics and the chemistry. Many important properties of the full problem are determined by the fluid dynamical part, which is described by a homogeneous conservation law. Therefore, we study numerical methods for homogeneous conservation laws in detail in Chapter 4 and Chapter 5. In Chapter 4 some basic numerical concepts are introduced, such as discrete conservation, discretization error, consistency, convergence and stability. Furthermore, we present the Lax-Wendroff theorem. This well-known theorem shows that if a sequence of approximations converges, then the limit is a weak solution. In this chapter we present some straightforward methods showing the typical behaviour of first order methods (smearing of the solution around the discontinuities) and standard second order methods (oscillations around the discontinuities). Furthermore, we present the numerical version of the entropy condition. This condition is used to ensure that the numerical solution converges to a weak solution satisfying the original entropy condition (and thus the physically relevant weak solution). Finally, we discuss the first order methods of Godunov and Roe for systems of homogeneous conservation laws.

In Chapter 5 a theory is presented, which, under certain assumptions, guarantees convergence of a method. It turns out that one class of convergent methods consists of monotone methods. However, monotone methods are only accurate of at most order one. The main objective of Chapter 5 is to develop high resolution methods for the homogeneous conservation law. High resolution methods are methods which are at least second order accurate for smooth solutions and yet resolve discontinuities quite well (without excessive smearing or oscillations). The main idea behind any high resolution method is to attempt to use a high order method, but to modify the method around discontinuities, thereby increasing the amount of numerical diffusion. In this chapter we consider two types of high resolution methods for scalar problems, namely flux limiter methods and slope limiter methods. The slope limiter methods are extended to nonlinear systems of equations.

In Chapter 6 we return to nonhomogeneous conservation laws. We discuss time splitting methods for hyperbolic conservation laws with source terms. Splitting methods approximate the exact solution at discrete time levels. We describe the general idea and present two well-known splitting methods. We show that these methods are first and second order accurate in time, respectively. Furthermore, by also using a spatial discretization we obtain numerical methods based on splitting methods. Finally, we discuss the local discretization error for those numerical methods and compute the order of accuracy.

As remarked before, for fast reactions (or large mesh widths) the numerical solution of the reactive Euler equations may approximate the wrong (nonphysical) weak solution. For studying numerical methods, the reactive Euler equations are often too complicated. Therefore, we introduce in Chapter 7 a simplified detonation model. This model relates to the reactive Euler equations in a similar way as Burgers' equation to the ordinary Euler equations. We observe the same difficulty of approximating nonphysical solutions for the simplified detonation model. For this model it is illustrated that nonphysical solutions are always weak detonation waves. This is used to obtain a simple criterion which ensures that, even for relatively large mesh widths, the numerical solution approximates the physically relevant weak solution. Furthermore, a high resolution method is developed for the simplified detonation model and the numerical results of this method clearly illustrate that all properties of point (iv) in Section 1.3 are satisfied in for this model problem.

Finally, in Chapter 8 we consider the numerical solution of the reactive Euler equations. We extend the criterion of Chapter 7 and present some numerical results illustrating the use of this criterion to exclude nonphysical weak solutions (i.e. weak detonation waves). Furthermore, we describe a high resolution method based on the second order splitting method of Chapter 6 and the slope limiter method of Chapter 5. We present results showing that all the properties, mentioned in Section 1.3, are fulfilled. Finally, for the sake of completeness we make some remarks on front tracking, since front tracking methods are often used to solve the numerical difficulty of approximating incorrect weak solutions.

2

THE ONE-DIMENSIONAL REACTIVE EULER EQUATIONS

In simulations of the flow of a reacting gas mixture, chemical reactions between the constituent gases need to be modelled together with the fluid dynamics. Problems of this form arise, for example, in combustion. The basic equations of combustion theory are the conservation equations for reacting gas flow together with chemical kinetics. These equations represent the conservation of mass, momentum and energy of the total mixture and the change of composition of the gas mixture due to reaction.

A considerable simplification of these equations is possible if we restrict ourselves to one-dimensional detonations. Since detonation waves are propagating with very high speeds, molecular diffusion, thermal conduction and viscous effects are usually unimportant, and therefore can be ignored. If effects of walls, heat sources and external forces are also ignored, we essentially obtain the Euler equations of gas dynamics, which are completed by the continuity equations for the various species. These latter equations include source terms describing the chemical reactions. The total system of equations is often referred to as the reactive Euler equations.

This chapter is organized as follows. In the first section we give the general system of equations modelling a one-dimensional reacting gas flow. In Section 2.2 we describe the assumptions made in order to derive the reactive Euler equations. Furthermore, a dimensionless formulation of the Euler equations is derived in Section 2.3. In Section 2.4 we obtain the well-known Rankine-Hugoniot equations. These equations relate the upstream and downstream conditions of travelling combustion waves. Finally, in Section 2.5 we consider the ZND model, which describes the travelling wave solution (and in particular the detonation wave solution) of the reactive Euler equations.

2.1. THE ONE-DIMENSIONAL CONSERVATION EQUATIONS FOR REACTING GAS FLOW

Consider a tube filled with a gas mixture consisting of N different chemical species, denoted by $\mathcal{M}_i$ ($i = 1, 2, \dots, N$), in which M chemical reactions take place. We assume the species to be continuously distributed in any control volume.

Suppose the gas mixture is uniformly distributed across the tube, so there is variation in one direction only; therefore, we can restrict ourselves to one space dimension. Further assume that a *combustion wave* is propagating in the positive x-direction. This combustion wave consists of a zone involving chemical reactions, heat conduction, mass diffusion and viscous effects (see Figure 2.1). Ahead of the combustion wave there is a mixture of reactants which are in equilibrium. In the combustion wave the gas is burning and all reactants are entirely converted into products such that at the end of the zone the mixture consists of products only. All quantities ahead of the combustion wave will be identified by the subscript u (the unburnt gas), while all quantities behind the wave are denoted by the subscript b (the burnt gas).

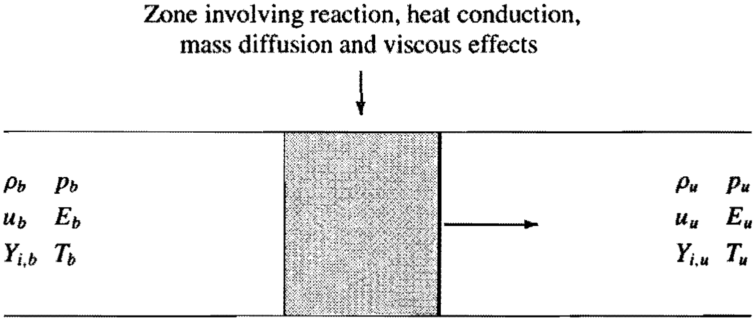

Figure 2.1. Schematic diagram of the combustion wave travelling through a tube.

For this kind of combustion problems, chemical reactions between the constituent gases need to be modelled together with the fluid dynamics. Therefore, we consider the *conservation equations for reacting gas flow*. These equations represent the conservation of mass, momentum and energy of the total mixture and the balance of mass for the various species. The latter equations include source terms which describe the chemical reactions that take place.

Below a brief description of the conservation equations for reacting gas flow is given. For a more detailed one, see e.g. [85, 96]. For *mass density* ρ, *mass-weighted average velocity* u, *mass fractions* Y_i, *diffusion velocities* U_i, *reaction rates* w_i, *stress* σ, *specific external forces* f_i, *specific total energy* E and *heat flux* q, the one-dimensional conservation equations for reactive gas flow may be written as follows [85, 92, 96]:
overall continuity (conservation of mass)

$$\frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho u) = 0; \tag{2.1a}$$

conservation of momentum

$$\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2) = \frac{\partial}{\partial x}\sigma + \rho \sum_{j=1}^{N} Y_j f_j; \tag{2.1b}$$

conservation of energy

$$\frac{\partial}{\partial t}(\rho E) + \frac{\partial}{\partial x}(\rho u E) = -\frac{\partial}{\partial x}q + \frac{\partial}{\partial x}(\sigma u) + \rho \sum_{j=1}^{N} Y_j f_j (u + U_j); \quad (2.1c)$$

and balance of mass for the various species

$$\frac{\partial}{\partial t}(\rho Y_i) + \frac{\partial}{\partial x}(\rho u Y_i) = -\frac{\partial}{\partial x}(\rho Y_i U_i) + w_i, \quad i = 1, 2, \dots, N. \quad (2.1d)$$

The reaction rate w_i is defined as the mass of species $\mathcal{M}_i$ created or destroyed by chemical reactions, per unit volume and per unit time. Since mass is neither created nor destroyed by chemical reactions but only converted from one species to another, it is obvious that

$$\sum_{j=1}^{N} w_j = 0. \quad (2.2)$$

The mass fraction Y_i of species $\mathcal{M}_i$ is given by $Y_i := \rho_i / \rho$, where ρ_i denotes the mass density of species $\mathcal{M}_i$. Clearly, the mass fractions satisfy

$$\sum_{j=1}^{N} Y_j = 1. \quad (2.3)$$

It is customary to write the flow velocity u_i of species $\mathcal{M}_i$ as

$$u_i = u + U_i,$$

where

$$u := \sum_{j=1}^{N} Y_j u_j.$$

In the equations above, u is the mass weighted flow velocity of the mixture, and U_i is called the diffusion velocity of species $\mathcal{M}_i$. Using the equations above it can easily be shown that

$$\sum_{j=1}^{N} Y_j U_j = 0. \quad (2.4)$$

The set of equations (2.1) has to be completed by models for U_i, w_i, σ, f_i and q. A brief description of these physical/chemical parameters follows in the next section.

System (2.1) describes a large class of combustion problems, but is generally quite complex. A considerable simplification of (2.1) is possible if we restrict ourselves to one-dimensional detonations. Therefore, in the next section several simplifying though realistic assumptions are made in order to have a more manageable form of (2.1).

2.2. DERIVATION OF THE ONE-DIMENSIONAL REACTIVE EULER EQUATIONS

In this section the reactive Euler equations will be derived. We start with the general system of equations (2.1) and list the assumptions to simplify (2.1).

The only external force which is usually of importance is *gravity*. In some low-speed combustion problems the gravity is not negligible. However, usually the influence of the gravity force is very small, and therefore neglected. We start with the following two assumptions.

A1 *The external forces f_i are negligible.*

A2 *The mass diffusion caused by pressure and thermal gradients is negligible.*

The mass diffusion caused by thermal gradients is known as the *Soret effect*. The *mole fraction* X_i of species $\mathcal{M}_i$ is defined by

$$X_i \;:=\; \frac{WY_i}{W_i}, \quad i = 1, 2, \ldots, N, \tag{2.5a}$$

where W_i is the *molecular mass* of species $\mathcal{M}_i$ and W is the *average molecular mass* of the gas mixture, given by

$$W \;:=\; \left(\sum_{j=1}^{N} Y_j / W_j \right)^{-1}. \tag{2.5b}$$

Under the assumptions A1 and A2 the diffusion velocities U_i can be determined from the *Stephan-Maxwell equations* [85, 96]

$$\frac{\partial}{\partial x} X_i \;=\; \sum_{j=1}^{N} \frac{X_i X_j}{D_{ij}} (U_j - U_i), \quad i = 1, 2, \ldots, N, \tag{2.6}$$

where D_{ij} is the so-called *binary diffusion coefficient* for species $\mathcal{M}_i$ and $\mathcal{M}_j$. This diffusion coefficient is defined as a proportionality constant which relates the net transport of species to the concentration gradient under conditions where the total net transport of molecules is zero. Equation (2.6) can be simplified considerably by assuming that the binary diffusion coefficients do not differ appreciably.

A3 *All binary diffusion coefficients are equal, i.e. $D_{ij} = D$ for all i and j.*

It can be shown that assumption A3 is always satisfied for binary mixtures [96]. From (2.3), (2.4), (2.5), (2.6) and assumption A3, *Fick's law of mass diffusion* follows, i.e.

$$Y_i U_i \;=\; -D \frac{\partial}{\partial x} Y_i.$$

For combustion of gases the *bulk viscosity* is negligible; therefore we assume the following.

A4 *The mixture behaves like a Newtonian fluid for which the bulk viscosity can be neglected.*

The stress of the gas mixture is the sum of normal and viscous stress. Under assumption A4, σ can be written as

$$\sigma = -p + \tau, \tag{2.7}$$

where p is the *hydrostatic pressure* and τ is the *viscous stress*. According to assumption A4, τ is defined by

$$\tau := \frac{4}{3}\mu \frac{\partial}{\partial x} u, \tag{2.8}$$

where μ is the *viscosity* of the gas mixture. In general μ is a function of the temperature, the pressure and the mole fractions. Using (2.7) and (2.8), the stress terms in the momentum equation and in the energy equation can be reduced to

$$\frac{\partial}{\partial x}\sigma = -\frac{\partial}{\partial x}p + \frac{4}{3}\frac{\partial}{\partial x}(\mu \frac{\partial}{\partial x}u), \tag{2.9a}$$

$$\frac{\partial}{\partial x}(\sigma u) = -\frac{\partial}{\partial x}(pu) + \frac{4}{3}\frac{\partial}{\partial x}(\mu u \frac{\partial}{\partial x}u). \tag{2.9b}$$

In many applications the viscous term is not an important mechanism for energy transport.

A5 *The viscous term* $\frac{\partial}{\partial x}(\mu u \frac{\partial}{\partial x}u)$ *in the energy equation is negligible.*

This assumption implies that (2.9b) simplifies to

$$\frac{\partial}{\partial x}(\sigma u) = -\frac{\partial}{\partial x}(pu). \tag{2.9c}$$

If temperature gradients give rise to diffusion velocities (the Soret effect), then concentration gradients must produce a heat flux. This reciprocal cross-transport process is known as the *Dufour effect*. Since we neglect the Soret effect (assumption A2), the Dufour effect is also neglected.

A6 *Heat transfer caused by radiation and concentration gradients (known as the Dufour effect) are negligible.*

This implies that the heat flux q of the gas mixture is given by

$$q = -\lambda \frac{\partial}{\partial x}T + \rho \sum_{j=1}^{N} h_j Y_j U_j,$$

where T is the *absolute temperature* of the gas mixture, λ is the *thermal conductivity* of the gas mixture and h_i is the *specific enthalpy* of species $\mathcal{M}_i$, which is defined by the caloric equation of state

$$h_i := h_i^0 + \int_{T_0}^{T} c_{p,i}(\xi)\mathrm{d}\xi, \quad i = 1, 2, \ldots, N. \tag{2.10}$$

The parameter h_i^0 is the *standard heat of formation* per unit mass for species $\mathcal{M}_i$ at a reference temperature T_0, and $c_{p,i} = c_{p,i}(T)$ is the *specific heat at constant pressure* for species $\mathcal{M}_i$.

After substituting the above models into (2.1) we arrive at the following set of equations for a reacting gas flow

$$\begin{aligned}
\frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho u) &= 0, && (2.11a)\\
\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) &= \frac{4}{3}\frac{\partial}{\partial x}(\mu\frac{\partial}{\partial x}u), && (2.11b)\\
\frac{\partial}{\partial t}(\rho E) + \frac{\partial}{\partial x}(\rho u E + pu) &= \frac{\partial}{\partial x}(\lambda\frac{\partial}{\partial x}T) + \frac{\partial}{\partial x}(\rho D\sum_{j=1}^{N} h_j\frac{\partial}{\partial x}Y_j), && (2.11c)\\
\frac{\partial}{\partial t}(\rho Y_i) + \frac{\partial}{\partial x}(\rho u Y_i) &= \frac{\partial}{\partial x}(\rho D\frac{\partial}{\partial x}Y_i) + w_i, \quad i = 1, 2, \dots, N. && (2.11d)
\end{aligned}$$

System (2.11) consists of $N + 3$ equations, however it follows from (2.2) and (2.3) that (2.11a) is the sum of (2.11d) for the individual species. Therefore, only $N + 2$ of these equations are independent. The independent variables are: ρ, u, $N - 1$ mass fractions Y_i, p, E and T; thus we have $N + 4$ unknowns. Therefore, two extra equations are required to complete the system. These equations are the *equation of state* and the *thermodynamic identity*.

A7 *The gas mixture behaves like an ideal gas.*

There are very few combustion problems where this ideal gas law is not effectively true and this assumption greatly reduces the complexity of the conservation equations. Under assumption A7 the equation of state becomes

$$p = \rho RT/W, \tag{2.12a}$$

where R is the *universal gas constant*. Furthermore, the thermodynamic identity for an ideal gas is given by

$$h := e + \frac{p}{\rho} = \sum_{j=1}^{N} Y_j h_j, \tag{2.12b}$$

where h is the *specific enthalpy of the mixture* and e the *specific internal energy*. The specific internal energy e is related to the specific total energy E by the relation

$$E = e + \tfrac{1}{2}u^2. \tag{2.13}$$

The term $u^2/2$ in (2.13) represents the specific kinetic energy of the gas mixture. Equation (2.12b) defines e as a function of T and Y_i through the caloric equation of state (2.10) and the equation of state (2.12a). The *stagnation enthalpy* H of the gas mixture is defined as

$$H := h + \tfrac{1}{2}u^2 = E + \frac{p}{\rho}. \tag{2.14}$$

Let the specific heat c_p at constant pressure and the specific heat c_v at constant volume for the gas mixture be defined as

$$\begin{aligned} c_p(T, Y_1, \ldots, Y_N) &:= \sum_{j=1}^{N} Y_j c_{p,j}(T), \\ c_v(T, Y_1, \ldots, Y_N) &:= \sum_{j=1}^{N} Y_j c_{v,j}(T), \end{aligned}$$

where $c_{v,i} = c_{v,i}(T)$ is the *specific heat at constant volume* for species $\mathcal{M}_i$. If a mixture behaves like an ideal gas, then we have

$$c_v(T, Y_1, \ldots, Y_N) = c_p(T, Y_1, \ldots, Y_N) - \frac{R}{W}. \tag{2.15}$$

A8 *All chemical species have constant and equal specific heats c_p at constant pressure.*

Assumption A8 is not essential because the important qualitative results to be obtained do not depend on this assumption. However, it enables us to elaborate (2.12b) in more detail. Using (2.10), (2.12b) and assumption A8, we derive

$$h = \sum_{j=1}^{N} Y_j h_j^0 + c_p(T - T_0). \tag{2.16}$$

Furthermore, we assume the following.

A9 *The gas is a binary mixture in which only one chemical reaction takes place.*

A10 *The chemical reaction is exothermic.*

Of course assumption A9 is often not true. However, the global chemical behaviour of a mixture can often be modelled quite adequately by a single reaction. Consider therefore, the one-step reaction in which a reactant $\mathcal{R}$ is converted into a product $\mathcal{P}$. Note that in this case $W = W_1 = W_2$ and let $Y_1 = Y$ denote the mass fraction of the reactant (and consequently $Y_2 = 1 - Y$). Since $N = 2$, we have $\sum_{j=1}^{2} h_j \partial Y_j / \partial x = (h_1 - h_2) \partial Y / \partial x$. Using assumptions A8 and A9, the *heat release* Q of the reaction per unit mass is given by [85, 96]

$$Q = h_1^0 - h_2^0. \tag{2.17}$$

Since we consider an exothermic reaction, $Q > 0$. In general Q depends on the initial and final composition of the gas mixture. Equation (2.16), together with assumptions A8 and A9, and (2.17) gives

$$h = QY + c_p T, \tag{2.18}$$

where, for convenience sake, we choose $h_b = c_p T_b$. It follows from the latter equation, (2.12a), (2.13) and (2.15) that we can also write the thermodynamic identity (2.12b) as

$$p = (\gamma - 1)\rho(E - \tfrac{1}{2}u^2 - QY), \tag{2.19}$$

where $\gamma = c_p/c_v$ is the *specific heat ratio.*

Until now we have not specified how the reaction rates w_i depend on the other variables. Since Y denotes the mass fraction of the reactant we write $w := w_1 = -w_2$ (see (2.2)). We assume that the one-step reaction is described by the *ignition model* and the *law of mass action* [92, 96]

$$w = \begin{cases} 0, & T < T_{ign}, \\ -k\rho Y, & T \geq T_{ign}, \end{cases} \tag{2.20a}$$

where T_{ign} is the *ignition temperature* and k is the *specific rate constant* for the reaction. If the temperature is below T_{ign} no chemical reaction takes place. If we omit the ignition model, then a slow chemical reaction will take place ahead of the combustion wave, which implies that the problem is ill-posed as $x \to \infty$. This phenomenon is known as the *cold-boundary difficulty* [96]. Hence, if the ignition temperature satisfies $T_u < T_{ign} \leq T_b$, then the introduction of this model ensures that the cold boundary difficulty does not appear. In practically all realistic cases $T_{ign} \gg T_u$. This can be thought of as being caused by a kinetic competition for radical species, which for low temperatures is completely on the side of the radical consuming reactions. In other words, at low temperatures the reactions are kinetically quenched, due to a lack of radicals. We assume that k satisfies *Arrhenius' law*

$$k = A(T)\exp(-\frac{E_a}{RT}), \tag{2.20b}$$

$$A(T) = BT^{\alpha}. \tag{2.20c}$$

The coefficients A and E_a in (2.20) are called the *frequency factor* and the *activation energy*, respectively. Further B is some positive fixed constant.

When we consider combustion waves propagating with high speeds, molecular diffusion, thermal conduction and viscous effects are usually unimportant transport mechanisms. Furthermore, for this kind of reactions the dependence of the frequency factor A on the temperature is rather unimportant. Thus, for such reactions the following assumptions seem reasonable.

A11 *The molecular diffusion, the thermal conduction and the viscous effects are negligible (* $D = 0$, $\lambda = 0$ *and* $\mu = 0$*).*

A12 *The temperature dependence of the frequency factor in the reaction rate is negligible (*$\alpha = 0$*, thus* $A(T) = A$*).*

Combining all results and assumptions above, we obtain from (2.11) the *reactive Euler equations* [26, 85]

$$\frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho u) = 0, \tag{2.21a}$$

$$\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \tag{2.21b}$$

$$\frac{\partial}{\partial t}(\rho E) + \frac{\partial}{\partial x}(\rho u E + p u) = 0, \tag{2.21c}$$

$$\frac{\partial}{\partial t}(\rho Y) + \frac{\partial}{\partial x}(\rho u Y) = w. \tag{2.21d}$$

The system of equations (2.12a), (2.19), (2.20) and (2.21) consists of 8 equations for the variables ρ, u, Y, p, E, T, k and w.

This section is concluded by introducing two quantities which will be useful later on. First, we define the *specific entropy* S for fixed Y by the first law of thermodynamics [18], i.e.

$$T\mathrm{d}S = \mathrm{d}h - \frac{1}{\rho}\mathrm{d}p. \tag{2.22}$$

The entropy variation from a reference state, indicated by the subscript $*$, is obtained from (2.12a), (2.15), (2.18) and (2.22) as

$$S - S_* = c_v \ln \frac{p/p_*}{(\rho/\rho_*)^\gamma}. \tag{2.23}$$

Finally, it will also be convenient to introduce the *frozen speed of sound* c, which for an ideal gas is given by

$$c = \sqrt{\frac{\gamma p}{\rho}}. \tag{2.24}$$

2.3. DIMENSIONLESS EQUATIONS

Often it is convenient to use *dimensionless variables*. In this section we like to derive a dimensionless formulation of the reactive Euler equations (2.21) together with the ideal gas law (2.12a), the thermodynamic identity (2.19) and the reaction rate (2.20). Let g_{ref} be a representative value of a generic variable g, then the associated dimensionless variable $\tilde{g}$ is defined by $\tilde{g} := g/g_{ref}$. The mass fraction Y is not scaled since it is dimensionless already.

Let ρ_{ref}, u_{ref} and x_{ref} be some given reference values of ρ, u and x, respectively. Suppose that x_{ref} is a typical length scale based on the convection of the flow, for instance the length of a finite tube (see Figure 2.1) in a laboratory and suppose that u_{ref} is of the same order as the speed of the combustion wave. We introduce the following dimensionless variables [26]

$$\begin{aligned} \tilde{E} &:= \frac{E}{u_{ref}^2}, & \tilde{p} &:= \frac{p}{\rho_{ref} u_{ref}^2}, & \tilde{\rho} &:= \frac{\rho}{\rho_{ref}}, \\ \tilde{t} &:= t\frac{u_{ref}}{x_{ref}}, & \tilde{u} &:= \frac{u}{u_{ref}}, & \tilde{w} &:= w\frac{t_{ref}}{\rho_{ref}}, \\ \tilde{x} &:= \frac{x}{x_{ref}}. \end{aligned} \tag{2.25}$$

In (2.25), $E_{ref} := u_{ref}^2$ is used as a characteristic total energy, since the kinetic energy forms an important contribution to the total energy. Furthermore, the choice of p_{ref} is motivated by Bernoulli's law, i.e. $p_{ref} := \rho_{ref} u_{ref}^2$ and $t_{ref} := x_{ref}/u_{ref}$ can be interpreted as a characteristic convection time scale. If the dimensionless variables (2.25) are used, then a straightforward substitution into (2.21) shows that the dimensionless reactive Euler equations are given by

$$\begin{aligned} \frac{\partial}{\partial \tilde{t}}(\tilde{\rho}) + \frac{\partial}{\partial \tilde{x}}(\tilde{\rho}\tilde{u}) &= 0, && (2.26a) \\ \frac{\partial}{\partial \tilde{t}}(\tilde{\rho}\tilde{u}) + \frac{\partial}{\partial \tilde{x}}(\tilde{\rho}\tilde{u}^2 + \tilde{p}) &= 0, && (2.26b) \\ \frac{\partial}{\partial \tilde{t}}(\tilde{\rho}\tilde{E}) + \frac{\partial}{\partial \tilde{x}}(\tilde{\rho}\tilde{u}\tilde{E} + \tilde{p}\tilde{u}) &= 0, && (2.26c) \\ \frac{\partial}{\partial \tilde{t}}(\tilde{\rho}Y) + \frac{\partial}{\partial \tilde{x}}(\tilde{\rho}\tilde{u}Y) &= \tilde{w}. && (2.26d) \end{aligned}$$

Let the dimensionless activation energy $\tilde{E}_a$, the dimensionless heat release of the chemical reaction $\tilde{Q}$ and the dimensionless temperature $\tilde{T}$ be introduced as

$$\tilde{E}_a := E_a \frac{\rho_{ref}}{W p_{ref}}, \qquad \tilde{Q} := Q \frac{\rho_{ref}}{p_{ref}}, \qquad \tilde{T} := T \frac{R\rho_{ref}}{W p_{ref}},$$

where the choice of T_{ref} and $(E_a)_{ref}$ is suggested by the equation of state for an ideal gas (2.12a) and the choice of Q_{ref} is motivated by the thermodynamic identity (2.19). The dimensionless equation of state and the dimensionless thermodynamic identity are then given by, respectively,

$$\begin{aligned} \tilde{p} &= \tilde{\rho}\tilde{T}, && (2.27) \\ \tilde{p} &= (\gamma - 1)\tilde{\rho}(\tilde{E} - \tfrac{1}{2}\tilde{u}^2 - \tilde{Q}Y). && (2.28) \end{aligned}$$

Let k_{ref} be the rate constant k at a typical reference temperature T_* (see (2.20b)). If the dimensionless rate constant $\tilde{k}$ is defined as $\tilde{k} := k/k_{ref}$, then

$$\tilde{k} = \exp(\tilde{E}_a(\frac{1}{\tilde{T}_*} - \frac{1}{\tilde{T}})), \qquad (2.29)$$

where $k_{ref} := A\exp(-E_a/(RT_*))$ and $\tilde{T}_* := T_* R\rho_{ref}/(W p_{ref})$. In Chapter 8 we choose $\tilde{T}_*$ equal to the von Neumann temperature, which is discussed in Section 2.5. Using $\tilde{k} := k/k_{ref}$, $\tilde{w} = w t_{ref}/\rho_{ref}$ and (2.20a), the dimensionless reaction rate can be given by

$$\tilde{w} = \begin{cases} 0, & \tilde{T} < \tilde{T}_{ign}, \\ -Da\, \tilde{k}\tilde{\rho}Y, & \tilde{T} \geq \tilde{T}_{ign}, \end{cases} \qquad (2.30)$$

where $\tilde{T}_{ign} := T_{ign}R\rho_{ref}/(Wp_{ref})$ is the dimensionless ignition temperature and the dimensionless constant Da is defined by

$$Da := t_{ref}k_{ref}. \tag{2.31}$$

The constant Da is referred to as the *Damköhler number* [85]. Recall that x_{ref} is a typical length scale based on the convection of the flow and u_{ref} is of the same order as the speed of the combustion wave. Hence, u_{ref}/k_{ref} can be interpreted as a characteristic reaction length [21, 85]. It follows from $t_{ref} = x_{ref}/u_{ref}$ and (2.31) that the Damköhler number is the ratio of the convection length scale and the reaction length scale [85]. Obviously, if Da is small the reaction occurs slowly relative to the specified time scale and if Da is large, the reaction zone is thin and the reaction occurs quickly relative to the specified time scale t_{ref}.

It will be useful to introduce the dimensionless specific enthalpy $\tilde{h}$, the dimensionless stagnation enthalpy $\tilde{H}$, the dimensionless specific entropy $\tilde{S}$ and the dimensionless frozen speed of sound $\tilde{c}$ as

$$\begin{aligned} \tilde{h} &:= h\frac{\rho_{ref}}{p_{ref}}, & \tilde{H} &:= H\frac{\rho_{ref}}{p_{ref}}, \\ \tilde{S} &:= S\frac{W}{R}, & \tilde{c} &:= c\sqrt{\frac{\rho_{ref}}{p_{ref}}}, \end{aligned}$$

where h_{ref}, H_{ref} and S_{ref} are suggested by (2.12b), (2.14) and (2.22), respectively. Furthermore, $c_{ref} = \sqrt{p_{ref}/\rho_{ref}}$ is motivated by the definition of the speed of sound (2.24). It follows directly from (2.15) and (2.18) that

$$\tilde{h} = \tilde{Q}Y + \frac{\gamma}{\gamma - 1}\tilde{T}. \tag{2.32}$$

From (2.14) it follows that

$$\tilde{H} = \tilde{E} + \frac{\tilde{p}}{\tilde{\rho}}.$$

Also it is easy to see that (2.22) gives

$$\tilde{T}\mathrm{d}\tilde{S} = \mathrm{d}\tilde{h} - \frac{1}{\tilde{\rho}}\mathrm{d}\tilde{p}.$$

Furthermore, the entropy variation from the reference state $\tilde{S}_* := WS_*/R$ is given by (see (2.23))

$$\tilde{S} - \tilde{S}_* = \frac{1}{\gamma - 1}c_v \ln\frac{\tilde{p}}{(\tilde{\rho})^\gamma}.$$

Finally, one can easily verify that (2.24) implies

$$\tilde{c} = \sqrt{\frac{\gamma\tilde{p}}{\tilde{\rho}}}.$$

In the remainder it is assumed that all variables are dimensionless, where, for shortness of notation, the tilde is suppressed.

2.4. THE REACTIVE RANKINE-HUGONIOT EQUATIONS

2.4.1. Derivation of the Reactive Rankine-Hugoniot Equations

We assume that a combustion wave is propagating with a constant velocity s in the positive x-direction of the tube (the direction of the unburnt gas, see Figure 2.1). It is clear that $s > u_u$ and $s > u_b$, since otherwise the wave will never pass the unburnt gas. The main goal of this section is to derive equations relating the state of the unburnt gas at the downstream side of the tube ($x = +\infty$) with the state of the completely burnt gas, at the upstream side of the tube ($x = -\infty$).

We make the following assumption [18, 26, 96].

A13 *The flow is steady with respect to a coordinate system moving with the combustion wave ($s > 0$ is constant).*

The reactive Rankine-Hugoniot equations can also be derived if the flow around the reaction zone is unsteady, i.e. s is not constant. However, for clarity of approach we make assumption A13. According to A13, it is natural to introduce a coordinate system which is stationary with respect to the wave. To this end, the variable ξ is introduced as

$$\xi(x,t) := x - st. \tag{2.33}$$

Using (2.33) and assumption A13 we write $g(x,t) = g(x - st) = g(\xi)$ for a generic variable g. Subsequently, (2.26) can be rewritten as a system of ordinary differential equations, i.e.

$$-s\frac{\mathrm{d}}{\mathrm{d}\xi}(\rho) + \frac{\mathrm{d}}{\mathrm{d}\xi}(\rho u) = 0, \tag{2.34a}$$

$$-s\frac{\mathrm{d}}{\mathrm{d}\xi}(\rho u) + \frac{\mathrm{d}}{\mathrm{d}\xi}(\rho u^2 + p) = 0, \tag{2.34b}$$

$$-s\frac{\mathrm{d}}{\mathrm{d}\xi}(\rho E) + \frac{\mathrm{d}}{\mathrm{d}\xi}(\rho u E + pu) = 0, \tag{2.34c}$$

$$-s\frac{\mathrm{d}}{\mathrm{d}\xi}(\rho Y) + \frac{\mathrm{d}}{\mathrm{d}\xi}(\rho u Y) = w. \tag{2.34d}$$

Now we are able to obtain the reactive Rankine-Hugoniot equations. After integrating (2.34a) from $\xi = -\infty$ to $\xi = +\infty$, we deduce

$$\rho_u(u_u - s) = \rho_b(u_b - s) =: -m, \tag{2.35a}$$

where $m > 0$ is the so-called dimensionless *mass flux*. Similarly, (2.34b) is integrated, which gives, using (2.35a),

$$\rho_u(u_u - s)^2 + p_u = \rho_b(u_b - s)^2 + p_b. \tag{2.35b}$$

Integrating (2.34c) implies $-mE_u+p_uu_u = -mE_b+p_bu_b$. It follows from (2.27), (2.28) and (2.32) that $h = E - u^2/2 + p/\rho$. Using this, together with (2.35a) and (2.35b), the latter equation can be rewritten as

$$h_u + \tfrac{1}{2}(u_u - s)^2 = h_b + \tfrac{1}{2}(u_b - s)^2. \tag{2.35c}$$

The equations (2.35) are called the *reactive Rankine-Hugoniot equations* [18, 26, 85, 96].

Integrating (2.34d) gives, using (2.35a), $Y_u = 1$ and $Y_b = 0$

$$-m = \int_{-\infty}^{\infty} w(\xi)\mathrm{d}\xi.$$

This provides the requirements

$$w_u = w_b = 0.$$

It follows from $T_{ign} > T_u$, $Y_b = 0$ and (2.30) that the latter equations are always satisfied. The equation of state (2.27)

$$\frac{p_u}{\rho_u T_u} = \frac{p_b}{\rho_b T_b} \tag{2.36}$$

and the thermodynamic relation (2.32)

$$h_b - \frac{\gamma}{\gamma - 1}T_b = h_u - Q - \frac{\gamma}{\gamma - 1}T_u = 0, \tag{2.37}$$

constitute further relations between the variables at $\xi = -\infty$ and $\xi = +\infty$.

The set of states at $\xi = -\infty$ (with fixed parameters at $\xi = +\infty$) for which equations (2.35a) and (2.35b) are satisfied is often referred to as the *Rayleigh line*. It follows directly from (2.35a) and (2.35b) that the Rayleigh line is given by

$$\frac{p_u - p_b}{(1/\rho_u) - (1/\rho_b)} = -m^2 < 0. \tag{2.38}$$

Using equation (2.35a) to express $u - s$ in terms of m and ρ in equation (2.35c) yields

$$h_b - h_u = \frac{1}{2}\left(\frac{1}{\rho_b} + \frac{1}{\rho_u}\right)(p_b - p_u),$$

where equation (2.38) is used to eliminate m^2. Since the velocities have been eliminated, the latter equation is a relationship among thermodynamic properties alone. In summary, equations (2.35), (2.36) and (2.37) complete the independent relations between the burnt and unburnt quantities. If all unburnt quantities (values of all quantities as $\xi \to +\infty$) are specified, then these 5 equations completely determine the variables u_b, ρ_b, p_b, T_b and h_b.

2.4.2. The Hugoniot Curve

Suppose that all quantities in the unburnt gas are specified. We want to describe the set of possible values for the quantities in the completely burnt gas, such that all relations given in the previous section are fulfilled. Therefore, the subscript b is suppressed. Since Y_b vanishes and Q is constant, (2.27) and (2.32) imply that the specific enthalpy h of the completely burnt gas can be expressed in terms of $1/\rho$ and p, i.e. $h = h(1/\rho, p)$. For shortness of notation, the *specific volume* v is introduced as

$$v := 1/\rho. \tag{2.39}$$

We define the *Hugoniot function* $\mathcal{H}$ by

$$\mathcal{H}(v, p) := h(v, p) - h_u - \frac{1}{2}(v + v_u)(p - p_u), \tag{2.40}$$

where (2.32) implies $h_u = Q + T_u\gamma/(\gamma - 1)$. It is straightforward to see that the equation $\mathcal{H} = 0$ defines a curve in (v, p)-space. This curve is called the *Hugoniot curve.* It follows from (2.27), (2.32) and $Y_u = 1$, $Y_b = 0$ that

$$h(v, p) - h_u = -Q + \frac{\gamma}{\gamma - 1}(pv - p_u v_u).$$

After substituting the latter equation into (2.40) we obtain

$$\mathcal{H}(v, p) = -Q + \frac{\gamma}{\gamma - 1}(pv - p_u v_u) - \frac{1}{2}(v + v_u)(p - p_u).$$

One can easily verify that the Hugoniot curve $\mathcal{H} = 0$ can be written as

$$p(v) = \frac{2Q - p_u v + \frac{\gamma+1}{\gamma-1} p_u v_u}{\frac{\gamma+1}{\gamma-1} v - v_u}. \tag{2.41}$$

Furthermore, the Rayleigh line (2.38) is rewritten as

$$p(v) = -m^2(v - v_u) + p_u. \tag{2.42}$$

The intersection of the Hugoniot curve with the Rayleigh line determines the final thermodynamic state, after m has been obtained from ρ_u, u_u and s for the particular experiment. The value u_b may then be calculated from (2.35a). In Figure 2.2 the Hugoniot curve (2.41) is drawn for $\gamma = 1.4$ and several values of the heat release Q.

Since the pressure p and the specific volume v should be positive, we require

$$\begin{gathered}\frac{\gamma - 1}{\gamma + 1} v_u < v < \frac{2Q}{p_u} + \frac{\gamma + 1}{\gamma - 1} v_u, \\ 0 < p < \infty,\end{gathered}$$

where the upper bound of v corresponds to the limit $p \to 0$.

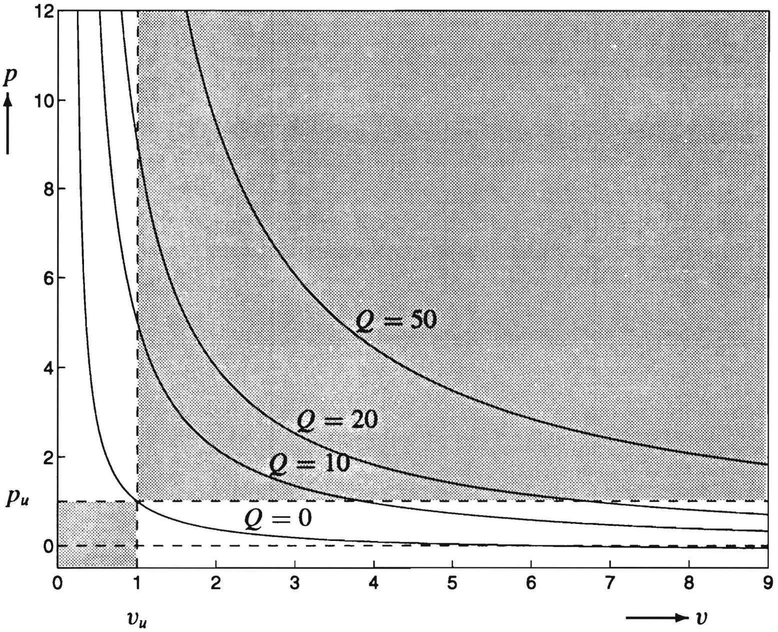

Figure 2.2. Hugoniot curves (2.41) with $v_u = p_u = 1$, $\gamma = 1.4$ and several values of Q.

2.4.3. Various Types of Processes.

The intersection between the Hugoniot curve (2.41) and the Rayleigh line (2.42) determines the conditions in the burnt gas. Since $m^2 > 0$, the slope of the Rayleigh line is negative and final states lying in the two shaded regions in Figure 2.2 are physically meaningless. Each Hugoniot curve is therefore divided into two distinct branches. The upper branch, which is given by (2.41) with

$$\begin{aligned} \frac{(\gamma-1)Q}{v_u} + p_u \;&\le\; p \;<\; \infty, \\ \frac{\gamma-1}{\gamma+1}v_u \;&<\; v \;\le\; v_u, \end{aligned}$$

is called the *detonation branch.* The lower branch, which is given by (2.41) with

$$\begin{aligned} 0 \;<\; &p \;\le\; p_u, \\ v_u + \frac{Q(\gamma-1)}{\gamma p_u} \;\le\; &v \;<\; \frac{2Q}{p_u} + \frac{\gamma+1}{\gamma-1}v_u, \end{aligned}$$

is called the *deflagration branch.* The only acceptable values for v_b and p_b are values such that the point (v_b,p_b) lies on one of these two branches. Combustion waves are termed *detonation waves* or *deflagration waves* according to the branch of the Hugoniot curve upon which the point (v_b,p_b) falls. An understanding of the difference between detonation and deflagration waves is best obtained by contrasting the characteristics of each. Thus, in passing through a detonation wave the gas speeds up, and its pressure and density increase (i.e. $p_b > p_u$ and $\rho_b > \rho_u$). On the other hand, in going through a deflagration the gas is slowed down and expands, and its pressure decreases (i.e. $p_b < p_u$ and $\rho_b < \rho_u$).

It can be shown that there are at most two points of intersection between the Rayleigh line and the detonation branch of the Hugoniot curve. There is a unique slope of the Rayleigh line such that it is tangent to the detonation branch. This point of tangency (point B in Figure 2.3) separates the detonation branch in two parts and is called the *upper Chapman-Jouguet point* (upper CJ point). Any straight line through (v_u, p_u) with a slope less than that of the line through B intersects the Hugoniot curve in two points (the steepest dashed line in Figure 2.3). Depending on the final conditions of the detonation we can distinguish between three different processes [18, 26, 85].

(i) Detonation waves with final conditions on the curve AB are called *strong detonations* .

(ii) Detonation waves with final conditions at point B are called *Chapman-Jouguet detonations* (CJ detonations).

(iii) Detonation waves with final conditions on the curve BC are called *weak detonations.*

There is also a unique slope of the Rayleigh line, such that it is tangent to the deflagration branch of the Hugoniot curve. This point of tangency (point E in Figure 2.3) separates the deflagration branch in two parts and is called the *lower Chapman-Jouguet point* (lower CJ point). Similarly to detonation waves we can distinguish between three different deflagration waves [18, 26, 85].

(i) Deflagration waves with final conditions on the curve DE are called *weak deflagrations.*

(ii) Deflagration waves with final conditions at point E are called *Chapman-Jouguet deflagrations* (CJ deflagrations).

(iii) Deflagration waves with final conditions on the curve EF are called *strong deflagrations.*

In the remainder we restrict ourselves to detonation waves. Next we elaborate the detonation processes in more detail [26, 85]. Let quantities in the completely burnt gas behind a strong, Chapman-Jouguet or weak detonation be denoted by the subscripts st, CJ or we, respectively. It follows from (2.41) and (2.42) that the mass flux $m_{CJ} > 0$, is given by

$$m_{CJ}^2 = \gamma \frac{p_u}{v_u} + (\gamma^2 - 1)\frac{Q}{v_u^2}\left\{1 + \sqrt{1 + \frac{2\gamma p_u v_u}{(\gamma^2 - 1)Q}}\right\}. \tag{2.43}$$

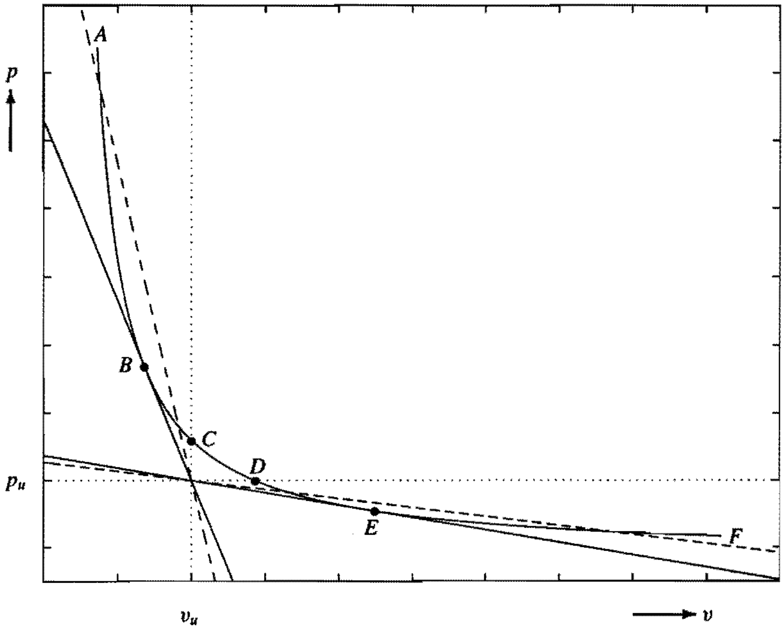

Figure 2.3. Different sections of the Hugoniot curve (2.41).

For all $m < m_{CJ}$ there will be no detonation. If $m = m_{CJ}$, then there will be a CJ detonation with

$$p_{CJ} = \frac{m^2 v_u + p_u}{\gamma + 1}, \tag{2.44a}$$

$$v_{CJ} = \frac{\gamma(m^2 v_u + p_u)}{m^2(\gamma + 1)}, \tag{2.44b}$$

$$u_{CJ} = s_{CJ} - v_{CJ} m. \tag{2.44c}$$

If $m > m_{CJ}$, then there will be a detonation with

$$p_{st} = \frac{m^2 v_u + p_u}{\gamma + 1} + \frac{1}{\gamma + 1}\sqrt{(m^2 v_u - \gamma p_u)^2 - 2(\gamma^2 - 1)m^2 Q}, \tag{2.45a}$$

$$v_{st} = \frac{\gamma(m^2 v_u + p_u)}{m^2(\gamma + 1)} - \frac{1}{m^2(\gamma + 1)}\sqrt{(m^2 v_u - \gamma p_u)^2 - 2(\gamma^2 - 1)m^2 Q}, \tag{2.45b}$$

$$u_{st} = s_{st} - v_{st} m, \tag{2.45c}$$

in case of a strong detonation and

$$p_{we} = \frac{m^2 v_u + p_u}{\gamma + 1} - \frac{1}{\gamma + 1}\sqrt{(m^2 v_u - \gamma p_u)^2 - 2(\gamma^2 - 1)m^2 Q}, \tag{2.46a}$$

$$v_{we} = \frac{\gamma(m^2 v_u + p_u)}{m^2(\gamma+1)} + \frac{1}{m^2(\gamma+1)}\sqrt{(m^2 v_u - \gamma p_u)^2 - 2(\gamma^2-1)m^2 Q}, \quad (2.46b)$$

$$u_{we} = s_{we} - v_{we} m, \quad (2.46c)$$

in case of a weak detonation.

Finally, we present, without a proof, some characteristic properties by which we can distinguish between the various detonation waves. These properties are referred to as Jouguet's rule [18]. Let c_u denote the speed of sound ahead of the reaction front and let c_b denote the speed of sound behind the reaction front.

Jouguet's rule:
The gas flow relative to the reaction front is

supersonic ahead of a detonation front (i.e. $s - u_u > c_u$),

subsonic behind a strong detonation front (i.e. $0 < s - u_b < c_b$),

sonic behind a Chapman-Jouguet detonation front (i.e. $s - u_b = c_b$),

supersonic behind a weak detonation front (i.e. $s - u_b > c_b$).

2.5. THE ZND MODEL FOR DETONATION WAVES

The Rankine-Hugoniot equations give no insight into the internal structure of detonation waves. Independently from each other, Zel'dovich, von Neumann and Döring developed a model which explains the internal structure of detonation waves, the so-called *ZND model* [18, 26, 85]. Apart from the assumptions A1-A13, the ZND model also presumes the following.

A14 *A detonation wave travelling with constant speed s has the internal structure of an ordinary (nonreacting) precursor fluid dynamical shock wave followed by a deflagration wave.*

A15 *The reaction rate is zero ahead of the shock and finite behind the shock.*

The shock wave is assumed to be much thinner than the zone of chemical reaction. Thus the shock can be considered as a discontinuous jump. This assumption is physically reasonable, since a few collisions in the material will establish a thermal equilibrium behind the shock, but many collisions are required for creating enough energy to initiate the chemical reaction [26, 85]. Hence, due to a nonreacting shock wave the temperature of the unburnt gas T_u jumps to a value larger than T_{ign} and a reaction is initiated.

A16 *The material emerging from the precursor fluid dynamical shock wave is assumed to be in thermochemical equilibrium and thus described by a thermodynamic equation of state.*

Hence, for a constant composition of the gas mixture, the thermodynamic equations hold. We still assume that effects of mass diffusion, thermal conductivity and viscosity are negligible (assumption A11) and that the flow is steady with respect to a coordinate system moving with the detonation wave (assumption A12). Assumption A16 implies that the Rankine-Hugoniot equations (2.35) should hold between any constant state ahead of the shock and any interior point of the reaction zone behind the shock. The Hugoniot curve now depends on the extent of the chemical reaction (reactant mass fraction Y), which varies continuously from 1 to 0, giving the generalization of (2.41)

$$p(v, Y) = \frac{2Q(1-Y) - p_u v + \frac{\gamma+1}{\gamma-1} p_u v_u}{\frac{\gamma+1}{\gamma-1} v - v_u}. \tag{2.47}$$

The Hugoniot curves (2.47) for different values of Y are drawn in Figure 2.4. The curve with $Y = 0$ corresponds to the states where the reaction is completed and all heat is released (see Figure 2.2).

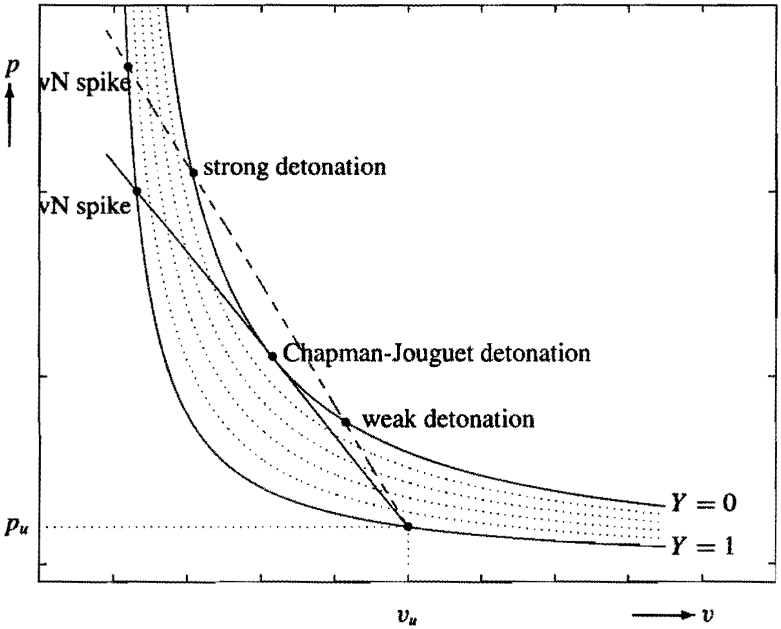

Figure 2.4. The Hugoniot curves (2.47) corresponding to the ZND theory.

The single variable Y completely defines the state as the state point moves down the Rayleigh line. First, due to a nonreacting shock wave the pressure and density (and temperature) jump to a higher value on the Hugoniot curve $p(\cdot, 1)$, called the *von Neumann spike* (vN spike) [26, 85]. The von Neumann spike is the state immediately behind the nonreacting shock wave and would be the final state if no chemical reaction takes place.

It is clear that the temperature at the von Neumann spike must lie above the ignition temperature. As the reaction proceeds, the state point moves down the Rayleigh line (pressure and density decrease) until the reaction is completed and the final state on the Hugoniot curve $p(\cdot, 0)$ is reached. At each point on the Rayleigh line between the von Neumann spike and the final state there is a unique Y determined from the Rankine-Hugoniot equations. The corresponding values for the pressure p and specific volume v can be obtained from the Rankine-Hugoniot equations and the value of Y. It can easily be verified that for a CJ or strong detonation

$$p(Y) = \frac{m^2 v_u + p_u}{\gamma + 1} + \frac{1}{\gamma + 1}\beta(Y), \tag{2.48a}$$

$$v(Y) = \frac{\gamma(m^2 v_u + p_u)}{m^2(\gamma + 1)} - \frac{1}{m^2(\gamma + 1)}\beta(Y), \tag{2.48b}$$

$$u(Y) = s - v(Y)m, \tag{2.48c}$$

where, for shortness of notation, $\beta(Y)$ is introduced as

$$\beta(Y) := \sqrt{(m^2 v_u - \gamma p_u)^2 - 2(\gamma^2 - 1)m^2 Q(1 - Y)}. \tag{2.48d}$$

Note that $p(0)$, $v(0)$ and $u(0)$ correspond to the final states given by (2.44) or (2.45).

Suppose that at time $t = 0$ the precursor shock is located at $x = 0$. Hence, at time t the variable $\xi = x - st$ measures the distance between the point x and the precursor shock. Therefore, $g(\xi) = g_u$ for all $\xi > 0$ and all variables g. Still the dependence of Y on the distance ξ has to be determined. Note that all variables can be expressed in terms of Y, and, subsequently, also the reaction rate w. Since we consider a steady flow, equation (2.34) should hold. Using (2.34a) and (2.35a), equation (2.34d) implies that the mass fraction of the reactant Y is given by the following ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{d}\xi}Y(\xi) = -\frac{w(Y(\xi))}{m}, \quad \forall \xi < 0, \tag{2.49a}$$

$$Y(0) = 1, \tag{2.49b}$$

where $\xi = 0$ corresponds to the position of the precursor shock. Note that $T(Y(\xi)) \geq T_{ign}$ for all $\xi < 0$. Hence, $w(Y(\xi))$ is continuous for all $\xi < 0$. In general, (2.49) can not be solved exactly and the solution must be obtained numerically. If we have computed Y, then we can determine all other variables from (2.48).

Note that for the ZND model the final state is a strong or CJ detonation. There is no path from the von Neumann spike to a point on the Hugoniot curve with $Y = 0$, corresponding to a weak detonation. It is obvious from Figure 2.4 that there can also be a shockless steady state solution as the state point moves up the Rayleigh line from the initial point (v_u, p_u) to a weak detonation point. In the present context the reaction rate would have to be finite in the initial state, without a shock to start it (contrary to assumption A15). Therefore, we are only interested in techniques for describing strong or CJ detonations [18, 85].

The minimum speed for a detonation wave is the speed s_{CJ} of a CJ detonation [18, 26, 85]. It will be useful to define a quantity which measures the overdrive of a strong detonation. Therefore, let the *degree of overdrive* f be defined by [11]

$$f := (s/s_{CJ})^2, \tag{2.50}$$

from which it directly follows that $f \geq 1$. In practice the degree of overdrive f is not a known parameter of the problem. Instead of f, the state of one variable behind the detonation wave is given, for example the temperature of the burnt gas. However, in order to study the difference between CJ detonations and strong detonations in detail, it will be useful to consider f as a given parameter and compute the corresponding states in the completely burnt gas. Therefore, suppose that all states ahead of the detonation wave are known (the unburnt gas) and let the parameters Da, E_a, f, Q and γ be given. First, we compute m_{CJ} using (2.43). Subsequently, m_{CJ} and (2.35a) give the speed s_{CJ} of a CJ detonation. Using the degree of overdrive f we can compute the detonation speed as $s = s_{CJ}\sqrt{f}$. After computing m by (2.35a) and solving (2.49) the complete ZND solution is derived (see (2.48)).

Finally, it is convenient to introduce the *half-reaction length* $L_{1/2}$. The half-reaction length is the distance for half completion of the reaction starting from the front of the detonation wave [26]. Often $L_{1/2}$ is given and (2.30) is used to compute the corresponding Damköhler number Da [11, 26]. It is easy to see that (2.49) implies that $L_{1/2}$ is given by

$$L_{1/2} = -m \int_{1/2}^{1} \frac{1}{w(Y)} \mathrm{d}Y. \tag{2.51}$$

In general, the half-reaction length has to be computed by some numerical method, since it is not possible to solve the integral above exactly.

Example 2.1. As an example of the preceding theory we describe the ZND solution of the CJ detonation discussed in [10]. Suppose that the dimensionless preshock state is given by

$$p_u = 1, \quad \rho_u = 1, \quad u_u = 0.$$

Furthermore, we take the following parameter values

$$E_a = 14, \quad Q = 14, \quad f = 1, \quad \gamma = 1.4.$$

Finally, $L_{1/2} = 1$ and the corresponding Damköhler number $Da = 0.6488$. We choose a relatively small Da, since otherwise the reaction is very fast and the plot of the ZND profile is not very clarifying. It follows from (2.35a), (2.43) and (2.44) that the final state for the CJ detonation is given by

$$p_b = p_{CJ} = 12.756, \quad \rho_b = \rho_{CJ} = 1.6583, \quad u_b = u_{CJ} = 2.1602,$$

where the CJ detonation is propagating with speed $s = s_{CJ} = 5.4419$. In Figure 2.5 the steady ZND solution is drawn. The pressure reaches its maximum value right behind the precursor shock. As mentioned before this value is called the von Neumann spike, which in this particular case satisfies $p_{vN} = p(1) = 24.512$ (see (2.48a)). The von Neumann temperature is given by $T_{vN} = 5.0509$. Furthermore, $c_b = 3.2817$ and, subsequently, $s - u_b = c_b$, as predicted by Jouguet's rule.

The maximum of the temperature near the end of the reaction zone can be explained by the geometry of the isotherms near the CJ point (see Figure 2.6). The Rayleigh line is tangent to the Hugoniot curve at the CJ point. The isotherms are less steep but concave upward, so precisely one of them will be tangent to the Rayleigh line somewhere above the CJ point [21, 26]. If this point of tangency lies below the von Neumann spike, the steady state solution will have a maximum in the temperature. Finally, Figure 2.5 clearly shows that the reaction rate w is zero ahead of the shock and finite behind it (as assumed in A15). □

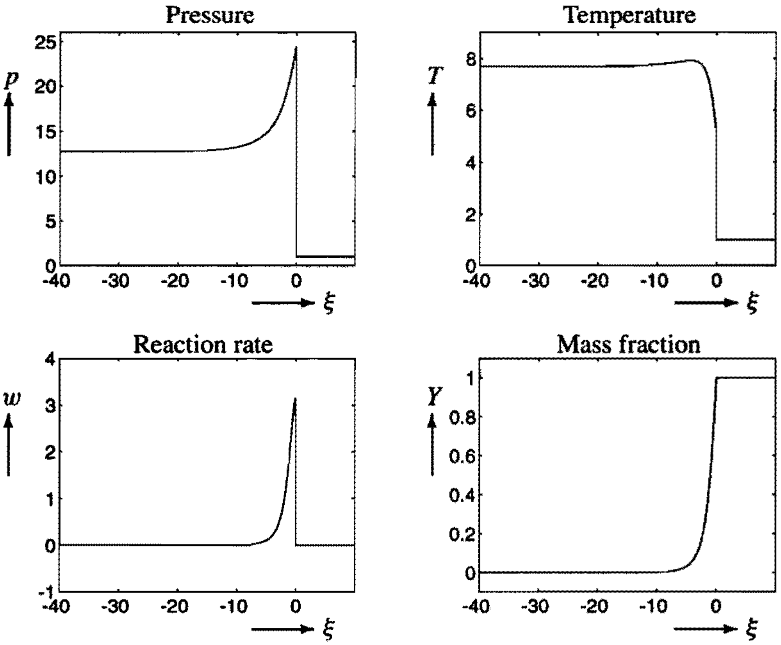

Figure 2.5. ZND solution of (2.26) with $E_a = 14$, $Q = 14$, $f = 1$, $\gamma = 1.4$ and $Da = 0.6488$ ($L_{1/2} = 1$).

Example 2.2. As a second example the ZND solution of a strong detonation is described [10, 11, 26]. Again, all quantities are made dimensionless with respect to the unburnt gas. Hence, the dimensionless preshock state is given by

$$p_u = 1, \quad \rho_u = 1, \quad u_u = 0.$$

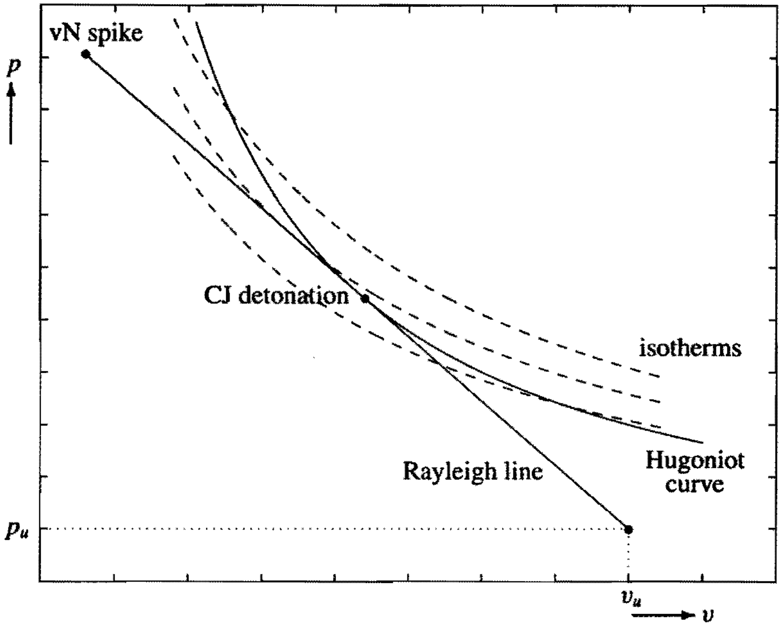

Figure 2.6. Explanation of temperature maximum in the ZND solution.

The dimensionless parameters are chosen as

$$E_a = 150, \quad Q = 50, \quad f = 1.8, \quad \gamma = 1.2.$$

Finally, $L_{1/2} = 1$ and the corresponding Damköhler number $Da = 0.6329 \cdot 10^{-1}$. It follows from (2.35a), (2.43) and (2.45) that the final state for the strong detonation is given by

$$p_b = p_{st} = 63.680, \quad \rho_b = \rho_{st} = 4.0158, \quad u_b = u_{st} = 6.8609,$$

where the strong detonation is propagating with a speed $s = s_{st} = 9.1359$. In Figure 2.7 the steady ZND solution is drawn. In this particular case the von Neumann spike satisfies $p_{vN} = p(1) = 75.786$ (see (2.48a)). For this particular example the von Neumann temperature is given by $T_{vN} = 7.8801$. Furthermore, $c_b = 3.0751$ and, subsequently, $s - u_b = 2.2750 < c_b$, as predicted by Jouguet's rule.

Figure 2.7 clearly shows that the gas requires considerable heating before the main exothermic reaction takes place. This is a consequence of the high activation energy and the characteristic exponential (Arrhenius-type) behaviour of the reaction rate (see (2.29)). In general, for detonative combustion processes the activation energy is high and the exponential term in the reaction rate guarantees the rate to be small even for temperatures relatively close to the burnt gas temperature. This important property of detonation waves is used later on in developing a reliable detonation capturing method. □

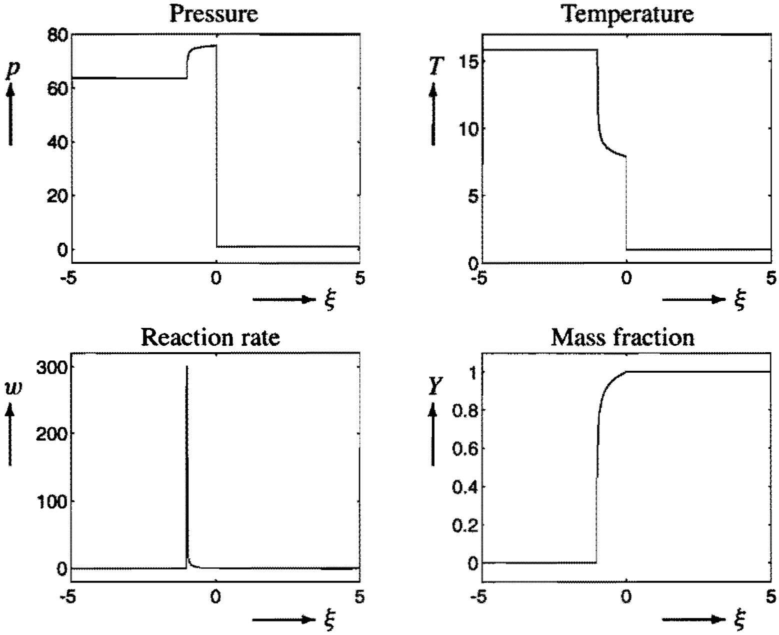

Figure 2.7. ZND solution of (2.26) with $E_a = 150$, $Q = 50$, $f = 1.8$, $\gamma = 1.2$ and $Da = 0.6329 \cdot 10^{-1}$ ($L_{1/2} = 1$).

3

ONE-DIMENSIONAL HYPERBOLIC CONSERVATION LAWS

Many (practical) problems in science and engineering involve conservation laws. A special class are the so-called hyperbolic conservation laws, which can be formulated as a system of first order partial differential equations. An important example are the (reactive) Euler equations of gas dynamics. Other examples arise in meteorology and astrophysics. In general it is not possible to derive exact solutions of these equations, and therefore, we have to devise and study numerical methods to approximate their solutions. This is done in the subsequent chapters. In this chapter we present some mathematical properties of one-dimensional hyperbolic conservation laws.

This chapter is organized as follows. The first section is of preliminary nature. We introduce one-dimensional hyperbolic conservation laws and present two important examples. In Section 3.2 weak solutions of hyperbolic conservation laws are defined. These weak solutions turn out to be nonunique and therefore an extra condition (i.e. the entropy condition) is necessary to characterize the physically relevant solution. In Section 3.3, the Riemann problem for homogeneous conservation laws is introduced and we present some characteristic features of the solution. This Riemann problem is important, because it forms the underlying physical model for the Godunov-type methods. Finally, in the last section we briefly describe the analytical solution of the Riemann problem for the nonreactive Euler equations. This solution is used later on to solve the reactive Euler equations numerically.

3.1. INTRODUCTION

In the sequel we consider one-dimensional conservation laws with source terms. It is assumed that the source terms are only dependent on the solution u. The general form of such conservation laws is

$$\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{x_L}^{x_R} \mathrm{u}(x,t)\mathrm{d}x = \mathrm{f}(\mathrm{u}(x_L,t)) - \mathrm{f}(\mathrm{u}(x_R,t)) + \int\limits_{x_L}^{x_R} \mathrm{q}(\mathrm{u}(x,t))\mathrm{d}x. \qquad (3.1)$$

Let the solution $\mathrm{u} : \mathbb{R} \times [0, \infty) \to \mathbb{R}^m$ and the flux function $\mathrm{f} : \mathbb{R}^m \to \mathbb{R}^m$ be continuously differentiable and the source term $\mathrm{q} : \mathbb{R}^m \to \mathbb{R}^m$ be continuous. Since (3.1) should hold for arbitrary x_L and x_R, it is clear that u satisfies

$$\frac{\partial}{\partial t}\mathrm{u}(x,t) + \frac{\partial}{\partial x}\mathrm{f}(\mathrm{u}(x,t)) = \mathrm{q}(\mathrm{u}(x,t)). \tag{3.2a}$$

This is the differential form of the conservation law. In order to obtain an initial value problem we add initial data to (3.2a), i.e.

$$\mathrm{u}(x,0) = \mathrm{u}^0(x), \quad \forall x \in \mathbb{R}. \tag{3.2b}$$

We rewrite (3.2a) as

$$\frac{\partial}{\partial t}\mathrm{u}(x,t) + A(\mathrm{u}(x,t))\frac{\partial}{\partial x}\mathrm{u}(x,t) = \mathrm{q}(\mathrm{u}(x,t)),$$

where $A(\mathrm{u})$ is the *Jacobian matrix* of $\mathrm{f}(\mathrm{u})$, defined by

$$A(\mathrm{u}) := \frac{\partial}{\partial \mathrm{u}}\mathrm{f}(\mathrm{u}). \tag{3.3}$$

For cases where the Jacobian matrix $A(\mathrm{u})$ is defined for all $\mathrm{u} \in \Omega \subset \mathbb{R}^m$ we can now define a *hyperbolic conservation law on* Ω as follows [5, 28, 49].

Definition 3.1. *Let a domain $\Omega \subset \mathbb{R}^m$ be given, such that $A(\mathrm{u})$ is well defined by* (3.3) *for all $\mathrm{u} \in \Omega$. The system* (3.2a) *is called a* hyperbolic conservation law on Ω *if there exists a real diagonal matrix $\Lambda(\mathrm{u})$ and a nonsingular real matrix $R(\mathrm{u})$ such that*

$$A(\mathrm{u})R(\mathrm{u}) = R(\mathrm{u})\Lambda(\mathrm{u}), \quad \forall \mathrm{u} \in \Omega.$$

Here $\Lambda(\mathrm{u}) = \mathrm{diag}(\lambda_1(\mathrm{u}), \lambda_2(\mathrm{u}), \ldots, \lambda_m(\mathrm{u}))$ is the diagonal matrix of the eigenvalues of $A(\mathrm{u})$ and $R(\mathrm{u}) = (\mathrm{r}^{(1)}(\mathrm{u}), \mathrm{r}^{(2)}(\mathrm{u}), \ldots, \mathrm{r}^{(m)}(\mathrm{u}))$ is the matrix of the corresponding right eigenvectors of $A(\mathrm{u})$. We assume that the eigenvalues are labelled in nondecreasing order, i.e. $\lambda_1(\mathrm{u}) \le \lambda_2(\mathrm{u}) \le \ldots \le \lambda_m(\mathrm{u})$.

We consider two examples. In the first example the *inviscid Burgers' equation* is introduced. This scalar conservation law is probably the simplest model that includes nonlinear effects of fluid dynamics.

Example 3.2 (Burgers' equation). Burgers' equation is by far the most famous scalar model problem for hyperbolic conservation laws. In this model equation the flux function f is given by $f(u) := \frac{1}{2}u^2$. Hence, the initial value problem for Burgers' equation is given by

$$\frac{\partial}{\partial t}u + u\frac{\partial}{\partial x}u = 0, \tag{3.4a}$$

$$u(x,0) = u^0(x). \tag{3.4b}$$

Consider curves in the $x - t$ plane satisfying the ordinary differential equations

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}x(t) &= u(x(t), t), && (3.5a) \\ x(0) &= x^0. && (3.5b) \end{aligned}$$

These curves are known as the *characteristics* of the equation. Along the characteristics u is constant, since

$$\frac{\mathrm{d}}{\mathrm{d}t}u(x(t), t) = \frac{\partial}{\partial t}u(x(t), t) + x'(t)\frac{\partial}{\partial x}u(x(t), t) = \frac{\partial}{\partial t}u + u\frac{\partial}{\partial x}u = 0.$$

Moreover, since u is constant on each characteristic, the slope $x'(\cdot)$ is constant by (3.5a) and so the characteristics are straight lines, determined by the initial data. □

The second example is a very important system of hyperbolic conservation laws, namely: the reactive Euler equations (2.26), which are described in Chapter 2.

Example 3.3 (The reactive Euler equations). Let the vector of conservative variables $\mathbf{u}$, the flux vector $\mathbf{f}(\mathbf{u})$ and the source vector $\mathbf{q}(\mathbf{u})$ be defined by, respectively,

$$\begin{aligned} \mathbf{u} &:= (\rho\,,\ \rho u\,,\ \rho E\,,\ \rho Y)^T, \\ \mathbf{f}(\mathbf{u}) &:= (\rho u\,,\ \rho u^2 + p\,,\ \rho u H\,,\ \rho u Y)^T, \qquad (3.6) \\ \mathbf{q}(\mathbf{u}) &:= (0\,,\ 0\,,\ 0\,,\ w)^T, \end{aligned}$$

then the reactive Euler equations can be written in the general form (3.2a). For the Jacobian matrix $A(\mathbf{u})$ as defined in (3.3), a straightforward computation shows that

$$A(\mathbf{u}) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{2}(\gamma - 3)u^2 & (3-\gamma)u & \gamma - 1 & -(\gamma - 1)Q \\ u(\frac{1}{2}(\gamma - 1)u^2 - H) & H - (\gamma - 1)u^2 & \gamma u & -(\gamma - 1)Qu \\ -uY & Y & 0 & u \end{pmatrix}. \qquad (3.7)$$

Furthermore the eigenvalues λ_k and right eigenvectors $\mathbf{r}^{(k)}$ $(k = 1, \ldots, 4)$ of $A(\mathbf{u})$ are given by

$$\lambda_1(\mathbf{u}) = u - c, \quad \lambda_2(\mathbf{u}) = u, \quad \lambda_3(\mathbf{u}) = u, \quad \lambda_4(\mathbf{u}) = u + c \qquad (3.8)$$

and

$$\begin{aligned} \mathbf{r}^{(1)}(\mathbf{u}) &= (1\,,\ u - c\,,\ H - uc\,,\ Y)^T, \\ \mathbf{r}^{(2)}(\mathbf{u}) &= (1\,,\ u\,,\ \tfrac{1}{2}u^2\,,\ 0)^T, \\ \mathbf{r}^{(3)}(\mathbf{u}) &= (0\,,\ 0\,,\ Q\,,\ 1)^T, \qquad (3.9) \\ \mathbf{r}^{(4)}(\mathbf{u}) &= (1\,,\ u + c\,,\ H + uc\,,\ Y)^T. \end{aligned}$$

Obviously, the reactive Euler equations are a hyperbolic system of conservation laws. □

3.2. WEAK SOLUTIONS

The assumption that the solution of (3.1) is continuously differentiable is too strong, since in practice discontinuous solutions u of (3.1) also occur [49, 79]. For instance detonation waves have a discontinuous structure including a strong leading shock wave (see e.g. Figure 2.5). This is the reason why *weak solutions* of the initial value problem (3.2) are interesting. These weak solutions are obtained from multiplying (3.2a) by an arbitrary test function $\varphi \in C_0^1(I\!R \times [0,\infty))$ (i.e. φ vanishes for $|x| + t$ large) and, subsequently, partially integrating this equation in space and time. This leads to the following definition.

Definition 3.4. *A bounded function* u *is called a* weak solution *of the conservation law* (3.2a) *with bounded initial data* (3.2b) *if*

$$\int_0^\infty \int_{-\infty}^\infty \left\{ \mathrm{u}(x,t)\frac{\partial}{\partial t}\varphi(x,t) + \mathrm{f}(\mathrm{u}(x,t))\frac{\partial}{\partial x}\varphi(x,t) \right\} \mathrm{d}x\mathrm{d}t = - \int_{-\infty}^{\infty} \mathrm{u}^0(x)\varphi(x,0)\mathrm{d}x - \int_0^\infty \int_{-\infty}^\infty \mathrm{q}(\mathrm{u}(x,t))\varphi(x,t)\mathrm{d}x\mathrm{d}t$$

for all functions $\varphi \in C_0^1(I\!R \times [0,\infty))$.

From now on by a solution of (3.2) a weak solution of (3.2) in the sense of Definition 3.4 is meant. It can be shown that a solution of (3.1) is always a weak solution of (3.2) [49, 56, 79]. Thus discontinuous solutions of (3.2) are also allowed. Let u have a discontinuity along a smooth curve Γ, i.e. u has well defined limits on both sides of Γ. Let Γ be given by $x = x(t)$, then the values $\mathrm{u}_L = \mathrm{u}(x(t)-0, t)$ and $\mathrm{u}_R = \mathrm{u}(x(t)+0, t)$ are well defined. Suppose the differential equation (3.2a) holds on both sides of Γ and also $x_L < x(t) < x_R$ for some fixed $t \geq 0$. Let $s = x'(t)$ be the speed of the discontinuity, then

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_{x_L}^{x_R} \mathrm{u}(x,t)\mathrm{d}x &= \frac{\mathrm{d}}{\mathrm{d}t}\{\int_{x_L}^{x(t)} \mathrm{u}(x,t)\mathrm{d}x + \int_{x(t)}^{x_R} \mathrm{u}(x,t)\mathrm{d}x\} \\
&= \int_{x_L}^{x(t)} \frac{\partial}{\partial t}\mathrm{u}(x,t)\mathrm{d}x + \int_{x(t)}^{x_R} \frac{\partial}{\partial t}\mathrm{u}(x,t)\mathrm{d}x + (\mathrm{u}_L - \mathrm{u}_R)s \\
&= \mathrm{f}(\mathrm{u}(x_L,t)) - \mathrm{f}(\mathrm{u}_L) + \mathrm{f}(\mathrm{u}_R) - \mathrm{f}(\mathrm{u}(x_R,t)) + (\mathrm{u}_L - \mathrm{u}_R)s \\
&\quad + \int_{x_L}^{x_R} \mathrm{q}(\mathrm{u}(x,t))\mathrm{d}x.
\end{aligned}$$

Thus, (3.1) shows that

$$s(\mathrm{u}_L - \mathrm{u}_R) = \mathrm{f}(\mathrm{u}_L) - \mathrm{f}(\mathrm{u}_R) \tag{3.10}$$

must hold at each point on Γ. Relation (3.10) is called the *jump condition.* In nonreactive gas dynamics the system of equations (3.10) is known as the *Rankine-Hugoniot equations.*

A difficulty is that the weak solutions of (3.2a) turn out to be nonunique for a given set of initial data. This is illustrated for Burgers' equations (3.4) in the following example.

Example 3.5. In this example we consider again the initial value problem for (3.4a) with initial data

$$u^0(x) = \begin{cases} -1, & x < 0, \\ 1, & x > 0. \end{cases} \tag{3.11}$$

The solution of problem (3.4a),(3.11) is not unique. This problem has a continuous solution given by (see Figure 3.1)

$$u(x,t) = \begin{cases} -1, & x < -t, \\ x/t, & -t < x < t, \\ 1, & x > t. \end{cases} \tag{3.12}$$

Moreover, for each $\alpha \geq -1$, this problem has discontinuous solutions u_α given by

$$u_\alpha(x,t) = \begin{cases} -1, & 2x < (-\alpha - 1)t, \\ -\alpha, & (-\alpha - 1)t < 2x < 0, \\ \alpha, & 0 < 2x < (\alpha + 1)t, \\ 1, & 2x > (\alpha + 1)t. \end{cases} \tag{3.13}$$

Thus problem (3.4a),(3.11) has a continuum of discontinuous solutions (see Figure 3.1). Note that all discontinuities satisfy (3.10) with $s = (u_L + u_R)/2$. □

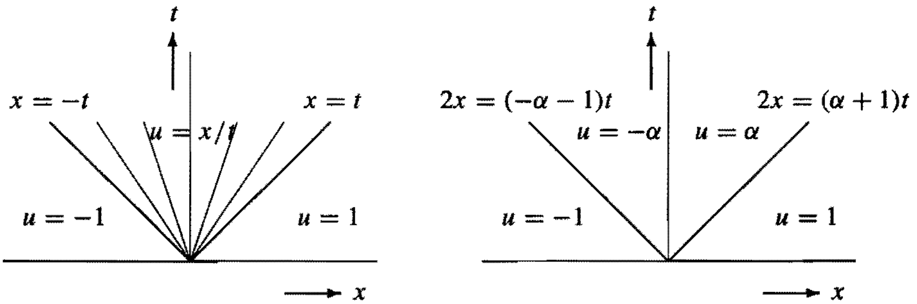

Figure 3.1. Continuous and discontinuous solutions of (3.4a),(3.11) in different sections of the $x - t$ space.

Conservation laws describe physical phenomena and so we must have some mechanism to characterize the "physically relevant" weak solution. Hyperbolic conservation laws often arise in models of physical processes which ignore the effects of various dissipative mechanisms. In more accurate models, these mechanisms introduce higher order derivatives in the equation multiplied by small coefficients called *viscosity coefficients*, as in gas dynamics. The consistency of the two models would then require solutions of the two sets of equations to be 'close' in some sense. In the limit, as the viscosity coefficient approaches zero, the solution of the higher order equations should converge to the solution of the first order conservation laws. We therefore remark that (3.2a) may be obtained, in the limit for $\mu \downarrow 0$, from the equation

$$\frac{\partial}{\partial t}\mathrm{u}_\mu(x,t) + \frac{\partial}{\partial x}\mathrm{f}(\mathrm{u}_\mu(x,t)) = \mu\frac{\partial^2}{\partial x^2}\mathrm{u}_\mu(x,t) + \mathrm{q}(\mathrm{u}_\mu(x,t)), \tag{3.14}$$

with μ the viscosity coefficient ($\mu > 0$). Hence the unique, *physically relevant weak solution* is, roughly speaking, defined as the stable limit of this vanishing viscosity mechanism [47, 49, 79]. Since the vanishing viscosity method requires studying more complicated systems of equations, we like to find other conditions which can be imposed directly on weak solutions of hyperbolic conservation laws in order to pick out the physically correct solution.

In nonreactive gas dynamics the entropy jumps to a higher value as the gas crosses a shock discontinuity. It follows from the second law of thermodynamics that the entropy never jumps to a lower value. This gives the extra condition, the so-called *entropy condition*, which picks out the correct weak solution in gas dynamics (see the Lax entropy condition in Section 3.3.4). For the ZND model for detonation waves, the relevant weak solution is characterized by Jouguet's rule, which says that the flow in a front attached frame of reference is supersonic ahead of the detonation wave and subsonic or sonic behind the detonation wave (see Section 2.4.3).

For general hyperbolic conservation laws we also impose an extra condition upon the solution, such that a physically relevant solution is obtained [47, 79]. As in gas dynamics this condition is called the entropy condition. For the sake of completeness we briefly describe the derivation of this entropy condition.

In order to find the general entropy condition it is useful to consider a twice continuously differentiable function $\eta : \mathbb{R}^m \to \mathbb{R}$. The function η is called *convex* if its Hessian (denoted by η_{uu}) is symmetric positive semidefinite. Thus, for a convex function η the following inequality holds (note that $\eta_{\mathrm{uu}}(\mathrm{v})$ is a $m \times m$-matrix)

$$(\eta_{\mathrm{uu}}(\mathrm{v})\mathrm{v})^T\mathrm{v} \geq 0, \quad \forall\, \mathrm{v} \in \mathbb{R}^m. \tag{3.15}$$

Next we define the *entropy function* and the *entropy flux*. These concepts are used to find the entropy condition.

Definition 3.6. *A twice continuously differentiable, convex function $\eta : \mathbb{R}^m \to \mathbb{R}$ is called an* entropy function *for the conservation law* (3.2a), *if there exists a continuously differentiable function $\psi : \mathbb{R}^m \to \mathbb{R}$, such that*

$$\nabla\psi(\mathrm{u})^T = \nabla\eta(\mathrm{u})^T\frac{\partial}{\partial \mathrm{u}}\mathrm{f}(\mathrm{u}), \quad \forall\, \mathrm{u}. \tag{3.16}$$

The function ψ is called an entropy flux.

Note that all convex functions serve as entropy functions in the scalar case. A straightforward computation shows that

$$\frac{\partial}{\partial t}\eta(\mathrm{u}(x,t)) + \frac{\partial}{\partial x}\psi(\mathrm{u}(x,t)) = \nabla\eta(\mathrm{u}(x,t))^T\mathrm{q}(\mathrm{u}(x,t))$$

holds, if u is a continuously differentiable solution of (3.2a).

The system of equations (3.16) has two unknowns, η and ψ. If the system has too many equations, then it may have no solution. If the Jacobian matrix $A(\mathrm{u})$ is symmetric for all u, i.e. $\partial f_i/\partial u_j = \partial f_j/\partial u_i$, then there exists a function $g : \mathbb{R}^m \to \mathbb{R}$, such that $\partial g/\partial u_l = f_l$. Now it is clear that η and ψ can be chosen as $\eta(\mathrm{u}) = \frac{1}{2}\sum_j u_j^2$ and $\psi(\mathrm{u}) = \sum_j u_j f_j - g(\mathrm{u})$. In [79] other examples are given for which nontrivial solutions of (3.16) exist.

For the solution of the viscous equation (3.14), we can associate the (small but) positive viscosity term with an entropy inequality. If it is assumed that the solution of the parabolic equation (3.14) is twice continuously differentiable, then equation (3.16) leads to

$$\frac{\partial}{\partial t}\eta(\mathrm{u}_\mu) + \frac{\partial}{\partial x}\psi(\mathrm{u}_\mu) = \mu\nabla\eta(\mathrm{u}_\mu)^T\frac{\partial^2}{\partial x^2}\mathrm{u}_\mu + \nabla\eta(\mathrm{u}_\mu)^T\mathrm{q}(\mathrm{u}_\mu). \tag{3.17}$$

Let $\varphi \in C_0^1(\mathbb{R}\times[0,\infty))$ be an arbitrary test function such that $\varphi(x,t) \geq 0$ for all $x \in \mathbb{R}$ and $t \in [0,\infty)$, and assume that the solution u_μ in (3.17) is bounded. Using (3.15) and (3.17) it is easy to see that

$$\begin{aligned}
&\int_0^\infty\int_{-\infty}^\infty \left\{\frac{\partial}{\partial t}\eta(\mathrm{u}_\mu) + \frac{\partial}{\partial x}\psi(\mathrm{u}_\mu)\right\}\varphi\,\mathrm{d}x\mathrm{d}t &=\\
&\int_0^\infty\int_{-\infty}^\infty \left\{\mu\nabla\eta(\mathrm{u}_\mu)^T\frac{\partial^2}{\partial x^2}\mathrm{u}_\mu + \nabla\eta(\mathrm{u}_\mu)^T\mathrm{q}(\mathrm{u}_\mu)\right\}\varphi\,\mathrm{d}x\mathrm{d}t &=\\
&\int_0^\infty\int_{-\infty}^\infty \left\{\mu\frac{\partial^2}{\partial x^2}\eta(\mathrm{u}_\mu) - \mu(\eta_{\mathrm{uu}}(\mathrm{u}_\mu)\frac{\partial}{\partial x}\mathrm{u}_\mu)^T\frac{\partial}{\partial x}\mathrm{u}_\mu + \nabla\eta(\mathrm{u}_\mu)^T\mathrm{q}(\mathrm{u}_\mu)\right\}\varphi\,\mathrm{d}x\mathrm{d}t &\leq\\
&\mu\int_0^\infty\int_{-\infty}^\infty \frac{\partial^2}{\partial x^2}\eta(\mathrm{u}_\mu)\varphi\mathrm{d}x\mathrm{d}t + \int_0^\infty\int_{-\infty}^\infty \nabla\eta(\mathrm{u}_\mu)^T\mathrm{q}(\mathrm{u}_\mu)\,\varphi\,\mathrm{d}x\mathrm{d}t.
\end{aligned}$$

Recall that the physically relevant solution u is defined as, roughly speaking, "$\mathrm{u} = \lim_{\mu\downarrow 0}\mathrm{u}_\mu$". Next an *entropy solution* is defined [47, 79].

Definition 3.7. *A weak solution* u *of* (3.2a) *is called an* entropy solution *if, for all convex entropy functions* η *and corresponding entropy fluxes* ψ*, the inequality*

$$\frac{\partial}{\partial t}\eta(\mathrm{u}(x,t)) + \frac{\partial}{\partial x}\psi(\mathrm{u}(x,t)) \leq \nabla\eta(\mathrm{u}(x,t))^T\mathrm{q}(\mathrm{u}(x,t)) \tag{3.18}$$

is satisfied in the weak sense (for all nonnegative test functions). The inequality (3.18) *is called the* entropy condition.

In a similar way as we derived (3.10) it can be shown that (3.18) is equivalent to the condition that

$$s(\eta(\mathbf{u}_L) - \eta(\mathbf{u}_R)) \leq \psi(\mathbf{u}_L) - \psi(\mathbf{u}_R) \tag{3.19}$$

holds at every discontinuity of a piecewise continuous solution $\mathbf{u}$. Hence this criterion is also often used as the definition of entropy solutions. For more details, see [49, 79].

Consider the scalar nonlinear conservation law, i.e. (3.2a) with $m = 1$. Suppose that an entropy function η and a corresponding entropy flux ψ are given by

$$\begin{aligned} \eta(u(x,t)) &= |u(x,t) - z|, \\ \psi(u(x,t)) &= \{f(u(x,t)) - f(z)\}\mathrm{sgn}(u(x,t) - z), \end{aligned}$$

where z is an arbitrary real number and $\mathrm{sgn}(x) = 1$ for $x > 0$ and $\mathrm{sgn}(x) = -1$ for $x < 0$. It has been shown by Krushkov that the entropy solution is unique and is characterized by these choices for the entropy functions and the entropy fluxes. Furthermore, this unique entropy stable solution is equal to the physically relevant solution [28, 47].

Therefore, by analogy with the scalar case, condition (3.18) or (3.19) is often imposed in order to identify the physically relevant solution. Moreover, to our knowledge, there is no detailed analysis of the entropy condition for general systems yet.

Example 3.8. In this example we consider again the solution of problem (3.4a),(3.11) and we want to derive the unique entropy solution (see Example 3.5). Therefore, we consider (3.19) with $f(u) = u^2/2$ and $\eta(u) = u^2$. Then (3.16) gives $\psi'(u) = 2u^2$ and hence $\psi(u) = 2u^3/3$. It follows from (3.10) that $s = (u_L + u_R)/2$. After substituting this into (3.19) we obtain that the inequality

$$-\tfrac{1}{6}(u_L - u_R)^3 \leq 0$$

should hold at every discontinuity of a piecewise continuous solution u. Hence, the only allowable discontinuities have $u_L > u_R$. After using this for the solutions u_α in (3.13), we find that $-1 \geq -\alpha \geq \alpha \geq 1$. Hence, there exists no α such that u_α is an entropy solution of (3.4a),(3.11). Therefore, the unique entropy solution of (3.4a),(3.11) is given in (3.12). □

3.3. THE RIEMANN PROBLEM FOR HOMOGENEOUS CONSERVATION LAWS

3.3.1. Preliminaries

In this section the *Riemann problem* is introduced. The Riemann problem is very important because it provides the underlying physical model of many numerical methods for hyperbolic conservation laws. For instance, the well-known *Godunov upwind methods* use the exact solution of the Riemann problem [40, 56].

Definition 3.9. *The* Riemann problem *for a homogeneous hyperbolic conservation law is the following initial value problem*

$$\frac{\partial}{\partial t}\mathbf{u}(x,t) + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}(x,t)) = \mathbf{0} \tag{3.20a}$$

with initial data

$$\mathbf{u}^0(x) = \begin{cases} \mathbf{u}_L, & x < 0, \\ \mathbf{u}_R, & x > 0, \end{cases} \tag{3.20b}$$

where $\mathbf{u}_L \in I\!R^m$ *and* $\mathbf{u}_R \in I\!R^m$ *are given constant states.*

Since (3.20a) is assumed to be a hyperbolic system, the Jacobian matrix $A(\mathbf{u})$ has m real eigenvalues $\lambda_1(\mathbf{u}), \ldots, \lambda_m(\mathbf{u})$ and m linearly independent right eigenvectors $\mathbf{r}^{(1)}(\mathbf{u}), \ldots, \mathbf{r}^{(m)}(\mathbf{u})$. For calculating the solution of the Riemann problem (3.20) the following concepts are introduced [49, 79].

Definition 3.10. *Consider the hyperbolic system* (3.20a). *Let* $\mathbf{r}^{(k)}(\mathbf{u})$ *be a right eigenvector of* $A(\mathbf{u})$ *and* $\lambda_k(\mathbf{u})$ *the corresponding eigenvalue. The eigenvector* $\mathbf{r}^{(k)}(\mathbf{u})$ *is called* genuinely nonlinear *if*

$$(\nabla\lambda_k(\mathbf{u}), \mathbf{r}^{(k)}(\mathbf{u})) \neq 0, \quad \forall\, \mathbf{u}. \tag{3.21}$$

The eigenvector $\mathbf{r}^{(k)}(\mathbf{u})$ *is called* linearly degenerate *if*

$$(\nabla\lambda_k(\mathbf{u}), \mathbf{r}^{(k)}(\mathbf{u})) = 0, \quad \forall\, \mathbf{u}. \tag{3.22}$$

Here $\nabla\lambda_k(\mathbf{u}) = (\frac{\partial}{\partial u_1}\lambda_k(\mathbf{u}), \ldots, \frac{\partial}{\partial u_m}\lambda_k(\mathbf{u}))^T$ and $(\cdot,\cdot)$ denotes the usual inner product in $I\!R^m$. It is assumed that $\mathbf{f}$ is twice continuously differentiable, so that $\nabla\lambda_k(\mathbf{u})$ exists for all k.

Next a theorem is given, which shows whether the eigenvectors belonging to the Euler equations are genuinely nonlinear or linearly degenerate. Its proof is a straightforward calculation.

Theorem 3.11. *Consider the nonreactive Euler equations, i.e.* (2.26) *with* $w = 0$. *Let the corresponding eigenvectors* $\mathbf{r}^{(k)}(\mathbf{u})$, *(*$k = 1, \ldots, 4$*) be given by* (3.9). *Then* $\mathbf{r}^{(k)}(\mathbf{u})$ *is linearly degenerate for* $k = 2, 3$ *and genuinely nonlinear for* $k = 1, 4$.

The solution of the Riemann problem for a nonlinear hyperbolic system is hard to find in general. But for certain pairs $(\mathbf{u}_L, \mathbf{u}_R)$ the solution of the Riemann problem can be derived easily. In [49, 79] it is proved that, if $\mathbf{u}_L$ and $\mathbf{u}_R$ are sufficiently close, the initial value problem (3.20) has a unique solution. Here we only present some characteristic features of the solution. In Section 3.4 we briefly describe the solution of the Riemann problem for the nonreactive Euler equations.

In the remainder of this chapter it is assumed that each eigenvector is either linearly degenerate or genuinely nonlinear.

Note that Theorem 3.11 ensures that this assumption is fulfilled for the Euler equations. The following theorem describes the general form of solutions of the Riemann problem.

Theorem 3.12. *Suppose that there exists a unique solution* u *of the Riemann problem* (3.20). *Then this solution can be written in the similarity form* $\mathrm{u}(x,t) =: \tilde{\mathrm{u}}(x/t) =: \tilde{\mathrm{u}}(\xi)$.

Proof. Define $\mathrm{u}_\alpha(x,t) := \mathrm{u}(\alpha x, \alpha t)$ with $\alpha > 0$. Then it is easily verified that $\mathrm{u}_\alpha(x,t)$ is also a solution of the Riemann problem. Hence, $\mathrm{u}(x,t) = \mathrm{u}(\alpha x, \alpha t)$ for all $\alpha > 0$, so $\mathrm{u}(x,t) = \tilde{\mathrm{u}}(x/t)$. □

In order to analyse the Riemann problem, the part of a solution associated with a single eigenvector is considered. Here different possibilities exist. If the eigenvector is linearly degenerate, then a *contact discontinuity* appears. In the genuinely nonlinear case there are two possibilities: first $\lambda(\mathrm{u}_L) < \lambda(\mathrm{u}_R)$ in which case a *rarefaction wave* is found, and secondly, $\lambda(\mathrm{u}_L) > \lambda(\mathrm{u}_R)$ in which case a *shock wave* is found.

To calculate the contact discontinuity or the rarefaction wave solution, integral curves in the *phase space* are considered. This phase space is simply the m-dimensional space containing all values of $\mathrm{u} = (u_1, u_2, \ldots, u_m)^T$. In general, starting at each point u_L there are m curves consisting of points u_R which can be connected to u_L by a rarefaction wave or contact discontinuity. These turn out to be subsets of the integral curves of the vector fields $\mathrm{r}^{(k)}$.

For each right eigenvector $\mathrm{r}^{(k)}$ we define an integral curve in the phase space such that it starts in some arbitrary given state u_L and has the property that the tangent to the curve at any point u lies in the direction $\mathrm{r}^{(k)}$. The existence of smooth curves of this form follows from smoothness of f and hyperbolicity, since $\mathrm{r}^{(k)}$ is then a smooth function of u.

3.3.2. Rarefaction Waves

In this section we briefly describe the rarefaction wave solution of the Riemann problem (3.20). For details we refer to [79]. Suppose that $\mathrm{r}^{(k)}(\mathrm{u})$ is a genuinely nonlinear eigenvector and $\lambda_k(\mathrm{u}_L) < \lambda_k(\mathrm{u}_R)$. Then $\mathrm{r}^{(k)}(\mathrm{u})$ can be normalized such that

$$(\nabla\lambda_k(\mathrm{u}), \mathrm{r}^{(k)}(\mathrm{u})) = 1, \quad \forall\, \mathrm{u}.$$

For an arbitrary state u_L and $k \in \{1, \ldots, m\}$, we consider the following initial value problem

$$\left\{ \begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\xi}\tilde{\mathrm{u}}(\xi) &= \mathrm{r}^{(k)}(\tilde{\mathrm{u}}(\xi)), \\ \tilde{\mathrm{u}}(0) &= \mathrm{u}_L. \end{aligned} \right. \tag{3.23}$$

Let ξ_R be chosen such that $\mathrm{u}_R = \tilde{\mathrm{u}}(\xi_R)$ is well defined and (3.23) holds for all $\xi \in [0, \xi_R]$. So $\tilde{\mathrm{u}}(\xi)$ is a parameterization (for $\xi \in [0, \xi_R]$) of an integral curve in the phase space corresponding to the kth eigenvector. Since

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\lambda_k(\tilde{\mathrm{u}}(\xi)) = (\nabla\lambda_k(\tilde{\mathrm{u}}(\xi)), \mathrm{r}^{(k)}(\tilde{\mathrm{u}}(\xi))) = 1, \tag{3.24}$$

it is obvious that $\lambda_k(\tilde{\mathbf{u}}(\xi)) = \xi + \lambda_k(\mathbf{u}_L)$ for all $\xi \in [0, \xi_R]$. Define the function $\mathbf{u}$ by

$$\mathbf{u}(x,t) := \begin{cases} \mathbf{u}_L, & x/t < \lambda_k(\mathbf{u}_L), \\ \tilde{\mathbf{u}}(x/t - \lambda_k(\mathbf{u}_L)), & \lambda_k(\mathbf{u}_L) < x/t < \lambda_k(\mathbf{u}_R), \\ \mathbf{u}_R, & \lambda_k(\mathbf{u}_R) < x/t. \end{cases} \tag{3.25}$$

It will be verified that $\mathbf{u}(x,t)$ defined in (3.25) is the solution of the Riemann problem (3.20). From now on we restrict ourselves to the case $\lambda_k(\mathbf{u}_L) < x/t < \lambda_k(\mathbf{u}_R)$. The cases $x/t < \lambda_k(\mathbf{u}_L)$ or $x/t > \lambda_k(\mathbf{u}_R)$ are trivial. It is easy to see that

$$\begin{aligned} \lambda_k(\mathbf{u}(x,t)) &= \lambda_k(\tilde{\mathbf{u}}(x/t - \lambda_k(\mathbf{u}_L))) \\ &= x/t - \lambda_k(\mathbf{u}_L) + \lambda_k(\mathbf{u}_L) \;=\; x/t. \end{aligned} \tag{3.26}$$

Therefore, using (3.23) with $\xi = x/t - \lambda_k(\mathbf{u}_L)$ and the previous equation, we have

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{u}(x,t) + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}(x,t)) &= \frac{\partial}{\partial t}\mathbf{u}(x,t) + A(\mathbf{u}(x,t))\frac{\partial}{\partial x}\mathbf{u}(x,t) \\ &= -\frac{x}{t^2}\mathbf{r}^{(k)}(\mathbf{u}(x,t)) + \frac{1}{t}A(\mathbf{u}(x,t))\mathbf{r}^{(k)}(\mathbf{u}(x,t)) \\ &= -\frac{x}{t^2}\mathbf{r}^{(k)}(\mathbf{u}(x,t)) + \frac{1}{t}\lambda_k(\mathbf{u}(x,t))\mathbf{r}^{(k)}(\mathbf{u}(x,t)) \;=\; \mathbf{0}. \end{aligned}$$

Thus $\mathbf{u}(x,t)$ defined in (3.25) is the solution of the Riemann problem with initial states $(\mathbf{u}_L, \mathbf{u}_R)$ indeed. This solution is called a k-rarefaction wave. An illustration is given in Figure 3.2.

In gas dynamics rarefaction waves are often called expansion waves. They represent smooth variations of the pressure and the density of the gas.

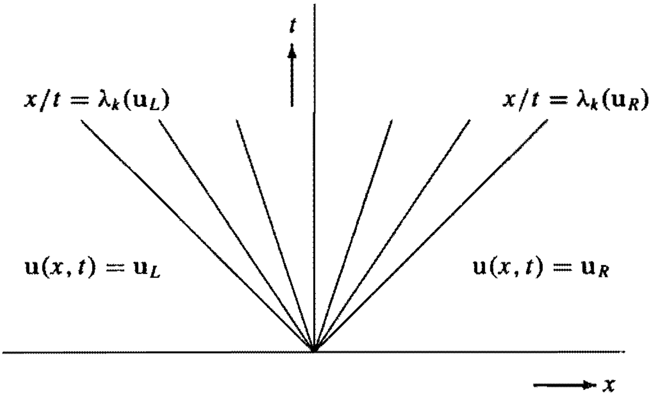
t
$x/t = \lambda_k(\mathbf{u}_L)$
$x/t = \lambda_k(\mathbf{u}_R)$
$\mathbf{u}(x,t) = \mathbf{u}_L$
$\mathbf{u}(x,t) = \mathbf{u}_R$
x

Figure 3.2. Characteristics of a k-rarefaction wave solution (3.25) of the Riemann problem (3.20).

3.3.3. Contact Discontinuities

In this section we want to derive the contact discontinuity solution of the Riemann problem (3.20). Suppose that $\mathbf{r}^{(k)}(\mathbf{u})$ is a linearly degenerate eigenvector. Let $\tilde{\mathbf{u}}(\xi)$ be the solution of (3.23) and suppose that the value ξ_R is chosen such that $\mathbf{u}_R = \tilde{\mathbf{u}}(\xi_R)$ is well defined. Since

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\lambda_k(\tilde{\mathbf{u}}(\xi)) = (\nabla\lambda_k(\tilde{\mathbf{u}}(\xi)), \mathbf{r}^{(k)}(\tilde{\mathbf{u}}(\xi))) = 0,$$

it is obvious that $\lambda_k(\tilde{\mathbf{u}}(\xi)) = \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}_R)$ for all $\xi \in [0, \xi_R]$. Hence, $\lambda_k(\tilde{\mathbf{u}})$ is constant along integral curves in the phase space corresponding to the kth eigenvector. Of course the value of $\lambda_k(\tilde{\mathbf{u}})$ might vary from one integral curve to the next. Define the discontinuous function $\mathbf{u}$ by

$$\mathbf{u}(x,t) := \begin{cases} \mathbf{u}_L, & x/t < \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}_R), \\ \mathbf{u}_R, & x/t > \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}_R). \end{cases} \tag{3.27}$$

It will be shown that the function $\mathbf{u}$ defined in (3.27) is the solution of the Riemann problem (3.20). For this it suffices to show that the jump condition (3.10) is satisfied. Since

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\xi}\left\{\mathbf{f}(\tilde{\mathbf{u}}(\xi)) - \lambda_k(\tilde{\mathbf{u}}(\xi))\tilde{\mathbf{u}}(\xi)\right\} &= A(\tilde{\mathbf{u}}(\xi))\frac{\mathrm{d}}{\mathrm{d}\xi}\tilde{\mathbf{u}}(\xi) - \lambda_k(\tilde{\mathbf{u}}(\xi))\frac{\mathrm{d}}{\mathrm{d}\xi}\tilde{\mathbf{u}}(\xi) \\ &= A(\tilde{\mathbf{u}}(\xi))\mathbf{r}^{(k)}(\tilde{\mathbf{u}}(\xi)) - \lambda_k(\tilde{\mathbf{u}}(\xi))\mathbf{r}^{(k)}(\tilde{\mathbf{u}}(\xi)) \\ &= \lambda_k(\tilde{\mathbf{u}}(\xi))\mathbf{r}^{(k)}(\tilde{\mathbf{u}}(\xi)) - \lambda_k(\tilde{\mathbf{u}}(\xi))\mathbf{r}^{(k)}(\tilde{\mathbf{u}}(\xi)) = 0, \end{aligned}$$

it is easy to see that (3.10) holds with $s = \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}_R)$. Thus $\mathbf{u}(x,t)$ defined in (3.27) is the solution of the Riemann problem with initial states $(\mathbf{u}_L, \mathbf{u}_R)$ indeed. This solution is called a k-contact discontinuity. An illustration is given in Figure 3.3.

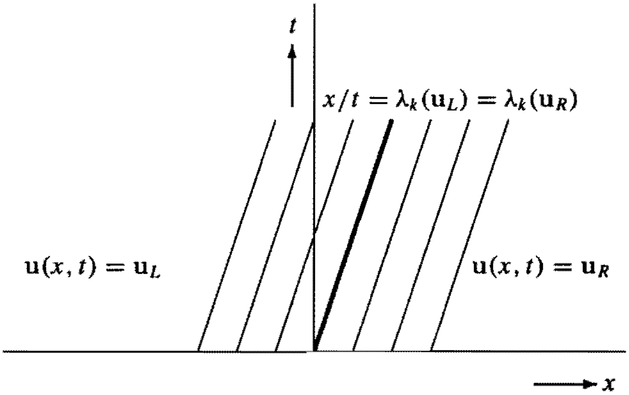
t
$x/t = \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}_R)$
$\mathbf{u}(x,t) = \mathbf{u}_L$
$\mathbf{u}(x,t) = \mathbf{u}_R$
x

Figure 3.3. Characteristics of a k-contact discontinuity solution (3.27) of the Riemann problem (3.20).

In gas dynamics a contact discontinuity represents an interface between two fluid regions of different densities but equal pressure. Since the contact interface moves with the fluid particles (e.g. $\lambda_k(\mathrm{u}_L) = \lambda_k(\mathrm{u}_R) = u$ for the Euler equations), the velocity has to be continuous over a contact discontinuity.

3.3.4. Shock Waves and the Lax Entropy Condition

Another elementary type of solution of the Riemann problem is given by shock wave solutions. Suppose that $\mathrm{r}^{(k)}(\mathrm{u})$ is a genuinely nonlinear eigenvector and $\lambda_k(\mathrm{u}_L) > \lambda_k(\mathrm{u}_R)$. Recall that if a discontinuity propagating with constant speed s has constant values u_L and u_R on either side of the discontinuity, then the jump condition (3.10) must hold. Now suppose we fix the point $\mathrm{u}_L \in I\!R^m$ and attempt to determine the set of all points u_R which can be connected to u_L by a discontinuity satisfying (3.10) for some s. This gives a system of m equations in $m+1$ unknowns: u_R and s, leading to a one parameter family of solutions. In the phase space there exists an integral curve through the point u_L describing the possible shock wave solutions. For a detailed description of shock wave solutions, see [49, 79]. Here only a short description is given, since a detailed description does not contribute very much to the understanding of the numerical methods which will be discussed.

Define the discontinuous function u by

$$\mathrm{u}(x,t) := \begin{cases} \mathrm{u}_L & x/t < s, \\ \mathrm{u}_R & x/t > s, \end{cases} \tag{3.28}$$

where the speed of discontinuity s is given by (3.10). If the characteristics corresponding to λ_k disappear into the shock as time advances, then we call this shock a *compression shock* (see Figure 3.4). If not, then the shock is called an *expansion shock.* In Example 3.5 we have described an expansion shock for Burgers' equation connecting the values $-\alpha$ and α. In gas dynamics shocks are solutions of the Rankine-Hugoniot equations with nonzero mass flow through the shock. Consequently, pressure and normal velocity undergo discontinuous variations, while the tangential velocity remains constant. The Rankine-Hugoniot equations imply a discontinuous entropy variation through the shock. This variation has to be positive, corresponding to compression shocks. It can be shown that across expansion shocks the entropy jumps to a lower value, which is not allowed by the second principle of thermodynamics [18]. To exclude expansion shocks as possible solutions we assume that the following inequalities hold

$$\begin{aligned} \lambda_{k-1}(\mathrm{u}_L) &< s < \lambda_k(\mathrm{u}_L), \\ \lambda_k(\mathrm{u}_R) &< s < \lambda_{k+1}(\mathrm{u}_R). \end{aligned} \tag{3.29}$$

These inequalities assert that $m-k+1$ characteristics impinge on the curve of discontinuity from the left and k from the right, a total of $m+1$ (see Figure 3.4). Now the jump condition (3.10) gives m equations connecting the values of the solution on both sides of the discontinuity with speed s. If $\mathrm{u}_L \neq \mathrm{u}_R$, then we may eliminate s from these equations to get $m-1$ equations between u_L and u_R. Hence, a total of $2m$ equations are obtained for the $2m$ variables u_L and u_R. If the function u defined in (3.28) satisfies (3.10) and (3.29), then u is the solution of the Riemann problem.

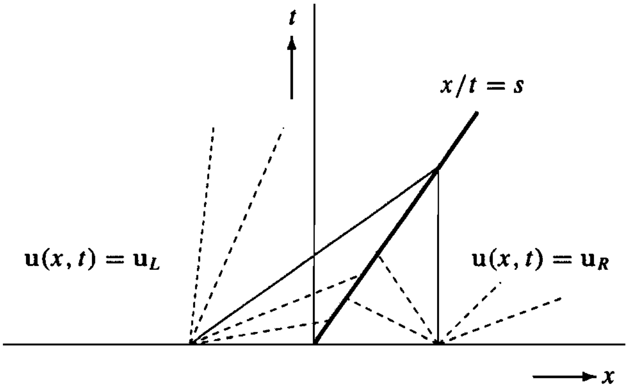

Figure 3.4. Characteristics of a k-shock wave solution (3.28) of the Riemann problem (3.20).

This solution is called a k-shock wave (in e.g. [79] it is proved that shock waves may occur for the Riemann problem). Condition (3.29) is called the *Lax entropy condition.* It is called an entropy condition since it can be proved that (3.29) holds if and only if (3.18) holds [79]. Hence, condition (3.29) is often used to define the entropy solution.

3.3.5. General Solution of the Riemann Problem

The theory described in the previous sections can be used to solve the Riemann problem (3.20). The Riemann problem can be solved provided $\|\mathrm{u}_L - \mathrm{u}_R\|$ is small in some particular norm. The following theorem describes the total solution of the Riemann problem (for a proof we refer to [79]).

Theorem 3.13. *Let* $\mathrm{u}_L \in \mathbb{R}^m$ *be given and suppose that the system* (3.20) *is hyperbolic. Further assume that each eigenvector of the Jacobian matrix of* f *is either genuinely nonlinear or linearly degenerate. Then there exists a neighbourhood* $\Omega \subset \mathbb{R}^m$ *of* u_L *such that, if* $\mathrm{u}_R \in \Omega$*, the Riemann problem* (3.20) *has a unique solution. This solution consists of at most* $m+1$ *constant states separated by shocks, rarefaction waves or contact discontinuities.*

In the following example the solution of the Riemann problem for linear systems is derived [49, 56, 81]. This solution is used in the derivation of Roe's numerical method for nonlinear conservation laws (see Section 4.5.3, [74]).

Example 3.14. Let the flux function f be linear, i.e. $\mathrm{f}(\mathrm{u}) = A\mathrm{u}$, where A is a constant $m \times m$-matrix. Hence (3.20) simplifies to

$$\frac{\partial}{\partial t}\mathrm{u}(x,t) + A\frac{\partial}{\partial x}\mathrm{u}(x,t) = 0 \tag{3.30a}$$

with initial data

$$\mathrm{u}^0(x) = \begin{cases} \mathrm{u}_L, & x < 0, \\ \mathrm{u}_R, & x > 0, \end{cases} \tag{3.30b}$$

where $\mathrm{u}_L \in I\!R^m$ and $\mathrm{u}_R \in I\!R^m$ are given constant states. Since in the linear case all eigenvalues of A are constant, all eigenvectors are linearly degenerate. Thus, only contact discontinuities appear in the solution. The eigenvectors $\mathrm{r}^{(k)}$, $k = 1, \ldots, m$, are linearly independent and therefore, $\{\mathrm{r}^{(1)}, \ldots, \mathrm{r}^{(m)}\}$ can be used as a basis for $I\!R^m$. A solution of (3.30) can be expressed with respect to this basis, i.e.

$$\mathrm{u}(x,t) = \mathrm{u}_L + \sum_{k=1}^{m} \beta_k(x,t)\mathrm{r}^{(k)},$$

where $\beta_k : I\!R \times [0, \infty) \to I\!R$ for all k with $1 \leq k \leq m$. Substitution of this expression into (3.30a) leads to

$$\frac{\partial}{\partial t}\mathrm{u}(x,t) + A\frac{\partial}{\partial x}\mathrm{u}(x,t) = \sum_{k=1}^{m} \left\{\frac{\partial}{\partial t}\beta_k(x,t) + \lambda_k \frac{\partial}{\partial x}\beta_k(x,t)\right\}\mathrm{r}^{(k)} = 0.$$

Since the eigenvectors are linearly independent, the following equalities should hold

$$\frac{\partial}{\partial t}\beta_k(x,t) + \lambda_k \frac{\partial}{\partial x}\beta_k(x,t) = 0, \quad k = 1, \ldots, m.$$

The general solutions of these equations are given by

$$\beta_k(x,t) = \beta_k^0(x - \lambda_k t), \quad k = 1, \ldots, m,$$

where $\beta_k^0 : I\!R \to I\!R$. Hence, the general solution of (3.30) reads

$$\mathrm{u}(x,t) = \mathrm{u}_L + \sum_{k=1}^{m} \beta_k^0(x - \lambda_k t)\mathrm{r}^{(k)}. \tag{3.31}$$

Let the initial states be decomposed as $\mathrm{u}_R - \mathrm{u}_L = \sum_{k=1}^{m} \alpha_k \mathrm{r}^{(k)}$ and recall the definition of the *Heavyside function* H by $H(x) = 1$ if $x > 0$ and $H(x) = 0$ if $x < 0$. Then,

$$\mathrm{u}(x,0) = \mathrm{u}_L + H(x)(\mathrm{u}_R - \mathrm{u}_L) = \mathrm{u}_L + \sum_{k=1}^{m} \alpha_k H(x)\mathrm{r}^{(k)}.$$

Comparing this equation and (3.31), the solution of the Riemann problem (3.30) is obviously,

$$\mathrm{u}(x,t) = \mathrm{u}_L + \sum_{k=1}^{m} \alpha_k H(x - \lambda_k t)\mathrm{r}^{(k)}. \tag{3.32}$$

In Figure 3.5 an illustration is given of the solution of (3.30) with $m = 3$. □

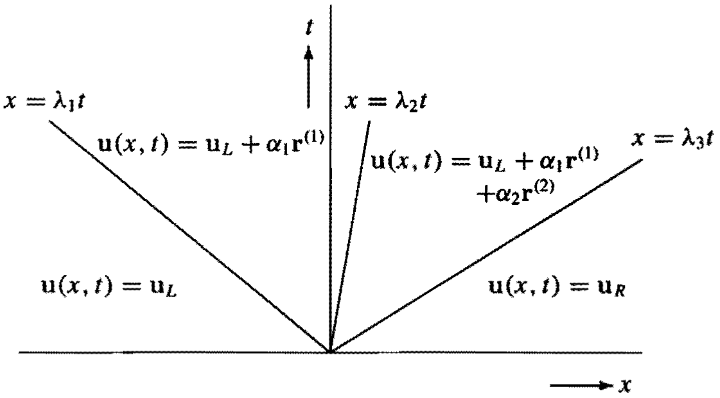
t
$x = \lambda_1 t$
$x = \lambda_2 t$
$x = \lambda_3 t$
$\mathrm{u}(x,t) = \mathrm{u}_L + \alpha_1 \mathrm{r}^{(1)}$
$\mathrm{u}(x,t) = \mathrm{u}_L + \alpha_1 \mathrm{r}^{(1)} + \alpha_2 \mathrm{r}^{(2)}$
$\mathrm{u}(x,t) = \mathrm{u}_L$
$\mathrm{u}(x,t) = \mathrm{u}_R$
x

Figure 3.5. The solution given by (3.32) of the linear Riemann problem (3.30) with $m = 3$.

3.3.6. Riemann Invariants

For the construction of the solution of Riemann problems the so-called *Riemann invariants* are useful; they are defined as follows.

Definition 3.15. *Consider the hyperbolic system* (3.20a). *Let* $\mathrm{r}^{(k)}(\mathrm{u})$ *be the kth right eigenvector of the Jacobian matrix* $A(\mathrm{u})$. *A* k-Riemann invariant *is a continuously differentiable function* $w_k : \mathbb{R}^m \to \mathbb{R}$ *such that*

$$(\nabla w_k(\mathrm{u}), \mathrm{r}^{(k)}(\mathrm{u})) = 0, \quad \forall \mathrm{u}.$$

Note that if $\mathrm{r}^{(k)}(\mathrm{u})$ is linearly degenerate, then the eigenvalue $\lambda_k(\mathrm{u})$ is a k-Riemann invariant (see (3.22)). For the construction of rarefaction waves or contact discontinuities (3.23) has to be solved. Let $\tilde{\mathrm{u}}(\xi)$, $0 \le \xi \le \xi_R$, be the solution of (3.23). Then

$$\frac{\mathrm{d}}{\mathrm{d}\xi} w_k(\tilde{\mathrm{u}}(\xi)) = (\nabla w_k(\tilde{\mathrm{u}}(\xi)), \mathrm{r}^{(k)}(\tilde{\mathrm{u}}(\xi))) = 0.$$

Therefore, a k-Riemann invariant is constant along the curve described by (3.23). If there are $m-1$ k-Riemann invariants $w_k^1, w_k^2, \ldots, w_k^{m-1}$, with linearly independent gradients, then it is easily seen that the curve described by (3.23) is part of the curve described by

$$\left\{\mathrm{u} \in \mathbb{R}^m \mid w_k^1(\mathrm{u}) = w_k^1(\mathrm{u}_L), w_k^2(\mathrm{u}) = w_k^2(\mathrm{u}_L), \ldots, w_k^{m-1}(\mathrm{u}) = w_k^{m-1}(\mathrm{u}_L)\right\}.$$

The following theorem gives the Riemann invariants corresponding with the eigenvectors belonging to the nonreactive Euler equations. The proof is just a straightforward computation.

Theorem 3.16. *Consider the nonreactive Euler equations, i.e.* (2.26) *with* $w = 0$. *Let the corresponding eigenvectors* $\mathbf{r}^{(k)}(\mathbf{u})$, $(k = 1, \dots, 4)$ *be given by* (3.9). *Then the Riemann invariants* w_k^i *corresponding to the eigenvectors* $\mathbf{r}^{(k)}(\mathbf{u})$ *are given by, respectively,*

$$\begin{array}{rclrclrcl}
w_1^1(\mathbf{u}) & = & u + \dfrac{2}{\gamma - 1}c, & w_1^2(\mathbf{u}) & = & S, & w_1^3(\mathbf{u}) & = & Y, \\
w_2^1(\mathbf{u}) & = & u, & w_2^2(\mathbf{u}) & = & p, & & & \\
w_3^1(\mathbf{u}) & = & u, & w_3^2(\mathbf{u}) & = & p, & & & \\
w_4^1(\mathbf{u}) & = & u - \dfrac{2}{\gamma - 1}c, & w_4^2(\mathbf{u}) & = & S, & w_4^3(\mathbf{u}) & = & Y.
\end{array}$$

3.4. THE RIEMANN PROBLEM FOR THE NONREACTIVE EULER EQUATIONS

A simple example illustrating the interesting behaviour of the solution of a Riemann problem is the *shock tube problem* of gas dynamics. The physical set-up is a tube filled with two gases (denoted by G_1 and G_2). Initially the tube is divided by a membrane in two sections, where in each section only one gas is present. The density and the pressure of gas G_1 at the left side of the membrane are larger than the density and the pressure of gas G_2 at the right side of the membrane, and the velocity is zero everywhere. If only one gas is present, then G_1 and G_2 simply denote two different states for the same gas. At time $t = 0$ the membrane is suddenly removed or broken, and the gas flows. It is expected that the gas moves in the direction of lower pressure. Assuming that the flow is uniform across the tube, there is variation in only one direction and if no reaction takes place, then the Riemann problem (3.20) corresponding to the one-dimensional nonreactive Euler equations is relevant (i.e. (2.26) with $w = 0$). The eigenvectors belonging to these equations are given in (3.9), where Y is the mass fraction of gas G_2. As described in Theorem 3.13 the solution consists of at most 5 constant states separated by shocks, rarefaction waves or contact discontinuities. Since $\lambda_2(\mathbf{u}) = \lambda_3(\mathbf{u}) = u$, the solution consists of 4 constant states denoted by $\mathbf{u}_L$, $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{u}_R$. A schematic diagram of this solution is given in Figure 3.6.

The first eigenvector $\mathbf{r}^{(1)}(\mathbf{u})$ is genuinely nonlinear (see Theorem 3.11), so the 1-wave is always a rarefaction wave or shock wave, depending on $\mathbf{u}_L$ and $\mathbf{u}_1$. If $\mathbf{u}_L$ and $\mathbf{u}_1$ satisfy

$$u_L + \frac{2}{\gamma - 1}c_L = u_1 + \frac{2}{\gamma - 1}c_1, \qquad S_L = S_1, \qquad Y_L = Y_1 \tag{3.33}$$

and

$$u_L - c_L < u_1 - c_1, \tag{3.34}$$

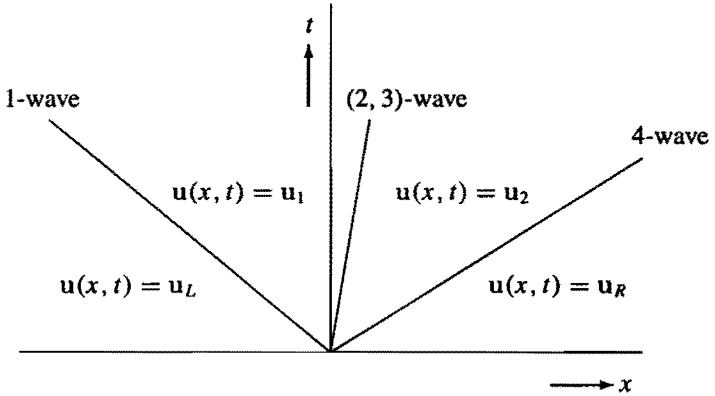

Figure 3.6. Schematic diagram of the solution of the Riemann problem for the nonreactive Euler equations.

then a 1-rarefaction wave exists. This solution is given by

$$\begin{array}{rcll} \mathbf{u} &=& \mathbf{u}_L, & x/t < u_L - c_L, \\ \left.\begin{array}{rcl} u + \dfrac{2}{\gamma-1}c &=& u_L + \dfrac{2}{\gamma-1}c_L \\ S &=& S_L \\ Y &=& Y_L \\ u - c &=& x/t \end{array}\right\} & & , & u_L - c_L < x/t < u_1 - c_1, \\ \mathbf{u} &=& \mathbf{u}_1, & u_1 - c_1 < x/t. \end{array} \qquad (3.35)$$

Note that $u - c = x/t$ follows from (3.26). If $\mathbf{u}_L$ and $\mathbf{u}_1$ are such that (3.33) holds, while $u_L - c_L > u_1 - c_1$, then the solution given by (3.35) corresponds to a multivalued solution, called a *compression wave*. Although such a compression wave has no physical meaning (the physical solution is a shock wave in this case), it is shown in [69] that by allowing compression waves, an approximate solution of the Riemann problem can be obtained, which leads to an excellent numerical method for the Euler equations.

If $\mathbf{u}_L$ and $\mathbf{u}_1$ are such that (3.33) holds and $u_L - c_L > u_1 - c_1$, then a 1-shock wave, corresponding to $\mathbf{r}^{(1)}(\mathbf{u})$ exists. This solution is given by

$$\mathbf{u}(x,t) = \begin{cases} \mathbf{u}_L & x/t < s, \\ \mathbf{u}_1 & x/t > s, \end{cases} \qquad (3.36)$$

where s is defined by the Rankine-Hugoniot equations (3.10) (with $\mathbf{u}_R = \mathbf{u}_1$). The theory of the nonreactive Rankine-Hugoniot equations (3.10) is completely analogous to that presented for the reactive Rankine-Hugoniot equations in Section 2.4. For a detailed description, see [18, 79].

The second and the third eigenvector belonging to the Euler equations are linearly degenerate (see Theorem 3.11). Thus, the (2, 3)-wave is always a contact discontinuity. If $\mathbf{u}_1$ and $\mathbf{u}_2$ are such that $u_1 = u_2$ and $p_1 = p_2$, then a contact discontinuity, corresponding to $\mathbf{r}^{(2)}(\mathbf{u})$, $\mathbf{r}^{(3)}(\mathbf{u})$ exists, which is given by

$$\mathbf{u}(x,t) = \begin{cases} \mathbf{u}_1 & x/t < u_1 = u_2, \\ \mathbf{u}_2 & x/t > u_1 = u_2. \end{cases} \tag{3.37}$$

Hence, it is impossible for gas particles to cross a contact discontinuity. Therefore, at the left of the contact discontinuity there is only gas G_1, while at the right only gas G_2 is present (see the mass fraction in Figure 3.7).

Next, we consider the 4-wave. The fourth eigenvector $\mathbf{r}^{(4)}(\mathbf{u})$ is genuinely nonlinear (see Theorem 3.11), so the 4-wave is always a rarefaction wave or shock wave, depending on $\mathbf{u}_2$ and $\mathbf{u}_R$. If $\mathbf{u}_2$ and $\mathbf{u}_R$ are such that

$$u_2 - \frac{2}{\gamma-1}c_2 = u_R - \frac{2}{\gamma-1}c_R, \qquad S_2 = S_R, \qquad Y_2 = Y_R \tag{3.38}$$

and

$$u_2 + c_2 < u_R + c_R, \tag{3.39}$$

then a 4-rarefaction wave, corresponding to $\mathbf{r}^{(4)}(\mathbf{u})$ exists. This solution is given by

$$\begin{array}{rcll} \mathbf{u} & = & \mathbf{u}_2, & x/t < u_2 + c_2, \\ u - \frac{2}{\gamma-1}c & = & u_R - \frac{2}{\gamma-1}c_R & \\ S & = & S_R & \\ Y & = & Y_R & \Bigg\}, \quad u_2 + c_2 < x/t < u_R + c_R, \\ u + c & = & x/t & \\ \mathbf{u} & = & \mathbf{u}_R, & u_R + c_R < x/t. \end{array} \tag{3.40}$$

If $u_2 + c_2 > u_R + c_R$ and (3.38) still hold, then (3.40) generates a compression wave.

Finally, if the pair $\mathbf{u}_2$ and $\mathbf{u}_R$ is such that (3.38) and $u_2 + c_2 > u_R + c_R$ hold, then a 4-shock wave, corresponding to $\mathbf{r}^{(4)}(\mathbf{u})$ exists. This solution is given by

$$\mathbf{u}(x,t) = \begin{cases} \mathbf{u}_2 & x/t < s, \\ \mathbf{u}_R & x/t > s, \end{cases} \tag{3.41}$$

where s is obtained from the Rankine-Hugoniot equations (3.10) (with $\mathbf{u}_L = \mathbf{u}_2$).

A combination of the equations (3.35)-(3.41) gives the total solution. An example is given in Figure 3.7.

The following theorem describes whether the Riemann problem for the Euler equations has a solution or not (for a proof we refer to [79]).

Theorem 3.17. *Consider the one-dimensional nonreactive Euler equations, i.e.* (2.26) *with* $w = 0$. *Let* $\mathbf{u}_L$ *and* $\mathbf{u}_R$ *be any two states (not necessarily close). Then there is a unique solution to the Riemann problem* (3.20), *if and only if*

$$u_R - u_L < \frac{2}{\gamma-1}(c_L + c_R). \tag{3.42}$$

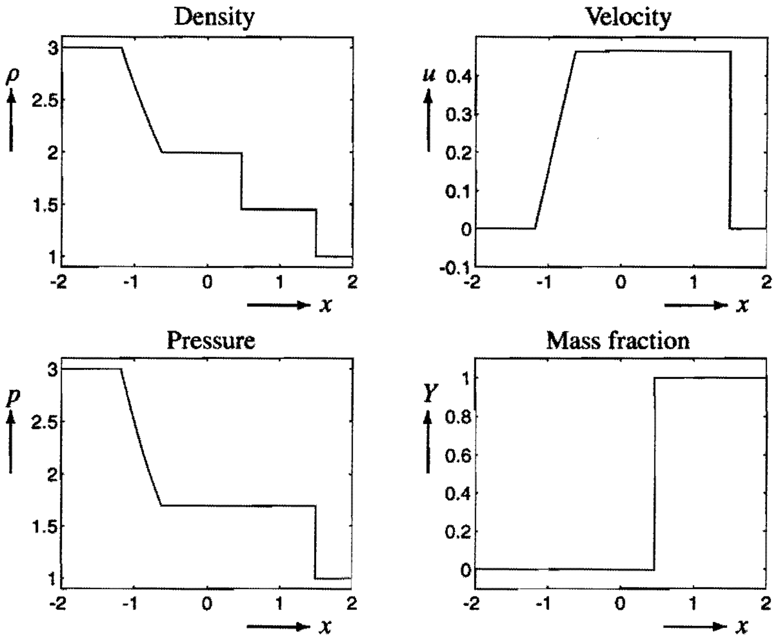

Figure 3.7. Solution at time $t = 1$ of a shock tube problem for the one-dimensional nonreactive Euler equations, i.e. (2.26) with $w = 0$. The initial conditions are $p(x,0) = 3$, $\rho(x,0) = 3$, $u(x,0) = 0$, $Y(x,0) = 0$ if $x < 0$ and $p(x,0) = 1$, $\rho(x,0) = 1$, $u(x,0) = 0$, $Y(x,0) = 1$ if $x > 0$.

Theorem 3.17 is a global theorem, in that two states are not required to be close to each other (see Theorem 3.13). If (3.42) is violated, then the relative velocities on both sides of the membrane are so large that a vacuum is formed.

4

FINITE DIFFERENCE METHODS FOR CONSERVATION LAWS

Apart from practical applications, there are two other reasons for studying numerical methods for hyperbolic conservation laws. First, there are special difficulties associated with solving hyperbolic conservation laws (e.g. shock formation) which must be dealt with carefully in developing numerical methods. Secondly, a great deal is known about the mathematical structure of these equations and their solutions [49, 79]. This theory can be used to develop reliable numerical methods.

In this thesis only finite difference methods will be considered. Methods developed using straightforward finite difference discretizations are inappropriate near discontinuities, since they are based on truncated Taylor series expansions. A survey of these methods (with applications to the Euler equations) is given in [39, 40]. Another important class of numerical methods are the first order Godunov-type methods [24, 36, 53]. These methods use, in some way, the exact solution of the Riemann problem and do not produce oscillations around discontinuities. However, since these methods are only first order, the solutions are smoothed around discontinuities. Therefore, other methods have been developed. A very popular class of methods are the high resolution methods, which are described in the next chapter. In this chapter we present some finite difference methods for hyperbolic conservation laws and elaborate some theoretical aspects. Clearly, this survey is not complete.

This chapter is organized as follows. The first section is of a preparatory nature. Some basic numerical concepts are introduced, such as discrete conservation, discretization error, consistency, convergence and stability. Furthermore the well-known theorem of Lax and Wendroff is presented and two straightforward methods are discussed for the scalar, linear convection equation. In Section 4.2 the modified equation is introduced. The modified equation is useful for studying the behaviour of numerical methods. In Section 4.3 the numerical entropy condition is introduced. Finally, in the last section first order Godunov-type methods are discussed. We describe the basic Godunov method and a method developed by Roe [74].

4.1. SOME BASIC NUMERICAL CONCEPTS

A variety of numerical methods can be developed for conservation laws with source terms (3.2). A very natural (and popular) way to solve (3.2) is a time splitting method. In the time splitting method we alternate between solving the homogeneous conservation laws without source terms,

$$\frac{\partial}{\partial t}\mathrm{u}(x,t) + \frac{\partial}{\partial x}\mathrm{f}(\mathrm{u}(x,t)) = 0 \tag{4.1}$$

and solving the ordinary differential equations with no space derivatives,

$$\frac{\partial}{\partial t}\mathrm{u}(x,t) = \mathrm{q}(\mathrm{u}(x,t)). \tag{4.2}$$

In the following we restrict ourselves to the initial value problem for (4.1), since it is well-known that many important properties of (3.2a) are determined by the homogeneous part. In Chapter 6 we describe the time splitting method in more detail and return to conservation laws with source terms.

When solutions of (4.1) are calculated numerically, new problems arise. A finite difference discretization developed for smooth solutions is expected to be inappropriate near discontinuities. Indeed, if discontinuous solutions of conservation laws are computed using standard finite difference methods, poor numerical results are obtained [28, 39, 56]. Later in this section a short description of two standard methods is given, since they are the starting point for more sophisticated methods.

For a given time step Δt, the discrete time levels t^n are defined by

$$t^n := n\Delta t, \quad n = 0, 1, 2, \ldots .$$

For a given mesh width Δx, the spatial mesh points x_i are defined by

$$x_i := i\Delta x, \quad i = \ldots, -2, -1, 0, 1, 2, \ldots .$$

It will also be useful to define intermediate points

$$x_{i+1/2} = (i + \tfrac{1}{2})\Delta x.$$

The finite difference methods we shall consider, produce approximations $\mathbf{U}_i^n \in \mathbb{R}^m$ to the true solution $\mathrm{u}(x_i, t^n)$ at the discrete points (x_i, t^n). The average of $\mathrm{u}(\cdot, t^n)$ on the cell $[x_{i-1/2}, x_{i+1/2})$ is defined by

$$\bar{\mathrm{u}}_i^n := \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \mathrm{u}(x, t^n)\mathrm{d}x. \tag{4.3}$$

For conservation laws it is often convenient to view $\mathbf{U}_i^n$ as an approximation to this average, since the integral form of the conservation laws describes the evolution in time of integrals such as (4.3). As initial data for a numerical method we use a given function

$\mathbf{u}^0 = \mathbf{u}(\cdot, 0)$ to define $\mathbf{U}^0$ by cell averages, i.e. $\mathbf{U}_i^0 := \bar{\mathbf{u}}_i^0$. In the following it is assumed that, for a given constant $\tau > 0$, the mesh width Δx and time step Δt satisfy

$$\frac{\Delta t}{\Delta x} = \tau.$$

For the sake of convenience, we construct a piecewise constant function $\mathbf{U}_{\Delta t}$ for all x and t from the discrete values $\mathbf{U}_i^n$ by

$$\mathbf{U}_{\Delta t}(x, t) := \mathbf{U}_i^n, \quad \forall\, (x, t) \in [x_{i-1/2}, x_{i+1/2}) \times [t^n, t^{n+1}). \tag{4.4}$$

Many difficulties for numerical methods for hyperbolic conservation laws are caused by the fact that a discontinuous weak solution of (4.1) may occur. It is not surprising that a method might converge to a wrong solution, since in general a weak solution is not unique. Therefore the discrete solution of problem (4.1) is often required to satisfy a discrete form of the entropy condition, as defined in Definition 3.7. More surprisingly, a method may converge to a function that is not a weak solution at all. The latter problem is avoided by only considering conservative methods, which are consistent with the conservation law (4.1) [50].

Definition 4.1. *Let a* $(2k + 1)$*-point finite difference method, with* 2 *time levels, for the hyperbolic conservation law* (4.1) *be given. The numerical method is said to be* conservative, *if the corresponding scheme can be written as*

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \tau(\mathbf{F}_{i+1/2}^n - \mathbf{F}_{i-1/2}^n), \tag{4.5}$$

where $\mathbf{F}$ *is a continuous function of the values of* $\mathbf{U}$ *at* $2k$ *points, i.e.*

$$\mathbf{F}_{i+1/2}^n = \mathbf{F}(\mathbf{U}_{i+k}^n, \ldots, \mathbf{U}_{i-k+1}^n). \tag{4.6}$$

$\mathbf{F}$ *is called the* numerical flux function.

Another important concept is the *local discretization error*. The local discretization error $\mathbf{D}_{\Delta t}(x, t)$ is a measure how well the difference equation approximates the differential equation locally (at the point (x, t)). Let the conservative, $(2k + 1)$-point scheme (4.5) be written as

$$\mathbf{U}_i^{n+1} = \mathcal{L}_{\Delta t}(\mathbf{U}_{i+k}^n, \ldots, \mathbf{U}_{i-k}^n), \tag{4.7}$$

where $\mathcal{L}_{\Delta t}$ is a finite difference operator. Using the piecewise constant function $\mathbf{U}_{\Delta t}$, as defined in (4.4), equation (4.7) can be rewritten in the functional form

$$\mathbf{U}_{\Delta t}(\cdot, t + \Delta t) = \mathcal{L}_{\Delta t}\mathbf{U}_{\Delta t}(\cdot, t), \tag{4.8}$$

where $\mathcal{L}_{\Delta t}$ is an operator between two function spaces. We will use the same symbol $\mathcal{L}_{\Delta t}$ to denote both the discrete and the continuous operator. Note that for a $(2k + 1)$-point method, $\mathbf{U}_{\Delta t}(\cdot, t)$ is evaluated at $2k + 1$ points. If now each $\mathbf{U}_{\Delta t}(\cdot, t)$ in (4.8) is replaced by the exact solution of (4.1), then in general the equality will not hold exactly. This leads to the following definition.

Definition 4.2. *Consider a conservative,* $(2k+1)$*-point method written in the generic form* (4.8). *The* local discretization error $\mathbf{D}_{\Delta t}$ *of this method at the point* (x,t) *is defined by*

$$\mathbf{D}_{\Delta t}(x,t) := \frac{1}{\Delta t}\Big\{\mathbf{u}(x,t+\Delta t) - (\mathcal{L}_{\Delta t}\mathbf{u}(\cdot,t))(x)\Big\}, \tag{4.9}$$

where $\mathbf{u}$ *is the exact solution of* (4.1).

In the previous definition $(\mathcal{L}_{\Delta t}\mathbf{u}(\cdot,t))(x)$ denotes the value of the function $\mathcal{L}_{\Delta t}\mathbf{u}(\cdot,t)$ at the point x. Note that the local discretization error depends on the exact solution $\mathbf{u}$ of (4.1). Often $\mathbf{u}$ is assumed to be a smooth solution, since Taylor series expansions are used to calculate the local discretization error. Using the local discretization error we can define the concept of *consistency* [56].

Definition 4.3. *Consider a conservative,* $(2k+1)$*-point method. The numerical method is called* consistent of order p *with the conservation law* (4.1) *in some particular norm* $\|\cdot\|$, *if for each time* $T > 0$ *there exists a constant* C *and a value* k_0 *such that*

$$\|\mathbf{D}_{\Delta t}(\cdot,t)\| \leq C\Delta t^p, \quad \forall t \leq T, \ \Delta t < k_0$$

with $p > 0$. *The method is called consistent with the conservation law* (4.1), *if it is consistent of order* p.

It can be shown (see [28]) that for consistency of a conservative method it is sufficient to require the flux function $\mathbf{F}$ of the corresponding scheme (4.5) to be Lipschitz continuous and to satisfy

$$\mathbf{F}(\mathbf{u},\ldots,\mathbf{u}) = \mathbf{f}(\mathbf{u}). \tag{4.10}$$

The importance of the concepts that are introduced in this section is shown by the well-known theorem of Lax-Wendroff. Lax and Wendroff proved that if we use a conservative method and the numerical solution converges to some function $\mathbf{u}$, then this function is a weak solution of the conservation law (4.1). Before we present this theorem, it is explained in more detail what is meant by convergence.

The *global discretization error* $\mathbf{E}_{\Delta t}(x,t)$ of a conservative, $(2k+1)$-point method is defined for arbitrary x and t as

$$\mathbf{E}_{\Delta t}(x,t) := \mathbf{U}_{\Delta t}(x,t) - \mathbf{u}(x,t).$$

With this definition, convergence of a numerical method is defined.

Definition 4.4. *Consider a conservative,* $(2k+1)$*-point method. The method is called* convergent *in some particular norm* $\|\cdot\|$, *if*

$$\|\mathbf{E}_{\Delta t}(\cdot,t)\| \to 0, \quad as\ \Delta t \to 0,$$

for any fixed $t \geq 0$ *and for all initial data* $\mathbf{u}_0$ *with* $\|\mathbf{u}_0\|$ *finite.*

The Lax-Wendroff theorem requires convergence with respect to the L_1-norm on compact subsets. For the sake of simplicity we restrict ourselves to rectangular compact subsets $\Omega \subset I\!R \times [0, \infty)$, i.e. there exists x_L, x_R and T such that $\Omega := [x_L, x_R] \times [0, T]$. The $L_1(\Omega)$-norm $\|\cdot\|_{1,\Omega}$ is now defined as

$$\|v\|_{1,\Omega} := \int_0^T \int_{x_L}^{x_R} |v(x,t)| \mathrm{d}x\mathrm{d}t.$$

A numerical method is called *L_1^{loc}-convergent,* if it is convergent in the $L_1(\Omega)$-norm for all compact $\Omega \subset I\!R \times [0, \infty)$. Now we can present the theorem of Lax-Wendroff [28, 35, 50].

Theorem 4.5 (Lax-Wendroff). *Suppose the finite difference method* (4.5) *has a Lipschitz continuous flux function* $\mathbf{F}$ *that satisfies* (4.10). *Let* $\mathbf{U}_i^n$ *be a solution of* (4.5) *with given initial values* $\mathbf{U}_i^0 = \bar{\mathbf{u}}_i^0$, *as defined in* (4.3). *Define the piecewise constant function* $\mathbf{U}_{\Delta t}$ *as in* (4.4). *Suppose that there exists a sequence* $\Delta t_m \downarrow 0$ *for* $m \rightarrow \infty$ *for which the limit*

$$\lim_{m\to\infty} \mathbf{U}_{\Delta t_m}(x,t) = \mathbf{u}(x,t)$$

exists in the sense of bounded, L_1^{loc}-convergence, i.e.

$$\|\mathbf{U}_{\Delta t_m}\|_{L_\infty(I\!R\times[0,\infty))} \le C,$$

$$\mathbf{U}_{\Delta t_m} \textit{ converges towards } \mathbf{u} \textit{ in } L_1^{loc} \textit{ for } m \to \infty.$$

Then the limit $\mathbf{u}$ *is a weak solution of* (4.1).

The final concept we introduce in this section is *stability.* For simplicity it is assumed that the numerical method is linear, i.e. $\mathcal{L}_{\Delta t}$ is a linear difference operator. Note that (4.9) can be rewritten in the form

$$\mathbf{u}(x, t+\Delta t) = (\mathcal{L}_{\Delta t}\mathbf{u}(\cdot, t))(x) + \Delta t\mathbf{D}_{\Delta t}(x,t).$$

Since the numerical solution satisfies (4.8), after subtracting these two equations a simple recurrence relation for the global discretization error $\mathbf{E}_{\Delta t}$ is obtained,

$$\mathbf{E}_{\Delta t}(x, t+\Delta t) = (\mathcal{L}_{\Delta t}\mathbf{E}_{\Delta t}(\cdot, t))(x) - \Delta t\mathbf{D}_{\Delta t}(x,t).$$

Note that linearity is essential here. The latter equation can be rewritten in the functional form

$$\mathbf{E}_{\Delta t}(\cdot, t+\Delta t) = \mathcal{L}_{\Delta t}\mathbf{E}_{\Delta t}(\cdot, t) - \Delta t\mathbf{D}_{\Delta t}(\cdot, t).$$

The global error $\mathbf{E}_{\Delta t}$ at time $t + \Delta t$ consists of two parts. One is the new local error $-\Delta t\mathbf{D}_{\Delta t}$ introduced in the last time step. The other part is the cumulative error from the previous time steps. By applying this relation recursively we obtain an expression for the global error at time t^n

$$\mathbf{E}_{\Delta t}(\cdot, t^n) = \mathcal{L}_{\Delta t}^n\mathbf{E}_{\Delta t}(\cdot, 0) - \Delta t\sum_{j=1}^{n} \mathcal{L}_{\Delta t}^{n-j}\mathbf{D}_{\Delta t}(\cdot, t^{j-1}). \tag{4.11}$$

Here superscripts for $\mathcal{L}_{\Delta t}$ represent powers of the linear operator obtained by repeated applications. In order to obtain a bound on the global error, we must ensure that the local error $\mathbf{D}_{\Delta t}(\cdot, t^{j-1})$ is not unduly amplified by applying $n - j$ steps of the method. Note that a bound is always with respect to some given norm $\|\cdot\|$.

Next, the concept of stability is introduced [56].

Definition 4.6. *Consider a conservative, $(2k+1)$-point linear method written in the generic form* (4.8) *for arbitrary x and t. The numerical method is called* stable *in some particular norm $\|\cdot\|$, if for each time $T > 0$ there exists a constant $C > 0$ and a value k_0 such that*

$$\|\mathcal{L}^n_{\Delta t}\| \le C, \quad \forall n\Delta t \le T, \ \Delta t < k_0$$

holds.

In practice, instead of Definition 4.6, often the *von Neumann method* for stability analysis is used [39, 73]. This method gives necessary conditions for a numerical method to be stable. Unfortunately, these conditions are not always sufficient for stability.

In the remainder of this thesis only conservative numerical methods, which are consistent with the conservation law (4.1), are considered. For hyperbolic conservation laws the numerical results near discontinuities are of paramount importance. In the following we present two basic examples illustrating the poor behaviour of standard finite difference methods around discontinuities. Discontinuities are smeared out as time evolves, or oscillations occur near discontinuities. All these phenomena can be observed in scalar problems. Therefore, in the following examples we restrict ourselves to the scalar, linear convection equation

$$\frac{\partial}{\partial t}u(x,t) + a\frac{\partial}{\partial x}u(x,t) = 0. \tag{4.12}$$

Hence, the flux function $f : I\!R \to I\!R$ is given by $f(u) := au$. It can easily be shown that the exact solution of (4.12) is given by $u(x,t) = u^0(x - at)$. We assume that the solution u is three times continuously differentiable. We briefly describe two well-known finite difference methods: the *basic upwind method* (a first order method) and the *Lax-Wendroff method* (a second order method). Numerical results for both methods are presented in Figure 4.1. It is useful to introduce the *Courant number*, which is defined by

$$\sigma := a\tau = a\frac{\Delta t}{\Delta x}. \tag{4.13}$$

Example 4.7 (Basic upwind). Upwind methods depend on the stream direction of the fluid. If in (4.12) $a > 0$, then the information is propagating in the positive x-direction. Thus the information in, say, point x_i has reached point x_{i-1} before. Therefore, in this case, it is meaningful to replace $\partial u/\partial x$ by a backward difference. Similarly $\partial u/\partial x$ is replaced by a forward difference if $a < 0$. In both cases $\partial u/\partial t$ is replaced by a forward difference. Let a^+ and a^- be defined by

$$\begin{aligned} a^+ &:= \max(a, 0) \ge 0, \\ a^- &:= \min(a, 0) \le 0, \end{aligned}$$

then the *basic upwind method* is given by

$$U_i^{n+1} = U_i^n - \tau\left\{a^+(U_i^n - U_{i-1}^n) + a^-(U_{i+1}^n - U_i^n)\right\}. \qquad (4.14)$$

This method is stable in the $L_1(\Omega)$-norm under the *CFL condition* [39, 56]

$$|\sigma| \le 1. \qquad (4.15)$$

If the numerical flux function F is defined by

$$F_{i+1/2}^{(BU)} := a^+U_i^n + a^-U_{i+1}^n, \qquad (4.16)$$

then it follows immediately that the basic upwind method (4.14) is a conservative method, which is consistent with the conservation law (4.12). Extensions of the upwind method for nonlinear conservation laws are given in Section 4.5. ☐

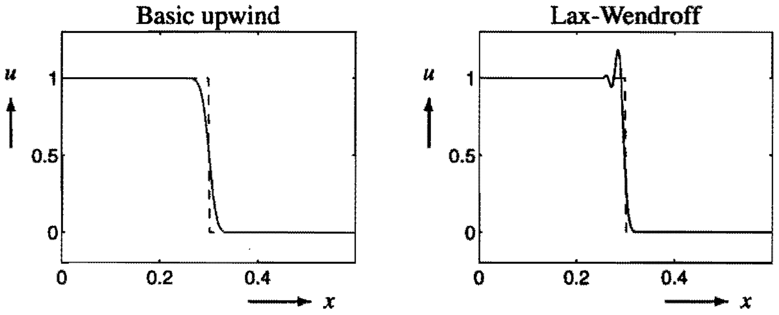

Figure 4.1. Exact solution (dashed line) and numerical solution (solid line) of (4.12) at $t = 0.3$ with $a = 1$, $\Delta t = 0.002$, $\Delta x = 0.0025$ and the initial condition $u(x,0) = 1$ if $x < 0$ and $u(x,0) = 0$ if $x > 0$.

The second order method presented below is based on the Taylor series expansion

$$u(x, t+\Delta t) = u(x,t) + \Delta t\frac{\partial}{\partial t}u(x,t) + \tfrac{1}{2}\Delta t^2\frac{\partial^2}{\partial t^2}u(x,t) + \mathcal{O}(\Delta t^3). \qquad (4.17)$$

Since it is assumed that u is three times continuously differentiable, it follows from $\partial u/\partial t = -a\partial u/\partial x$ that

$$\frac{\partial^2}{\partial t^2}u(x,t) = -a\frac{\partial^2}{\partial t\partial x}u(x,t) = -a\frac{\partial^2}{\partial x\partial t}u(x,t) = a^2\frac{\partial^2}{\partial x^2}u(x,t).$$

Using this equality, equation (4.17) becomes

$$u(x, t+\Delta t) = u(x,t) - \Delta t a\frac{\partial}{\partial x}u(x,t) + \tfrac{1}{2}\Delta t^2 a^2\frac{\partial^2}{\partial x^2}u(x,t) + \mathcal{O}(\Delta t^3). \qquad (4.18)$$

Example 4.8 (Lax-Wendroff). The *Lax-Wendroff method* results from retaining only the first three terms of (4.18) and using central difference approximations for the derivatives appearing there. Therefore, the corresponding finite difference scheme is

$$U_i^{n+1} = U_i^n - \tfrac{1}{2}\sigma(U_{i+1}^n - U_{i-1}^n) + \tfrac{1}{2}\sigma^2(U_{i+1}^n - 2U_i^n + U_{i-1}^n). \qquad (4.19)$$

The Lax-Wendroff method is stable under the CFL condition (4.15) [39]. If the numerical flux function F is defined by

$$F_{i+1/2}^{(LW)} = \tfrac{1}{2}a(U_{i+1}^n + U_i^n) - \tfrac{1}{2}a\sigma(U_{i+1}^n - U_i^n), \qquad (4.20)$$

then it is obvious that the Lax-Wendroff method is a conservative method, which is consistent with the conservation law (4.12). □

4.2. DISCRETE CONSERVATION

The basic principle underlying a homogeneous conservation law is that the total quantity of a conserved variable in any region changes only due to the flux through the boundaries (note that we assume $\mathbf{q} = \mathbf{0}$). This gives the integral form of the conservation law, i.e. (3.1) with $\mathbf{q} = \mathbf{0}$. Notice in particular that if $\mathbf{u}$ is constant outside some finite interval during the time interval $a \leq t \leq b$, say $\mathbf{u} = \mathbf{u}_L$ for $x \leq x_L$ and $\mathbf{u} = \mathbf{u}_R$ for $x \geq x_R$, then integrating (3.1) in time over $[a, b]$ gives

$$\int_{x_L}^{x_R} \mathbf{u}(x, b)\mathrm{d}x = \int_{x_L}^{x_R} \mathbf{u}(x, a)\mathrm{d}x + (b-a)(\mathbf{f}(\mathbf{u}_L) - \mathbf{f}(\mathbf{u}_R)). \qquad (4.21)$$

For a finite propagation speed this will be the case if the initial data is constant outside some finite interval.

Next we show that a consistent, conservative method will have a form of conservation analogous to (4.21). If $\mathbf{u}^0$ is constant outside some finite interval, so is $\mathbf{U}^n$ (see (4.5) and (4.6)). Consider a $(2k+1)$-point method and let $n > 0$ be fixed, then there exist indices I_1 and I_2 such that $\mathbf{U}_i^n = \mathbf{u}_L$ for all $i \leq I_1+k-1$ and $\mathbf{U}_i^n = \mathbf{u}_R$ for all $i \geq I_2-k+1$. Hence, using this and the consistency of the flux function $\mathbf{F}$ (see (4.10)), we obtain $\mathbf{F}_{I_1-1/2}^n = \mathbf{f}(\mathbf{u}_L)$ and $\mathbf{F}_{I_2+1/2}^n = \mathbf{f}(\mathbf{u}_R)$ at time t^n. We start by substituting j for $n+1$ in (4.5). After multiplying the resulting scheme by Δx, summing over i and using the above we obtain

$$\Delta x \sum_{i=I_1}^{I_2} \mathbf{U}_i^j = \Delta x \sum_{i=I_1}^{I_2} \mathbf{U}_i^{j-1} + \Delta t(\mathbf{f}(\mathbf{u}_L) - \mathbf{f}(\mathbf{u}_R)).$$

Summing the equations above over all j with $1 \leq j \leq n$, we see that

$$\Delta x \sum_{i=I_1}^{I_2} \mathbf{U}_i^n = \Delta x \sum_{i=I_1}^{I_2} \mathbf{U}_i^0 + n\Delta t(\mathbf{f}(\mathbf{u}_L) - \mathbf{f}(\mathbf{u}_R)).$$

Suppose that

$$\Delta x \sum_{i=I_1}^{I_2} \mathrm{U}_i^0 = \int_{x_{I_1-1/2}}^{x_{I_2+1/2}} \mathrm{u}(x,0)\mathrm{d}x,$$

which will hold for example if we take initial values $\mathrm{U}_i^0 = \bar{\mathrm{u}}_i^0$ as defined in (4.3). From this and (4.21) with $a = 0$ and $b = n\Delta t$ it follows that

$$\Delta x \sum_{i=I_1}^{I_2} \mathrm{U}_i^n = \int_{x_{I_1-1/2}}^{x_{I_2+1/2}} \mathrm{u}(x,t^n)\mathrm{d}x,$$

for all n small enough to ensure that the solution remains constant in the neighbourhood of $x_{I_1-1/2}$ and $x_{I_2+1/2}$. Using the piecewise constant function $\mathrm{U}_{\Delta t}$ as defined in (4.4), gives

$$\int_{x_{I_1-1/2}}^{x_{I_2+1/2}} \mathrm{U}_{\Delta t}(x,t^n)\mathrm{d}x = \int_{x_{I_1-1/2}}^{x_{I_2+1/2}} \mathrm{u}(x,t^n)\mathrm{d}x. \tag{4.22}$$

Hence, we have *discrete conservation*. If a "numerical shock" is propagating at the wrong speed, then the integral of $\mathrm{U}_{\Delta t}$ is increasing at the wrong rate and (4.22) is violated [56]. Therefore, any shocks we compute must have the correct location. The solution computed with a conservative method might have the shock smeared out, but since the integral (4.22) is correct, it must at least be smeared around the correct location.

4.3. MODIFIED EQUATIONS

A useful technique for studying the behaviour of numerical solutions is to model the difference equation by a differential equation. Of course the difference equation was originally derived by approximating (4.1), but as it turns out, the former approximates a different differential equation to even higher order accuracy sometimes; these are the so-called *modified equations* [28, 39, 56].

The modified equation is derived by a two-step procedure [94]. For the sake of simplicity, we describe this procedure for the scalar, linear convection equation (4.12) only. In this analysis it is assumed that there exists a smooth function $U = U(x,t)$, which is an exact solution of a given finite difference scheme (4.8). The first step is to expand each term of the given finite difference scheme in a Taylor series expansion around $U(x,t)$. Substituting the Taylor series expansions in the scheme gives a partial differential equation which includes an infinite number of both space and time derivatives.

In the second step, all time derivatives appearing in the previously derived equation, are eliminated, with the exception of the $\partial/\partial t$ term. The equation obtained after the second step, is called the modified equation. This two-step procedure is illustrated for the Lax-Wendroff method (4.19).

Example 4.9. For the Lax-Wendroff method (4.19), the first step gives the differential equation (using $\Delta t = \sigma \Delta x / a$)

$$\begin{aligned}\frac{\partial}{\partial t}U(x,t) + a\frac{\partial}{\partial x}U(x,t) + \frac{\sigma}{2a}\Delta x\frac{\partial^2}{\partial t^2}U(x,t) - \frac{a\sigma}{2}\Delta x\frac{\partial^2}{\partial x^2}U(x,t) \\ + \frac{\sigma^2}{6a^2}\Delta x^2\frac{\partial^3}{\partial t^3}U(x,t) + \frac{a}{6}\Delta x^2\frac{\partial^3}{\partial x^3}U(x,t) + \frac{\sigma^3}{24a^3}\Delta x^3\frac{\partial^4}{\partial t^4}U(x,t) \\ - \frac{a\sigma}{24}\Delta x^3\frac{\partial^4}{\partial x^4}U(x,t) + \cdots = 0.\end{aligned} \tag{4.23}$$

In the second step we have to eliminate all time derivatives appearing in (4.23), with the exception of the $\partial/\partial t$ term. Suppose that we want to eliminate, for example, the term $\partial^2 U/\partial t^2$ in this equation. Therefore, the operator $-(\Delta t/2)\partial/\partial t$ is applied to (4.23), and the result is added to (4.23). The resulting equation has a term $-(\sigma \Delta x/2)\partial^2 U/\partial t \partial x$ which, in turn, can be eliminated by applying the operator $(\sigma \Delta x/2)\partial/\partial x$ to equation (4.23) and adding the result to the new equation. Similarly, the other time derivatives appearing in (4.23) can be removed. Finally, the following equation is obtained

$$\frac{\partial}{\partial t}U(x,t) + a\frac{\partial}{\partial x}U(x,t) = \frac{a}{6}(\sigma^2 - 1)\Delta x^2\frac{\partial^3}{\partial x^3}U(x,t) + \mathcal{O}(\Delta x^3). \tag{4.24}$$

Equation (4.24) is called the modified equation of the Lax-Wendroff method [39, 56]. □

In general, for a given finite difference scheme corresponding with (4.12), the procedure described above provides the modified equation

$$\frac{\partial}{\partial t}U(x,t) + a\frac{\partial}{\partial x}U(x,t) = \sum_{j=p+1}^{\infty} \mu(j)\Delta x^{j-1}\frac{\partial^j}{\partial x^j}U(x,t),$$

for a method of order p. The coefficients $\mu(j)$ appearing in the sum denote the coefficients of the jth spatial derivatives. These derivatives do not occur in the original partial differential equation and constitute a form of discretization error introduced by the finite difference method.

Example 4.10. The modified equation of the basic upwind method (4.14) for the scalar convection equation (4.12) is given by [36, 39, 56]

$$\frac{\partial}{\partial t}U(x,t) + a\frac{\partial}{\partial x}U(x,t) = \frac{|a|}{2}(1 - |\sigma|)\Delta x\frac{\partial^2}{\partial x^2}U(x,t) + \mathcal{O}(\Delta x^2).$$

□

For a first order method like the basic upwind method, the modified equation is a *convection-diffusion equation* of the form

$$\frac{\partial}{\partial t}U(x,t) + a\frac{\partial}{\partial x}U(x,t) = \mu\Delta x\frac{\partial^2}{\partial x^2}U(x,t) + \mathcal{O}(\Delta x^2), \tag{4.25}$$

with a diffusion coefficient $\mu\Delta x$. The quantity $\mu\Delta x\ \partial^2 U(x,t)/\partial x^2$ is called the ***numerical diffusion*** of the scheme. To study the behaviour of the numerical solution of these two methods, the solution U is developed in a Fourier series. Since linear schemes with constant coefficients are considered, it is sufficient to consider only a single Fourier mode of this series

$$U^{\xi\omega}(x,t) := \exp(i(\xi x + \omega t)), \tag{4.26}$$

where ξ is called the *wavenumber*, ω is called the *frequency* and $i^2 := -1$. Note that if ξ is large, then (4.26) is a highly oscillatory Fourier mode. These highly oscillatory Fourier modes appear around discontinuities. Furthermore, if $U^{\xi\omega}$ satisfies (4.12), then we obtain the dispersion relation $\omega(\xi) = -a\xi$. Substitution of the Fourier mode into the modified equation (4.25) gives

$$\omega(\xi) = -a\xi + \mu\Delta x i\xi^2.$$

Note that $\mu > 0$ is a necessary condition for stability. If $\mu > 0$ (i.e. if $|\sigma| < 1$ for the basic upwind method), then especially the highly oscillatory Fourier modes at $t = 0$ are damped as time evolves. Hence, it is expected that the solution of the modified equation is smeared out as time evolves (see Figure 4.1). This indicates why the basic upwind method approximate discontinuities in solutions too smooth. In general, first order methods have the disadvantage to smear out the solution around discontinuities [39, 56].

For a second order method like the Lax-Wendroff method, the modified equation is a *dispersive equation* of the form (see 4.24)

$$\frac{\partial}{\partial t}U(x,t) + a\frac{\partial}{\partial x}U(x,t) = \mu\Delta x^2\frac{\partial^3}{\partial x^3}U(x,t) + \mathcal{O}(\Delta x^3). \tag{4.27}$$

The quantity $\mu\Delta x^2\ \partial^3 U(x,t)/\partial x^3$ is called the *numerical dispersion* of the scheme. Using the same arguments as for the convection-diffusion equation, again a single Fourier mode (4.26) is considered. If this mode is substituted into the modified equation (4.27), it is seen that

$$\omega(\xi) = -a\xi - \mu\Delta x^2\xi^3.$$

Suppose that $a > 0$. If $\mu < 0$ (i.e. if $|\sigma| < 1$ for the Lax-Wendroff method), then $\omega(\xi) > -a\xi$. Hence, the (highly oscillatory) Fourier modes propagate with a numerical speed less than the exact speed a. Thus, oscillations occur behind the discontinuity (see Figure 4.1). If $\mu > 0$, then $\omega(\xi) < -a\xi$, and the (highly oscillatory) Fourier modes travel too fast. Thus, the oscillations are ahead of the discontinuity (for instance if the Beam-Warming method is used [56]). Most standard second order methods produce oscillations around discontinuities.

4.4. THE NUMERICAL ENTROPY CONDITION

Suppose that a conservative numerical method converges to some function $\mathbf{u}$. The Lax-Wendroff theorem (Theorem 4.5) does not guarantee that $\mathbf{u}$ satisfies the entropy condition (3.18) with $\mathbf{q} = \mathbf{0}$, and there are many examples of conservative numerical methods which converge to weak solutions and yet violate the entropy condition. In this section we look for conditions which guarantee that the limit function u is an entropy solution of the conservation law. Therefore, a numerical variant of Definition 3.7 with $\mathbf{q} = 0$ is given [28, 36, 56].

Definition 4.11. *Let Ψ be a function of the values of* U *at* $2k$ *points, i.e.*

$$\Psi_{i+1/2} = \Psi(\mathrm{U}^n_{i+k}, \dots, \mathrm{U}^n_{i-k+1}).$$

Ψ is called the numerical entropy flux. *It is assumed that the numerical entropy flux is consistent with the entropy flux ψ, i.e. Ψ is Lipschitz continuous and*

$$\Psi(\mathrm{u}, \dots, \mathrm{u}) = \psi(\mathrm{u}).$$

Then a conservative numerical method is called an entropy consistent method *if, for all convex entropy functions η and corresponding entropy fluxes ψ, the inequality*

$$\eta(\mathrm{U}^{n+1}_i) \le \eta(\mathrm{U}^n_i) - \tau(\Psi_{i+1/2} - \Psi_{i-1/2}) \tag{4.28}$$

is satisfied.

In the next section we will study an upwind method for which the discrete entropy inequality (4.28) can be easily proved. The following theorem shows the importance of the concepts which are introduced in this section. It is a simple extension of the Lax-Wendroff theorem and is proved in [28, 36].

Theorem 4.12. *Consider the conservative finite difference method* (4.5) *and suppose that all the assumptions of Theorem 4.5 hold. Furthermore, suppose that* (4.5) *is an entropy consistent method. Then the weak solution* u *is an entropy solution of* (4.1).

Note that we need the entropy condition to exclude nonphysical expansion shocks, across which the entropy jumps to a lower value. If a certain amount of numerical diffusion is added to the numerical method around discontinuities, then entropy consistent methods are obtained, at the cost of smearing out the physical discontinuities [61, 89]. Hence, expansion shocks are excluded. In the next chapter we discuss high resolution methods. The main idea behind these methods is to attempt to use a high order method, but to modify the method and increase the amount of numerical diffusion around discontinuities. We expect that these methods do not produce expansion shocks. In the scalar case, there exists an easier requirement for a numerical method to converge to the entropy solution [28, 67].

Definition 4.13. *Consider a conservative, $(2k+1)$-point finite difference method with* 2 *time levels, which is consistent with the scalar conservation law* (4.1). *If the corresponding numerical flux $F_{i+1/2}$ of the method satisfies*

$$\operatorname{sgn}(U_{i+1} - U_i)(F_{i+1/2} - f(u)) \le 0, \tag{4.29}$$

for all u between U_i and U_{i+1}, then the method is called an E-method.

The following theorem is proved by Osher in [67] and clarifies why E-methods are useful.

Theorem 4.14. *Suppose that the conservative difference method* (4.5) *is consistent with the conservation law* (4.1). *If the method is an E-method, then the method is convergent in the sense of bounded, L_1^{loc}-convergence and its limit is the unique entropy solution of the scalar conservation law* (4.1).

E-methods have the following important disadvantage [28, 67].

Theorem 4.15. *An E-method is consistent of at most order one.*

4.5. GODUNOV-TYPE METHODS

4.5.1. Introduction

An important class of numerical methods for hyperbolic conservation laws are the *Godunov-type methods*. Godunov suggests to solve Riemann problems forward in time. Solutions of Riemann problems are relatively easy to compute, give substantial information about the characteristic structure and lead to conservative methods, since they are themselves exact solutions of the conservation laws and hence conservative. The previous reasons are, among others, important reasons why Godunov methods are quite popular. In first order Godunov-type methods, the numerical solution is assumed piecewise constant in each mesh cell $[x_{i-1/2}, x_{i+1/2})$ and at each time level $t^n = n\Delta t$. The evolution of the solution to the next time level t^{n+1} results from the wave interactions originating at the boundaries between adjacent cells. The cell interface at $x_{i+1/2}$ separates two constant states $\mathbf{U}_i$ at the left and $\mathbf{U}_{i+1}$ at the right hand side, thus the resulting local interaction can be resolved exactly, since the initial conditions at the time t^n correspond to the Riemann problem (3.20). As was shown in Section 3.3, this problem has an exact solution consisting of constant states separated by shocks, contact discontinuities or rarefaction waves (see Theorem 3.13). The new piecewise constant approximation at time t^{n+1} is then obtained by averaging over each cell, the exact solution of the Riemann problem.

However, the computational costs to obtain this exact solution are high in general [28, 56, 81]. Therefore, *approximate Riemann solutions* are considered in order to reduce the computational work. The *approximate Riemann solver* to be described in this section is developed by Roe [74]. Another popular approximate Riemann solver is introduced by Osher [69]. However, we will restrict ourselves to Roe's solver.

Since we shall apply the theory of Section 3.3, it is assumed that each eigenvector is either linearly degenerate or genuinely nonlinear.

Only conservative methods are considered. Such methods are completely determined by their numerical flux function (see (4.5)). Therefore, we restricts ourselves to the computation of the numerical flux function for all the considered methods.

4.5.2. Basic Godunov Method

Three steps are involved in the *basic Godunov method* in order to calculate the numerical solution at time level t^{n+1} from the known numerical solution at time level t^n [40].

In the first step the numerical solution $\mathbf{U}_i^n$ is used to define a piecewise constant function $\tilde{\mathbf{U}}^n$ by

$$\tilde{\mathbf{U}}^n(x,t^n) := \mathbf{U}_i^n, \quad \forall x \in [x_{i-1/2}, x_{i+1/2}). \tag{4.30}$$

At time t^n this function is equal to the piecewise constant function $\mathbf{U}_{\Delta t}$ which has already been introduced in (4.4). Unlike $\mathbf{U}_{\Delta t}$, the function $\tilde{\mathbf{U}}_i^n$ will not be constant for $t^n \leq t < t^{n+1}$. Because of the piecewise constant approximation of $\mathbf{u}$, the Godunov method is first order accurate in space.

In the second step we use $\tilde{\mathbf{U}}^n(\cdot, t^n)$ as initial data for the conservation law

$$\frac{\partial}{\partial t}\tilde{\mathbf{U}}^n(x,t) + \frac{\partial}{\partial x}\mathbf{f}(\tilde{\mathbf{U}}^n(x,t)) = \mathbf{0}, \tag{4.31}$$

which we solve exactly to obtain $\tilde{\mathbf{U}}^n(\cdot, t)$ for $t^n < t \leq t^{n+1}$. This initial value problem can be solved exactly over a short time interval because the initial data $\tilde{\mathbf{U}}^n(\cdot, t^n)$ is piecewise constant, and hence defines a sequence of Riemann problems. The exact solution, up to the first time when two waves from neighbouring Riemann problems interact, is obtained by simply "piecing together" these Riemann solutions. Hence, in the second step we compute the solution of the local Riemann problem at each cell interface. Let the Riemann problem at the cell interface $x_{i+1/2}$ be given by (see (3.20))

$$\frac{\partial}{\partial t}\mathbf{v}_{i+1/2}^n(x,t) + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{v}_{i+1/2}^n(x,t)) = \mathbf{0} \tag{4.32a}$$

with initial data

$$\mathbf{v}_{i+1/2}^n(x,t^n) = \begin{cases} \mathbf{U}_i^n, & x < x_{i+1/2}, \\ \mathbf{U}_{i+1}^n, & x > x_{i+1/2}. \end{cases} \tag{4.32b}$$

Note that $\mathbf{v}_{i+1/2}^n(x,t^n) = \tilde{\mathbf{U}}^n(x,t^n)$ for all $x \in [x_{i-1/2}, x_{i+3/2})$. Let the solution of (4.32) be denoted by (see Section 3.3)

$$\mathbf{v}_{i+1/2}^n(x,t) =: \mathbf{v}^{(R)}\left(\frac{x - x_{i+1/2}}{t - t^n}; \mathbf{U}_i^n, \mathbf{U}_{i+1}^n\right), \tag{4.33}$$

for all $t > t^n$. We assume that adjacent Riemann problems do not interfere as $t^n \leq t \leq t^{n+1}$. If the inequality

$$\Delta t |\lambda|_{\max} < \tfrac{1}{2}\Delta x$$

holds, where $|\lambda|_{\max} = \max(|\lambda_1|, |\lambda_2|, \ldots, |\lambda_m|)$, then this assumption is fulfilled [40] and so $\tilde{\mathbf{U}}^n$ is given by

$$\tilde{\mathbf{U}}^n(x,t) = \mathbf{v}_{i+1/2}^n(x,t), \quad \forall (x,t) \in [x_i, x_{i+1}) \times [t^n, t^{n+1}]. \tag{4.34}$$

Finally, in the third step the approximate solution $\mathbf{U}_i^{n+1}$ at time level t^{n+1} is defined by averaging the exact solution $\tilde{\mathbf{U}}^n$ at time t^{n+1}, thus

$$\mathbf{U}_i^{n+1} := \frac{1}{\Delta x} \int\limits_{x_{i-1/2}}^{x_{i+1/2}} \tilde{\mathbf{U}}^n(x, t^{n+1})\mathrm{d}x. \tag{4.35}$$

Note that in this latter equation two different Riemann problems are involved. Using (4.33) this equation can be rewritten as

$$\mathbf{U}_i^{n+1} = \frac{1}{\Delta x} \int\limits_0^{\Delta x/2} \mathbf{v}^{(R)}(\frac{y}{\Delta t}; \mathbf{U}_{i-1}^n, \mathbf{U}_i^n)\mathrm{d}y + \frac{1}{\Delta x} \int\limits_{-\Delta x/2}^{0} \mathbf{v}^{(R)}(\frac{y}{\Delta t}; \mathbf{U}_i^n, \mathbf{U}_{i+1}^n)\mathrm{d}y \tag{4.36}$$

with respectively $y = x - x_{i-1/2}$ in the first integral, and $y = x - x_{i+1/2}$ in the second integral. The values computed in (4.35) are then used to define a new piecewise constant function $\tilde{\mathbf{U}}^{n+1}$ (see (4.30)) and the procedure is repeated.

The numerical flux $\mathbf{F}^{(G)}$ of the Godunov method can be computed from an integral form of the conservation law (4.31). Since $\tilde{\mathbf{U}}^n$ is the exact solution of (4.31) with initial data (4.30), it is easy to see that

$$\begin{aligned}\int\limits_{x_{i-1/2}}^{x_{i+1/2}} \tilde{\mathbf{U}}^n(x, t^{n+1})\mathrm{d}x &= \int\limits_{x_{i-1/2}}^{x_{i+1/2}} \tilde{\mathbf{U}}^n(x, t^n)\mathrm{d}x + \int\limits_{t^n}^{t^{n+1}} \mathbf{f}(\tilde{\mathbf{U}}^n(x_{i-1/2}, t))\mathrm{d}t \\ &\quad - \int\limits_{t^n}^{t^{n+1}} \mathbf{f}(\tilde{\mathbf{U}}^n(x_{i+1/2}, t))\mathrm{d}t,\end{aligned}$$

by integrating (4.31) in space and time. After dividing by Δx, using (4.30) and (4.35) this equation reduces to

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \tau\left\{\frac{1}{\Delta t}\int\limits_{t^n}^{t^{n+1}} \mathbf{f}(\tilde{\mathbf{U}}^n(x_{i+1/2}, t))\mathrm{d}t - \frac{1}{\Delta t}\int\limits_{t^n}^{t^{n+1}} \mathbf{f}(\tilde{\mathbf{U}}^n(x_{i-1/2}, t))\mathrm{d}t\right\}.$$

Hence, the Godunov method can be written in the conservative form (4.5) with the numerical flux

$$\mathbf{F}_{i+1/2}^{(G)} = \mathbf{F}_{i+1/2}^{(G)}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n) := \frac{1}{\Delta t}\int\limits_{t^n}^{t^{n+1}} \mathbf{f}(\tilde{\mathbf{U}}^n(x_{i+1/2}, t))\mathrm{d}t.$$

Using (4.33) and (4.34), it is easy to see that the integrand in the above equation is independent of t. Therefore, the numerical flux can be rewritten as

$$\mathbf{F}_{i+1/2}^{(G)} = \mathbf{f}(\mathbf{v}^{(R)}(0; \mathbf{U}_i^n, \mathbf{U}_{i+1}^n)). \tag{4.37}$$

If every Riemann solution $\mathbf{v}^{(R)}$ is an entropy solution, then it can be shown that the Godunov method is an entropy consistent method [36, 56]. From (4.37) it directly follows that the Godunov method is consistent. Thus, all hypotheses of Theorem 4.12 are satisfied and therefore, if the method is convergent in the sense of bounded, L_1^{loc}-convergence, then the limit is an entropy solution of (4.1).

Example 4.16. In this example we consider Burgers' equation as described in Example 3.2. Using Section 3.3, it is easy to see that the corresponding Riemann problem has the following solution. If $u_L < u_R$, then the solution is a rarefaction wave which is given by

$$u(x,t) = u^{(R)}(\frac{x}{t}; u_L, u_R) := \begin{cases} u_L, & x/t < u_L, \\ x/t, & u_L < x/t < u_R, \\ u_R, & u_R < x/t. \end{cases} \tag{4.38}$$

In Example 3.5 a rarefaction wave is given with $u_L = -1$ and $u_R = 1$ (see (3.12)). If $u_L > u_R$, then the solution is a shock wave propagating with speed $s = \frac{1}{2}(u_L + u_R)$ (see (3.10)). This solution is given by

$$u(x,t) = u^{(R)}(\frac{x}{t}; u_L, u_R) := \begin{cases} u_L, & x/t < s, \\ u_R, & x/t > s. \end{cases} \tag{4.39}$$

Note that the scalar flux function in (3.4a) is $f(u) = \frac{1}{2}u^2$. Using this and (4.37) it is not difficult to see that Godunov's numerical flux is given by [40, 53]

$$F_{i+1/2}^{(G)} = \begin{cases} \frac{1}{2}(U_{i+1}^n)^2, & U_i^n < 0,\ U_{i+1}^n < 0, \\ \frac{1}{2}(U_i^n)^2, & U_i^n > 0,\ U_{i+1}^n > 0. \end{cases} \tag{4.40a}$$

If U_i^n and U_{i+1}^n have opposite signs, then the numerical flux is given by

$$F_{i+1/2}^{(G)} = \begin{cases} 0, & U_i^n < 0 < U_{i+1}^n, \\ \frac{1}{2}(U_i^n)^2, & U_i^n > 0 > U_{i+1}^n \text{ and } s_{i+1/2}^n > 0, \\ \frac{1}{2}(U_{i+1}^n)^2, & U_i^n > 0 > U_{i+1}^n \text{ and } s_{i+1/2}^n < 0, \end{cases} \tag{4.40b}$$

where $s_{i+1/2}^n := \frac{1}{2}(U_i^n + U_{i+1}^n)$ is the propagation speed of the shock wave. □

4.5.3. Roe's Method

The basic Godunov method requires the solution of Riemann problems (4.32) at every cell interface at each time step. Although in theory these Riemann problems can be solved exactly, doing so in practice is expensive and typically requires some iterative method for solving nonlinear equations [79].

A popular approach to decrease the computational costs of the basic Godunov method is to solve an approximate Riemann problem at the cell interfaces instead of (4.32). Note that in the basic Godunov method $\mathbf{U}_i^{n+1}$ is defined by averaging the exact solution $\tilde{\mathbf{U}}^n$ at time t^{n+1} (see (4.35)). Now $\tilde{\mathbf{U}}^n$ is replaced by an approximate solution $\hat{\mathbf{U}}^n$, which is obtained by "piecing together" approximate Riemann solutions, just as $\tilde{\mathbf{U}}^n$ is defined for the exact Riemann solutions (see (4.34)). Therefore, consider the following Riemann problem at the cell interface $x_{i+1/2}$

$$\frac{\partial}{\partial t}\hat{\mathbf{v}}^n_{i+1/2}(x,t) + \frac{\partial}{\partial x}\hat{\mathbf{f}}(\hat{\mathbf{v}}^n_{i+1/2}(x,t)) = 0 \tag{4.41a}$$

with initial data

$$\hat{\mathbf{v}}^n_{i+1/2}(x,t^n) = \begin{cases} \mathbf{U}^n_i, & x < x_{i+1/2}, \\ \mathbf{U}^n_{i+1}, & x > x_{i+1/2}. \end{cases} \tag{4.41b}$$

Here $\hat{\mathbf{f}}$ is an approximation of $\mathbf{f}$. Let the solution be denoted by (see Section 3.3)

$$\hat{\mathbf{v}}^n_{i+1/2}(x,t) =: \hat{\mathbf{v}}^{(R)}(\frac{x - x_{i+1/2}}{t - t^n}; \mathbf{U}^n_i, \mathbf{U}^n_{i+1}),$$

for all $t > t^n$. If we assume that adjacent Riemann problems do not interfere as $t^n \leq t \leq t^{n+1}$, then the global approximate solution $\hat{\mathbf{U}}^n$ is given by (see (4.34))

$$\hat{\mathbf{U}}^n(x,t) = \hat{\mathbf{v}}^n_{i+1/2}(x,t), \quad \forall\,(x,t) \in [x_i, x_{i+1}) \times [t^n, t^{n+1}].$$

Subsequently the approximate solution $\mathbf{U}^{n+1}_i$ at time level t^{n+1} is defined by averaging $\hat{\mathbf{U}}^n$ at time t^{n+1}, thus

$$\mathbf{U}^{n+1}_i := \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \hat{\mathbf{U}}^n(x, t^{n+1})\mathrm{d}x. \tag{4.42}$$

Since in the latter equation two different Riemann problems are involved, equation (4.42) can be rewritten as (see (4.36))

$$\mathbf{U}^{n+1}_i = \frac{1}{\Delta x}\int_0^{\Delta x/2} \hat{\mathbf{v}}^{(R)}(\frac{y}{\Delta t}; \mathbf{U}^n_{i-1}, \mathbf{U}^n_i)\mathrm{d}y + \frac{1}{\Delta x}\int_{-\Delta x/2}^{0} \hat{\mathbf{v}}^{(R)}(\frac{y}{\Delta t}; \mathbf{U}^n_i, \mathbf{U}^n_{i+1})\mathrm{d}y \tag{4.43}$$

with respectively $y = x - x_{i-1/2}$ in the first integral, and $y = x - x_{i+1/2}$ in the second integral. If the solution $\hat{\mathbf{v}}^n_{i+1/2}$ of the approximate Riemann problem (4.41) has the following property [36, 56]

$$\int_{x_i}^{x_{i+1}} \hat{\mathbf{v}}^n_{i+1/2}(x, t^{n+1})\mathrm{d}x = \frac{1}{2}\Delta x(\mathbf{U}^n_i + \mathbf{U}^n_{i+1}) + \Delta t\mathbf{f}(\mathbf{U}^n_i) - \Delta t\mathbf{f}(\mathbf{U}^n_{i+1}), \tag{4.44}$$

then (4.43) is conservative with the numerical flux function $\hat{\mathbf{F}}$ given by

$$\hat{\mathbf{F}}_{i+1/2} := \mathbf{f}(\mathbf{U}^n_i) - \frac{1}{\Delta t}\int_{x_i}^{x_{i+1/2}} \hat{\mathbf{v}}^{(R)}(\frac{x - x_{i+1/2}}{\Delta t}; \mathbf{U}^n_i, \mathbf{U}^n_{i+1})\mathrm{d}x + \frac{\Delta x}{2\Delta t}\mathbf{U}^n_i. \tag{4.45}$$

It follows from (4.44) and (4.45) that (4.43) is consistent with the conservation law (4.1). In [36] conditions are given such that (4.43) is an entropy consistent method.

A popular approximate Riemann solver is due to Roe [74]. The idea is to determine $\hat{\mathbf{v}}^n_{i+1/2}$ by solving a linear system of conservation laws. Therefore, let $\hat{\mathbf{f}}$ be given by

$$\hat{\mathbf{f}}(\hat{\mathbf{v}}^n_{i+1/2}(x,t)) := \hat{A}(\mathbf{U}^n_i, \mathbf{U}^n_{i+1})\hat{\mathbf{v}}^n_{i+1/2}(x,t),$$

where $\hat{A}(\mathbf{U}^n_i, \mathbf{U}^n_{i+1})$ is a constant $m \times m$-matrix. Thus, the Riemann problem (4.41) can be rewritten as

$$\frac{\partial}{\partial t}\hat{\mathbf{v}}^n_{i+1/2}(x,t) + \hat{A}(\mathbf{U}^n_i, \mathbf{U}^n_{i+1})\frac{\partial}{\partial x}\hat{\mathbf{v}}^n_{i+1/2}(x,t) = 0 \tag{4.46a}$$

$$\hat{\mathbf{v}}^n_{i+1/2}(x,t^n) = \begin{cases} \mathbf{U}^n_i, & x < x_{i+1/2}, \\ \mathbf{U}^n_{i+1}, & x > x_{i+1/2}. \end{cases} \tag{4.46b}$$

Roe requires that the matrix $\hat{A}$ has the following properties:

(i) If $\mathbf{U}^n_i, \mathbf{U}^n_{i+1} \to \bar{\mathbf{u}}$, then $\hat{A}(\mathbf{U}^n_i, \mathbf{U}^n_{i+1}) \to A(\bar{\mathbf{u}})$ smoothly;

(ii) $\hat{A}(\mathbf{U}^n_i, \mathbf{U}^n_{i+1})(\mathbf{U}^n_{i+1} - \mathbf{U}^n_i) = \mathbf{f}(\mathbf{U}^n_{i+1}) - \mathbf{f}(\mathbf{U}^n_i)$;

(iii) $\hat{A}(\mathbf{U}^n_i, \mathbf{U}^n_{i+1})$ is diagonalizable with real eigenvalues.

Condition (i) is necessary to recover the linearized algorithm from the nonlinear version smoothly. Condition (ii) has two effects. First it ensures the method to be conservative (i.e. (4.44) is satisfied) and secondly, in the special case that $\mathbf{U}^n_i$ and $\mathbf{U}^n_{i+1}$ are connected by a single shock wave or contact discontinuity, the approximate Riemann solution agrees with the exact Riemann solution [56, 74]. Finally, condition (iii) is clearly required for the problem to be hyperbolic and solvable.

Instead of the original Riemann problem (4.32), Roe considers a linear Riemann problem (4.46), hence, the approximate Riemann solver recognizes only contact discontinuities [40]. Moreover, condition (i) guarantees that the method behaves reasonably for smooth solutions, since $\|\mathbf{U}^n_{i+1} - \mathbf{U}^n_i\| = \mathcal{O}(\Delta x)$ implies that the linearized equation $(\mathbf{v}^n_{i+1/2})_t + A(\mathbf{U}^n_{i+1})(\mathbf{v}^n_{i+1/2})_x = 0$ is approximately valid. It is natural to require that the linear system (4.46a) agrees with the linearization in this case. Since (ii) guarantees that the method behaves reasonably around an isolated discontinuity, it is only when a Riemann problem has a solution with more than one strong shock or contact discontinuity that the approximate Riemann problem will differ significantly from the exact solution. In practice this happens infrequently (near points where two shocks collide, for example).

It is very easy to construct $\hat{A}(\mathbf{U}^n_i, \mathbf{U}^n_{i+1})$ such that condition (i) is satisfied [74]. Condition (iii) can be easily checked a posteriori. The difficulty arises entirely from condition (ii). In [36] it is shown that for a general system with an entropy function, a complicated averaging of the Jacobian matrix can be used for $\hat{A}$. This shows that a matrix $\hat{A}$ exists which satisfies the conditions (i)-(iii), but unfortunately, it appears that the computed matrix is too complicated to use in practice. Fortunately, for special systems of equations it is possible to derive suitable matrices which are very efficient to use relative to the exact Riemann solution. For example in [74] a suitable matrix $\hat{A}$ is derived for the Euler equations and in [56] a matrix is derived for the isothermal Euler equations. In the following it is assumed that a matrix $\hat{A}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$ exists satisfying conditions (i)-(iii).

Condition (iii) implies that there exists a real diagonal matrix $\hat{\Lambda}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$ and a nonsingular real matrix $\hat{R}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$ such that

$$\hat{A}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)\hat{R}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n) = \hat{R}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)\hat{\Lambda}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n).$$

Here $\hat{\Lambda}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$ is the matrix of the eigenvalues of $\hat{A}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$, where the eigenvalues are labelled in increasing order, and $\hat{R}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$ is the matrix of the corresponding right eigenvectors of $\hat{A}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$. Hence,

$$\begin{aligned}\hat{\Lambda}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n) &:= \operatorname{diag}(\hat{\lambda}_1(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n), \hat{\lambda}_2(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n), \ldots, \hat{\lambda}_m(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)),\\ \hat{R}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n) &:= (\hat{\mathbf{r}}^{(1)}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n), \hat{\mathbf{r}}^{(2)}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n), \ldots, \hat{\mathbf{r}}^{(m)}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)).\end{aligned}$$

For shortness of notation, $\hat{A}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$, $\hat{\lambda}_k(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$ and $\hat{\mathbf{r}}^{(k)}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n)$ will be denoted by $\hat{A}_{i+1/2}$, $\hat{\lambda}_{k,i+1/2}$ and $\hat{\mathbf{r}}_{i+1/2}^{(k)}$. For all k with $1 \leq k \leq m$, $\hat{\lambda}_{k,i+1/2}^{+}$ and $\hat{\lambda}_{k,i+1/2}^{-}$ are defined by

$$\hat{\lambda}_{k,i+1/2}^{+} = \max(\hat{\lambda}_{k,i+1/2}, 0) \geq 0 \quad \text{and} \quad \hat{\lambda}_{k,i+1/2}^{-} = \min(\hat{\lambda}_{k,i+1/2}, 0) \leq 0. \tag{4.47}$$

Since all eigenvectors are linearly independent, the initial states $\mathbf{U}_i^n$ and $\mathbf{U}_{i+1}^n$ of (4.46) can be decomposed as

$$\mathbf{U}_{i+1}^n - \mathbf{U}_i^n = \sum_{k=1}^{m} \hat{\alpha}_{k,i+1/2}\hat{\mathbf{r}}_{i+1/2}^{(k)}, \tag{4.48}$$

where $\hat{\alpha}_{k,i+1/2} \in I\!R$ for all k with $1 \leq k \leq m$. The solution of (4.46) is then given by (analogously to (3.32))

$$\hat{\mathbf{v}}^{(R)}(x,t) = \mathbf{U}_i^n + \sum_{k=1}^{m} \hat{\alpha}_{k,i+1/2} H(x - x_{i+1/2} - \hat{\lambda}_{k,i+1/2}(t - t^n))\hat{\mathbf{r}}_{i+1/2}^{(k)},$$

for all $t > t^n$. After substituting this solution into (4.45), Roe's numerical flux is derived

$$\mathbf{F}_{i+1/2}^{(R)} = \mathbf{F}_{i+1/2}^{(R)}(\mathbf{U}_i^n, \mathbf{U}_{i+1}^n) := \mathbf{f}(\mathbf{U}_i^n) + \sum_{k=1}^{m} \hat{\lambda}_{k,i+1/2}^{-}\hat{\alpha}_{k,i+1/2}\hat{\mathbf{r}}_{i+1/2}^{(k)}. \tag{4.49}$$

Roe's method may include a physically inadmissible expansion shock. This is a direct consequence of the admission of an expansion shock in the underlying approximate Riemann solution. Harten and Hyman proposed a modification of the numerical flux function for a transonic expansion which excludes expansion shocks [28, 40, 53].

Example 4.17. In this example again Burgers' equation (3.4a) is considered, with a solution given by (4.38) or (4.39). Let $s^n_{i+1/2} := \frac{1}{2}(U^n_i + U^n_{i+1})$ (see Example 4.16). Note that $A(u) = u$ and let $\hat{A}_{i+1/2}$ be defined by

$$\hat{A}_{i+1/2} := \tfrac{1}{2}(U^n_i + U^n_{i+1}) = s^n_{i+1/2}.$$

It is easy to see that the conditions (i)-(iii) are satisfied. For $f(u) = \frac{1}{2}u^2$ equation (4.49) becomes

$$F^{(R)}_{i+1/2}(U^n_i, U^n_{i+1}) = \tfrac{1}{2}(U^n_i)^2 + (s^n_{i+1/2})^-(U^n_{i+1} - U^n_i),$$

where $(s^n_{i+1/2})^- = \min(s^n_{i+1/2}, 0) \leq 0$. Using this, it is not difficult to see that Roe's numerical flux for Burgers' equation is given by [40, 53]

$$F^{(R)}_{i+1/2} = \begin{cases} \frac{1}{2}(U^n_{i+1})^2, & s^n_{i+1/2} < 0, \\ \frac{1}{2}(U^n_i)^2, & s^n_{i+1/2} > 0. \end{cases} \tag{4.50}$$

The numerical flux function deviates from the Godunov flux function (4.40) only in the case of a transonic expansion wave. In [53] it is shown that in this case Roe's method replaces the transonic expansion wave by an expansion shock. □

Example 4.18. In the last example of this section Roe's method is applied to the shock tube problem for the one-dimensional nonreactive Euler equations (i.e. (2.26) with $w = 0$). For every pair $(\mathbf{U}^n_i, \mathbf{U}^n_{i+1})$ the matrix $\hat{A}_{i+1/2}$ is given by [74]

$$\hat{A}_{i+1/2} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{2}(\gamma - 3)\hat{u}^2 & (3-\gamma)\hat{u} & \gamma - 1 & -(\gamma - 1)Q \\ \hat{u}(\frac{1}{2}(\gamma - 1)\hat{u}^2 - \hat{H}) & \hat{H} - (\gamma - 1)\hat{u}^2 & \gamma\hat{u} & -(\gamma - 1)Q\hat{u} \\ -\hat{u}\hat{Y} & \hat{Y} & 0 & \hat{u} \end{pmatrix}, \tag{4.51}$$

where the quantities $\hat{u}$, $\hat{H}$ and $\hat{Y}$ are defined as

$$\begin{aligned} \hat{u} &= \frac{(u\sqrt{\rho})^n_{i+1} + (u\sqrt{\rho})^n_i}{\sqrt{\rho^n_{i+1}} + \sqrt{\rho^n_i}}, \quad \hat{H} = \frac{(H\sqrt{\rho})^n_{i+1} + (H\sqrt{\rho})^n_i}{\sqrt{\rho^n_{i+1}} + \sqrt{\rho^n_i}}, \\ \hat{Y} &= \frac{(Y\sqrt{\rho})^n_{i+1} + (Y\sqrt{\rho})^n_i}{\sqrt{\rho^n_{i+1}} + \sqrt{\rho^n_i}}. \end{aligned} \tag{4.52}$$

Note that (4.51) shows a very remarkable result, the matrix $\hat{A}_{i+1/2}$ is identical to the Jacobian matrix $A(\mathbf{u})$ given in (3.7). In order to derive the eigenvectors and the eigenvalues of the matrix $\hat{A}_{i+1/2}$ the following quantity is useful

$$\hat{c}^2 = (\gamma - 1)(\hat{H} - \frac{\hat{u}^2}{2} - Q\hat{Y}). \tag{4.53}$$

Now the computation of the eigenvalues and the eigenvectors is straightforward. They are given by

$$\hat{\lambda}_{1,i+1/2} = \hat{u} - \hat{c}, \quad \hat{\lambda}_{2,i+1/2} = \hat{u}, \quad \hat{\lambda}_{3,i+1/2} = \hat{u}, \quad \hat{\lambda}_{4,i+1/2} = \hat{u} + \hat{c} \tag{4.54}$$

and

$$\begin{aligned} \hat{\mathbf{r}}^{(1)}_{i+1/2} &= (1\,,\, \hat{u} - \hat{c}\,,\, \hat{H} - \hat{u}\hat{c}\,,\, \hat{Y})^T, \\ \hat{\mathbf{r}}^{(2)}_{i+1/2} &= (1\,,\, \hat{u}\,,\, \tfrac{1}{2}\hat{u}^2\,,\, 0)^T, \\ \hat{\mathbf{r}}^{(3)}_{i+1/2} &= (0\,,\, 0\,,\, Q\,,\, 1)^T, \\ \hat{\mathbf{r}}^{(4)}_{i+1/2} &= (1\,,\, \hat{u} + \hat{c}\,,\, \hat{H} + \hat{u}\hat{c}\,,\, \hat{Y})^T. \end{aligned} \tag{4.55}$$

Hence, for every pair $(\mathbf{U}^n_i, \mathbf{U}^n_{i+1})$ Roe's numerical flux $\mathbf{F}^{(R)}_{i+1/2}$ at the cell interface $x_{i+1/2}$ is derived in three steps. The first step is the computation of the quantities defined in (4.52) and (4.53). In the second step the eigenvalues (4.54) and eigenvectors (4.55) are computed. Finally, in the third step, (4.48) is used to compute $\hat{\alpha}_{1,i+1/2}$, $\hat{\alpha}_{2,i+1/2}$, $\hat{\alpha}_{3,i+1/2}$ and $\hat{\alpha}_{4,i+1/2}$. The computation of Roe's numerical flux (4.49) is now straightforward.

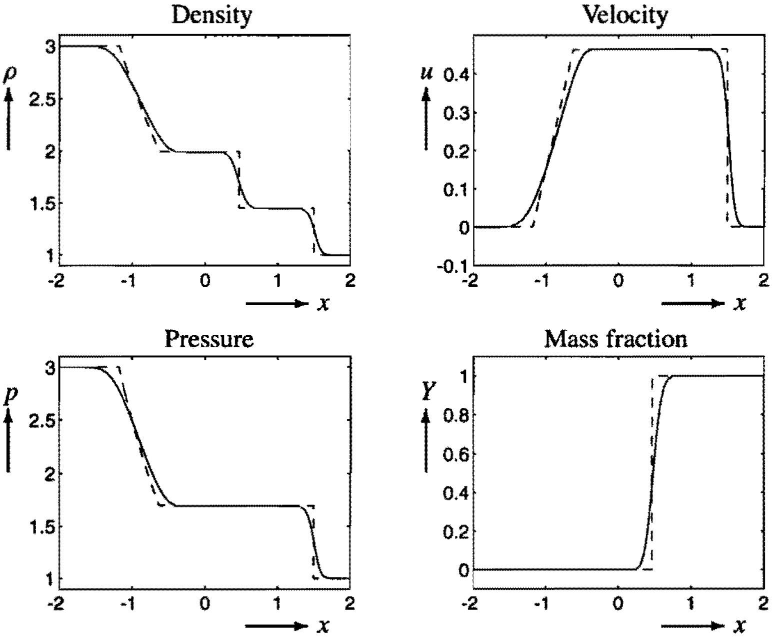

Figure 4.2. Numerical results for Roe's method (4.49); exact solution (dashed line) and numerical solution (solid line) of (2.26) with $w = 0$ at time $t = 1$ with $\Delta t = 0.002$, $\Delta x = 0.02$ and the initial conditions $p(x,0) = 3$, $\rho(x,0) = 3$, $u(x,0) = 0$, $Y(x,0) = 0$ if $x < 0$ and $p(x,0) = 1$, $\rho(x,0) = 1$, $u(x,0) = 0$, $Y(x,0) = 1$ if $x > 0$.

The numerical results in Figure 4.2 clearly illustrate that Roe's method is a first order method, since the discontinuities are smeared out. □

5

HIGH RESOLUTION METHODS

The Lax-Wendroff theorem (Theorem 4.5) does not say anything about convergence of a method. We only know that if a sequence of approximations converges, then the limit is a weak solution. In this section we will present a theory, which under certain assumptions, guarantees convergence of a method. It turns out that one class of convergent methods are the so-called monotone methods. However, it can be shown that monotone methods are consistent of at most order one.

In the previous chapter we observed that first order methods give poor accuracy in smooth regions of the flow. Moreover, discontinuities tend to smear out heavily and are poorly resolved on the grid. These effects are due to a large amount of numerical diffusion in first order methods (see Section 4.3). In contrast, second order methods give good results in smooth regions of the flow. However, oscillations may occur around discontinuities. Hence, some numerical diffusion is still needed to give nonoscillatory discontinuities and to ensure convergence, but first order methods (like monotone methods) tend to diffuse too much.

In this chapter we will study some high resolution methods. This term applies to methods which are at least second order accurate for smooth solutions and yet resolve discontinuities quite well. The main idea behind any high resolution method is to attempt to use a high order method, but to modify the method around discontinuities, thereby increasing the amount of numerical diffusion.

This chapter is organized as follows. In the first section we present a form of nonlinear stability that allows us to prove convergence for a wide class of practical methods. Furthermore, the important monotone and total variation diminishing (TVD) methods are introduced. In Section 5.2 and Section 5.3 two classes of high resolution methods are introduced which are quite popular. We consider flux limiter methods for nonlinear scalar conservation laws in Section 5.2. Finally, slope limiter methods are discussed in Section 5.3. We describe the basic idea for a scalar, linear conservation law and present one possibility to extend the slope limiter method to nonlinear systems of conservation laws. Section 5.3 is concluded with numerical results of a slope limiter method applied to the one-dimensional nonreactive Euler equations.

5.1. MONOTONE AND TVD METHODS FOR SCALAR CONSERVATION LAWS

Until now, we have not discussed whether a method converges. In this section we will give a theory from which convergence can be established [28]. This theory has been completely successful so far however, for scalar problems only. To our knowledge, for general systems of equations with arbitrary initial data, no numerical method has been proved to be convergent, although convergence results have been obtained in some special cases (see e.g. [20]).

Obviously, the scalar case has limited direct applicability to real-world problems. However, many of the most successful numerical methods for systems have been developed by first inventing good methods for the scalar case (where the theory provides good guidance) and then extending them in a relatively straightforward way to systems of equations. The fact that we can prove they work well for scalar problems is no guarantee that they will work at all for systems, but in practice this approach has been very successful.

So, consider the nonlinear scalar conservation law

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}f(u(x,t)) = 0. \tag{5.1}$$

Let $a(u)$ be defined by $a(u) := f'(u)$. To calculate solutions of (5.1) numerically, we consider only conservative, $(2k+1)$-point finite difference methods with 2 time levels, which are consistent with the conservation law (5.1) (see Section 4.1), i.e.

$$U_i^{n+1} = U_i^n - \tau(F_{i+1/2}^n - F_{i-1/2}^n). \tag{5.2}$$

Let the function $U_{\Delta t}$ be defined by (4.4), then the numerical scheme can be written as in (4.8), i.e.

$$U_{\Delta t}(\cdot, t+\Delta t) = \mathcal{L}_{\Delta t}U_{\Delta t}(\cdot, t). \tag{5.3}$$

Let $T > 0$ be a given constant. First some new concepts are introduced. For a given function $u = u(x,t)$ the *total variation* over $[0,T]$ is defined by

$$\begin{aligned} \mathrm{TV}_T(u) \;:=\; & \limsup_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\int_0^T\int_{-\infty}^{+\infty}|u(x+\varepsilon,t)-u(x,t)|\mathrm{d}x\mathrm{d}t \\ & + \limsup_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\int_0^T\int_{-\infty}^{+\infty}|u(x,t+\varepsilon)-u(x,t)|\mathrm{d}x\mathrm{d}t. \end{aligned} \tag{5.4}$$

The total variation over $[0,T]$ of the function $U_{\Delta t}$ is derived after substituting this function into (5.4) with $T = N\Delta t$ for some integer N, which gives

$$\mathrm{TV}_T(U_{\Delta t}) = \sum_{n=0}^{N-1}\sum_{i=-\infty}^{+\infty}\left\{\Delta t|U_{i+1}^n - U_i^n| + \Delta x|U_i^{n+1} - U_i^n|\right\}.$$

Analogously to (5.4), the one-dimensional total variation at time t is defined by

$$\mathrm{TV}(u(\cdot,t)) := \limsup_{\varepsilon\downarrow 0} \frac{1}{\varepsilon} \int_{-\infty}^{+\infty} |u(x+\varepsilon,t) - u(x,t)| \mathrm{d}x. \tag{5.5}$$

The total variation of the function $U_{\Delta t}$ at time t^n is defined by substituting this function into (5.5), which gives

$$\mathrm{TV}(U_{\Delta t}(\cdot,t^n)) = \sum_{i=-\infty}^{+\infty} |U_{i+1}^n - U_i^n|.$$

To guarantee convergence, we need some form of stability. Note that linearity of the numerical method is essential for the stability concept defined in Definition 4.6. In this section we consider two forms of nonlinear stability that allow us to prove convergence results for a wide class of practical methods. Another form of nonlinear stability is introduced by Einfeldt in [23]. We start by introducing an L_∞*-stable method* [28].

Definition 5.1. *Consider a conservative,* $(2k+1)$*-point method written in the generic form* (5.3) *for arbitrary t. The numerical method is called* L_∞-stable, *if for each time $T > 0$ there exists a constant C and a value k_0 such that*

$$\|U_{\Delta t}(\cdot,t^n)\|_{L_\infty(\mathbb{R})} \leq C, \quad \forall n\Delta t \leq T, \ \Delta t < k_0$$

holds.

Another important form of nonlinear stability is *TV-stability* [28, 33].

Definition 5.2. *Consider a conservative,* $(2k+1)$*-point method written in the generic form* (5.3) *for arbitrary t. The numerical method is called* TV-stable, *if for each time $T > 0$ there exists a constant C and a value k_0 such that*

$$\mathrm{TV}(U_{\Delta t}(\cdot,t^n)) \leq C, \quad \forall n\Delta t \leq T, \ \Delta t < k_0 \tag{5.6}$$

holds.

The following theorem shows the importance of the stability concepts which are introduced in this section. The proof is given in [28, 33].

Theorem 5.3. *Suppose that the finite difference method* (5.2) *is an entropy consistent method with a Lipschitz continuous flux function F that satisfies* (4.10). *Let U_i^n be a solution of* (5.2) *with given initial values $U_i^0 = \bar{u}_i^0$, as defined in* (4.3). *Define the piecewise constant function $U_{\Delta t}$ as in* (4.4) *and suppose that the method is L_∞-stable and TV-stable. Then the method is convergent (for $\Delta t \to 0$) in the sense of bounded, L_1^{loc}-convergence and its limit is the unique entropy solution of* (5.1).

In the remainder of this section we assume that (5.1) has initial data $u(x,0) = u^0(x)$ with finite total variation. An easy way to ensure that condition (5.6) is fulfilled, is to require that the total variation is nonincreasing as time evolves, so that the total variation of $U_{\Delta t}$ at any time $t > 0$ is bounded by the total variation of the initial data. This requirement gives rise to the following definition [33, 56].

Definition 5.4. *The numerical method* (5.3) *is called a* total variation diminishing method *(abbreviated TVD method) if*

$$\mathrm{TV}(U_{\Delta t}(\cdot, t^{n+1})) \leq \mathrm{TV}(U_{\Delta t}(\cdot, t^n)),$$

for all grid functions $U_{\Delta t}(\cdot, t^n)$.

An important argument to consider TVD methods is that the exact solution to the scalar conservation law (5.1) has also this TVD property [33, 56]. Any weak solution of (5.1) satisfies

$$\mathrm{TV}(u(\cdot, t^2)) \leq \mathrm{TV}(u(\cdot, t^1)), \quad \forall t^2 \geq t^1.$$

If a TVD method is used, then the following inequalities hold

$$\mathrm{TV}(U_{\Delta t}(\cdot, t^n)) \leq \mathrm{TV}(U_{\Delta t}(\cdot, 0)) \leq \mathrm{TV}(u^0),$$

for all $n \geq 0$. Since the initial function u^0 is assumed to have a finite total variation, (5.6) holds and the method is TV-stable. If the initial data u^0 has a compact support, then the whole sequence U^n has a compact support (see (5.2)). For a numerical solution U^n, we can write

$$-U_i^n = \sum_{j=i}^{+\infty}(U_{j+1}^n - U_j^n)$$

so that

$$\|U_{\Delta t}(\cdot, t^n)\|_{L_\infty(I\!R)} \leq \mathrm{TV}(U_{\Delta t}(\cdot, t^n)) \leq \mathrm{TV}(u^0).$$

Therefore, if the initial data of a TVD method has compact support and bounded total variation, then the method is TV-stable and L_∞-stable and, subsequently, convergent. In [28, 32] some examples of TVD methods are given.

It has been shown earlier that one difficulty associated with numerical approximations of discontinuous solutions is that oscillations may appear near a discontinuity. It can be proved that the exact solution does not have these oscillations. More precisely, if u is a weak solution of the scalar conservation law (5.1) with initial data u^0 with finite total variation, then u has the following *monotonicity preserving* properties as a function of t [32]: (i) no new extrema in x are created; (ii) the value of a local minimum is nondecreasing and the value of a local maximum is nonincreasing. We emphasize that we only consider scalar problems here. Since the exact solution has this property, it seems natural to require that the numerical solution has this same property [28, 56].

Definition 5.5. *The numerical method* (5.3) *is called* monotonicity preserving *if the following statement holds. If* u^0 *is monotone (either nonincreasing or nondecreasing), then* $U_{\Delta t}(\cdot, t)$ *is also monotone for all* $t > 0$.

In [28, 81] it is shown that a linear $(2k+1)$-point finite difference scheme

$$U_i^{n+1} = \sum_{j=-k}^{k} \alpha_j U_{i+j}^n$$

is monotonicity preserving if and only if $\alpha_j \geq 0$ for all j with $-k \leq j \leq k$.

Another useful property of the entropy solution of (5.1) is given by the following theorem [28, 46].

Theorem 5.6. *Suppose that u and v are two entropy solutions of* (5.1). *If $u(\cdot,0) - v(\cdot,0) \in L_1(\mathbb{R})$ and if $u(\cdot,0) - v(\cdot,0)$ has bounded total variation, then*

$$\|u(\cdot,t^2) - v(\cdot,t^2)\|_1 \leq \|u(\cdot,t^1) - v(\cdot,t^1)\|_1, \tag{5.7}$$

for all t^1, t^2 with $t^2 \geq t^1 \geq 0$. Here $\|\cdot\|_1$ denotes the L_1-norm in the space variable.

The property (5.7) is called *L_1-contraction*. In analogy to this, an *L_1-contracting numerical method* is defined as follows [56].

Definition 5.7. *The numerical method* (5.3) *is called* L_1-contracting *if, for any two functions $U_{\Delta t}(\cdot,t^n)$ and $V_{\Delta t}(\cdot,t^n)$ satisfying* (5.3) *for which $U_{\Delta t}(\cdot,t^n) - V_{\Delta t}(\cdot,t^n)$ has compact support, the following inequality holds:*

$$\|U_{\Delta t}(\cdot,t^{n+1}) - V_{\Delta t}(\cdot,t^{n+1})\|_1 \leq \|U_{\Delta t}(\cdot,t^n) - V_{\Delta t}(\cdot,t^n)\|_1.$$

The last property of the entropy solution of (5.1) that is used is the following [49].

Theorem 5.8. *If u and v are two entropy solutions of* (5.1) *with initial data satisfying $v^0(x) \geq u^0(x)$ for all x, then the solutions u and v satisfy $v(x,t) \geq u(x,t)$ for all x and $t > 0$.*

A numerical method which has the same property is called a *monotone method* and is defined as follows [28, 56].

Definition 5.9. *The numerical method* (5.3) *is called* monotone *if the following statement holds*

$$V_{\Delta t}(x,t^n) \geq U_{\Delta t}(x,t^n) \ \forall x \quad \Rightarrow \quad V_{\Delta t}(x,t^{n+1}) \geq U_{\Delta t}(x,t^{n+1}) \ \forall x.$$

To prove the monotonicity of a method, it is sufficient to check whether the difference operator $\mathcal{L}_{\Delta t}$ is a nondecreasing function of each argument [19]. The basic upwind method is an example of a monotone method. In [67] it is shown that every E-method (see Definition 4.13) is monotone.

The relations between all the concepts which are introduced in this section are given by the following theorem [28, 56].

Theorem 5.10. *If the numerical method* (5.3) *is monotone, then it is L_1-contracting. A numerical method* (5.3) *which is L_1-contracting, is always TVD and furthermore, a numerical method* (5.3) *which is TVD, is always monotonicity preserving.*

These relations can be summarized as

$$\text{monotone} \;\Rightarrow\; L_1\text{-contracting} \;\Rightarrow\; \text{TVD} \;\Rightarrow\; \text{monotonicity preserving}.$$

An easy application of Theorem 5.3 and Theorem 5.10 shows that a monotone method converges. Monotone numerical methods have the satisfying property that we do not have to worry about the entropy condition, since a monotone method contains enough numerical diffusion to exclude expansion shocks and, subsequently, converge always to the entropy solution. The following theorem shows this property [28, 35].

Theorem 5.11. *If the numerical method* (5.3) *is monotone, then the method is convergent (for $\Delta t \to 0$) in the sense of bounded, L_1^{loc}-convergence and its limit is the unique entropy solution of* (5.1).

Although the monotonicity requirement is easy to check and monotone methods always converge to the entropy solution, the class of monotone methods is seriously restricted as the following theorem shows [28, 35].

Theorem 5.12. *A monotone numerical method is consistent of at most order one.*

A monotone method is not accurate enough in regions where the solution is smooth. Therefore, TVD methods are used more frequently. To derive a higher order TVD method is not trivial. In [93] it is shown that any 3-point TVD method is at most first order accurate. Thus methods with more than 3 points are required to achieve second order accuracy. Also in [93] a 5-point TVD method is derived, which is entropy consistent and second order accurate in regions where the solution is smooth.

5.2. FLUX LIMITER METHODS FOR SCALAR CONSERVATION LAWS

In the flux limiter approach, we choose a high order flux (e.g. the Lax-Wendroff flux) which works well in regions where the solution is smooth, and a low order flux (e.g. the flux from some monotone method) which behaves well near discontinuities. The main idea is the hybridization of these two flux functions into a single flux in such a way that this single flux reduces to the high order flux in smooth regions and to the low order flux near discontinuities. This idea is elaborated in this section. For details we refer to [88] and references in there.

Let a conservative 3-point method be given, which is consistent with the conservation law (5.1). The corresponding finite difference scheme is given by (see (5.2))

$$U_i^{n+1} = U_i^n - \tau(F_{i+1/2}^{(E)} - F_{i-1/2}^{(E)}), \tag{5.8}$$

where $\tau = \Delta t/\Delta x$ and $F_{i+1/2}^{(E)} = F(U_i^n, U_{i+1}^n)$ denotes the numerical flux of some arbitrary E-method (see Definition 4.13). Recall that every E-method is monotone. We define the following flux differences

$$\begin{aligned}(\delta f^n_{i+1/2})^+ &:= f(U^n_{i+1}) - F^{(E)}_{i+1/2},\\ (\delta f^n_{i+1/2})^- &:= F^{(E)}_{i+1/2} - f(U^n_i).\end{aligned} \tag{5.9}$$

Note that $(\delta f^n_{i+1/2})^+ + (\delta f^n_{i+1/2})^- = f(U^n_{i+1}) - f(U^n_i)$. These flux differences in turn are used to define the local Courant numbers,

$$\begin{aligned}\sigma^+_{i+1/2} &:= \tau\frac{(\delta f^n_{i+1/2})^+}{U^n_{i+1} - U^n_i},\\ \sigma^-_{i+1/2} &:= \tau\frac{(\delta f^n_{i+1/2})^-}{U^n_{i+1} - U^n_i}.\end{aligned} \tag{5.10}$$

It is not difficult to see that the inequality (4.29), which defines an E-method, implies [28, 32, 88]

$$\sigma^-_{i+1/2} \le 0 \le \sigma^+_{i+1/2}. \tag{5.11}$$

Let the following 3-point finite difference method be given to approximate (5.1) numerically,

$$U^{n+1}_i = U^n_i - (D^n_{i+1/2}(U^n_{i+1} - U^n_i) - C^n_{i-1/2}(U^n_i - U^n_{i-1})), \tag{5.12}$$

where $C^n_{i-1/2}$ and $D^n_{i+1/2}$ are data dependent coefficients, i.e. $C^n_{i-1/2} = C(U^n_i, U^n_{i-1})$ and $D^n_{i+1/2} = D(U^n_{i+1}, U^n_i)$. In [32] the following theorem is proved, which gives sufficient conditions for the method above to be TVD.

Theorem 5.13. *If the coefficients in* (5.12) *satisfy*

$$C^n_{i+1/2} \le 0, \qquad D^n_{i+1/2} \le 0, \qquad -(C^n_{i+1/2} + D^n_{i+1/2}) \le 1,$$

then the numerical method (5.12) *is a TVD method.*

From (5.9) it is seen that

$$\begin{aligned}F^{(E)}_{i+1/2} - F^{(E)}_{i-1/2} &= (\delta f^n_{i+1/2})^- + (\delta f^n_{i-1/2})^+\\ &= \frac{1}{\tau}(\sigma^-_{i+1/2}(U^n_{i+1} - U^n_i) + \sigma^+_{i-1/2}(U^n_i - U^n_{i-1})),\end{aligned}$$

and therefore, one possibility of writing a general scheme (5.8) in the form (5.12) is

$$U^{n+1}_i = U^n_i - (\sigma^-_{i+1/2}(U^n_{i+1} - U^n_i) + \sigma^+_{i-1/2}(U^n_i - U^n_{i-1})),$$

i.e. taking $C^n_{i+1/2} = -\sigma^+_{i+1/2}$ and $D^n_{i+1/2} = \sigma^-_{i+1/2}$. Using (5.11) and Theorem 5.13 it is obvious that (5.8) is a TVD method, if it is an E-method and the CFL-like condition

$$\sigma^+_{i+1/2} - \sigma^-_{i+1/2} \le 1 \tag{5.13}$$

is fulfilled.

For clarity of approach we first consider the scalar, linear equation (see (4.12))

$$\frac{\partial}{\partial t}u(x,t) + a\frac{\partial}{\partial x}u(x,t) = 0. \tag{5.14}$$

Hence, the flux function $f : \mathbb{R} \to \mathbb{R}$ is given by $f(u) := au$. Subsequently, the method will be extended for nonlinear equations. Let $F^{(LW)}_{i+1/2}$ and $F^{(BU)}_{i+1/2}$ denote the numerical flux corresponding to, respectively, the Lax-Wendroff method and the basic upwind method, both applied to (5.14) (see (4.16) and (4.20)). Furthermore, let $a > 0$, then it is easy to see that

$$F^{(LW)}_{i+1/2} = F^{(BU)}_{i+1/2} + \tfrac{1}{2}(1-\sigma)(f(U^n_{i+1}) - F^{(BU)}_{i+1/2}),$$

where σ is given by (4.13). Hence, the Lax-Wendroff flux function is composed of the first order basic upwind flux plus an additional flux, which is often called an *antidiffusive flux*. Since it is well-known that the Lax-Wendroff method is not TVD, we try to remedy this by adding only a limited amount of the antidiffusive flux to the first order upwind flux, i.e.

$$F^n_{i+1/2} = F^{(BU)}_{i+1/2} + \varphi(\theta^+_{i+1/2})\tfrac{1}{2}(1-\sigma)(f(U^n_{i+1}) - F^{(BU)}_{i+1/2}), \tag{5.15}$$

where the function φ is called a *limiter*. To detect where the amount of the antidiffusive flux is large, we have to measure the "smoothness" of the data in some way. One possibility is to consider the limiter φ as a function of the following ratio of slopes in the upwind direction

$$\theta^+_{i+1/2} = \frac{U^n_i - U^n_{i-1}}{U^n_{i+1} - U^n_i}. \tag{5.16}$$

We want to choose the limiter φ such that the limited antidiffusive flux is maximised, while the resulting scheme remains TVD. In the above description we assumed $a > 0$. Obviously, a similar method can be defined when $a < 0$, by again considering Lax-Wendroff as a modification of the basic upwind method. We can unify these methods (for $a > 0$ and $a < 0$) into the single formula

$$\begin{aligned} F^n_{i+1/2} &= F^{(BU)}_{i+1/2} + \varphi(\theta^+_{i+1/2})\tfrac{1}{2}(1-\sigma)(f(U^n_{i+1}) - F^{(BU)}_{i+1/2}) \\ &\quad - \varphi(\theta^-_{i+1/2})\tfrac{1}{2}(1+\sigma)(F^{(BU)}_{i+1/2} - f(U^n_i)). \end{aligned} \tag{5.17}$$

Again we consider the limiter φ as a function of the ratio of slopes in the upwind direction, i.e. $\theta^+_{i+1/2}$ $(a > 0)$ is given by (5.16) and $\theta^-_{i+1/2}$ $(a < 0)$ is given by

$$\theta^-_{i+1/2} = \frac{U^n_{i+2} - U^n_{i+1}}{U^n_{i+1} - U^n_i}. \tag{5.18}$$

Now we return to the nonlinear equation (5.1). We take the underlying first order method to be an E-method and add both limited positive and negative antidiffusive fluxes. Using (5.9) and (5.10) we generalize (5.17) as

$$\begin{aligned} F^n_{i+1/2} &= F^{(E)}_{i+1/2} + \varphi(\theta^+_{i+1/2})\tfrac{1}{2}(1-\sigma^+_{i+1/2})(\delta f^n_{i+1/2})^+ \\ &\quad - \varphi(\theta^-_{i+1/2})\tfrac{1}{2}(1+\sigma^-_{i+1/2})(\delta f^n_{i+1/2})^-. \end{aligned} \tag{5.19}$$

To detect where the amount of the antidiffusive flux is large in the nonlinear case, the limiters are considered as functions of the following ratios [40, 88]:

$$\begin{aligned} \theta^+_{i+1/2} &:= \frac{(1-\sigma^+_{i-1/2})(\delta f^n_{i-1/2})^+}{(1-\sigma^+_{i+1/2})(\delta f^n_{i+1/2})^+}, \\ \theta^-_{i+1/2} &:= \frac{(1+\sigma^-_{i+3/2})(\delta f^n_{i+3/2})^-}{(1+\sigma^-_{i+1/2})(\delta f^n_{i+1/2})^-}. \end{aligned}$$

If $f(u) = au$ and $F^{(E)}_{i+1/2} = F^{(BU)}_{i+1/2}$, then (5.19) reduces to (5.17). Thus (5.19) is a generalization of (5.17) for nonlinear conservation laws.

The limiter φ is taken to be nonnegative, so that the sign of the antidiffusive flux is maintained, i.e.

$$\varphi(\theta) \geq 0 \ \forall \theta. \tag{5.20}$$

The numerical method defined by the flux (5.19) is called a *flux limiter method.* An easy calculation, using Taylor series expansions shows that this flux defines a numerical method, which is second order consistent in space if $\varphi = 1$ [28].

If we want to apply Theorem 5.13, then the numerical method given by the flux (5.19), has to be written in the same form as (5.12). One possibility is to take $C^n_{i+1/2}$ and $D^n_{i+1/2}$ as

$$\begin{aligned} C^n_{i+1/2} &= -\sigma^+_{i+1/2}\left\{1 + \tfrac{1}{2}(1-\sigma^+_{i+1/2})\left[\frac{\varphi(\theta^+_{i+3/2})}{\theta^+_{i+3/2}} - \varphi(\theta^+_{i+1/2})\right]\right\}, \\ D^n_{i+1/2} &= \sigma^-_{i+1/2}\left\{1 + \tfrac{1}{2}(1+\sigma^-_{i+1/2})\left[\frac{\varphi(\theta^-_{i-1/2})}{\theta^-_{i-1/2}} - \varphi(\theta^-_{i+1/2})\right]\right\}. \end{aligned}$$

Suppose there exists a constant Φ with $0 < \Phi \leq 2$, such that

$$\left|\frac{\varphi(\theta^+_{i+3/2})}{\theta^+_{i+3/2}} - \varphi(\theta^+_{i+1/2})\right| \leq \Phi \quad \text{and} \quad \left|\frac{\varphi(\theta^-_{i-1/2})}{\theta^-_{i-1/2}} - \varphi(\theta^-_{i+1/2})\right| \leq \Phi. \tag{5.21}$$

If the following CFL-like condition is satisfied (see (5.13))

$$\sigma^+_{i+1/2} - \sigma^-_{i+1/2} \leq \frac{2}{2+\Phi},$$

then all assumptions of Theorem 5.13 are fulfilled and the method is TVD. If in addition to (5.20) it is also required that

$$\varphi(\theta) = 0 \ \forall \theta \leq 0,$$

then the bound (5.21) reduces to

$$0 \leq \frac{\varphi(\theta)}{\theta} \leq \Phi \ \forall \theta \quad \text{and} \quad 0 \leq \varphi(\theta) \leq \Phi \ \forall \theta. \tag{5.22}$$

The last condition on the limiter is given by the following theorem [28, 68].

Theorem 5.14. *The flux limiter method with flux* (5.19) *is consistent with the conservation law* (5.1) *provided* φ *is a bounded function. It is a second order TVD method (on smooth solutions with* $\partial u/\partial x$ *bounded away from zero) provided* φ *satisfies* (5.22), $\varphi(1) = 1$ *and* φ *is Lipschitz continuous at* $\theta = 1$.

Hence the flux limiter method (5.19) is second order and TVD except around extreme points (where the limiter $\varphi = 0$). In [88] it appears that the best choice for φ is a convex combination of 1 and θ, i.e.

$$\varphi(\theta) \;=\; 1 \;+\; \zeta(\theta)(\theta - 1),$$

with $0 \le \zeta(\theta) \le 1$ for all θ. Other choices appear to give too much compression, i.e. smooth initial data such as a sine wave tend to turn into a square wave as time evolves [88]. Note that with this choice of φ the condition $\varphi(1) = 1$ is automatically satisfied.

Example 5.15. Roe chooses $\varphi(\theta)$ as large as possible such that all conditions of Theorem 5.14 are fulfilled. This limiter is called the *superbee limiter* and is given by [56]

$$\varphi(\theta) \;:=\; \max(0,\; \min(1, 2\theta),\; \min(\theta, 2)). \tag{5.23}$$

On the other hand, if we choose $\varphi(\theta)$ as small as possible such that all conditions of Theorem 5.14 are fulfilled, then we obtain Roe's *minmod limiter*, which is given by [56]

$$\varphi(\theta) \;:=\; \max(0,\; \min(\theta, 1)). \tag{5.24}$$

A smoother limiter function is introduced by van Leer [51] and is given by

$$\varphi(\theta) \;:=\; \frac{|\theta| + \theta}{1 + |\theta|}. \tag{5.25}$$

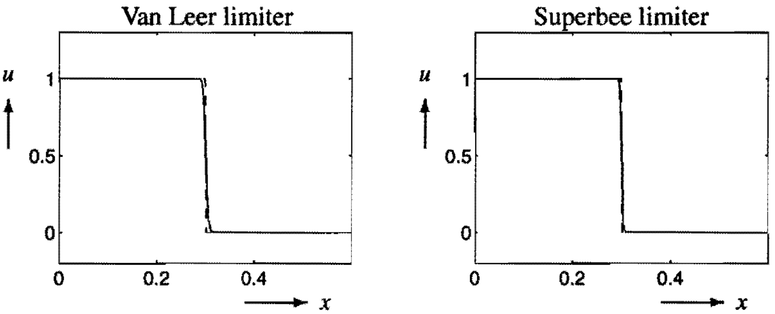

Figure 5.1. Numerical results for method (5.17): exact solution (dashed line) and numerical solution (solid line) of (4.12) at $t = 0.3$ with $a = 1$, $\Delta t = 0.002$, $\Delta x = 0.0025$ and the initial condition $u(x, 0) = 1$ if $x < 0$ and $u(x, 0) = 0$ if $x > 0$.

In Figure 5.1 numerical results are given for the scalar, linear equation (5.14). In these results the underlying E-method is simply the basic upwind method and van Leer's limiter and Roe's superbee limiter are used, respectively. Comparing the results with the results presented in Figure 4.1, we clearly observe less numerical diffusion around shocks for the flux limiter methods. □

In [88] some other examples of limiters and the corresponding numerical results are given. Two questions remain open: under which conditions are flux limiter methods entropy consistent methods and can flux limiter methods be extended to systems of nonlinear conservation laws? The flux limiter method does not necessarily converge to the entropy solution, since for some peculiar data, expansion shocks may occur. A possible remedy consists in adding some extra diffusion when a sonic point occurs in an expansion region [28]. We refer to [68], where a particular flux limiter method is described for nonlinear systems. Further, in the scalar case it is proved, that this method converges to the unique entropy solution. A natural way to generalize a scalar flux limiter method to systems of equations is to linearize the nonlinear system. The generalization is then obtained by diagonalizing the resulting linear system and applying the scalar method to each of the resulting scalar equations [32]. In the next section we describe a high resolution method for nonlinear systems, which is based on this idea.

5.3. SLOPE LIMITER METHODS

5.3.1. Introduction

In this section the *slope limiter method* is described. Methods of this type were first introduced by van Leer [51]. A variety of similar methods have been proposed since then (see e.g. [30, 40, 80]). In many cases the slope limiter method can be converted into a flux limiter method. However, the slope limiter method is more geometric in nature and it can be extended to systems of equations in a relatively straightforward way.

The basic idea of a slope limiter method is to generalize a Godunov method by replacing the piecewise constant representation of the solution $\tilde{\mathrm{U}}^n$ (see (4.30)) by a more accurate representation, say piecewise linear. Recall that three steps are involved in the basic Godunov method in order to calculate the numerical solution at time level t^{n+1} from the known numerical solution at time level t^n (see Section 4.5.2),

step 1: use U_i^n to construct a piecewise continuous function $\tilde{\mathrm{U}}^n(\cdot, t^n)$;

step 2: solve the conservation law exactly with initial data $\tilde{\mathrm{U}}^n(\cdot, t^n)$;

step 3: define the numerical solution U_i^{n+1} at time level t^{n+1} by averaging the resulting solution over each mesh cell.

In the first order Godunov method as described in Section 4.5, $\tilde{\mathrm{U}}^n(\cdot, t^n)$ is chosen to be piecewise constant (see (4.30)). This procedure is generalized by defining a piecewise linear function $\tilde{\mathrm{U}}^n$, instead of a piecewise constant function, i.e.

$$\tilde{\mathrm{U}}^n(x, t^n) := \mathrm{U}_i^n + \delta_i^n (x - x_i), \quad \forall\, x \in [x_{i-1/2}, x_{i+1/2}). \tag{5.26}$$

Here $\boldsymbol{\delta}_i^n \in \mathbb{R}^m$ is a vector of slopes in the ith cell for each component of $\tilde{\mathbf{U}}^n$. The slopes are based on the data $\mathbf{U}^n$. Note that if $\boldsymbol{\delta}_i^n = \mathbf{0}$, then (5.26) reduces to (4.30). One of the main problems in obtaining a slope limiter method is the choice of the slopes $\boldsymbol{\delta}_i^n$. The reconstruction of the first step of the basic Godunov method may be replaced by more accurate approximations as well. We can attempt to obtain better accuracy by using quadratics, as in the piecewise parabolic method (PPM) of Colella and Woodward [17] or even higher order reconstructions as in the essentially nonoscillatory (ENO) methods [34, 37]. However, we restrict ourselves to piecewise linear functions.

Remark that the cell average of $\tilde{\mathbf{U}}^n(\cdot, t_n)$ over the cell $[x_{i-1/2}, x_{i+1/2})$ is equal to $\mathbf{U}_i^n$ for any choice of $\boldsymbol{\delta}_i^n$. Since step 2 and step 3 are also conservative, the overall method is conservative for any choice of $\boldsymbol{\delta}_i^n$.

In step 2 piecewise linear initial data are used instead of piecewise constant initial data. This implies that the Riemann problem cannot be solved exactly in general (except for linear conservation laws). However, it is possible to approximate the solution in a suitable way. One possibility is to approximate the Riemann problem by a two-step procedure. The linear profile is used to compute two constant reference states $\mathbf{u}_L$ and $\mathbf{u}_R$ at each cell interface. These states are used as piecewise constant initial data for the Riemann problem at the cell interface, which then may be solved exactly [28, 40, 52]. Another possibility is to use the piecewise linear profiles as initial data for the Riemann problem at the cell interfaces and, subsequently, approximate the solution of the Riemann problem [4]. We will use a method introduced by LeVeque [56]. In this method the nonlinear flux function $\mathbf{f}$ is replaced by some linear function $\hat{\mathbf{f}}$ in the neighbourhood of each cell interface. Subsequently, the resulting linear Riemann problem with piecewise linear initial data is solved exactly. This type of approximation has already been introduced in the discussion of Roe's method in Section 4.5.3. First, we describe the slope limiter method for the scalar, linear conservation law (5.14) in more detail. Subsequently, the above linearization process is used to extend the method for systems of nonlinear conservation laws.

5.3.2. Scalar, Linear Conservation Laws

In this section we consider the scalar, linear convection equation (5.14) with $a > 0$. For this linear problem we can perform step 2 exactly, where the exact solution $\tilde{U}^n(\cdot, t^{n+1})$ is given by $\tilde{U}^n(x, t^{n+1}) = \tilde{U}^n(x - a\Delta t, t^n)$. Using this together with (4.35), (5.26) and $\sigma \leq 1$, it is straightforward to see that

$$U_i^{n+1} = U_i^n - \sigma(U_i^n - U_{i-1}^n) - \tfrac{1}{2}\sigma(1-\sigma)(\Delta x\delta_i^n - \Delta x\delta_{i-1}^n), \tag{5.27}$$

where σ is given by (4.13). The numerical flux for the method above is given by

$$F_{i+1/2}^n = F_{i+1/2}^{(BU)} + \tfrac{1}{2}a(1-\sigma)\Delta x\delta_i^n,$$

which has exactly the same form as the flux limiter method (5.15) if we set

$$\delta_i^n = \frac{U_{i+1}^n - U_i^n}{\Delta x}\varphi(\theta_{i+1/2}^+).$$

In this way the "flux limiter" φ can be interpreted as a "slope limiter". Furthermore, if $\varphi = 1$, then $F^n_{i+1/2}$ reduces to the Lax-Wendroff flux function (4.20). This shows that it is possible to obtain second order accuracy by this approach. Obviously, a method similar to (5.27) can be obtained when $a < 0$. We can unify both methods (for $a > 0$ and $a < 0$) into the single formula

$$U_i^{n+1} = U_i^n - \sigma(U_j^n - U_{j-1}^n) - \tfrac{1}{2}\sigma(\operatorname{sgn}(a) - \sigma)(\Delta x\delta_j^n - \Delta x\delta_{j-1}^n), \tag{5.28}$$

where $\operatorname{sgn}(x) = 1$ for $x > 0$, $\operatorname{sgn}(x) = -1$ for $x < 0$ and j is defined by

$$j := \begin{cases} i+1, & a < 0, \\ i, & a > 0. \end{cases} \tag{5.29}$$

It is straightforward to see that the numerical flux of method (5.28) is given by

$$F^n_{i+1/2} = F^{(BU)}_{i+1/2} + \tfrac{1}{2}a(\operatorname{sgn}(a) - \sigma)\Delta x\delta_j^n. \tag{5.30}$$

A possible choice for δ_i^n is

$$\delta_i^n = \begin{cases} \dfrac{U_i^n - U_{i-1}^n}{\Delta x}\varphi\Big(\dfrac{U_{i+1}^n - U_i^n}{U_i^n - U_{i-1}^n}\Big), & a < 0, \\ \dfrac{U_{i+1}^n - U_i^n}{\Delta x}\varphi\Big(\dfrac{U_i^n - U_{i-1}^n}{U_{i+1}^n - U_i^n}\Big), & a > 0, \end{cases} \tag{5.31}$$

in which case the slope limiter method (5.30) reduces to the flux limiter method (5.17).

The oscillations which arise with the Lax-Wendroff method (see Figure 4.1) can be interpreted geometrically as being caused by a poor choice of slopes, leading to a piecewise linear reconstruction $\tilde{U}^n(\cdot, t^n)$ with much larger total variation than the given data U^n. Similarly to the flux limiter approach we try to remedy this by choosing an appropriate limiter φ, which reduces the slope δ_i^n near discontinuities or extreme points.

Theorem 5.16. *Suppose the initial data of method* (5.28) *have compact support and bounded total variation and define the piecewise constant function* $U_{\Delta t}$ *as in* (4.4). *If the slope* δ_i^n *is chosen such that*

$$\mathrm{TV}(\tilde{U}^n(\cdot, t^n)) \le \mathrm{TV}(U_{\Delta t}(\cdot, t^n)) \tag{5.32}$$

holds, then the method (5.28) *is a TVD method under the CFL condition* (4.15).

Since steps 2 and 3 are TVD, it is clear that imposing (5.32) in step 1 results in a method that is overall TVD. For the proof of Theorem 5.16 we refer to [28]. If van Leer's limiter (5.25) or Roe's minmod limiter (5.24) is used, then (5.32) is satisfied. However, it is possible to violate (5.32) and still obtain a TVD method, since step 3 tends to reduce the total variation and may eliminate overshoots caused in the previous steps.

5.3.3. Linear Systems of Conservation Laws

In this section we want to generalize the scalar flux function (5.30) to linear systems of equations. Therefore, let the flux function $\mathbf{f}$ be linear, i.e. $\mathbf{f}(\mathbf{u}) = A\mathbf{u}$, where A is a constant $m \times m$-matrix and consider the linear system of hyperbolic conservation laws

$$\frac{\partial}{\partial t}\mathbf{u}(x,t) + A\frac{\partial}{\partial x}\mathbf{u}(x,t) = 0. \tag{5.33}$$

Since the system (5.33) is assumed to be hyperbolic, we have $AR = R\Lambda$, where the matrix $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ is the diagonal matrix of the eigenvalues of A and $R = (\mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \dots, \mathbf{r}^{(m)})$ is the nonsingular matrix of the corresponding right eigenvectors of A. The natural generalization of (5.30) to linear systems is obtained by diagonalizing the system and applying the slope limiter method to each of the resulting scalar equations. It is straightforward to see that (5.33) is diagonalized as follows

$$\frac{\partial}{\partial t}\mathbf{v}(x,t) + \Lambda\frac{\partial}{\partial x}\mathbf{v}(x,t) = 0, \tag{5.34}$$

where $\mathbf{v}$ is defined by $\mathbf{v} := R^{-1}\mathbf{u}$. Note that (5.34) is a system of m decoupled scalar equations. Let the vector $\mathbf{V}_i^n = (V_{1,i}^n, V_{2,i}^n, \dots, V_{m,i}^n)^T$ denote the numerical approximation of $\mathbf{v}$, which is obtained by applying the scalar method (5.28) to each component of $\mathbf{v}$, i.e.

$$V_{k,i}^{n+1} = V_{k,i}^n - \sigma_k(V_{k,j(k)}^n - V_{k,j(k)-1}^n) - \tfrac{1}{2}\sigma_k(\operatorname{sgn}(\lambda_k) - \sigma_k)(\Delta x\delta_{k,j(k)}^n - \Delta x\delta_{k,j(k)-1}^n), \tag{5.35}$$

where $\sigma_k := \lambda_k \Delta t/\Delta x$, $\delta_{k,i}^n$ is the slope of $V_{k,\cdot}^n$ in the ith cell and $j(k)$ is defined by (see (5.29))

$$j(k) := \begin{cases} i+1, & \lambda_k < 0, \\ i, & \lambda_k > 0. \end{cases} \tag{5.36}$$

Since $\mathbf{v} = R^{-1}\mathbf{u}$ it seems natural to define $\mathbf{U}_i^n$ by $\mathbf{U}_i^n := \sum_{k=1}^m V_{k,i}^n \mathbf{r}^{(k)}$. After multiplying (5.35) by $\mathbf{r}^{(k)}$ and summing over k we obtain

$$\begin{aligned} \mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \tau \Big\{ & \sum_{k=1}^m [\,\lambda_k(V_{k,j(k)}^n - V_{k,j(k)-1}^n)\mathbf{r}^{(k)}\,] \\ & + \tfrac{1}{2}\sum_{k=1}^m [\,\lambda_k(\operatorname{sgn}(\lambda_k) - \sigma_k)(\Delta x\delta_{k,j(k)}^n - \Delta x\delta_{k,j(k)-1}^n)\mathbf{r}^{(k)}\,] \Big\} \end{aligned} \tag{5.37}$$

Using the notation above, it can be shown that for the linear system (5.33), the numerical flux of the first order basic Godunov method (4.37) is given by [56]

$$\mathbf{F}_{i+1/2}^{(G)} = \sum_{k=1}^m \lambda_k V_{k,j(k)}^n \mathbf{r}^{(k)}.$$

Now it can be shown that the numerical flux for the method (5.37) is given by

$$\mathbf{F}_{i+1/2}^n = \mathbf{F}_{i+1/2}^{(G)} + \tfrac{1}{2}\sum_{k=1}^m \lambda_k(\operatorname{sgn}(\lambda_k) - \sigma_k)\Delta x\delta_{k,j(k)}^n \mathbf{r}^{(k)}, \tag{5.38}$$

which is the generalization of the scalar flux function (5.30) to linear systems. The remaining question is the choice of the slope $\delta^n_{k,i}$ of $\mathbf{V}^n_{k,\cdot}$. A possible choice for $\delta^n_{k,i}$ is (5.31), i.e.

$$\delta^n_{k,i} = \begin{cases} \dfrac{V^n_{k,i} - V^n_{k,i-1}}{\Delta x}\,\varphi\Big(\dfrac{V^n_{k,i+1} - V^n_{k,i}}{V^n_{k,i} - V^n_{k,i-1}}\Big), & \lambda_k < 0, \\ \dfrac{V^n_{k,i+1} - V^n_{k,i}}{\Delta x}\,\varphi\Big(\dfrac{V^n_{k,i} - V^n_{k,i-1}}{V^n_{k,i+1} - V^n_{k,i}}\Big), & \lambda_k > 0. \end{cases} \tag{5.39}$$

Since all the eigenvectors are linearly independent the states $\mathbf{U}^n_i$ and $\mathbf{U}^n_{i+1}$ can be decomposed as

$$\mathbf{U}^n_{i+1} - \mathbf{U}^n_i = \sum_{k=1}^{m} \alpha_{k,i+1/2}\mathbf{r}^{(k)}_{i+1/2},$$

where $\alpha_{k,i+1/2} \in I\!R$ for all k with $1 \le k \le m$. Using this, it follows that $\alpha_{k,i+1/2} = V^n_{k,i+1} - V^n_{k,i}$ and, subsequently, (5.39) is rewritten as

$$\delta^n_{k,i} = \begin{cases} \dfrac{\alpha^n_{k,i-1/2}}{\Delta x}\,\varphi\Big(\dfrac{\alpha^n_{k,i+1/2}}{\alpha^n_{k,i-1/2}}\Big), & \lambda_k < 0, \\ \dfrac{\alpha^n_{k,i+1/2}}{\Delta x}\,\varphi\Big(\dfrac{\alpha^n_{k,i-1/2}}{\alpha^n_{k,i+1/2}}\Big), & \lambda_k > 0. \end{cases} \tag{5.40}$$

5.3.4. Nonlinear Systems of Conservation Laws

A natural way to generalize the slope limiter method (5.38) to nonlinear systems is to linearize the nonlinear system. Therefore, the nonlinear flux function $\mathbf{f}$ is replaced by some linear function $\hat{\mathbf{f}}$ in the neighbourhood of each cell interface. This type of approximation has already been discussed in Section 4.5.3. Hence, at each cell interface $x_{i+1/2}$, we compute an $m \times m$-matrix $\hat{A}_{i+1/2}$ which satisfies the requirements $(i) - (iii)$, as described in Section 4.5.3. Subsequently, around $x_{i+1/2}$, $\mathbf{f}$ is replaced by $\hat{\mathbf{f}}(\mathbf{u}) = \hat{A}_{i+1/2}\mathbf{u}$. Analogously to Section 4.5.3, we denote the eigenvalues of $\hat{A}_{i+1/2}$ by $\hat{\lambda}_{k,i+1/2}$ and the corresponding right eigenvectors by $\hat{\mathbf{r}}^{(k)}_{i+1/2}$.

First of all we remark that Roe's numerical flux function $\mathbf{F}^{(R)}_{i+1/2}$ (see (4.49)) results from applying the first order basic Godunov method to the linearized system. Now the straightforward generalization of method (5.38) to nonlinear systems becomes

$$\mathbf{F}^n_{i+1/2} = \mathbf{F}^{(R)}_{i+1/2} + \tfrac{1}{2}\sum_{k=1}^{m} \hat{\lambda}_{k,i+1/2}(\operatorname{sgn}(\hat{\lambda}_{k,i+1/2}) - \hat{\sigma}_{k,i+1/2})\Delta x \hat{\delta}^n_{k,j(k)}\hat{\mathbf{r}}^{(k)}_{i+1/2}, \tag{5.41}$$

where $\hat{\sigma}_{k,i+1/2} := \hat{\lambda}_{k,i+1/2}\Delta t/\Delta x$ and $\mathbf{F}^{(R)}_{i+1/2}$ is Roe's first order numerical flux function given by (4.49). Since all right eigenvectors are assumed to be linearly independent, the states $\mathbf{U}^n_i$ and $\mathbf{U}^n_{i+1}$ can be decomposed as in (4.48). Using the coefficients $\hat{\alpha}^n_{k,i+1/2}$ in this decomposition, the slopes $\hat{\delta}^n_{k,i}$ are chosen analogous to (5.39), giving

$$\hat{\delta}^n_{k,i} := \begin{cases} \dfrac{\hat{\alpha}^n_{k,i-1/2}}{\Delta x}\,\varphi\Big(\dfrac{\hat{\alpha}^n_{k,i+1/2}}{\hat{\alpha}^n_{k,i-1/2}}\Big), & \hat{\lambda}_{k,i+1/2} < 0, \\ \dfrac{\hat{\alpha}^n_{k,i+1/2}}{\Delta x}\,\varphi\Big(\dfrac{\hat{\alpha}^n_{k,i-1/2}}{\hat{\alpha}^n_{k,i+1/2}}\Big), & \hat{\lambda}_{k,i+1/2} > 0, \end{cases} \tag{5.42}$$

where φ is some limiter function satisfying (5.22) and $\varphi(1) = 1$.

Until now we have ignored the entropy condition. Since Roe's method is used, method (5.41) may include physically inadmissible expansion shocks and, analogously to the first order case, some modification is needed to exclude these expansion shocks [28].

Example 5.17. In this example again Burgers' equation (3.4a) is considered, with a solution given by (4.38) or (4.39). Let $s^n_{i+1/2} := \frac{1}{2}(U^n_i + U^n_{i+1})$ (see Example 4.16). Furthermore, $\hat{A}_{i+1/2}$ is defined by $\hat{A}_{i+1/2} := \frac{1}{2}(U^n_i + U^n_{i+1}) = s^n_{i+1/2}$ (see Example 4.17). Hence, $\hat{r}_{i+1/2} = 1$ and $\hat{\lambda}_{i+1/2} = s^n_{i+1/2}$. Using this, it is not difficult to see that numerical flux of the slope limiter method (5.41) for Burgers' equation is given by

$$F^n_{i+1/2} = F^{(R)}_{i+1/2} + \tfrac{1}{2}s^n_{i+1/2}(\operatorname{sgn}(s^n_{i+1/2}) - \hat{\sigma}_{i+1/2})\Delta x \hat{\delta}^n_{j(k)},$$

where $F^{(R)}_{i+1/2}$ is Roe's first order numerical flux given by (4.50), $\hat{\sigma}_{i+1/2} := s^n_{i+1/2}\Delta t/\Delta x$, $j(k)$ is given by (5.36) and the slopes $\hat{\delta}^n_i$ are computed by (5.42). For Burgers' equation (5.42) can be rewritten as (see (5.31))

$$\hat{\delta}^n_i = \begin{cases} \dfrac{U^n_i - U^n_{i-1}}{\Delta x}\,\varphi\Big(\dfrac{U^n_{i+1} - U^n_i}{U^n_i - U^n_{i-1}}\Big), & s^n_{i+1/2} < 0, \\ \dfrac{U^n_{i+1} - U^n_i}{\Delta x}\,\varphi\Big(\dfrac{U^n_i - U^n_{i-1}}{U^n_{i+1} - U^n_i}\Big), & s^n_{i+1/2} > 0, \end{cases}$$

where φ is some limiter function. □

Example 5.18. In this example the high resolution method (5.41) with slopes (5.42) is applied to the shock tube problem for the one-dimensional nonreactive Euler equations, i.e. (2.26) with $w = 0$. Note that numerical results of Roe's first order method applied to the Euler equations are presented in Example 4.18. The matrix $\hat{A}_{i+1/2}$ is defined by (4.51) and the eigenvalues and eigenvectors of $\hat{A}_{i+1/2}$ are given by (4.54) and (4.55), respectively. Finally, (4.48) is used to compute $\hat{\alpha}_{1,i+1/2}$, $\hat{\alpha}_{2,i+1/2}$, $\hat{\alpha}_{3,i+1/2}$ and $\hat{\alpha}_{4,i+1/2}$. The computation of the numerical flux (5.41) is now straightforward, if the slopes $\hat{\delta}$ are defined by (5.42). In Figure 5.2 the numerical results are given, where the slopes are computed using Roe's superbee limiter (5.23). Comparing the results with Figure 4.2, we clearly observe a better resolution of the smooth part of the solution and a sharper resolution of the discontinuities. □

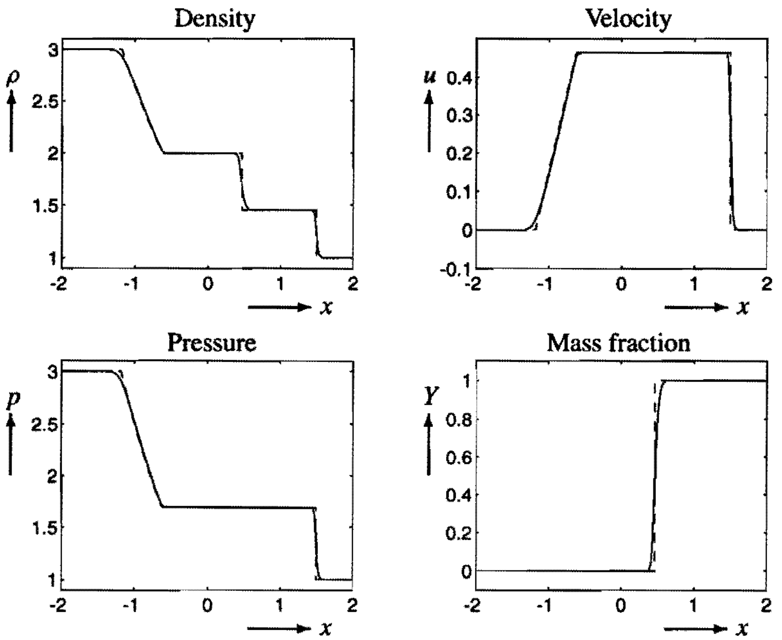

Figure 5.2. Numerical results for the high resolution method (5.41) with slopes (5.42) and Roe's superbee limiter (5.23); exact solution (dashed line) and numerical solution (solid line) of (2.26) with $w = 0$ at time $t = 1$ with $\Delta t = 0.002$, $\Delta x = 0.02$ and the initial conditions $p(x, 0) = 3$, $\rho(x, 0) = 3$, $u(x, 0) = 0$, $Y(x, 0) = 0$ if $x < 0$ and $p(x, 0) = 1$, $\rho(x, 0) = 1$, $u(x, 0) = 0$, $Y(x, 0) = 1$ if $x > 0$.

Method (5.41) with slopes (5.42) is one example of a high resolution method for nonlinear systems of conservation laws. Especially, the choice of the slopes and the choice of the limiter function strongly depends on the particular problem. We have experienced that our slopes (5.42) works well for one-dimensional detonation waves (see Chapter 8). Moreover, for flows with strong discontinuities, Roe's superbee limiter appears to be the best choice for φ. For many problems the high resolution method (5.41) with slopes (5.42) is neither the most sophisticated nor the best. However, it illustrates many of the basic ideas used in the wide variety of other methods available in literature [4, 17, 23, 34, 37, 40, 68].

6

TIME SPLITTING METHODS

In this chapter we discuss and analyse time splitting methods for hyperbolic conservation laws with source terms. Splitting methods for time dependent partial differential equations have most frequently been studied in the context of spatial splittings [84]. Some attention has also been given to splitting or fractional step methods for problems where the differential operator is split into parts corresponding to different physical processes, which are most naturally handled by different techniques [57].

Time splitting methods are often used for detonation waves, where it seems natural to deal with the fluid dynamics and the chemistry in a different way [10, 16]. In this chapter we discuss two splitting methods for hyperbolic conservation laws with source terms. In both splitting methods we alternate at each time level between solving the homogeneous conservation law without source term (i.e. the fluid dynamics) and solving the conservation laws without convection (i.e. the chemistry), giving a system of ordinary differential equations.

Naturally, the main question is how well the time splitting method approximates the exact solution at each time level. For dimensional splitting this question has been answered in [84]. However, to our knowledge, splitting methods for nonhomogeneous conservation laws have not been analysed so far. In this chapter we analyse two splitting methods. We show that they are first and second order accurate in time, respectively. Both time splitting methods approximate the exact solution at discrete time levels t^n. Subsequently, using also a spatial discretization, we obtain a numerical version of both splitting methods. Finally, we analyse the corresponding discretization errors.

This chapter is organized as follows. In the first section we describe the general idea behind a time splitting method. In the second section two well-known splitting methods are introduced. Furthermore, we introduce and analyse the splitting error for both methods. In Section 6.3 we introduce two numerical methods based on the splitting methods of Section 6.2. Furthermore, the corresponding errors are discussed and analysed. Finally, we present some numerical results illustrating the preceding analysis.

6.1. INTRODUCTION

A variety of numerical methods can be developed for initial value problems for conservation laws with source terms (see (3.2)) [3, 9, 12, 58]

$$\frac{\partial}{\partial t}\mathrm{u}(x,t) + \frac{\partial}{\partial x}\mathrm{f}(\mathrm{u}(x,t)) = \mathrm{q}(\mathrm{u}(x,t)), \tag{6.1a}$$

$$\mathrm{u}(x,0) = \mathrm{u}^0(x), \quad \forall x \in \mathbb{R}. \tag{6.1b}$$

A very natural (and popular) way to solve (6.1) is a *time splitting method* [57, 84, 91]. The basic idea behind splitting methods is to split a complicated problem into subproblems, which are easier to handle. Splitting methods are often used for detonation waves, where it seems natural to deal with the fluid dynamics and the chemistry in a different way [10, 72].

Time splitting methods approximate the exact solution at discrete time levels. For a given time step Δt, the discrete time levels t^n are defined by (see Section 4.1) $t^n := n\Delta t$ for $n = 0, 1, 2, \ldots$. The splitting methods we shall consider produce approximations $\tilde{\mathrm{u}}(\cdot, t^n)$ to the true solution $\mathrm{u}(\cdot, t^n)$ of (6.1) at the discrete time levels t^n.

Before we describe the splitting technique in more detail, we introduce two initial value problems, which are used later on. The first problem is obtained by neglecting the source term in (6.1) (i.e. $\mathrm{q} = 0$), giving a homogeneous conservation law. Hence, we deduce the following initial value problem

$$\frac{\partial}{\partial t}\mathrm{v}(x,t) + \frac{\partial}{\partial x}\mathrm{f}(\mathrm{v}(x,t)) = 0, \tag{6.2a}$$

$$\mathrm{v}(x,\tau_1) = \mathrm{v}^{\tau_1}(x), \quad \forall x \in \mathbb{R}, \tag{6.2b}$$

where $\tau_1 = t^n$ for some $n \geq 0$, in general. Differential equation (6.2a) describes the fluid dynamical part of (6.1a). The second initial value problem is obtained by neglecting the convection in (6.1) (i.e. $\partial\mathrm{f}/\partial x = 0$), giving a system of ordinary differential equations for all x. The corresponding initial value problem reads

$$\frac{\partial}{\partial t}\mathrm{w}(x,t) = \mathrm{q}(\mathrm{w}(x,t)), \tag{6.3a}$$

$$\mathrm{w}(x,\tau_2) = \mathrm{w}^{\tau_2}(x), \tag{6.3b}$$

for all $x \in \mathbb{R}$. In general, $\tau_2 = t^n$ for some $n \geq 0$. However, τ_1 is not necessarily equal to τ_2. Differential equation (6.3a) may be interpreted as the chemical part of (6.1a).

Next we explain the general idea behind the splitting approach. Let the approximate solution $\tilde{\mathrm{u}}(\cdot, t^n)$ at time level t^n be given. In the time splitting method, the approximate solution $\tilde{\mathrm{u}}(\cdot, t^{n+1})$ at time level t^{n+1} is computed by solving a sequence of initial value problems of type (6.2) or (6.3). The initial value problems occurring in this sequence are related to each other and to $\tilde{\mathrm{u}}(\cdot, t^n)$ via the initial conditions. For example, consider the following splitting method to obtain an approximation $\tilde{\mathrm{u}}(\cdot, t^{n+1})$ (for given $\tilde{\mathrm{u}}(\cdot, t^n)$)

(i) solve problem (6.2a) for $t^n < t \leq t^{n+1}$ (i.e. $\tau_1 = t^n$) with initial data $\mathrm{v}^{\tau_1} = \tilde{\mathrm{u}}(\cdot, t^n)$ and denote the solution by v;

(ii) solve problem (6.3a) for $t^n < t \leq t^{n+1}$ (i.e. $\tau_2 = t^n$) with initial data $\mathrm{w}^{\tau_2} = \mathrm{v}(\cdot, t^{n+1})$ and denote the solution by w;

(iii) define $\tilde{\mathrm{u}}(\cdot, t^{n+1})$ by $\tilde{\mathrm{u}}(\cdot, t^{n+1}) := \mathrm{w}(\cdot, t^{n+1})$.

A detailed description of the splitting method (i)-(iii) is given in the next section.

The splitting approach is frequently used to solve reacting flow problems. At first glance it may appear to be less satisfactory than a standard method, since the fluid dynamics and chemistry are strongly coupled in our problem (see e.g. (2.26)). However, time splitting methods are frequently used, since high quality numerical methods have been developed both for systems of homogeneous conservation laws and for (stiff) ordinary differential equations. By decomposing the problem (6.1) into the subproblems (6.2) and (6.3), one can take advantage of positive aspects of both methods. Moreover, to some extent the underlying mathematical theory may also be carried over. By alternatingly applying a high resolution method to (6.2) and a stable (stiff) solver to the system of ODEs (6.3), one can easily derive a method with excellent stability properties for the full problem.

Naturally, an important question is how well a splitting method approximates the exact solution of (6.1). In other words, we are interested in the error $\tilde{\mathrm{u}}(\cdot, t^n) - \mathrm{u}(\cdot, t^n)$. In the next section we show that method (i)-(iii) above is first order accurate, i.e. $\|\tilde{\mathrm{u}}(\cdot, t^n) - \mathrm{u}(\cdot, t^n)\| = \mathcal{O}(\Delta t)$. Furthermore, we describe in Section 6.2 the so-called Strang splitting method for problem (6.1). This method is shown to be second order accurate, i.e. $\|\tilde{\mathrm{u}}(\cdot, t^n) - \mathrm{u}(\cdot, t^n)\| = \mathcal{O}(\Delta t^2)$. Finally, we describe in the last section how time splitting methods can be used in developing numerical methods for (6.1).

6.2. THE SPLITTING ERROR

In this section we like to analyse the error $\tilde{\mathrm{u}}(\cdot, t^n) - \mathrm{u}(\cdot, t^n)$ for two particular splitting methods. As for the discretization error, it turns out to be useful to define a local error first. Therefore, we introduce the *local splitting error*. The local splitting error $\mathrm{D}_{\Delta t}^{spl}(x, t)$ measures how well the splitting method approximates the solution of problem (6.1) locally after a time step Δt. First, we introduce some notation. We recall that a function $\mathrm{v} \in L_1^{loc}(\mathbb{R})$, if $\mathrm{v} \in L_1(\Omega)$ for all compact $\Omega \subset \mathbb{R}$. Let the subset $D_f \subset L_1^{loc}(\mathbb{R})$ be defined by

$$D_f := \{ \mathrm{g} \in L_1^{loc}(\mathbb{R}) \mid \text{(6.2a) with initial data } \mathrm{v}^{\tau_1} = \mathrm{g} \text{ has a unique solution } \mathrm{v}(\cdot, t) \in L_1^{loc}(\mathbb{R}) \text{ for } \tau_1 < t \leq \tau_1 + \Delta t \}.$$

Next, we define an operator $\mathcal{S}_{\Delta t}^f$ which relates initial data $\mathrm{v}^{\tau_1} \in D_f$ to the solution of (6.2) at time $\tau_1 + \Delta t$. The operator $\mathcal{S}_{\Delta t}^f : D_f \to L_1^{loc}(\mathbb{R})$ is called the *exact solution operator* of problem (6.2a) at $t = \tau_1 + \Delta t$, if

$$\mathrm{v}(\cdot, \tau_1 + \Delta t) = \mathcal{S}_{\Delta t}^f \mathrm{v}^{\tau_1}, \tag{6.4}$$

where $\mathrm{v}(\cdot, \tau_1 + \Delta t)$ is the exact solution of (6.2a) at time $t = \tau_1 + \Delta t$ with initial data v^{τ_1}. Similarly, we define the subset $D_c \subset L_1^{loc}(\mathbb{R})$ as

$$D_c \;:=\; \{\, \mathrm{g} \in L_1^{loc}(\mathbb{R}) \mid \text{(6.3a) with initial data } \mathrm{w}^{\tau_2} = \mathrm{g} \text{ has a unique solution } \mathrm{w}(\cdot, t) \in L_1^{loc}(\mathbb{R}) \text{ for } \tau_2 < t \le \tau_2 + \Delta t \,\}.$$

Next, we define an operator $\mathcal{S}^c_{\Delta t}$ which relates initial data $\mathrm{w}^{\tau_2} \in D_c$ to the solution of (6.3a) at time $\tau_2 + \Delta t$. The operator $\mathcal{S}^c_{\Delta t} : D_c \to L_1^{loc}(\mathbb{R})$ is called the exact solution operator of problem (6.3a) at time $t = \tau_2 + \Delta t$, if

$$\mathrm{w}(\cdot, \tau_2 + \Delta t) \;=\; \mathcal{S}^c_{\Delta t}\,\mathrm{w}^{\tau_2}, \tag{6.5}$$

where $\mathrm{w}(\cdot, \tau_2 + \Delta t)$ is the exact solution at $t = \tau_2 + \Delta t$ of (6.3a) with initial data w^{τ_2}. For the sake of convenience we denote the value of $\mathcal{S}^f_{\Delta t}\mathrm{v}(\cdot, t)$ at the point x by $\mathcal{S}^f_{\Delta t}\mathrm{v}(x, t)$, i.e. $\mathcal{S}^f_{\Delta t}\mathrm{v}(x, t) := (\mathcal{S}^f_{\Delta t}\mathrm{v}(\cdot, t))(x)$. Similarly, we write $\mathcal{S}^c_{\Delta t}\mathrm{w}(x, t) := (\mathcal{S}^c_{\Delta t}\mathrm{w}(\cdot, t))(x)$.

We consider splitting methods consisting of some product of the solution operators $\mathcal{S}^f_{\Delta t}$ and $\mathcal{S}^c_{\Delta t}$. As noted before, splitting methods produce approximations $\tilde{\mathrm{u}}(\cdot, t^n)$ to the true solution $\mathrm{u}(\cdot, t^n)$. Let $\mathcal{S}^{tot}_{\Delta t} : L_1^{loc}(\mathbb{R}) \to L_1^{loc}(\mathbb{R})$ be some product of the solution operators $\mathcal{S}^f_{\Delta t}$ and $\mathcal{S}^c_{\Delta t}$, then the time splitting method is given by

$$\tilde{\mathrm{u}}(\cdot, t + \Delta t) \;=\; \mathcal{S}^{tot}_{\Delta t}\,\tilde{\mathrm{u}}(\cdot, t). \tag{6.6}$$

It is assumed that $\mathcal{S}^{tot}_{\Delta t}$ is well defined. For example, if we consider the splitting method (i)-(iii) of Section 6.1, then $\mathcal{S}^{tot}_{\Delta t} = \mathcal{S}^c_{\Delta t}\mathcal{S}^f_{\Delta t}$. Now $\mathcal{S}^{tot}_{\Delta t}$ is well defined, provided $\mathcal{S}^f_{\Delta t} : D_f \to D_c$. For convenience sake we write $\mathcal{S}^{tot}_{\Delta t}\tilde{\mathrm{u}}(x, t) := (\mathcal{S}^{tot}_{\Delta t}\tilde{\mathrm{u}}(\cdot, t))(x)$. If now $\tilde{\mathrm{u}}$ is replaced in (6.6) by the exact solution of (6.1), then in general the equality will not hold exactly. This leads to the following definition (see Definition 4.2).

Definition 6.1. *Consider a splitting method written in the generic form* (6.6). *The* local splitting error $\mathrm{D}^{spl}_{\Delta t}$ *of this method at the point* (x, t^n) *is defined by*

$$\mathrm{D}^{spl}_{\Delta t}(x, t^n) \;:=\; \frac{1}{\Delta t}\{\mathrm{u}(x, t^n + \Delta t) \;-\; \mathcal{S}^{tot}_{\Delta t}\,\mathrm{u}(x, t^n)\}, \tag{6.7}$$

where u *is the exact solution of* (6.1).

Note that the local splitting error depends on the exact solution of (6.1). The *global splitting error* $\mathrm{E}^{spl}_{\Delta t}(x, t)$ is defined as

$$\mathrm{E}^{spl}_{\Delta t}(x, t^n) \;:=\; \tilde{\mathrm{u}}(x, t^n) - \mathrm{u}(x, t^n) \;=\; (\mathcal{S}^{tot}_{\Delta t})^n\,\mathrm{u}(x, 0) - \mathrm{u}(x, t^n),$$

where $\tilde{\mathrm{u}}(\cdot, 0) = \mathrm{u}^0$ and the superscript n for $\mathcal{S}^{tot}_{\Delta t}$ represents the nth power of the operator, i.e. n times applied.

It follows from (6.6) and (6.7) that the global splitting error satisfies (rewritten in the functional form)

$$\mathrm{E}^{spl}_{\Delta t}(\cdot, t^{n+1}) \;=\; \mathcal{S}^{tot}_{\Delta t}\tilde{\mathrm{u}}(\cdot, t^n) - \mathcal{S}^{tot}_{\Delta t}\mathrm{u}(\cdot, t^n) \;-\; \Delta t\,\mathrm{D}^{spl}_{\Delta t}(\cdot, t^n).$$

Now we assume that for some given norm $\|\cdot\|$ there exists a constant $C \geq 0$ such that

$$\|\mathcal{S}^{tot}_{\Delta t}\tilde{\mathrm{u}}(\cdot,t^n) - \mathcal{S}^{tot}_{\Delta t}\mathrm{u}(\cdot,t^n)\| \;\leq\; (1+\Delta t C)\|\tilde{\mathrm{u}}(\cdot,t^n) - \mathrm{u}(\cdot,t^n)\|. \tag{6.8}$$

For the operator $\mathcal{S}^c_{\Delta t}$ the above requirement is fulfilled if q is a Lipschitz continuous function. Using the latter equations, we obtain

$$\|\mathrm{E}^{spl}_{\Delta t}(\cdot,t^{n+1})\| \;\leq\; (1+C\Delta t)\|\mathrm{E}^{spl}_{\Delta t}(\cdot,t^n)\| \;+\; \Delta t\|\mathrm{D}^{spl}_{\Delta t}(\cdot,t^n)\|.$$

The global error $\mathrm{E}^{spl}_{\Delta t}$ at time t^{n+1} consists of two parts. One is the new local error $\Delta t \mathrm{D}^{spl}_{\Delta t}$ introduced in the last time step. The other part is the global error from the previous time steps. By applying this relation recursively we obtain an expression for the global error at time t^n

$$\|\mathrm{E}^{spl}_{\Delta t}(\cdot,t^n)\| \;\leq\; (1+C\Delta t)^n\|\mathrm{E}^{spl}_{\Delta t}(\cdot,0)\| \;+\; \Delta t\sum_{j=1}^{n}(1+C\Delta t)^{n-j}\|\mathrm{D}^{spl}_{\Delta t}(\cdot,t^{j-1})\|. \tag{6.9}$$

In order to obtain a bound on the global error, we must ensure that the local error $\mathrm{D}^{spl}_{\Delta t}(\cdot,t^{j-1})$ is not unduly amplified by applying $n-j$ steps of the method. It follows from (6.9) and $(1+C\Delta t)^n \leq \exp(Cn\Delta t)$ that (6.8) is a sufficient condition to obtain a bound on the global error. Note that a bound is always with respect to some given norm $\|\cdot\|$. Next, the concept of *stability* is introduced.

Definition 6.2. *Consider the splitting method written in the generic form* (6.6). *The splitting method is called* stable *in some norm* $\|\cdot\|$, *if for each time* $T > 0$ *there exists a constant* $C \geq 0$ *and a constant* k_0 *such that*

$$\|\mathcal{S}^{tot}_{\Delta t}\mathrm{g}_1 - \mathcal{S}^{tot}_{\Delta t}\mathrm{g}_2\| \;\leq\; (1+C\Delta t)\|\mathrm{g}_1 - \mathrm{g}_2\|, \quad \forall n\Delta t \leq T, \;\; \Delta t \;<\; k_0$$

holds for all useful $\mathrm{g}_1,\ \mathrm{g}_2 \in L^{loc}_1(\mathbb{R})$.

It follows from (6.9) that for smooth solutions the order of the local splitting error is equal to the order of the global splitting error, provided the method is stable. Therefore, we study the local splitting error in more detail. Note that the splitting error is independent of the numerical methods used to solve (6.2) and (6.3). As noted before, discontinuous solutions of (6.1) may also occur (e.g. detonation waves). For nonsmooth solutions, a detailed analysis of a particular splitting method for scalar conservation laws is given in [91]. However, to our knowledge, a more general analysis of the local splitting error for problems with nonsmooth solutions is still an open problem. In order to analyse the splitting error for nonlinear systems, we restrict ourselves to smooth solutions. Therefore, we assume that u, v and w are three times continuously differentiable for all x and t, i.e. $\mathrm{u}, \mathrm{v}, \mathrm{w} \in C^3(\mathbb{R} \times [0,\infty))$. For ease of notation, in the remainder of this chapter the partial derivatives are denoted by subscripts. The analysis of the local splitting error is based on the Taylor series expansion

$$\mathrm{u}(x,t+\Delta t) \;=\; \mathrm{u}(x,t) \;+\; \Delta t \mathrm{u}_t(x,t) \;+\; \tfrac{1}{2}\Delta t^2 \mathrm{u}_{tt}(x,t) \;+\; \mathcal{O}(\Delta t^3). \tag{6.10}$$

Recall that the Jacobian matrix $A(\mathrm{u})$ of $\mathbf{f}(\mathrm{u})$ is defined by (3.3). Similarly, we define the Jacobian matrix $B(\mathrm{u})$ of $\mathrm{q}(\mathrm{u})$ by

$$B(\mathrm{u}) := \frac{\partial}{\partial \mathbf{u}} \mathrm{q}(\mathrm{u}).$$

For later purposes it will be useful to introduce the $m \times m$-matrix $\partial A(\mathbf{g})$ for all $\mathbf{g} \in \mathbb{R}^m$ by

$$\partial A(\mathbf{g}) := \begin{pmatrix} \mathbf{g}^T (f_1)_{\mathbf{uu}} \\ \mathbf{g}^T (f_2)_{\mathbf{uu}} \\ \vdots \\ \mathbf{g}^T (f_m)_{\mathbf{uu}} \end{pmatrix},$$

where $\mathbf{f} := (f_1, f_2, \ldots, f_m)^T$ is the flux function and the $m \times m$-matrix $(f_j)_{\mathbf{uu}}$ denotes the Hessian of f_j for $j = 1, \ldots, m$. After a rather technical but straightforward computation, it follows from (6.1) that

$$\begin{aligned} \mathrm{u}_t &= -A(\mathrm{u})\mathrm{u}_x + \mathrm{q}(\mathrm{u}), \\ \mathrm{u}_{tt} &= -\partial A(\mathrm{u}_x)\mathrm{q}(\mathrm{u}) - A(\mathrm{u})B(\mathrm{u})\mathrm{u}_x - B(\mathrm{u})A(\mathrm{u})\mathrm{u}_x + A^2(\mathrm{u})\mathrm{u}_{xx} \\ &\quad + \partial A(\mathrm{u}_x)A(\mathrm{u})\mathrm{u}_x + A(\mathrm{u})\partial A(\mathrm{u}_x)\mathrm{u}_x + B(\mathrm{u})\mathrm{q}(\mathrm{u}). \end{aligned}$$

Subsequently, we substitute the expressions above for u_t and u_{tt} into (6.10), giving

$$\begin{aligned} \mathrm{u}(x, t + \Delta t) &= \mathrm{u}^* + \Delta t \{-A(\mathrm{u}^*)\mathrm{u}^*_x + \mathrm{q}(\mathrm{u}^*)\} \\ &\quad + \tfrac{1}{2}\Delta t^2 \Big\{ - \partial A(\mathrm{u}^*_x)\mathrm{q}(\mathrm{u}^*) - A(\mathrm{u}^*)B(\mathrm{u}^*)\mathrm{u}^*_x - B(\mathrm{u}^*)A(\mathrm{u}^*)\mathrm{u}^*_x \\ &\quad + A^2(\mathrm{u}^*)\mathrm{u}^*_{xx} + \partial A(\mathrm{u}^*_x)A(\mathrm{u}^*)\mathrm{u}^*_x + A(\mathrm{u}^*)\partial A(\mathrm{u}^*_x)\mathrm{u}^*_x \\ &\quad + B(\mathrm{u}^*)\mathrm{q}(\mathrm{u}^*) \Big\} + \mathcal{O}(\Delta t^3), \end{aligned} \tag{6.11}$$

where, for simplicity of notation, we denote the value of the function u at the point (x, t) by u^*, i.e. $\mathrm{u}^* := \mathrm{u}(x, t)$. Furthermore, using (6.2), (6.4) and $\mathrm{v} \in C^3(\mathbb{R} \times [0, \infty))$, it can be shown that

$$\begin{aligned} \mathcal{S}^f_{\Delta t}\mathrm{v}^{\tau_1}(x) &= \mathrm{v}^* - \Delta t A(\mathrm{v}^*)\mathrm{v}^*_x \\ &\quad + \tfrac{1}{2}\Delta t^2 \{A^2(\mathrm{v}^*)\mathrm{v}^*_{xx} + \partial A(\mathrm{v}^*_x)A(\mathrm{v}^*)\mathrm{v}^*_x + A(\mathrm{v}^*)\partial A(\mathrm{v}^*_x)\mathrm{v}^*_x\} + \mathcal{O}(\Delta t^3), \end{aligned} \tag{6.12}$$

where $\mathrm{v}^* := \mathrm{v}^{\tau_1}(x)$. Finally, it follows from (6.3), (6.5) and $\mathrm{w} \in C^3(\mathbb{R} \times [0, \infty))$ that

$$\mathcal{S}^c_{\Delta t}\mathrm{w}^{\tau_2}(x) = \mathrm{w}^* + \Delta t \mathrm{q}(\mathrm{w}^*) + \tfrac{1}{2}\Delta t^2 B(\mathrm{w}^*)\mathrm{q}(\mathrm{w}^*) + \mathcal{O}(\Delta t^3), \tag{6.13}$$

where $\mathrm{w}^* := \mathrm{w}^{\tau_2}(x)$. Next we like to study the local splitting error for two popular splitting methods.

First, we describe the local splitting error for probably the most obvious splitting method. In this case $\mathcal{S}^{tot}_{\Delta t}$ is simply obtained by successively applying $\mathcal{S}^f_{\Delta t}$ and $\mathcal{S}^c_{\Delta t}$ to given data $\tilde{\mathrm{u}}(\cdot, t^n)$ (see method (i)-(iii) in Section 6.1). Hence this particular splitting method is given by

$$\tilde{\mathrm{u}}(\cdot, t^{n+1}) \;=\; \mathcal{S}^{tot}_{\Delta t}\,\tilde{\mathrm{u}}(\cdot, t^n) \;:=\; \mathcal{S}^{c}_{\Delta t}\,\mathcal{S}^{f}_{\Delta t}\,\tilde{\mathrm{u}}(\cdot, t^n), \tag{6.14}$$

where it is assumed that $\mathcal{S}^f_{\Delta t} : D_f \to D_c$, $\mathcal{S}^c_{\Delta t} : D_c \to D_f$ and $\tilde{\mathrm{u}}(\cdot, 0) \in D_f$. For the method above we are now able to prove the following theorem.

Theorem 6.3. *Let* $t > 0$ *be given and assume that the initial value problem* (6.1) *has a solution* $\mathrm{u} \in C^3(\mathbb{R} \times [t, t + \Delta t])$. *Furthermore, assume that the result after one step of the splitting method* (6.14) *and the final result are also three times continuously differentiable on* $\mathbb{R} \times [t, t + \Delta t]$. *Then the local splitting error at the point* (x, t) *of the splitting method* (6.14), *as defined in* (6.7), *satisfies*

$$\begin{aligned}
\mathrm{D}^{spl}_{\Delta t}(x, t) \;&=\; \frac{1}{\Delta t}\{\mathrm{u}(x, t + \Delta t) \;-\; \mathcal{S}^c_{\Delta t}\,\mathcal{S}^f_{\Delta t}\,\mathrm{u}(x, t)\} \\
&=\; \tfrac{1}{2}\Delta t\{-\partial A(\mathrm{u}^*_x)\mathrm{q}(\mathrm{u}^*) - A(\mathrm{u}^*)B(\mathrm{u}^*)\mathrm{u}^*_x + B(\mathrm{u}^*)A(\mathrm{u}^*)\mathrm{u}^*_x\} + \mathcal{O}(\Delta t^2),
\end{aligned} \tag{6.15}$$

where $\mathrm{u}^* := \mathrm{u}(x, t)$ *and* u *is the exact solution of* (6.1).

If (6.1a) is linear, i.e. $f(\mathrm{u}) = A\mathrm{u}$, then $\partial A(\mathrm{g}) = 0$ for all $\mathrm{g} \in \mathbb{R}^m$. Moreover, if the matrices A and $B(\mathrm{u})$ commute, then $\mathrm{D}^{spl}_{\Delta t}(x, t) = \mathcal{O}(\Delta t^2)$. Obviously, this is true for scalar, linear conservation laws. For these scalar, linear problems it can even be shown that the local splitting error of method (6.14) is zero. Equation (6.15) suggests that using splitting method (6.14) at each time step Δt of a numerical method will introduce an error of magnitude $\mathcal{O}(\Delta t)$, leading to a method which is consistent of order one. Hence, independent of the numerical method which is used to approximate (6.2) and (6.3), the splitting method (6.14) is consistent of at most order one. Next we will proof Theorem 6.3.

Proof. We start the proof of Theorem 6.3 by introducing the function $\mathrm{g} \in D_c$ as $\mathrm{g} := \mathcal{S}^f_{\Delta t}\mathrm{u}(\cdot, t)$. From (6.13) it follows directly that

$$\mathcal{S}^c_{\Delta t}\mathcal{S}^f_{\Delta t}\mathrm{u}(x, t) \;=\; \mathrm{g}^* \;+\; \Delta t\,\mathrm{q}(\mathrm{g}^*) \;+\; \tfrac{1}{2}\Delta t^2 B(\mathrm{g}^*)\mathrm{q}(\mathrm{g}^*) \;+\; \mathcal{O}(\Delta t^3), \tag{6.16}$$

where $\mathrm{g}^* \in \mathbb{R}^m$ is given by $\mathrm{g}^* := \mathcal{S}^f_{\Delta t}\mathrm{u}(x, t)$. Let $\mathrm{u}^* := \mathrm{u}(x, t)$. Since (6.12) holds with $\mathrm{v}^* = \mathrm{v}^{\tau_1}(x) = \mathrm{u}^*$ (i.e. $\tau_1 = t$), a straightforward Taylor series expansion shows

$$\mathrm{q}(\mathrm{g}^*) \;=\; \mathrm{q}(\mathrm{u}^*) \;-\; \Delta t B(\mathrm{u}^*)A(\mathrm{u}^*)\mathrm{u}^*_x \;+\; \mathcal{O}(\Delta t^2)$$

and

$$B(\mathrm{g}^*)\mathrm{q}(\mathrm{g}^*) \;=\; B(\mathrm{u}^*)\mathrm{q}(\mathrm{u}^*) \;+\; \mathcal{O}(\Delta t).$$

Subsequently, we substitute the latter two equations and (6.12), with $\mathrm{v}^* = \mathrm{u}^* = \mathrm{u}(x, t)$, into (6.16), which gives

$$\begin{aligned}
\mathcal{S}^c_{\Delta t}\,\mathcal{S}^f_{\Delta t}\,\mathrm{u}(x, t) \;&=\; \mathrm{u}^* \;+\; \Delta t\,\{-A(\mathrm{u}^*)\mathrm{u}^*_x + \mathrm{q}(\mathrm{u}^*)\} \\
&\quad + \tfrac{1}{2}\Delta t^2\Big\{ -2B(\mathrm{u}^*)A(\mathrm{u}^*)\mathrm{u}^*_x + A^2(\mathrm{u}^*)\mathrm{u}^*_{xx} + \partial A(\mathrm{u}^*_x)A(\mathrm{u}^*)\mathrm{u}^*_x \\
&\qquad + A(\mathrm{u}^*)\partial A(\mathrm{u}^*_x)\mathrm{u}^*_x + B(\mathrm{u}^*)\mathrm{q}(\mathrm{u}^*)\Big\} \;+\; \mathcal{O}(\Delta t^3).
\end{aligned}$$

If the latter equation is subtracted from (6.11), then it follows immediately that (6.15) holds. This completes the proof. □

Strang [84] pointed out that the accuracy of the splitting method (6.14) can be increased if we use a slightly different product of the solution operators $\mathcal{S}^f_{\Delta t}$ and $\mathcal{S}^c_{\Delta t}$. The *Strang splitting* approximates the exact solution $\mathbf{u}$ of (6.1) by the following three-step splitting method

$$\tilde{\mathbf{u}}(\cdot, t^{n+1}) = \mathcal{S}^{tot}_{\Delta t}\, \tilde{\mathbf{u}}(\cdot, t^n) := \mathcal{S}^c_{\Delta t/2}\, \mathcal{S}^f_{\Delta t}\, \mathcal{S}^c_{\Delta t/2}\, \tilde{\mathbf{u}}(\cdot, t^n), \tag{6.17}$$

where it is assumed that $\mathcal{S}^f_{\Delta t} : D_f \to D_c$, $\mathcal{S}^c_{\Delta t/2} : D_c \to D_f$ and $\tilde{\mathbf{u}}(\cdot, t^n) \in D_c$ for all $n \geq 0$. The latter assumptions imply that $\mathcal{S}^{tot}_{\Delta t}$ is well defined. We are now able to prove the following theorem.

Theorem 6.4. *Let $t > 0$ be given and assume that the initial value problem* (6.1) *has a solution $\mathbf{u} \in C^3(I\!R \times [t, t + \Delta t])$. Furthermore, assume that the results after the first and second step of the splitting method* (6.17) *and the final result $\tilde{\mathbf{u}}$ are also three times continuously differentiable on $I\!R \times [t, t + \Delta t]$. Then the splitting error at the point (x, t) of the splitting method* (6.17), *as defined in* (6.7), *satisfies*

$$\mathbf{D}^{spl}_{\Delta t}(x, t) = \frac{1}{\Delta t}\{\mathbf{u}(x, t + \Delta t) - \mathcal{S}^c_{\Delta t/2}\, \mathcal{S}^f_{\Delta t}\, \mathcal{S}^c_{\Delta t/2}\mathbf{u}(x, t)\} = \mathcal{O}(\Delta t^2), \tag{6.18}$$

where $\mathbf{u}$ is the exact solution of (6.1).

Also for method (6.17) it can be shown that for scalar, linear conservation laws no splitting error error is made. It follows from equation (6.18) that the splitting method (6.17) in each time step Δt of a numerical method will introduce an error $\mathcal{O}(\Delta t^2)$ resulting in a method of order two. Therefore, independent of the numerical method used to approximate (6.2) and (6.3), the splitting method (6.17) is consistent of at most order two. In the following chapters we will discuss, among other things, a numerical method which alternates between a high resolution method for (6.2) and a stable (stiff) solver for the system of ODEs (6.3), using the Strang splitting (6.17). At a first look a numerical method using the Strang splitting (6.17) seems to be more expensive to implement than a numerical method using the first order method (6.14). In the next section we will see that the computational costs of method (6.17) are only slightly higher than that of the first order method (6.14).

Proof. We start the proof of Theorem 6.4 by introducing the function $\mathbf{g} \in D_c$ as $\mathbf{g} := \mathcal{S}^f_{\Delta t}\mathcal{S}^c_{\Delta t/2}\mathbf{u}(\cdot, t)$. From (6.13) it follows directly that

$$\mathcal{S}^c_{\Delta t/2}\mathcal{S}^f_{\Delta t}\mathcal{S}^c_{\Delta t/2}\mathbf{u}(x, t) = \mathbf{g}^* + \tfrac{1}{2}\Delta t\mathbf{q}(\mathbf{g}^*) + \tfrac{1}{8}\Delta t^2 B(\mathbf{g}^*)\mathbf{q}(\mathbf{g}^*) + \mathcal{O}(\Delta t^3), \tag{6.19}$$

where $\mathbf{g}^* \in I\!R^m$ is given by $\mathbf{g}^* := \mathcal{S}^f_{\Delta t}\mathcal{S}^c_{\Delta t/2}\mathbf{u}(x, t)$. In the remainder of the proof $\mathbf{u}^* = \mathbf{u}(x, t)$. Furthermore, (6.12) holds with $\mathbf{v}^* = \mathcal{S}^c_{\Delta t/2}\mathbf{u}(x, t)$. Equation (6.13) implies that $\mathbf{v}^*$ satisfies

$$\mathbf{v}^* = \mathcal{S}^c_{\Delta t/2}\mathbf{u}(x, t) = \mathbf{u}^* + \tfrac{1}{2}\Delta t\mathbf{q}(\mathbf{u}^*) + \tfrac{1}{8}\Delta t^2 B(\mathbf{u}^*)\mathbf{q}(\mathbf{u}^*) + \mathcal{O}(\Delta t^3).$$

In order to eliminate the higher order terms in (6.12), we use the equation above, (6.1a) and Taylor series expansions. A technical but straightforward computation reveals that

$$A(\mathrm{v}^*)\mathrm{v}^*_x = A(\mathrm{u}^*)\mathrm{u}^*_x + \tfrac{1}{2}\Delta t\{A(\mathrm{u}^*)B(\mathrm{u}^*)\mathrm{u}^*_x + \partial A(\mathrm{u}^*_x)\mathrm{q}(\mathrm{u}^*)\} + \mathcal{O}(\Delta t^2)$$

and

$$\begin{aligned} A^2(\mathrm{v}^*)\mathrm{v}^*_{xx} &= A^2(\mathrm{u}^*)\mathrm{u}^*_{xx} + \mathcal{O}(\Delta t), \\ \partial A(\mathrm{v}^*_x)A(\mathrm{v}^*)\mathrm{v}^*_x &= \partial A(\mathrm{u}^*_x)A(\mathrm{u}^*)\mathrm{u}^*_x + \mathcal{O}(\Delta t), \\ A(\mathrm{v}^*)\partial A(\mathrm{v}^*_x)\mathrm{v}^*_x &= A(\mathrm{u}^*)\partial A(\mathrm{u}^*_x)\mathrm{u}^*_x + \mathcal{O}(\Delta t), \end{aligned}$$

where $\mathrm{v}^* := \mathcal{S}^c_{\Delta t/2}\mathrm{u}(x, t)$. After substituting the equations above into (6.12) we obtain

$$\begin{aligned} \mathrm{g}^* = {} & \mathrm{u}^* + \Delta t\,\{-A(\mathrm{u}^*)\mathrm{u}^*_x + \tfrac{1}{2}\mathrm{q}(\mathrm{u}^*)\} \\ & + \tfrac{1}{2}\Delta t^2\Big\{ - \partial A(\mathrm{u}^*_x)\mathrm{q}(\mathrm{u}^*) - A(\mathrm{u}^*)B(\mathrm{u}^*)\mathrm{u}^*_x + A^2(\mathrm{u}^*)\mathrm{u}^*_{xx} + \partial A(\mathrm{u}^*_x)A(\mathrm{u}^*)\mathrm{u}^*_x \\ & \qquad + A(\mathrm{u}^*)\partial A(\mathrm{u}^*_x)\mathrm{u}^*_x + \tfrac{1}{4}B(\mathrm{u}^*)\mathrm{q}(\mathrm{u}^*)\Big\} + \mathcal{O}(\Delta t^3). \end{aligned}$$

Using this we deduce

$$\mathrm{q}(\mathrm{g}^*) = \mathrm{q}(\mathrm{u}^*) + \Delta t\{-B(\mathrm{u}^*)A(\mathrm{u}^*)\mathrm{u}^*_x + \tfrac{1}{2}B(\mathrm{u}^*)\mathrm{q}(\mathrm{u}^*)\} + \mathcal{O}(\Delta t^2)$$

and

$$B(\mathrm{g}^*)\mathrm{q}(\mathrm{g}^*) = B(\mathrm{u}^*)\mathrm{q}(\mathrm{u}^*) + \mathcal{O}(\Delta t).$$

Subsequently, substituting the latter three equations into (6.19) leads to

$$\begin{aligned} \mathcal{S}^c_{\Delta t/2}\mathcal{S}^f_{\Delta t}\mathcal{S}^c_{\Delta t/2}\,\mathrm{u}(x,t) = {} & \mathrm{u}^* + \Delta t\,\{-A(\mathrm{u}^*)\mathrm{u}^*_x + \mathrm{q}(\mathrm{u}^*)\} \\ & + \tfrac{1}{2}\Delta t^2\Big\{ - \partial A(\mathrm{u}^*_x)\mathrm{q}(\mathrm{u}^*) - A(\mathrm{u}^*)B(\mathrm{u}^*)\mathrm{u}^*_x - B(\mathrm{u}^*)A(\mathrm{u}^*)\mathrm{u}^*_x \\ & \qquad + A^2(\mathrm{u}^*)\mathrm{u}^*_{xx} + \partial A(\mathrm{u}^*_x)A(\mathrm{u}^*)\mathrm{u}^*_x + A(\mathrm{u}^*)\partial A(\mathrm{u}^*_x)\mathrm{u}^*_x \\ & \qquad + B(\mathrm{u}^*)\mathrm{q}(\mathrm{u}^*)\Big\} + \mathcal{O}(\Delta t^3). \end{aligned}$$

Finally, subtracting the latter equation from (6.11) implies that (6.18) holds. This completes the proof. □

Example 6.5. In this example we like to illustrate Theorem 6.3 and Theorem 6.4. Before we illustrate the order of the splitting methods we remark that the natural norm for conservation laws is the L_1-norm $\|\cdot\|_1$, defined for a general function $\mathrm{v}(x)$ by

$$\|\mathrm{v}\|_1 := \int_{-\infty}^{\infty} |\mathrm{v}(x)|\mathrm{d}x.$$

This norm is natural since it essentially requires only just integrating the function; note that the integral form of the conservation law (3.1) often allow us to estimate these integrals.

As noted before, it follows from (6.9) that for smooth solutions the order of the local splitting error is equal to the order of the global splitting error, provided the method is stable. Next we check for a particular conservation law, whether the order of the splitting method (6.14) equals one, as appears from (6.15) and whether the order of the splitting method (6.17) equals two, as appears from (6.18). Therefore, we consider the following scalar initial value problem

$$\frac{\partial}{\partial t}u + u\frac{\partial}{\partial x}u = -\mu u(1-u)^2, \tag{6.20a}$$

$$\mathrm{u}(x,0) = \frac{\exp(\mu x)}{1+\exp(\mu x)}, \quad \forall x \in I\!R. \tag{6.20b}$$

where $\mu > 0$ is some constant. The source term in (6.20a) admits two equilibrium states as solutions of the underlying characteristic equation, namely: $u = 0$ and $u = 1$. The state $u = 0$ is a stable equilibrium state. The exact solution of (6.20) is given by

$$u(x,t) = \frac{\exp(\mu(x-t))}{1+\exp(\mu(x-t))}.$$

Note that $\lim_{x\to-\infty} u(x,t) = 0$ and $\lim_{x\to\infty} u(x,t) = 1$ for all fixed t. Since these limits are also valid for the solution of the splitting method $\tilde{\mathrm{u}}$, the L_1-norm of the global splitting error $E^{spl}_{\Delta t}(\cdot, t^n)$ is expected to be finite for all finite n.

	global splitting error $\|E^{spl}_{\Delta t}(\cdot, t^n)\|_1$	
Δt	method (6.14)	method (6.17)
1/5	$3.01 \cdot 10^{-2}$	$5.10 \cdot 10^{-4}$
1/10	$1.52 \cdot 10^{-2}$	$1.28 \cdot 10^{-4}$
1/20	$7.69 \cdot 10^{-3}$	$3.27 \cdot 10^{-5}$
1/40	$3.91 \cdot 10^{-3}$	$8.22 \cdot 10^{-6}$
1/80	$2.02 \cdot 10^{-3}$	$2.11 \cdot 10^{-6}$

Table 6.1. The global splitting error at $t^n = 1$ for the splitting methods (6.14) and (6.17) applied to problem (6.20) with $\mu = 5$, $n = 1/\Delta t$.

The results in Table 6.1 clearly illustrate the first order behaviour of (6.14) and the second order behaviour of (6.17), as was predicted by Theorem 6.3 and 6.4. □

6.3. NUMERICAL METHODS BASED ON TIME SPLITTING

In this section we describe a possible way to develop numerical methods for (6.1) based on splitting methods. The basic idea is to replace the solution operators $\mathcal{S}^f_{\Delta t}$ and $\mathcal{S}^c_{\Delta t}$

by approximate operators. Let $\mathcal{L}^f_{\Delta t} : L_1^{loc}(\mathbb{R}) \to L_1^{loc}(\mathbb{R})$ represent the *approximate solution operator* of problem (6.2) at time $\tau_1 + \Delta t$, i.e. (see (4.8))

$$\mathbf{V}_{\Delta t}(\cdot, \tau_1 + \Delta t) = \mathcal{L}^f_{\Delta t} \mathbf{V}_{\Delta t}(\cdot, \tau_1).$$

Similarly, let $\mathcal{L}^c_{\Delta t} : L_1^{loc}(\mathbb{R}) \to L_1^{loc}(\mathbb{R})$ represent the approximate solution operator of problem (6.3) at time $\tau_2 + \Delta t$, i.e.

$$\mathbf{W}_{\Delta t}(\cdot, \tau_2 + \Delta t) = \mathcal{L}^c_{\Delta t} \mathbf{W}_{\Delta t}(\cdot, \tau_2).$$

Here $\mathbf{V}_{\Delta t}$ and $\mathbf{W}_{\Delta t}$ are piecewise continuous functions (e.g. as (4.4)). Again we consider some product of the operators $\mathcal{L}^f_{\Delta t}$ and $\mathcal{L}^c_{\Delta t}$. We first study the local error of the total method. The local error $\mathbf{D}^{tot}_{\Delta t}(x, t)$ measures how well the numerical method approximates the exact solution of (6.1) locally after a time step Δt. We expect the local error to be composed of the local splitting error $\mathbf{D}^{spl}_{\Delta t}(x, t)$ and the local discretization errors made by approximating (6.2) (i.e. $\mathbf{D}^f_{\Delta t}(x, t)$) and (6.3) (i.e. $\mathbf{D}^c_{\Delta t}(x, t)$). Let $\mathcal{L}^{tot}_{\Delta t} : L_1^{loc}(\mathbb{R}) \to L_1^{loc}(\mathbb{R})$ be some product of the approximate solution operators $\mathcal{L}^f$ and $\mathcal{L}^c$, then the numerical method is given by (see (4.8))

$$\mathbf{U}_{\Delta t}(\cdot, t + \Delta t) = \mathcal{L}^{tot}_{\Delta t} \mathbf{U}_{\Delta t}(\cdot, t). \tag{6.21}$$

Equation (6.21) is the numerical version of (6.6). Similarly as in the previous section we write $\mathcal{L}^j_{\Delta t}\mathbf{g}(x, t) = (\mathcal{L}^j_{\Delta t}\mathbf{g}(\cdot, t))(x)$ for $j = f, c, tot$. If now $\mathbf{U}_{\Delta t}$ is replaced by the exact solution of (6.1), then the equality will not hold exactly in general. This leads to the following definition (see Definition 6.1).

Definition 6.6. *Consider a numerical splitting method written in the generic form* (6.21). *The* local error $\mathbf{D}^{tot}_{\Delta t}$ *of this method at the point* (x, t^n) *is defined by*

$$\mathbf{D}^{tot}_{\Delta t}(x, t^n) := \frac{1}{\Delta t}\{\mathbf{u}(x, t^n + \Delta t) - \mathcal{L}^{tot}_{\Delta t}\, \mathbf{u}(x, t^n)\}, \tag{6.22}$$

where $\mathbf{u}$ *is the exact solution of* (6.1).

Note that the local error depends on the exact solution $\mathbf{u}$. Furthermore, we remark that if $\mathbf{q} = \mathbf{0}$, the local error can be interpreted as the local discretization error (see Definition 4.2). It follows from (6.7) and (6.22) that

$$\mathbf{D}^{tot}_{\Delta t}(\cdot, t^n) = \mathbf{D}^{spl}_{\Delta t}(\cdot, t^n) + \frac{1}{\Delta t}\{\mathcal{S}^{tot}_{\Delta t}\, \mathbf{u}(\cdot, t^n) - \mathcal{L}^{tot}_{\Delta t}\, \mathbf{u}(\cdot, t^n)\}. \tag{6.23}$$

Since $\mathbf{D}^{spl}_{\Delta t}$ is studied in detail in Section 6.2, the analysis in this section is restricted to the term $(\mathcal{S}^{tot}_{\Delta t}\mathbf{u}(\cdot, t^n) - \mathcal{L}^{tot}_{\Delta t}\mathbf{u}(\cdot, t^n))/\Delta t$.

The local discretization error $\mathbf{D}^f_{\Delta t}$ made in approximating (6.2a) is defined by

$$\mathbf{D}^f_{\Delta t}(x, t^n) := \frac{1}{\Delta t}\{\mathbf{v}(x, t^n + \Delta t) - \mathcal{L}^f_{\Delta t}\, \mathbf{v}(x, t^n)\}, \tag{6.24}$$

where $\mathbf{v}$ is the exact solution of (6.2a) with initial data $\mathbf{v}^{\tau_1} = \mathbf{v}(\cdot, t^n)$ (i.e. $\tau_1 = t^n$). Note that (6.24) is completely similar to the local discretization error defined in Definition 4.2, if a $(2k+1)$-point method written in the generic form (4.8) is used. We define the local discretization error $\mathbf{D}^c_{\Delta t}$ made in approximating (6.3a) by

$$\mathbf{D}^c_{\Delta t}(x, t^n) := \frac{1}{\Delta t}\{\mathbf{w}(x, t^n + \Delta t) - \mathcal{L}^c_{\Delta t}\, \mathbf{w}(x, t^n)\},$$

where $\mathbf{w}$ is the exact solution of (6.3a) with initial data $\mathbf{w}^{\tau_2} = \mathbf{w}(\cdot, t^n)$ (i.e. $\tau_2 = t^n$). The *global error* $\mathbf{E}^{tot}_{\Delta t}(x, t)$ is defined as

$$\mathbf{E}^{tot}_{\Delta t}(x, t^n) := \mathbf{U}_{\Delta t}(x, t^n) - \mathbf{u}(x, t^n) = (\mathcal{L}^{tot}_{\Delta t})^n\, \mathbf{U}_{\Delta t}(x, 0) - \mathbf{u}(x, t^n), \qquad (6.25)$$

where the superscript n for $\mathcal{L}^{tot}_{\Delta t}$ represents the nth power of the operator.

Again it can be shown that for smooth solutions the order of the local error is equal to the order of the global error, provided all operators are stable. Therefore, we study the local error in more detail.

First, we describe the local error for the numerical method based on the splitting method (6.14), i.e.

$$\mathbf{U}_{\Delta t}(\cdot, t^{n+1}) = \mathcal{L}^{tot}_{\Delta t}\, \mathbf{U}_{\Delta t}(\cdot, t^n) := \mathcal{L}^c_{\Delta t}\, \mathcal{L}^f_{\Delta t}\, \mathbf{U}_{\Delta t}(\cdot, t^n). \qquad (6.26)$$

For the method above we are now able to prove the following theorem.

Theorem 6.7. *Assume that all the assumptions of Theorem 6.3 hold. Let some norm* $\|\cdot\|$ *be given and suppose that* $\mathcal{S}^c_{\Delta t}$ *is stable in this norm, as defined in Definition 6.2. Furthermore, assume that* $\mathcal{L}^f_{\Delta t}$ *is consistent of order* p_1 *with the conservation law* (6.2) *and* $\mathcal{L}^c_{\Delta t}$ *is consistent of order* p_2 *with the differential equation* (6.3) *(see Definition 4.3). Then for each time* $T > 0$ *there exists constants* C_1, C_2, C_3 *and* k_0 *such that the local error of method* (6.26) *satisfies*

$$\begin{aligned} \|\mathbf{D}^{tot}_{\Delta t}(\cdot, t^n)\| &= \frac{1}{\Delta t}\|\mathbf{u}(\cdot, t^n + \Delta t) - \mathcal{L}^c_{\Delta t}\, \mathcal{L}^f_{\Delta t}\, \mathbf{u}(\cdot, t^n)\| \\ &\leq C_1 \Delta t + C_2 \Delta t^{p_1} + C_3 \Delta t^{p_2}, \quad \forall n\Delta t \leq T, \ \Delta t < k_0, \end{aligned} \qquad (6.27)$$

where $\mathbf{u}$ *is the exact solution of* (6.1).

As we expected, method (6.26) is consistent of at most order one. Therefore, it is sufficient to use first order operators $\mathcal{L}^f_{\Delta t}$ and $\mathcal{L}^c_{\Delta t}$. In other words, it makes no sense to use some high resolution method for $\mathcal{L}^f_{\Delta t}$, whenever the first order splitting method (6.26) is used.

Proof. From (6.23) and the triangular inequality it follows that

$$\begin{aligned} \|\mathbf{D}^{tot}_{\Delta t}(\cdot, t^n)\| &\leq \|\mathbf{D}^{spl}_{\Delta t}(\cdot, t^n)\| + \frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t}\mathcal{S}^f_{\Delta t}\mathbf{u}(\cdot, t^n) - \mathcal{L}^c_{\Delta t}\mathcal{L}^f_{\Delta t}\mathbf{u}(\cdot, t^n)\| \\ &\leq \|\mathbf{D}^{spl}_{\Delta t}(\cdot, t^n)\| + \frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t}\mathcal{S}^f_{\Delta t}\mathbf{u}(\cdot, t^n) - \mathcal{S}^c_{\Delta t}\mathcal{L}^f_{\Delta t}\mathbf{u}(\cdot, t^n)\| \\ &\quad + \frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t}\mathcal{L}^f_{\Delta t}\mathbf{u}(\cdot, t^n) - \mathcal{L}^c_{\Delta t}\mathcal{L}^f_{\Delta t}\mathbf{u}(\cdot, t^n)\|. \end{aligned} \qquad (6.28)$$

Theorem 6.3 implies that there exists a constant C_1 such that $\|\mathbf{D}^{spl}_{\Delta t}(\cdot, t^n)\| \leq C_1 \Delta t$. Since $\mathcal{S}^c_{\Delta t}$ is stable and $\mathcal{L}^f_{\Delta t}$ is consistent of order p_1, we deduce

$$\begin{aligned}\frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t}\mathcal{S}^f_{\Delta t}\mathrm{u}(\cdot, t^n) - \mathcal{S}^c_{\Delta t}\mathcal{L}^f_{\Delta t}\mathrm{u}(\cdot, t^n)\| &\leq (1 + C\Delta t)\frac{1}{\Delta t}\|\mathcal{S}^f_{\Delta t}\mathrm{u}(\cdot, t^n) - \mathcal{L}^f_{\Delta t}\mathrm{u}(\cdot, t^n)\| \\ &\leq C_2 \Delta t^{p_1},\end{aligned}$$

for some positive constants C and C_2. Finally, it follows from the consistency of order p_2 of $\mathcal{L}^c_{\Delta t}$ that there exists a constant C_3 such that

$$\frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t}\mathcal{L}^f_{\Delta t}\mathrm{u}(\cdot, t^n) - \mathcal{L}^c_{\Delta t}\mathcal{L}^f_{\Delta t}\mathrm{u}(\cdot, t^n)\| \leq C_3 \Delta t^{p_2}.$$

Substituting the above inequalities into (6.28) gives (6.27). This completes the proof. □

Subsequently, we replace the exact solution operators $\mathcal{S}^f_{\Delta t}$ and $\mathcal{S}^c_{\Delta t}$ by numerical approximations $\mathcal{L}^f_{\Delta t}$ and $\mathcal{L}^c_{\Delta t}$ in the Strang splitting method (6.17). The numerical method based on the Strang splitting is given by [84]

$$\mathrm{U}_{\Delta t}(\cdot, t^{n+1}) = \mathcal{L}^{tot}_{\Delta t}\mathrm{U}_{\Delta t}(\cdot, t^n) := \mathcal{L}^c_{\Delta t/2}\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathrm{U}_{\Delta t}(\cdot, t^n). \qquad (6.29)$$

The computational costs of method (6.29) seem to be much higher than the computational costs of the first order method (6.26), since three applications of the numerical operators are required at each time step rather than two. In practice, however, several time steps are combined to yield

$$\mathrm{U}_{\Delta t}(\cdot, t^{n+1}) = \mathcal{L}^c_{\Delta t/2}(\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t})^n \mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathrm{U}_{\Delta t}(\cdot, 0).$$

In this form the costs of this method are only slightly higher than that of the first order method (6.26), i.e.

$$\mathrm{U}_{\Delta t}(\cdot, t^{n+1}) = (\mathcal{L}^c_{\Delta t}\mathcal{L}^f_{\Delta t})^n \mathrm{U}_{\Delta t}(\cdot, 0).$$

Only at the beginning and end of the computation (and at any intermediate times where output is desired) the "half-step operators" $\mathcal{L}^c_{\Delta t/2}$ must be employed. For method (6.29) we proof the following theorem.

Theorem 6.8. *Assume that all the assumptions of Theorem 6.4 be satisfied. Let some norm $\|\cdot\|$ be given and assume that $\mathcal{S}^f_{\Delta t}$ and $\mathcal{S}^c_{\Delta t}$ are stable in this norm, as defined in Definition 6.2. Furthermore, assume that $\mathcal{L}^f_{\Delta t}$ is consistent of order p_1 with the conservation law* (6.2) *and $\mathcal{L}^c_{\Delta t}$ is consistent of order p_2 with the differential equation* (6.3) *(see Definition 4.3). Then for each time $T > 0$ there exists constants C_1, C_2, C_3 and k_0 such that the local error of method* (6.29) *satisfies*

$$\begin{aligned}\|\mathbf{D}^{tot}_{\Delta t}(\cdot, t^n)\| &= \frac{1}{\Delta t}\|\mathrm{u}(\cdot, t^n + \Delta t) - \mathcal{L}^c_{\Delta t/2}\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathrm{u}(\cdot, t^n)\| \\ &\leq C_1\Delta t^2 + C_2\Delta t^{p_1} + C_3\Delta t^{p_2}, \quad \forall n\Delta t \leq T, \quad \Delta t < k_0,\end{aligned} \qquad (6.30)$$

where u *is the exact solution of* (6.1).

As we expected method (6.29) is consistent of at most order two. Therefore, it is sufficient to use second order operators $\mathcal{L}^f_{\Delta t}$ and $\mathcal{L}^c_{\Delta t}$.

Proof. We start the proof by using (6.23) and the triangular inequality to obtain

$$\begin{aligned}
\|\mathbf{D}^{tot}_{\Delta t}(\cdot, t^n)\| \;\le\;& \|\mathbf{D}^{spl}_{\Delta t}(\cdot, t^n)\| + \frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t/2}\mathcal{S}^f_{\Delta t}\mathcal{S}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{L}^c_{\Delta t/2}\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\| \\
\le\;& \|\mathbf{D}^{spl}_{\Delta t}(\cdot, t^n)\| + \frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t/2}\mathcal{S}^f_{\Delta t}\mathcal{S}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{S}^c_{\Delta t/2}\mathcal{S}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\| \\
&+ \frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t/2}\mathcal{S}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{S}^c_{\Delta t/2}\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\| \\
&+ \frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t/2}\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{L}^c_{\Delta t/2}\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\|.
\end{aligned} \tag{6.31}$$

From Theorem 6.4 it follows that there exists a constant C_1 such that $\|\mathbf{D}^{spl}_{\Delta t}(\cdot, t^n)\| \le C_1 \Delta t^2$. Since $\mathcal{S}^f_{\Delta t}$ and $\mathcal{S}^c_{\Delta t}$ are stable in the sense of Definition 6.2 and $\mathcal{S}^c_{\Delta t}$ is consistent of order p_2, it is straightforward to see that there exist constants A_1, A_2 and A_3 such that

$$\begin{aligned}
&\frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t/2}\mathcal{S}^f_{\Delta t}\mathcal{S}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{S}^c_{\Delta t/2}\mathcal{S}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\| \;\le \\
&\qquad (1 + A_1\Delta t)\,\frac{1}{\Delta t}\|\mathcal{S}^f_{\Delta t}\mathcal{S}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{S}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\| \;\le \\
&\qquad (1 + A_1\Delta t)\,(1 + A_2\Delta t)\frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\| \;\le\; A_3\Delta t^{p_2}.
\end{aligned}$$

The stability of $\mathcal{S}^c_{\Delta t}$ and the consistency of order p_1 of $\mathcal{S}^f_{\Delta t}$ imply that the second term at the right hand side of (6.31) can be estimated by

$$\begin{aligned}
&\frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t/2}\mathcal{S}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{S}^c_{\Delta t/2}\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\| \;\le \\
&\qquad (1 + A_4\Delta t)\,\frac{1}{\Delta t}\|\mathcal{S}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\| \;\le\; C_2\Delta t^{p_1},
\end{aligned}$$

where A_4 and C_2 are positive constants. Finally, we estimate the last term of inequality (6.31). Since $\mathcal{S}^c_{\Delta t}$ is consistent of order p_2, there exists a constant A_5 such that

$$\frac{1}{\Delta t}\|\mathcal{S}^c_{\Delta t/2}\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n) - \mathcal{L}^c_{\Delta t/2}\mathcal{L}^f_{\Delta t}\mathcal{L}^c_{\Delta t/2}\mathbf{u}(\cdot, t^n)\| \;\le\; A_5\Delta t^{p_2}.$$

If we substitute the latter inequalities into (6.31) and C_3 is defined as $C_3 := A_3 + A_5$, then we obtain (6.30). This completes the proof. □

Example 6.9. In this example we like to illustrate Theorem 6.7 and Theorem 6.8. We illustrate the order of the numerical splitting methods (6.26) and (6.29), using the L_1-norm as described in Example 6.5. As remarked before, for smooth solutions the order of the local error $\mathbf{D}^{tot}_{\Delta t}$ is equal to the order of the global error $\mathbf{E}^{tot}_{\Delta t}$, provided the numerical method is stable. In this example we check for model problem (6.20), whether the order of the numerical method (6.26) equals one, as appears from (6.27) and whether the order of the method (6.29) equals two, as appears from (6.30).

For method (6.26) we use Roe's first order method (see Section 4.5) and the forward Euler method. For method (6.29) we use Roe's second order method (see Example 5.17) and the trapezoidal method. In all the results the ratio of Δt and Δx remains constant.

Δt	global error $\|E_{\Delta t}^{tot}(\cdot, t^n)\|_1$	
	method (6.26)	method (6.29)
1/5	$1.24 \cdot 10^{-1}$	$7.57 \cdot 10^{-2}$
1/10	$6.95 \cdot 10^{-2}$	$2.78 \cdot 10^{-2}$
1/20	$3.75 \cdot 10^{-2}$	$8.35 \cdot 10^{-3}$
1/40	$1.98 \cdot 10^{-2}$	$2.25 \cdot 10^{-3}$
1/80	$1.02 \cdot 10^{-2}$	$5.79 \cdot 10^{-4}$

Table 6.2. Global error at $t^n = 1$ for the numerical splitting methods (6.26) and (6.29) applied to problem (6.20) with $\mu = 5, n = 1/\Delta t, \Delta t/\Delta x = 0.5$; in method (6.26), $\mathcal{L}_{\Delta t}^f$ is given by Roe's first order method and $\mathcal{L}_{\Delta t}^c$ is given by the forward Euler method and in method (6.29), $\mathcal{L}_{\Delta t}^f$ is given by Roe's second order method and $\mathcal{L}_{\Delta t}^c$ is given by the trapezoidal method.

In Table 6.2 we take a relatively small μ, since otherwise we observe the first order (second order, respectively) behaviour only for very small time steps. The results in Table 6.2 clearly illustrate the first order behaviour of (6.26) and the second order behaviour of (6.29), as predicted by Theorem 6.7 and 6.8. □

In Theorem 6.7 it is shown that method (6.26) is consistent of at most order one. Therefore, it is sufficient to use first order operators $\mathcal{L}_{\Delta t}^f$ and $\mathcal{L}_{\Delta t}^c$. In other words, it makes no sense to use some high resolution method for $\mathcal{L}_{\Delta t}^f$, whenever the first order splitting method (6.26) is used.

Method (6.29) is shown to be consistent of order two for smooth solutions. Hence, the best results are obtained if we combine in (6.29) a high resolution method for $\mathcal{L}_{\Delta t}^f$ with some second order (stiff) ODE solver for $\mathcal{S}_{\Delta t}^c$. On the one hand if we use a first order method for $\mathcal{L}_{\Delta t}^f$, then the second order accuracy of (6.29) is lost (see (6.30)) and from a computational point of view it is better to use the first order splitting. On the other hand, if we use a third order method for $\mathcal{L}_{\Delta t}^f$ (or $\mathcal{L}_{\Delta t}^c$), the overall method (6.29) remains second order, whereas the computational costs to obtain third order approximations for $\mathcal{L}_{\Delta t}^f$ (or $\mathcal{L}_{\Delta t}^c$) are relatively high.

In Chapter 7 we will apply the first order method (6.26) and the second order method (6.29) to a simplified detonation model (i.e. Burgers' equation with source term). In Chapter 8, both methods are applied to the one-dimensional reactive Euler equations (2.26).

7

NUMERICAL SOLUTION OF A SIMPLIFIED DETONATION MODEL

When attempting to solve the reactive Euler equations (2.26) numerically, we encounter problems which are not present in nonreacting flows. For fast reactions it is possible to obtain stable numerical solutions which look reasonable and yet are completely wrong, because the discontinuities have the wrong locations. Thus, the numerical reaction waves are propagating with nonphysical wave speeds. These "wrong solutions" turn out to be approximations of nonphysical weak solutions. Although this phenomenon has been observed by several other authors [3, 16, 31, 58], a detailed analysis of the occurrence of wrong wave speeds has not been given so far.

Since the reactive Euler equations are complicated, it is not surprising that simpler qualitative models have been developed. In this chapter the simplified detonation model is introduced and studied [60]. This 2×2 system of equations essentially consists of Burgers' equation completed by an extra equation describing the chemical reaction. The chemical reaction is described by an ignition model. In this model there exists an ignition value u_{ign} such that the reaction rate is large when $u \geq u_{ign}$ and zero otherwise.

We observe the same essential numerical difficulty of approximating incorrect weak solutions for low ignition values in the simplified detonation model. However, in most practical applications the ignition value is much higher than the value of u in the unburnt gas. Our numerical results illustrate that for these ignition values, the nonphysical weak solutions will not occur.

For a scalar model problem the numerical wave speed has been studied in [31, 58]. However, we have experienced that for analysing numerical methods the simplified detonation model is a much better model problem. In this chapter we present new theoretical insights for explaining the strong influence of the ignition value on the numerical solution of the simplified detonation model. The incorrect weak solution appears to be a weak detonation wave followed by an ordinary shock wave. We use this to obtain a simple criterion on the ignition value of the chemical model, which ensures that even for relatively coarse meshes, the numerical solution approximates the physically correct weak solution.

This chapter is organized as follows. In the first section the simplified detonation model is presented. Furthermore, we describe the analogues of the reactive Rankine-

Hugoniot equations (see Section 2.4) and the ZND model (see Section 2.5). In Section 7.2 we present a numerical method based on the first order splitting method (6.14), where, for the sake of simplicity, we only describe a combination of Roe's method and the backward Euler method. Numerical results show the occurrence of nonphysical wave speeds. In Section 7.3 we show that nonphysical solutions are always weak detonation waves. The latter property is used to obtain the desired extra criterion which excludes the nonphysical weak solutions. Furthermore, we present numerical results showing that the extra criterion is a useful one to exclude nonphysical solutions. In Section 7.4 we describe a high resolution method based on the second order splitting method (6.17) and present numerical results for this method.

7.1. A SIMPLIFIED DETONATION MODEL

7.1.1. Introduction

For studying numerical methods, the reactive Euler equations are often too complicated. Therefore, simpler qualitative models for (2.26) have been developed. Although physically not very realistic, these model problems are interesting for testing and analysing numerical methods. Clearly, simplified models are inadequate as a full test problem for any numerical method. However, a study of these problems suffices to analyse some of the difficulties that may arise in more complicated systems.

The model we study is a 2×2 system of equations [25]

$$\frac{\partial}{\partial t}u + \frac{\partial}{\partial x}(\tfrac{1}{2}u^2) = -Qw(u, Y), \tag{7.1a}$$

$$\frac{\partial}{\partial t}Y = w(u, Y). \tag{7.1b}$$

In the model above, Y may be interpreted as the mass fraction of the unburnt gas and $Q > 0$ may be interpreted as the heat release of the chemical reaction. If we consider the reactive Euler equations, then the reaction rate w depends on the temperature T via the Arrhenius' law (2.29). The exponential behaviour of the Arrhenius' law guarantees the reaction rate to be exponentially small even for temperatures close to the von Neumann temperature. Hence, the reaction rate is very large when the temperature is sufficiently high, but negligible for small T. As in (2.29), the reaction rate w in (7.1) should be an exponential function of u [60]. However, for simplicity, we approximate this by a model, in which the Arrhenius' behaviour is idealized to

$$w(u, Y) = \begin{cases} 0, & u < u_{ign}, \\ -Da\, Y, & u \geq u_{ign}, \end{cases} \tag{7.2}$$

where u_{ign} is the ignition value of the chemical reaction and, similar to (2.30), Da is called the Damköhler number. In practically all realistic cases the ignition temperature is much higher than the temperature of the unburnt gas. Hence, the ignition value u_{ign} in (7.2) should satisfy $u_{ign} \gg u_u$, where u_u is the value of u in the unburnt gas. Model problem (7.1) is referred to as the *simplified detonation model*. We emphasize that (7.1) should be viewed as a qualitative model which incorporates many features of the original system of reactive Euler equations. We are not aware of any systematic derivation of this model from the full system. A systematic derivation is possible when $\partial Y/\partial t$ is replaced by $\partial Y/\partial x$ [16]. However, we have experienced that for testing and analysing numerical methods the numerical behaviour of (7.1) shows more similarity to the reactive Euler equations.

Note that if u, f(u) and q(u) are defined by, respectively,

$$\mathrm{u} := (u\,,\,Y)^T, \qquad \mathrm{f(u)} := (\tfrac{1}{2}u^2\,,\,0)^T, \qquad \mathrm{q(u)} := (-Qw\,,\,w)^T,$$

the simplified detonation model (7.1) can be written in the general form (3.2a). For a detailed description of this model problem see [60]. For problem (7.1) we can develop a theory similar to the theory developed in Section 2.4 and 2.5. This theory turns out to be useful for testing and analysing several numerical methods. We start with the analogue of the Rankine-Hugoniot equations (2.35).

7.1.2. The Analogue of the Reactive Rankine-Hugoniot Equations

In this section we assume that a travelling wave is propagating with a constant velocity $s > 0$ in the positive x-direction. As in Section 2.4 we call this wave a detonation wave. Furthermore, we assume that the flow is steady with respect to a coordinate system moving with the detonation wave. All quantities ahead of the detonation wave will be identified by the subscript u (the unburnt gas), while the quantities behind the wave are denoted by the subscript b (the burnt gas). The variable ξ is defined as $\xi(x,t) := x - st$ (see (2.33)) and thus, (7.1) can be rewritten as a system of ordinary differential equations, i.e.

$$-s\frac{\mathrm{d}}{\mathrm{d}\xi}u + \frac{\mathrm{d}}{\mathrm{d}\xi}(\tfrac{1}{2}u^2) = -Qw(u,Y), \tag{7.3a}$$

$$-s\frac{\mathrm{d}}{\mathrm{d}\xi}Y = w(u,Y). \tag{7.3b}$$

We assume u and Y to satisfy

$$\lim_{\xi\to\infty}(u(\xi),Y(\xi)) = (u_u,1), \tag{7.4a}$$

$$\lim_{\xi\to-\infty}(u(\xi),Y(\xi)) = (u_b,0), \tag{7.4b}$$

where $u_u + Q < u_b$ and $0 \le u_u < u_{ign} \le u_b$. After integrating (7.3) from $\xi = -\infty$ to $\xi = +\infty$ and using (7.4) we deduce

$$s(u_b - u_u - Q) = \tfrac{1}{2}u_b^2 - \tfrac{1}{2}u_u^2. \tag{7.5}$$

Consistent with (2.35) the equation above is called the reactive Rankine-Hugoniot equation. We like to describe the set of possible values for u_b, such that for given u_u, Q and

s, equation (7.5) is satisfied. Therefore, the subscript b is suppressed. Let the curve $\mathcal{H}$ and the line $\mathcal{R}$ be defined by, respectively,

$$\mathcal{H}(u) := \tfrac{1}{2}u^2, \tag{7.6}$$

$$\mathcal{R}(u) := s(u - u_u - Q) + \tfrac{1}{2}u_u^2. \tag{7.7}$$

Note that the intersection point of $\mathcal{H}$ and $\mathcal{R}$ determines the final value for u_b.

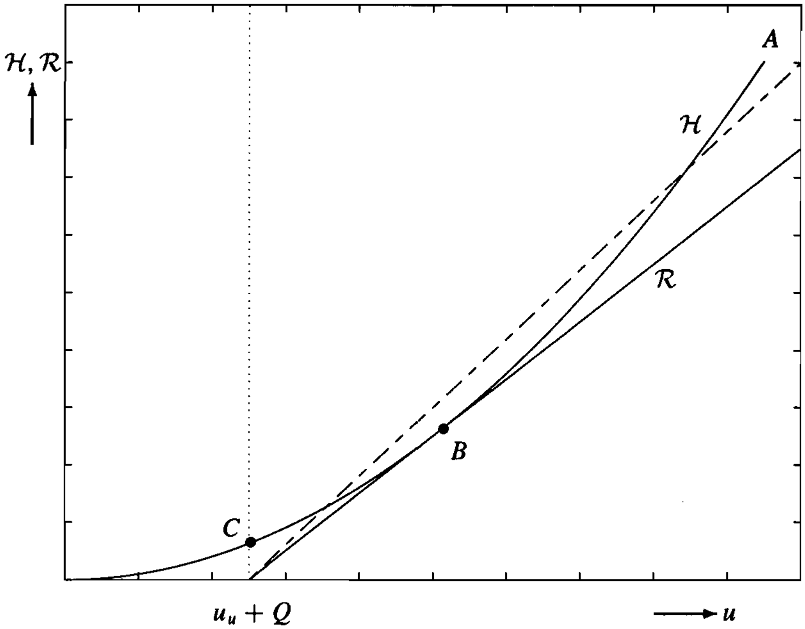

Figure 7.1. Different sections of the curve $\mathcal{H}$ (7.6) (the analogue of Figure 2.3).

It can be shown that there are at most two positive points of intersection between the line $\mathcal{R}$ and the curve $\mathcal{H}$, such that $u_b > u_u + Q$. There is a unique slope of the line $\mathcal{R}$ such that it is tangent to $\mathcal{H}$. This point of tangency (point B in Figure 7.1) separates $\mathcal{H}$ into two parts. Any straight line through $(u_u + Q, u_u^2/2)$ with a slope larger than that of the line through B intersects the curve $\mathcal{H}$ in two points (the dashed line in Figure 7.1). Depending on the final value of u_b we can distinguish between three different processes, which are called strong, Chapman-Jouguet and weak detonations.

(i) Detonation waves with the point $(u_b, u_b^2/2)$ on the curve AB are called strong detonations.

(ii) Detonation waves with the point $(u_b, u_b^2/2)$ equal to point B are called Chapman-Jouguet detonations (CJ detonations).

(iii) Detonation waves with the point $(u_b, u_b^2/2)$ on the curve BC are called weak detonations.

We replace u_b by u_{st}, u_{CJ} or u_{we} in case of a strong, Chapman-Jouguet or weak detonation, respectively. It follows from (7.6) and (7.7) that $s_{CJ} > 0$ is given by

$$s_{CJ} = u_u + Q + \sqrt{Q^2 + 2u_u Q}. \tag{7.8}$$

For all $s < s_{CJ}$ there will be no detonation. If $s = s_{CJ}$, then there will be a CJ detonation with $u_{CJ} = s_{CJ}$. If $s > s_{CJ}$, there will be a detonation with

$$u_{st} = s + \sqrt{(s - u_u)^2 - 2sQ}, \tag{7.9}$$

in case of a strong detonation and

$$u_{we} = s - \sqrt{(s - u_u)^2 - 2sQ}, \tag{7.10}$$

in case of a weak detonation. Finally, detonation waves can be characterized by the following properties.

For detonation waves $s > u_u$,

for strong detonations $s < u_{st} = u_b$,

for Chapman-Jouguet detonations $s = u_{CJ} = u_b$,

for weak detonations $s > u_{we} = u_b$.

The properties above are the analogues of Jouguet's rule (see Section 2.4.3). Again we consider Chapman-Jouguet and strong detonations as the only relevant detonations.

7.1.3. The Analogue of the ZND Model

Next we briefly develop the analogue of the ZND theory, as described in Section 2.5. Again we assume that a detonation wave travelling with constant speed s has the internal structure of an ordinary (nonreacting) precursor shock wave followed by a reaction zone (see Assumption A14). Hence the front of a detonation wave is a shock wave which initiates a chemical reaction behind it. Equation (7.5) should hold between any state ahead of the shock and any interior point of the reaction zone behind the shock. The curve $\mathcal{H}$ now depends on the variable Y, which varies continuously from 1 to 0, giving the generalization of (7.5)

$$\mathcal{H}(u, Y) := \tfrac{1}{2}u^2 - sQY. \tag{7.11}$$

The curves (7.11) for different values of Y are drawn in Figure 7.2.

First, due to a shock wave the variable u jumps to a higher value on the curve $\mathcal{H}(\cdot, 1)$, called the von Neumann spike (vN spike) as in reactive gas dynamics. It is straightforward to see that $u_{vN} = 2s - u_u$ and thus, $u_u < s < u_{vN}$. The von Neumann spike is the state immediately behind the shock wave and would be the final state if no reaction

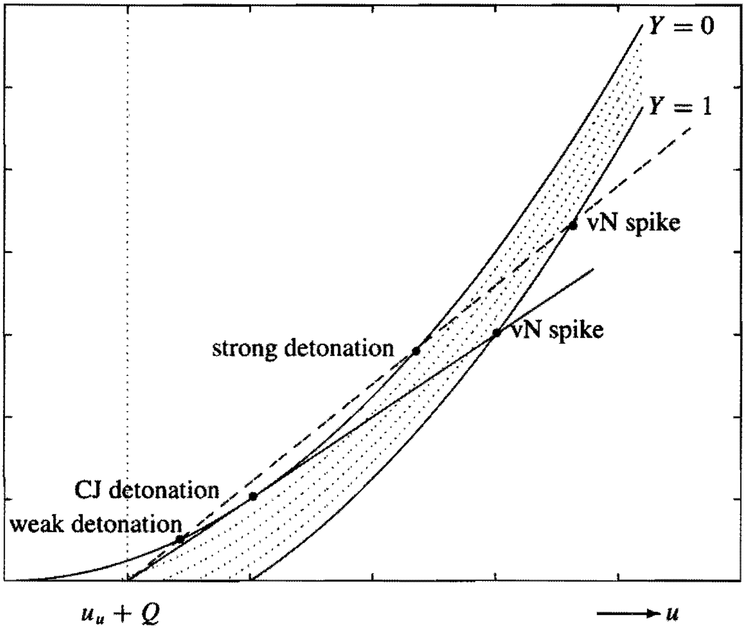

Figure 7.2. The curves $\mathcal{H}$ (7.11) corresponding to the ZND theory (the analogue of Figure 2.4).

would take place. It is clear that $u_{vN} > u_b \geq u_{ign}$. As the reaction proceeds the state point moves down the line $\mathcal{R}$ until the final state on the curve $\mathcal{H}(\cdot, 0)$ is reached. At each point on the line $\mathcal{R}$ between the von Neumann spike and the final state there is a unique Y determined from the Rankine-Hugoniot equations. It can easily be verified that for a CJ or strong detonation

$$u(Y) = s + \sqrt{(s-u_u)^2 - 2sQ(1-Y)}. \tag{7.12}$$

Suppose that at time $t = 0$ the precursor shock is located at $x = 0$. Hence, at time t the variable $\xi = x - st$ measures the distance between the point x and the precursor shock. Therefore, $u(\xi) = u_u$ for all $\xi > 0$. The dependence of Y on the distance ξ is characterized by the following ordinary differential equation (see (7.2) and (7.3b))

$$\frac{\mathrm{d}}{\mathrm{d}\xi}Y(\xi) = -\frac{Da\,Y(\xi)}{s}, \quad \forall\,\xi < 0, \tag{7.13a}$$

$$Y(0) = 1, \tag{7.13b}$$

where $\xi = 0$ corresponds to the position of the precursor shock. We can solve (7.13) exactly and obtain the exact ZND solution of (7.1) (see (7.12))

$$Y(x,t) = Y(\xi) = \begin{cases} 1, & \xi > 0, \\ \exp(Da\,\xi/s), & \xi \leq 0, \end{cases} \tag{7.14a}$$

$$u(x,t) = u(\xi) = \begin{cases} u_u, & \xi > 0, \\ s + \sqrt{(s-u_u)^2 - 2sQ(1-Y(\xi))}, & \xi \leq 0. \end{cases} \tag{7.14b}$$

Note that for the ZND model the final state is a strong or CJ detonation. Therefore, we are interested in techniques for describing strong or CJ detonations. The minimum speed for a detonation is the speed s_{CJ} of a CJ detonation. We define the degree of overdrive f completely similar to (2.50), i.e. $f := (s/s_{CJ})^2$, from which it follows directly that $f \geq 1$.

Finally, the half-reaction length $L_{1/2}$ is introduced as the distance for half completion of the reaction starting from the front of the detonation wave. It is easy to see that (7.14a) implies that $L_{1/2}$ is given by

$$L_{1/2} = \frac{s}{Da} \ln(2). \tag{7.15}$$

Often $L_{1/2}$ is given and (7.15) is used to compute the corresponding Damköhler number Da.

Example 7.1. As an example of the preceding theory we describe the ZND solution of the CJ detonation with

$$u_u = 0, \qquad Q = 0.5, \qquad f = 1, \qquad Da = 0.6931.$$

The half-reaction length is given by $L_{1/2} = 1$. It follows from (7.8) that the CJ detonation is propagating with a speed $s = s_{CJ} = 1$. The final state for the CJ detonation is given by $u_b = u_{CJ} = s_{CJ} = 1$. In Figure 7.3 the steady ZND solution is drawn. The variable u reaches its maximum value right behind the precursor shock. As mentioned before this value is called the von Neumann spike, which in this particular case satisfies $u_{vN} = u(1) = 2$ (see (7.12)). ☐

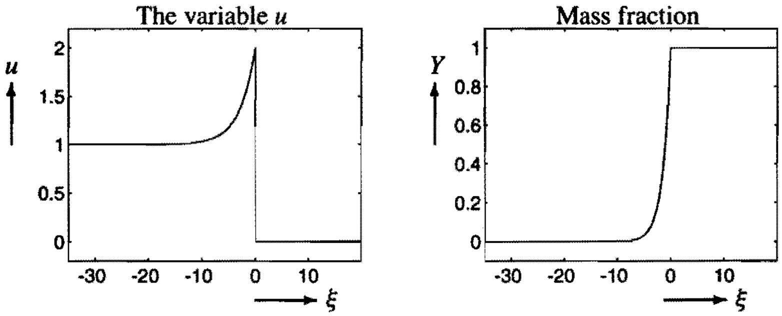

Figure 7.3. ZND solution of (7.1) with $Q = 0.5$, $f = 1$ and $Da = 0.6931$ ($L_{1/2} = 1$).

Example 7.2. As a second example we describe the ZND solution of a strong detonation with

$$u_u = 0, \qquad Q = 2, \qquad f = 1.265625, \qquad Da = 3.1192.$$

The half-reaction length is given by $L_{1/2} = 0.1$ and the final state for this strong detonation is given by $u_b = 6$ (see (7.9)). The exact ZND solution is propagating with $s = 4.5$ and $u_{vN} = u(1) = 9$ (see (7.12)). In Figure 7.4 the steady ZND solution (7.14) is drawn. ☐

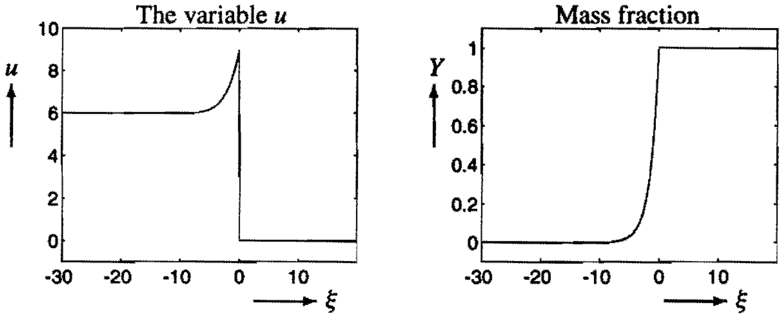

Figure 7.4. ZND solution of (7.1) with $Q = 2$, $f = 1.265625$ and $Da = 3.1192$ ($L_{1/2} = 1$).

7.2. NUMERICAL COMPUTATION OF STRONG OR CJ DETONATION WAVES

In this section we want to compute the ZND solution of strong or CJ detonation waves propagating with a constant wave speed $s > 0$. In practice the time scale of the chemical reaction is very small compared with the time scale of the fluid dynamics (i.e. Da is large) and in many applications the detonation wave may be considered as a discontinuity. Moreover, we are rather interested in the global behaviour of the detonation wave (i.e. a correct capturing of the detonation wave) and not so much in the inner structure of the wave (like the von Neumann spike). Therefore, we consider numerical methods for spatial meshes and time steps appropriate to resolve the fluid dynamics but not for the rapid chemical reaction. If the propagation speed is correct, even on relatively coarse meshes, adaptive refinement of the spatial mesh may be used to describe any details of the detonation.

To obtain an initial value problem we add initial data $\mathbf{u}^0 = (u^0, Y^0)^T$ to differential equation (7.1). For the sake of simplicity, we assume that the initial data correspond to the exact ZND solution, i.e. (see (7.14))

$$Y^0(x) = \begin{cases} 1, & x > 0, \\ \exp(Da\, x/s), & x \le 0, \end{cases} \tag{7.16a}$$

$$u^0(x) = \begin{cases} u_u, & x > 0, \\ s + \sqrt{(s-u_u)^2 - 2sQ(1-Y^0(x))}, & x \le 0. \end{cases} \tag{7.16b}$$

Hence, the exact solution of (7.1),(7.16) is the strong or CJ detonation wave (7.14) propagating with a constant wave speed $s > 0$.

Let a time step Δt and a mesh width Δx be given. As in Section 4.1, we define discrete time levels $t^n := n\Delta t$ for $n = 0, 1, 2, \ldots$ and discrete mesh points $x_i := i\Delta x$ for $i = \ldots, -2, -1, 0, 1, 2, \ldots$. The simplified detonation model (7.1) is solved numerically using a method based on the first order splitting method (6.14). In this method the numerical solution at each time level is derived by a two-step procedure. In the first step we assume that no reaction occurs (i.e. $w = 0$ in (7.1)) and approximate the solution of the remaining homogeneous equation, i.e. Burgers' equation. In the second step we assume no convection (i.e. $\partial u^2/\partial x = 0$ in (7.1)) and solve the corresponding ordinary differential equations numerically. For a detailed description of splitting methods we refer to Chapter 6.

Let the numerical solution at time level t^n be given. In the first step we have to approximate the solution of Burgers' equation (the mass fraction Y remains constant during the first step) at time level t^{n+1}. We use Roe's conservative three-point method, described in Example 4.17. For later purposes it will be useful to denote the result by C_i^{n+1}, so

$$C_i^{n+1} \;:=\; U_i^n \,-\, \tau\{F_{i+1/2}^n - F_{i-1/2}^n\}, \tag{7.17}$$

where the numerical flux function F is given by (4.50). For fast reactions the stability of an explicit method for the ordinary differential equation would imply a much more severe time step restriction than the usual CFL condition $\tau \max_i |U_i^n| \leq 1$. Therefore, we solve the ODE in the second step by the backward Euler method. The complete finite difference scheme then reads

$$U_i^{n+1} \;=\; U_i^n \,-\, \tau\{F_{i+1/2}^n - F_{i-1/2}^n\} \,-\, \Delta t Q w(U_i^{n+1}, Y_i^{n+1}), \tag{7.18a}$$

$$Y_i^{n+1} \;=\; Y_i^n \,+\, \Delta t w(U_i^{n+1}, Y_i^{n+1}). \tag{7.18b}$$

Hence, using the notation of Section 6.3, $\mathcal{L}_{\Delta t}^f$ is given by Roe's method and $\mathcal{L}_{\Delta t}^c$ is given by the backward Euler method. As usual, the time step Δt is restricted by the CFL stability condition.

Remember that due to a shock wave propagating into the unburnt gas, u increases above the ignition value u_{ign} and a reaction is started. Therefore, it seems reasonable to assume that no chemical reaction occurs in the cell $[x_{i-1/2}, x_{i+1/2})$ during the $(n+1)$st time step, if $C_i^{n+1} < u_{ign}$. After substituting (7.18a) into (7.18b), using (7.2) and (7.17), we can rewrite (7.18) as

$$U_i^{n+1} \;=\; C_i^{n+1} + \frac{\Delta t Da}{1 + \Delta t Da} Q H(C_i^{n+1} - u_{ign}) Y_i^n, \tag{7.19a}$$

$$Y_i^{n+1} \;=\; \frac{1}{1 + \Delta t Da H(C_i^{n+1} - u_{ign})} Y_i^n, \tag{7.19b}$$

where H is the Heavyside function defined by $H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ if $x < 0$.

Next we present numerical results for method (7.19). In the first example it is shown that for small mesh widths Δx, the numerical solution approximates the exact solution very well. However, if Δx is increased to more practical values, then the numerical solution of method (7.19) becomes totally wrong.

Example 7.3. In this example we consider the strong detonation of Example 7.2. We increase Da to 31.192, so the half reaction length is given by $L_{1/2} = 0.1$. The exact ZND solution of (7.1),(7.16) is given by (7.14) with $s = 4.5$.

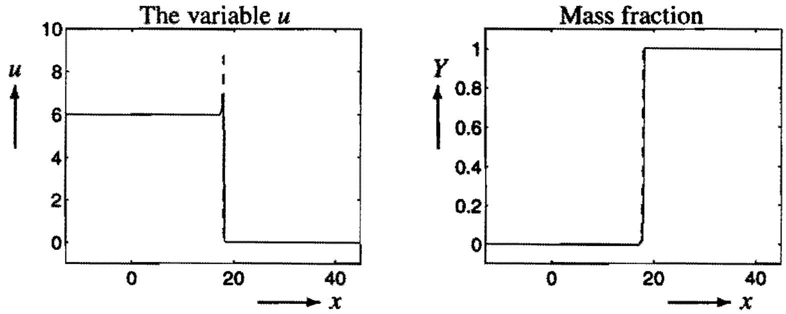

Figure 7.5. Numerical results for method (7.19) (with $\Delta t = 0.01$ and $\Delta x = 0.1$); numerical solution (solid line) and exact solution (dashed line) of (7.1),(7.16) at $t = 4$ of a strong detonation with $Q = 2$, $f = 1.265625$, $Da = 31.192\,(L_{1/2} = \Delta x)$ and $u_{ign} = 0.1$.

In Figure 7.5 the numerical results are compared with the exact solution. Due to numerical diffusion the peak in u is almost disappeared. However, the numerical ZND profile is essentially correct. However, the mesh width is relatively small ($L_{1/2} = \Delta x$) and in most practical cases we cannot afford such fine meshes. Therefore, we increase Δx and Δt and keep $\tau = \Delta t/\Delta x$ fixed.

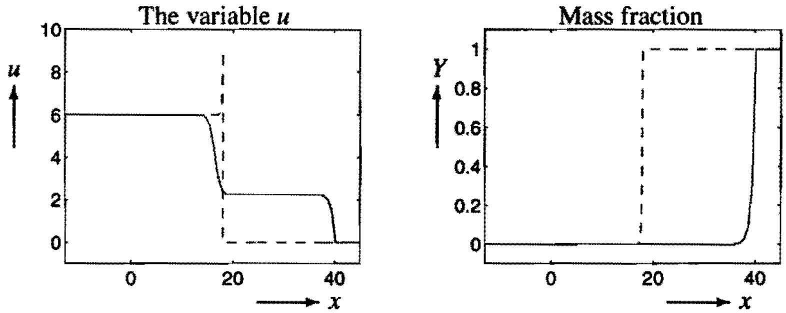

Figure 7.6. Numerical results for method (7.19) (with $\Delta t = 0.08$ and $\Delta x = 0.8$); numerical solution (solid line) and exact solution (dashed line) of (7.1),(7.16) at $t = 4$ of a strong detonation with $Q = 2$, $f = 1.265625$, $Da = 31.192\,(L_{1/2} = 0.125\Delta x)$ and $u_{ign} = 0.1$.

Figure 7.6 clearly illustrates that the numerical solution is completely wrong. The (incorrect) numerical solution appears to be a weak detonation wave propagating with speed 10 (one cell per time step), followed by an ordinary shock wave, while the physically correct solution is a strong detonation wave propagating with speed $s = 4.5$. In the weak detonation wave all heat is released and the gas is completely burnt.

In the previous figures we choose a low and nonphysical ignition value $u_{ign} \approx u_u$. As remarked before in practice $u_{ign} \gg u_u$ and we increase u_{ign} to 1.5. In Figure 7.7 the solution is drawn. The numerical detonation wave solution is the correct strong detona-

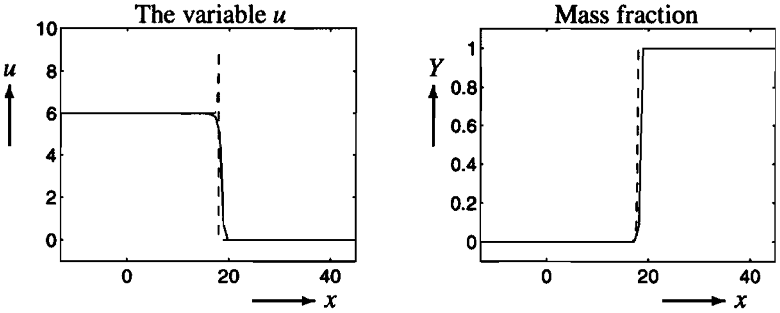

Figure 7.7. Numerical results for method (7.19) (with $\Delta t = 0.08$ and $\Delta x = 0.8$); numerical solution (solid line) and exact solution (dashed line) of (7.1),(7.16) at $t = 4$ of a strong detonation with $Q = 2$, $f = 1.265625$, $Da = 31.192$ ($L_{1/2} = 0.125\Delta x$) and $u_{ign} = 1.5$.

tion wave. The peak in the variable u has completely disappeared. This is caused by the combination of a large mesh width and a thin reaction zone ($\Delta x = 8L_{1/2}$). The reaction is so fast (relative to the mesh width) that even in the initial data $\mathrm{U}_i^0 = \bar{\mathrm{u}}_i^0$ no peak can be seen. Comparing the results in Figure 7.7 with the results in Figure 7.6, the difference is clear. ☐

Numerical detonation waves may propagate at nonphysical wave speeds (see Figure 7.6). In the next section we will show that these "wrong solutions" are approximations of nonphysical weak solutions of (7.1) (i.e. weak detonation waves).

The basic explanation for the occurrence of nonphysical wave speeds is that the numerical propagation of the precursor shock wave always results in a smeared representation of this shock wave including intermediate values $u_u < u < u_b$ in front of it (see Figure 7.8). If the ignition value u_{ign} is close to u_u, then, due to numerical diffusion, u is raised above the ignition value and an artificial reaction is started ahead of the shock wave (the light grey area in Figure 7.8). If Da is large enough, then the gas is completely burnt in the next time step Δt and the discontinuity is shifted to a cell boundary. Therefore, it is not surprising that nonphysical wave speeds of one cell per time step may be observed for large Da [10, 16, 58].

We emphasize that the main goal of this thesis is to develop a numerical method that automatically captures the detonation wave. Clearly, in Figure 7.6 the detonation wave is captured completely wrong. Therefore, we like to study the occurrence of nonphysical solutions and especially their wrong wave speeds, in more detail. To this end, we define two quantities S_1^n and S_2^n at time t^n by

$$S_1^n n\Delta t(u_b - u_u) \;:=\; \Delta x \sum_{i=-\infty}^{\infty} \left(U_i^n - U_i^0\right), \tag{7.20a}$$

$$-S_2^n n\Delta t \;:=\; \Delta x \sum_{i=-\infty}^{\infty} \left(Y_i^n - Y_i^0\right). \tag{7.20b}$$

Since u^0 is constant outside some finite interval so is U_i^n (see (7.19)). Hence, the right-

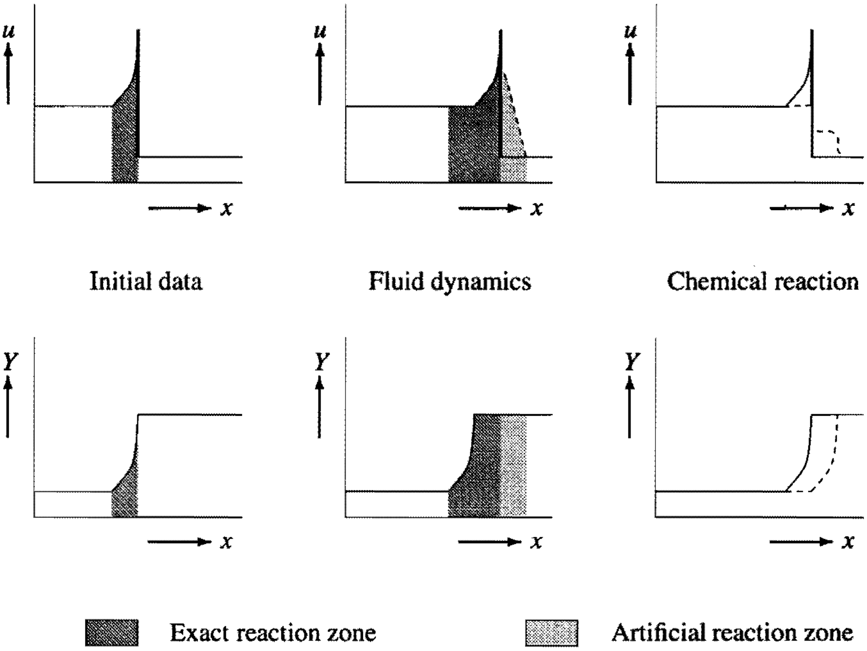

Figure 7.8. Global explanation of the occurrence of nonphysical solutions for the simplified detonation model (7.1); exact solution (solid line) and numerical solution (dashed line).

hand side of (7.20) is finite and S_1^n and S_2^n are well defined. The quantity S_2^n may be interpreted as the average speed of the numerical detonation wave. Some other authors define the numerical wave speed for a finite difference method by an expression of the form $m\Delta x/(l\Delta t)$, where l and m are relatively prime numbers [31]. In other words, $\mathbf{U}_i^n = \mathbf{U}_{i-m}^{n-l}$ for all i and for all $n \geq l$. However, in general m and l are hard to compute from numerical results. On the other hand, S_1^n and S_2^n can be computed easily by (7.20).

Our numerical experiments show that the sequence S_2^n always converges for $n \to \infty$. Therefore, we assume that there exists a positive constant S_2 such that $\lim_{n\to\infty} S_2^n = S_2$. To our knowledge, the influence of the ignition value on the numerical wave speed has not been studied so far. Considering the previous example, we expect that for a fixed Damköhler number, the speed S_2 depends on the ignition value u_{ign} and the mesh width Δx. This is illustrated in the following example.

Example 7.4. In this example we consider the strong detonation of Example 7.2. We recall that the exact detonation wave is propagating with speed $s = 4.5$. In Table 7.9 the limit value $S_2 = \lim_{n\to\infty} S_2^n$ is presented for several values of u_{ign} and Δx.

As expected, for low ignition values the average numerical wave speed S_2 is only a good approximation of the exact wave speed for small mesh widths ($\Delta x \approx L_{1/2}$). If we increase Δx and keep $L_{1/2}$ (or Da) fixed, we observe completely wrong wave speeds of

	Mesh width Δx				
u_{ign}	$L_{1/2}$	$10 \cdot L_{1/2}$	$10^2 \cdot L_{1/2}$	$10^3 \cdot L_{1/2}$	$10^4 \cdot L_{1/2}$
0.1	4.5000	10.000	10.000	10.000	10.000
0.2	4.5000	5.7143	10.000	10.000	10.000
0.3	4.5000	5.0000	6.6666	6.6666	6.6666
0.4	4.5000	5.0000	5.0000	5.0000	5.0000
0.5	4.5000	5.0000	5.0000	5.0000	5.0000
0.6	4.5000	4.5000	5.0000	5.0000	5.0000
0.7	4.5000	4.5000	5.0000	5.0000	5.0000
0.8	4.5000	4.5000	4.5455	4.6154	4.6154
0.9	4.5000	4.5000	4.5000	4.5000	4.5000
1.0	4.5000	4.5000	4.5000	4.5000	4.5000

Table 7.9. Average numerical wave speed $\lim_{n\to\infty} S_2^n$ for method (7.19) with $Q = 2$, $f = 1.265625$, $Da = 31.192$ ($L_{1/2} = 0.1$), $\tau = \Delta t/\Delta x = 0.1$ and initial data (7.16).

one cell per time step. However, for more realistic values of u_{ign}, the numerical wave speed is correct, independent of the mesh width Δx being used. The last two columns in Table 7.9 are the same, since in both cases the gas is completely burnt during a time step Δt. Hence, Table 7.9 shows the remarkable result that the wrong wave speeds are caused by using a nonphysical ignition value. Although the wrong wave speeds have been studied in several papers (e.g. [16, 58]), this result has not been noticed by other authors yet. In the next section we present theoretical insights assessing why a correct ignition value will exclude nonphysical weak detonation waves, even for relatively coarse meshes. □

Finally, we like to study the global error. We showed in Section 6.3 that method (7.19) is consistent of order 1 in the L_1-norm. In Chapter 6 we have defined the global error $\mathbf{E}_{\Delta t}$ (or $\mathbf{E}_{\Delta t}^{tot}$) for arbitrary x and t as (see (6.25))

$$\mathbf{E}_{\Delta t}(x,t) := \mathbf{U}_{\Delta t}(x,t) - \mathbf{u}(x,t),$$

where $\mathbf{u} = (u, Y)^T$ is the exact solution of (7.1),(7.16) and $\mathbf{U}_{\Delta t}$ is given by (4.4). Since the 1-norm of this error is hard to compute in general, we consider the pointwise global error

$$\mathbf{E}_i^n := \mathbf{U}_i^n - \mathbf{u}(x_i, t^n).$$

The corresponding norm is a discrete 1-norm which can be applied to the pointwise error $\mathbf{E}^n$,

$$\|\mathbf{E}^n\|_1 = \Delta x \sum_{i=-\infty}^{\infty} |\mathbf{U}_i^n - \mathbf{u}(x_i, t^n)|. \tag{7.21}$$

Next we check for a particular detonation, whether the order of method (7.19) equals one.

Example 7.5. In this example we consider the strong detonation of Example 7.2. Only the first component of the global error $\mathbf{E}^n = (E_1^n, E_2^n)^T$ is considered, i.e.

$$\|E_1^n\|_1 = \Delta x \sum_{i=-\infty}^{\infty} |U_i^n - u(x_i, t^n)|. \tag{7.22}$$

The results in Table 7.10 clearly illustrate the first order behaviour of method (7.19). □

l	Δt_l	global error $\|E_1^{n_l}\|_1$
0	1/25	$1.7763 \cdot 10^{+0}$
1	1/50	$1.0912 \cdot 10^{+0}$
2	1/100	$3.6659 \cdot 10^{-1}$
3	1/200	$1.8273 \cdot 10^{-1}$
4	1/400	$9.1220 \cdot 10^{-2}$
5	1/800	$4.5576 \cdot 10^{-2}$

Table 7.10. Global error for method (7.19) at $t = 1$ with $Q = 2$, $f = 1.265625$, $Da = 3.1192\,(L_{1/2} = 1)$, $u_{ign} = 1$, $\Delta t_l = (1/25)\cdot(1/2)^l$, $\Delta x_l = (10/25)\cdot(1/2)^l$, $E_1^{n_l}$ defined by (7.22) and initial data (7.16).

7.3. DETONATION CAPTURING FOR STIFF COMBUSTION CHEMISTRY

7.3.1. The Avoidance of Nonphysical Detonation Waves

In this section we explain the strong influence of the ignition value on the numerical solution of the simplified detonation model (7.1). In Example 7.3 the incorrect weak solution appears to be a weak detonation wave followed by an ordinary shock wave, while the physically correct solution is a strong or CJ detonation wave propagating with speed $s > 0$.

For given states $\mathbf{u}_u$ at large distances corresponding to a strong or CJ detonation propagating with speed s, there is a whole family of nonphysical intermediate states. In this section we show that these intermediate states are parametrized by the speed S_2 of the leading weak detonation [16]. The value of u between the leading weak detonation and the subsequent shock decreases with S_2. The intermediate state with the highest possible u-value of all nonphysical solutions in the family is attained when $S_2 = s$. The value of u is then still smaller than the value of u in the burnt gas behind the correct strong detonation. Hence, fake numerical detonation waves will be excluded, if the value of u in the burnt gas directly behind the detonation wave (the intermediate value) is larger than the value of u behind a weak detonation wave propagating with speed s. We use this to obtain the desired criterion, which ensures that the numerical wave speed equals the physically correct wave speed. In the next chapter this criterion is extended to the reactive Euler equations.

The extra criterion simply states that u_{ign} should exceed a certain threshold value determined by the value of u behind a weak detonation wave propagating with speed s and by the heat release Q of the chemical reaction. Our numerical computations convincingly illustrate that the correct detonation speed is obtained, even if the Damköhler number is very large (i.e. $L_{1/2} \approx 10^{-5}\Delta x$). Certainly, in this situation any information about the detailed structure of the detonation wave is lost. However, the major flaw of the generation of nonphysical weak detonations is overcome and the detonation wave is captured correctly.

Since we like to study detonation waves in which the thickness of the reaction zone occupies a tiny fraction of Δx (Da is very large), we may restrict ourselves to piecewise constant initial data, i.e. (7.16) with $Da \to \infty$,

$$(u^0(x), Y^0(x))^T = \begin{cases} (u_b, 0)^T, & x \le 0, \\ (u_u, 1)^T, & x > 0. \end{cases} \tag{7.23}$$

The following lemma gives a relation between S_1^n and S_2^n. This relation may be interpreted as the numerical analogue of the reactive Rankine-Hugoniot equation (7.5).

Lemma 7.6. *Assume the finite difference method* (7.18) *is used to approximate the simplified detonation model* (7.1) *with initial data* (7.23). *Let* $\mathrm{U}_i^n = (U_i^n, Y_i^n)^T$ *be a solution of* (7.18) *with given initial values* $\mathrm{U}_i^0 = \bar{\mathrm{u}}_i^0$, *as defined in* (4.3). *Then* S_1^n *and* S_2^n *satisfy the relation*

$$S_1^n(u_b - u_u) - S_2^n Q = \tfrac{1}{2}u_b^2 - \tfrac{1}{2}u_u^2. \tag{7.24}$$

Proof. It follows from (7.23) that

$$\lim_{i\to-\infty} F_{i-1/2}^n = \tfrac{1}{2}u_b^2 \quad \text{and} \quad \lim_{i\to\infty} F_{i+1/2}^n = \tfrac{1}{2}u_u^2,$$

for all $n \ge 0$. First we replace $n+1$ by j in (7.18). After multiplying the resulting scheme by Δx, summing over i and using the limits above we obtain

$$\begin{aligned} \Delta x \sum_{i=-\infty}^{\infty} \left(U_i^j - U_i^{j-1}\right) &= \Delta t \left\{\tfrac{1}{2}u_b^2 - \tfrac{1}{2}u_u^2\right\} - \Delta t \Delta x\, Q \sum_{i=-\infty}^{\infty} w(U_i^j, Y_i^j), \\ \Delta x \sum_{i=-\infty}^{\infty} \left(Y_i^j - Y_i^{j-1}\right) &= \Delta t \Delta x \sum_{i=-\infty}^{\infty} w(U_i^j, Y_i^j). \end{aligned}$$

Summing the equations above over all j with $1 \le j \le n$, gives

$$\begin{aligned} \Delta x \sum_{i=-\infty}^{\infty} (U_i^n - U_i^0) &= n\Delta t \left\{\tfrac{1}{2}u_b^2 - \tfrac{1}{2}u_u^2\right\} - \Delta t \Delta x\, Q \sum_{j=1}^{n} \sum_{i=-\infty}^{\infty} w(U_i^j, Y_i^j), \\ \Delta x \sum_{i=-\infty}^{\infty} (Y_i^n - Y_i^0) &= \Delta t \Delta x \sum_{j=1}^{n} \sum_{i=-\infty}^{\infty} w(U_i^j, Y_i^j). \end{aligned}$$

After substituting the second equation into the first we obtain (see (7.20))

$$S_1^n n\Delta t(u_b - u_u) = n\Delta t\left\{\tfrac{1}{2}u_b^2 - \tfrac{1}{2}u_u^2\right\} + S_2^n n\Delta t Q.$$

Finally we divide the latter equation by $n\Delta t$ and the result (7.24) follows immediately. This completes the proof. □

As shown in Section 7.2, for large *Da* it is possible to obtain numerical solutions that approximate a nonphysical weak solution of (7.1), i.e. a weak detonation wave followed by an ordinary shock wave (see Figure 7.6). In order to approximate the correct weak solution, the final state of the burnt gas directly behind the numerical detonation wave should be equal to u_b for $n \to \infty$. Let $v_b^n > 0$ be given such that

$$S_2^n(v_b^n - u_u - Q) = \tfrac{1}{2}(v_b^n)^2 - \tfrac{1}{2}u_u^2, \tag{7.25}$$

where we assume that $S_2^n \geq s_{CJ}$, since otherwise v_b^n will not exist (see Section 7.1.2). Furthermore, it follows from $v_b^n > 0$, $S_2^n > 0$ and (7.25) that $v_b^n > u_u + Q$. The constant v_b^n is the final state of a detonation wave propagating with speed S_2^n. If $S_2^n > s_{CJ}$, there are two possible values for v_b^n satisfying (7.25), namely: $v_b^n > S_2^n$ for a strong detonation and $v_b^n < S_2^n$ for a weak detonation. The wrong weak solutions mentioned earlier, consist of a detonation wave linking the state $(u_u, 1)^T$ to $(v_b^n, 0)^T$, followed by an ordinary shock wave linking the state $(v_b^n, 0)^T$ to $(u_b, 0)^T$. Let $\bar{S}^n$ denote the speed of this shock wave, i.e.

$$\bar{S}^n := \tfrac{1}{2}(v_b^n + u_b). \tag{7.26}$$

The existence of numerical solutions approximating this nonphysical weak solution is illustrated by the following theorem.

Theorem 7.7. *Let all assumptions of Lemma 7.6 be fulfilled and let* $\mathbf{U}_{\Delta t} = (U_{\Delta t}, Y_{\Delta t})^T$ *be given by* (4.4). *Furthermore, let* $n > 0$ *and assume that* $S_2^n \geq s_{CJ}$ *(i.e. there exists a* $v_b^n > 0$ *such that* (7.25) *holds). Then the function* $\tilde{\mathbf{u}}^n(\cdot, t^n)$ *defined by*

$$\tilde{\mathbf{u}}^n(x, t^n) := \begin{cases} (u_b, 0)^T, & x < \bar{S}^n t^n, \\ (v_b^n, 0)^T, & \bar{S}^n t^n < x < S_2^n t^n, \\ (u_u, 1)^T, & x > S_2^n t^n \end{cases} \tag{7.27}$$

satisfies

$$\int_{-\infty}^{\infty} (\mathbf{U}_{\Delta t}(x, t^n) - \tilde{\mathbf{u}}^n(x, t^n))\mathrm{d}x = 0. \tag{7.28}$$

Furthermore, for given S_2^n *and* v_b^n *satisfying* (7.25), $\tilde{\mathbf{u}}^n(\cdot, t^n)$ *is the only piecewise constant function consisting of at most three constant states* $(c_1, 0)^T$, $(c_2, 0)^T$ *and* $(c_3, 1)^T$, *such that* (7.28) *holds.*

For homogeneous conservation laws we have discrete conservation with respect to the exact solution (see (4.22)). However, Theorem 7.7 shows that for problem (7.1) we have discrete conservation with respect to the function $\tilde{\mathbf{u}}^n$ given in (7.27). Unfortunately, $\tilde{\mathbf{u}}^n$ is not equal to the exact solution in general. If the numerical solution consists of three constant states $(c_1, 0)^T$, $(c_2, 0)^T$ and $(c_3, 1)^T$, then Theorem 7.7 implies that any shock we compute at time t^n must, in a sense, have the same location as the shocks in $\tilde{\mathbf{u}}^n$ (see Section 4.2). Moreover, if $v_b^n < u_b$, then (7.27) consists of a detonation wave followed by an ordinary shock wave and $\tilde{\mathbf{u}}^n$ is the nonphysical weak solution observed in our numerical experiments. Furthermore, $v_b^n = u_b$ implies $S_2^n = s$ and (7.27) is the physically correct weak solution.

Proof. Since $\mathbf{u}^0$ is piecewise constant, $\mathbf{U}_{\Delta t}$ is constant outside some finite interval. Furthermore, since (7.18) is a three-point method, initial data (7.23) are used, $\bar{S}^n > 0$ and $S_2^n < \Delta x/\Delta t$,

$$\int_{-\infty}^{\infty} (\mathbf{U}_{\Delta t}(x, t^n) - \tilde{\mathbf{u}}^n(x, t^n))\,\mathrm{d}x = \int_{-(n+1/2)\Delta x}^{(n+1/2)\Delta x} (\mathbf{U}_{\Delta t}(x, t^n) - \tilde{\mathbf{u}}^n(x, t^n))\,\mathrm{d}x. \tag{7.29}$$

Note that (7.28) consists of two equations, one equation for U and one equation for Y. First we prove the second equality in (7.28). After replacing the summation in (7.20b) by an integral, using (4.4), and subsequently applying (7.29), we arrive at

$$\int_{-(n+1/2)\Delta x}^{(n+1/2)\Delta x} Y_{\Delta t}(x, t^n)\,\mathrm{d}x = (n + \tfrac{1}{2})\Delta x - S_2^n t^n. \tag{7.30}$$

The second equation in (7.28) directly follows from (7.27), (7.29) and (7.30). We now replace the summation in (7.20a) by an integral and obtain, using (7.23) and (7.29),

$$\int_{-(n+1/2)\Delta x}^{(n+1/2)\Delta x} U_{\Delta t}(x, t^n)\,\mathrm{d}x = (n + \tfrac{1}{2})\Delta x u_b + S_1^n t^n (u_b - u_u) + (n + \tfrac{1}{2})\Delta x u_u. \tag{7.31}$$

Using (7.5), (7.24), (7.25) and (7.26), we deduce

$$\begin{aligned} S_1^n t^n (u_b - u_u) &= s t^n (u_b - u_u - Q) + S_2^n t^n Q \\ &= t^n (\tfrac{1}{2}u_b^2 - \tfrac{1}{2}(v_b^n)^2) + t^n (\tfrac{1}{2}(v_b^n)^2 - \tfrac{1}{2}u_u^2) + S_2^n t^n Q \\ &= \bar{S}^n t^n (u_b - v_b^n) + S_2^n t^n (v_b^n - u_u). \end{aligned}$$

Using this together with (7.27) and (7.31) gives

$$\begin{aligned} \int_{-(n+1/2)\Delta x}^{(n+1/2)\Delta x} U_{\Delta t}(x, t^n)\,\mathrm{d}x &= (n + \tfrac{1}{2})\Delta x u_b + \bar{S}^n t^n u_b + (S_2^n - \bar{S}^n) t^n v_b^n \\ &\quad + ((n + \tfrac{1}{2})\Delta x - S_2^n t^n) u_u \\ &= \int_{-(n+1/2)\Delta x}^{(n+1/2)\Delta x} \tilde{u}^n(x, t^n)\,\mathrm{d}x. \end{aligned}$$

This completes the proof of (7.28). It remains to prove that $\tilde{\mathbf{u}}^n(\cdot, t^n)$ is the only piecewise constant function consisting of at most three constant states, satisfying (7.28). Two constant states are given by $(u_u, 1)^T$ and $(u_b, 0)^T$ (see (7.23)). Denote the third constant state by $(c, 0)^T$ and let the function $\tilde{\mathbf{w}}^n(\cdot, t^n) = (\tilde{w}_1^n(\cdot, t^n), \tilde{w}_2^n(\cdot, t^n))^T$ be defined by

$$\tilde{\mathbf{w}}^n(x, t^n) = (\tilde{w}_1^n(x, t^n), \tilde{w}_2^n(x, t^n))^T := \begin{cases} (u_b, 0)^T, & x < a, \\ (c, 0)^T, & a < x < b, \\ (u_u, 1)^T, & x > b, \end{cases}$$

where $0 < a \leq b$. We have to show that $a = \bar{S}^n t^n$, $b = S_2^n t^n$ and $c = v_b^n$ (see (7.27)). It follows from (7.30) and the definition of $\tilde{\mathbf{w}}^n$ that $b = S_2^n t^n$. Since $\tilde{\mathbf{w}}^n$ satisfies (7.28) and (7.29) holds, it follows from (7.31) and $a = \frac{1}{2}(u_b + c)t^n$ that

$$\begin{aligned} 0 &= \int_{-(n+1/2)\Delta x}^{(n+1/2)\Delta x} (U_{\Delta t}(x, t^n) - \tilde{w}_1^n(x, t^n))\,\mathrm{d}x \\ &= S_1^n t^n (u_b - u_u) + a(c - u_b) + S_2^n t^n (u_u - c) \\ &= S_2^n t^n Q + t^n(\tfrac{1}{2}u_b^2 - \tfrac{1}{2}u_u^2) + t^n(\tfrac{1}{2}c^2 - \tfrac{1}{2}u_b^2) + S_2^n t^n (u_u - c) \\ &= S_2^n t^n (-c + u_u + Q) + t^n(\tfrac{1}{2}c^2 - \tfrac{1}{2}u_u^2). \end{aligned}$$

Using this and (7.25) we obtain $c = v_b^n$ and, subsequently $a = \bar{S}^n t^n$. This completes the proof. □

The main goal of this section is to derive a criterion which excludes nonphysical weak solutions. So, numerical approximations of (7.27) should be excluded, except for the case $v_b^n = u_b$ and $S_2^n = s$. Recall that numerical experiments show that the sequence S_2^n always converges for $n \to \infty$. Therefore, we assume that there exists a positive constant S_2 such that $\lim_{n\to\infty} S_2^n = S_2$. It follows from (7.25) that $\lim_{n\to\infty} v_b^n = v_b$ for some v_b and hence, S_2 and v_b are related by

$$S_2(v_b - u_u - Q) = \tfrac{1}{2}v_b^2 - \tfrac{1}{2}u_u^2. \tag{7.32}$$

An important question is whether the piecewise constant solution (7.27) is stable as time evolves. Here by stability is meant that $\tilde{\mathbf{u}}^n$ remains a piecewise constant function (for $n \to \infty$) consisting of the three constant states $(u_u, 1)^T$, $(v_b^n, 0)^T$ and $(u_b, 0)^T$ with $\lim_{n\to\infty} v_b^n = v_b$. This is important since an apparently wrong solution $\tilde{\mathbf{u}}^n$ might converge to the correct weak solution for $n \to \infty$. For convenience's sake, we replace v_b^n, S_2^n and $\bar{S}^n$ in (7.27) by the corresponding limit values, v_b, S_2 and $\bar{S} := \frac{1}{2}(v_b + u_b)$.

Theorem 7.8. *Let* u_u, s *and* S_2 *be given and assume that* u_b *is the final state of a strong or CJ detonation wave propagating with speed* s *and* v_b *is the final state of a detonation wave propagating with speed* S_2, *i.e.* (7.5) *and* (7.32) *hold. Suppose the initial data* u^0 *are given by (see* (7.27)*)*

$$(u^0(x), Y^0(x))^T = \begin{cases} (u_b, 0)^T, & x < a, \\ (v_b, 0)^T, & a < x < b, \\ (u_u, 1)^T, & x > b, \end{cases} \tag{7.33}$$

where $a < b$ *are given constants. Let* $\mathrm{u} = (u, Y)^T$ *be a weak solution of* (7.1) *with initial data* (7.33) *and suppose that* u *consists of at most three constant states for all* t. *Then for* t *sufficiently large the following holds,*

(i) *if* $v_b > u_{we}$, *then the weak solution* u *consist of two constant states separated by a strong or CJ detonation wave propagating with a speed* s, *i.e.*

$$\mathrm{u}(x,t) = (u(x,t), Y(x,t))^T = \begin{cases} (u_b, 0)^T, & x < d + st, \\ (u_u, 1)^T, & x > d + st, \end{cases} \tag{7.34}$$

for some constant $d > a$;

or,

(ii) *if* $v_b \leq u_{we}$, *then the weak solution* u *consist of a weak detonation wave propagating with speed* $S_2 \geq s$, *followed by an ordinary shock wave propagating with speed* $\bar{S} = (u_b + v_b)/2 < s$, *i.e.*

$$\mathrm{u}(x,t) = (u(x,t), Y(x,t))^T = \begin{cases} (u_b, 0)^T, & x < a + \bar{S}t, \\ (v_b, 0)^T, & a + \bar{S}t < x < b + S_2 t, \\ (u_u, 1)^T, & b + S_2 t < x. \end{cases} \tag{7.35}$$

Here u_{we} *denotes the final state of the weak detonation propagating with speed* s *(see* (7.10)*).*

Note that if $v_b = u_b$ (i.e. $S_2 = s$) and $b = 0$, then (7.33) reduces to (7.23) and the physically correct weak solution is given by (7.34) with $d = 0$. If $v_b > u_{we}$, then (7.33) converges to the correct weak solution for $t \rightarrow \infty$. This property is used to derive a criterion which excludes the weak solutions described by (ii). In Theorem 7.8 we have assumed that the solution consists of at most three constant states for all t. In other words, there are no "new" constant states created as time evolves. This assumption is not very restrictive, since in numerical experiments we have never observed these "new" constant states. Moreover, we believe that Theorem 7.8 also holds without this assumption, since "new" constant states, not equal to u_u, v_b or u_b, will probably not remain constant as time evolves.

Proof. As noted before, the minimum speed for a detonation wave is the speed s_{CJ} of a CJ detonation. It will be useful to consider u_{st} and u_{we} as a function of s. Therefore, we define two functions $g_{st} : [s_{CJ}, \infty) \to I\!R$ and $g_{we} : [s_{CJ}, \infty) \to I\!R$ as (see (7.9) and (7.10))

$$\begin{aligned} g_{st}(s) &:= s + \sqrt{(s-u_u)^2 - 2sQ}, \\ g_{we}(s) &:= s - \sqrt{(s-u_u)^2 - 2sQ}. \end{aligned}$$

Note that $s_{CJ} = u_{CJ} = g_{st}(s_{CJ}) = g_{we}(s_{CJ})$. From $s - u_u - Q > \sqrt{(s-u_u)^2 - 2sQ}$ it follows that $g'_{we}(s) < 0$ for all $s > s_{CJ}$ and, subsequently, $g_{we}(s) \le g_{we}(s_{CJ}) = u_{CJ}$. Using this together with $g_{st}(s) \ge s \ge s_{CJ} = u_{CJ}$, we derive

$$u_{we} \le u_{CJ} \le u_{st}.$$

Clearly, (7.5), (7.32) and $v_b = u_b$ imply $S_2 = s$, i.e. (i) holds and u is given by (7.34) with $d = b$. The remainder of the proof ($v_b \neq u_b$) is given in two steps. In step 1 it is shown that if $v_b > u_b$, then for t sufficiently large the weak solution of (7.1),(7.33) is given by (7.34) with $d > b$. In step 2 it is shown that for $v_b < u_b$, we must distinguish between two cases. If $S_2 < s$, then the weak solution of (7.1),(7.33) is given by (7.34) with $a < d < b$. On the other hand, if $S_2 \ge s$, then (7.35) describes the weak solution of (7.1),(7.33) (i.e. (ii) holds).

Step 1. In this step we assume $v_b > u_b$. It follows from this, $u_b \ge s \ge s_{CJ} = u_{CJ}$ and $u_{we} \le u_{CJ} \le u_{st}$ that v_b is the final state of a strong detonation and therefore, $v_b > S_2$. Furthermore, $v_b > u_b$ implies that the detonation wave, which connects the state $(u_u, 1)^T$ with $(v_b, 0)^T$, is followed by a rarefaction wave consisting of a smooth transition from $(v_b, 0)^T$ to $(u_b, 0)^T$. However, this solution is unstable, since the head of the rarefaction wave is propagating with a speed $v_b > S_2$. Hence, as t increases the rarefaction wave will overtake the detonation wave and slows it down until the propagation speed is equal to s and the state behind the detonation wave becomes u_b (see left figure in Figure 7.11). Hence, the weak solution of (7.1),(7.33) is given by (7.34) with $d > b$ for t sufficiently large.

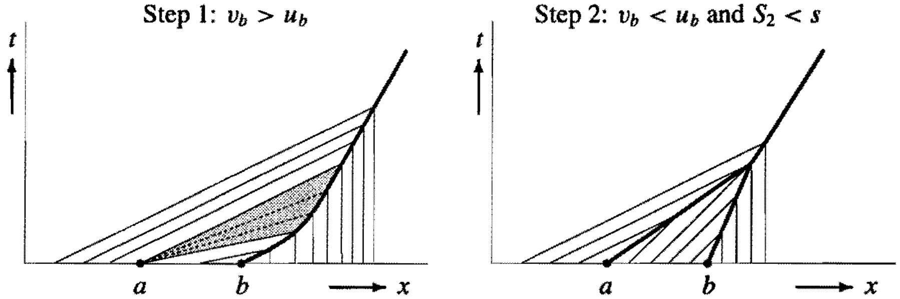

Figure 7.11. Characteristics corresponding to the formation of the solution (7.34) as described in the proof of Theorem 7.8.

Step 2. Suppose $v_b < u_b$ and $S_2 < s$. In this case the detonation wave is followed by a shock wave, which connects $(v_b, 0)^T$ with $(u_b, 0)^T$. Let $\bar{S}$ denote the speed of this shock wave, i.e. $\bar{S} := (u_b + v_b)/2$. Using this together with (7.5) and (7.32), we obtain

$$\begin{aligned} \bar{S}(u_b - v_b) &= \tfrac{1}{2}u_b^2 - \tfrac{1}{2}v_b^2 = \tfrac{1}{2}u_b^2 - \tfrac{1}{2}u_u^2 - S_2(v_b - u_u - Q) \\ &= s(u_b - u_u - Q) - S_2(v_b - u_u - Q) \\ &> S_2(u_b - u_u - Q) - S_2(v_b - u_u - Q) = S_2(u_b - v_b). \end{aligned}$$

Hence $\bar{S} > S_2$ and therefore this solution is also unstable, since the shock wave will overtake the detonation wave and will accelerate it until the detonation is propagating with speed s and the state behind the detonation wave is equal to u_b (see right figure in Figure 7.11). Therefore, for t sufficiently large, the weak solution of (7.1),(7.33) is given by (7.34) with $a < d < b$.

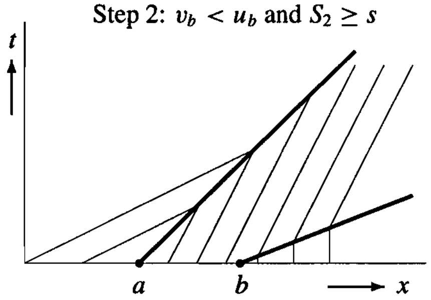

Figure 7.12. Characteristics corresponding to the solution (7.35) as described in the proof of Theorem 7.8.

Finally we consider the case $v_b < u_b$ and $S_2 \geq s$. Suppose $v_b \geq S_2$. Since g_{st} is increasing and $S_2 \geq s$, we have $v_b = g_{st}(S_2) \geq g_{st}(s) = u_b$. This is in contradiction with $v_b < u_b$, so $v_b < S_2$ and the detonation wave is a weak detonation wave (see Section 7.1.2). This weak detonation wave is followed by a shock wave, which connects $(v_b, 0)^T$ to $(u_b, 0)^T$. Again we denote the speed of the shock wave by $\bar{S}$, i.e. $\bar{S} = (u_b + v_b)/2$. In a similar way as the previous case ($S_2 \leq s$), we can prove that $\bar{S} \leq S_2$ and the shock will not overtake the detonation wave as time evolves (see Figure 7.12). This is the stable solution described by (ii). Obviously, in this case the weak solution u is given by (7.35). Since g_{we} is decreasing and $S_2 \geq s$, we obtain $v_b = g_{we}(S_2) \leq g_{we}(s) = u_{we}$. This completes the proof. □

Hence, weak detonation waves will not occur or disappear as time evolves, if $v_b > u_{we}$. Consider method (7.19) and suppose that in some cell $[x_{i_0-1/2}, x_{i_0+1/2})$ the gas is burnt during the $(n+1)$st time step, i.e. $C_{i_0}^{n+1} \geq u_{ign}$ and $Y_{i_0}^n = 1$. Then (7.19a) implies that

$$U_{i_0}^{n+1} \geq u_{ign} + \frac{\Delta t Da}{1 + \Delta t Da} Q. \tag{7.36}$$

Note that $U_{i_0}^{n+1}$ is the state immediately behind the detonation wave and therefore may be interpreted as the quantity v_b^n defined in (7.25). Using Theorem 7.7 and Theorem 7.8 we should require that $v_b^n > u_{we}$ in order to exclude nonphysical weak detonations. Hence, using $U_{i_0}^{n+1} \approx v_b^n > u_{we}$ and (7.36) it seems useful to require that the following inequality holds

$$u_{ign} + \frac{\Delta t Da}{1 + \Delta t Da} Q \; > \; u_{we}. \tag{7.37}$$

If (7.37) is satisfied, we expect $v_b^n > u_{we}$ and thus, $S_2^n \geq s_{CJ}$. In general $\Delta t Da$ will be very large and (7.37) reduces to $u_{ign} + Q > u_{we}$. It appears that the nonphysical weak detonation waves are only observed if u_{ign} is close to u_u. Since $u_u + Q < u_{we}$ (see (7.5) and $s > 0$), inequality (7.37) is not satisfied and due to numerical diffusion, u is raised above the ignition value and an artificial reaction is started (see Figure 7.8).

However, our numerical experiments in the next section illustrate that for physically more realistic values of u_{ign}, (7.37) is satisfied and $v_b = u_b$. We emphasize that criterion (7.37) is only useful for stiff combustion chemistry (i.e. the gas is completely burnt during a time step Δt). Therefore, it may be used as a criterion on the ignition value but not on the time step Δt.

7.3.2. Numerical Results

In this section we show that criterion (7.37) is a useful one to exclude nonphysical solutions. In Example 7.3 and Example 7.4 it has been shown that for small mesh widths, the numerical solution is correct. However, if the mesh width is increased to more practical values, then the solution becomes a nonphysical weak detonation wave. Moreover, in Figures 7.5 and 7.6 we have used a nonphysical ignition value u_{ign} (close to u_u) and (7.37) is not satisfied. However, in practice the ignition value is much higher and (7.37) will be satisfied. Naturally, we consider fast reactions (or large mesh widths), since then the wrong solutions occur.

Example 7.9. In this example we consider the strong detonation of Example 7.2. In order to investigate the practical use of (7.37), we choose initial data corresponding to the nonphysical solution (7.27). Then we examine whether this solution has a temporally constant profile or transforms into the physically correct detonation wave as time evolves. Let the initial data be given by

$$(u^0(x), Y^0(x))^T = \begin{cases} (u_{st}, 0)^T, & x < -30, \\ (u_{we}, 0)^T, & -30 < x < 0, \\ (u_u, 1)^T, & x > 0, \end{cases} \tag{7.38}$$

where $u_{st} = 6$, $u_{we} = 3$ and $u_u = 0$. Analogously to Theorem 7.7, v_b^n denotes the value of u behind the numerical detonation wave, so initially $v_b^0 = u_{we}$. Furthermore, $\Delta t = 0.125$, $Da = 3.1192 \cdot 10^5$ and $Q = 2$, so (7.37) is rewritten as $u_{ign} + 2 > u_{we} = 3$. It follows from (7.15) that $L_{1/2} = 10^{-5}$.

u_{ign}	S_2	v_b	(7.37) satisfied
0.2	9.0000	2.2917	no
0.4	5.5800	2.6017	no
0.8	4.5000	3.0000	no
1.0	4.5000	3.0000	no
1.1	4.5000	6.0000	yes
1.2	4.5000	6.0000	yes

Table 7.13. Numerical results for method (7.19) with $Q = 2$, $f = 1.265625$, $Da = 3.1192 \cdot 10^5 \, (L_{1/2} = (9/8) \cdot 10^{-5} \Delta x)$, $\Delta x = 1.125$, $\Delta t = 0.125$ and initial data (7.38).

In Table 7.13 we have printed the limit values $S_2 = \lim_{n\to\infty} S_2^n$ and $v_b = \lim_{n\to\infty} v_b^n$. The results clearly show that if (7.37) is satisfied, the weak detonation wave is unstable and after some period (7.19) will approximate the correct strong detonation wave. □

Example 7.10. In the last example of this section we consider the CJ detonation of Example 7.1. We increase the Damköhler number to $Da = 6.9315 \cdot 10^5$ and use initial data (7.23). The exact ZND solution propagates with speed $s = 1$. Furthermore, $\Delta t = 0.25$, $\Delta x = 1$ (i.e. $\Delta x = 10^6 L_{1/2}$) and (7.37) is rewritten as $u_{ign} + 0.5 > 1$. In Figure 7.14 the numerical results are compared with the exact solution for $u_{ign} = 0.51$. In this case (7.37) is fulfilled, since $u_{ign} + 0.5 = 1.01 > 1$. The results in Figure 7.14 clearly illustrate that the numerical detonation wave approximates the correct CJ detonation wave. So method (7.19) captures the detonation wave correctly, even for fast reactions ($\Delta x = 10^6 L_{1/2}$).

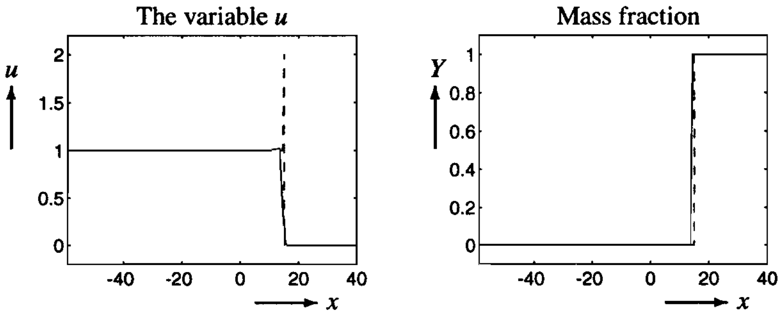

Figure 7.14. Numerical results for method (7.19) (with $\Delta t = 0.25$ and $\Delta x = 1.0$); numerical solution (solid line) and exact solution (dashed line) of (7.1),(7.23) at $t = 16$ of a CJ detonation with $Q = 0.5$, $f = 1.0$, $Da = 6.9315 \cdot 10^5 \, (L_{1/2} = 10^{-6} \Delta x)$ and $u_{ign} = 0.51$.

In Table 7.15 we present the corresponding limit values $S_2 = \lim_{n\to\infty} S_2^n$ and $v_b = \lim_{n\to\infty} v_b^n$. The results in Table 7.15 convincingly illustrate that (7.37) is a sufficient condition to obtain the correct CJ detonation wave. □

u_{ign}	S_2	v_b	(7.37) satisfied
0.01	4.0000	0.5359	no
0.1	1.6000	0.6202	no
0.2	1.1429	0.7388	no
0.3	1.0286	0.8571	no
0.4	1.0000	1.0000	no
0.51	1.0000	1.0000	yes
0.6	1.0000	1.0000	yes

Table 7.15. Numerical results for method (7.19) with $Q = 0.5$, $f = 1.0$, $Da = 6.9315 \cdot 10^5$ ($L_{1/2} = 10^{-6}\Delta x$), $\Delta x = 1.0$, $\Delta t = 0.25$ and initial data (7.23).

For stiff combustion chemistry (i.e. the gas is completely burnt during a time step Δt), method (7.19) captures the detonation wave correctly, provided the correct ignition value $u_{ign} \gg u_u$ is used. Due to the chemical reaction we observe a sharp resolution of the detonation wave without excessive smearing or oscillations. Finally, we like the method to be second order accurate in smooth parts of the flow (see Section 1.3). To this end we develop a high resolution method for the simplified detonation model in the next section.

7.4. A HIGH RESOLUTION METHOD

In this section we describe a high resolution method for the simplified detonation model (7.1), i.e. a method which achieves second order accuracy in smooth parts of the flow (except perhaps near extrema). The method is based on the second order splitting method of Strang (6.17). First we describe method (6.29) in the present context. Suppose the numerical solution at time level t^n is known. The numerical solution at the next time level $t^{n+1} = t^n + \Delta t$ is computed in three steps (corresponding to $\mathcal{L}^c_{\Delta t/2}$, $\mathcal{L}^f_{\Delta t}$ and $\mathcal{L}^c_{\Delta t/2}$).

Step 1. In the first step the convection is neglected. The remaining ODE with initial data $\mathbf{U}_i^n$ is solved exactly. Denote the result at time $t^{n+1/2} := t^n + \Delta t/2$ by $\mathbf{U}_i^{n+1/2} = (U_i^{n+1/2}, Y_i^{n+1/2})^T$.

Step 2. In the second step we assume that no reaction occurs and approximate the solution of the remaining homogeneous equation, i.e. Burgers' equation (Y remains constant in the second step). In this approximation initial data $\mathbf{U}_i^{n+1/2}$ are used, as results from step 1. The numerical solution at time $t^n + \Delta t$ is computed using the slope limiter method of Example 5.17. The result is denoted by $\tilde{\mathbf{U}}_i^{n+1/2} = (\tilde{U}_i^{n+1/2}, \tilde{Y}_i^{n+1/2})^T$ (i.e. $\tilde{Y}_i^{n+1/2} = Y_i^{n+1/2}$).

Step 3. Finally, in the last step we obtain the numerical solution $\mathbf{U}^{n+1}$ at time level t^{n+1} by applying step 1 again (i.e. no convection) with initial data $\tilde{\mathbf{U}}_i^{n+1/2}$.

Using the notation of Section 6.3, $\mathcal{L}^f_{\Delta t}$ is given by the slope limiter method and $\mathcal{L}^c_{\Delta t/2} = \mathcal{S}^c_{\Delta t/2}$ is the exact solution operator. As usual, the time step Δt is restricted by the CFL stability condition, i.e. $\tau \max_i |U_i^n| \leq 1$.

In the following examples numerical results are presented for the method above. It turns out that the behaviour of the high resolution method corresponds very much with the behaviour of method (7.19). Again, we observe nonphysical wave speeds for large mesh widths and low ignition values.

Example 7.11. In this example we consider the strong detonation of Example 7.2. Furthermore, we use Roe's superbee limiter (5.23) in the computation of the numerical flux function in step 2.

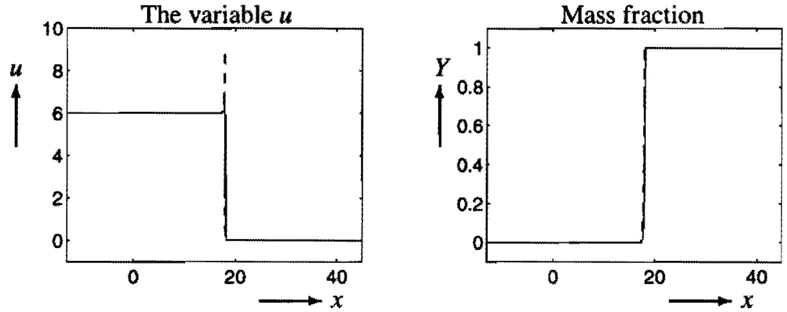

Figure 7.16. Numerical results for the high resolution method (with $\Delta t = 0.01$ and $\Delta x = 0.1$); numerical solution (solid line) and exact solution (dashed line) of (7.1),(7.16) at $t = 4$ of a strong detonation with $Q = 2$, $f = 1.265625$, $Da = 31.192$ ($L_{1/2} = \Delta x$) and $u_{ign} = 0.1$.

In Figure 7.16 the numerical results are compared with the exact solution. The mesh width $\Delta x = L_{1/2} = 0.1$ and the time step $\Delta t = 0.01$. The numerical ZND profile is essentially correct. Comparing the results with those in Figure 7.5, no important differences are observed.

For larger mesh widths ($L_{1/2} = 0.125\Delta x$), the numerical solution should still propagate with a wave speed $s = 4.5$. However, Figure 7.17 clearly illustrates that the numerical solution is completely wrong. Again, we observe a weak detonation wave propagating with a wave speed 10 (one cell per time step). Since the slope limiter method also introduces intermediate states $u_u < u < u_b$ ahead of the leading shock wave, the global explanation described in Figure 7.8 still holds. Comparing the results with those of Figure 7.6, we clearly observe a sharper resolution of the ordinary shock wave. However, this improvement is irrelevant, since the solution remains totally wrong.

As expected, for a physically more realistic value $u_{ign} = 1.5$ (see Figure 7.18), the numerical solution is the correct strong detonation wave. However, there is no clear improvement noticeable with respect to Figure 7.7. ☐

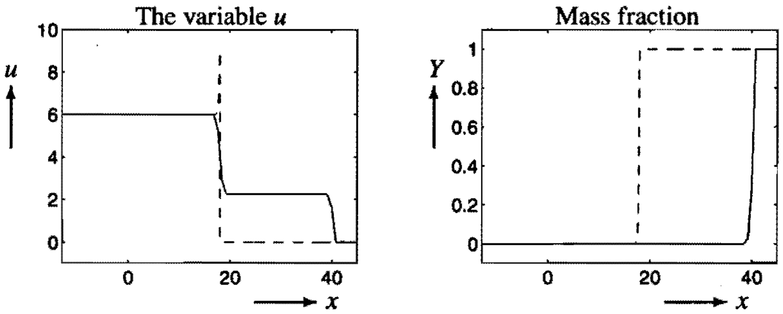

Figure 7.17. Numerical results for the high resolution method (with $\Delta t = 0.08$ and$\Delta x = 0.8$); numerical solution (solid line) and exact solution (dashed line) of (7.1),(7.16) at $t = 4$ of a strong detonation with $Q = 2$, $f = 1.265625$, $Da = 31.192\,(L_{1/2} = 0.125\Delta x)$ and $u_{ign} = 0.1$.

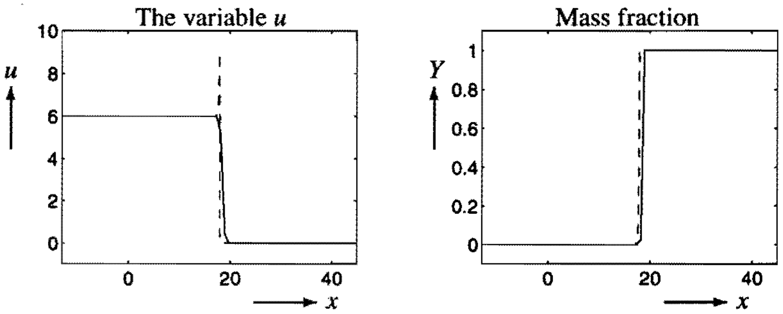

Figure 7.18. Numerical results for the high resolution method (with $\Delta t = 0.08$ and $\Delta x = 0.8$); numerical solution (solid line) and exact solution (dashed line) of (7.1),(7.16) at $t = 4$ of a strong detonation with $Q = 2$, $f = 1.265625$, $Da = 31.192\,(L_{1/2} = 0.125\Delta x)$ and $u_{ign} = 1.5$.

Now we consider Example 7.5 again, and investigate whether the order of accuracy of the high resolution method is larger than one.

Example 7.12. In this example we consider the strong detonation of Example 7.2 and use initial data (7.16). Again Roe's superbee limiter (5.23) is used in the computation of the numerical flux function in step 2. We only consider the first component of the global error $\mathbf{E}^n$, as defined in (7.22).

It follows from Table 7.19 that the accuracy of the high resolution method is far less than satisfactory. The results are only slightly better than those of Table 7.10 (method (7.19)). This should be expected, however, since the use of a slope limiter tends to clip the extreme point (the vN spike) behind the ordinary shock wave. Moreover, the method is in fact first order accurate around extreme points (see Chapter 5). ☐

l	Δt_l	global error $\|E_1^{n_l}\|_1$
0	1/25	$1.5778 \cdot 10^{+0}$
1	1/50	$7.6731 \cdot 10^{-1}$
2	1/100	$2.1002 \cdot 10^{-1}$
3	1/200	$1.0720 \cdot 10^{-1}$
4	1/400	$5.2625 \cdot 10^{-2}$
5	1/800	$2.6427 \cdot 10^{-2}$

Table 7.19. Global error for the high resolution method at $t = 1$ with $Q = 2$, $f = 1.265625$, $Da = 3.1192$ ($L_{1/2} = 1$), $u_{ign} = 1$, $\Delta t_l = (1/25) \cdot (1/2)^l$, $\Delta x_l = (10/25) \cdot (1/2)^l$, $E_1^{n_l}$ defined by (7.22) and initial data (7.16).

There are several reasons for the absence of real differences between the results for method (7.19) and the high resolution method described in this section. First, due to the limiter, the high resolution method is modified around discontinuities, thereby increasing the amount of numerical diffusion (see Chapter 5). Moreover, almost every numerical method for homogeneous conservation laws spreads the shock over a couple of mesh points (except, for example, front tracking methods). Hence, the numerical solution always contains intermediate values $u_u < u < u_b$ ahead of the detonation wave (see Figure 7.8). If the ignition value u_{ign} is close to u_u, then (also for the high resolution method) an artificial reaction is started. Therefore, it is no surprise that for low ignition values and large Damköhler numbers, the wrong wave speeds occur.

Secondly, if $u > u_{ign}$ and Da is large enough, the gas is completely burnt in the next time step and u is shifted to the equilibrium state u_b. So, even with the first order method (7.19) a sharp resolution of the detonation wave is obtained (see Figure 7.7 and Figure 7.14). Obviously, nonreactive discontinuities are represented better by the high resolution method (see Chapter 5). This is illustrated in Figure 7.6 and Figure 7.17. The resolution of the reactive shock (i.e. detonation) in both figures is the same. However, we observe a sharper resolution of the nonreactive shock in Figure 7.17.

8

NUMERICAL SOLUTION OF THE REACTIVE EULER EQUATIONS

In this chapter we consider detonation waves for the reactive Euler equations. As noted before, the reactive Euler equations are a system of first order hyperbolic conservation laws. Since detonation waves have a discontinuous structure, including a strong leading shock front, we consider weak solutions of hyperbolic differential equations. A difficulty is that weak solutions turn out to be nonunique (see Chapter 3) and we have to characterize the unique "physically relevant" weak solution. In case of the ZND model for detonation waves, this unique solution is derived by considering strong or Chapman-Jouguet detonation waves only (see Chapter 2).

As for the simplified detonation model, also for the reactive Euler equations stable numerical solutions may occur, which are yet completely wrong, because the discontinuities have the wrong locations. Thus the numerical reaction waves are propagating with nonphysical wave speeds. These "wrong solutions" turn out to be approximations of nonphysical weak solutions (i.e. weak detonation waves) [16].

In general the nonphysical weak detonation waves are observed only if the ignition temperature is close to the temperature of the unburnt gas. For these relatively low values of the ignition temperature, numerical experiments show that only on very fine meshes the computed solution will approximate the correct weak solution. In most practical cases we cannot afford such fine meshes and therefore we want to exclude the nonphysical weak detonation waves also on relatively coarse meshes.

In practical applications the ignition temperature is much higher than the temperature of the unburnt gas and our numerical results illustrate that for these ignition temperatures the nonphysical weak solutions will not occur. To our knowledge, this result has not been noticed before. Moreover, our study appears to be the first attempt to study the influence of the ignition temperature on the numerical solution. In this chapter we extend the theory of Chapter 7 to the reactive Euler equations and explain why a correct ignition temperature will exclude the wrong solutions.

This chapter is organized as follows. In the first section, we present a numerical method based on the first order splitting method (6.14). In this method the fluid dynamical part is solved by Roe's first order method and the chemistry is solved exactly using a fixed

temperature approximation. Numerical results show the occurrence of wrong solutions (weak detonation waves) for low ignition temperatures and correct solutions (strong or CJ detonation waves) for realistic ignition temperatures. In Section 8.2 we extend the criterion (7.37) to the reactive Euler equations in order to explain the influence of the ignition temperature on the numerical solution. Furthermore, we present numerical results supporting our claim that a correct ignition temperature excludes the nonphysical weak solutions. In Section 8.3 we describe a high resolution method based on the second order splitting method (6.17) and present some numerical results for this method. For the sake of completeness, in Section 8.4 we make some remarks on front tracking. Front tracking methods are often used to solve the numerical difficulty of approximating incorrect weak solutions [10, 76].

8.1. NUMERICAL COMPUTATION OF STRONG OR CJ DETONATION WAVES

In this chapter we consider the reactive Euler equations (2.26). As in Chapter 7 we assume that the initial data correspond to the exact ZND solution of a strong or CJ detonation wave, where the ordinary shock wave is located at $x = 0$. To obtain the exact ZND solution at $t = 0$, we have to solve (2.49). In general (2.49) cannot be solved exactly and the solution Y must be obtained numerically. If we have computed Y, then all other variables can be determined from (2.48). Hence, the exact solution is the strong or CJ detonation described in Section 2.5. If u, f and q are defined by (3.6), then the reactive Euler equations (2.26) can be written in the general form (3.2a). The reaction rate w depends on the temperature T via the Arrhenius' law (see (2.29) and (2.30))

$$w = \begin{cases} 0, & T < T_{ign}, \\ -Da\ \rho Y \exp(E_a(\frac{1}{T_{vN}} - \frac{1}{T})), & T \geq T_{ign}. \end{cases} \tag{8.1}$$

Detonation waves are characterized by a shock heating the gas above a certain temperature (the von Neumann temperature) and a subsequent rapid heat release. In practically all realistic cases T_{ign} is much higher than the temperature of the unburnt gas, i.e. $T_u \ll T_{ign} \leq T_{vN}$. From a mathematical point of view, this is a consequence of a characteristic exponential (Arrhenius-type) behaviour of the reaction rate laws, which guarantees the rates to be exponentially small for low temperatures. On the other hand, from a chemical point of view, this can be thought of as being caused by a kinetic competition for radical species, which for low temperatures is completely on the side of the radical consuming reactions. In other words, at low temperatures the reactions are kinetically quenched, due to a lack of radicals.

We choose a time step Δt and a mesh width Δx to define discrete time levels $t^n := n\Delta t$ for all $n \geq 0$ and discrete mesh points $x_i := i\Delta x$ for all i. We are rather interested in detonation capturing and do not want to describe all physical details. Therefore, we consider mesh widths Δx and time steps Δt appropriate to resolve the fluid dynamics but not for the chemical reaction. What we should demand is that our method produces a sharp representation of the detonation wave at the correct location. The reactive Euler equations (2.26) are solved numerically using a method based on the first order splitting method (6.14). In this method the numerical solution at each time level is derived in two steps. In the first step we assume that no reaction occurs (i.e. (3.2a) with $\mathrm{q} = 0$) and approximate the solution of the remaining homogeneous system, i.e. the nonreactive Euler equations. In the second step we assume no convection (i.e. (3.2a) with $\partial \mathbf{f}/\partial x = 0$) and solve the corresponding ordinary differential equations numerically. For a detailed description of splitting methods we refer to Chapter 6.

Next we describe in more detail the method used in this section. Let the numerical solution at time level t^n (i.e. $\mathbf{U}_i^n$ for all i) be given. In the first step we solve the nonreactive Euler equations numerically by Roe's first order method for gas dynamics (see Example 4.18). Let us denote the result by $\mathbf{C}_i^{n+1} \in \mathbb{R}^4$, so

$$\mathbf{C}_i^{n+1} \;:=\; \mathbf{U}_i^n \;-\; \tau\{\mathbf{F}_{i+1/2}^n - \mathbf{F}_{i-1/2}^n\},$$

where the numerical flux function $\mathbf{F}_{i+1/2}^n$ is given by (4.49). Let $\tilde{g}_i^{n+1}$ denote the result in the ith cell after the first step, for all variables g, i.e. the vector $\mathbf{C}_i^{n+1}$ is given by $\mathbf{C}_i^{n+1} = (\tilde{\rho}_i^{n+1}, \tilde{\rho}_i^{n+1}\tilde{u}_i^{n+1}, \tilde{\rho}_i^{n+1}\tilde{E}_i^{n+1}, \tilde{\rho}_i^{n+1}\tilde{Y}_i^{n+1})^T$. Recall that a leading shock wave initiates a chemical reaction. Therefore, we assume that no chemical reaction occurs in the cell $[x_{i-1/2}, x_{i+1/2})$ during the $(n+1)$st time step, if $\tilde{T}_i^{n+1} < T_{ign}$. Furthermore, since the first three equations in (3.2a) are independent of q, $\rho_i^{n+1} = \tilde{\rho}_i^{n+1}$, $u_i^{n+1} = \tilde{u}_i^{n+1}$ and $E_i^{n+1} = \tilde{E}_i^{n+1}$. Hence, in the second step we have to solve the following scalar initial value problem (see (8.1))

$$\frac{\mathrm{d}}{\mathrm{d}t}Y(x_i,t) \;=\; -\,Da\; H(\tilde{T}_i^{n+1} - T_{ign})Y(x_i,t)\exp\Big(E_a\Big(\frac{1}{T_{vN}} - \frac{1}{T(Y(x_i,t))}\Big)\Big), \quad (8.2a)$$

$$Y(x_i,t^*) \;=\; \tilde{Y}_i^{n+1}, \quad (8.2b)$$

for all i, where H is the Heavyside function and $T(Y)$ is given by (see (2.27) and (2.28))

$$T(Y) \;=\; (\gamma-1)(\tilde{E}_i^{n+1} - \tfrac{1}{2}(\tilde{u}_i^{n+1})^2 - QY).$$

Furthermore, t^* is some fixed initial time, which is equal to t^n or $t^n + \Delta t/2$ in general.

As remarked before, we don't want to resolve the chemistry in detail and an explicit method for the ordinary differential equation (8.2) would imply a severe time step restriction (compared with the usual CFL condition). Therefore, we have to solve the ODE in the second step by an implicit method. However, the Arrhenius' law is complicated and even the backward Euler method for (8.2) results in a nonlinear equation for Y_i^{n+1} with no unique solution. Therefore, we replace $T(Y)$ by $\tilde{T}_i^{n+1/2}$ in (8.2a) and integrate the resulting linear ODE for Y over $[t^*, t^* + \Delta t]$ exactly, i.e.

$$Y_i^{n+1} \;=\; \begin{cases} \tilde{Y}_i^{n+1}, & \tilde{T}_i^{n+1} < T_{ign}, \\ \tilde{Y}_i^{n+1}\exp(-r\Delta t\; Da), & \tilde{T}_i^{n+1} \geq T_{ign}, \end{cases} \quad (8.3a)$$

where for shortness of notation $r > 0$ is given by

$$r \;:=\; \exp\Big(E_a\Big(\frac{1}{T_{vN}} - \frac{1}{\tilde{T}_i^{n+1}}\Big)\Big). \quad (8.3b)$$

Since E_a is large in general, r is large when $\tilde{T}_i^{n+1}$ is sufficiently high and negligible for low $\tilde{T}_i^{n+1}$. At first glance (8.3) may appear to be less satisfactory than an implicit numerical method for (8.2a), since in practice the temperature is not constant. However, in our applications a detailed description of the chemistry is not necessary and the fixed temperature approach produces satisfying results.

Using the notation of Section 6.3, $\mathcal{L}^f_{\Delta t}$ is given by Roe's first order method and $\mathcal{L}^c_{\Delta t}$ is the exact solution operator of a modified problem (obtained by a fixed temperature approximation).

Next we present numerical results for method (8.3). In the first example it is shown that also for the reactive Euler equations, nonphysical wave speeds may occur. For simplicity, only the pressure and mass fraction are drawn.

Example 8.1. In this example we approximate the ZND solution of a strong detonation. Initially the preshock state is given by

$$p_u = 1, \quad \rho_u = 1, \quad u_u = 0.$$

Furthermore, we choose the following parameter values:

$$E_a = 10, \quad Q = 10, \quad f = 1.1, \quad \gamma = 1.4 \quad Da = 66.201.$$

The half-reaction length is given by $L_{1/2} = 10^{-2}$ (see (2.51)). The final burnt state for the strong detonation is given by (see (2.45))

$$p_b = p_{st} = 13.481, \quad \rho_b = \rho_{st} = 2.0741, \quad u_b = u_{st} = 2.5423,$$

where the strong detonation is propagating with a speed $s = 4.9093$. For this particular example the von Neumann pressure and the von Neumann temperature are given by $p_{vN} = 19.918$ and $T_{vN} = 4.2838$, respectively.

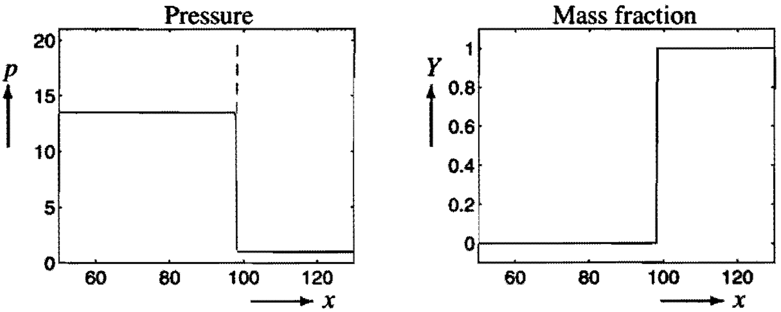

Figure 8.1. Numerical results for method (8.3) (with $\Delta t = 0.01$ and $\Delta x = 0.1$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 20$ of a strong detonation with $E_a = 10$, $Q = 10$, $f = 1.1$, $\gamma = 1.4$, $Da = 66.201$ ($L_{1/2} = 0.1\Delta x$) and $T_{ign} = 1.01$.

In Figure 8.1 the numerical results are compared with the exact solution. The pressure peak has completely disappeared, since we are using a large mesh width

($\Delta x = 10L_{1/2}$), on which the chemical reaction cannot be solved. Since we are interested in the global behaviour only, the results in Figure 8.1 are satisfying. The ignition temperature in Figure 8.1 is unrealistic (too low). However, the small time step and the exponential term in (8.1) guarantee a negligible reaction rate for low temperatures. This will be explained in the next section. We observe a sharp resolution of the detonation, since the chemical reaction shifts all intermediate values (with sufficiently high temperatures) to the equilibrium state $\mathbf{u}_b$.

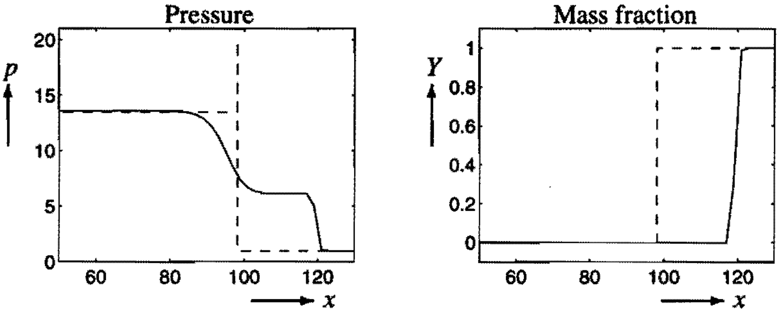

Figure 8.2. Numerical results for method (8.3) (with $\Delta t = 0.2$ and $\Delta x = 2.0$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 20$ of a strong detonation with $E_a = 10$, $Q = 10$, $f = 1.1$, $\gamma = 1.4$, $Da = 66.201$ ($L_{1/2} = 5 \cdot 10^{-3} \Delta x$) and $T_{ign} = 1.01$.

We increase Δx to 2.0 (i.e. $\Delta x = 200L_{1/2}$) and keep $\tau = \Delta x / \Delta t$ fixed. The numerical solution should still propagate with a wave speed $s = 4.9093$. However, Figure 8.2 clearly illustrates that the numerical solution is completely wrong. As in Figure 7.6, there is a weak detonation wave propagating faster than the exact detonation wave. In this weak detonation wave all energy is released and the gas is completely burnt.

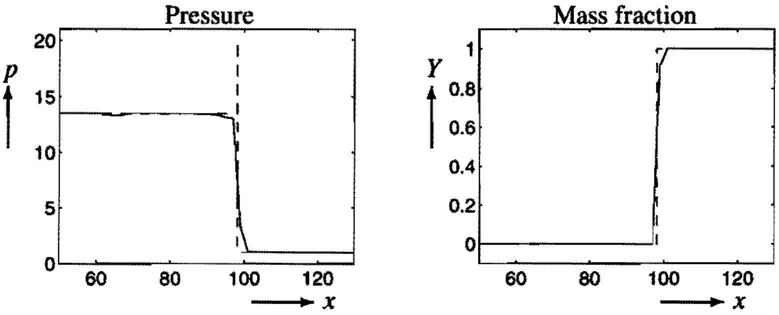

Figure 8.3. Numerical results for method (8.3) (with $\Delta t = 0.2$ and $\Delta x = 2.0$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 20$ of a strong detonation with $E_a = 10$, $Q = 10$, $f = 1.1$, $\gamma = 1.4$, $Da = 66.201$ ($L_{1/2} = 5 \cdot 10^{-3} \Delta x$) and $T_{ign} = T_{vN} = 4.2838$.

As remarked before, in practice $T_u \ll T_{ign} \leq T_{vN}$. In Figure 8.3 the solution is drawn for $T_{ign} = T_{vN}$. Although we observe some noise behind the detonation wave, the numerical solution approximates the global behaviour of the exact detonation wave very well. The disturbance behind the pressure jump is caused by the splitting method. This will be explained later on. ☐

The previous example shows the possible occurrence of nonphysical solutions. We like to draw attention to the fact that the reactive Euler equations and the simplified detonation model have a similar numerical behaviour (compare the results of Example 7.3 with those of Example 8.1). As in Section 7.2 (see Figure 7.8), the basic explanation for the occurrence of nonphysical wave speeds is the smeared representation of the leading shock wave, which includes intermediate temperatures $T_u < T < T_b$ in front of it. If $T_{ign} \approx T_u$, due to numerical diffusion, the temperature is raised above the ignition temperature and an artificial reaction is started in front of the shock wave. Apparently, $r\Delta t Da$ is large enough to burn a substantial part of the gas (see (8.3a)) and cause an increase of the numerical wave speed.

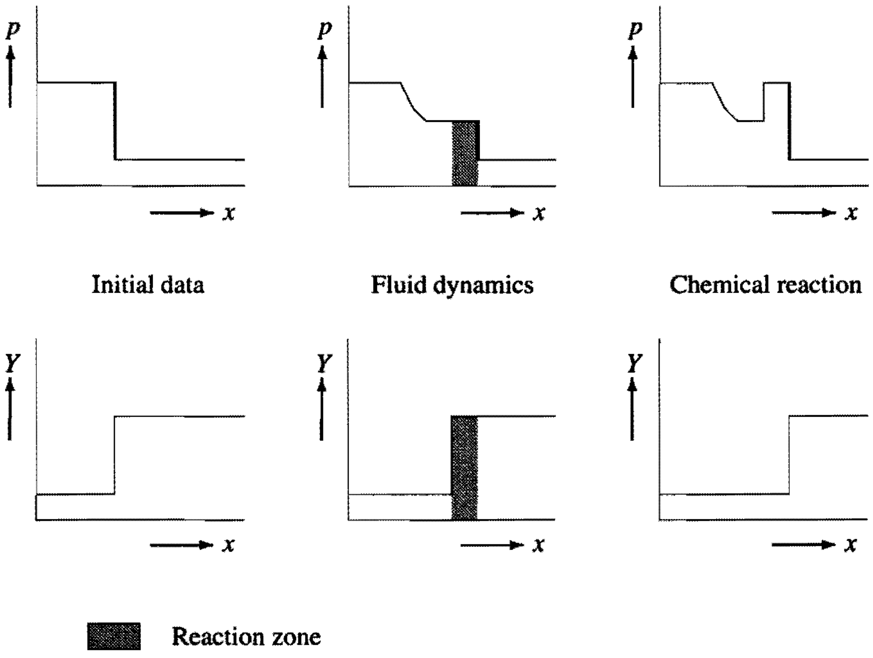

Figure 8.4. Global explanation of the occurrence of small disturbances in the pressure profile behind the detonation wave.

In Figure 8.3 we observe a small disturbance in the numerical solution behind the detonation wave. Often, we observe a lot of disturbances behind the detonation wave (see Figure 8.7). They are caused by the splitting method. Suppose the fluid dynamics and the chemistry are solved exactly and consider the Riemann problem at the boundary of two adjacent cells separating the constant states $\mathbf{u}_b$ and $\mathbf{u}_u$. In the first step of the splitting method we derive the exact solution of the Riemann problem at the cell boundary. Suppose the solution consists of a simple wave, a contact discontinuity and a shock wave, respectively (see Figure 8.4). For the sake of simplicity we restrict ourselves to the pressure. Note that in the first step Y is simply propagated along the contact discontinuity. In the second step, the gas is burnt in the grey area and the pressure jumps to a higher value in this region. Since in the rest of our computational domain the pressure remains constant, it is obvious that we may observe oscillations in the pressure behind the reacting shock wave. In order to avoid these oscillations it is necessary to know the position of the front accurately. One possibility is to use a front tracking method in the first step. In the front tracking method the piecewise constant profile is simply propagated in the positive x-direction, without creating simple waves or contact discontinuities.

In Example 8.1 the activation energy is relatively low, so the reaction rate is not dominated by the exponential term. In the following example we consider a strong detonation wave with high activation energy.

Example 8.2. In this example we consider the strong detonation of Example 2.2. We use initial data corresponding to the exact ZND profile and increase the Damköhler number to $6.3293 \cdot 10^3$ (i.e. $L_{1/2} = 10^{-5}$). The exact detonation wave is propagating with speed $s = 9.1359$. The results in Figure 8.5 illustrate the correct propagation speed of the numerical detonation wave. Again, no pressure peak can be seen and the pressure jump is smeared over a couple of mesh points. Although we use a low ignition temperature, the numerical solution represents the global behaviour of the exact solution very well.

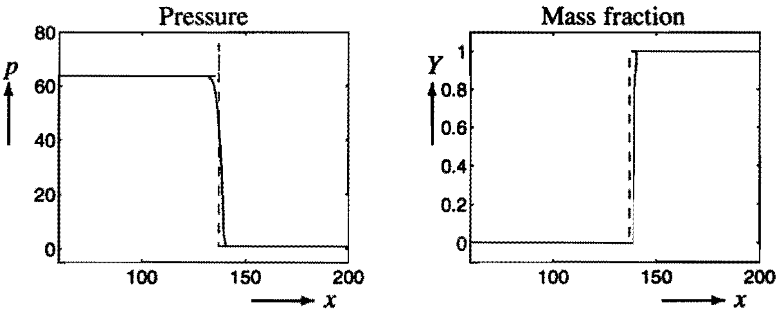

Figure 8.5. Numerical results for method (8.3) (with $\Delta t = 0.05$ and $\Delta x = 1.0$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 15$ of a strong detonation with $E_a = 150$, $Q = 50$, $f = 1.8$, $\gamma = 1.2$, $Da = 6.3293 \cdot 10^3$ ($L_{1/2} = 10^{-5}\Delta x$) and $T_{ign} = 1.01$.

To observe nonphysical wave speeds, we must increase Da to $6.3293 \cdot 10^{+15}$. In this extremely stiff case, the reaction rate (8.1) is no longer dominated by the exponential term. □

The numerical experiments in this section show that for high ignition temperatures we obtain the correct solution. This is explained in the next section. On the other hand, for low ignition temperatures the Damköhler number, the activation energy and the time step play a crucial role in obtaining the correct solution. Therefore, in the next section we also study the influence of these parameters on the numerical solution (and especially the location of the reaction zone). Naturally, we restrict ourselves to stiff combustion chemistry (i.e. the gas is completely burnt during a time step Δt), since then the wrong numerical solutions occur.

8.2. THE AVOIDANCE OF NONPHYSICAL DETONATION WAVES

As shown in Example 8.1, it is also possible to obtain a numerical solution with a reaction zone at the wrong location for the reactive Euler equations. As remarked before, these "wrong solutions" are nonphysical weak detonation waves. In this section we study the occurrence of nonphysical wave speeds. To this end we define a quantity S^n at time t^n as (see (7.20b))

$$-S^n n \Delta t = \Delta x \sum_{i=-\infty}^{\infty} \left(Y_i^n - Y_i^0\right), \tag{8.4}$$

where Y is the mass fraction of the unburnt gas. Again, S^n may be interpreted as the average speed of the numerical detonation wave at time level t^n.

For the simplified detonation model we have developed a criterion which explains the occurrence of nonphysical weak solutions for low ignition temperatures. Next we extend this criterion (7.37) to the reactive Euler equations. Suppose that in some cell $[x_{i_0-1/2}, x_{i_0+1/2})$ the gas is burnt during the $(n+1)$st time step, i.e. $\tilde{T}_{i_0}^{n+1} \geq T_{ign}$ and $\tilde{Y}_{i_0}^{n+1} > 0$. Contrary to the simplified detonation model, $\tilde{Y}_{i_0}^{n+1} \neq 1$ in general for the reactive Euler equations. In the first step of the splitting method Y is simply propagated along the contact discontinuity and therefore in some "unburnt cells" Y will decrease to values below 1. Hence, we only know that $0 \leq \tilde{Y}_{i_0}^n \leq 1$. It follows from (2.27) and (2.28) that

$$\tilde{T}_{i_0}^{n+1} = (\gamma - 1)(\tilde{E}_{i_0}^{n+1} - \tfrac{1}{2}(\tilde{u}_{i_0}^{n+1})^2 - Q\tilde{Y}_{i_0}^{n+1}).$$

Using this, it is easy to see that (see (8.3))

$$\begin{aligned} T_{i_0}^{n+1} &= \tilde{T}_{i_0}^{n+1} + (\gamma - 1)\, Q\,(\tilde{Y}_{i_0}^{n+1} - Y_{i_0}^{n+1}) \\ &\geq T_{ign} + (\gamma - 1)\, Q\, \tilde{Y}_{i_0}^{n+1}(1 - \exp(-r\Delta t Da)), \end{aligned}$$

where r is given by (8.3b). As in (7.37) we require the following inequality to hold

$$T_{ign} + (\gamma - 1)\, Q\, \tilde{Y}_{i_0}^{n+1}(1 - \exp(-r\Delta t Da)) > T_{we}, \tag{8.5}$$

where $T_{we} = p_{we}/\rho_{we}$ is the final temperature of the corresponding weak detonation wave propagating with speed s (see (2.46)).

We expect that for higher ignition temperatures T_{ign} (8.5) will be satisfied and the nonphysical weak detonations will not occur. Moreover, we show that (8.5) may also be used to study the influence of $\Delta t Da$ and E_a on the numerical wave speed. Obviously, it is not a desirable situation if the method produces the correct wave speed for one particular ignition temperature only. Considering the theory in Section 7.3, we expect that this is not the case. We now present some numerical results supporting this statement.

Example 8.3. In this example we consider the strong detonation of Example 8.1. In Table 8.6 the limit value $S = \lim_{n\to\infty} S^n$ is presented for several values of T_{ign} and Δx.

T_{ign}	Mesh width Δx				
	$10L_{1/2}$	$10^2 \cdot L_{1/2}$	$10^3 \cdot L_{1/2}$	$10^4 \cdot L_{1/2}$	$10^5 \cdot L_{1/2}$
1.01	4.9093	5.3404	7.5000	10.000	10.000
1.5	4.9093	5.1520	6.6666	7.5000	7.5000
2.0	4.9093	5.0000	5.0388	5.0388	5.0388
2.5	4.9093	4.9093	5.0000	5.0000	5.0000
3.0	4.9093	4.9093	4.9093	4.9093	4.9093
3.5	4.9093	4.9093	4.9093	4.9093	4.9093

Table 8.6. Average numerical wave speed $\lim_{n\to\infty} S^n$ for method (8.3) with $Q = 10, f = 1.1, \gamma = 1.4, Da = 66.201$ ($L_{1/2} = 0.01$), $\tau = \Delta t/\Delta x = 0.1$ and initial data corresponding to the exact ZND profile.

For low ignition temperatures the average numerical wave speed S is a good approximation of the exact wave speed for small mesh widths only ($\Delta x \approx 10L_{1/2}$). If we increase Δx and keep $L_{1/2}$ (or Da) fixed, we observe completely wrong wave speeds. However, if $T_{ign} \geq 3.0$ the numerical wave speed is correct, independent of the mesh width used. This interesting observation can be explained by (8.5). If $T_{ign} = 3.0$, (8.5) becomes approximately $3 + 4\tilde{Y}_{i_0}^{n+1} > 5.4506$, since $\exp(-r\Delta t Da)$ is negligible. This inequality is satisfied as $\tilde{Y}_{i_0}^{n+1} > 0.61265$. Our numerical results illustrate that (8.5) is satisfied in general if $T_{ign} \geq 3.0$. □

We present a second example illustrating that for high ignition temperatures the correct solution is obtained.

Example 8.4. In this example we consider the CJ detonation of Example 2.1. We use initial data corresponding to the exact initial profile (see Figure 2.5) and increase the Damköhler number to $Da = 6.4880 \cdot 10^4$ ($L_{1/2} = 10^{-5}$). The numerical solution should propagate with speed $s = 5.4419$. Note that $T_{vN} = 5.0509$. Furthermore, $\Delta t = 0.1$ and $\Delta x = 1$ (i.e. $L_{1/2} = 10^{-5}\Delta x$). For $T_{ign} = 4.0$, (8.5) becomes $3 + 5.6\tilde{Y}_{i_0}^{n+1} > 7.6921$, which is satisfied as $\tilde{Y}_{i_0}^{n+1} > 0.6593$.

The results in Figure 8.7 show that (for $T_{ign} = 4.0$) the numerical solution approximates the physically correct detonation wave. As noted before, in the first step of our method in some "unburnt cells" Y will decrease below 1. However, long before Y reaches

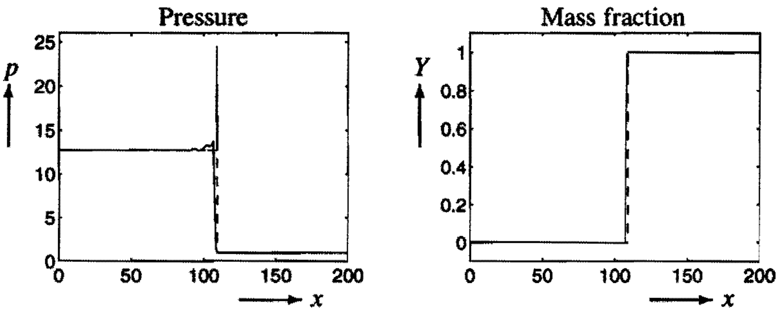

Figure 8.7. Numerical results for method (8.3) (with $\Delta t = 0.1$ and $\Delta x = 1.0$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 20$ of a CJ detonation with $E_a = 14$, $Q = 14$, $f = 1.0$, $\gamma = 1.4$, $Da = 6.4880 \cdot 10^4$ ($L_{1/2} = 10^{-5}\Delta x$) and $T_{ign} = 4.0$.

0.6593, the temperature increases above T_{ign} and the gas is burnt. So, (8.5) is satisfied in general, as is clearly illustrated by the results in Figure 8.7. The reaction zone has the correct location. The small disturbances behind the pressure jump are caused by the splitting method as explained in Figure 8.4. □

Hence, for realistic ignition temperatures (8.5) will be satisfied and nonphysical solutions will not occur. Some other authors call the nonphysical weak detonation waves a purely numerical artifact, since they disappear as the mesh is refined. We claim that the occurrence of wrong solutions is caused by using an unrealistic ignition temperature. The question that remains is why for low ignition temperatures, the correct solution is obtained for $\Delta t \rightarrow 0$. This may not be expected, since we are using a wrong physical model. Suppose $T_{ign} \approx T_u$ and recall that the wrong wave speeds are caused by the numerical diffusion in front of the leading shock wave (see Figure 7.8). Due to this diffusion the temperature increases above the ignition temperature and an artificial reaction is started ahead of the shock. Moreover, if $r\Delta t Da$ is large for this artificial reaction, then it follows from (8.3a) that during a time step Δt a substantial part of the gas is burnt and the numerical wave speed increases (see (8.4)). Hence, for low ignition temperatures, we expect S^n to depend on $r\Delta t Da$. To study for which temperatures $r\Delta t Da$ is large, we normalize $r\Delta t Da$ such that it reduces to an exponential term only (multiplied by one). Therefore, let the quantity Da^* be defined by

$$Da^* := \Delta t\, Da \exp(E_a(\frac{1}{T_{vN}} - \frac{1}{T^*})),$$

where $T_u < T^* < T_b$. Then $r\Delta t Da$ can be rewritten as

$$r\Delta t Da = Da^* \exp(E_a(\frac{1}{T^*} - \frac{1}{\tilde{T}_i^{n+1}})).$$

Subsequently, we choose $T^* \in [T_u, T_b]$ such that $Da^* = 1$, i.e.

$$T^* = \min(T_b, \max(\frac{T_{vN} E_a}{E_a + T_{vN} \ln(\Delta t Da)}, T_u)), \tag{8.6}$$

where we assume $\Delta t Da \neq \exp(-E_a/T_{vN})$. If $\tilde{T}_i^{n+1} < T^*$, then $r\Delta t Da$ is small and $Y_i^{n+1} \approx \tilde{Y}_i^{n+1}$. However, if $\tilde{T}_i^{n+1} \geq T^*$, then $r\Delta t Da$ is large and $Y_i^{n+1} \ll \tilde{Y}_i^{n+1}$. Hence, T^* may be interpreted as the effective "ignition temperature" of the chemical reaction.

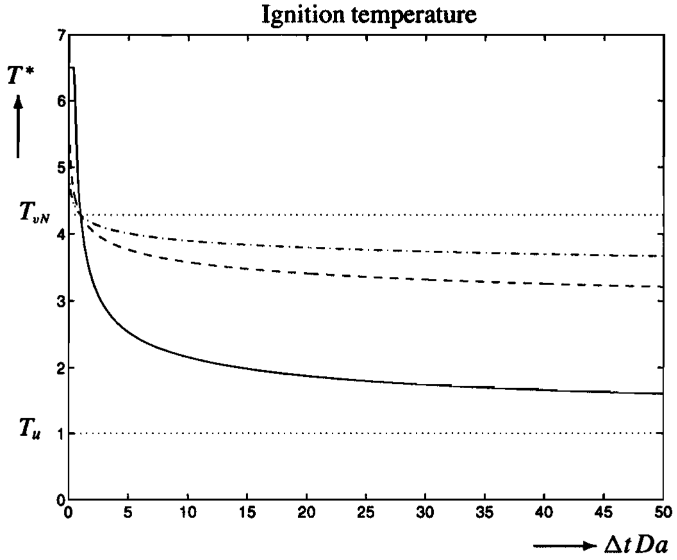

Figure 8.8. The "ignition temperature" T^* (8.6) with $T_{vN} = 4.2838$, $T_u = 1.0$, $T_b = 6.4997$, $E_a = 10$ (solid line), $E_a = 50$ (dashed line) and $E_a = 100$ (dashed-dotted line).

In Figure 8.8 we have drawn T^* as a function of $\Delta t Da$ for several values of E_a. If E_a is small (solid line) and $\Delta t Da$ is large, then $T^* \approx T_u$. On the other hand, if E_a is large (dashed-dotted line) and $\Delta t Da$ is small, then $T^* \approx T_b$ (see Figure 8.8). Therefore, if we decrease Δt (and thus $\Delta t Da$), then (8.6) implies that T^* increases (see Figure 8.8). Hence, if we identify T^* with T_{ign}, then for Δt sufficiently small, we expect that (8.5) is satisfied and therefore, the wrong wave speeds will not occur. This illustrated by the following example.

Example 8.5. First, we consider the strong detonation of Example 8.1. Table 8.6 shows that the correct wave speeds are obtained if $T_{ign} \geq 3.0$. In other words, we like to choose Δt such that $T^* \geq 3.0$. Subsequently, since $E_a = 10$ and $T_{vN} = 4.2838$, (8.6) implies that $\Delta t \leq 2.7155/Da$. In Figure 8.1 and Figure 8.2 we have chosen $Da = 66.201$, and thus $\Delta t \leq 0.0410$. As the results clearly illustrate this time step restriction is fulfilled in Figure 8.1. On the other hand, in Figure 8.2 the time step is too large, as we expected.

Secondly, we consider the strong detonation of Example 8.2. Numerical experiments of method (8.3) show that we obtain the correct wave speed, if $T_{ign} \geq 2.9$. Using, $E_a = 150$ and $T_{vN} = 7.8801$, we obtain the time step restriction $\Delta t \leq 1.6 \cdot 10^{14}/Da$. Clearly, this is not very restrictive except for very large Damköhler number. In Figure 8.5 $Da = 6.3293 \cdot 10^3$ and $\Delta t = 0.05$. As expected for these values the time step restriction is satisfied. ☐

In this section we showed that (8.5) can be used to explain the influence of $\Delta t\, Da$, E_a and T_{ign} on the numerical solution. If we identify T^* with T_{ign}, then for low activation energies and large Damköhler numbers we have to take very small time steps to satisfy (8.5). Moreover, we emphasize that these low ignition temperatures are not very realistic. For low T_{ign}, we have to decrease the time step to increase T^* and to be close to the real physics. Obviously this is not very satisfying.

Method (8.3) captures the detonation wave correctly and produce satisfying results (even for stiff combustion chemistry), provided the correct ignition temperature is chosen. In order to obtain second order accuracy in smooth parts of the flow, we develop a high resolution method in the next section.

8.3. A HIGH RESOLUTION METHOD

In this section we present a high resolution method for the reactive Euler equations (2.26). The numerical method is based on the second order splitting method of Strang (6.17). Let the numerical solution at time level t^n be given. The numerical solution at the next time level $t^{n+1} = t^n + \Delta t$ is computed in three steps (corresponding to $\mathcal{L}^c_{\Delta t/2}$, $\mathcal{L}^f_{\Delta t}$ and $\mathcal{L}^c_{\Delta t/2}$).

Step 1. In the first step we neglect convection (i.e. $\partial/\partial x = 0$ in (2.26)). As for the first order method in Section 8.1, we have to solve the ODE (8.2a) for all i with initial data $Y(x_i, t^*) = Y_i^n$. Again, we assume a constant temperature (i.e. we replace $T(Y)$ by T_i^n in (8.2a)) and integrate the resulting linear differential equation over $[t^*, t^*+\Delta t/2]$ exactly. Denote the result by $Y_i^{n+1/2}$ and define $\rho_i^{n+1/2} := \rho_i^n$, $u_i^{n+1/2} := u_i^n$ and $E_i^{n+1/2} := E_i^n$.

Step 2. In the second step we assume that no reaction takes place and solve (2.26) with $w = 0$ (the nonreactive Euler equations). We use initial data $\mathbf{U}_i^{n+1/2}$ as results from step 1 and compute the numerical solution after a time step Δt of the nonreactive Euler equations by the slope limiter method of Example 5.18. The final result is denoted by $\tilde{\mathbf{U}}_i^{n+1/2}$.

Step 3. Finally, in the last step we obtain the numerical solution $\mathbf{U}^{n+1}$ at time level t^{n+1} by applying step 1 again (i.e. no convection) with initial data $\tilde{\mathbf{U}}_i^{n+1/2}$.

Using the notation of Section 6.3, $\mathcal{L}^f_{\Delta t}$ is given by the slope limiter method and $\mathcal{L}^c_{\Delta t/2} = \mathcal{S}^c_{\Delta t/2}$ is the exact solution operator for problem (8.2a) with constant temperature. As usual, the time step Δt is restricted by the CFL stability condition.

Next results are presented for the method above. In the first example we show that for low ignition temperatures also with the slope limiter method nonphysical wave speeds may occur for large mesh widths.

Example 8.6. In this example we consider the strong detonation of Example 8.1. Recall that the detonation wave propagates with speed $s = 4.9093$. We use the parameter values of Figure 8.2 and apply the slope limiter method described above with Roe's superbee limiter (5.23). In Figure 8.9 the numerical results are compared with the exact solution.

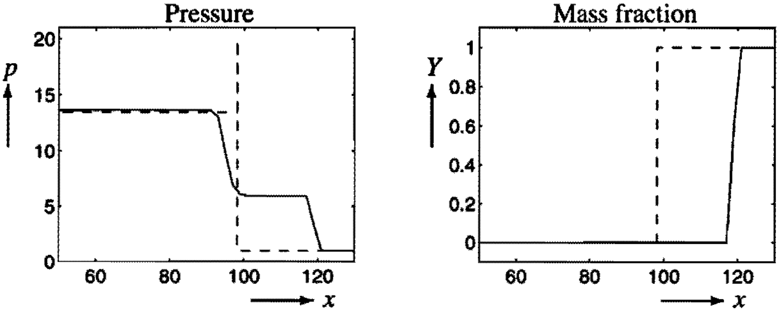

Figure 8.9. Numerical results for the high resolution method (with $\Delta t = 0.2$ and $\Delta x = 2.0$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 20$ of a strong detonation with $E_a = 10$, $Q = 10$, $f = 1.1$, $\gamma = 1.4$, $Da = 66.201$ ($L_{1/2} = 5 \cdot 10^{-3}\Delta x$) and $T_{ign} = 1.01$.

As in Figure 8.2, the numerical solution is totally wrong. Although the superbee limiter suppresses the numerical diffusion, the temperature will increase ahead of the detonation wave and a fake detonation wave is started. The fact that we are using a high resolution method is clearly illustrated by the sharp resolution of the ordinary shock wave (compared with the resolution in Figure 8.2).

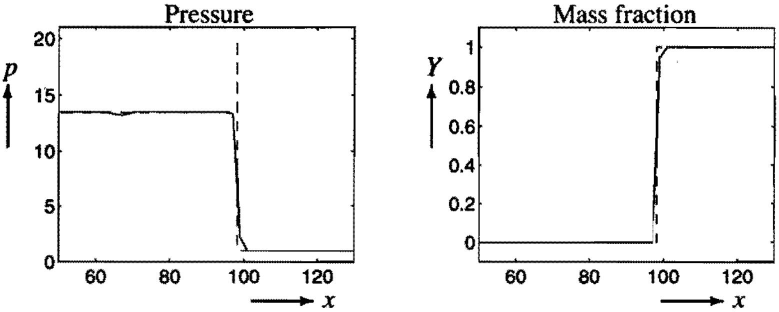

Figure 8.10. Numerical results for the high resolution method (with $\Delta t = 0.2$ and $\Delta x = 2.0$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 20$ of a strong detonation with $E_a = 10$, $Q = 10$, $f = 1.1$, $\gamma = 1.4$, $Da = 66.201$ ($L_{1/2} = 5 \cdot 10^{-3}\Delta x$) and $T_{ign} = T_{vN} = 4.2838$.

For a physically realistic ignition temperature $T_{ign} = T_{vN} = 4.2838$, the numerical solution is correct (see Figure 8.10). Comparing the results with those in Figure 8.3 we observe a slightly better representation of the detonation wave. ☐

Example 8.7. As the last example of this section we consider the CJ detonation of Example 2.1. We apply the high resolution method with Roe's superbee limiter, using the parameter values of Figure 8.7.

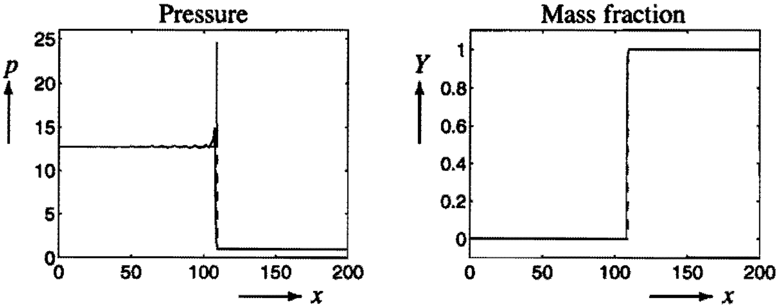

Figure 8.11. Numerical results for the high resolution method (with $\Delta t = 0.1$ and $\Delta x = 1.0$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 20$ of a CJ detonation with $E_a = 14$, $Q = 14$, $f = 1.0$, $\gamma = 1.4$, $Da = 6.4880 \cdot 10^4$ ($L_{1/2} = 10^{-5}\Delta x$) and $T_{ign} = 4.0$.

The results in Figure 8.11 clearly show that (for $T_{ign} = 4.0$) the numerical solution approximates the physically correct detonation wave. As expected we observe a better resolution of the detonation wave (compared with Figure 8.7). Furthermore, the pressure peak is resolved slightly better. However, for a good resolution of the pressure peak one needs a finer grid or a front tracking method (see Section 8.4). Finally, also for the high resolution method we observe small disturbances behind the pressure jump caused by the splitting method (as explained in Figure 8.4). ☐

Hence, the global behaviour of the first order method (8.3) and the high resolution method described in this section is the same. For both methods nonphysical wave speeds may occur for low ignition temperatures. The only real distinction is that the high resolution method represents the possible discontinuities sharper. However, in general this improvement is very small for detonation waves (with stiff combustion chemistry). If $T > T_{ign}$ and the reaction is fast enough, the gas is completely burnt in the next time step and $\mathbf{u}$ is shifted to the equilibrium state $\mathbf{u}_b$. So, often the first order method represents the detonation wave very well. Naturally, nonreactive discontinuities are represented much better (sharper) by the high resolution method (as is shown in Chapter 5).

In Section 1.3 we have mentioned that the main goal of this thesis is to develop a numerical method with the following properties

- at least second order accuracy in smooth parts of the flow;
- a sharp resolution of the discontinuities without excessive smearing or oscillations;
- numerically stable with the usual time step and mesh width restrictions;
- approximating the physically relevant weak solution.

The theory and the results in Chapter 7 and Chapter 8 show that the last property is the most interesting one. However, the high resolution method presented in this section has all the desired properties, provided we use a correct ignition temperature.

In this chapter we derived a criterion on the ignition temperature which guarantees that the numerical solution approximates the physically relevant weak solution. Furthermore, this criterion is used to study the influence of $\Delta t\,Da$ and E_a on the numerical wave speed. We believe that the analysis presented in Chapter 7 and Chapter 8 gives a better insight in the behaviour of detonation capturing methods for the reactive Euler equations.

For the sake of completeness we make some remarks on front tracking methods in the next section.

8.4. SOME REMARKS ON FRONT TRACKING

In Section 8.2 we have shown that a correct ignition temperature will exclude nonphysical weak solutions. Another way to enforce the correct wave speed (even for low ignition temperatures) is to approximate the solution by a front tracking method. The basic idea behind any front tracking method is to follow the front position in some way. Often, one adds a new mesh point to the computational mesh, which follows the front position. Another approach is to consider the front as a level set of the scalar field $G(x, t)$ [64, 77, 86] and update this level set as time evolves. If we model the detonation wave as a gas dynamic discontinuity, we may consider to track the detonation wave instead of the leading shock wave [77]. However, we restrict ourselves to cases where the leading shock wave of the detonation is tracked.

If we are capable to compute the shock position accurately, the front tracking method has many advantages compared to the previous methods. For instance, the leading shock is spread over at most two mesh cells. Furthermore, if we like to describe the chemical reaction in detail and we know the front position, then we also know where the spatial grid should be refined. Here we consider the problem of nonphysical wave speeds only. The previous methods automatically capture possible discontinuities. The jumps in the solution are replaced by sharp transitions over a few mesh cells, using a sufficient amount of numerical diffusion. In the front tracking method, all discontinuities are captured indeed except for the leading ordinary shock of the detonation wave, which is tracked explicitly.

Suppose the detonation wave propagates into an undisturbed, unburnt, uniform gas. Since the position of the leading shock is known, the numerical solution is not smoothed around the shock and equal to $\mathbf{u}_u$ ahead of the detonation wave. So, in contrast to the previous methods, even for low ignition temperatures, no artificial reaction is started. Hence, the front tracking method will always produce the physically correct detonation wave, provided the front position is computed accurately.

The major disadvantage of tracking methods is that they are difficulty to generalize to higher space dimensions. For instance, in one space dimension it is sufficient to add a point to the grid, which follows the front position. However, in two dimensions the front becomes a curve. Representing this curve on a two-dimensional spatial grid and updating its position as time evolves, may be very complicated [13, 76, 87]. The level set approach is much easier in higher space dimensions and allows the representation of complex front topologies [64, 77]. However in many applications a front tracking method is too complicated and methods on fixed grids are to be preferred.

In [10] a method is developed for detonation waves, which uses the front tracking method of Chern and Colella [13] in combination with the piecewise parabolic method. Furthermore, the stability properties of ZND solutions for detonation waves are studied. In [76] a front tracking method is introduced, based on Roe's linearization. Also in [76] the front tracking method is successfully applied to the reactive Euler equations, to study stability properties of ZND solutions. Finally, we refer to [64, 77, 78] where a new algorithm for the tracking of detonation waves is presented, which uses the level set approach.

For the sake of completeness we present some numerical results for a front tracking method. In the present context we use the second order splitting method (6.17) as described in Section 8.3. Only the second step is changed. Here, we solve the nonreactive Euler equations in the second step by the front tracking method described in [13] in combination with the slope limiter method (see Example 5.18). The front tracking method in [13] has the important property to be globally conservative and the time step Δt is restricted by the usual CFL stability condition.

Example 8.8. In this example we consider again the strong detonation of Example 8.1. In Figure 8.12 the numerical solution is compared with the exact solution. The numerical solution approximates the exact ZND profile very well. Comparing the results with those in Figure 8.1, the resolution of the pressure peak is much better. This is caused by the tracking of the leading shock wave.

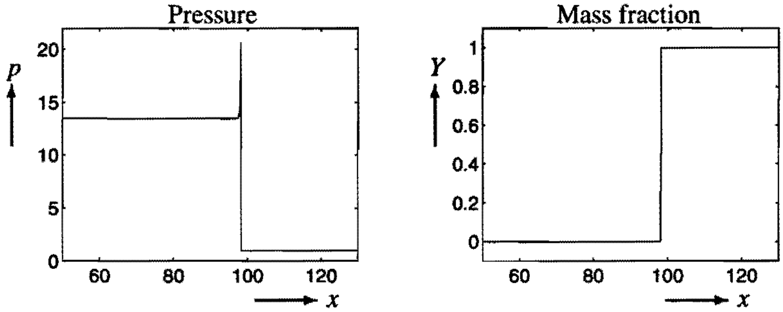

Figure 8.12. Numerical results for the front tracking method (with $\Delta t = 0.01$ and $\Delta x = 0.1$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 20$ of a strong detonation with $E_a = 10$, $Q = 10$, $f = 1.1$, $\gamma = 1.4$, $Da = 66.201$ ($L_{1/2} = 0.1\Delta x$) and $T_{ign} = 1.01$.

Figure 8.13 clearly illustrates the strength of the front tracking method. We use a low ignition temperature ($T_{ign} = 1.01 \approx T_u$), a low activation energy ($E_a = 10$) and a large mesh width ($\Delta x = 200 L_{1/2}$). Contrary to the results of Figure 8.2 (and Figure 8.9), the numerical solution in Figure 8.13 is correct. The numerical reaction zone is larger than the exact reaction zone, since the fast reaction cannot be resolved very well on the large mesh. □

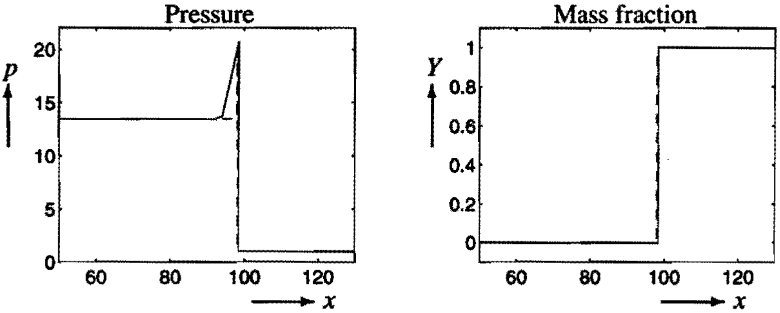

Figure 8.13. Numerical results for the front tracking method (with $\Delta t = 0.2$ and $\Delta x = 2.0$); numerical solution (solid line) and exact solution (dashed line) of (2.26) at $t = 20$ of a strong detonation with $E_a = 10$, $Q = 10$, $f = 1.1$, $\gamma = 1.4$, $Da = 66.201$ ($L_{1/2} = 5 \cdot 10^{-3}\Delta x$) and $T_{ign} = 1.01$.

For low ignition temperatures the front tracking method behaves better than the previous methods. However, in this thesis we have shown that for realistic ignition temperatures all methods produce similar wave speeds. Sometimes an accurate representation of the leading front (especially the von Neumann spike) is necessary. For instance, the front interface plays an important role in the stability mechanism of the ZND solution. Excessive numerical diffusion around the front can modify the stability properties of the detonation sometimes to the extent of transforming a physically unstable detonation into a numerically stable one. For this kind of problems a method which tracks the front explicitly seems unavoidable [10, 76, 78]. However, in general the major flaw of, namely the generation of nonphysical detonation waves, can be overcome by using a correct ignition temperature. Hence, we have shown that for approximating the correct weak solution, one does not need complex methods like front tracking.

NOMENCLATURE

symbol	description
A	frequency factor (s^{-1})
B	constant in frequency factor ($s^{-1}K^{-\alpha}$)
c	speed of sound (m/s)
c_p	specific heat at constant pressure of the gas mixture (J/(kg K))
$c_{p,i}$	specific heat at constant pressure for species $\mathcal{M}_i$ (J/(kg K))
c_v	specific heat at constant volume of the gas mixture (J/(Kg K))
$c_{v,i}$	specific heat at constant volume for species $\mathcal{M}_i$ (J/(kg K))
D	constant binary diffusion coefficient for all pairs of species (m^2/s)
D_{ij}	binary diffusion coefficient for species $\mathcal{M}_i$ and $\mathcal{M}_j$ (m^2/s)
Da	Damköhler number
e	specific internal energy per unit mass for the gas mixture (J/kg)
E	specific total energy per unit mass for the gas mixture (J/kg)
E_a	activation energy (J/mol)
f	degree of overdrive
f_i	external force per unit mass on species $\mathcal{M}_i$ (N/kg)
h	specific enthalpy of the gas mixture (J/kg)
h_i	specific enthalpy of species $\mathcal{M}_i$ (J/kg)
h_i^0	standard heat of formation per unit mass for species $\mathcal{M}_i$ at temperature T_0 (J/kg)
H	stagnation enthalpy of the gas mixture (J/kg)
k	specific rate constant (s^{-1})
$L_{1/2}$	half-reaction length (m)
m	mass flux (kg/(m^2 s))
p	hydrostatic pressure (N/m^2)
q	heat flux for the gas mixture (J/(m^2 s))
Q	heat release of the chemical reaction per unit mass (J/kg)
R	universal gas constant (J/(mol K))
s	propagation speed of the combustion wave (m/s)

symbol	description
S	specific entropy of the gas mixture (J/(kg K))
T	absolute temperature of the gas mixture (K)
T_0	fixed reference temperature (K)
T_{ign}	ignition temperature of the chemical reaction (K)
u	mass weighted average velocity of the gas mixture (m/s)
u_i	flow velocity of species $\mathcal{M}_i$ (m/s)
U_i	diffusion velocity of species $\mathcal{M}_i$ (m/s)
v	specific volume of the gas mixture (m^3/kg)
w_i	reaction rate of species $\mathcal{M}_i$ (kg/(m^3 s))
W	average molecular weight of the gas mixture (kg/mol)
W_i	molecular weight of species $\mathcal{M}_i$ (kg/mol)
X_i	mole fraction of species $\mathcal{M}_i$
Y_i	mass fraction of species $\mathcal{M}_i$
α	constant determining the temperature dependence of the frequency factor
λ	thermal conductivity of the gas mixture (J/(m s K))
μ	viscosity coefficient (kg/(m s))
γ	specific heat ratio
ρ	mass density of the gas mixture (kg/m^3)
ρ_i	mass density of species $\mathcal{M}_i$ (kg/m^3)
σ	stress of the gas mixture (N/m^2)
τ	viscous stress of the gas mixture (N/m^2)

BIBLIOGRAPHY

[1] W.F. Ames, Numerical Methods for Partial Differential Equations, Academic Press, New York (1977).

[2] J.B. Bdzil and R. Klein, *Weakly nonlinear dynamics of near-CJ detonation waves*, in Proc. of the 2nd ICASE/LaRC workshop on Combustion, Newport News, VA (1992).

[3] M. Ben-Artzi, *The generalized Riemann problem for reactive flows*, J. Comput. Phys. **55** (1984), pp. 1-32.

[4] M. Ben-Artzi and J. Falcovitz, *A second-order Godunov-type scheme for compressible fluid dynamics*, J. Comput. Phys. **81** (1989), pp. 70-101.

[5] A.C. Berkenbosch, E.F. Kaasschieter and J.H.M. ten Thije Boonkkamp, *Finite-difference methods for one-dimensional hyperbolic conservation laws*, Num. Meth. for Part. Diff. Eq. **10** (1994), pp. 225-269.

[6] A.C. Berkenbosch, E.F. Kaasschieter and J.H.M. ten Thije Boonkkamp, *The numerical wave speed for one-dimensional scalar hyperbolic conservation laws with source terms*, RANA 94-01 of the Department of Mathematics and Computing Science, Eindhoven University of Technology (1994).

[7] A.C. Berkenbosch, E.F. Kaasschieter and J.H.M. ten Thije Boonkkamp, *The numerical computation of one-dimensional detonation waves*, RANA 94-21 of the Department of Mathematics and Computing Science, Eindhoven University of Technology, submitted (1994).

[8] A.C. Berkenbosch, E.F. Kaasschieter, J.H.M. ten Thije Boonkkamp and R. Klein, *Detonation capturing for stiff combustion chemistry*, RANA 95-06 of the Department of Mathematics and Computing Science, Eindhoven University of Technology, submitted (1995).

[9] J.P. Boris and E.S. Oran, Numerical Simulation of Reactive Flow, Elsevier, New York (1987).

[10] A. Bourlioux, Numerical Study of Unstable Detonations, Ph.D. Thesis, Princeton University, Princeton (1991).

[11] A. Bourlioux, A. Majda and V. Roytburd, *Theoretical and numerical structure for unstable one-dimensional detonations*, SIAM J. Appl. Math. **51** (1991), pp. 303-343.

[12] A. Chalabi, *Stable upwind schemes for hyperbolic conservation laws with source terms*, IMA J. of Num. Anal. **12** (1992), pp. 217-241.

[13] I. Chern and P. Colella, *A conservative front tracking method for hyperbolic conservation laws*, Preprint UCRL-97200, Lawrence Livermore National Laboratory (1987).

[14] A.J. Chorin, *Random choice methods with applications to reacting gas flow*, J. Comput. Phys. **25** (1977), pp. 253-272.

[15] J.F. Clarke, S. Karni, J.J. Quirk, P.L. Roe, L.G. Simmonds and E.F. Toro, *Numerical Computation of two-dimensional unsteady detonation waves in high energy solids*, J. Comput. Phys. **106** (1993), pp. 215-233.

[16] P. Colella, A. Majda and V. Roytburd, *Theoretical and numerical structure for reacting shock waves*, SIAM J. Sci. Stat. Comput. **7** (1986), pp. 1059-1080.

[17] P. Colella and P.R. Woodward, *The piecewise parabolic method (PPM) for gas-dynamical simulations*, J. Comput. Phys. **54** (1984), pp. 174-201.

[18] R. Courant and K.O. Friedrichs, Supersonic Flow and Shock Waves, Wiley, New York (1948).

[19] M.G. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp. **34** (1980), pp. 1-21.

[20] R.J. DiPerna, *Finite difference schemes for conservation laws*, Comm. Pure Appl. Math. **35** (1982), pp. 379-450.

[21] J.W. Dold, *Emergence of a detonation within a reacting medium*, in: M. Onofri and A. Tesei (eds), Fluid Dynamical Aspects of Combustion Theory Vol. 223, Longman, Harlow (1991), pp. 161-183.

[22] D.H. Edwards, G.T. Williams and J.C. Breeze, *Pressure and velocity measurements on detonation waves in hydrogen-oxygen mixtures*, J. of Fluid Mech. **6** (1959), pp. 497-517.

[23] B. Einfeldt, *On Godunove-type methods for gas dynamics*, SIAM J. Numer. Anal. **25** (1988), pp. 294-318.

[24] B. Engquist and S. Osher, *One-sided difference approximations for nonlinear conservation laws*, Math. Comp. **36** (1981), pp. 321-351.

[25] W. Fickett, *Detonation in miniature*, American J. of Phys. **47** (1980), pp. 1050-1059.

[26] W. Fickett and W.C. Davis, Detonation, University of California Press, Berkeley (1979).

[27] P.R. Garabedian, Partial Differential Equations, J. Wiley, New York (1964).

[28] E. Godlewski and P.A. Raviart, Hyperbolic Systems of Conservation Laws, Mathematiques & Applications, SMAI, Ellipses, Paris (1991).

[29] J. Goettgens, F. Mauss and N. Peters, *Analytic Approximations of Burning Velocities and Flame Thicknesses of Lean Hydrogen, Methane, Ethylene, Acethylene and Propane Flames*, in Proc. of the 24th Intl. Symp. on Combustion, The Combustion Institute, Pittsburgh (1992), pp. 129-135.

[30] J.B. Goodman and R.J. LeVeque, *A geometric approach to high resolution TVD schemes*, SIAM J. Numer. Anal. **25** (1988), pp. 268-284.

[31] D.F. Griffiths, A.M. Stuart and H.C. Yee, *Numerical wave propagation in an advection equation with a nonlinear source term*, SIAM J. Numer. Anal. **29** (1992), pp. 1244-1260.

[32] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys. **49** (1983), pp. 357-393.

[33] A. Harten, *On a class of high resolution total-variation-stable finite-difference schemes*, SIAM J. Numer. Anal. **21** (1984), pp. 1-23.

[34] A. Harten, B. Engquist, S. Osher and S. Chakravarthy, *Uniformly high order accurate essentially nonoscillatory schemes, III*, J. Comput. Phys. **71** (1987), pp. 231-303.

[35] A. Harten, J.M. Hyman and P.D. Lax, *On finite-difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math. **29** (1976), pp. 297-322.

[36] A. Harten, P.D. Lax and B. van Leer, *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Review **25** (1983), pp. 35-61.

[37] A. Harten and S. Osher, *Uniformly high order accurate nonoscillatory schemes. I*, SIAM J. Numer. Anal. **24** (1987), pp. 279-309.

[38] L. He and P. Clavin, *Stability and nonlinear dynamics of one-dimensional overdriven detonations in gases*, J. Fluid Mech. **277** (1994), pp. 227-248.

[39] C. Hirsch, Numerical Computation of Internal and External Flows, volume 1: Fundamentals of Numerical Discretization, J. Wiley, Chichester (1988).

[40] C. Hirsch, Numerical Computation of Internal and External Flows, volume 2: Computational Methods for Inviscid and Viscous Flows, J. Wiley, Chichester (1990).

[41] A. Jeffrey and T. Taniuti, Non-linear Wave Propagation, Academic Press, New York (1964).

[42] C. Johnson, A. Szepessy and P. Hansbo, *On the convergence of shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws*, Math. Comp. **54** (1990), pp. 107-129.

[43] R. Klein, *Curved detonations in explosive gas mixtures with high temperature sensitivity*, Intl. Conf. on Combustion, 80th Birthday Memorial for Ya.B. Zeld'dovic, Moskow (1994).

[44] R. Klein, *Analysis of accelerating detonation using large activation energy asymptotics*, 2nd Workshop on Microscopic and Macroscopic Approaches in Detonation Theory, St. Malo (1994).

[45] R. Klein and D.S. Stewart, *The relation between curvature, rate state dependence and detonation velocity*, SIAM J. Appl. Math. **53** (1993), pp. 1401-1435.

[46] H.O. Kreiss, *Difference approximations for initial-boundary value problems for hyperbolic differential equations*, in Numerical Solutions of Nonlinear Partial Differential Equations, D. Greenspan (ed.), J. Wiley, New York, 1966, pp. 140-166.

[47] S.N. Krushkov, *First order quasi-linear equations in several independent variables*, Math. USSR Sb. **10** (1970), pp. 217-243.

[48] B. Larrouturou, *How to preserve the mass fractions positivity when computing compressible multi-component flows*, J. Comput. Phys. **95** (1991), pp. 59-84.

[49] P.D. Lax, Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves, SIAM Regional Conference Series in Applied Mathematics, volume 11, SIAM, Philadelphia (1973).

[50] P.D. Lax and B. Wendroff, *Systems of conservation laws*, Comm. Pure Appl. Math. **13** (1960), pp. 217-237.

[51] B. van Leer, *Towards the ultimate conservative difference scheme II. Monotonicity and conservation combined in a second order scheme*, J. Comput. Phys. **14** (1974), pp. 361-370.

[52] B. van Leer, *Towards the ultimate conservative difference scheme V. A second-order sequel to Godunov's method*, J. Comput. Phys. **32** (1979), pp. 101-136.

[53] B. van Leer, *On the relation between the upwind-differencing schemes of Godunov, Engquist-Osher and Roe*, SIAM J. Sci. Stat. Comput. **5** (1984), pp. 1-20.

[54] R.J. LeVeque, *Convergence of a large time step generalization of Godunov's method for conservation laws*, Comm. Pure Appl. Math. **37** (1984), pp. 463-477.

[55] R.J. LeVeque, *A large time step generalization of Godunov's method for systems of conservation laws*, SIAM J. Numer. Anal. **22** (1985), pp. 1051-1073.

[56] R.J. LeVeque, Numerical Methods for Conservation Laws, Lectures in Mathematics, Birkhäuser Verlag, Basel (1990).

[57] R.J. LeVeque and J. Oliger, *Numerical methods based on additive splittings for hyperbolic partial differential equations*, Math. Comp. **40** (1983), pp. 469-497.

[58] R.J. LeVeque and H.C. Yee, *A study of numerical methods for hyperbolic conservation laws with stiff source terms*, J. Comput. Phys. **86** (1990), pp. 187-210.

[59] H.W. Liepmann and A. Roshko, Elements of Gas dynamics, J. Wiley, London (1957).

[60] A. Majda, *A qualitative model for dynamic combustion*, SIAM J. Appl. Math. **41** (1981), pp. 70-93.

[61] A. Majda and S. Osher, *Numerical viscosity and the entropy condition*, Comm. Pure Appl. Math. **32** (1979), pp. 797-838.

[62] A. Majda and V. Roytburd, *Numerical study of the mechanisms for initiation of reacting shock waves*, SIAM J. Sci. Stat. Comput. **11** (1990), pp. 950-974.

[63] S.B. Margolis, *Time-dependent solution of a premixed laminar flame*, J. Comput. Phys. **27** (1978), pp. 410-427.

[64] V. Moser, F. Zhang and P. Thibault, *Detonation front tracking via in-cell reconstruction*, in Proc. of the 3rd Conf. of the CFD Soc. of Canada, Banff, Alberta (1995).

[65] W. Mulder, S. Osher and J.A. Sethian, *Computing interface motion in compressible gas dynamics*, J. Comput. Phys. **100** (1992), pp. 209-228.

[66] M.A. Nettleton, Gaseous Detonations, their Nature, Effects and Control, Chapman and Hall, New York (1987).

[67] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM J. Numer. Anal. **21** (1984), pp. 217-235.

[68] S. Osher and S. Chakravarthy, *High resolution schemes and the entropy condition*, SIAM J. Numer. Anal. **21** (1984), pp. 955-984.

[69] S. Osher and F. Solomon, *Upwind difference schemes for hyperbolic systems of conservation laws*, Math. Comp. **38** (1982), pp. 339-374.

[70] S. Osher and E. Tadmor, *On the convergence of difference approximations to scalar conservation laws*, Math. Comp. **50** (1988), pp. 19-51.

[71] G. Paczko and R. Klein, *Reduced chemical kinetics schemes for hydrogen-air-steam detonation simulations*, preliminary text, submitted for presentation at 14th ICODERS, Portugal (1993)

[72] R. Pember, *Numerical methods for hyperbolic conservation laws with stiff relaxation I. Spurious solutions*, SIAM J. Appl. Math. **53** (1993), pp. 1293-1330.

[73] R.D. Richtmeyer and K.W. Morton, Difference Methods for Initial-value Problems, Wiley-Interscience, New York (1967).

[74] P.L. Roe, *Approximate Riemann solvers, parameter vectors and difference schemes*, J. Comput. Phys. **43** (1981), pp. 357-372.

[75] V. Roytburd, *On detonation instability*, in: M. Onofri and A. Tesei (eds), Fluid Dynamical Aspects of Combustion Theory Vol. 223, Longman, Harlow (1991), pp. 184-195.

[76] K.M. Shyue, Front Tracking Methods Based on Wave Propagation, Ph.D. Thesis, University of Washington, Seattle (1993).

[77] V. Smiljanovski and R. Klein, *Flame front tracking via in-cell reconstruction*, submitted to Proc. of the 5th Intl. Conf. on Hyperbolic Problems, USB, Stony Brook (1994).

[78] V. Smiljanovski and R. Klein, *Simulation of gasdynamic flame instability and DDT using in-cell reconstruction*, preliminary text, submitted for presentation at the 6th Intl. Conf. on Num. Fluid Mech., Lake Tahoe, Nevada (1995).

[79] J. Smoller, Shock Waves and Reaction-Diffusion Equations, Grundlehren der mathematischen Wissenschaften, Vol. 258, Springer-Verlag, New York (1983).

[80] A.C. Sod, *A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws*, J. Comput. Phys. **27** (1978), pp. 1-31.

[81] S.P. Spekreyse, Multigrid Solution of the Steady Euler Equations, volume 46 of CWI tracts, CWI, Amsterdam (1988).

[82] J.L. Steger and R.F. Warming, *Flux vector splitting of the inviscid gas dynamic equations with applications to finite-difference methods*, J. Comput. Phys. **40** (1981), pp. 263-293.

[83] D.S. Stewart and J.B. Bdzil, *The shock dynamics of stable multi-dimensional detonation*, Comb. & Flame **72** (1988), pp. 311-323.

[84] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal. **5** (1968), pp. 506-517.

[85] R.A. Strehlow, Combustion Fundamentals, McGraw-Hill, New York (1984).

[86] M. Sussman, P. Smereka and S. Osher, *A level set approach for computing solutions to incompressible two-phase flow*, J. Comput. Phys. **114** (1994), pp. 146-159.

[87] B.K. Swartz and B. Wendroff, *Aztec: A front tracking code based on Godunov's method*, Appl. Numer. Math. **2** (1986), pp. 385-397.

[88] P.K. Sweby, *High resolution schemes using flux limiters for hyperbolic conservation laws*, SIAM J. Numer. Anal. **21** (1984), pp. 995-1011.

[89] E. Tadmor, *The numerical viscosity of entropy stable schemes for systems of conservation laws I*, Math. Comp. **49** (1987), pp. 91-103.

[90] E. Tadmor, *Convergence of spectral methods for nonlinear conservation laws*, SIAM J. Numer. Anal. **26** (1989), pp. 30-44.

[91] T. Tang and Z.H. Teng, *Error bounds for fractional step methods for conservation laws with source terms*, SIAM J. Numer. Anal. **32** (1995), pp. 110-127.

[92] J.H.M. ten Thije Boonkkamp, The conservation equations for reacting gas flow, EUT Report 93-WSK-01, Eindhoven (1993).

[93] J.P. Vila, *High-order schemes and entropy condition for nonlinear hyperbolic systems of conservation laws*, Math. Comp. **50** (1988), pp. 53-73.

[94] R,F, Warming and B.J. Hyett, *The modified equation approach to the stability and accuracy analysis of finite difference methods*, J. Comput. Phys. **14** (1974), pp. 159-179.

[95] G.B. Whitham, Linear and Nonlinear Waves, J. Wiley, New York (1974).

[96] F.A. Williams, Combustion Theory, The Fundamental Theory of Chemically Reacting Flow Systems, Addison-Wesley, Redwood City (1985).

INDEX

Absolute temperature, 17
Activation energy, 4, 20
Antidiffusive flux, 86
Arrhenius' law, 20

Basic upwind method, 62, 66
Binary diffusion coefficient, 16
Bulk viscosity, 16
Burgers' equation, 38, 41, 72, 76, 94

Caloric equation of state, 17
CFL condition, 63
Characteristic, 39
CJ deflagration wave, 28
CJ detonation wave, 28, 30, 32
CJ model, 3
Cold-boundary difficulty, 20
Combustion wave, 14
Compression shock, 49
Conductivity, thermal, 17
Conservation equations for reacting gas flow, 14
Conservative method, 59
Consistent method, 60
Contact discontinuity, 46, 48
Contracting method, L_1-, 83
Convergence, L_1^{loc}-, 61
Convergent method, 60
Courant number, 62

Damköhler number, 23
Deflagration wave, 28
  Chapman-Jouguet (CJ), 28
  strong, 28
  weak, 28
Degree of overdrive, 33
Density, 14
Detonation wave, 1, 28, 30
  Chapman-Jouguet (CJ), 28, 30, 32
  strong, 28, 30, 32
  weak, 28, 30
Diffusion velocity, 14, 15
Dimensionless variable, 21
Discrete conservation, 65
Discretization error
  global, 60
  local, 60
Dufour effect, 17

E-method, 68, 84
Energy
  activation, 4, 20
  internal, 18
  kinetic, 18
  total, 14, 18
Enthalpy
  specific, 17, 18
  stagnation, 19
Entropy, 21

Entropy condition, 7, 43
  Lax, 50
Entropy consistent method, 68
Entropy flux, 42
  numerical, 68
Entropy function, 42
Entropy solution, 43
Equation of state, 18, 22
  caloric, 17
Error
  global, 108
  local, 107
Euler equations
  nonreactive, 45, 53, 76, 94
  reactive, 20, 22, 39, 142
Expansion shock, 49

Fick's law of mass diffusion, 16
Flux limiter method, 84, 87
Force, external, 14, 16
Frequency factor, 20
Front tracking method, 155

Genuinely nonlinear eigenvector, 45
Godunov method, 70, 71
Gravity, 16

Half-reaction length, 33
Heat flux, 14, 17
Heat of formation, standard, 18
Heat release, 19
High resolution method, 9, 10, 84, 89, 136, 152
Hugoniot curve, 26, 31
Hugoniot function, 26
Hyperbolic conservation law, 5, 38

Ideal gas, 18
Ignition delay time, 4
Ignition temperature, 20
Internal energy, 18

Jacobian matrix, 38, 102
Jouguet's rule, 30
Jump condition, 41

Kinetic energy, 18

Law of mass action, 20
Lax entropy condition, 50
Lax-Wendroff method, 64, 66
Lax-Wendroff theorem, 61
Limiter, 86
Linearly degenerate eigenvector, 45

Mass flux, 24
Mass fraction, 14, 15
Minmod limiter, 88
Modified equation, 65
Mole fraction, 16
Molecular mass, 16
Monotone method, 83
Monotonicity preserving method, 82

Nonreactive Euler equations, 45, 53, 76, 94
Nonreactive Rankine-Hugoniot equations, 41, 54
Numerical diffusion, 67
Numerical dispersion, 67
Numerical entropy flux, 68
Numerical flux function, 59
Numerical wave speed, 124, 148

Pressure, hydrostatic, 17

Rankine-Hugoniot equations
  nonreactive, 41, 54
  reactive, 25
Rarefaction wave, 46, 47
Rate constant, 20
Rayleigh line, 25, 32
Reaction rate, 14, 20
Reactive Euler equations, 20, 22, 39, 142
Reactive Rankine-Hugoniot equations, 25
Riemann invariant, 52
Riemann problem, 45, 50
Roe's method, 74, 75

Shock capturing method, 9
Shock tube problem, 53, 76, 94
Shock wave, 46, 49

Simplified detonation model, 114
Slope limiter method, 89, 93
Solution operator
    approximate, 107
    exact, 99, 100
Soret effect, 16
Specific heat at constant pressure, 18, 19
Specific heat at constant volume, 19
Specific heat ratio, 20
Speed of sound, frozen, 21
Splitting error
    global, 100
    local, 100
Stable method, 62
    L_∞-, 81
    TV-, 81
Stable splitting method, 101
Stagnation enthalpy, 19
Stephan-Maxwell equations, 16
Strang's time splitting method, 104, 109
Stress, 14, 17
    viscous, 17
Strong deflagration wave, 28
Strong detonation wave, 28, 30, 32
Superbee limiter, 88

Temperature
    absolute, 17
    ignition, 20
Thermodynamic identity, 18, 20, 22
Time splitting method, 8, 98, 100
    Strang, 104, 109
Total energy, 14, 18
Total variation, 80
TVD method, 82, 85

Universal gas constant, 18

Van Leer limiter, 88
Velocity
    diffusion, 14, 15
    mass-weighted average, 14
Viscosity, 17
    bulk, 16
    coefficient, 42
Viscous stress, 17
Volume, specific, 26
Von Neumann spike, 31

Weak deflagration wave, 28
Weak detonation wave, 28, 30
Weak solution, 6, 40

ZND model, 3, 30

SAMENVATTING

Het doel van dit proefschrift is de ontwikkeling van een betrouwbare numerieke methode voor het simuleren van eendimensionale detonatiegolven. Detonatiegolven zijn zeer snel voortbewegende verbrandingsgolven (vlammen) waarin explosieve reacties optreden en grote hoeveelheden energie omgezet worden. Deze eigenschappen maken detonaties gemakkelijk te onderscheiden van andere verbrandingsprocessen.

In veel toepassingen worden detonaties gemodelleerd door de Euler-vergelijkingen met daaraan toegevoegd balansvergelijkingen voor de verschillende stoffen in het gasmengsel. De balansvergelijkingen bevatten brontermen die de chemische reacties beschrijven. Het totale stelsel vergelijkingen wordt de reagerende Euler-vergelijkingen genoemd. De chemische reacties worden beschreven door een Arrhenius-model. Dit model bevat een ontbrandingstemperatuur T_{ign}, zodanig dat voor temperaturen boven T_{ign} de reactiesnelheid groot is en voor temperaturen beneden T_{ign} de reactiesnelheid nul is. In bijna alle praktische toepassingen is de ontbrandingstemperatuur veel hoger dan de temperatuur van het koude gas T_u, omdat bij lage temperaturen de detonatie uitdooft door een tekort aan radicalen.

In dit proefschrift beschrijven we detonatiegolven met het ZND-model (een model ontwikkeld door Zel'dovich, von Neumann en Döring). Het ZND-model neemt aan dat de detonatiegolf bestaat uit een schokgolf (een compressieschok) gevolgd door een reactiezone die allebei met dezelfde snelheid voortbewegen. Dus, door een schokgolf stijgt de temperatuur boven de ontbrandingstemperatuur en wordt er een chemische reactie gestart.

De reagerende Euler-vergelijkingen zijn een stelsel eerste orde hyperbolische behoudswetten met bronterm. In dit proefschrift beschrijven we enkele karakteristieke eigenschappen van hyperbolische behoudswetten. Aangezien detonatiegolven discontinu zijn (ze bevatten een sterke compressieschok), beschouwen we zwakke oplossingen. Een probleem is echter dat deze zwakke oplossingen niet uniek zijn. In het algemeen worden de unieke fysisch relevante zwakke oplossingen gekarakteriseerd door een zogenaamde entropieconditie. Bij het ZND-model bepaalt "Jouguet's rule" de unieke zwakke oplossing. "Jouguet's rule" impliceert dat de stroming in een met de detonatiegolf meebewegend assenstelsel supersonisch is met betrekking tot het niet verbrande gas (voor de detonatie) en subsonisch of sonisch met betrekking tot het verbrande gas (achter de detonatie). De detonatie heet een Chapman-Jouguet (CJ) detonatie bij een sonische uitstroming en een sterke detonatie bij een supersonische uitstroming.

In dit proefschrift lossen we de reagerende Euler-vergelijkingen numeriek op. We doen dit met behulp van een "time splitting method". In de "time splitting method" benaderen we in iedere discrete tijdstap de stroming (de gewone Euler-vergelijkingen) en de chemie apart. Omdat we vooral geïnteresseerd zijn in het globale gedrag gebruiken we maaswijdten en tijdstappen die geschikt zijn om de stroming te beschrijven maar niet de veel snellere chemie. Om de chemie ook goed te beschrijven moeten we het rooster (lokaal) verfijnen. Belangrijk is dat de numerieke detonatie met de goede snelheid voortbeweegt ("detonation capturing"), zonder dat we de locatie van de detonatie expliciet volgen ("detonation tracking").

De stroming lossen we numeriek op met een Godunov-methode. In deze methode wordt de ruimte verdeeld in discrete cellen en veronderstellen we de oplossing constant in iedere cel. Op deze manier wordt een rij Riemann-problemen gedefinieerd die exact (of benaderend) opgelost worden. Deze oplossing wordt vervolgens gebruikt om een benadering op het volgende tijdstip te bepalen. Dit is een eerste orde methode en tweede orde wordt bereikt door de oplossing stuksgewijs lineair te veronderstellen ("slope limiter methods"). We beschrijven de chemie door een gewone differentiaalvergelijking. Deze wordt opgelost door haar te vereenvoudigen en vervolgens exact te integreren.

Door nu alternerend de "chemie-solver" en de "stromings-solver" (de "slope limiter method") toe te passen ("Strang splitting") krijgen we een "high resolution method". Deze methode is tweede orde nauwkeurig voor gladde oplossingen, geeft een scherpe representatie van eventuele discontinuïteiten en is numeriek stabiel onder de gangbare CFL-conditie. De belangrijkste vraag is of deze methode ook altijd de fysisch relevant zwakke oplossing benadert (de sterke of CJ-detonatie).

Bij snelle reacties kan de numerieke methode een niet-fysische oplossing benaderen (de verkeerde zwakke oplossing). Er is in de literatuur dan ook veel aandacht besteed aan het ontwikkelen van nieuwe en vaak complexe methoden om dit probleem op te lossen. In dit proefschrift bestuderen we de invloed van verschillende numerieke en fysische parameters op de numerieke oplossing. Uit onze analyse volgt dat de ontbrandingstemperatuur een cruciale rol speelt. We laten zien dat de niet-fysische oplossingen alleen voorkomen bij een onrealistisch lage T_{ign}. Verder laten we zien dat bij een correcte T_{ign}, onze relatief eenvoudige "detonation capturing method" altijd de fysisch correcte detonatie benadert en dat het niet nodig is om (complexe) "detonation tracking methods" te ontwikkelen.

CURRICULUM VITAE

De schrijver van dit proefschrift is geboren op 6 september 1968 te Leiden. Na het behalen van het diploma Atheneum B aan het Christelijk Lyceum te Alphen aan den Rijn is hij in 1986 wiskunde gaan studeren aan de Rijksuniversiteit te Leiden. Het afstudeerproject ging over contractiviteit van algemene lineaire methoden en is uitgevoerd onder begeleiding van dr. J.F.B.M. Kraaijevanger. In augustus 1991 is hij afgestudeerd in de numerieke wiskunde.

Van september 1989 tot juni 1991 werkte hij als student-assistent bij de afdeling Wiskunde en Informatica van de Rijksuniversiteit Leiden. Van 1 september 1991 tot 1 september 1995 is hij als assistent in opleiding in dienst geweest van de vakgroep analyse, faculteit Wiskunde en Informatica van de Technische Universiteit Eindhoven. In deze functie heeft hij onder leiding van prof.dr. R.M.M. Mattheij het onderzoek verricht dat geleid heeft tot dit proefschrift.

Stellingen

behorende bij het proefschrift

CAPTURING DETONATION WAVES FOR THE REACTIVE EULER EQUATIONS

door

A.C. Berkenbosch

1. Beschouw een beginwaarde probleem voor een stelsel hyperbolische behoudswetten met bronterm

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{u}(x,t) + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}(x,t)) &= \mathbf{q}(\mathbf{u}(x,t)), \\ \mathbf{u}(x,0) &= \mathbf{u}^0(x), \quad \forall\, x \in \mathbb{R}. \end{aligned}$$

Stel dat we de oplossing numeriek benaderen met het conservatieve differentie schema

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{\Delta t}{\Delta x}(\mathbf{F}_{i+1/2}^n - \mathbf{F}_{i-1/2}^n) + \Delta t \mathbf{Q}_i^{n-j},$$

waar $\mathbf{F}_{i+1/2}^n := \mathbf{F}(\mathbf{U}_{i+k}^n, \ldots, \mathbf{U}_{i-k+1}^n)$, $\mathbf{Q}_i^{n-j} := \mathbf{Q}(\mathbf{U}_{i+l}^{n-j}, \ldots, \mathbf{U}_{i-l}^{n-j})$, $k, l \in \mathbb{N}$ en $j \in \{0, 1\}$. Stel dat $\mathbf{F}$ een Lipschitz continue functie is en $\mathbf{Q}$ een continue functie is en neem aan dat $\mathbf{F}(\mathbf{u}, \ldots, \mathbf{u}) = \mathbf{f}(\mathbf{u})$ en $\mathbf{Q}(\mathbf{u}, \ldots, \mathbf{u}) = \mathbf{q}(\mathbf{u})$.

Dan geldt: als de numerieke oplossing convergeert in L_1^{loc}-zin voor $\Delta t \to 0$, dan convergeert zij naar een zwakke oplossing van de behoudswet.

2. Het verkrijgen van een numerieke detonatiegolf met de fysisch correcte snelheid wordt ten onrechte gebruikt als reden om geavanceerde (en vaak complexe) methoden als 'front tracking' of 'subcell resolution' te gebruiken.

3. Bij het ZND-model voor detonatiegolven kan 'Jouguet's rule' als entropieconditie gebruikt worden om de fysisch relevante zwakke oplossingen te karakteriseren.

4. Zij $r > 0$ en beschouw n getallen $z_i \in \mathbb{C}$ met $|z_i| < 1$ voor $i = 1, \ldots, n$ en $z_i \neq z_j$ als $i \neq j$. Dan is de $n \times n$-matrix A met matrixelementen a_{ij} gedefinieerd door

$$a_{ij} = (1 - \bar{z}_i z_j)^{-r},$$

positief definiet.

5. Laat de afbeeldingen $S : C(\mathbb{R}) \to C(\mathbb{R})$ en $T : C(\mathbb{R}) \to C(\mathbb{R})$ gedefinieerd zijn door

$$(Sf)(t) = \int_0^t f(\tau)\mathrm{d}\tau \quad \text{en} \quad (Tf)(t) = \int_1^t f(\tau)\mathrm{d}\tau.$$

Dan geldt $R(S^k) \subseteq C^k(\mathbb{R})$, $R(T^k) \subseteq C^k(\mathbb{R})$ en $R(S^k) + R(T^k) = C^k(\mathbb{R})$ voor alle $k \in \mathbb{N}$.

6. Een tekort aan zuurstof in de inademingslucht heeft naast een stimulerende ook een deprimerende werking op de longventilatie (zie [1]).

[1] A. Berkenbosch, A. Dahan, J. DeGoede en I.C.W. Olievier, *The ventilatory response to* CO_2 *of the peripheral and central chemoreflex loop before and after sustained hypoxia in man*, J. Physiol. **456** (1992), pp. 71-83.

7. Er zijn in Nederland te veel universiteiten waar een studie wiskunde gevolgd kan worden.

8. Bij de aanstelling van een promovendus dient een vast bedrag gereserveerd te worden voor het bezoeken van (inter)nationale congressen.

9. Het gebrek aan daadkracht bij parlementariërs komt tot uitdrukking in de exorbitante bedragen die bij grootschalige projecten zoals de Betuwelijn worden uitgetrokken voor mitigerende maatregelen.

10. De komst van e-mail heeft de drempel voor communicatie te laag gemaakt.

11. De belofte in een reisgids dat men na een reis in een slaapbus uitgerust op de vakantiebestemming arriveert, ontstijgt alleen het niveau van een flauwe grap wanneer deze belofte vergezeld gaat van een 'niet goed geld terug garantie'.